Estudo Computacional do Nelfinavir incorporado na HIV-1 Protease

António Cesário

Faculdade de Ciências da Universidade do Porto
Mestrado em Física Experimental
Bioquímica Computacional

23 de janeiro de 2022

Resumo

Neste trabalho foram exploradas os primeiros passos com rumo à simulação de sistemas biológicos, tendo sido abordada teoria, trabalho computacional e vários métodos de análise da dados.

A incorporação do Nelfinavir na HIV-1 protease foi o nosso objeto de estudo e foi realizada uma pequena simulação de 120 ps de Dinâmica Molecular, que se mostrou termodinamicamente estável, embora não se tenha conseguido chegar a uma estrutura convergida.

A análise de dados foi bastante diversificada, permitindo-me aprender maneiras diferentes de extrair vários tipos de informação das estruturas obtidas, desde comprimentos e ângulos de ligação das pontes de hidrogénio do sistema às funções de distribuição radial de moléculas de água de certos átomos em resíduos de referência.

1 Introdução

1.1 Inibidores da HIV-1 protease

A Protease do HIV-1 é uma enzima responsável por um passo essencial no ciclo de vida do vírus HIV. Para que as proteínas do HIV-1 desempenhem as suas funções, é necessário que elas sejam clivadas em pontos específicos. Dada a especificidade deste processo, foram procuradas formas de interferir com ele, afim de interromper o ciclo de vida do vírus.

Neste relatório foi abordado um dos métodos desenvolvidos para concretizar esta ideia, incorporando o inibidor *Nelfinavir* na estrutura da protease.

Em primeiro irei apresentar algum fundamento teórico para explicar a física subjacente aos cálculos computacionais. Depois mostro como representar a estrutura estudada e explico como foram executadas as simulações computacionais. Por último irei apresentar os resultados das simulações e explorar vários métodos de análise dos dados obtidos.

2 Fundamento Teórico

A base para a entendermos como se comportam estes sistemas biológicos está em estabelecer modelos computacionais que nos permitam simular propriedades de interesse. Nesta secção apresento alguns destes métodos e refiro ideias que encontrei durante a minha pesquisa e tenha achado interessantes.

Os princípios físicos utilizados dependem de modelo para modelo, sendo os mais importantes derivados de Mecânica Molecular, Dinâmica Molecular e Mecânica Quântica Molecular. Como o sistema estudado possui centenas de átomos, as simulações baseadas em Mecânica Quântica não são passíveis de se concluírem em tempo útil, pelo que neste trabalho apenas foram estudados apenas os primeiros dois métodos computacionais.

2.1 Mecânica Molecular

O método mais comum para descrever a energia de um sistema biológico serve-se de potenciais empíricos derivados da Mecânica Molecular. Se considerarmos um sistema de N átomos, a sua energia potencial, $U(r_1, r_2, \ldots, r_N)$, pode ser descomposta como a soma de de interações a uma partícula, duas partículas, e assim sucessivamente:

$$U(r_1, r_2, \ldots, r_N) = \sum_{i=1}^{N} U(r_i) + \sum_{i=1}^{N}\sum_{j>i}^{N} U(r_i, r_j) + \sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{k>j>i}^{N} U(r_i, r_j, r_k) + \ldots \tag{1}$$

onde r_i representa o vetor posição do átomo i. Aproximações são então impostas para simplificar o problema. O termo de primeira ordem, $U(r_i)$, está associado às interações com campos externos ao sistema, pelo que normalmente não é considerado.

Os termos de segunda e terceira ordem, por outro lado, contêm as interações mais energéticas do sistema, pelo que estes são os mais estudados, enquanto que os de quarta ordem e superior são desprezados. No entanto, como as interações a três partículas são muito dispendiosas a nível computacional,

foram encontradas maneiras de omitir estes termos reajustando os campos de forças e os parâmetros que governam as interações a duas partículas.

O problema inicial foi então reduzido a entender como é que o sistema se comporta face ao potencial:

$$U(r_1, r_2, \ldots, r_N) \approx \sum_{i=1}^{N} \sum_{j>i}^{N} U(r_i, r_j) \tag{2}$$

onde cada termo pode ser decomposto em vários potenciais empíricos, caracterizando interações diferentes, resumidas abaixo e esquematizadas na Fig. 1.

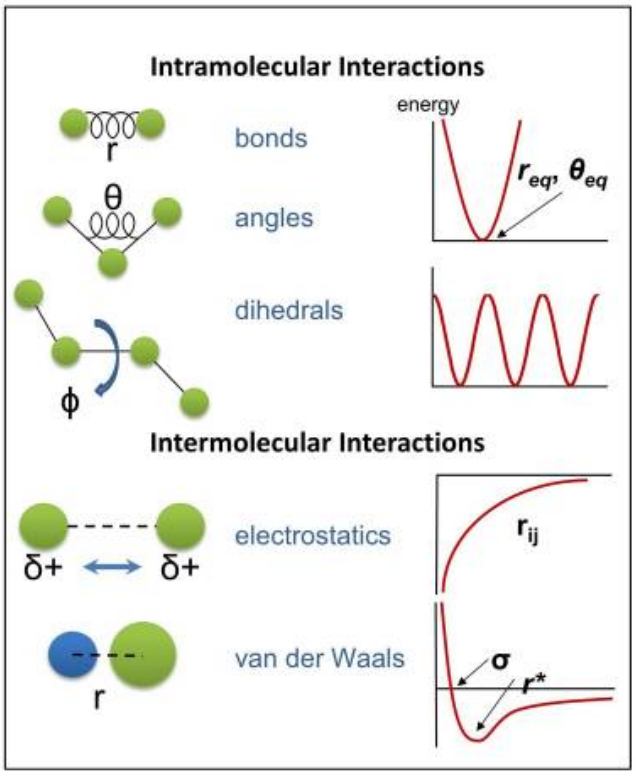


Figura 1: Resumo da origem das interações mais importantes para o termo potencial da Eq. (2). À direita encontram-se as formas genéricas dos respetivos potenciais. Imagem de [1].

2.1.1 Interações Intramoleculares

As interações intramoleculares descrevem a estrutura covalente de uma molécula, de onde se distinguem a energia de ligação, os ângulos de ligação, e os diedros/torção.

A energia de ligação entre dois átomos é um termo harmónico, resultante da expansão do potencial num polinómio de Taylor, até segunda ordem, em função da distância entre as duas partículas, r. No entanto, esta aproximação não prevê corretamente a energia das ligações para desvios significativos da posição de equilíbrio, como esquematizado na Fig. 2. Para contornar este problema podem ser incluídos termos de maior ordem do polinómio de Taylor.

Analogamente, a energia do ângulo de ligação também apresenta um comportamento harmónico, com origem nas (pequenas) oscilações do ângulo que dois átomos fazem com um terceiro, θ.

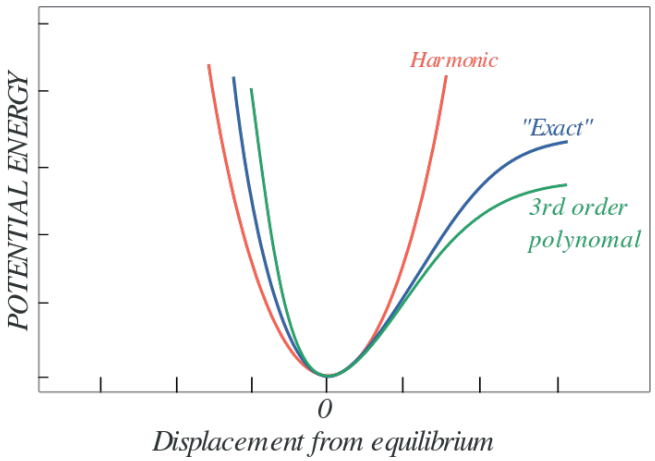


Figura 2: Validade das aproximações de primeira e segunda ordem em r e em θ na determinação da energia de ligação entre duas partículas. Imagem de [2].

