Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Departamento de Química
Programa de Pós Graduação em Química

Tese de Doutorado

Estudo de interações Proteína-ligante e do Enovelamento de Proteínas através de Dinâmica Molecular, Modelos de Estado de Markov, Cálculos Quânticos e Descritores de Reatividade

Acassio Rocha Santos

Orientador: Prof. Dr. Gerd Bruno da Rocha

Coorientador: Dr. Gabriel Aires Urquiza de Carvalho

João Pessoa

-Dezembro de 2021-

Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Departamento de Química
Programa de Pós Graduação em Química

Tese Doutorado

Estudo de interações Proteína-ligante e do Enovelamento de Proteínas através de Dinâmica Molecular, Modelos de Estado de Markov, Cálculos Quânticos e Descritores de Reatividade

Acassio Rocha Santos

Tese de Doutorado submetida ao Programa de Pós Graduação em Química da Universidade Federal da Paraíba, como parte dos requisitos para obtenção do título de Doutor em Ciências, área de concentração: Físico-Química

Orientador: Prof. Dr. Gerd Bruno da Rocha

Coorientador: Dr. Gabriel Aires Urquiza de Carvalho

João Pessoa

-Dezembro de 2021-

Catalogação na publicação
Seção de Catalogação e Classificação

S237e Santos, Acassio Rocha.
Estudo de interações proteína-ligante e do enovelamento de proteínas através de dinâmica molecular, modelos de estado de Markov, cálculos quânticos e descritores de reatividade / Acassio Rocha Santos. - João Pessoa, 2021.
230 f. : il.

Orientação: Gerd Bruno da Rocha.
Coorientação: Gabriel Aires Urquiza de Carvalho.
Tese (Doutorado) - UFPB/CCEN.

1. Química. 2. Ricina. 3. QM/MM. 4. Enovelamento de proteínas. 5. MSMs. 6. Descritores globais. 7. Locais de reatividade. I. Rocha, Gerd Bruno da. II. Carvalho, Gabriel Aires Urquiza de. III. Título.

UFPB/BC CDU 54(043)

Elaborado por Larissa Silva Oliveira de Mesquita - CRB-15/746

Estudo de interações proteína-ligante e do enovelamento de proteínas através de dinâmica molecular, modelos de estado de Markov, cálculos quânticos e descritores de reatividade.

Tese de Doutorado apresentada pelo aluno Acássio Rocha Santaos e aprovada pela banca examinadora em 10 de dezembro de 2021.

Prof. Dr. Gerd Bruno da Rocha
Departamento de Química – CCEN/UFPB
Orientador/Presidente

Dr. Gabriel Aires Urquiza de Cavalho
Pesquisador junto ao DQF/UFPE
2º. Orientador

gov.br Documento assinado digitalmente
Leonardo Henrique Franca de Lima
Data: 16/12/2021 17:03:38-0300
Verifique em https://verificador.iti.br

Prof. Dr. Leonardo Henrique França de Lima
UFSJ – Sete Lagoas – MG
Examinador

Prof. Dr. Paulo Augusto Netz
Instituto de Química – UFRGS-RS
Examinador

Profa. Dra. Karen Cacilda Weber
Departamento de Química – UFPB-PB
Examinadora

Prof. Dr. Wagner de Mendonça Faustino
Departamento de Química – UFPB-PB
Examinador

Assinaturas da Banca realizadas em modo Webconferência em 10/12/2021, digitalizadas e certificadas pelo Prof. Dr. Gerd Bruno da Rocha (SIAPE 1520134) em 10/12/2021

AGRADECIMENTOS

- A Deus, por ter me dado forças e ânimo para não desistir dos meus sonhos e projetos.
- À minha esposa, Ana Virgínia, pelo seu amor, força, motivação e por estar presente em todas as etapas da minha carreira acadêmica. Sem você eu não teria conseguido chegar ao fim desta tese!
- Ao meu filho, Rafael, por tornar os meus dias mais felizes.
- À minha mãe, Zélia Rocha, pela confiança, força e amor.
- Ao meu pai, Dilmar Brito, por sempre acreditar em mim e em meus sonhos.
- Ao meu irmão, Éder Rocha, pelo apoio, pelas conversas constantes e por estar sempre próximo, apesar de estarmos em continentes diferentes.
- Ao meu sogro, Antonio José, pelo apoio durante o tempo em que estive em João Pessoa.
- À minha sogra, Maria Elivânia, pelo cuidado e pelas conversas no campus da UFPB.
- Ao meu amigo Daniel Marcos, pela amizade, pelo apoio e pelos conselhos, tanto em assuntos profissionais quanto pessoais.
- Ao meu orientador, professor Gerd Bruno da Rocha, pelas orientações, correções, auxílio e disponibilidade durante todo o meu doutorado.
- Ao meu coorientador, Gabriel Aires Urquiza de Carvalho, pelas contribuições para esta tese.
- Ao professor Jadson Cláudio Belchior, pela oportunidade de iniciação científica, mesmo sabendo da minha realidade de trabalho e estudos na época da graduação. Essa oportunidade fez toda a diferença na minha carreira acadêmica.
- Aos colegas de laboratório Igor Barden Grillo e Elton Chaves, pelas contribuições dadas para esta tese.
- Aos amigos Emerson, Larissa e José Cícero, pelas conversas e partilhas sobre pesquisa e sobre questões pessoais.

- À CAPES/FAPESQ pela bolsa concedida ao longo do doutorado.
- Ao NPAD-UFRN, CENAPAD-SP e LMMRQ-UFPB, pelo tempo computacional disponibilizado.
- Aos professores Karen Cacilda Weber, Wagner de Mendonça Faustino, Leonardo Henrique Franca de Lima, Paulo Augusto Netz e Sidney Ramos de Santana por aceitarem o convite para a composição da banca de doutorado.

Muito Obrigado!

RESUMO

Nesta tese, buscamos avaliar aspectos da estrutura eletrônica de sistemas biológicos através de cálculos termoquímicos e de descritores quânticos de reatividade (QCMDs - *Quantum Chemical Molecular Descriptors*). Esses estudos foram aplicados a dois problemas biológicos distintos: 1) interação proteína-ligante; 2) enovelamento de proteínas.

Em relação ao primeiro trabalho, realizamos um estudo de candidatos a inibidores da subunidade RTA da ricina, uma proteína citotóxica produzida na semente da mamona (*Ricinus communis*) e pertence à família de proteínas inativadoras de ribossomos (tipo 2). Trata-se de uma das toxinas biológicas mais potentes entre as conhecidas, sendo constituidas de duas subunidades, RTA e RTB, unidas por uma ponte de dissulfeto.

Realizamos um estudo para avaliar as interações entre a subunidade da toxina A da ricina (RTA) e alguns de seus inibidores, utilizando métodos modernos de Química Quântica semiempírica e Mecânica Quântica/Mecânica Molecular ONIOM (QM/MM). Duas abordagens foram seguidas: cálculo da entalpia de ligação, ΔH_{bind}, e descritores químico-quânticos de reatividade, QCMDs). Essas abordagens foram comparadas com dados experimentais da metade da concentração inibitória máxima (IC_{50}), a fim de obter informações sobre os inibidores RTA e verificar qual método químico-quântico descreve melhor as interações RTA-ligante. Descobrimos que o ΔH_{bind}, obtido via cálculos *single-point* de energia com os métodos semiempíricos PM6-DH+, PM6-D3H4 e PM7; bem como o ΔE_{bind}, obtido via método QM/MM ONIOM, apresentaram boa correlação com os dados de IC_{50}. Observamos, no entanto, que a correlação diminuiu significativamente quando calculamos o ΔH_{bind} após a otimização da geometria *full atom* com todos os métodos semiempíricos. Com base nos resultados dos cálculos dos QMCDs para os casos estudados, notamos que ambos os tipos de interações, sobreposição molecular e interações eletrostáticas, desempenham papéis significativos na afinidade geral desses ligantes para o sítio de ligação da RTA.

No segundo estudo, avaliamos aspectos da estrutura eletrônica no processo do enovelamento das proteínas NTL9, BBA e α3D. Nesse trabalho, avaliamos três proteínas de rápido enovelamento (NTL9, BBA e α3D) através de DM, MSMs e descritores globais e locais de reatividade, obtidos através de cálculos DFT-D3 e método semiempírico PM7.

Os resultados demonstraram que os MSMs foram capazes de caracterizar em detalhes diversos caminhos de enovelamento, fornecendo informação que possibilitaram inferir o tipo

de mecanismo dos sistemas estudados. Dados dos cálculos quânticos apontaram que a entalpia de formação é uma boa coordenada de reação para o processo do enovelamento. Além disso, foi observado que a integração de MD, DFT-D3, método semiempírico e QCMDs oferecem o potencial de revelar aspectos-chave no processo do enovelamento de proteínas que não são descritos por nenhuma outra abordagem. Verificamos ainda que a dureza local por resíduo é capaz de distinguir estruturas não nativas de estruturas nativas, revelando que aspectos intrínsecos da estrutura eletrônica desempenham um papel altamente relevante no processo de enovelamento de proteínas. Essa observação pode ser uma assinatura na estrutura eletrônica durante o enovelamento de proteínas.

Palavras-chave: Ricina, QM/MM, Enovelamento de Proteínas; MSMs ; Descritores Globais e Locais de Reatividade.

ABSTRACT

In this thesis, we seek to evaluate aspects of the electronic structure of biological systems through thermochemical calculations and quantum reactivity descriptors (QCMDs). These studies were applied to two distinct biological problems: 1) protein-ligand interaction; 2) protein folding.

Regarding the first work, we performed a study of candidates for ricin RTA subunit inhibitors, a cytotoxic protein produced in castor bean seeds Ricinus communis and belongs to the family of ribosome-inactivating proteins (type 2). It is one of the most potent biological toxins among those known, consisting of two subunits, RTA and RTB, joined by a disulfide bridge.

We performed a study to assess the interactions between the ricin toxin A (RTA) subunit of ricin and some of its inhibitors using modern Semiempirical Quantum Chemistry and ONIOM Quantum Mechanics/Molecular Mechanics (QM/MM) methods. Two approaches were followed: (calculation of the binding enthalpies, ΔH_{bind}, and reactivity quantum chemical descriptors, QCMDs). These approaches were compared with the respective half-maximal inhibitory concentration (IC_{50}) experimental data, in order to gain insight into RTA inhibitors and verify which quantum chemical method would better describe RTA-ligand interactions. We found that single-point energy calculations of ΔH_{bind}, with the PM6-DH+, PM6-D3H4, and PM7 semiempirical methods; as well as the ΔE_{bind}, obtained via the ONIOM QM/MM method, presented a good correlation with the IC_{50} data. We observed, however, that the correlation decreased significantly when we calculated ΔH_{bind} after optimizing the *full atom* geometry with all semi-empirical methods after full-atom geometry optimization with all semiempirical methods. Based on the results from reactivity descriptors calculations for the cases studied, we noted that both types of interactions, molecular overlap and electrostatic interactions, play significant roles in the overall affinity of these ligands for the RTA binding pocket.

In the second study, we evaluated aspects of the electronic structure in the folding process of NTL9, BBA and α3D proteins. In this work we evaluated three fast folding proteins (NTL9, BBA and α3D) through DM, MSMs and global and local reactivity descriptors obtained through DFT-D3 calculations and PM7 semi-empirical method.

The results showed that the MSMs were able to characterize in details several folding paths, providing information that made it possible to infer the type of mechanism of the studied systems. Data from quantum calculations indicated that the enthalpy of formation is

a good reaction coordinate for the folding process. In addition, our results demonstrate that the integration of MD, DFT-D3, semiempirical method and quantum reactivity descriptors offers the potential to reveal key aspects in protein folding process that are not described by any other approach. It was observed that local hardness per residue is able to distinguish non-native from native-like structures, revealing that intrinsic aspects of electronic structure play a highly relevant role in protein folding process. This observation may be a signature on the electronic structure during protein folding.

Keywords: Ricin; QM/MM; Protein Folding; MSMs; Global and Local Reactivity Quantum Chemical Descriptors.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

CAPÍTULO 1

INTRODUÇÃO

Nessa tese, buscamos encontrar assinaturas da densidade eletrônica em dois problemas biológicos bastante estudados na literatura: 1) interação proteína-ligante e enovelamento de proteínas. Ao longo da tese, demonstramos que descritores quânticos de reatividade baseados na densidade eletrônica como: funções de Fukui, dureza e moleza química revelam aspectos importantes nas interações proteína-ligante e no processo do enovelamento de proteínas. Em relação à interação proteína-ligante, realizamos estudo de candidatos a inibidores da subunidade RTA da ricina, já no segundo estudo, estudamos o processo do enovelamento das proteínas NTL9, BBA e α3D.

1.1 Ricina e abordagens computacionais aplicadas ao estudo de interações proteína-ligante

A ricina é uma proteína citotóxica produzida na semente da mamona (*Ricinus communis*), e pertence à família de proteínas inativadoras de ribossomo (tipo 2) e é uma das toxinas biológicas mais potentes conhecidas.[1] Essa proteína consiste em duas subunidades unidas por uma ligação dissulfeto, a saber, toxina de ricina A (RTA) e toxina de ricina B (RTB).[2–4] A RTA (267 resíduos) é uma N-glicosidase que inativa ribossomos eucarióticos através da depurinação de uma adenina específica localizada no motivo de alça sarcina-ricina (SRL) da subunidade 28S rRNA.[5] A RTB (262 resíduos) é uma lectina que medeia a captação de holoricina nas células por meio do reconhecimento de galactose e N-6-acetilgalactosamina.[6] Uma vez internalizada, a holoricina sofre transporte retrógrado vesicular até atingir o lúmen do retículo endoplasmático;[7] então, uma isomerase reduz a ligação dissulfeto entre as subunidades e a RTA é

translocada para o citosol e, subsequentemente, ataca eficazmente os ribossomos.[8,9]

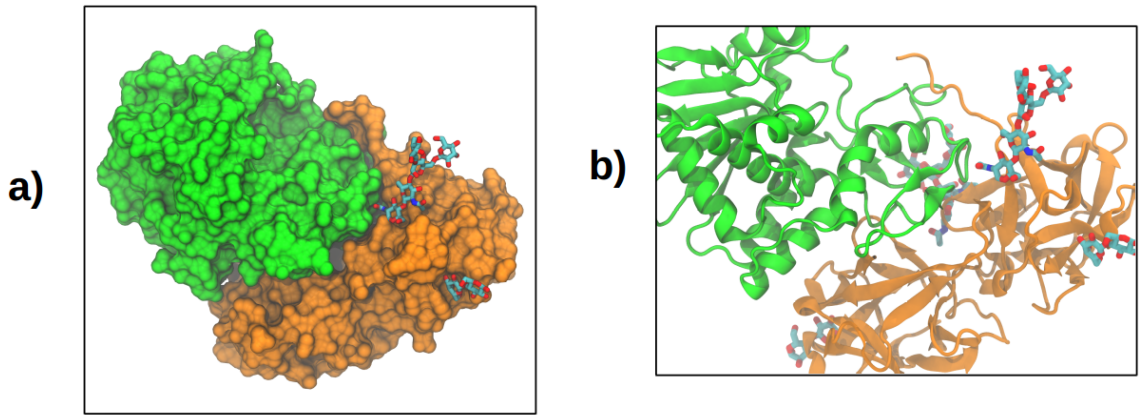


Figura 1.1: Estrutura da ricina. Em verde está a RTA e em alaranjado a RTB. Em (a) destacamos as glicosilações e em (b) a estrutura secundária da ricina.

Há uma preocupação global por parte das autoridades mundiais em relação à toxicidade da ricina devido ao seu potencial uso como arma química, principalmente por grupos terroristas e ativistas.[10] Além disso, a ausência de contramedidas ao envenenamento por ricina contribui ainda mais para essa preocupação. Dessa forma, métodos teóricos, como o *docking* molecular, têm guiado grupos de pesquisa na busca por candidatos a inibidores para o sítio ativo do RTA. No entanto, essa busca não tem sido uma tarefa fácil, dado que o sítio ativo do RTA e seus arredores são amplamente polares, impondo restrições de polaridade que não são contabilizadas por métodos teóricos clássicos.[11]

Até o momento, os métodos de química teórica e computacional têm desempenhado um papel fundamental no estudo de sistemas biológicos e/ou bioquímicos, fornecendo uma orientação adequada para a descoberta de novos fármacos.[12] O conhecimento das interações intermoleculares dos sistemas proteína-ligante é uma característica importante para o desenvolvimento de novos medicamentos.[13] Dessa forma, algumas abordagens, por exemplo, *docking molecular*, têm sido usadas para prever a pose bioativa de ligantes nos sítios ativos de alvos biológicos com interesse terapêutico;[10] no entanto, é bastante ineficaz prever a afinidade de ligação relativa e classificar os ligantes de acordo com dados experimentais.[12,14,15]

Atualmente, a estimativa da energia livre de ligação ($\Delta G_{lig.}$) e entalpia de ligação ($\Delta H_{lig.}$) de um ligante em um complexo proteína-ligante consiste em um grande desafio no campo da química computacional.[12] Para resolver este problema, vários grupos têm usado métodos de química computacional para prever melhores perfis de energias de interações.[12,13,16–34] Entre essas abordagens estão aquelas inteiramente baseadas em campos de força clássicos, como MM-G(P)BSA,[35,36] LIE,[37] SMD[38,39] e FEP,[40] bem como suas versões envolvendo potenciais

híbridos: mecânica quântica/mecânica molecular (QM/MM) ou aqueles totalmente calculados por mecânica quântica (QM), tais como: QM-MM-G(P)BSA,[41,42] QM-MM-LIE,[24,43,44] QM-FEP[26,45] and QM-SMD.[46,47] Todas essas abordagens requerem flexibilidade do complexo proteína-ligante resultando em altos custos computacionais. Portanto, o uso de tais abordagens é limitado a um pequeno conjunto de ligantes e complexos de pequeno e médio porte.

Uma alternativa que requer menor custo computacional é o uso de métodos teóricos para uma única estrutura, em que a flexibilidade do receptor (R), ligante (L) e complexo receptor-ligante (R-L) pode ser parcialmente explicada pela realização de otimização de geometria para R, L e R – L. Nesse cenário, se o interesse é aumentar a precisão das previsões da energia de ligação do complexo receptor-ligante, o uso de métodos QM é obrigatório. Existem duas abordagens para calcular a energia de ligação de um ligante em um alvo biológico usando métodos QM. A primeira abordagem considera uma seleção dos átomos do complexo R – L (QM-*cluster*), e a segunda considera todos os átomos do complexo R – L no cálculo QM.

Na estratégia de QM-*cluster*, um conjunto representativo de resíduos (variando entre 20 e 200 átomos) é selecionado e cortado do local ativo. Este conjunto é estudado separadamente pela aplicação de métodos QM no vácuo ou em um modelo de solvente contínuo.[48,49] A vantagem dessa estratégia é que poucos átomos são considerados no cálculo, permitindo a avaliação de um grande conjunto de ligantes para um mesmo sítio ativo. No entanto, a desvantagem desta abordagem é que a remoção de resíduos importantes durante a seleção da região QM pode levar a resultados não acurados.[50,51] Na estratégia que considera todos os átomos R-L para cálculos QM, é possível acelerar o cálculo das propriedades termoquímicas e permitir a exploração de grandes bancos de dados de ligantes. Isso é possível através do uso de técnicas de escalonamento linear juntamente unidades de processamento gráfico (GPU).[52,53]

Na literatura, alguns estudos realizaram modelagem e simulação molecular envolvendo ricina. Olson, M.A.[54] aplicou métodos de dinâmica molecular (DM) para analisar os aspectos estruturais e energéticos de três ligantes polinucleotídicos (análogos de substrato de rRNA) que se ligam ao sítio ativo do RTA. As simulações dos três complexos R-L, bem como sua comparação com dados experimentais, permitiram um melhor entendimento da interação do RTA com o rRNA.

Em outro estudo, Olson, M.A. e Cuff, L.[55] expandiram o estudo anterior[54] analisando os determinantes de energia livre para a formação de complexos RTA com o substrato rRNA e vários pequenos ligantes. Os autores observaram que as energias livres absolutas de formação, obtidas para o complexo RTA-RNA, bem como para vários mutantes de proteínas, apresentaram boa concordância com os dados experimentais. Os componentes individuais

(por resíduo de aminoácido) da energia livre de ligação do complexo RTA-RNA revelaram interações eletrostáticas altamente relevantes, decorrentes da complementaridade carga-carga das argininas interfaciais com a estrutura de fosfato de RNA. Além do mais, foi observado que a complementaridade hidrofóbica é dominada pelas interações de base da estrutura do *tetraloop* GAGA.

Yan e coautores[56] conduziram estudos sobre as interações de pequenos anéis com o sítio ativo do RTA, a fim de entender melhor como a ricina reconhece os anéis de adenina. Os cálculos das geometrias e energias de interação foram realizados usando métodos MM para alguns complexos entre o sítio ativo da RTA e modificações tautoméricas de adenina, formicina, guanina e pterina. Os resultados indicaram que as energias de interação entre o anel de pterina e RTA são mais fortes do que as da formicina com a RTA. Também foi descoberto que a formicina se liga mais fortemente à RTA do que a adenina. Essas informações apresentaram boa concordância com os dados experimentais. Além disso, os resultados do trabalho experimental e de modelagem molecular sugerem que o sítio de ligação da ricina é bastante rígido e pode reconhecer apenas uma pequena gama de anéis semelhantes à adenina.

Em um estudo recente, Chaves e colaboradores[10] realizaram o *re-docking* de seis inibidores da RTA conhecidos e, em seguida, realizaram simulações de dinâmica molecular dirigida (SMD) para avaliar a afinidade de ligação relativa destes ligantes. A abordagem de *docking molecular* foi capaz de prever a poste bioativa dos ligantes, no entanto, a função de pontuação não foi capaz de classificar esses inibidores de acordo com os dados experimentais. A simulação SMD foi usada para desacoplar esses ligantes da bolsa de ligação, e os perfis de força foram estimados e apresentaram uma forte correlação com os dados experimentais ($R^2 > 0.9$).

Existem poucos estudos de modelagem molecular ou simulação sobre a ricina e, em nossas revisões bibliográficas, não encontramos nenhum estudo aplicando métodos químicos quânticos para complexos RTA-ligante inteiros. Assim, no nosso primeiro estudo, realizamos cálculos do ΔH_{bind} e de descritores quânticos entre a RTA e alguns de seus inibidores usando métodos quânticos semi-empíricos e QM/MM ONIOM para comparar os resultados com dados experimentais (metade da concentração inibitória máxima - IC_{50}) e obter informações de interação local entre os resíduos da RTA e inibidores de RTA. Na próxima seção apresentamos uma introdução sobre o nosso segundo estudo, enolvendo o problema do enovelamento de proteínas.

1.2 Enovelamento de proteínas

Um dos grandes desafios dos cientistas atualmente consiste em compreender os mecanismos que regem os sistemas biomoleculares.[57] Nesse contexto, as proteínas são componentes fundamentais de todos os organismos vivos, pois desempenham diversas funções, tais como o processo de construção da célula, o controle de entrada e saída de íons, atividades de transporte e a sinalização celular.[57] O processo pelo qual cada sequência de proteínas assume sua estrutura única (tridimensional), a partir da síntese da estrutura primária no ribossomo, recebe o nome de enovelamento, estágio final que traduz a informação genética em forma de uma proteína funcional. Trata-se de um processo complexo gerado em temperatura e pressão constantes dentro da célula.[58] O enovelamento anômalo das proteínas, bem como as anormalidades nos dispositivos de controle das células estão diretamente relacionados a diversas doenças, a exemplo do mal de Alzheimer, da encefalopatia espongiforme, da diabetes tipo II e da doença de Huntington.[57]

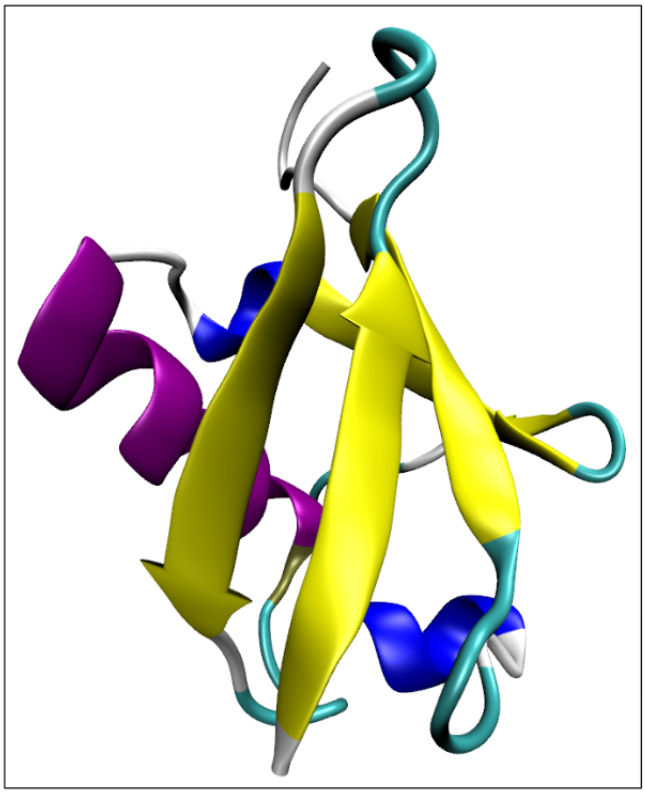
As proteínas tratam-se de macromoléculas em que as suas unidades monoméricas são resíduos de aminoácidos. Elas diferenciam-se por sua sequência de aminoácidos e por sua estrutura tridimensional, sendo essa última o determinante das funcionalidades da célula.[59]

As proteínas podem ser associadas a polímeros lineares de aminoácidos; entre as quais as proteínas naturais são constituídas pelo agregado de cerca de 20 aminoácidos. Por sua vez, os aminoácidos formam-se por um átomo de carbono tetraédrico na cadeia central, chamado de Cα, que apresenta ligações covalentes com um grupo amino, um grupo ácido carboxílico, uma cadeia lateral e um átomo de hidrogênio. Eles se diferenciam em virtude da cadeia lateral que pode apresentar variações no tipo de grupo funcional, na forma e no tamanho; e ligam-se aos seus vizinhos por uma ligação peptídica. A combinação dos diversos aminoácidos forma as distintas proteínas.[60,61]

As proteínas são capazes de assumir quatro níveis estruturais distintos. A estrutura primária consiste em uma sequência linear de aminoácidos. Já a estrutura secundária diz respeito à organização local da proteína e é constituída principalmente por hélice α e folhas β. Na hélice α, a estrutura se contorce assumindo a forma de um bastão condensado; em seu interior ocorre a interação do tipo ligação de hidrogênio entre o grupo CO de cada resíduo com o grupo NH, situado quatro resíduos à frente na sequência de aminoácidos. As folhas β constituem-se por duas ou mais cadeias polipeptídicas, denominadas fitas β. Ao contrário das hélices α, as fitas β apresentam estruturas praticamente estendidas, nas quais a junção entre duas ou mais fitas β ocorre através de ligações de hidrogênio entre os grupos CO e NH de cada fita. A estrutura terciária define-se como um arranjo tridimensional da proteína. Por último, a estrutura qua-

ternária é constituída pela união de duas ou mais estruturas terciárias.[60,61] Na Figura 1.2, a seguir, é apresentada a estrutura primária, secundária e terciária da ubiquitina.

Figura 1.2: Estrutura terciária da ubiquitina: em lilás, temos a hélice α em formato espiral; em amarelo, as folhas β em formato de setas; em azul escuro, os loops e em azul claro/cinza, estruturas sem forma definida. Figura retirada do PDB (1UBQ).



A busca de uma teoria ou modelo apropriado para os processos de enovelamento é de grande interesse, devido à possibilidade que eles fornecem de prever a estrutura *in vivo* de uma sequência de aminoácidos. Uma compreensão mais ampla desse processo resultaria em avanço significativo para as áreas de Biotecnologia e de Ciências da Saúde, por meio da utilização de grandes quantidades de dados de sequenciamentos, com o fim de prever estruturas de várias proteínas e estudar suas funcionalidades sem altos custos experimentais.

O enovelamento de proteínas tem sido tema de pesquisa há mais de cinco décadas, entretanto, sempre houve dificuldades de estudo desse processo na abordagem experimental, bem como da construção de um modelo que englobe todas as suas características.

Anfinsen *et al.* apresentaram contribuições importantíssimas para o entendimento de como as proteínas se enovelam. Em seus experimentos com a enzima ribonuclease pancreática bovina, observaram que essa proteína, após desnaturada, poderia voltar espontaneamente ao seu estado nativo, após as suas condições iniciais serem reestabelecidas. Desse modo, concluíram que toda a informação para a proteína assumir a sua conformação tridimensional

está contida unicamente na sequência de aminoácidos.[62]

Outra contribuição importante de Anfinsen *et al.* foi a hipótese de que uma proteína, para assumir sua conformação nativa, deva possuir um mínimo de energia livre de Gibbs. Essa é a chamada "Hipótese Termodinâmica",[63] que levou os pesquisadores a observarem o problema do enovelamento não somente como caráter biológico, mas também como um problema de busca de mínimo de energia livre.[57] Considerando-se que a sequência dos aminoácidos possui toda a informação para a estrutura se enovelar, pesquisadores passaram a abordar o problema por métodos de simulação computacional.

Dado que o enovelamento de proteínas ocorre em uma escala de tempo muito grande, somente com o avanço computacional no estágio atual é que se tornou possível realizar simulações com modelos completamente atomísticos. Além das dificuldades práticas apresentadas no início da exploração dessa área de pesquisa, sempre houve dúvidas sobre os principais efeitos que contribuem energeticamente e os possíveis mecanismos propostos para a investigação. Os primeiros resultados dos estudos termodinâmicos mostraram que a diferença de energia livre era bem pequena entre o estado nativo e o desnaturado, sugerindo a ocorrência de uma grande perda entrópica, o que pode indicar efeitos antagônicos no processo.

Através da abordagem cinética, Cyrus Levinthal concluiu que o tempo necessário para se varrer todo o espaço conformacional, a fim de encontrar a estrutura com menor energia livre (estado nativo), seria maior que o tempo do universo.[63] Como a proteína se enovela na escala de mili a microssegundos, supõe-se que o seu enovelamento segue um processo de busca direcionada, em vez de buscas aleatórias sobre todo o espaço conformacional. Essa observação contraditória ficou conhecida com o "Paradoxo de Levinthal".[64–66]

A perda entrópica se dá em decorrência do grande número de configurações possíveis, ou graus de liberdade, que o estado desnaturado pode apresentar. Esse número é estimado em 10^{30} conformações possíveis, em função do número de aminoácidos, de suas conformações e combinações entre si. Com base no grande número de configurações, observou-se que era necessário um caminho, como o de uma reação química para a proteína se enovelar, pois uma busca aleatória em sua superfície de energia potencial não condiz com o tempo real que a proteína se enovela de fato. Isso também é conhecido como o requisito cinético para o enovelamento, em que um modelo deve reproduzir o tempo correto do processo.

Uma grande perda entrópica acompanhada de grande perda entálpica indica que o processo de enovelamento é semelhante a uma transição de primeira fase, como uma mudança de estado físico. No entanto, o número de conformações possíveis e a natureza das forças presentes indicam que há um caminho de enovelamento no qual a proteína passa por estruturas

intermediárias.[67–70]

Na década de 1980, surgiu uma abordagem alternativa para o processo de enovelamento, propondo que ele seria organizado por um grupo de estruturas semelhantes ao longo da superfície de energia e não por estruturas intermediárias em uma via de enovelamento. Isso abriu caminho para que abordagens estatísticas fossem utilizadas para a descrição dessa superfície no processo em questão.[71,72] Surgiram então modelos teóricos voltados à descrição desse processo, entre os quais podemos destacar o colapso hidrofóbico, em que a força de atração entre os resíduos hidrofóbicos geram uma grande compactação da cadeia principal da proteína, de modo que o enovelamento possa ocorrer em um espaço confinado, reduzindo assim a busca conformacional para o estado nativo. Outro modelo proposto foi o mecanismo de difusão-colisão, em que as estruturas secundárias se enovelam primeiro, seguido da ancoragem dessas estruturas entre si, formando a estrutura terciária até chegar à estrutura nativa. Além dos dois modelos citados, surgiu um novo mecanismo, denominado nucleação-condensação, no qual são formadas interações hidrofóbicas de longo alcance e outras interações nativas no estado de transição, com o fim de estabilizar a formação de estruturas secundárias.[72]

Nesse sentido, os modelos buscavam responder à questão: *quais são os principais requisitos para que uma sequência de aminoácidos se enovele assumindo sua conformação nativa?* Tal problemática foi e continua a ser investigada, principalmente por meio de métodos de simulação computacional.

1.2.1 Modelos Teóricos e Superfície de Energia Livre

A criação de modelos que se adéquem às características dos processos de enovelamento de proteínas faz parte de uma abordagem fenomenológica que postula características da superfície de energia. Baseados nos princípios da mecânica estatística,[58,73–75] esses modelos usaram métodos de simulação simplificados com funções de energia derivados de distribuições aleatórias, para assim explorar as consequências termodinâmicas e cinéticas do enovelamento.[57,59] As informações cedidas por esses modelos iniciais permitiram a formulação da Teoria da Superfície de Energia Livre (*Free Energy Landscape*),[76–80] a qual abrangia todos os tipos de mecanismos gerais propostos na época, tanto um caminho de estruturas discretas, quanto conjuntos de configurações como intermediários, em consonância com os requisitos cinéticos e termodinâmicos para o processo de enovelamento.[58,81] Através da Teoria da Superfície de Energia Livre, muito tempo de simulação pode ser salvo, graças a métodos de simulação baseados em suas premissas, como as simulações de reenovelamento e as de desenovelamento.[71,82]

Uma contribuição que apresenta fundamentos para compreender o processo de enovela-

mento de proteínas diz respeitos aos vidros de spins, grupos bastante estudados na mecânica estatística.[83] Os vidros de spins são sistemas de átomos com distribuições aleatórias de momento magnético que apresentam um elevado grau de interações conflitantes, denominado frustrações. Os heteropolímeros aleatórios e os sistemas magnéticos, por sua vez, são modelos de vidros de spins que apresentam diversos estados metaestáveis de mínimo local com uma superfície rugosa de energia, assemelhando-se ao comportamento estrutural de um vidro. Esses sistemas de átomos apresentam uma temperatura de transição (Tg), relacionada ao caráter rugoso da superfície de energia, sendo que, para temperaturas menores que Tg, o sistema fica preso em regiões de mínimo local.[57] A formulação mecânica estatística de materiais vítreos para o enovelamento de proteínas ocorreu devido à semelhança do comportamento termodinâmico e cinético, revelado por estudos físico-químicos de estruturas nativas e experimentos de reenovelamento in vitro.[71,82,83]

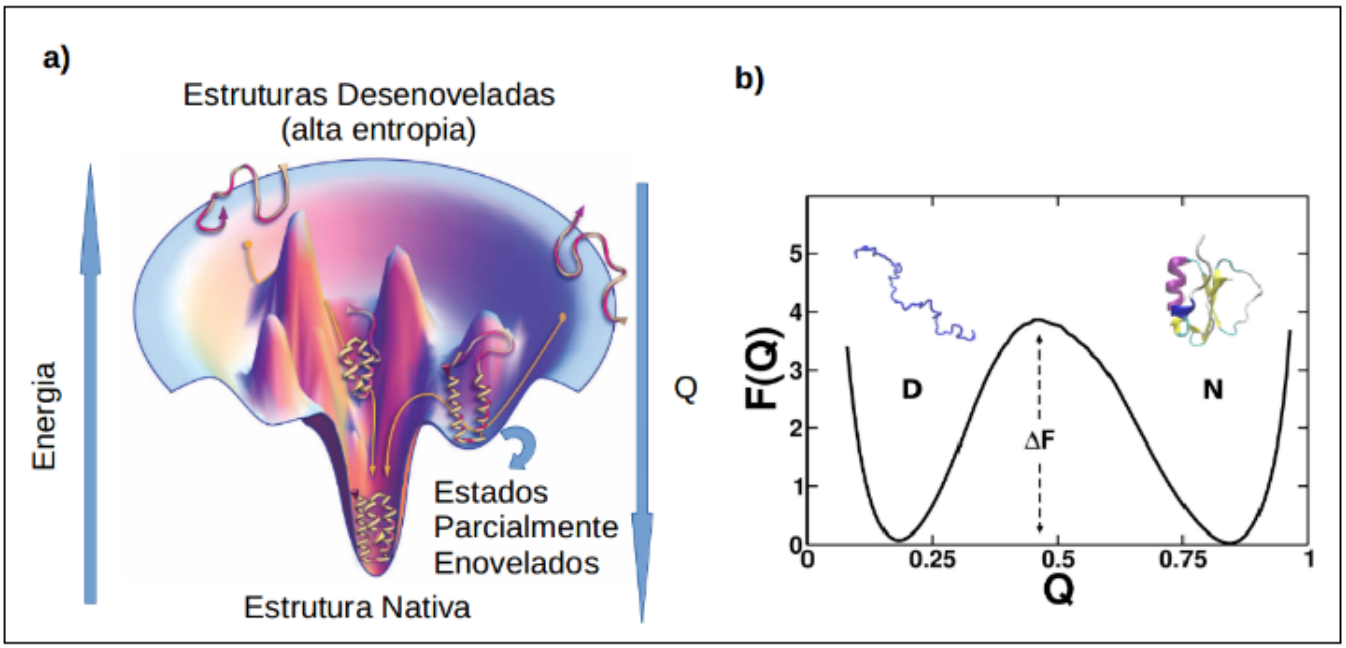
As proteínas também possuem rugosidade na superfície, porém, diferentemente dos heteropolímeros aleatórios, elas apresentam sequências e conformações nativas que, durante a evolução biológica, foram selecionadas seguindo o princípio de mínimas frustrações.[57,83] Segundo esse princípio, as proteínas evoluíram para um estado de frustração mínima, ou seja, a estrutura nativa representa o estado de menor energia, conforme apontado pela hipótese de Anfinsen.[57]

Devido ao número de coordenadas atômicas que uma proteína pode possuir (reflexo do elevado número de átomos em sua estrutura molecular), faz-se necessário o uso de um parâmetro que represente bem o avanço do processo de enovelamento na superfície de energia potencial como uma coordenada de reação. Um parâmetro bastante utilizado é a fração de contatos nativos (Q), que resulta da razão entre número de contatos nativos formados por uma determinada conformação em relação aos contatos presentes na estrutura nativa.[84] Desse modo, a fração de contatos nativos assume valores entre "zero" e "um", sendo que o valor igual a "zero" indica que a estrutura está totalmente desnaturada e o valor igual a "um" indica que a estrutura está na conformação nativa.[57]

Estudos de Best, R. B. *el al.*, apontam que Q foi capaz de detectar o mecanismo de enovelamento para maioria das proteínas pequenas, demonstrando estar de acordo com a Teoria da Superfície de Energia Livre.[84] Logo, a estrutura nativa será a com menor valor de energia livre na superfície; ao passo que as estruturas correspondentes às configurações possíveis para a estrutura desnaturada serão as de maior energia livre. Nesse sentido, a superfície de energia livre é apresentada como uma funil com um gradiente de energia apontado na direção da estrutura enovelada.[57,85,86] A Figura 1.3, a seguir, representa o funil de enovelamento e a

coordenada de reação para um sistema de enovelamento de proteínas.

Figura 1.3: Temos: a) funil de enovelamento representando o processo de enovelamento de proteínas em função da energia e da fração de contatos nativos (Q). b) Representação da energia livre F(Q) em função da coordenada de reação (Q). O estado desenovelado é representado por (D); o estado nativo por (N) e a barreira energética do estado de transição por (ΔF). A Figura a) foi adaptada da referência [86] e a Figura b) foi retirada da referência [57].



Na parte superior do funil, temos as estruturas desenoveladas com alta entropia e alto valor energético. À medida que a estrutura vai descendo no funil, ela começa a perder entropia, assumindo conformações parcialmente enoveladas, até atingir a estrutura nativa, que possui o menor valor de energia livre de Gibbs. Também é observado que a fração de contatos nativos aumenta à medida que ela se aproxima da estrutura enovelada no funil. É observado também que não há somente um caminho possível para processo de enovelamento, e sim caminhos que levam ao mesmo estado enovelado.[81,85]

1.2.2 Estrutura eletrônica e enovelamento de proteínas

Atualmente, um dos principais desafios para a compreensão do enovelamento de proteínas consiste no entendimento sobre a sua estrutura eletrônica. Ressaltamos que tanto as propriedades estruturais quanto dinâmicas das proteínas são extremamente importantes para o entendimento de como ocorrem as mudanças conformacionais dessa durante o enovelamento, bem como a predição da sua função biológica.[87] Como a estrutura eletrônica está

fundamentalmente correlacionada com a sua representação atômica, uma análise mais aprofundada da estrutura eletrônica durante o enovelamento pode auxiliar no desenvolvimento de métodos de previsão da atividade biológica da proteína.[87] Contudo, o entendimento do real papel desempenhado pela estrutura eletrônica no processo do enovelamento de proteínas se configura como um desafio teórico-computacional muito difícil de ser superado.

Nesse aspecto, uma estratégia válida é a de avaliar quantitativamente o papel das interações intermoleculares em sistemas biológicos usando a densidade eletrônica molecular para se estudar o problema do enovelamento de proteínas. E nesse aspecto, já existem algumas ferramentas em que isso pode ser feito, como é o caso do software NCIPLOT.[88]

Recentemente, Boto, R.A. e colaboradores[89] apresentaram uma nova versão, o NCIPLOT4. Nesse trabalho é proposto a noncovalent interaction density integral, uma quantidade proposta para obter uma imagem semiquantitativa de interações não covalentes em sistemas complexos com uma resolução temporal. Nesse trabalho, os autores estudaram a dinâmica de formação da estrutura secundária (hélice-α) da *mature amyloid fibril* (2M4J)[90] através de dinâmica molecular clássica de 200ns. As protodensidades moleculares foram calculadas ao longo da trajetória e a integral que quantifica as interações intermoleculares NCI foram correlacionados com a energia de estabilização calculada com o campo de força AMBER. Os autores mostraram que quando a estrutura secundária foi formada é acompanhada por um aumento significativo da densidade de interação não covalente integral. Exemplos adicionais do uso dessa quantidade para outros sistemas moleculares podem ser conferidos no trabalho de Peccati, F..[91]

Outros pesquisadores apresentaram estudos na tentativa de avançar na compreensão da estrutura eletrônica no processo do enovelamento de proteínas.

Dwyer, D. S., calculou as propriedades eletrônicas das cadeias laterais de um banco de dados com 118 proteínas com o intuito de avaliar as preferências dos aminoácidos para a formação de hélice-α, fita-β ou *coil*.[92] Para quantificar os efeitos eletrônicos dos aminoácidos no enovelamento de proteínas, foram utilizados valores de pKa no grupo amino, constantes substituintes de efeito localizado (e_σ) e uma métrica composta por essas duas escalas (ε).[92] Ele observou, de modo geral, que a propensão de um aminoácido para uma determinada conformação estava relacionada às características eletrônicas da sua cadeia lateral.[92]

Em um segundo trabalho, Dwyer, D. S. avaliou as propriedades eletrônicas das cadeias laterias dos aminoácidos através dos métodos semi-empíricos PM3, AM1 e MNDO.[93] Nesse trabalho, a análise da distribuição de elétrons e efeitos de polarizabilidade contabilizados com método PM3 foram usados para derivar escalas quantitativas que descrevem fatores estéricos bem como efeitos indutivos, de ressonância e de campo e fatores de polarizabilidade

nas cadeias laterais dos aminoácidos.[93] Foi observado que os três métodos semi-empíricos apresentaram resultados semelhantes quanto aos efeitos eletrônicos nas cadeias laterais. Uma análise de regressão revelou que os valores da população de Mulliken no Cα e os efeitos indutivos apresentaram maior propensão à formação de hélice-α. O autor concluiu que os dados fornecem uma caracterização inicial dos efeitos dos substituintes das cadeias laterais de aminoácidos e sugerem que essas propriedades afetam a densidade eletrônica ao longo do *backbone* do peptídeo.[93]

Um preditor de reatividade química conhecido pelos seus ótimos resultados é o conceito de ácidos e bases duros e macios (HSAB).[94] Nesse sentido, Faver, J. *et al.*[94] avaliaram esse conceito via função de Fukui (*um descritor de reatividade da DFT*) para três problemas biológicos conhecidos: ancoragem do ligante, detecção de sítio ativo e enovelamento de proteínas. Em relação ao enovelamento de proteínas, os autores sugeriram a hipótese de que as proteínas mais bem enoveladas deveriam apresentar interações duras e moles mais favoráveis entre os resíduos do que as mal enoveladas. Para testar essa hipótese, foi introduzida uma função pontuação dependente da distância que compara os índices de Fukui entre os distintos aminoácidos. Eles observaram que apesar da estrutura nativa apresentar uma das melhores pontuações da função de Fukui, essa não foi capaz de discriminar a nativa do conjunto de *decoys*. Os autores inferiram que as interações de dureza e moleza parecem não incluir contribuições eletrostáticas que são importantes no ranqueamento de *decoys* de proteínas.[94]

Urquiza-Carvalho, G. A. e colaboradores[95] apresentaram uma abordagem diferente da citada por Faver, J.[94] em relação ao ranqueamento de um conjunto de *decoys* de proteínas. Eles utilizaram as entalpias de formação obtidas com os métodos PM6, PM6-DH+, PM6-D3 e PM7 em solução aquosa, como função de pontuação para discriminar estruturas nativas ou bons modelos em 33 conjuntos de *decoys* de proteínas. Foi observado que o método PM7 apresentou melhores resultados, sendo capaz de identificar a estrutura nativa entre todos os conjuntos de *decoys*. Além disso o PM7 descreveu melhor estatisticamente as características esperadas para um funil de enovelamento, simulando o poço em que a hipótese termodinâmica prevê que a conformação nativa esteja na superfície de energia livre.[95]

No trabalho de Momen, R. e colaboradores[96] é apresentada uma nova interpretação via QTAIM do mapa de Ramachandran. O mapa de Ramachandran é um gráfico em que são mostrados os ângulos torcionais mais importantes da cadeia polipeptídica de uma proteína (ϕ e ψ) e, que representam a conformação de uma proteína. Ou seja, olhar para o mapa de Ramachandran é estar olhando diretamente para a conformação da proteína naquele instante. E, como em muitas situações se define o problema do enovelamento das proteínas como sendo

um problema de busca conformacional da cadeia polipeptídica, a redefinição do mapa de Ramachandran para incorporar efeitos da densidade eletrônica trata-se de uma metodologia inovadora. A nova interpretação do mapa de Ramachandran foi na tentativa de gerar novas coordenadas de enovelamento baseadas em descritores QTAIM. Um foco especial foi dado para o papel que ligações de hidrogênio possuem quando certos tipos de aminoácidos estão presentes na estrutura.

Ianeselli e colaboradores[97] recentemente publicaram um estudo que tentou demonstrar que simulações de dinâmica molecular de caminhos de enovelamento combinadas com avaliação quântica do espectro de dicroísmo circular consegue complementar a técnica de dicroísmo circular aplicadas no processo de enovelamento, pois fornece informações dependentes do tempo que são impossíveis de serem capturadas experimentalmente. Isto foi exemplificado para o enovelamento da proteína da lisozima do leite canino com excelentes resultados.

Recentemente M. Culka e colaborador publicaram três interessantes estudos em que abordam o problema do enovelamento de proteína através do uso combinado de métodos de MD e QM.[98–100]

No primeiro artigo,[98] os autores avaliaram a propensão de estrutura secundária inerente de peptídeos curtos através da junção da perspectiva da bioinformática e cálculos de química quântica. Os autores realizaram a identificação das sequências de tripeptídeos (pró-helicoidal e pró-estendido) em proteínas que apresentam preferência pela formação de estrutura helicoidal ou estendida. Essas trincas de aminoácidos foram convertidas em tripeptídeos isolados, sendo submetidos a uma extensa amostragem conformacional e otimização de geometria, sendo que o valor da energia livre foi obtida via DFT-D3/COSMO-RS. Eles observaram que a maioria dos tripeptídeos isolados de baixa energia apresentaram a tendência de formação de estruturas helicoidais, independente de serem pró-helicoidais ou pró-estendidos na estrutura protéica. No entanto, foi observado que os tripeptídeos pró-helicoidais apresentam uma preferência helicoidal um pouco mais acentuada. Além disso, se a simulação for realizada com um solvente hidrofóbico, imitando as partes menos polares de uma proteína, os tripeptídeos pró-estendidos são favorecidos. Os autores sugeriram, a partir da análise dos cálculos realizados, que proteínas complexas podem começar a emergir a partir do nível de pequenas unidades oligopeptídicas, o que está de acordo com a natureza hierárquica do enovelamento de proteínas.[98]

No segundo trabalho,[99] os autores usaram o mesmo protocolo para avaliar a propensão da cadeia-beta de diferentes tipos de enovelamentos do domínio WW encontrados no banco de dados *SCOPe 2.0719*. Eles mostraram que a energia de deformação interna é maior nas folhas beta e menor nos *loops*, enquanto a energia de interação tem uma tendência oposta.

Com base nos seus resultados os autores chegam a interessante conclusão que a energia de interação interna é a quantidade física ajustada pela evolução para definir o enovelamento de proteínas da folha-β.

No terceiro trabalho dessa série,[100] mais uma vez os autores usaram o protocolo de cálculos desenvolvidos anteriormente[98] para saber quais interações são a chave para a formação e evolução da estrutura proteica: as de longo alcance ou os efeitos de curto alcance. Essas quantidades foram correlacionadas com a conservação de resíduos de aminoácidos em elementos da estrutura secundária e também com o grau de internalização do resíduo na estrutura tridimensional da proteína. Mesmo afirmando textualmente que *"Entende-se, e também foi mostrado neste trabalho, que o problema do enovelamento de proteínas é muito complexo para ser descrito por quantidades simples, como a interação e as energias de deformação apresentadas aqui."* (tradução livre), os autores apontam que seus resultados podem representar uma contribuição importante para a compreensão do enovelamento de proteínas a partir de métodos de primeiros princípios (*ab initio*).

Descritores moleculares são usados para predizer propriedades de moléculas e materiais, classificar estruturas químicas ou procurar similaridades entre elas[101].

Em princípio, qualquer quantidade calculada com métodos de química quântica, sejam propriedades físicas observáveis ou não, podem ser classificadas como sendo um *Quantum-Chemical Molecular Descriptor* (QCMD).[102] A lista é muito ampla pois além de contar com as quantidades moleculares em si, é possível propor transformações ou composições dessas para gerar novos descritores. Assim, alguns desses podem ser: cargas atômicas, quantidades termoquímicas, energias eletrônicas, energias dos orbitais moleculares, momento dipolar, potencial de ionização, localização de orbitais moleculares de fronteira, densidade eletrônica em átomos, etc.

Atualmente os QCMDs baseados na densidade eletrônica têm despertado o interesse de vários pesquisadores em estudos de QSAR/QSPR,[103,104] possibilitando a modelagem de propriedades biofísicas e biológicas no ambiente da proteína. A razão apontada para isso vem do fato de que a densidade eletrônica é fundamentalmente relacionada ao Hamiltoniano molecular e, portanto, é a fonte de todas as propriedades moleculares, seja no estado fundamental quanto nos estados excitados.[105] Outro ponto importante é que a densidade eletrônica tanto pode ser calculada quanto obtida experimentalmente por diversas técnicas.

Todavia, diante do poder computacional que dispomos atualmente e aliado ao uso de métodos e algoritmos eficientes que passaram a permitir a modelagem de sistemas moleculares com muitos átomos, o desafio que é posto agora é o de calcular QCMDs baseados na densidade

eletrônica para biomoléculas, onde é muito comum encontrar estruturas que ultrapassam milhares de átomos.

Nesse sentido, Grillo, I. B. e colaboradores desenvolveram o PRIMoRDiA,[106] (https://github.com/igorChem/PRIMoRDiA1.0v.), uma ferramenta de código aberto destinada a calcular uma abundância de QCMDs para grandes sistemas.

Usando o PRIMoRDiA, Grillo, I. B. e colaboradores realizaram um estudo em que foram calculadas funções de Fukui (um tipo de descritor quântico baseado na densidade eletrônica) para oito pequenos polipeptídeos usando cinco métodos quântico-químicos semiempíricos modernos, considerando orbitais moleculares localizados e os canônicos, bem como dois métodos de aproximação para o cálculo de derivadas usadas no cálculo das funções de Fukui.[107] Após comparar com os resultados de um método de referência escolhido (B3LYP), foi revelado que a combinação de métodos semiempíricos com a Aproximação de Orbitais Congelados permite o cálculo desses descritores de reatividade para sistemas biológicos com precisão e velocidade razoáveis, permitindo seu uso para esses sistemas biomoleculares. Também foi mostrado o potencial desse protocolo em aplicá-lo aos complexos ligantes-proteína da protease do HIV-1, prevendo quais ligantes são inibidores ativos.[107]

Em outro trabalho,[108] Grillo, I. B. e colaboradores propuseram o uso de descritores quânticos de reatividade para caracterizar a catálise enzimática, delineando o seu perfil de reação via métodos computacionais de baixo nível, como Hamiltonianos semiempíricos. Foram simulados três caminhos de reações enzimáticas e calculados os descritores de reatividade para todas as estruturas obtidas. Os resultados desse trabalho sugerem que a análise da estrutura eletrônica pode permitir a proposição e/ou previsão de novos mecanismos, fornecendo assim caracterização química das regiões ativas da enzima, acelerando o processo de transformação das estruturas tridimensionais da proteína resolvida em informações catalíticas.

Neste trabalho, calculamos diversos QCMDs para estruturas ao longo de trajetória de enovelamento das proteínas BBA[109](1FME), NTL9[110](2HBA) e α3D[111](2A3D) estudadas por Lindorff-Larsen, K. e colaboradores,[112] o que vai permitir avaliar aspectos importantes da estrutura eletrônica durante o processo de enovelamento de proteínas.

CAPÍTULO 2

OBJETIVOS

2.1 Objetivo Geral

O objetivo geral desse trabalho consiste em avaliarmos o papel da estrutura eletrônica de sistemas biológicos (interação proteína-ligante e enovelamento de proteínas) através de cálculos termoquímicos, métodos QM e híbridos QM/MM modernos e de descritores quânticos de reatividade *(Quantum-Chemical Molecular Descriptor* - QCMDs).

2.2 Objetivos Específicos

i. Realização de cálculos termoquímicos e de QCMDs pra obtenção de informações sobre candidatos a inibidores da toxina A da ricina (RTA).

i. Realizar o tratamento dos dados de dinâmica molecular do enovelamento de proteínas de forma estatisticamente adequada.

ii. Realizar estudos cinéticos e termodinâmicos do processo de enovelamento de proteínas.

iii. Encontrar caminhos de envovelamento através de modelos de estados de Markov, teoria do estado de transição (TPT) e analise de componentes independentes do tempo (tICA).

iv. Avaliar o papel da estrutura eletrônica durante o caminho de enovelamento através de QCMDs globais e locais.

v. Correlacionar dados obtidos pelos descritores quânticos com descritores estruturais amplamente utilizados.

CAPÍTULO 3

MÉTODOS E FUNDAMENTAÇÃO TEÓRICA

3.1 Formalismo da Dinâmica Molecular Clássica

Nesta seção, são descritas de forma compacta as equações de movimento utilizadas na DM clássica e os algoritmos para integração dessas equações. Informações mais detalhadas podem ser verificadas nas referências.[113,114]

A simulação da DM clássica é um método de modelagem computacional baseado no movimento de cada partícula de um determinado sistema. As partículas são descritas pela equação de movimento de Newton, obtendo-se a posição e a velocidade de cada uma durante cada passo de simulação. A equação de Newton para uma partícula simples é descrita pela equação 3.1:

$$F_i(t) = m_i \alpha_i. \tag{3.1}$$

em que F_i é a força atuante sobre uma dada partícula do sistema em um dado tempo t, m_i e α_i a massa e a aceleração do átomo i, respectivamente.[114]

O potencial de interação entre os átomos do sistema é calculado com base em parâmetros que constituem o campo de força. Esse é uma função da energia potencial total do sistema,V(r),composta por termos atribuídos aos átomos ligados (estiramento de ligação, deformação angular e de torção) e termos atribuídos aos átomos não ligados (interações de van der Waals e de Coulomb).[113] Um campo de força típico pode ser descrito pela equação 3.2, a seguir:

$$V(r) = \sum V_{lig.}(r) + \sum V_{angular}(\theta) + \sum V_{diedro}(\phi) + \sum V_{VdW} + \sum V_{elet}. \tag{3.2}$$

Fazendo-se uso das aproximações dos potenciais harmônicos, os termos referentes ao estiramento de ligação e deformação angular são dados, respectivamente, pelas seguintes

equações:

$$\sum V_{lig.}(r) = \sum_{lig.} \frac{1}{2} K_r (r - r_0)^2; \tag{3.3}$$

$$\sum V_{angular}(\theta) = \sum_{angular} \frac{1}{2} K_r (\theta - \theta_0)^2; \tag{3.4}$$

onde r e θ correspondem aos comprimentos e ângulos de ligação em torno das posições de equilíbrio, r_θ e θ_0, respectivamente. Os termos K_r e K_θ são as constantes da força restauradora do sistema.[113,114] Uma forma bem utilizada para representar o potencial de torção é descrito pela equação 3.5:

$$V_{diedro}(\phi) = \frac{V_n}{2}(1 + cos(n\phi - \gamma)). \tag{3.5}$$

Nessa equação V_n, n, ϕ e γ correspondem, em ordem: à barreira energética para a torção, número de máximos ou mínimos de energia da torção, ângulo diedro e ângulo de fase. Em alguns campos de força são também inclusas aproximações harmônicas para definir o potencial torcional impróprio.

As interações intermoleculares, compostas pelos átomos não ligados, são dadas pelos termos de van der Waals e eletrostático,[113,114] os quais são geralmente representados pelos potenciais de Lennard-Jones e de Coulomb, conforme equações 3.6 e 3.7:

$$V(_{vdW}) = 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]; \tag{3.6}$$

$$V(_{elet}) = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}. \tag{3.7}$$

Na equação 3.6, ε_{ij}, σ_{ij} e r_{ij} correspondem, respectivamente, à profundidade do potencial, à largura do poço de potencial e à separação entre os átomos i e j. O primeiro termo corresponde à parte repulsiva e o segundo à atrativa. Na equação 3.7, q_i e q_j correspondem às cargas parciais dos átomos i e j, respectivamente.[114] O termo ε_0 é a permissividade do meio; e r_{ij} possui o mesmo sentido dado na equação 3.6.

Diversos campos de força têm sido desenvolvidos nas últimas décadas, sendo que os mais utilizados para proteínas são o CHARMM, GROMOS, AMBER, OPLS, entre outros.[114]

Tendo-se definido o campo de força, calculam-se as forças atuantes sobre cada átomo através da derivada primeira da energia potencial em relação às posições dos átomos, conforme a equação 3.8, a seguir:

$$F_i(t) = -\frac{\partial V(r_i)}{\partial r_i}. \tag{3.8}$$

Através da equação 3.8, obtém-se a aceleração da partícula, e integrando-se as equações de movimento, obtêm-se as velocidades e as novas posições dos átomos. Com essas novas posições e velocidades, é calculada a energia potencial e cinética do sistema. Com a aplicação sucessiva desse procedimento, teremos um conjunto de posições e velocidades ao longo do tempo, compondo uma trajetória. Existem diversos algoritmos para a integração das equações de movimento, entre os quais é bastante utilizado na DM o algoritmo de Verlet,[113,114] descrito na equação 3.9:

$$r(t + \Delta t) = 2r(t) - r(t + \Delta t) + \alpha(t)\Delta t^2. \tag{3.9}$$

A equação 3.9 é obtida considerando-se que a posição da partícula no tempo t, "r(t)" e a posição da partícula no tempo $t + \Delta t$, "$r(t + \Delta t)$" são obtidas através da expansão em série de Taylor, que é aplicada tanto para frente quanto para trás, conforme as equações 3.10 e 3.11, respectivamente:

$$r(t + \Delta t) = r(t) + v(t)\Delta t + \frac{1}{2}\alpha(t)\Delta t^2 + ... \tag{3.10}$$

$$r(t - \Delta t) = r(t) - v(t)\Delta t + \frac{1}{2}\alpha(t)\Delta t^2 - ... \tag{3.11}$$

A equação 3.9 é obtida pela soma das equações 3.10 e 3.11 e isolamento do $r(t + \Delta t)$.

3.2 Formalismo dos Métodos hibridos QM/MM

Uma alternativa para reduzir o custo computacional é realizar o tratamento quântico apenas na região envolvida no processo em estudo. Nesse caso, considerando que o subsistema de interesse está imerso com uma perturbação, o restante do sistema pode ser tratado de forma clássica. Essa é a ideia dos métodos híbridos QM/MM, caracterizados por combinarem a mecânica quântica (QM) e a mecânica clássica ou molecular (MM). Warshel e Levitt[115] foram os percussores do primeiro cálculo realizado por meio de um potencial híbrido. No trabalho desenvolvido por eles, a reação química na parte interna de uma enzima foi tratada de forma híbrida, sendo que o centro reativo foi tratado por meio de mecânica quântica, ao passo que o restante da enzima e do substrato foram tratados classicamente, através de um campo de força empírico. Além disso, as moléculas de água ao redor do sistema foram tratadas como dipolos

pontuais. Seguindo o modelo de Warshel e Levitt, outros trabalhos surgiram, a exemplo dos de Singh e Kollman[116] e o de Field, Bash e Karplus.[117] Esses trabalhos destacam que, para o tratamento de sistemas muito grandes, como proteínas em solução aquosa, a região quântica pode apresentar diversas subdivisões. À parte mais importante devem ser aplicados cálculos de alto nível; às demais partes devem ser aplicados cálculos com nível mais baixo. Essas medidas são adotadas para reduzir o custo computacional do sistema. Diversos pesquisadores têm se dedicado a desenvolver métodos computacionais que possibilitem tratamento diferenciado da região quântica. Dentre eles, merece nossa atenção para o método ONIOM de Morokuma *et.al*.[118–120] Esse tema de pesquisa rendeu o prêmio Nobel de Química em 2013 para os pesquisadores Martin Karplus, Michael Levitt e Arieh Warshel.[121]

Apesar de existirem atualmente uma enorme variedade de métodos QM/MM com aproximações diferentes, a maioria desses métodos podem ser divididos em duas categorias: convencional e sequencial, as quais explicaremos a seguir.

3.2.1 Método QM/MM convencional

Na abordagem convencional, os tratamentos clássico e quântico ocorrem de forma simultânea. Desse modo, todo o sistema deve ser tratado através de um Hamiltoniano efetivo, conforme a equação 3.12:

$$\hat{H}_{ef} = \hat{H}_{QM} + \hat{H}_{MM} + \hat{H}_{QM/MM}. \tag{3.12}$$

O termo $\hat{H}_{QM/MM}$ pode adotar formas variadas, responsáveis por diferir os diversos métodos QM/MM convencionais, atualmente. A energia total do sistema é descrita da seguinte forma:

$$E_{tot} = E_{QM} + E_{MM} + E_{QM/MM}; \tag{3.13}$$

$$E_{tot} = \langle \Psi | \hat{H}_{QM} | \Psi \rangle + \langle \Psi | \hat{H}_{MM} | \Psi \rangle + \langle \Psi | \hat{H}_{QM/MM} | \Psi \rangle . \tag{3.14}$$

Uma diferença importante entre os métodos QM/MM consiste no modo em que a parte eletrostática é descrita na interação entre as duas regiões ($\hat{H}_{QM/MM}$). Tal interação pode ocorrer de duas formas: acoplamento mecânico ou acoplamento eletrostático. No acoplamento mecânico, a interação entre as regiões se dá por meio da atribuição de cargas parciais aos átomos da região quântica a fim de que eles possam interagir com as cargas parciais dos átomos da região clássica. Já no acoplamento eletrostático, essa interação se dá pela introdução de um potencial gerado pelas cargas parciais da região clássica no Hamiltoniano quântico.[113]

A principal dificuldade entre os métodos QM/MM convencionais, consiste na escolha das regiões que terão tratamento quântico e clássico, dado que, em decorrência do custo computacional, a região quântica é a menor possível. Geralmente, no método convencional, a escolha da divisão entre a parte clássica e quântica se dá de duas maneiras: (1) a superfície que separa as regiões isola moléculas inteiras, sendo realizado o corte do sistema em partes com pouca densidade eletrônica; (2) a superfície corta as ligações químicas deixando a região quântica com orbitais com valência incompleta.[113]

A inclusão de moléculas inteiras na região quântica é a mais simples das metodologias QM/MM porque não necessita cortar ligações químicas, tornando as suas implementações mais simples e também computacionalmente mais eficiente. Além disso, as partes moleculares das duas regiões (QM e MM) interagem essencialmente por meio de interações intermoleculares e como essas interações são mais fracas, podem ser modeladas de maneira mais fácil por potenciais clássicos, com a ressalva de que interações não-covalentes só podem ser modeladas corretamente com o emprego de correlação eletrônica e grandes conjuntos de base.[122]

Nesse sentido, uma aproximação QM/MM com um pequeno acoplamento entre as regiões clássica e quântica pode ser realizada de forma simples, assumindo um determinado nível de cálculo (QM) para o soluto e um campo de força para o solvente (MM).[113] A interação soluto-solvente utilizando o acoplamento mecânico e contendo apenas termos intermoleculares pode ser feito por:

$$\hat{H}_{QM/MM} = \sum_a \sum_b 4\varepsilon_{ab} \left[\left(\frac{\sigma_{ab}}{r_{ab}} \right)^{12} - \left(\frac{\sigma_{ab}}{r_{ab}} \right)^{6} \right] + \sum_a \sum_b \frac{q_a q_b}{r_{ab}}; \tag{3.15}$$

onde “a” corresponde ao sítio do soluto (QM) e “b” corresponde ao sítio do solvente (MM). O parâmetro Lennard-Jones é utilizado para modelar os sítios do soluto, enquanto as cargas parciais do soluto qa são obtidas por meio de tratamento quântico. A implementação dessa técnica foi realizada por Kaminski e Jorgensen,[123] usando o método semi-empírico AM1 para o tratamento quântico. Já para o tratamento clássico, foi utilizado o campo de força OPLS. Para a obtenção das cargas parciais por meio da função de onda AM1, foi utilizado o método CM1,[124] descrito na literatura como AM1/OPLS/CM1 ou AOC. Nesse método, as cargas parciais são colocadas de modo que o momento de dipolo de solutos neutros aumente em 20%. Essa estratégia foi adotada na tentativa de tentar descrever a polarização do soluto pelo meio.

O método AOC apresentou eficiência satisfatória na descrição de efeitos de solventes polares no equilíbrio rotamérico de determinadas moléculas como o 1,2-dicloroetano e o 2-furfural.[113] Temos, porém, que em situações onde os aumentos no momento de dipolo

provocados pela polarização do soluto atingem patamares superiores a 20%, faz-se necessária a utilização de um acoplamento soluto-solvente mais efetivo. Uma maneira de tentar corrigir essa limitação é pela utilização de um acoplamento eletrostático para a descrição da polarização do soluto pelo solvente.[113] Nessa situação, o Hamiltoniano de interação pode ser descrito da seguinte forma:

$$\hat{H}_{QM/MM} = \sum_a \sum_b 4\varepsilon_{ab} \left[\left(\frac{\sigma_{ab}}{r_{ab}} \right)^{12} - \left(\frac{\sigma_{ab}}{r_{ab}} \right)^{6} \right] + \sum_m \sum_b \frac{Z_m q_b}{r_{mb}} - \sum_i \sum_b \frac{q_b}{r_{ib}}. \tag{3.16}$$

O potencial Lennard-Jones continua inalterado, entretanto, ocorre mudança na descrição da contribuição eletrostática composta pelos dois últimos termos. O primeiro corresponde à interação das cargas parciais do solvente com o termo clássico correspondente aos núcleos do soluto, já o segundo termo corresponde à interação entre as cargas parciais e os elétrons do soluto. Ressaltamos, porém, que a utilização de métodos semiempíricos na região quântica é bem frequente devido ao mencionado tipo de simulação ter um alto custo computacional.[113] A utilização de métodos ab intio e DFT também têm sido aplicados aos cálculos QM/MM, mas isso ocorre com maior frequência em cálculos single point.[113] No intuito de tentar reduzir o custo computacional envolvido nos cálculos QM/MM, diversos trabalhos têm sido realizados, como o de Cubero, E. *et al.*,[125] Minõ, P. L. e Callis, P. R.[126] e Cui, Q. e Karplus, M..[127]

Merece nossa atenção a situação em que se considera tanto a polarização do soluto quanto a do solvente. Nesse caso, ocorre a atribuição de um tensor de polarizabilidade (α) que está localizado no centro de massa das moléculas do solvente ou pode-se também considerar que cada átomo possua uma polarizabilidade associada, de modo que o momento de dipolo induzido em cada centro seja dado por:

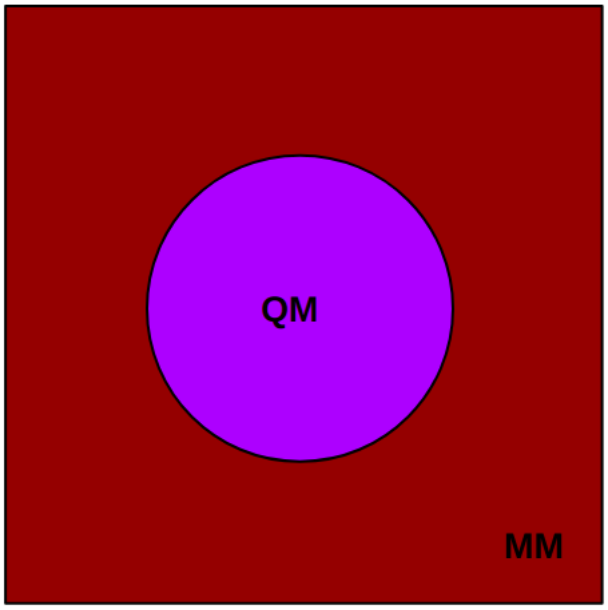
$$\mu_{ind} = \alpha \cdot E; \tag{3.17}$$

onde o vetor campo elétrico (E) de uma determinada molécula é gerado pelas cargas parciais, pelos momentos induzidos das demais moléculas, bem como pela distribuição de carga na região quântica. Os trabalhos de Thompson, M. A. *et al.*[128,129] utilizam essa abordagem, mas ela é pouco usual, também em decorrência do alto custo computacional. Já cálculos *single-point* utilizando métodos QM/MM inteiramente acoplados são mais utilizados.[113]

Uma situação mais complexa, em que se necessita de outra abordagem, refere-se ao estudo de uma reação que ocorre no sítio ativo de uma enzima. Nesse caso, é preciso cortar ligações químicas, pois uma parte da proteína receberá um tratamento quântico (QM), enquanto o restante da proteína receberá um tratamento clássico (MM). Esse tipo de abordagem exige que se decidam quais ligações podem ou devem ser cortadas e o local em que é melhor cortar;

portanto, requer uma "sensibilidade química" mais aprofundada. Parece mais adequado cortar ligações pouco polares e fracamente polarizáveis, como ligações simples C-C em carbonos sp^3. Com o objetivo de buscar representar da melhor forma possível a densidade eletrônica entre a ligação QM-MM, duas aproximações distintas, que compartilham ligações químicas, merecem destaque no acoplamento QM/MM: átomos de ligação (*link atoms*) e orbitais congelados (*frozen orbitals*).[113]

Figura 3.1: Esquema de partição de um sistema com o método QM/MM.



Na abordagem dos átomos de ligação, ocorre o acréscimo de um átomo (hidrogênio ou valência um) para completar cada valência insaturada na fronteira. A visibilidade desses átomos ocorre somente no subsistema quântico (QM), sem interação com a parte clássica (MM). Dessa forma, os átomos são inclusos no Hamiltoniano ($\hat{H}_{QM}$).[113] Considerando-se que a inclusão do átomo deve ocorrer ao longo da ligação QM-MM, surgiram várias propostas acerca da ligação QM-H (em que H representa a ligação do átomo) as quais remetem a um mesmo pensamento que consiste em igualar o raio R(QM-H) a uma fração do raio de R(QM-MM).[113] Na referida abordagem, as interações intermoleculares QM/MM são tratadas de forma semelhante à modelagem MM, o que simplifica a implementação, todavia, as interações eletrostáticas intermoleculares apresentam uma complexidade maior, pois leva em consideração as interações entre o subsistema quântico e a carga parcial do átomo MM na fronteira. Para se evitar uma polarização de cargas da região quântica pelos átomos MM da fronteira, devido à sua proximidade, as cargas dos átomos MM da fronteira são zeradas.[113]

Já na abordagem da aproximação de orbital congelado, ocorre a transformação das cargas pontuais próximas da fronteira em distribuição de cargas. Essa distribuição é tratada quanti-

camente e, depois de calculada, é mantida fixa para evitar a polarização artificial da parte MM próxima. Apesar de bastante atraente do ponto de vista teórico, esse método é mais complexo e mais caro computacionalmente. Por fim, entre as metodologias que utilizam esse tipo de aproximação, podem-se destacar o método do campo autoconsistente localizado (LSCF)[130,131] e o método do orbital híbrido generalizado (GHO).[132–134]

3.2.2 Método QM/MM sequencial

Uma nova metodologia chamada QM/MM sequencial foi desenvolvida por diversos pesquisadores.[113] Nessa metodologia, o tratamento da parte clássica (MM) e da parte quântica (QM) é realizada de modo separado. Inicialmente, é feito o tratamento clássico e, em seguida, ocorre o tratamento quântico.

Dentre esses trabalhos que utilizam a metodologia citada, destacamos o de Coutinho, K. e Canuto, S.[135] e Canuto e colaboradores.[136] Os autores desenvolveram um método QM/MM sequencial que, além de buscar minimizar algumas dificuldades apresentadas pelo QM/MM convencional, também garante que os resultados finais obtidos sejam médias estatisticamente convergidas.

Observamos que, atualmente, há sofisticados métodos de química quântica disponíveis, estando consolidado o tratamento de moléculas isoladas. Por outro lado, o tratamento de moléculas em meio líquido não apresenta a mesma situação, havendo ainda grandes desafios no sentido de desenvolver técnicas e algoritmos para se estudar "efeitos de solvente" e a modificação das propriedades moleculares, devido a interação com o meio. Vale a pena destacarmos o método geral (Gen-Ew) de Tatiana Vasilevskaya e Walter Thiel.[137]

3.3 Simulação Molecular do Enovelamento de Proteínas

A simulação computacional do enovelamento de proteínas foi uma alternativa de estudo desde o surgimento da área, por haver grande dificuldade de se obter um mecanismo com as técnicas instrumentais da época.[138] O objetivo dessa simulação no enovelamento de uma dada sequência é caracterizar a superfície de energia potencial do processo, desde a estrutura desnaturada até a estrutura nativa, ou de desenovelamento, partindo da estrutura nativa até uma estrutura desenovelada.[139] Com isso, é possível estabelecer caminhos entre conjuntos de estruturas e determinar as interações e movimentos da cadeia principal durante o enovelamento.[58] Essa abordagem é realizada através da obtenção da densidade de estados, isto é, do número de configurações com mesma energia, de forma que todas as propriedades ter-

modinâmicas podem ser calculadas através da função de partição e da definição da entropia configuracional.[140]

As propriedades do sistema podem ser amostradas do *ensemble* de configurações possíveis, utilizando-se uma média de um número representativo dessas configurações. Isso é realizado por meio de técnicas de geração estocástica de configurações nas condições do *ensemble*, separando somente aquelas que podem ser termicamente povoadas nas devidas condições macroscópicas. Essas técnicas estão inseridas na classe dos métodos de simulação de Monte Carlo.[141] Por outro lado, as técnicas de simulação de dinâmica molecular geram as configurações possíveis em função dos movimentos atômicos no tempo, governados pelos gradientes correspondentes aos termos de energia potencial considerados, que são as forças. A média do *ensemble* é tratada como a média temporal dessas configurações geradas.[142]

No período em que as simulações de enovelamento se iniciaram, a área de simulação computacional de moléculas estava em pleno desenvolvimento de métodos aproximados, uma vez que havia limitação computacional. Para o estudo de proteínas, a restrição era ainda maior devido ao número e natureza de interações envolvidas.[140]

Além de tratar as moléculas e os ambientes usando modelos que correspondessem ao tratamento de Física Clássica, sem contabilizar efeitos não negligenciáveis inerentes a sua estrutura eletrônica, foram necessárias outras aproximações, chamadas de modelos mínimos, assim nomeados para referir-se ao conjunto de aproximações feitos para tratar sistemas grandes como os biológicos.[59] As aproximações consistem basicamente em negligenciar certas interações e estimá-las na função de energia. Nos modelos iniciais para métodos de simulação, as interações entre os átomos de cada resíduo eram desconsideradas, a energia do resíduo era contabilizada a partir de um valor médio, e a função de energia se encarregava de calcular a energia potencial das interações entre os resíduos para cada conformação.

Atualmente, com o poder computacional disponível, a técnica mais indicada para realizar experimentos de dinâmica de enovelamento por simulação é a técnica de Dinâmica Molecular (DM),[143] por corresponder mais intrinsecamente à física do processo. Os campos de força modernos já contam com todos os tipos conhecidos de interações presentes na proteína que podem ser descritas classicamente. Apesar da estrutura eletrônica descrita pela mecânica quântica não ser contemplada, todas as propriedades médias são baseadas em parâmetros calculados por métodos quânticos como: distâncias de equilíbrio entre ligações específicas, ângulos de ligação, ângulos de torção e etc. Todavia, destaque-se que os campos de força ainda falham ao contabilizar os efeitos de polarização e transferência de carga, o que no futuro, com o esperado aumento de poder computacional, pode ser resolvido usando descrições das

estruturas eletrônicas dos átomos.[144]

Uma limitação da DM, reconhecida por alguns no campo do enovelamento de proteínas, é a falta de consideração sobre a cooperatividade que cada sequência deve apresentar para minimizar as frustrações.[145]

Ainda como um desafio computacional, o processo de enovelamento é muito longo para ser simulado inteiramente por DM de forma eficiente.[146] Para proteínas pequenas, já foi mostrado que é possível enovelar as estruturas por métodos *ab intio*[1] com simulação de dinâmica molecular.[147] No entanto, para estruturas com número maior de resíduos, tipicamente acima de 20, que se constituem da maioria das proteínas com funcionalidade de interesse nos organismos vivos, faz-se necessária a aplicação de estratégias de simulação para amostrar estruturas suficientes, com o objetivo de prover informações para o estudo do mecanismo do enovelamento.[148]

Existem duas técnicas que já foram muito utilizadas: a do desenovelamento e a do reenovelamento. Como a pesquisa sobre o enovelamento consiste em entender de que forma a proteína chega à estrutura nativa, os estudos são realizados sabendo-se inicialmente da estrutura cristalográfica para se comparar com os resultados do experimento. Essas estratégias citadas se valem do conhecimento prévio das estruturas nativas, a fim de gerar conformações iniciais para a simulação e uma trajetória no tempo de simulação estipulado.[149]

A estratégia de desenovelamento força a desnaturação da estrutura nativa identificando estruturas intermediárias e até de transição, considerando que essas estruturas seriam as mesmas nos dois sentidos do processo.[84] Já no reenovelamento, a proteína é também desenovelada até certo ponto e depois simula-se sua dinâmica normalmente para verificar se há o reenovelamento, podendo assim estudar o mecanismo com a trajetória gerada.[143]

Diversas técnicas experimentais e estratégias teóricas têm sido aplicadas para estudar o problema de enovelamento de proteínas. Entre as técnicas experimentais, podemos listar cristalografia de raios-X,[150] espectroscopia de RMN,[150] saltos de temperatura a laser,[151,152] espectroscopia de estrutura fina de absorção de raios-X próximo da borda (NEXAFS),[87,153] dicroísmo circular,[152] medições de fluorescência[152] e pinça ótica.[154] As estratégias teóricas incluem simulações de dinâmica molecular (DM),[143,149,155] métodos de amostragem ampliada[156] como *replica exchange molecular dynamics* (REMD)[157–159] , *umbrella sampling*[160] e metadinâmica.[161–163] Além dessas técnicas, destacamos as baseadas em aprendizado de máquina (*Machine Learning*) como o AlphaFold.[164] Ressaltamos porém, que essas técnicas

[1] O método *ab initio* aqui não se refere a cálculos de mecânica quântica. Essa nomenclatura é utilizada na modelagem de proteínas para designar que trata-se de um método que é independente de estruturas molde.

possuem um foco maior na predição da estrutura nativa e não nos mecanismos que regem processo do enovelamento.

3.4 Estudo do Enovelamento de Proteínas com o Emprego de Dinâmica Molecular

Após as contribuições apontadas por Anfinsen em 1961 e 1973, já discutidas na Introdução, diversos pesquisadores começaram a aplicar métodos de simulação molecular para compreender o processo de enovelamento das proteínas.

Em um dos primeiros trabalhos empregando DM no estudo do enovelamento de proteínas, Michel Levitt[59] apontou que o conceito de forças de tempo médio, introduzidas no ano anterior por ele e por Warshel, A., pode ser usado de modo a simplificar os cálculos de energia conformacional em proteínas globulares.[59] O autor descreveu detalhadamente a parametrização do campo de força da geometria molecular simplificada e apresentou testes para escapar de mínimos locais. Foram realizadas simulações por representações simplificadas do enovelamento do inibidor da tripsina pancreática bovina (BPTI), a partir de conformações de cadeia aberta (estado desenovelado) e da estrutura nativa.

As simulações de diversos estados desnaturados forneceram estruturas compactas que apresentaram várias características da estrutura nativa. Os cálculos por meio da representação simplificada reduziram o número de átomos e consideraram apenas os graus de liberdade efetivos, ou seja, os que apresentaram maior efeito sobre a conformação.[59] Tais cálculos buscavam economizar no número de amostragem de estruturas que apresentassem pequenas diferenças conformacionais, metodologia que também simplificou a inclusão do efeito do solvente e de movimento térmico atômico, o que tornou o cálculo de energia menos sensível aos parâmetros acurados de energia. As simulações de DM propiciaram o estudo das etapas do processo de reenovelamento, obtendo-se informações úteis sobre interações locais, contribuições energéticas e conformações parcialmente enoveladas.[59]

As simulações capturaram dois momentos importantes no enovelamento, com rápida compactação até o desvio quadrático médio (rms) de 6 Å, e depois um ganho entálpico com interações entre cadeias laterais, através de ligações de hidrogênio e forças de Van der Waals. Apesar de serem encontrados alguns bons resultados, as simulações computacionais simplificadas apresentaram falhas na obtenção de estruturas próximas das nativas em que o rms era menor que 2,5 Å. Essa limitação se deu devido à falta de detalhamento das cadeias laterais que assumem o formato esférico nas simplificações realizadas. Caso as cadeias laterais tivessem

vários centros de interação, como no tratamento full atom, certamente encontrariam estruturas mais próximas da nativa.[59]

Outra limitação das simulações computacionais simplificadas foi que a energia da estrutura mais próxima da nativa nem sempre apresentou o menor valor energético. Além disso, surgiram diversos estados de mínimo distintos com energias bastante similares. Esses problemas só podem ser contornados por tratamento full atom,no entanto, seria bastante difícil realizar esse tratamento com os computadores da época. Desse modo, vale destacar que o tratamento simplificado do enovelamento de proteínas tinha sua importância, uma vez que apresentavam cálculos rápidos de energia, facilidades na inclusão de efeitos de perturbações do solvente, vibração de grupo e energia térmica. Ademais, a superfície de energia possuía menor número de mínimos, o que facilitava bastante o tratamento dos sistemas.[59] Essa abordagem foi importante, em especial, porque introduziu pesquisas sobre o enovelamento à luz da DM; por outro lado, não atende às demandas atuais, o que a torna ultrapassada.

Com a evolução dos computadores contendo processadores *multicore*, a utilização de GPUs e a mudança de paradigmas, em que a programação serial passou a ser substituída por programação paralela de alta performance, já é possível realizarmos simulações de DM de enovelamento/desenovelamento *full atom* de proteínas, tornando as aproximações dos sistemas desnecessárias e obsoletas. Apresentaremos a seguir alguns trabalhos mais atuais sobre os estudos do enovelamento de proteínas empregando técnicas de DM.

A primeira questão a ser analisada refere-se à possibilidade de estudar o enovelamento de proteínas de forma eficiente, por meio de dinâmicas moleculares clássicas. Para verificá-la, diversos pesquisadores realizaram testes avaliando a performance da DM, bem como os modelos de campos de forças e efeitos do solvente. Os resultados obtidos foram comparados com dados experimentais. Nessa mesma perspectiva, Snow, C. D. *et al.*[152] realizaram simulações de DM de enovelamento de mutantes da mini proteína BBA5 (23 resíduos) e compararam os resultados com dados experimentais. Eles perceberam que a dificuldade em simular o enovelamento consistia no custo e nas limitações computacionais da época. Já era conhecido que as estruturas secundárias como hélice α e folhas β se formavam entre 0,1 e 10 μs, ao passo que pequenas proteínas demoravam na ordem de dezenas de microssegundos para se enovelarem.

