UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE QUÍMICA

ESTUDOS QSAR DE COMPOSTOS COM ATIVIDADE LEISHMANICIDA

Universidade Estadual de Campinas
Instituto de Química - Depto. Físico-Química

Kesley Moraes Godinho de Oliveira

Orientador: Prof. Dr. Yuji Takahata
Co-Orientador: Prof. Dr. Rogério Custódio

Área: Físico-química

Campinas
2009

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO INSTITUTO DE QUÍMICA DA UNICAMP

OL42e

Oliveira, Kesley Moraes Godinho de.
Estudos QSAR de compostos com atividade leishmanicida / Kesley Moraes Godinho de Oliveira. -- Campinas, SP: [s.n], 2009.

Orientador: Yuji Takahata.
Co-orientador: Rogério Custódio.

Tese - Universidade Estadual de Campinas, Instituto de Química.

1. QSAR(Bioquímica). 2. Leishmaniose visceral. 3. Nucleosídeos. 4. Antifúngicos. I. Takahata, Yuji. II. Custódio, Rogério. III. Universidade Estadual de Campinas. Instituto de Química. IV. Título.

Título em inglês: QSAR studies of compounds with leishmanicidal activity

Palavras-chaves em inglês: QSAR, Visceral leishmaniasis, Nucleosides, Antifungals

Área de concentração: Físico-Química

Titulação: Doutor em Ciências

Banca examinadora: Yuji Takahata (orientador), Anderson Coser Gáudio (IF-UFES), Antonia Tavares do Amaral (IQ-USP-SP), Pedro Antonio Muniz Vazquez (IQ-UNICAMP), Munir Salomão Skaf (IQ-UNICAMP)

Data de defesa: 26/06/2009

À Amaurinda Flora de Oliveira
cujo amor, compreensão e companheirismo
faz com que os desafios pareçam menores.

Porque para Deus não haverá impossíveis
em todas as Suas promessas.
Lc. 1:37

Agradecimentos

À Deus por ter tornado possível todas as coisas em minha vida.

Ao prof. Yuji Takahata pela oportunidade, incentivo constante, compreensão, paciência e pelos preciosos ensinamentos não apenas científicos, mas também para a vida.

Ao prof. Rogério Custódio pela co-orientação deste trabalho e pelo exemplo de amabilidade.

Ao meu marido pelo amor, companheirismo, apoio e respeito demonstrados, principalmente, nos meus momentos mais difíceis.

Aos meus pais, Rubens e Leonice, pelo suporte, por tudo que vocês representam em minha vida e pelo exemplo de luta, persistência e amor.

À Amaurinda e Paulo (*in memoriam*) pela extraordinária dedicação à minha família, amor, cumplicidade e o auxílio em todos os momentos.

Aos meus filhos, Ana Beatriz e Ian, fontes de motivação, vida, amor e alegria.

Ao Luciano e Rogério pelo apoio.

Aos colegas de grupo Edílson Borges, Maria Cristina Costa, Maximiliano Segala e André Okamoto pela amizade.

Ao Dr. Edson Lima, Diretor da Divisão Farmoquímica do laboratório Cristália, pelo total apoio e incentivo à concretização deste trabalho.

À Fundação de Amparo a Pesquisa e Desenvolvimento do Estado de São Paulo (FAPESP) pelo suporte financeiro.

Curriculum Vitæ

Dados Pessoais

Nome: Kesley Moraes Godinho de Oliveira
Nome em citações bibliográficas: Oliveira, KMG

Atuação Profissional

2002 – Atual: Pesquisadora na área de Modelagem Molecular, Divisão Farmoquímica, Laboratório Cristália Produtos Químicos Farmacêuticos Ltda.

Especialização

2006 – 2007: Capacitação em Gerenciamento de Projetos.
Universidade Estadual de Campinas, UNICAMP, Campinas, Brasil
Projeto de Conclusão: Ampliação e Automação de Salas de Pesagem.

Formação Acadêmica/Titulação

1996 – 1999: Mestrado em Física.
Universidade Estadual de Campinas, UNICAMP, Campinas, Brasil
Título: Estudos de Difração de Raios-X a alta resolução da Beta-lactoglobulina bovina
Ano de obtenção: 1999
Orientador: Igor Polikarpov
Bolsista da: Fundação de Amparo à Pesquisa do Estado de São Paulo
Palavras-chave: Cristalografia, Proteínas, Difração de Raios-X, Beta lactoglobulina bovina, Transição de Tanford
Áreas do conhecimento: Difração de Raios-X, Biofísica molecular

1992 – 1996: Bacharelado em Física.
Universidade Estadual de Campinas, UNICAMP,Campinas, Brasil
Bolsista da: Fundação de Amparo à Pesquisa do Estado de São Paulo

Patentes (Depósito PCT)

WO 2005111006 Ritonavir analogous compound useful as retroviral protease inhibitor, preparation of the ritonavir analogous compound and pharmaceutical composition for the ritonavir analogous compound.

Artigos completos em periódicos

1.Oliveira, K.M.G., Valente.Mesquita, V. L., Botelho, M. M., Sawyer, L., Ferreira, S. T., Polikarpov, I. Crystal structures of bovine beta.lactoglobulin in the orthorhombic space group $C22_21$. Structural differences between genetic variants A and B and features of the Tanford transition. *Eur. J. Bioche*m., 268, p.477.483, 2001.

2. Botelho, M.M., Valente.Mesquita, V.L., Oliveira, K.M.G., Polikarpov, I., Ferreira, S. T. Pressure denaturation of beta.lactoglobulin. Different stabilities of isoforms A and B, and an investigation of the Tanford transition. *Eur. J. Bioche*m., 267, p.2235.2241, 2000

3. Guedes, S. ; Hadler, J. C. ; Sarkis, J. E. S. ; Oliveira, K. M. G. ; Kakazu, M. H. ; Iunes, P. J. ; Saiki, M. ; Tello S., C.A. ; Paulo, S. R. Spontaneous-fission decay constant of U-238 measured by nuclear track techniques without neutron irradiation. *Journal of Radioanalytical and Nuclear Chemistry,* Dordrecht, v. 258, n. 1, p. 117-122, 2003.

4. Guedes, S. ; Hadler, J. C. ; Iunes, P. J. ; Oliveira, K. M. G. ; Moreira, P. A.; F. P. ; Tello S., C.A. Kinetic model for the annealing of fission tracks in zircon. *Radiation Measurements,* Oxford, v. 40, n. 2-6, p. 517-521, 2005.

5. Guedes, S. ; Hadler, J. C. ; Oliveira, K. M. G. ; Moreira, P. A. F. P. ; Iunes, P. J.; Tello S., C.A. Kinetic model for the annealing of fission tracks in minerals and its application to apatite. *Radiation Measurements,* Oxford, v. 41, n. 4, p. 392-398, 2006.

6. Oliveira, K. M. G.; Takahata, Y. QSAR modeling of nucleosides against amastigotes of *Leishmania donovani* using logistic regression and classification tree. *QSAR Comb. Sci.*, v. 27, n. 8, p. 1020-1027, 2008.

Trabalhos resumidos publicados em anais de evento

1.Oliveira, K.M.G., Takahata, Y. Estudo sobre tiosemicarbazonas de 3-carboxi.beta.carbolinas In: 23a. Reunião Anual da Sociedade Brasileira de Química, 2000, Poços de Caldas. Livro de Resumos/ Vol. 1, Brasil. São Paulo: Sociedade Brasileira de Química, 2000. QT041

2.Oliveira, K.M.G., Takahata, Y.Determinação da estrutura química de derivados de 4-(3'-4'-X-fenilamino)-1,3-dimetil 1H pirazolo[3,4.b]piridinas In: X Simpósio Brasileiro de Química Teórica, 1999, Caxambu. Livro de Resumos. Rio de Janeiro: Divisão gráfica da UFRJ, 1999. p.372.

Resumo

ESTUDOS QSAR DE COMPOSTOS COM ATIVIDADE LEISHMANICIDA

O objeto de estudo desta tese é a forma mais severa e letal de Leishmaniose: a Leishmaniose Visceral, LV, cujo principal agente etiológico é a espécie *Leishmania donovani*. O objetivo deste trabalho é o desenvolvimento de modelos Quantitativos das Relações Estrutura-Atividade para duas séries de compostos com atividade anti-leishmaniose contra formas amastigotas de *Leishmania donovani*. A primeira série envolve vinte e um análogos de nucleosídeos pirazolo-pirimidínicos e a segunda série compreende oito antifúngicos. Para garantir a robustez dos modelos, além dos compostos que compõem as respectivas séries, foram selecionadas moléculas com conhecida atividade contra LV para compor um conjunto de teste e avaliar a capacidade de previsão dos respectivos modelos. Para a série dos nucleosídeos foram desenvolvidos modelos de regressão logística e árvore de classificação. Em ambas as abordagens os descritores Mor26v e o GAP(HOMO, HOMO-1) se mostraram relevantes para a explicação da atividade leishmanicida destes compostos. O modelo de regressão logística atingiu 90,5% para acurácia de classificação para o conjunto de trabalho e 58% para o conjunto de teste após a análise do domínio de aplicabilidade do modelo. O modelo para árvore de classificação alcançou 95% para acurácia de classificação para o conjunto de trabalho e 83% para o conjunto de teste. Para a série dos antifúngicos foi utilizado um modelo de regressão linear múltipla onde a energia eletrônica e a área da superfície polar se mostraram importantes para a atividade leishmanicida da série. Os valores previstos exibem 98% de correlação com os valores experimentais. Os valores encontrados para o conjunto de teste também estão de acordo com a literatura. Finalmente, novos compostos foram propostos para síntese e avaliação da atividade leishmanicida.

Abstract

QSAR STUDIES OF COMPOUNDS WITH LEISHMANICIDAL ACTIVITY

The object of study of this thesis is the most severe and lethal form of leishmaniasis: visceral leishmaniasis, LV, whose main etiological agent is the species *Leishmania donovani*. The goal of this work is the development of Quantitative Structure-Activity Relationship models for two series of compounds presenting anti-leishmanial activity against amastigotes of *Leishmania donovani*. The first series includes twenty-one analogues of pyrazolo-pyrimidine nucleosides and the second series comprises eight antifungals. To ensure the robustness of the models, in addition to the compounds that make up their series, molecules with known activity against LV were selected to compose a testset and evaluate the predictive power of their models. For the series of nucleosides logistic regression models and tree classification were developed. In both approaches, the Mor26v and GAP(HOMO, HOMO-1) descriptors were relevant to explain the leishmanicidal activity of these compounds. The logistic regression model reached 90.5% for classification accuracy to the workingset and 58% for testset after analysis of the domain of applicability of the model. The classification tree model reached 95% for classification accuracy of the workingset and 86% for the testset. A multiple linear regression model was applied to the series of antifungal agents. The electron energy and polar surface area were important for the leishmanicidal activity of this series. The predicted values showed 98% of correlation with the experimental values. The values found for the testset are also in agreement with the literature. Finally, new compounds were proposed for synthesis and evaluation of leishmanicidal activity.

Índice

Lista de Abreviaturas

AIDS Sigla inglesa para SÍndrome da Imunodeficiência Adquirida.

AUC Sigla inglesa para Área Sob a Curva.

C Concentração molar de um composto.

ED_{50} Concentração molar correspondente a 50% da dose efetiva.

HIV Sigla inglesa para Vírus Humano da Imunodeficiência.

HOMO Orbital molecular mais alto ocupado.

IC_{50} Concentração inibitória média (concentração que reduz o efeito em 50%).

Ki Constante de dissociação de um inibidor.

LC Leishmaniose cutânea.

LD_{50} Dose letal média (concentração letal para 50% da população de animais de teste em um tempo prescrito).

LUMO Orbital molecular mais baixo desocupado.

LV Leishmaniose visceral.

MCMM Sigla inglesa *Monte Carlo Multiple Minimum* usada para especificar um método de análise conformacional.

MMFF Sigla inglesa *Molecular Merck Force Field* usada para especificar um tipo de campo de força na Mecânica Molecular.

PKDL Sigla inglesa para Leishmaniose dérmica pós-kalazar.

PM3 Sigla Inglesa para o método semiempírico *Parametric Model 3*.

QSAR Sigla inglesa para Relações quantitativas entre estrutura química e atividade biológica.

ROC Sigla inglesa *Receiver Operating Curve*.

RMN Sigla inglesa para Ressonância Magnética Nuclear.

RMSD Sigla inglesa para Desvio Quadrático Médio.

SM5.42R Sigla inglesa *Solvent Model 5.42R* utilizada para especificar o modelo de solvatação.

UFS Sigla inglesa *Unsupervised Foward Selection* utilizada para especificar um tipo de método para seleção de variáveis.

π Constante de lipofilicidade relativa.

σ Parâmetro eletrônico (constante de Hammett).

Lista de Tabelas

Lista de Figuras

Capítulo 1

Doenças infecciosas tropicais — A Leishmaniose

Introdução

Este capítulo mostra o quadro atual da leishmaniose no Brasil e no mundo bem como as opções terapêuticas disponíveis. O objeto de estudo deste trabalho é a forma mais severa e letal de leishmaniose: a leishmaniose visceral, LV. O principal agente etiológico da LV é a espécie *Leishmania donovani*.

Duas séries de compostos com atividade leishmanicida contra formas amastigotas de *Leishmania donovani* serão utilizadas na elaboração de relações quantitativas estrutura-atividade (modelos QSAR). Ao final do capítulo, serão apresentadas as metas para alcançar estes objetivos.

1.1 Aspectos epidemiológicos

A leishmaniose é causada pelo protozoário do gênero *Leishmania* (ver Figura 1-1). É uma patologia endêmica em 88 países na zona tropical e sub-tropical afetando 12 milhões de pessoas nas regiões mais pobres do globo e causa 51.000 mortes anualmente (ver Figura 1-2).

Dentre as 30 espécies existentes, 20 infectam humanos que são hospedeiros casuais ao contrário de outros mamíferos, como roedores e cães, que são reservatórios naturais do parasito.

A transmissão ocorre através da picada da fêmea do mosquito do gênero *Phlebotomus*. De um modo geral, a doença se manifesta em duas formas: cutânea (variações: leishmaniose cutâneo-mucosa e cutâneo difusa) e visceral (variações: leishmaniose dérmica pós-calazar).

A forma cutânea (ver Figura 1-3) causa desfigurações e a forma visceral (ver Figura 1-4), também conhecida como calazar e febre negra, é a mais severa apresentando altos índices de mortalidade. A leishmaniose visceral é causada pelo complexo de espécies *Leishmania donovani* (i.e., *L donovani* e *L infantum* na Europa e *L chagasi* nas Américas), *Leishmania tropica* (Europa) e *Leishmania amazonensis* (Américas) que afetam baço, fígado, mucosa do intestino delgado, medula óssea e nodos linfáticos (WHO, 2004; HERWALDT, 1999).

No Brasil, são confirmados 3500 casos por ano de leishmaniose visceral[1] sendo que 56% ocorrem na região Nordeste (CEARÁ, 2007) em função da precariedade das condições sanitárias, mas a doença existe na maioria dos outros estados do país. Inicialmente, as áreas rurais eram as mais afetadas, mas os registros de casos em centros urbanos aumentaram em função da intensificação da urbanização (GOTIJO e MELO, 2004).

[1] No Brasil a Leishmaniose é uma doença de notificação obrigatória.

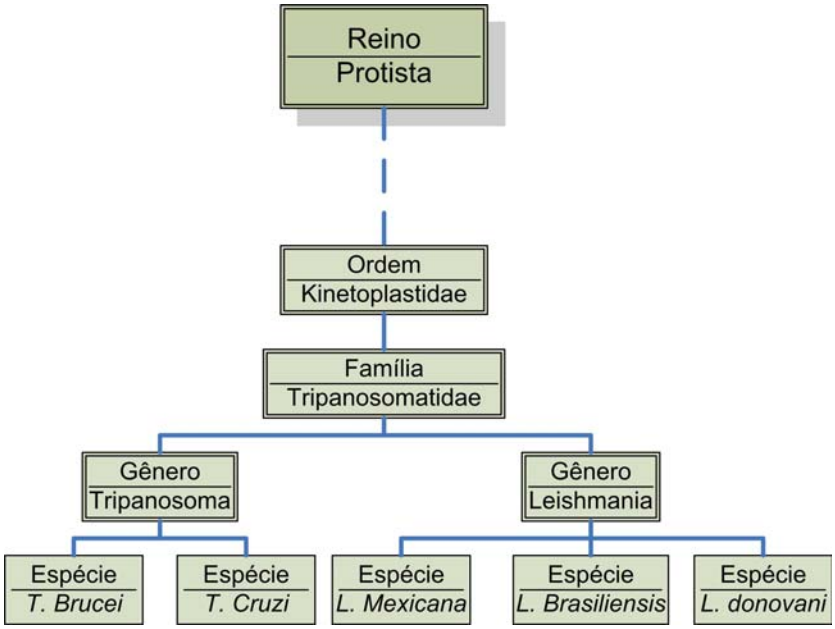


Figura 1-1: Taxonomia do parasito *Leishmania*.

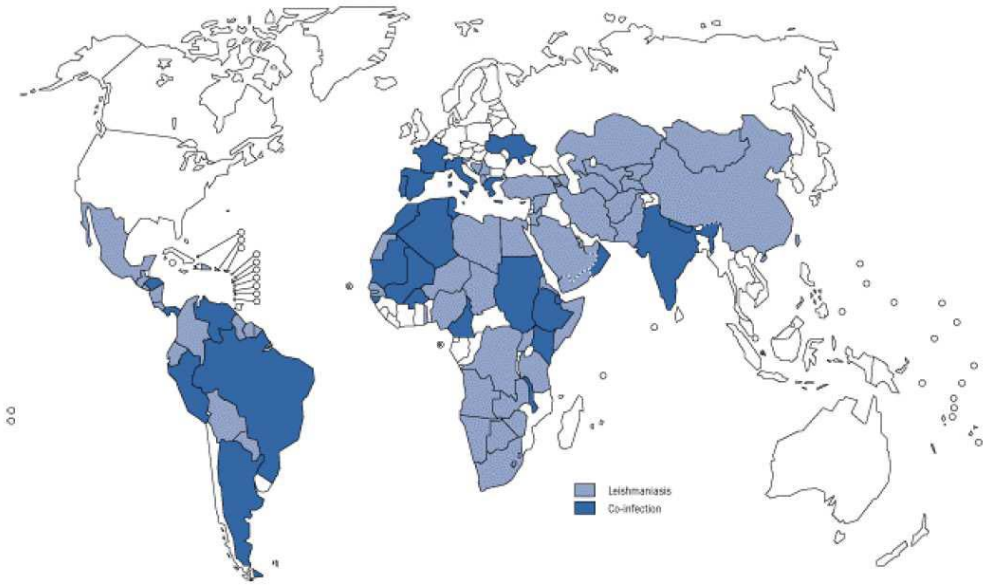


Figura 1-2: Mapa de prevalência de leishmaniose e co-infecção leishmaniose/HIV entre 1990 e 1998[2].

[2] Extraído de WHO-2000.

Figura 1-3: Leishmaniose cutânea[3].

Figura 1-4: Esplenomegalia na Leishmaniose visceral[4].

O diagnóstico da doença se baseia na demonstração de amastigotas de *Leishmania* na medula óssea ou através de biópsia do baço ou do fígado. Estes procedimentos são invasivos e, freqüentemente, não sensíveis dificultando o diagnóstico rápido em estágios precoces da doença.

A leishmaniose visceral tem despertado o interesse internacional em função da infecção oportunista freqüente em pacientes portadores do vírus HIV-1 e a letalidade de pacientes que sofreram transplantes, principalmente, renais acarretando a migração desta forma patológica para novas áreas geográficas (ver Figura 1-2). No sudeste da Europa, 70% dos casos de LV entre adultos estão associados com infecção por HIV. Usuários de drogas injetáveis são os mais expostos a este risco (WHO, 2007; HARMS e FELDMEIER, 2002). Na França,

[3] Extraído de <http://www.ops-oms.org/English/AD/DPC/CD/leish-fotos2.htm>. Acesso em 19.Mar.2008.
[4] Extraído de <http://www.keele.ac.uk/depts/aep/staff/rdcmres.htm>. Acesso em 19.Mar.2008.

77% dos casos de LV são de pacientes provenientes do sudeste europeu que sofreram transplantes renais (BASSET et al., 2005).

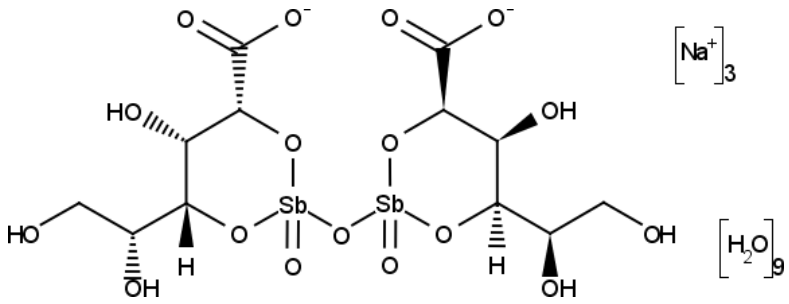
1.2 Terapias disponíveis

A forma mais eficiente de combate a leishmaniose ainda é a eliminação do vetor. As terapias amplamente utilizadas foram introduzidas há mais de 50 anos (ver Figura 1-5). Apresentam alta toxicidade e efeitos colaterais indesejados somados ao modo de administração parenteral e o tempo prolongado de tratamento, tipicamente de várias semanas (HERWALDT, 1999; CROFT et al., 2006).

