

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Curso de Biomedicina

Estágio em Pesquisa e Trabalho de Conclusão de Curso em Biomedicina

**Desenvolvimento de um modelo computacional aplicado à família das serino  
endopeptidases**

Kaluti Rossi De Martini Moraes

Porto Alegre

Dezembro/2013

Kaluti Rossi De Martini Moraes

**Desenvolvimento de um modelo computacional aplicado à família das serino  
endopeptidases**

Trabalho de conclusão de curso de  
graduação apresentado ao Instituto de  
Ciências Básicas da Saúde da Universidade  
Federal do Rio Grande do Sul, como  
requisito parcial para obtenção do título de  
Bacharel em Biomedicina.

Orientador: Prof. Dr. Walter Filgueira de  
Azevedo Jr.

Porto Alegre

Julho/2013

# Índice

1. Introdução expandida	4
2. Breve justificativa metodológica	6
3. Base teórica	9
3.1. Dinâmica molecular	9
3.2. Minimização de energia	13
4. Artigo	17
4.1. Resumo	17
4.2. Introdução	18
4.3. Metodologia	19
4.4. Discussão	21
5. Conclusão e perspectivas	27
6. Referências	28

## 1. Introdução expandida

O papel da computação no tratamento de investigações científicas, em especial na Biologia, Biofísica e Bioquímica, ao longo das últimas décadas, tem sido cada vez mais relevante (Van Gunsteren et al., 2006). O contínuo aumento da capacidade de processamento tem viabilizado a análise, comparação e caracterização de grandes e complexos conjuntos de dados obtidos de experimentação em sistemas biológicos. Houve uma transição entre o caráter passivo e ativo da computação, do ponto de vista experimental, visto que atualmente esta não lida apenas com o tratamento dos dados experimentais, mas também é capaz de atuar na geração destes, através da construção de modelos para os sistemas biológicos de interesse e desenvolvimento de simulações utilizando os mesmos.

Um dos problemas centrais no desenvolvimento de novos fármacos é a construção de modelos computacionais capazes de prever a afinidade de um dado ligante por uma proteína alvo, tendo como base a análise das coordenadas atômicas do complexo proteína-ligante. Tal análise pode fazer uso somente da informação estrutural estática, disponível na estrutura cristalográfica de tais complexos. Contudo, tal abordagem, é ingênua, pois a interação de um dado ligante com uma proteína é um processo de reconhecimento molecular, onde aspectos dinâmicos devem ser considerados para que o sistema apresente bases físicas. Portanto, é grande o interesse em compreender as forças que estabilizam uma determinada interação proteína-ligante, para o desenvolvimento racional de novos fármacos. As principais forças moleculares que governam a interação proteína-ligante são: ligações de hidrogênio, interações hidrofóbicas e de van der Waals. Quando há o propósito de se estudar um determinado modelo para um sistema biológico de interesse, é de fundamental importância avaliar algumas características inerentes ao mesmo, tais como os graus de liberdade associados, a escala temporal dos fenômenos e propriedades físicas a serem consideradas no modelo (Van Gunsteren et al., 2006).

É provável que a relevância da contribuição de cada uma das forças na estabilidade da interação proteína - ligante seja distinta conforme lidamos com diferentes famílias de proteínas, mesmo porque diferentes famílias geralmente

implicam em distintos padrões moleculares caracterizando o sítio de ligação. Nesse sentido, o presente projeto de pesquisa visa a simulação de dinâmica molecular de complexos binários oriundos de uma família de proteínas em especial, pois a partir das simulações computacionais será possível acessar aspectos dinâmicos relevantes para a afinidade dos ligantes neste contexto. Os resultados devem convergir para a proposição de um modelo matemático que seja capaz de prever a afinidade de diferentes ligantes, utilizando uma função polinomial empírica para estimativa da interação.

A base de dados PDBBind (Wang et al., 2004) apresenta informação estrutural de milhares de estruturas cristalográficas, para os quais dados de afinidade do ligante pela proteína estão disponíveis (constante de inibição). Tal facilidade permite que usemos as estruturas dos complexos binários proteína-ligante para a simulação de dinâmica molecular. Ao avaliar a base de dados citada, foi observada a abundância de resultados para uma família de proteínas em especial, no que diz respeito aos critérios de seleção adotados para o tratamento do problema em estudo. Essa família corresponde à classificação enzimática EC 3.4.21, conjunto de serino endopeptidases, da qual podemos destacar a tripsina, além da elastase, fator de ativação Xa, urocinase e trombina. Desta forma se justificam os objetivos do presente trabalho: simulação de dinâmica molecular para conjuntos de complexos proteína-ligante, para os quais há informação experimental não só para estrutura cristalográfica do complexo, como também para a afinidade do ligante pela proteína; preparação de uma base de dados estruturais com os modelos dos complexos proteína-ligantes; utilização desta base de dados para elaboração de modelos computacionais para previsão da afinidade de ligantes por proteínas da família da tripsina.

## 2. Breve justificativa metodológica

A dinâmica molecular utiliza as leis da física clássica para prever as estruturas e propriedades das moléculas. Há diversos métodos que realizam tais predições através da dinâmica molecular, cada qual é caracterizado pelo seu campo de força em particular (Bartels, 2000).

Um campo de força consiste em um conjunto de equações que definem como a energia potencial de uma molécula varia de acordo com as localizações de seus átomos. Além destas equações se faz necessária uma série de definições para os diferentes tipos de átomos, uma vez que as características que um elemento manifesta dependem do contexto químico em que o mesmo se encontra. Portanto, diferentes tipos de átomos (inclusive para um mesmo elemento químico) irão descrever um comportamento distinto dependendo do contexto em questão, como por exemplo, diferentes hibridizações, cargas e vizinhos químicos (Hess et al., 2013).

É imprescindível a presença de conjuntos de parâmetros que irão adequar as equações e tipos de átomos para que haja correspondência entre o modelo *in silico* e os dados experimentais. Inerentes a esses parâmetros estão as constantes de força, que consistem em valores usados nas equações para relacionar características atômicas com suas respectivas componentes de energia e dados estruturais, tais como: comprimento de ligações químicas e seus respectivos ângulos.

Cálculos envolvendo dinâmica molecular não tratam os elétrons no sistema molecular de forma explícita, mas sim atuam com base em interações entre os núcleos atômicos. Os efeitos decorrentes da presença eletrônica estão implícitos nos campos de força através de parametrizações. Esta aproximação reduz de forma relevante o custo computacional dos cálculos na dinâmica molecular, viabilizando na prática o tratamento de grandes sistemas, contendo alguns milhares de átomos, por exemplo o caso de macromoléculas biológicas como proteínas.

