UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE QUÍMICA

Programa de Pós-Graduação em Ciências Biológicas (Bioquímica)

Ariane Ferreira Nunes Alves

**Um método computacional para estimar
afinidades entre proteínas flexíveis e
pequenos ligantes**

Versão original da dissertação defendida

São Paulo

28/03/2013

Ariane Ferreira Nunes Alves

**Um método computacional para estimar
afinidades entre proteínas flexíveis e
pequenos ligantes**

*Dissertação apresentada ao Instituto de
Química da Universidade de São Paulo para
obtenção do Título de Mestre em
Ciências (Bioquímica)*

Orientador: Prof. Dr. Guilherme Menegon Arantes

São Paulo

2013

Ficha Catalográfica

Elaborada pela Divisão de Biblioteca e
Documentação do Conjunto das Químicas da USP.

Alves, Ariane Ferreira Nunes

A474m Um método computacional para estimar afinidades entre proteínas flexíveis e pequenos ligantes / Ariane Ferreira Nunes Alves. -- São Paulo, 2013.
72p.

Dissertação (mestrado) – Instituto de Química da Universidade de São Paulo. Departamento de Bioquímica.
Orientador : Arantes, Guilherme Menegon

1. Proteína : Interação molecular : Bioquímica I. T. II Arantes, Guilherme Menegon, orientador.

574.19245

CDD

Aos meus pais, Maria Elisa e Heli.
Obrigada pelo amor, carinho, educação e apoio.

A Javier.

Obrigada pelo amor, carinho e apoio.

Agradecimentos

Agradeço ao meu orientador, prof. Dr. Guilherme Menegon Arantes, que sempre acompanhou de perto o meu trabalho e contribuiu com inúmeras sugestões, críticas construtivas e recomendações de leitura. A orientação dele foi determinante para meu crescimento científico e intelectual durante a realização do mestrado.

Agradeço aos meus colegas e ex-colegas de laboratório, Rafael, João e Gustavo, e aos meus amigos e colegas de estudo, Bruno e Valquíria, que também contribuíram para a minha formação.

Agradeço ao meu namorado, Javier, por suas sugestões para melhorar a dissertação.

Agradeço aos criadores do abnTeX, um pacote de classes L^AT_EX para a criação e formatação de documentos conforme as normas ABNT.

Agradeço também ao Instituto de Química da Universidade de São Paulo por prover um bom ambiente para a realização do meu mestrado.

Por fim, agradeço às agências que financiaram a minha pesquisa e a minha bolsa de mestrado, o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp).

Um pouco de ciência nos afasta de Deus. Muito, nos aproxima.

Louis Pasteur

Resumo

Alves, A.F.N. **Um método computacional para estimar afinidades entre proteínas flexíveis e pequenos ligantes.** 2013. 72p. Dissertação - Programa de Pós-Graduação em Bioquímica. Instituto de Química, Universidade de São Paulo, São Paulo.

Métodos computacionais são usados para gerar estruturas de complexo proteína-ligante e estimar suas afinidades. Esse trabalho investigou como as diferentes representações da flexibilidade proteica afetam as poses obtidas por ancoragem molecular e as afinidades atribuídas a essas poses. Os mutantes L99A e L99A/M102Q da lisozima T4 foram escolhidos como sistemas modelo. Um descritor para predição de afinidades baseado na aproximação de energia de interação linear (LIE) foi parametrizado especificamente para ligantes da lisozima e foi usado para estimar as afinidades. A proteína foi representada como um grupo de estruturas cristalográficas ou de estruturas de trajetória de dinâmica molecular. O campo de força OPLS-AA para modelar a proteína e os ligantes e a aproximação de Born generalizada para modelar o solvente foram empregados. O descritor de afinidades parametrizado resultou em desvios médios entre afinidades experimentais e calculadas de 1,8 kcal/mol para um conjunto de testes. O descritor teve desempenho satisfatório na separação entre poses cristalográficas e poses falso-positivo e na identificação de poses falso-positivo. Experimentos de agrupamento de complexos realizados com o objetivo de reduzir o custo computacional para estimar afinidades apresentaram resultados insatisfatórios. As melhores aproximações da teoria do ligante implícito propostas aqui para estimar afinidades consideram conjuntos de estruturas de receptor com o mesmo peso. Configurações de ligante também apresentam o mesmo peso ou são dominadas por uma única configuração. A representação da flexibilidade requer um tratamento estatístico adequado para estimativa de afinidades. Aqui, a associação entre LIE e a teoria do ligante implícito mostrou-se frutífera.

Palavras-chave: afinidade ligante-proteína, ancoragem molecular, energia de interação linear (LIE), flexibilidade conformacional, lisozima T4.

Abstract

Alves, A.F.N. **A computational method to estimate affinities between flexible proteins and small ligands.** 2013. 72p. Master Thesis - Graduate Program in Biochemistry. Instituto de Química, Universidade de São Paulo, São Paulo.

Computational methods are used to generate protein–ligand complex structures and estimate their binding affinities. This work investigated how different representations of protein flexibility affect poses obtained by molecular docking and the affinities attributed to these poses. T4 lysozyme mutants L99A and L99A/M102Q were chosen as model systems. A descriptor for prediction of affinities based on linear interaction energy (LIE) approximation was parametrized specifically to lysozyme ligands and was used to estimate affinities. The protein was represented as a group of crystal structures or as structures from a molecular dynamics trajectory. OPLS–AA force field was used to model protein and ligands and the Generalized Born approximation was used to model solvent. The parametrized affinity descriptor resulted in average deviations between experimental and calculated affinities of 1.8 kcal/mol for a test set. Descriptor performance was satisfactory in the separation between crystal poses and false–positive ones and in the identification of false–positive poses. Clustering of complexes was tried out to reduce computational cost to estimate affinities, but results were poor. The best approximations to the ligand implicit theory proposed here in order to estimate affinities consider groups of receptor structures with the same weight. Ligand configurations also have the same weight or are dominated by only one configuration. The representation of a protein flexibility requires an adequate statistical treatment when used to estimate affinities. Here, the linking between LIE and the implicit ligand theory proved itself useful.

Keywords: conformational flexibility, docking, ligand–protein affinity, linear interaction energy (LIE), T4 lysozyme.

Lista de Figuras

1.1	Estrutura da lisozima do bacteriófago T4.	p. 11
1.2	Mutantes da lisozima usados no estudo da formação de complexos.	p. 11
4.1	Esqueletos proteicos alinhados do conjunto de estruturas cristalográficas de L99A (a) e de M102Q (b) e do conjunto de estruturas da trajetória de L99A (c) e de M102Q (d).	p. 26
4.2	Sítio de ligação putativo entre domínios.	p. 27
4.3	Aminoácidos associados a obstrução do sítio de ligação.	p. 29
4.4	Trajetoórias de deslocamento químico predito para os átomos de TYR88 indicados na legenda.	p. 30
4.5	Estruturas de alguns ligantes de M102Q escolhidos para o estudo.	p. 32
4.6	Comparação entre energias livres de ligação experimentais e calculadas por Vina (a) ou pelo descritor de afinidades (b).	p. 39
4.7	Poses do não-ligante NCF obtidas por ancoragem à estrutura do PDB 3HT6.	p. 40
4.8	Superposição das poses do ligante JZ4 (M102Q) obtidas por ancoragem a um conjunto de estruturas cristalográficas.	p. 42
4.9	Histogramas de energias livres de ligação para complexos com BNZ (L99A) e JZ4 (M102Q).	p. 43
4.10	Desvios da energia média [(a), $RMSD_{ene}$] e da dispersão [(b), $RMSD_{dis}$] para todos os ligantes testados.	p. 46
4.11	Dispersão máxima de $E_{desc}(cc)$ (DM) para todos os ligantes testados.	p. 47
4.12	Desvios da energia média [(a), $RMSD_{ene}$] e da dispersão [(b), $RMSD_{dis}$] para todos os ligantes testados.	p. 48
4.13	Dispersão máxima de $E_{desc}(cc)$ (DM) para todos os ligantes testados.	p. 48
4.14	Diferença entre (a) $RMSD_{ene}$ e (b) $RMSD_{dis}$ dos grupos obtidos do agrupamento por RMSD ($RMSD^{RMSD}$) e do agrupamento por contatos ($RMSD^{contatos}$) para todos os ligantes testados.	p. 49
4.15	Diferença entre a dispersão máxima de $E_{desc}(cc)$ obtida pelo agrupamento por RMSD (DM^{RMSD}) e pelo agrupamento por contatos ($DM^{contatos}$).	p. 50
4.16	Diferença entre a população do grupo com dispersão máxima de $E_{desc}(cc)$ obtida pelo agrupamento por RMSD (P^{RMSD}) e pelo agrupamento por contatos ($P^{contatos}$).	p. 50

4.17	Superposição das poses do ligante JZ4 (M102Q) obtidas por ancoragem a um conjunto de estruturas da trajetória.	p. 52
4.18	Sequência de aproximações utilizadas para obter energias livres de ligação usando a teoria do ligante implícito.	p. 52
4.19	Estrutura de complexos com lisozima.	p. 58

Lista de Tabelas

4.1	RMSD (em ppm) entre deslocamentos químicos experimentais e preditos para a trajetória de M102Q ou para o conjunto de estruturas cristalográficas de M102Q.	p. 25
4.2	Média $\pm$ desvio padrão dos deslocamentos químicos calculados (em ppm) para M102Q.	p. 27
4.3	Média $\pm$ desvio padrão dos deslocamentos químicos calculados (em ppm) para M102Q.	p. 28
4.4	Ligantes e não–ligantes (em negrito) de L99A e M102Q escolhidos para o estudo, com a estrutura cristalográfica (código de PDB) e sigla correspondentes indicadas.	p. 32
4.5	Valores de momento de dipolo e seus componentes (em Debye) calculados usando as cargas fornecidas pelo OPLS-AA, o método HF ou o método AM1.	p. 34
4.6	Energias livres de ligação (em kcal/mol) para complexos do conjunto de treino.	p. 37
4.7	Energias livres de ligação (em kcal/mol) para complexos do conjunto de teste.	p. 38
4.8	Ligantes e não–ligantes (em negrito) de L99A e M102Q ordenados de forma decrescente a partir de energias livres de ligação.	p. 41
4.9	Energias livres de ligação (em kcal/mol) de complexos oriundos de ancoragem com estrutura <i>apo</i> ou com conjunto de estruturas cristalográficas.	p. 45
4.10	Energias livres de ligação (em kcal/mol) estimadas pela teoria do ligante implícito.	p. 55
4.11	Energias livres de ligação (em kcal/mol) estimadas por diferentes aproximações para a teoria do ligante implícito.	p. 59
4.12	Ligantes e não–ligantes (em negrito) de L99A ordenados de forma decrescente a partir de energias livres de ligação estimadas.	p. 61
4.13	Ligantes e não–ligantes (em negrito) de M102Q ordenados de forma decrescente a partir de energias livres de ligação estimadas.	p. 61
4.14	Desvios médios (em kcal/mol) entre energias livres de ligação experimentais e estimadas.	p. 62

Lista de abreviaturas, siglas e símbolos

1AN	2-fluoroanilina
2AP	2-aminofenol
2EP	2-etoxifenol
3MP	3-metilpirrol
4CP	4-cloro-1h-pirazol
4VP	4-vinilpiridina
A	área de superfície do soluto acessível ao solvente
AM1	Austin Model 1
ANL	anilina
B	energia de interação efetiva média para uma configuração de receptor
B^D	valor de B calculado usando a equação 4.2
B^M	valor de B calculado usando a equação 4.1
BNZ	benzeno
C	carbono da ligação peptídica
$C\alpha$	carbono α
CAQ	catecol
CHX	cicloexano
CMI	5-cloro-2-metilfenol
δ	deslocamento químico do sinal de ressonância magnética nuclear
$\Delta E_{desc}(e)$ -fs	diferença entre $E_{desc}(e)$ máximo e mínimo para complexos em que o ligante estava fora sítio de ligação
$\Delta E_{desc}(e)$ -s	diferença entre $E_{desc}(e)$ máximo e mínimo para complexos em que o ligante estava no sítio de ligação
ΔG^D	valor de ΔG_{lig} calculado usando a equação 4.4
ΔG_{desc}^M -fs	ΔG_{desc}^M calculado considerando complexos em que o ligante estava fora do sítio de ligação

$\Delta G_{desc}^M - s$	ΔG_{desc}^M calculado considerando complexos em que o ligante estava no sítio de ligação
ΔG_{exp}	energia livre de ligação experimental do complexo proteína–ligante
ΔG_{lig}	energia livre de ligação estimada para o complexo proteína–ligante usando a teoria do ligante implícito
ΔG^M	valor de ΔG_{lig} calculado usando a equação 4.3
ΔG_{ξ}	energia livre de confinamento do ligante ao sítio de ligação
E	energia livre de ligação estimada para o complexo proteína–ligante
$E(apo)^+$	valor mais favorável das energias livres de ligação calculadas para complexos resultantes de ancoragem com a estrutura cristalográfica <i>apo</i>
$E(c)$	energia livre de ligação calculada para complexos que compõe o conjunto de treino ou de teste do descritor de afinidades
$E(cc)$	energia livre de ligação calculada para complexos resultantes de ancoragem com um conjunto de estruturas cristalográficas
$E(cc)^+$	valor mais favorável das energias livres de ligação calculadas para complexos resultantes de ancoragem com um conjunto de estruturas cristalográficas
$E(e)$	energia livre de ligação calculada para complexos resultantes de ancoragem com um conjunto de estruturas da trajetória
$E(n)$	energia livre de ligação calculada para complexos resultantes de ancoragem com a estrutura cristalográfica correspondente ao ligante ou com a estrutura de código de PDB 1NHB (L99A) ou 3HT6 (M102Q)
E_{desc}	energia livre de ligação do complexo proteína–ligante calculada usando a equação 3.1 e os parâmetros obtidos na seção 4.2.4
EIP	2–etilfenol
E_{Su}	energia livre de ligação do complexo proteína–ligante calculada usando a equação 3.1 com parâmetros descritos no item 2 da seção 3.3
E_{tot}	energia livre de ligação do complexo proteína–ligante calculada usando a equação 3.1 com parâmetros descritos no item 3 da seção 3.3
ETP	2–etil tiofenol
E_{Vina}	energia livre de ligação do complexo proteína–ligante aproximada pela função de energia do Vina
FEP	perturbação da energia livre
G_{cav}	energia requerida para formar a cavidade do soluto dentro do solvente
G_{GB}	energia livre de polarização

G_{NP}	energia livre de interação não-polar do soluto com o solvente implícito
H α	hidrogênio ligado ao carbono α
HF	Hartree–Fock
HN	hidrogênio ligado ao nitrogênio da ligação peptídica
I4B	isobutilbenzeno
IND	indol
IPH	fenol
J0Z	benzil acetato
J1Z	tieno[3,2-b]tiofeno
JZ0	orto-cresol
JZ3	2-metoxifenol
JZ4	2-propilfenol
k_B	constante de Boltzmann
L99A	mutante L99A da lisozima do bacteriófago T4
LIE	energia de interação linear
M102Q	mutante L99A/M102Q da lisozima do bacteriófago T4
MBN	tolueno
MEM	3-etiltolueno
MXY	meta-xileno
N	nitrogênio da ligação peptídica
N3B	propilbenzeno
N4B	N-butilbenzeno
NBE	nitrosobenzeno
NCF	N-(O-tolil)cianoformamida
OEM	2-etiltolueno
Ω	volume do sítio de ligação
OPLS-AA	Optimized Potentials for Liquid Simulations – All-Atom
OXE	orto-xileno
PAN	(fenilamino)acetonitrila

PDB	Protein Data Bank
PEM	4-etiltolueno
PHD	fenilidrazina
PMF	potencial de força média
ψ	energia de interação efetiva do complexo proteína–ligante
PXY	para–xileno
PYL	etilbenzeno
RMN	ressonância magnética nuclear
RMSD	raiz do desvio quadrático médio
T	temperatura
TBB	ter–butilbenzeno
TMB	1,3,5–trimetilbenzeno
V^o	inverso da concentração padrão
V_{elet}	energia de interação eletrostática
V_{vdW}	energia de interação de van der Waals

Conteúdo

1	Introdução	p. 1
1.1	Ancoragem Molecular	p. 1
1.2	Representações da Estrutura Proteica	p. 3
1.3	Mecânica Molecular	p. 3
1.4	Solvente Implícito	p. 5
1.5	Métodos para Estimar Afinidades	p. 6
1.5.1	Teoria do Ligante Implícito	p. 7
1.5.2	Energia de Interação Linear	p. 8
1.5.2.1	Energia de Interação Linear Usando Solvente Implícito	p. 9
1.6	Lisozima do Bacteriófago T4	p. 10
2	Objetivos	p. 13
3	Métodos	p. 14
3.1	Ancoragem Molecular	p. 14
3.2	Agrupamento de Complexos	p. 15
3.2.1	Agrupamento por RMSD	p. 15
3.2.2	Agrupamento por Contatos	p. 15
3.3	Estimativa de Afinidades	p. 15

3.3.1	Calibração do Descritor de Afinidades	p. 16
3.3.2	Obtenção de Contribuições Energéticas	p. 17
3.4	Cálculos de Momentos de Dipolo e Cargas Parciais	p. 18
3.5	Dinâmica Molecular	p. 19
3.5.1	Dinâmica Estocástica em Solvente Implícito	p. 19
3.6	Predição de Deslocamento Químico	p. 20
3.7	Raiz do Desvio Quadrático Médio	p. 21
4	Resultados e Discussão	p. 22
4.1	Representações da Estrutura Proteica	p. 22
4.1.1	Seleção de Mutantes	p. 22
4.1.2	Escolha do Conjunto de Estruturas de Receptor	p. 24
4.1.3	Comparação Entre os Conjuntos de Estruturas Obtidos	p. 24
4.1.3.1	Esqueleto Proteico	p. 25
4.1.3.2	Posicionamento dos Domínios	p. 25
4.1.3.3	Cadeias Laterais	p. 28
4.2	Estimativa de Afinidades	p. 31
4.2.1	Seleção de Ligantes	p. 31
4.2.2	Parametrização de Ligantes	p. 31
4.2.3	Obtenção de Poses Falso-positivo	p. 35
4.2.4	Calibração e Teste do Descritor de Afinidades	p. 36
4.2.5	Comparação do Descritor de Afinidades Usando Ancoragem Nativa	p. 38
4.3	Complexos com Conjuntos de Estruturas Cristalográficas	p. 41

4.3.1	Comparação do Descritor de Afinidades	p. 42
4.3.2	Comparação Entre Estruturas <i>Apo</i> e os Conjuntos de Estruturas . . .	p. 44
4.3.3	Agrupamento de Complexos	p. 45
4.3.3.1	Agrupamento por RMSD	p. 46
4.3.3.2	Agrupamento por Contatos	p. 48
4.3.3.3	Comparação Entre os Métodos de Agrupamento Testados .	p. 49
4.4	Complexos com Estruturas da Trajetória de Dinâmica Molecular	p. 51
4.4.1	Aproximações para a Teoria do Ligante Implícito	p. 51
4.4.2	Análise Inicial pela Teoria do Ligante Implícito	p. 54
4.4.3	Aproximações Dentro da Teoria do Ligante Implícito	p. 57
5	Conclusões	p. 64
	Bibliografia	p. 67

1 *Introdução*

1.1 Ancoragem Molecular

O reconhecimento molecular entre um ligante e um sítio em uma proteína receptora constitui o primeiro passo de qualquer via de sinalização celular. Por causa disso, esse fenômeno tem sido explorado pela indústria farmacêutica para combater doenças e infecções. O reconhecimento molecular pode ser afetado por meio da presença de ligantes exógenos, ou seja, não produzidos pelo corpo, que interferem com a atividade do receptor, causando sua ativação ou inativação. A partir da estrutura do sítio ativo de um receptor, é possível usar modelagem computacional para criar novos ligantes, melhorar os que já existem e identificar os que possuem maior afinidade pela proteína, o que constitui um método mais racional, rápido e econômico para otimizar as propriedades de interação de uma dada molécula com uma proteína comparado a outros, como a triagem de alto desempenho (*high throughput screening*) [1].

Devido ao seu baixo custo computacional, o método mais usado atualmente é o *docking*, ou a ancoragem molecular do ligante ao receptor [2]. Neste método, diferentes orientações e conformações do ligante são exploradas para posicioná-lo em algum potencial sítio de ligação da proteína. Uma vez formados os complexos proteína–ligante, eles são ordenados a partir de uma função de pontuação, ou seja, um modelo para classificar a afinidade ou energia livre de ligação entre proteína e ligante (ΔG_{exp}).

