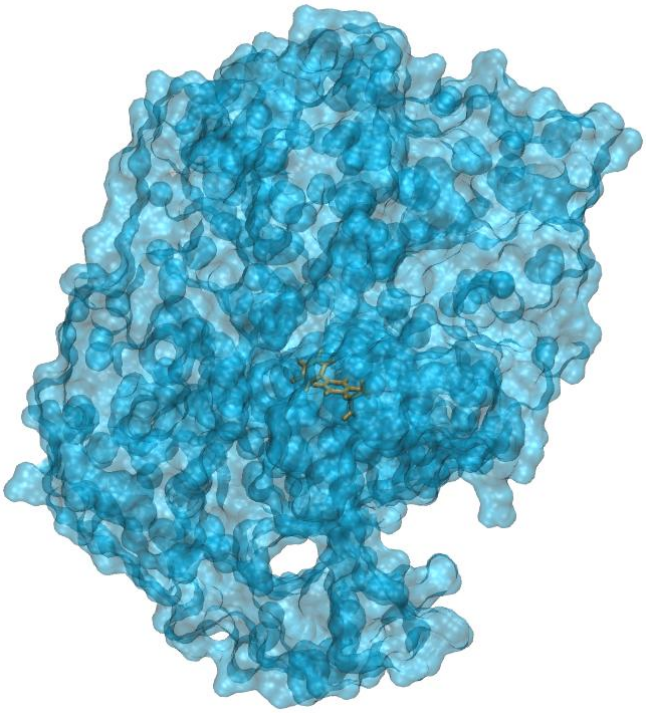
DEPARTAMENTO DE CIÊNCIAS DA VIDA

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Uma estratégia para a discriminação entre compostos activos e inactivos em experiências de rastreio virtual: COX-1 como caso de estudo



Pedro Miguel Serrano Germano Calado Carreiras

2012

DEPARTAMENTO DE CIÊNCIAS DA VIDA

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Uma estratégia para a discriminação entre compostos activos e inactivos em experiências de rastreio virtual: COX-1 como caso de estudo

Dissertação apresentada à Universidade de Coimbra para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Biologia, realizada sob a orientação científica do Professor Doutor Rui Manuel Pontes Meireles Ferreira de Brito (Universidade de Coimbra) e do Professor Doutor João Carlos Mano Castro Loureiro (Universidade de Coimbra)

Pedro Miguel Serrano Germano Calado Carreiras

2012

Agradecimentos

Em primeiro lugar queria agradecer ao Doutor Rui M. M. Brito pela disponibilidade em me aceitar sob a sua orientação como aluno de mestrado, permitindo-me conhecer uma nova área de investigação que tanto me estimulou e fez crescer em todos os aspectos do meu ser.

Queria também agradecer ao Dr. João Loureiro por ter-me aceitado sob a sua co-orientação tão em cima da hora e que se disponibilizou em ajudar no que fosse necessário.

Agradeço especialmente à Cândida Silva por todo o trabalho, ajuda, apoio, preocupação quer no trabalho desenvolvido quer no meu crescimento como aluno e como pessoa. Pelas horas disponibilizadas e as dores de cabeça que lhe proporcionei o meu profundo e sincero obrigado, sem dúvida nunca chegaria onde cheguei sem a tua ajuda. Este trabalho é também teu. E nunca poderei demonstrar a minha sincera gratidão pelo que fizeste por mim.

Queria também agradecer aos restantes membros do grupo do RMBLab, Carlos, Catarina, Daniela, Elsa, Pedro, Tiago e Zaida por me terem aceitado de braços aberto e contribuído para o meu trabalho dentro do grupo tanto directa como indirectamente demonstrando total disponibilidade para me ajudar e criticar quando assim foi necessário. Um sincero obrigado a todos pois sem vocês este trabalho não faria sentido.

Um agradecimento muito especial aos meus pais por me terem apoiado, compreendido (o que nem sempre foi fácil) e confiado em mim pois sem vocês nada na vida faria sentido e eu nunca teria tido oportunidade de chegar onde cheguei. Aos meus irmãos, tios, primos e família em geral por me aturarem e apoiarem incondicionalmente.

Um agradecimento muito especial à Filipa e à Tânia as duas pessoas que mais me apoiaram neste ano e ao longo do meu trabalho pois estiveram sempre lá para me ouvir e aconselhar quer fosse em momentos bons como naqueles menos bons. Vocês sabem o que significam.

Um agradecimento a todos os meus amigos e ao pessoal do KickGym pois sem vocês estes últimos anos de nada valeriam a pena e foram vocês que animaram toda a minha vida.

E a ti por tudo o que significas para mim e porque sem ti nada faria sentido.

Índice

Índice de Figuras

Índice de Tabelas

Lista de Abreviaturas

- AINES, Anti-inflamatórios não esteróides
- AUC, Área abaixo da curva ROC
- COX-1, Ciclooxigenase 1
- COX-2, Ciclooxigenase-2
- DUD, A Database of Useful Decoys
- EF, factores de enriquecimento
- RMSD, Raiz dos desvios médios quadráticos
- ROC, *Receiver Operating Characteristic*
- SVM, Máquinas de Vectores de Suporte
- VinaCluster, Conjunto de melhores poses seleccionadas com base nos resultados de uma análise de grupos
- VinaFP, Conjunto de melhores poses seleccionado com base na ordenação dada pela função de pontuação do AutoDock Vina

Lista de Traduções

Palavras no Inglês	Tradução Utilizada
➢ *3-fold cross validation*	Validação cruzada *3-fold*
➢ *Clustering*	Análise de grupos
➢ *Complete linkage*	Vizinho mais distante
➢ *Consensus scoring*	Pontuação consenso
➢ *Decoys*	Inactivos
➢ *Docking*	Acoplamento molecular
➢ *Force-Field Based*	Baseadas em campos de forças
➢ *Hierarchical clustering*	Análise de grupos hierárquica
➢ *Knowledge-based*	Baseadas em conhecimento
➢ *Root-mean-square-deviation*	Raiz dos desvios médios quadráticos
➢ *Support Vector Machines*	Máquinas de Vectores de Suporte
➢ *Virtual Screening*	Rastreio Virtual

Resumo

Um dos grandes desafios para a realização de experiências de rastreio virtual aplicando técnicas de acoplamento molecular está em encontrar ferramentas capazes de prever boas poses de compostos no local activo de uma proteína e de as pontuar correctamente, de uma forma rápida e com um baixo custo. Neste trabalho foram testadas diferentes estratégias para obter uma melhor discriminação entre compostos activos e inactivos em experiências de rastreio virtual baseadas em técnicas de acoplamento molecular utilizando a COX-1 (ciclooxigenase-1) como caso de estudo. A COX-1 foi escolhida como caso de estudo porque a sua actividade pode ser afectada por diferentes fármacos sem que estes tenham sido desenvolvidos para esse propósito, sendo por isso importante desenvolver estratégias para a identificação desses fármacos.

O acoplamento molecular dos compostos no pacote da DUD (*A Database of Useful Decoys*) para a COX-1 foi realizado com o objectivo de se validar a capacidade do programa AutoDock Vina de prever e pontuar resultados de acoplamento molecular utilizando a COX-1. Adicionalmente, os resultados do acoplamento molecular foram analisados para obter os valores dos parâmetros constituintes da função de pontuação do programa. Estes foram utilizados para gerar classificadores através da utilização do SVM-light, um programa que implementa um algoritmo de Máquinas de Vectores de Suporte (SVM). A avaliação do desempenho da função de pontuação do AutoDock Vina e dos classificadores obtidos com o SVM-light foi realizada para dois conjuntos de "melhores" poses seleccionadas com base: (i) na ordenação dada pela função de pontuação do AutoDock Vina, e (ii) nos resultados de análise de grupos (*clustering*), aplicando uma análise de curvas ROC, das áreas abaixo das curvas ROC (AUC) e de curvas de enriquecimento e factores de enriquecimento.

Os resultados obtidos mostram que a utilização de SVM para o desenvolvimento de classificadores a partir dos parâmetros constituintes da função de pontuação do AutoDock Vina apresenta melhorias significativas na discriminação de compostos activos e inactivos. Adicionalmente, os resultados demonstram que a utilização de novas estratégias como a utilização de uma análise de grupos para seleccionar as "melhores" poses pode melhorar significativamente os resultados do acoplamento molecular.

Abstract

One of the challenges to perform virtual screening when using docking is finding tools capable of predicting good poses of compounds in the active site of a protein and scoring them correctly, in a fast and cheap way. In this work, different strategies were tested to obtain a better discrimination between active and inactive compounds in virtual screening based on docking techniques using COX-1 (cyclooxygenase-1). COX-1 was chosen as a case study because its activity can be affected by different pharmaceutical drugs, even if these drugs have not been developed for that purpose, and therefore it is important to develop new strategies to identify such active compounds.

All active and inactive compounds for COX-1 were obtained from DUD (*A Database of Useful Decoys*) and docked to COX-1 with the aim of validating the ability of the AutoDock Vina program to predict and score the results. Additionally all the docking results were analyzed to obtain the values of Vilna's scoring function parameters. These parameters were then used to train classification models with SVM-light, a program that implements an algorithm of support vector machines (SVM). The performance of the AutoDock Vina scoring function and the classification models obtained from the SVM-light were evaluated on two sets of "best" poses selected based on (i) the order given by the AutoDock Vina scoring function, and (ii) the results of clustering analyze of the poses, and then applying analysis of ROC curves, area under the curve ROC (AUC), enrichment curves and enrichment factors.

The results show that the use of SVM to development models of classification using the constituent parameters of the AutoDock Vina Scoring function shows significant improvement in discrimination of active and inactive compounds. Also, the results show that the use of others strategies like the clustering analyze of the poses to select the "best" pose besides the one given from the scoring functions of the docking programs can significantly improve the results of the docking.

Capítulo 1 Introdução

1 – Objectivos

Um dos desafios para a realização de experiências de rastreio virtual aplicando técnicas de acoplamento molecular está em encontrar ferramentas capazes de prever boas poses de compostos no local activo de uma proteína e de as pontuar correctamente, de uma forma rápida e com um baixo custo.

Com a realização deste trabalho pretendeu-se explorar soluções para a resolução destes problemas. Primeiro, validar a capacidade do programa AutoDock Vina – um programa de livre acesso - de prever e pontuar resultados de acoplamento molecular utilizando a COX-1, uma enzima com importantes funções na síntese de prostaglandinas, como caso de estudo. Segundo, desenvolver e testar diferentes estratégias para a construção de modelos de classificação utilizando os parâmetros constituintes da função de pontuação do AutoDock Vina, que permitam uma discriminação efectiva entre poses de compostos activos e inactivos, após o acoplamento molecular realizado pelo AutoDock Vina.

2 – Rastreio Virtual

Rastreio virtual (no inglês, *Virtual Screening*) é um termo criado nos finais dos anos 90 e aplicado ao conjunto de métodos computacionais utilizados na triagem de grandes bibliotecas virtuais de compostos químicos (Walters *et al.*, 1998). O rastreio virtual é utilizado na descoberta de novos fármacos e tem como objectivo encontrar em bibliotecas de compostos químicos, que actualmente contêm informação de milhões de compostos, aqueles que melhor possam interagir com determinada molécula alvo (Lazarova, 2008). A utilização deste tipo de metodologias computacionais permite reduzir custos, desperdício de material e tempo despendido no estudo e análise dos compostos por técnicas experimentais (Delaglio, 2001). A triagem dos compostos pode ser feita de diversas formas, dependendo do método de rastreio virtual utilizado.

Embora várias das tecnologias relacionadas com o rastreio virtual só tenham surgido por volta de 1997 (Oprea *et al.*, 2004), estas têm sofrido uma evolução constante e todos os anos surgem novos desenvolvimentos e programas (comerciais e de código aberto). As técnicas de rastreio virtual dividem-se em duas categorias principais dependendo da abordagem utilizada, podendo ser baseadas no ligando ou no receptor, sendo neste caso também por vezes designadas por baseadas na estrutura (Jackson, 1995; Ripphausen *et al.*, 2010; Sousa *et al.*, 2010).

As técnicas de rastreio baseadas no ligando utilizam informação acerca de compostos que apresentam actividade com uma molécula alvo determinada experimentalmente e procuram em bases de dados compostos com características físico-químicas e/ou estruturais semelhantes (Barril *et al*., 2004; Pérez-Nueno *et al*,. 2008). Estas técnicas baseiam-se no pressuposto de que compostos com características semelhantes apresentam actividades semelhantes.

Por sua vez, as técnicas de rastreio virtual baseadas no receptor (ou estrutura) envolvem a utilização da estrutura tridimensional (3D) da molécula alvo (receptor), normalmente obtida por ressonância magnética nuclear ou por cristalografia de raios-X (Oprea *et al.*, 2004), e permitem estudar o local activo da molécula alvo e as interacções estabelecidas com compostos que apresentem afinidade para esse local. Assim, o objectivo é procurar nas bases de dados compostos que possam potencialmente apresentar afinidade para o local activo da molécula de interesse (Andricopulo *et al.*, 2009).

As duas metodologias acima descritas englobam várias técnicas. No grupo de técnicas de rastreio virtual baseadas no ligando encontram-se as técnicas de similaridade 2D (Duan *et al.,* 2010), de similaridade 3D (Jenkins *et al.,* 2004) e a técnica de modelos de farmacóforos baseados em ligandos (Yang, 2010; Sun, 2008). Entre as técnicas de rastreio virtual baseadas no receptor encontram-se a técnica de modelos de farmacóforo baseados no receptor (Yang, 2010; Sun, 2008) e a técnica de acoplamento molecular (no inglês, *Molecular Docking*) (Reddy *et al.,* 2007; Yuriev *et al.,* 2009).

De um modo geral, as técnicas de rastreio virtual baseadas no receptor envolvem quatro passos:

1. Identificação da molécula alvo (receptor) e do seu local activo;
2. Identificação de um conjunto de potenciais compostos, que liguem ao local activo da molécula alvo que sirvam como modelo;
3. Identificação das estruturas receptor-ligando que apresentem modos de ligação com valores de energia mais baixos;
4. Repetição dos passos 2 e 3 para obter as características que melhor determinam a interacção receptor-ligando com o intuito de rastrear bases de dados de compostos e obter os que apresentam características mais semelhantes (Lazarova, 2008).

3 - Acoplamento Molecular

3.1 – Fundamentos do Acoplamento Molecular

O acoplamento molecular é uma técnica computacional que procura prever a melhor conformação de um ligando e a sua orientação no local activo da molécula alvo em estudo (Kitchen *et al*., 2004; Yuriev *et al.,* 2009). A técnica de acoplamento molecular permite estudar vários tipos de interacções moleculares tais como proteína-ligando, proteína-proteína e de proteínas com outras biomoléculas como o DNA e RNA (Lengauer *et al.,* 1996). Esta técnica divide-se em dois passos principais. No primeiro passo, chamado de acoplamento ou posicionamento, o algoritmo tenta encontrar quais as melhores conformações e orientações do(s) ligando(s) no local activo da molécula alvo recorrendo para tal a uma busca conformacional extensa do(s) ligando(s) e eventualmente do local activo. A cada um dos modos de ligação obtidos neste primeiro

passo designa-se por pose. No segundo passo, o objectivo é seleccionar as melhores poses obtidas. Para tal, as poses são ordenadas numa relação de ordem-afinidade, com base numa função de pontuação que avalia a afinidade dos compostos para o local de ligação da molécula alvo (Coupez *et al.,* 2006; Onodera *et al,* 2007; Huang *et al.,* 2010). Dependendo da função de pontuação utilizada podem ser considerados apenas parâmetros do ligando tais como a conformação, orientação e hidrofobicidade, e/ou parâmetros relativos à interacção entre o ligando e o local activo como por exemplo ligações de hidrogénio e forças de Van der Waals (Stahl *et al.,* 2001).

Em suma, quando a técnica de acoplamento molecular é aplicada pretendem-se alcançar dois objectivos distintos: primeiro, prever a melhor orientação estrutural do ligando relativamente ao receptor, e segundo, obter uma correcta previsão e pontuação da afinidade de ligação (Kitchen*et al*., 2004).

3.2 – Vantagens e desvantagens da técnica de Acoplamento Molecular

Algumas das vantagens da técnica de acoplamento molecular relativamente as outras técnicas são (i) a capacidade de incorporar a flexibilidade dos ligandos no acoplamento; (ii) envolver processos físicos próximos do processo de ligação receptor-ligando, permitindo rastrear compostos de uma maneira menos tendenciosa; (iii) possibilitar o estudo de compostos para os quais não exista qualquer informação experimental; e (iv) fornecer previsões geométricas dos ligandos no local activo em estudo, permitindo optimizar compostos capazes de interagir com esse local activo sem criar modelos com base na expectativa das características que os ligandos possam ou não ter (Doman *et al.,* 2002; Pérez-Nueno *et al.,* 2007; Sousa *et al.,* 2010).

No entanto, a utilização da técnica de acoplamento molecular para rastreio virtual apresenta também algumas limitações associadas tanto ao passo de acoplamento como ao passo de ordenação das poses obtidas para cada composto. Relativamente ao passo de acoplamento destacamos duas limitações importantes. A primeira está associada à necessidade de existir uma estrutura 3D determinada experimentalmente, de preferência com boa qualidade, da molécula alvo o que nem sempre acontece. A segunda limitação está relacionada com a inclusão do factor de flexibilidade da estrutura da molécula alvo e dos ligandos nas simulações de acoplamento molecular. Hoje em dia, e com o avanço da computação, muitos programas já têm em consideração a flexibilidade da estrutura do ligando mas na sua esmagadora maioria continuam a

considerar a molécula alvo como uma estrutura rígida (Taylor *et al.*, 2002; Halperin *et al.*, 2002). O ideal seria considerar flexível tanto a estrutura da molécula alvo como a do(s) ligando(s), uma vez que é aceite que a interacção molécula alvo-ligando é dinâmica. No entanto, entrar em consideração com a flexibilidade da estrutura da molécula alvo aumenta muito o espaço de pesquisa, o que aumenta o tempo e os recursos computacionais necessários para a realização do rastreio virtual (Coupez*et al.*, 2006; Yuriev*et al.*, 2009).

