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

DAVI F. DUARTE*†, CAMILA S. DE MAGALHÃES*, ANTONIO 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
Rua Manoel Vitorino, 625, Piedade
20740-280 - Rio de Janeiro, RJ, Brasil

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

Abstract— The rapid and accurate determination of the binding affinity between ligands molecules 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 analysis of the binding affinity 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 75 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. The ligand structures were represented in a three-dimensional grid and used as input to the network. 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, agilizando 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 75 inibidores da enzima HIV-1 protease, com afinidade de ligação conhecida, foram obtidos do banco de estruturas moleculares Protein Data Bank (PDB) e utilizadas para treinamento e teste da rede neural. Foi desenvolvido um modelo para a representação das estruturas dos ligantes em uma matriz tridimensional, utilizada 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) (Waszkowycz, 2008). O DRBE visa à identificação e 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 atividade distintos, podem ser utilizados tanto para a descoberta de novas substâncias bioativas – através de técnicas conhecidas como virtual screening – quanto para o refinamento e 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 tenham sido propostas, 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 (Kitchen *et al.*, 2004; Leach *et al.*, 2006).

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 (Haykin, 1999). A principal vantagem das RNAs em relação aos métodos tradicionais está na capacidade de "aprenderem" com a experiência, dotando sistemas computacionais de aspectos cognitivos. Essas características tornam as RNAs metodologias promissoras para a análise e

predição da afinidade de ligação de moléculas ligantes (Fabry-Asztalos *et al*., 2008).

Neste trabalho, uma Rede Neural de Regressão Genérica (GRNN) 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 atividade distintos. A enzima HIV-1 protease é um alvo molecular importante para o tratamento da AIDS (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 75 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 foram obtidas do banco de estruturas moleculares Protein Data Bank (PDB) (Berman *et al.*, 2000), com resolução inferior a 2.8 Å. Um arquivo PDB contém as coordenadas (x, y, z) de todos os átomos do ligante e da proteína. Somente as estruturas dos ligantes foram utilizadas, sendo retiradas as coordenadas dos átomos da proteína.

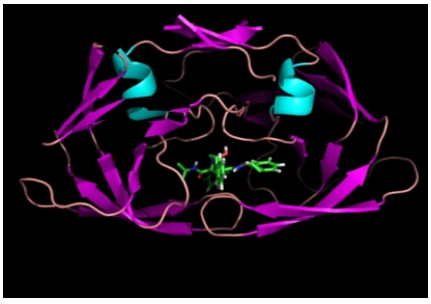

Figura 1. Exemplo do complexo HIV-1 Protease-ligante, com o ligante DMP no sítio ativo da enzima HIV-1 protease.

As afinidades de ligação (K_i) foram obtidas dos bancos de afinidades de ligação: BindingDB (Liu *et al.*, 2007), BindingMoad (Hu *et al.*, 2005) e da literatura (Blum *et al.*, 2007). Os valores de K_i dos ligantes selecionados variam de 0.0027 nM a 2150 nM, com a média de 88.43 nM. Valores mais baixos de K_i indicam ligantes com melhor afinidade de ligação. Os 75 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.

Tabela 1. Arquivo PDB das estruturas HIV-1 protease-ligante obtidas do Protein Data Bank

Arquivos PDB
1A8G, 1AAQ, 1AID, 1AJV, 1AJX, 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, 1WBM, 1XL2, 1XL5, 1ZP8, 1ZPA, 1ZSF, 1ZSR, 2BPY, 2BQV, 2FDE, 2I4U, 2PQZ, 2PWC, 2PWR, 2QNN, 2QNP, 2QNQ, 3TLH, 4HVP, 7HVP, 7UPJ, 8HVP, 9HVP

2.2 Preparação do Conjunto

As estruturas de todos os inibidores foram sobrepostas em uma mesma referência, preservando o modo de ligação do ligante no sítio ativo (Figura 2). A sobreposição dos ligantes foi feita utilizando como referência a estrutura do ligante indinavir (arquivo PDB 1HSG), com a utilização do programa Swiss-Pdb Viewer (Guex *et al*., 1997).