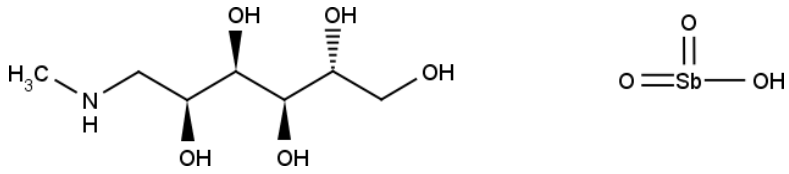
A maior preocupação com estes fármacos é o desenvolvimento de resistência clínica como é o caso da pentamidina e os antimônios pentavalentes estibogluconato de sódio e antimoniato de meglumina (SINGH et al., 2006). O medicamento lançado mais recentemente é a miltefosina (ver Figura 1-6) que tem demonstrado excelentes resultados de cura (taxa de cura de 94% na Fase Clínica III). No entanto, ela apresenta certas desvantagens como teratogenicidade e toxicidades gastrointestinal e renal reversíveis (SUNDAR e CHATTERJEE, 2006). Infelizmente, pacientes portadores do vírus HIV não apresentam respostas satisfatórias aos tratamentos disponíveis.

1.3 Considerações para o desenvolvimento de antiparasitários

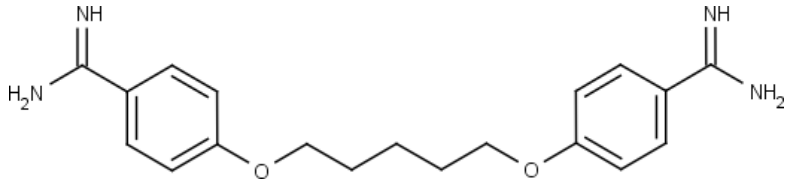
Durante seu ciclo de vida, o parasito se apresenta em duas formas distintas morfológica e bioquimicamente de acordo com o hospedeiro (ver Figura 1-7). A forma promastigota se manifesta no inseto (vetor) e é caracterizada por sua extrema mobilidade assemelhando-se aos hemoflagelados. Nos mamíferos se converte na forma amastigota após ser interiorizada por macrófagos. Esta forma é desprovida de movimento e é resistente às enzimas presentes no interior dos vacúolos fagolisossomais (BALANÃ-FOUCE et al., 1998).



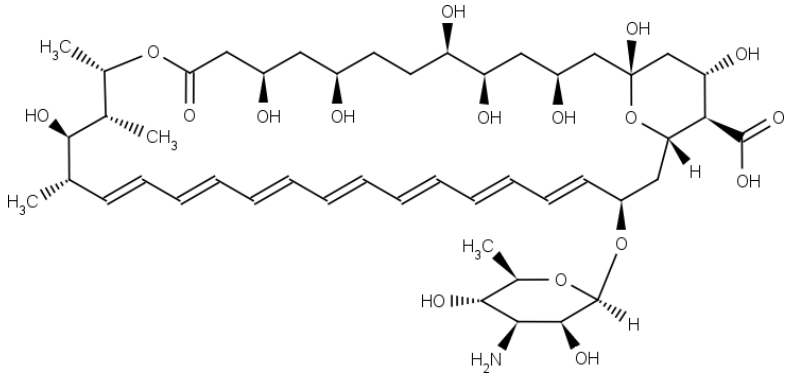
(a)



(b)



(c)



(d)

Figura 1-5: Fármacos utilizados no tratamento de Leishamniose. (a) Estibogluconato de sódio, (b) Antimoniato de meglumina, (c) Pentamidina e (d) Anfotericina B.

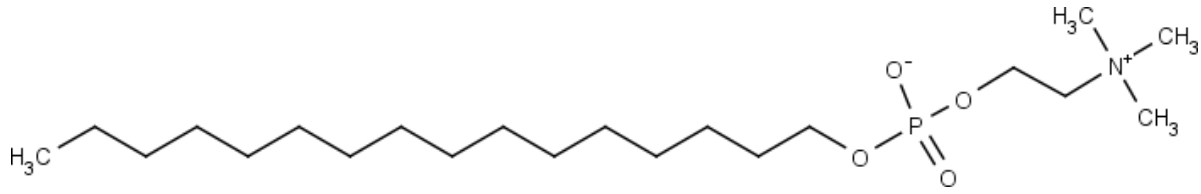


Figura 1-6: Miltefosina, fármaco recentemente lançado para o tratamento de leishmaniose visceral.

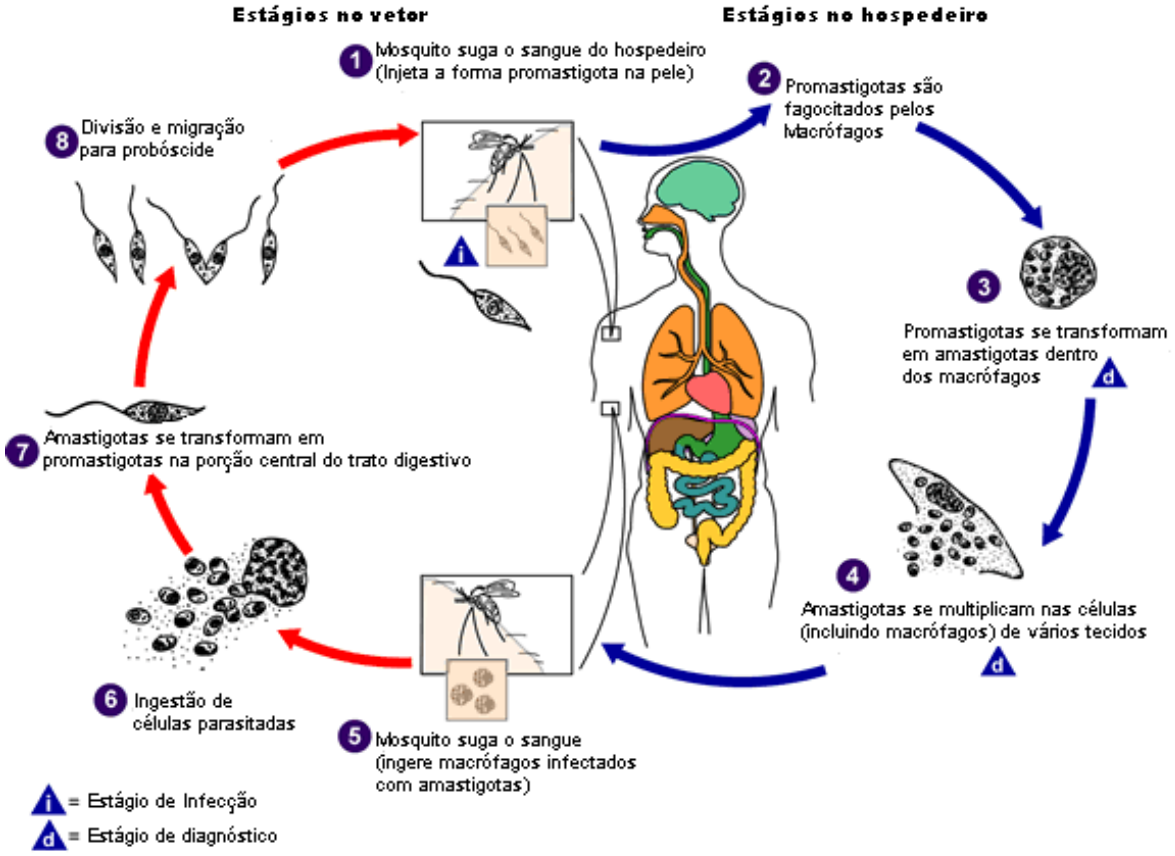


Figura 1-7: Ciclo de vida do parasito *Leishmania*.[5]O parasito muda de forma morfológica e bioquímica de acordo com o hospedeiro. A forma promastigota se manifesta no inseto e a forma amastigota no hospedeiro humano. A transmissão do parasito aos humanos ocorre pela picada do inseto.

[5] Extraído e adaptado de http://www.dpd.cdc.gov/dpdx. Acesso em 19.Mar.08.

Estas diferenças conferem sensibilidade diferenciada frente aos fármacos. A forma de maior interesse nos testes de atividade é a forma amastigota, entretanto, existem dificuldades quanto à obtenção e manutenção deste estágio *in vitro* daí a importância do desenvolvimento de culturas de amastigotas em meios axênicos[6] que viabilizou a avaliação do potencial leishmanicida de compostos (CALLAHAN, 1997 apud CROFT e YARDLEY, 2002).

O desenvolvimento de agentes quimioterápicos deve considerar as diferenças de sensibilidade entre as espécies de *Leishmania* (ver Figura 1-8), o tecido de atuação preferencial de cada espécie do parasito, pois os macrófagos da pele e do fígado, por exemplo, apresentam propriedades diferentes e, finalmente, propriedades farmacocinéticas e farmacodinâmicas adequadas para permear as diferentes membranas celulares incluindo a do próprio parasito.

De acordo com Wang e Wang (2007), os alvos terapêuticos para o tratamento de doenças parasitárias são eleitos considerando as seguintes estratégias: (a) enzimas essenciais encontradas somente nos parasitos, (b) enzimas presentes no hospedeiro e no parasito, mas indispensável somente para o parasito e (c) funções bioquímicas comuns em ambos, mas com diferentes propriedades farmacológicas.

1.4 Seleção do conjunto de dados

Os protozoários são incapazes de sintetizar nucleotídeos purínicos (via *de novo*) sendo dependentes do hospedeiro para obter bases purínicas e nucleosídeos através de vias de salvação[7]. As enzimas envolvidas na via de salvação são exclusivas ou indispensáveis somente para os parasitos. Portanto, análogos de nucleosídeos são candidatos interessantes para as estratégias a e b mencionadas na seção anterior.

[6] Meio livre de contaminantes.

[7] Vias de salvação são vias bioquímicas utilizadas para recuperar bases e nucleosídeos que são formados durante a degradação de RNA e DNA. As bases e nucleosídeos recuperados são convertidos novamente em nucleotídeos.

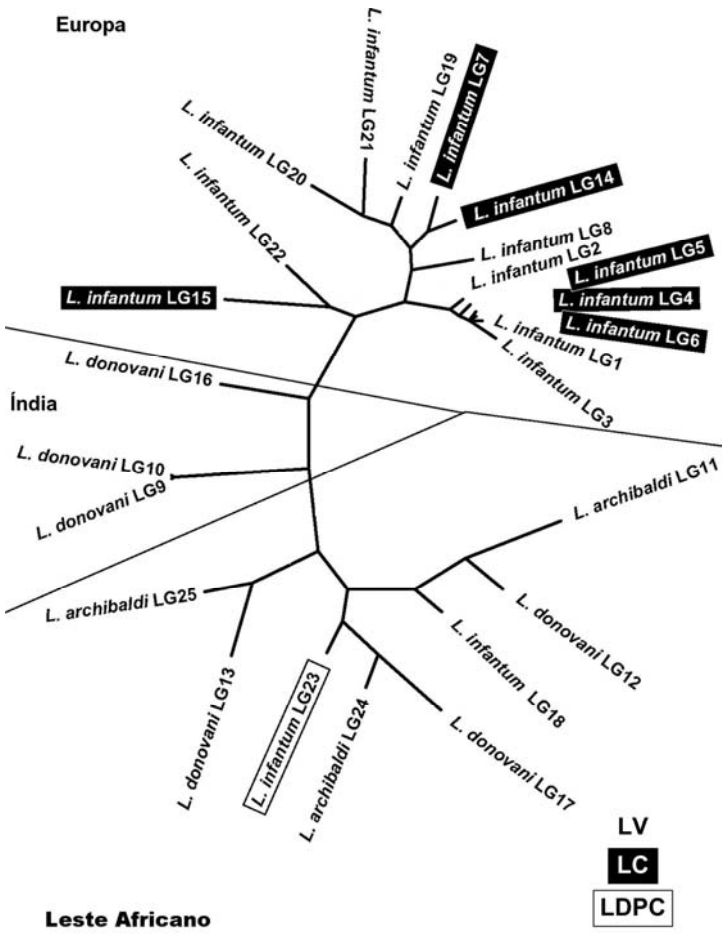


Figura 1-8: Diversidade genética do complexo *Leishmania donovani*. As espécies sem caixa causam a leishmaniose visceral (LV), as espécies nas caixas pretas causam leishmaniose cutânea (LC) e as espécies nas caixas brancas causam leishmaniose dérmica pós-calazar (LDPC)[8]. Na parte superior, central e inferior encontram-se as espécies mais comuns na Europa, Índia e Leste Africano respectivamente. Os códigos LG enumerados de 1 a 25 representam as cepas de parasitos.

No passado, foram conduzidos testes clínicos com o alopurinol — um nucleosídeo utilizado no tratamento da hiperuricemia primária da gota (ver Figura 1-9). Entretanto, o rápido metabolismo e excreção contribuíram para a falta de êxito deste fármaco. Outras tentativas foram conduzidas sem sucesso também com o alopurinol ribosídeo (CROFT e YARDLEY, 2002) (ver Figura 1-10).

[8] Extraído e adaptado de Lukes et al. (2007).

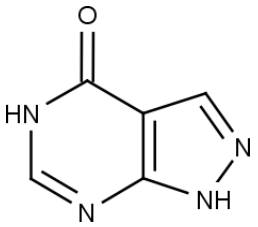


Figura 1-9: Alopurinol.

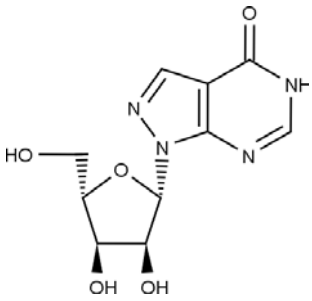


Figura 1-10: Alopurinol ribosídeo.

Insuficiências farmacocinéticas, como estas apresentadas pelos nucleosídeos alopurinol e alopurinol ribosídeo, podem eventualmente ser contornadas através do desenvolvimento de pró-farmacos. Pró-fármacos são compostos farmacologicamente inativos que se convertem para a forma ativa do fármaco dentro do corpo humano. Um exemplo atual de latenciação de nucleosídeo é o tenofovir disoproxila um pró-fármaco do tenofovir que se encontra entre os fármacos de primeira linha no tratamento contra AIDS (ver Figura 1-11). O tenofovir apresenta propriedades anti-HIV, porém baixa biodisponibilidade oral e, somente com o desenvolvimento de um pró-fármaco, veio a se tornar um medicamento (DE CLERCQ, 2005).

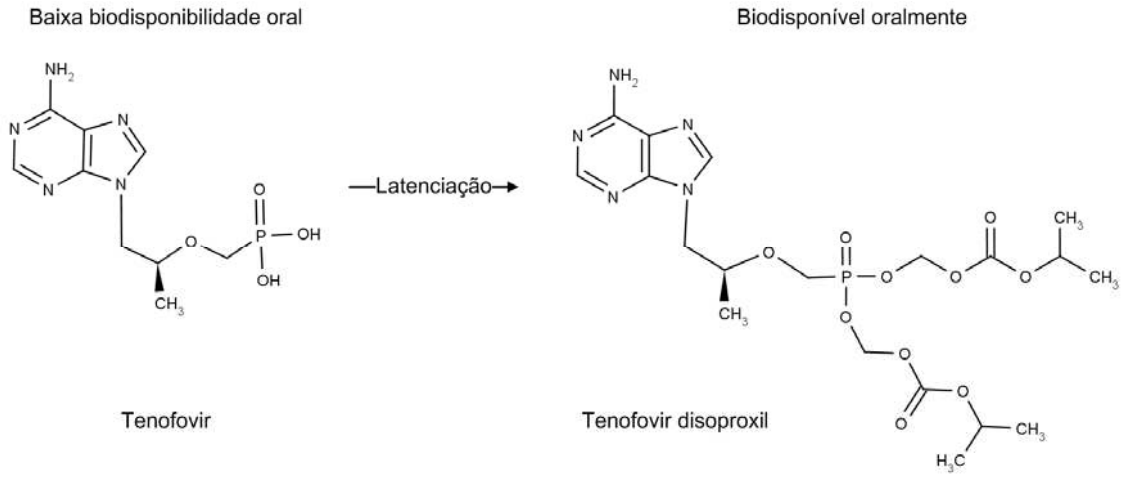


Figura 1-11: Estratégia para tornar o tenofovir biodisponível através da latenciação.

Assim sendo, o primeiro conjunto de dados (ver Figura 1-12) envolve vinte e um análogos de nucleosídeos pirazolo-pirimidínicos obtidos dos trabalhos de Bhakuni e colaboradores (1989, 1990). A atividade biológica foi determinada por meio de testes *in vivo* (hamsters) avaliando a percentagem de inibição da replicação de formas amastigotas de *Leishmania donovani*. O alopurinol foi utilizado como fármaco de referência para os testes. Os ensaios foram conduzidos em uma única dose e expressos em percentagem de inibição o que caracteriza experimentos típicos de triagem exploratória, ou seja, prospecção de protótipos. O alvo macromolecular destes compostos não é conhecido, mas Bhakuni e colaboradores (1989, 1990) sugerem alguma proteína envolvida na via de salvação purínica. Neste caso, o modelo QSAR resultante será do tipo global (ver seção 1.5) e não mecanístico como nas análises clássicas.

As membranas celulares de fungos e *Leishmania* apresentam em sua composição ergosterol em contraste com as células de mamíferos compostas por colesterol. A inibição de enzimas que participam da biossíntese do ergosterol resulta no aumento da permeabilidade da membrana ocasionando a perda de componentes celulares vitais para estes organismos. Assim, fármacos que inibem a síntese do ergosterol, como os antifúngicos, são candidatos apropriados para a estratégia c (ver seção anterior) e mencionados para o tratamento de leishmaniose desde o início da década de 1980.

Na prática clínica atual há relatos de cura da leishmaniose visceral com antifúngicos normalmente em terapia combinada com alopurinol (COLAKOGLU et al., 2006 a e b; KUYUCU et al., 2001, LLORENTE et al., 2000, HALIM et al., 1999). Esta abordagem tem a vantagem de utilizar medicamentos que já estão consolidados no mercado para outras indicações o que caracteriza o chamado uso

off-label[9]. No Brasil, a Agência Nacional de Vigilância Sanitária entende que o uso *off-label* é, por definição, não autorizado, mas isso não implica que seja incorreto (ANVISA, 2005).

O segundo conjunto de dados (ver Figura 1-13) compreende oito antifúngicos desenvolvidos pela companhia Hoffman La Roche e testadas in vitro por Gebre-Hiwot e Frommel (1993). A dose efetiva (ED_{50}) foi calculada a partir da percentagem de inibição de células infectadas em relação ao controle (células infectadas sem a presença dos fármacos). Os valores obtidos estão apresentados na Tabela 1-1. Estes antifúngicos possuem o mesmo modo de ação, ou seja, são inibidores da síntese do ergosterol, porém por mecanismos diferentes que serão detalhados no Capítulo 4. O modelo QSAR para esta série também se enquadra na categoria de modelo global (ver seção 1.5).

[9] Uso *off-label* consiste na prescrição de medicamentos para um propósito diferente do uso original aprovado pela agência regulatória de um determinado país.

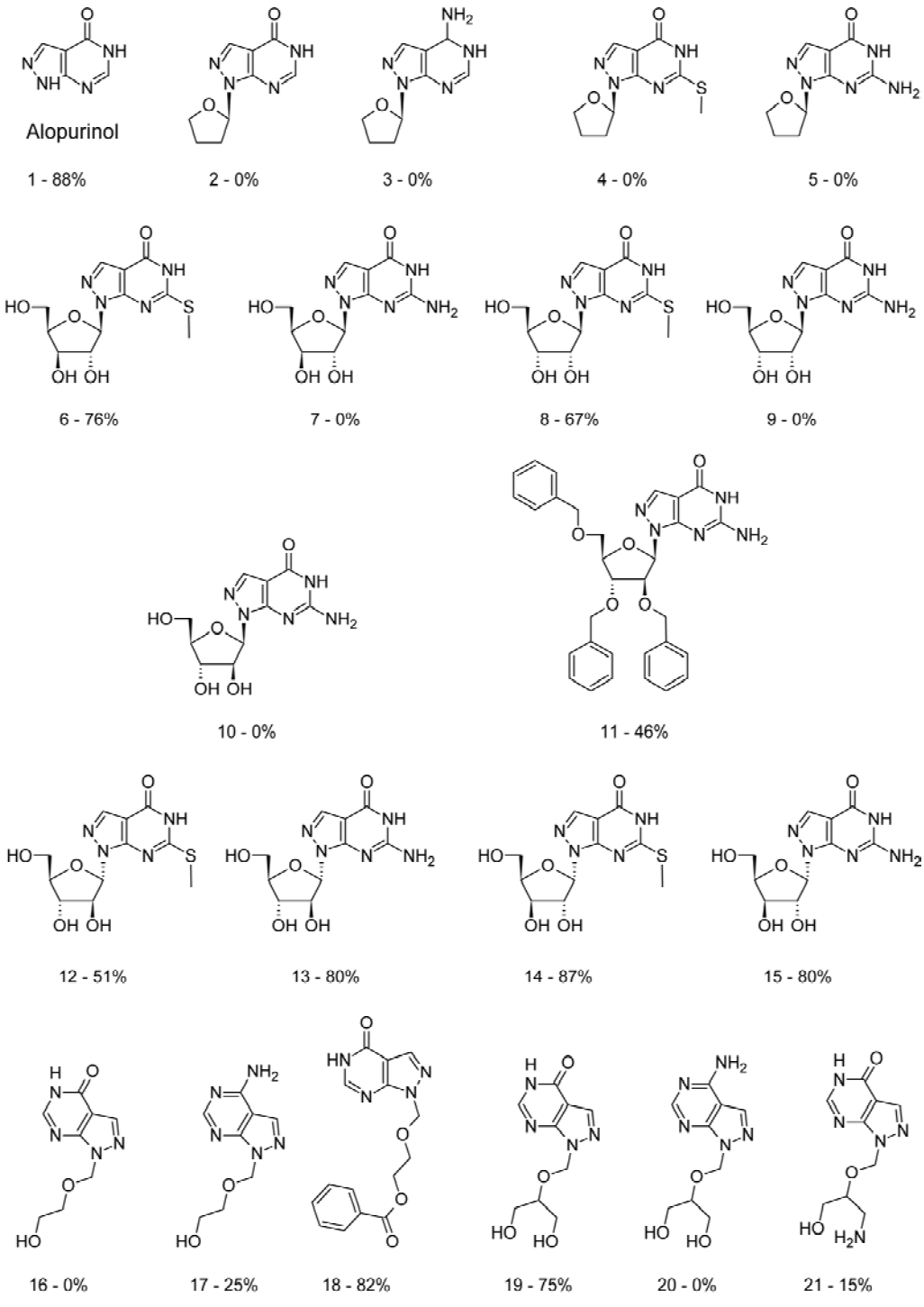


Figura 1-12: Série de nucleosídeos análogos ao alopurinol e avaliados contra formas amastigotas de *Leishmania donovani* (Bhakuni et al., 1989 e 1990).

