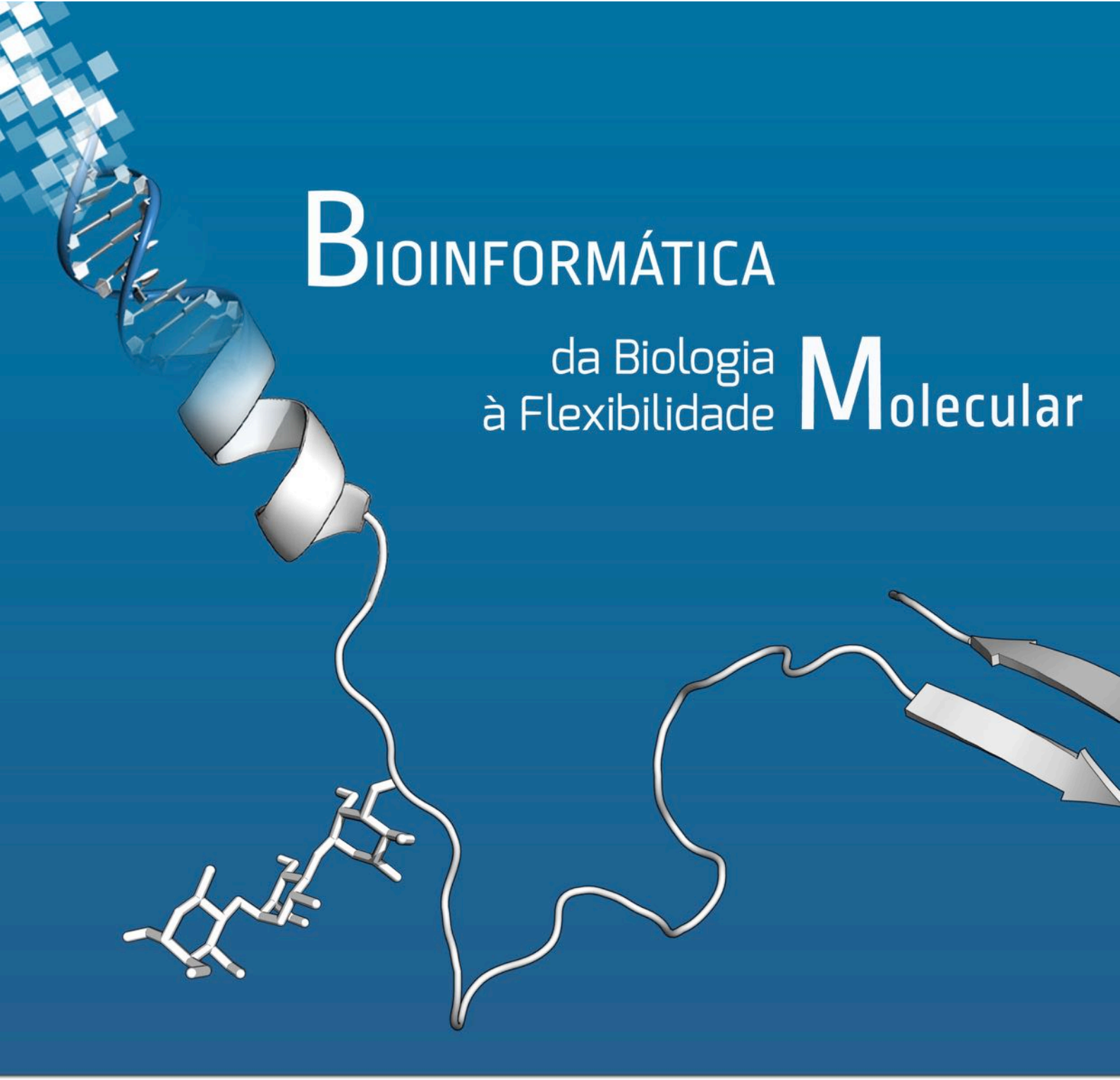


Bioinformática
da Biologia à Flexibilidade Molecular

Hugo Verli (Org.)

1ª edição
São Paulo, 2014

ISBN 978-85-69288-00-8
9 788569 288008

Sociedade Brasileira de Bioquímica e Biologia Molecular — SBBq

Apoio:
CNPq
CAPES
FAPERGS
UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Hugo Verli Organizador

Bioinformática: da Biologia à Flexibilidade Molecular

1ª Edição

São Paulo

Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq

2014

Ficha catalográfica elaborada por Rosalia Pomar Camargo CRB 856/10

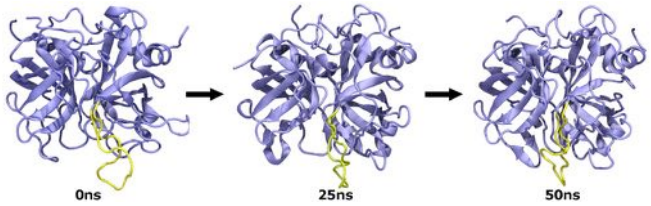
B615 Bioinformática da Biologia à flexibilidade molecular / organização de Hugo Verli. - 1. ed. - São Paulo : SBBq, 2014. 282 p. : il.

1. Bioinformática 2. Biologia Molecular

CDU 575.112
ISBN 978-85-69288-00-8

8. Dinâmica Molecular

Hugo Verli



Flexibilidade da enzima trombina evidenciada através de simulação por dinâmica molecular.

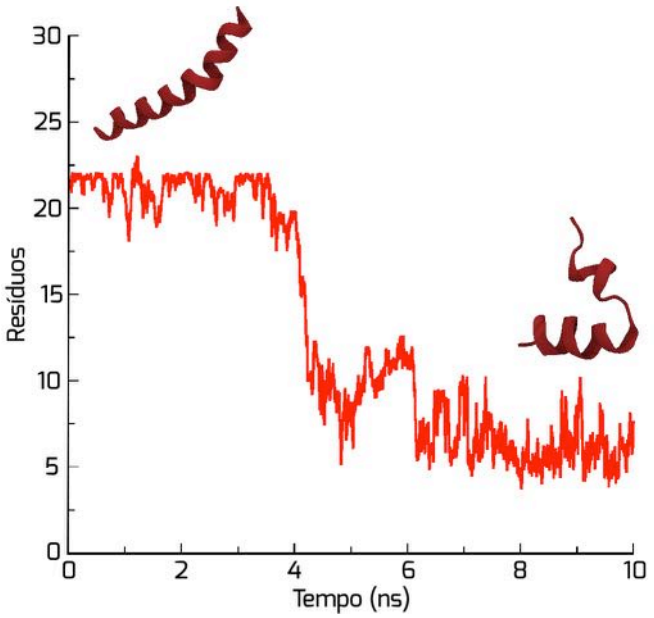


Figura 1-8: Variação do conteúdo de estrutura secundária da melitina, peptídeo da abelha *Apis mellifera*, como função do tempo. A forma inicial é encontrada no ambiente cristalino, enquanto a final é observada em condições próximas às plasmáticas.

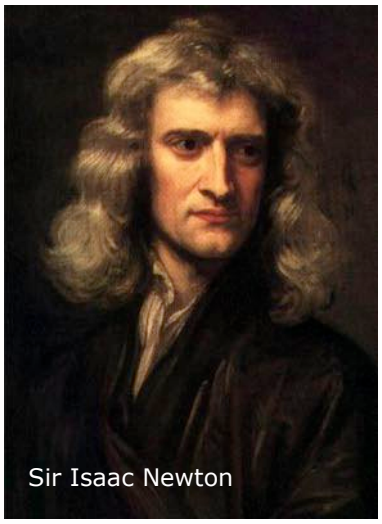
8.1. Introdução

Segundo a IUPAC (*International Union of Pure and Applied Chemistry*), a "dinâmica molecular é um procedimento de simulação que consiste na computação do movimento dos átomos em uma molécula ou de átomos individuais ou moléculas em sólidos, líquidos e gases, de acordo com as leis de movimento de Newton". Em outras palavras, a dinâmica molecular (DM) descreve a variação do comportamento molecular como função do tempo (Figura 1-8).

Quando mencionamos "comportamento molecular", nos referimos a quaisquer propriedades de uma molécula em estudo, tais como seu conteúdo de estrutura 2^{ária}, orientação de cadeias laterais, conformação de alças e a energia de interação entre diferentes moléculas (enzima e substrato, proteína e proteína, proteína e DNA ou fármaco e receptor). Por outro lado, a ideia de que estas propriedades variam como função do tempo indica que as mesmas não são estáticas, mas se modificam em soluções biológicas. Isto aproxima em muito a DM de métodos experimentais como a Ressonância Magnética Nuclear (RMN, Capítulo 12), que geram medidas representando, de fato, médias temporais, colhidas durante a realização do experimento. Assim, ao final de uma simulação de DM, buscamos estas propriedades médias, representativas de comportamentos biológicos medidos experimentalmente.