A energia de torção tem origem na rotação das ligações e varia sinusoidalmente com o ângulo de torção, ϕ. As interações intramoleculares podem, portanto, ser escritas na forma:

$$U_{\text{intra}}\left(r_i, r_j\right) = U_{\text{intra}}\left(r, \theta, \phi\right) = K_{\text{lig}}\left(r - r_{eq}\right)^2 + K_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_n \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \quad (3)$$

onde r_{eq}, θ_{eq}, e γ são parâmetros de equilíbrio da geometria, e os parâmetros K_{lig}, K_θ e V_n são constantes de força, que caracterizam a dependência da energia face os desvios à geometria de equilíbrio.

O somatório em n no termo de torção diz-nos que podemos ter vários ângulos de rotação estáveis, onde o n está associado à periodicidade/simetria do sistema, que determina o de mínimos de energia contidos entre 0 e 2π.

As interações intramoleculares entre duas partículas do sistema são, portanto, caracterizadas pelo conjunto de parâmetros referidos, tipicamente sob a assunção de que o campo de forças é transferível de vizinhança para vizinhança, ou seja: uma ligação C-C num propano deve de ser semelhante a C-C em butano. Porém, esta transferibilidade nem sempre é possível, pelo que existem campos de forças que só podem ser aplicados num conjunto restrito de sistemas. Também existem casos em que as ligações são especificadas tendo em conta um terceiro átomo, por exemplo: um conjunto de parâmetros se em C-C um dos C's estiver ligado a um H e outro conjunto se esse C estiver ligado a um S.

Os comprimentos e ângulos de ligação são determinados experimentalmente por cristalografia (como difração de neutrões, ou outras técnicas de espectroscopia), com constantes de força inferidas a partir de outras experiências de micro-ondas ou infravermelhos, por exemplo. Os termos de torção, por outro lado, são muito mais difíceis de determinar experimentalmente, por envolverem a interação de quatro partículas, embora a ligação que estejamos a analisar seja só entre duas. Estes parâmetros costumam ser os últimos a ser determinados, obtidos pela conjugação de dados experimentais, por exemplo através de Ressonância Magnética Nuclear, e cálculos de Mecânica Quântica [1, 3, 4].

2.1.2 Interações Intermoleculares

As interações intermoleculares envolvem forças de maior alcance. Quando dois átomos i e j com cargas parciais q_i e q_j, respetivamente, se encontram a uma distância r_{ij}, a sua energia potencial eletroestática é dada pela Lei de Coulomb:

$$U_C\left(r_i, r_j\right) = \frac{q_i q_j}{4\pi\varepsilon}\frac{1}{r_{ij}} \tag{4}$$

onde ε é a permitividade elétrica do meio. Para ter uma descrição mais completa da energia do sistema, é introduzido o potencial de Lennard-Jones:

$$U_{\mathrm{LJ}} = \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \tag{5}$$

onde o primeiro termo representa a repulsão das nuvens eletrónicas de i e j, e o segundo é um termo de atração resultante de forças de dispersão de eletrões. Para simular este tipo de interações também é necessário determinar os parâmetros q_i, q_j, A_{ij}, B_{ij} [4].

O problema de Mecânica Molecular da Eq. (1) foi, portanto, reduzido à minimização do potencial:

$$\begin{gathered} U = \sum_{\text{bonds}} K_{\text{lig}}^{ij}\left(r_{ij} - r_{eq}^{ij}\right)^2 + \sum_{\text{angles}} K_{\theta}^{ij}\left(\theta_{ij} - \theta_{eq}^{ij}\right)^2 + \\ \sum_{\text{dihedrals}}\sum_{n}\frac{V_n^{ij}}{2}\left[1+\cos\left(n\phi_{ij} - \gamma^{ij}\right)\right] + \sum_{\text{non-bond}}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right] \end{gathered} \tag{6}$$

que corresponde à forma como estão parametrizados os campos de força AMBER, que possuem informação sobre milhares de combinações possíveis de parâmetros, que são introduzidos nesta equação para determinar, por exemplo, a configuração de distâncias, ângulos e torções intramoleculares que minimizam a energia de uma molécula. Mais informações sobre estes campos de força podem ser encontradas na página web e no seu manual de 2021.

Existem, no entanto, várias desvantagens nestes métodos. Para além de termos que determinar e ricamente uma grande parte dos campos de forças, a Mecânica Molecular não nos permite obter com confiança a conformação de menor energia de um sistema. Isto acontece devido à elevada complexidade da superfície de potencial e o grande número de graus de liberdade. Assim, não é possível determinar o mínimo global do potencial sem uma amostragem exaustiva do imenso espaço das configurações possíveis. É neste contexto que surgem outros métodos de simulação alternativos/-complementares como a Dinâmica Molecular [1, 5] ou até mesmo simulações de Monte-Carlo [6].

2.2 Dinâmica Molecular

As simulações de dinâmica molecular baseiam-se na integração das equações clássicas de movimento, as equações de Newton, para todos os átomos no sistema. A simulação inicia-se com a atribuição estocástica de velocidades às N partículas, baseada na distribuição de Boltzmann para uma dada

temperatura T tal que:

$$\frac{1}{2}\sum_{i}^{N} m_i v_i^2 = \frac{3Nk_{\mathrm{B}}T}{2} \tag{7}$$

Seja U_{ij} o potencial entre os átomos i e j como vimos na Sec. (2.1), a força que j exerce sobre i é dada por

$$F_i = -\nabla_i \sum_{j}^{N} U_{ij} \tag{8}$$

logo, pela segunda lei de Newton:

$$m_i \frac{d^2 r_i}{dt^2} = -\nabla_i \sum_{j}^{N} U_{ij} \tag{9}$$

A dinâmica é então propagada pela integração numérica da Eq. (9), de onde se extrairá a velocidade (integrar 1x) e a posição (integrar 2x) de cada uma das partículas do sistema. Os métodos numéricos tipicamente implementados são o das velocidades de Verlet e o *leapfrog* [7]. Este último algoritmo foi o utilizado nas nossas simulações, onde a cada iteração são calculadas as forças, posições e velocidades de cada átomo, pelo que é necessário escolher um bom compromisso entre tempo de computação e precisão, atendendo sempre ao facto de que a escala temporal do nosso passo de integração tem que ser menor que o fenómeno mais rápido que queremos estudar. Os cálculos das forças podem ser muito exigentes, pelo que certos algoritmos podem optar por simplificar U_{ij} relativamente à Mecânica Molecular. Também é conveniente distinguir que os campos de força mencionados anteriormente na Sec. (2.1) estão otimizados para a estrutura e não para dinâmica.

A estabilidade das simulações está intimamente ligada ao passo de integração, que por sua vez está associado à frequência esperada da dinâmica. As ligações que envolvem hidrogénio são muito rápidas, sugerindo-nos passos de integração extremamente pequenos, < 1 fs, para ter uma boa amostragem da dinâmica. Felizmente, o algoritmo SHAKE [8] permite-nos restringir graus de liberdade dessas ligações tal que seja possível utilizar tempos de integração maiores.

Uma limitação significativa da DM é o tempo de simulação necessário para o sistema ultrapassar grandes barreiras de potencial. À temperatura ambiente, barreiras de 1 e 5 kcal/mol conseguem ser ultrapassadas em ps e ns de simulação, respetivamente, enquanto que para 10 kcal/mol já são necessárias simulações da ordem dos μs [1]. Isto impede as simulações de concluírem uma amostragem completa do espaço de estados em tempo útil de computação.

2.3 RMSD - análise da convergência da estrutura

Depois de todos estes métodos de simulação, precisamos de ter uma forma quantitativa de avaliar a convergência dos resultados. Como as posições dos átomos variam de uma iteração para a outra, podemos obter os desvios sofridos por cada átomo em *snapshots* diferentes do sistema e representar a média destes desvios. A isto dá-se o nome de "Root Mean Square Deviation" (RMSD) e tem a forma:

$$RMSD(v,w) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left((v_{ix}-w_{ix})^2 + (v_{iy}-w_{iy})^2 + (v_{iz}-w_{iz})^2\right)} \tag{10}$$

onde v_i e w_i são as posições do átomo i nos *snapshots* v e w, respetivamente.

Estando estabelecidos os princípios teóricos necessários para compreender as bases por detrás dos cálculos computacionais, estamos agora em condições de explorar o trabalho computacional e de simulação.