Até a época da publicação de Snow, C. D. *et al.*,[152] a simulação mais longa tinha sido de 1 μs da proteína villin headpiece. Porém, essa simulação única poderia apresentar o problema de não contemplar todas as características do processo de enovelamento devido aos diversos estados de transição possíveis.[152] Para tentar contornar esse problema, os autores adotaram a estratégia de realizarem DM de dezenas de milhares de trajetórias de 5 a 20 ns, totalizando

700μs de simulação para a BBA5 mutante. Essas dinâmicas rápidas de relaxamento foram comparadas com dados experimentais de salto de temperatura a laser, sendo que os tempos médios de enovelamento da DM e as constantes de equilíbrio apresentaram plena concordância com as previsões experimentais. Desse modo, foi observado que a BBA5 se enovela rapidamente porque a estrutura secundária se forma também rapidamente.[152]

As simulações foram realizadas com o programa Tinker, utilizando o campo de força OPLS e o modelo implícito de solvente.[152] Isso só foi possível devido ao projeto de computação distribuída, chamado de Folding@Home, isto é, um grupo de computadores voluntários distribuídos em diversos países. Em 2002, o projeto possuía mais de 30.000 computadores e uma capacidade acumulada de 1 milhão de dias de CPU de tempo de simulação.[152]

Rhee, Y. M. *et al.*[165] buscaram responder aos seguintes questionamentos que estavam em aberto sobre o papel do solvente no processo de enovelamento de proteínas:

1. *A água apenas induz forças hidrofóbicas entre os resíduos das cadeias laterais da proteína (colapso hidrofóbico) ou a natureza discreta da água desempenha um papel estrutural no enovelamento?*

2. *Os efeitos da água, que não são contemplados nos modelos de solvatação implícita, alteram significativamente o mecanismo do enovelamento?*

Para responder a essa problemática, Rhee, Y. M. *et al.*[165] realizaram simulações de DM da BBA5 semelhantes às tratadas por Snow, C. D. *et al.*,[152] porém com a adoção do modelo de solvente explícito. A proteína foi solvatada com 3.938 moléculas de água, aplicando-se o modelo TIP3P e o campo de força AMBER94.[113,114] As cadeias laterais ácidas e básicas da proteína foram protonadas, assumindo o PH neutro, e adicionou-se um íon cloreto para refletir a neutralidade da carga do sistema solvatado, o que totalizou um sistema de 12.200 átomos em uma caixa cúbica 50 Å de lado. As simulações ocorreram a temperatura e pressão constantes com o uso do pacote GROMACS modificado para a infraestrutura Folding@Home.[165]

Por meio da análise das trajetórias, observaram-se vários eventos de enovelamento, de modo que foi possível avaliar diversos aspectos da cinética do enovelamento de proteínas, desde a estrutura desenovelada até o estado nativo. A conclusão apresentada por Rhee, Y. M. et al. foi a de que, tanto as taxas de enovelamento quanto a taxa da constante de difusão corrigida, apresentaram resultados com excelentes concordâncias com os dados experimentais.[165]

Ainda nesse trabalho, em busca de verificar o papel da água no processo de enovelamento, Rhee *et al.* compararam o modelo explícito e os modelos implícitos do solvente. Análises da densidade do solvente próximo aos grupos hidrofóbicos da BBA5 indicaram a existência de efeitos induzidos pela água não capturados por modelos de solvatação implícita. Nesse

sentido, mostraram-se presentes sinais que indicam que a densidade do solvente ao redor do núcleo diminui simultaneamente com o colapso hidrofóbico do núcleo da proteína, o que sugere a expulsão de moléculas de água no processo de enovelamento. Embora haja divergências entre os resultados dos modelos implícitos e explícitos do solvente, o mecanismo encontrado para o enovelamento da BBA5 foi o de difusão-colisão, em que as estruturas secundárias se formam primeiro independentemente e, em seguida, se colidem para formar a estrutura nativa.[165]

Apesar de o mecanismo geral ser o mesmo, Rhee, Y. M. *et al.* observaram que não houve correlação entre os valores de P_{fold}[165] (probabilidade de se enovelar ou desenovelar) das simulações com solvente explícito em relação às com solvente implícito. Esses valores consistem em uma métrica importante para determinação do estado de transição do enovelamento e referem-se à probabilidade de uma determinada proteína se enovelar dentro de um intervalo de tempo curto, que nesse trabalho foi de 5ns.[165]

O estudo de Rhee *et al.* contou com uma amostragem de 25 estruturas em trajetórias de enovelamento com desvio quadrático médio de carbono α (RMSD-Cα) entre 1,7 e 7,5 Å. Para cada conformação, ocorreram 100 simulações independentes com vetor velocidade aleatórios. Para uma transição de 2 estados (desenovelado e enovelado), as estruturas com valores de $P_{fold} < 0{,}5$ fizeram parte do grupo de estado de transição, sendo o valor de P_{fold} foi utilizado para ordenar as estruturas ao longo da coordenada de reação. Os autores propuseram que o estado de transição no modelo implícito está mais próximo da estrutura nativa do que o encontrado pelo modelo de solvatação explícita, provavelmente em razão da natureza discreta da água que está presente apenas nos modelos explícitos.[165]

A simulação da DM *full atom* do processo de enovelamento de proteínas promoveu análises com alto nível de detalhamento. Ressaltamos, entretanto, que em 2008 ainda havia dificuldades em amostrar tempos na escala de microssegundos necessários para o enovelamento, além da necessidade de realizar testes de campos de força para avaliar a performance no processo de enovelamento de proteínas.[147]

Fredolino, P. L. *et al.*[147] avaliaram a Fip35, uma proteína mutante de enovelamento rápido da Pin1 WW domain. Os autores estimaram que o tempo experimentalmente de enovelamento seria de 13,3 μs; e também realizaram uma simulação de 10μs, com o objetivo de obterem dados mais detalhados do processo de enovelamento da Fip35. Eles observaram que a proteína simulada apresentava diversos estados metaestáveis com topologia incorreta, não assumindo o estado nativo. As simulações da Fip35 iniciaram-se no estado desenovelado, revelando estruturas helicoidais bem diferentes da enovelada. Segundo os autores, apesar de haver

diversos trabalhos que apontavam o sucesso do enovelamento de um polipeptídeo desnaturado até o seu estado nativo ou quase nativo usando DM, até então, encontrar estruturas com RMSD-Cα < 2,0 Å em relação à estrutura cristalográfica foi raramente alcançado.[147]

A trajetória de 10 μs do WW domain representava uma das mais longas simulações de DM até o ano de 2008, o que só foi possível por haver o paralelismo programado no código do NAMD.[147] Fredolino, P. L. *et al.* destacam que se faz necessária uma série de simulações de acompanhamento e experimentos para entender por que, especificamente, essa trajetória de enovelamento em solvente explícito (modelo TIP3P) e campo de força CHARMM22 não alcançou uma conformação semelhante à nativa. Na realidade, essa situação pode apontar falhas no campo de força, uma vez que o tempo de simulação de enovelamento foi bem próximo do experimental. Todavia, o tempo de simulação de DM pode ser também influenciado pela estrutura de partida, uma vez que há proteínas cujo comportamento de enovelamento é variável, dependendo das condições iniciais.[147]

Em face do cenário descrito, surgiu a seguinte pergunta: quão robustas são as simulações do enovelamento de proteínas com respeito à parametrização de campo de força? Buscando respondê-la, Piana, S., *et al.*[166] simularam o equilíbrio de uma variante da proteína villin headpiece com quatro campos de força distintos pertencentes às famílias CHARMM e Amber: CHARMM27, CHARMM22*, AMBER ff03 e AMBER ff99SB*-ILDN. Foram realizadas simulações longas de 100, 300, 100 e 117μs com os campos de força ff03, ff99SB*-ILDN, CHARMM27 e CHARMM22*, respectivamente.

Como resultado, a proteína se enovelou e desenovelou reversivelmente por mais de 50 vezes em cada simulação. A estrutura nativa considerada foi o centróide do cluster mais populoso, obtido por métodos de clusterização com um raio de corte (RMSD = 1,0 Å). Compararam-se as estruturas do centróide com a estrutura cristalográfica, apresentando RMSDs de 1,3 Å, 0,7 Å, 0,6 Å e 0,7 Å para os campos de forças FF03, ff99SB*-ILDN, CHARMM27 e CHARMM22*, respectivamente.[166] Isso demonstra que, do ponto de vista estrutural, todos os campos de força encontraram estruturas bem próximas do estado nativo (RMSD < 2 Å).

Os tempos de enovelamento também estão de acordo com o experimental, que é de aproximadamente 1μs, e a dependência da taxa de enovelamento em relação à temperatura apresentou-se relativamente modesta. Portanto, concluíram os autores, que todos os campos de força retornam resultados satisfatórios em relação à estrutura e a cinética do enovelamento da villin.[166]

Em se tratando das entalpias de enovelamento, verificou-se que os campos de força ff99SB*-ILDN, CHARMM22* e CHARMM27 apresentaram concordância razoável com as experimen-

tais (aproximadamente 25 kcal/mol), enquanto que o ff03 apresentou resultados menores (9,7 kcal/mol). Mesmo havendo concordâncias, Piana, S., *et al.* descobriram que os quatro campos de força diferem em suas superfícies de energia livre, em que o CHARMM27 e o ff03 (em menor grau) favorecem um estado de desenovelamento helicoidal e o mecanismo do tipo difusão-colisão.

Nos campos de força (ff99SB*-ILDN e CHARMM22*), a formação das hélices e da estrutura secundária acontece de forma mais cooperativa, semelhante ao mecanismo de nucleação-condensação. Os autores recomendaram muita cautela ao tentar prever mecanismos de enovelamento a partir de dinâmicas moleculares, pois isso só pode ser feito se houver possibilidade de comparação com dados experimentais.[166]

Com base na discussão apresentada, podemos afirmar que, apesar dos avanços obtidos nos últimos anos, tanto na área experimental quanto em simulações computacionais, compreender o processo pelo qual as proteínas se enovelam até a conformação nativa ainda é um desafio para diversas áreas do conhecimento. Centenas de estudos já abordaram o tema, porém, a caracterização experimental do caminho completo que leva uma sequência primária a assumir sua conformação tridimensional é uma tarefa bastante árdua, porque as proteínas apresentam tamanho, topologia e estabilidades bastante distintas. Logo, simular o enovelamento de uma pequena proteína com detalhamento em nível atômico é uma tarefa bastante complexa.[112]

Por essa razão, tanto estudos experimentais quanto computacionais têm focado geralmente em uma proteína por vez, abordando o problema sob condições diversas ou por meio de técnicas distintas. Diversas teorias foram propostas para descrever o processo de enovelamento, mas não se chegou a nenhum consenso, uma vez que há muitas opiniões diferentes, mesmo para os princípios básicos do enovelamento obtidos por interpretações variadas de um grande número de experimentos de enovelamento.[167]

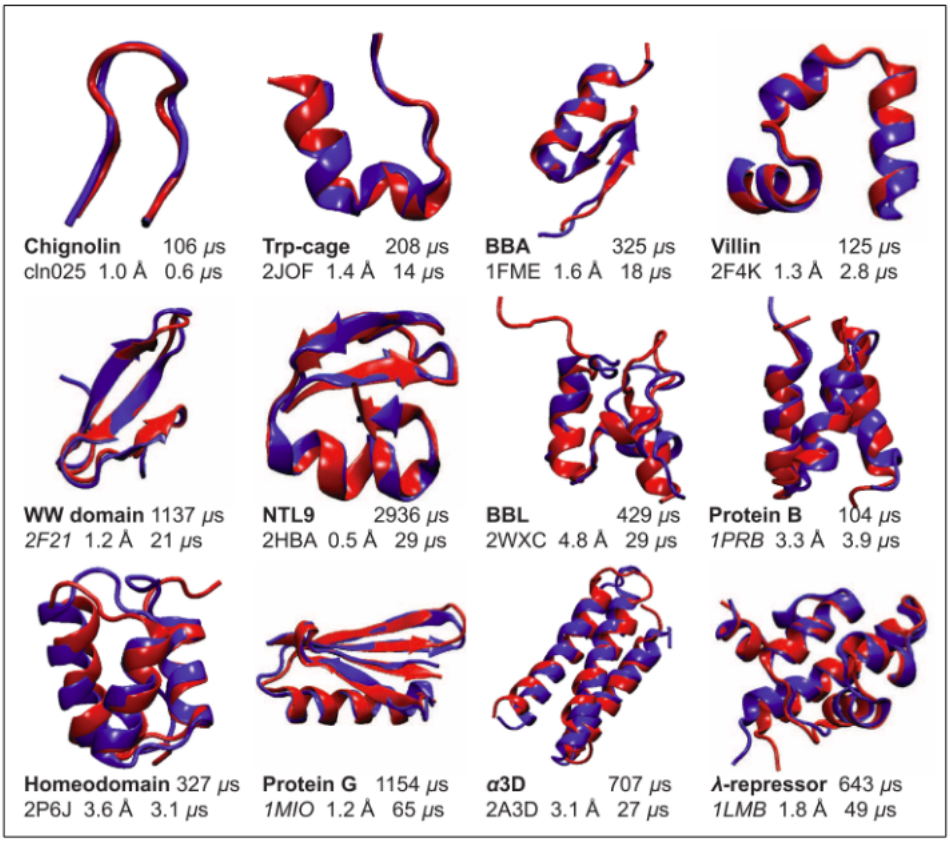
Lindorff-Larsen, K. L., *et al.*,[112] preocupados em elucidar eventos gerais que possam fornecer ajuda para se chegar a um consenso no processo de enovelamento, realizaram a simulação de DM *full atom* para um conjunto de 12 proteínas de rápido enovelamento. Essas proteínas apresentam tamanhos entre 10 e 80 resíduos, os quais incluem membros das três principais classes estruturais (hélice α, folhas β e misto α/β), formando um conjunto estrutural diversificado de proteínas.

As proteínas foram simuladas em solvente explícito TIP3P, com uso do campo de força CHARMM22*, em períodos que variam de 100μs a 1ms e que revelaram diversos princípios comuns ao enovelamento das 12 proteínas. Constatou-se que 11 entre as 12 proteínas se enovelaram de forma espontânea para a estrutura nativa.[112] A estrutura *homeodomain Engrailed*

demonstrou instabilidade durante a simulação, porém, os pesquisadores conseguiram enovelar uma mutante da *homeodomain* que mantêm as mesmas características gerais da *homeodomain Engrailed*.

Além disso, constatou-se também pelo menos 10 eventos de enovelamento e 10 de desenovelamento por estrutura, sendo coletados aproximadamente 8 ms de simulação, com um total de mais de 400 eventos desse tipo. Os resultados apontaram que 8 das 12 proteínas simuladas apresentaram RMSD-Cα dentro da faixa de 2 Å em relação à estrutura cristalográfica. Na Figura 3.2, a seguir, é apresentada a superposição entre as estruturas cristalográfica e as obtidas da simulação de Lindorff-Larsen, K. L., *et al.*[112] A estrutura da DM tida como a nativa resultou do método de clusterização, em que o centroide do *cluster* mais populoso foi considerado como a estrutura nativa.[112]

Figura 3.2: Sobreposição das estruturas representativas do estado enovelado observadas em simulações reversíveis de 12 proteínas em relação à estrutura nativa. A estrutura enovelada obtida da simulação está em azul e a cristalográfica está em vermelho. Também é mostrado o tempo total de simulação, a identificação PDB da estrutura experimental, o RMSD-Cα (sobre todos os resíduos) entre as duas estruturas e o tempo de enovelamento. As entradas do PDB, em itálico, indicam que a estrutura não foi determinada para a sequência simulada e que, em vez disso, a comparação foi realizada com a estrutura do homólogo mais próximo no PDB. Figura retirada de referência [112].



Por meio da análise das trajetórias, os autores conseguiram fornecer uma análise unificada sobre o enovelamento das 12 proteínas estudadas. Eles observaram que são transitoriamente formados elementos da estrutura local nativisada (nativelike) no estado desenovelado, ao mesmo tempo em que se formam alguns elementos de estrutura secundária e um número pequeno de contatos não-locais, os quais promovem a estabilização dos elementos estruturais, iniciando-se assim o estado de transição do enovelamento.[112] Lindorff-Larsen, K. L., *et al.* acrescentaram que, em grande parte dos casos, o processo do enovelamento segue por um caminho dominante, em que os elementos da estrutura nativa são formados em uma ordem bem definida e altamente correlacionada com a sua predisposição a se formar no estado desenovelado.

Contudo, verificou-se para duas proteínas (NTL9 e a variante da *Protein G*) que o mecanismo de enovelamento é heterogêneo, pois contém grupos distintos de estado de transição. Além disso, verificou-se também que um único campo de força foi capaz de enovelar um grande número de proteínas, abrangendo as três principais classes estruturais até atingir o estado nativo. Isso sugere que os campos de força da atualidade apresentam acurácia suficiente para que uma simulação de DM de longa escala possa ser empregada na caracterização de mudanças conformacionais em proteínas.[112]

As proteínas de enovelamento rápido são aquelas que se enovelam na escala de microssegundos, tratando-se de uma classe especial. Após a aplicação de simulações longas de DM bem-sucedidas, que elucidaram os mecanismos que desencadeiam o processo de enovelamento de proteínas, surgiu a seguinte pergunta: *até que ponto as conclusões obtidas por análises realizadas para proteínas de rápido enovelamento aplicam-se a proteínas com enovelamento na faixa de milissegundos ou de enovelamento lento?*

Com o objetivo de abordar essa questão, Piana, S. *et al.*[151] realizaram simulações moleculares para avaliar o processo de enovelamento da ubiquitina, uma proteína com 76 resíduos, amplamente estudada experimentalmente e que apresenta um tempo de enovelamento na faixa de milissegundos. Eles realizaram oito simulações de DM usando o campo de força CHARMM22* e solvatação explicita com moléculas de água do modelo TIP3P.

Seis simulações foram iniciadas no estado enovelado, partindo-se da estrutura cristalográfica (PDB ID: 1UBQ), e duas partiram de um estado desenovelado (com a cadeia polipeptídica estendida). Com a aplicação de simulações atomísticas de equilíbrio, foram determinados o mecanismo, a cinética e a termodinâmica do enovelamento da ubiquitina. Piana, S. et al. constataram que, de um modo geral, o mecanismo da ubiquitina é consistente com um grande conjunto de dados experimentais.[151]

Ao compararem esse mecanismo com o trabalho sobre as 12 proteínas de rápido enovelamento,[112] eles também constataram que, embora a ubiquitina demore muito mais tempo para se enovelar, os mecanismos que governam as proteínas de rápido enovelamento parecem se aplicar também à ubiquitina. Desse modo, os estudos experimentais e computacionais de pequenas proteínas de rápido enovelamento podem fornecer informações valiosas sobre o processo de enovelamento de proteínas em geral, em que as de ocorrência natural são inclusas.[151]

A integração entre métodos experimentais e computacionais de resolução atômica apresentam um alto potencial para auxiliar no entendimento de aspectos centrais no processo de enovelamento de proteínas. Considerando-se que, tanto os métodos experimentais quanto

computacionais apresentam capacidades e limitações distintas, faz-se necessária sua integração, de modo a serem obtidas informações mais acuradas relativas ao entendimento do processo de enovelamento.

Atualmente, as técnicas experimentais já são capazes de produzir resolução em nível atômico através da escala de tempo de milissegundos; todavia, em escalas abaixo de milissegundos, tal resolução só é possível com o uso de sondas espectroscópicas *coarse-grained*.[168] Por outro lado, já é possível gerar trajetórias de DM de enovelamento e de desenovelamento com bastante riqueza de detalhes, através de simulações *full atom* na escala de microssegundos. Entretanto, isso apresenta alto custo computacional e necessita de aproximações físicas com campos de força clássicos, ainda não totalmente validados.[168]

O processo do enovelamento de proteínas se dá através da formação de redes cooperativas de interações fracas entre os resíduos, que são o ponto chave para o entendimento do mecanismo de enovelamento.[84] A seguinte questão se impõe nesse contexto: *quais são os principais elementos que compõem a rede de interações resíduo-resíduo no processo de enovelamento da proteína?*

Sborgi, L. *et al.*,[168] na tentativa de esclarecer quais são os principais elementos que compõem essa rede de interação, avaliaram o enovelamento da proteína gpW de domínio único, utilizando a combinação entre a ressonância magnética nuclear (RMN) e DM de longa escala. É preciso destacar que a gpW possui 62 resíduos que se enovelam em uma topologia mista (hélice α + folha β), antiparalela, na escala de microssegundos. Assim, a análise ocorreu por RMN do desenovelamento térmico, com a gpW em pH 3,5, e com a obtenção do deslocamento químico completo do desenovelamento da proteína em alta temperatura, sendo que a desnaturação térmica da gpW foi medida pela técnica espectroscópica do dicroísmo circular[169] (CD) de UV distante.

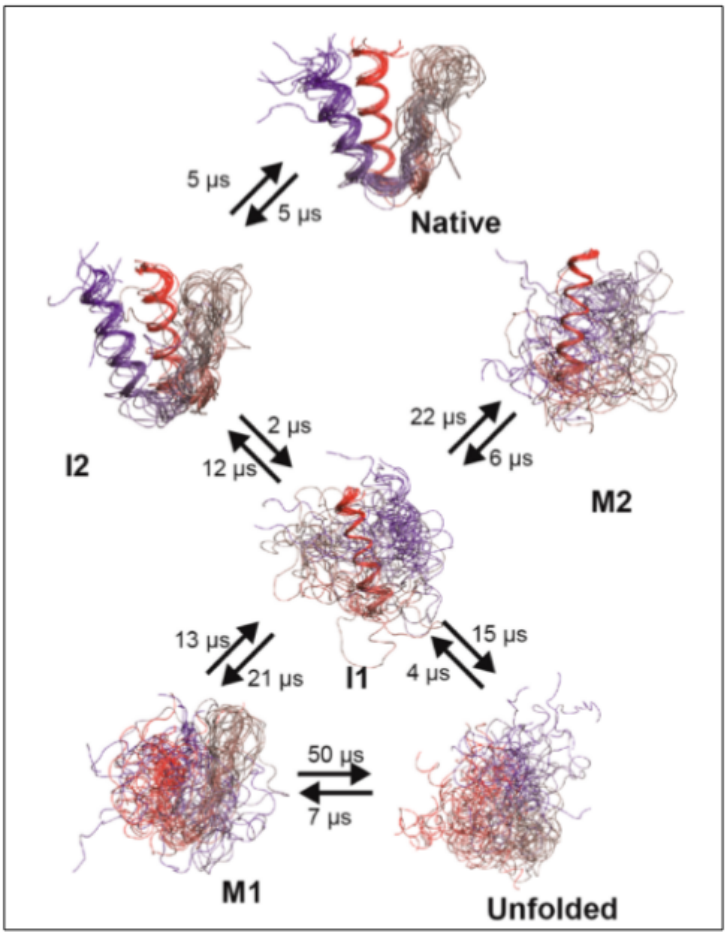
A RMN é uma técnica muito versátil na investigação de alterações conformacionais de proteínas com resolução atômica, pois fornece informações estruturais dos estados de transição do enovelamento que raramente são detectados, diretamente em decorrência da sua baixa ocupação e vidas curtas, isto é, os chamados "estados invisíveis". Com a evolução das simulações de DM para a escala de tempo de milissegundos, tornou-se possível a obtenção de trajetórias que possuem diversos eventos de enovelamento e desenovelamento de pequenas proteínas.[112]

Como os campos de força são baseados em modelos físicos imperfeitos (física clássica), é fácil perceber que o cruzamento entre os resultados obtidos por métodos experimentais e de DM para a gpW são úteis não só para compreender mais profundamente o processo de enovelamento, como também para avaliar a performance dos campos de força atuais. Ainda

nesse trabalho, Sborgi *et al*[168] realizaram uma simulação de 250μs de enovelamento reversível (enovelamento e desenovelamento) da gpW a 340K, com o objetivo de caracterizar a cinética do enovelamento. Também foram realizadas quatro simulações independentes (aproximadamente 200 μs) com temperatura simulada (ST), para calcular propriedades dependentes da temperatura, bem como comparar com os dados obtidos pelos experimentos de RMN. Por meio da comparação entre o experimento e as cinco simulações, avaliou-se a convergência dessas, bem como os cálculos dos erros estatísticos associados.

As simulações iniciaram-se na conformação estendida, sendo observados durante a trajetória um conjunto de eventos consistentes com o estado nativo (RMSD-Cα = 1 Å da estrutura obtida por RMN). Foi observado também, através dos dados de RMSD-Cα, um alto grau de heterogeneidade estrutural, sugerindo que a gpW amostra diversos estados metaestáveis e não somente o estado nativo e desenovelado (dinâmica de dois estados). Verificou-se também que, próximo ao ponto médio de desnaturação, há evidências de um rápido equilíbrio entre a estrutura nativa e um intermediário (I2), em que estão bem formados o núcleo hidrofóbico e as duas hélices α, porém a folha β não está estruturada (Figura 3.3). Por fim, o conjunto de conformações obtido pela DM estava em acordo com os obtidos experimentalmente (variação média de 0,02 Å), demonstrando que as estruturas nativas nas simulações de DM são bem semelhantes às obtidas por RMN.[168]

Figura 3.3: Modelo de Markov da superfície de energia livre do enovelamento obtido a partir da análise cinética de todas as trajetórias de simulações combinadas. A hélice 1 é mostrada em vermelho e a hélice 2 em azul.Também são relatadas as taxas de interconversão entre os clusters, estimadas a partir da simulação de equilíbio da DM em 340K. Figura retirada da referência [168].



Um tratamento mais detalhado do desenovelamento da gpW se deu através da análise dos acoplamentos termodinâmicos resíduo-resíduo. Nessa abordagem, os acoplamentos primários ocorrem através da interação de resíduo-resíduo na estrutura nativa, enquanto que os acoplamentos secundários se dão através de resíduos que estão indiretamente conectados por meio de interações mútuas com outros resíduos. Verificou-se que, tanto as matrizes de acoplamento dos experimentos de RMN quanto nas baseadas nas simulações, exibiram um padrão complexo de acoplamentos. Isso indica que há variações significativamente acentuadas nas interações resíduo-resíduo durante o processo de desenovelamento. Por meio do tratamento cinético obtido das simulações, observou-se que a folha β é flexível no estado nativo, sendo que não está estruturada nos outros estados.[168]

As matrizes de acoplamento simuladas e experimentais indicaram um forte acoplamento

entre as extremidades da hélice 1 e a alça e o início de hélice 2 da gpW. Além disso, foi observado um acoplamento marginal entre as extremidades da proteína, bem como entre a folha β e o restante da proteína. Esses resultados demonstram que as simulações são capazes de reproduzir os principais aspectos estruturais e mecanísticos do processo de enovelamento da gpW. Apesar das semelhanças, constataram-se que os elementos da estrutura secundária apresentaram maior estabilidade na simulação, resultando em uma superfície de energia livre com um número maior de estados metaestáveis quando comparados com o experimental. Os autores inferiram que isso ocorre devido às aproximações nos campos de força e nos modelos de solvatação atuais. O experimento de RMN surgiu como uma nova abordagem para a caracterização da rede de interações do processo de enovelamento de forma complementar às análises de simulação de DM.[168]

3.5 Desafios Atuais sobre a DM no Processo de Enovelamento de Proteínas

Os avanços recentes nas técnicas de desenvolvimento de software e de hardware tornaram possíveis as simulações de longa escala. Conforme apresentado até esta seção, o desenvolvimento de campos de força tem recebido bastante atenção da comunidade científica. Embora haja muito a ser realizado para desenvolver e validar campos de força com uma maior acurácia, já existem aqueles capazes de descrever o enovelamento de proteínas com um grau positivo de conformidade em relação aos dados experimentais.[143] Uma vez que já é possível gerar uma amostragem na escala de milissegundos em um campo de força relativamente acurado, colocando uma diversidade de proteínas ao alcance desses métodos, surge a seguinte pergunta: *quais os desafios atuais na simulação do enovelamento de proteínas?*

Apesar de ser significativamente desafiador, gerar trajetórias em um campo de força acurado está longe de ser o fim do caminho para o processo do enovelamento. Na realidade, em decorrência do fato de as trajetórias tornarem-se muito longas, surgiu outro grande desafio: transformar dados em conhecimento. Para se ter uma ideia, as simulações clássicas integram as equações de movimento de Newton em intervalos de tempo da ordem de femtossegundos (10^{-15}s), logo, simulações de enovelamento requerem aproximadamente 10^{12} passos no tempo para atingir a escala de milissegundos ($10^{-3}s$).[143]

Os *insights* obtidos a partir das simulações ajudaram a moldar o campo do enovelamento de proteínas através da integração entre dados simulados e experimentais como RMN.[168][60] Porém, estudos revelam que a análise de dados oriundos de simulações é complexa, uma vez

que já ocorreram situações em que certas técnicas levaram os pesquisadores a acreditar em resultados inconsistentes com seus dados de simulação bruta.[143]

Dessa forma, a comunidade de simulação tem se empenhado em desenvolver ferramentas de análise de dados gerais, que sejam robustas o suficiente para tratar o problema do enovelamento. Com o avanço das simulações na escala de milissegundos, a amostragem se tornou um problema menor, todavia, com isso surgiu outro desafio: *como tratar o grande volume de dados fornecidos pela amostragem da DM de modo a se obter informações relevantes para o enovelamento?*

Com o grande volume de dados gerados, as técnicas de análise tornaram-se o novo fator limitante para o entendimento do enovelamento de proteínas através de simulações de DM, consistindo no chamado desafio de *Big Data*.[143] Nesse sentido, a eficiência de um método de análise depende da redução dos dados de simulação, preservando as informações essenciais com o menor grau de simplificações possível. Atualmente, há duas classes de técnicas de análise de dados utilizadas pela comunidade de simulação; a saber:

- Métodos de busca de coordenadas de reação e estados de transição;
- Modelos de estado de Markov (MSMs).

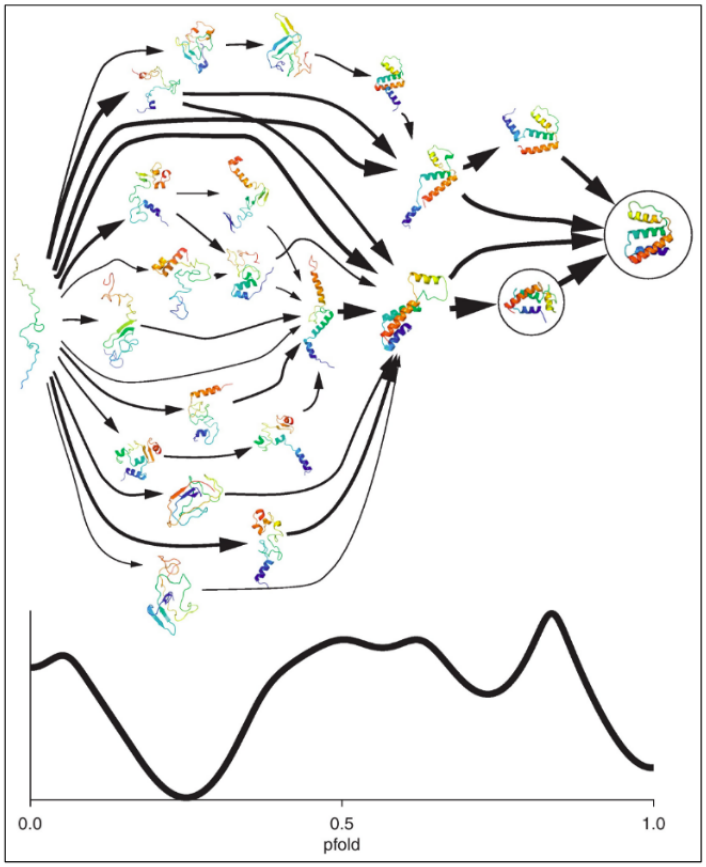
Nos métodos de coordenada de reação, ocorre a busca por uma coordenada única para descrever o progresso, desde o estado desenovelado até a estrutura nativa, construindo-se um modelo para a cinética ao longo da coordenada. Essa metodologia permite a localização de conjunto(s) de estado(s) de transição, composto(s) pelas estruturas ao longo da trajetória que apresentam a probabilidade de 50% de se enovelarem ou desenovelarem (P_{fold}).[170]

Como já discutido anteriormente, através do P_{fold} é possível ordenar as estruturas ao longo do caminho de reação. As estruturas do estado de transição apresentam relevância para a cinética do processo como em uma reação química convencional. Esses métodos são muito interessantes, uma vez que reduzem as informações em uma única coordenada, com intermediários (estados de transição) que podem ser investigados, descartando-se estruturas que não integram esse caminho.[143]

Por sua vez, os MSMs são representados por uma cinética de primeira ordem, através de um grupo de estados discretos. Nesse modelo, diversas simulações pequenas independentes são geradas e agrupadas estatisticamente, formando um modelo completo do sistema. Os MSMs fazem o agrupamento das trajetórias por meio da seleção de eventos que descrevam uma pequena parte do espaço de fase da proteína, semelhante a partes de um quebra-cabeça que, ao serem montadas todas as peças, formam a imagem completa. Tais métodos simplificam a análise das trajetórias, descartando eventos muito rápidos, abaixo do chamado tempo de

atraso (*lag time*). Como esse tempo é ajustável, pode-se alterar a resolução dos MSMs desde ajustes finos, da ordem de nanossegundos, até ajustes grosseiros da ordem de milissegundos ou mais. As trajetórias, quando combinadas em MSMs, são capazes de prever fenômenos cinéticos em uma escala de tempo muito maior do que as trajetórias individuais utilizadas no modelo.[143] Na Figura 3.4, a seguir, são apresentados exemplos de coordenada de reação e MSMs para a proteína Acil-CoA (ACBP):

Figura 3.4: Representação dos métodos MSMs (superior) e a coordenada de reação (inferior) para o sistema (ACBP). Na coordenada de reação, o enovelamento da proteína é representado por um progresso ao longo de um único grau de liberdade. Nos MSMs, a imagem é mais detalhada, uma vez que é capaz de capturar diversos caminhos paralelos para se chegar à estrutura nativa. Figura retirada da referência [143].



Estudos recentes apontam que a análise por coordenada de reação é capaz de reproduzir o tempo correto de enovelamento, bem como tornar possível a proteína se enovelar e desenovelar ao longo do caminho (P_{fold}). Esse método de análise permite-nos encontrar o caminho de enovelamento de maior fluxo, mas apresenta dificuldades de encontrar caminhos paralelos para o enovelamento. Por outro lado, os MSMs têm o potencial de fornecer diversos caminhos paralelos para o enovelamento, entretanto, apresentam ainda alguns desafios como a escolha da forma ideal para particionar o espaço de configuração e do tempo de atraso (*lag time*).[143]

3.6 Modelos de Markov

3.6.1 Dinâmica Molecular e Modelos de Markov

Os sistemas moleculares apresentam alta sensibilidade a detalhes no nível atômico, por exemplo, uma única mutação pontual pode ocasionar efeitos expressivos no enovelamento ou função de uma proteína. Para um entendimento detalhado, far-se-ia necessário a utilização de modelos atomicamente detalhados que fossem capazes de captar tanto a cinética quanto a termodinâmica do processo de interesse. Apesar de haver diversos métodos experimentais para sondar a estrutura e a dinâmica de grandes sistemas como as proteínas, nenhum desses é capaz de fornecer uma compreensão completa do sistema. Por exemplo, ao monitorar o relaxamento de um grupo de proteínas do estado desenovelado até a estrutura nativa, geralmente é observado um comportamento simples que pode ser bem ajustado por uma suavização exponencial simples ou dupla.

Para avançar em relação a esses modelos altamente grosseiros, seria preciso fazer perturbações como mutações ou tentar incorporar outros dados experimentais. Porém, devido à sensibilidade atomística de diversos processos moleculares, a interpretação dos efeitos das mutações torna-se complexa, sendo que a junção de tipos de dados experimentais distintos também não é fácil, devido à dificuldade de ponderar as contribuições relativas de tipos de dados para um determinado modelo. Desse modo, apresar dos avanços, não há atualmente um caminho claro para construção de modelos atomicamente detalhados para todo um sistema a partir de dados experimentais isolados. Nesse sentido faz-se necessário o desenvolvimento de modelos computacionais que complementem os dados experimentais, fornecendo uma descrição inequívoca dos movimentos atômicos de um sistema. Assim sendo, pode-se obter tanto informações estruturais quanto cinéticas do modelo que podem ser utilizadas para explicar as origens dos resultados experimentais e gerar hipóteses para guiar o planejamento de novos experimentos. Nesse contexto, as técnicas de dinâmica molecular são as mais utilizadas.[171] Os métodos de simulação molecular são uma forma poderosa para o entendimento de sistemas moleculares, principalmente quando esses são difíceis de investigar experimentalmente, como é o caso do enovelamento de proteínas.

Para extrairmos de forma satisfatória todo o potencial das simulações, faz-se necessário a aplicação de métodos que possam fornecer entendimento, estabelecerem uma conexão quantitativa com o experimento e impulsionar simulações eficientes. Nesse contexto, os modelos de estado de Markov (MSMs) tem se mostrado promissores no atendimento desses três requisitos.[171]

Um MSM, trata-se de uma rede de estados conformacionais e uma matriz de probabilidades de transição que descreva a probabilidade de um estado saltar para o outro em um pequeno intervalo de tempo (equivalente às equações mestras de tempo discreto).[172] Além disso, destacamos que os MSMs são baseados em critérios cinéticos ao invés de geométricos, logo, torna-se possível identificar os limites entre as bacias de energia livre com alto grau de precisão e modelar processos como o relaxamento para o equilíbrio.[171] O modelo de Markov é um granulado grosseiro (*coarse-graining*) que reflete o cenário de energia livre subjacente que determina a estrutura e a dinâmica do sistema. Os modelos de Markov são capazes de fornecer perspectivas importantes sobre uma molécula, porque promovem uma visão intuitiva muito melhor para estados e taxas (probabilidade de transição) do que um grande número de conformações geradas por simulações de dinâmica molecular.[171]

3.6.2 Formalismo das Cadeias de Markov

Um processo é chamado markoviano se dado um estado presente, a probabilidade de acessar um estado futuro seja independente do passado (processo sem memória). Na linguagem matemática dizemos que dado um espaço que possui k elementos $S = \{1, 2, 3, 4, \ldots, k\}$, imaginando que uma partícula salte entre esses estados em tempos discretos, a probabilidade de uma partícula estando no estado i saltar para o estado j é denominada **probabilidade de transição** p_{ij} e a matriz $P = [p_{ij}]$ é chamada **matriz de transição** da cadeia de Markov. Nesse processo, o $\sum_{i}^{k} P_{ij} = 1$ para todo $i \, \epsilon \, [1, k]$ e as probabilidades de transição não se alteram com o tempo, ou seja, essa cadeia de Markov é dita homogênea no tempo. De maneira técnica, um processo discreto $\left(X_n\right)_{n \, \epsilon \, N}$ é dito de Markov se a probabilidade condicional satisfizer:

$$P(X_{n+1} = X_{n+1} | X_0 = x_0, ..., X_n = x_n) = P(X_{n+1} | X_n = x_n); \tag{3.18}$$

em que para todo $n \geqslant 1$ e para toda a sequência de elementos no espaço de estados S. Desse modo, a incerteza do processo no tempo $t = n + 1$ (estado futuro) depende somente do estado atual $t = n$ (presente) e não dos estados ou trajetórias em tempos anteriores, conforme já afirmado anteriormente.[173,174]

Por exemplo, em uma cadeia de Markov com três estados, a matriz de transição (T) é apresenta o seguinte formato:

[1.....2....3] Estado Precedente

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$ Novo estado

Nessa matriz, p_{32} é a probabilidade de o sistema mudar do estado *3* para o estado *2* e p_{33} é a probabilidade do sistema permanecer no estado *3* imediatamente após ter sido observado no estado *3*.

No exemplo a seguir, retirado da referência[174] ilustramos uma aplicação da cadeia de Markov:

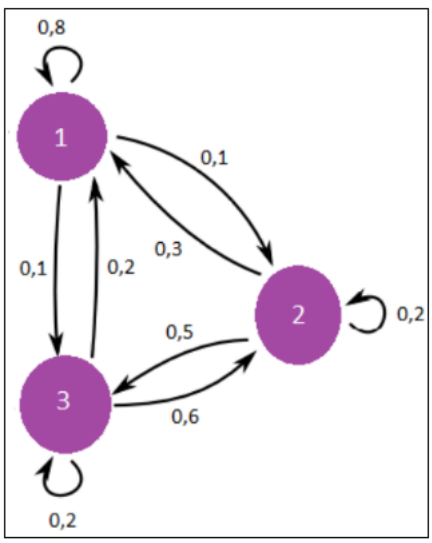
Exemplo 1: Uma locadora de automóveis tem três lojas de atendimento, denotadas por 1,2 e 3. Um cliente pode alugar um carro em qualquer uma das três lojas e devolver o carro para qualquer uma das três lojas. O gerente nota que os clientes costumam devolver os carros de acordo com as seguintes probabilidades:

[1.....2.....3] Alugado da loja

$$\begin{bmatrix} 0,8 & 0,1 & 0,1 \\ 0,3 & 0,2 & 0,5 \\ 0,2 & 0,6 & 0,2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$ Devolvido à loja

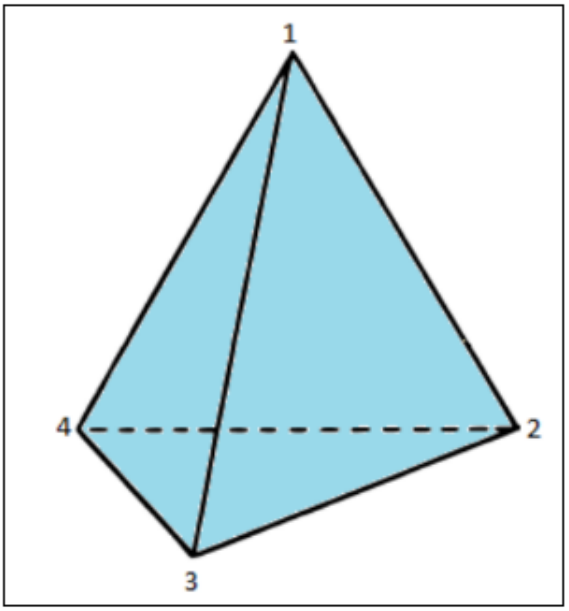
Essa é a matriz de transição se o sistema for considerado uma cadeia de Markov. Através da análise dessa matriz, podemos verificar que a probabilidade de um carro ser alugado na loja 2 e ser devolvido na loja 3 é de 0,5 e a probabilidade de um carro ser alugado na loja 3 e ser devolvido na loja 3 é de 0,2. Através da Teoria dos Grafos, podemos utilizar a matriz de transição para visualizar os grafos dirigidos dos pares ordenados, conforme figura a seguir:

Figura 3.5: Grafo com probabilidades de transição para um sistema com 3 estados.



Exemplo 2: Considere uma partícula que realiza movimentos aleatórios sobre os vértices de um tetraedro regular. Imagine que a cada passo a partícula possa saltar para um vértice vizinho com igual probabilidade. Considerando que o sistema segue uma cadeia de Markov, dê a matriz de transição.

Figura 3.6: Tetraedro regular



Como o tetraedro regular possui 4 vértices, o espaço do sistema é dado por: $\mathrm{S} = \{1, 2, 3, 4\}$. Considerando X_n como a posição da partícula no tempo n, logo:

$$P\left(X_{n+1} = j | X_n = i\right) = \begin{cases} \frac{1}{3} \text{ se } i \text{ e } j \text{ estão conectados,} \\ 0 \text{ caso contrário} \end{cases}$$

Como $P\left(X_{n+1} = j | X_n = i\right)$ é independente de n, a cadeia de Markov é homogênea, ou seja, as probabilidades de transição não mudam ao longo do tempo(estacionárias). Desse modo, a matriz de probabilidades ficará da seguinte forma:

$$P = p_{ij} = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix}$$

Verifique que a probabilidade da partícula sair do vértice *1* e ir para o vértice *3* é $p_{13} = \frac{1}{3}$.

3.6.3 Equação de Chapman-Kolmogorov

Como já discutido anteriormente, em uma cadeia de Markov, P_{ij} representa a probabilidade de um sistema no estado i salte para o estado j na próxima transição. Podemos também definir a probabilidade de transição de um sistema estando no estado i deslocar-se para o estado j após duas transições consecutivas $\left(P_{ij}^{(2)}\right)$ da seguinte forma:

$$P_{ij}^{(2)} = \sum_{k=0}^{M} P\left\{X_2 = j, X_1 = k | X_0 = i\right\}; \tag{3.19}$$

$$P_{ij}^{(2)} = \sum_{k=0}^{M} P\left\{X_2 = j, X_1 = k, X_0 = i\right\} P\left\{X_1 = k | X_0 = i\right\}; \tag{3.20}$$

$$P_{ij}^{(2)} = \sum_{k=0}^{M} P_{kj} P_{ik}. \tag{3.21}$$

Usualmente, definimos a matriz de probabilidades de n transições,$\left(P_{ij}^{(n)}\right)$ da seguinte forma:

$$P_{ij}^{(n)} = P\left\{X_{n+m} = j | X_m = i\right\}. \tag{3.22}$$

Para se calcular o valor de $\left(P_{ij}^{(n)}\right)$, utiliza-se a equação de **Chapman-Kolmogorov** conforme descrita e demonstrada a seguir.[175]

$$P_{ij}^{(n)} = \sum_{k=0}^{M} P_{ik}^{(r)} P_{kj}^{(n-r)}; \tag{3.23}$$

Tal que $0 < r < n$.

Demonstração:

$$P_{ij}^{(n)} = P\left\{X_n = j | X_0 = i\right\}; \tag{3.24}$$

$$P_{ij}^{(n)} = \sum_k P\left\{X_n = j, X_r = k | X_0 = i\right\}; \tag{3.25}$$

$$P_{ij}^{(n)} = \sum_k P\left\{X_n = j, X_r = k, X_0 = i\right\} P\left\{X_r = k | X_0 = i\right\}; \tag{3.26}$$

$$P_{ij}^{(n)} = \sum_k P_{kj}^{(n-r)} P_{ik}^{(r)}. \tag{3.27}$$

Podemos afirmar de forma simplificada, que as equações de Chapman-Kolmogorov fornecem uma maneira de se computar a matriz de n passos no tempo, ou seja, de m para m+2, m para m+3,..., m para m+n. Para cadeias de Markov homogêneas (probabilidades de transição constantes no tempo), pode-se calcular facilmente a matriz de transição do salto no tempo n, elevando-se a matriz de transição do salto um para a potência n, ou seja:

$$P^{(n)} = P^n. \tag{3.28}$$

3.6.4 Aplicações das cadeias de Markov nas áreas da química

As cadeias de Markov foram utilizadas amplamente para resolução de problemas principalmente nas áreas da Física e da Matemática, porém essa aplicação tem se estendido para outros campos da ciência como é o caso da química. Nessa área as aplicações são mais direcionadas para o campo da cinética química no sentido de descrever a dinâmica das reações.[172]

Há diversos métodos na literatura utilizados para descrever a cinética das reações químicas, sendo que o método determinístico é o mais conhecido. Porém, para que esse método apresente bons resultados, é necessário que o número de moléculas seja muito elevado, permitindo uma abordagem através de um ponto de vista contínuo, inferido a partir do uso de concentrações nas equações de velocidade.[172] Quando o número de moléculas é muito pequeno (da ordem de centenas), a aleatoriedade não pode ser ignorado, sendo necessário aplicação de métodos

estocásticos como as equações mestras (equivalentes à equação da Chapman-Kolmogorov, porém na forma diferencial).

No campo da química destacam-se o uso de equações mestras na modelagem de catálise enzimática,[176] em mecanismos do controle do relógio biológico de organismos,[177] na cinética de difusão de átomos em ligas metálicas fora do equilíbrio,[178] produtos de combustão em reações radicalares,[179,180] entre outros.

Embora a teoria das cadeias de Markov tenha mais de um século, a aplicação dos MSMs a DM foi introduzida em 1999 por Schutte e colaboradores.[181] Com o avanço computacional, essas ideias foram adotadas e aprimorados por alguns grupos de pesquisa na comunidade de DM em meados do ano 2000.[182] Desde então, os MSMs e técnicas relacionadas tem sido utilizados no estudo da cinética e termodinâmica de sistemas complexos como o enovelamento de proteínas, ligação proteína-ligante, dinâmica peptídica, agregação peptídica e alterações na conformação da proteína.[182]

3.6.5 Construção, validação e análise dos MSMs

Nos últimos anos, diversos grupos de pesquisa se empenharam no desenvolvimento e melhorias de métodos para construção, validação e análise de modelos cinéticos.[182] Destacamos a determinação de coordenadas coletivas e métricas adequadas,[182] o desenvolvimento de métodos eficientes: de clusterização,[182,183] para estimar taxa e matriz de transição,[182,183] de estimativa de erros estatísticos e sistemáticos,[182,184] para seleção de modelos,[185] de granulação grosseira (*coarse-graining methods*)[182] e de análise dos MSMs com a teoria do caminho de transição (TPT).[182]

Além disso, foram desenvolvidos métodos de estimativa de MSM para dados produzidos em estados termodinâmicos distintos (por exemplo, temperaturas e potenciais de polarização). Esses métodos proporcionaram um caminho para integrar simulações de amostragem aprimoradas e simulações MD diretas.[186–189] Vale a pena destacar que a construção, validação e análise de modelos cinéticos como os MSMs constituem-se de uma tarefa complexa, sendo necessário o uso de softwares confiáveis e eficientes. Atualmente os dois pacotes de softwares mais completos e eficientes para essa finalidade são o PyEMMA[182] e o MSMBuilder.[190]

Para construirmos um MSM, faz-se necessário aplicarmos em etapas um conjunto de redução de dimensionalidade do sistema. A seguir, apresentamos um conjunto básico de etapas para construção, validação e análise de um MSM:

1. Configure um sistema para dinâmica molecular e execute uma ou mais simulações pelo maior tempo

possível.

A DM é a etapa inicial para construção do MSM, sendo necessário uma DM longa ou várias simulações para confecção do modelo.

2. Adicione as coordenadas de entrada que deseja trabalhar.

Nessa etapa, escolhem-se os recursos (*features*) que se deseja aplicar no estudo do sistema. Para proteínas é mais comum utilizar os ângulos de torção do backbone (ϕ/ψ) e ângulos de torção da cadeia lateral (χ1,χ2, etc). Ressaltamos porém, que há inúmeros outros recursos como: distância entre os átomos, distância entre todos os Cα , rmsd mínimo em relação a uma estrutura de referência, centro de massa por resíduo, etc.

3. Reduza a dimensionalidade do sistema.

Os vetores de recursos selecionados na etapa anterior podem apresentar alta dimensionalidade. Por exemplo, há quase 5000 distâncias de Cα em uma pequena proteína de 100 resíduos. Tentar discretizar espaços com muitas dimensões através de métodos de clusterização não é eficiente, tendendo a produzir dados de baixa qualidade.[191] Uma abordagem comum bastante utilizada para redução de dimensões é a análise de componentes principais (PCA).[192,193] É conhecido que a PCA é ótima dentro dos métodos de transformações lineares em relação à maximização da variância retida. Porém, o objetivo não é necessariamente manter a variância, mas descrever a cinética molecular.

Desse modo, o interesse maior consiste na preservação dos movimentos lentos em detrimento aos de grande amplitude.[182] Como exemplo, considere um peptídeo não estruturado que possui extremidades bastante flexíveis, mas que passa por uma transição de torção de evento raro em seu centro. Como identificar a transição de evento raro ao invés de flutuações rápidas e de alta variância dos terminais?

Para esse fim, a análise de componentes independentes de estrutura do tempo (time-structure independent components analysis-tICA) tem apresentado ótimos resultados.[143,168,194] A tICA implementa a abordagem variacional da dinâmica de conformação[195] e apresenta excelentes resultados quando comparada com demais métodos lineares de recuperação das coordenadas de reação lenta e suas escalas de tempo de relaxamento.[182] Tanto o *software* PyEMMA quanto o MSMbuilder recomendam a tICA como método padrão para redução de

dimensionalidade. Além disso, sugerem o escalonamento das coordenadas tICA para obter um mapa cinético,[191] pois as coordenadas resultantes definem um espaço métrico no qual as distâncias geométricas são proporcionais às distâncias cinéticas, sendo desse modo preparadas otimamente para a próxima etapa que é a clusterização geométrica.

A seguir apresentamos um exemplo do programa MSMbuilder que compara a performance entre a PCA e a tICA.

Exemplo 3: Observe a seguinte equação de energia potencial:

$$E(x,y,z) = 5 \cdot (x-1)^2 \cdot (x+1)^2 + y^2 + z^2. \tag{3.29}$$

Ao analisarmos essa equação, é fácil perceber que ao longo da dimensão x, o potencial é um duplo poço e ao longo das dimensões y e z trata-se de um potencial harmônico. Logo devemos esperar que x seja a coordenada lenta, ao passo que o sistema deve-se equilibrar rapidamente ao longo de y e z. Após a realização de algumas dinâmicas e aplicando-se a redução de dimensionalidade PCA e tICA, obteve-se a seguinte figura:

Figura 3.7: PCA x tICA aplicado a uma função de energia potencial.

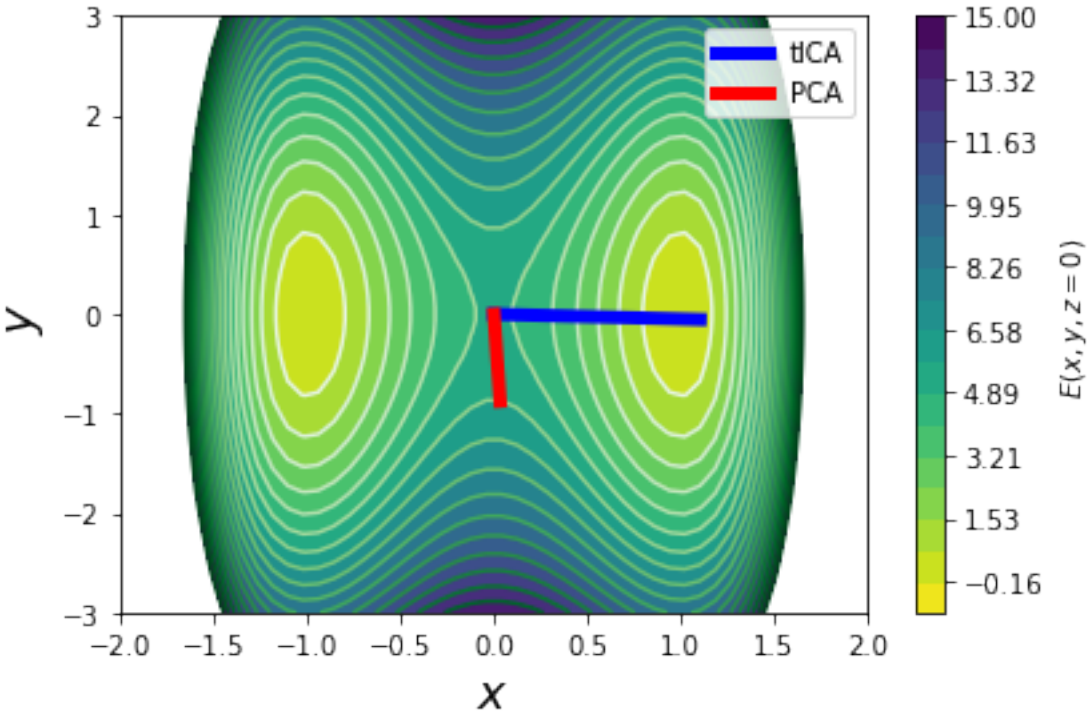


Figura gerada através do exemplo disponível em:
http://msmbuilder.org/3.8.0/examples/tICA-vs-PCA.html (acesso em 17/05/2022).

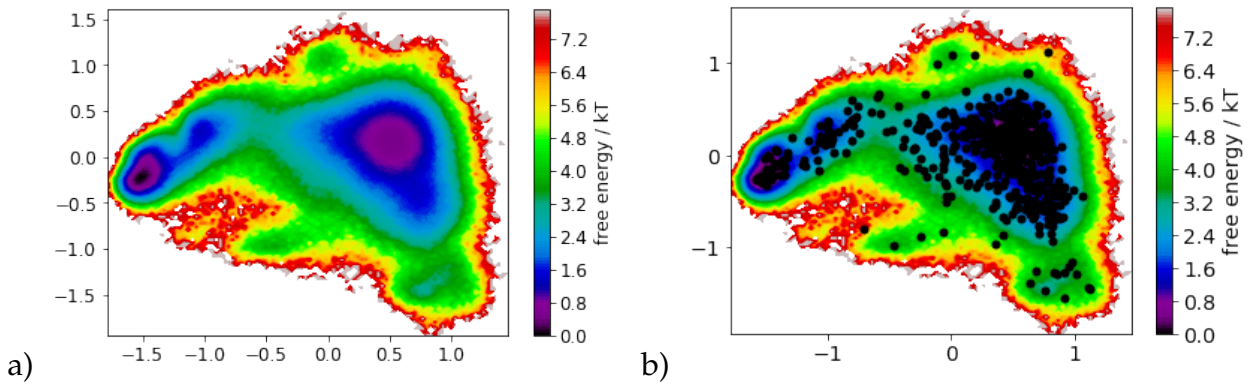
Note que o tICA encontra uma projeção capaz de resolver a coordenada lenta x ao passo que o PCA falha nessa tarefa.

4. Discretize os estados em microestados.

Diversos modelos cinéticos, como os MSMs são construídos em espaços de estados discretos, sendo necessário a discretização dos dados. Pode acontecer que mesmo após a redução de dimensionalidade tICA, o sistema ainda permaneça com muitas dimensões, principalmente quando trabalhamos com proteínas. Uma discretização que pode ser utilizada em um espaço multidimensional é a discretização de Voronoi,[196] onde um conjunto de centros k é determinado pelo método de clusterização, pegando-se a estrutura representativa mais próxima ao centroide.

Apesar de haver outras métricas, a distância euclidiana é a mais utilizada para medir as distâncias entre as estruturas. Se os mapas cinéticos baseados na tICA forem utilizados, essas distâncias aproximam-se das distâncias cinéticas.[182] Há na literatura vários métodos de clusterização com performances distintas para diversos sistemas estudados,[197] porém o método recomendado pelo programa PyEMMA é o *k-means*[171,197] em combinação com o procedimento de inicialização *k-means* ++.[198] Essa abordagem apresenta a vantagem de ter um baixo custo computacional, gerando resultados reprodutíveis com apenas algumas iterações *k-means* (< 10). A Figura 3.8 a seguir apresenta a combinação do mapa cinético baseado na tICA com o método de clusterização *k-means* (n = 550 *clusters*) para a proteína BBA (pdb: 1FME).

Figura 3.8: Em (a) mapa cinético baseado na tICA e em (b) centróide dos clusters com o método *k-means* e mapa cinético de fundo para a proteína BBA (pdb: 1FME).



5. Estime um modelo de MSM a partir dos dados clusterizados.

Nesse ponto é gerado o MSM a partir do espaço de estados discretos $S(t)$ pulando entre os n microestados (n é igual ao número de centroides obtidos na etapa de clusterização). O MSM prediz a cinética em escalas de tempo maiores em termos das potências da matriz de transição. (propriedades das cadeias de Markov descritas na seção 2.5.2) Em teoria, estimar a matriz de transição requer uma tarefa simples de contagem. A primeira etapa consiste em atribuir dados aos *clusters* que serão numerados de 0 a $(n-1)$.

Desse modo, cada trajetória pode ser considerada como uma série de atribuições em microestados ao invés de uma série de conformações. Sendo assim, o número de transições entre cada par de estados pode ser contado e armazenado em uma matriz de contagem de transição (C) em que C_{ij} é o número de transições do estado i para o estado j.

Com um conjunto de dados infinitos, pode-se usar a estimativa de máxima verossimilhança para as probabilidades de transição entre cada par de estados de modo a converter a matriz de contagem de transição (C) em uma matriz de probabilidade de transição (P), ou seja:

$$P_{ij}(\tau) = \frac{C_{ij}}{\sum_k C_{ik}}; \tag{3.30}$$

onde τ é o tempo de atraso (*lag time*).

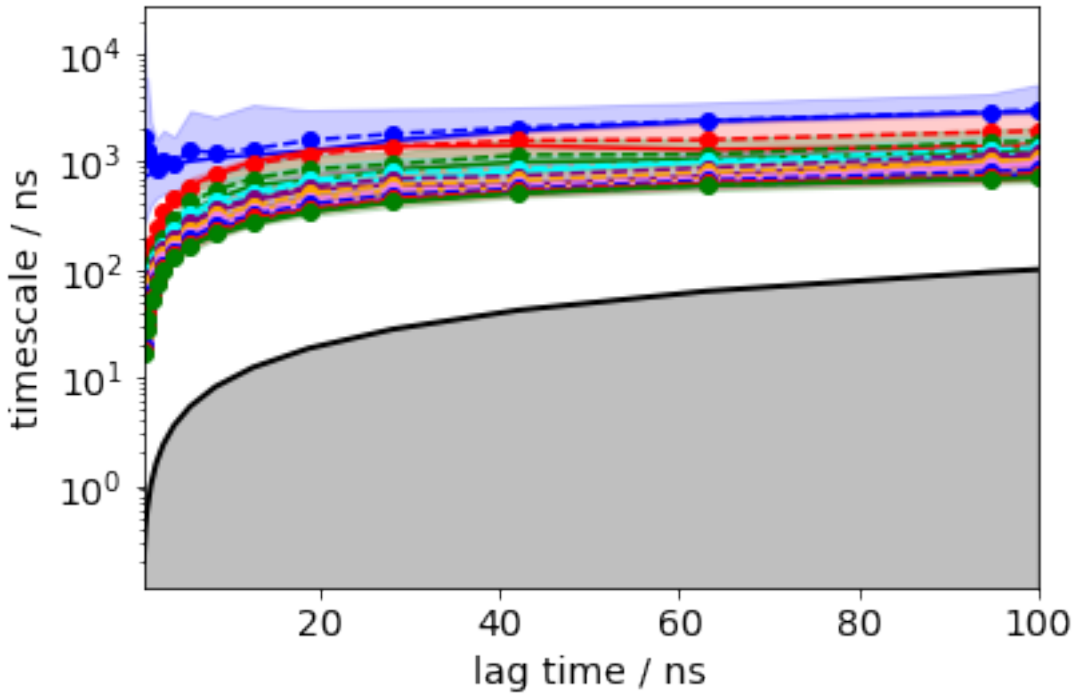
Ressaltamos, porém, que na prática a estimativa de matrizes de transição torna-se mais complicada devido a um conjunto de fatores como amostragem finita e imperfeições nas definições de macroestados. Além dos estimadores de máxima verossimilhança, também são utilizados estimadores bayesianos para construção da matriz de transição de um MSM.

Um parâmetro extremamente importante para a acurácia de um MSM é o *lag time*. Esse deve ser escolhido de forma que as escalas de tempo de relaxação sejam constantes dentro do erro estatístico.[182] Os tempos de relaxação de um modelo são uma função de autovalores de sua matriz de probabilidade de transição e é dado pela equação (3.31) a seguir:

$$t_i = \frac{\tau}{ln\,|\lambda_i(\tau)|}; \tag{3.31}$$

onde t_i é o tempo de relaxação, τ é o *lag time* e $\lambda_i(\tau)$ é o i-ésimo maior autovalor do MSM estimado em τ. A Figura 3.9 a seguir apresenta estimativas de escala de tempo em nanossegundos para a proteína BBA (pdb: 1FME)

Figura 3.9: Escalas de tempo de relaxação implícitas para a proteína BBA (pdb: 1FME). Os intervalos de confiança são representados pelas áreas sombreadas que contém 95% das amostras geradas pelo método MSM bayesiano.



Observe que a partir de 20 *ns* as regiões sombreadas que representam 95% do intervalo de confiança estão constantes, sendo que são resolvidos os 4 processos lentos. A região em cinza representa processos que são mais rápidos que *lag time* escolhido. Logo se houvesse algum processo nessa região, eles não seriam resolvidos.

6. Valide o seu MSM.

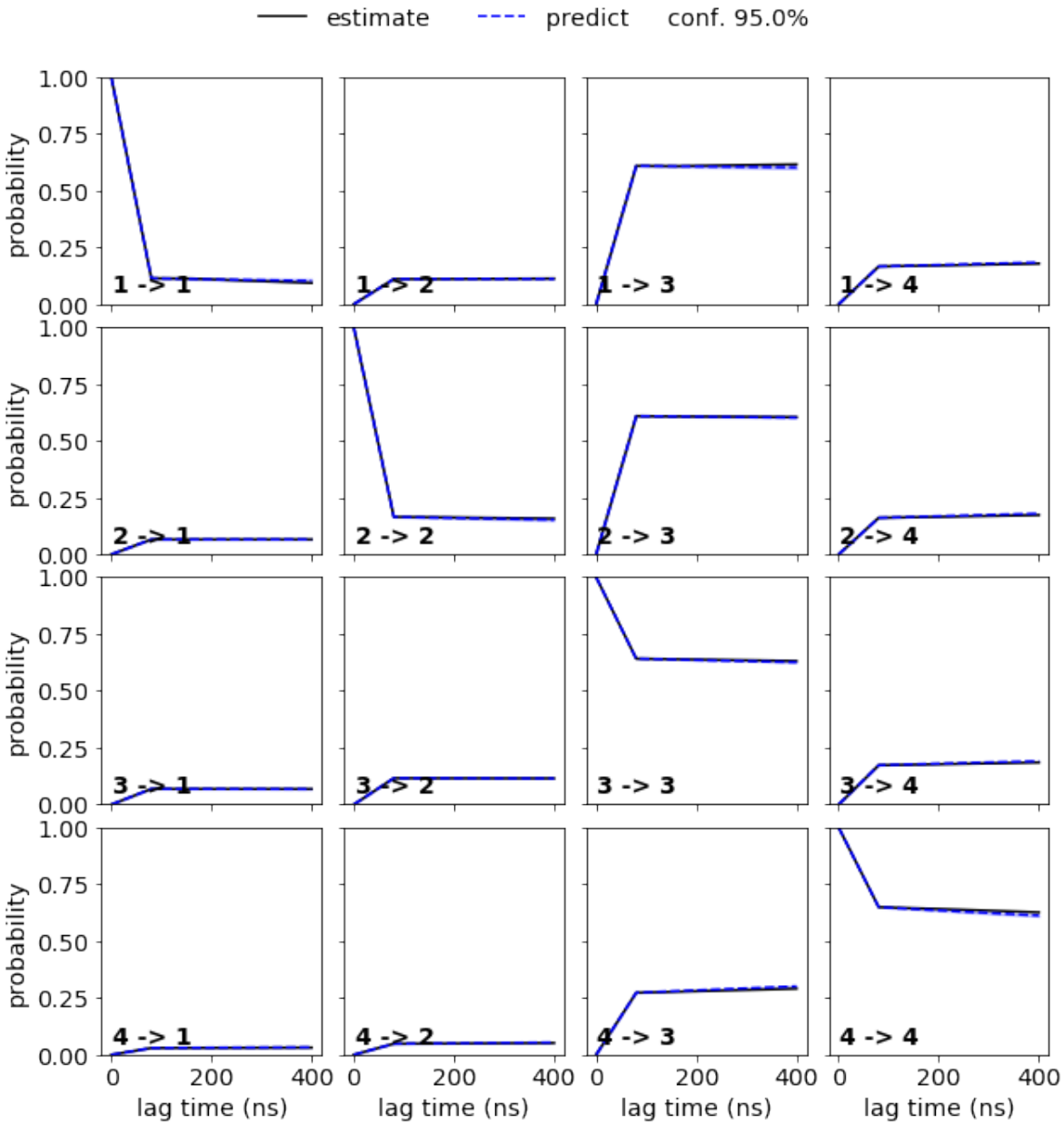
A validação do modelo de Markov é realizado segundo a equação de Chapman-Kolmogorov descrita em mais detalhes na seção 3.5.3. Essa equação descrita em função do *lag time* (τ) fica da seguinte forma:

$$P(n\tau) = P(\tau)^n; \tag{3.32}$$

onde n é um número inteiro de passos, sendo cada um com um tempo de atraso τ de comprimento. Recapitulando o que foi mencionado na seção 3.5.3, essa equação captura o fato

de que tomar n passos com um MSM com um tempo de atraso de τ deve ser equivalente a um MSM com um *lag time* $n\tau$. Na Figura 3.10 a seguir é apresentado um teste de passagem para um modelo de 4 estados para a proteína BBA(pdb: 1FME).

Figura 3.10: Teste de Chapman-Kolmogorov para um MSM de 4 estados para a proteína BBA.

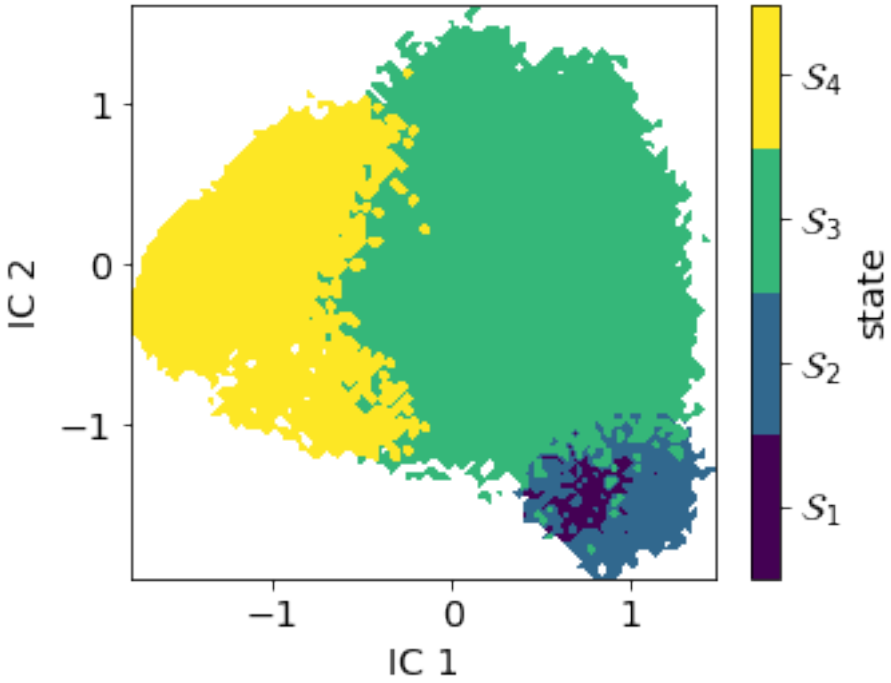


Observe que o valor estimado (linha preta) e o valor predito (linha azul) estão sobrepostos, mostrando que os caminhos estimados e preditos coincidem, ou seja, os dois lados da equação 3.32 são equivalentes para um intervalo de confiança de 95% desse conjunto de dados. Logo o MSM passou no teste para o *lag time* escolhido.

7. Selecione um conjunto de macroestados, um granulado groseiro (coarse-graining) e um caminho de transição (TPT) para o seu MSM.

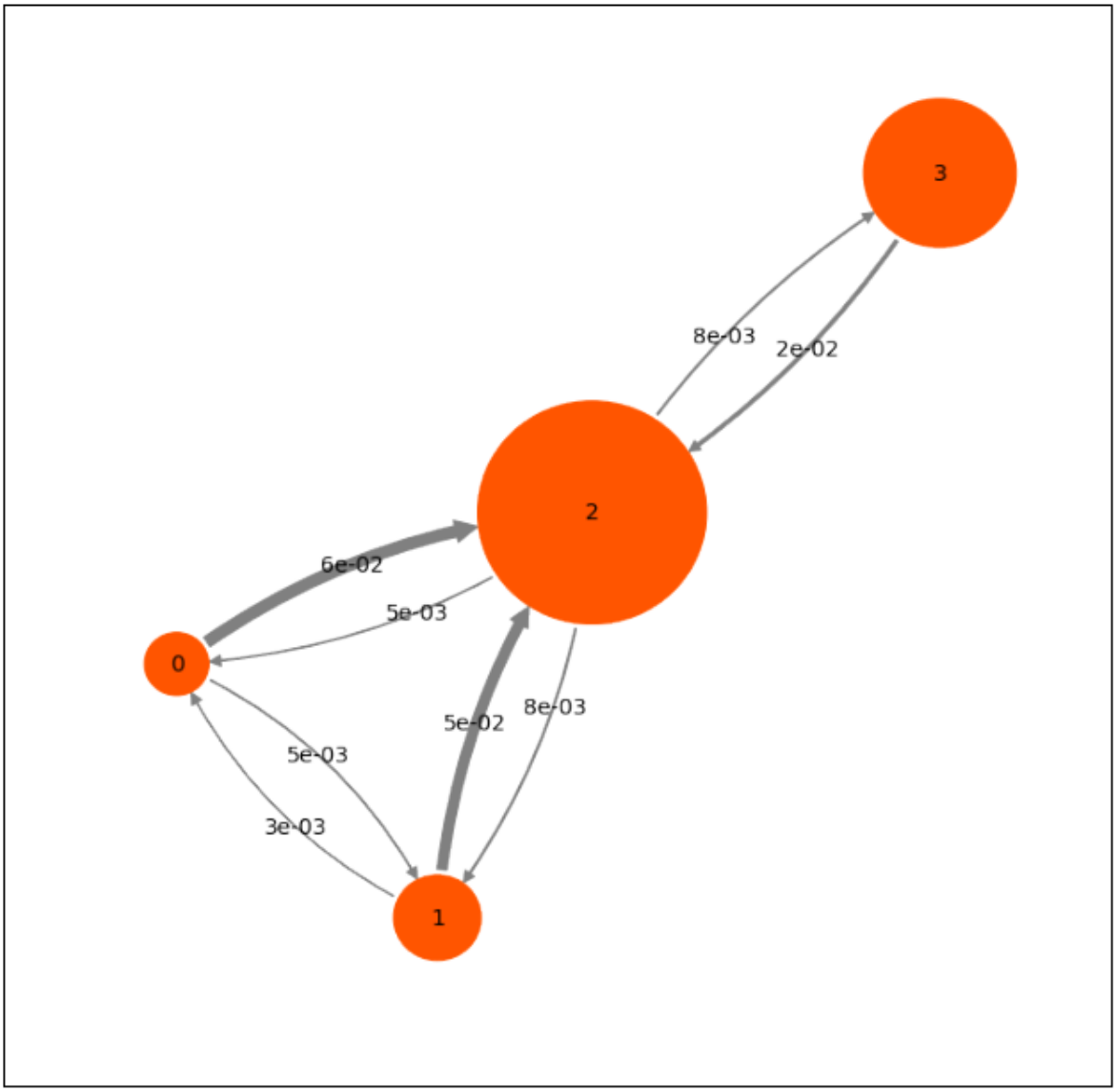
Quando estamos interessados em analisar caminhos, modelos com menos estados são mais desejáveis, pois são mais fáceis de entender. Um método utilizado para esse fim é a versão mais robusta do método de *Perron Cluster Cluster Analysis*, denominada (PCCA++).[199] O método PCCA++ calcula os chamados membros, ou seja, verifica a probabilidade de cada microestado pertencer a um determinado macroestado. Os macroestados correspondem às bacias da paisagem de energia livre do sistema. Na Figura 3.11 a seguir, é apresentado a separação dos microestados da proteína BBA em 4 bacias (macroestados) na projeção da paisagem de energia livre baseada nas duas primeiras coordenadas tICA.

Figura 3.11: Modelo de 4 estados metaestáveis obtidos para a proteína BBA através do método PCCA++.



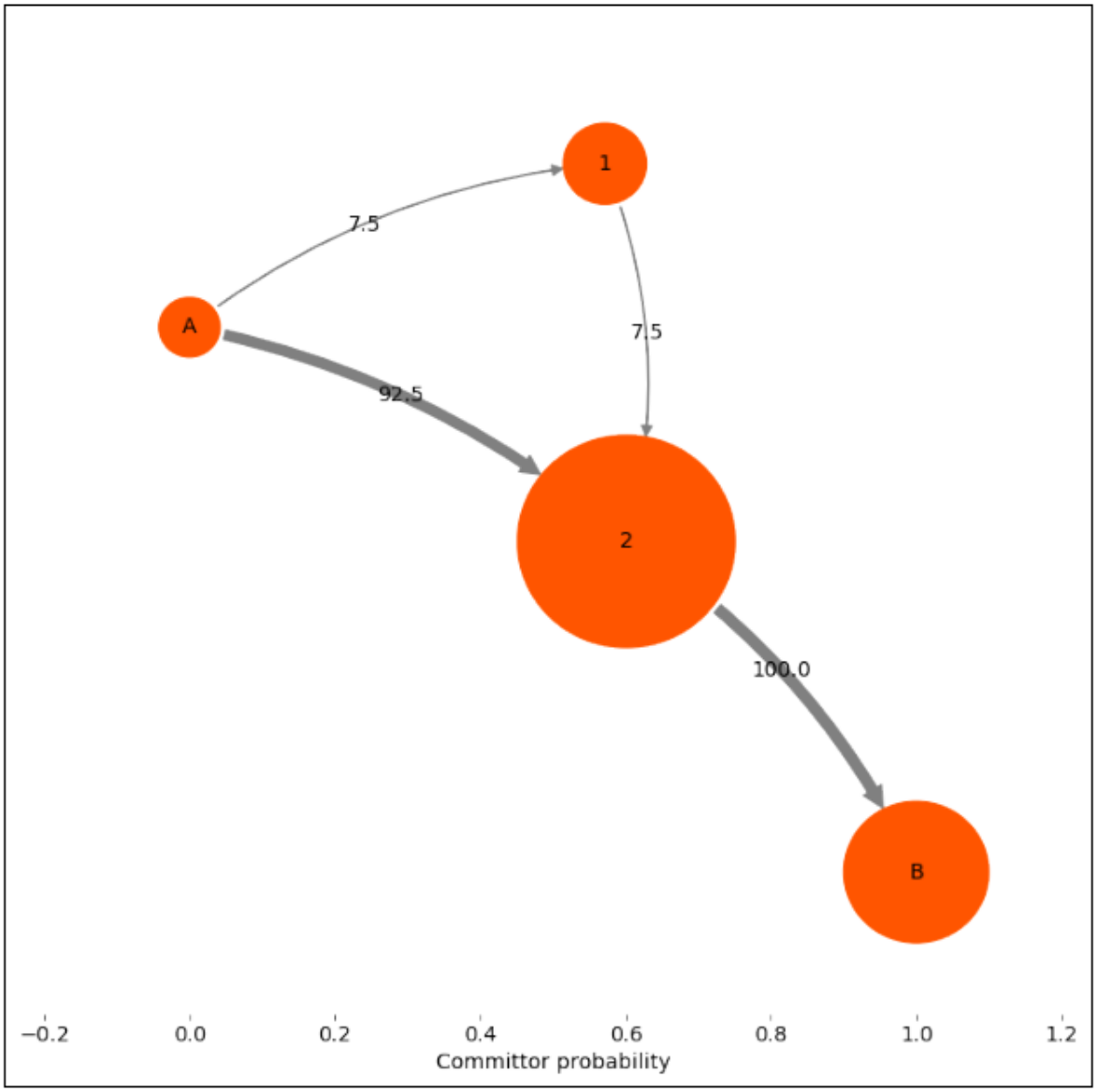
Nesse momento é possível amostrar um conjunto de estruturas representativas de cada uma das bacias de paisagem de energia livre. Para se obter um modelo cinético grosseiro entre os quatro estados metaestáveis, utiliza-se a *coarse-graining* baseada nos modelos ocultos de Markov (HMM).[200] Através dessa técnica obtém-se a matriz de transição entre os estados metaestáveis. De possa da matriz de transição, pode-se plotar os grafos com as probabilidades de transição entre os estados. Na Figura 3.12 é apresentado o grafo com as probabilidades de transição entre os 4 estados metaestáveis para a proteína BBA.

Figura 3.12: Grafo com a probabilidade de transição entre os quatro estados metaestáveis para a proteína BBA através do método *coarse-graining*.



Pode acontecer, porém de estarmos interessados não na transição entre todos os macroestados, mas sim no fluxo entre um macroestado A de origem para um sumidouro ou produto B. Dessa maneira, poderíamos extrair informações sobre o mecanismo estrutural e a cinética de transição de $A \rightarrow B$. Para obtermos essas informações fazemos o uso da Teoria do Caminho de Transição (TPT).[182] A TPT tem apresentado boa performance no estudo de processos como o enovelamento de proteínas[140,201,202] e a ligação proteína-ligante.[203,204] Na Figura 3.13 a seguir apresentamos o fluxo total normalizado do primeiro estado (0) para o quarto estado (3) através da TPT pra a proteínas BBA.

Figura 3.13: Caminho do estado de transição entre o estado desenovelado e nativo para a proteína BBA.



Nesse caso a proteína sai do estado **A** de maior energia (desenovelado) para o estado **B** de menor energia (estrutura nativa) com probabilidades de transição entre os possíveis caminhos de $A \rightarrow B$.

Informações adicionais acerca do formalismo e dos procedimentos para construção, validação e análise de MSMs podem ser consultadas na referência.[182]

CAPÍTULO 4

PROCEDIMENTOS COMPUTACIONAIS

4.1 Conjunto de Procedimentos Aplicados à Interação Proteína-Ligante

Inicialmente, recuperamos as seguintes estruturas dos complexos RTA-ligante no Protein Data Bank (PDB): 4HUP (RTA-19M),[205] 4HUO (RTA-RS8),[205] 4ESI (RTA-0RB),[206] 4MX1 (RTA-1MX),[207] 3PX8 (RTA-JP2),[208] e 3PX9 (RTA-JP3).[208] A Figura 4.1 apresenta as estruturas químicas dos seis ligantes do RTA estudados neste trabalho.

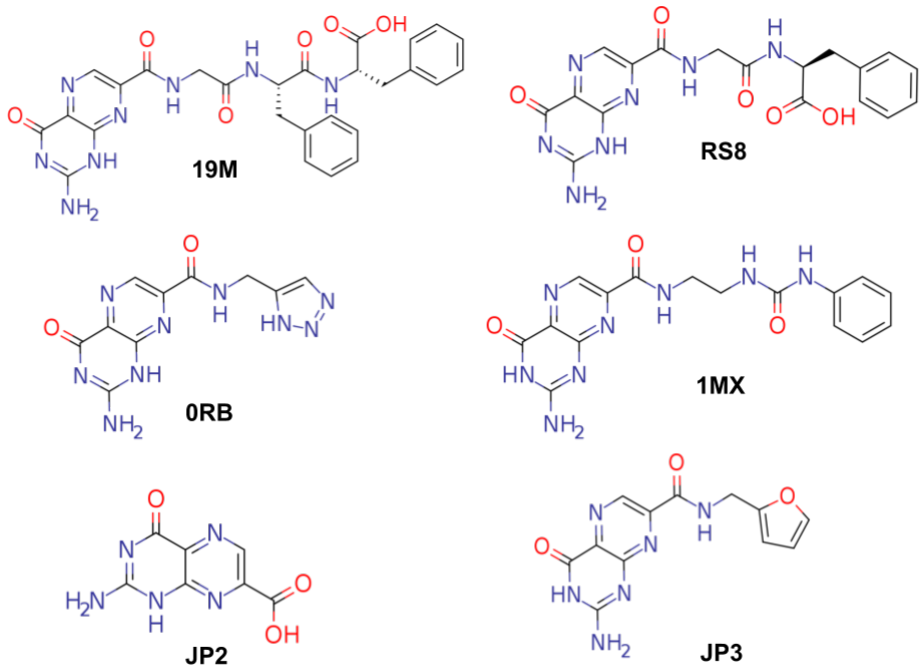


Figura 4.1: Estruturas dos seis ligantes da RTA estudados neste trabalho. O código de três letras corresponde aos códigos PDB ID para esses ligantes.

Todos os complexos RTA-ligante foram relaxados e, em seguida, equilibrados usando simulações de dinâmica molecular. Definimos as simulações de DM de acordo com as configurações usadas na referência.[10] O último *frame* de cada simulação de DM foi usado para calcular entalpias de ligação usando métodos de mecânica quântica semiempírica (SQM), considerando dois cenários: (i) um cálculo *single-point* direto de energia com os métodos: RM1,[209] PM6,[210] PM6-DH+,[211] PM6-D3H4 e PM7[212] e (ii) após a otimização da geometria *full atom* com os métodos PM6-DH+, PM6-D3H4 e PM7.

Para todos os cálculos semiempíricos, usamos o algoritmo de escalonamento linear MOZYME,[213] disponível no pacote MOPAC2016.[214] Para os critérios de convergência SCF, usamos as configurações padrão e um raio de corte de 10 Å para o algoritmo MOZYME. Para as otimizações de geometria *full atom*, usamos o algoritmo BFGS, considerando uma norma de gradiente de 5,0 $kcal \cdot mol^{-1} \cdot Å^{-1}$ como critério de parada para a RTA e os complexos RTA-Ligante e 0,01 $kcal \cdot mol^{-1} \cdot Å^{-1}$ para os ligantes livres, onde nenhuma restrição de movimento foi considerada para os átomos. Para todos os cálculos (cálculos *single-point* e de otimização de geometria), consideramos o modelo de solvatação implícita (COSMO) com uma permissividade relativa de 78,4 e um raio efetivo da molécula de solvente 1,3 Å. As entalpias de ligação para os complexos RTA-ligante foram calculadas de acordo com a Equação 4.1 . Para todas as proteínas não complexadas e complexos R-L, assumimos uma carga total de +2e.

$$\Delta H_{binding} = \Delta H_f^{ligand-RTA_{complex}} - \left(\Delta H_f^{ligand} + \Delta H_f^{RTA}\right). \tag{4.1}$$

A RTA tem 4200 átomos e os ligantes 19M, RS8, 0RB, 1MX, JP2 e JP3 possuem 67, 47, 30, 43, 20 e 31 átomos, respectivamente. A soma entre o número de átomos da RTA e cada ligante resulta no número de átomos para o respectivo complexo RTA-ligante.

Para fins de comparação e para testar o desempenho dos métodos semiempíricos, também realizamos cálculos *single-point* considerando o método híbrido QM/MM ONIOM,[215,216] usando as mesmas geometrias da DM para os seis complexos RTA-ligante. Usamos o programa Gaussian 09[217] com o funcional híbrido GGA B3LYP[218,219] e o funcional híbrido GGA ωB97X-D,[220] que inclui correções de dispersão DFT-D2.[221] Em ambos os casos, usamos o conjunto de bases 6-31+G(d)[222–224] para a região QM da proteína RTA e dos complexos RTA-ligante. O campo de força universal UFF[225] foi usado para o restante da proteína, que é denotado como a parte MM. Além disso, átomos de hidrogênio foram usados como *link atoms*, conectando essas duas partes. Para cálculos QM/MM ONIOM, foram utilizados os critérios padrão do programa Gaussian 09.[217]

A parte QM da proteína RTA inclui os seguintes resíduos: Glu-177, Arg-180 (importantes

para a catálise enzimática), Tyr-80, Val-81, Gly-121 e Tyr-123 (importantes para o atracamento e reconhecimento do ligante no sítio ativo).[226] Além disso, incluímos os resíduos Asn-122, Ser-176, Asn-209, Gly-212 e Arg-213 porque eles estão a cerca de 3,0 Åde distância de qualquer um dos seis ligantes. Desse modo, foi produzido um modelo QM para a RTA com 189 átomos, 716 elétrons e carga +1. Para os complexos RTA-ligante, também incluímos os respectivos ligantes na parte QM. Portanto, os modelos foram gerados com (256 átomos, 1008 elétrons e carga +1), (236 átomos, 930 elétrons e carga +1), (219 átomos, 864 elétrons e carga +1), (232 átomos, 908 elétrons e + 1 carga), (209 átomos, 822 elétrons e carga +1) e (220 átomos, 864 elétrons e carga +1) para os complexos RTA-19M, RTA-RS8, RTA-0RB, RTA-1MX, RTA-JP2 e RTA-JP3, respectivamente. As energias de ligação para os complexos RTA-ligante foram calculadas de acordo com a Equação 4.2.

$$\Delta E_{ligacão} = E_{QM/MM}^{Complexo_{RTA-ligante}} - \left(E_{QM}^{ligante} + E_{QM/MM}^{RTA}\right). \tag{4.2}$$

Na Figura 4.2, apresentamos a geometria do complexo RTA-19M com ênfase nos resíduos do sítio ativo e no ligante (em amarelo).

