UNIVERSIDADE FEDERAL DE JUIZ DE FORA

ICE / ENGENHARIA

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

Vinicius Carius de Souza

Combinação de diferentes métodos de redução de dimensionalidade e de agrupamento para a detecção automática de conformações moleculares preferenciais

Juiz de Fora

2020

Vinicius Carius de Souza

Combinação de diferentes métodos de redução de dimensionalidade e de agrupamento para a detecção automática de conformações moleculares preferenciais

Tese ao doutorado apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Doutor em Modelagem Computacional.

Orientadora: Prof.(a) Dr.(a) Priscila Vanessa Zabala Capriles Goliatt

Coorientador: Prof. Dr. Leonardo Goliatt da Fonseca

Juiz de Fora

2020

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Souza, Vinicius Carius.
Combinação de diferentes métodos de redução de dimensionalidade e de agrupamento para a detecção automática de conformações moleculares preferenciais / Vinicius Carius de Souza. – 2020.
182 f. : il.

Orientadora: Priscila Vanessa Zabala Capriles Goliatt
Coorientador: Leonardo Goliatt da Fonseca
Tese (Doutorado) – Universidade Federal de Juiz de Fora, ICE / Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2020.

1. *Schistosoma mansoni*. 2. NTPDases. 3. Dinâmica molecular. 4. Proteínas. 5. Agrupamento. I. Goliatt, Priscila Vanessa Zabala Capriles, orient. II. da Fonseca, Leonardo Goliatt, coorient. III. Título.

Vinicius Carius de Souza

Combinação de diferentes métodos de redução de dimensionalidade e de agrupamento para a detecção automática de conformações moleculares preferenciais

Tese ao doutorado apresentada ao Programa de Pos-Graduacao em Modelagem Computacional da Universidade Federal de Juiz de Fora, como requisito parcial para obtencao do titulo de Doutor em Modelagem Computacional:

Aprovada em 16 de Abril de 2020

BANCA EXAMINADORA

Professora Dra. Priscila Vanessa Zabala Capriles Goliatt - Orientadora

Universidade Federal de Juiz de Fora

Professor Dr. Leonardo Goliatt da Fonseca - Coorientador

Universidade Federal de Juiz de Fora

Pesquisador Dr. Eduardo Krempser da Silva

Fundacao Oswaldo Cruz (Fiocruz)

Professor Dr. Laurent Emmanuel Dardenne

Laboratorio Nacional de Computacao Cientifica (LNCC)

Professora Dra. Priscila de Faria Pinto

Universidade Federal de Juiz de Fora

Professor Dr. Heder Soares Bernardino

Universidade Federal de Juiz de Fora

Dedico este trabalho aos meus pais pelo amor e confiança.
E acima de tudo pelos conselhos e incentivos.

AGRADECIMENTOS

Agradeço aos meu pais Valter e Imaculada pelo amor incondicional, oportunidades e conselhos que me guiaram até aqui;

Aos meus irmãos Vitor e Igor e demais membros da minha família pelo suporte em todos os momentos, participando e torcendo por cada passo meu;

À minha orientadora Prof^{a}. Dr^{a}. Priscila Capriles Goliatt e ao Prof. Dr. Leonardo Goliatt pela crença em meu potencial, oportunidade tão valiosa, disponibilidade e ajuda na realização deste trabalho;

Ao Prof. Dr. Marcelo de Oliveira Santos por acreditar em mim e proporcionar a oportunidade de iniciar na ciência;

Agradeço a Dr.^{a} Deborah Antunes, a Dr.^{a} Ana Carolina Ramos Guimarães e a Dr.^{a} Teca Calcagno Galvão, pesquisadoras do Laboratório de Genômica Funcional e Bioinformática - Instituto Oswaldo Cruz, pela oportunidade de colaborar no estudo sobre a enzima EthA;

Ao meu amigo Vinicius Schmitz Nunes pelo companheirismo, ajuda e dicussões enriquecedoras;

A todos meus amigos de graduação, os quais têm apoiado muito nessa jornada acadêmica;

As secretárias do Programa de Pós-graduação em Modelagem Computacional (PPGMC) por todo trabalho e ajuda com a burocracia ao longo do desenvolvimento dessa tese;

Aos demais professores e alunos do PPGMC que por tornarem a convivência mais agradável;

Ao Núcleo de Identificação e Pesquisa em Princípios Ativos Naturais da UFJF por ceder a molécula com potencial no tratamento de esquistossomose;

Meu agradecimento a UFJF, FAPEMIG, CAPES e CNPq pelo apoio financeiro durante o curso.

‘A alegria está na luta, na tentativa,
no sofrimento envolvido
e não na vitória propriamente dita.’

Mahatma Gandhi

RESUMO

A simulação de dinâmica molecular (DM) é uma técnica usada para estudar os movimentos de átomos e moléculas, permitindo a análise de conformações recorrentes e estados de transição. Porém, um grande número de conformações é necessário para estudos de predição de propriedades físico-químicas e geométricas de moléculas. Devido ao número de parâmetros considerados na descrição dos movimentos moleculares (*e. g.* distâncias intra e inter-atômicas, ângulos diedrais) os conjuntos de trajetórias apresentam uma alta-dimensionalidade, sendo este o principal fator que torna difícil a análise de longas simulações por DM.

A utilização de técnicas como o aprendizado de máquina têm sido usadas para encontrar um espaço dimensional reduzido que representa os movimentos essenciais das moléculas, permitindo identificar movimentos representativo e facilitando a análise de longas simulações. Em geral, a análise das componentes principais (PCA), um método de transformação linear, tem sido frequentemente usado para reduzir a dimensionalidade do problema em estudos de DM essencial. Contudo, a literatura propõe o uso de métodos não-lineares para a detecção do espaço de fase de moléculas proteicas.

Assim, o objetivo desta tese é desenvolver um fluxograma automatizado foi desenvolvido para a obtenção das conformações preferenciais de proteínas, trazendo para discussão os métodos de redução de dimensionalidade de dados (RDD): Autoencoder, Isomap, t-SNE, MDS e Spectral. Adicionalmente, nós propomos a combinação desses métodos com algoritmos de agrupamento para descobrir conformações representativas da trajetória de DM. E finalmente uma análise estrutural e de inibição enzimática das proteínas alvo-terapêutico no tratamento da esquistossomose. Para seleção de estruturas representativas é gerado um perfil de energia livre (FEL) usando o método *Weighted Histogram Analysis Method* (WHAM) para verificar a superfície de energia obtida por cada RDD e desta forma encontrar a conformação com maior convergência.

A flutuação atômica das proteínas foi representada pelas distâncias euclidianas entre os átomos $C\alpha$ intra-moleculares em cada conformação. A matriz de características obtida foi usada como entrada para os redutores de dimensionalidade combinadas com algoritmos de agrupamento (K-means, Ward, Meanshift e *Affinity Propagation*). O parâmetro de define o número de grupos do K-means e Ward foi predito usando os métodos BIC, elbow, GAP e máxima silhueta. E a análise de qualidade dos grupos detectados foi avaliado por métricas de validação interna de agrupamento (*e.g.*, Calinski-Harabasz (CH), *Davies-Bouldin index* (DBI) e Silhueta).

Como conjunto de testes, nós usamos como simulações as DM da miniproteína *Trp-cage* (PDB1L2Y) e da calmodulina (PDB1CLL) nas temperaturas de 310K e 510K. De acordo com os resultados, os métodos Spectral e Isomap foram capazes de gerar

espaços de dimensionalidades reduzidas que fornecem um bom discernimento sobre a separação de classes de conformações. Por serem métodos não-lineares, o espaço gerado representa melhor os movimentos proteicos que o PCA, e ,portanto, podem ser considerados alternativas promissoras para a análise de DM por dinâmica essencial.

Para a validação desses resultados, aplicamos o fluxograma em conformações da proteína HIV-1 protease obtidas por simulações de DM essencial e acelerada. Os resultados obtidos apresentaram novamente os métodos Spectral e Isomap como as melhores abordagens para a separação de classes de conformações.

Por fim, aplicamos essas técnicas em estudo de caso com proteínas avaliadas por nosso grupo de pesquisa como alvos moleculares para o tratamento da esquistossomose, as isoformas 1 (smNTPDase1) e 2 (smNTPDase2) da ATP-Difosfohidrolase de *Schistossoma mansoni*. Para as estruturas de menor energia obtidas pelo método Spectral, foram realizados estudos de *docking* molecular contra o composto LS1 sintetizado e cedidos pelo Núcleo de Identificação e Pesquisa em Princípios Ativos Naturais da UFJF, previamente estudado experimentalmente e apresentado como inibidor da smNTPDase1. Os resultados obtidos foram melhores do que os previamente publicados com o modelo de smNTDase1 e apontam que o composto LS1 possui grande potencial de inibição para ambas enzimas smNTPDases.

Palavras-chave: dinâmica molecular, agrupamento, redução de dimensionalidade, *Schistossoma mansoni*, *docking* molecular.

ABSTRACT

Molecular dynamics simulation (MD) is a technique used to study atoms and molecules' movements, allowing the analysis of recurring conformations and transition states. However, many conformations are necessary for studies of the prediction of physical-chemical and geometric properties of molecules. Due to the number of parameters considered in the description of movements molecular (*e. g.* intra and inter-atomic distances, dihedral angles), the sets of trajectories present a high-dimensionality, this being the main factor that makes the analysis of long simulations by DM difficult.

Use of techniques such as machine learning has been used to find a reduced dimensional space representing the essential movements of molecules, allowing them to identify representative movements and facilitate extended simulation analysis. In general, the principal component analysis (PCA), a linear transformation method, has often been used to reduce the problem's dimensionality in essential DM studies. However, the literature proposes the use of non-linear methods to detect the phase space of protein molecules.

Thus, the objective of this thesis is to develop an automated workflow was developed to obtain the preferential conformations of proteins, bringing to discussion the methods of reducing the dimensionality of data (RDD): Autoencoder, Isomap, t-SNE, MDS, and Spectral. Additionally, we propose to combine these methods with algorithms of grouping to discover representative conformations of the DM trajectory. And finally, structural analysis and enzymatic inhibition of target-therapeutic proteins in the treatment of schistosomiasis. To select representative structures, a free energy profile (FEL) is generated using the *Weighted Histogram Analysis Method* (WHAM) method to check the energy surface obtained by each RDD and thus find the conformation with greater convergence.

The atomic fluctuation of proteins was represented by Euclidean distances between the $C\alpha$ intra-molecular atoms in each conformation. The characteristic matrix obtained was used as an input for dimensionality reducers combined with clustering algorithms (K-means, Ward, Meanshift, and *Affinity Propagation*). The parameter defines the number of K-means groups, and Ward was predicted using the BIC, elbow, GAP, and maximum silhouette. And the quality analysis of the detected groups was evaluated by internal cluster validation metrics (*e.g.*, Calinski-Harabasz (CH), *Davies-Bouldin index* (DBI) and Silhouette).

As a set of tests, we used the DMs of the mini protein *Trp-cage* (PDB1L2Y) and calmodulin (PDB1CLL) as simulations in temperatures of 310K and 510K. According to the results, the Spectral and Isomap methods were able to generate dimensional spaces that provide a good insight into the separation of conformations classes. As they are non-linear methods, the space generated better represents protein movements than PCA

and, therefore, can be considered promising alternatives for the analysis of MD by essential dynamics.

We applied the workflow to HIV-1 protease conformations obtained by essential and accelerated MD simulations to validate these results. The results obtained again presented the Spectral and Isomap methods as the best approaches for separating classes of conformations.

Finally, we apply these techniques in a case study with proteins evaluated by our research group as molecular targets for the treatment of schistosomiasis, isoforms 1 (smNTPDase1) and 2 (smNTPDase2) from ATP-Diphosphohydrolase de *Schistossoma mansoni*. For the lower energy structures obtained by the Spectral method, molecular *docking* studies against the LS1 compound synthesized and provided by Núcleo de Identificação e Pesquisa em Princípios Ativos Naturais da UFJF, previously studied experimentally and presented as a smNTPDase1 inhibitor. The results obtained were better than those previously published with the smNTDase1 model and point out that the compound LS1 has great potential for inhibition for both smNTPDases enzymes.

Keywords: Molecular dynamics, Clustering, Dimensionality reduction, *Schistossoma mansoni*, Molecular *docking*.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ALA	Alanina
ASX	Asparagina ou Aspartato
CYS	Cisteína
ASP	Aspartato (Ácido aspartico)
GLU	Glutamato (Ácido glutâmico)
PHE	Fenilalanina
GLY	Glicina
HIS	Histidina
ILE	Isoleucina
LYS	Lisina
LEU	Leucina
MET	Metionina
ASN	Asparagina
PRO	Prolina
GLN	Glutamina
ARG	Arginina
SER	Serina
THR	Treonina
VAL	Valina
TRP	Triptofano
TYR	Tirosina
GLX	Glutamina ou Glutamato

LISTA DE SÍMBOLOS

A	Alanina
B	Asparagina ou Aspartato
C	Cisteína
D	Aspartato (Ácido aspartico)
E	Glutamato (Ácido glutâmico)
F	Fenilalanina
G	Glicina
H	Histidina
I	Isoleucina
K	Lisina
L	Leucina
M	Metionina
N	Asparagina
P	Prolina
Q	Glutamina (Glutamida)
R	Arginina
S	Serina
T	Treonina
V	Valina
W	Triptofano (Triptofana)
Y	Tirosina
Z	Glutamina ou Glutamato

SUMÁRIO

1 INTRODUÇÃO

Nesta seção serão apresentados a caracterização do problema explorado nesta tese, a motivação e seus os objetivos.

1.1 Caracterização do Problema

As proteínas são biomoléculas que desempenham funções importantes nos organismos, como por exemplo a composição de estruturas celulares, transporte de íons e moléculas, além de participarem de reações biológicas (106). Cada uma das funções que uma proteína pode exercer está intimamente relacionada com a estrutura geométrica tridimensional assumida em seu ambiente natural (106)(111). Essa estrutura, denominada nativa (ou funcional), em geral é considerada como um conjunto de estados conformacionais de baixa energia-livre e biologicamente relevantes (93)(106).

Uma vez que as macromoléculas proteícas possuem dinamismo estrutural intrínseco, estudar suas funções requer predizer e analisar a mudança de estados funcionais. As predições do comportamento dinâmico de um sistema molecular biológico pode ser realizado por meio de simulações de dinâmica molecular (DM), a qual permite obter um conjunto de trajetórias que descrevem o comportamento de um sistema de partículas a partir do seu estado inicial.

A análise dessas trajetórias fornece informações das propriedades físico-químicas e geométricas das moléculas, assim como a identificação de estados recorrentes e suas transições (35). Simulações por DM são importantes para preencher detalhes não obtidos por métodos experimentais, desempenhando um papel fundamental na análise da dinâmica de proteínas permitindo, por exemplo, a identificação de sítios (cavidades) crípticos - aqueles identificados apenas nas estruturas interagindo com substrato - e alostéricos - cavidades regulatórias. Ou ainda, auxiliar na descoberta de novas drogas baseadas em estruturas alvo, junto com metodologias de docagem molecular proteína-ligante e *virtual-screening*, fornecendo flexibilidade aos receptores alvo (30).

No contexto de desenvolvimento racional de fármacos, o uso de diferentes conformações advindas das simulações por DM representam uma alternativa para otimizar a seleção de moléculas com interação possivelmente favorável com o receptor (103)(3). Uma vez que as informações de flexibilidade deste, obtidas por subconjuntos conformacionais, aumentam a confiabilidade dos resultados e reduzem erros de predição dos locais de ligação em relação aos métodos nos quais o receptor é tratado como rígido (100). Além disso, proteínas podem assumir distintos estados em intervalos de tempo e desenhar ligantes para um estado desconhecido é uma tarefa difícil sem o uso de DM (16).

Contudo, problemas podem ser encontrados durante a análise conformacional de

proteínas, quando usadas técnicas de simulações por DM. O primeiro deles é o tempo necessário para realizar a triagem de estados conformacionais, considerados significativos, entre os vários encontrados durante uma simulação. Outro problema é que as informações relacionadas as mudanças de estados (fornecidas por várias conformações) podem não ser facilmente observadas ou tratáveis, necessitando de abordagens adequadas para serem efetivamente exploradas (10).

Conforme foi exposto por Machado 2011 (26), para um experimento de *docking* molecular, dado um banco de dados de pequenas moléculas (ligantes) como o ZINC, com mais de 20 milhões de compostos disponíveis. Se analisarmos *in silico* a interação de todas essas moléculas com uma proteína-alvo que possui pelo menos 3.000 conformações distintas, estima-se que seriam necessários pelo menos 650.000.000 horas (mais de 74 mil anos) até o término da execução do experimento (considerando que cada simulação de interação possui tempo de 1 minuto). Assim, maneiras menos custosas de incorporar a flexibilidade dos receptores nas simulações de *docking* molecular é uma das principais preocupações na área.

Estudos anteriores têm abordado possíveis estratégias para solucionar tais problemas, propondo abordagens para a seleção de estados proteícos representativos das simulações por DM. Entre as principais abordagens propostas está o uso de métodos de inteligência computacional não-supervisionados, por exemplo, algoritmos de agrupamento (*clustering*) (98). Os quais têm permitido obter resultados valiosos em relação à evolução das proteínas, contribuindo para uma melhor compreensão das transições conformacionais destas biomoléculas.

Torda e van Gunsteren em 1994 (98) e Shao *et al.* 2007 (88), foram os primeiros trabalhos que empregaram algoritmos de agrupamento em um conjunto de conformações proteícas de uma trajetória de DM. Em ambos trabalhos, a infomação usada para determinar os grupos de conformações foi o valor de pRMSD (*pairwise Root Means Square Deviation*), uma medida bem conhecida de similaridade estrutural para agrupar trajetórias. A partir desses estudos, novas abordagens vêm sendo desenvolvidas para detecção de grupos de estados proteícos em simulações por DM.

De Paris *et. al.* 2015 (21) comparou o uso do pRMSD da proteína inteira em relação diferentes propriedades do sítio de interação com substrato (volume, número de átomos pesados e a distância do pRMSD dos resíduos do sítio catalítico), como métricas de similaridade para agrupamento. De acordo com os resultados o uso do RMSD, para todos os ou parte das estruturas de dinâmica molecular, não é o medidor mais apropriado para agrupar conformações quando a proteína alvo possui um sítio catalítico com muita flutuação, pois essa métrica é fortemente influenciada por alterações que ocorrem em outras partes das estruturas. Assim, a busca por propriedades capazes de descrever a evolução estrutural de biomoléculas vêm sendo considerada um desafio para análises automatizadas

de simulações po DM.

O uso da acessibilidade relativa ao solvente (RSA - *Relative Solvent Accessibility*) é uma das propriedades sugeridas por Teletin *et. al.* (2018) (95) como forma de representação das mudanças e conformacionais em proteínas. Usando diferentes métodos de agrupamento e métricas de avaliação, observou-se que esta medida de variação interna foi capaz de fornecer informações relevantes sobre o enovelamento das proteínas testadas. Porém, os autores ressaltam a importância de se testar outras medidas para obter mais informações sobre transições conformacionais de proteínas.

Nesse sentido, nosso grupo também tem avançado neste campo, com o uso de matrizes de distância euclidiana (EDM - *Euclidean Distance Matrix*) interna entre átomos do *backbone* proteíco. Nunes 2015 (72) aplicou a combinação dos métodos GLCM (*Gray Level Cooccurrence Matrix*) e de agrupamento para obter conformações proteícas oriundas de DM, para testes de *docking* molecular. Os resultados obtidos com esta combinação foram tão bons quanto àqueles usando o programa *g_cluster*(79), quando avaliado o valor de drugabilidade das cavidades catalítica e "críptica" das conformações proteícas selecionadas como representativas.

Em outro trabalho de nosso grupo, Souza *e. al.* 2017 (92) utilizou as EDM e propriedades como SASA (*Solvent-Accessible Surface Area*) e indíce temporal como informações para obtenção de grupos coesos, formados por conformações similares. A estratégia usada foi primeiro aplicar métodos para redução de dimensionalidade das EDM, permitindo gerar um novo espaço de representação intrínseca dos estados proteícos e a seguir aplicar algoritmos de agrupamento. Conforme as análises dos grupos obtidos, essa abordagem permitiu detectar conformações distintas que representam bem a variação de estados das proteínas avaliadas, especialmente em processos de desenovelamento.

Dado os avanços em nosso grupo de pesquisa, mostrados nos parágrafos anteriores, este trabalho é mais uma contribuição para a seleção de conformações proteícas oriundas da simulações por DM. Permitindo acelerar e otimizar experimentos de seleção de fármacos direcionados a alvos-terapêuticos específico por técnicas de *docking* e *virtual screening*, a partir da aplicação de métodos de redução de dimensionalidade combinados com algoritmos de agrupamento sobre matrizes de contato.

1.2 Motivação

A variação conformacional em proteínas é influenciada por uma série de fatores referentes ao ambiente que se encontram (*e.g.*, temperatura, pressão e concentração iônica), e da interação com outras moléculas (*e.g.* substratos, inibidores e outras proteínas). Essas variações permitem que as proteínas adotem várias configurações estruturais durante sua vida útil, além de poder fazer transições rápidas entre estas. O estudo da dinâmica

na estrutura tridimensional das proteínas é essencial para compreender suas funções, o resultado de mutações nos organismos, bem como fornecer *insights* sobre mecanismos de ativação ou inibição de suas atividades.

No contexto do desenvolvimento de novos fármacos baseados em alvos, a principal motivação deste trabalho está em contribuir para a redução do conjunto de conformações proteicas consideradas nesse tipo de estudo, mantendo informações explícitas da dinâmica estrutural. Para isso, as diversas conformações moleculares obtidas foram agrupadas de forma automatizada, usando medidas de similaridade e métodos de aprendizado não-supervionado. Essa abordagem permite contornar um dos desafios em estudos de *docking* e *virtual screening* que é explorar diferentes estados de um alvo-terapêutico.

Outra motivação deste trabalho está relacionado ao estudo das enzimas Nucleosídeo Trifosfato Difosfohidrolases (NTPDases) como alvo-terapeutico no tratamento da esquistossomose. A esquistossomose é uma doença tropical causada por parasitos do gênero *Schistomosoma* associada principalmente à falta de saneamento. De acordo com a Organização Mundial da Saúde (OMS), cerca de 99 milhões de pessoas em todo o mundo receberam tratamento para doença no ano de 2017, e pelo menos 200 mil morrem ao ano devido a infecção (113). A principal forma de tratamento da doença é o uso do medicamento praziquantel, porém, há relatos na literatura de linhagens resistentes do parasito ao medicamento, o que leva à necessidade pela busca de novos alvos moleculares e compostos para o tratamento da doença.

Para atingir os objetivos propostos, primeiramente foi realizada uma análise das melhores combinações de métodos para redução de dimensionalidade das matrizes de contato e algoritmos de agrupamento. Nesta etapa foram usados dois sistemas teste e um de validação, considerando proteínas distintas e sob diferentes condições. Após isso, foram utilizados as melhores abordagens para a análise dos sistemas alvo-terapeuticos.

1.3 Objetivos

Objetivo Geral

O objetivo deste trabalho foi propor um fluxograma para o estudo de grandes conjuntos de conformações moleculares, em particular para simulações de dinâmica molecular (DM) de proteínas, trazendo para discussão diferentes métodos de redução de dimensionalidade de dados (RDD). Adicionalmente, nós propomos a combinação desses redutores com algoritmos de agrupamento para descobrir conformações representativas da trajetória de DM, contribuindo assim com a redução do tempo necessário para a análise deste tipo de experimento.

Objetivos Específicos

- Avaliar métodos de redução de dimensionalidade não-lineares para análise do espaço intrínseco do movimento de diferentes proteínas.
- Criar mapas de energia livre usando o método WHAM (*Weight Histogram Analysis Method*).
- Agrupar conformações proteicas obtidas por simulações de DM usando como coordenadas internas as matrizes de distância euclidiana (EDM) gerada a partir das distâncias entre átomos $C\alpha$ de cada estado.
- Contribuir para a redução do tempo de execução das análises de simulações de dinâmica molecular, permitindo uma seleção mais rápida e eficiente de conformações de proteínas para estudos de docagem molecular.
- Desenvolver um fluxograma para análise automatizada de conjuntos conformacionais de proteínas.

2 TRABALHOS RELACIONADOS

O presente estudo apresentou a relevância científica no contexto de explorar e trazer para discussão diferentes abordagens para análise de simulações de dinâmica molecular (DM), usando técnicas não-supervisionadas de aprendizado de máquina. Conforme a revisão bibliográfica realizada revelou, métodos de inteligência artificial que não necessitam de conhecimento prévio dos dados ainda são pouco explorados para análise de transições de proteínas. Além disso, diferente do proposto aqui, poucos estudos têm realizado uma análise sobre métodos não-lineares de redução de dimensionalidade para detecção de mapas de energia referentes aos estados assumidos por biomoléculas ao longo de trajetórias de DM.

O trabalho publicado por Gordon e Somorjai (1992) (41) foi um dos primeiros estudos a utilizar métodos de aprendizado de máquina não supervisionado para análise de conformações de proteínas oriundas de simulações de DM. Neste trabalho foi usada uma abordagem *fuzzy* (lógica difusa) para agrupar diferentes fragmentos do hormônio paratireóide (PTH) simulados por DM. Para cada conjunto teste usado no trabalho, foram obtidas 1.000 conformações e a matriz de RMSD foi usada como informação de similaridade para a análise de agrupamento pelo método *fuzzy*. Os resultados obtidos pelos autores mostram que essa abordagem é promissora já que não necessita de um *cutoff* para determinar em qual grupo uma conformação está, isto é feito aplicando o princípio da incerteza inerente à técnica de lógica difusa.

Torda e van Gunsteren (1994) (98) aplicaram dois diferentes algoritmos, o *Single Linkage* e o *Hierarquical divisive*, em dois conjuntos de dados extraídos da simulação da proteína serino protease: um referente ao *backbone* de 12 resíduos aminoácidos de regiões bem conhecidas por representar as conformações e outro utilizando o *backbone* de todos os resíduos da proteína. No trabalho, foram usadas 2.000 conformações oriundas de DM e assim como em (41) o RMSD foi usado como métrica para determinar a similaridade entre os dados. Como resultado, os autores apontaram que o algoritmo *Hierarquical divisive* pareceu obter melhores resultados em relação ao *single linkage*, uma vez que este é baseado em uma distância mínima entre os pontos. Além disso, observaram que o conjunto de dados referente apenas aos 12 resíduos de aminoácidos parece ser mais informativo sobre as mudanças individuais de cada conjunto conformacional.

Com relação a caracterização de diferentes algoritmos de agrupamento, dois trabalhos têm grande similaridade ao proposto nesta tese. O primeiro deles é a publicação de Shao *et. al.* (2007) (88), no qual 11 algoritmos foram utilizados para o agrupamento de dados de diferentes simulações de DM usando a função de similaridade RMSD entre as conformações obtidas. Os autores realizaram um ensaio para avaliar se o número de grupos era um parâmetro dependente do tamanho da simulação, e concluiram que em

simulações curtas de 500 ps (100 conformações) o número de grupos ideal foi de 5, enquanto que em simulações de 36 ns (3.644 conformações) um número variável de grupos pode ser encontrado e a performance dos algoritmos é altamente dependente da escolha deste parâmetro. Além disso, concluiram que os algoritmos K-means, *linkage*, *average-linkage* e *Self-Organizing Maps* possuem melhores resultados, especialmente quando avaliadas as métricas DBI e pSF, e assim como verificado por (98) o *single linkage* é sensível a presença de *outliers* devido sua abordagem por distãncia mínima.

O trabalho de Machado (2011) (26) também está intimamente relacionado ao proposto na presente tese. Nele, a autora aplica os mesmos algoritmos usados por (88) porém a proposta do trabalho foi a seleção de conformações para ensaios de *Docking*. No estudo foram usadas 3.100 conformações obtidas da trajetória de DM da enzima 2-trans-enoil ACP(CoA) Redutase de *Mycobacterium tuberculosi*. Diferente de Shao *et al.* (2007), Machado (2011) utilizou diferentes medidas de similaridade para agrupamento, relacionadas ao total de contatos da conformação. Contudo, assim como em trabalhos anteriores, os algoritmos K-means e *average linkage* novamente apresentaram melhores resultados.

Outro trabalho com abordagem similar ao desenvolido aqui é o artigo de Teletin *et. al.* (2018)(95). As autoras realizaram experimentos usando diferentes métodos de agrupamento, K-means e *hierarchical agglomerative clustering*(HAC), e de avaliação, silhueta e *V-measue*. Foram usados no trabalho um conjunto de 10.000 conformações proteicas, divididas em quatro classes, representadas pelos valores de RSA (*relative solvent accessible*). Os resultados obtidos indicam, assim como na presente tese, que o método *HAC* detecta grupos melhores do que o algoritmo K-means com bons valores de *V-measure* e silhueta. No entanto, a abordagem usada diferencia-se do presente estudo, pelo fato que o RSA informa a exposição ao solvente dos resíduos, enquanto o SASA (usado em (92)) e EDM fornecem informação tanto da exposição de toda a proteína e suas mudanças conformacionais.

Albert *et. al.* (2018) (2), conduziu um estudo a cerca da influencia de representações vetoriais das proteínas e o impacto sobre a detecção de suas estruturas usando métodos não-supervisionados como o algoritmo *self-organazing map*. De acordo com os autores, a combinação de valores de RSA e letras de um alfabeto estrutural, o qual representa as conformações proteicas, foi capaz de proporcionar informações significativas sobre a transição de estados de proteínas. Isto levanta a possibilidade de novos estudos, acerca de novas representações e métodos de agrupamento de moléculas biológicas, em nosso grupo de estudo.

Além da detecção de estruturas significativas, outra abordagem discutida neste trabalho foi o uso de métodos de redução de dimensionalidade para analisar a superfície de energia das proteínas alvo aqui estudadas. Em estudos anteriores como Das *et. al.* (2006)

(17), Brown *et. al.* (2008) (9) e Stamati *et. al.* (2010) (94), os autores propõem o uso de redutores não-lineares de redução de dimensionalidade para explorar melhor o espaço de fase de proteínas. Essa premissa parte do pressuposto que o PCA (*Principal Component Analysis*), método amplamente usado para dinâmica essencial, tende a falhar na detecção do espaço intrínseco de movimento de moléculas biológicas, pois os mesmo tentem a ser altamente não-lineares(94).

Em Brown *et. al.* (2008) (9), os autores aplicaram diferentes métodos de redução (PCA, isomap, LLE (*local linear embedding*) e autoencoder) sobre os dados de coordenadas cartesianas e angulos diedrai de um conjunto conformacional da molécula trans-1,2,4-trifluorociclo-octano, oriundas de simulações de dinâmica molecular. Com base nos resultados, os métodos LLE e Isomap apresentaram foram similares entre. E quando comparados ao PCA, foi observado que na maioria dos testes os abordagens não-lineares foram melhores.

Em contrapartida, Duan *et. a.* (2013) (28) observou que o método PCA não foi tão diferente em relação métodos não-lineares. Os autores aplicaram os redudores isomap, *diffusion maps*, LLE e PCA sobre o conjunto de trajetórias da proteína $\beta-hairpin$ e observaram que, embora métodos não-lineares sejam capazes de detectar com sucesso bacias de energia nativas da proteína e separar bem grupos, a relação de distância entre conformações não é mantida em baixa dimensionalidade(28).

Em um trabalho aplicando uma grande variedade de métodos de redução de dimensionalidade sobre trajetórias de simulações MD, Tribello e Gasparotto (2019) (101) apontam que, em geral, tanto métodos lineares quanto não-lineares são capazes de detectar bem conformações e descrever as superficíes energéticas. Quando um desses algoritmos de redução de dimensionalidade obtem melhor resultado em relação a outro, geralmente é devido a características dos dados que apenas um dos algoritmos pode reconhecer. Por exemplo, o isomap superará o PCA quando se trata de projetar dados que estão em um *manifold* não-linear. Desta forma, os autores concluem que o ideal é analisar o conjunto de trajetórias com diferentes abordagens e então definir qual a melhor escolha, baseando-se em usando critérios.

Apesar dos trabalhos citados anteriormente estarem relacionados ao proposto nesta tese, os mesmos, o diferencial do estudo realizado nesta tese é o uso de diferentes combinações de métodos de redução de dimensionalidade e detecção de grupos a partir das EDM entre os $C\alpha$ de cada conformação obtida por DM. Outro diferencial desta tese relaciona-se às proteínas alvo-terapeutico analisadas. Trabalhos previamente publicados, nos quais foram realizados experimentos *in vitro* envolvendo a inibição e análise de atividade catalítica, têm apontado a importância das enzimas NTPDases de *S. mansoni* no tratamento da esquistossomose (104)(86)(8)(19)(20). Porém, poucos estudos têm sido desenvolvidos com o intuito de fornecer informações à nível molecular sobre as mudanças

estruturais e interações com possíveis inibidores para as smNTPDases.

Nunes (2015) (72), em sua tese de doutorado, realizou um estudo de comparativo entre as isoformas 1 de *Homo sapiens*(E-NTPDase1 ou CD39) e *S. mansoni* (smNTPDase 1). Em seu trabalho, Nunes empregou o método GLCM (*Grey Level Co-occurrence Matrices*) sobre as matrizes de distância euclidiana dos átomos Cα de cada conformação para caracteristicas a serem usadas pelo método de agrupamento K-means. Sobre as conformações medóides obtidas, foram realizados ensaios de *docking* molecular.

Contudo, embora esta tese esteja intrinsicamente relacionada ao trabalho de Nunes (2015) (72), as abordagens para detecção de grupos foram diferentes. O presente estudo também focou em simulações convencinais longas (250 nanossegundos) e aceleradas (50 nanossegundos) para avaliar os estados conformacionais de cada isoforma de NTPDase de *S. mansoni*, considerando a ausência e presença de substrato no sítio catalíco. Além disso, foram avaliados juntamente com as mudanças estruturais e análises de *docking*, os mapas de energia livre para cada conjunto de trajetórias.

3 REFERENCIAL TEÓRICO

3.1 Nucleosídeo Trifosfato Difosfohidrolases

As Nucleosídeo Trifosfato Difosfohidrolases (NTPDases; EC 3.6.1.5) são enzimas que catalisam a hidrólise de molecúlas de nucleotídeos di- e trifosfatados, como por exemplo o ATP (Adenosina Trifosfato) e o ADP (Adenosina Difosfato), para suas formas monofosfatadas, usando como cofatores íons bivalente (especialemente cálcio e magnésio). Essas moléculas participam do processo de sinalização celular, compondo o sistema purinérgico que atua na comunicação célula-célula, diferenciação celular, coagulação sanguínea, etc (53). Uma característica comum das enzimas pertencentes a esta família é a presença de cinco regiões conservadas denominadas de *apyrase conserverd regions*(ACR), as quais estão envolvidas na ligação e hidrólise dos substratos di- e trifosfatados (80)(85).

Embora o presente estudo esteja fundamentalmente focado nas NTPDases de *S. mansoni*, essas enzimas são amplamente distribuídas em diferentes organismos vivos: bactérias, protistas, fungos, plantas e animais (71). A presença destas proteínas em parasitos parece estar relacionado com mecanismos de invasão ao organismo hopedeiro ou ainda como forma de "driblar" as respostas imunológicas deste (104)(86). Quanto a localização celular, as NTPDases podem ocorrer no meio intracelular, ancoradas na membrana plasmática na superfície das células ou ainda serem secretadas (85)(53).

Com relação ao *S. mansoni*, a primeira evidência da presença destas enzimas foi relatado por Vasconcelos *et. al. 1993*(105) em frações isoladas do tegumento do verme, sendo posteriormente identificadas duas isoformas com massa molecular de 65KDa e 55KDa denominadas de smNTPDase1 e smNTPDase2, respectivamente. Estas isoformas são expressas e ativas em todos os estágios do ciclo de vida do verme, indicando assim uma possível importância fisiológica para o parasito(32)(59). Além disso, a smNTPDase1 compõem o segundo grupo mais abundante de proteínas no tegumento de espécimes na fase adulta, reforçando essa teoria (13)

Em estudo utilizando siRNA (*small interfering Ribonucleic acid*), (19) demonstrou que em vermes adultos de *S. mansoni* que tiveram a expressão de smNTPDase suprimida, apresentaram redução da atividade de clivar ATP exógeno e liberar fosfato inorgânico (Pi). De acordo com esses resultados, uma vez que a smNTPDase 1 é capaz de hidrolisar ATP e ADP exógeno, a mesma poderia estar relacionada com a capacidade do parasito modular eventos tromborregulatórios e ativação do sitema imune. Já que tais nucleotídeos são sinalizadores para agregação plaquetária e molécular pro-inflamatórias (19)(8)(20).

Em outro estudo, também utilizando RNAi (*Ribonucleic acid interference*), (20) demonstrou que em vermes adultos nos quais a expressão de smNTPDase2 foi suprimida, mantendo-se a expressão de smNTPDase1, os níveis de hidrólise de ATP e ADP exógenos

não foram drásticamente reduzidos. Isto poderia indicar que a principal enzima que catalisa o metabolismo dessas moléculas no tegumento é a smNTPdase 1. Porém, devido a alta expressão de smNTPDase 2 nas fases de mirácidio e cercária, formas infectantes em caramujos e humanos, respectivamente, sugere-se que tal enzima tenha papel fundamental na invasão realizada parasito em ambos hospedeiros (37).

Com relação a estrutura destas enzimas, (24) realizou a caracterização do gene responsável pela síntese da SmATPDase1 e levantou a hipotese de que tal isoforma estaria localizada na superfície externa do tegumento e sua estrutura seria semelhante a isoforma 1 (CD39) em humanos, com um domínio ECD e dois domínios transmembranares (TM1 e TM2). Com base nessas informações, nosso grupo de pesquisa propôs o primeiro modelo da estrutura tridimensional da smNTPDase 1 e CD39 com o uso da técnica de modelagem por homologia (71). Além disso, foi realizada uma análise estrutural de ambas isoformas usando simulações por dinâmica molecular (72).

Quanto a smNTPDase 2, (59) publicou a caracterização do gene que codifica esta isoforma, propondo que a mesma, diferente da smNTPDase 1, possuiria apenas uma hélice transmembranar que seria clivada após a síntese, sendo assim sintetizada e secretada pelo tegumento de vermes adultos. Foi sugerido também que essa isoforma apresentaria homologia com as isoformas NTPDase 5 e 6 de humanos (59). Em 2014, nosso grupo propôs os modelos referentes a smNTPdase e a isoforma NTPDase 6 de humanos, usando modelagem por homologia (23).

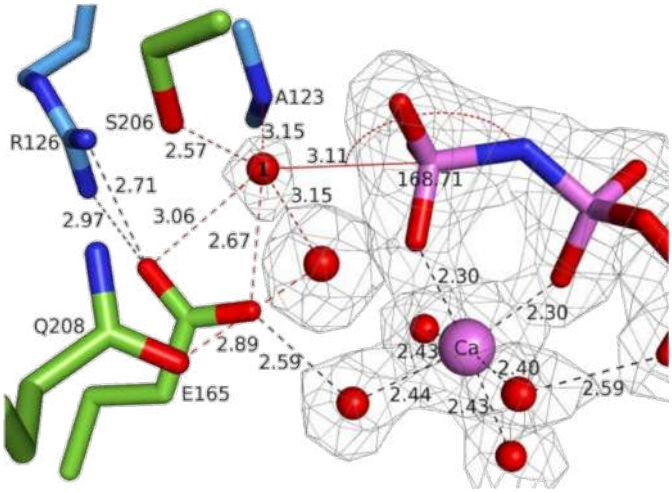
O estudo dos modelos obtidos para smNTPDase 1 e smNTPDase 2, bem como das isoformas humanas, tem permitido ao nosso grupo avanços referentes a possíveis formas de inibição destas enzimas e desenvolvimento de terapias alternativas à esquistossomose. Recentemente, foi relatado a atividade antiesquistossomáticas de derivados de chalconas *in vitro* e através de métodos de simulação da interação ligante-prteína (*docking*) sugeriu-se que tais moléculas são possíveis inibidoras da atividade de smNTPDase 1 (76).

3.1.1 Mecanismos Catalíticos das NTPDases

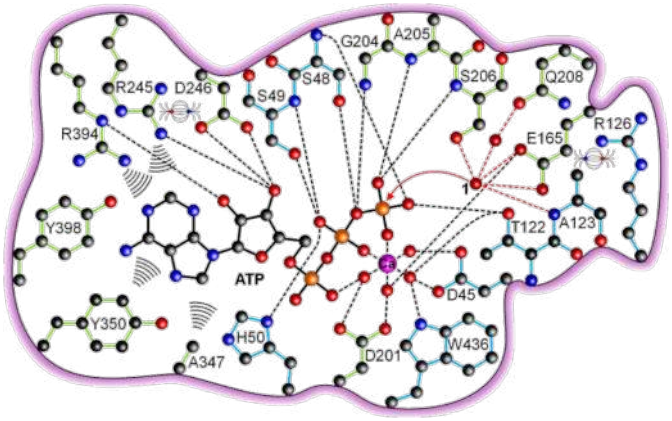
Considerando o fato de que o presente estudo irá abordar o potencial uso terapeutico das smNTPDases, vale expor os principais modelos descritos na literatura sobre o mecanismo catalítico desta família de enzimas. Em 2008, foram descritos e publicados dois modelos distintos sobre o mecanismo de hidróise realizado por enzimas NTPDases, sendo eles: (i) o modelo de Zebish e Sträter (2008) (117), o qual foi baseado na estrutura da isoforma NTPDase2 (PDB3CJA) do organismo *Rattus novergicus*(rato wistar ou de laboratório) e (ii) o modelo proposto por Kozakiewicz *et. al.*(2008), sendo baseado no modelo da apirase de batata (*Solanum tuberosum*) (56).

De acordo com Zebisch e Sträter (2008), a hidrólise do grupamento fosfato terminal da molécula de ATP (γ-fosfato) ocorre com o ataque nucleofílico de uma molécula de

água que é orientada por interações com resíduos A123, E165 e S206, presentes no sítio catalítico a enzima (figura 1). Segundo os autores, a água nucleofílica é ativada pelo resíduo E165 e estabilizada graças as interações com outra molécula de água na vizinhança e pelo resíduo Q208. Além disso, a presença do cátion bivalente funcionaria como um catalizador polarizando uma das ligações P-O do grupo fosfato terminal. Durante o ataque nucleofílico, as cargas negativas dos estados de transição seriam estabilizadas pelo íon metálico. E a posição do mesmo seria coordenada pela presença de quatro moléculas de água (117).



(a)



(b)

Figura 1 – Representação do mecanismo catalítico das NTPDases segundo Zebisch e Sträter (2008). Na figura (a) é apresentado a região do sítio ativo da NTPDase2 de *Rattus norvegicus*. Destaque para a água nucleofílica (água 1 - em vermelho) posicionada a uma distância de 3,11 Å do grupo fosfato terminal do ligante ATP (em rosa), os resíduos (em azul e verde) que direcionam a posição da água nucleofílica e íon cálcio (em rosa) coordenado por quatro moléculas de água. Em (b) esta representado o modelo esquemático do sitio ativo, destacando o ataque do nucleófilo ao γ-fosfato da molécula de ATP. As ligações de hidrogênio são mostradas como linhas tracejadas, as interações hidrofóbicas são as linhas de onda e as pontes salinas estão mostradas como linhas de campo.(Extraído de (117))

Quanto ao modelo de Kozakiewicz (figura 2), mecanismo catalítico envolvendo dois conjuntos de resíduos, onde o primeiro grupo (S54, T127 e E170) seria responsável pela hidrólise do $\gamma-$fosfato, enquanto que o segundo grupo (T55 e E78) atuaria na hidrólise do β-fosfato. O resíduo S54 seria responsável pelo posicionamento do ATP no sítio ativo da enzima ao passo que o resíduo T127 seria o responsável pela ativação de uma molécula de água responsável pelo ataque nucleofílico ao grupo γ-fosfato. Tal molécula de água seria estabilizada graças a interação com o resíduo E170 (56).

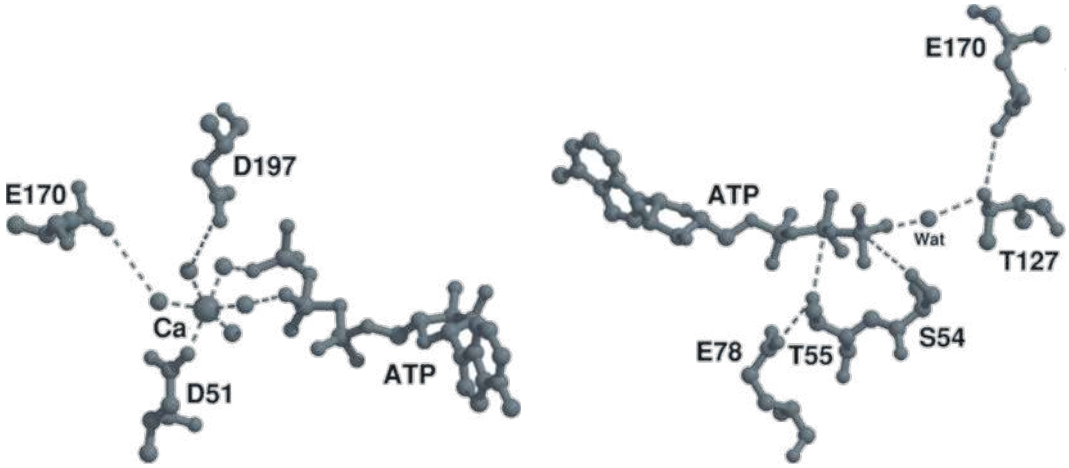


Figura 2 – Representação do mecanismo catalítico das NTPDases segundo Kozakiewicz *et. al.* (2008). Em (a) estão representados os resíduos catalíticos da apirase de batata. O sítio apresenta os principais resíduos envolvidos na interação da enzima com o substrato (ATP) e com a molécula de água nucleofílica (Wat). Já em (b) é apresentado o posicionamento do íon metal e moléculas de água. Destaque para o íon cálcio, e as moléculas de água e resíduos que atuam no posicionamento do íon no sítio ativo. (Extraído de (56))

Ainda com relação ao segundo modelo de catálise, a hidrólise dos grupamento β-fosfato ocorreria através de um ataque nucleofílico realizado pelo resíduo T55. Similar ao proposto por Zebisch e Sträter (2008), o íon metálico também não interagiria diretamente com os resíduos da enzima, mas sim por meio de moléculas de água. Estas, juntamente com os resíduos D50, D197 e E170, seriam responsáveis pelo posicionamento do íon metálico no sítio ativo da enzima (56).

As informações fornecidas por ambos modelos a cerca dos resíduos catalítico e cofatores, tem permitido o avanço no desenvolvimento de possíveis drogas para o tratamento de doenças relacionadas a estas enzimas. Tais drogas atuariam como inibidores, interagindo no sítio catalítico especialmente com os resíduos sitados nos modelos de catálise.

3.2 Dinâmica Molecular

A dinâmica molecular (DM) é uma técnica que permite gerar dados sobre movimentos de um sistema de partículas em função do tempo, constituindo trajetórias, as

quais são dependentes do potencial de interação entre as partículas e da resolução das equações de movimento da mecânica clássica (50)(39). Desde a sua origem, as técnicas de DM sofreram grande avanço permitindo a simulação de sistemas biológicos relevantes com centenas de átomos, incluindo proteínas inteiras inseridas em membranas, ou ainda sistemas complexos como nucleossomos e ribossomos (47).

As forças que atuam em cada átomo do sistema pode ser obtida pela derivação de um conjunto de equações, onde a energia potencial pode ser deduzida da estrutura molecular. Este conjunto de equações juntamente com parâmetros atômicos obtidos por experimentos físicos ou químicos, constituem os campos de força. A forma funcional básica de um campo de força, apresentada na equação 3.1, inclui termos de átomos ligados covalentemente, e de átomos não-ligados que descrevem as interações eletrostáticas de longo alcance e forças de van der Waals.

$$\begin{aligned} E_{\text{total}} &= E_{\text{ligados}} + E_{\text{não-ligados}} \\ E_{\text{ligados}} &= E_{\text{ligação}} + E_{\text{ângulo}} + E_{\text{diedral próprio}} + E_{\text{diedral impróprio}} \\ E_{\text{não-ligados}} &= E_{\text{electrostático}} + E_{\text{van de Waals}} \end{aligned} \tag{3.1}$$

O termo que descreve o potencial de interação entre átomos ligados pode ser dividido em quatro subtermos: $E_{\text{ligação}}$ é o potencial harmônico linear para ligações químicas e seus movimentos lineares, $E_{\text{ângulo}}$ descreve o potencial harmônico angular para os ângulos formados por três átomos ligados consecutivamente, $E_{\text{diedral próprio}}$ e $E_{\text{diedral impróprio}}$ representam os potenciais diedrais próprio e impróprio, respectivamente. O potencial diedral próprio descreve as rotações das ligações químicas formadas, enquanto que o diedral impróprio representa os ângulos entre os planos formados por quatro átomos. Quanto o termo dos átomos não ligados pode ser dividido em: $E_{\text{electrostático}}$ representa as interações eletrostáticas descritas pelo potencial de coulomb e $E_{\text{van de Waals}}$ é que representa as interações de van de Waals descritas pelo potencial de Lennard-Jones.

A forma explícita das equações de uma campo de força pode ser escrita como (39):

$$\begin{aligned} E_{\text{total}} = &\underbrace{\frac{1}{2}\sum_{n=1}^{N_b} K_{bn}\left(b_n - b_{0n}\right)^2}_{I} + \underbrace{\frac{1}{2}\sum_{n=1}^{N_\theta} K_\theta\left(\theta_n - \theta_{0n}\right)^2}_{II} + \underbrace{\frac{1}{2}\sum_{n=1}^{N_\xi} K_\xi\left(\xi_n - \xi_{0n}\right)^2}_{III} + \\ &\underbrace{\sum_{n=1}^{N_\phi} K_{\phi n}\left[1 + \cos\left(n_n\phi_n - \delta_n\right)\right]}_{IV} + \underbrace{\sum_{n\leq k}^{N_{atoms}} \left(\frac{A_{ij}}{r_{ij}}\right)^6 + \left(\frac{B_{ij}}{r_{ij}}\right)^{12}}_{V} + \underbrace{\sum_{n\leq k}^{N_{atoms}} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}}}_{VI} \end{aligned} \tag{3.2}$$

onde, o termo I representa o potencial harmônico para as N_b ligações químicas e seus movimentos vibracionais, sendo K_{bn} a constante de Hooke, b e b_0 os comprimentos instantâneo e de equilíbrio, respectivamente. O termo II descreve o potencial angular para os

N_θ ângulos formados por três átomos ligados consecutivamente na molécula, sendo K_θ a constante de Hooke, θ o ângulo entre as ligações e θ_0 o ângulo de equilíbrio. III denota o potencial diedral impróprio entre os N_ξ ângulos formados pelos planos das ligações que envolvem quatro átomos, os termos K_ξ, ξ e ξ_0 representam respectivamente a constante de Hooke, o ângulo entre os planos e o ângulo de equilíbrio.

O quarto termo da equação 3.2 refere-se ao diedral próprio entre as N_ϕ rotações em torno das ligações químicas, sendo K_ϕ a constante que define a barreira de rotação das ligações químicas formadas, n é o número de mínimos assumido pela função, ϕ é variação do ângulo e δ é o ângulo de diferença de fase (0° ou 180°). Os termos seguintes descrevem a interação entre átomos não ligados, sendo que IV representa o potencial de Lennard-Jones, o qual possui um termo atrativo entre os átomos (interação de van de Waals) que é definido por $\left(\frac{A_{ij}}{r_{ij}}\right)^6$, e um termo repulsivo descrito por $\left(\frac{B_{ij}}{r_{ij}}\right)^{12}$, onde temos que A_{ij} e B_{ij} são as distâncias finitas mínimas nas quais o potencial inter-atômico é iguail a zero, ambos termos são dependentes dos tipos de átomos i e j, e r_{ij} é a distância entre esses átomos. Já o termo VI representa as interações de Coulomb, sendo q_i e q_j as cargas dos átomos i e j, respectivamente, ϵ_0 a permissividade elétrica do vácuo, ϵ_r a constante dielétrica do meio e r_{ij} a distância entre os átomos interagentes.

O principal objetivo de um campo de força, aplicado a simulações de DM, é descrever em termos clássicos as interações intra e interatômicas. Além da forma funcional dos potenciais, os campos de força incluem um conjunto de parâmetros para diferentes tipos de átomos, ligações químicas e ângulos. Em um campo de força, por exemplo, o oxigênio presente em um grupo carbonila possui parametrização distinta daquele que está em um grupo hidroxila. Apesar de simulações de DM envolver moléculas biológicas como proteínas, DNA e RNA, estes parâmetros são em geral obtidos por métodos empíricos com pequenas moléculas orgânicas que são facilmente tratáveis em estudos experimentais e cálculos quânticos (106).

Uma vez definida a forma de interação dos átomos de um sistema, é necessário determinar as condições iniciais para posição e velocidade das partículas. Para estruturas proteicas, a posição inicial dos átomos é dada pelas coordenadas espaciais obtidas por métodos de predição tridimensional experimental ou *in silico*. Já as velocidades podem ser obtidas a partir da distribuição de Boltzmann dependente da temperatura do sistema (65). Assim, podemos obter a aceleração dos átomos resolvendo a equação clássica do movimento, descrita como:

$$m_\mathrm{i}\frac{d^2\mathbf{s}_\mathrm{i}}{dt^2} = \mathbf{f}_\mathrm{i} = -\frac{\partial}{\partial \mathbf{s}_\mathrm{i}} U(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_\mathrm{N}) \tag{3.3}$$

onde $U(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_\mathrm{N})$ é a energia potencial dependente das coordenadas das N partículas do sistema. Este sistema é composto por N equações não-lineares de segunda ordem acopladas e não possui solução exata, assim, a equação 3.3 deve ser resolvida numericamente passo a

passo por algoritmos de integração (*e.g.* Verlet, Meio-Passo ou Leap-Frog, Velocity-Verlet e Beeman).

3.2.1 *Ensembles* termodinâmicos

As equações de integração mantêm um número constante de partículas (N), o volume (V) da caixa de simulação e a energia (E) total do sistema. Isto implica que as trajetórias obtidas foram simuladas em um *ensemble* micro-canônico ou NVE. No entanto, apesar da energia total ser constante as energias cinética e potencial podem variar, indicando que o sistema não está em equilíbrio e o mesmo irá sofrer flutuações na temperatura. Assim, é desejável o controle da temperatura, mantendo-a constante. Além disso, em determinados casos pode ser interessante manter a pressão do sistema constante. Portanto, diferentes *ensembles* termostáticos têm sido propostos:

- Simulações com temperatura constante: também conhecido por NVT, o número de partículas, volume e temperatura são mantidos constantes. Conforme abordado anteriormente, a temperatura de um sistema está diretamente relacionada com a energia cinética das partículas. Para manter a temperatura em uma faixa considerada constante, devemos redimensionar a velocidade dos átomos usando um fator. Para isso, diferentes abordagens podem ser aplicadas, sendo os métodos mais comuns: o *Nosé-Hoover thermostat*, *Nosé-Hoover chains*, *Berendsen thermostat*, *Andersen thermostat* e *Langevin dynamics*.

- Simulações com pressão e temperatura constantes: neste tipo de simulação o número de partículas (N), pressão (P) e temperatura (T) são mantidos constantes. As abordagens usadas para controle da pressão são similares aquelas usadas para controle da temperatura. Neste caso, a pressão é controlada redimensionando o volume da célula de simulação nas três direções espaciais. $\sqrt{T_{\mathrm{B}}/T_{\mathrm{t}}}$, onde T_{B} é a temperatura desejada para o sistema e T_{t} o temperatura instantânea antes de redimensionar a velocidade.

3.2.2 Etapas da Dinâmica Molecular

A simulação de DM envolve basicamente quatro etapas:

1. Minimização: a função de energia potencial de uma biomolécula é muito complexa apresentando diversos mínimos e máximos locais. Apesar de existir um mínimo global de energia, atingí-lo pode ser algo difícil. Assim a minimização tem como objetivo encontrar um mínimo local em que a molécula pode ser biologicamente encontrada. Além disso, a etapa de minimização de um sistema é altamente recomendada, pois permite resolver problemas de sobreposição de átomos, interações instáveis ou

desfavoráveis e outras distorções na estrutura que podem levar a resultados não confiáveis.

2. Banho térmico: é a etapa de tratamento térmico do sistema. Este passo é necessário para, a partir da temperatura de resolução experimental da biomolécula, atingir a temperatura desejada para simulação. Durante a fase de aquecimento, as velocidades iniciais dos átomos são atribuídas em baixa temperatura e a cada passo de tempo as velocidades são recalculadas para temperaturas maiores. Este passo é repetido até atingir a temperatura necessária.

3. Equilibração: as etapas anteriores não consideram o controle e estabilidade de parâmetros como pressão, volume e energia. O propósito de realizar a fase de equilibração é estabilizar propriedades do sistema em valores desejáveis, com o intuito de reduzir variações bruscas responsáveis por instabilidade da biomolécula e por resultados não confiáveis.

4. Produção: é a fase final da simulação de dinâmica molecular, onde o sistema é simulado por um período de tempo necessário para observar as mudanças estruturais de relevância biológica. Durante essa fase, as coordenadas do sistema em diferentes intervalos de tempo são armazenadas sob a forma de trajetórias. Estas serão então usadas para cálculos de energia média, flutuações estruturais, entre outros.

3.2.3 Desafios da Análise de Simulações de Dinâmica Molecular

O maior desafio das simulações de DM é a sua amostragem, a qual se refere o quão bem o comportamento do sistema de partículas em estudo é descrito em uma escala de tempo. Sistemas complexos, como por exemplo aqueles que envolvem biomoléculas, exigem um maior tempo de simulação. O que gera dezenas de gigabytes de dados dificultando as análises e interpretação das trajetórias. Além disso, mesmo aplicando tempo de simulação maior, não há garantia de que o sistema não irá sair de um provável mínimo local. Desta forma, técnicas como simulação usando campos de força do tipo átomo unido ou *coarse-grained*, análises de modos normais e métodos de mineração de dados envolvendo abordagens estatísticas e de inteligência artificial, podem ser aplicados para solucionar os desafios gerados pela amostragem da simulação de DM.

Os modelos *coarse-grained* são amplamente usados para redução do tempo e custo computacional das simulações de sistemas proteicos complexos (40)(52). No entanto, esta técnica apresenta o desafio de retornar um modelo simplificado para o modelo *all-atom* (83)(89). Quanto mais simplificada a representação, mais difícil será a reconstrução de todos os átomos do sistema.

Uma forma alternativa aos métodos de simulações de DM para predizer o conjunto de conformações que uma biomolécula pode assumir é a técnica de análise de modos normais.

A qual investiga os movimentos vibracionais de um sistema de oscilação harmônica na vizinhança do ponto de equilíbrio, isto é, a soma das forças atuando sobre o sistema é igual a zero (63)(91). Além disso, as conformações obtidas pela técnica são independentes entre si, o que dificulta observar mudanças conformacionais relacionadas a escala temporal (63).

3.2.4 Dinâmica Molecular Acelerada

A complexidade e os altos graus de liberdade dos sistemas biológicos configuram uma superfície de energia potencial "rugosa", com um conjunto de mínimos locais separados por barreiras energéticas (73). Estas são suficientemente altas, de forma a impedir que as pequenas flutuações no sistema inicial, geradas por uma cMD, acessem todos os mínimos. Em geral os estados acessados são próximos ao mínimo inicial de energia (60)(7).

De acordo com o teorema da ergordicidade (90), sob um período prolongado de tempo, todos os microestados de energia acessíveis são igualmente prováveis. Porém, uma vez que em simulações convencionais o sistema pode ficar estagnado em regiões de mínimo local, as propriedades termodinâmicas de interesse para sistemas biológicos podem não ser medidas corretamente. Sendo necessário longos períodos de simulação para se atingir a ergodicidade do sistema. Ou ainda, o uso de técnicas que possibilitem explorar o espaço conformacional em um tempo de simulação mais curto, como por exemplo a DM acelerada (aMD).

A dinâmica molecular acelerada (aMD) é um método de amostragem que serve para melhorar a exploração do espaço conformacional de um sistema (44). O método de aMD modifica o formato da superfície de energia potencial ao elevar os valores de energia nos poços de mínimo que estão abaixo de um certo limiar (*threshold*), enquanto que valores de energia acima desse limiar não são afetados (44)(108). Como resultado, as barreiras que separam as bacias energéticas são reduzidas, permitindo ao sistema amostrar espaços conformacionais que não poderiam ser facilmente acessados em uma simulação clássica de DM (6).

No modelo padrão de simulação aMD proposto por Hamelberg *et. al* 2004 , o sistema evolui naturalmente até atingir um valor de energia potencial inferior ao limiar estabelecido previamente. Este valor, em geral, pode ser definido como o valor médio da energia potencial de equilibração do sistema (E). Quando o valor pré-estabelecido é atingido, adiciona-se à energia potencial $V(\mathbf{r})$ do sistema um impulso $\Delta V(\mathbf{r})$ e obtendo um novo valor de energia potencial $V^*(\mathbf{r})$, descrito pelas equações 3.4 e 3.5:

$$V^*(\mathbf{r}) = V(\mathbf{r}) + \Delta V(\mathbf{r}) \tag{3.4}$$

$$\Delta V(\mathbf{r}) = \begin{cases} 0 & V(\mathbf{r}) \geq \mathbf{E} \\ \frac{(\mathbf{E}-V(\mathbf{r}))^2}{\alpha+\mathbf{E}-V(\mathbf{r})} & V(\mathbf{r}) < \mathbf{E} \end{cases} \tag{3.5}$$

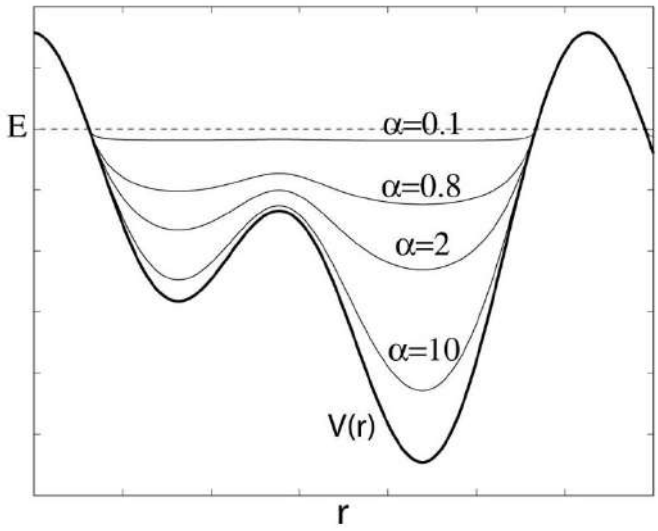


Figura 3 – Esquema do método de aMD. Energia potencial original $V(\mathbf{r})$ (linha espessa), limite de energia (*threshold*) $\mathbf{E}$ (linha tracejada) e perfis de energia modificados (linhas finas) variando de acordo com o parâmetro α. Quanto menor os valores de α, menores são as barreiras que separam as bacias de energia no espaço de fase (6).

