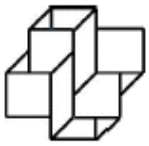


Laboratório Nacional de Computação Científica
Programa de Pós Graduação em Modelagem Computacional

Construção de Funções Empíricas Utilizando Rede Neural para Determinação de Constantes de Afinidade Receptor-Ligante

Por

Thaís Gaudencio do Rêgo

B.Sc. Ciências Biológicas - UFPB/2003

PETRÓPOLIS, RJ - BRASIL

AGOSTO DE 2008

CONSTRUÇÃO DE FUNÇÕES EMPÍRICAS UTILIZANDO REDE NEURAL PARA

DETERMINAÇÃO DE

CONSTANTES DE AFINIDADE RECEPTOR-LIGANTE

Thaís Gaudencio do Rêgo

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DE RECURSOS HUMANOS DO LABORATÓRIO NACIONAL DE COMPUTAÇÃO CIENTÍFICA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA OBTENÇÃO DO TÍTULO DE MESTRE EM MODELAGEM COMPUTACIONAL COM ÊNFASE EM BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL.

Aprovada por:

_____

Laurent Emmanuel Dardenne, D.Sc. - LNCC

_____

Hélio J.C. Barbosa, D.Sc. - LNCC

_____

Carlos Maurício Rabello de Sant'anna, D.Sc. UFRRJ

_____

Fernanda Maria Pereira Raupp, D.Sc. - LNCC/PUC-RJ

PETRÓPOLIS, RJ - BRASIL

AGOSTO DE 2008

Rêgo, Thaís Gaudencio

R343c Construção de Funções Empíricas Utilizando Rede Neural para Determinação de Constantes de Afinidade Receptor-Ligante / Thaís Gaudencio do Rêgo -- Petrópolis, RJ.: Laboratório Nacional de Computação Científica, 2008.

xxiii, 139 p. : il. ; 29 cm.

Orientadores : Laurent Emmanuel Dardenne ; Hélio José Correa Barbosa

Dissertação (Mestrado) - Laboratório Nacional de Computação Científica, 2008.

1. Biomoléculas – Estrutura – Simulação por computador 2. Proteínas - Estrutura – Simulação por computador 3. Atracamento molecular 4. Desenvolvimento de fármacos 5. Energia livre I. Dardenne, Laurent Emmanuel II. Barbosa, Hélio José Correa III. MCT/LNCC IV.Título

CDD – 572.33

"O destino do homem é ser livre."

Nietzsche

Dedico à minha família e amigos.

Agradecimentos

Gostaria de agradecer primeiramente a meus pais, que são exemplos de vida, de pessoas honestas, educadas, inteligentes e maiores incentivadores de minha vida acadêmica. Amo-os incondicionalmente e espero retribuir por tudo que fizeram por mim.

Não menos importantes são meus irmãos Matheus e Gabriel (Pirra), meninos perfeitos, extremamente alegres e que me fizeram sorrir e chorar de saudades muitas vezes ao longo destes anos. À minha avó, minhas tias e tios (Badêi, Badé e Júnior, principalmente) e primos que tanto amo, Isabel, Érica, Elisa, Ernesto e Felipe. Todos loucos na medida certa. E à Carol, Simone, Vovó Bita e Solange.

A Felipe que ao longo dos últimos 3 anos esteve sempre do meu lado, com a maior paciência do mundo. Aos novos ventos...

Ao meu orientador Laurent e meu co-orientador Hélio, por estarem sempre tão presentes e cuidarem de mim e de nosso trabalho como se fôssemos únicos, adotando-nos como nenhum acadêmico poderia fazer de forma mais entusiasmada.

Aos meus antigos orientadores de monografia e projetos, Christina Bonato e Demétrius Araújo por despertarem essa fascinação que eu tenho pelas ciências e por até hoje acompanharem meus passos.

Aos amigos, os mais novos e mais importantes nessa etapa de minha vida, Papablo, Marco, Tutuca e Mula, parceirões de madrugadas, de café, de filmes, de conversas, de conquistas, amigos que vou guardar para a vida toda. Ao Pablo em especial por fazer minha vida bem mais divertida.

As amigas mais lindas e divertidas que uma pessoa pode ter, em ordem alfabética para evitar brigas: Adri, Anabelle, Camila, Chandra, Claudinha, Elen, Fabíola, Flavinha, Isabelle, Juliana Peba, Karina, Lígia (apesar de ter sempre coisa melhor pra fazer), Marbella, Mara, Mari, Marina, Marisa, Monika, Patrícia Uni, Patrícia, Priscila, Rafa, Renata Baratinha, Renata e Salomé. Renata,

Rafinha e Marina, principalmente, por tornarem meus domingos em Petrópolis bem mais alegres.

Aos amigos, tão lindos e divertidos quanto às meninas, Alexandre, André Enxaqueca, Arthur (Marmita), Bidu, Dadado, David, Dioguinho, Diogão, Flávio Embriagado e agora pai, Flávio, Giorgio, Glauberto, Héllio, Itácio, Italo, João, Luíz, Marcel (Mocotó), Murilo, Marx, Neto, Oberdan, Pablo Hess, Rangel, Reinaldo, Sávio, Tércio, Thibério, Tito, Toni (Delegado), Vandique e Vinícius (Binha). Ao longo desses anos passamos por muitas mudanças, mas a única que aconteceu em relação a amizade com esses transeuntes foi a intensidade, que aumentou.

Ao LNCC, pela estrutura, pelo incentivo e pelo exemplo de instituição pública que funciona e que deveria ser seguido pelo resto do Brasil.

A todos os professores com quem tive o prazer de estudar, Regina, Hélio, Raul, Ana Tereza, Bidu e Bordoni. Às secretárias Ana Paula e Ana Neri, ao pessoal da limpeza, aos seguranças e a todos que de alguma forma contribuíram para a realização deste trabalho.

A CAPES, pelo primeiro ano de bolsa e à FAPERJ pelo segundo ano de bolsa, esta com adicional, como prêmio de Aluno nota 10, pelo Mestrado de Modelagem Computacional do LNCC!!

Resumo da Dissertação apresentada ao LNCC;MCT como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

CONSTRUÇÃO DE FUNÇÕES EMPÍRICAS UTILIZANDO REDE NEURAL PARA DETERMINAÇÃO DE CONSTANTES DE AFINIDADE RECEPTOR-LIGANTE

Thaís Gaudencio do Rêgo

Agosto, 2008

Orientador: Laurent Emmanuel Dardenne, D.Sc. - LNCC

Co-orientador: Hélio José Correa Barbosa, D.Sc. - LNCC

A compreensão dos mecanismos de reconhecimento molecular receptor-ligante é um dos aspectos centrais na descoberta e planejamento de novos fármacos baseado em estrutura. Uma metodologia chave é o atracamento de pequenas moléculas em sítios de ligação de proteínas, o atracamento molecular (em inglês, "*molecular docking"*). Existem dois pontos chaves em qualquer programa de atracamento: a busca da "melhor" conformação ligante-proteína e o cálculo da energia livre desta associação, ou sua constante de afinidade. Foi construído neste trabalho um conjunto-teste formado por 50 complexos proteína-ligante, com valores de K_i ou K_d determinados experimentalmente, para a construção de uma função empírica específica para o programa DOCKTHOR, utilizando como variáveis de entrada valores de energias de interação eletrostática e de Lennard-Jones, área de contato ligante-receptor da superfície acessível ao solvente, presença de ligações hidrogênio, e o número de ligações torcionáveis do ligante. Estes variáveis foram utilizados para a construção de dois tipos de funções de cálculo de energia livre. Através de regressão múltipla, foi avaliada a importância de cada uma das variáveis utilizadas como dados de entrada na construção desta função. Utilizando uma rede neural, buscou-se construir o melhor modelo para o cálculo de constantes de afinidade. O programa DOCKTHOR atualmente tem poder de predição correspondente a r =

0,4245, o que mostra a importância de se melhorar sua função de avaliação. A função construída com a metodologia de regressão múltipla que obteve melhor resultado foi a que utilizou as 5 variáveis de entrada apresentando termos lineares, cruzados e quadráticos, com r igual a 0,7542. Funções empíricas construídas por redes neurais também foram avaliadas neste trabalho. Utilizando a metodologia de validação cruzada de grupo (VCG) chegou-se à conclusão que a melhor arquitetura para a rede neural é constituída por 9 neurônios na camada oculta, pois possui o menor erro de generalização e a maior homogeneidade nos erros. No teste com esta arquitetura de rede neural, com a função construída utilizando os 50 complexos proteína-ligante no treinamento e os mesmos, no teste, observamos que 66% dos complexos tiveram uma diferença menor que 1,0 dos valores observados em relação aos esperados. O erro de generalização, obtido por VCG, de uma rede neural utilizando 9 neurônios na camada oculta foi cerca de dez vezes menor ao obtido utilizando uma função polinomial. Isto é um indicativo da superioridade da metodologia de rede neural, com relação a metodologia de regressão multivariada, principalmente em uma função empírica desenvolvida para estimar afinidades relativas à uma ampla gama de complexos receptor-ligante.

Abstract of Dissertation presented to LNCC/MCT as a partial fulfillment of the requirements for the degree of Master of Sciences (M.Sc.)

CONSTRUCTION OF EMPIRICAL SCORING FUNCTIONS USING ARTIFICIAL NEURAL NETWORK FOR DETERMINATION OF AFFINITIES CONSTANTS BETWEEN RECEPTOR-LIGAND

Thaís Gaudencio do Rêgo

August, 2008

Advisor: Laurent Emmanuel Dardenne, D.Sc. - LNCC

Co-advisor: Hélio José Correa Barbosa, D.Sc. - LNCC

The understanding of receptor-ligand molecular recognition is one of the central aspects in structure-based design and discovery of new drugs. The key methodology is the docking of small molecules in active sites of proteins. There are two aims in any program of molecular docking: the search for the best ligand-protein conformation and the calculation of the free energy of this association, or its affinity constant. The test set used in this work was composed by 50 protein-ligand complexes, with experimentally measured Ki or Kd values for the construction of an empirical function specific to the DOCKTHOR program, using as input variables: energy of electrostatic interaction and Lennard-Jones, contact area of ligand-receptor on the surface accessible to the solvent, the presence of hydrogen bridges, and the number of the ligand rotatable bonds that were frozen in the process of docking. These variables were used for the construction of two types of free energy scoring functions. The importance of each variable used as input data for the construction of those functions was rated by means of multiple regression. A neural network was

also used to try to build the best model for the calculation of the affinity constant. The DOCKTHOR program currently has a prediction power leading to r = 0.4245, which indicates the importance of improving its scoring. The function built with the multiple regression methodology used 5 input variables and had linear, quadratic, and cross-product terms leading to r = 0.7542. Using the methodology of group cross-validation (VCG), it was concluded that the best architecture for the neural network consists of 9 neurons in the hidden layer, as it has the smallest error of generalization and greater consistency in errors. In the tests with this neural network architecture built using the same 50 protein-ligand complexes in training and test, 66% of the complexes had a difference smaller than 1.0 in the observed values. The generalization error (obtained by VCG) of a neural network that uses 9 neurons in the hidden layer was about ten times lower than that obtained by using a polynomial function. This is an indication of the superiority of the neural network methodology with respect to the multivariate regression methodology, specially for an empirical function developed for a broad range of receptor-ligand complexes.

Sumário

Lista de Figuras

Figura

Lista de Gráficos

Gráficos

Lista de Tabelas

Tabelas

Lista de Siglas e Abreviaturas

λ: Parâmetro de Acoplamento

Å: Ångstroms

A: Energia livre de Helmholtz

ADA: Adenosina Deaminase

AG: Algoritmo Genético

ANN ("*Artificial Neural Network*"): Rede Neural Artificial

Asp: Ácido Aspártico

Ca^{2+}: Íon Cálcio

cLogP: Medida de lipofilicidade

CV: Validação Cruzada

Da: Dalton

DRBE: Desenho Racional de Fármacos Baseado em Estrutura

E: Enzima

EI: Complexo Enzima-Inibidor

EL: Interação Eletrostática

Err: Erro de generalização calculado em toda a população

Êrr: Erro de generalização calculado em uma amostra da população

ELE: Energia Eletrostática

FKBP: Proteína do grupo imunofilina que tem afinidade com fármacos FK506

G: Energia livre de Gibbs

Glu: Ácido Glutâmico

GVC: validação cruzada de grupo

H: Hamiltoniana

H: Entalpia

His: Histidina

I: Inibidor

IT: Integração Termodinâmica

K_a: Constante de Associação

K_{eq}: Constante de Equilíbrio

K_d: Constante de Dissociação

K_i: Constante de Inibição

LIE ("*Linear Interaction Energy*"): Energia de Interação Linear

LH: Ligações Hidrogênio

pK_d = log Kd: Logaritmo da Constante de Dissociação

pKi = log K_i: Logaritmo da Constante de Inibição

MDS ("*Molecular Dot Surface"*): Superfície Molecular Pontual

NTOR: Número de ligações torcionáveis congeladas

PDB ("*Protein Data Bank*"): Banco de Dados de Proteínas

PDE: Fosfodiesterase

PEL: Perturbação da Energia Livre

pKa: Logaritmo da Constante de Associação

r: Coeficiente de Correlação Linear

RMN: Ressonância Magnética Nuclear

RMSD: Desvio Médio Quadrático

S: Entropia

SAS: Superfície Acessível ao Solvente

SCOP ("*Structural Classification of Proteins*"): Classificação Estrutural de Proteínas

STRM: Seleção por Torneio Restrito Modificado

SVM ("*Support Vector Machine*"): Máquina de Vetor de Suporte

TIM: Triose-fosfato isomerases

T: Temperatura absoluta (em graus Kelvin).

VDW: Energia de van der Waals

VP: Proteínas do revestimento viral e capsídeo

WDI ("*World Drug Index*"): Catálogo Mundial de Drogas

CAPÍTULO 1

Este trabalho tem como objetivo a construção de uma função empírica para o cálculo do valor de afinidade entre receptor-ligante utilizando uma rede neural treinada com o algoritmo de retropropagação. Esta função empírica foi construída objetivando o seu uso no programa de atracamento molecular receptor-ligante desenvolvido pelo Grupo de Modelagem Molecular de Sistemas Biológicos/LNCC (de Magalhães, 1995). Este trabalho está dividido em seis capítulos. No Capítulo introdutório, podemos encontrar um breve histórico do processo de descoberta e desenvolvimento de fármacos, dos primórdios do processo até a incorporação de ferramentas computacionais para a descoberta e melhoramento destas moléculas em uma intervenção terapêutica direcionada e racional. São também descritos os pontos-chaves da técnica de atracamento molecular de um ligante ao sítio ativo de um receptor.

Como o foco deste trabalho compreendeu o desenvolvimento de funções empíricas para o cálculo da constante de afinidade no atracamento molecular, um tópico sobre energia livre e termodinâmica foi incluído, bem como uma breve descrição de algumas metodologias atualmente existentes para o seu cálculo. Em seguida, detivemo-nos na descrição dos tipos de funções de cálculo de energia livre de ligação acoplada aos programas de atracamento molecular e, por fim, apresentamos um breve resumo de todas as etapas executadas por estes programas.

Como utilizamos uma rede neural para a construção de uma função empírica que represente a energia livre, descrevemos o funcionamento desta técnica com maiores detalhes no Capítulo 2.

No Capítulo 3, destacamos os objetivos deste trabalho. A metodologia que utilizamos é detalhadamente apresentada no Capítulo 4.

No Capítulo 5, apresentamos os resultados obtidos. As conclusões deste trabalho, assim como nossas perspectivas de pesquisas futuras, estão presentes no Capítulo 6.

1. INTRODUÇÃO

Desde tempos imemoráveis, a humanidade tem utilizado a flora e a fauna que a cerca na busca pela cura de suas doenças (Koehn e Carter, 2005). A experiência acumulada ao longo dos anos foi, por séculos, nossa única fonte de conhecimento farmacêutico. No último deles, no entanto, esta dependência foi quebrada por duas razões. Primeiramente, os avanços na biologia ampliaram nossa compreensão acerca dos mecanismos moleculares através dos quais as doenças instalam-se em nossos organismos, sugerindo alvos moleculares para uma intervenção terapêutica racional. Além disso, avanços na química orgânica permitiram o projeto e a síntese de moléculas sofisticadas para atuarem como fármacos. Na década de 1950, tais medicamentos puramente sintéticos mantinham o mecanismo funcional similar a algum produto natural já conhecido. Não obstante, os produtos naturais são considerados ainda uma fonte valiosa de possíveis fármacos e o teste de extratos de produtos naturais é praticado extensamente na indústria farmacêutica (Ortholand e Ganesan, 2004).

Ao contrário do método histórico de descoberta de novos fármacos por testes de tentativa e erro de substâncias químicas em animais, o desenho racional de fármacos é realizado com o conhecimento das respostas químicas específicas no corpo ou no organismo alvo, e com a utilização destas informações para o ajuste de um tratamento ideal. Um exemplo particular de desenho racional de fármacos envolve o uso da informação tridimensional de biomoléculas obtido com técnicas como difração de Raios-X em cristais e espectrometria de Ressonância Magnética Nuclear (RMN) e é referida normalmente como Desenho Racional de Fármacos Baseado em Estrutura (DRBE).

Como o número de proteínas com sua estrutura tridimensional conhecida vem crescendo enormemente, tanto pela melhoria das técnicas de determinação estrutural, como a difração de raios-X em cristais de alto rendimento (Blundell e Patel, 2004; Blundell et al. 2002), quanto em

razão da disponibilidade destes dados surgidos de projetos de genômica estrutural, vem crescendo também o número de proteínas que podem ser utilizadas como alvo no desenho racional de fármacos baseado em estruturas (Kitchen et al., 2004).

Sistemas biológicos contêm “apenas” quatro tipos de macromoléculas alvo para as quais podem ser desenvolvidas pequenas moléculas como agentes terapêuticos que podem interferir nas suas funções, são elas: proteínas, polissacarídeos, lipídeos e ácidos nucléicos. No entanto, para avaliar a possibilidade de serem utilizadas pela indústria farmacêutica, devem ser estudadas as propriedades que são requeridas para viabilizar o uso destas moléculas como fármacos, inclusive do ponto de vista comercial (Hopkins e Groom, 2002).

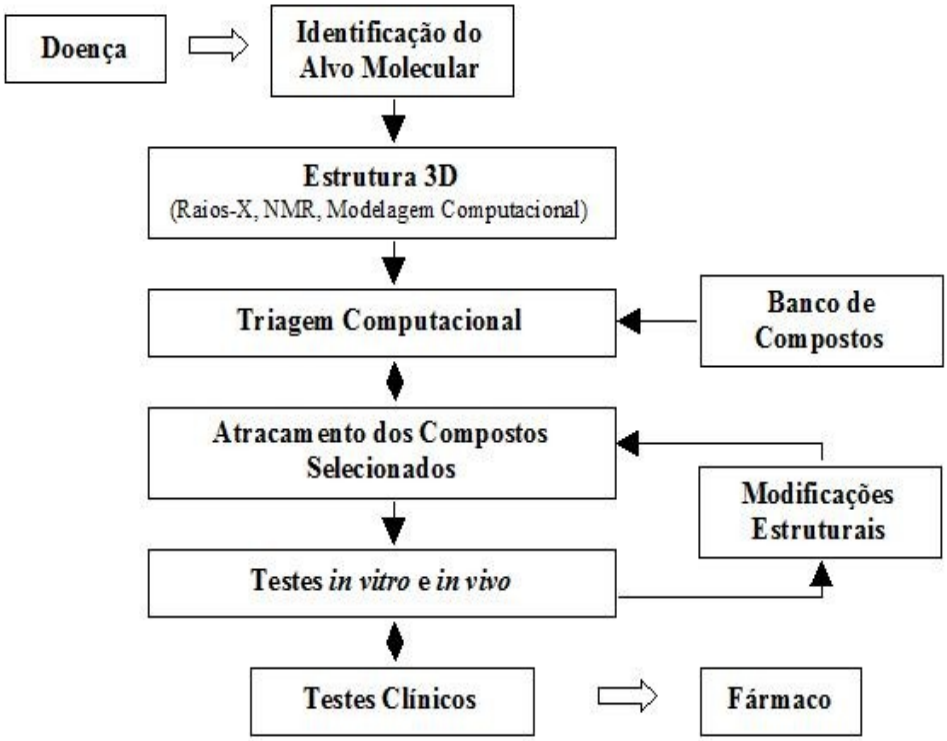
O mais bem conhecido estudo nesta área é o trabalho de Lipinski e co-autores da Pfizer, que fizeram a análise estatística de 2.200 fármacos do Catálogo Mundial de Drogas (WDI, "*World Drug Index*"). Eles avaliaram um conjunto de características presentes na maior parte dos fármacos considerados neste estudo, normalmente conhecido como "Regra Pfizer" ou "regra dos cinco", a qual afirma que a absorção ou permeabilidade de um fármaco (que não é um substrato para um transportador biológico) é provavelmente enfraquecida quando o peso molecular é maior que 500 Da, a lipoficilidade é alta (expressa como cLogP > 5), o número de grupos doadores de hidrogênio é maior que 5 e o número de grupos aceptores de hidrogênio é maior que 10 (Lipinski et al., 1997). Estas regras não podem ser aplicadas a produtos naturais, para os quais outros mecanismos de absorção estão envolvidos (Hopkins e Groom, 2002).

No genoma humano, o número de genes que expressam proteínas susceptíveis à ação de fármacos é de aproximadamente 30.000. As similaridades de seqüência e função dentro de uma família gênica são usualmente indicativas de uma conservação geral da arquitetura do sítio de ligação entre membros de uma mesma família de proteínas. Isto poderia sugerir que se um membro de uma família gênica é capaz de se ligar a um fármaco, outros membros poderiam também ligar-se a um composto com propriedades físico-químicas ou estruturais similares. Usando esta afirmação,

3.051 das 30.000 proteínas preditas apresentam algum precedente para ligação de um fármaco. Por fim, estimativas recentes propõem que existam de 3.000 a 10.000 genes relacionados à doenças, e estudos de grande escala de nocaute em rato revelaram que apenas aproximadamente 10% dos genes que sofreram nocaute poderiam ter potencial para interferir na doença. Os potenciais alvos de fármacos que a indústria farmacêutica pode explorar estão na interseção do conjunto do genoma que pode ser ligado a algum fármaco e os genes relacionados a doenças, entre valores de 600 e 1.500 alvos de fármacos para o homem (Hopkins e Groom, 2002).

A descoberta, desenvolvimento e registro de um fármaco é uma operação imensamente custosa e representa um desafio único. Em 2003, os custos na descoberta e no desenvolvimento de um fármaco eram da ordem de 800 milhões de dólares, levando até 16 anos para que este chegasse ao mercado (Dickson e Gagnon, 2003; Preziosi, 2004). Para cada grupo de 5.000 a 10.000 compostos sintetizados ou isolados com potencial terapêutico, apenas um, em média, alcançará o mercado (Preziosi, 2004). As etapas do desenho racional de fármacos baseado em estrutura desde a identificação do alvo molecular relativo a uma doença, até a geração do fármaco a ela associado, está ilustrado na Figura 1.1 (de Magalhães, Barbosa e Dardenne, 2007) e descrito em seguida (Barreiro, 2007).

Figura 1.1: Etapas do Desenho Racional de Fármacos Baseado em Estrutura.



Fonte: de Magalhães, Barbosa e Dardenne, 2007.

O primeiro passo deste processo envolve o estudo da doença para a busca da cura ou a redução significativa de sintomas. Este processo envolve o entendimento do estado atual da etiologia da doença, o conhecimento científico a seu respeito, a necessidade médica e a oportunidade comercial.

Como discutido anteriormente, há muito tempo o conhecimento de nossos antepassados é utilizado para o tratamento de doenças a partir de plantas e outros organismos (Koehn e Carter, 2005). Isto ocorre em virtude da presença de um princípio ativo que apresenta atividade farmacológica. Este mesmo princípio ativo pode ser isolado e utilizado pela indústria farmacêutica no desenvolvimento ou melhoramento de um novo fármaco. Além da medicina tradicional e de pesquisas realizadas na natureza, a descoberta de compostos ativos pode ocorrer a partir de

fármacos já existentes ou mesmo por acaso, como ocorreu na descoberta da penicilina.

As modernas estratégias usadas no planejamento racional de novos compostos-protótipos se baseiam na abordagem fisiológica. Essa abordagem permite planejar a estrutura química de uma nova molécula com base na definição prévia do mecanismo de ação terapêutica, ou seja, a biomacromolécula à qual o novo fármaco irá se ligar para alterar um determinado processo bioquímico.

Conhecendo-se a estrutura molecular tridimensional do alvo terapêutico eleito, em particular a região responsável pela interação química, é possível identificar um composto capaz de se ligar àquele sítio ativo e alterar suas propriedades utilizando-se técnicas computacionais de modelagem molecular. O desenho racional dessa molécula é realizado utilizando-se técnicas de simplificação molecular (redução de complexidade estrutural de um composto), hibridação molecular (obtenção de nova estrutura química a partir de partes de duas ou mais substâncias bioativas distintas) e bioisosterismo (substituição de partes da estrutura de uma substância bioativa por outra ou outras com comportamento eletrônico similar), além do uso da intuição química de um pesquisador experiente para propor novas modificações.

Uma nova abordagem computacional vem sendo utilizada principalmente pelos laboratórios de pesquisas industriais, onde é feita a busca de compostos protótipos em bancos de dados que contêm a descrição de uma imensa variedade de compostos naturais e sintéticos, puros ou combinados para a realização de bioensaios. O objetivo é descobrir compostos-protótipos ativos que possam, ao serem avaliados experimentalmente, apresentar um nível de afinidade receptor-ligante em nível micromolar (μM), ou seja, ativo em uma concentração de um milionésimo de mol por litro, ou nanomolar (nM), ativo em uma concentração de um bilionésimo de mol por litro.

Além desta metodologia de filtragem *in silico*, essa busca é realizada aplicando a "Regra dos Cinco", onde podem ser eliminadas moléculas que violam uma ou mais regras, são ainda aplicados um conjunto de regras empíricas e sistemas baseados em decisão. Estas filtragens podem ser

utilizadas como seleção positiva e negativa e estão sendo cada vez mais complementados por algoritmos que predizem propriedades físico-químicas, habilidade para penetrar a barreira cérebro-sangue, toxicidade baixa, entre outras (Schneider e Böhm, 2002).

Ainda pode ser realizada a filtragem *in vivo*, onde a exclusão de determinadas moléculas acontece depois de aplicados alguns testes experimentais, como de citotoxicidade ou de determinadas propriedades citocinéticas (Walum et al., 2005). Após a descoberta do composto-protótipo desejado, uma série congênere (compostos similares, com pequenas variações estruturais) deve ser construída e avaliada farmacologicamente, para orientar a otimização do futuro fármaco. Essa etapa é fundamental, pois representa a busca da melhor caracterização da eficácia terapêutica do candidato a fármaco.

Deve então ser realizada a etapa de testes para a validação do ligante selecionado, determinando se este é, de fato, a molécula ideal para a obtenção do resultado esperado com a avaliação das propriedades farmacocinéticas do ligante, relacionadas ao seu trajeto no organismo, desde a absorção, distribuição e eliminação.

Os testes pré-clínicos envolvem quatro passos. O primeiro, de estudos farmacológicos primários de natureza variável, depende do tipo de fármaco em desenvolvimento. Em geral, os estudos de farmacodinâmica são baseados em testes tanto *in vitro* (testes molecular e celular), como *in vivo* (modelos animais espontâneos ou experimentais da doença) (Preziosi, 2004).

O segundo teste refere-se a estudos farmacológicos secundários, os quais estão relacionados à segurança do fármaco, em termos de efeitos potenciais da molécula em determinados órgãos e no organismo todo. O terceiro estágio compreende estudos farmacocinéticos iniciais, focados na absorção, metabolismo e excreção da molécula candidata à fármaco. Testes deste tipo requerem o desenvolvimento de processos analíticos para o cálculo dos níveis da molécula e de seus metabólitos em fluidos biológicos. O quarto e último estágio de estudos pré-clínicos é o teste toxicológico (Preziosi, 2004). Esses estudos são importantes porque permitem definir a via de

administração do futuro fármaco, as doses a serem administradas e os intervalos de tempo em que isso deve ser feito (Barreiro, 2007).

Depois dos testes pré-clínicos são avaliados os resultados da atuação dos ligantes e, caso estes não atinjam as expectativas buscadas, novas alterações estruturais são realizadas nestas moléculas até que se encontre o fármaco ideal que possa ser aprovado em testes clínicos.

O período de testes clínicos é dividido em três fases. Na primeira fase os testes são conduzidos em um pequeno número de voluntários saudáveis para determinação da dose segura a ser administrada e a toxicidade de um composto. Caso uma determinada molécula seja considerada promissora, esta alcança a fase 2, na qual é testada em um número maior de voluntários que apresentam a doença que o fármaco se propõe curar ou controlar seus sintomas. Novamente, se for constatada sua atuação, esta irá para a fase 3, onde será submetida a um número maior de pessoas com a doença de interesse, nos quais serão testadas doses diferentes da fase 2. A principal proposta da fase 3 é demonstrar a eficácia da molécula, entretanto, como a fase 3 envolve mais voluntários que a fase 2, espera-se encontrar mais eventos adversos. Se ao fim da fase 3 o composto for considerado promissor, ele será submetido ao órgão responsável pela aplicação de novos fármacos do país onde a pesquisa foi desenvolvida (Dickson e Gagnon, 2004).

Cada estágio desse processo requer tempo e investimento, sendo de grande interesse identificar o mais cedo possível os agentes que são provavelmente menos promissores, permitindo uma concentração de esforços nos compostos que têm maior probabilidade de chegar ao mercado.

Na descoberta e planejamento de novos fármacos dentro da área de Desenho Racional de Fármacos Baseado em Estrutura (DRBE), a compreensão dos mecanismos de reconhecimento molecular receptor-ligante, além de ser um dos principais desafios da biologia molecular, é um dos aspectos centrais para seu sucesso (de Magalhães, Barbosa e Dardenne, 2007).

Com o intuito de melhor entender e simular este mecanismo, metodologias computacionais tornaram-se componentes cruciais de muitos programas utilizados na produção de fármacos,

buscando, por ligantes ou estruturas ideais, com técnicas de filtragem em bancos de dados como também no refinamento e otimização dos compostos previamente identificados. Uma metodologia chave, o atracamento de pequenas moléculas em sítios de ligação de proteínas, foi criada no início da década de 1980 e tornou-se uma área de pesquisa altamente ativa (Kitchen et al., 2004).

A abordagem computacional possibilita a realização de testes que agilizam o processo manual gerando uma economia considerável nos custos relativos à produção de novos fármacos. (Bock e Gough, 2002; Brooihmans e Kuntz, 2003)

1.1. Atracamento Molecular

O atracamento molecular (em inglês, "*molecular docking*") pode ser definido como a predição da estrutura de complexos receptor-ligante, onde o receptor normalmente é uma proteína ou um oligômero protéico e o ligante é uma molécula pequena natural ou sintética, peptídeos, proteínas ou ainda ácidos nucléicos. O sucesso nos estudos iniciais, ainda na década de 1970, e em estudos desenvolvidos por Levinthal et al. (1975) envolvendo a interação de moléculas de hemoglobina em microtúbulos ou fibras, guiou a exploração de atracamentos moleculares como uma ferramenta na descoberta de fármacos, para encontrar e otimizar compostos, freqüentemente por filtragem em bancos de dados (Brooijmans e Kuntz, 2003).

Existem dois pontos chaves em qualquer programa de atracamento: a busca da "melhor" conformação resultante da formação do complexo ligante-proteína e o cálculo da energia livre desta associação, ou sua constante de afinidade (Åqvist e Marelius, 2001).