3 Preparar e Visualizar a Estrutura

A estrutura inicial foi obtida no website de RCSB Protein Data Bank, onde se encontra identificada pelo nome *1ohr*, que pode visitar diretamente através deste link.

O primeiro passo foi procurar desvendar a informação contida em `1ohr.pdb`. Ao abrir o ficheiro com um editor de texto, podemos encontrar a descrição do sistema em estudo estruturada como na Fig. 3, onde está expressa a informação relativa a cada átomo que constitui o sistema. Identificamos pelo menos duas cadeias, A e B, divididas no ficheiro pela linha inciada por `TER`. Conseguimos também observar a composição e distribuição dos elementos do resíduo PRO da cadeia B, identificado como o seu primeiro resíduo. Com base nestas coordenadas e associações a resíduos, cadeias, entre outros, será possível estabelecer representações gráficas do sistema com o auxílio do *free software* VMD (Visual Molecular Dynamics) [9].

```
ATOM    863  CZ  PHE A  99      13.741  10.524  26.807  1.00 27.49           C
ATOM    864  OXT PHE A  99      11.666  12.580  31.808  1.00 29.65           O
ATOM    865  H   PHE A  99      10.270  11.707  30.465  1.00  0.00           H
TER     866      PHE A  99
ATOM    867  N   PRO B   1      12.544  14.339  30.054  1.00 32.33           N
ATOM    868  CA  PRO B   1      11.746  15.278  29.223  1.00 27.77           C
ATOM    869  C   PRO B   1      10.368  14.698  29.023  1.00 25.91           C
ATOM    870  O   PRO B   1      10.122  13.581  29.466  1.00 22.46           O
ATOM    871  CB  PRO B   1      12.446  15.447  27.860  1.00 26.34           C
ATOM    872  CG  PRO B   1      13.226  14.133  27.786  1.00 28.43           C
ATOM    873  CD  PRO B   1      13.707  13.944  29.231  1.00 29.49           C
ATOM    874  H2  PRO B   1      12.792  14.664  31.014  1.00  0.00           H
ATOM    875  H3  PRO B   1      11.955  13.488  30.267  1.00  0.00           H
ATOM    876  N   GLN B   2       9.472  15.451  28.403  1.00 26.18           N
ATOM    877  CA  GLN B   2       8.104  14.996  28.126  1.00 26.53           C
ATOM    878  C   GLN B   2       7.987  15.084  26.620  1.00 28.74           C
ATOM    879  O   GLN B   2       8.193  16.181  26.094  1.00 30.52           O
ATOM    880  CB  GLN B   2       7.090  15.931  28.761  1.00 26.02           C
```

Figura 3: Secção da informação contida em 1ohr.pdb. As colunas correspondem a: (1) Identificação do tipo de campo; (2,3) número e nome do átomo no sistema; (4,6) nome e número do resíduo; (5) cadeia; (7-9) coordenadas (x,y,z); (10) fator de ocupação; (11) fator temperatura; (12) elemento químico. Podemos observar a transição da cadeia A para a B e os dados referentes ao primeiro resíduo (PRO) da cadeia B.

Na Fig. 4 temos representada a estrutura inicial, onde estão destacadas as cadeias (A e B), as flaps das cadeias, o inibidor e os asparatos catalíticos.

3.1 Preparação do Ficheiro PDB

Embora tenhamos uma representação do sistema que queremos estudar, esta serve apenas como uma aproximação incical à sua estrutura. Isto deve-se ao facto destes dados terem sido obtidos experimentalmente, possivelmente em condições diferentes daquelas em que estamos interessados. É muito

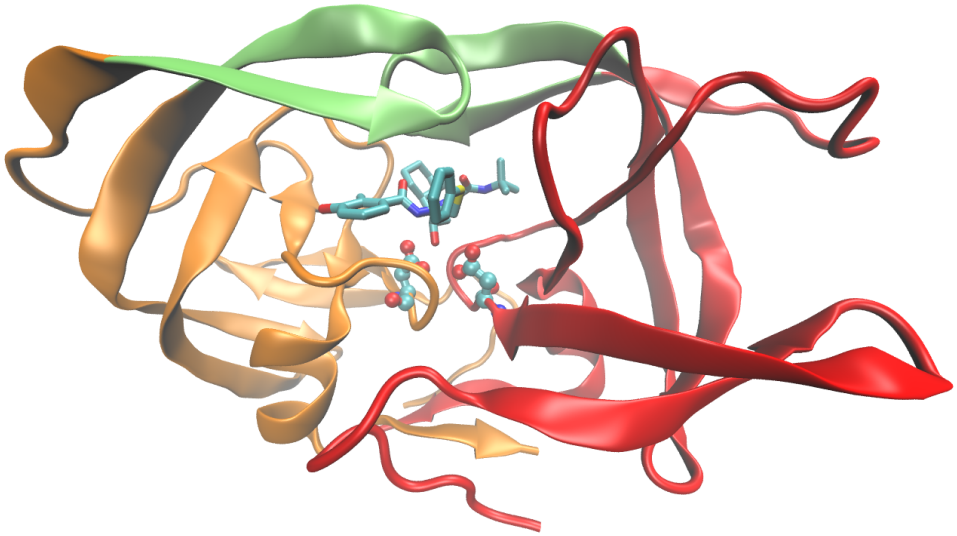


Figura 4: Ilustração da estrutura inicial. A proteína está representada no modo *NewCartoon* e dividida nas cadeias A (laranja + verde) e B (vermelho + verde), cujas *flaps* foram destacadas a verde. Abaixo das *flaps* observamos o inibidor *Nelfinavir*. Os resíduos ASP25 das das duas cadeias também estão visíveis abaixo do inibidor e ligadas às respetivas cadeias.

importante levar isto em conta porque podemos querer simular meios e/ou condições de pressão-temperatura específicos, alterando comprimentos ou ângulos de ligação e, possivelmente, quantidades termodinâmicas em que estamos interessados.

3.1.1 Remover Hidrogénios

Como a estrutura da Fig. 4 foi obtida por cristalografia de raios-x, uma técnica que tem origem no *scattering* de raios-x na nuvem eletrónica, e o hidrogénio possui apenas um eletrão de valência, é provável que as posições detetadas para estes átomos sejam pouco precisas ou que até não tenham sido detetados todos os hidrogénios da estrutura [10]. O próximo passo foi, portanto, remover todas as linhas em `1ohr.pdb` onde se encontram átomos de H, resultando no ficheiro `1ohr-no-H.pdb`. Estes átomos serão posteriormente adicionados pelo AMBER.

3.1.2 Identificar Interações no Inibidor

Para lidar com elementos estruturais diferentes, distinguir por exemplo uma prolina (PRO) de uma glutamina (GLN), o AMBER recorre a bibliotecas que possuem modelos destas estruturas básicas. No entanto, o inibidor Nelfinavir, que designaremos por 1UN, não está presente nestas bibliotecas, pelo que teremos que o adicionar ao AMBER antes de prosseguir com as simulações. A criação dos *scripts* adequados é um processo complexo, pelo que os professores simplesmente nos forneceram os ficheiros necessários já preparados: `1UN.mod` e `1UN.prepc`.

3.1.3 ASP25 Protonado

Como é referido em [11–13], o sistema constituído pela protease e o inibidor só é estável se um dos dois ácidos aspárticos (ASP25) for protonado. Para o AMBER interpretar um deles como tal, alterou-se `ASP` (desprotonado) para `ASH` (protonado) na cadeia A.

4 Simulações Computacionais

A concretização destas simulações passa pela elaboração de *scripts* escritos especificamente para as várias funções de AMBER utilizadas. Para facilitar a partilha do código utilizado e o apresentar de uma maneira conveniente, disponibilizei todo o trabalho computacional num repositório de GitHub, permitindo também a reprodução dos resultados obtidos.

Antes de submeter o sistema às simulações de dinâmica, ainda é necessário neutralizá-lo, adicionar um solvente, e proceder à minimização da sua energia por mecânica molecular. Estes cálculos foram realizados através dos pacotes `xleap` e `sander` do AMBER-12, cujo manual pode ser encontrado neste link. Foram usados os campos de forças AMBER FF03 [14, 15], para simular certos parâmetros dos resíduos, e o GAFF (*General Amber Force Field*), que nos permite estudar um grande número de moléculas orgânicas compostas por C, N, O, H, S, P, F, Cl, Br e I. Os ficheiros referidos na Sec.(3.1.2) também tiveram que ser dados como *input*.