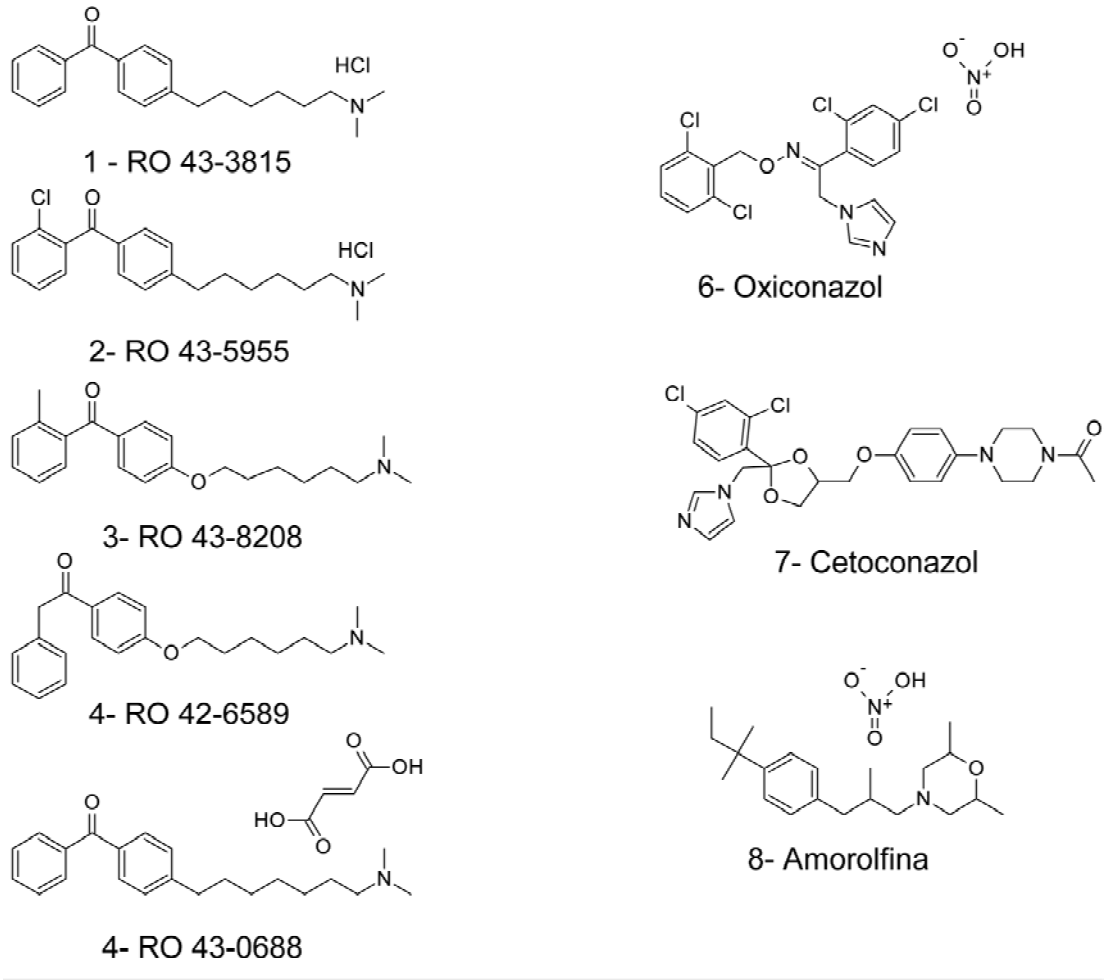


Figura 1-13: Série de antifúngicos avaliados contra *Leishmania donovani* (Gebre-Hiwot e Frommel, 1993).

Tabela 1-1: Atividade in vitro dos antifúngicos contra amastigotas de *Leishmania donovani* (Gebre-Hiwot e Frommel, 1993).

N°	Molécula	ED_{50} (µM)	LD_{50} (µM)	Índice Terapêutico (LD_{50}/ED_{50})
1	RO 43-3815	1,7	7,8	4,6
2	RO-43-5955	2,37	13,1	5,5
3	RO 43-8208	10	29,5	3,0
4	RO 42-6589	6,22	19,7	3,2
5	RO 43-0688	1,54	10,2	6,6
6	Oxiconazol	6,69	13,4	2,0
7	Cetoconazol	> 300	17,7	< 0,1
8	Amorolfina	4,19	38,4	9,2

ED_{50} = Quantidade de composto que produz resposta terapêutica em 50% das células em relação às células de controle. LD_{50} = Dose letal para 50% das células em teste em relação às células de controle.

1.5 Modelos QSAR globais e locais

Mesmo mecanismo de ação é uma restrição imposta ao conjunto de trabalho usado na proposição de modelos QSAR. O principal objetivo desta restrição é trabalhar sob condições controladas garantindo a homogeneidade do conjunto de dados com o intuito de assegurar correlação entre os parâmetros descritores e a atividade biológica. Na ausência de comprovação do preciso mecanismo de ação, usualmente, assume-se que moléculas similares atuam pelo mesmo mecanismo de ação. Modelos QSAR gerados sob esta condição de congenericidade e mesmo mecanismo de ação são denominados modelos QSAR locais (*QSAR local model*).

Uma das vertentes de QSAR é previsão de toxicidade de compostos. Em países desenvolvidos, existem controles regulatórios para substâncias químicas produzidas em larga escala (> 1 ton/ano). Substâncias produzidas nesta escala necessitam de caracterização toxicológica e ecotoxicológica. Esta exigência regulatória imprime uma forte demanda por ensaios. Por esta razão, há uma grande concentração de esforços na elaboração de modelos QSAR com aplicação regulatória que sejam suficientemente robustos para inferir toxicidade de compostos. Uma motivação legítima para o desenvolvimento de modelos deste tipo é evitar a morte desnecessária de animais de laboratório (BENFENATI, 2007).

Neste contexto, mesmo mecanismo de ação foi ampliado para mesmo modo de ação. Por modo de ação, entende-se a mesma resposta fisiológica ou mesmo efeito no organismo. O efeito narcótico, por exemplo, é causado por diferentes mecanismos de ação. Séries não congenéricas (heterólogas) passaram a compor o conjunto de trabalho, pois modelos para previsão de toxicidade contemplam maior variabilidade de estrutural para de fato possuírem aplicação prática. Modelos obtidos sob estas condições são denominados modelos QSAR globais (*QSAR global models*).

Outras propriedades, além da toxicidade, utilizam modelos QSAR globais, as quais são citadas a seguir. Identificação de possíveis ligantes de uma determinada família de proteínas como cinases (XIA et al., 2004) e receptores acoplados à proteína G (*GPCR*) (ROLLAND, 2005) que são grandes famílias de proteínas da maior relevância por estarem envolvidas em vários processos biológicos. Previsões para efeitos farmacodinâmicos e farmacocinéticos como Absorção, Distribuição, Metabolismo e Excreção (também conhecidas por propriedades *ADME*) (BUGRIM et al., 2004).

Tipicamente, 39 % dos compostos que entram em fase clínica falham por problemas de farmacocinética (KENNEDY, 1997). A média dos custos divulgados no desenvolvimento de novos fármacos está em torno de 880 milhões de dólares. Cerca de 75% destes custos foram atribuídos a falhas durante o desenvolvimento (TOLLMAN, 2001 apud SEIFERT, 2003). Embora haja muito para se evoluir em termos de previsão para os sistemas *in silico* (EGAN et al., 2004), a maior motivação para o uso destes modelos está baseada na eliminação de protótipos inaptos logo no início do desenvolvimento. O adágio *fail early, fail cheap* traduz claramente esta necessidade. Estima-se que é possível economizar aproximadamente 15% do montante total investido no desenvolvimento de novos fármacos através de simulações *in silico* (TOLLMAN, 2001 apud SEIFERT, 2003).

Uma das premissas no desenvolvimento de novos fármacos é a busca por ligantes seletivos, ou seja, que interagem com um único alvo macromolecular (proteína). Esta premissa está baseada no fato de que a seletividade reduz potenciais efeitos tóxicos. Para que esta premissa seja aplicada é necessária a validação do alvo terapêutico, ou seja, é preciso demonstrar que uma determinada proteína está envolvida no processo de desenvolvimento de uma patologia e que sua modulação produz efeitos terapêuticos. Técnicas genéticas que bloqueiam ou inativam a expressão de um gene específico (*knockout* de genes) são utilizadas na

validação de alvos terapêuticos. A inativação de um gene torna o organismo incapaz de produzir a proteína que este gene codificava e, desta forma, é possível investigar os efeitos associados a esta proteína. Estudos deste tipo no genoma de ratos demonstraram que apenas 10% dos genes *knockout* podem ser úteis para validação de alvos terapêuticos. A explicação para isto é a presença de funções redundantes e rotas de sinalização alternativas no organismo. Por esta razão, a abordagem mais adequada para compreender os efeitos biológicos é através de uma rede de interações. É neste cenário que surge o conceito de polifarmacologia que é a ligação de um composto a mais de um alvo macromolecular. Estes compostos são denominados fármacos multi-alvo (*multi-target drugs*) e seus efeitos farmacológicos são resultantes da combinação de interação com diferentes alvos (HOPKINS, 2008). Este é o novo paradigma no desenho de novos fármacos atualmente. Multi-alvo não significa inespecífico e sim alvos que são relevantes na obtenção do efeito terapêutico desejado. Normalmente, fármacos multi-alvo interagem com afinidade menor do que fármacos que interagem com um único alvo (*single-target drugs*). No entanto, é a combinação de efeitos, devido a interação com múltiplos alvos, que produz os efeitos farmacológicos superiores (CSERMELY, 2005).

Nestes casos, a seleção de um único alvo macromolecular para medir atividade e propor um modelo QSAR nos moldes clássicos provavelmente produzirá um modelo pouco efetivo para futuras previsões. O que esta sendo discutido aqui não é a superioridade da abordagem global em detrimento da local na proposição de modelos QSAR, mas sim que ambas as abordagens são válidas e que cada uma tem o seu propósito específico apesar de suas limitações. Feher e Ewing (2009), por exemplo, combinaram estas abordagens para efetuar previsões

de inibição dos citocromos P450 CYP3A4 e CYP2D6[10] que são enzimas envolvidas no metabolismo de xenobióticos bem como interações entre fármacos. Normalmente, modelos QSAR locais são mais apropriados para otimização de compostos (*lead optimization*) e modelos globais são mais adequados para prospecção de protótipos, ou seja, filtros preliminares para triagem.

1.6 Objetivos

O principal objetivo deste trabalho é disponibilizar modelos QSAR globais para triagem *in silico* de compostos, ou seja, avaliar o potencial de novos compostos como agentes leishmanicidas, em especial, para *Leishmania donovani*. Os modelos são relativos a duas classes de compostos: nucleosídeos e antifúngicos.

Com a série de nucleosídeos pretende-se elaborar um modelo de regressão logística para acessar a probabilidade de que um novo nucleosídeo seja ativo. Para este fim, pretende-se determinar quais são os descritores que apresentam maior poder discriminatório entre moléculas ativas e inativas. Os resultados obtidos serão comparados com um segundo método muito utilizado no suporte a decisões que é a árvore de classificação (*classification tree*).

Para a série de antifúngicos, pretende-se propor um modelo capaz de efetuar previsões para inibidores da síntese do ergosterol. Como os mecanismos de ação que levam à inibição da síntese do egosterol são distintos, pretende-se avaliar se é possível identificar características moleculares relevantes para atividade e que sejam comuns a compostos que atuam por esta via bioquímica.

Para atingir o objetivo principal foram estabelecidas as seguintes metas:

1. Efetuar análise conformacional por meio de uma abordagem apropriada para tratar sistemas com muitos graus de liberdade;

[10] Cinco citocromos são os principais responsáveis pelo metabolismo de fármacos CYP1A2, CYP2C9, CYP2D6 e CYP3A4. Este último está presente em grande abundância no fígado (> 30%) e possui a maior variedade de substratos. Uma breve introdução pode ser encontrada em http://www.aafp.org/afp/980101ap/cupp.html.

2. Capturar informações moleculares através do cálculo de descritores moleculares;
3. Selecionar variáveis para reduzir o conjunto inicial de descritores moleculares, para um subconjunto de descritores que sejam relevantes para o propósito do modelo, utilizando técnicas de redução de dados;
4. Estabelecer relações quantitativas entre a estrutura química e a atividade biológica observada para estes compostos;
5. Verificar a consistência entre os descritores eleitos e o conjunto de dados.
6. Analisar o domínio de aplicabilidade do modelo, ou seja, determinar a região do espaço descritor onde é possível fazer previsões;
7. Validar o modelo efetuando previsões para um conjunto de dados externo ao modelo, ou seja, que não tenha sido usado para gerá-lo.

Capítulo 2

Fundamentos Teóricos

Introdução

O objetivo final de estudos QSAR é a elaboração de um modelo que descreva a variação da atividade biológica em termos de propriedades moleculares relevantes e que apresente boa capacidade de previsão de atividade para novos compostos. Para construir um modelo QSAR é necessário: coletar dados biológicos para uma série de compostos; efetuar análise conformacional das estruturas químicas; gerar descritores ou parâmetros moleculares; selecionar os descritores relevantes; escolher o tipo de modelo de regressão que será utilizado na modelagem dos dados (linear, parabólico, bilinear, logística, etc.); analisar o domínio de aplicabilidade do modelo e validar do modelo.

Este capítulo trata dos fundamentos teóricos das etapas acima mencionadas e que foram seguidas durante o desenvolvimento dos modelos QSAR do presente trabalho.

2.1 Princípios de QSAR

Estudos Quantitativos das Relações entre a Estrutura Química e a Atividade Biológica, cuja sigla inglesa é QSAR, procuram quantificar a atividade biológica em termos das propriedades da estrutura molecular.

Historicamente, Crum-Brown e Fraser em 1868 foram pioneiros na área ao assumir a atividade biológica como uma função da estrutura química (CRUM-BROWN e FRASER, 1868). Entretanto, o método tornou-se popular em química medicinal a partir da década de 60 com os trabalhos de Hansch e Fujita e Free e Wilson (KUBINYI, 1993). Os respectivos métodos ficaram conhecidos como análise de Hansch e análise de Free-Wilson.

Na análise de Hansch, a atividade biológica de uma série de compostos é correlacionada com propriedades físico-químicas de substituintes ou parâmetros globais que representam efeitos hidrofóbicos e eletrônicos. Matematicamente, o modelo é descrito por:

$$\log\left(\frac{1}{C}\right) = -a\pi^2 + b\pi + \rho\sigma + c \tag{2.1}$$

onde C é a concentração molar que causa uma determinada resposta biológica, π é o valor relativo de lipofilicidade do substituinte, σ corresponde a constante eletrônica relativa ao substituinte e a, b, ρ e c são coeficientes de regressão. O valor de π é definido por:

$$\pi = \log P_X - \log P_H \tag{2.2}$$

O termo P_X é o coeficiente de partição, $\log P$, para o composto com um substituinte X e P_H é o coeficiente de partição para a molécula sem o substituinte, ou seja, $X = H$.

Na análise de Free-Wilson, adota-se como premissa que a introdução de um substituinte em qualquer posição da molécula sempre muda a potência relativa da mesma quantidade independente dos outros substituintes. Neste modelo são utilizadas variáveis indicadoras que assumem valores nulos ou unitários para representar, respectivamente, a ausência ou presença de um dado substituinte, *X*. A equação correspondente é dada por:

$$\log\left(\frac{1}{C}\right) = \sum_i a_i X_i + \mu \tag{2.3}$$

onde X_i é uma variável indicadora, a_i é coeficiente de regressão e μ é atividade geral média do composto de referência.

Há ainda o modelo misto que combina as variáveis da análise Hansch e de Free-Wilson bem como outras variações em torno destes modelos. Os modelos apresentados permitem a descrição da atividade biológica em termos de variáveis que fornecem interpretações claras. No entanto, o número de descritores disponíveis atualmente aumentou consideravelmente e, por vezes, são difíceis de interpretar, mas em compensação, podem contribuir com o poder de previsão do modelo. A equação correspondente que descreve a atividade biológica foi generalizada para:

$$\log\left(\frac{1}{C}\right) = \sum_i \beta_i V_i + \beta_0 \tag{2.4}$$

onde β_i eβ_0 são coeficientes de regressão e V_i são os descritores moleculares.

É importante salientar que dados biológicos confiáveis são necessários para a construção de modelos com boa capacidade de previsão. Isto significa que os dados devem ser obtidos de uma forma consistente através de um único protocolo, de preferência, pelo mesmo grupo de pesquisa (CRONIN e SCHULTZ, 2003).

Além da concentração molar, C, outros parâmetros podem ser utilizados em análises QSAR tais como: IC_{50}, ED_{50}, LD_{50}, K_i, atividades biológicas *in vitro*, parâmetros farmacocinéticos e farmacodinâmicos. Porém, percentagem de inibição não é uma forma apropriada de expressar a atividade biológica nestes estudos devido à natureza não linear das relações dose-resposta (KUBINYI, 1993).

Outro aspecto importante sobre estudos QSAR se refere aos requerimentos de representatividade e homogeneidade para aplicabilidade das análises (ERIKSSON et al., 2003). Para que um modelo QSAR seja aplicado, os compostos devem apresentar variações em torno de uma estrutura básica, ou seja, devem ser congenéricos e apresentar o mesmo modo ou mecanismo de ação (homogeneidade). Simultaneamente, estes compostos devem cobrir adequadamente o espaço químico (representatividade). Vighi (1998) questiona a severidade com o requerimento de congenericidade argumentando que é viável o tratamento de séries heterólogas, pois se todas as características estruturais de uma molécula podem ser descritas então todos os padrões de comportamento podem ser descritos por meio de um modelo adequado. O desafio é desenvolver sistemas de descritores capazes de expressar, de maneira quantitativa, todas as características estruturais das substâncias químicas.

Finalmente, as condições para avaliar a validade dos modelos QSAR foram reunidas na forma de princípios os quais foram denominados originalmente princípios de Setubal[1], cidade portuguesa onde ocorreu a conferência organizada pelo ICCA (*International Council of Chemical Associations*) e CEFIC (*European Chemical Industry Council*). De acordo com estes princípios, modelos QSAR deveriam (idealmente):

[1] Atualmente, estes princípios são conhecidos por Princípios OECD (*Organisation for Economic Co-operation and Development*)

1º Princípio: Estar associados com uma finalidade definida. Entende-se por finalidade qualquer propriedade físico-química, efeito biológico ou parâmetro ambiental relacionado à estrutura química que pode ser medido e modelado.
2º Princípio: Tomar a forma de um algoritmo compreensível;
3º Princípio: Ter domínio de aplicabilidade definido;
4º Princípio: Estar associados com medidas apropriadas para os índices de confiabilidade, robustez e previsibilidade;
5º Princípio: Estar associados, se possível, com uma interpretação mecanística.

2.2 Análise conformacional

As propriedades físicas, químicas e biológicas das moléculas dependem, em geral, de suas conformações. Explorar a superfície de energia potencial de uma molécula é uma tarefa extremamente árdua dependendo do número de graus de liberdade envolvido. Diversos métodos têm sido propostos a fim de identificar a conformação bioativa mais provável das moléculas. De acordo com os estudos conduzidos por BOSTRÖM (2001) cada metodologia apresenta certo nível de eficiência na reprodução da conformação do ligante no complexo protéico. Apesar da conformação que corresponde ao mínimo de energia global ser um arranjo atômico de grande interesse, as propriedades observadas em um sistema molecular correspondem a uma distribuição térmica de estados de baixa energia que representam diversas regiões do espaço conformacional.

Cabe ressaltar que a conformação biologicamente ativa não necessariamente corresponde a um mínimo de energia calculado para a molécula isolada. Perola e Charifson (2004) estudaram um conjunto de 150 complexos protéicos farmaceuticamente relevantes e confirmaram que ligantes flexíveis raramente interagem com a macromolécula na conformação de menor energia.

Os métodos utilizados na amostragem conformacional usualmente dividem-se em: sistemáticos, fragmentação molecular, Monte Carlo, *Simulated Annealing*, algoritmo genético, distância geométrica, dinâmica molecular (LEACH, 1996). Na grande maioria, o estudo conformacional é feito em quatro etapas:

1. Geração de uma estrutura primária;
2. Minimização de energia;
3. Eliminação de confôrmeros duplicados;
4. Reinício do processo.

Em geral, o que distingue as abordagens entre si são a primeira etapa e o sistema de coordenadas utilizado (internas, cartesianas ou matrizes de distâncias internucleares). Entretanto, a segunda etapa é a que mais consome tempo computacional. Tipicamente, 95% (ou mais) do tempo total gasto na busca conformacional é destinado à segunda etapa.

A remoção dos confôrmeros duplicados ocorre com base na diferença entre as coordenadas através do desvio quadrático médio, RMSD, da estrutura recém-gerada e as demais já armazenadas como únicas. O RMSD é expresso por:

$$\mathrm{RMSD}(v,w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|v_i - w_i\right\|^2} \qquad (2.5)$$

onde $\boldsymbol{v}$ e $\boldsymbol{w}$ representam n pontos equivalentes dois a dois nas estruturas em comparação. Não há um consenso a respeito do valor ideal para o valor de corte do desvio quadrático médio, porém aceitam-se valores acima de 0,4Å (aproximadamente 15° para ângulos diedros).