Embora seja muito utilizada, a ancoragem apresenta duas aproximações que podem ser fontes de erro na ordenação ou na procura por sítios de ligação. A primeira delas é manter a proteína rígida. Isso possibilita rapidez na geração de complexos, mas pode levar a erros, como

falha no reconhecimento de um ligante ou estimativas errôneas da afinidade entre o ligante e a proteína. Desta forma, o paradigma chave-fechadura para a interação entre um ligante e uma proteína, que trata a proteína como se ela tivesse uma conformação única e bem definida, dever ser substituído pela consideração da flexibilidade da proteína. Métodos para incluir a flexibilidade da proteína na ancoragem serão discutidos na seção 1.2.

A segunda fonte de erro é a função de pontuação, que é aproximada e pode desconsiderar termos importantes para o processo de complexação. No programa de ancoragem AutoDock Vina [3], por exemplo, o valor da função de pontuação (E) da conformação de ligante com maior afinidade é calculado da seguinte forma:

$$E = \frac{c}{1 + 0,0585N} \quad (1.1)$$

onde N representa o número de torsões do ligante e c é a interação intermolecular, representada pela soma das contribuições energéticas oriundas de choque estérico (3 primeiros termos da equação 1.2), interações hidrofóbicas (h) e ligações de H (b):

$$c = \sum_{i < j} -0,0356e^{-(d_{ij}/0,5)^2} - 0,00516e^{-[(d_{ij}-3)/2]^2} + 0,84k - 0,0351h - 0,587b \quad (1.2)$$

$$d_{ij} = r_{ij} - R_i - R_j \quad (1.3)$$

$$k = \begin{cases} d_{ij}^2 & \text{se } d_{ij} < 0 \\ 0 & \text{se } d_{ij} \geq 0 \end{cases} \quad (1.4)$$

$$h = \begin{cases} 1 & \text{se } d_{ij} < 0,5\text{\AA} \\ 0 & \text{se } d_{ij} > 1,5\text{\AA} \end{cases} \quad (1.5)$$

$$b = \begin{cases} 1 & \text{se } d_{ij} < -0,7\text{\AA} \\ 0 & \text{se } d_{ij} > 0 \end{cases} \quad (1.6)$$

onde r_{ij} representa a distância entre os átomos i e j e R é o raio de van der Waals. h e b variam linearmente em função de d_{ij} entre os valores extremos das equações 1.5 e 1.6. A função de energia da equação 1.1 desconsidera termos importantes, como as interações eletrostáticas e a entropia, que está sub-representada pelo valor de N . Métodos mais rigorosos para calcular

afinidades serão apresentados na seção 1.5.

1.2 Representações da Estrutura Proteica

Diferentes metodologias foram criadas para incluir a flexibilidade de uma proteína na ancoragem. Entre as primeiras tentativas estão o *soft docking* [4] e a incorporação da mobilidade das cadeias laterais usando uma biblioteca de rotâmeros [5]. A primeira técnica permite que ocorra certa sobreposição entre ligante e proteína na ancoragem, abordando de forma limitada a flexibilidade proteica, enquanto a segunda não inclui a mobilidade do esqueleto da proteína, somente das cadeias laterais dos aminoácidos e, além disso, pode gerar conformações que não representam a estrutura da proteína em solução. Um método mais recente foi desenvolvido, que estabelece conexões entre domínios, interligando-os por meio de dobradiças e permitindo movimentação somente em torno dessas dobradiças [6]. Tal método, no entanto, apresenta a desvantagem de considerar como flexível apenas parte da estrutura do esqueleto proteico, desconsiderando a movimentação das cadeias laterais.

Por outro lado, existem métodos que permitem incluir a flexibilidade da estrutura proteica inteira, como a amostragem dos diferentes estados conformacionais usando dinâmica molecular [7–9] ou representação da proteína por diferentes estruturas cristalográficas [10].

1.3 Mecânica Molecular

Mecânica molecular usa campos de força para calcular a energia potencial de um sistema. Os campos de força mais usados para simular biomoléculas incluem termos que tratam de interações covalentes e não-covalentes. As interações covalentes são descritas pela soma dos termos correspondentes às energias de ligação, ângulo, diedral próprio e distorção fora do plano ou diedral impróprio. As energias de estiramento de ligação (V_{lig}) e de ângulo (V_{ang}) são geralmente aproximadas por funções harmônicas [11]:

$$V_{lig} \approx \sum_{lig} \frac{1}{2} k_b (b - b_0)^2 \quad (1.7)$$

$$V_{ang} \approx \sum_{ang} \frac{1}{2} k_{\theta} (\theta - \theta_0)^2 \quad (1.8)$$

onde k_b e k_{θ} são constantes de força, b é o comprimento da ligação entre dois átomos, θ é o ângulo de ligação entre três átomos, e b_0 e θ_0 são os valores de equilíbrio. A energia de diedral próprio (V_{died}) pode ser aproximada por uma função periódica [11]:

$$V_{died} \approx \sum_{died} \frac{1}{2} k_d [1 + \cos(n_d \phi - \delta_d)] \quad (1.9)$$

onde k_d é uma constante de força, n_d representa a periodicidade do ângulo, δ_d representa a fase do ângulo e ϕ é o ângulo de diedral próprio. A mesma equação pode ser usada para descrever a energia de diedral impróprio.

As interações não-covalentes incluem usualmente interações entre pares (i,j) de partículas, descritas pela soma dos termos eletrostáticos e de van der Waals. O cálculo da energia eletrostática (V_{elet}) é baseado na lei de Coulomb [11]:

$$V_{elet} = k_e \sum_{i < j} \frac{q_i q_j}{r_{ij}} \quad (1.10)$$

onde k_e é uma constante que depende da permissividade dielétrica do meio, q_i e q_j são as cargas parciais dos átomos i e j e r_{ij} é a distância entre os dois átomos. O cálculo da energia de van der Waals (V_{vdW}) é aproximado pela função de Lennard-Jones [11]:

$$V_{vdW} \approx \sum_{i < j} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \quad (1.11)$$

onde A_{ij} e B_{ij} são constantes cujos valores dependem dos tipos dos átomos i e j. O termo $1/r_{ij}^{12}$ se refere a interações entre nuvens de elétrons próximas, o que causa repulsão entre os átomos, enquanto o termo $1/r_{ij}^6$ se refere a energia de dispersão devida a flutuações correlacionadas nas distribuições das cargas dos dois átomos, o que promove atração entre eles [11].

Idealmente, a energia potencial de sistemas microscópicos deveria ser calculada por equações da mecânica quântica. A resolução dessas equações, no entanto, apresenta custo computacional elevado para moléculas grandes, como proteínas. A mecânica molecular realiza algumas aproximações que a tornam computacionalmente mais barata do que a mecânica quântica. Em

geral, a mecânica molecular representa os átomos com carga fixa e pontual e, portanto, não há efeito de indução, transferências de elétron ou quebras de ligação [12, 13].

1.4 Solvente Implícito

Em mecânica molecular, o solvente pode ser representado explicitamente, por meio dos átomos que o compõe, ou de forma implícita.

Formalmente, a energia de interação com o solvente implícito, G_{sol} , é dada pelo seguinte potencial de força média (PMF) [14]:

$$G_{sol} = -k_B T \ln \left(\frac{\int e^{-V_{int}^{rs}/k_B T} e^{-V^s/k_B T} dc_s}{\int e^{-V^s/k_B T} dc_s} \right) \quad (1.12)$$

onde k_B é a constante de Boltzmann, T é a temperatura em Kelvin, c_s representa as coordenadas do solvente, V_{int}^{rs} representa a energia de interação entre solvente e soluto e V^s representa a energia potencial do solvente. O PMF representa G_{sol} como o logaritmo da média de $e^{-V_{int}^{rs}/k_B T}$ com pesos dados por $e^{-V^s/k_B T}$. Dessa forma, G_{sol} fornece a energia livre de interação do soluto com as diversas configurações possíveis do solvente, dispensando a amostragem explícita dessas configurações, o que reduz o custo computacional. PMFs podem ser usados também para representar a energia de interação entre ligante e receptor, conforme será visto na seção 1.5.1.

Em simulações que usam mecânica molecular, G_{sol} pode ser estimada por [15]:

$$G_{sol} \approx G_{GB} + G_{NP} + G_{cav} \quad (1.13)$$

onde G_{GB} representa a energia livre de polarização segundo a aproximação de Born generalizada, G_{NP} representa a energia livre de interação não-polar do soluto com o solvente implícito e G_{cav} é a energia requerida para formar a cavidade do soluto dentro do solvente, incluindo o trabalho para reorganizar as moléculas de solvente ao redor do soluto e o trabalho feito contra a pressão do solvente para criar a cavidade [15].

O termo não-eletrostático da equação 1.13 pode ser calculado por [16]:

$$G_{NP} + G_{cav} = \eta A \quad (1.14)$$

onde A é a área de superfície do soluto acessível ao solvente e η é uma constante.

G_{GB} pode ser obtido pela aproximação de Born generalizada. A formulação dada por Still *et al.* [16] é usada em diversos programas de simulação:

$$G_{GB} = -\frac{1}{2} \left(1 - \frac{1}{\varepsilon} \right) \sum_{i=1}^N \sum_{j=1}^N \frac{q_i q_j}{f(r_{ij}, a_{ij})} \quad (1.15)$$

onde N é o número de átomos, ε é a constante dielétrica do meio, e $a_{ij} = (a_i a_j)^{1/2}$, onde a_i e a_j são os raios de Born dos átomos i e j respectivamente. $f(r_{ij}, a_{ij}) = (r_{ij}^2 + a_{ij}^2 e^{-D})$, onde $D = r_{ij}^2 / (2a_{ij})^2$. Devido à forma da função f , quando $i = j$, G_{GB} resulta no modelo de Born [17], que estima a energia livre de polarização para uma carga esférica. Quando duas cargas estão próximas, como um dipolo ($r_{ij} < 0,1a_{ij}$), G_{GB} é semelhante ao modelo de Onsager [18]. Quando duas cargas estão separadas ($r_{ij} > 2,5a_{ij}$), G_{GB} corresponde à soma das expressões dos modelos de Born e Coulomb [16].

1.5 Métodos para Estimar Afinidades

A afinidade de um ligante (L) por uma proteína (R) é dada pela constante de dissociação, K_d . Considerando o sistema $R + L \rightleftharpoons RL$, K_d é dado por:

$$K_d = \frac{[R][L]}{[RL]} \quad (1.16)$$

onde $[X]$ representa a concentração da espécie X . A energia livre de ligação de um complexo em condições de equilíbrio, ΔG^o , pode ser estimada por métodos computacionais e se relaciona a K_d da seguinte forma:

$$\Delta G^o = RT \ln K_d \quad (1.17)$$

onde R é a constante dos gases.

Um dos métodos computacionais mais rigorosos para estimar afinidades é a perturbação da energia livre (FEP) [19]. Nesse método, a afinidade é estimada pelo cálculo de mudanças na energia ao longo da simulação de um caminho mutacional possivelmente arbitrário e não-físico que conecta o estado do ligante livre no solvente ao estado do ligante complexado à proteína. Se esses dois estados são muito diferentes, envolvendo, por exemplo, grandes mudanças conformacionais na proteína, a convergência das energias livres será demorada e o sistema pode ficar preso em mínimos locais de energia. Nesse caso, há necessidade de longos tempos de simulação para uma boa estimativa da afinidade, o que apresenta custo computacional elevado. A seguir, serão apresentados dois métodos para estimar afinidades que resolvem de formas diferentes o problema de custo computacional elevado apresentado por FEP.

1.5.1 Teoria do Ligante Implícito

A teoria do ligante implícito, criada por Minh [20], propõe um modo rigoroso de calcular afinidades considerando diversas configurações de ligante e receptor amostradas separadamente. Esse modo de amostragem tenta resolver o problema de longos tempos de simulação necessários para realizar boas amostragens de um complexo.

Nessa teoria, ψ , a energia de interação efetiva para um determinado complexo, é dada pela subtração entre a energia potencial do complexo e as energias potenciais do ligante e da proteína isolados em solvente implícito. Nesse caso, o potencial inclui a energia livre de solvatação.

Energias de interação de diversas configurações de ligante para uma única configuração de receptor são combinadas, resultando em B , a energia de interação efetiva média para uma dada configuração de receptor:

$$B = -k_B T \ln \left(\frac{\int I_\xi e^{-\psi/k_B T} e^{-V^l/k_B T} dc_l d\xi_l}{\int I_\xi e^{-V^l/k_B T} dc_l d\xi_l} \right) \quad (1.18)$$

onde ξ_l representa os graus de liberdade que descrevem a rotação e a translação do ligante em relação a proteína, V^l representa a energia potencial do ligante e c_l representa as coordenadas do ligante. I_ξ é uma função indicadora, assumindo valor 1 quando o complexo está formado e

O quando não está. B é um PMF calculado como o logaritmo da média de $e^{-\psi/k_B T}$ com pesos dados por $I_\xi e^{-V^l/k_B T}$. Esse PMF representa o ligante do mesmo modo que o PMF de solvente implícito representa o solvente, originando o nome da teoria.

Valores de B para diferentes configurações amostradas do receptor são combinados, resultando em ΔG_{lig} , a afinidade ou energia livre de ligação estimada para o complexo:

$$\Delta G_{lig} = -k_B T \ln \left(\frac{\int e^{-B/k_B T} e^{-V^p/k_B T} dc_p}{\int e^{-V^p/k_B T} dc_p} \right) + \Delta G_\xi \quad (1.19)$$

onde c_p representa as coordenadas do receptor e V^p representa a energia potencial da proteína (as siglas ΔG_{lig} e E representam estimativas para ΔG_{exp} , a energia livre de ligação experimental; o uso da sigla ΔG_{lig} ficou restrito a estimativas baseadas na teoria do ligante implícito). A primeira parte da soma que fornece ΔG_{lig} é dada pelo logaritmo da média de $e^{-B/k_B T}$ com pesos dados por $e^{-V^p/k_B T}$. A segunda parte, ΔG_ξ , representa a energia livre de confinamento do ligante ao sítio de ligação e é dada por:

$$\Delta G_\xi = -k_B T \ln \left(\frac{\Omega}{8\pi^2 V^o} \right) \quad (1.20)$$

onde Ω é o volume do sítio de ligação e V^o é o inverso da concentração padrão.

1.5.2 Energia de Interação Linear

Energia de interação linear (LIE) é um método criado por Åqvist *et al.* [21] para estimar afinidades baseado na aproximação de resposta linear. LIE supera o problema de custo computacional elevado de FEP empregando somente simulações do ligante livre no solvente e complexado à proteína, abdicando do uso de caminhos mutacionais.

A partir de simulações de dinâmica molecular, obtêm-se contribuições para a seguinte equação:

$$E = \alpha(\langle V_{vdW}^c \rangle - \langle V_{vdW}^f \rangle) + \beta(\langle V_{elet}^c \rangle - \langle V_{elet}^f \rangle) + \epsilon(\langle A^c \rangle - \langle A^f \rangle) \quad (1.21)$$

onde $\langle \dots \rangle$ representa média configuracional, V^c se refere a energia potencial de interação do

ligante com o restante do sistema no complexo proteína–ligante, V^f se refere a energia de interação do ligante com o sistema quando o ligante está livre no solvente e A^c e A^f designam a área de superfície do ligante acessível ao solvente quando o ligante está complexado com a proteína e livre no solvente respectivamente. Como aproximação, a equação acima considera que a resposta linear é a mesma quando o solvente é a proteína e a água ou somente a água.

Na formulação original da equação 1.21 [21], o termo $\epsilon(\langle A^c \rangle - \langle A^f \rangle)$ não existe, $\beta = 0,5$ e α é calibrado para reproduzir valores de afinidades experimentais. O valor de β é baseado em um resultado conhecido: a contribuição eletrostática para a energia livre de solvatação de um íon é igual a metade de V_{elet} entre o íon e o solvente [22]. Outro resultado conhecido era a dependência linear entre a energia livre de solvatação de hidrocarbonetos e o tamanho de suas cadeias carbônicas [23], o que sugeriu que a contribuição não–polar para a energia livre de solvatação poderia estar relacionada linearmente com V_{vdW} entre o hidrocarboneto e a água.

α , β e ϵ podem ser calibrados a partir de complexos com afinidades conhecidas ou podem ser derivados teoricamente. Na derivação proposta por Su *et al.* [24], considera-se que V_{elet} e V_{vdW} apresentam distribuição Gaussiana, e chega-se aos valores $\alpha = 1$ e $\beta = 0,5$. $\epsilon = 73$ cal/molÅ², o valor teórico estimado para o trabalho de formação da cavidade do ligante em solvente [25].

O custo computacional de LIE pode ser ainda mais reduzido pelo uso de solvente implícito (descrito na seção 1.4), que dispensa a amostragem das configurações das moléculas de água.

1.5.2.1 Energia de Interação Linear Usando Solvente Implícito

A equação de LIE para solvente implícito apresenta uma formulação diferente, mas que segue a mesma ideia de resposta linear para solvente explícito [24]:

$$\begin{aligned}
 E = & \alpha(\langle V_{vdW}^c \rangle + \langle G_{NP}^c - G_{NP}^p \rangle - \langle G_{NP}^f \rangle) + \beta(\langle V_{elet}^c \rangle + 2\langle G_{GB}^c - G_{GB}^p \rangle - \langle G_{GB}^f \rangle) \\
 & + \omega(\langle G_{cav}^c - G_{cav}^p \rangle - \langle G_{cav}^f \rangle)
 \end{aligned} \tag{1.22}$$

onde G^c se refere a energia livre de interação do complexo com o solvente, G^p se refere a energia livre de interação da proteína com o solvente quando o complexo está formado, G^f se refere a energia livre de interação do ligante com o solvente quando o ligante está livre no solvente e V^c se refere a energia potencial de interação entre proteína e ligante. $(G_{GB}^c - G_{GB}^p)$ e $(G_{NP}^c - G_{NP}^p)$ representam as energias livres de interação polar e não-polar do ligante com o solvente implícito quando o ligante está complexado com a proteína. Como aproximações, a equação acima considera que a resposta linear é a mesma para $\langle V_{vdW}^c \rangle$, $\langle G_{NP}^c - G_{NP}^p \rangle$ e $\langle G_{NP}^f \rangle$, que a resposta linear é a mesma para $\langle V_{elet}^c \rangle$ e $\langle G_{GB}^f \rangle$ e o dobro para $\langle G_{GB}^c - G_{GB}^p \rangle$. A derivação teórica de α e β apresentada anteriormente também se aplica a equação 1.22.

As equações 1.21 e 1.22 podem incluir também o parâmetro τ (representado pela letra δ em trabalhos anteriores), que inclui outras contribuições possivelmente não consideradas nas equações.

1.6 Lisozima do Bacteriófago T4

A lisozima do bacteriófago T4 (número de acesso P00720 no UniProt) pertence à família das glicosil hidrolases, possui 164 aminoácidos e massa de 18.635 Da. Ela contribui para o ciclo do vírus quebrando a parede celular de bactérias pela hidrólise de ligações $\beta(1 \rightarrow 4)$ entre ácido N-acetilmurâmico e N-acetilglicosamina [26]. A lisozima é uma proteína globular, composta por uma cadeia polipeptídica que contem dois domínios ligados por uma α -hélice (figura 1.1). A atividade catalítica está contida no domínio N-terminal [27].

Após determinação da estrutura da lisozima por cristalografia [28], diversos mutantes foram criados para testar sua estabilidade estrutural [29–31]. Um desses mutantes, L99A [32], cria uma cavidade hidrofóbica de $150 \text{ \AA}^3$ no domínio C-terminal (figura 1.2a) e não altera significativamente a estrutura proteica, comparado à forma selvagem. Tal cavidade não existe na forma selvagem e não tem relação com o sítio catalítico, que está localizado em outro domínio. Essa cavidade se mostrou propícia para o estudo da formação de complexos com pequenas moléculas apolares e guiou a criação de outros mutantes para esse fim. Um exemplo é o mu-

tante L99A/M102Q [33] (figura 1.2b), que incorpora na cavidade um parceiro para ligação de H, permitindo a complexação de moléculas polares.

Figura 1.1: Estrutura da lisozima do bacteriófago T4.

Figura 1.2: Mutantes da lisozima usados no estudo da formação de complexos. (a) Mutante L99A complexado com benzeno (laranja), evidenciando a localização do sítio de ligação e da hélice F (amarelo). (b) Mutante L99A/M102Q com as posições das mutações destacadas: 99 em vermelho e 102 em rosa. Somente o domínio C-terminal está representado.