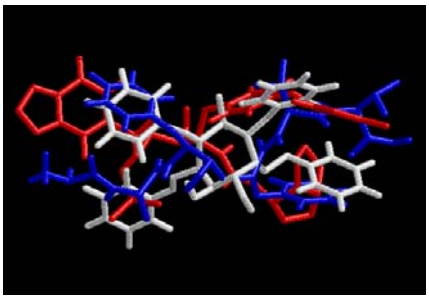

Figura 2. Exemplo de três ligantes sobrepostos em uma mesma orientação espacial.

O programa PRODRG 2.5 Beta (Schuettelkopf e Aalten, 2004), foi utilizado para a inclusão de hidrogênios nos átomos polares e nos anéis aromáticos dos ligantes.

2.3 Representação das Estruturas

Dois principais tipos de representação foram analisados: 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 do ligante.

Representação em Malha Tridimensional. Uma malha tridimensional englobando todos os ligantes sobrepostos foi gerada (Figura 3), com base nas coordenadas (x, y, z) dos átomos dos ligantes. A malha engloba a região do sítio ativo da proteína, contendo todos os átomos do ligante. Foi desenvolvido um programa, em linguagem C++, para gerar a representação das estruturas dos ligantes como entrada para a RNA. A malha gerada possui dimensões de 27 Å x 22 Å x 18 Å e espaçamento de 2.5 Å, com um total de 693 pontos. 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. Dois tipos de representação em malha foram analisados neste trabalho: uma informando o tipo de átomo da molécula (MA) e outra a carga parcial do átomo (MC).

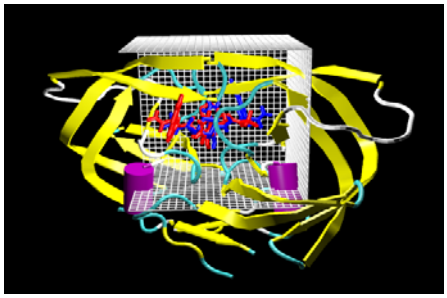

Figura 3. Malha tridimensional com complexo HIV-1 protease-ligante, englobando a região do sítio ativo da proteína.

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 é atribuída a sua carga parcial atômica. As cargas parciais foram calculadas com o programa MOE (Molecular Operating Environment, 2004) utilizando o campo de força MMFF94.

Representação por Descritores Moleculares. Outro modo de representação dos ligantes para a RNA foi analisado, com utilização de descritores moleculares (RD). Foi desenvolvido um programa para calcular 15 descritores moleculares, abrangendo a constituição molecular do ligante e alguns grupamentos (Tabela 3).

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)

Além das representações acima, também foram analisadas combinações das representações: Malha tridimensional com tipo de átomo e com descritores moleculares (MA+RD); e Malha tridimensional com cargas parciais atômicas juntamente com descritores moleculares (MC+RD), totalizando cinco formas de representar os ligantes para a RNA.

2.4 Rede Neural

Uma Rede Neural de Regressão Genérica (GRNN) (Specht, 1991), que utiliza algoritmos genéticos para minimizar o erro médio quadrático, foi utilizada para discriminação de compostos em faixas de afinidade. Foi utilizado o simulador de redes neurais NeuroShell (Ward Systems Group Inc, 1996). 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 utilizada em cada representação: 693 neurônios na representação em malha tridimensional; 15 neurônios na representação com descritores; e 708 neurônios quando combinada as duas representações utilizadas. Uma camada oculta, com o número de neurônios corresponde à quantidade de ligantes utilizados para o treinamento da rede. E uma camada de saída com um neurônio.

O desempenho da RNA foi avaliado com dois tipos de validações: com a utilização de um conjunto de dados externo e com a validação cruzada (Leave-One-Out). Na validação com conjunto externo, o conjunto de dados com 75 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. 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, cada vez deixando de fora um ligante diferente para a validação.