A respeito de algumas limitações deste método, é bem reconhecido que cada campo de força alcança bons resultados apenas para uma limitada classe de moléculas, de acordo com as quais o mesmo foi parametrizado. Portanto, nenhum

campo de força pode ser usado indistintamente para todos os sistemas moleculares de interesse. Além de que, a aproximação assumida em relação aos elétrons implica que o método não pode tratar de problemas onde a influência eletrônica seja preponderante. Logo, não há como descrever processos em que haja formação ou quebra de ligações químicas. As propriedades moleculares que dependam intrinsecamente da distribuição eletrônica não são reproduzíveis através da dinâmica molecular.

A aplicação da dinâmica molecular no presente estudo, que visa investigar a interação proteína-ligante, consiste em um excelente modelo computacional, visto que neste contexto não há quebra nem formação de ligações químicas. A energia de interação neste caso é influenciada essencialmente por forças de ação à distância, tais como dipolo-dipolo e íon-íon, ligações de hidrogênio e forças de dispersão (Sant'Anna, 2009). As simulações de sistemas contendo um número grande de átomos (como no caso das proteínas), do ponto de vista de capacidade de processamento, são viáveis no que tange à aplicação da dinâmica molecular.

A definição de um campo de força é um aspecto inerente à aplicação do método de dinâmica molecular, apesar do reconhecimento da impossibilidade de um único campo ser capaz de contemplar uma gama ilimitada de classe de moléculas (Hess et al, 2013). Portanto, para o presente trabalho foi utilizado um campo de força recente, parametrizado em especial para o tratamento de proteínas. Dessa forma, surge um problema teórico pois se faz necessária a inclusão de parâmetros de topologia para os ligantes, uma vez que os mesmos não são cobertos pela referência padrão do campo de força escolhido. Com a finalidade de contornar essa carência na parametrização dos ligantes, a aplicação de métodos de estrutura eletrônica aparece como uma alternativa interessante, visto que os ligantes apresentam poucas dezenas de átomos.

Métodos de estrutura eletrônica - *eletronic structure methods* - são métodos que utilizam as leis da mecânica quântica como base para seus cálculos. Estados quânticos são aqueles em que a energia e outras propriedades relacionadas podem ser obtidas solucionando a equação de Schrödinger:  $H\Psi=E\Psi$ , que consiste em aplicar um funcional sobre a função de onda correspondente ao sistema. Contudo, para quaisquer sistemas que não sejam mínimos, soluções

exatas para a equação de Schrödinger não é computacionalmente prática. Logo, essa classe de métodos é caracterizada pelo uso de diversas aproximações matemáticas para suas soluções (Frisch et al., 2009).

O modelo do funcional de densidade - *density functional theory* (DFT) - dentre os métodos de estrutura eletrônica disponíveis, é aquele que apresenta um ótimo retorno quando suas previsões são comparadas ao seu custo computacional (Mourik et al., 2002). Seu diferencial está no tratamento dos efeitos de correlação eletrônica, devido ao fato dos elétrons reagirem a presença uns dos outros no contexto molecular. Os cálculos feitos através da abordagem de Hartree-Fock - base teórica que implica em menor custo computacional dentre aquelas disponíveis para o método *ab initio* - consideram que cada elétron reage a uma densidade eletrônica. Portanto o DFT é um método interessante, visto que é capaz de prover resultados com os benefícios de alguns métodos *ab initio* (de maior custo computacional) requerendo a exigência computacional do método de Hartree-Fock (Mourik et al., 2002).

Considera-se que a energia de um conjunto de elétrons sob influência de um campo externo é um funcional único da densidade eletrônica. Esta dependência aparece em dois termos da energia eletrônica, designados por funcional de troca e funcional de correlação (Hohenberg et al., 1964). Alguns funcionais foram desenvolvidos a partir da mecânica quântica fundamental e outros a partir da parametrização de funções que melhor reproduzem resultados experimentais. Desse modo, pode-se dizer que há versões *ab initio* e semi-empíricas do modelo DFT. Atualmente, um dos modelos mais utilizados é o modelo do funcional de troca híbrido de 3 parâmetros de Becke e do funcional de correlação de Lee-Yang-Parr (B3LYP), devido à qualidade dos seus resultados, particularmente para moléculas orgânicas (Sant'Anna, 2009).

### 3. Base teórica

#### 3.1. Dinâmica molecular

A consideração no que diz respeito às condições inerentes à descrição de um sistema de  $n$  partículas, no qual as suas constituintes têm seu movimento restrito umas em relação às outras, é de fundamental importância. A Terceira Lei de Newton desempenha um papel importante na dinâmica de um sistema de partículas em decorrência das forças internas entre as partículas do sistema (Thornton, 2011). A relevância desta lei é justificada na independência da natureza das forças internas de um sistema. Pois, enquanto estas seguirem a forma "fraca" da Terceira Lei de Newton, o centro de massa deste sistema se move, sob ação da força total externa, como se fosse uma única partícula de massa igual à massa total do sistema (somatório das massas das  $n$  partículas). A denominada forma "fraca", citada anteriormente, considera que as forças exercidas por duas partículas entre si são iguais em magnitude, apesar de ocorrerem em sentidos opostos (Thornton, 2011). No caso da aplicação no presente estudo, considera-se que a força total externa aplicada sobre o sistema é nula, logo devem ser avaliadas certas características para este contexto em especial, que dizem respeito à distribuição da quantidade de movimento linear e energia.

A quantidade de movimento linear do sistema é a mesma de uma única partícula de massa  $M$  (somatório das massas das  $n$  partículas) localizada na posição do centro de massa e se movendo da mesma maneira deste. Como consequência da condição de um sistema livre de forças externas, a quantidade de movimento linear total deste sistema é constante e igual àquela do centro de massa, aspecto coerente com a lei da conservação da quantidade de movimento linear de um sistema.

A energia cinética total do sistema é igual à soma da energia cinética de uma partícula de massa  $M$  movendo-se com a velocidade do centro de massa e da energia cinética de movimento das partículas individuais em relação ao centro de massa. Para um sistema conservativo, onde todas as forças envolvidas são deriváveis de potenciais que não dependem explicitamente do tempo, a sua energia total é uma constante.