A flexibilidade dos mutantes de lisozima é parcialmente conhecida. Dados experimentais e computacionais sugerem maior mobilidade da hélice F (figura 1.2a) [34–36], que pode sofrer deslocamentos de até 5 Å, tornando o sítio de ligação artificial mais acessível [35]. A distância entre os domínios N e C-terminal varia de acordo com o mutante [37, 38], e pode sofrer mudanças de até 8 Å em simulações de dinâmica molecular [39]. Estruturas cristalográficas dos mutantes de lisozima sugerem que as áreas que viabilizam a entrada de ligantes no sítio

criado por mutação são pequenas. Desconhece-se, no entanto, o mecanismo de “respiração conformacional”, necessário para permitir a entrada de ligantes no sítio de ligação.

Os mutantes de lisozima apresentam inúmeras vantagens como modelo experimental, como tamanho pequeno, simplicidade dos sítios de ligação e disponibilidade de dados estruturais de alta qualidade [40]. Cálculos computacionais indicam que é energeticamente desfavorável manter moléculas de água no sítio de ligação artificial [41], o que indica que o sítio está dessolvatado na ausência de ligante. Dessa forma, moléculas de solvente não são deslocadas pela formação de complexos, simplificando os estudos computacionais. Os mutantes de lisozima já foram usados para testar diversas metodologias computacionais para cálculos de afinidade, como FEP [42, 43], integração termodinâmica [44, 45] e LIE [46].

2 *Objetivos*

O objetivo desse trabalho é investigar como a representação da flexibilidade de uma proteína afeta a previsão dos modos de ligação de pequenas moléculas e suas respectivas afinidades. Os mutantes de lisozima foram usados como sistema modelo e a ancoragem molecular foi empregada na obtenção de complexos.

Esse objetivo foi realizado por meio dos seguintes objetivos específicos, que visavam superar possíveis fontes de erro da ancoragem:

1. montagem de conjuntos para representar a estrutura proteica, possibilitando a inclusão de flexibilidade estrutural;
2. calibração de uma equação semelhante a equação de LIE para descrever as afinidades dos complexos obtidos, substituindo a função de pontuação da ancoragem.

O presente estudo é, portanto, retrospectivo, pois usa dados experimentais já disponíveis na literatura para estudar os melhores métodos e aproximações para estimar afinidades.

3 *Métodos*

3.1 Ancoragem Molecular

Para obtenção de complexos proteína–ligante, o programa de ancoragem AutoDock Vina [3] foi usado, com *grid* cúbico de 80 pontos e centro no sítio de ligação artificial da lisozima. 20 poses de ligante foram geradas para cada estrutura de receptor.

Para obtenção de complexos contendo potenciais poses falso–positivo (definidas na seção 4.2.3), o programa de ancoragem AutoDock 4.0 [47] foi usado, com *grid* cúbico de 60 pontos e centros variados, estabelecidos de forma a excluir o sítio de ligação do espaço disponível para ancoragem. 100 poses de ligante foram geradas para cada estrutura de receptor.

Para fazer ancoragem, não houve preparação prévia de estruturas oriundas de trajetória (seção 3.5.1). As estruturas cristalográficas foram usadas após remoção das águas, formas alternativas e moléculas além da proteína. Os hidrogênios da proteína foram construídos pelo processador de arquivos PDB do GROMACS, *pdb2gmx*. As cadeias laterais incompletas foram construídas pelo servidor WHAT IF [48, 49]. As estruturas dos ligantes foram construídas no programa Molden [50] e otimizadas usando o método Austin Model 1 (AM1) [51].

Antes da ancoragem, as estruturas da proteína foram alinhadas usando os carbonos α ($C\alpha$) dos aminoácidos LEU84, VAL87, ARG95, ALA98, ALA99, VAL111, LEU118, GLN122, ALA129 e LEU133.

3.2 Agrupamento de Complexos

3.2.1 Agrupamento por RMSD

O agrupamento ou clusterização por RMSD foi feito segundo o método particional descrito por Daura *et al.* [52], com *cutoff* de 0,5 Å. Neste método o RMSD (raiz do desvio quadrático médio, definido na seção 3.7) das coordenadas de cada par de estruturas é calculado. A estrutura com maior número de vizinhos é retirada do conjunto de estruturas junto com todos os seus vizinhos, formando o primeiro grupo, e é considerada o centro do grupo formado. O método é repetido até que todas as estruturas sejam agrupadas. Nesse trabalho, somente as coordenadas dos átomos pesados e dos hidrogênios polares dos ligantes foram consideradas no cálculo de RMSD.

3.2.2 Agrupamento por Contatos

Para medir as distâncias que caracterizavam os contatos entre proteína e ligante, foram escolhidos 23 átomos da proteína que permitissem separar diferentes configurações de receptor e 4 átomos do ligante, sendo 2 ou 3 do anel de benzeno e 2 ou 1 de grupos substituintes do anel. No total, 92 distâncias foram calculadas. O agrupamento foi feito a partir dessas distâncias pelo programa GeneCluster 2.1.7 [53], que usa um algoritmo de mapas auto-organizáveis. Este mapa é composto por nós e, nesse trabalho, cada nó estava associado a um vetor contendo 92 distâncias, a mesma dimensão dos vetores que descrevem os complexos. Os nós foram organizados em uma matriz 20x20, e cada complexo foi associado ao nó cujo vetor apresentasse a menor distância do vetor do complexo. Ao final, o conjunto de complexos de cada nó formou um grupo do agrupamento por contatos.

3.3 Estimativa de Afinidades

Os valores de E foram obtidos a partir da seguinte equação:

$$E = \alpha V_{vdW}^c + \beta_1 V_{elet}^c + \beta_2 (G_{GB}^c - G_{GB}^p) - \beta_3 G_{GB}^f + \gamma_1 (G_{NP}^c - G_{NP}^p) - \gamma_2 G_{NP}^f + \tau \quad (3.1)$$

A equação 3.1 é semelhante a equação 1.22, mas não apresenta médias configuracionais porque as contribuições energéticas foram obtidas de ponto único (detalhes na seção 3.3.2). Outra diferença da equação 1.22 é que não foi pressuposto que V_{elet}^c e G_{GB} ou V_{vdW}^c e G_{NP} apresentam a mesma resposta linear. G_{cav} está incluído no termo G_{NP} porque o programa usado para obter as contribuições energéticas calcula esses termos conjuntamente e o parâmetro τ foi incluído na equação. Por fim, na equação 3.1 cada contribuição energética tem sua resposta descrita por um parâmetro diferente.

Três conjuntos de parâmetros foram usados para obter valores de E a partir da equação 3.1:

1. parâmetros calibrados nesse trabalho para complexos com mutantes de lisozima (seção 4.2.4);
2. $\beta_1 = 0,25$, $\alpha = \beta_2 = \beta_3 = \gamma_1 = \gamma_2 = 0,50$ e $\tau = 7,73$, que são oriundos da equação de LIE calibrada por Su *et al.* [24] para descrever as afinidades de complexos entre transcriptase reversa de HIV e ligantes da classe HEPT em solvente implícito. A equação 1.22, proposta no mesmo trabalho, foi usada para obter os parâmetros, com as diferenças de que o parâmetro τ estava presente e foi pressuposto que $\langle G_{GB}^f \rangle$ apresentava o dobro da resposta linear vista para $\langle V_{elet}^c \rangle$;
3. $\alpha = \beta_1 = \beta_2 = \beta_3 = \gamma_1 = \gamma_2 = 1$ e $\tau = 0$, que equivale à soma das energias de interação totais, sem parametrização, como no cálculo de ψ proposto por Minh na teoria do ligante implícito (seção 1.5.1) [20].

3.3.1 Calibração do Descritor de Afinidades

A calibração dos parâmetros da equação 3.1 foi feita usando uma combinação dos algoritmos genético e simplex [54]. A otimização por simplex encontra mínimos locais para a função erro, que mede a diferença entre os valores de E usados como referência e obtidos pelos parâmetros da equação 3.1, e atua usando uma figura geométrica chamada simplex com $n + 1$ vértices, onde n é o número de parâmetros. A cada passo, a função erro é medida para as combinações possíveis de vértices, e o vértice de pior qualidade é substituído. O algoritmo

genético é um método para encontrar mínimos globais para a função erro, e atua imitando o processo de seleção natural: uma população de indivíduos com diferentes parâmetros, representados por cromossomos, é otimizada por meio da sobrevivência do indivíduo melhor adaptado a cada geração.

Para obtenção dos parâmetros da equação 3.1, o algoritmo genético foi utilizado primeiro. O código fonte do algoritmo genético escrito por Carroll [55] foi empregado. Usou-se uma população de 10 indivíduos, os cromossomos foram representados por 12 dígitos binários e os parâmetros α , β_1 , β_2 , β_3 , γ_1 e γ_2 variaram de -10 a 10, enquanto τ variou de -40 a 40. A população evoluiu por 10^6 gerações.

Os parâmetros dos 8 indivíduos com o menor valor de função erro obtidos pelo algoritmo genético foram usados para iniciar o algoritmo simplex [56], que realizou otimizações até a diferença entre valores da função erro de ciclos consecutivos de otimização ser inferior a 10^{-5} kcal/mol. A faixa de variação de parâmetros usada no simplex foi a mesma do algoritmo genético. Os parâmetros adotados para a equação 3.1 foram aqueles que resultaram no menor valor de função erro do último ciclo de otimização do simplex.

Optou-se por chamar a equação 3.1 de descritor de afinidades ao invés de equação de LIE devido a faixa de variação dos parâmetros, que foi maior do que a prevista teoricamente pela aproximação de resposta linear (seção 1.5.2).

3.3.2 Obtenção de Contribuições Energéticas

As contribuições energéticas necessárias para a equação 3.1 foram obtidas de um ponto único (*single-point*), usando GROMACS versão 4.5 [57] e o campo de força OPLS-AA (Optimized Potentials for Liquid Simulations – All-Atom) [58], que foi escolhido por ter sido parametrizado para reproduzir propriedades termodinâmicas. O solvente foi representado implicitamente usando a aproximação de Born generalizada (seção 1.4) [16] e o modelo Onufriev–Bashford–Case (OBC) [59] para calcular os raios de Born. Tal modelo foi escolhido por ter sido parametrizado para reproduzir as energias de solvatação de peptídeos. Complexos

tiveram suas energias minimizadas usando o algoritmo de gradiente conjugado com tolerância de $12 \text{ kcal mol}^{-1} \text{ nm}^{-1}$. As contribuições energéticas do complexo foram obtidas a partir da estrutura otimizada. As contribuições energéticas da proteína foram obtidas recalculando as contribuições para a estrutura otimizada sem o ligante, e as contribuições energéticas do ligante foram obtidas recalculando as contribuições para a estrutura otimizada sem a proteína.

Para complexos oriundos de ancoragem, não houve preparação prévia da estrutura antes da minimização de energia. Para complexos oriundos de estrutura cristalográfica, o preparo foi igual ao das estruturas cristalográficas da seção 3.1, mas sem excluir o ligante. Além disso, os hidrogênios do ligante foram construídos pelo programa Babel 2.2 [60].

O método GBr⁶ [61] foi usado para calcular as contribuições G_{GB} , pois as oscilações para o termo ($G_{GB}^c - G_{GB}^p$) obtidas por esse método em trajetórias de dinâmica molecular foram muito menores do que as oscilações calculadas pelo GROMACS.

3.4 Cálculos de Momentos de Dipolo e Cargas Parciais

Momentos de dipolo e cargas parciais de moléculas foram calculados pelos programas Gaussian 09 [62] e AMSOL 7.1 [63]. Para o programa Gaussian, foi usado o método *ab initio* Hartree–Fock (HF) e a base 6–31G*. O uso de HF foi motivado pela sua capacidade de reproduzir momentos de dipolo experimentais (veja seção 4.2.2). Para o programa AMSOL, foi usado o método semi-empírico AM1 [51], com a molécula imersa no modelo de solvente implícito Charge Model 2 (CM2) [64]. As cargas parciais calculadas por esse método assemelham-se às cargas do OPLS-AA, o que motivou seu uso. Os momentos de dipolo para moléculas com cargas parciais atribuídas por OPLS-AA foram calculados usando o programa *g_dipoles* do GROMACS. Os cálculos de dipolo e cargas parciais foram feitos com uma única configuração da molécula, obtida após otimização de geometria usando AM1.

3.5 Dinâmica Molecular

Trajetórias foram obtidas usando GROMACS 4.5 e o campo de força OPLS-AA. As estruturas iniciais de complexos usadas para as simulações foram resultantes de ancoragem. O complexo foi incluído em caixa dodecaédrica cujas arestas distavam 8 Å do complexo. As caixas foram solvatadas usando o modelo de água SPC/E (Extended Simple Point Charge) [65], escolhido por melhor reproduzir propriedades termodinâmicas de hidratação [66]. Para neutralização do sistema foram adicionados íons cloreto. Foram usadas condições periódicas de contorno e o método v-rescale [67] para controle de temperatura. O controle de pressão só foi iniciado nas simulações para amostragem de dados, e foi feito usando o método Parrinello-Rahman [68], que manteve a pressão em 1 bar. Antes da coleta de dados, foram feitas uma minimização de energia e simulações curtas de dinâmica molecular com aumento gradual da temperatura (10 K, 50 K, 100 K, 200 K e 300 K) e redução gradual da constante de força sobre os átomos pesados da proteína (239 kcal/nm^2 , 119 kcal/nm^2 , 24 kcal/nm^2 , 2 kcal/nm^2 e 0), que restringiu suas posições às coordenadas iniciais por meio de um potencial harmônico. Cada passo de aumento da temperatura ou de redução da constante de força foi precedido por uma simulação de 20 ps. Estes passos promoveram a relaxação do sistema, desfazendo contatos desfavoráveis entre a água e o complexo. Simulações para amostragem de dados duraram 10 ou 20 ns e foram feitas usando PME para tratar as energias eletrostáticas e potencial trocado (*cutoffs* de 0,8 e 1,2 nm) para tratar as energias de van der Waals. O passo de tempo foi 2 fs.

3.5.1 Dinâmica Estocástica em Solvente Implícito

Trajetórias de dinâmica estocástica foram usadas para procurar potenciais poses falso-positivo (definidas na seção 4.2.3) e para obter estruturas dos mutantes L99A e L99A/M102Q.

A coleta de dados foi precedida de minimização de energia da estrutura. As condições de simulação foram semelhantes às descritas na seção 3.5, com o solvente representado implicitamente (modelo idem ao descrito na seção 3.3.2).

Para procurar potenciais poses falso-positivo, as estruturas iniciais usadas nas simulações

foram complexos resultantes de ancoragem. Simulações para amostragem de dados duraram 2 ns.

As estruturas cristalográficas dos códigos do Protein Data Bank (PDB) 3DMV [69] para L99A e 1LI3 [33] para L99A/M102Q foram usadas para iniciar as trajetórias para obtenção de estruturas, após preparação idem a das estruturas cristalográficas da seção 3.1. Simulações para a amostragem de dados duraram 167 ns para L99A e 200 ns para L99A/M102Q.

O tempo necessário para que os valores de RMSD (definido na seção 3.7) das coordenadas atômicas dos $C\alpha$ da proteína e do domínio que contem o sítio de ligação (aminoácidos LYS83 até LEU164) se tornassem estáveis foi 5,5 ns para a trajetória de L99A e 1,5 ns para L99A/M102Q.

3.6 Predição de Deslocamento Químico

Valores de deslocamento químico (δ) do sinal de ressonância magnética nuclear (RMN) foram calculados usando o programa SPARTA+ [70], considerado um dos preditores que melhor reproduz valores de δ experimentais [71].

Para a predição de δ de uma estrutura de trajetória, não houve preparação prévia da estrutura. As estruturas cristalográficas foram preparadas da mesma forma que as estruturas cristalográficas da seção 3.1.

A reprodução dos valores de δ experimentais pelas trajetórias usadas para obtenção de estruturas (seção 3.5.1) foi testada usando um conjunto composto por estruturas coletadas a cada 500 ps após o tempo de estabilização do RMSD das coordenadas dos $C\alpha$ para L99A e a cada 50 ps após o tempo de estabilização do RMSD das coordenadas dos $C\alpha$ para L99A/M102Q.

3.7 Raiz do Desvio Quadrático Médio

A raiz do desvio quadrático médio (RMSD, da sigla em inglês) mede as diferenças entre valores de referência e calculados e é obtida pela seguinte equação:

$$RMSD(\theta_1, \theta_2) = \sqrt{\frac{\sum_{i=1}^n (x_{1,i} - x_{2,i})^2}{n}} \quad (3.2)$$

onde θ_1 e θ_2 são os conjuntos de valores que se deseja comparar, n é o número de dados de um dos conjuntos θ e $x_{m,i}$ é o valor que ocupa a posição i do conjunto θ_m . Neste trabalho, θ_1 e θ_2 podem ser conjuntos de valores de afinidades, δ ou coordenadas atômicas.

4 *Resultados e Discussão*

A seguir, serão discutidas as diferentes formas escolhidas para representar a estrutura proteica nos experimentos de ancoragem (seção 4.1). Depois, a calibração de um descritor para estimar as afinidades dos complexos será discutida (seção 4.2). Por fim, os resultados referentes a obtenção de complexos usando as diferentes representações da estrutura proteica e as aproximações para cálculo de afinidades serão apresentados e discutidos (seções 4.3 e 4.4).

4.1 Representações da Estrutura Proteica

Esta seção descreve a seleção de mutantes de lisozima (seção 4.1.1) e a montagem dos conjuntos de estruturas cristalográficas e de estruturas obtidas da trajetória (seção 4.1.2), usados para representar a estrutura dos receptores nos experimentos de ancoragem. Os conjuntos de estruturas são comparados e valores de deslocamentos químicos de sinal de RMN experimentais obtidos para L99A [36] são usados para analisar a qualidade dos modelos estruturais (seção 4.1.3).

4.1.1 Seleção de Mutantes

A lisozima possui mais de 400 estruturas cristalográficas no Protein Data Bank, com diversos mutantes e estruturas complexadas com ligantes. Foram encontrados 4 tipos de mutantes propícios para estudo de afinidades:

1. mutante L99A, criado por Eriksson *et al.* [32], que contem também as mutações C54T e C97A;

2. mutante criado por Wei *et al.* [33], denominado L99A/M102Q;
3. mutante criado por Boyce *et al.* [41], também chamado de L99A/M102Q, mas que contém também as mutações S38D e N144D;
4. mutante L99A/M102E, criado por Liu *et al.* [72], que contém também as mutações T21C, S38D, E108V, S117V, T142C, N144D e C97 modificado para S,S-(2-hidroxi-etil)tiocisteína.

L99A cria um sítio de ligação apolar, conforme mencionado na seção 1.6, enquanto os demais mutantes criam um sítio de ligação ligeiramente polar devido às mutações de M102 para GLN ou GLU.

Para decidir quais mutantes seriam utilizados neste estudo, os mutantes criados por Wei *et al.* e Boyce *et al.* foram comparados, com especial atenção às diferenças localizadas no domínio do sítio de ligação artificial. O aminoácido da posição 38 está localizado no domínio que não compõe o sítio de ligação criado pela mutação L99A; o aminoácido 144 encontra-se no domínio do sítio de ligação, mas sua cadeia lateral está orientada para fora do sítio, a uma distância de 19,7 Å do ligante benzil acetato (código de PDB 3HUK) [41]. Esses mutantes podem ser incluídos no mesmo grupo, chamado de agora em diante de M102Q, pois apresentam sítio de ligação praticamente idêntico. Já L99A/M102E foi excluído porque apresenta várias diferenças em relação a M102Q: os aminoácidos 97, 108, 117, 142 e 144 estão localizados no domínio do sítio de ligação. As cadeias laterais dos aminoácidos 108 e 117 apresentam distâncias inferiores a 10 Å do ligante benzeno (código de PDB 3GUJ) [72] e são mutações que alteram a natureza da cadeia lateral, que deixa de ser polar e passa a ser apolar nos dois casos.

Portanto, os mutantes L99A e M102Q nos itens 1, 2 e 3 acima foram escolhidos para o trabalho.

4.1.2 Escolha do Conjunto de Estruturas de Receptor

Somente complexos foram considerados, incluindo 26 estruturas cristalográficas de L99A e 32 de M102Q. As 20 estruturas escolhidas para cada mutante foram aquelas com maior RMSD de C α entre si, aquelas correspondentes aos complexos com ligantes e não–ligantes escolhidos para o estudo (tabela 4.4), e aquelas que representassem a diversidade de posições da hélice F (seção 1.6) e de valores de diedral χ_1 de VAL103, VAL111 e LEU118. A hélice F pode deixar o sítio de ligação mais ou menos fechado, enquanto as cadeias laterais dos aminoácidos mencionados podem se reorientar ao ocorrer complexação [41, 73].