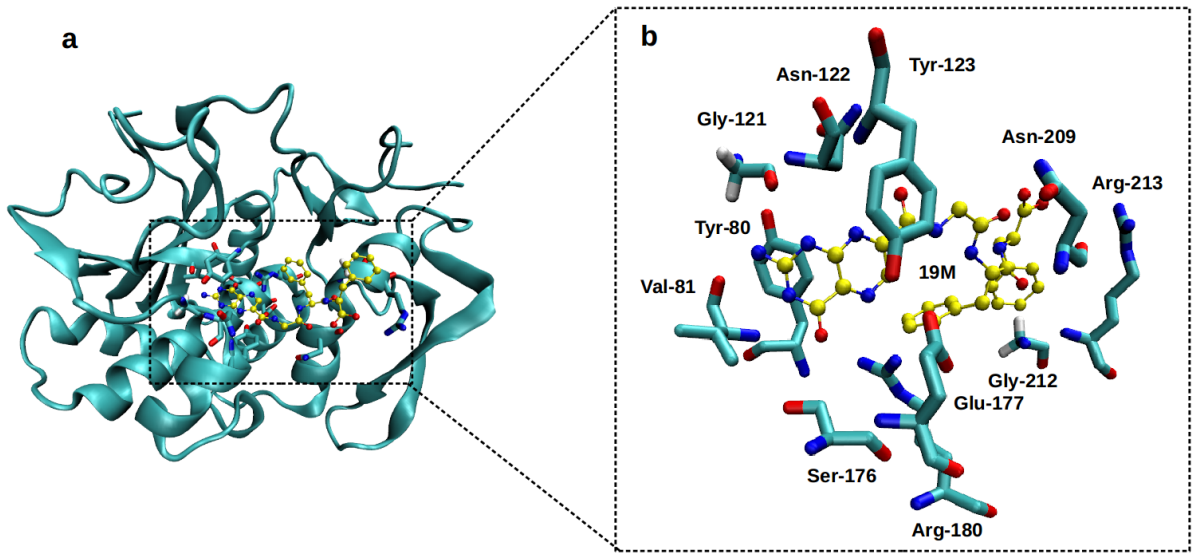


Figura 4.2: (a) Geometria do complexo RTA-19M usado no cálculo QM/MM ONIOM. (b) Vista ampliada do sítio ativo mostrando os 11 resíduos e o ligante 19M (em amarelo).

Na Figura 4.3, apresentamos os modelos usados na parte QM para todos os complexos. Consideramos todos os átomos dos resíduos Tyr-80, Val-81, Gly-121, Asn-122, Tyr-123, Asn-209, Gly-212 e Arg-213 porque para esses resíduos há interações entre os átomos do *backbone* da RTA com os ligantes. Para os resíduos Ser-176, Glu-177 e Arg-180, consideramos apenas as cadeias laterais, uma vez que as interações intermoleculares com os ligantes ocorrem apenas

nesta região dos aminoácidos.

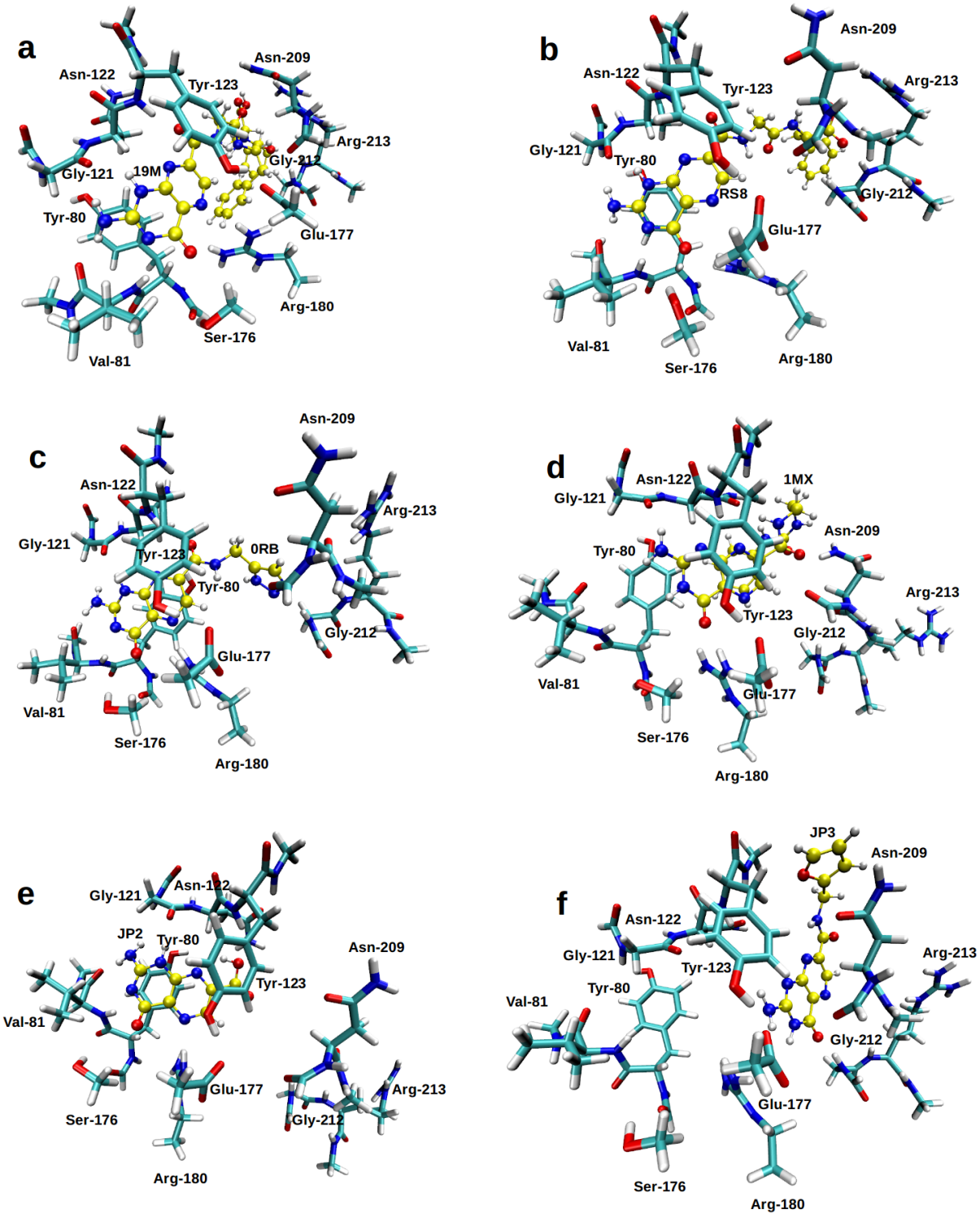


Figura 4.3: Modelos QM para cálculos QM/MM ONIOM para complexos: (a) RTA-19M, (b) RTA-RS8, (c) RTA-0RB, (d) RTA-1MX, (e) RTA-JP2 e (f) RTA-JP3.

4.1.1 Cálculos dos Descritores de Reatividade

Os descritores quânticos mais bem-sucedidos vêm da teoria do funcional da densidade conceitual (CDFT), que, por meio do desenvolvimento matemático da teoria do funcional da densidade, deu definições quantitativas válidas para conceitos químicos bem conhecidos,[227] como dureza e moleza. A partir dessa teoria, as principais interações entre duas moléculas são resumidas em dois tipos de processos: polarização mútua seguida de interações de sobreposição de orbital molecular e forças coulombianas.[228] A propensão desses efeitos nas moléculas pode ser rastreada localmente usando a função de Fukui para o primeiro tipo de interação e a dureza local para o segundo.[229]

Para representar localmente as interações de sobreposição de orbitais moleculares, usamos a função de Fukui.[229] Este descritor foi dividido em duas funções, a saber, a função de Fukui esquerda (f^-), específica para suscetibilidade ao ataque eletrofílico (que para a aproximação orbital congelada é igual à densidade do orbital molecular ocupado de mais alta energia (*HOMO*),[228] e a função Fukui direita (f^+), específica para suscetibilidade ao ataque nucleofílico (que é igual à densidade do orbital molecular desocupada de mais baixa energia (*LUMO*).

Uma representação conveniente das funções Fukui é onde os valores são atribuídos para cada centro de átomo k na molécula. O f^- definido na Equação 4.3 é a soma dos coeficientes ν dos orbitais atômicos quadrados (AO) que compõem o HOMO e pertencem ao $k-simo$ átomo, mais a soma do produto entre os coeficientes dos índices ν, μ e os correspondentes aos elementos da matriz de *overlap* $S_{\mu\nu}$.[230] Uma definição equivalente para o f^+ é apresentada na Equação 4.4, usando o *LUMO* em vez do *HOMO*.

$$f^-(k) = \sum_{\nu \in k}^{AO} |C_{\nu HOMO}|^2 + \sum_{\mu \notin \nu}^{AO} |C_{\nu HOMO} C_{\mu HOMO}| S_{\mu\nu}; \tag{4.3}$$

$$f^+(k) = \sum_{\nu \in k}^{AO} |C_{\nu LUMO}|^2 + \sum_{\mu \notin \nu}^{AO} |C_{\nu LUMO} C_{\mu LUMO}| S_{\mu\nu}. \tag{4.4}$$

Neste estudo, os efeitos gerais computados nessas duas funções de Fukui foram usados em combinação, através da função de Fukui média, definida na Equação 4.5 .

$$f^0 = \frac{f^+(k) + f^-(k)}{2}. \tag{4.5}$$

Esses descritores químicos quânticos foram identificados como úteis para determinar a toxicidade e a atividade biológica de vários ligantes potenciais.[107,231] Para as enzimas, esses

descritores têm sido usados para determinar sítios de reatividade, como funções de pontuação em *docking molecular* e para pesquisar estruturas nativas de proteínas[232] e encontrar estruturas-chave em simulações de caminhos de reação.[108,233]

A estrutura eletrônica de sistemas protéicos apresenta diversos orbitais moleculares relevantes para interações químicas com a energia eletrônica próxima ao HOMO e ao LUMO.[107,108] Assim, neste estudo, calculamos as funções de Fukui não apenas usando os orbitais HOMO e LUMO, mas também considerando todos os orbitais moleculares de uma faixa de 3eV do HOMO e LUMO. Como Fukushima e colaboradores mostraram em seu trabalho, esse valor pode variar de 1 a 5 eV dependendo do sistema em estudo.[234]

Usamos a dureza local ($H(k)$) como um descritor quântico para calcular as interações da força Coulômbica. Esta definição particular é baseada na contribuição elétron-elétron do potencial eletrostático molecular, como mostrado na Equação 4.6.[235] O cálculo da dureza local no $k-simo$ centro atômico é definido como a soma da função de Fukui à esquerda para cada $l-simo$ centro atômico dividido por sua distância euclidiana R_{kl}. Essas grandezas teóricas foram calculadas usando o software PRIMoRDiA.[106]

$$H(k) = \sum_{l \neq k}^{átomos} \frac{f^{-}(l)}{R_{kl}}. \tag{4.6}$$

4.2 Conjunto de Procedimentos Aplicados ao Estudo do Enovelamento de Proteínas

Inicialmente realizamos o alinhamento das estruturas da DM de três proteínas (NTL9, BBA e α3D), obtidas da referência,[112] fornecidas pelo pesquisador D. E. Shaw, com o programa MDTraj.[236] Ressaltamos que o fato de termos utilizados trajetórias prontas ao invés de gerarmos as nossas próprias trajetórias se deu devido a dois fatores:

1. *Não disponibilizamos de recursos computacionais adequados para rodarmos DM de proteínas com tempo na faixa de μs.*

2. *Há diversos artigos recentes de alto impacto[182,190,237–240] que utilizaram essas ou outras trajetórias de enovelamento geradas pelo grupo do pesquisador D. E. Shaw com o auxílio do supercomputador Anton,[241] que foi construído exclusivamente para estudos de DM.*

Desse modo, como essas trajetórias já apontaram alta performance, não faria sentido gerar trajetórias menores e com qualidade inferior para o nosso estudo. As trajetórias das proteínas NTL9, BBA e α3D apresentam tamanhos de 377 μs, 223 μs e 346 μs respectivamente,

tempo compatível com dados experimentais para esse conjunto de proteínas de rápido enovelamento.[112] Após o alinhamento das trajetórias realizamos dois procedimentos para a caracterização do caminho de enovelamento: na primeira abordagem, utilizamos o programa PyEMMA e na segunda, o programa MSMbuilder. Esses dois programas apresentam ferramentas semelhantes para redução de dimensionalidade do sistema e construção de MSMs.

4.2.1 Abordagem utilizando o programa PyEMMA

Na abordagem utilizando o PyEMMA, de posse das trajetórias alinhadas na etapa anterior, apresentamos essas em um vetor apropriado de recursos. Foram escolhidos os seguintes recursos (*features*) para a composição do nosso sistema: todos os ângulos de torção do *backbone* da proteína, ângulos $\chi 1$ das cadeias laterais e o RMSD mínimo entre os *frames* da trajetória. Em seguida, utilizamos a o método tICA (*time-structure based Independent Components Analysis*) para redução de dimensionalidade do sistema estudado e geração do mapa cinético. O tempo de retardo (*lag time*) utilizado para a redução de dimensionalidade foi de 10ns, 0,5μs e 10 ns para as proteínas NTL9, BBA e α3D respectivamente. Os valores de *lag time* utilizados apresentam a mesma ordem de magnitude dos dados de *lag time* utilizados em trabalhos anteriores para esse mesmo conjunto de proteínas.[140,237] Após essa etapa, os dados foram agrupados com o método *k-means* em combinação com o procedimento de inicialização *k-means* ++. Nessa etapa, o centroide de cada *cluster* corresponde a um microestado do nosso sistema. Em seguida foi estimado um MSM bayesiano a partir dos dados agrupados (*clusters*), sendo que esse foi validado pelo teste de Chapman-Kolmogorov.

Após esse procedimento, utilizamos o método PCCA++ para agruparmos os microestados em macroestados correspondentes. De cada macroestado, foram amostradas 100 conformações para o tratamento quântico posterior. A partir dos dados dos macroestados, aplicamos a técnica *coarse-graining* para obtenção da matriz de transição e o grafo de probabilidade de transição entre os estados metaestáveis. Em seguida, selecionamos o macroestado desenovelado (reagentes) e o enovelado (produtos) e aplicamos a TPT para encontrarmos os possíveis caminhos para o enovelamento e suas respectivas probabilidades de transição.

Após esse procedimento, realizamos cálculos *single-point* das estruturas amostradas com o PCCA++ via MOPAC2016[242] com o método PM7,[212] algoritmo de escalonamento linear MOZYME[213] e modelo de solvatação implícita COSMO (EPS=78.4, RSOL=1.3). Foram realizados cálculos de 500 conformações para as proteínas NTL9 e α3D e de 400 estruturas para a BBA, totalizando 1400 cálculos *single-point* de energia. A partir dos cálculos *single-point* foram extraídos a energia total (E_{TOT}) e o calor de formação (ΔH_f). Além disso, foram calculados

os seguintes QMCDs globais com o programa PRIMORDIA:[106] potencial de ionização (PI), afinidade eletrônica (AE), potencial químico (μ), dureza química (η), moleza química (S), eletrofilicidade (ω) e o número máximo de elétrons (n_{MAX}). Além disso realizamos os cálculos da fração de contatos nativos (Q) e o RMSD-Cα de todas as conformações em relação à estrutura cristalográfica.

4.2.2 Abordagem utilizando o progrma MSMBuilder

Na abordagem utilizando o programa MSMBuilder, foram escolhidos como *features*, os ângulos de diedros ψ e ϕ do *backbone* da proteína. Aqui não utilizamos ângulos das cadeias laterais, uma vez que queremos ter o mínimo de dimensões possíveis. Após essa etapa, aplicamos a análise tICA para a redução da dimensionalidade e projetamos o mapa cinético das duas primeiras coordenadas tICA. Como observamos que o eixo x apresentou ser uma boa coordenada de enovelamento/desenovelamento para as proteínas NTL9, BBA e α3D, realizamos a amostragem de 100 conformações para cada proteína ao longo desse eixo para o tratamento quântico.

Para esse conjunto de estruturas, estendemos a nossa abordagem e realizamos cálculos *single-point* via DFT-D3 e método semiempírico PM7.

Os cálculos *single-point* DFT-D3 de energia para as estruturas amostradas da tICA foram realizados com o programa TeraChem[243,244] com o funcional híbrido GGA B3LYP,[218,219] considerando correções de dispersão D3[245] e modelo de solvatação PCM[246] (EPS = 78,39). Os cálculos semiempíricos seguiram o mesmo protocolo utilizado na abordagem com o programa PyEMMA. Desse modo, foram realizados 300 cálculos *single-point* via DFT e 300 cálculos *single-point* via PM7, totalizando 600 cálculos quânticos.

A partir dos cálculos *single-point* de energia dos *frames* selecionados na etapa anterior, calculamos dois conjuntos de QCMDs: o global e o local por resíduo. OS QCMDs globais foram os mesmos obtidos na abordagem com o programa PyEMMA, sendo que a energia total (E_{TOT}) e o calor de formação (ΔH_f) foram obtidos via método PM7 e os demais descritores globais foram obtidos tanto via PM7 quanto via DFT-D3. Além disso, com o programa PRIMoRDIA,[106] os seguintes QMCDs locais por resíduo foram obtidos (via DFT-D3 e PM7): densidade eletrônica local, dureza local (*local hardness*) obtida por 4 métodos distintos implementados no PRIMoRDIA,[106] índice de localização do orbital, eletrofilicidade, suscetibilidade ao ataque eletrofílico (EAS), suscetibilidade ao ataque nucleofílico (NAS) e suscetibilidade ao ataque radicalar (RAS). Detalhes adicionais sobre como calcular os descritores de reatividade e o formalismo matemático podem ser encontrados nas seguintes referências.[106–108,233]

A dureza local calculada a partir da interação elétron-elétron da equação do potencial eletrostático molecular é a mais bem sucedida em encontrar padrões nas trajetórias de enovelamento entre as várias alternativas de dureza locais disponíveis no software PRIMoRDiA.[106] Esta definição de dureza local em particular não se integra à quantidade global e também não é o inverso direto da moleza local.

Usando coordenadas de enovelamento/desenovelamento, também calculamos a fração de contatos nativos (Q) e o RMSD-C_α de todas as conformações, tomando a estrutura cristalográfica como referência. Finalmente, realizamos os cálculos DSSP, RMSF e RMSD por resíduo de aminoácido para comparação com os resultados dos descritores de reatividade local.

Os QCMDs locais são geralmente obtidos na representação condensada por átomo, isto é, onde os valores das quantidades locais são atribuídos individualmente a cada átomo com a coordenada do núcleo como a posição de referência. No software PRIMoRDiA[106] há também a representação de resíduos condensados, onde os valores atribuídos aos átomos são somados para cada resíduo de aminoácido, o que melhora a interpretação no contexto de polímeros biológicos e torna os modelos estatísticos mais simples.

Na figura 4.4 a seguir é apresentado um fluxograma com as etapas dessa abordagem. As etapas podem ser resumidas da seguinte modo: i) Amostragem dos dados brutos para obter estruturas não correlacionadas com o tempo; ii) Para cada estrutura da amostra, foram realizados cálculos de química quântica de ponto único (DFT e método semiempírico) de campo autoconsistente para produzir estruturas eletrônicas a serem analisadas pelo software PRIMoRDiA; iii) Cálculo de descritores químicos quânticos para todas as estruturas amostradas; iv) Análise estatística de todas as estruturas da amostra.

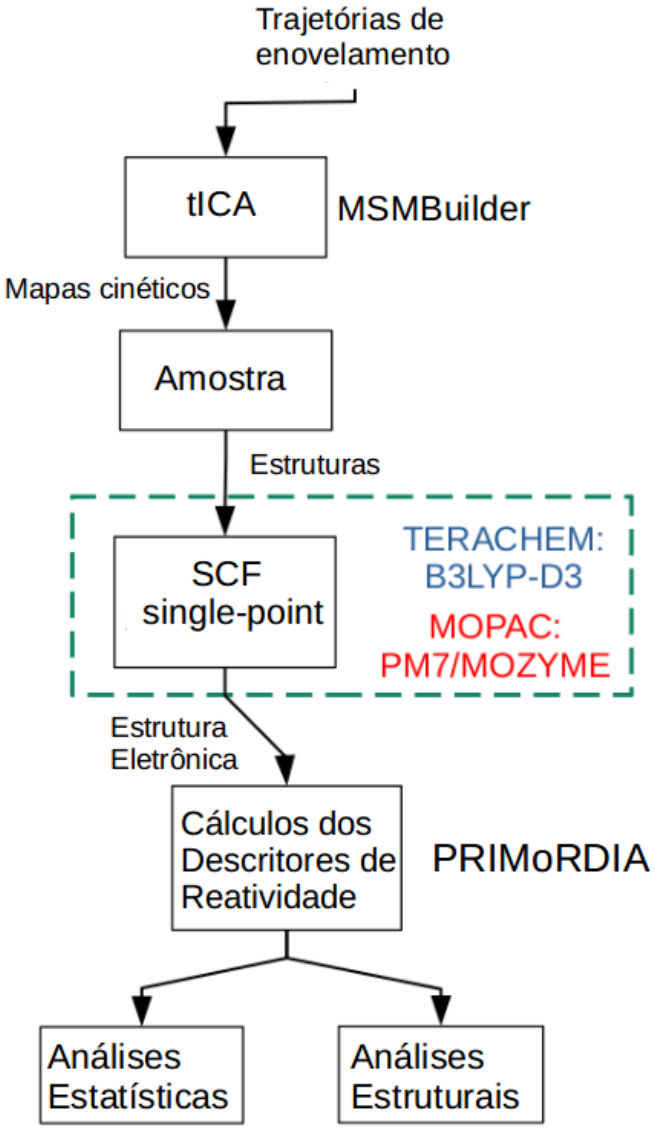


Figura 4.4: Fluxograma mostrando o pipeline de informações e os procedimentos computacionais realizados neste trabalho.

CAPÍTULO 5

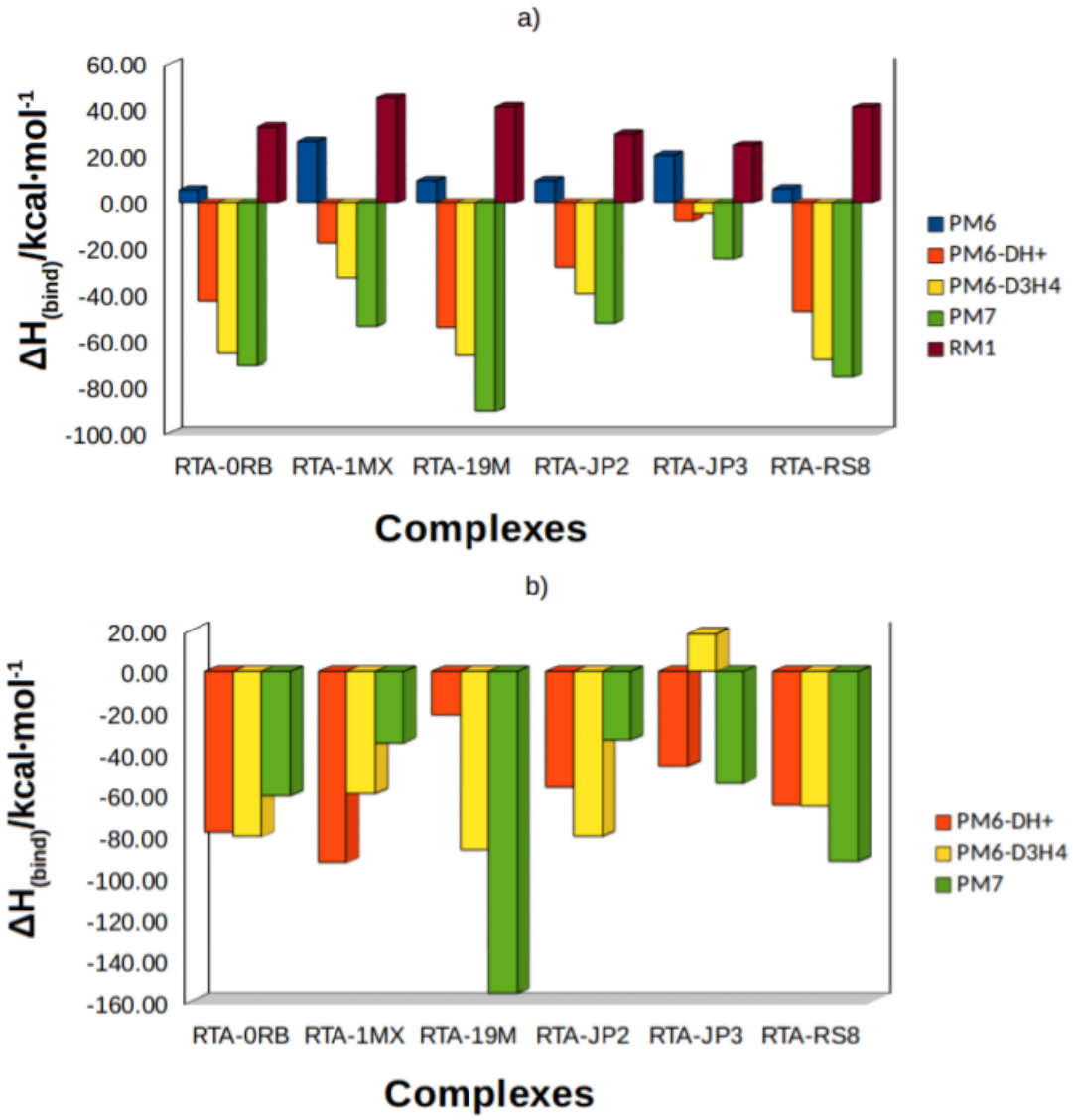
RESULTADOS E DISCUSSÕES

Nessa seção, discutiremos os resultados obtidos através de cálculos quânticos e descritores de reatividade aplicados a dois problemas biológicos distintos: 1) interação proteína-ligante e 2) enovelamento de proteínas, conforme apresentados nas seções de introdução e metodologia.

5.1 Interação Proteína-Ligante: candidatos a inibidores da toxina A da ricina (RTA)

Na Figura 5.1 , apresentamos os resultados de ΔH_{bind} de todos os complexos RTA-ligante estudados neste trabalho, considerando que ambas as estratégias são realizadas: (i) cálculo *single-point* de energia e (ii) cálculo do ΔH_{bind} após otimização de geometria *full atom* a RTA (R), ligante (L) e e complexo RTA-Ligante (R-L).

Figura 5.1: Entalpias de ligação,ΔH_{bind}, para os seguintes complexos: RTA-0RB, RTA-1MX, RTA-19M, RTA-JP2, RTA-JP3 e RTA-RS8. Em (a), os resultados são representados a partir de cálculos *single-point*, e em (b), os resultados são obtidos após a otimização da geometria *full atom* de R, L e R-L.



.

Ao analisarmos os resultados do ΔH_{bind} obtidos via cálculos *single-point* (Figura 5.1 a), podemos ver que os métodos PM6-DH+, PM6-D3H4 e PM7 previram valores negativos de ΔH_{bind} para todos os complexos RTA-ligantes estudados. A faixa de valores de ΔH_{bind} também é consistente com os resultados esperados para para complexos R-L a partir de ensaios termoquímicos, cerca de dezenas de $kcal \cdot mol^{-1}$.[247]

Os métodos PM6 e RM1 apresentaram baixo desempenho, apresentando valores positivos

de ΔH_{bind} para todos os complexos RTA-ligantes. Comparando o desempenho entre os métodos PM6, PM6-DH+ e PM6-D3H4, verificamos que há diferenças nas entalpias de ligação quando as correções para dispersão e ligação de hidrogênio são consideradas. Esses resultados são consistentes com estudos recentes que sugerem que a qualidade dos métodos semiempíricos apresenta melhorias significativas quando tais correções são consideradas.[17,248,249] Esses estudos também indicam que os métodos semiempíricos que incorporam dispersão e ligações de hidrogênio produzem resultados semelhantes aos das correções do tipo D para os métodos da teoria do funcional da densidade (DFT).

A figura 5.1 b mostra os resultados de ΔH_{bind} depois da optimização da geometria *full atom* para todos os complexos utilizando os métodos PM6-DH+, PM6-D3H4, e PM7. Verificamos que os métodos seguiram a mesma tendência dos cálculos *single-point*, ou seja, as entalpias de ligação calculadas foram negativas para todos os complexos RTA-ligante. A única exceção foi o ligante RTA-JP3, que apresentou um valor de ΔH_{bind} positivo com o método PM6-D3H4. Esse resultado sugere que os métodos PM6-DH+, PM6-D3H4 e PM7 são capazes de descrever, pelo menos qualitativamente, as entalpias de ligação dos sistemas estudados. Observamos que a otimização da geometria *full atom* mostrou uma tendência de diminuir o valor do ΔH_{bind} dos complexos. Dos 18 cálculos de ΔH_{bind} realizados (otimização de seis complexos com três métodos semi-empíricos distintos: PM6-DH+, PM6-D3H4 e PM7), 12 apresentaram um valor mais baixo de ΔH_{bind} em comparação com aqueles obtidos através de cálculos *single-point*. Ocorreram exceções para os complexos RTA-0RB (ΔH_{bind} = -60,50 $kcal \cdot mol^{-1}$), RTA-1MX (ΔH_{bind} = -34,41 $kcal \cdot mol^{-1}$) e RTA-JP2 (ΔH_{bind} = -33,23 $kcal \cdot mol^{-1}$) com o método PM7, complexos RTA-JP3 (ΔH_{bind} = 18,35 $kcal \cdot mol^{-1}$) e RTA-RS8 (ΔH_{bind} = -65,33 $kcal \cdot mol^{-1}$) com o método PM6-D3H4 e o complexo RTA-19M (ΔH_{bind} = -20,97 $kcal \cdot mol^{-1}$) com o método PM6-DH+.

Na Tabela 5.1 , apresentamos os valores de IC_{50} [205–207,250] e ΔH_{bind} para cada complexo RTA-ligante avaliado neste trabalho.

Tabela 5.1: Entalpias de ligação ($kcal \cdot mol^{-1}$) para os seis complexos RTA-ligante avaliados neste estudo.

Ligante	IC_{50} μM	ΔH_{bind} via cálculos single-point					ΔH_{bind} com otimização de geometria		
		PM6	PM6-DH+	PM6-D3H4	PM7	RM1	PM7	PM6-DH+	PM6-D3H4
19M	15 (1)	9.58 (3)	-54.37 (1)	-66.47 (2)	-90.76 (1)	41.27 (5)	-156.11 (1)	-20.97 (6)	-86.38 (1)
RS8	20 (2)	6.10 (2)	-47.31 (2)	-68.23 (1)	-75.59 (2)	41.12 (4)	-93.33 (2)	-64.95 (3)	-65.33 (4)
0RB	70 (3)	5.55 (1)	-42.88 (3)	-65.53 (3)	-70.5 (3)	32.74 (3)	-60.5 (3)	-78.21 (2)	-80.12 (2)
1MX	209 (4)	26.28 (6)	-18.06 (5)	-33.01 (5)	-53.65 (4)	45.24 (6)	-34.41 (5)	-92.76 (1)	-59.54 (5)
JP2	230 (5)	9.71 (4)	-28.5 (4)	-39.53 (4)	-52.39 (5)	-29.51 (1)	-33.23 (6)	-56.1 (4)	-79.94 (3)
JP3	380 (6)	20.54 (5)	-8.12 (6)	-5.16 (6)	-24.89 (6)	24.78 (2)	-54.16 (4)	-45.87 (5)	18.35 (6)

Ao comparar os resultados do ΔH_{bind} obtidos por cálculos *single-point* usando os métodos PM6, PM6-DH+, PM7 e RM1 com dados experimentais de IC_{50} para os ligantes (19M, RS8, 0RB, 1MX, JP2 e JP3),[10] observamos que o método PM6 apresentou uma correlação de 0,688 e o método RM1 não apresentou correlação com o IC_{50} . Por outro lado, os métodos PM6-DH+ e PM7 foram capazes de identificar o ligante 19M, que possui o menor valor de IC_{50} (15 μM), como o melhor ligante.

O método PM6-D3H4 mudou a ordem de classificação dos ligantes 19M e RS8 (ΔH_{bind} = -66,47 e -68,23 $kcal \cdot mol^{-1}$, respectivamente), mas isso pode ter ocorrido porque os valores de IC_{50} desses dois ligantes são muito próximos (15 e 20 μM), de modo que o método PM6-D3H4 não foi sensível o suficiente para classificar o melhor ligante para pequenas variações de IC_{50}. O método PM6-DH+ também apresentou inversões dos valores das entalpisa de ligação entre os ligantes 1MX (ΔH_{bind} = -18,06 $kcal \cdot mol^{-1}$) e JP2 (ΔH_{bind} = -28,50 $kcal \cdot mol^{-1}$). O método PM7 foi capaz de classificar corretamente todos os ligantes, sendo mais sensível às pequenas variações do IC_{50}.

Quando realizamos o tratamento estatístico entre os dados de IC_{50} e o ΔH_{bind} obtidos através de cálculos *single-point*, encontramos valores de R de 0,962, 0,975 e 0,984 para os métodos PM6-DH+, PM7 e PM6-D3H4, respectivamente. Assim, o método PM6-D3H4 apresentou a melhor correlação com os dados de IC_{50}. Ressaltamos que os resultados para os métodos PM6-DH+ e PM6-D3H4 são bastante satisfatórios. Na Figura 5.2, apresentamos os gráficos de correlação entre os dados de IC_{50} e ΔH_{bind} obtidos via cálculos *single-point* com os métodos PM6-DH+, PM6-D3H4 e PM7.

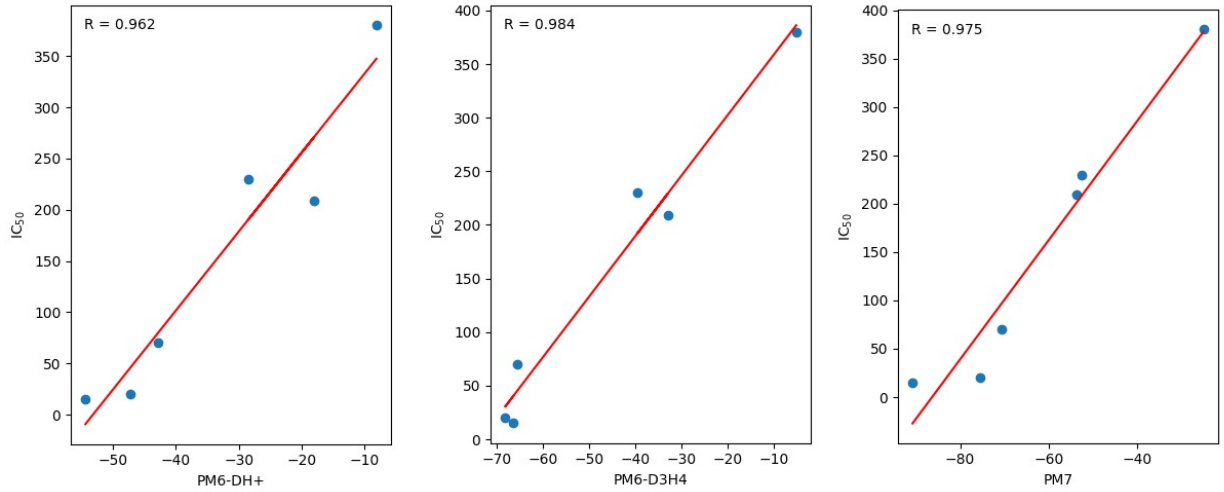


Figura 5.2: Gráfico de correlação entre dados de IC_{50} e ΔH_{bind} obtidos via cálculos *single-point* para os métodos: **a)** PM6-DH+, **b)** PM6-D3H4 e **c)** PM7.

Ao analisarmos os dados do ΔH_{bind} com otimização da geometria *full atom* (Tabela 5.1), verificamos que o método PM6-DH+, embora apresente bons resultados via cálculos *single-point*, não apresentou o mesmo desempenho quando as geometrias foram otimizadas. Este método não foi capaz de identificar o melhor ligante (19M) e tampouco apresentou boa correlação com dados de IC_{50}.

O método PM7 conseguiu classificar os melhores ligantes (de 1 a 3), mas houve inversões na classificação dos demais ligantes. Além disso, houve uma distorção muito acentuada entre a variação nos resultados das entalpias de ligação e de valores IC_{50}. O método PM6-D3H4 foi capaz de identificar o melhor ligante (19M), mas apresentou inversões na classificação entre ligantes RS8 e 0RB e os ligantes 1MX e JP2. Porém, em comparação com o método PM7, o método PM6-D3H4 apresentou distorções menores entre os dados de ΔH_{bind} e IC_{50}.

Quando o tratamento estatístico foi realizado entre os dados de IC_{50} e ΔH_{bind}, encontramos valores de R de -0,085, 0,672 e 0,789 para os métodos PM6-DH+, PM7 e PM6-D3H4, respectivamente. Portanto, o método PM6-D3H4 também apresentou a melhor correlação com o IC_{50} quando realizamos uma estratégia de otimização de geometria *full atom*. Além disso, observamos que na otimização da geometria *full atom*, a correlação entre os dados IC_{50} e as entalpias de ligação diminuiu consideravelmente.

Sulimov *et al.*[251] recalculou as energias de aproximadamente 8.000 estruturas melhores classificadas via *docking molecular* com o campo de força MMFF94 através de cálculos *single-point* e otimização de ligante usando o método PM7 e considerando um modelo de solvente implícito COSMO. Os autores observaram que as energias obtidas por cálculos *single-point* melhoraram muito a classificação de estruturas próximas ao estado nativo. Porém, ao otimizar

os ligantes com o método PM7 no vácuo e recalcular suas energias com o modelo COSMO, os autores observaram uma piora dos resultados para vários complexos em relação aos obtidos via cálculos *single-point* com o método PM7. Em conformidade com esses resultados, verificamos (de forma mais abrangente) que as otimizações da geometria reduziram o desempenho de todos os métodos semi-empíricos considerados.

Indicamos os métodos PM7 e PM6-D3H4 como os melhores métodos semi-empíricos para a obtenção de entalpias de ligação e classificações de ligantes. Não estamos afirmando que os cálculos quânticos no *frame* final da simulação de dinâmica molecular são melhores para obtenção da energias de ligação do ligante do que aqueles que usam otimização de geometria *full atom*. No entanto, isso foi verdade para nosso estudo. Efeitos devido à baixa amostragem, falta de entropia ou cancelamento de erros podem estar por trás desses resultados, mas a investigação dessas questões está fora do escopo deste estudo. Uma excelente discussão sobre essas questões pode ser encontrada no *review* de Ryde e Soderhjelm.[12]

Na Tabela 5.2, apresentamos os dados do desvio quadrático médio (RMSD) entre as estruturas da DM e aquelas otimizadas com os métodos PM6-DH+, PM6-D3H4 e PM7.

Tabela 5.2: Resultados do RMSD (em angstroms) entre o último *frame* de cada uma das 6 simulações de DM e a otimização desses mesmos *frames* por métodos semiempíricos.

Complexo	PM6-DH+ (Å)	PM6-D3H4 (Å)	PM7 (Å)
RTA-19M	1,554	1,296	2,108
RTA-RS8	1,747	1,533	2,024
RTA-0RB	1,567	1,401	1,950
RTA-1MX	1,832	1,468	1,919
RTA-JP2	1,581	1,626	2,036
RTA-JP3	1,745	1,511	2,057

Ao analisar os dados da Tabela 5.2, verificamos que o método PM6-D3H4 apresentou os menores valores de RMSD, exceto para o complexo RTA-JP2, onde o método PM6-DH+ apresentou menor valor de RMSD. O método PM7 apresentou os maiores valores de RMSD para todos os complexos analisados. Esta observação sugere que o método PM7 é capaz de relaxar a estrutura inicial mais do que os métodos PM6-DH+ e PM6-D3H4. A Figura 5.3 mostra a superposição entre a estrutura inicial e as estruturas otimizadas pelo método PM7.

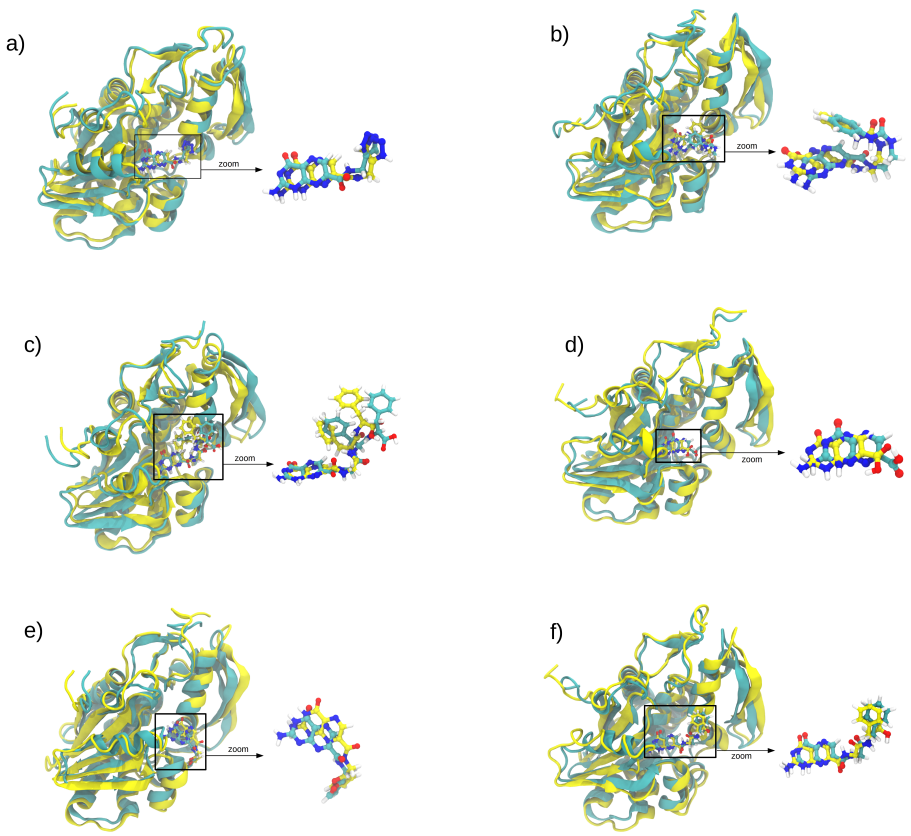


Figura 5.3: Superposição dos complexos RTA-ligante otimizados: **a)** RTA-0RB, **b)** RTA-1MX, **c)** RTA-19M, **d)** RTA-JP2, **e)** RTA-JP3 e **f)** RTA-RS8. As estruturas otimizadas são representadas em verde e as estruturas iniciais são representadas em amarelo.

A partir da análise das estruturas da RTA nos complexos, verificamos que não houve alterações significativas após a otimização, com conservação das estruturas secundárias. Todos os complexos R-L apresentaram RMSDs próximos a 2,0 , o que é consistente com a resolução dos dados experimentais.[252]

Ao compararmos as poses dos ligantes através de optimização geometria *full atom* com estruturas da DM, observou-se que o ligante 0RB (Figura 5.3a) apresentou rotação do grupo triazol, sendo aproximadamente perpendicular à posição preferida deste grupo na proteína. O ligante 1MX (Figura 5.3b) apresentou pequena torção do grupo pterina e variação do anel benzênico. Devido à sua alta flexibilidade conformacional, o ligante 19M (Figura 5.3c) passou por uma grande modificação após a otimização da geometria, principalmente no que diz respeito às torções no grupo pterina, anéis benzênicos e grupo ácido carboxílico. O ligante JP2 (Figura 5.3d) exibiu rotação de seu grupo ácido carboxílico e pequenos desvios em outros grupos da estrutura. O ligante JP3 (Figura 5.3e) apresentou modificação em sua estrutura, com deslocamentos nas posições dos átomos ao longo do mesmo plano. O ligante RS8 (Figura 5.3f)

exibiu uma pequena variação em comparação com o observado no último *frame* da DM.

Na Tabela 5.3, apresentamos os dados de RMSD entre os ligantes dos complexos das simulações MD e os ligantes dos complexos otimizados com os métodos PM6-DH+, PM6-D3H4 e PM7.

Tabela 5.3: Resultados do RMSD (em angstroms) entre os ligantes dos complexos otimizados por métodos semiempíricos e os ligantes dos complexos das simulações de DM.

Complexo	PM6-DH+ (Å)	PM6-D3H4 (Å)	PM7 (Å)
19M	0,675	0,657	1,530
RS8	0,808	1,291	0,598
0RB	0,761	0,755	1,062
1MX	0,932	0,477	0,812
JP2	0,889	0,500	0,909
JP3	0,581	0,510	0,612

Ao analisarmos os dados da Tabela 5.3 , verificamos que as poses dos ligantes apresentaram diversas variações após a otimização da geometria dos complexos R-L. Para o método PM6-DH+, o ligante 1MX apresentou a maior variação (RMSD = 0,932 Å) e o ligante JP3 apresentou a menor variação para esse método (RMSD = 0,581 Å). Para o método PM6-D3H4, observou-se que os ligantes RS8 e 1MX apresentaram as maiores e menores variações, respectivamente, com RMSD's iguais a 1,291 Å e 0,477 Å. Para o método PM7, o ligante 19M apresentou a maior variação (RMSD = 1,530 Å), e a menor variação para esse método foi do ligante RS8 (RMSD = 0,598 Å). Observamos que o método PM7 apresentou maiores valores de RMSD para quatro dos seis ligantes, sendo as duas exceções os ligantes RS8 e 1MX.

Na Tabela 5.4, apresentamos os dados de IC_{50} [205–207,250] e os valores da variação de energia de ligação (ΔE_{bind}) para cada complexo RTA-ligante calculado através do método QM/MM ONIOM com os funcionais B3LYP e ωB97X-D.

Tabela 5.4: Energias de ligação, ΔE_{bind}, (em hartrees) para os seis complexos RTA-ligante obtidas via cálculos *single-point* QM/MM ONIOM. Os dados de IC_{50} estão em (micromolar).

Complexo	IC_{50} (μM)	ΔE_{bind}/B3LYP (ua)	ΔE_{bind}/ωB97X-D (ua)
19M	15 (1)	-0,136 (1)	-0,230 (1)
RS8	20 (2)	-0,119 (2)	-0,202 (2)
0RB	70 (3)	-0,106 (3)	-0,178 (3)
1MX	209 (4)	-0,062 (6)	-0,130 (4)
JP2	230 (5)	-0,070 (4)	-0,129 (5)
JP3	380 (6)	-0,057 (5)	-0,088 (6)

Ao analisar os dados da Tabela 5.4, observamos que os cálculos *single-point* QM/MM ONIOM com o funcional B3LYP foram capazes de identificar os três melhores ligantes (19M, RS8 e 0RB). A única exceção foi a inversão entre os ligantes 1MX e JP2. A correlação entre os dados de IC_{50} e ΔE_{bind} apresentou um valor de R igual a 0,929. Embora os resultados via QM/MM ONIOM com o funcional B3LYP tenham demonstrado boa correlação com os dados experimentais de IC_{50} , os métodos semiempíricos PM6-DH+, PM7 e PM6-D3H4 apresentaram melhores resultados.

Para fins de teste, também realizamos cálculos *single-point* QM/MM ONIOM com o funcional B3LYP considerando uma região QM menor para cada complexo. Neste teste, consideramos na parte QM apenas os seis resíduos mais importantes para o sítio ativo da RTA e o ligante. Quando fazemos isso, a correlação diminui drasticamente para 0,693.

Quando trocamos o funcional B3LYP pelo funcional ωB97X-D, que inclui correções de dispersão DFT-D2,[221] observamos que o valor de R aumentou de 0,929 para 0,972, mostrando uma melhora na correlação com dados de IC_{50} . Além disso, com o uso desse funcional, todos os ligantes foram classificados corretamente de acordo com os valores de ΔE_{bind} e IC_{50}.

Considerando esses resultados e o baixo custo computacional dos métodos semiempíricos, podemos afirmar que os métodos PM7, PM6-DH+ e PM6-D3H4 apresentaram melhor desempenho do que o cálculo QM/MM ONIOM com o funcional B3LYP para complexos RTA estudados neste trabalho. Ao usarmos o funcional ωB97X-D, o método QM/MM apresentou valor R e desempenho semelhantes ao método PM7, mas com um custo computacional superior. A vantagem do método QM/MM ONIOM é que sempre se pode melhorar os resultados aumentando a região QM e usando conjuntos de base maiores e / ou levando em consideração os efeitos do solvente e as correções de dispersão, mas isso acaba sendo caro do ponto de vista

computacional.

5.1.1 Descritores de reatividade aplicados aos complexos proteína-ligante

Os descritores de reatividade para os complexos proteína-ligante foram calculados e usados para realização de uma análise gráfica com o objetivo de caracterizarmos teoricamente as interações mais importantes que ocorrem entre os inibidores e os resíduos do sítio ativo. Conforme mostrado na Figura 5.4, a dureza local tem valores mais altos na bolsa de ligação para o complexo RTA-19M, com seu valor mais alto em Arg-180 e Trp-211. Para o complexo RTA com o ligante JP3, as interações eletrostáticas não são colocadas dentro da região do ligante. Os valores significativos da função Fukui foram calculados nos resíduos da bolsa de ligação para os três complexos de ligantes mais ativos e com pequenos valores para o ligante JP2. A Figura 5.5 descreve a função Fukui para o segundo e terceiro casos mais ativos. Para o complexo RTA-RS8, a função Fukui se localiza em um dos anéis de ligante do grupo pterina e em Tyr-80. No complexo RTA-0RB, todos os átomos do grupo pterina apresentam valores elevados do descritor fornecido, assim como Arg-180.

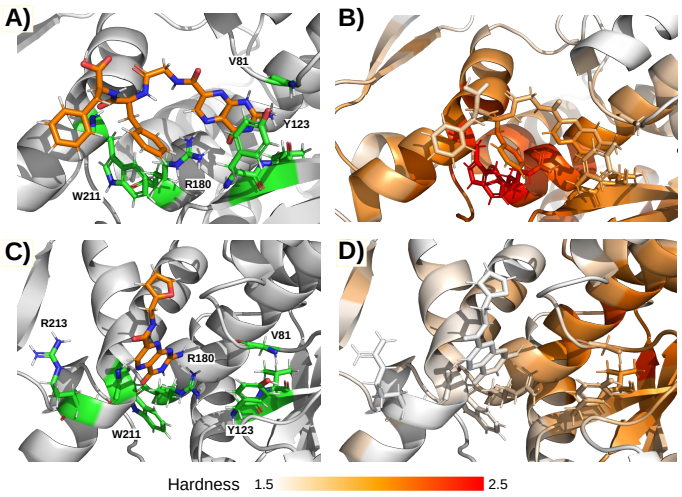


Figura 5.4: Dureza local calculada para complexos RTA com os ligantes 19M e JP3 (laranja). (A) Rótulos para os resíduos mais próximos do ligante 19M (em verde); (B) dureza local para o complexo RTA-19M; (C) rótulos para os resíduos mais próximos do ligante JP3 (verde); e (D) dureza local para o complexo RTA-JP3.

Na Figura 5.6, a função Fukui é apresentada para os ligantes 1MX e JP2, mostrando a distribuição do descritor no grupo pterina e nenhuma distribuição nos resíduos mais próximos. A estrutura molecular desses ligantes é baseada no grupo pterina,[207] o mesmo grupo presente na adenina que é hidrolisada no ribossomo pela ação da catálise do RTA.[5] Nos complexos considerados pelos cálculos, a reatividade neste grupo é sempre indicada pela função Fukui.

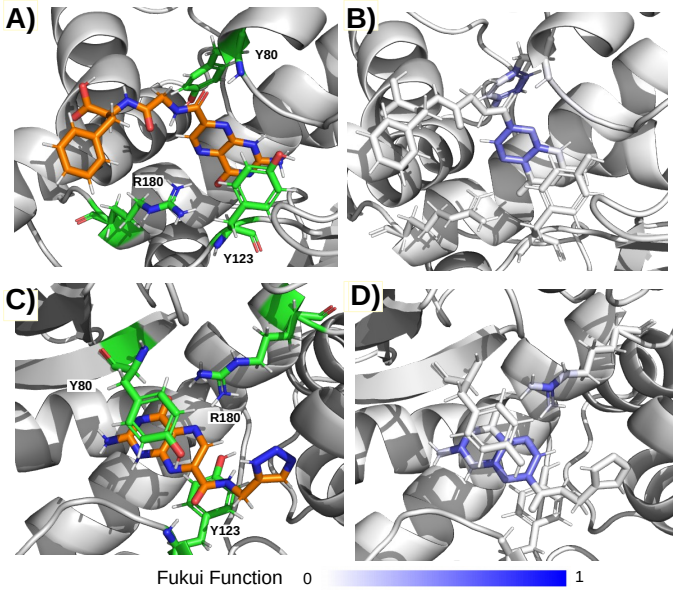


Figura 5.5: Função de Fukui calculada para complexos RTA com os ligantes RS8 e 0RB (alaranjado). (A) Rótulos para os resíduos mais próximos do ligante RS8 (verde); (B) Função Fukui para o complexo RTA-RS8; (C) rótulos para os resíduos mais próximos do ligante 0RB (verde); e (D) dureza local para o complexo RTA-0RB.

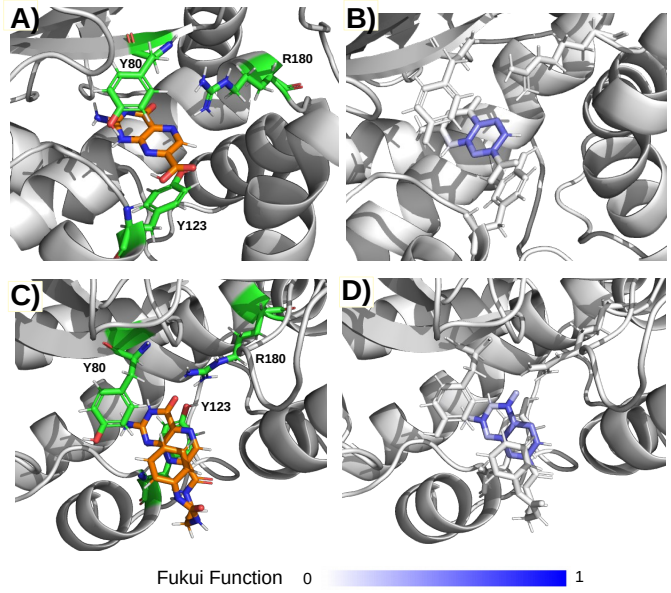


Figura 5.6: Função de Fukui calculada para complexos RTA com os ligantes 1MX e JP2 (laranja). (A) Rótulos para os resíduos mais próximos do ligante 1MX (verde); (B) Função Fukui para o complexo RTA-1MX; (C) rótulos para os resíduos mais próximos do ligante JP2 (verde); e (D) dureza local para o complexo RTA-JP2

Os modelos construídos a partir de correlações de propriedades termoquímicas poderiam classificar os melhores inibidores, mas apenas descritores de reatividade podem representar, que são as características da estrutura molecular que novos ligantes devem ter para serem mais eficientes.

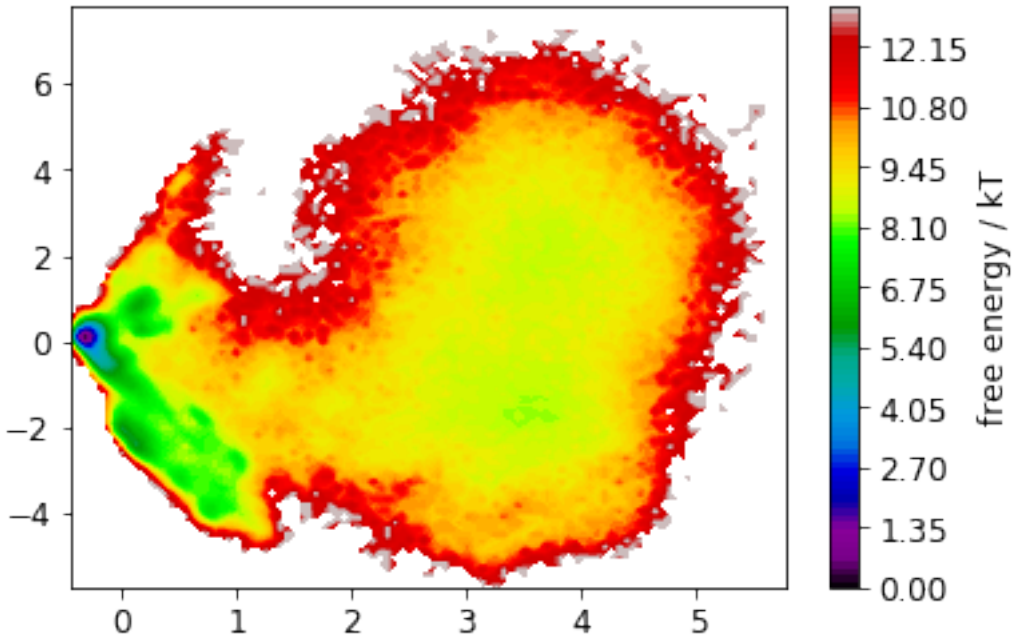
5.2 Enovelamento de Proteínas

5.2.1 Proteína NTL9: Abordagem com o programa PyEMMA

A proteína NTL9 possui 624 átomos e 39 resíduos. A DM utilizada para esses estudo possui aproximadamente 377 μs, sendo que o intervalo de registro dos *frames* foi de 0,2 ns, totalizando 1.883.599 conformações. Nós aplicamos as seguintes *features* para construção da matriz de dimensões do sistema: todos ângulos de torção do *backbone* da proteína, ângulos $\chi 1$ das cadeias laterais e o RMSD mínimo entre os *frames* da trajetória. Após a aplicação desse protocolo, o nosso sistema ficou com um total de 209 dimensões.

Na etapa seguinte, aplicamos a tICA para reduzir a dimensionalidade do sistema, com um *lag time* de 10 ns e uma variância que explica 95 % dos nossos dados, reduzindo o sistema para 51 dimensões. Na figura 5.7, apresentamos o mapa de energia livre plotado a partir da tICA.

Figura 5.7: Mapa cinético baseado na tICA para a proteína NTL9.

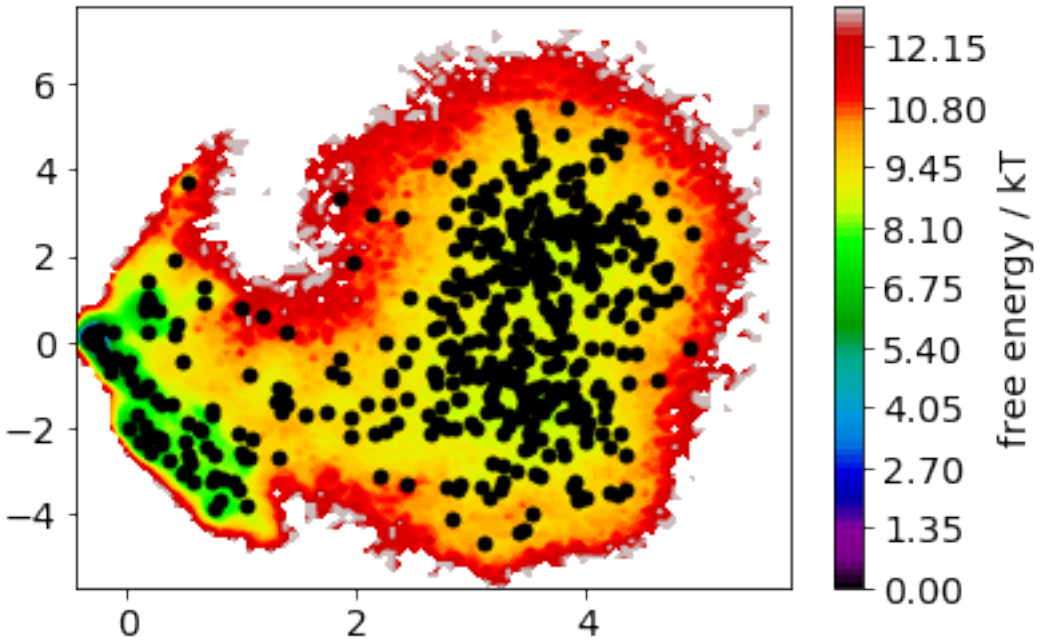


Através do mapa cinético plotado a partir das duas primeiras coordenadas tICA, verificamos que a estrutura com a menor energia livre está concentrada bem à esquerda do

gráfico de energia. Podemos suspeitar que a coordenada x da tICA fornece uma trajetória de desenovelamento, partindo da estrutura de menor energia para a de maior energia. (*Essa abordagem será tratada na próxima subseção com o auxílio do programa MSMbuilder*).

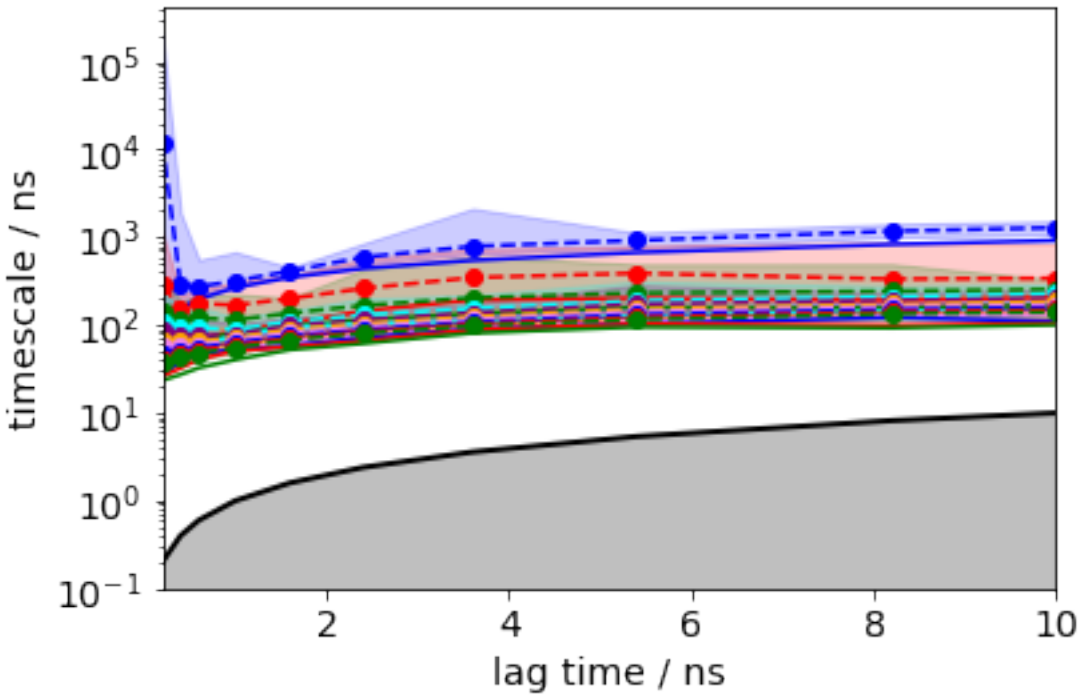
Apesar da redução da dimensionalidade, o nosso sistema ainda possui 1.883.599 conformações, porém agora com 51 ao invés de 209 dimensões. Para resolver essa dificuldade, utilizamos o mapa cinético como base para a etapa de clusterização, em que foi utilizada uma quantidade de 500 *clusters*, ou seja, agrupamos as nossas conformações em 500 grupos e pegamos o centroide (*estrutura representativa*) de cada um. Desse modo, reduzimos as nossas conformações para uma quantidade de 500 estruturas. Na figura 5.8 é apresentado o mapa cinético com os *clusters* sobre a superfície de energia livre.

Figura 5.8: Centroide dos *clusters* com método *k-means* e mapa cinético de fundo para a proteína NTL9.



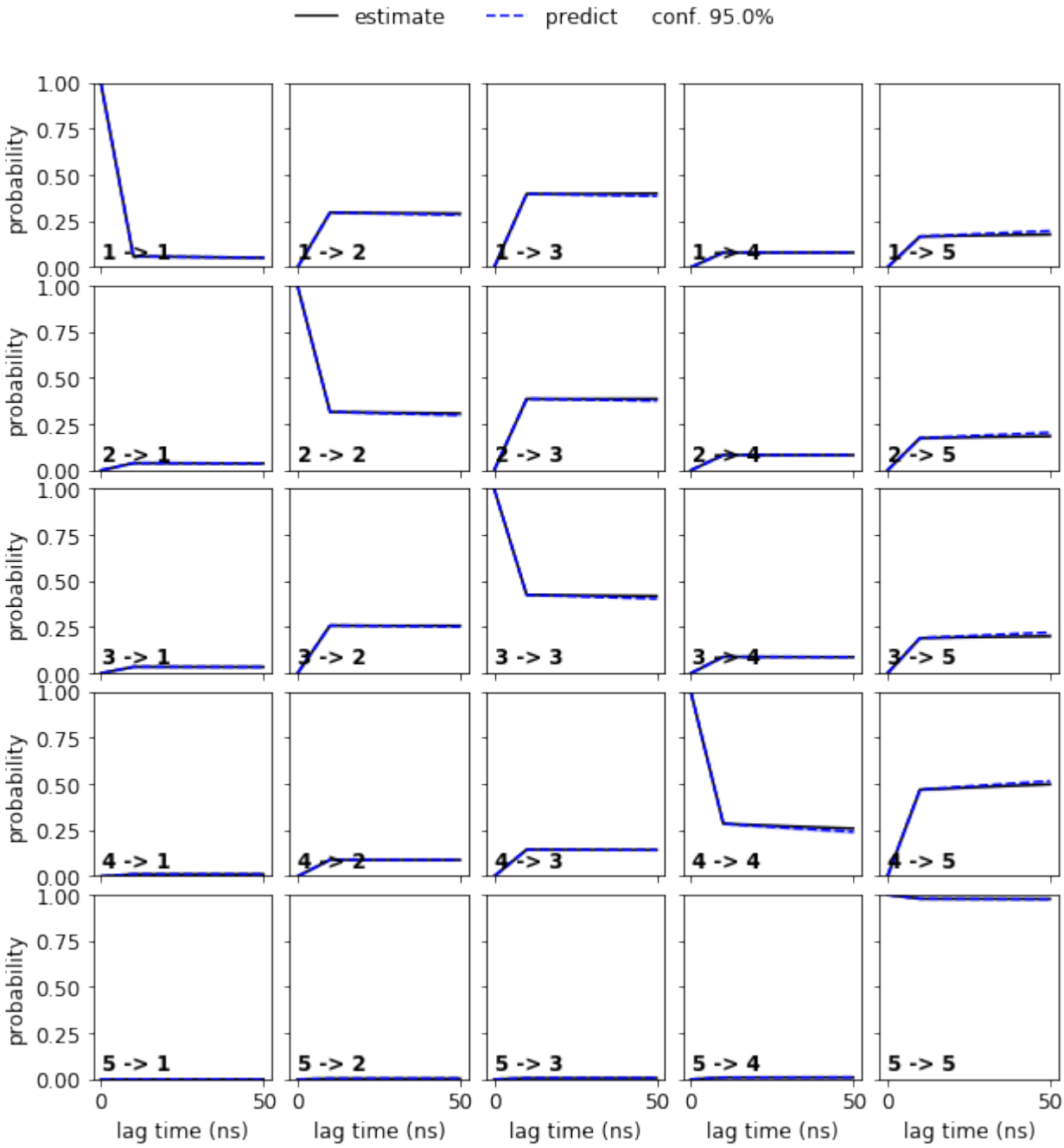
Observamos que os *clusters* cobriram bem a superfíce de energia livre, apontando que a junção da técnica tICA e do método *k-means* fornecem uma boa discretização dos nossos dados. Na figura 5.9 é apresentado o gráfico das escalas de tempo de relaxação implícitas em função do *lag time*.

Figura 5.9: Escalas de tempo de relaxação implícitas para a proteína NTL9



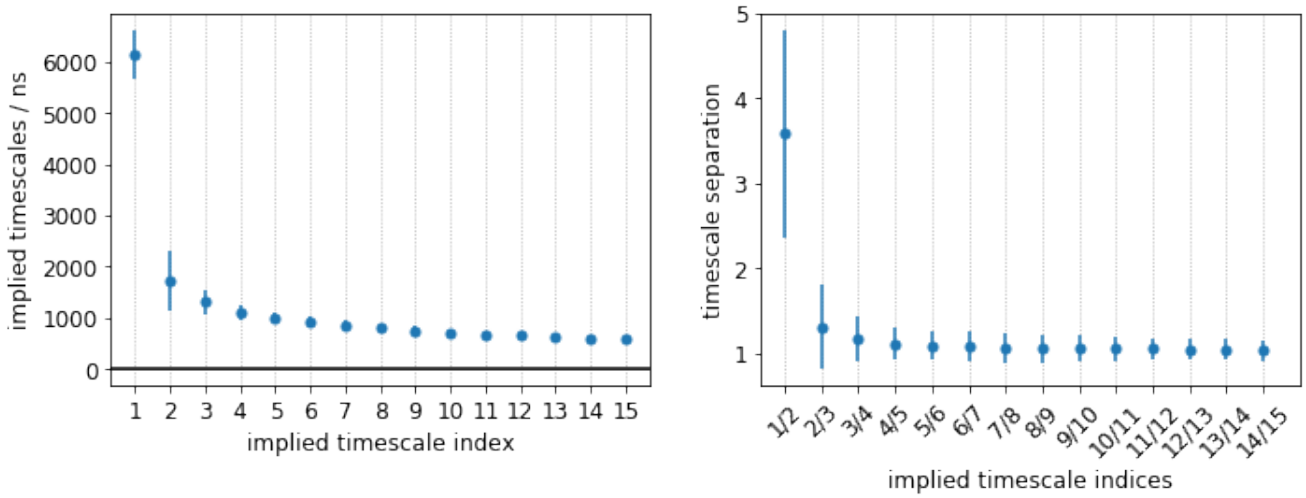
Verificamos que as escalas de tempo convergem rapidamente quando o *lag time* assume valores maiores que 2 ns. Utilizamos o tempo de latência de 10 ns porque verificamos que nesse tempo, o MSM apresentou melhores resultados no teste de passagem. Na figura5.10 é apresentado o teste de Chapman-Kolmogorov para 5 estados.

Figura 5.10: Teste de Chapman-Kolmogorov para um MSM com 5 estados para a proteína NTL9.



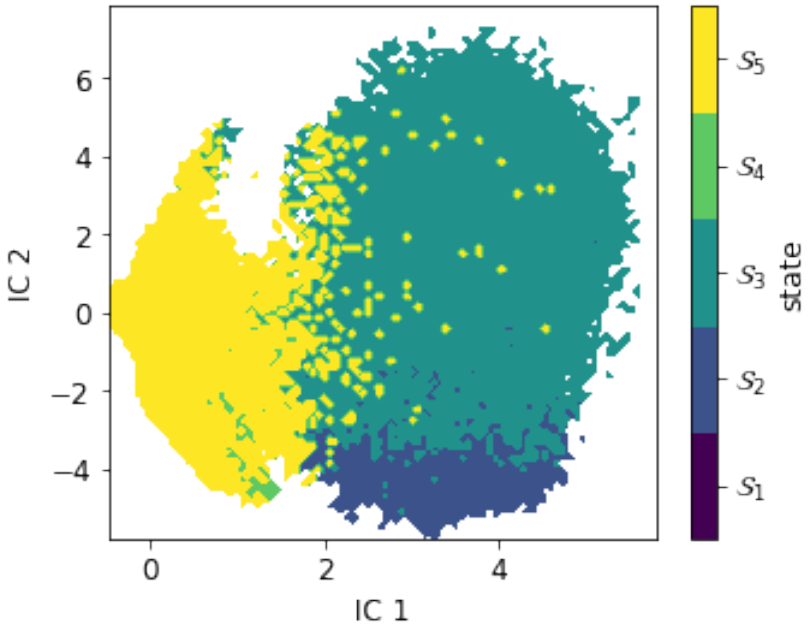
Verificamos que para um MSM com 5 estados, o teste de Chapman-Kolmogorov foi satisfatório. Desse modo, temos um modelo validado dos nossos dados. A partir do MSM é possível avaliarmos as escalas de tempo implícitas que são classificadas em ordem decrescente. A figura 5.11 apresenta as escalas de tempo implícitas.

Figura 5.11: Escalas de tempo implícitas para a NTL9.



Ao analisarmos as escalas de tempo implícitas, verificamos que há uma separação relativamente grande entre o segundo e terceiro processo. Isso sugere que 3 estados metaestáveis são uma boa escolha para um *coarse-graining*. Porém, como estamos interessados em analisar o maior número de caminhos possíveis para o enovelamento, escolhemos 5 macroestados para descrevermos o nosso modelo ao invés de somente 3 com maior fluxo. Na figura 5.12 apresentamos a separação dos microestados em 5 macroestados através do método PCCA++.

Figura 5.12: Modelo de 5 estados metaestáveis obtidos para a proteina NTL9.



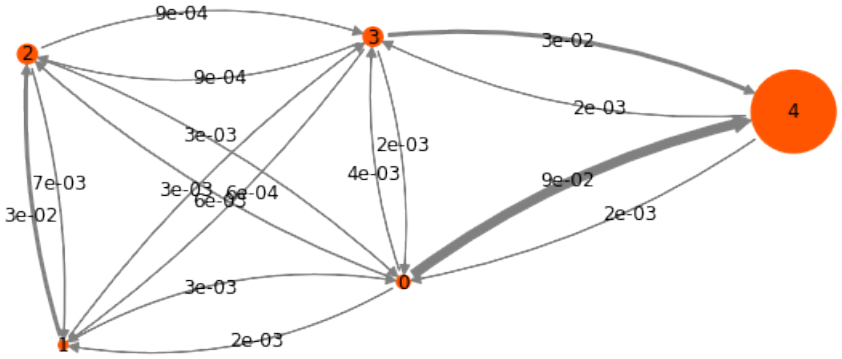
.

Foram amostradas 100 conformações de cada uma das bacias, totalizando 500 estruturas para análise da estrutura eletrônica. Através do MSM obtivemos a matriz de transição entre os 5 macroestados pelo método *coarse-graining*, conforme a seguir:

$$\begin{bmatrix} 8.99 \cdot 10^{-1} & 1.63 \cdot 10^{-2} & 6.07 \cdot 10^{-3} & 3.59 \cdot 10^{-3} & 8.97 \cdot 10^{-2} \\ 3.27 \cdot 10^{-3} & 9.65 \cdot 10^{-1} & 2.93 \cdot 10^{-2} & 2.62 \cdot 10^{-3} & 0.00 \\ 2.84 \cdot 10^{-3} & 6.86 \cdot 10^{-3} & 9.89 \cdot 10^{-1} & 8.60 \cdot 10^{-4} & 0.00 \\ 1.71 \cdot 10^{-3} & 6.24 \cdot 10^{-4} & 8.75 \cdot 10^{-4} & 9.65 \cdot 10^{-1} & 3.18 \cdot 10^{-4} \\ 2.35 \cdot 10^{-3} & 0.00 & 0.00 & 1.75 \cdot 10^{-3} & 9.96 \cdot 10^{-1} \end{bmatrix}$$

A partir da matriz de transição, é possível plotar o grafo com as probabilidades de transição entre os 5 macroestados. Na figura 5.13 é apresentado o grafo de probabilidades de transição entre os 5 macroestados.

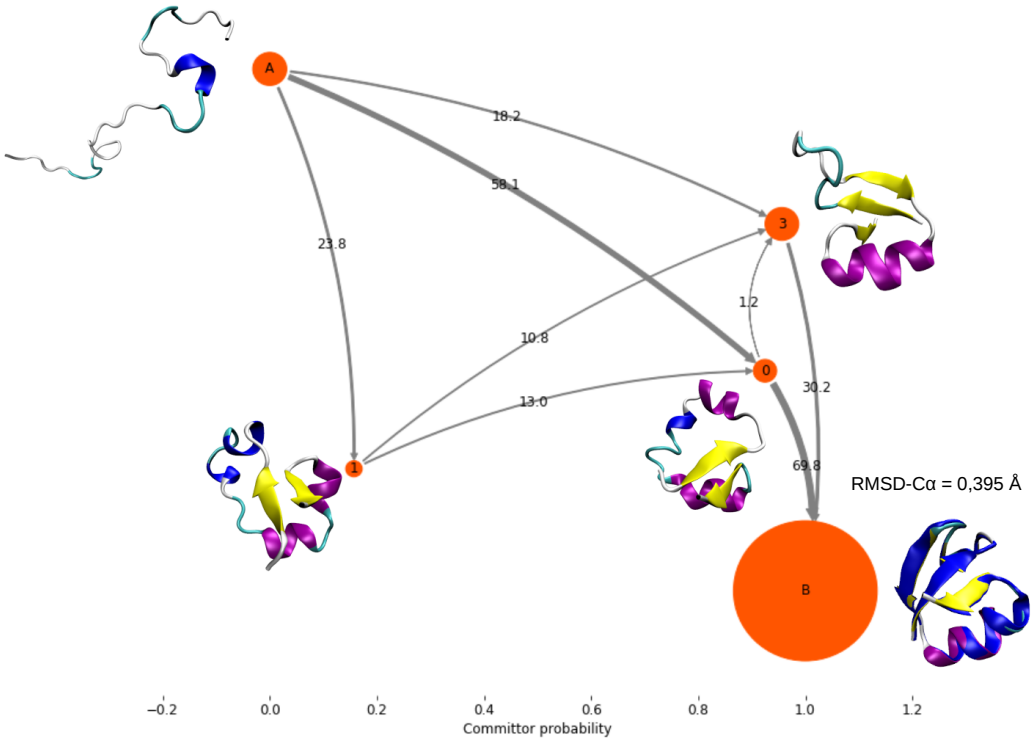
Figura 5.13: Grafo com 5 estados metaestáveis obtidos para a proteína NTL9 através do método *coarse-graining*.



.

Ressaltamos porém que o nosso interesse não está em verificar a probabilidade de transição entre todos os estados, mas a probabilidade de transição entre dois estados específicos que consiste em sair do estado desenovelado da proteína para o estado enovelado. Desse modo, escolhemos o caminho de transição, partindo do macroestado 2 (*desenovelado*) para o macroestado 4 (*estado enovelado*) através da TPT. Para uma melhor interpretação dos dados, plotamos os caminhos de transição normalizados para o percentual de 100%. Desse modo foi possível apontar os dados percentuais para cada caminho do enovelamento. Na figura 5.14 é apresentado o caminho do estado de transição com uma estrutura representativa para cada macroestado.

Figura 5.14: Caminho de transição entre os 5 macroestados para a proteína NTL9.



Observamos que o caminho de maior probabilidade passa por três estados, saindo do estado **A** (*estrutura desenovelada*), passando pelo estado **0** (*estado intermediário*), chegando ao estado B (*estrutura nativa*).O RMSD-Cα da estrutura enovelada em relação à estrutura cristalográfica foi de 0,395 Å. A conformação do estado **B** está representada sobreposta à estrutura cristalográfica (*em azul*). Isso demonstra que o nosso MSM foi capaz de caracterizar bem os possíveis caminhos para o enovelamento, demonstrando que não há somente um caminho para o enovelamento, mas diversos caminhos possível ao longo da superfície de energia livre. Percebemos também que as imagens representativas das macroestados da figura 5.14 sugerem que o enovelamento da NTL9 segue o mecanismo de difusão-colisão. Observa-se que as estruturas secundárias se formam primeiro, seguindo da ancoragem dessas entre si formando a estrutura terciária até chegar ao estado nativo. Analisando o estado 3, é fácil perceber que enquanto a terceira folha β não está completamente formada, a proteína não assume o seu estado de menor energia.

Além do caminho de maior fluxo, podemos observar também caminhos alternativos para o enovelamento de proteínas. Na tabela 5.5 são apresentados os resultados para o fluxo e o percentual do caminho para cada uma das vias de enovelamento.

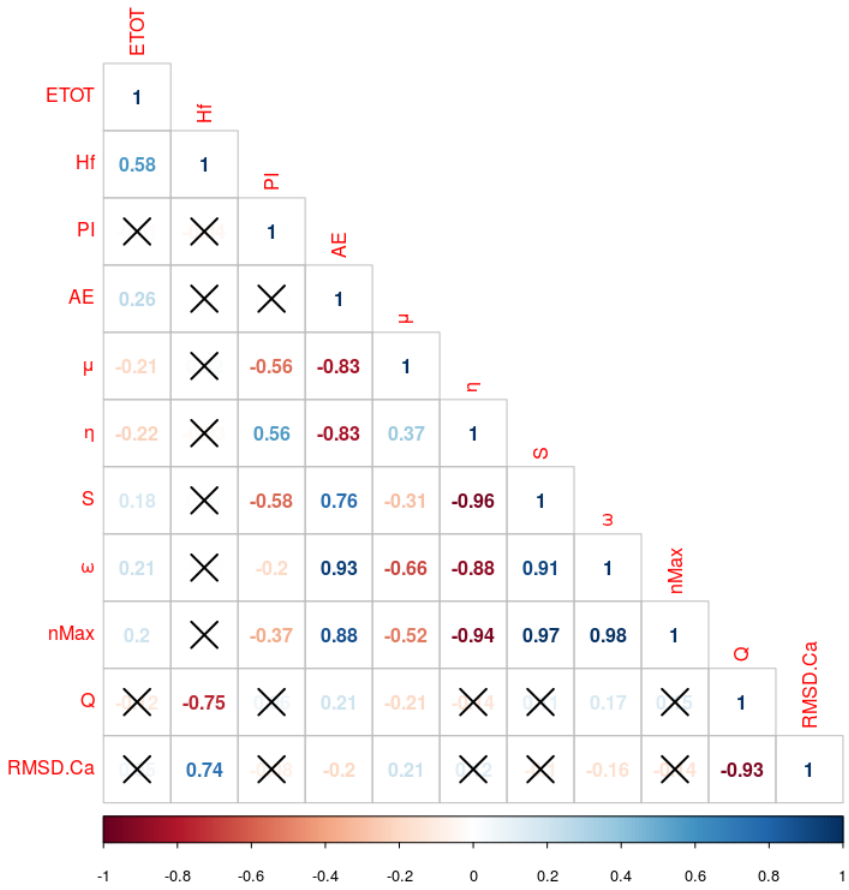
Tabela 5.5: Fluxo e percentagem de caminho para as diversas rotas do enovelamento da proteína NTL9.

Proteína NTL9		
Fluxo do caminho	**Percentagem do caminho**	**Caminho**
2,56E-06	58,1%	[2 0 4]
8,01E-07	18,2%	[2 3 4]
5,20E-07	11,8%	[2 1 0 4]
4,77E-07	10,8%	[2 1 3 4]
5,16E-08	1,2%	[2 1 0 3 4]

Avaliando os dados da tabela 5.5, nota-se que o caminho de maior fluxo [2 0 4] apresenta uma probabilidade de ocorrência de 58,1%, ao passo que o segundo caminho de maior fluxo [2 3 4] apresenta uma probabilidade de 18,2%. Além desses dois caminhos, vale destacar que há mais duas vias para o enovelamento com probabilidade de aproximadamente 11% e uma rota pouco provável, com 1,2% de probabilidade de ocorrência.

A partir desse momento, realizamos o cálculo *single-point* de 100 estruturas representativas de cada macroestado da figura 5.14, totalizando 500 conformações, através do método PM7 e modelo implícito de solvente COSMO, com o programa MOPAC. Através desses dados obtivemos os valores da energia total (E_{TOT}) e do calor de formação $\Delta H_f)$, sendo também possível os QCMDs globais com o programa PRIMORDIA. Além disso, realizamos os cálculos de descritores estruturais como RMSD-C_α e fração de contatos nativos (Q) com o objetivo de avaliarmos se esses apresentam alguma correlação com os descritores globais de reatividade ou com valores de (E_{TOT}) e (ΔH_f). Os QCMDs globais escolhidos foram: potencial de ionização (PI), afinidade eletrônica (AE), potencial químico (μ), dureza química (η), moleza química (S), eletrofilicidade (ω) e número máximo de elétrons (n_{Max}) Na figura 5.15 são apresentados os dados de correlação entre todos os descritores.

Figura 5.15: Mapa de correlação entre descritores globais de reatividade, fração de contatos nativos e RMSD-Cα de um MSM de 5 estados para a NTL9.



