UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
CENTRO DE BIOTECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA
CELULAR E MOLECULAR

Felipe Castro Nepomuceno

Análise Estrutural e Conformacional de Carboidratos Depositados no Protein Data Bank

Porto Alegre
2019

Felipe Castro Nepomuceno

Análise Estrutural e Conformacional de Carboidratos Depositados no Protein Data Bank

Dissertação submetida ao Programa de Pós-Graduação em Biologia Celular e Molecular do Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul como parte dos requisitos necessários para a obtenção do grau de Mestre em Biologia Celular e Molecular.

Orientador: Hugo Verli

Porto Alegre

2019

Nepomuceno, Felipe Castro
Análise Estrutural e Conformacional de Carboidratos Depositados no Protein Data Bank/ Felipe Castro Nepomuceno. – Porto Alegre, 2019-
117 f.

Orientador: Hugo Verli

Dissertação (Mestrado) – Universidade Federal do Rio Grande do Sul, Centro de Biotecnologia do Estado do Rio Grande do Sul, Programa de Pós-Graduação em Biologia Celular e Molecular, Porto Alegre, BR-RS, 2019.

1, Carboidratos. 2, Protein Data Bank. 3, Glicobiologia Estrutural. 4, Metadinâmica. I. Verli, Hugo, orient. II. Título

Felipe Castro Nepomuceno

Análise Estrutural e Conformacional de Carboidratos Depositados no Protein Data Bank

Dissertação submetida ao Programa de Pós-Graduação em Biologia Celular e Molecular do Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul como parte dos requisitos necessários para a obtenção do grau de Mestre em Biologia Celular e Molecular.

Banca Examinadora:

Prof. Hugo Verli
Orientador

Prof. Dr. Márcio Dorn
Instituto de Informática - UFRGS

Prof. Dr. José Fernando Ruggiero Bachega
Departamento de Farmacociências - UFCSPA

Dr. Pablo Ricardo Arantes
PPG-Bio - UFCSPA

Prof. Dr. Rodrigo Ligabue-Braun (Suplente)
Departamento de Farmacociências - UFCSPA

Porto Alegre, 1 de Outubro

2019

Agradecimentos

Nessa seção dedicada a agradecimentos, gostaria de agradecer a todos que fizeram parte da minha jornada durante a graduação e mestrado.

Primeiramente, agradecer o Prof. Dr. Hugo Verli, que me recebeu em seu laboratório e confiou em mim com projetos autônomos desde o início permitindo meu desenvolvimento desde então. Graças a sua orientação, conversas e ensinamentos, sempre me apontando o caminho, me tornei o cientista que hoje sou. Muito obrigado!

Agradeço também à minha comissão de acompanhamento, Prof. Dr. Rogério Margis e Prof. Dr. Rodrigo Ligabue-Braun, pelo acompanhamento do projeto durante seu desenvolvimento e pelo amparo em situações em que precisei de auxílio.

Meus agradecimentos aos membros da banca, por terem aceitado esse convite.

À secretaria do PPGBCM, em especial à Silvinha, pela enorme ajuda, paciência, carinho e atenção, sempre que precisei durante todo o processo do mestrado, sempre com um sorrisão no rosto.

À UFRGS, ao Centro de Biotecnologia e ao Programa de Pós-Graduação em Biologia Celular e Molecular pela excelência do ensino que permitiu o meu desenvolvimento acadêmico, mesmo em tempos onde o incentivo à ciência é escasso.

Um agradecimento especial aos colegas, ex-colegas e agregados do GBE. Graças a todos, cada um da sua maneira especial e única, pude crescer e me tornar um cientista e pessoa melhor. Meus grandes amigos que vou levar para vida toda, integrantes do "Old-GBE": Conrado, Pablo e Marcelo, pelo enorme apoio e ajuda, com muitas conversas e ensinamentos, sempre com um excelente humor e amizade; Bianca e Fábio, por se tornarem meus grandes amigos de cafézinho, de cuidado da sanidade mental e de almoço, sempre me fazendo companhia; Rodrigo pelo exemplo de pessoa de por todas as ajudas, que muitas vezes pareciam impossíveis; Elisa e Juliano pelos muitos excelentes assuntos, teorias e histórias trazidos para o grupo. Aos integrantes do "New-GBE"também: Crisciele, Laura e Jota, por me ensinarem a ensinar, sempre estarem dispostos a uma conversa, manterem a chama do GBE acesa e carregar esse legado aos demais que virão. À todos vocês, meu sincero muito obrigado e, espero que saibam, que vou levá-los no coração.

Aos meus amigos de todas as horas: Lúcio, Rodrigo, Arthur, Ian, Manoela e Laura, que sempre que eu precisei de um ombro, ajuda ou mesmo só uma distração,

estiveram lá para me apoiar sem falhar.

Às minhas amadas irmãs mais novas que sempre me motivaram a ser o melhor irmão mais velho possível e em quem me espelho pela coragem e determinação que têm.

Não poderia faltar um agradecimento aos meus pais Adriano, Maura, Roselaine e Rodrigo, minha fonte inesgotável de conselhos, motivação e inspiração. É por causa de vocês que eu cheguei onde estou e, com certeza, vai ser por causa de vocês que chegarei mais longe. Eu amo muito vocês 4 e não tenho palavras pra agradecer o que vocês fizeram e fazem por mim. Muito obrigado do fundo do meu ser.

Por último, mas de maneira alguma menos importante, minha namorada Betina. Eu tenho certeza que só sou a pessoa que sou e só sonho o que sonho por tua causa. Tu me ensinou a sonhar, me tornar uma pessoa melhor, a cuidar de mim e me dar valor. Era chegar em casa e te ver, que o mundo deixava de existir e éramos só nós. Isso foi o que me manteve inteiro. Obrigado pela tua confiança, apoio, amor, diversão, motivação pra continuar e todo o resto. Tu é meu suporte, minha base, meu tudo, eu te amo, meu amor. Obrigado por tudo! Esse trabalho é um pouco teu também.

"Um grande número de estranhos
pode cooperar de maneira eficaz
se acreditar nos mesmos mitos."
Yuval Noval Harari

Resumo

Os carboidratos são bem conhecidos por suas características físico-químicas, biológicas, funcionais e terapêuticas. Infelizmente, sua natureza química impõe sérios desafios para a elucidação estrutural desses fenômenos, prejudicando não apenas a profundidade de nossa compreensão dos carboidratos, mas também o desenvolvimento de novas aplicações biotecnológicas e terapêuticas baseadas nessas moléculas. No passado recente, a quantidade de informações estruturais, obtidas principalmente da cristalografia de raios-X, aumentou progressivamente, assim como sua qualidade. Nesse contexto, o presente trabalho apresenta uma análise global das informações sobre carboidratos disponíveis em todo o principal banco de dados de estruturas cristalográficas, Protein Data Bank (PDB). A partir de estruturas de alta qualidade, fica claro que a maioria dos dados está altamente concentrada em alguns tipos de resíduos, principalmente em suas formas monossacarídicas e com um nível limitado de ramificação. As geometrias adotadas pelas ligações glicosídicas podem estar principalmente associadas aos tipos de ligações em vez dos resíduos, enquanto o nível de distorção dos monossacarídeos, baseado em medições do *puckering*, foi caracterizado, quantificado e localizado em uma paisagem de equilíbrio pseudo-rotacional - não apenas para mínimos locais, mas também para estados de transição. Além disso, foi realizado um ajuste de parâmetros já existentes para simulações de dinâmica molecular de hexopiranose (GROMOS53a6$GLYC$), a fim de descrever corretamente o equilíbrio de conformações presentes nos monossacarídeos *in silico*. Essas análises qualitativas e quantitativas oferecem uma imagem global do conteúdo estrutural de carboidratos no PDB, potencialmente apoiando a construção de novos modelos para fenômenos biológicos relacionados a carboidratos no nível atomístico. Além disso, o cuidado com a representação correta dessas moléculas auxilia em estudos futuros sobre as propriedades terapêuticas e atomísticas dessas importantes biomoléculas.

Palavras-chaves: Carboidratos. Protein Data Bank. Metadinâmica. Glicobiologia Estrutural.

Abstract

Carbohydrates are well known for their physico-chemical, biological, functional and therapeutic characteristics. Unfortunately, their chemical nature impose severe challenges for the structural elucidation of these phenomena, impairing not only the depth of our understanding of carbohydrates, but also the development of new biotechnological and therapeutic applications based on these molecules. In the recent past, the amount of structural information, obtained mainly from X-ray crystallography, has increased progressively, as well as its quality. In this context, the current work presents a global analysis of the carbohydrate information available on the entire Protein Data Bank. From high quality structures, it is clear that most of the data is highly concentrated on a few set of residue types, mainly on their monosaccharidic forms and with a limited level of branching. The geometries adopted by glycosidic linkages can be mostly associated to the types of linkages instead of the residues, while the level of puckering distortion was characterized, quantified and located in a pseudorotational equilibrium landscape - not only to local minima, but also to transitional states. Furthermore an adjustment of already existing parameters for hexopyranoses molecular dynamics simulations (GROMOS53a6$GLYC$) was performed, in order to correctly describe the equilibrium of conformations present in monosaccharides *in silico*. These qualitative and quantitative analyses offer a global picture of carbohydrate structural content on PDB, potentially supporting the building of new models for carbohydrate related biological phenomena at the atomistic level. In addition, the care for the correct representation of these molecules *in silico* aids in future studies regarding the therapeutical and atomistic properties of such important biomolecules.

Keywords: Carbohydrates. Protein Data Bank. Metadynamics. Structural Glycobiology.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

BGC	β-D-Glicose
CV	Variáveis Coletivas (*Collective Variables*
DM	Dinâmica Molecular
fs	Femtossegundo
GAL	β-D-Galactose
GLC	α-D-Glicose
IUPAC	União Internacional de Química Pura e Aplicada (*International Union of Pure and Applied Chemistry*)
LAT	β-Lactose
LJ	Lennard-Jones
MM	Mecânica Molecular
NAG	N-Acetil-β-D-Glicosamina
NOE	*Nuclear Overhauser Effect*
NPT	*Ensemble* Isotérmico-Isobárico (número de partículas, pressão e temperatura constantes)
ns	Nanossegundo
NVT	*Ensemble* Canônico (número de partículas, volume e temperatura constantes)
PDB	Banco de dados de estruturas proteicas (*Protein Data Bank*)
PME	*Particle Mesh Ewald*
ps	Picossegundo
QM	Mecânica Quântica (*Quantum Mechanics*)
RMN	Ressonância Magnética Nucelar
SEL	Superfície de Energia Livre

Lista de símbolos

α	Alpha
Å	Ângstrom
β	Beta
μ	Micro
ϕ	Phi
ψ	Psi
σ	Sigma
θ	Theta
J	Joules
K	Kelvin

Sumário

1 Introdução

1.1 Carboidratos

Atualmente, são conhecidas 4 tipos de biomoléculas: Lipídeos, Proteínas, Ácidos Nucleicos e Carboidratos. Carboidratos, além de serem os mais abundantes na biosfera terrestre, se diferem das demais classes em dois importantes aspectos: podem ser altamente ramificados e as suas unidades monoméricas podem ser ligados entre si por diversos tipos de ligações [1]. Ademais, carboidratos (também chamados de glicanas ou açúcares) possuem centenas de unidades monoméricas na natureza, uma variedade muito maior do que os 22 aminoácidos naturais e as 5 bases nitrogenadas dos ácidos nucleicos [2,3]. Os monossacarídeos (monômeros de glicanas, suas unidades mais simples) possuem diferentes configurações (D e L), equilíbrios pseudo-rotacionais e estados anoméricos (α e β) [2,4].

Os açúcares, além disso, podem adotar diversos níveis organizacionais, sendo encontrados na natureza tanto de maneira monomérica, quanto organizado em pequenas (oligossacarídeos) e grandes cadeias de carboidratos (polissacarídeos). A partir disso, mais níveis de complexidade são adicionados às suas estruturas, devido a existência de diferentes tipos de ligações entre monômeros e diversos valores de ângulos diedrais que podem ser adotados por elas [2]. Isso por si só eleva a complexidade estrutural atrelada a essas biomoléculas, tanto no nível polimérico quanto no nível monomérico, fazendo com que essas biomoléculas sejam capazes de desempenhar inúmeras funções biológicas [5] e possuam uma grande diversidade estrutural e conformacional.

1.1.1 Diversidade Funcional e Estrutural

Carboidratos, de uma maneira geral, possuem uma diversidade estrutural muito elevada. Mesmo no seu nível mais simples, os monossacarídeos, existem centenas de moléculas que variam de acordo com o número de átomo de carbonos presentes (trioses, tetroses, pentoses, hexoses, e assim por diante), a orientação dos substituintes (axial ou equatorial), os estados anoméricos, configurações (D ou L), e formas tautoméricas [4]. A grande maioria desse monômeros são encontrados na configuração D e sendo compostos por 6 átomos de carbono (D-hexoses) [6]. Monossacarídeos podem ser classificados, também, de acordo com o tipo de grupamento carbonílico presente na molécula (aldose para grupamento aldeído e cetose para grupamento cetona) [7].

Além disso, monossacarídeos podem se organizar em polímeros de diferentes complexidades, desde a lactose, um dissacarídeo de β-galactose e β-glicose, até a celulose, um polímero de inúmeras repetições de β-glicose (figura 1). Esses diferentes níveis organizacionais adicionam mais graus de liberdade à informação estrutural de carboidratos, devido à presença de distintos tipos de ligação glicosídica entre os monômeros (por exemplo, β-(1→4) e α-(1→4)) e os ângulos diedrais dessas ligações, phi (ϕ) e psi (ψ), que são adotados por elas [2]. Essa diversa gama de variáveis estruturais atreladas aos carboidratos permite que eles desempenhem uma grande variedade de funções, o que os torna moléculas extremamente dinâmicas (figura 1).

Quando em meio biológico, carboidratos desempenham um amplo espectro de funções, muitas delas ligadas à glicosilação de proteínas, como resposta imune, sinalização celular, enovelamento de proteínas, tráfego celular, entre outras. As formas complexas, funcionalidade, e propriedades dinâmicas de oligo- e polissacarídeos os permitem funcionar em interações intermoleculares como codificadores de informação biológica [8]. Por exemplo, o reconhecimento de carboidratos é uma parte integral do desenvolvimento biológico natural [9] e da defesa imune contra patógenos externos através da identificação de glicanas exógenas [10, 11]. Em contrapartida, bactérias e patógenos virais aderem, inicialmente, ao tecido do hospedeiro através de ligações especificamente com carboidratos presentes na superfície das células do hospedeiro [12].

Outra função importante desempenhada por carboidratos é de estrutura. Grandes cadeias poliméricas de carboidratos, como a celulose, que é composta por várias β-glicoses conectadas por ligações glicosídicas β-(1→4), é um dos principais componentes das paredes celulares de plantas e compõe grande parte da biomassa terrestre [13, 14]. Da mesma forma, a quitina, um polímero de N-acetil-glicosamina, desempenha o papel primordial na formação de componentes estruturais da parede celular de fungos e leveduras, bem como do citoesqueleto de artrópodes [15]. Além disso, carboidratos podem auxiliar na manutenção de determinadas estruturas proteicas, como a estrutura quaternária da imunoglobulina G, ou no ancoramento de proteínas à membrana celular [1].

Moléculas como o amido e o glicogênio, que servem como repositório de energia e fonte de carbono para, virtualmente, todos os organismos vivos, também são carboidratos [16]. O glicogênio, que é sintetizado durante condições de abundância para ser utilizado em momentos de necessidade energética, é formado pela polimerização da α-glicose através de ligações α-(1→4) com eventuais ramificações α-(1→6) [17]. Já o amido é composto por dois polímeros de carboidratos, a amilose e a amilopectina, que são moléculas praticamente lineares (ligações α-(1→4) com poucas ramificações em α-

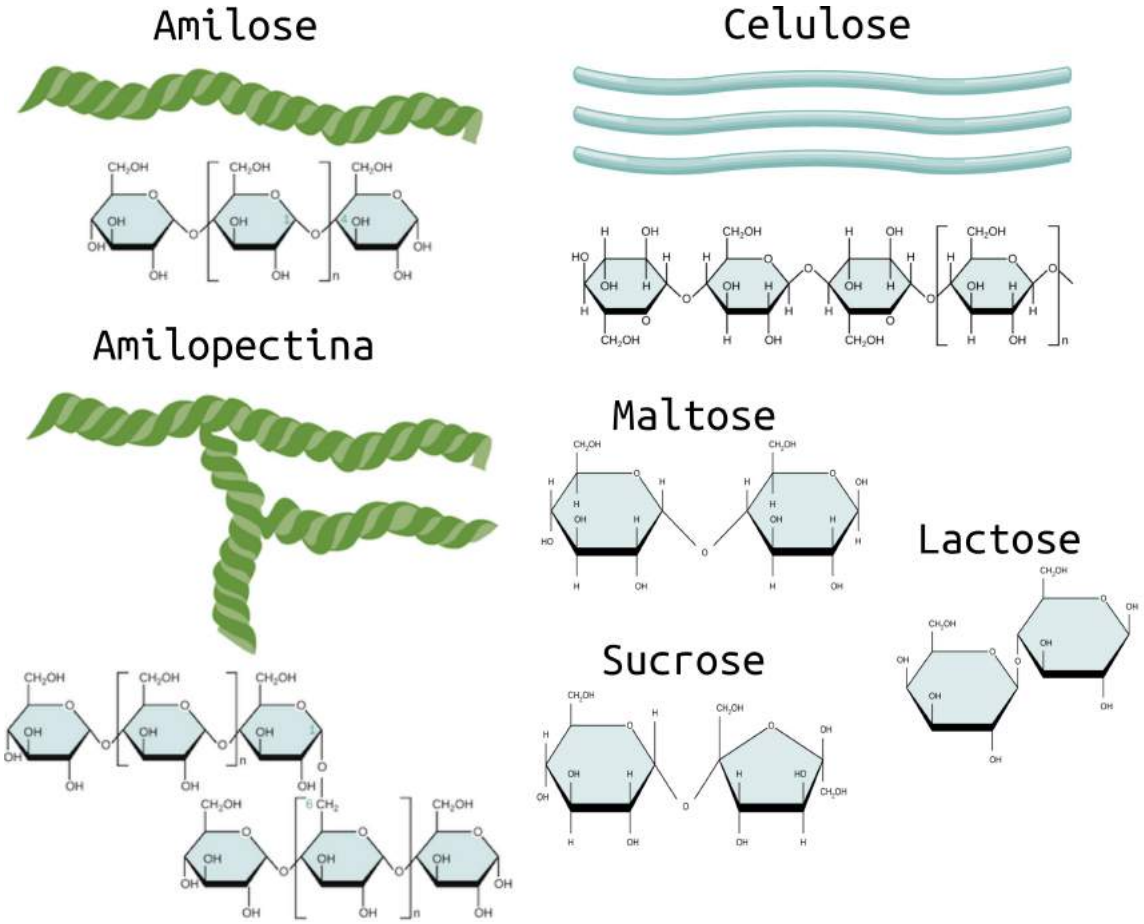


Figura 1 – Esquema ilustrando a variedade de estruturas que podem ser adotadas por carboidratos

(1→6)) e altamente ramificadas (curtas cadeias de ligações α-(1→4) ligadas entre si por ligações α-(1→6)) respectivamente [18].

1.1.2 Diversidade Conformacional

Além da diversidade estrutural e funcional apresentada anteriormente, no que se trata de unidades monoméricas de carboidratos de anéis de 6 membros, há também a variedade de conformações que podem ser adotadas por eles. Esses conformações vem sendo foco de estudos tanto experimental, quanto computacionalmente [19–23], devido ao conhecimento de que o estado em que se encontram influencia, diretamente, o comprimento das ligações e a forma dos cadeias sacarídicas [24]. As conformações mais recorrentes e, mais energeticamente estáveis observadas em solução são as duas formas de cadeiras: $^4C_1$ para D-açúcares (onde o carbono 4 fica acima do plano do anel e o carbono 1 abaixo) e $^1C_4$ para L-açúcares (onde o inverso acontece) (figura 2).

Existem outras conformações que podem ser adotadas. No entanto, elas são

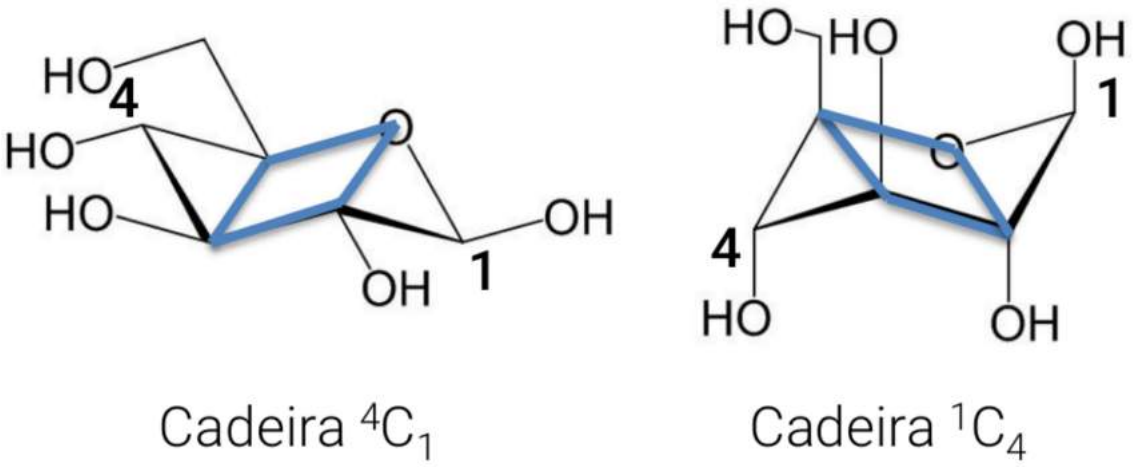


Figura 2 – Ilustração dos planos formados entre os átomos dos anéis de monossacarídeos nas conformações $^4C_1$ e $^1C_4$

energeticamente menos favoráveis e ocorrem, via de regra, entre a transição de uma cadeira para a outra. Sabe-se, adicionalmente, que em cadeias oligossacarídicas, como a heparina, a conformação de alguns monômeros é não apenas uma cadeira, mas também um bote torcido $^2S_0$ [25, 26] (figura 3). Em alguns casos, esse tipo de distorção do anel sacarídico possui uma consequência mecanística favorável na hidrólise de ligações glicosídicas, posicionando o oxigênio glicosídico próximo ao resíduo ácido/básico do sítio catalítico [27, 28].

1.1.2.1 *Puckering*

Os estados conformacionais adotados por piranoses são classificados de acordo com a nomenclatura descrita por Schwartz [29] e adotada pela União Internacional de Química Pura e Aplicada (IUPAC) [30]. As 38 conformações canônicas de *puckering* representam todas as formas únicas que um anel de piranose pode adotar como cadeira (C), envelope (E), meia-cadeira ou *half-chair* (H, às vezes chamado de meio-bote), torcido ou *skew* (S, às vezes chamado de bote-torcido) e bote (B). Para facilitar a discriminação das diferentes conformações dos anéis sacarídicos, foi estabelecida uma maneira quantitativa de avaliar seu grau de dobramento, chamado de *puckering* [31]. Nessa definição, um sistema de coordenadas esféricas foi utilizado para fazer a representação das diferentes conformações de anéis de 6 membros de monossacarídeos, definido por 3 variáveis (θ, ϕ e Q).

O método baseia-se na definição de um único plano médio passando pelo centro geométrico da molécula e, a partir dele, pode-se calcular os ângulos dos átomos que estejam fora desse plano. Para tanto, as variáveis θ e ϕ especificam o "tipo de distor-

ção"presente no anel e Q indica a amplitude com que ela ocorre. Cada conformação possível de ser adotada por um anel de 6 membros é descrita por uma esfera conformacional (figura 3), sendo os pólos dessa esfera as duas conformações mais estáveis (Cadeiras $^4C_1$ e $^1C_4$). Entre elas, existem conformações de transição, que são adotadas de acordo com a via de interconversão de cada monossacarídeo, sendo a diferença de energia entre esses estados conformacionais também variável de acordo com a molécula de estudo [21].

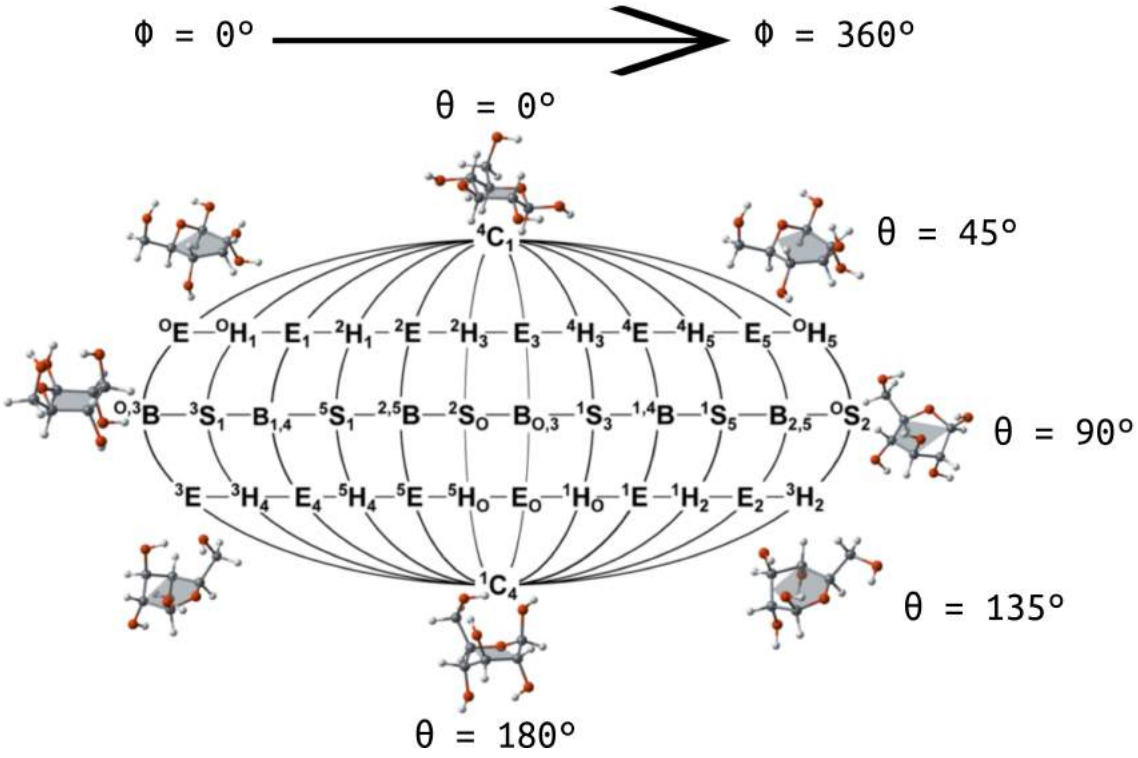


Figura 3 – Esfera de conformações de monossacarídeos em função das coordenadas de *puckering* θ e ϕ. Adaptado de [21]

1.2 Caracterização estrutural de Carboidratos

A área de estudo envolvendo a determinação de estruturas de biomoléculas surgiu em meados do śeculo XX e desde então vem agregando contribuições expressivas para o avanço científico [32]. O estudo da estrutura de macromoléculas é um ponto crucial no entendimento da biologia, visto que funções biológicas são consequência de interações moleculares que estão diretamente ligadas à estrutura macromolecular [33]. Nesse contexto, novas técnicas de elucidação estrutural vêm sendo desenvolvidas com o propósito de incrementar o conhecimento estrutural de entidades biológicas em diferentes níveis de resolução. Técnicas como o espalhamento de raios-X a baixos ângulos [34]

e o dicroísmo circular [35] são capazes de descrever forma, tamanho, conteúdo de estrutura secundária e enovelamento proteico, porém não possuem a resolução atomística necessária para a descrição de moléculas menores, como carboidratos.

A determinação de uma estrutura exata de carboidratos é, muitas vezes, um desafio, porque mesmo no seu nível monomérico, existem diferenças de configuração, estereoquímica e conformações, resultando em dados altamente similares [36]. As principais técnicas utilizadas para a determinação estrutural macromoléculas biológicas são a cristalografia de raios-X e a ressonância magnética nuclear (figura 4), sendo a primeira a mais dominante. No Protein Data Bank (PDB) [37], o principal banco de dados de estruturas tridimensionais de biomoléculas, das 156.101 estruturas depositadas, cerca de 89% foram resolvidas utilizando a cristalografia de raios-x (acessado em setembro de 2019). Nesse banco de dados, no entanto, grande parte das estruturas de carboidratos presentes estão covalentemente ligados à proteínas ou como ligantes não-covalentemente ligados [38].

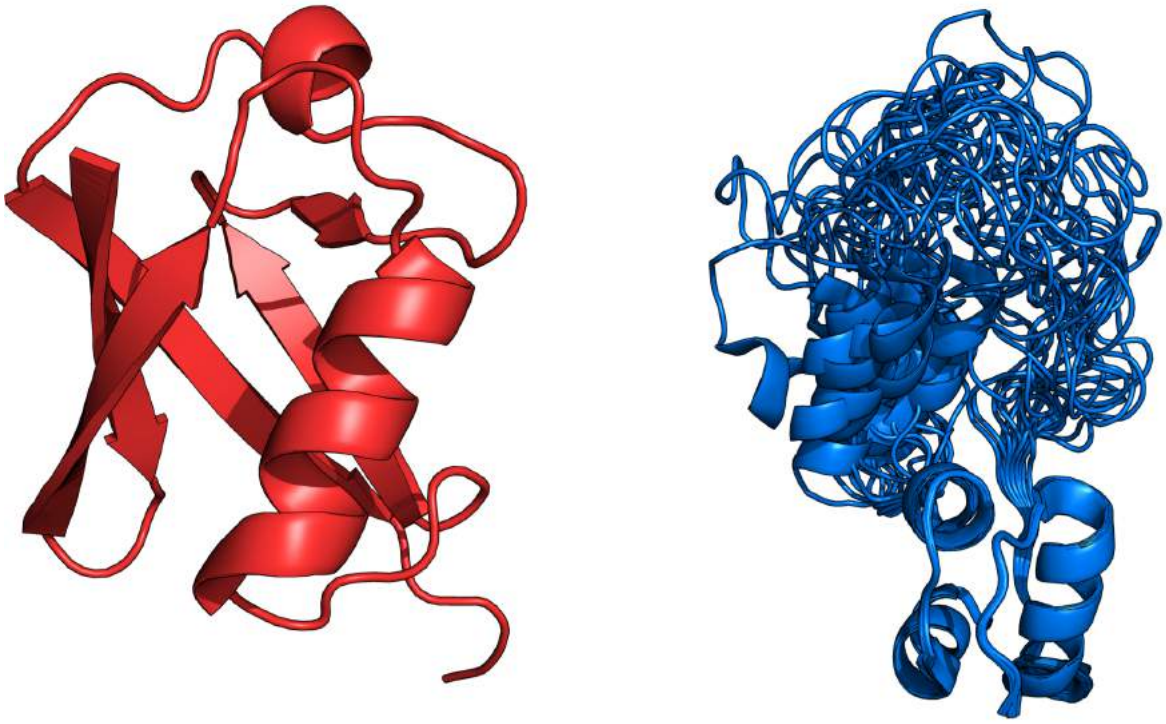


Figura 4 – Estruturas tridimensionais de biomoléculas obtidas pelas duas metodologias mais empregadas atualmente. Em vermelho: Cristalografia de Raios-X - Ubiquitina (ID 1UBQ) [39]. Em azul: Ressonância Magnética Nuclear - Proinsulina (ID 2KQP) [40].

1.2.1 Cristalografia de Raios-X

A cristalografia de raios-X é uma técnica usada para elucidar a estrutura atômica e molecular de um cristal, onde um feixe incidente de raios-X é difratado por essa estrutura cristalina. Medindo-se a intensidade e o ângulo com os quais esses raios são difratados, pode-se reconstruir o perfil tridimensional de densidade eletrônica do cristal. A partir dessa densidade eletrônica, a posição média dos átomos do cristal pode ser determinada, assim como outras propriedades, tais quais ligações químicas, comprimento de ligações, etc. Essa é a principal técnica aplicada para a elucidação estrutural de macromoléculas biológicas e ofereceu uma importante contribuição no entendimento de diversos processos e mecanismos biológicos [41].

Devido ao seu alto impacto, essa técnica vem sido cada vez mais desenvolvida e aprimorada. No entanto, a qualidade da estrutura obtida como resultado está diretamente ligada à qualidade do cristal da macromolécula de interesse, tornando o processo de cristalização o principal gargalo dessa técnica [42]. A ocorrência da cristalização é dada em condições controladas, incluindo uma solução supersaturada da molécula de interesse, agentes precipitantes, condições específicas de temperatura, força iônica e em pequenos intervalos de variação de pH [43].

Apesar da sua aplicação bem difundida, essa técnica conta com algumas limitações [44]: (i) A formações de cristais da molécula de interesse, devido à multiplicidade de fatores que participam no processo; (ii) A indução de estabilidade conformacional pelo ambiente cristalino, que pode acarretar erros na estrutura final; (iii) A impossibilidade de obter-se um conjunto de conformações, sendo a estrutura resolvida uma conformação média. Devido a isso, a aplicação dessa metodologia juntamente com outras técnicas de elucidação estrutural (como RMN ou Crio-Microscopia Eletrônica), garantem a qualidade do estrutura final obtida e permitem a detecção de artefatos que possam estar contidos no cristal [45].