O conjunto L99A contem as seguintes estruturas cristalográficas: códigos de PDB 182L [34], 183L [34], 184L [34], 185L [34], 186L [34], 187L [34], 188L [34], 1NHB [34], 2OU0 [73], 2OTY [73], 2RB2 [74], 2RAY [74], 1L83 [32], 3DN1 [69], 3DN2 [69], 3DN3 [69], 3DN4 [69], 3DN6 [69], 3HH3 [75] e 3HH5 [75].

O conjunto M102Q contem as seguintes estruturas cristalográficas: códigos de PDB 1LGX [33], 1LGW [33], 1LI2 [33], 1LI3 [33], 2RBN [74], 2RBP [74], 2RBR [74], 1XEP [76], 3HT6 [41], 3HT7 [41], 3HT8 [41], 3HT9 [41], 3HTB [41], 3HTD [41], 3HTF [41], 3HTG [41], 3HUA [41], 3HUK [41], 3HUQ [41] e 3HU8 [41].

Para montar o conjunto a partir da trajetória de dinâmica molecular, foram coletadas estruturas a cada 3 ns da trajetória de L99A e a cada 4 ns da trajetória de M102Q (seção 3.5.1), descartando o tempo de equilíbrio. Um total de 50 estruturas da trajetória foram usadas para cada mutante.

4.1.3 Comparação Entre os Conjuntos de Estruturas Obtidos

A seção a seguir descreve diferenças estruturais entre os conjuntos de estruturas cristalográficas e os conjuntos de estruturas da trajetória descritos na seção 4.1.2 que podem ter implicações nas poses obtidas por ancoragem. Serão considerados o esqueleto da proteína (seção 4.1.3.1), o posicionamento relativo dos domínios (seção 4.1.3.2) e as cadeias laterais de aminoácidos (seção 4.1.3.3). As diferenças estruturais foram associadas a valores de desloca-

mento químico de RMN (δ) preditos, que foram comparados com valores experimentais [36].

4.1.3.1 Esqueleto Proteico

A figura 4.1 mostra que os conjuntos de estruturas da trajetória possuem maior variabilidade conformacional do que os conjuntos de estruturas cristalográficas. A tabela 4.1 mostra que os RMSDs entre δ experimentais e preditos são parecidos para os dois conjuntos de estruturas. O mesmo resultado foi obtido para L99A. Os RMSDs obtidos estão próximos dos erros atribuídos ao cálculo do programa SPARTA+, indicando que os dois conjuntos representam razoavelmente bem a distribuição conformacional do esqueleto proteico.

A maior variabilidade conformacional do conjunto de estruturas da trajetória pode levar a presença de mais sítios disponíveis para complexação, chamados de sítios crípticos [9]. Tais sítios já foram encontrados, por exemplo, para β -lactamase, interleucina-2 e RNase H usando trajetórias de dinâmica molecular [77].

Tabela 4.1: RMSD (em ppm) entre deslocamentos químicos experimentais e preditos para a trajetória de M102Q ou para o conjunto de estruturas cristalográficas de M102Q.

tipo atômico	cristais	trajetória
$C\alpha^1$	$1,4 \pm 2,7$	$1,5 \pm 2,9$
$H\alpha^2$	$0,2 \pm 0,3$	$0,3 \pm 0,4$
HN^3	$0,4 \pm 0,6$	$0,5 \pm 0,9$
N^4	$2,7 \pm 4,4$	$3,0 \pm 4,7$
C^5	$1,0 \pm 1,2$	$1,0 \pm 1,2$

¹ $C\alpha$: carbono α .

² $H\alpha$: hidrogênio ligado a $C\alpha$.

³ HN : hidrogênio ligado ao nitrogênio da ligação peptídica.

⁴ N : nitrogênio da ligação peptídica.

⁵ C : carbono da ligação peptídica.

4.1.3.2 Posicionamento dos Domínios

A principal diferença entre as estruturas cristalográficas e da trajetória é a distância entre os dois domínios do receptor. A figura 4.2a representa a distância entre os domínios como a distância entre os C dos aminoácidos LYS35 e PHE104, cada um localizado em um domínio. Essa distância é aproximadamente 16 Å nas estruturas cristalográficas e é sempre menor na

Figura 4.1: Esqueletos proteicos alinhados do conjunto de estruturas cristalográficas de L99A (a) e de M102Q (b) e do conjunto de estruturas da trajetória de L99A (c) e de M102Q (d). Somente o domínio que contém o sítio de ligação está mostrado.

trajetória. Além disso, a distância entre os domínios varia na trajetória, corroborando resultados de outros autores [37–39]. A tabela 4.2 mostra que os δ preditos para a trajetória são diferentes dos preditos para o conjunto de estruturas cristalográficas somente para LYS35-N e PHE104-N considerando os desvios padrão obtidos. Para LYS35-N, somente a trajetória reproduziu os dados experimentais. Para PHE104-N, nem a trajetória nem as estruturas cristalográficas reproduziram os resultados experimentais. Os mesmos resultados foram obtidos para L99A. Isso indica que as estruturas da trajetória representam melhor a distância entre os dois domínios do receptor.

Figura 4.2: Sítio de ligação putativo entre domínios. (a) Distância entre os domínios que compõe o receptor. Para as estruturas cristalográficas, a linha tracejada representa a média e a linha pontilhada representa o desvio padrão. (b) Estrutura cristalográfica e (c) correspondente ao tempo de 54 ns da trajetória de M102Q, representadas com o esqueleto e a superfície molecular. O quadrado preto indica a localização do falso sítio.

Tabela 4.2: Média $\pm$ desvio padrão dos deslocamentos químicos calculados (em ppm) para M102Q.

átomo	experimental	cristais	trajetória
LYS35-C	176,7	$176,6 \pm 0,2$	$176,4 \pm 0,4$
LYS35-HN	8,6	$8,3 \pm 0,0$	$8,4 \pm 0,2$
LYS35-N	124,9	$121,8 \pm 0,2^1$	$123,8 \pm 1,6$
PHE104-C	176,9	$176,7 \pm 0,2$	$176,9 \pm 0,5$
PHE104-HN	7,9	$8,5 \pm 0,1$	$8,4 \pm 0,3$
PHE104-N	126,5	$119,8 \pm 0,2$	$117,7 \pm 1,6$

¹Deslocamento químico predito que difere do experimental acima do erro atribuído ao cálculo do SPARTA+ está indicado em negrito.

O posicionamento dos domínios nas estruturas cristalográficas pode gerar um sítio de ligação (figura 4.2b), conforme discutido na seção 4.3.1. Esse sítio não aparece na trajetória (figura 4.2c), sugerindo que ele é um artefato da cristalização.

4.1.3.3 Cadeias Laterais

A média do número de poses no sítio de ligação obtido por ancoragem foi menor para os complexos obtidos usando estruturas da trajetória. Mudanças nos diedrais χ_1 de ILE78, LEU84 e TYR88 podem obstruir transitoriamente o sítio de ligação durante a dinâmica molecular, conforme demonstrado na figura 4.3 e explicado em detalhe nos próximos parágrafos. Resultados semelhantes foram vistos para a trajetória de L99A. As transições de diedral χ_1 mencionadas não foram vistas nos conjuntos de estruturas cristalográficas, já que os receptores foram cristalizados na presença de um ligante que ocupava o sítio.

A cadeia lateral de ILE78 bloqueia o sítio de ligação quando $\chi_1 = -180^\circ$. Por exemplo, nos tempos 90, 94 e 102 ns (figura 4.3) não são observadas poses de ligante no sítio. A tabela 4.3 mostra que somente δ predito para C de ILE78 na trajetória reproduz o valor de δ experimental, sugerindo que a trajetória representa melhor a conformação de ILE78.

Tabela 4.3: Média $\pm$ desvio padrão dos deslocamentos químicos calculados (em ppm) para M102Q.

átomo	experimental	cristais	trajetória
ILE78-C	177,2	178,6 $\pm$ 0,0¹	177,7 $\pm$ 0,6
ILE78-HN	7,5	7,8 $\pm$ 0,1	7,7 $\pm$ 0,3
ILE78-N	121,8	123,0 $\pm$ 0,3	122,4 $\pm$ 0,8

¹Deslocamento químico predito que difere do experimental acima do erro atribuído ao cálculo do SPARTA+ está indicado em negrito.

A cadeia lateral de LEU84 bloqueia o sítio de ligação quando $\chi_1 = -180^\circ$, como por exemplo no tempo 74 ns (figura 4.3). Mudanças no valor de diedral χ_1 de LEU84 durante a trajetória não causam alterações nos δ preditos para os átomos desse aminoácido, o que levou os δ preditos para a trajetória e para o conjunto de estruturas cristalográficas a apresentarem valores semelhantes. Esses valores, por sua vez, se assemelham aos experimentais.

Figura 4.3: Aminoácidos associados a obstrução do sítio de ligação. (a) Número médio de poses no sítio de ligação para cada estrutura da trajetória de M102Q usada nos experimentos de ancoragem. (b, c e d) Trajetórias de diedral χ_1 dos aminoácidos indicados na legenda.

A cadeia lateral de TYR88 bloqueia o sítio de ligação quando $\chi_1 = -60^\circ$, como nos tempos 182, 186, 190 e 194 ns (figura 4.3), em que não são observadas poses de ligante no sítio. A figura 4.4 mostra que a mudança no valor de diedral χ_1 de TYR88 na trajetória de M102Q piora a descrição de δ em 3 casos ($C\alpha$, N e HN). Isso é um indicativo de que o diedral alternativo de TYR88 visto na trajetória é pouco frequente e sua visitação pode ser fruto de imprecisões no campo de força usado.

Portanto, os resultados obtidos sugerem que as trajetórias e os conjuntos de estruturas cristalográficas representam a distribuição conformacional do receptor. Apesar do comportamento da trajetória de diedral χ_1 de TYR88, a representação da distância entre os dois domínios e da flexibilidade da cadeia lateral de ILE78 são mais apropriadas na trajetória.

Figura 4.4: Trajetórias de deslocamento químico predito para os átomos de TYR88 indicados na legenda. A linha que alterna traços e pontos indica o deslocamento químico experimental. As linhas tracejada e pontilhada indicam respectivamente a média e o desvio padrão dos deslocamentos químicos preditos para o conjunto de estruturas cristalográficas de M102Q.

4.2 Estimativa de Afinidades

Esta seção descreve a calibração (seção 4.2.4) e uma aplicação inicial (seção 4.2.5) do descritor de afinidades. Antes, a seleção (seção 4.2.1) e parametrização dos ligantes (seção 4.2.2) e a obtenção de poses falso-positivo (seção 4.2.3), necessárias para a calibração, são relatadas.

4.2.1 Seleção de Ligantes

Para L99A, são conhecidos os ΔG_{exp} de 21 ligantes, dentre os quais 14 possuem também estruturas cristalográficas, e 51 moléculas foram identificadas como não-ligantes [33, 69, 73, 74, 78, 79]. Para M102Q, são conhecidos os ΔG_{exp} de 17 ligantes, dentre os quais 15 possuem também estruturas cristalográficas, e 19 moléculas foram identificadas como não-ligantes ou ligantes fracos, 4 delas possuindo estruturas cristalográficas [33, 41, 74, 80]. A tabela 4.4 mostra os ligantes e não-ligantes escolhidos para esse trabalho, enquanto a figura 4.5 mostra as estruturas de alguns deles. A maioria dos ligantes (a palavra “ligantes” se refere a ligantes e não-ligantes, a não ser que seja explicitamente declarado) é constituída de um anel de benzeno com um ou dois substituintes. Ligantes cujas estruturas cristalográficas disponíveis continham moléculas de água no sítio de ligação, como fenol e 4,5,6,7-tetraidro-1h-indol para M102Q, foram excluídos do estudo, pois o uso de solvente implícito prejudica o cálculo das contribuições energéticas para essas estruturas.

4.2.2 Parametrização de Ligantes

Optou-se por construir os parâmetros dos ligantes manualmente, pois há relatos de erros em programas como Antechamber ou o servidor PRODRG, usados para essa finalidade [41, 81]. Os ligantes 2EP, J0Z, CHX, 4VP, PHD, 4CP, 3MP, J1Z, PAN, NCF e JZ3 não apresentavam parâmetros para determinados ângulos, ligações ou diedrais próprios no OPLS-AA. Tais parâmetros foram incluídos como aproximação de funções químicas semelhantes. Alguns tipos atômicos não apresentavam parametrização disponível para solvente implícito. Tais parâmetros

Tabela 4.4: Ligantes e não-ligantes (em **negrito**) de L99A e M102Q escolhidos para o estudo, com a estrutura cristalográfica (código de PDB) e sigla correspondentes indicadas.

L99A			M102Q		
ligante	sigla	estrutura	ligante	sigla	estrutura
benzeno	BNZ	1L83	catecol	CAQ	1XEP
etilbenzeno	PYL	1NHB	(fenilamino)acetonitrila	PAN	2RBN
orto-xileno	OXE	188L	tieno[3,2-b]tiofeno	J1Z	3HUQ
para-xileno	PXY	187L	benzil acetato	J0Z	3HUK
N-butilbenzeno	N4B	186L	2-etoxifenol	2EP	3HU8
isobutilbenzeno	I4B	184L	2-propilfenol	JZ4	3HTB
indol	IND	185L	5-cloro-2-metilfenol	CMI	3HT8
meta-xileno	MXY	- ¹	2-etilfenol	EIP	3HT7
propilbenzeno	N3B	-	orto-cresol	JZ0	3HT6
2-etiltolueno	OEM	-	2-fluoroanilina	1AN	1LGW
3-etiltolueno	MEM	-	3-metilpirrol	3MP	-
4-etiltolueno	PEM	-	tolueno	MBN	-
tolueno	MBN	-	2-metoxifenol	JZ3	3HT9
fenol	IPH	-	nitrosobenzeno	NBE	3HU9
cicloexano	CHX	-	4-cloro-1h-pirazol	4CP	3HTF
1,3,5-trimetilbenzeno	TMB	-	4-vinilpiridina	4VP	-
ter-butilbenzeno	TBB	-	2-aminofenol	2AP	-
2-fluoroanilina	1AN	-	fenilidrazina	PHD	-
anilina	ANL	-	2-etil tiofenol	ETP	-
3-metilpirrol	3MP	-	N-(O-tolil)cianofor- mamida	NCF	-

¹Complexo sem estrutura cristalográfica disponível.

Figura 4.5: Estruturas de alguns ligantes de M102Q escolhidos para o estudo. Os ligantes estão coloridos de acordo com o tipo atômico: carbono em verde, hidrogênio em branco, nitrogênio em azul, oxigênio em vermelho e enxofre em amarelo.

também foram derivados de tipos atômicos semelhantes.

No OPLS-AA, as cargas para cada tipo atômico são atribuídas segundo sua função orgânica. Para ligantes apolares ou com somente um grupo polar, as cargas disponíveis no OPLS-AA foram usadas na parametrização. No entanto, foi pressuposto que as cargas oferecidas pelo OPLS-AA seriam inadequadas para ligantes com mais de um grupo polar. Isso motivou a realização de cálculos de dipolo para estes ligantes usando as cargas atribuídas pelo OPLS-AA, o método *ab initio* HF (referência quanto-mecânica) ou o método semi-empírico AM1. O cálculo de dipolo foi escolhido, em detrimento de outras expansões como quadrupolo e octapolo, porque é o termo de maior relevância na determinação das interações eletrostáticas.

A seguinte metodologia foi empregada para determinar as cargas parciais dos ligantes com mais de um grupo polar:

1. comparação do momento de dipolo do ligante e seus componentes calculados na referência quanto-mecânica e obtidos usando as cargas do OPLS-AA para descrever o ligante; se estes fossem parecidos qualitativamente e quantitativamente (diferença inferior a 50%), as cargas fornecidas pelo OPLS-AA eram mantidas, caso contrário prosseguia-se para o próximo passo;
2. comparação do momento de dipolo do ligante e seus componentes calculados na referência quanto-mecânica e em AM1; se estes fossem parecidos, as cargas do AM1 eram usadas na parametrização do ligante, caso contrário prosseguia-se para o próximo passo;
3. uso das cargas calculadas na referência quanto-mecânica (método HF).

16 ligantes foram submetidos ao procedimento acima (tabela 4.5): 4VP, 4CP, EIP, CMI e J1Z permaneceram com as cargas fornecidas pelo OPLS-AA; JZ0, 2EP, J0Z, CAQ, PHD, 1AN, PAN, NCF e JZ3 foram parametrizados com as cargas calculadas por AM1; e 2AP e NBE foram parametrizados com as cargas calculadas pela referência quanto-mecânica. Valores experimentais de momento de dipolo estão disponíveis somente para PHD e CAQ, e são de 1,65–1,70 Debye [82] e $2,62 \pm 0,03$ Debye [83] respectivamente, semelhantes aos valores obtidos usando

a referência quanto-mecânica, o que sugere que os momentos de dipolo calculados por esse método são confiáveis.

Tabela 4.5: Valores de momento de dipolo e seus componentes (em Debye) calculados usando as cargas fornecidas pelo OPLS-AA, o método HF ou o método AM1.

ligante	origem da carga	componentes do dipolo			
		total	x	y	z
4VP	OPLS	2,7	-2,6	-0,6	-0,1
	HF_{OPLS}^1	2,6	-2,6	-0,4	-0,1
4CP	OPLS	2,8	2,3	-1,5	0,0
	HF_{OPLS}	2,5	2,3	-0,9	0,0
EIP	OPLS	2,5	1,6	-1,9	-0,4
	HF_{OPLS}	1,8	1,2	-1,3	-0,2
CMI	OPLS	0,5	0,5	-0,1	0,0
	HF_{OPLS}	0,9	0,9	0,1	0,0
J1Z	OPLS	0,0	0,0	0,0	0,0
	HF_{OPLS}	0,0	0,0	0,0	0,0
JZ0	OPLS	1,6	-1,6	0,0	0,0
	HF_{OPLS}	1,3	-1,2	0,4	0,0
	AM1	0,9	0,6	0,7	0,0
	HF_{AM1}^2	1,3	1,0	0,8	0,0
2EP	OPLS	3,3	2,6	1,8	0,9
	HF_{OPLS}	2,5	1,1	1,9	1,1
	AM1	2,0	1,6	0,8	-1,0
	HF_{AM1}	2,5	2,2	0,7	-1,0
J0Z	OPLS	5,1	2,3	-4,5	-0,2
	HF_{OPLS}	4,8	1,4	-4,6	0,4
	AM1	4,3	3,6	2,3	0,6
	HF_{AM1}	4,8	4,0	2,6	0,6
CAQ	OPLS	3,5	-2,8	-2,1	0,0
	HF_{OPLS}	2,1	-1,7	-1,3	0,0
	AM	2,1	-1,8	-1,1	0,0
	HF_{AM1}	2,1	-1,8	-1,1	0,0
PHD	OPLS	3,4	-3,3	-0,9	0,2
	HF_{OPLS}	1,4	1,3	-0,6	0,2
	AM1	0,9	-0,2	0,9	0,1
	HF_{AM1}	1,4	-0,4	1,4	0,2
1AN	OPLS	2,2	0,2	-0,1	2,2
	HF_{OPLS}	1,9	-1,0	1,1	1,2
	AM1	2,0	1,2	0,9	1,2
	HF_{AM1}	1,9	1,4	0,6	1,2

continua

¹ HF_{OPLS} : cálculo com HF realizado usando a mesma orientação do ligante utilizada no cálculo usando as cargas do OPLS-AA.

² HF_{AM1} : cálculo com HF realizado usando a mesma orientação do ligante utilizada no cálculo com AM1.

continuação					
ligante	origem da carga	componentes do dipolo			
		total	x	y	z
PAN	OPLS	2,9	0,3	2,7	1,1
	HF_{OPLS}	3,0	-0,1	2,7	1,2
	AM1	3,0	-0,5	0,1	3,0
	HF_{AM1}	3,0	-0,3	0,1	3,0
NCF	OPLS	6,6	5,3	3,9	1,0
	HF_{OPLS}	5,1	4,3	2,7	0,3
	AM1	4,9	-3,2	2,8	2,4
	HF_{AM1}	5,1	-3,4	2,9	2,5
JZ3	OPLS	2,4	-1,1	1,1	1,8
	HF_{OPLS}	2,4	-1,3	-1,2	1,7
	AM1	2,1	1,4	1,0	-1,2
	HF_{AM1}	2,4	2,0	0,8	-1,2
2AP	OPLS	2,4	0,7	1,2	-1,9
	HF_{OPLS}	1,5	-0,6	1,4	0,2
	AM1	1,2	-1,2	0,1	0,2
	HF_{AM1}	1,5	-1,4	-0,5	0,2
NBE	OPLS	2,9	1,1	0,0	-2,7
	HF_{OPLS}	3,5	2,8	0,0	-2,1
	AM1	1,8	-1,3	1,2	0,0
	HF_{AM1}	3,5	-3,5	-0,4	0,0

4.2.3 Obtenção de Poses Falso-positivo

Um dos objetivos do trabalho é classificar um número grande de poses geradas por ancoragem. Assim, poses falso-positivo, ou seja, poses instáveis de moléculas que são genuinamente ligantes, geradas por imprecisões da ancoragem e em sítios diferentes do modo de ligação cristalográfico ou “nativo”, foram incluídas na calibração e teste do descritor de afinidades. Desse modo, espera-se que o descritor seja capaz de discriminar entre poses “nativas” e poses falso-positivo do mesmo ligante. As potenciais poses falso-positivo foram geradas por ancoragem em estruturas cristalográficas e submetidas a dinâmicas moleculares. A pose foi considerada falso-positivo quando o ligante passava pelo menos 20% do tempo total da trajetória (desconsiderando o tempo de equilíbrio) dissociado da proteína. O ligante foi considerado dissociado quando apresentava valor máximo de área de superfície exposta ao solvente. No total, 10 poses falso-positivo de diferentes ligantes foram obtidas para cada mutante.