Num à parte, e porque queria ter uma noção do tipo de informação necessária para construir um campo de forças eficiente, encontrei um repositório no GitHub com os ficheiros leaprc.ff03 e leaprc.gaff, semelhantes aos que foram carregados pelo `xleap`. Estes dependem, por sua vez, de outros mais extensos como gaff.dat e all_aminoct03.lib, respetivamente, que possuem uma longa lista de dados que caracterizam as interações/ligações entre vários átomos/estruturas. Todos estes ficheiros têm que ser lidos pelo AMBER, e encontram-se algures nos sub-diretórios da variável local `$AMBERHOME`. Muita informação útil também pode ser encontrada na página web e no manual de 2021 do AMBER, que já foram referidos na Sec. (2.1).

4.1 Adição de um Solvente

O sistema atual encontra-se em vácuo, pelo que precisamos de simular uma situação mais próxima daquelas em que a molécula se encontra na realidade, adicionando, por exemplo, uma vizinhança constituída por água.

Antes de adicionar a água é necessário neutralizar o sistema. Com o `xleap` adicionaram-se iões de Cl^- e/ou Na^+, conforme determinado pelo programa, e logo depois, introduziu-se no sistema uma caixa de águas (explícitas) com dimensões tais que as distâncias do exterior da proteína às paredes da caixa sejam no mínimo 12Å. Os valores tipicamente usados para as dimensões da caixa encontra-se entre $8 - 12$Å, quando condições fronteira periódicas são aplicadas aos cálculos [5], que serão incluídas a seguir na mecânica e dinâmica moleculares. O valor escolhido é um *trade-off* entre precisão e custo computacional, que aumentam com as dimensões da caixa. Deste processo resultam

os ficheiros seguintes: `prot-nelf.pdb`, `prot-nelf.prmtop`, e `prot-nelf.inpcrd`. O primeiro é simplesmente o `.pdb` anterior mas com a caixa de águas e os iões neutralizadores. No prmtop encontram-se os parâmetros do campo de forças e a topologia do sistema. Esta informação permanecerá inalterada ao longo das simulações. Por último, `inpcrd` possui as coordenadas dos vários átomos e até das velocidades, quando pedido para dar *print* nos cálculos de dinâmica molecular. Ao longo das simulações estas posições irão evoluir, cada uma resultando num novo conjunto de coordenadas.

4.2 Mecânica Molecular - Minimização da Energia

A introdução de certa forma grosseira das águas e dos iões pelo `xleap` coloca o sistema num estado de elevada energia e com as ligações entre os átomos muito pouco otimizadas, pelo que precisamos de fazer as águas e os iões reagir à presença da molécula molécula. Se iniciássemos neste ponto a dinâmica molecular, os ângulos e distâncias interatómicas resultariam em forças irrealisticamente elevadas em vários átomos. Para as simulações de dinâmica, esta condição inicial faria com que esses elementos se deslocassem significativamente da sua posição na estrutura inicial, tendo energia suficiente para se aproximar significativamente de outros átomos, transmitindo-os energia para estes também se deslocaram consideravelmente da sua posição inicial. Este efeito em cadeia resultaria em conformações finais instáveis e incorretas [1].

Aqui entra o processo de minimização de energia por mecânica molecular, utilizado para aproximar o sistema atual do seu estado fundamental. Este método não tem a capacidade de determinar com precisão a conformação convergida, mas idealmente irá aproximá-la o suficiente para a podermos alcançar por dinâmica molecular [1, 5].

O sistema passou por três processos de minimização: no primeiro relaxou-se as posições e geometria das moléculas de água; em segundo foram otimizadas as posições dos hidrogénios; em terceiro relaxou-se a posição de todos os átomos.

Cada um destes processos começou com 250 iterações de *steepest descent* e acabou com 250 de um algoritmo de gradiente conjugado. Foram admitidas condições fronteira periódicas e usou-se um raio de corte de 10Å para interações não ligantes, ou seja, entre moléculas distintas.

4.3 Dinâmica Molecular - Equilibração e Produção

As simulações de dinâmica molecular foram divididas em duas partes: equilibração e produção. Na equilibração, correspondendo ao ensemble canónico, o sistema foi aquecido a volume constante desde 0 K até à temperatura de produção de 310.15 K (37 °C) ao longo de 20 ps com intervalo de integração de 1 fs. Para controlo de temperatura usou-se o termostato de Langevin com fator $\gamma = 1$ ps^{-1}, que corresponde à frequência das colisões simuladas, responsáveis pela troca de energia cinética entre os átomos. Foi considerado um raio de corte de 10Å para as interações não ligantes. Nesta fase pretendemos preparar o sistema para a simulação biologicamente relevante, tendo-se mantido o vo-

lume constante ao longo da simulação. Face à exigência computacional destes cálculos, eles foram realizados através do SANDER compilado em paralelo: `sander.MPI`.

Na produção, por outro lado, a temperatura e a pressão foram mantidas constantes, a 310.15 K e 1 bar, respetivamente, e o sistema foi evoluído ao longo de 100 ps com intervalo de integração de 1 fs. O controlo de temperatura e o raio de corte foram iguais ao da primeira dinâmica, enquanto que o controlo de pressão foi realizado através de "MD with isotropic position scaling" com um tempo de relaxação de 2 ps. Esta parte da DM também foi compilada pelo `sander.MPI` usando 4 nodos, demorando cerca de 10h para finalizar o cálculo.

Como não consegui com que cálculos tão longos corressem sem interrupção nas máquinas da FCUP (por alguma razão a sessão expirava em etapas diferentes da simulação), a partir daqui irei usar os resultados de DM dos professores.

5 Análise de Dados

Depois de finalizados os 120 ps de dinâmica, queremos analisar como evolui o sistema neste período. O nosso primeiro objetivo é procurar saber se ocorreram alterações significativas na estrutura da proteína. Depois foi analisada a evolução das variáveis termodinâmicas e do RMSD para tirar conclusões quanto à convergência da simulação.

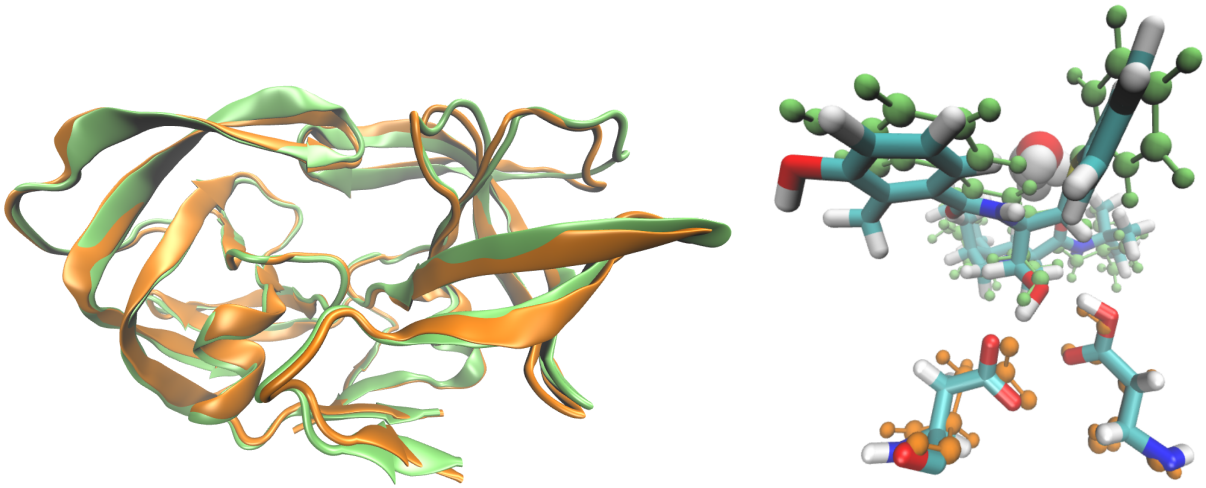
Os dados utilizados nesta análise obtiveram-se pelos resultados obtidos pela simulação de 120 ps dos professores. Os professores escolheram alterar no seu `.pdb` a protonação do ASP25 da cadeia B para ASH, enquanto que acima eu fiz-lo para a cadeia A.