1.2.2 Ressonância Magnética Nuclear

A ressonância magnética nuclear é uma técnica de espectroscopia que tem como princípio observar o campo magnético ao redor de núcleos atômicos. Nessa técnica, uma amostra é colocado em um campo magnético e o sinal de RMN é produzido pela excitação dos núcleos da amostra por ondas de rádio, que são detectadas por radio-receptores sensíveis. A partir disso, é possível identificar a presença de núcleos de átomos que estejam espacialmente próximos, ou seja, na vizinhança um outro, gerando mapas de restrição espacial que são a base para a construção dos modelos tridimensional dessas moléculas. O efeito Overhauser nuclear (NOE, *nuclear Overhauser effect*) - que ocorre

quando átomos estão próximos em estados de magnetização diferentes - é empregado em espectros de resolução bidimensional [46] com intuito de detectar a proximidade entre átomos não-covalentemente ligados (distância menor que 5 Å) [47].

Devido ao fato de as amostras serem submetidas à espectroscopia de RMN em solução, essa metodologia oferece uma caracterização que permite observar a flexibilidade da estrutura e os diferentes estados conformacionais das biomoléculas [48]. Nesse contexto, a RMN é a principal técnica para a elucidação estrutural de carboidratos, dada a sua alta flexibilidade e dinâmica. A partir de dados de RMN foi possível observar e compreender o complexo processo de equilíbrio conformacional e de interconversão entre estados anoméricos dessas monossacarídeos [3, 4]. No entanto, dada a baixa utilização dessa técnica para a obtenção de modelos tridimendionais depositados no PDB e a dificuldade de analisar as estruturas resultantes (compostas por vários diferentes modelos que são observados em solução), a informação estrutural de carboidratos nesse banco de dados baseia-se principalmente na cristalografia de raios-X.

1.2.3 Acurácia das Estruturas de Carboidratos

No servidor do PDB [49], o nível de resolução e de acurácia de anotação das estruturas contendo carboidratos varia amplamente [50]. A quantidade de informação e dados experimentais a respeito da estrutura de carboidratos e sua composição variam de acordo com o tipo de açúcar (isto é, mono-, oligo- e polissacarídeos, glicosaminoglicanos, glicoproteínas ou outros compostos glicoconjugados), o grande número de isômeros e confôrmeros, a complexa interação entre diferentes forças condutoras afetando as populações de confôrmeros em equilíbrio e a dificuldade de extrair informação inequívoca de bancos de dados experimentais [51].

A cristolografia de raio-X, a principal metodologia empregada na resolução de estruturas depositadas no PDB, não funciona muito bem com sistemas altamente flexíveis, já que requer cristais regulares e apenas alguns oligossacarídeos não-derivativos cristalizam. Em suma, a complexidade estrutural das cadeias glicanas, sua micro heterogenicidade, flexibilidade e a não disponibilidade desses compostos em quantidades suficientes prejudicam os estudos estruturais [52] e uma compreensão mais detalhada da dinâmica de estruturas sacarídicas em solução permanece um desafio.

Preocupações sobre a acurácia de estruturas cristalográficas contendo carboidratos foram levantadas, anteriormente, por diferentes autores [53–56]. Quando moléculas tão pequenas estão presentes em estruturas macromoleculares, elas, frequentemente, são reportadas com erros estereo- e regioquímicos e em conformações incomuns [57]. Apesar de que distorções conformacionais podem resultar de interações ocorrendo em um com-

plexo [58], elas também podem ser decorrentes de anotações errôneas dessas moléculas. Isso resulta em um mau entendimento químico e na falta de restrições estereoquímicas apropriadas no refinamento, geralmente contra dados com baixa resolução [59].

Ademais, em contraste com proteínas e ácidos nucleicos, não há uma nomenclatura padrão para resíduos de carboidratos em arquivos PDB. [60]. Nesse formato, a nomenclatura do resíduo é determinada por apenas 3 letras, o que é o suficiente para os 22 aminoácidos que compõe proteínas, porém insuficiente para codificar centenas de diferentes monossacarídeos. A abreviação utilizada para representar essas moléculas, portanto, frequentemente não é relacionada ao nome comum do resíduo [50]. Além disso, em alguns casos, cadeias sacarídicas inteiras (oligossacarídeos e glicoconjugados) são representados pelo mesmo código de três letras, apesar de serem compostos por diferentes monômeros. Por exemplo, os resíduos GAL (β-galactose) e BGC (β-glucose) são ambos denominados LAT (β-Lactose) em entradas do PDB onde essa moléculas está presente [50, 61].

1.3 Métodos para Simulação de Carboidratos

Existem métodos computacionais capazes de auxiliar o estudo estrutural dessas moléculas altamente diversas. Uma dessas técnicas é a dinâmica molecular (DM), que foi desenvolvida na década de 1970 [62] e vem sendo empregada para abordar diversos problemas estruturais [63–66]. Além da dinâmica, pode-se utilizar uma técnica de amostragem ampliada chamada metadinâmica [67] (uma simulação de DM enviesada), a qual permite que conformações, que não seriam alcançadas no tempo de simulação limitado da DM, sejam estudadas. A aplicação de ambas técnicas permite que carboidratos sejam estudados sem que haja a dependência da qualidade das estruturas resolvidas. Esse movimento vêm ocorrendo, onde são realizados estudos estruturais de carboidratos aplicando técnicas computacionais, observando tanto mudanças conformacionais, preferências de ângulos diedrais, dinâmica dessas moléculas livres e em interação com outras biomoléculas [25, 51, 68, 69].

1.3.1 Dinâmica Molecular

O método de dinâmica molecular (DM) permite que sejam realizados experimentos com resolução atomística dos fenômenos estudados, além da observação de propriedades tempo-dependentes. A DM avalia o comportamento físico-químico (baseado em um conjunto de parâmetros chamados campo de força) de um determinado sistema de átomos, baseando-se em forças que podem induzir a movimentação desses

átomos. Essa técnica baseia-se em integrações sucessivas da equação de movimento de Newton aplicada a todos os átomos do sistema, empregando um tempo de integração denominado *dt* [70]. Esse processo é realizado para todas as posições e interações dos átomos por um determinado período de tempo, sendo atualizados constantemente e formando o que é chamado de trajetória. Nessa trajetória, está contido o conjunto de diferentes conformações adotadas pela molécula simulada (*ensemble*) que descrevem o perfil conformacional encontrado para tal sistema [33].

A DM está diretamente atrelada ao conceito de Mecânica Molecular (MM) onde apenas os núcleos dos átomos são levados em consideração nos cálculos, excluindo a descrição direta de propriedades eletrônicas e, por conseguinte, a possibilidade da ocorrência de quebra e formação de ligações químicas. Dessa maneira, a DM é regida pela seguinte equação: $F_i = m_i \cdot a_i$. O termo a*i* pode ser descrito como $a_i = d^2 r_i t_{i(t)} / dt^2$, onde $r_{i(t)}$ corresponde à posição do átomo *i* e *d*t ao tempo de integração discretizado [70]. A força do sistema também pode ser descrita de acordo com o termo: $F_i = -dV/dr_i$, onde V representa a energia potencial do sistema (calculada para todo átomo *i* na posição r_i). Isso é primordial na determinação da intensidade e direção de F_i, visto que esse termo é colocado em função das coordenadas cartesianas dos átomos e da energia potencial (V) do sistema. Sendo assim, integra-se a equação considerando a variação temporal (t); as forças que agem sobre o átomo i, calculadas pelo conjunto de interações descritas para esse átomo (campo de força); a partir dessas forças, ocorre uma variação no espaço cartesiano (*d*r), em função do tempo de integração (no presente caso, 2 fs). Resolvendo-se essa equação para cada átomo do sistema, obtém-se a trajetória, sobre a qual o comportamento do dado sistema durante um determinado período de tempo de simulação será avaliado [70].

Os cálculos de energia potencial dos átomos que compõe os sistemas simulados são realizados com base em um conjunto de funções que definem os chamados termos ligados e termos não-ligados (figura 5). Esse conjunto de parâmetros e funções é chamado de campo de força e a forma funcional básica desses campos de força é dada pelo somatório desses termos [70]. Dentre os termos ligados estão: termos de ligação covalente entre dois átomos ($V_{ligação}$), descritos por um potencial harmônico (k_b) que penaliza energeticamente variações de *b* em relação ao valor de referência b_0; termos de descrição de ângulos envolvendo três átomos ($V_{ângulo}$), que de forma semelhante emprega um potencial harmônico (k_θ) para evitar variações excessivas, onde θ é o ângulo calculado e θ_0 é o ângulo de referência; termos torcionais relativos à rotação de ângulos diedrais formados entre quatro átomos ($V_{diedrais}$). O diedro próprio é descrito por uma função cosseno com *n* sendo o valor de periodicidade ou multiplicidade (indicando o número de mínimos de energia), ϕ é o valor do diedro, k_ϕ determina a barreira energética

para que ocorra mudança de mínimo de energia e δ indica o máximo de energia no perfil rotacional do diedro. O diedro impróprio varia para cada campo de força, podendo tanto seguir o padrão periódico encontrado para diedros próprios (AMBER e OPLS), como o padrão de potenciais restritivos encontrado para ligações e ângulos (CHARMM e GROMOS). Os termos não-ligados compõem a soma de dois outros termos: os termos coulômbicos relativos a componentes eletrostáticos dos átomos (V_{eletro}), calculados como a soma de interações entre pares de cargas atômicas, utilizando a Lei de Coulomb; os termos de van der Waals (V_{LJ}), calculados aplicando um potencial de Lennard-Jones 12-6 entre pares de átomos neutros. Apesar dessa forma básica, cada campo de força tem suas peculiaridades para lidar com a resolução dessas equações. Atualmente, os campos de força mais comumente utilizados são o AMBER [71], CHARMM [72], GROMOS [73] e OPLS-AA [74]. Ademais, o Grupo de Bioinformática Estrutural desenvolveu um conjunto de parâmetros, complementar ao campo de força GROMOS, específicos para hexopiranoses [75–77], permitindo a aplicação da DM a esse grupo de biomoléculas.

1.3.2 Metadinâmica

Nas simulações de dinâmica molecular, os resultados obtidos são significantes a partir do momento que há um tempo de simulação extenso o suficiente para que o sistema visite todos as configurações energeticamente relevantes. Em alguns casos, dependendo do tamanho do sistema e da velocidade com que mudanças conformacionais ocorrem e da barreira energética entre elas, o tempo de simulação pode se tornar um fator limitante [79]. Nesse contexto, surge a metadinâmica, uma técnica de amostragem ampliada comumente utilizada para descrever os perfis de energia livre de sistemas capaz de acelerar eventos raros que são descritos por Hamiltonianos complexos, tanto em um nível clássico (MM) quanto em um nível quântico (QM) [80].

Essa metodologia é baseada na inclusão de variáveis coletivas (Collective Variables - CV) [81] durante os cálculos de dinâmica, que devem ser previamente identificadas como capazes de descrever o processo de interesse [80]. Dessa forma, o espaço conformacional que será adotado pelo sistema é ditado por essas variáveis, de maneira que potenciais de energia são incluídos ao longo do tempo de simulação e o sistema fica impedido de retornar a estados conformacionais já adotados, permitindo assim que a superfície de energia livre de um determinado processo em função de uma determinada variável seja completamente populada (figura 6).

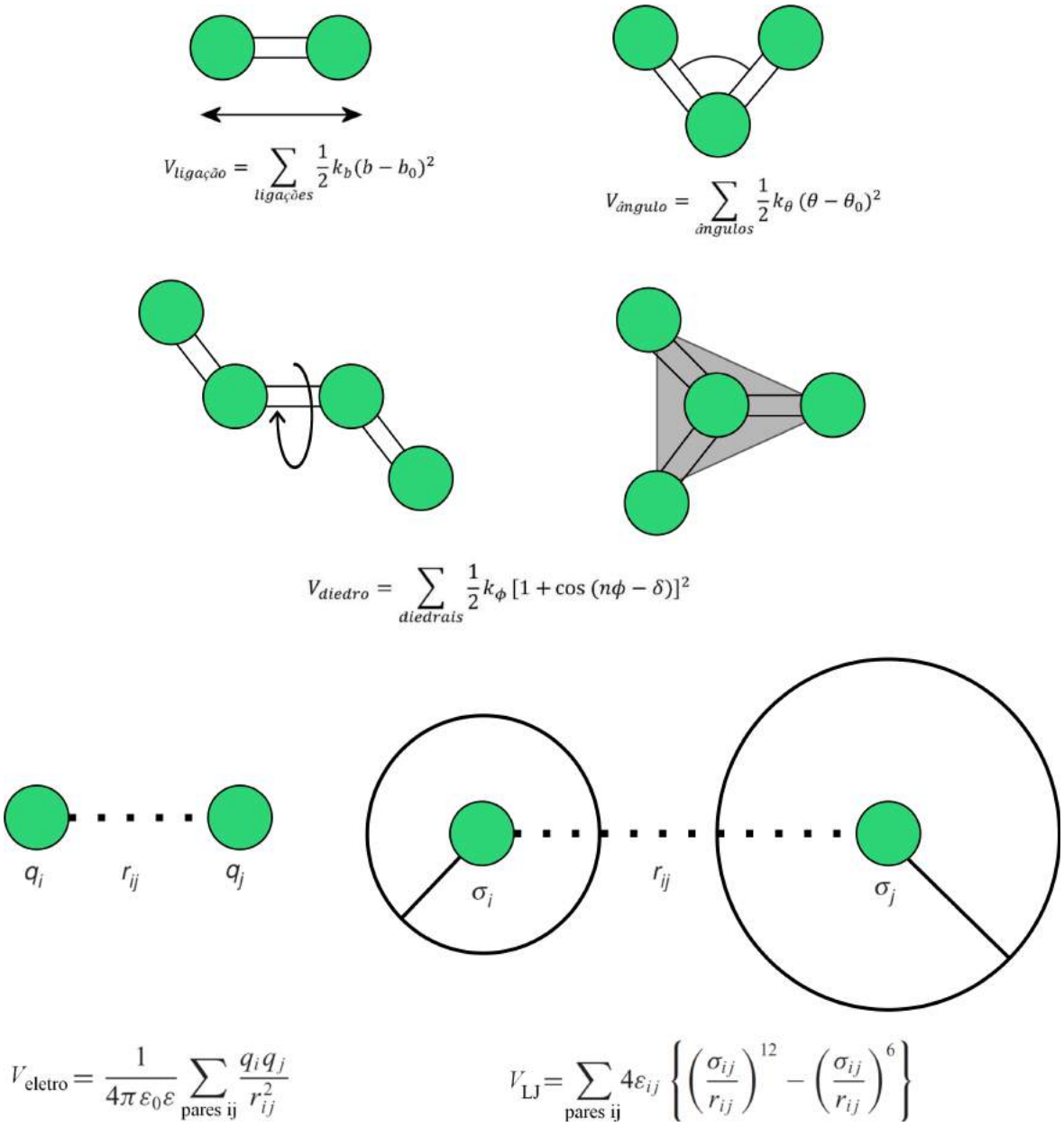


Figura 5 – Esquema representativo das funções que descrevem potenciais que regem o funcionamento de um campo de força. Adaptado de Pedebos [78]

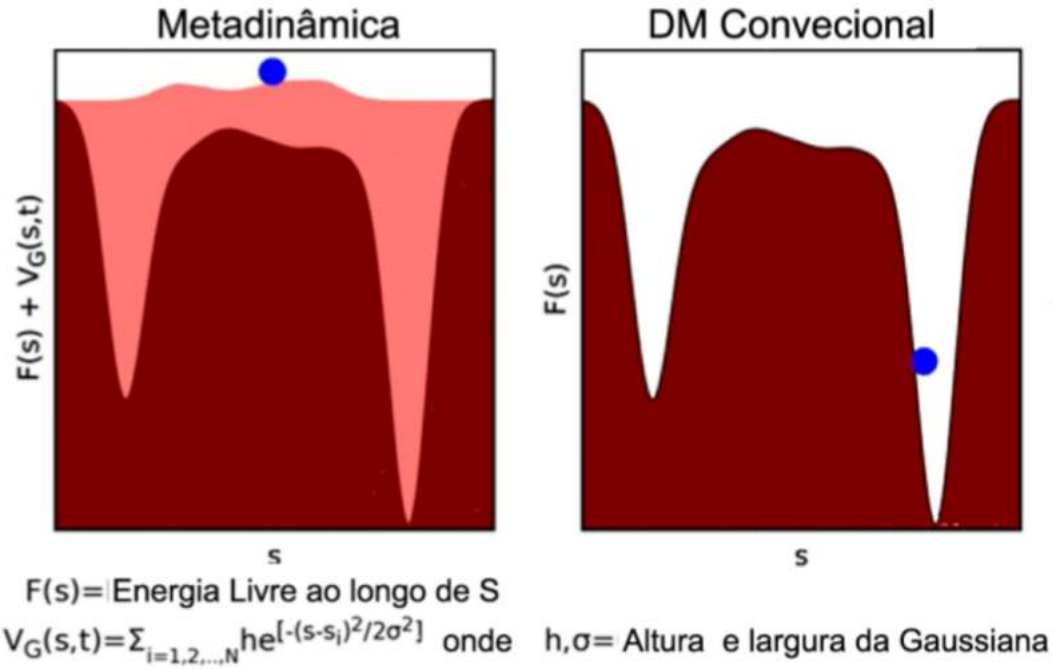


Figura 6 – Representação do cálculo de metadinâmica e a equação aplicada ao sistema durante a simulação.

2 Justificativa

Carboidratos, por possuírem um espectro de funções tão diverso, tem sido o crescente alvo de estudos. A sua participação em processos de sinalização e em glicosilações de proteínas os tornam alvos interessantes para o desenvolvimento de fármacos. No entanto, devido a sua elevada heterogenicidade e flexibilidade, a dinâmica estrutural dessas biomoléculas não é inteiramente compreendida.

Estudos baseados em modelos computacionais, que fornecem uma resolução atomística dos fenômenos, são importantes ferramentas para o estudo dessas moléculas. Abordagens desse tipo são um dos poucos métodos que permitem o estudo conformacional de carboidratos, já que tanto o RMN como a cristalografia tem dificuldades ao lidar com esse tipo de biomolécula. No entanto, sua representatividade está diretamente ligada ao correto desenvolvimento do modelo.

Estruturas presentes em bancos de dados, quando levada em consideração suas qualidades, oferecem informações estruturais de carboidratos obtidas experimentalmente. A extração dessas informações de maneira acurada permite o auxílio de um melhor entendimento da glicobiologia. Ademais, garante que o desenvolvimento de modelos para o estudo dessas moléculas sejam o mais acurados possível, reproduzindo o comportamento observado experimentalmente.

3 Objetivos

O objetivo desse trabalho é contribuir com o crescimento do conhecimento referente à glicobiologia estrutural. Para isso, o trabalho conta com os seguintes objetivos específicos:

- Extrair corretamente informações estruturais de carboidratos do PDB, considerando os possíveis erros;
- Identificar os principais monossacarídeos presentes no banco de dados e os seus estados conformacionais de preferência;
- Identificar a abundância dos diferentes tipos de ligação glicosídica e a preferência dos ângulos diedrais adotados;
- Ajustar o valor de barreira energética entre os principais estados conformacionais de monossacarídeos ($^4C_1$ e $^1C_4$) no campo de força GROMOS53a6GLYC [75].

4 Procedimentos metodológicos

Durante o trabalho a nomenclatura dos átomos de todos os resíduos de carboidratos seguiu a lógica representada na figura abaixo.

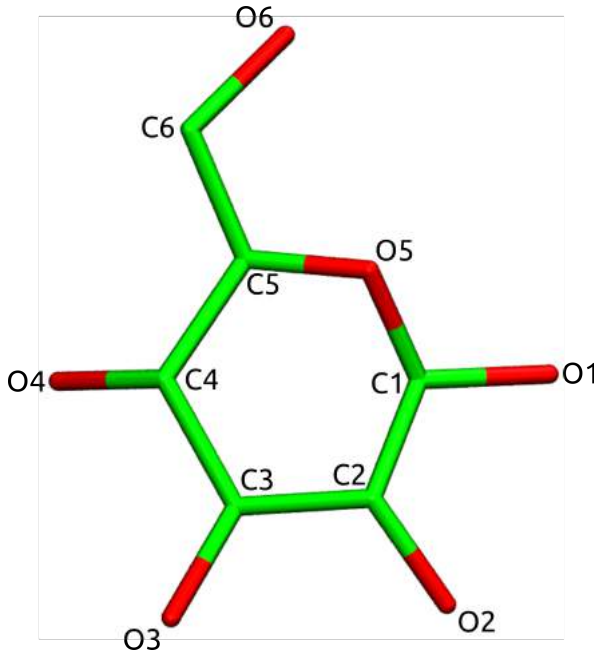


Figura 7 – Nomeclatura padrão dos átomos de anéis de hexopiranoses.

4.1 Análise das Estruturas do PDB

Buscando analisar sistematicamente a informação de carboidratos depositada no PDB, foi implementado um fluxo de trabalho que separa as diferentes entradas em suas múltiplas unidades de informação. Nele, é feita a filtragem por carboidratos, identificação de monossacarídeos (mesmo em estruturas oligossacarídicas), identificação do tipo de ligação presente entre monossacarídeos, bem como a medição de ângulos diedrais e ângulos de *puckering*. Todos esses passos de análises e curadoria seguem a lógica do pseudo-código presente na figura 8.

4.1.1 Obtenção dos dados e Filtragem

Para garantir a análise completa do banco de dados de estruturas, todas as entradas disponíveis nele foram baixadas no dia 28 de Junho de 2018 (totalizando 140.547 arquivos) para passarem, posteriormente, por um processo de filtragem. Tanto o

```
1: Download do Protein Data Bank;
para cada arquivo_pdb:
      2: Filtragem por identificadores_carboidratos;
            se TAG_HETNAM em identificadores_carboidratos:
                  tag_carbo = TAG_HETNAM;
                  seleciona(arquivo_pdb);
      3: Filtro de Resolução;
            se RESOLUTION < 1.5 Å:
                  seleciona(arquivo_pdb);
      4: Identificação de carboidratos e separação da proteína;
            se carbo_tag == TAG_RES_HETATM:
                  arquivo_pdb_carbo = átomos_carboidratos;
para cada arquivo_pdb_carbo:
      5: Identificação da ligação glicosídica (BioPDB);
            encontra(O_ligação);
            procura_vizinhos(O_ligação);
            define(átomos_diedros);
      6: Cálculo dos ângulos diedrais (Φ, Ψ);
            calcula_diedros(átomos_diedros);
      7: Separação de monômeros;
            separa_monômeros(arquivo_pdb_carbo);
para cada arquivo_pdb_monômero:
      8: Cálculo de puckering (θ, Φ, Q);
            calcula_puckering(arquivo_pdb_monômero);
9: Organização dos dados e análise.
```

Figura 8 – Pseudo-código que ilustra a lógica utilizada no fluxo de trabalho na análise de estruturas de carboidratos depositadas no PDB.

download dos arquivos, quanto o restante dos passos dessa seção, foram feitos utilizando um *script* desenvolvido com o auxílio da biblioteca de funções BioPython [82] da linguagem de programação Python. No processo de filtragem, buscava-se selecionar apenas as estruturas que contivessem moléculas de carboidratos, ligadas ou não a proteína cristalizada. Para isso, todos os arquivos foram abertos em forma de texto e, em todos eles, buscou-se por pelo identificador "HETNAM", presente no cabeçalho de arquivos no formato PDB que identifica os heteroátomos do sistema presente no arquivo. Nessa *flag* são definidos os nomes de todos os resíduos do sistema que não pertencem a proteínas e, portanto, permite identificar a presença de carboidratos.

A identificação das biomoléculas de interesse foi feita com base nos nomes dos resíduos que eram considerados heteroátomos, já que átomos pertencentes a proteínas são identificados apenas como "ATOM". Era feita a comparação entre esses nomes e uma série

Tabela 1 – Lista dos nomes, prefixos e sufixos utilizados na filtragem de carboidratos dos arquivos baixados do PDB.

Nome de Carboidrato	Abequose, Arabinose, Fructose, Fucose, Galactosamine, Galactose, Glucosamine, Glucose, Iduronic, Lactose, Mannose, Maltopyranoside, Maltoside, Rhamnose, Xylose.
Sufixo ou Prefixo de Carboidrato	"Ose", "Gluc", "Uronic", "Saccharide".

de nomes, prefixos e sufixos comuns em carboidratos. Para uma entrada ser selecionada (ou seja, ela possuía pelo menos uma estrutura nomeada como um carboidrato dentro do arquivo) as palavras descritas na tabela 1 deveriam ser identificadas em alguma porção do nome do resíduo.

Em adição à filtragem em busca de carboidratos, foi realizada outra etapa de seleção, porém, dessa vez, usando a resolução das estruturas depositadas como guia. Visando garantir a boa qualidade das estruturas a serem analisadas, moléculas com resoluções menores do que 1.5 Å foram deixadas de lado. A resolução da estrutura de uma molécula definida por cristalografia e difração de raios-x representa a menor distância capaz de ser resolvida naquele cristal, ou seja, com resoluções maiores do que 1.5 Å é possível definir, com precisão, a posição de cada átomo dentro de um espaço de 1.5 Å. Para isso, varreu-se os arquivos PDB até o identificador "RESOLUTION"ser encontrado e, caso o valor definido a ele fosse inferior a 1.5 Å, a entrada era armazenada. Todos as estruturas não selecionadas nessa etapa foram salvas separadamente, para realizar algumas das etapas aplicadas no conjunto de dados de interesse para fins comparativos. Estruturas caracterizadas por ressonância magnética nuclear (RMN), por não conterem resolução, por serem uma estrutura média do comportamento de uma molécula visto em solução e por serem uma pequena porcentagem dos dados obtidos, não foram consideradas no conjunto de dados.

4.1.2 Identificação e Isolamento de Carboidratos

Tendo em vista que os cálculos e análises posteriores seriam realizados apenas sobre os átomos dos resíduos de carboidratos, a identificação dessas biomoléculas e o seu isolamento dos demais átomos presentes no sistema (proteínas, co-fatores, etc) fez-se necessário. Nessa etapa, todos os átomos que não fossem correspondentes a um resíduo de carboidrato identificado no passo anterior eram descartados. Isso foi realizado a partir da nomenclatura identificada por "HETNAM". Nela, além do nome de cada um dos resíduos diferentes aos resíduos das proteínas, há também uma *tag* de três letras

que está presente na definição de cada um dos átomos do resíduos (figura 9). Com as *tags* de todos os carboidratos presentes em um arquivo, ele era percorrido comparando cada um dos conjuntos de três letras de cada um dos átomos de todo o sistema. Quando um átomo possuía uma das *tags* referentes aos carboidratos achados, ele era salvo em um novo arquivo, onde estavam contidos apenas os átomos de carboidratos da entrada em questão. Por exemplo, caso em uma entrada fosse identificado a presença de uma N-Acetil-Glicosamina (NAG), apenas os átomos marcados com essa *tag* seriam selecionados e salvos em um novo arquivo. Para cada um dos arquivos selecionados na etapa anterior, um novo arquivo foi gerado, contendo os açúcares encontrados no seu espaço tridimensional. No nome desse arquivo era composto pelo código PDB da entrada original mais uma marcação "_carbo", a fim de demonstrar que, naquele novo arquivo, apenas os carboidratos estavam presentes.

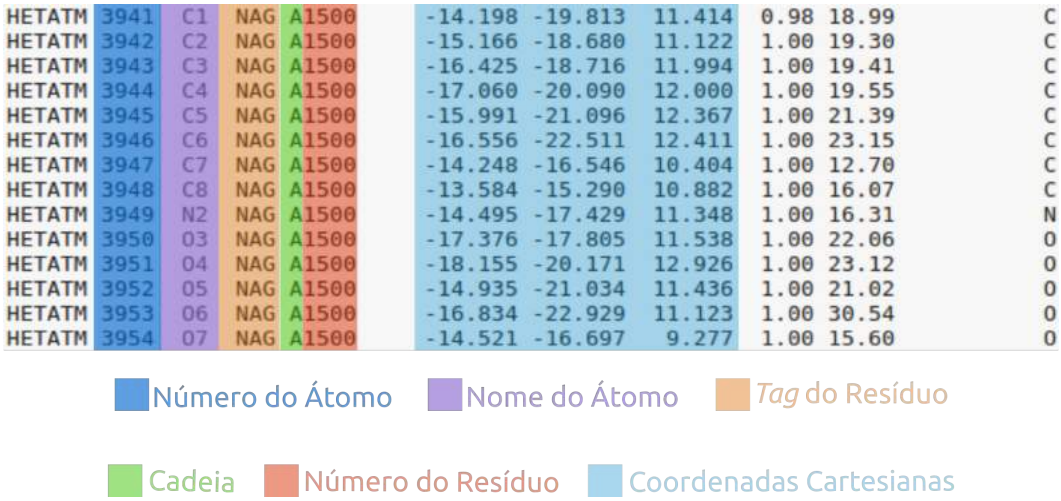
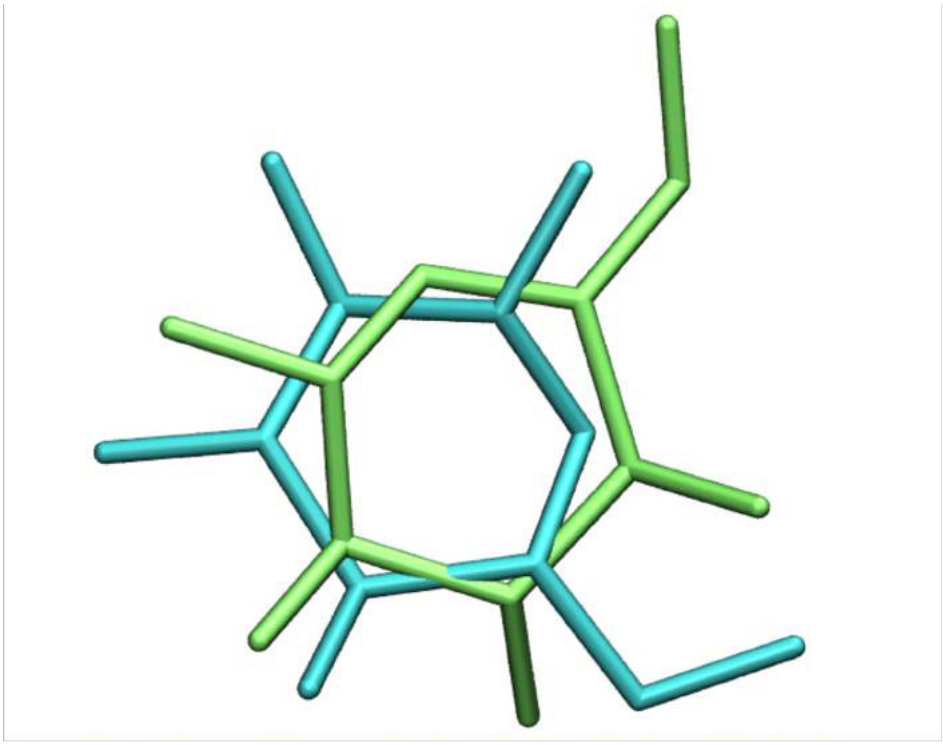


Figura 9 – Identificação das colunas do arquivo PDB e o que cada uma delas representa.

Entradas que possuíssem estruturas de dupla conformação, ou seja, com duas posições possíveis para o mesmo conjunto de átomos, distinguidas por um ' ou uma letra maiúscula ao lado do nome do átomo (C1' e C1B, respectivamente), foram tratadas como estruturas distintas. Alternativamente, poderia-se encontrar resíduos inteiros com diferentes localizações, sendo eles distinguidos através de letras maiúsculas ao lado da *tag* do resíduo (figura 10). Com o auxílio das ferramentas do BioPDB presentes na biblioteca BioPython [82], foi possível identificar a presença e agrupar os átomos com as mesmas marcas, salvando em um arquivo todos os átomos com ' ou "B"e em outro arquivo os átomos sem marcações ou com um "A". Nesse passo, foi primordial checar o número do resíduo atrelado aos átomos desordenados, a fim de garantir que tratava-se, de fato, de organizações diferentes do mesmo resíduo.



```
HETATM    1  C1 AGLC A1605      43.568  -0.082  16.118  0.50 21.93           C
HETATM    2  C1 BGLC A1605      42.208  -1.059  16.157  0.50 42.26           C
HETATM    3  C2 AGLC A1605      42.287  -0.039  15.311  0.50 20.96           C
HETATM    4  C2 BGLC A1605      42.686  -2.170  15.232  0.50 41.46           C
HETATM    5  C4 AGLC A1605      42.926  -2.458  15.030  0.50 19.30           C
HETATM    6  C4 BGLC A1605      44.983  -1.066  15.717  0.50 41.32           C
HETATM    7  C5 AGLC A1605      44.328  -2.080  15.570  0.50 19.21           C
HETATM    8  C5 BGLC A1605      44.198   0.091  16.391  0.50 42.59           C
HETATM    9  C6 AGLC A1605      45.508  -2.859  14.976  0.50 19.51           C
HETATM   10  C6 BGLC A1605      44.885   1.446  16.167  0.50 43.33           C
```

Figura 10 – Representação de resíduos com dupla conformação. Para cada átomo há mais de uma possível localização e mais de uma linha o definindo no arquivo PDB. Nesse exemplo, a distinção é feita por diferentes letras maiúsculas ao lado da *tag* do resíduo.

4.1.3 Identificação de Ligações Glicosídicas e Separação de Diferentes Cadeias

Para que fosse possível extrair mais informações, como determinar o número de cadeias de carboidratos ou qual o nível organizacional das cadeias selecionadas, foi preciso, em primeiro lugar, separar as diferentes cadeias que estavam contidas nos novos arquivos gerados na etapa anterior. Esse processo foi realizado a partir da identificação da presença de ligação glicosídica entre os resíduos. Caso houvesse, eles eram considerados partes da mesma cadeia. Em casos onde não havia ligação glicosídica entre um resíduo e algum possível vizinho, esse era considerado um monômero isolado. Em todos os casos, novos arquivos eram gerados, salvando separadamente cada uma das cadeias identificadas, ou seja, pelo menos um novo arquivo era criado para cada

entrada.

