Learning and Nonlinear Models (L&NLM) – Journal of the Brazilian Neural Network Society, Vol. 8, Iss. 1, pp. 52-62, 2010

ESTUDO DE REPRESENTAÇÕES EM REDES NEURAIS PARA ANÁLISE DA AFINIDADE DE LIGAÇÃO DE COMPOSTOS ANTI-HIV

Davi F. Duarte*†º, Camila S. De Magalhães*§, Antônio C. A. Mol†, Ernesto R. Caffarena*

*Programa de Computação Científica – Fundação Oswaldo Cruz (PROCC, Fiocruz)
Av. Brasil 4365 - 21045-900 Rio de Janeiro, RJ, Brasil

†Departamento de Ciência da Computação, Universidade Gama Filho (UGF)
Rua Manoel Vitorino, 625, Piedade - 20740-280 Rio de Janeiro, RJ, Brasil

§Departamento de Matemática, Universidade Federal Rural do Rio de Janeiro (UFRRJ)
BR-465, Km 7 - 23890-000 Seropédica, RJ, Brasil

ºPrograma de Pós-Graduação em Sistemas e Computação, Instituto Militar de Engenharia (IME)
Praça General Tibúrcio, 80, Praia Vermelha - 22290-270 Rio de Janeiro, RJ, Brasil

E-mails: davi.faisca@yahoo.com.br, camilasm@ufrrj.br, mol@ien.gov.br, ernesto@fiocruz.br

Abstract: The rapid and accurate determination of the binding affinity between ligands and their receptors would be of enormous benefit in structure-based rational drug design. This fact would allow the analysis of the affinity of a large number of compounds before they were chemically synthesized and experimentally evaluated. In this work, we evaluate the use of Artificial Neural Networks (ANNs) for the binding affinity analysis of compounds with potential anti-HIV activity. The method developed uses a General Regression Neural Network (GRNN) for binding affinity analysis based on structural information and binding mode of the compounds. A data set of 90 experimental structures of HIV-1 protease inhibitors, with known binding affinity were obtained from the Protein Data Bank (PDB) and used for training and testing the neural network. Several ways of ligand structure representation were developed, using a three-dimensional grid and molecular descriptors as input to the ANN. The ANN was studied for classification of compounds in activity levels. The results indicate that the method can become a useful tool in computer-aided drug design area.

Keywords: Ligand-receptor Binding Affinity, Artificial Neural Networks, HIV-1 Protease, AIDS.

Resumo: A determinação rápida e acurada da afinidade de ligação entre ligantes e seus receptores, seria de enorme benefício para a área de desenho racional de fármacos. Este fato possibilitaria a análise da afinidade de um grande número de compostos antes que eles fossem quimicamente sintetizados e avaliados experimentalmente, tornando mais ágil o processo como um todo. Neste trabalho, a utilização de Redes Neurais Artificiais (RNA) para a análise da afinidade de ligação de compostos com potencial atividade anti-HIV é investigada. O método desenvolvido utiliza uma rede neural de regressão genérica (GRNN – General Regression Neural Network) para análise da afinidade de ligação com base nas informações estruturais e no modo de ligação dos compostos. Estruturas experimentais de 90 inibidores da enzima HIV-1 protease, com afinidade de ligação conhecida, foram obtidas do banco de estruturas moleculares Protein Data Bank (PDB) e utilizadas para treinamento e teste da rede neural. Foram desenvolvidas diversas maneiras de representação das estruturas dos ligantes, a partir de uma malha tridimensional e por descritores moleculares, utilizadas como entrada para a rede. A RNA foi estudada para discriminação de compostos em níveis de atividade. Os resultados indicam que o método desenvolvido pode se tornar uma ferramenta útil para o desenho racional de fármacos.

Palavras-chave: Afinidade de Ligação Receptor-ligante, Redes Neurais Artificiais, HIV-1 Protease, AIDS.

1 Introdução

O processo de reconhecimento molecular receptor-ligante é a base para o desenvolvimento de fármacos. Os métodos para a análise e predição da afinidade de ligação receptor-ligante são uma parte importante da área de Desenho Racional de Fármacos Baseado em Estrutura (DRBE) [1]. O DRBE visa à identificação e a uma maior compreensão das interações moleculares entre receptor e ligante, envolvendo a utilização de métodos computacionais baseados nas estruturas tridimensionais das moléculas interagentes para o desenvolvimento de compostos candidatos a novos fármacos. Métodos computacionais que possam predizer a afinidade de ligação receptor-ligante ou, em uma fase inicial do processo, diferenciar compostos em níveis de afinidade distintos, podem ser utilizados tanto para a descoberta de novas substâncias bioativas – através de técnicas conhecidas como *virtual screening* – quanto para o refinamento e a otimização de compostos bioativos previamente identificados. O objetivo desses métodos é a obtenção de uma estimativa acurada da afinidade de ligação, *i.e.*, da constante de

inibição (K_i), observada experimentalmente. Entretanto, os processos relacionados à afinidade de ligação receptor-ligante são complexos, envolvendo uma combinação de efeitos entálpicos e entrópicos. Embora uma grande diversidade de abordagens teóricas e empíricas tenha sido proposta, o desenvolvimento de métodos rápidos e acurados para a análise/predição da afinidade de ligação receptor-ligante permanece como um dos principais desafios da área [2,3].