O algoritmo de busca deve investigar a hipersuperfície de energia da forma mais eficiente possível, tentando encontrar o mínimo global de energia livre. No atracamento considerando o receptor rígido, isso significa que o algoritmo de busca explora diferentes posições para o ligante no sítio ativo do receptor, utilizando os graus de liberdade translacionais, rotacionais e

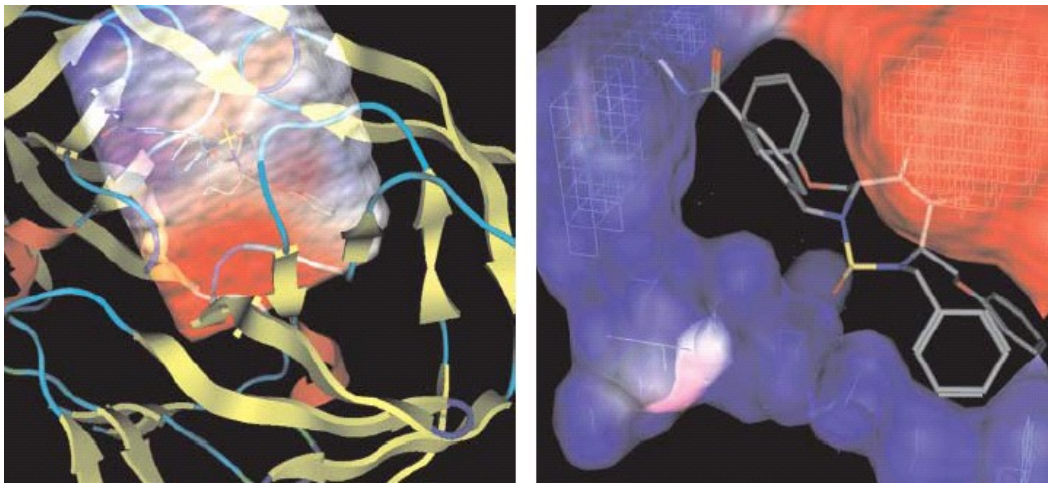
conformacionais do mesmo (para ligantes flexíveis) (Brooijmans e Kuntz, 2003).

Esse problema pode ser melhor entendido a partir do modelo chave-fechadura proposto por Emil Fischer em 1894. Nele, o ligante seria a chave e a fechadura seria o sítio ativo da uma proteína, na qual a chave deve se encaixar com perfeição. No entanto, o modelo chave-fechadura não reproduz com fidelidade o mecanismo existente na natureza, onde, durante o processo de interação, mudanças conformacionais ocorrem tanto no ligante quanto no sítio ativo da proteína. É justamente a exploração dessas modificações no receptor, um dos grandes desafios dos novos programas de atracamento, com receptores flexíveis.

O tratamento da flexibilidade do ligante é algo corriqueiro nos atuais métodos de atracamento molecular, sendo que seu desempenho cai drasticamente com o aumento dos graus de liberdade conformacionais (*i.e.,* aumento do número de ligações químicas torcionáveis) do ligante (de Magalhães et al., 1995). Alguns métodos têm tentado incluir a flexibilidade do receptor levando em consideração os graus de liberdade torcionais das cadeias laterais dos aminoácidos no sítio ativo da proteína (Leach et al., 2006), ou considerando diversas conformações da proteína obtidas de diferentes estruturas cristalográficas (*i.e*., onde a mesma proteína está complexada com distintos ligantes) ou obtidas de cálculos de dinâmica molecular (Apostolakis et al., 1998; Trosset e Scheraga, 1999).

Existem três tipos básicos para a representação do receptor: atômica, superfície molecular e em grade (Figura 1.2). Entre estas, a representação atômica é usada geralmente em conjunção com uma função de energia potencial e é frequentemente utilizada apenas durante o processo final de "ranqueamento" dos modos de ligação do ligante devido à alta complexidade do cálculo das interações interatômicas.

Figura 1.2: Representações do receptor em grade.



Fonte: Kitchen et al., 2004.

À esquerda é mostrada um esquema de uma grade capturando o potencial eletrostático da HIV protease (código PDB: 1BVE) ao redor de seu sítio ativo (com o inibidor Dmp323). Áreas em vermelho e azul indicam o potencial eletrostático negativo e positivo, respectivamente. Á direita, podemos observar o raio de corte da grade de potencial eletrostático da enzima ao redor do inibidor atracado.

Os programas baseados na superfície molecular são tipicamente, mas não exclusivamente, usados em programas de atracamento proteína-proteína. Estes métodos tratam de pontos alinhados nas superfícies para a minimização do ângulo formado entre as superfícies das distintas moléculas. Portanto, uma aproximação de corpo rígido é mantida normalmente em muitas técnicas de atracamento proteína-proteína.

O uso de grades de potencial de energia foi iniciado por Goodforf (1985) e vários programas de atracamento utilizam este tipo de representação para os cálculos de energia (Morris et al., 1998; Rarey et al., 1996; Friesner et al., 2004; Verdonk et al., 2003; Ewing et al., 2002). A idéia básica nesta metodologia é, considerando o receptor rígido, armazenar a informação acerca das contribuições energéticas do receptor em pontos da grade que só serão necessárias durante a avaliação do ligante. Nas formas mais básicas, os pontos da grade armazenam dois tipos de potenciais: eletrostático e de Lennard Jones (*i.e.* van der Waals mais o de repulsão interatômica)

(Kitchen et al., 2004).

1.2. Interações Receptor-Ligante

Um grande número de interações inter- e intramoleculares estão envolvidos no processo de reconhecimento molecular receptor-ligante. Como postulado por Pauling e Delbruck (1940), as interações de van der Waals, interações eletrostáticas e ligações hidrogênio, são importantes para a estabilização dos complexos biomoleculares (Nakamura, 1996). No entanto, é importante citarmos ainda o efeito hidrofóbico, efeitos entrópicos, interações do tipo cátion-π envolvendo grupamentos positivamente carregados e anéis aromáticos; interações do tipo empilhamento π e empilhamento-T entre grupamentos aromáticos e interações envolvendo metais.

A energia interna das moléculas também tem que ser levada em conta. A energia interna do receptor e/ou ligante pode ser maior no complexo do que é nos estados não complexados, e esta "energia de deformação" desestabiliza os complexos (Brooijmans e Kuntz, 2003).

1.2.1. Potencial Eletrostático

As forças eletrostáticas são de longo alcance e dependem das cargas elétricas dos átomos. A energia potencial de uma interação eletrostática é dada por:

$$V(r) = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{Dr} \tag{1.1}$$

onde V é a energia potencial eletrostática, q_1 e q_2 são as cargas parciais atômicas dos dois átomos (em unidades de carga eletrônica), r é a distância entre os dois átomos (em Ångstroms), D é a constante dielétrica relativa do meio, que reflete a tendência do meio em blindar uma carga da

outra, e ε_0 = 8,854238837 x 10^{-12} C^2/Nm^2 é a constante de permissividade do vácuo (Berg, Tymoczko e Stryer, 2004).

1.2.2. Potencial de Lennard-Jones

Átomos e moléculas neutras estão sujeitas a duas forças distintas: uma força atrativa (*i.e.*, força de dispersão de London), que tem energias entre 0,2 e 1,0 kcal/mol e uma força de repulsão de curto alcance, que resulta da repulsão entre os núcleos atômicos e da repulsão entre as densidades eletrônicas devido ao princípio de repulsão de Pauli. O potencial de Lennard-Jones (LJ), proposto em 1931 por John Lennard-Jones é um dos modelos matemáticos que representa este comportamento.

O potencial LJ entre dois átomos i e j é dado por:

$$LJ = 4\varepsilon\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \qquad (1.2)$$

onde ε é a profundidade do poço de potencial e σ é uma constante com a dimensão do comprimento. Estes parâmetros podem ser ajustados para reproduzirem dados experimentais ou deduzidos de resultados acurados obtidos a partir de cálculos de química quântica. O termo com dependência $(1/r)^{12}$ descreve a repulsão e o termo com dependência $(1/r)^6$ descreve a atração entre os dois átomos.

O potencial LJ é aproximado. A forma do termo atrativo de van der Waals possui justificativa teórica, enquanto que a forma do termo de repulsão não possui. Na realidade, a força de repulsão depende exponencialmente da distância, mas o termo de repulsão da fórmula de LJ é mais conveniente por facilitar a eficiência de computar r^{12} como o quadrado de r^6 (Chiquito e de Almeida,

1999).

1.2.3. Ligações hidrogênio

Uma ligação hidrogênio é uma interação de origem essencialmente eletrostática e ocorre quando dois átomos eletronegativos (*e.g.,* nitrogênio, oxigênio) interagem com o mesmo hidrogênio. O átomo eletronegativo (densidade de carga negativa) ligado covalentemente ao hidrogênio (densidade de carga positiva) é chamado de doador e o outro átomo eletronegativo (densidade de carga negativa) que interage com o mesmo hidrogênio é chamado de aceptor.

As ligações hidrogênio, apesar de mais fortes que as interações de van der Waals, são muito mais fracas do que as ligações covalentes. Possuem energias entre 3 e 9 kcal mol^{-1} (aproximadamente 10 a 40 kJ mol^{-1}), em comparação a 100 kcal mol^{-1} (418 kJ mol^{-1}) para uma ligação covalente carbono-hidrogênio. Elas são também mais longas que as covalentes. Seus comprimentos (medidos a partir do átomo de hidrogênio até o átomo aceptor) vão de 1,5 a 2,6 Å, portanto, distâncias que vão de 2,4 a 3,5 Å separam os dois átomos não-hidrogênio em uma ligação. Possuem também direcionalidade, sendo que as mais fortes tendem a formar ângulos de 180º, de tal modo que o átomo de hidrogênio, o átomo doador e o aceptor se disponham ao longo de uma linha reta (Berg e Tymoczko, 2004), no entanto em sistemas com ligações hidrogênio intramoleculares, esse ângulo pode ser bem diferente.

1.2.4. Efeito Hidrofóbico

Efeitos entrópicos como, por exemplo, o efeito hidrofóbico, também contribuem para a formação do complexo receptor-ligante. O efeito hidrofóbico é definido como a preferência de moléculas apolares por ambientes não aquosos, minimizando seus contatos com moléculas de água, sendo este um dos principais fatores de estabilização de proteínas, ácidos nucléicos e membranas

lipídicas. O efeito hidrofóbico está intimamente ligado à incapacidade das moléculas apolares de formarem ligações hidrogênio com moléculas de água e não devido a possíveis forças de repulsão. O efeito hidrofóbico resulta em uma tendência das moléculas apolares de formarem agregados de tal forma a minimizar o custo entrópico associado à formação de cavidades de moléculas de água em torno de solutos apolares.

A aproximação das superfícies apolares durante a interação ligante-receptor, libera e desorganiza as moléculas de água, aumentando a entropia do sistema e favorecendo a formação do complexo.

De forma geral, o processo de reconhecimento molecular receptor-ligante é dirigido por uma combinação de efeitos entálpicos e entrópicos. Estes efeitos podem ser estimados através da energia livre de Gibbs (de Magalhães, 2006).

1.2.5. Efeito Entrópicos Conformacionais, Rotacionais e Translacionais

Além do aumento na entropia do solvente, gerada como conseqüência do efeito hidrofóbico durante a interação ligante com a proteína, existem também mudanças na entropia dos solutos. Primeiro, existem perdas na entropia translacional e rotacional porque estas duas moléculas se tornarão um complexo. Segundo, existem mudanças na entropia vibracional por causa também da formação do complexo. Terceiro, existem mudanças na entropia conformacional por causa da restrição dos ângulos diedrais com a formação do complexo.

1.3. Campos de Força Clássicos

Muitos dos problemas que gostaríamos de resolver na modelagem molecular envolvem um

número extremamente grande de átomos para ser considerado em cálculos quânticos de estrutura eletrônica, cujos métodos tratam explicitamente os elétrons do sistema. Dessa forma, se alguns dos elétrons são ignorados (como em esquemas semi-empíricos), um grande número de partículas pode ser considerado e os cálculos têm seu tempo reduzido.

Métodos que utilizam campos de força clássicos (também conhecidos como métodos de mecânica molecular) ignoram os movimentos dos elétrons e calculam a energia de um sistema como uma função apenas das posições dos núcleos. Mecânica molecular, portanto, é normalmente utilizada na execução de cálculos em sistemas contendo um grande número de átomos. Em alguns casos, campos de forças podem fornecer respostas que são tão acuradas quanto às obtidas realizando cálculos de alto nível de mecânica quântica, em um tempo computacional bastante reduzido. Entretanto, a mecânica molecular não pode, obviamente, determinar propriedades que dependam da distribuição eletrônica em uma molécula.

Os trabalhos de mecânica molecular são todos apoiados em algumas hipóteses. A primeira delas é a da aproximação de Born-Oppenheimer, que permite resolver a equação de Schrödinger para o movimento eletrônico em relação aos núcleos de uma molécula, considerando estes últimos fixos. Deste modo, ela permite que a equação de Schrödinger para o movimento nuclear seja resolvida considerando-se a influência eletrônica como um potencial dependente das posições nucleares. Outra aproximação importante associada ao uso da mecânica/dinâmica molecular é que os movimentos dos átomos não são mais regidos pela equação de Schrödinger da mecânica quântica, mas sim pela segunda lei de Newton da mecânica clássica. Os campos de força clássicos são baseados em modelos simples de interações dentro de um sistema com contribuições do afastamento de ligações, a abertura e fechamento de ângulos e as rotações de ligações simples (Leach, 2001).

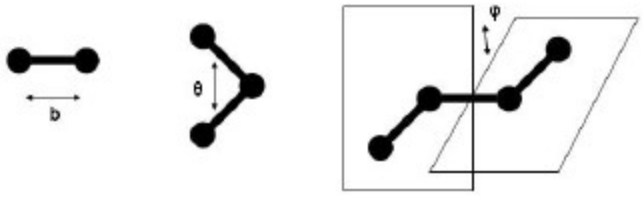
Muitos dos campos de força de modelagem molecular atualmente em uso podem ser interpretados em termos de uma função relativamente simples de forças intra e intermoleculares

dentro do sistema. Penalidades de energia estão associadas com o desvio das ligações e ângulos de seus valores de referência ou equilíbrio. Existe uma função que descreve como a energia varia quando as ligações são rotacionadas e, finalmente, o campo de força contém termos que descrevem as interações entre partes não-ligadas do sistema. Abaixo estão descritos os primeiros quatro termos (ligados) usualmente presentes em um campo de força clássico:

$$V_{ligados}(r_1,\ldots,r_N)=\sum_{i=1}^{N_b}(\frac{1}{2})K_{bi}(b_n-b_{0i})^2+\sum_{i=1}^{N_\theta}(\frac{1}{2})K_{\theta i}(\theta_n-\theta_{0i})^2$$
$$\sum_{i=1}^{N_\xi}(\frac{1}{2})K_{\xi i}(\xi_n-\xi_{0i})^2+\sum_{i=1}^{N_\varphi}(\frac{1}{2})K_{\varphi i}[1+\cos(n_i\varphi_i-\delta_i)] \quad (1.3)$$

onde K_b é a constante harmônica de energia associada à ligação química entre dois átomos e b_0 a distância de equilíbrio da mesma; K_θ é a constante harmônica de energia associada ao ângulo entre duas ligações químicas e θ_0 é o ângulo de equilíbrio; K_ξ é constante harmônica de energia associada ao ângulo formado entre dois planos (definidos por quatro átomos) e ξ_0 é o ângulo de equilíbrio entre estes dois planos. K_φ é uma constante de energia associada ao termo torcional (rotação de uma ligação química) entre dois planos definidos por quatro átomos, φ é o ângulo de torção entre estes dois planos, n a multiplicidade e δ é o ângulo de fase.

Figura 1.3: Termos de Energia do Campo de Força para Átomos Ligados.



Fonte: de Magalhães, 2006.

Temos que b é a distância de equilíbrio entre dois átomos, θ é o ângulo de equilíbrio e φ é o ângulo de torção entre estes dois planos.

Os termos de interação entre átomos não-ligados são expressos como potências do inverso da distância entre pares de átomos, sendo que, para interações intramoleculares, apenas as contribuições entre átomos separados por pelo menos três ligações químicas são consideradas:

$$V_{(não-ligados)}(r_1, \ldots, r_N) = \sum_{i<j}^{N} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i<j}^{N}\left[\frac{(q_i q_j)}{D r_{ij}}\right], \quad (1.4)$$

onde: r_{ij} é a distância entre os átomos i e j; ε_{ij} e σ_{ij} são os parâmetros de Lennard-Jones cujo valor depende dos tipos de átomos envolvidos; q_i e q_j são as cargas parciais localizadas nos respectivos átomos e D é a constante dielétrica relativa do meio. Os dois primeiros termos da Equação acima formam o potencial 6-12 de Lennard-Jones e estão associados a dois tipos de interações distintas.

Os campos de força comumente empregados em programas de atracamento molecular e mecânica/dinâmica molecular são o GROMOS (van Gunsteren e Berendsen, 1987), AMBER (Weiner et al., 1984; Cornell et al., 1995), CHARMM (Brooks et al., 1983) e MMFF94 (Halgren, 1996a,b,c,d,e).

1.4. Termodinâmica e Energia Livre

Segundo Rudolf Clausius (1865), a desordem dos componentes de um sistema químico é expressa como entropia, denotada por S. Qualquer mudança da desordem do sistema é expressa como variação da entropia (ΔS), a qual, por definição, tem um valor positivo quando a desordem aumenta. J. Willard Gibbs, que desenvolveu a teoria da variação de energia durante reações químicas, mostrou que o índice de energia livre, G, de qualquer sistema fechado pode ser definido

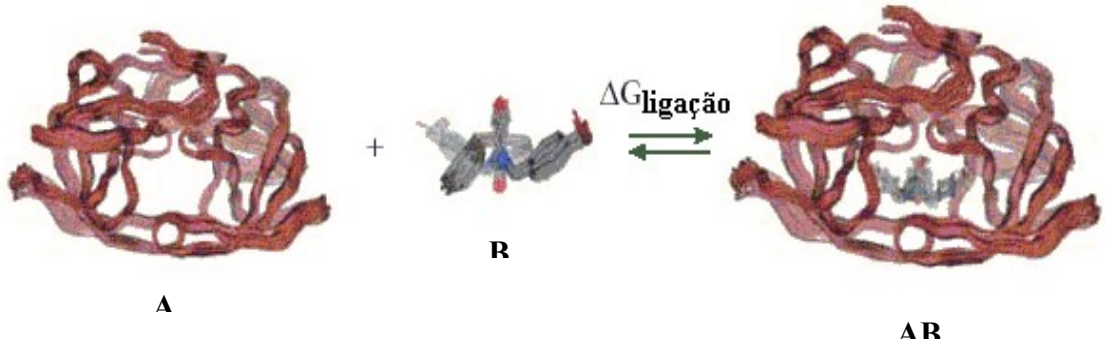
por três quantidades: entalpia H, refletindo o número e tipos das ligações; entropia, S; e a temperatura absoluta, T (em graus Kelvin). A definição da energia livre é $G = H - TS$.

A energia livre é usualmente expressa como a função de Helmholtz (A) ou a função de Gibbs (G). A energia livre de Helmholtz é apropriada para um sistema com número de partículas, temperatura e volume constantes. A energia livre de Gibbs é apropriada para sistemas onde o número de partículas, a temperatura e a pressão são constantes (Leach, 2001). As reações ocorridas em nosso organismo acontecem em condições de temperatura e pressão constantes, onde a função de Gibbs é a quantidade de energia livre apropriada.

A variação de energia livre de Gibbs, ΔG, é determinada pela mudança de entalpia, ΔH, refletindo os tipos e os números de ligações químicas e de interações não-covalentes quebradas e formadas, e pela variação da entropia, ΔS, descrevendo a mudança na desordem do sistema: $\Delta G = \Delta H - T\Delta S$.

A tendência de uma reação química ocorrer pode ser expressa como uma constante de equilíbrio. Para a reação mostrada na Figura 1.3.:

Figura 1.4: Processo de atracamento do ligante Dmp323 na proteína HIV-Protease (Código PDB: 1BVE)



Fonte: Kitchen et al. (2004).

$$aA + bB \leftrightarrows cAB \tag{1.5}$$

a constante de equilíbrio é dada por:

$$K_{eq} = \frac{[AB_{eq}]^c}{([A_{eq}]^a [B_{eq}]^b)} \quad (1.6)$$

onde $[A_{eq}]$ é a concentração de A, $[B_{eq}]$ a concentração de B, e assim por diante, quando o sistema alcança o equilíbrio. Para uma reação onde ocorre a interação de um ligante com um receptor, os reagentes são as duas moléculas livres e o produto é o complexo ligante-proteína. Um valor alto de K_{eq} significa que a reação tende a prosseguir até que os reagentes estejam quase que completamente convertidos em seus produtos.

Gibbs mostrou que ΔG para qualquer reação química é uma função da mudança padrão da energia livre, ΔG^0 - uma constante que é característica de cada reação - e um termo que expressa as concentrações iniciais dos reagentes e dos produtos:

$$\Delta G = \Delta G^0 + RTln[AB_i]^c / [A_i]^a [B_i]^b \quad (1.7)$$

onde $[A_i]$ é a concentração inicial de A, e de maneira equivalente para os outros elementos; R é a constante universal dos gases, 1,987 cal K^{-1} mol^{-1} e T é a temperatura absoluta, 298,15 K.

Quando uma reação alcança o equilíbrio, temos: $\Delta G = 0$. Para este caso especial, $[A_i] = [A_{eq}]$, e assim por diante, para todos os reagentes e produtos, e

$$[AB_i]^c / ([A_i]^a [B_i]^b) = [AB_{eq}]^c / ([A_{eq}]^a [B_{eq}]^b) = K_{eq} \quad (1.8)$$

Substituindo ΔG por 0 e K_{eq} por $[C_i]^c/[A_i]^a[B_i]^b$ na Equação (1.7), obtemos a relação

$$\Delta G^0 = - RTlnK_{eq} \quad (1.9)$$

que é definida pela reação de associação, E + I ⇆ EI, onde E é a enzima e I o inibidor. No entanto, como as constantes de inibição (K_i) estão relacionadas com o processo oposto, de dissociação entre o ligante e o receptor, EI ⇆ E + I, o cálculo do ΔG perde o sinal negativo passando a ser (Morris et al., 1998):

$$\Delta G^0 = RT\ln K_i \tag{1.10}$$

1.5. Cálculo das Diferenças de Energia Livre

A energia livre de ligação de Gibbs pode ser obtida através de métodos teóricos, embora a obtenção de estimativas mais precisas de ΔG envolva um custo computacional muitas vezes proibitivo. O objetivo do cálculo da energia livre de ligação é a obtenção de uma estimativa acurada das constantes de inibição observadas experimentalmente.

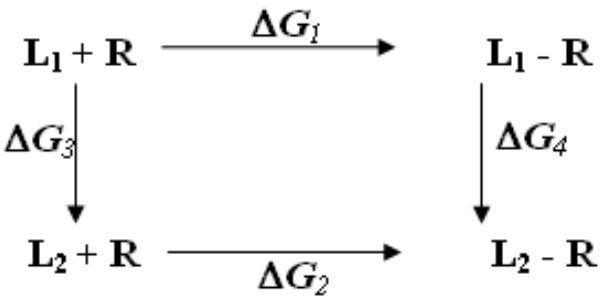
Três métodos, que utilizam simulações de dinâmica molecular utilizando um campo de força molecular clássico, são geralmente utilizados para o cálculo da energia livre: o método de perturbação da energia livre (PEL); o método de integração termodinâmica (IT) e o de crescimento lento. Embora esses métodos sejam precisos, com erros de aproximadamente 1 kcal/mol, o alto custo computacional envolvido limita a sua utilização. Além disso, esses métodos tendem a ter um pior desempenho quando os compostos envolvidos diferem de muitos átomos.

Para a predição da energia livre de ligação, Åqvist et al. (1994) desenvolveu o método de energia de interação linear (LIE - "*Linear Interaction Energy*"). O método LIE é um método semiempírico, baseado na teoria de resposta linear, o qual é menos custoso computacionalmente quando comparado aos métodos citados anteriormente.

1.5.1. Ciclos Termodinâmicos

Considerando a ligação de dois diferentes ligantes (L_1 e L_2) na molécula receptora (R), L_1 e L_2 poderiam ser possíveis inibidores de uma enzima R ou dois "hóspedes" para um hospedeiro R. O ciclo termodinâmico para os dois processos de ligação é mostrado na Figura 1.4 e a afinidade de ligação de L_1 e L_2 é igual a $\Delta G_2 - \Delta G_1$ e é comumente escrito como $\Delta\Delta G$. Em princípio, seria possível calcular valores de $\Delta\Delta G_1$ e $\Delta\Delta G_2$ pela simulação do processo de associação real. Para fazer isto, seria necessário tratar o ligante e o receptor juntos com uma separação grande inicial para gradualmente formar o complexo intermolecular. Entretanto, em muitos casos isto envolveria uma maior reorganização do receptor, do ligante e do solvente e seria difícil de assegurar uma amostragem adequada do espaço de fase.

Figura 1.5: Ciclo Termodinâmico



Fonte: Leach, 2001.

Para calcular a energia livre de ligação de dois ligantes, estes deverão ser trasformados de L_1 em L_2 em solução (L1 – R e L2 – R) e L_1 para L_2 interagindo com o receptor R (L1 + R e L2 + R).

A energia livre é uma função de estado e, portanto, este valor em um ciclo termodinâmico deve ser igual à zero. Assim, $\Delta G_2 - \Delta G_1 = \Delta G_3 - \Delta G_4$ (Figura 1.4). A diferença entre as afinidades de ligação dos dois ligantes em solução é igual ao ΔG_4 e ΔG_3 é comumente descrito para o

complexo receptor-ligante. As variações de ΔG_3 e ΔG_4 não correspondem a nenhuma transformação que pode ser realizada em laboratório, mas elas são completamente possíveis no computador. A diferença de energia livre só depende dos pontos finais e, portanto, podem ser realizadas mudanças nas Hamiltonianas da forma desejada.

As diferenças de energia livre obtidas de formas não-físicas normalmente são mais confiáveis que os processos "plausíveis fisicamente" já que elas envolvem uma menor reorganização do sistema. Isto acontece particularmente se os dois ligantes L_1 e L_2 têm estruturas similares. Para calcular a energia livre de ligação entre dois ligantes devemos transformar L_1 para L_2 em solução e L_1 para L_2 dentro do receptor. Na prática este raciocínio permite que se calculem as variações de energia livre relativas entre dois ligantes diferentes em relação a um mesmo alvo molecular.

1.5.2. Método de perturbação da energia livre (PEL)

Se focalizarmos nas diferenças de energia livre entre os sistemas relacionados A e B ($\Delta G = \Delta G_B - \Delta G_A$), representados pelas Hamiltonianas, H_A e H_B, esta diferença pode ser dada pela Equação:

$$\Delta G = -RT \ln \left\langle e^{\left(\frac{-\Delta H}{RT}\right)} \right\rangle_c \qquad (1.11)$$

onde $\Delta H = H_B - H_A$ e $\langle\rangle_C$ se refere a um conjunto médio sobre um sistema representado pela Hamiltoniana, H_C. A Equação 1.11 é a equação fundamental dos cálculos de perturbação de energia livre. Nós podemos, entretanto, generalizar o problema e descrever a Hamiltoniana $H(\lambda)$, como na Equação 1.12.

$$H(\lambda) = \lambda H_B + (1 - \lambda) H_A \qquad (1.12)$$

com variação de 0 (H = H_A) a 1 (H = H_B). Nós podemos generalizar a Equação 1.11 como:

$$\Delta G = G_B - G_A = \sum_{\lambda=0}^{1} (-RTln\langle e^{(-\Delta H'/RT)}\rangle_\lambda) \tag{1.13}$$

onde $\Delta H' = H_{\lambda+d\lambda}$ - H_λ. O cálculo de energia livre é dividido em janelas, cada uma envolvendo um intervalo suficientemente pequeno em λ, para que seja permitido o cálculo acurado da energia livre (Kollman, 1993).

Na prática, podemos supor que estamos usando uma função empírica de energia como a seguinte, para descrever as interações inter e intramoleculares em nosso sistema A→B (*e.g.* etanol→etanotiol).

$$\begin{aligned} r^N = & \sum_{i-1}^{N_b} (\frac{1}{2}) K_{bi} (b_n - b_{0i})^2 + \sum_{i-1}^{N_\theta} (\frac{1}{2}) K_{\theta i} (\theta_n - \theta_{0i})^2 \\ & \sum_{i-1}^{N_\xi} (\frac{1}{2}) K_{\xi i} (\xi_n - \xi_{0i})^2 + \sum_{i-1}^{N_\varphi} (\frac{1}{2}) K_{\varphi i} (\varphi_n - \varphi_{0i})^2 \\ & + \sum_{i<j}^{N} (4\varepsilon_{ij}) [(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}] + \sum_{i<j}^{N} [\frac{(q_i q_j)}{D r_{ij}}] \end{aligned} \tag{1.14}$$

A relação entre os estados inicial, final e intermediário é descrito em termos de um parâmetro de acoplamento, λ. Como λ é mudado de 0 para 1, a Hamiltoniana varia de H_x para H_y. Cada um dos termos no campo de força para um estado intermediário λ, pode ser escrito como uma combinação linear dos valores para X e Y:

1. Ligações: $k_l(\lambda) = \lambda k_l(Y) + (1- \lambda)k_l(X)$ (1.15)

$$l_0(\lambda) = \lambda l_0 (Y) + (1- \lambda) l_0(X) \tag{1.16}$$

2. Ângulos: $k_\theta(\lambda) = \lambda k_\theta(Y) + (1- \lambda)k_\theta(X)$ (1.17)

$\theta_0(\lambda) = \lambda\theta_0(Y) + (1- \lambda)\theta_0(X)$ (1.18)

3. Diedrais: $v_\omega(\lambda) = \lambda v_\omega(Y) + (1- \lambda)v_\omega(X)$ (1.19)

4. Eletrostáticos: $q_i(\lambda) = \lambda\, q_i(Y) + (1- \lambda)q_i(X)$ (1.20)

5. van der Waals: $\varepsilon(\lambda) = \lambda\, \varepsilon(Y) + (1- \lambda)\varepsilon(X)$ (1.21)

$\sigma(\lambda) = \lambda\sigma(Y) + (1- \lambda)\sigma(X)$ (1.22)

Para cada valor de $\lambda(\lambda_i)$, uma simulação é realizada (usando Monte Carlo ou dinâmica molecular) com os parâmetros de campo de força apropriados. Primeiro, o sistema é equilibrado usando os parâmetros de campo de força apropriados para λ_i. Uma fase de produção é então executada durante a qual a diferença de energia livre ΔG ($\lambda_i \rightarrow \lambda_{i+1}$) é acumulada como $k_B T \ln\langle\exp(-\Delta H_i / (k_B T))\rangle$, onde $\Delta H_i = \Delta H_{i+1} - H_i$. A energia livre total que muda de $\lambda = 0$ para $\lambda = 1$ é então a soma das variações de energia livre para os vários valores de λ (Leach et al., 2004).

1.5.3. Método de Energia de Interação Linear (LIE)

Outra metodologia bastante utilizada para a obtenção de energias livres de ligação é a chamada Energia de Interação Linear (LIE, "*Linear Interaction Energy*"). Os cálculos de energia livre com esta metodologia envolvem simulações de dinâmica molecular somente nos estados livre (ligante em solução) e ligado (complexo receptor-ligante), podendo reduzir os problemas de convergência e custos computacionais associados às técnicas descritas anteriormente. A idéia principal é considerar as contribuições polares e não polares separadamente. A parte polar ou eletrostática pode ser tratada usando a aproximação de resposta linear, enquanto que a não-polar é calculada usando uma fórmula empírica calibrada sobre um conjunto de dados experimentais

$$\Delta G_{lig} = \alpha \, (<V^{LJ}>_{ligado} - <V^{LJ}>_{livre}) + \beta \, (<V^{el}>_{ligado} - <V^{el}>_{livre}) \qquad (1.23)$$

onde α é o fator empírico que surge das interações não-polares e β é o correspondente às interações eletrostáticas. <V> representa os valores médios da energia de interação entre o ligante e o meio circundante, tanto para o termo eletrostático (EL) como para o de Lennard-Jones (LJ) (Åqvist e Marelius, 2001).

1.6. Funções de Avaliação ("*Scoring*")

A avaliação e o "ranqueamento" de conformações preditas do ligante são um aspecto crucial da seleção virtual baseada em estrutura. Mesmo quando as conformações são corretamente previstas, os cálculos não farão realmente efeito se não diferenciarem posições corretas das incorretas e se os ligantes ativos não puderem ser identificados. A construção de funções de avaliação confiáveis e rápidas o suficiente para avaliarem centenas ou milhares de ligantes em poucos minutos é de importância fundamental. Este fato tem levado ao desenvolvimento de um grande número de funções que fazem uso de aproximações para a avaliação da constante de afinidade de um complexo receptor-ligante.

A maioria dos programas de atracamento molecular utiliza modelos simples de funções de energia potencial, geralmente durante a fase de execução e, posteriormente, avaliam as conformações obtidas com funções de avaliação mais sofisticadas, com inclusão de alguns termos entrópicos e de solvaltação e desolvatação. Essencialmente, quatro tipos ou classes de funções de avaliação são aplicados atualmente: (i) empíricas, (ii) baseadas em campo de força, (iii) funções de avaliação baseadas em conhecimento, (iv) consenso. A descrição destas classes de funções de avaliação foi feita com base no trabalho de Kitchen et al. (2004).