A descrição conformacional oferecida pela DM, para uma determinada molécula ou

conjunto de moléculas, baseia-se na solução da 2ª Lei de Newton, onde Fx_i é a força aplicada ao átomo i na posição x, t é o tempo, v a velocidade e a_i a aceleração do átomo i. Por ser baseada na física desenvolvida por Sir. Isaac Newton, a DM faz parte dos métodos denominados Clássicos (também chamados de métodos de mecânica molecular), em oposição aos métodos baseados na física quântica (que deram origem aos denominados métodos de mecânica quântica).


Sir Isaac Newton

$$F_{x_i} = \frac{d^2 x_i}{dt^2} m_i = \frac{\Delta v_i}{\Delta t} m_i = a_i m_i$$

Assim, a DM nos possibilita obter modelos de moléculas muito mais próximos da realidade biológica, pois inclui diretamente características como a flexibilidade molecular (através da variação temporal de propriedades) e a temperatura (através da aceleração dos átomos). A maioria dos fenômenos biológicos estão associados à flexibilidade de biomoléculas, como a catálise e a modulação de canais iônicos e de receptores acoplados à proteína G. De fato, muitos destes processos vêm sendo descritos com sucesso por simulações de DM ao longo dos anos.

Outros tipos de simulação estão disponíveis, tais como o Método de Monte Carlo, a Dinâmica Estocástica e a Dinâmica Browniana. Iremos, contudo, nos ater à DM em decorrência de seu maior uso, nos últimos anos, no estudo de biomoléculas.

Muitos programas (Tabela 1-8) estão disponíveis para a realização de simulações por DM diferindo, por exemplo, quanto a seu acesso (gratuito ou pago), custo computacional (isto é, tempo necessário para a execução de um mesmo cálculo) e tipos de campos de força disponíveis (ver adiante).

8.2. Campos de força

Como visto no item anterior, para descrever a variação da posição x de um átomo i como função do tempo precisamos conhecer o valor da massa de cada átomo, m_i (essa é fácil, vem da tabela periódica) e a força (Fx_i) sobre cada átomo i em uma determinada posição x. A temperatura fornece energia para que os átomos sofram uma aceleração, mudando suas posições no espaço. Contudo,

Tabela 1-8: Alguns dos principais programas disponíveis para simulações por DM.

Programa	Distribuição
Abalone	Gratuito
ADUN	Gratuito
AMBER	Pago
Ascalaph Designer	Gratuito
CHARMM	Pago
Discovery Studio	Pago
GROMACS	Gratuito
GROMOS	Pago
GULP	Gratuito
LAMMPS	Gratuito
MDynaMix	Gratuito
MOE	Pago
MOIL	Gratuito
MOLDY	Gratuito
NAMD	Gratuito
RedMD	Gratuito
TeraQuem	Pago
TINKER	Gratuito
YASARA	Pago

como os átomos não estão isolados, mas ligados a outros átomos formando moléculas que, por sua vez, interagem com outras moléculas, eles estão sujeitos a forças interatômicas e inter-moleculares. O cálculo destas forças é realizado por uma outra função matemática, denominada campo de força.

O campo de força, seguindo a definição da IUPAC, pode ser descrito brevemente como "um conjunto de funções e parametrização usadas em cálculos de mecânica molecular". Cada campo de força estabelece um conjunto de equações matemáticas dedicadas a reproduzir aspectos do comportamento molecular, como o estiramento de ligações químicas, a deformação de um ângulo de ligação ou a torção de um diedro, como podemos observar em um espectro de infravermelho. Estas equações, por sua vez, são calibradas (ou seja, parametrizadas) para reproduzir o comportamento dos compostos de interesse (Figura 2-8).

Equações e parametrizações diferentes podem ser empregadas, dando origem a campos de força diferentes, com vantagens e também limitações. Por exemplo, enquanto um tipo de campo de força pode descrever com elevada fidelidade proteínas, ele pode ser bastante limitado na reprodução da geometria de carboidratos ou ácidos nucleicos. Desta forma, ao iniciarmos um estudo por DM, devemos ter em mente qual o tipo de molécula com o qual pretendemos trabalhar e qual o melhor campo de força para descrevê-la.

A escolha de um campo de força não é, contudo, baseada somente no tipo de molécula com o qual queremos lidar. Diversos outros aspectos podem influenciar esta escolha. Existem, por exemplo, diferentes níveis de simplificação na descrição dos átomos (Figura 3-8). O campo de força pode descrever todos os átomos do sistema (em inglês são denominados campos de força *all atom*), mas isto implica em um maior custo computacional, o que pode se tornar proibitivo no estudo de grandes sistemas moleculares se não temos acesso a grandes estruturas de processamento em paralelo (os chamados *clusters*).

Como o elemento encontrado em maior quantidade é o átomo de hidrogênio, uma primeira simplificação é denominada de átomo unido (em inglês são denominados campos de força *united atom*). Neste

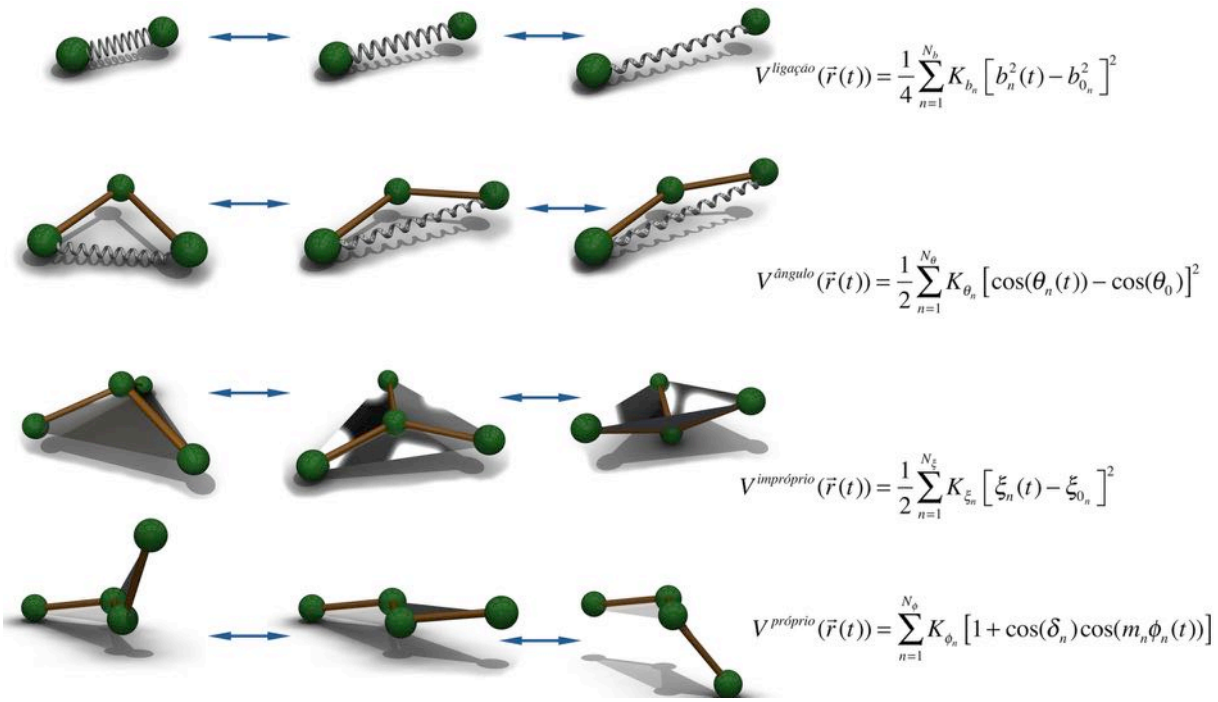


Figura 2-8: Representação de alguns termos que compõem o campo de força GROMOS96. Termos semelhantes são também encontrados em diversos outros campos de força.