A aplicação para o presente estudo se dá sobre um sistema cujas partículas constituintes são restritas na manutenção das suas posições relativas, visto que os átomos estão vinculados aos seus vizinhos imediatos através de ligações químicas, para as quais não há tratamento teórico quanto à possibilidade de quebra ou formação durante as simulações (limitação conhecida da dinâmica molecular). Neste caso, para qualquer processo que envolva o sistema, a energia potencial interna permanece constante e, logo, pode ser ignorada no cálculo da energia potencial total do sistema. Sua influência seria simplesmente redefinir o zero para a posição da energia potencial, mas, efetivamente, essa posição possui valor arbitrário, uma vez que somente a diferença na energia potencial agrega valor físico à interpretação do sistema. Portanto, o valor absoluto da energia potencial é uma quantidade arbitrária e irrelevante para correlação física com o comportamento esperado na realidade (Thornton, 2011).

Em um complexo binário, correspondente à interação entre um ligante e uma proteína, cada átomo, apesar de apresentar massa, é considerado como um corpo pontual, cujo movimento é influenciado pelas forças de interação com os átomos da vizinhança. Este consiste em um sistema molecular capaz de ser descrito pelas equações do movimento da Física Clássica (McCammon, 1987). Apesar desta abordagem consistir em uma aproximação da dinâmica real, ela é capaz de prover uma descrição detalhada do comportamento do sistema molecular em muitos casos (Field, 2007).

Para o presente trabalho, será considerado sistema molecular como aquele descrito pelas coordenadas atômicas de um complexo binário (interação proteína - ligante). Cada átomo  $i$  na estrutura é representado pelo vetor  $\mathbf{r}_i$ , apresenta velocidade  $\mathbf{v}_i$ , massa  $\mathbf{m}_i$ , está submetido a uma força  $\mathbf{F}_i$  e apresenta momento linear  $\mathbf{p}_i$ . O Hamiltoniano clássico para este sistema pode ser descrito como a soma das energias potencial e cinética e apresenta a seguinte forma:

$$H(\mathbf{p}_i, \mathbf{r}_i) = \sum_{i=1}^N \frac{1}{2m_i} p_i^2 + V(\mathbf{r}_i)$$

onde  $V$  se refere à energia potencial e o somatório abrange todos os  $N$  átomos do sistema. Este Hamiltoniano é uma função de  $6N$  variáveis independentes, sendo que dentre elas há  $3N$  para as coordenadas e  $3N$  para os momenta atômicos.

Equações do movimento para as variáveis da equação anterior são obtidas utilizando as chamadas canônicas hamiltonianas:

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial r_i}$$

Desde que o primeiro termo do hamiltoniano clássico para o sistema não dependa explicitamente de  $\mathbf{r}_i$ , as derivadas com respeito ao tempo para  $\mathbf{p}_i$  e  $\mathbf{r}_i$  são respectivamente:

$$\frac{dp_i}{dt} = -\frac{\partial V}{\partial r_i} = F_i \quad \frac{dr_i}{dt} = \frac{\partial H}{\partial p_i} = \frac{p_i}{m_i}$$

Observamos que as hamiltonianas para o sistema consistem em um conjunto de  $2N$  equações diferenciais de primeira ordem que, ao serem resolvidas, apresentam  $2N$  variáveis  $\mathbf{p}_i$  e  $\mathbf{r}_i$  como funções do tempo, sendo que se fazem necessárias  $2N$  constantes arbitrárias para garantir resolução particular para este conjunto (correspondem às condições iniciais do sistema em questão). Uma equação diferencial de segunda ordem pode ser obtida ao se utilizar a segunda lei da mecânica clássica ( $\mathbf{p}_i = m_i \cdot \mathbf{v}_i$ ), considerando que os átomos do sistema não apresentam alteração de massa, é possível observar que:

$$\begin{aligned} \frac{dp_i}{dt} &= m_i \frac{dv_i}{dt} = -\frac{\partial V}{\partial r_i} = F_i \\ m_i \frac{d^2 r_i}{dt^2} &= F_i \end{aligned}$$

Em termos práticos, resolver as equações citadas anteriormente para cada um dos átomos do sistema é o objetivo das simulações de dinâmica molecular. Existem diferentes algoritmos para lidar com os processos de integração inerentes às soluções destas equações. Em geral, estes partem de valores definidos inicialmente para as coordenadas e momenta atômicos e promovem a integração das equações de movimento por um intervalo de tempo definido (passo de integração, que corresponde a uma das iterações), para o caso da simulação de proteínas este intervalo é tipicamente da ordem de femtosegundos (Field, 2007).

Se a posição de uma partícula é dada por uma função que depende do tempo  $x(t)$  e apresenta valor conhecido neste caso em especial, sabe-se que a nova

posição após um determinado intervalo de tempo pode ser definida por uma típica expansão em séries de Taylor:

$$x(t + \Delta t) = x(t) + \frac{dx(t)}{dt} \Delta t + \frac{d^2x(t)}{dt^2} \frac{\Delta t^2}{2} + \dots$$

A partir desta expansão, se faz necessária a avaliação de um ponto de corte para as derivadas de mais altas ordens, desta forma estamos assumindo uma ordem de aproximação para tornar viável sua aplicação nos cálculos de dinâmica molecular. Para o presente estudo, o programa utilizado para gerar as simulações aplica um algoritmo denominado de *leap-frog* (Hockney, 1974), que será descrito através das seguintes considerações para a expansão em séries de Taylor para a função horária da velocidade.

$$v_i(t + \frac{\Delta t}{2}) = v_i(t) + \frac{dv_i(t)}{dt} \frac{\Delta t}{2} + \frac{1}{2} \frac{d^2v_i(t)}{dt^2} \left(\frac{\Delta t}{2}\right)^2 + \dots$$

$$v_i(t - \frac{\Delta t}{2}) = v_i(t) - \frac{dv_i(t)}{dt} \frac{\Delta t}{2} + \frac{1}{2} \frac{d^2v_i(t)}{dt^2} \left(\frac{\Delta t}{2}\right)^2 + \dots$$

Realizando uma subtração entre as expansões descritas e rearranjando os termos temos que:

$$v_i(t + \frac{\Delta t}{2}) = v_i(t - \frac{\Delta t}{2}) + \frac{dv_i(t)}{dt} \Delta t + \dots$$

Neste contexto, desenvolvendo a mesma expansão para a função horária da posição, temos que:

$$r_i(t + \Delta t) = r_i(t) + v_i(t + \frac{\Delta t}{2}) \Delta t + \dots$$

Estas expansões representam o algoritmo citado, que é reconhecido como aquele de melhor precisão para aplicação no método de dinâmica molecular (Hinchliffe, 2003). Neste algoritmo, primeiro as velocidades são calculadas para o tempo " $t + \Delta t/2$ " e então utilizadas para o cálculo das coordenadas atômicas para o tempo " $t + \Delta t$ ". Desta forma, esta abordagem permite o cálculo das duas grandezas físicas de interesse - velocidades e coordenadas atômicas - ao longo das iterações (número de passos definidos para o desenvolvimento da simulação).