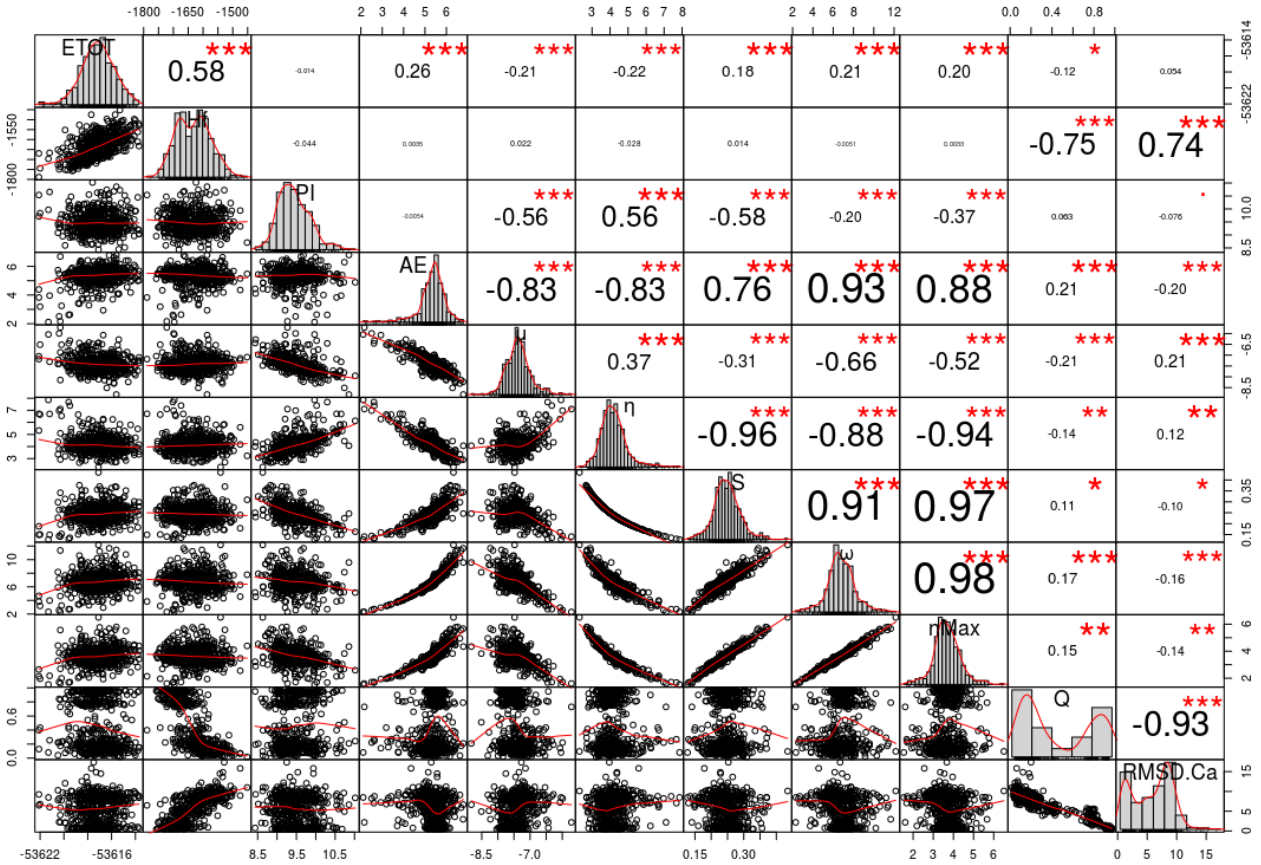
Na figura 5.15, os valores em vermelho representam correlações negativas e em azul correlações positivas, sendo que quando mais intensa for a cor, maior será a correlação correspondente. Como já esperado, observamos uma dependência inversa entre o valor de Q e de RMSD-Cα, uma vez que ao longo do caminho de enovelamento, o RMSD tende a diminuir e o valor de Q tende a aumentar. Também é fácil perceber uma dependência positiva entre o ΔH_f e o RMSD-Cα $(R = 0,58)$, uma vez que o sistema deve diminuir a sua entalpia à medida que descemos no funil de enovelamento em direção ao estado nativo.

Ressaltamos porém, que não foi observado correlação entre os QCMDs globais e valores de RMSD. Esperávamos, por exemplo, que o valor de η fosse aumentando ao longo do caminho de enovelamento, mas isso não ocorreu. O que observamos foi que a densidade eletrônica global variou muito pouco ao longo do funil de energia livre, ou seja, uma estrutura completamente desenovelada apresentou densidade muito próxima da estrutura nativa, o que é muito surpreendente. Esses resultados corroboram com dados experimentais recentes, onde foi observado que os orbitais de fronteira permanecem fortemente localizados durante o eno-

velamento.[87] Desse modo como todas as propriedades eletrônicas calculadas são derivadas da densidade eletrônica, a correlação apresentada foi muito baixa.

Uma maneira útil de ampliarmos a nossa análise consiste em plotarmos, além da correlação, um histograma com a linha de densidade e linha tendência entre as variáveis dos nossos dados. Na figura 5.16 apresentamos o histograma de todas as variáveis do nosso conjunto de dados.

Figura 5.16: Histograma com correlação, linhas de densidade e de tendência de um MSM de 5 estados para a NTL9.



Na parte superior desse gráfico é apresentado um histograma com uma linha de densidade para cada uma das variáveis. Além disso há a comparação entre cada uma das variáveis em que o valor em negrito corresponde ao coeficiente de correlação. Os asteriscos correspondem ao valor p do seguinte modo: *3 asteriscos* $(p < 0,001)$, *2 asteriscos* $(0,001 < p < 0,01)$, *1 asterisco* $(0,01 < p < 0,05)$, *1 ponto* $(0,05 < p < 0,10)$ e caso $(p > 0,10)$ o coeficiente de correlação aparece sem nenhum asterisco ou ponto. Na parte inferior do gráfico há uma plotagem entre cada uma das variáveis em que no eixo x temos uma variável e no eixo y outra variável. Há os pontos para verificarmos como essas variáveis se relacionam e uma linha de tendência vermelha mostrando como esses pontos deveriam estar se a relação fosse perfeita.

Analisando as linhas de tendência entre o RMSD-C_α com as demais propriedades, veri-

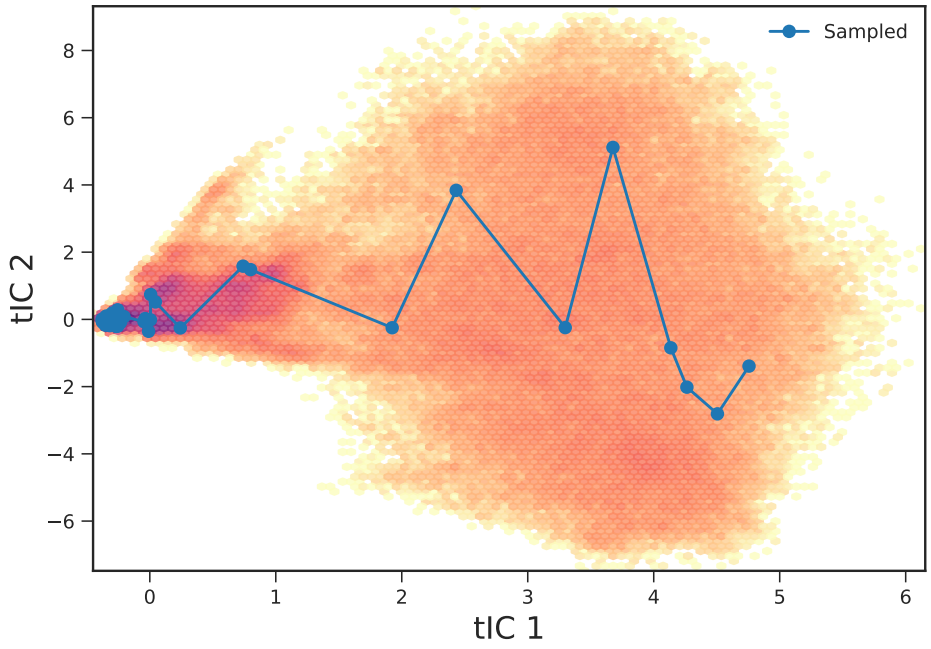
ficamos que a (E_{TOT}), o PI, a η e a S apresentam praticamente uma linha de tendência na horizontal. Logo como, não há uma variância significativa desses descritores ao longo do enovelamento, conforme dito anteriormente. Somente o ΔH_f apresentou alguma correlação $(R = 0,58)$ com RMSD-C$_\alpha$ ou Q. Aqui, evitamos comentar a correlação entre os QCMDs globais, uma vez que são propriedades oriundas da densidade eletrônica e o nosso interesse está na comparação dos descritores globais de reatividade com propriedades estruturais conhecidas (RMSD-C$_\alpha$ e Q).

5.2.2 Proteína NTL9: Abordagem com o programa MSMBuilder

Conforme mencionado anteriormente, a projeção das 2 primeiras coordenadas tICA para a NTL9 na Figura 5.7 fornece um mapa cinético bastante interessante. Verificamos que o eixo x da coordenada tICA fornece uma trajetória de desenovelamento, partindo da estrutura mais enovelada à esquerda para a conformação mais desenovelada à direita.

O programa MSMbuilder possui uma ferramenta para amostrar uma quantidade pré-estabelecida de conformações ao longo de uma coordenada tICA. Para realização do tratamento quântico, escolhemos uma quantidade de 100 conformações ao longo da coordenada x do mapa cinético. Na Figura 5.17 é apresentada a amostragem aleatória ao longo da coordenada x para a proteína NTL9.

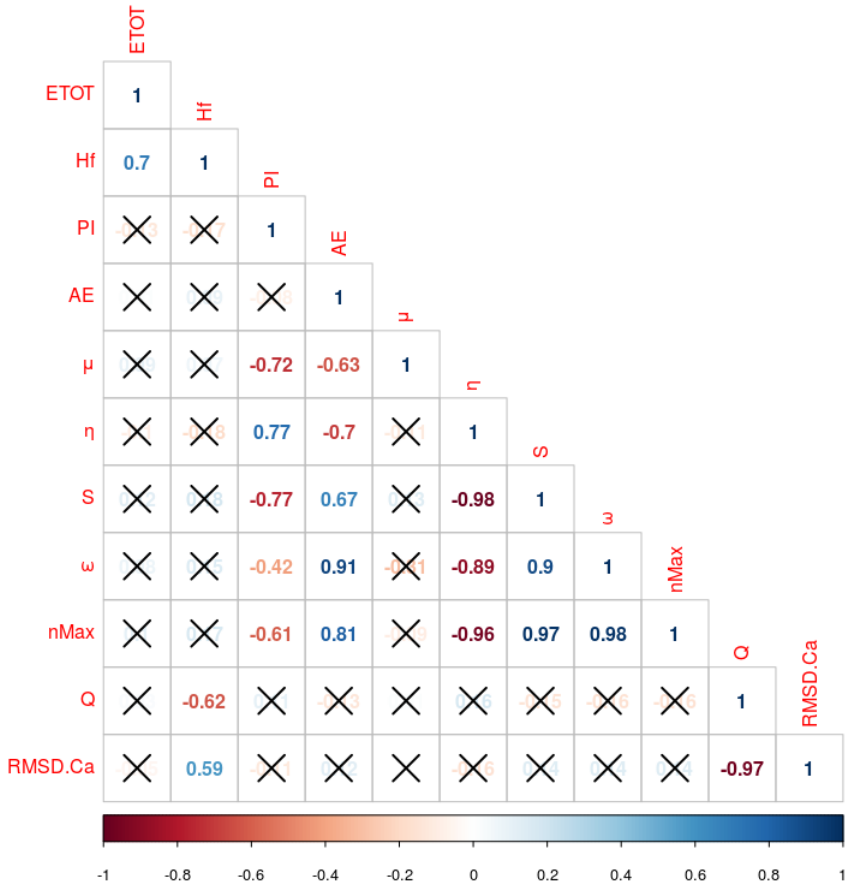
Figura 5.17: Amostra de 100 conformações ao longo da coordanada x da tICA.



Foi realizado o mesmo tratamento quântico da seção anterior, ou seja, método PM7 com

modelo de solvatação implícita COSMO via pacote MOPAC. O intuito desse novo tratamento consiste em verificarmos se a escolha do caminho de enovelamento influencia na avaliação dos descritores de reatividade. Na Figura 5.18 é apresentado o mapa de correlação entre todos os descritores globais, E_{TOT}, ΔH_f, RMSD-Cα e Q para as 100 conformações nesse caminho de enovelamento. Além disso, fizemos o tratamento quântico dessas estruturas via DFT-D3, conforme descrito no capítulo 4. O nosso objetivo foi avaliar se a performance dos QCMDs globais obtidos via DFT-D3 apresentariam melhorias em relação aos obtidos via PM7. Aqui, apresentamos os resultados dos QCMDs globais obtidos via PM7 para compararmos com os resultdos da seção anterior. Os resultados dos QCMDs globais obtidos via DFT-D3 para a proteína NTL9 podem ser consultados nas Figuras A.7 e A.10 do Apêndice A.

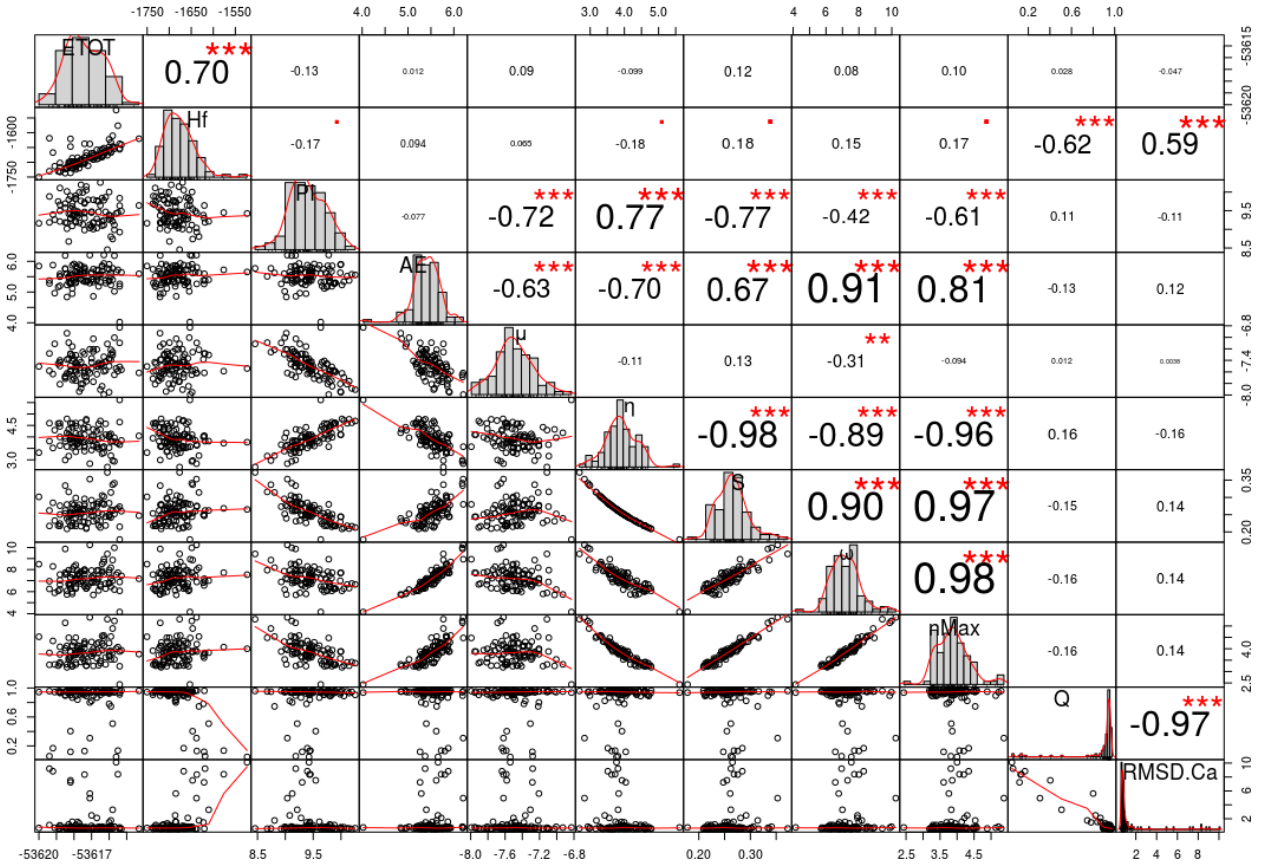
Figura 5.18: Mapa de correlação entre os descritores globais de reatividade, E_{TOT}, ΔH_f, Q e RMSD-Cα para a primeira coordenada tICA da NTL9.



Ao analisarmos o gráfico da Figura 5.18, verificamos que esse segue a mesma tendência observada para os macroestados da Figura 5.15. Somente o ΔH_f apresenta alguma correlação com o RMSD-Cα ou com a fração de contatos nativos (Q). Esses resultados estão de acordo com o observado por Urquiza-Carvalho, G. A. *et al.*[95] em que apontaram que as entalpias

de formação semi-empíricas são uma boa função de pontuação para discriminar estruturas nativas de um conjunto de *decoy*. A correlação entre o RMSD-Cα e Q foi de $-0,93$. Já o H_f apresentou $(R = -0,62)$ e $(R = 0,59)$ em relação a Q e RMSD-Cα respectivamente. Desse modo, mesmo para um caminho de enovelamento distinto, os descritores globais de reatividade apresentaram uma baixa correlação com propriedades estruturais. Na Figura 5.19 é apresentado o histograma das conformações da coordenada x da tICA para a proteína NTL9.

Figura 5.19: Histograma com correlação, linhas de densidade e de tendência referente aos dados da coordenada x da tICA para a NTL9.

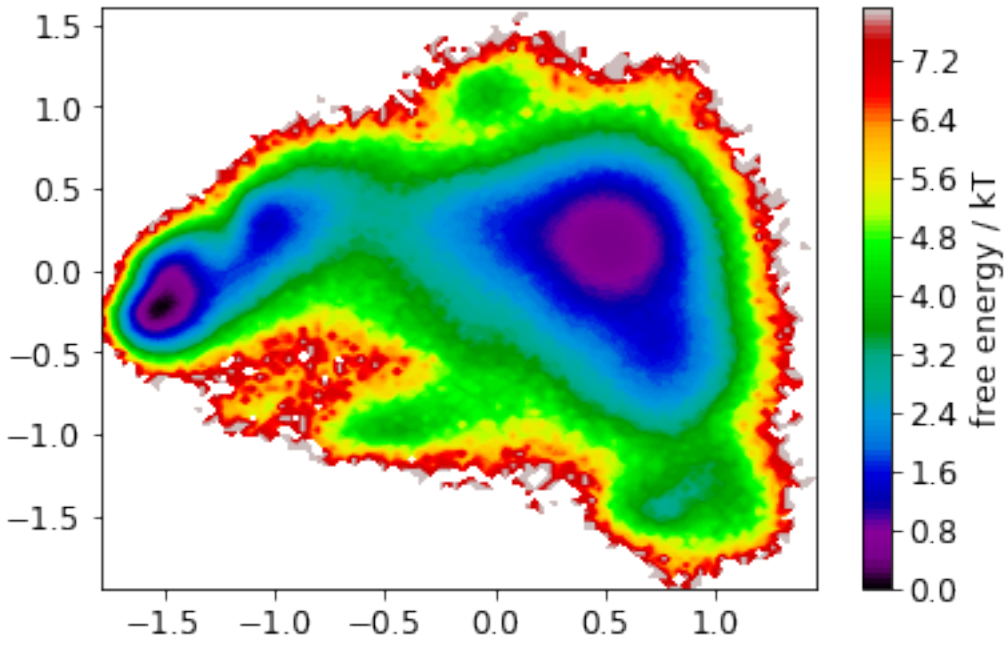


Verificamos nesse gráfico, que os descritores globais não variam ao longo do processo de enovelamento tanto em relação ao RMSD-Cα quanto em relação a Q, uma vez que apresentaram uma linha de tendência horizontal para todos os descritores de reatividade. Desse modo, não há variação da densidade eletrônica para a proteína NTL9, o que sugere que QCMDs globais não fornecem uma coordenada de reação satisfatória para o caminho do enovelamento. Os resultados obtidos via DFT-D3 também não apresentaram nenhuma melhoria em relação aos obtidos via PM7 (figruas A.7 e A.10 do Apêndice A).

5.2.3 Proteína BBA: Abordagem com o program PyEMMA

A proteína BBA possui 504 átomos e 28 resíduos. A DM utilizada para esses estudo possui aproximadamente 223 μs, sendo que o intervalo de registro dos *frames* foi de 0,2 ns, totalizando 1.114.545 conformações. Após a aplicação das mesmas *features* utilizadas para a proteína NTL9, o nosso sistema ficou com 159 dimensões. Na etapa seguinte, aplicamos a tICA para reduzir a dimensionalidade do sistema, com um *lag time* de 500 ns e uma variância que explica 95 % dos nossos dados, reduzindo o sistema para 43 dimensões. Para essa proteína em específico, o passo no tempo foi muito superior ao das demais proteínas, conforme dados retirados da referência [89], tratando-se portanto de um chute inicial. Ao plotarmos as escalas de tempo implícitas, será possível observarmos o comportamento do *lag time* e avaliarmos a possibilidade de escolhermos um passo menor de modo a melhorarmos a resolução do MSM. Na Figura 5.20 apresentamos o mapa de energia livre plotado a partir da tICA para a proteína BBA.

Figura 5.20: Mapa cinético baseado na tICA para a proteína BBA.



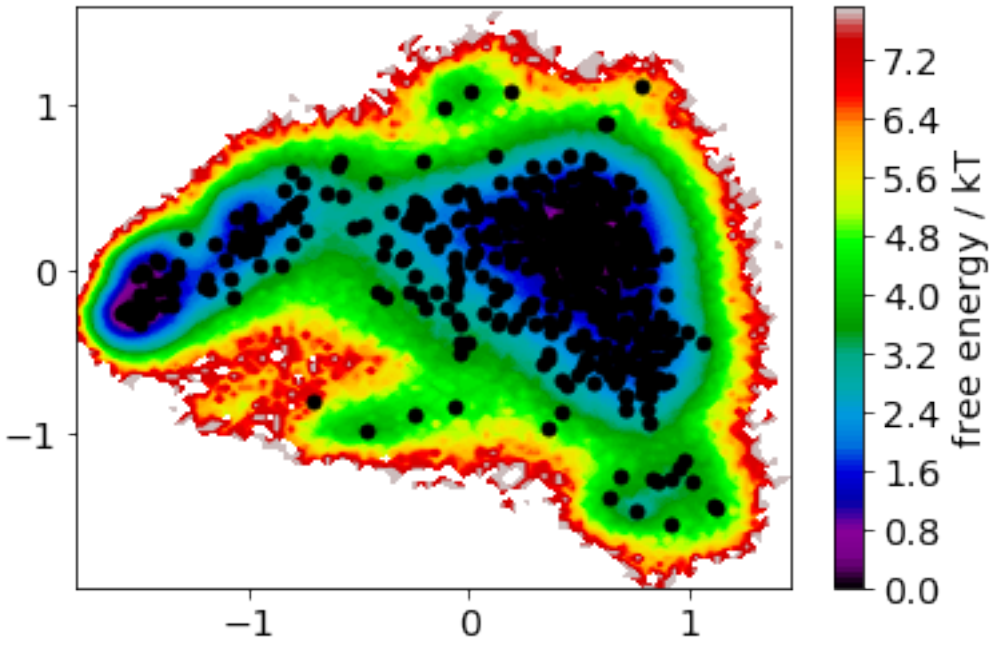
.

O mapa cinético nos fornece informações importantes sobre o nosso sistema. Podemos observar duas bacias mais fundas, uma mais à direita e outra mais à esquerda do gráfico. Observamos que a bacia de energia da direita, apesar de possuir um mínimo, esse não é global e sim local, pois a estrutura de menor energia não está nesse poço. Olhando para a bacia mais à esquerda, verificamos que trata-se do mínimo global, uma vez que a conformação mais

estável encontra-se nessa região.

Após à redução de dimensionalidade, o nosso sistema saiu de 159 para 43 dimensões, porém ainda há 1.114.545 conformações. Para encontrarmos estruturas representativas ao longo do mapa de energia livre, selecionamos 550 *clusters* e pegamos o centróide de cada conjunto de conformações. Desse modo, passamos a ter um total de 550 estruturas representativas de todo o nosso sistema. Na Figura 5.21 é apresentado os *clusters* ao longo do mapa cinético.

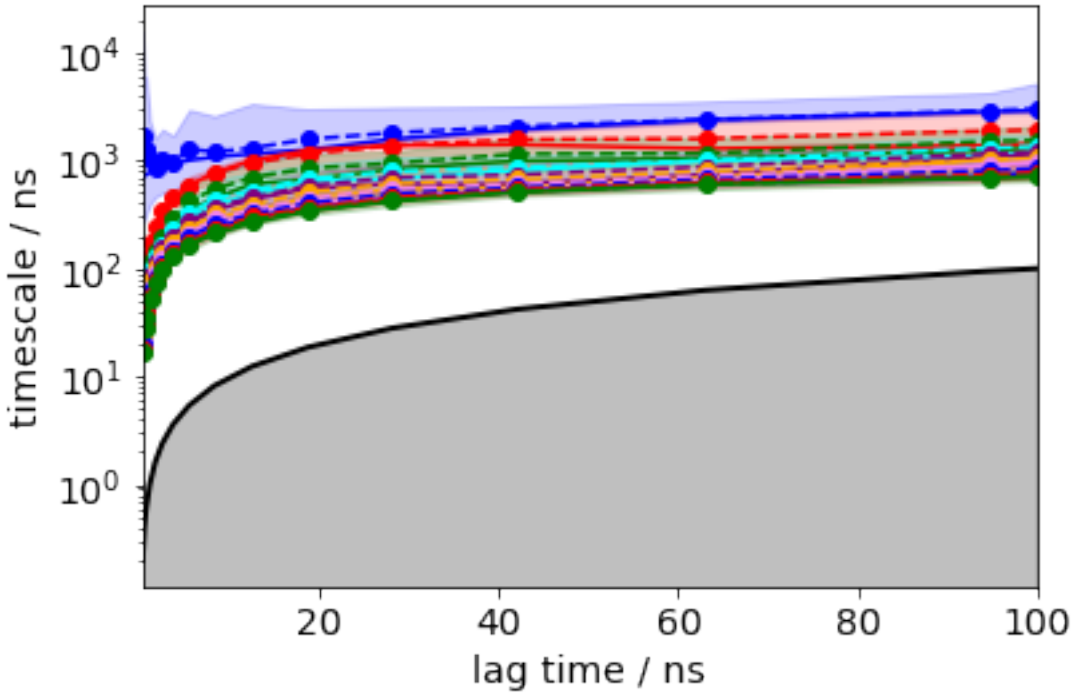
Figura 5.21: Centroide dos *clusters* com o método *k-means* e mapa cinético de fundo para a proteína BBA.



.

Percebemos que os *clusters* se distribuiram bem ao longo da superfície de energia livre, mas percebemos algumas regiões com menor densidade de conformações. Tentamos aumentar o número de *clusters* para 700, porém percebemos que as regiões continuam menos densas, não fazendo sentido aumentarmos para mais de 550, uma vez que o custo computacional aumentaria bastante. A tICA e o método de clusterização *k-means* são bem correlacionados, uma vez que não há presença de *clusters* fora do mapa cinético. Caso tivéssemos utilizado a PCA ao invés da tICA, essa situação poderia não ocorrer. Na Figura 5.22 é apresentado o gráfico das escalas de tempo de relaxação implícita em função do *lag time*.

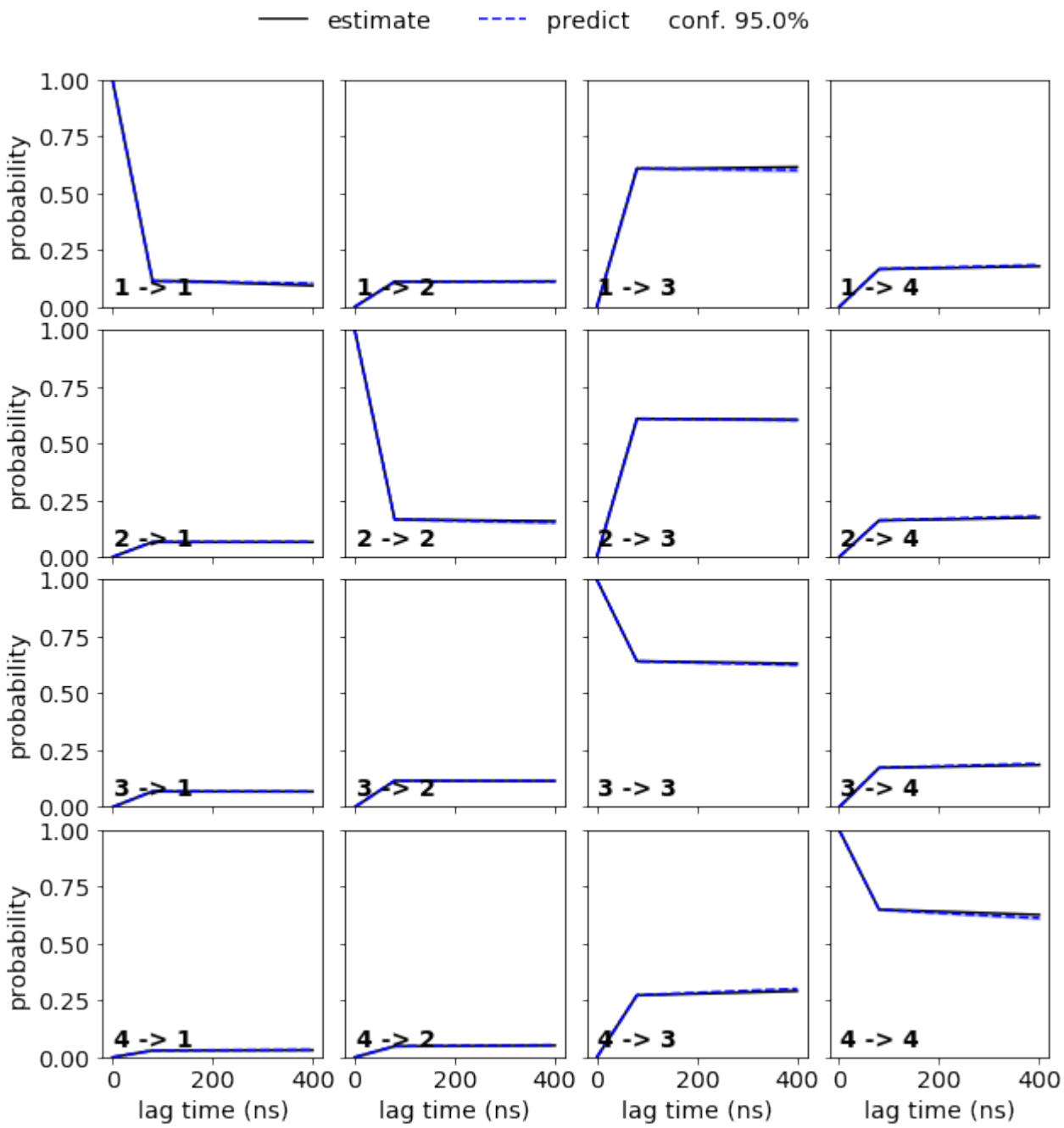
Figura 5.22: Escalas de tempo de relaxação implícitas para a proteína BBA.



.

Percebemos que a partir de um *lag time* igual a 20 ns, as escalas de tempo convergem rapidamente. Nos nossos testes, quando o tempo de latência assumiu o valor de 80 ns, o teste de Chapman-Kolmogorov apresentou melhores resultados. Desse modo, escolhemos esse tempo para as próximas etapas do processo. Na Figura 5.23 é apresentado o teste de Chapman-Kolmogorov para 4 estados.

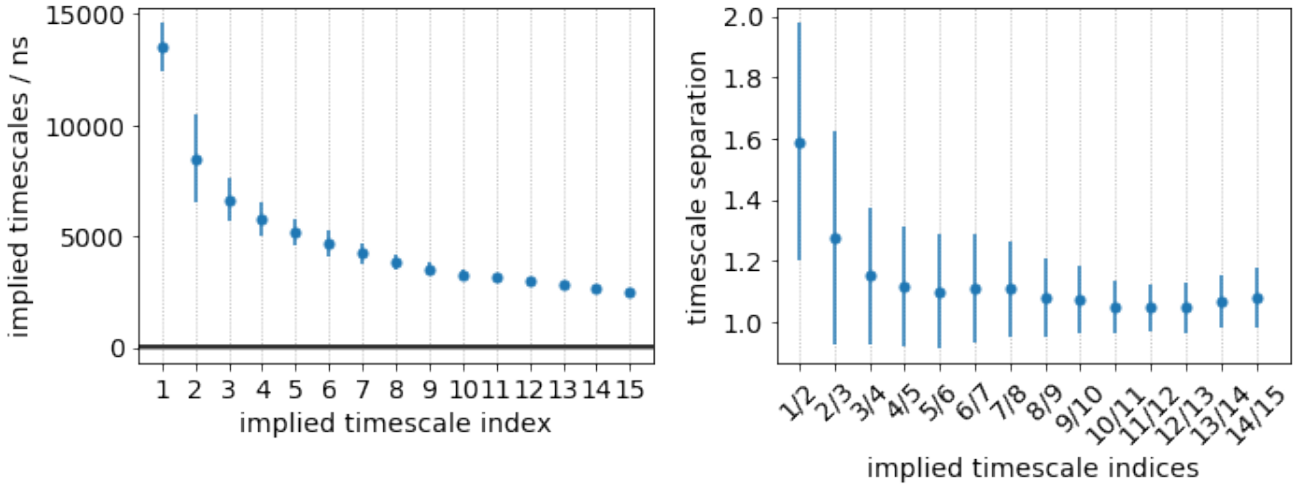
Figura 5.23: Teste de Chapman-Kolmogorov para um MSM com 4 estados para a proteína BBA.



.

Observamos portanto, que o teste de Chapman-Kolmogorov foi validado para um MSM com 4 estados, ou seja, o valor estimado (*linha preta*) e o valor predito (*linha azul*) estão praticamente sobrepostos. A partir desse momento, podemos avaliar as escalas de tempo implícitas que estão classificadas em ordem decrescente. Na Figura 5.24 são apresentadas as escalas de tempo implícitas.

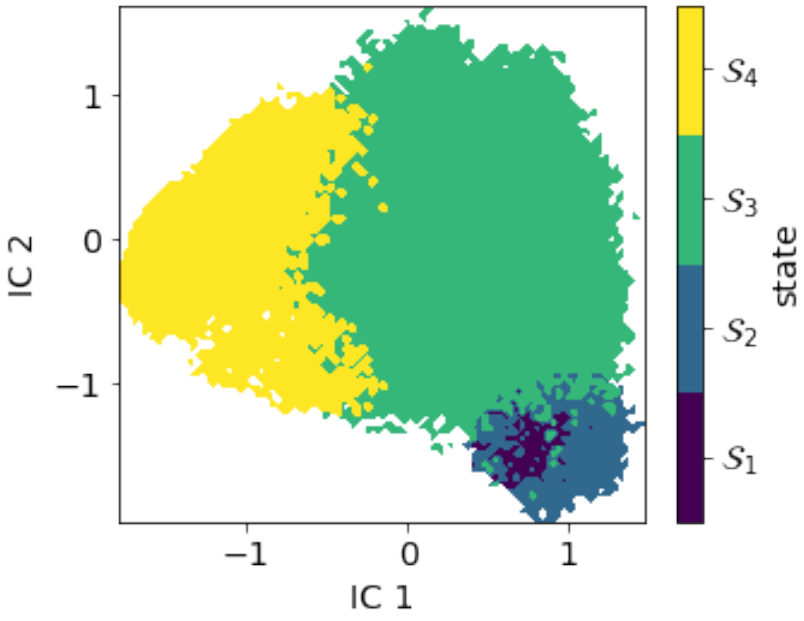
Figura 5.24: Escalas de tempo implícitas para a BBA.



.

Nesse gráfico, verificamos a partir da separação da escala de tempo, que há um intervalo relativamente grande entre o 2^o e o 3^o processo, sugerindo que um modelo com 3 estados metaestáveis seria uma boa escolha para aplicação do método *coarse-graining*. Como temos interesse também em rotas alternativas, escolhemos 4 estados metaestáveis para a nossa análise. Na Figura 5.25 apresentamos a superfície de energia livre que separa os microestados em 4 macroestados através do método PCCA++.

Figura 5.25: Modelo de 4 estados metaestáveis obtidos para a proteina BBA.



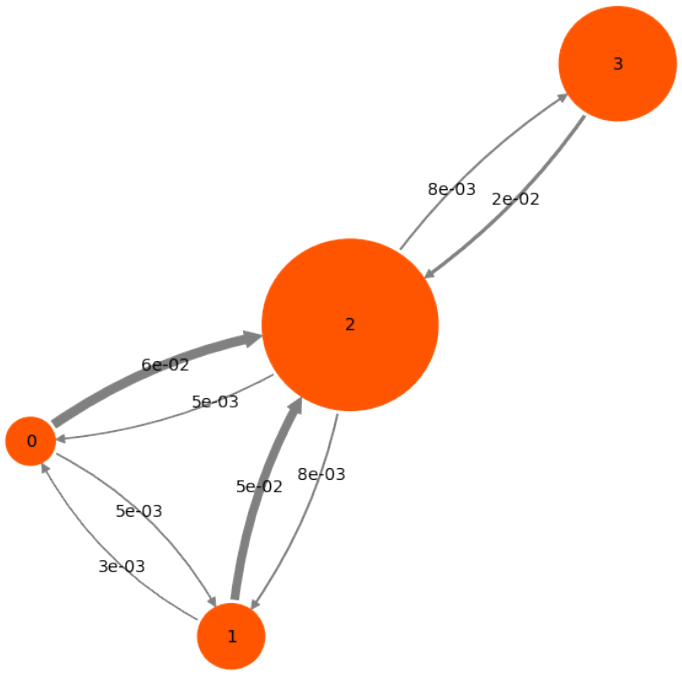
.

Observamos que os limites entre as 4 bacias metaestáveis estão muito bem definidas, sendo que o macroestado de menor energia é o S_4. Inferimos essa informação a partir da análise do mapa cinético da Figura 5.20. Foram amostradas 100 conformações de cada macroestado, totalizando 400 estruturas para tratamento quântico posterior. Através do MSM foi calculada a matriz de transição entre os 4 estados metaestáveis pelo método *coarse-graining*, conforme a seguir:

$$\begin{bmatrix} 9,33 \cdot 10^{-1} & 5,20 \cdot 10^{-3} & 6,14 \cdot 10^{-2} & 7,20 \cdot 10^{-63} \\ 2,82 \cdot 10^{-3} & 9,42 \cdot 10^{-1} & 5,47 \cdot 10^{-2} & 7,03 \cdot 10^{-175} \\ 4,86 \cdot 10^{-3} & 8,00 \cdot 10^{-3} & 9,79 \cdot 10^{-1} & 7,91 \cdot 10^{-3} \\ 1,28 \cdot 10^{-63} & 2,31 \cdot 10^{-175} & 1,78 \cdot 10^{-2} & 9,82 \cdot 10^{-1} \end{bmatrix}$$

De posse da matriz de transição, conforme a teoria das cadeias de Markov, é possível plotarmos o grafo de probabilidades entre os 4 macroestados. Na Figura 5.26 é apresentado o grafo com as probabilidades de transição entre os 4 macroestados:

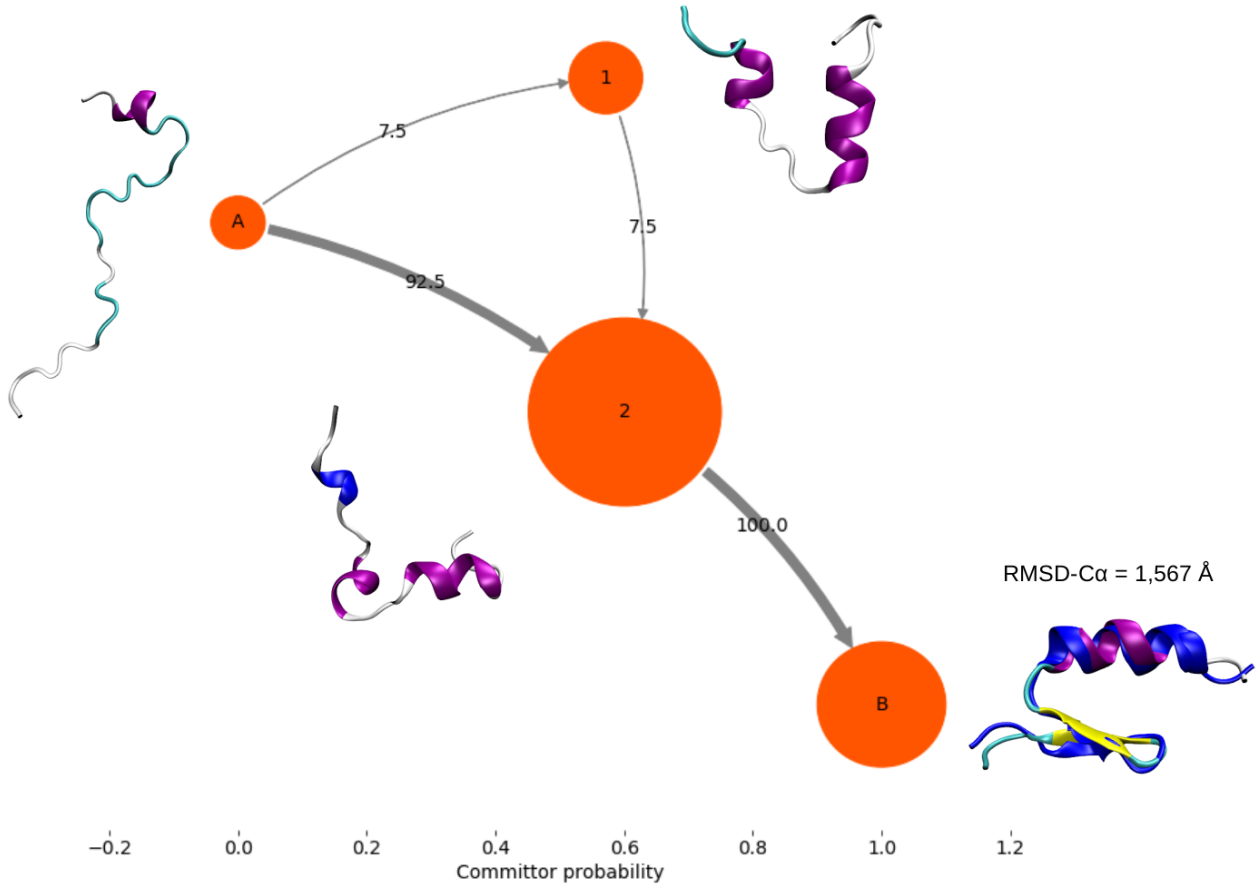
Figura 5.26: Grafo com 4 estados metaestáveis obtidos para a proteina BBA através do método *coarse-graining*.



Observamos pela análise do grafo que a proteína estando no estado zero ou no estado

um, apresenta alta probabilidade de ir para o estado 2, sugerindo que esse é um estado de transição para a proteína, uma vez que sabemos que o estado 3 é o de menor energia. Devido à configuração do programa pyEMMA, a matriz de transição começa a contar de zero a três (representação dos 4 macroestados). Na Figura 5.25 a contagem vai de S_1 até S_4. Desse modo o estado zero aqui, corresponde ao estado S_1 e assim sucessivamente. Desse modo, escolhemos o caminho de transição partindo do macroestado zero (*estado desenovelado*) para o macroestado 3 (*estado nativo*) a partir da TPT. Para facilitar a análise, as probabilidades de transição foram normalizadas para o percentural de 100%. Na Figura 5.27 é apresentado o caminho do estado de transição com uma estrutura representativa para cada macroestado:

Figura 5.27: Caminho de transição entre o estado desenovelado e o estado nativo para a proteína BBA.



A partir da análise da Figura 5.27, verificamos uma alta probabilidade (92%) da proteína sair do estado **A** (*estrutura desenovelada*) e seguir para o estado **2** (*estado de transição*). Do estado **2** para o estado **B** (*estrutura nativa*), a probabilidade foi de 100%. Logo, podemos inferir que essa proteína segue uma dinâmica de 3 estados. Percebemos uma rota alternativa em que a BBA passa primeiro pelo estado **1** para depois seguir para o estado **2**, porém essa probabilidade é muito baixa (7,5%). O RMSD-Cα entre a estrutura cristalográfica e o estado nativo foi de 1,567

Å, o que está condizente com dados da referência.[112] A estrutura nativa do estado **B** está representada sobreposta à estrutura cristalográfica (*em azul*).

Apesar de poucos macroestados, podemos inferir que primeiro formam-se as estruturas secundárias para depois seguir para o estado nativo, estando de acordo com o mecanismo de difusão-colisão. Vale destacar que na imagem está representada somente uma estrutura representativa de cada estado, mas na realidade o que temos é um *ensamble* de 100 conformações para cada macroestado. Na tabela 5.6 apresentamos os resultados para o fluxo do caminho, percentagem do caminho e as possíveis vias para o enovelamento da proteína BBA.

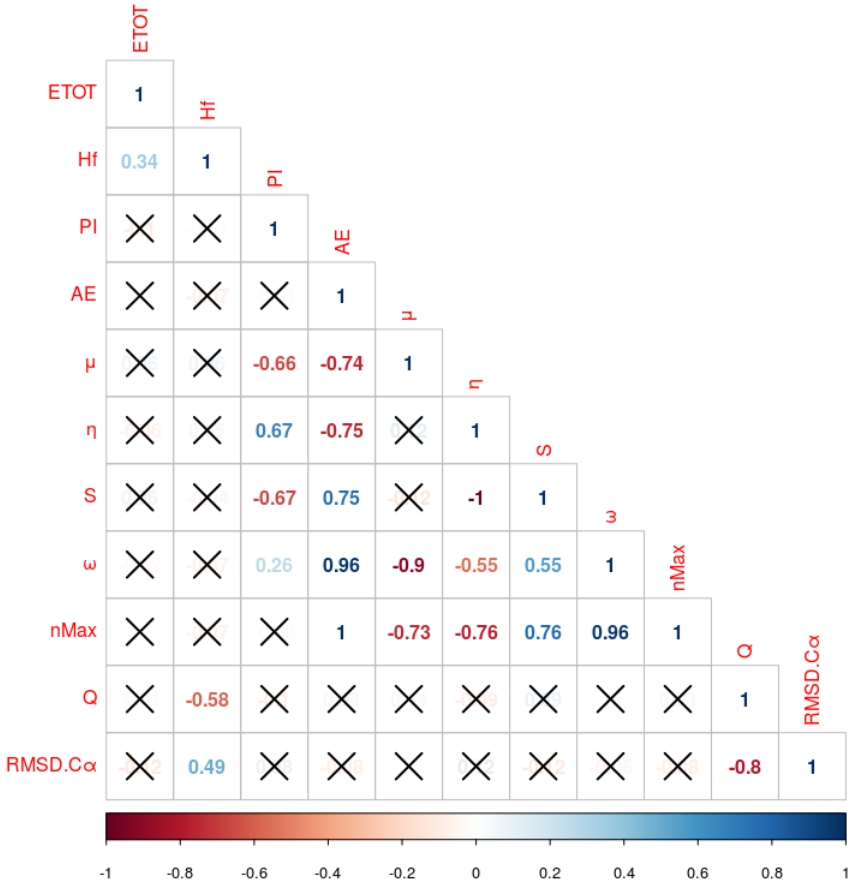
Tabela 5.6: Fluxo e percentagem de caminho para as diversas rotas do enovelamento da proteína BBA.

Proteína BBA		
Fluxo do caminho	**Percentagem do caminho**	**Caminho**
4,38E-06	92,5%	[0 2 3]
3,53E-07	7,5%	[0 1 2 3]

Conforme já mencionado, observamos que há somente dois caminhos possíveis para a proteína BBA: [0 2 3] com probabilidade de 92,5% e [0 1 2 3] com probabilidade de 7,5%.

A partir desse momento, realizamos o cálculo *single-point* com o método PM7 e modelo implícito de solvente COSMO com o pacote MOPAC para cada uma das 100 conformações representativas de cada macroestado da Figura 5.27, totalizando 400 conformações. Com os dados do MOPAC, obtivemos os valores de E_{TOT} e ΔH_f e calculamos os descritores de reatividade global com o programa PRIMORDIA. Além disso realizamos cálculos da fração de contatos nativos (Q) e do RMSD-Cα para cada uma das conformações. Na Figura 5.28 é apresentado o mapa de correlação entre os descritores globais, Q e RMSD-Cα.

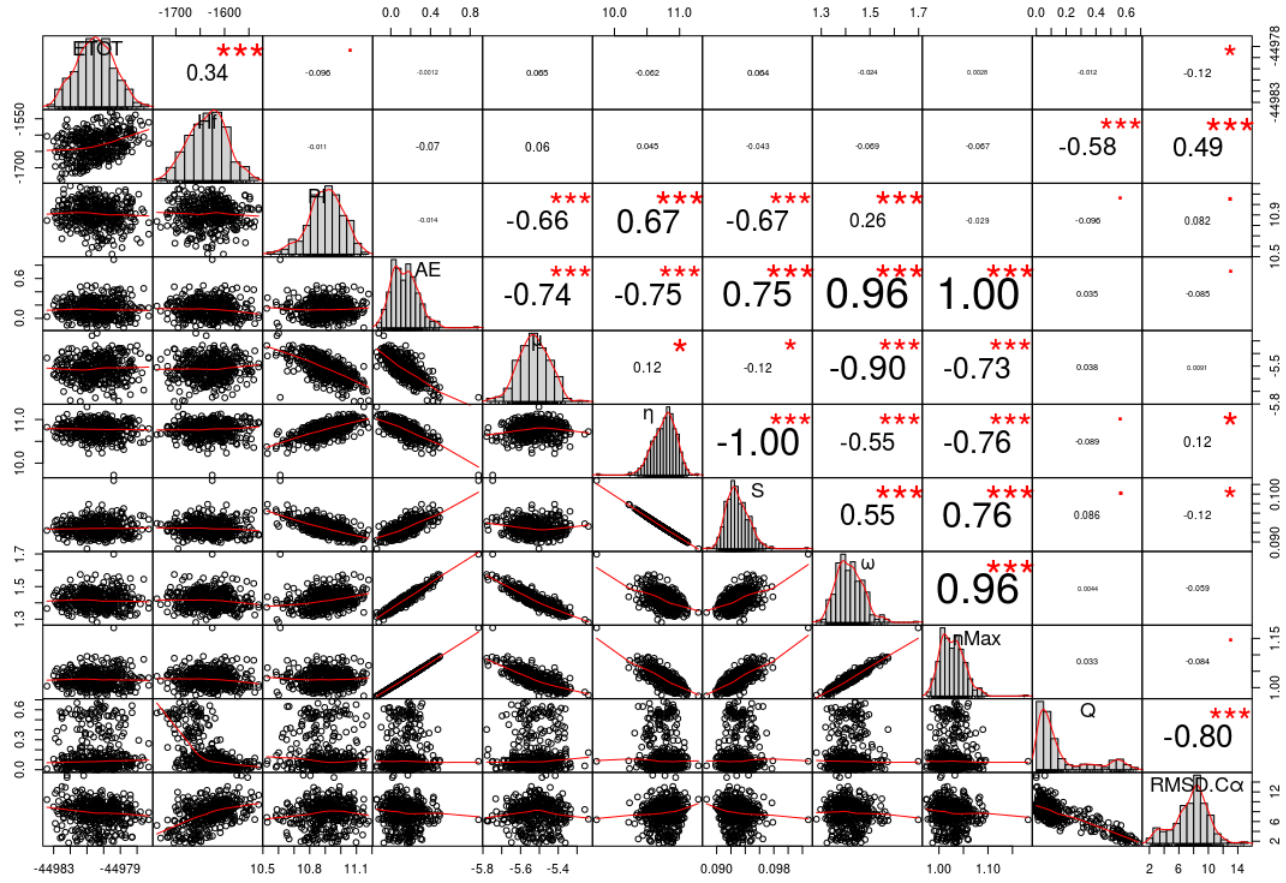
Figura 5.28: Mapa de correlação entre descritores globais de reatividade, fração de contatos nativos e RMSD-Cα de um MSM de 4 estados para a BBA.



Verificamos que os dados de correlação para a proteína BBA seguem a mesma tendência observada para a NTL9. A correlação entre Q e RMSD-Cα foi de (R = -0,8) e entre Q e ΔH_f foi (R = -0,58). Esses valores já eram esperados uma vez que ao descermos no funil de enovelamento, a fração de contatos nativos tende a aumentar e o RMSD-Cα tendo a diminuir. Também esperamos uma diminuição da entalpia de formação quando caminhamos para a estrutura nativa. Porém, do mesmo modo que para a NTL9, não observamos nenhuma correlação entre o RMSD-Cα ou Q com os descritores globais de reatividade.

Uma análise mais detalhada pode ser feita através do histograma com linhas de densidade e de tendência entre as variáveis do sistema. Esses dados são apresentados na Figura 5.29 apresentada a seguir:

Figura 5.29: Histograma com correlação, linhas de densidade e de tendência de um MSM de 4 estados para a BBA.



Através da análise do histograma, verificamos que a energia total apresenta uma correlação muito baixa em relação ao RMSD-Cα (R = -0,12), enquanto esperávamos que a energia total diminuísse ao passo que nos aproximamos da conformação nativa, ou seja, deveria apresentar uma alta correlação positiva com o RMSD-Cα e uma alta correlação negativa com Q. Se observarmos as duas últimas colunas verticais do histograma, verificamos que com exceção do ΔH_f, todas as outras propriedades apresentaram uma correlação próxima de zero tanto em relação ao RMSD-Cα quanto para Q.

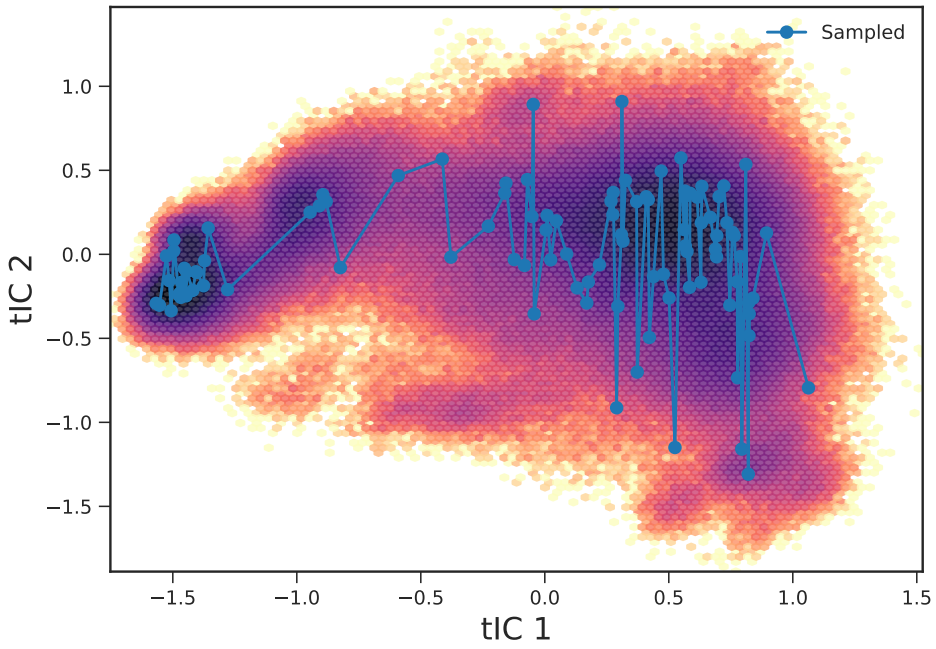
Olhando agora para as duas últimas linhas horizontais, observa-se que a linha de tendência é praticamente horizontal para todas os descritores globais de reatividade em relação ao RMSD-Cα e Q, com exceção do ΔH_f. Desse modo, de forma análoga ao ocorrido para a NTL9, a densidade eletrônica da BBA parece não variar ao longo do enovelamento. Esse afirmativa é muito desafiadora, uma vez que sai do padrão esperado. Tais dados reforçam que o processo do enovelamento é fenômeno extremamente complexo e desafiador.

5.2.4 Proteína BBA: Abordagem com o programa MSMBuilder

Analisando o mapa cinético da Figura 5.30, percebemos que de forma semelhante ao ocorrido com a NTL9, o eixo x da tICA parece ser uma boa coordenada de desenovelamento, partindo da estrutura nativa (*mais a esquerda*) para a estrutura desenovelada (*mais à direita do mapa cinético*). Desse modo, se estivermos interessados somente em encontrar um caminho para o enovelamento, podemos amostar uma quantidade de conformações ao longo do eixo x que teremos uma trajetória de desenovelamento ou enovelamento.

Essa abordagem é interessante por ser rápida, uma vez que não utiliza métodos de clusterização, construção do MSM e validação dos dados que elevam bastante o custo computacional. A desvantagem desse tipo de análise é que não é possível aplicarmos essa abordagem em qualquer proteína, mas somente para aquelas que apresentam um comportamento específico ao longo do eixo x ou do eixo y na duas primeiras coordenadas tICA. Além disso, nesse caso, temos somente um caminho para o enovelamento (de mais fluxo), não sendo possível avaliarmos possíveis caminhos e probabilidades de transição. Na Figura 5.30 é apresentada uma amostragem de 100 conformações ao longo da coordenada x da tICA para a proteína BBA:

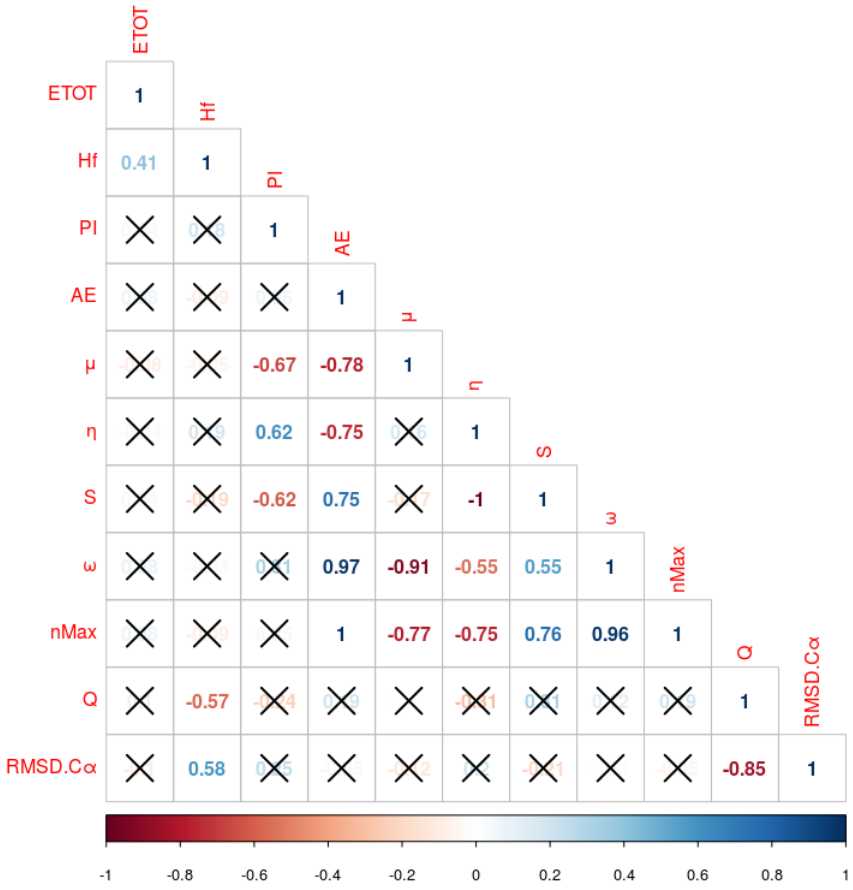
Figura 5.30: Amostra de 100 conformações ao longo da coordanada x da tICA para a BBA.



De posse dessas estruturas, realizamos o mesmo protocolo aplicado nas etapas anteriores para a obtenção do mapa de correlação entre os QCMDs globais, E_{TOT}, ΔH_f, RMSD-Cα e Q. Na Figura 5.31 é apresentado o mapa de correlação entre todas as variáveis avaliadas obtidas

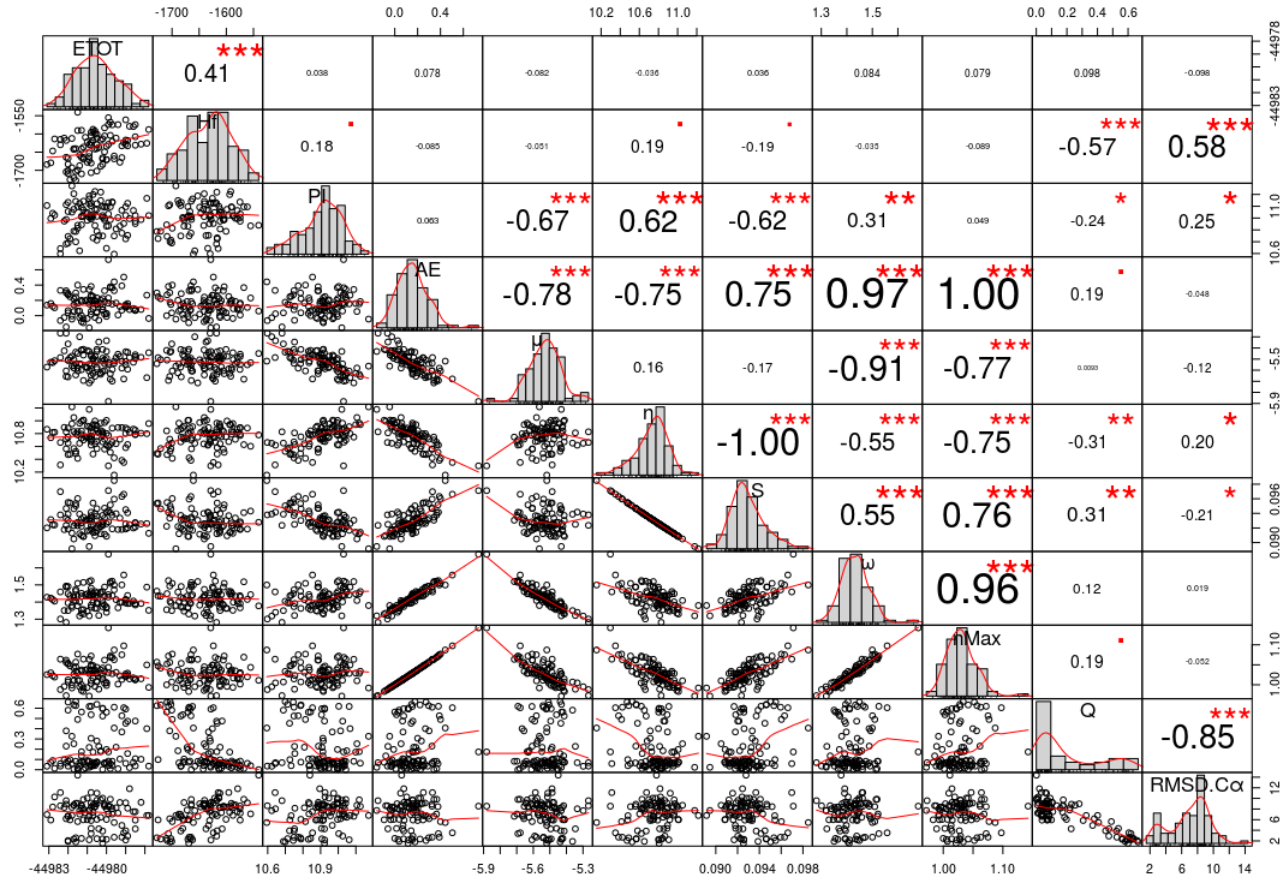
via método PM7. Os resultados dos QCMDs globais obtidos via DFT-3 para a proteína BBA podem ser consultados nas Figuras A.8 e A.11 do Apêndice A.

Figura 5.31: Mapa de correlação entre descritores globais de reatividade,E_{TOT}, ΔH_f, fração de contatos nativos e RMSD-Cα para a primeira coordenada tICA da BBA.



Percebemos que para esse novo caminho de enovelamento, a correlação entre RMSD-Cα e Q aumentou de (R = -0,80) na abordagem com MSM para (R = -0,85) na abordagem atual. A correlação entre ΔH_f e RMSD-Cα também aumentou de (R = 0,49) para (R = 0,58). Porém percebemos que não houve dependência entre os descritores globais de reatividade e as propriedades estruturais, seguindo tendência análoga ao observado para a NTL9. Para analisarmos melhor essas tendências, podemos utilizar o histograma dos dados. Na Figura 5.32 é apresentado o histograma com dados referentes às conformações da coordenada x da tICA para a proteína BBA:

Figura 5.32: Histograma com correlação, linhas de densidade e de tendência referente aos dados da coordenada x do mapa cinético para a BBA.



Verificamos que as variações são muito sutis quando comparadas com os dados da Figura 5.29. Ao analisarmos as duas últimas colunas da Figura 5.32, percebemos o mesmo comportamento já mencionado, em que com exceção do ΔH_f, as demais propriedades apresentam correlação muito baixa em relação ao RMSD-Cα ou Q. Ao observarmos as duas últimas linhas horizontais, percebemos que há flutuações mais acentuadas nas linhas de tendência, mas continuam apresentando baixos valores de correlação.

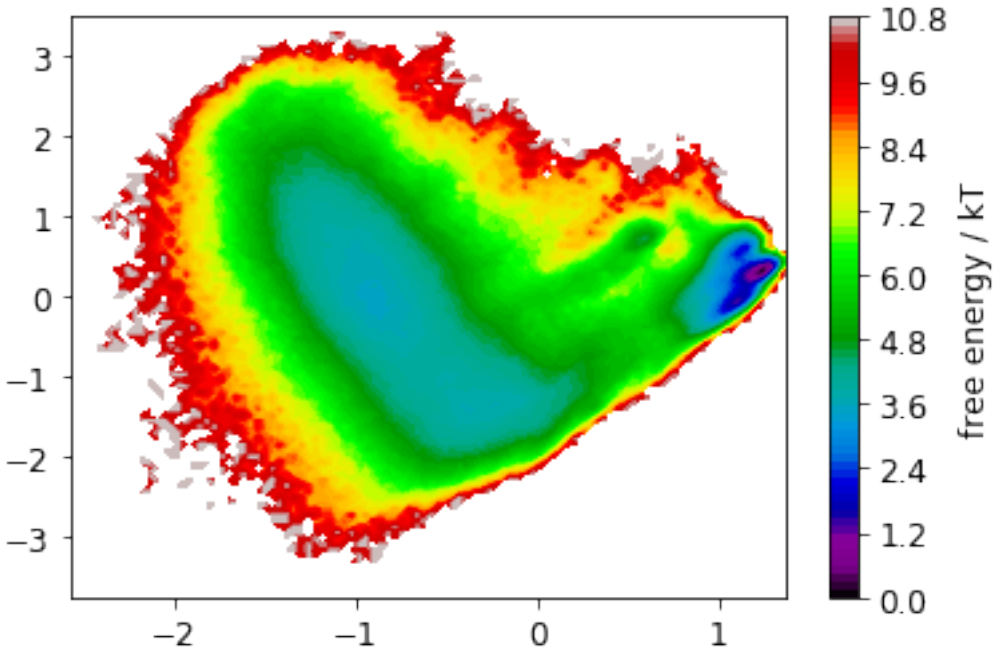
Logo verificamos mais uma vez que como para a proteína NTL9, os dados para a BBA apontam para a mesma direção, ou seja, os QCMDs globais não variam significativamente durante o processo de enovelamento, sendo que essa tendência parece ser independente da rota de enovelamento escolhida ao longo do funil de energia livre. Os resultados dos QCMDs obtidos via DFT-D3 apresentaram performance semelhante, ou seja, não foi observado correlação com descritores estruturais.

5.2.5 Proteína α3D: Abordagem com o program PyEMMA

A proteína α3D, é bem maior que as analisadas anteriormente, possuindo 1.140 átomos e 73 resíduos. A DM utilizada para esses estudo possui aproximadamente 346 μs, sendo que intervalo de registro dos *frames* foi de 0,2 ns, totalizando 1.732.710 conformações. Após a aplicação das mesmas *features* utilizadas para análise das proteínas anteriores, o nosso sistema ficou com 395 dimensões. Na etapa seguinte, aplicamos a tICA para reduzir a dimensionalidade do sistema, com um *lag time* de 10 ns e uma variância que explica 95 % dos nossos dados, reduzindo o sistema para 119 dimensões.

Observamos que mesmo após a redução de dimensionalidade, devido à complexidade da proteína, o sistema ainda ficou com muitas dimensões. Poderíamos escolher menos *features* como somente os ângulos de torção do *backbone* da proteína para termos menos dimensões, mas como queremos estabelecer correlação com a análise feita para as demais proteínas, mantivemos o mesmo protocolo para todas as análises. Na Figura 5.33, apresentamos o mapa de energia livre plotado a partir da tICA.

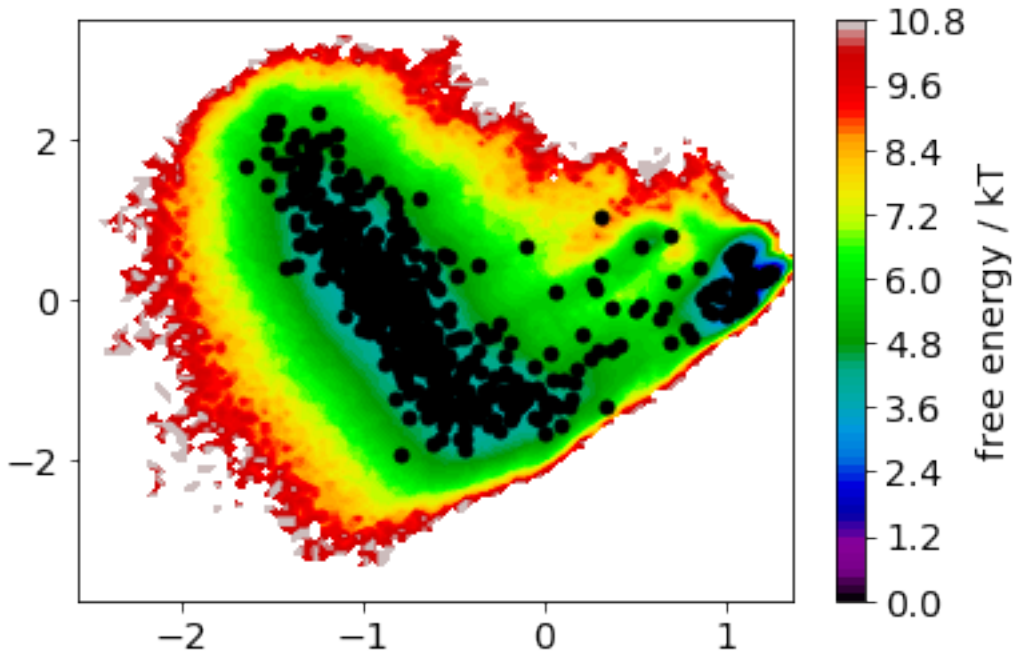
Figura 5.33: Mapa de cinético baseado na tICA para a proteína α3D.



No mapa da Figura 5.33, há uma escala de cores que varia do vermelho (*regiões de mais alta energia*) até o roxo (*região de mais baixa energia*). Analisando esse mapa, percebemos a presença de uma bacia bem definida na região central da Figura correspondente a um mínimo local, e outra bacia menor, bem à direita do mapa, correspondente ao mínimo global, uma vez que são observadas conformações de menor energia (*cor roxa*) nessa segunda bacia. Na etapa seguinte,

realizamos a clusterização com o método *k-means* com 550 *clusters* e pegamos o centroide de cada um desses para representar cada conjunto. Desse modo, o sistema passou a ter 550 estruturas representativas, sendo que a distribuição dessas conformações ao longo do mapa cinético é apresentado na Figura 5.34.

Figura 5.34: Centroide dos *clusters* com o método *k-means* e mapa cinético de fundo para a proteína α3D.

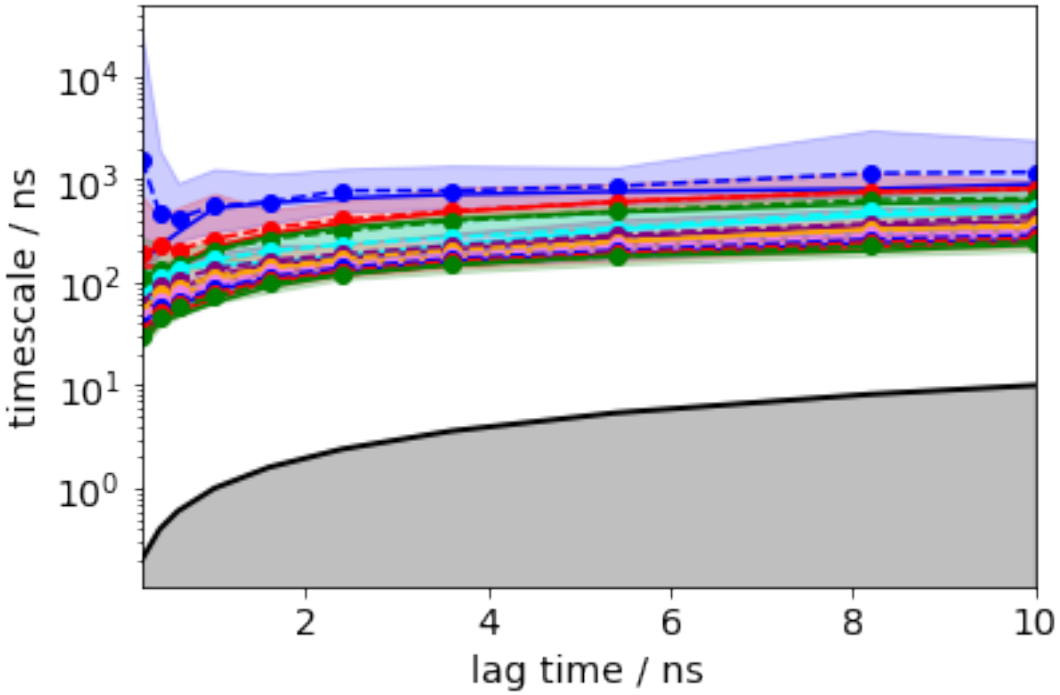


Verificamos que os *clusters* ocuparam mais a região central do mapa cinético, sendo que as áreas mais periféricas não foram cobertas. Aumentamos o número de *clusters* para 700, porém percebemos que não houve alteração na região de ocupação. Talvez fosse necessário amentarmos muito o número de conjuntos, mas para isso, nos deparamos com uma barreira computacional, uma vez que a memória necessária seria enorme (ordem de centenas de mega bytes de memória RAM) para realizarmos essa tarefa para mais de 1.000 agrupamentos. Como não disponibilizamos desse recurso, acabamos validando o nosso método com 550 *clusters*, que corresponde a um valor razoável para o tamanho do sistema.

Ressaltamos porém, que não temos certeza se o aumento do número de *clusters*, mudaria essa disposição encontrada, porém não testamos essa hipótese para valores superiores a 700 grupos. Para contornarmos esse problema, fizemos também uma análise das estruturas ao longo do eixo x com o programa MSMBuilder. Conforme mencionado anteriormente, podemos suspeitar que a coordenada x da tICA corresponde a uma trajetória de enovelamento, saindo da estrutura mais desenovelada (mais à esquerda) para a conformação nativa (mais à direita) da Figura 5.33. Na Figura 5.35 é apresentado o gráfico com as escalas de tempo de relaxação

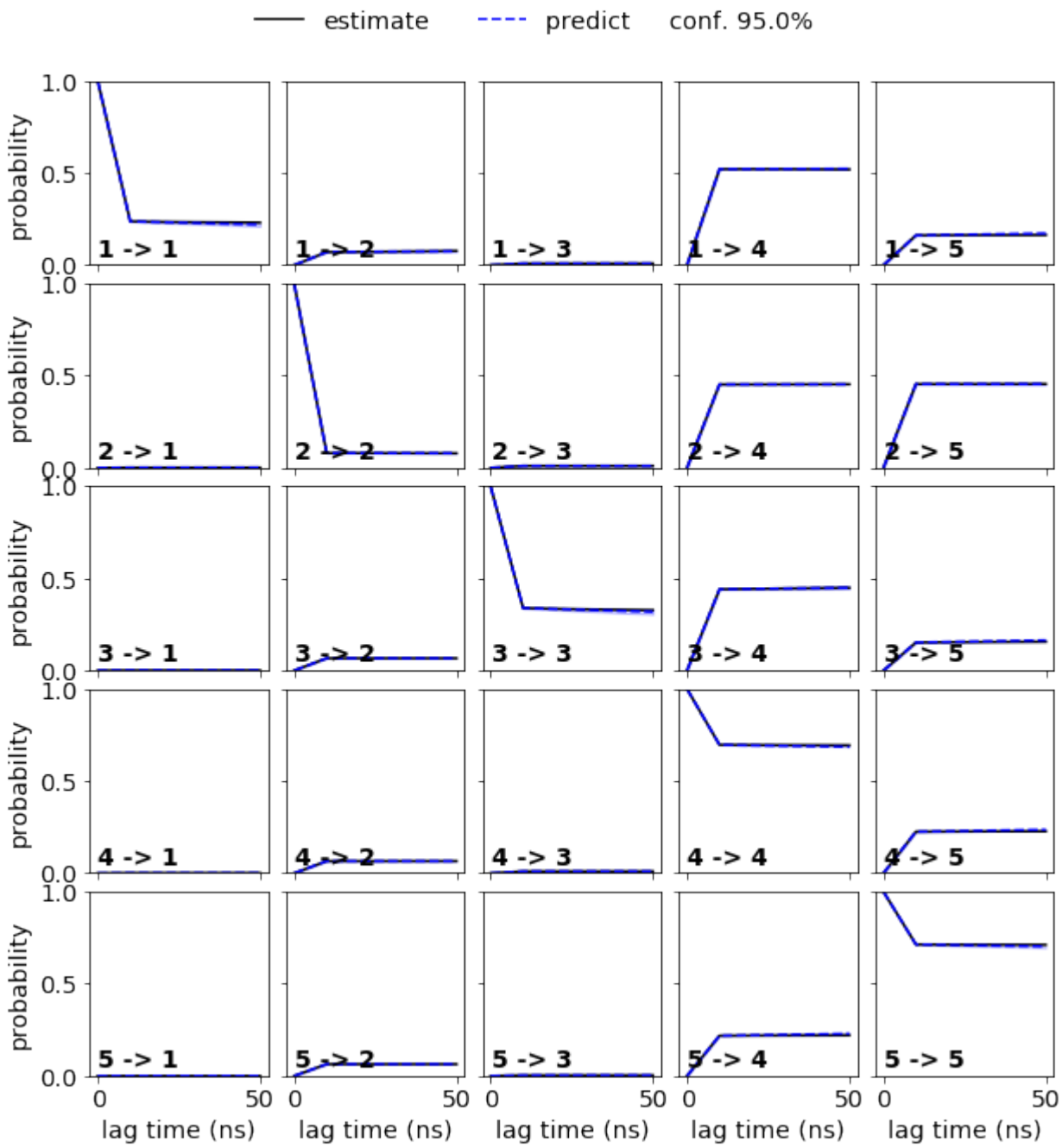
implícitas em função do *lag time*.

Figura 5.35: Escalas de tempo de relaxação implícitas para a proteína α3D.



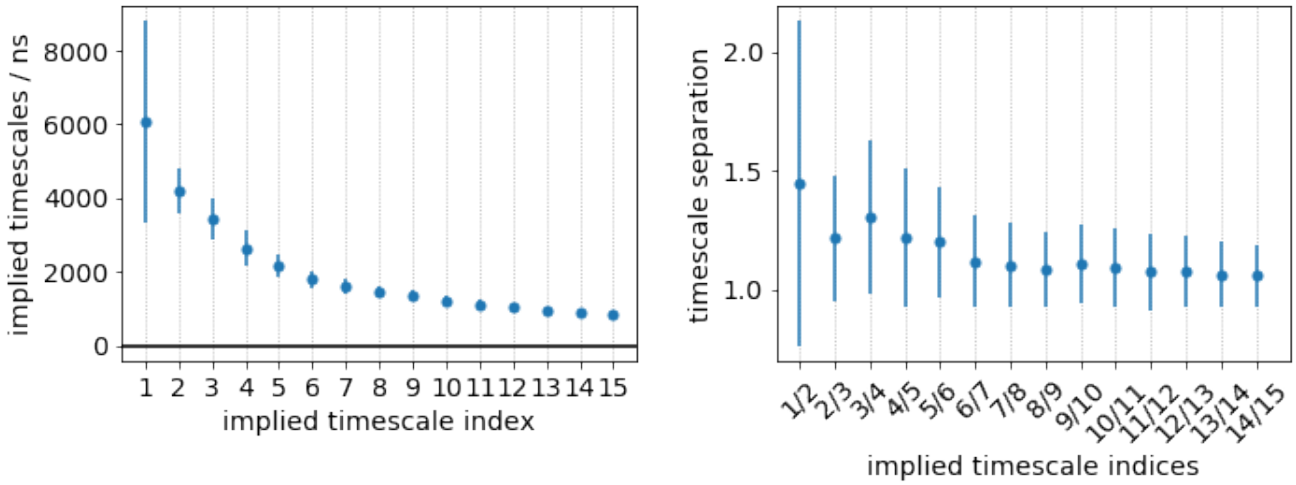
Observamos que de maneira análoga ao apresentado para a NTL9, quando o *lag time* assume valor acima de 2 *ns*, as escalas de tempo apresentam rápida convergência. Nos testes realizados, os melhores resultados para o MSM foram obtidos com tempo de latência de 10 *ns*. Desse modo, o MSM foi obtido com o lag time de 2 *ns*. Na Figura 5.36 é apresentado o teste de Chapman-Kolmogorov para um MSM com 5 estados para a proteína α3D.

Figura 5.36: Teste de Chapman-Kolmogorov para um MSM com 5 estados para a proteína BBA.



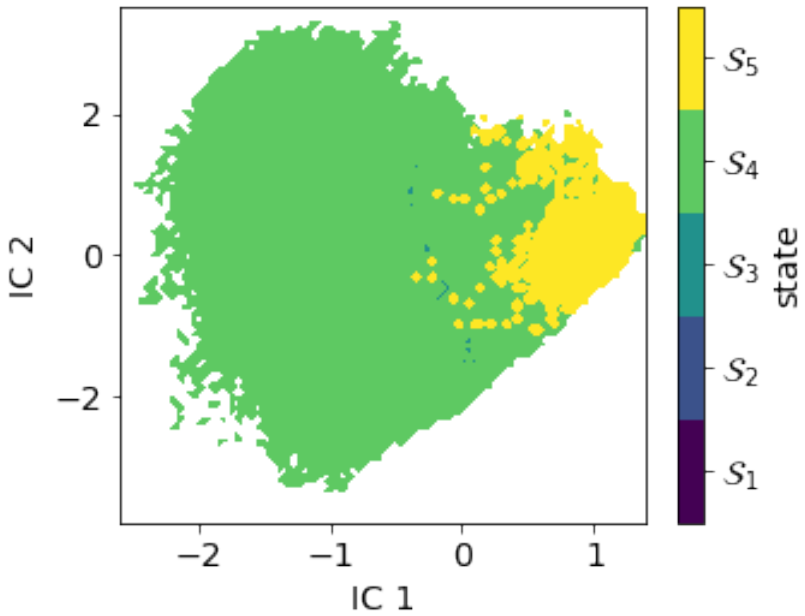
Da mesma forma como aconteceu com as proteínas NTL9 e BBA, o teste de Chapman-Kolmogorov foi validado para um MSM com 5 estados, sendo que o valor estimado (*linha preta*) e o valor predito (*linha azul*) apresentaram convergência muito boa entre si. Com o MSM é possível fazer uma avaliação do comportamento das escalas de tempo implícitas do sistema. Na Figura 5.37 é apresentado um gráfico com as escalas de tempo implícitas para a α3D:

Figura 5.37: Escalas de tempo implícitas para a proteína α3D.



A partir da análise da separação entre as escalas de tempo, percebemos que há uma entre o 4^o e o 5^o e entre o 5^o e o 6^o processos. Desse modo, podemos realizar um MSM com 5 ou 6 estados metaestáveis. Nos nossos testes o modelo com 5 bacias foi capaz de descrever melhor os estados de transição, logo utilizamos 5 macroestados para a aplicação dos métodos PCCA+ e *coarse-graining*. Na Figura 5.38 é apresentado o mapa cinético separado entre os 5 macroestados através do método PCCA+.

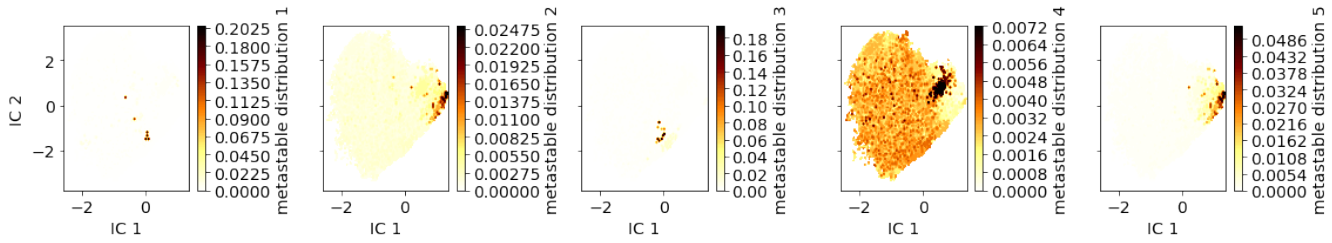
Figura 5.38: Modelo de 5 macroestados obtidos para a proteína α3D.



Na Figura 5.38, a superfície de energia foi separada em 5 estados metaestáveis, porém diferentemente do que foi observado para as proteínas NTL9 e BBA, as fronteiras de separação dos macroestados (S_1, S_2 e S_3) não estão bem definidas nas duas primeiras coordenadas

da tICA devido ao grande número de dimensões do sistema. Para melhor visualizarmos as informações desse sistema, na Figura 5.39 é mostrada a distribuição das estruturas para cada um dos 5 macroestados separadamente.

Figura 5.39: Distribuição das conformações por estado metaestável para a proteína α3D.