Redes Neurais Artificiais (RNAs) são métodos computacionais inspirados no funcionamento e comportamento do cérebro humano que têm sido aplicados com sucesso em problemas complexos de várias áreas do conhecimento [4]. A principal vantagem das RNAs em relação aos métodos tradicionais está na capacidade de "aprenderem" com a experiência, dotando aos sistemas computacionais de aspectos cognitivos. Essas características tornam as RNAs metodologias promissoras para a análise e predição da afinidade de ligação de moléculas ligantes [5].

Neste trabalho, uma Rede Neural de Regressão Genérica (GRNN) [6] foi utilizada para a análise da afinidade de ligação de compostos anti-HIV. As estruturas tridimensionais e os dados de afinidades de ligação de complexos HIV-1 protease-ligante foram utilizados para treinamento e testes de uma RNA para discriminar compostos em níveis de afinidade distintos. A enzima HIV-1 protease é um alvo molecular importante para o tratamento da AIDS (SIDA - Síndrome da Imunodeficiência Adquirida) e está diretamente relacionada ao processo de reprodução do vírus HIV. O método desenvolvido se baseia principalmente na utilização de uma malha tridimensional, onde são armazenadas informações sobre os ligantes, que são posteriormente utilizadas como entrada para a rede neural.

2 Metodologia

2.1 Conjunto de Dados

Estruturas experimentais de 90 inibidores da enzima HIV-1 protease e dados de afinidade de ligação foram utilizados para treinamento e teste de uma rede neural. A Figura 1 apresenta um exemplo da estrutura da enzima HIV-1 protease complexada com o ligante DMP. As estruturas tridimensionais, com resolução inferior a 2.8 Å, foram obtidas do banco de estruturas moleculares Protein Data Bank (PDB) [7].

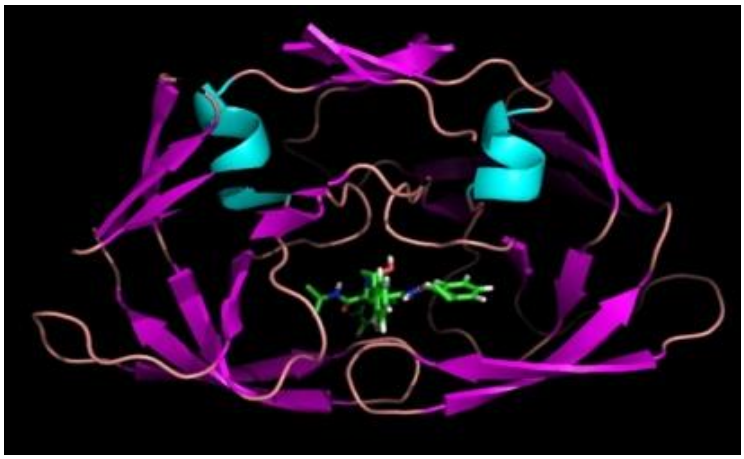
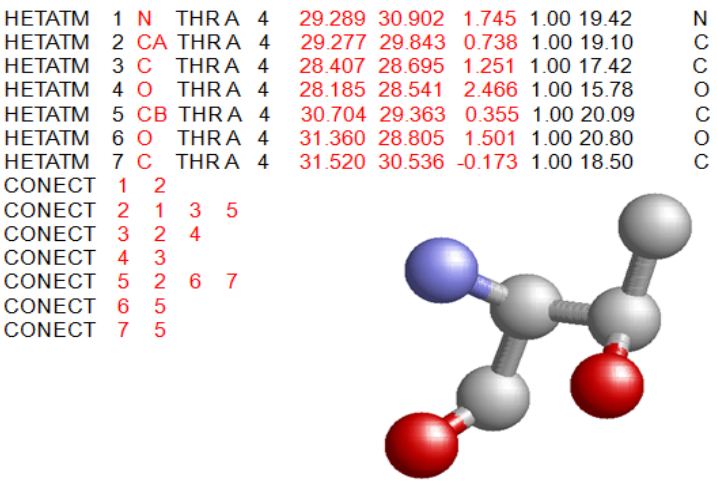


Figura 1. Exemplo do complexo proteína-ligante, com o ligante DMP (em verde, no centro) no sítio ativo da enzima HIV-1 protease (em rosa e azul) (Programa de visualização molecular RasMol [8]).

Um arquivo PDB contém as coordenadas cartesianas (x, y, z) de todos os átomos do ligante e da proteína, e também as ligações entre esses átomos, que definem a estrutura tridimensional (Figura 2). Somente os dados moleculares dos ligantes, sem a proteína complexada, foram utilizados neste trabalho.

As constantes de inibição (K_i) foram obtidas dos bancos de afinidades de ligação: BindingDB [9], BindingMoad [10] e também de outros trabalhos disponíveis na literatura especializada [11]. Os valores de K_i dos ligantes selecionados variam de 0.0021 nM a 9600 nM, com a média de 202.62 nM. Valores mais baixos de K_i indicam ligantes com maior capacidade de inibição. Os 90 compostos selecionados (Tabela 1) possuem diferentes características físico-químicas e estruturais. Esses inibidores foram classificados em 12 famílias de ligantes análogos, através de análise visual. Ligantes com grupamentos químicos semelhantes foram classificados como pertencentes à mesma família.



```
HETATM  1  N   THR A  4     29.289  30.902   1.745  1.00 19.42      N
HETATM  2  CA  THR A  4     29.277  29.843   0.738  1.00 19.10      C
HETATM  3  C   THR A  4     28.407  28.695   1.251  1.00 17.42      C
HETATM  4  O   THR A  4     28.185  28.541   2.466  1.00 15.78      O
HETATM  5  CB  THR A  4     30.704  29.363   0.355  1.00 20.09      C
HETATM  6  O   THR A  4     31.360  28.805   1.501  1.00 20.80      O
HETATM  7  C   THR A  4     31.520  30.536  -0.173  1.00 18.50      C
CONECT  1  2
CONECT  2  1  3  5
CONECT  3  2  4
CONECT  4  3
CONECT  5  2  6  7
CONECT  6  5
CONECT  7  5
```

Figura 2. Exemplo de um arquivo PDB representando o aminoácido Treonina (Thr), (Programa de visualização molecular RasMol [8]).