### 3.2. Minimização de energia

O objetivo dos métodos de minimização de energia do sistema é reduzir possíveis problemas na geometria das moléculas, tais como a presença não condizente com a realidade física de comprimento e ângulo de ligações químicas, bem como diedros relacionados. Do ponto de vista computacional, a maioria dos algoritmos de minimização de energia consistem na classe geral de métodos não lineares de otimização. Neste contexto, o objetivo é descobrir qual conjunto de valores para as variáveis independentes da função definida garantem um valor mínimo para tal função. Para o caso de um complexo binário (representação da interação proteína-ligante), com  $N$  átomos, temos  $3N$  variáveis, que correspondem às coordenadas atômicas, e a função que deseja-se minimizar corresponde à energia potencial deste sistema (McCammon, 1987).

Muitos dos campos de força modernos inerentes à dinâmica molecular (inclusive o utilizado no presente estudo), podem ser interpretados em termos de alguns parâmetros relativamente simples. A energia potencial para um sistema molecular dependente das coordenadas atômicas de seus constituintes pode ser expressa pelas seguintes contribuições:

$$V(r) = \frac{1}{2} \sum_{\text{ligação}} k_b (b_i - b_{i,o})^2 + \frac{1}{2} \sum_{\text{ângulo}} k_\theta (\theta_i - \theta_{i,o})^2 + \frac{1}{2} \sum_{\text{diedro}} V_n [1 + \cos(n\omega - \gamma)] + \sum_{i=1}^N \sum_{j=i+1}^N \left\{ 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}$$

O primeiro termo desta equação aparece devido ao estiramento das ligações químicas, o segundo devido às variações em relação aos ângulos destas, já o terceiro termo representa as mudanças nos diedros. O quarto termo representa a contribuição das interações à distância devido às forças de van der Waals, conhecido como potencial de Lennard-Jones (as constantes envolvidas foram determinadas empiricamente). O último termo também se refere à interação à distância, descreve a contribuição eletrostática. Programas modernos de dinâmica molecular empregam campos de força com pequenas variações nas constantes utilizadas para parametrização de cada um dos termos descritos anteriormente (Spoel et al., 2005). Portanto, essa característica repercute diretamente em como a energia potencial irá variar de acordo com as parametrizações feitas.

O método de minimização utilizado no presente estudo, assim como na maioria daqueles utilizados nos programas atuais de dinâmica molecular, deve contar com uma função bem comportada matematicamente. Neste sentido, essa determinada função deve ser contínua e diferenciável em todos os pontos, visto que se faz necessário um gradiente contínuo. O gradiente é um vetor que consiste na diferenciação parcial da função de interesse para cada uma de suas componentes e aponta para a direção da maior taxa de crescimento desta função. Se considerarmos esta função de interesse como  $f(x)$ , logo ao calcularmos  $-\text{grad } f(x)$  temos um vetor que aponta na direção do maior decréscimo para  $f(x)$ . Para um contexto de  $n$  dimensões, o cálculo para o gradiente da função  $f(x)$  fica da seguinte forma:

$$\text{grad } f = \left[ \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3}, \dots, \frac{\partial f}{\partial x_n} \right]$$

A função  $f(x)$  pode ser expandida em termos de uma série de Taylor em torno de um determinado ponto da seguinte maneira:

$$f(x) = f(x_0) + (x - x_0) \frac{df(x_0)}{dx} + \frac{(x - x_0)^2}{2} \frac{d^2 f(x_0)}{dx^2} + \dots$$

É possível generalizar esta relação para  $n$  dimensões substituindo  $x$  por um vetor  $r$  e introduzindo uma matriz para as derivadas correspondentes. Um método de otimização pode ser classificado de acordo com sua ordem, que se refere à derivada de mais alta ordem empregada pelo método dentre aquelas apresentadas na expansão de séries de Taylor. Portanto, métodos classificados como de primeira ordem, cortam a expansão logo após seu segundo termo. Há respaldo físico para tal aproximação, visto que a primeira derivada da energia potencial é a própria força (para sistemas conservativos, como aquele avaliado no presente estudo).

Os algoritmos atuam de forma iterativa, logo, iniciam em algum ponto da superfície molecular e prosseguem em ciclos de acordo com suas características. Cada um destes ciclos é definido como uma iteração, sendo representado através de um contador pela letra "k". A estrutura inicial do complexo binário é obtida a partir de um banco de dados de informações cristalográficas. O método *steepest*

*descent*, representante dos métodos de primeira ordem, foi utilizado durante as etapas de minimização de energia dos complexos proteína-ligante no presente estudo. Este método realiza sua busca por um ponto de mínimo através de três passos: segue-se a direção que aponta para o maior decréscimo para a função; o tamanho do passo, indicado por um escalar  $\lambda$ , é determinado; finalmente, o passo é realizado de acordo com a seguinte equação:

$$\mathbf{r}_k = \mathbf{r}_{k-1} + \lambda_k \mathbf{t}_k \quad \mathbf{t}_k = \frac{\mathbf{F}_k}{\max|\mathbf{F}_k|}$$

Na aplicação do algoritmo na minimização de energia, o tamanho do passo, representado por um escalar na equação descrita, é inicialmente determinado como condição inicial. Se a primeira iteração conduzir à uma redução na energia do sistema, esse escalar recebe um fator multiplicativo para a realização da segunda iteração. Este processo ocorre ciclicamente enquanto ocorrerem reduções na energia do sistema. Quando um passo promove um aumento na função da energia potencial, entende-se que o algoritmo passou por uma região de mínimo local (Radaport, 2004). Conseqüentemente, o tamanho do passo que será aplicado na próxima iteração é submetido à uma redução por um fator multiplicativo.

O algoritmo cessa as iterações quando o contador associado indica um número determinado de ciclos ou quando o valor para o gradiente encontrado é menor que uma referência (também determinada nas condições iniciais do método). Este se apresenta como uma poderosa ferramenta, para rapidamente atingir uma região de mínimo local, apesar de apresentar dificuldade em chegar ao ponto de mínimo valor para a função de interesse, quando nesta região.