1.6.1. Funções de avaliação empíricas

As funções de avaliação empíricas são construídas para reproduzir dados experimentais. O desenho de funções de avaliação empíricas é baseado na idéia de que energias de ligação podem ser aproximadas pelo somatório de diferentes variáveis não correlacionadas. Os coeficientes de vários termos são obtidos por análises de regressão usando energias livres de ligação (ou constantes de afinidade receptor-ligante) e a partir das correspondentes informações estruturais, ambas obtidas experimentalmente.

Uma desvantagem destes métodos é a sua dependência dos conjuntos de dados moleculares usados na análise de regressão e ajuste. O tamanho e a qualidade do conjunto treinamento afetará a forma final da função avaliação (Wang et al., 2002). Deve-se ter cuidado na seleção dos complexos, buscando um conjunto que tente abranger as diferentes variações observadas na natureza, tanto para o ligante, quanto para o receptor. Caso o intuito seja a criação de uma função avaliação para uma determinada família de proteínas, as diferenças do sítio ativo destas serão menores do que as observadas em proteínas de diferentes famílias. Neste sentido, é de se esperar que funções empíricas construídas para uma família específica de receptores tenham uma maior capacidade de predição. Entretanto, termos de funções empíricas ajustados diferentemente não podem ser facilmente combinados em uma nova função empírica (Wang et al., 2002).

Muitas das funções empíricas de energia livre geralmente são uma modificação de uma "equação maior" para modelar a energia livre de ligação, adicionando termos entrópicos para equações de mecânica molecular, por exemplo:

$$\Delta G = \Delta G_{L\text{-}J} + \Delta G_{hbond} + \Delta G_{elet} + \Delta G_{conform} + \Delta G_{tor} + \Delta G_{sol} \qquad (1.24)$$

onde os primeiros quatro termos são termos típicos de mecânica molecular para dispersão/repulsão,

ligações hidrogênio, eletrostáticos e penalidades de energia conformacional, respectivamente; ΔG_{tor} modela o termo entrópico associado às restrições conformacionais (*i.e.,* restrições dos ângulos de torção) impostas ao ligante no processo de ligação; e ΔG_{sol} modela a desolvatação durante o processo de ligação e o efeito hidrofóbico (a entropia do solvente que varia nas interfaces soluto-solvente) (Ajay e Murcko, 1995).

Descrevemos a seguir os tratamentos geralmente utilizados para alguns dos termos descritos na equação 1.24.

Na função empírica, a formação de ligações hidrogênio é um dos fatores mais importantes para a ligação específica de um ligante a seu receptor. Entretanto, átomos de hidrogênio normalmente não são revelados em análises de cristalografia de difração de raios-X, podendo ser adicionados para então ser submetidos à minimização de energia para encontrar suas possíveis posições originais. Esta prática torna-se problemática especialmente quando os átomos de hidrogênio apresentam posições múltiplas de menor energia e dependendo do pK_a local, o tratamento deste tipo de interação varia enormemente entre as diferentes funções de avaliação empírica (Wang et al., 2003)

Na função empírica do programa AutoDock (Morris et al., 1998) a energia relacionada a esse tipo de interação é dependente de um fator de peso angular. Na função empírica PLP (Gehlhaar et al., 1995) utilizada no programa Cerius, as interações entre átomos doadores e aceptores são tratadas como ligações hidrogênio e dependem do ângulo e da distância entre eles. Já na função LUDI (Böhm, 1994) implementada no programa Cerius2, ligações hidrogênio do tipo "neutra" e "iônica" são tratadas separadamente e sofrem penalidades em relação à distância e ângulos das conformações ideais, o que também é observado na função F-Score (Rarey et al., 1996) utilizada no programa SYBYL, onde também foi implementada a função ChemScore (Eldridge et al., 1997), que, por sua vez, não trata as ligações hidrogênio como "neutras" e "iônicas", assim como a função empírica do X-Score (Wang et al., 2002).

Durante o atracamento, tanto o ligante quanto a proteína apresentam restrições nas suas conformações quando comparadas com seus estados livres no solvente. Estas mudanças entrópicas adversas devem ser superadas durante o processo de ligação. Normalmente nas funções de avaliação empíricas, o efeito de deformação do ligante é frequentemente estimado pela quantificação do número de ligações rotacionáveis que se tornam "congeladas" durante o processo de ligação, assumindo que cada uma destas ligações está associada com um número discreto de conformações de menor energia e, portanto, uma quantidade aproximada da entropia conformacional. Se existe mais que uma ligação rotacionável no ligante, suas contribuições são usualmente tratadas como aditivas (Wang, et al., 2002). No programa AutoDock, este termo é proporcional ao número de ligações "congeladas" do tipo sp^3 no ligante (Morris et al., 1998), enquanto que na função ChemScore a contribuição de cada uma delas é dependente de uma complicada função que reflete a natureza química de seu ambiente (Wang et al., 2003).

A ligação do ligante na proteína é acompanhada pelo processo de desolvatação que modifica tanto a entropia como a entalpia do sistema. Um dos resultados é que grupos não-polares tendem a se aproximar em um processo conhecido como efeito hidrofóbico, anteriormente descrito. Este efeito é de difícil caracterização, porque envolve interações entre a água e o ligante; proteína e a água e entre moléculas de água antes e após a ligação. Diferentes metodologias, bastante simplificadas e aproximadas, são utilizadas para a determinação deste termo para as diferentes funções de avaliação empíricas.

Uma aproximação utilizada é considerar a contribuição de energia associada ao efeito hidrofóbico proporcional à superfície hidrofóbica "enterrada" do ligante, que pode ser calculado pela Equação 1.25, onde ocorre o somatório da superfície acessível ao solvente (SAS_i) de todos os átomos do ligante. Esta abordagem foi adotada pela função avaliação de Böhm (Böhm, 1994). É importante lembrar que tecnicamente existem diversos tipos de superfícies moleculares, no caso da função empírica do X-Score (Wang et al., 2002), também é usada a superfície acessível ao solvente.

$$HS = \sum_{i=1}^{N} SAS_i \tag{1.25}$$

onde N é o número de átomos do ligante e SASi é a superfície enterrada de átomos hidrofóbicos. Qualquer parte da superfície do ligante é considerada "enterrada" se esta penetra na superfície acessível ao solvente da proteína. Isto é feito para todos os átomos do ligante, como mostrado na Equação 1.25, levando em consideração apenas átomos hidrofóbicos (Wang et al., 2002). Esta mesma metodologia foi adotada na função empírica LUDI, implementada no programa Cerius2 (Wang et al., 2003).

Na função empírica ChemScore, o efeito hidrofóbico é calculado pela soma dos pares de átomos hidrofóbicos formados entre o ligante e a proteína. A função distância implementada reflete a intuição de que o tamanho da "interação hidrofóbica" irá chegar ao máximo quando dois átomos hidrofóbicos formam contanto de van der Waals e diminuem gradualmente com o aumento da distância inter-atômica.

Existe ainda uma outra metodologia, adotada pela função empírica SCORE (Wang et al., 1998). De acordo com este método, partes diferentes do ligante "sentem" a proteína de forma diferenciada por causa da natureza heterogênea do sítio de ligação. Se um átomo hidrofóbico do ligante é colocado em um sítio hidrofóbico da proteína, então isto favorece o processo de ligação. O cálculo total do efeito hidrofóbico entre o ligante e a proteína é calculado como:

$$HM = \sum_{i}^{ligante} logP_i \times HM_i \tag{1.26}$$

Onde HM_i é uma variável indicadora, isto é, ela recebe o valor 1 se o átomo hidrofóbico i é colocado em um ambiente hidrofóbico, caso contrário, recebe valor 0. Log P_i se refere a escala de

hidrofobicidade do átomo *i*, o qual é a contribuição do átomo *i* no coeficiente de partição n-octanol/ água (Log *P*) da molécula. O ambiente de um dado átomo do ligante é definido por todos os átomos na proteína que estão dentro de uma esfera de 6 Å (centrada no átomo do ligante). A hidrofobicidade do ambiente é determinada pelo somatório das escalas hidrofóbicas de todos os átomos.

A função empírica implementada no programa FlexX (Rarey et al., 1996), além de considerar alguns dos termos já descritos, como interação entre átomos polares e o número de ligações rotacionáveis congeladas, considera ainda interações do tipo ambíguo, formadas entre átomos polares e não-polares e uma variável com a soma para diferentes contribuições energéticas, como ligação hidrogênio, contato com metal e interação especificamente entre grupos aromáticos, todos multiplicados por duas funções de penalidades lineares para desvios de distância e ângulos de valores ideais pré-definidos (Stahl e Rarey, 2001).

Uma nova função empírica para estimar as afinidades de ligação foi desenvolvida e implementada no programa Glide 4.0 XP (Friesner et al., 2004) e apresenta modificações no tratamento de algumas das interações descritas. São aplicadas grandes penalidades de desolvatação para átomos polares da proteína e do ligante e também para certos grupos carregados. Esta função empírica também efetua a identificação de motivos estruturais específicos (e.g., bolsões hidrofóbicos) que fornecem grandes contribuições para o aumento da afinidade de ligação (Friesner et al., 2006).

1.6.2. Funções de Avaliação baseadas em campo de força

Existem diversos métodos estabelecidos para estudos de mecânica e dinâmica molecular,

como AMBER, CHARMM, DISCOVER, ECEPP e GROMOS. Muitos desses tradicionais campos de força modelam a energia de interação de um sistema molecular com termos para dispersão/repulsão, ligações hidrogênio, eletrostática e penalidades para afastamentos das distâncias de ligação e de ângulos ideais. Estes métodos são excelentes para o estudo de processos moleculares, usualmente considerando o solvente explicitamente, como na otimização de conformações e no cálculo de energia livre no método de perturbação que, como explicado anteriormente, é realizado entre moléculas com a mudança de poucos átomos. Normalmente estas metodologias requerem consideráveis investimentos de tempo computacional e, infelizmente, tendem a ser menos eficientes no "ranqueamento" das energias livres de ligação de compostos que diferem em muitos átomos (Morris et al., 1998).

Várias funções de avaliação empíricas são baseadas em diferentes conjuntos de parâmetros de campo de força. Por exemplo, G-Score é baseado no campo de força Tripos e AutoDock no AMBER.

As funções baseadas em campos de força são mais complicadas de se tratar pelo fato de que geralmente requerem a introdução de distâncias de corte para o tratamento de interações não-ligadas, as quais são escolhidas mais ou menos arbitrariamente, complicando o tratamento acurado dos efeitos de longa distância envolvidos na ligação (Kitchen et al., 2004).

1.6.3. Funções de avaliação baseadas em conhecimento

Ao invés de reproduzir a energia de ligação e outros parâmetros, funções avaliação baseadas em conhecimento tentam obter através de uma análise estatística das geometrias de interação átomo-átomo, "pseudo-potenciais" relativamente simples, que descrevam as geometrias preferenciais de interação receptor-ligante. Estas funções são modeladas utilizando potenciais de interação átomo a átomo, e devido à simplicidade de sua implementação, essa metodologia é

especialmente rápida (de Magalhães, Dardenne e Barbosa, 2007).

As metodologias baseadas em conhecimento tendem a capturar efeitos de ligação específicos que são difíceis de modelar explicitamente e podem ser facilmente encontrados com a varredura eficiente de um grande conjunto de dados de compostos. Entretanto, devido às limitações dos conjuntos de estruturas disponíveis, estas funções podem sofrer de sérias limitações.

Implementações populares deste tipo de função incluem o Potencial de Força Média (PMF, "*Potential Mean Force*") (Muegge, 2000) e DrugScore (Sotriffer et al., 2002), as quais também incluem um potencial com correções de acessibilidade ao solvente. O SMoG (DeWitte e Shakhnovich, 1996) é outra função avaliação encontrada nesta classe e que utiliza potenciais de pares atômicos para calcular as interações proteína-ligante (Kitchen et al., 2004).

1.6.4. Funções de avaliação consenso

Os erros e imperfeições das funções de avaliação atuais têm levado à utilização de uma combinação de diferentes tipos de função avaliação, para gerar uma única função: função consenso. X-SCORE (Wang et al., 2002) é um exemplo de função consenso que combina as funções de avaliações do programa GOLD (Verdonk et al., 2003), DOCK (Ewing et al., 2002), FlexX (Rarey et al., 1996), além de PMF (Muegge, 2000) e ChemScore (Eldridge et al., 1997). Resultados da literatura mostram que o uso de funções consenso pode aumentar significativamente a taxa de sucesso na identificação da conformação do ligante observada experimentalmente. Entretanto, a utilização de funções consenso deve ser realizada com cuidado, para que haja um balanceamento dos erros associados às diferentes funções de avaliação e não uma amplificação dos mesmos.

1.6.5. Funções de avaliação baseadas em reconhecimento de padrões

Novos métodos têm sido criados para a estimativa do valor de energia livre de ligação do

complexo proteína-ligante, alguns deles têm utilizado metodologias de reconhecimento de padrão, como redes neurais, ou máquina de vetor de suporte (SVM, "*Support Vector Machine*"), onde um conjunto de dados de entrada, calculados utilizando dados de estruturas determinadas experimentalmente, é utilizado para a determinação de um padrão de saída associado a valores experimentais conhecidos. Desta forma, a partir de valores de entrada extraídos de complexos proteínas-ligante, são determinados os valores correspondentes de energia livre de ligação, a partir deste padrão pré-estabelecido.

Bock e Gough (2002) desenvolveram um novo método de estimar valores de energias livre entre ligante-receptor. Cada complexo receptor-ligante foi transformado em um vetor de características numéricas que eram tidas como importantes para estimar os valores de energias entre ligante-receptor. Os vetores de características do receptor e do ligante construídos eram concatenados e classificados com o valor de sua energia livre de ligação. Estes vetores são submetidos a um treinamento a partir da regressão da máquina de vetor de suporte e testados por validação por cruzamento para calcular de forma precisa o valor de energia. Os resultados foram comparados com outros 9 trabalhos com o mesmo objetivo, utilizando dados de estruturas tri-dimensionais como entrada e concluíram que seu método era o quarto melhor em relação aos outros, quando avaliado o erro médio quadrado, apesar de um erro médio quadrático de 35,106 kcal/mol, extremamente alto comparado ao segundo maior, 6,683 e justificado pelo alto valor médio da amostra e seu grande tamanho.

Em outros trabalhos, estas metodologias de reconhecimento de padrão foram utilizadas não para a determinação do valor de energia de ligação, mas para a classificação de fármacos que poderiam ser utilizados para determinadas situações. No trabalho de Zernov e colaboradores (2003), foi provado que ao contrário do que rege na área de inteligência artificial, o SVM nem sempre retorna o melhor modelo, mas quando aplicado a uma situação específica, no caso de uma enzima em particular, este método fornece dados interessantes, como no caso de inibidores da enzima

anidrase carbônica II.

Byvatov et al. (2003), buscando o mesmo objetivo de classificar fármacos ideais para um quadro particular, avaliou além da metodologia de SVM, a rede neural artificial (ANN, "*Artificial Neural Network*"). Os resultados encontrados mostraram que o SVM apresentou-se mais robusto com um erro padrão menor comparado ao obtido utilizando uma rede neural. Geralmente, o classificador SVM apresenta um rendimento mais alto em relação a alta precisão nas previsões, independente dos tipos de descritores utilizados para as moléculas, o tamanho do conjunto de dados de treinamento e o algoritmo empregado para o treinamento da rede neural. No entanto, quando observamos seus resultados encontramos que para um conjunto de 525 descritores validados por cruzamento a classificação por SVM prediz corretamente 82% dos casos, enquanto a rede neural artificial 80%, o que não mostra ser uma mudança tão significativa.

1.6.6. Performance e limitações de funções empíricas

Um dos aspectos mais importantes na construção e na análise da performance de uma função de avaliação para o cálculo de uma constante de afinidade e de um programa de atracamento é a seleção do conjunto de proteínas utilizado para estas etapas.

No trabalho desenvolvido por Warren et al. (2006) observamos o quanto é laboriosa a seleção e a preparação das estruturas envolvidas na avaliação e na construção destas funções para o cálculo das constantes de afinidade e dos programas de atracamento molecular. No trabalho destes autores, as estruturas das proteínas foram selecionadas por químicos computacionais com vasta experiência em cada uma das proteínas-alvos e foram preparadas para os cálculos de atracamento por um único químico computacional. Depois de preparadas, as estruturas alvos foram passadas para químicos computacionais especialistas em cada um dos algoritmos de atracamento.

Para cada proteína-alvo, o especialista de cada proteína selecionava uma estrutura

representativa desta família de proteína para ser usada em todos os cálculos de atracamento. Para tanto, esse especialista levava em consideração não só a alta qualidade da estrutura com boa resolução, como a capacidade desta para acomodar todas as classes de compostos relevantes para a indústria farmacêutica. Os resíduos de aminoácidos do sítio de ligação e seus estados de ionização eram identificados automaticamente e eram modificados quando o especialista destas proteínas achava necessário. Estes mesmos especialistas também forneciam informações a respeito de moléculas de água consideradas importantes para a ligação dos compostos. Esta coleção de informações era passada para cada um dos especialistas dos diferentes programas de atracamento molecular para a configuração e cálculos de atracamento.

Ainda sobre a análise de funções de avaliação, Friesner et al. (2006) afirmam o quanto é crítica a seleção e preparação das proteínas e ligantes, ainda mais quando a construção destas funções é baseada na aplicação de variáveis físico-químicas destas estruturas. Quanto ao aspecto da preparação da proteína e do ligante, os autores citam o quanto é problemática a determinação dos estados de protonação do ligante e da proteína no sítio ativo da proteína, ponto levantado por Warren et al. (2006), que cita ainda a importância do relaxamento do sítio ativo da proteína de modo que este acomode pelo menos o ligante nativo.

Uma das principais conclusões encontradas por Warren et al. (2006) é a respeito da avaliação dos programas de atracamento, onde os autores observam que normalmente estes programas têm sucesso na geração de múltiplos modos que incluem ligações similares aos determinados cristalograficamente ligados a proteína. Enquanto os programas de atracamento obtiveram grande sucesso na reprodução de modos de ligação, as funções avaliação obtiveram pouco sucesso na identificação correta do modo de ligação.

Warren e seus colaboradores (2006) afirmam que, no entanto, o objetivo principal de sua avaliação foi quantificar a relação entre os valores de atracamento e a afinidade do composto sendo demonstrado que para as oito proteínas dos sete tipos de alvo estudados na avaliação, não foi

observada nenhuma relação estatisticamente relevante existente entre os valores de atracamento e a afinidade do ligante.

Este resultado não foi inesperado dado o grande número de aproximações utilizadas para os valores de atracamento em busca do aumento do desempenho computacional. Foi mostrado que quanto à predição do modo de ligação, os programas de atracamento poderiam reproduzir os modos de ligação dos ligantes observados experimentalmente, mas que não é observado nenhum dado consistente na correlação entre os valores de atracamento e a medida de afinidade. Portanto, o bom desempenho na reprodução dos modos de ligação determinados experimentalmente não tem nenhum impacto no sucesso da predição da afinidade ou no "ranqueamento" de compostos pela afinidade dentro de séries congenéricas (Warren et al., 2006).

A partir dos dados gerados pela avaliação, não fica claro o que causa as falhas da baixa predição da afinidade do ligante por valores de atracamento visto que o desempenho foi baixo para todos os tipos de alvo e para todas as funções avaliação testadas. A falha não ocorreu na reprodução do modo de ligação observado experimentalmente pelos programas de atracamento, mas na inabilidade das funções de avaliação atuais para distinguir, diferenciar e quantificar sutis diferenças que podem alterar a afinidade de um ligante de altamente ativo para inativo (Warren et al., 2006).

Desta forma, conclui-se que os programas de atracamento são capazes de reproduzir os modos de ligação encontrados experimentalmente e, em muitos casos, são capazes de determinar o melhor modo de ligação e distinguir os ligantes ativos dos inativos. No entanto, ainda não se desenvolveu uma função empírica capaz de determinar com exatidão os valores de afinidade de ligação e de ranquear os diferentes modos de ligação com base nas suas afinidades em relação as encontradas experimentalmente, mostrando a importância de estudos nessa área.

CAPÍTULO 2

2. REDES NEURAIS ARTIFICIAIS

As redes neurais artificiais são metodologias computacionais que executam análises multifatoriais que podem ser utilizadas para a predição de um valor de saída, na classificação de objetos, na aproximação de função e no reconhecimento de padrões de dados multifatoriais.

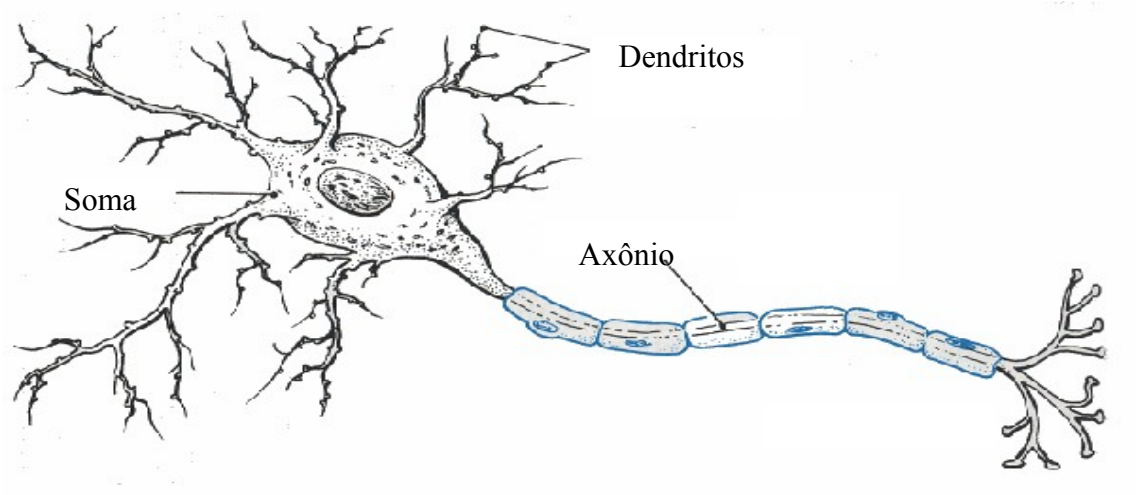
A técnica de redes neurais artificiais é um campo ativo de pesquisa que tem amadurecido grandemente nos últimos 40 anos. Os primeiros treinamentos de uma rede neural foram desenvolvidos em 1959 por Rosenblatt, bem como em 1960 por Widrow e Hoff. Atualmente as redes neurais têm sido utilizadas nas mais variadas áreas da ciência e tecnologia (Robert et al., 1998; Anderson et al., 1999; Choudhury, 1999; Paquet et al., 2000; Poulton, 2002; Bose, 2007; Gobel et al., 2007). Na medicina, esta metodologia tem seu uso na análise de imagens médicas em sistemas para detecção de tumores e sistemas de classificação de células normais e malignas em testes citológicos (Dayhoff e DeLeo, 2001).

Inspirados em redes neurais biológicas, os modelos de redes neurais artificiais são considerados como a coleção de unidades computacionais interconectadas chamadas neurônios artificiais. Assim como nas redes neurais biológicas, as artificiais podem ser definidas por diferentes fatores, como número e arranjo de seus neurônios, suas interconexões etc., simulando tanto sua arquitetura como sua função.

O neurônio biológico (Figura 2.1) consiste de três partes – o soma (o corpo celular com o núcleo); os dendritos (que conduzem os impulsos até o soma) e o axônio (que conduz o impulso a partir do soma). A conexão entre os neurônios é fornecida através de espaços chamados junções sinápticas. As sinapses são unidades estruturais e funcionais elementares que medeiam as interações entre os neurônios. O tipo mais comum de sinapse é a química, onde um processo pré-sináptico

libera uma substância transmissora que se difunde através da junção sináptica entre neurônios e então age sobre um processo pós-sináptico. Assim, uma sinapse converte um sinal pré-sináptico em um sinal químico e volta em um sinal elétrico pós-sináptico. Neste processo, os axônios funcionam como linhas de transmissão e os dendritos como zonas receptivas (Haykin, 2001).

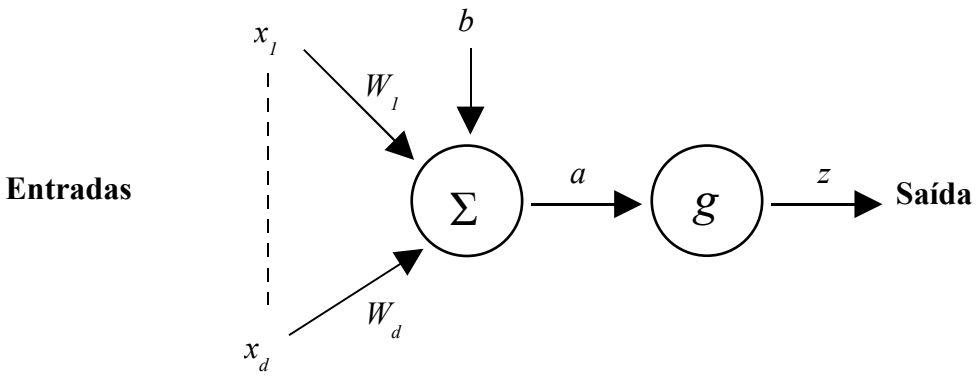
Figura 2.1: Estrutura de um neurônio biológico.



Fonte: Carlson, 1992.

O conceito de um neurônio artificial (Figura 2.2) surgiu a partir do neurônio biológico. Cada neurônio artificial apresenta um determinado número de entradas (x_R), sendo que cada uma delas apresenta um peso específico (w_R), que indica sua importância. No neurônio, a soma de entradas ponderadas (a) é calculada e processada usando uma função de ativação (g). O resultado desta função é distribuído através da saída (z) para o próximo neurônio artificial (Dohnal, 2005).

Figura 2.2: Modelo de um neurônio, que forma a base para o projeto de redes neurais (artificiais).



O modelo neural da Figura 2.2 inclui ainda um viés (do inglês, "*bias*") aplicado externamente, representado por b, que tem o efeito de aumentar ou diminuir a entrada das variáveis da função de ativação, dependendo de ser ele positivo ou negativo, respectivamente.

O processo de determinação dos pesos para a ponderação das entradas do neurônio é chamado de treinamento ou aprendizagem e pode ser uma tarefa computacionalmente custosa. No entanto, uma vez que os pesos são fixados, novos dados podem ser processados pela rede rapidamente.

Nós podemos fazer uma analogia de redes neurais com a técnica padrão de ajuste baseada em funções polinomiais. Um polinômio pode ser tratado como um mapeamento de uma única variável de entrada para uma única variável de saída. Os coeficientes em um polinômio são análogos aos pesos de uma rede neural e a determinação destes coeficientes (pela minimização do erro quadrático) corresponde ao processo de treinamento de uma rede neural (Bishop, 1994).

Além de oferecer uma alta velocidade de processamento, redes neurais têm uma grande capacidade de adaptar seus pesos sinápticos a modificações do meio ambiente. Em particular, uma rede neural treinada para operar em um ambiente específico pode ser facilmente "retreinada" para lidar com pequenas modificações nas condições operativas do ambiente (Haykin, 2001).

As principais desvantagens de uma rede neural são originadas da necessidade do fornecimento de um conjunto apropriado de dados de treinamento e potenciais problemas podem surgir se uma rede é requisitada para extrapolar novas regiões no espaço de entrada, as quais podem ser significativamente distintas daquelas cobertas pelos dados de treinamento utilizados (Bishop, 1994).

As vantagens e as limitações das redes neurais são freqüentemente complementares àquelas das técnicas de processamento de dados convencionais. Em geral, as redes neurais devem ser consideradas como possíveis candidatas para resolver problemas que possuem algumas, ou todas, destas características: (i) existe um amplo conjunto de dados para o treinamento da rede; (ii) é difícil fornecer uma solução adequada baseada em modelo ou por primeiros-princípios; (iii) novos dados devem ser processados em alta velocidade, ou por causa do grande volume de dados que devem ser analisados ou por causa de alguma restrição de tempo; (iv) os dados do método de processamento precisam ser robustos para níveis modestos de ruído nos dados de entrada (Bishop, 1994).

Um modelo matemático simples de um único neurônio foi introduzido por McCulloch e Pitts em 1943, e pode ser observado como uma função não-linear a qual transforma um conjunto de variáveis de entrada x_i, (i = 1, ... , d) em uma variável de saída z. No modelo de McCulloch-Pitts, o i-ésimo sinal x_i de entrada é primeiramente multiplicado por um parâmetro w_i, conhecido como peso, e é então adicionado aos outros sinais de entrada já "pesados" para resultar em uma entrada total para a unidade da seguinte forma:

$$a = \sum_{i=1}^{d} w_i x_i + b\,, \tag{2.1}$$

onde o parâmetro w_0 é chamado de viés. Formalmente, o viés pode ser tratado como um caso especial de peso de uma entrada extra, onde o valor x_0 é permanentemente ajustado por +1.

Portanto, nós podemos escrever a Equação acima na forma

$$a = \sum_{i=0}^{d} w_i x_i , \tag{2.2}$$

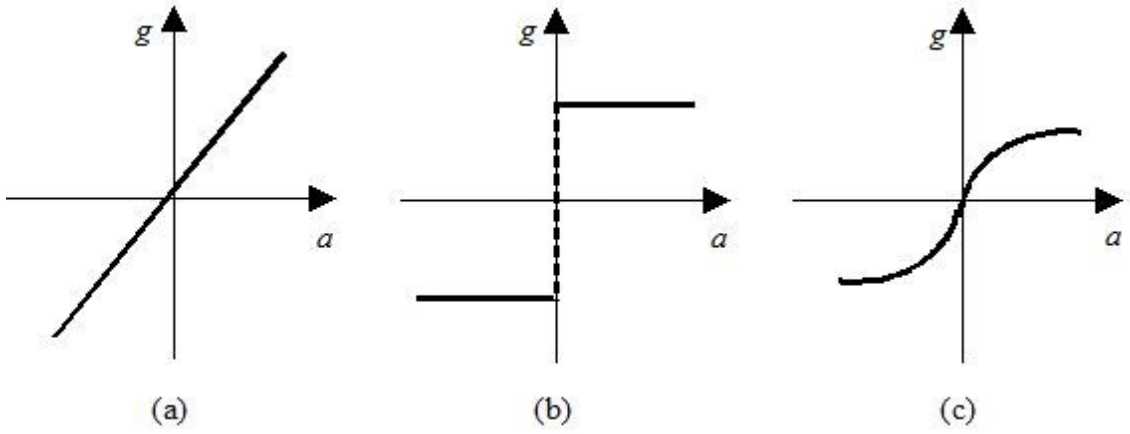
onde x_0 = 1. Note que os pesos (e o viés) podem ter qualquer sinal, correspondendo às sinapses excitatória ou inibitória. A saída da unidade z é resultado da aplicação de uma função ativação não linear g em a

$$z = g(a) \tag{2.3}$$

Algumas formas possíveis para a função g são mostradas a seguir:

Figura 2.3: Funções de ativação típicas: (a) linear; (b) por partes e (c) sigmoidal.

Fonte: Bishop, 1994.



O modelo de McCulloch-Pitts original usou a função mostrada na Figura 2.3 (a). A maioria das redes de interesse prático emprega função de ativação sigmoidal do tipo mostrado na Figura 2.3

(c).

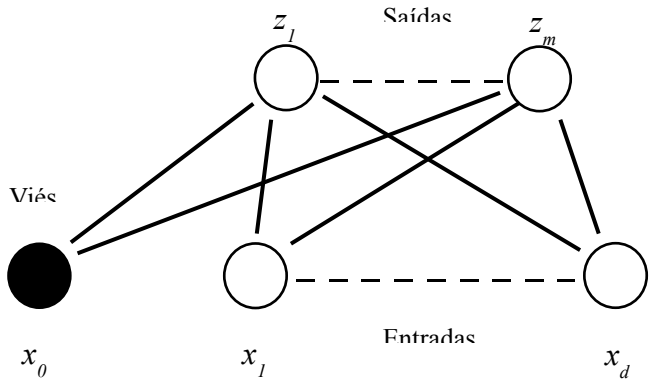
A função sigmoidal, normalmente denominada função sinal cujo gráfico tem a forma de um *s*, é definida como uma função estritamente crescente. Normalmente, as funções estão definidas no domínio [0,1] no entanto, algumas vezes é desejável que a função ativação esteja definida em [-1,1], neste caso, o gráfico da função é simétrica em relação à origem. Para obter o gráfico correspondente a uma função sigmóide podemos utilizar a função tangente hiperbólica (Haykin, 2001).