4.2.4 Calibração e Teste do Descritor de Afinidades

Para calibrar e testar o descritor de afinidades usa-se um conjunto de treino, composto por ligantes que são usados na calibração da equação 3.1 (seção 3.3), e um conjunto de teste, composto por ligantes que não estavam incluídos na calibração. Os 10 ligantes usados no conjunto de treino da equação do descritor foram escolhidos baseado em existência de estrutura cristalográfica do complexo (tabela 4.4) e diversidade de ΔG_{exp} e de estrutura. O conjunto de treino também incluiu 10 poses falso-positivo (5 de cada mutante) dos mesmos ligantes incluídos na calibração.

Para obter as contribuições energéticas da equação 3.1, a estrutura cristalográfica do complexo proteína-ligante (ligantes com estrutura cristalográfica disponível estão mostrados na tabela 4.4) foi utilizada. Para ligantes sem estrutura cristalográfica disponível, o complexo foi obtido a partir da pose com pontuação mais favorável resultante da ancoragem do ligante às estruturas *holo* 1NHB (L99A) ou 3HT6 (M102Q), escolhidas por apresentarem ligantes com estrutura semelhante a dos ligantes sem estrutura cristalográfica disponível. As afinidades calculadas a partir das estruturas mencionadas são chamadas de $E_{desc}(c)$, E_{desc} por usar o descritor de afinidades a ser calibrado e $E(c)$ por usar estruturas cristalográficas (nas siglas $E_A(B)$, A se refere ao descritor de afinidades e B se refere ao complexo utilizado). A amostragem do espaço configuracional ficou restrita aos diferentes modos de ligação cristalográficos e falso-positivo. Nos casos em que foi considerado mais de um modo de ligação cristalográfico para o mesmo ligante, $E_{desc}(c)$ total foi calculada a partir do logaritmo natural da soma dos exponenciais das $E_{desc}(c)$ calculadas para cada um dos diferentes modos. O não-ligante 4CP, apesar de apresentar estrutura cristalográfica em que cada um dos modos de ligação tinha 100% de ocupância, teve seu valor de $E_{desc}(c)$ total calculado segundo o método descrito.

Os parâmetros obtidos para o descritor de afinidades da equação 3.1 foram $\alpha = 0,30$, $\beta_1 = 0,09$, $\beta_2 = -0,53$, $\beta_3 = 2,63$, $\gamma_1 = 1,18$, $\gamma_2 = -3,43$ e $\tau = -29,32$. Estes resultaram em desvio médio entre ΔG_{exp} e $E_{desc}(c)$ de 0,7 kcal/mol para o conjunto de treino (tabela 4.6) e de 1,8 kcal/mol para o conjunto de teste (tabela 4.7). Além do descritor de afinidades, os desvios

podem ser atribuídos a um possível erro do Vina, que pode não ter encontrado a geometria mais relevante do complexo para os ligantes que não possuíam estrutura cristalográfica.

Tabela 4.6: Energias livres de ligação (em kcal/mol) para complexos do conjunto de treino.

L99A			M102Q		
ligante	ΔG_{exp}	$E_{desc}(c)$	ligante	ΔG_{exp}	$E_{desc}(c)$
I4B	-6,4	-7,7	JZ0	-4,7	-5,2
OXE	-4,6	-5,7	J0Z	-4,7	-4,7
PEM* ¹	-5,4	-5,4	1AN	-5,5	-4,6
IND	-4,9	-4,0	CAQ	-4,4	-4,5
BNZ	-5,2	-3,8	CMI	-5,3	-3,8
<u>IND</u> * ²	> -2,0	-2,4	<u>JZ0</u> *	> -2,0	-2,0
<u>OXE</u> *	> -2,0	-2,2	<u>1AN (1)</u> *	> -2,0	-1,7
<u>I4B</u> *	> -2,0	-2,0	<u>CAQ</u> *	> -2,0	-1,4
<u>PEM</u> *	> -2,0	-1,7	<u>1AN (2)</u> *	> -2,0	1,3
<u>BNZ</u> *	> -2,0	-1,6	<u>CMI</u> *	> -2,0	2,2

¹ Asteriscos marcam complexos sem estrutura cristalográfica disponível.

² Poses falso-positivo estão sublinhadas.

A tabela 4.7 mostra que as poses falso-positivo de PEM, N3B e J1Z não foram reconhecidas. Contudo, o desvio dessas poses foi de no máximo 1 kcal/mol, o que não prejudica a separação entre poses falso-positivo e poses genuinamente ligantes, pois as últimas sempre receberam $E_{desc}(c)$ mais favoráveis.

Os não-ligantes 3MP e IPH (L99A), e 4VP e NBE (M102Q) foram reconhecidos. Por outro lado, há não-ligantes que apresentaram $E_{desc}(c)$ muito favoráveis, como ETP (M102Q) e TBB (L99A), indicando uma falha do descritor de afinidades. A falta de reconhecimento desses não-ligantes pode ser devida a um desbalanço entre energias de interação do não-ligante com a proteína e com o solvente implícito, conforme discutido na seção 4.2.5.

Portanto, os resultados obtidos nessa seção mostram que o descritor de afinidades reproduz razoavelmente bem os dados experimentais para ligantes do conjunto de treino e de teste e é capaz de distinguir energeticamente poses genuinamente ligantes de poses falso-positivo.

Tabela 4.7: Energias livres de ligação (em kcal/mol) para complexos do conjunto de teste.

L99A			M102Q		
ligante	ΔG_{exp}	$E_{desc}(c)$	ligante	ΔG_{exp}	$E_{desc}(c)$
N4B	-6,7	-7,5	J1Z	-4,9	-7,5
N3B* ¹	-6,5	-6,3	JZ4	-5,6	-5,5
PYL	-5,7	-5,6	EIP	-4,8	-5,1
OEM*	-4,5	-5,5	MBN*	-5,2	-4,9
MEM*	-5,1	-5,5	3MP*	-5,2	-4,2
PXY	-4,6	-4,9	2EP	-4,3	-4,0
MXY*	-4,7	-4,6	PAN	-5,8	-3,3
MBN*	-5,5	-3,8	J1Z*	> -2,0	-2,9
PEM* ²	> -2,0	-3,0	JZ4 (1)*	> -2,0	-0,1
N3B*	> -2,0	-2,7	JZ4 (2)*	> -2,0	0,0
MBN*	> -2,0	-2,2	2EP*	> -2,0	0,5
PXY*	> -2,0	0,1	PAN*	> -2,0	3,2
OXE*	> -2,0	0,4	ETP*	> -2,0	-7,2
TBB* ³	> -2,0	-6,5	JZ3	> -2,0	-5,2
CHX*	> -2,0	-5,8	4CP	> -2,0	-4,7
TMB*	> -2,0	-4,7	NCF*	> -2,0	-4,4
1AN*	> -2,0	-3,7	2AP*	> -2,0	-4,2
ANL*	> -2,0	-3,4	PHD*	> -2,0	-3,7
3MP*	> -2,0	-3,2	4VP*	> -2,0	-3,2
IPH*	> -2,0	-2,8	NBE	> -2,0	-3,0

¹Asteriscos marcam complexos sem estrutura cristalográfica disponível.

²Poses falso-positivo estão sublinhadas.

³Não-ligantes estão indicados em negrito.

4.2.5 Comparação do Descritor de Afinidades Usando Ancoragem Nativa

As afinidades calculadas pelo descritor obtido também foram comparadas a função de energia do Vina [3]. Nestes testes o ligante foi ancorado à estrutura de receptor correspondente ao cristal do complexo (tabela 4.4), também conhecido como ancoragem nativa [84]. Caso não houvesse estrutura nativa disponível para o complexo, a estrutura de código de PDB 1NHB (L99A) ou 3HT6 (M102Q) era usada.

O complexo de cada ligante com energia mais favorável atribuída pelo Vina foi eleito como modo nativo e escolhido para os testes dessa seção. Foi denominado $E_{Vina}(n)$ o valor atribuído pela função de energia do Vina para este modo nativo, E_{Vina} devido ao método para obter o valor da afinidade e $E(n)$ devido ao uso de complexos obtidos de ancoragem nativa. A afinidade a-

tribuída a esses complexos pelo descritor de afinidades foi denominada $E_{desc}(n)$. Logo, a única variação entre $E_{Vina}(n)$ e $E_{desc}(n)$ é o método para calcular afinidades.

A figura 4.6 compara os valores de ΔG_{exp} com os valores de $E_{Vina}(n)$ e $E_{desc}(n)$. $E_{Vina}(n)$ atribui valores mais favoráveis do que ΔG_{exp} . Os desvios médios para os dados apresentados na figura, considerando apenas os ligantes (não–ligantes excluídos) presentes no conjunto de teste do descritor, foram de 1,14 kcal/mol para $E_{Vina}(n)$ e 1,07 kcal/mol para $E_{desc}(n)$, sugerindo que $E_{desc}(n)$ reproduz ligeiramente melhor os valores de ΔG_{exp} . O bom desempenho do Vina também é devido à presença de parte dos mesmos complexos no banco de dados PDBbind, usado para calibração de sua função de energia [3]. Além da contribuição da função de energia, os desvios podem ser atribuídos a geometria encontrada pelo Vina, que pode não ser relevante em uma distribuição de equilíbrio. A geometria com $E_{Vina}(n)$ mais favorável encontrada para NCF, por exemplo, está fora do sítio de ligação cristalográfico (figura 4.7), sendo provavelmente pouco relevante para a distribuição.

Figura 4.6: Comparação entre energias livres de ligação experimentais e calculadas por Vina (a) ou pelo descritor de afinidades (b). Quadrados indicam o conjunto de treino do descritor de afinidades, enquanto losangos marcam o conjunto de teste. A linha pontilhada corresponde a equação $x=y$.

O resultado a seguir é motivado por experimentos feitos na indústria farmacêutica para separar possíveis ligantes de não–ligantes. Nesse experimento as moléculas testadas são ordenadas de forma decrescente em função da afinidade estimada, e aquelas que ocupam as primeiras posições da lista são consideradas ligantes e eventualmente testadas experimental-

Figura 4.7: Poses do não–ligante NCF obtidas por ancoragem à estrutura do PDB 3HT6. Azul marca a pose de maior afinidade, vermelho marca a pose localizada no sítio de ligação cristalográfico.

mente. Espera-se que não–ligantes apresentem energia livre de ligação menos favorável do que a dos ligantes, ocupando as últimas posições do ordenamento relativo.

A tabela 4.8 mostra que, nos ordenamentos obtidos, $E_{Vina}(n)$ apresenta melhor desempenho para L99A, reconhecendo 6 não–ligantes, enquanto $E_{desc}(n)$ reconhece 4. No entanto, $E_{desc}(n)$ apresenta melhor desempenho para M102Q, reconhecendo 5 não–ligantes, enquanto $E_{Vina}(n)$ reconhece 4.

L99A apresenta não–ligantes apolares (TBB, CHX e TMB) e polares (3MP, ANL, 1AN e IPH). O ordenamento usando $E_{desc}(n)$ deixa de reconhecer somente os não–ligantes apolares. Uma possível razão é o desbalanço entre as energias de interação entre proteína e ligante e deste com o solvente implícito. Um argumento a favor dessa ideia é que os 3 não–ligantes são apolares e devem apresentar energias de interação mais favoráveis com o sítio de ligação apolar de L99A do que com a água, o que os leva a ocupar boas posições no ordenamento. TBB, CHX e TMB podem ser não–ligantes devido a ausência de um caminho que permita a entrada no sítio de ligação. A lisozima tem um sítio de ligação fechado, e precisa sofrer um movimento de respiração ou abertura para permitir a entrada de ligantes (seção 1.6). Tal movimento pode não ser suficiente para permitir a passagem desses não–ligantes, já que TBB apresenta substituinte volumoso no anel de benzeno, e CHX não apresenta anel planar. Os não–ligantes polares de L99A, por outro lado, são capazes de acessar o sítio de ligação, pois são ligantes de M102Q.

Tabela 4.8: Ligantes e não–ligantes (em negrito) de L99A e M102Q ordenados de forma decrescente a partir de energias livres de ligação.

L99A			M102Q		
ΔG_{exp}	$E_{Vina}(\mathbf{n})$	$E_{desc}(\mathbf{n})$	ΔG_{exp}	$E_{Vina}(\mathbf{n})$	$E_{desc}(\mathbf{n})$
N4B	N4B	I4B	PAN	J0Z	ETP
N3B	IND	N4B	JZ4	JZ4	J1Z
I4B	I4B	TBB	1AN	PAN	MBN
PYL	N3B	N3B	CMI	JZ0	JZ4
MBN	OEM	OEM	MBN	MBN	EIP
PEM	PYL	MEM	3MP	2AP	JZ0
BNZ	MEM	CHX	J1Z	PHD	CAQ
MEM	OXE	OXE	EIP	2EP	3MP
IND	1AN	PEM	J0Z	1AN	4VP
MXY	MBN	PYL	JZ0	NBE	J0Z
PXY	MXY	PXY	CAQ	4VP	2EP
OXE	PEM	MXY	2EP	EIP	NBE
OEM	PXY	TMB	PHD	ETP	1AN
IPH ¹	ANL	MBN	4CP	CMI	PHD
CHX	IPH	IND	4VP	CAQ	CMI
TMB	BNZ	3MP	2AP	NCF	JZ3
TBB	TBB	ANL	NCF	JZ3	PAN
ANL	CHX	1AN	ETP	J1Z	NCF
3MP	TMB	BNZ	JZ3	3MP	4CP
1AN	3MP	IPH	NBE	4CP	2AP

¹ A linha pontilhada separa ligantes (acima) de não–ligantes (abaixo) denominados de acordo com o experimento de ordenamento.

4.3 Complexos com Conjuntos de Estruturas Cristalográficas

Esta seção descreve experimentos de ancoragem cruzada [84] obtidos das estruturas cristalográficas descritas na seção 4.1.2. As afinidades calculadas pelo descritor obtido foram comparadas a função de energia do Vina [3] (seção 4.3.1), como na seção anterior (4.2.5). Em seguida, afinidades estimadas por ancoragem com estruturas *apo* e com conjuntos de estruturas cristalográficas foram comparadas (seção 4.3.2). Por último, diferentes métodos foram testados para agrupar os complexos gerados pela ancoragem com estruturas cristalográficas (seção 4.3.3), com o objetivo de reduzir o número de cálculos necessários para descrever as afinidades.

4.3.1 Comparação do Descritor de Afinidades

Os complexos obtidos com os conjuntos de estruturas de receptor (seção 4.1.2) tiveram suas afinidades determinadas pelo Vina, denominada $E_{Vina}(cc)$, ou pelo descritor de afinidades, $E_{desc}(cc)$.

Para cada estrutura de receptor 20 poses foram geradas pelo Vina, em um total de 400 poses (20 x 20 estruturas de receptor) para cada ligante. A figura 4.8 mostra a diversidade de poses obtidas dentro e fora do sítio de ligação. As poses com valores mais favoráveis de $E_{desc}(cc)$ encontram-se concentradas no sítio de ligação cristalográfico, enquanto as poses fora dele apresentam valores menos favoráveis. Assim, o descritor de afinidades calibrado é capaz de reconhecer o sítio de ligação. Resultados semelhantes foram obtidos para os demais ligantes.

Figura 4.8: Superposição das poses do ligante JZ4 (M102Q) obtidas por ancoragem a um conjunto de estruturas cristalográficas. Somente uma das estruturas de receptor do conjunto está representada. Poses estão coloridas de acordo com $E_{desc}(cc)$, com a escala de mais para menos favorável: vermelho, laranja, amarelo, verde, ciano e azul.

Para alguns ligantes, foram atribuídas $E_{Vina}(cc)$ e $E_{desc}(cc)$ favoráveis para poses fora do sítio de ligação. Essas poses encontram-se no falso sítio de ligação descrito na seção 4.1.3.2. Portanto, embora com afinidade elevada, tais poses não devem ser observadas experimentalmente. A presença dessas poses não interfere nos resultados obtidos na seção 4.3.

A figura 4.9 mostra histogramas das afinidades calculadas para os complexos de receptor

com BNZ (L99A) ou JZ4 (M102Q). Resultados semelhantes foram obtidos para outros ligantes não mostrados. Os valores de $E_{Vina}(cc)$ são pouco dispersos, e os valores atribuídos para complexos contendo pose falso-positivo e pose cristalográfica chegam a se sobrepor em JZ4. Os valores de $E_{desc}(cc)$, por outro lado, são mais diversificados, ocorrendo maior separação entre complexos com pose cristalográfica e com poses falso-positivo. Portanto, o descritor de afinidades calibrado nesse trabalho se mostrou mais adequado do que a função de energia do Vina para separar poses cristalográficas (experimentalmente observadas) de poses falso-positivo (fruto de imprecisões na ancoragem).

Figura 4.9: Histogramas de energias livres de ligação para complexos com BNZ (L99A) e JZ4 (M102Q). Complexos contendo pose cristalográfica estão em vermelho e complexos contendo pose falso-positivo estão em verde. Os complexos foram considerados cristalográficos quando o RMSD em relação à pose cristalográfica era inferior a 2,5 Å e falso-positivo quando o RMSD em relação a uma pose falso-positivo caracterizada para aquele ligante (veja seção 4.2.3) era inferior a 2,5 Å.

A comparação entre valores de $E_{Vina}(cc)$ e $E_{desc}(cc)$ atribuídos a complexos contendo poses falso-positivo mostra que $E_{desc}(cc)$ identifica todas as poses falso-positivo, atribuindo a elas valores mais desfavoráveis do que -2 kcal/mol. $E_{Vina}(cc)$, por outro lado, não é capaz de identificá-las.

Portanto, o descritor de afinidades (E_{desc}) foi usado para substituir a função de energia do Vina no restante do trabalho devido a sua capacidade de atribuir afinidades mais próximas das experimentais (veja seção 4.2.5), de distinguir energeticamente poses cristalográficas e poses falso-positivo e de identificar corretamente todas as poses falso-positivo encontradas.

4.3.2 Comparação Entre Estruturas *Apo* e os Conjuntos de Estruturas

O complexo com energia mais favorável atribuída pelo descritor de afinidades calibrado foi escolhido para a análise dessa seção. Complexos foram obtidos por ancoragem a estruturas *apo* (códigos de PDB 2B70 [85] para L99A e 1LGU [33] para M102Q) e ao conjunto de estruturas (os mesmos complexos da seção anterior). As afinidades calculadas foram denominadas $E_{desc}(apo)^+$ e $E_{desc}(cc)^+$, respectivamente.

A tabela 4.9 mostra que $E_{desc}(apo)^+$ apresenta valores menos favoráveis do que ΔG_{exp} , enquanto $E_{desc}(cc)^+$ apresenta valores mais favoráveis do que ΔG_{exp} .

$E_{desc}(cc)^+$ pode estar superestimada porque é obtida de um único complexo, o que não representa uma distribuição de equilíbrio. Métodos para obter afinidades que respeitem uma distribuição de equilíbrio serão discutidos na seção 4.4.1.

$E_{desc}(apo)^+$ apresenta valores menos favoráveis do que ΔG_{exp} devido ao pequeno volume do sítio de ligação nas estruturas *apo*. Essas estruturas não apresentam modificações no esqueleto ou cadeias laterais para acomodar os ligantes, gerando choques estéricos ou interações desfavoráveis e, logo, $E_{desc}(apo)^+$ desfavoráveis. O pequeno volume do sítio das estruturas *apo* é evidenciado pela ausência total de poses no sítio de ligação das estruturas *apo* para os complexos obtidos com I4B, N3B, PEM e MXY.