Um algoritmo interessante para uso complementar ao *steepest descent*, no processo de minimização de energia, é o algoritmo conhecido como *conjugate gradient*. Este método produz um conjunto de direções que não apresenta o comportamento oscilatório do *steepest descent*, visto que para as iterações seguintes são considerados os valores definidos para o gradiente do ciclo anterior. Para a primeira iteração, a direção de busca se dá através do cálculo do gradiente inicialmente (grad1), considerando o sinal oposto devido ao interesse de seguir para o sentido de maior decréscimo energético para o sistema (como discutido

anteriormente). A equação que orienta a direção de busca para este algoritmo consiste na seguinte forma:

$$s_k = -g_k + b_k s_{k-1} \quad b_k = \frac{g_k \cdot g_k}{g_{k-1} \cdot g_{k-1}}$$

Este algoritmo apresenta uma característica de levar mais tempo nos primeiros passos do processo de minimização de energia, devido ao seu método de cálculo ser recorrente, ou seja, depender explicitamente de valores das iterações anteriores. Em contrapartida, se apresenta como uma poderosa ferramenta nas proximidades do ponto de mínimo, sendo capaz de atingir valores mais próximos do mínimo local que o *steepest descent*.

Portanto uma combinação dos dois algoritmos apresentados é uma excelente aplicação para otimizar o processo de minimização de energia, tanto em exigência computacional, quanto na viabilidade de alcance de pontos mais próximos possíveis ao de menor energia para o sistema. Desta forma, é provável que os artifícios técnicos presentes na informação estrutural sejam evitados ao máximo, aproximando o modelo em estudo da realidade física.

# Desenvolvimento de um modelo computacional aplicado à família das serino endopeptidases

K. R. M. Moraes<sup>1</sup> ; W. F. de Azevedo Jr. <sup>1,2</sup>

<sup>1</sup> Instituto Nacional de Ciência e Tecnologia em Tuberculose-CNPq-MCT, Faculdade de Biociências, Laboratório de Bioquímica Estrutural, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre – RS, Brasil

<sup>2</sup> Programa de Pós-Graduação em Medicina e Ciências da Saúde, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre – RS, Brasil

## 4.1. Resumo

O processo de pesquisa e desenvolvimento de fármacos é complexo, longo e de alto custo. A descoberta de novos medicamentos tem suas raízes profundamente ligadas às inovações científicas e tecnológicas. Neste contexto, o desenvolvimento da computação, visto nos últimos anos, tem tido um papel de grande relevância. Os testes *in silico* constituem uma importante etapa a partir da identificação do problema alvo, pois permitem uma abordagem racional no processo de desenvolvimento de fármacos e, portanto, uma redução no tempo e custo envolvidos. Nesse sentido, o presente trabalho pretende propor um modelo computacional capaz de estimar a interação existente entre proteína - ligante, que leve em consideração aspectos dinâmicos deste sistema. Para tal, foi alvo de estudo uma família de proteínas em especial: as serino endopeptidases. Com a finalidade de avaliar os aspectos de caráter dinâmico responsáveis predominantemente pela estabilidade da interação do complexo biológico, foram realizadas simulações de dinâmica molecular, partindo de informações estruturais cristalográficas disponíveis em bancos de dados. A partir destas simulações será proposto um modelo matemático capaz de estimar a afinidade do ligante pela proteína, usando uma função *score* empírica, que terá inicialmente a forma polinomial, onde cada termo do polinômio será formado por termos que caracterizam a interação intermolecular entre o ligante e a proteína. Os coeficientes associados aos termos para a função de interesse serão determinados através de métodos de regressão linear multivariável.

Palavras chave: proteína, ligante, dinâmica molecular, função *score*.

## 4.2. Introdução

Novos fármacos são desenvolvidos no intuito de modular proteínas de interesse, ativando ou inibindo as suas funções. Os avanços na compreensão da interação entre proteínas e ligantes são de grande importância para contribuição com a eficiência dos processos de desenho e descoberta de fármacos. Com o objetivo de identificar candidatos em potencial para a interação com as proteínas em estudo, a aplicação de docagem molecular é um método interessante. O modelo em questão permite a estimativa da interação do complexo binário através das informações estruturais das macromoléculas, aliadas à aplicação de uma função de energia capaz de identificar conformações favoráveis para o sistema heterogêneo. Entretanto, nesta abordagem, o solvente, quando considerado, é tratado de forma implícita, além da limitação intrínseca ao método quanto à falta de flexibilidade no sítio alvo da proteína. Substituir a característica rígida da proteína (clássica da visão chave - fechadura para a interação desta com os ligantes) por um alvo flexível é

coerente com o conhecimento atual acerca da importância das mudanças conformacionais para a função protéica. Neste contexto, o emprego de simulações de dinâmica molecular tem sido uma excelente ferramenta computacional para compreensão das bases físicas envolvidas com a estrutura e, por conseguinte, com a função de macromoléculas biológicas. Estas simulações contribuem complementando os dados experimentais acerca das estruturas das proteínas de interesse, através da provisão de informações sobre a dinâmica destas a nível atômico. É relevante citar a consideração das moléculas do solvente de forma explícita nos sistemas, para o contexto biológico, tratamos do solvente universal: água. Portanto, é viabilizada uma abordagem do problema da estimativa da interação entre ligante e proteína, mesmo que ainda *in silico*, considerando aspectos dinâmicos para os sistemas em questão. O conjunto de dados gerados, oriundo das dinâmicas dos sistemas em estudo, é capaz de orientar a proposição de funções matemáticas com o mesmo propósito

das utilizadas no método de docagem molecular. Tal proposição, contudo, representa uma tentativa de tratamento com maior aproximação da realidade física.

### 4.3. Metodologia

#### Base de dados estruturais e de afinidade

A base de dados PDBind (Wang et al., 2004) foi utilizada como fonte de informação estrutural. Como forma de selecionar dentre as estruturas disponíveis, utilizamos os seguintes critérios de seleção:

- 1) Estruturas cristalográficas com resolução melhor que 2,0 Å ( $10^{-10}$  m);
- 2) Estruturas cristalográficas publicadas após 1º de janeiro de 2000;
- 3) Dados de afinidade expressos na forma de constante de inibição ( $K_i$ );
- 4) Estruturas de complexos 1 proteína + 1 ligante;
- 5) Estruturas sem metais no sítio de ligação;
- 6) Estruturas cristalográficas sem resíduos faltando na região do sítio de ligação ( $< 10,0$  Å);

#### Parametrização dos ligantes

As coordenadas atômicas dos ligantes foram obtidas a partir da estrutura cristalográfica do complexo binário. Os arquivos referentes às topologias dos ligantes foram gerados a partir do PRODRG (Schüttelkopf & van Aalten 2004). As cargas foram calculadas segundo o método ChelpG (Breneman et al., 1990) através do programa Gaussian, que também viabilizou a otimização da estrutura dos mesmos (Frisch et al., 2009). O cálculo se deu através do seguinte protocolo: DFT/B3LYP/6-311G/2dp.