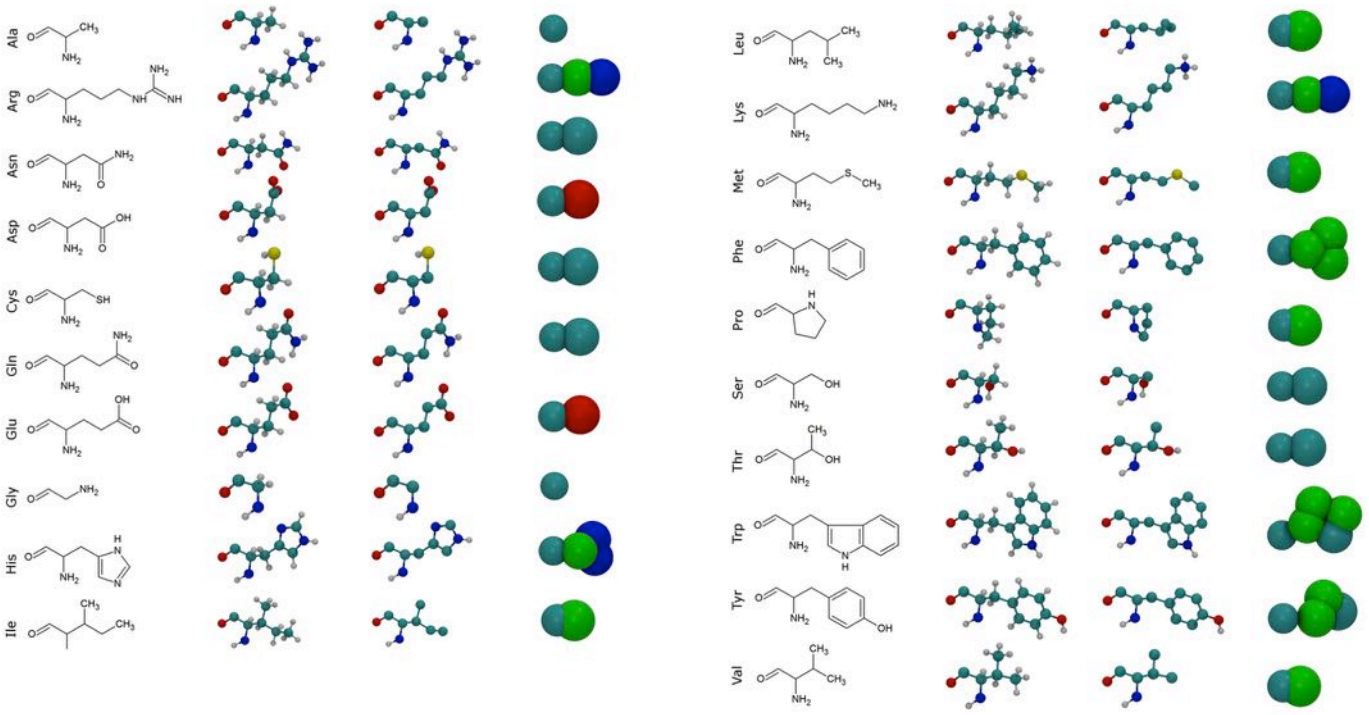


Figura 3-8: Representação dos 20 aminoácidos, codificados no genoma para síntese proteica, em um campo de força descrevendo todos os átomos, em um campo de força de átomo unido e *coarse-grained*.

caso, os átomos de hidrogênio apolares, ou seja, aqueles ligados a átomos de carbono, são unidos a este elemento, dando origem a um pseudoátomo representando as propriedades de grupos CH, CH_2 ou CH_3. Exceção se dá para o grupo CH de anéis aromáticos, que tem os átomos de hidrogênio descritos explicitamente nos campos de força de átomo unido mais modernos, como o GROMOS96.

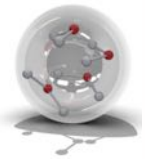
Há, por fim, um terceiro nível de simplificação, denominado *coarse-grained* (CG). Neste campo de força, vários átomos podem ser agregados em uma única partícula, análoga ao pseudoátomo do modelo de átomo unido. Por exemplo, todo um aminoácido pode ser considerado como uma única partícula, como é o caso da alanina e da glicina no campo de força MARTINI. Em outros resíduos, este campo de força considera o esqueleto peptídico como uma partícula e a cadeia lateral de uma (como na cisteína, treonina e serina) a três (histidina e fenilalanina) ou quatro (triptofano) partículas.

Quanto maior a simplificação, menor custo computacional do cálculo. Em outras palavras, podemos simular sistemas com maior número de átomos por mais tempo em computadores mais baratos. Infelizmente, estas simplificações trazem consigo algumas limitações. No caso do CG, perde-se a capacidade de descrever elementos de estrutura 2ária, mantendo-se somente a forma global da molécula em estudo. Assim, em estudos onde são esperadas mudanças no conteúdo de estrutura 2ária o método de CG não é indicado. Mas, por ser muito rápido, pode descrever movimentos entre diferentes domínios de uma dada proteína, o que é difícil de ser observado, usualmente, nos demais campos de força. Por outro lado, o caso dos modelos de átomo unido traz limitações como a dificuldade em se utilizar estes campos de força na obtenção e refinamento de modelos 3D de macromoléculas a partir de dados de RMN (Capítulo 12).

Outra diferença entre os campos de força diz respeito à descrição das moléculas de água, o principal solvente de biomoléculas (Tabela 2-8). De fato, uma das grandes vantagens do método de DM é a capacidade de incluir a presença de moléculas de água nos modelos gerados, descrevendo as suas interações, como função do tempo, com os compostos em estudo. Da mesma forma que visto para os campos de força, existem diversos modelos para descrição de moléculas de água, por vezes com mais de uma opção para um mesmo campo de força.

Estes organizam-se em dois grandes grupos: os modelos explícitos e os implícitos.

Tabela 2-8: Alguns dos modelos de água mais comumente empregados em simulações por DM[a].

Modelo	Campos de força onde são empregados	Tipo
SPC SPC/E TIP3P	AMBER, GROMOS, OPLS	
TIP4P	AMBER, CHARMM, OPLS	
TIP5P		
MARTINI	Martini	

[a]Uma revisão mais completa pode ser encontrada no site: www1.lsbu.ac.uk/water/models.html

Enquanto os modelos explícitos incluem os átomos da molécula de água, fisicamente, na simulação, os modelos implícitos (também chamados de modelos contínuos ou *continuum models*) não incluem estas moléculas diretamente, mas indiretamente, através da representação das propriedades dielétricas do solvente. Os átomos que compõem a água não participam das simulações, tornando o cálculo extremamente rápido (usualmente, a grande maioria dos átomos em um sistema a ser simulado por DM se refere ao solvente). Infelizmente, enquanto estes modelos implícitos são bastante eficientes no estudo de proteínas e ácidos nucleicos, o mesmo não vem se mostrando para carboidratos, compostos altamente polares que interagem intensamente com o solvente.

Embora os principais campos de força empregados atualmente (AMBER, CHARMM, OPLS e GROMOS) sejam compostos por equações bastante semelhantes (ver a seguir), cada um foi construído a partir de decisões metodológicas distintas apresentando, portanto, particularidades importantes. Como consequência, normalmente os parâmetros de um campo de força não são transferíveis para outro campo de força.

A importância de conhecermos estas características, reconhecendo cada campo de força como entidade única, reside no fato de que um grande número de compostos de interesse biológico não é descrito nos parâmetros atuais, o que pode limitar o seu estudo computacional. Dentre estes compostos com carências de parâmetros podemos citar aminoácidos modificados (além dos 20 codificados no genoma), neurotransmissores, hormônios, fosfolipídeos, carboidratos, produtos naturais e, por fim, fármacos. Como simulações por DM podem ser cálculos extremamente demorados, deixar para descobrir no meio do trabalho que seu modulador de interesse não tem parâmetros no campo de força escolhido pode lhe custar alguns meses de trabalho.

Em linhas gerais, tanto a distância entre 2 átomos ligados quanto o ângulo entre 3 átomos consecutivos é descrita a partir de $V_{ligação/ângulo} = K_n\,[n - n_o]^2$, onde V é a energia, n é a distância ou ângulo em um dado momento, n_o é a distância ou ângulo de referência e K_n é a constante de força da mola que mantém esses valores ao redor dos valores de referência (Figura 2-8).

Para diedros, a função mais usualmente empregada é baseada em $V_{diedro} = K_\chi\,[1 + cos(n_\chi - \delta)]$, sendo V a energia, $_\chi$ o valor do diedro e K_χ a altura da barreira de energia entre diferentes estados conformacionais. Estes estados surgem porque um diedro pode rodar 360º e, ao longo desta rotação, apresentar múltiplos mínimos de energia. Assim não há, necessariamente, uma única geometria de referência. O perfil rotacional dos diedros tem a adição do parâmetro n, que descreve a multiplicidade do diedro (ou seja, o número de mínimos de energia) e δ, que diz respeito à mudança de fase e à localização do máximo de energia ao longo do perfil da rotação do diedro.