5.1 Variações Estruturais

Na Fig. 5a encontra-se representada a sua estrutura quaternária no primeiro e último *frames* da fase de produção, a verde e laranja, respetivamente. Os *loops* correspondem às regiões mais desorganizadas e flexíveis, pelo que são as estruturas secundárias que mais variaram no decorrer da simulação. Por outro lado, as conformações α e β sofreram poucas alterações, um possível indicador de que a simulação não correu tempo suficiente para se observarem variações estruturais significativas do sistema.

Uma outra representação que pode ser importante é a das posições relativas dos ASP25 e do inibidor relativamente à proteína. Na Fig. 5b estão representados os resíduos ASP25 (`resid = 25 124`), a água catalítica (`resid = 25`) e o inibidor (`resid = 199`) antes e depois da produção. Como vemos, todas estas estruturas encontram-se muito próximas das suas posições iniciais, tendo sofrido algumas deformações estruturais. Atendendo às Figs. (4,5a), como as posições das *flaps* e do *backbone* se mantiveram aproximadamente as mesmas, então podemos concluir que as estruturas representadas na Fig. 5b se mantêm aproximadamente na localização inicial, pelo que o inibidor continua associado à enzima e no seu centro ativo.

A molécula de água estrutural também foi observada com um pouco mais de detalhe, tendo-se



(a) (b)

Figura 5: **(a)** Estrutura quaternária antes e depois da fase de produção, a verde e laranja, respetivamente. As estruturas foram alinhadas através da funcionalidade "Extensions - Analysis - RMSD Trajectory Tool" do VMD. **(b)** Comparação das posições relativas do inibidor, da água catalítica e dos resíduos ASP 25 antes e depois da fase de produção. As moléculas representadas a múltiplas cores correspondem às obtidas depois da produção. A laranja e verde estão representados os ASP25 e o nelfinavir, respetivamente, no início da produção. Aproximadamente ao centro da imagem, por cima do inibidor, consegue-se observar uma molécula de água a cores e, aproximadamente na mesma posição, uma outra molécula de água a cinzento.

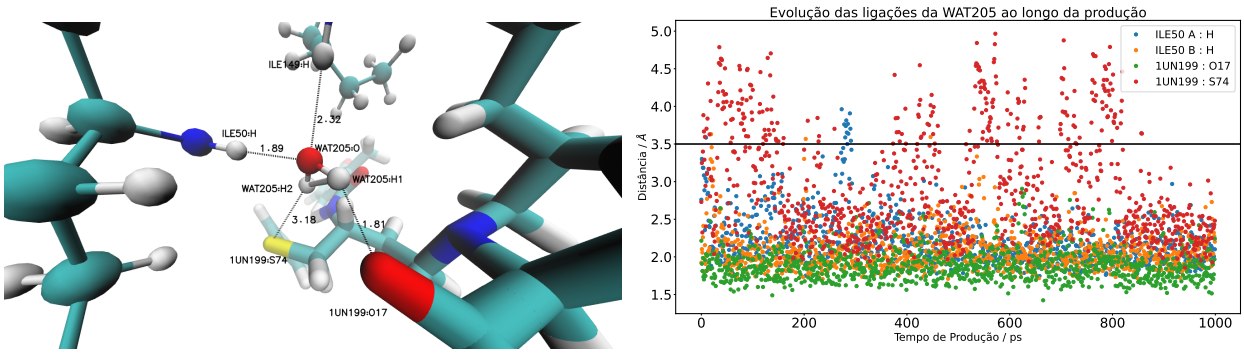
verificado que manteve uma posição praticamente fixa ao longo da simulação. Além disso, usou-se o VMD para determinar as pontes de hidrogénio em que esta água intervém, tendo-se observado duas com os resíduos ILE50 e duas com o inibidor - ver Fig. 6a - cujos comprimentos foram extraídos pelo VMD para todas as iterações da produção - Fig. 6b.

A Fig. 6b leva-nos a concluir que as pontes de hidrogénio com as ILE50 são muito estáveis, embora haja uma quebra da ligação por volta dos 300 ps, mas que pode ser justificada por uma configuração inicial não otimizada, dado que depois verifica-se que os pontos a azul concentram-se muito mais para distância pequenas. O facto destes resíduos pertencerem à proteína leva-me a concluir que estes são responsáveis pela grande estabilidade da água.

A água também se encontra fortemente ligada ao O17 do inibidor. No entanto, o S74 é o que quebra a ligação mais frequentemente, acabando por estabilizar abaixo do *treshold* a partir dos 800 ps de produção. Mais tempo de simulação é necessário para determinar se esta relação se mantém ou se surgem pontes de hidrogénio entre átomos diferentes.

5.2 Propriedades Termodinâmicas e RMSD

Como este tipo de sistemas biológicos é estável na natureza, só faz sentido tirar conclusões de um sistema termodinamicamente estável isto é, um sistema cujas propriedades termodinâmicas estejam convergidas. Essas propriedades são: volume, massa, temperatura, pressão e energia. O AMBER foi instruído a imprimir nos ficheiros de output todas as variáveis termodinâmicas que nos interessaram a cada 100 iterações, permitindo-nos construir os gráficos da Fig. 7a, onde podemos verificar que todas



(a) **(b)**

Figura 6: **(a)** Representação das quatro pontes de hidrogénio formadas através da molécula de água catalítica. **(b)** Gráfico das distâncias (Å) em função do tempo de produção (ps). As ligações ao inibidor são com os átomos O17 (verde) e S74 (vermelho), e a azul e amarelo estão as ligações com os resíduos ILE50 das cadeias A e B, respetivamente. A linha vertical, $d = 3.5$Å, assinala o *threshold* a partir do qual não identificamos uma ligação como correspondendo a uma ponte de hidrogénio.

as variáveis termodinâmicas estudadas convergiram.

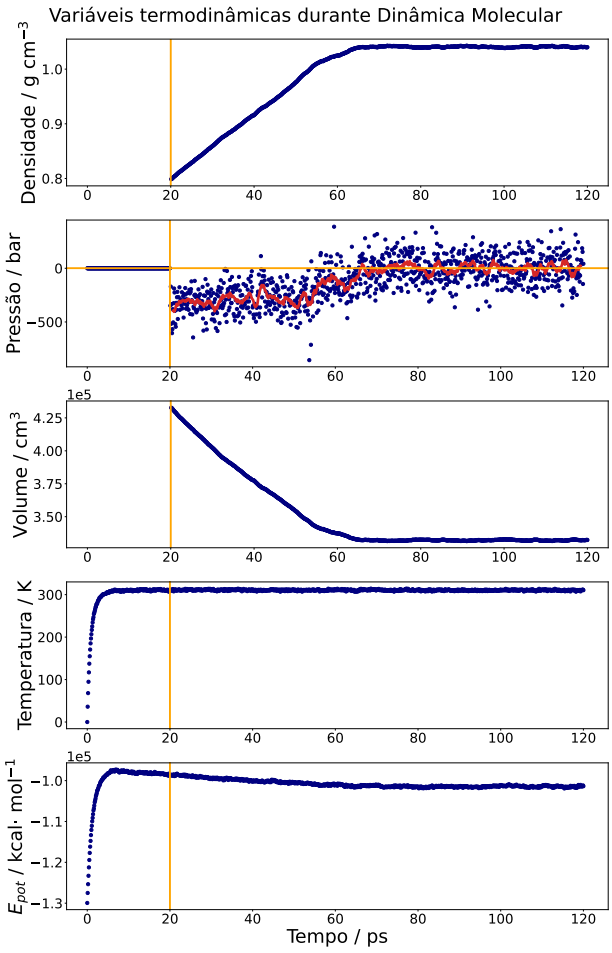
Numa discussão do trabalho com o meu colega Pedro Sousa, surgiram-nos dúvidas quanto à viabilidade da fase produção, que supostamente teria sido realizada a pressão constante. São evidentes oscilações significativas na pressão do sistema, na ordem das centenas de bar, de iteração para iteração. Consultando o capítulo "2.5.8 Pressure Regulation" do manual do AMBER, ficou claro que este tipo de oscilações é esperado do algoritmo, sendo que com o número de iterações a média das pressões irá tender para a pressão objetivo, no nosso caso 1 bar (linha amarela horizontal na Fig. 7a). A vermelho foi representada a média móvel de 15 pontos (intervalo de 14 fs) para ilustrar como de facto a média das pressões tende para o valor esperado.