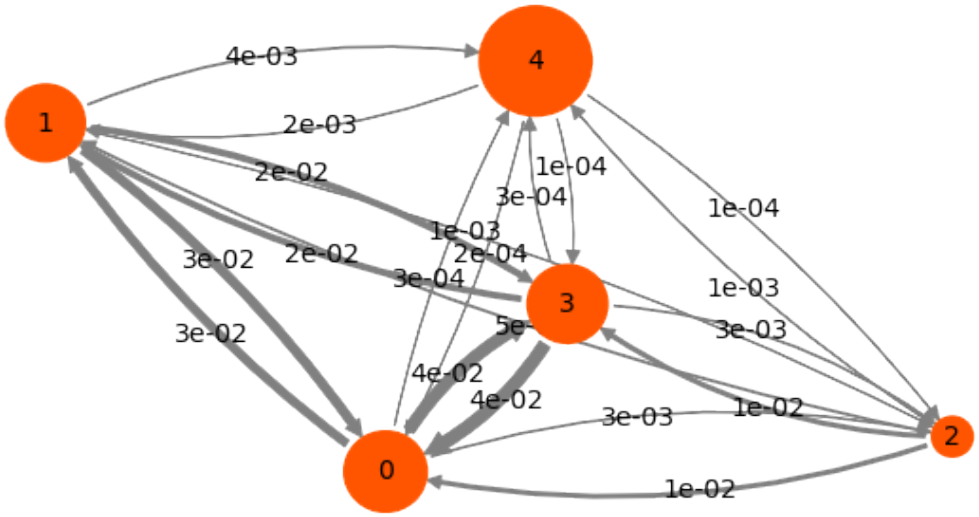


A partir da análise da Figura 5.39 percebemos que os estados metaestáveis (S_2 e S_3) são menos povoados justificando porque não visualizamos eles na Figura anterior. Após a aplicação do PCCA+, foram amostradas 100 conformações de cada macroestado para análises posteriores. A partir da análise das estruturas amostrados e das Figuras 5.38 e 5.39, inferimos que o estado desenovelado corresponde ao S_1 e o estado com a conformação nativa é o S_5. A seguir apresentamos a matriz de transição obtida a partir do MSM com 5 estados por meio do método *coarse-graining*:

$$\begin{bmatrix} 9,26 \cdot 10^{-1} & 2,83 \cdot 10^{-2} & 3,13 \cdot 10^{-3} & 4,21 \cdot 10^{-2} & 3,16 \cdot 10^{-4} \\ 3,11 \cdot 10^{-2} & 9,44 \cdot 10^{-1} & 1,37 \cdot 10^{-3} & 1,96 \cdot 10^{-2} & 4,01 \cdot 10^{-3} \\ 1,25 \cdot 10^{-2} & 4,99 \cdot 10^{-3} & 9,68 \cdot 10^{-1} & 1,31 \cdot 10^{-2} & 1,09 \cdot 10^{-3} \\ 4,49 \cdot 10^{-2} & 1,91 \cdot 10^{-2} & 3,49 \cdot 10^{-3} & 9,32 \cdot 10^{-1} & 2,81 \cdot 10^{-4} \\ 1,73 \cdot 10^{-4} & 2,00 \cdot 10^{-3} & 1,48 \cdot 10^{-4} & 1,44 \cdot 10^{-4} & 9,98 \cdot 10^{-1} \end{bmatrix}$$

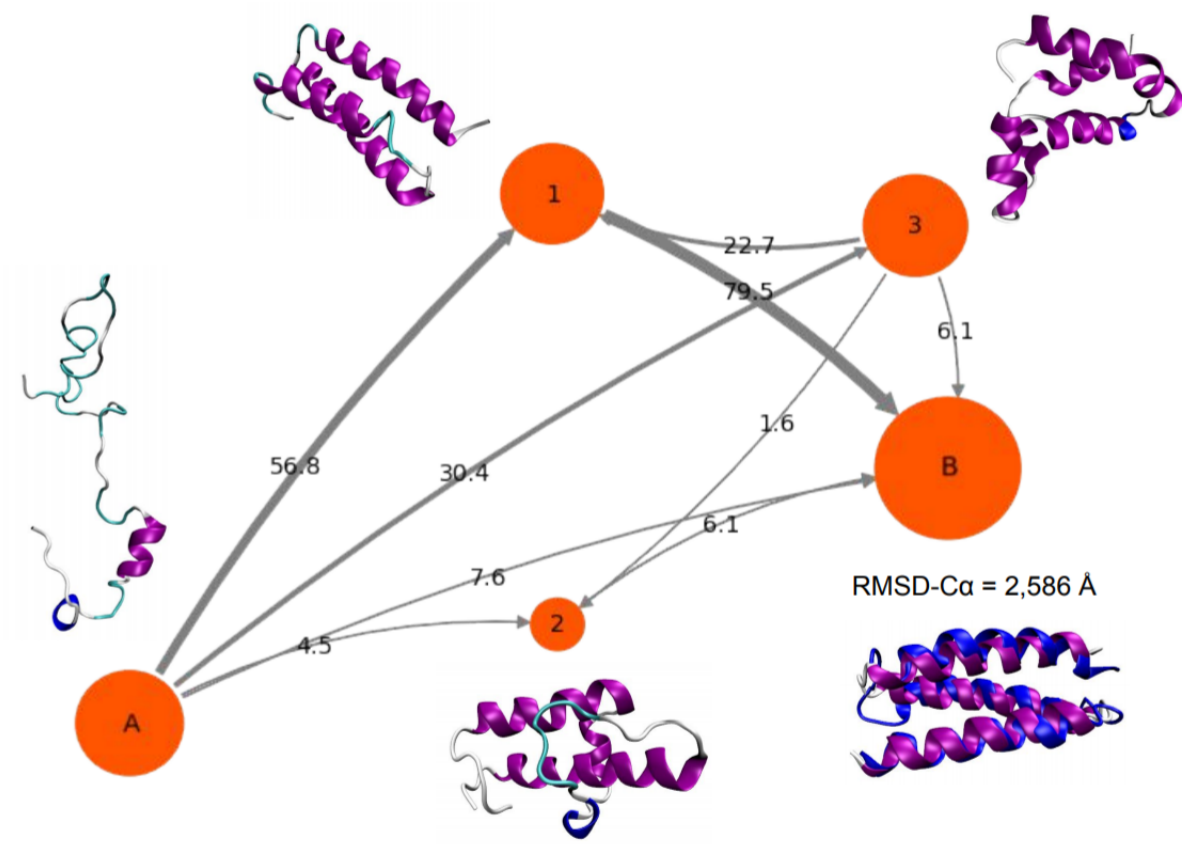
Com os dados da matriz de transição, podemos plotar o grafo de probabilidades de transição entre os 5 estados metaestáveis da proteína α3D conforme Figura 5.40 a seguir:

Figura 5.40: Grafo com 5 estados metaestáveis obtidos para a proteina α3D através do método *coarse-graining*.



Observamos que há uma alta probabilidade de transição do sistema estando no estado **zero** (S_1) seguir para os estados **1** (S_2) e **3** (S_4). Como sabemos que o estado **4** (S_5) é o de menor energia, podemos inferir que os estados **1** e **3**, são possíveis caminhos de transição. Para confirmarmos essa suspeita, podemos avaliar os caminhos de transição partindo de **0** $\rightarrow$ **4** através da TPT. Na Figura 5.41 apresentamos o caminho do estado de transição normalizado partindo do macroestado **zero** (*desenovelado*) para o **4** (*nativo*) com uma estrutura representativa para cada macroestado.

Figura 5.41: Caminho do estado de transição entre o estado desenovelado e nativo para a proteína α3D.



Observamos que o caminho de maior probabilidade passa por três estados, saindo de **A** (*estado desenovelado*), passando por **1** (*estado de transição*) e seguindo para **B** (*estado enovelado*), sugerindo que essa proteína apresenta uma via de três estados para o enovelamento. O RMSD-Cα entre o conformação nativa e a estrutura cristalográfica foi de 2,586 Å, estando de acordo com dados da referência.[112] A Figura representativa do estado metaestável **B** está representado em sobreposição à estrutura cristalográfica (*em azul*).

O segundo caminho de maior probabilidade passa por outro possível estado de transição (**3**), porém apresenta baixa probabilidade de seguir para o estado nativo (6,1%), preferindo retornar para (**1**) com 22,7% de probabilidade e depois seguir para o estado nativo. Se compararmos as estruturas representativas dos macroestados (**3**) e (**1**), percebemos que (**3**) possui as hélices mais bem formadas, sugerindo que enquanto as estruturas secundárias não estiverem bem formadas, a proteína não segue para o estado nativo. Esse dado está de acordo com observações feitas por Sborgi *et al.* para a proteína gpW na referência.[168]

Desse modo, a α3D parece seguir o mecanismo de difusão-colisão, sendo que as três hélices α se formam primeiro, seguido da ancoragem entre si, formando a estrutura terciária até chegar ao estado nativo. Os demais caminhos apresentam baixa possibilidade de ocorrência.

Na tabela 5.7 são apresentados os dados para o fluxo do caminho, percentagem do caminho e as possíveis rotas para o enovelamento da proteína α3D:

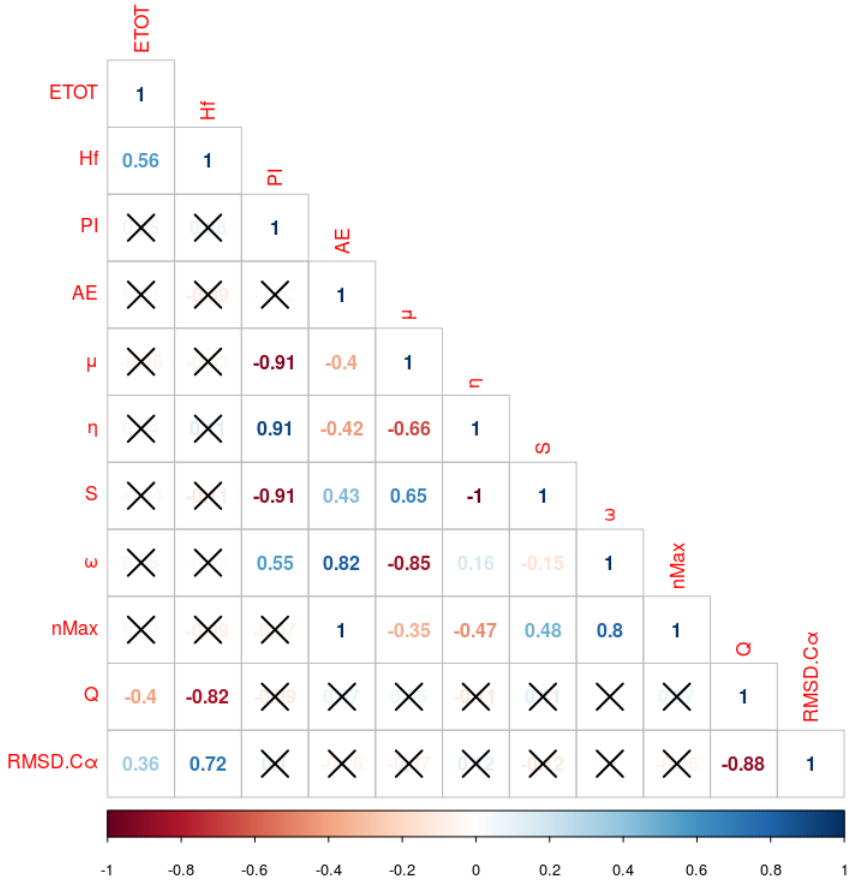
Tabela 5.7: Fluxo e percentagem de caminho para as diversas rotas do enovelamento da proteína α3D.

Proteína α3D		
Fluxo do caminho	**Percentagem do caminho**	**Caminho**
9,63E-06	56,8%	[0 1 4]
3,84E-06	22,7%	[0 3 1 4]
1,28E-06	7,6%	[0 4]
1,04E-06	6,1%	[0 3 4]
7,64E-07	4,5%	[0 2 4]
2,71E-07	1,6%	[0 3 2 4]

De acordo com os dados da tabela 5.7, há 6 vias para o enovelamento da α3D. O caminho de maior fluxo [0 1 4] possui uma probabilidade de 56,8% e o segundo caminho com maior probabilidade de ocorrência [0 3 1 4] apresenta um percentual de 22,7%. Além desses dois caminhos mais prováveis, percebemos 4 vias alternativas para o enovelamento: [0 4], [0 3 4], [0 2 4] e [0 3 2 4] com probabilidades de 7,6%, 6,1%, 4,5% e 1,6% respectivamente.

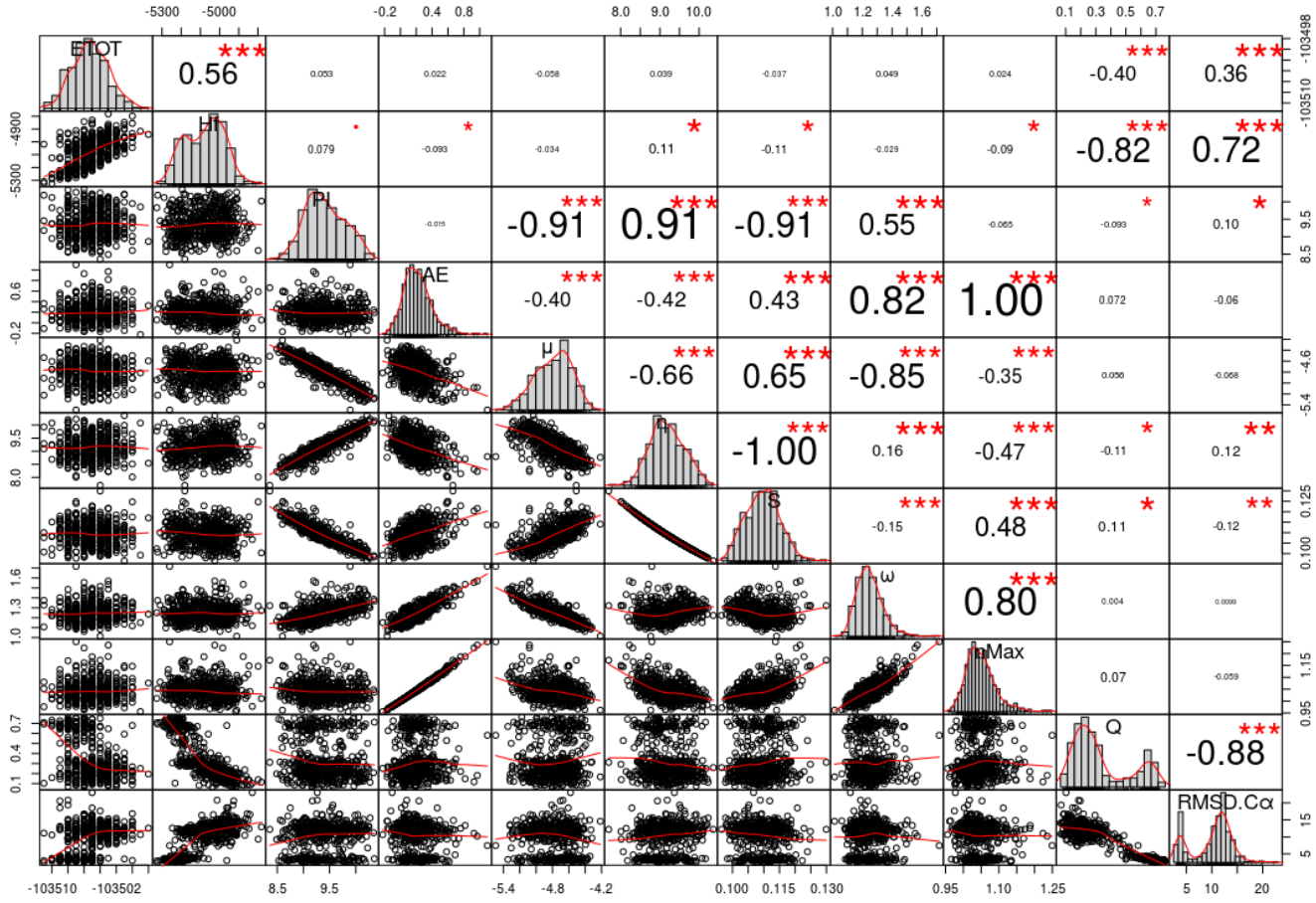
Percebemos portanto que o MSM com 5 estados para a α3D forneceu diversos *insights* importantes para a compreensão do mecanismo de enovelamento. Na Figura 5.42 é apresentado o mapa de correlação entre os descritores globais, E_{TOT}, ΔH_f, Q e RMSD-Cα para as 500 conformações amostradas dos macroestados da proteína α3D:

Figura 5.42: Mapa de correlação entre descritores globais de reatividade, E_{TOT}, ΔH_f, Q e RMSD-Cα de um MSM de 5 estados para a α3D.



Através da análise dos dados da Figura 5.42, percebemos que esse segue a mesma tendência geral observada para as proteínas NTL9 e BBA. Porém, a correlação entre Q e ΔH_f foi a melhor (R = -0,82) dentre as três proteínas avaliadas. A dependência entre o RMSD-Cα e ΔH_f também apresentou valores mais elevados (R = 0,72) para a α3D. Esses dados reforçam que o ΔH_f corresponde a uma boa função de pontuação, estando de acordo com o obtido na referência.[95] Os valores mais acentuados apontam que a contribuição entrópica para a α3D é menos pronunciada que para a NTL9 e BBA, ou seja, o ΔH_f apresenta valores próximos aos de energia livre. Desse modo, o calor de formação apresentou-se como uma boa coordenada de reação para o MSM de 5 macroestados. Para maior detalhamento das informações, na Figura 5.43 é apresentado o histograma com linhas de densidade e de tendência para todos os descritores avaliados:

Figura 5.43: Histograma com correlação, linhas de densidade e de tendência para um MSM de 5 estados da α3D.



Apesar das boas correlações entre Q, RMSD-Cα e Q, os descritores globais de reatividade não apresentaram associações significativas. Percebemos um aumento considerável na correlação entre dados de RMSD-Cα e de Q em função da E_{TOT} com (R = 0,36) e (R = -0,40) respectivamente. Apesar da melhoria, esses resultados são pouco expressivos.

Se olharmos para as gráficos com linhas de tendência entre as variáveis (lado esquerdo da Figura 5.43), percebemos que há uma variação muito pequena dos descritores globais ao longo do enovelamento. Portanto, de modo similar ao encontrado para a NTL9 e BBA, os descritores globais de reatividade, baseados na densidade eletrônica, não são boas coordenadas de reação para o processo do enovelamento das proteínas avaliadas nesse trabalho. Esses dados mais uma vez, estão de acordo com o observado por Milosvljevic, A. R. *et al.*[87] em que os orbitais de fronteira estão fortemente localizados ao longo do enovelamento.

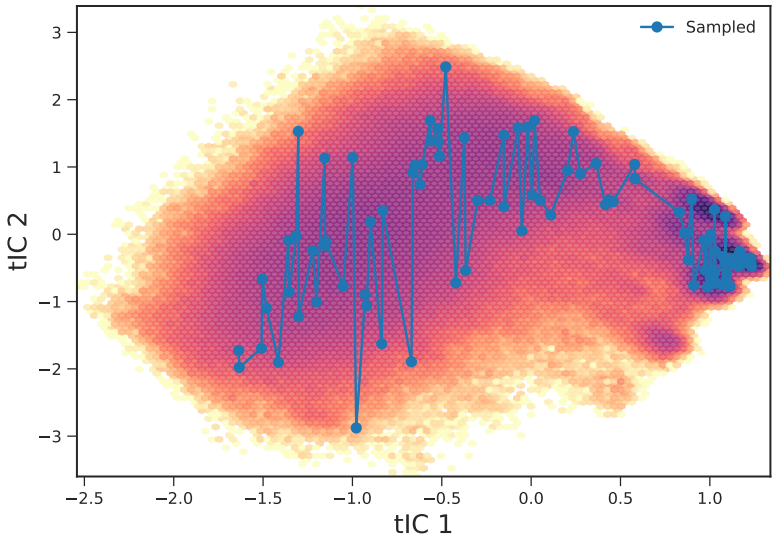
Esses dados, sugerem que esse comportamento se estenda para qualquer outro conjunto de proteínas. Ressaltamos entretanto, que as proteínas avaliadas são de rápido enovelamento

(*se dobram na ordem de microssegundos*), todavia não sabemos se isso é verdade para proteínas de enovelamento lento com a ubiquitina (*se enovela na ordem de segundos*).

5.2.6 Proteína α3D: Abordagem com o programa MSMBuilder

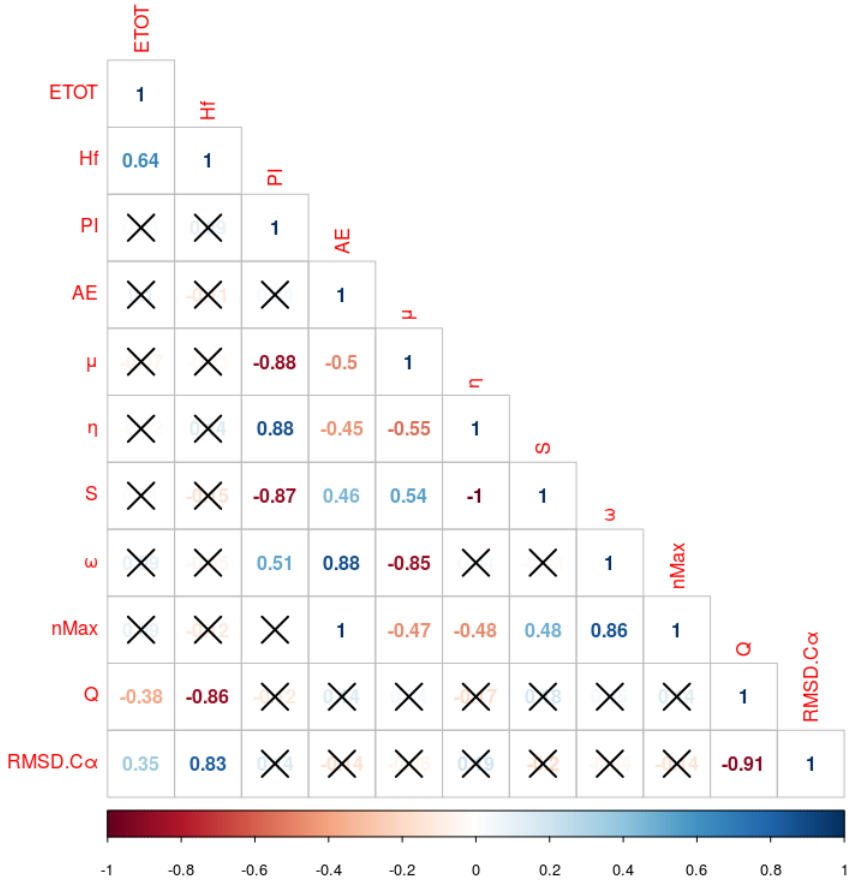
Aplicando o mesmo protocolo utilizado para as proteínas NTL9 e BBA, fizemos a amostram de 100 estruturas do longo da primeira coordenada tICA. Percebemos entretanto, conforme a Figura 5.33, que diferentemente das duas proteínas avaliadas anteriormente, a conformação nativa está mais à direita. Desse modo, podemos suspeitar que a coordenada x seja uma trajetória de enovelamento. Na Figura 5.44 é apresentada a amostra de 100 conformações ao longo de tIC1 como o mapa cinético de fundo.

Figura 5.44: Amostra de 100 conformações ao longo da coordenada tIC1 para a α3D.



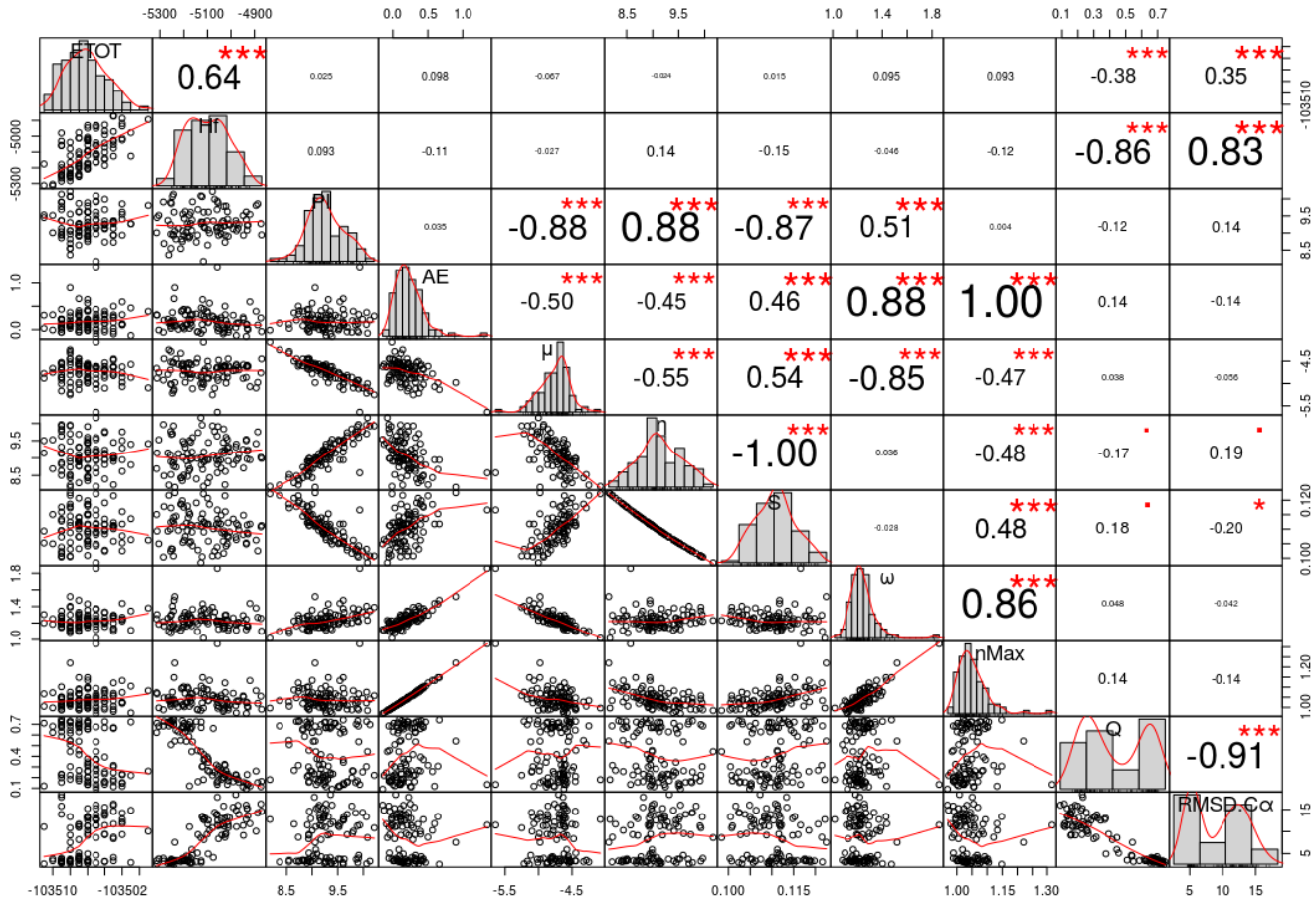
Após essa etapa, usamos o mesmo protocolo aplicado nas proteínas NTL9 e BBA para obtenção dos QCMDs globais Os resultados apresentados são obtidos via método PM7. Os resultados obtidos via DFT-D3 para a proteína α3D podem ser consultados nas Figuras A.9 e A.12 do Apêndice A. Na Figura 5.45 é exibido o mapa de correlação entre os descritores globais de reatividade: E_{TOT}, ΔH_f, RMSD-Cα e Q.

Figura 5.45: Mapa de correlação entre descritores globais de reatividade, E_{TOT}, ΔH_f, Q e RMSD-Cα de MSM de 5 estados para a α3D.



A partir dos dados da Figura 5.45, podemos verificar que a correlação entre Q e ΔH_f foi de (-0,86) e entre o RMSD-Cα e ΔH_f foi de (0,83). Isso confirma que independente do caminho de enovelamento, o ΔH_f corresponde a uma boa coordenada de reação para o enovelamento da proteína α3D. Porém mais uma vez, não podemos dizer o mesmo para os descritores globais associados ao hamiltoniano semi-empírico. O histograma da Figura 5.46 aponta mais detalhes de correlação entre os descritores avaliados:

Figura 5.46: Histograma com correlação, linhas de densidade e de tendência para a coordenada x da tICA para a proteína α3D.



Ao analisarmos os valores das duas últimas colunas, verificamos que Q e RMSD-Cα apresentam uma correlação de (-0,91), conforme é esperado para essas duas propriedades. Porém verificamos as correlações de Q e RMSD-Cα com os descritores PI, AE, μ, η, S, ω e n_{Max} apresentaram valores próximos de zero. Se olharmos para as duas últimas linhas da Figura 5.46, verificamos flutuações nas linha de tendência (*em vermelho*), apontando o comportamento aleatório dos dados ao longo do enovelamento para a maior parte dos descritores avaliados.

Como os QMCDs globais não apresentaram correlação com os descritores estruturais RMSD-C$_\alpha$ e Q para nenhuma das três proteínas estudas, o próximo passo foi realizarmos a análise dos QCMDs locais (por resíduo), obtidos com o programa PRIMoRDIA.[106]

5.2.7 Avaliação dos Descritores Químico-Quânticos Moleculares locais (QCMDs)

Nossa principal discussão sobre o papel da estrutura eletrônica no enovelamento de proteínas está focada nos QCMDs locais, que foram obtidos para todos os resíduos de aminoácidos

ao longo das trajetórias obtidas através da coordenada tICA com o programa MSMBuilder, conforme descritos no capítulo 4. Como os resultados obtidos via DFT e PM7 foram semelhantes, toda a discussão dos resultados foi baseada em cálculos DFT-D3 e os resultados via PM7 são apresentados Apêndice A. A discussão dos resultados é apresentada de forma conjunta para as três proteínas estudadas.

Foram gerados mapas de calor para todos os QCMDs locais listados no capítulo 4 para todas as 100 estruturas de cada proteína e os comparamos com os mapas de calor das propriedades estruturais por resíduo: RMSF, RMSD e DSSP.

Os mapas de calor mostram que os QCMDs locais EAS, NAS, RAS, orbitais moleculares de fronteira e eletrofilicidade não apresentaram padrões relevantes ao longo das trajetórias de enovelamento/desenovelamento para as três proteínas estudadas. No entanto, a dureza local e a densidade eletrônica mostraram comportamento interessante, que pode ser correlacionado com propriedades estruturais como RMSD, RMSF e DSSP. Os mapas de calor obtidos via dureza local e densidade eletrônica apresentaram o mesmo padrão, porém os dados de dureza local proporcionam melhor visualização, apresentando maior sensibilidade à pequenas variações de valores entre os resíduos de aminoácidos. Desse modo, a nossa discussão foi baseada na dureza local e os resultados da densidade eletrônica podem ser vistos nas Figuras A.18, A.19, A.27 e A.28 (Apêndice A).

Podemos observar na Figura 5.47 (a) que existem regiões claramente apresentando maiores valores de dureza do que outras, mais especificamente regiões de hélice-α e folhas-β. No entanto, não é homogêneo, pois há menos resíduos duros nas estruturas secundárias. Também descobrimos que ocorrem variações nas regiões mais duras da proteína ao longo do desenovelamento.

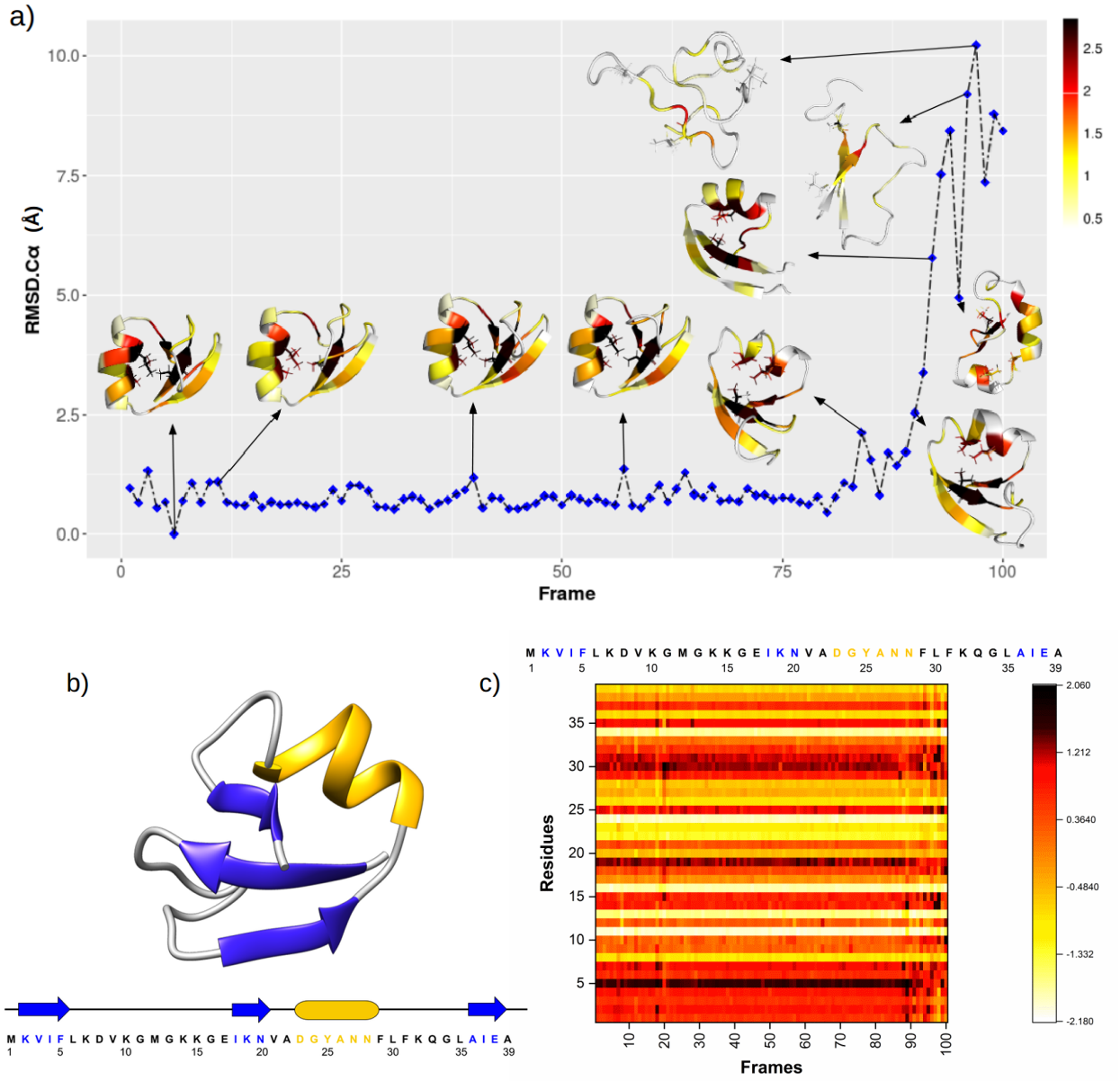


Figura 5.47: (a) Representação de 10 conformações ao longo do RMSD-Cα da trajetória obtida da coordenada x da tICA para o desenovelamento da NTL9. A dureza local é representada como uma paleta de cores aplicada ao *backbone* representado, em que a cor preta representa as regiões mais duras e a cor branca, as regiões mais moles; (b) Representação estrutural do estado nativo da proteína NTL9 (PDB-ID: 2HBA). As estruturas secundárias são representadas por cores diferentes. Em azul, são apresentadas as estruturas das folhas-β e em dourado, é apresentada a estrutura da hélice-α. Além disso, é apresentada a sequência da estrutura primária para a proteína NTL9, onde os aminoácidos pertencentes às estruturas secundárias são destacados por meio de cores e figuras geométricas; (c) Mapa de calor da dureza local para as 100 conformações do caminho de desenovelamento da NTL9.

Ao analisarmos a Figura 5.47 (c), observamos que o padrão de dureza muda à medida que o RMSD-C$_\alpha$ diminui. Quando a proteína NTL9 assume RMSD-C$_\alpha$ $\leq$ 2,5 Å(*frame* 90), a

dureza local assume valores mais estáveis, ou seja, a moleza e a dureza dos resíduos param de flutuar. Desta forma, podemos distinguir entre estruturas não nativas (*frames* 100-91) e estruturas semelhantes à nativa (*frames* 90-1), revelando uma visão muito mais detalhada da superfície de energia livre.

É possível prever se a estrutura está descendo no funil de energia livre em direção à estrutura nativa monitorando os padrões de dureza por resíduo. Observou-se que o comportamento de estabilização da dureza local se repete para RMSD, RMSF e DSSP (Figuras A.13 (a), A.14 (a) e A.15 do Apêndice A). Da direita para a esquerda (enovelamento), o RMSD e o RMSF diminuem rapidamente, assumindo valores estáveis por resíduo do frame 90 ao frame 1, correspondendo ao segmento da trajetória onde a dureza local para de flutuar.

A formação das estruturas secundárias hélice-α e folhas-β parece seguir o mesmo padrão, ou seja, a formação de todas as estruturas secundárias destacadas na Figura 5.47 (b) começa no frame 90 e segue com pequenas variações até o *frame* 1.

Ao contrário do que foi observado para a proteína NTL9, a superfície de energia livre da proteína BBA é mais rugosa, com a proteína visitando estados metaestáveis intermediários ao longo de uma via de enovelamento. De forma semelhante ao observado no NTL9, verificamos que a dureza local está mais concentrada nas estruturas secundárias, porém os resíduos nas hélices-α parecem apresentar uma dureza local maior que os resíduos nas folhas-β.

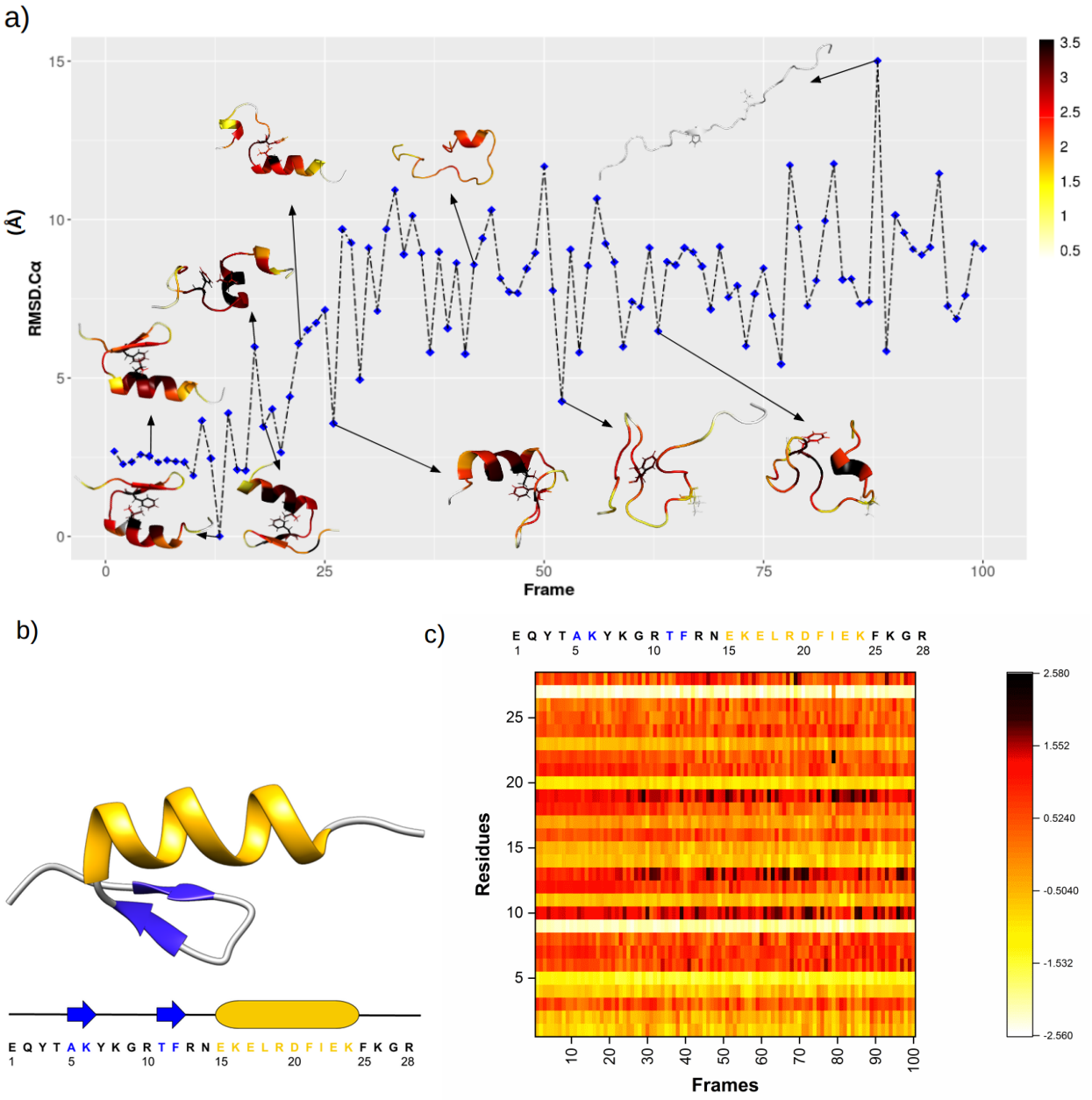


Figura 5.48: (a) Representação de 10 conformações ao longo do RMSD-Cα da trajetória obtida da coordenada x da tICA para o desenovelamento da proteína BBA. A dureza local é representada como uma paleta de cores aplicada ao *backbone* representado, em que a cor preta representa as regiões mais duras e a cor branca, as regiões mais moles; (b) Representação estrutural do estado nativo da proteína BBA (PDB-ID: 1FME). As estruturas secundárias são representadas por cores diferentes. As estruturas secundárias são representadas por cores diferentes. Em azul, são apresentadas as estruturas das folhas-β e em dourado, é apresentada a estrutura da hélice-α. Além disso, é apresentada a sequência da estrutura primária da proteína BBA, onde os aminoácidos pertencentes às estruturas secundárias são destacados por meio de cores e figuras geométricas. (c) Mapa de calor da dureza local para as 100 conformações do caminho de desenovelamento da BBA.

O mapa de calor da Figura 5.48 (c) juntamente com os dados das estruturas secundárias (Figura A.16 do Apêndice A), observamos que há uma maior estabilidade da dureza do *frame* 25 ao *frame* 1, onde a hélice-α é parcialmente formada (resíduos: 15-19) e uma folha β é formada (resíduos: 10-12). Neste ponto, o RMSD-C_α se reduz a um estado de mínimo local (*frame* 20), onde temos a formação completa da hélice-α (resíduos: 15-24) e a formação de duas folhas-β (resíduos: 3-4 e 11-12).

A partir do *frame* 19, a folha-β migra dos resíduos 3-4 para os resíduos 5-7, apresentando semelhança com a posição da estrutura cristalográfica (resíduo: 5-6). Além disso, observamos que o RMSD-C_α aumenta a partir do *frame* 19, onde as duas folhas- *beta* (resíduos: 5-7 e 10-12) são formadas e a hélice-α é parcialmente formada (resíduos: 15 -21) e diminui novamente a partir do *frame* 16, quando além das folhas-β, há um aumento no número de resíduos da hélice-α (15-23), atingindo o estado de energia mais baixa no *frame* 13 quando o número de resíduos da hélice-α é maximizado (15-25) e a dureza local permanece estável até o final da trajetória (*frame* 1). Desta forma, as regiões de dureza locais podem ser divididas em duas partes: 1) estruturas não nativas (*frame* 100-26) e estruturas semelhantes à nativa (*frame* 25-1), como pode ser visto pela mudança no padrão de dureza local da Figura 5.48 (c).

Na Figura 5.49 (a), vemos uma tendência semelhante à observada para as proteínas NTL9 e BBA, ou seja, embora a dureza global não varie muito ao longo da trajetória, verificamos que a dureza local muda consideravelmente ao longo do enovelamento. Podemos observar que regiões de estruturas secundárias apresentam maior dureza local, mas não há homogeneidade, pois existem regiões desenoveladas com alta dureza.

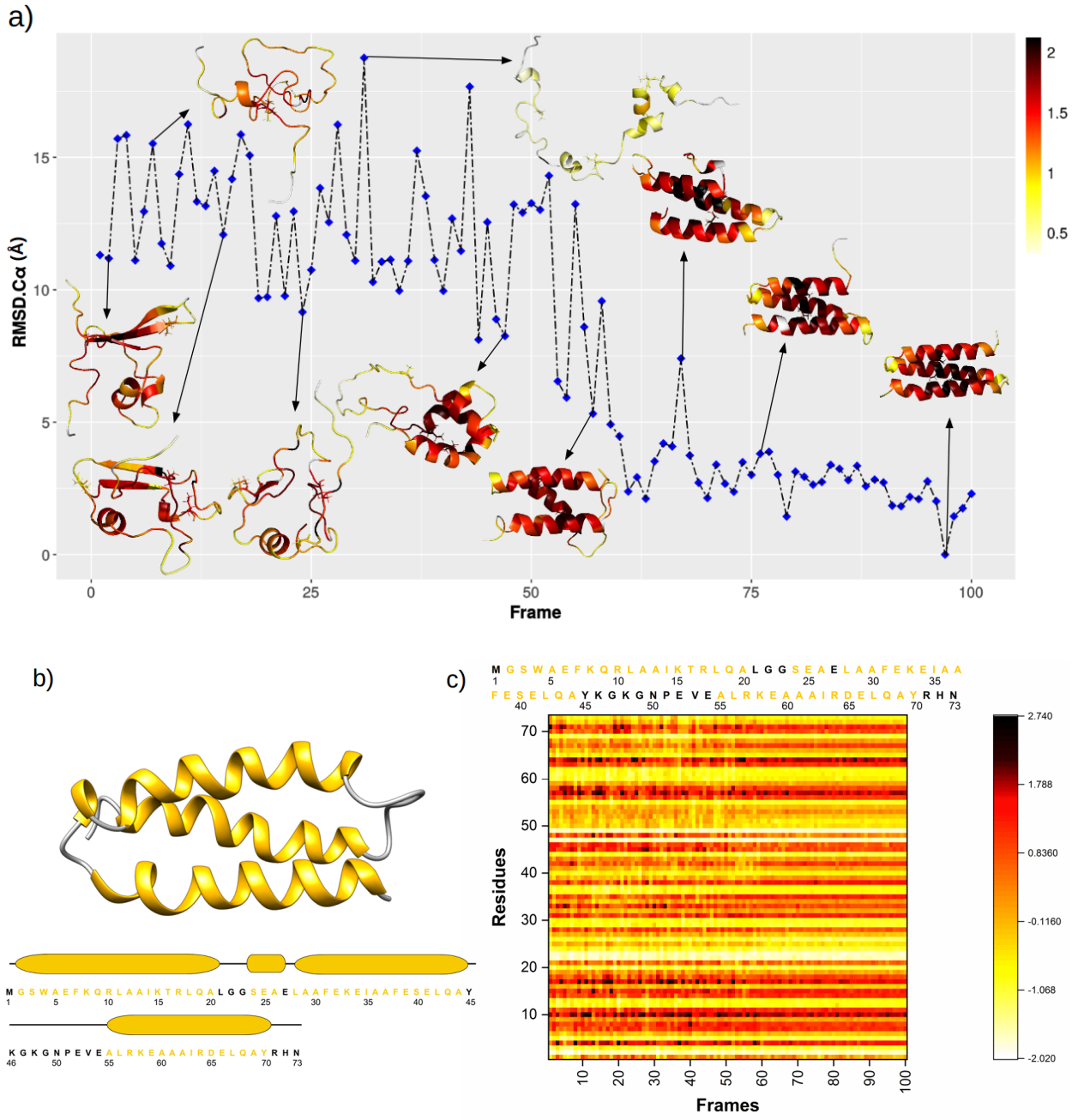


Figura 5.49: (a) Representação de 10 conformações ao longo do RMSD-C_α da trajetória obtida da coordenada x da tICA para o enovelamento da α3D. A dureza local é representada como uma paleta de cores aplicada ao *backbone* representado, em que a cor preta representa as regiões mais duras e a cor branca, as regiões mais moles; (b) Representação estrutural do estado nativo da proteína α3D (PDB-ID: 2A3D). As estruturas secundárias hélices-α são representadas em dourado. Além disso, é apresentada a sequência da estrutura primária da proteína α3D, onde os aminoácidos pertencentes às estruturas secundárias são destacados por meio de cores e figuras geométricas. (c) Mapa de calor da dureza local para as 100 conformações do caminho de enovelamento da α3D.

De forma semelhante ao observado para as proteínas NTL9 e BBA, observamos que à

medida que a estrutura se aproxima da nativa, os valores de dureza tornam-se mais estáveis. Notamos que as estruturas não nativas começam no início do caminho de enovelamento (*frame* 1) e vão até o *frame* 58. A partir do *frame* 59, notamos uma mudança nos padrões de dureza, onde os valores locais de dureza se estabilizam, tornando-se relativamente constantes.

Portanto, podemos separar os dados da Figura 5.49 (c) em duas regiões: estruturas não nativas (*frames* 1:58) e a região onde a dureza local se estabiliza, chamadas de estruturas semelhantes à nativas (*frames* 59-100). Observamos que esses dados corroboram o que foi verificado nos mapas de calor (RMSD e RMSF) por resíduo (Figuras A.13 (c) e A.14 (c) do apêndice A.) onde eles assumem valores relativamente constantes a partir do *frame* 59. Uma similaridade com a análise DSSP também é clara (Figura A.17 do apêndice A), onde a formação das três hélices-α ocorre exatamente a partir do frame 59, levando a assumir o estado nativo. Esses dados apontam que para a proteína α3D se enovelar, é necessário formar primeiro todas as estruturas secundárias, sugerindo um mecanismo de difusão-colisão.

Para uma melhor avaliação dos padrões de dureza ao longo da trajetória, realizamos um mapa de calor do $\Delta(\eta)$ local para as três proteínas estudadas. O valor de $\Delta(\eta)$ é calculado subtraindo os valores de (η) dos resíduos da trajetória pelos valores de (η) da estrutura nativa, ou seja, a com menor valor de RMSD-C_α da trajetória (*frames*: 6, 13 e 97 para as proteínas NTL9, BBA e α3D respectivamente). Os resultados para as três proteínas estão resumidos na Figura 5.50.

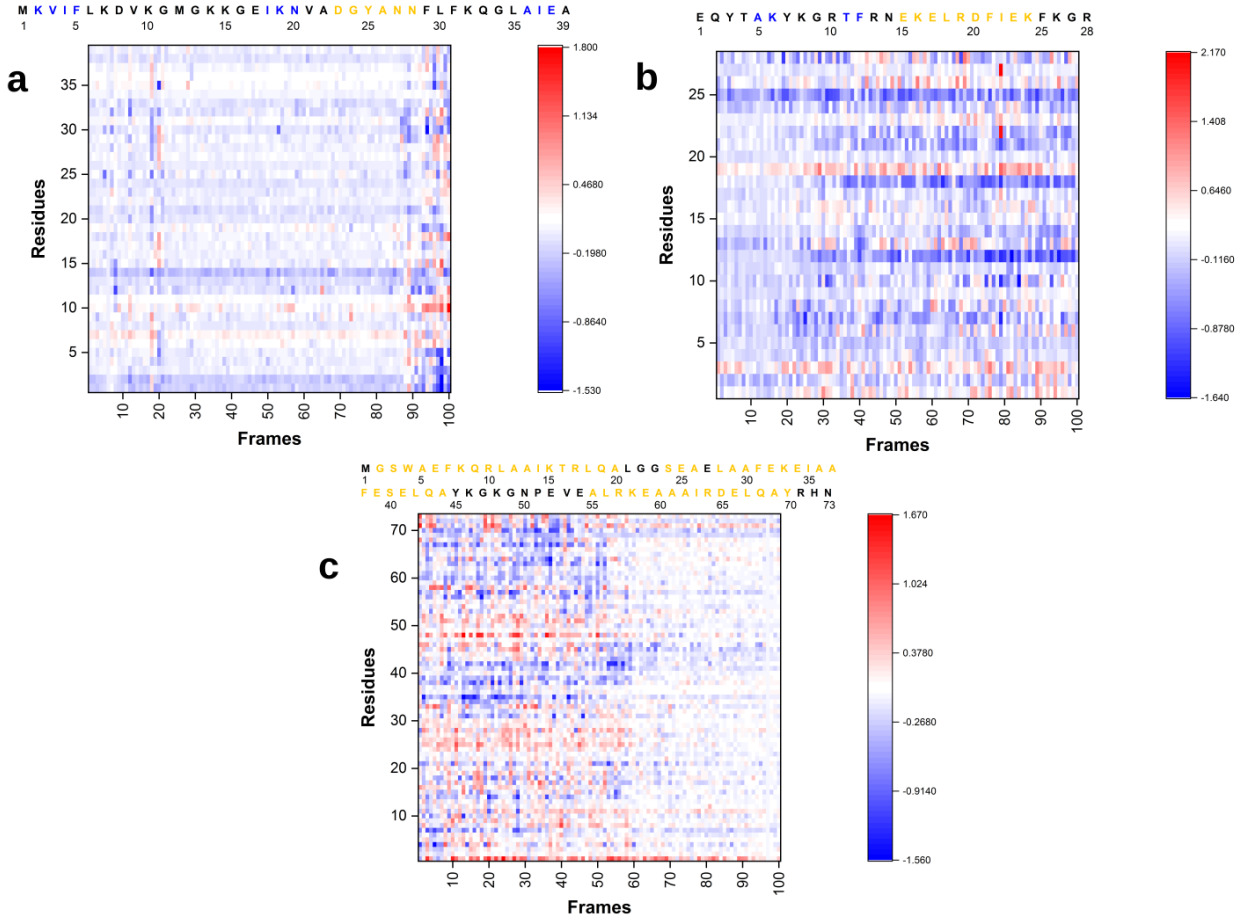


Figura 5.50: Mapa de calor da variação da dureza local ($\Delta(\eta)$) para as proteínas NTL9, BBA e α3D.

Ao analisarmos os dados da Figura 5.50 (a), observamos que o mapa de calor do $\Delta(\eta)$ foi muito semelhante aos mapas de calor gerados para o RMSD e RMSF por resíduo (Figuras A.13 (a) e A.14 (a) do Apêndice A, respectivamente), indicando ainda mais que este descritor segue o processo de enovelamento de proteínas. Além disso, na região de conformações não nativas (*frames*: 100-91), existem resíduos que são mais duros do que sua dureza esperada (em vermelho) e resíduos que são mais moles do que sua dureza esperada (em azul), enquanto $\Delta(\eta)$ assume valores constantes (próximos de zero) na região de estruturas semelhantes à nativa (*frames*: 90-1).

Isso mostra que a flutuação da dureza não é monotônica. Os resíduos moles não estão ficando mais moles à medida que a trajetória avança até atingirem seus níveis esperados de moleza e nem os resíduos duros estão aumentando a dureza ao longo da trajetória. Em vez disso, é um processo com mais nuances, pelo qual a dureza flutua tanto acima quanto abaixo dos valores nativos esperados para cada resíduo. Esses dados justificam porque Faver, J. e colaboradores[94] falharam em discriminar a estrutura nativa de um conjunto de estruturas *decoy*. A questão não é que as estruturas mais bem enoveladas tenham interações duras ou

moles mais favoráveis entre os resíduos, mas que a dureza dos resíduos se torna mais estável à medida que a conformação se aproxima da conformação nativa.

As Figuras 5.50 (b) e (c) mostram que o $\Delta(\eta)$ se comporta de forma semelhante para as proteínas BBA e α3D como para NTL9, embora as três proteínas tenham superfícies de energia livre com rugosidade distinta, mostrando que o $\Delta(\eta)$ pode levar à identificação de conformações importantes durante o enovelamento. Os dados também sugerem que a dureza local pode ser usada não apenas no processo de enovelamento, mas pode ser estendida a outros problemas de busca conformacional.

Os dados na Figura 5.50 também revelam que alguns resíduos apresentam pequenas variações de $\Delta(\eta)$ durante o caminho do enovelamento, enquanto outros resíduos exibem mudanças mais pronunciadas. Os resíduos Phe-5, Lys-19 e Leu-30 na proteína NTL9, Figura 5.50 (a), mostraram um aumento considerável em seus valores de $Delta(\eta)$ quando eles vão da região não nativa a estruturas semelhantes à nativas. Os resíduos Phe-5 e Lys-19 estão presentes nas duas primeiras folhas-β respectivamente, enquanto o resíduo Leu-30 está localizado logo após a hélice-α. Isso sugere que esses resíduos podem desempenhar um papel importante na manutenção da estabilidade das estruturas secundárias.

Na proteína BBA, Figura 5.50 (b), o $\Delta(\eta)$ para os resíduos Phe-12, Leu-18 e Phe-21 apresentou um aumento acentuado quando a proteína assumiu valores mais baixos de RMSD em relação ao estrutura nativa (estruturas semelhantes à nativa). O resíduo Phe-12 contribui para a formação da segunda folha-β enquanto os resíduos Leu-18 e Phe-21 estão presentes na hélice-α.

Finalmente, na proteína α3D, Figura 5.50 (c), os resíduos Phe-7, Phe-31, Ile-35, Phe-38 e Gln-67, apresentaram aumento do $\Delta(\eta)$ quando passaram da região não nativa para região de estruturas semelhantes a nativas. Observamos que o resíduo Phe-7 está presente na primeira hélice-α, os resíduos Phe-31, Phe-35 e Phe-38 estão presentes na segunda hélice-α e o resíduo Gln-67 pertence à terceira hélice-α. Como esses resíduos se tornaram mais duros, eles podem desempenhar um papel importante na formação e manutenção de estruturas secundárias.

Para verificarmos se existem padrões de dureza intrínsecos a cada resíduo, geramos um gráfico da variação da dureza local por tipo de resíduo para cada uma das três proteínas estudadas. Esses resultados são mostrados na Figura 5.51.

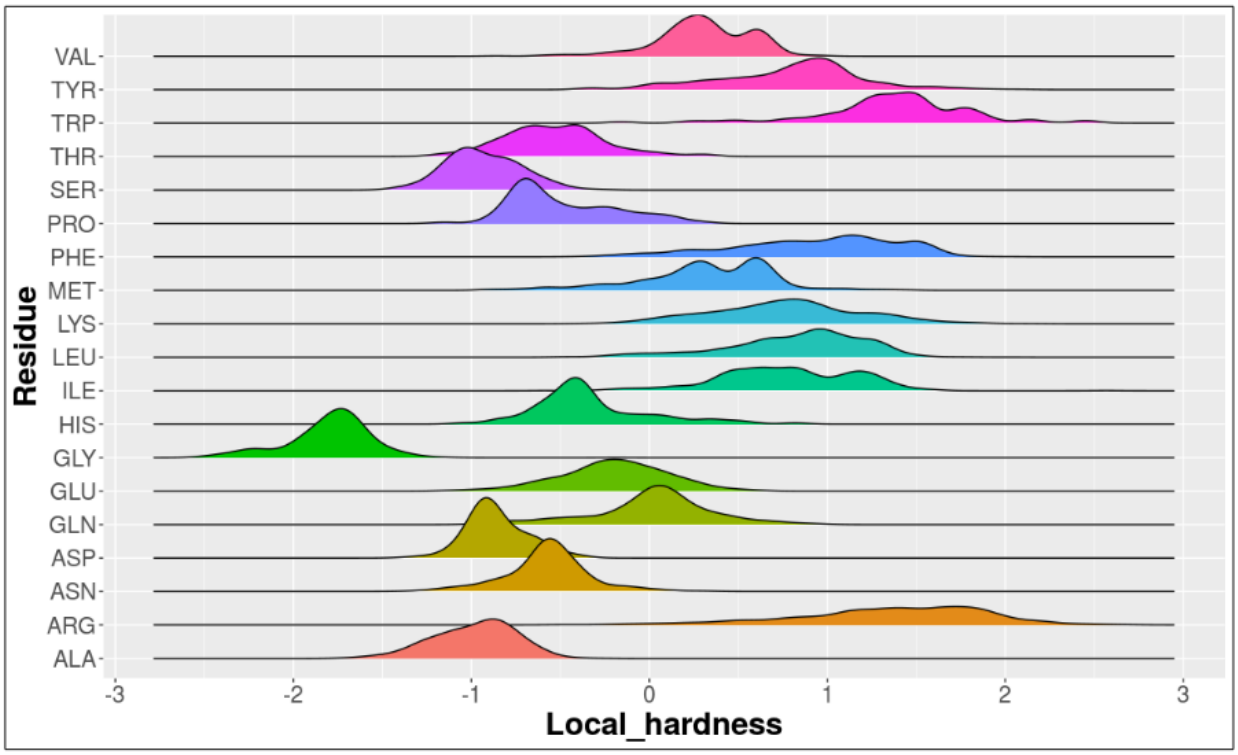


Figura 5.51: Variações de dureza local por tipo de resíduo para as proteínas NTL9, BBA e α3D.

A dureza dos resíduos flutua de maneira semelhante em todas as três proteínas em estudo. Portanto, inferimos que os resíduos possuem padrões intrínsecos de dureza. Os resíduos Ala, Asn, Asp, Gly, His, Pro e Ser são mais moles, enquanto Arg, Ile, Leu, Lys, Phe, Trp e Tyr são mais duros. Os resíduos Gln, Glu, Met, Thr e Val possuem dureza intermediária.

Embora existam alguns resíduos que apresentaram mudanças mais pronunciadas na dureza local, essa variação é significativa ao longo de todo o caminho de dobramento. Ressaltamos, no entanto, que as conformações assumem valores constantes quando vão para a região de estruturas semelhantes à nativa, ou seja, os resíduos assumem estados de dureza química de acordo com as tendências apresentadas nos resultados da Figura 5.51.

Para uma discussão mais detalhada das contribuições por resíduo, classificamos os aminoácidos em cinco grupos de acordo com suas propriedades da cadeia lateral: 1) alifático não polar, 2) aromático, 3) polar não carregado, 4) polar carregado positivamente e 5) polar carregado negativamente. Nas três proteínas estudadas, tomamos para cada grupo os valores médios de dureza dos aminoácidos na região semelhante à nativa e na região não nativa. O objetivo foi avaliar se existem padrões comuns às três proteínas de acordo com o grupo ao qual pertencem os aminoácidos e se a posição do aminoácido na sequência primária afeta a variação local da dureza.

Ao analisarmos o grupo 1 (Figura A.20 do Apêndice A), vemos que alguns resíduos desempenham um papel importante na estabilidade da proteína. Os resíduos (Met-1, Val-3 e

Leu-30), (Leu-18 e Ile-22) e (Ile-14, Ile-35, Leu-42, Leu-63 e Leu-67) presentes nas proteínas NTL9, BBA e α3D respectivamente, possuem alta dureza local média e baixa variabilidade nas regiões de estruturas semelhantes à nativa. Os demais resíduos não apresentaram variações significativas entre as duas regiões avaliadas.

No grupo 2 (Figura A.21 do Apêndice A), observamos que os resíduos de fenilalanina possuem uma propensão geral a se tornarem mais duros quando vão da região não nativa para a região semelhante à nativa ao passo que as tirosinas exibem o comportamento oposto. Para a proteína NTL9, é importante notar o grande aumento na dureza média (de 1,00 para 1,52) no resíduo Phe-5 quando a proteína assume o estado semelhante à nativa. Esta situação não ocorre com os resíduos Phe-29 e Phe-31, que apresentam ligeiras reduções e aumentos na dureza média local, respectivamente. A razão pode ser porque o resíduo Phe-5 está na região de formação da primeira folha-β, enquanto os resíduos Phe-29 e Phe-30 estão na região do *loop*, como mostrado na Figura 5.47 b. Este mesmo comportamento é observado na proteína BBA onde os resíduos Phe-12 (pertencente à segunda folha-β) e Phe-21 (pertencente à hélice-α) apresentam um aumento muito mais pronunciado na dureza média local na região semelhante à nativa da trajetória quando comparada com o resíduo Phe-25 (região de *loop*), como mostrado na Figura 5.48 b. Para a proteína α3D, os resíduos Phe-7, Phe-31 e Phe-38 apresentam um aumento na dureza média quando vão da região não nativa para a região semelhante à nativa, uma vez que todos esses resíduos pertencem à estruturas secundárias do tipo hélice-α.

Porém, o resíduo Phe-38 apresentou um maior aumento na dureza média. Portanto, pode se tratar de um resíduo importante para que a proteína chegue ao seu estado nativo. Esses dados sugerem que alguns resíduos da estrutura secundária assumem maior dureza local quando a proteína atinge seu estado nativo.

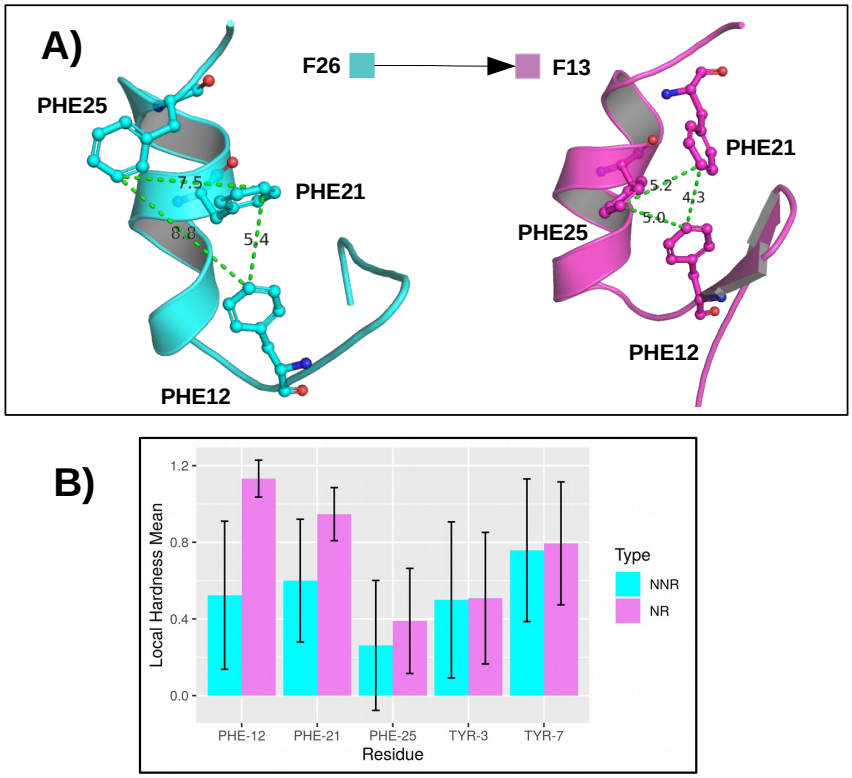


Figura 5.52: Em (A): conformações da proteína BBA para os *frames* 13 e 26 com ênfase nos resíduos Phe-12, Phe-21 e Phe-25. Em (B): Dureza Média Local para resíduos do grupo 2 da proteína BBA na região não nativa (NNR) e região semelhante à nativa (NR).

Do ponto de vista estrutural, os resíduos com maior média de dureza local (Phe-12, Phe-21 e Phe-25) apresentam distâncias muito menores entre si no *frame* 13 quando comparados ao *frame* 26. Este mesmo comportamento foi observado para os resíduos Phe-5 e Phe-31 da proteína NTL9 e para os resíduos Phe-7, Phe-31 e Phe-38 da proteína α3D, como pode ser visto nas Figuras A.29 e A.30 (Apêdice A), respectivamente. Isso sugere que os resíduos que apresentam maior média de dureza local na região semelhante à nativa controlam o processo do enovelamento de proteínas. Além disso, as barras de erro na figura 5.52 (b) representam a variabilidade da dureza local para esses resíduos, o que mostra que os valores nas estruturas semelhantes à nativa são maiores e também mais estáveis para os resíduos Phe-12, Phe-21 e Phe-25.

Ao analisarmos o grupo 3 (Figura A.22 do apêndice A), descobrimos que as cadeias laterais (Asn, Gln, Thr e Ser) tendem a ter valores médios de dureza muito baixos (a maioria com valores negativos) e que a região não nativa tem valores médios de dureza ligeiramente maiores que a região semelhante à nativa. Para a proteína NTL9, verificamos que os resíduos Asn-27 e Asn-28 (pertencentes à hélice-α) apresentaram um maior decréscimo na dureza média quando se deslocaram para a região semelhante à nativa. Este mesmo comportamento é

observado para a proteína BBA no resíduo Asn-14 (pertencente a hélice-α). Assim, sugerimos que esses resíduos desempenham um papel importante na formação da hélice-α enquanto as proteínas NTL9 e BBA seguem para o estado nativo. Para a proteína α3D, observamos que os resíduos Asn-50, Asn-73 (regiões de *loop*), Thr-4, Ser-24 e Ser-40 (regiões de hélices-α) apresentaram dureza média inferior a na região de estruturas semelhantes à nativa.

Ao analisarmos o grupo 4 Figura A.23 do Apêndice A), vemos que, em geral, tanto as lisinas quanto as argininas tendem a não apresentar variações significativas na dureza média quando vão da região não nativa para a região semelhante à nativa. As argininas têm uma dureza média muito superior às lisinas, com exceção da Arg-28. Isso ocorre porque Arg-28 é o último resíduo na sequência (região de *loop*) e, portanto, é mais flexível.

Um dado interessante ocorre na proteína NTL9, onde verificamos variações bruscas na dureza média dos resíduos Lys-10 e Lys-19, porém inversamente nas duas regiões analisadas. Verificamos inicialmente na sequência primária (Figura 5.47) que Lys-2 (resíduo da primeira folha-β) e Lys-7 (resíduo próximo à primeira folha-β) apresentam dureza média maior na região semelhante à nativa. Também notamos uma redução acentuada na dureza média da Lys-10 (região de *loop* longe das folhas-β) nesta região. A dureza média aumenta novamente nos resíduos Lys-14 e Lys-15 (resíduos próximos à segunda folha-β), tendo um valor máximo de dureza no resíduo Lys-19 (resíduo pertencente à segunda folha-β). Assim, para a proteína NTL9, sugerimos que a dureza média tende a ser maior nas folhas-β devido a um efeito indutivo de dureza nos resíduos próximos a essas estruturas secundárias, uma vez que esses resíduos são carregados positivamente. Para a proteína α3D, merece destaque o resíduo Lys-48, que apresentou maior redução na dureza média ao passar para a região semelhante à nativa.

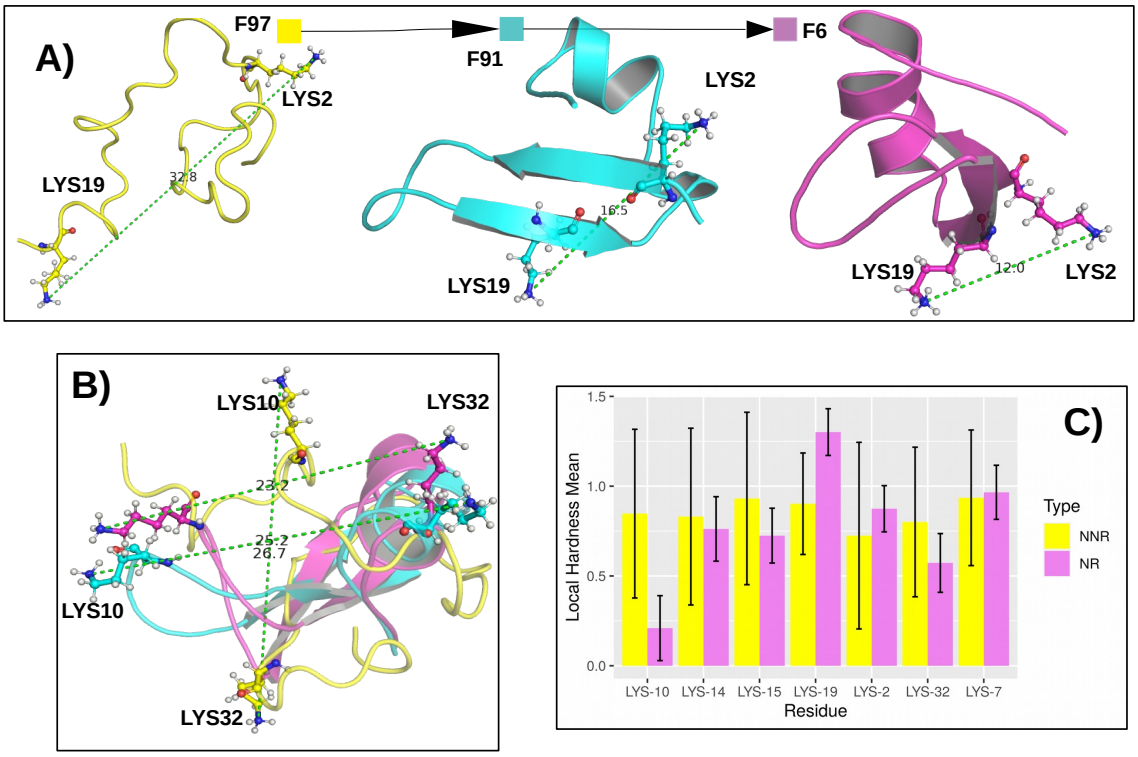


Figura 5.53: Em (A): conformações da proteína NTL9 para os *frames* 97, 91 e 6 com ênfase nos resíduos Lys-2 e Lys-19. Em (B): Estruturas alinhadas dos *frames* 97, 91 e 6 com ênfase nos resíduos 10 e 32. Em (C): Dureza Média Local para resíduos do grupo 4 da proteína NTL9 na região não nativa (NNR) e região semelhante à nativa (NR).

Na Figura 5.53, observamos que os resíduos com maiores valores de dureza média local na região semelhante à nativa (Lys-2 e Lys-19) apresentaram maiores variações estruturais ao passar do estado desenovelado para o estado nativo (Figura 5.53 A). Os resíduos Lys-10 e Lys-32, que apresentaram baixas médias de dureza local na região semelhante à nativa, não apresentaram alterações estruturais significativas. Esses resultados corroboram com os apresentados na análise dos resíduos do grupo 2, em que os resíduos com maior dureza média local são os mais importantes no processo de enovelamento de proteínas.

Ao analisarmos o Grupo 5 (Figura A.24), verificamos que para as proteínas NTL9 e BBA, os resíduos de glutamato e aspartato apresentaram durezas médias locais negativas, apresentando valores semelhantes nas regiões de estruturas não nativas e de conformações semelhante à nativa. Ressalta-se que os resíduos de aspartato apresentaram valores médios de dureza local inferiores aos resíduos de glutamato nas três proteínas estudadas. Para a proteína α3D, descobrimos que os resíduos Glu-25 e Glu-27 têm dureza local média muito menor na região próxima da nativa, enquanto os resíduos Glu-34 e Glu-39 têm dureza local média muito maior nessa mesma região . Isso sugere que esses resíduos são importantes no processo de dobramento da proteína α3D.

CAPÍTULO 6

CONCLUSÕES

As conclusões são divididas em duas seções, onde a primeira seção corresponde aos resultados referentes ao sistema RTA-ligante e a segunda seção ao problema do enovelamento de proteínas.

6.1 Sistema Proteína-Ligante: candidatos a inibidores da Ricina

O objetivo principal desse trabalho, foi apresentar um estudo onde duas abordagens computacionais foram aplicadas (cálculo do ΔH_{bind} e descritores químicos quânticos de reatividade) para obter *insights* teóricos sobre as interações entre a subunidade RTA da ricina e alguns de seus inibidores.

Em nosso estudo, observamos que os cálculos *single-point* de energia e do ΔH_{bind} com os métodos PM6-DH+, PM6-D3H4 e PM7 apresentaram uma excelente correlação com os dados experimentais de IC_{50}. Além disso, embora o método PM7 tenha apresentado uma correlação um pouco menor do que o método PM6-D3H4, apenas o método PM7 foi capaz de classificar corretamente todos os ligantes analisados. Este resultado sugere que o método PM7 é mais sensível a pequenas variações de IC_{50}.

Ao compararmos os dados de IC_{50} com os valores de ΔH_{bind} obtidos após a otimização *full atom* com os métodos PM6-DH+, PM6-D3H4 e PM7, verificamos que a correlação diminui significativamente. Embora a correlação tenha sido bastante reduzida com a otimização das estruturas, os métodos PM6-D3H4 e PM7 foram capazes de classificar corretamente o melhor ligante. Esses resultados impactaram em vários trabalhos que se sucederam no grupo,[253,254]

permitindo a aplicação da abordagem (DM + 1SCF) como etapa de triagem virtual e também na parte do enovelamento de proteínas. Desse modo, eliminamos uma possível necessidade de otimização de geometria das estruturas no caminho do enovelamento. Os dois trabalhos realizados nessa tese ocorreram em tempos distintos: primeiro, realizamos o estudo da interação proteína-ligante e segundo, do enovelamento de proteínas. Desse modo, o resultado do primeiro trabalho foi decisivo para avançarmos no estudo do enovelamento de proteínas, permitindo assim a conclusão da tese.

Observamos também que, na otimização da geometria, a estrutura adota uma conformação local mínima que contribui menos para a posição preferencial do ligante no sítio ativo. Por outro lado, a estrutura representativa da dinâmica molecular representa a posição preferencial do ligante no sítio ativo. Assim, concluímos que para o conjunto de dados estudado, é melhor usar estruturas equilibradas da DM para realizar cálculos *single-point* de energia para obter o ΔH_{bind}.

Os cálculos *single-point* vi QM/MM ONIOM com o funcional B3LYP mostraram boa correlação com IC_{50}, entretanto, os métodos PM6-DH+, PM6-D3H4 e PM7 apresentaram melhor desempenho. Ao usarmos o funcional ΩB97X-D que inclui correção de dispersão DFT-D2,[221] o método QM/MM ONIOM apresentou um valor de R e desempenho semelhante ao método PM7. Além disso, os cálculos QM/MM ONIOM para o ΔH_{bind} apresentaram um custo computacional muito mais elevado em comparação com os métodos semiempíricos, cerca de 250 vezes maior para o funcional (B3LYP) e 1400 vezes maior para o funcional (ωB97X-D).

Além disso, observamos que para os casos estudados, os descritores de reatividade apontaram que ambos os tipos de interações, sobreposição molecular (*molecular overlap*) e interações eletrostáticas, desempenham um papel importante na afinidade geral desses ligantes para o sítio de ligação da RTA.

6.2 Enovelamento de Proteínas

Verificamos que os MSMs são uma ferramenta importante para a avaliação de dados oriundos da trajetória de DM. Os MSMs foram capazes de fornecer informações importantes dando detalhes de diversos caminhos para o enovelamento e as probabilidades de transição, fornecendo não somente o caminho de maior fluxo, mas também caminhos alternativos para enovelamento das proteínas avaliadas.

Além disso, forneceu informações importantes sobre o estado de transição entre a estrutura desenovelada e a conformação nativa. Isso possibilitou a inferência sobre o tipo de mecanismo

que rege as proteínas avaliadas. Nós sugerimos que a NTL9, a BBA e a α3D seguem o mecanismo de difusão-colisão, em que as estruturas secundárias são formadas primeiro, seguido da associação dessas para formar o arranjo terciário até a composição do estado nativo.

Neste trabalho, também avaliamos aspectos da estrutura eletrônica de três proteínas de rápido enovelamento (NTL9, BBA e α3D) por meio de QCMDs obtidos pelos métodos DFT-D3 e PM7. A análise dos QCMDs locais revelou aspectos importantes sobre o papel da estrutura eletrônica ao longo do enovelamento de proteínas. Observamos que a dureza local e a densidade eletrônica tornam-se constantes quando a proteína se aproxima do estado nativo, correlacionando-se muito bem com RMSF e RMSD por resíduo. Desta forma, é possível entender aspectos da superfície de energia livre e prever se as conformações estão indo para o estado nativo ou não, avaliando a dureza local. Os dados do $\Delta(\ eta)$ local (Figura 5.50), produziram um mapa de calor muito semelhante aos obtidos com os dados do RMSF e RMSD por resíduo, evidenciando ainda mais a correlação entre a dureza local e propriedades estruturais. Além disso, os resultados também se correlacionam com a formação de estruturas secundárias, como pode ser visto nas análises DSSP (Figuras A.15-A.17 do apêndice A).

Sugerimos que os resíduos que apresentam maior dureza média durante o enovelamento são os que controlam o processo e, portanto, devem ter suas propriedades monitoradas individualmente ao longo do enovelamento das proteínas. Validamos todos os resultados via DFT-D3, mas ressaltamos que os cálculos utilizando o método semiempírico PM7 apresentaram o mesmo padrão, portanto é possível avaliar todos os descritores com o PM7 com qualidade semelhante e com baixo custo computacional.

O uso de descritores de reatividade obtidos por meio do método semiempírico PM7 pode desbloquear as fronteiras do estudo de sistemas macromoleculares, permitindo que o estudo de estruturas eletrônicas possa ser executado em processadores de baixo custo. Observamos que o cálculo *single-point* para a proteína NTL9 em um único processador de notebook (Intel Core i5-7300HQ 250 GHz) via PM7 com MOZYME foi cerca de 1700 vezes mais rápido do que o cálculo DFT-D3 realizado em uma GPU Nvidia Tesla K40 para o mesmo estrutura.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] W. P. Bozza, W. H. Tolleson, L. A. R. Rosado, and B. Zhang, "Ricin detection: Tracking active toxin," *Biotechnology Advances*, vol. 33, no. 1, pp. 117–123, 2015.

[2] W. Montfort, J. E. Villafranca, A. F. Monzingo, S. R. Ernst, B. Katzin, E. Rutenber, N. H. Xuong, R. Hamlin, and J. D. Robertus, "The three-dimensional structure of ricin at 2.8 a.," *Journal of Biological Chemistry*, vol. 262, no. 11, pp. 5398–5403, 1987.

[3] E. Rutenber, B. J. Katzin, S. Ernst, E. J. Collins, D. Mlsna, M. P. Ready, and J. D. Robertus, "Crystallographic refinement of ricin to 2.5 å," *Proteins: Structure, Function, and Bioinformatics*, vol. 10, no. 3, pp. 240–250, 1991.

[4] J. D. Robertus and A. F. Monzingo, "The structure of ribosome inactivating proteins.," *Mini Reviews in Medicinal Chemistry*, vol. 4, no. 5, pp. 477–486, 2004.

[5] Y. Endo and K. Tsurugi, "Rna n-glycosidase activity of ricin a-chain. mechanism of action of the toxic lectin ricin on eukaryotic ribosomes.," *Journal of Biological Chemistry*, vol. 262, no. 17, pp. 8128–8130, 1987.

[6] K. L. May, Q. Yan, and N. E. Tumer, "Targeting ricin to the ribosome," *Toxicon*, vol. 69, pp. 143–151, 2013.

[7] J. Audi, M. Belson, M. Patel, J. Schier, and J. Osterloh, "Ricin poisoning: a comprehensive review," *JAMA: The Journal of the American Medical Association*, vol. 294, no. 18, pp. 2342–2351, 2005.

[8] R. H. Argent, A. M. Parrott, P. J. Day, L. M. Roberts, P. G. Stockley, J. M. Lord, and S. E. Radford, "Ribosome-mediated folding of partially unfolded ricin a-chain," *Journal of Biological Chemistry*, vol. 275, no. 13, pp. 9263–9269, 2000.

[9] J. C. Simpson, L. M. Roberts, K. Römisch, J. Davey, D. H. Wolf, and J. M. Lord, "Ricin a chain utilises the endoplasmic reticulum-associated protein degradation pathway to enter the cytosol of yeast," *FEBS Letters*, vol. 459, no. 1, pp. 80–84, 1999.

[10] E. J. Chaves, I. Q. Padilha, D. A. Araújo, and G. B. Rocha, "Determining the Relative Binding Affinity of Ricin Toxin A Inhibitors by Using Molecular Docking and Nonequilibrium Work," *Journal of Chemical Information and Modeling*, vol. 58, no. 6, pp. 1205–1213, 2018.

[11] Y. Bai, B. Watt, P. G. Wahome, N. J. Mantis, and J. D. Robertus, "Identification of new classes of ricin toxin inhibitors by virtual screening," *Toxicon*, vol. 56, no. 4, pp. 526–534, 2010.

[12] U. Ryde and P. Soderhjelm, "Ligand-binding affinity estimates supported by quantum-mechanical methods," *Chemical Reviews*, vol. 116, no. 9, pp. 5520–5566, 2016.

[13] E. Nikitina, V. Sulimov, V. Zayets, and N. Zaitseva, "Semiempirical calculations of binding enthalpy for protein–ligand complexes," *International Journal of Quantum Chemistry*, vol. 97, no. 2, pp. 747–763, 2004.

[14] G. L. Warren, C. W. Andrews, A. M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff, and M. S. Head, "A critical assessment of docking programs and scoring functions," *Journal of Medicinal Chemistry*, vol. 49, no. 20, pp. 5912–5931, 2006.

[15] J. B. Cross, D. C. Thompson, B. K. Rai, J. C. Baber, K. Y. Fan, Y. Hu, and C. Humblet, "Comparison of several molecular docking programs: pose prediction and virtual screening accuracy," *Journal of Chemical Information and Modeling*, vol. 49, no. 6, pp. 1455–1474, 2009.

[16] N. D. Yilmazer and M. Korth, "Comparison of molecular mechanics, semi-empirical quantum mechanical, and density functional theory methods for scoring protein–ligand interactions," *The Journal of Physical Chemistry B*, vol. 117, no. 27, pp. 8075–8084, 2013.

[17] N. D. Yilmazer, P. Heitel, T. Schwabe, and M. Korth, "Benchmark of electronic structure methods for protein–ligand interactions based on high-level reference data," *Journal of Theoretical and Computational Chemistry*, vol. 14, no. 01, p. 1540001, 2015.

[18] J. Fanfrlík, A. K. Bronowska, J. Řezáč, O. Přenosil, J. Konvalinka, and P. Hobza, "A reliable docking/scoring scheme based on the semiempirical quantum mechanical PM6-DH2

method accurately covering dispersion and H-bonding: HIV-1 protease with 22 ligands," *Journal of Physical Chemistry B*, vol. 114, no. 39, pp. 12666–12678, 2010.

[19] A. Pecina, J. Brynda, L. Vrzal, R. Gnanasekaran, M. Hořejší, S. M. Eyrilmez, J. Řezáč, M. Lepšík, P. Řezáčová, P. Hobza, P. Majer, V. Veverka, and J. Fanfrlík, "Ranking Power of the SQM/COSMO Scoring Function on Carbonic Anhydrase II–Inhibitor Complexes," *ChemPhysChem*, vol. 19, no. 7, pp. 873–879, 2018.

[20] P. Mikulskis, S. Genheden, K. Wichmann, and U. Ryde, "A semiempirical approach to ligand-binding affinities: Dependence on the hamiltonian and corrections," *Journal of Computational Chemistry*, vol. 33, no. 12, pp. 1179–1189, 2012.

[21] R. González, C. F. Suárez, H. J. Bohórquez, M. A. Patarroyo, and M. E. Patarroyo, "Semi-empirical quantum evaluation of peptide–mhc class ii binding," *Chemical Physics Letters*, vol. 668, pp. 29–34, 2017.

[22] K. Kamel and A. Kolinski, "Assessment of the free binding energy of 1,25-dihydroxyvitamin D3 and its analogs with the human VDR receptor model," *Acta Biochimica Polonica*, vol. 59, no. 4, pp. 653–660, 2012.

[23] K. Wichapong, A. Rohe, C. Platzer, I. Slynko, F. Erdmann, M. Schmidt, and W. Sippl, "Application of docking and qm/mm-gbsa rescoring to screen for novel myt1 kinase inhibitors," *Journal of Chemical Information and Modeling*, vol. 54, no. 3, pp. 881–893, 2014.

[24] S. Natesan, R. Subramaniam, C. Bergeron, and S. Balaz, "Binding affinity prediction for ligands and receptors forming tautomers and ionization species: inhibition of mitogen-activated protein kinase-activated protein kinase 2 (mk2)," *Journal of Medicinal Chemistry*, vol. 55, no. 5, pp. 2035–2047, 2012.

[25] C. N. Alves, S. Martí, R. Castillo, J. Andres, V. Moliner, I. Tunon, and E. Silla, "A quantum mechanics/molecular mechanics study of the protein–ligand interaction for inhibitors of hiv-1 integrase," *Chemistry–A European Journal*, vol. 13, no. 27, pp. 7715–7724, 2007.