Apesar da semelhança nesses termos, existem diferenças importantes que devem ser consideradas. O CHARMM, por exemplo, emprega uma equação adicional na descrição dos ângulos de ligação, chamada

Urey-Bradley, que busca preservar a distância entre o primeiro e o terceiro átomos de um ângulo. Outra diferença se refere aos termos que descrevem a planaridade ou quiralidade em um conjunto de quatro átomos, o que é usualmente chamado de diedro impróprio (Figura 2-8). Enquanto AMBER e OPLS os descrevem da mesma forma que os demais diedros (também chamados de diedros próprios), CHARMM e GROMOS aplicam uma equação diferente, que se assemelha àquela empregada para distâncias e ângulos.

Abordar com profundidade a construção de parâmetros para campos de força está além do objetivo deste livro. Mas em muitos casos há uma solução um pouco mais simples para o problema. Uma característica importante de campos de força é a chamada transferabilidade. Isto significa que grupos químicos semelhantes possuem propriedades semelhantes que podem, assim, serem transferidas de uma molécula para outra. Por exemplo, o grupo hidroxila de um resíduo de Ser é equivalente ao grupo hidroxila de um resíduo de Thr. Assim, há uma redução enorme na necessidade de construção de parâmetros para novos compostos, se respeitarmos a semelhança química entre eles.

8.3. Minimização de energia

Quando iniciamos um estudo baseado em simulações por DM, podemos empregar estruturas de partida de diferentes origens, como modelos teóricos (ver capítulo 7) ou ainda dados experimentais de cristalografia de raios-X (ver capítulo 13) ou de RMN (ver capítulo 12). Independente de sua origem estas estruturas, ao serem solvatadas, criam interações soluto-solvente até então inexistentes (seja pelo dado ser teórico obtido no vácuo, em ambiente cristalino ou como uma média de diferentes conformações). Mas o solvente precisa se adaptar ao redor de seu soluto, e isto precisa ser corrigido antes que a simulação por DM se inicie. Por exemplo, quando o programa insere uma molécula de água, esta pode ter seu hidrogênio apontando para um átomo de hidrogênio da cadeia lateral de uma arginina, promovendo uma repulsão eletrostática pela proximidade de duas cargas de sinais iguais. Se isto não for corrigido antes do início da DM, a liberação desta energia na simulação pode gerar uma explosão da simulação (Figura 4-8) ou, de forma mais sutil (mas nem por isso menos perigosa para o estudo), promover mudanças conformacionais na proteína, ou mesmo desnaturações. Em outros casos, como na obtenção de modelos teóricos para a estrutura 3D de proteínas, a construção de cadeias laterais de aminoácidos pode aproximá-las artificialmente (e excessivamente) de outros resíduos.

Assim, uma das principais formas de tentar eliminar estes problemas reside no cálculo de minimização de energia (Figura 5-8). Durante este cálculo, a energia global do sistema é reduzida, alcançando por fim uma conformação mais estável para o sistema em estudo (ou seja, um estado de mínimo de energia).

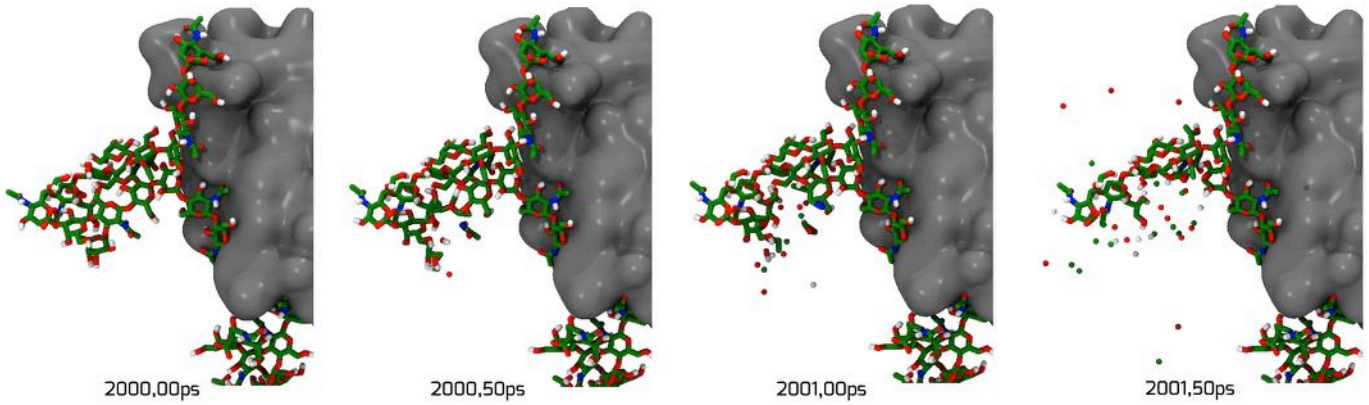


Figura 4-8: Explosão em uma simulação por DM.

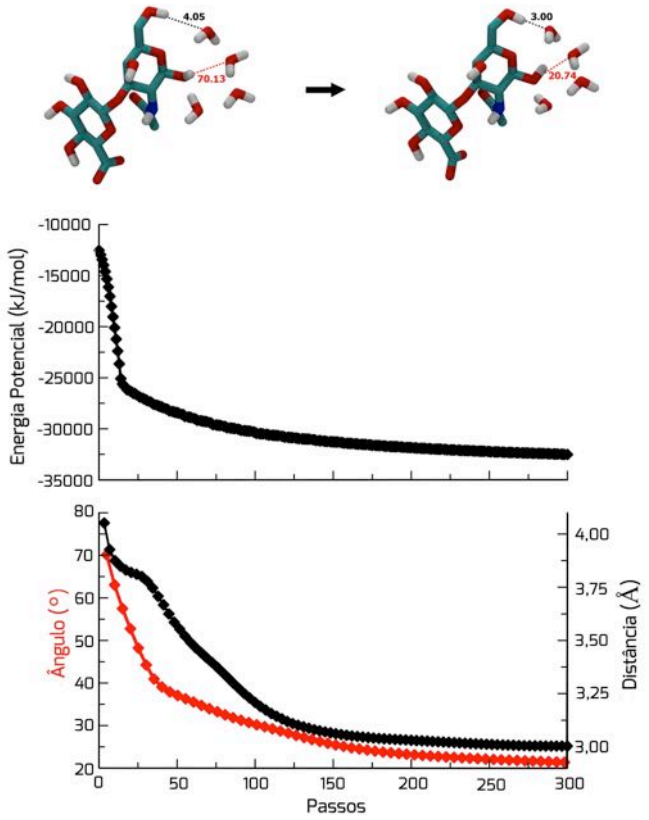


Figura 5-8: Exemplo da evolução de propriedades moleculares no decorrer de uma minimização de energia. A cada passo, a energia do sistema diminui, com a redução de contatos desfavoráveis e a formação de interações intra- e inter-moleculares como ligações de hidrogênio.

8.4. Simulações por DM

Além da escolha do campo de força e do modelo de água, o preparo e a análise de uma simulação por DM deve considerar alguns aspectos metodológicos importantes, dentre os quais destacaremos as condições periódicas de contorno, a equilibração, a amostragem, o tempo de integração e o cálculo de interações não ligadas. Uma escolha inadequada destas propriedades pode significar desde um maior custo computacional (isto é, uma simulação demorando mais do que precisaria) a resultados que não representam situações reais.

Condições periódicas de contorno

Quanto maior o número de moléculas incluídas em uma simulação, maior será o tempo necessário para realizar o cálculo. Por isso, buscamos sempre incluir o menor número de moléculas possível capaz de descrever as condições experimentais ou fisiológicas de referência. No caso da proteína, estamos na maioria das vezes ainda limitados a simulação de uma única molécula (salvo no caso de oligômeros). Contudo, a proteína não costuma ser a parte mais cara computacionalmente do cálculo, mas sim a inclusão do solvente (explícito). Uma otimização no número de moléculas de água pode representar uma grande otimização no tempo de máquina para conclusão da simulação (o que permite aumentar o tamanho da amostragem do estudo, ver adiante).

Uma forma de controlar o número de moléculas de água é controlando o tipo de "caixa" onde o sistema será simulado. Por caixa entendemos o espaço tridimensional onde soluto (biomolécula) e solvente (normalmente água) são colocados. O tamanho e a forma desta caixa, usualmente centralizada no soluto, definirá a quantidade de solvente a ser inserida.