3.3 Redução de Dimensionalidade em Simulações de Proteínas

Dados com alta dimensionalidade, isto é, grande número de características, representam um desafio para os algoritmos de agrupamento, pois aumentam o custo computacional e dificultam a obtenção de gruppos coesos. Este problema pode ser agravado quando os grupos apresentam formas, tamanho e densidade amplamente diferente, os quais podem ser resultantes de ruídos e valores atípicos dos dados. Assim, para estes conjuntos de dados é comum aplicar métodos de redução de dimensionalidade. O objetivo destes métodos é representar o conjunto de dados original em um espaço com dimensão reduzida, sem perda de informação significativa.

Conforme citado anteriormente, a técnica de PCA é amplamente usada para análise de movimentos de proteínas oriundas de simulações por dinâmica molecular. Esta abordagem foi introduzida por Ichiye e Karplus (1991) com o propósito de obter um conjunto de vetores ortogonais, os quais pertencem a um subespaço denominado "subespaço essencial", expressando o movimento proteico em uma dimensão reduzida e em termos singulares (49)(33). O uso do PCA como redução de dimensionalidade para análise de trajetórias de proteínas tem se tornado comum devido a dois fatores: (i) as bases e princípios matemáticos da técnica são bem estabelecidos na literatura e, (ii) as

componentes principais podem de certa forma ser mapeadas de volta à estrutura proteica em estudo (96).

Contudo, embora o PCA consiga detectar conformações ao longo de uma trajetória, têm sido proposto e explorado na literatura o uso de novos métodos, especialmente não-lineares para obtenção de um espaço intrínseco das trajetórias proteicas (17). Conforme descrito por Whitford e Onuchic (2015), o processo de enovelamento de uma proteína envolve mudanças na cadeia principal (*backbone*) e reorientações das cadeias laterais dos aminoácidos, onde resíduos menos estáveis podem assumir estados de desordem enquanto a molécula permanece predominantemente enovelada (112). Assim, o uso de abordagens não-linerares popdem explorar melhor o espaço de fase de proteínas.

3.4 Métodos de Agrupamento

O agrupamento (ou *clustering*) é uma técnica de aprendizado não-supervisionado que permite capturar padrões e correlações entre os dados de um conjunto, a partir da identificação de regiões densas e esparsas geradas pela distribuição dos dados em um espaço multidimensional (26). Em teoria, ao realizar o agrupamento de objetos em um conjunto de dados, os objetos dentro de um mesmo grupo (*cluster*) possuem alta similaridade entre si e alta dissimilaridade de objetos que estão em outros grupos. Esta similaridade é baseada em métricas de distância geradas usando os valores dos atributos que descrevem os objetos (45).

Vários algoritmos de agrupamento usando diferentes abordagens têm sido desenvolvidos. Desta forma, pode-se categorizá-los em (81):

1. Métodos de particionamento: considerando n amostras ou objetos, este tipo de método constrói k partições, onde cada partição representa um grupo e $k < n$. Fornecido k, a determinação das partições é realizada por meio de uma técnica de realocação iterativa que busca melhorar o particionamento movendo objetos de um grupo para outro.

2. Métodos hierárquicos: os objetos do conjunto de dados são separados hierarquicamente por sua similaridade, resultando em uma representação semelhante a um dendograma que pode expressar o processo de união ou divisão entre os grupos e todos seus níveis intermediários. Esse método pode ser aglomerativo, quando cada objeto é considerado uma folha da árvore, isto é, um grupo unitário e então são unidos até que estejam em grupos homogêneos ou que satisfaça a condição de agrupamento; ou pode ser divisivo, quando segue um processo *top-down* da árvore, todos objetos em um grupo e a cada iteração subgrupos são formados até cada objeto estar em um grupo ou uma condição de parada seja satisfeita. O resultado dos métodos hierárquicos é obtido cortando-se o dendograma gerado em um nível

de similaridade desejado entre os objetos. Esses métodos podem usar diferentes funções de similaridade e portanto podem ser sub-categorizados em (31)(68): (i) *Single-linkage*, o qual considera que a distância entre dois diferentes grupos é igual a menor distância de algum objeto de um dos grupos em relação a algum objeto presente no outro grupo; (ii) *Complete-linkage*, de forma contrária ao *single-linkage*, este considera que a distância entre dois grupos é a maior distância entre seus objetos; (iii) *Average-linkage*, este método diferente dos anteriores considera a distância entre dois grupos como a distância média entre os seus objetos; (iv) *Ward-linkage*, o qual realiza a minimização da variância entre dois grupos; (v) *Centroid-linkage*, a distância entre os grupos é representada pela distância entre as suas estruturas representativas (*centroids*), isto é, por seus centros geométricos.

3. Métodos baseados em densidade: estes métodos assumem que os objetos podem ser separados em grupos a partir de uma distribuição de probabilidade que represente cada grupo, sendo o conjunto de dados considerado uma mistura de distribuições. O crescimento de determinado grupo ocorre até que a densidade (número de objetos no grupo) na vizinhança exceda um limiar (ou *threshold*), isto é, para cada objeto de um cluster a sua vizinhança, em um determinado raio, deve conter um número mínimo de objetos.

4. Métodos baseados em malhas (*grids*): esse método quantifica o espaço dos objetos em um número finito de células que formam uma estrutura de malha. A vantagem desta abordagem é o processamento rápido uma vez que é independente do número de objetos e dependente somente do número de células em cada dimensão.

5. Métodos baseados em modelo: são definidos modelos hipotéticos para cada um dos grupos e é buscado o melhor encaixe dos dados nos modelos. O algoritmo baseado em modelo então deve estabelecer grupos construindo uma função de densidade que reflita a distribuição dos objetos.

Os algoritmos de agrupamento podem usar diferentes atributos para determinar a proximidade dos dados do conjunto de entrada. Estes atributos são usados como coordenadas em um espaço multidimensional, onde funções de distâncias são aplicadas para calcular a similaridade entre os pontos. Para dados providos de DM o principal atributo usado é o RMSD (desvio médio quadrático, do inglês *Root Mean Square Deviation*) das distâncias das trajetórias dos átomos de cada conformação (74)(62)(77). Em geral é calculada uma matriz de RMSD entre todas as conformações. O RMSD é uma interessante propriedade para determinar a semelhança de estruturas em uma trajetória ao longo de diferentes escalas de tempo e para avaliar a convergência da estrutura média.

Neste trabalho, foram utilizadas as matrizes de distância euclidiana entre átomos dentro de cada conformação. Esta métrica, considerada como coordenada interna, possui

vantagem sobre a métrica RMSD uma vez que permite analisar diretamente as flutuações atômicas do movimento proteico e obter mapas de distribuição de energia livre.

3.5 Determinando o Número de Grupos em Conjuntos de Dados

Alguns algoritmos de agrupamento requerem um conjunto de parâmetros os quais devem ser informados pelo usuário. Estes parâmetros variam entre as diferentes abordagens de agrupamento, alguns métodos necessitam que o número de grupos seja informado *a priori*. No entanto, obter este valor, comumente denominado como k, requer um conhecimento detalhado e prévio do conjunto de dados, o que em geral pode ser difícil e demandar um grande período de tempo (84). Assim, diferentes formas de detectar automaticamente o número de grupos têm sido criadas. De acordo com Maimon e Rokach, 2010, os métodos de detecção do número de k podem ser categorizados em três classes (81):

1. Métodos baseados na dispersão intra-grupos: estes métodos utilizam o princípio que à medida que o número de grupos aumenta, a curva de dispersão cai rapidamente. Após certo k, a curva é suavizada. Quando isto ocorre, este valor de k é considerado o valor apropriado.

2. Métodos baseados na dispersão intra e inter-grupos: apesar dos métodos baseados apenas na dispersão intra-grupos serem capazes de detectar o número de grupos esperados em um conjunto de dados de forma automática, em determinados casos podem falhar. Especialmente quando um determinado grupo esperado pode ser dividido em sub-grupos. Nestes casos, minimizar o valor da dispersão intra-grupo e maximizar a dispersão inter-grupos pode ser interessante para manter a coesão dos grupos gerados.

3. Métodos baseados em probabilidade: esta abordagem de detecção do número de k baseia-se em uma função de máxima verossimilhança para avaliação do modelo de agrupamento. Conforme a adição de parâmetros, neste caso o número de k, o valor de verossimilhança aumenta até determinado momento em que o resultado obtido não apresenta mais mudanças.

3.6 Métricas de Avaliação do Agrupamento

Avaliar em um método de agrupamento a qualidade da separação dos objetos em um conjunto de dados pode ser controverso e problemático (81). Não há uma definição exata sobre o que é um bom agrupamento, o que compromete determinar uma métrica universal para avaliar um algoritmo de agrupamento (88). A fim de solucionar este impasse, várias medidas têm sido propostas, as quais podem ser classificadas em (55):

1. Abordagens internas de avaliação: utilizam alguma medida de similaridade para determinar a compacidade dos grupos obtidos. Em geral, estas medidas avaliam a homogeneidade intra-grupos e a separação inter-grupos.

2. Abordagens externas de avaliação: utilizam informações dos possíveis grupos, obtidas previamente por algum especialista no conjunto de dados. Estas métricas podem ser comparadas àquelas usadas para avaliar métodos de classificação supervisionados.

3.7 Atracamento Molecular (*Molecular Docking*)

Na natureza, mais especificamente em sistemas biológicos, os mecanismos e processos celulares não ocorrem envolvendo apenas uma molécula, mas em geral a interação entre várias resultando em uma resposta ou sinal (106). E a compreensão de tais fenômenos requer determinar e estudar como as moléculas biológicas interagem entre si de forma a atuar, por exemplo, como agonista/antagonista em algum processo fisiológico. Para tal, técnicas de atracamento molecular (*ou docking molecular*) têm sido amplamente usadas (106).

O *docking* molecular tem como objetivo predizer possíveis conformações que uma molécula pode assumir no sítio de ligação de um receptor-alvo de interesse (43) e o grau de afinidade de interação entre eles. No geral, estes métodos incluem um algoritmo de busca para gerar as poses do ligante e uma função de energia usada para avaliar a afinidade com o receptor (43). O sucesso de um algoritmo na predição das interações receptor-ligante depende de como os métodos de busca lidam com a flexibilidade molecular e o quão bem as funções de energia descrevem os contatos entre as moléculas (43).

As metodologias de atracamento molecular proteína-ligante baseiam-se no clássico modelo chave-fechadura, proposto por Emil Fischer em 1894 (34), o qual considera que a interação entre o ligante e a proteína está relacionada a uma das possíveis ações de "abrir ou trancar" uma porta (106). Sendo, neste modelo, o receptor proteico associado à uma fechadura e o ligante à chave. Apesar de sua importância teórica, biologicamente as estruturas do receptor proteico e do ligante são flexíveis, de tal maneira que durante a interação tanto o ligante como o receptor podem apresentar mudanças em suas conformações.

Desta forma foi proposto o modelo de encaixe induzido, no qual o ligante ao interagir com o receptor induz alterações conformacionais neste que otimizam o complexo receptor-ligante (54)(27). Em contrapartida, foi proposto também o modelo da seleção conformacional, o qual sustenta a hipótese de que o receptor apresenta um conjunto de estados similares energeticamente e o ligante interage com um destes. Essa interação promove o deslocamento do equilíbrio químico de tal forma que a proporção de estados favoráveis à interação aumenta (67)(27).

De acordo com Meireles *et. al.* (2011), apesar das diferenças dos modelos citados

anteriormente, ambos coexistem (67). A de seleção conformacional, descrita pelo modelo de chave-fechadura, tem papel dominante na definição das mudanças estruturais de larga escala e que são exploradas pelo ligante, especialmente na fase inicial da ligação. Enquanto rearranjos de cadeia lateral dos resíduos próximos ao substrato, consideradas alterações locais e específicas, são explicadas pelo modelo de encaixe induzido (67)(27).

De maneira geral, o processo de reconhecimento molecular possui duas características importantes, sendo a primeira a especificidade, que distingue substratos altamente específicos daqueles menos específicos. A segunda é a afinidade, a qual determina que mesmo em alta concentração, moléculas com fraca interação não são capazes de induzir o mesmo efeito de ligantes específicos e com alta força de interação, mesmo que estes estejam em concentrações baixas (27). Ambas características dependem de propriedades físico-químicas, definidas por interações intermoleculares (*e.g.*, ligações de hidrogênio e interações de van der Waals, eletrostáticas e hidrofóbicas), e propriedades estruturais associadas com as variações na orientação espacial das ligações químicas (106). Além disso, tal processo é dirigido por uma combinação de efeitos entálpicos e entrópicos que podem ser estimados através da energia livre de ligação de Gibbs (eq. 3.6) e que está relacionada com a constante de equilíbrio de ligação K_{eq} (106)(72):

$$\Delta G_{lig} = \Delta H - T\Delta S = -RT \ln K_{eq} \tag{3.6}$$

na qual, ΔH é a variação de energia total do sistema, isto é, a variação da entalpia; T é a temperatura do sistema; ΔS é a variação entrópica; e R é a constante universal dos gases. A constante de ligação K_{eq} é determinada experimentalmente e pode ser calculada conforme apresentada a seguir (12)(106).

Considerando que a reação enzimática é definida por:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_1}{\rightarrow} E + P \tag{3.7}$$

sendo que E denota o receptor, S, ES o complexo receptor-ligante, P o produto da reação, k_1 a constante de associação, k_{-1} constante de dissociação e k_2 a constante de dissociação do complexo receptor-produto, e considerando apenas a formação e dissociação do complexo ES, $(E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES)$ temos que:

$$K_{eq} = \frac{[E]\,[S]}{[ES]} = \frac{k_{-1}}{k_1} \tag{3.8}$$

na qual, $[E]$ é a concentração enzimática, $[S]$ a concentração do substrato e $[ES]$ a concentração do complexo enzimático.

Em experimentos de atracamento molecular, os algoritmos de busca são usados para explorar a superfície da energia livre de ligação, descrita acima, a fim de encontrar

as melhores conformações do ligante (43)(72). Desta forma, considerando que os efeitos entálpicos e entrópicos foram corretamente modelados pela função de energia, então o mínimo global da superfície de energia estará associado ao modo de ligação receptor-ligante encontrado experimentalmente (72). Devido às aproximações introduzidas no modelo de interação molecular, nem sempre o mínimo global satisfaz este importante requisito (43).

4 MATERIAL E MÉTODOS

4.1 Simulações por Dinâmica Molecular

4.1.1 Preparação dos Sistemas

Cinco proteínas distintas foram usadas para simulação por dinâmica molecular:

- A calmodulina de humanos (*Homo sapiens*), obtida no banco de dados de proteínas PDB (*protein data bank*) sob o código 1CLL (14).
- O peptídeo sintético TRP-cage, obtido no banco PDB sob o código 1L2Y (69).
- A protease do HIV-1, obtida no PDB sob o código 2HB4, a qual está em um estado não ligado ao subtrato ou inibidores (46).
- smNTPDase1 (71), isoformas 1 da ATP-Difosfohidrolase (EC 3.6.1.5) de *Schistosoma mansoni*;
- smNTPDase2 (23), isoformas 2 da ATP-Difosfohidrolase (EC 3.6.1.5) de *Schistosoma mansoni*.

As três primeiras proteínas foram obtidas a partir do *Protein Data Bank* (PDB) [1] e usadas nas análises das abordagens de seleção das conformações proteícas. As outras duas proteínas foram modeladas anteriormente por nosso grupo e representam alvos terapêuticos para o tratamento da esquistossomose (23)(71).

Durante o preparo dos sistemas de simulação para as proteínas 1L2Y, 1CLL e 2HB4, todos os heteroátomos foram removidos com o objetivo de aumentar a busca conformacional. Contudo, a fim de verificar a influência dos heteroátomos na estruturas das proteínas smNTPDase1 e smNTPDase2, dois diferentes sistemas foram construídos para cada enzima. No primeiro foi considerado os heteroátomos presentes na proteína, enquanto que no segundo os mesmos foram removidos.

Uma vez que a enzima smNTPDase1 possui duas hélices transmembrares, no preparo dos seus sistemas foi gerada uma membrana de fosfatidilcolina (POPC). Este fosfolipídio é o principal constituinte da membrana plasmática do tegumento de *S. mansoni* (72).

Todas as proteínas foram inseridas em caixas octaédricas solvatadas com água modelo TIP3. Cada caixa foi extendida por 15Å a partir da proteína, nos eixos X, Y e Z. Após a adição de moléculas de água, os sistemas foram neutralizados com 150mM de NaCl. Para os sistemas smNTPDase1 e smNTPDasee2, foram adicionados também

[1] https://www.rcsb.org/

5mM de KCl, 2 mM de $MgCl_2$ e 5mM de $CaCl_2$, simulando as concentrações referentes ao ambiente sanguíneo do hospedeiro humano.

Todos os sistemas de simulação foram preparados no programa VMD versão 1.9.2 (48). O total de átomos para os sistemas testes foi: Para a proteína 1CLL - 48.083 átomos, dos quais 45.711 são referentes à moléculas de água, 66 íons sódio e 43 são íons cloro; 1L2Y, 10,873 átomos, sendo que 10.548 pertencem à moléculas de água, 10 íons sódio e 11 são íons de cloro. Para o sistema de validação 2HB4, o número de átomos foi de 16.544, sendo que 13.368 átomos pertencem às moléculas de água, 13 íons sódio e 19 são íons de cloro. Para os sistemas de estudo de caso, o total de átomos foi: smNTPDase 1, 275.959 átomos, dos quais 214.332 são referentes à moléculas de água, 52.394 pertencentes aos lipídeos de membrana, 8.728 da proteína, 8 íons cálcios, 7 potássio, 3 de magnésio, 202 de sódio, 241 de cloro e 44 que pertencem ao ligante ANP; Para a smNTPDase 2, o sistema completo contém 151.332 átomos, sendo 143.391 referente as moléculas de água, 7.592 de proteína, 5 íons cálcio, 2 de potássio, 2 de magnésio, 135 de sódio, 162 de cloro e 43 átomos referente ao ligante AU1.

4.1.2 Cálculos de Dinâmica Molecular

Os cálculos de dinâmica molecular para os sistemas de teste foram realizados usando 2 cristais de proteínas obtidas no banco de dados de proteínas PDB (*Protein Data Bank*): 1CLL, referente a calmodulina de *Homo sapiens* com 148 resíduos aminoácidos e 1L2Y, um peptídeo desenhado artificialmente com 20 resíduos. Afim de obter um conjunto de avaliação, as proteínas foram simuladas em duas diferentes temperaturas (310K e 510K), assumindo que em altas temperaturas as proteínas assumem maior conjunto conformacional, passando por processos de desnaturação e renaturação, por exemplo. Ao assumirem maior conjunto de estados conformacionais, as proteínas podem ultrapassar barreiras energéticas e explorar mais regiões de sua superfície de energia livre.

Já para as simulações de DM do sistema de validação, realizado com a proteína protease do HIV-1(PDB2HB4), e dos estudos de caso com as enzimas smNTPDase 1 e smNTPdase 2, uma abordagem diferente foi usada para explorar as mudanças de estados e assim a superfície do espaço de fase. Como alternativa as simulações em alta temperatura, realizaou-se simulações aceleradas, as quais permitem a redução das barreiras de energia que separam os diferentes estados de um sistema através da adição de um potencial extra (potencial de *boost* - impulso) aos ângulos diedrais da molécula.

Após o preparo dos sistemas, foram realizados passos de minimização de energia de 100ps, seguido por um passo de aquecimento aumentando a temperatura do cristal até 310K ou 510K, nos casos das proteínas 1CLL e 1L2Y. Para os sistemas 2HB4 e das NTPDases, o aquecimento foi realizado apenas até 310K. Uma vez atingida a temperatura de interesse, foram realizados passos de equilibração e dinâmica de produção. As simulações

foram executadas usando o programa NAMD v2.12, com o campo de força CHARMM27.

Conforme descrito na subseção sobre dinâmica molecular acelerado da seção 3, dois parâmetros são necessários para determinar o potencial de impulso (*boost*): o *treshold* da energia potencial **E** e α. No presente estudo, os valores referente a ambos parâmetros foram definidos baseando-se na média da energia diedral do sistema após equilibração seguindo as seguintes equações:

$$\mathbf{E} = \bar{V}\left(\mathbf{r}\right)_{dihed} + 4 * N_{residues} \tag{4.1}$$

$$\alpha = \frac{4 * N_{residues}}{5} \tag{4.2}$$

onde $\bar{V}\left(\mathbf{r}\right)_{dihed}$ e $N_{residues}$ são a média da energia potencial diedral do sistema e o número de resíduos aminoácidos da proteína-alvo, respectivamente.

Durante as simulações de DM, tanto aMD quanto cMD, as interações de Coulomb de longo alcance foram calculadas pelo método PME (*Particle Mesh-Ewald*), com um *cutoff* de 12Å e *switching* igual a 10Å. O algoritmo SHAKE foi aplicado usando um passo de integração de 2fs no algoritmo de Verlet. Para o controle de temperatura foi usado o método de Langevin com coeficiente de *damping* de 1ps. Para controle da temperatura foi aplicado o método Langevin piston.

Todos os sistemas foram simulados usando o *ensemble* NPT com pressão de 1atm. As simulações dos sistemas referentes a 1CLL e 1L2Y foram realizadas em um computador com quatro CPU AMD Opteron™ processador 6272 (64 cores de 2.1GHz e memória cache de 2MB), 128 Gb de RAM, HD de 500 Gb, quatro aceleradores gráficos Nvidia Tesla M2090 com 6 Gb cada e com sistema operacional CentOS *release* 6.5. Já os sistemas referentes as proteínas 2HB4, smNTPDase1 e smNTPDase2 foram executadas no supercomputador Santos Dumont [referenciar].

4.2 Análise das Simulações

A análise de estabilidade das proteínas foi realizada a partir das energias potenciais. Para isso foi utilizado o *plugin Namd energy* do programa VMD (48), no qual foram selecionados apenas átomos referentes às proteínas. Adicionalmente, foram avaliados a variação das medidas de raio de giro, RMSD (do inglês, *Root Mean Square Deviation*), RMSF (do inglês, *Root Mean Square Fluctuation*) e SASA (do inglês, *Solvent-Accessible Surface Area*) para cada proteína ao longo de cada simulação. Para isto, implementou-se *scripts* na linguagem Python utilizando os pacotes Pyemma(87) e MDtraj(66).

4.3 Conjunto de Dados para Agrupamento

Os dados das simulações de DM foram coletados a cada 2ps de simulação, resultando em 10.000 frames para as simulações da 1CLL e 1L2Y. Estas foram separadas a cada 20 passos, o que resultou em 500 conformações para cada conjunto de trajetórias, as quais foram utilizadas nos testes de agrupamento e detecção de conformações representativas.

De forma similar, para o sistema de validação usando a proteína protease do HIV-1 (PDB2HB4) foi gerado um total de 500.000 e 200.000 frames para as simulações cMD e aMD, respectivamente. Na simulação convencional, selecionou-se uma conformação a cada 50 passos, totalizando um total de 10.000 conformações onde cada pose representa 100ps de simulação. Já na simulação acelerada, o conjunto de poses apresentou 10.000 frames, sendo que as conformação foram separadas a cada 20 passos, de forma que elas representassem 40ps de simulação.

Nos sistemas de estudo de caso, tanto para a smNTPDase1 quanto smNTPDase2, as separações de conformações seguiram as mesmas regras. Nas simulações convencionais, o total de frames foi de 120.500 dos quais foram selecionados 6025 poses, representando 40ps de simulação cada uma delas. Enquanto que nas simulações aMD, foram obtidos 25.000 frames e todos foram usados nos testes de agrupamento para detecção de conformações significativas.

Em todos os conjuntos de dados, foi calculada a matriz de distâncias euclidiana (*Euclidean Distance Matrix* - EDM) entre os átomos de carbono α para cada conformação. Conforme apresentado na figura 4, as EDM foram usadas como entrada para os métodos de redução de dimensionalidade, afim de encontrar o espaço essencial que descreve os movimentos proteicos. Após isso, foram aplicados algoritmos de agrupamento sobre o espaço reduzido para separar os diferentes estados e movimentos significativos das proteínas.

4.4 Normalização dos Dados

A normalização refere-se a um conjunto de técnicas de pré-processamento dos dados no qual cada amostra com pelo menos um elemento diferente de zero tem seus valores em diferentes escalas ajustados para uma escala comum (45). Neste estudo foi aplicada a técnica de normalização Min-Max, a qual realiza a transformação linear dos dados de forma que uma vez fornecida uma amostra A com valores de mínimo e máximo iguais a min_A e max_A, os valores v_i de A são mapeados para v_i' em um novo intervalo [min_B, max_B], onde min_B e max_B são valores informados pelo usuário, preservando a relação original (45). Tal transformação pode ser escrita como:

$$v_i' = \frac{v_i - min_A}{max_A - min_A}\,(max_B - min_B) + min_B \tag{4.3}$$

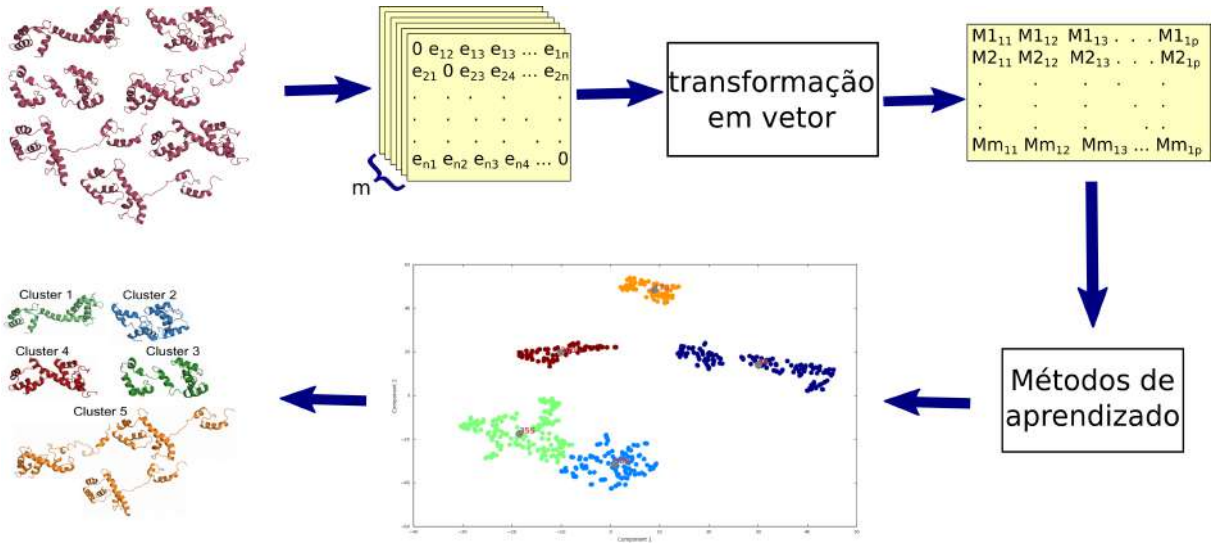


Figura 4 – Esquema obtenção da matriz de características para redução de dimensionalidade e agrupamento de conformações proteícas. Os detalhes do processo serão descritos ao longo do texto.

4.5 Redução de Dimensionalidade dos Dados

Para redução da dimensionalidade intrínseca dos dados oriundos de simulações de DM, foram usados 6 diferentes métodos, incluindo o método linear PCA e não-lineares. A seguir, uma breve descrição dos métodos utilizados:

1. AutoEncoder (AE): é um tipo de rede neuronal artificial capaz de aprender a representação de um conjunto de dados de maneira não-supervisionada (61). A proposta deste método é compactar dados de uma camada de entrada com representação multi-dimensional para uma camada intermediária com uma dimensão reduzida e em seguida recuperar a representatividade dos dados. Usando esta ideia o AutoEncoder consegue ignorar os ruídos e resolver o problema do mal dimensionamento (61). No presente trabalho, foi utilizado uma rede AutoEncoder onde a quantidade de neurônios na camada de entrada foi igual ao número de distâncias entre os átomos Cα de cada uma das proteínas usadas aqui. Após isto, cinco camadas intermediárias foram usadas com as seguintes quantidades de neurônios: 1024, 512, 128, 32 e a camada de codificação com 2 neurônios (figura 5). As funções de ativação usadas foram função linear para camada de codificação, sigmóide para decodificação e exponencial linear para as demais camadas (figura 5). Além disso, foram usadas 200 épocas de treinamento da rede neuronal.

2. *Multi Dimensional Scaling* (MDS): é um algoritmo clássico de redução de dimensionalidade não-linear que projeta as coordenadas dos pontos de um conjunto de entrada com alta dimensão para um espaço reduzido, mantendo as relações de distância par a par entre os pontos $\left|\vec{X}_i - \vec{X}_j\right|^2$ (110). MDS tem sido amplamente usado para

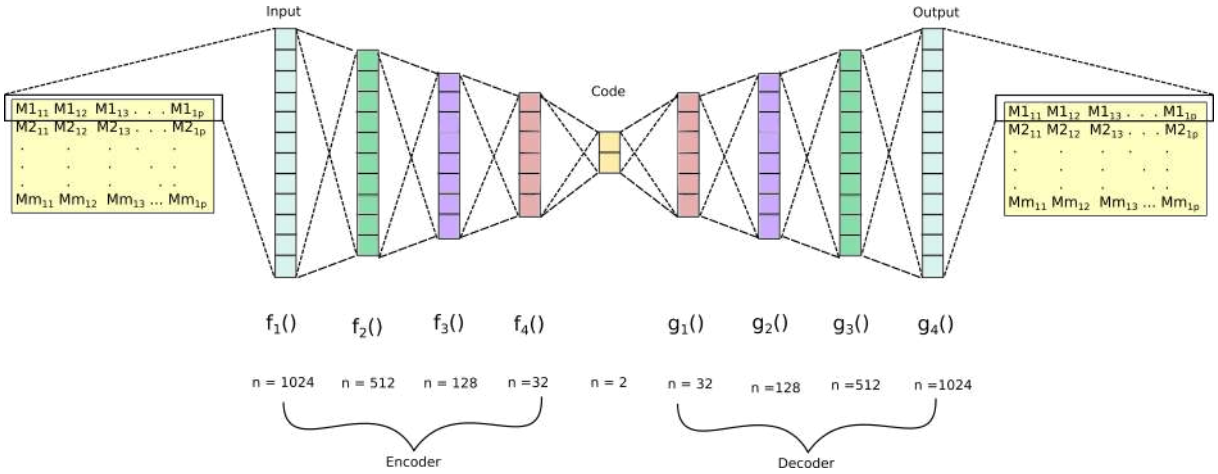


Figura 5 – Representação da rede neuronal autoencoder usada neste estudo. Uma rede autoencoder com 4 camadas para codificação e decodificação. Nas camadas de codificação as funções f_1 até f_3 são *Exponential Linear Unit* (ELU) e f_4 é uma função linear. Já para a camada de decodificação foi implementadas de g_1 até g_3 funçõe ELU e em g_4 uma função sigmóide. Na figura, n representa o número de neurônios em cada camada (22).

solucionar problemas envolvendo EDM, como por exemplo, problemas em encontrar a melhor representação de um conjunto de pontos dado um conjunto de distâncias (25).

3. Isomap: outro algoritmo não-linear amplamente usado para redução de altas dimensões, pode ser compreendido como uma extensão do MDS em um espaço de distâncias geodésicas. O método consiste em definir um número n de vizinhos mais próximos e criar um grafo de vizinhança onde cada ponto está conectado a outro se este é algum de seus vizinhos mais próximos, sendo o tamanho das arestas definido pela distância euclidiana. Após isso, é calculado o caminho mais curto entre dois nós e então aplicada a redução de dimensionalidade realizada pelo algoritmo MDS (38).

4. t-SNE: este método é uma variação do *Stochastic Neighbor Embedding* (SNE) o qual converte distâncias euclidiana dos pontos em alta dimensão para probabilidades gaussianas que modelem a similaridade entre os pontos. t-SNE usa uma distribuição t-student para representar as afinidades entre os pontos de um conjunto de dados em um espaço reduzido, o que torna o método sensível à estrutura local dos dados (64).

5. *Spectral embedding*: este método usa a decomposição espectral de um grafo laplaciano gerado a partir da matriz de similaridade dos pontos de um conjunto de dados (5). Este grafo é considerado uma aproximação discreta do *manifold* de baixa dimensão em um espaço de alta dimensionalidade (5). Após a decomposição espectral, os primeiros autovetores da matriz laplaciana são usados para representar os pontos em uma baixa dimensão (5).

6. Análise das componentes principais (PCA): o PCA é um método de redução de dimensionalidade linear, diferente dos apresentados anteriormente. O método usa a decomposição de valores singulares (SVD) dos dados para projetá-los em um subespaço que representa a maior variabilidade dos dados. Este subespaço linear pode ser determinado pelos vetores ortogonais que formam um novo sistema de coordenadas, comumente chamado de "componentes principais" (1).

4.6 Algoritmos de Agrupamento

Neste trabalho foram usados quatro diferentes algoritmos de agrupamento, como descrito abaixo:

1. *k-means*: considerando n como o número de objetos de uma amostra, este método constrói k partições, onde cada uma destas representa um grupo, satisfazendo a condição $k \leq n$. Dado k, o particionamento é realizado por uma técnica de realocação iterativa dos centróides que busca aumentar a separação dos objetos distintos movendo-os de um grupo a outro (78).

2. *Agglomerative Ward*: os objetos são hierarquicamente decompostos, resultando em uma representação similar a um dendograma, o qual realiza o processo de união dos objetos do conjunto de dados em k grupos, fornecido pelo usuário (109).

3. *Affinity propagation*: é um método baseado no conceito de "passagem de mensagem" por grafos interconectados, os quais são gerados a partir dos pontos do conjunto de dados (vertex) (36). Este algoritmo diferente dos métodos de particionamento e hierárquicos, não requer o parâmetro "número de grupos" (k) *a priori*.

4. *Mean-shift*: é um método baseado em densidade. O objetivo do algoritmo é detectar regiões similares a "bolhas" presentes em amostras de densidade suavizada. A cada passo de iteração, os candidatos a centróides são atualizados. O *Mean-shift* determina o número de grupos k automaticamente usando o parâmetro *bandwidth*, o qual determina o tamanho da região a ser pesquisada (15).

Todos os algoritmos de agrupamentos usados neste trabalho foram implementados usando o pacote scikit-learn v0.18 em python v2.7. Os experimentos computacionais foram realizados em um computador intel® core™ i7 860 2.8 GHz, 8Gb de RAM, HD de 860 Gb, um acelerador gráfico Nvidia Geforce GTX 285 com 1Gb e sistema operacional Fedora *release* 23.

4.7 Predição do Número de Grupos

Neste trabalho foram usadas quatro diferentes abordagens de detecção do número de grupos: *elbow*, *Bayesian information criterion* (BIC), estatística GAP e maximização do valor de silhoueta. Os valores preditos por estes métodos foram usados como parâmetro para os algoritmos k-means e *agglomerative ward*.

1. *Bayesian information criterion*: é um método estatístico usado para seleção de modelos baseado na função de verossimilhança. Essa abordagem pode ser aplicada na seleção de modelos com diferentes números de parâmetros. Quando usado em modelos de agrupamento, o BIC avalia o aumento da verossimilhança como uma função dependente do número de grupos. A fórmula geral do BIC é dada pela equação 4.4:

$$BIC = L(\theta) - \frac{1}{2} k \log n \tag{4.4}$$

onde $L(\theta)$ é a função de verossimilhança definida para cada modelo, k é o número de grupos e n é o tamanho do conjunto de dados. Estendendo a função de verossimilhança, conforme mostrado por Zhao e colaboradores (118), temos:

$$BIC = \sum_{i=1}^{k} \left(n_{\mathrm{i}} \log n_{\mathrm{i}} - n_{\mathrm{i}} \log n - \frac{n_{\mathrm{i}} * d}{2} \log(2\pi) - \frac{n_{\mathrm{i}}}{2} \log \sum\nolimits_{\mathrm{i}} - \frac{n_{\mathrm{i}} - k}{2} \right) - \frac{1}{2} k \log n \tag{4.5}$$

onde d é a dimensão do conjunto de dados, n_{i} é o tamanho do grupo e $\sum_{\mathrm{i}}$ é a estimativa da variância da máxima verossimilhança para o i-ésimo grupo, calculado usando a equação 4.6:

$$\sum\nolimits_{\mathrm{i}} = \frac{1}{n_{\mathrm{i}} - k} \sum_{j=1}^{n_{\mathrm{i}}} \| x_{\mathrm{j}} - C_{\mathrm{i}} \|^2 \tag{4.6}$$

sendo que n_{i} o número de objetos no grupo, x_{j} o j-ésimo objeto dentro do grupo e C_{i} é o centro do i-ésimo grupo.

2. *Elbow*: é um método visual de predição de número de grupos, o qual se baseia no fato de que a variância entre grupos pode ser explicada em função do número de grupos do conjunto de dados. Assim, dado o valor de número de grupos inicial k, conforme este aumenta a cada passo de iteração do método o valor de variância diminui drasticamente até atingir um platô, onde se pode observar a formação de um "cotovelo" no gráfico gerado entre o valor de k e a variância. Segundo o método, o número de grupos ideal é valor onde se observa o "cotovelo".

3. Estatística GAP: este método foi proposto por Tibshirani e colaboradores (2001) (97), e baseia-se no cálculo de uma medida de erro (qualidade) W_k do agrupamento para k grupos, a qual pode ser definida como monótona decrescente conforme os valores de k aumentam. Uma vez definida uma função W_k apropriada é possível determinar o padrão de distribuição dos pontos e assim encontrar o melhor número de grupos observando onde a função torna-se monótona. A medida de erro proposta por Tribshirani pode ser escrita como (97):

$$W_k = \sum_{r=1}^{K} \frac{1}{2n_{\mathrm{r}}} D_{\mathrm{r}} \tag{4.7}$$

onde n_r representa o número de pontos pertencentes ao grupo r e

$$D_{\mathrm{r}} = \sum_{i,j \in C_{\mathrm{r}}}^{K} d_{\mathrm{ij}}$$

sendo a soma das distâncias ou dissimilaridade entre esses.

Conforme dito anteriormente, a proposta do método GAP é determinar a distribuição do conjunto de dados e a partir da padronização do gráfico de $\log(W_k)$ comparando-o com seu valor esperado de uma distribuição de referência dos dados. Isto pode ser definido como:

$$Gap_{\mathrm{n}}(k) = {E^*}_{\mathrm{n}} \{\log(W_k)\} - \log(W_k) = \frac{1}{B} \sum_{b} \log({W_{kb}}^*) - \log(W_k) \tag{4.8}$$

onde ${E^*}_{\mathrm{n}}$ denota tal referência. Para escolha do melhor valor de k, calcula-se:

$$sd_{\mathrm{k}} = \left[(\frac{1}{B}) \sum_{b} \left\{ \log({W_{kb}}^*) - \vec{l} \right\}^2 \right]^{\frac{1}{2}} \tag{4.9}$$

com

$$\vec{l} = \left(\frac{1}{B} \right) \sum_{b} \log({W_{kb}}^*) \tag{4.10}$$

4. Maximização da silhueta: dado um conjunto de valores inteiros de K, os quais representam possíveis números de grupos de um conjunto de dados, esta abordagem baseia-se em encontrar o $k \in K$ para o qual a silhueta apresenta seu maior valor. Assim como os métodos descritos anteriormente, esta é uma técnica exaustiva e necessita de várias iterações do algoritmo de agrupamento para diferentes k.

4.8 Métricas de Avaliação de Agrupamento

A avaliação do quão bom está um determinado agrupamento pode ser uma tarefa difícil, uma vez que tal definição não é exata, variando de acordo com o conjunto de dados.

Assim, uma alternativa na avaliação das técnicas de agrupamento é a utilização de duas ou mais métricas de qualidade de forma a obter um consenso sobre as mesmas. Neste estudo, foram usadas quatro diferentes medidas de avaliação dos agrupamentos, as quais são descritas a seguir:

1. *Calinski-Harabaz Index* (CHI): criada por Calinski e Harabaz em 1974, esta métrica é definida como a razão entre a dispersão dos objetos dentro de uma mesma classe, pela dispersão entre aqueles pertencentes a classes distintas (11). A definição matemática pode ser expressa como:

$$CH(k) = \frac{B_{\mathrm{c}}(k)}{(k-1)} / \frac{W_{\mathrm{c}}(k)}{(n-1)} \tag{4.11}$$

$$B_{\mathrm{c}}(k) = \sum_{i=1}^{k} n_{\mathrm{i}} \left(c_{\mathrm{i}} - C\right) \left(c_{\mathrm{i}} - C\right)^{\mathrm{T}} \tag{4.12}$$

$$W_{\mathrm{c}}(k) = \sum_{i=1}^{k} \sum_{x \in c_{\mathrm{i}}} \left(x - c_{\mathrm{i}}\right) \left(x - c_{\mathrm{i}}\right)^{\mathrm{T}} \tag{4.13}$$

2. *Davies-Bouldin Index*: desenvolvida por Davies e Bouldin em 1979, pode ser definida como uma medida baseada na similaridade média de cada grupo com seu similar. Sendo matematicamente definida como (18):

$$DBI = \frac{1}{N} \sum_{i=1, i \neq j}^{N} max \left(\frac{S_{\mathrm{i}} + S_{\mathrm{j}}}{M_{\mathrm{ij}}} \right) \tag{4.14}$$

onde N é o número de grupos, S_i e S_j representam a distância média de todos os objetos dos grupos i e j em relação aos seus respectivos centroides, enquanto que M_{ij} é a distância entre os centroides dos grupos i e j. A dispersão intra e extra grupos podem ser escritas como:

$$S_{\mathrm{i}} = \left(\frac{1}{n_{\mathrm{i}}} \sum_{j=1}^{n_{\mathrm{i}}} |X_{\mathrm{j}} - A_{\mathrm{i}}|^{p} \right)^{\frac{1}{p}} \tag{4.15}$$

onde n_{i} é a dimensão do grupo, X_{j} é um dos objetos que pertencem o grupo i e A_{i} representa seu centróide.

$$M_{\mathrm{ij}} = \|A_{\mathrm{i}} - A_{\mathrm{j}}\|_{\mathrm{p}} = \left(\sum_{k=1}^{N} |a_{\mathrm{k,i}} - a_{\mathrm{k,j}}|^{p} \right)^{\frac{1}{p}} \tag{4.16}$$

sendo A_{i} e A_{j} os centróides obtidos para os grupos i e j, respectivamente, e $a_{\mathrm{k,i}}$ e $a_{\mathrm{k,j}}$ os elementos de tais grupos. A variável de p pode assumir qualquer valor, sendo que quando $p = 2$ temos a distância euclidiana.

3. Silhueta: é uma medida de validação da consistência de um agrupamento, a qual expressa o quão similar um objeto está de seu respectivo grupo e separado dos outros, baseando-se na diferença pareada entre as dispersões intra e extra grupos (82). A variação da silhueta é de -1 a 1, de forma que maiores valores representam maior similaridade dos objetos dentro do grupo, enquanto que aqueles menores indicam uma baixa coesão entre os objetos. Considerando um conjunto de objetos X com K partições e i um objeto que pertence uma partição $x_k \in X$, com $k \in \{1, 2, 3, ..., K\}$, considera-se o valor de silhueta s_i como (82):

$$s_{\mathrm{i}} = \frac{b_{\mathrm{i}} - a_{\mathrm{i}}}{max\left(a_{\mathrm{i}}, b_{\mathrm{i}}\right)} \tag{4.17}$$

onde a_{k} e b_{k} representam as médias das distâncias euclidianas de i em relação aos objetos do mesmo grupo k e entre aqueles contidos em outros grupos. O coeficiente de silhueta para um conjunto de objetos X é definido como a média dos valores de silhueta de cada objeto

$$SC = \frac{\sum_{i=1}^{N} s_{\mathrm{i}}}{N} \tag{4.18}$$

onde N é o número total de objetos no conjunto.

4.9 Mapas de Energia

A fim de obter mapas de energia livre para cada conjunto conformacional de cada simulação de DM, foi aplicado o método *Weight Histogram Analysis Method* (WHAM). Este método baseia-se no fato de que dado um conjunto de possíveis estados discretos de uma molécula, a energia livre pode ser calculada utilizando um histograma com um determinado número de partições (*bins*) que fornecem a probabilidade relativa de um estado ocorrer ao longo da trajetória (58). De acordo como o método, quanto maior a densidade dos estados em uma região do histograma, maior a probabilidade daquele conjunto de estados representarem um mínimo de energia. A ideia do WHAM é fundamentalmente originada da mecânica estatística em que função de energia livre $F(\cdot)$, considerando uma conformação ξ, temos que sua energia é (58):

$$F(\xi) = -k_B T \ln Z(\xi) \tag{4.19}$$

sendo:

$$Z(\xi) = \int e^{\beta U(\xi)} d\Omega \tag{4.20}$$

onde $\beta = 1/k_B T$, k_B é a constante de Boltzmann, T é a temperatura em Kelvin, $U(\xi)$ é a energia potencial e $Z(\xi)$ é a função de partição, a qual é proporcional a densidade de conformações nas células do histograma, ou seja, quanto maior o número de conformações

em uma partição do histograma, maior a probabilidade das conformações representarem um mínimo de $U(\xi)$.

Para um determinado espaço reduzido (espaço intrínseco), obtido por métodos de redução aplicados às coordenadas internas das conformações proteicas, o FEL pode ser obtido pela distribuição de probabilidade dos pontos para os k "principais" componentes (33):

$$F = -k_B T \ln \hat{P}\left(\left\{\vec{\Psi}\right\}_{i=2}^{k+1}\right) + C \tag{4.21}$$

4.10 Testes Estatísticos

Com o intuito de comparar as diferentes abordagens apresentadas, os experimentos computacionais de agrupamento foram executados 30 vezes, obtendo para cada execução os valores referentes as métricas de qualidade. Para verificar se o conjunto de valores de qualidade apresentavam distribuição normal, o teste de Anderson-Darling (4) foi aplicado. O teste de Kruskal-Wallis (KW) (57) foi usado como alternativa ao ANOVA para a análise de dados não-paramétricos. A hipótese nula (H0) testada por esse método afirma que as populações das quais os grupos de dados foram amostrados possuem a mesma distribuição. Para avaliação *post-hoc* dos resultados do KW empregou-se o teste de Dunn (29) com a correção de bonferroni. Em ambos métodos, o nível de significância (*p-value*) foi de 0,05 (5%).

4.11 Atracamento Proteína Ligante

No presente trabalho, foram avaliados a interação e potencial de inibição do composto LS1, pertencente ao grupo das chalconas, em relação as proteínas smNTPdases 1 e 2. Para os ensaios de atracamento molecular foram usadas as conformações referentes aos estados de mínimos globais de energia, para tais proteínas, obtidos pela aplicação do método WHAM sobre o espaço intrínseco gerado pelos métodos de redução de dimensionalidade. Cada um dos estados foi submetido ao programa Dockthor v2.0 (42) e os resultados foram avaliados de acordo com os *scores* gerados pelo programa.

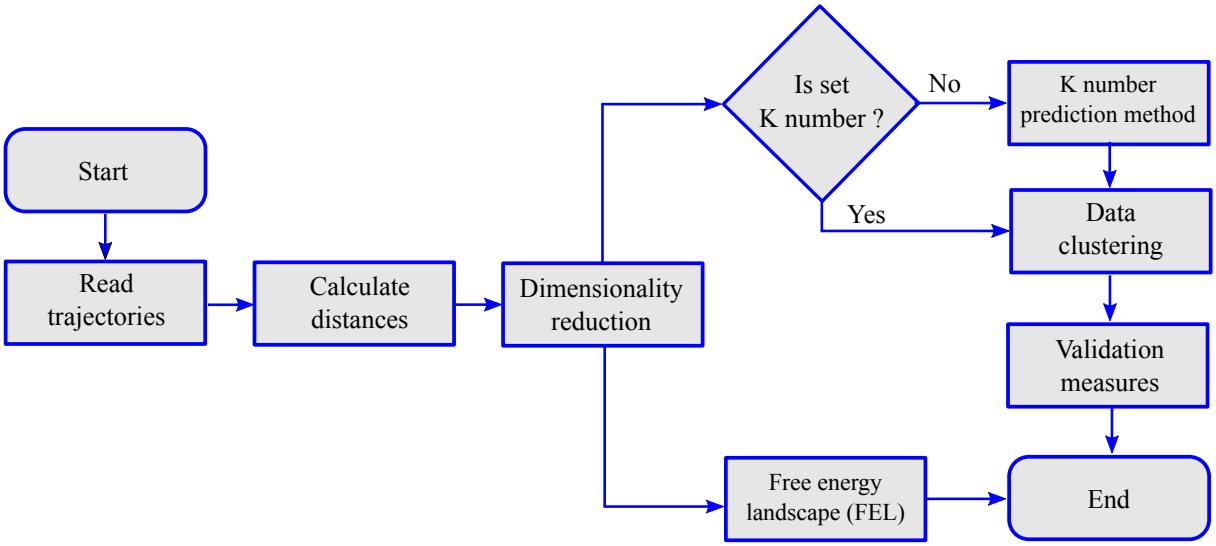
Tabela 1 – Parâmetros usados no programa Dockthor para simulações de atracamento molecular.

Parâmetro	Valor
Nº de avaliações	1000000
Tamanho da população	750
Nº de rodadas	24
seed	-1985

A *grid* usada para as simulações de atracamento molecular foram definidas no próprio programa *Dockthor*, usando as mesmas informações de centro de dimensões X, Y e Z disponíveis em Pereira *et. al.* (2018) (76). Além disso, o valor de discretização da grid foi de 0,5, e os parâmetros do algoritmo genético necessários ao programa estão descritos na tabela 1.

5 Fluxograma do Trabalho

Uma das propostas deste trabalho é o desenvolvimento de um fluxograma automatizado para a obtenção das conformações preferenciais de moléculas, seguindo os métodos aqui propostos, conforme mostrado na Figura 6. Em particular, iremos analisar conformações proteicas advindas de simulações de DM.



Os passos do presente fluxograma foram desenvolvidos utilizando a linguagem python v2.7. para leitura dos arquivos foi utilizado o pacote MDtraj(66), enquanto que os métodos de redução de dimensionalidade e agrupamento foram implementados usando o pacote scikit-learn v0.18 (75).

Figura 6 – Representação do fluxograma desenvolvido para a análise de conjunto de conformações moleculares.

Para a obtenção dos dados de entrada do fluxograma, foram executadas quatro simulações de dinâmica molecular, usando dois diferentes cristais obtidos no banco de dados PDB, os quais foram simulados em duas temperaturas (310K e 510K). Após a execução das simulações de DM, as trajetórias das proteínas no formato .dcd foram lidas, é importante ressaltar que além desse formato o fluxograma consegue ler também arquivos em .trr, .trj e .pdb. Após a importação das simulações, as distâncias entre os átomos de $C\alpha$ de cada pose foi calculada, afim de obter uma matriz EDM A com dimensão $m \times m$, na qual m representa o número de resíduos da proteína. Além desta abordagem, o fuxograma também possui a opção de utilizar a matriz de RMSD de todas as trajetórias entre si ou as coordenadas cartesianas dos átomos para determinar as conformações significativas. No entanto, estas duas últimas ainda não foram avaliadas até o momento.

As diferentes EDM obtidas são então linearizadas de forma que cada elemento A_{ij} será adicionado às colunas da matriz B de características (*features*) para o agrupamento

com dimensão $n \times p$, de forma que n é o número de conformações da trajetória e p é o número de distâncias entre os átomos de Cα. A nova matriz B representa o conjunto de coordenadas internas de cada conformação oriunda da simulação de DM e desta forma descreve as flutuações estruturais da proteína em um espaço multidimensional.

Com o intuito de obter as conformações representativas do movimento proteico, métodos de inteligência computacional não-supervisionados como os algoritmos de agrupamento são aplicados à matriz de características B. Uma vez que estas técnicas baseiam-se em observação dos dados e descoberta de padrões por si próprio, sendo importantes para separação de classes em casos nos quais não há conhecimento prévio do conjuntos de dados. Cada um dos métodos de agrupamento foi acoplado a técnicas de redução de dimensionalidade, afim de reduzir a complexidade do conjunto de características, diminuir o custo computacional associado, e em particular reduzir os ruídos oriundos da dinâmica molecular bem como também detectar o espaço essencial de características para determinar os movimentos ao longo das simulações.

Após a redução de dimensionalidade, o fluxograma segue por dois caminhos. O primeiro é a determinação dos grupos conformacionais, no qual são aplicados algoritmos de agrupamento, os quais podem ser dependentes ou não do número de grupos *a priori*. Para aqueles que são dependentes do número de gruppos k foram aplicados métodos preditores, os quais usam uma abordagem de "força bruta" na qual o agrupador é testado para um determinado intervalo de possíveis valores de k. Já para os algoritmos independentes de k nenhuma abordagem foi utilizada afim de determinar o melhor conjunto de grupos. Uma vez obtidos os grupos, são calculadas métricas de qualidade para validação da separação espacial das conformações. O segundo caminho que o fluxograma apresentado aqui segue é a obtenção de gráficos de energia livre (*Free Energy Landscape* - FEL) utilizando o método WHAM (*Weight Histogram Analysis Method*). Ao final de todo o processo, o fluxograma retorna ao usuário os grupos obtidos e suas estruturas representativas, bem como também os perfis de energia livre obtidos.

6 RESULTADOS E DISCUSSÃO

Durante o desenvolvimento do fluxograma proposto neste trabalho, foram usadas duas proteínas para a fase de teste (PDB 1CLL e 1L2Y) e uma proteína para a fase de validação (PDB 2HB4). Após a validação do fluxograma desenvolvido, o mesmo foi aplicado em duas proteínas alvos-terapêutico para o tratamento da esquistossomose, as isoformas 1 (smNTPDase1) e 2 (smNTPDase2) da ATP-Difosfohidrolase (EC 3.6.1.5) de *Schistossoma mansoni*, às quais já têm sido estudadas por nosso grupo de pesquisa.

Os resultados relativos à fase de construção do fluxograma proposto foram publicadas em (92) e (22), e os referentes à fase de validação estão em elaboração. Nas seções a seguir apresentaremos os resultados e discussões sobre os testes realizados nas 5 proteínas citadas.

6.1 Sistema de Teste 1: 1L2Y

Para avaliar os métodos aplicados neste estudo foram usadas simulações de dinâmica molecular (MD) de proteínas com diferentes propriedades e tamanho. A primeira molécula é a mini-proteína artificial Trp-cage com 20 resíduos de aminoácidos (NLYIQWLKDGGPSSGRPPPS), a qual está depositada no bando de dados de estruturas proteicas PDB (*Protein Data Bank*) sob o código 1L2Y. Esta pequena molécula foi desenhada por (69) e é considerada uma das moléculas com padrão de auto-enovelamento mais rápido, sendo assim importante para estudos de predição da influência de diferentes fatores na estabilidade de estrutura secundária e terciária de proteínas.

6.1.1 Perfil da Dinâmica Molecular

No intuito de aumentar o número de possíveis conformações da proteína 1L2Y, além da DM à 310K, realizou-se uma simulação de DM à 510K. Antes se iniciar a seleção de conformações de simulações por DM, é essencial a análise de determinadas informações sobre a estabilidade do sistema e comportamento geral da molécula estudada. De acordo com a proposta de estudo, as propriedades a serem observadas podem variar, porém quando estamos tratando de sistemas biológicos envolvendo proteínas algumas métricas são essenciais para o entendimento do comportamento destas moléculas. Entre elas podemos citar: o raio de giro, variação do RMSD e RMSF, acessibilidade ao solvente de resíduos, mudança de estruturas secundárias e estabilidade das energias potenciais.

Os resultados referentes às análises desses parâmetros são apresentados no Apêndice 9. Verificada a estabilidade do sistema e a diversidade conformacional pretendida, iniciou-se a etapa de inspeção visual das DM no intuito de determinar um número aproximado de grupos indentificável por profissional especialista.

6.1.1.1 Inspenção Visual das Conformações

Na Figura 7 é apresentada a variação do RMSD e as estruturas representativas de cada grupos selecionado pela inspeção visual. Em 310K a estrutura sofreu poucas variações conformacionais, e desta forma observou-se a formação de apenas um grupo para 1L2Y (Figura 7(a)). Apesar de uma leve mudança na estrutura secundária por conta da movimentação da região linear (*coil* - região sem estrutura secundária definida) considerou-se apenas um grupo conformacional já que a variação do RMSD assume valores abaixo de 3.5Å, e desta forma não representa uma flutuação significativa para peptídeos.

Diferente do resultado apresentado no parágrafo anterior, a figura 7(b) mostra que na simulação em 510K há grande flutuação conformacional, sendo observado nove possíveis grupos para a proteína 1L2Y. Observamos que ao longo da simulação em alta temperatura, a proteína alterna entre estados conformacionais enovelados (com estrutura 3D bem compacta) e completamente desnaturados (descompactada). Este tipo de variação estrutural torna difícil a análise de grupos mesmo por profissional especialista, já que requer experiência e um longo período de tempo.

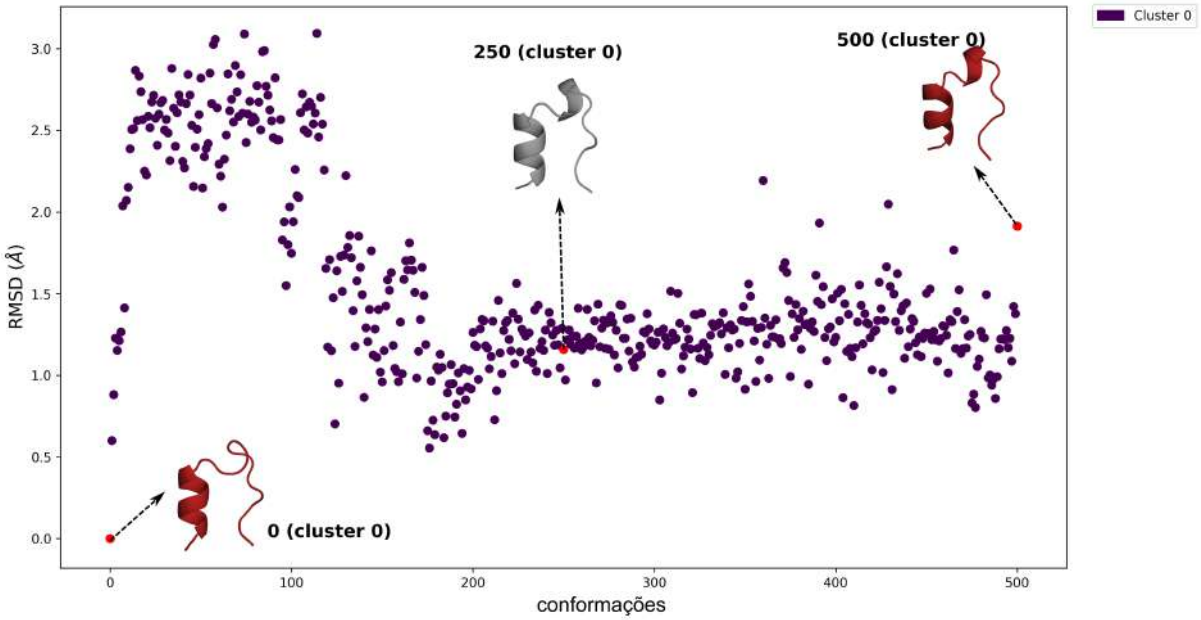
6.1.2 Testes de Agrupamento

6.1.2.1 Simulações em 310K

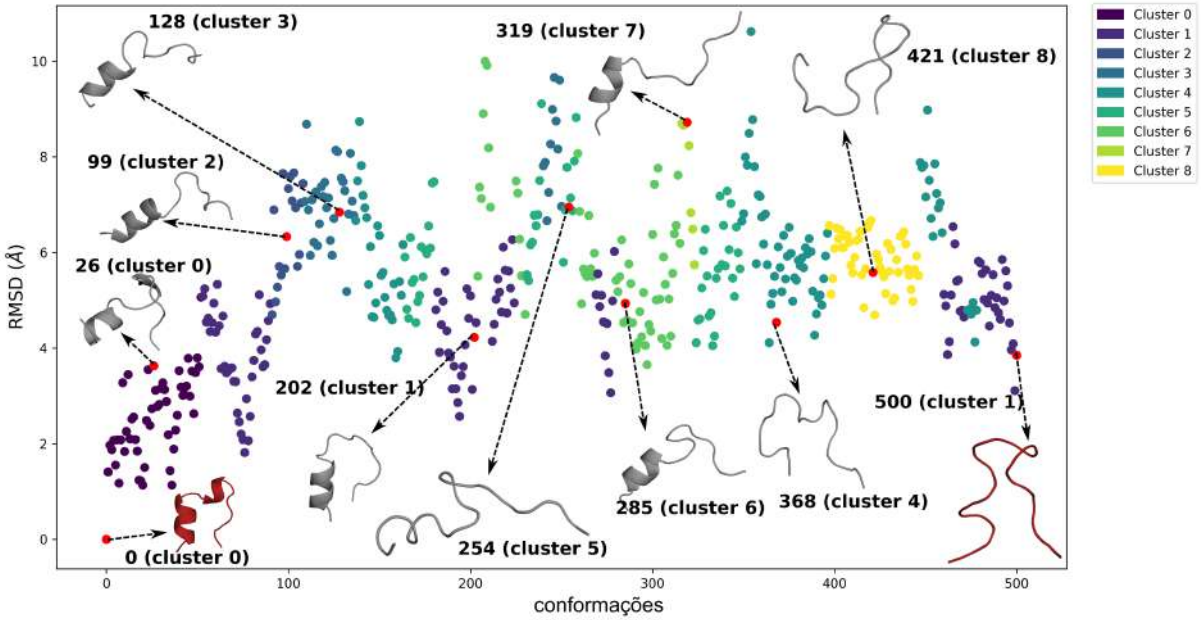
Na Tabela 2 são apresentados os valores medianos de cada métrica de qualidade para 20 melhores resultados entre as 70 possíveis combinações de métodos de redução de dimensionalidade, predição de k e agrupamento. A escolha foi realizada baseando-se nos valores de SI (*Synthetic Index*). Observamos que para este primeiro conjunto teste, os melhores resultados foram obtidos pelos redutores de dimensionalidade Isomap e Spectral. Para o primeiro observou-se que para ambos algoritmos de agrupamento a melhor predição de k foi realizada pelos métodos BIC e maximização da silhueta. Já quanto aos resultados obtidos com o método Spectral, observou-se que os três métodos de predição de k usados aqui obtiveram resultados satisfatórios quanto as métricas de qualidade.