A avaliação da presença e do tipo de ligação glicosídica, bem como a estereoquímica relativa da posição anomérica de cada ligação foram realizadas por um *script* desenvolvido utilizando a capacidade de manipulação e medição sobre as estruturas do PDB concedida pelo BioPDB [82]. A identificação das ligações glicosídicas foi feita através de uma busca de vizinhos no espaço tridimensional de cada um dos oxigênios existentes nos monossacarídeos (figura 11). Para cada resíduo diferente, os oxigênios foram encontrados e, a partir deles, foi feita uma busca por átomos próximos que cumprissem os requisitos que seguem:

- Estar a no máximo 2 Å de distância do átomo de oxigênio;
- O número de vizinhos ser igual a dois (Já que apenas dois carbonos estão ligados ao oxigênio em uma ligação glicosídica);
- Os dois vizinhos devem ser, obrigatoriamente, carbonos;
- Os dois vizinhos devem ser de resíduos diferentes, para evitar a seleção de átomos do mesmo monômero.

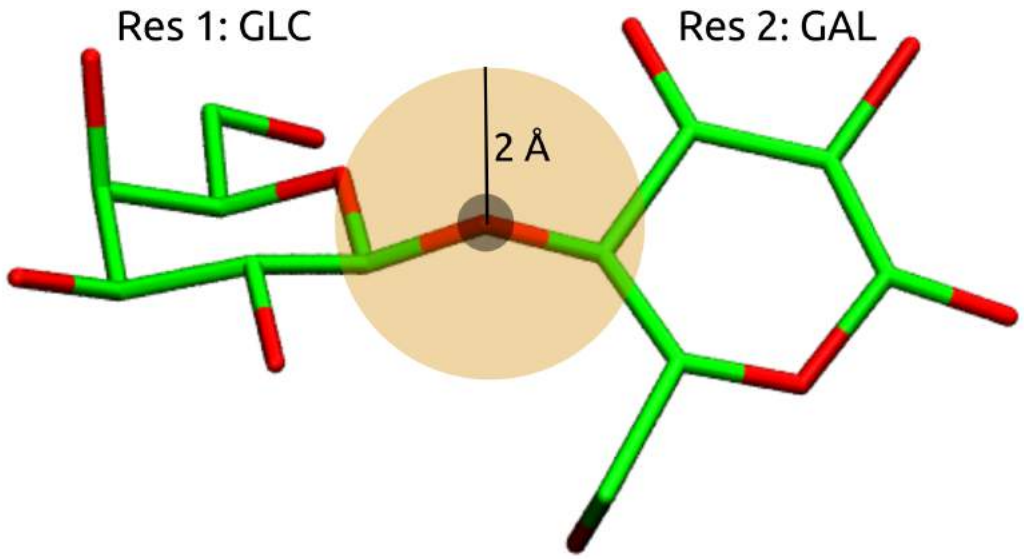


Figura 11 – Identificação da ligação glicosídica. A partir da identificação do átomo de oxigênio (circulado em cinza) foi feita a busca por átomos dentro de um raio de 2 Å (em laranja).

Após ter sido identificada a presença de uma ligação glicosídica, os nomes dos átomos de carbono participantes dela foram obtidos para definir o tipo de ligação existente entre os dois resíduos (caso os átomos fossem um C1 e um C4, era formada

Tabela 2 – Tabela contendo os átomos presentes na definição dos diedros impróprios para os diferentes centros anoméricos.

Centro Anomérico	Definição do Impróprio
C1	O5-C2-O1-C1
C2	C1-C3-O2-C2
C3	C2-C4-O3-C3
C4	C3-O5-O4-C4
C5	C4-O5-C6-C5

uma ligação 1-4). A partir disso, todos os resíduos que fossem considerados ligados e que pertencessem à mesma cadeia (identificada por uma coluna contendo uma letra maiúscula no arquivo PDB), eram considerados uma única molécula e eram salvos em um novo arquivo. Caso não houvesse ligação entre dois monômeros e eles não possuíssem o mesmo número de resíduo, eles eram considerados monossacarídeos isolados e eram armazenados em novos arquivos distintos. Esse processo foi realizado para cada um dos arquivos criados na etapa anterior (cujo nome continha "_carbo"). Além disso, o número de monômeros que compunha cada molécula de carboidrato foi obtido realizando a contagem dos diferentes números de resíduos de uma mesma cadeia presentes em uma mesma molécula. Para cada resíduo identificado, era somada uma unidade a uma variável contadora e, após percorrido todo o arquivo, esse número era salvo no nome do arquivo a fim de elucidar o número de monossacarídeos que haviam nele. O nome dos arquivos gerados nessa etapa era composto pelo código PDB, a cadeia a qual pertencia a molécula e o número de monômeros que a compunha.

Ademais, para distinguir entre ligações alfa (α) ou beta (β), a estereoquímica relativa do carbono anomérico e do estereocentro do C5 foram obtidas realizando o cálculo de diedro impróprio de cada uma dessas posições. Partindo de uma ordem idêntica dos átomos na definição dos dois impróprios a serem calculados e realizando os cálculos com a função *"calc_dihedral"* do BioPDB, pôde-se comparar os valores resultantes. No caso de diedros impróprios de estereocentros, existem duas possibilidades: possuir um valor positivo ou um valor negativo, o que difere é a posição do substituinte em relação ao plano do anel do monômero. Caso os dois centros possuíssem o mesmo valor de ângulo de diedro impróprio, isto é, tanto o substituinte do carbono anomérico (C1) quanto o do C5 estão em um mesmo lado, em relação ao plano do anel, a ligação realizada por esse monossacarídeo (caso houvesse) era considerado uma β-ligação. Se os valores fossem opostos, no entanto, era considerado uma α-ligação. A definição dos diedros utilizados nessa etapa estão descritos na tabela 2.

4.1.4 Cálculo dos Ângulos Diedrais

Buscando aumentar a informação sobre a estrutura de carboidratos, os ângulos diedrais (phi:ϕ, psi:ψ e ômega:ω) adotados por diferentes tipos de ligações glicosídicas e entre diferentes monômeros foram avaliados. Para a medição de cada ângulo, partindo das ligações glicosídicas previamente identificadas, os átomos que compunham a ligação deviam ser inequivocadamente selecionados. Utilizando os átomos de carbonos previamente identificados como componentes da ligação glicosídica entre dois monômeros, uma busca por átomos vizinhos desses carbonos foi realizada, a fim de obter os átomos restantes para compor os diedros a serem calculados. Nessa busca, realizada com um raio de 2 Å, a seleção do átomo a ser considerado parte do diedro era regida pelo número do carbono de qual a varredura partia. Caso o carbono inicial fosse um C1, o átomo a ser adicionado na composição do diedro era o O5. No entanto, partindo-se de um C4, por exemplo, o átomo a ser incorporado no diedro seria um carbono cuja numeração fosse menor, nesse caso o C3. Os diedros considerados, tanto para ϕ quanto para ψ e ω, dos principais tipos de ligações glicosídicas estão descritos na tabela 3.

Em seguida, o conjunto de coordenadas no espaço cartesiano de cada um dos átomos que compunham os diedros foi recuperado do arquivo PDB criado no passo anterior. Em posse das coordenadas, o cálculo do ângulos diedrais para cada uma das ligações glicosídicas foi realizado utilizando, também, a função *"calc_dihedral"*, que toma como argumentos o conjunto de coordenadas (X, Y, Z) de quatro átomos. Os valores resultantes dessa medição foram separados de acordo com o tipo de ligação glicosídica. Por exemplo, os valores de ângulos para uma ligação $\alpha(1 \rightarrow 4)$ (previamente identificada pelas etapas anteriores), foram separados de ângulos obtidos para ligações $\beta(1 \rightarrow 4)$. Ademais, dentro de cada tipo de ligação, diferentes populações de conjuntos de valores de ângulos diedrais foram identificadas, baseado na distribuição dos valores de ϕ, ψ e ω. *Outliers* foram identificados pelo método de intervalo interquartil e foram removidos de análises futuras.

4.1.5 Separação de Monossacarídeos

Tendo em vista a realização da medição das coordenadas de *puckering* sobre os anéis de carboidratos, a separação de cada um dos monômeros em arquivos distintos fez-se necessária. Nessa etapa, cada anel monossacarídico foi separado em novos arquivos distintos, possibilitando que o cálculo de *puckering* fosse realizado em um único anel de 6 membros de cada vez. Partindo dos arquivos criados no passo de "Identificação de Ligações Glicosídicas e Separação de Diferentes Cadeias", que continham uma única molécula de carboidrato por arquivo, um *script* percorria cada átomo de cada arquivo e

Tabela 3 – Tabela dos átomos, bem como a ordem, dos diedros utilizados para fazer a medição de ϕ, ψ e ω para cada ligação.

Ligação	ϕ	ψ	ω
1→1	O5-C1-O1-C1'	C1-O1-C1'-O5'	-
1→2	O5-C1-O1-C2'	C1-O1-C2'-C1'	-
1→3	O5-C1-O1-C3'	C1-O1-C3'-C2'	-
1→4	O5-C1-O1-C4'	C1-O1-C4'-C3'	-
1→6	O5-C1-O1-C6'	C1-O1-C6'-C5'	O1-C6'-C5'-O5'
3→3	C2-C3-O3-C3'	C3-O3-C3'-C2'	-
4→4	C3-C4-O4-C4'	C4-O4-C4'-C3'	-

Átomos marcados com um ' pertencem ao próximo resíduo.

separava-os de acordo com o número do resíduo, selecionando átomos com o mesmo número de resíduo e salvando-os em um novo arquivo. Por exemplo, no caso de um arquivo contendo um dissacarídeo de α-glicose (GLC com o número de resíduo 500 e β-galactose (GAL, com o número de resíduo 501), um novo arquivo contendo apenas átomos com a *tag* "GLC"e outro contendo apenas átomos com a *tag* "GAL"seriam criados.

4.2 Cálculo do *Puckering* de Monossacarídeos

Para avaliar a conformação de cada resíduo de carboidrato, dado que a conformação preferida e o grau de distorção do anel variam de monômero para monômero, uma etapa de cálculo das coordenadas de *puckering* [31] foi aplicado. Esse passo foi realizado utilizando uma versão adaptada do *script* desenvolvido por Hill *et al.* [83]. Após a separação de todos monossacarídeos de todas as cadeias em arquivos separados, os átomos que formam o anel piranosídico foram obtidos a partir da busca pelos nomes de átomos correspondentes no arquivo PDB em uma ordem pré-definida (O5, C1, C2, C3, C4, e C5), para que as coordenadas de *puckering* (θ, ϕ and Q) [31] de cada anel fossem medidas. Para cada arquivo contendo apenas um monossacarídeo, o nome do resíduo foi obtido a partir da *tag* presente em seus átomos, a fim de agrupar os resultados das medições por monômero.

4.3 Metadinâmica

Na tentativa de obter as superfícies de energia livre tanto de diferentes dissacarídeos, quanto de diferentes conformações de monossacarídeos, cálculos de metadinâmica foram realizados sobre diferentes sistemas. O pacote GROMACS 5.1.4 em conjunto com o pacote PLUMED 2.4 [84], foram utilizados para realizar os cálculos. Assim como em

simulações de dinâmica molecular (DM), é necessário algumas etapas de preparação dos sistemas antes que sejam executados, de fato, as etapas de equilibração e produção. Para a descrição dos parâmetros físico-químicos das moléculas estudadas, foi empregado o campo de força GROMOS53a6GLYC [75]. O modelo de água utilizado para solvatar as caixas dodecaédricas foi o SPC/E [85] na presença de condições periódicas de contorno. As minimizações de energia foram conduzidas com o algoritmo de *Steepest Descent* até que houvesse uma diferença de energia desprezível entre os passos do cálculo. O tratamento eletrostático escolhido foi o PME. Os comprimentos das ligações covalentes foi fixado pelo método de Lincs [86], permitindo a aplicação do tempo de integração de 2 fs.

Na etapa de equilibração do sistema, aplica-se uma força de restrição sobre a molécula de interesse, permitindo que as demais moléculas do sistema (em ambos os casos apenas moléculas de água) orientem-se gradualmente. Duas simulações de equilibração foram realizadas sobre os sistemas: uma com um *ensemble* canônico (volume e temperatura constantes - NVT) para permitir o ajuste adequado da temperatura do sistema e outra com um *ensemble* isobárico-isotérmico (pressão e temperatura constante - NPT) para permitir o ajuste adequado da densidade e o pressão do sistema. A força de restrição aplicada às biomoléculas como penalidade energética, a fim de restringir o movimento, foi de 1000 kJ/mol para o NVT e de 500 kJ/mol para o NPT. Ambos passos de equilibração foram simulados por 1 ns, em todos os sistemas.

Após essa etapa, é realizada a etapa de produção (também em NPT), quando ocorre a coleta dos resultados de perfil de energia livre. Nessa fase, os sistemas completos tem seu movimento irrestrito, sendo retirada a força de restrição anteriormente imposta, permitindo analisar sua dinâmica. Para isso, os sistemas de dissacarídeos foram simulados por 200 ns e os de monossacrídeos por 100 ns, permitindo uma completa descrição do seu perfil energético. O acoplamento de temperatura utilizado foi o termostato V-rescale [87] em todas as etapas, mantendo a temperatura dos sistemas em 298 K. O algoritmo de acoplamento de pressão utilizado foi o barostato Parrinello-Ramham [88, 89].

4.3.1 Ângulos Diedrais de Dissacarídeos

Nas simulações de metadinâmica dos sistemas de dissacarídeos, foram utilizados os átomos dos diedros da ligação glicosídica (ϕ e ψ) como variáveis coletivas (*collective variables* - CV). Cada simulação foi realizada por 200 ns com uma Gaussiana de altura 1,2 e um σ (largura) de 0,35. As superfícies de energia livre para os pares de ângulos foram obtidas aplicando o comando *sum_hills*, presente no pacote PLUMED [84]. O erro atrelado a cada uma das CVs foi calculado utilizando análise da média de blocos

ao longo das simulações e foi insignificante.

4.3.2 *Puckering* de Monossacarídeos

Nas simulações de metadinâmica dos sistemas de monossacarídeos, foram utilizadas as coordenadas de *puckering* dos anéis de 6 membros de cada um dos monossacarídeos simulados como variáveis coletivas. O pacote PLUMED permite utilizar as coordenadas θ, ϕ e Q como CVs, bastando apenas definir os átomos que compõe o anel do carboidrato. Dessa forma, a cada variação das posições dos átomos, o programa realiza o cálculo das coordenadas, bem como o valor de energia livre atrelado a elas. Nesse trabalho, foram utilizados apenas os ângulos θ e ϕ como variáveis. Cada simulação foi realizada por 100 ns com uma Gaussiana de altura 0,5 e um σ (largura) de 0,1. As superfícies de energia livre para os pares de ângulos foram obtidas aplicando o comando *sum_hills*, presente no pacote PLUMED [84]. O erro atrelado a cada uma das CVs foi calculado utilizando análise da média de blocos ao longo das simulações e foi insignificante.

4.4 Seleção Automatizada de Potenciais de Torcionais

No desenvolvimento do programa que seleciona o potencial torcional mais adequado a diferentes hexopiranoses, o fluxo de trabalho empregado está ilustrado na figura 12. As proporções conformacionais dos casos em que o presente programa foi testado foram obtidas a partir de trabalhos experimentais previamente conhecidos [20]. Os diedros selecionados como responsáveis pela flexibilidade dos anéis monossacarídicos eram conhecidos a partir do trabalho de desenvolvimento do GROMOS53a6GLYC [75], um conjunto de parâmetros que representam o comportamento de carboidratos em meio biológico, incluindo potenciais torcionais que descrevessem esses diedros. As informações de proporções conformacionais experimentais e o conjunto de potenciais torcionais a serem ajustados foram repassados ao programa que realiza a seleção a partir da execução de um algoritmo genético não-canônico em conjunto com simulações de dinâmica molecular.

4.4.1 Implementação do Algoritmo Genético

Para implementação do algoritmo genético não-canônico, visando fazer uma seleção de potenciais torcionais de monossacarídeos, seguiu-se a lógica do pseudo-código presente na figura 12.

Valores gerados aleatoriamente, dentro de um intervalo de -10 a 10, foram os indivíduos usados inicialmente na execução do algoritmo genético. Cada indivíduo era

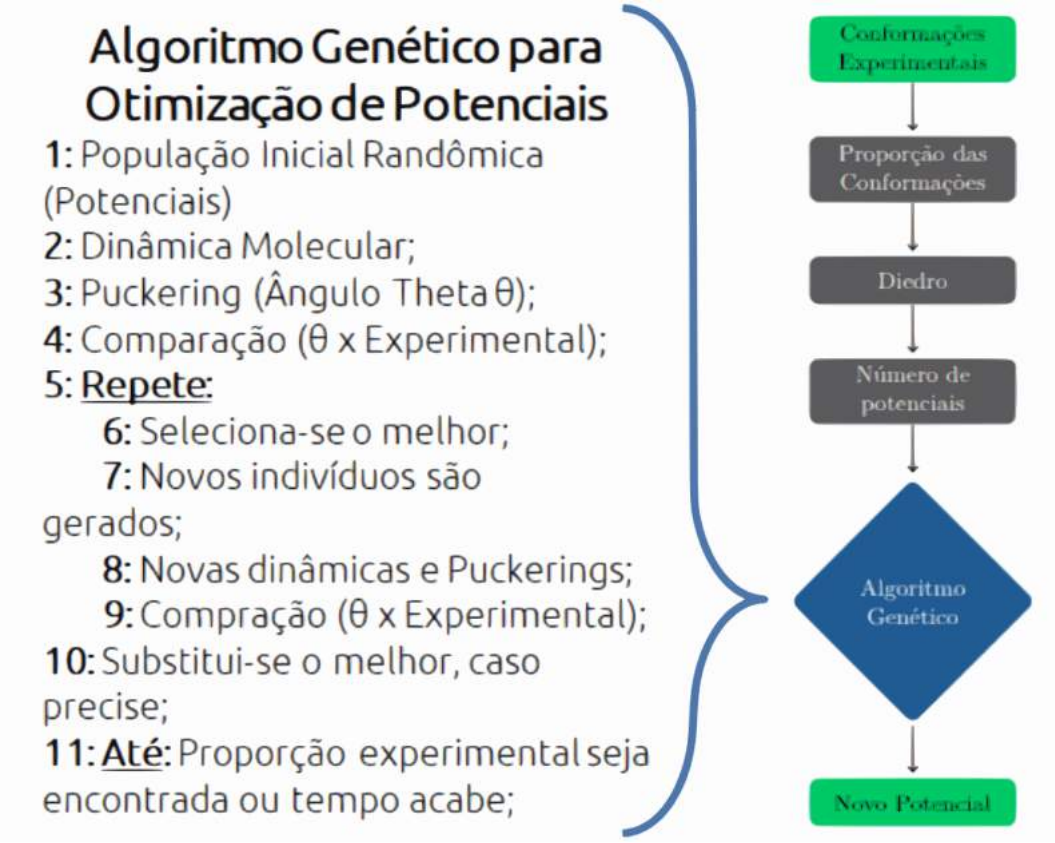


Figura 12 – Fluxograma da lógica empregada no desenvolvimento do programa de seleção automatizada de potenciais torcionais juntamente com o pseudo-código da implementação de algoritmo genético utilizada.

composto por apenas um valor dentro dos limites descritos e cada geração era composta por 16 indivíduos. Eles representavam diferentes valores de barreira energética atrelado aos potenciais torcionais utilizados para descrever os diedros, que seriam otimizados. Após a geração dos primeiros indivíduos, cada novo valor de barreira energética passou por uma etapa de dinâmica molecular (descrita na próxima seção). Tendo a trajetória das moléculas de interesse durante a fase de produção, foram realizados cálculos das coordenadas de *puckering* segundo Cremer-Pople [31].

A partir dos valores das coordenadas de *puckering* (θ e ϕ) ao longo da dinâmica, foi possível calcular a distribuição desses valores durante o tempo de execução e, portanto, obter a proporção de conformações das biomoléculas simuladas. Esse valor de proporções conformacionais era então comparado ao valor experimental, previamente conhecido. A comparação entre os valores experimentais e os simulados atuava como a função de *fit* do algoritmo genético. O indivíduo que mais se aproximasse da proporção experimental era armazenado como o melhor e era passado para a próxima geração de indivíduos. A geração de novos indivíduos foi dividida em três castas:

- Os melhores 20% dos indivíduos eram mantidos e repassados à próxima geração;
- 30% dos indivíduos eram gerados a partir do cruzamento dos 20% melhores com os 50% piores;
- Os 50% menos próximos eram gerados novamente aleatoriamente, da mesma maneira que foram gerados os indivíduos iniciais.

Essa divisão em castas visa popular adequadamente o espaço amostral, evitando que o algoritmo fique restrito a um mínimo local. Por fim, após a criação da segunda geração de indivíduos, o programa realiza novamente os passos descritos anteriormente, substituindo o melhor indivíduo, caso haja algum que melhor represente os dados experimentais. No presente trabalho, o número de gerações utilizado como limitante para a execução do algoritmo genético foi de 50. Após isso, o melhor indivíduo era retornado, juntamente com as proporções observadas durante a sua dinâmica molecular, bem como a convergência dos valores dos indivíduos, a fim de observar se houve, ou não, a restrição a um mínimo local.

4.4.2 Dinâmica Molecular

As simulações de dinâmica molecular realizadas na execução do algortimo genético, também passaram por etapas de preparação do sistema, equilibração e produção. O pacote GROMACS 5.1.1 foi utilizado para realizar esses processos e, assim como na seção anterior, o campo de força empregado para descrever o correto comportamento físico-químico das moléculas estudadas foi o GROMOS53a6GLYC [75]. Para a solvatação das caixas de simulação dodecaédricas foi utilizado o modelo de água SPC/E [85] com a presença de condições periódicas de contorno. As minimizações de energia foram conduzidas com o algoritmo de *Steepest Descent* até que houvesse uma diferença de energia desprezível entre os passos do cálculo. O tratamento eletrostático escolhido foi o PME. Os comprimentos das ligações covalentes foi fixado pelo método de Lincs [86], permitindo a aplicação do tempo de integração de 2 fs.

Nas simulações de equilibração do sistema aplicou-se uma força de restrição de 1000 kJ/mol durante a fase NVT e uma força de 500 kJ/mol na fase NPT. Ambas as fases de equilibração foram simuladas por 100 ps. Tendo em vista a grande quantidade de simulações de dinâmica molecular a serem executadas e visando a agilidade da execução do programa, na etapa de produção os sistemas foram simulados por 1 ns (hexopiranoses) e 5-10 ns (ciclohexano e seus derivados). O acoplamento de temperatura utilizado foi o termostato V-rescale [87] em todas as etapas, mantendo a temperatura dos sistemas em 298 K. O algoritmo de acoplamento de pressão utilizado foi o barostato

Parrinelo-Ramhan [88, 89]. Nos sistemas simulados de ciclohexanos e seus derivados foram mantidos os mesmos parâmetros, exceto pela diferença entre os solventes utilizados que, nesse caso, foi o Triclorifluormetano ($CFCl_3$).

5 Resultados

Os resultados presentes nessa dissertação foram divididos nos seguintes capítulos:

I Análise das estruturas de carboidratos depositadas no PDB

Felipe Nepomuceno, João Luiz Meirelles, Jorge Peña-García, Ricardo Rodríguez Schmidt, Horacio Pérez-Sánchez & Hugo Verli: Current status of carbohydrates information on Protein Data Bank

Nesse capítulo, redigido em formato de manuscrito pronto para a submissão, são analisadas as estruturas de carboidratos disponíveis no maior banco de dados estruturais da atualidade, bem como a representatividade do campo de força desenvolvido pelo grupo sobre os dados experimentais.

II Ajuste de potenciais torcionais utilizando algoritmo genético

Nessa seção da dissertação é abordado o ajuste dos potenciais torcionais de aldohexopiranoses, previamente desenvolvidos pelo grupo, através da aplicação conjunta de um algoritmo genético não-canônico e simulações de dinâmica molecular.

5.1 Capítulo I: Análise das estruturas de carboidratos depositadas no PDB

Nesse capítulo foi feita a análise das estruturas cristalizadas depositadas no *Protein Data Bank* (PDB) contendo carboidratos, a fim de obter uma maior informação sobre a dinâmica e estrutura dessas moléculas que são um dos maiores desafios da biologia estrutural. O trabalho aborda a qualidade das estruturas depositadas, a abundância de estruturas contendo cada monômero e os níveis de complexidade, além de análises de ângulos diedrais de ligações glicosídicas (ϕ e ψ) e medições de estados conformacionais de cada monossacarídeo através das coordenadas de *puckering* de Cremer-Pople (θ e ϕ) [31]. Ademais, comparações entre os dados experimentais e dados obtidos por simulações amostrando a superfície de energia livre (utilizando o conjunto de parâmetros desenvolvido pelo grupo: GROMOS53a6*GLYC* [75]) dos ângulos diedrais e coordenadas de *puckering* também foram realizadas.

Os resultados desse trabalho serão apresentados em formato de artigo que será submetido ao periódico *Journal of Chemical Information and Modeling*.

Current status of carbohydrates information on Protein Data Bank

Felipe C. Nepomuceno,[†] João L. de Meireles,[†] Jorge Peña-García,[‡] Ricardo Rodríguez Schmidt,[‡] Horacio Pérez-Sánchez,[‡] and Hugo Verli[*,¶]

[†]*Programa de Pos-Graduacao em Biologia Celular e Molecular (PPGBCM), Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Goncalves, 9500, Porto Alegre - Brazil*

[‡]*Bioinformatics and High Performance Computing Research Group (BIO-HPC), Computer Engineering Department, Universidad Catlica de Murcia (UCAM), Murcia, Spain*

[¶]*Programa de Pos-Graduacao em Biologia Celular e Molecular, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Av. Bento Goncalves, 9500, Porto Alegre - Brazil*

E-mail: hverli@cbiot.ufrgs.br

Phone: +55 51 3308 6068

Abstract

Carbohydrates are well known for their physico-chemical, biological, functional and therapeutic characteristics. Unfortunately, their chemical nature impose severe challenges for the structural elucidation of these phenomena, impairing not only the depth of our understanding of carbohydrates, but also the development of new biotechnological and therapeutic applications based on these molecules. In the recent past, the amount of structural information, obtained mainly from X-ray crystallography, has increased progressively, as well as its quality. In this context, the current work presents a global analysis of the carbohydrate information available on the entire Protein Data Bank. From high quality structures, it is clear that most of the data is highly concentrated on a few set of residue types, mainly on their monosaccharidic forms. The geometries adopted by glycosidic linkages can be mostly associated to the types of linkages instead of residues, while the level of puckering distortion was characterized, quantified and located in a pseudorotational equilibrium landscape - not only to local minima, but also to transitional states. These qualitative and quantitative analyses offer a global picture of carbohydrate structural content on PDB, potentially supporting the building of new models for carbohydrate related biological phenomena at the atomistic level, including new developments on force field parameters.

Introduction

Amongst biological macromolecules, carbohydrates (often referred to as glycans) are the most abundant ones in nature, differing from the other biomacromolecules in several properties; for example, they can be highly branched and their composing units can be linked to each other by different linkage types. Moreover, carbohydrates have hundreds of monomeric units in nature, a much more complex chemical space when compared to the 20 amino acid residues found in proteins and 5 the nitrogen bases found in nucleic acids.[1–3] Monosaccharides (sugar monomers, their simplest unit) have different configurations (D and L), pseudo-rotational equilibrium, and anomeric states.[2,4] In addition, carbohydrates can be organized in polymeric chains, oligosaccharides and polysaccharides, adding glycosidic linkages as another level for their structural information diversity, including linkage types, such as α-(1-4) and β-(1-2) and glycosydic bond dihedral angles (ϕ, ψ and ω).[2] These characteristics contribute for the broad spectrum of structures and conformations on carbohydrates, enabling them to play several roles in living organisms, many related to protein glycosylation, such as immune defense response, cell signaling, protein folding and cell trafficking, among others.[5] Additionally, polysaccharides can be employed as energy storage units (glycogen and starch), as well as structural components, such as cellulose in plants and chitin in arthropods.[6]

On the Protein Data Bank (PDB),[7] the level of resolution and the accuracy of annotation of carbohydrate-containing entries varies widely.[8] The amount of information and experimental data regarding carbohydrates structure and/or composition varies depending on the type of carbohydrate (that is, mono-, oligo- and polysaccharides, glycosaminoglycans, glycoproteins or other glycoconjugated compounds), the large number of isomers and conformers, the complex interplay between the different driving forces affecting the populations of conformers at equilibrium and the difficulty in extracting unambiguous information from experimentally derived databases.[9] X-ray crystallography, the main methodology employed to resolve structures deposited on the PDB, does not work well in highly flexible systems since it requires regular crystals, and only a few underivatized oligosaccharides crystallize.

In short, the structural complexity of the glycan chains, their microheterogeneity, flexibility, and the non-availability of these compounds in sufficient quantities have hindered their structural studies,[10] whereas a more detailed understanding of the dynamics of saccharidic structures in solution remains a challenge.[11,12]

Concerns about the accuracy of carbohydrate containing crystallographic structures were previously raised by different authors.[13–15] When such small molecules are present in macromolecular structures, they are often reported with stereo- and regiochemical errors and in unlikely conformations.[16] Although conformational distortions may reflect interactions taking place in a complex,[17] they may also reflect erroneous annotation of these molecules. This results in a poor chemical understanding and a lack of appropriate stereochemical restraints in refinement, often against low-resolution data.[18] Furthermore, in contrast to proteins or nucleic acids, there is no standard nomenclature for carbohydrate residues within the PDB-format files.[19] In this format, the residue nomeclature is determined by only three letters, which is enough for the 20 amino acid residues that compose proteins, but not to encode hundreds of different monosaccharides. Therefore, the abbreviations used to represent these molecules are often not related to the common residue name.[8] Other than that, in some cases, entire carbohydrate chains (oligosaccharides and glycoconjugates) are represented by the same three letter code, even though they are composed by different monomers. For example, the residues GAL (β-galactose) and BGC (β-glucose) are under the residue LAT (β-lactose) in PDB entries, as components of lactose,[8,20] creating difficulties to clearly identify and quantify specific structural information for residues and glycan chains.

So considering the continuous increase in the amount and quality of PDB structural information, the biological relevance of carbohydrates as well as their particular structural features, the current work aims to obtain a state of the art characterization of saccharidic PDB structural information landscape taking in to account, for high resolution structures: a) residue name lack of uniformity; b) glycosidic linkage errors; c) correction of atom numbering to IUPAC standards; d) measurement of each glycosidic dihedral linkage angles; e)

evaluation of the degree of puckering in each residue; f) location of both experimental glycosidic linkages and puckering within molecular mechanics energy surfaces from metadynamics. We understand that, by analyzing the extracted data, it is possible to obtain a clearer understanding of the conformational space of these molecules, so potentially impacting our comprehension of their biological roles and aiding in force field parameterization, modelling and drug design endeavours.

Results and discussion

PDB Overview

Starting from the data obtained by downloading the entire PDB, at first, the total number of files was about 140,000. This number, after selecting only those entries containing an identifier for carbohydrates, has dropped to around 10,000, representing 7% of the whole structural information deposited in the database. Even though one fifth of the proteins in nature are expected to be glycosylated,[21] a significant lack of representativeness can be observed in the main repository for biomacromolecules 3D structures. While the difficulty to crystallize conformationally flexible molecules is known,[22] it does impair our understanding of carbohydrates biological roles from the structural perspective. Still, in each of those entries more than one glycan chain, conformation or configuration could be present, as they were only separated by the PDB ID. So the level of carbohydrates structural information is larger than the PDB representativeness of glycans in nature.

Moreover, the quality of the information in these 10,000 files could impair the information extracted from them. Aiming to obtain a clear picture of high quality carbohydrates structural information, an evaluation of the resolution of these structures was performed (figure 1). Accordingly, the majority of the information had resolution lower than 1.5Å, the length of a covalent carbon-carbon bond applied as a resolution filter in the current work. This filtering returned only 7.2% of the glycan containing entries, which comprises only 0.5%

of all entries found in the PDB.

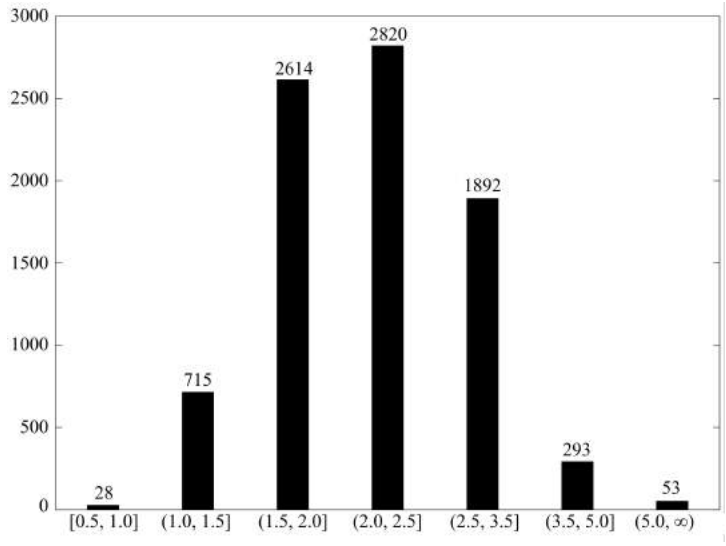


Figure 1: Resolution distribution of the carbohydrate containing entries obtained from the Protein Data Bank (PDB). Number of entries deposited in a given resolution interval.