O modelo simples de um neurônio forma o elemento matemático básico em muitos modelos de redes neurais artificiais. Com a conexão destes vários elementos, é possível construir uma grande classe geral de mapeamento não-linear, a qual pode ser aplicada em uma larga escala de problemas práticos. A adaptação dos valores dos pesos, de acordo com o algoritmo de treinamento apropriado, pode permitir a aprendizagem da rede em resposta a dados externos.

2.1. Redes Neurais de Única Camada e de Multicamadas

Em uma rede neural os neurônios são organizados em camadas, onde cada camada possui um determinado número de neurônios. Na forma mais simples desse tipo de rede, temos uma camada de entrada de nós de fonte que se projeta sobre uma camada de saída de neurônios (nós computacionais). A rede mostrada na Figura 2.4 possui dois nós tanto na camada de entrada como na de saída e é chamada de rede de camada única, sendo que a designação "camada única" se refere à camada de saída de nós computacionais (neurônios). A camada de entrada de nós de fonte não é contada porque nesta camada não é realizado qualquer processamento (Haykin, 2001).

Figura 2.4: Rede neural com uma única camada



Fonte: Bishop, 1994.

As variáveis de saída são denotadas por z_j e são dadas por

$$z_j = g(\sum_{i=0}^{d} w_{ji} x_i), \tag{2.4}$$

onde w_{ij} é o peso da entrada i para a unidade j (*i.e.*, neurônio), e g é a função ativação. Novamente, os parâmetros viés são incluídos como casos especiais de pesos de uma entrada extra $x_0 = 1$.

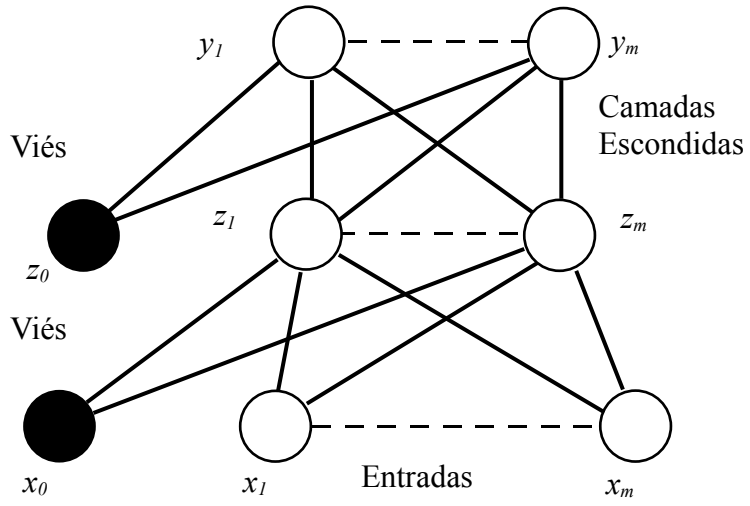
As redes neurais de camada única foram extensamente estudadas e usadas na década de 1960. No entanto, para a melhora de seu desempenho foram realizadas alterações tanto na utilização de outras funções de ativação como na adição de novas camadas à rede, surgindo as redes neurais multicamadas, que têm sido atualmente usadas como base na maioria das aplicações práticas de redes neurais.

É importante notar que utilizaremos aqui redes em que as saídas dos neurônios de uma dada camada se conectam apenas como entradas para os neurônios da camada seguinte, não havendo retro-alimentação. Em inglês diz-se que a rede é do tipo "feed-forward".

A Figura 2.5 mostra uma rede com duas camadas sucessivas de unidades e, portanto, duas

camadas de pesos. As unidades da camada do meio são conhecidas como unidades escondidas já que seus valores de ativação não são diretamente acessíveis de fora da rede.

Figura 2.5: Uma rede neural multicamadas contendo 2 camadas de pesos.



Fonte: Bishop, 1994.

No diagrama, cada círculo da parte inferior representa uma das entradas x_i; cada círculo no topo representa uma das saídas z_j, e as linhas conectando os círculos representam os pesos correspondentes w_{ji}. A entrada extra $x_0 = 1$ é mostrada por um círculo preto e as linhas conectando esta unidade à saída representam os parâmetros do tipo viés w_{j0}.

A ativação destas unidades é dada pela Equação 2.5, para o caso de uma rede de camada única. As saídas da rede são obtidas pela ação de outra função ativação sobre os *z*'s com uma segunda transformação, como dado em

$$y_k = \tilde{g}\left(\sum_{j=0}^{m} \tilde{w}_{kj} z_j\right), \tag{2.5}$$

onde $\tilde{w}_{kj}$ denota um peso da segunda camada, conectando a unidade *j* escondida com a unidade de

saída k, notando que foi introduzido uma unidade extra escondida com ativação $z_0 = 1$ para fornecer um viés para as unidades de saída. Os termos viés (tanto para as unidades escondidas, como para as unidades de saída) têm um papel importante, assegurando que a rede pode representar mapeamentos não-lineares gerais.

Minsky e Papert analisaram matematicamente e demonstraram que redes de uma única camada não são capazes de solucionar problemas de classificação que não sejam linearmente separáveis. Como não acreditavam na possibilidade de se construir um método de treinamento para redes com mais de uma camada, eles concluíram que as redes neurais seriam sempre suscetíveis a essa limitação (Fernandes, 2003). Contudo, o desenvolvimento do algoritmo de treinamento de retropropagação ("*backpropagation*"), por Rumelhart, Hinton e Williams em 1986, precedido por propostas semelhantes ocorridas nos anos 1970 e 1980, mostrou que é possível treinar eficientemente redes com camadas intermediárias. Nessas redes, cada camada tem uma função específica. A camada de saída recebe os estímulos da camada intermediária e constrói o padrão que será a resposta. As camadas intermediárias funcionam como extratoras de características, seus pesos são uma codificação de características apresentadas nos padrões de entrada e permitem que a rede crie sua própria representação, mais rica e complexa, do problema (Fernandes, 2003).

Se existirem as conexões corretas entre as unidades de entrada e um conjunto suficientemente grande de unidades intermediárias, pode-se sempre encontrar a representação que irá produzir o mapeamento correto da entrada para a saída através das unidades intermediárias. Como provou Cybenko (1989), são necessárias no máximo duas camadas intermediárias, com um número suficiente de unidades por camada, para se produzir quaisquer mapeamentos. Também foi provado que apenas uma camada intermediária é suficiente para aproximar qualquer função contínua (Fernandes, 2003).

Um conjunto preestabelecido de regras bem-definidas para a solução de um problema de aprendizagem é denominado um algoritmo de aprendizagem. Como se pode esperar, não há um

algoritmo de aprendizagem único para o projeto de redes neurais. Basicamente, estes algoritmos diferem entre si pela forma como é formulado o ajuste de um peso sináptico de um neurônio (Haykin, 2001).

2.2. Algoritmo de Aprendizagem de Retropropagação

Focamos o presente trabalho no algoritmo de aprendizagem de retropropagação, utilizado por nós. Para ilustrar esta regra de aprendizagem, podemos observar o grafo de sinal correspondente, incorporando duas fases, para frente e para trás, no caso de uma rede com 2 camadas, sendo 3 neurônios em cada uma delas e também para a camada de entrada.

A parte superior do grafo de fluxo de sinal corresponde ao passo para frente, enquanto a parte inferior se refere ao passo para trás, que é referido como o grafo de sensibilidade para o cálculo dos gradientes locais do algoritmo de retropropagação.

A atualização seqüencial dos pesos é o método preferido para a implementação em tempo de execução do algoritmo de retropropagação (Figura 2.6). Para este modo de operação, o algoritmo circula através da amostra de treinamento $\{x(n), d(n)\}_{(n=1)}^{N}$ como segue (Haykin, 2001):

1. Inicialização. Assumindo que nenhuma informação prévia esteja disponível, são retirados os pesos sinápticos e limiares de uma distribuição uniforme, cuja média é zero e cuja variância é escolhida para que o desvio padrão dos neurônios se encontre na transição entre as partes linear e saturada da função de ativação sigmóide.
2. Apresentação dos exemplos de treinamento. Para cada exemplo do conjunto, ordenado de alguma forma, é realizada a seqüência de computações para frente e para trás, descritas nos pontos 3 e 4, respectivamente.
3. Computação para frente (Propagação). Suponha que um exemplo de treinamento seja

representado por (x(n),d(n)), com o vetor de entrada x(n) aplicado à camada de entrada de nós sensoriais e o vetor respostas desejado d(n) apresentado à camada de saída de nós computacionais na interação n. Devem ser calculados os campos locais induzidos e os sinais funcionais da rede prosseguindo para frente através da rede, camada por camada. O campo local induzido $v_j^{(l)}(n)$ para o neurônio j na camada l é definido por

$$v_j^{(l)}(n) = \sum_{i=0}^{m_0} w_{ji}^{(l)}(n)\, y_i^{(l-1)}(n) \tag{2.6}$$

onde $y_i^{(l-1)}(n)$ é o sinal de saída (imagem da função de ativação do neurônio i na camada anterior l – 1), na interação n, e $w_{ji}^{(l)}(n)$ é o peso sináptico do neurônio j da camada l, que é alimentado pelos neurônios i da camada l – 1. Para i = 0, temos $y_0^{(l-1)}(n) = +1$ e $w_{j0}^{(l-1)}(n) = b_j^{(l)}(n)$ é o viés aplicado ao neurônio j na camada l. Assumindo-se o uso de uma função sigmóide, o sinal de saída do neurônio j na camada l é

$$y_j^{(l)}(n) = \varphi_j(l\text{-}1) \tag{2.7}$$

Se o neurônio j está na primeira camada oculta (*i.e.*, l = 1), fazemos

$$y_j^{(0)}(n) = x_j(n) \tag{2.8}$$

onde $x_j(n)$ é o j-ésimo elemento do vetor de entrada x(n). Se o neurônio j está na camada de saída (*i.e.*, l = L, onde L é denominado a profundidade da rede), fazemos

$$y_j^{(L)}(n) = o_j(n)\,, \tag{2.9}$$

Calculando o sinal de erro

$$e_j(n) = d_j(n) - o_j(n) \tag{2.10}$$

onde $d_j(n)$ é o j-ésimo elemento do vetor resposta desejada d(n).

4. Computação para Trás (Retropropagação). Devem ser calculados os δs (*i.e.*, gradientes locais) da rede, denotados e definidos por

$$\delta_j^{(l)}(n) = e_j^{(L)}(n)\varphi_j'\left(v_j^{(L)}(n)\right) \text{ para o neurônio } j \text{ da camada de saída L} \tag{2.11}$$

$$\varphi_j'(v_j^{(L)}(n)) = \sum_k \delta_k^{(l+1)}(n) w_{wj}^{(l+1)}(n) \text{ para o neurônio } j \text{ camada oculta } l \tag{2.12}$$

onde o apóstrofo em $\varphi'_j(\cdot)$ representa a diferenciação em relação ao argumento. Os pesos sinápticos da rede na camada l devem ser ajustados de acordo com a regra delta generalizada:

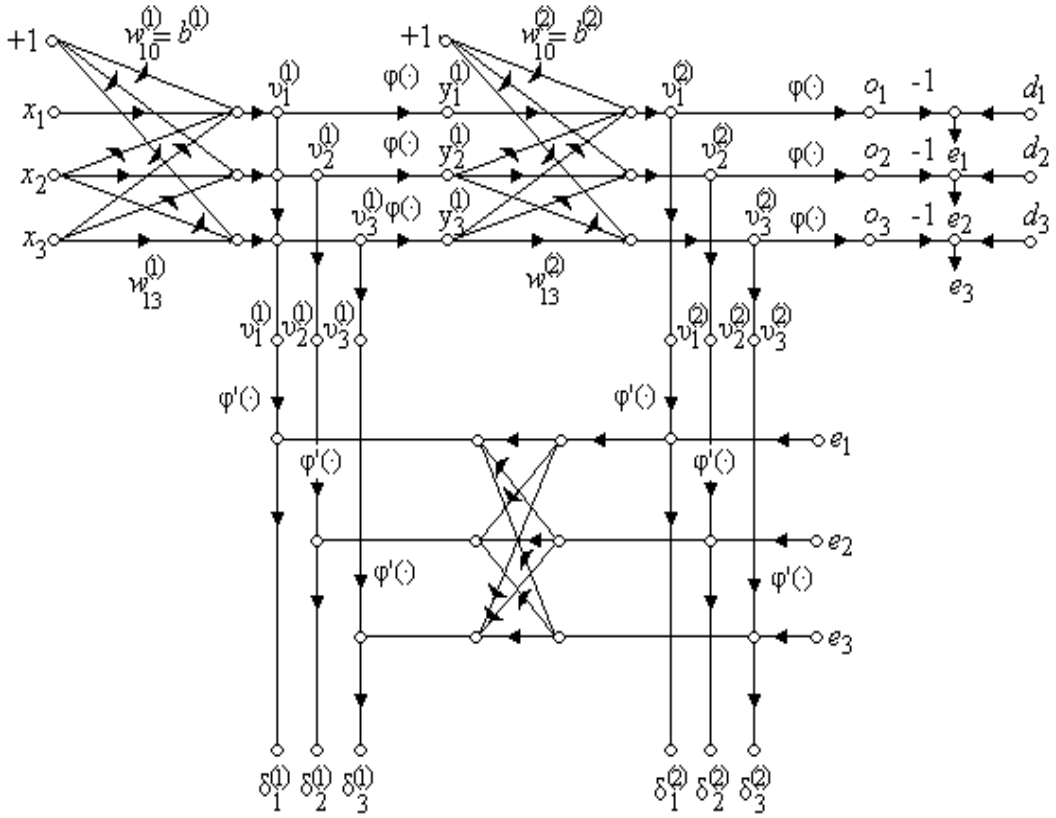
$$w_{ji}^{(l)}(n+1) = w_{ji}^{(l)}(n) + \alpha\left[w_{ji}^{(l)}(n-1)\right] + \eta\delta_j^{(l)}(n) y_i^{(l-1)}(n) \tag{2.13}$$

onde η é a taxa de aprendizagem e α é a constante de momento.

5. Iteração. Devem ser realizadas as computações para frente e para trás dos pontos 3 e 4, apresentando novas épocas de exemplos de treinamento para a rede até que seja satisfeito o

critério de parada.

Figura 2.6: Resumo gráfico do fluxo de sinal da aprendizagem por retropropagação.



Fonte: Haykin, 2001.

Parte superior do grafo: passo pra frente. Parte inferior do grafo: passo para trás.

2.3. Validação de um Modelo

Para serem eficientes, os modelos criados devem ser validados e testados de forma que garantam uma representação ou interpretação o mais próxima possível de um determinado

fenômeno. Isto é verdade especialmente para modelos empíricos, como redes neurais e modelos estatísticos, os quais são criados a partir de dados observados. De forma ideal, quantidades substanciais de dados devem estar disponíveis tanto para a construção como para a validação dos modelos empíricos. No entanto, para alguns sistemas, a definição dos dados é um processo custoso.

Neste trabalho serão descritos sucintamente três métodos de estimativa de erro utilizados para a avaliação de modelos de predição por Twomey e Smith (1998), são eles: (i) ressubstituição, (ii) treinamento e teste e (iii) validação cruzada e validação cruzada de grupo. Com os três métodos podemos observar o erro embutido na construção do modelo e avaliar a interferência da escolha das variáveis de entrada para a construção do modelo e na realização do teste.

De qualquer forma, antes da avaliação do modelo, qualquer metodologia de rede neural deve atender a três desafios (Yuan e Fine, 1998):

1) o grande número de potenciais variáveis de entrada;
2) a quantidade limitada de dados para o treinamento e teste que são estatisticamente homogêneos;
3) o tempo consumido no processo de treinamento que inibe uma exploração exaustiva dos desenhos alternativos da rede.

Dessa forma, podemos observar a dificuldade encontrada por nós no desenvolvimento de uma rede neural para o cálculo de um valor de afinidade utilizando para tanto um conjunto de treinamento e teste de apenas 50 complexos proteínas-ligantes, que além de tudo apresentam sítios ativos com características extremamente diversas.

Considerando o nosso problema, onde temos a variável esperada *Y*, um vetor de entrada *X* e um modelo de predição $\hat{f}(x)$ que é estimado de uma amostra treinamento, a função perda para o cálculo de erros entre o valor esperado, *Y* e $\hat{f}(X)$ é chamada de $L(Y, \hat{f}(X))$. As escolhas típicas são:

$$L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2, \text{ erro quadrático;} \tag{2.14}$$

$$L(Y, \hat{f}(X)) = |Y - \hat{f}(X)|, \text{ erro absoluto;} \tag{2.15}$$

O erro de generalização, também chamado de erro teste, é o erro de predição esperado, calculado sobre uma amostra teste independente:

$$\text{Err} = \text{E}[L(Y, \hat{f}(X))]. \tag{2.16}$$

onde o conjunto de pares de variáveis $t_i = (X_i, y_i)$, é escolhido aleatoriamente e T_n corresponde a uma amostra da população F, de tamanho n, $Tn = \{(X_1,y_1),(X_2,y_2),...,X_n,y_n)\}$, onde X_i é um vetor de variáveis características, e y_i é a variável de saída correspondente (Twomey e Smith,1998). O erro de treinamento é dado pela perda média da amostra de treinamento:

$$\hat{E}rr = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}[T_n, X_i])^{2.} \tag{2.17}$$

Gostaríamos de saber o erro de generalização do nosso modelo de predição $\hat{f}$. À medida que o modelo torna-se mais e mais complexo, ele é capaz de se adaptar às mais complicadas estruturas subjacentes (uma diminuição no viés), mas às custas de um aumento do erro de generalização (um aumento da variância). Um modelo ideal é aquele que possui uma complexidade tal que permite a obtenção de um erro de generalização mínimo.

O erro de treinamento diminui com a complexidade do modelo, tipicamente tendendo a zero se aumentarmos a complexidade do modelo o suficiente. O erro de treinamento igual a zero é

superadaptado para os dados de treinamento e o modelo irá tipicamente generalizar mal.

Os métodos mencionados aqui servem para estimar a curva do erro de generalização para um modelo. Um modelo pode possuir um ou mais parâmetros de ajuste, A, e assim podemos escrever as nossas previsões como $\hat{f}_A(x)$. Os parâmetros de ajuste modificam a complexidade do nosso modelo, e buscamos os valores dos parâmetros que minimizem o erro, ou seja, produzem o mínimo da curva do erro de generalização.

Neste trabalho, utilizaremos a seguinte terminologia:

- Modelo de predição: estima o desempenho de um modelo com uma dada complexidade;
- Modelo de validação: para um dado modelo de predição, é construído para estimar o erro de generalização correspondente.

2.3.1. Ressubstituição

No método de ressubstituição, o modelo de predição é criado a partir do treinamento do mesmo conjunto utilizado para teste, até que o conjunto de pesos encontrado seja aquele que retorne o menor erro possível para aquela amostra.

O método de ressubstituição é computacionalmente eficiente e requer a construção de apenas um único modelo, o qual é usado tanto para a predição quanto para a validação. A ressubstituição ocorre necessariamente na construção do modelo na rede neural, uma vez que este é criado a partir da minimização do erro para um determinado conjunto de dados de entrada.

A metodologia de ressubstituição é extremamente dependente dos dados utilizados no treinamento para a construção do modelo, sendo que quanto maior e diverso o conjunto de treinamento, mais perto do esperado será o valor estimado para um conjunto-teste.

2.3.2. Treinamento e teste

O método mais comumente utilizado para a avaliação do desempenho de uma rede neural faz uso da estimativa de erro de um conjunto independente de dados não utilizados para a construção do modelo. Tem-se uma amostra de tamanho n, subdividida em 2 sub-amostras, tal que $n_1 + n_2 = n$. Uma amostra de tamanho de n_1 é utilizada para construção do modelo $\hat{f}[T_{n1},X]$ e uma amostra de tamanho n_2 é usada para a validação. No modelo $\hat{f}[T_{n1},X]$, T_n corresponde a uma amostra da população F de tamanho n, $Tn = \{(X_1,y_1),(X_2,y_2),...,X_n,y_n)\}$, onde $t_i = (X_i,y_i)$ consiste de um vetor de variáveis características X_i e a variável de saída correspondente y_i, $i = 1,...,n$ (Twomey e Smith,1998).

Ao contrário do método de ressubstituição, o método de treinamento e teste não utiliza todos os dados disponíveis na construção dos modelos de predição. Em vez disso, o modelo de predição é construído em um conjunto reduzido de dados n_1. Como no método de ressubstituição, apenas um modelo é construído. O erro de generalização é calculado da seguinte forma:

$$\hat{Err}_{T/T} = \frac{1}{n_2} \sum_{j=n_1+1}^{n} (y_j - \hat{f}[T_{n1}, X_j])^2 \qquad (2.18)$$

Enquanto muitos autores têm comparado vários esquemas de particionamento para o treinamento e teste, existem trabalhos que utilizam formas de validação computacionalmente mais custosas. Validação Cruzada e Validação Cruzada de Grupo (VCG) são sugeridas para isso em modelos de análise de desempenho de redes neurais (Twomey e Smith,1998).

2.3.3. Validação Cruzada e Validação Cruzada de Grupo

Na validação cruzada, o erro de generalização é determinado pela construção de *n* modelos de validação a partir de cada uma das combinações das partições dos dados de toda a população, deixando sempre um de fora (tamanho $n - 1$), e então este único ponto omitido é testado. Neste método, n modelos de validação e 1 modelo de predição (construído utilizando todos os n dados da amostra) são construidos (Twomey e Smith,1998).

Uma variação computacionalmente menos pesada do CV é a validação cruzada de grupo, na qual (normalmente) grupos de tamanhos iguais de dados são removidos de cada modelo de validação, em vez de apenas uma única observação (Twomey e Smith,1998).

Quando estamos lidando com diferentes propostas de modelos, com distintas complexidades ou arquiteturas de redes neurais, o modelo que apresentar o menor erro de generalização, obtido por validação cruzada, deve ser o escolhido para se gerar o modelo de predição.

A predição do erro de generalização utilizando-se validação cruzada de grupo, para um dado modelo de predição, pode ser resumida nos seguintes passos:

1. Divide-se o conjunto de dados da amostra em *M* partes iguais;

2. Para cada parte *i*, um modelo de validação é gerado utilizando-se os dados das outras (M-1) partes. O erro de predição do modelo de validação é calculado utilizando-se os dados da parte *i*.

3. Calcula-se o erro de generalização utilizando-se a seguinte fórmula:

$$\hat{Err}_{VCG} = \frac{1}{n} \sum_{m=1}^{M} \sum_{h=1}^{H} \left| y_{(m-1)\,H+h} - \hat{f}\left[T_{(m)}, x_{(m-1)\,H+h} \right] \right| \tag{2.19}$$

onde m é o índice para o grupo excluído na construção do modelo de validação e $T_{(m)} = \{(x_1,y_1,), (x_2,y_2,),...,(x_{(m-1)H},y_{(m-1)H},), (x_{(m)H+1},y_{(m)H+1}),...,(x_n,y_n)\}$ é usado para a seleção do modelo de aplicação. O

valor n é o número de indivíduos da população, *M* é o número total de grupos e *H* é o número de observações por grupo, portanto $n = MH$.

CAPÍTULO 3

3. OBJETIVOS

3.1. Objetivo Geral

Neste trabalho, o objetivo geral é construir uma função empírica para calcular o valor de afinidade receptor-ligante utilizando uma rede neural treinada com o algoritmo de retropropagação. A construção desta função empírica será direcionada pelo contexto de implementação do programa de atracamento molecular DOCKTHOR (Magalhães, 2006) que utiliza o campo de força clássico GROMOS96 e uma metodologia de atracamento na rede (interpolação dos valores dos termos eletrostático e de van der Waals) para o cálculo da função energia.

3.2. Objetivos Específicos

Neste trabalho, consideramos como objeticos específicos:

(i) Definição, caracterização e construção de um conjunto-teste com 50 complexos proteína-ligante para construção de uma função empírica visando estimar constantes de afinidade receptor-ligante;

(ii) Definição das variáveis físico-químicas envolvidas no processo de atracamento molecular a serem utilizadas na implementação da função empírica;

(iii) Análise da influência das variáveis físico-químicas escolhidas na determinação da constante de afinidade receptor-ligante;

(iv) Avaliação dos modelos construídos utilizando redes neurais, visando a implementação de uma nova função empírica para estimar valores de afinidade receptor-ligante no programa DOCKTHOR.

CAPÍTULO 4

4. METODOLOGIA

A primeira parte deste trabalho envolveu definição, construção e caracterização de um conjunto-teste para a elaboração de uma função empírica para o cálculo da constante de afinidade receptor-ligante utilizando uma rede neural construída com auxílio da metolodologia de regressão múltipla e treinada com o algoritmo de retropropagação. É importante ressaltar que o objetivo é a construção de uma função empírica de caráter geral e não específica para uma determinada família de proteínas, e portanto o conjunto-teste deve ter uma certa abrangência no número de famílias distintas de proteínas consideradas.

Foram utilizadas para o treinamento da rede neural e para a regressão múltipla, que tem como valor de saída a constante de afinidade encontrada experimentalmente, cinco variáveis físico-químicas envolvidas no processo de atracamento molecular: os valores de energias de interação eletrostática, os valores de energia associados ao termo de Lennard-Jones, a área de contato ligante-receptor da superfície acessível ao solvente, o número de ligações hidrogênio receptor-ligante, e o número de ligações torcionáveis congeladas do ligante no processo de interação com a proteína.

Descreveremos neste capítulo como foi construído o conjunto-teste, como foram determinadas as variáveis físico-químicas a partir dos complexos proteína-ligante e, por fim, a estruturação da rede neural para a construção de uma função empírica para o cálculo da constante de afinidade receptor-ligante junto com a descrição de como foram realizados os diferentes testes para a avaliação da rede neural construída.

4.1. Definição e Construção do Conjunto-Teste

O conjunto-teste utilizado para a construção de uma função empírica para diferentes famílias de proteínas foi selecionado a partir de outros conjuntos-teste utilizados na avaliação do desempenho de programas de atracamento molecular descritos na literatura. Entre eles, 45 complexos proteínas-ligante do AutoDock (Morris et al., 1998); 15 complexos do DOCK 4.0 (Ewinga et al., 2001); 200 do FlexX (Kramer e Rarey, 1999); 305 complexos do CCDC/Astex utilizados para a validação da mais nova versão da função empírica do programa de atracamento GOLD e 99 complexos proteínas-ligante do GOLD (Willet et al., 1997). Também foram utilizados conjuntos-teste descritos em artigos que procuram validar as funções empíricas mais conhecidas, como o conjunto de 100 complexos proteínas-ligante descritos em Wang e Lu (2003) e 37 utilizados por Bursulaya et al. (2003). No total foram avaliados 601 complexos proteína-ligante para a construção do conjunto-teste final.

Com a descrição de todos os conjuntos-teste, foi desenvolvido um programa para observar a freqüência dos complexos proteínas-ligante. Depois de encontradas as mais freqüentes, observou-se os seguintes critérios de seleção do cristal: (i) se a resolução era menor que 3,00 Å; (ii) a presença de apenas um ligante no sítio ativo do receptor; (iii) a variação do tamanho e flexibilidade dos ligantes, que deveriam ter até 28 ligações torcionáveis e (iv) variação grande de pKi (em ordens de grandeza) para que a função empírica não ficasse restrita a um intervalo pequeno de valores de constante de afinidade.

Depois de selecionados os complexos proteína-ligante que atendiam a estas restrições, foi analisada a presença de moléculas de água, de íons metálicos ou cofatores. As águas cristalográficas foram removidas e os íons metal, quando presentes no sítio ativo da proteína, foram mantidos juntos à proteína. Estruturas que apresentavam cofatores no sítio ativo da proteína foram descartadas.

O conjunto-teste final conteve 50 estruturas, onde as proteínas foram parametrizadas com o

campo de força GROMOS96 (van Gunsteren e Berendsen, 1987), próprio para a parametrização de moléculas protéicas, com o auxílio do programa PDBTHORBOX (Dardenne, 2000), sendo nesta etapa adicionados os átomos de hidrogênio aos átomos polares dos receptores.

Utilizou-se a versão Beta do sítio do PRODRG (Schuettelkopf e Aalten, 2004), para a parametrização dos ligantes (APÊNDICE C). Neste mesmo sítio web, as moléculas tiveram os átomos de hidrogênio adicionados e foram criados os arquivos de topologia a partir do campo de força GROMOS96 (tipos de átomos e termos associados aos ângulos diedrais), para então parametrizar suas cargas com o campo de força MMFF94 (Halgren, 1996abcde), no programa Spartan (Wavefunction Inc., 2006). Tivemos um cuidado especial com a ionização e hibridização dos átomos, verificando-as com informações obtidas no banco de dados de ligantes AffinDB (http://pc1664.pharmazie.uni-marburg.de/affinity/) (Block et al., 2006). Neste mesmo sítio, do AffinDB, foram obtidas as informações das constantes de afinidades dos complexos receptor-ligante.

4.2. Determinação das Variáveis de Entrada da Rede Neural

Para a determinação das variáveis de entrada da rede neural: número de ligações hidrogênio, energia eletrostática, energia de Lennard-Jones, porcentagem da superfície acessível ao solvente do ligante em contato com a proteína e número de ligações torcionáveis congeladas no processo de interação intermolecular, realizamos em uma etapa anterior, a minimização do ligante no sítio receptor, utilizando o programa DOCKTHOR (de Magalhães, 2006).

O processo de minimização e posterior cálculo das variáveis físico-químicas é um aspecto importante no contexto da construção de uma função empírica a ser utilizada por um programa de atracamento molecular receptor-ligante. Ou seja, a minimização dentro do programa de certa forma reproduz o melhor resultado possível do algoritmo/método de atracamento em um experimento computacional real

(já incluindo as aproximações metodológicas inerentes ao método em questão).

4.2.1. DOCKTHOR

O DOCKTHOR (de Magalhães, 2006) é um programa para atracamento molecular e nele foram implementadas e analisadas várias estratégias de algoritmos genéticos (AG's) e para o problema de atracamento receptor-ligante. Os métodos implementados foram testados em estudos de "*re-docking*" e "*cross-docking*" de inibidores da enzima HIV-1 protease com alto grau de flexibilidade conformacional (de Magalhães, 2006).

O programa DOCKTHOR tem como interesse solucionar o problema de atracamento de ligantes flexíveis em receptores rígidos (aproximação de receptor rígido). Dessa forma, a estrutura da proteína é mantida fixa na conformação da estrutura observada experimentalmente. Nesse caso, uma vez que se tenha a conformação de ligação da proteína e a localização do sítio-ativo da mesma, o problema consiste em encontrar a conformação de ligação da molécula ligante correspondente à melhor energia de interação ligante-receptor (de Magalhães, 2006).

As diferentes características de cada conformação são representadas por um vetor de valores reais, o cromossomo. Cada indivíduo da população corresponde a um cromossomo, que é constituído por três genes representando os graus de liberdade translacionais, quatro genes representando os graus de liberdade rotacionais (*i.e.*, um quaternion, onde três graus de liberdade representam o vetor unitário que dá a direção de rotação, e um grau de liberdade dá o ângulo de rotação) e N genes representando os graus de liberdade conformacionais da molécula ligante (de Magalhães, 2006).

A população inicial de indivíduos é gerada aleatoriamente, ou seja, para cada gene de cada cromossomo na população, um valor real aleatório entre o limite inferior e superior da variável é gerado. A grade de energia foi construída com dimensões de [23Å, 23Å, 23Å] e centrada na posição

experimental do átomo de referência do ligante.

No caso dos genes rotacionais correspondentes à rotação de corpo rígido da molécula, um valor aleatório em [0°, 360°] é gerado para o ângulo de rotação. Além disso, para os três genes rotacionais correspondentes ao vetor unitário, são atribuídos valores em [-1, 1]. O vetor é normalizado durante a fase de avaliação da solução; entretanto, essa modificação não altera os valores dos genes no cromossomo (de Magalhães, 2006).

A distribuição da população inicial é feita utilizando uma distribuição de Cauchy. Os genes translacionais são gerados no centro da grade de energia e são deslocados por um fator, que é um número real aleatório obtido através de uma distribuição de Cauchy, que é dada por:

$$C(\alpha, \beta, x) = \frac{\beta}{\pi (\beta^2 + (x - \alpha)^2)} \tag{4.1}$$

onde α e β são 0 e 0,75, respectivamente.