Portanto, os resultados mostram que o conjunto de estruturas cristalográficas, que leva em

Tabela 4.9: Energias livres de ligação (em kcal/mol) de complexos oriundos de ancoragem com estrutura *apo* ou com conjunto de estruturas cristalográficas.

L99A				M102Q			
ligante	ΔG_{exp}	$E_{desc}(\text{apo})^+$	$E_{desc}(\text{cc})^+$	ligante	ΔG_{exp}	$E_{desc}(\text{apo})^+$	$E_{desc}(\text{cc})^+$
N4B	-6,7	-7,1	-7,7	J1Z	-4,9	-7,1	-7,6
OEM	-4,5	-5,4	-7,1	JZ4	-5,6	-5,4	-6,1
I4B	-6,4	-5,1* ¹	-7,8	CAQ	-4,4	-4,4	-4,9
MEM	-5,1	-5,0	-6,1	MBN	-5,2	-4,3	-5,3
OXE	-4,6	-5,0	-6,2	1AN	-5,5	-4,2	-4,6
PYL	-5,7	-4,9	-5,8	EIP	-4,8	-4,2	-6,0
MBN	-5,5	-4,5	-4,9	JZ0	-4,7	-4,1	-5,0
PXY	-4,6	-4,4	-5,1	CMI	-5,3	-3,8	-4,8
IND	-4,9	-4,0	-4,7	J0Z	-4,7	-3,8	-5,4
N3B	-6,5	-4,0*	-7,2	2EP	-4,3	-3,6	-4,6
BNZ	-5,2	-3,9	-4,3	3MP	-5,2	-3,5	-4,7
PEM	-5,4	-3,4*	-6,0	PAN	-5,8	-2,8	-3,8
MXY	-4,7	-2,4*	-5,3				

¹ Asteriscos indicam afinidades obtidas de complexos em que o ligante estava fora do sítio de ligação.

consideração a flexibilidade da proteína, é mais adequado para representar a proteína do que uma estrutura *apo*, pois somente o conjunto foi capaz de acomodar todos os ligantes no sítio de ligação.

4.3.3 Agrupamento de Complexos

Como o conjunto de estruturas cristalográficas usado nas seções anteriores resulta em um número grande de complexos gerados para cada ligante testado (400 neste caso), buscou-se alternativas para reduzir o número de cálculos e, conseqüentemente, o esforço computacional necessário para estimar afinidades para conjuntos de complexos. Assim, métodos de agrupamento dos complexos foram testados e procurou-se por um procedimento capaz de gerar grupos energeticamente homogêneos, com pequena dispersão interna dos valores de $E_{desc}(\text{cc})$, e que ao mesmo tempo pudessem ter sua dispersão descrita pela diferença de $E_{desc}(\text{cc})$ entre o complexo no centro e o complexo mais distante estruturalmente do centro do grupo. Dessa forma, a média de $E_{desc}(\text{cc})$ foi comparada a $E_{desc}(\text{cc})$ do complexo no centro do grupo.

A seção a seguir descreve os resultados obtidos com dois métodos de agrupamento, o agru-

pamento por RMSD (seção 4.3.3.1) e o agrupamento por contatos (seção 4.3.3.2), e compara esses resultados (seção 4.3.3.3). Nenhum dos métodos mostrou-se totalmente satisfatório.

4.3.3.1 Agrupamento por RMSD

Como o agrupamento gera aproximadamente 200 grupos para cada ligante, foi calculado RMSD_{ene} entre a média de $E_{desc}(cc)$ dentro de cada grupo e $E_{desc}(cc)$ do centro do mesmo grupo para cada ligante (figura 4.10a). Da mesma forma, foi calculado RMSD_{dis} entre a diferença de $E_{desc}(cc)$ média e $E_{desc}(cc)$ mais distante da média no mesmo grupo e a diferença de $E_{desc}(cc)$ do centro e $E_{desc}(cc)$ da pose mais distante estruturalmente do centro no mesmo grupo (figura 4.10b). Somente grupos com mais de 2 membros foram considerados nessa análise. Considerando um limite de 0,5 kcal/mol para valores aceitáveis de RMSD, os RMSDs ficam adequados para a maioria dos ligantes de L99A, mas para M102Q os RMSDs ficam acima do limite adotado para os ligantes 2EP, CAQ, CMI, JZ4, EIP e JZ0 e para os não-ligantes 2AP, PHD e JZ3.

Figura 4.10: Desvios da energia média [(a), RMSD_{ene}] e da dispersão [(b), RMSD_{dis}] para todos os ligantes testados. A linha tracejada vertical separa L99A (à esquerda) de M102Q (à direita). A linha tracejada horizontal marca o limite de 0,5 kcal/mol estabelecido para um RMSD aceitável.

A figura 4.11 mostra a dispersão de $E_{desc}(cc)$ do grupo energeticamente menos homogêneo para cada ligante. A dispersão foi calculada como a diferença entre $E_{desc}(cc)$ média e $E_{desc}(cc)$ mais distante da média no mesmo grupo. A dispersão máxima de 1 kcal/mol foi considerada

aceitável, pois valores muito acima caracterizam a presença de grupos energeticamente heterogêneos. O agrupamento por RMSD gerou grupos homogêneos para a maioria dos ligantes de L99A e ao menos um grupo heterogêneo para a maioria dos ligantes de M102Q. Em grupos energeticamente heterogêneos as estimativas por informações estruturais são piores e, portanto, RMSD_{ene} e RMSD_{dis} são maiores para ligantes de M102Q. Todos os ligantes desse mutante que apresentaram estimativas por informações estruturais insatisfatórias apresentaram também ao menos um grupo energeticamente heterogêneo.

Figura 4.11: Dispersão máxima de $E_{desc}(cc)$ (DM) para todos os ligantes testados. Linhas tracejadas como na figura 4.10.

As dispersões máximas são pequenas para os ligantes apolares de L99A e elevadas para os ligantes polares de M102Q. Nos ligantes polares, por exemplo, pequenas mudanças na posição de uma ligação de H, equivalentes a pequenas mudanças no valor do RMSD estrutural usado para agrupamento, podem gerar grandes diferenças na contribuição eletrostática de poses parecidas e, conseqüentemente, nos valores de $E_{desc}(cc)$ calculados, resultando em grupos energeticamente heterogêneos.

As dispersões elevadas poderiam ser resolvidas por um agrupamento que também considere as interações ou contatos que o ligante realiza com a proteína. Isso levou ao teste do método de agrupamento por contatos.

4.3.3.2 Agrupamento por Contatos

RMSD_{ene} (figura 4.12a) e RMSD_{dis} (figura 4.12b) são adequados para a maioria dos ligantes de L99A. No entanto, as estimativas são insatisfatórias para os ligantes de M102Q 2EP, CAQ, JZ4, EIP, JZ0 e J0Z e para os não-ligantes 2AP, PHD e JZ3. Alguns ligantes são os mesmos mencionados na seção 4.3.3.1, para os quais as estimativas também ficaram insatisfatórias.

Figura 4.12: Desvios da energia média [(a), RMSD_{ene}] e da dispersão [(b), RMSD_{dis}] para todos os ligantes testados.

A figura 4.13 mostra que o agrupamento por contatos gerou somente grupos homogêneos para alguns ligantes de L99A e ao menos um grupo heterogêneo para todos os ligantes de M102Q.

Figura 4.13: Dispersão máxima de $E_{desc}(cc)$ (DM) para todos os ligantes testados.

4.3.3.3 Comparação Entre os Métodos de Agrupamento Testados

Um método de agrupamento mostra-se útil ao gerar poucos grupos, pois assim diminui a dimensão do problema analisado. Os grupos devem ter mais de 2 membros para que ocorra redução de cálculos, pois as estimativas de média e dispersão de $E_{desc}(cc)$ requerem o cálculo de $E_{desc}(cc)$ para 2 membros do grupo. Comparado ao agrupamento por RMSD (seção 4.3.3.1), o agrupamento por contatos (seção 4.3.3.2) gerou menos grupos no total e maior número de grupos com mais de 2 membros. Conclui-se que o agrupamento por contatos é mais eficiente na redução do custo computacional.

Nas figuras 4.14a e 4.14b os valores são negativos para a maioria dos ligantes, o que indica que o agrupamento por contatos gera grupos com maior RMSD energético para as estimativas de média e dispersão de $E_{desc}(cc)$. As diferenças entre RMSDs, no entanto, não são significativas.

Figura 4.14: Diferença entre (a) $RMSD_{ene}$ e (b) $RMSD_{dis}$ dos grupos obtidos do agrupamento por RMSD ($RMSD^{RMSD}$) e do agrupamento por contatos ($RMSD^{contatos}$) para todos os ligantes testados.

Na figura 4.15 os valores em geral são negativos, indicando que o grupo com dispersão máxima obtido pelo agrupamento por contatos tem valor de dispersão superior ao valor apresentado pelo grupo com dispersão máxima obtido pelo agrupamento por RMSD. Isso é um indicativo de que os grupos gerados pelo agrupamento por contatos tendem a ser menos homogêneos energeticamente do que aqueles gerados pelo agrupamento por RMSD. Esse resul-

tado foi inesperado. A expectativa era que o agrupamento por contatos produzisse grupos mais homogêneos, já que ele foi testado justamente por melhorar a descrição de contatos entre ligante e proteína. Uma possível explicação para a dispersão máxima de $E_{desc}(cc)$ ser maior nos grupos obtidos por agrupamento por contatos é a presença de um maior número de membros no grupo que apresenta a dispersão máxima, comparado ao mesmo número no agrupamento por RMSD, o que levaria a presença de $E_{desc}(cc)$ mais diversas dentro do grupo. Na figura 4.16 os valores em geral são positivos, indicando que o grupo com dispersão máxima é mais populoso no agrupamento por RMSD, o que invalida a explicação proposta. Outra possível explicação é o algoritmo usado no agrupamento por contatos, que não foi desenhado para o estudo de complexos receptor–ligante.

Figura 4.15: Diferença entre a dispersão máxima de $E_{desc}(cc)$ obtida pelo agrupamento por RMSD (DM^{RMSD}) e pelo agrupamento por contatos ($DM^{contatos}$).

Figura 4.16: Diferença entre a população do grupo com dispersão máxima de $E_{desc}(cc)$ obtida pelo agrupamento por RMSD (P^{RMSD}) e pelo agrupamento por contatos ($P^{contatos}$).

Apesar do agrupamento por RMSD gerar mais grupos e menos grupos com mais de 2 membros, ele gera grupos mais homogêneos energeticamente. Nenhum dos métodos de agrupamento, no entanto, tem desempenho totalmente satisfatório considerando $RMSD_{ene}$ e $RMSD_{dis}$. Portanto, agrupamento não foi utilizado no restante do trabalho.

4.4 Complexos com Estruturas da Trajetória de Dinâmica Molecular

A seção a seguir descreve os experimentos realizados com complexos obtidos da ancoragem dos ligantes às estruturas da trajetória de dinâmica molecular descritas na seção 4.1.2. Primeiro, são descritas as aproximações usadas para estimar afinidades segundo a teoria do ligante implícito [20] (seção 4.4.1), apresentada na seção 1.5.1. Depois, uma das aproximações é testada e empregada na escolha de alguns ligantes (seção 4.4.2), que são então usados para comparar as diferentes aproximações aos resultados experimentais (seção 4.4.3).

4.4.1 Aproximações para a Teoria do Ligante Implícito

A afinidade atribuída aos complexos obtidos por ancoragem às estruturas da trajetória pelo descritor de energias calibrado foi denominada $E_{desc}(e)$. Para cada uma das 50 estruturas de receptor foram geradas 20 poses de complexos usando Vina, totalizando 1000 poses para cada ligante. Por exemplo, a diversidade de poses encontradas dentro e fora do sítio de ligação para JZ4 é mostrada na figura 4.17. Assim como na figura 4.8, as poses com $E_{desc}(e)$ mais favoráveis encontram-se na região do sítio de ligação cristalográfico, enquanto as poses fora dessa região apresentam valores menos favoráveis. Dessa forma, o descritor de afinidades é capaz de reconhecer o sítio de ligação, conforme já demonstrado na seção 4.3.1. Resultados semelhantes foram obtidos para os demais ligantes.

As diferentes aproximações para a teoria do ligante implícito usadas para estimar afinidades estão resumidas na figura 4.18. Diferente das seções anteriores, aqui todos os complexos obtidos são considerados nas estimativas de afinidade.

Figura 4.17: Superposição das poses do ligante JZ4 (M102Q) obtidas por ancoragem a um conjunto de estruturas da trajetória. Somente uma das estruturas de receptor do conjunto está representada. Poses estão coloridas de acordo com $E_{desc}(e)$, com a escala de mais para menos favorável: vermelho, laranja, amarelo, verde, ciano e azul.

Figura 4.18: Sequência de aproximações utilizadas para obter energias livres de ligação usando a teoria do ligante implícito. O significado de cada sigla está descrito no texto.

Para calcular ψ (seção 1.5.1), 4 aproximações foram empregadas:

- E_{desc} (seção 4.2.4)
- E_{Vina} (seção 4.2.5)
- E_{Su} : ψ é aproximado pelo descritor de afinidades calibrado por Su *et al.* [24] (equação 3.1 com parâmetros descritos no item 2 da seção 3.3);
- E_{tot} : ψ é aproximado pela soma das energias de interação totais, sem parametrização (equação 3.1 com parâmetros descritos no item 3 da seção 3.3).

O valor de B (seção 1.5.1) foi calculado por 2 aproximações:

$$B^M \approx -k_B T \ln \frac{1}{P} \sum_{i=1}^P e^{-\psi_i/k_B T} \quad (4.1)$$

$$B^D \approx \psi_{min} \quad (4.2)$$

Na equação 4.1, P representa o número de configurações de ligante consideradas para um dado receptor e tem valor máximo de 20. A aproximação feita é a atribuição de pesos iguais para as diferentes configurações do ligante. No cálculo de B^M foram usadas configurações do ligante com valor de ψ até 2 kcal/mol mais desfavoráveis do que o valor de ψ mais favorável para o receptor considerado. Tal corte foi realizado para remover poses pouco relevantes. Essas poses podem ser geradas pelo Vina porque sua função de pontuação é aproximada e as poses são obtidas por otimização e não por uma amostragem por importância [3]. O corte evita que B^M torne-se demasiadamente desfavorável e, possivelmente, deixa o conjunto de poses obtido mais próximo de uma distribuição de equilíbrio.

Na equação 4.2, ψ_{min} representa o valor mais favorável de ψ obtido para uma dada configuração de receptor. A aproximação considera que há uma configuração de ligante dominante para cada configuração de receptor.

O valor de ΔG_{lig} (seção 1.5.1) foi calculado por 2 aproximações:

$$\Delta G^M \approx -k_B T \ln \frac{1}{N} \sum_{n=1}^N e^{-B_n/k_B T} + \Delta G_\xi \quad (4.3)$$

$$\Delta G^D \approx B_{min} + \Delta G_\xi \quad (4.4)$$

Na equação 4.3, $N = 50$ representa o número de configurações do receptor consideradas e B é dado pelas equações 4.1 ou 4.2. ΔG_ξ (equação 1.20) representa uma correção à restrição de amostragem do ligante ao *grid* usado na ancoragem. Considerando $\Omega = 27.000 \text{ \AA}^3$ e $V^o = 1.668 \text{ \AA}^3$ [40], obteve-se $\Delta G_\xi = 0,9 \text{ kcal/mol}$. A aproximação feita é a atribuição de pesos iguais para as diferentes configurações de receptor. Diferente das configurações de ligante, as configurações de receptor foram obtidas por uma amostragem de simulação, estando, portanto, próximas de uma distribuição de equilíbrio.

Na equação 4.4, B_{min} representa o valor mais favorável de B obtido para um determinado complexo. A aproximação aqui é considerar que há uma configuração de receptor dominante dentro das 50 avaliadas.

As aproximações usadas para calcular B e ΔG_{lig} são limites opostos quanto ao peso das configurações. De um lado, em B^M e ΔG^M supõe-se que as configurações têm o mesmo peso. Do outro lado, em B^D e ΔG^D supõe-se que apenas uma das configurações tem peso 1, e as demais têm peso 0. A distribuição de configurações ideal, no entanto, deve ser intermediária a esses extremos.

4.4.2 Análise Inicial pela Teoria do Ligante Implícito

Inicialmente 20 ligantes de cada mutante foram usados para calcular afinidades segundo o método ΔG_{desc}^M (figura 4.18). Este foi escolhido por ser mais rigoroso e porque usa um descritor calibrado para estimar as afinidades dos complexos de ligantes com os mutantes utilizados nesse trabalho. O número de ligantes testado é elevado para permitir que os comportamentos ou tendências gerais se destaquem, em contraposição a variações como a qualidade da parametrização, etc.

Na tabela 4.10 é possível observar os seguintes comportamentos gerais, que serão discutidos a seguir: ΔG_{desc}^M é menos favorável do que $E_{desc}(c)$ e do que ΔG_{desc}^M -s e mais favorável do

que ΔG_{desc}^M -fs; ΔG_{desc}^M -s é mais favorável do que ΔG_{desc}^M -fs; e $\Delta E_{desc}(e)$ -s é menor do que $\Delta E_{desc}(e)$ -fs.

Tabela 4.10: Energias livres de ligação (em kcal/mol) estimadas pela teoria do ligante implícito.

ligante	ΔG_{exp}	$E_{desc}(c)^1$	ΔG_{desc}^M	ΔG_{desc}^M -s ²	ΔG_{desc}^M -fs ³	$\Delta E_{desc}(e)$ -s ⁴	$\Delta E_{desc}(e)$ -fs ⁵
L99A							
BNZ	-5,2	-3,8	-2,4	-2,7	-1,4	$0,3 \pm 0,3$	$2,8 \pm 1,0$
I4B	-6,5	-7,7	-5,1	-5,5	-3,8	$0,9 \pm 1,0$	$3,8 \pm 1,0$
IND	-4,9	-4,0	-2,8	-2,7	-2,9	$0,7 \pm 0,5$	$4,7 \pm 1,3$
MBN	-5,5	-3,8	-2,9	-3,2	-2,1	$0,6 \pm 0,4$	$3,3 \pm 1,0$
MEM	-5,0	-5,5	-3,7	-4,0	-2,7	$0,7 \pm 0,6$	$3,5 \pm 1,1$
MXV	-4,7	-4,6	-3,0	-3,3	-2,1	$0,7 \pm 0,6$	$3,2 \pm 1,1$
N3B	-6,5	-6,3	-4,3	-4,6	-3,0	$0,7 \pm 0,6$	$3,7 \pm 0,9$
N4B	-6,7	-7,5	-4,8	-5,2	-3,7	$0,8 \pm 0,8$	$4,2 \pm 1,0$
OEM	-4,5	-5,5	-4,4	-4,7	-3,6	$0,8 \pm 0,5$	$3,6 \pm 1,2$
OXE	-4,6	-5,7	-3,8	-4,1	-3,0	$0,6 \pm 0,4$	$3,1 \pm 1,0$
PEM	-5,4	-5,4	-3,6	-3,9	-2,9	$0,5 \pm 0,4$	$3,9 \pm 1,0$
PXY	-4,6	-4,9	-3,0	-3,4	-2,2	$0,5 \pm 0,4$	$3,5 \pm 1,3$
PYL	-5,7	-5,6	-3,4	-3,7	-2,7	$0,5 \pm 0,5$	$3,5 \pm 0,9$
1AN	> -2,0	-3,7	-2,0	-2,2	-1,6	$1,0 \pm 0,7$	$3,7 \pm 1,0$
3MP	> -2,0	-3,3	-2,8	-2,4	-2,8	$0,7 \pm 0,5$	$4,4 \pm 1,1$
ANL	> -2,0	-3,4	-1,8	-2,0	-1,4	$0,6 \pm 0,4$	$3,9 \pm 1,1$
CHX	> -2,0	-5,8	-5,1	-5,3	-4,0	$0,7 \pm 0,5$	$1,6 \pm 0,6$
IPH	> -2,0	-2,8	-1,4	-1,7	-1,1	$0,8 \pm 0,5$	$4,2 \pm 1,0$
TBB	> -2,0	-6,5	-5,4	-5,8	-4,7	$0,7 \pm 0,6$	$3,6 \pm 1,5$
TMB	> -2,0	-4,7	-3,0	-3,5	-2,3	$0,7 \pm 0,6$	$3,0 \pm 1,4$
M102Q							
1AN	-5,5	-4,6	-2,2	-2,5	-1,7	$1,3 \pm 0,5$	$3,9 \pm 1,0$
2EP	-4,3	-4,0	-2,8	-3,4	-2,0	$1,4 \pm 1,2$	$5,3 \pm 1,0$
PAN	-5,9	-3,3	-1,8	-2,2	-0,6	$0,7 \pm 1,1$	$6,1 \pm 1,3$
3MP	-5,2	-4,2	-2,5	-2,6	-2,5	$0,8 \pm 0,4$	$3,8 \pm 1,0$
CAQ	-4,4	-4,5	-2,6	-2,7	-2,5	$1,7 \pm 0,8$	$4,7 \pm 1,2$
CMI	-5,3	-3,8	-2,6	-2,5	-2,6	$1,2 \pm 0,7$	$5,0 \pm 1,2$
EIP	-4,8	-5,2	-3,0	-3,3	-2,1	$1,0 \pm 0,6$	$4,0 \pm 1,0$
J0Z	-4,8	-4,8	-2,2	-2,6	-1,3	$0,3 \pm 0,6$	$5,7 \pm 1,3$
J1Z	-5,0	-7,5	-5,2	-5,8	-4,1	$0,3 \pm 0,3$	$2,8 \pm 0,9$
JZ0	-4,7	-5,2	-2,7	-3,0	-2,2	$1,1 \pm 0,7$	$3,7 \pm 0,9$
JZ4	-5,7	-5,5	-3,4	-3,7	-2,9	$0,6 \pm 0,6$	$4,4 \pm 1,1$

continua

¹ $E_{desc}(c)$ como calculada na seção 4.2.4.