After filtering for the molecules of interest and for the appropriate resolutions, the carbohydrate chains comprised in the models were separated from anything that was not a glycan (protein atoms, co-factors, ions, other synthetic molecules and etc). Furthermore, each individual carbohydrate chain in a same entry was also separated meaning that, if an entry had three independent lactose molecules, they were considered as three distinct chains increasing, as a consequence, the amount of structural information available for analysis. This was also performed for cases where there were different conformations for a set of atoms in a given carbohydrate residue. The total number of files generated after this separation step was of 1,755 files containing a single glycan chain, with variable sizes.

The number of occurrences of each monosaccharide in the PDB entries was, then, assessed (table 1), including both anomeric states (α and β) in the same residue. In this case, the 'Count' column represents the number of files in which the monomer is encountered at

least once, either isolated or complexed. The two most abundantly found monosaccharide residues in the PDB structures are N-Acetyl-D-Glucosamine (NAG/NDG) and D-Glucose (GLC/BGC) residues. While NAG is present in the majority of eukariotes protein glycosylation backbones,[23] hence its elevated presence in the database, GLC/BGC are known to be the most abundant monosaccharides in nature,[24] being present in a wide variety of organisms and carbohydrate molecules.

Table 1: Most abundant monosaccharides in structures the deposited on the PDB.

Name	Count
N-Acetyl-D-Glucosamine	590
Glucose	539
Galactose	370
Mannose	322
Fucose	164
Xylopyranose	57
Other Molecules	388

For each of the residues 'Count' indicates the total number of structures with a resolution of 1.5Åor higher, in which a given residue is found at least once. The α and β configurations are presented together within the same monosaccharide.

In order to evaluate the amount of oligo- and polysaccharides in the crystal structures, the size of their chains was measured, based on their number of monomeric units (table 2). Accordingly, isolated monosaccharides represent 64% of the structures, and small oligosaccharides sum a total of about 30%. The information regarding these molecules, therefore, is highly concentrated, not only in a few types of monosaccharides, but also in lower levels of complexity.

Table 2: Length of glycan chains found in PDB by relative abundance.

Size	Quantity	Percentage
Monosaccharide	1036	64.1%
Disaccharide	224	13.9%
Trisaccharide	184	11.4%
Tetrasaccharide	80	5.0%
Pentasaccharide	45	2.8%
Hexasaccharide	16	1.0%
Heptasaccharide	20	1.2%
Eight or more	10	0.6%

Glycosidic linkage information

Having the separated files, each one containing a single carbohydrate chain, the glycosidic linkage information was then assessed and quantified (figure 2). The bond that showed to be the most recurrent was the $\beta(1 \rightarrow 4)$, with the most abundant occurrence of connected saccharides being two NAG (N-Acetyl-β-D-Glucosamine) residues, followed by a BGC (β-D-Glucose) disaccharide. The former type of linkage is a common structural element found in numerous carbohydrate structures, including plant cell walls[25] and insect exoskeletons,[26] and it is in the common core oligosaccharide in N-linked glycoproteins.[27]

Due to the high number of rotational degrees of freedom in oligo- and polysaccharides, the characterization of their the conformational properties could represent a key step to establish conformational-function relations.[28] Moreover, different types of linkages and different monomers composing the linkage lead to distinct patterns of glycosidic torsional dihedral angles,[29,30] adding to the complexity of these polymers. To aid in the comprehension of these torsional angles, the values of the dihedral angles ϕ and ψ of each individual glycosidic linkage were measured and grouped by linkage type. Additionally, for some of the linkages more than one population of ϕ or ψ angles was observed and, accordingly, grouped separately (table 3). We then compared these crystallographic geometries to metadynamics simulations using GROMOS53a6*GLYC* set of parameters[31](figure 3).

The most abundant glycosidic linkage on the PDB, the β-(1→4) (figures 3A, 3B and 3C),

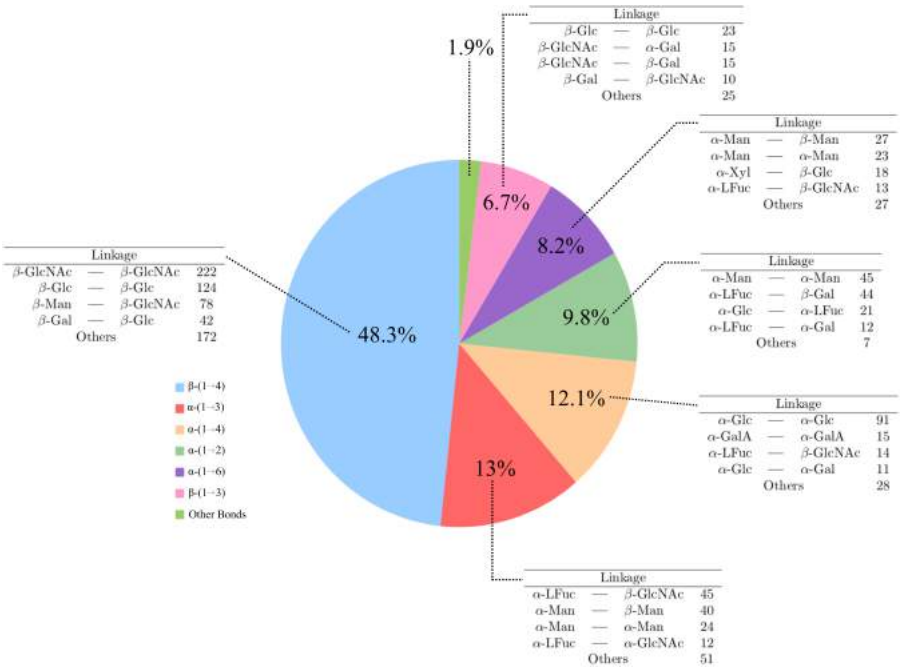


Figure 2: Distribution of the type of glycosidic linkages on the PDB. Each linkage shows the most recurrent connected residues. Absolute numbers: $\beta(1 \rightarrow 4)$: 638; $\alpha(1 \rightarrow 3)$: 172; $\alpha(1 \rightarrow 4)$: 159; $\alpha(1 \rightarrow 2)$: 129; $\alpha(1 \rightarrow 6)$: 108; $\beta(1 \rightarrow 3)$: 88; Other: 26.

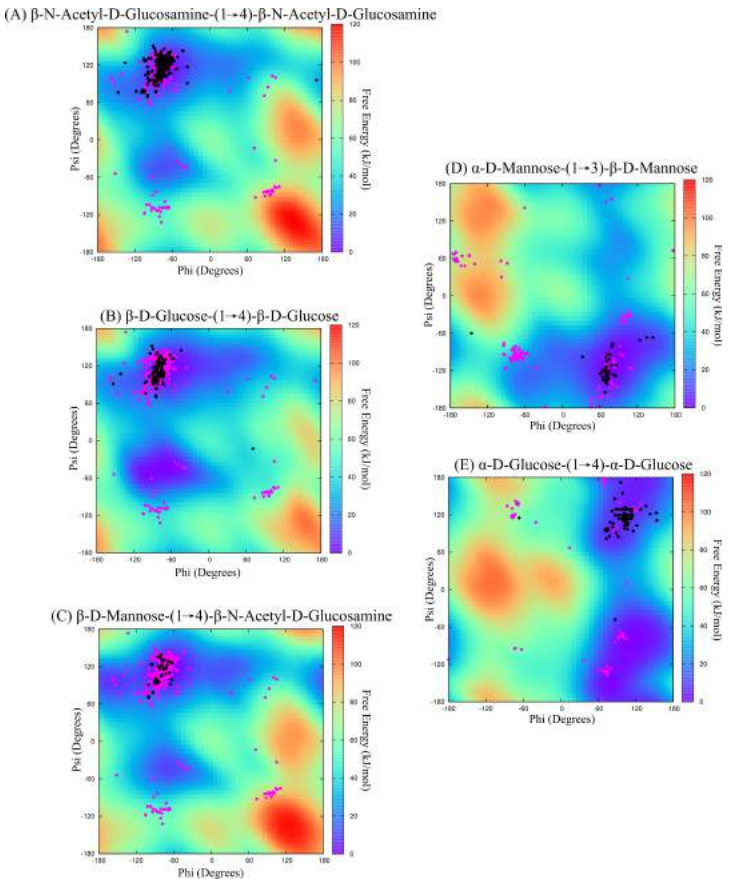


Figure 3: Overlap of Free-Energy Landscapes (FEL) obtained by metadynamics simulations of different disaccharides and Ramachandran plots obtained from the PDB measurement of dihedral angles. The black dots represent the measured dihedral angle values of the simulated disaccharide, while the magenta dots represent the values of other disaccharides with the same bond type.

is mostly explained by disaccharides formed by N-Acetyl-β-Glucosamine (35%), β-Glucose (19%) and β-Man-(1→4)-GlcNAc (12%), corresponding to about 2/3 of the conformational information around this linkage. Its geometry is highly concentrated around $\phi = -80°$ and $\psi = 115°$ (table 3), encompassing more than 90% of this linkage geometry, the main minimum energy region at the western-northern quadrant of the associated free-energy landscape (figure 3). These values are similar to those obtained from experimental data, being $\phi = -88°$ and $\psi = 99°$ (NMR[32,33]) and $\phi = -76°$ and $\psi = 108°$ (X-ray crystallography[29,33]). Still, some geometries may be found in less favourable or even transitional states (such as PDB ID: 5ELD,[34] 5ELC[34] and 4WZK[35]) suggesting specific conformational stabilization at crystallographic complexes.

The third most abundant glycosidic linkage on PDB, α-(1→4), represents about 25% of the amount of conformational information for the β-(1→4)(table 3), in an example of the limitation of carbohydrates experimental structural information. Within such small chemical space, more then 50% is related to α-Glucose disaccharides, highly concentrayed on the eastern-nothern quadrant of the free-energy landscape (figure 3).

From the observation of the experimental geometry for these linkages (table 3 and figure 3), two main aspects could be highlighted: *i*) conformational information on carbohydrates glycosidic linkages is also highly concentrated and, in most types of linkages, imposing severe limitations for an experimental support to describe most types of connections between carbohydrates; *ii*) while possibly associated to a limited amount of data, conformational preferences are mainly associated to the linkage type rather then to the connected residues, whereas some small conformer populations seem to appear for specific residue compositions. While the first case impairs an experimental support to carbohydrates modelling or development of probability functions for stochastic search methods,[36] the second observation indicates that higher amounts of experimental data might be not pivotal to obtain reliable carbohydrate models, at least by quantum mechanics or molecular mechanics.

In this sense, these conformational preferences have been observed by both computational

Table 3: Glycosidic dihedral angles ϕ and ψ populations average values.

		ϕ		ψ		ω	
		Average (°)	Count	Average (°)	Count	Average (°)	Count
$\alpha(1 \rightarrow 1)$	Pop 1	74.3	1	67.6	1	-	-
$\alpha(1 \rightarrow 2)$	Pop 1	86.5 ± 15.4	72	115.1 ± 31.7	82	-	-
	Pop 2	-75.2 ± 17.5	51	-90.6 ± 13.1	40	-	-
$\alpha(1 \rightarrow 3)$	Pop 1	75.1 ± 10.5	78	-108.2 ± 24.6	135	-	-
	Pop 2	-75.2 ± 9.5	55	60.6 ± 7.8	20	-	-
	Pop 3	-167.1 ± 12.8	19	-	-	-	-
$\alpha(1 \rightarrow 4)$	Pop 1	95.7 ± 15.9	129	119.1 ± 14.8	120	-	-
	Pop 2	-73.9 ± 5.3	22	-105.8 ± 30.0	29	-	-
$\alpha(1 \rightarrow 6)$	Pop 1	70.8 ± 10.3	72	177.7 ± 15.9	70	-63.6 ± 9.3	63
	Pop 2	-65.4 ± 11.0	25	95.7 ± 22.2	21	72.8 ± 20.3	41
	Pop 3	-	-	-117.5 ± 2.1	15	-	-
$\beta(1 \rightarrow 1)$	Pop 1	-150.1 ± 23.5	6	-152.0 ± 9.6	8	-	-
	Pop 2	167.2 ± 6.6	5	170.8 ± 9.0	4	-	-
$\beta(1 \rightarrow 2)$	Pop 1	87.0 ± 23.0	2	83.6 ± 52.9	3	-	-
	Pop 2	-76.7 ± 5.7	2	-93.6	1	-	-
$\beta(1 \rightarrow 3)$	Pop 1	-77.8 ± 12.8	80	-122.6 ± 20.1	85	-	-
	Pop 2	119.4 ± 37.9	4	-	-	-	-
$\beta(1 \rightarrow 4)$	Pop 1	-79.5 ± 12.0	609	115.8 ± 17.1	588	-	-
	Pop 2	95.3 ± 10.1	21	-90.5 ± 24.8	44	-	-
$\beta(1 \rightarrow 6)$	Pop 1	70.4 ± 13.4	4	148.3 ± 33.3	6	-85.2 ± 16.3	6

methods[28,30] and limited amount of experimental information in different levels of resolution or variation ranges. As a general feature, the average value of ϕ for α-linked glucose-based disaccharides was observed to be around [60°:100°], while for β-linked disaccharides around [-90°:-60°], both regions commonly associated to main energy minima in the free energy landscapes for glycosidic linkages and strongly related to the exo-anomeric effect.[37] The ψ angle, however, has a much more heterogeneous behaviour, which could account for the differences between crystallographic and simulated data, due to the different chains, conditions and processes in which the experimental data was obtained. In this sense, the data presented here offer a precise and quantitative measurement based on the majority of the currently available experimental information for linkages between carbohydrate residues.

Puckering

Since the 70's[38] different approaches were developed to quantitatively measure the level of distortion of carbohydrate residues.[39,40] Such developments may be interconnected to the recognition of unusual conformational states as important to several biological phenomena, such as glycosidic linkage hidrolysis[41,42] and biological roles performed by glycosamioglycans.[43] Also, when evaluating the pseudorotational space of carbohydrates, the crystal resolution could highly impact any analysis, as subtle conformational differences could be lost in middle to low resolutions. Accordingly, we employed Cremer-Pople puckering coordinates measurement[38] on carbohydrate structures with 1.5$\mathring{A}$ or higher resolution to obtain a fine characterization of the puckering information on PDB, which could in turn contribute to the understanding of these no-chair conformations to biological phenomena.

Globally, the analysis of the puckering from the most abundant monosaccharide residues (table 4) clearly indicates that conformations distant from the global minima ($^1C_4$ or $^4C_1$) are rare, from 1.5 to 6.7% of the PDB, depending on the type of the residue. These numbers are directly correlated to the resolution of the PDB entry, and can increase up to 6.5 fold when structures with resolutions lower than 1.5$\mathring{A}$ are considered. Such dependence on

the resolution was expected, but not objectively measured neither quantified to our knowledge, and offer a warning to inferences of the biological relevance of non-chair conformations depending on the experimental data quality.

Table 4: θ angle averages at ideal chair conformation and at distorted conformations for each of the most abundant monosaccharides.

	Ideal Conformation		Other Conformation		Deviation From Ideal Conformation (%)	
Monosaccharides	Average (°)	Count	Average (°)	Count	Resolution <1.5 Å	Resolution >1.5 Å
N-Acetyl-D-Glucosamine	7.4 ±4.6	892	63.2 ± 22.3	31	3.3	10.9
D-Glucose	7.6 ± 4.7	565	58.6 ± 26.4	34	5.7	9.5
D-Mannose	7.1 ± 4.8	471	65.9 ± 36.7	32	6.4	15.4
D-Galactose	6.7 ± 3.8	272	46.0 ± 24.2	4	1.4	6.0
L-Fucose	174.8 ± 3.5	221	93.6 ± 43.4	7	3.1	19.9
D-Xylopyranose	6.8 ± 3.5	180	96.6 ± 40.3	13	6.7	11.1

Monosaccharide rings have two main minima ($^4C_1$ and $^1C_4$) and several local minima or transitional states (Boat, Skew-Boat, Envelope, Half-Chair).[44] For instance, these multiple conformations have been suggested to play an essential role in the hydrolysis of glycosidic linkages by glycoside hydrolases (or glycosidases);[45] when the hexopyranosidic ring undergoes certain conformational changes, the glycosidic oxygen atom is placed near the acid/base catalytic residue.[41,46] Given that, the observed transitional conformations in figure 4 (above the black line), are potentially induced or selected conformational states upon interaction with proteins and enzymes, rather than experimental errors of the deposited structures in the database (considering the cut-off of 1.5$\mathring{A}$ employed in our analyses).

Aiming to locate these experimental conformations to points at their free-energy landscape (FEL), metadynamics simulations of the uncomplexed residues were performed and the results supperimposed to the conformations found in the PDB (figure 5). Accordingly, most conformations are in the global minima, whereas conformations are found at local minima or at transitional states. For instance, one of the energy minima found in the N-Acetyl-D-Glucosamine (figure 5A) corresponds to a ^{1,4}B conformation, which was pointed out previously to be induced by chitinases.[47,48]

For Glucose (figure 5B), local minima corresponding to $^1S_3$ conformation can be observed, as well as the pathway leading to that minima (passing through E_3, $^4H_3$ and E_4 conforma-

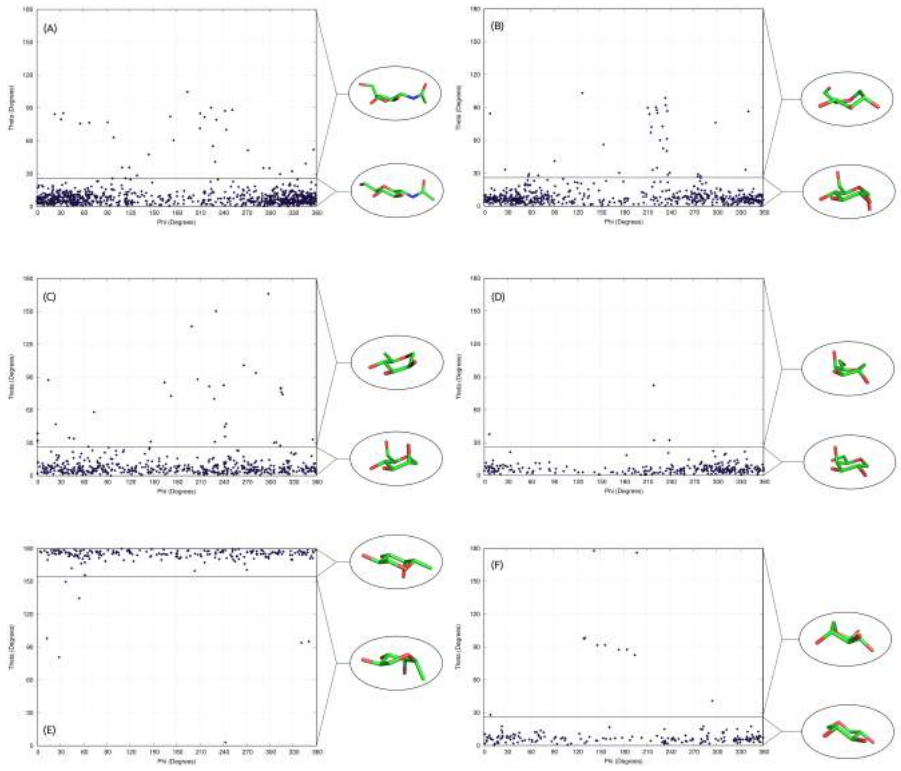


Figure 4: θ puckering coordinate values of the six most abundant monosaccharides found in the PDB. The region between θ=0°and θ=26°(except for L-Fucose where it is θ=154°and θ=180°) delimited by the horizontal line shows the most stable conformational space for each molecule. Representative chair and non-chair conformations are shown besides each case, to illustrate the ring distortion. (A) N-Acetyl-D-Glucosamine (B) D-Glucose (C) D-Mannose (D) D-Galactose (E) L-Fucose (F) D-Xylose. In every case, the alpha and beta conformations are considered together.

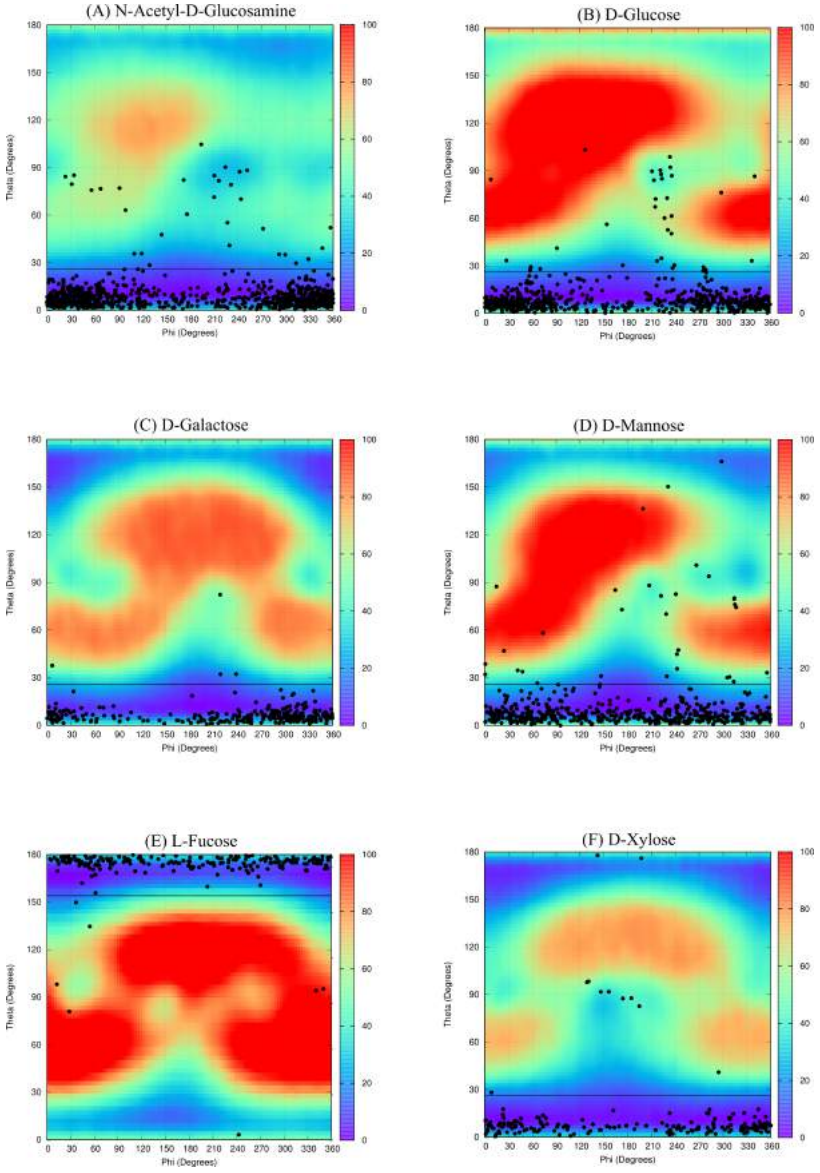


Figure 5: Overlap of Free-Energy Landscapes (FEL) obtained by metadynamics simulations of different monosaccharides and puckering coordinates values obtained from the PDB for the most abundant monosaccharides presented in figure 4

tions), in agreement with previous works.[49,50] Another conformations as $^{1}S_{5}$, $B_{2,5}$ and $^{O}S_{2}$, accessed as local minima by the FEL, were also related in previous works to the catalytic processes performed by different families of enzymes.[42,51] In these situations, it seems that a conformational selection takes place, whereas other conformations in non-stationary regions suggest a conformational induction process or a conformation particularly associated to the original experiment.

To additionally explore the influence of the protein environment on the carbohydrates pseudorotational equilibrium, we assessed the free energy profile in function of θ for two different monosaccharides (α-Glucose and α-Mannose) when interacting with glycosyl hidrolases (PDB ID 2BHY[52] and 2BY2[52] for α-Glucose and PDB ID 5A7V[53] for α-Mannose) compared to when free in solution (figure 6). In these situations, two different processes seem to occur: i) the enzime does not to interfere in the monosaccharide puckering (figure 6A), and ii) the enzyme stabilizes the associated free energy landscape by up to 10kJ/mol (figure 6B).

The conformational stabilization produced by the enzyme is evident (figure 6B), and the experimental conformation of the monosaccharide is located in a local minima with a minor energy stabilization (θ around 80 degrees), probably the conformation with the largest complementarity with the enzyme active site. In the α-Glucose complexes, on the other hand, no clear energy stabilization was observed upon complexation, which was checked for two different crystallographic structures (figure 6A). That, added by the fact that the same residue complexed with the same enzyme presents different levels of puckering distortion in non-stationary states suggest more a conformation derived by hostile environment needed for the crystallization process to take place, which can result in distortions or modifications of the ligand molecules,[16] then a conformational induction or selection promoted by the enzyme.

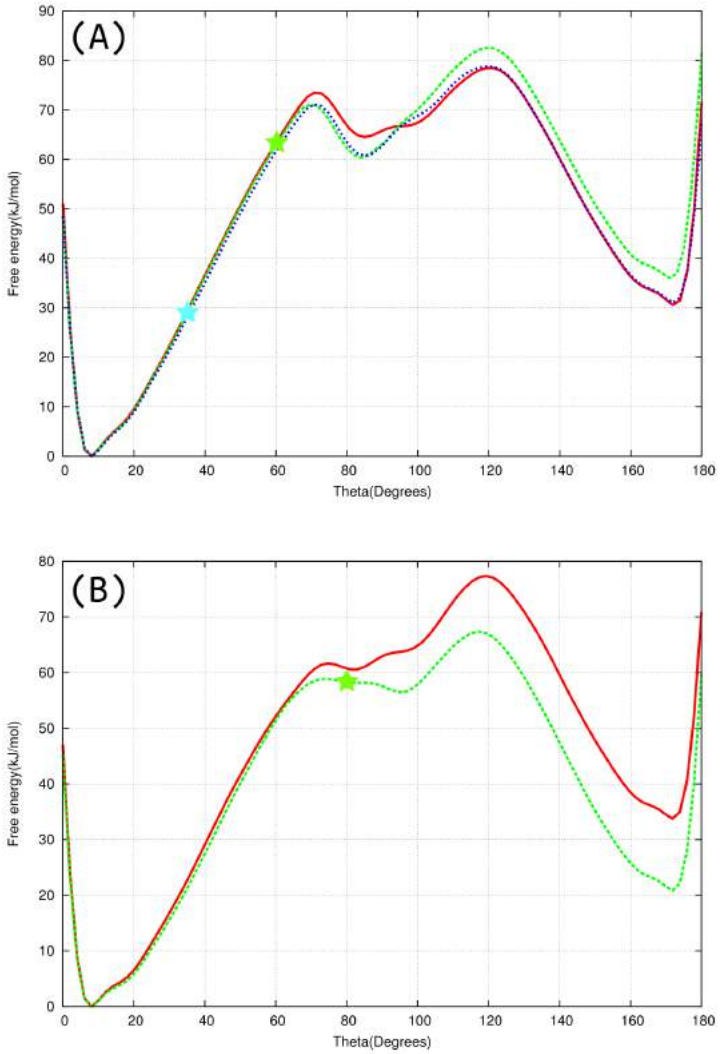


Figure 6: Free energy profile in function of θ for two monosaccharides residues while free in solution (red line) and interacting with a target protein (green and blue lines). (A) α-Glucose (2BHY: green line, 2BY2: blue line) (B) α-Mannose. The star mark the experimental θ value for the residue when complexed.

About carbohydrate related databases

The effort to unravel the information about glycobiology contained in different databases is not recent. Other studies have been tackling this subject with different approaches. In order to better extract the information from the PDB, given the great variety and number of errors in carbohydrate structures, Lutteke and co-workers[20,54,55] have developed a series of algorithms, that were compiled in a server called the Carbohydrate Structure Suite (CSS). In it, the glycosidic information could be extracted more easily, given that the server could check for the integrity of PDB-format files, identify carbohydrate structures in them, asses torsional angles and deposit them in a torsional angles database of its own, retrieve carbohydrate-protein vicinity interaction, among other functionalities. Although thorough, the CSS did not presented information regarding monosaccharide ring conformations, an important aspect when assessing carbohydrate structural information. Furthermore the information, while corrected for possible errors, does not have a resolution threshold, limiting the precise definition of the atomic positions of each atom in the molecules.

Other two unified databases were developed in view of supplying the need for a glycobiology dedicated tool for assessing carbohydrates and their properties. The Carbohydrates Structure Database (CSDB) is one of them and stores structural, bibliographic, taxonomic and NMR, as well as other data on natural carbohydrates and their derivatives published in the scientific literature.[56] However, structural information regarding torsional angles, conformational states, interaction to proteins and other aspects that are important for structural glycobiology are not contemplated in this database.

The GlycomeDB follows a similar approach to the CSBD. Unifying the structural and taxonomic data information of all major public carbohydrate databases, as well as carbohydrates contained in the Protein Data Bank, GlycomeDB is the database with the most comprehensive and unified resource for carbohydrate structures worldwide.[57] Even though in this work there is the addition of several taxonomy and crystallographic data from the PDB, there is little care for the distinct errors that can be found in these structures. It

helps to understand the structures and functions of well known carbohydrates, but when it comes to the generation of new knowledge based on analysis and observations of systematic structural behaviours these molecules might have, a more limited scenario is available.

Conclusions

In summary, in this work we were able to extract and analyze high quality structural information regarding the entire content of carbohydrate structural information present in the PDB. Accordingly, information such as monosaccharide abundance and conformations, preferred glycosidic dihedral angles and saccharidic chain size were analyzed, and could come to aid in the development, refinement or even validation of force field parameters, aiming a better representation of experimental observations. From the combination of experimentally observed conformational states for hexopyranoses with studies of energy barriers and transitional states pathways,[58–60] the work does could support a more accurate representation of the pseudorotational equilibrium in carbohydrate force fields, as well as the development of more specific torsional potentials from experimental knowledge regarding the ϕ and ψ populations to acknowledge distinct behaviours, one or more main conformational states and distinct levels of flexibility[2] from different types of linkages. So we expect that the obtained data could offer new perspectives for the knowledge of carbohydrate structures based on experimentally obtained information, which in turn could be usefull from both modelling and biological phenomena perspectives.

Experimental

Overview

In order to perform a systematic analysis of the entire carbohydrate information on PDB we implemented a pipeline aiming to separate entries into its multiple information units (see

further), identify each monosaccharide even in oligosaccharide structures without monosaccharidic content indication, identify each type of connectivity between monosaccharides, measure dihedral angles and puckering angles.

Data retrieving and filtering

All entries within the PDB were downloaded in 28 of June of 2018, in a total of 140,547 entries, which were filtered based on the identification of heteroatoms. If in any section of each PDB file the name of a carbohydrate was present in the heteroatom name section ('HETNAM' in the beginning of the PDB file, as opposed to 'HETATM' at the atoms description), the given file was considered to have a carbohydrate structure. This layer of filtering was applied in order to select as many carbohydrates as possible, using their most commonly known residue names, suffixes and prefixes: Abequose, Arabinose, Fructose, Fucose, Glucose, Galactose, Galactosamine, Glucosamine, Idoronic, Lactose, Maltopyranoside, Maltoside, Mannose, Rhamnose, Saccharide, Xylose, the prefix 'Gluc', and the suffixes 'Ose' and 'Uronic'. In cases where the name of a carbohydrate molecule was also the identification tag for each residue (right next to 'HETATM'), the files were manually inspected and the correct tags for the residues were added. In PDB files with β-Lactose, for example, the residues were marked with the tag 'LAT', but then were correctly renamed to 'GAL' (β-Galactose) and 'BGC' (β-Glucose).

Furthermore, a step of resolution-based filtering was employed. From the entries selected by the carbohydrate nomenclature, the resolution of the structures was encountered by searching for the 'RESOLUTION' tag in the PDB file. If the resolution was lower than 1.5 $\AA$, this entry was selected and stored. All the entries where the structure was determined with Nuclear Magnetic Resonance (NMR) were left out of this analysis.

The glycan PDB structures present in the selected files were ordered by number of occurrences of each different monosaccharide, and those with less than 10 occurrences in the database were removed from further analysis. Any linear form of carbohydrates, such as lin-

ear glucose and linear xylose (when specified in the sugar's name) were removed, along with 5-membered rings. Proper handling of these files was achieved through an in-house script developed using the programming language Python 2.7 following the pseudo-code represented in figure 7.

```
1: Download from Protein Data Bank;
for every pdb_file:
        2: Filtering for carbohydrate identifiers;
                if HETNAM_TAG in carbohydrates_identifiers:
                        carbo_tag = HETNAM_TAG;
                        select(pdb_file);
        3: Resolution filtering;
                if RESOLUTION < 1.5 Å:
                        select(pdb_file);
        4: Carbohydrate identification and separation from protein;
                if carbo_tag == HETATM_RES_TAG:
                        carbo_pdb_file = carbohydrate_atoms;
for every carbo_pdb_file:
        5: Glycosidic linkage identification (BioPDB);
                find(linkage_O_atom);
                search_neighbours(linkage_O_atom);
                define(dihedral_atoms);
        6: Dihedral angles calculations (φ, ψ);
                calc_dihedral(dihedral_atoms);
        7: Monomers separation;
                separate_monomers(carbo_pdb_file);
for every monomer_pdb_file:
        8: Puckering measurements (θ, φ, Q);
9: Data organization and analisys.
```

Figure 7: Pseudo-code for the pipeline employed to extract and analyze PDB carbohydrate information.

Glycosidic linkage identification and glycan identification

Aiming to remove the protein atoms and co-factors that were also present in the PDB file, a search for the carbohydrate tags identified by the 'HETNAM' label was performed. In it, every atom of the file was checked to assess whether the residue identifier (in each atom

definition) was the same as the carbohydrate tag (defined at the beginning of the file in the 'HETNAM' section). If true, the atom was saved in a new file. If an entry was said to have Glucose (GLC) in the 'HETNAM' section , for example, then only atoms with the 'GLC' residue identifier were selected in this step.