Embora existam algumas dificuldades em obter previsões de boas poses de compostos, actualmente é nas funções de pontuação que se encontra a verdadeira limitação dos programas de acoplamento molecular (Stahl *et al.*, 2001; Lill *et al.*, 2011). Idealmente, as funções de pontuação deveriam permitir pontuar e discriminar com precisão as melhores poses de cada composto, ao mesmo tempo permitindo distinguir os compostos que verdadeiramente ligam dos que não ligam ao local activo da molécula alvo (Kitchen *et al.*, 2004; Coupez *et al.*, 2006; Jain *et al.*, 2006; Huang *et al.*, 2010). No entanto, e apesar dos requisitos que as funções de pontuação devem cumprir estarem bem definidos, ainda não foi possível definir uma função de pontuação que os satisfaça na totalidade. Na maioria dos casos, as funções de pontuação cumprem apenas alguns desses requisitos. Uma das razões para que tal ocorra, está associado ao facto das funções de pontuação assumirem que a afinidade da ligação entre um composto e uma molécula alvo pode ser descrita como a soma de um número limitado de termos independentes, quando na realidade a ligação é influenciada por muitos termos, que por vezes não são independentes mas estão correlacionados entre si. Adicionalmente, a maioria das funções de pontuação não entra em conta com os efeitos entrópicos das interacções, uma vez que considera as estruturas das moléculas alvo como sendo rígidas, não tendo em conta as restantes poses, e ignorando também efeitos específicos de solvatação e desolvatação (Schulz-Gasch *et al.*, 2004).

Definir uma função de pontuação que considere todos os aspectos que caracterizam as ligações receptor-ligando seria computacionalmente dispendioso tanto em recursos computacionais como em tempo, o que tornaria a realização do rastreio virtual inviável. Na busca de um compromisso, as funções de pontuação apresentadas na literatura têm apenas em conta algumas características, aumentando assim a sua rapidez em detrimento da sua precisão (Sousa *et al.*, 2010). Dependendo do tipo de características incluídas nas funções de pontuação, estas são designadas por baseadas em campos de forças (no inglês, *Force-Field Based*), e empíricas ou baseadas em

conhecimento (no inglês, *knowledge-based*) (Kitchen *et al*., 2004; Jain 2006; Huang *et al.*, 2010).

Como os métodos para a definição de funções de pontuação são variados e as funções existentes apresentam lacunas, a combinação de várias funções de pontuação é uma das estratégias utilizada para pontuar poses dos compostos obtidas por acoplamento molecular. Uma das finalidades desta estratégia é compensar os erros que cada uma das funções apresenta e obter resultados mais precisos obtendo-se assim a chamada pontuação consenso (no inglês, *consensus scoring*) (Kitchen *et al*., 2004). Contudo o potencial deste tipo de pontuação consenso é limitado pela eventual utilização de funções com termos parecidos, o que poderá aumentar o erro em vez de o diminuir (Coupez*et al.*, 2006).

Recentemente, e tendo em consideração todos estes factos, têm-se procurado formas diferentes de melhorar as funções de pontuação ou estratégias alternativas que permitam uma boa discriminação entre compostos activos e inactivos. É neste contexto que surgem soluções como a utilização de Máquinas de Vectores de Suporte (no inglês, *Support Vector Machine* - SVM), métodos lineares, análise de grupos (*clustering*), métodos estatísticos de aprendizagem Bayesiana, redes neuronais e árvores de decisão (Plewczynski *et al.*, 2005; Melville *et al.*, 2009).

4 – Máquinas de Vectores de Suporte

As Máquinas de Vectores de Suporte (SVM; Cortes *et al.*, 1995) são uma das técnicas de *machine learning* que tem sido aplicada para a resolução de problemas de reconhecimento de padrões, classificação e regressão (Burges *et al.*, 1998). No contexto da Biologia e da Química Computacional, tem sido utilizada para resolver problemas de reconhecimento de padrões como por exemplo o reconhecimento e análise de genes, a detecção de homologia de proteínas, identificação de péptidos através da análise de dados de espectrometria de massa, identificação e previsão de interacções proteína-proteína, detecção da permeabilidade da barreira hemato-encefálica por fármacos, entre outros (Noble, 1998; Plewczynski *et al.*, 2005). No caso particular do rastreio virtual, tem sido aplicada para classificação e identificação de compostos activos e inactivos (Deng *et al.*, 2004; Ballester *et al.*, 2010).

As SVM são uma técnica de aprendizagem supervisionada em que os algoritmos computacionais têm a capacidade de generalizar um modelo com base num conjunto de exemplos. Quando perante um problema de classificação, dado um conjunto de exemplos e a sua classificação (X_i, Y_i) esta técnica produz um classificador capaz de prever qual a classe a que pertencem novos dados. Este processo é conhecido por treino. O classificador final pode também ser visto como uma função f que recebendo como argumento um novo dado x fornece uma previsão y (Figura 1).

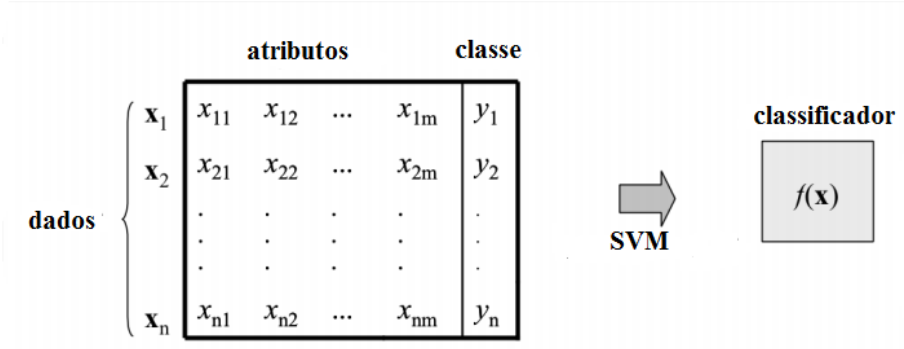


Figura 1 – Modelo de construção de um classificador. (Imagem adaptada de Lorena *et al.*, 2007).

Segundo a aprendizagem estatística, um conjunto de treino de padrões será linearmente separável se existir pelo menos um classificador linear definido pelo par (w,b) que classifique correctamente todos os padrões de treino. Este classificador é representado pelo hiperplano H ($f(x)=w.x+b=0$) e define uma região para a classe +1 ($w.x+b>0$) e outra para a classe -1 ($w.x+b<0$). Idealmente, o hiperplano será equidistante das 2 classes (Jorissen *et al.*, 2005; Noble, 2006; Hasegawa *et al.*, 2010).

Existem várias possibilidades de classificadores lineares que podem separar os dados. Contudo, existe apenas um que maximiza as margens, ou seja, a distância entre o hiperplano e o ponto mais próximo de cada classe (Figura 2). Este hiperplano é designado por hiperplano óptimo de separação. Assim, esta fronteira é tida como a que apresenta melhor capacidade para classificar correctamente novos exemplos.

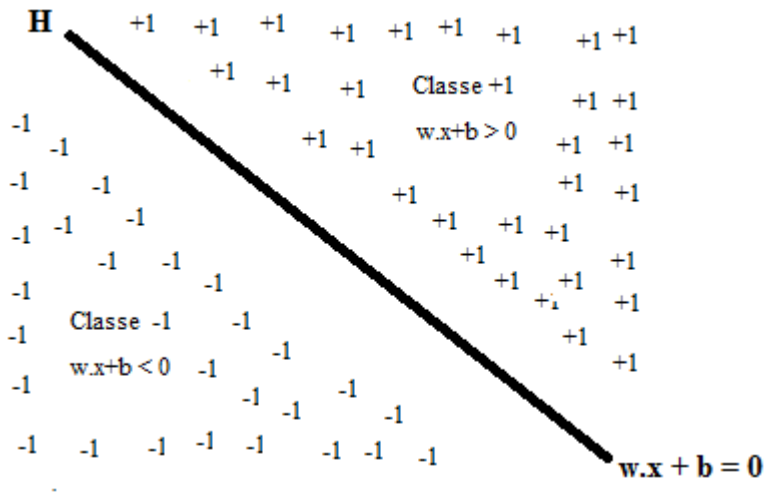


Figura 2 – Hiperplano de separação das classes -1 e +1.

Após o treino, o classificador será capaz de prever a classe a que pertencem novos exemplos, diferentes dos utilizados no treino. A classe de um exemplo x_k será determinada pela seguinte equação:

$$classe\ (x_k) = \begin{cases} +1\ se\ w.x_k + b > 0 \\ -1\ se\ w.x_k + b < 0 \end{cases}$$

Assim, a classificação dos novos dados irá depender apenas do sinal da expressão *w.x+b*.

As SVM apresentam baixa sensibilidade ao *overfitting* (sobre-ajustamento) dos dados mesmo com a utilização de um grande conjunto de dados redundantes, uma vez que são baseadas no princípio de minimização do risco estrutural para diminuir erros gerais e de treino (Han *et al.*, 2007). Contudo, as SVM tendem a necessitar de uma grande quantidade de dados de treino para a construção de um classificador (Yap *et al.*, 2005).

5 – Ciclooxigenases: um caso de estudo

A ciclooxigenase é a principal enzima envolvida no processo de síntese de prostanóides, tais como as prostaglandinas e tromboxanos, a partir de ácido araquidónico, e actua ao nível do sistema imunitário como resposta a um processo inflamatório (Dannhardt *et al.*, 2001). A ciclooxigenase-1 (COX-1) e a ciclooxigenase-2

(COX-2) são as duas isoformas que se conhecem desta enzima, sendo a COX-1 uma isoforma constitutiva e a COX-2 uma isoforma induzida predominantemente por uma resposta inflamatória. Estas enzimas apresentam uma identidade de mais de 60% da sequência de aminoácidos (Garavito *et al.,* 2002; Gupta *et al.,* 2004; Carvalho *et al.,* 2004).

O estudo das ciclooxigenases ganhou particular relevância quando se descobriu que são o alvo de fármacos anti-inflamatórios não esteróides (AINES), e que a sua inibição apresenta promissoras melhorias na prevenção da doença de Alzheimer e do cancro do colo-rectal (Vane *et al.,* 1998; Dannhardt, 2001; Garavito *et al.,* 2002; Gupta *et al.,* 2004).

A COX-1, caso de estudo no trabalho aqui apresentado, está presente em quase todas as células do corpo humano em condições fisiológicas, principalmente nos vasos sanguíneos, rins, estômago e plaquetas, encontrando-se envolvida em diversas funções vitais na resposta à inflamação, no sistema cardiovascular, na tumorigénese, no sistema gastrointestinal e no funcionamento renal (Morita, 2002). A inibição da actividade da COX-1 pode causar por isso graves problemas no organismo (Kummer *et al.,* 2002).

Actualmente conhecem-se diversos fármacos capazes de inibir a actividade da COX-1, mesmo quando não tenham sido desenvolvidos com esse objectivo. A aspirina é um desses casos (Dannhardt, 2001; Garavito *et al.,* 2002). Dadas as consequências nefastas associadas à inibição da COX-1, torna-se necessário conhecer os compostos que inibem ou alteram o funcionamento da COX-1. É neste contexto que a técnica de acoplamento molecular e a proposta de estratégias que permitam uma discriminação efectiva entre compostos activos e inactivos podem desempenhar um papel fundamental.

5.1 – COX-1: Caracterização estrutural e funcional.

A COX-1 é uma proteína homodimérica que se encontra predominantemente associada à membrana do retículo endoplasmático (Figura 3). A primeira estrutura tridimensional da COX-1 foi obtida por cristalografia de raios-X e descrita em 1994 pelo investigador Picot e seus colaboradores (Picot *et al.,* 1994). Actualmente, existem 231 estruturas cristalográficas da COX-1 no *Protein Data Bank* (PDB; Berman *et al.,* 2000), algumas das quais ligadas a inibidores.

Cada um dos monómeros da COX-1 é constituído por 576 aminoácidos organizados em três domínios estruturais (Figura 4), um domínio na região N-terminal semelhante ao factor de crescimento epidérmico (EGF), um domínio de ligação à membrana, e um domínio catalítico na região C-terminal que contém os dois locais activos desta enzima - o local com actividade de ciclooxigenase e o local com actividade de peroxidase (Smith *et al.,* 2000; Dannhardt, 2001; Garavito *et al.*, 2002; Gupta *et al.,* 2004).

Neste trabalho, focaremos a nossa atenção no local activo com actividade de ciclooxigenase pois é neste local activo que actuam inibidores da actividade da COX-1 como os AINES (Smith *et al.,* 2000; Dannhardt, 2001; Carvalho *et al.,* 2004). O local activo é formado por um longo canal hidrofóbico com cerca de 25 Å, confinado por várias hélices-α, que vai desde o domínio de ligação à membrana até ao centro do domínio catalítico. Este canal pode ser dividido em duas regiões distintas a região onde os AINES se ligam, e que engloba a metade superior do canal estendendo-se do aminoácido Arg-120 até perto da Tyr-385; e a região inicial formada pela metade inferior do canal e que forma uma "boca" no domínio de ligação à membrana permitindo a entrada directa do ácido araquidônico e de O_2 pela zona apolar da bicamada lipídica (Dannhardt, 2001; Garavito *et al.,* 2002; Carvalho *et al.,* 2004). A inibição deste local activo por fármacos como a Aspirina, o Ibuprofeno, o Flurbiprofeno ou outros AINES depende de uma eficiente ligação dos mesmos à Arg-120 e que a substituição ou supressão deste aminoácido reduz ou impede a inibição por este tipo de fármacos, no caso da Aspirina a acetilação ocorre mais especificamente na Ser-530 neste caso a Aspirina compete de uma forma rápida e reversível com o ácido araquidónico pelo local de ligação de ciclooxigenase promovendo uma modificação covalente (acetilação) da Ser-530. (Garavito *et al.,* 1999; Garavito *et al.,* 2002).

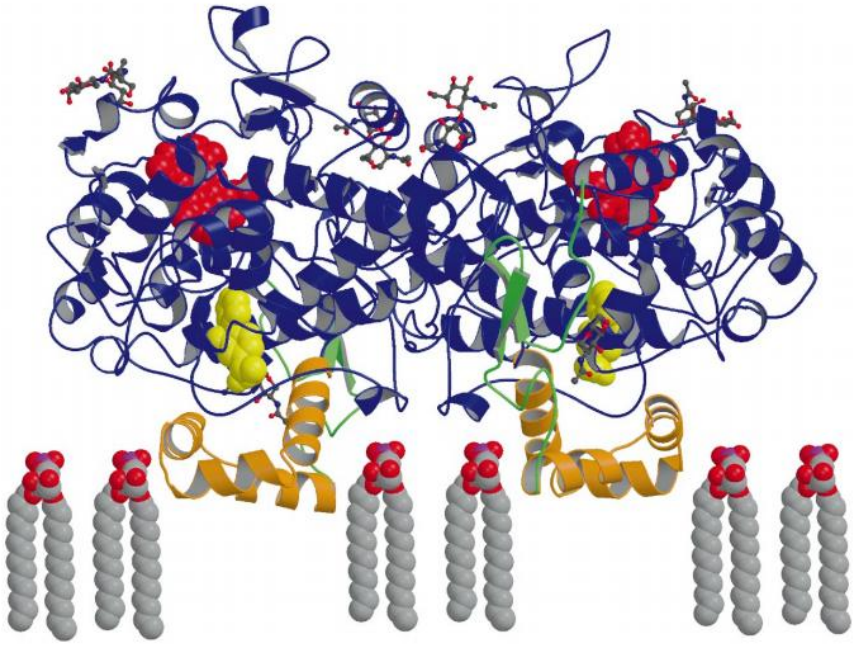


Figura 3 – Representação da estrutura da COX-1 de *Ovis aries.* A figura ilustra a posição do grupo heme (em esferas a vermelho), o local de ligação com o flurbiprofeno (amarelo) e a relação da proteína com a membrana do retículo endoplasmático. O domínio EGF, de ligação à membrana, e os domínios catalíticos estão representados coloridos a verde, laranja e azul respectivamente. (Imagem adaptada de Garavito e DeWitt.,1999)

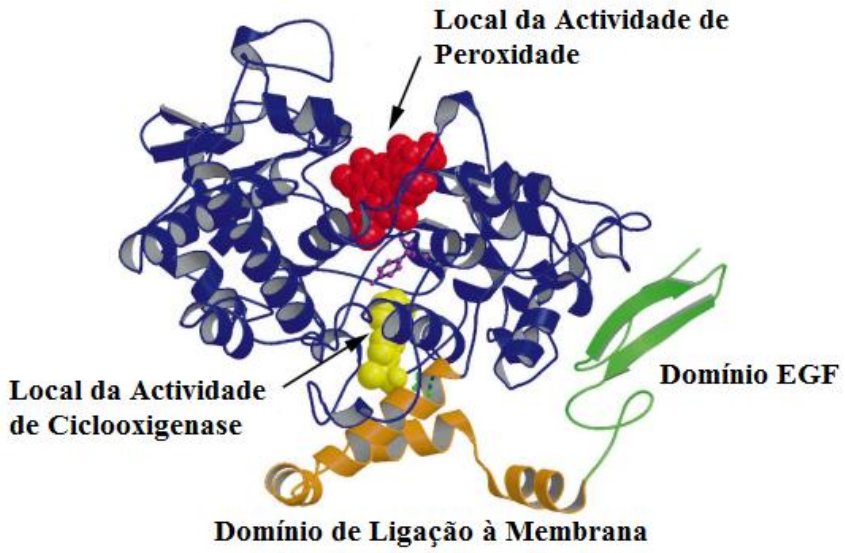


Figura 4 – Representação da estrutura dos domínios constituintes do monómero da COX-1 de *Ovis aries.* (Imagem adaptada de Garavito e DeWitt., 1999).

Capítulo 2 Ferramentas Computacionais

1 – AutoDock Vina

O programa AutoDock Vina (Trott *et al.*, 2010) foi utilizado para realizar o acoplamento molecular da COX-1 com os compostos activos e inactivos (no inglês *decoys*) presentes no pacote para a COX-1 disponível na DUD – "*A Database of Useful Decoys"* (Huang *et al.*, 2006). O intuito é validar a capacidade do programa de prever as poses e pontuar correctamente esses compostos e obter os valores dos parâmetros que constituem a função de pontuação usada pelo AutoDock Vina para o cálculo da afinidade de cada ligando.

O AutoDock Vina é um programa de código livre para a realização de acoplamento molecular e rastreio virtual desenvolvido por Oleg Trott do "*Molecular Graphics Lab"* no *"The Scripps Research Institute"*, La Jolla, EUA. A sua rapidez, precisão e livre acesso fazem do AutoDock Vina um programa rápido e fácil de usar.

A função de pontuação do AutoDock Vina tem em conta as seguintes contribuições: interacções estéricas (representadas pelos termos gauss 1, gauss 2, e repulsão), hidrofobicidade, ligações de hidrogénio e número de ângulos de torção de rotação livre dos ligandos (termo *Nrot*). Os valores dos pesos associados por defeito a cada um dos termos estão descritos na Tabela 1. O algoritmo de optimização global implementado no AutoDock Vina é o método de *Iterated Local Search* (Trott *et al.*, 2010).

Tabela I – Termos e respectivos pesos por defeito usados na função de pontuação do AutoDock Vina (Trott et. al., 2010)

Termos	Pesos
gauss 1	-0.0356
gauss 2	-0.00516
repulsão	0.840
hidrofobicidade	-0.0351
ligações de hidrogénio	-0.587
Nrot	0.0585

2 – MGLTools/AutoDockTools

Desenvolvido pelo "*Molecular Graphics Lab"* do *"The Scripps Research Institute"*, o MGLTools engloba um conjunto de programas e recursos computacionais necessários para a realização de acoplamento molecular e de rastreio virtual com o AutoDock Vina. Entre os vários programas disponibilizados encontra-se a ferramenta AutoDockTools (Michel *et al.* 1999).

O AutoDockTools é uma interface gráfica que permite executar, analisar e preparar o acoplamento molecular a realizar no AutoDock Vina. Este programa inclui as seguintes funcionalidades: visualização de ficheiros de proteínas e compostos, definição do tamanho da "caixa" de acoplamento que especifica o espaço cartesiano de procura em que o programa irá realizar o acoplamento, selecção dos ângulos torcionais de rotação livre dos compostos e adição ou remoção dos átomos de hidrogénio.

3 – SVM-light

O SVM-light (Joachims, 1999) é uma implementação de um algoritmo de Máquinas de Vectores de Suporte em C, que permite resolver problemas de regressão e classificação com aprendizagem e que é capaz de lidar com uma grande quantidade de dados. O SVM-light é um programa de livre acesso disponível no sítio da internet http://www.svmlight.joachims.org e desenvolvido por Thorsten Joachims do Departamento de Ciências da Computação da Universidade de Cornell.

Este programa disponibiliza dois módulos, um de aprendizagem (*svm_learn*) e outro de classificação (*svm_classify*). É necessário executar primeiro o módulo de aprendizagem com o conjunto de dados de treino para se produzirem os modelos de classificação. Após terem sido gerados, os modelos de classificação serão posteriormente utilizados para a execução do módulo de classificação para classificar os compostos nos conjuntos de dados de teste.

Neste trabalho, o SVM-light foi utilizado para treinar modelos de classificação a partir de conjuntos de treino constituídos por compostos presentes no pacote para a COX-1 disponível na DUD. A cada composto nos conjuntos de treino é atribuída a classe a que pertence (+1 se activo; -1 se inactivo) e o vector com os termos individuais dos parâmetros constituintes da função de pontuação do AutoDock Vina (Tabela 1). O objectivo é desenvolver um classificador que permita uma discriminação efectiva entre

compostos activos e inactivos da COX-1 baseado em dados originados do acoplamento molecular realizado pelo AutoDock Vina.

Capítulo 3 Protocolo Experimental

1 – Descrição do conjunto de dados

DUD – *A Directory of Useful Decoys* (Huang *et al.,* 2006) – é uma base de dados pública que reúne uma colecção de compostos activos e inactivos para diferentes alvos moleculares e que permite a realização de análises comparativas de programas de rastreio virtual (Huang *et al.,* 2006). Esta base de dados é constituída por 2950 ligandos activos para 40 proteínas alvo diferentes. A cada composto activo estão associados 36 compostos inactivos (*decoys*), perfazendo um total de 95316 compostos inactivos. Cada um dos 36 compostos inactivos assemelha-se a um composto activo em algumas das suas propriedades físico-químicas, como por exemplo peso molecular ou número de ligações de hidrogénio, apresentando no entanto propriedades topológicas diferentes.

DUD inclui uma biblioteca de compostos para a COX-1 constituída por 25 compostos activos e 911 compostos inactivos. Dos 911 compostos inactivos, 62 compostos foram excluídos do conjunto de compostos analisados uma vez que representam estruturas repetidas (um composto inactivo pode estar associado a vários compostos activos), obtendo-se um número final de compostos inactivos de 849. Para além da biblioteca de compostos activos e inactivos, também faz parte do pacote da COX-1 a estrutura de um complexo da proteína com o ligando ácido 2-(1,1'-bifenil-4-il) propanóico (BFL). Esta estrutura foi obtida por cristalografia de raios-x com uma resolução de 2 Å, encontrando-se disponível no *Protein Data Bank* (PDB) com o código 1Q4G (Gupta *et al.*, 2004).

2 – Análise de homologia da sequência de aminoácidos da COX-1

No PDB não se encontra disponível uma estrutura cristalográfica de COX-1 de *Homo sapiens* (Humana). Sendo as únicas estruturas disponíveis pertencentes às espécies *Ovis aries* (Ovelha) e *Mus musculus* (Ratinho). Dado que a estrutura disponível no pacote da DUD é de *Ovis aries* e que uma elevada similaridade da sequência de aminoácidos de proteínas pode implicar uma significativa semelhança estrutural, o que permitirá extrapolar os resultados obtidos de uma espécie para outra, procedeu-se a uma análise de homologia entre a sequência de aminoácidos da COX-1 de *Ovis aries* (código UniProt P05979) e a sequência de aminoácidos da COX-1 de *Homo sapiens* (código UniProt P23219).

A análise foi realizada recorrendo ao programa "Clustal O" disponível em http://www.uniprot.org (Consortium, 2012) e utilizando todos os valores por defeito dos parâmetros do programa.

3 – Acoplamento Molecular com AutoDock Vina

O protocolo para realizar simulações de acoplamento molecular com o AutoDock Vina envolve passos como a preparação do receptor e a definição da "caixa" onde o programa irá realizar as simulações e que deverá compreender o local activo de interesse.

Para uma melhor visualização do local activo, a estrutura da COX-1 foi orientada de forma a que o eixo principal coincidisse com o eixo dos ZZ. Em seguida, adicionaram-se os átomos de hidrogénio. A estrutura cristalográfica do complexo da COX-1 com o ligando BFL (PDB 1Q4G, cadeia B) incluída no pacote da DUD foi utilizada como modelo para a definição dos parâmetros da "caixa" seleccionada: centro nas coordenadas (x,y,z) = (26.6, 33.8, 201.5), com as dimensões (em Ångström) 18 × 18 × 20 (Figura 5). A "caixa" seleccionada foi testada através da realização de uma simulação de acoplamento molecular da COX-1 com o ligando BFL utilizando o AutoDock Vina. Para os restantes parâmetros de simulação foram utilizados os valores por defeito do programa. Em seguida, a validação da "caixa" foi realizada visualizando as poses geradas para o ligando BFL com o programa AutoDock Vina. Adicionalmente, o programa "Fconv" (Stahura *et al.*, 2004; Neudert *et al.*, 2011) foi utilizado para calcular a raiz dos desvios médios quadráticos (RMSD no inglês, root-mean-square deviation; Eq. 1) entre as poses geradas e a pose do ligando na estrutura cristalográfica.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N_{átomos}} (X_{1,i} - X_{2,i})^2 + (Y_{1,i} - Y_{2,i})^2 + (Z_{1,i} - Z_{2,i})^2}{N_{átomos}}} \quad \text{(Eq. 1)}$$

Uma vez preparado o receptor e validada a "caixa" de simulação, foi realizado o acoplamento molecular da COX-1 com os 25 compostos activos e os 849 compostos inactivos fornecidos no pacote da DUD, impondo-se a geração de um número máximo de 15 poses por composto (parâmetro "num_modes") e fazendo variar o valor que determina a máxima diferença de energia entre a melhor e a pior pose classificadas de 3 para 10 (parâmetro "energy_range", em kcal/mol). As poses obtidas para todos os compostos foram validadas por inspecção visual com o programa AutoDockTool.

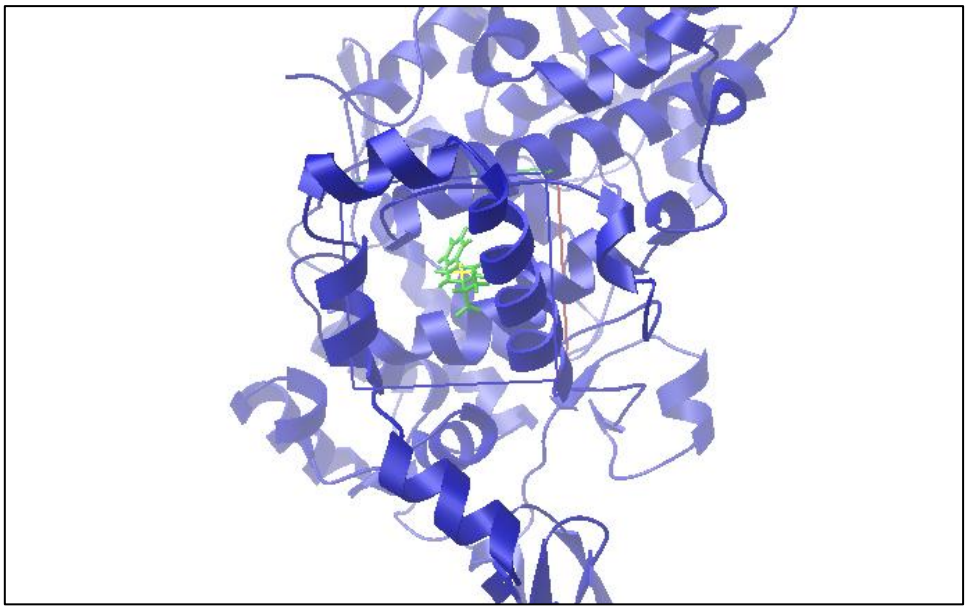


Figura 5 – "Caixa" seleccionada em torno do local activo da COX-1. Representação da "caixa" seleccionada para realizar o acoplamento com a COX-1 (a azul) e o ligando BFL no centro (a verde) criada recorrendo ao AutoDockTools

4 – Construção de modelos de classificação de compostos activos e inactivos da COX-1

Após terem sido obtidos os resultados do acoplamento molecular entre a COX-1 e todos os compostos seleccionados da DUD, procedeu-se à construção dos modelos de classificação que permitissem uma melhor discriminação entre compostos activos e inactivos. Pretendeu-se com este passo obter um melhor classificador das poses obtidas pelo acoplamento molecular. A construção dos modelos de classificação foi feita com base nos parâmetros da função de pontuação do AutoDock Vina.

4.1 – Selecção das melhores poses para cada composto

Foram seleccionadas duas "melhores" poses para cada um dos compostos com base em dois critérios diferentes. Uma primeira pose foi obtida com base na ordenação das poses dada pela função de pontuação do AutoDock Vina, a segunda pose foi obtida com base nos resultados de uma análise de grupos (clustering) aplicada às várias poses de cada composto (Cheng *et al.* 2009, Huang *et al.* 2010). Para este segundo critério, foi utilizado o método de análise de grupos hierárquico (no inglês, hierarchical clustering) com o critério do vizinho mais distante (no inglês, complete linkage) utilizando o programa "Fconv" (Stahura *et al.*, 2004; Neudert *et al.*, 2011), que permite agrupar as poses dos compostos utilizando o RMSD (Eq. 1) como medida de similaridade (Wang *et al.*, 2003, Bouvier *et al.* 2010). Para cada composto foi seleccionada a pose com valor mais negativo de afinidade no grupo (cluster) com maior número de poses.

Para cada uma das melhores poses seleccionadas dos 874 compostos foram obtidos os valores de cada um dos 5 termos da função de pontuação, utilizando a opção "score_only" do AutoDock Vina. Em seguida, os valores dos termos para cada pose foram normalizados recorrendo à Eq. 2 (Graf *et al.*, 2001).

$$\tilde{\chi} = \frac{\chi}{\|\chi\|} \in \mathrm{R}^{\mathrm{N}} \qquad \text{(Eq. 2)}$$

Na Eq. 2, χ é o vector com os valores dos parâmetros da função de pontuação do AutoDock Vina, $\|\chi\|$ é a norma do vector e $\tilde{\chi}$ representa o vector normalizado.

4.2 – Construção dos conjuntos de dados de treino e de teste

O rácio entre compostos activos e inactivos nos pacotes da DUD é de 1:36, o que faz com que o número de exemplos negativos seja muito maior do que o número de exemplos positivos. Para testar qual a proporção entre compostos activos e inactivos a incluir nos conjuntos de treino que produz um melhor modelo de classificação, os conjuntos de treino e teste foram construídos seguindo os seguintes passos (Figura 6):

1. Divisão dos compostos activos e inactivos em dois conjuntos diferentes de dados;

2. Divisão aleatória do conjunto de compostos activos (Figura 6, A1, A2 e A3) e do conjunto dos compostos inactivos (Figura 6, I1, I2 e I3) em três partes, com aproximadamente o mesmo número de compostos;
3. Cada um dos três conjuntos de compostos inactivos foi subdividido aleatoriamente em "n" subconjuntos, com o valor "n" a variar entre 5 e 36 (Figura 6, caso particular de n=5).
4. Em seguida e para cada iteração do método de validação cruzada *3-fold* foram produzidos "n" subconjuntos de treino contendo compostos activos e inactivos tal como é apresentado na Figura 6. Por exemplo, quando o conjunto de teste é formado pelos compostos activos em A3 e pelos compostos inactivos I3, os "n" subconjuntos de treino contêm todos os compostos activos A1 e A2, um dos "n*"* subconjuntos de I1 e um dos "n" subconjuntos de I2.

A técnica de validação cruzada *3-fold* (no inglês, *3-fold cross validation*), foi utilizada para avaliar a capacidade de generalização dos modelos a partir do conjunto de dados fornecidos testando a precisão dos modelos criados.

4.3 – Treino e teste dos modelos de classificação com o SVM-light

O treino de classificadores com o SVM-light envolve a utilização do módulo *svm_learn* sobre os conjuntos de treino produzidos anteriormente. Após a execução deste módulo, são criados "n" modelos de classificação para cada iteração do método de validação cruzada. Em seguida, utiliza-se o módulo *svm_classify* para classificar os compostos dos conjuntos de teste com os modelos correspondentes.

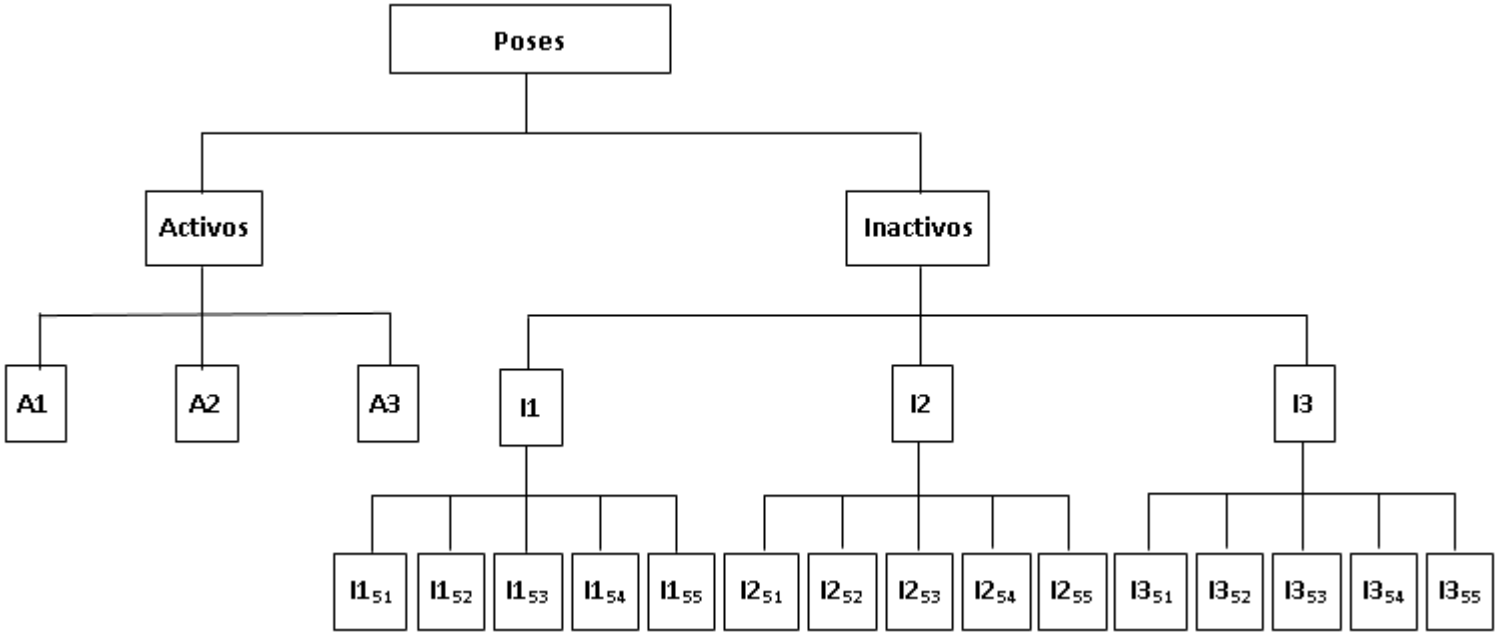
4.4 – Avaliação do desempenho dos classificadores

Em problemas de classificação utilizando SVMs, as classes de compostos activos e inactivos são separadas pelo valor 0, correspondendo um valor positivo a compostos activos e um valor negativo a compostos inactivos. Como tal, no fim da classificação obtêm-se "n" previsões, cada uma com um valor de classificação, para

Passo 1:

Passo 2:

Passo 3:



Passo 4:

Conjunto de Teste: TS3= A3 ∪ I3 com n=5

Conjuntos de Treino:

TrS_{51} = A1 ∪ A2 ∪ $I1_{51}$ ∪ $I2_{51}$

TrS_{52} = A1 ∪ A2 ∪ $I1_{52}$ ∪$I2_{52}$

TrS_{53} = A1 ∪ A2 ∪ $I1_{53}$ ∪ $I2_{53}$

TrS_{54} = A1 ∪ A2 ∪ $I1_{54}$ ∪ $I2_{54}$

TrS_{55} = A1 ∪ A2 ∪ $I1_{55}$ ∪ $I2_{55}$

Figura 6 – Esquema de construção dos conjuntos de treino e de teste. Exemplo para uma iteração do método de validação cruzada para o valor de n=5,. correspondendo $I1_{5(1\text{-}5)}$ às cinco subdivisões do conjunto I1. Analogamente para I2 e I3

cada um dos compostos no conjunto de teste. Para cada composto nos conjuntos de teste foi obtida uma previsão consenso somando os valores de classificação obtidos dos vários modelos de classificação: quando o valor final obtido é positivo, o composto é considerado activo e quando o valor final é negativo, o composto é considerado inactivo (Huang *et al.,* 2010; Kinnings *et al.*, 2011).

O valor do *F-score* (Eq. 3) foi calculado para verificar a precisão da previsão consenso de cada uma das 3 iterações, e o valor médio de *F-score* foi atribuído a cada valor de "n".

$$F - score = \frac{(2 \times precisão \times sensibilidade)}{(precisão + sensibilidade)} \quad \text{(Eq. 3)}$$

Na equação anterior, precisão representa o número de resultados positivos correctos dividido pelo total de previsões positivas e a sensibilidade representa o número de resultados positivos correctos dividido pelo número total de resultados que deviam ter sido previstos como positivos (Cannon *et al.,* 2007; Kinnings *et al.,* 2011).

Os valores mais elevados de *F-scores* indicam qual o valor de "n" que dá a proporção de compostos activos e compostos inactivos a incluir nos conjuntos de treino que produzem o melhor modelo de classificação para compostos que interajam com a COX-1.

5 – Métodos de avaliação da função de pontuação e dos classificadores

Actualmente, não existe uma metodologia padrão para a análise, avaliação e comparação estatística de resultados gerados pelas técnicas de rastreio virtual e que permita a partilha de novos resultados de forma fácil e concisa. Como tal os trabalhos apresentados por diferentes grupos de investigação nem sempre reportam as mesmas métricas para avaliar os métodos utilizados e/ou desenvolvidos, o que dificulta a comparação entre os resultados obtidos nos diferentes trabalhos (Triballeau *et al.*, 2005; Jain *et al.,* 2008).

Com base na revisão da literatura (Truchon *et al.,* 2007; Nicholls, 2008) foram escolhidas as seguintes métricas para avaliar o desempenho da função de pontuação do AutoDock Vina e dos classificadores: área abaixo da curva ROC (AUC), curvas de factores de enriquecimento e os factores de enriquecimento correspondentes a 1%, 5% e 10%, sensibilidade, especificidade, precisão e *F-score*.

Sensibilidade, especificidade, precisão e *F-score* são medidas estatísticas para avaliar o desempenho de classificadores binários. A sensibilidade (também designada de *Recall;* Eq. 4) mede a proporção de compostos activos que foram correctamente classificados. Por seu lado, especificidade (Eq. 5) mede a proporção de compostos inactivos que foram correctamente classificados. A precisão (Eq. 6) quantifica a proporção de compostos activos classificados correctamente entre todos os compostos classificados como activos. A *F-score* (Eq. 3) combina a medida de precisão com sensibilidade (recall).

$$Sensibilidade = \frac{T_P}{T_P + F_N} \text{ (Eq. 4)}$$

$$Especificidade = \frac{T_N}{T_N + F_P} \text{ (Eq. 5)}$$

$$Precisão = \frac{T_P}{T_P + F_P} \text{ (Eq. 6)}$$

Nas equações 4 a 6, T_P representa o número de compostos activos classificados correctamente, T_N representa o número de compostos inactivos classificados correctamente, F_P representa o número de compostos activos classificados incorrectamente e F_N representa o número de compostos inactivos classificados incorrectamente.

Actualmente, a melhor forma de comparar o desempenho de funções de pontuação e classificadores binários é através da análise das curvas ROC (no inglês, *Receiver Operating Characteristic*) e o reconhecimento precoce de compostos activos através da análise das curvas dos factores de enriquecimento (no inglês, *Enrichment Factor*) (Truchon *et al.,* 2007; Jain *et al.,* 2008; Nicholls, 2008).

As curvas ROC são uma representação gráfica da sensibilidade (proporção de verdadeiros positivos) em função de 1-especificidade (proporção de falsos positivos). O valor da área abaixo da curva ROC (AUC) fornece uma medida objectiva do desempenho global de um classificador. Um valor de AUC igual a 1 (ou 100%) indica uma discriminação perfeita entre compostos activos e inactivos, enquanto um valor de

0,5 (ou 50%) é interpretado como um desempenho aleatório. Em termos práticos, para estudos de rastreio virtual que apresentam um desempenho melhor que o aleatório obtêm-se valores de AUC entre 0,5 e 1, enquanto valores de AUC inferiores a 0,5 são obtidos para métodos que tendem a dar melhor pontuação a compostos inactivos do que a compostos activos. Genericamente, a exactidão do método de classificação pode ser avaliado com a seguinte escala: 0,9 ≤ AUC ≤ 1 é excelente; 0,8 ≤ AUC <0,9 é bom; 0,7 ≤ AUC <0,8 é razoável; 0,7 ≤ AUC <0,5 é mau; e AUC <0,5 corresponde a uma falha completa.

Um dos problemas apontados ao valor de AUC é o facto de esta ser uma medida muito global não apresentando qualquer informação sobre o reconhecimento precoce de compostos activos (Triballeau *et al.*, 2005; Jain *et al.*, 2008; Hamza *et al.*, 2012). Por outro lado, os factores de enriquecimento (EF) quantificam o rácio de compostos activos identificados no topo X% do conjunto total de compostos ordenados (Eq. 7):

$$EF^{x\%} = \frac{Compostos\ Activos^{x\%}_{Seleccionados} / Compostos^{x\%}_{Seleccionados}}{Compostos\ Activos^{x\%}_{Total} / Compostos^{x\%}_{Total}} \quad \text{(Eq. 7)}$$

Onde $Compostos\ Activos^{x\%}_{Seleccionados}$ é o número de compostos activos no topo X% do conjunto de compostos, $Compostos^{x\%}_{Seleccionados}$ é o total de compostos no topo X%, $Compostos\ Activos^{x\%}_{Total}$ é o número de compostos activos no conjunto total de compostos e $Compostos^{x\%}_{Total}$ é o número total de compostos analisados.

A análise das AUC (das curvas ROC) e dos factores de enriquecimento deve ser feita de forma complementar. Os valores de AUC fornecem uma medida do desempenho global da função de pontuação ou do classificador, enquanto os factores de enriquecimento indicam a eficácia com que os compostos activos estão a ser reconhecidos e identificados. (Truchon *et al.*, 2007; Nicholls, 2008; Hamza *et al.*, 2012).

Estas análises foram feitas recorrendo aos pacotes *enrichvs* (Yabuuchi, 2011) e *pROC* (Robin *et al.*, 2001) disponível no programa R (R Development Core Team, 2009). O pacote *pROC* permitiu obter as curvas ROC bem como os valores de AUC, e o o pacote *enrichvs* permitiu obter as curvas dos factores de enriquecimento e os factores de enriquecimento correspondentes a 1%, 5% e 10% de cada uma das curvas.

Capítulo 4 Resultados e Discussão

Neste capítulo são apresentados e discutidos os resultados obtidos. O capítulo começa com a descrição e análise dos resultados de acoplamento molecular obtidos com o programa AutoDock Vina. Em seguida, são avaliados e discutidos os diferentes métodos utilizados para a classificação de compostos activos e inactivos para a COX-1.

1 – Análise de homologia da sequência da COX-1

A análise de homologia entre a sequência de aminoácidos da proteína COX-1 de *Ovis aries* e de *Homo sapiens* foi realizada para averiguar a similaridade entre as duas sequências, em particular na região do local activo com função de ciclooxigenase. Esta análise é importante porque não sendo ainda conhecida a estrutura da proteína COX-1 humana e sabendo que uma elevada similaridade da sequência de aminoácidos de proteínas pode implicar uma significativa semelhança estrutural, os resultados aqui descritos para a COX-1 de ovelha podem ser extrapolados de uma espécie para a outra.

Na Figura 7 mostra-se o alinhamento das sequências lineares da COX-1 das espécies *Ovis aries* e *Homo sapiens*. Os resíduos do local activo estão assinalados a negrito e sublinhados. Observa-se que as sequências de aminoácidos da COX-1 nas duas espécies apresentam uma identidade de 100% nos resíduos constituintes do local activo, e conforme descrito na literatura a homologia da COX-1 entre espécies diferentes é de cerca de 85-90% (Smith *et al.*, 2000; Dannhardt, 2001; Carvalho *et al.*, 2004). Com base nos resultados obtidos da análise de homologia da sequência de aminoácidos das duas espécies, espera-se que os resultados obtidos no presente trabalho com a estrutura de *Ovis* aries, possam ser extrapolados e utilizados com a COX-1 humana.

2 – Acoplamento Molecular

O programa AutoDock Vina foi utilizado para realizar as simulações de acoplamento molecular dos compostos activos e inactivos do pacote da DUD para a COX-1. Os primeiros passos envolveram a preparação da proteína e dos compostos para o acoplamento molecular, bem como a definição dos parâmetros das simulações com base na estrutura cristalográfica do complexo da COX-1 com o ligando BFL.

```
Ovis aries   MSRQSISLRFPLLLLLLSPSPVFSADPGAPAPVNPCCYYPCQHQGICVRFGLDRYQCDCT 60
Homo sapiens -MSRSLLLWFLLFLLLLPPLPVLLADPGAPTPVNPCCYYPCQHQGICVRFGLDRYQCDCT 59
                :*: * * *:**** * **: ******:****************************

Ovis aries   RTGYSGPNCTIPEIWTWLRTTLRPSPSFIHFMLTHGRWLWDFVNATFIRDTLMRLVLTVR 120
Homo sapiens RTGYSGPNCTIPGLWTWLRNSLRPSPSFTHFLLTHGRWFWEFVNATFIREMLMRLVLTVR 119
             ************ :*****.:******* **:******:*:********: *********

Ovis aries   SNLIPSPPTYNIAHDYISWESFSNVSYYTRILPSVPRDCPTPMGTKGKKQLPDAEFLSRR 180
Homo sapiens SNLIPSPPTYNSAHDYISWESFSNVSYYTRILPSVPKDCPTPMGTKGKKQLPDAQLLARR 179
             *********** **********************:****************::*:**

Ovis aries   FLLRRKFIPDPQGTNLMFAFFAQHFTHQFFKTSGKMGPGFTKALGHGVDLGHIYGDNLER 240
Homo sapiens FLLRRKFIPDPQGTNLMFAFFAQHFTHQFFKTSGKMGPGFTKALGHGVDLGHIYGDNLER 239
             ************************************************************

Ovis aries   QYQLRLFKDGKLKYQMLNGEVYPPSVEEAPVLMHYPRGIPPQSQMAVGQEVFGLLPGLML 300
Homo sapiens QYQLRLFKDGKLKYQVLDGEMYPPSVEEAPVLMHYPRGIPPQSQMAVGQEVFGLLPGLML 299
             ***************:*:**:**************************************

Ovis aries   YATIWLREHNRVCDLLKAEHPTWGDEQLFQTARLILIGETIKIVIEEYVQQLSGYFLQLK 360
Homo sapiens YATLWLREHNRVCDLLKAEHPTWGDEQLFQTTRLILIGETIKIVIEEYVQQLSGYFLQLK 359
             ***:**************************:**************************

Ovis aries   FDPELLFGAQFQYRNRIAMEFNQLYHWHPLMPDSFRVGPQDYSYEQFLFNTSMLVDYGVE 420
Homo sapiens FDPELLFGVQFQYRNRIAMEFNHLYHWHPLMPDSFKVGSQEYSYEQFLFNTSMLVDYGVE 419
             ********.************:************:** *:*******************
```

Figura 7 - Alinhamento das sequências de aminoácidos de COX-1 de *Ovis aries* e *Homo sapiens*. Os resíduos do local activo responsável pela actividade de ciclooxigenase em estudo estão assinalados a negrito e sublinhados.

O acoplamento molecular do ligando BFL com a COX-1 (Figura 8) foi realizado para verificar a capacidade do AutoDock Vina prever boas poses para ligandos da COX-1, e para validar parâmetros de simulação, tais como a dimensão da "caixa" que define o espaço de pesquisa para futuras experiências de acoplamento molecular. Na Figura 8 (A) está representada a estrutura do complexo da COX-1 com o ligando BFL (PDB 1Q4G, cadeia B). Com base na análise deste complexo, o local activo com função

de ciclooxigenase é constituído por 19 resíduos, dos quais os resíduos Arg-120, Ser-353 e Ser-530 são os únicos resíduos polares. A "caixa" que delimita este local tem centro no ponto (26.6, 33.8, 201.5) e dimensões 18 × 18 × 20 Å (Figura 8, B). Do acoplamento molecular do ligando BFL com a COX-1 utilizando o programa AutoDock Vina foram geradas 8 poses. Tal como se pode observar na Figura 9 a melhor pose obtida para o ligando BFL pelo AutoDock Vina (a verde) é semelhante à pose do ligando obtida por cristalografia (a laranja).

Após a análise visual das poses obtidas para o ligando BFL, calculou-se o RMSD entre todas as poses obtidas pelo AutoDock Vina para o ligando BFL e a pose do ligando na estrutura cristalográfica do complexo formado com a COX-1, o que permitiu avaliar se o programa foi capaz de prever boas poses. A Figura 10 mostra os valores de RMSD das oito poses obtidas relativamente à estrutura cristalográfica e as respectivas afinidades calculadas pela função de pontuação do programa. Observa-se que os valores de RMSD variam aproximadamente entre 1,9 e 6,4 Å e os valores de afinidade variam entre -9,8 e -6,8 kcal/mol. A melhor pose (Figura 9, a verde) apresenta uma afinidade de -9.8 kcal/mol e um valor RMSD de 1,89 Å. As duas poses classificadas em segundo e terceiro lugares pela função de pontuação do AutoDock Vina apresentam valores de afinidades de -9,8 e -8.9 kcal/mol e valores de RMSD de 1,88 e 6,4 Å respectivamente. Os dados obtidos permitem concluir que os parâmetros definidos para as simulações de acoplamento molecular com o AutoDock Vina são válidos e podem ser usados nos acoplamentos moleculares da COX-1.

De seguida realizou-se o acoplamento molecular para todos os compostos do pacote da DUD com a "caixa" seleccionada, impondo-se que fossem geradas no máximo 15 poses por composto (parâmetro "num_modes") e fazendo variar o valor que determina a máxima diferença de energia entre a melhor e a pior pose classificadas de 3 para 10 (parâmetro "energy_range", em kcal/mol). Foi necessário alterar o parâmetro "num_modes" porque se pretendia obter mais poses por cada composto, necessárias para a análise de grupos. Isto obrigou também à alteração do parâmetro "energy_range" uma vez que o número de poses geradas é também dependente desse valor. Mesmo com um valor de diferença máxima de energia entre a melhor e a pior pose de 10 kcal/mol, para 244 compostos de um total de 874 não foram geradas 15 poses. No entanto, optou-se por não se aumentar mais o limiar de energia pois isso originava poses com valores de afinidade muito positivos. Utilizando o programa AutoDockTools para visualizar a

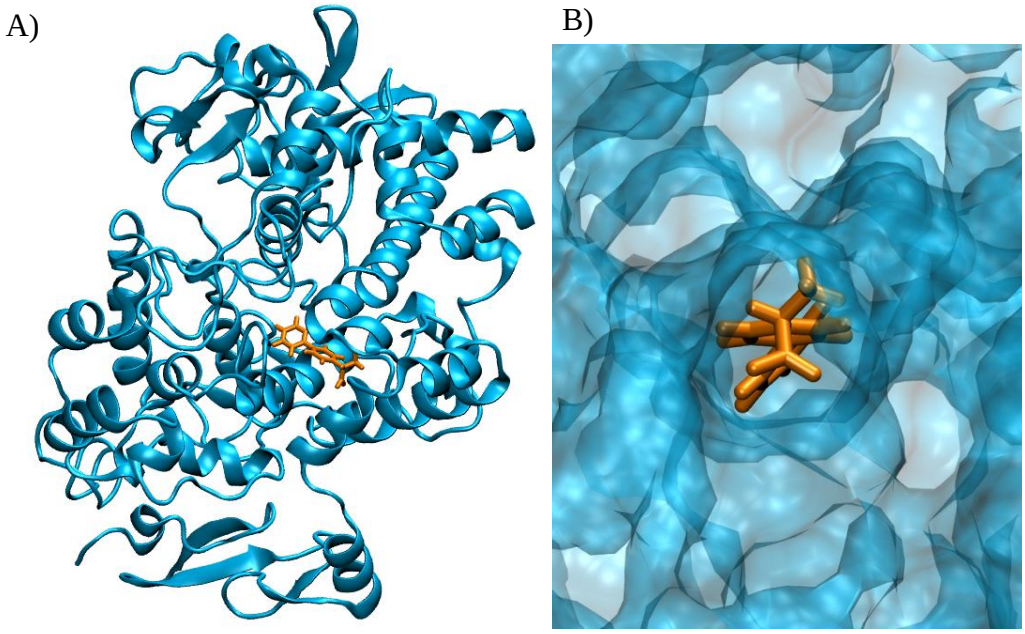


Figura 8 – Complexo formado pela COX-1 e o ligando ácido 2-(1,1'-bifenil-4-il) propanóico (BFL, PDB 1Q4G). A) Representação da estrutura da COX-1 obtida por cristalografia de raios-X com o ligando BFL no local activo em estudo. B) Representação da estrutura da COX-1 obtida por cristalografia de raios-X representando a superfície molecular com o ligando BFL no local activo em estudo.

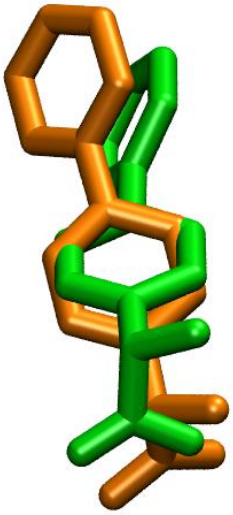


Figura 9 – Sobreposição de duas poses do ligando BFL. Representação da estrutura cristalográfica do ligando BFL (a laranja) e a estrutura da pose com melhor valor de afinidade (a verde) no acoplamento molecular com a COX-1 utilizando o programa AutoDock Vina.

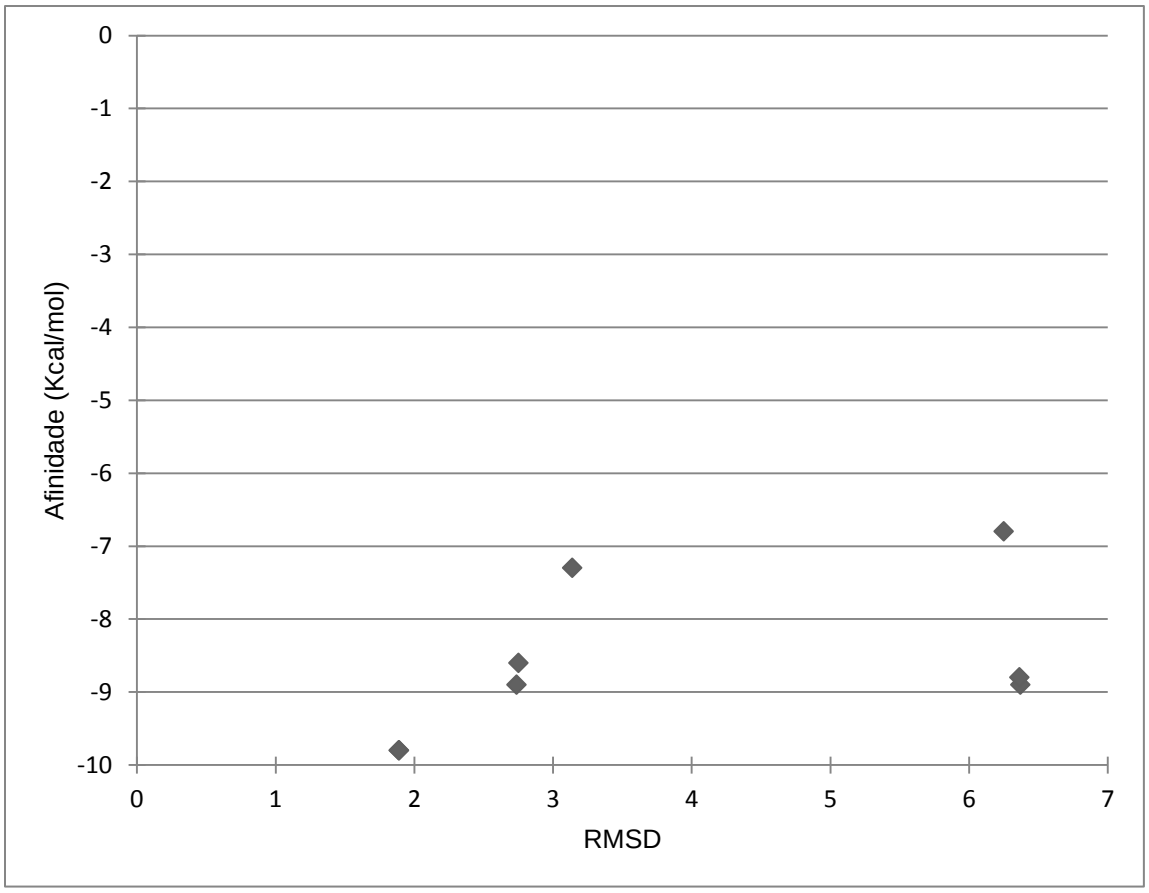


Figura 10 – Resultados do acoplamento molecular do ligando BFL. Distribuição dos valores das afinidades das poses obtidas para o ligando BFL em função do RMSD calculado entre cada pose e a estrutura cristalográfica. No gráfico encontram-se dois pontos sobrepostos na região de RMSD=1,9Å para duas poses com valores das afinidades (-9,8 kcal/mol) e RMSD muito semelhantes (1,89 e 1,88 Å).

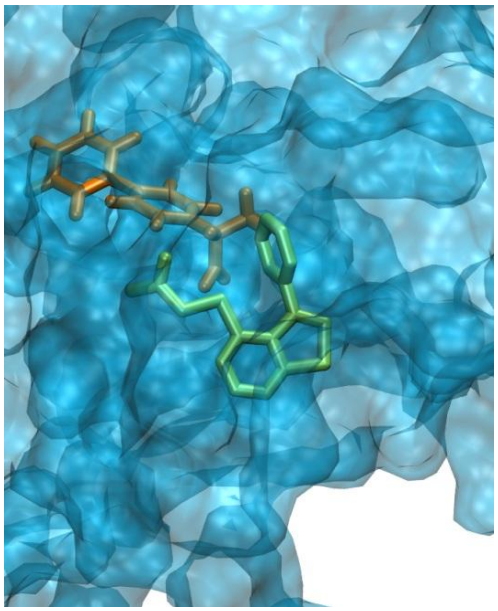


Figura 11 - Representação de uma pose fora do local activo da COX-1. Representação da estrutura da COX-1 obtida por cristalografia de raios-X representando a superfície molecular com o ligando BFL (a laranja) no local activo em estudo e a estrutura demonstrativa de uma pose gerada pelo AutoDock Vina que se encontra fora do local activo (a verde).

melhor pose gerada para cada um dos compostos foi possível constatar que para 35 compostos esta não se situava no centro do local activo mas antes ligeiramente deslocada do centro. Um exemplo de uma destas poses encontra-se representado na Figura 11. Como se pode observar a pose fora do local activo (a verde) apresenta uma estrutura com 3 anéis que não é tão pequena como o ligando BFL (a laranja) o que impossibilita a entrada no local activo. Estas características parecem "impossibilitar" o programa de "colocar" o composto dentro do local activo com o valor de "energy_range" definido. As melhores poses seleccionadas pelo AutoDock Vina para os 35 compostos que surgem fora do local activo, apresentam estruturas semelhantes à pose do exemplo da Figura 11 (a verde) não se encontrando dentro do local activo e sendo grandes de mais para lá "caber".

3 – Avaliação da função de pontuação do AutoDock Vina

O desempenho da função de pontuação do AutoDock Vina foi avaliado para dois conjuntos de "melhores" poses recorrendo aos valores de AUC e das curvas ROC, e à análise de curvas e factores de enriquecimento.

3.1 – Selecção das melhores poses para cada composto

A escolha das "melhores" poses para cada um dos compostos foi feita com base em dois critérios diferentes. Uma primeira "melhor" pose foi obtida com base na ordenação das poses dada pela função de pontuação do AutoDock Vina, a segunda pose foi obtida com base nos resultados de uma análise de grupos (clustering) aplicada às várias poses de cada composto (Cheng *et al.,* 2009, Huang *et al.,* 2010). Com a selecção de uma segunda "melhor" pose para cada composto, pretendia-se substituir poses que se encontravam "fora" do local activo em estudo, mas a que a função de pontuação do AutoDock Vina atribuiu um valor de afinidade melhor. Um exemplo de uma destas poses encontra-se representado na Figura 11.

Para a análise de grupos utilizou-se o programa "Fconv" (Stahura *et al.,* 2004; Neudert *et al.,* 2011) que implementa o método de análise de grupos hierárquico com o critério do vizinho mais distante. Este programa permitiu assim agrupar as poses dos compostos utilizando o RMSD como medida de similaridade (Wang *et al.,* 2003, Bouvier *et al.,* 2010). Muito resumidamente, o que este método faz, é adicionar uma

pose a um grupo (cluster) quando o valor de RMSD entre essa pose e qualquer elemento desse grupo for inferior a 2 Å, maximizando a distância entre grupos. Para cada composto foi seleccionada a pose com melhor afinidade no grupo (cluster) com o maior número de poses.

Das poses seleccionadas pela análise de grupos, 617 poses (19 activos + 598 inactivos) são diferentes da melhor pose escolhida pela função de pontuação do AutoDock Vina. Para além disso, a aplicação do método de análise de grupos reduziu o número de poses "fora" do local activo de 35 para 32.

Na Figura 12 são apresentados dois gráficos com a distribuição dos valores de afinidades correspondendo às duas melhores poses seleccionadas pelos métodos em cima descritos. A Figura 12 (A) apresenta as afinidades das melhores poses seleccionadas pela função de pontuação do AutoDock Vina. Pode observar-se que as afinidades dos compostos activos variam entre -9,8 e -3,7 kcal/mol enquanto os valores de afinidade dos compostos inactivos variam entre -10 e 0,7 kcal/mol. Os compostos inactivos apresentam uma maior variação dos valores de afinidades, bem como vários valores atípicos e extremos. Na Figura 12 (B) são apresentados os valores de afinidades das poses seleccionadas pelo método de análise de grupos, para as quais se observa que a afinidade dos compostos activos varia entre -9,4 e 0,4 kcal/mol, enquanto os valores de afinidade dos compostos inactivos variam entre -10 e 3,5 kcal/mol. Neste caso quer os compostos activos quer os compostos inactivos apresentam uma distribuição semelhante dos valores de afinidades e contêm valores atípicos. Mais uma vez, as poses dos compostos inactivos apresentam maior número de extremos. É também possível observar que nas poses seleccionadas pela análise de grupos passa a existir uma maior dispersão nos valores de afinidade, em particular para os compostos inactivos.

Ao longo do capítulo, o conjunto de poses seleccionadas apenas com base na função de pontuação do AutoDock Vina será designado por conjunto VinaFP, e o conjunto de poses seleccionadas pela análise de grupos será designado por conjunto VinaCluster.

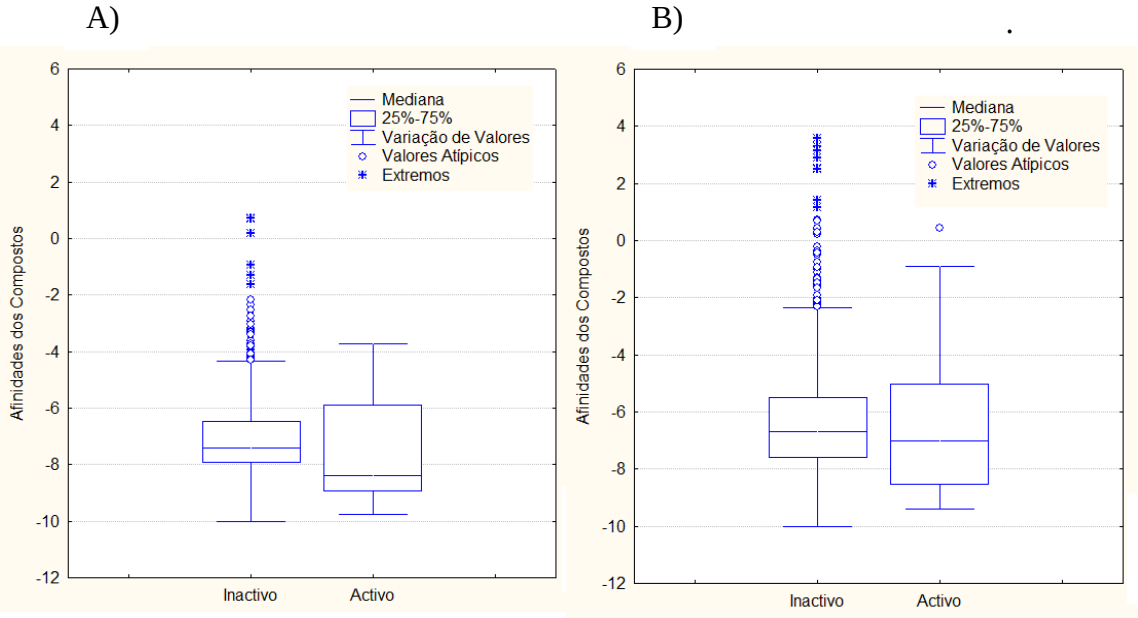


Figura 12 – Boxplots dos valores de afinidades das duas melhores poses seleccionadas. A) Boxplot dos valores de afinidade das poses no conjunto VinaFP B) Boxplot dos valores de afinidade das poses no conjunto VinaCluster.

3.2 – Análise do desempenho da função de pontuação com base nos valores de área abaixo da curva ROC (AUC) e de factores de enriquecimento

A avaliação do desempenho da função de pontuação do AutoDock Vina foi realizada recorrendo a análise das curvas ROC e da área abaixo da curva (AUC) correspondente. De seguida para avaliar a eficácia da função de pontuação, isto é para avaliar se os compostos activos são "rapidamente" colocados no topo da lista de ordenação, recorreu-se à análise das curvas de enriquecimento e dos factores de enriquecimento correspondentes a 1%, 5% e 10% da lista de ordenação dos compostos.

As curvas ROC permitem uma comparação directa de classificadores e a AUC é uma medida global para avaliar o desempenho dos classificadores (Triballeau *et al.*, 2005). Neste caso, os valores de AUC permitem avaliar para qual dos dois métodos de selecção de poses, a função de pontuação do AutoDock Vina dá uma melhor discriminação entre compostos activos e inactivos. As curvas ROC produzidas com valores de afinidades obtidos pela função de pontuação do AutoDock Vina para as poses seleccionadas pelos dois métodos descritos anteriormente são apresentadas na Figura 13. A curva ROC para os valores de afinidade das melhores poses no conjunto VinaFP (a azul) apresenta uma AUC de 63,44% (Tabela 2) enquanto a curva ROC para

os valores de afinidade das poses no conjunto VinaCluster (a vermelho) tem um valor de AUC de 55,06% (Tabela 2).

Os resultados obtidos não são surpreendentes uma vez que os valores de afinidade para as poses seleccionadas pelos dois métodos são calculados pela função de pontuação do AutoDock Vina, e as poses no conjunto VinaCluster terão um valor de afinidade sempre menor ou igual que a melhor pose seleccionada por defeito. No entanto, convém relembrar que a aplicação do método de análise de grupos é genericamente realizada com o objectivo de encontrar as poses dos compostos geradas computacionalmente mais parecidas com a *pose nativa*. Assim, e embora a curva ROC dos valores de afinidade das poses seleccionadas por este método apresente uma AUC menor, em termos práticos melhores poses podem estar de facto a ser seleccionadas.

Após a análise do desempenho da função de pontuação utilizando dois métodos de selecção de poses, procedeu-se à análise das curvas e factores de enriquecimento para avaliar se algum dos métodos de selecção torna a função mais eficaz no reconhecimento precoce de compostos activos. Estas curvas quantificam o rácio de compostos activos identificados no topo X% do conjunto total de compostos ordenados. Aqui são apresentados e discutidos os factores de enriquecimento correspondentes ao topo 1%, 5% e 10% de todo o conjunto de compostos testados.

No caso das curvas de enriquecimento é possível observar que a curva para as poses no conjunto VinaCluster (Figura 14, a vermelho) é ligeiramente melhor do que a obtida para as poses no conjunto VinaFP (Figura 14, a azul). Tal pode ser confirmado pelos valores de factores de enriquecimento apresentados na Tabela 2. Ao analisar-se os factores de enriquecimento correspondentes à lista de ordenação dos compostos, observa-se que para o topo 1% da base de dados nenhum dos métodos de selecção revela se a função de pontuação foi eficaz no reconhecimento precoce de compostos activos. No caso dos resultados para os topos 5% e 10% pode observar-se que a função de pontuação obteve melhores factores de enriquecimento no conjunto VinaCluster (3,2 e 2 respectivamente) do que no conjunto VinaFP (0,8 e 1,6).

Pela observação dos resultados apresentados, a função de pontuação do AutoDock Vina apresenta globalmente uma melhor capacidade de discriminar compostos activos de inactivos para o conjunto VinaFP. No entanto, quando avaliado o seu desempenho para o conjunto VinaCluster, o reconhecimento precoce de compostos activos é superior, em particular se referente ao topo 5%. Embora para ambos os métodos de selecção de poses, a função de pontuação apresente um desempenho melhor

que o aleatório (AUC=50%), nenhum se destaca nem consegue reconhecer compostos activos no topo 1%.

Para tentar melhorar estes resultados, recorreu-se a SVMs para construir um modelo de classificação utilizando os parâmetros envolvidos na função de pontuação do AutoDock Vina, esperando que discrimine melhor os compostos activos dos inactivos e que seja mais eficiente a colocá-los no topo da base de dados.

Tabela II – Métricas de avaliação dos valores de afinidade da função de pontuação do AutoDock Vina. Medidas de AUC e dos factores de enriquecimento para o topo 1%, 5% e 10% da base de dados para os valores de afinidade das melhores poses do conjunto VinaFP e do conjunto VinaCluster.

Medidas	**Métodos de Selecção**	
	VinaFP	VinaCluster
AUC %	63,44	55,06
EF a 1%	0	0
EF a 5%	0,8	3,2
EF a 10%	1,6	2

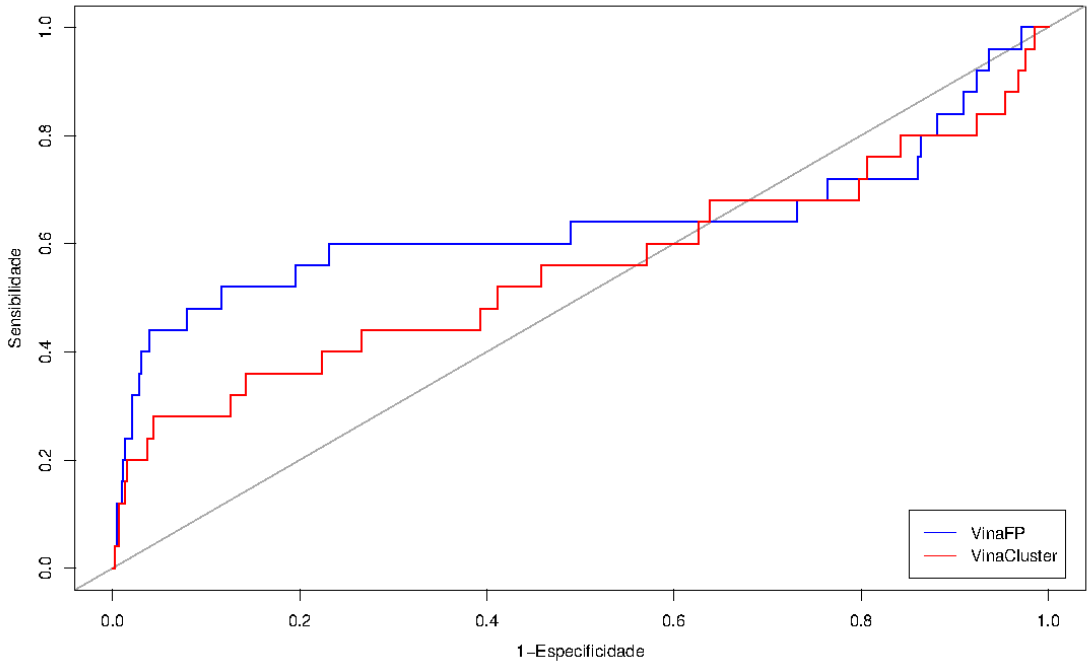


Figura 13 – Curvas ROC para os valores de afinidade. A linha na diagonal representa uma performance aleatória. A azul está representada a curva ROC para os valores de afinidade das poses do conjunto VinaFP. A vermelho está representada a curva de enriquecimento para os valores de afinidade das poses no conjunto VinaCluster.

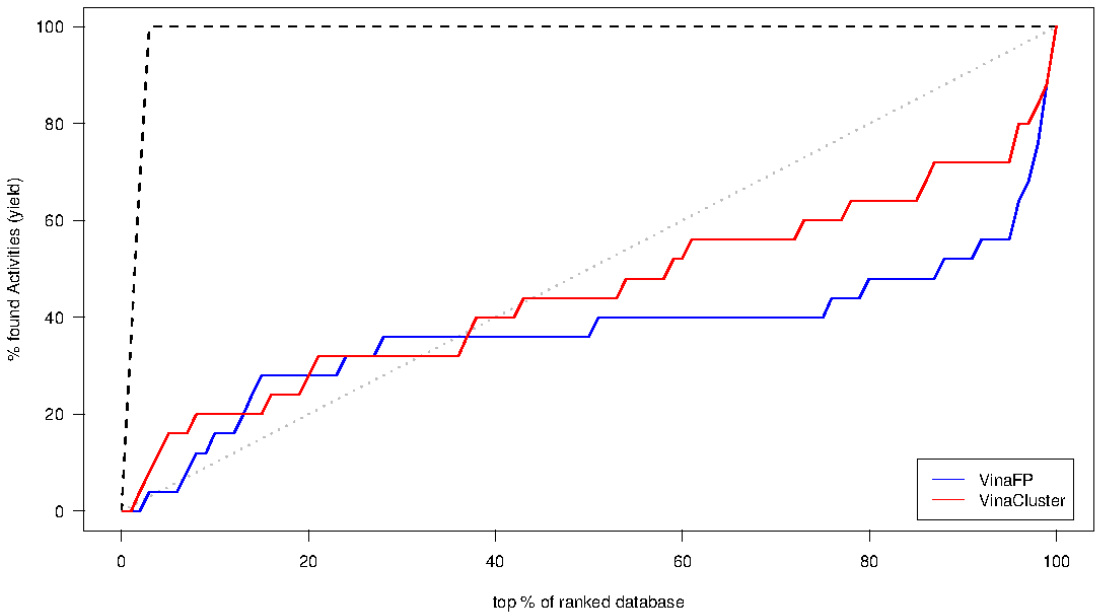


Figura 14 – Curvas de Enriquecimento para os valores de afinidade. A linha na diagonal representa uma performance aleatória e a linha a tracejado no topo superior da imagem representa o caso ideal. As curvas apresentam a percentagem de compostos activos identificados na percentagem X da base de dados ordenados. A azul está representada a curva de enriquecimento para os valores de afinidade das melhores poses do conjunto VinaFP. A vermelho está representada a curva de enriquecimento para os valores de afinidade das melhores poses no conjunto VinaCluster.

4 – Avaliação do desempenho dos classificadores obtidos com o SVM-light

A construção dos modelos de classificação obtidos pelo SVM-light foi feita com base nos parâmetros da função de pontuação do AutoDock Vina para as poses nos conjuntos VinaFP e VinaCluster. Uma vez que no conjunto total de compostos existem muito mais compostos inactivos do que activos, a primeira tarefa consistiu na escolha da proporção de compostos activos e compostos inactivos a incluir nos conjuntos de treino que produza o melhor modelo de classificação.

4.1 – Selecção da melhor divisão de compostos a incluir nos conjuntos de treino

Para seleccionar qual a proporção óptima de compostos activos e inactivos a incluir nos conjuntos de treino, os conjuntos VinaFP e VinaCluster foram divididos em três conjuntos aleatórios de compostos activos e inactivos. Em seguida, cada um dos conjuntos de compostos inactivos foi aleatoriamente dividido em "n" subconjuntos ("n" a variar entre 5 e 36). Os conjuntos finais de treino e teste foram obtidos tal como descrito no Capítulo 3 e esquematizado na Figura 6. A validação cruzada *3-fold* foi utilizada para testar a capacidade de generalização dos modelos gerados para cada valor de "n".

Uma vez que "n" modelos prevêem uma classificação para cada composto num determinado conjunto de teste, uma previsão consenso foi calculada a partir dos valores de classificação obtidos dos vários modelos de classificação. Em seguida, foi calculado o *F-score* para verificar a precisão da previsão consenso de cada uma das 3 iterações da validação cruzada, tendo sido o valor médio de *F-score* atribuído a cada valor de "n". Para o conjunto VinaFP a melhor divisão corresponde a "n"=33, enquanto para o conjunto VinaCluster a melhor divisão corresponde a "n"=32. Embora não tenha sido referido é de salientar que foram realizadas 12 repetições de todo o processo aleatório de criação dos conjuntos de treino e teste, seleccionando-se a repetição que obteve o melhor F-*score* final.

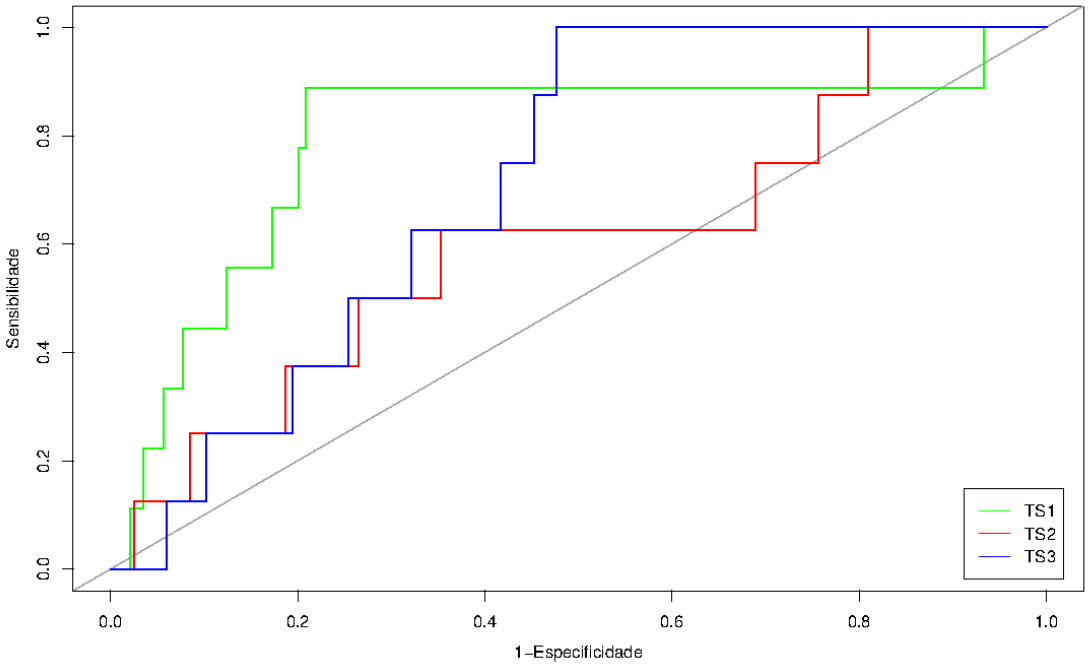
4.2 – Análise do desempenho dos classificadores

Após se escolher o melhor "n", foram calculados os valores de AUC das curvas das classificações obtidas para cada um dos conjuntos de teste com o objectivo se seleccionar o modelo de classificação obtido com melhor desempenho. As curvas ROC para os 3 melhores classificadores, foram produzidas para os conjuntos de teste correspondentes (Figura 15). De seguida, estes classificadores foram utilizados para classificar o conjunto total de compostos activos e inactivos presentes no pacote da DUD, e o seu desempenho nas classificações dos compostos foi avaliado através da análise das curvas ROC (Figura 16). Os valores de AUC para cada uma destas curvas ROC é apresentado na Tabela 3. Para o conjunto VinaFP, o melhor classificador (Figura 15A, a verde) apresenta uma AUC de 79,7% no conjunto de teste e de 74,5% no conjunto total de compostos. Por outro lado, para o conjunto VinaCluster, o melhor classificador (Figura 15B, a verde) apresenta uma AUC de 77,2% no conjunto de teste e de 76,4% no conjunto total de compostos. Os classificadores com melhor desempenho para o conjunto total de compostos (Figura 16A, a azul: Figura 16B, a azul) tem uma AUC de 74,9% e de 76,9% para o conjunto VinaFP e VinaCluster respectivamente.

Analisando a Tabela 3, podemos observar que globalmente o desempenho dos classificadores obtidos a partir dos dados das poses do conjunto VinaCluster é ligeiramente superior aos obtidos com os dados das poses do conjunto VinaFP. Além disso, observa-se também que o desempenho dos classificadores melhora quando é classificado um grande conjunto de compostos.

Para os melhores classificadores obtidos da comparação e avaliação do desempenho pelas curvas ROC e a AUC, fez-se também uma análise recorrendo à construção de curvas de enriquecimento e dos factores de enriquecimento correspondentes ao topo 1%, 5% e 10% tal como tinha sido realizado para a avaliação da função de pontuação do AutoDock Vina. No entanto neste caso pretende-se avaliar e comparar a eficácia que os classificadores seleccionados têm em reconhecer precocemente compostos activos nos topos escolhidos.

A)



B)

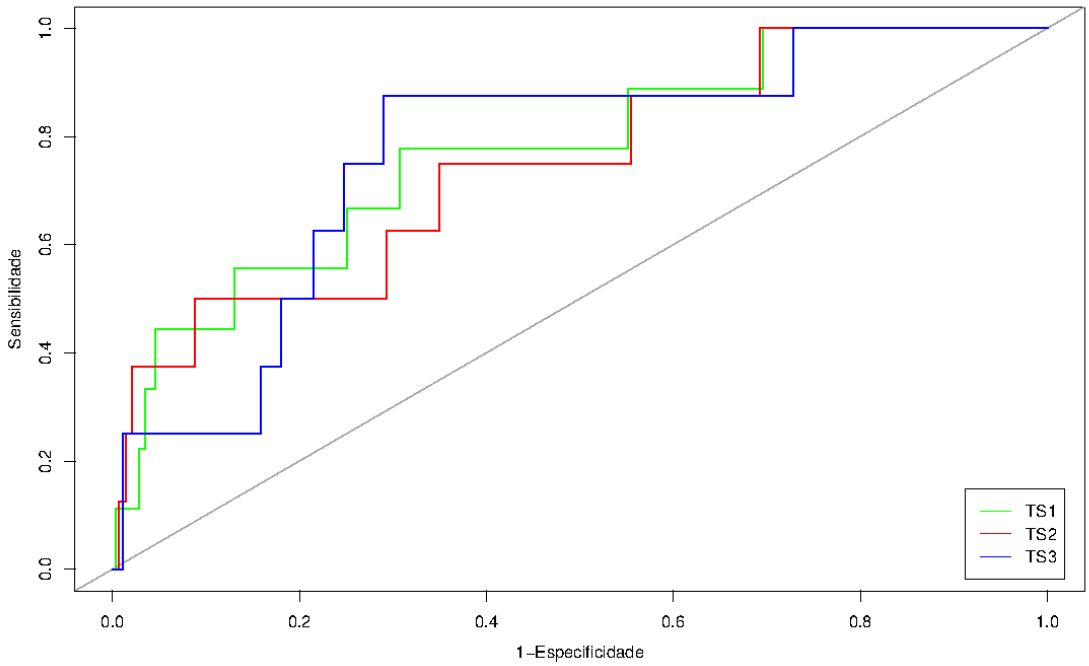
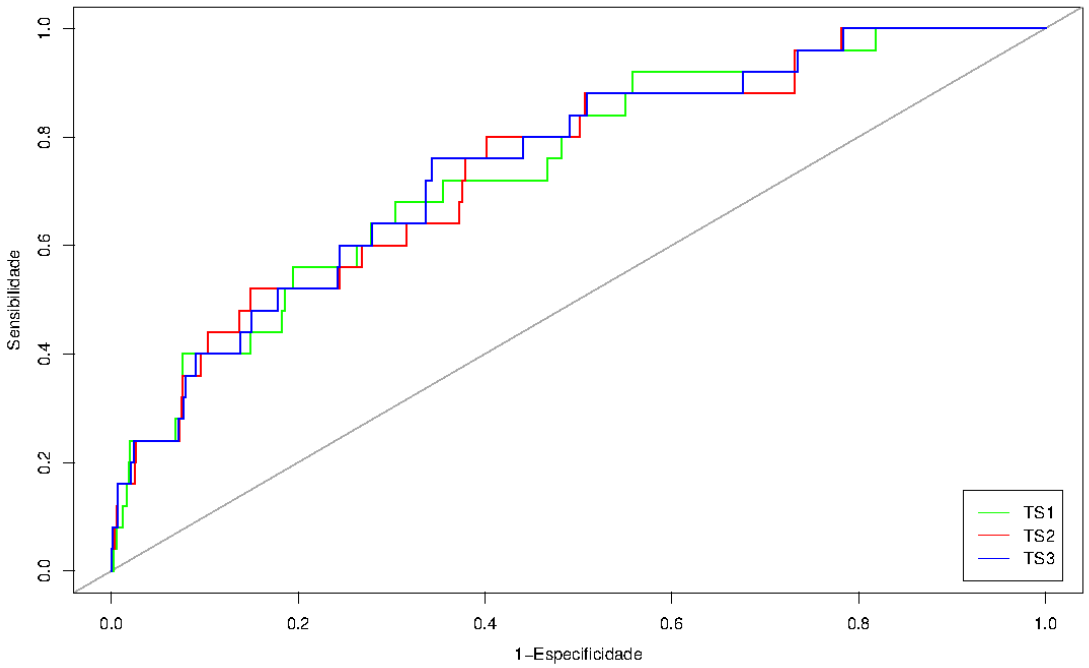


Figura 15 – Curvas ROC para os conjuntos de teste 1, 2 e 3. A linha na diagonal representa uma performance aleatória. A verde, vermelho e azul encontram-se representadas as curvas ROC para os modelos de classificação obtidos com os conjuntos de teste 1, 2 e 3 respectivamente. A) Curvas ROC dos melhores classificadores das poses do conjunto VinaFP. B) Curvas ROC dos melhores classificadores das poses do conjunto VinaCluster.

A)



B)

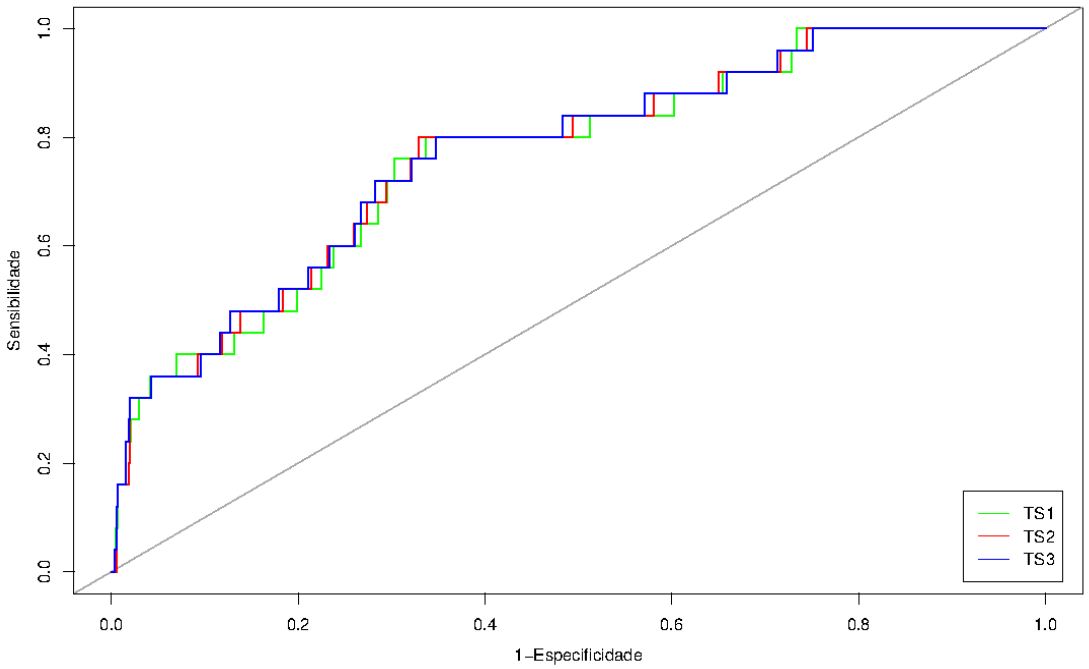


Figura 16 – Curvas ROC para o conjunto de teste total. A linha na diagonal representa uma performance aleatória. A) Curvas ROC dos melhores classificadores das poses do conjunto VinaFP. B) Curvas ROC dos melhores classificadores das poses do conjunto VinaCluster.

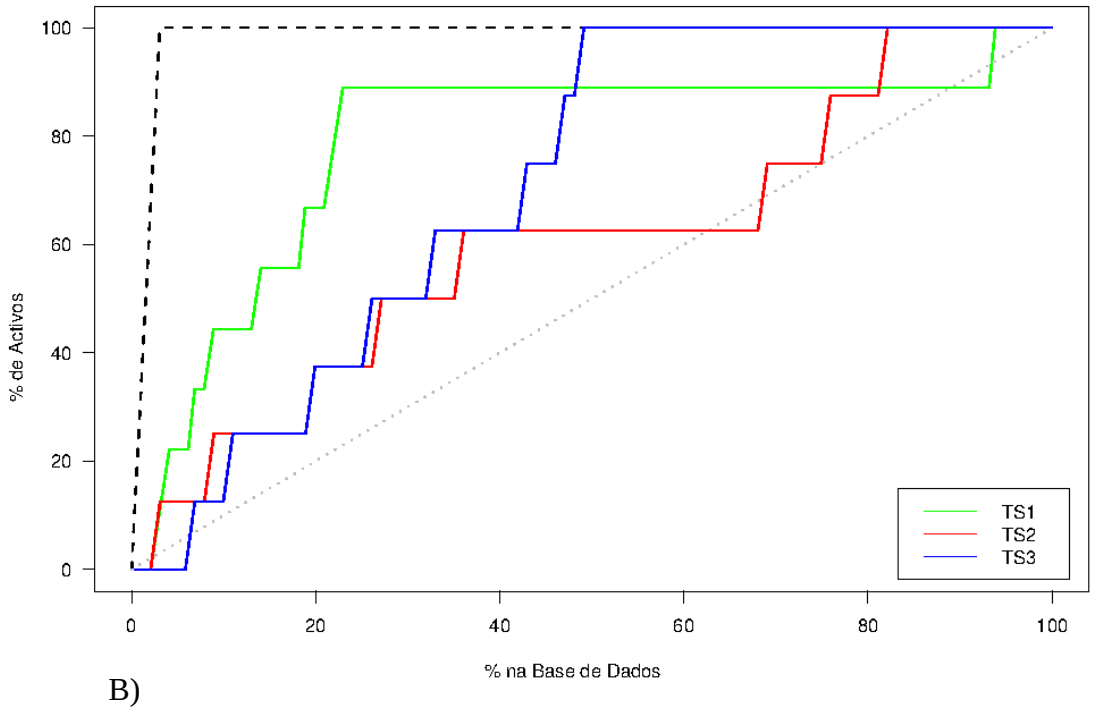
A Figura 17 apresenta as curvas de enriquecimento geradas para os melhores classificadores obtidos para os conjuntos de teste das poses do conjunto VinaFP (Figura 17A) e pelo método de selecção baseado na análise de grupos (Figura 17B). Analisando estas curvas observa-se que para o conjunto VinaFP, o classificador com melhor eficácia (Figura 17A, a verde) apresenta factores de enriquecimento de 0, 4,4 e 4,4 respectivamente no topo 1, 5 e 10% para o conjunto de teste e factores de enriquecimento de 8, 4,8 e 4 para o conjunto total de compostos. Para o conjunto VinaCluster, o classificador com melhor eficácia (Figura 17B, a vermelho) apresenta factores de enriquecimento de 11,4; 7,5 e 5 também no topo 1, 5 e 10% para o conjunto de teste e factores de enriquecimento de 12, 6,4 e 3,8 para o conjunto total de compostos.

Na Figura 18 são apresentadas as curvas de enriquecimento dos classificadores com melhores eficácia para o conjunto total de compostos em cada um dos conjuntos (Figura 18, A e B). Pela observação das curvas não é perceptível qual o classificador que apresenta melhor precisão em cada um dos métodos de selecção de poses. Recorrendo à análise dos factores de enriquecimento (Tabela 3) percebe-se que dentro de cada método os factores de enriquecimento dos diferentes classificadores são muito parecidos, mas em ambos os métodos o classificador mais eficaz é o criado com o conjunto de teste 2 (Figura 18A, a vermelho; Figura 18B, a vermelho).

Na Tabela 3 são também apresentados os valores de sensibilidade, especificidade, precisão e *F-scores* calculados tendo como base a melhor linha de corte para a função de decisão dos melhores classificadores para o conjunto total de compostos activos e inactivos. Nos dois casos estes classificadores correspondem ao conjunto de teste TS_{3T} (Tabela 3). Das oito métricas consideradas, o classificador construído com base no conjunto VinaFP apenas teve um valor ligeiramente melhor para a especificidade.

Embora os valores entre os dois melhores classificadores construídos a partir dos dois conjuntos não apresentem diferenças muito significativas, estes resultados demostram que os classificadores obtidos com as poses seleccionadas utilizando a análise de grupos (VinaCluster) apresentam melhores valores de sensibilidade, precisão, F-*score*, AUC e EF (para todos os topos) do que os classificadores construídos com as poses seleccionadas pela função de pontuação (VinaFP). Isto parece indicar que aplicar a análise de grupos na selecção das poses obtidas com o AutoDock Vina e a incluir nos

A)



B)

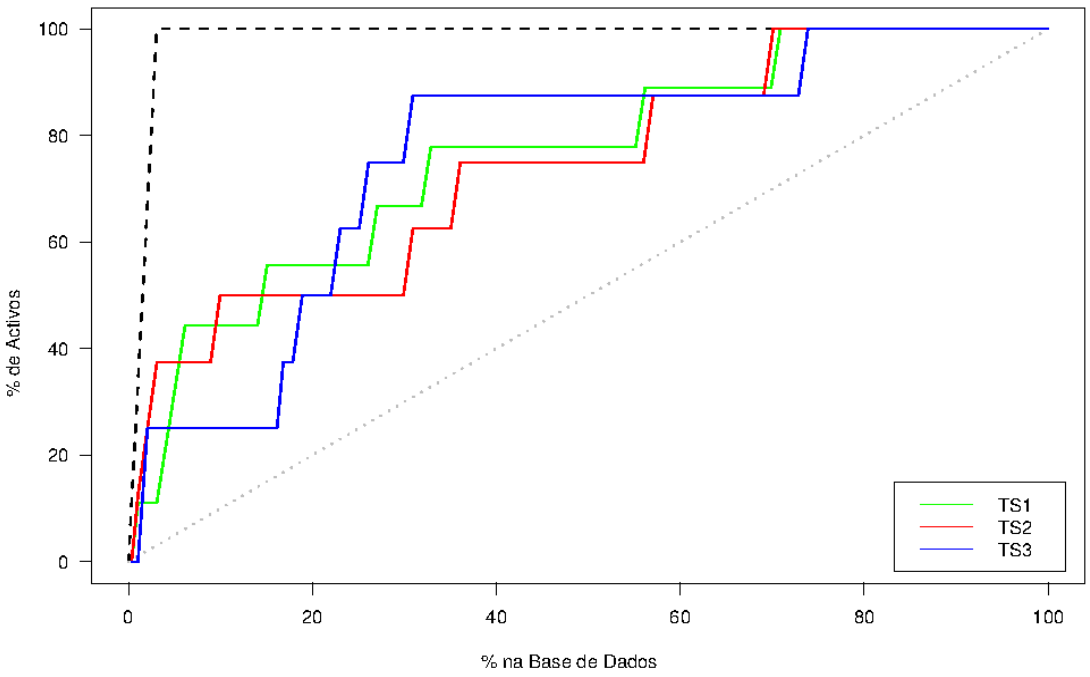


Figura 17 – Curvas de Enriquecimento para os conjuntos de teste 1, 2 e 3. A linha na diagonal representa uma performance aleatória e a linha a tracejado no topo superior da imagem representa o caso ideal. A) Curvas de enriquecimento dos melhores classificadores das poses do conjunto VinaFP B) Curvas de enriquecimento dos melhores classificadores das poses do conjunto VinaCluster.

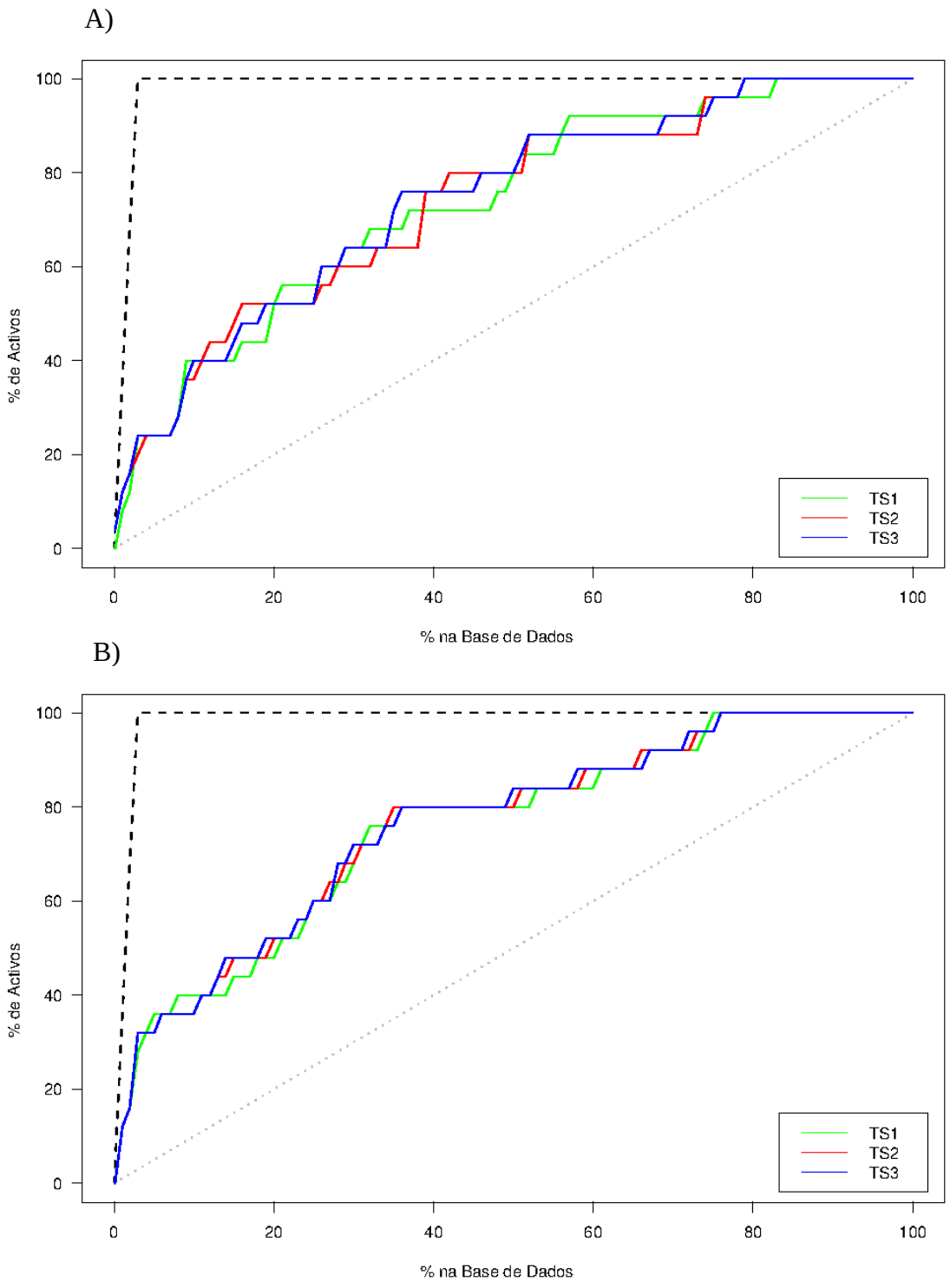


Figura 18 – Curvas de Enriquecimento para o conjunto de teste total. A linha na diagonal representa uma performance aleatória e a linha a tracejado no topo superior da imagem representa o caso ideal. A) Curvas de enriquecimento dos melhores classificadores das poses no conjunto VinaFP. B) Curvas de enriquecimento dos melhores classificadores das poses no conjunto VinaCluster.

conjuntos de treino para SVM, gera classificadores com melhor capacidade em discriminar e reconhecer eficientemente compostos activos de inactivos para todo o conjunto para a COX-1 presente no pacote da DUD.

Tabela III – Diferentes métricas de avaliação dos melhores classificadores. Sensibilidade, especificidade, previsão, *F-score*, AUC e factores de enriquecimento (EF) obtidos para as melhores linhas de corte das curvas produzidas para os melhores classificadores obtidos pelo treino com o SVM-light. Os valores das métricas dos melhores classificadores para cada método encontram-se assinalados a negrito e sublinhados. TS_{XT} representa o classificador obtido pelo conjunto de treino X foi utilizado para classificar o conjunto total de dados T.

Medidas	**Métodos de classificação SVM**											
	VinaFP						VinaCluster					
	Conjuntos de testes			Conjunto de testes total			Conjuntos de testes			Conjunto de testes total		
	TS_1	TS_2	TS_3	TS_{1T}	TS_{2T}	**TS_{3T}**	TS_1	TS_2	TS_3	TS_{1T}	TS_{2T}	**TS_{3T}**
Sensibilidade %	88,9	62,5	100	68	80	**76**	77,8	50	87.5	80	80	**80**
Especificidade %	79,2	64,7	52,3	69.6	59,8	**65,7**	69.3	91.2	71	66.3	67.1	**65,3**
Precisão %	11,9	48	56	66	56	**61**	74.5	13.8	79	65	67	**63**
F-Score %	20,9	89	10,6	12,0	10,5	**11,3**	76.1	21.6	14.4	12.1	12.3	**11,7**
AUC %	79,7	60,4	71,5	74,5	74,4	**74,9**	77,2	74,7	76,9	76,4	76,8	**76,9**
EF a 1%	0	0	0	8	12	**10,96**	11,11	11,38	0	10,96	12	**12**
EF a 5%	4,4	2,5	0	4,8	4,8	**4,8**	6,7	7,5	5	7	6,4	**6,4**
EF a 10%	4,4	2,5	1,3	4	3,6	**4**	4,4	5	2,5	4	3,8	**3,6**

Capítulo 5 Conclusão

O acoplamento molecular é uma técnica computacional de rastreio virtual que tenta prever a melhor conformação de um ligando e a sua orientação, no local activo de uma molécula alvo. A procura de ferramentas e metodologias capazes de prever boas poses de compostos no local activo de uma proteína e de as pontuar de uma forma rápida tem sido um dos maiores desafios na realização de experiências de rastreio virtual.

O presente trabalho foi realizado com o intuito de explorar soluções para a resolução destes problemas utilizando a COX-1 como um caso de estudo. A COX-1 foi a proteína alvo escolhida porque está presente em quase todas as células do corpo humano e encontra-se envolvida em diversas funções vitais nomeadamente, nos processos de síntese de prostanóides a partir de ácido araquidónico que actuam ao nível do sistema imunológico como resposta a um processo inflamatório. Outro motivo é o facto de esta proteína ser fortemente afectada por AINES, que muito embora não sejam desenvolvidos com esse propósito alteram ou inibem a sua função. Assim, é importante desenvolver estratégias que permitam uma melhor discriminação de compostos activos e inactivos para a COX-1 de maneira a se evitar desenvolver fármacos que inibam ou afectem a sua actividade indirectamente.

Primeiramente, avaliou-se a capacidade do programa de livre acesso AutoDock Vina prever e pontuar resultados de acoplamento molecular para a COX-1 com base no ligando BFL. As poses obtidas foram avaliadas através da sua visualização e do cálculo do RMSD entre as poses obtidas pelo AutoDock Vina para o ligando BFL e a sua pose na estrutura cristalográfica. As duas melhores poses obtidas apresentam um RMSD de 1,89 e 1,88 Å em relação à estrutura cristalográfica, e um mesmo valor de afinidade de -9,8 kcal/mol, o que significa que estas poses geradas pelo programa são muito semelhantes à estrutura do ligando BFL e apresentam boa capacidade de afinidade com o local activo. Destas análises ficou claro que o programa consegue prever boas poses para ligandos da COX-1.

Foi depois realizado um acoplamento molecular para todos os compostos do pacote da DUD para a COX-1 tentando-se obter mais poses para cada um dos compostos. As melhores poses seleccionadas pela função de pontuação do AutoDock Vina foram visualizadas e constatou-se que para 35 compostos de um total de 849, a melhor pose seleccionada não se situava na região "óptima" do local activo. O tamanho e forma dessas poses parecem "impossibilitar" o programa de as "colocar" dentro do local activo e como tal recorreu-se a uma estratégia alternativa para tentar seleccionar

melhores com base na sua estrutura e não nas afinidades obtidas pela função de pontuação do programa.

De seguida, o desempenho da função de pontuação do AutoDock Vina foi avaliado para dois conjuntos de "melhores" poses, seleccionadas após o acoplamento molecular para a COX-1 com todos os compostos presentes no pacote da DUD, através da análise de curvas ROC e respectivos valores de AUC e das curvas de enriquecimento e factores de enriquecimento. Um dos conjuntos foi seleccionado com base na ordenação dada pela função de pontuação do AutoDock Vina (VinaFP) e o outro conjunto foi seleccionado com base numa análise de grupos (VinaCluster). Os resultados obtidos foram apenas razoáveis em qualquer um dos conjuntos de poses considerados, uma vez que para o conjunto de poses VinaFP o valor de AUC foi de 63,44% e para o conjunto de poses VinaCluster foi de apenas 55,06%. Em relação à eficácia no reconhecimento precoce de compostos activos pela função de pontuação do AutoDock Vina, foi para o conjunto de poses VinaCluster que se obteve os melhores factores de enriquecimento. É contudo claro que a pontuação e discriminação obtida com base na função de pontuação do AutoDock Vina é insuficiente para obter resultados fidedignos e confiáveis numa campanha de rastreio virtual com compostos desconhecidos.

Tendo em conta os resultados iniciais obtidos, delineou-se uma estratégia de desenvolvimento de modelos de classificação usando os parâmetros constituintes da função de pontuação do AutoDock Vina. Neste caso, os resultados obtidos levaram a concluir que os classificadores treinados com o conjunto de poses VinaCluster apresentaram melhor desempenho e eficácia do que os classificadores treinados com o conjunto de poses VinaFP. O melhor classificador gerado a partir do conjunto de poses VinaCluster apresenta uma capacidade de discriminação de compostos activos e inactivos para a COX-1 (AUC= 76,9) e um reconhecimento precoce de compostos activos muito superiores (EF1%= 12) quer em relação à função de pontuação do AutoDock Vina (VinaFP: AUC= 63,44; EF10%= 1,6 / VinaCluster: AUC=55,06; EF10%= 2) quer em relação aos classificadores gerados pelas poses no conjunto VinaFP (AUC= 74,9; EF1%= 10,96). Estes resultados claramente demonstram que a estratégia aqui desenvolvida com base na escolha de "melhores" poses a partir de método de análise de grupos e na utilização de SVM para treinar classificadores melhorou significativamente a classificação de compostos. Estes resultados mostram

também que as funções de pontuação ainda se encontram longe de obter bons resultados e que estratégias alternativas devem ser exploradas.

Assim, é possível concluir que a utilização de SVMs para o desenvolvimento de classificadores apresenta melhorias significativas na classificação de resultados de acoplamento molecular, o que já tinha sido sugerido noutros trabalhos (Cannon *et al*., 2007; Kinnings *et al*., 2011). É possível também concluir que nem sempre as "melhores" poses seleccionadas pelos programas de acoplamento molecular são as melhores para o local activo em estudo, podendo a utilização de outras técnicas de selecção de poses, como a análise de grupos utilizada no presente trabalho melhorar significativamente os resultados do acoplamento molecular.

Em termos de utilidade a metodologia descrita neste trabalho poderá ser aplicada noutros alvos de interesse para futuras experiências de rastreio virtual que utilizem técnicas de acoplamento molecular, tal como pode ser utilizada para estimular a criação de novas e melhoradas funções de pontuação ou aplicada a outros programas já existentes de acoplamento molecular com o intuito de classificar melhor os resultados por eles obtidos.

Como perspectivas de trabalho futuro as hipóteses podem passar por testar a capacidade da função de pontuação do programa AutoDock Vina utilizando mais compostos com dados experimentais e estruturas cristalográficas disponíveis para além do composto BFL enriquecendo o conjunto de dados que foi utilizado. Outra hipótese interessante seria utilizar outros programas de acoplamento molecular e/ou funções de pontuação com capacidade de gerar mais parâmetros de caracterização das poses obtidas, com o objectivo de se tentar obter um classificador com mais poder discriminativo entre compostos activos e inactivos.

Bibliografia

Andrade, C. (2004). Analgésicos Inibidores Específicos da Ciclooxigenase-2: Avanços Terapêuticos *. *Revista Brasileira de Anestesiologia*, *54*, 448-464.

Andricopulo, A. D., Salum, L. B., & Abraham, D. J. (2009). Structure-based drug design strategies in medicinal chemistry. *Current topics in medicinal chemistry*, *9*(9), 771-790.

Ballester, P. J., & Mitchell, J. B. O. (2010). A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics (Oxford, England)*, *26*(9), 1169-75.

Barril, X., Hubbard, R. E., & Morley, S. D. (2004). Virtual screening in structure-based drug discovery. *Mini reviews in medicinal chemistry*, *4*(7), 779-791.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. *et al*. (2000). The Protein Data Bank. *Nucleic acids research*, *28*(1), 235-42.

Bouvier, G., Evrard-Todeschi, N., Girault, J.-P., & Bertho, G. (2010). Automatic clustering of docking poses in virtual screening process using self-organizing map. *Bioinformatics (Oxford, England)*, *26*(1), 53-60.

Burges, Christopher J. C., (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2), 121–167.

Cannon, E. O., Amini, A., Bender, A., Sternberg, M. J. E., Muggleton, S. H., Glen, R. C., & Mitchell, J. B. O. (2007). Support vector inductive logic programming outperforms the naive Bayes classifier and inductive logic programming for the classification of bioactive chemical compounds. *Journal of computer-aided molecular design*, *21*(5), 269-80.

Carvalho W.A., Carvalho R.D.S., Rios-Santos F. (2004). Analgésicos Inibidores Específicos da Ciclooxigenase-2: Avanços Terapêuticos. *Revista Brasileira de Anestesiologia*, *54(3)*, 448 – 464

Cheng, T., Li, X., Li, Y., Liu, Z., & Wang, R. (2009). Comparative assessment of scoring functions on a diverse test set. *Journal of chemical information and modeling*, *49*(4), 1079-93.

Consortium, T. U. (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research*, *40* (Database issue), D71-5.

Cortes C, Vapnik V (1995) Support-vector networks. *Machine Learning*, 20, 273-297

Coupez, B., & Lewis, R. A. (2006). Docking and scoring--theoretically easy, practically impossible? *Current medicinal chemistry*, *13*(25), 2995-3003.

Dannhardt, G., & Kiefer, W. (2001). Review Cyclooxygenase inhibitors – current status and future prospects. *European Journal of Medicinal Chemistry*, *36*, 109-126.

Delaglio, F. (2001). Virtual Screening Methods for Drug Discovery. *Pharmaceutical Sciences*.

Deng, W., Breneman, C., & Embrechts, M. J. (2004). Predicting protein-ligand binding affinities using novel geometrical descriptors and machine-learning methods. *Journal of chemical information and computer sciences*, *44*(2), 699-703.

Doman, T. N., McGovern, S. L., Witherbee, B. J., Kasten, T. P., Kurumbail, R., Stallings, W. C., Connolly, D. T. *et al*. (2002). Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *Journal of medicinal chemistry*, *45*(11), 2213-21.

Duan, J., Dixon, S. L., Lowrie, J. F., & Sherman, W. (2010). Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *Journal of molecular graphics & modelling*, *29*(2), 157-170.

Ewing, T. J. A., Makino, S., Skillman, A. G. & Kuntz, I. D. (2001). DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des. 15*, 411–428.

Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perr y JK, Shaw DE, Francis P, Shenkin PS. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem. 47,* 1739–1749.

Garavito, R. M., & DeWitt, D. L. (1999). The cyclooxygenase isoforms: structural insights into the conversion of arachidonic acid to prostaglandins. *Biochimica et biophysica acta*, *1441*(2-3), 278-87.

Garavito, R. M., Malkowski, M. G., & DeWitt, D. L. (2002). The structures of prostaglandin endoperoxide H synthases-1 and -2. *Prostaglandins & other lipid mediators*, *68-69*, 129-52.

Gohlke, H., Hendlich, M. & Klebe, G. (2000) Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol. 295*, 337–356.

Graf, A. B. A., & Borer, S. (2001). Normalization in Support Vector Machines. *Neural Computation*, 277-282.

Gupta, K., Selinsky, B. S., Kaub, C. J., Katz, A. K., & Loll, P. J. (2004). The 2.0Å Resolution Crystal Structure of Prostaglandin H2 Synthase-1: Structural Insights into an Unusual Peroxidase. *Journal of Molecular Biology*, *335*(2), 503-518.

Halperin, I., Ma, B., Wolfson, H., & Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, *47*(4), 409-443.

Hamza, A., Wei, N.-N., & Zhan, C.-G. (2012). Ligand-Based Virtual Screening Approach Using a New Scoring Function. *Journal of chemical information and modeling*. 52, 963-974.

Han, L. Y., Zheng, C. J., Xie, B., Jia, J., Ma, X. H., Zhu, F., Lin, H. H. *et al*. (2007). Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. *Drug discovery today*, *12*(7-8), 304-13.

Hasegawa, K., & Funatsu, K. (2010). Non-linear modeling and chemical interpretation with aid of support vector machine and regression. *Current computer-aided drug design*, *6*(1), 24-36.

Hiroaki Yabuuchi (2011). enrichvs: Enrichment assessment of virtual screening approaches.R package version 0.0.5. http://CRAN.R-project.org/package=enrichvs

Huang, N., Shoichet, B. K., & Irwin, J. J. (2006). Benchmarking sets for molecular docking. *Journal of medicinal chemistry*, *49*(23), 6789-801

Huang, S.-Y., Grinter, S. Z., & Zou, X. (2010). Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Physical chemistry chemical physics : PCCP*, *12*(40), 12899-908.

Humphrey, W., Dalke, a, & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of molecular graphics*, *14*(1), 33-8, 27-8.

Jackson, R. C. (1995). Update on computer-aided drug design. *Current opinion in biotechnology*, *6*(6), 646-651.

Jain, A. N. (2006). Scoring functions for protein-ligand docking. *Current protein & peptide science*, *7*(5), 407-20.

Jain, A. N., & Nicholls, A. (2008). Recommendations for evaluation of computational methods. *Journal of computer-aided molecular design*, *22*(3-4), 133-9.

Jain, N. (1996). Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. *Journal of computer-aided molecular design*, *10*(5), 427-40.

Jenkins, J. L., Glick, M., & Davies, J. W. (2004). A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. *Journal of medicinal chemistry*, *47*(25), 6144-6159.

Joachims, T., Schölkopf, B. Burges, C. and Smola, A. (1999) Advances in Kernel Methods - Support Vector Learning (ed.), in: *Making large-Scale SVM Learning Practical*. MIT Press, 1999.

Jorissen, R. N., & Gilson, M. K. (2005). Virtual screening of molecular databases using a support vector machine. *Journal of chemical information and modeling*, *45*(3), 549-61.

Kinnings, S. L., Liu, N., Tonge, P. J., Jackson, R. M., Xie, L., & Bourne, P. E. (2011). A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *Journal of chemical information and modeling*, *51*(2), 408-19.

Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews. Drug discovery*, *3*(11), 935-949.

Kummer C. L., Coelho T. C. R. B. (2002). Antiinflamatórios Não Esteróides Inibidores da Ciclooxigenase-2 (COX-2): Aspectos Atuais. *Revista Brasileira de Anestesiologia*, *52*, 498-512.

Kurt Hornik (2011). The R FAQ. ISBN} 3-900051-08-9 (http://CRAN.R-project.org/doc/FAQ/R-FAQ.html)

Lazarova, M. (2008). Virtual Screening – Models , Methods and Software Systems. *Current*, 55-60.

Lengauer, T., & Rarey, M. (1996). Computational methods for biomolecular docking. *Current opinion in structural biology*, *6*(3), 402-406.

Lill, M. A. (2011). Efficient incorporation of protein flexibility and dynamics into molecular docking simulations. *Biochemistry*, *50*(28), 6157-6169.

Lorena, A. C., & Carvalho, A. C. P. L. F. D. (2007). Uma Introdução às Support Vector Machines. *RITA*, 16(2), 43-67

Melville, J. L., Burke, E. K., & Hirst, J. D. (2009). Machine learning in virtual screening. *Combinatorial chemistry & high throughput screening*, *12*(4), 332-43.

Michel F. Sanner (1999) Python: A Programming Language for Software Integration and Development. *J. Mol. Graphics Mod., 17,* 57-61.

Morita, I. (2002). Distinct functions of COX-1 and COX-2, *69*, 165-175.

Morris G.M., Huey R., Lindstrom W., Sanner MF., Belew R.K., Goodsell D.S, Olson A.J. (2009). AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem. 30*, 2785–2791.

Neudert, G., & Klebe, G. (2011). fconv: Format conversion, manipulation and feature computation of molecular data. *Bioinformatics (Oxford, England)*, *27*(7), 1021-2.

Nicholls, A. (2008). What do we know and when do we know it? *Journal of computer-aided molecular design*, *22*(3-4), 239-55.

Noble, W. S. (2004). Support vector machine applications in computational biology. *MIT Press*, 71-92

Noble, W. S., & Street, P. (2006). What is a support vector machine ?, *24*(12), 1565-1568.

Onodera, K., Satou, K., & Hirota, H. (2007). Evaluations of molecular docking programs for virtual screening. *Journal of chemical information and modeling*, *47*(4), 1609-1618.

Oprea, T. I., & Matter, H. (2004). Integrating virtual screening in lead discovery. *Current opinion in chemical biology*, *8*(4), 349-358.

Pérez-Nueno, V. I., Ritchie, D. W., Rabal, O., Pascual, R., Borrell, J. I., & Teixidó, J. (2008). Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand-receptor docking. *Journal of chemical information and modeling*, *48*(3), 509-533

Picot D, Loll PJ, Garavito RM. (1994) The X-ray crystal structure of the membrane protein prostaglandin H2 synthase-1. *Nature*; 367:243–9.

Plewczynski, D., Spieser, S. a H., & Koch, U. (2009). Performance of machine learning methods for ligand-based virtual screening. *Combinatorial chemistry & high throughput screening*, *12*(4), 358-68.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rarey M, Kramer B, Lengauer T, Klebe G. (1996). A fast flexible docking method using an incremental construction algorithm. *J Mol Biol, 261*,470 – 489.

Reddy, A. S., Pati, S. P., Kumar, P. P., Pradeep, H. N., & Sastry, G. N. (2007). Virtual screening in drug discovery -- a computational perspective. *Current protein & peptide science*, *8*(4), 329-351.

Ripphausen, P., Nisius, B., Peltason, L., & Bajorath, J. (2010). Quo vadis, virtual screening? A comprehensive survey of prospective applications. *Journal of medicinal chemistry*, *53*(24), 8461-8467.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Jean-Charles Sanchez, J-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics,12,* 77.

Schulz-Gasch, T., & Stahl, M. (2004). Scoring functions for protein–ligand interactions: a critical perspective. *Drug Discovery Today: Technologies*, *1*(3), 231-239.

Smith, W. L., DeWitt, D. L., & Garavito, R. M. (2000). Cyclooxygenases: structural, cellular, and molecular biology. *Annual review of biochemistry*, *69*, 145-82.

Sousa, S. F., Cerqueira, N. M. F. S. a, Fernandes, P. a, & Ramos, M. J. (2010). Virtual screening in drug design and development. *Combinatorial chemistry & high throughput screening*, *13*(5), 442-453.

Stahl, M., & Rarey, M. (2001). Detailed analysis of scoring functions for virtual screening. *Journal of medicinal chemistry*, *44*(7), 1035-1042.

Stahura, F. L., & Bajorath, J. (2004). Virtual screening methods that complement HTS. *Combinatorial chemistry & high throughput screening*, *7*(4), 259-69.

Sun, H. (2008). Pharmacophore-Based Virtual Screening. *Current*, (973), 1018-1024.

Taylor, R. D., Jewsbury, P. J., & Essex, J. W. (2002). A review of protein-small molecule docking methods. *Journal of computer-aided molecular design*, *16*(3), 151-166.

Triballeau, N., Acher, F., Brabet, I., Pin, J.-P., & Bertrand, H.-O. (2005). Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *Journal of medicinal chemistry*, *48*(7), 2534-47.

Trott, O., Olson, A. J., (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading, *Journal of Computational Chemistry 31 455-461*.

Truchon, J.-F., & Bayly, C. I. (2007). Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of chemical information and modeling*, *47*(2), 488-508.

Vane, J. R., Bakhle, Y. S., & Botting, R. M. (1998). Cyclooxygenases 1 and 2. *Annual review of pharmacology and toxicology*, *38*, 97-120.

Walters, W. P., Stahl, M. T., & Murcko, M. A. (1998). Virtual screening – an overview. *Science*, *3*(4), 160-178.

Wang, R., Lai, L., & Wang, S. (2002). Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of computer-aided molecular design*, *16*(1), 11-26.

Wang, R.; Fang, X.; Lu, Y.; Wang, S.(2004) "The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures", *J. Med. Chem.*, *47*(12); 2977-2980.

Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. (2005) "The PDBbind Database: Methodologies and updates", *J. Med. Chem.*, *48*(12); 4111-4119.

Yang, S.-Y. (2010). Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug discovery today*, *15*(11-12), 444-450.

Yap, C. W., & Chen, Y. Z. (2005). Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *Journal of chemical information and modeling*, *45*(4), 982-92.

Yuriev, E., Agostino, M., & Ramsland, P. a. (2011). Challenges and advances in computational docking: 2009 in review. *Journal of molecular recognition : JMR*, *24*(2), 149-164.