[26] M. R. Reddy and M. D. Erion, "Relative binding affinities of fructose-1, 6-bisphosphatase inhibitors calculated using a quantum mechanics-based free energy perturbation method," *Journal of the American Chemical Society*, vol. 129, no. 30, pp. 9296–9297, 2007.

[27] P. Mikulskis, D. Cioloboc, M. Andrejić, S. Khare, J. Brorsson, S. Genheden, R. A. Mata, P. Söderhjelm, and U. Ryde, "Free-energy perturbation and quantum mechanical study

of sampl4 octa-acid host–guest binding energies," *Journal of Computer-Aided Molecular Design*, vol. 28, no. 4, pp. 375–400, 2014.

[28] R. Rathore, R. N. Reddy, A. Kondapi, P. Reddanna, and M. R. Reddy, "Use of quantum mechanics/molecular mechanics-based fep method for calculating relative binding affinities of fbpase inhibitors for type-2 diabetes," *Theoretical Chemistry Accounts*, vol. 131, no. 2, p. 1096, 2012.

[29] K. Świderek, S. Martí, and V. Moliner, "Theoretical studies of hiv-1 reverse transcriptase inhibition," *Physical Chemistry Chemical Physics*, vol. 14, no. 36, pp. 12614–12624, 2012.

[30] F. R. Beierlein, J. Michel, and J. W. Essex, "A simple qm/mm approach for capturing polarization effects in protein- ligand binding free energy calculations," *The Journal of Physical Chemistry B*, vol. 115, no. 17, pp. 4911–4926, 2011.

[31] C. J. Woods, K. E. Shaw, and A. J. Mulholland, "Combined quantum mechanics/molecular mechanics (qm/mm) simulations for protein–ligand complexes: free energies of binding of water molecules in influenza neuraminidase," *The Journal of Physical Chemistry B*, vol. 119, no. 3, pp. 997–1001, 2014.

[32] S. Genheden, U. Ryde, and P. Söderhjelm, "Binding affinities by alchemical perturbation using qm/mm with a large qm system and polarizable mm model," *Journal of Computational Chemistry*, vol. 36, no. 28, pp. 2114–2124, 2015.

[33] M. A. Olsson, P. Söderhjelm, and U. Ryde, "Converging ligand-binding free energies obtained with free-energy perturbations at the quantum mechanical level," *Journal of Computational Chemistry*, vol. 37, no. 17, pp. 1589–1600, 2016.

[34] Z. Cournia, B. Allen, and W. Sherman, "Relative binding free energy calculations in drug discovery: recent advances and practical considerations," *Journal of Chemical Information and Modeling*, vol. 57, no. 12, pp. 2911–2937, 2017.

[35] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham, "Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models," *Accounts of Chemical Research*, vol. 33, no. 12, pp. 889–897, 2000.

[36] T. Hou, J. Wang, Y. Li, and W. Wang, "Assessing the performance of the mm/pbsa and mm/gbsa methods. 1. the accuracy of binding free energy calculations based on

molecular dynamics simulations," *Journal of Chemical Information and Modeling*, vol. 51, no. 1, pp. 69–82, 2011.

[37] B. O. Brandsdal, F. Österberg, M. Almlöf, I. Feierberg, V. B. Luzhkov, and J. Åqvist, "Free energy calculations and ligand binding," in *Advances in Protein Chemistry*, vol. 66, pp. 123–158, Elsevier, 2003.

[38] A. V. Marenich, C. J. Cramer, and D. G. Truhlar, "Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions," *The Journal of Physical Chemistry B*, vol. 113, no. 18, pp. 6378–6396, 2009.

[39] R. F. Ribeiro, A. V. Marenich, C. J. Cramer, and D. G. Truhlar, "Prediction of sampl2 aqueous solvation free energies and tautomeric ratios using the sm8, sm8ad, and smd solvation models," *Journal of Computer-Aided Molecular Design*, vol. 24, no. 4, pp. 317–333, 2010.

[40] S. Miyamoto and P. A. Kollman, "Absolute and relative binding free energy calculations of the interaction of biotin and its analogs with streptavidin using molecular dynamics/-free energy perturbation approaches," *Proteins: Structure, Function, and Bioinformatics*, vol. 16, no. 3, pp. 226–245, 1993.

[41] S. K. Mishra and J. Koča, "Assessing the Performance of MM/PBSA, MM/GBSA, and QM-MM/GBSA Approaches on Protein/Carbohydrate Complexes: Effect of Implicit Solvent Models, QM Methods, and Entropic Contributions," *Journal of Physical Chemistry B*, vol. 122, no. 34, pp. 8113–8121, 2018.

[42] P.-C. Su, C.-C. Tsai, S. Mehboob, K. E. Hevener, and M. E. Johnson, "Comparison of radii sets, entropy, qm methods, and sampling on mm-pbsa, mm-gbsa, and qm/mm-gbsa ligand binding energies of f. tularensis enoyl-acp reductase (f abi)," *Journal of Computational Chemistry*, vol. 36, no. 25, pp. 1859–1873, 2015.

[43] A. Khandelwal, V. Lukacova, D. Comez, D. M. Kroll, S. Raha, and S. Balaz, "A combination of docking, qm/mm methods, and md simulation for binding affinity estimation of metalloprotein ligands," *Journal of Medicinal Chemistry*, vol. 48, no. 17, pp. 5437–5447, 2005.

[44] M. Xiang, Y. Lin, G. He, L. Chen, M. Yang, S. Yang, and Y. Mo, "Correlation between biological activity and binding energy in systems of integrin with cyclic rgd-containing

binders: a qm/mm molecular dynamics study," *Journal of Molecular Modeling*, vol. 18, no. 11, pp. 4917–4927, 2012.

[45] C. Cave-Ayland, C.-K. Skylaris, and J. W. Essex, "Direct validation of the single step classical to quantum free energy perturbation," *The Journal of Physical Chemistry B*, vol. 119, no. 3, pp. 1017–1025, 2014.

[46] L. Rao, I. Y. Zhang, W. Guo, L. Feng, E. Meggers, and X. Xu, "Nonfitting protein–ligand interaction scoring function based on first-principles theoretical chemistry methods: Development and application on kinase inhibitors," *Journal of Computational Chemistry*, vol. 34, no. 19, pp. 1636–1646, 2013.

[47] J. Fanfrlík, F. X. Ruiz, A. Kadlčíková, J. Řezáč, A. Cousido-Siah, A. Mitschler, S. Haldar, M. Lepšík, M. H. Kolář, P. Majer, A. D. Podjarny, and P. Hobza, "The Effect of Halogen-to-Hydrogen Bond Substitution on Human Aldose Reductase Inhibition," *ACS Chemical Biology*, vol. 10, no. 7, pp. 1637–1642, 2015.

[48] M. R. Blomberg, T. Borowski, F. Himo, R.-Z. Liao, and P. E. Siegbahn, "Quantum chemical studies of mechanisms for metalloenzymes," *Chemical Reviews*, vol. 114, no. 7, pp. 3601–3658, 2014.

[49] P. E. Siegbahn and F. Himo, "Recent developments of the quantum chemical cluster approach for modeling enzyme reactions," *JBIC Journal of Biological Inorganic Chemistry*, vol. 14, no. 5, pp. 643–651, 2009.

[50] C. V. Sumowski and C. Ochsenfeld, "A convergence study of qm/mm isomerization energies with the selected size of the qm region for peptidic systems," *The Journal of Physical Chemistry A*, vol. 113, no. 43, pp. 11734–11741, 2009.

[51] L. Hu, J. Eliasson, J. Heimdal, and U. Ryde, "Do quantum mechanical energies calculated for small models of protein-active sites converge?," *The Journal of Physical Chemistry A*, vol. 113, no. 43, pp. 11793–11800, 2009.

[52] W. A. De Jong, E. Bylaska, N. Govind, C. L. Janssen, K. Kowalski, T. Müller, I. M. Nielsen, H. J. Van Dam, V. Veryazov, and R. Lindh, "Utilizing high performance computing for chemistry: Parallel computational chemistry," *Physical Chemistry Chemical Physics*, vol. 12, no. 26, pp. 6896–6920, 2010.

[53] W. Jia, J. Wang, X. Chi, and L. W. Wang, "GPU implementation of the linear scaling three dimensional fragment method for large scale electronic structure calculations," *Computer Physics Communications*, vol. 211, pp. 8–15, 2017.

[54] M. A. Olson, "Ricin a-chain structural determinant for binding substrate analogues: A molecular dynamics simulation analysis," *Proteins: Structure, Function, and Bioinformatics*, vol. 27, no. 1, pp. 80–95, 1997.

[55] M. A. Olson and L. Cuff, "Free energy determinants of binding the rrna substrate and small ligands to ricin a-chain," *Biophysical Journal*, vol. 76, no. 1, pp. 28–39, 1999.

[56] X. Yan, P. Day, T. Hollis, A. F. Monzingo, E. Schelp, J. D. Robertus, G. Milne, and S. Wang, "Recognition and interaction of small rings with the ricin a-chain binding site," *Proteins: Structure, Function, and Bioinformatics*, vol. 31, no. 1, pp. 33–41, 1998.

[57] V. d. Godoi Contessoto, A. B. d. Oliveira, J. Chahine, R. J. d. Oliveira, and V. B. Pereira Leite, "Introdução ao problema de enovelamento de proteínas: uma abordagem utilizando modelos computacionais simplificados," *Revista Brasileira de Ensino de Física*, vol. 40, 2018.

[58] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, "Funnels, pathways, and the energy landscape of protein folding: a synthesis," *Proteins: Structure, Function, and Bioinformatics*, vol. 21, no. 3, pp. 167–195, 1995.

[59] M. Levitt, "A simplified representation of protein conformations for rapid simulation of protein folding," *Journal of Molecular Biology*, vol. 104, no. 1, pp. 59–107, 1976.

[60] D. L. Nelson and M. M. Cox, *Princípios de Bioquímica de Lehninger*. Artmed, Porto Alegre, 6 ed., 2014.

[61] J. M. Berg, L. Stryer, and J. L. Tymoczko, *Bioquimica*. Guanabara Koogan, 7 ed., 2014.

[62] C. B. Anfinsen, E. Haber, M. Sela, and F. White Jr, "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain," *Proceedings of the National Academy of Sciences*, vol. 47, no. 9, p. 1309, 1961.

[63] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[64] C. Levinthal, "Are there pathways for protein folding?," *Journal de Chimie Physique*, vol. 65, pp. 44–45, 1968.

[65] C. Levinthal, "Mössbaun spectroscopy in biological systems proceedings, urbana, 1968," 1969.

[66] L. Martínez, "Introducing the levinthal's protein folding paradox and its solution," *Journal of Chemical Education*, vol. 91, no. 11, pp. 1918–1923, 2014.

[67] E. I. Shakhnovich, "Theoretical studies of protein-folding thermodynamics and kinetics," *Current Opinion in Structural Biology*, vol. 7, no. 1, pp. 29–40, 1997.

[68] N. Darby, C. Van Mierlo, and T. Creighton, "The 5–55 single-disulphide intermediate in folding of bovine pancreatic trypsin inhibitor," *FEBS letters*, vol. 279, no. 1, pp. 61–64, 1991.

[69] J. P. Staley and P. S. Kim, "Complete folding of bovine pancreatic trypsin inhibitor with only a single disulfide bond.," *Proceedings of the National Academy of Sciences*, vol. 89, no. 5, pp. 1519–1523, 1992.

[70] A. L. Fink, "Compact intermediate states in protein folding," *Annual Review of Biophysics and Biomolecular Structure*, vol. 24, no. 1, pp. 495–522, 1995.

[71] V. Uversky, "Protein folding revisited. a polypeptide chain at the folding–misfolding–nonfolding cross-roads: which way to go?," *Cellular and Molecular Life Sciences CMLS*, vol. 60, no. 9, pp. 1852–1871, 2003.

[72] V. Daggett and A. R. Fersht, "Is there a unifying mechanism for protein folding?," *Trends in Biochemical Sciences*, vol. 28, no. 1, pp. 18–25, 2003.

[73] E. Shakhnovich and A. Gutin, "Formation of unique structure in polypeptide chains: theoretical investigation with the aid of a replica approach," *Biophysical Chemistry*, vol. 34, no. 3, pp. 187–199, 1989.

[74] J. D. Bryngelson and P. G. Wolynes, "A simple statistical field theory of heteropolymer collapse with application to protein folding," *Biopolymers: Original Research on Biomolecules*, vol. 30, no. 1-2, pp. 177–188, 1990.

[75] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, "Optimal protein-folding codes from spin-glass theory," *Proceedings of the National Academy of Sciences*, vol. 89, no. 11, pp. 4918–4922, 1992.

[76] P. E. Leopold, M. Montal, and J. N. Onuchic, "Protein folding funnels: a kinetic approach to the sequence-structure relationship.," *Proceedings of the National Academy of Sciences*, vol. 89, no. 18, pp. 8721–8725, 1992.

[77] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, "Theory of protein folding: the energy landscape perspective," *Annual Review of Physical Chemistry*, vol. 48, no. 1, pp. 545–600, 1997.

[78] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, "The protein folding problem," *Annual Review of Biophysics*, vol. 37, pp. 289–316, 2008.

[79] D. Thirumalai, E. P. O'Brien, G. Morrison, and C. Hyeon, "Theoretical perspectives on protein folding," *Annual Review of Biophysics*, vol. 39, pp. 159–183, 2010.

[80] J. Wang, R. J. Oliveira, X. Chu, P. C. Whitford, J. Chahine, W. Han, E. Wang, J. N. Onuchic, and V. B. Leite, "Topography of funneled landscapes determines the thermodynamics and kinetics of protein folding," *Proceedings of the National Academy of Sciences*, vol. 109, no. 39, pp. 15763–15768, 2012.

[81] A. R. Dinner, A. Šali, L. J. Smith, C. M. Dobson, and M. Karplus, "Understanding protein folding via free-energy surfaces from theory and experiment," *Trends in Biochemical Sciences*, vol. 25, no. 7, pp. 331–339, 2000.

[82] C. M. Dobson, "Protein folding and misfolding," *Nature*, vol. 426, no. 6968, pp. 884–890, 2003.

[83] J. D. Bryngelson and P. G. Wolynes, "Spin glasses and the statistical mechanics of protein folding," *Proceedings of the National Academy of sciences*, vol. 84, no. 21, pp. 7524–7528, 1987.

[84] R. B. Best, G. Hummer, and W. A. Eaton, "Native contacts determine protein folding mechanisms in atomistic simulations," *Proceedings of the National Academy of Sciences*, vol. 110, no. 44, pp. 17874–17879, 2013.

[85] S. W. Englander and L. Mayne, "The nature of protein folding pathways," *Proceedings of the National Academy of Sciences*, vol. 111, no. 45, pp. 15873–15880, 2014.

[86] K. A. Dill and J. L. MacCallum, "The protein-folding problem, 50 years on," *Science*, vol. 338, no. 6110, pp. 1042–1046, 2012.

[87] A. R. Milosavljević, C. Nicolas, M. L. Ranković, F. Canon, C. Miron, and A. Giuliani, "K-Shell Excitation and Ionization of a Gas-Phase Protein: Interplay between Electronic Structure and Protein Folding," *Journal of Physical Chemistry Letters*, vol. 6, no. 16, pp. 3132–3138, 2015.

[88] J. Contreras-García, E. R. Johnson, S. Keinan, R. Chaudret, J. P. Piquemal, D. N. Beratan, and W. Yang, "NCIPLOT: A program for plotting noncovalent interaction regions," *Journal of Chemical Theory and Computation*, vol. 7, no. 3, pp. 625–632, 2011.

[89] R. Boto, F. Peccati, R. Laplaza, chaoyu Quan, A. Carbone, J.-P. Piquemal, Y. Maday, and J. Contreras-García, "NCIPLOT4: A New Step Towards a Fast Quantification of Noncovalent Interactions," vol. 6498, no. 2010, pp. 1–31, 2019.

[90] J. X. Lu, W. Qiang, W. M. Yau, C. D. Schwieters, S. C. Meredith, and R. Tycko, "XMolecular structure of β-amyloid fibrils in alzheimer's disease brain tissue," *Cell*, vol. 154, no. 6, p. 1257, 2013.

[91] F. Peccati, "NCIPLOT4 Guide for Biomolecules: An Analysis Tool for Noncovalent Interactions," *Journal of Chemical Information and Modeling*, vol. 60, no. 1, pp. 6–10, 2020.

[92] D. S. Dwyer, "Electronic properties of the amino acid side chains contribute to the structural preferences in protein folding," *Journal of Biomolecular Structure and Dynamics*, vol. 18, no. 6, pp. 881–892, 2001.

[93] D. S. Dwyer, "Electronic properties of amino acid side chains: quantum mechanics calculation of substituent effects," *BMC Chemical Biology*, vol. 5, no. 1, pp. 1–11, 2005.

[94] J. Faver and K. M. Merz Jr, "Utility of the hard/soft acid- base principle via the fukui function in biological systems," *Journal of Chemical Theory and Computation*, vol. 6, no. 2, pp. 548–559, 2010.

[95] G. A. Urquiza-Carvalho, W. D. Fragoso, and G. B. Rocha, "Assessment of semiempirical enthalpy of formation in solution as an effective energy function to discriminate native-like structures in protein decoy sets," *Journal of Computational Chemistry*, vol. 37, no. 21, pp. 1962–1972, 2016.

[96] R. Momen, A. Azizi, L. Wang, P. Yang, T. Xu, S. R. Kirk, W. Li, S. Manzhos, and S. Jenkins, "The role of weak interactions in characterizing peptide folding preferences using a qtaim interpretation of the ramachandran plot (ϕ-ψ)," *International Journal of Quantum Chemistry*, vol. 118, no. 2, p. e25456, 2018.

[97] A. Ianeselli, S. Orioli, G. Spagnolli, P. Faccioli, L. Cupellini, S. Jurinovich, and B. Mennucci, "Atomic Detail of Protein Folding Revealed by an Ab Initio Reappraisal of Circular Dichroism," *Journal of the American Chemical Society*, vol. 140, no. 10, pp. 3674–3682, 2018.

[98] M. Culka, J. Galgonek, J. Vymetal, J. Vondrasek, and L. Rulisek, "Toward ab initio protein folding: Inherent secondary structure propensity of short peptides from the bioinformatics and quantum-chemical perspective," *The Journal of Physical Chemistry B*, vol. 123, no. 6, pp. 1215–1227, 2019.

[99] M. Culka and L. Rulíšek, "Factors Stabilizing β-Sheets in Protein Structures from a Quantum-Chemical Perspective," *Journal of Physical Chemistry B*, vol. 123, no. 30, pp. 6453–6461, 2019.

[100] M. Culka and L. Rulíšek, "Interplay between Conformational Strain and Intramolecular Interaction in Protein Structures: Which of Them Is Evolutionarily Conserved?," *The Journal of Physical Chemistry B*, vol. 124, no. 16, pp. 3252–3260, 2020.

[101] A. Mauri, V. Consonni, and R. Todeschini, *Molecular Descriptors. In Handbook of Computational Chemistry*. Springer, Cham, 2017.

[102] M. Karelson, V. S. Lobanov, and A. R. Katritzky, "Quantum-chemical descriptors in QSAR/QSPR studies," *Chemical Reviews*, vol. 96, no. 3, pp. 1027–1044, 1996.

[103] C. F. Matta and A. A. Arabi, "Electron-density descriptors as predictors in quantitative structure-activity/property relationships and drug design," *Future Medicinal Chemistry*, vol. 3, no. 8, pp. 969–994, 2011.

[104] C. F. Matta, "Modeling biophysical and biological properties from the characteristics of the molecular electron density, electron localization and delocalization matrices, and the electrostatic potential," *Journal of Computational Chemistry*, vol. 35, no. 16, pp. 1165–1198, 2014.

[105] D. Chopra, "Advances in understanding of chemical bonding: Inputs from experimental and theoretical charge density analysis," *Journal of Physical Chemistry A*, vol. 116, no. 40, pp. 9791–9801, 2012.

[106] I. B. Grillo, G. A. Urquiza-Carvalho, and G. B. Rocha, "Primordia: A software to explore reactivity and electronic structure in large biomolecules," *Journal of Chemical Information and Modeling*, vol. 60, no. 12, pp. 5885–5890, 2020.

[107] I. B. Grillo, G. A. Urquiza-Carvalho, E. J. F. Chaves, and G. B. Rocha, "Semiempirical methods do fukui functions: Unlocking a modeling framework for biosystems," *Journal of Computational Chemistry*, 2020.

[108] I. B. Grillo, G. Urquiza-Carvalho, J. F. R. Bachega, and G. B. Rocha, "Elucidating enzymatic catalysis using fast quantum chemical descriptors," *Journal of Chemical Information and Modeling*, 2020.

[109] C. A. Sarisky and S. L. Mayo, "The $\beta\beta\alpha$ fold: Explorations in sequence space," *Journal of Molecular Biology*, vol. 307, no. 5, pp. 1411–1418, 2001.

[110] J. H. Cho, W. Meng, S. Sato, E. Y. Kim, H. Schindelin, and D. P. Raleigh, "Energetically significant networks of coupled interactions within an unfolded protein," *Proceedings of the National Academy of Sciences*, vol. 111, no. 33, pp. 12079–12084, 2014.

[111] S. T. Walsh, H. Cheng, J. W. Bryson, H. Roder, and W. F. Degrado, "Solution structure and dynamics of a de novo designed three-helix bundle protein," *Proceedings of the National Academy of Sciences*, vol. 96, no. 10, pp. 5486–5491, 1999.

[112] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How fast-folding proteins fold," *Science*, vol. 334, no. 6055, pp. 517–520, 2011.

[113] N. H. Morgon and K. Coutinho, *Métodos de Química Teórica e Modelagem Molecular*. Livraria da Física, São Paulo, 2007.

[114] A. M. Namba, V. B. da Silva, and C. Da Silva, "Dinâmica molecular: teoria e aplicações em planejamento de fármacos," *Eclética Química*, vol. 33, pp. 13–24, 2008.

[115] A. Warshel and M. Levitt, "Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme," *Journal of Molecular Biology*, vol. 103, no. 2, pp. 227–249, 1976.

[116] U. C. Singh and P. A. Kollman, "A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the ch3cl+ cl- exchange reaction and gas phase protonation of polyethers," *Journal of Computational Chemistry*, vol. 7, no. 6, pp. 718–730, 1986.

[117] M. J. Field, P. A. Bash, and M. Karplus, "A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations," *Journal of Computational Chemistry*, vol. 11, no. 6, pp. 700–733, 1990.

[118] S. Humbel, S. Sieber, and K. Morokuma, "The IMOMO method: Integration of different levels of molecular orbital approximations for geometry optimization of large systems: Test for n-butane conformation and SN2 reaction: RCl+Cl-," *Journal of Chemical Physics*, vol. 105, no. 5, pp. 1959–1967, 1996.

[119] R. D. Froese and K. Morokuma, "The imomo and imonm methods for excited states. a study of the adiabatic s0→ t1, 2 excitation energies of cyclic alkenes and enones," *Chemical Physics Letters*, vol. 263, no. 3-4, pp. 393–400, 1996.

[120] M. Svensson, S. Humbel, R. D. Froese, T. Matsubara, S. Sieber, and K. Morokuma, "ONIOM: A multilayered integrated MO + MM method for geometry optimizations and single point energy predictions. A test for Diels-Alder reactions and Pt(P(t-Bu)3)2 + H2 oxidative addition," *Journal of Physical Chemistry*, vol. 100, no. 50, pp. 19357–19363, 1996.

[121] M. S. Skaf, "O Prêmio Nobel de Química 2013," *Química Nova na escola*, vol. 35, no. 4, pp. 243–246, 2013.

[122] J. Rezac, J. Fanfrlik, D. Salahub, and P. Hobza, "Semiempirical quantum chemical pm6 method augmented by dispersion and h-bonding correction terms reliably describes various types of noncovalent complexes," *Journal of Chemical Theory and Computation*, vol. 5, no. 7, pp. 1749–1760, 2009.

[123] G. A. Kaminski and W. L. Jorgensen, "A quantum mechanical and molecular mechanical method based on cm1a charges: applications to solvent effects on organic equilibria and reactions," *The Journal of Physical Chemistry B*, vol. 102, no. 10, pp. 1787–1796, 1998.

[124] J. W. Storer, D. J. Giesen, C. J. Cramer, and D. G. Truhlar, "Class iv charge models: A new semiempirical approach in quantum chemistry," *Journal of Computer-Aided Molecular Design*, vol. 9, no. 1, pp. 87–110, 1995.

[125] E. Cubero, F. J. Luque, M. Orozco, and J. Gao, "Perturbation approach to combined qm/mm simulation of solute- solvent interactions in solution," *The Journal of Physical Chemistry B*, vol. 107, no. 7, pp. 1664–1671, 2003.

[126] P. L. Muiño and P. R. Callis, "Hybrid simulations of solvation effects on electronic spectra: indoles in water," *The Journal of Chemical Physics*, vol. 100, no. 6, pp. 4093–4109, 1994.

[127] Q. Cui and M. Karplus, "Molecular properties from combined qm/mm methods. i. analytical second derivative and vibrational calculations," *The Journal of Chemical Physics*, vol. 112, no. 3, pp. 1133–1149, 2000.

[128] M. A. Thompson and G. K. Schenter, "Excited states of the bacteriochlorophyll b dimer of rhodopseudomonas viridis: a qm/mm study of the photosynthetic reaction center that includes mm polarization," *The Journal of Physical Chemistry*, vol. 99, no. 17, pp. 6374–6386, 1995.

[129] M. A. Thompson, "Qm/mmpol: A consistent model for solute/solvent polarization. application to the aqueous solvation and spectroscopy of formaldehyde, acetaldehyde, and acetone," *The Journal of Physical Chemistry*, vol. 100, no. 34, pp. 14492–14507, 1996.

[130] G. G. Ferenczy, J.-L. Rivail, P. R. Surján, and G. Náray-Szabó, "Nddo fragment self-consistent field approximation for large electronic systems," *Journal of Computational Chemistry*, vol. 13, no. 7, pp. 830–837, 1992.

[131] D. M. Philipp and R. A. Friesner, "Mixed ab initio qm/mm modeling using frozen orbitals and tests with alanine dipeptide and tetrapeptide," *Journal of Computational Chemistry*, vol. 20, no. 14, pp. 1468–1494, 1999.

[132] J. Gao, P. Amara, C. Alhambra, and M. J. Field, "A generalized hybrid orbital (gho) method for the treatment of boundary atoms in combined qm/mm calculations," *The Journal of Physical Chemistry A*, vol. 102, no. 24, pp. 4714–4721, 1998.

[133] M. Garcia-Viloca and J. Gao, "Generalized hybrid orbital for the treatment of boundary atoms in combined quantum mechanical and molecular mechanical calculations using the semiempirical parameterized model 3 method," *Theoretical Chemistry Accounts*, vol. 111, no. 2, pp. 280–286, 2004.

[134] J. Pu, J. Gao, and D. G. Truhlar, "Generalized hybrid orbital (gho) method for combining ab initio hartree- fock wave functions with molecular mechanics," *The Journal of Physical Chemistry A*, vol. 108, no. 4, pp. 632–650, 2004.

[135] K. Coutinho and S. Canuto, "Solvent effects from a sequential monte carlo-quantum mechanical approach," in *Advances in Quantum Chemistry*, vol. 28, pp. 89–105, Elsevier, 1997.

[136] S. Canuto, K. Coutinho, and M. C. Zerner, "Including dispersion in configuration interaction-singles calculations for the spectroscopy of chromophores in solution," *The Journal of Chemical Physics*, vol. 112, no. 17, pp. 7293–7299, 2000.

[137] T. Vasilevskaya and W. Thiel, "Periodic boundary conditions in qm/mm calculations: Implementation and tests," *Journal of Chemical Theory and Computation*, vol. 12, no. 8, pp. 3561–3570, 2016.

[138] P. L. Freddolino, C. B. Harrison, Y. Liu, and K. Schulten, "Challenges in protein-folding simulations," *Nature Physics*, vol. 6, no. 10, pp. 751–758, 2010.

[139] T. Veitshans, D. Klimov, and D. Thirumalai, "Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties," *Folding and Design*, vol. 2, no. 1, pp. 1–22, 1997.

[140] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, "Molecular simulation of ab initio protein folding for a millisecond folder ntl9 (1- 39)," *Journal of the American Chemical Society*, vol. 132, no. 5, pp. 1526–1528, 2010.

[141] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov chain Monte Carlo in practice*. CRC press, 1995.

[142] W. G. Hoover, *Molecular Dynamics*. Springer-Verlag, 1986.

[143] T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande, "To milliseconds and beyond: challenges in the simulation of protein folding," *Current Opinion in Structural Biology*, vol. 23, no. 1, pp. 58–65, 2013.

[144] S. Piana, J. L. Klepeis, and D. E. Shaw, "Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations," *Current Opinion in Structural Biology*, vol. 24, pp. 98–105, 2014.

[145] K. A. Dill, K. M. Fiebig, and H. S. Chan, "Cooperativity in protein-folding kinetics.," *Proceedings of the National Academy of Sciences*, vol. 90, no. 5, pp. 1942–1946, 1993.

[146] J. L. Klepeis, K. Lindorff-Larsen, R. O. Dror, and D. E. Shaw, "Long-timescale molecular dynamics simulations of protein structure and function," *Current Opinion in Structural Biology*, vol. 19, no. 2, pp. 120–127, 2009.

[147] P. L. Freddolino, F. Liu, M. Gruebele, and K. Schulten, "Ten-microsecond molecular dynamics simulation of a fast-folding ww domain," *Biophysical Journal*, vol. 94, no. 10, pp. L75–L77, 2008.

[148] M. E. Karpen, D. J. Tobias, and C. L. Brooks III, "Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of ypgdv," *Biochemistry*, vol. 32, no. 2, pp. 412–420, 1993.

[149] V. Daggett, "Protein folding- simulation," *Chemical Reviews*, vol. 106, no. 5, pp. 1898–1916, 2006.

[150] F. C. Almeida, K. Sanches, R. Pinheiro-Aguiar, V. S. Almeida, and I. P. Caruso, "Protein surface interactions—theoretical and experimental studies," *Frontiers in Molecular Biosciences*, vol. 8, 2021.

[151] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, "Atomic-level description of ubiquitin folding," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5915–5920, 2013.

[152] C. D. Snow, H. Nguyen, V. S. Pande, and M. Gruebele, "Absolute comparison of simulated and experimental protein-folding dynamics," *Nature*, vol. 420, no. 6911, pp. 102–106, 2002.

[153] G. Hähner, "Near edge x-ray absorption fine structure spectroscopy as a tool to probe electronic and structural properties of thin organic films and liquids," *Chemical Society Reviews*, vol. 35, no. 12, pp. 1244–1255, 2006.

[154] C. Bustamante, L. Alexander, K. Maciuba, and C. M. Kaiser, "Single-molecule studies of protein folding with optical tweezers," *Annual Review of Biochemistry*, vol. 89, pp. 443–470, 2020.

[155] H. A. Scheraga, M. Khalili, and A. Liwo, "Protein-Folding Dynamics: Overview of Molecular Simulation Techniques," *Annual Review of Physical Chemistry*, vol. 58, no. 1, pp. 57–83, 2007.

[156] Y. I. Yang, Q. Shao, J. Zhang, L. Yang, and Y. Q. Gao, "Enhanced sampling in molecular dynamics," *The Journal of Chemical Physics*, vol. 151, no. 7, p. 070902, 2019.

[157] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chemical Physics Letters*, vol. 314, no. 1-2, pp. 141–151, 1999.

[158] H. Kamberaj, "Faster protein folding using enhanced conformational sampling of molecular dynamics simulation," *Journal of Molecular Graphics and Modelling*, vol. 81, pp. 32–49, 2018.

[159] K. Jain, O. Ghribi, and J. Delhommelle, "Folding free-energy landscape of α-synuclein (35–97) via replica exchange molecular dynamics," *Journal of Chemical Information and Modeling*, vol. 61, no. 1, pp. 432–443, 2020.

[160] H. Meshkin and F. Zhu, "Thermodynamics of protein folding studied by umbrella sampling along a reaction coordinate of native contacts," *Journal of Chemical Theory and Computation*, vol. 13, no. 5, pp. 2086–2097, 2017.

[161] F. Baftizadeh, P. Cossio, F. Pietrucci, and A. Laio, "Protein folding and ligand-enzyme binding from bias-exchange metadynamics simulations," *Current Physical Chemistry*, vol. 2, no. 1, pp. 79–91, 2012.

[162] Y. Bian, J. Zhang, J. Wang, and W. Wang, "On the accuracy of metadynamics and its variations in a protein folding process," *Molecular Simulation*, vol. 41, no. 9, pp. 752–763, 2015.

[163] D. Kimanius, I. Pettersson, G. Schluckebier, E. Lindahl, and M. Andersson, "Saxs-guided metadynamics," *Journal of Chemical Theory and Computation*, vol. 11, no. 7, pp. 3491–3498, 2015.

[164] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[165] Y. M. Rhee, E. J. Sorin, G. Jayachandran, E. Lindahl, and V. S. Pande, "Simulations of the role of water in the protein-folding mechanism," *Proceedings of the National Academy of Sciences*, vol. 101, no. 17, pp. 6456–6461, 2004.

[166] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, "How robust are protein folding simulations with respect to force field parameterization?," *Biophysical Journal*, vol. 100, no. 9, pp. L47–L49, 2011.

[167] T. R. Sosnick and D. Barrick, "The folding of single domain proteins—have we reached a consensus?," *Current Opinion in Structural Biology*, vol. 21, no. 1, pp. 12–24, 2011.

[168] L. Sborgi, A. Verma, S. Piana, K. Lindorff-Larsen, M. Cerminara, C. M. Santiveri, D. E. Shaw, E. de Alba, and V. Munoz, "Interaction networks in protein folding via atomic-resolution experiments and long-time-scale molecular dynamics simulations," *Journal of the American Chemical Society*, vol. 137, no. 20, pp. 6506–6516, 2015.

[169] H. O. Verli, *Bioinformática: da Biologia à Flexibilidade Molecular*. Sociedade Brasileira de Bioquímica e Biologia Molecular, São Paulo, 2007.

[170] G. Hummer, "From transition paths to transition states and rate coefficients," *The Journal of Chemical Physics*, vol. 120, no. 2, pp. 516–523, 2004.

[171] R. Bowman Gregory, "An introduction to markov state models and their application to long timescale molecular simulation. bowman gr, pande vs, noé f, editors," 2014.

[172] T. C. Correra and A. d. C. Frank, "A equação-mestra: atingindo o equilíbrio," *Química Nova*, vol. 34, pp. 346–353, 2011.

[173] P. Rufino, *Uma Iniciação aos Sistemas Dinâmicos Estocásticos: 27 Colóquio Brasileiro de Matemática*. IMPA, Rio de Janeiro, 2009.

[174] H. Anton and C. Rorres, *Álgebra linear com aplicações*, vol. 8. Bookman Porto Alegre, 2001.

[175] M. A. Campos, L. C. Rêgo, and A. F. de Mendonça, *Métodos probabilísticos e estatísticos com aplicações em engenharias e ciências exatas*. Grupo Gen-LTC, 2017.

[176] B. Das and G. Gangopadhyay, "Master equation approach to single oligomeric enzyme catalysis: Mechanically controlled further catalysis," *The Journal of Chemical Physics*, vol. 132, no. 13, p. 135102, 2010.

[177] J. M. Vilar, H. Y. Kueh, N. Barkai, and S. Leibler, "Mechanisms of noise-resistance in genetic oscillators," *Proceedings of the National Academy of Sciences*, vol. 99, no. 9, pp. 5988–5992, 2002.

[178] V. Vaks, "Master equation approach to the configurational kinetics of nonequilibrium alloys: Exact relations, h-theorem, and cluster approximations," *Journal of Experimental and Theoretical Physics Letters*, vol. 63, no. 6, pp. 471–477, 1996.

[179] M. J. Pilling and S. H. Robertson, "Master equation models for chemical reactions of importance in combustion," *Annual Review of Physical Chemistry*, vol. 54, no. 1, pp. 245–275, 2003.

[180] S. H. Robertson, M. J. Pilling, L. C. Jitariu, and I. H. Hillier, "Master equation methods for multiple well systems: application to the 1-, 2-pentyl system," *Physical Chemistry Chemical Physics*, vol. 9, no. 31, pp. 4085–4097, 2007.

[181] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, "A direct approach to conformational dynamics based on hybrid monte carlo," *Journal of Computational Physics*, vol. 151, no. 1, pp. 146–168, 1999.

[182] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, "Pyemma 2: A software package for estimation, validation, and analysis of markov models," *Journal of Chemical Theory and Computation*, vol. 11, no. 11, pp. 5525–5542, 2015.

[183] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics: Generation and validation," *The Journal of Chemical Physics*, vol. 134, no. 17, p. 174105, 2011.

[184] N. Djurdjevac, M. Sarich, and C. Schütte, "Estimating the eigenvalue error of markov state models," *Multiscale Modeling & Simulation*, vol. 10, no. 1, pp. 61–81, 2012.

[185] R. T. McGibbon and V. S. Pande, "Variational cross-validation of slow dynamical modes in molecular kinetics," *The Journal of Chemical Physics*, vol. 142, no. 12, p. 03B621_1, 2015.

[186] H. Wu and F. Noé, "Optimal estimation of free energies and stationary densities from multiple biased simulations," *Multiscale Modeling & Simulation*, vol. 12, no. 1, pp. 25–54, 2014.

[187] A. S. Mey, H. Wu, and F. Noé, "xtram: Estimating equilibrium expectations from time-correlated simulation data at multiple thermodynamic states," *Physical Review X*, vol. 4, no. 4, p. 041018, 2014.

[188] E. Rosta and G. Hummer, "Free energies from dynamic weighted histogram analysis using unbiased markov state model," *Journal of Chemical Theory and Computation*, vol. 11, no. 1, pp. 276–285, 2015.

[189] H. Wu, A. S. Mey, E. Rosta, and F. Noé, "Statistically optimal analysis of state-discretized trajectory data from multiple thermodynamic states," *The Journal of Chemical Physics*, vol. 141, no. 21, p. 12B629_1, 2014.

[190] K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, "Msmbuilder2: modeling conformational dynamics on the picosecond to millisecond scale," *Journal of Chemical Theory and Computation*, vol. 7, no. 10, pp. 3412–3419, 2011.

[191] F. Noé and C. Clementi, "Kinetic distance and kinetic maps from molecular dynamics simulation," *Journal of Chemical Theory and Computation*, vol. 11, no. 10, pp. 5002–5011, 2015.

[192] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[193] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of Educational Psychology*, vol. 24, no. 6, p. 417, 1933.

[194] T. F. Reubold, K. Faelber, N. Plattner, Y. Posor, K. Ketel, U. Curth, J. Schlegel, R. Anand, D. J. Manstein, F. Noé, *et al.*, "Crystal structure of the dynamin tetramer," *Nature*, vol. 525, no. 7569, pp. 404–408, 2015.

[195] F. Noé and F. Nuske, "A variational approach to modeling slow processes in stochastic dynamical systems," *Multiscale Modeling & Simulation*, vol. 11, no. 2, pp. 635–655, 2013.

[196] G. Voronoi, "Recherches sur les paralléloèdres primitives," *Journal fur die Reine und Angewandte Mathematik*, vol. 134, pp. 198–287, 1908.

[197] J. Shao, S. W. Tanner, N. Thompson, and 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.

[198] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," tech. rep., Stanford, 2006.

[199] S. Röblitz and M. Weber, "Fuzzy spectral clustering by pcca+: application to markov state models and data classification," *Advances in Data Analysis and Classification*, vol. 7, no. 2, pp. 147–179, 2013.

[200] F. Noé, H. Wu, J.-H. Prinz, and N. Plattner, "Projected and hidden markov models for calculating kinetics and metastable states of complex molecules," *The Journal of Chemical Physics*, vol. 139, no. 18, p. 11B609_1, 2013.

[201] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, "Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 19011–19016, 2009.

[202] M. Pirchi, G. Ziv, I. Riven, S. S. Cohen, N. Zohar, Y. Barak, and G. Haran, "Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein," *Nature Communications*, vol. 2, no. 1, pp. 1–7, 2011.

[203] M. Held, P. Metzner, J.-H. Prinz, and F. Noé, "Mechanisms of protein-ligand association and its modulation by protein mutations," *Biophysical Journal*, vol. 100, no. 3, pp. 701–710, 2011.

[204] D.-A. Silva, G. R. Bowman, A. Sosa-Peinado, and X. Huang, "A role for both conformational selection and induced fit in ligand binding by the lao protein," *PLoS Computational Biology*, vol. 7, no. 5, p. e1002054, 2011.

[205] R. Saito, J. M. Pruet, L. A. Manzano, K. Jasheway, A. F. Monzingo, P. A. Wiget, I. Kamat, E. V. Anslyn, and J. D. Robertus, "Peptide-conjugated pterins as inhibitors of ricin toxin a," *Journal of Medicinal Chemistry*, vol. 56, no. 1, pp. 320–329, 2013.

[206] J. M. Pruet, R. Saito, L. A. Manzano, K. R. Jasheway, P. A. Wiget, I. Kamat, E. V. Anslyn, and J. D. Robertus, "Optimized 5-membered heterocycle-linked pterins for the inhibition of ricin toxin a," *ACS Medicinal Chemistry Letters*, vol. 3, no. 7, pp. 588–591, 2012.

[207] P. A. Wiget, L. A. Manzano, J. M. Pruet, G. Gao, R. Saito, A. F. Monzingo, K. R. Jasheway, J. D. Robertus, and E. V. Anslyn, "Sulfur incorporation generally improves ricin inhibition in pterin-appended glycine-phenylalanine dipeptide mimics," *Bioorganic & Medicinal Chemistry Letters*, vol. 23, no. 24, pp. 6799–6804, 2013.

[208] J. M. Pruet, K. R. Jasheway, L. A. Manzano, Y. Bai, E. V. Anslyn, and J. D. Robertus, "7-Substituted pterins provide a new direction for ricin A chain inhibitors," *European Journal of Medicinal Chemistry*, vol. 46, no. 9, pp. 3608–3615, 2011.

[209] G. B. Rocha, R. O. Freire, A. M. Simas, and J. J. Stewart, "Rm1: A reparameterization of am1 for h, c, n, o, p, s, f, cl, br, and i," *Journal of Computational Chemistry*, vol. 27, no. 10, pp. 1101–1111, 2006.

[210] J. J. Stewart, "Optimization of parameters for semiempirical methods v: modification of nddo approximations and application to 70 elements," *Journal of Molecular Modeling*, vol. 13, no. 12, pp. 1173–1213, 2007.

[211] M. Korth, "Third-generation hydrogen-bonding corrections for semiempirical qm methods and force fields," *Journal of Chemical Theory and Computation*, vol. 6, no. 12, pp. 3808–3816, 2010.

[212] J. J. Stewart, "Optimization of parameters for semiempirical methods vi: more modifications to the nddo approximations and re-optimization of parameters," *Journal of Molecular Modeling*, vol. 19, no. 1, pp. 1–32, 2013.

[213] J. J. Stewart, "Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations," *International Journal of Quantum Chemistry*, vol. 58, no. 2, pp. 133–146, 1996.

[214] J. J. Stewart, "Mopac: a semiempirical molecular orbital program," *Journal of Computer-Aided Molecular Design*, vol. 4, no. 1, pp. 1–103, 1990.

[215] S. Dapprich, I. Komáromi, K. S. Byun, K. Morokuma, and M. J. Frisch, "A new ONIOM implementation in Gaussian98. Part I. The calculation of energies, gradients, vibrational frequencies and electric field derivatives," *Journal of Molecular Structure: THEOCHEM*, vol. 461-462, pp. 1–21, 1999.

[216] L. W. Chung, W. M. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. Ke, F. Liu, H. B. Li, L. Ding, and K. Morokuma, "The ONIOM Method and Its Applications," *Chemical Reviews*, vol. 115, no. 12, pp. 5678–5796, 2015.

[217] M. Frisch, G. Trucks, H. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. Petersson, and *et al. Gaussian 09,*. Revision D. 01; Gaussian, Inc.: Wallingford, CT, 2009.

[218] C. Lee, W. Yang, and R. G. Parr, "Development of the colle-salvetti correlation-energy formula into a functional of the electron density," *Physical Review B*, vol. 37, no. 2, p. 785, 1988.

[219] A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange," *The Journal of Chemical Physics*, vol. 98, no. 7, pp. 5648–5652, 1993.

[220] J.-D. Chai and M. Head-Gordon, "Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections," *Physical Chemistry Chemical Physics*, vol. 10, no. 44, pp. 6615–6620, 2008.

[221] S. Grimme, "Semiempirical gga-type density functional constructed with a long-range dispersion correction," *Journal of Computational Chemistry*, vol. 27, no. 15, pp. 1787–1799, 2006.

[222] W. J. Hehre, R. Ditchfield, and J. A. Pople, "Self—consistent molecular orbital methods. xii. further extensions of gaussian—type basis sets for use in molecular orbital studies of organic molecules," *The Journal of Chemical Physics*, vol. 56, no. 5, pp. 2257–2261, 1972.

[223] R. Krishnan, J. S. Binkley, R. Seeger, and J. A. Pople, "Self-consistent molecular orbital methods. xx. a basis set for correlated wave functions," *The Journal of Chemical Physics*, vol. 72, no. 1, pp. 650–654, 1980.

[224] M. M. Francl, W. J. Pietro, W. J. Hehre, J. S. Binkley, M. S. Gordon, D. J. DeFrees, and J. A. Pople, "Self-consistent molecular orbital methods. xxiii. a polarization-type basis set for second-row elements," *The Journal of Chemical Physics*, vol. 77, no. 7, pp. 3654–3665, 1982.

[225] A. K. Rappé, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, "UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations," *Journal of the American Chemical Society*, vol. 114, no. 25, pp. 10024–10035, 1992.

[226] P. Grela, M. Szajwaj, P. Horbowicz-Drozdzal, and M. Tchórzewski, "How ricin damages the ribosome," *Toxins*, vol. 11, no. 5, pp. 1–16, 2019.

[227] P. Geerlings, F. De Proft, and W. Langenaeker, "Conceptual density functional theory.," *Chemical Reviews*, vol. 103, no. 5, pp. 1793–1873, 2003.

[228] R. G. Pearson, "Recent advances in the concept of hard and soft acids and bases," *Journal of Chemical Education*, vol. 64, no. 7, p. 561, 1987.

[229] M. Torrent-Sucarrat, F. De Proft, P. Geerlings, and P. W. Ayers, "Do the local softness and hardness indicate the softest and hardest regions of a molecule?," *Chemistry–A European Journal*, vol. 14, no. 28, pp. 8652–8660, 2008.

[230] J. Sánchez-Márquez, D. Zorrilla, V. García, and M. Fernández, "Introducing a new methodology for the calculation of local philicity and multiphilic descriptor: an alternative to the finite difference approximation," *Molecular Physics*, vol. 116, no. 13, pp. 1737–1748, 2018.

[231] U. Sarkar, D. Roy, P. Chattaraj, R. Parthasarathi, J. Padmanabhan, and V. Subramanian, "A conceptual dft approach towards analysing toxicity," *Journal of Chemical Sciences*, vol. 117, no. 5, pp. 599–612, 2005.

[232] K. M. Merz and J. Faver, "Utility of the hard/soft acid-base principle via the fukui function in biological systems," *Journal of Chemical Theory and Computation*, vol. 6, no. 2, pp. 548–559, 2010.

[233] I. B. Grillo, J. F. R. Bachega, L. F. S. Timmers, R. A. Caceres, O. N. de Souza, M. J. Field, and G. B. Rocha, "Theoretical characterization of the shikimate 5-dehydrogenase reaction from mycobacterium tuberculosis by hybrid qc/mm simulations and quantum chemical descriptors," *Journal of Molecular Modeling*, vol. 26, no. 11, pp. 1–12, 2020.

[234] K. Fukushima, M. Wada, and M. Sakurai, "An insight into the general relationship between the three dimensional structures of enzymes and their electronic wave functions: Implication for the prediction of functional sites of enzymes," *Proteins: Structure, Function, and Bioinformatics*, vol. 71, no. 4, pp. 1940–1954, 2008.

[235] M. Torrent-Sucarrat, F. De Proft, P. W. Ayers, and P. Geerlings, "On the applicability of local softness and hardness.," *Physical Chemistry Chemical Physics*, vol. 12, no. 5, pp. 1072–1080, 2010.

[236] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, "Mdtraj: a modern open library for the analysis of molecular dynamics trajectories," *Biophysical Journal*, vol. 109, no. 8, pp. 1528–1532, 2015.

[237] K. A. Beauchamp, R. McGibbon, Y.-S. Lin, and V. S. Pande, "Simple few-state models reveal hidden complexity in protein folding," *Proceedings of the National Academy of Sciences*, vol. 109, no. 44, pp. 17807–17813, 2012.

[238] A. Dickson and C. L. Brooks III, "Native states of fast-folding proteins are kinetic traps," *Journal of the American Chemical Society*, vol. 135, no. 12, pp. 4729–4734, 2013.

[239] T. J. Lane, G. R. Bowman, K. Beauchamp, V. A. Voelz, and V. S. Pande, "Markov state model reveals folding and functional dynamics in ultra-long md trajectories," *Journal of the American Chemical Society*, vol. 133, no. 45, pp. 18413–18419, 2011.

[240] E. Suárez, J. L. Adelman, and D. M. Zuckerman, "Accurate estimation of protein folding and unfolding times: beyond markov state models," *Journal of Chemical Theory and Computation*, vol. 12, no. 8, pp. 3473–3481, 2016.

[241] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, *et al.*, "Anton, a special-purpose machine for molecular dynamics simulation," *Communications of the ACM*, vol. 51, no. 7, pp. 91–97, 2008.

[242] J. J. Stewart, "Stewart computational chemistry," *http://openmopac. net/*, 2007.

[243] S. Seritan, C. Bannwarth, B. S. Fales, E. G. Hohenstein, S. I. Kokkila-Schumacher, N. Luehr, J. W. Snyder Jr, C. Song, A. V. Titov, I. S. Ufimtsev, *et al.*, "Terachem: Accelerating electronic structure and ab initio molecular dynamics with graphical processing units," *The Journal of Chemical Physics*, vol. 152, no. 22, p. 224110, 2020.

[244] S. Seritan, C. Bannwarth, B. S. Fales, E. G. Hohenstein, C. M. Isborn, S. I. Kokkila-Schumacher, X. Li, F. Liu, N. Luehr, J. W. Snyder Jr, *et al.*, "Terachem: A graphical processing unit-accelerated electronic structure package for large-scale ab initio molecular dynamics," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, p. e1494, 2020.

[245] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu," *The Journal of Chemical Physics*, vol. 132, no. 15, p. 154104, 2010.

[246] F. Liu, N. Luehr, H. J. Kulik, and T. J. Martínez, "Quantum chemistry for solvated molecules on graphical processing units using polarizable continuum models," *Journal of Chemical Theory and Computation*, vol. 11, no. 7, pp. 3131–3144, 2015.

[247] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, "Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1045–D1053, 2015.

[248] A. Pecina, J. Brynda, L. Vrzal, R. Gnanasekaran, M. Hořejší, S. M. Eyrilmez, J. Řezáč, M. Lepšík, P. Řezáčová, P. Hobza, P. Majer, V. Veverka, and J. Fanfrlík, "Ranking Power of the SQM/COSMO Scoring Function on Carbonic Anhydrase II–Inhibitor Complexes," *ChemPhysChem*, vol. 19, no. 7, pp. 873–879, 2018.

[249] N. D. Yilmazer and M. Korth, "Enhanced semiempirical qm methods for biomolecular interactions," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 169–175, 2015.

[250] J. M. Pruet, K. R. Jasheway, L. A. Manzano, Y. Bai, E. V. Anslyn, and J. D. Robertus, "7-substituted pterins provide a new direction for ricin a chain inhibitors," *European Journal of Medicinal Chemistry*, vol. 46, no. 9, pp. 3608–3615, 2011.

[251] A. V. Sulimov, D. C. Kutov, E. V. Katkova, and V. B. Sulimov, "Combined docking with classical force field and quantum chemical semiempirical method pm7," *Advances in Bioinformatics*, vol. 2017, 2017.

[252] J. H. Carra, C. A. McHugh, S. Mulligan, L. A. M. Machiesky, A. S. Soares, and C. B. Millard, "Fragment-based identification of determinants of conformational and spectroscopic change at the ricin active site," *BMC Structural Biology*, vol. 7, pp. 1–11, 2007.

[253] E. J. F. Chaves, L. E. Gomes da Cruz, I. Q. M. Padilha, C. H. Silveira, D. A. M. Araujo, and G. B. Rocha, "Discovery of rta ricin subunit inhibitors: a computational study using pm7 quantum chemical method and steered molecular dynamics," *Journal of Biomolecular Structure and Dynamics*, pp. 1–19, 2021.

[254] R. E. Rocha, E. J. Chaves, P. H. Fischer, L. S. Costa, I. B. Grillo, L. E. da Cruz, F. C. Guedes, C. H. da Silveira, M. T. Scotti, A. D. Camargo, *et al.*, "A higher flexibility at the sars-cov-2 main protease active site compared to sars-cov and its potentialities for new inhibitor virtual screening targeting multi-conformers," *Journal of Biomolecular Structure and Dynamics*, pp. 1–21, 2021.

APÊNDICE A

APÊNDICE A

A.1 QCMDs globais obtidos via método semiempírico PM7

Foram realizados cálculos single-point de 100 conformações representativas (para cada uma das proteínas estudadas) do caminho de enovelamento/desenovelamento , através do método PM7 e modelo implícito de solvente COSMO com o programa MOPAC. Através desses dados obtivemos os valores da energia total (E_TOT) e do calor de formação (ΔH_f). Com o uso do programa PRIMoRDIA, obtivemos os seguintes descritores globais de reatividade: potencial de ionização (PI) , afinidade eletrônica (AE), potencial químico (μ), dureza química (η), moleza química (S), eletrofilicidade (ω) e número máximo de elétrons (n_{Max}) Além disso, realizamos os cálculos de descritores estruturais: RMSD-Cα e fração de contatos nativos (Q) com o objetivo de avaliarmos se esses apresentam alguma correlação com os descritores globais de reatividade ou com valores de (E_{TOT}) e (ΔH_f) . Nas figuras A.1, A.2 e A.3 a seguir são apresentados os dados de correlação entre todos os descritores.

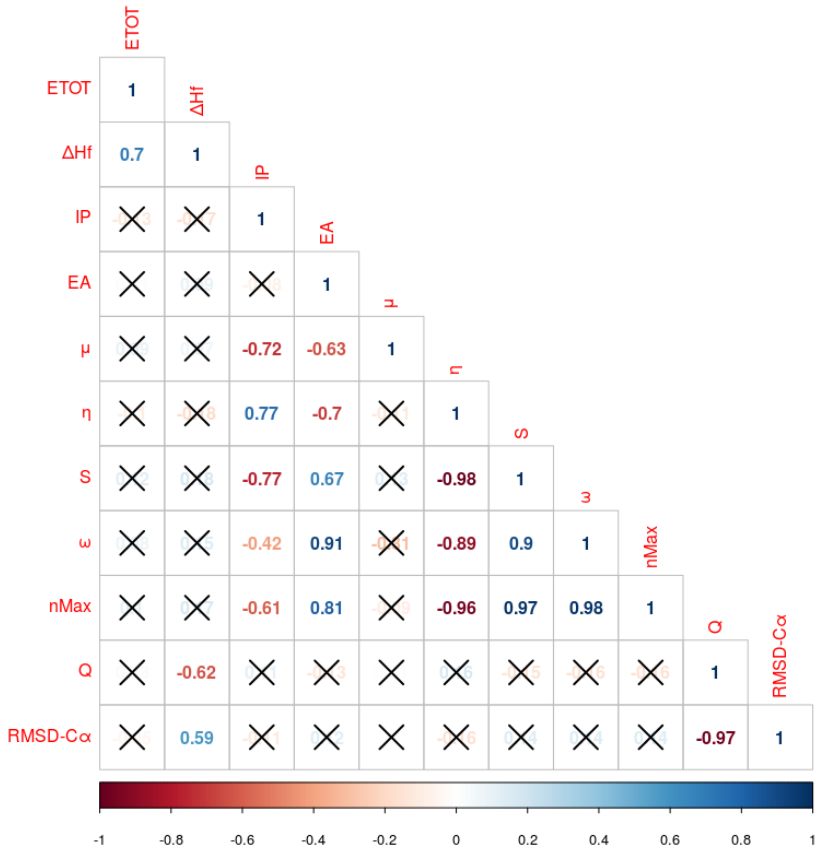


Figura A.1: Mapa de correlação entre os descritores globais de reatividade, E_{TOT} e ΔH_f (obtidos via PM7) e os descritores estruturais Q e RMSD-Cα para a primeira coordenada TICA da NTL9.

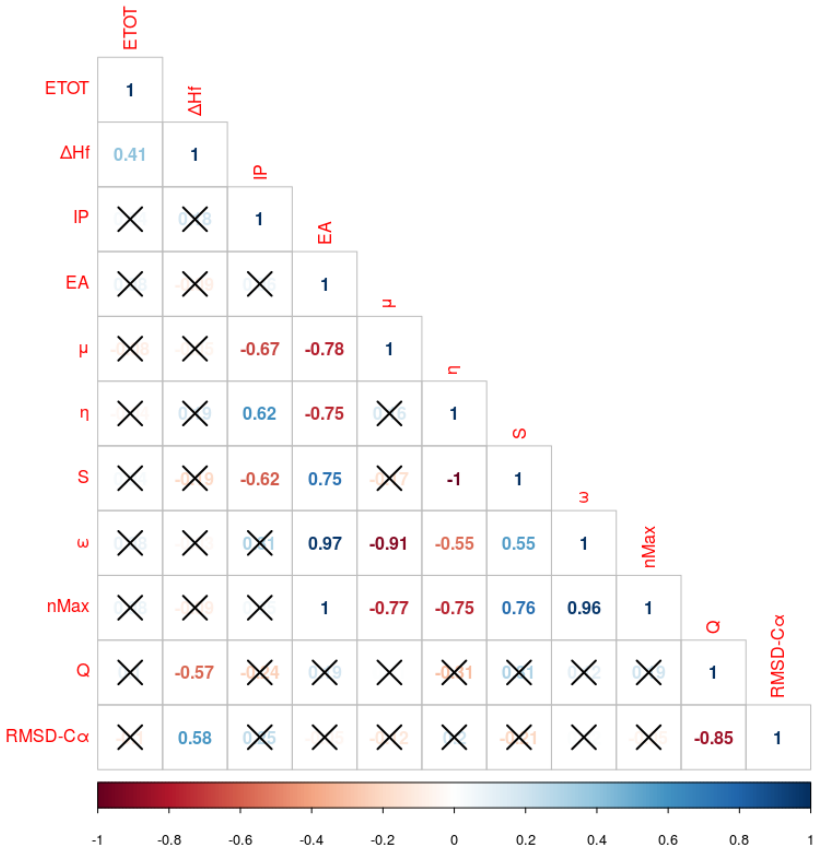


Figura A.2: Mapa de correlação entre os descritores globais de reatividade, E_{TOT} e ΔH_f (obtidos via PM7) e os descritores estruturais Q e RMSD-Cα para a primeira coordenada TICA da BBA.

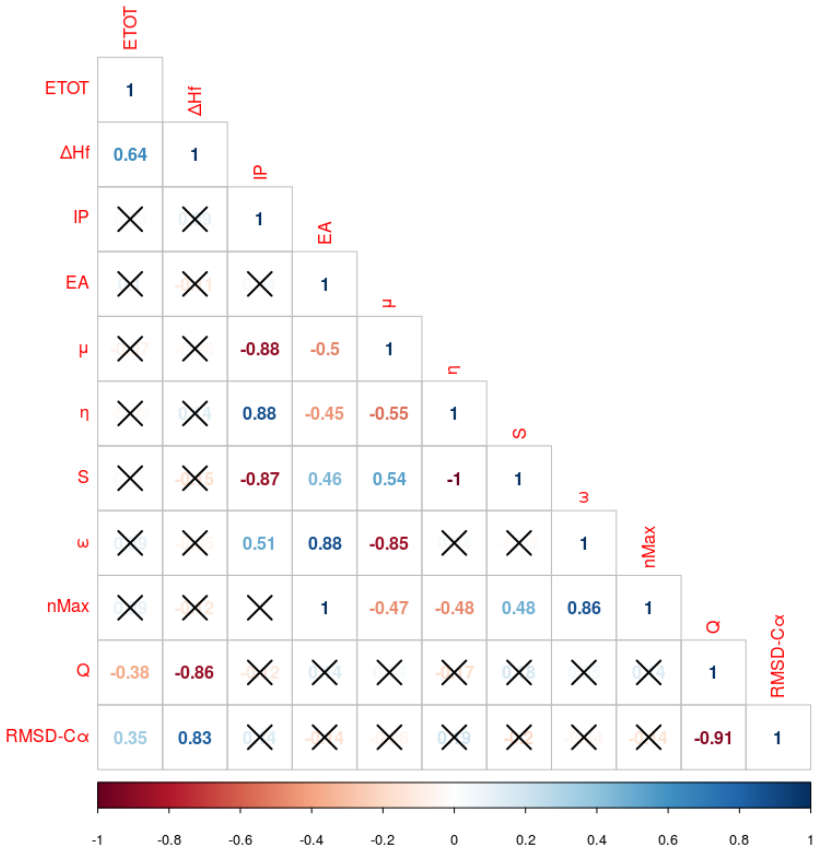


Figura A.3: Mapa de correlação entre os descritores globais de reatividade, E_{TOT} e ΔH_f (obtidos via PM7) e os descritores estruturais Q e RMSD-Cα para a primeira coordenada TICA da α3D.

Uma maneira útil de ampliarmos a nossa análise consiste em plotarmos, além da correlação, um histograma com a linha de densidade e linha tendência entre as variáveis dos nossos dados. Na figuras A.4, A.5 e A.6, apresentamos o histograma de todas as variáveis do nosso conjunto de dados para as proteínas NTL9, BBA e α3D respectivamente.

Na parte superior desse gráfico é apresentado um histograma com uma linha de densidade para cada uma das variáveis. Além disso há a comparação entre cada uma das variáveis em que o valor do em negrito corresponde ao coeficiente de correlação. Os asteriscos correspondem ao valor p do seguinte modo: 3 asteriscos ($p < 0,001$) , 2 asteriscos ($0.001 < p < 0.01$) , 1 asterisco ($0.01 < p < 0.05$) , 1 ponto ($0.05 < p < 0.10$) e caso ($p > 0.10$), o coeficiente de correlação aparece sem nenhum asterisco ou ponto. Na parte inferior do gráfico há uma plotagem entre cada uma das variáveis em que temos uma variável no eixo x e outra variável no eixo y. Há pontos para verificarmos como essas variáveis se relacionam e uma linha de tendência vermelha mostrando como esses pontos deveriam estar se a relação fosse perfeita.

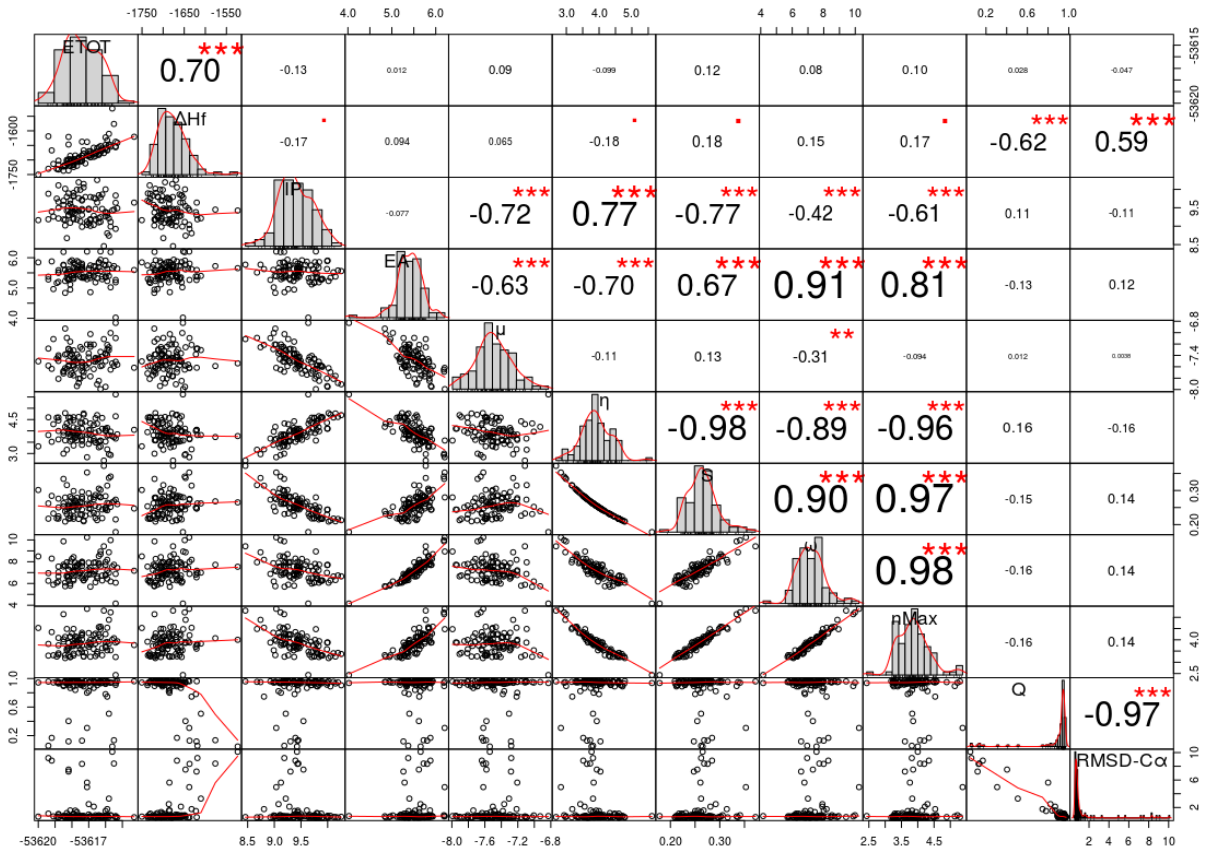


Figura A.4: Histograma com correlação, linhas de densidade e de tendência referente aos dados da coordenada de desenovelamento para a proteína NTL9. Os descritores globais foram obtidos via PM7.

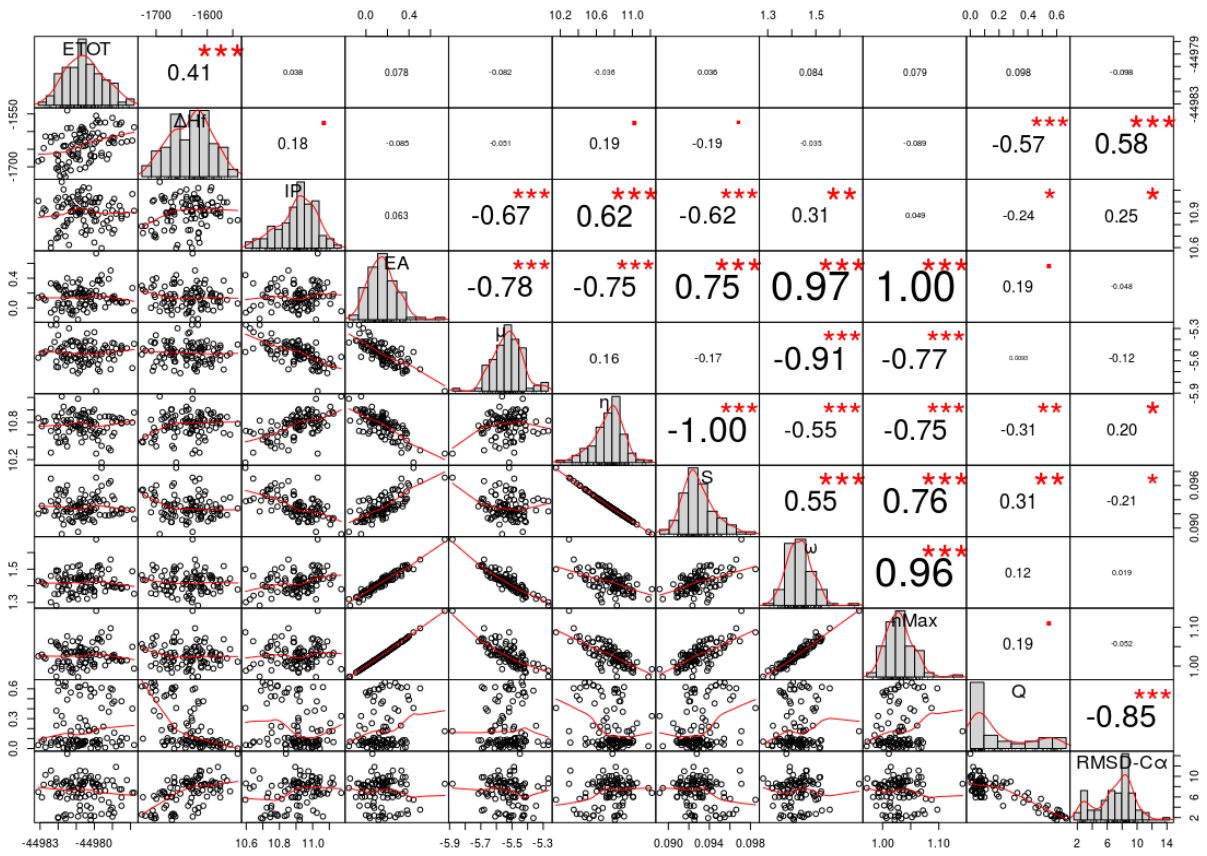


Figura A.5: Histograma com correlação, linhas de densidade e de tendência referente aos dados da coordenada de desenovelamento para a proteína BBA. Os descritores globais foram obtidos via PM7.

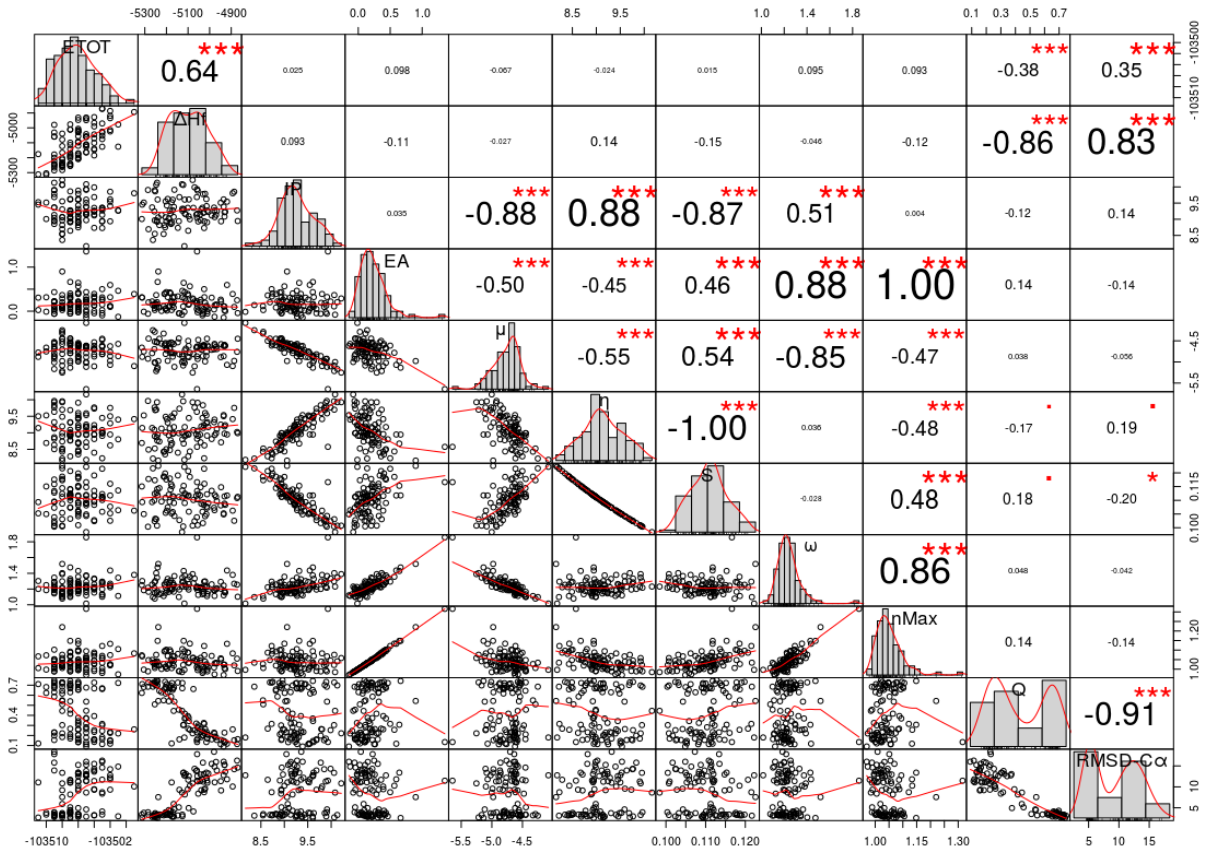


Figura A.6: Histograma com correlação, linhas de densidade e de tendência referente aos dados da coordenada de enovelamento para a proteína α3D. Os descritores globais foram obtidos via PM7.

A.2 QCMDs globais obtidos via método DFT-D3

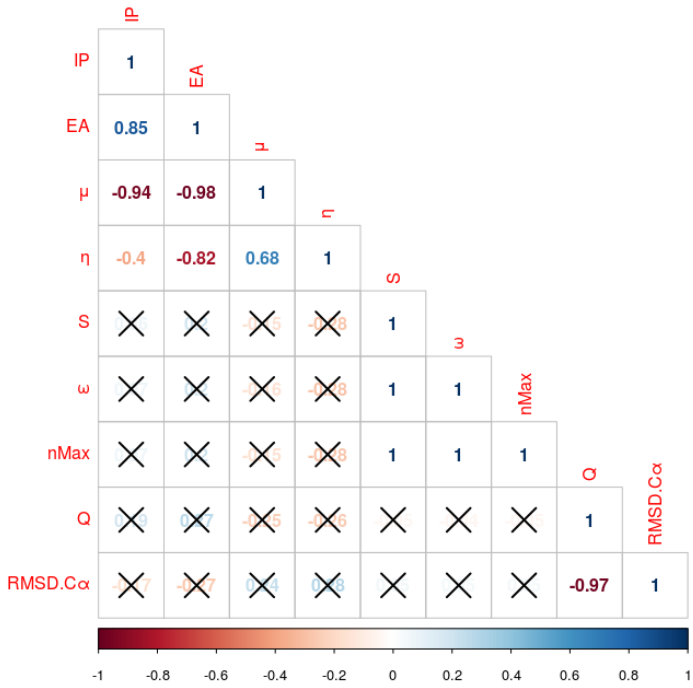


Figura A.7: Mapa de correlação entre os descritores globais de reatividade, E_{TOT} e ΔH_f (obtidos via DFT-D3) e os descritores estruturais Q e RMSD-Cα para a primeira coordenada TICA da NTL9.

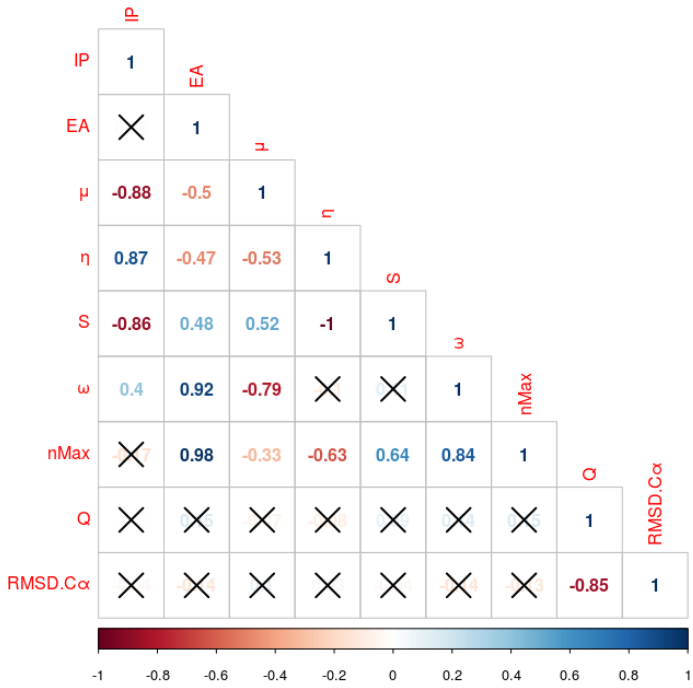


Figura A.8: Mapa de correlação entre os descritores globais de reatividade, E_{TOT} e ΔH_f (obtidos via DFT-D3) e os descritores estruturais Q e RMSD-Cα para a primeira coordenada TICA da BBA.

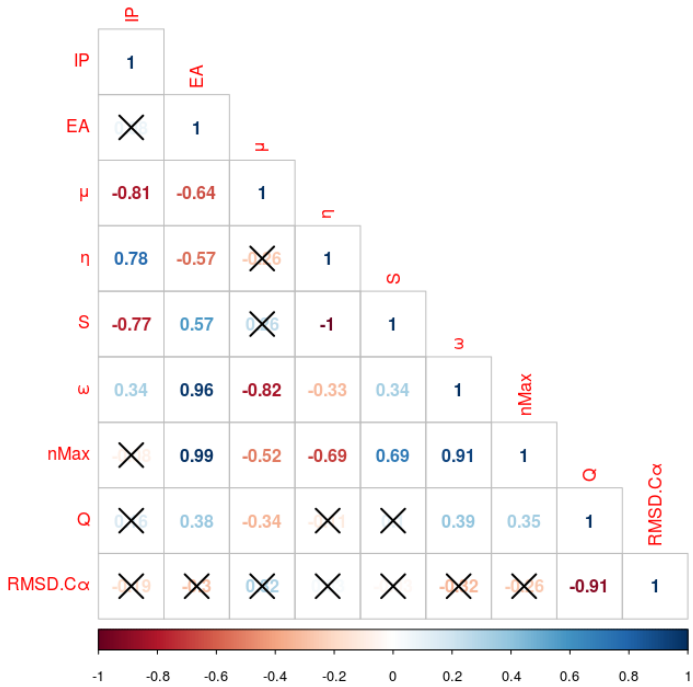


Figura A.9: Mapa de correlação entre os descritores globais de reatividade, E_{TOT} e ΔH_f (obtidos via DFT-D3) e os descritores estruturais Q e RMSD-Cα para a primeira coordenada TICA da α3D.

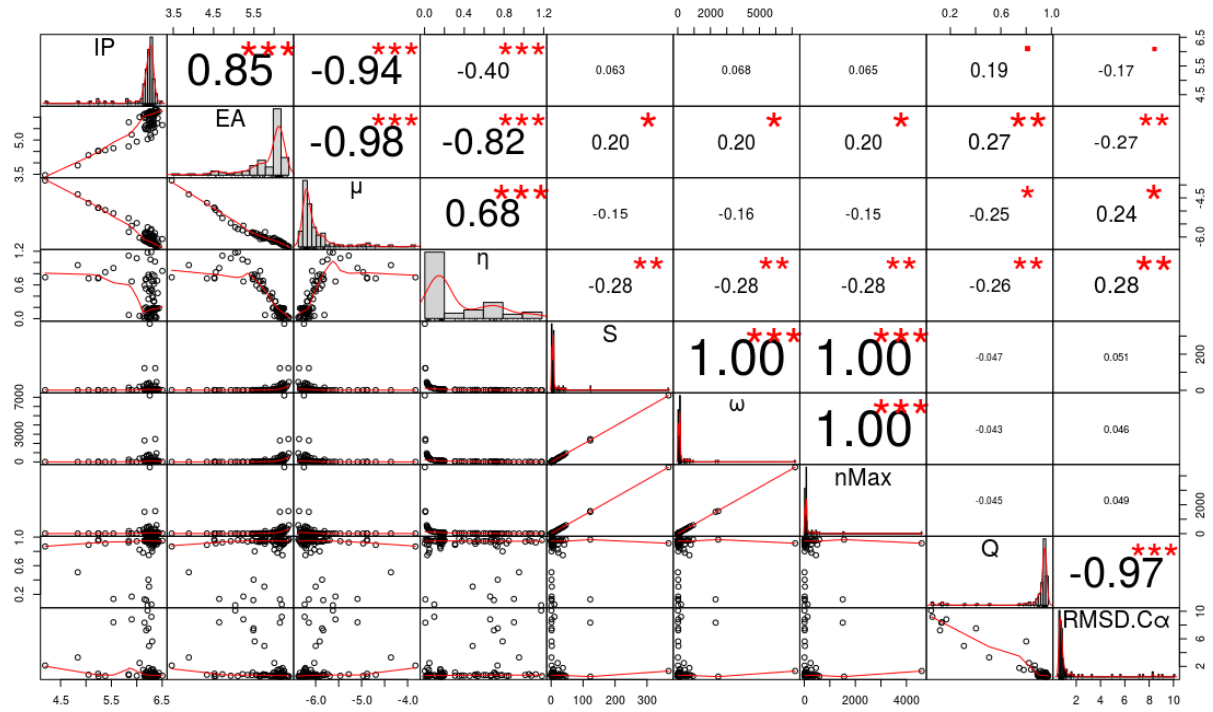


Figura A.10: Histograma com correlação, linhas de densidade e de tendência referente aos dados da coordenada de desenovelamento para a proteína NTL9. Os descritores globais foram obtidos via DFT-D3.

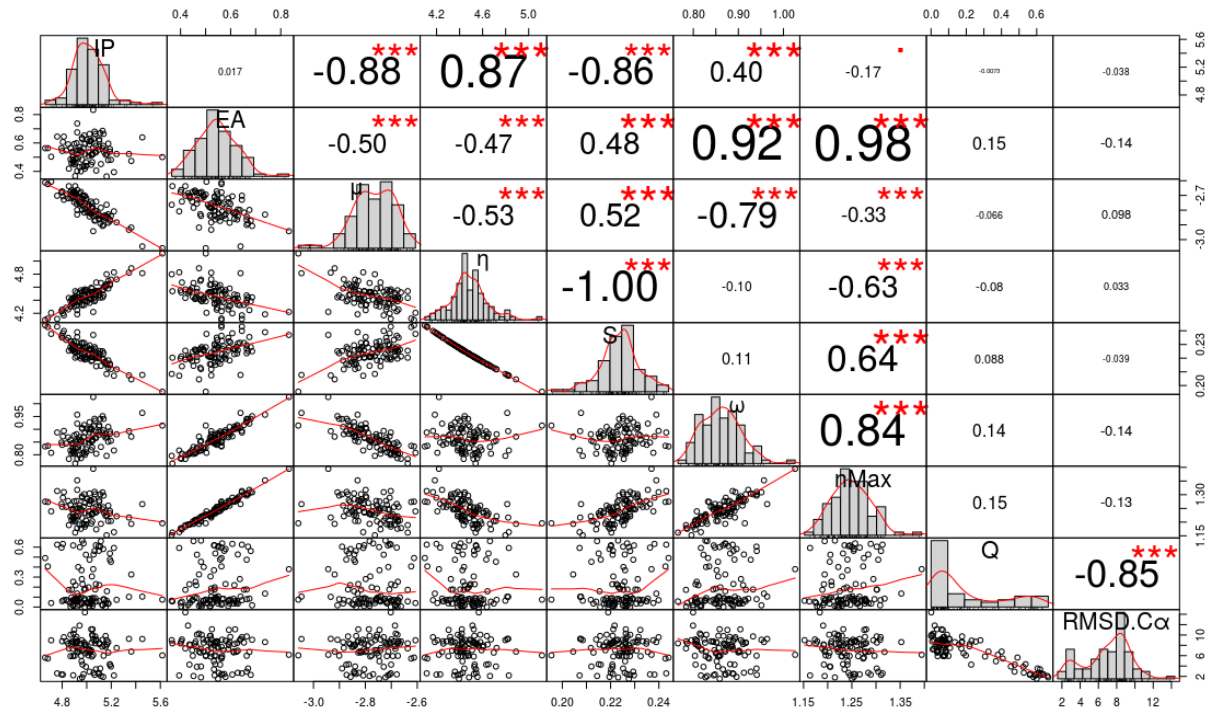


Figura A.11: Histograma com correlação, linhas de densidade e de tendência referente aos dados da coordenada de desenovelamento para a proteína BBA. Os descritores globais foram obtidos via DFT-D3.

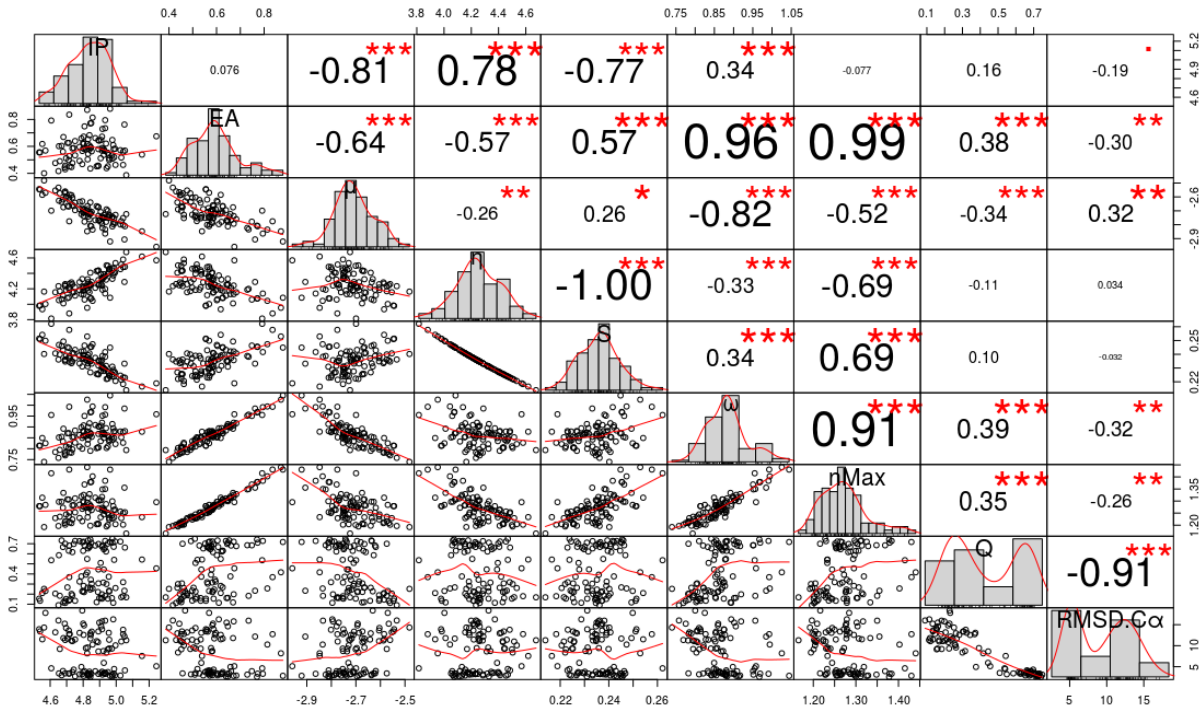


Figura A.12: Histograma com correlação, linhas de densidade e de tendência referente aos dados da coordenada de enovelamento para a proteína α3D. Os descritores globais foram obtidos via DFT-D3.