Entries with disordered structures, with two possible positions for the same atoms, often identified by a ' symbol or a capital letter beside the atom name (C1' or C1B, respectively), were treated as different entries. With the aid of the BioPython library BioPDB,[61] the separation of these cases was achieved by identifying the residues that had this atom name marks and grouping all the ones with ' or capital letter in one file and the unmarked in another. As a consequence, one PDB entry could offer more then one unique carbohydrate set of coordinates, one for each conformation retrieved in the experiment.

Moreover, to identify different glycan chains present in a same PDB entry, another step was performed, checking each atom of each residue and identifying the glycosidic linkages between residues in the structure. To assess the presence and type of each linkage and the relative stereochemistry of the anomeric position, an in-house script based on the BioPython library BioPDB[61] was written, allowing manipulation and measurements over PDB structures. By going through every oxygen atom in each monomer, the script performed a neighbour search in a 2 $\mathring{A}$ radius for each of them and, if a set of conditionals were true, the atoms were considered to be in a glycosidic bond. These were: (i) the number of neighbours were only two (as in the two carbons that are connected in a glycosidic linkage); (ii) both atoms were carbons, and (iii) they were from different numbered residues, to avoid selecting atoms from the same residue that did not participate in the bond. After that, by getting the name of each of the atoms composing the bond (for example: C1, C2, C3, C6), the linkage type was determined and stored. Afterwards, if the residues belonged to the same chain and there was a bond between them, they were considered as a single molecule and were stored in a new separate file. This resulted in several files from every PDB entry, where only one glycan chain was present.

The identification of the size of each oligosaccharidic structure was, then, performed by accessing each separated file and counting the number of residues within, using the same 'bond searching' strategy as described above. Furthermore, to distinguish between α and β-glycosidic linkages, the relative stereochemistry of the anomeric position and of the C5 stereocenter were affered by measuring the improper dihedral of each of those positions (table 5); by comparing these two dihedrals, it was possible to assess if they were on the same side of the ring plane (β orientation) or in opposite sides (α orientation). If both of the carbon atoms had the same improper dihedral (that is, the same stereochemistry), the bond was considered a β-glycosidic bond. On the other hand, if the dihedral was different, the linkage was considered an α-glycosidic bond.

Table 5: Table of the atoms composing improper dihedral definition for each anomeric center.

Anomeric Center	Improper Definition
C1	O5-C2-O1-C1
C2	C1-C3-O2-C2
C3	C2-C4-O3-C3
C4	C3-O5-O4-C4
C5	C4-O5-C6-C5

Dihedral angle calculations

The dihedral angles phi (ϕ), psi (ψ) and omega (ω) adopted by different types of glycosidic linkages and monomers were assessed. For the measurement of each angle, from the glycosidic linkages previously identified, the atoms from each linkage had to be unambiguously identified. By using the carbon atoms already selected in the glycosidic linkage identification step, a neighbour search was employed to identify the remaining atoms composing a given dihedral. In order to account for all types of glycosidic linkages, we searched in a radius of 2 $\mathring{A}$ around each of the carbon atoms. If the neighbour atom was a lower numbered carbon atom (for example a C3 when compared to the C4 carbon that participated on the linkage), this atom was added to the dihedral atoms. However, if the carbon atom was a C1, the

selected atom was O5, allowing us to automatically identify dihedrals from the main types of glycosidic linkages (table 6), as well as unusual ones. After that, the coordinates for the four atoms of each dihedral (ϕ and ψ) were retrieved from the PDB file. Having the coordinates of the dihedrals' atoms, the calculation of the angles were performed also using the *'calc_dihedral'* command from BioPython (BioPDB).[61] The angles were then organized by the linkage type. Within each linkage type, different angle populations were identified based in the distribution of the values of ϕ and ψ. Outliers were identified by the interquartile range method and removed from further analysis.

Table 6: Table of the atoms composing dihedral angle measurement (ϕ, ψ and ω) for each bond.

Bond	ϕ	ψ	ω
1→1	O5-C1-O1-C1'	C1-O1-C1'-O5'	-
1→2	O5-C1-O1-C2'	C1-O1-C2'-C1'	-
1→3	O5-C1-O1-C3'	C1-O1-C3'-C2'	-
1→4	O5-C1-O1-C4'	C1-O1-C4'-C3'	-
1→6	O5-C1-O1-C6'	C1-O1-C6'-C5'	O1-C6'-C5'-O5'
3→3	C2-C3-O3-C3'	C3-O3-C3'-C2'	-
4→4	C3-C4-O4-C4'	C4-O4-C4'-C3'	-

The atoms marked with a ' belong to the next residue.

Puckering calculation

Intending to perform puckering measurements over the carbohydrate rings, a monomer separation from oligosaccharidic chains was necessary. Starting from the files generated in the 'Glycosidic linkage and glycan identification' step, which contained only a single glycan chain in each file, separated them selecting atoms which had the same residue number, saving them in a separate file. To assess the conformation of each carbohydrate residue, as well as its level of distortion, a step of puckering evaluation was applied. This step used an adapted version of the script developed by Hill and co-workers.[39] After the separation of each carbohydrate found on PDB, the atoms that form a pyranose ring were identified by searching for the corresponding atom names, and a predefined order of their atoms (O5, C1, C2, C3, C4, and

C5) was selected to measure the puckering coordinates (θ, ϕ and Q) of every ring as given by Cremer-Pople.[38]

Metadynamics

In order to obtain energy landscapes as reference for the interpretation and analysis of carbohydrates experimental properties, well-tempered metadynamics simulations[62] were carried out using the GROMACS 5.1.4 package[63] and the open-source, community-developed PLUMED library, version 2.5.[64] The simulated systems were divided into two groups: (i) disaccharides, in order to access ϕ and ψ values of glycosidic linkages, and (ii) monosaccharides, in order to access θ and ϕ values. Both groups of systems underwent steps of system preparation and equilibration before the actual production phase. The set of physicochemical properties used to describe the molecules of this study was given by the GROMOS53a6*GLYC* force field, previously developed by our group[31] and successfully employed to characterize multiple examples of carbohydrates in nature.[65,66]

Dodecahedron simulation boxes were solvated using the SPC/E water model[67] in the presence of periodic boundary conditions. Energy minimization was carried out by the steepest descent algorithm, until the difference between two steps was negligible. The PME electrostatic treatment applied, and the bond lengths were constrained by the Lincs method,[68] allowing a 2 fs integration step. In the equilibration phase, during both the NVT step (volume and temperature constant) and the NPT step (pressure and temperature constant) a restriction force of 1000 kJ/mol and 500 kJ/mol was applied, respectively. Both steps were simulated for 1 ns.

After that, the systems underwent a production step, where no restriction forces were applied. The system with disaccharides were simulated for 200 ns, while the ones with the monosaccharides were simulated for 100 ns. The temperature coupling algorithm used was the V-rescale,[69] and pressure was kept constant by employing the Parrinello-Ramhan algorithm.[70,71] The collective variables used for the disaccharides were the dihedral angles

(ϕ and ψ), using a Gaussian height of 1.2 and a sigma of 0.35. In the monosaccharides systems, the puckering coordinates (θ and ϕ) were used as CVs, with a Gaussian height of 0.5 and a sigma of 0.1. The errors related to each CV of each system were calculated by the block-averaging analysis and comprised only up to 3% of the value of the barrier.

Acknowledgement

This research received funding by the Fundao de Amparo a Pesquisa do Estado do Rio Grande do Sul (FAPERGS), Conselho Nacional de Desenvolvimento Cientfico e Tecnolgico (CNP1), from CAPES/Drug Discovery grant number 23038.007777/2014-87 and from Fundacin Seneca, under grant number 19948/IV/15.

References

(1) Dwek, R. A. Glycobiology: Toward Understanding the Function of Sugars. *Chemical Reviews* **1996**, *96*, 683–720.

(2) Wormald, M. R.; Petrescu, A. J.; Pao, Y.-L.; Glithero, A.; Elliott, T.; Dwek, R. A. Conformational Studies of Oligosaccharides and Glycopeptides: Complementarity of NMR, X-ray Crystallography, and Molecular Modelling. *Chemical Reviews* **2002**, *102*, 371–386.

(3) Jens Ø. Duus,; Charlotte H. Gotfredsen,; ; Bock, K.; Charlotte H. Gotfredsen,; ; Bock, K. Carbohydrate Structural Determination by NMR Spectroscopy: Modern Methods and Limitations. *Chemical Reviews* **2000**, *100*, 4589–4614.

(4) Bubb, W. A. NMR spectroscopy in the study of carbohydrates: Characterizing the structural complexity. *Concepts in Magnetic Resonance Part A: Bridging Education and Research* **2003**, *19*, 1–19.

(5) DeMarco, M. L.; Woods, R. J. Structural glycobiology: a game of snakes and ladders. *Glycobiology* **2008**, *18*, 426–440.

(6) Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Bertozzi CR, H. G.; Etzler ME, *Essentials of Glycobiology*, 2nd ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, 2009.

(7) Berman, H.; Henrick, K.; Nakamura, H. Announcing the worldwide Protein Data Bank. *Nature Structural Biology* **2003**, *10*, 980.

(8) Joosten, R. P.; Lütteke, T. Carbohydrate 3D structure validation. *Current Opinion in Structural Biology* **2017**, *44*, 9–17.

(9) Kräutler, V.; Müller, M.; Hünenberger, P. H. Conformation, dynamics, solvation and relative stabilities of selected β-hexopyranoses in water: a molecular dynamics study with the gromos 45A4 force field. *Carbohydrate Research* **2007**, *342*, 2097–2124.

(10) Lakshmanan, T.; Sriram, D.; Priya, K.; Loganathan, D. On the structural significance of the linkage region constituents of N-glycoproteins: An X-ray crystallographic investigation using models and analogs. *Biochemical and Biophysical Research Communications* **2003**, *312*, 405–413.

(11) Pol-Fachin, L.; Fernandes, C. L.; Verli, H. GROMOS96 43a1 performance on the characterization of glycoprotein conformational ensembles through molecular dynamics simulations. *Carbohydrate Research* **2009**, *344*, 491–500.

(12) Fernandes, C.; Sachett, L.; Pol-Fachin, L.; Verli, H. GROMOS96 43a1 performance in predicting oligosaccharide conformational ensembles within glycoproteins. *Carbohydrate Research* **2010**, *345*, 663–671.

(13) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. Conformational changes of small molecules binding to proteins. *Bioorganic and Medicinal Chemistry* **1995**, *3*, 411–428.

(14) Davis, A. M.; Teague, S. J.; Kleywegt, G. J. Application and limitations of x-ray crystallographic data in structure-based ligand and drug design. *Angewandte Chemie - International Edition* **2003**, *42*, 2718–2736.

(15) Liebeschuetz, J.; Hennemann, J.; Olsson, T.; Groom, C. R. The good, the bad and the twisted: A survey of ligand geometry in protein crystal structures. *Journal of Computer-Aided Molecular Design* **2012**, *26*, 169–183.

(16) Agirre, J.; Davies, G.; Wilson, K.; Cowtan, K. Carbohydrate anomalies in the PDB. *Nature Chemical Biology* **2015**, *11*, 303.

(17) Davies, G. J.; Planas, A.; Rovira, C. Conformational analyses of the reaction coordinate of glycosidases. *Accounts of chemical research* **2012**, *45*, 308–316.

(18) Reynolds, C. H. Protein-ligand cocrystal structures: We can do better. *ACS Medicinal Chemistry Letters* **2014**, *5*, 727–729.

(19) Petrescu, A. J.; Petrescu, S. M.; Dwek, R. A.; Wormald, M. R. A statistical analysis of N- and O-glycan linkage conformations from crystallographic data. *Glycobiology* **1999**, *9*, 343–352.

(20) Lütteke, T.; Frank, M.; von der Lieth, C.-W. Data mining the protein data bank: automatic detection and assignment of carbohydrate structures. *Carbohydrate Research* **2004**, *339*, 1015–1020.

(21) Khoury, G. A.; Baliban, R. C.; Floudas, C. A. Proteome-wide post-translational modification statistics: Frequency analysis and curation of the swiss-prot database. *Scientific Reports* **2011**, *1*, 1–5.

(22) Yu, L.; Reutzel-Edens, S. M.; Mitchell, C. A. Crystallization and polymorphism of conformationally flexible molecules: Problems, patterns, and strategies. *Organic Process Research and Development* **2000**, *4*, 396–402.

(23) Schwarz, F.; Aebi, M. Mechanisms and principles of N-linked protein glycosylation. *Current Opinion in Structural Biology* **2011**, *21*, 576–582.

(24) Abraham J. Domb, Joseph Kost, D. W. *Handbook of Biodegradable Polymers*, 1st ed.; CRC Press, 1998.

(25) Gabius, H. J. The sugar code: Why glycans are so important. *BioSystems* **2018**, *164*, 102–111.

(26) Rinaudo, M. Chitin and chitosan: Properties and applications. *Progress in Polymer Science* **2006**, *31*, 603–632.

(27) Jockusch, R. A.; Kroemer, R. T.; Talbot, F. O.; Snoek, L. C.; Çarçabal, P.; Simons, J. P.; Havenith, M.; Bakker, J. M.; Compagnon, I.; Meijer, G.; Von Helden, G. Probing the Glycosidic Linkage: UV and IR Ion-Dip Spectroscopy of a Lactoside. *Journal of the American Chemical Society* **2004**, *126*, 5709–5714.

(28) Perić-Hassler, L.; Hansen, H. S.; Baron, R.; Hünenberger, P. H. Conformational properties of glucose-based disaccharides investigated using molecular dynamics simulations with local elevation umbrella sampling. *Carbohydrate Research* **2010**, *345*, 1781–1801.

(29) Rao, V. S. R.; Qasba, P. K.; Balaji, P. V.; Chandrasekaran, R. *Conformation of Carbohydrates*, 1st ed.; Amsterdam : Harwood Academic: Amsterdam, The Netherlands, 1998.

(30) Plazinski, W.; Drach, M. The influence of the hexopyranose ring geometry on the conformation of glycosidic linkages investigated using molecular dynamics simulations. *Carbohydrate Research* **2015**, *415*, 17–27.

(31) Pol-Fachin, L.; Rusu, V. H.; Verli, H.; Lins, R. D. GROMOS 53A6 GLYC, an improved GROMOS force field for hexopyranose-based carbohydrates. *Journal of Chemical Theory and Computation* **2012**, *8*, 4681–4690.

(32) Cheetham, N. W.; Dasgupta, P.; Ball, G. E. NMR and modelling studies of disaccharide conformation. *Carbohydrate Research* **2003**, *338*, 955–962.

(33) Pereira, C. S.; Kony, D.; Baron, R.; Müller, M.; van Gunsteren, W. F.; Hünenberger, P. H. Conformational and Dynamical Properties of Disaccharides in Water: a Molecular Dynamics Study. *Biophysical Journal* **2006**, *90*, 4337–4344.

(34) Heggelund, J. E.; Burschowsky, D.; Bjørnestad, V. A.; Hodnik, V.; Anderluh, G.; Krengel, U. High-Resolution Crystal Structures Elucidate the Molecular Basis of Cholera Blood Group Dependence. *PLOS Pathogens* **2016**, *12*, e1005567.

(35) Singh, B. K.; Leuthold, M. M.; Hansman, G. S. Human Noroviruses' Fondness for Histo-Blood Group Antigens. *Journal of Virology* **2015**, *89*, 2024–2040.

(36) Borguesan, B.; e Silva, M. B.; Grisci, B.; Inostroza-Ponta, M.; Dorn, M. APL: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. *Computational Biology and Chemistry* **2015**, *59*, 142–157.

(37) Pérez, S.; Marchessault, R. H. The exo-anomeric effect: experimental evidence from crystal structures. *Carbohydrate Research* **1978**, *65*, 114–120.

(38) Cremer, D.; Pople, J. A. A General Definition of Ring Puckering Coordinates. *Journal of the American Chemical Society* **1975**, *97*, 1354–1358.

(39) Hill, A. D.; Reilly, P. J. Puckering coordinates of monocyclic rings by triangular decomposition. *Journal of Chemical Information and Modeling* **2007**, *47*, 1031–1035.

(40) Bérces, A.; Whitfield, D. M.; Nukada, T. Quantitative description of six-membered ring conformations following the IUPAC conformational nomenclature. *Tetrahedron* **2001**, *57*, 477–491.

(41) Davies, G. J.; Williams, S. J. Carbohydrate-active enzymes: sequences, shapes, contortions and cells. *Biochemical Society Transactions* **2016**, *44*, 79–87.

(42) Saharay, M.; Guo, H.; Smith, J. C. Catalytic Mechanism of Cellulose Degradation by a Cellobiohydrolase, CelS. *PLoS ONE* **2010**, *5*, e12947.

(43) Verli, H.; Guimarães, J. A. Insights into the induced fit mechanism in antithrombinheparin interaction using molecular dynamics simulations. *Journal of Molecular Graphics and Modelling* **2005**, *24*, 203–212.

(44) Angyal, S. J. The Composition and Conformation of Sugars in Solution. *Angewandte Chemie International Edition in English* **1969**, *8*, 157–166.

(45) Gilbert, H. J.; Smith, N. L.; Stick, R. V.; Scaffidi, A.; Davies, G. J.; Money, V. A. Substrate Distortion by a Lichenase Highlights the Different Conformational Itineraries Harnessed by Related Glycoside Hydrolases. *Angewandte Chemie International Edition* **2006**, *45*, 5136–5140.

(46) Biarnés, X.; Ardèvol, A.; Planas, A.; Rovira, C.; Laio, A.; Parrinello, M. The conformational free energy landscape of β-D-glucopyranose. Implications for substrate preactivation in β-glucoside hydrolases. *Journal of the American Chemical Society* **2007**, *129*, 10686–10693.

(47) Jitonnom, J.; Limb, M. A. L.; Mulholland, A. J. QM/MM Free-Energy Simulations of Reaction in Serratia marcescens Chitinase B Reveal the Protonation State of Asp142 and the Critical Role of Tyr214. *The Journal of Physical Chemistry B* **2014**, *118*, 4771–4783.

(48) Jitonnom, J.; Lee, V. S.; Nimmanpipug, P.; Rowlands, H. A.; Mulholland, A. J. Quantum Mechanics/Molecular Mechanics Modeling of Substrate-Assisted Catalysis in Family 18 Chitinases: Conformational Changes and the Role of Asp142 in Catalysis in ChiB. *Biochemistry* **2011**, *50*, 4697–4711.

(49) Brás, N. F.; Santos-Martins, D.; Fernandes, P. A.; Ramos, M. J. Mechanistic Pathway

on Human α-Glucosidase Maltase-Glucoamylase Unveiled by QM/MM Calculations. *The Journal of Physical Chemistry B* **2018**, *122*, 3889–3899.

(50) Raich, L.; Borodkin, V.; Fang, W.; Castro-López, J.; van Aalten, D. M. F.; Hurtado-Guerrero, R.; Rovira, C. A Trapped Covalent Intermediate of a Glycoside Hydrolase on the Pathway to Transglycosylation. Insights from Experiments and Quantum Mechanics/Molecular Mechanics Simulations. *Journal of the American Chemical Society* **2016**, *138*, 3325–3332.

(51) Mayes, H. B.; Knott, B. C.; Crowley, M. F.; Broadbelt, L. J.; Ståhlberg, J.; Beckham, G. T. Who's on base? Revealing the catalytic mechanism of inverting family 6 glycoside hydrolases. *Chemical Science* **2016**, *7*, 5955–5968.

(52) Timmins, J.; Leiros, H.-K. S.; Leonard, G.; Leiros, I.; McSweeney, S. Crystal Structure of Maltooligosyltrehalose Trehalohydrolase from Deinococcus radiodurans in Complex with Disaccharides. *Journal of Molecular Biology* **2005**, *347*, 949–963.

(53) Cuskin, F.; Baslé, A.; Ladevèze, S.; Day, A. M.; Gilbert, H. J.; Davies, G. J.; Potocki-Véronèse, G.; Lowe, E. C. The GH130 Family of Mannoside Phosphorylases Contains Glycoside Hydrolases That Target β-1,2-Mannosidic Linkages in Candida Mannan. *The Journal of biological chemistry* **2015**, *290*, 25023–33.

(54) Lütteke, T.; von der Lieth, C.-W. pdb-care (PDB CArbohydrate REsidue check): a program to support annotation of complex carbohydrate structures in PDB files. *BMC Bioinformatics* **2004**, *5*, 69.

(55) Lütteke, T.; Frank, M.; von der Lieth, C.-W. Carbohydrate Structure Suite (CSS): analysis of carbohydrate 3D structures derived from the PDB. *Nucleic acids research* **2005**, *33*, D242–246.

(56) Toukach, P. V.; Egorova, K. S. Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts. *Nucleic Acids Research* **2016**, *44*, D1229–D1236.

(57) Ranzinger, R.; Herget, S.; Von Der Lieth, C. W.; Frank, M. GlycomeDB-A unified database for carbohydrate structures. *Nucleic Acids Research* **2011**, *39*, 373–376.

(58) Wang, L.; Berne, B. J. Efficient sampling of puckering states of monosaccharides through replica exchange with solute tempering and bond softening. *Journal of Chemical Physics* **2018**, *149*, 072306.

(59) Mayes, H. B.; Broadbelt, L. J.; Beckham, G. T. How sugars pucker: Electronic structure calculations map the kinetic landscape of fivebiologically paramount monosaccharides and their implications for enzymatic catalysis. *Journal of the American Chemical Society* **2014**, *136*, 1008–1022.

(60) Autieri, E.; Sega, M.; Pederiva, F.; Guella, G. Puckering free energy of pyranoses: A NMR and metadynamics-umbrella sampling investigation. *Journal of Chemical Physics* **2010**, *133*, 1–15.

(61) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; De Hoon, M. J. L. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423.

(62) Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Physical Review Letters* **2008**, *100*, 020603.

(63) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19–25.

(64) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Computer Physics Communications* **2014**, *185*, 604–613.

(65) Ligabue-Braun, R.; Sachett, L. G.; Pol-Fachin, L.; Verli, H. The Calcium Goes Meow: Effects of Ions and Glycosylation on Fel d 1, the Major Cat Allergen. *PLOS ONE* **2015**, *10*, e0132311.

(66) Velasquez, S. M. et al. O-Glycosylated Cell Wall Proteins Are Essential in Root Hair Growth. *Science* **2011**, *332*, 1401–1403.

(67) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The missing term in effective pair potentials. *The Journal of Physical Chemistry* **1987**, *91*, 6269–6271.

(68) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation* **2008**, *4*, 116–122.

(69) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **2007**, *126*, 014101.

(70) Nosé, S.; Klein, M. Constant pressure molecular dynamics for molecular systems. *Molecular Physics* **1983**, *50*, 1055–1076.

(71) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **1981**, *52*, 7182–7190.

5.2 Capítulo II: Ajuste de potenciais torcionais utilizando algoritmo genético

Nessa seção do trabalho, serão apresentados os resultados obtidos no desenvolvimento de um programa visando fazer o ajuste de potenciais torcionais de aldohexopiranoses utilizando um algoritmo genético não-canônico em conjunto com simulações de dinâmica molecular. Partindo do conjunto de parâmetros desenvolvidos para carboidratos de anéis de 6 membros GROMOS53a6*GLYC* [75], buscou-se, primeiramente avaliar sua representatividade no que tange o equilíbrio conformacional dessas biomoléculas.

A partir de simulações de dinâmica molecular (DM) com três diferentes monossacarídeos (α-Glicose, α-Idose e β-Idose) foi possível verificar as conformações adotadas por essas moléculas ao longo do tempo de execução utilizando as medições das coordenadas de *puckering*. Cada um dos sistemas foi simulado por 5 μs e as distribuições dos valores de θ adotados foram calculadas (figura 13). As proporções conformacionais observadas experimentalmente são as seguintes [20, 75]:

- α-Glicose: $100(^4C_1)$:$0(^1C_4)$
- α-Idose: $20(^4C_1)$:$80(^1C_4)$
- β-Idose: $75(^4C_1)$:$25(^1C_4)$

Para todos os casos testados, os potenciais torcionais utilizados não foram capazes de representar o comportamento observado na natureza, onde há um equilíbrio pseudorotacional entre as duas conformações de cadeira de cada monômero, apresentando apenas um dos estados durante toda a DM. Esse comportamento é decorrente de uma alta barreira energética entre os possíveis estados a serem adotados pelos diedros do anel dessas moléculas, que impede (em uma simulação não-enviesada) a troca de um estado conformacional para o outro. Partindo de um dos sistemas apresentados acima (α-idose), foi realizada uma execução do programa visando fazer o correto ajuste da barreira energética presente entre os estados conformacionais.

O foco central do programa era um ajuste empírico dos potenciais torcionais dos diedros responsáveis por esse equilíbrio. Conhecendo previamente os potenciais a serem ajustados para o caso teste, novos valores aleatórios de barreiras energéticas eram gerados e passavam por uma DM curta. Os resultados obtidos das DM de cada um dos valores eram passados por uma heurística que selecionava progressivamente a barreira que melhor representasse a proporção experimental em questão. Após uma execução do programa, o perfil de distribuição das conformações adotadas pelo monossacarídeos

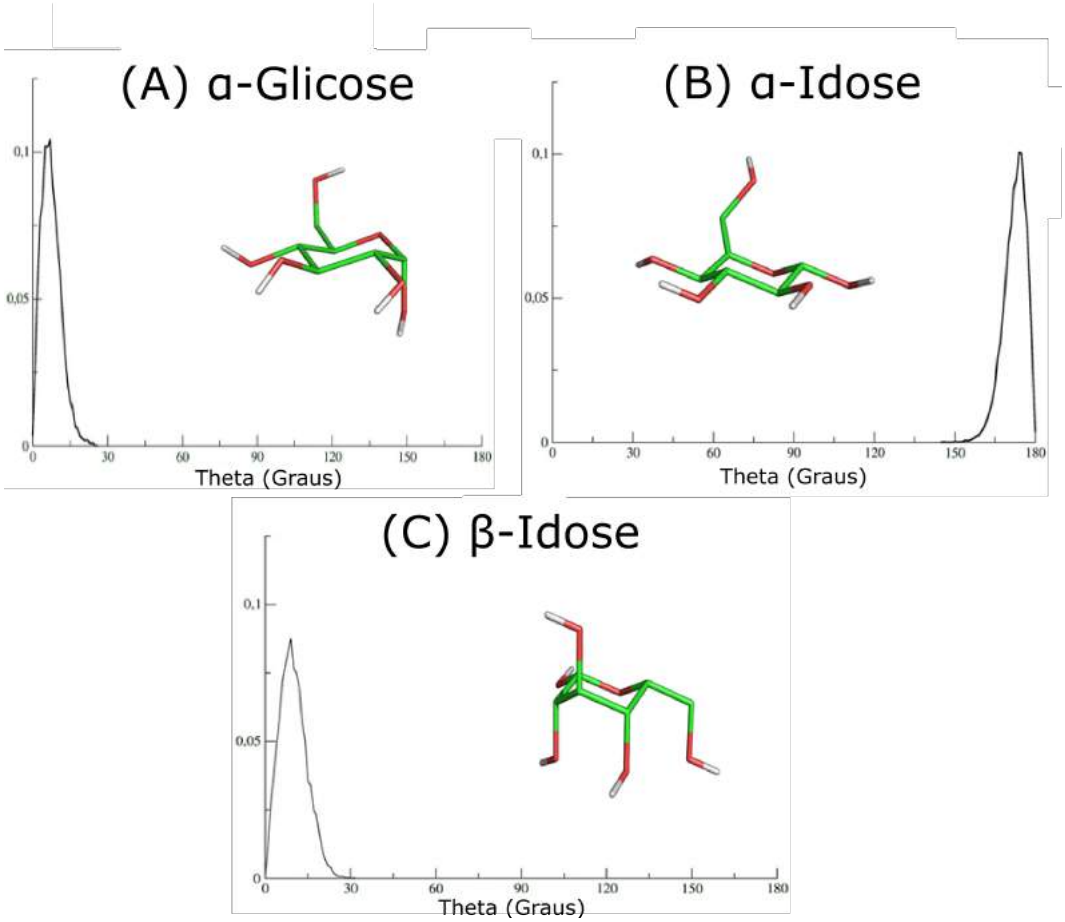


Figura 13 – Distribuição dos valores do ângulo θ ao longo dos 5 μs de simulação. Para todos os sistemas não houve a presença de outra conformação além da mais energeticamente estável. Sendo para A e C $^4C_1$ e para B $^1C_4$.

ao longo do tempo de simulação com o melhor resultado selecionado está apresentado na figura 14A. A nova proporção obtida pelo programa foi de 16($^4C_1$):78($^1C_4$) em comparação com a proporção experimental de 20($^4C_1$):80($^1C_4$), sendo o restante da porcentagem estados transicionais populados.

Na figura 14B foi avaliado o valor do ângulo θ para os *frames* calculados pela DM. Nela é possível observar que há a transição de estados conformacionais. No entanto, essa transição não ocorre em um equilíbrio onde há uma contínua troca entre estados, como seria esperado. Mesmo quando o tempo de simulação foi aumentado para 1 μs a interconversão entre os estados não foi observada da maneira esperada. Esse fenômeno pode ser justificado pelo tempo de interconversão de uma conformação de cadeira para a outra, que é da ordem de 10^3 ns [69] devido às interações entre os substituintes polares dessas moléculas e as moléculas de solvente.

Tendo em vista que a agilidade na execução do programa era um dos aspectos primordiais do trabalho e que execuções de simulações de DM da ordem de 10^3 ns para

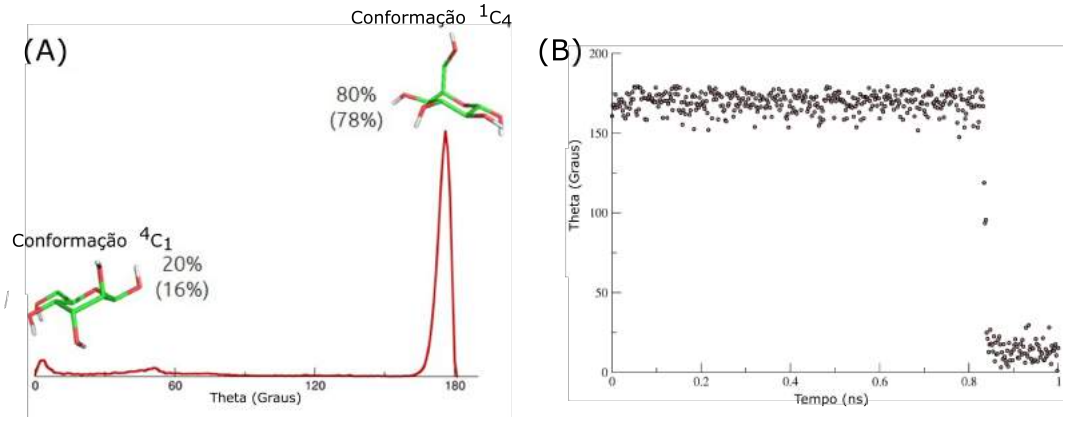


Figura 14 – Resultados obtidos após a primeira execução do programa. (A): Distribuição dos valores do ângulo θ adotados ao longo da simulação do indivíduo selecionado pelo programa, onde as proporções conformacionais experimentais encontram-se fora dos parênteses e as proporções obtidas pelo novo potencial torcional selecionado entre parênteses. (B): Valores de θ adotados pelo monossacarídeo ao longo da simulação, representando a troca de estados conformacionais de $^1C_4$ para $^4C_1$.

todos os indivíduos de todas as gerações do algoritmo genético elevaria exacerbadamente o custo computacional, mudou-se o caso teste do programa. Devido à sua falta de substituintes (apenas átomos de hidrogênios ligados ao anel), menor complexidade e, por conseguinte, menor tempo de transição, o ciclohexano passou a ser a molécula de estudo desse trabalho, na tentativa de validar a metodologia desenvolvida.

O resultado da execução do programa com o ciclohexano está apresentado na figura 15A. Nela é possível observar que, ao contrário do comportamento observado com a α-idose, há uma troca contínua de estados conformacionais, o que demonstra o equilíbrio esperado dessas estruturas em solução. A distribuição dos valores do ângulo θ obtidos pelo potencial selecionado pelo programa, $48(^4C_1)$:$48(^1C_4)$ (com o restante em estados transicionais), são também muito próximas do valor experimental conformações adotadas, $50(^4C_1)$:$50(^1C_4)$ [90]. Além disso na figura 15B foram plotados todos os valores de potencial de cada indivíduo a cada geração do algoritmo genético, mostrando que houve uma ampla ocupação do espaço amostral disponibilizado.