Tabela 1. Códigos dos arquivos PDB das estruturas HIV-1 protease-ligante obtidas do Protein Data Bank.

Arquivos PDB
1A8G, 1AAQ, 1AID, 1AJV, 1AJX, 1C6Z, 1C70, 1D4H, 1D4I, 1D4J, 1DIF, 1DMP, 1EBW, 1EBY, 1EBZ, 1EC0, 1EC1, 1EC2, 1G2K, 1G35, 1GNO, 1HBV, 1HEF, 1HEG, 1HIH, 1HIV, 1HOS, 1HPS, 1HPV, 1HPX, 1HSG, 1HVH, 1HVI, 1HVJ, 1HVK, 1HVL, 1HVR, 1HWR, 1HXB, 1HXW, 1IIQ, 1M0B, 1MTR, 1ODW, 1ODY, 1OHR, 1PRO, 1QBR, 1QBS, 1QBT, 1QBU,1SBG, 1T7K, 1VIJ, 1WBK, 1WBM, 1XL2, 1XL5, 1ZP8, 1ZPA, 1ZSF, 1ZSR, 2BPV, 2BPX, 2BPY, 2BQV, 2CEJ, 2CEM, 2CEN, 2FDE, 2I4D, 2I4U, 2I4W, 2P3B, 2PQZ, 2PWC, 2PWR, 2QNN, 2QNP, 2QNQ, 2UXZ, 3BGB, 3BGC, 3TLH, 4HVP, 5HVP, 7HVP, 7UPJ, 8HVP, 9HVP

2.2 Preparação do Conjunto de Dados

As estruturas de todos os inibidores foram sobrepostas em uma mesma referência, preservando o modo de ligação do ligante no sítio ativo (Figura 3). A sobreposição dos ligantes foi feita com o programa Swiss-Pdb Viewer [12], utilizando-se como referência a estrutura do ligante indinavir (arquivo PDB 1HSG).

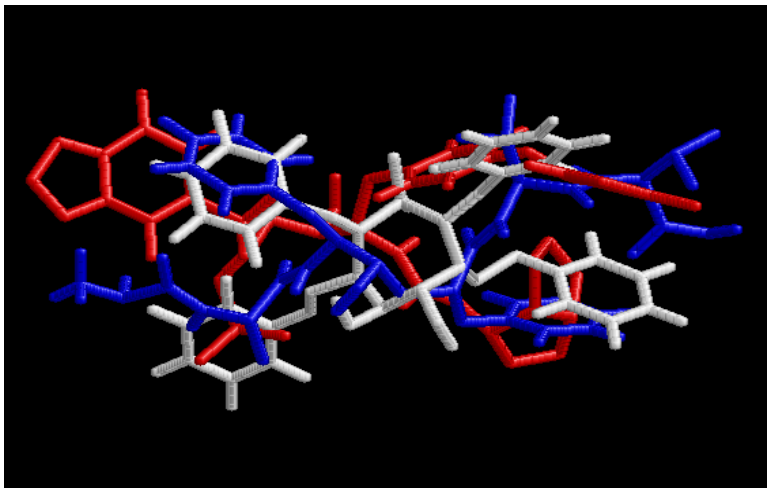


Figura 3. Exemplo de três ligantes sobrepostos em uma mesma orientação espacial (Programa de visualização molecular RasMol [8]).

O servidor PRODRG 2.5 Beta [13], foi utilizado para a inclusão de hidrogênios nos átomos polares e nos anéis aromáticos dos ligantes. As cargas parciais atômicas dos átomos de cada ligante foram calculadas com o programa MOE [14] utilizando o campo de força MMFF94 [15].

2.3 Representação das Estruturas

As representações desenvolvidas e analisadas neste trabalho podem ser classificadas em dois principais tipos: representação em malha tridimensional, levando em consideração a posição do ligante na região de ligação da enzima (sítio ativo); e representação por descritores moleculares, abrangendo características físico-químicas e estruturais do ligante.

2.3.1 Representação em Malha Tridimensional

Uma malha tridimensional englobando todos os ligantes sobrepostos foi gerada (Figura 4), com base nas coordenadas cartesianas dos átomos dos ligantes. A malha engloba a região do sítio ativo da proteína, contendo todas as coordenadas dos átomos do ligante. Neste trabalho foram desenvolvidos programas em linguagem C++ para gerar a representação das estruturas dos ligantes em malha 3D. Dois tipos de representação em malha foram analisados: uma informando o tipo de átomo da molécula (MA) e outra a carga parcial do átomo (MC). Para cada átomo de cada ligante, o ponto da malha mais próximo a esse átomo é identificado. Nesse ponto identificado, é atribuída uma informação referente ao átomo do ligante associado.