A análise conformacional é um processo que requer alta demanda computacional e a eficiência dos algoritmos é avaliada segundo a capacidade de encontrar todos os mínimos (ou o maior número deles) em um intervalo de tempo razoável. Este é um ponto crítico, pois não se conhece, a priori, todos os mínimos

da superfície de energia potencial e, portanto, a ampla amostragem é difícil de ser garantida exceto para a abordagem sistemática. Porém, a abordagem sistemática não é aplicável para sistemas com muitos graus de liberdade (n > 5) devido à explosão combinatória. O tamanho e o tipo do sistema fornecem, por outro lado, a extensão de aplicação do método. Os sistemas cíclicos, por exemplo, requerem cuidados especiais devido ao acoplamento dos diedros. É importante salientar que o desenvolvimento de *hardware* e de algoritmos de minimização de energia mais eficientes contribui enormemente com a viabilidade de técnicas computacionalmente custosas.

A seguir, cada uma das principais técnicas utilizadas no estudo conformacional serão apresentadas, segundo os princípios envolvidos e a evolução dos mesmos.

2.2.1 Abordagem sistemática

A abordagem sistemática classifica-se como determinística. Os métodos determinísticos exploram o espaço conformacional em intervalos regulares (Figura 2-1). Neste tipo de abordagem, o sistema de coordenadas internas é o mais apropriado na exploração do espaço conformacional, pois é possível reduzir o número de graus de liberdade identificados no sistema cartesiano aos ângulos torsionais independentes da molécula. Por exemplo, um sistema linear com N átomos possui 3N-5 graus de liberdade no sistema cartesiano (para sistemas cíclicos, este valor é de 3N-6) e, no sistema de coordenadas internas, este número se reduz a N´ ($N' < N$) ângulos torsionais independentes porque o comprimento das ligações entre átomos e os ângulos entre ligações consecutivas não variam drasticamente a conformação molecular (SAUNDERS et al., 1990 e GOODMAN e STILL, 1991).

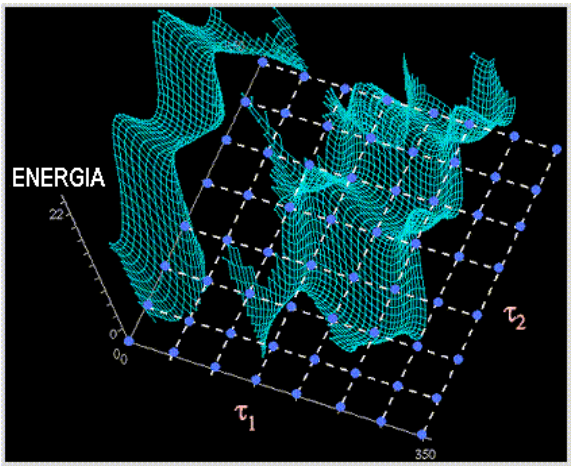


Figura 2-1: Mapeamento do espaço conformacional em espaços regulares[2].

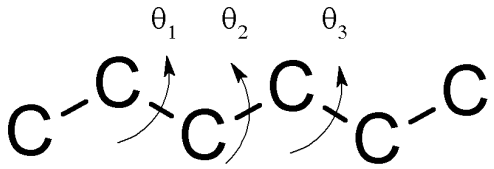


Figura 2-2: A combinação de ângulos torsionais gera novas conformações.

Diferentemente dos métodos aleatórios, os métodos determinísticos são uma garantia de que o espaço conformacional será inteiramente explorado. Cada ponto desta superfície corresponde a uma conformação que é formada a partir da combinação de ângulos torsionais (Figura 2-2). No entanto, um fator limitante inerente ao método é um fenômeno conhecido como explosão combinatória. Esta expressão é utilizada para enfatizar o crescimento vertiginoso do número de conformações geradas em função do aumento do número de ângulos torsionais. Matematicamente, este número é definido pela equação abaixo:

[2] Figura extraída e adaptada de:
http://cmm.info.nih.gov/modeling/guide_documents/conformation_document.html (Acessada em 25/06/2008).

$$N_{conformações} = \prod_{i=1}^{n} \frac{360^{\circ}}{\theta_i} \tag{2.6}$$

onde n é o número de ângulos torsionais independentes e θ_i é o intervalo angular que se pretende explorar na superfície de energia.

Outro ponto crítico a ser considerado é a resolução (determinada por θ_i) utilizada durante o mapeamento do espaço conformacional. Os dois extremos de resolução são inadequados, pois a demanda computacional pode se tornar intensa se a resolução for alta ou certos mínimos podem ser perdidos se a resolução for baixa. Portanto, para sistemas com muitos graus de liberdade (> 5) a busca sistemática torna-se impraticável. Contudo, esta metodologia ainda pode ser aplicada para sistemas flexíveis se algumas estratégias forem incluídas a fim de reduzir o número de conformações a serem otimizadas. Pode-se, por exemplo, eliminar combinações de ângulos que produzam impedimentos estereoquímicos ou ainda, para sistemas cíclicos, acrescentar restrições geométricas como distâncias entre átomos não ligados, distâncias entre átomos tomados como referência para marcar o início e o fim do anel. As duas últimas são importantes porque, devido à dificuldade de tratamento de anéis, as moléculas cíclicas são transformadas temporariamente em pseudo-acíclicas durante a execução da grande maioria das técnicas (Figura 2-3).

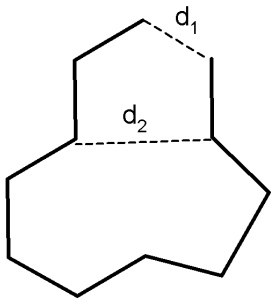


Figura 2-3: Restrições geométricas em anéis. A distância d_1 refere-se à referência para marcar o início e o fim do anel e a distância d_2 corresponde a restrição geométrica imposta a átomos não ligados.

O número de possíveis conflitos entre átomos, que devem ser avaliados, aumenta com a resolução, θ_i, e com o número de átomos na molécula, *N*, de acordo com a equação abaixo (BEUSEN et al., 1996):

$$V = \left(\frac{360^\circ}{\theta_i} \right)^T \frac{N(N-1)}{2} \tag{2.7}$$

onde *T* é o número de diedros independentes.

Saunders e colaboradores (1990) exploraram condições como as mencionadas acima e conseguiram reduzir o número de conformações iniciais do cicloheptadecano[3] ($C_{17}H_{34}$) de aproximadamente 10^8 para 10^4. Outras estratégias relacionam-se com implementações no algoritmo de modo a controlar a busca tornando-a mais eficiente.

O algoritmo mais utilizado neste tipo de busca exaustiva é o *depth-first*. No *depth-first* todas as possibilidades a um dado nível são analisadas e caso a condição imposta ao sistema não seja satisfeita retorna-se ao nível anterior como ilustrado na árvore da Figura 2-4. Cada ramo da árvore é analisado independentemente e determinados ramos podem ser eliminados se forem impostas condições estereoquímicas. Maiores detalhes a respeito de outros métodos de busca em árvores podem ser encontrados em:
http://www.rci.rutgers.edu/~cfs/305_html/Computation/comptoc.html.

[3] O cicloheptadecano possui 45 graus de liberdade ou, no espaço torsional, 11 ângulos independentes.

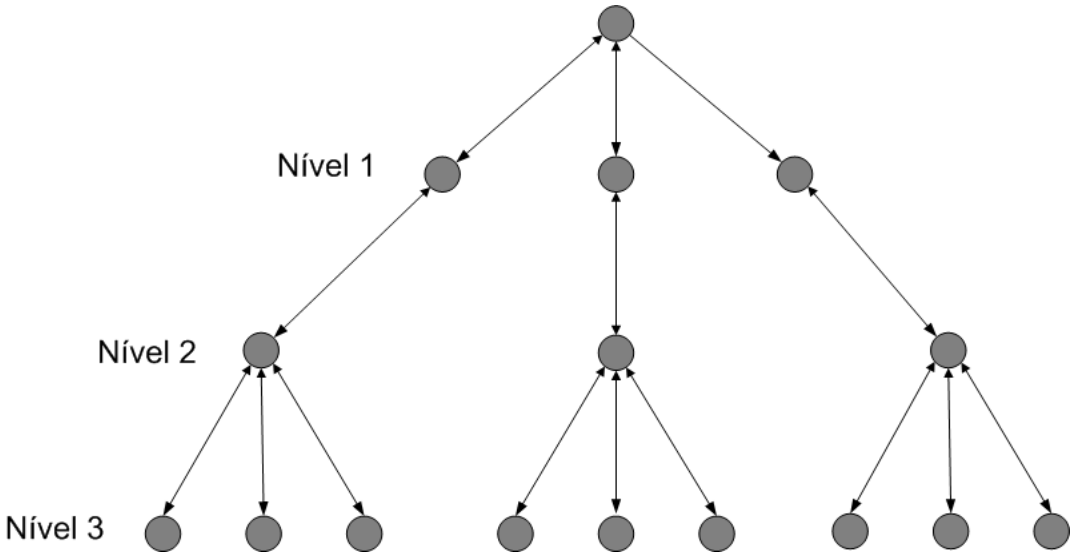


Figura 2-4: Busca sistemática usando o algoritmo *depth-first*. A seqüência de análise é dada pelas setas; os ramos são independentes entre si. As setas bidirecionais indicam a possibilidade de mudança de nível.

2.2.2 Abordagem aleatória (ou estocástica)

Explora o espaço conformacional gerando alterações aleatórias (ou pseudo-aleatórias) na geometria molecular. Na abordagem estocástica, as coordenadas cartesianas superam as coordenadas internas em termos de simplicidade e generalidade, pois o mesmo algoritmo pode ser utilizado para qualquer sistema, cíclico ou acíclico (SAUNDERS et al., 1990).

No sistema de coordenadas cartesianas, as estruturas são geradas por meio de translações aleatórias em cada átomo. No sistema de coordenadas internas, as transformações são aplicadas nos ângulos torsionais. Entretanto, aplicar translações aleatórias aos átomos aumenta a chance de gerar geometrias com altas energias dificultando o processo de otimização (LEACH, 1996, CHANG et al., 1989, GOTO e OSAWA, 1993).

Chang e colaboradores (1989) compararam diferentes maneiras de conduzir análises conformacionais através simulações de Monte Carlo para o n-octano, o ciclodecano e o ciclotetradecano. Os quesitos avaliados foram: o modo de gerar a

estrutura inicial, o número de coordenadas modificadas e o sistema de coordenadas utilizado. Os melhores resultados foram obtidos quando a estrutura inicial era selecionada de maneira uniforme dentro de um intervalo de energia até 3,5 kcal acima no mínimo global instantâneo e apenas uma fração dos ângulos diedros modificada aleatoriamente. Neste estudo, o sistema de coordenadas internas mostrou-se superior ao sistema de coordenadas cartesianas.

É importante salientar que o estudo foi conduzido aplicando-se deformações globais. Isto significa que os diedros são escolhidos e modificados aleatoriamente e, posteriormente, as restrições estereoquímicas são testadas. Este conjunto de estratégias é conhecido por MCMM (*Monte Carlo Multiple Minimum*).

Um estudo realizado por Weinberg e Wolfe (1994) apresentou resultados interessantes sobre deformações globais no método MCMM. O estudo avaliou o número de estruturas aceitáveis geradas por tais deformações em anéis. Para anéis cujo comprimento estava entre 10 e 100 átomos, a probabilidade de encontrá-los aleatoriamente variava de 5×10^{-3} a 2×10^{-4}. A partir destes resultados, Weinberg e Wolfe propuseram um novo algoritmo o qual se baseia na variação aleatória de diedros dentro de um intervalo de ângulos que favorecem a formação de anéis. Em uma dada estrutura, N/2 diedros são alterados aleatoriamente e os remanescentes são obtidos através de restrições que garantam a formação do anel. As respectivas probabilidades mencionadas acima passaram a variar de $3{,}7\times10^{-1}$ a $2{,}9\times10^{-1}$.

Nesta mesma direção, surgiram os trabalhos de Baysal e Meirovitch (1996 e 1997). Eles propuseram deformações locais em diedros não adjacentes escolhidos e modificados aleatoriamente. Uma conseqüência direta destas deformações é o rompimento temporário de ligações que são recuperadas após a otimização de geometria.

Goto e Osawa (1993) autores do programa CONFLEX criaram uma metodologia apropriada para anéis que difere da MCMM quanto ao modo de gerar

novas conformações e com relação aos critérios para selecionar a estrutura inicial. A candidata à estrutura inicial é escolhida entre as geometrias de menor energia dentro de uma janela de energia e não perturbada previamente. No passo seguinte, perturbações são aplicadas segundo as ilustrações da Figura 2-5 em conjunto ou separadamente. Entretanto, há certa tolerância com estruturas geradas cujas energias são ligeiramente acima do limite estabelecido, pois elas podem ser precursoras de geometrias aceitáveis como mostrado na Figura 2-6. Este algoritmo é conhecido por *reservoir-filling* em função do preenchimento gradual entre os limites estabelecidos. No caso do cicloheptadecano, CONFLEX encontrou duas conformações a mais (262) que o MCMM (260) devido a esta característica.

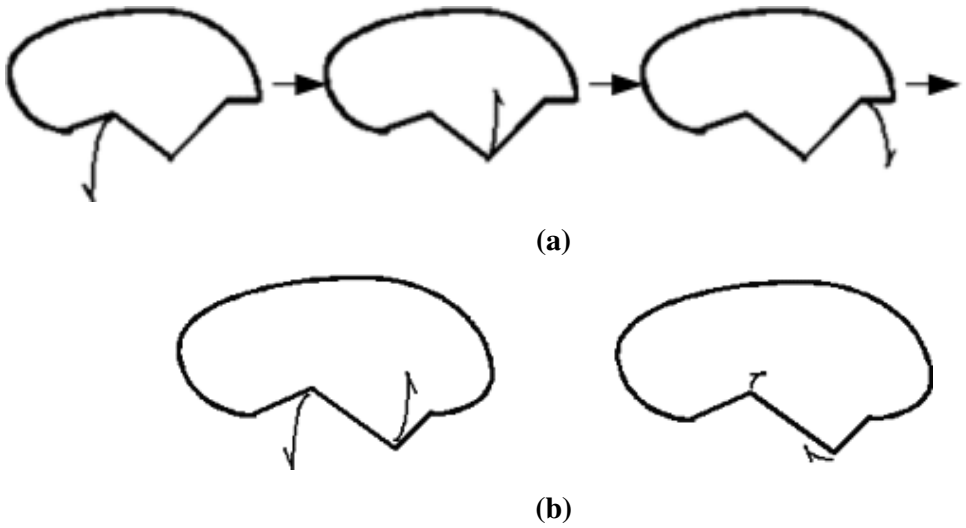


Figura 2-5: Modificações aplicadas a anéis usando as perturbações do tipo (a) *corner-flap* e (b) *edge-flip*[4].

[4] Figuras extraídas e adaptadas de:
http://www.conflex.us/prod_conflex.asp. Acesso em 8/07/2008.

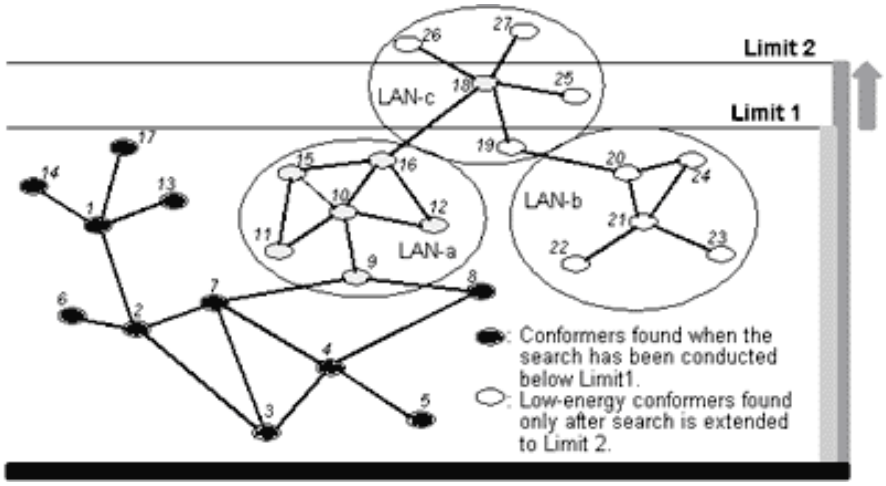


Figura 2-6: Ilustração esquemática do algoritmo *reservoir-filling*[4].

2.2.3 Abordagem por meio de Distância Geométrica

Embora esta técnica seja aleatória em essência, seu grande diferencial quanto à forma de gerar estruturas e a amostragem do espaço conformacional merecem uma seção a parte.

As conformações das moléculas são descritas por meio das distâncias entre todos os possíveis pares de átomos, *i* e *j*, dispostas em uma matriz $N \times N$. Cria-se uma matriz que contem distâncias mínimas e máximas aceitáveis entre pares de átomos. As conformações possíveis das moléculas estariam neste intervalo. O limite inferior é dado pela soma dos raios de van der Waals e é colocado abaixo da diagonal. O superior, acima da diagonal. É difícil estabelecer um limite superior para átomos separados por mais que três ligações. Assim, um número grande, algo como 999Å é colocado no lugar. Em seguida, a matriz é *suavizada* a fim de refinar o conjunto inicial de distâncias e torná-las geometricamente viáveis. Este procedimento é conhecido por suavização triangular (*triangle smoothing*). A suavização triangular utiliza restrições geométricas em grupos de três átomos conforme mostradas na equação abaixo:

$$S_{AC} \leq S_{AB} + S_{BC}$$
$$I_{AC} \geq I_{AB} - S_{BC} \tag{2.8}$$

A primeira restrição estabelece que a maior distância entre dois átomos A e C não pode ser maior que a soma dos valores máximos das distâncias entre os átomos AB e BC. A segunda restrição estabelece que a mínima distância entre dois átomos A e C não pode ser menor do que a diferença entre a menor distância entre os átomos AB e a maior distância entre BC (ver Figura 2-7).

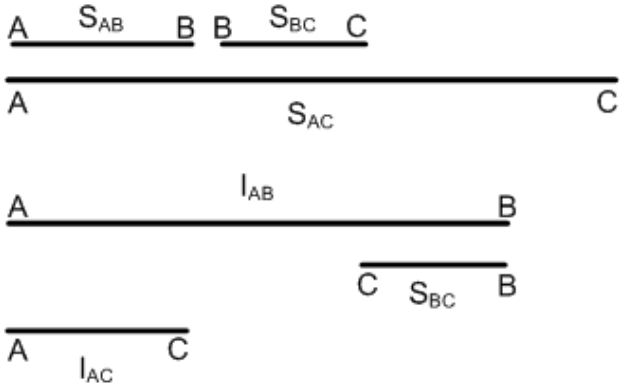


Figura 2-7: Desigualdades triangulares. S_{mn} corresponde à máxima distância entre os átomos i e j (limite superior). I_{mn} corresponde à mínima distância entre os átomos i e j (limite inferior)[5]

Em seguida, atribuem-se valores aleatórios aos elementos da matriz de distâncias dentro dos limites estabelecidos. Os valores, d_{ij}, contidos nesta matriz são usados para gerar a matriz métrica, G, definida por:

$$G_{ij} = \frac{d_{i0}^2 + d_{j0}^2 - d_{ij}^2}{2} \tag{2.9}$$

onde d_{i0} e d_{j0} são as respectivas distâncias dos átomos i e j a origem.

[5] Figura extraída e adaptada de LEACH, 1996.

A matriz G é uma matriz simétrica quadrada e, portanto, pode ser decomposta em:

$$G = VL^2V^T \tag{2.10}$$

onde *V* e *L* são, respectivamente, as matrizes dos autovalores e autovetores de G. A estrutura tridimensional representada por esta matriz em coordenadas cartesianas é obtida através da relação:

$$X = VL \tag{2.11}$$

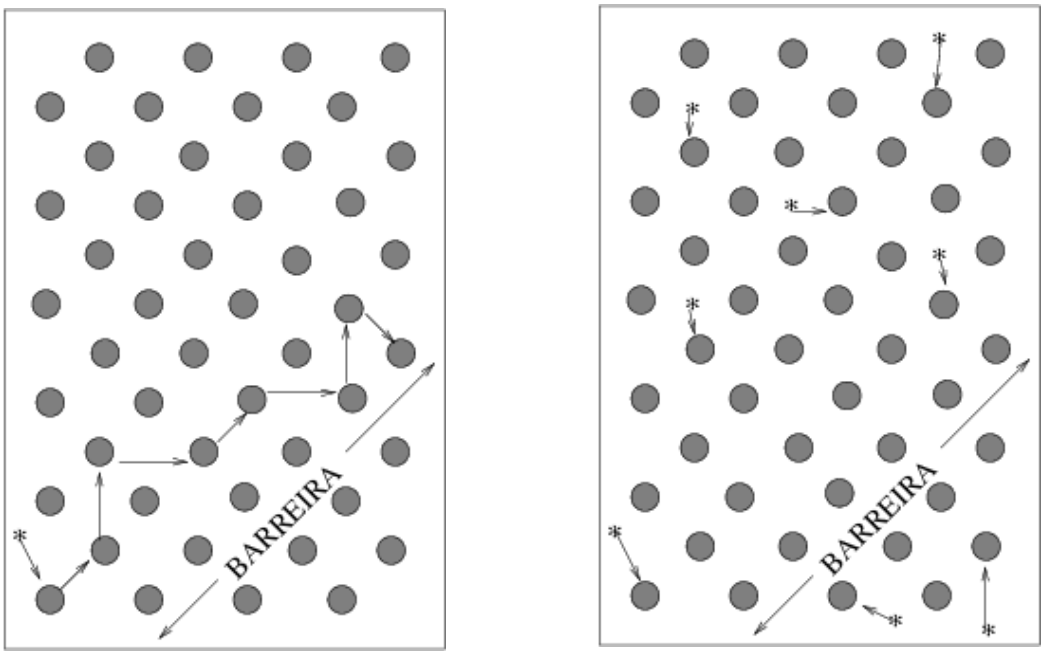
Em seguida, estas coordenadas são refinadas contra uma função semelhante à Eq. 2.12:

$$E = \sum_i \sum_{j>i} \begin{cases} \left(d_{ij}^2 - S_{ij}^2\right)^2 & d_{ij} > S_{ij} \\ 0 & I_{ij} \le d_{ij} \le S_{ij} \\ \left(I_{ij}^2 - d_{ij}^2\right)^2 & d_{ij} < I_{ij} \end{cases} \tag{2.12}$$

onde I_{ij} e S_{ij} são, respectivamente, os limites inferior e superior de distâncias entre os átomos *i* e *j*. Portanto, as estruturas geradas não dependem de estruturas prévias e estão distribuídas aleatoriamente no espaço conformacional. Esta característica permite a identificação de conformações separadas por uma larga barreira energética que, eventualmente, podem ser perdidas aplicando-se MCMM. A Figura 2-8 ilustra esta idéia.