Os operadores genéticos utilizados no programa para a geração de novos indivíduos foram o de recombinação ("*crossover*") e mutação. Os operadores de recombinação utilizados foram: (i) de 2-pontos, que gera duas soluções filhas a partir de duas soluções pais, trocando os genes entre dois pontos de corte aleatoriamente escolhidos nos cromossomos-pais; (ii) recombinação discreta, que gera uma nova solução filha a partir de duas soluções pais, onde para cada gene é realizada uma escolha aleatória, para decidir de qual dos dois pais o gene correspondente será herdado; (iii) recombinação discreta modificada, onde duas novas soluções filhas são geradas a partir de duas soluções pais, trocando um conjunto de genes do mesmo tipo (translacional, rotacional ou conformacional) entre as solução pais e (iv) a recombinação binária simulada que gera duas soluções filhas a partir de duas soluções pais (de Magalhães, 2006).

Os operadores de mutação utilizados no programa são três: (i) aleatória, que gera um novo indivíduo modificando o valor de um gene da solução pai, aleatoriamente escolhido no

cromossomo; (ii) não-uniforme, que quando aplicado a um indivíduo em determinada geração gera um novo indivíduo modificando uma variável aleatoriamente escolhida. O operador de mutação não-uniforme reduz a amplitude das perturbações nos genes com o aumento do número de gerações e (iii) mutação de Cauchy, que modifica um gene sorteado aleatoriamente de um indivíduo, somando a esse gene um número real aleatório obtido por uma distribuição de Cauchy (4.1), sendo que os parâmetros utilizados por padrão para α e β são 0 e 1, respectivamente (de Magalhães, 2006).

Um método adaptativo foi utilizado para definir dinamicamente as probabilidades dos operadores genéticos. O método adaptativo implementado consiste em coletar informações sobre a qualidade das soluções geradas por cada operador e atribuir "crédito", de maneira que a probabilidade dos operadores com melhor desempenho seja aumentada em detrimento daqueles que não foram tão produtivos durante o processo de busca (de Magalhães, 2006).

Depois de aplicados os operadores, os indivíduos da população eram avaliados por uma função aptidão e para tanto devemos saber um pouco mais de um dos principais aspectos de um programa de atracamento molecular que está associado ao tratamento dado às interações de longo alcance. Na metodologia de atracamento na grade, os valores relacionados ao potencial de LJ e eletrostático, de todos os átomos do receptor rígido são pré-calculados e armazenados em cada ponto de uma grade/malha tridimensional construída de tal forma a englobar o sítio-ativo do receptor. Neste programa, no entanto, a energia de interação eletrostática e os termos relativos ao potencial de LJ são calculados para cada tipo de átomo do campo de força, e não apenas para cada tipo de átomo do ligante, como nas outras implementações.

Embora o potencial de Lennard-Jones (LJ) modele de forma adequada as interações repulsivas e atrativas, o termo repulsivo (r^{-12}) muitas vezes se torna um problema, fazendo com que a energia de interação do ligante com o receptor tenha um valor muito alto, mesmo para conformações muito próximas às conformações com interação "ótima" com o receptor.

Vários métodos têm sido empregados pelos programas de atracamento para a "suavização"

do potencial de LJ. Neste trabalho, um multiplicador (*p*) que varia com o contador de avaliações de função aptidão (*neval*), foi introduzido no potencial de LJ, onde p é dado por:

$$p = min\{1, \frac{neval}{(c \times maxeval)}\} \tag{4.2}$$

onde *maxeval* é o número máximo de avaliações de função, e *c* foi escolhido como 0,5. Esse parâmetro foi introduzido para que, no início do algoritmo, o termo de energia intramolecular do ligante seja priorizado em relação ao termo de energia de interação proteína-ligante. O termo relativo às interações proteína-ligante, modeladas pelo potencial de LJ, é então introduzido de maneira gradual (de Magalhães, 2006).

A função aptidão utilizada no programa é a energia de interação proteína-ligante. Como o objetivo do programa é minimizar a função energia, o melhor indivíduo será aquele que possuir o menor valor de função aptidão, que compreende termos da energia de interação para átomos não-ligados proteína-ligante e da energia intramolecular do ligante, além do termo relativo aos ângulos diedrais da molécula ligante. A função utilizada possui a forma:

$$\sum_{proteína}\sum_{ligante} p\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + \frac{q_i \times q_j}{(D(r_{ij}) r_{ij})} \quad + \quad \sum_{ligante}\sum_{ligante} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i \times q_j}{(D(r_{ij}) r_{ij})} \quad + \quad \sum_{diedrais} \gamma_k (1 + \cos(w_k \theta_k - \theta_{0k})) \tag{4.3}$$

onde r_{ij} é a distância entre os átomos *i* e *j*, A_{ij} e B_{ij} são os parâmetros de Lennard-Jones, q_i e q_j são as cargas parciais atômicas dos átomos e *i* e *j* respectivamente, e D é a função dielétrica sigmoidal dependente da distância r_{ij}, que modela o efeito de blindagem do solvente nas interações eletrostáticas (Arora e Jayaram, 1997). O parâmetro γ_k é a constante de energia associada com a rotação de uma ligação química, θ_k é o ângulo de torção, w_k é a periodicidade e θ_{0k} é o ângulo de fase.

Depois de aplicada a função aptidão ocorria a seleção e inserção de novos indivíduos na população com o método de Seleção por Torneio Restrito modificado (STRM).

Na Seleção por Torneio Restrito Modificado, a população é ordenada por critério de energia, e, tomando-se como referência a energia do novo indivíduo gerado, dois conjuntos de indivíduos da população são selecionados aleatoriamente. Um conjunto contendo indivíduos com energias melhores que o indivíduo gerado e outro conjunto contendo indivíduos com energias piores. Para cada um dos dois conjuntos é selecionado o indivíduo mais próximo (utilizando-se a distância euclidiana no espaço dos genótipos, onde para cada tipo de gene utilizam-se pesos específicos) do novo indivíduo gerado. Se o desvio médio quadrático (RMSD, "*Root Mean Square Deviation*") do novo indivíduo com o mais próximo do primeiro conjunto for maior que 2,0 Å, ocorre a substituição do indivíduo do segundo conjunto, senão se o desvio é menor que 2,0 Å, o novo indivíduo é descartado. Nessa técnica, os nichos de melhor aptidão são fortalecidos com a utilização de um critério de RMSD < 2,0 Å, que evita a inserção de um indivíduo quando outro indivíduo melhor e muito similar a este já existe na população (de Magalhães, 2006).

Para a formação de nichos na população implementadas via STRM, um tamanho de torneio (*w*) dinâmico foi utilizado. A utilização de tamanho de torneio dinâmico para o método STRM foi implementada com um contador do número de avaliações, de maneira que no início do algoritmo, w seja grande, e que diminua ao longo da evolução. O valor inicial do parâmetro *w* é 1000 e decai linearmente até um valor mínimo (de Magalhães, 2006).

Além dos operadores já descritos, o programa também utiliza o método de busca local de Hooke-Jeeves, com uma probabilidade de aplicação após a aplicação dos operadores genéticos. Neste método, o procedimento de ir de um ponto no espaço de busca para outro é chamado de um movimento com o operador Hooke-Jeeves com 100% de probabilidade de atuação. O método de Hooke-Jeeves consiste basicamente num processo iterativo de combinação de movimentos exploratórios com um movimento padrão. Os movimentos exploratórios são realizados

sistematicamente na vizinhança da solução atual, visando encontrar a melhor solução ao redor daquela. Quando uma nova solução melhor do que a atual é encontrada, as duas são utilizadas para a realização de um movimento padrão. Este movimento utiliza a informação adquirida nos movimentos exploratórios para indicar uma direção provável para um movimento com "sucesso" (de Magalhães, 2006).

No movimento exploratório, cada componente do vetor solução é perturbada positiva e negativamente, de acordo com um determinado tamanho de passo. O valor da coordenada correspondente ao menor valor da função f é então retido. No final do processo, após todas as componentes terem sido investigadas, o ponto atual é comparado com a solução inicial. Se o ponto atual diferir do original, o movimento exploratório é considerado um sucesso, caso contrário, uma falha é considerada. O movimento exploratório retorna o melhor ponto obtido durante todo o processo (de Magalhães, 2006).

Já o movimento padrão gera um novo ponto, a partir do melhor ponto atual, na direção do melhor obtido anteriormente. Isso é realizado modificando-se todas as coordenadas do ponto atual, por um valor igual à diferença entre o ponto atual e o anterior. O resultado do movimento padrão pode ser um sucesso ou uma falha (de Magalhães, 2006).

4.2.2. Minimização de Energia

Uma vez selecionados os complexos proteína-ligante, com suas estruturas cristalográficas determindas experimentalmente, realizamos uma etapa de minimização de energia dos ligantes em suas conformações originais (dado cristalográfico), no sítio ativo das proteínas correspondentes.

A minimização da energia é uma técnica que visa encontrar um conjunto de coordenadas que minimizam a energia potencial do sistema de interesse. O procedimento básico consiste em caminhar sobre a superfície de potencial na direção em que a energia decresce de maneira que o

sistema é levado a um mínimo de energia local próximo. A minimização da energia explora somente uma pequena parte do espaço de configurações. Porém, pelos ajustes nas posições atômicas, ela relaxa as distorções nas ligações químicas, nos ângulos entre ligações e nos contatos de van der Waals.

Para a realização desta etapa, modificamos o programa DOCKTHOR, onde criamos a população de 1.000 indivíduos a partir da conformação original do ligante no cristal do complexo com uma grade de energia, centralizada no átomo de referência do ligante, de tamanho 23 Å nas três direções e com uma discretização de 0,25 Å, permitindo o deslocamento translacional de 0,1 Å nas três direções e 0,5º na variação do ângulo rotacional. Para reduzir o espaço de busca e permitir de fato a atuação do método de busca local de Hooke-Jeeves (Hooke-Jeeves, 1960), alteramos um dos parâmetros da distribuição de Cauchy, que determina um número real aleatório somado a um dos valores passíveis de mutação no algoritmo genético.

Os parâmetros padrões utilizados na distribuição de Cauchy (Equação 4.1) para α e β são 0 e 0,75. No entanto, para minimizar a variação de nossa população, reduzimos o valor β para 0,15. O método de Hooke-Jeeves normalmente aplicado com uma probabilidade de 0,002, teve sua freqüência alterada para 1,000, confirmando sua participação na minimização da estrutura alvo.

O método de busca padrão de Hooke-Jeeves consiste em aplicar iterativamente movimentos exploratórios e um movimento padrão. Como estamos realizando a minimização da estrutura, reduzimos o tamanho do passo, para 0,001 Å, nos movimentos translacionais, nas três direções, e 0,1º na variação do ângulo rotacional, aplicando pequenas variações para a busca da melhor solução (de Magalhães, 2006). A minimização de energia no sítio ativo envolveu a utilização do algoritmo genético com o operador Hooke-Jeeves com probabilidade 1,0, garantindo que este atuasse sempre.

No final da etapa de minimização de energia dos ligantes no sítio ativo das proteínas foi escolhida a conformação de menor energia para a etapa de cálculo das variáveis independentes.

4.2.3. Ligações hidrogênio

Um programa para a determinação de possíveis ligações hidrogênio, formadas entre a proteína e o ligante no processo de interação intermolecular, foi desenvolvido neste trabalho. O programa foi escrito na linguagem C e utiliza entrada de arquivos no formato do Protein Data Bank (PDB).

O programa calcula as distâncias entre os potenciais doadores ou aceptores de ligações hidrogênio e retorna todas as interações que estão dento do intervalo definido pelo usuário (valores utilizados neste trabalho: limite inferior de 2,7 Å e o superior de 3,3 Å). O objetivo é mostrar as interações de ligações hidrogênio entre o ligante selecionado e os resíduos da proteína.

Além deste tipo de interação, analisamos ainda a interação entre o ligante e o receptor mediada por ligações hidrogênio envolvendo moléculas de água cristalográficas. Este último tipo de interação não foi utilizado como variável de entrada da rede neural, mas foi levado em consideração para avaliar a possível participação destas moléculas de águas cristalográficas no sítio ativo de algumas proteínas, fato que poderia influir no desempenho da função de energia de ligação para estes complexos.

4.2.4. Energia Eletrostática e Energia de Lennard-Jones

No cálculo das energias eletrostáticas e de Lennard-Jones, os termos de longo alcance para cada conformação investigada foram calculados, após o processo de minimização, através de uma interpolação trilinear dos valores de oito pontos localizados nos vértices de um cubo da grade de contribuições energéticas contendo um determinado átomo do ligante (de Magalhães, 2006).

O valor da energia de interação eletrostática associada a um determinado átomo do ligante foi obtido utilizando uma função dielétrica sigmoidal (Arora e Jayaram, 1997) multiplicando-se a

carga deste átomo pelo valor do potencial eletrostático EL, obtido pela interpolação trilinear (de Magalhães, 2006).

O valor da energia de interação relacionada ao potencial de Lennard-Jones (LJ) é obtido de forma análoga, diretamente da interpolação trilinear (considerando-se os valores de energia do potencial de LJ pré-calculados nos oito pontos da grade para o tipo de átomo em questão).

No cálculo destas variáveis, o programa DOCKTHOR (de Magalhães, 2006) foi modificado para o desacoplamento dos termos eletrostático e do potencial de Lennard-Jones (na saída do programa).

4.2.5. Área de Interface Entre Ligante e Proteína

É importante lembrar que tecnicamente existem diversos tipos de superfícies moleculares. Neste trabalho adotamos o somatório da superfície acessível ao solvente (SAS) de todos os átomos do ligante, já que em modelos aproximados/simplificados o efeito hidrofóbico é considerado proporcional à superfície hidrofóbica "enterrada" do ligante e por esta ser uma das formas mais simples de avaliar esse tipo de interação.

Para verificar a área de interface entre o ligante e a proteína, foi desenvolvido um programa baseado no algoritmo de Connoly (Conolly, 1983). Este programa determina a área de interseção das superfícies moleculares do ligante e da proteína, através da sobreposição das superfícies traçadas pelo centro de uma sonda esférica imaginária (*i.e.,* Superfície Acessível ao Solvente, SAS), representando uma molécula do solvente, ao ser rolada sobre a parte exterior das superfícies formadas pela união das esferas de van de Waals, dos átomos do ligante e da proteína (Lee e Richards, 1971).

O algoritmo de Connoly (Conolly, 1983) foi implementado no programa MDS (*Molecular Dot Surface*) (Connoly, 1986). Dado um arquivo de entrada contendo as coordenadas espaciais dos

átomos que constituem o complexo em estudo, os raios de van der Waals associados a estes átomos, o raio da sonda esférica e a densidade de pontos $\AA^2$ da superfície de interseção, o MDS gera um conjunto de pontos relativos à esta região e, para cada ponto, são geradas as coordenadas espaciais do vetor normal externo à superfície (n_x, n_y, n_z) e uma área aproximada da superfície associada a este ponto.

Neste trabalho, foi utilizado o programa SASCalc (Goliatt, 2007), escrito em linguagem C, que usa o MDS como rotina principal e como parâmetros de entrada um arquivo no formato PDB (com adição dos hidrogênios nos átomos polares), os valores do raio de sonda (R_s = 1,4 Å) e da densidade de pontos desejados. Os raios de van der Waals são associados aos respectivos átomos de forma automática, assumindo-se os valores calculados por Bondi (1964) (Apêndice B) e para o tratamento dos átomos unidos (*i.e.*, tipos de átomos do campo de força clássico GROMOS96 associados aos carbonos apolares, não aromáticos, CH, CH2 e CH3) foram utilizados valores determinados por Li e Nussinov (1998) (Apêndice B).

O valor usado como variável de entrada para a rede neural foi a porcentagem da área de superfície total do ligante, em contato com a proteína após o processo de minimização.

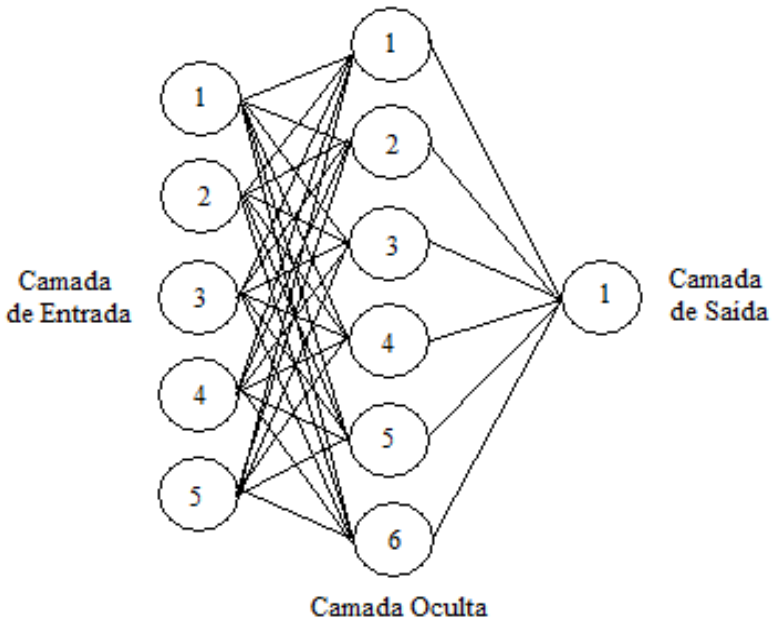
4.2.6. Número de Ligações Congeladas durante a Interação Proteína-Ligante

O número de ligações congeladas durante o processo de interação entre as moléculas de proteína e ligante foi determinado com a observação direta das estruturas tri-dimensionais do complexo, após o processo de minimização. Assim como no AutoDock, este termo reflete o número de ligações "congeladas" do tipo sp3 no ligante (Morris, et al., 1998), lembrando que ligações do tipo peptídicas não foram contabilizadas, uma vez que este tipo de ligação é bastante rígida devido ao seu caráter parcial de dupla ligação.

4.3. Arquitetura da Rede Neural

Foi criada uma rede neural no programa MATLAB, versão 7.0.0.19901 (R14) com a caixa de ferramentas para redes neurais, na versão 4.0.3 (R14). A arquitetura da rede foi de 3 camadas, sendo uma camada com 5 neurônios, uma oculta e uma camada de saída com 1 neurônio. Foram investigadas três possibilidades de arquitetura relativas ao número de neurônios da camada oculta. Uma com 3 neurônios, uma com 6 neurônios (Figura 4.1) e uma com 9 neurônios. As camadas possuíam a função de ativação do tipo sigmoidal. Os pesos desta rede foram ajustados com o algoritmo de aprendizagem de retropropagação. Para o treinamento e o teste da rede neural o conjunto de dados foi normalizado no intervalo [-1,1].

Figura 4.1: Modelo de rede neural com uma camada de entrada com 5 neurônios (variáveis de entrada), uma oculta com 6 neurônios e uma camada de saída com 1 neurônio.



Foram utilizados ainda como parâmetros para a rede neural, uma taxa de aprendizagem com valor de 0,1 e um fator momento (suavizante) igual a 0,9; erro de 10^{-3}; 1.000 épocas e gradiente de parada igual a 10^{-5}. A rede neural foi treinada para reproduzir os dados de constantes de afinidade, determinadas experimentalmente para o conjunto-teste de 50 ligantes.

Para a obtenção de cada modelo (de predição ou validação) a rede neural foi treinada 100 vezes e o conjunto de pesos que retornou as saídas mais próximas dos valores esperados, estimado a partir do valor de r (coeficiente de correlação linear), era selecionado para o teste com os valores de entrada ainda não utilizados. O conjunto de entrada possuía valores da porcentagem da superfície total do ligante em contato com a proteína, energia de Lennard-Jones, energia eletrostática, número de possíveis ligações hidrogênio existentes entre a proteína e o ligante e número de ligações torcionáveis congeladas na interação do ligante com a proteína e, como valores de saída, dados de - log K_i ou - log K_d de 50 complexos proteína-ligante, que eram normalizados antes do treinamento da rede neural.

Foram feitos 3 testes diferentes para construção e validação da rede neural. No primeiro deles utilizamos a metodologia de ressubstituição, onde assim como na regressão múltipla, todos os 50 complexos proteína-ligante foram utilizados no treinamento e na estimativa do erro de generalização. No segundo, utilizou-se a metodologia de treinamento e teste. Foram investigados 3 tipos de partição de dados, contendo 15, 25 e 45 complexos no conjunto de treinamento (e respectivamente 35, 25 e 5 complexos no conjunto de teste).

A partição dos dados foi feita de maneira aleatória. No terceiro teste, utilizou-se a metodologia de validação cruzada de grupo, para se estimar o erro de generalização de um modelo de predição obtido utilizando-se os 50 complexos receptor-ligante. Na validação cruzada de grupo o conjunto de dados foi dividido em 5 partes iguais, o erro de generalização foi calculado utilizando-se a fórmula 2.19 e o modelo de predição foi construído utilizando-se todos os 50 complexos. Nos 3 testes ainda foram analisadas a dependência do modelo com a variação do número (*i.e.*, 3, 6 e 9)

de neurônios na camada oculta (*i.e.*, análise na variação da complexidade da rede neural). Na construção de modelos utilizando-se regressão múltipla e rede neural, todos os dados foram normalizados no intervalo de -1 a 1.

4.4. Regressão Múltipla

No problema que estamos tratando temos os valores conhecidos, de cinco variáveis distintas e independentes: número de ligações hidrogênio, porcentagem de superfície acessível ao solvente, número de ligações torcionáveis congeladas durante a interação da proteína com o ligante e as energias de Lennard-Jones e eletrostática e a constante de afinidade como variável dependente a ser descoberta.

Supondo que os pontos têm coordenadas (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), onde **x** é a variável independente e *y* é a variável dependente, a curva de ajuste $f(\boldsymbol{x})$ tem o desvio (erro) *d* de cada ponto de dado, isto é, $d_1 = y_1 - f(x_1)$, $d_2 = y_2 - f(x_2)$, ..., $d_n = y_n - f(x_n)$. De acordo com o método dos mínimos quadrados, a melhor curva de ajuste minimiza:

$$II = d_1^2 + d_2^2 + \ldots + d_n^2 = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 \qquad (4.4)$$

A regressão múltipla estima a saída (variáveis dependentes) as quais podem ser afetadas por mais de um parâmetro de controle (variáveis independentes) (de Barros Neto et al., 1995).

Para exemplificar, consideremos uma função f(x,y) = z com duas variáveis independentes *x* e *y* e uma variável dependente *z* em um caso de relação linear:

$$z = a + bx + cy \tag{4.5}$$

onde as constantes a, b e c devem ser determinadas.

Para um determinado conjunto de dados (x_1, y_1, z_1), (x_2, y_2, z_2), ..., (x_n, y_n, z_n), onde $n \geq 3$, a melhor curva de ajuste $f(x)$ tem um erro quadrático mínimo, isto é,

$$II = min \sum_{i=1}^{n} [z_i - f(x_i, y_i)]^2 = min \sum_{i=1}^{n} [z_i - (a + bx_i + cy_i)]^2 \tag{4.6}$$

É importante notar que a, b e c são coeficientes desconhecidos enquanto todos os x_i, y_i e z_i são dados. Para obter o erro quadrático mínimo, os coeficientes desconhecidos a, b e c devem anular as primeiras derivadas do erro quadrático mínimo em relação a a, b e c (de Barros Neto et al., 1995).

$$\begin{cases} \frac{\partial_{II}}{\partial_a} = 2\sum_{i=1}^{n}[z_i - (a + bx_i + cy_i)] = 0 \\ \frac{\partial_{II}}{\partial_b} = 2\sum_{i=1}^{n} x_i[z_i - (a + bx_i + cy_i)] = 0 \\ \frac{\partial_{II}}{\partial_c} = 2\sum_{i=1}^{n} y_i[z_i - (a + bx_i + cy_i)] = 0 \end{cases} \tag{4.7}$$

Expandindo as equações acima, nós temos

$$\begin{cases} \sum_{i=1}^{n} z_i = a\sum_{i=1}^{n} 1 + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i z_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i z_i = a\sum_{i=1}^{n} y_i + b\sum_{i=1}^{n} x_i y_i + c\sum_{i=1}^{n} y_i^2 \end{cases} \tag{4.8}$$

74

Os coeficientes *a*, *b* e *c* podem, desta forma, ser obtidos pela resolução do sistema linear de equações mostrado acima (de Barros Neto et al., 1995).

Utilizamos MATLAB, versão 7.0.0.19901 (R14), para a construção de funções de regressão múltipla para a avaliação das variáveis de entrada utilizadas em nosso trabalho. Tratamos aqui, nas análises feitas com a regressão múltipla, com termos lineares, termos cruzados e termos quadráticos dos coeficientes de entrada do nosso problema.

4.5. Análise Estatística

Dois diferentes critérios de avaliação foram utilizados: o coeficiente de correlação linear e o erro de generalização (pág. 54 e 53). O valor *r* chamado de coeficiente de correlação linear é utilizado no programa MATLAB, adotado em nosso trabalho, como critério de avaliação da qualidade dos modelos gerados pela rede neural, uma vez que ele calcula a distância e a direção de uma relação linear entre duas variáveis, no caso os valores observados e os valores esperados.

Seu valor é determinado da seguinte forma:

$$r = \frac{[n\sum_{n=1}^{N} xy - (\sum_{n=1}^{N} x)(\sum_{n=1}^{N} y)]}{[\sqrt{n(\sum_{n=1}^{N} x^2) - (\sum_{n=1}^{N} x)^2} \times \sqrt{n(\sum_{n=1}^{N} y^2) - (\sum_{n=1}^{N} y)^2}]}, \qquad (4.9)$$

onde *n* é o número de pares de dados, *x* é o vetor de valores esperados e *y*, o vetor de valores observados correspondentes.

O valor de *r* varia entre $-1 \le r \le +1$. Os sinais de + e – correspondem a correlações lineares positivas e negativas, respectivamente.

Se x e y tem uma correlação linear positiva forte, r se aproxima de +1. Uma relação maior que 0,8 é geralmente descrita como forte, enquanto uma correlação menor que 0,5 é geralmente descrita como fraca.

CAPÍTULO 5

5. RESULTADOS E DISCUSSÃO

5.1. Caracterização do Conjunto-Teste

A primeira etapa deste trabalho foi a seleção e caracterização de 50 complexos com dados de cristalografia de difração de raios-X disponíveis no banco de dados de proteína (PDB – "*Protein Data Bank*"), e constantes de afinidade receptor-ligante determinadas experimentalmente (obtidas do sítio web AffinDB, Block et al., 2006). A partir das estruturas destes complexos, após o processo de minimização, foram calculados os valores de energia de Lennard-Jones; eletrostática; porcentagem da SAS do ligante que interage com a proteína; número de possíveis ligações hidrogênio existentes entre as duas moléculas e o número de ligações torcionáveis congeladas durante a interação do ligante com a proteína.

Tabela 5.1: Estruturas dos complexos proteína-ligante utilizadas na construção das funções empíricas para o cálculo da constante de afinidade.

A	B (%)	C (kcal/mol)	D (kcal/mol)	E	F	G	H
1ADD	95,745	-11,7465509	-35,3714197	5	2	2	6,74
1AI5	98,112	-9,02132321	-28,8863093	3	3	3	3,72
1AZM	90,43	-10,5620956	-26,7606016	3	3	1	6,04
1C83	81,898	-26,7180433	-26,3242276	8	4	4	4,85
1CBX	96,09	-66,9633041	-20,9482778	5	5	3	6,35
1CIL	85,098	-4,87789937	-40,2326366	6	3	2	9,43
1CPS	91,906	-16,3356861	-17,2043904	6	5	3	6,66
1DBB	77,025	-2,59077191	-37,7909913	3	1	1	9
1DBJ	83,4	-2,42038463	-39,617435	2	0	0	7,68
1DHF	81,414	-10,6225108	-45,1231975	6	9	9	7,4
1EBG	43,492	-31,4298116	-1,40730029	14	2	2	10,82
1EJN	74,18	-1,1356101	-30,1081015	7	4	4	5,61
1F0R	75,037	-6,94044015	-36,6489326	1	5	5	7,65
1FKF	51,116	-9,6401639	-56,2871929	4	7	0	9,7
1FL3	92,671	-0,394795861	-53,3212043	1	7	7	6,79
1H4N	67,285	0,69389769	-13,9138893	6	3	3	4,92
1HRI	95,559	1,08764133	-37,2412721	0	9	9	4,34
1HYT	92,053	-50,4709748	-24,0934075	5	5	3	5,2
1L82	72,016	-1,62378557	-13,5036381	0	7	1	4,85
1L86	66,993	-1,34344846	-13,3838601	0	7	1	3,37
1MDR	95,258	-3,97625311	-22,1768603	10	2	1	7,52
1MMQ	62,058	-5,12899775	-45,3502514	0	6	4	6,16
1MNC	69,991	-40,4084002	-40,8401273	10	9	5	9
1OYN	92,54	-1,25295313	-34,7499003	1	4	4	6,49
1PDZ	83,198	-9,59992249	-18,8430238	8	3	3	2,2
1RBP	97,323	0,106648394	-35,8010669	7	6	6	4,31
1RGL	59,821	-29,5932135	-25,0541675	4	4	3	6,22
1RNT	70,298	-16,7810048	-34,5748647	8	4	3	5,19
1SNC	75,451	-63,9970303	-19,9846242	8	7	7	6,7
1SRE	91,505	-12,0548911	-19,4587617	4	4	4	3,85
1TNG	90,141	-1,7858301	-16,4877397	3	1	0	2,93
1TNH	90,455	-2,26381273	-21,1884214	3	2	2	3,37
1TNJ	92,192	-0,577721388	-19,3900576	0	2	0	1,96

Continuação da Tabela 5.1.

1TNK	89,546	-0,950784264	-18,516278	2	3	0	1,49
1ULB	96,884	-10,3534358	-20,8277477	3	0	0	5,3
2AK3	83,082	-21,9590722	-34,6154228	6	4	4	3,86
2CTC	98,06	-27,9140836	-22,9057349	2	3	1	3,89
2H4N	79,132	-2,45686392	-21,7713857	6	3	3	8,69
2IFB	97,118	-8,0157996	-40,6280036	2	14	14	5,43
2PCP	92,939	-1,14863228	-36,9889989	1	2	1	8,69
2SIM	86,089	-20,4268431	-29,4199358	10	5	4	3,42
2SNS	69,197	-57,706666	-29,2236281	2	5	4	6,7
2XIS	96,604	-16,1590488	-16,5319142	9	4	4	5,82
2YPI	24,144	-5,62697136	-3,29275871	7	3	3	4,82
3PTB	88,3	-0,555652415	-20,3326639	5	1	0	4,74
4XIA	95,753	-20,7042542	-17,8391045	12	5	4	1,54
5ENL	86,431	-12,285367	-20,9232525	7	4	4	3,8
6RNT	58,582	-27,0775474	-25,5731435	5	4	4	2,37
6TIM	23,863	-4,85331693	-5,80507889	6	5	5	3,21
7TIM	95,754	-24,8020066	-19,0222371	8	3	3	5,4

A – Código PDB;
B – porcentagem da superfície total do ligante em contato com a proteína;
C – energia de Lennard-Jones;
D – energia eletrostática;
E – número de possíveis ligações hidrogênio existentes entre a proteína e o ligante;
F – número de ligações torcionáveis;
G – número de ligações torcionáveis congeladas e
H – - log K_i ou - log K_d.

Com a definição do conjunto-teste (Tabela 5.1) avaliamos o número de famílias de proteínas nele presentes, bem como o número de tipos de enovelamento destas mesmas proteínas, nos dando idéia do número de possíveis diferentes sítios ativos que estávamos tratando neste conjunto. Utilizamos para esta avaliação a classificação disponibilizada no sítio do banco de dados do PDB que utiliza, para tanto, as informações do sítio de classificação estrutural de proteínas - SCOP ("*Structural Classification of Proteins*"). A classificação é baseada nas relações evolutivas e em características que descrevem suas estruturas tri-dimensionais, como a disposição de suas estruturas secundárias e a topologia de suas cadeias de polipeptídeos. São consideradas de uma

mesma família, proteínas que apresentam uma origem evolutiva comum, com pelo menos 30 % de seus resíduos idênticos ou proteínas com uma identidade de seqüência menor que esta, mas com funções e estruturas muito similares, como acontece na família das proteínas globinas, com 15% de identidade de seqüência (Murzin et al., 1995).