Dado o sucesso do programa de selecionar um novo valor de barreira energética que fosse capaz de representar as trocas de conformações do ciclohexano, aumentou-se levemente o nível de complexidade das moléculas testes. Foram realizadas execuções do programa utilizando ciclohexanol e ciclohexilamina e os resultados obtidos para essas execuções estão demonstrados na figura 16. Nela é possível perceber que, tanto para o ciclohexanol (16A) quanto para o ciclohexilamina (16B) não é contemplado o equilíbrio de conformações ao longo do tempo de dinâmica, muito embora as proporções

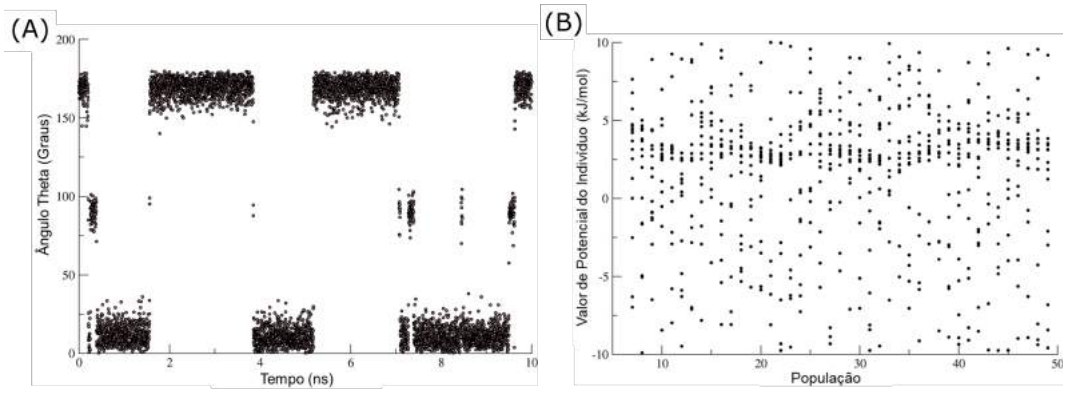


Figura 15 – Resultados obtidos após a execução do programa utilizando o ciclohexano como molécula teste. (A) Valores do ângulo θ adotados ao longo do tempo de simulação do indivíduo selecionado como o melhor. (B) Valores de barreiras energéticas gerados para cada indivíduo (kJ/mol) de acordo com cada geração criada.

experimentais e as calculadas *in silico* sejam muito próximas.

- COH: 6($^4C_1$):94($^1C_4$) [Experimental] [90]| 7($^4C_1$):92($^1C_4$) [DM]
- CNH: 96($^4C_1$):4($^1C_4$) [Experimental] [90]| 93($^4C_1$):6($^1C_4$) [DM]

A simples presença de um substituinte polar, além de mudar drasticamente a conformação preferencial do anel, impede a seleção de um potencial que seja representativo do equilíbrio conformacional em solução. Mesmo com uma ampla exploração do espaço amostral oferecido (figura 16), não foi possível obter um valor de barreira energética que permitisse esse equilíbrio.

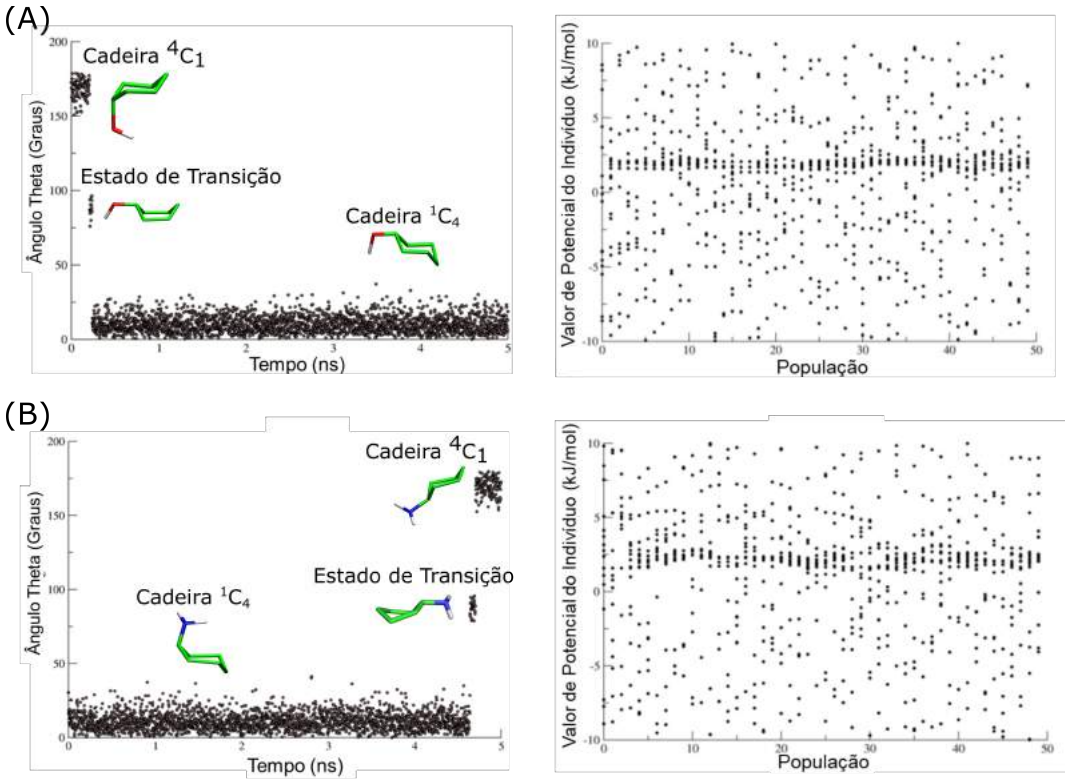


Figura 16 – Resultados obtidos após a execução do programa utilizando o ciclohexanol e a ciclohexilamina. Lado a lado os valores do ângulo θ adotados ao longo do tempo de simulação do indivíduo selecionado como o melhor e os valores de barreiras energéticas gerados para cada indivíduo (kJ/mol) de acordo com cada geração criada na execução com o ciclohexanol em (A) e com a ciclohexilamina em (B).

6 Discussão geral

Os resultados obtidos na análise do PDB, juntamente com as simulações de metadinâmica demonstram que, a partir de um certo nível de resolução, as estruturas de carboidratos depositadas no banco de dados possuem, de fato, uma boa qualidade e condizem com o comportamento esperado para os carboidratos avaliados. A informação obtida, no entanto, encontra-se condensada em alguns poucos monossacarídeos (45% das estruturas sendo de N-acetil-glicosamina e glicose) e nos níveis conformacionais mais simples (monossacarídeos e dissacarídeos compondo 78% das estruturas).

Isso deve-se, possivelmente, às diferentes abundâncias desses monômeros encontradas na natureza, visto que a N-acetil-glicosamina (NAG) é o principal componente do esqueleto de glicosilações de proteínas em eucariotos [91], uma das mais abundantes modificações co- e pós-traducionais da natureza [92]. Ademais, a glicose é o monossacarídeo mais abundantemente encontrado na natureza [93], justificando a sua presença elevada no banco de dados. Outra informação relevante é a dificuldade de cristalização de moléculas tão flexíveis quanto carboidratos, fazendo com que apenas alguns oligossacarídeos não ramificados cristalizem regularmente [52]. Ao aumentarem de tamanho, mais graus de complexidade são adicionados a essas moléculas, dificultando ainda mais a elucidação estrutural a partir da cristalografia de raios-X, principal técnica empregada na resolução de estruturas depositadas no PDB.

Ao observar as informações de ligações glicosídicas, é possível perceber, também, que há uma preferência por certos tipos de ligação. A maioria das estruturas estudadas encontra-se em uma ligação entre o C1 e C4, tanto na orientação α, quanto na orientação β. Além disso, dentre os tipos de ligação analisados utilizando metadinâmica, é possível perceber que, mesmo sendo compostos por diferentes dissacarídeos, há um preferência por certos valores de ângulos de acordo com a orientação dessa ligação. Ligações com o carbono anomérico na orientação α, tendem a popularem, tanto nos cristais do PDB quanto nos mapas de energia livre, valores de ϕ entre 60°e 120°, enquanto ligações beta tiveram a preferência de valores de ϕ entre -100°e -50°. Isso deve-se ao efeito exo-anomérico que ocorre entre os monossacarídeos, que dita as preferências atreladas a esse ângulo diedral, especificamente [94, 95]. Apesar disso, ainda é possível ver pequenas populações de ϕ que se opõe a esses valores, o que deve ocorrer devido às diferentes orientações dos substituintes dos monossacarídeos envolvidos. Como a maioria dos monômeros presentes nos cristais eram de derivados de Glicose e essa preferência do ângulo ϕ é ligada a monossacarídeos baseados em Glicose [68], pode-se supor que os

demais pertençam a monômeros mais complexos ou não baseados em Glicose.

No que tange o estudo de conformações dos monossacarídeos, a preferência pelos estados adotados pelas estruturas cristalizadas condiz com os valores de ângulo das conformações mais estáveis para cada monômero. Devido ao fato de a fucose ser encontra na natureza na sua configuração L, em oposição aos demais monossacarídeos que encontram-se na configuração D, seu estado conformacional mais estável é oposto aos demais. Os pontos que ultrapassam a linha que delimita as conformações mais estáveis podem ser decorrentes de erros atrelados a anotação desses carboidratos, algo que ocorre com certa frequência no PDB [57] ou mesmo devido ao ambiente não-biológico que é necessário para a formação do cristal, podendo causar modificações conformacionais nessas moléculas. Uma alternativa é a indução de novas conformações que, quando em solução não ocorrem espontaneamente, pelas proteínas cristalizadas junto com os monossacarídeos. Alguns estudos demonstram que existe uma modificação conformacional resultante da ação da proteína sobre os carboidratos, principalmente a fim de aproximar o centro anomérico do sítio ativo para facilitar a quebra da ligação glicosídica [27, 96, 97].

Quando observadas as superfícies de energia livre (SEL) das diferentes conformações adotadas pelos 6 monossacarídeos mais abundantes, percebe-se que as estruturas que se encontram em conformações menos estáveis estão, na grande maioria dos casos, em mínimos de energia locais. Isso indica a possível via de interconversão entre uma cadeira e outra ($^4C_1$ e $^1C_4$), mostrando que não necessariamente há erros na cristalização ou anotoção dessas moléculas. Nesse contexto, os anéis sacarídicos passariam por alguns desses pontos, adotando essa conformações durante o processo de transição. Alguns trabalhos foram realizados descrevendo a via de interconversão de diferentes monossacarídeos com diferentes níveis de resolução (métodos baseado em QM e outros baseados em MM). A partir desses trabalhos, observa-se que os mínimos de energia populados tanto pelos cristais do PDB quanto os populados pela metadinâmica são, na realidade, possíveis estados transicionais entre os dois mais estáveis [21, 23].

No caso do mapa de energia livre da NAG, por exemplo, a conformação identificada como um mínimo energético local pelas simulações de metadinâmica foi a ^{1,4}B (figura 17), correspondente aos valores de $\theta = 90°$e $\phi = [210°{:}240°]$. Essa mesma conformação foi identificada como um dos estados adotados pela NAG durante sua interação com quitinases em outros trabalhos que utilizaram metodologia de abordagem conjunta de QM/MM [96, 97]. Outros trabalhos também mostraram o mesmo comportamento para outros monossacarídeos, como é o caso da Glicose. É possível perceber um intervalo de conformações adotadas pela Glicose, de acordo com o seu

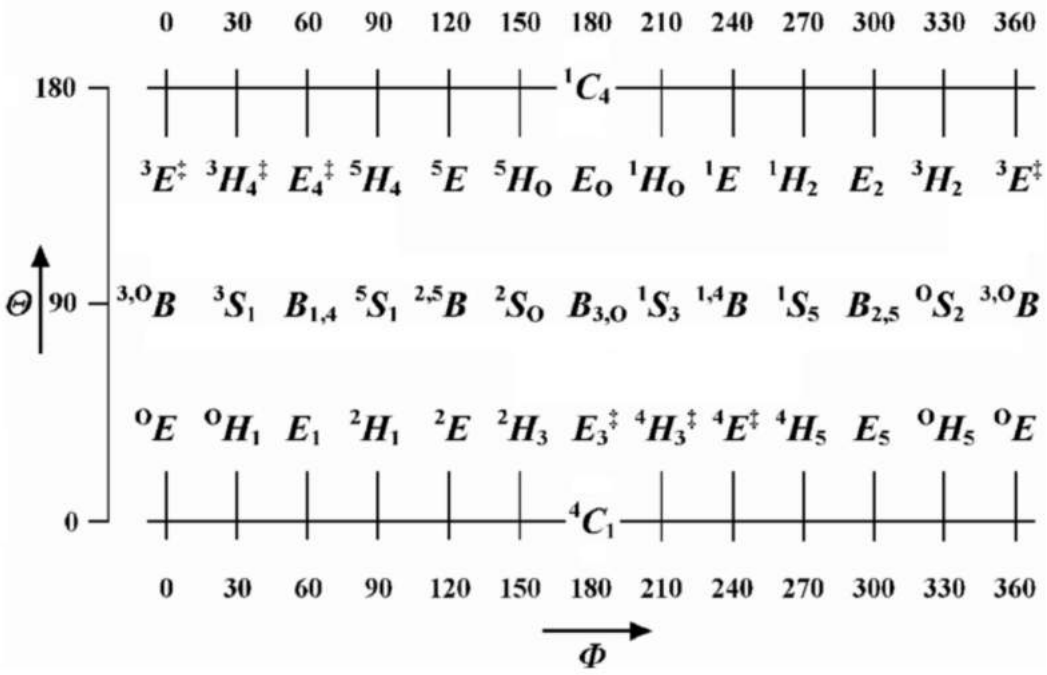


Figura 17 – Projeção planificada (Mercator) da esfera de coordenadas de *puckering* de Cremer-Pople [31].

mapa de energia livre ($^1S_5$, $B_{2,5}$, $^OS_2$ e $^1S_3$ ilustradas nas figura 17) que localizam-se em regiões de baixa energia. Essas conformações já foram descritas, também através de métodos híbridos, como possíveis de serem adotadas por esse monossacarídeo quando interagindo com uma enzima [97–105]. O fato de que existem cristais populando essas regiões de mínimo mostra que a cristalização dessas proteínas ocorreu com moléculas desses monômeros presentes no seu sítio ativo, justificando as distorções vistas pelos mapas de coordenadas de *puckering*. Isso é sustentado, ademais, pela diferença nas barreiras energéticas observadas nas simulações onde há um monossacarídeo livre e um monossacarídeo interagindo com a proteína. Há uma mudança da barreira energética entre os dois estados conformacionais e, também, uma mudança do valor de energia associado ao estado transicional, induzidos pela proteína.

Em respeito dos resultados apresentados sobre o ajuste de potenciais torcionais de hexopiranoses, fica claro que, mesmo que não houvesse a representação do enquilíbrio entre conformações, o algoritmo gerou uma mudança na barreira energética entre os estados conformacionais que permitiu que essa transição ocorresse. Em um trabalho desenvolvido por Sattelle *et al.* [69] mostrou-se, através de métodos de MM, que o tempo de interconversão entre as duas cadeiras de um monossacarídeo chega a ordem de 0.8 μs e, visto que as simulações de cada indivíduo realizadas pelo programa foram de apenas 1 ns, é improvável que houvesse essa troca contínua de estados como era esperado. Mesmo quando estendidas as simulações com a barreira energética do melhor

indivíduo até 2 μs não foi observada a troca de estado, mantendo-se apenas no estado adotado após a única transição, levando à conclusão de que o comportamento observado tratava-se de um artefato de simulação.

Ao adotar o ciclohexano como molécula teste foi possível observar a troca constante entre estados conformacionais, devido ao seu menor tempo de transição. Não só isso, mas a presença de um heterociclo na molécula do carboidrato faz com que a barreira energética entre os estados transicionais e as cadeiras seja mais elevada, dificultando a interconversão [23]. Apesar de não possuírem um heterociclo em sua estrutura, o ciclohexanol e a ciclohexilamina não popularam as duas cadeiras em equilíbrio. A presença de um grupamento extremamente polar em um dos seus substituintes, apesar de não interferir no equilíbrio da orientação dos susbtituintes [23], aumenta as interações que são realizadas com o solvente. Isso também interfere na magnitude da barreira energética que separa os dois estados, como pode ser percebido pelas diferentes proporções que correspondem a essas moléculas. Em casos onde foram extendidas as simulações com essas estruturas, o comportamento foi similar ao observado com os carboidratos, apenas uma troca de conformação e, depois, a manutenção em apenas um estado.

A adoção dos diedros internos do anel (compostos apenas por átomos de carbono) de todas as moléculas simuladas pode ser uma explicação para o comportamento observado. Quando há apenas átomos de hidrogênio, além dos átomos de carbono do anel, é possível fazer o ajuste da barreira energética que rege essa estrutura. Com a adição de grupamentos polares dentro e como substituintes do anel, o ajuste não ocorre com a mesma eficácia. Nesse caso, a barreira energética de outros diedros, compostos pelos susbtituintes e pelo heteroátomo do anel, pode complementar o ajuste realizado no interior do anel.

7 Conclusões

O presente trabalho tratou os objetivos traçados e permitiu:

- Fazer a correta extração da informação estrutural de carboidratos depositada no PDB, sendo possível observar diferentes níveis de precisão da informação de acordo com a resolução com que as estruturas foram resolvidas, bem como ilustrar a concentração dessa informação estrutural em alguns poucos níveis de organização.
- Identificar os estados conformacionais preferenciais dos 6 principais monossacarídeos encontrados no banco de dados;
- Identificar a abundância dos diferentes tipos de ligação glicosídica e a concentração, no que tange essa informação, em alguns tipos de ligação. Além disso, identificar a preferência dos ângulos diedrais adotados pelos três diferentes tipos de ligações mais abundantes de acordo com os diferentes monômeros que a compunham;
- Validar a representatividade do campo de força GROMOS53a6GLYC [75] frente ao conjunto de dados experimentais que reproduziu, tanto a superfície de energia livre dos diferentes estados conformacionais de monossacarídeos, como a dos ângulos diedrais das ligações glicosídicas;
- Ajustar o valor de barreira energética entre os principais estados conformacionais do ciclohexano sem substituintes ($^4C_1$ e $^1C_4$) no GROMOS53a6GLYC [75].

8 Perspectivas

- Ampliar o filtro de carboidratos utilizado, permitindo a obtenção de informação estrutural sobre mais moléculas dessa classe;
- Realizar simulações de metadinâmica com mais sistemas de monossacarídeos (avaliando suas conformações) e de dissacarídeos (avaliando os ângulos diedrais) para extender a validação o campo de força e ampliar o conhecimento frente a isso;
- Realizar mais simulações de metadinâmica e mais extensas de complexos proteicos interagindo com carboidratos e observar a possível indução conformacional induzida pela proteína;
- Fazer o ajuste da barreira energética entre conformações de diferentes monossacarídeos utilizando a análise da superfície de energia livre correspondente a um conjunto de valores de barreira energética.

Referências

1 DWEK, R. A. Glycobiology: Toward Understanding the Function of Sugars. *Chemical Reviews*, v. 96, p. 683–720, 1996. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.720.2160{&}rep=rep1{&}ty>. Citado 2 vezes nas páginas 16 e 17,

2 WORMALD, M. R. et al. Conformational Studies of Oligosaccharides and Glycopeptides: Complementarity of NMR, X-ray Crystallography, and Molecular Modelling. *Chemical Reviews*, v. 102, n. 2, p. 371–386, 2002. ISSN 0009-2665. Citado 2 vezes nas páginas 16 e 17,

3 Jens Ø. Duus et al. Carbohydrate Structural Determination by NMR Spectroscopy: Modern Methods and Limitations. *Chemical Reviews*, American Chemical Society, v. 100, n. 12, p. 4589–4614, 2000. Disponível em: <https://pubs-acs-org.ez45.periodicos.capes.gov.br/doi/abs/10.1021/cr990302n>. Citado 2 vezes nas páginas 16 e 23,

4 BUBB, W. A. NMR spectroscopy in the study of carbohydrates: Characterizing the structural complexity. *Concepts in Magnetic Resonance Part A: Bridging Education and Research*, v. 19, n. 1, p. 1–19, 2003. ISSN 15466086. Citado 2 vezes nas páginas 16 e 23,

5 VARKI, A. Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology*, v. 3, n. 2, p. 97–130, apr 1993. ISSN 0959-6658. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/8490246https://academic.oup.com/glycob/article-lookup/doi/10.1093/glycob/3.2.97>. Citado na página 16,

6 IUPAC. Nomenclature of Carbohydrates (Recommendations 1996). *Advances in Carbohydrate Chemistry and Biochemistry*, Academic Press, v. 52, p. 44–177, jan 1997. ISSN 0065-2318. Citado na página 16,

7 MCMURRY, J. *Organic Chemistry*. 7th. ed. [S.l.]: Physical Sciences, 2008. Citado na página 16,

8 DEMARCO, M. L.; WOODS, R. J. Structural glycobiology: a game of snakes and ladders. *Glycobiology*, NIH Public Access, v. 18, n. 6, p. 426–440, jun 2008. ISSN 1460-2423. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/18390826http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4203483>. Citado na página 17,

9 HALTIWANGER, R. S.; LOWE, J. B. Role of Glycosylation in Development. *Annual Review of Biochemistry*, v. 73, n. 1, p. 491–537, jun 2004. ISSN 0066-4154. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15189151http://www.annualreviews.org/doi/10.1146/annurev.biochem.73.011303.074043>. Citado na página 17,

10 BROWN, G. D.; GORDON, S. Immune Recognition. A new receptor for β-glucans. *Nature*, v. 413, n. 6851, p. 36–37, sep 2001. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/11544516>. Citado na página 17,

11 COBB, B. A.; KASPER, D. L. Coming of age: carbohydrates and immunity. *European Journal of Immunology*, v. 35, n. 2, p. 352–356, feb 2005. ISSN 0014-2980. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15682450http://doi.wiley.com/10.1002/eji.200425889>. Citado na página 17,

12 ROSTAND, K. S.; ESKO, J. D. Microbial adherence to and invasion through proteoglycans. *Infection and immunity*, v. 65, n. 1, p. 1–8, jan 1997. ISSN 0019-9567. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/8975885http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC174549>. Citado na página 17,

13 GABIUS, H. J. The sugar code: Why glycans are so important. *BioSystems*, v. 164, p. 102–111, 2018. ISSN 18728324. Citado na página 17,

14 KLEMM, D. et al. Cellulose: Fascinating Biopolymer and Sustainable Raw Material. *Angewandte Chemie International Edition*, John Wiley & Sons, Ltd, v. 44, n. 22, p. 3358–3393, may 2005. ISSN 1433-7851. Disponível em: <http://doi.wiley.com/10.1002/anie.200460587>. Citado na página 17,

15 RINAUDO, M. Chitin and chitosan: Properties and applications. *Progress in Polymer Science*, Pergamon, v. 31, n. 7, p. 603–632, jul 2006. ISSN 0079-6700. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0079670006000530>. Citado na página 17,

16 PELLERIN, P. et al. Glycogen in Methanothrix. *Archives of Microbiology*, Springer-Verlag, v. 146, n. 4, p. 377–381, jan 1987. ISSN 0302-8933. Disponível em: <http://link.springer.com/10.1007/BF00410939>. Citado na página 17,

17 ROACH, P. Glycogen and its Metabolism. *Current Molecular Medicine*, v. 2, n. 2, p. 101–120, mar 2002. Citado na página 17,

18 BULÉON, A. et al. Starch granules: structure and biosynthesis. *International journal of biological macromolecules*, v. 23, n. 2, p. 85–112, aug 1998. ISSN 0141-8130. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/9730163>. Citado na página 18,

19 RAO, V.; VIJAYALAKSHMI, K.; SUNDARARAJAN, P. Theoretical studies on the conformation of aldohexopyranoses. *Carbohydrate Research*, v. 17, n. 2, p. 341–352, apr 1971. ISSN 00086215. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0008621500825422>. Citado na página 18,

20 ANGYAL, S. J. The Composition and Conformation of Sugars in Solution. *Angewandte Chemie International Edition in English*, v. 8, n. 3, p. 157–166, 1969. ISSN 15213773. Citado 3 vezes nas páginas 18, 41 e 82,

21 MAYES, H. B.; BROADBELT, L. J.; BECKHAM, G. T. How sugars pucker: Electronic structure calculations map the kinetic landscape of fivebiologically paramount monosaccharides and their implications for enzymatic catalysis. *Journal of the American Chemical Society*, v. 136, n. 3, p. 1008–1022, 2014. ISSN 00027863. Citado 3 vezes nas páginas 18, 20 e 88,

22 WANG, L.; BERNE, B. J. Efficient sampling of puckering states of monosaccharides through replica exchange with solute tempering and bond softening. *Journal of Chemical Physics*, v. 149, n. 7, p. 072306, 2018. ISSN 00219606. Citado na página 18,

23 STORTZ, C. A. Conformational pathways of simple six-membered rings. *Journal of Physical Organic Chemistry*, v. 23, n. 12, p. 1173–1186, 2010. ISSN 08943230. Citado 3 vezes nas páginas 18, 88 e 90,

24 Rao, V. S. R.; Qasba, P. K.; Balaji, P. V.; Chandrasekaran, R. *Conformation of Carbohydrates*. 1st. ed. Amsterdam, The Netherlands: Amsterdam : Harwood Academic, 1998. Citado na página 18,

25 VERLI, H.; GUIMARÃES, J. A. Molecular dynamics simulation of a decasaccharide fragment of heparin in aqueous solution. *Carbohydrate Research*, v. 339, n. 2, p. 281–290, 2004. ISSN 00086215. Citado 2 vezes nas páginas 19 e 24,

26 BECKER, C. F.; GUIMARÃES, J. A.; VERLI, H. Molecular dynamics and atomic charge calculations in the study of heparin 0conformation in aqueous solution. *Carbohydrate Research*, v. 340, n. 8, p. 1499–1507, 2005. ISSN 00086215. Citado na página 19,

27 BIARNÉS, X. et al. The conformational free energy landscape of β-D-glucopyranose. Implications for substrate preactivation in β-glucoside hydrolases. *Journal of the American Chemical Society*, v. 129, n. 35, p. 10686–10693, 2007. ISSN 00027863. Citado 2 vezes nas páginas 19 e 88,

28 DAVIES, G. J.; WILLIAMS, S. J. Carbohydrate-active enzymes: sequences, shapes, contortions and cells. *Biochemical Society Transactions*, v. 44, n. 1, p. 79–87, 2016. ISSN 0300-5127. Disponível em: <http://biochemsoctrans.org/cgi/doi/10.1042/BST20150186>. Citado na página 19,

29 SCHWARZ, J. C. P. Rules for conformation nomenclature for five- and six-membered rings in monosaccharides and their derivatives. *Journal of the Chemical Society, Chemical Communications*, The Royal Society of Chemistry, n. 14, p. 505, jan 1973. ISSN 0022-4936. Disponível em: <http://xlink.rsc.org/?DOI=c39730000505>. Citado na página 19,

30 Dr. H. B. F. DIXON. Conformational Nomenclature for Fiveand Six-Membered Ring Forms of Monosaccharides and Their Derivatives. *International Union of Pure and Applied Chemistry*, v. 53, n. 10, p. 1901–1905, jan 1981. Disponível em: <http://www.degruyter.com/view/j/pac.1981.53.issue-10/pac198153101901/pac198153101901.xml>. Citado na página 19,

31 CREMER, D.; POPLE, J. A. A General Definition of Ring Puckering Coordinates. *Journal of the American Chemical Society*, v. 97, n. 6, p. 1354–1358, 1975. ISSN 15205126. Citado 5 vezes nas páginas 19, 39, 42, 46 e 89,

32 CURRY, S. Structural Biology: A Century-long Journey into an Unseen World. *Interdisciplinary Science Reviews*, Taylor & Francis, v. 40, n. 3, p. 308–328, jul 2015. ISSN 0308-0188. Disponível em: <http://www.tandfonline.com/doi/full/10.1179/0308018815Z.000000000120>. Citado na página 20,

33 GELPI, J. et al. Molecular dynamics simulations: advances and applications. *Advances and Applications in Bioinformatics and Chemistry*, Dove Press, v. 8, p. 37, nov 2015. ISSN 1178-6949. Disponível em: <https://www.dovepress.com/molecular-dynamics-simulations-advances-and-applications-peer-reviewed-article-AABC>. Citado 2 vezes nas páginas 20 e 25,

34 FISCHER, H. et al. Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale. *Journal of Applied Crystallography*, International Union of Crystallography, v. 43, n. 1, p. 101–109, feb 2010. ISSN 0021-8898. Disponível em: <http://scripts.iucr.org/cgi-bin/paper?S0021889809043076>. Citado na página 20,

35 GREENFIELD, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nature Protocols*, Nature Publishing Group, v. 1, n. 6, p. 2876–2890, dec 2006. ISSN 1754-2189. Disponível em: <http://www.nature.com/articles/nprot.2006.202>. Citado na página 21,

36 AGRAWAL, P. K. NMR spectroscopy in the structural elucidation of oligosaccharides and glycosides. *Phytochemistry*, v. 31, n. 10, p. 3307–30, oct 1992. ISSN 0031-9422. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/1368855>. Citado na página 21,

37 BERMAN, H. M. et al. The Protein Data Bank. *Nucleic Acids Research*, Oxford University Press, v. 28, n. 1, p. 235–242, jan 2000. ISSN 13624962. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.235>. Citado na página 21,

38 LÜTTEKE, T.; Von Der Lieth, C. W. Data mining the PDB for glyco-related data. *Methods in Molecular Biology*, 2009. ISSN 10643745. Citado na página 21,

39 VIJAY-KUMAR, S.; BUGG, C. E.; COOK, W. J. Structure of ubiquitin refined at 1.8 A resolution. *Journal of molecular biology*, v. 194, n. 3, p. 531–44, apr 1987. ISSN 0022-2836. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/0022283687906796http://www.ncbi.nlm.nih.gov/pubmed/3041007>. Citado na página 21,

40 YANG, Y. et al. Solution structure of proinsulin: connecting domain flexibility and prohormone processing. *The Journal of biological chemistry*, American Society for Biochemistry and Molecular Biology, v. 285, n. 11, p. 7847–7851, mar 2010. ISSN 1083-351X. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/20106974http:

//www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2832934>. Citado na página 21,

41 WLODAWER, A. et al. Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination. *FEBS Journal*, v. 280, n. 22, p. 5705–5736, nov 2013. ISSN 1742464X. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/24034303http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4080831http://doi.wiley.com/10.1111/febs.12495>. Citado na página 22,

42 Russo Krauss, I. et al. An Overview of Biological Macromolecule Crystallization. *International Journal of Molecular Sciences*, Multidisciplinary Digital Publishing Institute, v. 14, n. 6, p. 11643–11691, may 2013. ISSN 1422-0067. Disponível em: <http://www.mdpi.com/1422-0067/14/6/11643>. Citado na página 22,

43 MALUF, F. V. et al. Cristalografia de Proteínas. In: VERLI, H. (Ed.). *Bioinformática: da Biologia à Flexibilidade Molecular*. 1st. ed. São Paulo: Sociedade Brasileira de Bioquímica, 2014. cap. 13, p. 251. Disponível em: <https://www.ufrgs.br/bioinfo/ebook/>. Citado na página 22,

44 .HIGGINS, M. K.; LEA, S. M. On the state of crystallography at the dawn of the electron microscopy revolution. *Current Opinion in Structural Biology*, Elsevier Current Trends, v. 46, p. 95–101, oct 2017. ISSN 0959-440X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0959440X17300209>. Citado na página 22,

45 ACHARYA, K. R.; LLOYD, M. D. The advantages and limitations of protein crystal structures. *Trends in pharmacological sciences*, v. 26, n. 1, p. 10–4, jan 2005. ISSN 0165-6147. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15629199>. Citado na página 22,

46 KWAN, E. E.; HUANG, S. G. Structural Elucidation with NMR Spectroscopy: Practical Strategies for Organic Chemists. *European Journal of Organic Chemistry*, John Wiley & Sons, Ltd, v. 2008, n. 16, p. 2671–2688, jun 2008. ISSN 1434193X. Disponível em: <http://doi.wiley.com/10.1002/ejoc.200700966>. Citado na página 23,

47 ALMEIDA, M. d. S. RMN. In: VERLI, H. (Ed.). *Bioinformática: da Biologia à Flexibilidade Molecular*. 1st. ed. São Paulo: Sociedade Brasileira de Bioquímica, 2014. cap. 12, p. 236–251. Disponível em: <https://www.ufrgs.br/bioinfo/ebook/>. Citado na página 23,

48 ZIAREK, J. J.; BAPTISTA, D.; WAGNER, G. Recent developments in solution nuclear magnetic resonance (NMR)-based molecular biology. *Journal of Molecular Medicine*, Springer Berlin Heidelberg, v. 96, n. 1, p. 1–8, jan 2018. ISSN 0946-2716. Disponível em: <http://link.springer.com/10.1007/s00109-017-1560-2>. Citado na página 23,

49 BERMAN, H.; HENRICK, K.; NAKAMURA, H. Announcing the worldwide Protein Data Bank. *Nature Structural Biology*, v. 10, n. 12, p. 980, 2003. ISSN 10728368. Citado na página 23,

50 JOOSTEN, R. P.; LÜTTEKE, T. Carbohydrate 3D structure validation. *Current Opinion in Structural Biology*, v. 44, p. 9–17, jun 2017. ISSN 0959440X. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0959440X16301221>. Citado 2 vezes nas páginas 23 e 24,

51 KRÄUTLER, V.; MÜLLER, M.; HÜNENBERGER, P. H. Conformation, dynamics, solvation and relative stabilities of selected β-hexopyranoses in water: a molecular dynamics study with the gromos 45A4 force field. *Carbohydrate Research*, v. 342, n. 14, p. 2097–2124, 2007. ISSN 00086215. Citado 2 vezes nas páginas 23 e 24,

52 LAKSHMANAN, T. et al. On the structural significance of the linkage region constituents of N-glycoproteins: An X-ray crystallographic investigation using models and analogs. *Biochemical and Biophysical Research Communications*, v. 312, n. 2, p. 405–413, 2003. ISSN 0006291X. Citado 2 vezes nas páginas 23 e 87,

53 NICKLAUS, M. C. et al. Conformational changes of small molecules binding to proteins. *Bioorganic and Medicinal Chemistry*, v. 3, n. 4, p. 411–428, 1995. ISSN 09680896. Citado na página 23,

54 SITZMANN, M. et al. PDB ligand conformational energies calculated quantum-mechanically. *Journal of Chemical Information and Modeling*, v. 52, n. 3, p. 739–756, 2012. ISSN 15499596. Citado na página 23,