Desde sua concepção, em 1970, o método vem sendo aperfeiçoado no sentido de aprimorar a conversão da matriz de distâncias em coordenadas cartesianas e de assegurar que a geometria gerada é aceitável para a minimização. No início da década de 90, Saunders e colaboradores (1990) relataram os resultados obtidos para

o cicloheptadecano com esta metodologia. Comparativamente a outras técnicas, esta produziu resultados insatisfatórios naquela data, encontrando apenas 176 mínimos dos 262 conhecidos. No entanto, as relações geométricas não foram estabelecidas naquela época através das desigualdades triangulares devido ao custo computacional. Alternativamente, as distâncias eram correlacionadas entre si. Um ano depois, Peishoff e Dixon (1991) conseguiram 223 mínimos para o cicloheptadecano relaxando este critério de correlação e efetuando a amostragem em termos de diedros. Posteriormente, com as novas implementações (SPELLMEYER et al., 1997) este número subiu para 242.



(a) **(b)**

Figura 2-8: Diferenças na amostragem via (a) MCMM e (b) Distância geométrica. [6]

[6] Figuras extraídas e adaptadas de WEINBERG, 1994.

A maior aplicação desta técnica ocorre nas determinações estruturais via Ressonância Magnética Nuclear (RMN). Diversos artigos atestam a capacidade deste método em resolver sistemas com muitos graus de liberdade (como proteínas) aliada as informações experimentais. Embora algumas modificações ao algoritmo acima exposto, tenham sido propostas por Spellmeyer e colaboradores (1997), eles concluíram que o método tradicional, é o mais adequado para tratamento de dados de RMN. Em apenas quatro horas foi possível resolver a estrutura da enzima inibidora da tripsina pancreática bovina que contem cerca de 900 átomos!

2.2.4 Algoritmos Genéticos

Os algoritmos genéticos foram concebidos na década de 60 por John Holland. Os princípios envolvidos são inspirados na teoria de Darwin sobre a evolução das espécies.

Uma população inicial com N componentes é escolhida aleatoriamente. Cada indivíduo é caracterizado por seus cromossomos e ordenado em ordem decrescente de aptidão. No processo de reprodução, novos indivíduos são gerados através de recombinação (*crossover*) e eventuais mutações (*mutation*). Em processos ditos elitistas, os melhores classificados são escolhidos para reprodução. A aptidão dos indivíduos é avaliada de acordo com uma função característica do problema. O processo continua até um número preestabelecido de gerações ou de cópias de um mesmo cromossomo.

Cada cromossomo corresponde a um conjunto de palavras binárias (genes) com qualquer comprimento (ver Figura 2-9). Entretanto, aumentando o comprimento reduz-se a capacidade de localizar valores que se diferenciam em uma posição da palavra (MCGARRAH e JUDSON, 1993). Os processos de recombinação e mutação são exemplificados na Figura 2-10.

Figura 2-9: Ilustração de um cromossomo. Cada cor representa um gene diferente que por sua vez está associado a um ângulo diedro.

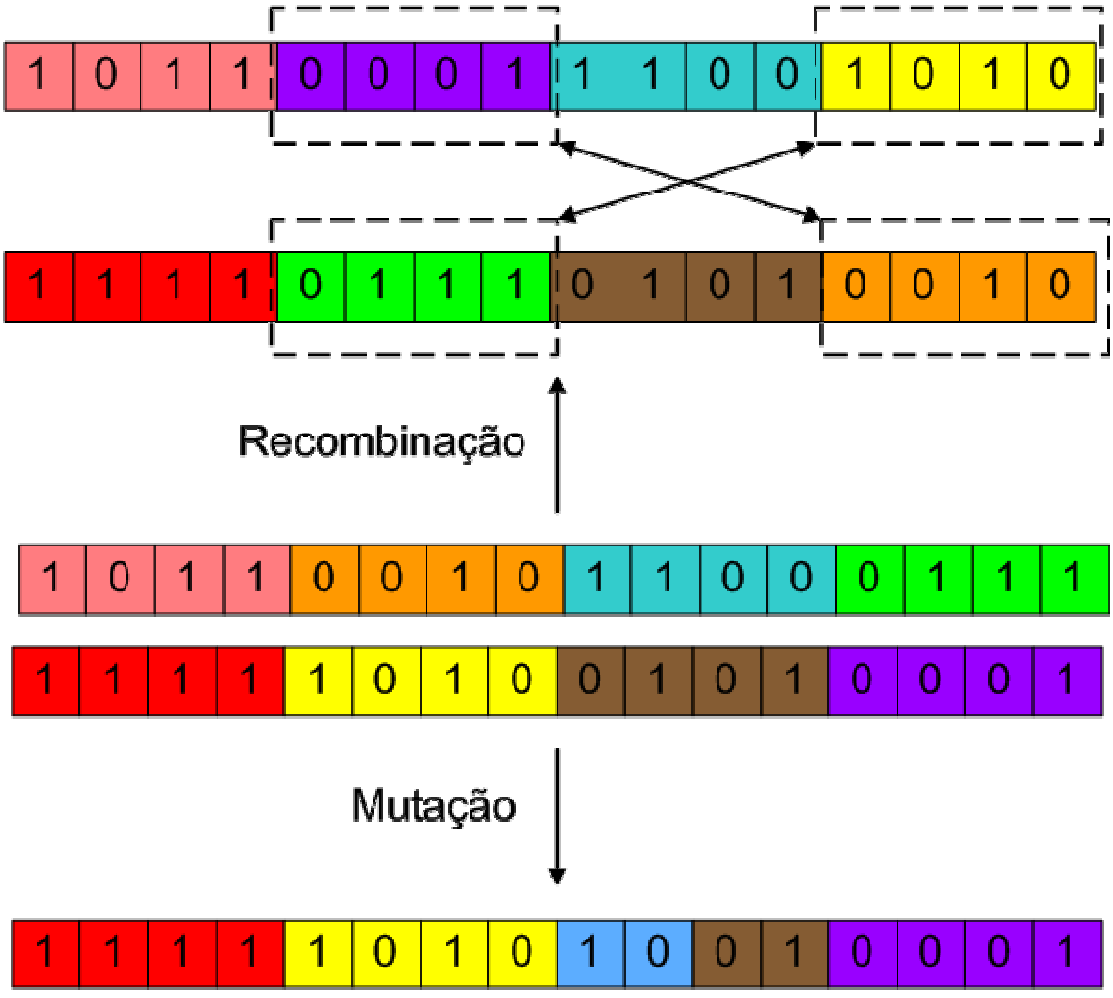


Figura 2-10: Ilustração dos processos de reprodução no algoritmo genético. Na recombinação, dois genes são trocados entre dois cromossomos. Na mutação, um gene sofre alterações.

No caso da busca conformacional em moléculas, cada gene relaciona-se com um diedro como mostrado na Figura 2-9. A função de energia potencial é usada para calcular a energia de uma dada conformação. O valor obtido é o critério usado na ordenação da população. As novas populações são formadas como descrito acima (vide Figura 2-10). Judson e colaboradores (1993) trataram 72 moléculas retiradas do banco de dados cristalográficos *Cambrigde Structure Database* com algoritmo genético. Os resultados mostraram que os mínimos obtidos diferiam pouco, em termos de energia, das estruturas cristalinas correspondentes comprovando, assim, a eficiência do método.

Uma introdução muito interessante e exemplos em algoritmos genéticos podem ser encontrados na página http://www.obitko.com/tutorials/genetic-algorithms/portuguese/.

2.2.5 Fragmentação molecular

Na técnica de fragmentação molecular (também conhecida por *Rule-based systems*), a molécula é decomposta em frações menores como mostrado na Figura 2-11. Cada unidade obtida é comparada com um banco de dados que contem as conformações preferenciais dos fragmentos. O banco de dados é formado por dados experimentais considerando a freqüência de ocorrência das diversas conformações. A molécula total é recomposta considerando-se a estereoquímica envolvida e depois submetida à otimização (KLEBE e MIETZNER, 1994). A grande desvantagem desta técnica é a dependência de um banco de dados que não necessariamente contempla todas as conformações de uma molécula qualquer.

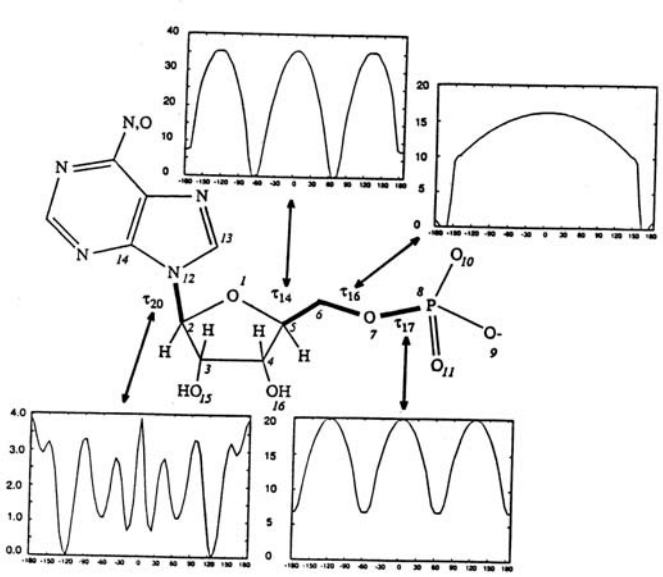


Figura 2-11: Exemplo de fragmentação molecular[7]. Os valores de cada ângulo torsional são provenientes de uma biblioteca que contém dados experimentais.

O programa OMEGA[8] que será utilizado no presente trabalho é um exemplo de sistema baseado em regras. A rotina computacional empregada para geração de conformações segue três etapas. Inicialmente, um pequeno conjunto de conformações é gerado pelo método de distância geométrica (ver seção 2.2.3). Os confôrmeros resultantes são encaminhados para a etapa seguinte. A próxima etapa consiste na divisão dos confôrmeros iniciais em fragmentos menores de até cinco ângulos consecutivos passíveis de rotação. Uma biblioteca, que contém ângulos torsionais e conformações de anéis permitidos, é usada para gerar as combinações de ângulos para cada fragmento. Na seqüência, estes fragmentos são reunidos para construir as conformações da molécula como um todo usando o algoritmo de busca *depth-first* (ver seção 2.2.1). Este procedimento é orientado pelas respectivas energias dos fragmentos para excluir ramificações da árvore de busca (ver figura

[7] Figura extraída de KLEBE, 1994.

[8] Os programas computacionais utilizados serão referenciados utilizando letras capitais.

Figura 2-4) com altos valores de energia. No último estágio, todas as conformações geradas são otimizadas utilizando mecânica molecular com o campo de força MMFF (*Merck Molecular Force Field*) que é um método confiável para estimar a diferença de energia entre conformações (YOUNG, 2001). Finalmente, as conformações duplicadas são removidas segundo critério de corte estabelecido para o RMSD.

2.2.6 Outros Métodos

Técnicas alternativas, como complementaridade molecular e dinâmica molecular, também figuram entre as propostas para resolver o problema de múltiplos mínimos.

A complementaridade molecular baseia-se em dois princípios: Primeiro, conformações de baixa energia possuem subestruturas de baixa energia; Segundo, conformações de baixa energia compartilham subestruturas similares. As diferentes conformações são construídas com dois operadores: combinação e espelhamento (WANG, 1997). A Figura 2-12 ilustra este procedimento.

Nas simulações de dinâmica molecular a molécula é descrita como uma estrutura dinâmica com coordenadas atômicas que se alteram com o tempo de acordo com as forças que agem nos átomos. A trajetória da molécula é determinada através das leis newtonianas (HOWARD e KOLLMAN, 1988). O comportamento físico da molécula é simulado em um banho térmico. Altas temperaturas são normalmente utilizadas para que as moléculas possam cruzar barreiras de potencial. Em certos intervalos de tempo, as conformações das moléculas são coletadas e submetidas à otimização de geometria.

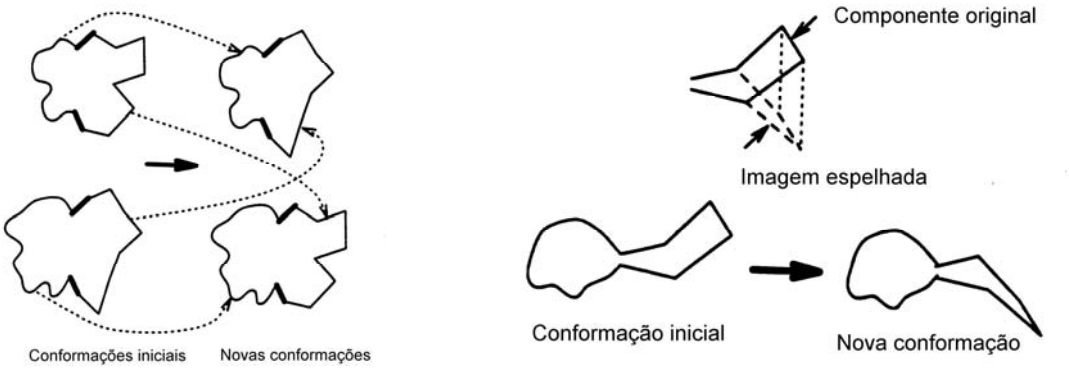


Figura 2-12: Análise conformacional utilizando complementaridade molecular. (a) Operador de combinação e (b) Operador de espelhamento[9].

2.2.7 Comparação entre métodos

Cada abordagem possui pontos fortes e fracos e, dependendo das necessidades envolvidas, as desvantagens podem ser contornadas. A análise conformacional com algoritmos sistemáticos é adequada para sistemas com poucos graus de liberdade (≤ 5), pois em sistemas maiores, a amostragem conformacional em uma resolução adequada ($\leq 60^{\circ}$) é impraticável. Ainda que certas restrições sejam impostas a fim de minimizar os custos computacionais, um conhecimento amplo das particularidades da molécula em estudo é necessário. Por outro lado, o rendimento (ou seja, o número de confôrmeros únicos gerados) é constante durante todo o processo. Além disso, como dito anteriormente, temos a garantia de que o espaço conformacional será mapeado por inteiro.

Nos métodos estocásticos, em princípio, não há restrição de tamanho para o sistema a ser tratado e um conjunto de soluções (talvez nem todas) pode ser

[9] Figuras extraídas e adaptadas de WANG, 1997.

encontrado em um intervalo de tempo relativamente menor. Entretanto, o rendimento não é constante e cai drasticamente no final do procedimento.

Nos últimos estágios, conformações previamente encontradas são redescobertas e não há uma fiança expressa de que todas as conformações possíveis foram exploradas. Normalmente, se análises independentes produzem o mesmo conjunto final de conformações, então se assume que houve convergência. Por ser um processo contínuo, é vantajoso porque não é necessário especificar, a priori, os parâmetros de busca e o processo termina quando os critérios de convergência são atingidos.

Dificilmente, os métodos disponíveis são hábeis para tratar sistemas com muitos graus de liberdade como proteínas. Grande parte das aplicações restringe-se ao caso do cicloheptadecano. Esta molécula tornou-se uma referência para avaliar a extensão e a efetividade de cada metodologia. A Tabela 2-1 mostra uma compilação da literatura dos resultados obtidos para a análise conformacional do cicloheptadecano utilizando as técnicas mais conhecidas.

Tabela 2-1: Número de mínimos encontrados por diferentes métodos para a molécula do cicloheptadecano. (*)Mínimo global.

Técnica	Tipo de Coordenadas	N° de mínimos encontrados	Idealizadores	Referência
Sistemática	internas	138	Houk/Wu	SAUNDERS et al.,1990
Sistemática	internas	211	Still/Lipton	SAUNDERS et al.,1990
Pseudosistemática	internas	257	Ngo/Karplus	Ngo-1997
MCMM	cartesianas	222	Saunders	SAUNDERS et al.,1990
MCMM(Random Walk)	internas	237	Still/Chang/Guida	SAUNDERS et al.,1990
MCMM(Usage directed)	internas	260	Still/Chang/Guida	SAUNDERS et al.,1990
MCMM(Early rejection)	internas	262	Weinberg/Wolfe	WEINBERG e WOLFE, 1994
MCMM(Flip/Flap)	internas	262	Goto/Osawa	GOTO e OSAWA, 1993
MCMM(Flex)	internas	232	Kolossvary/Guida	KOLOSSVARY e GUIDA, 1993
MCMM (LTD)	internas	247	Baysal/Meirovitch	BAYSAL e MEIROVITCH, 1997
Distânica Geométrica	internas	223	Peishoff/Dixon	PEISHOFF e DIXON, 1991
Distânica Geométrica	cartesianas	242	Spellmeyer *et al.*	SPELLMEYER et al., 1997
Complementaridade	-	256	Wang	WANG, 1997
Dinâmica Molecular	cartesianas	169	-	SAUNDERS et al.,1990
Simulated Annealing	internas	1 (*)	Guarnieri/Wilson	GUARNIERI e WILSON, 1992

2.2.8 Outros fatores que interferem na eficiência dos métodos

Como foi dito anteriormente, grande parte do tempo da análise conformacional é gasta com otimização de geometria. A concepção de algoritmos que reduzam este tempo pode, por exemplo, viabilizar técnicas marginalizadas. Saunders e colaboradores (1990) reduziram o tempo de otimização por um fator de três aplicando diferentes algoritmos de minimização de energia.

Guarnieri e Wilson (1992) sugeriram o uso de uma relação exata entre diedros para transformar anéis mantendo os comprimentos de ligação e ângulos de ligação intactos. Desta forma seria possível eliminar o passo de otimização.

Goto e Osawa (1993) propuseram uma comparação entre a estrutura em otimização com aquelas previamente armazenadas a cada determinado número de interações. Se as estruturas puderem ser superpostas o processo era interrompido economizando, assim, tempo computacional. Entretanto, este tipo de comparação não pode ser feito constantemente, pois o número de estruturas armazenadas aumenta com a evolução da análise conformacional reduzindo a eficácia desta estratégia.

O sistema de coordenadas utilizado na descrição da geometria molecular também interfere no tempo computacional de uma forma muito mais branda e não decisiva. Embora seja mais simples trabalhar com coordenadas cartesianas, é consenso geral, que as coordenadas internas são mais apropriadas. Isto porque a manutenção da estereoquímica das moléculas é mais eficiente.

2.3 Descritores moleculares

Os descritores moleculares são as variáveis independentes que serão utilizadas na formulação do modelo QSAR. Eles contêm a informação química das moléculas ou conforme a definição de Consonni e Todeschini (2000):

"Descritor molecular é o resultado final de um procedimento matemático e lógico o qual transforma informação química codificada dentro de uma representação simbólica de uma molécula em um número útil ou resultado de algum experimento padronizado".

Os descritores podem ser oriundos de dados experimentais ou calculados (*in silico*). Os descritores *in silico* dividem-se de acordo com a dimensionalidade da molécula usada para calculá-los, ou seja, unidimensional (1D), bidimensional (2D) e tridimensional (3D). Os descritores 1D e 2D são preferíveis em geral, pois independem da conformação ao contrário dos descritores 3D. Este tipo de dependência acarreta a necessidade de amostragens conformacionais e seleção da conformação que será usada na construção do modelo.

Os descritores utilizados no presente trabalho serão apresentados a seguir a fim de proporcionar um panorama geral das informações extraídas das moléculas em estudo. Os descritores foram divididos conforme a dimensionalidade envolvida nos cálculos.

Os descritores unidimensionais utilizados classificam-se em:

- Constitucionais;
- Grupos funcionais;
- Fragmentos átomo-centrado;
- Empíricos;
- Propriedades.

Os descritores constitucionais são baseados nos constituintes dos compostos, como por exemplo, número de átomos, número de ligações, número de anéis, peso molecular, etc.

Os descritores de grupos funcionais são baseados na contagem de grupos químicos funcionais presentes na molécula. Alguns exemplos são: número de ésteres alifáticos, número de aminas primárias aromáticas, número de tióis, etc.

Fragmentos átomo-centrado são descritores baseados na contagem de 120 fragmentos formados por átomos e suas possíveis conexões conforme definido por Ghose e Crippen (GHOSE et al., 1989)

Descritores empíricos são formados pelos descritores: índice de insaturação, Ui, fator hidrofílico, Hy, e razão aromática, ARR. Ui é fruto da contagem de ligações duplas e triplas. Hy é uma função da contagem de grupos hidrofílicos. ARR é uma razão entre o número de ligações aromáticas é o número total de ligação sem as ligações com hidrogênio.

Descritores de propriedades compreendem variáveis calculadas por meio de regressão linear. São eles: refratividade molar de Ghose-Crippen, MR, área da superfície polar baseada em fragmentos, PSA, e coeficiente de partição de Moriguchi, MlogP.

Os descritores bidimensionais utilizados classificam-se em:

- Descritores topológicos;
- Contagens de caminhos moleculares;
- BCUT;
- Índices topológicos de cargas de Gálvez;
- Autocorrelações 2D.