Ainda em relação à classificação do SCOP, utilizamos a categoria de tipo de enovelamento para análise dos complexos selecionados para o nosso conjunto-teste. Superfamílias e famílias são definidas como apresentando um mesmo tipo de enovelamento se suas proteínas possuem as mesmas estruturas secundárias em um mesmo arranjo, com as mesmas conexões topológicas. Para proteínas colocadas juntas em uma mesma categoria de enovelamento, as similaridades estruturais provavelmente aparecem devido à física e química das proteínas que favorecem certos arranjos e topologias de cadeia. Podem existir, no entanto, casos onde uma origem evolutiva comum é ofuscada devido à divergência na seqüência, estrutura e função de uma determinada proteína (Murzin et al., 1995).

Utilizando este critério, foram identificadas 27 famílias diferentes de proteínas para os 50 complexos proteína-ligante do conjunto-teste. O conjunto-teste apresentou complexos proteína-ligantes distribuídos em 22 diferentes categorias de tipos de enovelamento diferentes. Estas informações podem ser observadas na Tabela 5.2 abaixo.

Tabela 5.2: Estruturas de proteínas utilizadas na construção das funções empíricas para o cálculo da constante de afinidade.

A	B	C
1ADD	Beta TIM/Alfa-Barril	Adenosina Deaminase (ADA)
1AI5	Nitrogênio Hidrolase	Penicilina Acilase (Domínio Catalítico)
1AZM	Anidrase Carbônica	Anidrase Carbônica
1C83	Proteína Fosfotirosina Fosfatase II	Proteína Fosfotirosina Fosfatase de Alto Peso Molecular
1CBX	Fosforilase/Hidrolase	Carboxipeptidase Pancreática
1CIL	Anidrase Carbônica	Anidrase Carbônica

Continuação da Tabela 5.2.

1CPS	Fosforilase/Hidrolase	Carboxipeptidase Pancreática
1DBB	Imunoglobulina do tipo Sanduíche	Conjunto de Domínios do tipo V (Domínio do tipo anticorpo variável)
1DBJ	Imunoglobulina do tipo Sanduíche	Conjunto de Domínios do tipo V (Domínio do tipo anticorpo variável)
1DHF	Diidrofolato Redutase	Diidrofolato Redutase
1EBG	Beta TIM/Alfa-Barril	Enolase
1EJN	Serino Protease do Tipo Tripsina	Proteases Eucariótica
1F0R	Serino Protease do Tipo Tripsina	Proteases Eucariótica
1FKF	FKBP	Imunofilina FKBP/Prolina Isomerase
1FL3	Imunoglobulina do tipo Sanduíche	Conjunto de Domínios do tipo V (Domínio do tipo anticorpo variável)
1H4N	Anidrase Carbônica	Anidrase Carbônica
1HRI	Nucleoplasmina/VP (Proteínas do revestimento viral e capsídeo)	VP do tipo Picornavaridae (VP1, VP2, VP3 e VP4)
1HYT	Zincina	Termolisina
1L82	Lisozima	Lisozima fagocitária
1L86	Lisozima	Lisozima fagocitária
1MDR	Beta TIM/Alfa-Barril	D-gluconato deidratase
1MMQ	Zincina	Matriz Metaloprotease (Domínio Catalítico)
1MNC	Zincina	Matriz Metaloprotease (Domínio Catalítico)
1OYN	Domínio HD	PDEase
1PDZ	Beta TIM/Alfa-Barril	Enolase
1RBP	Lipocalina	Proteína do tipo ligante de Retinol
1RGL	Ribonuclease Microbiana	Ribonuclease Fúngica
1RNT	Ribonuclease Microbiana	Ribonuclease Fúngica
1SNC	Enovelamento OB	Nuclease Staphylococcal
1SRE	Streptavidina	Streptavidina/Avidina
1TNG	Serino Protease do Tipo Tripsina	Protease Eucariótica
1TNH	Serino Protease do Tipo Tripsina	Protease Eucariótica
1TNJ	Serino Protease do Tipo Tripsina	Protease Eucariótica
1TNK	Serino Protease do Tipo Tripsina	Protease Eucariótica
1ULB	Fosforilase/Hidrolase	Fosforilase Purina e Uridina
2AK3	Nucleosídeo Trifosfato Hidrolase contendo a Volta P	Nucleotídeo e Nucleosídeo Quinases
2CTC	Fosforilase/Hidrolase	Carboxipeptidase Pancreática
2H4N	Anidrase Carbônica	Anidrase Carbônica
2IFB	Lipocalina	Proteína do tipo ligante de Ácidos Graxos

Continuação da Tabela 5.2.

2PCP	Imunoglobulina do tipo Sanduíche	Conjunto de Domínios do tipo V (Domínio do tipo anticorpo variável)
2SIM	Hélice Beta 6-Laminada	Sialidase (neuroaminidases)
2SNS	Enovelamento OB	Nuclease Staphylococcal
2XIS	Beta TIM/Alfa-Barril	Xilose Isomerase
2YPI	Triose-Fosfato Isomerase (TIM)	Triose-Fosfato Isomerase (TIM)
3PTB	Serino Protease do Tipo Tripsina	Protease Eucariótica
4XIA	Beta TIM/Alfa-Barril	Xilose Isomerase
5ENL	Beta TIM/Alfa-Barril	Enolase
6RNT	Ribonuclease Microbiana	Ribonuclease Fúngica
6TIM	Beta TIM/Alfa-Barril	Triose-Fosfato Isomerase (TIM)
7TIM	Beta TIM/Alfa-Barril	Triose-Fosfato Isomerase (TIM)

A – Código PDB;

B – Tipo de Enovelamento;

C – Família.

Normalmente, no cálculo da energia livre de ligação entre um ligante e uma proteína, é levado em consideração o número de ligações torcionáveis congeladas durante esse evento, estimando a perda entrópica que ocorre quando ligações acíclicas, simples, do ligante tornam-se não rotacionáveis quando estão em contato com a proteína, como já descrito anteriormente.

O número de ligações torcionáveis afeta diretamente a dimensionalidade do espaço de busca do problema de atracamento e, como é esperado, afeta enormemente o desempenho das ferramentas de atracamento. Um exemplo disto pode ser observado nos dados obtidos com a ferramenta GlamDock, onde o impacto do aumento da flexibilidade na porcentagem de erros da amostra é imensamente pronunciado. Aproximadamente 10% dos ligantes com menos de 13 ligações rotacionáveis apresentaram RMSD > 2.0 Å em relação à estrutura determinada experimentalmente, sendo que esta porcentagem subiu para 40% dos ligantes que tinham 13 ou mais ligações

torcionáveis (Tietze e Apostolakis, 2007).

O conjunto-teste aqui utilizado apresenta no máximo 14 ligações torcionáveis congeladas. Nas Figuras 5.1 e 5.2, mostramos, respectivamente, o número de ligações torcionáveis e quantas destas são congeladas durante o processo de atracamento do ligante na proteína.

Figura 5.1: Número de complexos proteína-ligante em relação ao número de ligações torcionáveis do ligante.

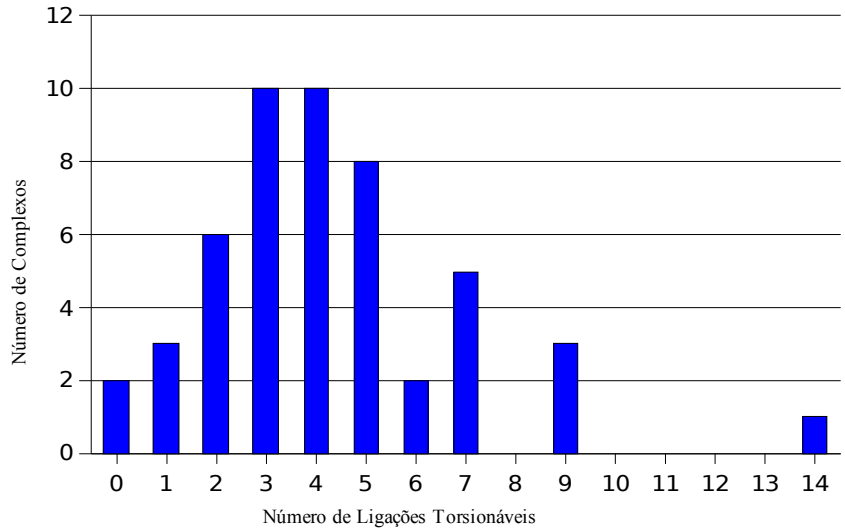
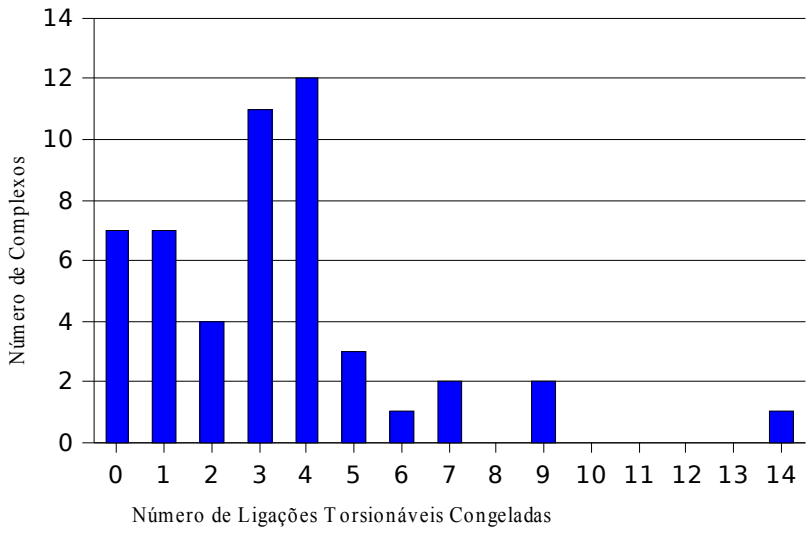


Figura 5.2: Número de complexos proteína-ligante em relação ao número de ligações torcionáveis congeladas do ligante durante o processo de atracamento à proteína.



Neste trabalho, a presença de grupamentos carregados do ligante formando ligações hidrogênio não recebeu tratamento especial. Ou seja, as ligações hidrogênio foram contabilizadas independentemente do fato de envolverem ou não grupos carregados da proteína ou do ligante.

A presença destas ligações hidrogênio pode estar embutida no valor de energia eletrostática, o que foi observado apenas na presença de ligantes com cargas -4 (Figura 5.3), mas que não ficou evidente para os outros ligantes não neutros. Na Figura 5.4, é mostrado o número de complexos proteína-ligante em relação à carga de seus ligantes.

Figura 5.3: Carga total dos ligantes e valores de energia eletrostática calculada no atracamento do ligante com a proteína.

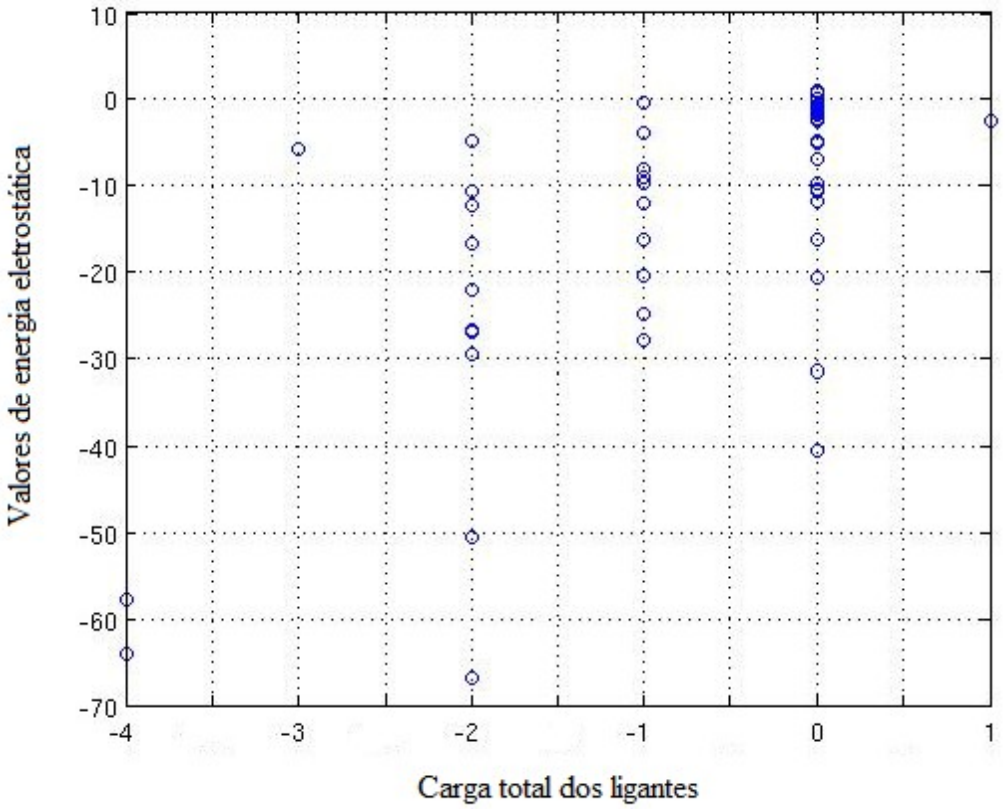
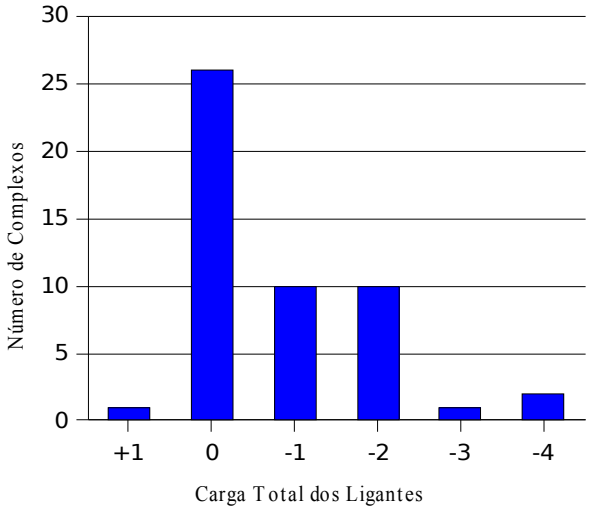


Figura 5.4: Número de complexos proteína-ligante em relação à carga total dos ligantes.



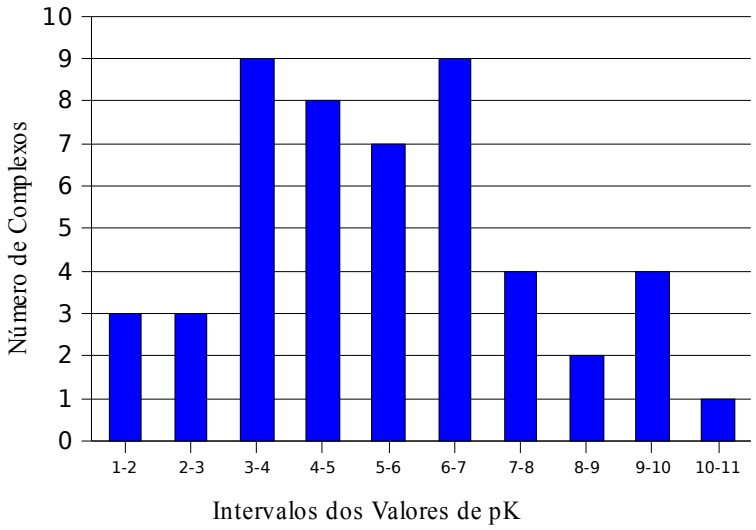
O número de possíveis ligações hidrogênio formadas entre o ligante e a proteína, sem levar em consideração as cargas de ambos, como já foi explicado, variou de 0 a 14. As energias eletrostática e de Lennard-Jones variaram entre 1,088 kcal/mol e -66,963 kcal/mol, e -1,407 kcal/mol e -56,287 kcal/mol, respectivamente e a porcentagem de superfície do ligante em contato com a proteína variou de 23,86% a 98,06%.

A determinação acurada e com um custo computacional razoável da afinidade de ligação (K_d ou K_i) entre ligantes e seus receptores, seria de enorme benefício para a área de desenho racional de fármacos. Este fato possibilitaria a predição da afinidade de um grande número de compostos antes que eles fossem quimicamente sintetizados, agilizando o processo como um todo (de Magalhães, 2006). É fundamental no entanto, que nosso conjunto-teste apresente um grande intervalo de valores de constante de afinidade, possibilitando que a função avaliação consiga discernir entre ligantes interessantes e não tão interessantes para uma determinada proteína.

Em nosso trabalho, as constantes de dissociação foram tratadas em unidades de -log K_d ou -log K_i, que variavam de 1,49 a 10,82, abarcando aproximadamente 9 ordens de magnitude. É

importante lembrar que -log K igual a 1,49 representa uma inibição da proteína com a concentração de milimolar (10^{-3} mol/L) do ligante , enquanto -log K igual a 10,82 significa uma concentração do ligante em nanomolar (10^{-9} mol/L). A distribuição dos complexos proteína entre os diferentes intervalos de pK, é mostrada na figura 5.5, onde a maioria dos complexos do conjunto teste apresenta pK entre 3 e 7.

Figura 5.5: Número de complexos proteína-ligante em relação aos diferentes intervalos de pK (-log K_d ou -log K_i) no processo de atracamento destas moléculas.



Sob os aspectos de diversidade do ligante, famílias, cargas, Ki, entre outras características, o conjunto teste utilizado em nosso trabalho possui uma grande variabilidade e, portanto, em princípio seria um bom conjunto teste para a construção de uma função empírica de caráter geral (não família ou ligante dependente).

5.2. Construção de uma Função Empírica por Regressão Múltipla

Além da construção da função empírica para a determinação da constante de afinidade entre receptor-ligante realizada com a metodologia de rede neural, foi construído também um segundo conjunto de funções encontradas através do método de regressão múltipla, cujos coeficientes podem ser computados a partir do ajuste dos mínimos quadrados.

Em uma rede neural o conjunto de pesos multiplicado pelas variáveis de entrada é relativamente grande, devido ao número de neurônios necessários para um ajuste ideal para os valores de saída. A escolha destes pesos não nos fornece uma função onde os coeficientes podem ser analisados nos dando informações sobre a importância das variáveis de entrada para a resposta ideal. Na construção de uma função empírica por regressão múltipla, é possível estimar a importância das variáveis utilizadas através da análise dos coeficientes a elas associados.

Na etapa de construção das funções empíricas por regressão múltipla, utilizando os dados extraídos dos 50 complexos receptor-ligante, foram testados distintos conjuntos de tipos de variáveis dependentes, mantendo-se sempre, em cada conjunto, as variáveis relativas à energia eletrostática e à energia do termo de Lennard-Jones. Utilizando estes conjuntos, distintos tipos de funções polinomiais foram testados: (1) com termos unicamente lineares, (2) com termos lineares e cruzados, e (3) com termos lineares, cruzados e quadráticos.

As variáveis de entrada para a regressão foram as mesmas cinco utilizadas na construção de uma função para a determinação da constante de afinidade entre proteína-ligante utilizando rede neural. Foram testados oito distintos conjuntos de variáveis independentes:

Tabela 5.3. Variáveis independentes de entrada para a construção das funções para a determinação do valor de afinidade entre proteína-ligante utilizando a metodologia de regressão.

Conjuntos utilizados na construção das funções	Váriaveis independentes
(i)	A , B
(ii)	A, B, C
(iii)	A, B, D
(iv)	A, B, E
(v)	A, B, C, E
(vi)	A, B, D, E
(vii)	A, B, C, D
(viii)	A, B, C, D, E

A – Energia eletrostática;
B – Energia de Lennard-Jones;
C – Porcentagem da superfície total do ligante em contato com a proteína;
D – Número de possíveis ligações existentes entre a proteína e o ligante;
E – Possíveis ligações torcionáveis congelados.

A avaliação das funções obtidas foi feita utilizando o valor do coeficiente de correlação linear, *r* e o erro de generalização, *Err*.

Tabela 5.4: Coeficiente de correlação, *r* e o erro de generalização, *Err*. Cada uma das oito combinações com as 5 variáveis de entrada: energias eletrostática e de Lennard-Jones, número de ligações torcionáveis congeladas, porcentagem da superfície total do ligante em contato com a proteína e número de possíveis ligações hidrogênio existentes entre a proteína e o ligante em funções com termos.

	(1)		**(2)**		**(3)**	
	r	***Err***	***r***	***Err***	***r***	***Err***
(i)	0,4345	0,1819	0,5026	0,1675	0,5870	0,1469
(ii)	0,5002	0,1681	0,5864	0,1471	0,6253	0,1365
(iii)	0,4835	0,1718	0,5823	0,1482	0,6327	0,1344
(iv)	0,4618	0,1764	0,5418	0,1584	0,6242	0,1368
(v)	0,5270	0,1619	0,6467	0,1304	0,7049	0,1128
(vi)	0,5186	0,1639	0,6521	0,1288	0,6954	0,1158
(vii)	0,5348	0,1601	0,6386	0,1328	0,6863	0,1186
(viii)	0,5686	0,1517	0,7127	0,1103	0,7542	0,0967

(1) unicamente lineares,

(2) com termos lineares e cruzados e por fim,

(3) com termos lineares, cruzados e quadráticos.

(i) Energias eletrostática, Lennard-Jones; (ii) Energias eletrostática, Lennard-Jones e porcentagem da superfície total do ligante em contato com a proteína; (iii) Energias eletrostática, Lennard-Jones e número de possíveis ligações hidrogênio existentes entre a proteína e o ligante; (iv) Energias eletrostática, Lennard-Jones e número de ligações torcionáveis congeladas; (v) Energias eletrostática, Lennard-Jones, número de ligações torcionáveis congeladas e porcentagem da superfície total do ligante em contato com a proteína; (vi) Energias eletrostática, Lennard-Jones, número de ligações torcionáveis congeladas e o número de possíveis ligações hidrogênio existentes entre a proteína e o ligante; (vii) Energias eletrostática, Lennard-Jones, porcentagem da superfície total do ligante em contato com a proteína e o número de possíveis ligações hidrogênio existentes entre a proteína e o ligante e (viii) Energias eletrostática, Lennard-Jones, número de ligações torcionáveis congeladas, porcentagem da superfície total do ligante em contato com a proteína e número de possíveis ligações hidrogênio existentes entre a proteína e o ligante.

É importante lembrar, inicialmente, que o programa DOCKTHOR tem sua função objetivo de energia livre derivada apenas dos valores de energias eletrostática e de Lennard-Jones. Mesmo em um polinômio com termos quadráticos, o valor de *r* é de 0,5870, mostrando a importância de novos termos que possam resultar em melhores resultados na avaliação das conformações dos ligantes no processo de atracamento molecular. No item (i)-(A) com *r* = 0.4245 temos uma estimativa do poder de previsão da função energia atual do DOCKTHOR. A baixíssima previsibilidade mostra a importância de se melhorar a função avaliação no programa.

Analisando o valor de *r* podemos ter uma noção da importância de algumas das variáveis de entrada na busca do valor da constante de afinidade. Na inclusão de mais um termo, além das energias eletrostática e de Lennard-Jones, vemos que na função de termos lineares, e lineares e cruzados, a variável que resultou em melhores valores foi a porcentagem da superfície total do ligante em contato com a proteína, com valor de *r* de 0,5864, que representa, de alguma forma, a energia de solvatação do processo de atracamento do ligante na proteína. Na inserção de um termo quadrático, no entanto, o melhor resultado (*r* = 0,6327), foi obtido com o número de possíveis ligações hidrogênio entre a proteína e o ligante, e em muitos casos.

A importância da inclusão de termos que não lineares em modelos para cálculo de energia livre pode ser bem observada no trabalho de (Oliveira et al., 2006) onde foi utilizado um termo de dependência quadrática na energia livre de solvatação do ligante, que provavelmente reflete o caráter dual hidrofílico/hidrofóbico do sítio ativo da enzima fosfodiesterase 4 (PDE4). O modelo empírico de energia livre desenvolvido pelos autores resultou em uma correlação com coeficiente de determinação, $r^2 = 0,92$ entre os dados experimentais e teóricos.

A função construída com a metodologia de regressão múltipla que obteve melhor resultado foi a que utilizou as 5 variáveis de entrada apresentando termos lineares, cruzados e quadráticos, com *r* igual a 0,7542.

Gráfico 5.1: Gráfico dos valores reais de pK e os melhores valores obtidos.

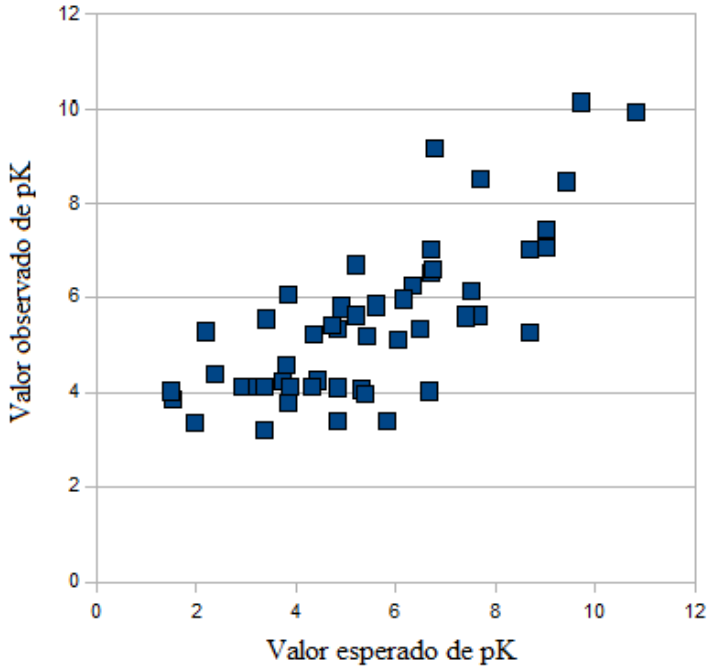


Gráfico dos valores reais de pK (eixo y) dos 50 complexos proteína-ligante do conjunto-teste utilizado neste trabalho (em laranja, e os melhores valores obtidos (em azul) com o polinômio com 5 variáveis de entrada com termos lineares, cruzados e quadráticos.

O polinômio resultante da regressão com as 5 variáveis de entrada pode ser observado a seguir:

$$pK = - 0{,}2566 - 0{,}1837\ ELE + 0{,}1924\ LJ - 0{,}0229\ LH + 0{,}4157\ SAS + 0{,}0737\ NTOR + 0{,}0408\ (ELE \times LJ) + 0{,}1557\ (ELE \times LH) - 0{,}1337\ (ELE \times SAS) - 0{,}1202\ (ELE \times NTOR) + 0{,}0771\ (LJ \times LH) - 0{,}6807\ (LJ \times SAS) + 0{,}9390\ (LJ \times NTOR) - 0{,}3186\ (LH \times SAS) - 0{,}7114\ (LH \times NTOR) - 0{,}5083\ (SAS \times NTOR) + 0{,}1987\ (ELE)^2 + 0{,}8455\ (LJ)^2 + 0{,}0711\ (LH)^2 - 0{,}7834\ (SAS)^2 + 0{,}3585\ (NTOR)^2 \qquad (5.1)$$

Podemos observar o alto valor dos coeficientes associados aos termos lineares da variável SAS (superfície acessível ao solvente), seguida em importância pelas variáveis ELE (eletrostático) e LJ (Lennard-Jones). Nos termos quadráticos observamos a maior importância das variáveis LJ, SAS e NTOR (número de ligações torcionáveis congeladas). As variáveis NTOR e LH (Número de ligações hidrogênio) apresentam uma contribuição importante nos termos cruzados. Deste modo, observamos a importância das 3 novas variáveis introduzidas (SAS, NTOR e LH) no cálculo da constante de afinidade receptor-ligante.

Utilizando esta função para o cálculo da constante de afinidade, obtivemos 20% dos complexos proteína-ligante com um erro menor que 1,00, representando uma diferença de concentração do ligante da ordem de 10^{-1} em relação a encontrada durante a medição do valor esperado de pK. 56% dos complexos apresentaram erros menores que 2,0 e 72% dos complexos, erros menores que 3,0, representando uma diferença de concentração do ligante da ordem de 10^{-3}, lembrando que de qualquer forma, ainda devemos levar em consideração o erro experimental embutido nos nossos dados.

Tabela 5.5: Estruturas de proteínas utilizadas na construção das funções empíricas para o cálculo da constante de afinidade.

Código PDB do complexo	Valor esperado	Valor absoluto da diferença entre o valor esperado e o valor observado
1ADD	6,74	3,6424
1AI5	3,72	6,4177
1AZM	6,04	1,6574
1C83	4,85	0,7094
1CBX	6,35	1,0745
1CIL	9,43	2,4186
1CPS	6,66	1,8507
1DBB	9	3,6424
1DBJ	7,68	0,6389
1DHF	7,4	4,0107
1EBG	10,82	6,7883
1EJN	5,61	0,3087
1F0R	7,65	1,8579
1FKF	9,7	4,4869
1FL3	6,79	2,6609
1H4N	4,92	4,2439
1HRI	4,34	2,6754
1HYT	5,2	0,3990
1L82	4,85	1,8362
1L86	3,37	0,0194
1MDR	7,52	3,2802
1MMQ	6,16	1,9095
1MNC	9	3,3721
1OYN	6,49	1,9048
1PDZ	2,2	1,8226
1RBP	4,31	1,3099
1RGL	5,43	1,3164
1RNT	5,19	1,0854
1SNC	6,7	3,3104
1SRE	3,85	2,2689
1TNG	2,93	1,1364
1TNH	3,37	2,4599

Continuação da Tabela 5.5.

1TNJ	1,96	4,6200
1TNK	1,49	1,7143
1ULB	5,3	1,1720
2AK3	3,86	4,5961
2CTC	3,89	1,3494
2H4N	8,69	2,6429
2IFB	4,43	0,5808
2PCP	8,69	2,4409
2SIM	3,42	0,7039
2SNS	6,7	1,5764
2XIS	5,82	0,3967
2YPI	4,82	5,1216
3PTB	4,74	0,7781
4XIA	1,54	4,4200
5ENL	3,8	2,7047
6RNT	2,37	5,0671
6TIM	3,21	0,5733
7TIM	5,4	1,2759

A – Código PDB;

B – Valor esperado;

C – Valores absolutos das diferenças entre os valores esperados e os valores observados calculados a partir da equação 5.1.

O maior erro observado no valor do pK foi de 6,7883, onde o valor esperado era igual a 10,82, representando uma concentração do ligante da ordem de nanomolar, chegando quase a picomolar (10^{-12}), evento muito raro de ser observado. De qualquer forma, como nosso conjunto-teste continha apenas um único complexo proteína-ligante com valor de afinidade deste tipo, é difícil enquadrar esta situação em um modelo.

Dos 50 complexos proteína-ligante utilizados na construção da função empírica para o

cálculo da constante de afinidade, 14 (28%) apresentaram valor absoluto maior que 3,0 na diferença entre o valor esperado e o valor obtido. Podemos esperar esse tipo de problema uma vez que muitas das enzimas em questão apresentam em seu sítio-ativo moléculas de água e mesmo íons metálicos, que participam diretamente na interação da proteína com o ligante. Esse tipo de interação não é incluída em nenhuma das variáveis de entrada de nossa função avaliação. Por isso, é necessário mencionar aqui a importância das águas (principalmente estruturais) na estimação das constantes de afinidade. Ou seja, a preparação do conjunto de treinamento é crucial e nenhuma técnica de construção empírica corrige este problema.

Em alguns dos maiores erros observados na Tabela 5.5, na diferença dos valores esperados e observados de pK nas proteínas com código PDB 1AI5, 1EBG, 1FKF, 1H4N, 1TNJ, 2YPI e 4XIA, observamos que estas apresentam valores extremos de alguma das variáveis de entrada, SAS, LH, LJ e ELE e também nos valores de saída, constatando o problema de valores extremos nas variáveis dependentes e independentes na construção de funções empíricas por regressão múltipla. Tanto nas proteínas 1AI5 e 2YPI, observamos o maior e o menor valor do termo SAS, comprovando mais uma vez, a importância dessa variável no cálculo da constante de afinidade.

Em dois dos casos citados acima, 2YPI e 1TNJ, observamos valores de entrada semelhantes aos de outros dois complexos, 6TIM e 1TNK, respectivamente. No entanto, o baixo valor do erro na estimativa do valor da constante de afinidade, pode estar relacionado à mudança de alguma outra variável que pode ser mais importante no processo de interação com os ligantes nestas proteínas, no primeiro caso, o número de ligações torcionáveis foi maior na proteína 6TIM e na proteína 1TNK, o número de ligações hidrogênio formadas no processo de atracamento molecular foi maior.