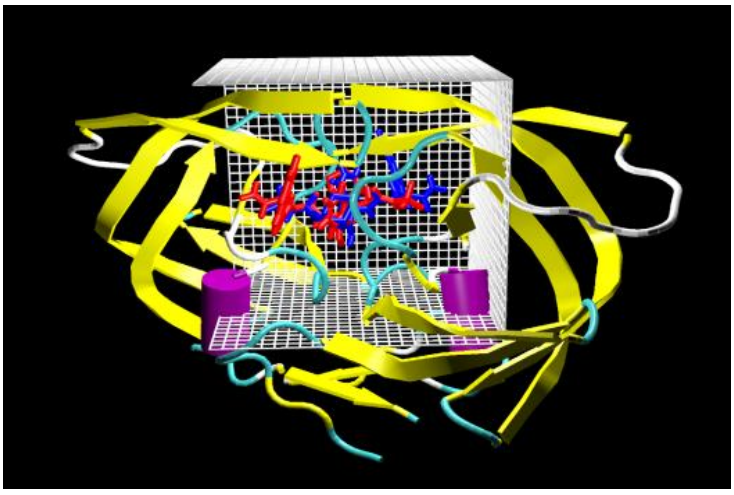


Figura 4. Malha tridimensional com complexo HIV-1 protease-ligante, englobando a região do sítio ativo da proteína (Figura retirada de [16] com autorização dos autores).

Na representação em malha tridimensional com o tipo de átomo (MA), no ponto identificado como o mais próximo ao átomo, é atribuído um determinado valor de 1 a 10, de acordo com o tipo de átomo correspondente (Tabela 2). Para os outros pontos da malha, que não tenham sido associados a nenhum átomo, o valor zero é atribuído.

Tabela 2. Valor atribuído à malha de acordo com o tipo de átomo do ligante.

Tipo de Átomo	Valor Correspondente
Carbono (C)	1
Hidrogênio (H)	2
Nitrogênio (N)	3
Oxigênio (O)	4
Enxofre (S)	5
Flúor (F)	6
Fósforo (P)	7
Bromo (Br)	8
Cloro (Cl)	9
Iodo (I)	10

Na representação em malha com cargas parciais atômicas (MC), no ponto identificado como o mais próximo ao átomo do ligante é atribuída a sua carga parcial atômica.

Para as duas representações em malha (MA e MC), a malha gerada possui dimensões de 27 Å x 22 Å x 18 Å com espaçamento de 2.5 Å. Assim, cada ligante é representado por um vetor de 693 posições (correspondentes aos pontos da malha), contendo informações sobre os tipos de átomos (MA) ou sobre as cargas dos átomos (MC), que são utilizados como entrada para a RNA.

2.3.2 Representação por Descritores Moleculares

O uso de descritores moleculares (RD) foi analisado para a representação dos ligantes para a RNA. Foi desenvolvido um programa em linguagem C++ para calcular 15 descritores moleculares, abrangendo a constituição molecular do ligante e alguns grupamentos químicos (Tabela 3). A entrada para a RNA é constituída por um vetor de 15 posições, onde cada posição corresponde a um descritor calculado.

Tabela 3. Descritores Moleculares.

Descritores Moleculares
Número de Total de Átomos
Número de Grupamentos Hidroxila OH
Número de Grupamentos NH
Número de Grupamentos Amina (NH2)
Número de Anéis Aromáticos
Número de Átomos Carbono (C)
Número de Átomos Hidrogênio (H)
Número de Átomos Nitrogênio (N)
Número de Átomos Oxigênio (O)
Número de Átomos Enxofre (S)
Número de Átomos Flúor (F)
Número de Átomos Fósforo (P)
Número de Átomos Bromo (Br)
Número de Átomos Cloro (Cl)
Número de Átomos Iodo (I)

Também foram analisadas combinações das representações, sendo: malha tridimensional com tipo de átomo e descritores moleculares (MA+RD), e malha tridimensional com cargas parciais atômicas juntamente com descritores moleculares (MC+RD). Nos dois casos, a entrada para a RNA é constituída por um vetor de 708 posições abrangendo os dois tipos de representação (693 malha + 15 descritores).

2.3.3 Representação por Descritores Moleculares pela Distância

Além das representações acima, um novo tipo de representação por descritores, que também considera a posição do ligante no sítio ativo da proteína, foi analisado. Nesta representação, por descritores moleculares pela distância (DD), vários níveis de distância foram definidos, e os descritores moleculares da Tabela 3, foram calculados separadamente para todos os átomos pertencentes a cada nível de distância, que foram definidos considerando-se a distância dos átomos do ligante em relação ao centro do sítio ativo da proteína.

O primeiro nível de distância inclui todos os átomos que estão a uma distância de 2 Å do ponto central da malha (utilizada nas representações anteriores), representando o centro do sítio ativo da proteína ($x = 11$, $y = 21.5$ e $z = 6.5$). O segundo nível de distância inclui todos os átomos do ligante que possuem distância do ponto central maior do que 2 Å e menor do que 4 Å. Assim, foram definidos oito níveis de distância, com intervalos de 2 Å cada. Em cada nível, todos os descritores da Tabela 3 foram calculados para os átomos do ligante correspondentes ao nível, com exceção da quantidade de anéis aromáticos, que foi obtida apenas uma vez para cada ligante. A entrada para a RNA é constituída por 113 variáveis, onde cada grupo de 14 variáveis corresponde aos descritores calculados nos oito níveis de distância.