Descritores topológicos são obtidos a partir de grafos moleculares invariantes sem contar ligações de hidrogênio. Os grafos moleculares são representações das

moléculas que codificam a conectividade molecular onde os átomos são representados por pontos (vértices) e as ligações covalentes por linhas (bordas).

Invariante é uma propriedade matemática de uma estrutura que não depende da forma de numerar ou desenhar os vértices. Exemplos são: índice de Winner, W, índice de Balaban, J e índice de Randic, CID (CONSONNI e TODESCHINI, 2000).

Contagens de caminhos moleculares são descritores baseados na matriz de adjacência de grafos, **A**. A matriz de adjacência é uma matriz quadrada (n x n) cujos elementos são dados por:

$$a_{ij} = \begin{cases} 1, \text{ se o átomo i está conectado ao j} \\ 0, \qquad \text{caso contrário} \end{cases} \tag{2.13}$$

A k-ésima potência de **A**, $\mathbf{A}^k$, são números inteiros que fornecem o número de caminhos de comprimento de k ligações do átomo i ao j. A soma da linha i de $\mathbf{A}^k$ resulta em todos os caminhos de comprimento k iniciando a partir do átomo i (RÜCKER et al., 1993). Diferentes potências de k são exemplos de descritores desta classe.

Descritores BCUT são obtidos a partir dos autovalores positivos e negativos da matriz de adjacência onde os elementos da diagonal são ponderados pelo peso molecular dos átomos. Autovalores maiores e menores da matriz de Adjacência ponderadas pelas massas atômicas são exemplos de descritores.

Índices topológicos de cargas de Gálvez são descritores que caracterizam a distribuição de cargas das moléculas a partir da matriz de adjacência e o inverso quadrático da matriz de distâncias entre vértices (GÁLVEZ et al., 1994). Dentre os descritores podem-se destacar os índices topológicos de carga de diferentes ordens e índice de carga topológico global.

Descritores de autocorrelações 2D são calculados através de grafos moleculares através da soma dos produtos dos pesos dos átomos terminais de todas as trilhas de um dado comprimento (CONSONNI e TODESCHINI, 2000). Descritores de autocorrelação segundo Moreau-Broto são exemplos desta classe.

Os descritores tridimensionais utilizados classificam-se em:

- Índices de aromaticidade;
- Perfis moleculares de Randic;
- Descritores Geométricos;
- Descritores 3D-MoRSE;
- Descritores RDF;
- Descritores WHIM;
- Descritores GETAWAY;
- Descritores eletrônicos.

Índices de aromaticidade são índices que dependem da distância entre átomos envolvidos em ligações aromáticas.

Perfis moleculares de Randic são descritores derivados da matriz de distância geométrica, D, de uma estrutura. Da matriz D extrai-se a soma das linhas (ou colunas) onde cada elemento da matriz está elevado a k-ésima potência e o somatório é normalizado para k! resultando no fator ^{k}D. O valor final do descritor corresponde ao somatório de ^{k}D para diferentes potências k. Este descritor está correlacionado com a forma molecular (RANDIC, 1995).

O grupo dos descritores geométricos corresponde a diferentes tipos de descritores baseados na geometria molecular como, por exemplo, excentricidade molecular, MEcc.

Os descritores 3D-MoRSE baseiam-se na equação para difração de elétrons e são definidos por:

$$I(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} p_i p_j \frac{sen(sr_{ij})}{sr_{ij}} \tag{2.14}$$

onde p é a propriedade atômica para os átomos *i* e *j*, *N* é o número de átomos na molécula, r_{ij} é a distância entre os átomos *i* e *j*. O parâmetro *s* é dependente do ângulo de espalhamento θ e do comprimento de onda, λ, de forma que $s = 4\pi sen(\theta/2)/\lambda$.

Os descritores RDF são ligeiramente correlacionados com os descritores 3D-MoRSE, pois são definidos matematicamente por:

$$g(r) = f \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_i p_j \exp\left(-B(r - r_{ij})^2\right) \tag{2.15}$$

onde *f* é um fator de escala, *B* é um parâmetro análogo ao fator de temperatura na teoria de difração e os demais termos foram definidos acima para 3D-MoRSE.

Descritores WHIM são baseados na decomposição da matriz de covariância em eixos de componentes principais. A matriz de covariância é calculada a partir das coordenadas geométricas da molécula e, pode ou não, ser ponderada por diferentes fatores como massa atômica, polarizabilidade atômica, volume de Van der Waals, eletronegatividade atômica e estado eletrotopológico atômico (CONSONNI e TODESCHINI, 2000). O estado eletrotopológico atômico é um descritor que unifica atributos topológicos e eletrônicos envolvendo elétrons sigma e de valência (KIER et al.,1991).

Descritores GETWAY são derivados da matriz de influência molecular (em inglês, *molecular influence matrix*), H, que é definida matematicamente por:

$$\mathbf{H} = \mathbf{M} \bullet \left(\mathbf{M}^{\mathrm{T}} \bullet \mathbf{M}\right)^{-1} \bullet \mathbf{M}^{\mathrm{T}} \tag{2.16}$$

onde a matriz **M** é constituída por N linhas representando as coordenadas cartesianas x, y e z (em relação ao centro geométrico da molécula) de cada átomo na molécula exceto hidrogênios.

Os elementos da diagonal, h_{ii}, codificam informação atômica e a influência de cada átomo na determinação do formato global da molécula. Os elementos fora da diagonal, h_{ij}, representam o grau de acessibilidade do átomo j para interações com o átomo i (CONSONNI et al., 2002 a e b).

Os descritores eletrônicos foram calculados via mecânica quântica com o método semi-empírico PM3 considerando efeitos do solvente (LEACH, 1996; BASHFORD et al., 2000; PLIEGO JR., 2006). No modelo de solvatação utilizado, o solvente é tratado implicitamente onde a molécula (soluto) encontra-se em uma cavidade circundada por um meio (solvente) cuja constante dielétrica é dada por ε.

A energia livre de solvatação (ΔG_{sol}) é definida como a variação na energia livre para transferir uma molécula do vácuo para o solvente. Matematicamente, é descrita por três termos:

$$\Delta G_{sol} = \Delta G_{elec} + \Delta G_{vdw} + \Delta G_{cav} \tag{2.17}$$

O termo ΔG_{elec} é o componente eletrostático que pode ser incorporado ao cálculo quântico através do modelo campo de reação (em inglês, *reaction field*) considerando que o dipolo do soluto dentro da cavidade induz um dipolo nas redondezas do meio que por sua vez induz um campo elétrico dentro da cavidade. Desta forma, o campo de reação é considerado uma perturbação do hamiltoniano para uma molécula isolada:

$$H_{tot} = H_0 + H_{RF} \tag{2.18}$$

onde H_0 é o hamiltoniano da molécula isolada e H_{RF} é a perturbação dada por:

$$H_{RF} = \hat{\mu}^T \frac{2(\varepsilon - 1)}{(2\varepsilon + 1)a^3} \langle \Psi | \hat{\mu} | \Psi \rangle \tag{2.19}$$

onde $\hat{\mu}$ é o operador momento de dipolo, ε é a constante dielétrica do solvente e a é o raio da cavidade.

Após a determinação da função de onda total do sistema, ψ, pode-se calcular ΔG_{elec} através da equação:

$$\Delta G_{elec} = \langle \psi | H_{tot} | \Psi \rangle - \langle \Psi_0 | H_0 | \Psi_0 \rangle + \frac{1}{2} \frac{2(\varepsilon - 1)}{(2\varepsilon + 1)a^3} \mu^2 \tag{2.20}$$

onde ψ_0 é a função de onda na fase gasosa e o último termo da Eq. 2.20 é um fator de correção correspondendo ao trabalho feito para criar a distribuição de carga do soluto dentro da cavidade no meio dielétrico.

Os termos ΔG_{cav} e ΔG_{vdw} são dependentes da área acessível ao solvente do soluto, A, pois as moléculas do solvente que compõem a primeira camada de solvatação são as mais afetadas em ambos os termos. O número de moléculas de solvente na primeira camada de solvatação é aproximadamente proporcional à área acessível ao solvente do soluto. Estes termos são relacionados conforme a equação definida abaixo:

$$\Delta G_{cav} + \Delta G_{vdw} = \gamma A + b \tag{2.21}$$

onde γ e b são determinados experimentalmente a partir da energia livre para a transferência de alcanos do vácuo para água.

Efeitos de solvatação são importantes para a energia relativa de isômeros e confôrmeros e para a estrutura eletrônica. Como os fenômenos biológicos ocorrem em meio aquoso, uma estratégia adotada no presente trabalho foi a reordenação dos

confôrmeros provenientes da análise conformacional em termos dos valores de energia em solução mantendo a geometria rígida. A manutenção da geometria rígida reside na confiabilidade das geometrias geradas durante a análise conformacional. Para este fim, o modelo de solvente SM5.42R é o mais indicado.

O modelo SM5 é um modelo de solvatação universal e o ponto chave desta aproximação é a dependência das tensões de superfície com as distâncias interatômicas. Modelos de solvatação universais são válidos para todos os solventes, pois as propriedades do solvente são representadas na constante dielétrica. O modelo SM5.42R é baseado em cargas pontuais do tipo classe IV (representado pelo .4 na nomenclatura) do modelo de cargas CM2 (representado pelo 2 na nomenclatura) e foi concebido para prever energia livre de solvatação usando as geometrias otimizadas em fase gasosa, ou seja, a otimização de geometria não é feita na fase líquida onde a molécula é mantida rígida (representado pelo R na nomenclatura) (TRUHLAR et al., 1999). As cargas classe IV fornecem uma descrição mais realista da distribuição de cargas molecular visto que as cargas são parametrizadas para reproduzirem observáveis experimentais como momentos de dipolo. O modelo de cargas CM2 é um tipo de parametrização para atingir valores mais acurados para as cargas classe IV (TRUHLAR et al., 1998).

Existem diversos tipos de descritores eletrônicos. Entretanto, no presente trabalho foi dado a preferência aos descritores que codificassem características da molécula como um todo ao invés de descritores locais. Nesta categoria se enquadram: o calor de formação, ΔH_f, a energia eletrônica, *EE*, a energia de repulsão cerne-cerne, *CCR*, a energia do HOMO, E_{homo}, a energia do LUMO, E_{lumo}, a diferença de energia entre o LUMO e o HOMO, *Gap(Lumo,Homo)*, a diferença de energia entre o HOMO e o HOMO-1, *Gap(Homo, Homo-1)*, o momento dipolar total, *Dip*, e a energia de solvatação (ΔG_{water}).

O calor de formação, ΔH_f, é utilizado para mensurar diferenças de energia, pois é proporcional a energia interna, ΔU, visto que o volume molecular e a pressão não variam consideravelmente com as alterações conformacionais de acordo com a relação abaixo (GAUDIO, 1992):

$$\begin{aligned} \Delta H_f &= \Delta U + \Delta(pV) \\ \Delta H_f &= \Delta U + \underbrace{\Delta pV}_{\Delta p \approx 0} + \underbrace{p\Delta V}_{\Delta v \approx 0} \\ \Delta H_f &\approx \Delta U \end{aligned} \tag{2.22}$$

A energia de repulsão cerne-cerne, *CCR*, expressa a energia de repulsão eletrostática entre cargas de cerne definidas como a carga nuclear pontual subtraída da carga dos elétrons da camada interna. A soma entre a energia eletrônica e a energia de repulsão cerne-cerne fornece a energia total do sistema.

A energia do HOMO, E_{homo}, corresponde à energia do orbital de mais alta energia ocupado e é proporcional ao primeiro potencial de ionização de acordo com o teorema de Koopmans. Caracteriza a suscetibilidade da molécula frente a um ataque eletrofílico.

A energia do LUMO, E_{lumo}, corresponde à energia do orbital de menor energia desocupado e é proporcional a afinidade eletrônica da molécula. Caracteriza a suscetibilidade da molécula frente a um ataque nucleofílico.

A diferença de energia entre o LUMO e o HOMO, *Gap(Lumo,Homo)*, fornece um índice de estabilidade da molécula, pois uma grande diferença entre o LUMO e o HOMO implica em menor reatividade nas reações químicas. Também se correlaciona com a polarizabilidade desde que quanto menor a diferença de energia entre HOMO e o LUMO maior a facilidade de polarização da molécula.

O *Gap(Homo,Homo-1)* tem se mostrado importante em diferentes tipos de atividade biológica, no entanto, o significado físico-químico deste descritor não é claramente compreendido (COLUCI et al., 2002).

A polaridade de uma molécula é um descritor molecular associado à distribuição de carga na molécula. O momento dipolar total reflete a polaridade total da molécula.

Um sumário dos descritores utilizados no presente trabalho podem ser encontrados na Tabela 2-2.

Tabela 2-2: Sumários dos descritores utilizados na extração de informações das moléculas em estudo.

Descritores 1D	
Tipo de descritor	Nº de descritores calculados
Constitucionais	47
Grupos funcionais	121
Fragmentos átomo-centrado	120
Empíricos	3
Propriedades	3
Descritores 2D	
Tipo de descritor	Nº de descritores calculados
Topológicos	266
Contagens de caminhos moleculares	21
BCUT	64
Índices topológicos de cargas de Gálvez	21
Descritores de autocorrelações 2D	96
Descritores 3D	
Tipo de descritor	Nº de descritores calculados
Índices de aromaticidade	4
Perfis moleculares de Randic	41
Geométricos	70
3D-MoRSE	160
RDF	150
WHIM	99
GETWAY	197
Eletrônicos	9
Total de descritores	1492

2.4 Seleção de variáveis

A seleção de variáveis é um conjunto de técnicas que têm por objetivo eliminar atributos (ou variáveis) considerados redundantes ou irrelevantes durante a modelagem matemática dos dados. Devido ao elevado número de variáveis, o processo é normalmente conduzido de forma automatizada e os critérios adotados para julgar a relevância dos atributos podem potencialmente eliminar variáveis importantes. Por esta razão é recomendável confrontar o modelo resultante com a noção intuitiva sobre o mecanismo de ação ou modo de ação.

Kubinyi (1993) e, posteriormente, Guyon e Elisseeff (2003) demonstraram por meio de exemplos que certas variáveis individualmente podem se mostrar irrelevantes, porém, quando combinadas, podem produzir modelos com boa capacidade de previsão. Esta constatação amplia as possibilidades a serem exploradas gerando um enorme desafio a concepção de metodologias eficientes de busca.

As metodologias desenvolvidas para seleção de variáveis procuram solucionar esta questão e, essencialmente, se dividem em filtros, envoltórios (*wrappers*) e embutidos (*embedded*). Há ainda uma classificação destes métodos em supervisionados e não-supervisionados. Métodos supervisionados selecionam variáveis baseados em algum critério pré-definido como a classe do objeto, valores da variável dependente, etc. Já os métodos não supervisionados trabalham apenas com o conjunto de atributos.

Métodos baseados em filtros selecionam um subconjunto de variáveis antes da modelagem propriamente dita, ou seja, funcionam como um pré-processamento dos dados. O peso de Fisher é um exemplo de filtro que se enquadra como método supervisionado, pois utiliza a informação de classe para avaliar o poder discriminatório de uma dada variável. Matematicamente, este filtro é defino por:

$$F_W = \frac{\left(\bar{x}_i^{(c_1)} - \bar{x}_i^{(c_2)}\right)^2}{\left(\sigma_i^{(c_1)}\right)^2 + \left(\sigma_i^{(c_2)}\right)^2} \tag{2.23}$$

onde C_1 e C_2 são as categorias, i é a variável em análise, $\sigma_i^{(c_j)}$ é o desvio padrão da variável i na categoria j (j = 1,2).

O coeficiente de correlação de Pearson pode ser utilizado tanto como um método supervisionado quanto não supervisionado. Na primeira opção, as variáveis são selecionadas de acordo com a correlação com a variável de resposta. Na segunda opção, variáveis correlacionadas são eliminadas segundo um critério de corte sem levar em consideração qualquer informação da variável de resposta. Matematicamente, este coeficiente é definido por:

$$C_{Pearson} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i' - \bar{x}')}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(x_i' - \bar{x}')^2\right]}} \tag{2.24}$$

onde $\bar{x}$ e $\bar{x}'$ correspondem à média aritmética das variáveis x e x' e n é o número de amostras.

Whitley e colaboradores (WHITLEY et al., 2000) desenvolveram um método não supervisionado conhecido como UFS (*Unsupervised Foward Selection*). Este método faz uso de diferentes tipos de filtros selecionando um subconjunto de descritores conforme descrito a seguir. Inicialmente, os descritores com variância próxima a zero são eliminados. Em seguida calcula-se a correlação entre os descritores remanescentes. Os dois descritores menos correlacionados são selecionados. Na seqüência, o próximo descritor a ser incluído deve apresentar máxima ortogonalidade até um limite de corte estabelecido em relação à base

ortonormal definida pelos dois vetores inicialmente selecionados. Este processo é repetido para cada descritor do conjunto original.

Métodos baseados em envoltórios utilizam o algoritmo de aprendizado de interesse como uma função de mérito para avaliar cada subconjunto proposto de variáveis até que um determinado critério seja satisfeito. Os resultados obtidos com estes métodos estão fortemente ligados ao algoritmo de aprendizado. Como exemplo de algoritmos de aprendizado pode-se citar K-NN (*K-Nearest Neighbors*), redes neurais, árvores de decisão, algoritmos de agrupamento (*clustering*), regressão logística, Naive Bayes, máquina de vetor suporte (SVM), etc.

As estratégias mais populares utilizadas na proposição de subconjuntos de variáveis são: a busca sistemática, o algoritmo genético, a seleção prospectiva (forward stepwise selection) e a eliminação retrospectiva (backward elimination).

A busca sistemática e o algoritmo genético seguem os princípios apresentados nas seções 2.2.1 e 2.2.4. A seleção prospectiva incorpora progressivamente as variáveis iniciando o processo a partir de um conjunto vazio enquanto que a eliminação progressiva trabalha no sentido inverso, ou seja, a partir do conjunto total de variáveis eliminam-se progressivamente as menos promissoras.

Finalmente, métodos embutidos agregam o processo de seleção como parte do algoritmo de mineração. Em outras palavras, a seleção de variáveis é guiada pelo processo de aprendizagem. Exemplos de algoritmos são ID3, C4.5 e CART. Estes três algoritmos são utilizados em árvores de classificação e regressão e são baseados no ganho de informação de cada variável (ver seção 2.6).

Discussões mais aprofundadas sobre os métodos apresentados nesta seção podem ser encontradas nos trabalhos de Guyon e Elisseeff (2003) e Gaudio e colaboradores (2002).

2.5 Análise de Regressão

A análise de regressão compreende técnicas estatísticas utilizadas para encontrar relações entre variáveis descrevendo-as por meio de uma equação matemática com o propósito de efetuar previsões futuras.

O presente trabalho faz uso de modelos de regressão linear e não-linear que serão apresentados nesta seção seguindo os princípios gerais na proposição, validação e análise de modelos QSAR discutidos por Gaudio e Zandonade (2001).

2.5.1 Regressão Linear Simples

A regressão linear simples é uma ferramenta estatística utilizada para estabelecer uma relação entre uma única variável explanatória, *x*, e uma única variável de resposta, *y*. Esta relação é expressa através de uma equação linear que irá fornecer um valor estimado, $\hat{y}$, para a variável de resposta:

$$\hat{y} = b_0 + b_1 x_1 + Err \tag{2.25}$$

onde b_0 é coeficiente linear, b_1 é o coeficiente angular e *Err* é o resíduo. O objetivo da regressão é encontrar os coeficientes b_0 e b_1 que minimizam a diferença entre *y* e $\hat{y}$.

O método utilizado para encontrar os coeficientes b_j (j = 0,1) é denominado método dos mínimos quadrados. Neste método a soma quadrática da diferença entre *y* e $\hat{y}$, ou seja, a soma quadrática dos erros para todas as *n* amostras do conjunto de dados deve ser mínima. O uso da soma quadrática ao invés da soma simples reside na premissa assumida de que os erros seguem uma distribuição aleatória com média zero acarretando, conseqüentemente, um valor nulo para a soma simples desde que diferenças negativas e positivas se cancelariam mutuamente.

Em notação simbólica tem-se:

$$SSE = \sum_{k=1}^{n} \left(y_k - \hat{y}_k \right)^2 \quad \text{ou} \quad SSE = \sum_{k=1}^{n} \left(y_k - b_0 - b_1 x_{1k} \right)^2 \tag{2.26}$$

onde SSE é a sigla inglesa para Soma dos Erros Quadráticos (*Sum of Squared Errors*).

O mínimo desta função é dado pela primeira derivada em relação aos coeficientes b_j (j = 0, 1):

$$\frac{\partial SSE}{\partial b_o} = 0 \text{ e } \frac{\partial \text{SSE}}{\partial b_1} = 0 \tag{2.27}$$

Os coeficientes que satisfazem as condições descritas na Eq. 2.16 são definidos como:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \text{ e } b_0 = \bar{y} - b_1 \bar{x} \tag{2.28}$$

A avaliação de qualidade do modelo obtido é realizada através de diagnósticos definidos pela análise de variância, estatística F e análise dos resíduos.

A avaliação do ajuste do modelo é feita por meio de índices que agregam conceitos diferentes na mensuração de qualidade do modelo. Os índices comumente utilizados são: R (coeficiente de correlação), R^2 (coeficiente de determinação), S_e (desvio padrão ou erro padrão da estimativa), teste F e teste t.

O coeficiente de correlação mede a intensidade e a direção da relação linear entre duas variáveis, mas não, necessariamente, implica relação de causa-efeito. Matematicamente, é definido por:

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.29}$$

De acordo com a Eq. 2.29, R assume valores no intervalo [-1,1] e quanto mais próximos dos extremos estiverem estes valores, maior será a correlação entre as variáveis independente e dependente.