5.3. Construção de uma Função Empírica por Rede Neural

Na construção da função empírica utilizando rede neural, diferentes abordagens foram

realizadas. A primeira delas foi a construção de uma rede neural utilizando a metodologia de validação de treinamento e teste. Neste experimento, avaliamos modelos gerados durante o treinamento com conjuntos contendo 25, 35 e 45 complexos proteína-ligante, com a validação (no conjunto teste respectivo de 25, 15 e 5 complexos) sendo realizada utilizando o melhor conjunto de pesos obtidos no treinamento. Nesta abordagem foi investigado o desempenho da rede com relação ao número de neurônios utilizados na camada oculta.

A segunda abordagem utilizou a metodologia de validação cruzada de grupo para estimar o erro de generalização de uma rede neural construída utilizando-se o conjunto inteiro de 50 complexos proteína-ligante. Nesta abordagem também foi investigado o desempenho da rede com relação ao número de neurônios utilizados na camada oculta.

5.3.1 Metodologia de Treinamento e Teste

Como podemos observar nas Tabelas 5.6, 5.7 e 5.8 existe uma relação entre os valores de correlação e de erro do conjunto teste e o número de neurônios da camada oculta. Os testes foram realizados com 3, 6 e 9 neurônios na camada oculta e com os conjuntos treinamentos com 45, 35 e 25 complexos proteína-ligante e os conjuntos testes com 5, 15 e 25 complexos, respectivamente.

A rede neural com 9 neurônios na camada oculta apresenta o menor erro de treinamento (calculado utilizando a fórmula 2.18) para as 3 partições de dados investigadas. Este resultado é esperado, pois com o aumento da complexidade do modelo é natural obter-se um menor erro de treinamento. Para a partição de dados 25/25 (Tabela 5.6) o menor erro de teste foi obtido utilizando-se 9 neurônios na camada oculta. Para a partição de dados 35/15 (Tabela 5.7) o menor erro de teste foi obtido utilizando-se 6 neurônios na camada oculta. Para a partição de dados 45/5 (Tabela 5.8) o

menor erro de teste foi obtido utilizando-se 3 neurônios na camada oculta. Ou seja, para os casos 35/15 e 45/5, o modelo com menor erro de treinamento não possui o menor erro de teste (superajustamento).

O superajustamento (“overfitting”), é um dos problemas que ocorrem comumente durante o treinamento de uma rede neural. O erro do conjunto treinamento é um valor muito pequeno, mas quando se testa a rede com novos dados, o erro aumenta consideravelmente. A rede memorizou os exemplos do treinamento, embora não tenha aprendido a generalizar as novas situações.

Um método para melhorar a generalização da rede é utilizar um conjunto de dados que seja grande o bastante para fornecer um ajuste adequado. Por outro lado, quanto maior o número de camadas e de neurônios na rede gerada, mais complexas são as funções criadas. Se utilizarmos uma rede com o número de camadas e neurônios suficientemente pequeno, ela não terá poder suficiente para superajustar os dados. Infelizmente, é difícil saber de antemão a complexidade ideal de uma rede para uma dada aplicação.

Se o número de parâmetros na rede é muito menor do que o número total de pontos do conjunto treinamento, há pouca ou nenhuma chance de superajustamento. O que não acontece no nosso problema.

A rede com 3 neurônios na camada oculta treinada na partição de dados 45/5 é a que possui o menor erro de teste. Isto pode refletir o fato de que o modelo foi treinado em um conjunto mais amplo e com uma versão mais simplificada da rede, o que parece minimizar o problema de superajustamento tornando o modelo mais previsível. Mas não se pode desprezar o fato da análise ser dependente do tipo de partição do conjunto treinamento/teste.

Tabela 5.6: Teste utilizando a metodologia de treinamento e teste, onde 25 complexos proteínas ligantes foram utilizados para o treinamento da rede neural e as outras 25 estruturas foram usadas no teste. Redes neurais foram treinadas com a camada oculta apresentando 3, 6 e 9 neurônios.

Neurônios da Camada Oculta	3 neurônios		6 neurônios		9 neurônios	
Estimativas de correlação e erro	r^a	Err^b	*r*	*Err*	*r*	*Err*
Treinamento	0,9057	0,0846	0,9281	0,0494	0,9581	0,0403
Teste	0,2178	0,5852	0,1041	0,7458	0,2978	0,4761

a-Coeficiente de correlação; b- erro de generalização, calculado a partir da fórmula 2.18.

Tabela 5.7: Teste utilizando a metodologia de treinamento e teste, onde 35 complexos proteínas ligantes foram utilizados para o treinamento da rede neural e as outras 15 estruturas foram usadas no teste. Redes neurais foram treinadas com a camada oculta apresentando 3, 6 e 9 neurônios.

Neurônios da Camada Oculta	3 neurônios		6 neurônios		9 neurônios	
Estimativas de correlação e erro	r^a	Err^b	*r*	*Err*	*r*	*Err*
Treinamento	0,9017	0,0542	0,9467	0,0330	0,9667	0,0217
Teste	0,3579	0,5335	0,4696	0,2820	0,2887	0,4602

a-Coeficiente de correlação; b- erro de generalização, calculado a partir da fórmula 2.18.

Tabela 5.8: Teste utilizando a metodologia de treinamento e teste, onde 45 complexos proteínas ligantes foram utilizados para o treinamento da rede neural e as outras 5 estruturas foram usadas no teste. Redes neurais foram treinadas com a camada oculta apresentando 3, 6 e 9 neurônios.

Neurônios da Camada Oculta	3 neurônios		6 neurônios		9 neurônios	
Estimativas de correlação e erro	r^a	Err^b	*r*	*Err*	*r*	*Err*
Treinamento	0,8037	0,1044	0,8738	0,0667	0,8836	0,0630
Teste	0,7540	0,2604	0,5539	0,6165	0,1910	0,7910

a-Coeficiente de correlação; b- erro de generalização, calculado a partir da fórmula 2.18.

Na adoção de uma única função para o cálculo de constantes de afinidade, o número de complexos proteínas-ligantes utilizados para o treinamento deve ser o maior possível. No entanto, novos testes devem ser realizados para definir o número de neurônios da camada oculta. A regra geral para este problema é escolher os parâmetros de forma a encontrar o melhor modelo possível com o menor número deles. Na prática, deve-se de fato experimentar arquiteturas diferentes e comparar os resultados para encontrar o modelo de rede neural mais adequada para o problema em mãos (Wolfram, 1992).

Fica claro que quanto maior o conjunto de treinamento, maior a porcentagem de acerto, levando-nos a crer que quanto maior o conjunto utilizado no treinamento, mais próximos serão os valores observados dos valores esperados (*i.e.*, calculados pela rede). Isso é esperado, uma vez que a rede neural, como uma metodologia de reconhecimento de padrão, busca um mesmo padrão para determinadas faixas nos valores de entrada e de saída e quanto maior o número de casos para um valor de saída e de entrada, melhor a rede estará modelada para diferentes casos com aquele valor observado de pK_i ou pK_d. É também esperado, que em modelos criados com complexos de uma mesma família, a previsibilidade para estruturas deste grupo seja maior, uma vez que o sítio ativo é homólogo e provavelmente os tipos importantes de forças intermoleculares envolvidas no processo de interação sejam mais semelhantes para distintos ligantes.

5.3.2 Metodologia de Validação Cruzada de Grupo

Neste segundo abordagem utilizada corresponde à metodologia de validação cruzada de grupo (VCG) para estimar o erro de generalização de uma rede neural construída utilizando-se o conjunto inteiro de 50 complexos proteína-ligante. Procurou-se também analisar a dependência dos

dados de entrada para a construção da rede, assim como a dependência com o número de neurônios na camada oculta. Para realizar a VCG o conjunto teste total foi particionado em 5 diferentes conjuntos de 10 complexos proteína-ligante. Durante o processo de treinamento, eram utilizados 40 complexos deixando sempre um dos conjuntos selecionados de fora.

Tabela 5.9: Análise de erro de generalização e coeficiente de correlação para distintas partições 40/10 treinamento/teste. Rede neural com camada oculta de 9 neurônios.

Camada oculta de 9 neurônios		
Estimativas de correlação e erro	r^a	Err^b
Teste (Grupo 1)	0,3931	0,4556
Teste (Grupo 2)	0,6207	0,4856
Teste (Grupo 3)	0,4880	0,4335
Teste (Grupo 4)	0,4921	0,4663
Teste (Grupo 5)	-0,0439	0,9037

a-Coeficiente de correlação; b- erro de generalização, calculado a partir da fórmula 2.18. O conjunto teste total foi particionado em 5 diferentes conjuntos de 10 complexos proteína-ligante e no processo de treinamento, eram utilizados 40 complexos deixando sempre um dos conjuntos selecionados de fora.

Tabela 5.10: Análise de erro de generalização e coeficiente de correlação para distintas partições 40/10 treinamento/teste. Rede neural com camada oculta de 6 eurônios.

6 neurônios		
Estimativas de correlação e erro	r^a	Err^b
Teste (Grupo 1)	0,3929	0,4507
Teste (Grupo 2)	0,6923	0,2830
Teste (Grupo 3)	0,0419	0,7205
Teste (Grupo 4)	0,2931	0,7357
Teste (Grupo 5)	-0,4291	1,3814

a-Coeficiente de correlação; b- erro de generalização, calculado a partir da fórmula 2.18. O conjunto teste total foi particionado em 5 diferentes conjuntos de 10 complexos proteína-ligante e no processo de treinamento, eram utilizados 40 complexos deixando sempre um dos conjuntos selecionados de fora.

Tabela 5.11: Análise de erro de generalização e coeficiente de correlação para distintas partições 40/10 treinamento/teste. Rede neural com camada oculta de 3 neurônios.

3 neurônios		
Estimativas de correlação e erro	r^a	Err^b
Teste (Grupo 1)	0,2975	0,5201
Teste (Grupo 2)	0,6779	0,4025
Teste (Grupo 3)	-0,3007	1,0940
Teste (Grupo 4)	0,5878	0,3434
Teste (Grupo 5)	0,1286	0,6813

a-Coeficiente de correlação; b- erro de generalização, calculado a partir da fórmula 2.18. O conjunto teste total foi particionado em 5 diferentes conjuntos de 10 complexos proteína-ligante e no processo de treinamento, eram utilizados 40 complexos deixando sempre um dos conjuntos selecionados de fora.

Nas Tabelas 5.9, 5.10 e 5.11 são analisados os erros de correlação e erro de conjunto teste nos cinco conjuntos utilizados para a validação cruzada de grupo (com diferentes números de neurônios para camada oculta). Cada conjunto de 10 complexos foi utilizado como conjunto teste para avaliar uma rede neural treinada utilizando os 40 complexos restantes. Analisando os resultados fica claro o quanto as funções construídas pelas redes neurais são dependentes dos dados de entrada e do conjunto teste avaliado. Encontramos uma grande variação dos valores de correlação e dos erros dos conjuntos teste (calculados utilizando a equação 2.18) nos 5 grupos, em todos os testes com diferentes números de neurônios na camada oculta da rede neural. Tanto os valores de correlação como o erro do conjunto teste foram calculados apenas com os indivíduos de cada grupo, não se levando em consideração o comportamento de todos eles para uma mesma função.

Na rede neural com a camada oculta com 9 neurônios, observamos os valores dos erros dos conjuntos teste mais homogêneos, com exceção do grupo 5. Nas arquiteturas com 3 e 6 neurônios na camada oculta, observou-se uma menor homogeneidade nos resultados, com relação aos obtidos com a arquitetura de 9 neurônios.

O cálculo do erro teste na metodologia de validação cruzada leva em consideração o desempenho da rede neural para todos os grupos, como demonstrado na Equação 2.19.

Podemos observar na Tabela 5.12, que a rede com a melhor arquitetura, ou seja, com menor erro de generalização, é constituída por uma camada oculta com 9 neurônios. O erro de generalização obtido com a arquitetura contendo 3 neurônios ficando bastante próximo. É interessante notar que na metodologia de treinamento e teste, o menor erro no conjunto teste (partição 45/5) foi observado com uma rede neural com uma arquitetura de 3 neurônios em sua camada oculta. Entretanto, o mesmo não ocorreu nos testes com as partições de dados 25/25 e 35/15, sendo que no treinamento e teste com a partição 25/25 a melhor arquitetura de rede neural foi com 9 neurônios na camada oculta.

Tabela 5.12: Erro de generalização calculado através de Validação Cruzada de Grupo.

Erro	*Erro de Treinamento*	*Err (Generalização)*
3 neurônios[a]	0,09118	0,60826
6 neurônios[a]	0,06982	0,71426
9 neurônios[a]	0,05682	0,54894

a- número de neurônios da camada oculta. O conjunto total foi particionado em 5 diferentes conjuntos de 10 complexos proteína-ligante

Analisando os erros de generalização da Tabela 5.12 e a homogeneidade dos erros nas diferentes partições teste utilizadas na validação cruzada de grupo (Tabelas 5.8, 5.9 e 5.10), chegamos à conclusão que a melhor arquitetura de rede neural é aquela constituída por 9 neurônios na camada oculta, pois possui o menor erro de generalização e a maior homogeneidade nos erros. O modelo de aplicação final, com esta arquitetura, foi construído utilizando os 50 complexos no treinamento da rede neural. Na Tabela 5.12 são mostrados os valores calculados de pK para cada um dos complexos estudados.

Gráfico 5.2: Gráfico dos valores reais de pK e os melhores valores obtidos.

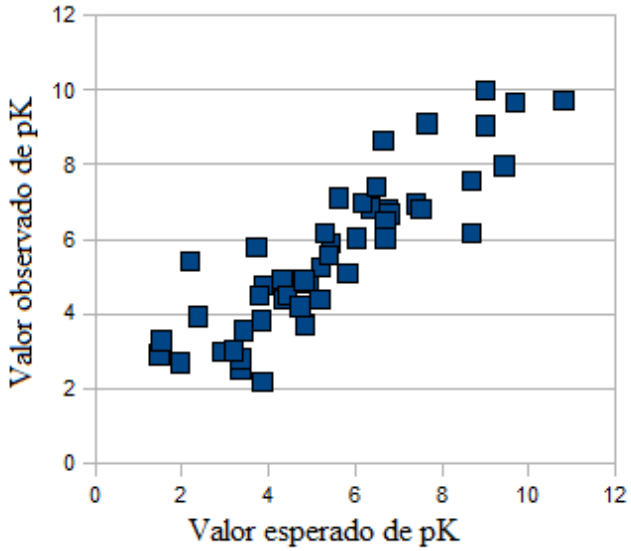


Gráfico dos valores reais de pK (eixo y) dos 50 complexos proteína-ligante do conjunto-teste utilizado neste trabalho (em laranja, e os melhores valores obtidos (em azul) com a melhor arquitetura de rede neural constituída por 9 neurônios na camada oculta, com o menor erro de generalização e a maior homogeneidade nos erros.

Assim como analisado no modelo construído pelo método de regressão, observamos quais complexos não tiveram o valor esperado previsto dentro do intervalo de até 3,0, representando uma diferença de concentração do ligante da ordem de 10^{-3} mol/L. Lembrando que, as medidas da constante de afinidade em nosso problema, são tratadas em pK, como ($-\log K_i$) ou ($-\log K_d$). No teste com a rede neural, 66% dos complexos tiveram uma diferença menor que 1,0 dos valores observados em relação esperados, 92% deles apresentaram diferenças menores que 2,0 e 98%, menor que 3,0, ficando de fora apenas o complexo 1PDZ.

Tabela 5.13: Valores de constantes de afinidade recpetor-ligante obtidos utilizando uma rede neural com 3 neurônios na camada oculta.

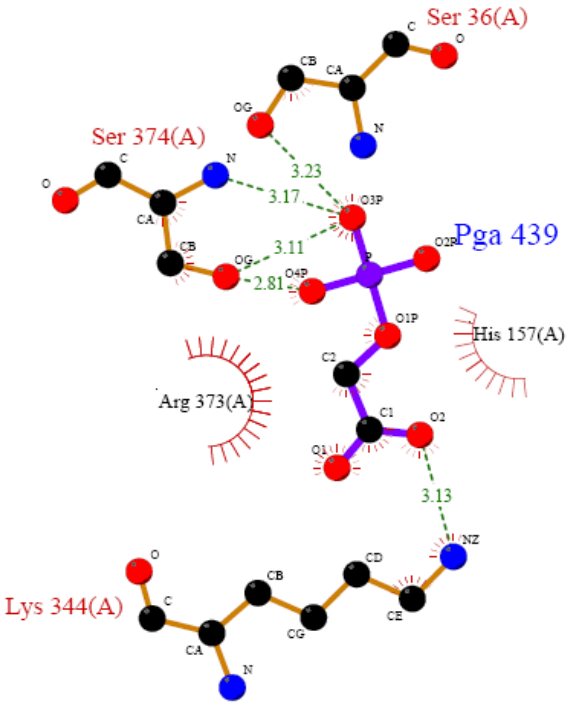
Código PDB do complexo	Valor esperado	Valor absoluto da diferença entre o valor esperado e o valor observado
1ADD	6,74	0,0504
1AI5	3,72	2,0560
1AZM	6,04	0,1013
1C83	4,85	1,1128
1CBX	6,35	0,4893
1CIL	9,43	1,7633
1CPS	6,66	1,9695
1DBB	9	0,9963
1DBJ	7,68	1,4379
1DHF	7,4	0,4653
1EBG	10,82	1,1158
1EJN	5,61	0,4870
1F0R	7,65	1,4973
1FKF	9,7	0,0227
1FL3	6,79	0,1135
1H4N	4,92	0,1095
1HRI	4,34	0,0519
1HYT	5,2	0,0563
1L82	4,85	1,1842
1L86	3,37	0,8711
1MDR	7,52	0,7015
1MMQ	6,16	0,8088
1MNC	9	0,0415
1OYN	6,49	0,9067
1PDZ	2,2	3,2200
1RBP	4,31	0,6117
1RGL	5,43	0,4720
1RNT	5,19	0,8183
1SNC	6,7	0,2537
1SRE	3,85	0,0477
1TNG	2,93	0,0407
1TNH	3,37	0,5659

Continuação da Tabela 5.13.

1TNJ	1,96	0,7180
1TNK	1,49	1,4148
1ULB	5,3	0,8628
2AK3	3,86	1,6751
2CTC	3,89	0,8728
2H4N	8,69	2,5262
2IFB	4,43	1,0714
2PCP	8,69	1,1207
2SIM	3,42	0,1283
2SNS	6,7	0,6665
2XIS	5,82	1,7169
2YPI	4,82	0,0812
3PTB	4,74	0,5405
4XIA	1,54	1,7424
5ENL	3,8	0,6918
6RNT	2,37	1,5474
6TIM	3,21	0,2034
7TIM	5,4	1,1707

No complexo 1PDZ (Figura 5.6), encontramos um mecanismo catalítico onde um íon Mn^{2+} está ligado a 3 grupamentos carboxilatos e 3 moléculas de água. O ligante está ligado próximo, mas não diretamente ao metal. A His 157, presente em um dos laços móveis, está em contato com o átomo C2 do ligante. Uma molécula de água forma uma ligação hidrogênio com o carboxilato do ligante e àqueles do Glu 166 e Glu 209. Foi sugerido que a His 157 é a base que recebe o próton do C2H, visto que a molécula de água é parte de um sistema de transmissão do próton que mantém o sistema na forma de ácido carboxílico, onde o pKa do grupo C2H é baixo o suficiente para transferência do próton para a His 157 (Duquerroy et al., 1995). A ausência dessa molécula de água no nosso sistema, acarretou na perda da acurácia deste mecanismo catalítico, refletida possivelmente no cálculo dos valores de energias e mesmo no número de ligações hidrogênio realizado pelo ligante.

Figura 5.6: Mecanismos de ação do ligante Ácido 2 Glicofosfólico (Pga439) no sítio ativo da proteína Enolase, com destaque para os aminoácidos Ser 36 , Ser 374, Arg 373, Lys 311 e His 157 presentes na cadeia A.



Assim como neste caso, outros tantos sítios ativos funcionam na presença de metais ou de moléculas de água estruturais, que estão ali presentes para a colaboração da interação do ligante na enzima, mostrando novamente a importância das águas (principalmente estruturais) na estimativa das constantes de afinidade.

O cálculo do erro de generalização utilizando validação cruzada de grupo foi realizado para a metodologia de regressão multivariada com 5 variáveis de entrada com termos lineares, quadráticos e cruzados (*i.e.*, superfície acessível ao solvente-SAS, energias eletrostática-EL e de Lennard-Jones-LJ, número de ligações hidrogênio formadas no processo de interação entre a proteína e o ligante-LH e o número de ligações torcionáveis congeladas nesse mesmo processo-NTOR). Além disso, esse mesmo cálculo foi realizado com a regressão multivariada envolvendo apenas as duas variáveis utilizadas na função original de energia do programa DOCKTHOR, com energias eletrostática-EL e de Lennard-Jones-L-J em termos lineares. Foi utilizado o mesmo

protocolo da validação cruzada de grupo da rede neural. Utilizando validação cruzada de grupo para as funções obtidas por regressão multivariada obteve-se um erro de generalização de 8,79644 para a função de 5 variáveis e um erro de 6,39720 para a função com duas variáveis.

O erro de generalização, obtido por VCG, de uma rede neural utilizando 9 neurônios na camada oculta (0,54894) é muito melhor do que os calculados para as funções encontradas por regressão. Este resultado é um indicativo da superioridade da metodologia de rede neural, com relação à metodologia de regressão multivariada, para estimar constantes de inibição receptor-ligante, principalmente numa função capaz de generalizar dados de entrada tão diversos, como os aqui utilizados.

CAPÍTULO 6

6. Conclusão e Perspectivas

Neste trabalho, foram construídas funções empíricas para o cálculo da constante de afinidade entre proteína-ligante através de duas metodologias distintas. Na primeira, através de regressão múltipla, foi avaliada a importância de cada uma das variáveis utilizadas como dados de entrada na construção desta função. Na segunda metodologia, através de rede neural, buscamos o melhor modelo/arquitetura para o cálculo de constantes de afinidade.

Antes da construção de cada um dos modelos, foi feita uma grande pesquisa e foi estabelecida uma metodologia adequada para a criação de um conjunto-teste ideal, de forma que este possuísse complexos receptor-ligante capazes de serem menos susceptíveis à geração de erros e representando uma grande diversidade em vários aspectos importantes para a construção de uma função empírica geral para estimar a afinidade receptor-ligante. Este mesmo conjunto-teste também serviu de base (de Magalhães, 2006) para o teste do programa de atracamento molecular (programa DOCKTHOR).

O conjunto-teste construído possui 50 complexos proteína-ligante de diferentes famílias e características em seu sítio ativo, desde estruturas marcadamente hidrofóbicas e hidrofílicas, como também com sítios mistos, com uma boa variabilidade nas constantes de afinidade e no número de ligações torcionáveis congeladas dos ligantes e ainda levando em consideração ligantes com distintos padrões de carga.

Com relação à importância das variáveis de entrada verificou-se que a introdução de três novas variáveis (e.g., porcentagem da superfície acessível ao solvente do ligante em contato com a proteína, número de ligações torcionáveis do ligante congeladas no processo de atracamento e número de ligações hidrogênio receptor-ligante formadas) possui o potencial de melhorar bastante o

poder de predição de constante de afinidade do programa DOCKTHOR.

A correlação obtida com a função atual do programa é de apenas 0,4245, mostrando a grande necessidade de se introduzir melhorias neste aspecto. A função empírica, construída com a metodologia de regressão múltipla, que obteve melhor resultado foi a que utilizou as 5 variáveis de entrada apresentando termos lineares, cruzados e quadráticos, com um coeficiente de correlação linear igual a 0,7542.

Funções empíricas constituídas por redes neurais com diferentes arquiteturas também foram avaliadas neste trabalho. Para a avaliação destas, utilizamos duas metodologias diferentes, a de treinamento e teste e a validação cruzada de grupo.

Na metodologia de treinamento e teste, observamos que a melhor arquitetura da rede neural foi dependente do tipo de partição de dados treinamento/teste utilizado. Também foram observados problemas de superajustamento (*i.e.*, modelos onde o erro de treinamento era melhor, mas com erro de generalização maior).

Os resultados obtidos neste trabalho indicam quatro principais linhas de pesquisa a serem exploradas em trabalhos futuros: A primeira seria a construção de um conjunto validação maior, como vem sendo feito na avaliação de funções de energia livre de ligação de outros programas de atracamento molecular. A segunda seria o desenvolvimento de funções de energia livre de ligação para proteínas de uma mesma família, observando os resultados e os desempenhos para um sítio ativo homólogo, onde possivelmente os mesmos tipos de interação entre os ligantes e as proteínas são observados.

A terceira perspectiva é a inclusão de novos termos de energia na construção da função de energia livre de ligação, onde podem ser testados outros termos já utilizados em outras funções de outros programas de atracamento molecular, como interações do tipo ambíguo, formadas entre átomos polares e não-polares, ou a consideração de diferentes tipos de ligações hidrogênio, dependente dos átomos envolvidos.

Por fim, testes com outras arquiteturas de rede neural também devem ser realizados, buscando melhores formas de se evitar o superajustamento, problema que deve ser suavizado com um maior conjunto teste, mas que deve ser também considerado. Além da utilização de outras metodologias de reconhecimento de padrão, como a máquina de vetor de suporte (SVM, "*Support Vector Machine*").

REFERÊNCIAS BIBLIOGRÁFICAS

K.A., Anderson, B.A. Magnuson, M.L. Tshirgi e B. Smith. Journal of Agricultural and Food Chemistry. 47(4): 1568-1575, 1999.

J. Apostolakis, A. Pluckthun e A. Caflisch. Docking small ligands in flexible bindings sites. **Journal of Computational Chemistry**. 19 (1): 21-37, 1998.

N. Arora e B. Jayaram. Strenght of hydrogen bonds in alpha-helices. **Journal of Computational Chemistry**, 18(9): 1245-1252, 1997.

J. Åqvist, C. Medina e J. E. Samuelson. A new method for predicting binding affinity in computer-aided drug design. **Protein Eng.,** 7: 385-391, 1994.

J. Åqvist e J. Marelius. The Linear Interaction Energy Method for Predicting Ligand Binding Energies. **Combinatorial Chemistry & High Throughput Screening**, 4: 613-626, 2001.

E. J. Barreiro. A Descoberta Racional de Fármacos. **Ciência Hoje,** 40(235): 26-31, 2007.

B. de Barros Neto, I. S. Scarminio e R. E. Bruns. **Planejamento e Otimização de Experimentos**. Editora da UNICAMP. 2ª Edição, 1995.

J. M. Berg, J. L. Tymoczko, e L. Stryer. **Bioquímica**. Editora Guanabara Koogan S.A. 5ª Edição, 2004.

C. M. Bishop. Neural Networks and Their Applications. **Rev. Sci. Instrum.**, 65 (6): 1803-1830, 1994.

P. Block, C. A. Sotriffer, I. Dramburg, e G. Kleber. AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. **Nucleic Acids Research**, 34(Database Issue):D522-6, 2006.

E. Byvatov, U. Fechner, J. Sadowski e G. Schneider. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. **J. Chem. Inf. Comput. Sci.**, 43(6):1882-9, 2003.

T. L. Blundell, H. Jhoti, e C. Abell. High-throughput crystallography for lead discovery in drug design. **Nature Reviews – Drug Discovery**, 1(1): 45-54, 2002.

T. L. Blundell e S. Patel. High-throughput X-ray crystallography for drug discovery. **Curr. Opin. Pharmacol.**, 4(5): 490-6, 2004.

J. R. Bock e D. A. Gough. A New Method to Estimate Ligand-Receptor Energetics. **Molecular & Cellular Proteomics,** 1(11): 904-10, 2002.

A. Bondi. van der Waals volumes and radii. **The Journal of Physical Chemistry**, 68(3): 441-451, 1964.

B.K. Bose. Neural network applications in power electronics and motor drives- An introduction and perspective. **IEEE Transactions on Industrial Electronic**s, 54(1): 14-33, 2007.

N. Brooijmans e I. Kuntz. Molecular Recognition and Docking Algorithms. **Annu. Rev. Biophys. Struct.**, 32: 335-73, 2003.

B. R. Brooks, R. E. Bruccoreli, B. D. Olafson, D. J. States, S. Swaminathan, e M. Karplus. CHARMM: a program for macromolecular energy minimization and dynamics calculation. **Journal of Computacional Chemistry**, 4(2): 187-217, 1983.

B. D. Bursulaya, M. Totrov, R. Abagyan, e C. L. Brooks III. Comparative Study of Several Algorithms for Flexibe Ligand Docking. **Journal of Computer-Aided Molecular Design**, 17: 755-763, 2003.

A. M. Cappalonga, R. S. Alexander, D.W. Christianson. Structural comparison of sulfodiimine and sulfonamide inhibitors in their complexes with zinc enzymes. **J.Biol.Chem**., 267**:** 19192-19197, 1992.

N. A., Carlson. **Foundations of Physiological Psychology.** Needham Heights, Massachusetts: Simon & Schuster, 1992.

S. K. Choudhury, V. K. Jain e C. V. V. R. Rao. On-line monitoring of tool wear in turning using a neural network. **International Journal of Machine Tools & Manufacture**. 39(3): 489-504, 1999.

R. Clausius. The Mechanical Theory of Heat with Its Applications to the Steam-Engine and to the Physical Properties of Bodies, 1865.

M. L. Connoly. Solvent-accessible surfaces of proteins and nucleic acids. **Science**, 221 (4612): 709-713, 1983.

M. L. Connoly. MDS Molecular Dot Surface. Ported to Macintosh and Language Systems Fortran in July,1993, 1986.

W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz Jr., D. M. Fergunson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, e P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. **Journal of the American Chemical Society**, 117: 5179-5197, 1995.

G. V. Cybenko. Approximation by Superpositions of a Sigmoidal Function. **Mathematics of Control Signals and Systems**, 2: 303-314, 1989.

L. E. Dardenne. Notas de aula da disciplina "Estrutura de Proteínas e Simulação Computacional de Macromoléculas Biológicas" oferecida pelo Laboratório Nacional de Computação Científica, 2000.

L. E. Dardenne. "**Propriedades Eletrostáticas do Sítio Ativo de Cisteíno Proteinases da Família da Papaína."** Tese de Doutorado em Ciências Biológicas (Biofísica). Universidade Federal do Rio de Janeiro, Rio de Janeiro – Rio de Janeiro, 2000.

M. Dickson e J. P. Gagno. Key Factors in the Rising Cost of New Drug Discovery and Development. **Nature Reviews – Drug Discovery**, 3: 417-429, 2004.

J. Ding, G. Koellner, H. P. Grunert, W. Saenger. Crystal structure of ribonuclease T1 complexed with adenosine 2`-monophosphate at 1.8-A resolution. **J. Biol.Chem.**, 266: 15128-34, 1991.

T. J. A. Ewinga, S. Makinoa, A. G. Skillmana, e I. D. Kuntz. DOCK 4.0: Search Strategies for Automated Molecular Docking of Flexible Molecule Databases. **Journal of Computer-Aided Molecular Design**, 15: 411-428, 2001.

A. M. R Fernandes. **Inteligência Artificial – Noções Gerais**. Editora Visual Books. Florianópolis, Brasil, 1ª Edição, 2003.

R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin e D. T. Mainz. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. **Journal of Med. Chem.**, 49: 6177-6196, 2006.

E. H. Fischer. **Dictionary of scientific biography**. 5: 1-5.

S. C. Gad. **Drug Discovery Handbook**. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2005.

W. Gobel, B.M. Kampa e F. Helmchen. Imaging cellular network dynamics in three dimensions using fast 3D laser scanning. **NATURE Methods**, 4(1): 73-79, 2007.

P. J. Gooford. A Computational Procedure for Determining Energetically Favorable Binding Site on Biologically Important Macromolecules. **J. Med. Chem.,** 28: 849-857, 1985.

P. Gratteri, C. Bonaccino, e F. Melani. Searching for a reliable orientation of ligands in their binding site: comparison between a structure-based (Glide) and a ligand-based (FIGO) approach in the case study of PDE4 inhibitors. **J. Med. Chem.,** 48(5): 1657-65, 2005.

T. A. Halgren. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. **Journal of Computational Chemistry**, 17 (5&6): 490-519, 1996a.

T. A. Halgren. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. **Journal of Computational Chemistry**, 17 (5&6): 520-552, 1996b.

T. A. Halgren. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. **Journal of Computational Chemistry**, 17 (5&6): 553-586, 1996c.

T. A. Halgren. Merck molecular force field. IV. Conformational energies and geometries for MMFF94. **Journal of Computational Chemistry**, 17 (5&6): 587-615, 1996d.