Atualmente, não é comum definir a forma da caixa como uma esfera, por motivos que explicaremos a seguir. As formas mais comuns são cúbica, octaédrica e dodecaédrica. A forma de um octaedro apresenta 77% do volume de um cubo, enquanto que o dodecaedro 71%, representando a forma mais próxima de uma esfera. Contudo, como a forma de proteínas e outras biomoléculas varia muito, devemos avaliar qual caixa se adequa melhor ao sistema em estudo. Por exemplo, a simulação de membranas é normalmente realizada em um cubo ou uma forma retangular, que pode ser uma boa alternativa também para proteínas em forma de bastão.

O uso de uma caixa em forma de esfera ao redor da proteína de interesse nos levaria a um aproveitamento do espaço tridimensional melhor do que o dodecaedro, economizando mais moléculas de água e, assim, liberando custo computacional. Contudo, as moléculas em uma simulação por DM podem se difundir ao longo da caixa. Como além da caixa de simulação temos condições de vácuo, o solvente iria progressivamente evaporar, a partir da face da esfera. A forma de

impedir isso é criar uma força que impeça as moléculas do sistema de ultrapassarem os limites desta esfera, o que representa a inclusão de forças artificiais, não observáveis em condições biológicas.

As formas geométricas empregadas mais frequentemente em em simulações por DM estão relacionadas a uma estratégia denominada condições periódicas de contorno (Figura 6-8). Estas formas permitem que uma caixa de simulação seja replicada em todas as suas dimensões, de forma periódica. Estas réplicas são idênticas à caixa construída, de forma que um movimento molecular em uma será idêntico ao movimento da mesma molécula na outra. Mas, agora, a face da caixa não está em contato com o vácuo, mas com solvente. E, caso uma molécula saia da caixa central, uma de suas imagens entrará pela face oposta, mantendo o número de moléculas constante. Isto representa uma continuidade da solução, nos aproximando de condições experimentais.

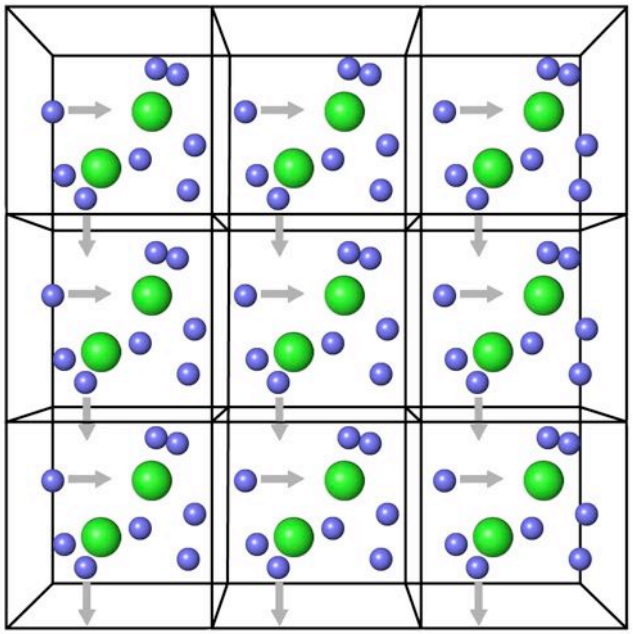


Figura 6-8: Representação das condições periódicas de contorno em uma simulação por DM. Somente a caixa central é simulada, enquanto que as réplicas garantem a continuidade do sistema, isto é, ausência de contato das moléculas com o vácuo.

Devemos, contudo, tomar cuidado para não definir uma caixa excessivamente pequena, buscando economizar custo computacional ao reduzir a quantidade de solvente excessivamente. Se a caixa for pequena demais, a proteína pode interagir com suas imagens, geradas pelas condições periódicas de contorno, criando uma situação artificial que provavelmente irá deturpar os resultados obtidos. É importante, assim, avaliar se o corte das interações não ligadas (ver adiante) é menor que a distância da proteína às suas imagens.

Equilibração

A ideia de equilibração de uma simulação por DM se refere à estabilização de suas propriedades, ou seja, que estas alcancem um estado de equilíbrio. Considera-se que, antes de estarem equilibradas, as propriedades em estudo apresentam variações ou comportamentos não representativos das situações de interesse. Assim, é necessário que o tempo de simulação seja suficientemente longo (tamanho da amostragem, ver adiante) para que as propriedades em estudo estejam adequadamente equilibradas. Na Figura 1-8, por exemplo, a simulação de um monômero de melitina demora em torno de 4 ns para se equilibrar.

Um dos motivos mais comuns para a necessidade de equilibração é devido ao uso de estruturas 3D derivadas de ambientes cristalinos, isto é, aquelas obtidas por cristalografia de raios-X. Este ambiente apresenta concentração de proteínas muito maior do que aquela observada, usualmente, nas condições biológicas de interesse, por vezes em estados oligoméricos não observados em condições biológicas. Assim, a remoção destes contatos e sua substituição por moléculas de água, acarretará em uma instabilidade inicial na simulação, envolvendo: 1) a perda de contatos cristalográficos, e 2) a formação de interações com moléculas de água.

Infelizmente, a busca por tempos de simulação "suficientemente longos" para equilibração das propriedades de interesse pode ser desafiadora, pois nem todas as propriedades moleculares equilibram a uma mesma velocidade. Por exemplo, a interação de uma proteína com o solvente equilibra usualmente mais rapidamente do que a perda ou a formação de estrutura 2ária. Estas, por sua vez, equilibram mais

rapidamente que o movimento de domínios em uma dada proteína.

Amostragem

A amostragem de uma simulação por DM se refere a quão bem ela é capaz de descrever o comportamento do sistema molecular em estudo. Idealmente, a amostragem de uma simulação deve ser longa o bastante para descrever os fenômenos de interesse. Contudo, a simulação de sistemas complexos como aqueles envolvendo biomoléculas frequentemente esbarra em amostragens ainda inalcançáveis em decorrência de seu elevado custo computacional.

A maneira mais simples de se entender a amostragem é considerando o tamanho da simulação em uma escala de tempo. Um maior tempo de simulação implica em uma maior amostragem. Contudo, diversos aspectos podem interferir neste entendimento. O aumento do número de moléculas e átomos no sistema aumenta o número de possíveis conformações a serem adotadas. Por outro lado, o uso de campos de força do tipo átomo unido ou ainda *coarse-grained*, ao reduzir o número de átomos, reduz o número de possíveis estados conformacionais a serem adotados pelo sistema, tornando assim a amostragem maior em uma mesma escala de tempo.

Tempo de integração

O cálculo de uma simulação por DM não gera informações contínuas, mas sim é dividida em pequenos passos, usualmente na escala de femtossegundos (fs). A sucessão destes passos dará origem ao nosso entendimento de trajetória, isto é, à evolução temporal do comportamento molecular na simulação realizada. O tamanho destas partes é o que chamamos de tempo de integração (Figura 7-8).

A definição de um valor apropriado para o tempo de integração está diretamente relacionada ao tamanho da amostragem da simulação e, por conseguinte, ao custo computacional da mesma. Conforme ilustrado na Figura 7-8, a descrição de uma determinada propriedade tempo-tempendente