Depois de verificada a convergência das variáveis termodinâmicas, o próximo gráfico a estudar é o da RMSD - ver Fig. 7b. Embora a RMSD aparente ter convergido para valores próximos de 1.3 Å, sabemos que os 120 ps de simulação são extremamente pequenos para sistemas biológicos deste tipo, onde evoluções de dinâmica de molecular de $3-4$ ns são tipicamente realizadas. Se continuássemos a simulação até 12 ns, como os professores fizeram, concluíamos que a estrutura ainda está longe de se encontrar convergida.

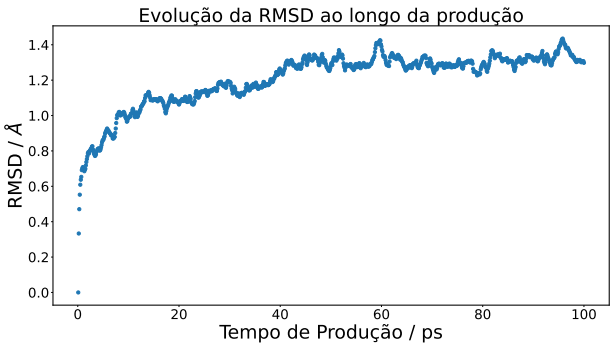
5.3 Função de Distribuição Radial

As distribuições radiais constituem um dos métodos mais importantes para a caracterização estrutural das interações entre um soluto e as moléculas do solvente. A "Radial Distribution Function" (RDF) entre dois tipos de partículas A e B, $g_{AB}(r)$, representa a densidade de partículas do tipo B à distância r do tipo A.

Estas funções foram determinadas usando o comando `ptraj` do AMBER para os resíduos ASP25, ASP29 e GLU21 da proteína. O solvente analisado foi a água que envolve a proteína. Os resultados destes cálculos estão representados na Fig. 8 assim como a contagem de moléculas de água analisadas



(a)



(b) RMSD da fase de produção. **(a)** Evolução das variáveis termodinâmicas no decorrer das simulações de Dinâmica Molecular. As linhas verticais a laranja indicam $t = 20$ ps, o instante de onde de passou da fase de equilibração para a produção. A linha horizontal laranja no gráfico da pressão indica $P = 1$ bar, a pressão objetivo estabelecida para o sistema. Nesse mesmo gráfico está representada a vermelho a média móvel de 15 pontos, correspondente a um intervalo de 14 fs.

Figura 7: Análise de convergência e estabilidade das simulações de Dinâmica Molecular.

à medida que nos afastamos dos átomos de referência. Os princípios físicos que governam os vários picos observados podem ser consultados em [16, 17], sendo que cada mínimo indica o fim de uma esfera de solvatação.

Na Fig. 8c conseguimos distinguir três esferas de solvatação para a cadeia B, presentes a $r \approx 7.1, 8.0$ e $9.0\mathring{A}$. Não consigo identificar com clareza numa destas esferas para o o ASP da cadeia A. Note-se que as RDFs começam a ganhar amplitude ligeiramente antes dos 6Å. Por outro lado, as distribuições para os resíduos ASP29 e GLU21 ganham amplitude por volta dos 2.2 Å e possuem duas esferas de solvatação a $r \approx 3.5$ e $r \approx 5.5\mathring{A}$, embora esta última seja pouco evidente.

Note-se que a Fig. 8c revela claras assimetrias entre os dois ácidos aspárticos, evidenciando o quanto a protonação do resíduo influencia a química dos processos na vizinhança desta zona da proteína. As distâncias elevadas observadas neste gráfico revelam que a água catalítica se encontra demasiado afastada destes resíduos para se poderem estabelecer pontes de hidrogénio entre eles. Um resultado curioso é que a primeira esfera de solvatação do ASP25 da cadeia B contém apenas uma molécula de água, a segunda contém duas, e a terceira três, todas separadas por aproximadamente 1Å. Em contrapartida, para as GLU21 e ASP29, a primeira molécula de água surge a 2.85 e 2.65 Å dos respetivos átomos de referência.

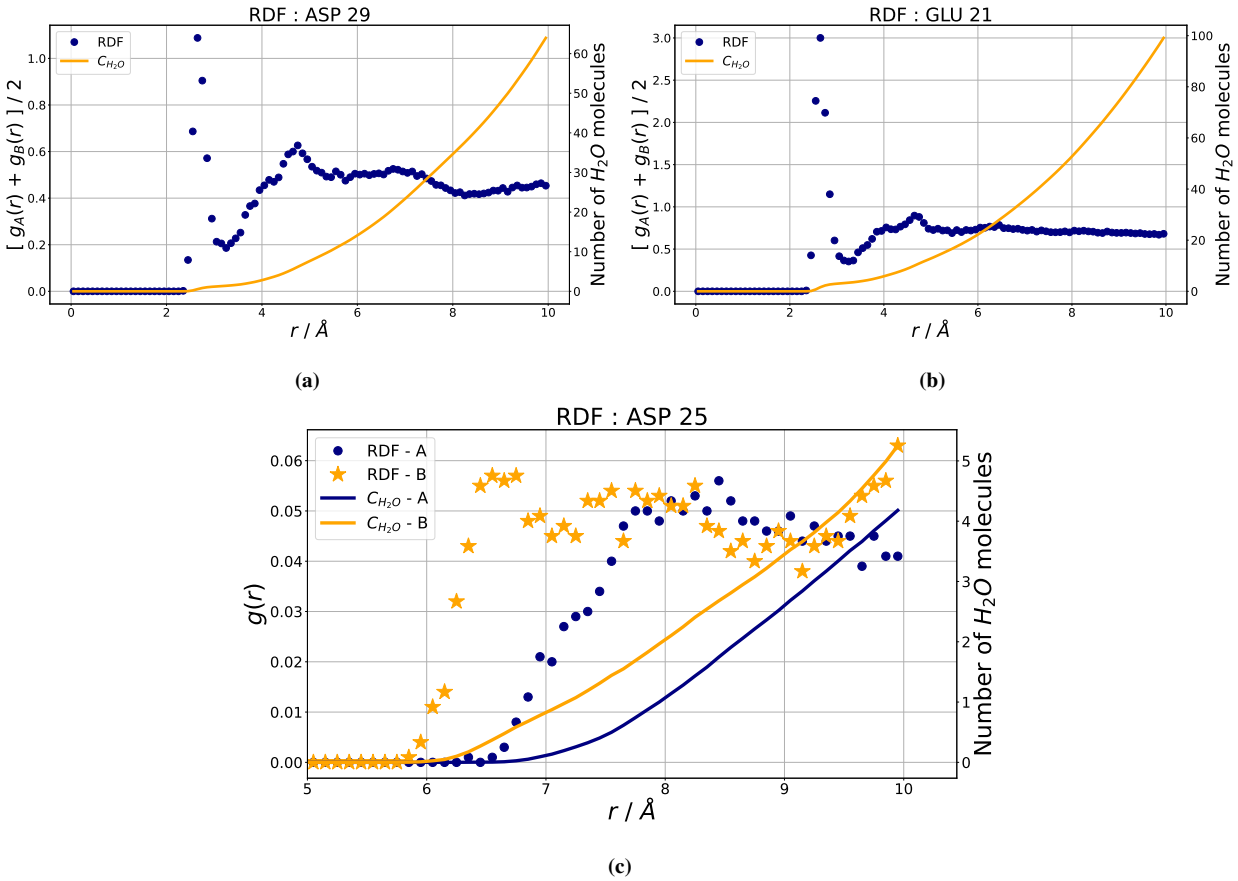


Figura 8: Funções de Distribuição Radial (pontos) e número de moléculas de água (linhas a cheio) em função da distância aos resíduos ASP29:OD* **(a)**, GLU21:OE* **(b)**, e ASP25:OD* **(c)**. Os dois primeiros gráficos foram obtidos através da média dos resultados para as cadeias A e B. Por outro lado, os dados para o ASP25 estão distinguidos por cadeia, uma vez que o ASP da cadeia B está protonado nesta estrutura.