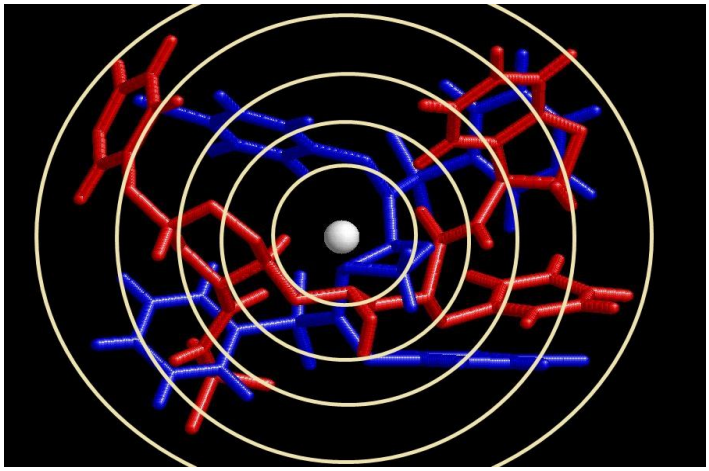


Figura 5. Exemplo da representação por Descritores Moleculares pela Distância, descritores calculados para os átomos pertencentes à distância de 2 Å a partir do ponto central (Programa de visualização molecular RasMol [8]).

2.4 Rede Neural

Para avaliação do modelo desenvolvido, a RNA foi estudada para classificação dos compostos em níveis de afinidade. Os compostos foram classificados de acordo com o seu valor da afinidade de ligação (K_i). Foram feitos dois tipos de testes: classificando os compostos em duas ou três faixas de afinidade. O número de compostos em cada classe, e o valor utilizado como saída desejada para a RNA, para os dois casos testados, são mostrados nas Tabelas 4 e 5, respectivamente.

Tabela 4. Classificação dos compostos em dois níveis de afinidade.

Afinidade (nM) [a]	Nº de compostos [b]	Valor [c]
$K_i \leq 10$	64	1
$K_i > 10$	26	2

[a] Faixa de valores utilizado para classificação;
[b] Número de compostos classificados por faixa de afinidade;
[c] Valor atribuído à rede como saída desejada.

Tabela 5. Classificação dos compostos em três níveis de afinidade.

Afinidade (nM) [a]	Nº de compostos [b]	Valor [c]
$K_i \leq 1$	37	1
$1 < K_i \leq 10$	27	2
$K_i > 10$	26	3

[a] Faixa de valores utilizado para classificação;
[b] Número de compostos classificados por faixa de afinidade;
[c] Valor atribuído à rede como saída desejada.

Foi utilizada uma Rede Neural de Regressão Genérica (GRNN) [6], que utiliza algoritmos genéticos para minimizar o erro médio quadrático, com o simulador de redes neurais NeuroShell [17]. A representação dos ligantes foi utilizada como entrada da rede e seus valores correspondentes à faixa de afinidade foram utilizados como saída desejada. A rede possui uma camada de entrada, com o número de neurônios correspondentes à quantidade de variáveis utilizadas em cada representação. A quantidade de neurônios para cada representação é mostrada na Tabela 6.

Os modelos de representação desenvolvidos foram validados com a utilização de um conjunto externo de dados e com validação cruzada (Leave-One-Out). Na validação com conjunto externo, o conjunto de dados com 90 compostos foi separado em dois subconjuntos: um subconjunto de treinamento e um subconjunto de testes, escolhido de forma aleatória. A rede foi avaliada com a utilização de cinco conjuntos de treinamento/teste distintos, onde foi calculada a média de acertos com os cinco conjuntos de teste. É importante ressaltar que os dados utilizados para teste da RNA não foram utilizados em nenhum

momento durante a fase de treinamento. Na validação cruzada (Leave-One-Out), todo o conjunto de ligantes é utilizado para treinamento, exceto um, que é utilizado para teste. Este procedimento é repetido para todos os ligantes, deixando de fora um ligante diferente por vez.

Tabela 6. Arquitetura da RNA.

Representações	**Número de Neurônios por Camada**			
	Entrada	Oculta		Saída
		Conj. Externo	Val. Cruzada	
MA	693	78	89	1
MC	693	78	89	1
RD	15	78	89	1
MA+RD	708	78	89	1
MC+RD	708	78	89	1
DD	113	78	89	1

3 Resultados

3.1 Classificação de Compostos em Duas Faixas de Afinidade

Os resultados obtidos para as seis representações analisadas, na validação cruzada e na validação com conjunto externo, são mostrados nas Figuras 6 e 7, respectivamente. Com a utilização do conjunto de treinamento a taxa de sucesso para discriminar compostos em dois níveis de afinidade foi de 100% para as seis representações.