² ΔG_{desc}^M -s: ΔG_{desc}^M calculada considerando complexos em que o ligante estava no sítio de ligação.

³ ΔG_{desc}^M -fs: ΔG_{desc}^M calculada considerando complexos em que o ligante estava fora do sítio de ligação.

⁴ $\Delta E_{desc}(e)$ -s: média $\pm$ desvio padrão da diferença entre $E_{desc}(e)$ máxima e mínima de cada receptor para complexos em que o ligante estava no sítio de ligação.

⁵ $\Delta E_{desc}(e)$ -fs: média $\pm$ desvio padrão da diferença entre $E_{desc}(e)$ máxima e mínima de cada receptor para complexos em que o ligante estava fora do sítio de ligação.

continuação							
ligante	ΔG_{exp}	$E_{desc}(c)$	ΔG_{desc}^M	ΔG_{desc}^M-s	ΔG_{desc}^M-fs	$\Delta E_{desc}(e)-s$	$\Delta E_{desc}(e)-fs$
MBN	-5,2	-4,9	-2,9	-3,2	-2,0	$0,7 \pm 0,4$	$3,1 \pm 1,1$
2AP	$>-2,0$	-4,2	-1,9	-1,9	-1,8	$1,6 \pm 0,9$	$5,1 \pm 1,3$
4CP	$>-2,0$	-4,7	-2,3	-1,5	-2,4	$1,1 \pm 0,7$	$5,5 \pm 1,5$
4VP	$>-2,0$	-3,2	-2,1	-2,3	-1,8	$0,7 \pm 0,6$	$4,3 \pm 0,9$
ETP	$>-2,0$	-7,2	-4,7	-5,2	-3,4	$1,0 \pm 0,6$	$3,7 \pm 0,9$
JZ3	$>-2,0$	-5,2	-2,6	-2,8	-2,1	$1,4 \pm 0,9$	$4,8 \pm 1,1$
NBE	$>-2,0$	-3,0	-2,2	-2,1	-2,3	$1,2 \pm 0,6$	$4,8 \pm 1,0$
NCF	$>-2,0$	-4,4	0,0	-1,3	-0,2	$0,8 \pm 0,8$	$5,5 \pm 1,4$
PHD	$>-2,0$	-3,7	-1,9	-2,2	-1,5	$0,9 \pm 0,7$	$5,0 \pm 1,0$

$\Delta E_{desc}(e)-s$ é menor que $\Delta E_{desc}(e)-fs$ porque o espaço de configurações que o ligante explora é maior fora do sítio de ligação, o que leva a uma maior diversidade de $E_{desc}(e)$.

ΔG_{desc}^M é em geral menos favorável do que $E_{desc}(c)$ porque $E_{desc}(c)$ é calculada com a estrutura cristalográfica nativa do complexo, que pode ter afinidade elevada mas ser pouco relevante para a distribuição estrutural de equilíbrio do receptor. ΔG_{desc}^M , por outro lado, é calculada considerando configurações de ligante e receptor e, portanto, algum tratamento estatístico mais próximo da distribuição de equilíbrio. Entre as configurações de receptor usadas no cálculo de ΔG_{desc}^M , estavam presentes configurações em que o sítio de ligação estava obstruído, conforme descrito na seção 4.1.3.3. Nesses casos, todas as poses de ligante se encontravam fora do sítio, recebendo, portanto, $E_{desc}(e)$ desfavoráveis, o que contribuiu para B^M desfavoráveis e, consequentemente, ΔG_{desc}^M menos favoráveis que $E_{desc}(c)$.

A diminuição de ΔG_{desc}^M resultante da inclusão de configurações do ligante sem uma explícita amostragem por importância é parcialmente contornada impondo um *cutoff* nos valores de $E_{desc}(e)$ considerados para calcular B , conforme mencionado acima. A validade ou a necessidade deste corte arbitrário poderia ser testada usando métodos de amostragem, como dinâmicas moleculares em que o receptor é mantido rígido e o ligante pode se mover.

O descritor de afinidades, embora calibrado com poses falso-positivo, atribui $E_{desc}(e)$ para algumas poses fora do sítio superiores aos $E_{desc}(e)$ de poses no sítio. Isso é evidenciado ao se comparar ΔG_{desc}^M-s com ΔG_{desc}^M-fs . Se o descritor de afinidades tivesse comportamento adequado, ΔG_{desc}^M-s seria maior que ΔG_{desc}^M-fs em todos os casos, mas isso não ocorre para os li-

gantes IND (L99A) e CMI (M102Q) e para os não–ligantes 3MP (L99A), 4CP e NBE (M102Q). CMI e 4CP são os únicos ligantes da tabela 4.10 que apresentam Cl em sua composição, o que indica que parte do problema pode ser a parametrização de Cl no campo de força usado.

A diferença entre ΔG_{desc}^M e ΔG_{desc}^M –s é pequena, sendo de no máximo 0,5 kcal/mol para L99A e 1,3 kcal/mol para M102Q, mostrando que, de forma geral, o descritor de afinidades e o *cutoff* imposto aos valores de $E_{desc}(e)$ funcionam em conjunto razoavelmente bem.

O não–ligante de M102Q NCF foi reconhecido como tal somente por ΔG_{desc}^M e não por $E_{desc}(c)$ devido a amostragem de configurações. Embora seja possível ancorar NCF no sítio de ligação da estrutura cristalográfica e existam estruturas de receptor ao longo da trajetória em que NCF cabe no sítio, essas estruturas são pouco frequentes (3 em 50 no conjunto de estruturas da trajetória). Desse modo, NCF está fora do sítio de ligação na maioria dos complexos, levando a $E_{desc}(e)$ desfavoráveis, B^M desfavoráveis para a maioria das configurações de receptor e, conseqüentemente, ΔG^M desfavorável.

O descritor atribui afinidades elevadas para algumas poses fora do sítio. Essas poses podem se tratar de poses falso–positivo (definidas na seção 4.2.3) não identificadas ou de poses metaestáveis, conforme já foi descrito por exemplo para o complexo entre β –tripsina e benza–midina [86, 87]. A presença de moléculas orgânicas ou adjuvantes que auxiliam na cristalização de proteínas complexadas nos potenciais sítios metaestáveis (figura 4.19) sugere que estes possam ser possíveis sítios alternativos de complexação. Portanto, poses com afinidade elevada fora do sítio de ligação não podem ser excluídas, pois podem ser relevantes para a distribuição de poses do complexo em solução no equilíbrio.

4.4.3 Aproximações Dentro da Teoria do Ligante Implícito

Foram escolhidos 4 ligantes e 4 não–ligantes de cada mutante cujas afinidades fossem razoavelmente bem descritas por $E_{desc}(c)$ e ΔG_{desc}^M para testar outras aproximações no cálculo de ψ , B e ΔG_{lig} (figura 4.18). As aproximações foram comparadas entre si e com valores de $E(c)$ calculados para complexos oriundos de estruturas cristalográficas.

Figura 4.19: Estrutura de complexos com lisozima. (a) Pose do ligante JZ4 (laranja) fora do sítio de ligação e com afinidade relevante. (b) Pose da molécula auxiliadora da cristalização 2-hidroxi-etil dissulfídeo (vermelho) na estrutura do PDB 3DKE.

A tabela 4.11 mostra que as afinidades obtidas usando E_{Su} e E_{tot} são muito favoráveis comparadas a ΔG_{exp} , mostrando que esses descritores de afinidades são inadequados para a proteína estudada. Esse comportamento era esperado para E_{tot} , pois ele não representa um descritor feito para reproduzir afinidades. E_{tot} é uma soma de energias potenciais e, por isso, desconsidera contribuições importantes para estimar afinidades, como a entropia.

E_{Su} poderia ser melhorado para a proteína utilizada nesse estudo por modificação do parâmetro τ (seção 1.5.2.1), que constitui um fator de correção dependente da proteína [88]. Além disso, E_{Su} poderia ser melhorado para os mutantes estudados também pela separação dos termos G_{NP} e G_{cav} , que são considerados conjuntamente como G_{NP} na equação 3.1, mas separadamente na parametrização obtida por Su *et al.* (equação 1.22) [24]. Por fim, Su *et al.* [24] considera em sua parametrização que V_{elet}^c e G_{GB} , assim como V_{vdW}^c e G_{NP} , apresentam a mesma resposta linear (seção 1.5.2.1), o que pode estar incorreto.

Tabela 4.11: Energias livres de ligação (em kcal/mol) estimadas por diferentes aproximações para a teoria do ligante implícito.

ligante	ΔG_{exp}	$E_{desc}(c)^1$	$E_{Su}(c)$	$E_{tot}(c)$	ΔG_{desc}^M ²	ΔG_{desc}^{DM}	ΔG_{desc}^{DD}	ΔG_{tot}^M	ΔG_{Su}^M	ΔG_{Vina}^M
L99A										
I4B	-6,4	-7,7	-14,8	-33,9	-5,1	-5,5	-6,4	-30,2	-11,4	-4,0
OXE	-4,6	-5,7	-11,7	-28,3	-3,8	-4,2	-4,9	-25,5	-9,3	-3,9
PEM	-5,4	-5,4	-12,3	-29,4	-3,6	-3,9	-4,7	-27,3	-10,2	-3,7
PXY	-4,6	-4,9	-11,6	-27,7	-3,0	-3,4	-4,1	-25,2	-9,2	-3,5
1AN	> -2,0	-3,7	-10,1	-26,5	-2,0	-2,6	-3,4	-31,9	-8,2	-3,7
3MP	> -2,0	-3,2	-8,5	-21,8	-2,8	-3,2	-4,3	-29,0	-7,5	-2,3
ANL	> -2,0	-3,4	-9,5	-24,7	-1,8	-2,3	-3,0	-32,7	-8,0	-3,4
IPH	> -2,0	-2,8	-9,7	-26,5	-1,4	-2,0	-2,8	-32,0	-8,1	-3,4
M102Q										
1AN	-5,5	-4,6	-11,0	-30,9	-2,2	-2,8	-3,7	-34,0	-8,4	-3,8
CAQ	-4,4	-4,5	-12,0	-31,4	-2,6	-3,1	-4,2	-41,8	-9,4	-3,4
JZ0	-4,7	-5,2	-12,6	-32,4	-2,7	-3,3	-4,3	-31,0	-9,1	-3,7
JZ4	-5,6	-5,5	-15,1	-36,1	-3,4	-3,8	-4,7	-38,0	-11,4	-3,5
4VP	> -2,0	-3,2	-11,1	-25,4	-2,1	-2,5	-3,5	-29,5	-9,2	-2,8
NBE	> -2,0	-3,0	-10,5	-24,9	-2,2	-2,7	-4,1	-30,1	-8,9	-3,3
NCF	> -2,0	-4,4	-16,4	-41,8	0,0	-0,5	-2,1	-40,3	-11,5	-3,4
PHD	> -2,0	-3,7	-11,5	-31,5	-1,9	-2,4	-3,0	-37,8	-9,1	-3,5

¹ $E_{desc}(c)$ como calculada na seção 4.2.4.

² ΔG_{desc}^M como calculada na seção 4.4.2.

A comparação entre $E(c)$ e ΔG^M para um mesmo descritor de afinidades mostra que as energias calculadas por ΔG^M são em geral menos favoráveis, conforme discutido na seção 4.4.2. A relação entre $E(c)$ e ΔG^M vista para o descritor de afinidades calibrado nesse trabalho, portanto, também é vista para outros descritores.

Para ΔG_{tot}^M , as afinidades calculadas são menos favoráveis do que $E_{tot}(c)$ somente para ligantes apolares (I4B, OXE, PEM e PXY). As interações eletrostáticas entre o ligante e o receptor são mais relevantes para complexos com ligantes polares e têm maior contribuição no cálculo de E_{tot} do que no cálculo de E_{desc} ou E_{Su} (os parâmetros que multiplicam a contribuição eletrostática nesses descritores são 1, 0,09 e 0,25 respectivamente). Como $E_{tot}(c)$ é calculado para complexos no sítio de ligação, $E_{tot}(e)$ e, logo, ΔG_{tot}^M são calculados para complexos dentro e fora do sítio, as interações que tornam $E_{tot}(e)$ mais favorável devem ocorrer fora do sítio. Assim, o uso de um descritor calibrado melhora a distinção entre configurações de ligante muito e pouco relevantes para a distribuição conformacional no caso de ligantes polares.

Para verificar a utilidade das diferentes aproximações na separação de ligantes e não-ligantes, as afinidades calculadas foram utilizadas em experimentos de ordenamento (explicados na seção 4.2.5). A tabela 4.12 mostra que nas aproximações ΔG_{tot}^M , ΔG_{Vina}^M e ΔG_{desc}^{DD} não houve separação adequada entre ligantes e não-ligantes para L99A. Já para M102Q, a tabela 4.13 mostra que as aproximações $E_{Su}(c)$, $E_{tot}(c)$, ΔG_{desc}^{DD} , ΔG_{tot}^M , ΔG_{Su}^M e ΔG_{Vina}^M não identificaram adequadamente os não-ligantes. Em muitos casos, o não-ligante NCF não foi reconhecido. Somente 3 métodos reconheceram todos os não-ligantes nos ordenamentos: $E_{desc}(c)$, ΔG_{desc}^M e ΔG_{desc}^{DM} .

Tabela 4.12: Ligantes e não–ligantes (em negrito) de L99A ordenados de forma decrescente a partir de energias livres de ligação estimadas.

ΔG_{exp}	$E_{desc}(\mathbf{c})$	$E_{Su}(\mathbf{c})$	$E_{tot}(\mathbf{c})$	ΔG_{desc}^M	ΔG_{desc}^{DM}	ΔG_{desc}^{DD}	ΔG_{tot}^M	ΔG_{Su}^M	ΔG_{Vina}^M
I4B	I4B	I4B	I4B	I4B	I4B	I4B	ANL	I4B	I4B
PEM	OXE	PEM	PEM	OXE	OXE	OXE	IPH	PEM	OXE
PXY	PEM	OXE	OXE	PEM	PEM	PEM	1AN	OXE	1AN
OXE	PXY	PXY	PXY	PXY	PXY	3MP	I4B	PXY	PEM
IPH ¹	1AN	1AN	IPH	3MP	3MP	PXY	3MP	1AN	PXY
ANL	3MP	IPH	1AN	1AN	1AN	1AN	PEM	IPH	IPH
3MP	ANL	ANL	ANL	ANL	ANL	ANL	OXE	ANL	ANL
1AN	IPH	3MP	3MP	IPH	IPH	IPH	PXY	3MP	3MP

¹ A linha pontilhada separa ligantes (acima) de não–ligantes (abaixo) denominados de acordo com o experimento de ordenamento.

Tabela 4.13: Ligantes e não–ligantes (em negrito) de M102Q ordenados de forma decrescente a partir de energias livres de ligação estimadas.

ΔG_{exp}	$E_{desc}(\mathbf{c})$	$E_{Su}(\mathbf{c})$	$E_{tot}(\mathbf{c})$	ΔG_{desc}^M	ΔG_{desc}^{DM}	ΔG_{desc}^{DD}	ΔG_{tot}^M	ΔG_{Su}^M	ΔG_{Vina}^M
JZ4	JZ4	NCF	NCF	JZ4	JZ4	JZ4	CAQ	NCF	1AN
1AN	JZ0	JZ4	JZ4	JZ0	JZ0	JZ0	NCF	JZ4	JZ0
JZ0	1AN	JZ0	JZ0	CAQ	CAQ	CAQ	JZ4	CAQ	PHD
CAQ	CAQ	CAQ	PHD	1AN	1AN	NBE	PHD	4VP	JZ4
PHD	NCF	PHD	CAQ	NBE	NBE	1AN	1AN	PHD	CAQ
4VP	PHD	4VP	1AN	4VP	4VP	4VP	JZ0	JZ0	NCF
NCF	4VP	1AN	4VP	PHD	PHD	PHD	NBE	NBE	NBE
NBE	NBE	NBE	NBE	NCF	NCF	NCF	4VP	1AN	4VP

A tabela 4.14 mostra que os desvios médios entre afinidades experimentais e estimadas foram maiores para os métodos dependentes de E_{Su} e E_{tot} e menores para métodos que usam a função de energia do Vina ou o descritor calibrado nesse trabalho.

Tabela 4.14: Desvios médios (em kcal/mol) entre energias livres de ligação experimentais e estimadas.

método	desvio médio
$E_{desc}(c)$	1,2
ΔG_{desc}^{DM}	1,2
ΔG_{desc}^{DD}	1,2
ΔG_{desc}^M	1,4
ΔG_{Vina}^M	1,4
ΔG_{Su}^M	5,9
$E_{Su}(c)$	8,4
$E_{tot}(c)$	26,4
ΔG_{tot}^M	29,2

ΔG_{desc}^{DM} teve um melhor desempenho, pois foi capaz de reconhecer todos os não-ligantes e apresentou o menor desvio médio. A mesma observação se aplica aos valores calculados para $E(c)$ pelos diferentes descritores de afinidades. Esse resultado era esperado, pois ΔG_{desc}^{DM} e $E_{desc}(c)$ foram calculados com um descritor calibrado especificamente para estimar afinidades de complexos entre L99A ou M102Q e seus ligantes.

Considerando as aproximações para calcular ΔG_{lig} que usam o descritor de afinidades calibrado nesse trabalho, a aproximação que melhor reproduziu os dados experimentais foi ΔG_{desc}^{DM} , sugerindo que a aproximação de 1 configuração de ligante dominante para cada configuração de receptor é adequada para a proteína estudada, ou que Vina obtém apenas 1 configuração de ligante relevante para cada configuração de proteína. ΔG_{desc}^M também reproduziu razoavelmente bem os dados experimentais, pois também identificou todos os não-ligantes e apresentou desvio médio somente 0,2 kcal/mol maior que o menor desvio médio obtido. A aproximação ΔG_{desc}^{DD} , apesar de apresentar desvio médio pequeno, teve problemas na identificação de não-ligantes, sugerindo que a aproximação de 1 complexo dominante é inadequada para a proteína estudada.

As estimativas realizadas por $E_{desc}(c)$, embora também reproduzam dados experimentais,

têm aplicação limitada, pois necessitam de conhecimentos prévios, como dados estruturais de complexos e a localização do sítio de ligação. As estimativas realizadas usando $E_{desc}(e)$, por outro lado, podem se tornar menos dependentes de parametrização pelo uso, por exemplo, de LIE adaptativo [46], cujos parâmetros (α , β , γ e τ) podem ser estimados a partir de descritores estruturais do ligante e do receptor.

Em conclusão, o método ΔG_{desc}^{DM} foi o que melhor reproduziu os dados experimentais, pois reconheceu todos os não-ligantes e apresentou o menor desvio médio entre afinidades experimentais e calculadas. Outras aproximações, como ΔG_{desc}^M e $E_{desc}(c)$, também se mostraram razoáveis na reprodução de dados experimentais.

5 *Conclusões*

Considerando os objetivos do trabalho (seção 2), os primeiros passos foram montar conjuntos de estruturas para representar a flexibilidade proteica e calibrar uma equação para descrever as afinidades de complexos com a proteína modelo, lisozima.

Duas formas de representar a flexibilidade proteica previamente usadas na literatura [7–10] foram adotadas: conjuntos de estruturas cristalográficas e de estruturas obtidas da trajetória de dinâmica molecular. Dados experimentais de RMN sugerem que os conjuntos montados representam razoavelmente bem a distribuição conformacional do receptor (seções 4.1.2 e 4.1.3).