Outro resultado interessante desta análise é o valor para o qual vemos que as densidades convergem para $r > 9\AA$. A convergência dá-se porque quanto mais nos afastamos do átomo de referência, menor a influência terão esses átomos sobre estas camadas exteriores e menor será a contribuição da densidade das camadas interiores para o cálculo médio da densidade, ou seja, aproximamo-nos da homogeneidade. No entanto, seria de esperar uma tendência para 1 (densidade da água) e não ≈ 0.5, como se verifica. Isto acontece porque aproximadamente metade do volume contido nas cascas esféricas para o cálculo da RDF é preenchido pela protease, que não contribui para a função densidade do solvente. Se expandíssemos a análise para raios maiores acabaríamos por ver o patamar esperado para a RDF. Porém, esse estudo não faria sentido porque nas nossas simulações só foram consideradas interações num raio de $10\AA$ de cada átomo.

Para ter uma melhor noção da origem destes resultados e da disposição dos resíduos na proteína, na Fig. 9 estão representados os resíduos estudados nesta secção assim como as suas águas mais próximas. Como podemos observar, as GLU21 e ASP29 encontram-se muito próximos das fronteiras da proteína, que contacta com o solvente, e daí a grande dispersão de nos resultados entre estes resíduos e os ASP, cuja água mais próxima é a catalítica.

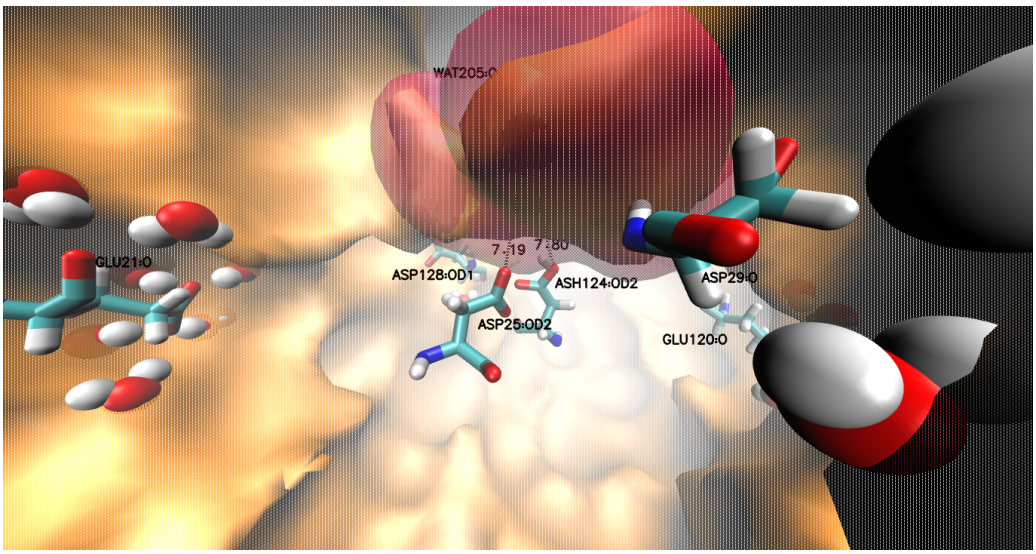


Figura 9: Representação dos resíduos estudados na RDF e das águas mais próximas de cada um deles. Para as GLU21 e ASP29 estão representadas as águas até 2.0 Å de si. A laranja e vermelho estão representadas a proteína e o inibidor em modo "QuickSurf".

5.4 Ligações de Hidrogénio

O `ptraj` foi novamente utilizado, desta vez para determinar os comprimentos (d) e ângulos (θ) de ligação de todas as pontes de hidrogénio presentes no sistema. Na Fig. 10 é apresentado um *print* do *output* destes cálculos com as cinco principais pontes de hidrogénio, onde o valor de ocupação indica a percentagem de *snapshots* do sistema em que a ligação em questão cumpria as condições para se qualificar como ponte de hidrogénio, um indicador da estabilidade da ligação, tendo-se detetado um total de 202 ligações deste tipo.

Para tentar encontrar os valores de d e θ que conferem maior estabilidade às ligações construiu-se o gráfico da Fig. 11. Embora esta análise seja muito primitiva e superficial, os valores de d mais estáveis parecem estar em torno dos $2.8 - 2.9$ Å, e os de θ mais próximos do intervalo $160 - 165°$.

Uma análise mais detalha passaria por distinguir as ligações por aceitadores e dadores, para podermos isolar as variações de d e θ das variações de vizinhança em que ocorrem.

```
HBOND SUMMARY:

    data was sorted, intra-residue interactions are NOT included,
    Distance cutoff is 3.50 angstroms, angle cutoff is 140.00 degrees
    Hydrogen bond information dumped for occupancies > 0.05

        DONOR          ACCEPTORH       ACCEPTOR
  atom# :res@atom   atom# :res@atom atom# :res@atom %occupied  distance       angle
|   950  :59@O    |  1179   :75@H    1178   :75@N    |  99.90  2.898 ( 0.12)  13.32 ( 7.07)
|  2515 :158@O    |  2744  :174@H    2743  :174@N    |  99.80  2.891 ( 0.11)  14.60 ( 7.20)
|   764  :47@O    |   833   :54@H     832   :54@N    |  99.50  2.892 ( 0.11)  18.80 ( 7.41)
|  2702 :170@O    |  2584  :163@H    2583  :163@N    |  99.40  2.895 ( 0.11)  14.47 ( 7.70)
|  2329 :146@O    |  2398  :153@H    2397  :153@N    |  99.40  2.931 ( 0.13)  18.55 ( 7.47)
```

Figura 10: Tabela com as cinco pontes de hidrogénio mais estáveis do sistema ao londo dos 120 ps de DM. Podemos facilmente identificar os átomos dadores e aceitadores de cada uma. Os valores entre "()" correspondem à incerteza da medição em questão.

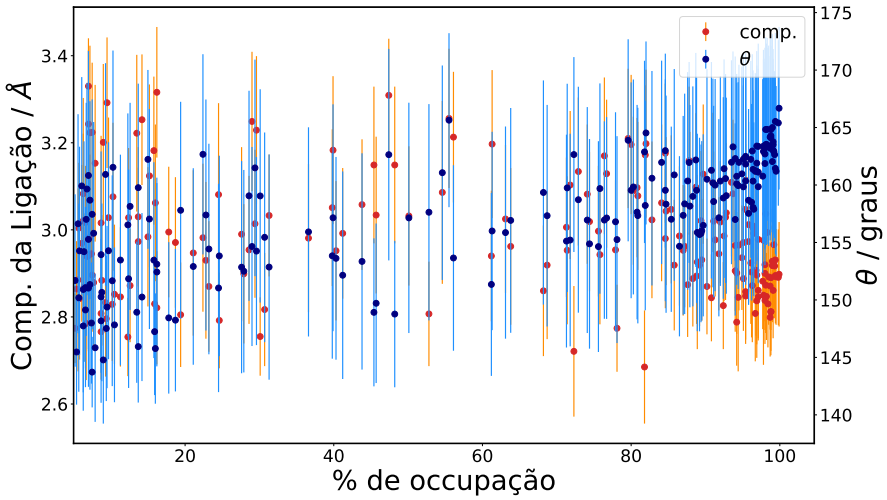
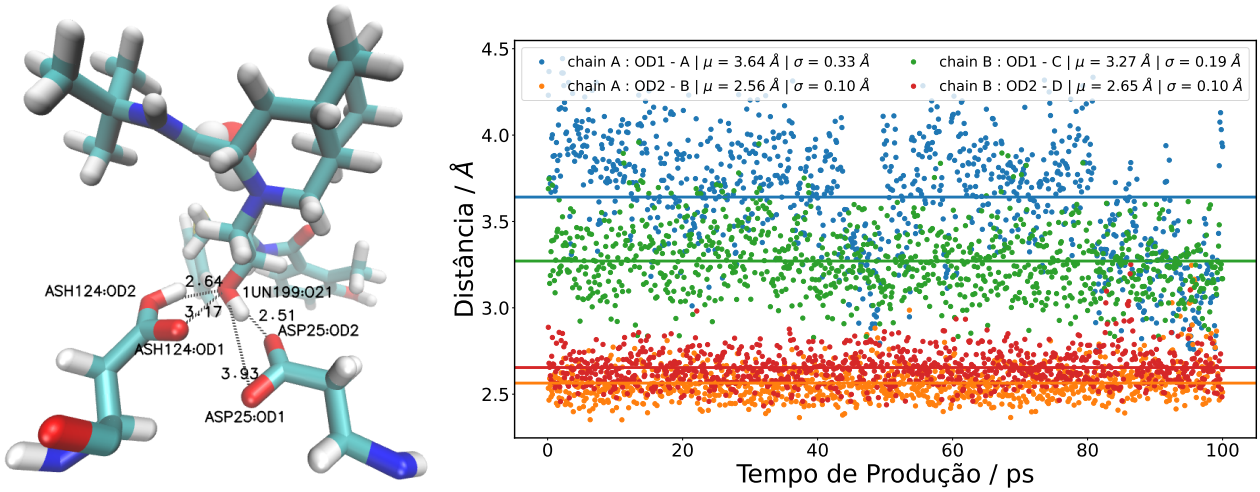