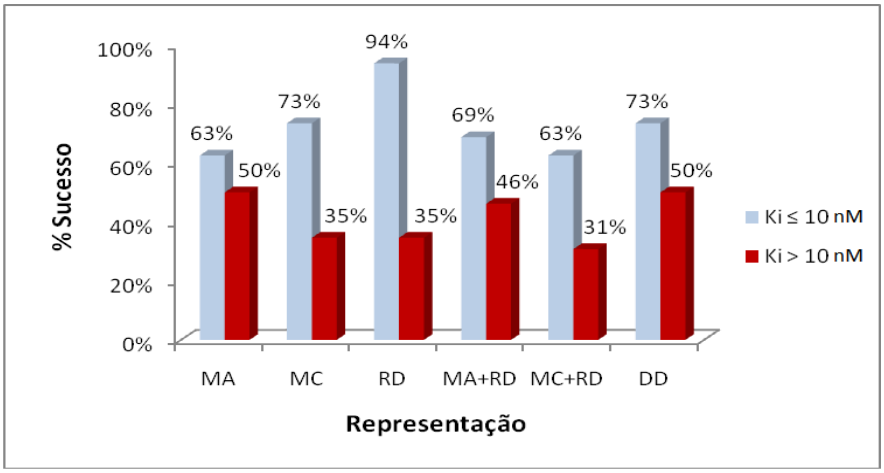


Figura 6. Porcentagem média de sucesso obtida com validação cruzada para as 6 representações analisadas para discriminação de compostos em duas faixas de afinidade. MA: representação malha tridimensional com tipo de átomo; MC: representação malha tridimensional com cargas parciais; RD: representação por descritores moleculares; MA+RD: representação MA juntamente com RD; MC+RD: representação MC juntamente com RD; DD: representação por descritores moleculares pela distância.

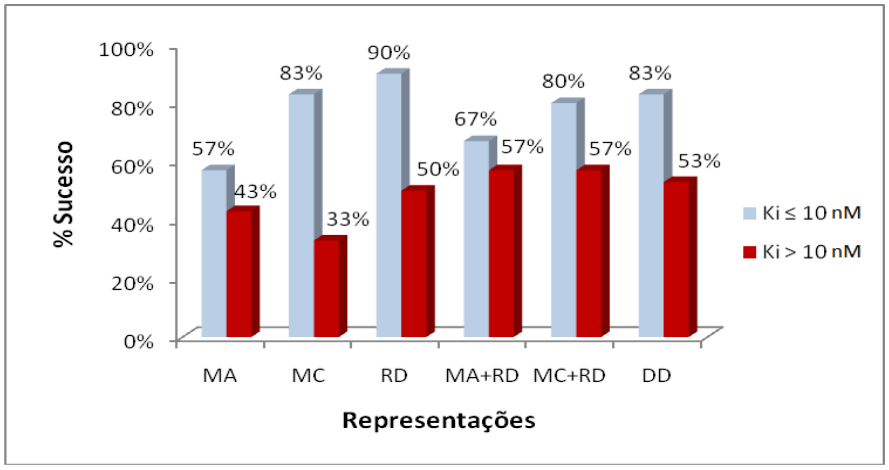


Figura 7. Porcentagem média de sucesso obtida com validação com conjunto externo para as 6 representações analisadas para discriminação de compostos em duas faixas de afinidade. MA: representação malha tridimensional com tipo de átomo; MC: representação malha tridimensional com cargas parciais; RD: representação por descritores moleculares; MA+RD: representação MA juntamente com RD; MC+RD: representação MC juntamente com RD; DD: representação por descritores moleculares pela distância.

Para efeito de comparação entre as representações analisadas foram considerados os resultados obtidos com a validação cruzada. Para classificação de compostos com $K_i \leq 10$ nM, a taxa de sucesso foi superior a 60% para todas as representações analisadas. Os melhores resultados foram obtidos com as representações descritores moleculares (RD), malha com cargas (MC) e descritores pela distância (DD), com taxa de sucesso de 94%, 73% e 73%, respectivamente.

Para classificação de compostos com $K_i > 10$ nM, a taxa de sucesso média obtida foi de aproximadamente 30% para as representações malha com cargas (MC), descritores moleculares (RD) e malha com cargas juntamente com descritores (MC+RD). Entretanto, para as outras representações (representação em malha com átomos (MA) e por descritores pela distância (DD)) a taxa de sucesso foi de 50%, com exceção da representação em malha com átomos juntamente com descritores (MA+RD), com 46% de sucesso.

Além da validação cruzada, o desempenho da RNA com as 6 representações utilizadas foi analisado com a utilização de um conjunto teste externo. Para classificação de ligantes com $K_i \leq 10$ nM, foi obtido 80% de sucesso ou mais para as representações malha com cargas (MC), descritores moleculares (RD), malha com cargas juntamente com descritores (MC+RD) e descritores pela distância (DD), sendo que para a representação descritores moleculares (RD), uma taxa de sucesso média de 90% foi obtida.

Assim como na validação cruzada, a taxa de sucesso média para classificar ligantes com $K_i > 10$ nM, foi inferior aos resultados obtidos para classificação de compostos com $K_i \leq 10$ nM. As representações por descritores moleculares (RD), malha com átomos juntamente com descritores (MA+RD), malha com cargas juntamente com descritores (MC+RD) e descritores pela distância (DD), alcançaram taxa de sucesso média de aproximadamente 54%, enquanto que, com as representações malha com átomos (MA) e malha com cargas (MC), foram obtidas taxas de sucesso médias de 43% e 33%, respectivamente.

3.2 Classificação de Compostos em Três Faixas de Afinidade

Neste teste, a capacidade da RNA em discriminar compostos em três níveis de afinidade foi avaliada. A taxa de sucesso com utilização do conjunto de treinamento foi de 100%. A porcentagem média de sucesso para as seis representações analisadas, na validação cruzada e na validação com conjunto externo, é mostrada nas Figuras 8 e 9, respectivamente.

Para discriminação dos compostos em três níveis de afinidade, também foram considerados os resultados obtidos com a validação cruzada para efeito de comparação entre as representações analisadas. Para classificação de compostos com $K_i \leq 1$ nM, os melhores resultados foram obtidos com as representações malha com átomos (MA) e descritores moleculares (RD) com taxa de sucesso de 59% e 62%, respectivamente.