Os demais índices são derivados da decomposição da variância total em variância explicada e não explicada pelo modelo (ver Figura 2-13):

$$\overbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}^{\text{Variância Total}} = \overbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}^{\text{Variância explicada}} + \overbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}^{\text{Variância não explicada}} \tag{2.30}$$

$$\downarrow \qquad \downarrow \qquad \downarrow$$

$$\text{SST} \quad = \quad \text{SSR} \quad + \quad \text{SSE}$$

O coeficiente de determinação R^2 mede a proporção da variância explicada pelo modelo e é definido por:

$$R^2 = \frac{SSR}{SST} \tag{2.31}$$

Da Eq. 2.31 observa-se que R^2 pertence ao intervalo [0,1] e quanto mais próximo do valor unitário, melhor é o ajuste da reta de regressão aos dados. É necessário cautela na interpretação de R^2, pois este índice indica o quanto à linha de regressão ($\hat{y}$) se aproxima dos dados reais, porém, não expressa:

- Se há relação de causalidade entre x e y;
- Se existe influência sobre o modelo de possíveis variáveis omitidas;

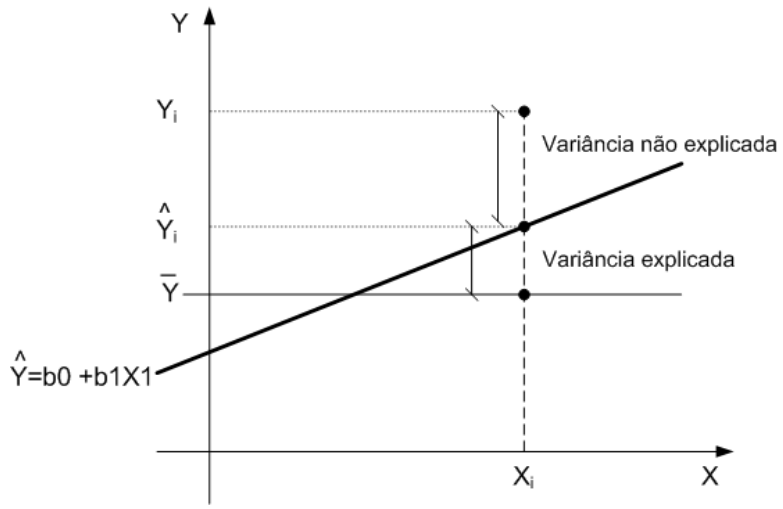


Figura 2-13: Explicação gráfica para a decomposição da variância total.

- Se a regressão correta foi utilizada;
- Se o conjunto de variáveis independentes utilizado é o mais apropriado;
- Se há colinearidade no conjunto de dados;
- Se o modelo pode ser melhorado pelo uso de variáveis independentes que sejam transformações das variáveis originais.

O desvio padrão dos dados ao redor da reta de regressão é denominado erro padrão da estimativa, S_e. Este índice mede a variância não explicada pela regressão e é definido por:

$$S_e = \sqrt{\frac{\left(y_i - \hat{y}_i\right)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}} \tag{2.32}$$

Quanto menor o valor de S_e melhor será o ajuste do modelo de regressão.

Os testes de significância para o modelo de regressão e os coeficientes individuais da regressão demonstram que os valores encontrados para os índices de confiabilidade não ocorreram ao acaso. Significância estatística é obtida através da rejeição da hipótese nula, H_0.

O teste *F* é utilizado para avaliar a significância estatística de R^2 e é definido matematicamente por:

$$F_0 = \frac{SSR}{SSE/n-2} \tag{2.33}$$

O teste *F* mede a razão entre a variância explicada e a não explicada. Se, para um dado nível de significância α, $F_0 > F_{\text{crítico}} = F_{\alpha,1,n-2}$ então se rejeita a hipótese nula de que o modelo de regressão não é significante. O teste *t* avalia a hipótese nula de que não há relação linear entre a variável independente e a dependente. Esta afirmação é equivalente à premissa de que os coeficientes b_j (j = 0, 1) são nulos. O teste *t* é definido por:

$$t_0 = \frac{b_j}{\sqrt{\left(\frac{SSE}{n-2}\right)\sum\left(x_i - \bar{x}\right)^2}} \qquad (\text{i} = 0,1) \tag{2.34}$$

Similarmente ao teste *F*, se $t_0 > t_{\alpha,n-2}$ existe relação linear entre a variável dependente e independente.

A análise de variância, cuja sigla inglesa é ANOVA, é freqüentemente apresentada na forma de uma tabela como mostrado na Tabela 2-3.

Tabela 2-3: Tabela ANOVA para o modelo de regressão linear.

Fonte de Variação	**N° Graus de liberdade**	**Soma dos Quadrados**	**Média da Soma dos Quadrados**	***F₀***	**Probabilidade P**
Modelo	1	SSR	MSR=SSR	MSR/MSE	$=P(H_0:F_0 \leq F_{\text{crítico}})$
Resíduo	n-2	SSE	MSE= SSE/(n-2)		
Total	n-1	SST	SST/(n-1)		

A melhor forma de avaliar a capacidade de previsão de um modelo é através de validação externa, ou seja, fazer previsões para um grupo de dados que não foram utilizados na construção do modelo. No entanto, nem sempre é possível dispor de um conjunto externo de validação de forma que o método da validação cruzada torna-se uma alternativa para a análise de previsibilidade do modelo.

Neste método, certo grupo de dados é separado do conjunto de dados inicial e em seguida reconstrói-se o modelo com os dados remanescentes. Com o modelo obtido fazem-se previsões acerca dos dados separados e mensura-se o desvio entre os valores observados, Y_i, e os estimados, $\hat{Y}_i$. Este processo é repetido variando-se o grupo de dados separados. Normalmente, retira-se um objeto do conjunto de dados ao que se chama validação *leave-one-out*.

Os indicadores utilizados no método são:

$$PRESS = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \tag{2.35}$$

$$Q^2 = 1 - \frac{PRESS}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2} \tag{2.36}$$

$$SPRESS = \frac{\sqrt{PRESS}}{n-2} \tag{2.37}$$

Um modelo com boa capacidade de previsão apresenta valores para o coeficiente de correlação da validação cruzada, Q^2, próximos de 1 e valores para o desvio padrão da validação, *SPRESS*, que tendem a zero.

Finalmente, modelos de regressão linear podem ser utilizados desde que quatro premissas principais sejam respeitadas:

1. Existência de linearidade entre as variáveis dependente e independente;
2. Os resíduos são independentes;

3. Os resíduos apresentam variância constante (homocedasticidade);
4. Os resíduos seguem uma distribuição normal $N(0,\sigma^2)$.

2.5.2 Regressão linear Múltipla

A regressão linear múltipla, RLM, é uma extensão da regressão linear simples para mais de uma dimensão. Neste caso, a melhor forma de expressar a relação entre as variáveis dependentes e independentes é através de matrizes:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1i} \\ 1 & x_{21} & x_{22} & \cdots & x_{2i} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{ni} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_i \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \tag{2.38}$$

A Eq. 2.38 pode ser reescrita como:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \tag{2.39}$$

Aplicar o método dos mínimos quadrados na Eq. 2.39 para obter a estimativa dos coeficientes $\mathbf{b}$ significa resolver a equação matricial abaixo:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2.40}$$

onde $\mathbf{X}'$ corresponde a matriz transposta de $\mathbf{X}$, $\mathbf{X}^{-1}$ é a matriz inversa de $\mathbf{X}$ e a matriz resultante do produto $\mathbf{X}'\mathbf{X}$ é denominada matriz de correlação.

Idealmente, as variáveis independentes $\mathbf{X}$ em um modelo de regressão devem ser ortogonais, ou seja, não correlacionadas. A razão para esta condição pode ser compreendida a partir da Eq. 2.40. Se o determinante de uma matriz for nulo, então esta matriz é singular e não possui inversa. Logo, se houver um alto grau de colinearidade entre as variáveis $\mathbf{X}$ não é possível calcular a matriz inversa da matriz de correlação, pois seu determinante é nulo.

O valor estimado para a variável dependente é calculado pela expressão:

$$\hat{\mathbf{y}} = \mathbf{X}\underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}_{\mathbf{b}} \tag{2.41}$$

A avaliação do ajuste do modelo é feita com os mesmos índices de confiabilidade apresentados para a regressão linear. A análise de variância correspondente é mostrada na Tabela 2-4.

Tabela 2-4: Tabela ANOVA para o modelo de regressão linear múltipla.

Fonte de Variação	N° Graus de liberdade	Soma dos Quadrados	Média da Soma dos Quadrados	F_0	Probabilidade P
Modelo	k	SSR	MSR=SSR/k	MSR/MSE	$=P(H_0:F_0 \leq F_{crítico})$
Resíduo	n-k-1	SSE	MSE= SSE/(n-k-1)		
Total	n-1	SST	SST/(n-1)		

Outro indicador importante para o modelo de regressão linear múltipla é o coeficiente de determinação ajustado, R_{ajust}^2. Este índice é uma modificação de R^2 que leva em consideração o número de coeficientes utilizados no modelo. É útil para avaliar se a inserção de novas variáveis contribui para a melhora do modelo.

Ao contrário de R^2, que no mínimo permanece constante ou aumenta quando novos termos são adicionados ao modelo, R_{ajust}^2 pode até diminuir tornando explícita a contribuição de cada novo termo adicionado ao modelo. Matematicamente, este parâmetro é definido por:

$$R_{ajust}^2 = 1 - \frac{(n-1)}{(n-k-1)}(1-R^2) \tag{2.42}$$

A capacidade de previsão do modelo é avaliada através dos mesmos indicadores *PRESS*, Q^2 e *SPRESS* utilizados na regressão linear simples.

Entretanto, a expressão para o desvio padrão da validação é dependente do número de graus de liberdade do resíduo:

$$SPRESS = \frac{\sqrt{PRESS}}{n-k-1} \tag{2.43}$$

A utilização de modelos de regressão linear múltipla está condicionada às mesmas quatro premissas apresentadas no final da seção anterior e a ausência de multicolinearidade entre as variáveis explicativas.

2.5.3 Regressão não-linear: regressão logística

A regressão logística é uma ferramenta estatística apropriada quando a variável de resposta é binária, por exemplo, ativo/inativo, sucesso/fracasso e assim por diante. Este tratamento de dados tem sido aplicado com sucesso em estudos QSAR (LI et al., 2007; CRONIN et al., 2002; HAJDUK et al., 2000).

Nestes casos procura-se pela probabilidade, *P*, de que a variável dependente assuma o valor unitário a partir de um conjunto de variáveis independentes. Para que seja possível utilizar variáveis dependentes deste modo é necessário aplicar uma transformação logística de *P*. Matematicamente define-se:

$$\log it(P) = \log\left(\frac{P}{1-P}\right) \tag{2.44}$$

Nesta forma assegura-se que as probabilidades assumem valores entre 0 e 1 para quaisquer variáveis independentes. O modelo de regressão correspondente é dado por:

$$\log it(P) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 \cdots + b_k x_k \tag{2.45}$$

onde x_i (i=1,…,k) são os descritores moleculares e b_i são os coeficientes de regressão. Assim, a probabilidade de que a variável dependente assuma o valor unitário é dada por:

$$P(Y|x_1,\cdots,x_k)=\frac{1}{1+\exp\left(-\left[b_0+b_1x_1+\cdots+b_kx_k\right]\right)} \tag{2.46}$$

Diferentemente da regressão linear onde os coeficientes da equação minimizam a soma quadrática dos erros, na regressão logística os coeficientes b_j (j = 0, 1, ..., k) são determinados de forma a maximizar a probabilidade de observar os valores das amostras.

O método utilizado na estimativa dos coeficientes b_j é denominado método da máxima verossimilhança. Neste método as observações independentes y_1, y_2, ..., y_N possuem uma densidade de probabilidade dada por:

$$f(y;b_0,b_1,b_2,...,b_k) \tag{2.47}$$

onde b_j (j = 0, 1, ..., k) são constantes desconhecidas que necessitam ser estimadas.

A função de probabilidade conjunta para cada y_i (i = 1, ..., N) é dada pelo produto das funções de densidade de probabilidade:

$$L(y_1,y_2,\cdots,y_n\,|\,b_0,b_1,\cdots,b_k)=\prod_{i=1}^{N}f(y_i,b_0,b_1,\cdots,b_k) \tag{2.48}$$

A Eq. 2.48 torna-se mais apropriada para determinação dos coeficientes b_j em sua forma logarítmica:

$$\Lambda=\ln L=\sum_{i=1}^{N}f(y_i,b_0,b_1,\cdots,b_k) \tag{2.49}$$

Os coeficientes b_j que maximizam a Eq. 2.49 são soluções de k equações dadas por:

$$\frac{\partial \Lambda}{\partial b_j} = 0 \qquad (\mathrm{j} = 0,1,\cdots,\mathrm{k}) \tag{2.50}$$

A maneira mais simplista de se analisar o ajuste de um modelo de regressão logística é por meio de uma tabela de classificação. Cada observação é classificada de acordo com o valor de probabilidade estimado pelo modelo. Se a probabilidade estimada for superior a 0,5 (critério de corte) então a observação é classificada como sucesso (um), caso contrário, como fracasso (zero). Após a classificação, as informações são organizadas como mostrado na Tabela 2-5.

A partir da tabela de classificação são definidos índices como Especificidade e Sensibilidade que são medidas estatísticas do desempenho de um teste de classificação binária.

A Especificidade é matematicamente definida por:

$$Especificidade = \frac{TN}{(FP + TN)} \tag{2.51}$$

onde *TN* é o número de negativos verdadeiros e *FP* o número de falsos positivos. Quando a especificidade assume o valor unitário significa que todos os casos de sucesso são reconhecidos adequadamente.

Similarmente, a Sensibilidade é matematicamente definida por:

$$Sensibilidade = \frac{TP}{(TP + FN)} \tag{2.52}$$

onde *TP* é o número de positivos verdadeiros e *FN* é o número de falsos negativos. O valor unitário para a sensibilidade indica que todos os casos de fracassos são reconhecidos como tais. Entretanto, os valores da tabela de classificação variam conforme o valor de corte estabelecido. Conseqüentemente, os valores para a Especificidade e a Sensibilidade também variam. Por esta razão, a tabela de classificação é a forma mais frágil de se analisar o ajuste do modelo.

Tabela 2-5: Tabela de classificação utilizada em regressão logística.

		Valor Real		
		Sucesso	***Fracasso***	Total
Previsão	***Sucesso***	Positivo Verdadeiro (TP)	Falso Positivo (FP)	P'
	Fracasso	Falso Negativo (FN)	Negativo Verdadeiro (TN)	N'
	Total	P	N	

A variação destes índices quando representada através de um gráfico bidimensional dá origem a curva ROC (*Receiver Operating Characteristic*). A curva ROC expressa a relação entre a Sensibilidade e $(1-\text{Especificidade})$ para diferentes valores de corte. Estatisticamente, a Área Sob a Curva ROC, AUC, pode ser interpretada como a probabilidade do teste distinguir casos de sucesso dos casos de fracasso. O valor unitário para AUC representa o teste perfeito.

A quantidade $-2\ln(\textit{Verossimilhança})$ é um indicador utilizado para avaliar o ajuste do modelo à verossimilhança. Um bom modelo resulta em valores altos para a verossimilhança dos resultados observados o que se traduz em valores baixos para $-2\ln(\textit{Verossimilhança})$.

Este indicador também é útil na seleção de modelos. A diferença entre os valores de $-2\ln(\textit{Verossimilhança})$ para um modelo completo e um reduzido segue aproximadamente uma distribuição χ^2. Se esta diferença for grande, então se rejeita a hipótese nula, ou seja, que parâmetros extras são nulos.

Quando dados são analisados com regressão logística, não há um indicador equivalente ao coeficiente de determinação, R^2. No entanto, existem definições de pseudo-R^2 que são similares a R^2 somente em termos de escala (ou seja, variam de

0 a 1) e quanto maior o valor melhor o ajuste do modelo. Uma definição comumente utilizada é conhecida como pseudo-R^2 de McFadden:

$$pseudo - R^2_{McFadden} = 1 - \frac{\ln \hat{L}(M_{total})}{\ln \hat{L}(M_{\text{int}ercepto})} \tag{2.53}$$

onde $\hat{L}$ é a verossimilhança estimada, M_{total} é o modelo com todos os regressores e $M_{intercepto}$ é o modelo sem os regressores.

O Critério de Informação de Akaike (AIC, em inglês) é uma medida de ajuste do modelo que é aplicado para seleção de modelos. Matematicamente é definido por:

$$AIC = N \ln(RSS) + 2k \tag{2.54}$$

onde N é o número de observações, k é o número de regressores utilizados na equação e RSS é a soma dos quadrados dos resíduos. Quando se quer decidir entre dois modelos, o melhor é o que produz menor valor do Critério de Akaike.

Outra medida utilizada na comparação de modelos é o Critério Bayesiano de Schwarz. Este critério é conceitualmente similar ao Critério de Informação de Akaike e é definido por:

$$BIC = N \ln\left(\frac{RSS}{N}\right) + k \ln(N) \tag{2.55}$$

onde N, RSS e k possuem o mesmo significado descrito para o Critério de Informação de Akaike.

2.6 Árvores de Classificação

Árvore de classificação é uma técnica utilizada para derivar regras a partir de um conjunto de dados formado por categorias, assim como na regressão logística.

O conjunto de dados é analisado utilizando o conceito de ganho de informação (em inglês, *information gain*) que é muito utilizado em teoria da informação.

O Ganho de informação de uma dada variável (ou atributo) X com respeito a um atributo de classe Y é a redução na incerteza de Y quando se conhece o valor de X. A incerteza sobre o valor de Y é definida em termos de entropia. Para compreender a entropia suponha que uma dada variável de classe Y assuma valores $V_1, V_2, \cdots, V_n$ e que as probabilidades de ocorrência destes valores são dadas por $P(Y = V_1) = p_1$, $P(Y = V_2) = p_2$,..., $P(Y = V_n) = p_n$ então a entropia de Y, $H(Y)$, é dada por:

$$H(Y) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{2.56}$$

A expressão acima também é utilizada na definição da entropia condicional, $H(Y|X)$, onde a expressão definida pela Eq. 2.56 é calculada apenas para os casos em que *X* assume o valor W_i. Desta forma o ganho de informação, IG, é dado por:

$$IG = H(Y) - \sum_{i=1}^{n} P(X = W_i) H(Y|X = W_i) \tag{2.57}$$

Cada variável independente é analisada e aquela que fornece o maior ganho de informação ocupa o primeiro nó da árvore. Se houver casos não explicados por esta variável, outros nós são acrescentados com as variáveis remanescentes que fornecem sequencialmente o maior ganho de informação.

2.7 Análise do Domínio de Aplicabilidade do modelo

Previsão de atividade por um modelo QSAR é limitada pelo espaço químico definido pelo conjunto de dados (também denominado conjunto de trabalho) utilizado na proposição do modelo. Quando o conjunto de trabalho é analisado no espaço de descritores do modelo, os compostos são representados

como pontos no espaço multidimensional. Previsões feitas neste espaço correspondem a interpolações no domínio definido pelo modelo. O conhecimento deste domínio propicia a qualificação da confiabilidade do modelo na previsão de novos compostos (JAWORSKA et al., 2005).

Durante a conferência ocorrida na cidade de Setúbal, mencionada no final da seção 2.1 deste capítulo, foi formulada uma definição para o domínio de aplicabilidade do modelo (JAWORSKA et al., 2005):

"O domínio de aplicabilidade de um modelo (Q)SAR é o espaço (físico-químico, estrutural ou biológico), conhecimento ou informação sobre o qual o conjunto de trabalho do modelo foi desenvolvido e para o qual é aplicável fazer previsões para novos compostos. É recomendável que o domínio de aplicabilidade de um modelo (Q)SAR seja descrito em termos dos parâmetros mais relevantes, ou seja, usualmente os descritores utilizados no modelo. Idealmente, o modelo (Q)SAR deveria ser usado para fazer previsões dentro do domínio por meio de interpolação e não extrapolação."

Em termos práticos, esta definição deve ser traduzida em procedimentos quantitativos para se determinar o domínio de aplicabilidade do modelo. Neste caso, existem diversos métodos para definir regiões de interpolação bem como para avaliar similaridade. Como conseqüência da falta de harmonização, diferentes métodos produzem diferentes domínios de aplicabilidade para o mesmo conjunto de dados.

Jaworska e colaboradores (2005) recomendam que o método eleito seja aquele cuja distribuição associada ao conjunto de dados satisfaça as premissas do método.

Para modelos de regressão e classificação os métodos mais apropriados para estimar o domínio de aplicabilidade são aqueles baseados (JAWORSKA et al., 2005; SCHROETER et al., 2007):

- No intervalo dos descritores (*range based*);
- Em definições geométricas (*geometrical based*);
- Em distâncias (*distance based*);
- Na distribuição da densidade de probabilidade (*probability density distribution based*).

Cada uma destas abordagens determina a região de interpolação através da construção de envelopes convexos ao redor do conjunto de dados no espaço multidimensional. Para modelos onde o espaço descritor é unidimensional a região de interpolação é definida pelo mínimo e o máximo valor do conjunto de trabalho.

Métodos baseados no intervalo dos descritores consideram a diferença entre os valores máximos e mínimos de cada descritor para definir a aresta do envelope convexo. No espaço multidimensional, o resultado será a formação de um hipercubo. Neste caso, assume-se como premissa que a distribuição dos dados é uniforme, caso contrário, haverá espaços vazios dentro da região de interpolação. Nos espaços vazios a previsão não é confiável, pois estas regiões não são cobertas pelo conjunto de dados. O pré-processamento dos dados através da análise de componentes principais (PCA) tende a reduzir o volume vazio dos dados originais.

Métodos baseados em definições geométricas são elaborados para calcular diretamente o envelope convexo analisando os limites do conjunto de dados. Entretanto, existem algoritmos eficientes apenas para duas e três dimensões. Nenhuma consideração sobre a distribuição dos dados é feita neste método, portanto, a região de interpolação pode ainda conter espaços vazios.