Figura 11: Representação dos comprimentos e ângulos de ligação para as 202 pontes de H estudadas, a vermelho e azul, respetivamente. As barras de erro estão representadas a tons mais leves das respetivas cores.

5.5 Cálculo de Distâncias e Ângulos

Na Sec. (5.1) foi apresentado uma breve análise dos comprimentos de ligação das pontes de hidrogénio da água catalítica usando o VMD - ver Fig. 6. O VMD também é capaz de determinar ângulos de ligação e diedros com a mesma facilidade. Estas grandezas podem ter muito interesse conforme o tipo de análise em que estamos interessados. No entanto, o método é pouco autonomizável, por ser necessário selecionar manualmente os átomos numa interface gráfica. É neste contexto que entra mais uma vez o `ptraj`, que nos permite determinar as grandezas referidas com grande facilidade.

Na Fig. 12a estão representadas as ligações que o inibidor forma com os OD* dos resíduos ASP25 (cadeia A) e ASH124 (cadeia B). Estes dois resíduos são aqueles que estão implícitos quando menciono "os resíduos ASP25". Os comprimentos de ligação foram representados graficamente na Fig. 12b, onde estão devidamente assinaladas a média e o desvio padrão para cada uma das ligações.



(a) **(b)**

Figura 12: **(a)** Representação das ligações entre os átomos OD* dos ASP25 ao O21 do inibidor. **(b)** Gráfico das distâncias (Å) em função do tempo de produção (ps). Na legenda pode-se observar o valor médio (μ) e o desvio padrão (σ) dos comprimentos de ligação. Na vertical também estão assinaladas, nas suas cores correspondentes, linhas que correspondem ao μ de cada ligação.

6 Conclusões

Neste trabalho foram exploradas os primeiros passos com rumo à simulação de sistemas biológicos, tendo sido abordada teoria, trabalho computacional e vários métodos de análise da dados.

A incorporação do Nelfinavir na HIV-1 protease foi o nosso objeto de estudo e foi realizada uma pequena simulação de 120 ps de Dinâmica Molecular, que se mostrou termodinamicamente estável, embora uma estrutura convergida só fosse possível atingir para simulações superiores a 3 ns.

A estrutura final e a sua evolução ao longo da etapa de produção foram analisados à luz das suas propriedades termodinâmicas, *root mean square deviation* e função de distribuição radial (RDF). Foram também estudadas todas as pontes de hidrogénio do sistema, num total de 202 detetadas, com ênfase atribuída às que uma molécula de água estrutural estabelece com o Nelfinavir e a proteína. Finalmente, foram analisados os comprimentos de ligação entre os ácidos aspárticos do centro ativo da proteína e o inibidor.

Em geral, achei todo este trabalho extremamente interessante e enriquecedor. Tive a oportunidade de aprender fundamentos de uma área que achava ser completamente ortogonal à Física mas que se veio a revelar ter imensa coisa em comum. Gostei do desafio de investigar vários conceitos como a protonação dos ASP25 e a construção dos campos de forças. No entanto, gostava de ter explorado mais a fundo outros pontos como as RDFs e a análise das pontes de hidrogénio, mas sinto que ter frequentado esta cadeira e escrever este relatório me proporcionaram a curiosidade e confiança para investigar estes assuntos por conta própria.

Muito obrigado aos professores pelas ótimas aulas, por promoverem a discussão com os alunos e estarem sempre disponíveis para esclarecer dúvidas.

Referências

[1] Thomas E. Cheatham, Bernard R. Brooks, and Peter A. Kollman. Molecular Modeling of Nucleic Acid Structure: Energy and Sampling. *Current Protocols in Nucleic Acid Chemistry*, 4(1), April 2001.

[2] https://chem.libretexts.org/@go/page/13415, nov 15 2021. [Online; accessed 2022-01-22].

[3] Scott J. Weiner, Peter A. Kollman, David A. Case, U. Chandra Singh, Caterina Ghio, Guliano Alagona, Salvatore Profeta, and Paul Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784, February 1984. Publisher: American Chemical Society.

[4] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, May 1995. Publisher: American Chemical Society.

[5] Rodrigo Galindo-Murillo, Christina Bergonzo, and Thomas E. Cheatham III. Molecular Modeling of Nucleic Acid Structure: Setup and Analysis. *Current Protocols in Nucleic Acid Chemistry*, 56(1):7.10.1–7.10.21, 2014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471142700.nc0710s56.

[6] Kiril Ivanov Tenekedjiev, Natalia Danailova Nikolova, and Krasimir Kolev. Applications of Monte Carlo Simulation in Modelling of Biochemical Processes. In Charles J. Mode, editor, *Applications of Monte Carlo Methods in Biology, Medicine and Other Fields of Science*, Wellcome Trust–Funded Monographs and Book Chapters. InTech, Rijeka (HR), 2011.

[7] M.E.J. Newman. *Computational physics*. Createspace Independent Pub, 2013.

[8] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, March 1977.

[9] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[10] Magdalena Woińska, Simon Grabowsky, Paulina M. Dominiak, Krzysztof Woźniak, and Dylan Jayatilaka. Hydrogen atoms can be located accurately and precisely by x-ray crystallography. *Science Advances*, 2(5):e1600192, May 2016.

[11] Unraveling HIV protease flaps dynamics by Constant pH Molecular Dynamics simulations - ScienceDirect.

[12] Rosemberg O. Soares, Pedro H.M. Torres, Manuela L. da Silva, and Pedro G. Pascutti. Dataset showing the impact of the protonation states on molecular dynamics of HIV protease. *Data in Brief*, 8:1144–1150, July 2016.

[13] Substrate Recognition in HIV-1 Protease: A Computational Study | The Journal of Physical Chemistry B.

[14] Yong Duan, Chun Wu, Shibasish Chowdhury, Mathew C. Lee, Guoming Xiong, Wei Zhang, Rong Yang, Piotr Cieplak, Ray Luo, Taisung Lee, James Caldwell, Junmei Wang, and Peter Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry*, 24(16):1999–2012, December 2003.

[15] Yangpeng Zhang, Hao Liu, Sheng Yang, Ray Luo, and Hai-Feng Chen. A Well-balanced Force Field ff03CMAP for Folded and Disordered Proteins. *Journal of chemical theory and computation*, 15(12):6769–6780, December 2019.

[16] Russell J. Boyd. Electron density partitioning in atoms. *The Journal of Chemical Physics*, 66(1):356–358, January 1977. Publisher: American Institute of Physics.

[17] Peter Politzer and Robert G. Parr. Separation of core and valence regions in atoms. *The Journal of Chemical Physics*, 64(11):4634–4637, June 1976. Publisher: American Institute of Physics.