A.3 Descritores Estruturais

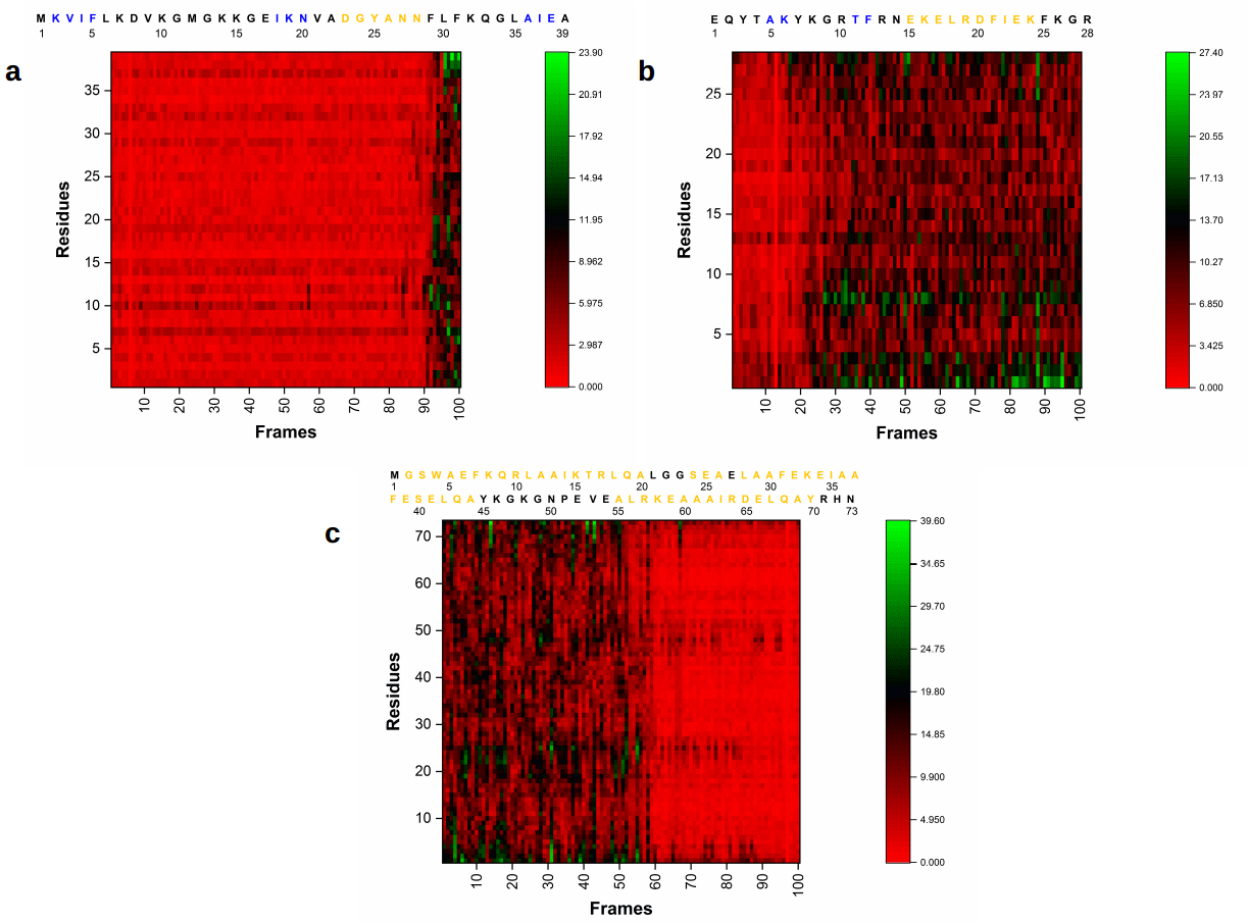


Figura A.13: RMSD por resíduo para as proteínas: (a) NTL9, (b) BBA e (c) α3D.

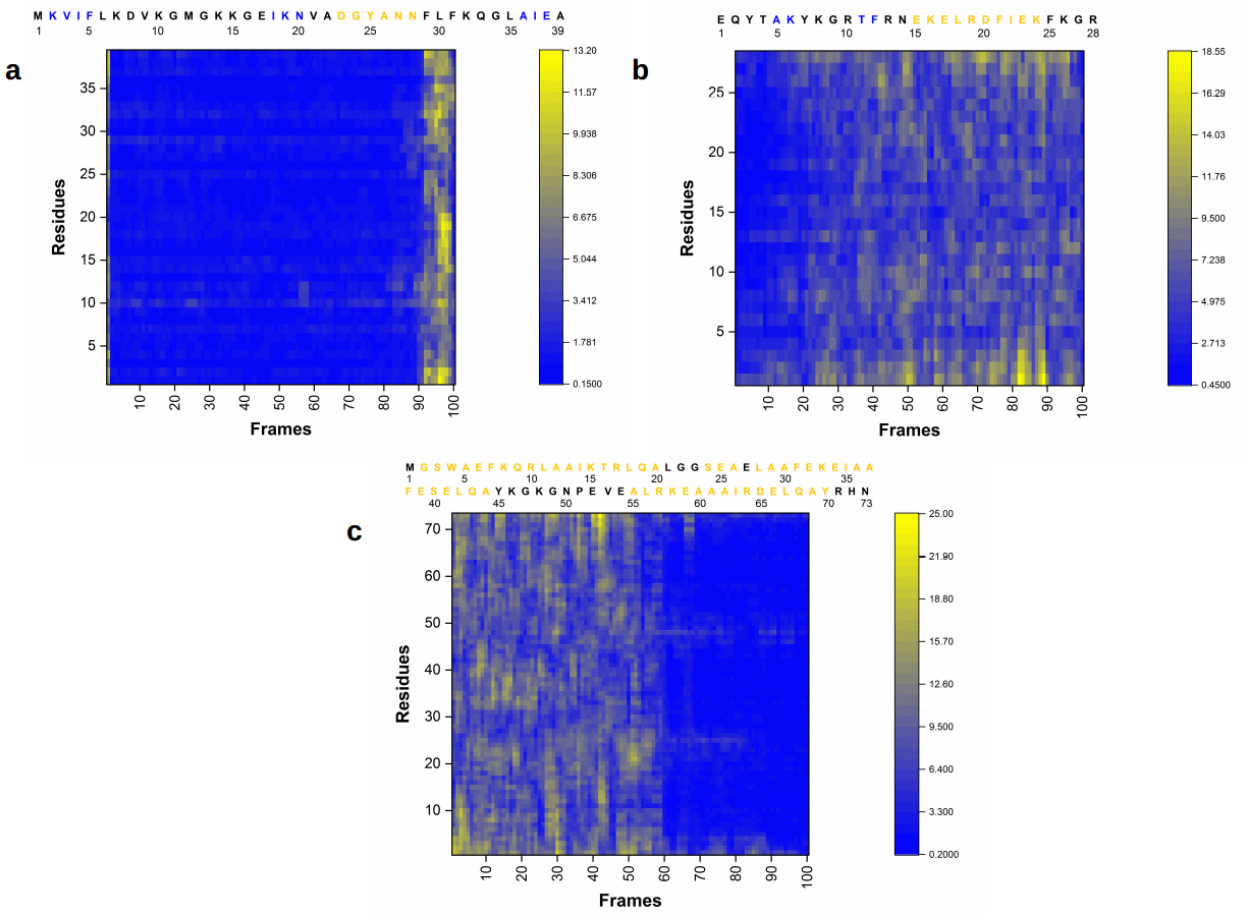


Figura A.14: RMSF por resíduo para as proteínas: (a) NTL9, (b) BBA e (c) α3D.

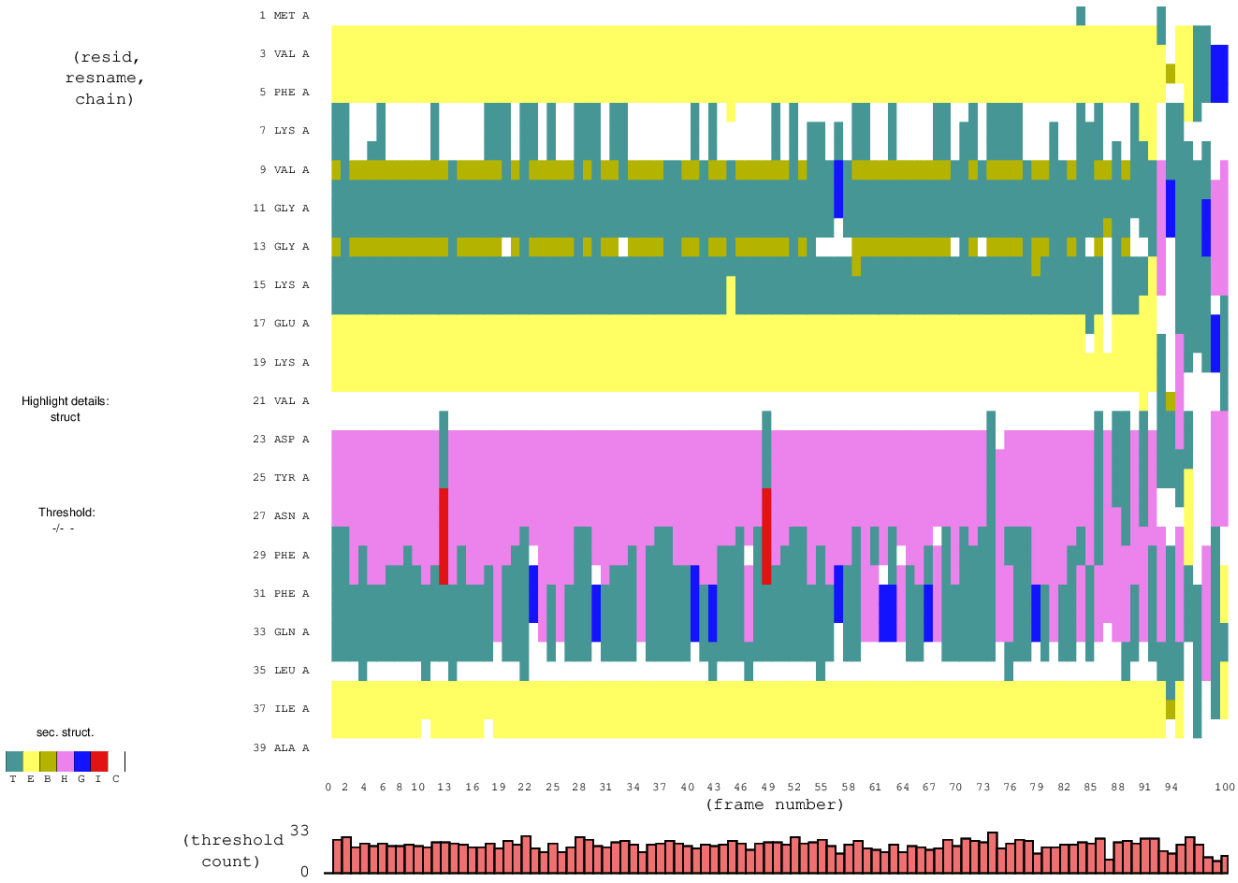


Figura A.15: Gráfico da estrutura secundária para a proteína NTL9.

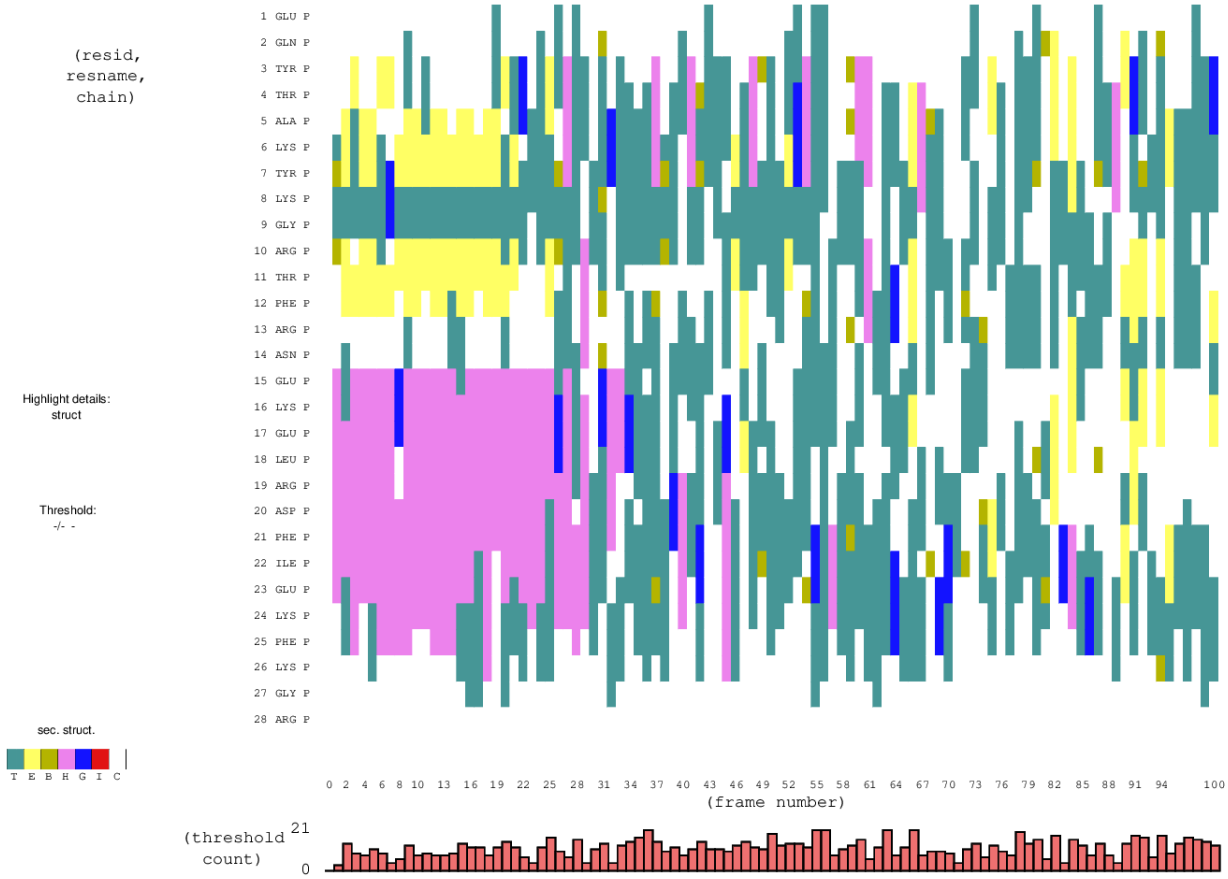


Figura A.16: Gráfico da estrutura secundária para a proteína BBA.

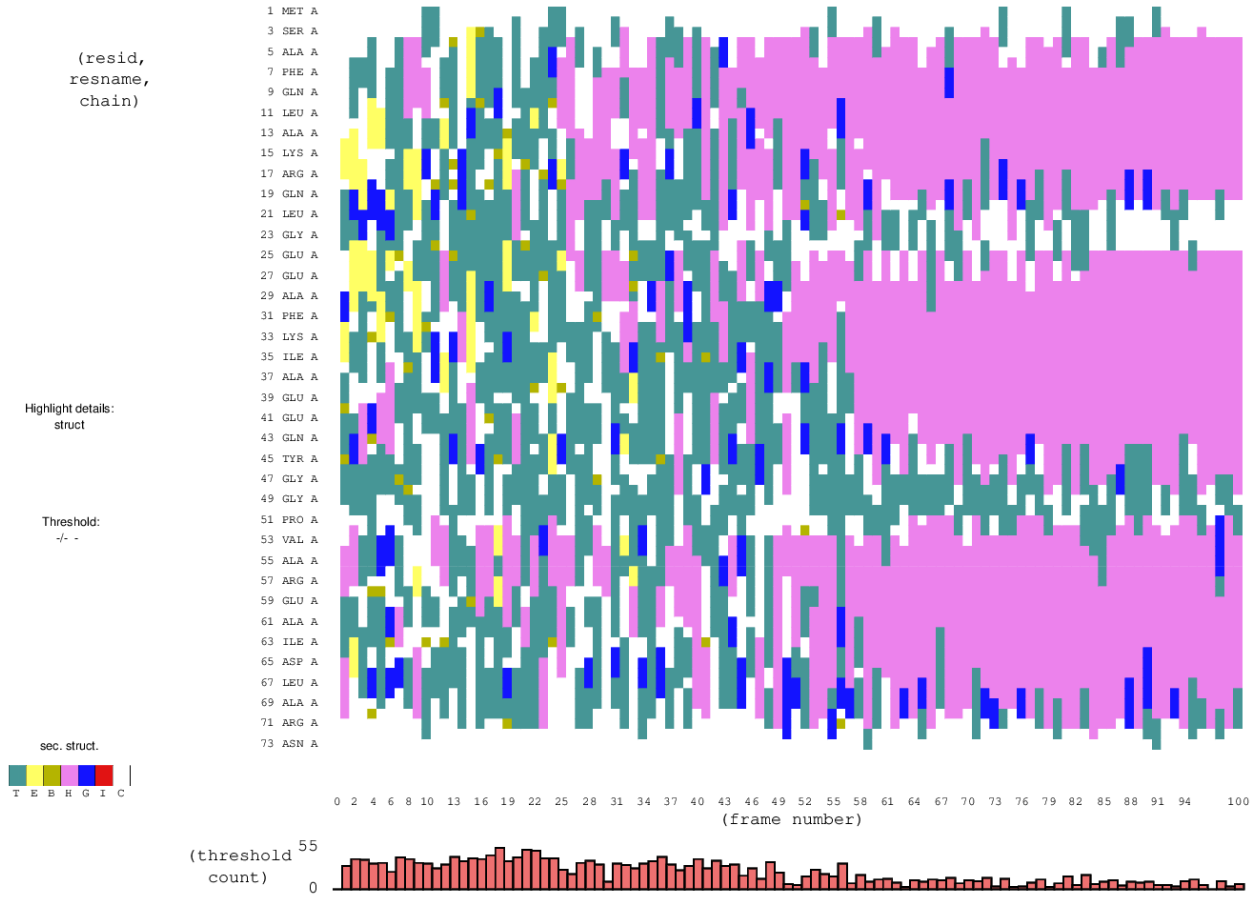


Figura A.17: Gráfico da estrutura secundária para a proteína α3D.

A.4 QCMDs Locais obtidos via método DFT-D3

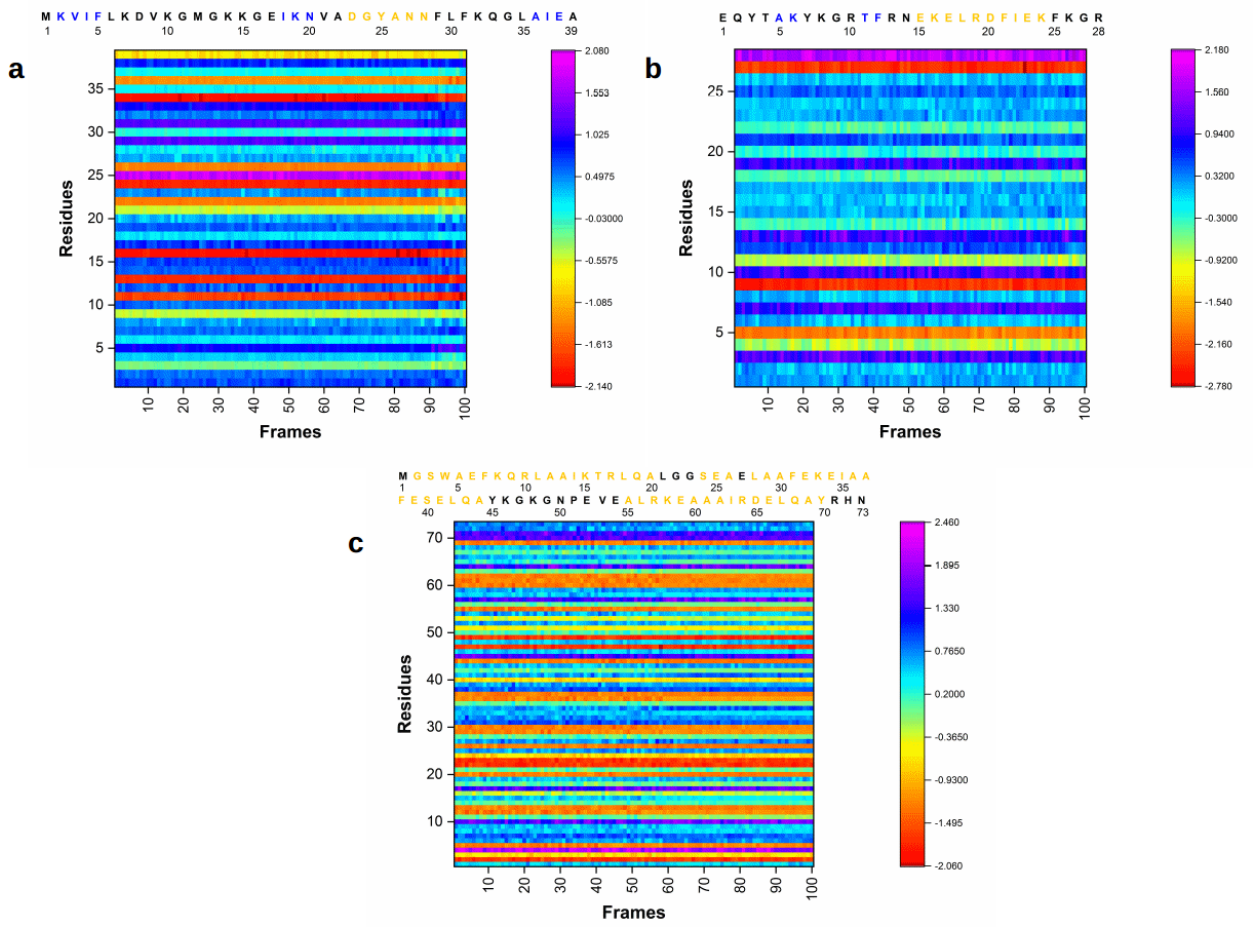


Figura A.18: Mapa de calor da densidade eletrônica para as proteínas: (a) NTL9, (b) BBA e (c) α3D obtidos via método DFT-D3.

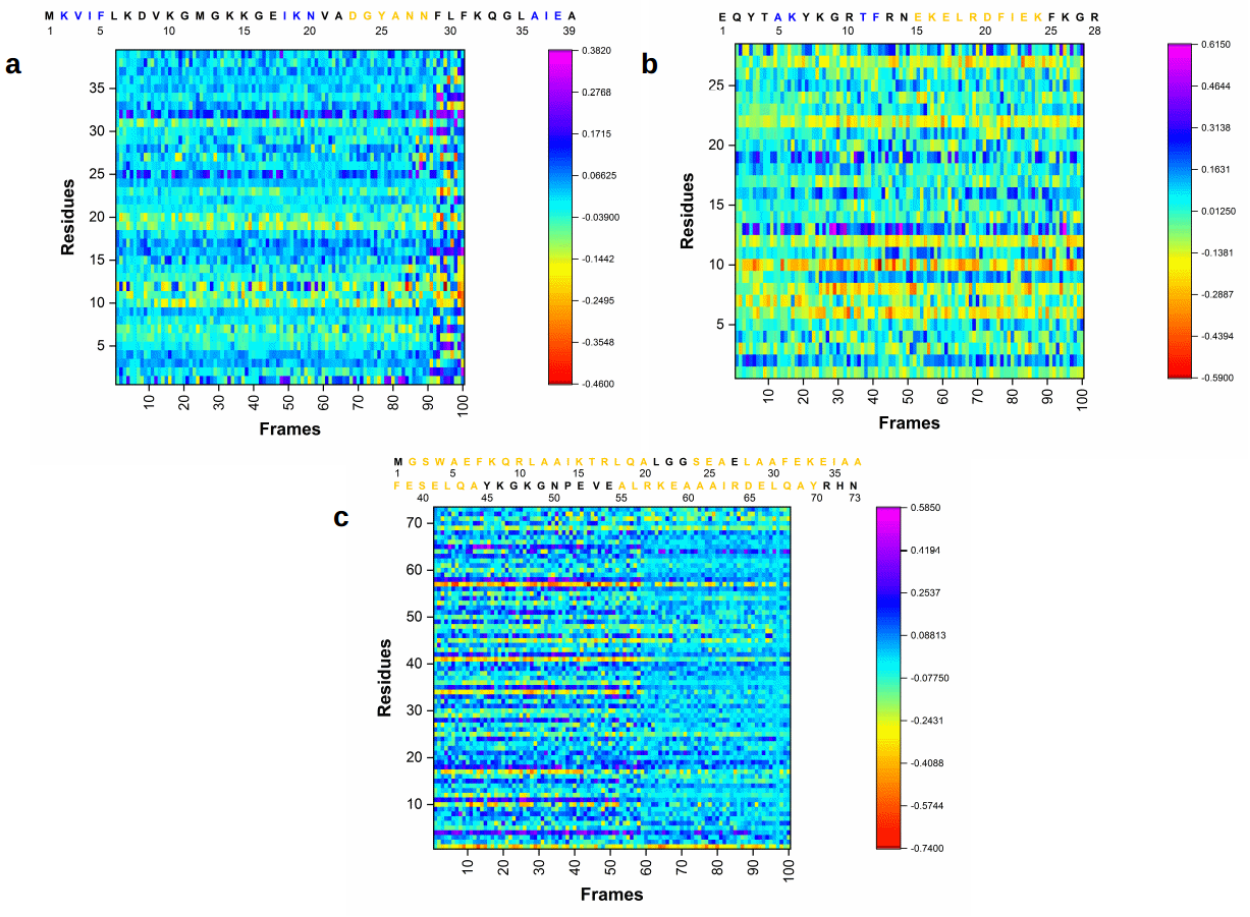


Figura A.19: Mapa de calor da variação da densidade eletrônica Δ_ρ) para as proteínas: (a) NTL9, (b) BBA e (c) α3D obtidos via método DFT-D3.

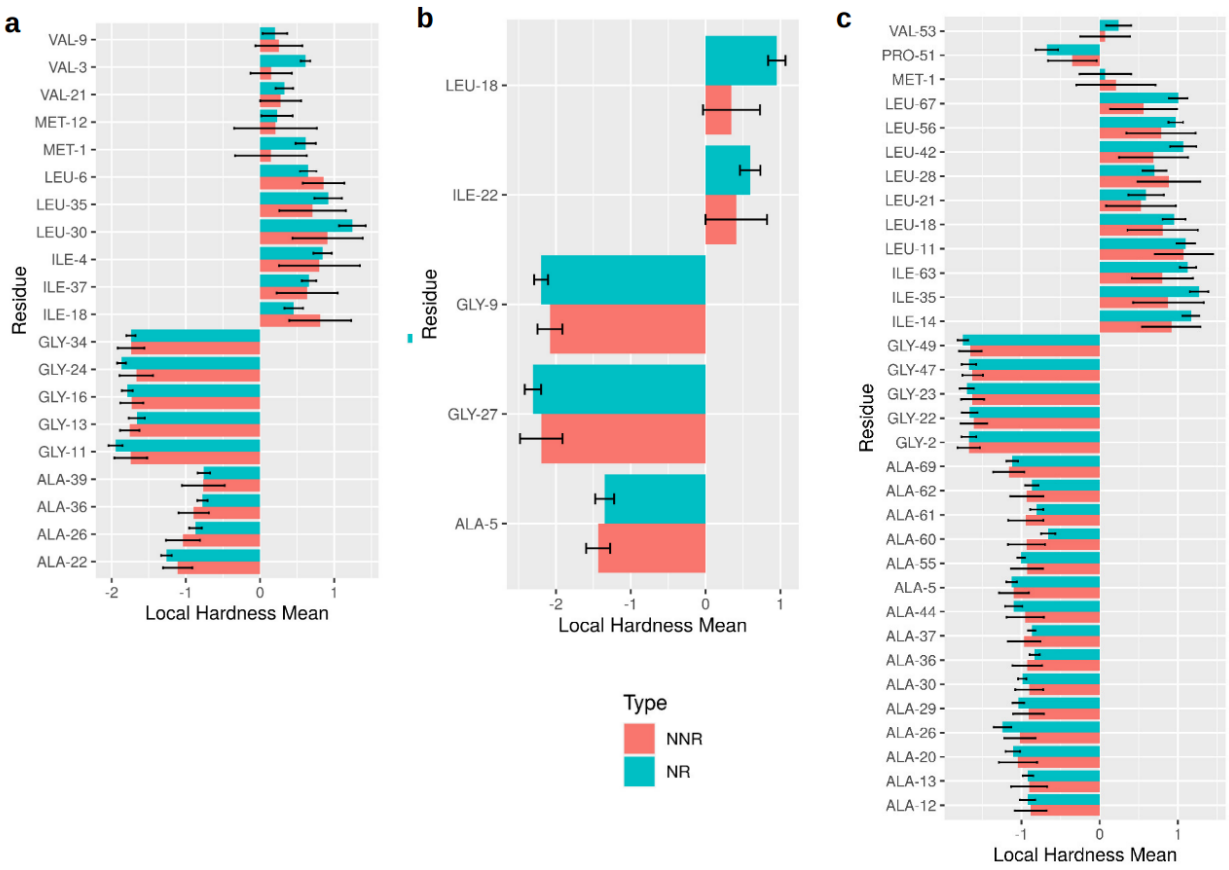


Figura A.20: Dureza média dos resíduos alifáticos não polares (grupo 1) para as proteínas: (a) NTL9, (b) BBA e (c) α3D obtidos via método DFT-D3.

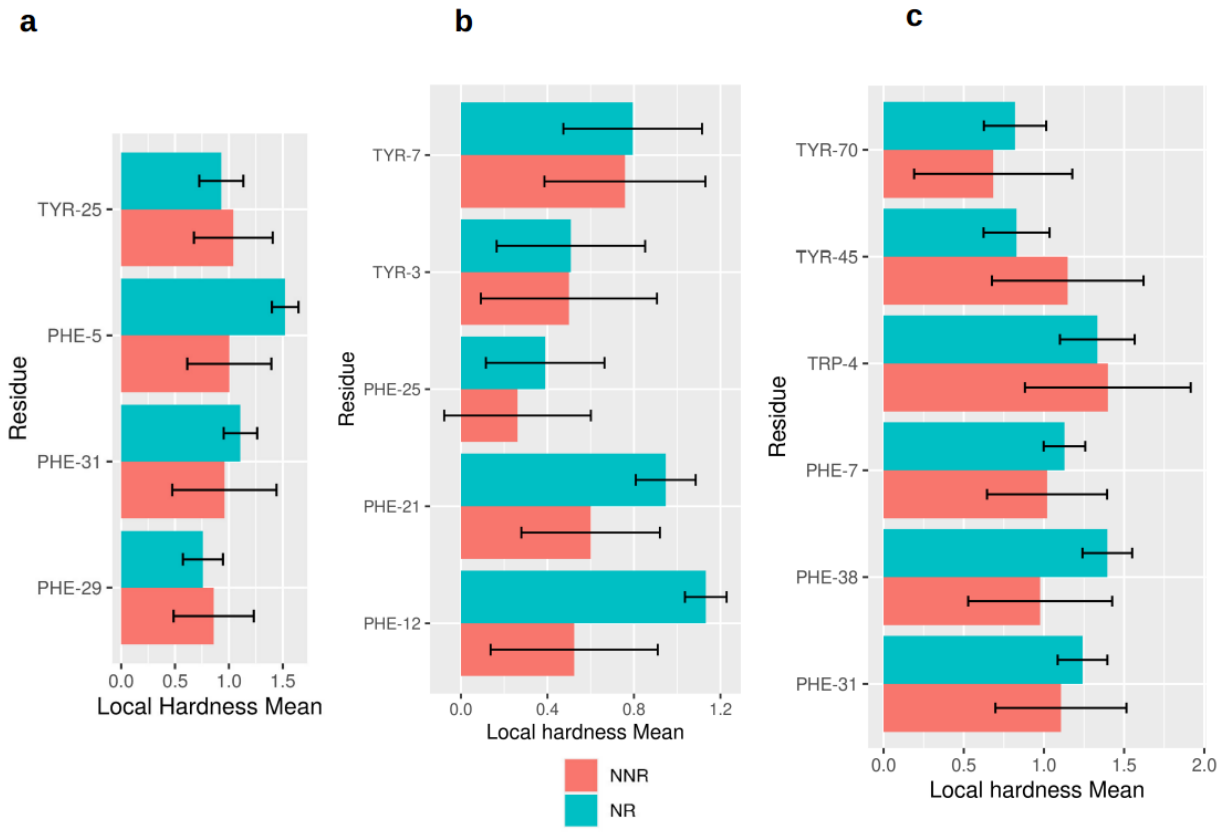


Figura A.21: Dureza média dos resíduos aromáticos (grupo 2) para as proteínas: (a) NTL9, (b) BBA e (c) α3D obtidos via método DFT-D3.

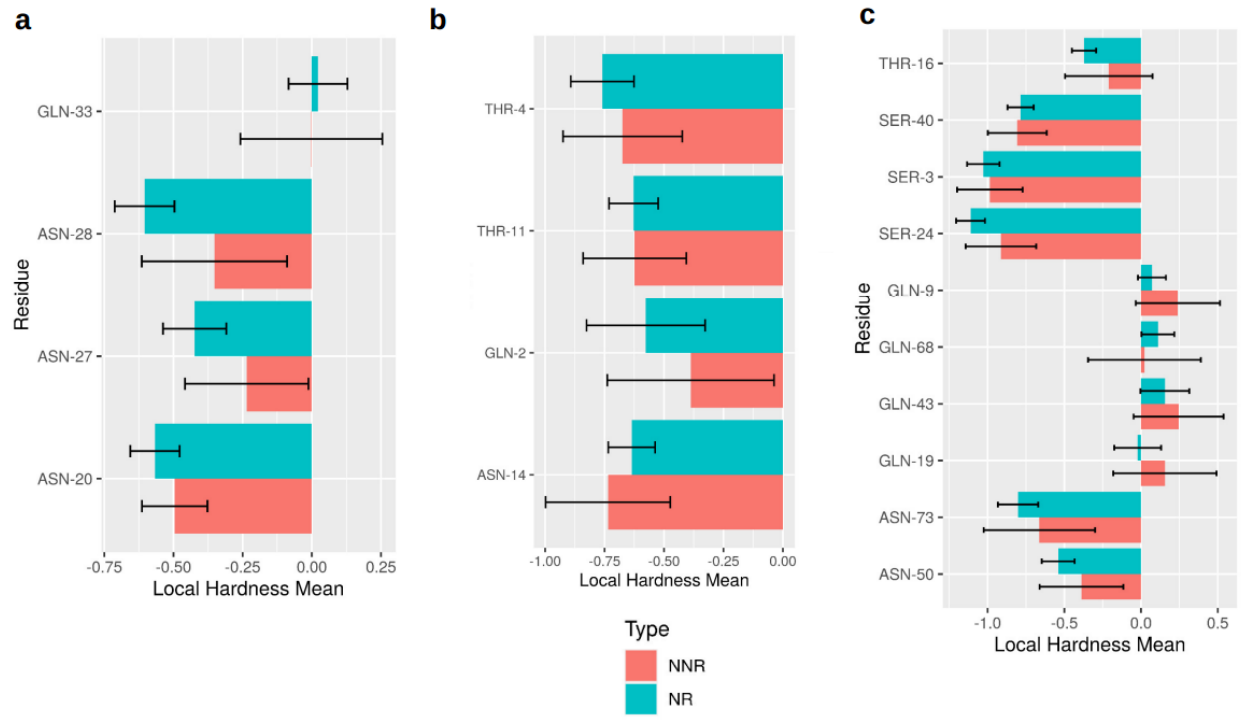


Figura A.22: Dureza média dos resíduos polares não carregados (grupo 3) para as proteínas: (a) NTL9, (b) BBA e (c) α3D obtidos via método DFT-D3.

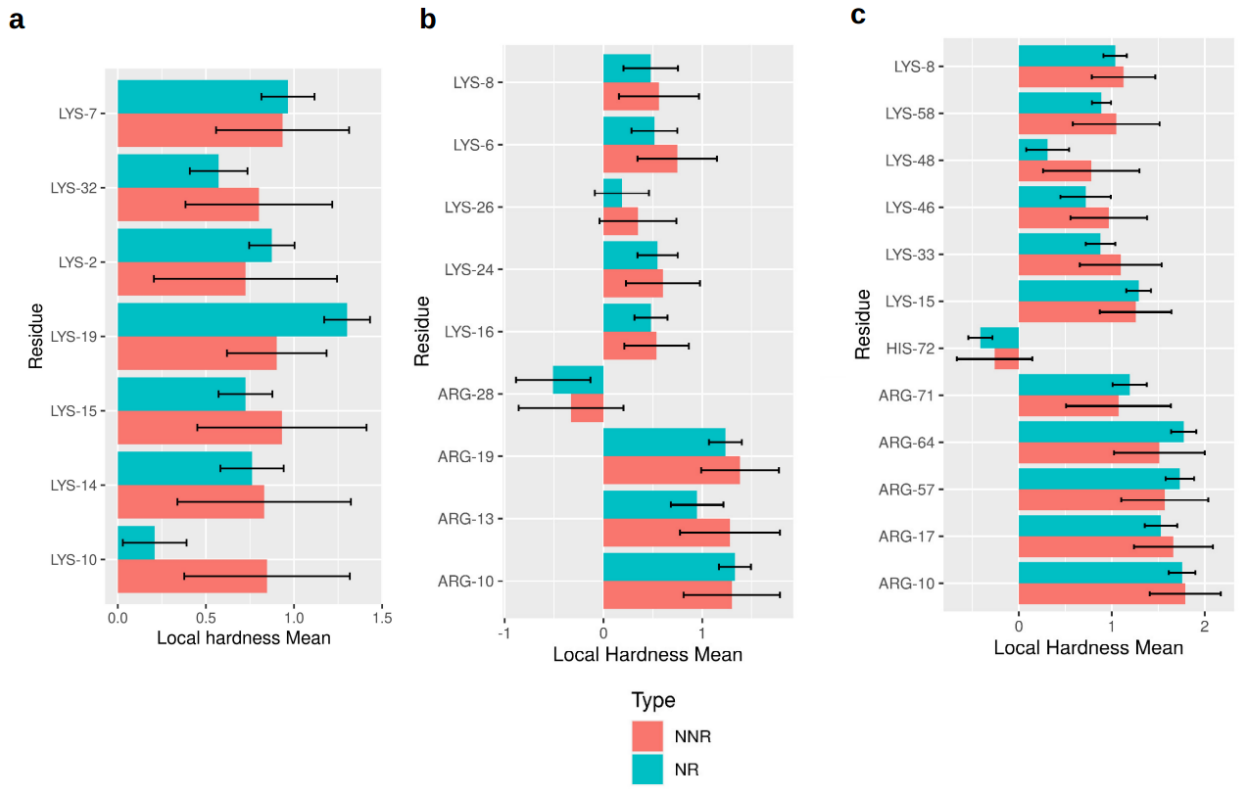


Figura A.23: Dureza média dos resíduos polares carregados positivamente (grupo 4) para as proteínas: (a) NTL9, (b) BBA e (c) α3D obtidos via método DFT-D3.

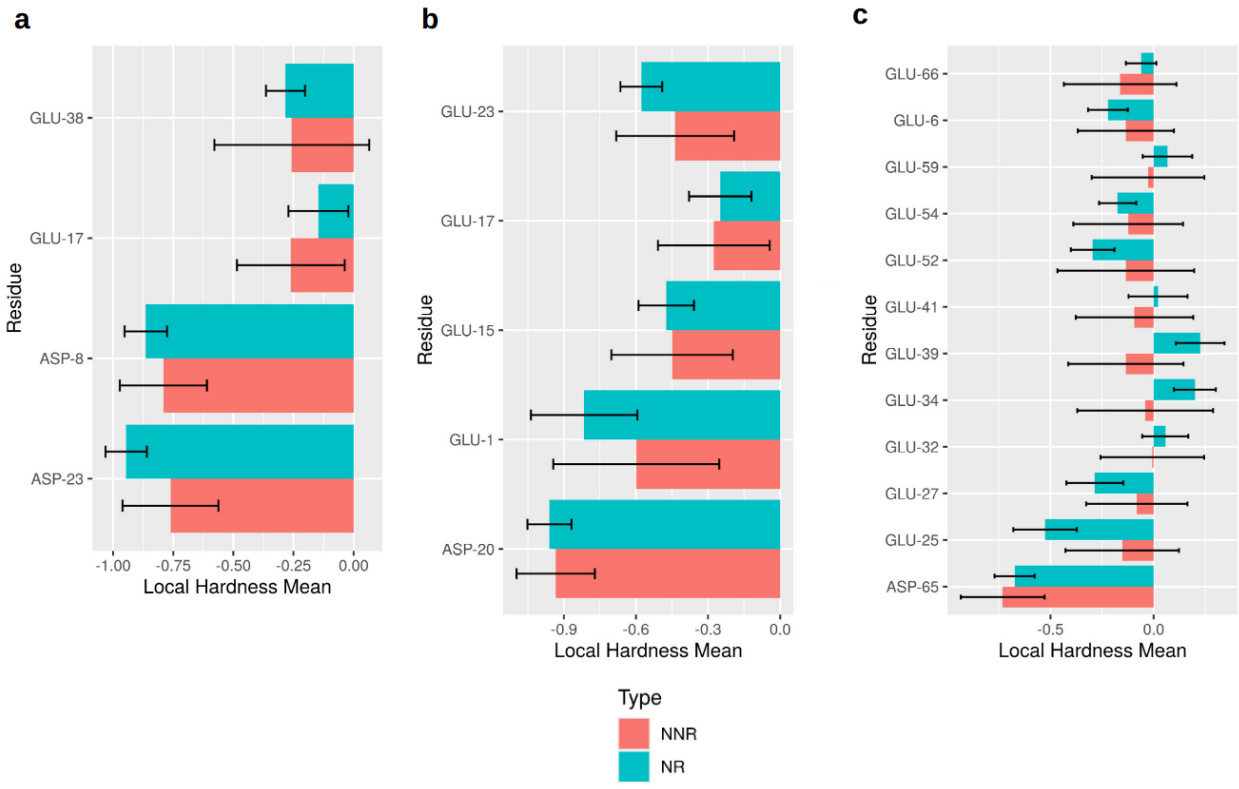


Figura A.24: Dureza média dos resíduos polares carregados negativamente (grupo 5) para as proteínas: (a) NTL9, (b) BBA e (c) α3D obtidos via método DFT-D3.

A.5 QCMDs Locais obtidos via método semiempírico PM7

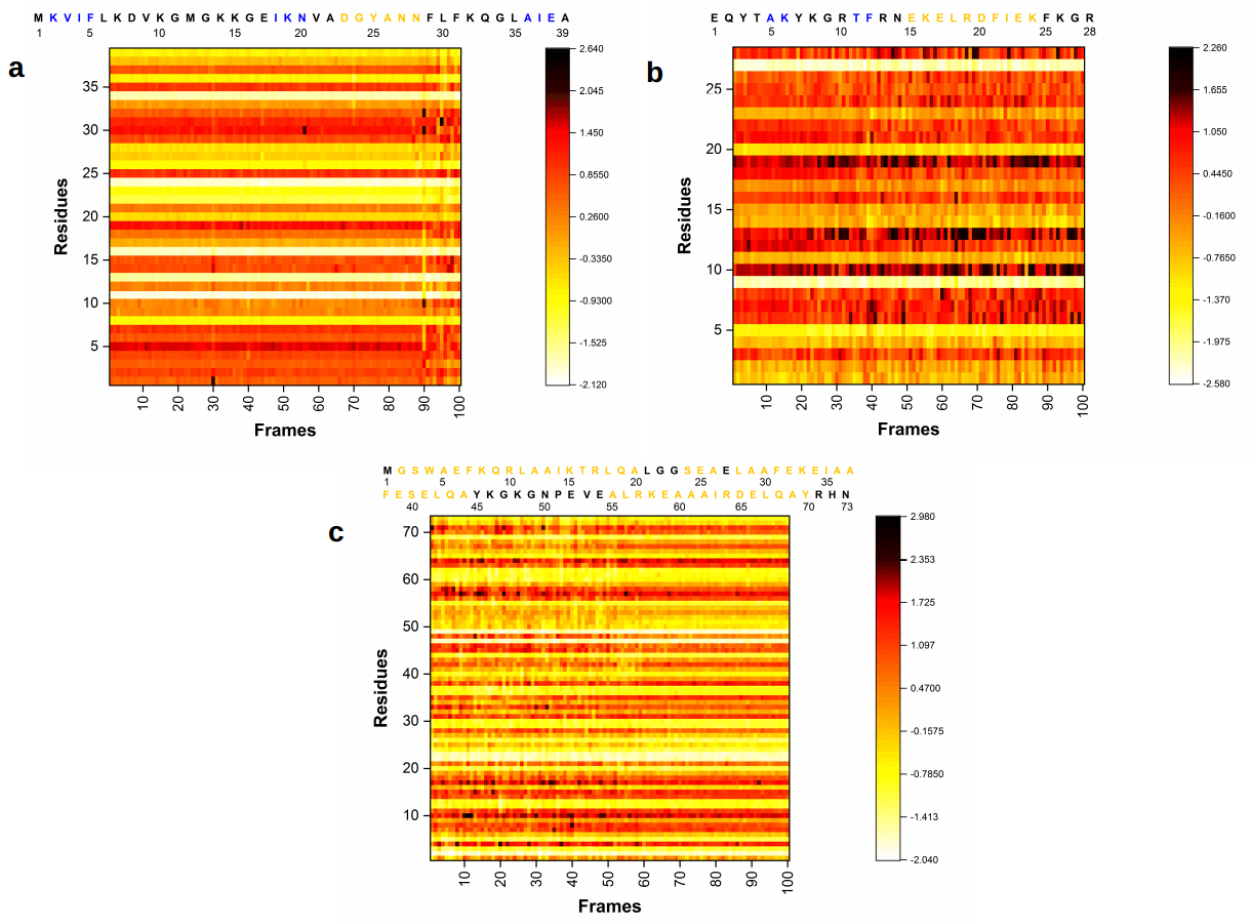


Figura A.25: Mapa de calor da dureza local para as proteínas: (a) NTL9, (b) BBA e (c) α3D obtidos via método semiempírico PM7.

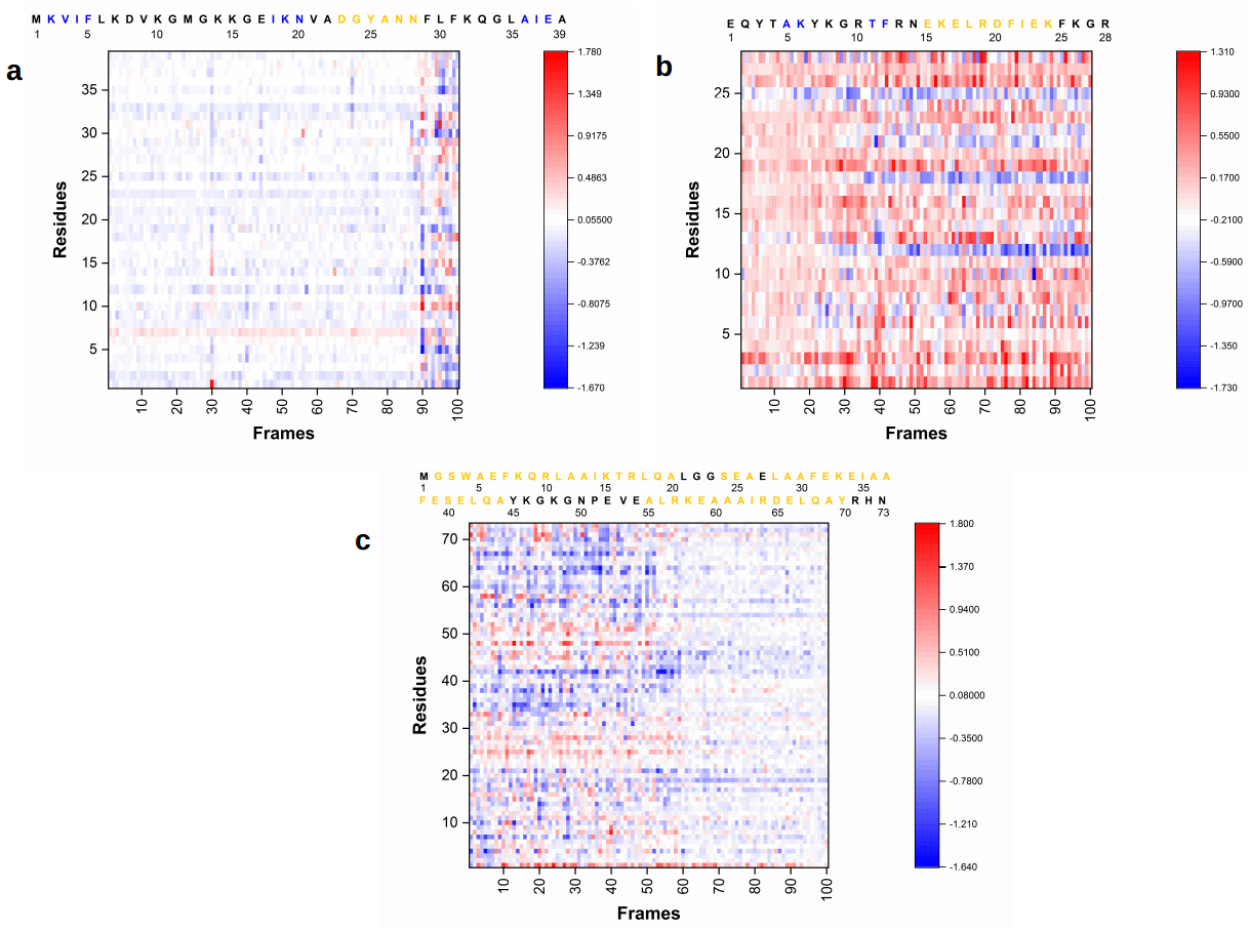


Figura A.26: Mapa de calor do Δ_η para as proteínas: (a) NTL9, (b) BBA e (c) α3D obtidos via método semiempírico PM7

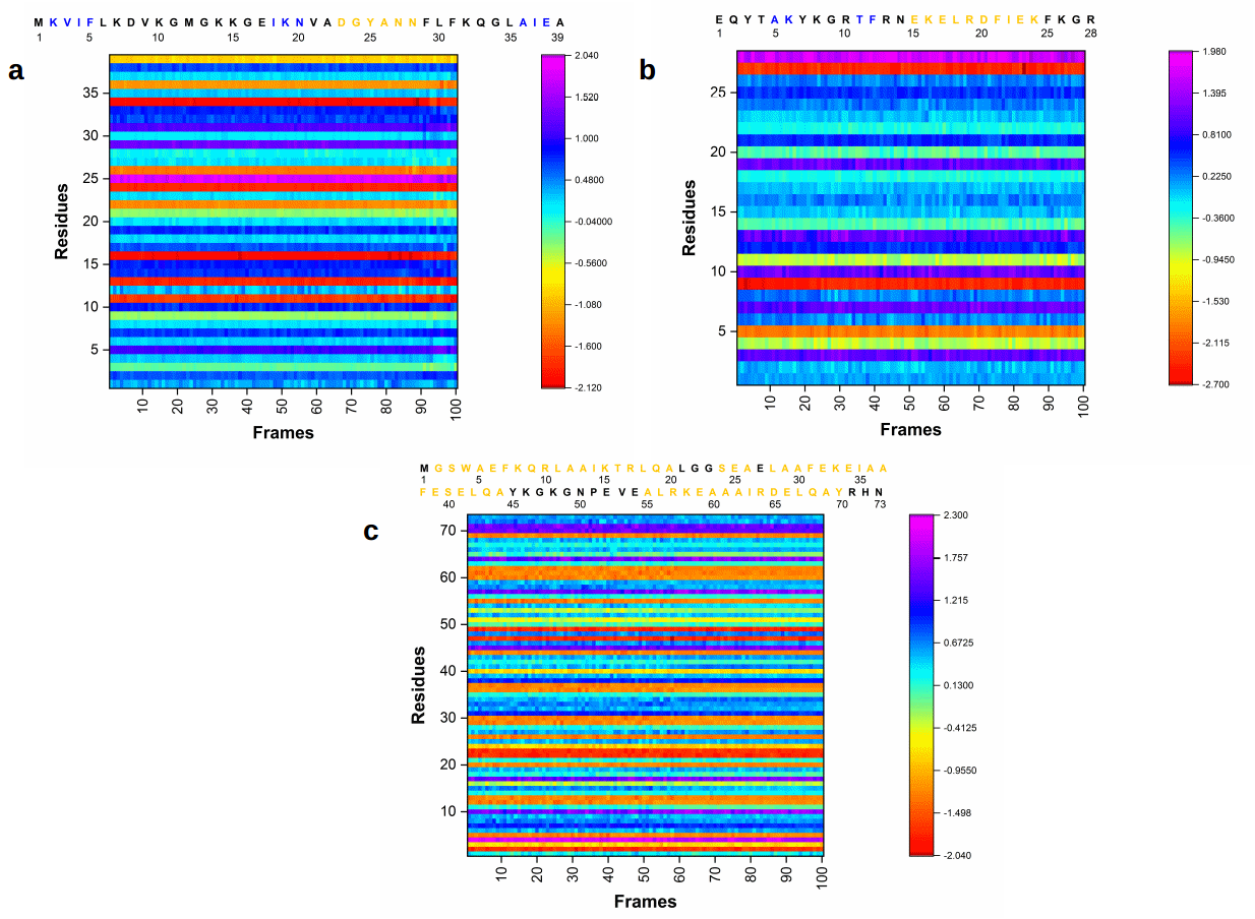


Figura A.27: Mapa de calor da densidade eletrônica para as proteínas: (a) NTL9, (b) BBA e (c) α3D obtidos via método semiempírico PM7.

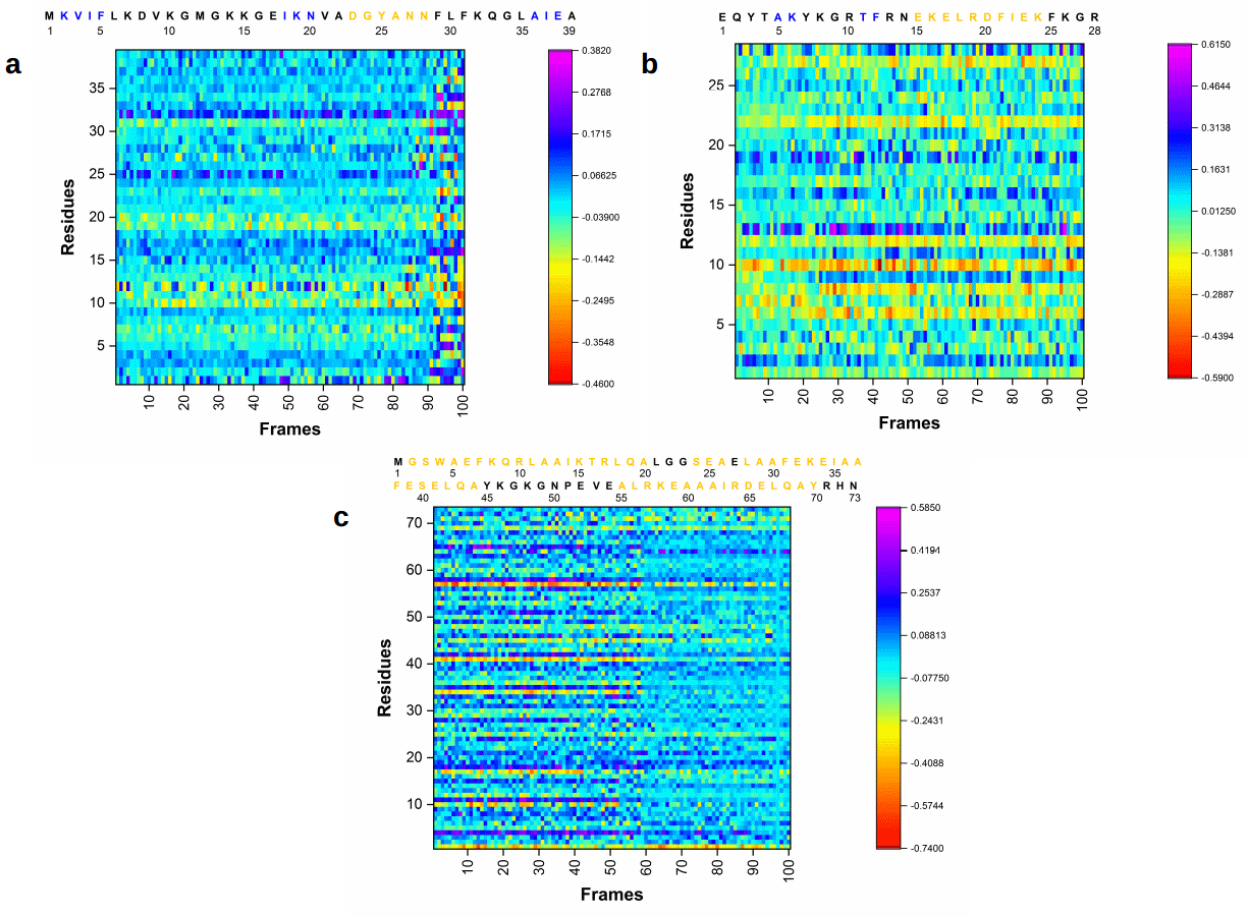


Figura A.28: Mapa de calor da variação da densidade eletrônica (Δ_ρ) para as proteínas: (a) NTL9, (b) BBA e (c) α3D obtidos via método semiempírico PM7.

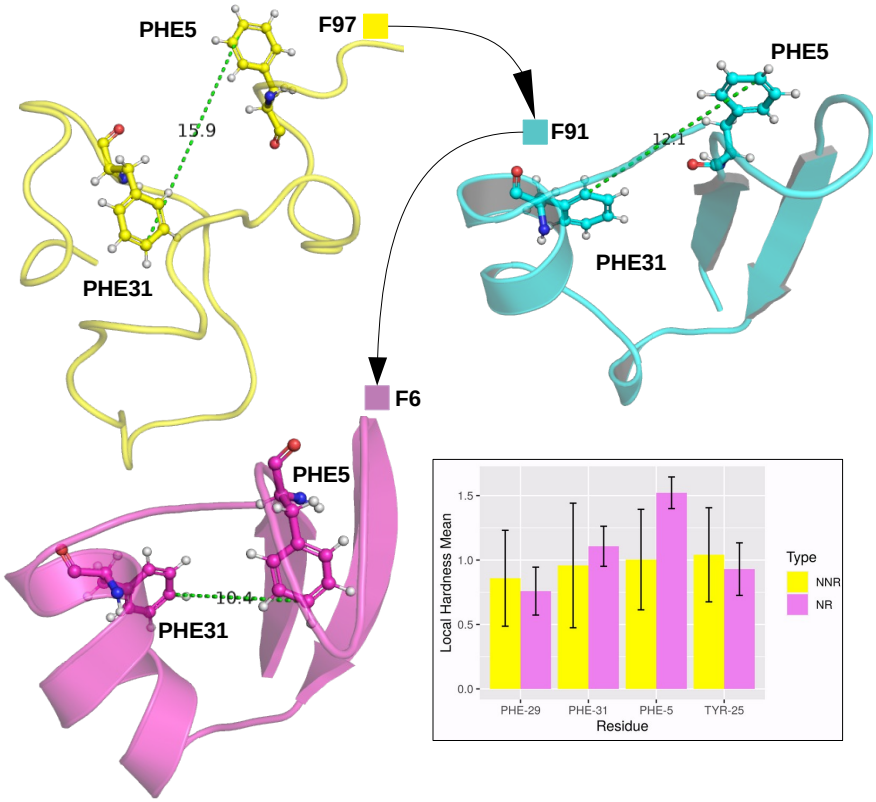


Figura A.29: Em (A): Conformações da proteína NTL9 para os *frames* 97, 91 e 6 com ênfase nos resíduos Phe-5 e Phe-31. Em (B): Dureza média local para os resíduos do grupo 2 da proteína NTL9 na região não nativa (NNR) e região semelhante à nativa (NR).

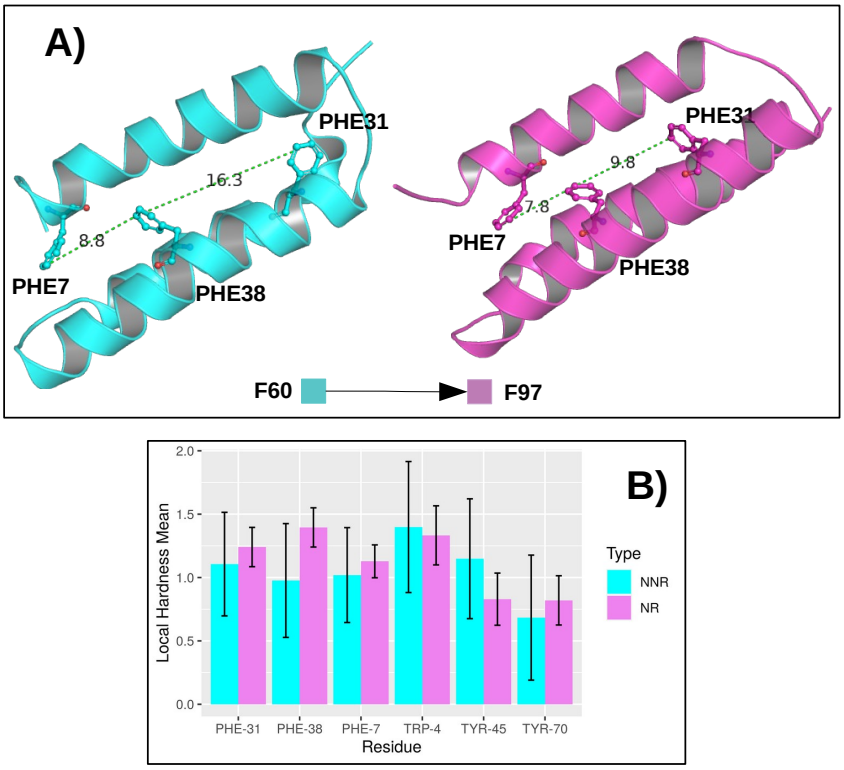


Figura A.30: Em (A): Conformações da proteína α3D para os *frames* 60 e 97 com ênfase nos resíduos Phe-7, Phe-31 e Phe-38. Em (B): Dureza média local para os resíduos do grupo 2 da proteína BBA na região não nativa (NNR) e região semelhante à nativa (NR).

ANEXOS

- **Anexo 1.** Resumo do trabalho apresentado no XIX Simpósio Brasileiro de Química Teórica (SBQT 2017).
- **Anexo 2.** Resumo do trabalho apresentado no III Simpósio Norte e Nordeste de Bioinformática (SNNB 2018).
- **Anexo 3.** Resumo do trabalho apresentado na IX Escola de Modelagem Molecular de Sistemas Biológicos (IX EMMSB 2018).
- **Anexo 4.** Resumo do trabalho apresentado no XX Simpósio Brasileiro de Química Teórica (SBQT 2019).
- **Anexo 5.** Resumo do trabalho apresentado no XXI Simpósio Brasileiro de Química Teórica (SBQT 2021).
- **Anexo 6.** Certificado de Comunicação oral Flash apresentada no SBQT 2019.
- **Anexo 7.** Certificado de premiação com menção honrosa recebida no SBQT 2019.
- **Anexo 8.** Certificado de premiação de melhor pôster no SBQT 2021.
- **Anexo 9.** Artigo publicado na revista *ACS Omega*.

XIX SBQT
Simpósio Brasileiro de Química Teórica 2017
12 a 17/Nov, 2017, Águas de Lindóia/SP, Brasil

Semiempirical ΔH_{bind} calculations for interactions between the RTA and RTB subunits and ricin inhibitors

Acassio Rocha Santos (PG)[1], Elton José Ferreira Chaves (PG)[2], Gabriel Urquiza de Carvalho (PQ)[1], Gerd Bruno da Rocha (PQ)[1]

[1]*Chemistry Departament, Univesidade Federal da Paraíba (UFPB), Cidade Universitária, s/n, Castelo Branco, (58051-900), João Pessoa, Paraíba, Brazil*
[2]*Biotechnology Departament, Universidade Federal da Paraíba (UFPB), Cidade Universitária, s/n, Castelo Branco, (58051-900), João Pessoa, Paraíba, Brazil*

Abstract: Industry, governments and the media have become increasingly more interested in the castor bean seeds (Ricinus communis L.). This stems from the unusual properties of its byproducts, such as castor oil and ricin. The ricinoleic acid comprises around 90% of all fatty acids extracted from the castor bean plant [1].

Besides the interest in its oil, the co-products generated during its production have garnered wide commercial attention. This is due to the production of around 1.5 million tons per year [1]. Hence, it is in the best interest of the industry to find an economically viable purpose for these co-products. A welcoming alternative regards the use of these co-products as animal food, though it is still not possible due to the presence of ricin, a ribosome-inactivating protein composed by two sub-units (known as RTA and RTB), in which the RTA serves as the catalytic sub-unit [2]. In addition to the problems related to co-products in the production of castor oil, terrorist groups yet utilize the ricin as a chemical weapon [3]. In this manner, the inhibition of the action mechanism within the ricin is of major economic, public and military interest, with the RTA being the target of inhibitors. Recently, fields of study within theoretical and computational chemistry have developed a capital role in the research of biological and/or biochemical systems, which provide with a proper orientation towards the conception of new drugs.

This study has carried out calculations for enthalpy of formation (ΔH_f) and ground state geometries of RTA and RTB subunit, both separately and joined to form complexes containing possible inhibitors, through semiempirical methods such as: RM1, PM6, PM6-DH+ and PM7. Crystallographic structures available at the PDB(ID) of the complexes containing inhibitors (0RB, PT1, EJ5, JP2) and of the RTA and RTB Ricin subunits (2AAI) were used in these studies. We also performed studies with two different inhibitor candidates synthesized by the group (Lv213 and Lv215). The structures were positioned in the active site of RTA through *Molecular Docking* [4]. The objective was to identify mechanisms that would favor ricin inhibition and to verify which semiempirical method would better describe the binding enthalpies (ΔH_{bind}) of the RTA-ligand and RTA-RTB complexes, at least from a qualitative viewpoint.

Semiempirical calculations of ΔH_f for ligands (0RB, PT1, EJ5, JP2, Lv213 and Lv215), RTA-0RB (4228 atoms), RTA-PT1 (4250 atoms), RTA-EJ5 (4234 atoms), RTA-JP2 (4218 atoms), RTA-Lv213 (4245 atoms) and RTA-Lv215 (4238 atoms) complexes were performed using the MOZYME [5] linear scaling technique implemented on the MOPAC program [6]. We carried out these calculations with the crystallographic structures, optimized by each of the mentioned methods. For the ricin structure (2AAI), since the RTB subunit presents glycosylations in its structure, ΔH_f calculations were conducted for the RTA-RTB systems without glycosylations (8212 atoms) and RTA-RTB with glycosylations (8444 atoms). Once the data for the ΔH_f of the ligands, RTA, RTB and the RTA-ligands and RTA-RTB complexes were calculated, one obtains the $\Delta H_{(bind)}$ values for the RTA-RTB systems (without glycosylations), RTA-RTB systems (with glycosylations) and the various RTA-ligands complexes. In all calculations, we considered the effects of the solvent through the implicit COSMO model for proteins solvated in water.

The ΔH_f results on the crystallographic geometry using the PM7 semiempirical method presented the following values for the RTA-ligand complexes: RTA-0RB2 [$\Delta H_{(bind)}$= -61.04 kcal/mol], RTA-1PT [$\Delta H_{(bind)}$ = -62.04 kcal/mol], RTA-EJ5 [$\Delta H_{(bind)}$ = -69.23 kcal/mol], RTA-JP2 [$\Delta H_{(bind)}$ = -60.39 kcal/mol], RTA-Lv213 [$\Delta H_{(bind)}$ = -32.31 kcal/mol] and RTA-Lv215 [$\Delta H_{(bind)}$ = -17.62 kcal/mol]. All values are consistent with those were experimentally observed for common enzyme-ligand complexes (at least in terms of the value range). Therefore, the data shows that, from an enthalpy viewpoint, the EJ5 ligand presented the lowest $\Delta H_{(bind)}$ when forming a complex with the RTA. On the other hand, the Lv215 ligand was the one presented the highest interaction value. Another interesting point is that the RTA-RTB complex without glycosylations presented an unfavorable interaction enthalpy ($\Delta H_{(bind)}$ = 80.91kj/mol) for its formation. However, the RTA-RTB complex with glycosylations presented a very pronounced favorable interaction ($\Delta H_{(bind)}$ = -6568.311kj/mol). Such results suggest the glycosylations play an important role in the formation of the ricin enzyme complex (RTA-RTB).

Key-words: Ricin, semi-empirical methods, enzyme-ligand interaction energy
Support: This work has been supported by CAPES/FAPESQ, CNPq and CAPES (biologia computacional, auxpe1375/2014)

References:

[1] Severino, L., Auld, D., *et al*. A*gronomy Journal*, **104** (4), 853–880 (2012).
[2] Endo, Y., Tsurugi, K. The RNA N-glycosidase activity of ricin A-chain. In *Nucleic acids symposium series*, **19**, 139–142 (1987).
[3] Franz, D., Jaax, N. Ricin toxin. *Medical aspects of chemical and biological warfare*, 631–642 (1997).
[4] Chaves, H. J. F. *Simulação Molecular de Inibidores da subunidade da ricina, RTA*. 2016. 80f. Dissertação (Mestrado em Biotecnologia) – Departamento de Biotecnologia, Universidade Federal da Paraíba, João Pessoa.2016.
[5] Stewart, J. P. J. *Int. Journal of Quantum Chemistry*, **58**, 133-146 (1996).
[6] Stewart, J. P. J. *Journal of Computer-Aided Molecular Design*, **4**, 1-105 (1990).

Proteínas e proteômica

CÁLCULOS DAS ENTALPIAS DE LIGAÇÃO ENTRE AS SUBUNIDADES RTA E RTB DA RICINA ATRAVÉS MÉTODOS SEMIEMPÍRICOS

Autores: Acassio Rocha Santos[1]; Gerd Bruno da Rocha[1]; Elton José Ferreira Chaves[2];

E-mail para correspondência: acassioroch@gmail.com

Instituições: [1]Departamento de Química - Universidade Federal da Paraíba (UFPB); [2]Departamento de Biotecnologia - Universidade Federal da Paraíba (UFPB);

Palavras-chave: ricina; entalpia de ligação; métodos semiempíricos

Apoio: CAPES/FAPESQ, CAPES (biologia computacional, auxpe 1375/2014), CNPq, CENAPAD -SP e NPAD-UFRN

A ricina é proteína inativadora de ribossomos composta por duas subunidades (RTA e RTB) ligadas por uma ponte de dissulfeto, em que a RTA é a unidade catalítica[1]. Por se tratar de uma proteína citotóxica, a ricina é utilizada como arma química, principalmente por grupos terroristas[2]. Esse trabalho consiste em realizar o cálculo das entalpias de formação ($?H_f$) e o cálculo da geometria das subunidades RTA e RTB separtadamente e unidas formando o complexo RTA-RTB através dos métodos semiempíricos de química quântica PM6, PM6-DH+, PM7 e RM1. A estrutura cristalográfica da ricina (2AAI), disponível no PDB foi utilizada nesse estudo. Como a subunidade RTB da ricina apresenta glicosilação em sua estrutura, os cálculos de $?H_f$ foram conduzidos para os sistemas RTA-RTB sem glicosilações (8212 átomos) e RTA-RTB com glicosilações (8444 átomos). Outro objetivo foi verificar se as glicosilações da RTB influenciam na energética do complexo RTA-RTB e qual método semiempírico descreve melhor as entalpias de ligação ($?H_{bind}$) dos complexos RTA-RTB, pelo menos do ponto de vista qualitativo. Uma vez calculados os dados para o $?H_f$ dos complexos RTA-RTB, obtém-se os valores de $?H_{bind}$ para os sistemas RTA-RTB (sem glicosilações) e RTA-RTB (com glicosilações). Foram realizados tando cálculos single-point quanto otimização de todos os átomos. Para todos os cálculos, usamos o algoritmo de escalonamento linear MOZYME[3] e modelo implícito de solvente COSMO para proteínas solvatadas em meio aquoso, disponíveis no pacote MOPAC[4]. Cálculos de energia na geometria experimental com o método PM7 para a RTA-RTB sem glicosilações apresentou entalpia de formação desfavorável ($?H_{bind}$ = 80,91 $kcalmol^{-1}$), porém para a RTA-RTB com glicosilações, a entalpia de formação foi altamente favorável ($?H_{bind}$ = -6.568,31 $kcalmol^{-1}$). Essa mesma tendência foi observado para os métodos PM6, PM6-DH+ e RM1. Otimizações de todos os átomos com o método PM6 apresentaram ($?H_{bind}$ = 223,44 $kcalmol^{-1}$) e ($?H_{bind}$ = -15,51 $kcalmol^{-1}$) para a RTA-RTB sem glicosilações e com glicosilações respectivamente. Isso sugere que as glicosilações são muito importantes para a energética da ricina.

Cálculos das entalpias de ligação entre a subunidade RTA da ricina e candidatos à inibidores através de métodos semiempíricos

Acassio Rocha Santos[1], Elton J. F. Chaves[2], Amanara S. de Freitas[1] e Gerd B. Rocha[1]
[1]Departamento de Química - UFPB, [2]Departamento de Biotecnologia - UFPB

A semente da mamoneira (Ricinus communis L.) tem apresentado elevado interesse da indústria, de governos e da mídia devido às propriedades incomuns dos seus derivados como o óleo e a ricina.[1] A ricina é uma glicoproteína citotóxica constituída por duas subunidades: RTA e RTB que estão ligadas por uma ponte de dissulfeto. A RTA com 267 resíduos é a unidade catalítica e a RTB é uma leucina com 262 resíduos, sendo responsável pela internalização do complexo RTA-RTB no citosol das células.[2] Além do interesse no óleo, as autoridades mundiais tem voltado o seu olhar para a ricina devido à sua utilização como arma química por grupos terroristas.[3] Desse modo tem-se despertado bastante interesse na obtenção de mecanismos para inibição da ricina, sendo a RTA como principal alvo dos candidatos a inibidores. O objetivo desse trabalho consistiu em realizar o cálculo das entalpias de ligação ($\Delta H_{(bind)}$) para os complexos RTA-ligantes e avaliar a performance dos métodos semiempíricos PM6, PM6-DH+, PM6-D3H4, PM7 e RM1 através de correlação desses métodos com resultados experimentais de IC_{50}. Primeiramente, obtivemos, do banco de dados do PDB, estruturas cristalográficas de seis complexos inibidores de RTA (PDB-IDs: 4ESI, 4MX1, 4HUP, 3PX8, 3PX9 e 4HUO). Depois as estruturas dos complexos de proteínas foram relaxadas usando simulações de dinâmica molecular.[4] As geometrias finais foram utilizadas para o cálculo de entalpias de ligação por meio de métodos quânticos semiempíricos. Uma vez calculados os dados para o ΔH_f para os ligantes, RTA e complexos RTA-ligantes, obtém-se os valores de $\Delta H_{(bind)}$ para os diversos sistemas. Foram realizados tanto cálculos *single-point* quanto otimização de geometria. Para todos os cálculos de química quântica, usamos o algoritmo de escalonamento linear MOZYME [5] disponíveis no pacote MOPAC [6]. Além disso, os efeitos do solvente foram considerados através do modelo implícito COSMO para proteínas solvatadas em meio aquoso. Os métodos PM6-DH+, PM7 e PM6-D3H4 foram os que apresentaram os melhores resultados tanto quando se usou as geometrias otimizadas quanto a partir de cálculos *single-point*, obtendo-se boa correlação com dados de IC_{50}. Os resultados de correlação com IC_{50} foram os seguintes: PM6-DH+(R^2 = 0,925), PM7 (R^2 = 0,950) e PM6-D3H4 (R^2 = 0,969). Isso demonstra que a dinâmica molecular estava bem equilibrada e esses métodos semiempíricos apresentaram bons resultados. Para a otimização de todos os átomos somente os métodos PM7 (R^2 = 0,447) e PM6-D3H4 (R^2 =0,623) apresentaram alguma correlação com os dados de IC_{50}, porém com resultados inferiores em relação aos cálculos *single-point*. Apesar disso, tanto o PM7 quanto o PM6-D3H4 acertaram no ranqueamento do melhor ligante para a RTA. Isso demonstra que os métodos semiempíricos apresentam um grande potencial para aplicação no desenvolvimento de novos fármacos e para o tratamento quântico completo de grandes sistemas biomoleculares.

Palavras-chave: RTA, métodos semiempíricos, entalpia de ligação
Agradecimentos: CAPES/FAPESQ, CAPES (biologia computacional auxpe 1375/2014), CNPQ, NPAD-UFRN e CENAPAD-SP

Referências

[1] Severino, L., Auld, D., *et al. Agronomy Journal*, 104 (4), 853–880 (2012).
[2] Endo,Y., Tsurugi,K. *Nucleic Acids Symp. Ser.*, 19, 139–142 (1988).
[3] Franz, D., Jaax, N. Ricin toxin. Medical aspects of chemical and biological warfare, 631-642 (1997).
[4] Chaves, E., *et al. J. Chem. Inf. Model*, 58 (6), 1205-1213 (2018).
[5] Stewart, J. P. J. *Int. Journal of Quantum Chemistry*, 58, 133-146 (1996).
[6] Stewart, J. P. J. *Journal of Computer-Aided Molecular Design*, 4, 1-105 (1990).

XX SIMPÓSIO BRASILEIRO DE QUÍMICA TEÓRICA

10 a 14 de novembro de 2019
Centro de Convenções de João Pessoa
Paraíba - Brasil

Análise do enovelamento da proteína α3D através de dinâmica molecular e modelo de estado de Markov

Acassio R. Santos (PG),[1]* Gabriel A. U. Carvalho (PQ),[2]Gerd B. Rocha (PQ).[1]

acassioroch@gmail.com

[1]Departamento de Química, Universidade Federal da Paraíba, João Pessoa-PB; [2]Departamento de Química Fundamental,Universidade Federal de Pernambuco, Recife-PE.

Palavras Chave: *Enovelamento de Proteínas, Dinâmica Molecular, Modelos de Estado de Markov, Proteína α3D*

INTRODUÇÃO

O enovelamento de proteínas tem sido tema de pesquisa há mais de cinco décadas, contudo, existem dificuldades na construção de um modelo teórico que englobe todas as suas características. Atualmente, um método bem sucedido para a construção desses modelos, consiste na aplicação da dinâmica molecular (DM)[1]. Os desafios da DM nesse caso têm sido no tratamento da trajetória, devido ao longo tempo de simulação. Uma ferramenta que tem apresentando bons resultados são os modelos de estado de Markov (MSMs)[2]. Os MSMs são representados por uma cinética de primeira ordem, através de um grupo de estados discretos em que a análise das trajetórias são simplificadas, descartando-se eventos muito rápidos, abaixo do chamado tempo de latência (*lagtime*)[3]. Nesse estudo utilizamos um modelo de estado de Markov para analisar a dinâmica do enovelamento da proteína α3D (pdb: 2a3d) [1].

MÉTODOS

Inicialmente alinhamos a trajetória de 346μs da α3D com o programa MDtraj[3]. Após isso, construímos o MSM bayesiano com o programa pyEMMA[4]. Para a construção do MSM, geramos uma matriz de dimensões com as seguintes *features*: todos os ângulos de torção do *backbone*, ângulos χ_1 das cadeias laterais e o RMSD mínimo entre os *frames* da trajetória. Em seguida, aplicamos a técnica de análise de componentes independentes do tempo (TICA) para redução da dimensionalidade do sistema. Após, realizamos a clusterização das conformações com o método *k-means* e pegamos o centroide de cada um dos 550 grupos representativos. O *lag time* escolhido foi de 10ns, sendo validado com o teste de Chapman-Kolmogorov. Após a validação do MSM com 550 microestados, usamos a técnica PCCA+[1] para obter 5 macroestados representativos do sistema. Após esse procedimento, aplicamos a teoria do caminho de transição (TPT)[3] para gerarmos a matriz de transição entre os macroestados e obtermos o grafo com os possíveis caminhos entre os macroestados desenovelado e enovelado.

RESULTADOS

Na figura 1 é apresentado o grafo com 5 macroestados obtidos a partir da matriz de transição, saindo do macroestado **A** (desenovelado) para o macroestado **B** (estrutura nativa).

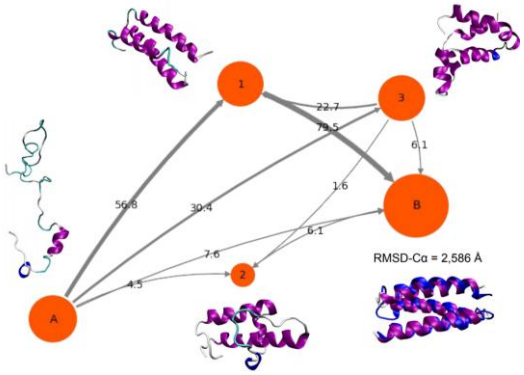


Figura 1: Grafo com probabilidades de transição entre os macorestados para a proteína α3D(pdb: 2a3d).

Verificamos que o caminho de maior fluxo, apresenta 56,8% de probabilidade de sair de (**A**), para o estado de transição (**1**) e 79,5% de chance de estando no estado (**1**) seguir para o estado enovelado em (**B**). Logo a α3D apresenta uma dinâmica de três estados, estando de acordo com dados da referência [2].

CONCLUSÕES

Observamos que o MSM apresentou dados detalhados sobre o enovelamento da α3D. Essas informações contribuíram para a análise dos possíveis caminhos do enovelamento possibilitando a inferência do tipo de mecanismo que a α3D assume durante esse processo.

REFERÊNCIAS

[1]Lindorff-Larsen, K. *et al*. *Science* **2011**, *334* (6055), 464–465.
[2]Beauchamp, K. A. *et al. PNAS* **2012**, 109, 17807−17813.
[3]McGibbon, R. T. *et al. Biophysical J.* **2015**, 109, 1528-1532.
[4]Scherer, M. K. *et al*. *JCTC* **2015**, *11* (11), 5525–5542.
[5]Sborgi, L. *et al. JACS* **2015**, *137*(20), 6506-6516.

AGRADECIMENTOS

CAPES/FAPESQ, CNPq, CENAPAD-SP, NPAD-UFRN, FACEPE

XXI Simpósio Brasileiro de Química Teórica
XXI Brazilian Symposium on Theoretical Chemistry
8 a 12 de Novembro de 2021
Virtual

Cálculos QM/MM ONIOM para a obtenção de informações sobre inibidores da toxina A da ricina (RTA)

Acassio Rocha-Santos (PG)[1], Gerd Bruno Rocha (PQ)[1]

acassioroch@gmail.com

[1]**Departamento de Química, Universidade Federal da Paraíba, João Pessoa – PB– Brasil.**
Palavras-Chave:*RTA, QM/MM ONIOM, Energias de Ligação.*

Introdução

A ricina é uma proteína citotóxica produzida na semente da mamona (*Ricinus communis*); pertence à família de proteínas inativadoras de ribossomos (tipo 2). Trata-se de uma das toxinas biológicas mais potentes conhecidas, sendo constituídas de duas subunidades, RTA e RTB, unidas por uma ponte de dissulfeto. A RTA (com 267 resíduos) é uma N-glicosidase que inativa ribossomos eucarióticos e a RTB (com 262 resíduos) é uma lectina responsável pela internalização do complexo RTA-RTB no citosol da célula [1]. Autoridades mundiais têm demonstrado preocupação em relação à toxicidade da ricina devido ao seu potencial uso como armas químicas por grupos terroristas. A ausência de medidas contra o envenenamento por ricina tem contribuído ainda mais para essas preocupações [1]. Desse modo, diversas abordagens de química teórica e computacional têm sido empregadas para obter informações acerca dos mecanismos de inibição da ricina, na qual a RTA é o principal alvo de candidatos a inibidores.

Metodologia

Nesse trabalho, realizamos cálculos *single-point* das energias de ligação para seis estruturas do complexo RTA-ligante: RTA-19M (PDB: 4HUP), RTA-RS8 (PDB: 4HUO), RTA-0RB (PDB: 4ESI), RTA-1MX (PDB: 4MX1), RTA-JP2 (PDB: 3PX8) e RTA-JP3 (PDB: 3PX9) e comparamos os resultados com dados experimentais de IC_{50}. Todos os complexos foram relaxados e equilibrados através de simulações de DM [2], sendo que o último frame da trajetória foi utilizado para realização dos cálculos do ΔE_{bind} através do método QM/MM ONIOM [3]. Para todos os cálculos, utilizamos os critérios padrão para cálculos QM/MM ONIOM [3] do programa Gaussian 09. Na parte QM da RTA e dos complexos RTA-ligante, foram utilizados os funcionais B3LYP [4], ωB97X-D [5] (que inclui correções de dispersão DFT-D2) e funções de base 6-31+G(d). Na parte MM foi utilizado o campo de força universal UFF. A parte QM da proteína RTA (189 átomos), incluiu os seguintes resíduos: Glu-177, Arg-180 (importantes para a catálise enzimática), Tyr-80, Val-81, Gly-121 e Tyr-123 (importantes para a ligação e reconhecimento do sítio ativo) [1]. Além desses, incluímos os resíduos Asn-122, Ser-176, Asn-209, Gly-212 e Arg-213 que estão a cerca de 3,0 Å de distância de algum dos seis ligantes, produzindo interações do tipo ligação de hidrogênio. As energias de ligação para os complexos RTA-ligante foram calculadas de acordo com a seguinte equação:

$$\Delta E_{bind} = \Delta E_{QM/MM}^{RTA-ligante} - \left(\Delta E_{QM}^{ligante} + \Delta E_{QM/MM}^{RTA}\right)$$

Resultados

Tabela 1: Energias de ligação, ΔE_{bind} (em hartrees) para os seis complexos RTA-ligante obtidos através de cálculos single-point QM/MM ONIOM.