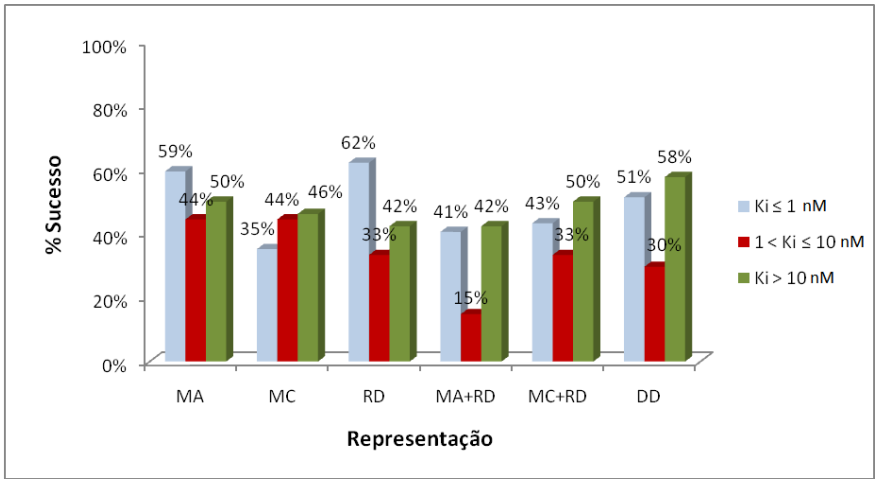


Figura 8. Porcentagem média de sucesso obtida com validação cruzada para as 6 representações analisadas para discriminação de compostos em três faixas de afinidade. MA: representação malha tridimensional com tipo de átomo; MC: representação malha tridimensional com cargas parciais; RD: representação por descritores moleculares; MA+RD: representação MA juntamente com RD; MC+RD: representação MC juntamente com RD; DD: representação por descritores moleculares pela distância.

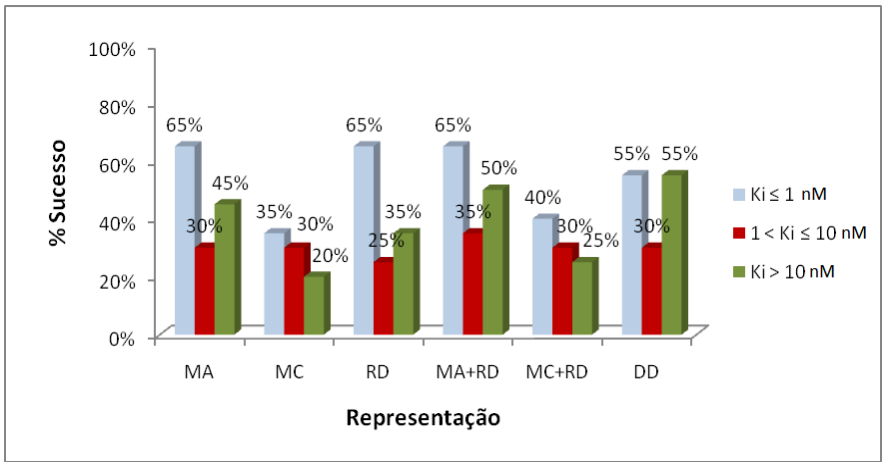


Figura 9. Porcentagem média de sucesso obtida com validação com conjunto externo para as 6 representações analisadas para discriminação de compostos em três faixas de afinidade. MA: representação malha tridimensional com tipo de átomo; MC: representação malha tridimensional com cargas parciais; RD: representação por descritores moleculares; MA+RD: representação MA juntamente com RD; MC+RD: representação MC juntamente com RD; DD: representação por descritores moleculares pela distância.

Para as outras representações (MC, MA+RD, MC+RD e DD), a taxa de sucesso obtida foi em torno de 42%. Para ligantes com K_i entre 1 nM e 10 nM, a taxa de sucesso obtida foi de 44% para as representações malha com átomos (MA) e malha com cargas (MC), 30% a 33% para as representações descritores moleculares (RD), malha com cargas juntamente com descritores (MC+RD) e descritores pela distância (DD). O pior resultado foi obtido com a representação malha com átomos juntamente com descritores (MA+RD), com taxa de sucesso de apenas 15%. Para ligantes com $K_i > 10$ nM, uma taxa de sucesso entre 50% e 58%, foi obtida para as representações malha com átomos (MA), malha com cargas juntamente com descritores (MC+RD) e descritores pela distância (DD). Para as outras representações (MC, RD e MA+RD), a taxa de sucesso obtida foi de aproximadamente 43%.

A validação com a utilização de um conjunto externo de teste também foi realizada para classificação de compostos em 3 faixas de afinidade. Para classificação dos ligantes com $K_i \leq 1$ nM, a taxa média de sucesso foi de 65% para as representações malha com átomos (MA), descritores moleculares (RD) e malha com átomos juntamente com descritores (MA+RD). Para as outras representações (MC, MC+RD e DD), a taxa de sucesso média obtida foi entre 35% e 55%. Para ligantes com K_i entre

1 nM e 10 nM, a taxa de sucesso média obtida foi de 30% para todas as representações. Neste caso, a representação malha com átomos juntamente com descritores (MA+RD) obteve o melhor resultado, com 35% de sucesso. Para ligantes com $K_i > 10$ nM, a taxa de sucesso média foi de 50% e 55%, para as representações malha com átomos juntamente com descritores (MA+RD) e descritores pela distância (DD), respectivamente. Para as outras representações (MA, MC, RD, MC+RD), a taxa de sucesso ficou entre 20% e 45%, e o pior resultado foi obtido com a representação malha com cargas (MC), com taxa média de sucesso de apenas 20%.