55 DAVIS, A. M.; TEAGUE, S. J.; KLEYWEGT, G. J. Application and limitations of x-ray crystallographic data in structure-based ligand and drug design. *Angewandte Chemie - International Edition*, v. 42, n. 24, p. 2718–2736, 2003. ISSN 14337851. Citado na página 23,

56 LIEBESCHUETZ, J. et al. The good, the bad and the twisted: A survey of ligand geometry in protein crystal structures. *Journal of Computer-Aided Molecular Design*, v. 26, n. 2, p. 169–183, 2012. ISSN 0920654X. Citado na página 23,

57 AGIRRE, J. et al. Carbohydrate anomalies in the PDB. *Nature Chemical Biology*, v. 11, p. 303, 2015. ISSN 15524469. Citado 2 vezes nas páginas 23 e 88,

58 DAVIES, G. J.; PLANAS, A.; ROVIRA, C. Conformational analyses of the reaction coordinate of glycosidases. *Accounts of chemical research*, v. 45, n. 2, p. 308–316, 2012. ISSN 15204898. Citado na página 24,

59 REYNOLDS, C. H. Protein-ligand cocrystal structures: We can do better. *ACS Medicinal Chemistry Letters*, v. 5, n. 7, p. 727–729, 2014. ISSN 19485875. Citado na página 24,

60 PETRESCU, A. J. et al. A statistical analysis of N- and O-glycan linkage conformations from crystallographic data. *Glycobiology*, v. 9, n. 4, p. 343–352, 1999. ISSN 09596658. Citado na página 24,

61 LÜTTEKE, T.; FRANK, M.; LIETH, C.-W. von der. Data mining the protein data bank: automatic detection and assignment of carbohydrate structures. *Carbohydrate Research*, Elsevier, v. 339, n. 5, p. 1015–1020, apr 2004. ISSN 0008-6215. Disponível em:

<https://www.sciencedirect.com/science/article/pii/S0008621503005755>. Citado na página 24,

62 WARSHEL, A.; LEVITT, M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology*, Academic Press, v. 103, n. 2, p. 227–249, may 1976. ISSN 0022-2836. Disponível em: <https://www.sciencedirect.com/science/article/pii/0022283676903119>. Citado na página 24,

63 STANSFELD, P. J. Computational studies of membrane proteins: from sequence to structure to simulation. *Current Opinion in Structural Biology*, v. 45, p. 133–141, aug 2017. ISSN 0959440X. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0959440X17300465>. Citado na página 24,

64 BEST, R. B. Computational and theoretical advances in studies of intrinsically disordered proteins. *Current Opinion in Structural Biology*, Elsevier Current Trends, v. 42, p. 147–154, feb 2017. ISSN 0959-440X. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0959440X17300246>. Citado na página 24,

65 HUBER, R. G. et al. Multiscale molecular dynamics simulation approaches to the structure and dynamics of viruses. *Progress in Biophysics and Molecular Biology*, v. 128, p. 121–132, sep 2017. ISSN 00796107. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/27765545https://linkinghub.elsevier.com/retrieve/pii/S0079610716300815>. Citado na página 24,

66 TINOCO, I.; WEN, J.-D. Simulation and analysis of single-ribosome translation. *Physical Biology*, IOP Publishing, v. 6, n. 2, p. 025006, jul 2009. ISSN 1478-3975. Disponível em: <http://stacks.iop.org/1478-3975/6/i=2/a=025006?key=crossref.bbc70d0e95351e0d37667a63277b2e01>. Citado na página 24,

67 BARDUCCI, A.; BUSSI, G.; PARRINELLO, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Physical Review Letters*, v. 100, n. 2, p. 020603, jan 2008. ISSN 0031-9007. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/18232845https://link.aps.org/doi/10.1103/PhysRevLett.100.020603>. Citado na página 24,

68 PERIĆ-HASSLER, L. et al. Conformational properties of glucose-based disaccharides investigated using molecular dynamics simulations with local elevation umbrella sampling. *Carbohydrate Research*, v. 345, n. 12, p. 1781–1801, 2010. ISSN 00086215. Citado 2 vezes nas páginas 24 e 87,

69 SATTELLE, B. M.; ALMOND, A. Is N-acetyl-d-glucosamine a rigid 4C 1 chair? *Glycobiology*, v. 21, n. 12, p. 1651–1662, 2011. ISSN 09596658. Citado 3 vezes nas páginas 24, 83 e 89,

70 LEACH, A. R. *Molecular modelling : principles and applications*. 2nd ed.. ed. Harlow England ;;New York: Prentice Hall, 2001. 744 p. ISBN 9780582382107. Disponível em:

<https://www.worldcat.org/title/molecular-modelling-principles-and-applications/oclc/45008511>. Citado na página 25,

71 CORNELL, W. D. et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, American Chemical Society, v. 117, n. 19, p. 5179–5197, may 1995. ISSN 0002-7863. Disponível em: <https://pubs.acs.org/doi/abs/10.1021/ja00124a002>. Citado na página 26,

72 MACKERELL, A. D. et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins [†]. *The Journal of Physical Chemistry B*, American Chemical Society, v. 102, n. 18, p. 3586–3616, apr 1998. ISSN 1520-6106. Disponível em: <https://pubs.acs.org/doi/10.1021/jp973084f>. Citado na página 26,

73 OOSTENBRINK, C. et al. Validation of the 53A6 GROMOS force field. *European Biophysics Journal*, Springer-Verlag, v. 34, n. 4, p. 273–284, jun 2005. ISSN 0175-7571. Disponível em: <http://link.springer.com/10.1007/s00249-004-0448-6>. Citado na página 26,

74 William L. Jorgensen, . et al. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. American Chemical Society, 1996. Disponível em: <https://pubs.acs.org/doi/abs/10.1021/ja9621760>. Citado na página 26,

75 POL-FACHIN, L. et al. GROMOS 53A6 GLYC, an improved GROMOS force field for hexopyranose-based carbohydrates. *Journal of Chemical Theory and Computation*, v. 8, n. 11, p. 4681–4690, 2012. ISSN 15499618. Citado 8 vezes nas páginas 26, 30, 40, 41, 43, 46, 82 e 91,

76 POL-FACHIN, L.; VERLI, H.; LINS, R. D. Extension and validation of the GROMOS 53A6 _{glyc} parameter set for glycoproteins. *Journal of Computational Chemistry*, John Wiley & Sons, Ltd, v. 35, n. 29, p. 2087–2095, nov 2014. ISSN 01928651. Disponível em: <http://doi.wiley.com/10.1002/jcc.23721>. Citado na página 26,

77 ARANTES, P. R. et al. Development of GROMOS-Compatible Parameter Set for Simulations of Chalcones and Flavonoids. *The Journal of Physical Chemistry B*, American Chemical Society, v. 123, n. 5, p. 994–1008, feb 2019. ISSN 1520-6106. Disponível em: <http://pubs.acs.org/doi/10.1021/acs.jpcb.8b10139>. Citado na página 26,

78 PEDEBOS, C.; VERLI, H. *Estrutura e dinâmica de oligossacariltransferases procarióticas*. Tese (Doutorado), 2017. Citado na página 27,

79 BARDUCCI, A.; BONOMI, M.; PARRINELLO, M. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, John Wiley & Sons, Ltd (10.1111), v. 1, n. 5, p. 826–843, sep 2011. ISSN 17590876. Disponível em: <http://doi.wiley.com/10.1002/wcms.31>. Citado na página 26,

80 LAIO, A.; GERVASIO, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, IOP Publishing, v. 71, n. 12, p. 126601, dec 2008. ISSN 0034-4885. Disponível em: <http://stacks.iop.org/0034-4885/71/i=12/a=126601?key=crossref.2dcba907622222368898e968e8d42594>. Citado na página 26,

81 LAIO, A.; PARRINELLO, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America*, National Academy of Sciences, v. 99, n. 20, p. 12562–12566, oct 2002. ISSN 0027-8424. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/12271136http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC130499>. Citado na página 26,

82 COCK, P. J. A. et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, v. 25, n. 11, p. 1422–1423, 2009. ISSN 13674803. Citado 3 vezes nas páginas 32, 34 e 36,

83 HILL, A. D.; REILLY, P. J. Puckering coordinates of monocyclic rings by triangular decomposition. *Journal of Chemical Information and Modeling*, v. 47, n. 3, p. 1031–1035, 2007. ISSN 15499596. Citado na página 39,

84 TRIBELLO, G. A. et al. PLUMED 2: New feathers for an old bird. oct 2013. Disponível em: <http://arxiv.org/abs/1310.0980http://dx.doi.org/10.1016/j.cpc.2013.09.018>. Citado 3 vezes nas páginas 39, 40 e 41,

85 BERENDSEN, H. J. C.; GRIGERA, J. R.; STRAATSMA, T. P. The missing term in effective pair potentials. *The Journal of Physical Chemistry*, American Chemical Society, v. 91, n. 24, p. 6269–6271, nov 1987. ISSN 0022-3654. Disponível em: <http://pubs.acs.org/doi/abs/10.1021/j100308a038>. Citado 2 vezes nas páginas 40 e 43,

86 HESS, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation*, v. 4, n. 1, p. 116–122, jan 2008. ISSN 1549-9618. Disponível em: <https://pubs.acs.org/doi/10.1021/ct700200b>. Citado 2 vezes nas páginas 40 e 43,

87 BUSSI, G.; DONADIO, D.; PARRINELLO, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, v. 126, n. 1, p. 014101, jan 2007. ISSN 0021-9606. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/17212484http://aip.scitation.org/doi/10.1063/1.2408420>. Citado 2 vezes nas páginas 40 e 43,

88 NOSÉ, S.; KLEIN, M. Constant pressure molecular dynamics for molecular systems. *Molecular Physics*, Taylor & Francis Group, v. 50, n. 5, p. 1055–1076, dec 1983. ISSN 0026-8976. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/00268978300102851>. Citado 2 vezes nas páginas 40 e 44,

89 PARRINELLO, M.; RAHMAN, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, American Institute of Physics, v. 52, n. 12, p. 7182–7190, dec 1981. ISSN 0021-8979. Disponível em: <http://aip.scitation.org/doi/10.1063/1.328693>. Citado 2 vezes nas páginas 40 e 44,

90 Eusebio Juaristi. *Conformational Behavior of Six-Membered Rings : Analysis, Dynamics and Stereoelectronic Effects*. John Wiley and Sons Ltd, 1995. ISBN 0471186058. Disponível em: <https://www.bookdepository.com/Conformational-Behavior-Six-Membered-Rings-Eusebio-Juaristi/9780471186052>. Citado 2 vezes nas páginas 84 e 85,

91 SCHWARZ, F.; AEBI, M. Mechanisms and principles of N-linked protein glycosylation. *Current Opinion in Structural Biology*, Elsevier Ltd, v. 21, n. 5, p. 576–582, 2011. ISSN 0959440X. Disponível em: <http://dx.doi.org/10.1016/j.sbi.2011.08.005>. Citado na página 87,

92 KHOURY, G. A.; BALIBAN, R. C.; FLOUDAS, C. A. Proteome-wide post-translational modification statistics: Frequency analysis and curation of the swiss-prot database. *Scientific Reports*, v. 1, p. 1–5, 2011. ISSN 20452322. Citado na página 87,

93 Abraham J. Domb, Joseph Kost, D. W. *Handbook of Biodegradable Polymers*. 1st. ed. [S.l.]: CRC Press, 1998. Citado na página 87,

94 MACKIE, W. Carbohydrates structure and biology. *Biochemical Education*, John Wiley & Sons, Ltd, v. 26, n. 3, p. 257, jul 1998. ISSN 03074412. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0307441298001496>. Citado na página 87,

95 PÉREZ, S.; MARCHESSAULT, R. H. The exo-anomeric effect: experimental evidence from crystal structures. *Carbohydrate Research*, Elsevier, v. 65, n. 1, p. 114–120, aug 1978. ISSN 0008-6215. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0008621500842184>. Citado na página 87,

96 THOMPSON, A. J. et al. Evidence for a Boat Conformation at the Transition State of GH76 α-1,6-Mannanases-Key Enzymes in Bacterial and Fungal Mannoprotein Metabolism. *Angewandte Chemie International Edition*, v. 54, n. 18, p. 5378–5382, apr 2015. ISSN 14337851. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/25772148http://doi.wiley.com/10.1002/anie.201410502>. Citado na página 88,

97 JITONNOM, J. et al. Quantum Mechanics/Molecular Mechanics Modeling of Substrate-Assisted Catalysis in Family 18 Chitinases: Conformational Changes and the Role of Asp142 in Catalysis in ChiB. *Biochemistry*, American Chemical Society, v. 50, n. 21, p. 4697–4711, may 2011. ISSN 0006-2960. Disponível em: <https://pubs.acs.org/doi/10.1021/bi101362g>. Citado 2 vezes nas páginas 88 e 89,

98 BRÁS, N. F. et al. Mechanistic Pathway on Human α-Glucosidase Maltase-Glucoamylase Unveiled by QM/MM Calculations. *The Journal of Physical Chemistry B*, American Chemical Society, v. 122, n. 14, p. 3889–3899, apr 2018. ISSN 1520-6106. Disponível em: <https://pubs.acs.org/doi/10.1021/acs.jpcb.8b01321>. Citado na página 89,

99 RAICH, L. et al. A Trapped Covalent Intermediate of a Glycoside Hydrolase on the Pathway to Transglycosylation. Insights from Experiments and Quantum

Mechanics/Molecular Mechanics Simulations. *Journal of the American Chemical Society*, American Chemical Society, v. 138, n. 10, p. 3325–3332, mar 2016. ISSN 0002-7863. Disponível em: <https://pubs.acs.org/doi/10.1021/jacs.5b10092>. Citado na página 89,

100 TANKRATHOK, A. et al. A Single Glycosidase Harnesses Different Pyranoside Ring Transition State Conformations for Hydrolysis of Mannosides and Glucosides. *ACS Catalysis*, American Chemical Society, v. 5, n. 10, p. 6041–6051, oct 2015. ISSN 2155-5435. Disponível em: <https://pubs.acs.org/doi/10.1021/acscatal.5b01547>. Citado na página 89,

101 JITONNOM, J.; LIMB, M. A. L.; MULHOLLAND, A. J. QM/MM Free-Energy Simulations of Reaction in Serratia marcescens Chitinase B Reveal the Protonation State of Asp142 and the Critical Role of Tyr214. *The Journal of Physical Chemistry B*, American Chemical Society, v. 118, n. 18, p. 4771–4783, may 2014. ISSN 1520-6106. Disponível em: <https://pubs.acs.org/doi/10.1021/jp500652x>. Citado na página 89,

102 SAHARAY, M.; GUO, H.; SMITH, J. C. Catalytic Mechanism of Cellulose Degradation by a Cellobiohydrolase, CelS. *PLoS ONE*, Public Library of Science, v. 5, n. 10, p. e12947, oct 2010. ISSN 1932-6203. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0012947>. Citado na página 89,

103 PETERSEN, L. et al. Mechanism of Cellulose Hydrolysis by Inverting GH8 Endoglucanases: A QM/MM Metadynamics Study. *The Journal of Physical Chemistry B*, American Chemical Society, v. 113, n. 20, p. 7331–7339, may 2009. ISSN 1520-6106. Disponível em: <https://pubs.acs.org/doi/10.1021/jp811470d>. Citado na página 89,

104 MAYES, H. B. et al. Who's on base? Revealing the catalytic mechanism of inverting family 6 glycoside hydrolases. *Chemical Science*, The Royal Society of Chemistry, v. 7, n. 9, p. 5955–5968, aug 2016. ISSN 2041-6520. Disponível em: <http://xlink.rsc.org/?DOI=C6SC00571C>. Citado na página 89,

105 BIARNÉS, X. et al. Substrate Distortion in the Michaelis Complex of Bacillus 1,3–1,4-$\beta$$-*Glucanase*.*Journal of Biological Chemistry*, *AmericanSocietyforBiochemistryandMo* $-1441, jan2006.ISSN0021-9258.Disponívelem: <>$. Citado na página 89,

Apêndices

Molecular Neurobiology
https://doi.org/10.1007/s12035-019-1528-z

Niemann-Pick Disease Type C: Mutation Spectrum and Novel Sequence Variations in the Human *NPC1* Gene

Márcia Polese-Bonatto[1,2] · Hugo Bock[1,2] · Ana Carolina S. Farias[1] · Rafaella Mergener[1] · Maria Cristina Matte[1] · Mirela S. Gil[1] · Felipe Nepomuceno[3] · Fernanda T. S. Souza[4,5] · Rejane Gus[4] · Roberto Giugliani[1,4,5,6] · Maria Luiza Saraiva-Pereira[1,2,4,6,7]

Received: 24 September 2018 / Accepted: 15 February 2019

Abstract
Niemann-Pick type C (NP-C) is a rare autosomal recessive disorder characterized by storage of unesterified glycolipids and cholesterol in lysosome and/or late endosome due to mutations in either *NPC1* or *NPC2* gene. This study aims to identify the spectrum of sequence alterations associated to NP-C in individuals with clinical suspicion of this disease. The entire coding region and flanking sequences of both genes associated to NP-C were evaluated in a total of 265 individuals that were referred to our laboratory. Clinical and/or biochemical suspicion of NP-C was confirmed by molecular analysis in 54 subjects. In this cohort, 33 different sequence alterations were identified in *NPC1* and one in *NPC2*. Among those, 5 novel alterations in *NPC1* gene were identified as follows: one deletion (p.Lys38_Tyr40del), one frameshift (p.Asn195Lysfs*2), and three missense mutations (p.Cys238Arg, p.Ser365Pro and, p.Val694Met) that are likely to be pathogenic through different approaches, including in silico tools as well as multiple sequence alignment throughout different species. We have also reported main clinical symptoms of patients with novel alterations and distribution of frequent symptoms in the cohort. Findings reported here contribute to the knowledge of mutation spectrum of NP-C, defining frequent mutations as well as novel sequence alterations associated to the disease.

Keywords Niemann-Pick type C disease · *NPC1* gene · *NPC2* gene · Mutation spectrum · Novel variation

Introduction

Niemann-Pick type C disease (NP-C disease; OMIM #257220) is a rare autosomal recessive neurodegenerative disorder characterized by storage of unesterified glycolipids and cholesterol in lysosome and/or late endosome (LE/L) due to mutations in either *NPC1* or *NPC2* genes [1]. This disorder causes premature death, and subjects from different ethnic groups can be affected [2–4]. NP-C prevalence is approximately 1/100,000 live births, but incidence can vary among different countries [4]. Hepatosplenomegaly, vertical supranuclear ophthalmoplegia, progressive ataxia, dystonia, and dementia are among symptoms characterized as the "classic" phenotype [5–7]. Mutations in genes coding for the large transmembrane endosomal NPC1 and a small soluble lysosomal NPC2 proteins result in intracellular sterol

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12035-019-1528-z) contains supplementary material, which is available to authorized users.

✉ Maria Luiza Saraiva-Pereira
mlpereira@hcpa.edu.br

[1] Laboratório de Identificação Genética, Centro de Pesquisa Experimental, Hospital de Clínicas de Porto Alegre, Porto Alegre, RS, Brazil

[2] Programa de Pós-Graduação em Ciências Biológicas: Bioquímica, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

[3] Programa de Pós-Graduação em Biologia Celular e Molecular, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

[4] Serviço de Genética Médica, Hospital de Clínicas de Porto Alegre, Porto Alegre, RS, Brazil

[5] Departamento de Genética, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

[6] INAGEMP—Instituto Nacional de Genética Médica Populacional, Porto Alegre, RS, Brazil

[7] Departamento de Bioquímica, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

Published online: 28 February 2019

cycling alterations [8]. Great majority of NP-C cases is due to mutations in *NPC1* gene (95%) whereas the remaining are caused by mutations in *NPC2* gene [4, 9]. The human *NPC1* gene is located at *locus* 18q11, spans more than 47 kb, and is organized into 25 exons. The transcript is 4.9 kb long encoding a protein with 1278 amino acids [10]. NPC1 protein has 13 transmembrane domains, 3 large and 4 small luminal loops, 6 small cytoplasmic loops, and a cytoplasmic tail [11]. High homology was observed among NPC1 protein and other NP-C orthologs, such as mouse, *Saccharomyces cerevisiae*, and *Caenorhabditis elegans* [12]. High sequence homology between NPC1 and other proteins that are involved in cholesterol metabolism was also observed [13, 14]. The human *NPC2* gene is located at *locus* 14q24.3, spans more than 13 kb, and is organized into 5 exons. The transcript of 0.9 kb produces a small soluble glycoprotein that contains 131 amino acid residues [5, 15]. *NPC1* and *NPC*2 genes have many mammalian orthologs with highly conserved primary sequences [16].

Diagnosis of NP-C requires biochemical evaluation, such as Filipin staining test in fibroblasts or plasma oxysterols evaluation, and/or molecular analysis of *NPC1* and *NPC2* genes [5, 9]. To date, more than 460 different sequence alterations have been reported to be associated to NP-C.

This study describes the mutation spectrum of a broad genetic analysis in a cohort of patients with NP-C, including five novel sequence variants and rare mutations.

Material and Methods

Patients

In this study, we have included biological samples from 265 individuals that were sent to our laboratory from different regions of Brazil, from 2011 to 2017, through the NPC Network. Inclusion criteria were positive or inconclusive result in the Filipin staining test or a strong clinical suspicion of NP-C, regardless the biochemical evaluation outcome. This study was approved by our local Institutional Review Board (project number 05168).

DNA Isolation and Amplification of *NPC1* and *NPC2* Genes

Genomic DNA was isolated from peripheral white blood cells using standard protocols and stored at – 20 °C. Polymerase chain reaction (PCR) was used to selectively amplify specific fragments of *NPC1* (NG_012795.1) and *NPC2* (NG_007117.1) genes. Primer sequences can be found in Supplementary Table S1. Coding sequences and flanking regions (exons 1 to 25 of *NPC1* gene and exons 1 to 5 of *NPC2* gene) were amplified by PCR using genomic DNA as template. The whole coding region of *NPC1* was divided into 24 different amplicons (exons 15 and 16 were amplified as one fragment). *NPC2* coding region was divided into 5 different amplicons. Amplification reaction was performed in final volumes of 25 μL containing 25 ng genomic DNA, 200 mM of each dNTP, 2.5 μM of each primer (forward and reverse), 2.5 mM of $MgCl_2$, 200 mM of Tris-HCl (pH 8.4), 50 mM of KCl, and 1.25 U of *Taq* DNA Polymerase (Invitrogen™, Carlsbad, CA, USA). Cycling conditions were initial denaturation at 95 °C for 5 min, followed by 30 cycles of denaturation at 95 °C for 30 s, annealing at 60 °C for 30 s, and extension at 72 °C for 1 min, followed by final extension at 72 °C for 10 min. Each PCR product was verified by electrophoresis on a 1.5% (*w*/*v*) agarose gel and visualization under UV light.

DNA Sequencing

Amplicons were purified using 2.5 U of Exonuclease I (USB, Cleveland, OH, USA) and 0.25 U of Shrimp Alkaline Phosphatase (USB, Cleveland, OH, USA). DNA sequencing was performed using BigDye® Terminator Cycle Sequencing kit v. 3.1 (Applied Biosystems, Foster City, CA, USA) from universal M13 (− 20) forward and reverse primers, following the manufacturer's instructions. Sequences were analyzed with DNA Sequencing Analysis software v. 5.2 (Applied Biosystems) in an ABI PRISM® 3130*xl* Genetic Analyzer. All identified sequence variations were confirmed by sequencing an independent sample from both forward and reverse primers. Sequence variations were compared to data available in the NP-C database in the Human Gene Mutation Database (HGMD® - http://www.hgmd.cf.ac.uk), the Exome Aggregation Consortium (ExAC) browser (http://exac.broadinstitute.org/), the Genome Aggregation Database (gnomAD) browser (http://gnomad.broadinstitute.org/), and 1000genomes (http://www.internationalgenome.org/home).

Evaluation of Novel Mutations

Amino acid sequences of *NPC1* from 10 different species were compared by multiple alignment in order to determine whether changes identified in their amino acid sequences were associated to conserved residues. Sequences were searched for using the protein database from the National Center for Biotechnology Information (NCBI - https://www.ncbi.nlm.nih.gov/). Amino acid sequences were aligned with Clustal Omega using FASTA format (https://www.ebi.ac.uk/Tools/msa/clustalo/). In order to assess their potential pathogenicity, novel sequence variations in the *NPC1* coding region were analyzed using eight web-based tools. Those tools were PolyPhen-2 (Polymorphism Phenotyping v2, http://genetics.bwh.harvard.edu/pph2/) [17], SNPs3D (http://www.snps3d.org/) [18], Align GVGD (http://agvgd.iarc.fr/) [19], Mutation Taster (http://www.mutationtaster.org/) [20], Mendelian Clinically Applicable Pathogenicity (M-CAP) Score (http://bejerano.

Table 1 Alleles defined by this study. Novel sequence variants are shown in italics

Mutation	cDNA nucleotide substitution	Exon	# of alleles	Allelic frequence
NPC1 gene				
p.Lys38_40Tyrdel	*c.114-122del9*	*2*	*1*	*0.011*
p.Gln117*	c.349C > T	4	1	0.011
p.Cys177Tyr	c.530G > A	5	2	0.023
p.Ala183Thr + *p.Ser365Pro*	c.547G > A + *c.1093 T > C*	5 and 8	1	0.011
p.Ser151Phefs*70	c.451_452delAG	4	1	0.011
*p.Asn195Lysfs*2*	*c.584dupA*	*5*	*1*	*0.011*
p.Cys238Arg	*c.712 T > C*	*6*	*1*	*0.011*
p.Cys247Tyr	c.740G > A	6	1	0.011
p.Arg372Trp	c.1114C > T	8	1	0.011
p.Arg615His	c.1844G > A	12	1	0.011
p.Val664Met	c.1990G > A	13	2	0.023
p.Ser667Leu	c.2000C > T	13	1	0.011
p.Val694Met	*c.2080G > A*	*13*	*1*	*0.011*
p.Gly710Alafs*19	c.2129delA	13	2	0.023
p.Pro733Serfs*10	c.2196_2197insT	14	1	0.011
p.Ala764Val	c.2291C > T	15	1	0.011
p.Ser865Leu	c.2594C > T	17	1	0.011
p.Ala926Thr	c.2776G > A	18	1	0.011
p.Trp942Cys	c.2826G > T	19	1	0.011
p.Asp945Asn	c.2833G > A	19	1	0.011
p.Cys957Tyr	c.2870G > A	19	1	0.011
p.Gly992Arg	c.2974G > C	20	1	0.011
p.Pro1007Ala	c.3019C > G	20	15	0.170
p.Ala1035Val	c.3104C > T	21	24	0.273
p.Ile1061Thr	c.3182 T > C	21	4	0.045
p.Gly1140Val	c.3419G > T	22	1	0.011
p.Asn1156Ile	c.3467A > T	22	1	0.011
p.Leu1157Pro	c.3470 T > C	22	1	0.011
p.Val1165Met	c.3493G > A	23	1	0.011
p.Glu1166Lys	c.3496G > A	23	1	0.011
p.Arg1186His	c.3557G > A	23	2	0.023
p.Phe1221Serfs*20	c.3662delT	24	13	0.148
Total			88	
NPC2 gene				
p.Glu20*	c.58G > T	2	2	1.000
Total			2	

stanford.edu/mcap/) [21], Combined Annotation Dependent Depletion (CADD) (http://cadd.gs.washington.edu/snv) [22], Rare Exome Variant Ensemble level (REVEL) (https://sites.google.com/site/revelgenomics/) [23], and Variant Effect Scoring Tool (VEST3) (http://hg19.cravat.us/CRAVAT/) [24]. Mutalyzer 2.0 was used as a reference for naming novel sequence variations (https://mutalyzer.nl/) [25]. Model structure of *NPC1* was generated by PyMOL 2.0 software (https://pymol.org/2/), and mutant models by Modeler 9.1 software [26], using PDB ID code 3JD8.

Results

Genotype of NP-C patients was defined through identification of novel and/or rare sequence alterations in *NPC1* or *NPC2* genes. Mutant alleles were confirmed in 54 out of 265 individuals, being 29 (53.7%) females and 25 (46.3%) males. Within NP-C confirmed patients, 18 (33.3%) were from consanguineous marriage. Among NP-C cases identified in this work, 52 (96.3%) patients have mutations in *NPC1* gene while the remaining 2 (3.7%) patients carry mutations in *NPC2*. In

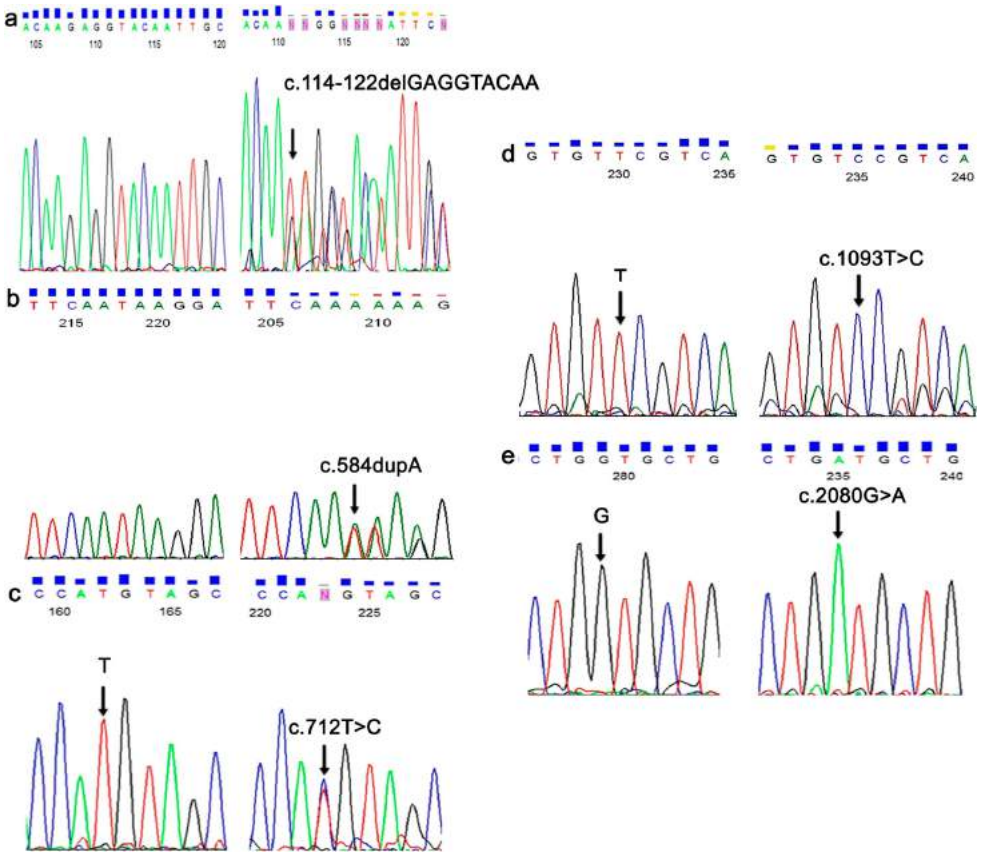


Fig. 1 DNA sequencing of novel sequence variations identified in *NPC1* gene. (a) Direct sequencing of exon 2 from the forward primer. Arrow indicates the beggiging of the deletion in the p.Lys38_Tyr40del (c.114_122delGAGGTACAA) variation. Patient is heterozygous for this variation; therefore, after sequence variation, two different profiles can be seen in the figure: one from wild-type allele and the other from mutant allele. (b) Direct sequencing of exon 5 from the forward primer. The arrow shows nucleotide duplication that characterizes the p.Asn195Lysfs*2 (c.584dupA) variation. Patient is heterozygous for this variation; therefore, after sequence variation, two different profiles can be seen in the figure: one from wild-type allele and the other from mutant allele. (c) Direct sequencing of exon 6 from the forward primer. The arrow indicates T to C substitution in the p.Cys238Arg (c.712 T > C) variation. Patient is heterozygous for this variation. (d) Direct sequencing of part of exon 8 from the forward primer. The arrow indicates T to C substitution in the p.Ser365Pro (c.1093 T > C) variation. Patient is homozygous for this variation. (e) Direct sequencing of part of exon 13 from the forward primer. The arrow indicates G to A substitution in the p.Val694Met (c.2080G > A) variation. Patient is homozygous for this variation

total, 34 different sequence alterations were identified, including 5 novel variations in *NPC1*, and a detailed distribution of mutations is shown in Table 1. Frequency of variants was estimated using unrelated chromosomes only; i.e., we have just considered one allele from homozygous patients of consanguineous marriages, giving a total of 90 alleles. The most frequent mutation was p.Ala1035Val (27.0%), followed by p.Pro1007Ala (16.9%), and p.Phe1221Serfs*20 (14.6%).

All novel variants described here are located in *NPC1* gene, and distributed as follows: one small deletion (p.Lys38_Tyr40del), one frameshift (p.Asn195Lysfs*2), and 3 missense mutations (p.Cys238Arg, p.Ser365Pro, and p.Val694Met). Sequencing profile of novel mutations is in Fig. 1. These novel changes were not found among 400 alleles from normal individuals. All 5 patients carrying novel mutations were previously evaluated by Filipin staining test and results were positive. A brief clinical description of patients

with novel sequence variations can be found in Table 2. Variations cited in this work as novel were not found in the Human Gene Mutation Database (HGMD®), ExaC, gnomAD, and 1000genomes. Findings regarding these novel changes are described below.