#### Minimização de energia

As estruturas de interesse, através da utilização do programa Gromacs 4.6.2 (Van Der Spoel et al., 2005), foram solvatadas utilizando o modelo SPC/E para as moléculas de água (Berendsen et al., 1987), em um contexto de uma caixa cúbica, garantindo uma distância mínima entre a proteína e a superfície da caixa no valor de 10 angstroms. Sempre que necessário, foi realizada uma neutralização das cargas do

sistema, através da adição de cátions (sódio) ou ânions (cloreto), utilizando a função Genion do próprio Gromacs. Os sistemas foram submetidos ao método de *steepest descent*, seguido do *conjugate gradient* para fins de minimização de energia dos mesmos. Foi definida uma tolerância energética de 1000 kJ/mol para o primeiro algoritmo e de 100 kJ/mol para o segundo.

### ***Position Restraints***

Uma etapa de curta duração (30 ps -  $10^{-12}$  s) se segue ao processo de minimização de energia. Esta etapa já se dá sob as condições da dinâmica molecular propriamente dita, apesar de que são aplicadas forças nos átomos do sistema para restringir o movimento deles. Desta forma, há uma liberdade apenas ao movimento das moléculas do solvente, para as quais não são parametrizadas forças restritivas.

### **Termalização**

Segue à dinâmica molecular de pequena duração (com as condições de restrição), consiste na etapa em que os sistemas são gradualmente aquecidos. O aquecimento se dá

através dos seguintes incrementos na temperatura: 50K (10ps), 100K (10ps), 150K (10ps), 200K (10ps), 250K (10ps) e 298K (20ps). Para cada etapa, as velocidades do átomos foram novamente geradas de acordo com uma distribuição de Maxwell-Boltzman.

### **Dinâmica Molecular**

Performance da etapa de produção propriamente dita. Ao final desta, são coletados dados sobre os sistemas de interesse. Segue com o mesmo protocolo a partir da última rampa de aquecimento da fase de termalização, entretanto não são mais geradas novas distribuições para as velocidades dos átomos dos sistemas. Etapa serve como critério de exclusão da seguinte forma: estruturas para as quais a simulação de dinâmica molecular não convirja para a estabilidade da estrutura nos primeiros 25 ns da simulação, além daquelas que apresentem interrupção inesperada durante o processo serão desconsideradas para os conjuntos finais.

## Função Escore

A grande quantidade de dados gerada devido às simulações computacionais servirá de base para a elaboração do modelo computacional capaz de prever a afinidade do ligante pela proteína. Há abordagens computacionais que utilizam análise de dinâmica molecular para avaliação da interação proteína - ligante, por exemplo a opção o cálculo da energia de interação (*Linear Interaction Energy*) (Åqvist et al., 1994; Hansson

et al., 1998) do pacote de programas GROMACS (van der Spoel et al., 2005). No presente projeto usaremos um polinômio similar ao utilizado para a implementação do programa POLSCORE (De Azevedo & Dias, 2008). O POLSCORE é usado para previsão da afinidade do ligante pela proteína, apesar de que este leva em consideração unidades de pKa. A novidade da presente abordagem é que aspectos dinâmicos da interação serão inseridos numa função polinomial para a previsão da afinidade.

## 4.4. Discussão

O presente trabalho teve como ponto de partida a análise do banco de dados PDBBind (Wang et al., 2004), no intuito de avaliar quais critérios de seleção seriam considerados para os complexos binários. Esta medida foi tomada com o objetivo de assegurar a qualidade das informações a serem utilizadas ao longo do estudo, bem como minimizar a interferência de possíveis variáveis de confusão. Logo, foi percebida a coerência em abordar

o problema da estimativa da interação proteína - ligante, considerando uma família de proteínas em especial, uma vez que é de se esperar que funções distintas venham a explicar o contexto das interações para diferentes tipos de famílias protéicas. Portanto, para que o estudo tivesse seguimento, se fez necessária a escolha da família que seria o foco deste. A opção pelas serino endopeptidases foi feita orientada exclusivamente por

motivos técnicos, visto que, dentre as informações dos bancos de dados consultados, correspondia àquela de melhor disponibilidade dentro dos critérios de seleção adotados.

Foram avaliadas as estruturas disponíveis para a família escolhida (EC 3.4.21.). A relação destas foi consultada em Janeiro de 2013, no seguinte site: [www.ebi.ac.uk/pdbe](http://www.ebi.ac.uk/pdbe) (European Bioinformatics Institute). No momento da consulta, haviam 1.717 resultados para a família de interesse.

De posse deste conjunto de informações estruturais, foram aplicados os critérios de seleção definidos para o presente trabalho. Portanto, houve uma redução para 197 estruturas em conformidade com os mesmos.

A partir dos arquivos escolhidos, foram extraídas informações estruturais dos ligantes e proteínas a partir dos complexos depositados no PDB ([rcsb.org](http://rcsb.org)). Foram desconsideradas as moléculas de água cristalográficas. Assim como demais moléculas remanescentes, artifício técnico da determinação estrutural cristalográfica das macromoléculas, por exemplo: glicerol, íons sulfato, ácido acético.

Logo, para cada arquivo do conjunto inicial de dados, foram salvos dois, um referente às coordenadas atômicas da proteína e o outro ao ligante, ambos no formato PDB. Este último foi submetido ao PRODRG (Schüttelkopf & van Aalten, 2004), com o propósito de obter um arquivo para sua topologia. Tal procedimento é de fundamental importância devido à ausência de parametrização para os ligantes no campo de força escolhido para as simulações realizadas no Gromacs. A distribuição de cargas nos ligantes é um ponto nevrálgico, logo, se faz necessária a aplicação de um método mais rigoroso no tratamento desta questão. Para tal, foi empregado um método de estrutura eletrônica, consistido na teoria de densidades eletrônicas (DFT), através do programa Gaussian (Frisch et al., 2009). O aumento significativo do custo computacional, devido à implementação desta etapa adicional para o tratamento dos ligantes, é justificado em virtude da qualidade advinda deste procedimento. Foram aproveitadas as informações resultantes das otimizações pelo Gaussian no que diz respeito aos comprimentos e ângulos de ligações, ângulos de torções, além da

distribuição das cargas propriamente dita.