Complexo	IC_{50} (μM)	ΔE_{bind}/ B3LYP (au)	ΔE_{bind}/ωB97X-D (au)
19M	15 (1)	-0,136 (1)	-0,230 (1)
RS8	20 (2)	-0,119 (2)	-0,202 (2)
0RB	70 (3)	-0,106 (3)	-0,178 (3)
1MX	209 (4)	-0,062 (6)	-0,130 (4)
JP2	230 (5)	-0,070 (4)	-0,129 (5)
JP3	380 (6)	-0,057 (5)	-0,088 (6)

Ao analisarmos os resultados da Tabela 1, verificamos que os cálculos *single-point* QM/MM ONIOM com o funcional B3LYP foram capazes de identificar os três melhores ligantes (19M, RS8 e 0RB), apresentando uma correlação de 0,929 com os dados de IC_{50}. A única exceção foi o ligante 1MX que apresentou $\Delta E_{(bind)}$ menor que com os ligantes JP2 e JP3. Quando mudamos o funcional B3LYP para o funcional ωB97X-D, que inclui correções de dispersão DFT-D2, verificamos que correlação aumentou de 0,929 para 0,972. Além disso, os resultados obtidos com funcional ωB97X-D classificaram corretamente todos os ligantes de acordo com dados de IC_{50}.

Conclusões

Os cálculos das energias de ligação obtidas através de métodos QM/MM ONIOM apresentaram boas correlações com dados experimentais de IC_{50}, sendo que a performance é aumentada com a utilização de funcionais com correção de dispersão.

Agradecimentos

Capes, Fapesq, CNPQ e NPAD-UFRN.

Referências

[1] Rocha-Santos, A. *et al.* ACS Omega **2021**, 6(13), 8764-8777.
[2] Chaves, E. J. F. *et al.* JCIM **2018**, 58(6),1205-1213.
[3] Dapprich, S. *et al.* THEOCHEM **1999**, 461, 1-21.
[4] Lee, C. *et al.* Phy. Rev. B **1998**, 37(2), 785.
[5] Chai, J-D. *et al.* PCCP **2008**, 10(44), 6615-6620.

SIMPÓSIO BRASILEIRO DE QUÍMICA TEÓRICA 2019

CERTIFICADO

Certificamos que o trabalho "*Análise do enovelamento da proteína α3D através de dinâmica molecular e modelo de estado de Markov*", de autoria de **Acassio R. Santos, Gabriel A. U. Carvalho, Gerd B. Rocha**, foi apresentado na forma de Apresentação Flash, no **XX Simpósio Brasileiro de Química Teórica**, realizado em João Pessoa-PB, no período de 10 a 14 de novembro de 2019.

Profª. Elizete Ventura do Monte
Coordenadora Geral

APOIO:

PATROCÍNIO:

XX SIMPÓSIO BRASILEIRO DE QUÍMICA TEÓRICA 2019

CERTIFICADO

Certificamos que o trabalho *"Análise do enovelamento da proteína α3D através de dinâmica molecular e modelo de estado de Markov"*, de autoria de **Acassio R. Santos; Gabriel A. U. Carvalho e Gerd B. Rocha**, apresentado na forma de Comunicação oral Flash, foi premiado com uma menção honrosa no **XX Simpósio Brasileiro de Química Teórica**, realizado em João Pessoa-PB, no período de 10 a 14 de novembro de 2019.

João Pessoa-PB, 14 de novembro de 2019.

Profª Elizete Ventura do Monte
Coordenadora Geral do Evento

Apoio:

Patrocínio:

XXI SBQT
SIMPÓSIO BRASILEIRO
DE QUÍMICA TEÓRICA

CERTIFICADO

Certificamos que o trabalho intitulado

"Cálculos QM/MM ONIOM para a obtenção de informações sobre inibidores da toxina A da ricina (RTA)",

apresentado por **Acassio Rocha Santos,**

foi agraciado com o prêmio de melhor pôster apresentado no XXI Simpósio Brasileiro de Química Teórica, realizado de 8 a 12 de novembro de 2021 no formato virtual.

Willian R. Rocha
Coordenador Geral - Presidente do Comitê Organizador
do XXI Simpósio Brasileiro de Química Teórica

Apoio:

Patrocínio:

ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Thermochemical and Quantum Descriptor Calculations for Gaining Insight into Ricin Toxin A (RTA) Inhibitors

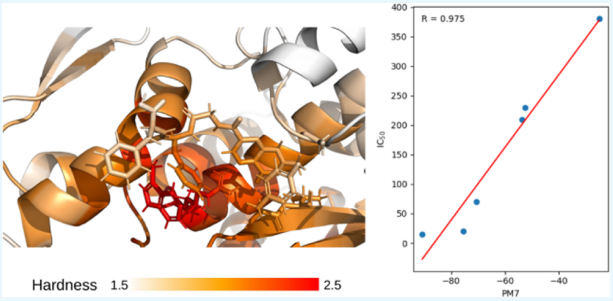
Acassio Rocha-Santos, Elton José Ferreira Chaves, Igor Barden Grillo, Amanara Souza de Freitas, Demétrius Antônio Machado Araújo, and Gerd Bruno Rocha*

Cite This: *ACS Omega* 2021, 6, 8764–8777

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: In this work, we performed a study to assess the interactions between the ricin toxin A (RTA) subunit of ricin and some of its inhibitors using modern semiempirical quantum chemistry and ONIOM quantum mechanics/molecular mechanics (QM/MM) methods. Two approaches were followed (calculation of binding enthalpies, ΔH_{bind}, and reactivity quantum chemical descriptors) and compared with the respective half-maximal inhibitory concentration (IC_{50}) experimental data, to gain insight into RTA inhibitors and verify which quantum chemical method would better describe RTA−ligand interactions. The geometries for all RTA−ligand complexes were obtained after running classical molecular dynamics simulations in aqueous media. We found that single-point energy calculations of ΔH_{bind} with the PM6-DH+, PM6-D3H4, and PM7 semiempirical methods and ONIOM QM/MM presented a good correlation with the IC_{50} data. We also observed, however, that the correlation decreased significantly when we calculated ΔH_{bind} after full-atom geometry optimization with all semiempirical methods. Based on the results from reactivity descriptors calculations for the cases studied, we noted that both types of interactions, molecular overlap and electrostatic interactions, play significant roles in the overall affinity of these ligands for the RTA binding pocket.



1. INTRODUCTION

Ricin is a cytotoxic protein produced in the seed of the castor bean plant (*Ricinus communis*), and it belongs to the ribosome-inactivating protein family (type-2) and is one of the most potent biological toxins known.[1] This protein consists of two subunits joined together by a disulfide bond, namely, ricin toxin A (RTA) and ricin toxin B (RTB).[2−4] RTA (267 residues) is an *N*-glycosidase that inactivates eukaryotic ribosomes via depurination of a specific adenine located in the sarcin-ricin loop (SRL) motif of the 28S rRNA subunit.[5] RTB (262 residues) is a lectin that mediates the uptake of holoricin into cells via recognition of galactose and *N*-6-acetylgalactosamine.[6] Once internalized, holoricin undergoes vesicular retrograde transport until reaching the endoplasmic reticulum lumen;[7] then, an isomerase reduces the disulfide bond between the subunits, and RTA is translocated to the cytosol and subsequently efficiently attacks ribosomes.[8,9]

There is a global concern by world authorities regarding ricin toxicity due to its potential use as a chemical weapon, mainly by terrorist and activist groups. In addition, the absence of countermeasures to ricin poisoning further contributes to this concern. In this way, theoretical methods, such as molecular docking, have guided research groups to find antagonist scaffolds for the RTA active site. However, this search has not been a trivial task because the active site of RTA and its surroundings are largely polar, imposing polarity constraints that in turn are not accounted for by the simplest theoretical methods.

To date, theoretical and computational chemistries have played a key role in studying biological and/or biochemical systems, providing proper direction toward drug discovery.[10] Knowledge of the intermolecular interactions of protein−ligand systems is an important feature for the development of new drugs.[11] In this way, some approaches, e.g., molecular docking, have been used to predict the bioactive pose of ligands in the active sites of biological targets with therapeutic interest;[12] however, it is quite ineffective to predict the relative binding affinity and rank ligands according to experimental data.[10,13,14]

Received: June 1, 2020
Accepted: December 30, 2020
Published: March 23, 2021

Currently, the estimation of the binding free energy (ΔG_{bind}) and binding enthalpy (ΔH_{bind}) of a ligand in a protein−ligand complex is a major challenge in computational chemistry.[10] To tackle this problem, several groups have been using computational chemistry methods to predict better energy profiles.[10,11,15−33] Among these approaches are those entirely based on classical force fields, such as MM-G(P)BSA,[34,35] LIE,[36] SMD,[37,38] and FEP,[39] as well as their versions involving hybrid potentials (quantum mechanics/molecular mechanics (QM/MM)) or those totally calculated by quantum mechanics (QM), such as QM-MM-G(P)-BSA,[40,41] QM-MM-LIE,[23,42,43] QM-FEP,[25,44] and QM-SMD.[45,46] All of these approaches require protein−ligand flexibility, either at their end-points or at the free-energy surface, resulting in high computational costs. Therefore, the use of such approaches is limited to a small set of ligands and small- and medium-sized complexes.

An alternative that requires lower computational cost is using theoretical methods for a single structure, in which the flexibility of the receptor (R), ligand (L), and receptor−ligand (R−L) complex can be partially accounted for by carrying out geometry optimization for R, L, and R−L. In this scenario, if the interest is to increase the accuracy of the receptor−ligand-binding energy predictions, the use of QM methods is mandatory. There are two approaches for calculating the binding energy of a ligand in a biological target using QM methods. The first approach considers a selection of the R−L complex atoms (QM cluster), and the second considers all atoms of the R−L complex in the QM calculation.

In the QM cluster strategy, a representative set of residues (ranging between 20 and 200 atoms) is selected and cut off from the active site. This set is studied separately by applying QM methods either in vacuum or in a continuous solvent model.[47,48] The advantage of this strategy is that few atoms are considered in the calculation, which allows us to assess a large set of ligands for the same active site. However, the disadvantage of this approach is that removing important residues during QM region selection can lead to inaccurate results.[49,50] In the strategy that considers all R−L atoms for QM calculations, it is possible to both hasten the computation of thermochemical properties and enable exploration of large ligand databases using linear scaling techniques coupled with graphics processing unit (GPU)-type accelerators.[51,52]

In the literature, some studies have performed molecular modeling and simulation involving ricin. Olson[53] applied molecular dynamics (MD) methods to analyze the structural and energetic aspects of three polynucleotide ligands (rRNA substrate analogues) that bound to the RTA active site. The results of the interaction energies showed that the overall binding is dominated by nonspecific interactions that occur in the region with a high degree of protein basicity through specific arginine contacts. The simulations of the three R−L complexes, as well as their comparison with experimental data, allowed a better understanding of the interaction of RTA with rRNA.

In another report, Olson and Cuff[54] expanded the previous study[53] by analyzing the free-energy determinants for the formation of RTA complexes with the rRNA substrate and several small ligands. The authors found that the absolute free energies of formation obtained for the RTA−RNA complex, as well as for several protein mutants, presented good agreement with the experimental data. In addition, it was observed that the terms of free energies presented unfavorable electrostatic contributions that were balanced by the favorable nonspecific hydrophobic effect, with free energies similar to those of protein−protein complexes. The individual components (by amino acid residue) of the binding free energy of the RTA−RNA complex revealed highly relevant electrostatic interactions arising from the charge−charge complementarity of the interfacial arginines with the RNA phosphate backbone. In addition, it has been observed that the hydrophobic complementarity of the domain is exerted by the base interactions of the GAGA loop structure.

Yan and co-workers[55] conducted studies on the interactions of small rings with the RTA active site to better understand how ricin recognizes adenine rings. The geometries and interaction energies were calculated using MM methods for some complexes between the RTA active site and tautomeric modifications of adenine, formycin, guanine, and pterin. The results indicated that the interaction energies between the pterin ring and RTA are stronger than those of formycin with RTA. It has also been found that formycin binds more strongly to RTA than adenine. This information presented good agreement with the experimental data. In addition, the results of experimental and molecular modeling work suggest that the binding site of ricin is quite rigid and can recognize only a small range of adenine-like rings.

In a recent study, Chaves and co-workers[12] carried out redocking of six known RTA inhibitors and then performed steered molecular dynamics (SMD) simulations to assess the relative binding affinity of these ligands. The molecular docking approach was able to predict the bioactive pose of the ligands; however, the score function was unable to rank these inhibitors according to experimental data. Steered MD simulation was used to decouple these ligands from the binding pocket, and the force profiles were estimated and presented a strong correlation with the experimental data ($R^2 > 0.9$).

As a matter of fact, there are few molecular modeling or simulation studies about ricin, and in our searches, we found no study applying quantum chemical methods for whole RTA−ligand complexes. Thus, in this work, we performed calculations of ΔH_{bind} and quantum descriptors between RTA and some of its inhibitors using semiempirical quantum chemical and QM/MM ONIOM methods to compare the results with experimental data (half-maximal inhibitory concentration (IC_{50})) and obtain local interaction information between the RTA residues and RTA inhibitors.

2. METHODS

We retrieved the following RTA−ligand complex structures from the Protein Data Bank (PDB): 4HUP (RTA−19M),[56] 4HUO (RTA−RS8),[56] 4ESI (RTA−0RB),[57] 4MX1 (RTA−1MX),[58] 3PX8 (RTA−JP2),[59] and 3PX9 (RTA−JP3).[59] Figure 1 presents the chemical structures of the six ligands of RTA studied in this work.

All RTA−ligand complexes were relaxed and then equilibrated using molecular dynamics simulations. We set the MD simulations according to the settings used in ref 12. The last frame of each MD simulation was used to calculate binding enthalpies using semiempirical quantum mechanical (SQM) methods considering two scenarios: (i) a direct single-point energy calculation with RM1,[60] PM6,[61] PM6-DH+,[62] PM6-D3H4, and PM7[63] and (ii) after full-atom geometry optimization with PM6-DH+, PM6-D3H4, and PM7.

https://doi.org/10.1021/acsomega.0c02588
ACS Omega 2021, 6, 8764−8777

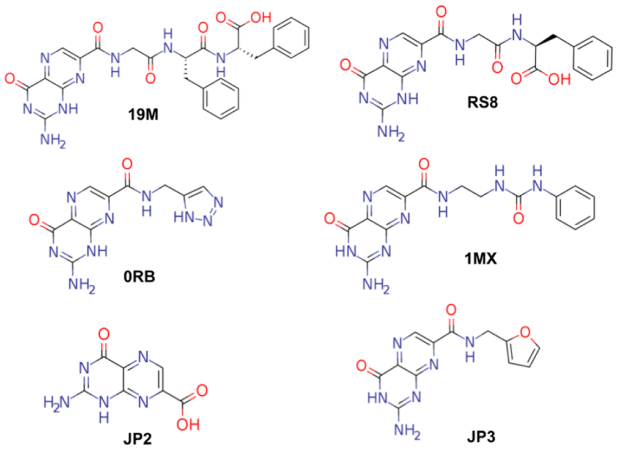


Figure 1. Structures of the six ligands of RTA studied in this work. The three-letter code corresponds to the PDB ID codes for these ligands.

For all semiempirical calculations, we used the linear scaling algorithm MOZYME,[64] available in the MOPAC2016 package.[65] For the SCF convergence criteria, we used standard settings and a cutoff radius of 10 Å for the MOZYME algorithm. For full-atom geometry optimizations, we used the limited-memory BFGS algorithm considering a norm of gradient of 5.0 kcal·mol^{-1}·Å^{-1} as stop criteria for complexes and noncomplexed proteins and 0.01 kcal·mol^{-1}·Å^{-1} for the free ligands, where no restrictions of movement were considered for the atoms. For all calculations (single-point and geometry optimization calculations), we considered the conductor-like screening model (COSMO) implicit solvent field with a relative permittivity of 78.4 and an effective solvent molecule radius of 1.3 Å. The binding enthalpies for the ligand−RTA complexes were calculated according to eq 1. For all noncomplexed proteins and R−L complexes, we assumed a total charge of +2 *e*.

$$\Delta H_{\text{binding}} = \Delta H_f^{\text{ligand-RTA}_{\text{complex}}} - (\Delta H_f^{\text{ligand}} + \Delta H_f^{\text{RTA}}) \quad (1)$$

The RTA has 4200 atoms and the ligands 19M, RS8, 0RB, 1MX, JP2, and JP3 have 67, 47, 30, 43, 20, and 31 atoms, respectively. The summation between the number of atoms for RTA and each ligand gives the number of atoms for the respective RTA−ligand complex.

For the sake of comparison and to test the performance of semiempirical methods, we also carried out single-point calculations considering hybrid QM/MM ONIOM method[66−68] using the same geometries from DM for the six RTA−ligand complexes. We used Gaussian 09 program[69] with the hybrid GGA B3LYP functional[70,71] and the hybrid GGA ωB97X-D functional,[72] which includes DFT-D2[73] dispersion correction. In both cases, we used the 6-31+G(d) basis set[74−76] for the QM region in both RTA protein and RTA−ligand complexes. UFF universal force field[77] was used for the remainder of the protein, which is denoted as the MM part. As requested, hydrogen atoms were used as link atoms connecting these two parts. We used all default criteria for ONIOM QM/MM calculations in Gaussian 09 program.

The QM part of the RTA protein includes the following residues: Glu-177, Arg-180 (important for enzyme catalysis), Tyr-80, Val-81, Gly-121, and Tyr-123 (important for attachment and recognition of the ligand at the active site).[78] In addition, we include residues Asn-122, Ser-176, Asn-209, Gly-212, and Arg-213 because they are about 3.0 Å apart from any of the six ligands, thereby forming hydrogen-bond interactions, producing a model with 189 atoms, 716 electrons, and +1 charge. For the RTA−ligand complexes, we also included the respective ligands in the QM part. Therefore, models were generated with (256 atoms, 1008 electrons and +1 charge), (236 atoms, 930 electrons and +1 charge), (219 atoms, 864 electrons and +1 charge), (232 atoms, 908 electrons and +1 charge), (209 atoms, 822 electrons and +1 charge), and (220 atoms, 864 electrons and +1 charge) for RTA−19M, RTA−RS8, RTA−0RB, RTA−1MX, RTA−JP2, and RTA−JP3 complexes, respectively. The binding energies for the ligand−RTA complexes were calculated according to eq 2.

$$\Delta E_{\text{binding}} = \Delta E_{\text{QM/MM}}^{\text{ligand-RTA}_{\text{complex}}} - (\Delta E_{\text{QM}}^{\text{ligand}} + \Delta E_{\text{QM/MM}}^{\text{RTA}}) \quad (2)$$

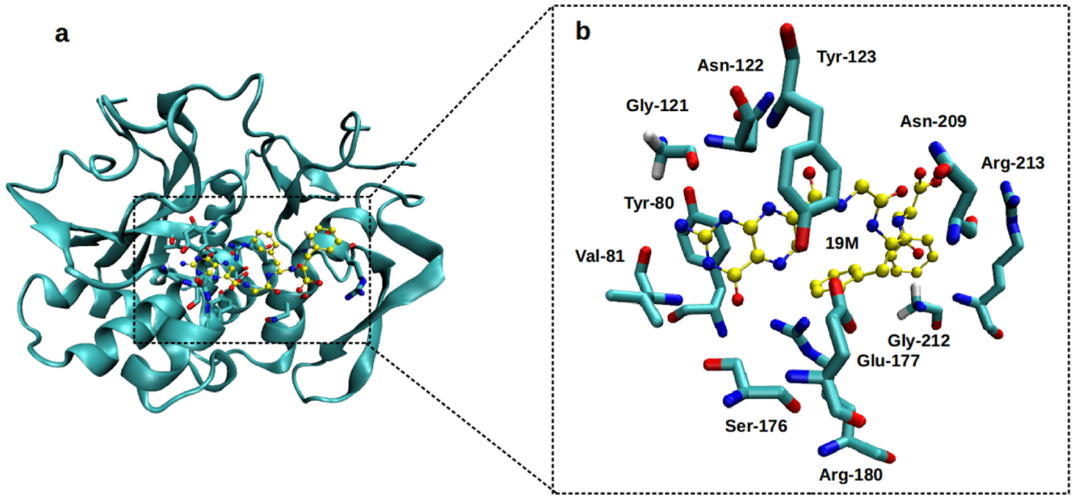


Figure 2. (a) Geometry of RTA−19M complex used in the ONIOM QM/MM calculation. (b) Zoomed view of the active site showing the 11 residues and 19M ligand (in yellow).

https://doi.org/10.1021/acsomega.0c02588
ACS Omega 2021, 6, 8764−8777

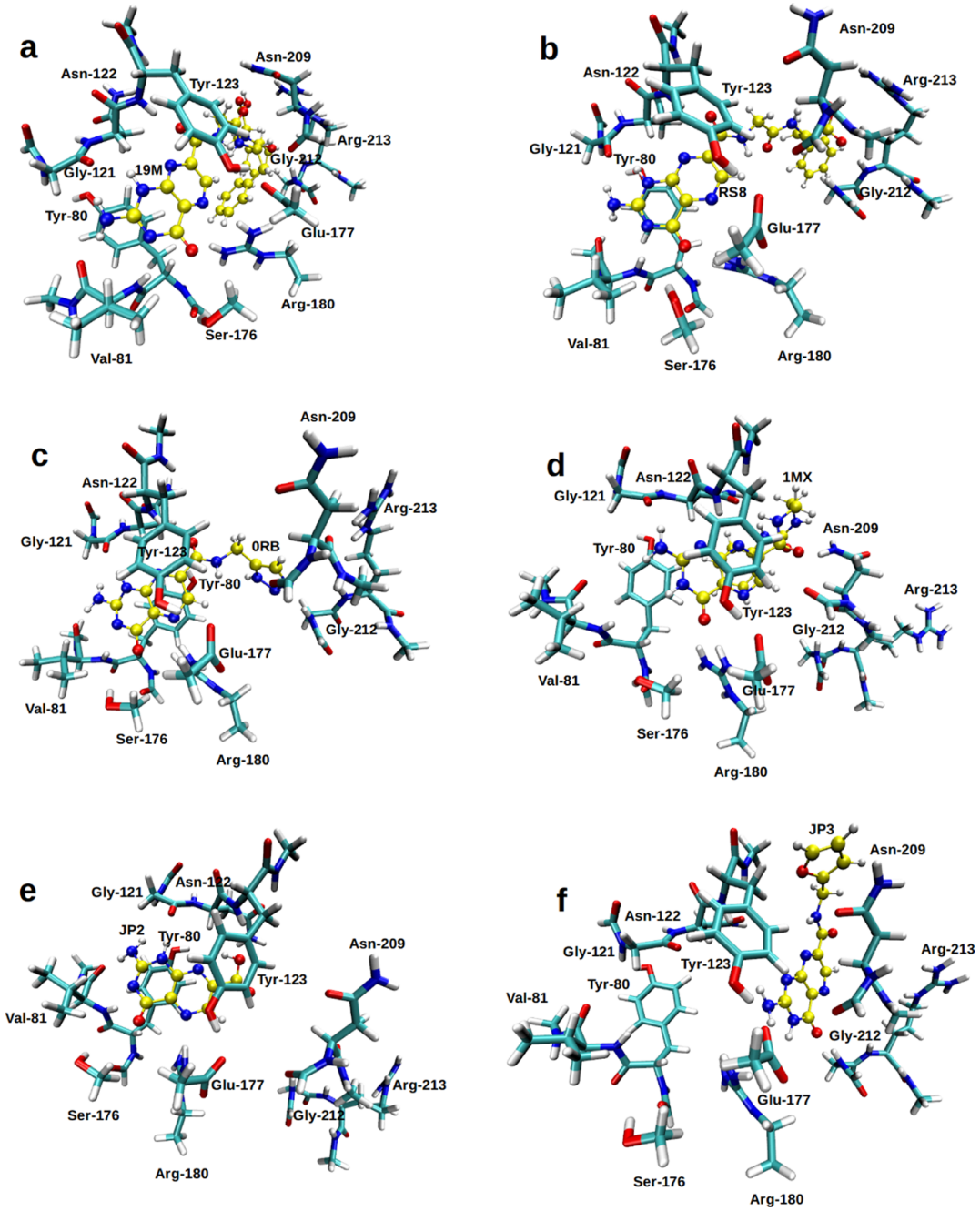


Figure 3. QM models for ONIOM QM/MM calculations for complexes: (a) RTA−19M, (b) RTA−RS8, (c) RTA−0RB, (d) RTA−1MX, (e) RTA−JP2, and (f) RTA−JP3.

In Figure 2, we present the geometry for the RTA−19M complex with emphasis on the residues of the active site and the ligand (in yellow).

In Figure 3, we present QM models used for all complexes. We considered all atoms for residues Tyr-80, Val-81, Gly-121, Asn-122, Tyr-123, Asn-209, Gly-212, and Arg-213 because there are interactions between atoms of protein backbone and the ligands for such residues. For residues Ser-176, Glu-177, and Arg-180, we consider only side chains since intermolecular interactions with ligands occur only in this region of amino acids.

3. REACTIVITY DESCRIPTOR CALCULATIONS

The most successful quantum descriptors come from conceptual density functional theory (CDFT), which through the mathematical development of density functional theory has given valid quantitative definitions for well-known chemical concepts,[79] such as hardness and softness. From this theory, the main interactions between two molecules are summarized in two types of processes: mutual polarization followed by molecular orbital overlap interactions and Coulombic forces.[80] The propensity of these effects in molecules can be traced locally using the Fukui function for

https://doi.org/10.1021/acsomega.0c02588
ACS Omega 2021, 6, 8764−8777

the first type of interaction and local hardness for the second.[81]

To locally represent the molecular orbital overlap interactions, we used the Fukui function.[81] This descriptor was further divided into two functions, namely, the left Fukui function (f^-), specific for electrophilic attack susceptibility (which for the frozen orbital approximation is equal to the highest-energy occupied molecular orbital (HOMO) density[80]), and the right Fukui function (f^+), specific for nucleophilic attack susceptibility (which is equal to the lowest-energy unoccupied molecular orbital (LUMO) density).

A convenient representation of the Fukui functions is where the values are assigned for each k atom center in the molecule. The f^- defined in eq 3 is the sum of the squared ν atomic orbital (AO) coefficients that comprise the HOMO and belong to the kth atom, plus the sum of the product between the coefficients of indices ν, μ, and the corresponding overlap matrix element $S_{\mu\nu}$.[82] An equivalent definition for the f^+ is presented in eq 4, using the LUMO instead of the HOMO.

$$f^-(k) = \sum_{\nu \in k}^{AO} |C_{\nu HOMO}|^2 + \sum_{\mu \notin \nu}^{AO} |C_{\nu HOMO} C_{\mu HOMO}| S_{\mu\nu} \quad (3)$$

$$f^+(k) = \sum_{\nu \in k}^{AO} |C_{\nu LUMO}|^2 + \sum_{\mu \notin \nu}^{AO} |C_{\nu LUMO} C_{\mu LUMO}| S_{\mu\nu} \quad (4)$$

In this study, the overall effects tallied in these two Fukui functions were used in combination via the average Fukui function, defined in eq 5.

$$f^0 = \frac{f^+(k) + f^-(k)}{2} \quad (5)$$

These quantum chemical descriptors have been identified as useful for determining the toxicity and biological activity of several potential ligands.[83,84] For the enzymes, these descriptors have been used to determine reactivity sites, as score functions in molecular docking, and to search protein native structures[85] and find key structures in reaction path simulations.[86,87]

The electronic structure of protein systems presents several relevant molecular orbitals for chemical interactions with the electronic energy near the HOMO and LUMO.[84,86] Thus, in this study, we computed the Fukui functions not only by using the HOMO and LUMO orbitals but also by considering all of the molecular orbitals from a range of 3 eV from the HOMO and LUMO. As Fukushima and co-workers showed in their work, this value can vary from 1 to 5 eV depending on the system under study.[88]

We used the local hardness ($H(k)$) as a quantum descriptor to compute the Coulombic force interactions. This particular definition is based on the electron−electron contribution of the molecular electrostatic potential, as shown in eq 6;[89] the calculation of local hardness at the kth atomic center is defined as a sum of the left Fukui function for each lth atomic center divided by their Euclidean distance R_{kl}. These theoretical quantities were calculated using the PRIMoRDiA software.[90]

$$H(k) = \sum_{l \neq k}^{atoms} \frac{f^-(l)}{R_{kl}} \quad (6)$$

4. RESULTS AND DISCUSSION

In Figure 4, we present the ΔH_{bind} results of all of the RTA−ligand complexes studied in this work, considering both strategies are carried out: (i) direct single-point energy calculation and (ii) calculation of ΔH_{bind} after full-atom geometry optimization of R, L, and R−L.

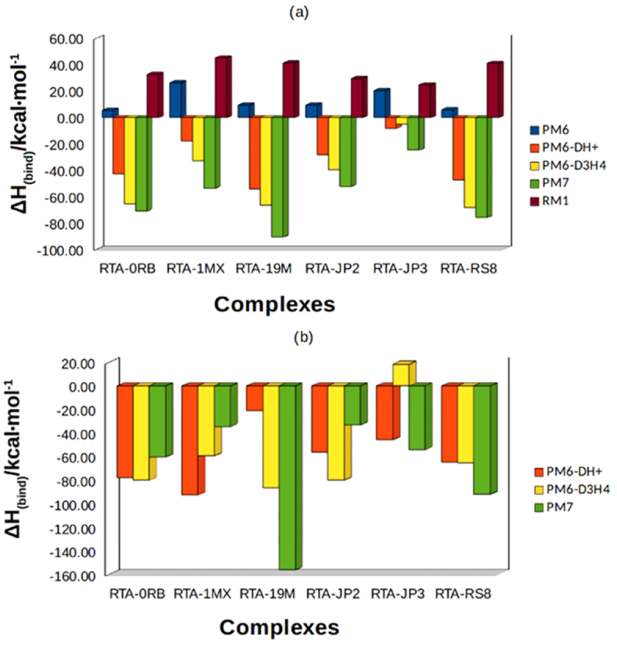


Figure 4. Binding enthalpies, ΔH_{bind}, for the complexes (RTA−0RB, RTA−1MX, RTA−19M, RTA−JP2, RTA−JP3, and RTA−RS8). In (a), the results are depicted from single-point calculations, and in (b), the results are after full-atom geometry optimization of R, L, and R−L.

Looking at the results for ΔH_{bind} calculated using the single-point strategy (Figure 4a), we can see that PM6-DH+, PM6-D3H4, and PM7 predicted negative ΔH_{bind} values for all studied RTA−ligand complexes. The range of ΔH_{bind} values is also consistent with the expected results from ligand−enzyme thermochemical assays, approximately a dozen of kcal·mol^{-1}.[91]

PM6 and RM1 showed low performance, presenting positive values for ΔH_{bind} for all RTA−ligand complexes. By comparing the performance among the PM6, PM6-DH+, and PM6-D3H4 methods, we verified that there are differences in binding enthalpies when corrections for dispersion and hydrogen bonding are considered. These results are consistent with recent studies suggesting that the quality of semi-empirical methods presents significant improvements when such corrections are considered.[16,18,92] These studies also indicate that SQM methods incorporating dispersion and hydrogen bonding yield results similar to those of D-type corrections for density functional theory (DFT) methods.

Figure 4b shows the results of ΔH_{bind} after full-atom geometry optimization for all complexes using the PM6-DH+, PM6-D3H4, and PM7 methods. We verified that the methods followed the same tendency of the single-point calculations, i.e., the calculated binding enthalpies were negative for all RTA−ligand complexes. The only exception was the RTA−JP3 ligand, which presented a positive ΔH_{bind} value with the PM6-D3H4 method. This result suggests that the PM6-DH+,

https://doi.org/10.1021/acsomega.0c02588
ACS Omega 2021, 6, 8764−8777

Table 1. Binding Enthalpies (kcal·mol^{-1}) for the Six RTA−Ligand Complexes Evaluated in This Study

		ΔH_{bind} via single-point calculations					ΔH_{bind} via geometry optimizations		
ligand	IC_{50} (μM)	PM6	PM6-DH+	PM6-D3H4	PM7	RM1	PM7	PM6-DH+	PM6-D3H4
19M	15 (1)	9.58 (3)	−54.37 (1)	−66.47 (2)	−90.76 (1)	41.27 (5)	−156.11 (1)	−20.97 (6)	−86.38 (1)
RS8	20 (2)	6.10 (2)	−47.31 (2)	−68.23 (1)	−75.59 (2)	41.12 (4)	−93.33 (2)	−64.95 (3)	−65.33 (4)
0RB	70 (3)	5.55 (1)	−42.88 (3)	−65.53 (3)	−70.5 (3)	32.74 (3)	−60.5 (3)	−78.21 (2)	−80.12 (2)
1MX	209 (4)	26.28 (6)	−18.06 (5)	−33.01 (5)	−53.65 (4)	45.24 (6)	−34.41 (5)	−92.76 (1)	−59.54 (5)
JP2	230 (5)	9.71 (4)	−28.5 (4)	−39.53 (4)	−52.39 (5)	−29.51 (1)	−33.23 (6)	−56.1 (4)	−79.94 (3)
JP3	380 (6)	20.54 (5)	−8.12 (6)	−5.16 (6)	−24.89 (6)	24.78 (2)	−54.16 (4)	−45.87 (5)	18.35 (6)

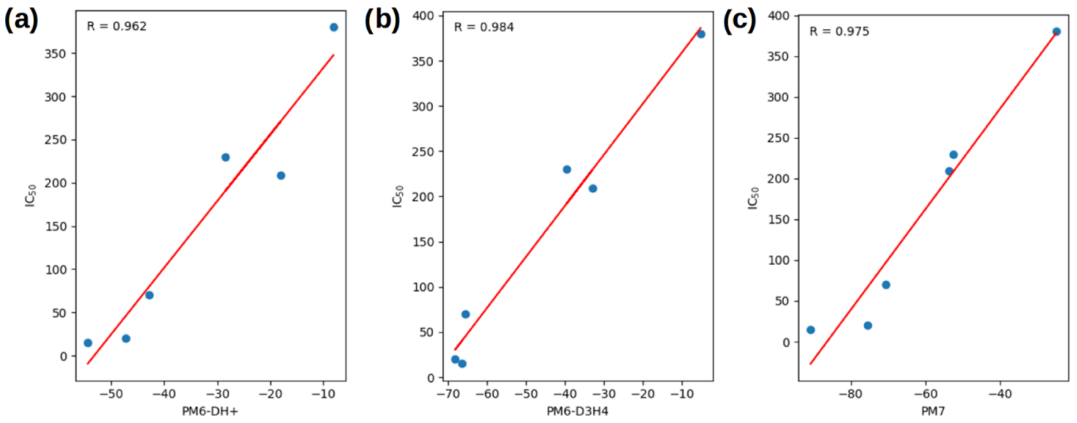


Figure 5. Correlation graphs between IC_{50} and ΔH_{bind} data obtained by single-point calculations for the (a) PM6-DH+, (b) PM6-D3H4, and (c) PM7 methods.

PM6-D3H4, and PM7 methods are able to describe, at least qualitatively, the binding enthalpies of the studied systems. We observed that the full-atom geometry optimization showed a tendency to decrease the ΔH_{bind} value of the complexes. From the 18 ΔH_{bind} calculations performed (optimization of six complexes with three distinct semiempirical methods: PM6-DH+, PM6-D3H4, and PM7), 12 presented a decreased ΔH_{bind} value compared to those obtained by single-point calculations. Exceptions occurred for the RTA−0RB (ΔH_{bind} = −60.50 kcal·mol^{-1}), RTA−1MX (ΔH_{bind} = −34.41 kcal·mol^{-1}) and RTA−JP2 (ΔH_{bind} = −33.23 kcal·mol^{-1}) complexes with the PM7 method, RTA−JP3 (ΔH_{bind} = 18.35 kcal·mol^{-1}) and RTA−RS8 (ΔH_{bind} = −65.33 kcal·mol^{-1}) complexes with the PM6-D3H4 method and the RTA−19M (ΔH_{bind} = −20.97 kcal·mol^{-1}) complex with the PM6-DH+ method.

In Table 1, we present the IC_{50}[56−59] and ΔH_{bind} values for each RTA−ligand complex evaluated in this work.

When comparing the ΔH_{bind} results obtained by single-point calculations using the PM6, PM6-DH+, PM7, and RM1 methods with IC_{50} experimental data for the ligands (19M, RS8, 0RB, 1MX, JP2, and JP3),[12] we found that the PM6 method showed correlation of 0.688 and RM1 method showed no correlation with the IC_{50}. On the other hand, the PM6-DH+ and PM7 methods were able to identify the 19M ligand, which has the lowest IC_{50} value (15 μM), as the best ligand.

The PM6-D3H4 method switched the rank order of the 19M and RS8 ligands (ΔH_{bind} = −66.47 and −68.23 kcal·mol^{-1}, respectively), but this may have occurred because the IC_{50} values of these two ligands are very close (15 and 20 μM) so that the PM6-D3H4 method was not sensitive enough to rank the best ligand for small IC_{50} variations. The PM6-DH+ method also showed inversions of binding enthalpy values between the 1MX (ΔH_{bind} = −18.06 kcal·mol^{-1}) and JP2 (ΔH_{bind} = −28.50 kcal·mol^{-1}) methods. The PM7 method was able to correctly rank all of the ligands without inversions, being more sensitive to the small variations of IC_{50}.

When we performed the statistical treatment between the IC_{50} and ΔH_{bind} data obtained by single-point calculations, we found R values of 0.962, 0.975, and 0.984 for the PM6-DH+, PM7, and PM6-D3H4 methods, respectively. Thus, the PM6-D3H4 method presented the best correlation with the IC_{50} data. So, it is worth mentioning that the results for PM6-DH+ and PM6-D3H4 are quite satisfactory. In Figure 5, we present the correlation graphs between the IC_{50} and ΔH_{bind} data obtained by single-point calculations when the PM6-DH+, PM6-D3H4, and PM7 methods were used.

When analyzing the binding enthalpy data with full-atom geometry optimization (Table 1), we verified that the PM6-DH+ method, although presenting good results with single-point calculations, did not show the same performance. This method was not able to identify 19M as the best ligand or present a good correlation with the IC_{50} data.

The PM7 method could rank the best ligands (from 1 to 3), but there were inversions in the ranking of the other ligands. Besides, there is a very marked distortion between the variation in the binding enthalpy results and IC_{50} values. The PM6-D3H4 method was able to identify the best ligand (19M) but showed inversions in the ranking between the RS8 and 0RB ligands and 1MX and JP2 ligands. However, compared to the PM7 method, the PM6-D3H4 method presented lower distortions between the ΔH_{bind} and IC_{50} data.

When the statistical treatment was performed between the IC_{50} and ΔH_{bind} data, we found R values of −0.085, 0.672, and 0.789 for the PM6-DH+, PM7, and PM6-D3H4 methods,

https://doi.org/10.1021/acsomega.0c02588
ACS Omega 2021, 6, 8764−8777

respectively. Therefore, the PM6-D3H4 method also presented the best correlation with the IC_{50} when we performed a full-atom geometry optimization strategy. Moreover, we observed that in the full-atom geometry optimization, the correlation between the IC_{50} data and binding enthalpies decreased considerably.

Sulimov et al.[93] recalculated the energies of approximately 8000 best-ranked structures in the molecular docking with the MMFF94 force field through single-point calculations and ligand optimization using the PM7 method, considering an implicit COSMO solvent model. The authors observed that the energies obtained by single-point calculations greatly improved the ranking of structures close to the native state. However, when optimizing the ligands with the PM7 method in vacuum and recalculating their energies with the COSMO model, the authors observed a worsening of the results for several complexes compared to those with PM7 (single-point energy calculation). In accordance with those results, we observed more comprehensively that the geometry optimizations reduced the performance of all considered semiempirical methods.

Although the correlation for full-atom geometry optimizations is lower than that obtained through single-point energy calculations, we emphasize that if the molecular dynamics simulation is not carried out properly to find an equilibrated structure, then this optimization is necessary to obtain some correlation with experimental data. We have indicated PM7 and PM6-D3H4 as the best semiempirical methods for obtaining binding enthalpies and ligand rankings. We are not stating that single-point QM calculations on the end frame from MD simulation are better for ligand-binding energies than ones that use geometry optimization at the QM level. However, this was true for our study. Effects due to low sampling, entropy lacking, or cancelation of errors could be behind these results, but the investigation of these issues is outside the scope of this study. Anyhow, an excellent discussion about these issues can be found in the review by Ryde and Soderhjelm.[10]

In Table 2, we present the root-mean-square deviation (RMSD) data between the MD structures and those optimized with the PM6-DH+, PM6-D3H4, and PM7 methods.

When analyzing the data in Table 2, we verified that the PM6-D3H4 method presented the lowest RMSD values, except for the RTA−JP2 complex, where the PM6-DH+ method presented a lower RMSD value. The PM7 method presented the highest RMSD values for all of the analyzed complexes. This observation suggests that the PM7 method is able to relax the initial structure more than the PM6-DH+ and PM6-D3H4 methods. Figure 6 shows the superposition between the initial structure and the structures optimized via the PM7 method.

Table 2. RMSD Results (in Å) between the End Frame of the MD Simulations and the Optimization of These Same End Frames by Semiempirical Quantum Methods[a]

complex	PM6-DH+ (Å)	PM6-D3H4 (Å)	PM7 (Å)	crystallographic (Å)
RTA−19M	1.554	1.296	2.108	2.176
RTA−RS8	1.747	1.533	2.024	2.266
RTA−0RB	1.567	1.401	1.950	1.839
RTA−1MX	1.832	1.468	1.919	2.676
RTA−JP2	1.581	1.626	2.036	1.917
RTA−JP3	1.745	1.511	2.057	2.103

[a]In the last column, we present RMSD results (in Å) between the end frame of the MD simulations and each crystallographic structure.

From the analysis of the RTA structures in the complexes, we verified that there were no significant changes after optimization, with conservation of the secondary structures. All R−L complexes presented RMSDs close to 2.0 Å, which is consistent with the resolution of the experimental data.[94]

When comparing the ligand poses via full-atom geometry optimization with structures from molecular dynamics, we observed that the 0RB ligand (Figure 6a) presented rotation of the triazole group, being approximately perpendicular to the preferred position of this group in the protein. The ligand 1MX (Figure 6b) presented small pterin group torsion and benzene ring variation. Due to its high conformational flexibility, the 19M ligand (Figure 6c) underwent a large modification after geometry optimization, mainly regarding torsions in the pterin moiety, benzene rings, and carboxylic acid group. The JP2 ligand (Figure 6d) exhibited rotation of its carboxylic acid group and small deviations in other groups of the structure. The JP3 ligand (Figure 6e) showed modification of its structure, with displacements in the positions of atoms along the same plane. The RS8 ligand (Figure 6f) exhibited a small variation in comparison to that observed in the last frame obtained from the molecular dynamics simulations.

In Table 3, we present the RMSD data between the ligands of complexes from the MD simulations and the ligands of the complexes optimized with the PM6-DH+, PM6-D3H4, and PM7 methods.

When analyzing the data of Table 3, we verified that the poses of the ligands presented several variations after the geometry optimization for all R−L complexes. For the PM6-DH+ method, the 1MX ligand presented the highest variation (RMSD = 0.932 Å), and the JP3 ligand presented the lowest variation for this method (RMSD = 0.581 Å). For the PM6-D3H4 method, it was observed that the RS8 and 1MX ligands presented the highest and lowest variations, respectively, with RMSDs equal to 1.291 and 0.477 Å. For the PM7 method, the 19M ligand showed the highest variation (RMSD = 1.530 Å), and the lowest variation for this method was for the RS8 ligand (RMSD = 0.598 Å). We observed that the PM7 method showed higher RMSD values for four of the six ligands, and the two exceptions were the RS8 and 1MX ligands.

In Table 4, we present the IC_{50}[56−59] and ΔE_{bind} values for each RTA−ligand complex calculated via ONIOM QM/MM with B3LYP and ωB97X-D functionals.

When analyzing the data in Table 4, we found that the single-point ONIOM QM/MM calculations with B3LYP functional were able to identify the three best ligands (19M, RS8, and 0RB). The only exception was the inversion between the ligands 1MX and JP2. The correlation between the IC_{50} and ΔE_{bind} data presented an R-value of 0.929. Although the ONIOM QM/MM results with B3LYP functional showed good correlation with IC_{50}, PM6-DH+, PM7, and PM6-D3H4 showed better results.

For the sake of testing, we also carried out single-point ONIOM QM/MM calculations with the B3LYP functional considering a smaller QM region for each complex. In this test, we considered in the QM part only the six most

https://doi.org/10.1021/acsomega.0c02588
ACS Omega 2021, 6, 8764−8777

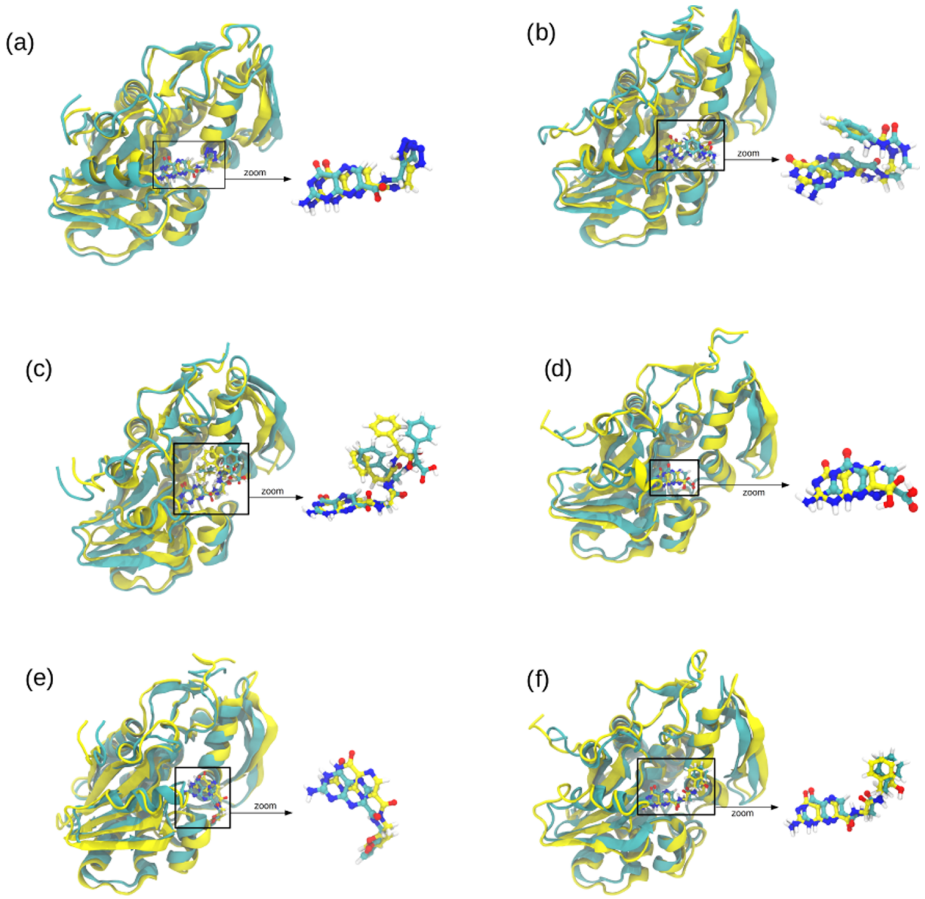


Figure 6. Superposition of the RTA−ligand optimized complexes: (a) RTA−0RB, (b) RTA−1MX, (c) RTA−19M, (d) RTA−JP2, (e) RTA−JP3, and (f) RTA−RS8. The optimized structures are represented in green, and initial structures are represented in yellow.

Table 3. RMSD Results (in Å) between the Ligands of the Complexes Optimized by Semiempirical Methods and the Ligands of the Complexes of the MD Simulations[a]

complex	PM6-DH+ (Å)	PM6-D3H4 (Å)	PM7 (Å)	crystallographic (Å)
19M	0.675	0.657	1.530	6.005
RS8	0.808	1.291	0.598	5.819
0RB	0.761	0.755	1.062	2.856
1MX	0.932	0.477	0.812	4.318
JP2	0.889	0.500	0.909	1.957
JP3	0.581	0.510	0.612	2.597

[a]In the last column, we present RMSD results (in Å) between the ligands of the MD simulations and the ligands from crystallographic structures.

Table 4. Binding Energies, ΔE_{bind}, (in hartrees) for the Six RTA−Ligand Complexes Calculated with ONIOM QM/MM Single-Point Calculations

complex	IC_{50} (μM)	ΔE_{bind}/B3LYP (au)	$\Delta E_{bind}/\omega$B97X-D (au)
19M	15 (1)	−0.136 (1)	−0.230 (1)
RS8	20 (2)	−0.119 (2)	−0.202 (2)
0RB	70 (3)	−0.106 (3)	−0.178 (3)
1MX	209 (4)	−0.062 (6)	−0.130 (4)
JP2	230 (5)	−0.070 (4)	−0.129 (5)
JP3	380 (6)	−0.057 (5)	−0.088 (6)

important residues of RTA plus the ligand. When we did this the correlation decreases dramatically to 0.693.

When we changed the B3LYP functional to the ωB97X-D functional, which includes DFT-D2[73] dispersion correction, we observe that the *R*-value increased from 0.929 to 0.972, showing a slight improvement in the correlation with IC_{50}. In addition, calculated results obtained with this functional, all ligands were ranked correctly according to ΔE_{bind} and IC_{50} values.

Considering these results and the low computational cost of semiempirical methods, we can state that methods PM7, PM6-DH+, and PM6-D3H4 showed better performance than the ONIOM QM/MM calculation with B3LYP functional for RTA complexes studied in this paper. When using the ωB97X-D functional, the QM/MM method presented *R*-value and performance similar to the PM7 method but with a higher computational cost. The advantage of ONIOM QM/MM is that one can always improve the results by increasing the QM region and using larger basis sets, and/or taking solvent effects and dispersion corrections into account, but this ends up being computationally expensive.

As a final experiment with binding affinities for these RTA complexes, we carried out new calculations using the dataset of binding affinities produced in the study carried out by Chaves and his collaborators.[12]

Our inspiration was a study carried out by Gupta and collaborators, which achieved a significant improvement for

https://doi.org/10.1021/acsomega.0c02588
ACS Omega 2021, 6, 8764−8777

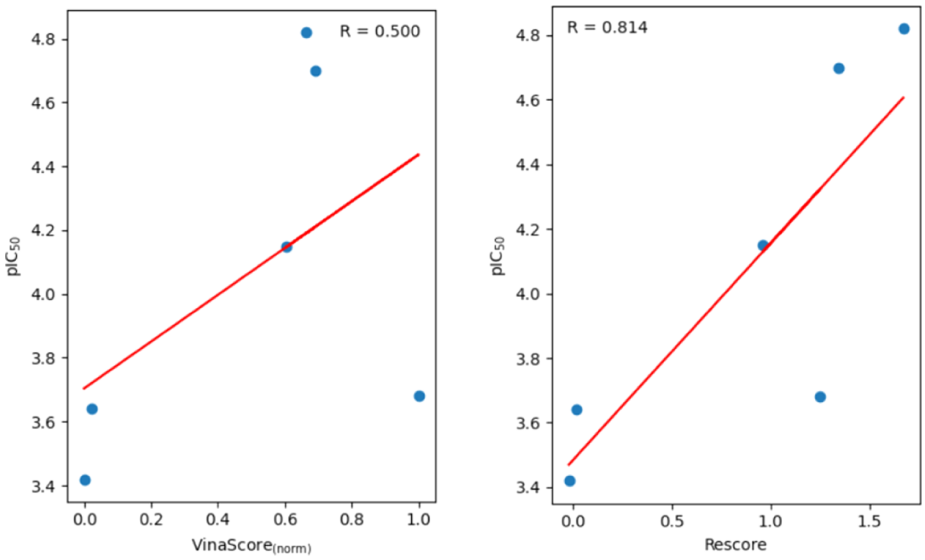


Figure 7. Pearson correlation between pIC_{50} vs Vina score and pIC_{50} vs re-score that is defined in eq 7.

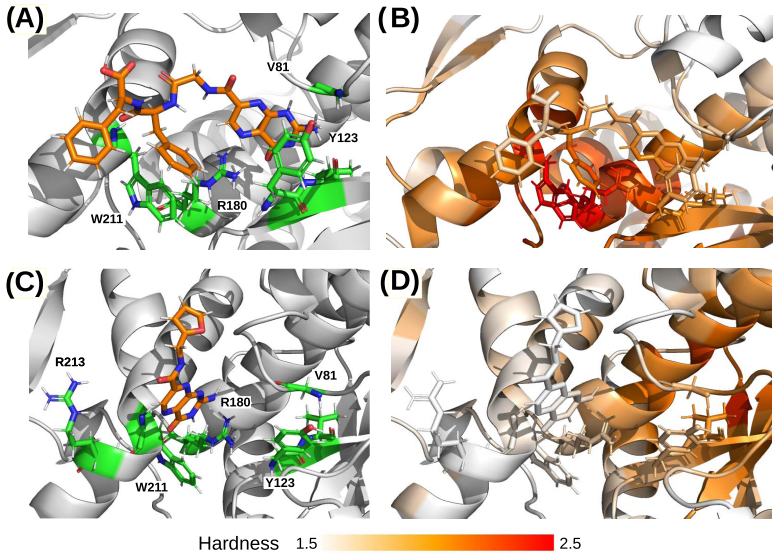


Figure 8. Calculated local hardness for RTA complexes with the 19M and JP3 ligands (orange). (A) Labels for the nearest residues (green) from 19M; (B) local hardness for the RTA−19M complex; (C) labels for the nearest residues (green) from JP3; and (D) local hardness for the RTA−JP3 complex.

the binding affinity predictions when they considered molecular descriptors in different molecular docking approaches for the inhibitors of microsomal prostaglandin E synthase-1.[95] In their study, molecular descriptors such as topological polar surface area (TPSA), partition coefficient (log *P*), volume (Vol), and the number of rotatable bonds (Nrtb) were used as docking scores. With this, the authors took into account desolvation penalties and conformational free-energy changes when a ligand binds to a protein. In their study, the re-score routine was considered as an empirical summation of normalized docking affinities. Similarly, we also incorporated a re-score routine to the binding affinity predictions provided by Vina software[96] for the same set of ligands (19M, RS8, 0RB, 1MX, JP2, and JP3). These binding affinity predictions were retrieved from another validation study carried out by our group; therefore, details about docking protocol can be reviewed in ref 12. In summary, the ligand molecular descriptors such as TPSA, Vol, and Nrtb were calculated using the web service Molinspiration (https://www.molinspiration.com). We also normalized the docking scores and the molecular descriptors to values between 0 and 1. Then, the following equation was used to perform re-score calculations

$$\text{rescore} = \text{docking score} + \text{TPSA} + \text{Vol} - \text{Nrtb} \quad (7)$$

We present in Figure 7 the Pearson correlations between pIC_{50} vs Vina score and pIC_{50} vs re-score, respectively. In short, our results showed that the consideration of such molecular descriptors improves the relative binding affinity predictions provided by Vina software, with an increase from 0.500 to 0.814 in the Pearson correlation.

https://doi.org/10.1021/acsomega.0c02588
ACS Omega 2021, 6, 8764−8777

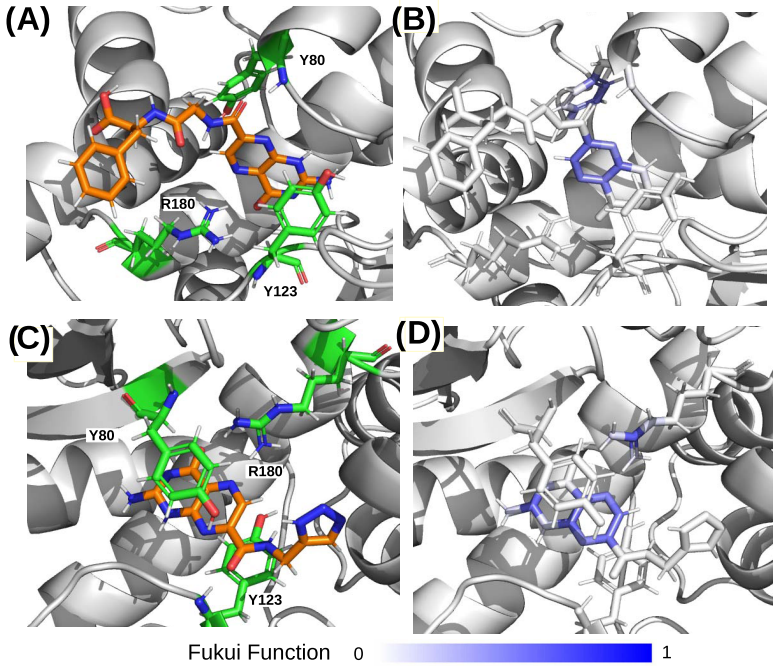


Figure 9. Calculated Fukui function for RTA complexes with the RS8 and 0RB ligands (orange). (A) Labels for the nearest residues (green) from RS8; (B) Fukui function for the RTA−RS8 complex; (C) labels for the nearest residues (green) from 0RB; and (D) local hardness for the RTA−0RB complex.

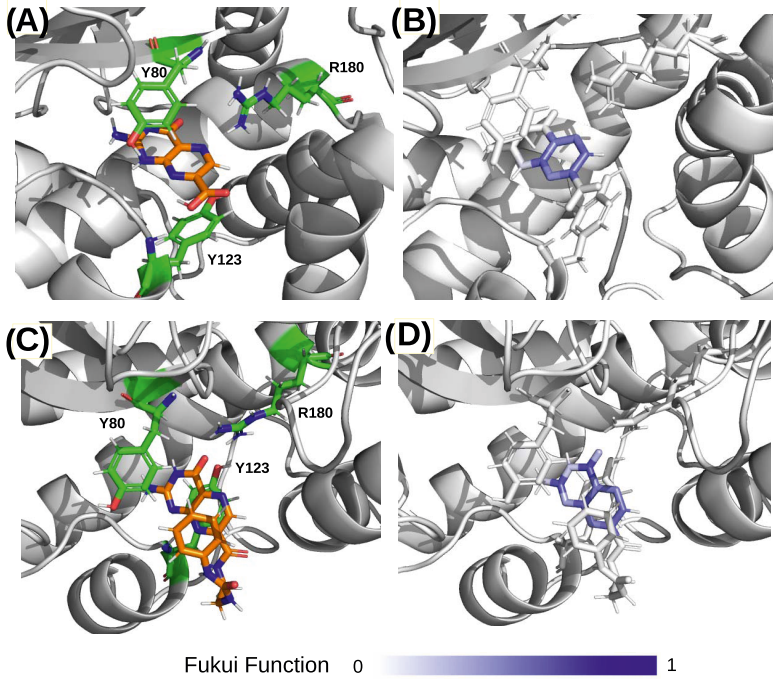


Figure 10. Calculated Fukui function for RTA complexes with the 1MX and JP2 ligands (orange). (A) Labels for the nearest residues (green) from 1MX; (B) Fukui function for RTA−1MX complex; (C) labels for the nearest residues (green) from JP2; and (D) local hardness for the RTA−JP2 complex.

5. REACTIVITY DESCRIPTORS

We calculated the reactivity descriptors for the protein−ligand complexes and used to perform a graphical analysis to theoretically characterize the most important interactions that occur between the inhibitors and the active site residues. As shown in Figure 8, the local hardness has higher values in the binding pocket for the RTA−19M complex, with its highest value at Arg-180 and Trp-211. For the RTA complex with the JP3 ligand, the electrostatic interactions are not placed within the ligand region. Significant Fukui function values were

https://doi.org/10.1021/acsomega.0c02588
ACS Omega 2021, 6, 8764−8777

computed in the residues of the binding pocket for the three most active ligand complexes and with small values for the JP2 ligand. Figure 9 depicts the Fukui function for the second and third most active cases. For the RTA−RS8 complex, the Fukui function localizes at one of the ligand rings of the pterin group and in Tyr-80. In the RTA−0RB complex, all atoms in the pterin group present high values of the given descriptor, as does Arg-180.

In Figure 10, the Fukui function is presented for 1MX and JP2, showing the distribution of the descriptor at the pterin group and none at the closest residues. The molecular structure of these ligands is based on the pterin group,[58] the same group present in adenine that is hydrolyzed in the ribosome by the action of RTA catalysis.[5] In the complexes considered by calculations, the reactivity in this group is always indicated by the Fukui function.

The models built from thermochemical properties correlations could rank the best inhibitors, but only reactivity descriptors can pose, which are the molecular structure features that new ligands must have to be more efficient.

6. CONCLUSIONS

The main objective of this work was to present a study where two computational approaches were applied (calculation of ΔH_{bind} and reactivity quantum chemical descriptors) to theoretically gain insights into the interactions between the RTA subunit of ricin and some of its inhibitors.

In our study, we observed that single-point energy calculations of ΔH_{bind} with the PM6-DH+, PM6-D3H4, and PM7 methods presented an excellent correlation with the IC_{50} data. In addition, although the PM7 method presented a slightly lower correlation than the PM6-D3H4 method, only the PM7 method was able to correctly rank all of the ligands analyzed. This result suggests that the PM7 method is more sensitive to small IC_{50} variations.

When comparing the IC_{50} data with the calculated ΔH_{bind} values obtained after full-atom optimization with the PM6-DH+, PM6-D3H4, and PM7 methods, we find that the correlation decreases significantly. Although the correlation was greatly reduced with the optimization of the structures, the PM6-D3H4 and PM7 methods were able to correctly rank the best ligand. This result suggests that if the molecular dynamics is not carried out properly to obtain an equilibrated structure, it is necessary to carry out a full-atom geometry optimization to obtain some correlation with the experimental data.

We also observed that in geometry optimization, the structure adopts a minimum local conformation that contributes less to the preferential position of the ligand at the active site. On the other hand, the representative structure from molecular dynamics represents the preferred position of the ligand in the active site. Thus, we conclude that for the dataset studied, it is preferable to use equilibrated MD structures to perform single-point energy calculations to obtain ΔH_{bind}.

ONIOM QM/MM single-point calculations with the B3LYP functional showed good correlation with IC_{50}; however, the methods PM6-DH+, PM6-D3H4, and PM7 showed better performance. When using the ωB97X-D functional that includes DFT-D2[73] dispersion correction, the QM/MM method presented R-value and performance similar to the PM7 method. Besides, ONIOM QM/MM calculations for ΔH_{bind} incurred a higher computational cost compared with semiempirical methods, about 250 times higher (B3LYP) and 1400 times higher (ωB97X-D).

In addition, we found for the cases studied, reactivity descriptors pointed out that both types of interactions, molecular overlap and electrostatic interactions, play an important role in the overall affinity of these ligands for the RTA binding pocket. From a relatively simple computational protocol, this approach opens the possibility to reveal additional information using structures treated with earlier simulations.

AUTHOR INFORMATION

Corresponding Author

Gerd Bruno Rocha − *Department of Chemistry, Federal University of Paraíba, João Pessoa, PB 58051-900, Brazil;* orcid.org/0000-0001-9805-9497; Phone: +55-83-3216-7437; Email: gbr@quimica.ufpb.br

Authors

Acassio Rocha-Santos − *Department of Chemistry, Federal University of Paraíba, João Pessoa, PB 58051-900, Brazil;* orcid.org/0000-0001-9928-9067

Elton José Ferreira Chaves − *Department of Biotechnology, Federal University of Paraíba, João Pessoa, PB 58051-900, Brazil*

Igor Barden Grillo − *Department of Chemistry, Federal University of Paraíba, João Pessoa, PB 58051-900, Brazil*

Amanara Souza de Freitas − *Department of Chemical Engineering, Federal University of Paraíba, João Pessoa, PB 58051-900, Brazil*

Demétrius Antônio Machado Araújo − *Department of Biotechnology, Federal University of Paraíba, João Pessoa, PB 58051-900, Brazil*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.0c02588

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support from the Brazilian agencies, institutes, and networks: Instituto Nacional de Ciência e Tecnologia de Nanotecnologia para Marcadores Integrados (INCT-INAMI), Conselho Nacional de Desenvolvimento Científico e Tecnolgico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES), Fundação de Apoio à Pesquisa do Estado da Paraíba (FAPESQ-PB), Programa de Apoio a Núcleos de Excelência (PRONEX-FACEPE), and Financiadora de Estudos e Projetos (FINEP). They also acknowledge the physical structure and computational support provided by Universidade Federal da Paraíba (UFPB), the computer resources of Centro Nacional de Processamento de Alto Desempenho em São Paulo (CENAPAD-SP), and Núcleo de Processamento de Alto Desempenho da Universidade Federal do Rio Grande do Norte (NPAD/UFRN). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES) through the research project Bioinformática Estrutural de Proteínas: Modelos, Algoritmos e Aplicaões Biotecnológicas (Edital Biologia Computacional 51/2013, processo AUXPE1375/2014 da CAPES). G.B.R. acknowledges support from the

https://doi.org/10.1021/acsomega.0c02588
ACS Omega 2021, 6, 8764−8777

Brazilian National Council for Scientific and Technological Development (CNPq grant no. 309761/2017-4).

REFERENCES

(1) Bozza, W. P.; Tolleson, W. H.; Rosado, L. A. R.; Zhang, B. Ricin detection: Tracking active toxin. *Biotechnol. Adv.* **2015**, *33*, 117−123.
(2) Montfort, W.; Villafranca, J. E.; Monzingo, A. F.; Ernst, S. R.; Katzin, B.; Rutenber, E.; Xuong, N. H.; Hamlin, R.; Robertus, J. D. The three-dimensional structure of ricin at 2.8 A. *J. Biol. Chem.* **1987**, *262*, 5398−5403.
(3) Rutenber, E.; Katzin, B. J.; Ernst, S.; Collins, E. J.; Mlsna, D.; Ready, M. P.; Robertus, J. D. Crystallographic refinement of ricin to 2.5 Å. *Proteins: Struct., Funct., Bioinf.* **1991**, *10*, 240−250.
(4) Robertus, J. D.; Monzingo, A. F. The structure of ribosome inactivating proteins. *Mini-Rev. Med. Chem.* **2004**, *4*, 477−486.
(5) Endo, Y.; Tsurugi, K. RNA N-glycosidase activity of ricin A-chain. Mechanism of action of the toxic lectin ricin on eukaryotic ribosomes. *J. Biol. Chem.* **1987**, *262*, 8128−8130.
(6) May, K. L.; Yan, Q.; Tumer, N. E. Targeting ricin to the ribosome. *Toxicon* **2013**, *69*, 143−151.
(7) Audi, J.; Belson, M.; Patel, M.; Schier, J.; Osterloh, J. Ricin poisoning: a comprehensive review. *JAMA, J. Am. Med. Assoc.* **2005**, *294*, 2342−2351.
(8) Argent, R. H.; Parrott, A. M.; Day, P. J.; Roberts, L. M.; Stockley, P. G.; Lord, J. M.; Radford, S. E. Ribosome-mediated folding of partially unfolded ricin A-chain. *J. Biol. Chem.* **2000**, *275*, 9263−9269.
(9) Simpson, J. C.; Roberts, L. M.; Römisch, K.; Davey, J.; Wolf, D. H.; Lord, J. M. Ricin A chain utilises the endoplasmic reticulum-associated protein degradation pathway to enter the cytosol of yeast. *FEBS Lett.* **1999**, *459*, 80−84.
(10) Ryde, U.; Soderhjelm, P. Ligand-binding affinity estimates supported by quantum-mechanical methods. *Chem. Rev.* **2016**, *116*, 5520−5566.
(11) Nikitina, E.; Sulimov, V.; Zayets, V.; Zaitseva, N. Semi-empirical calculations of binding enthalpy for protein-ligand complexes. *Int. J. Quantum Chem.* **2004**, *97*, 747−763.
(12) Chaves, E. J.; Padilha, I. Q.; Araújo, D. A.; Rocha, G. B. Determining the Relative Binding Affinity of Ricin Toxin A Inhibitors by Using Molecular Docking and Nonequilibrium Work. *J. Chem. Inf. Model.* **2018**, *58*, 1205−1213.
(13) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.
(14) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455−1474.
(15) Yilmazer, N. D.; Korth, M. Comparison of molecular mechanics, semi-empirical quantum mechanical, and density functional theory methods for scoring protein-ligand interactions. *J. Phys. Chem. B* **2013**, *117*, 8075−8084.
(16) Yilmazer, N. D.; Heitel, P.; Schwabe, T.; Korth, M. Benchmark of electronic structure methods for protein-ligand interactions based on high-level reference data. *J. Theor. Comput. Chem.* **2015**, *14*, No. 1540001.
(17) Fanfrlík, J.; Bronowska, A. K.; Řezáč, J.; Přenosil, O.; Konvalinka, J.; Hobza, P. A reliable docking/scoring scheme based on the semiempirical quantum mechanical PM6-DH2 method accurately covering dispersion and H-bonding: HIV-1 protease with 22 ligands. *J. Phys. Chem. B* **2010**, *114*, 12666−12678.
(18) Pecina, A.; Brynda, J.; Vrzal, L.; Gnanasekaran, R.; Hořejší, M.; Eyrilmez, S. M.; Řezáč, J.; Lepšík, M.; Řezáčová, P.; Hobza, P.; Majer, P.; Veverka, V.; Fanfrlík, J. Ranking Power of the SQM/COSMO Scoring Function on Carbonic Anhydrase II-Inhibitor Complexes. *ChemPhysChem* **2018**, *19*, 873−879.
(19) Mikulskis, P.; Genheden, S.; Wichmann, K.; Ryde, U. A semiempirical approach to ligand-binding affinities: Dependence on the Hamiltonian and corrections. *J. Comput. Chem.* **2012**, *33*, 1179−1189.
(20) González, R.; Suárez, C. F.; Bohórquez, H. J.; Patarroyo, M. A.; Patarroyo, M. E. Semi-empirical quantum evaluation of peptide-MHC class II binding. *Chem. Phys. Lett.* **2017**, *668*, 29−34.
(21) Kamel, K.; Kolinski, A. Assessment of the free binding energy of 1,25-dihydroxyvitamin D3 and its analogs with the human VDR receptor model. *Acta Biochim. Pol.* **2012**, *59*, 653−660.
(22) Wichapong, K.; Rohe, A.; Platzer, C.; Slynko, I.; Erdmann, F.; Schmidt, M.; Sippl, W. Application of docking and QM/MM-GBSA rescoring to screen for novel Myt1 kinase inhibitors. *J. Chem. Inf. Model.* **2014**, *54*, 881−893.
(23) Natesan, S.; Subramaniam, R.; Bergeron, C.; Balaz, S. Binding affinity prediction for ligands and receptors forming tautomers and ionization species: inhibition of mitogen-activated protein kinase-activated protein kinase 2 (MK2). *J. Med. Chem.* **2012**, *55*, 2035−2047.
(24) Alves, C. N.; Martí, S.; Castillo, R.; Andres, J.; Moliner, V.; Tunon, I.; Silla, E. A Quantum Mechanics/Molecular Mechanics Study of the Protein-Ligand Interaction for Inhibitors of HIV-1 Integrase. *Chem. − Eur. J.* **2007**, *13*, 7715−7724.
(25) Reddy, M. R.; Erion, M. D. Relative binding affinities of fructose-1, 6-bisphosphatase inhibitors calculated using a quantum mechanics-based free energy perturbation method. *J. Am. Chem. Soc.* **2007**, *129*, 9296−9297.
(26) Mikulskis, P.; Cioloboc, D.; Andrejić, M.; Khare, S.; Brorsson, J.; Genheden, S.; Mata, R. A.; Söderhjelm, P.; Ryde, U. Free-energy perturbation and quantum mechanical study of SAMPL4 octa-acid host-guest binding energies. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 375−400.
(27) Rathore, R.; Reddy, R. N.; Kondapi, A.; Reddanna, P.; Reddy, M. R. Use of quantum mechanics/molecular mechanics-based FEP method for calculating relative binding affinities of FBPase inhibitors for type-2 diabetes. *Theor. Chem. Acc.* **2012**, *131*, No. 1096.
(28) Świderek, K.; Martí, S.; Moliner, V. Theoretical studies of HIV-1 reverse transcriptase inhibition. *Phys. Chem. Chem. Phys.* **2012**, *14*, 12614−12624.
(29) Beierlein, F. R.; Michel, J.; Essex, J. W. A simple QM/MM approach for capturing polarization effects in protein- ligand binding free energy calculations. *J. Phys. Chem. B* **2011**, *115*, 4911−4926.
(30) Woods, C. J.; Shaw, K. E.; Mulholland, A. J. Combined quantum mechanics/molecular mechanics (QM/MM) simulations for protein-ligand complexes: free energies of binding of water molecules in influenza neuraminidase. *J. Phys. Chem. B* **2015**, *119*, 997−1001.
(31) Genheden, S.; Ryde, U.; Söderhjelm, P. Binding affinities by alchemical perturbation using QM/MM with a large QM system and polarizable MM model. *J. Comput. Chem.* **2015**, *36*, 2114−2124.
(32) Olsson, M. A.; Söderhjelm, P.; Ryde, U. Converging ligand-binding free energies obtained with free-energy perturbations at the quantum mechanical level. *J. Comput. Chem.* **2016**, *37*, 1589−1600.
(33) Cournia, Z.; Allen, B.; Sherman, W. Relative binding free energy calculations in drug discovery: recent advances and practical considerations. *J. Chem. Inf. Model.* **2017**, *57*, 2911−2937.
(34) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889−897.
(35) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.* **2011**, *51*, 69−82.
(36) Brandsdal, B. O.; Österberg, F.; Almlöf, M.; Feierberg, I.; Luzhkov, V. B.; Åqvist, J. Free Energy Calculations and Ligand Binding. *Adv. Protein Chem.* **2003**, *66*, 123−158.

https://doi.org/10.1021/acsomega.0c02588
ACS Omega 2021, 6, 8764−8777

(37) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **2009**, *113*, 6378−6396.
(38) Ribeiro, R. F.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Prediction of SAMPL2 aqueous solvation free energies and tautomeric ratios using the SM8, SM8AD, and SMD solvation models. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 317−333.
(39) Miyamoto, S.; Kollman, P. A. Absolute and relative binding free energy calculations of the interaction of biotin and its analogs with streptavidin using molecular dynamics/free energy perturbation approaches. *Proteins: Struct., Funct., Bioinf.* **1993**, *16*, 226−245.
(40) Mishra, S. K.; Koča, J. Assessing the Performance of MM/PBSA, MM/GBSA, and QM-MM/GBSA Approaches on Protein/Carbohydrate Complexes: Effect of Implicit Solvent Models, QM Methods, and Entropic Contributions. *J. Phys. Chem. B* **2018**, *122*, 8113−8121.
(41) Su, P.-C.; Tsai, C.-C.; Mehboob, S.; Hevener, K. E.; Johnson, M. E. Comparison of radii sets, entropy, QM methods, and sampling on MM-PBSA, MM-GBSA, and QM/MM-GBSA ligand binding energies of F. tularensis enoyl-ACP reductase (F abI). *J. Comput. Chem.* **2015**, *36*, 1859−1873.
(42) Khandelwal, A.; Lukacova, V.; Comez, D.; Kroll, D. M.; Raha, S.; Balaz, S. A combination of docking, QM/MM methods, and MD simulation for binding affinity estimation of metalloprotein ligands. *J. Med. Chem.* **2005**, *48*, 5437−5447.
(43) Xiang, M.; Lin, Y.; He, G.; Chen, L.; Yang, M.; Yang, S.; Mo, Y. Correlation between biological activity and binding energy in systems of integrin with cyclic RGD-containing binders: a QM/MM molecular dynamics study. *J. Mol. Model.* **2012**, *18*, 4917−4927.
(44) Cave-Ayland, C.; Skylaris, C.-K.; Essex, J. W. Direct validation of the single step classical to quantum free energy perturbation. *J. Phys. Chem. B* **2015**, *119*, 1017−1025.
(45) Rao, L.; Zhang, I. Y.; Guo, W.; Feng, L.; Meggers, E.; Xu, X. Nonfitting protein-ligand interaction scoring function based on first-principles theoretical chemistry methods: Development and application on kinase inhibitors. *J. Comput. Chem.* **2013**, *34*, 1636−1646.
(46) Fanfrlík, J.; Ruiz, F. X.; Kadlčíková, A.; Řezáč, J.; Cousido-Siah, A.; Mitschler, A.; Haldar, S.; Lepšík, M.; Kolář, M. H.; Majer, P.; Podjarny, A. D.; Hobza, P. The Effect of Halogen-to-Hydrogen Bond Substitution on Human Aldose Reductase Inhibition. *ACS Chem. Biol.* **2015**, *10*, 1637−1642.
(47) Blomberg, M. R.; Borowski, T.; Himo, F.; Liao, R.-Z.; Siegbahn, P. E. Quantum chemical studies of mechanisms for metalloenzymes. *Chem. Rev.* **2014**, *114*, 3601−3658.
(48) Siegbahn, P. E.; Himo, F. Recent developments of the quantum chemical cluster approach for modeling enzyme reactions. *JBIC, J. Biol. Inorg. Chem.* **2009**, *14*, 643−651.
(49) Sumowski, C. V.; Ochsenfeld, C. A convergence study of QM/MM isomerization energies with the selected size of the QM region for peptidic systems. *J. Phys. Chem. A* **2009**, *113*, 11734−11741.
(50) Hu, L.; Eliasson, J.; Heimdal, J.; Ryde, U. Do quantum mechanical energies calculated for small models of protein-active sites converge? *J. Phys. Chem. A* **2009**, *113*, 11793−11800.
(51) De Jong, W. A.; Bylaska, E.; Govind, N.; Janssen, C. L.; Kowalski, K.; Müller, T.; Nielsen, I. M.; Van Dam, H. J.; Veryazov, V.; Lindh, R. Utilizing high performance computing for chemistry: Parallel computational chemistry. *Phys. Chem. Chem. Phys.* **2010**, *12*, 6896−6920.
(52) Jia, W.; Wang, J.; Chi, X.; Wang, L. W. GPU implementation of the linear scaling three dimensional fragment method for large scale electronic structure calculations. *Comput. Phys. Commun.* **2017**, *211*, 8−15.
(53) Olson, M. A. A-chain structural determinant for binding substrate analogues: A molecular dynamics simulation analysis. *Proteins: Struct., Funct., Bioinf.* **1997**, *27*, 80−95.
(54) Olson, M. A.; Cuff, L. Free energy determinants of binding the rRNA substrate and small ligands to ricin A-chain. *Biophys. J.* **1999**, *76*, 28−39.
(55) Yan, X.; Day, P.; Hollis, T.; Monzingo, A. F.; Schelp, E.; Robertus, J. D.; Milne, G.; Wang, S. Recognition and interaction of small rings with the ricin A-chain binding site. *Proteins: Struct., Funct., Bioinf.* **1998**, *31*, 33−41.
(56) Saito, R.; Pruet, J. M.; Manzano, L. A.; Jasheway, K.; Monzingo, A. F.; Wiget, P. A.; Kamat, I.; Anslyn, E. V.; Robertus, J. D. Peptide-conjugated pterins as inhibitors of ricin toxin A. *J. Med. Chem.* **2013**, *56*, 320−329.
(57) Pruet, J. M.; Saito, R.; Manzano, L. A.; Jasheway, K. R.; Wiget, P. A.; Kamat, I.; Anslyn, E. V.; Robertus, J. D. Optimized 5-membered heterocycle-linked pterins for the inhibition of Ricin Toxin A. *ACS Med. Chem. Lett.* **2012**, *3*, 588−591.
(58) Wiget, P. A.; Manzano, L. A.; Pruet, J. M.; Gao, G.; Saito, R.; Monzingo, A. F.; Jasheway, K. R.; Robertus, J. D.; Anslyn, E. V. Sulfur incorporation generally improves Ricin inhibition in pterin-appended glycine-phenylalanine dipeptide mimics. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 6799−6804.
(59) Pruet, J. M.; Jasheway, K. R.; Manzano, L. A.; Bai, Y.; Anslyn, E. V.; Robertus, J. D. 7-Substituted pterins provide a new direction for ricin A chain inhibitors. *Eur. J. Med. Chem.* **2011**, *46*, 3608−3615.
(60) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. Rm1: A reparameterization of am1 for h, c, n, o, p, s, f, cl, br, and i. *J. Comput. Chem.* **2006**, *27*, 1101−1111.
(61) Stewart, J. J. Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173−1213.
(62) Korth, M. Third-generation hydrogen-bonding corrections for semiempirical QM methods and force fields. *J. Chem. Theory Comput.* **2010**, *6*, 3808−3816.
(63) Stewart, J. J. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **2013**, *19*, 1−32.
(64) Stewart, J. J. Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. *Int. J. Quantum Chem.* **1996**, *58*, 133−146.
(65) Stewart, J. J. MOPAC: a semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−103.
(66) Dapprich, S.; Komáromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. A new ONIOM implementation in Gaussian98. Part I. The calculation of energies, gradients, vibrational frequencies and electric field derivatives. *J. Mol. Struct.: THEOCHEM* **1999**, *461−462*, 1−21.
(67) Chung, L. W.; Sameera, W. M.; Ramozzi, R.; Page, A. J.; Hatanaka, M.; Petrova, G. P.; Harris, T. V.; Li, X.; Ke, Z.; Liu, F.; Li, H. B.; Ding, L.; Morokuma, K. The ONIOM Method and Its Applications. *Chem. Rev.* **2015**, *115*, 5678−5796.
(68) Chagas, M. A.; Pereira, E. S.; Godinho, M. P.; Da Silva, J. C. S.; Rocha, W. R. Base Mechanism to the Hydrolysis of Phosphate Triester Promoted by the Cd2+/Cd2+ Active site of Phospho-triesterase: A Computational Study. *Inorg. Chem.* **2018**, *57*, 5888−5902.
(69) Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. et al. *Gaussian 09*, revision D.01; Gaussian, Inc.: Wallingford, CT, 2009.
(70) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785.
(71) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652.
(72) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615−6620.
(73) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **2006**, *27*, 1787−1799.

https://doi.org/10.1021/acsomega.0c02588
ACS Omega 2021, 6, 8764−8777

(74) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self-consistent molecular orbital methods. XII. Further extensions of Gaussian-type basis sets for use in molecular orbital studies of organic molecules. *J. Chem. Phys.* **1972**, *56*, 2257−2261.
(75) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.* **1980**, *72*, 650−654.
(76) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *J. Chem. Phys.* **1982**, *77*, 3654−3665.
(77) Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.
(78) Grela, P.; Szajwaj, M.; Horbowicz-Drozdzal, P.; Tchórzewski, M. How ricin damages the ribosome. *Toxins* **2019**, *11*, No. 241.
(79) Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual density functional theory. *Chem. Rev.* **2003**, *103*, 1793−1873.
(80) Pearson, R. G. Recent advances in the concept of hard and soft acids and bases. *J. Chem. Educ.* **1987**, *64*, No. 561.
(81) Torrent-Sucarrat, M.; De Proft, F.; Geerlings, P.; Ayers, P. W. Do the local softness and hardness indicate the softest and hardest regions of a molecule? *Chem. − Eur. J.* **2008**, *14*, 8652−8660.
(82) Sánchez-Márquez, J.; Zorrilla, D.; García, V.; Fernández, M. Introducing a new methodology for the calculation of local philicity and multiphilic descriptor: an alternative to the finite difference approximation. *Mol. Phys.* **2018**, *116*, 1737−1748.
(83) Sarkar, U.; Roy, D.; Chattaraj, P.; Parthasarathi, R.; Padmanabhan, J.; Subramanian, V. A conceptual DFT approach towards analysing toxicity. *J. Chem. Sci.* **2005**, *117*, 599−612.
(84) Grillo, I. B.; Urquiza-Carvalho, G. A.; Chaves, E. J. F.; Rocha, G. B. Semiempirical methods do Fukui functions: Unlocking a modeling framework for biosystems. *J. Comput. Chem.* **2020**, 862.
(85) Merz, K. M.; Faver, J. Utility of the hard/soft acid-base principle via the fukui function in biological systems. *J. Chem. Theory Comput.* **2010**, *6*, 548−559.
(86) Grillo, I. B.; Urquiza-Carvalho, G.; Bachega, J. F. R.; Rocha, G. B. Elucidating Enzymatic Catalysis using Fast Quantum Chemical Descriptors. *J. Chem. Inf. Model.* **2020**, 578.
(87) Grillo, I. B.; Bachega, J. F. R.; Timmers, L. F. S.; Caceres, R. A.; de Souza, O. N.; Field, M. J.; Rocha, G. B. Theoretical characterization of the shikimate 5-dehydrogenase reaction from Mycobacterium tuberculosis by hybrid QC/MM simulations and quantum chemical descriptors. *J. Mol. Model.* **2020**, *26*, No. 297.
(88) Fukushima, K.; Wada, M.; Sakurai, M. An insight into the general relationship between the three dimensional structures of enzymes and their electronic wave functions: Implication for the prediction of functional sites of enzymes. *Proteins: Struct., Funct., Bioinf.* **2008**, *71*, 1940−1954.
(89) Torrent-Sucarrat, M.; De Proft, F.; Ayers, P. W.; Geerlings, P. On the applicability of local softness and hardness. *Phys. Chem. Chem. Phys.* **2010**, *12*, 1072−1080.
(90) Grillo, I. B.; Urquiza-Carvalho, G. A.; Rocha, G. B. PRIMoRDiA: A Software to Explore Reactivity and Electronic Structure in Large Biomolecules. *J. Chem. Inf. Model.* **2020**, 5885.
(91) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **2016**, *44*, D1045−D1053.
(92) Yilmazer, N. D.; Korth, M. Enhanced semiempirical QM methods for biomolecular interactions. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 169−175.
(93) Sulimov, A. V.; Kutov, D. C.; Katkova, E. V.; Sulimov, V. B. Combined docking with classical force field and quantum chemical semiempirical method PM7. *Adv. Bioinf.* **2017**, *2017*, No. 7167691.
(94) Carra, J. H.; McHugh, C. A.; Mulligan, S.; Machiesky, L. A. M.; Soares, A. S.; Millard, C. B. Fragment-based identification of determinants of conformational and spectroscopic change at the ricin active site. *BMC Struct. Biol.* **2007**, *7*, No. 72.
(95) Gupta, A.; Chaudhary, N.; Kakularam, K. R.; Pallu, R.; Polamarasetty, A. The augmenting effects of desolvation and conformational energy terms on the predictions of docking programs against mPGES-1. *PLoS One* **2015**, *10*, No. e0134472.
(96) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.

https://doi.org/10.1021/acsomega.0c02588
ACS Omega 2021, 6, 8764−8777