4 Conclusão

Neste trabalho, seis modos de representação de ligantes em uma Rede Neural de Regressão Genérica (GRNN) foram analisados para classificação de compostos anti-HIV em dois e três níveis de afinidade distintos.

Na classificação de compostos em dois níveis de afinidade, os resultados revelam que o modelo desenvolvido pode identificar ligantes altamente ativos ($K_i \leq 10$ nM) com mais de 90% de sucesso (representação por descritores moleculares - RD). Para ligantes com $K_i > 10$ nM, a taxa de sucesso obtida foi maior ou igual 46% para 3 das representações testadas (MA, MA+RD e DD). O melhor resultado obtido na classificação para ligantes com $K_i \leq 10$ nM, pode ser atribuído ao maior número de ligantes nesta classe em comparação à outra (Tabela 4). Na classificação em três níveis de afinidade, os resultados foram similares entre as representações analisadas. Os melhores resultados foram obtidos com as representações MA, RD e DD, com taxas de sucesso médias (entre as três faixas de afinidade) de 51%, 45.7% e 46.3%, respectivamente.

De maneira geral, as representações com utilização dos descritores moleculares (RD e DD) mostraram-se viáveis para classificação de compostos com afinidade anti-HIV. Nós acreditamos que a alta dimensionalidade (número de variáveis) em relação ao número de elementos de treinamento tenha influenciado a capacidade de generalização da GRNN. Os resultados obtidos sugerem que o modelo desenvolvido pode ser melhorado com a inclusão de um maior número de ligantes por faixa de afinidade, podendo se tornar uma ferramenta útil para a discriminação de compostos na área de desenho racional de fármacos baseado em estrutura.

5 Agradecimentos

Os autores agradecem à FAPERJ e ao CNPq pelos recursos financeiros.

6 Referências Bibliográficas

[1] Waszkowycz B., “Towards Improving Compound Selection in Structure-Based Virtual Screening". Drug Discovery Today, Vol. 13, No. 5-6., 219-226. 2008.

[2] Kitchen D. B., Decornez H., Furr J. R., Bajorath J., “Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications”. Nature Reviews Drug Discovery, Vol. 3, No. 11, 935-949. 2004.

[3] Leach A. R., Brian S. K., and Peishoff C. E., “Prediction of Protein−Ligand Interactions. Docking And Scoring: Successes and Gaps”. Journal of Computational Chemistry, 49 (20), pp 5851–5855. 2006.

[4] Haykin S., “Neural Networks. A Comprehensive Foundation”. New Jersey, Prentice-Hall. 1999.

[5] Fabry-Asztalos L., Andonie R., Collar CJ., Abdul-Wahid S. Salim N., “A Genetic Algorithm Optimized Fuzzy Neural Network Analysis of The Affinity of Inhibitors for HIV-1 Protease”. Bioorganic Medicinal Chemistry. 16(6):2903-11, 2008.

[6] Specht D.F., “A General Regression Neural Network, IEEE Transactions on Neural Networks”, Vol. 2, Issue 6 568–576. 1991.

[7] Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E., “The Protein Data Bank”. Nucleic Acids Research. 28:235-242. 2000.

[8] RasMol And OpenRasmol - http://www.openrasmol.org/

[9] Liu T., Lin Y., Wen X., Jorissen R.N. and Gilson M.K., “Bindingdb: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities”. Nucleic Acids Research. 35:198-201. 2007.

[10] Hu L., Benson M.L., Smith R.D., Lerner M.G., Carlson H.A.. Binding MOAD (Mother Of All Databases). Proteins 60, 333-40. 2005.

[11] Blum A., Böttcher J., Heine A., Klebe G., and Diederich W. E., “Structure-Guided Design of C2-Symmetric HIV-1 Protease Inhibitors Based on a Pyrrolidine Scaffold”. Journal of Medicinal Chemistry 51:2078–2087. 2008.

[12] Guex. N., Peitsch. M.C., "SWISS-MODEL and the Swiss-Pdbviewer: An Environment for Comparative Protein Modeling". Electrophoresis. 18:2714-2723. 1997.

[13] Schuettelkopf A. W., Van Aalten D. M. F., "PRODRG - A Tool For High-Throughput Crystallography of Protein-Ligand Complexes". Acta Crystallography, disponível no site: http://davapc1.bioch.dundee.ac.uk/cgi-bin/prodrg_beta. D60. 1355--1363. 2004.

[14] Molecular Operating Environment (MOE 2004.03); Chemical Computing Group Inc.. 2004.

[15] Halgren T. A., "Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94". Journal of Computational Chemistry, v. 17, p. 490–519. 1996.

[16] De Magalhaes C. S., Barbosa H. J. C., Dardenne L. E., "Métodos de Docking Receptor-Ligante para o Desenho Racional de Compostos Bioativos". In: Nelson H. Morgon; Kaline Coutinho. (Org.). Métodos de Química Teórica e Modelagem Molecular. 1ª ed. São Paulo, Livraria da Física, 489-531. 2007.

[17] Frederick; NeuroShell 2, release 3. Ward System Group Inc. 1996.