6.1.2.2 Simulações em 510K

Assim como na análise anterior, separamos as 20 melhores combinações de métodos aplicados aos dados da simulação em 510K da proteína 1L2Y. Na tabela 3, observamos que o redutor Spectral obtém os melhores resultados, tanto para algoritmos dependentes de k quanto independentes. Contudo, embora alguns métodos tenham obtido bons valores de métricas, o número de grupos detectados foi diferente do esperado pela análise visual do conjunto de dados que detectou entre 5 e 9 possíveis principais estados. permitiu a identificação de número de grupos próximos ao esperado pela análise por especialista. Este resultado pode está associado ao fato da estrutura alterar rapidamente sua conformação



(a) 1L2Y-310K



(b) 1L2Y-510K

Figura 7 – Detecção de grupos para proteína 1L2Y pela inspeção visual. Em 7(a) são apresentados a estrutura mediana (cinza) que representa o grupo verificado e as estruturas inicial e final da simulação em 310K. Já em 7(b) estão representados os medóides (cinza) de cada um dos grupos detectados para a simulação em 510K, as conformações em vermelho são o primeiro e o último frame da trajetória.

gerando grupos com muitas estruturas que poderiam ser consideradas *outliers* para esses espaços reduzidos.

Tabela 2 – Avaliação dos algoritmos de agrupamento aplicados a simulação da 1L2Y em 310K.

Redução	Método K	algorítimo	$n\_clusters$	CH	DBI	Silhueta	SI
Isomap	Silhueta	K-means	2,0	1311,921	0,341	0,675	0,925992
Isomap	BIC	K-means	2,0	1311,921	0,341	0,675	0,925992
Isomap	Silhueta	Ward	2,0	1106,476	0,297	0,665	0,865502
Isomap	BIC	Ward	2,0	1106,476	0,297	0,665	0,865502
Spectral	GAP	K-means	3,0	808,962	0,149	0,663	0,801706
Spectral	Silhueta	K-means	3,0	808,962	0,149	0,663	0,801706
Spectral	Elbow	Ward	5,0	873,202	0,172	0,643	0,797465
Spectral	Elbow	K-means	5,0	897,297	0,190	0,635	0,794067
Spectral	Silhueta	Ward	3,0	794,588	0,163	0,653	0,786257
Spectral	GAP	Ward	3,0	794,588	0,163	0,653	0,786257
Spectral	-	Meanshift	3,0	699,500	0,147	0,645	0,754454
Spectral	-	Affinity	15,0	1471,913	0,430	0,454	0,749254
t-SNE	GAP	K-means	5,0	1117,500	0,281	0,525	0,738659
NR	Silhueta	Ward	2,0	882,061	0,597	0,627	0,724235
PCA	Silhueta	Ward	2,0	882,061	0,597	0,627	0,724235
t-SNE	GAP	Ward	5,0	1051,880	0,262	0,525	0,723059
t-SNE	BIC	K-means	2,0	859,737	0,159	0,566	0,722851
t-SNE	Silhueta	K-means	2,0	849,278	0,159	0,568	0,721822
NR	Silhueta	K-means	2,0	888,563	0,624	0,626	0,721255
PCA	Silhueta	K-means	2,0	888,563	0,624	0,626	0,721255

6.1.3 Variação Conformacional e Mapas de Energia Livre - FEL

De acordo com os resultados, para cada método aplicado na obtenção do espaço essencial das moléculas o perfil de energia obtido foi diferente dentro do mesmo *dataset* (figura 8). Isto se deve ao fato de que cada uma destas técnicas apresenta diferentes abordagens para representar um conjunto de pontos em um espaço N-dimensional euclidiano para um espaço reduzido que é o *manifold*. Este resultado foi similar ao obtido no trabalho de Duan *et. al.* (2013), onde a comparação entre LLE (Locally Linear Embeding), PCA, Isomap e diffusion maps indicou que os perfis de energia e as estruturas nos prováveis mínimos são dependentes de cada redutor (28).

Em todos os perfis apresentados na figura 9, observamos a ocorrência de agrupamentos com maior frequência de microestados (conformações) do sistema simulado. Os agrupamentos de maior frequência aparecem como bacias (mínimos) definidas a partir da equação 3.20. No caso das simulações a 310K (A, B e C), é interessante observar que os mínimos correspondem às conformações cuja estruturra secundária (SS) aparece melhor conservada, ou seja, formando principalmente alfa-hélices. Por outro lado, na simulação à 510K, certamente devido à alta temperatura, as conformações de maior frequência são aquelas onde há maior desnaturação da estrutura.

As análises estatísticas entre os diferentes redutores de dimensionalidade Isomap,

Tabela 3 – Avaliação dos algoritmos de agrupamento aplicados a simulação da 1L2Y em 510K.

Redução	Método K	algorítimo	$n\_clusters$	CH	DBI	Silhueta	SI
Spectral	BIC	K-means	14,0	1598,085	0,474	0,425	0,817894
Spectral	-	Affinity	16,0	1625,728	0,515	0,410	0,808564
Spectral	Elbow	Ward	5,0	815,264	0,175	0,561	0,780517
Spectral	BIC	Ward	8,0	1188,363	0,338	0,457	0,774460
Spectral	Silhueta	Ward	4,0	706,597	0,172	0,586	0,771687
Spectral	Elbow	K-means	4,0	833,794	0,247	0,503	0,741617
Spectral	Silhueta	K-means	2,0	335,566	0,119	0,645	0,735470
Spectral	-	Meanshift	5,0	561,851	0,215	0,572	0,727173
AE	Silhueta	Ward	2,0	381,159	0,197	0,622	0,719146
AE	Silhueta	K-means	2,0	421,142	0,230	0,610	0,715310
t-SNE	Silhueta	K-means	2,0	706,244	0,197	0,476	0,710348
t-SNE	BIC	K-means	2,0	683,018	0,198	0,476	0,705409
t-SNE	GAP	K-means	2,0	666,666	0,199	0,464	0,695678
t-SNE	Silhueta	Ward	2,0	674,793	0,211	0,460	0,693156
AE	BIC	Ward	2,0	392,855	0,275	0,582	0,687087
t-SNE	Elbow	K-means	5,0	679,427	0,265	0,444	0,676294
AE	BIC	K-means	3,0	441,213	0,301	0,547	0,674319
t-SNE	-	Affinity	17,0	857,764	0,454	0,430	0,672220
t-SNE	BIC	Ward	2,0	606,881	0,207	0,440	0,669603
t-SNE	GAP	Ward	2,0	580,707	0,207	0,437	0,662686

PCA e Spectral combinados com o algoritmo Ward e apresentados nas figuras 9, apontaram que, em 310K, quando observadas as métricas de qualidade CH, silhueta e o número de grupos não há diferença significativa entre as abordagens. Já para os valores de DBI os métodos Isomap e PCA foram similares entre si, enquanto que o Spectral apresentou diferença significativa com esses redutores. Para os testes em 510K, os métodos Isomap e Spectral não apresentaram diferenças significativas entre quando comparadas as métricas CH, DBI e silhueta ao passo que foram diferentes do PCA. Contudo, comparando-se os valores de número de grupos detectados, não foram observadas diferenças significativas entre os três métodos. Os dados completos dos testes estatísticos realizados aqui podem ser verificados no link[1].

[1] `https://drive.google.com/drive/folders/1F8PpaDkprV7AwTemHNG_fcO_uV_4G3O5?usp=sharing`

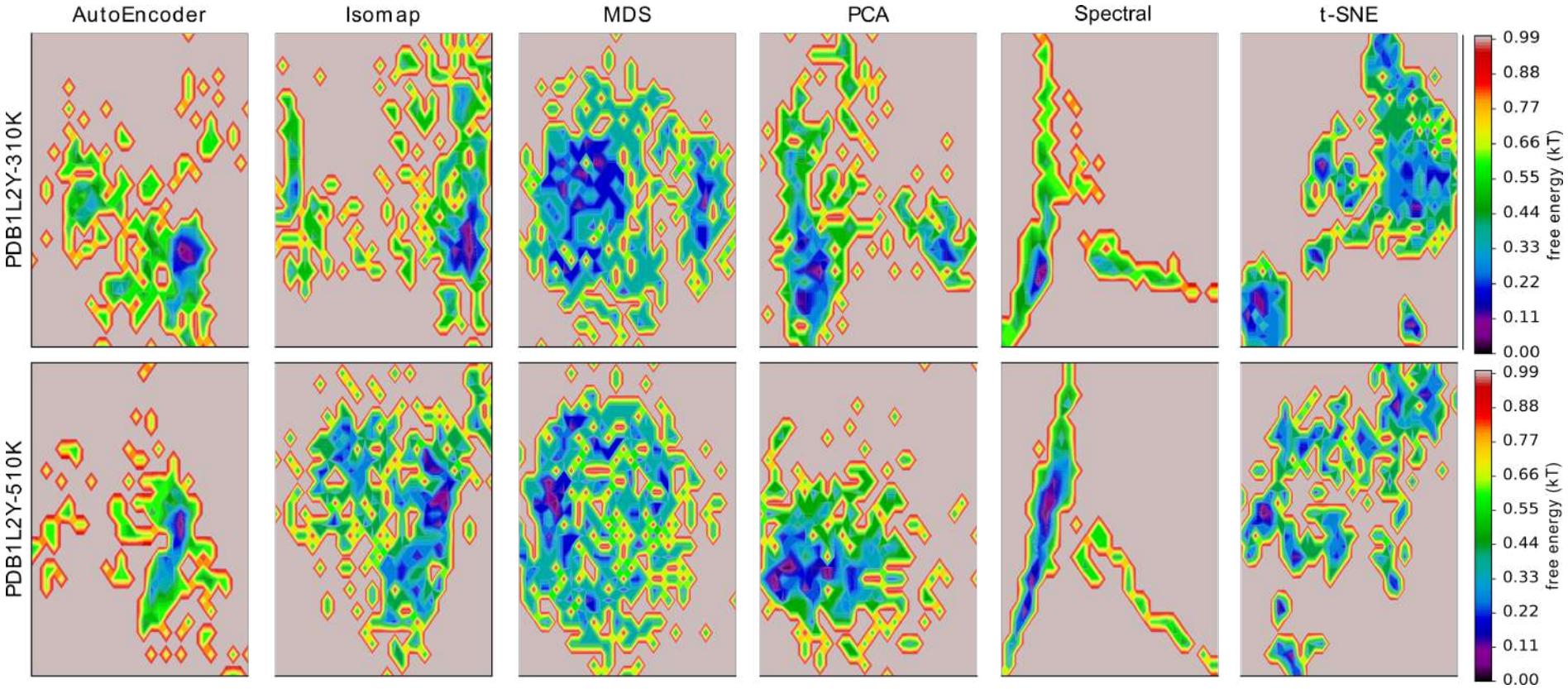


Figura 8 – Mapas de energia livre (FEL) obtidos para cada redutor de dimensionalidade para as trajeórias da proteína 1L2Y. Os mapas de energia foram obtidos usando o método WHAM com valor de bin igual 30. As regiões em roxo e azul nos mapas representam possíveis bacias de energia enquanto que regiões em vermelho são barreiras energéticas

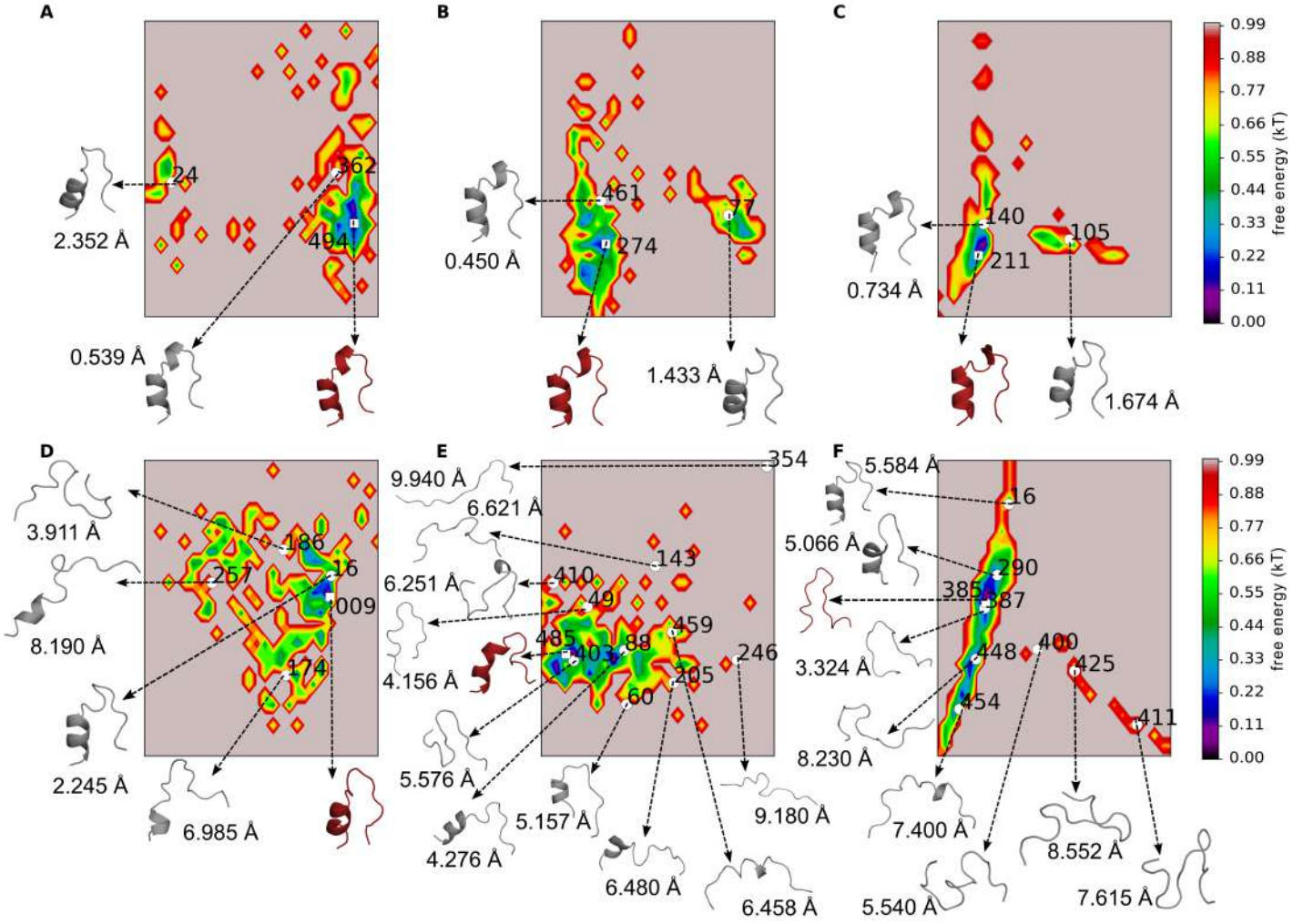


Figura 9 – Perfis de energia para 1L2Y em diferentes temperaturas. Nas Figuras estão representados os mapas de energia obtidos com os redutores Isomap, PCA e Spectral para as simulações em 310K (A, B e C) e 510K (D, E e F) da proteína 1L2Y. Os mapas foram gerados utilizando o método WHAM. As estruturas medóides (cinza) são comparadas com aquela na curva de nível de energia igual a zero (vermelho), usando-se os valores de RMSD. Os medóides dos grupos foram detectados usando o algoritmo Ward combinado com o método Elbow. As regiões em roxo e azul nos mapas representam possíveis bacias de energia enquanto que regiões em vermelho são barreiras energéticas.

6.1.4 Considerações Parciais

Esta seção apresentou os resultados das simulações da proteína 1L2Y em diferentes temperaturas. Foi mostrado os experimentos de agrupamento executados com diferentes configurações, usando como informação de similaridade a matriz de distância euclidiana entre os átomos de carbono α. Ao final foram obtidos os mapas de energia (*Free-Energy Landscape*) aplicando o método WHAM (*Weighted Histogram Analysis Method*) para cada abordagem de redução de dimensionalidade.

A partir das análises das simulações de dinâmica molecular da proteína 1L2Y, obtivemos um conjunto de testes com número suficente de estados distintos, para colocar em prova as diferentes abordagens sugeridas neste trabalho. Observamos que o método de Isomap, Spectral e t-SNE apresentaram melhores resultados em ambas simulações avaliadas. Os mesmos geraram espaços de dimensões reduzido (espaço intríseco), aos quais, quando aplicado os métodos de agrupamento obtivemos grupos de estados conformacionais com diferenças significativas entre si e com melhores valores de qualidade.

Quanto aos métodos de agrupamento, verificamos que os algoritmos K-means (combinado com BIC e a maximização da silhueta) e Ward (combinado com Elbow e BIC), foram capazes de detectar grupos com os melhores valores de métricas, nos dois conjuntos de trajetórias. Muito embora o número de grupos obtidos por esses métodos, na simulação em alta temperatura, tenha sido diferente do esperado pela análise visual, os mesmos fornceram bons *insights* de escolha de estruturas.

Quanto aos métodos independentes de k, observamos que apesar do algoritmo affinity obter bons resultados para as métricas de qualidade em ambas trajetórias, o número de grupos foi quase o dobro do esperado para simulação em 310K. Resultado oposto foi obtido pelo algoritmo meanshift, mesmo este tendo apresentado resultados bons de qualidade quando aplicado sobre o espaço gerado pelos métodos de redução citados acima.

6.2 Sistema de Teste 2: 1CLL

A calmodulina (CaM) é uma proteína responsável por capturar os íons Ca^{+2} do meio e está diretamente associada com diversos processos fisiológicos em nosso organismo, como por exemplo: sinalização celular(102), contração muscular(107) e metabolismo pela ativação de enzimas dependentes de calmodulina(70)(114). Devido esta importância, trabalhos anteriores têm proposto um estudo da dinâmica conformacional da CaM tanto em temperatura fisiológica quanto em temperaturas altas afim de analisar sua estabilidade térmica.

6.2.1 Perfil da Dinâmica Molecular

Assim como realizado para a 1L2Y, as simulações de DM para a 1CLL foram em temperatura à 310K e 510K. Os resultados referentes às análises de raio de giro, variação do RMSD e RMSF, acessibilidade ao solvente de resíduos e estabilidade das energias potenciais, são apresentados no Apêndice . Após a verificação da estabilidade do sistema e a diversidade conformacional pretendida, iniciamos a etapa de inspeção visual das DM no intuito de determinar um número aproximado de grupos indentificável por profissional especialista.

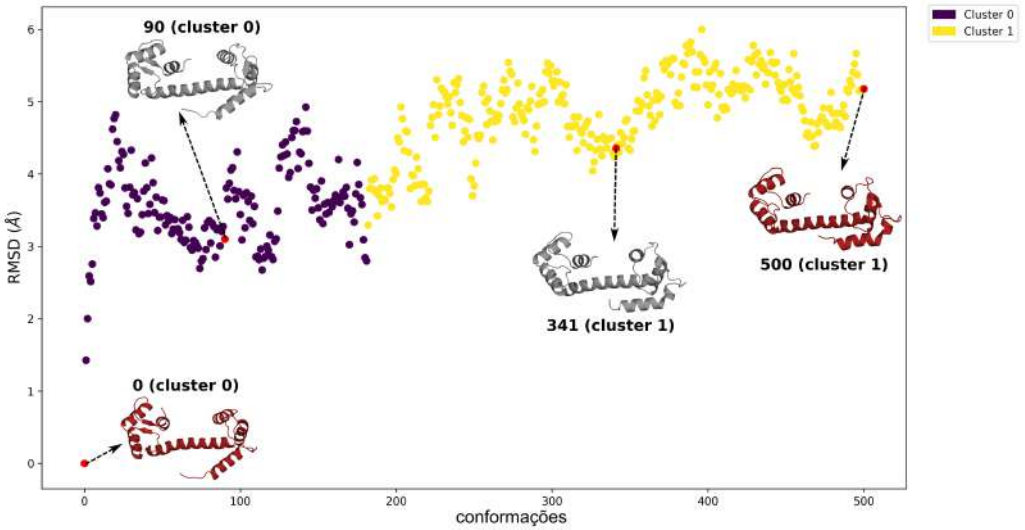
6.2.1.1 Inspeção Visual das Conformações

Assim como realizado para as análises das simulações da proteína 1L2Y, realizamos aqui uma análise visual das trajetórias obtidas para a 1CLL. Na Figura 10 é apresentada a variação do RMSD e as estruturas representativas de cada grupo selecionado pela inspeção visual por um especialista. Observou-se que em 310K a proteína sofre uma pequena movimentação na região dos lóbulos de interação com cálcio e, baseando-se nessa análise visual, detectamos dois possíveis grupos de conformações. Porém na simulação em alta temperatura, verificou-se que, conforme as métricas haviam indicado, a proteína sofre flutuações na geometria tridimensional, assumindo diferentes estados de compactação, por exemplo a conformação 500 (figura 10(b)) em que há desestruturação de um dos lóbulos. Para essa simulação, encontramos de 5 a 8 possíveis estados.

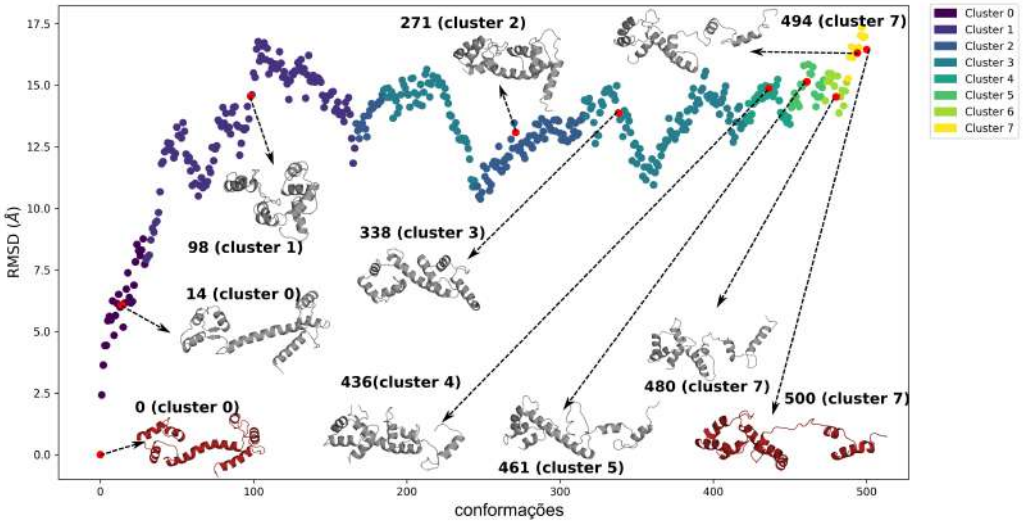
6.2.2 Testes de Agrupamento

6.2.2.1 Simulações em 310K

Na Tabela 4 são apresentados os valores de mediana para as mátricas de qualidade dos 20 melhores resultados entre as abordagens testadas aqui. A ordenação foi realizada baseando-se nos valores de SI, para os dados referentes a simulação em 310K da proteína 1CLL. Observamos que os melhores resultados foram obtidos quando aplicados os redutores de dimensionalidade Isomap e Spectral. Em geral, as abordagens detectaram entre valores



(a) 1CLL-310K



(b) 1CLL-510K

Figura 10 – Detecção de grupos para proteína 1CLL pela inspeção visual. Em 10(a) são apresentados a estrutura mediana (cinza) que representa o grupo verificado e as estruturas inicial e final da simulação em 310K. Já em 10(b) estão representados os medóides (cinza) de cada um dos grupos detectados para a simulação em 510K, as conformações em vermelho são o primeiro e o último frame da trajetória.

medianos de número de grupos entre 2 e 4, próximo ao obsrvado pela análise visual deste conjunto de trajetórias que foi de 2 possíveis grupos de estados conformacionais, embora o método affinity foi capaz de separar 9 ou 12 possíveis conjuntos de acordo com o redutor usado.

6.2.2.2 Simulações em 510K

Para as trajetórias da simulação à 510K da proteína 1CLL, a análise visual permitiu a detecção de 5 a 8 grupos conformacionais. Contudo, as diferentes abordagens avaliadas

Tabela 4 – Avaliação dos algoritmos de agrupamento aplicados a simulação da 1CLL em 310K.

Redução	Método K	algorítimo	$n\_clusters$	CH	DBI	Silhueta	SI
Isomap	Elbow	Ward	4,0	2843,715	0,135	0,678	0,904490
Isomap	GAP	Ward	4,0	2843,715	0,135	0,678	0,904490
Isomap	BIC	Ward	4,0	2843,715	0,135	0,678	0,904490
Isomap	-	Meanshift	4,0	2874,473	0,142	0,679	0,904475
Isomap	Elbow	K-means	4,0	2880,141	0,143	0,676	0,903148
Isomap	GAP	K-means	4,0	2880,141	0,143	0,676	0,903148
Isomap	BIC	K-means	2,0	2269,179	0,090	0,716	0,887833
Isomap	Silhueta	K-means	2,0	2269,179	0,090	0,716	0,887833
Isomap	Silhueta	Ward	2,0	2020,752	0,089	0,704	0,858224
Spectral	BIC	K-means	3,0	1745,131	0,139	0,722	0,814135
Spectral	Silhueta	K-means	3,0	1745,131	0,139	0,722	0,814135
Spectral	-	Meanshift	3,0	1721,580	0,135	0,719	0,812432
Spectral	Silhueta	Ward	3,0	1694,977	0,139	0,722	0,809174
Spectral	BIC	Ward	3,0	1694,977	0,139	0,722	0,809174
Spectral	Elbow	K-means	4,0	2139,463	0,221	0,687	0,795752
Spectral	Elbow	Ward	4,0	2102,417	0,222	0,686	0,791123
Isomap	-	Affinity	9,0	3082,933	0,347	0,527	0,751851
Spectral	-	Affinity	12,0	3370,153	0,470	0,522	0,716110
PCA	BIC	K-means	2,0	1362,214	0,198	0,655	0,715666
PCA	Silhueta	K-means	2,0	1362,214	0,198	0,655	0,715666

aqui detectaram em sua maioria valores entre 2 e 5 grupos. De acordo com a avaliação das métricas de qualidade e valores de SI (5), os melhores redutores foram Isomap, Spectral e PCA tanto para métodos dependentes quanto independentes de k. Porém, considerando também o número de grupos detectados, as melhores combinações foram: Isomap + Affinity, com valor mediano de 13 grupos; Isomap + GAP + K-means, com 8 grupos; e Spectral + Affinity, com mediana de 11 grupos.

6.2.3 Variação Conformacional e Mapas de Energia Livre - FEL

Na figura 12, assim como para a proteína 1L2Y, observamos que na simulação com calmodulina (1CLL) em todos os perfis apresentados observamos a ocorrência de agrupamentos com maior frequência de conformações obtidas do sistema simulado. Aqui é interessante observar que mesmo na simulação à 510K, as conformações apresentam uma alta conservação da SS. Certamente por se tratar de um proteína que tolera temperaturas mais altas (termo-estável). Aqui os agrupamentos formam mínimos bem definidos, principalmente com o PCA e Spectral. Uma proteína termo-estável apresenta menor variação conformacional de tal modo que um redutor linear consegue transformar boa parte da variação do sistema.

As análises estatísticas entre os diferentes redutores de dimensionalidade Isomap, PCA e Spectral combinados com o algoritmo Ward e apresentados nas figuras 12, apontaram

Tabela 5 – Avaliação dos algoritmos de agrupamento aplicados a simulação da 1CLL em 510K.

Redução	Método K	algorítimo	$n\_clusters$	CH	DBI	Silhueta	SI
Isomap	Silhueta	K-means	2,0	1516,868	0,077	0,666	0,832646
Isomap	BIC	K-means	2,0	1516,868	0,077	0,666	0,832646
Isomap	-	Affinity	13,0	3046,291	0,274	0,520	0,829209
Isomap	Silhueta	Ward	2,0	1458,071	0,077	0,665	0,825712
Isomap	BIC	Ward	2,0	1458,071	0,077	0,665	0,825712
Isomap	GAP	K-means	8,0	2214,711	0,265	0,512	0,738675
Spectral	Silhueta	K-means	3,0	972,493	0,151	0,632	0,719356
Isomap	Elbow	K-means	4,0	1679,513	0,183	0,508	0,718785
Spectral	Elbow	K-means	4,0	1290,072	0,212	0,621	0,718343
Spectral	-	Meanshift	3,0	971,975	0,157	0,632	0,716323
Spectral	Silhueta	Ward	3,0	943,136	0,151	0,628	0,714142
Spectral	Elbow	Ward	4,0	1222,049	0,207	0,611	0,708375
Isomap	GAP	Ward	8,0	1931,408	0,244	0,484	0,704078
Isomap	Elbow	Ward	3,0	1234,046	0,133	0,473	0,677325
Isomap	-	Meanshift	4,0	1141,474	0,138	0,497	0,676727
Spectral	-	Affinity	11,0	2027,808	0,380	0,511	0,660680
PCA	Silhueta	K-means	5,0	1058,335	0,275	0,585	0,643718
PCA	BIC	K-means	5,0	1058,335	0,275	0,585	0,643718
PCA	Elbow	K-means	5,0	1058,335	0,275	0,585	0,643718
PCA	Elbow	Ward	5,0	949,101	0,264	0,575	0,632216

que, em 310K, quando observadas as métricas de qualidade CH não há diferença significativa entre os redutores Isomap, PCA e Spectral. Para os valores de DBI, silhueta e número de grupos, apenas os métodos Isomap e PCA foram similares entre si. Para os testes em 510K, os três métodos de redução não apresentaram diferenças significativas quando avaliadas as métricas CH, silhueta e número de grupos. Contudo, comparando-se os valores de DBI, observou-se que o método Spectral foi significativamente diferente do Isomap e PCA. Os dados completos dos testes estatísticos realizados aqui podem ser verificados no link[2].

[2] `https://drive.google.com/drive/folders/1F8PpaDkprV7AwTemHNG_fcO_uV_4G3O5?usp=sharing`

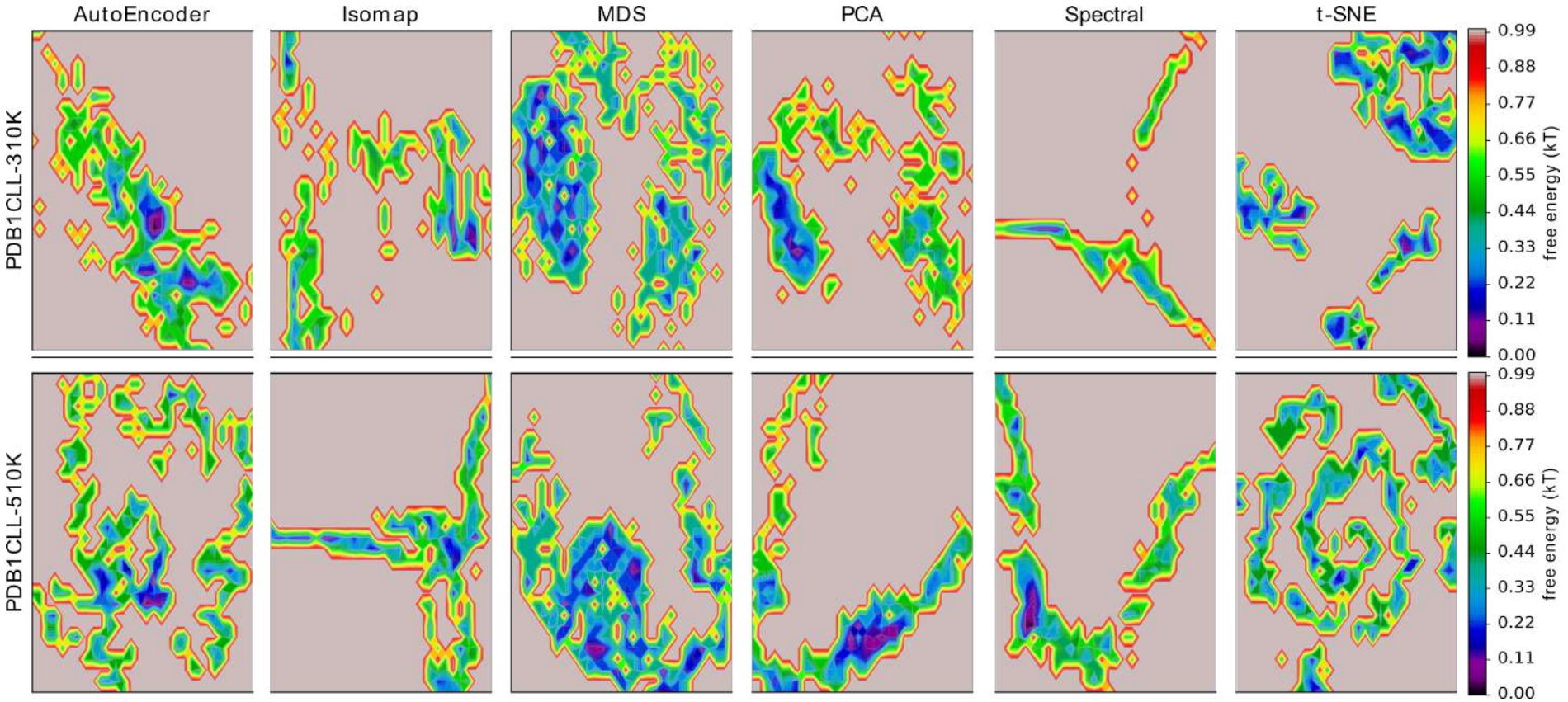


Figura 11 – Mapas de energia livre (FEL) obtidos para cada redutor de dimensionalidade para as trajetórias da proteína 1CLL. Os mapas de energia foram obtidos usando o método WHAM com valor de bin igual 30. As regiões em roxo e azul nos mapas representam possíveis bacias de energia enquanto que regiões próxima de vermelho são regiões de barreiras energéticas

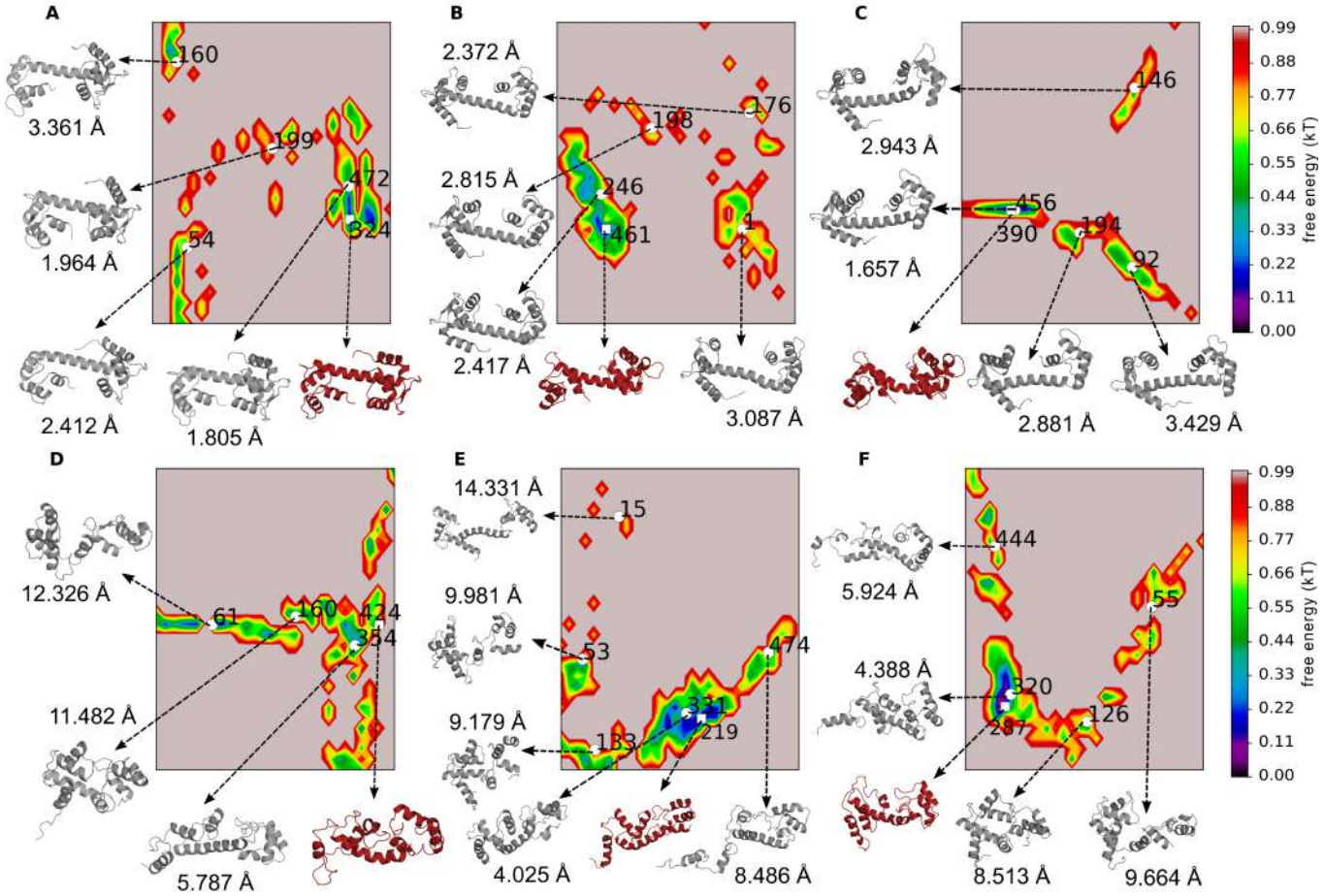


Figura 12 – Perfis de energia para 1CLL em diferentes temperaturas. Nas Figuras estão representados os mapas de energia obtidos com os redutores Isomap, PCA e Spectral para as simulações em 310K (A, B e C) e 510K (D, E e F) da proteína 1CLL. Os mapas foram gerados utilizando o método WHAM. As estruturas medóides (cinza) são comparadas com aquela na curva de nível de energia igual a zero (vermelho), usando-se os valores de RMSD. Os medóides dos grupos foram detectados usando o algoritmo Ward combinado com o método Elbow. As regiões em roxo e azul nos mapas representam possíveis bacias de energia enquanto que regiões próxima de vermelho são regiões de barreiras energéticas.

6.2.4 Conclusões Parciais

Esta seção apresentou os resultados das simulações da proteína 1CLL em diferentes temperaturas. Foi mostrado os experimentos de agrupamento executados com diferentes configurações, usando como informação de similaridade a matriz de distância euclidiana entre os átomos de carbono α. E ao final foram obtidos os mapas de energia (*Free-Energy Landscape*) aplicando o método WHAM (*Weighted Histogram Analysis Method*) para cada abordagem de redução de dimensionalidade.

A partir das análises das simulações de dinâmica molecular da 1CLL, observamos que esta proteína apresentou uma maior estabilidade quando comparada a 1L2Y. Além disso, as mudanças de estados de compactação e descompactação ocorreram de forma mais lenta do que observado para a proteína anterior. Devido a este perfil diferente de comportamento nas diferentes simulações, e por ser uma proteína com maior número de átomos, representaou bem um segundo conjunto de dados para teste dos métodos de agrupamento usados neste trabalho.

Observamos que, assim como para o primeiro conjunto de testes, os métodos de Isomap e Spectral apresentaram resultados satisfatórios em ambas conjuntos de trajetórias que compõem o sistema de testes 1CLL. No entanto, diferente das análises para 1L2Y, o método PCA também obteve bons resultados aqui. Contudo, embora os métodos de agrupamentos tenham obtido boas métricas para esses redutores, observamos que em geral o número de grupos na simulação em 510K foi abaixo do esperado.

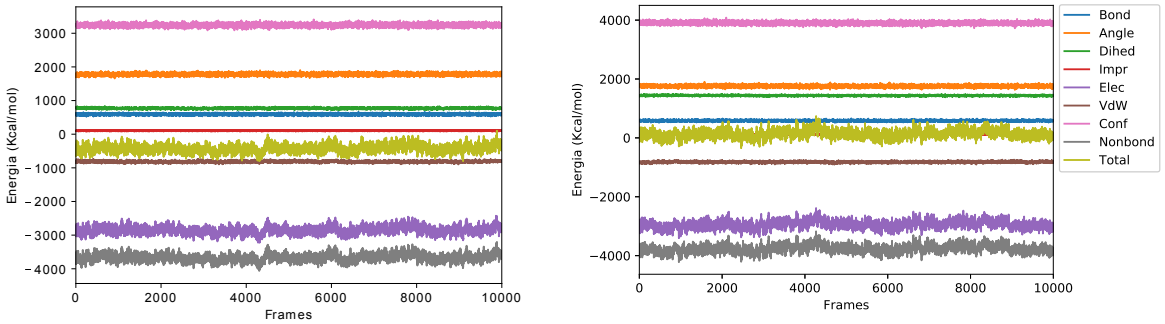
Assim como verificado para o conjunto de dados da 1L2Y, obervamos que as mesmas combinações dos métodos de agrupamento K-means (combinado com BIC ou maximização da silhueta) e Ward (combinado com Elbow ou BIC) detectaram os melhores conjuntos de estados de acordo com os valores de SI.

6.3 Sistema de Validação: Protease do HIV-1

A protease do HIV-1 é uma enzima com 99 aminoácidos que atua na hidrólise de ligações peptídicas em sítios de clivagem das poliproteínas *Gag* e *Gag-pol*, gerando assim componentes proteicos essenciais para a formação da *Virion* do HIV (forma infecciosa do vírus fora da célula-hospedeira). Esta enzima funciona como um homodímero e seu sítio ativo está localizado em uma cavidade central coberta por duas regiões de "*flaps*" flexíveis. Essas regiões consistem em duas alças expostas ao solvente (resíduos de 33 a 43 de cada cadeia) e duas regiões ricas em glicina (resíduos de 44 a 62 de cada cadeia) que são importantes para a ligação do substrato. De acordo com as regiões flexíveis do *flap*, o dímero da protease do HIV-1 pode apresentar três conformações diferentes envolvidas na ligação ou liberação do substrato: (i) ambos *flaps* estão abertos; (ii) uma subunidade possui uma conformação na qual o *flap* está fechado, enquanto o outro *flap* está aberto; e (iii) ambos os *flaps* estão fechados. Devido estes estados distintos, essa enzima tem um valor importante na validação de métodos para detecção automática de conformações significativas a partir de simulações de dinâmica molecular.

6.3.1 Perfil da Dinâmica Molecular

Nesse trabalho realizamos dois tipos de simulações por DM, a chamada simulação convencional (cDM) e a simulação acelerada (aDM), ambas apresentadas no capítulo 4. O objetivo foi o de explorar mais regiões da superfície de energia desta proteína na tentativa de verificar ao menos a variação entre os diferentes estados conformacionais citados acima.

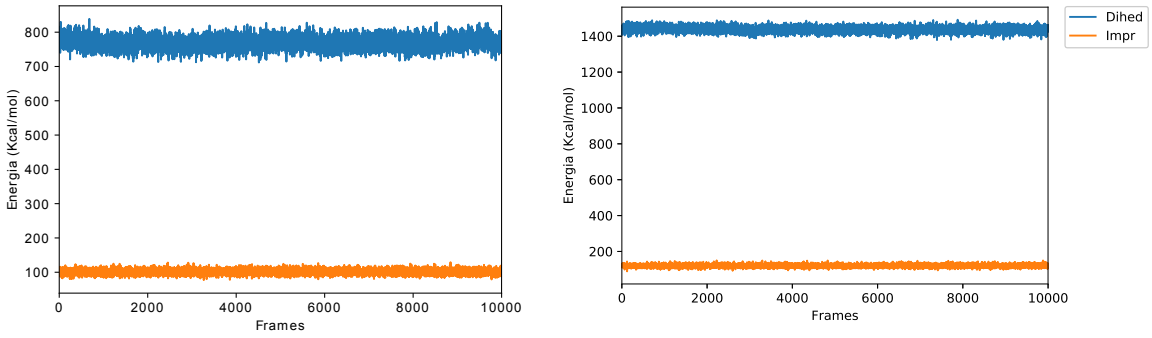


(a) HIV-cMD (b) HIV-aMD

Figura 13 – Flutuação das energias potenciais para simulações da protease do HIV-1. As energias calculadas com o plugin NAMD-energy dentro do programa VMD [3].

Na figura 13 e 14 são mostrados os valores de energia para cada um dos conjuntos de trajetórias da proteína 2HB4. Os resultados apontam que a proteína tanto na simulação convencional quanto acelerada apresenta-se equilibrada, não havendo grandes flutuações que indicariam provável instabilidade ou grandes flutuações conformacionais. Além disso,

como esperado, com a adição do potencial extra nas energias diedrais, observamos que os valores das energias são maiores nas simulações aMD.



(a) HIV-cMD (b) HIV-aMD

Figura 14 – Flutuação das energias potenciais para simulações da protease do HIV-1. As energias calculadas com o plugin NAMD-energy dentro do programa VMD [4].

Na figura 15 é apresentado uma análise comparativa do valores de RMSD entre as simulações de DM convencional e acelerada para a enzima protease do HIV-1. Observamos que após a adição do potêncial de *boost* aos ângulos diedrais, o valor médio de RMSD muda de 2.32Å para 4.56Å. Este resultado indica que a enzima foi capaz de assumir mais conformações, uma vez que as simulações aDM permitem que a proteína explore diferentes regiões de sua superfície energética.

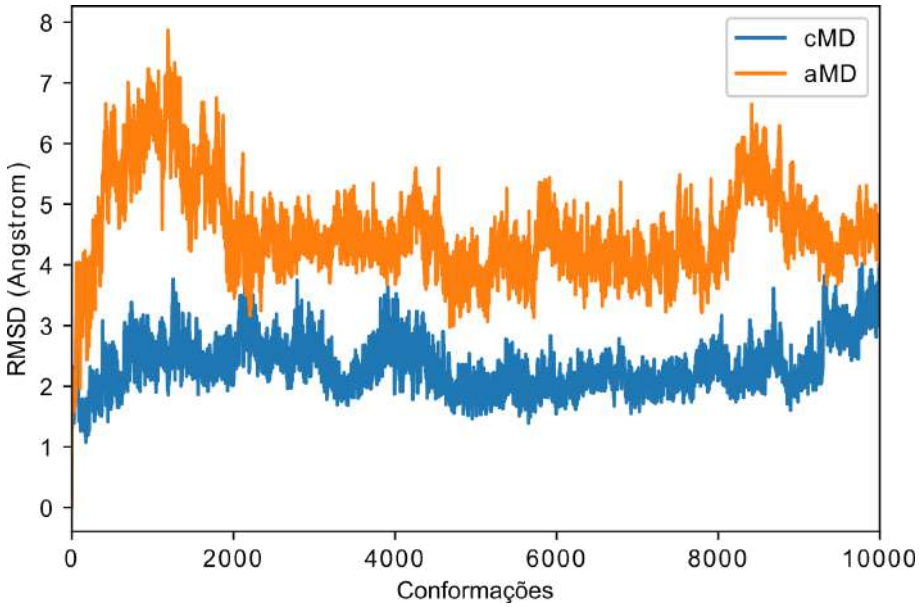


Figura 15 – Flutuação dos valores de RMSD nas simulações cMD e aMD para 2HB4.

Assim, para verificar quais resíduos ou regiões sofreram flutuações estruturais, o RMSF de duas simulações é comparado na figura 16. Comparando-se as duas simulações, percebemos que ocorre uma translação vertical dos valores do gráfico referente a simulação com adição do potencial. Foi observado também que os resíduos 50 e 149 (resíduo 50

da cadeia B) exibem maior flutuação na simulação aDM do que cDM, indicando que a amplitude de movimento das regiões de *flaps* flexíveis aumentou após a aplicação do potencial de impulso (*boost*).

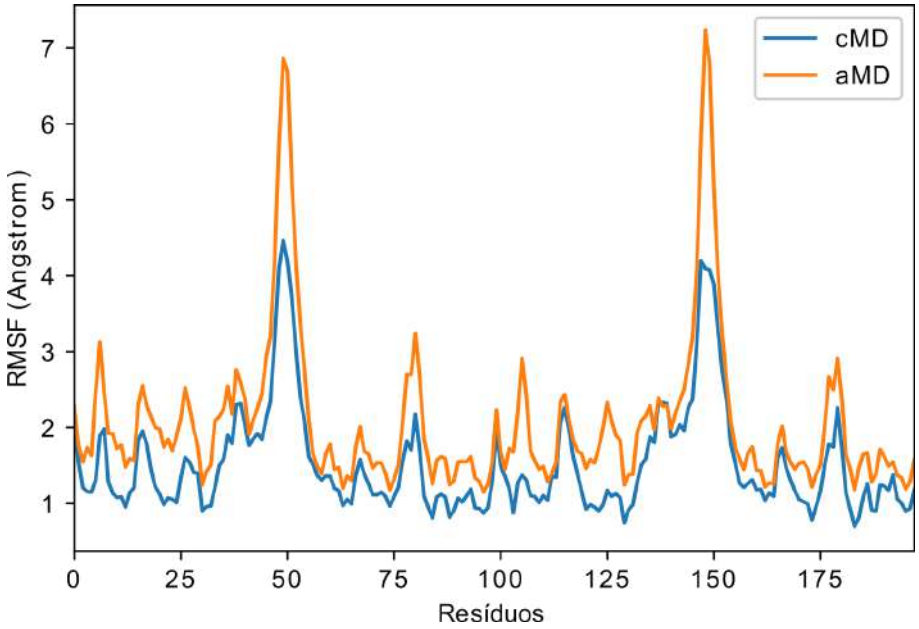


Figura 16 – Flutuação dos valores de RMSF nas simulações cMD e aMD para 2HB4.

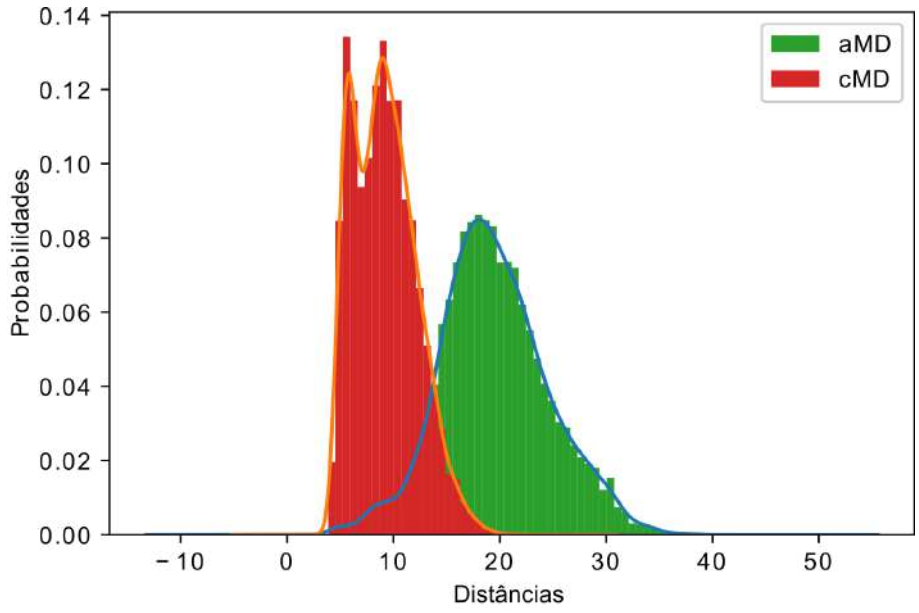


Figura 17 – Distribuição dos valores de distância entre os átomos de carbono α dos resíduos I50A e I50B nas simulações cMD e aMD para 2HB4.

Conforme apresentado na figura 17, observamos que durante a simulação convencional a distância entre os Cα dos resíduos I50A e I50B foi em média 9,16Å com descio padrão de 2,94Å. Estes valores estão de acordo com aqueles reportados para o estado semi-aberto (99), sugerindo que esta conformação possui maior representatividade no

conjunto de trajetórias obtidas pela simulação sem adição do potencial diedral. Contudo, quando analisada a simulação acelerada, o valor médio e desvio-padrão das distâncias entre os mesmos átomos foram, respectivamente, 21,82Å e 5,1Å, indicando que tal estado foi predominantemente visitado (99).

6.3.2 Testes de Agrupamento

6.3.2.1 Agrupamento Baseado em *Cutoff* Definido pelo Usuário

Utilizamos os algoritmos Ward e Elbow para predizer os possíveis grupos nas dinâmica cDM e aDM, e como métodos de redução foram usados os métodos Isomap, Spectral e t-SNE. A escolha destas combinações foram baseadas nos resultados obtidos com os sistemas de teste 1L2Y e 1CLL. O desempenho das mesmas foi comparado com os resultados obtidos quando aplica-se o PCA ou nenhum método redução.

Além disso, diferente do realizado para os sistemas 1L2Y e 1CLL, para o presente sistema de validação não foram usadas as matrizes de distâncias euclidianas (EDM) entre todos os átomos de carbono α intramoleculares. Uma vez que aumentamos o tamanho das proteínas, em número de aminoácidos, consequentemente o tamanho das matrizes aumenta, elevando o custo computacional (uso de memória RAM - do inglês, *Random Access Memory*). Por tanto, baseado-se nos valores de RMSF, foram usados apenas resíduos com flutuação acima de um *cutoff* igual a 5Å e assim obter uma matriz de contatos.

6.3.2.1.1 Simulação Convencional

Na tabela 6 são apresentados os resultados de qualidade dos agrupamentos obtidos com as diferentes abordagens testadas aqui. Usando como referência os valores de SI, observamos que os grupos detectados pelos algoritmos K-means e Ward combinados com o método Elbow tiveram melhores valores de qualidade quando aplicado o redutor Spectral. Observamos ainda que no geral o valor mediano de grupos detectados foi próximo a quantidade de estados descritos na literatura, que seria entre três e quatro.

6.3.2.1.2 Simulação Acelerada

Assim como observado nos testes realizados com a dinâmica molecular convencional, os algoritmos K-means e Ward obtiveram novamente os melhores resultados de qualidade aplicando o método Spectral (tabela 7). Além disso, o número de grupos foi próximo ao valor esperado. Contudo, comparativamente com os resultados anteriores, os valores de SI foram menores, isto pode está associado ao fato do espaço intríseco obtido apresentar-se mais compacto, ou seja, os pontos apresentam-se espacialmente muito próximos dificultando a detecção de grupos.

Tabela 6 – Avaliação dos algoritmos de agrupamento aplicados sobre a simulação convencional da 2HB4.

Redução	Algoritimo	*n_clusters*	CH	DBI	Silhueta	SI
Spectral	K-means	4,0	75109,843	0,794	0,902	0,820388
Spectral	Ward	4,0	71288,154	0,806	0,900	0,797275
Isomap	K-means	5,0	8292,120	0,551	0,471	0,422494
t-SNE	K-means	6,0	4439,539	0,350	0,304	0,411679
Isomap	Ward	5,0	7838,241	0,560	0,452	0,408302
t-SNE	Ward	6,0	3664,162	0,392	0,270	0,375974
PCA	K-means	5,0	5013,539	0,984	0,330	0,169432
PCA	Ward	4,0	4480,104	0,955	0,289	0,160420
NR	K-means	6,0	2044,198	1,063	0,183	0,058298
NR	Ward	5,0	1751,232	1,174	0,156	0,000000

Tabela 7 – Avaliação dos algoritmos de agrupamento aplicados sobre a simulação acelerada da 2HB4.

Redução	algorítimo	*n_clusters*	CH	DBI	Silhueta	SI
Spectral	K-means	5,0	17655,405	1,769	0,573	0,672161
Spectral	Ward	5,0	14055,648	1,793	0,519	0,549359
t-SNE	K-means	6,0	4405,839	0,337	0,304	0,513134
Isomap	K-means	6,0	5020,336	0,749	0,347	0,465332
t-SNE	Ward	6,0	3473,456	0,367	0,249	0,443714
Isomap	Ward	5,0	4103,503	0,731	0,303	0,415853
PCA	K-means	5,0	4557,316	0,972	0,334	0,394442
PCA	Ward	5,0	3768,558	1,049	0,307	0,339214
NR	K-means	5,0	2207,254	1,101	0,204	0,214015
NR	Ward	6,0	1646,879	1,123	0,148	0,153388

6.3.2.2 Agrupamento Baseado em *Cutoff* Automatizado

Conforme exposto anteriormente, o uso de EDM em simulações longas para proteínas com muitos resíduos de aminoácidos pode ocasionar em um alto gasto de memória RAM. Para isso, uma abordagem focada apenas em resíduos acima de determinado valor de *cut-off* de RMSF foi desenvolvida nos testes anteriores. Porém, esta abordagem pode enviezar os resultados a análise de regiões com muita flutuação e de pouco interesse biológico. Desta forma, uma segunda abordagem foi gerada, na qual se obtém uma distribuição dos valores de RMSF e partir desta foram selecionados apenas os resíduos dentro do intervalo $[\mu - \sigma, \mu + \sigma]$, onde μ é a média e σ o desvio padrão.

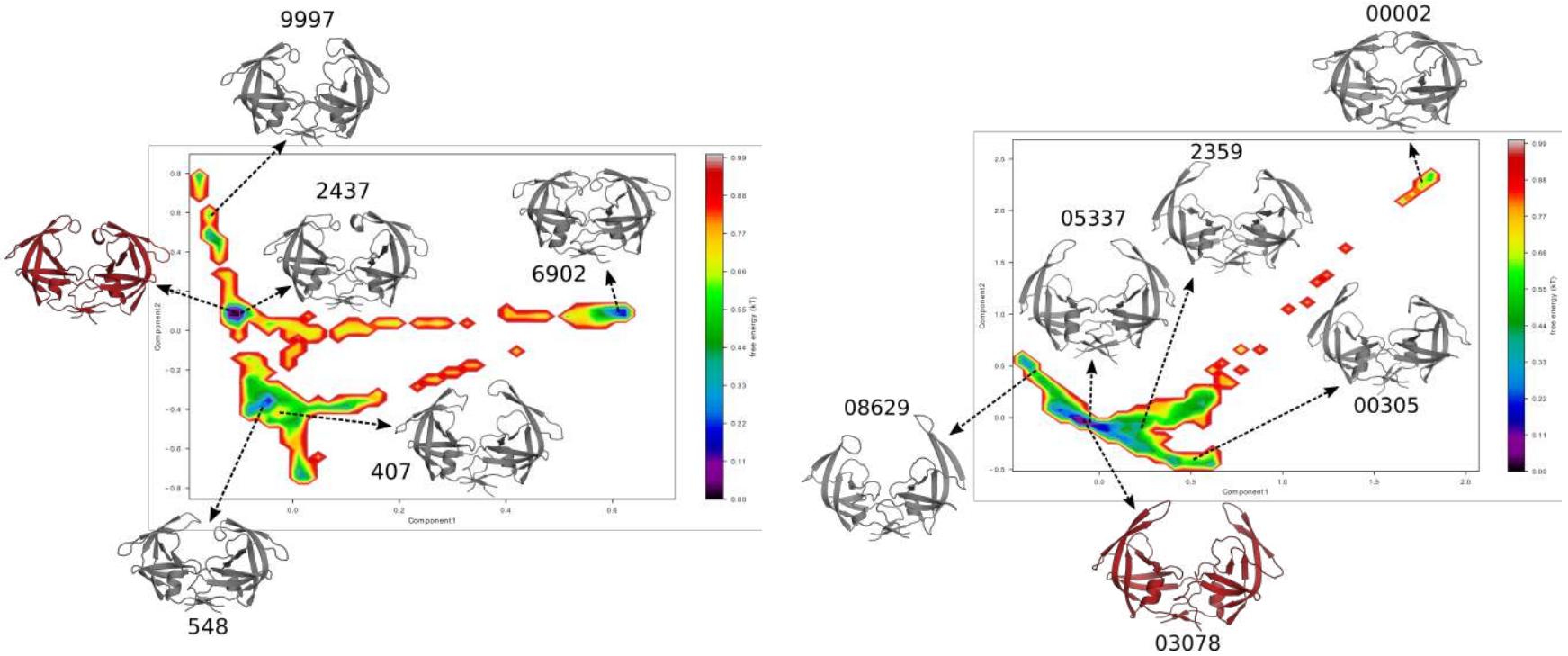
6.3.3 Variação Conformacional e Mapas de Energia Livre - FEL

Na figura 18 são apresentados os mapas de energia das conformações medianas de cada grupo, juntamente com a estrutura considerada no mínimo de energia. Comparando-se os mapas das simulações cMD e aMD, observamos que na primeira simulação há a ocorrência de dois mínimos bem definidos, sendo um deles referentes às estruturas fechadas

e o outro àquelas na forma semi-aberta. Enquanto na aMD, há apenas um mínimo, o qual se refere às estruturas abertas.

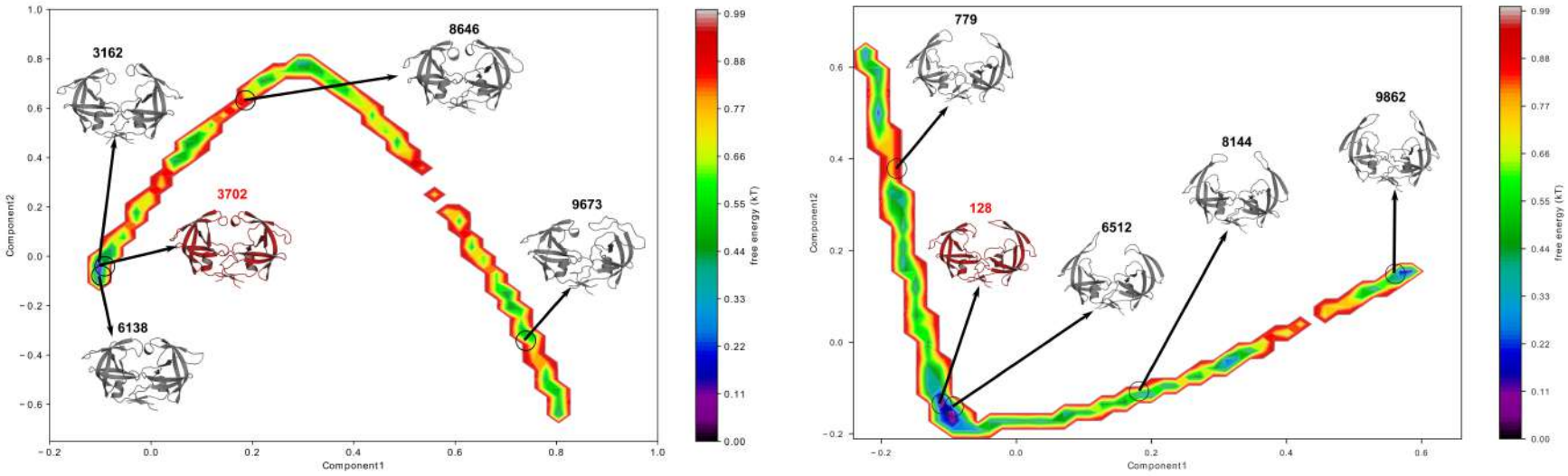
A diferença entre os perfis obtidos, pode ser explicada essencialmente pela amostragem do sistema. Conforme foi mostrado na figura 17, na simulação convencional a enzima assume preferencialmente estados fechado e semi-aberto . Assim, ao se obter os mapas de energia usando histogramas, essas conformações aparecem em mínimos devido a maior probabilidade de ocorrência. Enquanto que na simulação acelerada a enzima assume maior quantidade de estados abertos.

Já na figura 19 são apresentados os mapas de energia obtidos a partir das análises usando o *cutoff* automatizado para seleção de resíduos de interesse. Observamos que diferente da abordagem anterior, foram detectados quatro grupos distintos de conformações. Além disso, as bacias de energia foram menores e menos evidentes. Contudo, apesar do número menor de grupos detectados, os estados medóides representam bem a mudança que a proteína sofre ao longo das trajetórias. As diferenças observadas foram devido ao fato de que na primeira abordagem pegamos preferencialmente os resíduos que compõem os flaps, e na segunda aqueles que representam regiões mais estáveis da enzima.



(a) (b)

Figura 18 – Perfis de energia para 2HB4 gerados a partir do *cutoff* manual de RMSF. Nas Figuras estão representados os mapas de energia obtidos com o redutor Spectral para as simulações cMD ((a)) e aMD ((b)). Os mapas foram gerados utilizando o método WHAM. As estruturas medóides (cinza) são comparadas com aquela na curva de nível de energia igual a zero (vermelho), usando-se os valores de RMSD. Os medóides dos grupos foram detectados usando o algoritmo Ward combinado com o método Elbow. As regiões em roxo e azul nos mapas representam possíveis bacias de energia enquanto que regiões próxima de vermelho são regiões de barreiras energéticas.



(a) (b)

Figura 19 – Perfis de energia para 2HB4 gerados a partir do *cutoff* automatizado de RMSF. Nas Figuras estão representados os mapas de energia obtidos com o redutor Spectral para as simulações cMD ((a)) e aMD ((b)). Os mapas foram gerados utilizando o método WHAM. As estruturas medóides (cinza) são comparadas com aquela na curva de nível de energia igual a zero (vermelho), usando-se os valores de RMSD. Os medóides dos grupos foram detectados usando o algoritmo Ward combinado com o método Elbow. As regiões em roxo e azul nos mapas representam possíveis bacias de energia enquanto que regiões próxima de vermelho são regiões de barreiras energéticas.

6.3.4 Conclusões Parciais

De acordo com os resultados obtidos, observamos que a aplicação do redutor Spectral mostrou-se capaz de gerar um espaço intrínseco, no qual quando aplicado métodos de agrupamento, a separação dos grupos mostrou-se qualitativamente melhor analisando-se as métricas Calinski-Harabaz, Davis-Bouldin Index e Silhueta. Além disso, quando verificadas as conformações selecionadas como medóides (representantes de cada grupo) e mínimos de energia, foi observado que os mesmos foram bem distintos entre si e representativos dos estados esperados de acordo com a literatura. Desta forma, esta abordagem será usada para os passos seguintes desse trabalho.

6.4 Sistema de Estudo de Caso 1: smNTPDase 1

Conforme ja foi mencionado, uma das motivações deste trabalho foi analisar as simulações por dinâmica molecular das enzimas Nucleosídeo Trifosfato Difosfohidrolases (NTPdases) do parasito *Schsitosoma mansoni*, agente etiológico da doença esquistossomose (popularmente conhecida como barriga d'água). Estudos anteriores sugerem que estas enzimas estão envolvidas no processo de invasão do organismo hospedeiro e enfraquecimento de sua resposta imunológica, o que indica que as NTPDases possuem um importante potencial farmacológico no tratamento da esquistossomose (20)(37). Desta forma, serão apresentados nesta sessão os resultados obtidos refentes as simulações, seleção de conformações significativas e *docking* proteína-ligante com um dos compostos já testados por nosso grupo de pesquisa (76).

6.4.1 Análise Agrupamento

Para os testes de agrupamento para as simulações realizadas para a enzima smNTPDase1, foi usada a abordagem baseada na média e desvio-padrão do RMSF para detectar os resíduos aminoácidos de interesse e assim obter as matrizes de contato. Uma vez que a combinação de métodos Spectral + Elbow + Ward obtiveram resultados satisfatórios nos testes anteriores, os mesmos foram usados aqui. Além disso, utilizou-se 6.025 frames dos 120.500 referentes a 250 nanossegundos de simulação convencional e todos os 250.00 frames das simulações aceleradas.

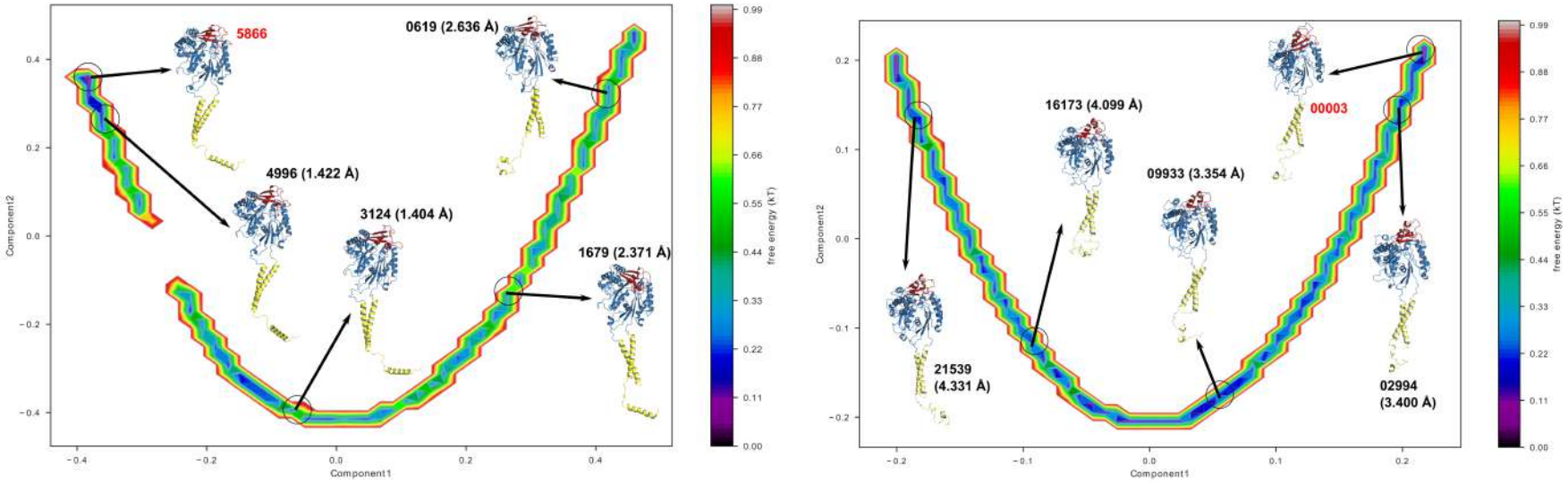
Na tabela 8 são apresentados os resultados das métricas de qualidade dos agrupamentos obtidos para cada uma das simulações da enzima smNTPDase 1. Observamos que no geral, as simulações sem a presença do ligante ANP apresentaram quatro grupos, enquanto que nas simulações considerando esta molécula foram encontrados 5 grupos. Este comportamento parece estar associado ao movimento das hélices transmembranares, as quais, de acordo com as variações de RMSD (figuras .1.2 e .2.2) e raio de giro (figuras .1.3 e .2.3), apresentam uma flutuação maior na presença de ANP.

Tabela 8 – Avaliação dos algoritmos de agrupamento aplicados sobre as simulações da enzima smNTPDase 1.

Simulação	$n\_clusters$	CH	DBI	Silhueta
sm1-cMD	4,0	8958,7202	1,1321	0,6087
sm1-aMD	4,0	29665,824	1,680	0,541
sm1_anp-cMD	5,0	10109,2225	1,6957	0,6340
sm1_anp-aMD	5,0	34274,9477	2,1404	0,5491

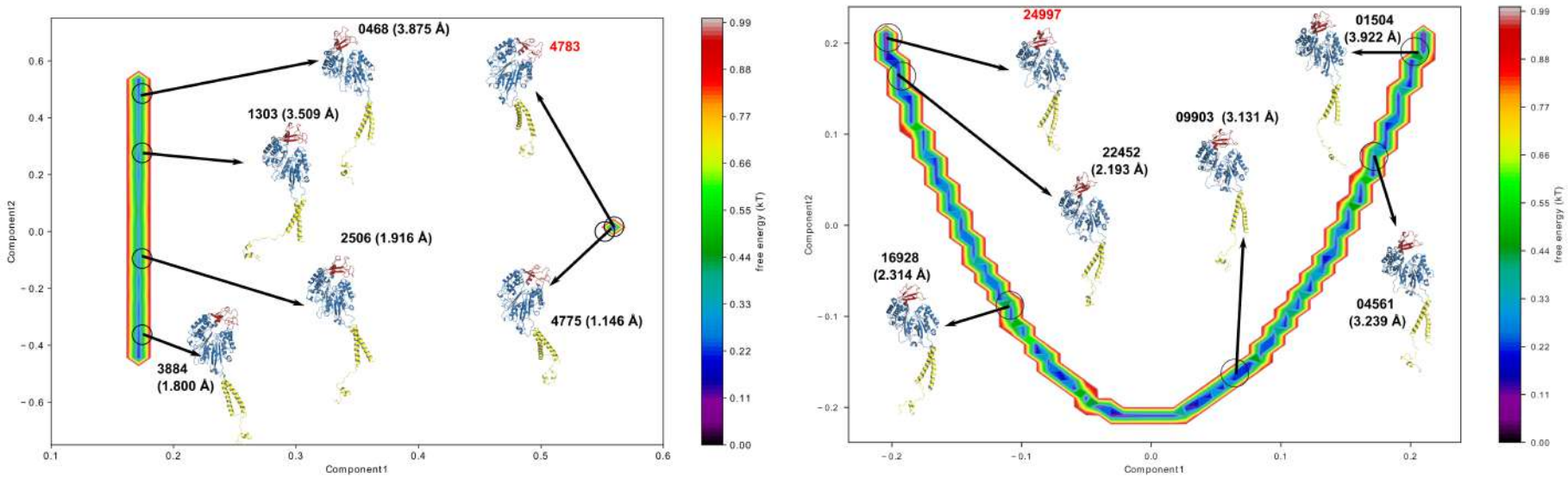
Nas figuras 20 e 21 são apresentados os perfis de energia obtidos para cada conjunto de trajetórias da smNTPDase1, usando o método WHAM. De acordo com os resultados obtidos, observamos que no geral o método spectral proporcionou perfis similares para os conjuntos de trajetórias. Isto parece estar relacionado as características usadas tanto para

o agrupamento quanto para obter os espaços intrínsecos e mapas de energia. As bacias de energia foram menores e pouco evidentes para as trajetórias cMD com ANP (figura 21(a)) do que aquelas verificadas nas outras simulações (figura20(a), 21(a) e 21(b)). Além disso, nos perfis das simulações convencionais existe um ponto de cela (região de máximo de energia), separando o mínimo de energia global do resto da superfície.



(a) (b)

Figura 20 – Perfis de energia das simulações cMD e aMD da proteína smNTPDase1 sem substrato. Nas Figuras estão representados os mapas de energia obtidos com o redutor Spectral para as simulações cMD (20(a)) e aMD (20(b)). Os mapas foram gerados utilizando o método WHAM. O número próximo a cada conformação representa o frame a qual pertence. As estruturas medóides (números em preto) são comparadas com aquela na curva de nível de energia igual a zero (númeração em vermelho), usando-se os valores de RMSD (valores entre parenteses). Os medóides dos grupos foram detectados usando o algoritmo Ward combinado com o método Elbow. As regiões em roxo e azul nos mapas representam possíveis bacias de energia enquanto que regiões próxima de vermelho são regiões de barreiras energéticas.



(a) (b)

Figura 21 – Perfis de energia das simulações cMD e aMD da proteína smNTPDase1 com substrato. Nas Figuras estão representados os mapas de energia obtidos com o redutor Spectral para as simulações cMD (21(a)) e aMD (21(b)). Os mapas foram gerados utilizando o método WHAM. O número próximo a cada conformação representa o frame a qual pertence. As estruturas medóides (números em preto) são comparadas com aquela na curva de nível de energia igual a zero (númeração em vermelho), usando-se os valores de RMSD (valores entre parenteses). Os medóides dos grupos foram detectados usando o algoritmo Ward combinado com o método Elbow. As regiões em roxo e azul nos mapas representam possíveis bacias de energia enquanto que regiões próxima de vermelho são regiões de barreiras energéticas.

Nas figuras 22 e 23 são apresentadas as comparações das variações dos valores de raio de giro para as regiões ECD e transmembranares (na presença e ausência de ANP) nas simulações cMD e aMD, respectivamente. De acordo com as análises do raio de giro do domínio ECD nas simulações cMD sem e com ANP, observamos que os valores desta métrica são menores quando o ligante está ausente no sítio ativo. Contudo, comparando-se o ângulo de rotação do domínio ECD1 dos estados medóides em relação ao modelo obtido por modelagem comparativa, verificamos que nas simulação sem ANP tal domínio teve uma amplitude de movimento maior do que na simulação considerando esta molécula (figura22(a) e 22(b)).

Quanto às hélices transmembranares, considerando as simulações cMD, as análises da flutuação do raio de giro e detecção das estruturas medóides permitiram observar que este domínio sofre maior variação de rotação em relação ao centro de massa quando comparado ao ECD. De acordo com as figuras 22(c) e 22(d), observamos que embora na simulação sem ANP os valores de raio de giro sejam maiores, esta métrica parece estabilizar próximo a 29Å, diferente da simulação considerando o ligante. Os valores dos ângulos entre as hélices indicam que a amplitude do movimento em "tesoura" foi maior quando considerado os heteroátomos no sítio catalítico da enzima, corroborando com a maior flutuação do raio de giro.

Em relação as simulações aMD, os resultados das análises do domínio extracelular indicam que, tanto na presença quanto ausência do ligante no sítio catalítico, os valores de raio de giro para ambas condições foram próximos entre si. Contudo, assim como observado para as simulações convencionais, a presença do ANP parece estabilizar o domínio ECD. Isto pode ser verificado pelas figuras 23(a) e 23(b), as quais mostram que, de acordo com a análise dos medóides, os valores do ângulo de rotação do subdomínio ECD1 em relação ao modelo 3D é maior nas simulações sem heteroátomos na cavidade catalítica.

Os resultados obtidos pelas análises para o domínio transmembranar mostram que diferente das simulações cMD, em ambas condições de simulação os valores de raio de giro desta região da proteína apresentou alta variação. Além disso, nas simulações aMD sem ANP os valores de raio de giro foram menores do que aqueles verificados para simulações considerando o ligante. Quanto aos ângulos entre as hélices transmembranares, assim como visto nas simulações convencionais, os valores indicam que a presença do substrato no sítio influencia no movimento em "tesoura" deste domínio.

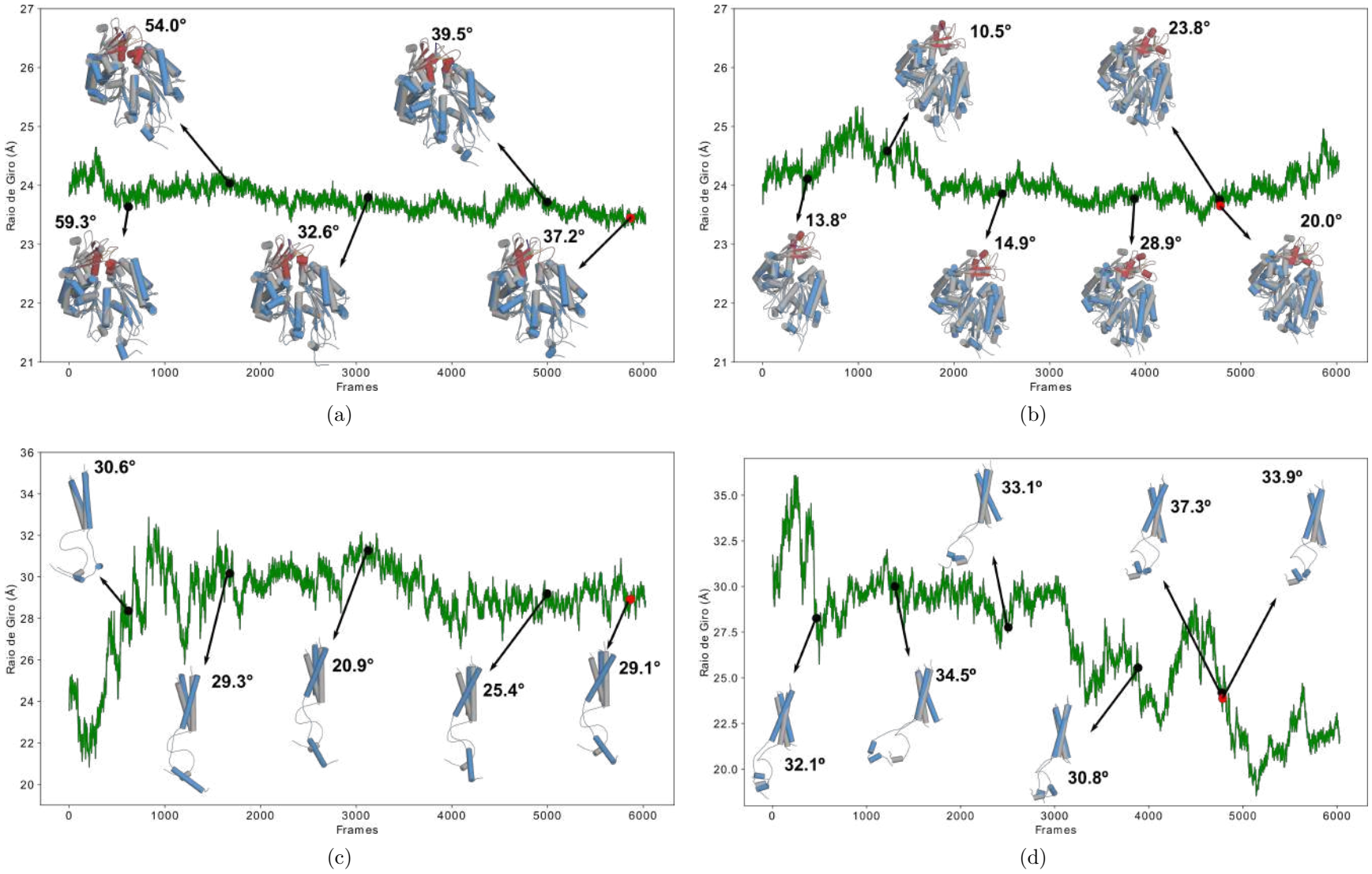


Figura 22 – Variação dos valores de raio de giro e ângulos de rotação dos domínios ECD e transmembranar nas simulações cMD. Nas figuras 22(a) e 22(b), são apresentados os valores de raio de giro e a variação do ângulo de rotação do subdomínio ECD1 das conformações medóides e mínimo de energia, em relação ao modelo publicado (71), para as simulações cMD sem e com ANP, respectivamente. Já nas figuras 22(c) e 22(d) são apresentados a flutuação dos valores de raio de giro da região transmembranar, bem como também, o ângulo entre as hélices, tanto para os medóides quanto para a estrutura mínimo de energia, para as simulações aMD sem e com ANP, respectivamente.

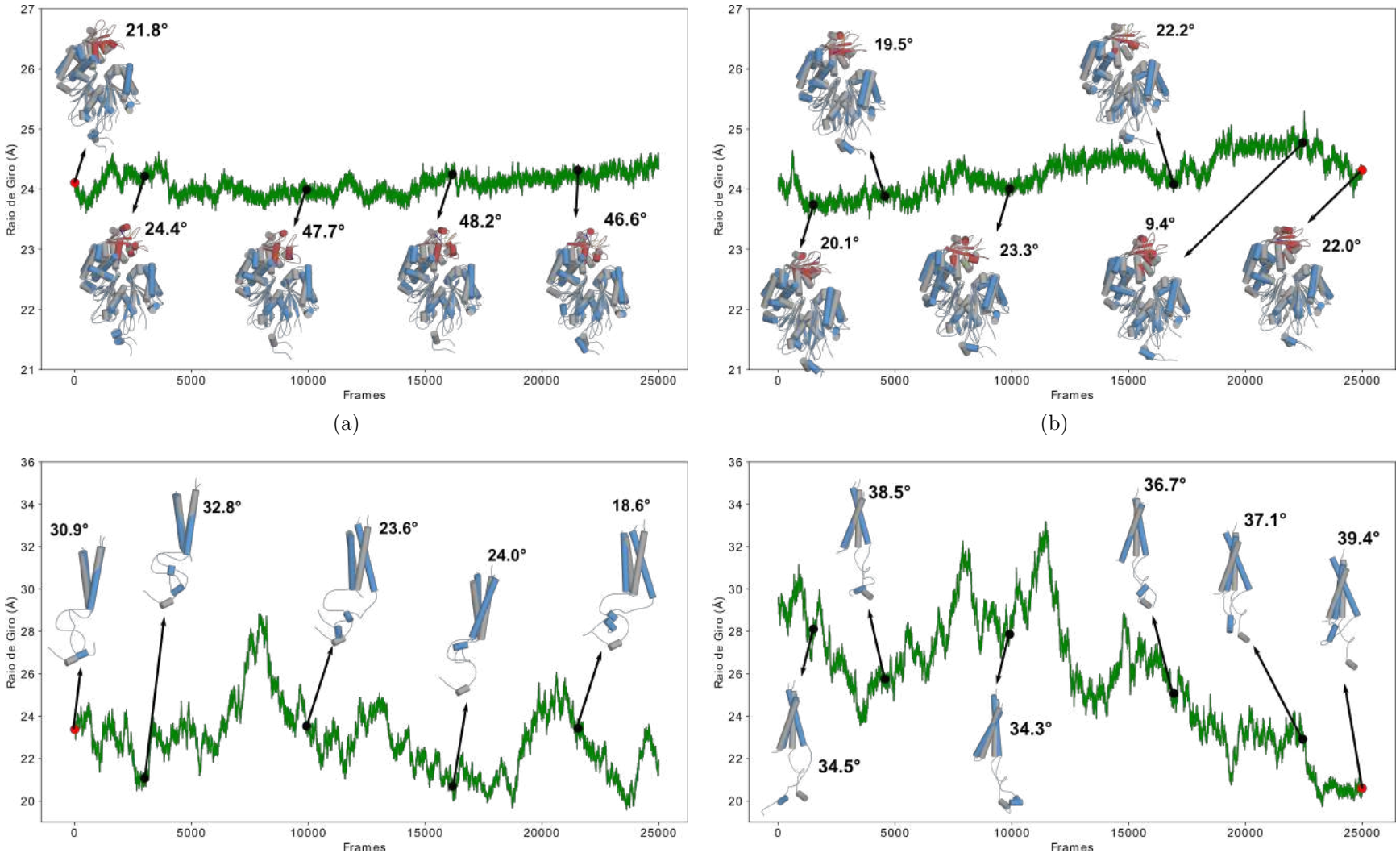


Figura 23 – Variação dos valores de raio de giro e ângulos de rotação dos domínios ECD e transmembranar nas simulações aMD.Nas figuras 23(a) e 23(b), são apresentados os valores de raio de giro e a variação do ângulo de rotação do domínio ECD1 das conformações medóides e mínimo de energia, em relação ao modelo publicado (71), para as simulações aMD sem e com ANP, respectivamente. Já nas figuras 23(c) e 23(d) são apresentados a flutuação dos valores de raio de giro da região transmembranar, bem como também, o ângulo entre as hélices, tanto para os medóides quanto para a estrutura mínimo de energia, para as simulações aMD sem e com ANP, respectivamente.

6.4.1.1 Análise de Dockings

Para a análise das interações proteína-ligante, foram usadas as conformações consideradas como mínimos de energia em cada conjunto de trajetórias. Os mínimos foram obtidos usando o método WHAM (*Weight Histogram Analysis Method*) aplicado sobre o espaço intrínseco obtido pelo redutor Spectral. Desta forma, considerando cada uma das configurações de dinâmica molecular, obtivemos 4 conformações no total. Além disso, para validar a caixa de simulação (*grid*) de *docking* foi usado o modelo tridimensional previamente gerado por nosso grupo para a smNTPDase1 (71).

Em estudo previamente publicado por Pereira *et. al.* 2018, foi sugerido que o composto LS1, o qual pertence a classe das chalconas, é um potencial candidato para inibição das smNTPDases. Na tabela 9 são apresentados os resultados de energia e as interações observadas entre o composto LS1 e cada um dos estados conformacionais da proteínas alvo. O resultado obtido para o modelo está energeticamente similar aos descritos no trabalho citado anteriormente, porém as interações observadas foram distintas. Estas podem estar relacionadas aos métodos empregados pelos algoritmos usados em cada trabalho. Além disso, as diferenças indicam que a molécula LS1 apresenta distintas formas de interação com a enzima smNTPDase1 com valores similares de energia.

Tabela 9 – Resultados de *dockings* obtidos para conformações em mínimos de energia da enzima smNTPDase 1.

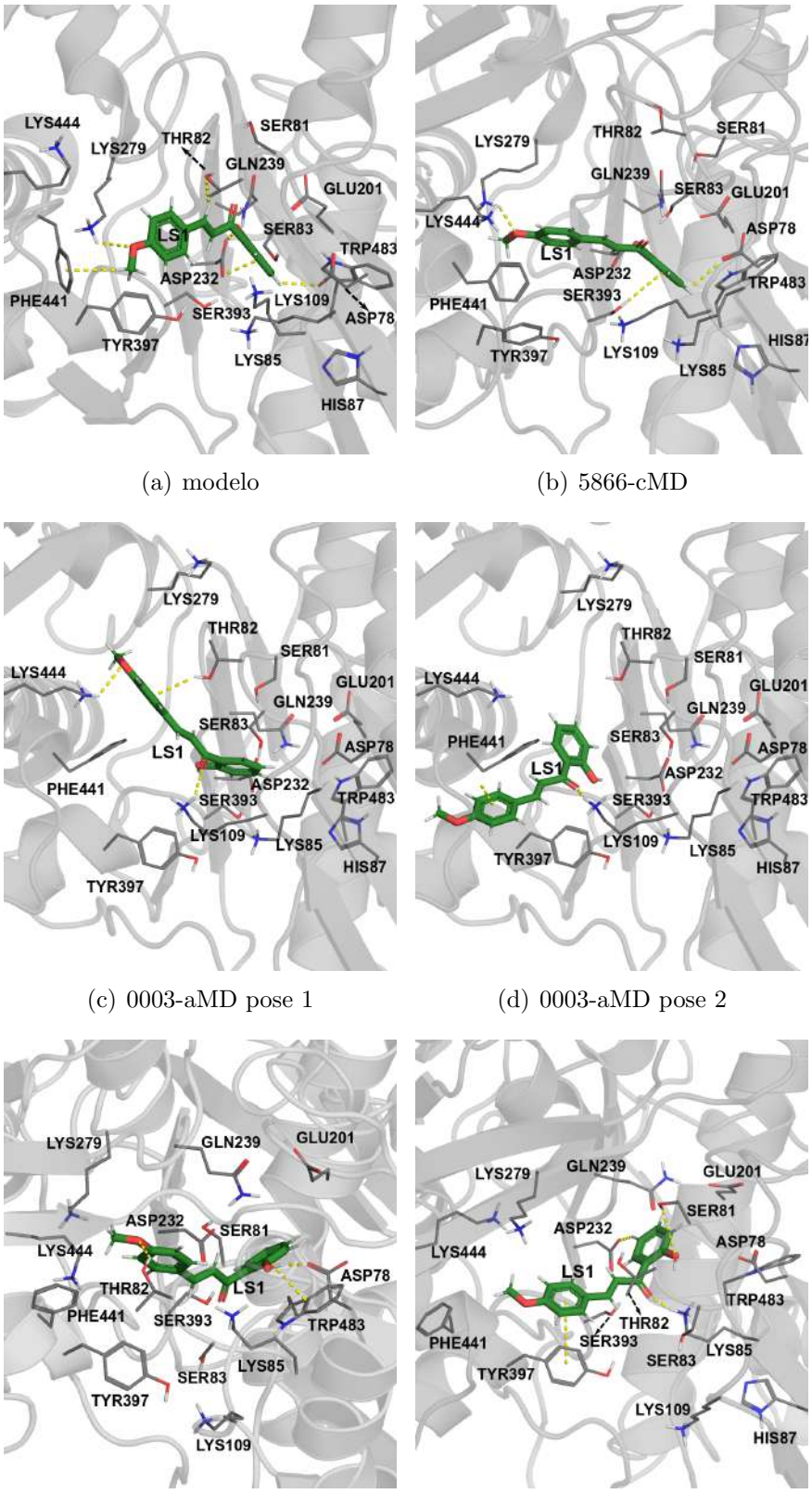
Simulação	Conformação	Score	cKi (μM)[a]	Interações
Modelo[b]	-	-7,3	4,45	S81, T82, S83, H87 e G235
Modelo[c]	-	-7,441	3,513	T82, K279, F441 e D232
sem ANP	5866 - cMD	-7,475	3,317	S393, D78, K279 e D232
	0003 - aMD	-7,340	4,166	T82, K109 e K444
		-8,31	0,810	K109 e F441
com ANP	4783 - cMD	-8,151	1,598	W483, T82 e D78
	24997 - aMD	-8,056	1,244	T82, K85, D232, Y397 e S81

[a] concentração de inibição calculada.

[b] resultado do *docking* realizado por Pereira *et. al. 2018* (76)

[c] resultado do *docking* realizado neste trabalho.

Em nosso estudo, os ensaios com o modelo apontam que o ligante interage por ligações de hidrogênio com os resíduos T82, K279 e D232 e com o resíduo F441 por meio de π-cátion do grupamento -O-CH_3 com o radical fenil (figura 24(a)). De acordo com Kozakiewicz *et. al.* (2008) , o resíduo T82 participa do processo de hidrólise do nucleotídeo, e portanto a interação observada favorece a inibição catalítica. Além disso, o resíduo F441 parece ser importante para a estabilidade do anel da base nitrogenada do substrato,



(a) modelo (b) 5866-cMD

(c) 0003-aMD pose 1 (d) 0003-aMD pose 2

(e) 4783-cMD (f) 24997-aMD

Figura 24 – Melhores interações moleculares proteína-ligante obtidas por *docking* para smNTPDase1.

reforçando o potencial inibidor das interações observadas. Embora Pereira *et. al.* (2018) tenha observado interações com os resíduos T81 e T82, similar ao encontrado aqui, em seus ensaios de *docking* molecular o composto não interagiu com os resíduos D232 e F441, resíduos que podem aumentar a estabilidade do composto LS1 no sítio catalítico. O maior número de interações com a proteína observados no presente resultado, levaram a melhores valores de *score* e concentração inibitória calculada (cK_i).

Quanto as análises de *docking* para as simulações cMD sem ANP, observamos que o composto interage com os resíduos S393, D78, K279 e D232 (figura 24(b)). Conforme o modelo descrito por Zebisch & Sträter 2008, os resíduos D45, D201 (D78 e D232 na smNTPDase 1) e R245 (K279 na smNTPDase 1) são responsáveis, respectivamente, pela estabilidade do íon bivalente no sítio, via interação com moléculas de água e do substrato nucleotídico via interação com o grupo hidroxila 3' do açúcar do substrato (ou como observado para o modelo da smNTPDase 1 com o anel da base nitrogenada (71)). Desta forma, embora o LS1 não tenha interagido com resíduos catalíticos, descritos na literatura, as interações observadas podem fornecer um impedimento do tipo estérico de forma a bloquear a entrada do substrato na cavidade catalítica (figura 24(b)).

Para a simulação de DM acelerada sem ANP, duas poses do composto LS1 parecem favorecer o potencial inibitório na enzima. Na primeira pose observamos que o ligante interagiu com os resíduos T82 via uma ligação do tipo OH-π, K109 e K444, via interações de hidrogênio (figura 24(c)). Já na segunda pose ocorreram interações do tipo CH-π com o carbono β do resíduo F441, responsável por estabilizar o anel da base nucleotídica, e interação de hidrogênio com o resíduo K109 (figura 24(d)). Estas interações, assim como discutido no parágrafo anterior, sugerem uma inibição por impedir a entrada do substrato no sítio catalítico, uma vez que não há ligações com resíduos catalíticos (figura (figura 24(c) e 24(d)).

Conforme já mencionado, foram realizadas também análises de *docking* molecular com o composto LS1 para os estados escolhidos das simulações com ANP. As análises para a simulação cMD nessa condição mostram que o composto interage com a proteína a partir de interações de hidrogênio com os resíduos D78 e T82 (figura 24(e)), os quais participam da estabilização do íon bivalente e da hidrólise, respectivamente. Além disso, observou-se também interação do tipo π-*stacking* com o resíduo W483, que também está associado a estabilidade do cátion bivalente via interação com molécula de água (figure figura 24(e)).

Na simulação acelerada com ANP, a melhor pose do composto LS1 apresentou interações com os resíduos S81, T82, K85, D232 e Y397 (figura 24(f)). Observamos que com relação aos resíduos catalíticos S81 e T82 o composto interage via ligações de hidrogênio com o grupo amina (R_1-NH-R_2) do *backbone* proteico. Já para os resíduos D232 e Y397, foram observadas interações de hidrogênio e π-*stacking*, respectivamente. O resíduo Y397 está relacionado, assim como F441, com a estabilidade do anel da base nitrogenada

do substrato, desta forma, tal interação atua impedindo a entrada e estabilização do nucleotídeo no sítio inativando a atividade catalítica.

6.4.1.2 Conclusões Parciais

Nesta seção foram apresentados resultados referentes às trajetórias obtidas por dinâmica molecular da isoforma 1 da enzima NTDase de *Schistosoma mansoni*. Foram realizadas aqui um total de quatro simulações, duas convencionais e duas aceleradas, adicionando um potêncial extra às energias diedrais da proteína, e além disso, considerando também a presença ou ausência do ligante ANP. Cada conjunto de conformações obtido foi submetido aos métodos de redução de dimensionalidade e agrupamento para detecção de estados significativos para análises de atracamento com possíveis inibidores.

Os resultados obtidos neste estudo mostram que, assim como observado em isoformas de outros organismos, a enzima smNTPDase 1 também apresenta movimentos de "tesoura" das hélices transmembranares, o qual possui maior amplitude na presença de substrato. A rotação do subdomínio ECD1, parece ocorrer naturalmente ao longo do tempo, porém a variação angular foi superior em simulações sem a presença da molécula substrato. Estes dados foram somente obtidos com o uso dos métodos de redução de dimensionalidade combinados com algoritmos de agrupamento.

A combinação entre os métodos Spectral, elbow e Ward, selecionado a partir de análises dos sistemas de teste e validação, permitiu a detecção de grupos conformacionais coesos, de acordo com os valores das métricas de qualidade. As estruturas medóides foram bem representativas em relação às mudanças características da enzima como descrito para esta família na literatura. Além disso, os FEL (*Free Energy Landscappe*), aplicando-se o método WHAM (*Weight Histogram Analysis Method*) sobre a espaço dimensional obtido pelo Spectral, mostraram que em simulações convencionais existem menos bacias de mínimo energético que naquelas aceleradas. Estes resultados podem sugerir que quando não há aplicação de potencial diedral extra, a enzima assume diferentes conformações porém dentro de uma mesma região, visitando número menor de regiões em seu espaço de fase. Embora essa conclusão possa ser interessante, ainda precisa de novos estudos aplicando outras informações como coordenadas de reação e também análise de parâmetros para os métodos utilizados aqui.

Com relação aos ensaios de *docking*, os resultados mostram que o composto LS1 possui um importante potencial de inibição da enzima smNTPDase 1, principalmente por interagir com resíduos essenciais para o mecanismo de catálise. Neste estudo, diferente do que foi previamente publicado sobre testes com o modelo da presente proteína, este potencial inibidor parece agir em diferentes estados conformacionais da smNTPDase 1. Contudo, o mesmo teve resultados relativamente melhores, quanto as interações e energia, em estruturas nas quais a enzima apresenta maior compactação do domínio ECD

e amplitude do ângulo entre as hélices transmembranares (movimento de tesoura).

Em conclusão, as análises aqui realizadas foram capazes de fornecer importantes *insights* estruturais e de inibição sobre a enzima smNTPDase 1 a nível molecular, os quais não foram publicados ou verificados anteriormente. As abordagens usadas para detecção automática de conformações proteícas mostraram-se essenciais para análise das simulações por dinâmica molecular, reforçando a importância da aplicação da inteligência computacional.

6.5 Sistemas de Estudo de Caso 2: smNTPDase 2

Nesta seção será apresentado e discutido os resultados obtidos das simulações e ensaios de atracamento proteína-ligante para a enzima smNTPDase 2. Como já foi abordado anteriormente, essa isoforma diferente da smNTPDase 1 é secretada e por tanto a região transmembranar é clivada.

6.5.1 Análise Agrupamento

Para os testes de agrupamento para as simulações realizadas para enzima smNTPDase2, assim como para a smNTPDase1, foi usada a abordagem baseada na média e desvio-padrão do RMSF para detectar os resíduos aminoácidos de interesse e assim obter as matrizes de contato. A combinação dos métodos Spectral, Elbow e Ward foi novamente aplicada, uma vez que mostrou reesultados satisfatórios em relação aos testes anteriores. Além disso, utilizou-se 6025 frames dos 120500 referentes as 250 nanossegundos de simulação convencional e todos os 25000 frames das simulações aceleradas.

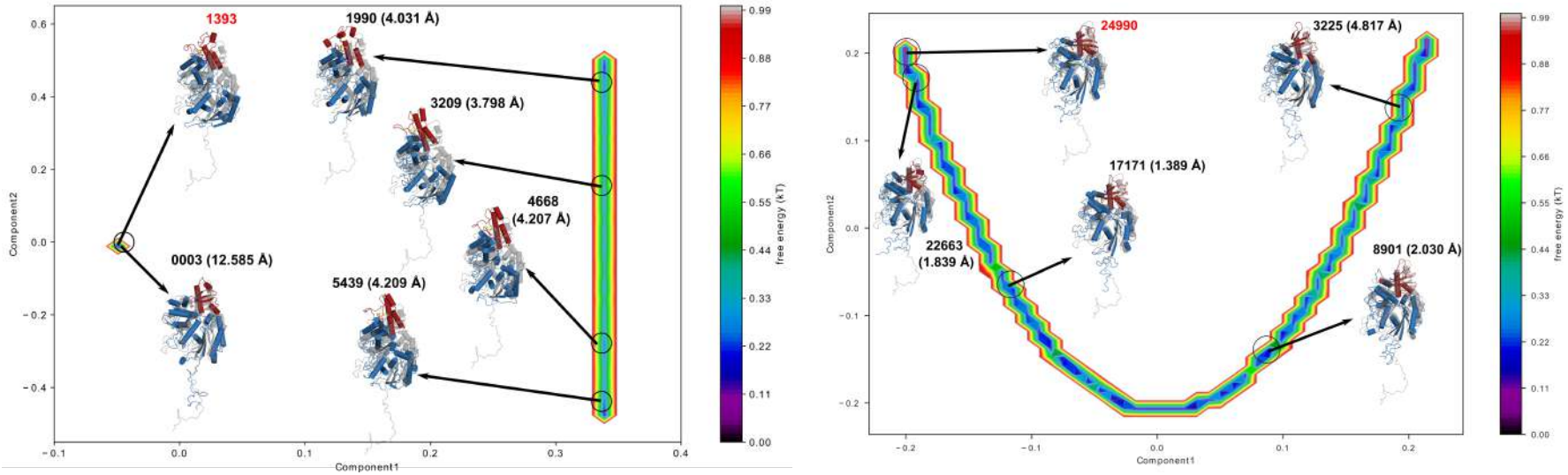
Na tabela 10 são apresentados os resultados das métricas de qualidade dos agrupamentos obtidos para cada uma das simulações da enzimas smNTPDase2. Conforme esperado pelas análises de RMSD e raio de giro, na simulação convencional sem ligante foram encontrados cinco grupos conformacionais.Já para a simulação acelerada, devido o fato da proteína relaxar sua estrutura mais rápido e apresentar menor flutuação conformacional, a abordagem de agrupamento detectou quatro grupos.

Tabela 10 – Avaliação dos algoritmos de agrupamento aplicados sobre as simulações da enzima smNTPDase 2.

Simulação	*n_clusters*	CH	DBI	Silhueta
sm2-cMD	5,0	11875,674	1,570	0,663
sm2-aMD	4,0	29609,809	1,837	0,542
sm2_au1-cMD	4,0	14535,372	0,996	0,730
sm2_au1-aMD	4,0	30238,346	1,704	0,546

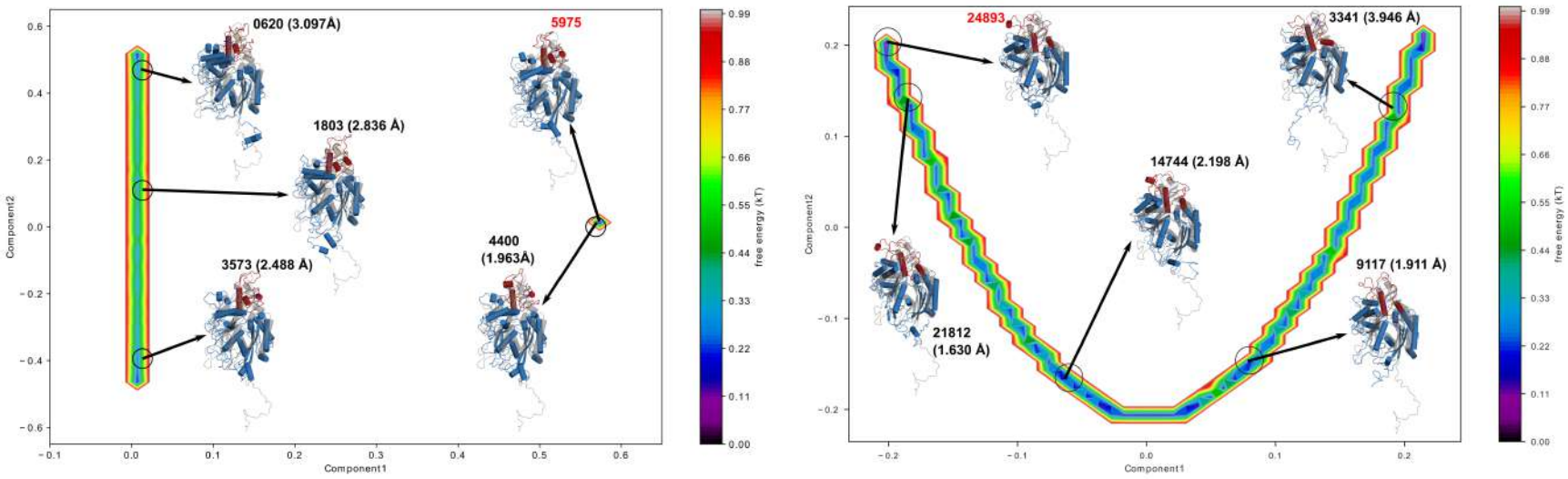
As análises de agrupamento para as simulações com a presença do ligante AU1 no sítio, mostraram que em geral a proteína assume 4 grupos conformacionais. Estes resultados podem estar relacionados ao fato do ligante estabilizar a estrutura a proteína, de modo que esta convirja em menor tempo em relação a estrutura sem ligante. Isso fica evidenciado nas análises dos valores de raio de giro e RMSD, os quais são relativamente menores nas simulações com ligante do que naquelas onde o mesmo foi removido.

Nas figuras 25 e 26 são apresentados os perfis de energia obtidos para cada conjunto de trajetórias da smNTPDase2, usando o método WHAM. Os resultados mostram que os perfis obtidos foram próximos aos observados para a enzima smNTPDase 1. Além disso, os mapas foram similares entre os diferentes conjuntos de trajetórias.



(a) (b)

Figura 25 – Perfis de energia das simulações cMD e aMD da proteína smNTPDase2 sem substrato. Nas Figuras estão representados os mapas de energia obtidos com o redutor Spectral para as simulações cMD (25(a)) e aMD (25(b)). Os mapas foram gerados utilizando o método WHAM. O número próximo a cada conformação representa o frame a qual pertence. As estruturas medóides (números em preto) são comparadas com aquela na curva de nível de energia igual a zero (númeração em vermelho), usando-se os valores de RMSD (valores entre parenteses). Os medóides dos grupos foram detectados usando o algoritmo Ward combinado com o método Elbow. As regiões em roxo e azul nos mapas representam possíveis bacias de energia enquanto que regiões próxima de vermelho são regiões de barreiras energéticas.



(a) (b)

Figura 26 – Perfis de energia das simulações cMD e aMD da proteína smNTPDase2 com substrato. Nas Figuras estão representados os mapas de energia obtidos com o redutor Spectral para as simulações cMD (26(a)) e aMD (26(b)). Os mapas foram gerados utilizando o método WHAM. O número próximo a cada conformação representa o frame a qual pertence. As estruturas medóides (números em preto) são comparadas com aquela na curva de nível de energia igual a zero (núméração em vermelho), usando-se os valores de RMSD (valores entre parenteses). Os medóides dos grupos foram detectados usando o algoritmo Ward combinado com o método Elbow. As regiões em roxo e azul nos mapas representam possíveis bacias de energia enquanto que regiões próxima de vermelho são regiões de barreiras energéticas.

Em geral, as bacias de energia para as simulações aceleradas foram mais evidentes do que aquelas observadas nas simulações convencionais. Observamos também que, nestas últimas, similar ao obtido para smNTPDase1, a região do mínimo de energia global está separado do resto da superfície por um ponto de cela.

Nas figuras 27(a) e 27(b) são mostrados os valores de RMSF para as simulações cMD na ausência e presença do ligante AU1, respectivamente. De acordo com os valores, observamos que a região com maior flutuação pertence a porção N-terminal, a qual não possui estrutura secundária definida no modelo 3D publicado para esta enzima. A partir das análises de agrupamento, foi possível verificar que tal porção tende a se enovelar, acompanhando o padrão de compactação da proteína verificado pelo raio de giro (figura 27(c) e 27(d)). Contudo, embora a porção N-terminal tenha se compactado nas simulações sem ANP, ocorreram perda de estrutura secundária nas regiões vizinhas e diretamente conectadas (figura 27(c)).

Quanto as simulações aMD, os resultados de RMSF novamente apontam que a região com maior flutuação é a porção N-terminal (28(a) e 28(b)). Assim como nas simulações cMD, tal região segue a compactação da proteína, porém diferente do observado anteriormente, não há perda de estrutura secundária das regiões próximas (28(c) e 28(d)). Isto pode estar associado ao fato de que durante as simulações aMD a proteína sofre um relaxamento mais rápido que nas simulações cMD, já que os valores de raio de giro destas últimas são maiores que nas primeiras.

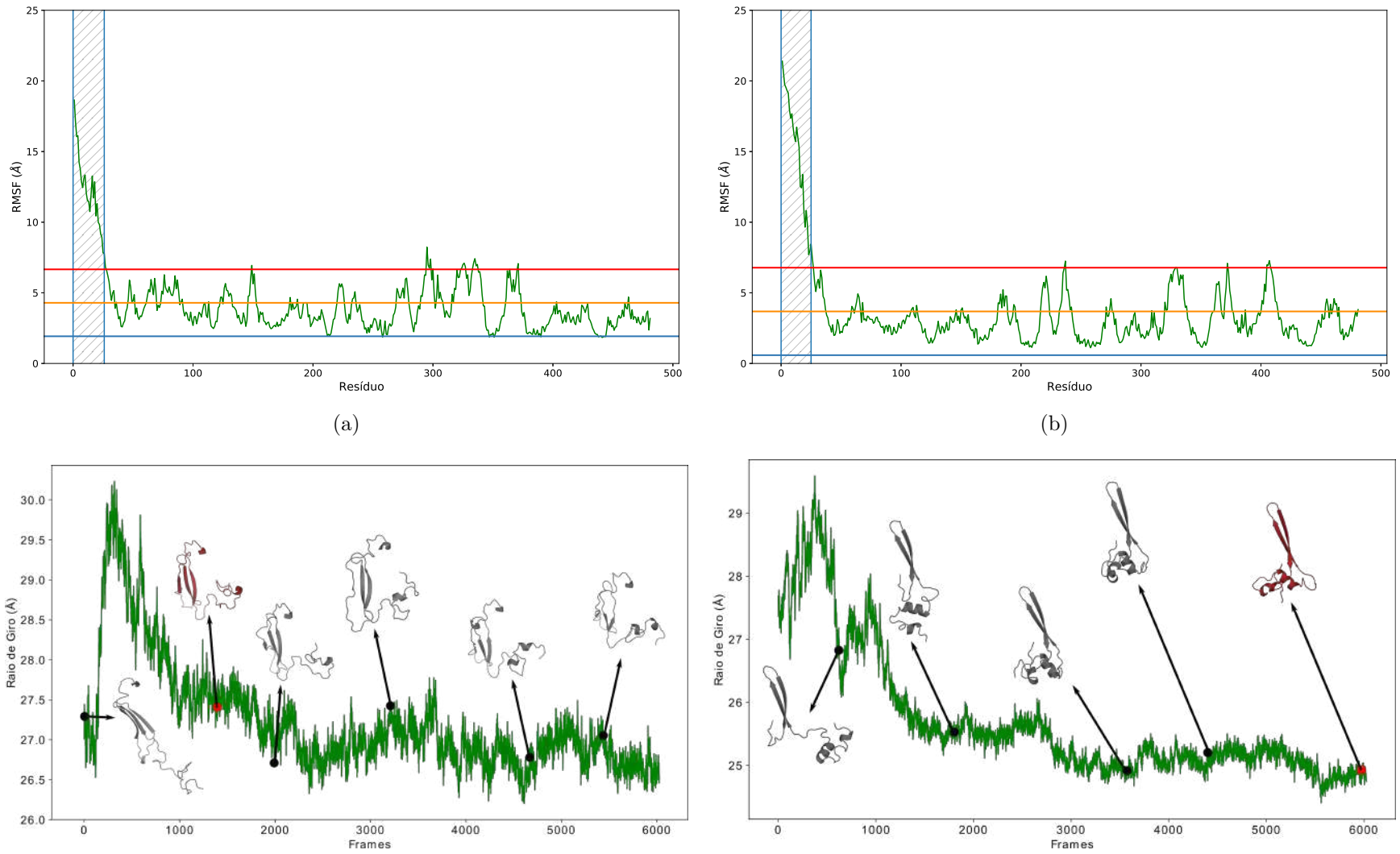


Figura 27 – Variação dos valores de RMSF e raio de giro nas simulações cMD da proteína smNTPDase2. Nas figuras 27(a) e 27(b) são apresentados os valores de flutuação em Å dos resíduos da proteína smNTPDase2 nas simulações cMD com e sem AU1, respectivamente. A região hachurada em destaque representa a porção N-terminal da enzima, a qual é apresentada nas figuras 27(c) e 27(d). Nestas são mostrados os valores de raio de giro para a proteína completa e o enovelamento da porção N-terminal das estruturas medóides (cinza) e mínimo de energia (vermelho) para simulações cMD sem AU1 (27(c)) e com esta molécula (27(d)).

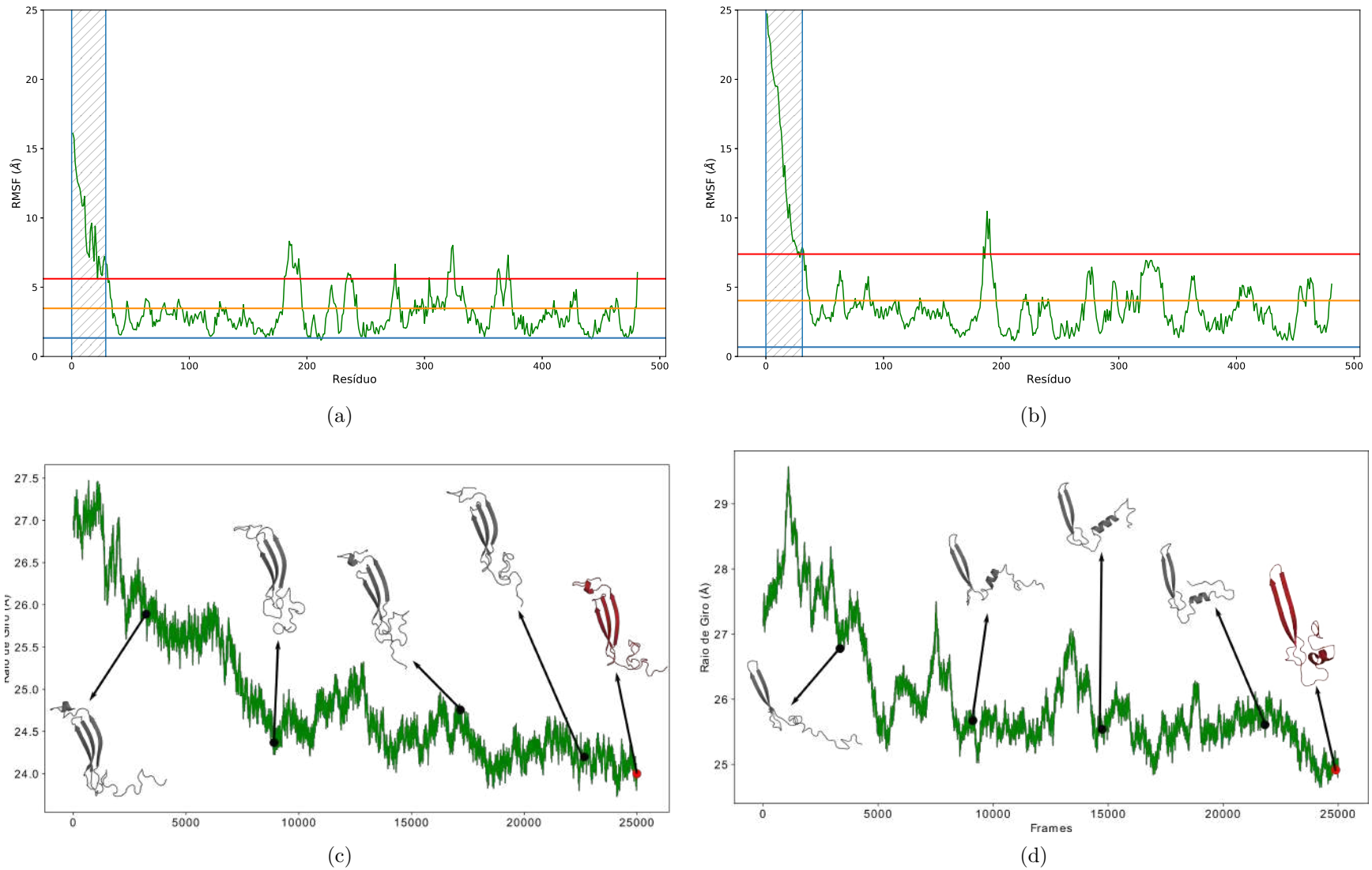


Figura 28 – Variação dos valores de RMSF e raio de giro nas simulações aMD da proteína smNTPDase2. Nas figuras 28(a) e 28(b) são apresentados os valores de flutuação em Å dos resíduos da proteína smNTPDase2 nas simulações cMD com e sem AU1, respectivamente. A região hachurada em destaque representa a porção N-terminal da enzima, a qual é apresentada nas figuras 28(c) e 28(d). Nestas são mostrados os valores de raio de giro para a proteína completa e o enovelamento da porção N-terminal das estruturas medóides (cinza) e mínimo de energia (vermelho) para simulações aMD sem AU1 (28(c)) e com esta molécula (28(d)).

Assim como verificado para a enzima smNTPDase 1, foi avaliado aqui o ângulo de rotação do subdomínio ECD1 da proteína smNTPDase 2. Esta região foi mapeada a partir do alinhamento entre a estrutura de ambas isoformas. Além disso, o ângulo de rotação foi calculado baseando-se na modelo publicado por (23).

Nas figuras 29(a) e 29(b) são apresentados os valores dos ângulos de rotação entre domínios para as estruturas medóides e o mínimo de energia das simulações cMD sem e com substrato, respectivamente. Os resultados mostram que na simulação convencional sem ligante ocorreu uma maior flutuação dos valores de rotação e raio de giro da proteína completa, juntamente com uma maior variação estrutural do subdomínio ECD1, do que nas simulações considerando ligante. Interessantemente, embora a estrutura proteíca apresente "abertura" do sítio catalítico ao longo da trajetória cMD sem ligante, não foi observada mudanças drásticas nos valores de superfície acessível ao solvente (*Solvent accessible surface area* - SASA).

As mesmas análises foram realizadas para simulações aceleradas sem e com AU1, e os resultados obtidos apontam que as mudanças conformacionais foram inversas aquelas observadas anteriormente (figura 29(c) e 29(d)). Nas simulações aMD sem AU1 os valores do ângulo de rotação do subdomínio ECD1 em relação ao modelo foram menores do que em simulações aMD com ligante (figura 29(c)). Estes resultados seguem o padrão dos valores de raio de giro nestas simulações, o qual tende a decrescer, indicando que a proteína se compacta ao longo das trajetórias. Contudo, em relação aos valores de SASA, observamos que nas simulações aMD com AU1 a superfície acessível apresentou flutuações relativamente menores, indicando que embora em ambos casos ocorra compactação da estrutura, a presença do ligante leva a proteína relaxar-se e estabilizar-se mais rapidamente.

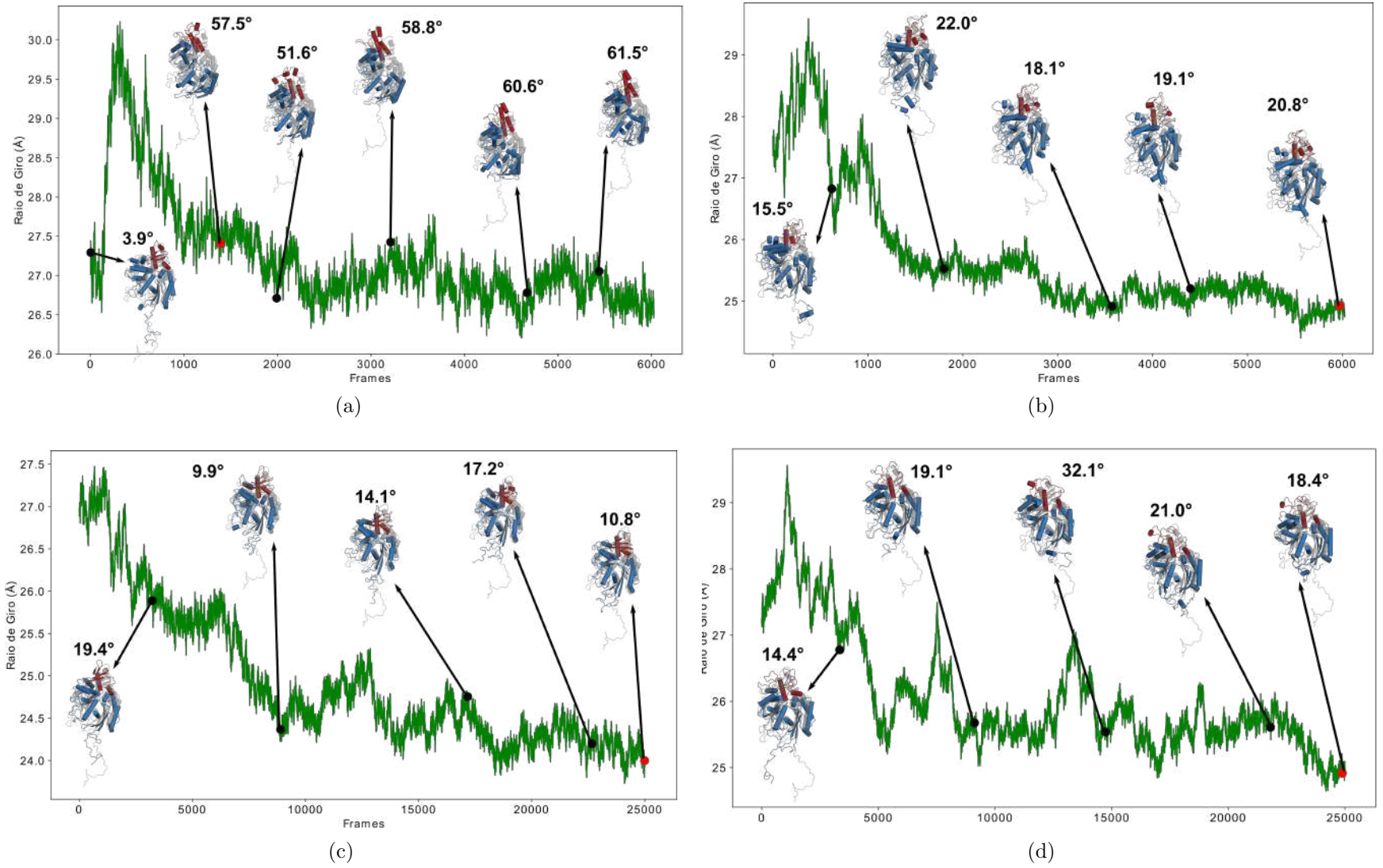
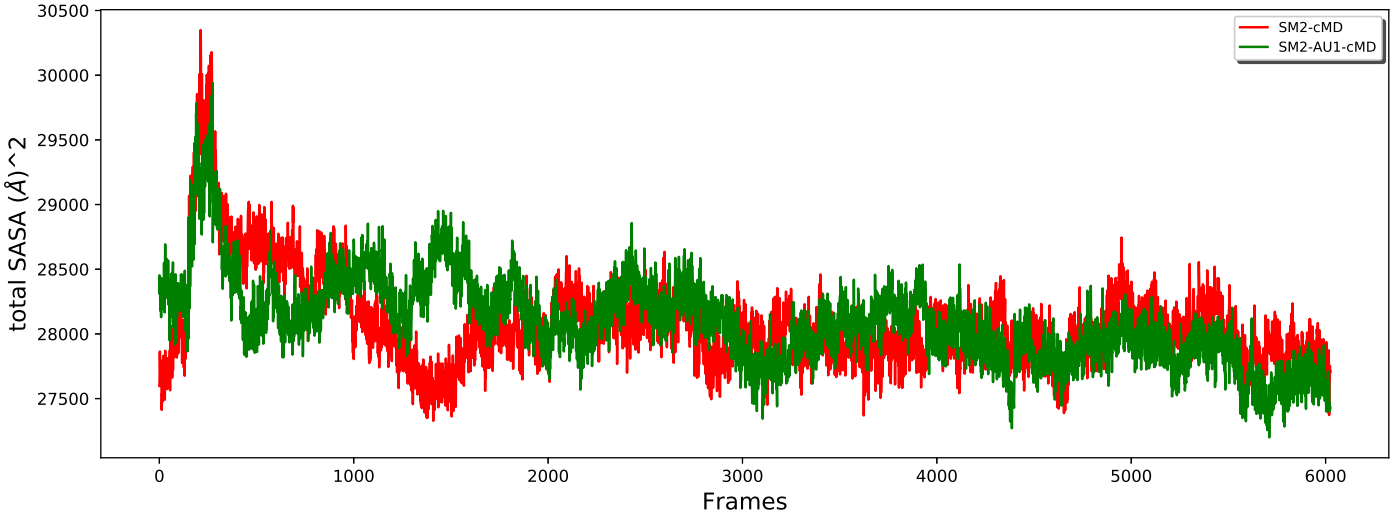
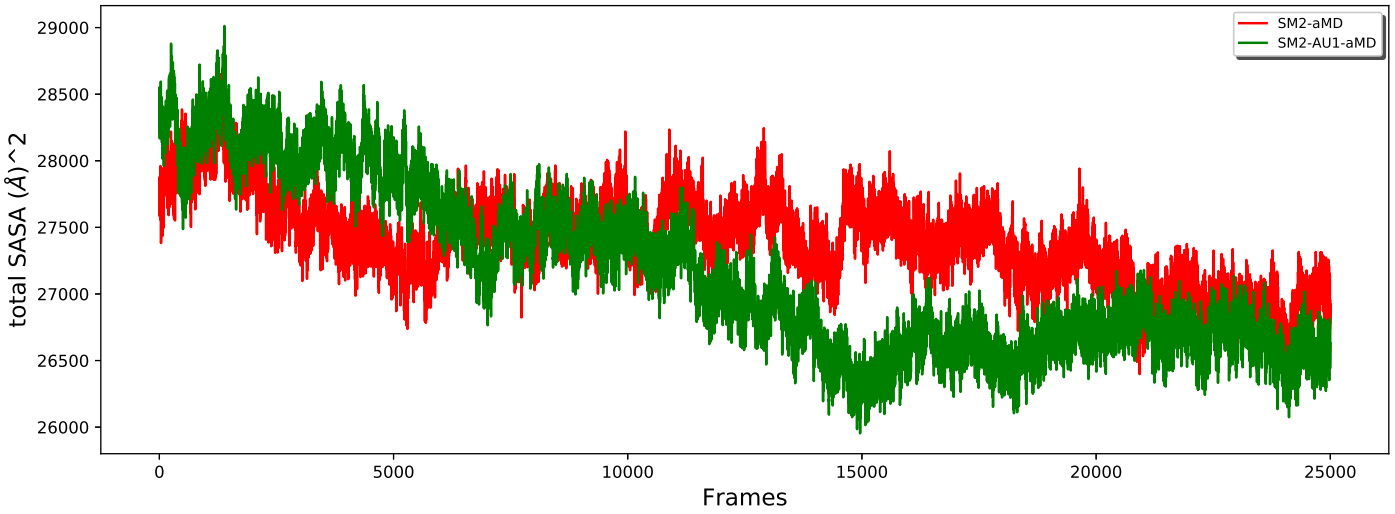


Figura 29 – Variação dos valores de raio de giro e ângulo de rotação do subdomínio ECD1 das simulações cMD e aMD, sem e com AU1. Nas figuras 29(a) e 29(c), são apresentados os valores de raio de giro e a variação do ângulo de rotação do domínio ECD1 das conformações medóides e mínimo de energia, em relação ao modelo publicado, para as simulações cMD e aMD sem ANP, respectivamente. Já nas figuras 29(b) e 29(d) são apresentados os valores de raio de giro e a variação do ângulo de rotação do domínio ECD1 das conformações medóides e mínimo de energia, em relação ao modelo publicado, para as simulações cMD e aMD com ANP, respectivamente.



(a)



(b)

Figura 30 – Variação dos valores da área de supperfície acessível ao solvente (SASA) para simulações cMD e aMD, sem e com AU1. Nas figuras (a) e (b), são apresentados os valores de SASA para as simulações cMD e aMD, respectivamente, considerando tanto a presença quanto ausência do ligante.

6.5.1.1 Análise de Dockings

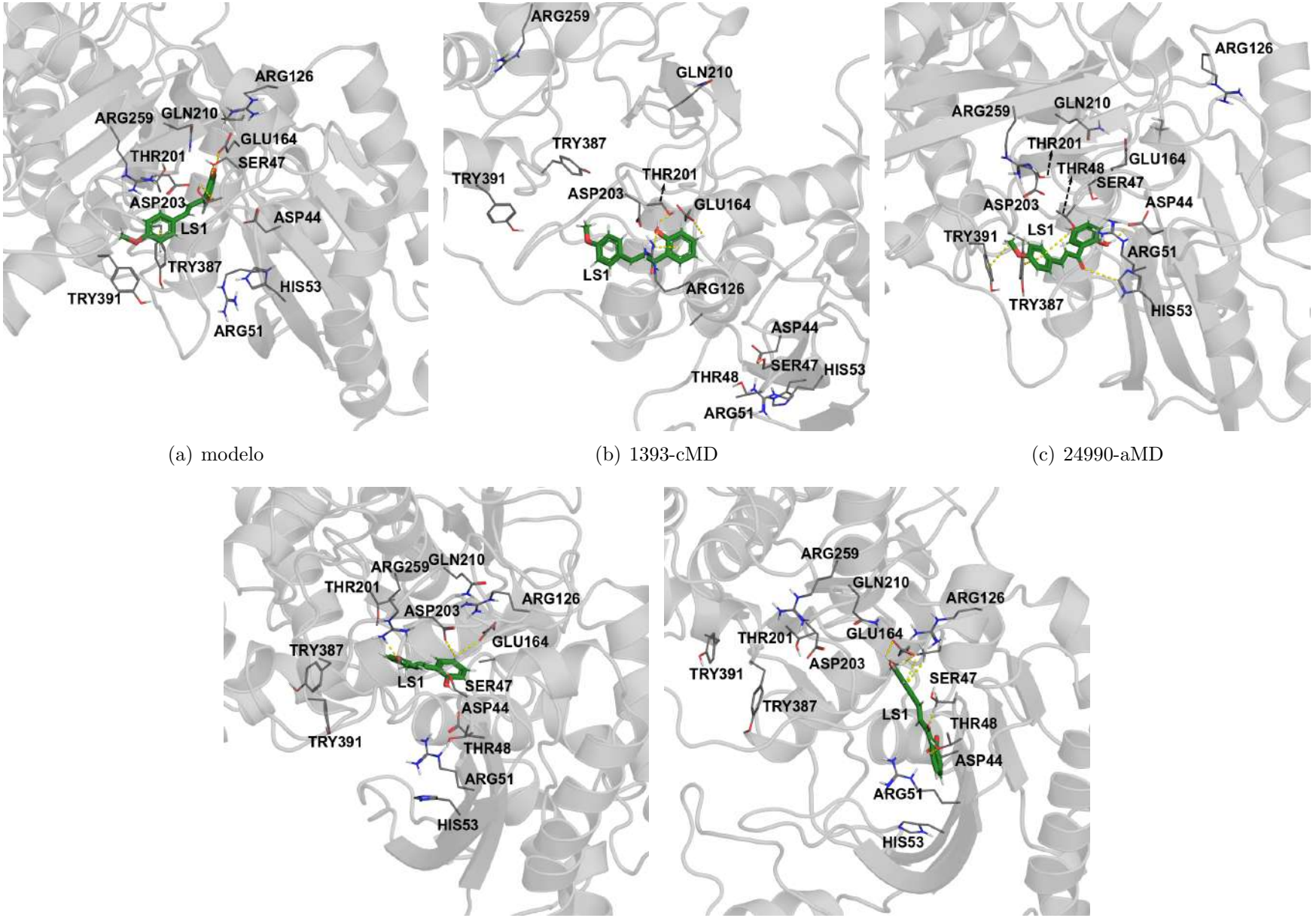
Nesta seção será apresentado os resultados de atracamento molecular do composto LS1 com as diferentes conformações obtidas para a proteína smNTPDase2. Na tabela 11 são apresentados os resultados de energia e as interações observadas entre tal molécula e cada um dos estados considerados mínimos da proteínas alvo em cada um dos conjuntos de trajetórias. Comparando-se os resultados obtidos para o modelo e outras conformações da smNTPDase2 com aqueles apresentados para smNTPDase1, observamos que energeticamente o composto LS1 parece interagir com esta enzima tão bem quanto com a a isoforma 1 de *S. mansoni*. Pereira *et. al.* (2018), mostraram que para o modelo de smNTPDase1 os valores do *score* e cK_i foram -7.3 Kca/mol e 4.45μM, respectivamente.

Os testes com o modelo mostram que o composto LS1 é capaz de interagir via ligações de hidrogênio com os resíduos T48 e E164 (figura 31(a)), os quais são considerados catalíticos de acordo com os modelos de Zebisch e Sträter (2008) e Kozakiewicz *et. al.* (2008), respectivamente. Além disso, foi obsservado pareamento do tipo T-*stacking* com o resíduo Y387 (figura 31(a)) que parece estar relacionado a estabilidade dos aneis da base nitrogenada do substrato, juntamente com o resíduo T391 (23). Estas interações apontam para uma potencial inibição da enzima via bloqueio da hidrólise do substrato, já que os resíduos catalíticos são bloqueados de interagir com o mesmo.

Tabela 11 – Resultados de *dockings* obtidos para conformações em mínimos de energia da enzima smNTPDase 2.

Simulação	Conformação	Score	cKi	Interações
modelo	-	-7,952	1,483	Y387, T48, E164
sem AU1	1393 - cMD	-8,075	1,205	E164, T201, R126
	24990 - aMD	-8,240	0,912	Y391, Y387, R51, D44, T48
com AU1	5975 - cMD	-7,095	6,299	R259, S47, D203, E164
	24893 - aMD	-7,334	4,208	E164, R126, A123, T48, D44

Nas análises da simulação convencional da smNTPDase 2 sem AU1, observamos que a proteína sofre uma grande mudança estrutural, expondo a região do sítio catalítico, o que poderia facilitar a entrada de substratos, bem como também do composto LS1. Os resultados de atracamento molecular mostram que, assim como observado para o modelo, a potencial molécula inibidora interage com o resíduo catalítico E164 via ligação de hidrogênio (figura 31(b)). Porém, além desta interação, foram observadas ligações de hidrogênio com os resíduos R126 e T201, os quais não apresentam ligações com o substrato nem mesmo com o íon bivalente na estrutura do modelo. Além disso, não ocorreram interações com os resíduos catalíticos S47 e T48, pois os mesmos se afastam da cavidade catalítica e são expostos ao solvente ao longo da simulação.



(a) modelo (b) 1393-cMD (c) 24990-aMD

(d) 5975-cMD (e) 24893-aMD

Figura 31 – Melhores interações moleculares proteína-ligante obtidas por *docking* para smNTPDase2.

Por outro lado, na simulação acelerada sem AU1, a estrutura da proteína permanece compactada, mesmo ocorrendo movimentações das ACR. Os resultados de atracamento mostraram que o composto LS1 não interagiu com o catalítico T48 via -OH-π e por ligações de hidrogênio com os resíduos D44 e R51. Além disso, o mesmo realizou interações do tipo τ-*stacking* e π-*stacking* com o anel 4-hidroxifenil dos resíduos Y387 e Y391, respectivamente (figura 31(c)).

Conforme um alinhamento estrutural entre a smNTPDase 2 e a NTPDase 1 de *Legionella pneumophila* (LpNTPDase 1 - PDB4BR7), verificamos que os resíduo D44, na primeira, e D49 na outra possuem a mesma função de estabilizar o íon metálico no sítio (conforme abordado para a smNTPDase 1). Já o resíduo R51 na smNTPDase 2, referente ao R56 da LpNTPDase 1, pode estar relacionado ao modo de ligação do substrato. Assim como apresentado por Zebisch *et. al.* (2013), tais resíduos (da mesma forma que a K85 na smNTPDase 1) representam uma substituição funcional da histidina 50 na NTPDase 2 de*Rattus novergicus* (RnNTPDase 2). A importancia destes resíduos pode ser justificada pelo fato de que formas mutantes R56S de LpNTPDase 1 apresentam uma diminuição de 250 vezes na eficiência catalítica (116).

Nas análises para as simulações com a presença do AU1 no sítio catalítico, observamos que os resultados foram energeticamente maiores e consequentemente a concentração inibitória (cK_i) foi superior aos realizados anteriormente. Porém, tanto para simulação cMD e aMD as melhores poses do composto LS1 apresentaram interações com resíduos importantes para a hidrólise do substrato. Além disso, tais resultados foram qualitativamente próximos com aqueles obtidos por Pereira *et. al.* (2018) para a smNTPDase 1.

Para a conformação mínima da simulação cMD com AU1, observamos que o composto LS1 realizou ligações de hidrogênio com os resíduos catalíticos S47 e E164, bem como também com os resíduos D203 e R259 (figura 31(d)). O resíduo D203 possui função similar ao D232 na smNTPDase 1, isto é, estabilizar o íon metálico no sítio. Já a arginina 259, tanto no modelo quanto ao longo das simulações, não observamos interações diretas com o substrato, porém a interação entre este resíduo e composto LS1 pode estabilizar esta molécula no sítio catalítico e desta forma promover inibição da hidrólise.

Quanto as análises para a simulação aMD com AU1, foram observadas interações com os resíduos catalíticos D44, T48, A123, R126 e E164 (figura 31(e)). Conforme descrito para a RnNTPDase 2, a cadeia lateral do resíduo R126 (mesma numeração para smNTPDase 2) está orientada para o resíduo E165 (E164 em nossa enzima) formando uma ponte salina auxiliando o posicionameto do resíduo catalítico. Estudos de mutação direcionada ao resíduo R143 (equivalente ao R126) em NTPDases 3 mostram que em mutantes nos quais a arginina foi substituída por alanina perderam atividade catalítica (115, 51). Devido a importância do resíduo R126, bem como, dos resíduos catalíticos (T48 e E164) e daqueles responsáveis pela estabilidade do cátion bivalente, concluímos que de

fato as interações do LS1 podem favorecer o processo de inibição da enzima smNTPDase 2.

6.5.1.2 Conclusões Parciais

Nesta seção foram apresentados resultados referentes às trajetórias obtidas por dinâmica molecular da isoforma 2 da enzima NTDase de *Schistosoma mansoni*. Foram realizadas aqui um total de quatro simulações, duas convencionais e duas aceleradas, adicionando um potêncial extra às energias diedrais da proteína, e além disso, considerando também a presença ou ausência do ligante AU1. Cada conjunto de conformações obtidas foram submetidas a métodos de redução de dimensionalidade e agrupamento para detecção de estados significativos para análises de atracamento com possíveis inibidores.

Os resultados obtidos aqui mostram que, diferente das isoformas da enzima NTPDase descrita para outros organismos, a smNTPDase 2 aparentou ser mais instável estruturalmente. Isto ficou evidenciado quando observado as variações de estrutura secundária da região equivalente ao subdomínio ECD1 da smNTPDase 1, especialmente nas simulações cMD sem AU1. Além disso, comparando as duas isoformas da NTPDase em *S. mansoni*, as flutuações dos valores de raio de giro para smNTPDase 2 foram maiores do que aqueles obtidos para a smNTPDase 1.

A aplicação dos métodos Spectral, elbow e Ward permitiram, em geral, a detecção de estados medóides e de mínimo global (via WHAM - *Weight Histogram Analysis Method*), os quais proporcionaram informações moleculares e estruturais importantes sobre a enzima aqui estudada. A primeira conclusão obtida foi com relação a região N-terminal, que na estrutura previamente publicada por Souza *et. al.* (2014), enovelou-se ao longo das trajetórias assumindo estados compactados. Uma vez que a smNTPDase 2 não apresenta regiões transmembranares em sua estrutura madura, os movimentos da região N-terminal parecem estar diretamente relacionados as mudanças do domínio equivalente ao ECD. Além disso, os resultados das simulações permitem propor que, devido a ausência de regiões transmembranares, as maiores flutuações observadas sejam necessárias para o processo de catálise.

Os FEL (*Free Energy Landscappe*) mostraram que em simulações convencionais existem menos bacias de mínimo energético que naquelas aceleradas, similar ao verificado para a smNTPDase 1. Corroborando com o fato de que com a aplicação de um potencial diedral extra, maior número de regiões do espaço de fase da proteína são visitados. Os resultados dos mapas de energia, juntamente com as análises de RMSD e raio de giro, perminte especular que as barreiras energéticas da superfíce energética das smNTPDases são pequenas e com as simulações aMD elas podem ser facilmente transpostas. Esta última faz sentido uma vez que as proteínas sofrem um relaxamento mais rápido nas simulações aceleradas.

Sobre os ensaios de *docking*, os resultados mostram que o composto LS1 também possui um potencial de inibição da enzima smNTPDase 2, interagindo com resíduos catalíticos, conforme observado para smNTPDase 1. As energias e valores de cK_i foram tão bons quanto os publicados por Pereira *et. al.* 2018 e observados na seção anterior referente a isoforma 1. Além disso, os resultados aqui não permitem associar as energias obtidas com os ensaios de *docking* e o grau de compactação da estrutura proteica, como especulada para smNTPDase 1.

Concluímos que as análises aqui realizadas permitiram obter importantes informações estruturais e moleculares sobre a da enzima smNTPDase 2 e possível inibição desta. O que proporciona *insights* para novos estudos teóricos e práticos sobre tratamentos alternativos para a esquistossomose.

7 Conclusões Finais

A proposta do presente estudo foi realizar uma análise comparativa entre as diferentes combinações de métodos de aprendizado de máquina para detecção de conformações representativas da trajetória das proteínas oriundas de simulações de dinâmica molecular em diferentes temperaturas. Esta abordagem visa contribuir na redução do conjunto de estruturas a serem analisadas e consideradas para trabalhos como análise de padrões conformacionais assumidos durante o enovelamento proteico ou ainda em ensaios de *ensemble docking*, nos quais diferentes estados da mesma molécula podem ser considerados para o desenho racional e seleção de fármacos através de *virtual screening*. Para isto, foram aplicados seis diferentes redutores de dimensionalidade, para obter possíveis "espaços essenciais" capazes de descrever o conjunto de movimentos proteicos que foram combinados com diferentes algoritmos de agrupamento.

Uma vez que o RMSD (*Root Mean Square Deviation*) não é capaz de representar de forma eficaz as mudanças conformacionais e associá-las a um perfil de energia livre, foi usada aqui as matrizes de distância euclidiana (EDM) entre os átomos $C\alpha$ dos resíduos aminoácidos para cada conformação. Do ponto de vista da dinâmica essencial, este conjunto de distâncias entre átomos pode ser considerada como uma coordenada interna e apresenta importante associação com a superfície energética assumida pela proteínas ao longo de suas dinâmicas conformacionais. As EDM obtidas foram então usadas para a busca do espaço essencial de movimentos proteicos e este como informação para o agrupamento das conformações de DM. Os resultados apontam que esta métrica foi capaz de determinar com eficiência as mudanças conformacionais das moléculas testadas aqui. Contudo, uma análise comparativa com outras métricas, como por exemplo as coordenadas cartesianas dos átomos, é necessária.

Sobre os redutores de dimensionalidade (RD) aplicados, observou-se que os métodos não-lineares Isomap e Spectral foram capazes de fornecer um discernimento sobre a separação de classes de conformações tão bem quanto o PCA. E apresentaram bons resultados quando analisadas as métricas de qualidade, isto pode ser observado pelo valor de SI obtido. Além disso, observou-se que o espaço essencial detectado por estes três redutores forneceu perfis de energia completamente distintos ente si, o que pode ser explicado pelo forma que cada método mapeia a dimensão original em um espaço reduzido. De forma geral, os mapas obtidos por PCA e Spectral fornecem um melhor entendimento da variação da superfície de energia, muito embora isto necessite de simulações longas com melhor exploração do espaço conformacional das proteínas.

Entre os diferentes métodos usados aqui, observou-se que os algoritmos de agrupamento K-means e Ward apresentaram os melhores resultados de agrupamento, independente dos métodos de redução de dimensionalidade, obtendo bons valores das métricas de qua-

lidade avaliadas aqui, além de melhores valores de SI (*Synthesis Index*). Os algoritmos Meanshift e Affinity propagation (AP), outro lado, apresentaram resultados opostos para os diferentes testes. Em geral Meanshift obtém poucos conjuntos de conformações, mesmo em simulações em alta temperatura, enquanto que o AP gera muitos grupos, mesmo em simulações onde observa-se pouca variação estrutural, os quais apresentam poucas estruturas e que poderiam ser unidos em outros grupos afim de obter agrupamentos mais coesos.

O fluxograma desenvolvido no presente estudo foi capaz de reduzir o tempo de análise e seleção de conformações representativas de simulações de DM por um especialista, o que pode demorar dias, além de ser uma tarefa exaustiva para o pesquisador. De acordo com os nossos resultados, o tempo para análise dos dados de forma automatizada utilizando o fluxograma pode ser reduzido para aproximadamente 9 minutos (539.2 segundos) em testes onde foi aplicado o método AutoEncoder ao conjunto de trajetórias da 1CLL à 310K. Mais testes ainda necessitam ser realizados para determinar a eficiência para cada combinação de métodos e associá-los com a complexidade e dimensionalidade dos dados de entrada.

Por último, os resultados apresentados no presente estudo mostram que o uso de métodos de redução de dimensionalidade combinados com algoritmos de agrupamento são capazes de auxiliar na análise da dinâmica conformacional de proteínas. Além disso, o uso do método Spectral embedding para determinar o espaço essencial dessas mudanças mostrou-se uma abordagem interessante como alternativa ao PCA e Isomap, gerando mapas de energia (FEL) capazes de fornecer uma boa definição das barreiras e bacias exploradas pelas moléculas aqui simuladas. O trabalho também possui grande importância por contribuir para o desenvolvimento de uma nova linha de pesquisa no grupo.

De acordo com as análises dos sistemas de teste e validação, a melhor abordagem para detecção dos mapas de energia e agruamento foi a combinação Spectral+Elbow+Ward. Estes métodos foram usados para detecção das estruturas medóides e os mínimos de energia para as trajetórias das enzimas smNTPDase 1 e 2. A partir da verificação de métricas relacionadas a estrutura (RMSD, raio de giro, ângulo de rotação entre domínios e SASA) dos estados representativos foi possível verificar padrões de movimento em cada isoformas em diferentes condições de simulação. Os resultados obtidos fornecem informações importantes para a novos estudos *in silico* e *in vitro*.

Usando as conformações em mínimo de energia, foram realizados ensaios de *docking* do composto LS1 contra cada uma das isoformas. Os resultados apontam que esta molécula possui importante potencial de inibição tanto para a smNTPDase 1 quanto smNTPDase 2, atuando em diferentes estados conformacionais. Além disso, os resultados corroboram com trabalhos previamente publicados sugerindo o composto LS1 como alternativa a drogas como praziquantel no tratamento da equistossomose.

De maneira geral, foram realizadas contribuições importantes a cerca dos métodos de inteligência computacional aplicados na análise do espaço conformacional de biomoléculas. Bem como também, na compreensão das enzimas com importância terapeutica estudadas por nosso grupo de pesquisa. Apesar de atacar diferentes problemas, este trabalho deixa em aberto determinadas questões, especialmente com relação a otimização de parâmetro de métodos, que abrem caminho para serem exploradas em novos estudos.

8 CONTRIBUIÇÕES

Parte dos resultados descritos no presente estudo estão publicados em artigos e resumos de congressos:

- artigo publicado durante o evento *Latin American Conference on Computational Intelligence* (LACCI), o qual foi indexado pela IEEE ();
- apresentação de resumo durante o evento *2nd Brazilian Student Council Symposium*;
- apresentação de resumo na XI Escola de Modelagem Molecular de Sistemas Biológicos (EMMSB), no qual o trabalho foi agraciado com o título de menção honrosa.
- artigo publicado durante o evento IWBBIO 2019 (*7th International Wrork-Conference on Bioinformatics and Biomedical Engineering*), o qual foi indexado pela revista *Lecture Notes in Computer Science*();
- artigo aceito para publicação, aplicando os métodos aqui etudados sobre a enzima EthA ().

9 TRABALHOS FUTUROS

Como proposta de trabalhos futuros:

- Avaliar a aplicação de outras medidas de distância (*e.g.* manhattan), ou somente o uso das coordenadas cartesianas dos átomos, para determinar flutuações conformacionais;
- Avaliar novas características para obter o espaço de fase das proteínas;
- Realizar seleção de conformações significativas baseando-se nos mínimos de energia por grupo;
- Aplicar o método de *grid search* para determinar os melhores parâmetros dos métodos de redução de dimensionalidade e agrupamento;
- Desenvolver uma interface, disponibilizando o programa desenvolvido neste trabalho para auxiliar a análise de conformações moleculares diferentes pesquisadores da área.

REFERÊNCIAS

1 Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

2 Albert, S., Czibula, G., and Teletin, M. (2018). Analyzing the impact of protein representation on mining structural patterns from protein data. In *2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pages 000533–000538. IEEE.

3 Alonso, H., Bliznyuk, A. A., and Gready, J. E. (2006). Combining docking and molecular dynamic simulations in drug design. *Medicinal research reviews*, 26(5):531–568.

4 Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain"goodness of fit"criteria based on stochastic processes. *The annals of mathematical statistics*, pages 193–212.

5 Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396.

6 Bernardi, R., Bhandarkar, M., Bhatele, A., Bohm, E., Brunner, R., Buelens, F., Chipot, C., Dalke, A., Dixit, S., Fiorin, G., et al. (2018). Namd user's guide. *Urbana, Illinois, USA: Theoretical and Computational Biophysics Group, Beckman Institute, University of Illinois*.

7 Bernardi, R. C., Melo, M. C., and Schulten, K. (2015). Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1850(5):872–877.

8 Bhardwaj, R. and Skelly, P. J. (2009). Purinergic signaling and immune modulation at the schistosome surface? *Trends in parasitology*, 25(6):256–260.

9 Brown, W. M., Martin, S., Pollock, S. N., Coutsias, E. A., and Watson, J.-P. (2008). Algorithmic dimensionality reduction for molecular structure analysis. *The Journal of chemical physics*, 129(6):064118.

10 Buonfiglio, R., Recanatini, M., and Masetti, M. (2015). Protein flexibility in drug discovery: From theory to computation. *ChemMedChem*, 10(7):1141–1148.

11 Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

12 Campbell, M. K., Farrell, S. O., Reyes, A. C., Gasco, J. H. C., Talavera, J. B., Bonilla, A. C., and Muñoz, G. R. (2004). *Bioquímica*, volume 4. Thomson.

13 Castro-Borges, W., Simpson, D. M., Dowle, A., Curwen, R. S., Thomas-Oates, J., Beynon, R. J., and Wilson, R. A. (2011). Abundance of tegument surface proteins in the human blood fluke schistosoma mansoni determined by qconcat proteomics. *Journal of Proteomics*, 74(9):1519–1533.

14 Chattopadhyaya, R., Meador, W. E., Means, A. R., and Quiocho, F. A. (1992). Calmodulin structure refined at 1.7 å resolution. *Journal of molecular biology*, 228(4):1177–1192.

15 Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619.

16 Cozzini, P., Kellogg, G. E., Spyrakis, F., Abraham, D. J., Costantino, G., Emerson, A., Fanelli, F., Gohlke, H., Kuhn, L. A., Morris, G. M., et al. (2008). Target flexibility: an emerging consideration in drug discovery and design. *Journal of medicinal chemistry*, 51(20):6237–6255.

17 Das, P., Moll, M., Stamati, H., Kavraki, L. E., and Clementi, C. (2006). Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proceedings of the National Academy of Sciences*, 103(26):9885–9890.

18 Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

19 Da'dara, A. A., Bhardwaj, R., Ali, Y. B., and Skelly, P. J. (2014a). Schistosome tegumental ecto-apyrase (smatpdase1) degrades exogenous pro-inflammatory and pro-thrombotic nucleotides. *PeerJ*, 2:e316.

20 Da'dara, A. A., Bhardwaj, R., and Skelly, P. J. (2014b). Schistosome apyrase smatpdase1, but not smatpdase2, hydrolyses exogenous atp and adp. *Purinergic signalling*, 10(4):573–580.

21 De Paris, R., Quevedo, C. V., Ruiz, D. D., and de Souza, O. N. (2015). An effective approach for clustering inha molecular dynamics trajectory using substrate-binding cavity features. *PloS one*, 10(7):e0133172.

22 de Souza, V. C., Goliatt, L., and Capriles, P. V. (2019). Insight about nonlinear dimensionality reduction methods applied to protein molecular dynamics. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 219–230. Springer.

23 de Souza, V. C., Nunes, V. S., Vasconcelos, E. G., Faria-Pinto, P., and Capriles, P. V. (2014). Structural comparative analysis of secreted ntpdase models of schistosoma mansoni and homo sapiens. In *Brazilian Symposium on Bioinformatics*, pages 91–98. Springer.

24 DeMarco, R., Kowaltowski, A. T., Mortara, R. A., and Verjovski-Almeida, S. (2003). Molecular characterization and immunolocalization of schistosoma mansoni atp-diphosphohydrolase. *Biochemical and Biophysical Research Communications*, 307(4):831–838.

25 Dokmanic, I., Parhizkar, R., Ranieri, J., and Vetterli, M. (2015). Euclidean distance matrices: Essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30.

26 dos Santos Machado, K. (2011). *Seleção eficiente de conformações de receptor flexível em simulações de docagem molecular*. PhD thesis, Pontifícia Universidade Católica do Rio Grande do Sul.

27 Du, X., Li, Y., Xia, Y.-L., Ai, S.-M., Liang, J., Sang, P., Ji, X.-L., and Liu, S.-Q. (2016). Insights into protein–ligand interactions: mechanisms, models, and methods. *International journal of molecular sciences*, 17(2):144.

28 Duan, M., Fan, J., Li, M., Han, L., and Huo, S. (2013). Evaluation of dimensionality-reduction methods from peptide folding–unfolding simulations. *Journal of chemical theory and computation*, 9(5):2490–2497.

29 Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64.

30 Durrant, J. D. and McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC biology*, 9(1):71.

31 El-Hamdouchi, A. and Willett, P. (1987). Techniques for the measurement of clustering tendency in document retrieval systems. *Journal of Information Science*, 13(6):361–365.

32 Faria-Pinto, P., Meirelles, M., Lenzi, H., Mota, E., Penido, M., Coelho, P., and Vasconcelos, E. (2004). Atp diphosphohydrolase from schistosoma mansoni egg: characterization and immunocytochemical localization of a new antigen. *Parasitology*, 129(1):51–57.

33 Ferguson, A. L., Panagiotopoulos, A. Z., Kevrekidis, I. G., and Debenedetti, P. G. (2011). Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach. *Chemical Physics Letters*, 509(1-3):1–11.

34 Fischer, E. (1894). Einfluss der configuration auf die wirkung der enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3):2985–2993.

35 Fraccalvieri, D., Pandini, A., Stella, F., and Bonati, L. (2011). Conformational and functional analysis of molecular dynamics trajectories by self-organising maps. *BMC bioinformatics*, 12(1):158.

36 Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814):972–976.

37 Garcia, J. C. L. (2008). *Caracterização bioquímica e imunológica das enzimas recombinantes ATP-difosfohidrolases 1 e 2 do parasita Schistosoma mansoni*. PhD thesis, Universidade de São Paulo.

38 Ghodsi, A. (2006). Dimensionality reduction a short tutorial. *Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada*, 37:38.

39 Goliatt, P. V. Z. C. (2007). *Técnicas de Bioinformática e Modelagem Computacional Aplicadas ao Estudo do Genoma de Trypanosoma cruzi e de Enzimas Consideradas de Interesse no Tratamento da Doença de Chagas: Estudo Particular das Cruzipaínas 1 e 2*. PhD thesis, Laboratório Nacional de Computação Científica.

40 Goliatt, P. V. Z. C. (2011). *Desenvolvimento e implementação de um modelo coarse-grained para predição de estruturas de proteínas*. PhD thesis, Ph. D. dissertation, LNCC, Rio de Janeiro/Brasil.

41 Gordon, H. L. and Somorjai, R. L. (1992). Fuzzy cluster analysis of molecular dynamics trajectories. *Proteins: Structure, Function, and Bioinformatics*, 14(2):249–264.

42 Guedes, I., Barreto, A., Miteva, M., and Dardenne, L. (2016). Development of empirical scoring functions for predicting protein-ligand binding affinity. *Soc. Bras. Bioquim. Biol. Mol*, pages 1–174.

43 Guedes, I. A., de Magalhães, C. S., and Dardenne, L. E. (2014). Receptor–ligand molecular docking. *Biophysical reviews*, 6(1):75–87.

44 Hamelberg, D., Mongan, J., and McCammon, J. A. (2004). Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *The Journal of chemical physics*, 120(24):11919–11929.

45 Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

46 Heaslet, H., Rosenfeld, R., Giffin, M., Lin, Y.-C., Tam, K., Torbett, B. E., Elder, J. H., McRee, D. E., and Stout, C. D. (2007). Conformational flexibility in the flap domains of ligand-free hiv protease. *Acta Crystallographica Section D: Biological Crystallography*, 63(8):866–875.

47 Hospital, A., Goñi, J. R., Orozco, M., and Gelpí, J. L. (2015). Molecular dynamics simulations: advances and applications. *Advances and applications in bioinformatics and chemistry: AABC*, 8:37.

48 Humphrey, W., Dalke, A., and Schulten, K. (1996). Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38.

49 Ichiye, T. and Karplus, M. (1991). Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Bioinformatics*, 11(3):205–217.

50 Karplus, M. and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9(9):646–652.

51 Kirley, T. L., Crawford, P. A., and Smith, T. M. (2006). The structure of the nucleoside triphosphate diphosphohydrolases (ntpdases) as revealed by mutagenic and computational modeling analyses. *Purinergic Signalling*, 2(2):379.

52 Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., and Kolinski, A. (2016). Coarse-grained protein models and their applications. *Chemical Reviews*, 116(14):7898–7936.

53 Knowles, A. F. (2011). The gda1_cd39 superfamily: Ntpdases with diverse functions. *Purinergic signalling*, 7(1):21–45.

54 Koshland Jr, D. E. (1995). The key–lock theory and the induced fit theory. *Angewandte Chemie International Edition in English*, 33(23-24):2375–2378.

55 Kovács, F., Legány, C., and Babos, A. (2005). Cluster validity measurement techniques. In *6th International symposium of hungarian researchers on computational intelligence*. Citeseer.

56 Kozakiewicz, A., Neumann, P., Banach, M., Komoszyński, M., and Wojtczak, A. (2008). Modeling studies of potato nucleoside triphosphate diphosphohydrolase ntpdase1: an insight into the catalytic mechanism. *Acta Biochimica Polonica*, 55(1):141–150.

57 Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

58 Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., and Kollman, P. A. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of computational chemistry*, 13(8):1011–1021.

59 Levano-Garcia, J., Mortara, R. A., Verjovski-Almeida, S., and DeMarco, R. (2007). Characterization of schistosoma mansoni atpdase2 gene, a novel apyrase family member. *Biochemical and Biophysical Research Communications*, 352(2):384–389.

60 Liao, C. and Zhou, J. (2014). Replica-exchange molecular dynamics simulation of basic fibroblast growth factor adsorption on hydroxyapatite. *The Journal of Physical Chemistry B*, 118(22):5843–5852.

61 Liou, C.-Y., Cheng, W.-C., Liou, J.-W., and Liou, D.-R. (2014). Autoencoder for words. *Neurocomputing*, 139:84–96.

62 Lyman, E. and Zuckerman, D. M. (2006). Ensemble-based convergence analysis of biomolecular trajectories. *Biophysical journal*, 91(1):164–172.

63 Ma, J. (2005). Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, 13(3):373–380.

64 Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

65 Martínez, L. (2007). *Simulações de dinâmica molecular dos receptores do hormônio tireoideano*. PhD thesis, Instituto de Química/Universidade Estadual de Campinas.

66 McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L.-P., Lane, T. J., and Pande, V. S. (2015). Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal*, 109(8):1528 – 1532.

67 Meireles, L., Gur, M., Bakan, A., and Bahar, I. (2011). Pre-existing soft modes of motion uniquely defined by native contact topology facilitate ligand binding to proteins. *Protein Science*, 20(10):1645–1658.

68 Murtagh, F. and Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, ii. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(6):e1219.

69 Neidigh, J. W., Fesinmeyer, R. M., and Andersen, N. H. (2002). Designing a 20-residue protein. *Nature structural biology*, 9(6):425–430.

70 Nishizawa, Y., Okui, Y., Inaba, M., Okuno, S., Yukioka, K., Miki, T., Watanabe, Y., and Morii, H. (1988). Calcium/calmodulin-mediated action of calcitonin on lipid metabolism in rats. *The Journal of clinical investigation*, 82(4):1165–1172.

71 Nunes, V. S., Vasconcelos, E. G., Faria-Pinto, P., Borges, C. C. H., and Capriles, P. V. (2015). Structural comparative analysis of ecto-ntpdase models from s. mansoni and h. sapiens. In *International Symposium on Bioinformatics Research and Applications*, pages 247–259. Springer.

72 Nunes, V. S. P. (2015). *Análise comparativa das Ecto-NTPDase 1 de Homo sapiens e Schistosoma mansoni por meio de modelagem tridimensional, dinâmica molecular e docking receptor-ligante*. PhD thesis, Ph. D. dissertation, UFJF, Juiz de Fora - MG/Brasil.

73 Onuchic, J. N., Luthey-Schulten, Z., and Wolynes, P. G. (1997). Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry*, 48(1):545–600.

74 Paris, R. D., Quevedo, C. V., Ruiz, D. D., Souza, O. N. d., and Barros, R. C. (2015). Clustering molecular dynamics trajectories for optimizing docking experiments. *Computational intelligence and neuroscience*, 2015:32.

75 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

76 Pereira, V. R., Junior, I. J. A., da Silveira, L. S., Geraldo, R. B., de F. Pinto, P., Teixeira, F. S., Salvadori, M. C., Silva, M. P., Alves, L. A., Capriles, P. V., et al. (2018). In vitro and in vivo antischistosomal activities of chalcones. *Chemistry & biodiversity*, 15(12):e1800398.

77 Phillips, J. L., Colvin, M. E., and Newsam, S. (2011). Validating clustering of molecular dynamics simulations using polymer models. *BMC bioinformatics*, 12(1):445.

78 Popat, S. K. and Emmanuel, M. (2014). Review and comparative study of clustering techniques. *International journal of computer science and information technologies*, 5(1):805–812.

79 Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., et al. (2013). Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854.

80 Robson, S. C., Sévigny, J., and Zimmermann, H. (2006). The e-ntpdase family of ectonucleotidases: structure function relationships and pathophysiological significance. *Purinergic signalling*, 2(2):409.

81 Rokach, L. (2009). A survey of clustering algorithms. In *Data mining and knowledge discovery handbook*, pages 269–298. Springer.

82 Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

83 Rzepiela, A. J., Schäfer, L. V., Goga, N., Risselada, H. J., De Vries, A. H., and Marrink, S. J. (2010). Reconstruction of atomistic details from coarse-grained structures. *Journal of computational chemistry*, 31(6):1333–1343.

84 Salvador, S., Chan, P., and Brodie, J. (2004). Learning states and rules for time series anomaly detection. In *FLAIRS conference*, pages 306–311.

85 Sansom, F. M. (2012). The role of the ntpdase enzyme family in parasites: what do we know, and where to from here? *Parasitology*, 139(8):963–980.

86 Sansom, F. M., Robson, S. C., and Hartland, E. L. (2008). Possible effects of microbial ecto-nucleoside triphosphate diphosphohydrolases on host-pathogen interactions. *Microbiol. Mol. Biol. Rev.*, 72(4):765–781.

87 Scherer, M. K., Trendelkamp-Schroer, B., Paul, F., Pérez-Hernández, G., Hoffmann, M., Plattner, N., Wehmeyer, C., Prinz, J.-H., and Noé, F. (2015). Pyemma 2: A software package for estimation, validation, and analysis of markov models. *Journal of chemical theory and computation*, 11(11):5525–5542.

88 Shao, J., Tanner, S. W., Thompson, N., and Cheatham, T. E. (2007). Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms. *Journal of Chemical Theory and Computation*, 3(6):2312–2334.

89 Shimizu, M. and Takada, S. (2018). Reconstruction of atomistic structures from coarse-grained models for protein–dna complexes. *Journal of chemical theory and computation*, 14(3):1682–1694.

90 Shiryayev, A. (2016). Probability, new york: Spri nger-verlag. *Multivariate Incomplete Failure Time Data by Modeling Marginal Dis*.

91 Skjaerven, L., Hollup, S. M., and Reuter, N. (2009). Normal mode analysis for proteins. *Journal of Molecular Structure: THEOCHEM*, 898(1-3):42–48.

92 Souza, V. C. d., Goliatt, L., and Goliatt, P. V. C. (2017). Clustering algorithms applied on analysis of protein molecular dynamics. In *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pages 1–6. IEEE.

93 Spyrakis, F., BidonChanal, A., Barril, X., and Javier Luque, F. (2011). Protein flexibility and ligand recognition: challenges for molecular modeling. *Current topics in medicinal chemistry*, 11(2):192–210.

94 Stamati, H., Clementi, C., and Kavraki, L. E. (2010). Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. *Proteins: Structure, Function, and Bioinformatics*, 78(2):223–235.

95 Teletin, M., Czibula, G., Albert, S., and Bocicor, M.-I. (2018). Using unsupervised learning methods for enhancing protein structure insight. *Procedia Computer Science*, 126:19–28.

96 Teodoro, M. L., Phillips Jr, G. N., and Kavraki, L. E. (2003). Understanding protein flexibility through dimensionality reduction. *Journal of Computational Biology*, 10(3-4):617–634.

97 Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

98 Torda, A. E. and van Gunsteren, W. F. (1994). Algorithms for clustering molecular dynamics configurations. *Journal of computational chemistry*, 15(12):1331–1340.

99 Tóth, G. and Borics, A. (2006). Flap opening mechanism of hiv-1 protease. *Journal of Molecular Graphics and Modelling*, 24(6):465–474.

100 Totrov, M. and Abagyan, R. (2008). Flexible ligand docking to multiple receptor conformations: a practical alternative. *Current opinion in structural biology*, 18(2):178–184.

101 Tribello, G. A. and Gasparotto, P. (2019). Using dimensionality reduction to analyze protein trajectories. *Frontiers in molecular biosciences*, 6:46.

102 Van Eldik, L., Van Eldik, L. J., and Watterson, D. M. (1998). *Calmodulin and signal transduction*. Gulf Professional Publishing.

103 Van Gunsteren, W. F. and Berendsen, H. J. (1990). Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. *Angewandte Chemie International Edition in English*, 29(9):992–1023.

104 Vasconcelos, E. G., Ferreira, S. T., De Carvalho, T. M., De Souza, W., Kettlun, A. M., Mancilla, M., Valenzuela, M. A., and Verjovski-Almeida, S. (1996). Partial purification and immunohistochemical localization of atp diphosphohydrolase from schistosoma mansoni immunological cross-reactivities with potato apyrase and toxoplasma gondii nucleoside triphosphate hydrolase. *Journal of Biological Chemistry*, 271(36):22139–22145.

105 Vasconcelos, E. G., Nascimento, P. S., Nazareth, M., Meirelles, L., Verjovski-Almeida, S., and Ferreira, S. T. (1993). Characterization and localization of an atp-diphosphohydrolase on the external surface of the tegument of schistosoma mansoni. *Molecular and biochemical parasitology*, 58(2):205–214.

106 Verli, H. (2014). *Bioinformática da Biologia à Flexibilidade Molecular*. Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq, São Paulo.

107 Walsh, M. P. (1994). Calmodulin and the regulation of smooth muscle contraction. *Molecular and cellular biochemistry*, 135(1):21–41.

108 Wang, Y., Harrison, C. B., Schulten, K., and McCammon, J. A. (2011). Implementation of accelerated molecular dynamics in namd. *Computational science & discovery*, 4(1):015002.

109 Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

110 Weinberger, K. Q. and Saul, L. K. (2006). Unsupervised learning of image manifolds by semidefinite programming. *International journal of computer vision*, 70(1):77–90.

111 Werner, T., Morris, M. B., Dastmalchi, S., and Church, W. B. (2012). Structural modelling and dynamics of proteins for insights into drug interactions. *Advanced drug delivery reviews*, 64(4):323–343.

112 Whitford, P. C. and Onuchic, J. N. (2015). What protein folding teaches us about biological function and molecular machines. *Current opinion in structural biology*, 30:57–62.

113 WHO (2020). Who data show unprecedented treatment coverage for bilharzia and intestinal worm. Acesso online em 10 de Fevereiro de 2020.

114 Woods, A., Dickerson, K., Heath, R., Hong, S.-P., Momcilovic, M., Johnstone, S. R., Carlson, M., and Carling, D. (2005). Ca2+/calmodulin-dependent protein kinase kinase-β acts upstream of amp-activated protein kinase in mammalian cells. *Cell metabolism*, 2(1):21–33.

115 Yang, F., Hicks-Berger, C. A., Smith, T. M., and Kirley, T. L. (2001). Site-directed mutagenesis of human nucleoside triphosphate diphosphohydrolase 3: the importance of residues in the apyrase conserved regions. *Biochemistry*, 40(13):3943–3950.

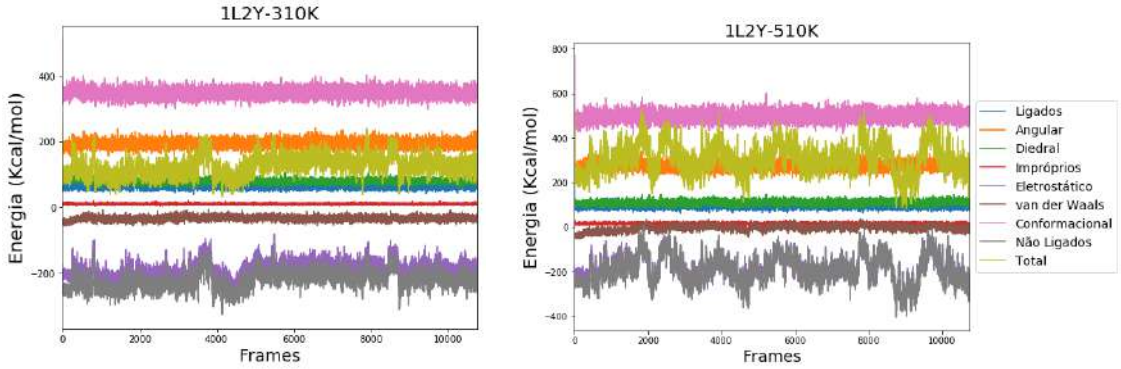
116 Zebisch, M., Krauss, M., Schäfer, P., Lauble, P., and Sträter, N. (2013). Crystallographic snapshots along the reaction pathway of nucleoside triphosphate diphosphohydrolases. *Structure*, 21(8):1460–1475.

117 Zebisch, M. and Sträter, N. (2008). Structural insight into signal conversion and inactivation by ntpdase2 in purinergic signaling. *Proceedings of the National Academy of Sciences*, 105(19):6882–6887.

118 Zhao, Q., Hautamaki, V., and Fränti, P. (2008). Knee point detection in bic for detecting the number of clusters. In *International conference on advanced concepts for intelligent vision systems*, pages 664–673. Springer.

Análises da Dinâmica Molecular da 1L2Y

Na figura .0.1 as diferentes componentes de energia potencial são comparadas entre as simulações realizadas em 310K e 510K para 1L2Y. De maneira geral, observou-se que para as simulações à 310K a energia potencial total foi ligeiramente menor em relação a em 510K, o que pode ser explicado pela introdução de energia no sistema com o aumento de temperatura. Além disso, o perfil "ocilatório" das energias dos átomos não-ligados e eletrostática indicam que a proteína em alguns momentos aumenta os contatos internos diminuindo em outros, o que sugere possível compactação e descompactação da mesma.



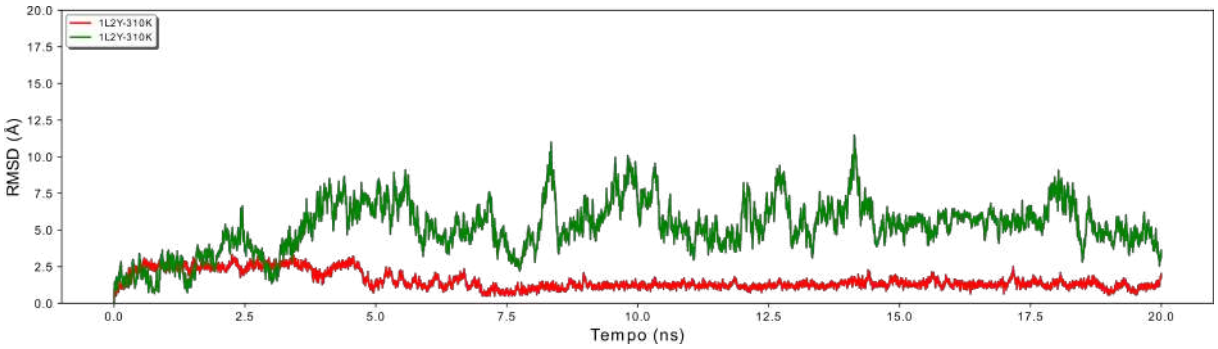
(a) 1L2Y-310K (b) 1L2Y-510K

Figura .0.1 – Flutuação das energias potenciais para simulações da 1L2Y. As energias calculadas com o plugin NAMD-energy dentro do programa VMD [1].

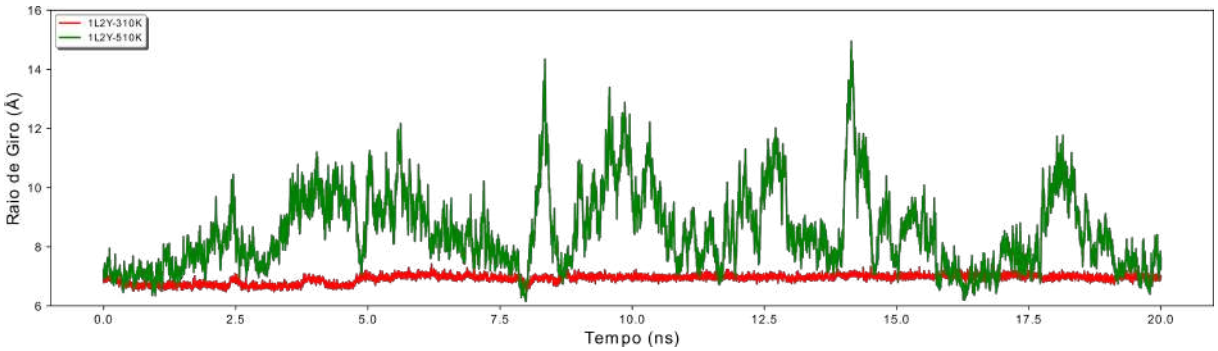
O *Root mean square deviation* (RMSD) é uma métrica de avaliação da similaridade entre duas estruturas tridimensionais a partir do cálculo das distâncias interatômicas em angstrons (Å), sendo amplamente usado para estudos de mudanças conformacionais. Para proteínas pequenas, com aproximadamente 150 aminoácidos, valores de RMSD menores que 3Å são considerados bem informativos e estruturas dentro desta faixa podem ser interessantes para estudos de mecanismos catalíticos. Valores entre 3Å e 5Å refletem estruturas menos informativas porém aceitáveis para correlação entre similaridade estrutural e função, bem como também, análise de possíveis motivos proteicos (40). Já valores acima de 5Å representam estruturas não informativas para estudos de função, apesar de fornecerem dados importantes para dinâmica de enovelamento estrutural das proteínas, especialmente em estudos analisando simulações em diferentes ambientes (40).

Nas Figuras 2(a) e 2(b) são mostradas as variações do RMSD e raio de giro, respectivamente, para as simulações nas temperaturas de 310K e 510K. De acordo com os resultados apresentados, observa-se que para simulação à 310K os valores de RMSD e raio de giro apresentam pouca variação, sendo o valor médio de ?Å, indicando que a proteína sofre pouca mudança estrutural. Estes resultados são opostos aqueles obtidos para simulação em 510K, na qual há grande flutuação nessas métricas.

Interessantemente, os valores de RMSD e raio de giro mostrados nas figuras 2(a) e 2(b) apresentam um compostamento de periodicidade e ao, final do tempo de simulação analisado, voltam a variar em uma faixa de valores próximos ao da simulação em 310K. Isto parece informar que ao final da simulação a proteína tende ao enovelamento novamente, após sua desnaturação. Tal fato pode ser corroborado com os valores de SASA mostrados na figura .0.3. 7.5ns tendem a estabilizarem.



(a) RMSD



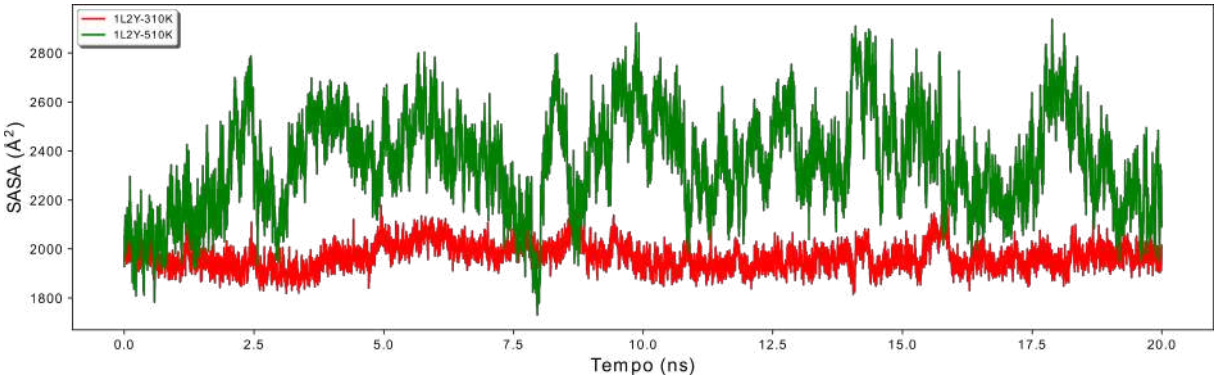
(b) Raio de Giro

Figura .0.2 – Os gráficos 2(a) e 2(b) referem-se a flutuação dos valores de RMSD e raio de giro, respectivamente, para a proteína 1L2Y sob temperaturas de 310K e 510K.

A figura .0.3 apresenta a variação do SASA para simulaçãoes em 310K e 510K da 1L2Y. Este valor O SASA é uma medida importante para avaliar a exposição dos resíduos ao meio que a proteína está sendo simulada e pode informar também juntamente com outras métricas (RMSD e raio de giro) o processo de desnaturação. Observa-se que na simulação à 310K, não ocorrem grandes flutuações no acessibilidade ao solvente, o que juntamente com os valores de RMSD e raio de giro mostram que a proteína mantêm a estabilidade estrutural. Quanto a simulação à 510K, observou-se que a proteína sofre ciclos de menor e maior SASA, os quais estão relacionados com a alternância entre estados conformacionais com diferentes graus de compactação estrutural, conforme indicado pelos valores de raio de giro.

A fim de uma análise mais refinada, envolvendo informações sobre o deslocamento dos resíduos à diferentes temperaturas, foi calculado o RMSF. Esta métrica pode ser compreendida como informação complementar ao RMSD e Rg, já que estes representam

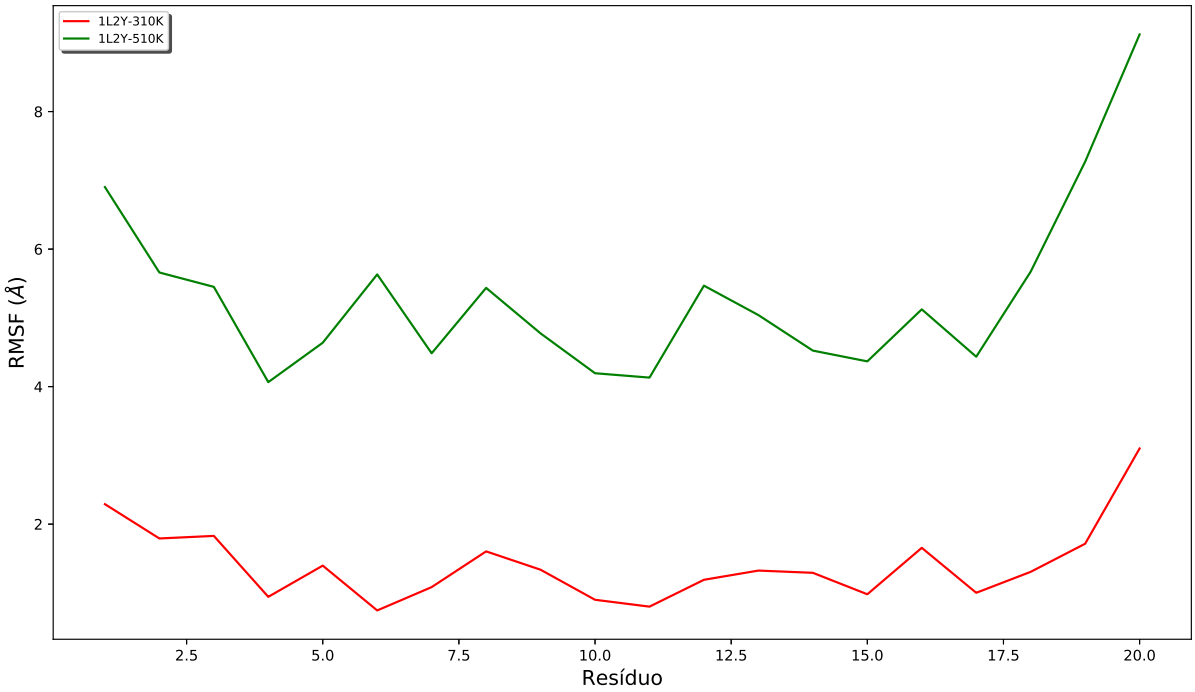
Figura .0.3 – Variação da exposição dos resíduos aminoácidos ao longo das simulações.



No gráfico acima é apresentado a flutuação do valor de área de superfície acessível ao solvente (SASA) para a proteína 1L2Y nas diferentes temperaturas.

mudanças globais ou de partes de uma proteína. Na Figura .0.4 é mostrado os resultados de RMSF obtidos aqui. De forma geral, observa-se que os resíduos apresentam menor flexibilidade em simulações à 310K comparado a 510K, isto pode ser explicado pelo aumento da energia cinética no sistema ocasionado pela temperatura elevada, conforme sugerido anteriormente.

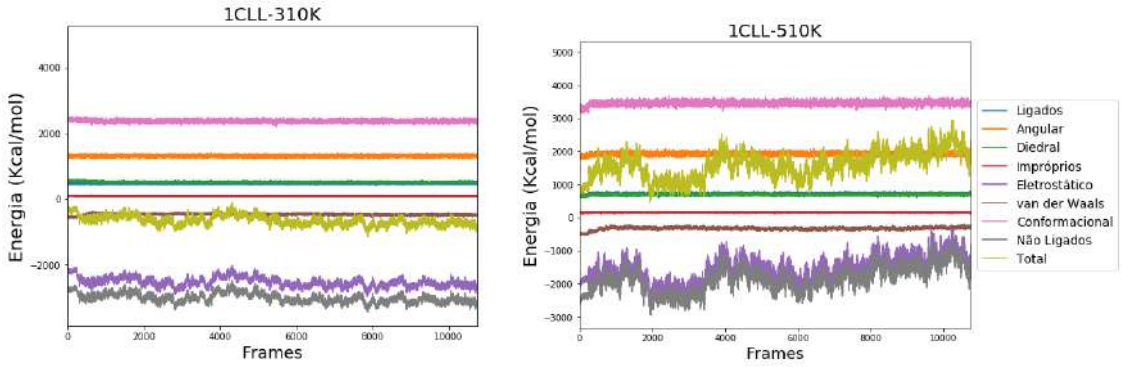
Figura .0.4 – Variação do RMSF ao longo das simulações



Os gráficos A e B referem-se a flutuação dos deslocamentos dos resíduos para 1CLL em 310K e 510K, respectivamente. Enquanto que em C e D são mostrados os deslocamentos dos resíduos aminoácidos para 1L2Y em ambas temperaturas.

Análises da Dinâmica Molecular da 1CLL

Na figura .0.1 as diferentes componentes de energia potencial são comparadas entre as simulações realizadas em 310K e 510K para 1CLL. Assim como observado nas simulações para a proteína 1L2Y, os valores das energias obtidos aqui indicam que na temperatura de 310K a energia potencial total é menor em relação a 510K, já que houve inserção de energia ao sistema. Porém, o perfil das energias dos átomos não-ligados e eletrostática em 510K não apresenta a "ocilação"vista para 1L2Y na mesma temperatura. Ao contrário, oberva-se que em 510K essas energias aumentam lentamente ao longo da simulação, indicando que a 1CLL tende manter ou aumentar sua compactação.



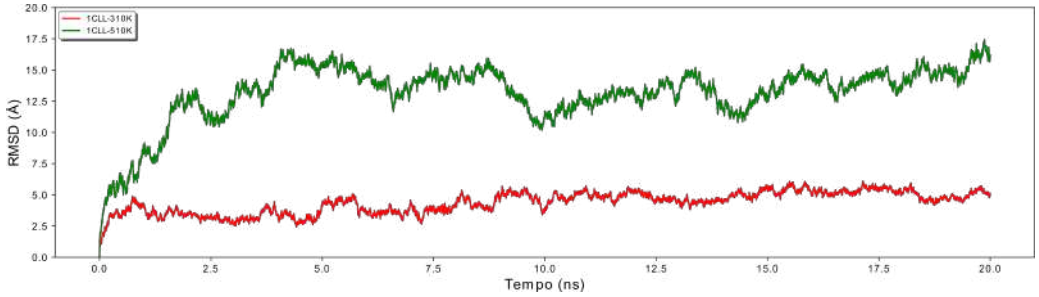
(a) 1CLL-310K (b) 1CLL-510K

Figura .0.1 – Flutuação das energias potenciais para simulações da 1CLL. As energias calculadas com o plugin NAMD-energy dentro do programa VMD[1].

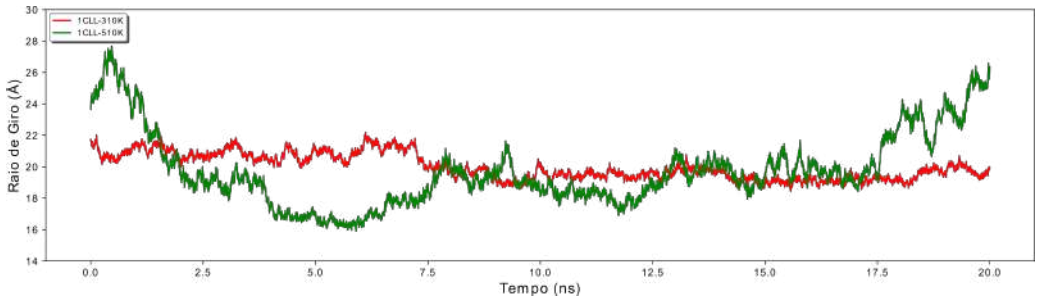
Na figura 2(a) é apresentada a variação dos valores de RMSD nas simulações em temperaturas de 310K e 510K. De acordo com os resultados apresentados na figura 2(a), observa-se que para simulação à 310K os valores de RMSD apresentam pouca variação, com valor médio de RMSD acima de 12.5Å, indicando que a proteína sofre pouca variação estrutural mesmo após a remoção dos íons de cálcio. O comportamento o oposto é verificado para a simulação em 510K, na qual observamos que há flutuações na estrutura da proteína, similar ao que foi verificado para 1L2Y.

Analisando as variações do raio de giro (figura 2(b)), observa-se que a compacidade da proteína altera muito pouco em relação ao seu centro de massa, nas simulações em 310K, mostrando que a mesma se mantém estável durante a simulação. Por outro lado, os resultados foram ligeiramente diferentes para simulação à 510K, na qual a proteína apresenta grande flutuação conformacional. Nessa temperatura, os valores de raio de giro, em boa parte da simulação, flutuam em faixas abaixo daqueles vistos em 310K, perminto sugerir que a proteína tende a se compactar ao longo da trajetória analisada.

observado é que apesar dos valores de Rg aumentarem inicialmente eles sofrem uma queda após 1.25ns o que parece estar relacionado ao processo de compactação da



(a) RMSD

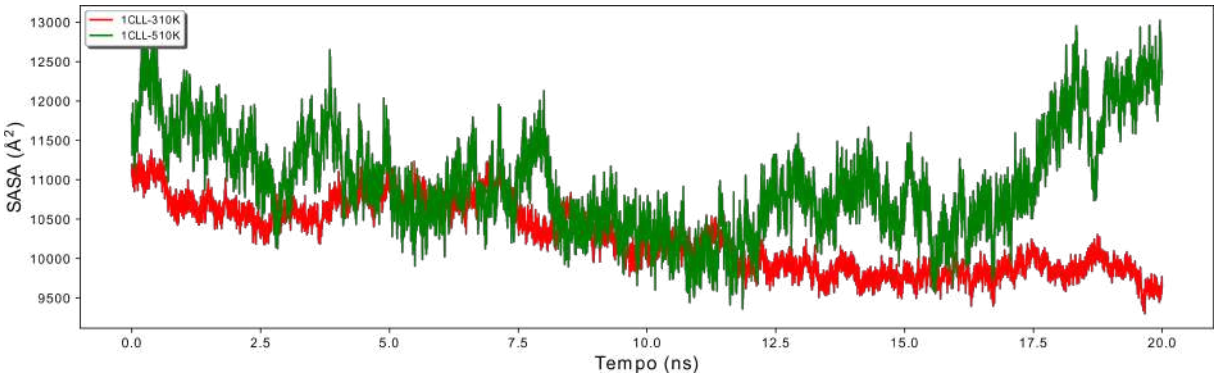


(b) Raio de Giro

Figura .0.2 – Os gráficos 2(a) e 2(b) referem-se a flutuação dos valores de RMSD e raio de giro, respectivamente, para a proteína 1CLL sob temperaturas de 310K e 510K.

proteína, esses valores voltam a aumentar apenas após 17.5ns de simulação.

Figura .0.3 – Variação da exposição dos resíduos aminoácidos ao longo das simulações.



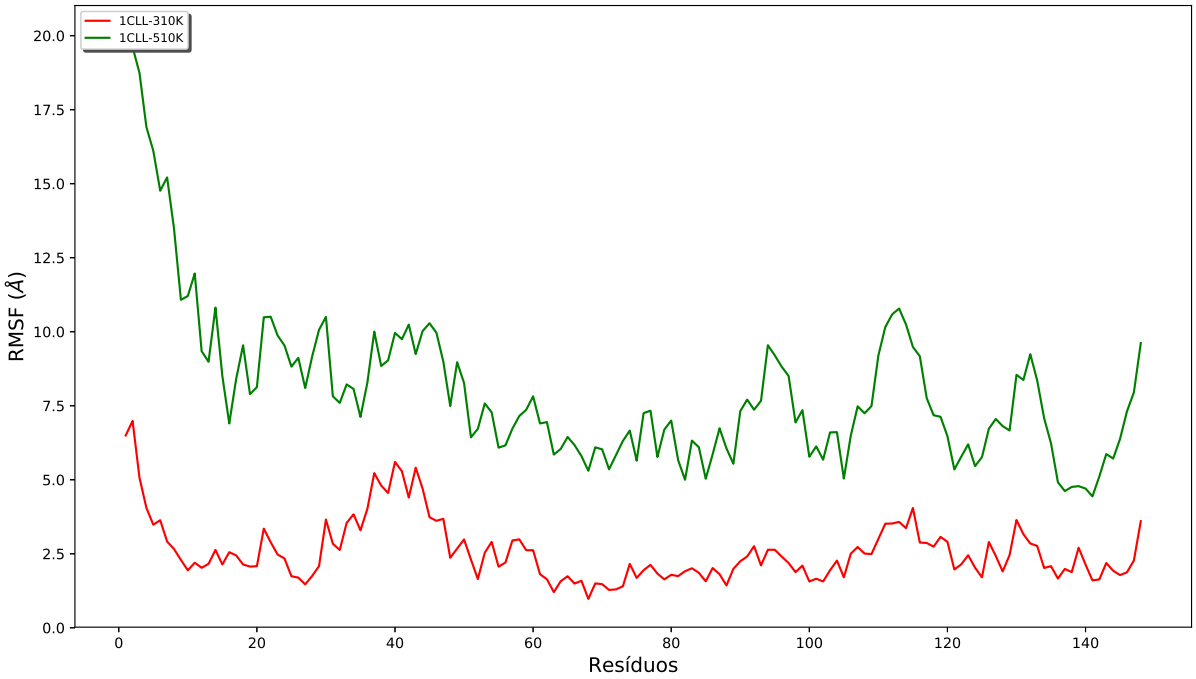
No gráfico acima é apresentado a flutuação do valor de área de superfície acessível ao solvente (SASA) para a proteína 1CLL nas diferentes temperaturas.

Os valores de SASA e RMSF foram calculados afim de corroborar o que foi observado a partir dos valores de RMSD e raio de giro. Conforme análise da variação dos valores de SASA (figura .0.3), concluimos que a proteína tende a assumir uma geometria tridimensional mais compacta, de forma que mesmo com aumento da temperatura os valores de acessibilidade ao solvente em 310K e 510K são relativamente próximos, salvo ao final das simulações onde divergem entre si. Quanto aos valores de RMSF (figura .0.4), verificamos que em média os valores em 510K foi maior que 310K mostrando que a proteína

possui grande flutuação e não somente nas reegiões dos lóbulos de interação com cálcio.

RMSF. Esta métrica pode ser compreendida como informação complementar ao RMSD e Rg, já que estes representam mudanças globais ou de partes de uma proteína.

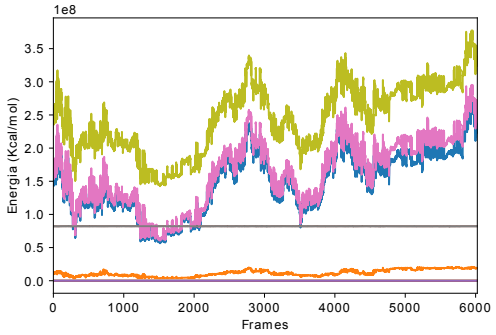
Figura .0.4 – Variação do RMSF ao longo das simulações



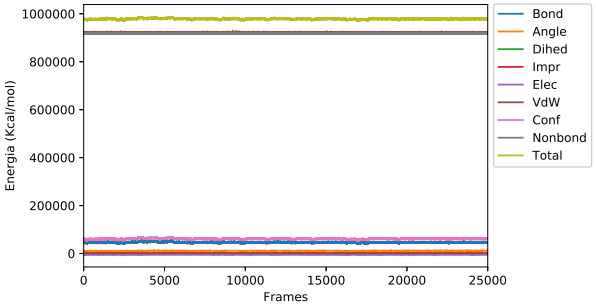
Os gráficos A e B referem-se a flutuação dos deslocamentos dos resíduos para 1CLL em 310K e 510K, respectivamente.

Análise das Simulações Referentes a SmNTPDase 1

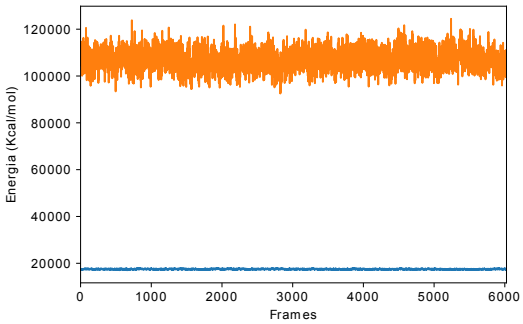
.1 Simulações sem ANP



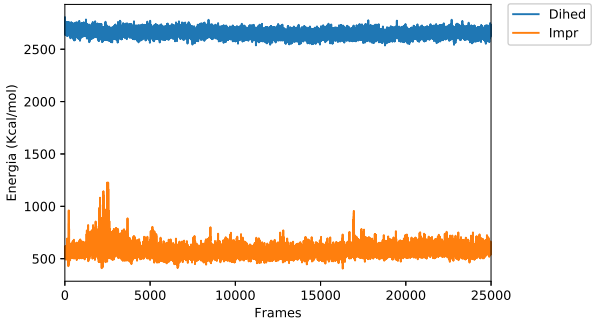
(a) sm1-cMD



(b) sm1-aMD



(c) sm1-cMD



(d) sm1-aMD

Figura .1.1 – Flutuação das energias potenciais para simulações para enzima smNTPDase 1 ao longo das simulações sem ANP. Em (a) e (b) são apresentadas as energias totais obtidas para a simulação cMD e aMD, respectivamente. Enquanto que nas figuras (c) e (d) são mostradas em ordem, para melhor observação, as energias potenciais diedrais (próprio em azul e impróprio em laranja), ao longo das simulações cMD e aMD. As energias calculadas com o plugin NAMD-energy dentro do programa VMD[1] considerando apenas os átomos referentes à proteína.

Na figura .1.1 são apresentadas as variações energias potenciais total (figuras 1(a) e 1(b)) e diedrais (figuras 1(c) e 1(d)) dos átomos da proteína smNTPDase1 ao longo das simulações convencional e acelerada, sem a presença do ligante ANP no sítio catalítico. Observamos que ao longo dos 250 nanossegundos(ns) de simulação convencional há uma grande flutuação da energia total da proteína, devido variações na energia conformacional (soma das energias dos átomos ligados), mais especificamente associada aos ângulos diedrais próprios e impróprios. Quando analisadas as mesmas energias ao longo dos 50ns de simulação acelerada, as variações foram menores, indicando que a proteína manteve-se mais estável mesmo com a adição de potenciais.

Conforme apresentado na figura .1.2, tanto o domínio ECD quanto as transs-membranas sofreram maiores variações de RMSD na simulação convencional do que na

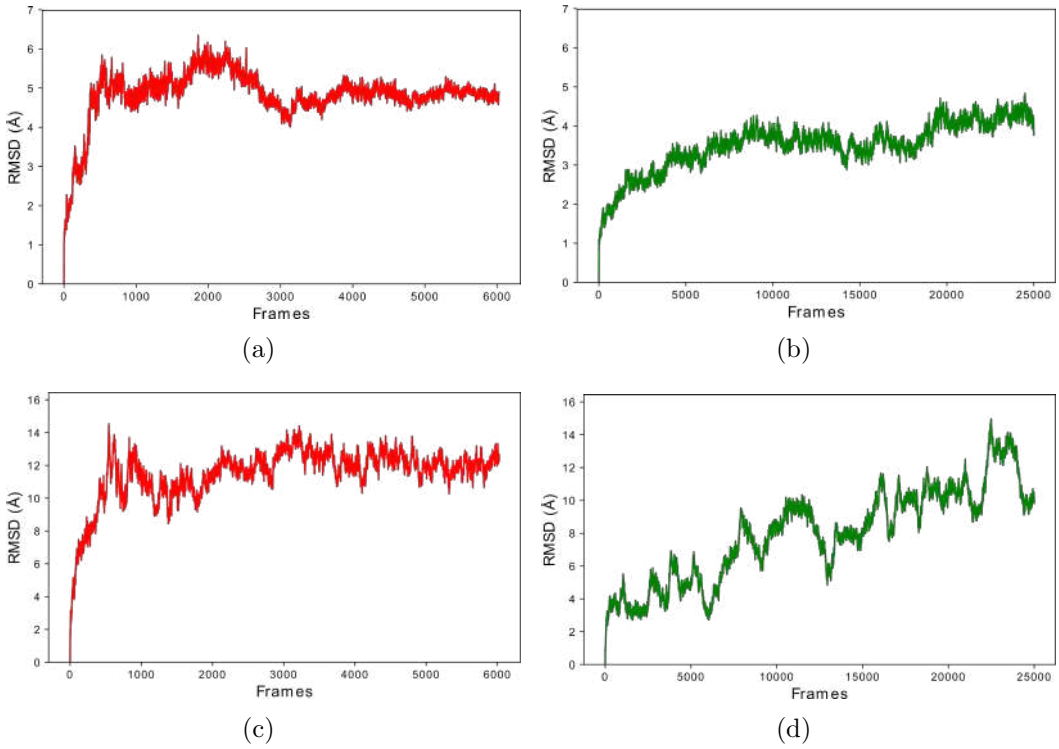


Figura .1.2 – Flutuação dos valores de RMSD das regiões ECD e TM em simulações sem ligante da smNTPDase1. Em **??** e **??** são apresentados os valores de RMSD da região ECD ao logo das simulações cMD e aMD, repectivamente. Já em **??** e (d) são mostradas flutuações do RMSD para as regiões transmembranares ao paras as trajetórias das simulações cM e aMD, respectivamente.

acelerada, apontando que de fato há maiores flutuações estruturais na primeira. Além disso, embora os valores de raio de giro (figura .1.3) do domínio ECD não foi tão diferente entre as duas simulações, porém as transmembranas tiveram maior flutuação na rotação em relação ao centro de massa na simulação cMD. Este resultado indica que nas simulações aMD a estrutura se estabiliza e relaxa mais rápido.

.2 Simulações com ANP

Na figuras .2.1 são apresentadas as variações das energias potenciais total e diedrais (próprio e impróprio) dos átomos da proteína smNTPDase1 ao longo das simulações convencional e acelerada na presença do ligante ANP no sítio catalítico. Diferente das simulações anteriores, observamos aqui que, tanto na simulação convencional quanto na acelerada, as energias potenciais apresentam comportamento mais estável. Este comportamento indica que não ocorrem grandes flutuações conformacionais na proteína, sugerindo que o ligante estabiliza a sua estrutura como sugerido pelo modelo de encaixe induzido.

A análise dos valores de RMSD para as regiões ECD e TM para ambas trajetórias cMD e aMD (figura .2.2), mostrou que, mesmo não ocorrendo grandes mudanças estru-

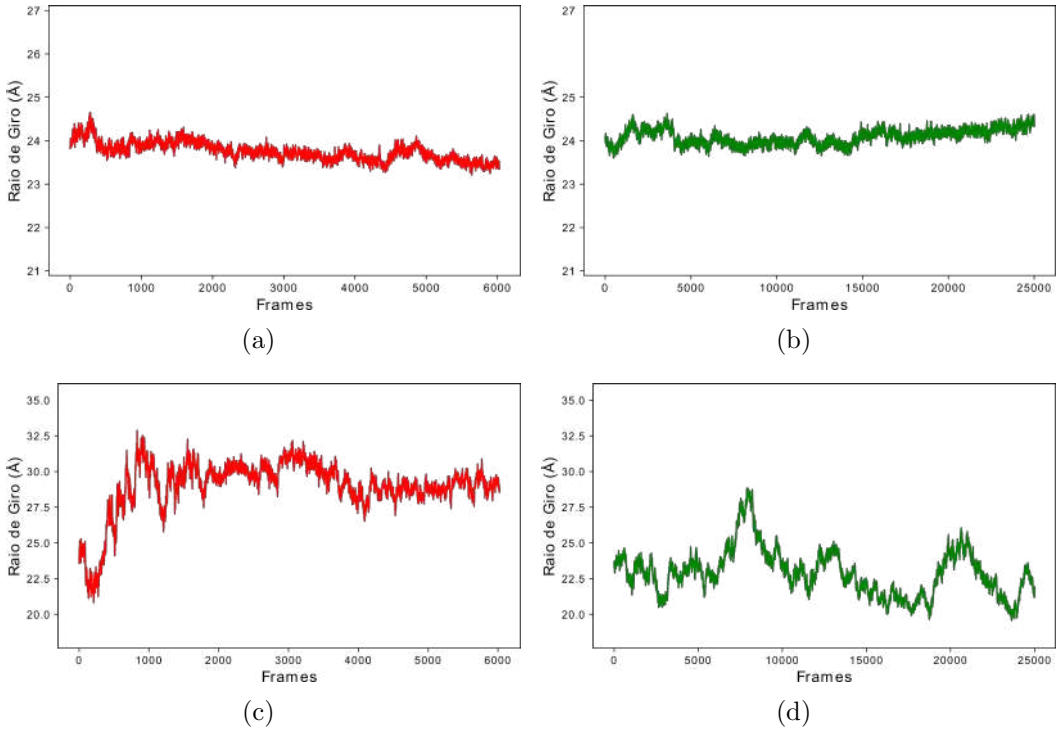
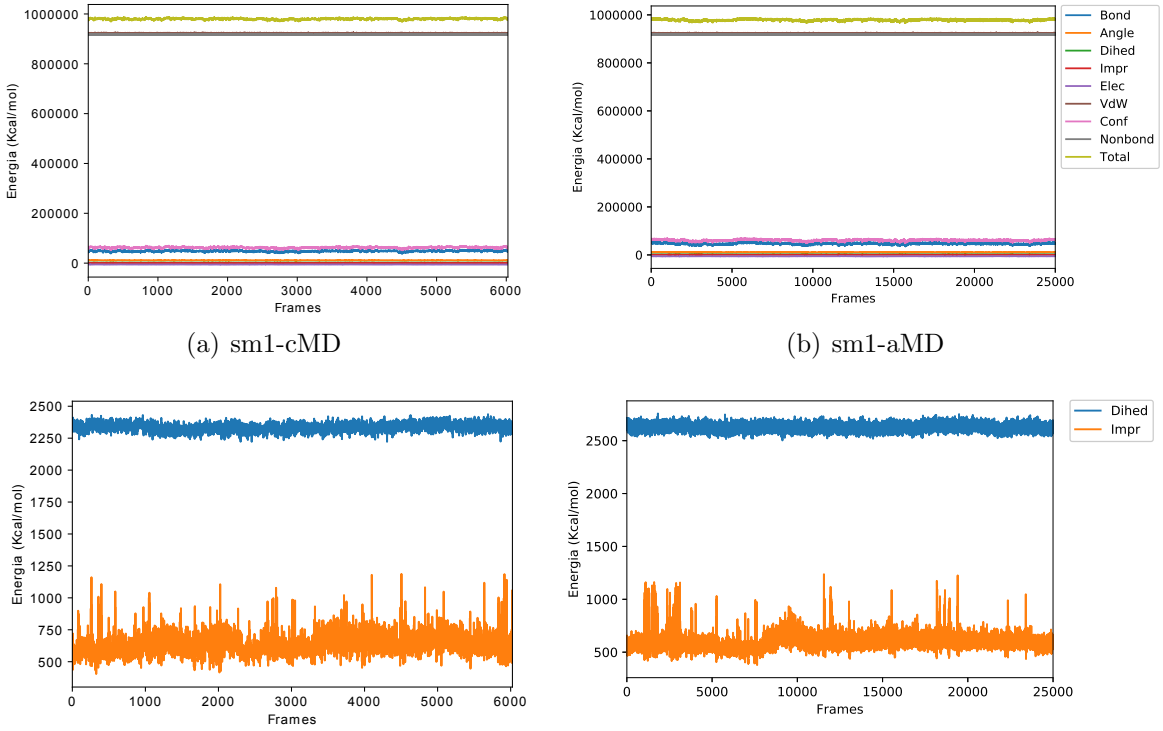


Figura .1.3 – Flutuação dos valores de raio de giro regiões ECD e TM em simulações sem ligante da smNTPDase1. Em (a) e (b) são apresentados os valores de raio de giro da região ECD ao logo das simulações cMD e aMD, repectivamente. Já em (c) e (d) são mostradas flutuações do raio de giro para as regiões transmembranares ao paras as trajetórias das simulações cM e aMD, respectivamente.

turais na proteína ao longo das simulações, esses domínios assumem maiores variações conformacionais nos 250ns de simulação convencional. Contudo, as variações nos valores de RMSD das TM apresentam um comportamento similar nas duas simulações, apontando mais uma vez para a conexão dos movimentos das hélices transmembranares atuarem sobre o movimento do domínio extracélular. Isto pode ser melhor evidenciado observando-se os comportamentos similares das flutuações do raio de giro (figura .2.3)



(a) sm1-cMD

(b) sm1-aMD

(c) sm1-cMD

(d) sm1-aMD

Figura .2.1 – Flutuação das energias potenciais para simulações para enzima smNTPDase 1 ao longo das simulações com ANP. Em (a) e (b) são apresentadas as energias totais obtidas para a simulação cMD e aMD, respectivamente. Enquanto que nas figuras (c) e (d) são mostradas em ordem, para melhor observação, as energias potenciais diedrais (próprio em azul e impróprio em laranja), ao longo das simulações cMD e aMD. As energias calculadas com o plugin NAMD-energy dentro do programa VMD[2] considerando apenas os átomos referentes à proteína.

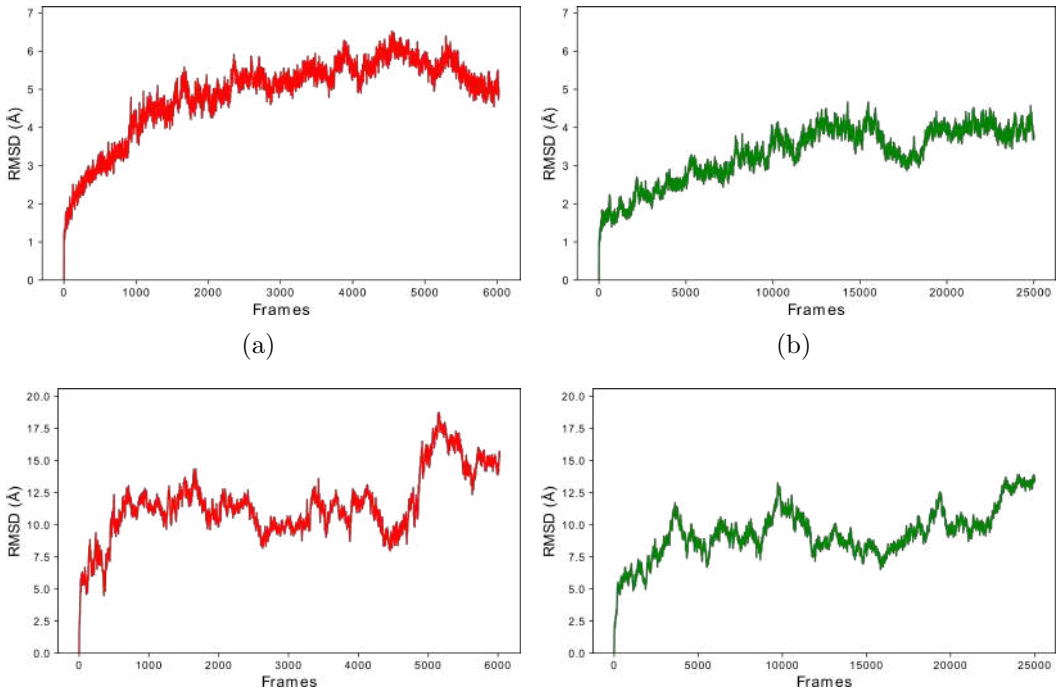


Figura .2.2 – Flutuação dos valores de RMSD das regiões ECD e TM em simulações com ligante da smNTPDase1. Em (a) e (b) são apresentados os valores de RMSD da região ECD ao logo das simulações cMD e aMD, repectivamente. Já em (c) e (d) são mostradas flutuações do RMSD para as regiões transmembranares ao paras as trajetórias das simulações cM e aMD, respectivamente.

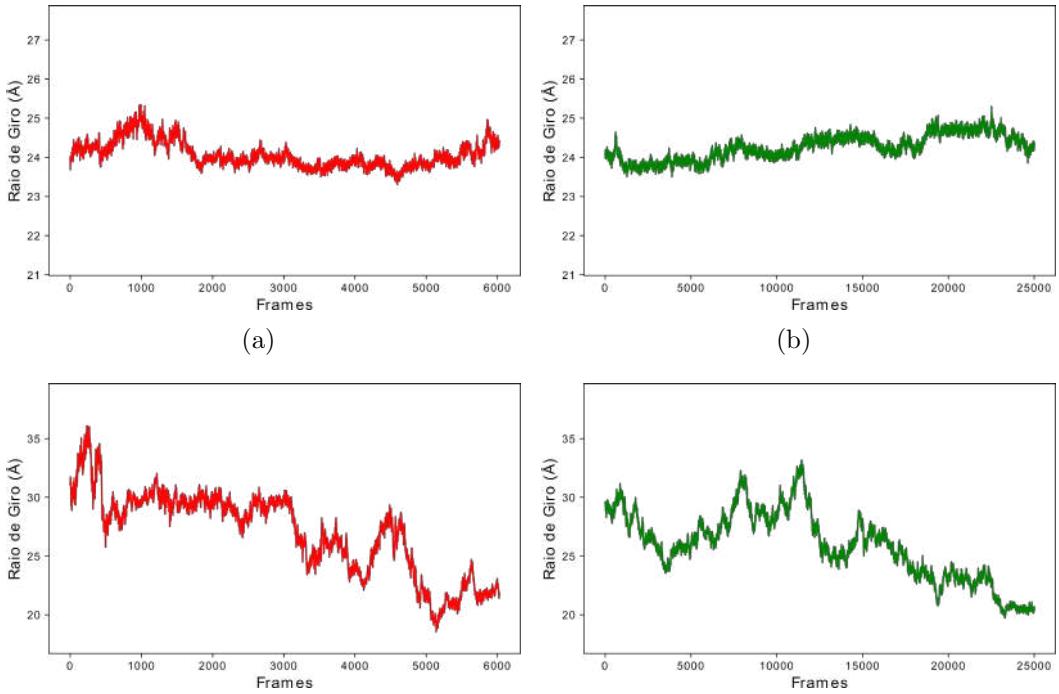
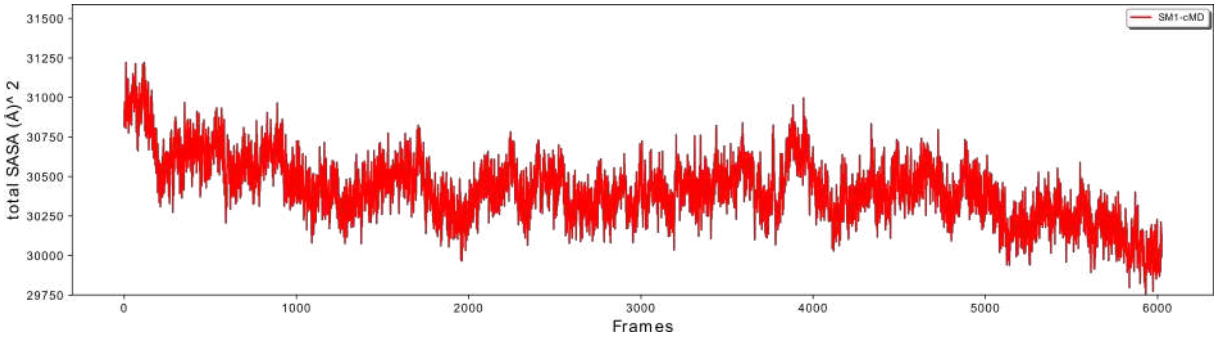
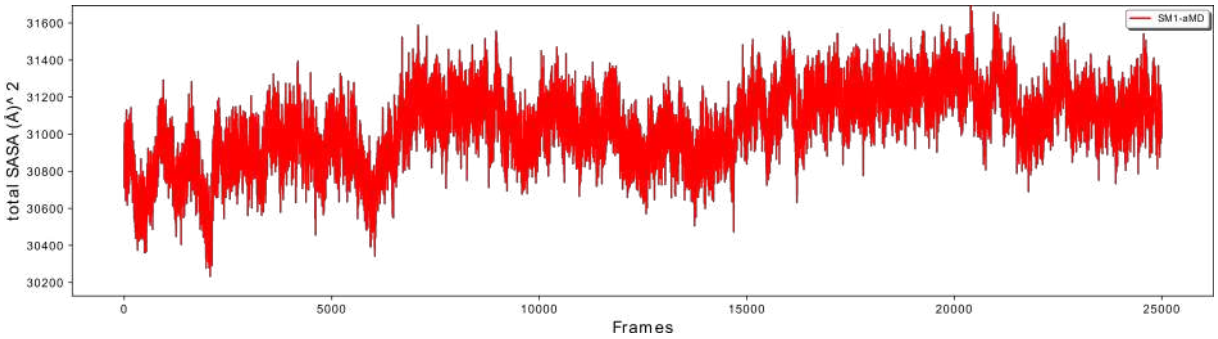


Figura .2.3 – Flutuação dos valores de raio de giro regiões ECD e TM em simulações com ligante da smNTPDase1. Em (a) e (b) são apresentados os valores de raio de giro da região ECD ao logo das simulações cMD e aMD, repectivamente. Já em (c) e (d) são mostradas flutuações do raio de giro para as regiões transmembranares ao paras as trajetórias das simulações cM e aMD, respectivamente.

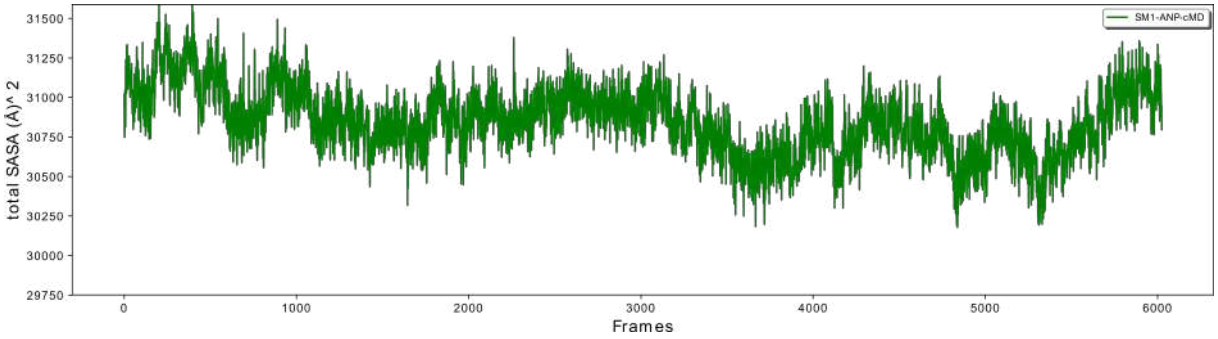


(a)

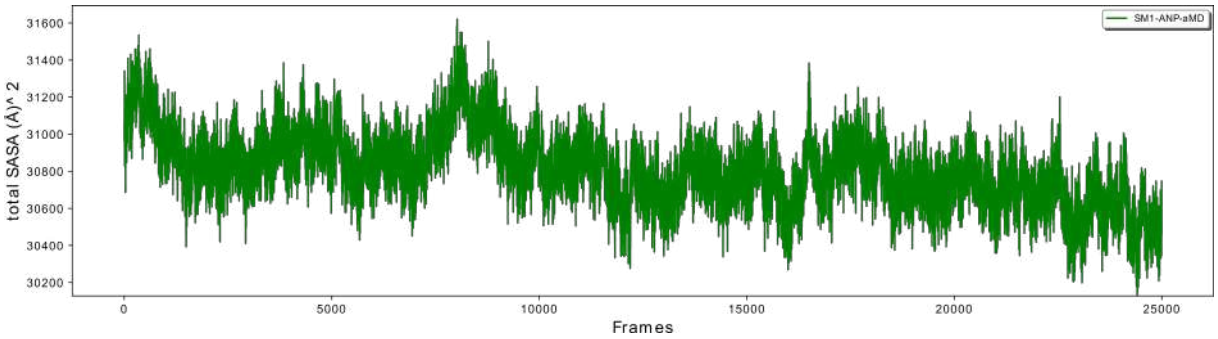


(b)

Figura .2.4 – Flutuação dos valores de SASA da proteína smNTPDase1 ao longo das simulações em ligante. A subfigura (a) refere-se as variações na simuação cMD, enquanto que (b) representa as mudanças de SASA na simulação aMD.



(a)



(b)

Figura .2.5 – Flutuação dos valores de SASA da proteína smNTPDase1 ao longo das simulações com ligante. A subfigura (a) refere-se as variações na simuação cMD, enquanto que (b) representa as mudanças de SASA na simulação aMD.

Análise das Simulações Referentes a SmNTPDase 2

.1 Simulações sem AU1

Na figura 1(b) são apresentadas as variações energias potenciais total e diedrais (próprio e impróprio) dos átomos da proteína smNTPDase2 ao longo das simulações convencional e acelerada sem a presença do ligante AU1 no sítio catalítico. Em geral, as energias total e diedrais apresentaram flutuações estáveis, isto é, dentro de um valor médio e sem grandes flutuações. Contudo, observamos que os valores de energia conformacional (soma das energias dos referentes aos átomos ligados) foram maiores que nas outras energias.

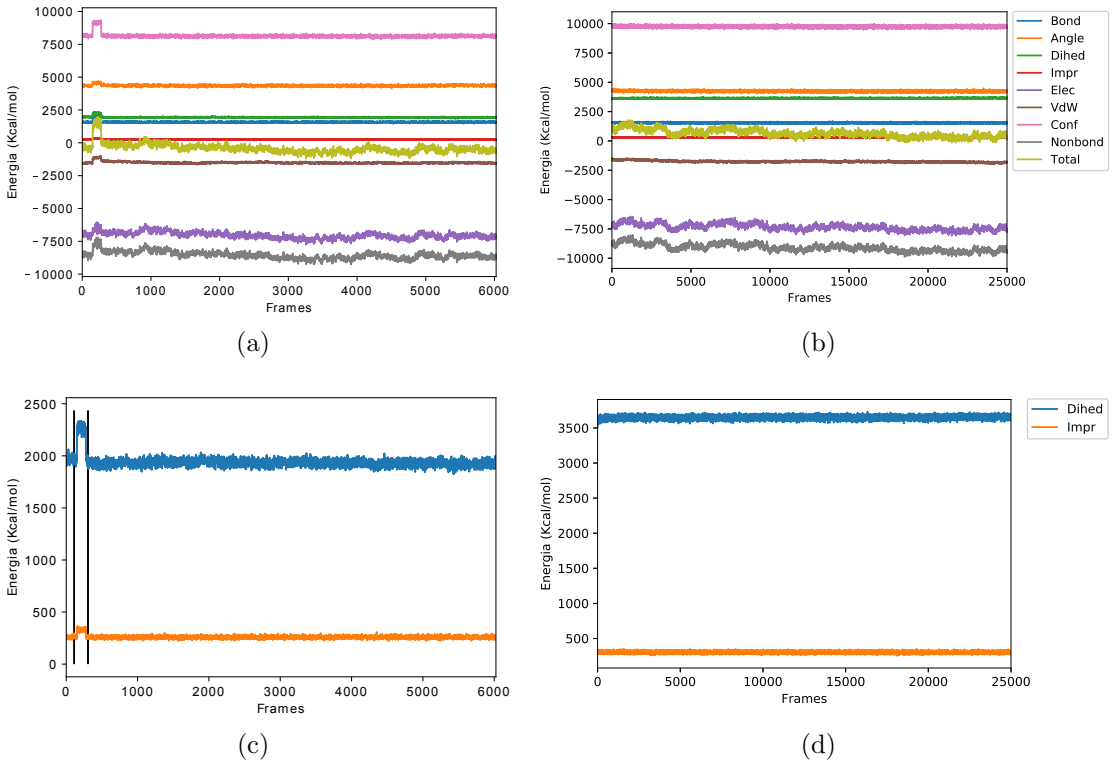


Figura .1.1 – Flutuação das energias potenciais para simulações para enzima smNTPDase 2 ao longo das simulações sem AU1. Em (a) e (b) são apresentadas as energias totais obtidas para a simulação cMD e aMD, respectivamente. Enquanto que nas figuras (c) e (d) são mostradas em ordem, para melhor observação, as energias potenciais diedrais (próprio em azul e impróprio em laranja), ao longo das simulações cMD e aMD. As energias calculadas com o plugin NAMD-energy dentro do programa VMD[1] considerando apenas os átomos referentes à proteína.

Os resultados de energia, acima abordados, indicam que a proteína smNTPDase2 apresenta variação em sua estrutura tridimensional tanto na simulação convencional quanto acelerada. Curiosamente, para a simulação convencional, observou-se picos nas energias entre os frames 110 e 305 (figura 1(c)), os quais parecem estar associados ao processo de enovelamento da porção N-terminal e abertura da cavidade catalítica da enzima. Embora

o enovelamento e acomodamento da porção N-terminal também tenha ocorrido ao final da simulação acelerada, não foi observadas abertura da cavidade catalítica nem a ocorrência de picos de energias.

Comparando-se a flutuação dos valores de RMSD e raio de giro (figura .1.2), os resultados indicam que, muito embora os valores de energia sejam próximos, a enzima sofre maiores variações conformacionais na simulação convencional do que na acelerada. Esse fato pode estar associado ao fato de que na simulação acelerada, com a aplicação do potencial de *boost* nos ânglulos diedrais, a proteína atinge estados de maior compactação e relaxamento mais rápido como pode ser evidenciado nas figuras 2(c) e 2(d).

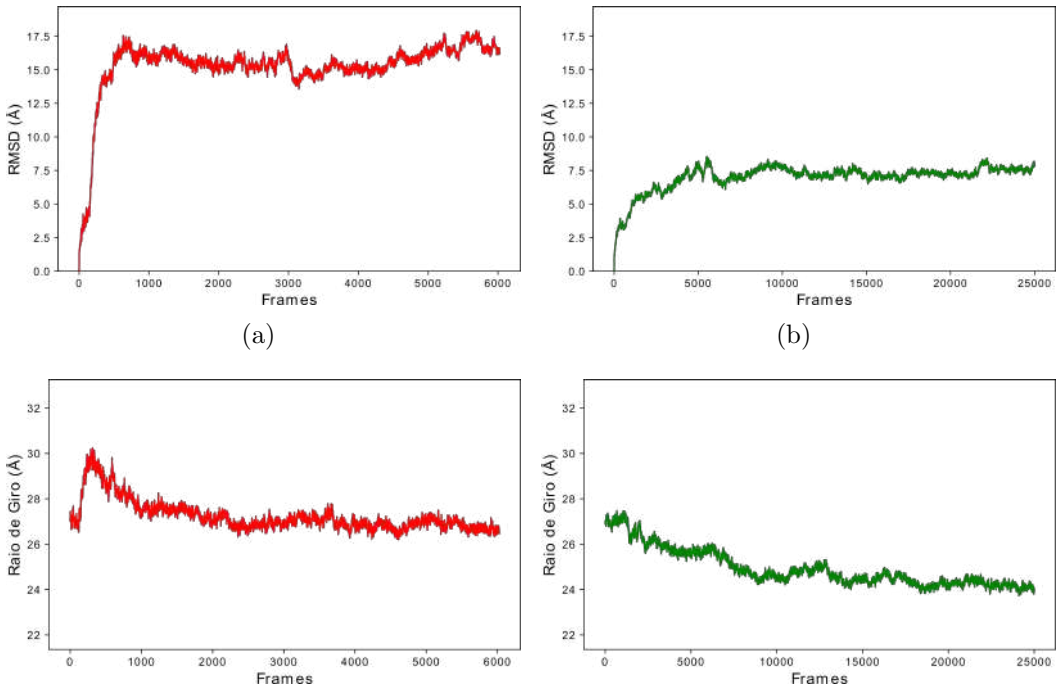


Figura .1.2 – Flutuação dos valores de RMSD e raio de giro da enzima smNTPDase 2 em simulações sem AU1. Em (a) e (b) são apresentados, em ordem, os valores de RMSD das simulações cMD e aMD. Já em (c) e (d) são mostradas flutuações do raio de giro ao longo das simulações cM e aMD, respectivamente.

.2 Simulações com AU1

Quanto as análises das energias para as simulações da smNTPDase2 com a presença do ligante AU1 no sítio catalítico, assim como observado para as simulações anteriores, não as energias potenciais total e diedrais não apresentaram grandes flutuações (figuras 1(a) e 1(b)). Assim como visto na simulação convencional, também foi observado picos nas energias entre os frames 110 e 305 que parecem estar associados aos movimentos da porção N-terminal e abertura da região catálitica para relaxamento do domínio referente ao ”ECD”nas NTPDases com transmembranas. Estes movimentos de abertura do sítio

foi observada em ambas simulações, porém não ocorreram picos nas enegias na simulação aMD. Além disso, devido tais movimentos, o ligante deixou de interagir com os resíduos catalítico e saiu do sítio de interação.

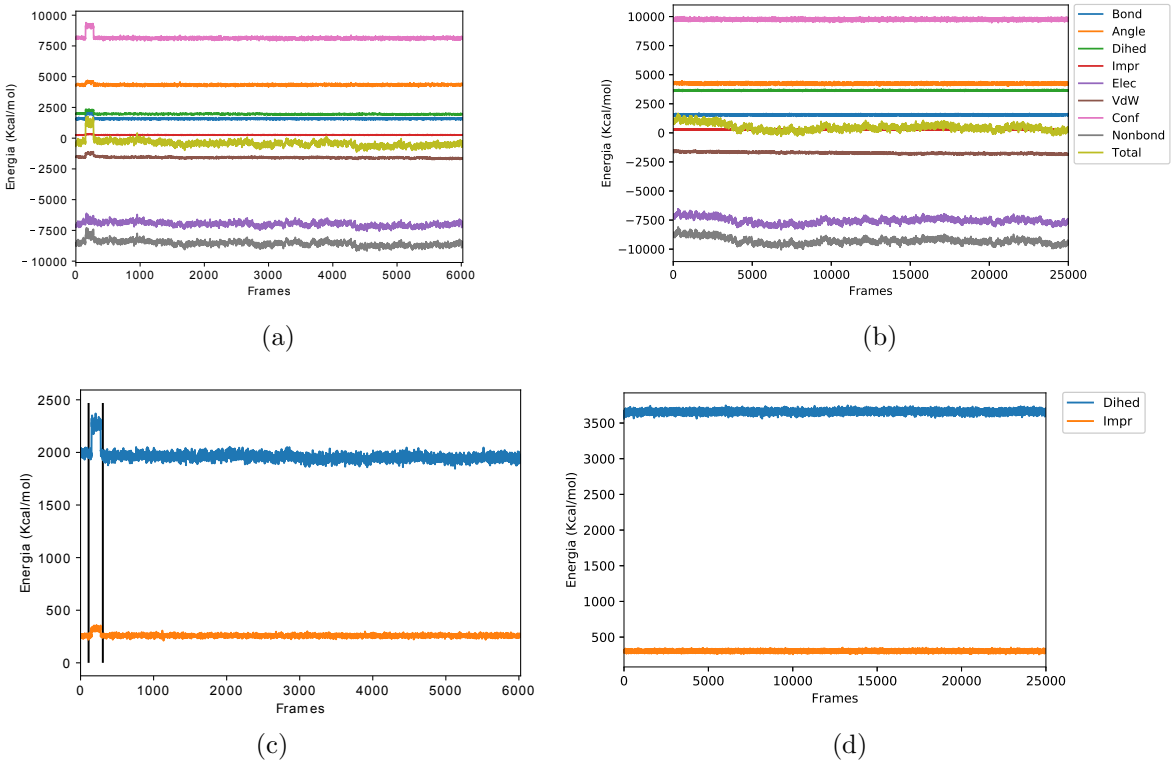


Figura .2.1 – Flutuação das energias potenciais para simulações para enzima smNTPDase 2 ao longo das simulações com AU1. Em (a) e (b) são apresentadas as energias totais obtidas para a simulação cMD e aMD, respectivamente. Enquanto que nas figuras (c) e (d) são mostradas em ordem, para melhor observação, as energias potenciais diedrais (próprio em azul e impróprio em laranja), ao longo das simulações cMD e aMD. As energias calculadas com o plugin NAMD-energy dentro do programa VMD[2] considerando apenas os átomos referentes à proteína.

Conforme observado nas figuras 2(a) e 2(b), os valores de RMSD são maiores ao longo da trajetória referente a simulação convencional, indicando que a proteína apresenta maior flutuação conformacional em relação ao primeiro frame ou ainda, que a mesma assume mais conformações. Quanto aos valores de raio de giro, figuras 2(c) e 2(d), os resultados indicam que de modo geral a proteína nas duas simulações tende ao enovelamento, já que os valores diminuem ao longo das trajetórias. Porém, assim como observado nas simulações anteriores, o acréscimo de um potencial nos ângulos diedrais permite um relaxamento mais rápido das estruturas, já que apenas 50ns os valores de raio de giro se estabilizam.

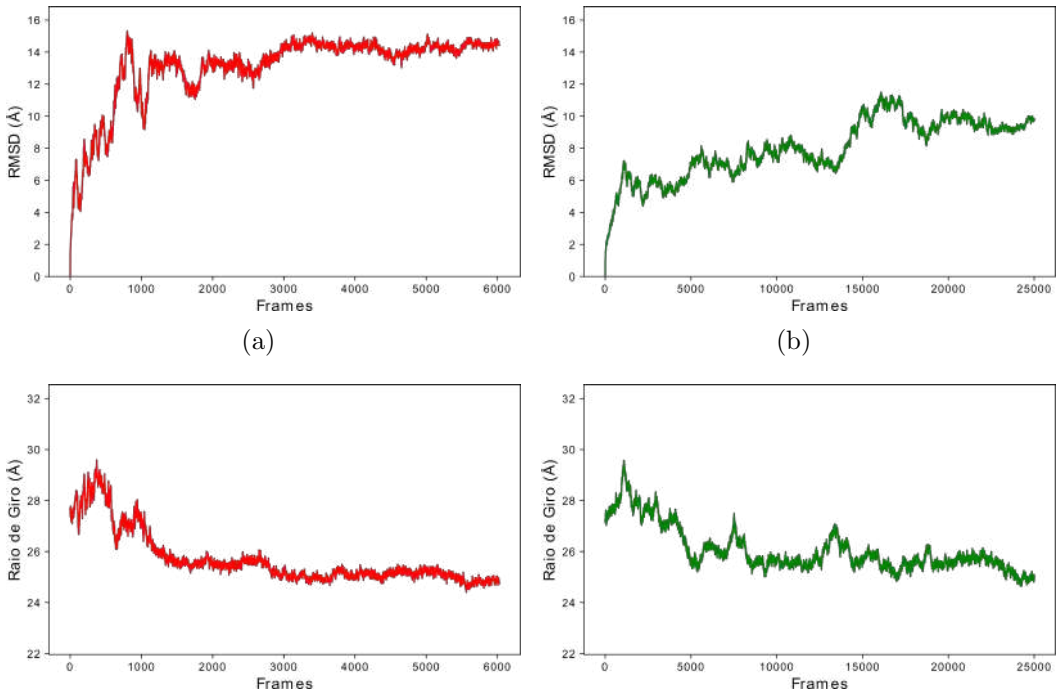
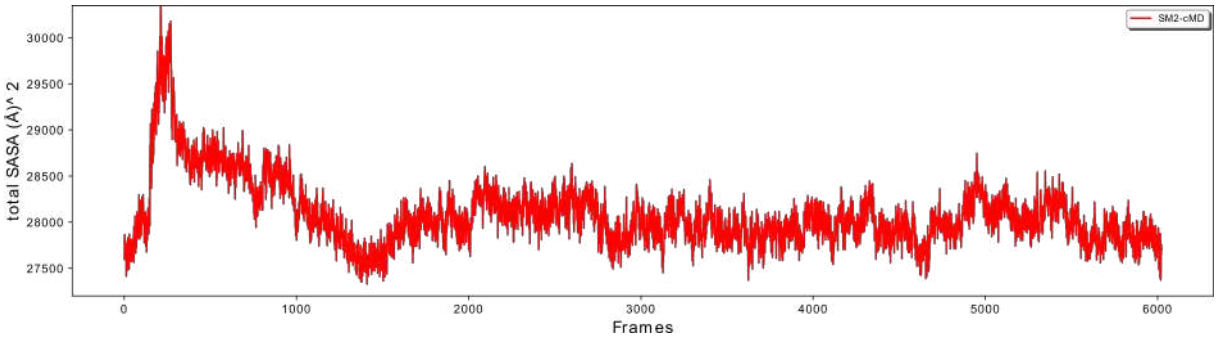
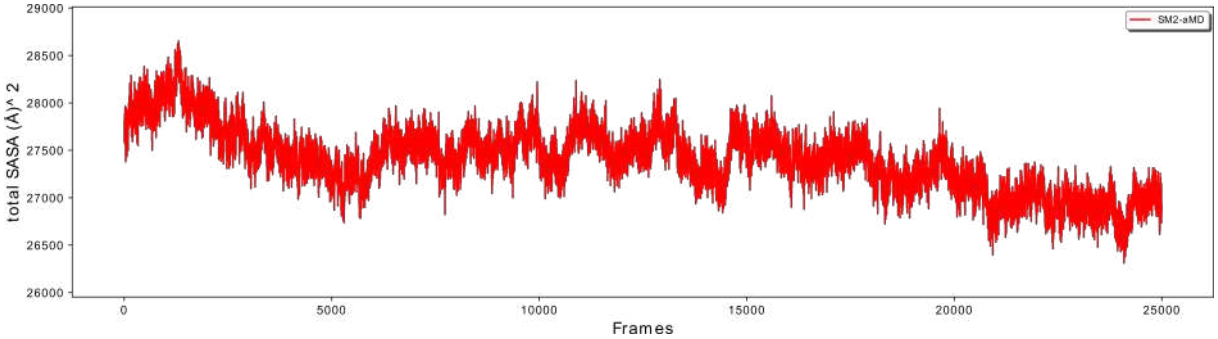


Figura .2.2 – Flutuação dos valores de RMSD e raio de giro da enzima smNTPDase 2 em simulações sem AU1. Em (a) e (b) são apresentados, em ordem, os valores de RMSD das simulações cMD e aMD. Já em (c) e (d) são mostradas flutuações do raio de giro ao longo das simulações cM e aMD, respectivamente.

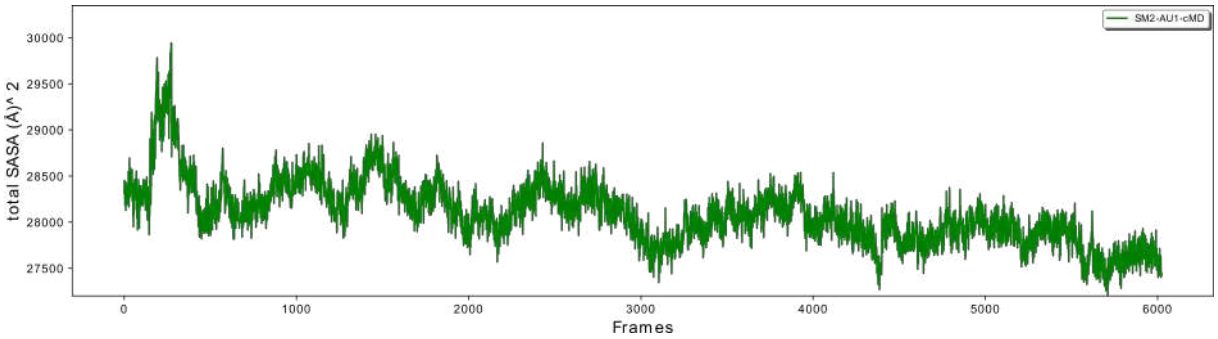


(a)

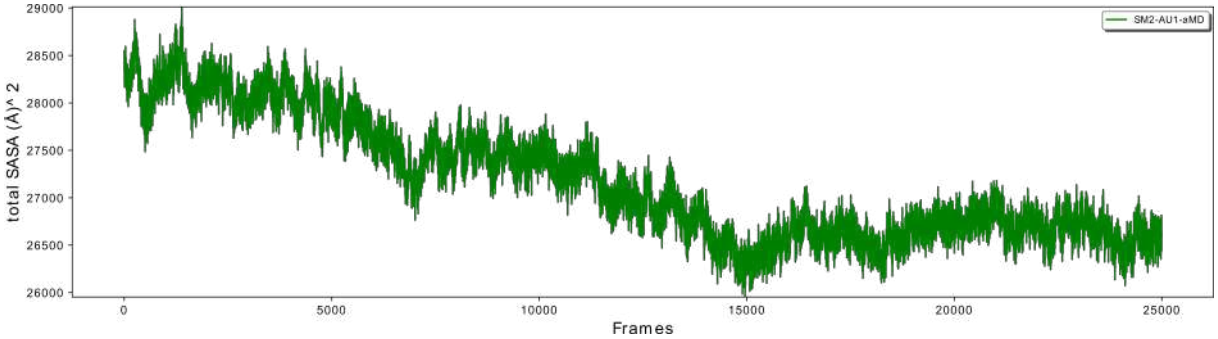


(b)

Figura .2.3 – Flutuação dos valores de SASA da proteína smNTPDase2 ao longo das simulações sem ligante. A subfigura (a) refere-se as variações na simuação cMD, enquanto que (b) representa as mudanças de SASA na simulação aMD.



(a)



(b)

Figura .2.4 – Flutuação dos valores de SASA da proteína smNTPDase2 ao longo das simulações com ligante. A subfigura (a) refere-se as variações na simuação cMD, enquanto que (b) representa as mudanças de SASA na simulação aMD.

ARTIGO 1

Clustering algorithms applied on analysis of protein molecular dynamics

Conference Paper · November 2017

DOI: 10.1109/LA-CCI.2017.8285695

3 authors:

Vinicius Carius De Souza
Federal University of Juiz de Fora
8 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Leonardo Goliatt
Federal University of Juiz de Fora
52 PUBLICATIONS **146** CITATIONS
SEE PROFILE

Priscila Capriles
Federal University of Juiz de Fora
23 PUBLICATIONS **147** CITATIONS
SEE PROFILE

Clustering Algorithms Applied on Analysis of Protein Molecular Dynamics

Vinicius Carius de Souza
Federal University of Juiz de Fora
Minas Gerais, Brazil
Email: carius.souza@gmail.com

Leonardo Goliatt
Federal University of Juiz de Fora
Minas Gerais, Brazil
Email: leonardo.goliatt@ufjf.edu.br

Priscila V. Z. Capriles Goliatt
Federal University of Juiz de Fora
Minas Gerais, Brazil
Email: priscilacapriles@gmail.com

***Abstract*—Analysis of molecular dynamic (MD) simulation has been difficult since this method generates a lot of conformations. Thus clustering algorithms have been applied to group similar structures from MD simulations, but the choice of the information to be clustered is still a challenge. In this work, we propose the use of Euclidean distance matrices (EDM) from conformations as input data to clustering algorithms. We used approaches combining non-reduction or reduction of data dimensionality (MDS and isomap methods), and different clustering algorithms (k-means, ward, mean-shift and affinity propagation). Results indicated that EDM could be a good information to be used in clustering conformations from MD. For data with small protein structure variation, the mean-shift algorithm had good results in both non-reduced and reduced data. However, for data with large protein structure variation, the methods that work better with smooth-density data (k-means and ward) had good results.**

I. Introduction

Molecular dynamic (MD) simulation is a powerful technique used to generate data about the movements of a particle system as a time function, constituting trajectories which are dependent on the potential of interaction between the atoms and the resolution of the equations of motion of classical mechanics [1].

MD simulations is well-suited for studying recurrent conformations, transitions states and predictions of physicochemical and geometric properties of proteins and other biomolecules, being important to characterize how these molecules perform their functions [2]. Understanding the dynamic (and conformations) of a protein is important to guide studies as the developing of compounds more efficient to a specific target protein. When no suitable crystallographic structures for a particular molecular target are available (*i.e.*, structures with inaccessible or poorly defined binding sites), MD can be applied to generates an ensemble of conformations for docking studies [3]. However, MD simulations generate a lot of conformations, which makes analysis difficult in practice [4]. Therefore, scientists have created novel techniques to reduce the number of MD structures without losing information about the protein dynamics[4][5][6].

The current research is supported by Brazilian Federal Agency for Postgraduate Education (CAPES).

978-1-5386-3734-0/17/$31.00

Clustering methods have been applied to MD results in order to partition protein ensembles into groups of structures with similar physicochemical and conformational features, constituting a protein sub-state (stable or transitional) [2][7][8][9][10]. This approach is useful, since it allows to characterize a conformational ensemble generated with MD and guides the analysis to focus on the most important changes [8]. Clustering trajectories typically require an understanding of possible states that a protein can assume. Additionally, the high dimensionality of conformational space, noise and other factors can avoid the formation of homogeneous clusters for real biomolecules. Approaches combining dimensionality reduction with subsequent clustering of trajectories provided by MD simulations are able to reduce noise and generate more homogeneous clusters [11].

Clustering algorithms use different metrics as input for similarity function, which determines the proximity of data. For data provided by MD, the main metric of structural similarity is the RMSD (Root Mean Square Deviation) value based on the atom distance between conformations [4][12]. The analysis using RMSD is interesting in to evaluate a convergence of the mean structure from MD. However, it is not possible to directly compute a time correction function from a series of RMSD values, since the difference between two RMSD values may not be indicative of how much structures are similar [14]. Therefore, new methods that replace RMSD may be interesting to identify or avoid problems of protein clustering. An interesting alternative is using the Euclidean Distances Matrices (EDM) to replace RMSD as similarity function. An EDM is a $n \times n$ matrix that represent the spacing of a set of n points in Euclidean space. These matrices can be applyed in problems such as wireless sensor network localization, statistics, dimensionality reduction (*e.g.* machine learning), and molecular conformation analysis [15]. In this paper, we propose the use of EDM between protein atoms as a alternative method to RMSD. EDM arises as an interesting alternative to RMSD (since it avoids the conformational overlap from similar RMSD) and can be straightforwardly used within the mathematical formulation of clustering algorithms and dimensionality reduction techniques. We performed different MD simulations at different temperatures to provide the trajectory data of proteins used to test and validate clustering algorithms. The obtained results are compared with manual analyses in

order to assess the applicability of the proposed approach.

II. Material and Methods

A. MD simulations details

The calculations were carried out using two crystal structure of proteins obtained from Protein Data Bank (PDB): 1CLL, a calmodulin of *Homo sapiens* (148 residues), and 1L2Y, a designed protein (20 residues). In order to obtain a diverse set of conformations for each protein, we performed simulations in 310K and 510K temperatures, assuming that high temperatures promoting more protein movements. The simulation systems were prepared using VMD program. To allow larger movements of 1CLL, we removed all Ca^{2+} ions bounded in the protein. Both proteins were embedded in a box containing TIP3 water, which extended to at least 15 Åbetween protein and the edge of the box. Finally, systems were neutralized adding 150mM of NaCl.

After systems preparations, internal constraints were relaxed by two energy minimization steps of 100ps, followed by an heating step raising the protein temperature from crystal conditions to 310K or 510K. Then, the equilibration step was performed by 1.5ns. Lastly, MD production step was performed for 20ns. The simulations were carried out using NAMD v2.12 program and CHARMM27 force field. During MD run, the long-range Coulomb interactions were calculated by the PME method, with a cutoff distance of 12Å and a switching function at 10Å. SHAKE algorithm was applied using a step size of 2fs in the Verlet algorithm. The temperature control was carried out by Langevin dynamics with a damping coefficient of 1ps. Pressure control was applied by a Langevin piston. All systems are run in the NPT ensemble at 1atm pressure.Simulations were carried out on a computer with four CPU AMD Opteron™ processor 6272 (64 cores of 2.1GHz and cache memory of 2MB), 128 Gb of RAM, HD of 500 Gb, four Nvidia Tesla M2090 video cards with 6 Gb and operational system CentOS release 6.5.

B. Data Set for Clustering the MD Trajectory

Data for the MD ensemble were collected at every 2 ps, resulting in a set of 10,000 trajectories for each simulation. For each data set, was selected 500 conformations (uniformly distributed) for clustering. In addition, we selected the last conformation of equilibration step, as a frame of reference. For each conformation were extracted the solvent-accessible surface area (SASA) in $Å^2$ and EDM between atoms of protein, using self-algorithms. We extracted EDM for protein backbone (N, C_α, C, and O) and C_α, for the purpose of analyzing which set of attributes represent better conformational changes in proteins. The information about SASA and time index of structures were used as connectivity to ward algorithm.

In order to organize MD data, we created a database using mysql v.7 and algorithms in python v2.7 were used to fill it. With this dataset, we seek to cluster different behaviors found along an MD simulation, which in turn may help to identify which of the clusters contain frames or structures that represent stable or transitory states of proteins.

C. Clustering Algorithms

In this study we used four clustering algorithms:

1) K-means: considering n samples, this method constructs k partitions, where each one represents a group and $k \leq n$. Given k, partitioning is accomplished by an iterative reallocation technique that seeks to improve partitioning by moving objects from one cluster to another [16].
2) Ward: the objects are hierarchically decomposed, resulting in a representation similar to a dendrogram that expresses the union process between the groups and all their intermediate levels. This method starts each object in a group and then joins them until they are in a single group or that satisfies the clustering condition [17].
3) Affinity propagation: it is based on the concept of "message passing" by interconnected graphs generated from the data points (vertex) [18]. This algorithm, unlike partition and hierarchical clustering methods, does not require the number of clusters *a priori*.
4) Mean-shift: it is a density-based algorithm. The aim is to discover "blobs" at smooth density samples updating candidates for centroids to be the mean of the points within a given region. It can estimates cluster number automatically using a parameter bandwidth, which dictates the size of the region to search through [19].

The elbow method was implemented to determinate cluster number for k-means and ward algorithms. This method is based on variance explained by the clusters against the number of clusters. According to the elbow method, you should choose a number of clusters so that adding another cluster the marginal gain will drop [20].

All clustering algorithms were implemented using scikit-learn package v0.18 in python v2.7. The computational experiments were carried out on a computer with intel® core™ i7 860 2.8 GHz, 8 Gbytes (Gb) of RAM, HD of 860 Gb, Nvidia Geforce GTX 285 video card with 1 Gb and operational system Fedora release 23.

D. Dimensionality Reduction of Data

The clustering of real biomolecules typically do not form such homogeneous groups, due basically to factors such as the high dimensionality of the space of conformations and noises produced by thermal variations [13]. Approaches combining dimensionality reduction with subsequent clustering for trajectory analysis provided by MD are able to reduce noise and generate groups of more homogeneous conformations [11].

In order to reduce the noise of data, two methods were used: Multidimensional scaling (MDS) and Isometric Mapping (Isomap). MDS is a classical algorithm for non-linear dimensionality reduction that projects the inputs with high dimensionality to a lower dimensional space, keeping their pairwise square distances $\left|\vec{X}_i - \vec{X}_j\right|^2$ [21]. MDS has been used to solve problems involving EDM, as the problem of finding the best point-set representation of a given set of distances [15]. Another method used to dimensionality reduction is isomap, which performs MDS in the geodesic space of the

non-linear data manifold to find a low-dimensional mapping that preserves pairwise distances of data [22].

III. Results and Discuss

To cluster MD data, each algorithm was applied to 1CLL and 1L2Y MD trajectories. All set of clusters obtained to four MD simulations were analyzed visually comparing the grouped structures and verifying the cohesion of groups using the silhouette score.

Results in Tables I and II provide a summary of relative performance of clustering runs as a number of clusters and silhouette score of cluster counts. In terms of clustering, values of silhouette near 1.0 suggest better clustering and values near -1.0 meaning non-cohesive clusters. We performed the manual analysis of 501 structures of each MD, in order to identify *a priori* the range of possible clusters to be generated. Through this analysis, we observed 1 or 2 clusters for 310K simulations, whereas for 510K simulation were observed a range of 8 to 10 possible clusters for 1L2Y and 5 to 8 possible clusters for 1CLL.

TABLE I
NUMBER OF CLUSTERS GENERATED BY CLUSTERING ALGORITHMS FOR 1L2Y SIMULATIONS

1L2Y	NR[5]		MDS		Isomap	
Temperature	310K	510K	310K	510K	310K	510K
K-means	5 (0.22)	10 (0.15)	4 (0.36)	5 (0.28)	6 (0.49)	5 (0.33)
Ward[1]	5 (0.21)	6 (0.14)	4 (0.31)	7 (0.21)	6 (0.48)	5 (0.29)
Ward[2]	9 (0.003)	10 (0.04)	7 (-0.02)	5 (0.13)	6 (0.18)	9 (0.06)
Ward[3]	4 (0.22)	4 (0.02)	4 (0.27)	4 (0.003)	4 (0.36)	6 (-0.13)
Ward[4]	4 (0.22)	4 (0.12)	5 (0.24)	11 (0.08)	4 (0.38)	4 (0.14)
Affinity	29 (0.07)	37 (0.14)	25 (0.25)	23 (0.26)	18 (0.36)	25 (0.28)
Mean-shift	2 (0.40)	1 (0)	2 (0.47)	1 (0)	3 (0.45)	2 (0.53)

[1]Ward without connectivity. [2]Ward using only SASA as connectivity. [3]Ward using time as connectivity. [4]Ward using time and SASA as connectivity. [5]Tests using non-reduced EDM. Numbers in parentheses are values of silhouette score.

TABLE II
NUMBER OF CLUSTERS GENERATED BY CLUSTERING ALGORITHMS FOR 1CLL SIMULATIONS

1CLL	NR[5]		MDS		Isomap	
Temperature	310K	510K	310K	510K	310K	510K
K-means	4 (0.30)	4 (0.38)	4 (0.33)	4 (0.44)	5 (0.53)	4 (0.53)
Ward[1]	4 (0.28)	4 (0.37)	4 (0.31)	4 (0.43)	5 (0.49)	4 (0.52)
Ward[2]	5 (0.21)	4 (0.28)	7 (0.18)	4 (0.36)	5 (0.44)	6 (0.34)
Ward[3]	4 (0.28)	4 (0.36)	4 (0.36)	4 (0.43)	5 (0.39)	4 (0.53)
Ward[4]	4 (0.30)	4 (0.38)	4 (0.32)	4 (0.43)	5 (0.41)	4 (0.51)
Affinity	22 (0.14)	29 (0.23)	20 (0.26)	19 (0.35)	13 (0.44)	14 (0.46)
Mean-shift	2 (0.48)	2 (0.40)	2 (0.53)	2 (0.40)	2 (0.68)	2 (0.64)

[1]Ward without connectivity. [2]Ward using only SASA as connectivity. [3]Ward using time as connectivity. [4]Ward using time and SASA as connectivity. [5]Tests using non-reduced EDM. Numbers in parentheses are values of silhouette score.

The K-means algorithm generated clustering sets with similar size in both reduction and non-reduction tests, excepting tests using non-reduced data of 1L2Y at 510K. Although silhouette score values had been relatively good in tests applied for 310K simulations, manual analyses indicated that K-means was very sensitive to small structure variations, both for reduced and non-reduced EDM. It promoted the separation of similar structures in different clusters. However, manual analyses of tests using 510K simulation data corroborated silhouette score results, indicating that K-means was able to generate cohesive clusters. We observed that in general K-means produced good results to non-reduced EDM.

In this study, we performed four different approaches for ward algorithm using the connectivity parameter (Fig. 1): (1) ward without connectivity; (2) SASA as connectivity; (3) time index as connectivity; and (4) time index and SASA as connectivity. In general, approaches using time as connectivity had been important to generate cohesive groups. It is probably because the information about temporal sequence binds similar data points, restricting how clustering algorithm groups data. For simulations at 310K, ward algorithm was sensitive to subtle structure variations similar to K-means. On the other hand, for simulations of 1L2Y at 510K, ward algorithm using time as connectivity to non-reduced EDM or reduced EDM by MDS, resulted in clusters with protein structures more similar than obtained clusters using reduced EDM by the isomap method. Furthermore, simulation of 1CLL at 510K did not present differences between approaches of the ward in both non-reduced and reduced EDM.

The affinity algorithm generated more groups than other algorithms. We presume that it is possible because this algorithm generates interconnected graphs generated from data points. Manual analyses of affinity clusters indicated that clusters could be regrouped to generate groups more cohesive, mainly for simulations at 310K where we previously detected 1 or 2 clusters.

Interestingly, the mean-shift algorithm had good results in simulations at 310K, being able to detect manually verified clusters. However, this method was not be able to detect good clusters in simulations to 510K (Tab. I and Tab. II). The mean-shift method proposed by Fukunaga and Hostetler in 1975 uses a kernel density to estimate the gradient of the data density. At every iteration, the kernel is shifted to a higher density region until convergence and a new centroid or mean is defined within it [19]. Consequently, mean-shift tends to generate better results for smooth-distribution data than higher distribution. We observed that simulation at 510K for 1L2Y and 1CLL generated higher-distribution data point, due to unstable conformations (Fig. 2). This property explains poor results obtained with mean-shift for simulations at 510K.

In general, analyses of results indicated that clustering algorithms used to non-reduced and reduced EDM was similar in simulations to 310K and 1CLL to 510K. Although approaches using isomap had better silhouette values than non-reduced and MDS methods. Concerning the simulation for 1L2Y to 510K, the ward algorithm had different behaviors when used non-reduced and reduced EDM (Tab. I). In addition, we observed that MDS method with time and SASA as connectivity was be able to generate cohesive clusters for simulation for 1L2Y to 510K. Similar results were obtained using non-reduced and isomap methods with only SASA as connectivity.

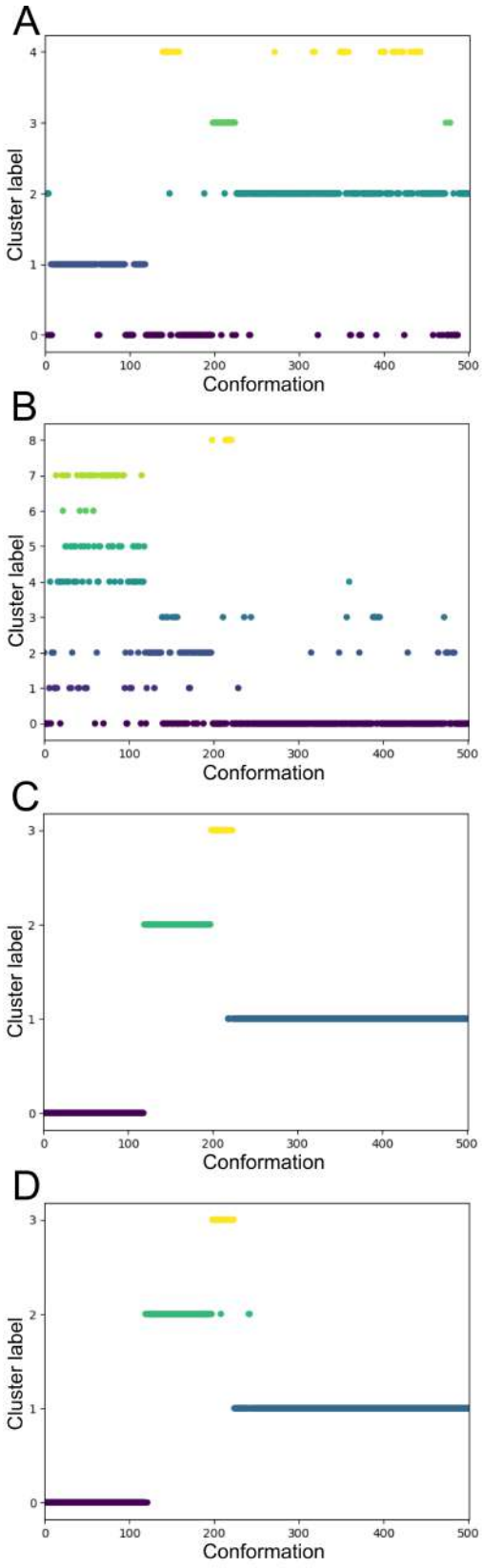


Fig. 1. Result comparing different approaches used to evaluate ward algorithm. These plots refer to clustering of non-reduced data from simulation of 1L2Y at 310K. (A) Ward without connectivity; (B) Ward with SASA as connectivity; (C) Ward using time as connectivity; and (D) Ward using time and SASA as connectivity.

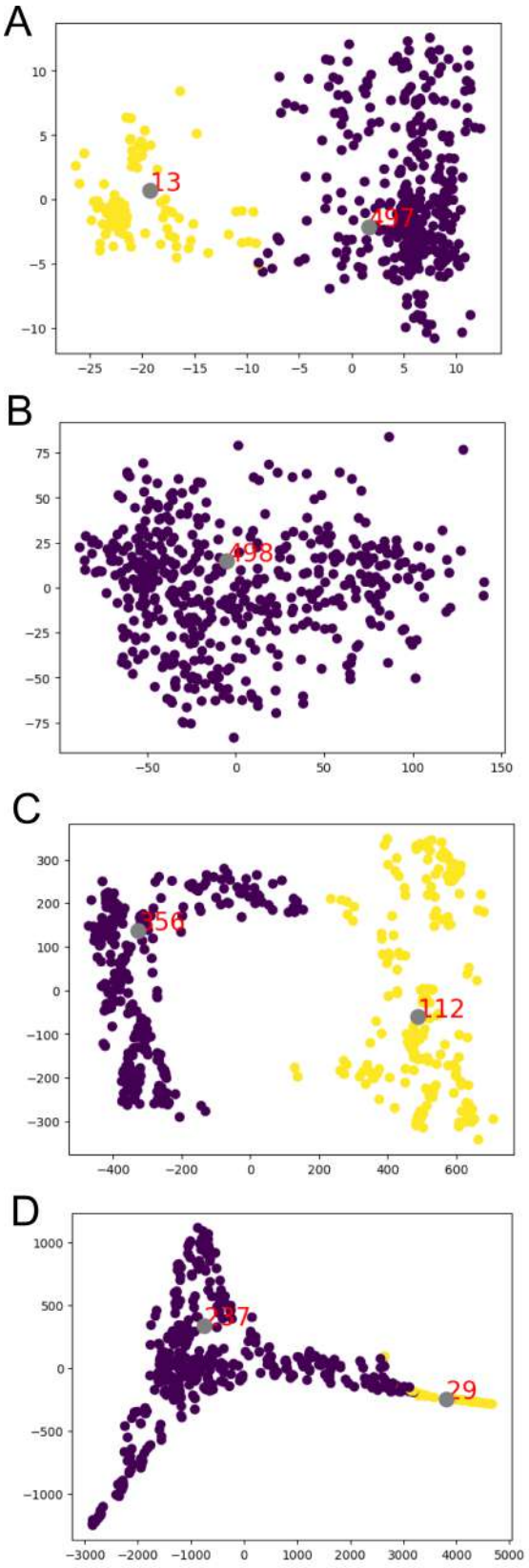


Fig. 2. 2D plots of mean-shift clustering tests. A and B represent data from simulations for 1L2Y to 310K and 510K, respectively. C and D represent data from simulations for 1CLL to 310K and 510K, respectively. Gray dots and labels are representative structure of cluster. It be able to see that 310K results have smooth density compared 510K.

In figures 3 and 4, we observed best algorithms for different simulations and representative conformations for each group. For simulations at 310K, representative conformations are very similar between them, because both 1L2Y and 1CLL explored only a stable conformation. However, for simulation at 510K, we observed that representative conformations are very different, because both proteins undergo denaturation process, assuming a lot of stable and unstable conformations. This process generates noises which make it difficult to obtain cohesive clusters.

Despite we have obtained good results using mean-shift (non-reduced and reduced EDM) for simulations at 310K and ward (connectivities and reduced EDM) for simulations at 510K. Two facts could suggest the poor results to other algorithms: (1) the dimensionality reduction used here has led to the loss of significant data features, which are important to generate clusters with more similar structures; and (2) in this work, tests were carried out using default parameters for both clustering algorithms and reduction methods.

IV. Conclusion and Future Work

In this work, we performed clustering protein conformations from MD simulations using EDM between C_α atoms of structures. We used different approaches combining methods of dimensionality reduction (MDS and isomap) with clustering algorithms (K-means, affinity propagation, mean-shift and agglomerative ward). The results indicated that different approaches should be used to group protein structures from MD simulations.

According to results, we observed that for data from MD where proteins assume stable conformations (310K), the mean-shift algorithm had the best results, because of smooth distribution of data. However, for data with largely unstable conformations (510K), methods that work better with higher distribution data (kmeans and ward) had the best results. In addition, we observed disagreement between manual analyses and values of silhouette score, *e.g.* mean-shift results for simulations at 510K had higher silhouette score but clusters were not consistent. This result indicates that new methods of clustering validity criteria (*e.g.* Calinski–Harabasz, Dunn's index and Davies–Bouldin) should be explored for clustering of MD simulations data.

Finally, EDM could be an interesting option to group protein structures from MD simulations. In future work, we propose to explore parameters of clustering algorithm and other methods of dimensionality reduction. Optimization algorithms (*e.g.* differential evolution and particle swarm optimization) are alternative forms to search the best parameters of clustering algorithms for different data sets.

Acknowledgment

The authors thank Postgraduate Program in Computational Modeling from Federal University of Juiz de Fora and CAPES.

References

[1] M Karplus, J A McCammon, "Molecular Dynamics Simulations of Biomolecules", *Nature Structural Molecular Biology*, vol. 9, no. 9, pp. 646-652, 2002.

[2] J L Phillips, M E Colvin, S Newsam, "Validating Clustering of Molecular Dynamics Simulations Using Polymer Models", *BMC bioinformatics*, vol. 12, no. 1, pp. 445, 2011.

[3] L G Ferreira, R N dos Santos, G Oliva, A D Andricopulo, "Molecular Docking and Structure-based Drug Design Strategies", *Molecules*, vol. 20, no. 7, pp. 13384-13421, 2015.

[4] R D Paris, C V Quevedo, D D Ruiz, O N D Souza, R C Barros, "Clustering Molecular Dynamics Trajectories for Optimizing Docking Experiments", *Computational intelligence and neuroscience*, vol. 2015, pp. 32, 2015.

[5] R D Paris, F A Frantz, O N de Souza, D D Ruiz, "wFReDoW: A Cloud-based Web Environment to Handle Molecular Docking Simulations of a Fully Flexible Receptor Model", *BioMed research international*, vol. 2013, 2013.

[6] C V Quevedo, R D Paris, D D Ruiz, O N De Souza, "A Strategic Solution to Optimize Molecular Docking Simulations Using Fully-flexible Receptor Models", *Expert Systems with Applications*, vol. 41, no. 16, pp. 7608-7620, 2014.

[7] A E Torda, W F Van Gunsteren, "Algorithms for Clustering Molecular Dynamics Configurations", *Journal of computational chemistry*, vol. 15, no. 12 pp. 1331-1340, 1994.

[8] J M Troyer, F E Cohen, "Protein Conformational Landscapes: Energy Minimization and Clustering of a Long Molecular Dynamics Trajectory", *Proteins: Structure, Function, and Bioinformatics*. vol. 23, no. 1, pp. 97-110, 1995.

[9] J Shao, S W Tanner, N Thompson, T E Cheatham, "Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms", *Journal of Chemical Theory and Computation*, vol. 3, no. 6, pp. 2312-2334, 2007.

[10] E Papaleo, P Mereghetti, P Fantucci, R Grandori, L De Gioia, "Free-energy Landscape, Principal Component Analysis, and Structural Clustering to Identify Representative Conformations from Molecular Dynamics Simulations: The Myoglobin Case", *Journal of molecular graphics and modelling*, vol. 27, no. 8, pp. 889-899, 2009.

[11] A Wolf, K N Kirschner, "Principal Component and Clustering Analysis on Molecular Dynamics Data of The Ribosomal L11 23S Subdomain", *Journal of molecular modeling*, vol. 19, no. 2, pp. 539-549, 2013.

[12] E Lyman, D M Zuckerman, "Ensemble-based Convergence Analysis of Biomolecular Trajectories", *Biophysical journal*, vol. 91, no. 1, pp. 164-172, 2006.

[13] J L Phillips, M E Colvin, S Newsam, "Validating Clustering of Molecular Dynamics Simulations Using Polymer Models", *BMC bioinformatics*, vol. 12, no. 1, pp. 445, 2011.

[14] D R Roe, T E Cheathmiii, "PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data", *Journal of Chemical Theory and Computation*, vol. 9, no. 7, 3084-3095, 2013.

[15] I Dokmanic, R Parhizkar, J Ranieri, M Vetterli, "Euclidean Distance Matrices: Essential Theory, Algorithms, and Applications", *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 12-30, 2015.

[16] S K Popat, M Emmanuel, "Review and Comparative Study of Clustering Techniques", *International Journal of Computer Science and Information Technologies*, vol. 5, no. 1, pp. 805-812, 2014.

[17] J H Ward Jr, "Hierarchical Grouping to Optimize an Objective Function", *Journal of the American statistical association*, vol. 58, no. 301, 236-244, 1963.

[18] B J Frey, D Dueck, "Clustering by Passing Messages between Data Points", *science*, vol. 315, no. 5814, pp. 972-976, 2007.

[19] D Comaniciu, P Meer, "Mean shift: A Robust Approach Toward Feature Space Analysis",. *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603-619, 2002.

[20] T S Madhulatha, "An Overview on Clustering Methods",*IOSR Journal of Engineering*, vol. 2, no. 4, pp. 719-725, 2012.

[21] K Q Weinberger, L K Saul, "Unsupervised Learning of Image Manifolds by Semidefinite Programming", *International Journal of Computer Vision*, vol. 70, no. 1, pp. 77-90, 2006.

[22] A Ghodsi, "Dimensionality reduction a short tutorial", *Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada*, vol. 37, pp. 38, 2006.

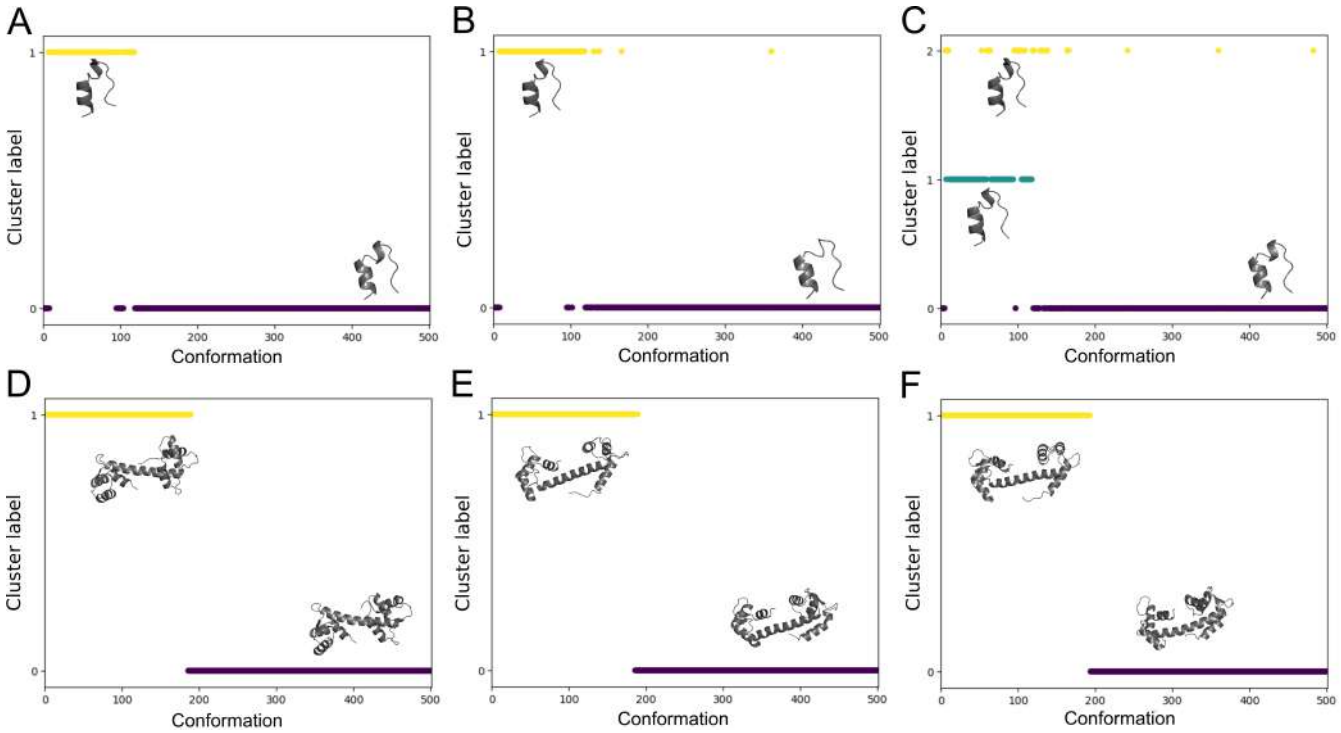


Fig. 3. Comparative results of mean-shift algorithm using different approaches of dimensionality reduction for data from simulation to 310K. For 1L2Y: (A) non-reduced EDM, (B) MDS method applied, (C) isomap method applied. For 1CLL: (D) non-reduced EDM, (E) MDS method applied, (F) isomap method applied.

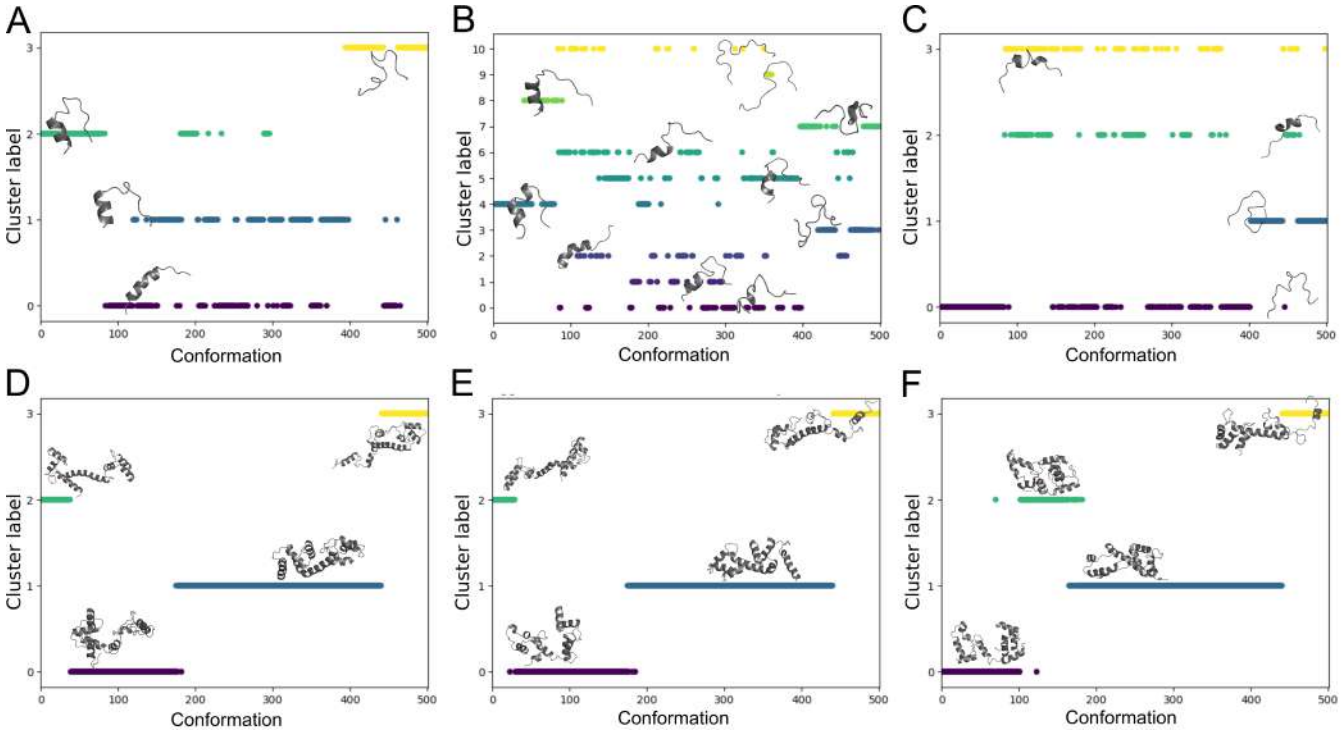


Fig. 4. Comparative results of ward algorithm using different approaches of dimensionality reduction for data from simulation to 510K. At all tests were used time and SASA as connectivity. For 1L2Y: (A) non-reduced EDM, (B) MDS method, (C) isomap method. For 1CLL: (D) non-reduced EDM, (E) MDS method, (F) isomap method.

ARTIGO 2

Insight About Nonlinear Dimensionality Reduction Methods Applied to Protein Molecular Dynamics

Vinicius Carius de Souza(✉), Leonardo Goliatt, and Priscila V. Z. Capriles

Federal University of Juiz de Fora, Juiz de Fora, MG, Brazil
vinicius.carius@ice.ufjf.br,
{leonardo.goliatt,priscila.capriles}@ufjf.edu.br

Abstract. The advance in molecular dynamics (MD) techniques has made this method common in studies involving the discovery of physicochemical and conformational properties of proteins. However, the analysis may be difficult since MD generates a lot of conformations with high dimensionality. Among the methods used to explore this problem, machine learning has been used to find a lower dimensional manifold called "intrinsic dimensionality space" which is embedded in a high dimensional space and represents the essential motions of proteins. To identify this manifold, Euclidean distance between intra-molecular C_α atoms for each conformation was used. The approaches used were combining data dimensionality reduction (AutoEncoder, Isomap, t-SNE, MDS, Spectral and PCA methods) and Ward algorithm to group similar conformations and find the representative structures. Findings pointed out that Spectral and Isomap methods were able to generate low-dimensionality spaces providing good insights about the classes separation of conformations. As they are nonlinear methods, the low-dimensionality generated represents better the protein motions than PCA embedding, so they could be considered alternatives to full MD analyses.

Keywords: Molecular dynamics · Manifold · Clustering algorithm

1 Introduction

Since its inception, molecular dynamics (MD) techniques have suffered important modifications leading to the simulation of complex and relevant systems with hundreds of different atoms [15]. In a biological context, MD simulation has proved to be suitable to study transition states and predictions of physicochemical and geometric properties of proteins, the key to characterize molecules functions [22]. Previous studies had shown that MD simulations can be applied to generate an ensemble of conformations for docking studies to protein structures with inaccessible or poorly defined binding sites [12].

I. Rojas et al. (Eds.): IWBBIO 2019, LNBI 11466, pp. 219–230, 2019.
https://doi.org/10.1007/978-3-030-17935-9_21

Despite its usefulness, MD analysis may be difficult as many conformations are generated and classifying them demands a lot of time and knowledge about protein behavior [20]. So, artificial intelligence have been auspicious to detect and classify conformations from a set of trajectories, at no risk of loss of information on the protein dynamics [8,20,23]. Such techniques include unsupervised methods like clustering algorithm that attempts to partition data set into groups with similar features without prior knowledge. Every feature used as input for a clustering algorithm is considered a coordinate in a space with n dimensionality.

When applied to protein conformations, different physicochemical and conformational properties can be used as input for clustering. However, the high dimensionality of conformational space, noise, and other factors lead to the well-known curse of dimensionality in statistical pattern recognition, preventing homogeneous clusters from forming. Therefore, the use of dimensionality reduction (DR) methods with subsequent clustering of trajectories from MD simulations have been able to reduce noise and generate more homogeneous clusters [34].

Principal component analysis (PCA) has been applied to analyze data from MD since 1991 [16]. It is very important to obtain a set of orthogonal vectors which are considered the "essential subspace" and are able to capture the largest amplitude of protein motions from a set of trajectories [6]. Although PCA is probably the best known linear technique able to acquire information about complex dynamics, by intrinsically incorporating a dimensional hyperplane, the low-dimensional embedding obtained may be distorted [5,11,31] when PCA is applied to the nonlinear space of conformational protein changes. The reason why nonlinear machine learning techniques have been proposed to pinpoint the underlying manifold structure so as to analyze the space explored by MD simulations and solve the inherent problem of PCA [24].

This study aims to contribute to new approaches of MD analysis, placing the AutoEncoder, t-SNE and Spectral embedding in the context of the of DR methods applied to MD simulation analyses. Combining these methods with the clustering Ward algorithm to identify representative structures, a comparative analysis of six different DR methods (including linear and nonlinear) was performed.

2 Materials and Methods

2.1 Data Set for Clustering the MD Trajectory

Euclidean distances were calculated between intra-molecular C_α for every 501 conformations from MD simulation (at 310 K and 510 K temperatures) of calmodulin (PDB 1CLL), previously published in [27], to obtain a Euclidean distance matrix (EDM) $E_{i_{n\times n}}$, where i is the matrix index and n is total the number of residues. The upper triangular part of each EDM was converted to 1D vector. Each 1D vector was added to a feature matrix $M_{m\times p}$, where m is the number of conformations and p is the number of distances that describe the

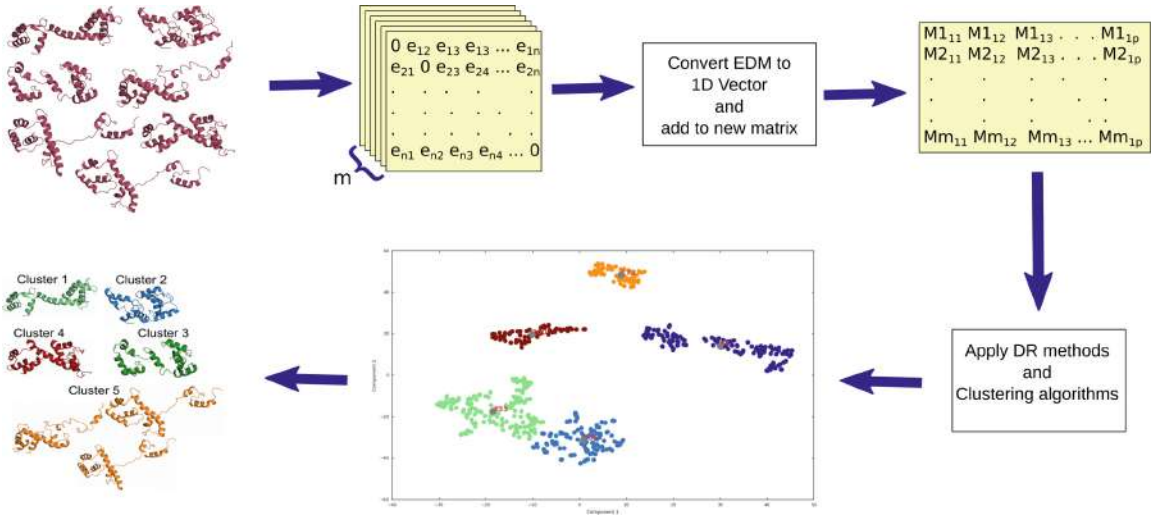


Fig. 1. Protein representation in computational experiments.

conformational fluctuations. The feature matrix was used as input for clustering and dimensionality reduction methods carried out in this paper (Fig. 1).

2.2 Dimensionality Reduction (DR)

In order to reduce the noise of data and find a new space of coordinates that represent the protein fluctuations, the following six methods were used:

(a) **Multidimensional scaling (MDS)** is a technique applied to nonlinear DR, which builds a projection in a lower dimensional space of n points in Euclidian space. In this new space obtained by MDS, elements are represented by points whose respective distances best approximate the initial distance [10,33]. In this work, was used an MDS variation called *Metric Multi-dimensional Scaling* (mMDS). This method minimizes the cost function called "Stress" which is a measure for the deviation from monotonicity between the distances d_{ij} and the observed dissimilarities [29],

$$S = \sqrt{\frac{\sum_{ij} \left(d_{ij} - d_{ij}^{*}\right)^2}{\sum_{ij} d_{ij}^2}} \tag{1}$$

where d_{ij} and d_{ij}^{*} are the predicted and target distances, respectively.

(b) **Isometric feature mapping (Isomap)** is considered an extension of MDS idea, incorporating the geodesic distances induced by a neighborhood graph embedded in the classical scaling [14,30]. This method performs three steps. Step 1, the algorithm determines the neighbors on the manifold M of each point in the input space X, based on the distances $d_X(ij)$. Using these distances, the method constructs an edge-weighted neighborhood graph G, where the weight of each edge is equal to the Euclidean distance $d_X(ij)$. Step 2, Isomap estimates the geodesic distances $d_M(ij)$ between all pairs of

points on the manifold M by computing their shortest path in the graph G. Step 3, lower-dimensional embedding is computed applying MDS method to the matrix of graph distances $D_G = \{d_G(ij)\}$ [30].

(c) **t-distributed Stochastic Neighbor Embedding (t-SNE)** is a variation of Stochastic Neighbor Embedding (SNE), which converts the high-dimensional Euclidean distances between points from the data set into Gaussian joint probabilities that represent similarities [18]. This method performs two main steps: Step 1, t-SNE starts by converting the high-dimensional Euclidean distances between points in the initial space into conditional probabilities p_{ij} (Eq. 2) that represent their similarities.

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N} \tag{2}$$

where p_{ij} is proportional to the similarity of objects x_i and x_j as follows:

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2/2\sigma_i^2\right)} \tag{3}$$

Step 2, a similar probability distribution q_{ij} (Eq. 4) is defined over the points in the low-dimensional map using a heavy-tailed Student-t distribution and minimizes the gradient of the Kullback-Leibler divergence between P and the Student-t based joint probability distribution Q [18].

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}} \tag{4}$$

(d) **Spectral Embedding (SE)** uses the spectral decomposition of the Laplacian graph generated by the similarity matrix of the data [2], being considered a discrete approximation of the low dimensional manifold in the high dimensional space [2]. SE algorithm calculates an affinity matrix $A \in \mathbb{R}^{nxn}$ defined by $A_{ij} = \exp\left(-\|s_i - s_j\|^2/2\sigma^2\right)$ if $i \neq j$, and $A_{ii} = 0$. After this, the method defines a diagonal matrix D whose (i,i)-element is the sum of i-th rows of A, and constructs the laplacian matrix L, where $L = D^{-1/2}AD^{-1/2}$. Applying a spectral decomposition, the k largest eigenvectors of L are chosen and form the matrix $X = [x_1, x_2, \ldots, x_k] \in \mathbb{R}^{nxk}$, by stacking the eigenvectors in columns. Finally, the matrix Y is obtained from X by renormalizing each row of X to have unit length [19].

(e) **Principal Component Analysis (PCA)** is a linear method that uses the singular value decomposition of the data to project them to a linear subspace attempting to maintain most of the variability of the data [1,26].

(f) **AutoEncoder (AE)** is a type of artificial neural network that is able to learn representations from data sets in an unsupervised manner [32]. Architecturally, an autoencoder consists of two parts, the encoder and the decoder. The encoder maps an input $x_i \in \mathbb{R}^{d_x}$ to a hidden layer $y_i \in \mathbb{R}^{d_N}$ with reduced dimensionality, using a function g, as described bellow.

$$y_i = g(Wx_i) \tag{5}$$

where g can be any function for a linear projection (*e.g.* identity function) or for a nonlinear mapping (*e.g.* sigmoid function). The parameter W is a $d_y x d_x$ weighted matrix. The decoder allows reconstructing $x'_i \in \mathbb{R}^{d_x}$ from the hidden layer with low-dimension y_i,

$$x'_i = f(W' y_i) \tag{6}$$

where W' is another $d_y x d_x$ weighted matrix defined as W^T. The function f is similar to g and can be a function for linear or nonlinear projection [32]. The Fig. 2 shows a structure of autoencoder built in this study. During training, we used 150 epochs and a batch size value equals 128. In addition, we also used the mean squared error (mse) as loss function and the optmizer Adaptive Moment Estimation (Adam).

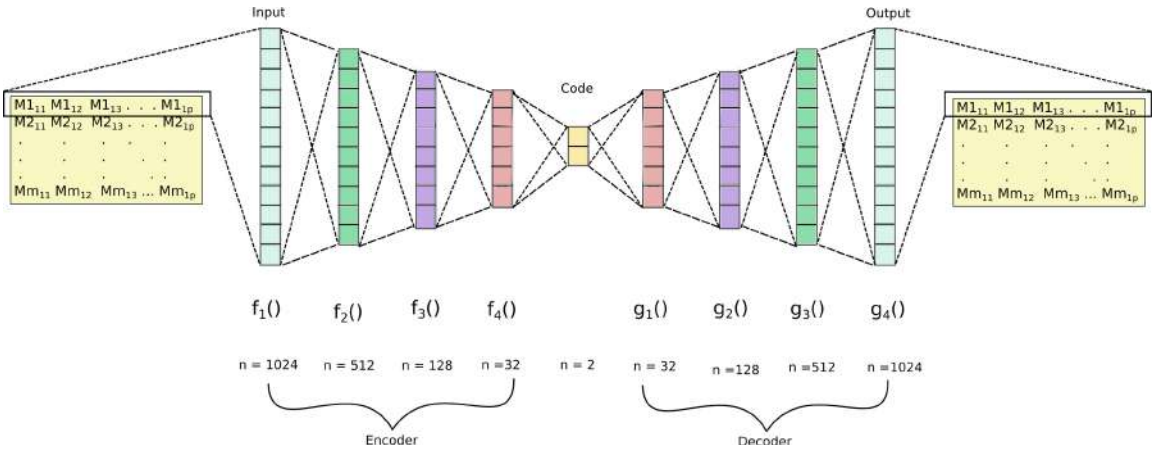


Fig. 2. Autoencoder representation. An autoencoder with 4 layers for encoder and decoder. For the encoder part, two functions was used: Exponential Linear Unit (ELU) from f_1 to f_3, and a linear function for f_4. For the decoder part, from g_1 to g_3 was used ELU and for g_4 was used a sigmoid function. In this figure, n represents the number of neurons in each layer.

2.3 Clustering Approach

Ward algorithm was used to cluster the protein conformations. Previous studies showed that this algorithm has generated good results when applied to data from MD simulations [9,27]. Ward is an agglomerative hierarchical clustering method that minimizes the total within-cluster variance for each pair of cluster centers or medoids found after merging. At the first step, all clusters contain a single point and then iterative steps are performed until all points are merged in the homogeneous group or that a condition of grouping is satisfied.

Ward algorithm is a K dependent method, which requires to provide the number of clusters *a priori*. In this work, we performed the elbow method [3] to find the best number of clusters (K number) for the data set after running each

dimensionality reduction method. To evaluate unsupervised classification performed by Ward, we calculated different validation measures: Calinski-Harabasz index (CH) [4], Davies-Bouldin index (DBI) [7], Fowlkes Mallows Index (FMI) [13], and Silhouette score [25]. The Fig. 3 shows a flowchart developed here.

Another analysis performed was the Root-Mean-Square-Deviation (RMSD) between the lower energy structure obtained by weighted Histogram Analysis Method (WHAM) and the medoids found using Ward clustering algorithm.

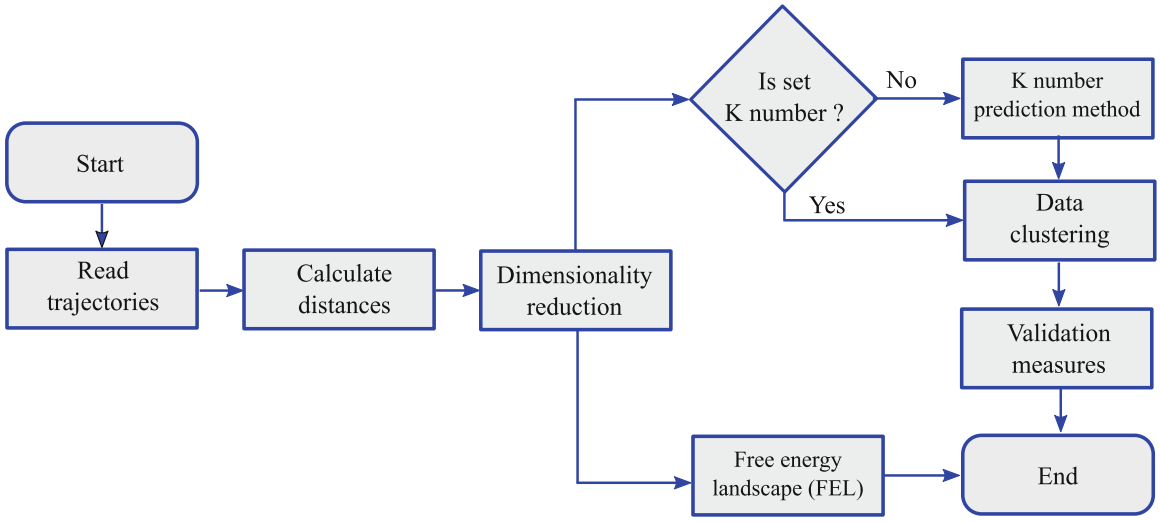


Fig. 3. Flowchart of the methodology. All steps were implemented using python v2.7. The different DR methods and Ward clustering algorithm were implemented using scikit-learn package v0.18 [21].

2.4 Statistical Evaluation

All tests were run thirty times to calculate the statistical differences between them. An Anderson-Darling test was performed to verify the normality of the answer variables. As non-parametric analysis, the Kruskal-Wallis followed by Dunn's *post-hoc* test were run to verify the statistical differences between the clustering scores calculated, using a significance set at the 5%. All computational experiments were carried out on intel® core™ i7 860 2.8 GHz processor, 8 Gbytes (Gb) of RAM, HD of 860 Gb, operating system Fedora release 23.

2.5 Free Energy Landscape

In order to have a free energy landscape (FEL) for each conformational set from MD simulations, we applied the Weighted Histogram Analysis Method (WHAM) [17]. This method is based on the fact that given a set of discrete states of a molecule, is possible to obtain a histogram with discrete bins that provide a relative probability of a state occurs along the trajectory [17]. So, the higher density of states in a histogram region provides the insight that the probability of this set representing an energy basin is greater. The WHAM idea is derived

from statistical mechanics and the function of free energy, considering a state ξ, is defined by

$$F(\xi) = -k_B T \ln Z(\xi) \tag{7}$$

so that $Z(\xi)$ is a partition function that is proportional to the density of the states in bins and is given by

$$Z(\xi) = \int e^{\beta U(\xi)} d\Omega \tag{8}$$

where $\beta = 1/k_B T$, k_B is the Boltzmann constant, T is the temperature in Kelvin, and $U(\xi)$ is the potential energy. Considering a reduced space (intrinsic space), found by DR methods applied to the internal coordinates of the protein conformations, the FEL can be generated by inverting the probability distribution ($\hat{P}$) of points (states) of a multidimensional histogram obtained from the n principal components ($\{\boldsymbol{\Psi}\}_{i=2}^{k+1}$) [11], as follows:

$$F = -k_B T \ln \hat{P}\left(\{\boldsymbol{\Psi}\}_{i=2}^{k+1}\right) + C \tag{9}$$

3 Results and Discussion

When clustering methods are applied to bio-molecule simulation data, it is expected that generated partitions include similar conformations which represent transition and meta-states of the system. Approaches which combine dimensionality reduction (DR) and clustering methods are able to generate homogeneous clusters when applied to molecular dynamics (MD) simulations, which assists the analysis of protein conformational fluctuations [34]. In this paper, we performed a comparative analysis of different DR approaches combined with the Ward clustering algorithm, applied to MD data.

Figure 4 shows the clustering results for the Ward algorithm using different DR methods applied to data from 1CLL simulations at 310 K and 510 K. In 310 K simulation, all methods found 3 or 4 clusters, whereas in 510 K simulations it was found a range from 4 to 6 clusters. It was slightly different from previous manual analyses in which was observed 2 and 8 clusters for 310 K and 510 K simulations, respectively [27]. Probably, it is due to the overlapping between some classes in a lower dimensionality space, maybe because the data sets were reduced to only two dimensions and some information was lost.

To evaluate the class prediction quality of each approach, the internal clustering validation metrics were analyzed, which are based on the intrinsic information of the data: Calinski-Harabasz index (CH), Davies-Bouldin index (DBI) and Silhouette score. In 310 K simulations, it was observed that the Isomap method had the best values of quality, followed by SE (Fig. 5). However, in 510 K simulations, the best CH and DBI values were obtained by Spectral and Isomap, but for the silhouette result, Spectral and PCA presented best values (Fig. 5).

Internal validation metrics provide an important insight into the quality of different clustering approaches. However, unsupervised methods are more liable

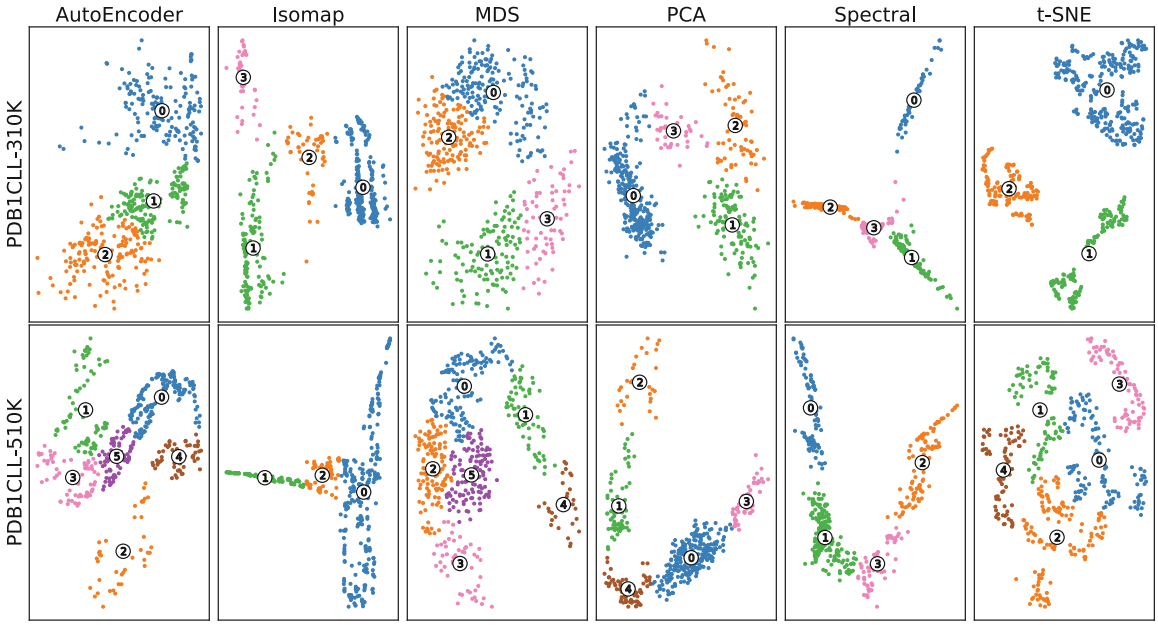


Fig. 4. Visualization of reduced space (2D) obtained from different DR method applied to 501 calmodulin (PDB 1CLL) conformations from MD simulations. Clusters were obtained by Ward algorithm.

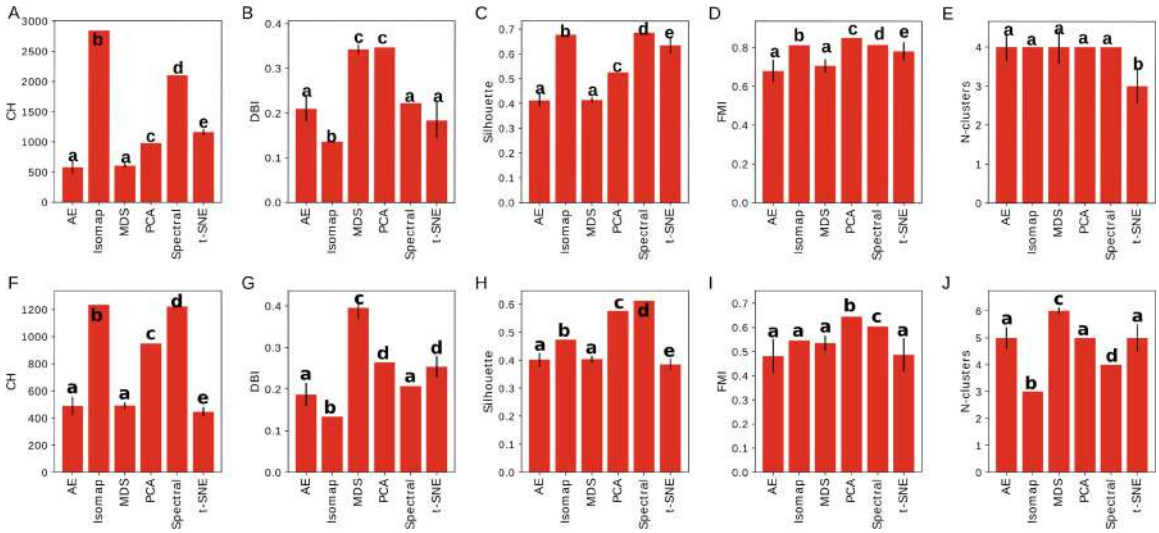


Fig. 5. Comparative analysis of the clustering evaluation metrics between DR approaches. Plots from A to E show metrics for 310 K simulations and those from F to J are results for 510 K simulations. Bars with the same letter were statistically similar, according to Dunn's test with a significance level of 5%.

to misclassification errors than supervised methods, especially for protein data sets. So, external measures based on previous knowledge about data have been used to evaluate clustering algorithms. Therefore, to verify the performance of each method when compared to the manual analyses, the Fowlkes Mallows Index (FMI) was calculated [13] using manual analyses as the reference (data not shown). In 310 K simulations, Isomap and PCA showed high FMI values,

whereas AE and MDS presented low values and were not statistically different (Fig. 5D). In 510 K simulations, PCA had the best value of FMI followed by Spectral (Fig. 5I).

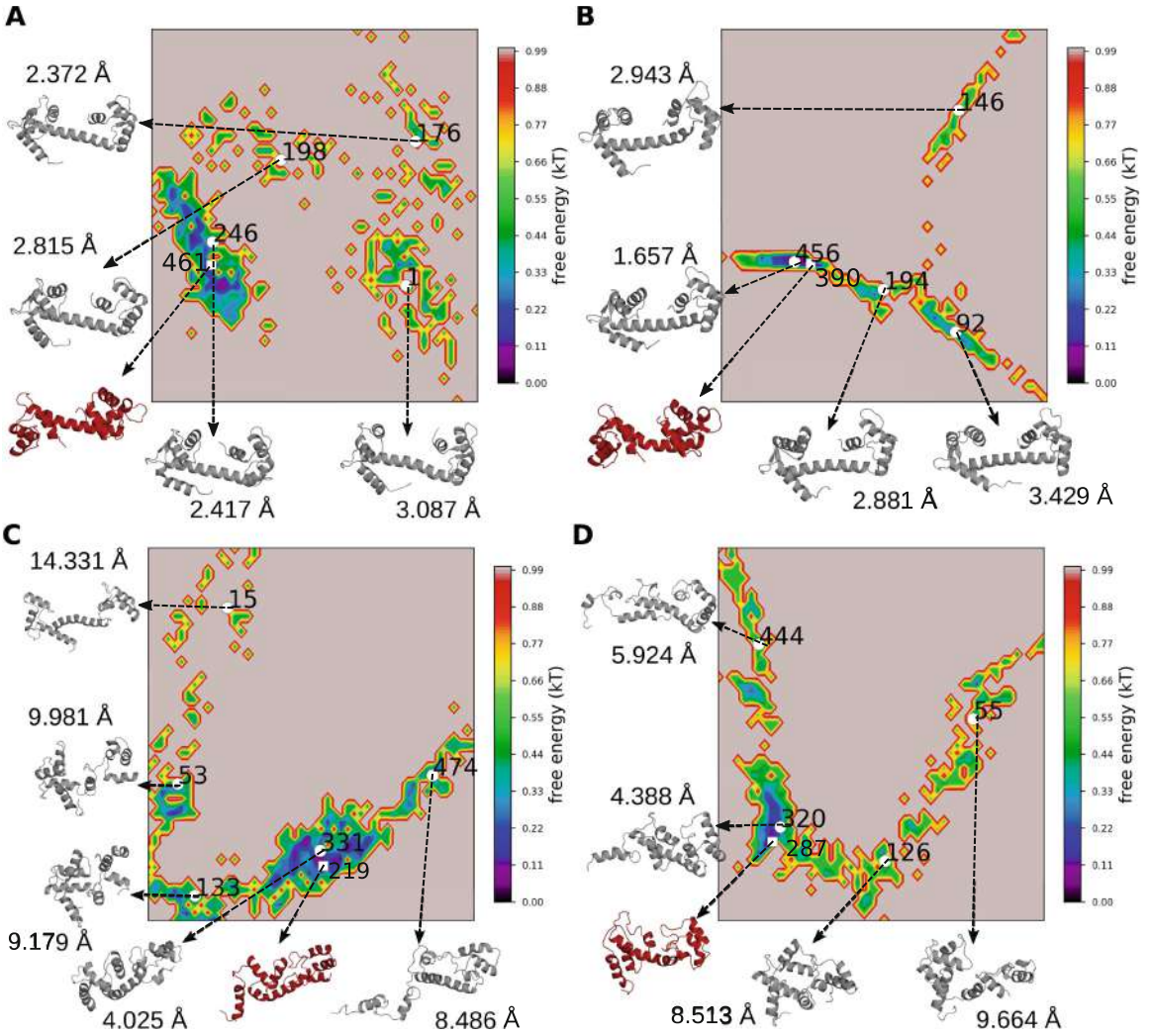


Fig. 6. Free Energy Landscape (FEL) for calmodulin (PDB 1CLL) and the representative structures predicted by the Ward algorithm for PCA and Spectral methods. Figures A and B show simulations at 310 K, whereas figures C and D show the FEL obtained for simulations at 510 K. Here the FEL was calculated using the Weighted Histogram Analysis Method (WHAM). In grey are represented the cluster medoids and in red are structures of lower energy value. (Color figure online)

According to our results, Spectral obtained good values in all evaluation measures. So, Free Energy Landscape (FEL) generated by PCA (classic method of DR) and Spectral (Fig. 6) were compared. In general, PCA gives more basins than Spectral, which could be explained by the fact that barriers and basins are influenced by the coordinate(s) in lower dimension space generated by different machine learning techniques. In addition, considering that the energy basins are regions of a greater density of states, we could say that the spectral method

was able to group more conformations for the same cluster than PCA, which generates fewer regions of minima in the FEL. A similar result has been reported by a previous study in which PCA found more energy minima than Isomap, even though the separation is less clear [28]. Our results also revealed that the FEL generated by PCA has more energy barriers than Spectral, which is highlighted by comparing both methods applied to 310 K simulations.

Although most representative structures predicted by the Ward algorithm are not in energy minima (Fig. 6), they reflect the general patterns assumed by protein during simulations. For example, Fig. 6 shows the structures in minimum energy (red) and medoid (grey) states. According to RMSD values calculated between the lower energy structure and medoids, it was observed that in simulations at 310 K the minimum and maximum values were 2.372 Å and 3.087 Å for PCA embedding, whereas for Spectral embedding the values were 1.657 Å and 3.429 Å. In these simulations, the highest RMSD values represent conformations in which rotations occurred in helices and loops of the lobes or even rotation in the central helix. For simulations at 510 K, the minimum and maximum RMSD values were 4.025 Å and 14.331 Å for PCA embedding and 4.388 Å and 9.664 Å for Spectral embedding. As expected in these simulations, the higher values of RMSD represent misfolded structures due to the high temperature.

4 Conclusions and Perspectives

The purpose of this work was to perform a comparative analysis between different machine learning approaches to find out the manifold that characterizes the protein motions and representative conformations from MD trajectories. For this, six different dimensionality reduction (DR) methods were applied to internal coordinates obtained from Euclidean distances between C_α atoms of structures, and the "intrinsic dimensionality space" found was used as input for agglomerative Ward algorithm to group similar conformations and identify those considered representatives within each cluster.

The results show that when considering the best values of external and internal validation metrics, Spectral and Isomap arise good alternatives to explore the conformational space of proteins from MD simulation, although these methods have failed to predict the K groups expected. Considering the K number predicted, AE, PCA and MDS methods presented the best performance. Another significant finding from this study is that AutoEncoder was able to identify lower dimension in a way that similar conformations were close, showing as a promising alternative to current DM analyzes.

Acknowledgment. The authors thank the Graduate Program in Computational Modeling from Federal University of Juiz de Fora and the Brazilian agencies FAPEMIG (grant 01606/15), CNPq (grant 429639/2016-3) and CAPES for the financial support.

References

1. Abdi, H., Williams, L.J.: Principal component analysis. Wiley Interdisc. Rev. Comput. Stat. **2**(4), 433–459 (2010)
2. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **15**(6), 1373–1396 (2003)
3. Bholowalia, P., Kumar, A.: EBK-means: a clustering technique based on elbow method and k-means in WSN. Int. J. Comput. Appl. **105**(9) (2014)
4. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Commun. Stat.-Theory Methods **3**(1), 1–27 (1974)
5. Das, P., Moll, M., Stamati, H., Kavraki, L.E., Clementi, C.: Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. Proc. Nat. Acad. Sci. **103**(26), 9885–9890 (2006)
6. David, C.C., Jacobs, D.J.: Principal component analysis: a method for determining the essential dynamics of proteins. In: Livesay, D. (ed.) Protein Dynamics, pp. 193–226. Springer, Heidelberg (2014). https://doi.org/10.1007/978-1-62703-658-0_11
7. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **2**, 224–227 (1979)
8. De Paris, R., Frantz, F.A., Norberto de Souza, O., Ruiz, D.D.: wFReDoW: a cloud-based web environment to handle molecular docking simulations of a fully flexible receptor model. BioMed Res. Int. **2013** (2013)
9. De Paris, R., Quevedo, C.V., Ruiz, D.D., de Souza, O.N.: An effective approach for clustering inha molecular dynamics trajectory using substrate-binding cavity features. PLoS ONE **10**(7), e0133172 (2015)
10. Dokmanic, I., Parhizkar, R., Ranieri, J., Vetterli, M.: Euclidean distance matrices: essential theory, algorithms, and applications. IEEE Sig. Process. Mag. **32**(6), 12–30 (2015)
11. Ferguson, A.L., Panagiotopoulos, A.Z., Kevrekidis, I.G., Debenedetti, P.G.: Nonlinear dimensionality reduction in molecular simulation: the diffusion map approach. Chem. Phys. Lett. **509**(1–3), 1–11 (2011)
12. Ferreira, L.G., dos Santos, R.N., Oliva, G., Andricopulo, A.D.: Molecular docking and structure-based drug design strategies. Molecules **20**(7), 13384–13421 (2015)
13. Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. J. Am. Stat. Assoc. **78**(383), 553–569 (1983)
14. Ghodsi, A.: Dimensionality reduction a short tutorial. Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada, vol. 37, p. 38 (2006)
15. Hospital, A., Goñi, J.R., Orozco, M., Gelpí, J.L.: Molecular dynamics simulations: advances and applications. Adv. Appl. Bioinf. Chem. AABC **8**, 37 (2015)
16. Ichiye, T., Karplus, M.: Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. Proteins Struct. Function Bioinf. **11**(3), 205–217 (1991)
17. Kumar, S., Rosenberg, J.M., Bouzida, D., Swendsen, R.H., Kollman, P.A.: The weighted histogram analysis method for free-energy calculations on biomolecules. I. the method. J. Comput. Chem. **13**(8), 1011–1021 (1992)
18. Maaten, L.V.D., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(Nov), 2579–2605 (2008)
19. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: analysis and an algorithm. In: Advances in Neural Information Processing Systems, pp. 849–856 (2002)
20. Paris, R.D., Quevedo, C.V., Ruiz, D.D., Souza, O.N.D., Barros, R.C.: Clustering molecular dynamics trajectories for optimizing docking experiments. Comput. Intell. Neurosci. **2015**, 32 (2015)

21. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
22. Phillips, J.L., Colvin, M.E., Newsam, S.: Validating clustering of molecular dynamics simulations using polymer models. BMC Bioinf. **12**(1), 445 (2011)
23. Quevedo, C.V., De Paris, R., Ruiz, D.D., De Souza, O.N.: A strategic solution to optimize molecular docking simulations using fully-flexible receptor models. Expert Syst. Appl. **41**(16), 7608–7620 (2014)
24. Rohrdanz, M.A., Zheng, W., Clementi, C.: Discovering mountain passes via torchlight: methods for the definition of reaction coordinates and pathways in complex macromolecular reactions. Ann. Rev. Phys. Chem. **64**, 295–316 (2013)
25. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
26. Shlens, J.: A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100 (2014)
27. de Souza, V.C., Goliatt, L., Goliatt, P.V.C.: Clustering algorithms applied on analysis of protein molecular dynamics. In: 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI), pp. 1–6. IEEE (2017)
28. Stamati, H., Clementi, C., Kavraki, L.E.: Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. Proteins Struct. Function Bioinf. **78**(2), 223–235 (2010)
29. Steyvers, M.: Multidimensional scaling. In: Encyclopedia of Cognitive Science, pp. 1–7 (2002)
30. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500), 2319–2323 (2000)
31. Teodoro, M.L., Phillips Jr., G.N., Kavraki, L.E.: A dimensionality reduction approach to modeling protein flexibility. In: Proceedings of the Sixth Annual International Conference on Computational Biology, pp. 299–308. ACM (2002)
32. Wang, W., Huang, Y., Wang, Y., Wang, L.: Generalized autoencoder: a neural network framework for dimensionality reduction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 490–497 (2014)
33. Weinberger, K.Q., Saul, L.K.: Unsupervised learning of image manifolds by semidefinite programming. Int. J. Comput. Vis. **70**(1), 77–90 (2006)
34. Wolf, A., Kirschner, K.N.: Principal component and clustering analysis on molecular dynamics data of the ribosomal L11·23S subdomain. J. Mol. Model. **19**(2), 539–549 (2013)

ARTIGO 3

catalysts

MDPI

Article

Insights into the Mechanism of Ethionamide Resistance in *Mycobacterium tuberculosis* through an *in silico* Structural Evaluation of EthA and Mutants Identi ed in Clinical Isolates

Vinicius Carius de Souza [1,†], Deborah Antunes [2,†], Lucianna H.S.Santos [3], Priscila Vanessa Zabala Capriles Goliatt [1], Ernesto Raul Ca arena [4], Ana Carolina Ramos Guimarães [2,*] and Teca Calcagno Galvão [2]

[1] Programa de Pós-graduação em Modelagem Computacional, Universidade Federal de Juiz de Fora–UFJF, Juiz de Fora, Minas Gerais, 36036-330, Brazil; carius.nara@gmail.com (V.C.d.S.); priscilacapriles@gmail.com (P.V.Z.C.G.)

[2] Fiocruz, Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Av. Brasil 4365, Rio de Janeiro, RJ 21040-360, Brazil; deborah.antuness@gmail.com (D.A.); carolg@ioc. ocruz.br (A.C.R.G.); teca@ioc. ocruz.br (T.C.G)

[3] Laboratório de Modelagem Molecular e Planejamento de Fármacos, Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG 31270-901, Brazil; luciannahss@gmail.com

[4] Fiocruz, Programa de Computação Cientí ca, Av. Brasil 4365, Rio de Janeiro, RJ 21040-360, Brazil; ernesto.ca arena@gmail.com

* Correspondence: carolg@ioc. ocruz.br

† These authors contributed equally to this work.

Received: date; Accepted: date; Published: date

Abstract: Mutation in the ethionamide (ETH) activating enzyme, EthA, is the main factor determining resistance to this drug, used to treat TB patients infected with MDR and XDR *Mycobacterium tuberculosis* isolates. Many mutations in EthA of ETH resistant (ETH-R) isolates have been described but their roles in resistance remain uncharacterized, partly because structural studies on the enzyme are lacking. Thus, we took a two-tier approach to evaluate two mutations (Y50C and T453I) found in ETH-R clinical isolates. First, we used a combination of comparative modeling, molecular docking, and molecular dynamics to build an EthA model in complex with ETH that has hallmark features of structurally characterized homologs. Second, we used free energy computational calculations for the reliable prediction of relative free energies between the wild type and mutant enzymes. The ΔΔG values for Y50C and T453I mutant enzymes in complex with $FADH_2$-NADP-ETH were 3.34 (+/−0.55) and 8.11 (+/−0.51) kcal/mol, respectively, compared to the wild type complex. The positive ΔΔG values indicate that the wild type complex is more stable than the mutants, with the T453I complex being the least stable. These are the rst results shedding light on the molecular basis of ETH resistance, namely reduced complex stability of mutant EthA.

Keywords: EthA; ethionamide resistance; BVMO; molecular dynamics; thermodynamic integration

1. Introduction

Although tuberculosis (TB) is a treatable disease, *Mycobacterium tuberculosis* is the single infectious agent causing the highest number of deaths [1]. Despite e orts by governments and bodies such as the WHO, the spread of drug-resistant strains continues. Factors underlying drug resistance include prolonged treatment schemes (ranging from 6 to 18 months), patient social vulnerability and the structure and e ectiveness of health systems [2]. Ethionamide (ETH) is used in treatment schemes of TB patients infected with drug-resistant *M. tuberculosis*. ETH has a low therapeutic index [3] and frequently causes dose-dependent adverse e ects (reviewed in [4]). Still, with the increase in the number of patients infected with isolates resistant to the range of drugs available [1], ETH is a critical resource in the clinic.

ETH is a prodrug activated by EthA, a avin adenine dinucleotide (FAD)-containing NADPH- and O_2-dependent Baeyer-Villiger monooxygenase (BVMO) [5–7]. Other proteins, such as MymA [8], EthR2 [9], Rv0565c [10] and MshA [11], have also been implicated in this process. However, because mutations in EthA are by far, the most commonly found in ETH resistant (ETH^R) *M. tuberculosis* clinical isolates, this protein is considered the major enzyme capable of forming the bactericidal NAD-ETH adduct (reviewed in [12]). The *in vivo* role of EthA in *M. tuberculosis* has not been fully elucidated but seems to involve modulation of cell wall composition. This is based on data showing that deletion of the *ethA-ethR* locus of *Mycobacterium bovis* BCG altered cell wall mycolic acid composition and increased adherence to host cells *in vitro*, a phenotype that can be modulated by cell wall components [13]. Because the mutant accumulates keto-mycolic acids, Alonso and colleagues have postulated a role for EthA in oxidizing keto-mycolic acids to wax ester mycolic acids, a reaction previously shown to be catalyzed by a BVMO in *Mycobacterium phlei* [13,14].

BVMOs use NADPH as an electron source and molecular oxygen as an oxidant to convert compounds with carbonyl groups into esters or lactones. This class of enzymes can transform a huge number of substrates with great regio- and enantioselectivity, making these enzymes highly relevant as biocatalysts. The type I oxygenation reaction catalyzed by FAD-dependent BVMOs depends on NADPH binding and reducing a stably bound FAD. NADPH, an electron donor, binds to FAD-bound BVMO, reduces FAD, and the reduced avin reacts with molecular oxygen, forming a reactive avin-peroxide intermediate that is stabilized by NADP. When the substrate is in the binding site, its electrophilic carbonyl su ers a nucleophilic attack by the peroxy avin intermediate, forming the Criegee intermediate (a tetrahedral species). Product formation occurs by rearranging the Crieege intermediate coupled to forming the product ester, the regeneration of the oxidized avin and NADP+ release (reviewed in [15–17]).

Nearly two hundred mutations in EthA have been reported in ETH^R clinical isolates (compiled in [18]). They are non synonymous substitutions, opal mutations, frameshits, and insertions, likely causing a wide range of changes in EthA that can a ect ETH activation. Enzymology studies have described intermediates of ETH activation by EthA [7,19]), without focusing on the BVMO reaction nor in EthA structural aspects. Thus, studies on EthA and on the impact of mutations found in ETH^R clinical isolates are sorely lacking. Mutations Y50C and T453I are likely to cause resistance as they were identi ed in ETH^R clinical isolates. Y50 lines the EthA binding pocket for $FADH_2$ and NADP [18], and mutation of the equivalent tyrosine in OTEMO reduces catalysis [20]. T453 is in the control loop region, whose movement is essential for catalysis. Here we use computational approaches to build and validate an EthA model and test the impact of Y50C and T453I.

2. Results and Discussion

2.1. Comparative Modeling

We modeled the three-dimensional (3D) structure of the 489 amino acid long EthA enzyme from *M. tuberculosis* (UNIPROT P9WNF9) by comparative modeling using as a template the structure of *Pseudomonas putida* OTEMO (PDB 3UOZ) [20]. Sequence alignment between EthA and 3UOZ revealed 25% identity, 39% similarity, 20% of gaps, and a query coverage of 450 amino acids (Figure 1a). EthA sequence similarity to cyclohexanone monooxygenase (CHMO), steroid monooxygenase (STMO) and phenylacetone monooxygenase (PAMO), the other BVMOs for which crystal structures are available, is lower than to OTEMO (data not shown). To optimize the overall structure, secondary structure constraints were inserted during the modeling process via Modeller v9.21 [21,22]. All generated models considered the presence of $FADH_2$ and NADP. The quality assessment results of the model were favorable according to ProSA-web, Whatcheck, and DOPE score. For the analysis of stereochemical quality, according to Molprobity, 95.2% of the residues were in the favorable or allowed regions of the Ramachandran map, and Procheck analysis showed that 95.7% of the residues were in the most favorable or allowed regions.

In another step to analyse the model, intermolecular interactions in the EthA FAD and NADPH binding regions were mapped by Protein Ligand Interaction Pro ler (PLIP) [23] and compared with those observed in 2-oxo-Δ 3 -4,5,5-trimethylcyclopentenylacetyl-coenzyme A monooxygenase (OTEMO). As shown in Figure 1a, the amino acids making up the OTEMO and EthA FAD and NADPH binding sites are mostly conserved. Three mobile regions characteristic of BVMOs are shown, the BVMO motif, the interdomain linker and the control loop [24]; reviewed in [25]. The overall organization of EthA and OTEMO is shown in Figure 1b, including four regions that, in OTEMO crystal structures, either appear disordered or adopt di erent conformations [20].

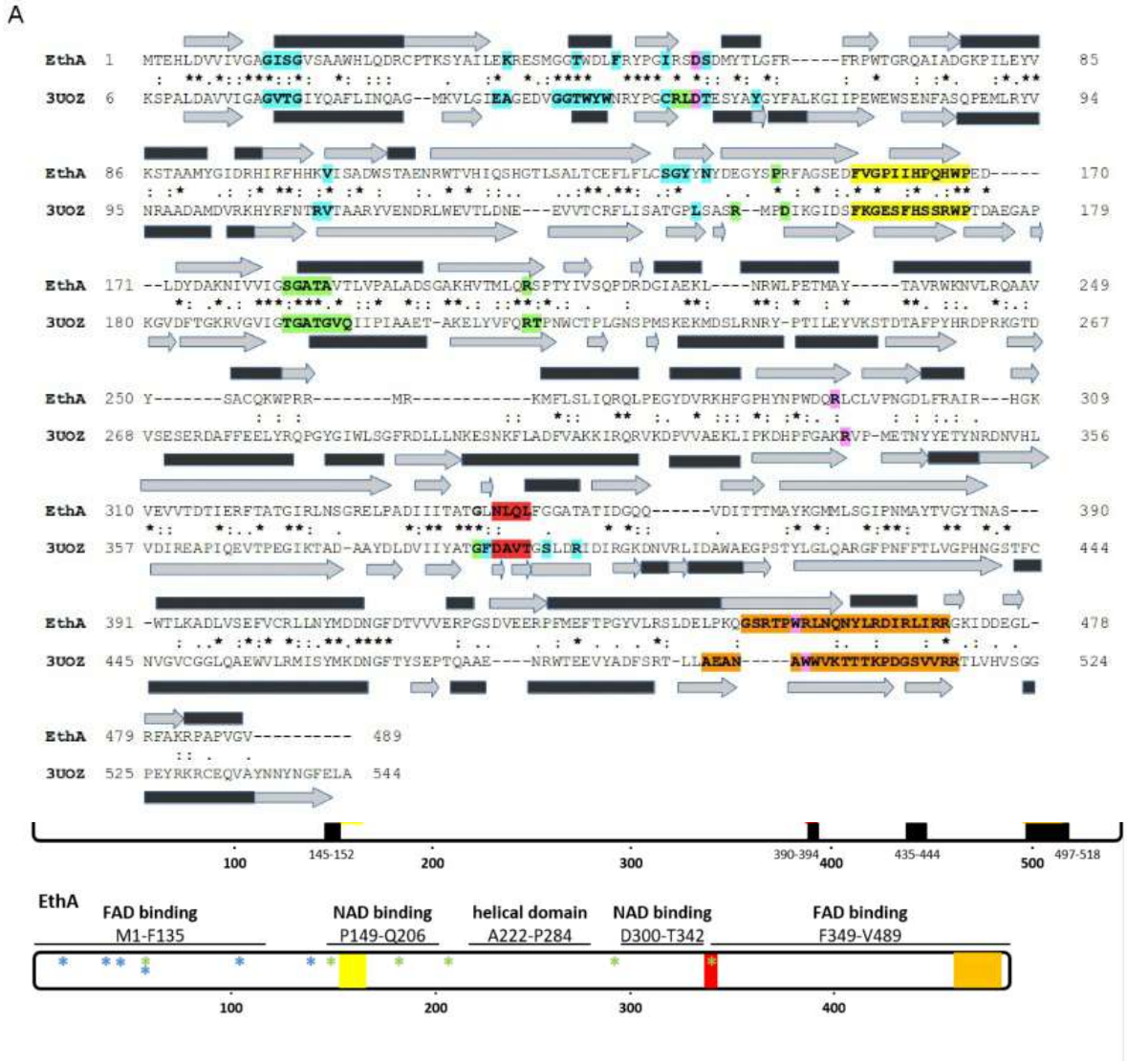
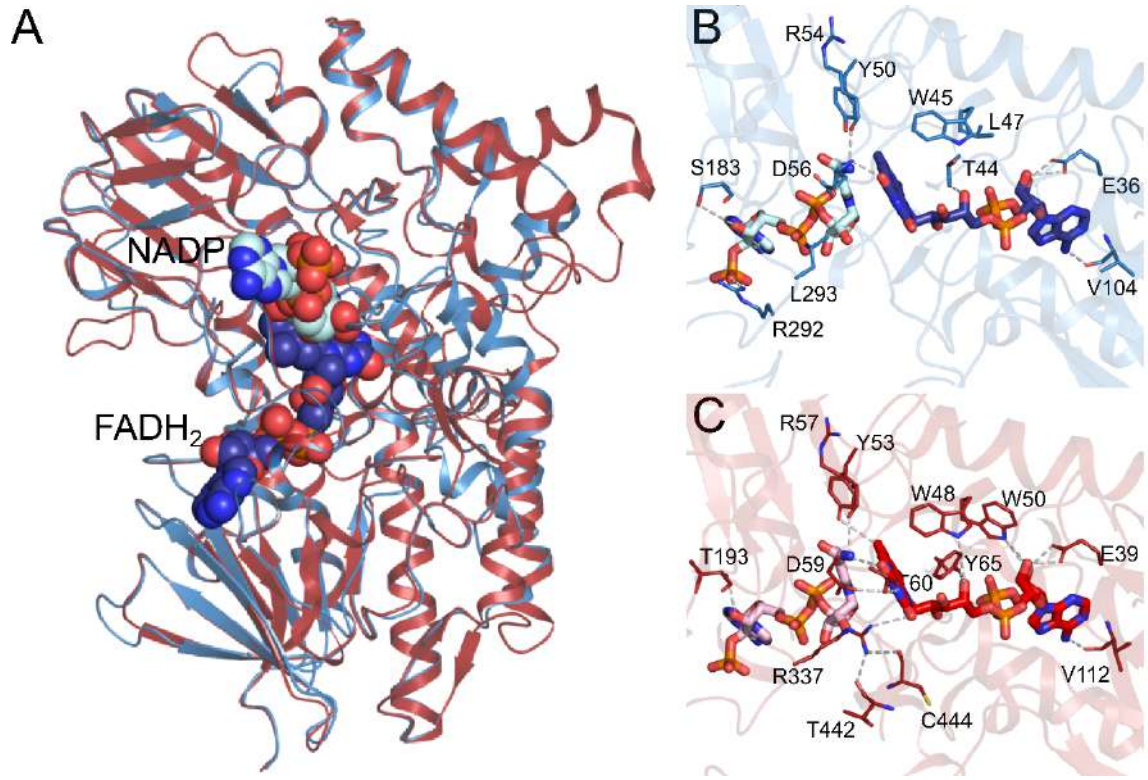


Figure 1. Comparison of EthA and OTEMO sequence and features. (**A**) Sequence alignment between EthA and 3UOZ visualized by Pymol Schrodinger llc v2.1.0. Rectangles and arrows represent helices and strands, respectively. Residues mapped by PLIP [23] as contributing to the $FADH_2$ and NADP binding sites are shown in cyan and green, respectively; catalytic arginines are shown in purple; residues in pink contribute to ETH binding in the *Acinetobacter radioresistens* EthA homolog [26]. Yellow, BVMO motif (EthA, F157-P168; 3UOZ, F160-P171); red, interdomain linker (EthA, G343-L348; 3UOZ, G388-T393); orange, control loop (EthA, G450-R470; 3UOZ, A496-R516). Secondary structure information was obtained with STRIDE [27]. (**B**) Domains and other features of OTEMO and EthA. Yellow, BVMO motif; red, interdomain linker; orange, control loop; black: OTEMO exible regions or that and/or undergo conformational transitions in crystal structures. Asterisks: amino acids that are part of the FAD (blue) and NADPH (green) binding sites in OTEMO structures (3UOZ, 3UOY, 3UOV, 3UOX, 3UP4, 3UP5) as mapped by PLIP and according to [20]. OTEMO domains are as in [20]. An alignment between PAMO, CHMO, STMO, OTEMO, and EthA [18] was used to infer EthA domains and EthA and OTEMO mobile functional regions.

Figure 2 shows the 3D alignment between the EthA model and 3UOZ. The amino acids contributing to $FADH_2$ and NADP binding in OTEMO are mostly conserved in EthA (Figure 2b,c), but there are two important di erences. First, in OTEMO, T442 and C444-V446 hold catalytic R337 near the $FADH_2$ isoalloxazine ring; this stretch of amino acids is missing in EthA, and, is not conserved in other BVMOs [18]. Second, the catalytic arginine, R292, is at a di erent position in EthA, possibly as a result of the absence of the above-mentioned amino acids. Alternatively, the di erence in position of the catalytic arginine may be in line with the observation that it adopts two conformations, competent for either intermediate stabilization



or for allowing an NADPH arrangement that is competent for reducing FAD [28].

Figure 2. 3D structure and active site of EthA and 3UOZ. (**A**) Structural alignment between the EthA model (blue) and 3UOZ (red) $FADH_2$ (blue) and NADP (cyan) are represented by spheres. The $FADH_2$ and NADP binding region is shown for (**B**) EthA and (**C**) OTEMO.

2.2. Clustering and Molecular Docking

MD simulations and clustering analysis helped improve and re ne EthA model in complex with cofactors and ligand ETH. We checked which EthA residues or regions underwent structural uctuations via RMSF calculations (Figure 3) and observed that residues M1, R483 and the loop P486-V489 exhibited the highest uctuations. A contact matrix with other residues up to 10 Å away was used to analyze conformational changes.

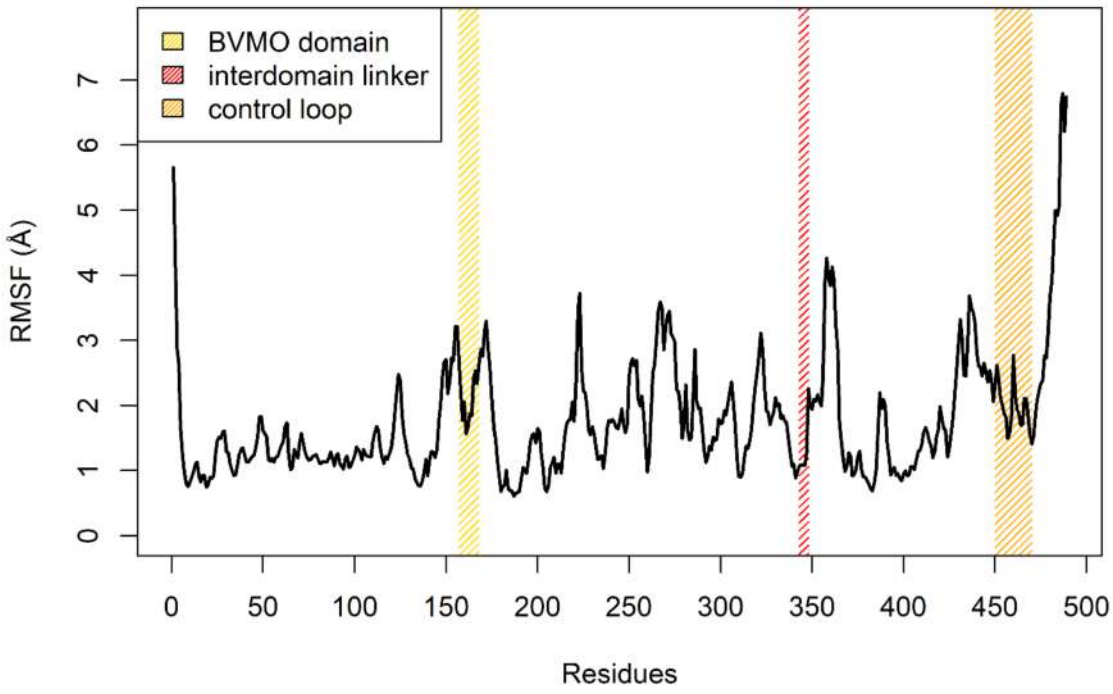


Figure 3. Root Mean Square Fluctuation (RMSF) to each residue of EthA. The positions of helices (black) and strands (grey) are indicated on the top and bottom axes of the uctuation plot.

Based on high dimensional data from contact matrices, the PCA and Spectral methods were applied to obtain a new dimensional space (intrinsic space), and Ward clustering was performed. The BVMO motif, interdomain linker, and control loop interact during BVMO catalysis [24]. Thus, the criteria to select representative structures in the clusters obtained was the position of these functional regions relative to each other. The lowest energy conformations from two clusters (01346 and 13841) and the conformation with the overall lowest energy (14053), as detected by the Spectral method, were chosen. While in conformation 01346 the control loop is away from the interdomain linker, in conformations 13841 and 14053 these regions are proximal (Figure 4a–c).

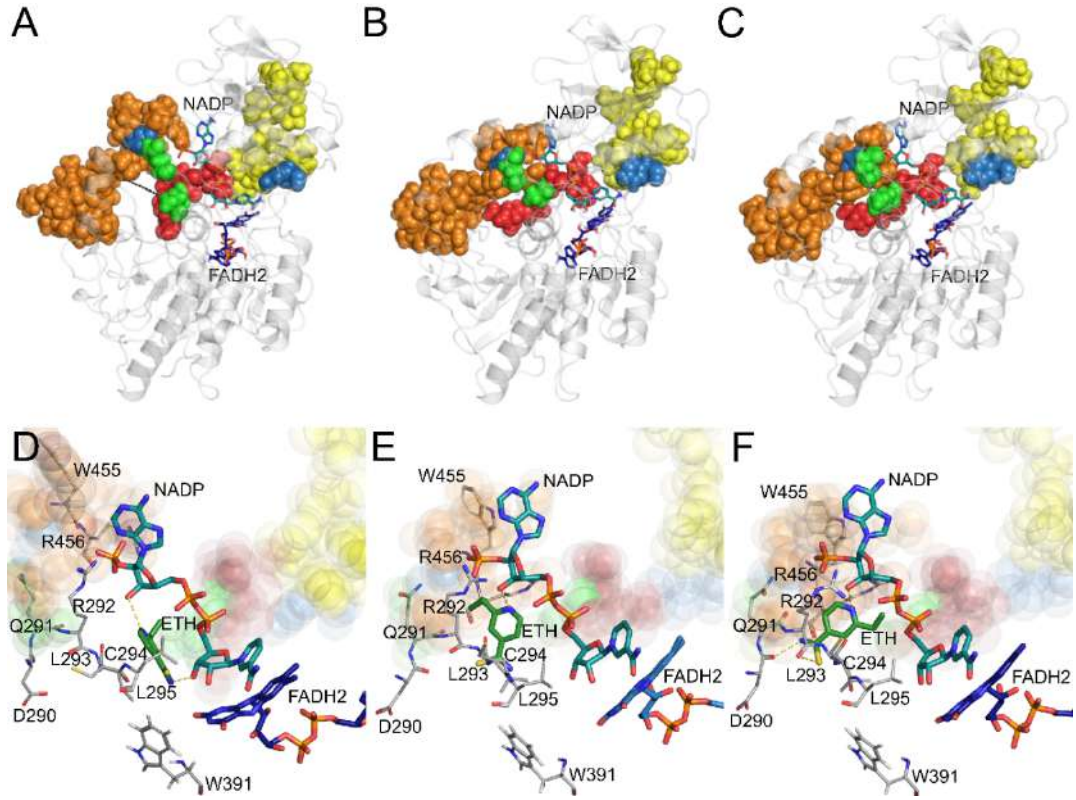


Figure 4. Control loop position and ETH docking in the selected conformations from the Spectral clustering results. BVMO motif (yellow), interdomain linker (red), control loop (orange), ETH binding region (green) are represented by spheres. Blue, Y50 and T453I. NADP (cyan) and $FADH_2$ (blue) are shown in stick representation. The distance between the interdomain linker (L348:CA) and control loop (N460:CA) in conformations (**a**) 01346 (12.7 Å), (**b**) 13841 (8.5 Å), and (**c**) 14053 (9.4 Å) are represented by black dashes. Best poses of ETH obtained from docking results in the selected conformations from the Spectral clustering results. (**d**) 01346, (**e**) 13841 and (**f**) 14053.

ETH was docked into the region encompassed by amino acids R292-L295, whose interaction with ETH has been proposed for an EthA homolog in *A. radioresistens* [26]. The putative binding site includes R292 of EthA, a conserved catalytic arginine (Figure 4d–e). Con gurations in which the interdomain linker and control loop remained closer (13841 and 14053) were also the ones yielding lower values of interaction energy, indicating more stable ETH binding (Table 1). In the *A. radioresistens* EthA homolog model, the interaction energy of ETH upon docking was higher [26], indicating a less stable interaction than found here for *M. tuberculosis* EthA. The docked systems were used for further characterization of the ETH-EthA interaction. It is interesting to point out that in con gurations with proximal interdomain linker and control loop (13841 and 14053) ETH interacted with R456. This arginine is adjacent to conserved W455, whose movement together with the rest of the control loop has been described as essential in BVMO catalysis [20,24].

Table 1. ETH docking results show more favorable interaction with EthA conformations 13841 and 14053.

EthA Conformation	ETH Interaction	Interaction Energy (Kcal/mol)
01346	NADP	–4.9
13841	R456	–5.6
14053	D290, Q291, L293, R456, NADP	–5.5

2.3. Assessing the Dynamic Properties of the EthA in Complex with ETH

The structural stability of ETH in di erent EthA conformations was evaluated by comparing the average RMSD of the systems during MD simulations (Table 2). During simulations, all systems presented EthA and NADP RMSD deviation of around 2 Å and 1.3 Å, respectively, showing a steady behavior throughout the triplicates (Figures S1 and S2). $FADH_2$ remained stable in the 13841 system (1.4 ± 0.3 Å) and less in 14053 (2.9 ± 0.7 Å). (Figure S2). RMSD values of ETH in the 14053 system remained low when compared to 01346 and 13841. Average and standard deviations were 2.9 ± 1.1 Å, 7.0 ± 1.2 Å, and 4.9 ± 3.0 Å, respectively (Figure S1).

Table 2. Root mean square deviations (Å) of EthA systems.

Molecules	01346 (Å)	13841 (Å)	14053 (Å)
EthA	2.3 ± 0.3	2.1 ± 0.4	2.0 ± 0.4
ETH	7.0 ± 1.2	4.9 ± 3.0	2.9 ± 1.1
NADP	1.6 ± 0.3	1.2 ± 0.3	1.2 ± 0.2
$FADH_2$	1.7 ± 0.8	1.4 ± 0.3	2.9 ± 0.7

We also calculated the hydrogen bond occupancy regarding pairs of residues involved in critical interactions (Table 3). System 01346 formed hydrogen bonds with di erent residues (C294, A237, and W240) and low occupancy values. ETH hydrogen-bonded to the residues (Q291, T453, and R456) in both the 13841 and 14053 systems. The occupancy values observed in the 14053 system were higher, reaching up to 64%.

Table 3. ETH interacts stably with conformation 14053.

Residue Pairs		Occupancy (%)		
Donor	Acceptor	01346	13841	14053
R456-Side	ETH	-	27.55	64.46
ETH	Q291-Main	0.11	13.51	62.47
ETH	T453-Side	-	11.79	26.12
ETH	C294-Main	12.70	-	-
ETH	A237-Main	19.24	-	-
W240-Main	ETH	10.59	-	-

The time evolution of the RMSD values show that ETH su ered minor conformational changes in system 14053 when compared to the others. Thus, to inspect the con gurational evolution of ETH over time, a clustering analysis was performed using all trajectories with a cut-o of 1.5 Å (Figure 5). ETH mobility and instability in systems 01346 and 13841 contributed to a broad exploration of the pocket, resulting in entirely di erent poses in the replicates. For the 14053 system, out of 72 clusters reported, the most signi cant (cluster 1) accounted for 58% of movements in the triplicates. In this system, ETH hydrogen-bonded to Q291, T453, and R456, in line with hydrogen bond occupancy (Table 3). Thus, a convergent con guration of ETH in the replicates was found in the 14053 system.

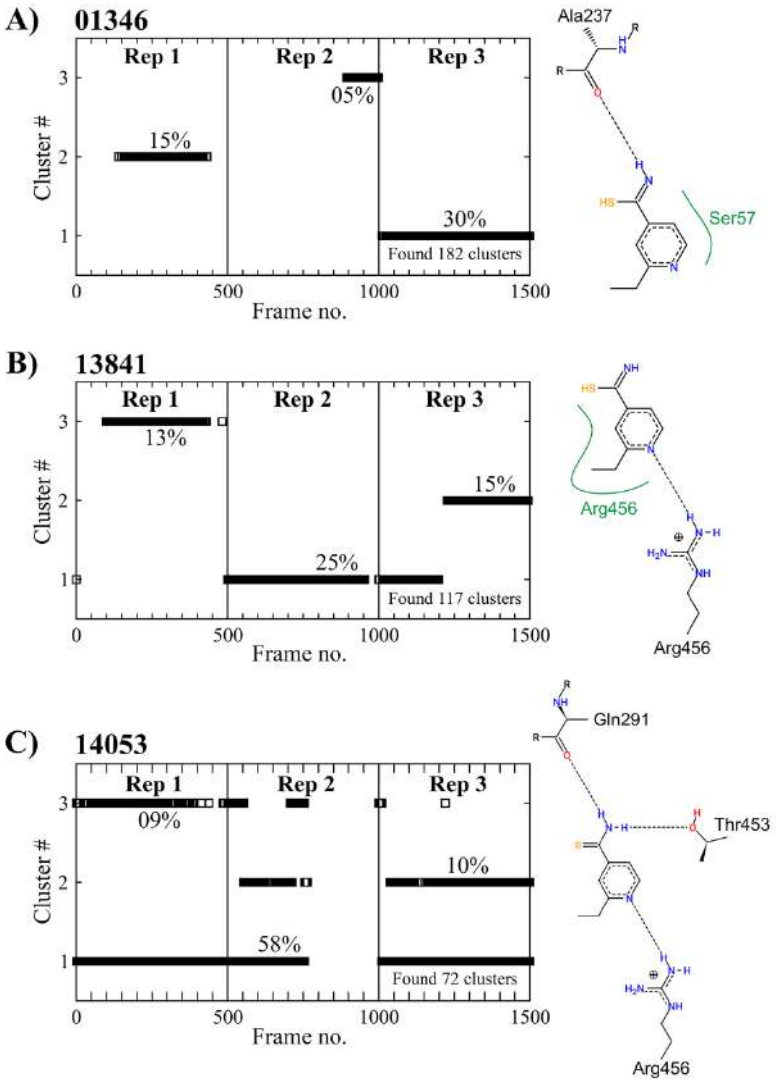


Figure 5. ETH clustering analysis from the MD simulations for systems (**a**) 01346, (**b**) 13841 and (**c**) 14053. The fractional contribution of the rst three clusters is shown in each graph. 2D representation of Protein-ETH interaction of the most representative structures resulting from clustering analysis is displayed next to each plot

(Poseview server). Black dashed lines and full green lines show hydrogen bonds and hydrophobic interactions, respectively.

In OTEMO, two con gurations of the control loop, β-hairpin and closed, result in an open or closed active site, respectively. Between these con gurations the control loop position shifts dramatically, and in the closed con guration, W501, a BVMO conserved residue, moves 9 Å to make a hydrogen bond with NADP (in agreement with experimental evidence for the role of this residue in catalysis [24]). The shift in the control loop, speci cally in the adjacent amino acid W502, is associated with changes in the conformation of the amino acids around catalytic R337 [20]. The observations that in system 14053 residue R456 (adjacent to W455, equivalent to W501) interacts with ETH (Tables 1–3; Figure 5) and that the interdomain linker and control loops are in contact are coherent with the importance of this mobile region. Also, these observations open speculation about a role for substrate binding in triggering the control loop movement.

Figure 6 presents the last ETH stabilization frame docked in the 14053 system after the molecular dynamics simulation. Despite the important role attributed to OTEMO W501 in NADP binding upon approximation between the control loop and interdomain linker [20], W455 in EthA remains distant from NADP throughout the molecular dynamics. Instead, R456 makes hydrogen bonds with NADP and ETH:N, suggesting a role in catalysis in place of W455. Interestingly, the conserved catalytic arginine, R292, appears to play an essential role in stabilizing NADP and forming the EthA-ETH complex. The R292-ETH:S interaction would favor electron transfer from the $FADH_2$-NADP system. Prior to molecular dynamics the position of R292 was away from NADP (Figure 2b,c), showing the importance of the molecular dynamics experiments to reveal EthA features. The model was built on an alignment of sequences with limited similarity, and the approach of optimizing the structure has replicated this important feature in BVMO catalysis.

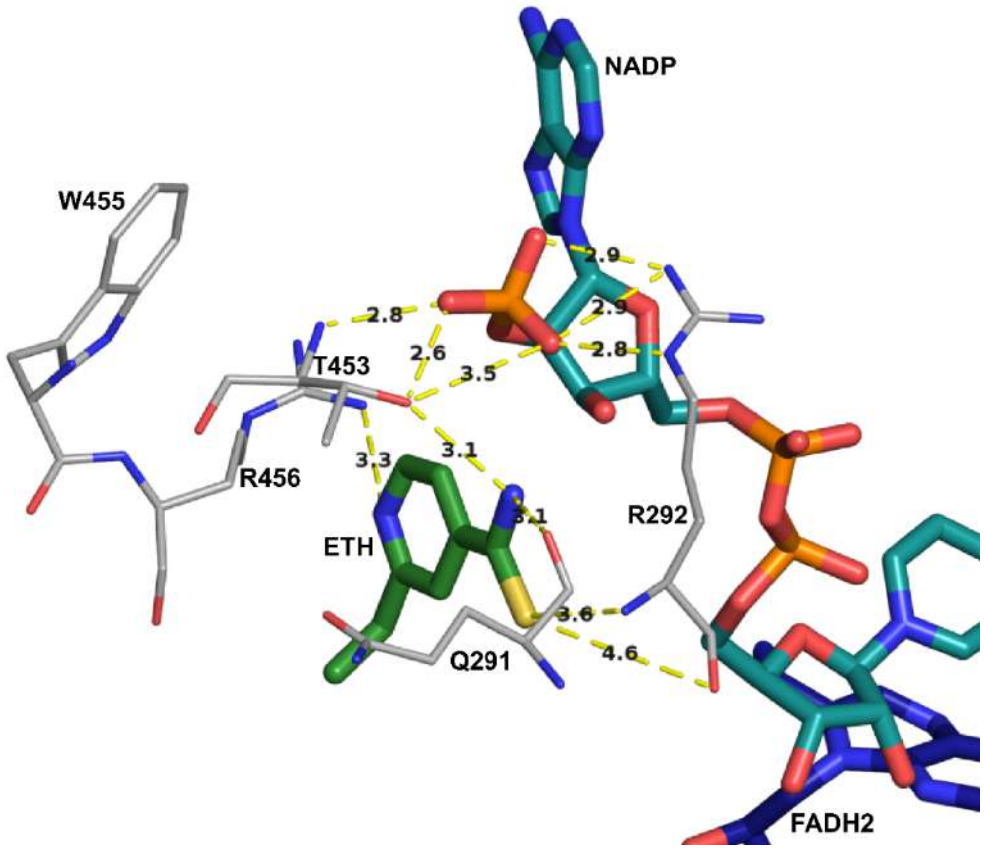


Figure 6. Representation of EthA-ETH complex stabilized during molecular dynamics. ETH (green), NADP (cyan), $FADH_2$ (blue), and residues (gray) involved in complex stabilization and catalytic mechanism are shown in stick representation. The distance (Å) between ETH, NADP and EthA residues are represented by yellow dashes.

2.4. Free Energy Changes for Y50C and T453I EthA Mutants

To evaluate the impact of mutations found in ETH^R clinical isolates, we chose to perform two relative alchemical free energies calculations in the FAD-binding domain. The in uence of mutations Y50C and T453I in ligand binding (NADP, $FADH_2$, and ETH) were computed using thermodynamic integration (TI). The rationale for choosing these mutations is as follows. The Y50C mutation has been found in three ETH^R clinical isolates [29] and Y50 is conserved in other BVMOs for which three-dimensional structures are available (CHMO, PAMO, OTEMO, and STMO; [18]). It is part of the EthA region that concentrates a large number of mutations in ETH^R isolates and which displays 67% similarity to an OTEMO stretch rich in $FADH_2$ and NADP binding amino acids [18]. Also, mutation of the equivalent amino acid in OTEMO, Y53, to phenylalanine, reduces catalysis to 30% [20]. T453 is in the control loop, and the T453I mutation has been identi ed in two clinical isolates [29]. Also, in system 14053 ETH hydrogen bonds with T453 and is in contact with NADP (Table 3; Figures 5

and 6). Thus, to test the in uence of mutations Y50C and T453I in ligand binding (NADP, $FADH_2$, and ETH), we used TI alchemical transformations.

Based on MD analysis showing stable ETH conformation in the -14053 system, we selected the most representative structure for TI simulations. When the Y50 ↔ C50 and T453 ↔ I453 transformations were assessed, positive values of ΔΔG of 3.34 ± 0.55 and 8.11 ± 0.51, respectively, were achieved (Table 4). Positive ΔΔGs implies substituting Y50 and T453 by C50 and I453 would be unfavorable to NADP, $FADH_2$, and ETH binding. Electrostatic transformation (Table 4) achieved similar ΔG values for both systems showing that charge a ected the systems equally. However, a substantial change in the van der Waals (vdW) transformation could be noticed for the NADP, $FADH_2$, and ETH bound systems. Spatially, Y50 is located at distances of 3.3 Å, 6.8 Å and 14 Å from NADP, $FADH_2$ and ETH, respectively, suggesting that the ΔΔG value in the Y50C transformation is unlikely to be related to ETH. Y50 lines the $FADH_2$-NADP pocket but mapping of their contacts by PLIP does not show bonding to either molecule. Thus, the reduced stability of the EthA Y50C complex detected by TI together with the reduced catalytic activity measured for the Y53F OTEMO mutant (without detectable loss of NADP a nity [20]) indicate that mutation in this position has a de nite, but indirect, role in catalysis. By contrast, distances between T453 and NADP and ETH were 1.8 Å and 3.2 Å, respectively (Figure 7). Thus, in this case the interactions with ETH and NADP are likely to contribute more to the ΔΔG of transformation T453I, which was more unfavorable than Y50C (8.11 and 3.34 kcal/mol, respectively) and resulted in a marked loss of complex stability. The underlying reason for this is that T453 interacts directly with ETH, via hydrogen bonding. The change to isoleucine likely destabilizes this bond, compromising the a nity between the substrate and the enzyme.

Table 4. Wild type EthA-FADH2-NADP-ETH complex is more stable than mutant complexes. TI outcomes for forward and backward paths in the two-step approach to determining the free energy change ($\Delta\Delta G = \Delta G_{HOLO} - \Delta G_{APO}$) for Y50C and T453I.

		Y50 ↔ C50		**T453 ↔ I453**	
		APO	HOLO	APO	HOLO
Forward	$\Delta G_{recharge}$	−17.92 ± 0.09	−17.45 ± 0.09	−27.29 ± 0.08	−29.32 ± 0.08
	ΔG_{vdw}	123.82 ± 0.27	127.45 ± 0.27	144.34 ± 0.22	155.08 ± 0.25
Backward*	$\Delta G_{recharge}$	−19.40 ± 0.08	−19.43 ± 0.08	−27.52 ± 0.07	−30.08 ± 0.07
	ΔG_{vdw}	124.93 ± 0.28	127.55 ± 0.23	144.20 ± 0.24	154.29 ± 0.27
	Final	105.72 ± 0.41	109.06± 0.38	116.87± 0.34	124.98 ± 0.38
	ΔΔG (kcal/mol)	**3.34 ± 0.55**		**8.11 ± 0.51**	

* Backward values are consistent with the forward process.

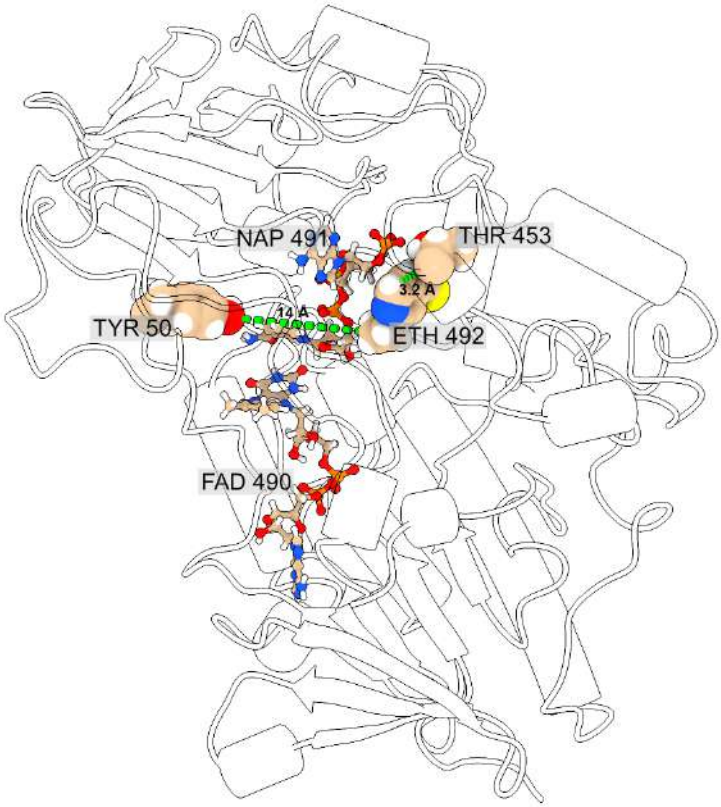


Figure 7. The spatial location of residues chosen for TI alchemical transformations in the most representative structure of EthA-14053 system. ETH is located at 14 Å and 3.2 Å from T50 and T453, respectively.

3. Materials and Methods

3.1. Comparative Modeling

Three-dimensional models of EthA (UNIPROT P9WNF9) were modeled using the program Modeller v9.21 [21,22]. The template was selected based on local alignment between the EthA sequence and of proteins deposited in Protein Data Bank (PDB). The structure 3UOZ was selected based on sequence identity, similarity, and maximum query coverage parameters. The sequence alignment between EthA and template was generated using ClustalΩ [30]. To verify the secondary structure consensus areas between the target sequence and template, the following programs were used: PSIPRED, NetSurfP, Jpred3, PORTER, SCRATCH and Jufo9D. Also, cysteine disul de bond analysis was performed by Cyspred and Disul nd programs. Based on the consensus regions predicted by these programs, secondary structure constraints were inserted during the modeling process via Modeller v9.21, performing two cycles of very slow optimization steps of VTFM and MD. 3D models were built considering all heteroatoms from the template. The best model was chosen according DOPE-HR and molpdf energies calculated by Modeller, structural quality evaluated by

ProSA-web and Whatcheck, and stereochemical quality assessed by Procheck and Molprobity programs.

3.2. Molecular Dynamics (MD)

MD simulations were carried out using AMBER 18.0 [31,32], and protein interactions were represented using 14SB force eld [33]. Bonded, electrostatic and Lennard-Jones parameters for ligands (NADP $FADH_2$ and ETH) were obtained using the generalized amber force eld (GAFF) [34] and AM1-BCC [35] tools while atomic partial charges were added using ANTECHAMBER [36]. Electrostatic interactions were treated using the Particle-Mesh Ewald (PME) algorithm with a cut-o of 10 Å. Each system was simulated in an octahedral box lled with TIP3P water molecules [37] under periodic boundary conditions, considering a distance of 14 Å from the outermost protein atoms in all cartesian directions. Protonation states of tritable residues were assigned using PDB2PQR software version 2.1.1. All systems were neutralized by adding 4 Cl^- counterions. Subsequently, a two-step energy minimization procedure was performed: (i) 2000 steps (1000 steepest descent + 1000 conjugate-gradient) with all heavy atoms harmonically restrained with a force constant of 5 kcal mol^{-1} $Å^{-2}$; (ii) 5000 steps (2500 steepest descent + 2500 conjugate-gradient) without position restraints. Next, initial atomic velocities were assigned using a Maxwell-Boltzmann distribution corresponding to an initial temperature of 20 K and the systems were gradually heated to 300 K over one nanosecond utilizing the Langevin thermostat. During this stage, all heavy atoms were harmonically restrained with a force constant of 10 kcal mol^{-1} $Å^{-2}$. Systems were subsequently equilibrated during nine successive 500 ps equilibration simulations where position restraints approached zero progressively. After this period, the systems were simulated with no restraints at 300 K in the Gibbs ensemble with a pressure of 1 atm. Two MD processes were performed: (i) a 500 ns simulation of the model (EthA, NADP, and $FADH_2$); (ii) Three independent MD simulations of 100 ns each of the ETH bound to protein in di erent conformations derived from clustering analysis (EthA, NADP, $FADH_2$, and ETH). Simulation trajectories were analyzed with GROMACS package tools version 2019.3 [38]. Root-mean-square deviation (RMSD) values were calculated separately for each system tting their backbone atoms, taking the initial structure of the production dynamics as a reference. Conformational clusterization for ETH was performed using the GROMOS method with a cut-o of 1.5 Å considering all atoms. Hydrogen bond formation was de ned using a geometric criterion with VMD software version 1.9.3. We considered a hit when the distance between two polar heavy atoms, with at least one hydrogen atom attached, was less than 3.5 Å using a D-Ĥ-A angle cuto of 30°.

3.3. Clustering and Free Energy Landscape

For a given set of 25,000 conformations of EthA from MD simulations, a contact matrix between residues was used as internal coordinates to analyze and detect structural cluster

centers. In order to reduce computational complexity, the root mean square fluctuation (RMSF) for each residue was calculated, and ones with values above 5 Å were considered to calculate the contact matrix using 10 Å as distance cutoff. The Spectral and PCA methods of dimensionality reduction were used to find out the intrinsic space prior to performing clustering. The obtained space was used to cluster protein conformations by ward algorithm. Elbow method was applied to determine the number of groups parameter in the ward algorithm. The free energy landscape (FEL) was calculated using the Weighted Histogram Analysis Method (WHAM) for each conformational set from MD simulations. According to this method, the bins of the histogram, obtained by discrete states of a molecule, provide a relative probability that a state occurs along the trajectory and regions with a higher density of states represent an energy basin [39]. Here, we calculated the FEL using intrinsic space found out by PCA and Spectral methods for two dimensions.

3.4. Docking

All molecular docking simulations were performed using Autodock vina software version 1.1.2 [40]. Using the AutoDock Tools (ADT) v 1.5.6, all hydrogens and Gasteiger charges were added to the EthA model for grid generation and docking. The grid was created with center coordinates in X = 44.180, Y = 47.593 and Z = 51.165 and size was X = 40 Å and Y = Z = 30 Å. During the grid preparation, the side chain of residues R292, L293, C294 and L295 were considered flexible. The 3D structure of ETH was downloaded from PubChem database [code: 2761171] and prepared using ADT software version 1.5.6., with the addition of Gasteiger charges and torsions, to allow flexibility.

3.5. Thermodynamic Integration (TI)

Free energy changes upon mutation of tyrosine to cysteine (Y50C) and threonine to isoleucine (T453I) for system 14053, were evaluated by TI to check how the mutation affects NADP $FADH_2$ and ETH binding. The newly GPU implementation, pmemdGTI [41], in AMBER 18.0 [32,33] was used in 11 equally spaced λ-windows from λ = 0 to λ = 1. The ΔΔG variation was calculated based on a thermodynamic cycle using two structures, a ligand-free 14053 and 14053 bound to NADP, $FADH_2$ and ETH (Figure S3) in aqueous solution. Computational calculations of thermodynamic Integrations were carried out using dual topology procedure. Alchemical transformations of λ-dependent potentials involving Coulomb and Lennard-Jones terms [42,43] were calculated in a set of λ-windows equally spaced in intervals of 0.1. Initial configurations were submitted to 1000 steps of the steepest descent algorithm. Thermalization stage was performed varying temperature from 0 K to 300 K over 100 ps in the NVT ensemble, followed by the equilibration stage in the NPT ensemble for 250 ns at a temperature of 300 K and a pressure of 1 atm. The electrostatic and vdW transformations took 10 ns for each λ-window, although only the last 9 ns were computed for calculations. For the analysis, the Alchemical analysis python package [44] was used to

calculate the free energy changes and corresponding errors. Final free energy change along the path was computed as a weighted sum of the ensemble averages of the derivative of the potential energy function with respect to λ using the trapezoidal rule and averaged by forward (λ = 0➔1) and backward (λ = 1➔0) paths.

4. Conclusion

A model for EthA, the main protein involved in resistance to ETH in *M. tuberculosis*, was built for the first time. MD simulations offered structural insight about the position of the control loop and showed that ETH binding to EthA involves contacts with the control loop, suggesting a role for substrate binding in control loop movement. TI calculations reveal that two mutations found in *M. tuberculosis* ETH^R clinical isolates, Y50C and T453I, result in lower stability of the mutant enzyme-ligands complex. The results indicate an essential role for T453 in the catalytic mechanism of EthA, interacting with NADP and ETH, and the evaluation of its mutation to isoleucine in this work showed a greater destabilization of the EthA-ETH complex. The results presented in this work shed light on the residues involved in the catalytic mechanism of EthA of *M. tuberculosis* and on the importance of mutants Y50C and T453I. These first steps helped to guide future experimental work and complementary computational studies.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: Protein and ETH RMSD of the EthA systems, **Figure S2:** NADP and $FADH_2$ RMSD of the EthA systems, **Figure S3:** Thermodynamic cycle of the forward free energy change upon a transformation of a wild residue (Y50 and T453) to a mutant residue (C50 and I453).

Author Contributions: A.C.R.G. and T.C.G. initiated the concept; A.C.R.G., D.A., E.R.C., P.V.Z.C.G. and T.C.G. designed and conceived the experiments; D.A., L.H.S, P.V.Z.C.G. and V.C.d.S performed the experiments; A.C.R.G., D.A., E.R.C., L.H.S, P.V.Z.C.G., T.C.G. and V.C.S analyzed the data; D.A., T.C.G. and V.C.S drafted the manuscript; E.R.C, A.C.R.G., P.V.Z.C.G. and T.C.G. reviewed the manuscript prior to submission. All authors approved the final manuscript.

Funding: This study was financed by Fiocruz, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES; Finance Code 001) and CNPq.

Acknowledgments: We thank CAPES and CNPq for continued support to students.

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

References

1. WHO. *Global Tuberculosis Report 2018*; World Health Organization: Geneva, Switzerland, 2018.
2. Gilpin, C.; Korobitsyn, A.; Migliori, G.B.; Raviglione, M.C.; Weyer, K. The World Health Organization standards for tuberculosis care and management. *Eur. Respir. J.* **2018**, *51*, 1800098.
3. API Consensus Expert Committee. API TB Consensus Guidelines 2006: Management of pulmonary tuberculosis, extrapulmonary tuberculosis and tuberculosis in special situations. *J. Assoc. Physicians India* **2006**, *54*, 219–234.
4. Ramachandran, G.; Swaminathan, S. Safety and tolerability profile of second-line anti-tuberculosis medications. *Drug Saf.* **2015**, *38*, 253–269.
5. Baulard, A.R.; Betts, J.C.; Engohang-Ndong, J.; Quan, S.; McAdam, R.A.; Brennan, P.J.; Locht, C.; Besra, G.S. Activation of the pro-drug ethionamide is regulated in mycobacteria. *J. Biol. Chem.* **2000**, *275*, 28326–28331.
6. DeBarber, A.E.; Mdluli, K.; Bosman, M.; Bekker, L.G.; Barry, C.E., 3rd. Ethionamide activation and sensitivity in multidrug-resistant *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 9677–9682.

7. Vannelli, T.A.; Dykman, A.; Ortiz de Montellano, P.R. The antituberculosis drug ethionamide is activated by a avoprotein monooxygenase. *J. Biol. Chem.* **2002**, *277*, 12824–12829.
8. Grant, S.S.; Wellington, S.; Kawate, T.; Desjardins, C.A.; Silvis, M.R.; Wivagg, C.; Thompson, M.; Gordon, K.; Kazyanskaya, E.; Nietupski, R.; et al. Baeyer-Villiger monooxygenases EthA and MymA are required for activation of replicating and non-replicating *Mycobacterium tuberculosis* inhibitors. *Cell Chem. Biol.* **2016**, *23*, 666–677.
9. Blondiaux, N.; Moune, M.; Desroses, M.; Frita, R.; Flipo, M.; Mathys, V.; Soetaert, K.; Kiass, M.; Delorme, V.; Djaout, K.; et al. Reversion of antibiotic resistance in *Mycobacterium tuberculosis* by spiroisoxazoline SMARt-420. *Science* **2017**, *17*, 1206–1211.
10. Hicks, N.D.; Carey, A.F.; Yang, J.; Zhao, Y.; Fortune, S.M. Bacterial genome-wide association identi es novel factors that contribute to ethionamide and prothionamide susceptibility in *Mycobacterium tuberculosis*. *mBio* **2019**, *10*, e00616–e00619.
11. Vilchèze, C.; Av-Gay, Y.; Attarian, R.; Liu, Z.; Hazbón, M.H.; Colangeli, R.; Chen, B.; Liu, W.; Alland, D.; Sacchettini, J.C.; et al. Mycothiol biosynthesis is essential for ethionamide susceptibility in *Mycobacterium tuberculosis*. *Mol. Microbiol.* **2008**, *69*, 1316–1329.
12. Laborde, J.; Deraeve, C.; Bernardes-Génisson, V. Update of antitubercular prodrugs from a molecular perspective: Mechanisms of action, bioactivation pathways, and associated resistance. *Chem. Med. Chem.* **2017**, *12*, 1657–1676.
13. Ang, M.L.; Siti, Z.Z.; Shui, G.; Dianišková, P.; Madacki, J.; Lin, W.; Koh, V.H.; Gomez, J.M.; Sudarkodi, S.; Bendt, A.; et al. An *ethA-ethR*-de cient *Mycobacterium bovis* BCG mutant displays increased adherence to mammalian cells and greater persistence *in vivo*, which correlate with altered mycolic acid composition. *Infect. Immun.* **2014**, *82*, 1850–1859.
14. Toriyama, S.; Imaizumi, S.; Tomiyasu, I.; Masui, M.; Yano, I. Incorporation of ^{18}O into long-chain, secondary alcohols derived from ester mycolic acids in *Mycobacterium phlei*. *BBA Lipid Lipid Met.* **1982**, *712*, 427–429.
15. Kamerbeek, N.M.; Janssen, D.B.; van Berkel, W.J.H.; Fraaije, M.W. Baeyer–Villiger monooxygenases, an emerging family of avin-dependent biocatalysts. *Adv. Synth. Catal.* **2003**, *345*, 667–678.
16. Torres Pazmiño, D.E.; Baas, B.-J.; Janssen, D.B.; Fraaije, M.W. Kinetic mechanism of phenylacetone monooxygenase from *Thermobi da fusca*. *Biochemistry* **2008**, *47*, 4082–4093.
17. Leisch, H.; Morley, K.; Lau, P.C.K. Baeyer−Villiger monooxygenases: More than just green chemistry. *Chem. Rev.* **2011**, *11*, 4165–4222.
18. da Silva, D.A.; Ferreira, N.V.; Rego, A.M.; Barbosa, P.C.; Machado, R.F.; Pimentel, A.; dos Reis, L.M.; de Pina, L.C.; Redner, P.; de Souza Caldas, P.C.; et al. Integrated analysis of ethionamide resistance loci in *Mycobacterium tuberculosis* clinical isolates. *Tuberculosis* **2018**, *113*, 163–174.
19. Fraaije, M.W.; Kamerbeek, N.M.; Heidekamp, A.J.; Fortin, R.; Janssen, D.B. The prodrug activator EtaA from *Mycobacterium tuberculosis* is a Baeyer-Villiger monooxygenase. *J. Biol. Chem.* **2004**, *279*, 3354–3360.
20. Leisch, H.; Shi, R.; Grosse, S.; Morley, K.; Bergeron, H.; Cygler, M.; Iwaki, H.; Hasegawa, Y.; Lau, P.C. Cloning, baeyer-villiger biooxidations, and structures of the camphor pathway 2-oxo-δ3-4, 5, 5-trimethylcyclopentenylacetyl-coenzyme a monooxygenase of *Pseudomonas putida* ATCC 17453. *Appl. Environ. Microbiol.* **2012**, *78*, 2200–2212.
21. Šali, A.; Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779–815.
22. Webb, B.; Sali, A. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinform.* **2016**, *54*, 5–6.
23. Salentin, S.; Schreiber, S.; Haupt, V.J.; Adasme, M.F.; Schroeder; M. PLIP: Fully automated protein–ligand interaction pro ler. *Nucleic Acids Res.* **2015**, *43*, W443–W447.
24. Mirza, I.A.; Yachnin, B.J.; Wang, S.; Grosse, S.; Bergeron, H.; Imura, A.; Iwaki, H.; Hasegawa, Y.; Lau, P.C.; Berghuis, A.M. Crystal structures of cyclohexanone monooxygenase reveal complex domain movements and a sliding cofactor. *J. Am. Chem. Soc.* **2009**, *131*, 8848–8854.
25. Fürst, M.J.; Fiorentini, F.; Fraaije, M.W. Beyond active site residues: Overall structural dynamics control catalysis in avin-containing and heme-containing monooxygenases. *Curr. Opin. Struct. Biol.* **2019**, *59*, 29–37.
26. Minerdi, D.; Zgrablic, I.; Sadeghi, S.J.; Gilardi, G. Identi cation of a novel Baeyer-Villiger monooxygenase from *Acinetobacter radioresistens*: Close relationship to the *Mycobacterium tuberculosis* prodrug activator EtaA. *Microb. Biotech.* **2012**, *5*, 700–716.
27. Heinig, M.; Frishman, D. STRIDE: A Web server for secondary structure assignment from known atomic coordinates of proteins. *Nucl. Acids Res.* **2004**, *32*, W500–W502.

28. Malito, E.; Al eri, A.; Fraaije, M.W.; Mattevi, A. Crystal structure of a Baeyer–Villiger monooxygenase. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 13157–13162.
29. Minerdi, D.; Zgrablic, I.; Sadeghi, S.J.; Gilardi, G. Identi cation of a novel Baeyer-Villiger monooxygenase from *Acinetobacter radioresistens*: Close relationship to the *Mycobacterium tuberculosis* prodrug activator EtaA. *Microb. Biotech.* **2012**, *5*, 700–716.
30. Leung, K.L.; Yip, C.W.; Yeung, Y.L.; Wong, K.L.; Chan, W.Y.; Chan, M.Y.; Kam, K.M. Usefulness of resistant gene markers for predicting treatment outcome on second-line antituberculosis drugs. *J. Appl. Microbiol.* **2010**, *109*, 2087–2094.
31. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539.
32. Case, D.A.; Cheatham, T.E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K.M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R.J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
33. Case, D.A.; Ben-Shalom, I.Y.; Brozell, S.R.; Cerutti, D.S.; Cheatham, T.E., III; Cruzeiro, V.W.D.; Darden, T.A.; Duke, R.E.; Ghoreishi, D.; Gilson, M.K.; et al. *AMBER*; University of California: San Francisco, CA, USA, 2018
34. Maier, J.A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K.E.; Simmerling, C. 14SB: Improving the accuracy of protein side chain and backbone parameters from 99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.
35. Wang, J.; Wolf, R.M.; Caldwell, J.W.; Kollman, P.A.; Case, D.A. Development and testing of a general amber force eld. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
36. Jakalian, A.; Jack, D.B.; Bayly, C.I. Fast, e cient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, *23*, 1623–1641.
37. Wang, J.; Wang, W.; Kollman, P.A.; Case, D.A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247–260.
38. Jorgensen, W.L.; Chandrasekhar, J.; Madura, J.D.; Impey, R.W.; Klein, M.L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.
39. Abraham, M.J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.C.; Hess, B.; Lindah, E. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.
40. Kumar, S.; Rosenberg, J.M.; Bouzida, D.; Swendsen, R.H.; Kollman, P.A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
41. Trott, O.; Olson, A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, e cient optimization and multithreading. *J. Comp. Chem.* **2010**, *31*, 455–461.
42. Lee, T.-S.; Hu, Y.; Sherborne, B.; Guo, Z.; York, D.M. Toward fast and accurate binding a nity prediction with pmemdGTI: An e cient implementation of GPU-accelerated thermodynamic integration. *J. Chem. Theory Comput.* **2017**, *13*, 3077–3084.
43. Kaus, J.W.; Pierce, L.T.; Walker, R.C.; McCammon, J.A. Improving the e ciency of free energy calculations in the amber molecular dynamics package. *J. Chem. Theory Comput.* **2013**, *9*, 4131–4139.
44. Mermelstein, D.J.; Lin, C.; Nelson, G.; Kretsch, R.; McCammon, J.A.; Walker, R.C. Fast and exible gpu accelerated binding free energy calculations within the amber molecular dynamics package. *J. Comput. Chem.* **2018**, *39*, 1354–1358.
45. Klimovich, P.V.; Shirts, M.R.; Mobley, D.L. Guidelines for the analysis of free energy calculations. *J. Comput. Aided Mol. Des.* **2015**, *29*, 397–411.