Para a calibração do descritor de afinidades, foi necessário construir os parâmetros dos ligantes para OPLS-AA. Os parâmetros covalentes dos ligantes, quando não disponíveis, foram aproximados de funções químicas semelhantes. Para ligantes que não tinham cargas bem descritas pelo OPLS-AA, as cargas eram obtidas usando o método AM1, que reproduz as cargas do OPLS-AA, ou o método HF, que reproduziu os momentos de dipolo experimentais disponíveis.

Para a calibração do descritor de afinidades, somente contribuições energéticas oriundas de estruturas cristalográficas ou de poses falso-positivo foram utilizadas. O uso de poses falso-positivo constitui uma inovação para a calibração de um descritor de energias. O uso de estruturas cristalográficas contorna a possibilidade de erro devido ao complexo obtido por ancoragem, que pode não ser o mais relevante para a distribuição de poses ligantes. Dessa forma, considerando a parametrização dos ligantes e o emprego de estruturas cristalográficas, é possível assumir que os desvios médios obtidos para as afinidades de ligantes do conjunto de treino têm como principal fonte a função de energia aproximada.

A comparação entre o descritor de afinidades calibrado e a função de energia do Vina foi feita usando complexos oriundos de ancoragem nativa e de ancoragem cruzada. Nesse caso, há duas fontes de erro, o descritor de energias usado e a geometria do complexo. Os resultados obtidos indicam que o descritor de afinidades reproduz os dados experimentais melhor do que a função de energia do Vina.

Métodos de agrupamento por RMSD e por contatos foram testados com o objetivo de reduzir o custo computacional para estimar afinidades. Tais métodos, no entanto, não foram usados no restante do trabalho, pois não mostraram resultados satisfatórios.

Então, procurou-se responder a pergunta central: qual a influência da flexibilidade proteica nas poses obtidas e nas afinidades estimadas?

A flexibilidade proteica interfere nas poses obtidas por ancoragem de diversas formas. Na comparação entre complexos obtidos por ancoragem com estruturas *apo* ou com conjuntos de estruturas cristalográficas, os conjuntos se mostraram melhores representantes da estrutura proteica, pois somente eles foram capazes de acomodar todos os ligantes no sítio de ligação. Esse caso exemplifica a importância de incorporar a flexibilidade proteica na ancoragem (seção 4.3.2).

Alguns complexos apresentaram afinidade elevada para sítios de ligação diferentes do sítio cristalográfico. Um desses sítios está localizado entre os dois domínios que constituem a proteína. Dados experimentais sugerem que este sítio, visto nas estruturas cristalográficas, é falso (seção 4.1.3.2).

Nos conjuntos de estruturas da trajetória, o sítio de ligação foi transitoriamente obstruído pelas cadeias laterais de ILE78, LEU84 e TYR88. Dados experimentais sugerem que a obstrução por ILE78 pode ser observada experimentalmente, enquanto a obstrução por TYR88 pode ser fruto de imprecisões do campo de força usado. Complexos obtidos com conjuntos de estruturas da trajetória sugerem a localização de potenciais sítios metaestáveis. Mais estudos, porém, são necessários para caracterizar esses sítios nos mutantes de lisozima (seção 4.4.2).

A flexibilidade proteica interfere também nas afinidades estimadas. A determinação da

afinidade utilizando a aproximação de pose dominante também para o conjunto de estruturas cristalográficas pode levar a superestimativas de afinidade.

A estimativa de afinidades a partir de complexos obtidos com conjuntos de estruturas requer um tratamento estatístico adequado, o que levou ao uso de aproximações para a teoria do ligante implícito. A associação de um descritor de afinidades como o LIE com a teoria do ligante implícito constitui uma inovação. No caso dos mutantes de lisozima, as aproximações que melhor reproduziram resultados experimentais foram aquelas em que as configurações de receptor obtidas tinham pesos iguais e as configurações de ligante obtidas tinham pesos iguais ou eram representadas somente pela configuração com afinidade mais favorável. Mais estudos são necessários para verificar se há apenas uma configuração dominante de ligante para cada configuração de lisozima. Por outro lado, imprecisões da ancoragem podem levar esse método a apontar somente uma das configurações de ligante relevante para cada configuração de receptor.

Nas aproximações que consideram configurações de receptor com pesos iguais, as estruturas devem ser obtidas por um método que se aproxime de uma amostragem por importância, como foi feito nesse trabalho, para que essas aproximações sejam razoáveis e levem a boas estimativas de afinidade.

O método para estimar afinidades proposto tem algumas limitações. Ele depende da calibração de um descritor de afinidades e, conseqüentemente, de dados estruturais e de afinidades experimentais disponíveis. Além disso, as configurações de ligante foram obtidas por ancoragem, o que não constitui uma amostragem por importância. O uso de um método de amostragem adequado pode verificar se os complexos obtidos são relevantes e se o uso de um *cutoff* para remover parte dos complexos é adequado (seção 4.4.1).

A combinação de um descritor de afinidades com a teoria do ligante implícito constitui um método rápido para estimar afinidades. Entretanto, sua utilidade será reconhecida somente depois de testes em outras proteínas.

Bibliografia

- 1 LAHANA, R. How many leads from HTS? *Drug Discov. Today*, v. 4, p. 447–448, 1999.
- 2 KUNTZ, I. D. et al. A geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.*, v. 161, p. 269–288, 1982.
- 3 TROTT, O.; OLSON, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, v. 31, p. 455–461, 2010.
- 4 JIANG, F.; KIM, S.-H. “Soft docking”: matching of molecular surface cubes. *J. Mol. Biol.*, v. 219, p. 79–102, 1991.
- 5 LEACH, A. R. Ligand docking to proteins with discrete side–chain flexibility. *J. Mol. Biol.*, v. 235, p. 345–356, 1994.
- 6 SANDAK, B.; WOLFSON, H. J.; NUSSINOV, R. Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers. *Proteins*, v. 32, p. 159–174, 1998.
- 7 BROUGHTON, H. B. A method for including protein flexibility in protein–ligand docking: improving tools for database mining and virtual screening. *J. Mol. Graphics Model.*, v. 18, p. 247–257, 2000.
- 8 CARLSON, H. A. et al. Developing a dynamic pharmacophore model for HIV–1 integrase. *J. Med. Chem.*, v. 43, p. 2100–2114, 2000.
- 9 ARANTES, G. M. Flexibility and inhibitor binding in Cdc25 phosphatases. *Proteins*, v. 78, p. 3017–3032, 2010.
- 10 KNEGTEL, R. M. A.; KUNTZ, I. D.; OSHIRO, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.*, v. 266, p. 424–440, 1997.
- 11 FIELD, M. J. *A practical introduction to the simulation of molecular systems*. 1st. ed. Cambridge: Cambridge University Press, 1999.
- 12 BEST, R. B. Atomistic molecular simulations of protein folding. *Curr. Opin. Struct. Biol.*, v. 22, p. 52–61, 2012.
- 13 STONE, A. J. Intermolecular potentials. *Science*, v. 321, p. 787–789, 2008.
- 14 GILSON, M. K. et al. The statistical–thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.*, v. 72, p. 1047–1069, 1997.

- 15 LEACH, A. R. *Molecular modelling: principles and applications*. 2nd. ed. Harlow: Prentice–Hall, 2001.
- 16 STILL, W. C. et al. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, v. 112, p. 6127–6129, 1990.
- 17 BORN, M. Volumes and hydration warmth of ions. *Zeitschrift fur physik*, v. 1, p. 45–48, 1920.
- 18 ONSAGER, L. Electric moments of molecules in liquids. *J. Am. Chem. Soc.*, v. 58, p. 1486–1493, 1936.
- 19 KIRKWOOD, J. G. Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, v. 3, p. 300–313, 1935.
- 20 MINH, D. D. L. Implicit ligand theory: rigorous binding free energies and thermodynamic expectations from molecular docking. *J. Chem. Phys.*, v. 137, p. 104106, 2012.
- 21 ÅQVIST, J.; MEDINA, C.; SAMMUELSSON, J.-E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.*, v. 7, p. 385–391, 1994.
- 22 WARSHEL, A.; RUSSELL, S. T. Calculations of electrostatic interactions in biological systems and in solutions. *Q. Rev. Biophys.*, v. 17, p. 283–422, 1984.
- 23 BEN-NAIM, A.; MARCUS, Y. Solvation thermodynamics of nonionic solutes. *J. Chem. Phys.*, v. 81, p. 2016–2027, 1984.
- 24 SU, Y. et al. Linear interaction energy (LIE) models for ligand binding in implicit solvent: theory and application to the binding of NNRTIs to HIV-1 reverse transcriptase. *J. Chem. Theory Comput.*, v. 3, p. 256–277, 2007.
- 25 GALLICCHIO, E.; KUBO, M. M.; LEVY, R. M. Enthalpy–entropy and cavity decomposition of alkane hydration freeenergies: numerical results and implications for theories of hydrophobic solvation. *J. Phys. Chem. B*, v. 104, p. 6271–6285, 2000.
- 26 ANDERSON, W. F. et al. Crystallographic determination of the mode of binding of oligosaccharides to T4 bacteriophage lysozyme: implications for the mechanism of catalysis. *J. Mol. Biol.*, v. 147, p. 523–543, 1981.
- 27 ANAND, N. N.; STEPHEN, E. R.; NARANG, S. A. Mutation of active site residues in synthetic T4-lysozyme gene and their effect on lytic activity. *Biochem. Biophys. Res. Commun.*, v. 153, p. 862–868, 1988.
- 28 MATTHEWS, B. W.; REMINGTON, S. J. The three dimensional structure of the lysozyme from bacteriophage T4. *Proc. Natl. Acad. Sci. U. S. A.*, v. 71, p. 4178–4182, 1974.
- 29 GRÜTTER, M. G. et al. Structural studies of mutants of the lysozyme of bacteriophage T4. The temperature-sensitive mutant protein Thr157→Ile. *J. Mol. Biol.*, v. 197, p. 315–329, 1987.
- 30 ALBER, T. et al. Contributions of hydrogen bonds of Thr 157 to the thermodynamic stability of phage T4 lysozyme. *Nature*, v. 330, p. 41–46, 1987.

- 31 MATSUMURA, M.; BECKTEL, W. J.; MATTHEWS, B. W. Hydrophobic stabilization in T4 lysozyme determined directly by multiple substitutions of Ile 3. *Nature*, v. 334, p. 406–410, 1988.
- 32 ERIKSSON, A. E. et al. A cavity-containing mutant of T4 lysozyme is stabilized by buried benzene. *Nature*, v. 355, p. 371–373, 1992.
- 33 WEI, B. Q. et al. A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.*, v. 322, p. 339–355, 2002.
- 34 MORTON, A.; MATTHEWS, B. W. Specificity of ligand binding in a buried nonpolar cavity of T4 lysozyme: linkage of dynamics and structural plasticity. *Biochemistry*, v. 34, p. 8576–8588, 1995.
- 35 WRAY, J. W. et al. Structural analysis of a non-contiguous second-site revertant in T4 lysozyme shows that increasing the rigidity of a protein can enhance its stability. *J. Mol. Biol.*, v. 292, p. 1111–1120, 1999.
- 36 BOUVIGNIES, G. et al. Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. *Nature*, v. 477, p. 111–114, 2011.
- 37 FABER, H. R.; MATTHEWS, B. W. A mutant T4 lysozyme displays five different crystal conformations. *Nature*, v. 348, p. 263–266, 1990.
- 38 ZHANG, X.; WOZNIAK, J. A.; MATTHEWS, B. W. Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme. *J. Mol. Biol.*, v. 250, p. 527–552, 1995.
- 39 SCHLICK, T. et al. Algorithmic challenges in computational molecular biophysics. *J. Comput. Phys.*, v. 151, p. 9–48, 1999.
- 40 GALLICCHIO, E.; LAPELOSA, M.; LEVY, R. M. Binding energy distribution analysis method (BEDAM) for estimation of protein–ligand binding affinities. *J. Chem. Theory Comput.*, v. 6, p. 2961–2977, 2010.
- 41 BOYCE, S. E. et al. Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site. *J. Mol. Biol.*, v. 394, p. 747–763, 2009.
- 42 DENG, Y.; ROUX, B. Calculation of standard binding free energies: aromatic molecules in the T4 lysozyme L99A mutant. *J. Chem. Theory Comput.*, v. 2, p. 1255–1273, 2006.
- 43 JIANG, W.; ROUX, B. Free energy perturbation Hamiltonian replica–exchange molecular dynamics (FEP/H-REMD) for absolute ligand binding free energy calculations. *J. Chem. Theory Comput.*, v. 6, p. 2559–2565, 2010.
- 44 RODINGER, T.; HOWELL, P. L.; POMÈS, R. Calculation of absolute protein–ligand binding free energy using distributed replica sampling. *J. Chem. Phys.*, v. 129, p. 155102, 2008.
- 45 GENHEDEN, S. et al. Nonpolar solvation free energies of protein–ligand complexes. *J. Chem. Theory Comput.*, v. 6, p. 3558–3568, 2010.
- 46 LINDER, M.; RANGANATHAN, A.; BRINCK, T. “Adapted linear interaction energy”: a structure–based LIE parametrization for fast prediction of protein–ligand affinities. *J. Chem. Theory Comput.*, v. 9, p. 1230–1239, 2013.

- 47 HUEY, R. et al. A semiempirical free energy force field with charge-based desolvation. *J. Comp. Chem.*, v. 28, p. 1145–1152, 2007.
- 48 VRIEND, G. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graphics*, v. 8, p. 52–56, 1990.
- 49 CHINEA, G. et al. The use of position-specific rotamers in model-building by homology. *Proteins*, v. 23, p. 415–421, 1995.
- 50 SCHAFTENAAR, G.; NOORDIK, J. H. Molden: a pre- and post-processing program for molecular and electronic structures. *J. Comput.-Aided Mol. Des.*, v. 14, p. 123–134, 2000.
- 51 DEWAR, M. J. S. et al. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.*, v. 107, p. 3902–3909, 1985.
- 52 DAURA, X. et al. Peptide folding: when simulation meets experiment. *Angew. Chem. Int. Ed.*, v. 38, p. 236–240, 1999.
- 53 REICH, M. et al. GeneCluster 2.0: an advanced toolset for bioarray analysis. *Bioinformatics*, v. 20, p. 1797–1798, 2004.
- 54 ARANTES, G. M.; RIBEIRO, M. C. C. A microscopic view of substitution reactions solvated by ionic liquids. *J. Chem. Phys.*, v. 128, p. 114503, 2008.
- 55 CARROLL, D. L. *Genetic algorithm driver, version 1.7*. 1998.
- 56 PRESS, W. H. et al. *Numerical Recipes in FORTRAN 77: The Art of Scientific Computing*. 2nd. ed. Cambridge: Cambridge University Press, 1992.
- 57 PRONK, S. et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, v. 29, p. 845–854, 2013.
- 58 JORGENSEN, W. L.; MAXWELL, D. S.; TIRADO-RIVES, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, v. 118, p. 11225–11236, 1996.
- 59 ONUFRIEV, A.; BASHFORD, D.; CASE, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized Born model. *Proteins*, v. 55, p. 383–394, 2004.
- 60 O'BOYLE, N. M. et al. Open Babel: an open chemical toolbox. *J. Cheminform.*, v. 3, 2011.
- 61 TJONG, H.; ZHOU, H.-X. *GBr*⁶: a parameterization-free, accurate, analytical Generalized Born method. *J. Phys. Chem. B*, v. 111, p. 3055–3061, 2007.
- 62 FRISCH, M. J. et al. *Gaussian 09, Revision A.1*. Gaussian, Inc., Wallingford CT, 2009.
- 63 HAWKINS, G. D. et al. *AMSOL, version 7.1*. 2004.
- 64 LI, J. et al. Accurate dipole moments from Hartree-Fock calculations by means of class IV charges. *J. Chem. Phys.*, v. 111, p. 885–892, 1999.

- 65 BERENDSEN, H. J. C.; GRIGERA, J. R.; STRAATSMA, T. P. The missing term in effective pair potentials. *J. Phys. Chem.*, v. 91, p. 6269–6271, 1987.
- 66 HESS, B.; VEGT, N. F. A. van der. Hydration thermodynamic properties of amino acid analogues: a systematic comparison of biomolecular force fields and water models. *J. Phys. Chem. B*, v. 110, p. 17616–17626, 2006.
- 67 BUSSI, G.; DONADIO, D.; PARRINELLO, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, v. 126, p. 014101, 2007.
- 68 PARRINELLO, M.; RAHMAN, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.*, v. 52, p. 7182–7190, 1981.
- 69 LIU, L.; BAASE, W. A.; MATTHEWS, B. W. Halogenated benzenes bound within a non-polar cavity in T4 lysozyme provide examples of I···S and I···Se halogen-bonding. *J. Mol. Biol.*, v. 385, p. 595–605, 2009.
- 70 SHEN, Y.; BAX, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. N.M.R.*, v. 48, p. 13–22, 2010.
- 71 ROBUSTELLI, P.; STAFFORD, K. A.; PALMER, A. G. Interpreting protein structural dynamics from NMR chemical shifts. *J. Am. Chem. Soc.*, v. 134, p. 6365–6374, 2012.
- 72 LIU, L. et al. Use of stabilizing mutations to engineer a charged group within a ligand-binding hydrophobic cavity in T4 lysozyme. *Biochemistry*, v. 48, p. 8842–8851, 2009.
- 73 MOBLEY, D. L. et al. Predicting absolute ligand binding free energies to a simple model site. *J. Mol. Biol.*, v. 371, p. 1118–1134, 2007.
- 74 GRAVES, A. P. et al. Rescoring docking hit lists for model cavity sites: predictions and experimental testing. *J. Mol. Biol.*, v. 377, p. 914–934, 2008.
- 75 LIU, L. et al. Boron mimetics: 1,2-dihydro-1,2-azaborines bind inside a nonpolar cavity of T4 lysozyme. *Angew. Chem. Int. Ed. Engl.*, v. 48, p. 6817–6819, 2009.
- 76 GRAVES, A. P.; BRENK, R.; SHOICHET, B. K. Decoys for docking. *J. Med. Chem.*, v. 48, p. 3714–3728, 2005.
- 77 BOWMAN, G. R.; GEISLER, P. L. Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proc. Natl. Acad. Sci. U. S. A.*, v. 109, p. 11681–11686, 2012.
- 78 MORTON, A.; BAASE, W. A.; MATTHEWS, B. W. Energetic origins of specificity of ligand binding in an interior nonpolar cavity of T4 lysozyme. *Biochemistry*, v. 34, p. 8564–8575, 1995.
- 79 SU, A. I. et al. Docking molecules by families to increase the diversity of hits in database screens: computational strategy and experimental evaluation. *Proteins*, v. 42, p. 279–293, 2001.
- 80 WEI, B. Q. et al. Testing a flexible-receptor docking algorithm in a model binding site. *J. Mol. Biol.*, v. 337, p. 1161–1182, 2004.

- 81 LEMKUL, J. A.; ALLEN, W. J.; BEVAN, D. R. Practical considerations for building GROMOS-compatible small-molecule topologies. *J. Chem. Inf. Model.*, v. 50, p. 2221–2235, 2010.
- 82 AUDRIETH, L. F.; NESPITAL, W.; ULRICH, H. Electric moments of hydrazine and its derivatives. *J. Am. Chem. Soc.*, v. 55, p. 673–678, 1933.
- 83 LANDER, J. J.; SVIRBELY, W. J. The dipole moments of catechol, resorcinol and hydroquinone. *J. Am. Chem. Soc.*, v. 67, p. 322–324, 1945.
- 84 FERRARA, P. et al. Assessing scoring functions for protein–ligand interactions. *J. Med. Chem.*, v. 47, p. 3032–3047, 2004.
- 85 COLLINS, M. D. et al. Cooperative water filling of a nonpolar protein cavity observed by high-pressure crystallography and simulation. *Proc. Natl. Acad. Sci. U.S.A.*, v. 102, p. 16668–16671, 2005.
- 86 BUCH, I.; GIORGINO, T.; FABRITIIS, G. D. Complete reconstruction of an enzyme–inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.*, v. 108, p. 10184–10189, 2011.
- 87 SÖDERHJELM, P.; TRIBELLO, G. A.; PARRINELLO, M. Locating binding poses in protein–ligand systems using reconnaissance metadynamics. *Proc. Natl. Acad. Sci. U.S.A.*, v. 109, p. 5170–5175, 2012.
- 88 ALMLÖF, M.; BRANDSDAL, B. O.; ÅQVIST, J. Binding affinity prediction with different force fields: examination of the linear interaction energy method. *J. Comput. Chem.*, v. 25, p. 1242–1254, 2004.