The p.Lys38_Tyr40del (c.114-122del9) mutation is a deletion of nine nucleotides within exon 2 (Fig. 1(a)) that leads to a protein with three missing amino acids (Lys, Arg, and Tyr). This mutation was found in a compound heterozygote carrying p.Phe1221Serfs*20 in the other allele. This male patient was diagnosed when he was 1 year old and clinical symptoms include dysphagia, cognitive decline, and developmental delay (Table 2).

Variation p.Asn195Lysfs*2 is due to a duplication of one nucleotide (adenine) in exon 5 of *NPC1* gene (Fig. 1(b)). This frameshift was identified in a female patient *in trans* with p.Phe1221Serfs*20. She was diagnosed at 6 months of age, and main symptoms were hepatosplenomegaly, hypotonia, and developmental delay. This mutation produces a truncated protein that is expected to have a defective function.

Regarding missense mutations, p.Cys238Arg is located at exon 6, due to T to C change (Fig. 1(c)), which was found in a female patient diagnosed at 2 years of age. Her clinical symptoms included cerebellar ataxia, developmental delay, and cataplexy. The p.Cys238Arg was found *in trans* with p.F1221Sfs*20 (patient was a compound heterozygote), and variants were confirmed in her parents.

Novel variation p.Ser365Pro is due to T to C change in exon 8 of *NPC1* (Fig. 1(d)) and it was found *in cis* with p.Ala183Thr. This complex allele was found in a homozygous male patient from a consanguineous marriage, and both mutations were confirmed in his parents.

The other novel missense mutation, p.Val694Met, is due to G to A change in exon 13 (Fig. 1(e)), and this alteration was detected in a female patient from consanguineous marriage. She was 14 years old at diagnosis and clinical symptoms include cerebellar ataxia, and vertical supranuclear gaze palsy. Patient was homozygous for this alteration, and variant was also confirmed in her parents.

All novel sequence variations were evaluated through alignment of amino acid sequences from 10 different organisms, and alterations are located within conserved residues, suggesting an effect on protein function or structure (Fig. 2). Pathogenicity of novel missense variations was evaluated using different web-based tools (supplementary Table S2), following the guidelines by the American College of Medical Genetics and Genomics (ACMG) to the interpretation of sequence variants [27]. Two missense mutations (p.Cys238Arg and p.Val694Met) were defined as being pathogenic by all different tools. p.Ser365Pro was also classified as pathogenic by great majority of tools except by one (CADD software). However, considering that this serine residue is conserved among species (Fig. 2), and that this substitution introduces a novel imino group in the protein, a pathogenic effect might be also expected in this case.

Finally, distribution of age at diagnosis ranged from 2 months to 46 years with an average of 11.3 years, and 32 cases (59.3%) were diagnosed in patients of up to 10 years of age. Patients included in the analysis were from different regions of Brazil, and more detailed description of symptoms was available from 33 confirmed cases. Therefore, more frequent symptoms, based on cases with complete clinical description, were splenomegaly, hepatomegaly, cerebellar ataxia, and vertical supranuclear gaze palsy, and distribution of symptoms according to age group is shown in supplementary Fig. S1.

Discussion

We identified mutations in 54 NP-C patients, being 96.3% in *NPC1* gene and 3.7% in *NPC2* gene. Distribution of mutations shown here is similar to the described in the literature from studies worldwide, where mutations in *NPC1* gene occur in 95% of the NP-C patients [2, 12, 15, 28].

The most frequent mutation in our sample population was p.Ala1035Val (27.0%) that is different from cohorts reported in North hemisphere, where p.Ile1061Thr is described as the most frequent one [29–38]. As previously reported, high frequency of p.Ile1061Thr in Hispanic patients suggest a founder effect originated in Western Europe [32, 39]. A prevalence of a different mutation in this studied cohort suggests a different

Table 2 Brief clinical description of patients that carry novel sequence variations identified by this study

Mutation	Genotype	Age at diagnosis	Gender	Clinical symptoms
p.Lys38_Tyr40del	p.Lys38_Tyr40del/p.Phe1221Serfs*20	1 year	Male	Dysphagia, cognitive decline, developmental delay
p.Ser365Pro	[p.Ser365Pro + p.Ala183Thr]/ [p.Ser365Pro + p.Ala183Thr]	27 years	Male	Neurological regression
p.Asn195Lysfs*2	p.N195Kfs*2/p.F1221Sfs*20	6 months	Female	Hepatomegaly, splenomegaly, hypotonia, developmental delay
p.Cys238Arg	p.Cys238Arg/p.Phe1221Serfs*20	2 years	Female	Cerebellar ataxia, developmental delay, cataplexy
p.Val694Met	p.Val694Met/p.Val694Met	14 years	Female	Cerebellar ataxia, vertical supranuclear gaze palsy

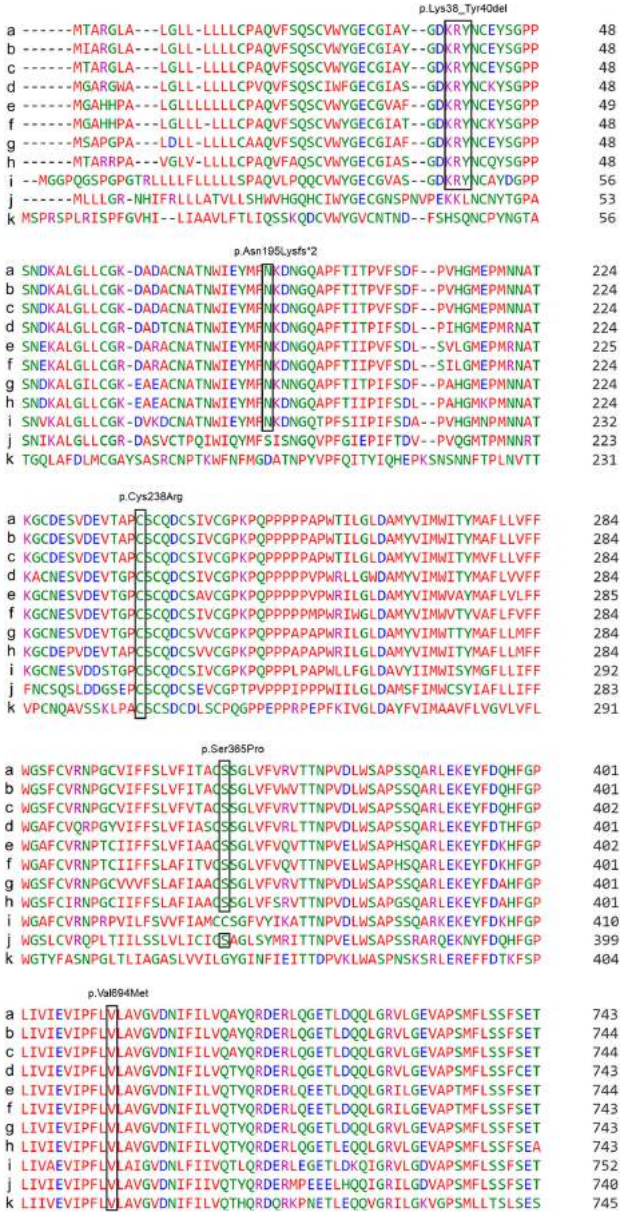


Fig. 2 Alignment of amino acid sequences of several organisms. Specific regions flanking mutations are represented. (a) Human (NP_000262.2); (b) Chimpanzee (XP_016788941.1); (c) Orangutan (XP_009250570.1); (d) Domestic guinea pig (XP_003474080.2); (e) Rat (NP_705888.2); (f) Mouse (NP_032746.2); (g) Horse (XP_005612821.1); (h) Dog (NP_001003107.1); (i) Chicken (XP_419162.3); (j) Zebrafish (NP_001230804.1); (k) Fruit fly (NP_001188769.1). Boxes indicate conserved residues associated to novel mutations. Designation of each mutation is given above the appropriate box

ethnic background of NP-C patients in Brazil. The second most common mutation in our study was p.Pro1007Ala (16.9%), and this alteration was also reported as being frequent in different European countries [5]. Frequency data of

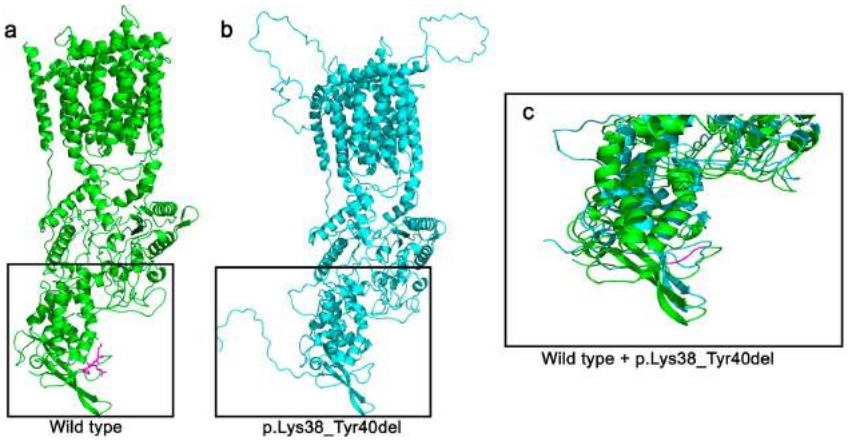


Fig. 3 Location of p.Lys38_40Tyr in the NPC1 protein. (a) Image represents the wild-type NPC1 protein; region of three amino acids (lysine, arginine, and tyrosine) involved in the deletion is represented in pink. (b) Image represents mutant NPC1 protein p.Lys38_40Tyr. (c) Close-up and superposed view of wild-type (in green) and mutant (in blue) NPC1 proteins. Figures were generated by PyMOL 2.0

this mutation reported in Portuguese, British, and German patients ranged from 15 to 20% [37].

We have observed higher frequency of some mutations in specific regions as follows: p.Ala1035Val was present in 32.0% of mutant alleles identified in São Paulo state, p.Pro1007Ala was found in 53.3% of mutant alleles from Paraná state, and p.Phe1221Serfs*20 was identified in 75.0% of alleles from Pernambuco state. Although preliminary, these data indicate that regional variation of ethnic background in a huge country as Brazil might determine higher frequencies of mutations in specific places. Several studies have reported specific disease-causing mutations among different populations and ethnic groups associated to NP-C [30, 31, 38–40]. Additional analyses are required to further investigate this issue.

Distribution of confirmed cases among geographical regions of Brazil was as follows: 46.3% of cases in Southeast, followed by 25.9% of cases from Northeast, 24.1% of cases from South, and 3.7% from West Central. This higher frequency of cases in Southeast might be a combination of highly populated region as well as a more facilitated access to health system.

Clinical presentation of NP-C can be heterogeneous and nonspecific, which makes more difficult to reach a correct diagnosis [9]. Symptoms of pediatric patients (≤4 years) described here are in agreement to a previous report that more discriminatory signs for NP-C in pediatric patients are splenomegaly, hepatomegaly, dysphagia, cognitive decline, delayed neuro-psychomotor, ataxia, and cataplexy [41]. Clinical findings reported here in adult patients, such as seizures, neurological regression, splenomegaly, cognitive decline, cerebellar ataxia, and vertical

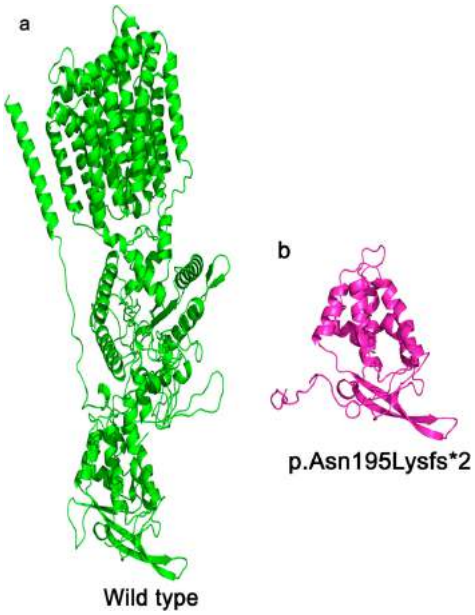


Fig. 4 Images representing wild-type NPC1 protein and NPC1 protein produced in the presence of p.Asn195Lysfs*2. (a) Wild-type NPC1 protein (1278 amino acids). (b) Mutant NPC1 protein produced by p.Asn195Lysfs*2 mutation (197 amino acids). Figures were generated by PyMOL 2.0

supranuclear gaze palsy, were described as more commonly found in older patients (supplementary Fig. S1).

Novel sequence variants appear to be widespread along different regions in the protein: p.Lys38_Tyr40del, p.Asn195Lysfs*2, and p.Cys238Arg are located within lumen A domain, p.Ser365Pro in the transmembrane II (TMII) domain, and p.Val694Met in the transmembrane V (TMV). Position of these variations can be visualized in NPC1 protein topology generated by Protter software [42] (supplementary Fig. S2).

Mutation p.Lys38_Tyr40del is an in-frame deletion that leads to a protein lacking three amino acids (Lys, Arg, and Tyr). This deletion is located within the N-terminal domain (NTD) (supplementary Fig. S2), which is the first luminal domain composed by 240 amino acids [43]. This type of mutation generates a mutant protein with a different tertiary structure (Fig. 3) that will likely affect protein function.

The frameshift variation p.Asn195Lysfs*2 produces a truncated small protein that is expected to have a defective function. The wild-type NPC1 protein has 1278 amino acids, and mutant protein produced by this variation would be expected to have 197 amino acids only (Fig. 4). Therefore, essential domains of NPC1 protein will be missing, and normal function highly impaired.

Missense mutation p.Cys238Arg is also located in a very relevant domain of the protein. The amino acid cysteine at position 238 establishes one of two disulfide bonds (C97–C238 and C227–C243) from Ψ loop. This particular loop has been reported before as being an important interface between N-terminal domain (NTD) and middle luminal domain (MLD) [44]. Therefore, the replacement of this cysteine residue by an arginine prevents the establishment of a disulfide bond, which changes protein conformation as can be observed in Fig. 5(b, c).

p.Ser365Pro variation introduces a novel imino group in the protein (Fig. 5(d, e)). This imino group will be part of a transmembrane domain (supplementary Fig. S2) and can cause a very relevant change in the protein structure by itself. In this particular case, as this mutation is found in a complex allele *in cis* with another mutation, an even stronger impact in protein function is expected due to a combination of effects.

The remaining missense mutation p.Val694Met is also located within a transmembrane region, which is important for NPC1 (Fig. 5(f) and Fig. 5(g)). This region shows high homology to the sterol-sensing domains (SSD) of HMG-Co A

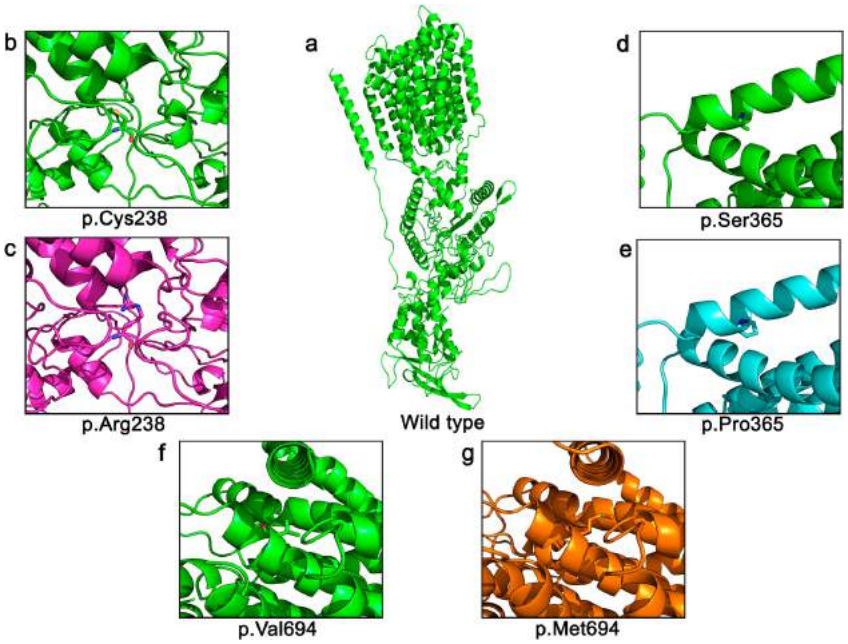


Fig. 5 Missense mutations in NPC1 when compared to the wild-type amino acido residue. Arrows indicate location of wild-type and mutant amino acid residue in each variant. (a) Wild-type NPC1 protein. (b) Wild-type residue Cys (cysteine) compared to (c) mutant residue Arg (arginine). (d) Wild-type residue Ser (serine) compared to (e) mutant residue Pro (proline). (f) Wild-type residue Val (valine) compared to (g) mutant residue Met (methionine). Figures were generated by PyMOL 2.0

reductase that is involved in cholesterol synthesis, and to the sterol regulatory element-binding protein (SREBP) cleavage-activating protein (SCAP), which is an activator of a transcription factor in cholesterol biosynthesis [11, 29, 30]. The insertion of cholesterol into the lysosomal membrane involves NPC1 transmembrane domains, including the sterol-sensing motif that has been identified in other proteins as involved in cholesterol homeostasis [45]. The majority of mutations in the SSD region is associated to a severe phenotype [30, 46].

All novel variants were located within conserved regions when multiple alignment analysis was performed with sequences of 10 different species (Fig. 2). These conserved regions imply a requirement of those amino acids for normal protein structure and/or activity throughout species. In addition, all novel variations were tested for hydrophobicity prediction and variations of hydrophobicity levels are expected in each one. This is further evidence related to protein topology susceptibility associated to novel sequences alterations reported here.

In summary, data provided here contribute to the knowledge of worldwide mutation spectrum associated to NP-C. Combination of molecular analyses with in silico tools, as well as molecular modeling, can generate a more comprehensive insight into NP-C associated proteins, with a potential to identify additional targets to the development of novel therapeutics for Niemann-Pick type C.

Acknowledgements We would like to thank the patients and their families for providing biological material for this study. The authors also thank investigators who enrolled patients in the NPC Brazil Network. This study was partially supported by grants from Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (CNPq), Fundação de Amparo a Pesquisa do Estado do Rio Grande do Sul (FAPERGS), Fundo de Incentivo a Pesquisa e Eventos do HCPA (FIPE-HCPA), and INAGEMP—National Institute of Population Medical Genetics (grant CNPq 573993/2008-4). The NPC Brazil Network is partially funded by an unrestricted grant from Actelion (05168). MPB, HB, and FN were supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES); RG and MLSP were supported by CNPq.

References

1. Vance JE, Karten B (2014) Niemann-Pick C disease and mobilization of lysosomal cholesterol by cyclodextrin. J Lipid Res 55:1609–1621. https://doi.org/10.1194/jlr.R047837
2. Vanier MT (2015) Complex lipid trafficking in Niemann-Pick disease type C. J Inherit Metab Dis 38:187–199. https://doi.org/10.1007/s10545-014-9794-4
3. Pandi S, Chandran V, Deshpande A, Kurien A (2014) Niemann-Pick disease type C or Gaucher's disease type 3? A clinical conundrum. BMJ Case Rep 2014:10–13. https://doi.org/10.1136/bcr-2014-203713
4. Vanier MT (2010) Niemann-Pick disease type C. Orphanet J Rare Dis 5:16. https://doi.org/10.1186/1750-1172-5-16
5. Vanier M, Millat G (2003) Niemann-Pick disease type C. Clin Genet 64:269–281. https://doi.org/10.1034/j.1399-0004.2003.00147.x
6. Patterson MC, Vanier MT, Suzuki K et al (2015) Part 16: Lysosomal disorders chapter 145: Niemann-Pick disease type C: a lipid trafficking disorder. In: Scriver CR, Beaudet AL, Sly WS, Valle D (eds) The metabolic and molecular bases of inherited disease. Accessed in 10/06/2015. https://doi.org/10.1036/ommbid.175
7. Walterfang M, Fietz M, Fahey M, Sullivan D, Leane P, Lubman DI, Velakoulis D (2006) The neuropsychiatry of Niemann-Pick type C disease in adulthood. J Neuropsychiatry Clin Neurosci 18:158–170. https://doi.org/10.1176/appi.neuropsych.18.2.158
8. Walterfang M, Abel LA, Desmond P, Fahey MC, Bowman EA, Velakoulis D (2013) Cerebellar volume correlates with saccadic gain and ataxia in adult Niemann-Pick type C. Mol Genet Metab 108:85–89. https://doi.org/10.1016/j.ymgme.2012.11.009
9. Patterson MC, Hendriksz CJ, Walterfang M, Sedel F, Vanier MT, Wijburg F, NP-C Guidelines Working Group (2012) Recommendations for the diagnosis and management of Niemann-Pick disease type C: an update. Mol Genet Metab 106:330–344. https://doi.org/10.1016/j.ymgme.2012.03.012
10. Morris JA, Zhang D, Coleman KG, Nagle J, Pentchev PG, Carstea ED (1999) The genomic organization and polymorphism analysis of the human Niemann-Pick C1 gene. Biochem Biophys Res Commun 261:493–498. https://doi.org/10.1006/bbrc.1999.1070
11. Davies JP, Ioannou YA (2000) Topological analysis of Niemann-Pick C1 protein reveals that the membrane orientation of the putative sterol-sensing domain is identical to those of 3-hydroxy-3-methylglutaryl-CoA reductase and sterol regulatory element binding protein cleavage-activating. J Biol Chem 275:24367–24374. https://doi.org/10.1074/jbc.M002184200
12. Carstea ED, Morris JA, Coleman KG, Loftus SK, Zhang D, Cummings C, Gu J, Rosenfeld MA et al (1997) Niemann-Pick C1 isease gene: Homology to mediators of cholesterol homeostasis. Science 277:228–231. https://doi.org/10.1126/science.277.5323.228
13. Hua X, Nohturfft A, Goldstein JL, Brown MS (1996) Sterol resistance in CHO cells traced to point mutation in SREBP cleavage-activating protein. Cell 87:415–426
14. Millard EE, Gale SE, Dudley N, Zhang J, Schaffer JE, Ory DS (2005) The sterol-sensing domain of the Niemann-Pick C1 (NPC1) protein regulates trafficking of low density lipoprotein cholesterol. J Biol Chem 280:28581–28590. https://doi.org/10.1074/jbc.M414024200
15. Naureckiene S, Sleat DE, Lackland H et al (2000) Identification of HE1 as the second gene of Niemann-Pick C disease. Science 290:2298–2301. https://doi.org/10.1126/science.290.5500.2298
16. Vanier MT, Millat G (2004) Structure and function of the NPC2 protein. Biochim Biophys Acta 1685:14–21. https://doi.org/10.1016/j.bbalip.2004.08.007
17. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249. https://doi.org/10.1038/nmeth0410-248
18. Yue P, Moult J (2006) Identification and analysis of deleterious human SNPs. J Mol Biol 356:1263–1274. https://doi.org/10.1016/j.jmb.2005.12.025
19. Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A et al (2005) Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J Med Genet 43:295–305. https://doi.org/10.1136/jmg.2005.033878
20. Schwarz JM, Cooper DN, Schuelke M, Seelow D (2014) MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 11:361–362. https://doi.org/10.1038/nmeth.2890

21. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, Bernstein JA, Bejerano G (2016) M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet 48:1581–1586. https://doi.org/10.1038/ng.3703
22. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310–315. https://doi.org/10.1038/ng.2892
23. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q et al (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet 99:877–885. https://doi.org/10.1016/j.ajhg.2016.08.016
24. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R (2013) Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 14:S3. https://doi.org/10.1186/1471-2164-14-S3-S3
25. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PEM (2008) Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 29:6–13. https://doi.org/10.1002/humu.20654
26. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815. https://doi.org/10.1006/jmbi.1993.1626
27. Richards S, Aziz N, Bale S et al (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17:405–423. https://doi.org/10.1038/gim.2015.30
28. Vanier MT, Duthel S, Rodriguez-Lafrasse C et al (1996) Genetic heterogeneity in Niemann-Pick C disease: a study using somatic cell hybridization and linkage analysis. Am J Hum Genet 58:118–125
29. Millat G, Marçais C, Rafi M a, et al (1999) Niemann-Pick C1 disease: the I1061T substitution is a frequent mutant allele in patients of Western European descent and correlates with a classic juvenile phenotype. Am J Hum Genet 65:1321–1329. https://doi.org/10.1086/302626
30. Yamamoto T, Ninomiya H, Matsumoto M, et al (2000) Genotype-phenotype relationship of Niemann-Pick disease type C: a possible correlation between clinical onset and levels of NPC1 protein in isolated skin fibroblasts. J Med Genet 37:707–712
31. Garver WS, Jelinek D, Meaney FJ, Flynn J, Pettit KM, Shepherd G, Heidenreich RA, Vockley CMW et al (2010) The national Niemann-Pick type C1 disease database: correlation of lipid profiles, mutations, and biochemical phenotypes. J Lipid Res 51:406–415. https://doi.org/10.1194/jlr.P000331
32. Park WD, O'Brien JF, Lundquist PA et al (2003) Identification of 58 novel mutations in Niemann-Pick disease type C: correlation with biochemical phenotype and importance of PTC1-like domains in NPC1. Hum Mutat 22:313–325. https://doi.org/10.1002/humu.10255
33. Bauer P, Knoblich R, Bauer C, Finckh U, Hufen A, Kropp J, Braun S, Kustermann-Kuhn B et al (2002) NPC1: Complete genomic sequence, mutation analysis, and characterization of haplotypes. Hum Mutat 19:30–38. https://doi.org/10.1002/humu.10016
34. Tarugi P, Ballarini G, Bembi B, Battisti C, Palmeri S, Panzani F, di Leo E, Martini C et al (2002) Niemann-Pick type C disease: mutations of NPC1 gene and evidence of abnormal expression of some mutant alleles in fibroblasts. J Lipid Res 43:1908–1919. https://doi.org/10.1194/jlr.M200203-JLR200
35. Ribeiro I, Marcão A, Amaral O, Sá Miranda M, Vanier MT, Millat G (2001) Niemann-Pick type C disease: NPC1 mutations associated with severe and mild cellular cholesterol trafficking alterations. Hum Genet 109:24–32
36. Meiner V, Shpitzen S, Mandel H, Klar A, Ben-Neriah Z, Zlotogora J, Sagi M, Lossos A et al (2001) Clinical-biochemical correlation in molecularly characterized patients with Niemann-Pick type C. Genet Med 3:343–348. https://doi.org/10.1097/00125817-200109000-00003
37. Yamamoto T, Nanba E, Ninomiya H, Higaki K, Taniguchi M, Zhang H, Akaboshi S, Watanabe Y et al (1999) NPC1 gene mutations in Japanese patients with Niemann-Pick disease type C. Hum Genet 105:10–16
38. Greer WL, Riddell DC, Murty S, Gillan TL, Girouard GS, Sparrow SM, Tatlidil C, Dobson MJ et al (1999) Linkage disequilibrium mapping of the Nova Scotia variant of Niemann-Pick disease. Clin Genet 55:248–255. https://doi.org/10.1034/j.1399-0004.1999.550406.x
39. Fernandez-Valero EM, Ballart a IC et al (2005) Identification of 25 new mutations in 40 unrelated Spanish Niemann-Pick type C patients: genotype-phenotype correlations. Clin Genet 68:245–254. https://doi.org/10.1111/j.1399-0004.2005.00490.x
40. Millat G, Marçais C, Tomasetto C, Chikh K, Fensom AH, Harzer K, Wenger DA, Ohno K et al (2001) Niemann-Pick C1 disease: correlations between NPC1 mutations, levels of NPC1 protein, and phenotypes emphasize the functional significance of the putative sterol-sensing domain and of the cysteine-rich luminal loop. Am J Hum Genet 68:1373–1385. https://doi.org/10.1086/320606
41. Yang C-C, Su Y-N, Chiou P-C, Fietz MJ, Yu CL, Hwu WL, Lee MJ (2005) Six novel NPC1 mutations in Chinese patients with Niemann-Pick disease type C. J Neurol Neurosurg Psychiatry 76:592–595. https://doi.org/10.1136/jnnp.2004.046045
42. Omasits U, Ahrens CH, Müller S, Wollscheid B (2014) Protter: interactive protein feature visualization and integration with experimental proteomic data. Bioinformatics 30:884–886. https://doi.org/10.1093/bioinformatics/btt607
43. Kwon HJ, Abi-Mosleh L, Wang ML, et al (2009) Structure of N-Terminal Domain of NPC1 Reveals Distinct Subdomains for Binding and Transfer of Cholesterol. Cell 137:1213–1224. https://doi.org/10.1016/j.cell.2009.03.049
44. Li X, Lu F, Trinh MN, Schmiege P, Seemann J, Wang J, Blobel G (2017) 3.3 Å structure of Niemann–Pick C1 protein reveals insights into the function of the C-terminal luminal domain in cholesterol transport. Proc Natl Acad Sci 114:201711716–201719121. https://doi.org/10.1073/pnas.1711716114
45. Vance JE (2010) Transfer of cholesterol by the NPC team. Cell Metab 12:105–106. https://doi.org/10.1016/j.cmet.2010.07.004
46. Millat G, Chikh K, Naureckiene S, et al (2001) Niemann-Pick disease type C: spectrum of HE1 mutations and genotype/phenotype correlations in the NPC2 group. Am J Hum Genet 69:1013–1021. https://doi.org/10.1086/324068

CURRICULUM VITÆ

Felipe Castro **Nepomuceno**

Bioinformatics BSc. · MSc. Student

Federal University of Rio Grande do Sul, Biotechnology Center Avenida Bento Gonçalves, 9500 Prédio 43421 Campus do Vale Agronomia 91501970 - Porto Alegre, RS - Brazil

(+55) 51-99107-4685 | fcastronepomuceno@gmail.com | www.ufrgs.br/bioinfo | felipe-castro-nepomuceno

Curriculum Vitae

August 22, 2019

September 16th, 1996

Brazilian

Education

2018-Current M. Sc. in Cellular & Molecular Biology
PPGBCM - UFRGS – Porto Alegre, BR
"Hexopyranoses Torsional Potential Adjustment using a Genetic Algorithm"
Supervisor: Dr. Hugo Verli

2014-2018 Bachelor of Biotechnology - Bioinformatics Habilitation
UFRGS – Porto Alegre, BR
"Development of a Thermoresistance Optimization Protocol for Proteins with Biotechnological and Industrial Applications"
Supervisor: Dr. Hugo Verli

Languages

- Portuguese Native speaker
- Fluent English
- Intermediate Spanish
- Basic German

Skills and Abilities

- Extensive experience in Molecular Dynamics Simulations;
- Extensive experience in Force Field Parameterization;
- Extensive experience in Molecular Modeling;
- Extensive experience in Python programming language
- Experience in Quantum mechanics calculations;
- Experience in Metadynamics calculations;
- Experience in Bash programming language;
- Experience in Structure Elucidation;
- Extensive knowledge in the GROMACS software suite;
- Extensive knowledge in Protein Structure, Dynamics and Conformation;
- Extensive knowledge in Structural Biology;
- Extensive knowledge in Carbohydrates Structures;
- Extensive knowledge in Glycoproteins and Oligosaccharides;
- Experience Supervising Undergraduate
- Verbal and Written Communication Skills;
- Work Effectively in an Independent Role and Collaboratively;
- Work as Part of a Multidisciplinary and Diverse Team.

Research Projects

1. Computational Strategies for Force Field Parametrization of Carbohydrates and Cyclic Molecules
2. Development of a Thermoresistance Protocol for Proteins with Biotehcnological and Industrial Applications
3. Protein Data Bank (PDB) Analysis Improving Carbohydrate Structural Information.

Research Articles

Márcia Polese-Bonatto; Hugo Bock; Ana Carolina S. Farias ; Rafaella Mergener; Maria Cristina Matte; Mirela S. Gil; **Felipe Nepomuceno**; Fernanda T.S. Souza; Rejane Gus; Roberto Giugliani; Maria Luiza Saraiva-Pereira: Niemann-Pick disease type C: mutation spectrum and novel sequence variations in the human NPC1 gene. Molecular Neurobiology (2019) - https://doi.org/10.1007/s12035-019-1528-z.

Felipe Nepomuceno; João Meirelles; Hugo Verli: Analysis of carbohydrates structural information on the Protein Data Bank. Journal of Chemical Information and Modeling (To be submitted)

Courses Ministrated

Nepomuceno, F.C.; Santos, L.A.; Faccioni, J.; Kampmann, P.F.: Basic Programming Python Course, 2018, at UFRGS, in Porto Alegre, RS, Brazil.

Nepomuceno, F.C.; Santos, L.A.; Faccioni, J.: Advanced Programming Python Course: Learning BioPython, 2018, at UFRGS, in Porto Alegre, RS, Brazil.

Nepomuceno, F.C.: Introduction to Python Programming in Bioinformatics, 2018, at UFRGS, in Porto Alegre, RS, Brazil.

Abstract published in proceedings of conferences

1. **Nepomuceno, F.C.**; Verli, H.: Hexopyranoses Torsional Potential Adjustment using a Genetic Algorithm. XIII School of Molecular Modeling in Biologic Systems, 2018, Petrópolis/RJ-BR.
2. **Nepomuceno, F.C.**; Verli, H.: Development of a Protocol for Thermo-resistance Optimization of Protein with Biotechnological and Industrial Applications. XLVII Annual Meeting of the Brazilian Society for Biochemistry and Molecular Biology, 2018, Joinville/SC-BR.
3. **Nepomuceno, F.C.**; Verli, H.: Development of a Thermoresistance Optimization Protocol for Proteins with Biotechnological and Industrial Applications. XIII School of Molecular Modeling in Biologic Systems, 2016, Petrópolis/RJ-BR.

Awards and Titles

1. XXVII Scientific Initiation Meeting, 2017: **Outstanding Session Presentation - Session: Bioinformatics II, UFRGS**. Development of a Thermoresistance Optimization Protocol for Proteins with Biotechnological and Industrial Applications. Advisor: Hugo Verli.