Os arquivos de entrada para cada simulação de dinâmica molecular foram: prot.pdb (arquivo PDB com as coordenadas atômicas da proteína), lig.pdb (coordenadas atômicas do ligante), lig.itp (topologia do ligante) e os arquivos de comandos do Gromacs (Spoel et al., 2013). Para cada caso, o arquivo contendo as coordenadas atômicas da proteína foi introduzido no programa Gromacs, onde as simulações de dinâmica molecular foram desenvolvidas. Foi promovida a solvatação das proteínas de interesse, criando uma caixa cúbica. Logo, as etapas de simulação tiveram seguimento tanto com, quanto sem a presença dos ligantes relacionados. Quando necessária, foi realizada uma etapa de neutralização das cargas do sistema, através da adição de cátions (sódio) ou ânions (cloreto), conforme o caso. A função utilizada para esse fim foi a Genion, do próprio Gromacs, a qual foi ajustada para que calculasse a densidade de cargas antes de inserir cada íon, no intuito de garantir a maior uniformidade eletrostática possível. Para tal, foi promovida a substituição de moléculas de solvente

para a introdução de cada íon, quando necessário.

O campo de força utilizado foi o Gromos 53a6, devido à sua parametrização para o tratamento de proteínas, além do seu recente estabelecimento em comparação aos outros incluídos no pacote Gromacs.

A minimização de energia foi uma etapa inicial na utilização do programa. O seu papel é justificado pela origem das informações estruturais, visto que existem artefatos inerentes à técnica de determinação cristalográfica por difração de raios de alta energia (raios x). Para a aplicação desta técnica, as proteínas devem ser cristalizadas, logo, ocorre um efeito de empacotamento que não pode ser desconsiderado no momento da avaliação da sua estrutura nativa ou ligada. A etapa de minimização foi aplicada no intuito de eliminar possíveis geometrias moleculares incoerentes com a realidade física das proteínas em meio fisiológico, exemplificadas através de comprimento ou ângulos de ligação incongruentes. Foram combinados dois algoritmos distintos para a realização desta etapa. Inicialmente foi aplicado o método de *steepest*

*descent*, para o qual foi definida uma tolerância energética de 1000 kJ/mol. Após a convergência do primeiro algoritmo, o seu resultado foi utilizado como ponto de partida para a aplicação de um segundo método: *conjugate gradient*, para o qual foi ajustada uma tolerância energética mais rigorosa, no valor de 100 kJ/mol.

Concluída a minimização de energia, segue a etapa de dinâmica com restrição ao movimento dos átomos da proteína e do ligante. Esta característica é assegurada pela definição de forças restritivas, aplicadas nos átomos do eixo central da proteína, assim como nos átomos do ligante. Esta etapa é importante, pois previne mudanças bruscas de partes da proteína devido à submissão às forças causadas por um desequilíbrio das moléculas do solvente. Logo, é permitida uma distribuição destas moléculas ao longo do sistema, sem que haja alteração relevante do complexo binário. Desta forma, é garantida uma aproximação do equilíbrio para o solvente antes que seja permitida a flexibilidade para o complexo binário, característica da dinâmica molecular propriamente dita.

Foram implementadas etapas consecutivas com gradual incremento na temperatura dos sistemas, constituindo numa fase sequencial de aquecimento. Nessa altura, são retiradas as forças restritivas sobre os componentes que atuavam durante a fase da dinâmica com restrições. A adoção desta estratégia metodológica visa evitar um choque térmico do sistema, prevenindo que a proteína assuma possíveis conformações incoerentes com àquelas do contexto fisiológico. Enquanto é realizado cada aumento de temperatura, são redistribuídas as velocidades associadas aos átomos do sistema. A temperatura final definida é aquela utilizada durante o ensaio enzimático para cada caso, correspondendo usualmente ao valor de 298,15 K.

A etapa subsequente corresponde à fase de produção propriamente dita, pois os sistemas, que já se encontram na temperatura relacionada aos ensaios enzimáticos específicos, são livres para o desenvolvimento da dinâmica molecular. É esperada uma convergência dos sistemas ao longo das simulações, no sentido de que não há mais restrições definidas para os

componentes dos mesmos, além de que não alterações na temperatura e pressão dos mesmos. O tempo para esta convergência é característico de cada sistema e é denominado como o tempo de relaxamento deste, referindo-se à estabilidade do comportamento após uma série de perturbações (incrementos de temperatura e, conseqüentemente, de energia total). A definição de quando o sistema atinge este tempo característico não é um consenso atualmente, quando realizada a partir da análise dos dados gerados durante as simulações, apesar de que usualmente o RMSD (desvio quadrático médio da posição dos átomos constituintes) é uma alternativa utilizada. Foi determinado um tempo total para a fase de dinâmica propriamente dita de 25 ns, como forma de balancear o custo computacional implicado e cobertura dos tempos de relaxamento para os sistemas estudados. De posse do conjunto de dados gerados após o desenvolvimento das simulações, chegaremos ao ponto de propor uma função matemática capaz de estimar a interação dos complexos através de métodos de regressão linear.

O uso de programas computacionais, para avaliar a interação proteína-ligante, baseados em funções *scores* empíricas, começou com o trabalho pioneiro de Böhm (Böhm, 1994). Nesse trabalho, foi proposto que a interação entre proteína e ligante fosse avaliada a partir de uma função *score* empírica. Tal função *score* é um polinômio que envolve aspectos estruturais da interação proteína-ligante, tais como, ligações de hidrogênio, contatos de van der Waals, contatos hidrofóbicos, interações eletrostáticas. Cada uma destas características estruturais entra na função *score* com um termo do polinômio, sempre de grau 1. Os pesos relativos de cada termo são determinados a partir de regressão linear, usando dados experimentais, disponíveis na literatura científica, sobre a afinidade proteína-ligante, como conjunto de treino (*training set*). A abordagem deste estudo não pretende, a princípio, restringir o grau para os termos assim como foi feito nos primeiros trabalhos sobre o tema. Esta função é uma expressão matemática que representa a afinidade de um ligante por uma proteína. Diversos programas computacionais trazem tais funções

escores que usam esta abordagem (Wang *et al.*, 2002; De Azevedo & Dias, 2008), constituindo a construção destas funções escores como uma alternativa muito utilizada atualmente no que diz respeito a avaliação de ligantes em potencial para os mais variados sistemas biológicos.