3 Resultados

Para avaliação do modelo desenvolvido, a RNA foi estudada para classificação dos compostos em níveis de atividade. Os compostos foram classificados de acordo com o seu valor da afinidade de ligação (K_i). Foram feitos dois tipos de testes: um para a RNA classificar os compostos em duas faixas de afinidade e outro para a RNA classificar os compostos em três faixas de afinidade. O número de compostos em cada classe, e o valor da 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]	Quantidade [b]	Valor [c]
$K_i \leq 10$	55	1
$K_i > 10$	20	2

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

Tabela 5. Classificação dos compostos em três níveis de afinidade

Afinidade (nM) [a]	Quantidade [b]	Valor [c]
$K_i \leq 1$	32	1
$1 < K_i \leq 10$	23	2
$K_i > 10$	20	3

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

4.1 Classificação de Compostos em Duas Faixas de Afinidade

Neste teste, a capacidade da RNA em discriminar compostos em dois 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 cinco representações analisadas, na validação com conjunto externo e na validação cruzada, são mostrados nas Figuras 4 e 5, respectivamente.

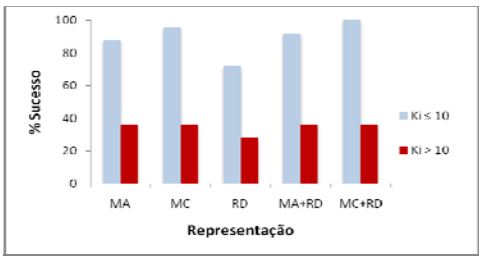

Figura 4. Resultado dos testes para discriminação de compostos em duas faixas de afinidade na validação com conjunto externo.[1]

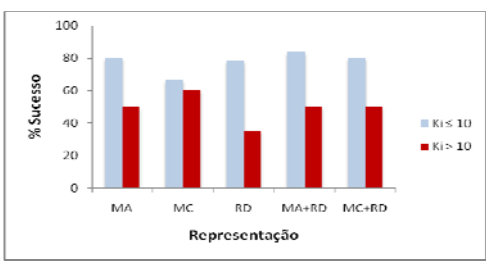

Figura 5. Resultado dos testes para discriminação de compostos em duas faixas de afinidade da validação cruzada.[1]

Nos testes para discriminação dos compostos com $K_i \leq 10$ nM, a taxa de sucesso média varia de 72% a 100% para as cinco representações analisadas, na validação com conjunto externo. Na validação cruzada a taxa de sucesso varia de 66% a 83% entre as representações. Nos testes para discriminação dos compostos com $K_i > 10$ nM, a taxa de sucesso média varia de 28% a 36% para as cinco representações analisadas, na validação com conjunto externo. Na validação cruzada a taxa de sucesso varia de 35% a 60%. As representações em malha tridimensional: MC, MA+RD e MC+RD apresentaram os melhores resultados para classificação dos compostos em duas faixas de afinidade.

4.2 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 cinco representações analisadas, na validação com conjunto externo e na validação cruzada, são mostrados nas Figuras 6 e 7, respectivamente.

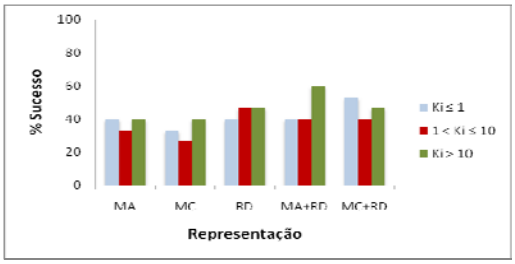

Figura 6. Resultado dos testes para discriminação de compostos em três faixas de afinidade da validação com conjunto externo.[1]

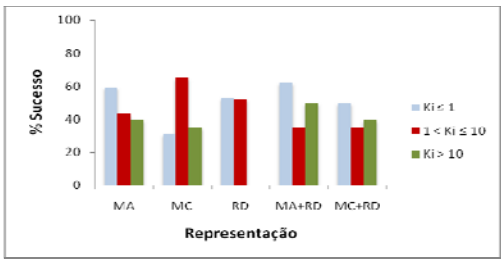

Figura 7. Resultado dos testes para discriminação de compostos em três faixas de afinidade da validação cruzada.[1]