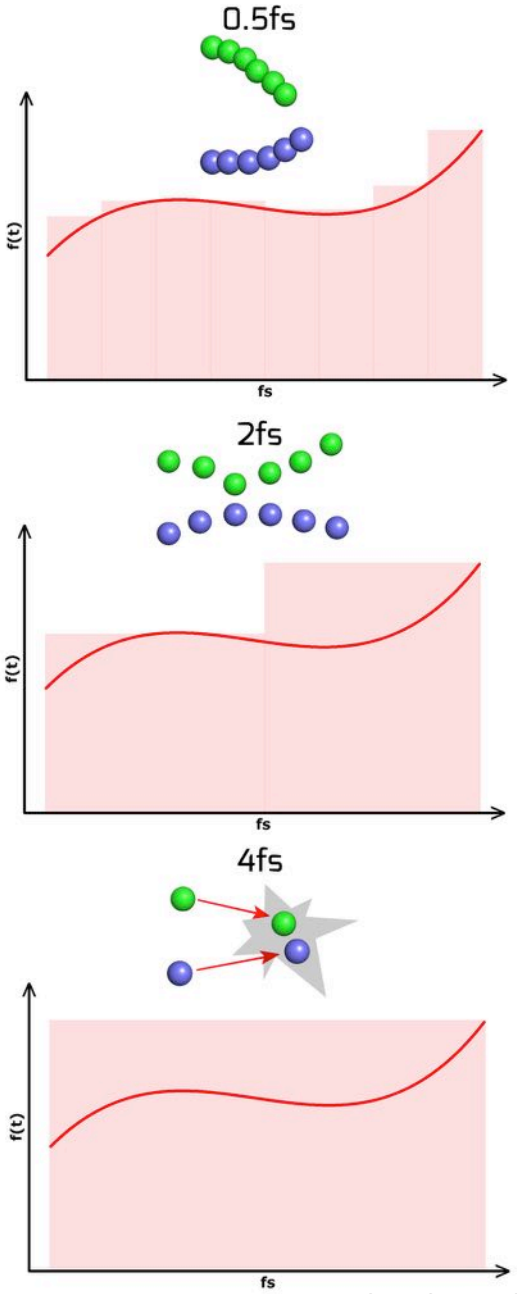


Figura 7-8: Representação do efeito de diferentes tempos de integração na amostragem de uma simulação por DM. Valores muito pequenos (0,5fs) descrevem fenômenos com maiores detalhes, mas mais lentamente. Valores muito grandes (4,0fs) apresentam menores custos computacionais, mas podem dar origem a instabilidades.

pode ser feita empregando-se diferentes valores de tempo de integração. Quanto maior este valor, menos passos de cálculo serão necessários à descrição do fenômeno e, por conseguinte, menor será o custo computacional associado. Quanto menor este valor,

mais passos serão necessários e, assim, maior o custo computacional. Infelizmente, o uso de tempos de integração muito elevados pode gerar instabilidades na trajetória, de forma que valores intermediários são usualmente empregados, no caso da Figura 7-8, 2fs.

Os valores de tempo de integração mais frequentemente empregados em simulações baseadas em campos de força atomísticos (isto é, todos os átomos são descritos) ou de átomo unido são 1fs, 2fs ou 5fs. O uso de 1fs é realizado quando as moléculas e suas ligações são tratadas como flexíveis durante a simulação, enquanto 2fs requerem o tratamento das ligações químicas como rígidas. Já para o uso de 5fs, toda a molécula é tratada como rígida (ou seja, ângulos e diedros não podem ser modificados), uma alternativa pouco utilizada no estudo de sistemas biológicos. Em algumas situações podem ser empregados tempos de integração menores que 1fs, mantida toda a flexiblidade da molécula. Em outros casos, como em simulações do tipo *coarse-grained*, tempos de integração de até 40fs.

Cálculo de interações não ligadas

Uma das partes mais custosas computacionalmente em simulações por DM envolve o cálculo das interações não ligadas, isto é, interações eletrostáticas (calculadas por termos de Coulomb) e de van der Waals (calculadas pelo potencial de Lennard-Jones). Para se ter uma ideia, enquanto o número de termos ligados (isto é, ligações, ângulos e diedros) é proporcional ao número de átomos, o número de interações não ligadas aumenta como função do quadrado do número de átomos do sistema. Assim, economizar custo computacional no cálculo destas interações representa uma significativa redução no custo da simulação como um todo. Como estas interações decrescem rapidamente em intensidade conforme dois átomos se distanciam no espaço, é possível realizar cortes nestas interações (*cut-off*). Em outras palavras, a partir da distância definida por estes cortes, nenhuma interação não ligada será calculada (Figura 8-8).

Por exemplo, consideremos dois possíveis raios de corte na simulação do soluto apresentado na Figura 8-8. O uso do raio **a** representaria um menor custo computacional, tendo em vista que nenhuma interação de Coulomb seria avaliada a partir desta distância. Já o uso do corte **b** traria um maior custo computacional, incluindo as interações entre o soluto e as moléculas na faixa cinza da figura. Contudo, ao reduzir o custo computacional, o corte **a** potencialmente implicará na perda de informações importantes, por ser muito próximo do soluto. Assim, a distância **b** seria preferível.

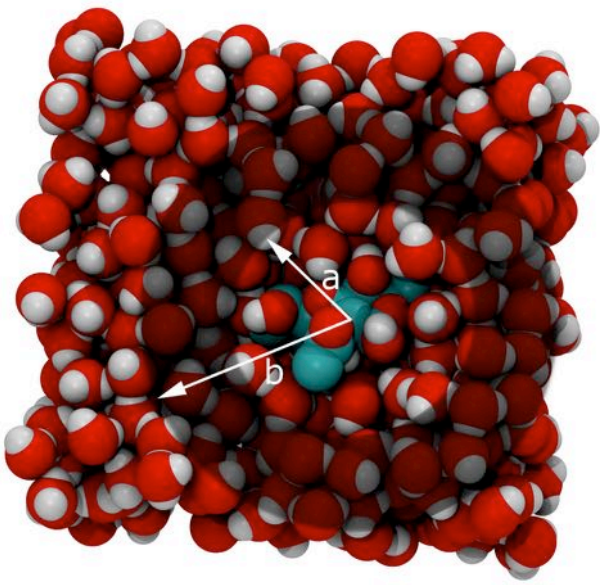


Figura 8-8: Representação de regiões de corte, a e b, a partir de um soluto, para cálculo de interações não ligadas.

A eliminação repentina da avaliação das interações não ligadas através de um *cut-off* pode gerar instabilidades ou erros na amostragem da simulação. Desta forma, estas interações a longas distâncias costumam ser descritas por outros tipos de métodos, como PME, Ewald ou Campo de Reação (*Reaction-Field*), dentre outros. Este tratamento é usualmente aplicado somente às interações de Coulomb, mais sensíveis a efeitos originados de cortes nas interações.

8.5. Estratégias de análise

Um dos maiores desafios em um estudo baseado em DM frequentemente reside mais na análise e interpretação dos resultados obtidos do que no preparo do sistema. De fato, simulações de proteínas em água podem gerar facilmente muitas dezenas de gigabytes de dados. Como retirar informações destas trajetórias, quais informações retirar e como interpretar estas informações, no contexto do

assunto em estudo, envolvem muitas vezes mais tempo do que a simulação computacional em si.

Os tipos de análises a serem empregadas estarão intrinsecamente relacionados à natureza do problema em estudo. Por exemplo, se estamos estudando uma proteína tentando mimetizar o ambiente nativo da mesma, em princípio, ela não pode se desnaturar durante a simulação. Por outro lado, o estudo de membranas elimina esta preocupação mas nos traz a necessidade de avaliar as propriedades dos lipídeos enquanto imersos num fluido. Adicionalmente, dados prévios sobre características estruturais e/ou funcionais das moléculas em estudo, obtidos tanto por métodos computacionais quanto por outras ferramentas experimentais são fundamentais na concepção, preparo, execução e análise de estudos por DM. Esta é, fundamentalmente, a razão pela qual este livro traz em si diversos métodos experimentais.

Neste momento, a adequação da amostragem às propriedades em estudo assume importância fundamental. Se buscamos estudar o movimento de domínios de uma proteína, simulações de dezenas de nanossegundos não serão suficientes, requerendo potencialmente tempos próximos de microssegundos, possivelmente inviabilizando o estudo por DM. De forma semelhante, a observação do enovelamento de proteínas por DM é impraticável na grande maioria dos casos, salvo em pequenas proteínas ou peptídeos, de qualquer forma, requerendo no mínimo centenas de nanossegundos. Por outro lado, reorientação ou refinamento de cadeias laderais de resíduos de aminoácidos ou de ligantes em complexos fármaco-receptor podem ser observados frequentemente em algumas dezenas de nanossegundos.

As análises de simulações por DM devem, preferencialmente, ser realizadas observando propriedades de complexidade crescente (o que costuma estar associado ao tempo requerido à equilibração desta propriedade). Assim, as primeiras propriedades a serem avaliadas são normalmente a pressão (no caso de simulações NPT, mais comuns em sistemas biológicos), o volume (no caso de simulações NVT), a densidade e a energia total do sistema. Todas estas propriedades devem alcançar um patamar estável, paralelo ao eixo x (tempo). Pode-se observar alguma variação no início da simulação mas, em seguida, devem atingir este patamar e se manter neste nível ao longo da simulação. Estas costumam ser propriedades de rápida equilibração em simulações por DM.

Garantidas estas propriedades, podemos passar à análise de aspectos mais complexos, como do comportamento da estrutura proteica ao longo da simulação. Neste grupo, as ferramentas mais comumente empregadas incluem o RMSD, o RMSF, o raio de giro, distâncias entre átomos ou grupamentos e a evolução do conteúdo de estrutura 2ária como função do tempo.

O RMSD (do inglês *root mean square deviation* ou desvio quadrático médio) é uma das principais estratégias de análise empregadas no estudo por DM de proteínas (Figura 9-8A). Indica o quanto a estrutura da proteína de interesse se modifica ao longo de uma simulação, em relação à estrutura de partida, normalmente cristalográfica. Assim, é usual que haja um aumento progressivo no RMSD de uma proteína, partindo de 0, até um patamar, o que pode indicar a equilibração do sistema. Este patamar pode variar em função das características da proteína mas, como um ponto de partida, podemos considerar um valor em torno de 3 Å quando todos os átomos do sistema são empregados na medida. Valores acima deste podem sugerir movimentos maiores de alças, em relação ao cristal, ou perda de estrutura 2ária, enquanto valores menores tendem a indicar sistemas mais semelhantes à referência cristalográfica.

Uma consideração importante quando realizamos análises de RMSD se refere ao fato de que esta análise oferece uma medida média de um conjunto de átomos, selecionados para a análise. Se todos os átomos de uma proteína são considerados, como no exemplo acima, os valores observados trazem consido influências de diferentes regiões da proteína. Por exemplo, normalmente conjuntos de hélices α se modificam menos durante uma simulação do que regiões de alças. Caso façamos uma análise de RMSD separada para estas regiões, veremos hélices α com valores menores e alças com valores maiores do que aqueles considerando

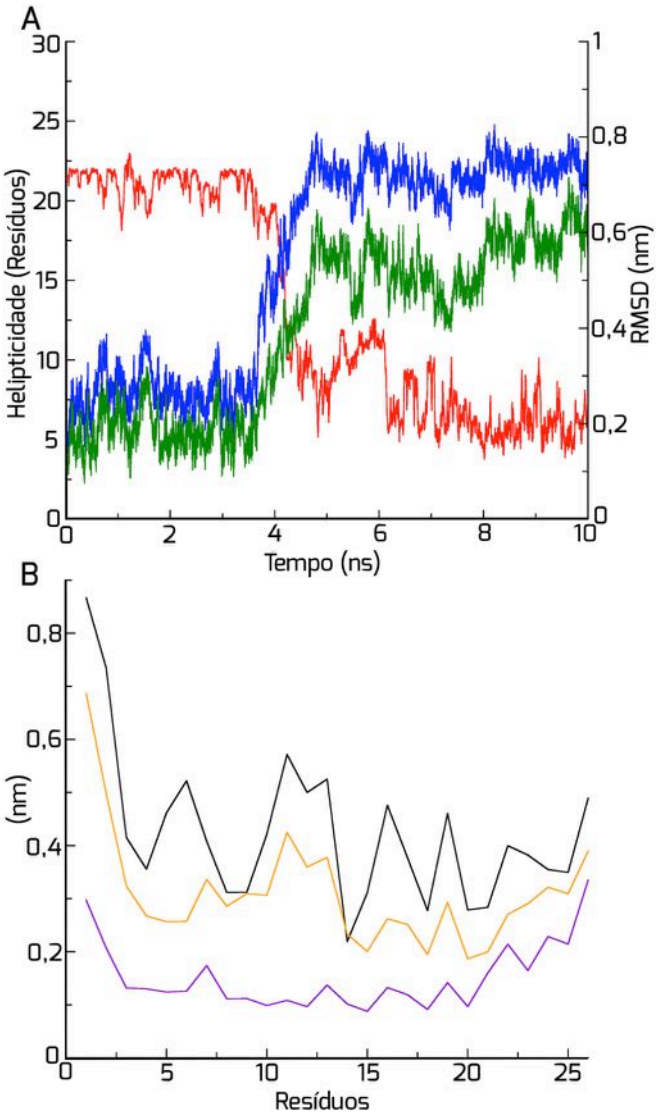


Figura 9-8: A) Helipticidade (vermelho) e RMSD, e B) RMSF para a melitina. O RMSD foi calculado para toda a proteína (azul) e para o esqueleto peptídico (verde). Já o RMSF foi medido como média para toda a trajetória (preto), para os primeiros 3 ns (roxo) e para os últimos 5 ns (laranja).

ambas regiões juntas. Processo similar ocorre caso consideremos todos os átomos do sistema (maior RMSD) ou simplesmente o esqueleto peptídico (menor RMSD) (Figura 9-8A).

Na análise por RMSD, todo resultado obtido irá depender da geometria de partida da simulação, usualmente cristalográfica. O RMSF (do inglês *root mean square fluctuation* ou flutuação quadrática média), em contrapartida, não apresenta esta dependência, mas descreve a variação da posição dos átomos (ou resíduos de aminoácidos) durante a simulação, indicando a flexibilidade do sistema (Figura 9-8B). Valores maiores de RMSF serão, portanto, usualmente observados para alças, e valores menores para hélices α. Por outro lado, regiões de hélices α apresentando valores elevados de RMSF podem estar passando, durante a simulação, por perda de sua estrutura 2ária.

Enquanto o RMSD apresenta um valor médio, a cada passo da simulação, para todos os átomos do sistema, o RMSF apresenta um valor médio, para cada átomo ou resíduo (usualmente mais útil para proteínas), ao longo de todos os passos da simulação. Assim, valores de RMSF para toda a trajetória podem diferir, por exemplo, daqueles observados no início e/ou no final da simulação (Figura 9-8B).

Ainda, ao observarmos o quanto uma proteína muda sua forma 3D em relação ao cristal ou a flexibilidade de cada resíduo ao longo da simulação, não temos informações diretas sobre o comportamento dos elementos de estrutura 2ária da proteína. Um valor de RMSD elevado pode tanto sugerir a desnaturação de uma hélice quanto uma reorientação da mesma que, contudo, pode se manter enovelada. Da mesma maneira, um resíduo muito flexível (conforme observado pelo RMSF) não necessariamente será encontrado somente em alças. Para tal, devemos empregar análises específicas capazes de indicar como a estrutura 2ária da proteína se comporta na simulação por DM.

Conforme observado no Capítulo 2, a definição da estutura 2ária não é algo tão simples e direto como possa parecer. Existe mais de uma forma de definir hélices e folhas, e diferentes estratégias podem oferecer resultados distintos. Por exemplo, o programa DSSP descreve a estrutura 2ária a partir do padrão de ligações de hidrogênio na sequência polipeptídica. À informação relacionada a interações por ligação de hidrogênio o programa STRIDE adiciona parâmetros torsionais relacionados ao esqueleto peptídico.

Outro aspecto importante quanto à análise do comportamento da estrutura 2ária diz respeito à escala de tempo na qual hélices e fitas se enovelam. Enquanto hélices usualmente se enovelam numa escala de tempo de centenas de nanossegundos, simulações de poucas dezenas de nanossegundos terão dificuldades em prever estes fenômenos. O caso de fitas é ainda mais complexo, exigindo escalas de tempo uma ordem de grandeza superiores.

Uso de estatística

Embora seja prática corriqueira, mesmo obrigatória, na grande maioria dos métodos experimentais empregados no estudo de sistemas biológios, o uso de métodos estatísticos não é, ainda, comum na análise de resultados obtidos em simulações por DM. Isto se deve ao fato de que, em uma mesma simulação, são normalmente gerados centenas de milhares ou mesmo milhões de dados para uma mesma variável (tamanho da simulação dividido pelo tempo de integração). O grande *n* assim obtido tenderá a tornar estatisticamente significativa mesmo variações bem pequenas nas propriedades de interesse.

Com a redução no custo dos computadores e aumento em sua velocidade, assim como na melhoria dos programas disponíveis, uma nova abordagem vem se apresentando, aproximando a análise de simulações por DM de estudos experimentais convencionais. Trata-se da realização de múltiplas simulações para um mesmo sistema. Assim, a informação a ser empregada nas análises é a média da informação gerada nas diversas simulações.

8.6. Limitações atuais da DM

Como toda técnica experimental, simulações por DM possuem limitações importantes que devem ser conhecidas pelos seus usuários de forma a reduzir a chance de interpretações equivocadas dos resultados obtidos.

Uma consequência direta da realização de cálculos baseados na mecânica molecular, ou seja, empregando campos de força, é a ausência de elétrons. Este tipo de cálculo não considera os elétrons e, por conseguinte, os resultados obtidos apresentam limitações em lidar com fenômenos envolvendo elétrons diretamente. Assim, simulações por DM não são capazes, por exemplo, de descrever reações químicas, como as observadas na ação de enzimas ou em processos de oxidação e redução. Uma alternativa recente para esta limitação envolve métodos denominados híbridos entre a mecânica molecular e a mecânica quântica.