T. A. Halgren. Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data and empirical rules. **Journal of Computacional Chemistry**, 17 (5&6): 616-641, 1996e.

S. Haykin. **Redes Neurais – Princípios e Prática**. Editora Bookman. Porto Alegre, Brasil, 2ª Edição, 2001.

R. Hooke e T. A. Jeeves. 'Direct Search' Solution of Numerical and Statistical Problems. **Journal of the ACM**, 8(2): 212-229, 1961.

D. B. Kitchen, H. Decornez, J. R. Furr, e J. Bajorath. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. **Nature Reviews – Drug Discovery**, 3(11): 935-49, 2004.

F. E. Koehn e G. T. Carter. The Evolving Role of Natural Products in Drug Discovery. **Nature Reviews – Drug Discovery**, 4: 206-220, 2005.

I. Kola e J. Landis. Can the pharmaceutical industry reduce attrition rates? **Nature Reviews – Drug Discovery**, 3(8): 711-715. 2004.

P. Kollman. Free Energy Calculations: Applications to Chemical and Biochemical Phenomena. **Chem. Rev.**, 93: 2395-2417, 1993.

B. Kramer, M. Rarey, e T. Lengauer. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. **Proteins**, 37(2): 228-41, 1999.

A. R. Leach. **Molecular Modeling – Principles and Applications**. Pearson Education Ltda., 2ª Edição, 2001.

A. R. Leach, B. K. Shoichet e C. E. Peishoff. Docking and Scoring - Perspective. **Journal of Medicinal Chemistry**, 49(20): 5851-5855, 2006.

R. Lee e F. M. Richards. The Implementation of protein structures: Estimation of static accessibility. **Journal of Molecular Biology**, 55: 379-400, 1971.

C. Levinthal, S. J. Wodak, P. Kahn, e A. K. Dadivanian. Hemoglobin Interaction in Sickle Cell Fibers. I. Theoretical Approaches to the Molecular Contacts. **Proc. Natl. Acad. Sci. USA**, 72: 1330-34, 1975.

C. A. Lesburg, C. Huang, D. W. Christianson, C. A. Fierke. Histidine --> carboxamide ligand substitutions in the zinc binding site of carbonic anhydrase II alter metal coordination geometry but retain catalytic activity. **Biochemistry**, 36:15780-91, 1997.

A. Li, e R. Nussinov. A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. **PROTEINS: Structure, Function and Genetics**, 32: 111-127, 1998.

R. D. Lins e P. H. Hunenberger. A New GROMOS Force Field for Hexopyranose-Based Carbohydrates. **J. Comput. Chem.**, 26: 1400-1412, 2005.

C. S. de Magalhães. **Algoritmos Genéticos para o Problema de Docking Proteína-Ligante**. Tese de Doutorado do Curso de Modelagem Computacional do Laboratório Nacional de Computação Científica, Petrópolis – Rio de Janeiro, 2006.

M. C. Medeiros, T. Terasvita, G. Rech. Building Neural Networks Models for Time Series: a Statistical Approach. **SSE/EFI Working Paper Series in Economics and Finance**, 508: 1-47, 2002.

G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, e A. J. Olson. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. **Journal of Computational Chemistry**, 19(14): 1639-1662, 1998.

A. A. Ajay e M.A. Murcko. Computational Methods to Predict Binding Free Energy in Ligand-Receptor Complexes. **Journal of Medicinal Chemistry**, 38(26): 4953-4967, 1995.

C. W. Murray, C.A. Baxter e A. D. Frenkel. The Sensitivity of the Results of Molecular Docking to Induced Effects: Aplication to thrombin, Thermolysin and Neuraminidase. **J. Comput. Aid. Mol. Des.**, 13: 547-62, 1999.

A. G. Murzin, S. E. Brenner, T. Hubbard e C. Chothia. SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. **J. Mol. Biol**, 247: 536-540, 1995.

H. Nakamura. Roles of electrostatic interaction in proteins. **Q. Rev. Biophys.**, 29(1): 1-90, 1996.

D. L. Nelson e M. M. Cox. **Lehninger – Principles of Biochemistry.** W. H. Freeman, 4ª edição, 2004.

F. G. Oliveira, C. M. Sant'Anna, E. R. Caffarena, L. E. Dardenne e E. J. Barreiro. Molecular docking study and development of an empirical binding free energy model for phosphodiesterase 4 inhibitors. **Bioorg Med. Chem.**, 14(17):6001-11, 2006.

J. Y. Ortholand e A. Ganesan. Natural products and combinatorial chemistry: back to the future. **Curr. Opin. Chem.**, 8(3): 271-80, 2004.

L. Pauling e M. Delbruck. The nature of the intermolecular forces operative in biological processes. **Science**, 92: 77-79, 1940.

J. Paquet, C. Lacroix e J. Thibault. Modeling of pH and acidity for industrial cheese production. **Journal of Dairy Science,** 83(11): 2393-2409, 2000.

M. M. Poulton. Neural networks as an intelligence amplification tool: A review of applications. **Geophysics,** 67(3): 979-993, 2002.

P. Preziosi. Science, Pharmacoeconomics and Etichs in Drug R&D: a Sustainable Future Scenario? **Nature Reviews – Drug Discovery**, 3: 521-526, 2004.

C. Robert, C. Guilpin e A. Limoge. Review of neural network applications in sleep research **Journal of Neuroscience Methods**, 79(2): 187-193, 1998.

W. R. Rocha. Interações Intermoleculares. **Cadernos Temáticos de Química Nova na Escola.** Número 4, 2001.

A. W. Schuettelkopf e D. M. F. Aalten. PRODRG – A Tool for High-Throughput Crystallography of Protein-Ligand Complexed. **Acta Crystallographica**, D60: 1355-1363, 2004. PRODRG is a available at http://davapc1.bioch.dundee.ac.uk/cgi-bin/prodrg_beta.html for academic use.

Spartan '04 Windows Tutorial and User's Guide. Wavefunction Inc.: Irvine, CA 92612. (http://www.wavefun.com/software/ spartan04_win/win_manual/main.html).

M. Stahl e M. Rarey. Detailed Analysis of Scoring Functions for Virtual Screening. **J. Med. Chem.**, 44: 1035-1042, 2001.

E. Stjernschantz, J. Marelius, C. Medina, M. Jacobson, N. P. E. Vermeulen e C. Oostenbrink. Are Automated Molecular Dynamics Simulation and Binding Free Energy Calculations Realistic Tools in Lead Optimization? An Evaluation of the Linear Interaction Energy (LIE) Method. **J. Chem. Inf. Model.**, 46: 1972-1983, 2006.

S. Tietze e J. Apostolakis. GlamDock: Development and Validation of a New Docking Tool on Several Thousand Protein-Ligand Complexes. **J. Chem. Inf. Model.**, 47(4): 1657-1672, 2007.

J. Y. Trosset e H. A. Scheraga. PRODOCK: Software package for protein modeling and docking . **Journal of Computational Chemistry**, 20 (4): 412-427, 1999.

J. M. Twomey e A. E. Smith. Bias and Variance of Validation Methods for Function Approximation Neural Networks Under Conditions of Sparse Data. **IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews**, 28(3): 417-430, 1998.

W. F. van Gunsteren e H. J. C. Berendsen. **GROMOS-87 Manual**. Biomos BV Nijenborgh 4, 9747 AG Groningen, The Netherlands, 1987.

J. Yuan e L. Terrence. Neural-Network design for small training sets of high dimension. **IEEE**

Transactions on neural networks, 9(2), 1998.

V. V. Zernov, K. V. Balakin, A. A. Ivaschenko, N. P. Savchuk e I. V. Pletnev. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. **J. Chem. Inf. Comput. Sci.** 43(6): 2048-56, 2003.

E. Walum, J. Hedander e P. Garberg. On the relevance of cytotoxicity measurements, barrier passage determinations and high throughput screening in vitro to select potentially hazardous compounds in large sets of chemicals. **Toxicology and Applied Pharmacology,** 207 (2-1): 393-397, 2005.

R. Wang, L. Lai, e S. Wang. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. **Journal of Computer-Aided Molecular Design**, 16: 11-26, 2002.

R. Wang, Y. Lu, e S. Wang. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. **J. Med. Chem.**, 46: 2287-2303, 2003.

G. L. Warren, C. W. Andrews, A. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff e M. S. Head. A Critical Assessment of Docking Programs and Scoring Functions. **J. Med. Chem.** 49: 5912-5931, 2006.

S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Guio, G. Alagona, S. Profeta, e P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. **Journal of**

the American Chemical Society, 106: 765-784, 1984.

B. Widrow e M. E. Hoff Jr. Adaptative switching circuits. **IRE WESCON Convention Record**, 96-104, 1960.

P. Willet, R. C. Glen, A. R. Leach, e R. Taylor. Development and Validation of a Genetic Algorithm for Flexible Docking – GOLD. **J. Mol. Biol.**, 267: 727-748, 1997.

S. Wolfram. **Mathematica: Reference Guide/for Mathematica Version 2.** Addison-Wesley, 1992.

APÊNDICE A

Manual para o Programa de Atracamento Molecular DOCKTHOR

O programa DOCKTHOR foi desenvolvido durante o programa de Doutorado em Modelagem Computacional da aluna Camila de Magalhães, sob orientação dos professores Laurent Dardenne e Hélio Barbosa, no Laboratório Nacional de Computação Científica. Assim como o AutoDock 3.0, este programa tem como entrada, arquivos de parâmetros tanto para a proteína como para os ligantes, para a construção de uma rede de contribuições energéticas e o atracamento molecular.

Na construção dos arquivos de entrada, a proteína foi parametrizada com o campo de força GROMOS 96, no programa PDBTHORBOX, desenvolvido pelos Professores Laurent Dardenne do Laboratório Nacional de Computação Científica, Ernesto Caffarena, da Fundação Oswaldo Cruz, Isabelle Ortmans da Universidade Federal do Rio de Janeiro e Michel Loos da Universidade de São Paulo. Os cuidados com as cargas de alguns aminoácidos, das moléculas de água e dos íons metais já foram descritos anteriormente.

Já os ligantes, eram submetidos à versão Beta do sítio do PRODRG (SCHUETTELKOPF e AALTEN, 2004) sem a realização de minimização de energia, para a construção de um arquivo de coordenadas com a adição de hidrogênios nos átomos polares e aromáticos (DRGAPH.PDB). Posteriormente, este arquivo de saída era submetido ao programa Babel 1.6 (WALTERS e STAHL, 1992-1996) para a geração de um arquivo de coordenadas internas do tipo Mopac (.int). Era então criada uma pasta de três letras identificando o ligante a ser usado, onde eram armazenados os arquivos de coordenadas cartesianas (.pdb) e o arquivo de coordenadas internas (.int), para então ser executado o programa GROTOP, desenvolvido pelo Professor Laurent Dardenne, para a a construção de um novo arquivo de coordenadas cartesianas (.pdb), um arquivo (.LOG) e um

arquivo sem extensão com as cargas dos átomos, ligações, ângulos de ligações, ligações diedrais próprias e impróprias, do campo de força GROMOS96, sendo os três arquivos com o mesmo nome do diretório citado anteriormente.

O campo de força para ligantes adotado neste trabalho, foi o MMFF94, portanto os valores encontrados no arquivo sem extensão eram mudados usando como referência o arquivo de saída do programa Spartan (.spartan) com a parametrização da mesma molécula com o campo de força MMFF94. Eram também checadas as ligações diedrais próprias para então se executar o programa PDBTHORBOX, que parametriza uma nova molécula de acordo com as cargas equivalentes ao tipo de molécula disponível no programa, que para o caso de nosso conjunto-teste tinha arquivos de carga do campo de força de nosso interesse, o MMFF94.

A saída deste programa, tanto no caso da proteína, como no caso do ligante, foram seis arquivos, dos tipos .IN, .TOP, .RAS, .MD, arquivos nomeados como RESUMO.OUT e INPUT.IN. Os arquivos de entrada para o programa DOCKTHOR, são dos tipos .TOP, .IN e .MD. O .IN é um arquivo de coordenadas que deve conter um título identificador, e 17 colunas com a descrição de cada átomo. Na primeira coluna descreve-se o número do átomo e na segunda o tipo, o que varia segundo o campo de força utilizado. As três colunas seguintes são usadas para as posições x, y e z. As outras seis colunas seguintes utilizam-se para descrever as ligações de cada átomo. Na coluna doze, coloca-se a carga parcial atômica e nas três seguintes as informações correspondentes ao código PDB. As colunas restantes são utilizadas para determinar o número de átomos de cada molécula.

O arquivo de topologia (.TOP) contem toda a informação molecular. Nele, detalham-se as ligações entre dois, três e quatro átomos, assim como a lista de vizinhos excluídos, terceiros vizinhos e as massas atômicas. Já o arquivo do tipo .MD é de parâmetros e contém a informação das condições nas quais será realizado o atracamento molecular, descrito abaixo:

```
  10.0000    TEMPERI  : Temperatura inicial p/ Termalizacao.(K)
     10.0000    DELTA_T  : Incremento na Temperatura.(K / ps)
     10.0000    TEMPSCAL : Re-escalona as veloc. quando T-To > TempScal.
    300.0000    TEMPERMAX: Temperatura maxima.(K)
      0.0005    HTIME    : Incremento no tempo de integracao.(ps)
   1.4584E-5    PRESSI   : Pressao inicial.(Kcal/(mol*A^3))
   1.4584E-5    PRESSF   : Pressao final.  (Kcal/(mol*A^3))
   0.0000E-3    DELTAP   : Incremento na Pressao. (Kcal/(mol*A^3*ps))
        1000    NSTEP    : Numero de passos no calculo MD.
           1    NTC      : (1 s/SHAKE;2 soh Hid.;3 completo. P/ SOLUTO)
           1    NTCC     : (1 solvente; 0 soluto + solvente)
          50    ISCALING : Freq. p/ re-escalonamento das velocidades.
          50    ISCALP   : Freq. p/ re-escalonamento da pressao
         100    IWRTOUT  : Freq. (em passos) p/ escr. no arquivo .OUT.
         100    IWRTDBK  : Freq. (em passos) p/ escr. em .xyz e  .DBK.
         500    IWRTDH   : Freq. (em passos) p/ escrever em .DH*.
           1    INVEL    : 0 utiliza MB; outro N veloc. anteriores.
           1    IAT      : Atomo inicial do loop da dinamica
           0    NFIXATM  : Numero de atomos fixos no calculo MD ou OTMZ.
```

Assim como o arquivo do tipo .MD, o arquivo .IN deve ser modificado com a inclusão de um cabeçalho:

```
MECANICA MOLECULAR

HOOK FTHET FDHP FDHI NBCVW GADOCK GROMOS2

    .1000000000     Valor inicial do passo DELTA-X usado na minimizacao.
    .0100000000     Valor da precisao no calculo do potencial.
    .0000000000     Posicao da interface no eixo X.
   4.0000000000     Valor de EPSLON1, correspondente ao meio aquoso.
    .0000000000     Valor de EPSLON2, correspondente ao meio apolar.
    .0000000000     Valor de XBOX, lado X da caixa para condicoes periodicas.
    .0000000000     Valor de YBOX, lado Y da caixa para condicoes periodicas.
    .0000000000     Valor de ZBOX, lado Z da caixa para condicoes periodicas.
   0.0000000000     Valor de RCUT, raio de corte.
   0.0000000000     Valor de RLIST, para lista de vizinhos.
```

São necessárias ainda informações adicionais no fim do arquivo com o número de ligações torcionáveis, a descrição destas ligações e do átomo de referência, junto com parâmetros necessários para o funcionamento do programa que são padrão.

```
 $LEDOCKPAR IAREF=16, NDIED=2   $END

    16    17
    16    13
 $DOCKINIT INITREF=3  $END
-3.96200   -2.67700   -9.89100

 $GSAPAR Ngsa=10, Temper0a=1.0,
  PardT0a=0.99, ParqV0a=1.3, ParqA0a=1.8, CteT0=1.0,
  ParqTa=2.0, ParqVa=1.3, ParqAa=1.8,  CteT=1.0,
  Nseriea=7, NStopMaxa=500, NStopMax2a=3000,
```

```
  NRANa =-43, NslideGSAa=50,
  Nmarkchaina=0, Nmarka=2000,
  MultiGSAa=.FALSE., NSola=10,
  Nmarkchain2a=0, Nmark2a=10000 &END


MultiGSAa =.FALSE.
NStopMaxa = numero de ciclos do GSA
Nmarkchaina = numero de cadeias de Markov a T cte nos 1os ciclos GSA
Nmarka = tamanho de cada cadeia de Markov (no de ciclos GSA a T cte)
(apos atingir Nmarkchain o GSA segue como o usual)

MultiGSAa=.TRUE.
NStopMaxa= numero maximo de passos a T cte para escolha das solucoes iniciais
para os ciclos GSA.
NSola= no de solucoes GSAs (no de configuracoes iniciais escolhi
NStopMax2a = numero de ciclos do GSA para cada configuracao
Nmarkchain2a = numero de cadeias de Markov a T cte nos 1os ciclos GSA para cada
configuracao inicial
Nmark2a = tamanho de cada cadeia de Markov (no de ciclos GSA a T cte p/cada
configuracao)
(apos atingir Nmarkchain o GSA segue como o usual)
```

É importante lembrar que a descrição destas ligações torcionáveis ocorre na direção a partir do átomo de referência. Por fim, todos os arquivos devem ser armazenados na pasta /data-in do programa DOCKTHOR.

Essas mudanças foram realizadas apenas para o tratamento dos ligantes, para a proteína, devemos mudar além do cabeçalho:

```
 MECANICA MOLECULAR

HOOK FTHET FDHP FDHI NBCVW FUNCDIE2 GRIDDOCK GROMOS2

     .1000000000     Valor inicial do passo DELTA-X usado na minimizacao.
     .0100000000     Valor da precisao no calculo do potencial.
     .0000000000     Posicao da interface no eixo X.
   80.0000000000     Valor de EPSLON1, correspondente ao meio aquoso.
     .0000000000     Valor de EPSLON2, correspondente ao meio apolar.
     .0000000000     Valor de XBOX, lado X da caixa para condicoes periodicas.
     .0000000000     Valor de YBOX, lado Y da caixa para condicoes periodicas.
     .0000000000     Valor de ZBOX, lado Z da caixa para condicoes periodicas.
     .0000000000     Valor de RCUT, raio de corte.
     .0000000000     Valor de RLIST, para lista de vizinhos.
```

Além disso, após a descrição dos átomos devemos adicionar a linha:

```
$LEREDEPAR XMIN=26.642 XMAX=53.642 YMIN=26.551 YMAX=53.551 ZMIN=17.003 ZMAX=44.003 RSTEP=0.25  $END
```

Com a descrição da coordenada mínima e máxima dos três eixos x, y e z da rede para o armazenamento da informação sobre as contribuições energéticas do receptor, com o espaçamento entre os pontos em Å.

Depois de criados e modificados todos os arquivos para a proteína e os ligantes, foi criada primeiramente a rede de contribuições energéticas do receptor, com isto serão criados dois arquivos de tipos .GRD e .VDW, que devem ser renomeados para GRDPOT.DAT e GDRVDW.DAT, respectivamente, e armazenados na pasta /data-in do programa DOCKTHOR.

Por fim, os parâmetros para o algoritmo genético e próprios para a execução do programa estão localizados no arquivo namelistgenin, localizado dentro da pasta /DOCKTHOR.

APÊNDICE B

Valores para o raio de van der Waals (em Å) utilizados pelo programa SASCalc (Goliatt, 2007).

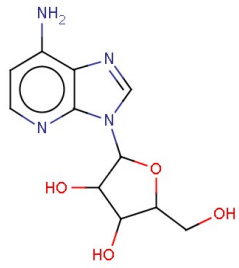
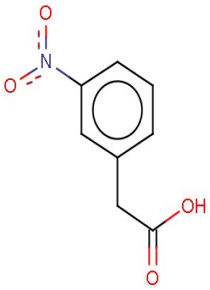
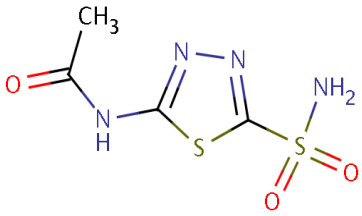
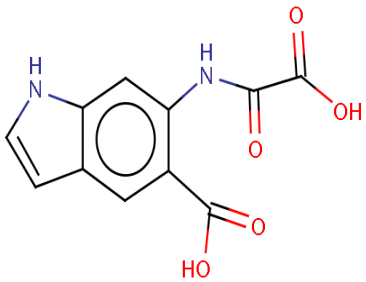
H	1,20	**Cl**	1,75
He	1,40	**Ar**	1,88
C	1,70	**As**	1,85
N	1,55	**Se**	1,90
O	1,52	**Br**	1,85
F	1,47	**Kr**	2,02
Ne	1,54	**Te**	2,06
Si	2,10	**I**	1,98
P	1,80	**Xe**	2,16
S	1,80		

Li	1,82	**Au**	1,66
Na	2,27	**Zn**	1,39
K	2,75	**Cd**	1,58
Mg	1,73	**Hg**	1,55
Ni	1,63	**Ga**	1,87
Pd	1,63	**In**	1,93
Pt	1,75-1,72	**Tl**	1,96
Cu	1,4	**Sn**	2,17
Ag	1,72	**Pb**	2,02

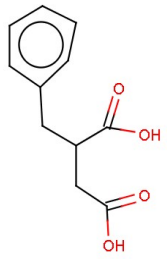
CH	2,01
CH2	1,92
CH3	1,92

APÊNDICE C

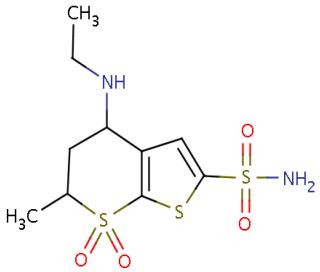
Estruturas dos ligantes (Formato Smiles) Fonte: (http://pc1664.pharmazie.uni-marburg.de/affinity/) (Block et al., 2006)

Código PDB do complexo	
1ADD	
1AI5	
1AZM	
1C83	

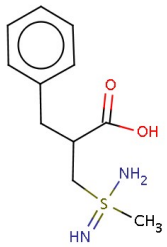
1CBX



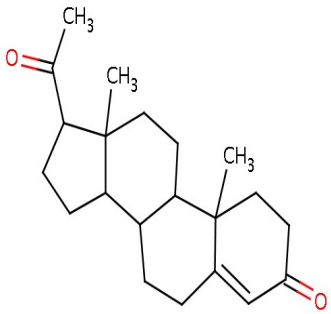
1CIL



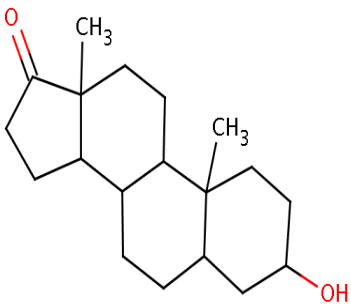
1CPS



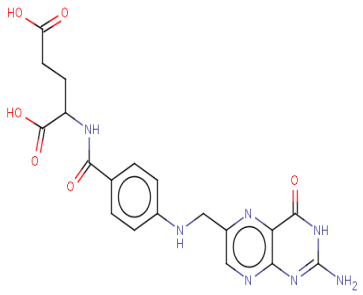
1DBB



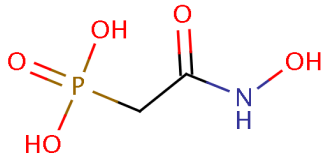
1DBJ



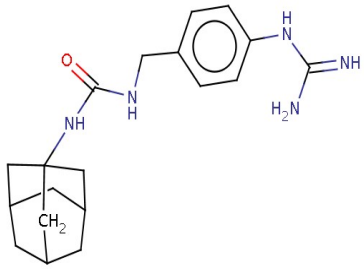
1DHF



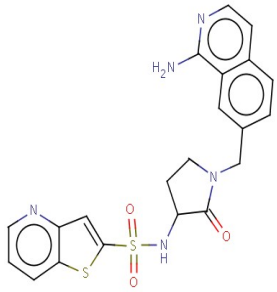
1EBG



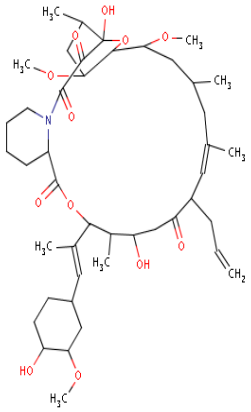
1EJN



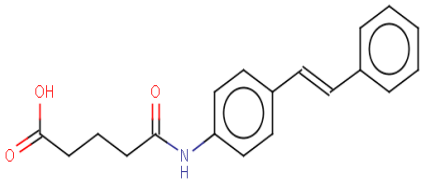
1F0R



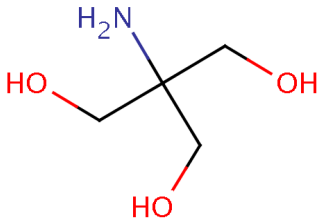
1FKF



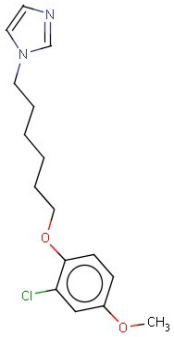
1FL3



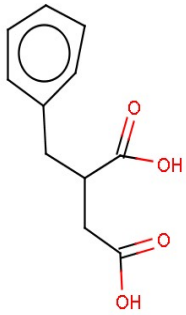
1H4N



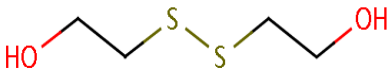
1HRI



1HYT

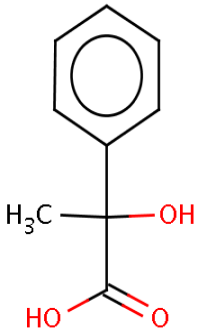


1L82

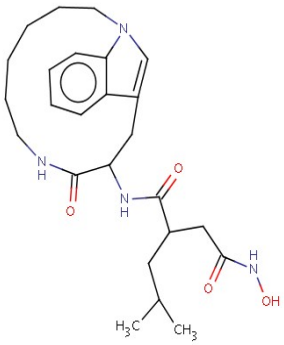


1L86

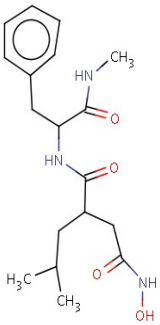
1MDR



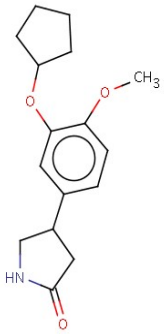
1MMQ



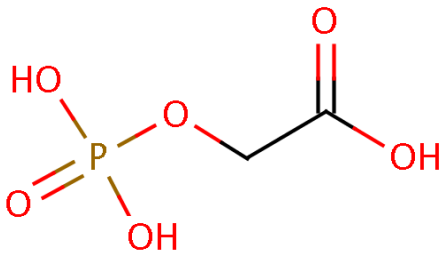
1MNC



1OYN

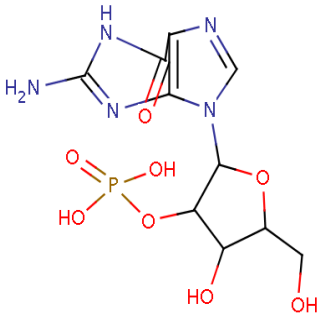


1PDZ

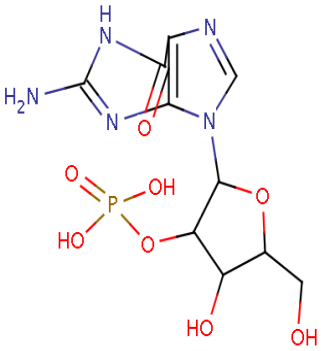


1RBP Não Disponível

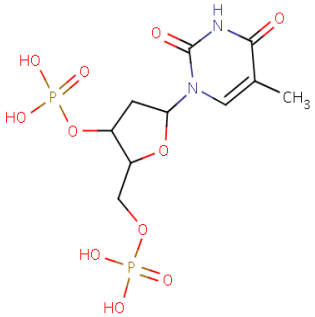
1RGL



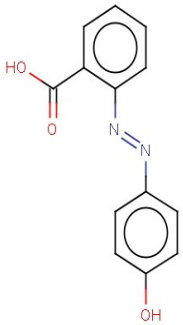
1RNT



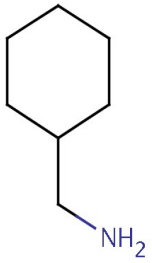
1SNC



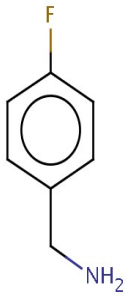
1SRE



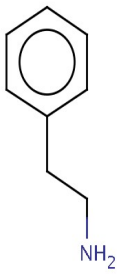
1TNG



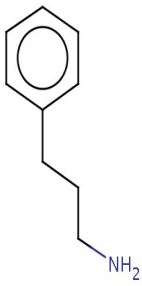
1TNH



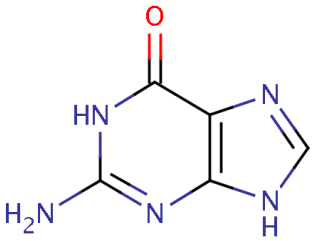
1TNJ



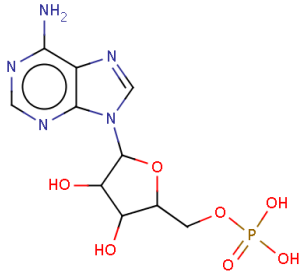
1TNK



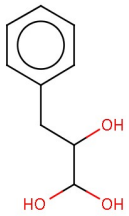
1ULB



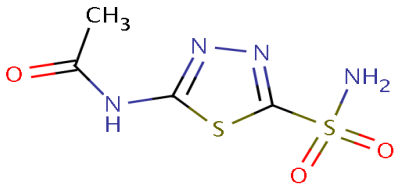
2AK3



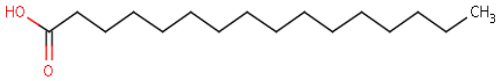
2CTC



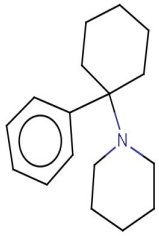
2H4N



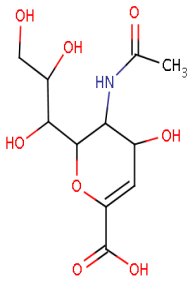
2IFB



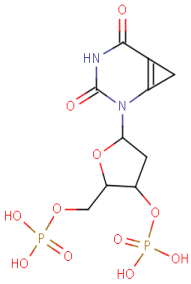
2PCP



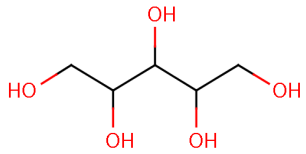
2SIM



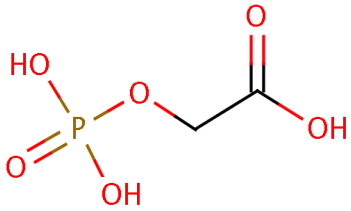
2SNS



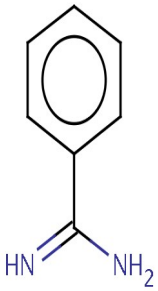
2XIS



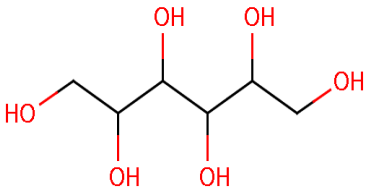
2YPI



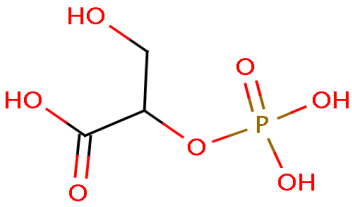
3PTB



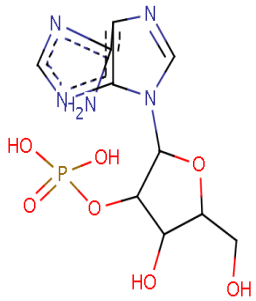
4XIA



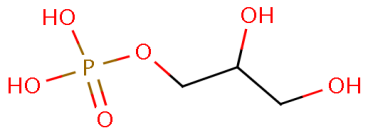
5ENL



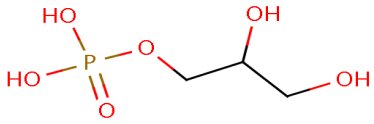
6RNT



6TIM



7TIM



Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo