Joana Filipa Monteiro de Sousa

RASTREIO VIRTUAL NA DESCOBERTA DE POSSÍVEIS INIBIDORES DA 5α–REDUTASE

Dissertação de Mestrado em Química Farmacêutica Industrial, orientada pela Doutora Cândida G. Silva e pelo Professor Doutor Jorge António Ribeiro Salvador e apresentada à Faculdade de Farmácia da Universidade de Coimbra

Setembro 2013

UNIVERSIDADE DE COIMBRA

Agradecimentos

Ao longo da realização deste trabalho inúmeras pessoas estiveram envolvidas, apresentando um papel fundamental. A todas elas os meus sinceros agradecimentos, em especial:

À Doutora Cândida G. Silva, que me acompanhou sempre ao longo do trabalho. Obrigada pela disponibilidade e acessibilidade demonstradas. Pelos bons ensinamentos, pela orientação fundamental na concretização desta dissertação. Pelos conselhos, compreensão e paciência no esclarecimento de qualquer dúvida.

Ao Professor Doutor Rui Brito, pela oportunidade em integrar no seu grupo de trabalho. Pela orientação, conselhos, compreensão e disponibilidade demonstrada.

Ao Professor Doutor Jorge Salvador, pela orientação, preocupação, acompanhamento e disponibilidade que demonstrou ao longo deste trabalho.

À Doutora Elsa Henriques, que me acompanhou numa parte inicial do trabalho. Pelos bons ensinamentos, orientação, conselhos e paciência no esclarecimento de qualquer dúvida.

Ao Doutor Carlos Simões, pelo material científico cedido, pela disponibilidade e conselhos dados.

A todos os restantes membros do grupo RMBLab, em especial, Catarina, Raquel, Zaida, Pedro, Tiago pelas muitas horas de companhia, dicas, ajuda e boa disposição.

A todos os meus amigos que de alguma forma deram-me ânimo e me incentivaram nos momentos mais difíceis. Em especial à Kristina, Sara e Cristina por serem as pessoas que me acompanharam desde o início desta fase, e tiveram sempre uma palavra de ânimo e alento nos momentos mais importantes.

Aos meus irmãos pelo carinho, compreensão e força que me transmitiram quando eu mais precisava. Principalmente à Ana Raquel, pela companhia e compreensão, mesmo quando via nela a única forma de aliviar o stress.

Aos meus pais, por todo o esforço e dedicação, pois sem eles não seria possível alcançar mais esta etapa. Por acreditarem e me incentivarem a seguir sempre o meu caminho.

Índice

Índice de figuras

Índice de Tabelas

Abreviaturas

3α-diol – 3 alfa-androstanodiol

3β-diol – 3 beta-androstanodiol

3D – tridimensional

3α-HSD –3 alfa-hidroxiesteróide desidrogenase

5α-redutase – 5 alfa-redutase

5α-R1 – 5 alfa-redutase 1

5α-R2 – 5 alfa-redutase 2

ASA – Área de superfície molecular acessível ao solvente de todos os átomos

ASA+ – Área de superfície molecular acessível ao solvente de todos os átomos com carga parcial positiva

ASA- – Área de superfície molecular acessível ao solvente de todos os átomos com carga parcial negativa

ASA-H – Área de superfície molecular acessível ao solvente de todos os átomos hidrofóbicos

ASA-P – Área de superfície molecular acessível ao solvente de todos os átomos hidrofílicos

CHO – Ovário de hamster chinês

DHT – Di-hidrotestosterona

ADN – Ácido desoxirribonucleico

HBP – Hiperplasia benigna da próstata

IC50 – Concentração inibitória de 50%

LBVS – Rastreio virtual baseado em Ligandos (do inglês, *Ligand-based virtual screening)*

MCS – Subestrutura máxima comum (do inglês, *Maximum Common Substructure*)

nM – Nanómetro

PCA – Análise de componentes principais

PC1 – Primeira componente principal

PC2 – Segunda componente principal

PC3 – Terceira componente principal

PC4 – Quarta componente principal

QSAR – Relação quantitativa de estrutura-atividade

RMN – Ressonância magnética nuclear

SBVS – Rastreio Virtual baseado na estrutura (do inglês, *Structure-based virtual screening*)

SVM – Máquinas de vetores de suporte

T – Testosterona

$\mu g.ml^{-1}$ – micromolar por mililitro

Resumo

A 5α-redutase é uma proteína microssomal, que converte a testosterona em di-hidrotestosterona (DHT) um androgénio mais potente que tem como função induzir a diferenciação durante o desenvolvimento do feto que conduz ao desenvolvimento dos genitais externos masculinos. Quando ocorrem distúrbios no desempenho desta enzima, podem surgir distúrbios como o pseudo-hermafroditismo, a calvície, a hiperplasia benigna e o cancro da próstata, entre outras. O cancro da próstata e a hiperplasia benigna da próstata são doenças que têm vindo a aumentar nos últimos anos. Existindo apenas dois fármacos comercializados, a finasterida e a dutasterida, que possuem no entanto efeitos colaterais, sendo por isso necessário o desenvolvimento de melhores inibidores para esta enzima.

Neste trabalho procedeu-se numa primeira etapa à tentativa de se obter informações relevantes sobre a possível estrutura da enzima 5α-redutase utilizando a técnica de modelação de proteínas por homologia. Numa segunda etapa tentou-se obter informação sobre a atividade dos inibidores desta enzima, de modo a contribuir para o desenvolvimento de antagonistas mais potentes, seletivos e menos tóxicos. Para esta etapa foram utilizando métodos computacionais de rastreio virtual baseados em ligandos (LBVS) entre as quais, técnicas de aprendizagem de máquina – numa tentativa de compreender como as características físico-químicas de inibidores já conhecidos contribuem para a definição da sua atividade. Por fim, numa terceira etapa, foram analisadas as subestruturas máximas comuns entre os inibidores das duas isoenzimas. Estas subestruturas foram posteriormente utilizadas para a pesquisa de novos possíveis inibidores da 5α.redutase, e a sua atividade foi prevista utilizando os métodos desenvolvidos na etapa anterior.

Na primeira etapa do trabalho, com a utilização da técnica de modelação de proteínas por homologia não foi possível obter informações relevantes sobre a possível estrutura da 5α-redutase, com os dados estruturais atualmente disponíveis.

Na segunda etapa do trabalho, os resultados obtidos pelos métodos de aprendizagem de máquina mostraram que as características físico-químicas (descritores moleculares) mais importantes para a construção de inibidores da 5α-R2 parecem ser as propriedades aromáticas dos compostos enquanto que para a 5α-R1 surgem propriedades como o LogP, ASA+, volume e área de superfície. A análise das subestruturas máximas comuns obtidas nas diferentes isoenzimas demonstrou que os esteroides 4-azasteroides e 6-azasteroides são bons inibidores de ambas as isoenzimas.

Palavras-Chave: 5α-redutase, aprendizagem de máquina, QSAR, PCA, seleção de atributos, árvores de decisão, SVM, MCS.

Abstract

The 5α-reductase is a microsomal protein that converts testosterone into dihydrotestosterone (DHT), a more potent androgen with the function of inducing differentiation during fetal development that leads to the development of the male external genitalia. When disturbances occur in the performance of this enzyme may arise disorders such as pseudohermaphroditism, baldness, benign hyperplasia and prostate cancer, among others. Prostate cancer and benign prostatic hyperplasia are diseases that have been increasing in recent years. There are only two marketed drugs, finasteride and dutasteride, which have side effects, which therefore stresses the need for the development of better inhibitors for this enzyme.

This work presented here was developed in three major steps. In a first step, an attempt was made to obtain relevant information about the possible structure of the enzyme 5α-reductase, using homology protein modeling techniques. In a second step, we tried to obtain information about activity the inhibitors of this enzyme, in order to contribute to the development of more powerful, selective and less toxic antagonists. In this step, we applied ligand based computational methods, the virtual screening (LBVS) computational methods – among which machine learning techniques – in an attempt to understand how the physico-chemical characteristics of the known inhibitors contribute to their activity. Finally, in a third step, the maximum common substructures among the inhibitors of both enzymes were analyzed. These substructures we then used to search to search in chemical databases for potential new inhibitors of 5α-reductase, and their activity predicted using the methods developed in the previous step.

In the first step of our work, using homology modeling techniques, it was not possible to obtain relevant information about the possible structure of 5α-reductase with structural data currently available.

In the second step of our work, the results obtained by the machine learning methods showed that the physic-chemical characteristics (molecular descriptors) most important for the construction of inhibitors of 5α-R2 seem to be the aromatic properties of the compound, while for the 5α-R1 the results suggest that the most important properties as LogP, ASA +, volume and surface area. The analysis of the maximum common substructure obtained with the analysis of compounds for the two isoenzymes showed that 4-azasteroids and 6-azasteroids are good inhibitors of both isoenzymes.

Key-words: 5α-reductase, machine learning, QSAR, PCA, Feature selection, Random forest, SVM, MCS.

A. Introdução

1. 5α-redutase

A enzima 5 alfa redutase (5α-redutase) é uma proteína microssomal que desempenha um papel fundamental na diferenciação sexual humana. A 5α-redutase catalisa a conversão da testosterona em di-hidrotestosterona (DHT) que é um androgénio mais potente e tem como função induzir a diferenciação durante o desenvolvimento do feto que conduz ao desenvolvimento dos genitais externos masculinos. Mutações no gene do esteroide 5α-redutase dão origem a uma forma de macho rara chamado de pseudo–hermafroditismo, onde os machos afetados desenvolvem o trato urogenital interno normal, mas não conseguem desenvolver as estruturas masculinas externas. A expressão do gene é regulada por androgénios em tecidos tais como a próstata e fígado. A 5α-redutase também tem um papel importante em vários distúrbios endócrinos, incluindo a hiperplasia benigna da próstata, calvície masculina, acne e hirsutismo. O papel central desempenhado pela 5α-redutase e pelo seu produto DHT nestas desordens tem resultado no desenvolvimento de inibidores desta enzima (Andersson & Russell 1990).

A hiperplasia benigna da próstata e o cancro da próstata são doenças que têm vindo a afetar cada vez mais a população mundial masculina no decorrer dos últimos anos. Atualmente, a finasterida e a dutasterida são os únicos inibidores da enzima 5α-redutase comercializados. Contudo, os efeitos colaterais provocados por estes inibidores levam à necessidade do desenvolvimento de novos inibidores para estas enzimas que sejam mais potentes e seletivos. No processo do desenvolvimento destes inibidores, os investigadores depararam-se com o problema da falta do conhecimento da estrutura tridimensional da enzima 5α-redutase, consequência da sua instabilidade durante o processo de purificação.

1.1. Biologia e propriedades bioquímicas da enzima 5α-redutase

O mecanismo de ação das hormonas esteroides inicia-se quando estas migram da corrente sanguínea para o interior da célula através da membrana celular através de um mecanismo de difusão. Uma vez dentro da célula, os esteroides formam complexos através da ligação com proteínas, chamadas de recetores (Van Eekelen 1996), o que desencadeia uma cascata de eventos que permitem a expressão de genes específicos (Aggarwal et al., 2010). Dentro da célula, as hormonas esteroides sofrem modificações na sua estrutura,

produzindo moléculas metabolicamente ativas, capazes de desenvolver uma cascata de reações. Estas mudanças moleculares são causadas pela presença de enzimas específicas em cada tecido. Na próstata, estes esteroides têm um papel importante na virilização masculina, na diferenciação e função sexual, anabolismo e comportamento. A glândula prostática depende fundamentalmente destas hormonas para o seu desenvolvimento, maturação e funcionamento. Contudo, a testosterona não é o androgénio primário responsável por esses eventos, mas um metabolito mais potente resultante da reação catalisada pela 5α-redutase, a DHT (Platz & Giovannucci 2004).

A superprodução de DHT pode levar ao desenvolvimento de hiperplasia benigna da próstata e a de cancro da próstata. A DHT é formada na próstata a partir da testosterona por ação da isoenzima 5α-redutase 2 (5α-R2). Sendo o androgénio mais potente no homem, a DHT liga-se ao recetor dos androgénios com alta afinidade (K_d= 10^{-11} M) e modula a expressão do gene (Rizner 2003). Dentro da próstata a DHT pode ser inativada pela desidrogenase 3α-hidroxiesteróide (3α-HSD) a partir da 3α-androstanodiol (3α-diol) ou pela 3β-HSD a partir da 3β-androstanodiol (3β-diol). Destas duas vias, a que envolve a 3α-HSD tem um papel dominante na redução da DHT. A posterior inativação da 3α-diol e da 3β-diol é alcançada através da glucuronidação, seguida de excreção (Fig. 1) (Rizner 2003).

Os mecanismos que podem levar ao excesso de DHT na próstata incluem: (1) aumento da síntese de DHT devido à elevada expressão da 5α-R2, (2) expressão elevada das isoformas oxidativas da 3α-HSD que convertem 3α-diol em DHT, e (3) diminuição da inativação de DHT devido à regulação negativa das 3-cetoesteroides redutases (Rizner 2003).

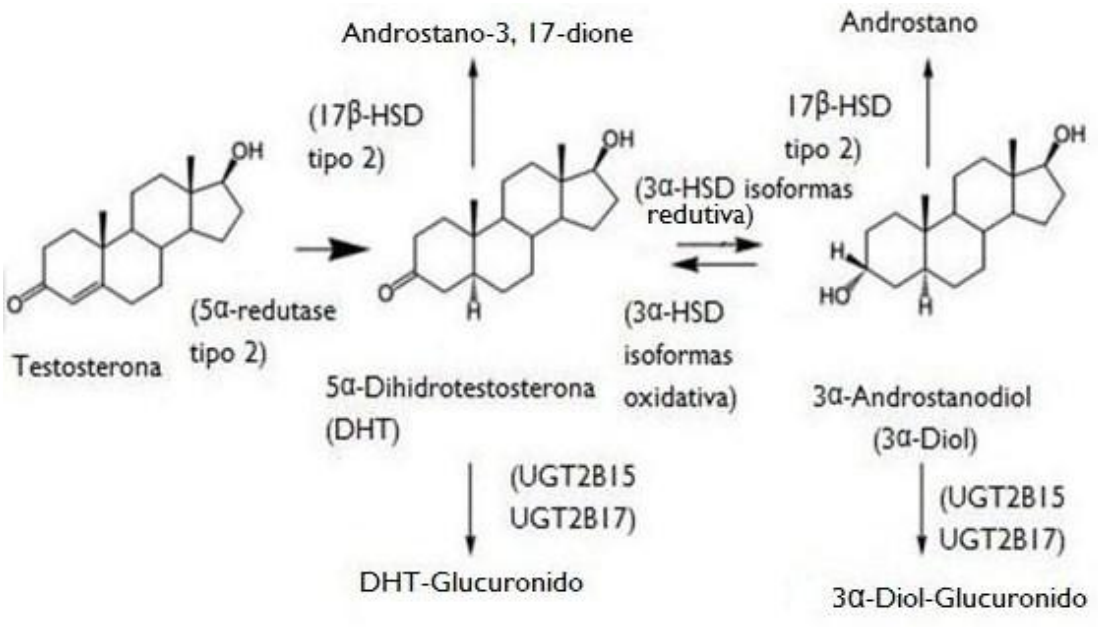


Figura 1. Metabolismo dos androgénios na próstata humana. (Adaptado de Rizner 2003).

1.2. Mecanismo de ação da enzima 5α-redutase β

A 5α-redutase é uma proteína microssomal que reduz estereoseletivamente a dupla ligação $\Delta^{4,5}$ nos carbonos C_{19} e C_{21} da testosterona usando a nicotinamida adenina dinucleótido fosfato (NADPH) como co-fator obtendo-se o correspondente 5α-3-oxosteroide DHT (Fig. 2) (Salvador *et al.* 2013). A catálise enzimática é iniciada pela formação de um complexo binário entre a enzima e o NADPH, seguida de um complexo ternário com o substrato. Um carbocatião deslocalizado é formado devido à ativação do sistema enona por uma forte interação com um resíduo eletrofilico (E^+) presente no local ativo.

O enolato de DHT é formado pela transferência direta do hidreto a partir do NADPH para a face α do carbocatião deslocalizado o que leva à redução seletiva em C_5. Este enolato, que é coordenado com o $NADP^+$ na face α, é atacado por um protão na face β em C_4 dando origem ao complexo ternário enzima-$NADP^+$-DHT. O complexo binário $NADP^+$-enzima é formado depois da saída de DHT e finalmente o $NADP^+$ deixa a enzima livre para novos ciclos catalíticos (Fig. 3) (Aggarwal et al. 2010).

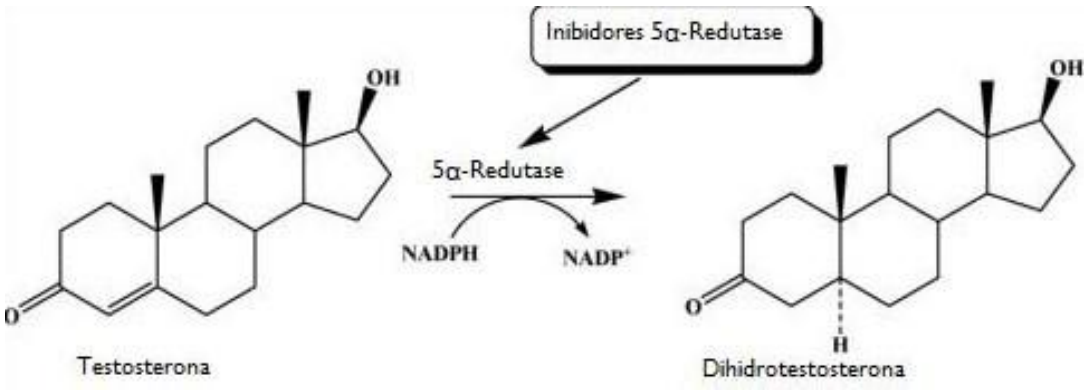


Figura 2. Local de ativação dos inibidores da 5α-redutase. (Adaptado de Aggarwal et al., 2010).

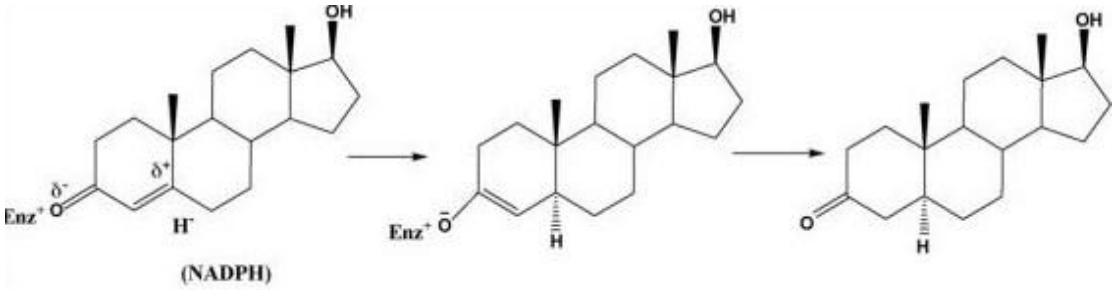


Figura 3. Mecanismo de redução da testosterona. (Adaptado de Aggarwal et al., 2010).

1.3. Isoformas da enzima 5α-redutase

Duas isoformas da enzima 5 α-redutase foram identificadas nos humanos, ratos, camundongos e macacos, sendo conhecidas como a isoenzima tipo 1 (5α-R1) e a isoenzima tipo 2 (5α-R2), nomenclatura baseada na ordem cronológica em que foram descobertas (Iehlé *et al.* 1999). Apesar de ambas as isoenzimas catalisarem a mesma reação, estas apresentam apenas uma baixa homologia na sequência peptídica, estão localizadas em diferentes cromossomas e cada uma possui propriedades bioquímicas distintas. Estas proteínas são compostas por 254 e 260 aminoácidos e têm com peso molecular estimado de 28 e 29 kDa, respetivamente. Estas enzimas possuem cerca de 44% de aminoácidos hidrofóbicos distribuídos ao longo de toda a cadeia polipeptídica da molécula, sugerindo que as mesmas são proteínas intrínsecas encontrando-se profundamente embebidas na bicamada lipídica. A média de homologia entre as sequências de aminoácidos nas isoenzimas dentro da mesma espécie é de aproximadamente 47%, enquanto que entre as mesmas isoenzimas de espécies diferentes é de 60% para a 5α-R1 e 77% para a 5α-R2 (Jin & Penning 2001).

Em estudos com lisados de células transfectadas, a 5α-R1 tem um pH ótimo para a sua reação que se estende até à faixa alcalina (pH 6-8,5), enquanto que a 5α-R2 tem um pH ótimo para a reação, restrito a um meio ácido (pH 5) (Poletti et al., 1997). A base bioquímica do pH ótimo ácido para a 5α-R2 não é completamente compreendida. Estudos com células transfectadas do ovário de hamster chinês (CHO) sugerem que a 5α-R2 pode efetivamente ter um pH ótimo neutro no seu estado nativo. Além disso, estudos imunocitoquímicos mostraram que ambas as isoenzimas 5α-R1 e 5α-R2 estão presentes no retículo endoplasmático das células CHO, um compartimento com pH neutro. Adicionalmente, análises utilizando células lisadas, células permeabilizadas e células intactas sugerem que a afinidade da 5α-R2 é mais elevada a pH neutro do que a pH ácido (pH 5), sugerindo assim que esta isoenzima atua a pH neutro na célula. Porém, com a lise física da membrana plasmática, a isoenzima passa a apresentar atividade somente a pH 5. A mudança no pH requerido pode refletir uma alteração conformacional da isoenzima, já que o pH ótimo de uma enzima é, normalmente, consequência do estado iónico dos aminoácidos carregados no local ativo (Imperato-McGinley & Zhu 2002).

O gene que codifica a 5α-R1 está localizado no cromossoma 5, enquanto que o gene que codifica a 5α-R2 encontra-se no cromossoma 2. Múltiplos polimorfismos foram identificados em ambos os genes, mas as suas consequências fisiológicas não são conhecidas. Nos humanos, ambas as isoenzimas são expressas no fígado, sem nenhuma diferença significativa entre os sexos. No entanto, encontram-se distribuídas em diferentes concentrações em diferentes órgãos (Fig. 4) (Steers 2001). A 5α-R1 encontra-se predominante nos folículos capilares e glândulas subcutâneas da pele, enquanto que a 5α-R2 encontra-se maioritariamente na próstata, genitais, vesículas seminais e epidídimo (Faragalla et al., 2003). Estudos apontam para uma distribuição estromal e epitelial diferente para as duas isoenzimas. Embora haja uma maior expressão da 5α-R2 na próstata normal bem como na hiperplasia benigna da próstata (HBP), no adenocarcinoma de próstata a expressão da 5α-R2 está diminuída, enquanto que a 5α-R1 pode estar aumentada, ao contrário do que ocorre quando não há cancro da próstata. De facto, tanto a 5α-R1 como a 5α-R2 estão presentes no estroma, embora com maior predominância da 5α-R2. No epitélio apenas há expressão da 5α-R1. O conhecimento de que a atividade da 5α-R regula o crescimento da próstata resultou no desenvolvimento de drogas, como a finasterida, um inibidor da 5α-R2 (Steers 2001).

Recentemente, uma terceira isoenzima da 5α-redutase foi identificada e designada como 5α-redutase 3 (5α-R3). Esta isoenzima foi originalmente identificada no tecido de cancro da próstata e atua como uma redutase poliprenol na glicosilação de proteínas, podendo também ser encontrada no pâncreas e cérebro (Kapp et al., 2012).

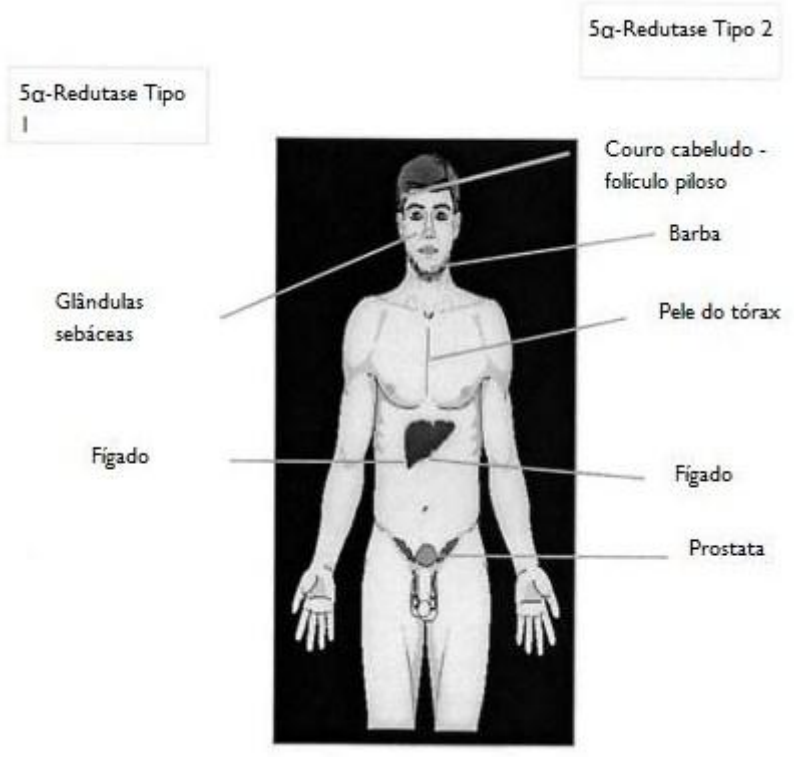


Figura 4. Órgãos do corpo humano onde se encontram maior concentração das enzimas 5α-R1 e 5α-R2. (Adaptado de Steers 2001).

1.4. Tipos de inibidores da enzima 5α-redutase

Três tipos de inibidores podem ser concebidos de acordo com a cinética do mecanismo de redução da testosterona: (i) os inibidores do tipo A que competem com o co-fator (NADPH) e o substrato (T) e interagem com a enzima livre; (ii) os inibidores do tipo B que competem com o substrato e encaixam no complexo NADPH-enzima; e (iii) os inibidores do tipo C que encaixam no complexo NADPH-enzima e não são competitivos com o substrato (Aggarwal et al., 2010).

Um fator importante que afeta a descoberta de novos inibidores é a ausência de informação sobre a estrutura do local ativo das duas isoenzimas da 5α-redutase. A estrutura tridimensional desta enzima ainda não é conhecida devido à sua instabilidade durante a

purificação, sendo que a única informação disponível é a sequência primária, estimada a partir do c-DNA (Aggarwal et al., 2010). Assim, ao longo dos anos, a maioria dos inibidores têm sido desenvolvidos considerando a estrutura do substrato (testosterona), o mecanismo de reação proposto para este processo enzimático e o conhecimento obtido a partir de estudos de relação quantitativa estrutura-atividade (QSAR, do Inglês *Quantitative Structure-Activity Relationship)* (Aggarwal et al., 2010).

Os primeiros inibidores têm sido desenvolvidos pela modificação da estrutura de substratos naturais, incluindo a substituição de um átomo de carbono do anel A ou B dos esteroides por um heteroátomo, o que levou à descoberta de potentes inibidores da 5α-redutase, tais como os 4-azasteróides (entre os quais a finasterida, comercializado para o tratamento da hiperplasia benigna da próstata), o 6-azasteroide, o 10-azasteroide e inibidores de ácidos carboxílicos esteróidais (Occhiato et al., 2004). Os inibidores da 5α-redutase mais importantes são os compostos esteroides. No entanto, vários inibidores não esteroides também foram desenvolvidos, imitando os inibidores esteroides. Os inibidores esteroidais podem ser estruturalmente divididos em três tipos principais: azasteroides, ácidos 3-carboxilicos e derivados de pregnano / androstano.

O primeiro e mais conhecido inibidor da 5α-redutase é o 4-aza-esteroide (finasterida) (Fig. 5) que inibe seletivamente a 5α-R2. Esta droga é um inativador baseado no mecanismo de reação, bloqueando a conversão da testosterona em DHT para reduzir a estimulação da próstata. Estudos clínicos de longo prazo com este fármaco em doentes com hiperplasia benigna da próstata demonstram uma redução sustentável do androgénio específico da próstata. No entanto, os efeitos secundários da finasterida e a necessidade de desenvolver novos inibidores da 5α-redutase mais potentes e seletivos bem como os estudos de QSAR têm levado os investigadores a procurar novos inibidores azasteroides baseados na estrutura da finasterida (Salvador *et al.* 2013).

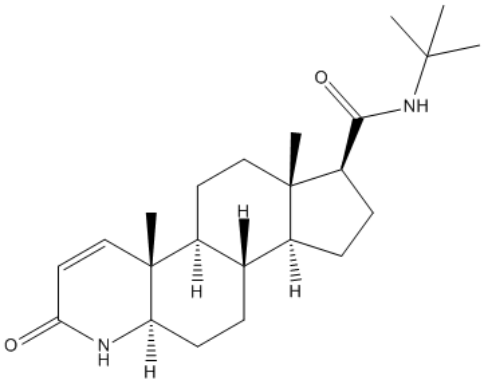


Figura 5. Estrutura do composto finasterida, um inibidor seletivo da 5α-redutase 2.

Uma melhoria significativa dos inibidores da 5α-redutase resultou do desenvolvimento de inibidores duplos, ou seja, inibidores de ambas as isoenzimas, o que significa que um bloqueio mais abrangente da síntese de DHT pode ser alcançado. A partir destes estudos, surgiu a dutasterida (Fig. 6) como um potente inibidor duplo, que permitiu reduzir os níveis de DHT em mais de 95% e melhorou os resultados clínicos de doentes com hiperplasia benigna da próstata. Além disso, a dutasterida é bem tolerado e tem apenas efeitos colaterais transientes (Amory *et al.* 2008). Por estas razões, a FDA aprovou a sua utilização no tratamento da hiperplasia benigna da próstata. De facto, como a única diferença entre a finasterida e a dutasterida está na cadeia lateral, na ligação C_{17}, uma série de modificações nesta parte da estrutura têm sido exploradas.

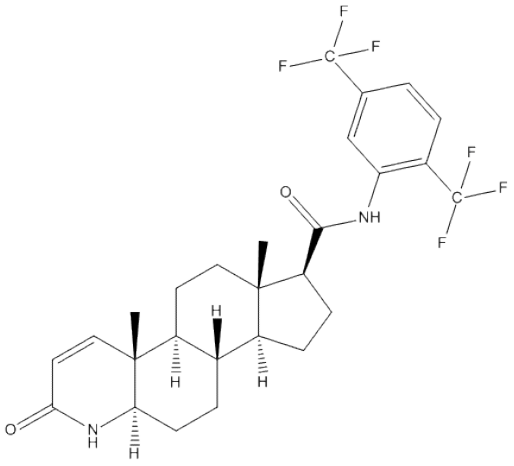


Figura 6. Estrutura da dutasterida, um potente inibidor duplo da 5α-redutase.

A enzima 5α-redutase também tem sido considerada como um importante alvo farmacológico no tratamento do cancro da próstata. De facto, níveis plasmáticos elevados do androgénio DHT têm sido associados não só à hiperplasia benigna da próstata, mas também ao cancro da próstata. Devido à sua capacidade de diminuir o volume da próstata, a finasterida e a dutasterida foram avaliadas para a prevenção e tratamento do cancro da próstata. Ambos se revelaram promissores na prevenção do cancro da próstata em homens com risco de desenvolver a doença. No entanto, o uso destes compostos na terapia do cancro da próstata ainda permanece controverso. De facto, observou-se que embora a incidência de cancro tenha sido reduzida pelo tratamento com estes inibidores da 5α-redutase, os cancros que foram detetados eram mais agressivos do que os cancros detetados em pacientes tratados com placebos. Por esta razão, é necessário a realização de mais estudos para o desenvolvimento de novos inibidores mais potentes e eficazes para a 5α-redutase (Amory *et al.* 2008; Salvador *et al.* 2013).

2. Métodos computacionais de rastreio virtual

O rastreio virtual é utilizado na descoberta de fármacos e baseia-se na pesquisa e análise de compostos disponíveis em bases de dados, tentando encontrar novos compostos e novos quimiotipos, com a atividade biológica requerida, como alternativa aos compostos já conhecidos (Lavecchia & Di Giovanni 2013). Pode ser geralmente descrito como um método que envolve a filtragem computacional de uma grande quantidade de moléculas para identificar as que têm uma maior probabilidade de exibir a atividade pretendida no sistema biológico de interesse. Um método de rastreio virtual começa com todas as moléculas que podem ser adquiridas (ou sintetizadas) e testadas, e em seguida seleciona as poucas que devem ser testadas experimentalmente (Chen *et al.* 2007).

Os métodos de rastreio virtual podem ser divididos em duas grandes categorias: métodos baseados em ligandos (LBVS, do inglês *ligand based virtual screening*) e métodos baseados na estrutura (SBVS, do inglês *structure based virtual screening*) (Lavecchia & Di Giovanni 2013).

2.1. Métodos baseados na estrutura

Os métodos de rastreio virtual baseados na estrutura (SBVS) requerem a disponibilidade de uma estrutura tridimensional do alvo biológico de interesse, determinada experimentalmente, através de cristalografia de raio-x ou ressonância magnética nuclear (RMN), ou computacionalmente através de modelação molecular. Esta informação permite o uso de métodos baseados no desenho ou no acoplamento molecular da proteína-ligando. Estes métodos têm-se tornado amplamente utilizados com o crescimento da disponibilidade dos dados tridimensionais de proteínas com interesse terapêutico (Chen *et al.* 2007).

A construção de modelos de farmacóforo baseados no recetor e o acoplamento molecular são duas das técnicas mais aplicadas de SBVS (Lyne 2002). Um farmacóforo define as características e os locais importantes de interação das ligações entre um ligando e o recetor (Valasani et al., 2013). Os modelos de farmacóforos derivados do mapeamento dos recetores desempenham um papel importante na construção de bibliotecas de compostos orientados para determinado alvo. Devido ao número crescente de estruturas resolvidas e armazenadas nas bases de dados, e devido ao desenvolvimento de métodos capazes de explorar e mapear possíveis locais de ligação na estrutura da proteína, a construção de modelos de farmacóforos baseados no recetor tem aumentado consideravelmente (Guido et al., 2008). O acoplamento molecular (Fig. 7) pode ser utilizado para identificar moléculas que se liguem ao(s) local(is) ativo(s) de uma proteína alvo (recetor) cuja estrutura é conhecida. Esta técnica envolve a previsão da conformação do ligando e sua orientação no local ativo da molécula alvo, seguindo-se uma avaliação da força ou afinidade da ligação entre as duas moléculas com base numa função de pontuação. Esta informação é então usada para ordenar e/ou classificar os compostos com o objetivo de selecionar e testar experimentalmente um pequeno subconjunto que se espera que apresentem com a atividade biológica pretendida (Lyne 2002).

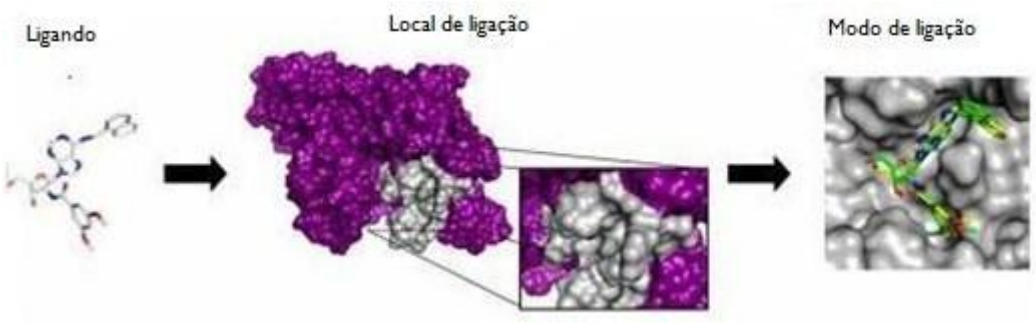


Figura 7. Procedimento geral do acoplamento molecular. (Adaptado de Guido *et al.* 2008).

Uma vez que estes métodos estão orientados pela e para a extração de conhecimento, estes estão fortemente dependente da quantidade e qualidade dos dados disponíveis sobre o sistema biológico em estudo. De facto, independentemente do tipo de proteína de interesse serão sempre colocadas questões importantes tais como a seleção da geometria mais relevante, a flexibilidade do recetor, a atribuição adequada de estados de protonação e análise de moléculas de água no local de ligação, entre outros. Outro passo importante para os métodos baseados na estrutura é a conceção apropriada das bibliotecas de pequenas moléculas candidatas ao processo de rastreio (Guido *et al.* 2008). Adicionalmente, a utilização da técnica de acoplamento molecular envolve ainda a seleção da ferramenta apropriada ao sistema em estudo, que apresentam ainda grandes limitações nas funções de pontuação usadas para prever energias de ligação.

2.2. Métodos baseados nos ligandos

O conceito das estratégias utilizadas nos métodos baseados em ligandos é utilizar dados da estrutura-atividade de um conjunto de compostos ativos que se sabe que ligam ao alvo desejado e usar esses dados para identificar outros compostos candidatos com propriedades semelhantes nas bases de dados para a avaliação experimental. Os métodos baseados em ligandos podem ser divididos em três grandes classes:

(1) Similaridade e pesquisa de subestruturas, incluindo impressões digitais (no inglês, *fingerprints*), coeficientes de Tanimato, Coseno, Hamming, Russel-Rao, e Forbes. Estes métodos atuam sob a premissa que moléculas estruturalmente relacionadas são suscetíveis de apresentar propriedades semelhantes, em particular, exibir a mesma atividade.

(2) Modelos de farmacóforo baseado em ligandos, podem ser definidos como arranjo tridimensional de características moleculares ou de fragmentos existentes em compostos ativos e que são necessários (mas não necessariamente suficientes), para que um composto seja capaz de estabelecer a ligação. Estes métodos envolvem a identificação do padrão farmacofórico comum a um conjunto de compostos ativos conhecidos, bem como a utilização desse padrão na pesquisa de subestruturas tridimensionais.

(3) Os métodos de aprendizagem de máquina (no inglês, *machine learning*) permitem a construção de regras de classificação a partir de um conjunto de treino contendo, por exemplo, moléculas ativas e inativas conhecidas. As técnicas de aprendizagem de máquina são utilizadas na previsão de propriedades químicas com base nas informações da estrutura molecular e em estudos de relações quantitativas de estrutura-atividade (QSAR) (Guido *et al.* 2008; Lavecchia & Di Giovanni 2013).

2.2.1. Modelos de farmacóforo baseado em ligandos

A identificação do modo de ligação entre uma proteína e um composto químico é fundamental para otimizar compostos candidatos a fármacos para o rastreio virtual de bibliotecas com milhões de compostos. Os modelos de farmacóforos são construídos a partir de grupos de átomos funcionais e essenciais, na posição tridimensional adequada, para interagir com um determinado recetor. Por este motivo um farmacóforo pode ser um modelo importante para o rastreio virtual, especialmente se a estrutura tridimensional da proteína não é conhecida e assim a técnica de acoplamento molecular não é aplicável. A grande vantagem do rastreio baseado no farmacóforo comparado com outras técnicas de rastreio baseadas em coeficientes de similaridade reside na capacidade de identificar um conjunto diversificado de potenciais compostos ativos com um suporte químico totalmente diferente, aumentando a possibilidade de alguns dos compostos identificados passarem todas as fases do desenvolvimento de drogas.

Tradicionalmente, os modelos de farmacóforos baseados em ligandos são calculados através da extração de características comuns entre as estruturas tridimensionais dos compostos que são conhecidos por interagir com a proteína de interesse (Khan et al., 2012; Lavecchia & Di Giovanni, 2013).

2.2.2. Relação quantitativa estrutura-atividade

Os estudos de relação quantitativa estrutura-atividade (QSAR) tem como finalidade gerar modelos capazes de correlacionar as propriedades estruturais e/ou físico-químicas dos compostos e a resposta biológica de interesse. O conhecimento das propriedades dos compostos, tais como a afinidade de ligação (Kd) ou a concentração inibitória (IC50), é necessário para aplicar os métodos de QSAR. As estruturas moleculares são normalmente representadas por conjuntos específicos de propriedades estruturais e físico-químicas (descritores moleculares) das moléculas consideradas mais relevantes para a atividade de ligação. As técnicas de QSAR partem do cálculo de descritores baseados nas estruturas moleculares e utiliza algoritmos computacionais para relacionar os descritores-chave com o valor de interesse da propriedade dependente. Estes estudos em Química Medicinal visam elucidar os aspetos fundamentais das relações entre os descritores estruturais ou propriedades e a atividade biológica, além de que são importantes para compreender a atividade de interesse e podem permitir a previsão da propriedade biológica de novos compostos (Guido *et al.* 2008).

O termo QSAR tem vindo a sofrer uma grande evolução desde as primeiras experiências em que foram utilizados modelos lineares com um pequeno número de descritores até a aplicação de técnicas sofisticadas não lineares de aprendizagem de máquina (*machine learning*) (Mitchell 2011).

A principal premissa que fundamenta os estudos de QSAR é que as interações dos fármacos com os seus alvos biológicos são determinadas por forças intermoleculares, ou seja, interações hidrofóbicas, polares, eletrostáticas e estéreas. Assim, os fármacos que exercem os seus efeitos biológicos interagem com um alvo específico (uma enzima, um recetor, um canal iónico, um ácido nucleico ou qualquer outra molécula biológica) e devem ter uma estrutura tridimensional tal que os grupos funcionais e as propriedades de superfície sejam mais ou menos complementares ao local de ligação. Desde que o sistema biológico seja mantido constante, a interação de dois fármacos diferentes com o local de ligação bem como as suas distribuições no sistema dependem apenas das estruturas químicas dos compostos. Se estas estruturas são intimamente relacionadas, as diferenças nas suas propriedades físico-químicas e assim as diferenças nas forças de interação podem ser facilmente descritas de uma maneira quantitativa, isto é, a variação nessas propriedades biológicas deve estar diretamente relacionada com as variações nessas propriedades. Assim,

todos os modelos quantitativos de relação estrutura-atividade são baseados na suposição de uma aditividade das contribuições de grupos aos valores da atividade biológica (Kubinyi, 1993).

Para que um modelo de QSAR seja desenvolvido com sucesso, vários aspetos devem ser considerados. Um dos mais importantes é a relação dos descritores moleculares que devem ser eficientes na descrição de parâmetros que sejam de facto relevantes para a atividade em estudo. Além disso, a realização cuidadosa da análise estatística, assim como a aplicação de testes de validação dos modelos são imprescindíveis para assegurar a confiança e a utilidade do modelo.

As técnicas de aprendizagem de máquina são poderosas para construir e otimizar os modelos preditivos. A aprendizagem de máquina é um ramo de inteligência artificial. Estes métodos assumem como entrada um conjunto de objetos que tenham sido previamente classificados em duas ou mais classes no contexto do rastreio de compostos, ou seja, um conjunto de moléculas que tenham sido previamente testados e demonstrado serem ativos ou inativos, por exemplo. Estas moléculas de treino são analisadas para desenvolver uma regra de decisão, que pode ser utilizada para classificar as novas moléculas (conjunto de teste) numa de duas ou mais classes (Chen et al., 2007; Cheng et al., 2012). De facto, a classificação não é restrita a escolhas binárias; a classificação também pode ser numérica, como a previsão dos valores de IC50. O termo estatístico regressão é utilizado para este tipo de problemas de aprendizagem.

Tal como já foi referido, para qualquer classificação ou regressão, a aprendizagem de máquina começa com um algoritmo e com exemplos de moléculas para as quais se conhecem a classe a que pertencem (por exemplo, ativa ou inativa) ou o valor que se pretende prever (por exemplo, IC50). Decorre então o processo de "treino" do algoritmo a partir dos exemplos conhecidos. Problemas de aprendizagem que exigem que os dados sejam pré-atribuídos para um determinado número de classes são chamados de supervisionados. Alternativamente, pode haver situações em que a pré-atribuição das classes não é possível ou desejável. Um exemplo seria a seleção de um pequeno número de moléculas a partir de uma base de dados maior, tentando-se de manter as propriedades globais de todo o conjunto. Cada molécula escolhida pode ser pensada como um representante de um certo número de outras moléculas num conjunto maior. A partição de um conjunto de dados desta forma, implica a existência de diferentes classes de moléculas, que não é conhecida *a priori*, e o significado físico das classes é improvável que seja tão

evidente como a atividade. Tais tarefas de aprendizagem são conhecidas como "não supervisionadas" (Melville et al., 2009).

Alguns exemplos de métodos de aprendizagem de máquina são redes neurais e máquinas de vetores de suporte, que permitem desenvolver modelos estatísticos capazes de prever atividades ou propriedades físico-químicas de estruturas, com um grau de diversidade química maior do que é possível utilizar com os métodos obtidos por QSAR clássicos.

2.2.3. Similaridade e pesquisa de subestruturas

Uma das técnicas mais simples e mais utilizada no rastreio virtual é a pesquisa por similaridade, em que uma estrutura bioativa de referência é procurada contra uma base de dados para identificar as moléculas vizinhas mais próximas, uma vez que estas são as mais propensas a apresentar bioatividade de interesse. Métodos comuns de pesquisa de similaridade estrutural incluem índices topológicos, impressões digitais moleculares (*fingerprints*), fragmentos, entre outros (Hert *et al.* 2004).

A pesquisa de relação de subestrutura e superestrutura encontram-se também entre as medidas de similaridade mais frequentes. Dadas duas estruturas químicas A e B se a estrutura A está totalmente contida na estrutura B, então A é uma subestrutura de B e B é uma superestrutura de A. Assim, de acordo com o princípio da similaridade, A e B podem partilhar propriedades que são relacionadas com a sua subestrutura comum (Fig. 8). Portanto, uma subestrutura que é supostamente associada com determinadas propriedades de interesse pode ser usada como consulta em base de dados para identificar todos os compostos que partilham esta subestrutura (ou superestrutura) e, eventualmente, as suas atividades (Cao et al., 2008).

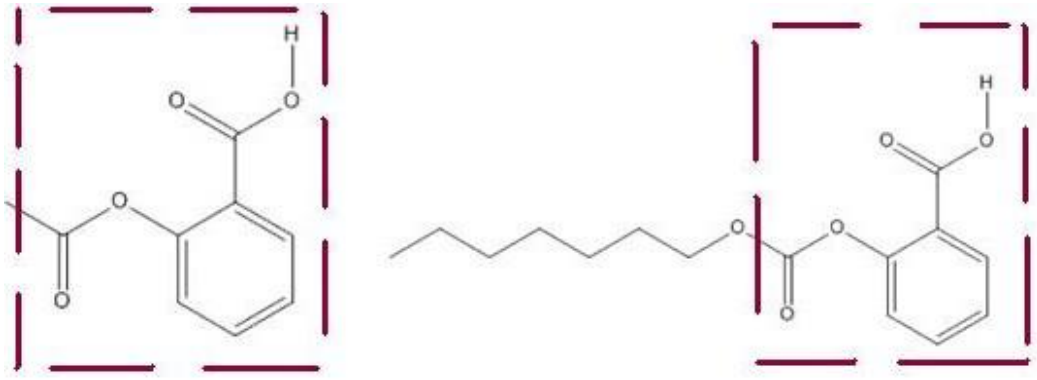


Figura 8. Local de similaridade entre os compostos. As duas estruturas partilham uma subestrutura comum (caixas a tracejado).

A subestrutura máxima comum (MCS) funciona como uma métrica na pesquisa de propriedades químicas semelhantes e as suas previsões de atividade. A MCS entre dois compostos é a maior subestrutura que aparece em ambas as estruturas. A utilização de MCS é vantajosa para medir a semelhança química de estruturas. Primeiro, porque é intuitivo, a maior subestrutura comum encontrada nas estruturas dos compostos é provavelmente um componente importante da sua atividade. Segundo, porque a semelhança das estruturas pode ser visualizado destacando o subgrafo máximo comum entre duas estruturas químicas. A MCS tem um bom desempenho na identificação de similaridade local em comparação com os métodos de descritores estruturais (Cao et al., 2008).

3. Modelação molecular por homologia estrutural

O processo de desenvolvimento racional de novos fármacos é colocado à prova sempre que se deseja descobrir um ligando para o qual a estrutura do seu alvo proteico ainda não foi determinada experimentalmente (Marshall 2004). Nestes casos, modelos computacionais do alvo terapêutico, em regra proteínas, podem ser elaborados por comparação da sua sequência de aminoácidos com as sequências de aminoácidos de proteínas homólogas que possuem estruturas tridimensionais experimentalmente resolvidas e que servem como molde (do inglês, *Template*). Este procedimento comparativo para a construção de modelos estruturais é conhecido como modelação molecular por homologia estrutural ou modelação comparativa (Liu te al., 2011).

Este método baseia-se no conceito de evolução molecular que tem como princípio a semelhança entre as estruturas primárias a proteína alvo que se pretende modelas e a proteína que serve de molde. Tem sido demonstrado que ao longo do processo evolutivo, a estrutura tridimensional de proteínas homólogas é muito mais conservada que a correspondente sequência primária, pelo que se demonstra que quando a identidade sequencial entre o alvo e o molde é elevado (> 70%), a modelação por homologia tem uma elevada probabilidade de sucesso.

A conservação da estrutura 3D entre as duas proteínas com sequências semelhantes permite a previsão da estrutura do alvo, utilizando as características estruturais do modelo. Assim sendo, os procedimentos de modelação de proteínas por homologia podem ser divididos em quatro passos sucessivos (Fig. 9):

1) Identificação e seleção do molde
2) Alinhamento de sequências
3) Construção das coordenadas do modelo
4) Validação

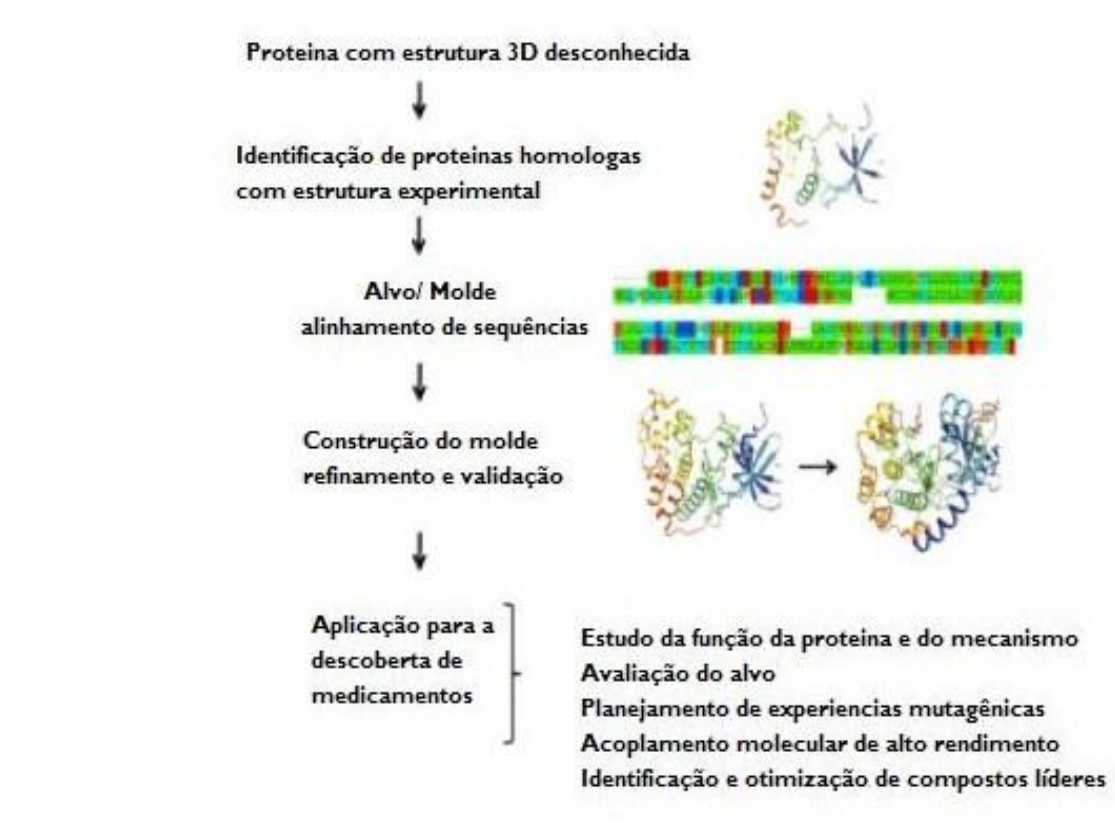


Figura 9. Esquema do processo da modelação de proteínas por homologia e as suas aplicações na descoberta de fármacos. (Adaptado de Cavasotto & Phatak 2009).

Para a execução de cada um destes passos existe um grande número de programas e servidores específicos disponíveis na internet (Liu *et al.* 2011). O primeiro passo na modelação de proteínas por homologia é a identificação e seleção de estruturas tridimensionais resolvidas, que possam atuar como uma base estrutural (molde) para a modelação do alvo. Uma das formas mais eficientes para realizar essa identificação é através da similaridade entre as sequências de aminoácidos das proteínas. Esta etapa tem como objetivo promover a identificação de proteínas com estruturas tridimensionais conhecidas que se correlacionam com o alvo e que possam funcionar como molde (no inglês, *template*) para a modelação do alvo com estrutura tridimensional desconhecida. Para esta identificação são utilizadas bases de dados como o GenBank(1), SwissProt(2), PDB(3), entre outras (Chi & City 2012).

(1) (http://www.ncbi.nlm.nih.gov/genbank/)
(2) (http://www.uniprot.org/)
(3) (http://www.rcsb.org)

Na segunda etapa, procede-se ao alinhamento das sequências entre a proteína alvo e a(s) proteinas(s) molde com o objetivo de identificar uma correlação ótima entre os resíduos de aminoácidos de cada sequência. Quando só é identificado um molde, é realizado apenas um alinhamento simples com o modelo. Quando várias proteínas molde são identificadas realiza-se um alinhamento múltiplo, onde estarão dispostas todas as sequências molde selecionadas na pesquisa e a proteína alvo. Este tipo de alinhamento, o alinhamento múltiplo, é considerado mais fiável do que o alinhamento simples, uma vez que o alinhamento múltiplo permite detetar mais facilmente as características estruturais comuns de proteínas homólogas. Para se obter este alinhamento final entre as sequências podem ser utilizados servidores web como BLAST[4], Clustal[5], FastA[6], Jalview[7], entre outros (Chi & City 2012).

Depois do alinhamento final das sequências, a construção do modelo pode ser realizada pelo programa MODELLER[8], Swiss-Pdbviewer[9], Swiss-Model[10] entre outros. Após a construção do modelo gerado por estes programas é feita a avaliação do modelo utilizando programas computacionais como o Procheck[11] e o Verify3D[12] (Chi & City 2012).

(4) (http://blast.ncbi.nlm.nih.gov)
(5) (http://www.ebi.ac.uk/Tools/msa/clustalw2/)
(6) (http://www.ebi.ac.uk/Tools/sss/fasta/)
(7) (http://www.jalview.org/)
(8) (http://salilab.org/modeller/)
(9) (http://spdbv.vital-it.ch/)
(10) (http://swissmodel.expasy.org/)
(11) (http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/)
(12) (http://www.chem.ac.ru/Chemistry/Databases/Verify3D.en.html)

4. Objetivos

O trabalho desenvolvido e aqui apresentado assenta em três objetivos principais que em seguida se descrevem:

1. Modelação molecular da estrutura tridimensional da enzima 5α-redutase uma vez que esta ainda não se encontrar determinada experimentalmente, o que representa um grande obstáculo para a conceção de novos inibidores para esta enzima.
2. Aplicação de métodos computacionais de rastreio virtual, nomeadamente métodos baseados em ligandos (LBVS) para tentar obter informações relevantes entre características estruturais e físico-químicas de inibidores das isoenzimas da 5α-redutase e a sua atividade, de modo a contribuir para o desenvolvimento de antagonistas mais potentes, seletivos e menos tóxicos.
3. Com base na informação obtida, pesquisar em base de dados de compostos potenciais, novos inibidores da 5α-redutase prevendo a sua atividade.

B. Protocolo experimental

I. Modelação da estrutura da proteína 5α-redutase por homologia

1.1. Sequência primária da 5α-redutase

Para obter a sequência de aminoácidos da proteína 5α-redutase (Fig.10), utilizou-se a base de dados UniProt. UniProt é uma base de dados abrangente, de alta qualidade e de livre acesso que contém as sequências e informação de proteínas.

>sp|P31213|S5A2_HUMAN 3-oxo-5-alpha-steroid 4-dehydrogenase 2 OS=Homo sapiens GN=SRD5A2 PE=1 SV=1

MQVQCQQSPVLAGSATLVALGALALYVAKPSGYGKHTESLKPAATRLPARAAWFLQELPS

FAVPAGILARQPLSLFGPPGTVLLGLFCVHYFHRTFVYSLLNRGRPYPAILILRGTAFCT

GNGVLQGYYLIYCAEYPDGWYTDIRFSLGVFLFILGMGINIHSDYILRQLRKPGEISYRI

PQGGLFTYVSGANFLGEIIEWIGYALATWSLPALAFAFFSLCFLGLRAFHHHRFYLKMFE

DYPKSRKALIPFIF

Figura 10. Sequência de aminoácidos da proteína 5α-redutase 2.

1.2. Identificação de sequências molde

Para a identificação e seleção das sequências molde foi utilizado o programa BLAST (*Basic Local Alignment Search tool*). O BLAST é um algoritmo para comparar sequências biológicas primárias, tais como sequências de aminoácidos de proteínas ou nucleótidos de sequências de DNA, calculando a significância estatística dos resultados e identificando as regiões de similaridade local entre as sequências.

No PDB realizou-se uma pesquisa para identificar quais proteínas molde tinham estrutura cristalográfica. O PDB é uma base de dados de livre acesso onde estão depositados dados de estruturas tridimensionais de muitas moléculas biológicas.

No programa Jalview procedeu-se ao alinhamento múltiplo das sequências molde e a sequência da proteína 5α-redutase. O Jalview é um programa gratuito para alinhamento múltiplo das sequências e sua análise.

1.3. Construção de um modelo estrutural da 5α-redutase

Na tentativa de construção de uma estrutura tridimensional para o alvo em estudo foram utilizadas duas ferramentas: o Phyre e o Swiss-Model. O Phyre [13] (**P**rotein **H**omology/analog**Y** **R**ecognition **E**ngine) tenta prever a estrutura e função de uma proteína, utilizando algoritmos como o BLAST (Kelley & Sternberg 2009). Por outro lado, o Swiss-Model é um servidor automático na construção da estrutura tridimensional por homologia. Para um dado alvo são pesquisadas estruturas tridimensionais determinadas experimentalmente em bases de dados. Através do alinhamento das sequências entre o alvo e o modelo é gerado um modelo tridimensional para o alvo.

(13) (http://www.sbg.bio.ic.ac.uk/~phyre/)

II. Rastreio virtual baseado em ligandos

1. Descrição do conjunto de dados

1.1. Origem dos dados

Para cada isoenzima da 5α-redutase, realizou-se uma pesquisa dos compostos que inibem as duas isoenzimas, na ChEMBL [(14)]. A ChEMBL é uma base de dados que congrega valores de bioatividade (IC50, KI, entre outros) para milhões de compostos sobre milhares de alvos terapêuticos distintos (Willighagen & Waagmeester 2013). Para a isoenzima 1 (5α-R1) foram encontrados 538 compostos com 725 valores de bioatividade reportada. Quanto à isoenzima 2 (5α-R2) foram encontrados 641 compostos com 793 valores de bioatividade reportados (Tabela1).

Tabela 1. Valores de IC50 e Ki encontrados na ChEMBL (acedida em dezembro de 2012) e IC50 finais depois de uma análise cuidada dos dados dos compostos.

Proteína	**Bioatividade**			**Número de compostos**	**Compostos estudados**
	Total	IC50	Ki		
5α-R1	725	425	187	538	302
5α-R2	793	466	102	641	354

(14) (http://wwwchembl.pt)

Neste trabalho foram alvo de análise os compostos para os quais eram apresentados dados de IC50, para os quais se procedeu a uma análise mais pormenorizada. Para tal, foram retirados da ChEMBL os dados destes compostos e lidos os artigos onde são descritos os ensaios de bioatividade realizados e que na ChEMBL são disponibilizados com o identificador na PubMed ou DOI. Após uma primeira análise dos dados, foi possível observar que 9 dos compostos da 5α-R1 indicavam conter dados de IC50 que na realidade não eram reportados nos artigos referenciados. O mesmo aconteceu para a 5α-R2, podendo-se encontrar 18 compostos nestas condições.

Para a 5α-R1 e para a 5α-R2, verificou-se que 42 e 45 compostos, respetivamente, apresentavam mais do que um valor de IC50 e para alguns destes compostos os valores de IC50 variavam significativamente. Estas observações levaram a pensar que estes compostos teriam sido submetidos a mais de um ensaio pelo mesmo ou por diferentes autores, e por isso fez-se uma análise cuidada dos respetivos artigos. Desta revisão da literatura, verificou-se que para o mesmo composto, os métodos de ensaio realizados variavam mais para a 5α-R1 do que para a 5α-R2. Além disso, e com base nos dados dos artigos referenciados observou-se que para a 5α-R2 existem compostos que têm associados valores de IC50 de facto obtidos para o composto finasterida, ou seja os valores de IC50 do composto finasterida apareciam como sendo valores de IC50 do composto em estudo, sendo que dois dos compostos teriam apenas valores de IC50 do composto finasterida. Por outro lado, na 5α-R1, o mesmo composto apresentava valores de IC50 em unidades diferentes: μg.mL-1 e nM. Procedeu-se à conversão dos valores de μg.mL-1 para nM e verificou-se que eram criadas repetições de valores de IC50 para o mesmo composto, ou seja, para os compostos com mais de um valor de IC50, estes tinham associado o valor de IC50 em μg.mL-1 e em nM.

Estas inconsistências, valores de IC50 do composto finasterida incorretamente atribuídos a outros compostos nos dados da 5α-R2 e os valores de IC50 que se repetiam nos compostos obtidos para a 5α-R1, foram retirados do conjunto de dados a analisar. No final os conjuntos de compostos analisados são constituídos por 304 e 406 compostos para a 5α-R1 e 5α-R2, respetivamente.

Para compreender melhor a variação dos valores de IC50 para compostos com vários valores reportados, procedeu-se à construção de um gráfico Box Plot, utilizando o programa Microsoft Office Excel 2007. Desta análise, concluiu-se que para o mesmo

composto a maior parte dos valores de IC50 estão próximos, tal como se pode observar nos gráficos seguintes (Fig. 11 a 15).

Na 5α-R1, para o composto ChEMBL710 (finasterida) foram obtidos 19 valores de IC50. Observou-se que 75% dos valores IC50 se encontravam entre 0 a 370 nM e que o valor mediano é 150 nM, verificando-se que os outros 25% dos valore de IC50 são bastante dispersos (Fig.11).

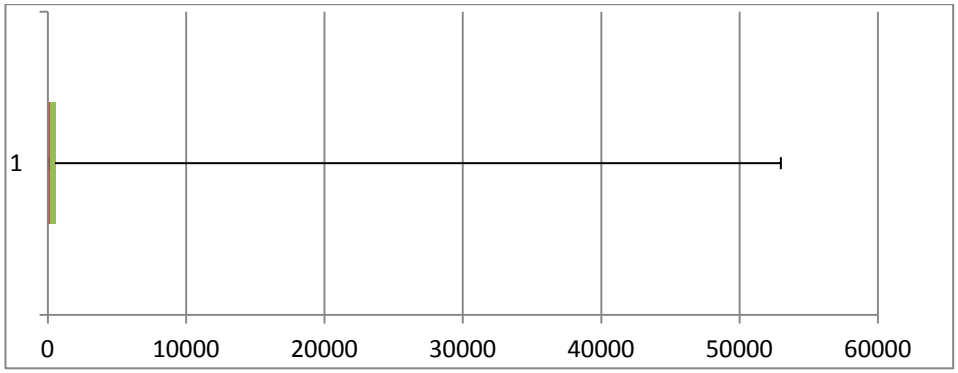


Figura 11. Box Plot do composto ChEMBL710 para a 5α-R1.

Para o composto ChEMBL29082 (sem nome atribuído), com 6 valores de IC50, verificou-se mais uma vez pela estrutura fina da caixa, que 50% dos valores de IC50 se encontram muito próximos, e a localização da caixa no gráfico indica que 75% dos valores de IC50 estão relativamente próximos, entre 18nM e 1475nM, comparativamente aos restantes 25% dos valores de IC50 (Fig. 12). O valor mediano de IC50 deste composto é 1050 nM.

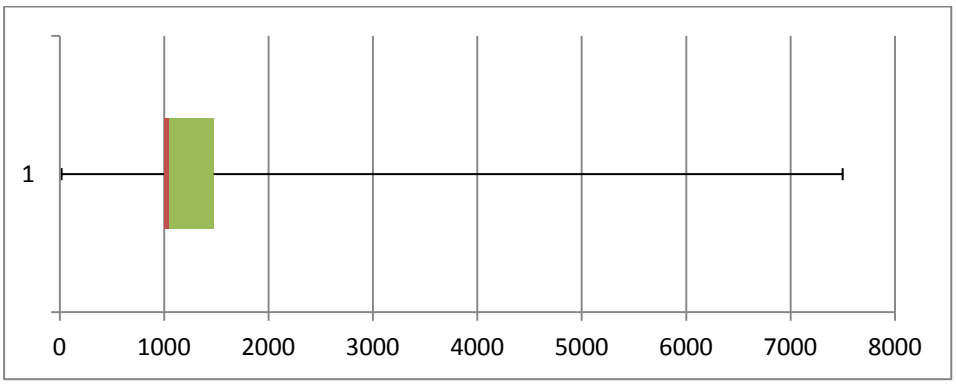


Figura 12. Box Plot do composto ChEMBL29082 para a 5α-R1.

Na 5α-R2 para o composto ChEMBL710 (finasterida) estão disponíveis 22 valores de IC50. Analisando a Box Plot (Fig. 13) observa-se que 75% dos valores IC50 se encontram muito próximos, entre 0 a 24 nM, e que o valor mediano de IC50 para este composto é de 4,365 nM, verificando-se que os restantes 25% dos valores de IC50 são muito dispersos.

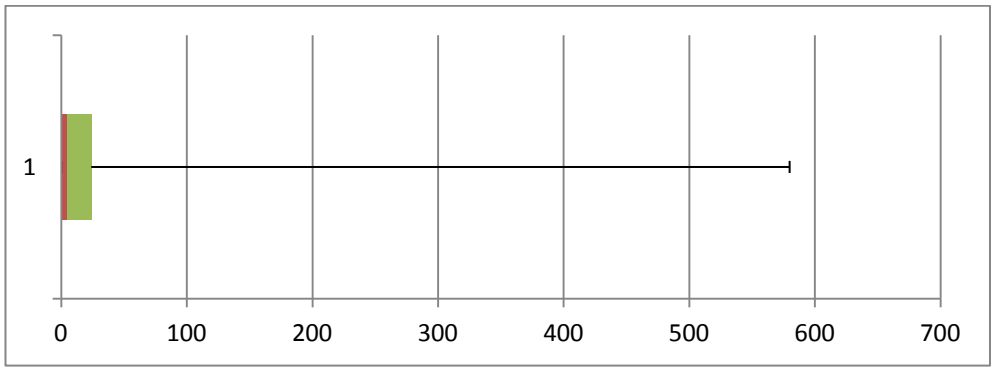


Figura 13. Box Plot do composto ChEMBL710 para a 5α-R2.

Para o composto ChEMBL1201841 (sem nome atribuído), com 7 valores de IC50 compilados, pode-se observar que 50% dos valores de IC50 são muito próximos variando entre 7,3 e 37nM estando os outros 50% dos valores de IC50 mais dispersos (Fig. 14), sendo a mediana observada para os valores de IC50 é de 37 nM.

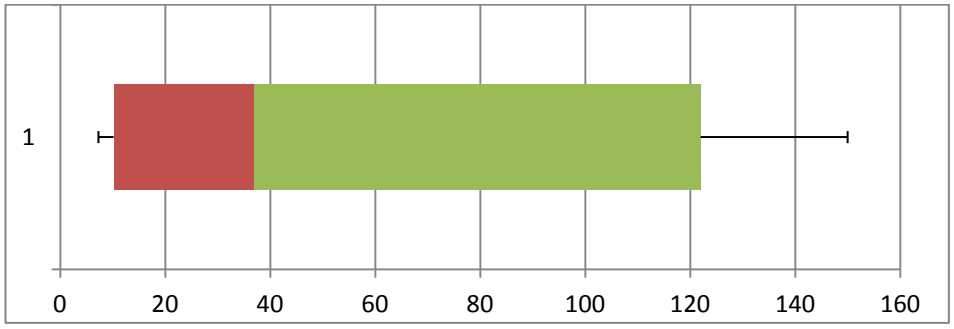


Figura 14. Box Plot do composto ChEMBL1201841 para a 5α-R2.

O composto ChEMBL290823 (Epristeride), com 5 valores de IC50, contém 75% dos valores de IC50 muito próximos, variando os valores entre 0 e 5,5nM e com mediana de 0,6 nM, sendo os outros 25% dos valores de IC50 dispersos em relação aos primeiros 75% dos valores, como pode ser observado na Box Plot a seguir (Fig. 15).

Figura 15. Box Plot do composto ChEMBL29082 para a 5α-R2.

Uma vez que, para os compostos referidos acima, os valores de IC50 são próximos, optou-se por incluir no conjunto de dados de estudo os valores medianos de IC50. Convém ainda referir que existem outros compostos com dois ou três valores de IC50 cujo os valores de IC50 são iguais ou muito próximos entre si, e que para estes compostos (CHEMBL118283, CHEMBL118389, CHEMBL118390, CHEMBL118447, CHEMBL118824, CHEMBL119722, CHEMBL16456, CHEMBL24464, CHEMBL27549, CHEMBL277403, CHEMBL277664, CHEMBL283564, CHEMBL300446, CHEMBL323996, CHEMBL326743, CHEMBL333684, CHEMBL57789, CHEMBL96006, CHEMBL99448) também se assumiu o valor mediano de IC50.

1.2. Cálculo de descritores moleculares

No software da Chemaxon [15] foi efetuado o cálculo dos descritores moleculares dos 304 compostos que inibem a 5α-R1 (Anexos, Tabela 37) e dos 354 compostos que inibem a 5α-R2 (Anexos, Tabela 38). A Chemaxon é uma empresa especializada em programas e aplicações quimioinformáticas.

O *plugin* cxcalc é um programa de linha de comando encontrado nos programas Marvin Beans e JChem da Chemaxon que permite o cálculo de um grande conjunto de descritores moleculares. Estes descritores moleculares encontram-se organizados por categorias (análise elementar, carga, conformação, geometria, isómeros, enumerações *Markush*, nome, partição, previsão, protonação e outros). Para este trabalho foram calculados 40 descritores moleculares, todos numéricos, distribuídos pelas diferentes classes acima enumeradas (Anexos, Tabela 39).

2. Métodos

Os métodos de aprendizagem de máquina utilizados no nosso trabalho englobaram: (1) o método de análise de componentes principais, sendo este método um método de redução de dimensões; (2) um método de seleção de atributos, utilizando uma estratégia *Backward elimination* e o classificador *Naïve Bayes*; e (3) dois métodos de classificação, árvores de decisão e máquinas de vetores de suporte; (4) um método de análise estrutural dos compostos.

(15) (http://www.chemaxon.com/)

2.1. Redução da dimensão dos dados e seleção de atributos

A inclusão de um elevado número de atributos desnecessários podem causar ruído nos resultados (Massart et al., 1988; Smith et al., 2002). O número de atributos em análise pode ser reduzido de duas formas. Através da redução da dimensão de dados, em que as variáveis originais são combinadas num pequeno número de componentes principais, ou utilizando técnicas de seleção de atributos (no inglês, *feature selection*) em que a partir de *m* variáveis se seleciona um subconjunto de variáveis que pareçam ser as mais discriminantes. As variáveis obtidas correspondem portanto a algumas das variáveis originais, enquanto que nos métodos de redução de dimensão as novas variáveis são obtidas pelo uso de todas as variáveis originais mas combinando-as num pequeno número de novas variáveis, reduzindo-se a complexidade dos dados e o tempo de processamento para extrair destes algum conhecimento.

2.1.1. Análise de componentes principais

A análise de componentes principais (PCA, do inglês *Principal Component Analysis*) é um método estatístico multivariado. Estes métodos são ferramentas estatísticas que estudam o comportamento de três ou mais variáveis ao mesmo tempo. São usados principalmente para encontrar as variáveis menos representativas de forma a eliminá-las, simplificando assim os modelos estatísticos e facilitando a compreensão da relação entre vários grupos de variáveis.

O PCA tem como princípio básico a análise dos dados com o objetivo de proceder à sua redução, eliminação de sobreposições e a escolha das formas mais representativas de dados a partir de combinações lineares das variáveis originais obtendo-se um novo conjunto de variáveis denominadas componentes principais (Anthon et al., 2004).

Considerando as variáveis x_1 e x_2, num espaço com duas dimensões, uma transformação linear transforma as variáveis x_1, x_2 em novas variáveis u_1, u_2. Esta transformação linear pode ser escrita como demonstrado nas equações seguintes.

$u_1 = ax_1 + bx_2$ (Eq. 1)

$u_2 = cx_1 + dx_2$, (Eq. 2)

onde *a, b, c, d* são os coeficientes.

Estas equações podem ser representadas geometricamente (Fig.16). O ponto P é definido pelos valores das variáveis x_1 e x_2, que são agora caracterizados pelos novos valores das variáveis u_1 e u_2.

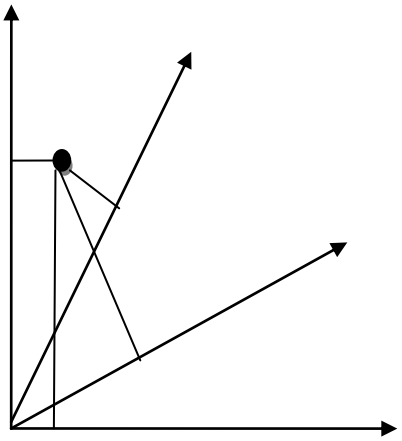


Figura 16. Representação geométrica da transformação linear das variáveis x1 e x2 nas variáveis u1 e u2.

Do ponto de vista geométrico, estas duas componentes principais não são correlacionadas, o que significa que as novas variáveis são ortogonais. Esta condição pode ser escrita como

$a.c + b.d = 0$ (Eq. 3)

O mesmo se aplica se foi considerado um espaço com *m*-dimensões,

$u_1 = v_{11}x_1 + v_{12}x_2 + \ldots + v_{1m}x_m$ (Eq. 4)

$u_2 = v_{21}x_1 + v_{22}x_2 + \ldots + v_{2m}x_m$ (Eq. 5)

$u_m = v_{m1}x_1 + v_{m2}x_2 + \ldots + v_{mm}x_m$, (Eq. 6)

onde v_{m1}, v_{m2}, v_{mm} são os coeficientes.

Existem duas condições que devem ser satisfeitas pelos coeficientes das componentes principais:

(1) Para cada par de componentes u_k e u_r

$v_{k1}v_{r1} + v_{k2}v_{r2} + \ldots + v_{km}v_{rm} = 0$ (Eq. 7)

(2) Para cada componente u_r

$V^2_{r1} + v^2_{r2} + \ldots + v^2_{rm} = 1$ (Eq. 8)

A primeira condição (Eq. 7) impõe que o produto de dois coeficientes é zero e portanto os vetores correspondentes são ortogonais. A segunda condição (Eq. 8) impõe que o vetor deve ter norma 1. Estas duas condições são explícitas para todas as dimensões (Massart et al., 1988).

Em estatística, existem várias análises que podem ser feitas sobre um conjunto de dados, como a média aritmética, o desvio padrão e a variância. Porém, todas essas medidas consideram separadamente cada tipo de dados. No entanto, a covariância é medida sempre entre duas dimensões; o cálculo da covariância entre uma dimensão e ela mesma resulta no cálculo da variância. Se forem consideradas mais de duas dimensões, é necessário calcular a covariância entre cada dimensão. A partir dessa ideia, surge o conceito de matriz de covariância (Anthon et al., 2004).

Assim, num espaço com *m*-dimensões, as condições descritas pelas equações 7 e 8 juntas são equivalentes a:

$V.V^T = 1$ (Eq. 9)

onde V^T é a matriz transposta da matriz V e I é a matriz identidade:

$$I = \begin{pmatrix} 1\ 0 \ldots 0 \\ 0\ 1 \ldots 0 \\ 0\ 0 \ldots 1 \end{pmatrix} \quad \text{(Eq. 10)}$$

A diagonal principal da matriz contém as variâncias e as demais posições a correlação entre as direções. A matriz é simétrica e real, de modo que é sempre possível encontrar um conjunto de vetores próprios ortogonais.

Os coeficientes das variáveis originais para cada uma das componentes principais são as coordenadas dos correspondentes vetores próprios. O fator (no inglês, *loading*) de uma variável numa componente principal é definido como esta coordenada multiplicada pela raiz quadrada dos valores próprios das componentes principais. Os próprios coeficientes são frequentemente chamados de fator. Quanto maior o fator (*loading*) de uma variável numa componente principal, mais a variável tem influência nesse componente. Estes factores (*loadings)* podem ser interpretados como a correlação entre as variáveis e os componentes principais. Na prática, as primeiras duas ou três componentes principais explicam uma importante parte do total da variância (Massart et al., 1988).

A primeira componente principal (PC1) extraída no PCA é responsável pela quantidade máxima de variância das variáveis analisadas. Isto significa que a primeira componente principal será correlacionada com algumas ou muitas das variáveis analisadas. A segunda componente principal (PC2) será ortogonal a PC1 e portanto não correlacionada com a primeira componente principal e responsável pela quantidade máxima de variância no conjunto de dados que não foi contabilizado pelo primeiro componente principal. Por outras palavras, a segunda componente principal está correlacionada com variáveis que não exibem forte correlação com a primeira componente principal. A projeção dos pontos obtidos das duas componentes principais, de facto, define um plano *m*-dimensional. Nas restantes componentes principais o processo é semelhante, e a cada novo componente principal a quantidade máxima de variância é menor, por isso para a interpretação dos dados só são escolhidos as primeiras duas ou três componentes principais (Maart, 1988; Jolliff, 2003).

2.2. Seleção de atributos importantes utilizando uma estratégia *Backward Elimination* e um classificador *Naïve Bayes*

Os métodos de seleção de atributos (no inglês, *feature selection*) dividem-se em duas grandes categorias: *Filter* e o *Wrapper* (Fig. 17). Os métodos *Filter* selecionam o subconjunto de variáveis avaliando apenas as características intrínsecas dos dados. Estes métodos usam a informação dos dados de treino para escolher as variáveis individualmente e escolhem um subconjunto selecionando as *K* variáveis melhor avaliadas. Para tal são utilizadas métricas como χ^2, ganho de informação (no inglês, *information gain*), CFS (no inglês, *correlation-based feature selection*) (Sayes et al., 2007; Goodarzi et al., 2012). Os métodos do tipo *Wrapper*

utilizam um algoritmo de classificação para avaliar os subconjuntos de atributos de acordo com a sua capacidade preditiva para determinada característica (por exemplo, atividade). Estes métodos realizam uma busca entre os possíveis subconjuntos a serem avaliados e o algoritmo de classificação é executado para cada subconjunto de variáveis, avaliando a sua capacidade preditiva. Métodos *Wrapper* produzem geralmente resultados melhores que os métodos *Filter*, uma vez que a seleção das variáveis é guiada pelo próprio algoritmo de classificação que é utilizado na fase de análise dos dados. O objetivo da busca é encontrar o subconjunto com a melhor qualidade, utilizando uma função heurística para guiá-lo (Lei et al., 200).

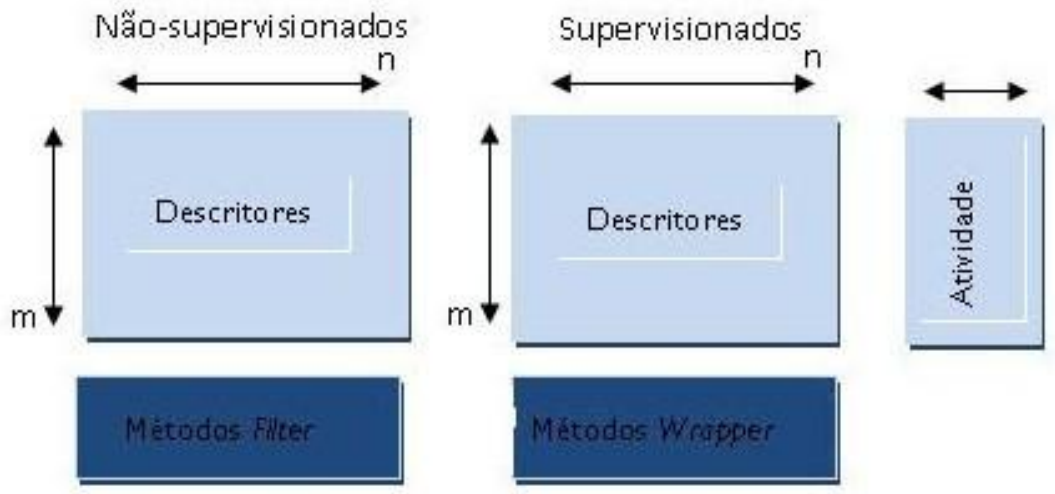


Figura 17. Métodos *Filter*: seleção de atributos com base apenas nos descritores sem a contribuição de qualquer algoritmo de aprendizagem. Métodos *Wrapper*: seleção de atributos com base nos descritores moleculares e na atividade usando um algoritmo de aprendizagem. (m) é o número de objetos (moléculas) e (n) número de variáveis (descritores). (Adaptado de Goodarzi et al., 2012).

A escolha dos subconjuntos de variáveis a avaliar depende da estratégia da pesquisa heurística, que pode ser *Forward selection* ou *Backward elimination* (Fig. 18). Nos métodos *Forward selection* a busca é iniciada sem variáveis e as mesmas são adicionadas uma a uma, aumentando a dependência até que seja encontrado o conjunto de variáveis com o valor máximo (ou mínimo) seja obtido para um critério especificado. Quando se aplica uma estratégia *Backward elimination*, a busca começa com todo o conjunto de atributos, vai-se eliminando um atributo a cada passo (Karnon et al., 2010).

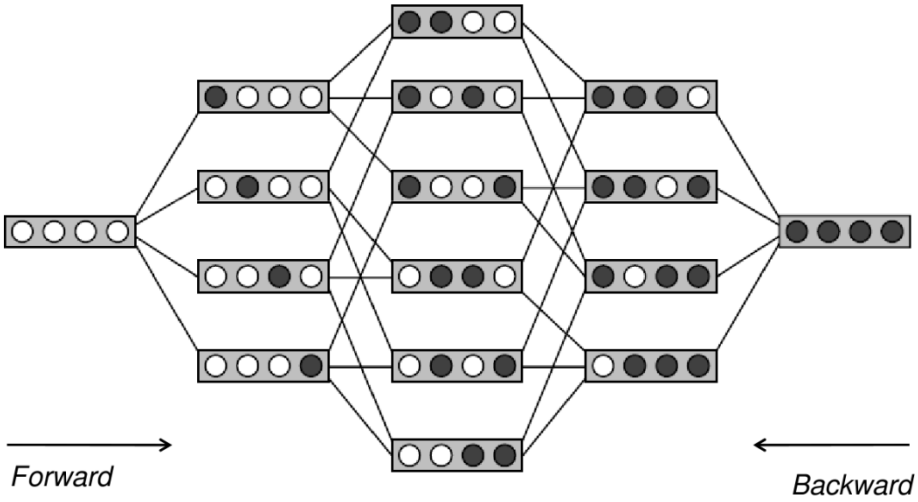


Figura 18. Esquema de *Forward selection* e *Backward elimination.*

O algoritmo de classificação utilizado neste trabalho foi o classificador *Naïve Bayes*,

$V_m = \text{argmax}_{vj \in v} P(vj)\ \pi P(a_i \backslash v_j)$ (Eq.11)

Este classificador é provavelmente o classificador mais utilizado em aprendizagem de máquina, é baseado no teorema de Bayes. É denominado *naïve* por assumir que os atributos são condicionalmente independentes, ou seja, a informação de um atributo não é dependente de nenhum outro. Além disso, este reporta o melhor desempenho em várias tarefas de classificação (Cortizo et al., 2006).

2.3. Modelos de classificação

Em aprendizagem de máquina, a classificação é considerada um exemplo de aprendizagem supervisionada, ou seja, aprendizagem onde um conjunto de treino corretamente identificado é válido. Este modelo representa essencialmente, uma relação entre os valores dos atributos de previsão e as classes, permitindo a previsão de uma classe num exemplo, dados os seus valores de atributos de previsão. Um dos principais objetivos de um algoritmo de classificação é maximizar a precisão da previsão obtida pelo modelo de classificação. Existem vários algoritmos de classificação tais como: o SVM, árvores de decisão, *Bayesian networks*, classificadores baseados em regras, Nearest *neighbour*, rede

neural artificial, *Rough sets*, *Fuzzy logic*, algoritmos genéticos, entre outros (Beniwal & Arora 2012). Neste trabalho os algoritmos de classificação utilizados foram o SVM e árvores de decisão, os algoritmos mais utilizados para estudos de QSAR e considerados mais eficientes na literatura. Niu e colaboradores (Niu et al. 2007) utilizaram o algoritmo SVM para discriminar 32 fenilaminas entre agonistas e antagonistas e prever a atividade desses compostos, tendo obtido bons resultados e concluído que o SVM é um bom algoritmo para estudos de SAR/QSAR. Num outro trabalho Darnag e colegas (Darnag et al. 2010), obtiveram melhores resultados com a utilização de SVM para os seus estudos de QSAR em derivados (TIBO) quando comparado com outros algoritmos tais como redes neurais artificiais. Já Kovalishyn et al. (2012), utilizou o algoritmo árvores de decisão nos seus estudos de QSAR, tendo obtido bons resultados. Adicionalmente Liu et al. (2012), num estudo que fez com o objetivo de identificar quais os melhores algoritmos de classificação concluiu que o algoritmo árvores de decisão é um dos melhores.

2.3.1. Árvores de decisão

Uma árvore de decisão representa um conjunto de regras que seguem uma hierarquia de classes e valores utilizados para classificar novos elementos (Fig.19). Um elemento é classificado a partir da raiz da árvore utilizando o atributo nela especificado. Em seguida, na sequência da ramificação correspondente o processo de decisão continua com base no valor do atributo especificado. Este processo é repetido até não se encontrar novas ramificações e é então tomada a decisão sobre a classe do novo elemento. Do algoritmo da árvore de decisão, crescem sucessivas árvores a partir de um conjunto de treino, subdividindo este conjunto até que a árvore é formada apenas por nós "puros", em que cada nó representa apenas uma única classe, ou até à satisfação de um critério.

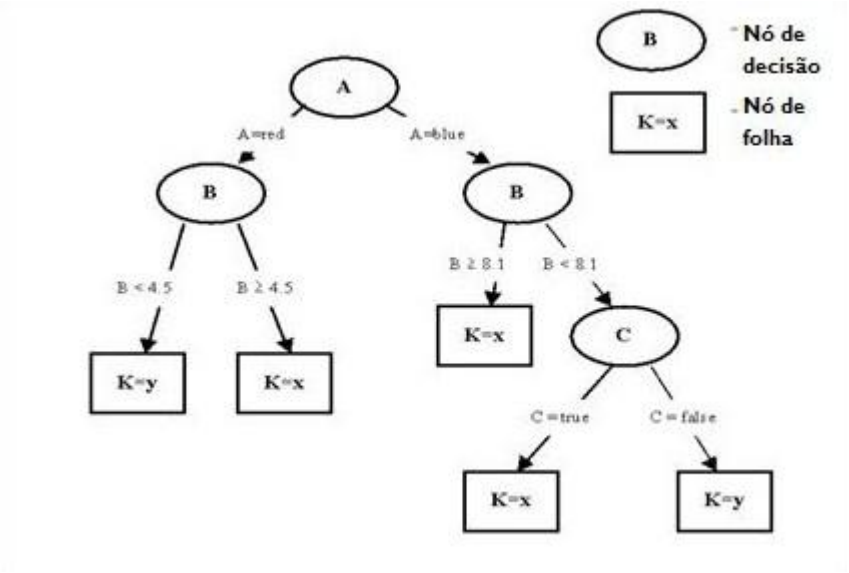


Figura 19. Esquema geral de uma árvore de decisão, onde os nós (B) representam nós de decisão e nós (K=y) representam nós de folha.

Uma árvore de decisão tem uma estrutura definida com base na conceção de uma árvore real: raiz, ramos (valores dos atributos/variáveis), nós internos (os atributos/variáveis, com duas ou mais subárvores que representam caminhos diferentes), e folhas ou nós puros (as classes) (Breiman, 2001).

A partir do conceito de árvore de decisão, Breiman (2001) formalizou o conceito de *random forest*. Uma *random forest* (floresta) é uma combinação de árvores de decisão, em que cada árvore gerada é utilizada para classificar um novo elemento, e a decisão final acerca da classe a que este novo elemento pertence, é tomada com base num voto por maioria.

2.3.2. Máquinas de vetores de suporte (SVM)

A técnica das máquinas de vetores de suporte (SVM, no inglês *support vector machine*) tornou-se popular na década de 1990, tendo sido desenvolvida por Vapnik e introduzida na área da quimioinformatica por Burbidge (Burbidge et al., 2001) e Czerminski (Czerminski et al., 2001) e colaboradores. Desde então as SVMs tornaram-se uma técnica popular de classificação. Esta técnica baseia-se na procura de uma fronteira ou um hiperplano que separa as duas classes, por exemplo uma biblioteca de composto ativos e não ativos. O hiperplano é posicionado usando exemplos no conjunto de treino que são conhecidos como

vetores de suporte. As moléculas no conjunto de teste são mapeados para as mesmas características espaciais e a sua atividade é prevista de modo a prever de que lado do hiperplano se encontram. O nível de confiança é dado através da distância para a fronteira, e quanto maior a distancia maior é a confiança na previsão (Fukunishi, 2009).

O problema inicial de classificação tratado pela SVMs foi o da classificação binária. O problema de classificação binária, trata da classificação de duas classes, através da definição de um hiperplano ótimo a partir de um conjunto de dados de treino linearmente separável. Um conjunto de treino é dito linearmente separável se for possível separar os padrões de classes diferentes contidos no mesmo por pelo menos um hiperplano.

Considerando o conjunto de treino $\{(xi, di)\}^{N}_{i=1}$, onde x_i é o padrão de entrada para o *i*-ésimo exemplo e d_i é a resposta desejada d_i ϵ {+1,−1} que representa as classes linearmente separáveis. A equação que separa os padrões através de hiperplanos pode ser definida por:

$$w^T x+b =0 \quad \text{(Eq. 12)}$$

onde, $w^T x$ é o produto escalar entre os vetores w e x, x é um vetor de entrada que representa os padrões de entrada do conjunto de treino, w é um vetor de pesos ajustáveis e *b* é um limiar, também conhecido como *bias*. A Figura 20 mostra o hiperplano de separação (w, *b*), num espaço bidimensional, para um conjunto de treino linearmente separável (Mining *et al.,* 1998).

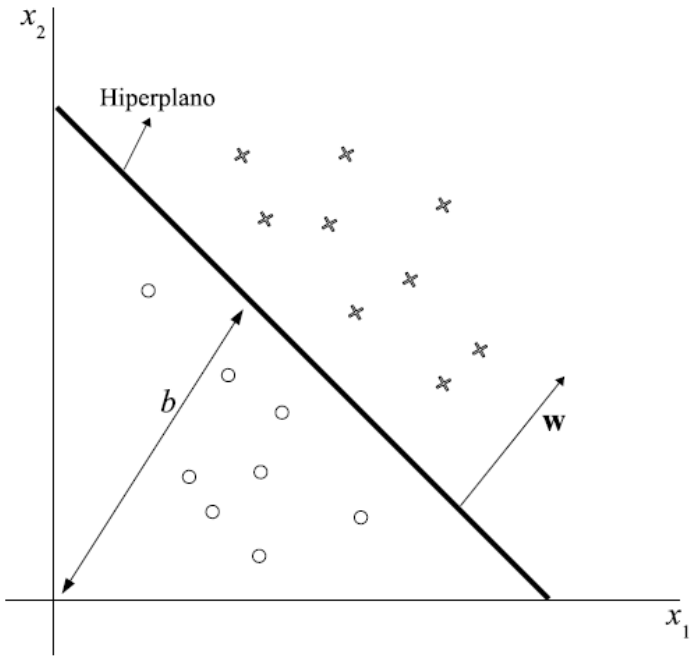


Figura 20. Hiperplano de separação (w, *b*) para um conjunto de treino bidimensional.

A equação anterior (Eq. 12) pode ser reescrita como:

$$\begin{cases} \mathbf{w}T\ \mathbf{x}i+b \geq 0, \text{ se } di = +1 \quad \text{(Eq. 13)} \\ \mathbf{w}T\ \mathbf{x}i+b < 0, \text{ se } di = -1 \quad \text{(Eq. 14)} \end{cases}$$

A margem de separação, distância entre o hiperplano definido na equação 12 e o ponto mais próximo de ambas as classes, é representado por ρ. O objetivo de uma SVM é encontrar um hiperplano que separe o conjunto de treino sem erro e que maximize a margem de separação. Nestas condições, o hiperplano é referido como hiperplano ótimo. A Figura 21 ilustra o hiperplano ótimo para um espaço de entrada bidimensional (Melville et al., 2009).

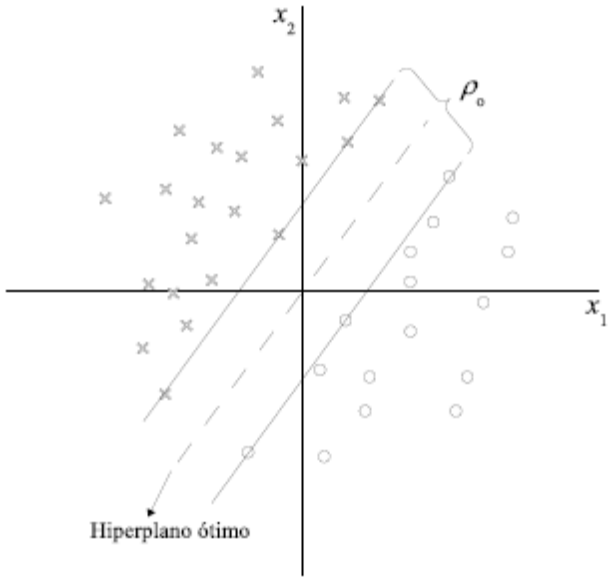


Figura 21. Hiperplano ótimo com máxima margem ρo de separação dos padrões linearmente separáveis.

Um problema de classificação binária, onde as classes distintas não são linearmente separáveis no espaço original, pode ser resolvido com um mapeamento não linear através de um produto interno (*kernel*). Este produto transforma o espaço original num espaço de características de dimensão maior, e o problema que era não linearmente separável no espaço original passa a ser linearmente separável no novo espaço de características (Fig.22).

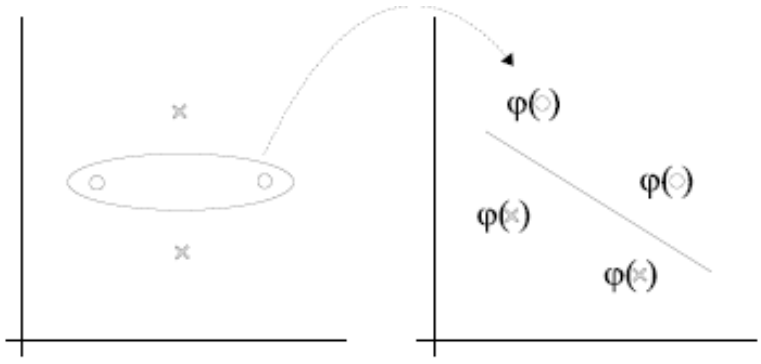


Figura 22. Mapeamento de características.

2.4. Medidas de desempenho

Atualmente, não existe uma metodologia padrão para a análise, avaliação e comparação estatística de resultados gerados pelas técnicas de aprendizagem de máquina 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.

Sensibilidade, especificidade, precisão e *F-score* são medidas estatísticas para avaliar o desempenho de classificadores. A sensibilidade (também designada de *Recall;* Eq. 15) mede a percentagem de previsões corretas. Por seu lado, especificidade (Eq. 2) mede a proporção de negativos que se previu corretamente. A precisão (Eq. 3) indica a percentagem de positivos observados que se previu corretamente. A *F-score* (Eq. 4) combina a medida de precisão com sensibilidade (*recall*).

$$\text{Sensibilidade} = \frac{T_P}{T_P+F_N} \qquad \text{(Eq. 15)}$$

$$\text{Especificidade} = \frac{T_N}{T_N+F_P} \qquad \text{(Eq. 16)}$$

$$\text{Precisão} = \frac{T_P}{T_P+F_P} \qquad \text{(Eq. 17)}$$

$$\text{F-score} = \frac{(2 \times \text{precision} \times \text{recall})}{(\text{precision} + \text{recall})} \qquad \text{(Eq.18)}$$

Nas equações 15 a 18, T_P representa o número de verdadeiros positivos, T_N representa o número de verdadeiros negativos, F_P representa o número falsos positivos e F_N representa o número de falsos negativos.

2.5. KNIME

Para a execução dos métodos de aprendizagem de máquina foi utilizado o programa KNIME (Konstanz Information Miner). O KNIME [16] disponibiliza uma interface gráfica de trabalho para todo o processo de análise de dados e integra vários componentes de aprendizagem de máquina.

Neste trabalho os métodos de aprendizagem de máquina utilizados foram a análise de componentes principais, seleção de atributos, árvores de decisão e máquina de vetores de suporte. Para o cálculo destes métodos foi necessário proceder à conversão dos valores de absolutos IC50 para classes de IC50 (Tabela 2).

Tabela 2. Atribuição de classes aos valores de IC50**.**

		Compostos	
Classes IC50	**Valores de IC50**	**5α-R1**	**5α-R2**
Muito bom	0-1	4	107
Bom	1-10	53	48
Médio	10-100	88	48
Mau	100, +∞	157	151

Todos os algoritmos de aprendizagem calculados no programa KNIME foram avaliados utilizando o método estatístico de validação cruzada (Fig. 23). Este método permite comparar e validar os algoritmos de aprendizagem dividindo os dados em dois segmentos: um utilizado como um modelo de treino e o outro usado para validar o modelo (Fig. 24) (Refaeilzadeh et al., 2012).

No KNIME o método estatístico de validação cruzada é calculado através de um meta nó. Os meta nós contêm *subworkflows*, ou seja, na área de trabalho estes parecem apenas um único nó, mas possuem muitos nós (e eventualmente, outros metas nós) englobados. Alguns nós dentro do meta nó de validação cruzada e apresentados na figura 24 podem ser substituídos, o nó de aprendizagem e o nó preditor podem ser qualquer nó do método que se pretenda avaliar (*SVM, RandomForest,…*).

(16) (http://www.knime.org/)

Para o método de validação cruzada os dados foram divididos em amostragens aleatórias com um k-*fold* (k=5). Para todos os métodos procedeu-se à normalização dos dados destes, exceto no método de árvores de decisão. A normalização utilizada foi a normalização *z-score* que consiste na transformação linear de cada um dos atributos de tal modo que os seus valores sigam uma distribuição normal com a média é de 0 e o desvio padrão de 1.

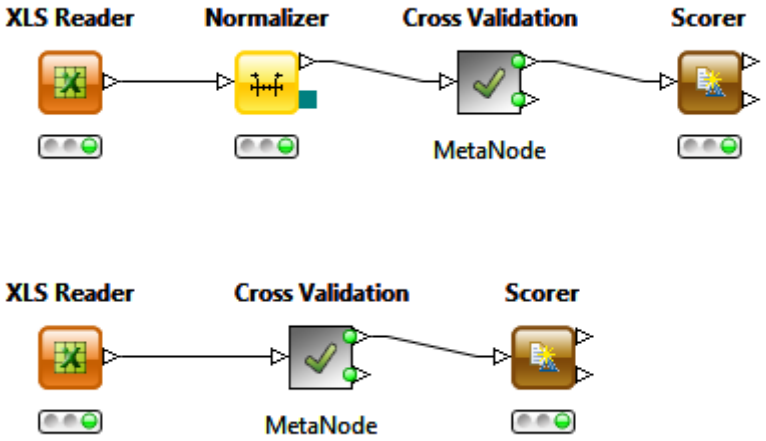


Figura 23. (A). Workflow do KNIME para os métodos PCA, *Feature Selection* e SVM. Para a aplicação destes métodos, os dados foram normalizados. (B). Workflow do KNIME para o método da *RandomForest*, que não necessita que os dados sejam normalizados.

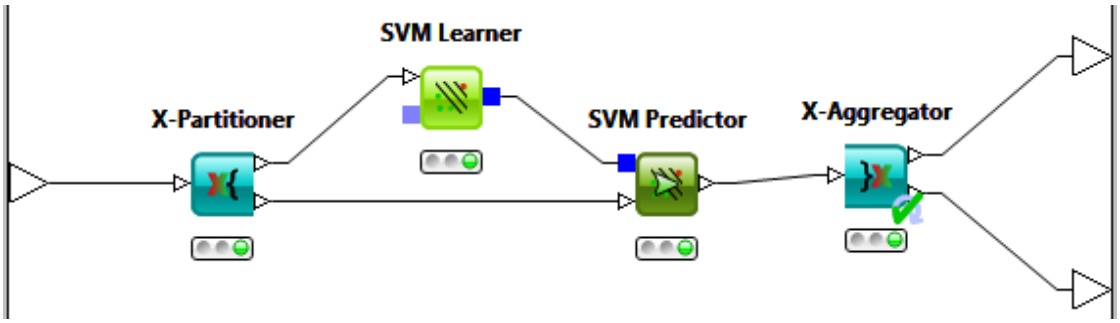
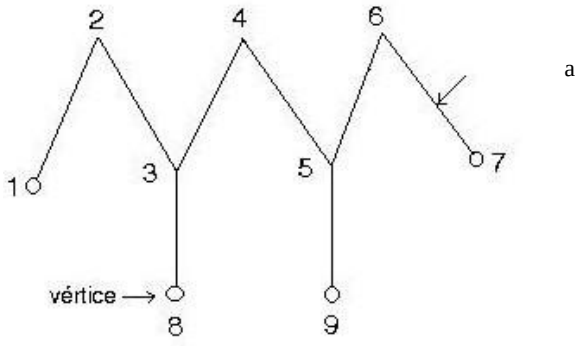


Figura 24. Meta nó da validação cruzada, para o exemplo de SVM.

2.6. Análise estrutural

A pesquisa da subestrutura máxima comum (MCS, do inglês *Maximum Common Substructure*) é o método mais sensível e preciso para determinar as semelhanças estruturais entre pequenas moléculas, sendo indispensável em muitas áreas de investigação como no

desenvolvimento e descoberta de fármacos e química genética. O conceito de máxima subestrutura comum consiste na similaridade baseada na teoria dos grafos, e que é definida como a maior subestrutura partilhada entre dois compostos. A teoria dos grafos é um ramo da matemática que estuda as relações entre os objetos de um determinado conjunto. Grafo é uma noção simples, abstrata e intuitiva, usada para representar a ideia de alguma relação entre os "objetos". Graficamente, aparece representado por uma figura com nós ou vértices, significando os objetos, unidos por um traço denominado de aresta configurando a relação definida entre os objetos (Fig. 25) (Hariharam et al., 2011).



a

Figura 25. Exemplo genérico de um grafo.

Por exemplo, se considerarmos uma estrutura de um composto em duas dimensões, os vértices do grafo representam os átomos, e as arestas representam as ligações entre pares de átomos covalentemente ligados. A pesquisa da MCS pode ser feita com base em dois critérios distintos (Raymond, 2002): (i) com base no maior número de vértices (átomos) em comum (MCIS), ou (ii) com base no maior número de arestas (ligações) em comum (MCES).

A LibMCS é uma biblioteca de pesquisa para subestruturas máximas comuns (MCS) de forma hierárquica, disponibilizada no software Chemaxon. Esta biblioteca considera a pesquisa da MCS entre duas moléculas com base na existência do maior número de ligações em comum, ou seja baseia-se numa estratégia MCES.

2.7. Pesquisa de novos compostos e previsão da sua atividade

Da pesquisa de subestruturas máximas comuns realizada na LibMCS do software Chemaxon, foram analisadas as subestruturas comuns a um maior número de compostos associados. Este processo foi realizado em ambas as isoenzimas 5α-R1 e 5α-R2, para as classes de IC50 muito bom e bom.

A partir das subestruturas máximas comuns a um maior número de compostos, procedeu-se na ChEMBL, à pesquisa de novos compostos. Estes novos compostos têm todos em comum a subestrutura de interesse. Nos compostos obtidos na ChEMBL encontravam-se os compostos utilizados na pesquisa inicial tendo-se prosseguido à remoção destes. De seguida, na Chemaxon para cada novo composto foram calculados 40 descritores moleculares, seguindo-se as mesmas condições do cálculo realizado para os compostos que inibem as duas isoenzimas. Preparados estes novos conjuntos de teste, procedeu-se no KNIME à sua análise através do modelo de classificação de árvores de decisão e SVM com o objetivo de prever a sua classe de IC50.

C. Resultados

I. Modelação da proteína 5α-redutase

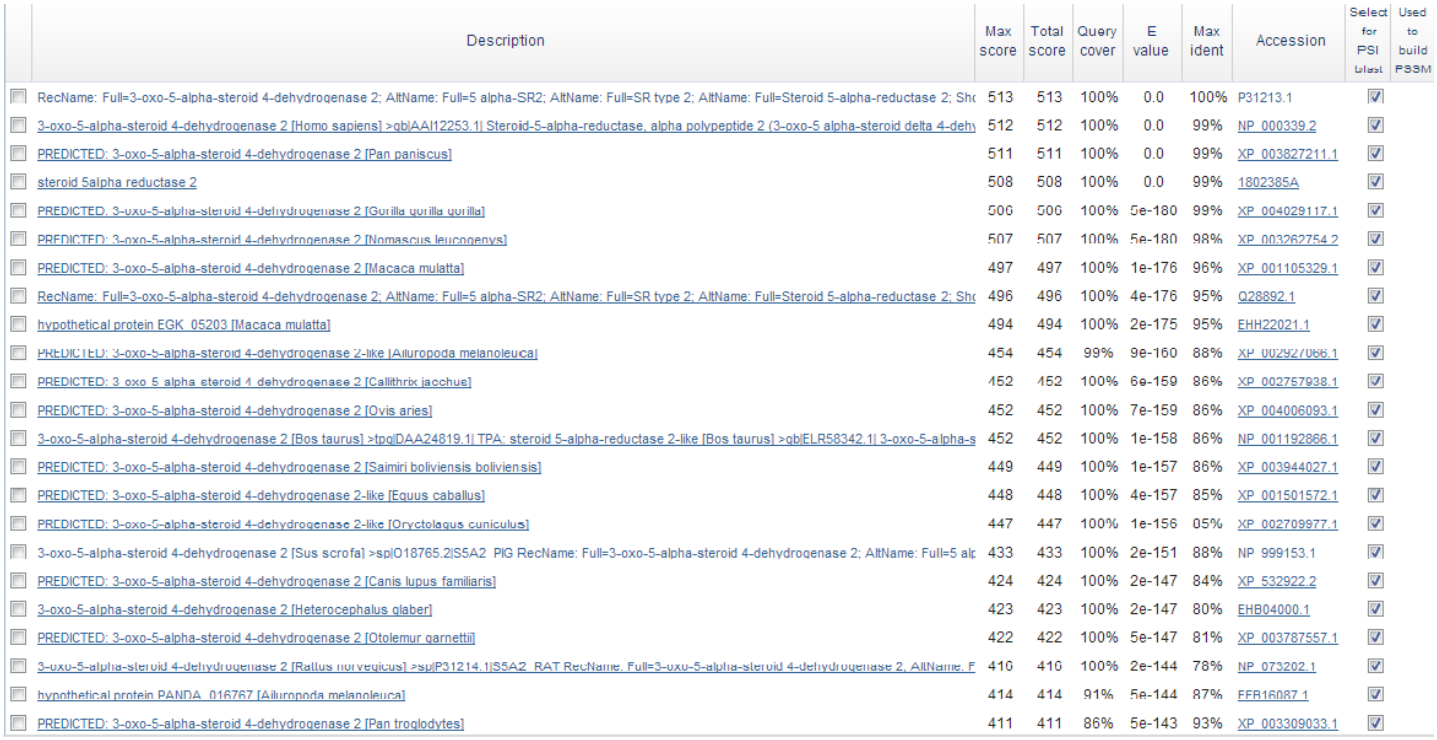
A proteína 5α-redutase é uma proteína transmembranar que não possui estrutura tridimensional determinada experimentalmente. Numa tentativa de obter um modelo estrutural 3D para esta proteína procedeu-se à modelação de proteínas por homologia.

Para a identificação e seleção de proteínas homólogas, o processo utilizado consistiu na pesquisa de sequências de aminoácidos de possíveis proteínas homólogas. Depois de obtida a sequência de aminoácidos da nossa proteína alvo no UniProt, indispensável para as futuras pesquisas nos diferentes programas utilizados ao longo do trabalho, procedeu-se então à identificação e seleção de proteínas molde. Foram utilizadas diversas aproximações para identificar e selecionar modelos de homologia para a 5α-redutase cujos resultados se descrevem em seguida.

1.1. BLAST e JALVIEW

No BLAST foram pesquisadas as possíveis proteínas homólogas à proteína alvo - a 5α-redutase, tendo-se obtido as sequências molde ordenadas de acordo com a percentagem de identidade sequencial, tal como se mostra na Figura 26. Quanto mais próximo de zero se encontrar o valor de parâmetro estatístico E-value (*expected-value*), maior o grau de similaridade da sequência. Este é um parâmetro importante, uma vez que a qualidade do modelo por homologia é dependente da qualidade do alinhamento da sequência e da estrutura do modelo. A adequabilidade do modelo diminui com a diminuição da identidade entre sequências.

Para a proteína 5α-redutase observou-se que as sequências de topo da lista apresentam uma identidade sequencial elevada (> 80%) e um E-*value* próximo de zero, levando a pensar que qualquer uma destas proteínas poderia ser um bom candidato a sequência molde. No entanto, quando se procedeu ao alinhamento múltiplo com o Jalview, observou-se que grande parte destas proteínas eram na realidade a proteína 5α-redutase que se encontra com nomes diferentes em diferentes bases de dados. Progredindo na lista, as proteínas com maior homologia à proteína alvo apresentam uma identidade sequencial baixa.



Description	Max score	Total score	Query cover	E value	Max ident	Accession	Select for PSI blast	Used to build PSSM
RecName: Full=3-oxo-5-alpha-steroid 4-dehydrogenase 2; AltName: Full=5 alpha-SR2; AltName: Full=SR type 2; AltName: Full=Steroid 5-alpha-reductase 2; Sh	513	513	100%	0.0	100%	P31213.1	☑	
3-oxo-5-alpha-steroid 4-dehydrogenase 2 [Homo sapiens] >gb|AAI12253.1| Steroid-5-alpha-reductase, alpha polypeptide 2 (3-oxo-5 alpha-steroid delta 4-deh	512	512	100%	0.0	99%	NP_000339.2	☑	
PREDICTED: 3-oxo-5-alpha-steroid 4-dehydrogenase 2 [Pan paniscus]	511	511	100%	0.0	99%	XP_003827211.1	☑	
steroid 5alpha reductase 2	508	508	100%	0.0	99%	1802385A	☑	
PREDICTED: 3-oxo-5-alpha-steroid 4-dehydrogenase 2 [Gorilla gorilla gorilla]	506	506	100%	5e-180	99%	XP_004029117.1	☑	
PREDICTED: 3-oxo-5-alpha-steroid 4-dehydrogenase 2 [Nomascus leucogenys]	507	507	100%	5e-180	98%	XP_003262754.2	☑	
PREDICTED: 3-oxo-5-alpha-steroid 4-dehydrogenase 2 [Macaca mulatta]	497	497	100%	1e-176	96%	XP_001105329.1	☑	
RecName: Full=3-oxo-5-alpha-steroid 4-dehydrogenase 2; AltName: Full=5 alpha-SR2; AltName: Full=SR type 2; AltName: Full=Steroid 5-alpha-reductase 2; Sh	496	496	100%	4e-176	95%	Q28892.1	☑	
hypothetical protein EGK_05203 [Macaca mulatta]	494	494	100%	2e-175	95%	EHH22021.1	☑	
PREDICTED: 3-oxo-5-alpha-steroid 4-dehydrogenase 2-like [Ailuropoda melanoleuca]	454	454	99%	9e-160	88%	XP_002927066.1	☑	
PREDICTED: 3-oxo-5-alpha-steroid 4-dehydrogenase 2 [Callithrix jacchus]	452	452	100%	6e-159	86%	XP_002757938.1	☑	
PREDICTED: 3-oxo-5-alpha-steroid 4-dehydrogenase 2 [Ovis aries]	452	452	100%	7e-159	86%	XP_004006093.1	☑	
3-oxo-5-alpha-steroid 4-dehydrogenase 2 [Bos taurus] >tpg|DAA24819.1| TPA: steroid 5-alpha-reductase 2-like [Bos taurus] >gb|ELR58342.1| 3-oxo-5-alpha-s	452	452	100%	1e-158	86%	NP_001192866.1	☑	
PREDICTED: 3-oxo-5-alpha-steroid 4-dehydrogenase 2 [Saimiri boliviensis boliviensis]	449	449	100%	1e-157	86%	XP_003944027.1	☑	
PREDICTED: 3-oxo-5-alpha-steroid 4-dehydrogenase 2-like [Equus caballus]	448	448	100%	4e-157	85%	XP_001501572.1	☑	
PREDICTED: 3-oxo-5-alpha-steroid 4-dehydrogenase 2-like [Oryctolagus cuniculus]	447	447	100%	1e-156	85%	XP_002709977.1	☑	
3-oxo-5-alpha-steroid 4-dehydrogenase 2 [Sus scrofa] >sp|O18765.2|S5A2_PIG RecName: Full=3-oxo-5-alpha-steroid 4-dehydrogenase 2; AltName: Full=5 alp	433	433	100%	2e-151	88%	NP_999153.1	☑	
PREDICTED: 3-oxo-5-alpha-steroid 4-dehydrogenase 2 [Canis lupus familiaris]	424	424	100%	2e-147	84%	XP_532922.2	☑	
3-oxo-5-alpha-steroid 4-dehydrogenase 2 [Heterocephalus glaber]	423	423	100%	2e-147	80%	EHB04000.1	☑	
PREDICTED: 3-oxo-5-alpha-steroid 4-dehydrogenase 2 [Otolemur garnettii]	422	422	100%	5e-147	81%	XP_003787557.1	☑	
3-oxo-5-alpha-steroid 4-dehydrogenase 2 [Rattus norvegicus] >sp|P31214.1|S5A2_RAT RecName: Full=3-oxo-5-alpha-steroid 4-dehydrogenase 2; AltName: F	416	416	100%	2e-144	78%	NP_073202.1	☑	
hypothetical protein PANDA_016767 [Ailuropoda melanoleuca]	414	414	91%	5e-144	87%	EFB16087.1	☑	
PREDICTED: 3-oxo-5-alpha-steroid 4-dehydrogenase 2 [Pan troglodytes]	411	411	86%	5e-143	93%	XP_003309033.1	☑	

Figura 26. Resultados da pesquisa no BLAST. Sequências ordenadas de acordo com a percentagem de identidade sequencial.

Numa tentativa de melhorar o alinhamento obtido, procedeu-se então ao alinhamento múltiplo de todas as sequências obtidas no BLAST utilizando o programa Jalview, considerando no alinhamento também os resíduos estruturalmente equivalentes, isto é com características estruturais comuns. Numa pesquisa paralela, foram obtidas as mutações mais relevantes da proteína 5α-redutase, em particular mutações que levam à má formação do feto e a doenças graves (Tabela 3). No alinhamento foram marcados os aminoácidos que podem sofrer tais mutações (Fig.27), de forma a ajudar a selecionar o melhor molde. Deste alinhamento múltiplo, concluiu-se que a região mais conservada da 5α-redutase é o C-terminal, local onde se liga o NADPH (Fig. 28).

Em simultâneo, realizou-se uma pesquisa no PDB para identificar quais das proteínas com maior homologia analisadas no Jalview possuíam estrutura determinada, tendo-se tentado obter manualmente um possível alinhamento alternativo que permitisse modelar uma estrutura para a proteína alvo, mas sem sucesso.

Tabela 3. Mutações da 5α-redutase e mal–formações e doenças associadas.

A49T	Alto impacto na atividade da enzima, aumentando a conversão de T>DHT
V89L	Reduz a atividade da enzima em 30%
R227Q	Reduz a atividade da enzima
Q71X (codão terminal)	Pseudo hermafroditismo no homem (impede o desenvolvimento dos genitais dos homens)
A228T	Pseudo hermafroditismo no homem
Q126R	Pseudo hermafroditismo, completa inativação da enzima
R145W	Pseudo hermafroditismo
G158	Pseudo hermafroditismo
G183S	Diminui a afinidade da enzima para o NADPH
G196S	Diminui a afinidade de 5AR2 para o co-factor NADPH
H231R	Afecta a capacidade da enzima se ligar a testosterona
G203S	Pseudo hermafroditismo
Y235F	Pseudo hermafroditismo
R246W	Pseudo hermafroditismo
L113V	Reduz a atividade da enzima
L224H	Pseudo hermafroditismo

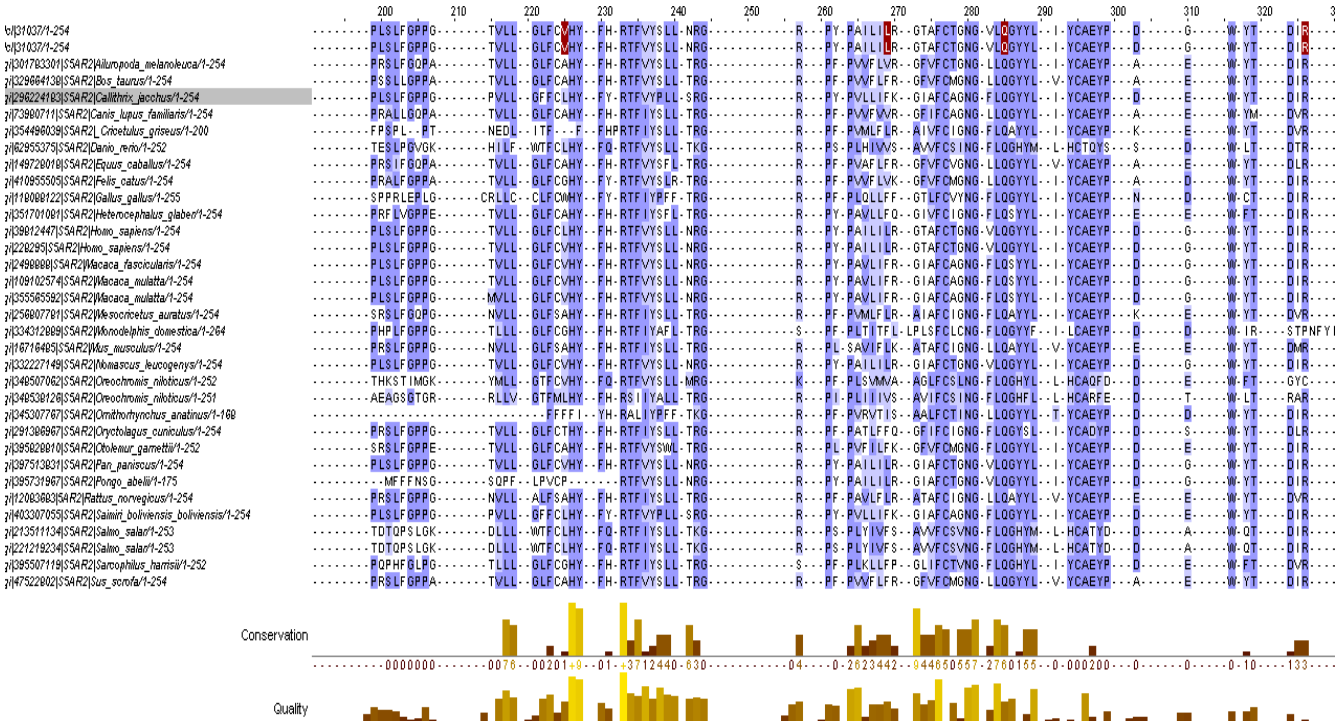


Figura 27. Resultados do alinhamento obtido com o Jalview. A vermelho encontram-se mutações pouco conservadas.

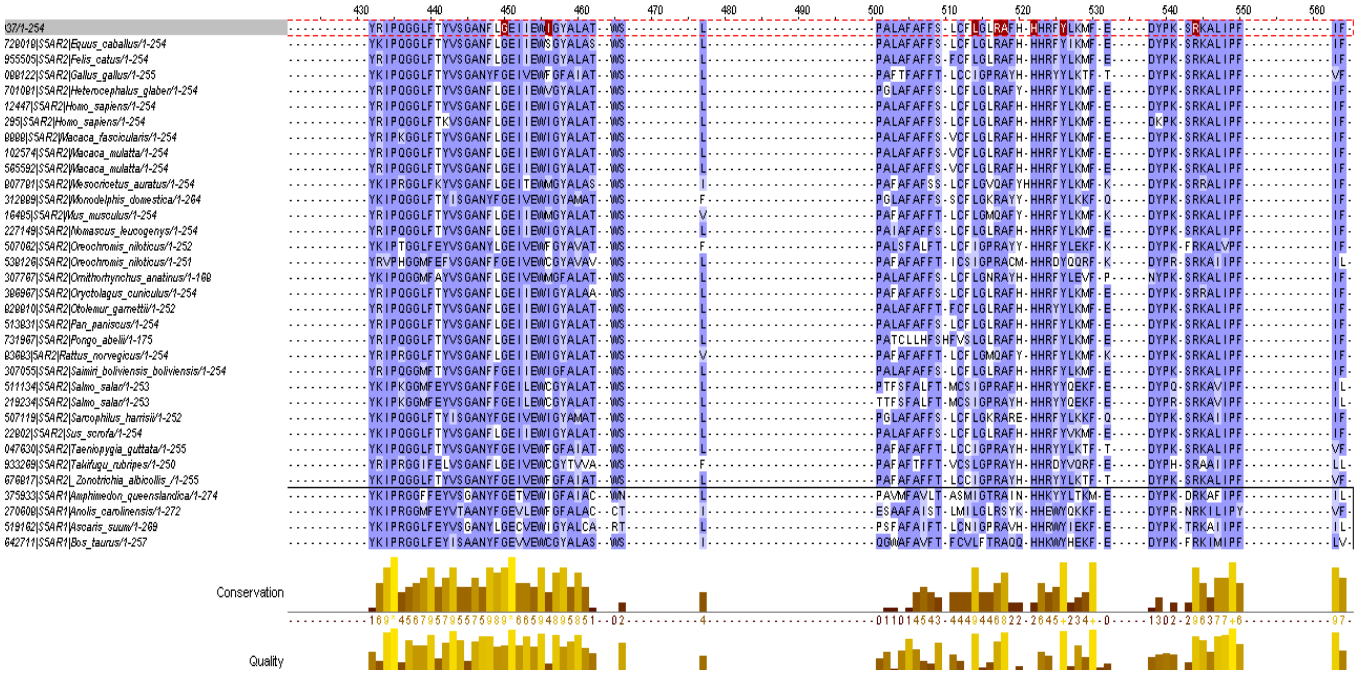


Figura 28. Alinhamento da região mais conservada da 5α-redutase, C-terminal local onde se liga o NADPH. A vermelho estão representados os resíduos com mutações da 5α-redutase.

1.2. Phyre

Uma vez que a utilização da combinação do BLAST com o Jalview não foi bem sucedida na identificação de um modelo de homologia, foi feita uma segunda tentativa com o programa Phyre. O Phyre é um dos programas mais utilizados para prever a estrutura das proteínas. Da nossa pesquisa, apenas se obteve um modelo com uma cobertura de 60% da sequência. No entanto, depois do alinhamento as sequências da proteína 5α-redutase e da proteína molde (PDB: 4A2N) (Fig. 29) observa-se que o grau de identidade sequencial é apenas de 10%, ou seja o melhor dos resultados no Phyre também não se revelou um bom candidato a molde. Além disso, a proteína selecionada para molde não possui a mesma atividade nem possui cofator.

```
1 c4a2nB_ Probab=99.86   E-value=9.2e-20  Score=128.39  Aligned_columns=175  Identities=10%

Q ss_pred             CHHHHHHHHHHHHHHHHHHHHHHCCCCCCC-----CHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCHHHHHHHHHH
Q ss_conf             124799999999999975566530168874-----0399999999999999999997137777631111158999999
Q 5_alpha_reduct   77 GPPGTVLLGLFCVHYFHRTFVYSLLNRGRP-----YPAILILRGTAFCTGNGVLQGYYLIYCAEYPDGWYTDIRFSLGVF  151 (254)
Q Consensus        77 ~~~~~ll~~l~~~HY~~R~~i~p~~~~~~~~~~~~~~~i~~~a~~F~~~ng~l~~~~l~~~~~~~~~~~~~~~~~iG~~  151 (254)
                      +.++.+++.++.+..+.+.+.+......+...     ....-......+-.+.......+..........+.......+|++
T Consensus         4 ~~w~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~g~~   83 (192)
T c4a2nB_           4 NLWKICFIVMFIIWVFVRKVYGTRAMKNKSKKKVRPNFEKSLVFLNFIGMVFLPLTAVFSSYLDSFNINLPDSIRLFALI   83 (192)
T ss_dssp             HHHHHHHHHHHHHHHHHHHHHHHHHSSSCEEECCSHHHHHHHHHHHHHHHTHHHHHHHSCSGGGGGCCCCCCHHHHHHHH
T ss_pred             CHHHHHHHHHHHHHHHHHHHHHHHHCCHHHHHHCCHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCHHHHHHHH
T ss_conf             689999999999999999997641004565310122368999999999999999998633204666543599999999


Q ss_pred             HHHHHHHHHHHHHHHHH---HCCCCCCCCCCCCCCHHHHHCCHHHHHHHHHHHHHHHHHCC-HHHHHHHHHHHHHHHH
Q ss_conf             99999999998799974---145877786323577468505845799999999999997164-69999999999998888
Q 5_alpha_reduct  152 LFILGMGINIHSDYILRQ---LRKPGEISYRIPQGGLFTYVSGANFLGEIIEWIGYALATWS-LPALAFAFFSLCFLGLR  227 (254)
Q Consensus       152 lf~~G~~~n~~sd~~L~~~~~lR~~~~~~~~~ip~gGlf~~v~~PnY~gEil~w~g~~l~~~s-~~~~~f~~~~~~l~~r  227 (254)
                      ++++|...+..++.++.+   .+.+.++++++.++|+|++|||||+||++.|+|.++++++ +.....+.....+..|
T Consensus        84 l~~~g~~~~~~a~~~l~~~~~~~~~~~~~~~~lv~~G~yr~vRHP~Y~g~~l~~~g~~l~~~~~~~l~~~~~~~~~~~~  163 (192)
T c4a2nB_          84 VTFLNIGLFTKIHKDLGNNWSAILEIKDGHKLVKEGIYKNIRHPMYAHLWLWVITQGIILSNWVVLIFGIVAWAILYFIR  163 (192)
T ss_dssp             HHHHHHHHHHHHHHHHGGGCCSSCCEETTCCCCCSSTTTTBSSHHHHHHHHHHHHHHHHHTCHHHHHHHHHHHHHHHH
T ss_pred             HHHHHHHHHHHHHHHHHHHCCCCCCCCCCCEECCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
T ss_conf             999999999999974232055556568762112446999840789999999999999832799999999999999961


Q ss_pred             HHHHHHHHHHHHH----HHCCCCCCCC
Q ss_conf             99999999977645----5304711047
Q 5_alpha_reduct  228 AFHHHRFYLKMFED----YPKSRKALIP  251 (254)
Q Consensus       228 a~~~~~wy~~~F~~~~~Y~~~r~alIP  251 (254)
                      +..+||...|+|+|    |.++++++||
T Consensus       164 ~~~EE~~l~~~fG~~Y~~Y~~~v~~~IP  191 (192)
T c4a2nB_         164 VPKEEELLIEEFGDEYIEYMGKTGRLFP  191 (192)
T ss_dssp             HHHHHHHHHHHHTHHHHHHHHCBSSSC
T ss_pred             CCHHHHHHHHCCHHHHHHHHHCCCCCC
T ss_conf             024479999856808999998575858 8
```

Figura 29. Alinhamento da proteína 5α-redutase e da proteína molde (PDB 4A2N) utilizando o programa Phyre.

1.3. Swiss-Model

Na pesquisa feita no servidor Swiss-Model, a proteína selecionada para molde foi a mesma que foi obtida com o Phyre, possuindo uma identidade sequencial apenas de 10.53% e um E-value de 1.90e-16. A estrutura modelada para a 5α-redutase pelo Swiss-Model utilizando como molde a proteína PDB: 4A2N (Fig. 30) é um modelo de cadeia única que no entanto apenas envolve os resíduos 77 a 252 da 5α-redutase (no total de 260 resíduos).

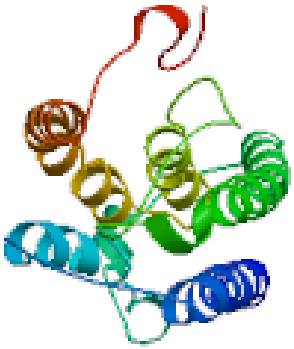


Figura 30. Modelo da proteína 5α-redutase e da proteína molde (PDB: 4A2N).

1.4. Por homologia com a proteína 5β-redutase

Numa pesquisa feita no PDB, observou-se que a proteína 5β-redutase possui estrutura cristalográfica e uma vez que esta possui a mesma função que a proteína 5α-redutase procedeu-se ao alinhamento desta com a proteína alvo no BLAST. No entanto, a 5β-redutase possui uma sequência de aminoácidos bastante mais longa (cerca de 100 aminoácidos) não sendo possível obter um bom alinhamento entre as proteínas, existindo apenas uma cobertura de 7%.

II. Rastreio virtual baseado em ligandos

1. Redução da dimensão dos dados e seleção de atributos

Com o objetivo de se identificar os descritores moleculares que levam um composto a ser um bom inibidor da 5α-redutase, neste trabalho foram analisadas para as isoenzimas 5α-R1 e 5α-R2, as seguintes classes de IC50: Muito Bom, Bom, e Médio (Tabela 3). No entanto para a 5α-R1 apenas foram analisadas as classes de IC50 Bom e Médio uma vez que a classe de IC50 Muito Bom apenas possui 4 compostos, não possuindo dados significativos para se proceder à análise desta classe.

1.1. Resultados da análise de componentes principais

O PCA é um método usado para reduzir o número de atributos usados para representar os dados. O benefício desta redução dimensional inclui o fornecimento de uma representação dos dados mais simples e uma classificação mais rápida.

A primeira componente principal (PC1) extraída é responsável pela quantidade máxima de variância total das variáveis observadas. A segunda componente principal (PC2) é responsável pela quantidade máxima de variância no conjunto de dados que não foram contabilizados pela primeira componente principal, e não será correlacionada com a primeira componente principal. As restantes componentes que são extraídas para análise exibem as mesmas duas características: cada componente é responsável pela quantidade máxima de variância do conjunto de dados que não foram contabilizados pelas componentes anteriores, não se correlacionando com as componentes anteriores. A cada nova análise das componentes principais os valores da variância são menores, por este motivo apenas se interpretam as primeiras componentes principais.

Neste trabalho, para a análise realizada com o método estatístico PCA no nosso trabalho foram interpretadas as primeiras 4 componentes principais, responsáveis por uma percentagem significativa e suficiente para explicar a maior parte da variância total dos dados, para ambas as isoenzimas. Na tabela 4 podemos observar que para a 5α-R1 na classe de IC50 Bom, as 4 primeiras componentes principais explicam aproximadamente 81,87% da

variância total, onde 41,64% da variância é explicada pela primeira componente principal. Na classe de IC50 Médio as primeiras 4 componentes principais descrevem 55,63% da variância total, explicando a primeira componente principal 41,87% da variância total.

Tabela 4. Percentagem da variância total em cada componente principal interpretada nas diferentes classes de IC50 na 5α-R1.

	Variância	
	Bom	**Médio**
PC1	41,64%	41,87%
PC2	23,47%	25,71%
PC3	11,26%	6,95%
PC4	5,5%	5,15%
Total	81,87%	55,63%

Quanto maior o fator de uma variável numa componente principal, maior a influência da variável nesse componente. Na 5α-R1, classe de IC50 Bom (Tabela 5), podemos observar os coeficientes com maior fator. Na primeira componente principal (PC1), componente mais significativa na interpretação dos dados, as variáveis mais influentes indicam que as propriedades alifáticas dos compostos são as mais relevantes, tendo também elevada influência a área da superfície molecular, o volume molecular e o número de átomos principalmente o número de átomos de hidrogénio. Nas restantes componentes principais (PC2, PC3 e PC4), pode-se destacar a influência do número de átomos de flúor e o número de dadores de hidrogénio.

Tabela 5. Análise de componentes principais da classe de IC50 Bom da 5α-R1.

Características	PC1	PC2	PC3	PC4
Número de átomos	0,236	-0,084	-0,083	0
Número de átomos (H)	0,238	0,01	-0,128	-0,02
Número de átomos alifaticos	0,242	0,001	0,059	-0,022
Número de ligações alifaticas	0,242	-0,019	0,076	-0,029
Número de átomos assimétricos	0,235	0,029	0,078	0,072
Número de anéis de carboalifáticos	0,225	0,01	0,163	-0,042
Número de anéis alifáticos unidos	0,219	0,062	0,176	-0,065
Área superficial (van der waals)	0,228	-0,114	-0,088	-0,016
Volume (van der waals)	0,215	-0,156	-0,062	0,006
Área de superfície polar	0,001	-0,256	0,259	-0,131
Número de átomos (N)	0,068	-0,097	0,301	-0,386
Número de átomos (F)	0,004	-0,081	0,319	0,307
ASA_P	-0,015	-0,17	0,298	0,259
Número de dadores de hidrogénio	0,023	-0,078	0,311	0,353

Quanto à classe de IC50 Médio, na tabela 6 pode-se observar que os descritores moleculares relacionados com as características alifáticas ainda possuem influência nas componentes principais, no entanto não de uma forma tão significativa como na classe de IC50 Bom. Nesta classe, também se pode observar influência das variáveis como a área de superfície molecular ASA, ASA+ e ASA_H. Por outro lado, e ao contrário do observado para a classe de IC50 Bom onde se destacava o número de dadores de hidrogénio, na classe de IC50 Médio destaca-se o número de aceitadores de hidrogénio.

Tabela 6. Análise das componentes principais da classe de IC50 Médio na 5α-R1.

Características	PC1	PC2	PC3	PC4
Número de átomos	0,239	0,027	0,087	-0,059
Número de átomos (C)	0,236	-0,056	0,065	0,001
Número de átomos (H)	0,221	0,09	0,14	-0,133
Massa exata	0,234	-0,058	-0,023	0,026
Polarizabilidade molecular	0,237	-0,051	0,064	-0,018
Número de átomos alifáticos	0,204	0,161	0,021	-0,01
Número de ligações alifáticas	0,209	0,155	0,01	0,02
Número de anéis alifáticos	0,137	0,248	-0,041	0,107
Número de anéis de carboalifáticos	0,166	0,204	-0,141	-0,085
Número de anéis alifáticos unidos	0,13	0,252	-0,025	0,134
Área de superfície molecular ASA	-0,143	0,153	-0,042	0,203
ASA+	0,208	-0,057	0,248	-0,063
Número de anéis	0,213	-0,002	-0,073	0,193
Volume (van der waals)	0,242	-0,016	0,06	-0,028
Número de anéis hetero	-0,096	-0,085	0,334	0,42
ASA_H	0,188	-0,115	0,276	-0,104
LogP	0,151	0,045	0,345	0,13
Número de átomos (N)	0,167	0,098	-0,104	0,228
Número de átomos (F)	0,005	-0,047	-0,166	0,245
Número de anéis heteroalifáticos	-0,107	0,06	0,274	0,484
Número de aceitadores de hidrogénio	0,13	-0,162	-0,128	0,225

A análise de componentes principais na 5α-R2 foi realizada para a classe de IC50 Muito Bom, classe de IC50 Bom e classe de IC50 Médio. Desta análise, e tal como acontecia na 5α-R1, as primeiras 4 componentes principais são suficientes para explicar a maior parte da variância dos dados. Na tabela 7 podemos observar que estas componentes explicam 72% da variância total para a classe de IC50 Muito Bom, onde 31% da variância é explicada pela primeira componente principal. Na classe de IC50 Bom as 4 primeiras componentes principais explicam 75% da variância total, sendo que a primeira componente principal explica 32% da variância. Já na classe de IC50 Médio 86% da variância é descrita pelas componentes principais, onde 40% da variância é explicada pela primeira componente principal.

Tabela 7. Percentagem da variância total em cada componente principal interpretada nas diferentes classes de IC50 na 5α-R2.

	Variância		
	Muito bom	**Bom**	**Médio**
PC1	31%	32%	40%
PC2	18%	23%	30%
PC3	13%	14%	10%
PC4	10%	6%	6%
Total	72%	75%	86%

Na classe de IC50 Muito Bom (5α-R2), podemos observar que as variáveis que mais influenciam a primeira componente principal são a área de superfície e volume (van der waals) (Tabela 8). As propriedades como a área de superfície molecular ASA+, ASA_H e as propriedades polarizabilidade molecular, massa molecular e o número de átomos nomeadamente o número de átomos de carbono têm também um elevado fator na primeira componente principal. Nas restantes componentes as variáveis mais influentes são as propriedades aromáticas, em contraste com o que acontecia na 5α-R1, onde as variáveis com propriedades alifáticas demonstravam ser as mais influentes nas características com maior relevância nos compostos que inibem a 5α-R1.

Tabela 8. Análise das componentes principais da classe de IC50 Muito Bom na 5α-R2.

Características	PC1	PC2	PC3	PC4
ASA_+	0,233	-0,136	0,029	0,113
ASA_H	0,218	-0,091	0,17	0,075
Número de átomos	0,245	-0,181	0,036	0,029
Número de átomos (C)	0,244	-0,112	0,181	-0,001
Massa exata	0,250	-0,028	-0,028	-0,135
Polarizabilidade molecular	0,257	-0,102	0,148	0,033
Área de superfície (van ser waals)	0,268	-0,124	0,008	-0,017
Volume (van der waals)	0,270	-0,114	0,034	-0,031
Número de átomos aromáticos	0,171	0,208	0,225	-0,001
Número de anéis aromáticos	0,158	0,213	0,238	0,035
Número de anéis aromáticos de 6 carbonos	0,17	0,203	0,221	-0,021
Número de átomos assimétricos	0,015	0,201	-0,107	0,212
Número de anéis carboaromáticos	0,17	0,203	0,221	-0,021
Número de átomos (N)	-0,009	0,046	0,207	-0,133
Número de anéis	0,039	-0,204	0,252	0,064
Número de anéis aromáticos de 5 carbonos	-0,041	0,071	0,104	0,277
Número de anéis aromáticos unidos	-0,037	0,1	0,161	0,242
Número de anéis hétero	0,038	0,077	-0,05	0,346
Número de anéis hétero alifáticos	0,075	0,044	-0,134	0,228
Número de anéis hétero aromáticos	-0,041	0,071	0,104	0,277

No entanto, na classe de IC50 Bom (5α-R2), as variáveis influentes na primeira componente principal são também as propriedades alifáticas (Tabela 9). Na quarta componente principal as variáveis influentes são as propriedades aromáticas, que possuem coeficientes com fator bastante elevado. Nas restantes componentes principais as variáveis que mais influenciaram estas componentes foram o número de átomos de azoto, área de superfície polar, número de anéis e o número de dadores de hidrogénio.

Tabela 9. Análise das componentes principais da classe de IC50 Bom na 5α-R2.

Características	PC1	PC2	PC3	PC4
Número de átomos alifáticos	0,215	-0,174	-0,064	-0,009
Número de ligações alifáticas	0,209	-0,191	0,018	-0,022
Número de anéis alifáticos	0,231	-0,093	0,172	-0,005
Número de anéis carboalifáticos	0,212	-0,098	-0,134	-0,094
Número de átomos (N)	0,139	0,008	0,234	0,198
Área de superfície polar	-0,113	0,159	0,250	-0,004
Número de anéis	-0,013	-0,124	0,373	0,003
Número de dadores de hidrogénio	0,045	0,089	0,294	0,071
Número de anéis aromáticos de 5 carbonos	-0,183	0,025	-0,084	0,433
Número de anéis aromáticos unidos	-0,183	-0,025	-0,084	0,433
Número de anéis hétero aromáticos	-0,183	-0,025	-0,084	0,433
Número de anéis hétero	0,081	-0,042	0,134	0,357

Para a classe de IC50 Médio (5α-R2), as variáveis influentes na primeira componente principal são as de propriedade alifáticas como o observado na tabela 10. Nas restantes componentes as variáveis influentes são o número de átomos de oxigénio, o número de átomos de flúor, número de aceitadores de hidrogénio, ASA_P, número de átomos em cadeia e área de superfície polar. Para a classe de IC50 Médio as variáveis relacionados com propriedades aromáticas dos compostos não têm influência sobre as componentes. Em suma, uma primeira análise dos dados utilizando a análise de compostos principais parece indicar que para que um composto seja um inibidor Muito Bom da 5α-R2, este deverá possuir propriedades aromáticas.

Tabela 10. Análise das componentes principais da classe de IC50 Médio na 5α-R2.

Características	PC1	PC2	PC3	PC4
Número de átomos alifáticos	0,220	-0,116	0,172	0,03
Número de ligações alifáticas	0,223	-0,12	0,145	-0,012
Número de anéis alifáticos	0,251	-0,023	0,011	-0,083
Número de átomos assimétricos	0,232	-0,053	0,031	-0,029
Número de anéis carboalifáticos	0,234	-0,028	0,119	-0,002
Número de anéis alifáticos unidos	0,244	-0,007	0,011	-0,027
Número de anéis unidos	0,209	-0,07	-0,13	0,016
Número de átomos (O)	-0,21	-0,011	0,228	-0,115
Número de átomos (F)	-0,095	0,027	0,283	-0,237
Número de aceitadores de hidrogénio	-0,175	-0,068	0,223	-0,168
ASA_P	-0,217	0,024	0,203	-0,17
Número de átomos em cadeia	-0,044	-0,179	0,292	0,255
Área de	-0,161	-0,057	0,269	-0,299

superfície polar				

1.2. Seleção de atributos

A seleção de atributos é o processo de seleção de um subconjunto de variáveis avaliando apenas as características relevantes dos dados. Neste trabalho procedeu-se à análise da seleção de atributos de ambas as isoenzimas 5α-R1 e 5α-R2 para as classes de IC50 Muito Bom, Bom e Médio, tal como executado para a análise de componentes principais. No entanto, enquanto na análise de PCA foram analisadas apenas as características intrínsecas dos dados sem qualquer outra informação adicional, o método de seleção de atributos foi aplicado para determinar quais os descritores moleculares que mais contribuem para a definição da classe de IC50 dos compostos.

Na tabela 11 observa-se o conjunto de variáveis relevantes para a classe de IC50 Bom e Médio da 5α-R1. Para a classe de IC50 Bom, o melhor modelo encontrado apresenta um erro de 0,6. Neste modelo constam das seis variáveis: a área de superfície polar, o número de anéis, a área de superfície (van der waals), o volume (van der waals), o LogP e o número de aceitadores de hidrogénio. Na classe de IC50 Médio, o modelo com o menor erro encontrado tem um erro de 0,375. O conjunto de seis variáveis que compõem o modelo com o menor erro associado é composto por: ASA+, número de anéis, área de superfície, volume, LogP e número de dadores de hidrogénio.

Segundo este método podemos verificar para a 5α-R1, que as diferenças entre a classe de IC50 Bom e a classe de IC50 Médio encontram-se em duas variáveis, sendo que para a classe de IC50 Médio, as variáveis área de superfície polar e número de aceitadores de hidrogénio encontradas na classe de IC50 Bom são substituídas pelas variáveis ASA e número de dadores de hidrogénio. Indicando que as variáveis ASA+ e o número de dadores de hidrogénio parecem diminuir a capacidade de um composto se tornar um bom inibidor da 5α-R1.

Tabela 11. Conjunto de variáveis presentes no modelo com o menor erro associado da classe de IC50 Bom e Médio na 5α-R1.

Bom (erro: 0,6)	**Médio (erro: 0,375)**
Área de superfície polar	ASA+
Número de anéis	Número de anéis
Área de superfície (van der waals)	Área de superfície (van der waals)
Volume (van der waals)	Volume (van der waals)
LogP	LogP
Número de aceitadores de hidrogénio	Número de dadores de hidrogénio

Para a 5α-R2 foram analisadas as classes de IC50 Muito Bom, Bom e Médio, com o objetivo de obter o conjunto de variáveis relevantes associado a cada uma destas classes (Tabela 12). Na classe de IC50 Muito Bom o conjunto de variáveis que compõem o modelo com o menor erro (erro de zero) é constituído por dezoito variáveis, onde cinco variáveis são referentes a propriedades aromáticas e quatro são referentes à área de superfície molecular.

Na classe de IC50 Bom o conjunto de variáveis do modelo com menor erro possui um erro de 0,5. Neste conjunto de seis variáveis consta a área de superfície polar, ASA+, LogP, número de aceitadores de hidrogénio, ASA_H e a área de superfície. As variáveis área de superfície polar, ASA+, ASA_H e o número de aceitadores de hidrogénio encontram-se em ambas as classes de IC50 Muito Bom e Bom, ou seja, do conjunto de variáveis encontrado para a classe de IC50 Muito Bom, estas serão as variáveis mais influentes na obtenção de um inibidor da 5α-R2 Muito Bom.

Quanto à classe de IC50 Médio, o conjunto de descritores que produz o modelo com menor erro possui um erro de 0,25. Este conjunto de quatro variáveis é composto por ASA_P, área de superfície, número de anéis hétero-aromáticos e número de anéis aromáticos. Para a classe de IC50 Médio são selecionadas três variáveis: o número de anéis hétero aromáticos, ASA_P e o número de anéis aromáticos unidos, que também estão presentes na classe de IC50 Muito Bom e uma variável, a área de superfície, que também se encontram no conjunto de variáveis da classe de IC50 Bom.

Tabela 12. Conjunto de variáveis presentes no modelo com o menor erro associado da classe de IC50 Muito Bom, Bom e Médio na 5α-R2.

Muito Bom (erro:0)	**Bom (erro:0,5)**	**Médio (erro:0,25)**
ASA+	Área de superfície polar	ASA_P
ASA_H	ASA+	Área de superfície (van der waals)
Número de anéis aromáticos	LogP	Número de anéis hétero aromáticos
Número de anéis aromáticos de 6 carbonos	Número de aceitadores de hidrogénio	Número de anéis aromáticos unidos
Número de anéis aromáticos de 5 carbonos	ASA_H	
Número de anéis aromáticos unidos	Área de superfície (van der waals)	
Número de anéis hétero		
Número de anéis hétero alifáticos		
Número de anéis hétero aromáticos		
Área de superfície polar		
Área de superfície molecular ASA		
Número de anéis unidos		
Número de dadores de hidrogénio		
Número de aceitadores de hidrogénio		
ASA_P		
Número de anéis carbo alifáticos		
Número de anéis alifáticos unidos		
Número de ligações alifáticas		

2. Descrição dos modelos de classificação

2.1. Árvores de decisão

Uma árvore de decisão representa um conjunto de regras que seguem uma hierarquia de classes e valores utilizados para classificar novos elementos. No programa KNIME procedeu-se à análise dos descritores moleculares dos compostos em todas as classes de IC50 para a 5α-R1 e para a 5α-R2 utilizando o algoritmo de aprendizagem de *Random Forest*. Os modelos de classificação de *Random Forest* são constituídos por 10 árvores de decisão, cada uma construída considerando 6 atributos aleatórios. Em seguida o conjunto de árvores de decisão (*Random Forest*) obtido foi utilizado para re-classificar o conjunto de compostos para testar a qualidade dos modelos de classificação produzidos. Na tabela 13 pode-se observar a contagem dos compostos para as classes de IC50 Bom, classe de IC50 Médio e classe de IC50 Mau, para a 5α-R1. Para a classe de IC50 Bom, todos os compostos bons foram re-classificados bons. Para a classe de IC50 Médio, dos 88 compostos com IC50 Médio, 2 compostos foram considerados pelo algoritmo de aprendizagem como maus e 86 compostos como pertencentes à classe de IC50 Médio. Na classe de IC50 Mau, dos 157 compostos maus, apenas 1 composto foi re-classificado como pertencendo à classe de IC50 Médio.

Tabela 13. Re-classificação dos compostos em análise que inibem a 5α-R1 aplicando os modelos de classificação baseados em árvores de decisão. Os conjuntos de treino e teste do algoritmo são o conjunto de compostos em análise.

	Bom	**Médio**	**Mau**	**Total**
Bom	53 (100%)	0	0	53
Médio	0	86 (97,7%)	2 (2,3%)	88
Mau	0	1 (0,64%)	156 (99,7%)	157

Adicionalmente, e como se pretende comparar os resultados dos dois métodos de classificação utilizados neste trabalho (Árvores de decisão e SVM), este algoritmo de aprendizagem foi avaliado utilizando o método estatístico de validação cruzada. Novos modelos de classificação de árvores de decisão foram gerados utilizando validação cruzada 5 vezes. Na tabela 14 observa-se que para a classe de IC50 Bom, dos 53 compostos neste

conjunto o algoritmo de aprendizagem apenas re-classifica 35 compostos como pertencentes à classe de IC50 Bom, sendo que dos restantes 18 compostos, 11 compostos foram considerados como pertencentes à classe de IC50 Médio e 7 compostos foram atribuídos à classe de IC50 Maus. Dos 86 compostos da classe de IC50 Médio, o algoritmo de aprendizagem, classifica 38 dos compostos como pertencentes à classe de IC50 Médio Dos 48 compostos restantes, 18 compostos foram atribuídos à classe de IC50 Bom e 32 compostos foram atribuídos à classe de IC50 Mau. Para a classe de IC50 Mau, dos 156 compostos, 133 foram atribuídos à classe de IC50 Mau. Dos 23 compostos restantes, 16 foram considerados como pertencentes à classe de IC50 Médio e 8 compostos foram atribuídos à classe de IC50 Bom.

Tabela 14. Re-classificação dos compostos em análise que inibem a 5α-R1 aplicando modelos de classificação baseados em árvores de decisão, construídas utilizando validação cruzada 5 vezes.

	Bom	**Médio**	**Mau**	**Total**
Bom	35 (66%)	11 (21%)	7 (13%)	53
Médio	18 (20%)	38 (43%)	32 (37%)	88
Mau	8 (5%)	16 (10%)	133 (85%)	157

Para a 5α-R2 os dados foram analisados pelo mesmo algoritmo árvores de decisão, e tal como se pode observar na tabela 15, a re-classificação dos compostos para as classes de IC50, dos 107 compostos pertencentes à classe de IC50 Muito Bom, 105 compostos classificados como tal, 1 composto foi classificado como Bom e 1 composto foi atribuído à classe de IC50 Mau. Segundo o algoritmo de aprendizagem, todos os compostos da classe de IC50 Bom foram corretamente classificados como fazendo parte da classe de IC50 Bom. Na classe de IC50 Médio dos 48 compostos 1 composto foi atribuído à classe de IC50 Bom e 47 compostos foram atribuídos à classe de IC50 Médio. Na classe de IC50 Mau todos os 151 compostos foram corretamente classificados pelo algoritmo de aprendizagem.

Tabela 15. Re-classificação dos compostos em análise que inibem a 5α-R2 aplicando os modelos de classificação baseados em árvores de decisão. Os conjuntos de treino e teste do algoritmo são o conjunto de compostos em análise.

	Muito bom	**Bom**	**Médio**	**Mau**	**Total**
Muito bom	105 (98%)	1 (0,9%)	1 (0,9%)	0	107
Bom	0	48 (100%)	0	0	48
Médio	0	1 (0,9%)	47 (98%)	0	48
Mau	0	0	0	151 (100%)	151

A avaliação do algoritmo de aprendizagem utilizando validação cruzada pode ser observada na tabela 16. Para a classe de IC50 Muito Bom dos 107 compostos pertencentes a esta classe, o algoritmo de aprendizagem atribuiu 91 compostos à classe de IC50 Muito Bom. Dos 16 compostos restantes, 10 compostos foram atribuídos à classe de IC50 Bom, 2 compostos foram atribuídos à classe de IC50 Médio e 4 compostos à classe de IC50 Mau. Para a classe de IC50 Bom, 29 compostos foram realmente considerados como pertencentes à classe de IC50 Bom. Dos 19 compostos restantes, 7 compostos foram atribuídos à classe de IC50 Muito Bom, 6 compostos foram atribuídos à classe de IC50 Médio e 6 compostos foram atribuídos à classe Mau. Dos 48 compostos pertencentes à classe de IC50 Médio, 15 compostos foram atribuídos à classe de IC50 Médio, 8 compostos foram atribuídos à classe de IC50 Bom, 4 compostos foram atribuídos à classe de IC50 Muito Bom e 21 compostos foram atribuídos à classe de IC50 Mau. Para a classe de IC50 Mau dos 151 compostos, 133 compostos foram classificados como pertencentes a esta classe, enquanto dos 18 compostos restantes, 7 foram atribuídos à classe de IC50 Médio, 6 compostos atribuídos à classe de IC50 Bom e 5 compostos foram atribuídos à classe de IC50 Muito Bom.

Tabela 16. Re-classificação dos compostos em análise que inibem a 5α-R2 aplicando modelos de classificação baseados em árvores de decisão, construídas utilizando validação cruzada 5 vezes.

	Muito bom	**Bom**	**Médio**	**Mau**	**Total**
Muito bom	91 (85%)	10 (9%)	2 (2%)	4 (4%)	107
Bom	7 (15%)	29 (60%)	6 (12,5)	6 (12,5%)	48
Médio	4 (4%)	8 (17%)	15 (31%)	21 (44%)	48
Mau	5 (3%)	6 (4%)	7 (5%)	133 (88%)	151

2.2. Máquinas de vetores de suporte (SVM)

Esta técnica baseia-se na procura de uma fronteira ou um hiperplano que separa as classes. Através deste algoritmo de aprendizagem procedeu-se à análise dos descritores moleculares calculados para os compostos das isoenzimas 5α-R1 e 5α-R2 para tentar perceber como estes podem ser utilizados para determinar as classes de IC50 para novos inibidores da 5α-R1 e 5α-R2. Na tabela 17 observa-se a re-classificação dos compostos para a 5α-R1. Na classe de IC50 Bom, o algoritmo de aprendizagem SVM, 36 dos compostos foram considerados como sendo pertencentes à classe de IC50 Bom, 8 compostos foram designados como pertencentes à classe de IC50 Médio e 9 compostos como pertencentes à classe de IC50 Mau.

Para a classe de IC50 Médio, 41 compostos foram considerados pelo algoritmo de aprendizagem como pertencentes à classe de IC50 Médio, 10 compostos foram atribuídos à classe de IC50 Bom e 37 compostos da classe de IC50 Mau. Na classe de IC50 Mau, 146 compostos foram identificados pelo algoritmo de aprendizagem como pertencentes à classe de IC50 Mau e 11 compostos como pertencentes à classe de IC50 Médio.

Tabela 17. Re-classificação dos compostos em análise que inibem a 5α-R1 aplicando os modelos de classificação baseados em SVM. Os conjuntos de treino e teste do algoritmo são o conjunto de compostos em análise.

	Bom	**Médio**	**Mau**	**Total**
Bom	36 (68%)	8 (15%)	9 (17%)	53
Médio	10 (12%)	41 (46%)	37 (42%)	88
Mau	0	11 (7%)	146 (93%)	157

Quando construímos os modelos de classificação SVM, utilizando validação cruzada, para a classe de IC50 Bom, 32 compostos foram considerados como pertencentes à classe de IC50 Bom, 11 compostos foram atribuídos à classe de IC50 Médio e 10 compostos foram atribuídos à classe de IC50 Mau (Tabela 18). Na classe de IC50 Médio, 33 compostos foram classificados com IC50 Médio, 12 compostos foram atribuídos à classe de IC50 Bom e 48 compostos foram atribuídos à classe de IC50 Mau. Para a classe de IC50 Mau, 138 compostos foram considerados como pertencentes à classe de IC50 Mau, 17 foram atribuídos à classe de IC50 Médio e 2 compostos foram atribuídos como IC50 Bom.

Tabela 18. Re-classificação dos compostos em análise que inibem a 5α-R1 aplicando modelos de classificação baseados em SVM, construídas utilizando validação cruzada 5 vezes.

	Bom	**Médio**	**Mau**	**Total**
Bom	32 (60%)	11 (21%)	10 (19%)	53
Médio	12 (14%)	33 (38%)	43 (49%)	88
Mau	2 (1%)	17 (19%)	138 (88%)	157

Na tabela 19 são apresentados os resultados da avaliação dos modelos de classificação SVM quando foi utilizado o conjunto total de compostos que inibem a 5α-R2 como conjunto de treino e de teste. Na classe de IC50 Muito Bom, 97 compostos foram corretamente classificados, 5 compostos foram atribuídos à classe de IC50 Bom, 1 composto foi atribuído à classe de IC50 Médio e 4 compostos foram atribuídos à classe de IC50 Mau. Para a classe de IC50 Bom, dos 48 compostos tidos como pertencentes à classe de IC50 Bom, 30 compostos foram atribuídos à classe de IC50 Bom, 8 compostos foram atribuídos à classe de IC50 Muito Bom, 1 composto foi atribuído à classe de IC50 Médio e 9 compostos foram atribuídos à classe de IC50 Mau. Na classe de IC50 Médio, dos 48 compostos designados como médios, 13 compostos foram considerados pelo algoritmo pertencentes a esta classe, 24 compostos foram atribuídos à classe de IC50 Mau, 6 compostos foram considerados como pertencentes à classe de IC50 Bom e 5 compostos foram atribuídos à

classe de IC50 Muito Bom. Para a classe de IC50 Mau, 140 compostos foram designados pelo algoritmo de aprendizagem como maus, 4 compostos foram considerados como pertencentes á classe de IC50 Médio, 6 compostos foram atribuídos à classe de IC50 Bom e 1 composto foi atribuído à classe de IC50 Muito Bom.

Tabela 19. Re-classificação dos compostos em análise que inibem a 5α-R2 aplicando os modelos de classificação baseados em SVM. Os conjuntos de treino e teste do algoritmo são o conjunto de compostos em análise.

	Muito bom	**Bom**	**Médio**	**Mau**	**Total**
Muito bom	97 (91%)	5 (5%)	1 (0,9%)	4 (4%)	107
Bom	8 (17%)	30 (63%)	1 (2%)	9 (19%)	48
Médio	5 (10%)	6 (12%)	13 (27%)	24 (50%)	48
Mau	1 (0,7%)	6 (4%)	4 (3%)	140 (93%)	151

A avaliação do algoritmo de aprendizagem pela validação cruzada para os compostos que inibem a 5α-R2 pode ser observada na tabela 20. Para a classe de IC50 Muito Bom, 93 compostos classificados como pertencendo à classe de IC50 Muito Bom. Dos restantes compostos, 4 compostos foram considerados como bons, 2 compostos como médios e 8 compostos como maus. Na classe de IC50 Bom, 27 compostos foram atribuídos à classe de IC50 Bom, 7 compostos foram considerados como pertencentes à classe de IC50 Muito Bom, 2 compostos foram atribuídos à classe de IC50 Médio e 12 compostos foram atribuídos à classe de IC50 Mau. Para a classe de IC50 Médio, apenas 7 em 48 compostos foram classificados como médios. Dos restantes compostos, 5 foram atribuídos à classe de IC50 Muito Bom, 10 compostos foram atribuídos à classe de IC50 Bom e 26 compostos foram atribuídos à classe de IC50 Mau. Para a classe de IC50 Mau, 136 foram corretamente classificados. Dos restantes compostos, 4 compostos foram atribuídos à classe de IC50 médio, 8 compostos foram atribuídos à classe de IC50 Bom e 3 compostos foram atribuídos à classe de IC50 Muito Bom.

Tabela 20. Re-classificação dos compostos em análise que inibem a 5α-R2 aplicando modelos de classificação baseados em SVM, construídas utilizando validação cruzada 5 vezes.

	Muito bom	**Bom**	**Médio**	**Mau**	**Total**
Muito bom	93 (87%)	4 (4%)	2 (2%)	8 (8%)	107
Bom	7 (15%)	27 (56%)	2 (4%)	12 (25%)	48
Médio	5 (10%)	10 (21%)	7 (16%)	26 (51%)	48
Mau	3 (2%)	8 (5%)	4 (3%)	136 (90%)	151

Através das métricas de desempenho da validação é possível avaliar a eficiência dos algoritmos de aprendizagem. Segundo as métricas de desempenho, sensibilidade e precisão, os algoritmos de aprendizagem demonstraram conter ambos uma eficiência semelhante. Como observado na tabela 21 e na tabela 22, referente à 5α-R1 e à 5α-R2, os algoritmos de aprendizagem árvores de decisão e SVM apresentam um desempenho semelhante.

Tabela 21. Valores de sensibilidade, precisão, especificidade e F-score dos algoritmos de aprendizagem SVM e árvores de decisão pela validação cruzada para a 5α-R1.

SVM				**A.D**			
Sensibilidade	**Precisão**	**Especificidade**	**F-score**	**Sensibilidade**	**Precisão**	**Especificidade**	**F-score**
0,375	0,541	0,867	0,443	0,432	0,585	0,871	0,497
0,604	0,696	0,943	0,646	0,66	0,574	0,894	0,614
0,879	0,723	0,624	0,793	0,847	0,773	0,723	0,809

Tabela 22. Valores de sensibilidade, precisão, especificidade e F-score dos algoritmos de aprendizagem SVM e árvores de decisão pela validação cruzada para a 5α-R2.

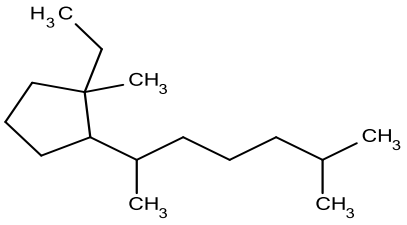
SVM				**A.D**			
Sensibilidade	**Precisão**	**Especificidade**	**F-score**	**Sensibilidade**	**Precisão**	**Especificidade**	**F-score**
0,869	0,861	0,939	0,865	0,85	0,85	0,935	0,85
0,562	0,551	0,928	0,557	0,604	0,547	0,922	0,574
0,146	0,467	0,974	0,222	0,312	0,5	0,951	0,385
0,901	0,747	0,773	0,817	0,881	0,881	0,847	0,844

3. Análise estrutural

Apesar dos métodos baseados nos descritores moleculares serem computacionalmente simples e eficazes na prática, possuem várias falhas, uma das mais importantes é a incapacidade de identificar as semelhanças estruturais entre as moléculas pequenas, muito importante para os químicos para compreender e sintetizar as moléculas.

A pesquisa de subestruturas máximas comuns (MCS) para a 5α-R1 e 5α-R2 foi realizada para as classes de IC50 Muito Bom, Bom e Médio usando a biblioteca LibMCS da ChemAxon. Na classe de IC50 Muito Bom para a 5α-R1, foi obtida 1 subestrutura comum para os 4 compostos pertencentes a esta classe (Tabela 23).

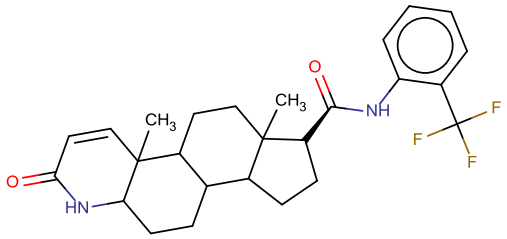
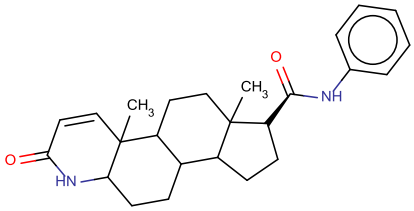
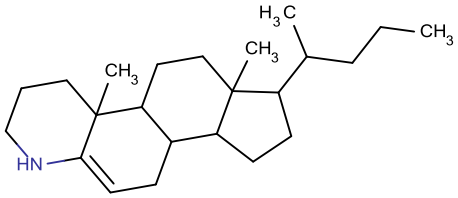
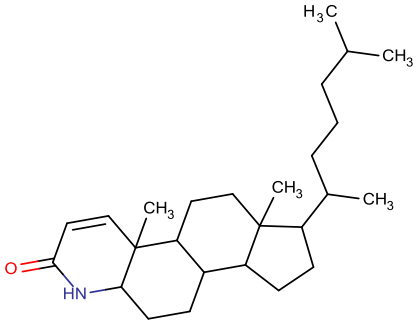
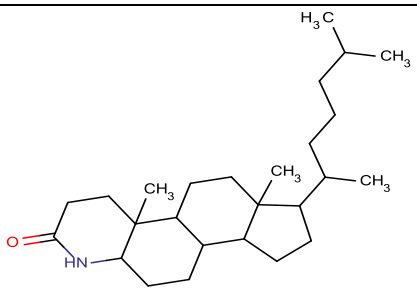
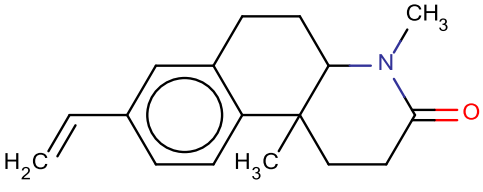
Tabela 23. Subestrutura máxima comum (MCS) dos compostos da classe IC50 Muito Bom para a 5α-R1.

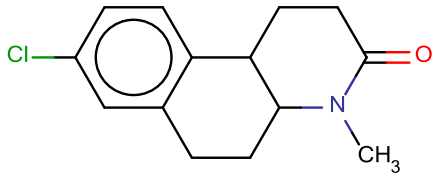
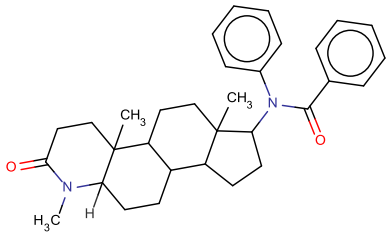
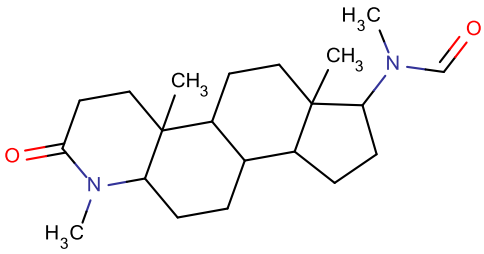
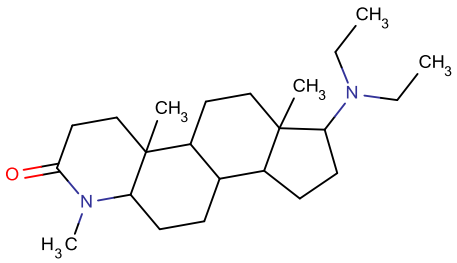
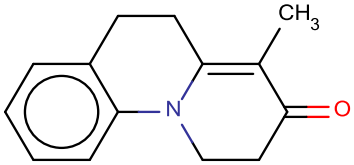
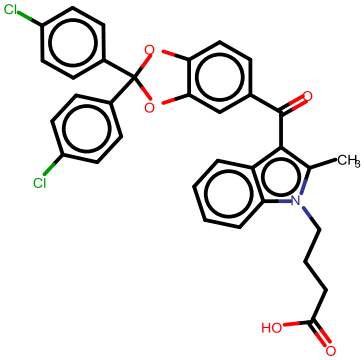
ID	MCS	# compostos
1		4

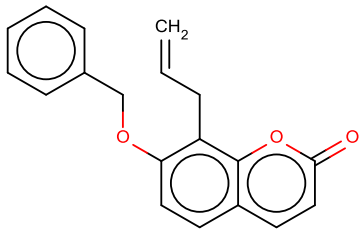
Para os 53 compostos pertencentes à classe de IC50 Bom, foram obtidas 13 MCS. Na tabela 24 podemos verificar que a maioria das subestruturas comuns encontradas são esteroides. As subestruturas esteroides encontradas são 4-azasteróides. Destas, as subestruturas 1 e 2 são muito semelhantes apenas variando na adição de um grupo trifluor.

A subestrutura 7 é um composto não esteroide, o composto LY-191,704, conhecido como bexlosteride e que é um análogo do 4-azasteróide, indicado como possuidor de potentes propriedades inibidoras da 5α-R1 (Salvador et al., 2013). A subestrutura 6 é semelhante à subestrutura 7, variando na substituição do cloro.

Tabela 24. Subestruturas máximas comuns (MCS) dos compostos da classe IC50 Bom para a 5α-R1.

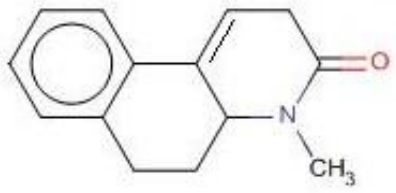
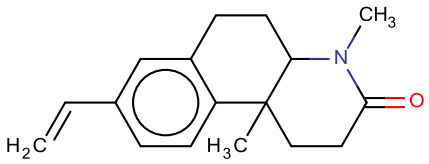
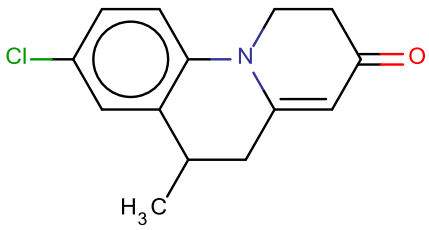
ID	MCS	# compostos
1		4
2		4
3		5
4		3
5		8
6		2

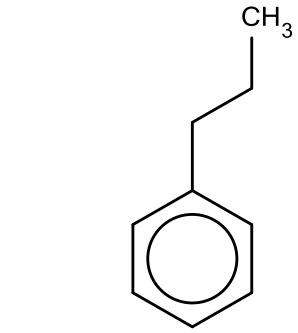
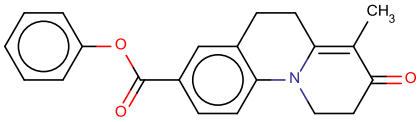
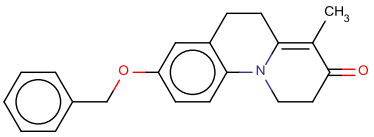
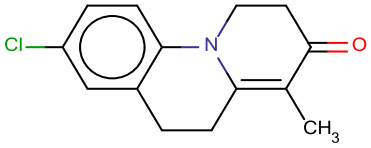
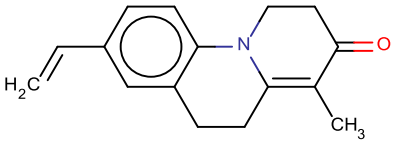
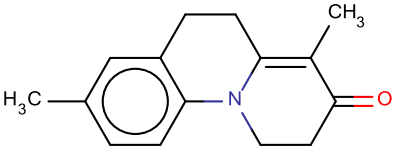
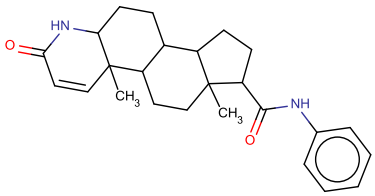
7		2
8		2
9		15
10		2
11		3
12		2

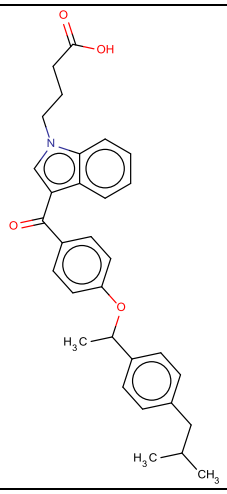
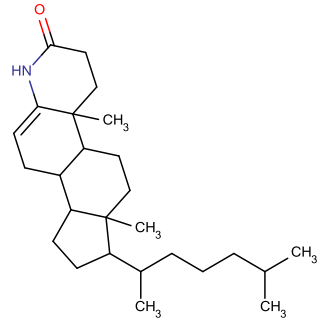
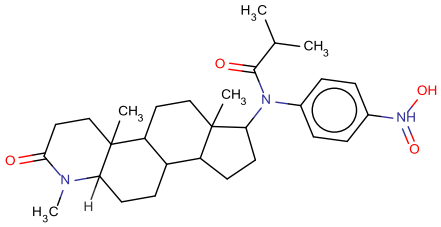
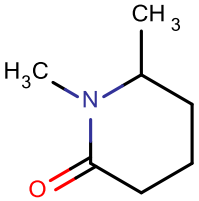
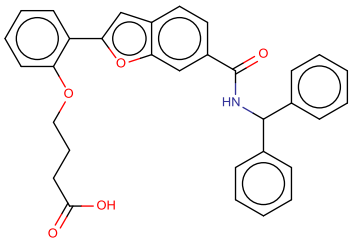
13		1

Para os 88 compostos incluídos na classe de IC50 Médio, foram obtidas 15 MCS. Na tabela 25 podemos observar que 29 dos compostos possuem em comum a subestrutura 10. Esta subestrutura é um esteroide, nomeadamente um 4-azasteróide. No entanto, a maioria das subestruturas comuns encontradas para esta classe são subestruturas não esteroides. A subestrutura 1 e 2 são semelhantes ao composto LY-191,704 encontrado como subestrutura comum para alguns dos compostos da classe IC50 Bom. A subestrutura 1 difere do composto LY-191,704 pela ausência do cloro no anel de benzeno, a subestrutura 3 possui alterações na mesma posição da ligação do cloro ao anel de benzeno. Quanto às subestruturas 5, 6, 7, 8 e 9 são subestruturas não esteroides e diferem entre si por alterações dos grupos químicos situadas na mesma ligação ao anel benzeno.

Tabela 25. Subestruturas máximas comuns (MCS) dos compostos da classe IC50 Médio para a 5α-R1.

ID	MCS	# compostos
1		2
2		3
3		2

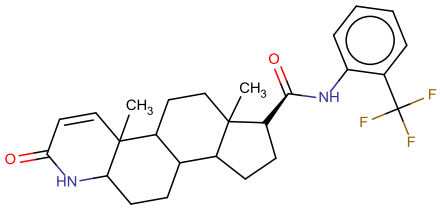
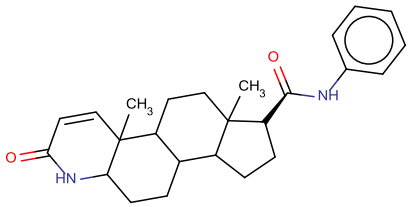
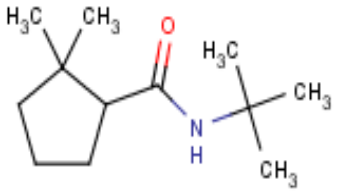
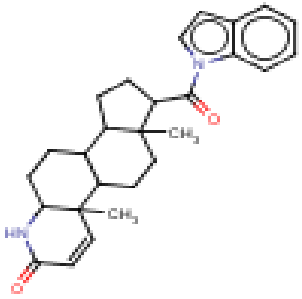
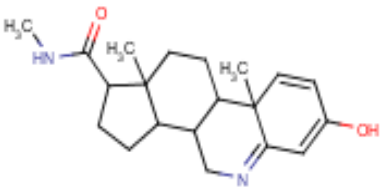
4		5
5		2
6		3
7		2
8		2
9		5
10		29

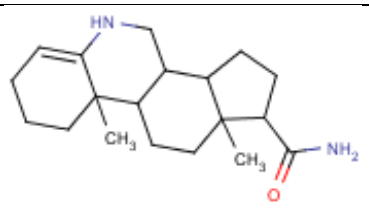
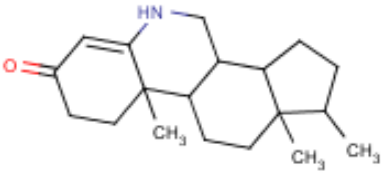
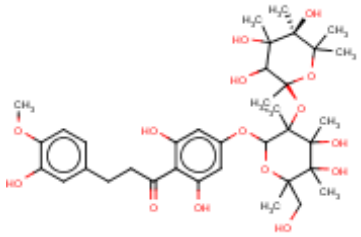
11		4
12		5
13		2
14		21
15		6

Na classe de IC50 Muito Bom para a 5α-R2 pode-se verificar na tabela 26 que a maioria das subestruturas comuns encontradas são esteroides nomeadamente 4-azasteróides e 6-azasteróides. As subestruturas 1 e 2 são muito semelhantes apenas variando na adição de um grupo trifluor ao anel de benzeno. As subestruturas 1, 2 e 4 são 4-azasteroides com variações na cadeia lateral que se liga ao C_{17}. As subestruturas 6-

azasteroides 5, 6 e 7 possuem variações nas cadeias laterais que se ligam ao C_{17} e C_3. A subestrutura 6 é a subestrutura com maior número de compostos associados.

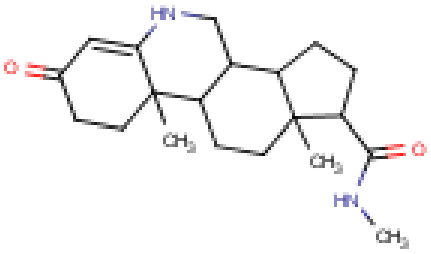
Tabela 26. Subestruturas máximas comuns (MCS) dos compostos da classe IC50 Muito Bom para a 5α-R2.

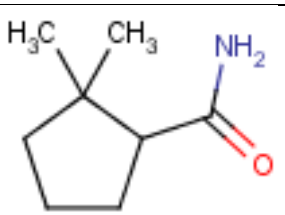
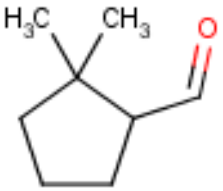
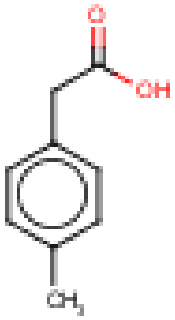
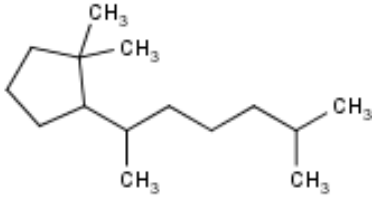
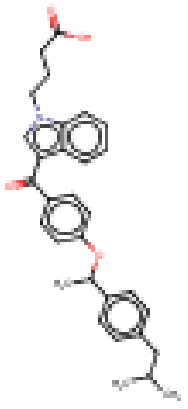
ID	MCS	# compostos
1		7
2		24
3		4
4		2
5		4

6		51
7		14
8		1

Na classe de IC50 Bom as subestruturas 1 e 2 abrangem uma grande parte dos compostos pertencentes a esta classe (Tabela 27), tendo 15 dos compostos a subestrutura comum 1 que é um 6-azasteroide e 23 dos compostos desta classe possuem em comum a subestrutura 2.

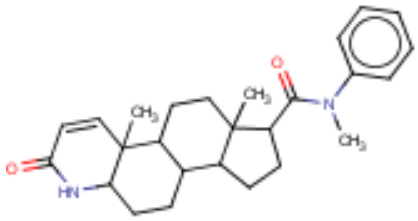
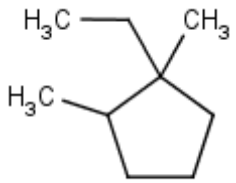
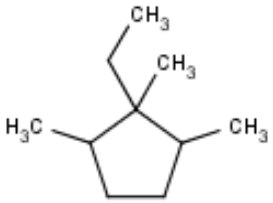
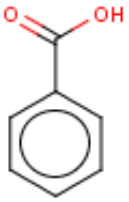
Tabela 27. Subestruturas máximas comuns (MCS) dos compostos da classe IC50 Bom para a 5α-R2.

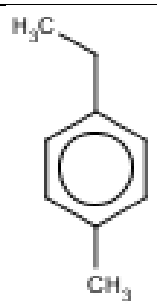
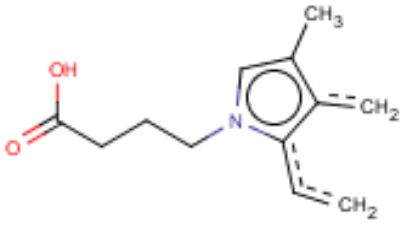
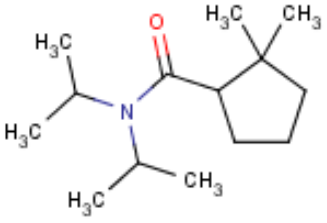
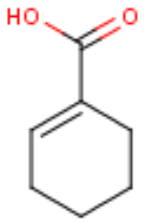
ID	MCS	# compostos
1		15

2		23
3		5
4		2
5		2
6		1

Na classe de IC50 Médio, como se pode observar na tabela 28, não existe nenhuma subestrutura que possua um número de compostos relevante. A subestrutura 1 é um 4-azasteroide, 6 compostos possuem esta estrutura em comum, as outras subestruturas são não esteroides.

Tabela 28. Subestruturas máximas comuns (MCS) dos compostos da classe IC50 Médio para a 5α-R2.

ID	MCS	# de compostos
1		7
2		6
3		9
4		5

5		7
6		6
7		5
8		2

4. Pesquisa de novos compostos e previsão da sua atividade

Com o intuito de se perceber a importância das subestruturas derivadas a partir de um maior número de compostos através análise de estruturas na relação com a sua atividade nas isoenzimas 5α-R1 e 5α-R2, procedeu-se a uma pesquisa e análise de outros compostos que possuíssem as subestruturas de interesse (Anexos, Tabela 40 a 45). Esta pesquisa foi

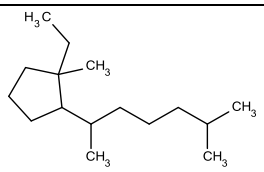
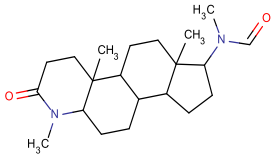
realizada para as subestruturas encontradas nas classes de IC50 Muito Bom, Bom e Médio em ambas as isoenzimas.

Assim, depois de se ter obtido o conjunto dos novos compostos para cada subestrutura de interesse a partir da ChEMBL, foram calculados os 40 descritores moleculares para cada composto, e em seguida, procedeu-se no KNIME à análise destes compostos, com o intuito de se prever a sua classe de atividade. A análise foi realizada em ambos os algoritmos árvores de decisão e SVM, uma vez que segundo os nossos dados ambos demonstraram ter um desempenho semelhante.

Na isoenzima 5α-R1 foi analisada a única subestrutura obtida para a classe de IC50 Muito Bom (subestrutura 1, Tabela 23) e a subestrutura com maior número de compostos associados na classe de IC50 Bom (subestrutura 9, Tabela24). Para a subestrutura da classe de IC50 Muito Bom, dos 89 novos compostos obtidos na ChEMBL e classificados com as árvores de decisão, 15 dos compostos foram previstos pertencerem à classe de IC50 Bom, a 29 dos compostos foi atribuída a classe de IC50 Médio e a 45 dos compostos foi atribuída a classe de IC50 Mau (Tabela 29).

A partir da subestrutura da classe de IC50 Bom foram obtidos da ChEMBL 91 novos compostos. Destes, a 80 compostos foi atribuída a classe de IC50 Bom, a 9 compostos foi atribuída a classe de IC50 Médio e a 2 dos compostos foi atribuída a classe de IC50 Mau.

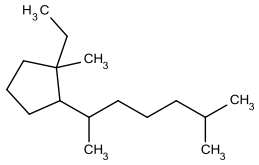
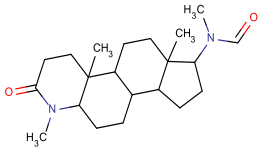
Tabela 29. Atribuição das classes de IC50 para os novos compostos na 5α-R1 pela análise realizada no algoritmo de aprendizagem árvores de decisão.

Compostos analisados				**Novos compostos**				
Classe IC50	**ID**	**MCS**	**# compostos**	**# compostos**	**Classe de IC50 prevista**			
					Muito Bom	**Bom**	**Média**	**Mau**
Muito bom	1		4	89	0	15	29	45
Bom	9		15	91	0	80	9	2

A mesma análise realizada com o algoritmo de aprendizagem SVM demonstrou que para a subestrutura da classe de IC50 Muito Bom (subestrutura 1, Tabela 23) todos os 89 novos compostos obtidos na ChEMBL pertencem à classe de IC50 Mau (Tabela 30).

Quanto à análise dos compostos obtidos a partir da subestrutura mais frequente da classe de IC50 Bom, dos 91 compostos obtidos na ChEMBL, apenas 2 compostos foram atribuídos à classe de IC50 Bom, 3 compostos foram atribuídos à classe de IC50 Médio e 86 compostos foram atribuídos à classe de IC50 Mau (Tabela 30).

Tabela 30. Atribuição das classes de IC50 para os novos compostos na 5α-R1 pela análise realizada no algoritmo de aprendizagem SVM.

Compostos analisados				**Novos compostos**				
Classe de IC50	**ID**	**MCS**	**# compostos**	**# compostos**	**Classe de IC50 prevista**			
					Muito Bom	**Bom**	**Médio**	**Mau**
Muito bom	1		4	89	0	0	0	89
Bom	9		15	91	0	2	3	86

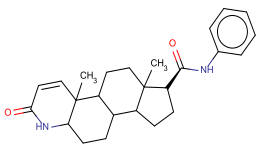
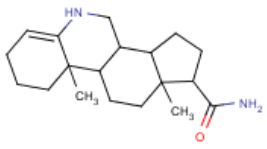
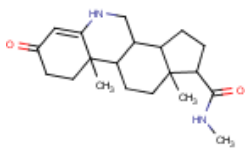
Na isoenzima 5α-R2 foram consideradas duas subestruturas para as classes de IC50 Muito Bom e Bom, a partir das quais se pesquisaram outros compostos na ChEMBL. Tal como para a 5α-R1, estes compostos foram classificados utilizando as árvores de decisão (Tabela 31) a subestrutura 2 da classe de IC50 Muito Bom (Tabela 26), dos 27 novos compostos obtidos na ChEMBL, 9 compostos foram previstos pertencer à classe de IC50 Muito Bom, 16 dos compostos foram previstos pertencerem à classe de IC50 Bom, e a 2 dos compostos foi atribuída a classe de IC50 Médio.

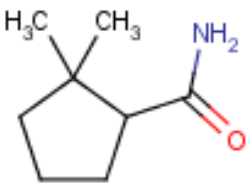
Para a subestrutura 6 da classe de IC50 Muito Bom (Tabela 26), dos 48 novos compostos obtidos na ChEMBL, 6 compostos foram previstos pertencer à classe de IC50 Muito Bom, a 24 dos compostos foi atribuída a classe de IC50 Bom, a 7 dos compostos foi atribuída a classe de IC50 Médio e a 11 dos compostos foi atribuída a classe de IC50 Mau.

A partir da subestrutura 1 da classe de IC50 Bom (Tabela27), foram obtidos da ChEMBL 69 novos compostos. Destes, 48 compostos foram previstos como pertencentes à classe de IC50 Muito Bom, a 8 dos compostos foi atribuída a classe de IC50 Bom, a 5 compostos foi atribuída a classe de IC50 Médio e a 8 dos compostos foi atribuída a classe de IC50 Mau.

Para a subestrutura 2 da classe de IC50 Bom (Tabela 27), foram obtidos da ChEMBL 92 novos compostos, sendo que a 22 dos compostos foi atribuída a classe de IC50 Muito Bom, a 18 dos compostos foi atribuída a classe de IC50 Bom, a 18 compostos foi atribuída a classe de IC50 Médio e a 34 dos compostos foi atribuída a classe de IC50 Mau.

Tabela 31. Atribuição das classes de IC50 para os novos compostos na 5α-R2 pela análise realizada no algoritmo de aprendizagem árvores de decisão.

Compostos analisados				**Novos compostos**				
Classe de IC50	**ID**	**MCS**	**# compostos**	**# compostos**	**Classe de IC50**			
					Muito Bom	**Bom**	**Médio**	**Mau**
Muito bom	2		21	27	9	16	2	0
Muito bom	6		46	48	6	24	7	11
Bom	1		19	69	48	8	5	8

Bom	2		27	92	22	18	18	34

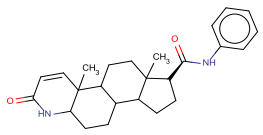
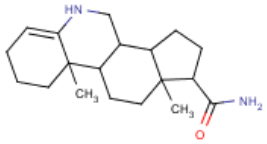
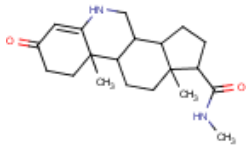
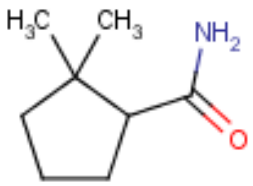
Na tabela 32 estão apresentados os resultados da classificação destes compostos utilizando os modelos de SVM. Para os compostos obtidos a partir da subestrutura 2 da classe de IC50 Muito Bom (Tabela 26), 5 compostos foram previstos pertencer à classe de IC50 Muito Bom, 7 dos compostos foram previstos pertencerem à classe de IC50 Bom, e a 15 dos compostos foi atribuída a classe de IC50 Mau.

Nas subestruturas da classe de IC50 Muito Bom (Tabela 26), dos 48 novos compostos obtidos na ChEMBL, 3 compostos foram previstos pertencer à classe de IC50 Muito Bom, a 13 dos compostos foi atribuída a classe de IC50 Bom, a 1 dos compostos foi atribuída a classe de IC50 Médio e a 31 dos compostos foi atribuída a classe de IC50 Mau.

Quanto à subestrutura 1 da classe de IC50 Bom (Tabela 27), dos 69 novos compostos, 42 compostos foram previstos como pertencentes à classe de IC50 Muito Bom, a 1 dos compostos foi atribuída a classe de IC50 Bom e a 26 dos compostos foi atribuída a classe de IC50 Mau.

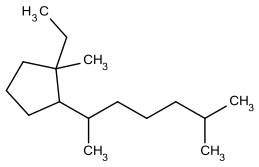
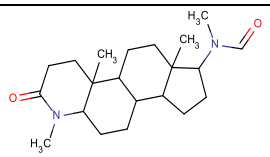
Para a subestrutura 2 da classe de IC50 Bom (Tabela 27), dos 92 novos compostos, a 13 compostos foi atribuída a classe de IC50 Muito Bom, a 2 dos compostos foi atribuída a classe de IC50 Bom, a 1 composto foi atribuída a classe de IC50 Médio e a 76 dos compostos foi atribuída a classe de IC50 Mau.

Tabela 32. Atribuição das classes de IC50 para os novos compostos na 5α-R2 pela análise realizada no algoritmo de aprendizagem SVM.

Compostos analisados				Novos compostos				
Classe de IC50	ID	MCS	# compostos	# compostos	Classe de IC50 prevista			
					Muito Bom	Bom	Médio	Mau
Muito bom	2		21	27	5	7	0	15
Muito bom	6		46	48	3	13	1	31
Bom	11		19	69	42	1	0	26
Bom	2		27	92	13	2	1	76

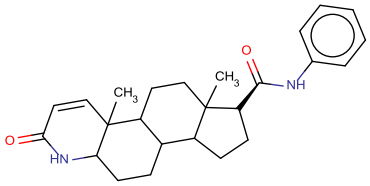
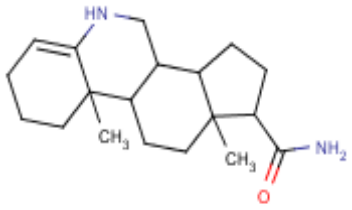
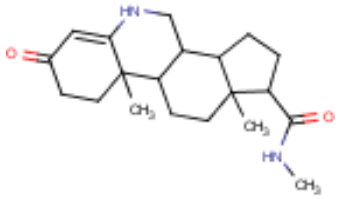
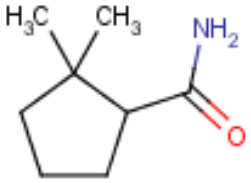
Na tabela 33 e tabela 34 podemos observar os compostos que foram classificados pelos dois métodos SVM e árvores de decisão na mesma classe de IC50 em cada uma das isoenzimas 5α-R1 e 5α-R2, respetivamente. Para a subestrutura 2 (Tabela 33), 45 compostos foram classificados pelos dois algoritmos de aprendizagem como pertencentes à classe de IC50 Mau. Para a subestrutura 9 (Tabela 33), podemos observar que 2 compostos foram classificados por ambos os algoritmos como pertencentes à classe de IC50 Mau, 3 compostos foram atribuídos à classe de IC50 Médio e 2 compostos foram atribuídos á classe de IC50 Bom.

Tabela 33. Número de compostos classificados pelos algoritmos de aprendizagem SVM e árvores de decisão na mesma classe de IC50 para a 5α-R1.

Compostos analisados			Compostos classificados pelos 2 métodos			
Classe de IC50	ID	MCS	Classe de IC50 prevista			
			Muito Bom	Bom	Médio	Mau
Muito bom	1		0	0	0	45
Bom	9		0	2	3	2

Na 5α-R2 para a subestrutura 2 (Tabela 34), 4 compostos foram atribuídos à classe de IC50 Muito Bom e 7 compostos foram classificados como pertencentes à classe de IC50 Bom, por ambos os algoritmos de aprendizagem. Para a subestrutura 6 (Tabela 34), 3 compostos foram classificados pelos dois métodos como pertencentes à classe de IC50 Muito Bom, 12 compostos à classe de IC50 Bom, 1 composto foi atribuído à classe de IC50 Médio e 11 compostos foram atribuídos à classe de IC50 Mau. Nas duas subestruturas analisadas para a classe de IC50 Bom, a subestrutura 1 possui 43 compostos classificados como IC50 Muito Bom, 1 compostos atribuído à classe de IC50 Bom e 8 compostos classificados como classe de IC50 Mau. Para a subestrutura 2 foram classificados por ambos os algoritmos, 13 compostos na classe de IC50 Muito Bom, 2 compostos classificados como classe de IC50 Bom, 1 composto atribuído à classe de IC50 médio e 40 compostos atribuídos à classe de IC50 Mau.

Tabela 34. Número de compostos classificados pelos algoritmos de aprendizagem SVM e árvores de decisão na mesma classe de IC50 para a 5α-R2.

Compostos analisados			Compostos classificados pelos 2 métodos			
Classe de IC50	ID	MCS	Classe de IC50 prevista			
			Muito Bom	Bom	Médio	Mau
Muito bom	2		4	7	0	0
Muito bom	6		3	12	1	11
Bom	11		43	1	0	8
Bom	2		13	2	1	40

D. Discussão

A proteína 5α-redutase tem como principal papel a diferenciação sexual humana, no entanto quando desregulada a função desta enzima, vários distúrbios endócrinos podem ocorrer, tais como calvície, acne, hiperplasia benigna da próstata, cancro da próstata, entre outros. O facto de nos últimos anos o número de população masculina com doenças como o cancro da próstata e a hiperplasia benigna da próstata ter aumentado, resultou no desenvolvimento de inibidores desta enzima.

Atualmente, apenas dois fármacos, o finasterida e o dutasterida se encontram comercializados como inibidores desta enzima. No entanto, possuem efeitos colaterais o que leva à necessidade de se desenvolverem novos inibidores mais potentes e seletivos. Essa investigação que não é facilitada pois a estrutura da 5α-redutase ainda não se encontra determinada experimentalmente, criando um grande obstáculo para o desenvolvimento de novos inibidores. Devido a este facto, é necessário investigar os inibidores já conhecidos para assim tentar encontrar novos inibidores da 5α-redutase.

A aplicação do método análise de componentes principais (PCA) e do método seleção de atributos, sobre os dados de compostos que inibem a atividade das isoenzimas 5α-R1 e 5α-R2 revelaram algumas variáveis (descritores moleculares) consideradas de elevada importância para na constituição de um bom inibidor para a 5α-redutase.

Da análise dos 40 descritores moleculares calculados para os compostos pertencentes à classe de IC50 Bom da 5α-R1, três características foram consideradas importantes em ambos os métodos como observado nas tabelas 5 e 11. Do conjunto de seis características obtidas para esta classe pelo método de seleção de atributos, a área de superfície polar, área de superfície e volume são identificadas pelo método de PCA com um fator de elevada relevância. A área de superfície e o volume molecular possuem um fator elevado na primeira componente principal, a área de superfície polar possui um fator elevado na terceira componente principal. Entre todas as variáveis encontradas como relevantes no PCA e na seleção de atributos podemos dizer que estas três características por se encontrarem aplicando os dois métodos poderão ser de maior importância, e das três características, as características pertencentes à primeira componente principal serão as mais relevantes para a classe de IC50 Bom.

Quanto à classe de IC50 Médio, 4 características foram consideradas importantes por estes métodos como se pode observar na tabela 6 e na tabela 11. Do conjunto de seis características selecionadas para esta classe com o método de seleção de atributos, a ASA+, o número de anéis, o volume e o LogP foram igualmente considerados pelo PCA como

características relevantes. A ASA+, número de anéis e volume possuem um fator elevado na primeira componente principal, enquanto o valor de LogP possui um fator elevado na terceira componente principal. O que significa que entre todas as características consideradas com importância com ambos os métodos, as 4 características que se encontram nos dois métodos terão entre todas uma relevância maior. Destas 4 características as que pertencem à primeira componente principal terão uma maior importância para a classe de IC50 Médio.

Das características obtidas como importantes nas classes de IC50 Muito Bom, Bom e Médio para a 5α-R2 com os métodos PCA e seleção de atributos algumas destas foram consideradas relevantes por estes dois métodos. Para a classe de IC50 Muito Bom das 18 características obtidas no método de seleção de atributos (Tabela 12), 9 destas características também foram consideradas relevantes pelo PCA (Tabela 8). Destas 9 características a ASA+ e a ASA_H possuem um fator elevado na primeira componente principal, o número de anéis aromáticos e o número de anéis aromáticos com 6 carbonos possuem um elevado fator na segunda e terceira componente principal e o número de anéis aromáticos com 5 carbonos, o número de anéis de aromáticos unidos, o número de anéis hétero, o número de anéis hétero alifáticos e o número de anéis hétero aromáticos possuem um elevado fator na quarta componente principal. Entre todas as variáveis encontradas como relevantes no PCA e na seleção de atributos podemos dizer que estas 9 características por se encontrarem nos dois métodos são das mais relevantes. Destas 9 características, as que pertencem à primeira componente principal serão as de maior importância para a classe de IC50 Muito Bom.

Na classe de IC50 Bom apenas uma das características é identificada pelos dois métodos. Como se pode observar a partir daa tabelas 9 e 12. Do conjunto de seis características obtidas para esta classe no método de seleção de atributos a área de superfície polar foi igualmente considerada pelo PCA como uma característica relevante. Tendo esta um fator elevado na terceira componente principal, o que nos leva a pensar que entre todas as características consideradas por estes métodos como relevantes para esta classe a área de superfície polar será a mais importante.

Quanto à classe de IC50 Médio também apenas uma característica é identificada como significativa pelos dois métodos (Tabela 10 e 12). Do conjunto de quatro características selecionadas pelo método de seleção de atributos, a ASA_P foi igualmente considerada pelo PCA como característica relevante. Tendo esta um fator negativo elevado

na primeira componente principal e um fator elevado positivo na terceira componente principal, o que nos leva a pensar que entre todas as características consideradas por estes métodos como relevantes para esta classe a ASA_P será a mais importante. O fator negativo indica que as variáveis são inversamente correlacionadas.

Podemos observar no método de seleção de atributos que para a 5α-R1, 4 características encontram-se em ambas as classes de IC50 Bom e Médio. Sendo elas o número de anéis, a área de superfície, o volume e o LogP. Uma vez que se encontram em ambas as classes estas características são segundo este método, muito importantes para o desenvolvimento dos inibidores da 5α-R1. Para a 5α-R2, encontram-se em ambas as classes de IC50 Muito Bom e Bom 4 características, a área de superfície polar, a ASA+, o número de aceitadores de hidrogénio e a ASA_H. Na classe de IC50 Bom e Médio podemos encontrar a característica área de superfície e nas classes de IC50 Muito Bom e Médio encontram-se as características ASA_P, número de anéis hetero aromáticos e o número de anéis aromáticos unidos. Estes resultados indicam que estas características são muito importantes no desenvolvimento de inibidores da 5α-R2.

Através do treino e teste dos algoritmos de aprendizagem SVM e árvores de decisão foi possível avaliar a utilização destes métodos para, a partir dos 40 descritores moleculares escolhidos, prever as classes de IC50 que novos compostos possam apresentar nas diferentes isoenzimas da 5α-redutase. Observando as tabelas 13 e 17, para a 5α-R1 os resultados obtidos pelo algoritmo de aprendizagem árvores de decisão aproximaram-se mais da classe de IC50 real dos compostos do que o método de SVM, quando todos os compostos foram utilizados para treino e teste. No entanto, quando foi utilizado o método validação cruzada na construção dos modelos de classificação (Tabela 14 e Tabela 18), os resultados em ambos os algoritmos de aprendizagem ficaram muito próximos uns dos outros e muito distantes das classes de IC50 reais. Dos resultados obtidos pelos algoritmos, a classe para a qual se obtém melhores resultados é a classe de IC50 Mau. Para este facto podem estar a contribuir o elevado número de compostos pertencentes à classe IC50 Mau, o que poderá estar a "ensinar" os algoritmos a reconhecer os compostos com valores de IC50 Mau.

Para a 5α-R2 também se observa que os resultados obtidos pelo algoritmo de aprendizagem árvores de decisão se aproximaram mais da classe de IC50 real dos compostos do que o SVM (Tabela 15 e 19). Contudo, quando aplicado o método de validação cruzada na construção dos modelos de classificação (Tabelas 16 e 20), em ambos

os algoritmos de aprendizagem, a classificação parece dar bons resultados para umas classes de IC50 e maus resultados para outros. Dos dados obtidos pelos algoritmos, as classes com melhores resultados são a classe de IC50 Muito Bom e a classe de IC50 Mau (Tabela 16 e 20). Mais uma vez estas, estas são as classes com maior número de compostos.

Nas tabelas 21 e 22 são apresentadas as métricas de desempenho para o conjunto de compostos da 5α-R1 e 5α-R2, respetivamente. Podemos observar apenas diferenças para a classificação na classe de IC50 Médio para a 5α-R1 e na classe de IC50 Mau para a 5α-R2. Podendo então concluir-se que os algoritmos de aprendizagem utilizados por nós neste trabalho são confiáveis e eficientes.

Da pesquisa de subestruturas máximas comuns (MCS) obtidas na 5α-R1 para a classe de IC50 Bom, dos 53 compostos pertencentes a esta classe 15 compostos possuem a subestrutura 9 (Tabela 24) esta subestrutura é um 4-azasteróide assim como a subestrutura 5. Grande parte dos compostos encontrados na ChEMBL são inibidores esteroides, sendo a maioria 4-azasteroides. De facto na literatura, podemos encontrar o dutasterida um inibidor duplo da 5α-R1 e da 5α-R2 (Salvador et al., 2013) que possui uma estrutura semelhante à subestrutura 1 e à subestrutura 2. No entanto é a subestrutura 9, um 4-azasteroide, que é comum a um maior número de compostos. Esta subestrutura possuí uma cadeia ligada ao C_{17} muito diferente da cadeia ligada no mesmo carbono na dutasterida, não contendo o anel aromático ligado a dois carbonos trifluor. Uma vez que da análise dos compostos com PCA a propriedade número de átomos de fluor foi considerada de elevada relevância na terceira componente principal e na quarta componente principal para esta classe podemos concluir que esta propriedade é importante na construção de inibidores desta isoenzima.

Na classe de IC50 Médio os 88 compostos que a compõem são na sua maioria subestruturas não esteroides, como observado na tabela 25, que têm na sua constituição estruturas semelhantes ou iguais ao composto LY-191,704 considerado um bom inibidor da 5α-R1 (Salvador et al., 2013). No entanto, a subestrutura 10 um 4-azasteroide é a subestrutura comum a um maior número de compostos, fazendo parte da constituição de 29 compostos. Esta subestrutura é igual à subestrutura 2 encontrada para 4 compostos na classe de IC50 Bom. O que nos leva a pensar que esteroides que possuam anéis aromáticos na cadeia ligada ao C_{17} não serão inibidores muito bons desta isoenzima.

Ao contrário do que acontecia na 5α-R1, na tabela 27, para a classe de IC50 Muito Bom na 5α-R2 podemos observar que as subestruturas comuns a um maior número de compostos são esteroides 4-azasteroides, possuidoras de anéis aromáticos e 6-azasteroides.

As propriedades aromáticas obtidas como relevantes no PCA para a classe de IC50 Muito Bom encontram-se expressas nas subestruturas desta classe. A subestrutura com maior número de compostos é um 6-azasteroide que possui ligado ao carbono C_{17} um grupo amida que favorece a ligação de heterocíclicos uma propriedade referida como relevante pelo método PCA.

Para a classe de IC50 Bom, podemos observar na tabela 27 que as subestruturas 6-azasteroides estão presentes numa grande parte dos compostos. A subestrutura comum a um maior número de compostos é uma subestrutura não-esteroide. No entanto esta subestrutura é um anel de 5 carbonos que possui ligado um grupo amida e dois grupos metil. Por esta razão esta subestrutura poderá ser um fragmento das subestruturas 6-azasteroides. Tal como observado, nas propriedades identificadas como importantes pelo PCA as propriedades aromáticas deixam de ter tanta relevância nesta classe.

Na classe de IC50 Médio não existe nenhuma subestrutura com um número de compostos associados relevantes. No entanto, podemos observar na tabela 28 que 6 compostos possuem em comum a subestrutura 1 que é um 4-azasteroide, sendo o resto das subestruturas não esteroides. São poucos os compostos nesta classe que possuem anéis aromáticos, a maioria dos compostos possui subestruturas com propriedades alifáticas, tendo estas sido consideradas no PCA de elevada relevância.

Em suma, comparando os nossos resultados obtidos na análise dos compostos pelos métodos de aprendizagem de máquina e as subestruturas máximas comuns obtidas para os compostos nas diferentes classes de IC50 e nas diferentes isoenzimas podemos dizer que os esteroides 4-azasteroides e 6-azasteroides são bons inibidores de ambas as isoenzimas. No entanto, para se obter um bom inibidor da 5α-R2 as propriedades aromáticas são muito importantes, uma vez que bons inibidores desta isoenzima possuem na sua estrutura anéis aromáticos. Por outro lado, na 5α-R1 são as propriedades alifáticas que se tornam importantes para a formação de um bom inibidor desta enzima.

A partir das subestruturas comuns a um maior número de compostos nas classes de IC50 Muito Bom e classe de IC50 Bom nas duas isoenzimas procedeu-se à identificação de outros compostos com estas subestruturas na ChEMBL e aplicaram-se os algoritmos de aprendizagem SVM e árvores de decisão para prever a classe de IC50 destes compostos em cada isoenzima. Na 5α-R1 dos compostos obtidos a partir da subestrutura da classe de IC50 Muito Bom, a nenhum composto foi atribuída a classe de IC50 Muito Bom. No entanto, 44 compostos foram atribuídos à classe de IC50 Bom e à classe de IC50 Médio e 45 compostos

foram atribuídos à classe de IC50 Mau com o classificador árvores de decisão (Tabela 29). Por outro lado, o classificador SVM atribuiu a todos os 89 compostos a classe de IC50 Mau (Tabela 30). Apesar destes resultados não serem animadores, esta subestrutura deve ser tida em consideração uma vez que todos os compostos com IC50 Muito Bom para a 5α-R1 têm esta subestrutura, e os classificadores obtidos terão sempre muita dificuldade em identificar compostos com classe de IC50 Muito Bom uma vez que no máximo apenas 4 exemplares fazem parte do conjunto de treino.

Para a classe de IC50 Bom, com os modelos de classificação das árvores de decisão, dos 92 novos compostos, 80 foram atribuídos à classe de IC50 Bom (Tabela 29). No entanto, com os modelos SVM 86 compostos foram atribuídos à classe de IC50 Mau (Tabela 30). Apesar do elevado número de compostos atribuído à classe de IC50 Mau pelo SVM, esta subestrutura torna-se interessante para futuras pesquisas devido ao elevado número de compostos atribuídos à classe de IC50 Bom pelos modelos das árvores de decisão e pelos resultados obtidos com os outros métodos. Uma vez, que esta subestrutura é um 6-azasteroide considerada importante para um bom inibidor da 5α-R1.

Para a 5α-R2, foram analisadas as duas subestruturas comuns a um maior número de compostos na classe de IC50 Muito Bom e na classe de IC50 Bom. Na subestrutura 2 (Tabela 26), 16 compostos dos 27 novos compostos foram atribuídos à classe de IC50 Bom com os classificadores baseados em árvores de decisão (Tabela 31). Com os classificadores SVM, 7 compostos foram atribuídos à classe de IC50 Bom, 5 à classe de IC50 Muito Bom (Tabela 32). Observando os nossos dados poderemos dizer que esta é uma subestrutura interessante para a formação de inibidores da 5α-R2. A segunda subestrutura comum a um maior número de compostos na classe de IC50 Muito Bom (subestrutura 6, Tabela 26) com os modelos das árvores de decisão, 6 dos novos compostos foram atribuídos à classe de IC50 Muito Bom, 24 compostos foram atribuídos à classe de IC50 Bom. Por outro lado, com o algoritmo SVM, 13 compostos foram atribuídos à classe de IC50 Bom. Quanto à subestrutura 1 (Tabela 27) pertencente à classe de IC50 Bom, 48 compostos identificados foram atribuídos à classe de IC50 Muito Bom e 8 compostos à classe de IC50 Bom com os modelos de árvores de decisão (Tabela 31). Enquanto que com os modelos SVM, 42 compostos foram atribuídos à classe de IC50 Muito Bom.

Para a subestrutura 2 (Tabela 27) da classe de IC50 Bom, dos 92 novos compostos identificados, as árvores de decisão, classificam 22 compostos como pertencendo à classe de

IC50 Muito Bom e 18 compostos com classe de IC50 Bom. Já com o algoritmo SVM, 13 compostos foram atribuídos à classe de IC50 Muito Bom.

A existência de compostos classificados pelos métodos SVM e árvores de decisão, dentro da mesma classe para a 5α-R1 e 5α-R2, leva-nos a pensar que a classe de IC50 foi devidamente atribuída a esses compostos. Assim sendo, para a 5α-R1 na subestrutura 1, (Tabela 33), 45 compostos foram classificados por ambos os algoritmos como pertencentes à classe de IC50 Mau. Na subestrutura 9, (Tabela 33), podemos observar que 2 compostos foram atribuídos à classe de IC50 Mau, 3 compostos à classe de IC50 Médio e 2 compostos à classe de IC50 Bom. Na 5α-R2, estes métodos classificaram para a subestrutura 2, (Tabela 34), 4 compostos como pertencendo à classe de IC50 Muito Bom e 7 compostos à classe de IC50 Bom. Na subestrutura 6, (Tabela 34), 3 compostos foram atribuídos à classe de IC50 Muito Bom, 12 compostos à classe de IC50 Bom, 1 composto à classe de IC50 Médio e 11 compostos à classe de IC50 Mau. Para a classe de IC50 Bom à subestrutura 1, (Tabela 34), foram atribuídos 43 compostos à IC50 Muito Bom e 8 compostos à classe de IC50 Mau. Enquanto na (subestrutura 2, Tabela 34), foram atribuídos 13 compostos à classe de IC50 Muito Bom, 2 compostos à classe de IC50 médio e 40 compostos à classe de IC40 Mau. Sendo estes os compostos corretamente classificados com a classe de IC50. Na tabela 35 pode-se observar as estruturas dos compostos atribuídos à classe de IC50 Bom para a 5α-R1 e na tabela 36 os compostos para a classe de IC50 Muito Bom para a 5α-R2.

E. Conclusão

O número crescente de doenças como o cancro da próstata e a hiperplasia benigna da próstata entre outras, causadas sobretudo por distúrbios na 5α-redutase, tem resultado no desenvolvimento de inibidores desta enzima. No entanto, atualmente apenas dois inibidores são comercializados (finasterida e dutasterida), e os efeitos colaterais provocados por estes, leva à necessidade de se desenvolverem novos inibidores mais potentes e seletivos. O motivo para a escassez de inibidores comercializados altamente eficazes, deve-se ao facto da estrutura tridimensional da enzima 5α-redutase não ser conhecida, consequência da sua instabilidade durante o processo de purificação.

Este trabalho teve como objetivo geral, a tentativa de se obter informações relevantes sobre a possível estrutura da enzima 5α-redutase e sobre os inibidores desta, de modo a contribuir para o desenvolvimento de antagonistas mais potentes, seletivos e menos tóxicos.

Na tentativa de obter informações relevantes sobre a possível estrutura da enzima em estudo, procedeu-se à modelação de proteínas por homologia, tendo-se em consideração que a modelação por homologia tem uma elevada probabilidade de sucesso quando a identidade sequencial entre a proteína alvo e o modelo é de pelo menos 70%. Diferentes aproximações foram executadas para a identificação e seleção de modelos de homologia para a 5α-redutase. No entanto, com os dados estruturais atualmente disponíveis tal não foi possível.

Uma vez que não foi possível obter uma estrutura da 5α-redutase, avançou-se então para métodos de rastreio virtual baseados em ligandos em busca de relações entre as propriedades dos inibidores da 5α-redutase conhecidos e da sua atividade. Com estes estudos é possível analisar quais as propriedades (descritores moleculares) relevantes para a construção de inibidores eficientes e seletivos da 5α-R1 e 5α-R2. Através dos métodos de aprendizagem de máquina utilizados neste trabalho, observou-se para a 5α-R1 que as propriedades como a área de superfície, o volume, a área de superfície polar, ASA+, LogP e número de anéis, são propriedades importantes no desenvolvimento de inibidores para esta isoenzima. Já para a 5α-R2 são propriedades como a área de superfície, ASA+, ASA_P, ASA_H, propriedades aromáticas, número de aceitadores de hidrogénio e área de superfície polar, as propriedades importantes no desenvolvimento de inibidores para esta isoenzima. Sendo algumas destas propriedades observadas nas subestruturas máximas comuns entre inibidores de ambas as enzimas.

Os métodos SVM e árvores de decisão foram testes de classificação das classes de IC50 dos compostos inibidores da 5α-redutase. Podemos observar para os dois algoritmos que a correta classificação é melhor conseguida para umas classes IC50 do que para outras. No entanto, as classes com maior número de compostos corretamente classificados foram as classes de IC50 Muito Bom e IC50 Mau para ambas as isoenzimas.

A comparação dos resultados obtidos entre as subestruturas máximas comuns e a análise dos compostos pelos métodos de aprendizagem de máquina permite-nos dizer que os esteroides 4-azasteroides e 6-azasteroides são bons inibidores de ambas as isoenzimas. Contudo, para se obter bons inibidores da 5α-R2 as propriedades aromáticas são muito importantes. Por outro lado, na 5α-R1 são as propriedades alifáticas que se tornam importantes para a formação de um bom inibidor desta enzima.

As subestruturas máximas comuns obtidas com maior número de compostos para cada isoenzima nas diferentes classes de IC50 foram avalidadas. Dos resultados obtidos podemos concluir, para a 5α-R1, que a única subestrutura obtida para a classe de IC50 Muito Bom deve de ser tida em consideração em futuras investigações e desenvolvimentos de inibidores porque apesar de não ter tido bons resultados todos os compostos classificados como muito bom têm esta subestrutura. Também a subestrutura da classe de IC50 Bom é interessante para futuras pesquisas devido ao elevado número de compostos atribuídos à classe de IC50 Bom pelos modelos das árvores de decisão e pelos resultados obtidos com os outros métodos. Uma vez, que esta subestrutura é um 6-azasteroide considerada importante para um bom inibidor da 5α-R1

Para a 5α-R2, na classe de IC50 Muito Bom podemos concluir que a subestrutura 2 (Tabela 26), é uma subestrutura interessante para a formação de inibidores da 5α-R2. Quanto à subestrutura 1 (Tabela 27) pertencente à classe de IC50 Bom, obteve-se bons resultados, podendo-se dizer que esta subestrutura será importante. Para a subestrutura 2 (Tabela 27) da classe de IC50 Bom, da observação dos resultados obtidos esta não será uma subestrutura muito importante para a construção de inibidores desta enzima.

Como perspetiva de trabalho futuro seria interessante avaliar laboratorialmente a atividade bioativa dos novos compostos retirados da ChEMBL e classificados dentro da mesma classe de IC50 Muito Bom e Bom pelos métodos SVM e árvores de decisão.

F. Bibliografia

Aggarwal, S., Thareja, S., Bhardwaj, T. R., Kumar, M. (2010). Self-organizing molecular field analysis on pregnane derivatives as human steroidal 5alpha-reductase inhibitors. *Steroids*, *75*(6) 411–8.

Aggarwal, S., Thareja, S., Verma, A., Bhardwaj, T. R., Kumar, M. (2010). An overview on 5alpha-reductase inhibitors. *Steroids 75*(2) 109–53.

Amory, J. K., Anawalt, B. D., Matsumoto, A. M., Page, S. T., Bremner, W. J., Wang, C., Swerdloff, R. S., *et al*. (2008). The effect of 5alpha-reductase inhibition with dutasteride and finasteride on bone mineral density, serum lipoproteins, hemoglobin, prostate specific antigen and sexual function in healthy young men. *The Journal of Urology 179*(6) 2333–8.

Andersson, S., Russell, D. W. (1990). Structural and biochemical properties of cloned and expressed human and rat steroid 5 alpha-reductases. *Proceedings of the National Academy of Sciences of the United States of America 87*(10) 3640–4.

Anton, H *et.al*. (2004). Algebra linear com aplicações. Bookman. Porto editora.

Beniwal, S., Arora, J. (2012). Classification and feature selection techniques in data mining. International Journal os Engineering Research and Technology. 6(1) 2278-2281.

Breiman, L. (2001). Random forests. *Machine learning* 45 5-32.

Burbidge, R. (2001). An evaluation of some SVM heuristics for predicting activity on *pKi* assays. (personal communication).

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

Cao, Y., Jiang, T., Girke, T. (2008). A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bionformatics* 24 366-374.

Cavasotto, C. N., Phatak, S. S. (2009). Homology modeling in drug discovery: current trends and applications. *Drug Discovery Today 14*(13-14) 676–83.

Chen, B., Harrison, R. F., Papadatos, G., Willett, P., Wood, D. J., Lewell, X. Q., Greenidge, P., *et al*. (2007). Evaluation of machine-learning methods for ligand-based virtual screening. *Journal of Computer-Aided Molecular Design 21*(1-3) 53–62.

Cheng, T., Li, Q., Zhou, Z., Wang, Y., Bryant, S. H. (2012). Structure-based virtual screening for drug discovery: a problem-centric review. *The AAPS Journal 14*(1) 133–41.

Chi, H., City, M. (2012). Homology Modeling Tutorial.

Cortizo, J. C *et.al* (2006). Multi criteria wrapper improvements to naive bayes learning. Springer – Verlag. (419) 420-427.

Czerminski, R. *et al*. (2001). Use of support vector machine in pattern classification: application to QSAR studies.

Darnag, R. *et al.* (2010). Support vector machines: development of QSAR models for predicting anti-HIV-1 activity of TIBO derivates. European Journal of Medicinal Chemistry. 45 1590-1597.

Faragalla, J., Bremner, J., Brown, D., Griffith, R., Heaton, A. (2003). Comparative pharmacophore development for inhibitors of human and rat 5-alpha-reductase. *Journal of Molecular Graphics & Modelling*, *22*(1) 83–92.

Fukunishi, Y. (2009). Structure based drug screening and ligand based drug screening with machine laerning. *Combinatorial Chemistry & High Throughput Screening.* 12 397-408.

Geppert, H. *et al.* (2010). Current trends ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *Journal Chemical Informatuion. and Modeling.*. 50 205–216

Guido, R. V. C., Oliva, G., & Andricopulo, A. D. (2008). Virtual screening and its integration with modern drug design technologies. *Current Medicinal Chemistry*, *15*(1) 37–46.

Goodarzini *et al.* (2012). Feature selection methods in QSAR studies. *Journal of AOAC International*. 95(3), 650-663.

Hariharan, R *et. al.* (2011). MultiMCS: A fast algorithm for the Maximum Common Substructure problem on multiple molecules. *Journal of chemical information and modeling*. 51 788–806

Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., & Schuffenhauer, A. (2004). Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Organic & Biomolecular Chemistry 2*(22), 3256–66.

Iehlé, C., Radvanyi, F., Gil Diez de Medina, S., Ouafik, L. H., Gérard, H., Chopin, D., Raynaud, J. P., et al. (1999). Differences in steroid 5alpha-reductase iso-enzymes expression between normal and pathological human prostate tissue. *The Journal of Steroid Biochemistry and Molecular Biology 68*(5-6), 189–95.

Imperato-McGinley, J., & Zhu, Y.-S. (2002). Androgens and male physiology the syndrome of 5alpha-reductase-2 deficiency. *Molecular and Cellular Endocrinology*, *198*(1-2), 51–9.

Jain, A. *et al*. (2008). Recommendations for evaluation of computational methods. *Journal of Comput-Aided Molecular Design*. 22133–139.

Jin, Y., Penning, T. M. (2001). Steroid 5alpha-reductases and 3alpha-hydroxysteroid dehydrogenases: key enzymes in androgen metabolism. *Best Practice & Research. Clinical Endocrinology & Metabolism*, *15*(1) 79–94.

Jolliffe, I.T.(2002). Principal component analysis. *Springer series in Statistics*. 2 470-489.

Kapp, F. G., *et al*. (2012). 5-alpha-reductase type I (SRD5A1) is up-regulated in non-small cell lung cancer but does not impact proliferation, cell cycle distribution or apoptosis. *Cancer Cell International*, *12*(1), 1-10.

Karnan, D (et.al). (2010), Atribute reduction using backward elimination algorithm.

Khan, H. N., Kulsoom, S., Rashid, H. (2012). Ligand based pharmacophore model development for the identification of novel antiepileptic compound. *Epilepsy Research*, *98*(1) 62–71.

Kubinyi, H (1933). QSAR: Hansch analysis and approaches (vol1). VCH Vergsgellachaf mbc, weinheim.

Lavecchia, A., Di Giovanni, C. (2013). Virtual Screening Strategies in Drug Discovery: A Critical Review. *Current Medicinal Chemistry*.

Lei. Y *et.al.* (2003). Feature selection for high – dimensional data: a fast correlation – based filter solution.

Liu, T., Tang, G. W., Capriotti, E. (2011). Comparative modeling: the state of the art and protein drug target structure prediction. *Combinatorial Chemistry &Hhigh Throughput Screening*, *14*(6) 532–47.

Liu, M. *et al.* (2013). Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. Sensors and Actuators B: Chemical. 177 970-980.

Lyne, P. D. (2002). Structure-based virtual screening: an overview. *Drug Discovery Today*, *7*(20) 1047–55.

Marshall, G. R. (2004). Introduction to Chemoinformatics in Drug Discovery – A Personal View, 1–22.

Massart, D.L *et .al*. (1988) – Data handling in science and technology chemometrics: a textbook. *Elsivier Science Publishers*. 2 1-488.

Melville, J. L., Burke, E. K., & Hirst, J. D. (2009). Machine learning in virtual screening. *Combinatorial Chemistry & High Throughput Screening*, *12*(4) 332–43.

Mitchell, J. B. O. (2011). Informatics, machine learning and computational medicinal chemistry. *Future Medicinal Chemistry*, *3*(4) 451–67.

Niu, B., *et al.* (2007). Support vector machine for SAR/QSAR of phenethyl-amines. Acta Pharmacologica Sinica. 28(7) 1075-1086.

Occhiato, E. G., Guarna, A., Danza, G., & Serio, M. (2004). Selective non-steroidal inhibitors of 5 alpha-reductase type 1. *The Journal of Steroid Biochemistry and Molecular Biology*, *88*(1) 1–16.

Platz, E. a, & Giovannucci, E. (2004). The epidemiology of sex steroid hormones and their signaling and metabolic pathways in the etiology of prostate cancer. *The Journal of Steroid Biochemistry and Molecular Biology 92*(4) 237–53.

Poletti, a, Celotti, F., Rumio, C., Rabuffetti, M., & Martini, L. (1997). Identification of type 1 5alpha-reductase in myelin membranes of male and female rat brain. *Molecular and CellularEendocrinology 129*(2) 181–90.

Plewczynski, D. *et al.* (2005). Assessinh different classification meyhods for virtual screening. *Journal Chemical. Information and. Modeling.* 46 1098-1106.

Refaeilzadeh, P. *et al.* (2012). Cross-validation.

Rizner, T. L. (2003). Human Type 3 3 -Hydroxysteroid Dehydrogenase (Aldo-Keto Reductase 1C2) and Androgen Metabolism in Prostate Cells. *Endocrinology 144*(7) 2922–2932.

Salvador, J. A. R., Carvalho, J. F. S., Neves, M. a C., Silvestre, S. M., Leitão, A. J., Silva, M. M. C., & Sá e Melo, M. L. (2013). *Anticancer steroids: Linking Natural and Semi Synthetic Compounds. Natural product reports* 30(2) 324–74.

Salvador, J. A.R., Pinto, R., Silvestre. S. (2013). Steroidal 5α-redutase and 17α-hydroxylase/17,20-lyase (CY17) inhibitors useful in the treatment of prostatic diseases. Journal of Steroid Biochemistry and Molecular Biology. 1-24.

Sayes, Y., *et al.* (2007). A review os feature selection tecniques in bioinformatics. *Bioinformatics.* 23(19) 2507-2517.

Steers, W. D. (2001). 5Alpha-Reductase Activity in the Prostate. *Urology 58*(6) 17–24.

Smith.L. (2002). A tutorial on principal componentes analysis.

Valasani, K. R., Chaney, M. O., Day, V. W., & Shidu Yan, S. (2013). Acetylcholinesterase Inhibitors: Structure Based Design, Synthesis, Pharmacophore Modeling and Virtual Screening. *Journal of Chemical Information and Modeling*.

Van Eekelen, C. C. E. M. (1996). Two Androgen Response Regions Cooperate in Steroid Hormone Regulated Activity of the Prostate-specific Antigen Promoter. *Journal of Biological Chemistry 271*(11) 6379–6388.

Willighagen, E., *et al.,* (2013). The ChEMBL database as linked open data. Journal of Cheminformatics. 23(1) 5-23.

G.Anexos

Tabela 35. Estrutura dos compostos atribuídos à classe de IC50 Bom para a 5α-R1 pelos métodos de classificação de árvores de decisão e SVM.

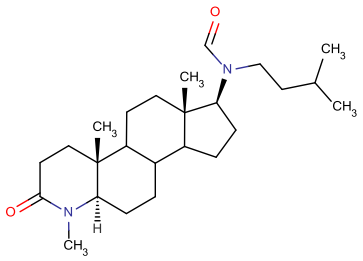
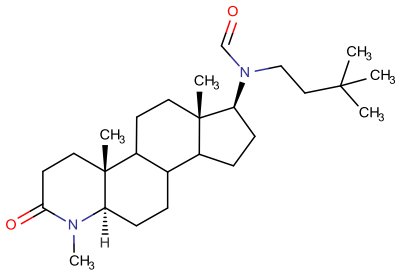
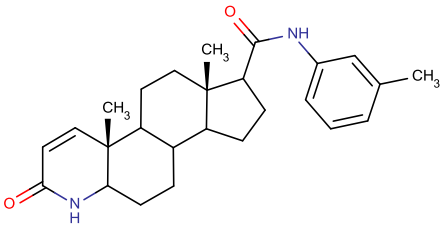
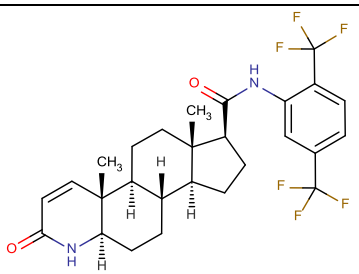
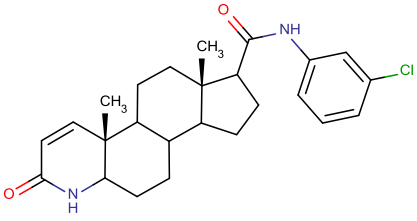
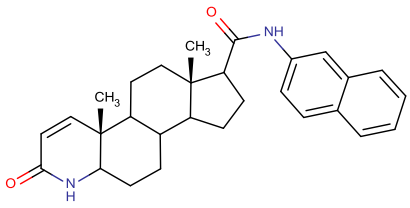
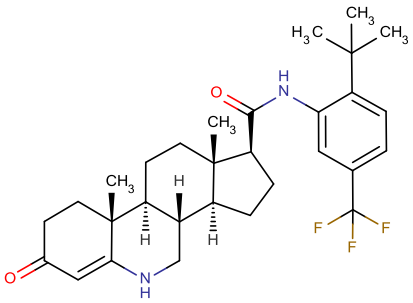
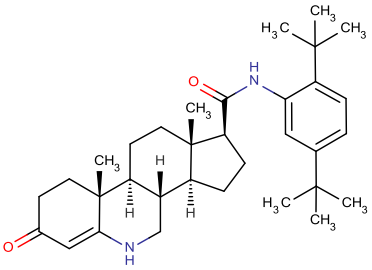
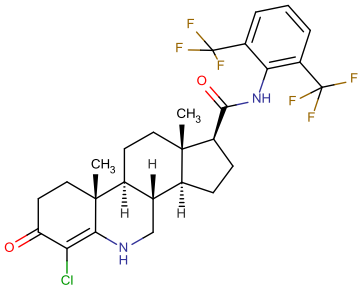
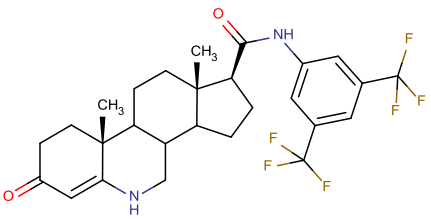
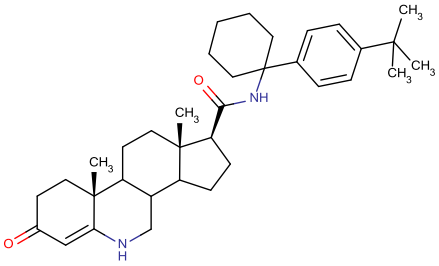
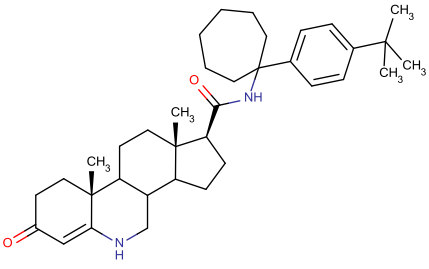
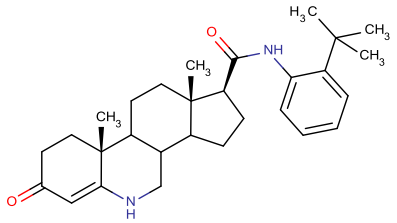
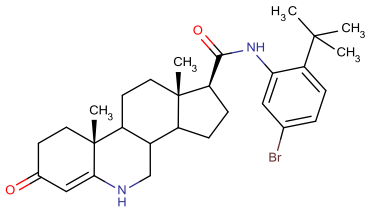
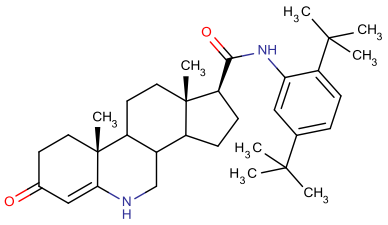
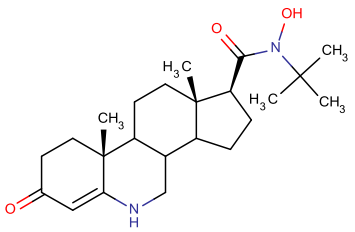
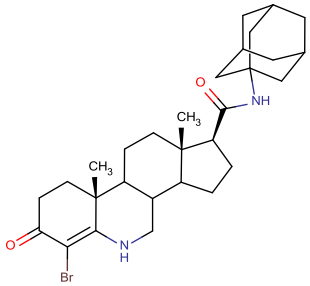
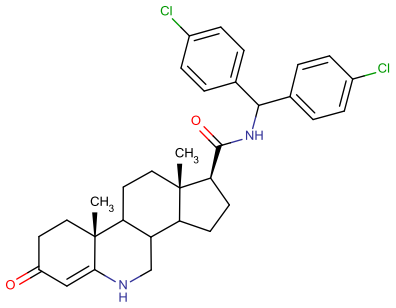
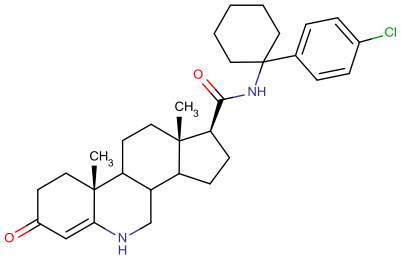
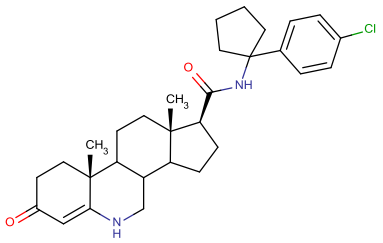
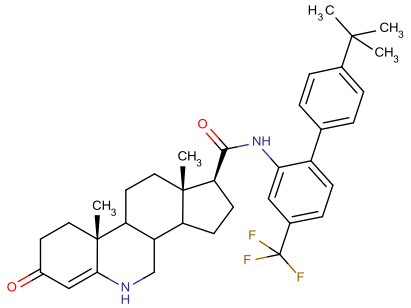
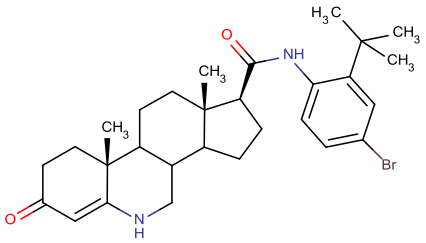
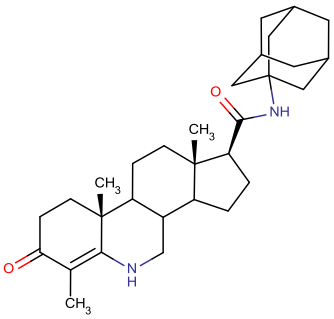
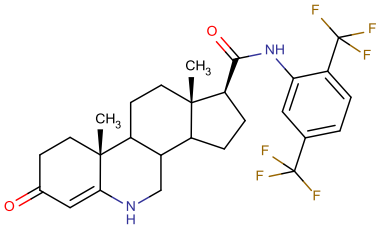
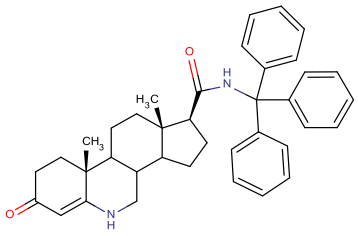
Classe de IC50	ID	ChEMBL ID	Estrutura
Bom	9	ChEMBL172222	
Bom	9	ChEMBL168663	

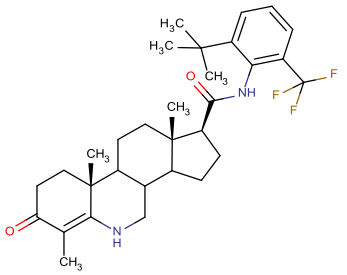
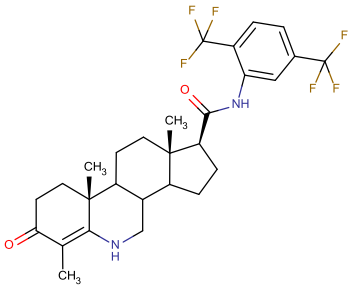
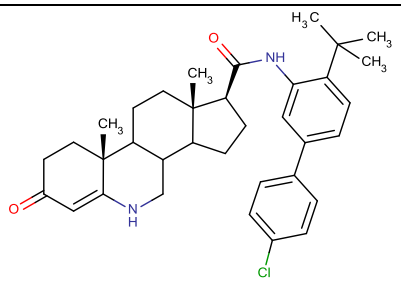
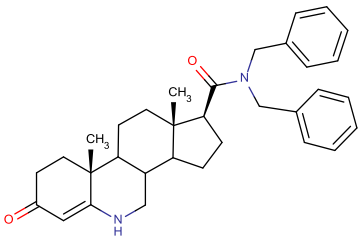
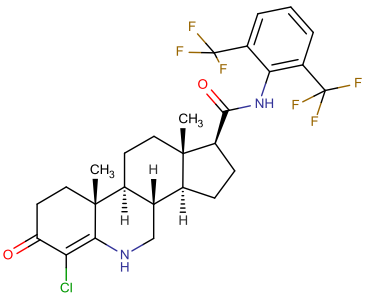
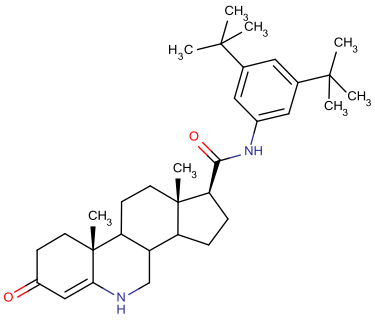
Tabela 36. Estrutura dos compostos atribuídos à classe de IC50 Muito Bom para a 5α-R2 pelos métodos de classificação de árvores de decisão e SVM.

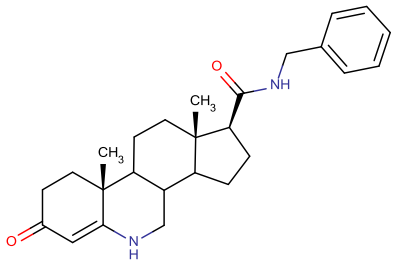
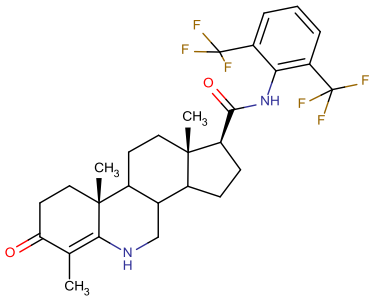
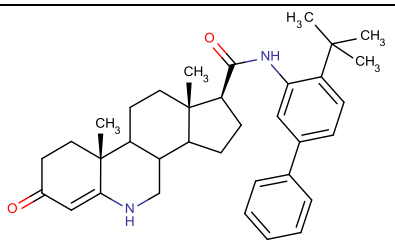
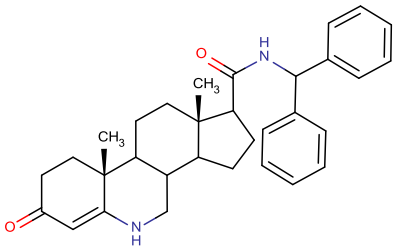
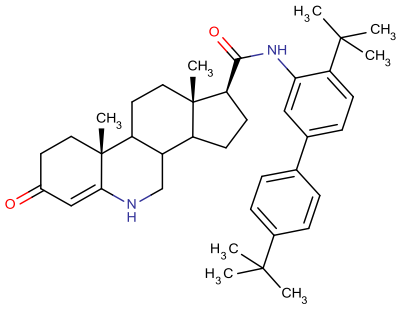
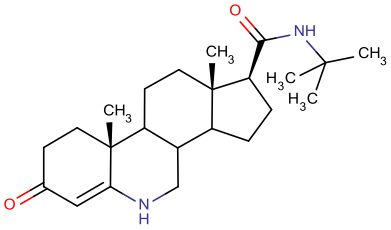
Classe de IC50	ID	ChEMBL ID	Estrutura
Muito Bom	2	ChEMBL108151	
Muito Bom	2	ChEMBL1200969	
Muito Bom	2	ChEMBL108325	

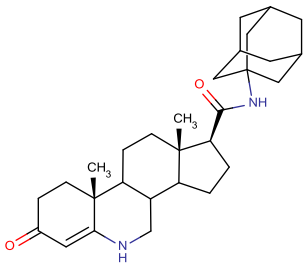
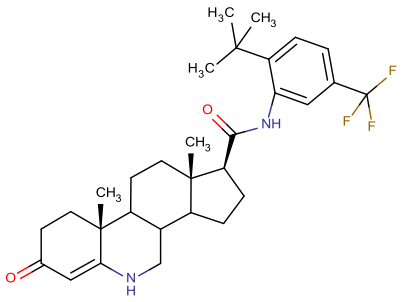
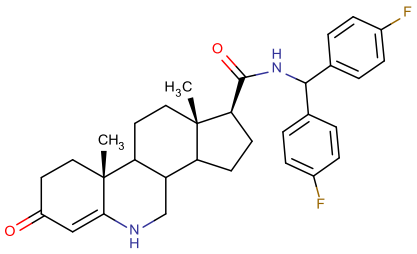
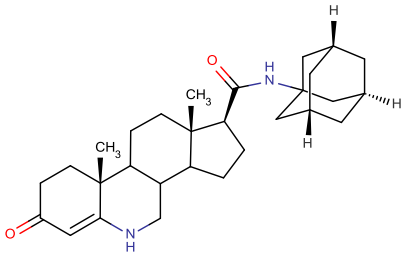
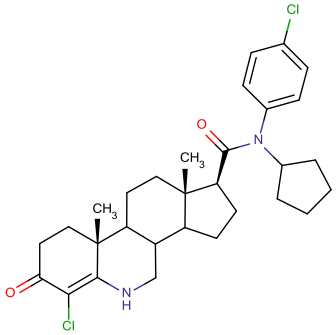
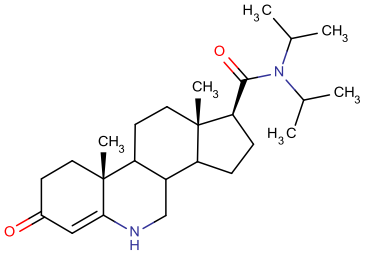
Muito Bom	2	ChEMBL107719	
Muito Bom	6	ChEMBL1762024	
Muito Bom	6	ChEMBL1762026	
Muito Bom	6	ChEMBL2115222	
Bom	1	ChEMBL89240	
Bom	1	ChEMBL88032	
Bom	1	ChEMBL88485	

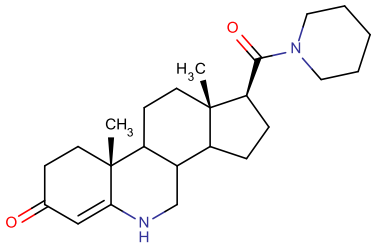
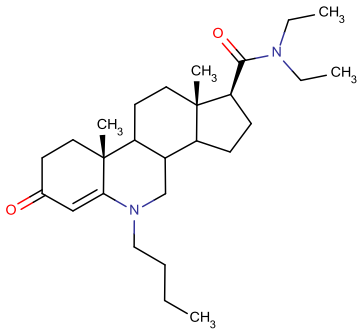
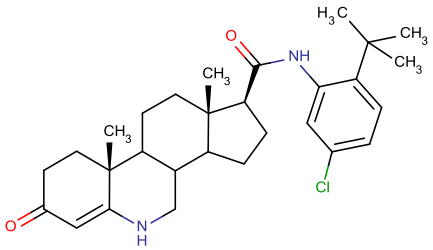
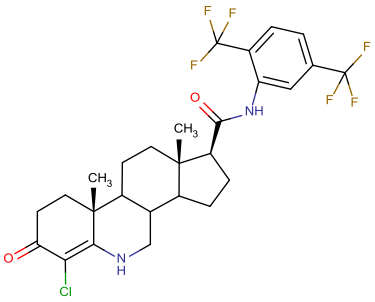
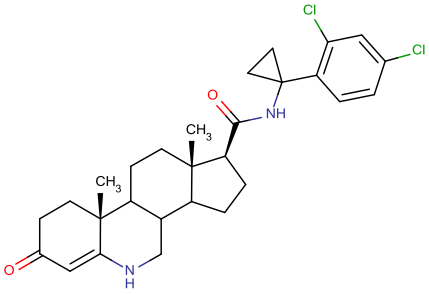
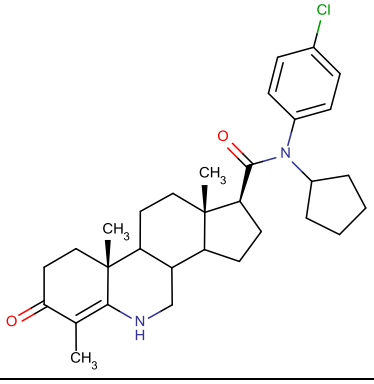
Bom	I	ChEMBL884944	
Bom	I	ChEMBL314725	
Bom	I	ChEMBL893277	
Bom	I	ChEMBL780644	
Bom	I	ChEMBL305855	
Bom	I	ChEMBL412425	
Bom	I	ChEMBL904440	

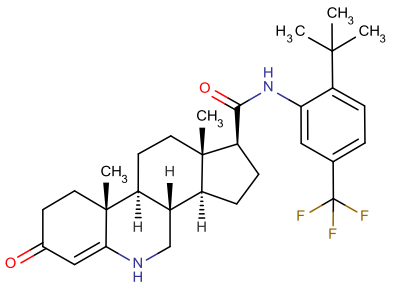
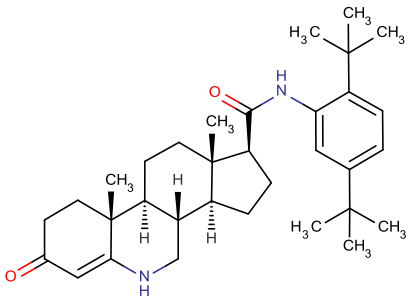
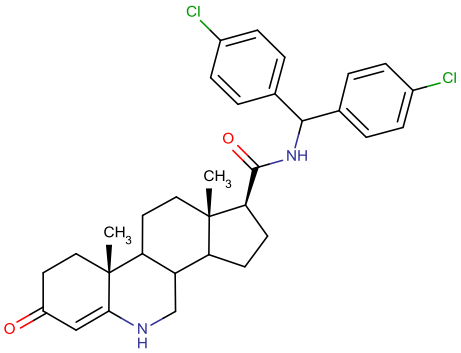
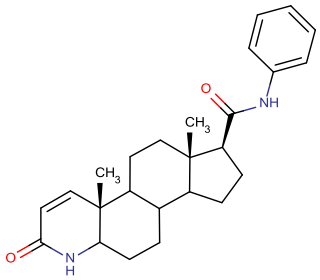
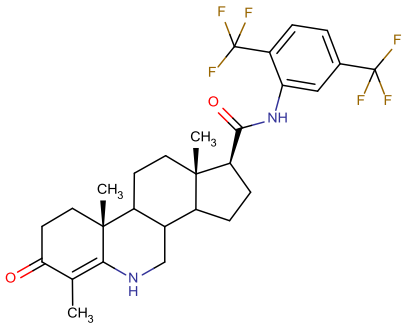
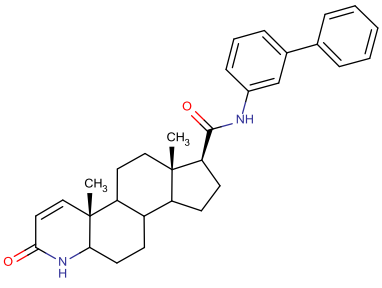
Bom	I	ChEMBL314953	
Bom	I	ChEMBL314069	
Bom	I	ChEMBL314977	
Bom	I	ChEMBL306289	
Bom	I	ChEMBL327702	
Bom	I	ChEMBL308258	

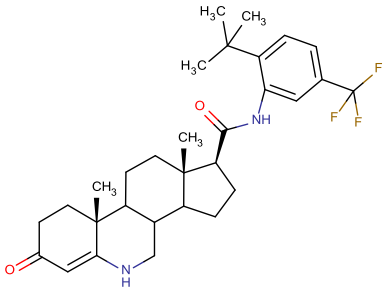
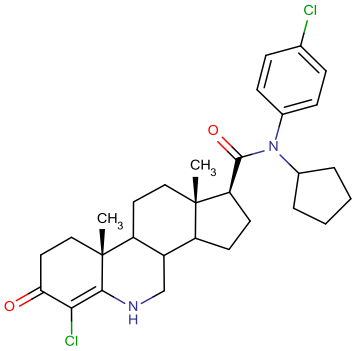
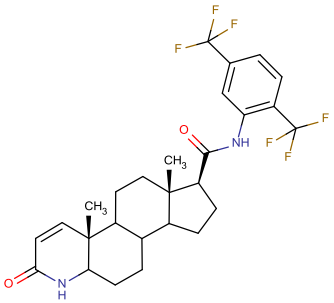
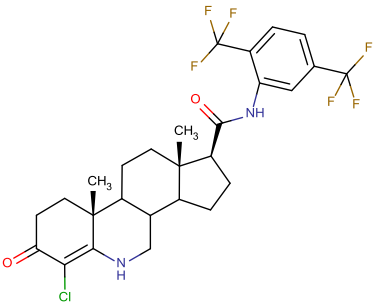
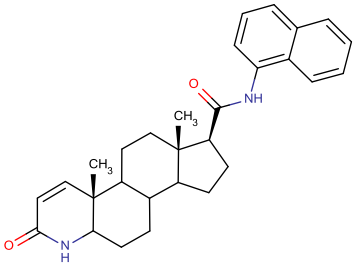
Bom	1	ChEMBL421735	
Bom	1	ChEMBL251833	
Bom	1	ChEMBL328670	
Bom	1	ChEMBL436215	
Bom	1	ChEMBL2115222	
Bom	1	ChEMBL25516	

Bom	I	ChEMBL135780	
Bom	I	ChEMBL445627	
Bom	I	ChEMBL87082	
Bom	I	ChEMBL323996	
Bom	I	ChEMBL87805	
Bom	I	ChEMBL76192	

Bom	I	ChEMBL770244	
Bom	I	ChEMBL277224	
Bom	I	ChEMBL306722	
Bom	I	ChEMBL137733	
Bom	I	ChEMBL283430	
Bom	I	ChEMBL137691	

Bom	I	ChEMBL771290	
Bom	I	ChEMBL773328	
Bom	I	ChEMBL870211	
Bom	I	ChEMBL282037	
Bom	I	ChEMBL895577	
Bom	I	ChEMBL283123	

Bom	1	ChEMBL1762024	
Bom	1	ChEMBL1762026	
Bom	2	ChEMBL412425	
Bom	2	ChEMBL25664	
Bom	2	ChEMBL25183	
Bom	2	ChEMBL264316	

Bom	2	ChEMBL277224	
Bom	2	ChEMBL283430	
Bom	2	ChEMBL283245	
Bom	2	ChEMBL282037	
Bom	2	ChEMBL285599	

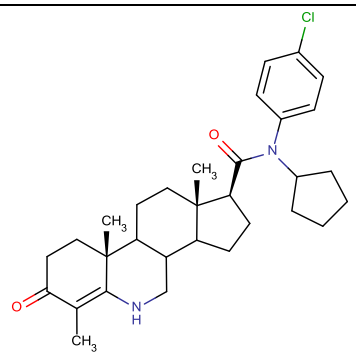
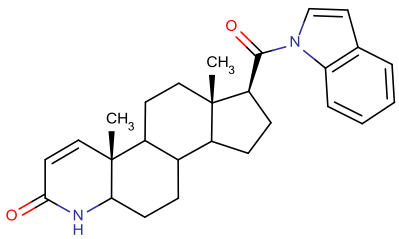
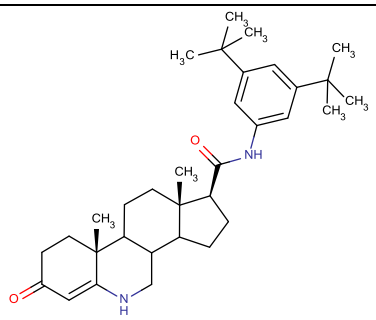
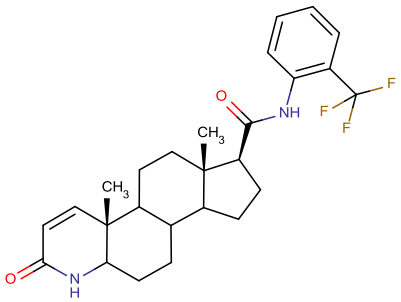
Bom	2	ChEMBL283123	
Bom	2	ChEMBL24465	
Bom	2	ChEMBL25516	
Bom	2	ChEMBL278787	

Tabela 37. ChEMBL IDs dos compostos analisados para a 5α-R1.

ChEMBL ID	ChEMBL ID	ChEMBL ID	ChEMBL ID	ChEMBL ID
CHEMBL107561	CHEMBL125899	CHEMBL108305	CHEMBL118091	CHEMBL120310
CHEMBL107654	CHEMBL126968	CHEMBL108559	CHEMBL118118	CHEMBL120923
CHEMBL107719	CHEMBL128023	CHEMBL108612	CHEMBL118283	CHEMBL122027
CHEMBL109170	CHEMBL132086	CHEMBL108625	CHEMBL118319	CHEMBL122290
CHEMBL109560	CHEMBL134357	CHEMBL108875	CHEMBL118390	CHEMBL122359
CHEMBL114245	CHEMBL135288	CHEMBL109088	CHEMBL118416	CHEMBL1237306
CHEMBL116432	CHEMBL136097	CHEMBL109089	CHEMBL118446	CHEMBL1237307
CHEMBL116830	CHEMBL155094	CHEMBL109232	CHEMBL118447	CHEMBL124803
CHEMBL117159	CHEMBL16027	CHEMBL110058	CHEMBL118452	CHEMBL124849
CHEMBL118100	CHEMBL16743	CHEMBL110288	CHEMBL118543	CHEMBL125073
CHEMBL118255	CHEMBL24955	CHEMBL110398	CHEMBL118573	CHEMBL125209

CHEMBL118389	CHEMBL25083	CHEMBL111142	CHEMBL118605	CHEMBL125256
CHEMBL118943	CHEMBL25578	CHEMBL111407	CHEMBL118822	CHEMBL125257
CHEMBL119255	CHEMBL25664	CHEMBL111476	CHEMBL118824	CHEMBL125260
CHEMBL119361	CHEMBL266519	CHEMBL114211	CHEMBL119138	CHEMBL125442
CHEMBL124750	CHEMBL274826	CHEMBL116370	CHEMBL119164	CHEMBL125481
CHEMBL125443	CHEMBL108028	CHEMBL116493	CHEMBL119259	CHEMBL125579
CHEMBL125599	CHEMBL108151	CHEMBL116604	CHEMBL119722	CHEMBL125668
CHEMBL125855	CHEMBL108181	CHEMBL116999	CHEMBL1201841	CHEMBL125769
CHEMBL126686	CHEMBL137040	CHEMBL17473	CHEMBL274238	CHEMBL279206
CHEMBL127702	CHEMBL138580	CHEMBL17245	CHEMBL274269	CHEMBL279417
CHEMBL127760	CHEMBL14575	CHEMBL17268	CHEMBL274276	CHEMBL279420
CHEMBL128303	CHEMBL152401	CHEMBL17422	CHEMBL27549	CHEMBL279131
CHEMBL128728	CHEMBL153258	CHEMBL17506	CHEMBL276527	CHEMBL280387
CHEMBL129017	CHEMBL155568	CHEMBL17514	CHEMBL277563	CHEMBL280954
CHEMBL129130	CHEMBL15917	CHEMBL180235	CHEMBL278049	CHEMBL281204
CHEMBL130845	CHEMBL16152	CHEMBL180451	CHEMBL278957	CHEMBL283245
CHEMBL131341	CHEMBL16248	CHEMBL181196	CHEMBL283564	CHEMBL283508
CHEMBL131803	CHEMBL16494	CHEMBL182391	CHEMBL290823	CHEMBL283981
CHEMBL134114	CHEMBL16330	CHEMBL182399	CHEMBL335919	CHEMBL285439
CHEMBL134810	CHEMBL16456	CHEMBL182787	CHEMBL336854	CHEMBL285599
CHEMBL134819	CHEMBL16533	CHEMBL182945	CHEMBL340961	CHEMBL29166
CHEMBL135014	CHEMBL16690	CHEMBL183007	CHEMBL414428	CHEMBL294630
CHEMBL135351	CHEMBL16927	CHEMBL183167	CHEMBL419602	CHEMBL300446
CHEMBL135651	CHEMBL16700	CHEMBL183185	CHEMBL56908	CHEMBL30223
CHEMBL135993	CHEMBL16784	CHEMBL183395	CHEMBL710	CHEMBL308610
CHEMBL25072	CHEMBL16863	CHEMBL183692	CHEMBL76959	CHEMBL308819
CHEMBL251109	CHEMBL17025	CHEMBL183448	CHEMBL77117	CHEMBL311013
CHEMBL251297	CHEMBL17193	CHEMBL183500	CHEMBL86418	CHEMBL311624
CHEMBL251506	CHEMBL17194	CHEMBL183820	CHEMBL99615	CHEMBL312466
CHEMBL251507	CHEMBL17206	CHEMBL24191	CHEMBL275153	CHEMBL321442
CHEMBL25193	CHEMBL17222	CHEMBL24193	CHEMBL276320	CHEMBL321617
CHEMBL25448	CHEMBL277664	CHEMBL24464	CHEMBL276355	CHEMBL321867
CHEMBL264316	CHEMBL278490	CHEMBL24465	CHEMBL276871	CHEMBL321923
CHEMBL269632	CHEMBL278787	CHEMBL248682	CHEMBL277053	CHEMBL322592
CHEMBL277403	CHEMBL278981	CHEMBL24958	CHEMBL277091	CHEMBL322749
CHEMBL279134	CHEMBL333684	CHEMBL344363	CHEMBL439908	CHEMBL56581
CHEMBL323031	CHEMBL334649	CHEMBL362379	CHEMBL439910	CHEMBL57789
CHEMBL323033	CHEMBL336200	CHEMBL36573	CHEMBL441287	CHEMBL58989
CHEMBL323788	CHEMBL338207	CHEMBL400059	CHEMBL441471	CHEMBL76373
CHEMBL323856	CHEMBL338376	CHEMBL411431	CHEMBL443995	CHEMBL76687
CHEMBL323996	CHEMBL338707	CHEMBL414427	CHEMBL51628	CHEMBL77167
CHEMBL324210	CHEMBL339912	CHEMBL415016	CHEMBL55260	CHEMBL77169
CHEMBL324687	CHEMBL340027	CHEMBL417109	CHEMBL55261	CHEMBL77649
CHEMBL325572	CHEMBL340181	CHEMBL417627	CHEMBL55402	CHEMBL77650
CHEMBL326439	CHEMBL340305	CHEMBL420415	CHEMBL56517	CHEMBL77789
CHEMBL326743	CHEMBL340684	CHEMBL421627	CHEMBL56518	CHEMBL80175
CHEMBL327042	CHEMBL340715	CHEMBL424084	CHEMBL432745	CHEMBL96006
CHEMBL327098	CHEMBL341004	CHEMBL425878	CHEMBL432977	CHEMBL99448
CHEMBL331253	CHEMBL341088	CHEMBL432439	CHEMBL332781	CHEMBL333183
CHEMBL332407	CHEMBL332413			

Tabela 38. ChEMBL IDs dos compostos analisados para a 5α-R2.

ChEMBL ID	ChEMBL ID	ChEMBL ID	ChEMBL ID	ChEMBL ID
CHEMBL710	CHEMBL1794821	CHEMBL118283	CHEMBL152401	CHEMBL264316
CHEMBL290823	CHEMBL269632	CHEMBL118389	CHEMBL153258	CHEMBL266519
CHEMBL1201841	CHEMBL274238	CHEMBL118390	CHEMBL155094	CHEMBL274269
CHEMBL24955	CHEMBL276871	CHEMBL118416	CHEMBL155568	CHEMBL274276
CHEMBL25448	CHEMBL107565	CHEMBL119722	CHEMBL1627395	CHEMBL274826
CHEMBL24291	CHEMBL107654	CHEMBL120923	CHEMBL1642918	CHEMBL275153
CHEMBL283981	CHEMBL107719	CHEMBL122359	CHEMBL1642919	CHEMBL276320
CHEMBL297524	CHEMBL108028	CHEMBL124750	CHEMBL1642920	CHEMBL276355
CHEMBL412425	CHEMBL108151	CHEMBL124803	CHEMBL16533	CHEMBL276527
CHEMBL1237306	CHEMBL108181	CHEMBL125209	CHEMBL166881	CHEMBL27658
CHEMBL135651	CHEMBL108305	CHEMBL125260	CHEMBL16690	CHEMBL277053
CHEMBL135993	CHEMBL108559	CHEMBL125579	CHEMBL16700	CHEMBL277563
CHEMBL136097	CHEMBL108612	CHEMBL126686	CHEMBL16743	CHEMBL278167
CHEMBL24464	CHEMBL108625	CHEMBL126968	CHEMBL16784	CHEMBL278490
CHEMBL24470	CHEMBL108875	CHEMBL127702	CHEMBL16863	CHEMBL278787
CHEMBL25083	CHEMBL109088	CHEMBL127760	CHEMBL168765	CHEMBL278860
CHEMBL25516	CHEMBL109089	CHEMBL128023	CHEMBL16927	CHEMBL279134
CHEMBL277224	CHEMBL109170	CHEMBL128303	CHEMBL169621	CHEMBL279206
CHEMBL277403	CHEMBL109232	CHEMBL128342	CHEMBL17025	CHEMBL279745
CHEMBL280155	CHEMBL109560	CHEMBL128728	CHEMBL17193	CHEMBL280015
CHEMBL283123	CHEMBL110058	CHEMBL129017	CHEMBL17194	CHEMBL280387
CHEMBL283430	CHEMBL110288	CHEMBL129130	CHEMBL17206	CHEMBL280954
CHEMBL283564	CHEMBL110398	CHEMBL130845	CHEMBL17222	CHEMBL282037
CHEMBL300446	CHEMBL111142	CHEMBL131341	CHEMBL17245	CHEMBL282133
CHEMBL306722	CHEMBL111407	CHEMBL132972	CHEMBL17268	CHEMBL283245
CHEMBL307626	CHEMBL111476	CHEMBL135313	CHEMBL17422	CHEMBL284434
CHEMBL308258	CHEMBL113726	CHEMBL135351	CHEMBL17473	CHEMBL284645
CHEMBL310710	CHEMBL114114	CHEMBL135780	CHEMBL17506	CHEMBL285439
CHEMBL312531	CHEMBL114594	CHEMBL136571	CHEMBL17514	CHEMBL285599
CHEMBL314769	CHEMBL114948	CHEMBL136863	CHEMBL180235	CHEMBL295396
CHEMBL323996	CHEMBL115019	CHEMBL137040	CHEMBL181196	CHEMBL296685
CHEMBL420020	CHEMBL115123	CHEMBL137691	CHEMBL182391	CHEMBL297174
CHEMBL73816	CHEMBL115298	CHEMBL137733	CHEMBL182399	CHEMBL297697
CHEMBL76192	CHEMBL115438	CHEMBL138173	CHEMBL182787	CHEMBL298709
CHEMBL76648	CHEMBL1159458	CHEMBL138225	CHEMBL182945	CHEMBL301725
CHEMBL77328	CHEMBL116370	CHEMBL138580	CHEMBL183167	CHEMBL305730
CHEMBL1237294	CHEMBL116621	CHEMBL201762	CHEMBL183185	CHEMBL305855
CHEMBL14220	CHEMBL117643	CHEMBL202871	CHEMBL183448	CHEMBL305932
CHEMBL14575	CHEMBL117793	CHEMBL203352	CHEMBL183500	CHEMBL306289
CHEMBL117881	CHEMBL14195	CHEMBL24091	CHEMBL183820	CHEMBL306507
CHEMBL25072	CHEMBL25572	CHEMBL24191	CHEMBL1908332	CHEMBL306554
CHEMBL25183	CHEMBL25578	CHEMBL24193	CHEMBL201138	CHEMBL306781
CHEMBL25193	CHEMBL25593	CHEMBL24465	CHEMBL201425	CHEMBL306948
CHEMBL25196	CHEMBL25664	CHEMBL24929	CHEMBL201699	CHEMBL307181
CHEMBL25282	CHEMBL262635	CHEMBL24958	CHEMBL201736	CHEMBL307398
CHEMBL307745	CHEMBL308107	CHEMBL308532	CHEMBL309261	CHEMBL307729
CHEMBL309585	CHEMBL327098	CHEMBL76671	CHEMBL420415	CHEMBL45671
CHEMBL310022	CHEMBL327702	CHEMBL76693	CHEMBL421627	CHEMBL47679
CHEMBL310336	CHEMBL328670	CHEMBL76803	CHEMBL421696	CHEMBL47785

CHEMBL310711	CHEMBL332781	CHEMBL76892	CHEMBL421735	CHEMBL47959
CHEMBL310822	CHEMBL334649	CHEMBL77024	CHEMBL422046	CHEMBL50113
CHEMBL311045	CHEMBL335805	CHEMBL77124	CHEMBL424066	CHEMBL55402
CHEMBL311577	CHEMBL338376	CHEMBL77129	CHEMBL425878	CHEMBL56518
CHEMBL311783	CHEMBL339912	CHEMBL77144	CHEMBL430617	CHEMBL56908
CHEMBL311818	CHEMBL340027	CHEMBL77290	CHEMBL432439	CHEMBL74661
CHEMBL311874	CHEMBL340181	CHEMBL77378	CHEMBL432641	CHEMBL75320
CHEMBL312143	CHEMBL340305	CHEMBL77765	CHEMBL432729	CHEMBL75926
CHEMBL314069	CHEMBL340684	CHEMBL77789	CHEMBL432751	CHEMBL75982
CHEMBL314309	CHEMBL340715	CHEMBL78003	CHEMBL432977	CHEMBL76030
CHEMBL314725	CHEMBL340961	CHEMBL78064	CHEMBL435332	CHEMBL76260
CHEMBL314953	CHEMBL341004	CHEMBL78108	CHEMBL436215	CHEMBL76536
CHEMBL314977	CHEMBL341088	CHEMBL78161	CHEMBL442747	CHEMBL76637
CHEMBL318512	CHEMBL344363	CHEMBL78275	CHEMBL370016	CHEMBL323033
CHEMBL321442	CHEMBL347059	CHEMBL78382	CHEMBL382214	CHEMBL323788
CHEMBL321617	CHEMBL352434	CHEMBL80160	CHEMBL382649	CHEMBL324210
CHEMBL321867	CHEMBL352942	CHEMBL80163	CHEMBL40831	CHEMBL324687
CHEMBL321923	CHEMBL353524	CHEMBL80520	CHEMBL411431	CHEMBL325572
CHEMBL322592	CHEMBL355567	CHEMBL80669	CHEMBL414427	CHEMBL326439
CHEMBL322749	CHEMBL362379	CHEMBL45241	CHEMBL414428	CHEMBL443995
CHEMBL323031	CHEMBL36392	CHEMBL36573	CHEMBL415016	CHEMBL445627
CHEMBL417627	CHEMBL44785			

Tabela 39. Descritores moleculares calculados para a 5α-R1 e 5α-R2.

Descritores Moléculares	**Descritores Moláculares**
Número de átomos	Número de átomos (C)
Número de átomos (H)	Número de átomos (N)
Número de átomos (O)	Número de átomos (Cl)
Número de átomos (F)	Número de átomos (Br)
Número de átomos (S)	Massa exacta
Polarizabilidade molecular	Número de átomos alifáticos
Número de ligações alifáticas	Número de anéis alifáticos
Número de átomos aromáticos	Número de anéis aromáticos
Número de anéis aromáticos com 6 carbonos	Número de anéis aromáticos com 5 carbonos
Número de anéis assimétricos	Número de anéis carboalifático
Número de anéis carboaromáticos	Número de átomos em cadeia
Número de anéis alifáticos unidos	Número de anéis aromáticos unidos
número de anéis unidos	Número de anéis heteroalifáticos
Número de anéis heteroaromático	Número de aneis hetero
ASA	ASA-
ASA H	ASA P
ASA+	Área de superfície polar
Número de anéis	Área de superfície Van der Waals (3D)
Volume van der Waals	LogP
Número de aceitadores de hidrogénio	Número de dadores de hodrigenio

Tabela 40. ChEMBL IDs dos novos compostos obtidos para a subestrutura 1 (Tabela 23) na 5α-R1.

ChEMBL ID	ChEMBL ID	ChEMBL ID	ChEMBL ID
CHEMBL451551	CHEMBL414198	CHEMBL8313	CHEMBL268149
CHEMBL274826	CHEMBL8315	CHEMBL269277	CHEMBL267696
CHEMBL273330	CHEMBL8333	CHEMBL266519	CHEMBL267913
CHEMBL273746	CHEMBL8417	CHEMBL267358	CHEMBL265076
CHEMBL274878	CHEMBL98904	CHEMBL269279	CHEMBL268918
CHEMBL275930	CHEMBL9104	CHEMBL298313	CHEMBL263957
CHEMBL275556	CHEMBL8646	CHEMBL8342	CHEMBL266867
CHEMBL414201	CHEMBL10077	CHEMBL8498	CHEMBL268748
CHEMBL6786	CHEMBL10106	CHEMBL10493	CHEMBL276665
CHEMBL7088	CHEMBL9672	CHEMBL10525	CHEMBL275727
CHEMBL6383	CHEMBL9199	CHEMBL268115	CHEMBL275764
CHEMBL6402	CHEMBL9948	CHEMBL9709	CHEMBL273496
CHEMBL6539	CHEMBL10018	CHEMBL37	CHEMBL6558
CHEMBL6579	CHEMBL10163	CHEMBL275284	CHEMBL10343
CHEMBL6600	CHEMBL10177	CHEMBL276004	CHEMBL266532
CHEMBL6629	CHEMBL10474	CHEMBL267703	CHEMBL10660
CHEMBL414006	CHEMBL10533	CHEMBL268209	CHEMBL440832
CHEMBL404519	CHEMBL9887	CHEMBL267720	CHEMBL9779
CHEMBL7365	CHEMBL10752	CHEMBL268340	CHEMBL9072
CHEMBL8254	CHEMBL417973	CHEMBL9114	CHEMBL10751
CHEMBL415301	CHEMBL430283	CHEMBL266350	CHEMBL273768
CHEMBL416516	CHEMBL427969	CHEMBL267668	CHEMBL441173
CHEMBL7447			

Tabela 41. ChEMBL IDs dos novos compostos obtidos para a subestrutura 9, (Tabela 24) na 5α-R1.

ChEMBL ID	ChEMBL ID	ChEMBL ID	ChEMBL ID
CHEMBL1077498	CHEMBL16533	CHEMBL17514	CHEMBL1077489
CHEMBL1077487	CHEMBL17193	CHEMBL1079183	CHEMBL1079512
CHEMBL1077618	CHEMBL17194	CHEMBL1079055	CHEMBL1077494
CHEMBL1077499	CHEMBL17206	CHEMBL1078880	CHEMBL1078856
CHEMBL1078047	CHEMBL17222	CHEMBL266519	CHEMBL1077530
CHEMBL1077496	CHEMBL17245	CHEMBL277053	CHEMBL1079056
CHEMBL276527	CHEMBL16700	CHEMBL280015	CHEMBL1077500
CHEMBL1077484	CHEMBL17268	CHEMBL1077526	CHEMBL1077505
CHEMBL1078531	CHEMBL17506	CHEMBL1077538	CHEMBL1077510
CHEMBL1077516	CHEMBL16863	CHEMBL1077509	CHEMBL1077522
CHEMBL1077492	CHEMBL16927	CHEMBL1077486	CHEMBL1077485
CHEMBL1078188	CHEMBL17422	CHEMBL275153	CHEMBL1077715
CHEMBL1077537	CHEMBL17473	CHEMBL278490	CHEMBL1078470
CHEMBL1077497	CHEMBL1077527	CHEMBL274269	CHEMBL1077523
CHEMBL1077531	CHEMBL1077507	CHEMBL274276	CHEMBL1077521
CHEMBL417627	CHEMBL1077511	CHEMBL276320	CHEMBL1077529
CHEMBL1078082	CHEMBL1077528	CHEMBL276355	CHEMBL1077524
CHEMBL1078664	CHEMBL1077520	CHEMBL1077400	CHEMBL1077514

CHEMBL1077488	CHEMBL1077536	CHEMBL1077535	CHEMBL1077532
CHEMBL1077504	CHEMBL1077493	CHEMBL1078653	CHEMBL1077534
CHEMBL1077506	CHEMBL1077501	CHEMBL1078432	CHEMBL1077533
CHEMBL1077508	CHEMBL1077525	CHEMBL1079154	CHEMBL1077495
CHEMBL1078605	CHEMBL1079298	CHEMBL1077515	

Tabela 42. ChEMBL IDs dos novos compostos obtidos para a subestrutura 2, (Tabela 26) na 5α-R2.

ChEMBL ID	ChEMBL ID	ChEMBL ID	ChEMBL ID
CHEMBL432977	CHEMBL1455412	CHEMBL1200969	CHEMBL323031
CHEMBL110058	CHEMBL1200969	CHEMBL326439	CHEMBL327098
CHEMBL321442	CHEMBL326439	CHEMBL108305	CHEMBL108612
CHEMBL108151	CHEMBL108305	CHEMBL109560	CHEMBL109232
CHEMBL1455412	CHEMBL109560	CHEMBL108325	CHEMBL111407
CHEMBL1200969	CHEMBL108325	CHEMBL432977	CHEMBL107565
CHEMBL326439	CHEMBL108875	CHEMBL110058	CHEMBL107719
CHEMBL108305	CHEMBL109088	CHEMBL321442	CHEMBL108181
CHEMBL109560	CHEMBL109089	CHEMBL108151	CHEMBL24191
CHEMBL108325	CHEMBL107654	CHEMBL108151	CHEMBL324687
CHEMBL432977	CHEMBL411431	CHEMBL1455412	CHEMBL321923
CHEMBL110058	CHEMBL325572	CHEMBL321442	

Tabela 43. ChEMBL IDs dos novos compostos obtidos para a subestrutura 6, (Tabela 26) na 5α-R2.

ChEMBL ID	ChEMBL ID	ChEMBL ID	ChEMBL ID
CHEMBL77765	CHEMBL76260	CHEMBL77378	CHEMBL24088
CHEMBL1762039	CHEMBL76892	CHEMBL76803	CHEMBL24291
CHEMBL80669	CHEMBL73816	CHEMBL80520	CHEMBL136863
CHEMBL307181	CHEMBL280155	CHEMBL1762023	CHEMBL308532
CHEMBL78275	CHEMBL312143	CHEMBL1762024	CHEMBL432729
CHEMBL80163	CHEMBL442747	CHEMBL1762025	CHEMBL2115222
CHEMBL77144	CHEMBL135313	CHEMBL1762026	CHEMBL75982
CHEMBL305932	CHEMBL307626	CHEMBL1762031	CHEMBL76648
CHEMBL310710	CHEMBL421696	CHEMBL1762032	CHEMBL1762035
CHEMBL311818	CHEMBL76671	CHEMBL1762033	CHEMBL1762036
CHEMBL311783	CHEMBL76536	CHEMBL1762034	CHEMBL1762037
CHEMBL24033	CHEMBL75320	CHEMBL1762038	
CHEMBL24729			

Tabela 44. ChEMBL IDs dos novos compostos obtidos para a subestrutura 1, (Tabela 27) na 5α-R2.

ChEMBL ID	ChEMBL ID	ChEMBL ID	ChEMBL ID
CHEMBL89240	CHEMBL25183	CHEMBL77129	CHEMBL137691
CHEMBL88032	CHEMBL24729	CHEMBL282037	CHEMBL421696
CHEMBL88485	CHEMBL328970	CHEMBL77378	CHEMBL76671
CHEMBL88494	CHEMBL436215	CHEMBL89557	CHEMBL77290
CHEMBL314725	CHEMBL2115222	CHEMBL283123	CHEMBL77328
CHEMBL89327	CHEMBL25516	CHEMBL80520	CHEMBL87021
CHEMBL78064	CHEMBL135780	CHEMBL1762023	CHEMBL310710
CHEMBL307398	CHEMBL445627	CHEMBL1762024	CHEMBL310711
CHEMBL305855	CHEMBL87082	CHEMBL1762025	CHEMBL311818
CHEMBL306554	CHEMBL323996	CHEMBL1762026	CHEMBL308258
CHEMBL412425	CHEMBL87805	CHEMBL1762031	CHEMBL311783
CHEMBL90440	CHEMBL76192	CHEMBL1762032	CHEMBL421735
CHEMBL314953	CHEMBL77024	CHEMBL1762034	CHEMBL24033
CHEMBL314069	CHEMBL277224	CHEMBL1762035	CHEMBL24088
CHEMBL314977	CHEMBL442747	CHEMBL1762036	CHEMBL314309
CHEMBL306289	CHEMBL306722	CHEMBL1762037	CHEMBL283430
CHEMBL327702	CHEMBL137733	CHEMBL1762038	CHEMBL1762033
CHEMBL1762039			

Tabela 45. ChEMBL IDs dos novos compostos obtidos para a subestrutura 2, (Tabela 27) na 5α-R2**.**

ChEMBL ID	ChEMBL ID	ChEMBL ID	ChEMBL ID
CHEMBL291158	CHEMBL27658	CHEMBL46353	CHEMBL27862
CHEMBL290372	CHEMBL283546	CHEMBL42362	CHEMBL295375
CHEMBL295240	CHEMBL283564	CHEMBL50325	CHEMBL283217
CHEMBL412425	CHEMBL300970	CHEMBL45169	CHEMBL293561
CHEMBL297601	CHEMBL301123	CHEMBL43832	CHEMBL290823
CHEMBL25664	CHEMBL284645	CHEMBL47855	CHEMBL279017
CHEMBL24033	CHEMBL298214	CHEMBL47892	CHEMBL278167
CHEMBL24088	CHEMBL292398	CHEMBL45394	CHEMBL282627
CHEMBL25072	CHEMBL300878	CHEMBL41529	CHEMBL431134
CHEMBL25183	CHEMBL296231	CHEMBL57762	CHEMBL278283
CHEMBL24729	CHEMBL297760	CHEMBL45539	CHEMBL262635
CHEMBL431331	CHEMBL422046	CHEMBL24249	CHEMBL57922
CHEMBL264316	CHEMBL28673	CHEMBL50382	CHEMBL40474
CHEMBL277224	CHEMBL24969	CHEMBL48105	CHEMBL36573
CHEMBL46716	CHEMBL24825	CHEMBL24465	CHEMBL40361
CHEMBL283430	CHEMBL25282	CHEMBL25516	CHEMBL40855
CHEMBL283245	CHEMBL25572	CHEMBL431730	CHEMBL47293
CHEMBL292699	CHEMBL25592	CHEMBL278787	CHEMBL47804
CHEMBL282037	CHEMBL25593	CHEMBL278860	CHEMBL48815
CHEMBL285599	CHEMBL42470	CHEMBL47470	CHEMBL277430
CHEMBL24191	CHEMBL57468	CHEMBL45227	CHEMBL276871
CHEMBL283123	CHEMBL44626	CHEMBL297697	CHEMBL60677
CHEMBL47261	CHEMBL24464	CHEMBL45468	CHEMBL47305