As funções escores empíricas terão inicialmente a forma de polinômios, onde cada termo do polinômio é formado por termos que caracterizem a interação intermolecular entre o ligante e a proteína. Os termos desses polinômios podem ser lineares, não lineares ou mistos, por exemplo a combinação de termos como contatos de van der Waals ou número de ligações de hidrogênio. (De Azevedo & Dias, 2008). É introduzida uma abordagem mais abrangente do problema em estudo, visto que a afinidade de um dado ligante por uma proteína será estimada por uma função escore específica, a qual apresentará termos ponderados pela contribuição das forças intermoleculares avaliadas ao longo das simulações de dinâmica molecular. É esperado que as funções que melhor ajustem os parâmetros

para correlação com os dados experimentais sejam aquelas que apresentem na sua constituição termos que correspondam às componentes físicas responsáveis pela estabilidade do ligante no sítio protéico.

## 5. Conclusão e perspectivas

O presente trabalho teve o objetivo de propor uma função matemática, capaz de avaliar a interação entre proteína e ligante que levasse em consideração aspectos dinâmicos deste complexo binário. Devido à diversidade estrutural relevante dentre as diferentes famílias de proteínas, o foco deste estudo foi a família das serino endopeptidases. Este estudo se encontra na fase de desenvolvimento das simulações de dinâmica molecular, portanto ainda está sendo gerado o conjunto de dados que será utilizado para o ajuste das funções matemáticas de interesse (cerne do modelo).

Aspectos médios relevantes, para interação proteína-ligante, serão calculados usando-se as trajetórias dos complexos obtidas durante as simulações de dinâmica molecular. Serão analisados aspectos como: número de ligações de hidrogênio intermoleculares, área de contato, área acessível ao solvente, RMSD (desvio médio quadrático) da estrutura do ligante, RMSD da estrutura da proteína entre outros. Tais aspectos serão combinados no intuito de constituir os coeficientes que otimizem as funções de correlação com dados experimentais.

Espera-se elaborar um modelo computacional para previsão da afinidade de ligantes proteínas, um problema central do desenho de novos fármacos. Tal modelo computacional será implementado num programa de computador a ser implementado na linguagem C++, usando uma abordagem computacional similar ao programa PROTLIGSCORE (094034 - Catálogo de Tecnologias disponíveis para licenciamento da PUCRS, organizado pelo Escritório de Transferência de Tecnologia da Universidade) desenvolvido no LaBioqEst-PUCRS.

## 6. Referências

- W. F. van Gunsteren, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.  
"Biomolecular Modeling: Goals, Problems, Perspectives", *Angew. Chem. Int. Ed.*  
2006, 45, 4064 – 4092.
- Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. "The PDBbind Database:  
Methodologies and updates", *J. Med. Chem.*, 2005; 48(12); 4111-4119.
- Wang, R.; Fang, X.; Lu, Y.; Wang, S. "The PDBbind Database: Collection of Binding  
Affinities for Protein-Ligand Complexes with Known Three-Dimensional  
Structures", *J. Med. Chem.*, 2004; 47(12); 2977-2980.
- M. J. Frisch, G. W. Trucks, H. B. Schlegel. "Gaussian 09 User's Reference", Gaussian,  
Inc., Wallingford CT, 2009.
- Van Mourik, Tanja; Gdanitz, Robert J. "A critical note on density functional theory  
studies on rare-gas dimers". *Journal of Chemical Physics* 116, 2002;(22): 9620–  
9623.
- Bartels, Christian "Analyzing biased Monte Carlo and molecular dynamics  
simulations". *Chemical Physics Letters* 331, 2000;(5–6): 446–454.
- Hohenberg, P; Kohn, W. "Inhomogeneous Electron Gas" *Phys. Rev.* 1964, 136, B864.
- B. Hess, D. van der Spoel, E. Lindahl, and the GROMACS development team  
"GROMACS User Manual version 4.6.4", 2013.
- Sant'Anna, Carlos M. R. "Molecular modeling methods in the study and design of  
bioactive compounds: An introduction" *Rev. Virtual Quim.*, 2009, 1 (1), 49-57.
- Stephen T. Thornton, Jerry B. Marion "Classical Dynamics of Particles and Systems  
- 5th edition", 2011; 291-330, Thomson Learning, Belmont.
- McCammon, J.A.; Harvey, S.C. "*Dynamics of protein and nucleic acids*", Cambridge  
University Press: Cambridge, 1987

Field, M.J. "A practical introduction to the simulation of molecular systems", 2nd Ed. Cambridge University Press: Cambridge, 2007.

Hockney, R.W., Goel, S.P., Eastwood, J. "Quiet high-resolution computer models of a plasma". *J. Comp. Phys.*, 1974, 14, 148–58.

Hinchliffe, A. "Molecular modelling for beginners". John Wiley & Sons Ltd: West Sussex, 2003.

van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A.E. Berendsen, H.J.C. "GROMACS: fast, flexible, and free". *J. Comp. Chem.*, 2005, 26, 1701–18.

Radaport, D.C. "The art of molecular dynamics simulations", 2nd Ed. Cambridge University Press: New York, 2004.

De Azevedo WF Jr. "Protein-drug interactions". *Curr Drug Targets* 2008; 9(12): 1030.

Barberato Filho, S. "Pesquisa e desenvolvimento de fármacos no Brasil: estratégias de fomento". 2006. 192f. Tese (Doutorado em Ciências Farmacêuticas) – Faculdade de Ciências Farmacêuticas, Universidade de São Paulo, São Paulo.

C.M. Breneman, K.B. Wiberg "Determining atom-centered monopoles from molecular electrostatic potentials, the need for high sampling density in formamide conformational analysis" *J. Comput. Chem.*, 11 (1990), pp. 361–373

H.J.C. Berendsen, J.R. Grigera, T.P. Straatsma. "The missing term in effective pair potentials". *J. Phys. Chem.*, 91 (1987), pp. 6269–6271

Åqvist, J.; Medina, C.; Samuelsson, J. E. (1994) *Protein Eng.* 7, 385–391.

Böhm, H. J. (1994) *J. Comput. Aid. Mol. Des.* 8, 243-256.

De Azevedo WF Jr, Dias R. "Evaluation of ligand-binding affinity using polynomial empirical scoring functions". *Bioorg Med Chem* 2008; 16(20): 9378 - 83.