Nos testes para discriminação dos compostos com $K_i \leq 1$ nM, a taxa de sucesso média varia de 33% a 53% na validação com conjunto externo, na validação cruzada a taxa de sucesso varia de 31% a 62%, para as cinco representações analisadas. Nos compostos com o K_i entre 1 nM e 10 nM, a taxa de

1 MA, representação malha tridimensional com tipo de átomo; MC, representação malha tridimensional com cargas parciais; RD, representação por descritores moleculares; MA+RD, representação malha tridimensional com tipo de átomo juntamente com descritores; MC+RD, representação malha tridimensional com cargas parciais juntamente com descritores.

sucesso média varia de 27% a 47% na validação com conjunto externo. E na validação cruzada taxa de sucesso entre 34.8% a 65.2%, para as cinco representações analisadas. Nos testes para discriminação dos compostos com $K_i > 10$ nM, a taxa de sucesso média varia de 40% a 60% na validação com conjunto externo e de 0% a 50% na validação cruzada. Assim como na classificação em duas faixas, as representações em malha tridimensional: MC, MA+RD e MC+RD apresentaram os melhores resultados para classificação dos compostos em três faixas de afinidade.

5 Conclusão

Neste trabalho, a utilização de uma Rede Neural de Regressão Genérica (GRNN) foi analisada para a classificação de compostos anti-HIV em níveis de atividade. Cinco modos de representação dos ligantes para a RNA foram analisados. O desempenho da RNA foi avaliado para classificação dos compostos em dois e três níveis de afinidade distintos.

Para classificação de compostos em dois níveis de afinidade, os resultados obtidos revelam que o modelo desenvolvido pôde identificar ligantes altamente ativos ($K_i \leq 10$ nM) com mais de 80% de sucesso (representação MA+RD). Para ligantes com $K_i > 10$ nM, a taxa de sucesso obtida com essa representação foi de 40%. O melhor resultado em relação às outras classes, obtido na classificação para ligantes com $K_i \leq 10$ nM, pode ser atribuído ao maior número de ligantes nesta classe em comparação às outras (Tabelas 4 e 5).

Na classificação em três níveis de afinidade, os melhores resultados foram obtidos com as representações MA+RD e MC+RD, com taxas de sucesso de 62.5% e 50%, respectivamente. Entretanto, em alguns casos, a taxa de sucesso obtida foi inferior a 30%. 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.

De maneira geral, a representação em malha tridimensional (com tipo de átomo e com cargas) mostrou-se viável para classificação de compostos com afinidade anti-HIV. 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.

Agradecimentos

Os autores agradecem à FAPERJ pelos recursos financeiros.

Referências Bibliográficas

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.

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.

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". Bioorg Med Chem.16(6):2903-11, 2008.

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

Guex. N. and Peitsch. M.C., "SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling". Electrophoresis. 18:2714-2723, 1997.

Haykin S., "Neural Networks. a Comprehensive Foundation". New Jersey. Prentice-Hall. 1999.

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.

Kitchen D. B., Decornez H., Furr J. R., Bajorath J.; Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov, Vol. 3, No. 11, pp. 935-949., 2004.

Leach A. R., S. Brian K., and Peishoff C. E.; Prediction of Protein−Ligand Interactions. Docking and Scoring: Successes and Gaps; J. Med. Chem., 49 (20), pp 5851–5855, 2006.

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.

Molecular Operating Environment, (MOE 2004.03); Chemical Computing Group Inc., 2004.

Schuettelkopf A. W. and 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.

Specht D.F., A general regression neural network, IEEE Transactions on Neural Networks, Vol. 2, Issue 6, pp. 568–576, 1991.

Waszkowycz B.; Towards improving compound selection in structure-based virtual screening. Drug Discov Today, Vol. 13, No. 5-6., pp. 219-226., 2008.