Simulações por DM apresentam grande dificuldade em descrever a energia livre de Gibbs associada a eventos moleculares. Portanto, informações sobre constantes de equilíbrio, constantes catalíticas ou afinidades entre moléculas não são usualmente acessíveis, com precisão, através destas técnicas. Embora diversas técnicas gerem estimativas de energia livre associadas à DM, como a perturbação da energia livre, o *linear interaction energy* e a metadinâmica, cada uma possui suas próprias limitações, dificultando seu uso amplo em estudos por DM.

Por fim, e não menos importante, temos a dificuldade em obter amostragens compatíveis com fenômenos observáveis em experimentos ou fisiologicamente. Mesmo nos maiores centros de supercomputação do mundo, ainda não chegamos, na grande maioria dos casos, em escalas de tempo compatíveis com o comportamento de proteínas em soluções biológicas. Por isso, devemos ter em mente que os resultados obtidos, por mais confiáveis e corretos que sejam, não necessariamente representam, estatisticamente, fenômenos medidos em solução.

8.7. E outras biomoléculas?

A maior parte da literatura, seja em livros seja em artigos, se refere ao estudo de proteínas. Ácidos nucleicos, membranas e carboidratos vêm sendo estudados com menos frequência, comparativamente, ao longo dos anos. Embora possa se justificar esta diferença em decorrência do fato de que as proteínas são as moléculas efetoras da informação genética, esta não é a única justificativa, tampouco proteínas são os únicos compostos biológicos importantes para a manutenção da vida.

O estudo de moléculas de DNA, por exemplo, vem ganhando importância com o desenvolvimento de compostos capazes de interagir, seletivamente, com regiões específicas do DNA, como é o caso dos agentes antineoplásicos. Enquanto moléculas de DNA apresentam estruturas mais ou menos bem definidas, moléculas de RNA são extremamente versáteis e complexas conformacio-

nalmente, a cada momento se mostrando como capazes de atuarem em mais fenômenos biológicos. Valorização semelhante vem sendo observada para membranas e carboidratos que, progressivamente, deixam de ter papéis passivos, simplesmente estruturais, passando a desempenhar papéis ativos, sinalizando diretamente múltiplas respostas em organismos.

Assim, a construção de modelos computacionais para o estudo de biomoléculas deve incluir o máximo de propriedades importantes ao desenvolvimento normal de suas funções, em condições nativas. Uma proteína inserida em membrana irá exigir a inclusão da membrana nas simulações, da mesma maneira que uma glicoproteína irá demandar a inclusão da parte sacarídica em seu estudo.

Do ponto de vista da disponibilidade de parâmetros de campos de força, diferentes classes de biomoléculas apresentam diferentes disponibilidades de parâmetros. Por isso, é importante considerar todos os componentes do sistema molecular quando da escolha do campo de força a ser empregado. Se a nossa molécula em estudo é uma glicoproteína, não adianta empregar um campo de força excelente para carboidratos se o mesmo não possui parâmetros para o estudo de proteínas.

Atualmente, os principais campos de força são capazes de descrever a grande maioria das classes de biomoléculas. Originalmente, no entanto, o campo de força AMBER foi desenvolvido para o estudo de ácidos nucleicos e proteínas, o CHARMM para proteínas, o GROMOS para lipídeos e o OPLS para líquidos e solventes. Com o passar do tempo, cada um desses parâmetros foi sendo aprimorado focando em diferentes biomoléculas, de forma que, hoje, alguns são empregados com maior freqüência para determinados sistemas por melhor descreverem suas propriedades (estruturais, conformacionais ou físico-químicas).

No caso específico de proteínas, os campos de força citados acima descrevem de forma semelhante sua estrutura, conformação e dinâmica. No caso de lipídeos, a maior parte dos estudos envolve os campos de força CHARMM e GROMOS, embora o último ofereça um ganho de velocidade de até nove vezes devido a sua natureza de átomo unido.

Para ácidos nucleicos, os campos de força mais amplamente utilizados são o AMBER e o CHARMM, tanto para DNA quanto para RNA.

A parametrização de carboidratos, por sua vez, está imersa em desafios devido à sua elevada complexidade estrutural e conformacional, de forma que uma sucessão de novos parâmetros vêm sendo desenvolvida.

Por fim, o grupo de compostos mais desafiadores com relação à disponibilidade prévia de parâmetros envolve os fármacos ou moduladores da função proteica que não estão sob uso terapêutico (genericamente chamados de ligantes). Em decorrência de sua variedade e originalidade química, é extremamente difícil ter, de antemão, parâmetros próprios à sua descrição. Assim, é frequente a necessidade de parametrização dos ligantes em estudo, seguindo as características do campo de força em uso.

Embora os quatro campos de força citados possuam parâmetros para um amplo espectro de grupamentos funcionais, para casos específicos ferramentas como o servidor PRODRG (para o GROMOS) e o GAFF (para o AMBER) são capazes de gerar parâmetros, com graus variados de precisão, que podem ser empregados no estudo de compostos orgânicos em geral.

8.8. Conceitos-chave

Amostragem: refere-se à descrição do comportamento conformacional de uma dada molécula em uma simulação.

Campo de força: conjunto de equações que descreve o comportamento molecular em cálculos de mecânica molecular. É ajustado para cada tipo de molécula a ser estudado.

Campo de força *all atom* (todos os átomos): considera todos os átomos do sistema explicitamente.

Campo de força *united atom* (átomo unido): transforma grupos CH, CH_2 e CH_3 em uma única partícula ou pseudoátomo, reduzindo o número de átomos a ser descrito.

Grupos CH de anéis aromáticos são descritos explicitamente.

Campo de força *coarse-grained*: transforma grupos de átomos em partículas, reduzindo o custo computacional ainda mais do que campos de átomo unido.

Condições periódicas de contorno: condição empregada em simulações por DM que impede o contato das moléculas do sistema com o vácuo, representando o sistema de forma periódica.

Cut-off: representa um corte no cálculo de interações não ligadas, reduzindo o custo computacional do cálculo. A partir da distância definida, estas interações não são mais calculadas.

Diedro próprio: ângulo formado por quatro átomos ligados em sequência. Os primeiros três átomos definem um plano, enquanto os últimos três definem outro plano. O ângulo formado por estes dois planos é o diedro.

Diedro impróprio: ângulo formado por quatro átomos que não estão ligados em sequência. É empregado para garantir, por exemplo, a quiralidade de átomos e a planaridade de anéis.

Dinâmica molecular: tipo de cálculo em que as coordenadas dos átomos variam como função do tempo.

Equilibração: período em que propriedades de uma simulação de DM demoram para atingir um patamar estável. Diferentes propriedades podem requerer tempos diferentes para equilibrar.

Mecânica molecular: tipo de cálculo em que o comportamento molecular é descrito a partir das equações da mecânica clássica ou de Newton.

Mecânica quântica: tipo de cálculo em que o comportamento molecular é descrito a partir das equações da mecânica quântica.

Minimização de energia: tipo de cálculo em que a energia do sistema é reduzida através da otimização das posições atômicas.

Modelo de água explícito: modelo no qual as moléculas de água são descritas pela presença física de seus átomos.

Modelo de água implícito: modelo no qual as moléculas de água são descritas sem a presença física de seus átomos.

NPT: condição de simulação na qual o número de partículas, a pressão e a temperatura permanecem constantes.

NVT: condição de simulação na qual o número de partículas, o volume e a temperatura permanecem constantes.

Tempo de integração: tamanho do passo empregado em cálculos de DM.

Transferabilidade: em um campo de força, se refere à manutenção das propriedades de um grupamento funcional em diferentes moléculas. Assim, uma hidroxila alcoólica de um resíduo de serina terá os mesmos parâmetros que a mesma hidroxila em uma treonina.

8.9. Leitura recomendada

MORGON, Nelson H.; COUTINHO, K. ***Métodos de Química Teórica e Modelagem Molecular***. São Paulo: Editora Livraria da Física, 2007.

LEACH, Andrew R. ***Molecular Modelling Principles and Applications***. 2.ed. Essex: Pearson Education Limited, 2001.

SANT'ANNA, Carlos Maurício R. Glossário de termos usados no planejamento de farmacos (recomendações da IUPAC para 1997). ***Quim. Nova***, 25, 505-512, 2002.