Métodos baseados em distância identificam a região de interpolação usando o conceito de distância entre pontos no espaço multidimensional. Há que se considerar a forma de coletar os valores de distância, a métrica que será

utilizada no cálculo da distância (Euclideana, Mahalanobis, Manhattan, etc), e o critério de corte para considerar se um ponto pertence ou não ao domínio.

A forma de coletar os valores de distância significa definir o que será mensurado, ou seja, se a distância será calculada entre o ponto em avaliação e o centro de massa do conjunto de dados, ou se será a distância média ou a máxima entre o ponto em avaliação e todos os pontos do conjunto de dados, por exemplo.

A métrica utilizada determina como a distância será mensurada. Os métodos que utilizam a métrica Euclideana e Mahalanobis assumem que a distribuição dos dados é uniforme enquanto que na métrica de Manhattan (também conhecida por City-block) assume-se distribuição triangular.

Finalmente, métodos fundamentados na distribuição da densidade de probabilidade procuram identificar as regiões de alta densidade do conjunto de dados onde serão feitas as previsões. Os métodos se dividem em paramétricos e não-paramétricos. Os métodos paramétricos assumem que a função de densidade tem a forma de uma distribuição padrão como Gaussiana, Poisson, etc. Os métodos não-paramétricos estimam a densidade de probabilidade a partir do conjunto de dados. A semelhança dos métodos baseados em distâncias geométricas, a estimativa de densidade restringe-se a no máximo três dimensões devido à demanda computacional envolvida para dimensões superiores.

Capítulo 3

Estudos Sobre Nucleosídeos

Introdução

O presente capítulo trata da obtenção do modelo QSAR através de regressão logística para uma série de compostos nucleosídicos. Fez-se necessário o uso de regressão logística, pois os dados originais de atividade foram obtidos como percentagem de inibição da replicação do parasito. A descrição do trabalho realizado engloba: a análise conformacional dos compostos pertencentes ao conjunto de trabalho, a geração e seleção de descritores moleculares, a proposição do modelo e a sua validação levando em consideração o respectivo domínio de aplicabilidade. Na abordagem pelo ligante, usada neste trabalho, a conformação bioativa geralmente não é conhecida. Por esta razão, adotou-se a conformação de menor energia encontrada para a proposição do modelo. A validade desta premissa para este conjunto de dados é avaliada e discutida aqui com o auxílio do modelo obtido.

3.1 Ensaios biológicos e transformação do conjunto de dados

A atividade biológica da série dos nucleosídeos mostrada na Figura 3-1 foi determinada por Bhakuni e colaboradores (1989, 1990) em hamsters infectados com amastigotas da cepa Dd-8 de *Leishmania donovani*. Os animais foram tratados por cinco dias consecutivos com uma única dose diária através de injeção intraperitoneal. Foram utilizados 2 a 3 animais para cada composto juntamente com 2 a 3 animais não tratados mantidos para controle. Uma semana após a interrupção do tratamento, foi realizada biópsia no fígado dos animais para contagem dos parasitos remanescentes. O percentual de inibição do crescimento do parasito foi determinado em relação aos animais de controle. O alopurinol foi utilizado como fármaco de referência durante os testes. A Tabela 3-1 mostra os valores obtidos para cada composto da série.

Conforme mencionado na seção 2.1, valores expressos em percentagem de inibição não são apropriados para o desenvolvimento de modelos QSAR. Portanto, estes valores não foram usados diretamente na proposição do modelo QSAR, mas transformados em valores binários de atividade. Desta forma, moléculas que expressaram percentagem de inibição não nula foram consideradas ativas e rotuladas com o valor unitário. As inativas receberam valor nulo. Esta transformação permite que os dados sejam tratados através de regressão logística que será detalhada mais adiante.

3.2 Análise conformacional

De acordo com o que foi apresentado na seção 2.2, a análise conformacional estuda a variação de energia da molécula em função das variações nos seus ângulos diedros. É neste estágio que é eleita a conformação da qual serão extraídas as propriedades moleculares.

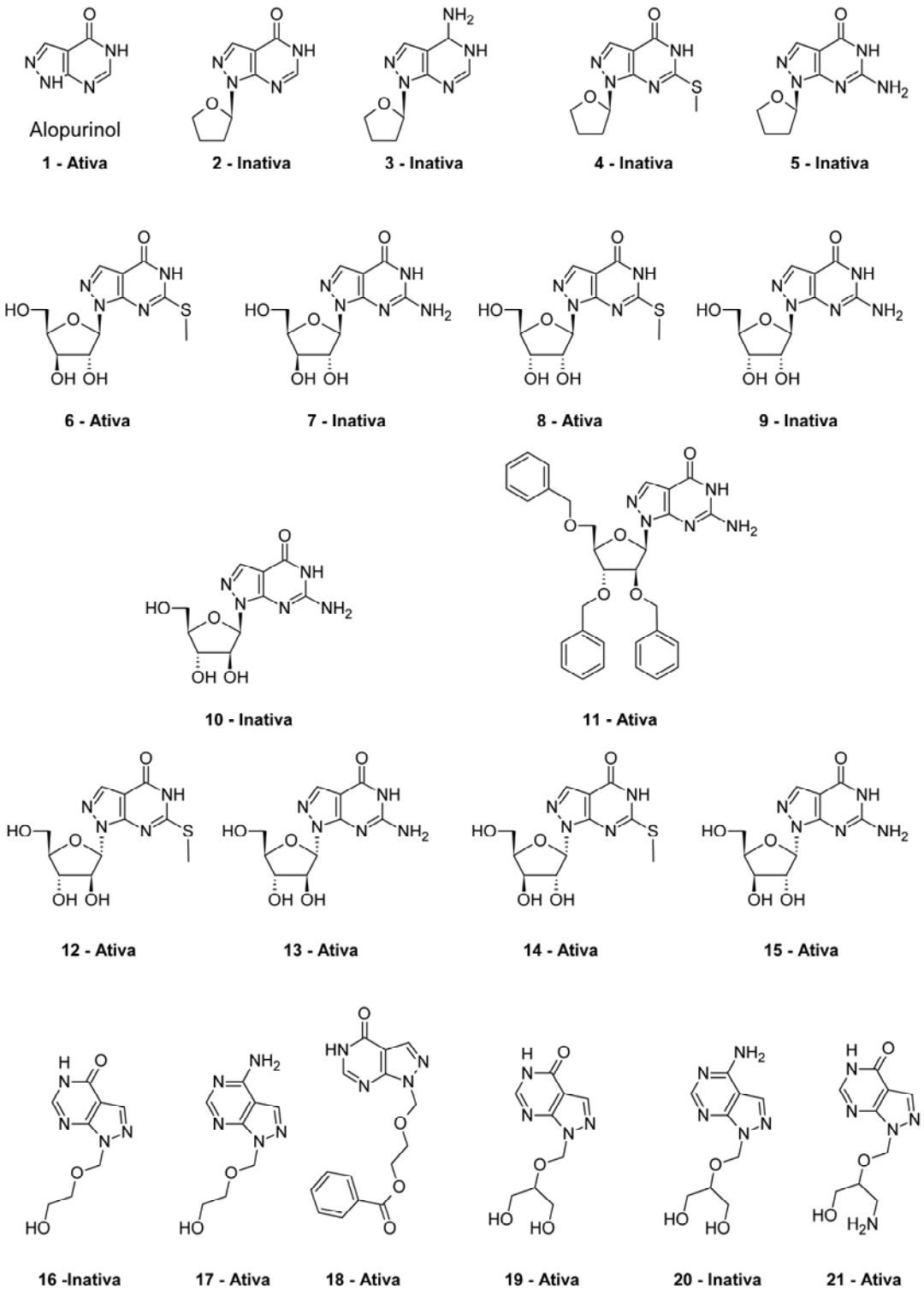


Figura 3-1: Série de nucleosídeos com atividade leishmanicida avaliada *in vivo* (BHAKUNI et al., 1989 e 1990).

Tabela 3-1: Percentagem de inibição do crescimento de formas amastigotas de *Leishmania donovani in vivo* (BHAKUNI et al.,1989 e 1990).

Nº Composto	Dose (mg/kg)	% de Inibição (*in vivo*)
1	25	88
2	25	0
3	25	0
4	25	0
5	25	0
6	25	76
7	25	0
8	10	67
9	25	0
10	25	0
11	25	46
12	25	51
13	25	80
14	25	87
15	25	80
16	25	0
17	25	25
18	25	82
19	25	75
20	25	0
21	25	15

A análise conformacional realizada no presente trabalho utilizou-se de um sistema baseado em regras (ver seção 2.2.5). O programa utilizado foi o OMEGA[1] que figura entre os mais aceitos pela comunidade científica em função de reproduzir conformações em boa concordância com estruturas cristalográficas a uma excelente razão de velocidade e desempenho (BOSTRÖM, 2001, PEROLA e CHARIFSON, 2004).

[1] Os programas computacionais utilizados serão referenciados utilizando letras capitais.

As quatro etapas que compõem o estudo conformacional mencionadas na seção 2.2 estão integradas à rotina deste programa e são tratadas através de palavras-chave antes da execução por meio de um arquivo de parâmetros de controle. As conformações geradas foram consideradas duplicadas se, na comparação com as conformações já aceitas como únicas, o desvio no RMSD (v. Apêndice I) entre as coordenadas fosse inferior a 0,8 Å.

Todas as conformações foram submetidas à otimização utilizando mecânica molecular com o campo de força MMFF corrigido com o termo de solvatação de Sheffield a fim de simular o meio aquoso. Os demais parâmetros do programa foram mantidos na definição padrão. Finalmente, as conformações foram aceitas no intervalo de energia até 13 kcal/mol do mínimo global encontrado e foram enumeradas em ordem crescente de energia.

No caso da série dos nucleosídeos, é importante mencionar que a forma tautomérica adotada para as moléculas fornecidas como entrada para a análise conformacional foi a forma ceto (ver Figura 3-1) por se tratar da forma mais estável na fase gasosa e em solução conforme o trabalho de Orozco et al. (1996). O número total de conformações geradas para esta série foi de 4699.

Em seguida, as conformações foram reordenadas por energia sem otimização de geometria utilizando valores provenientes dos cálculos semi-empíricos que incluíram efeitos de solvatação. Estes cálculos, para cada conformação, foram conduzidos em um único ponto (*single point*) utilizando método semi-empírico com hamiltoniano PM3 juntamente com solvatação implícita em água através do método contínuo do solvente (*solvent continuum method*), SM5.42R, implementado no programa AMSOL 7.1. Cálculos em um único ponto no vácuo também foram realizados para fins de comparação.

Efeitos de solvatação permitem a redistribuição de elétrons através da molécula e tendem a favorecer configurações com momentos dipolares mais

elevados. Como conseqüência, modelos em solução aquosa em geral prevêem confôrmeros estáveis diferentes quando comparados à fase gasosa. A Tabela 3-2 mostra que as conformações mais estáveis em solução aquosa e no vácuo foram diferentes para a grande maioria das moléculas. Também foi observado que, de um modo geral, as conformações mais estáveis em solução aquosa não necessariamente eram mais polares do que as moléculas mais estáveis no vácuo.

No presente trabalho, as moléculas de menor energia em solução aquosa serão utilizadas na confecção do modelo e a representatividade destas conformações em relação ao conjunto conformacional será avaliada por meio do modelo obtido.

3.3 Cálculo dos descritores moleculares

Uma gama de 1492 descritores moleculares foi calculada para cada conformação gerada (ver Tabela 2-2). A maioria dos descritores foi obtida com o programa DRAGON. Estes descritores foram divididos de acordo com a dimensionalidade da molécula envolvida nos cálculos:

- Descritores unidimensionais: 294;
- Descritores bidimensionais: 468;
- Descritores tridimensionais: 721;

Adicionalmente, foram incluídos nove descritores eletrônicos provenientes dos cálculos quânticos com o programa AMSOL 7.1 considerando efeitos de solvatação.

Tabela 3-2: Comparações entre as conformações mais estáveis obtidas em solução aquosa e no vácuo para a série dos nucleosídeos. Os valores de RMSD correspondem às diferenças entre as coordenadas das conformações.

Molécula N°	Conformação N°		$\Delta H_{Solução}$ (kcal/mol)	$\Delta H_{Vácuo}$ (kcal/mol)	RMSD (Å) Sol. Aq./Vac.
	Solução Aquosa	Vácuo			
1	1	1	28,28	46,68	0
2	2	2	-19,40	0,03	0
3	4	4	28,34	45,36	0
4	8	8	-13,12	3,457	0
5	1	1	-22,40	-0,48	0
6	16	3	-146,48	-119,28	1,24
7	17	1	-156,89	-122,95	2,82
8	69	6	-146,52	-117,78	0,80
9	11	4	-155,48	-117,96	2,82
10	13	3	-156,09	-117,78	2,61
11	2	2	-41,43	-12,52	0
12	77	9	-147,23	-118,12	1,28
13	7	7	-156,97	-123,42	0
14	47	27	-144,41	-113,78	2,48
15	13	6	-153,35	-116,89	2,71
16	60	42	-65,25	-40,56	1,48
17	64	42	-17,21	5,065	1,47
18	3596	1586	-66,94	-44,62	2,40
19	179	23	-113,35	-84,10	1,72
20	142	125	-65,02	-38,67	1,50
21	213	3	-64,56	-37,83	2,09

3.4 Seleção dos descritores

Dada a grande quantidade de descritores fez-se necessário a adoção de uma estratégia para a seleção de variáveis. Inicialmente, os dados foram pré-processados por um método não supervisionado baseado em filtros.

A abordagem é conhecida como UFS (ver seção 2.4). O limite de corte estabelecido para a projeção ortogonal durante o pré-processamento foi de 0,99. O número de descritores reduziu para vinte conforme mostrado na Tabela 3-3.

Nas análises regressão linear múltipla é necessário que as variáveis sejam independentes e, na prática corrente, recomenda-se o uso de variáveis cuja correlação não seja superior a 0,6 - 0,7 (KUBINYI, 1993). A colinearidade na regressão logística tem os mesmos efeitos que na regressão linear. A inclusão de variáveis fortemente correlacionadas pode resultar em modelos imprecisos. Métodos para identificação de colinearidade entre variáveis podem apresentar diferentes níveis de sofisticação e, normalmente, as publicações falham por insuficiência descritiva a respeito do método e critérios empregados na análise de multicolinearidade (BAGLEY et al., 2001; MIKOLAJCZYK, 2008). No presente trabalho, a colinearidade entre os descritores foi avaliada através do coeficiente de correlação de Pearson[2]. A Tabela 3-4 mostra que o subconjunto de variáveis selecionadas apresenta valores baixos de correlação. Os valores mais altos de correlação encontrados foram para os pares de descritores nH/R3m+, N-066/ISH com -0,620 e -0,645 respectivamente.

O processo de seleção de variáveis prosseguiu utilizando um método baseado em envoltórios (*wrappers*) disponível no programa ORANGE (módulo *Logistic Regression Classifier*). Neste caso, o algoritmo de aprendizado foi a própria

[2] Consultar Apêndice I.

regressão logística utilizada como uma função de mérito para avaliar cada subconjunto proposto de variáveis.

Tabela 3-3: Descritores selecionados após pré-processamento com o método UFS.

Código	Descrição	Tipo de Descritor
nH	Número de átomos de hidrogênio	Constitucional
nCOORPh	Número de ésteres aromáticos	Grupos Funcionais
nNH2Ph	Número de aminas primárias aromáticas	Grupos Funcionais
N-066	Al-NH2 (Ghose-Crippen)	Fragmentos átomo-centrado
MATS2e	Autocorrelação de Moran - 2º intervalo/ponderada pela eletronegatividade de Sanderson	Autocorrelações 2D
PJI3	Índice de forma Petijean 3D	Geométrico
Mor08u	3D-MoRSE – sinal 8/não ponderado	3D-MoRSE
Mor24u	3D-MoRSE – sinal 24/não ponderado	3D-MoRSE
Mor32u	3D-MoRSE – sinal 32/não ponderado	3D-MoRSE
Mor04m	3D-MoRSE – sinal 04/ponderado pelo peso molecular	3D-MoRSE
Mor17m	3D-MoRSE – sinal 17/ponderado pelo peso molecular	3D-MoRSE
Mor12v	3D-MoRSE – sinal 12/ponderado pelo volume de van der Waals	3D-MoRSE
Mor26v	3D-MoRSE – sinal 26/ponderado pelo volume de van der Waals	3D-MoRSE
Mor28v	3D-MoRSE – sinal 28/ponderado pelo volume de van der Waals	3D-MoRSE
Mor31v	3D-MoRSE – sinal 31/ponderado pelo volume de van der Waals	3D-MoRSE
E2u	2ª componente de acessibilidade direcional índice WHIM/não ponderado	WHIM
ISH	Conteúdo padronizado de informação sobre o leverage equality	GETWAY
HATS8u	*Leverage*-ponderado correlação de intervalo 8/não ponderado	GETWAY
R3m+	R máximo de correlação de intervalo 3/ponderada pela massa atômica	GETWAY
Dip	Momento dipolar total	Eletrônico

Tabela 3-4: Matriz de correlação para os descritores selecionados após pré-processamento com o método UFS. Os descritores com coeficiente de correlação superior a 0,6 foram destacados com fundo preto.

	nH	nCOORPh	nNH2Ph	N-066	MATS2e	PJI3	Mor08u	Mor24u	Mor32u	Mor04m	Mor17m	Mor12v	Mor26v	Mor28v	Mor31v	E2u	ISH	HATS8u	R3m+	Dip
nH	1	0,049	0,249	0,000	-0,049	0,216	-0,068	0,104	0,145	-0,114	0,566	-0,036	-0,214	-0,059	0,541	0,210	0,000	0,015	-0,620	0,189
nCOORPh	0,049	1	-0,213	-0,050	0,222	0,043	-0,235	0,145	0,192	-0,055	0,304	0,246	-0,328	-0,025	-0,054	0,289	0,091	-0,070	-0,200	-0,092
nNH2Ph	0,249	-0,213	1	-0,213	-0,045	0,075	-0,265	-0,065	0,227	-0,555	0,199	-0,155	0,120	-0,024	0,186	0,054	-0,111	0,112	-0,178	0,048
N-066	0,000	-0,050	-0,213	1	-0,123	-0,224	0,293	-0,169	0,330	-0,096	-0,293	-0,475	0,265	-0,029	-0,206	-0,034	-0,645	-0,072	-0,166	0,038
MATS2e	-0,049	0,222	-0,045	-0,123	1	-0,254	-0,562	-0,216	0,037	0,161	-0,043	0,181	0,237	0,511	0,454	0,144	0,157	-0,002	-0,152	-0,027
PJI3	0,216	0,043	0,075	-0,224	-0,254	1	-0,112	-0,284	0,169	0,133	0,437	-0,241	0,033	-0,285	0,137	-0,354	0,209	0,233	-0,157	-0,151
Mor08u	-0,068	-0,235	-0,265	0,293	-0,562	-0,112	1	0,066	-0,119	0,200	-0,093	-0,068	-0,186	-0,328	-0,296	-0,173	-0,244	-0,117	0,165	0,244
Mor24u	0,104	0,145	-0,065	-0,169	-0,216	-0,284	0,066	1	-0,340	-0,231	-0,057	0,412	-0,437	0,106	-0,168	0,330	-0,002	-0,019	0,404	0,270
Mor32u	0,145	0,192	0,227	0,330	0,037	0,169	-0,119	-0,340	1	-0,342	0,353	-0,364	0,006	-0,232	-0,150	0,164	-0,210	-0,186	-0,434	-0,049
Mor04m	-0,114	-0,055	-0,556	-0,096	0,161	0,133	0,200	-0,231	-0,342	1	0,018	0,078	-0,053	-0,048	0,046	-0,303	0,271	-0,203	-0,046	-0,143
Mor17m	0,566	0,304	0,199	-0,293	-0,043	0,437	-0,093	-0,057	0,353	0,018	1	0,130	-0,434	-0,259	0,173	0,196	0,060	-0,244	-0,406	0,170
Mor12v	-0,036	0,246	-0,155	-0,475	0,181	-0,241	-0,068	0,412	-0,364	0,078	0,130	1	-0,553	-0,209	0,009	0,403	0,323	-0,353	0,122	0,109
Mor26v	-0,214	-0,328	0,120	0,265	0,237	0,033	-0,186	-0,437	0,006	-0,053	-0,434	-0,553	1	0,289	0,011	-0,306	-0,029	0,378	0,028	-0,434
Mor28v	-0,059	-0,025	-0,024	-0,029	0,511	-0,285	-0,328	0,105	-0,232	-0,048	-0,259	-0,209	0,289	1	0,350	0,035	-0,282	0,366	0,248	0,181
Mor31v	0,541	-0,054	0,186	-0,206	0,454	0,137	-0,296	-0,168	-0,150	0,046	0,173	0,009	0,011	0,350	1	-0,069	0,098	0,143	-0,364	0,017
E2u	0,210	0,289	0,054	-0,034	0,144	-0,354	-0,173	0,330	0,164	-0,303	0,196	0,403	-0,306	0,035	-0,069	1	-0,116	-0,541	0,000	0,214
ISH	0,000	0,091	-0,111	**-0,645**	0,157	0,209	-0,244	-0,002	-0,219	0,271	0,060	0,323	-0,029	-0,282	0,098	-0,116	1	0,005	-0,105	-0,364
HATS8u	0,015	-0,070	0,112	-0,072	-0,002	0,233	-0,117	-0,019	-0,186	-0,203	-0,244	-0,353	0,378	0,366	0,143	-0,541	0,005	1	0,139	0,130
R3m+	**-0,620**	-0,200	-0,178	-0,166	-0,152	-0,157	0,165	0,404	-0,434	-0,046	-0,406	0,122	0,028	0,248	-0,364	0,000	-0,105	0,139	1	0,133
Dip	0,189	-0,092	0,048	0,038	-0,027	-0,151	0,244	0,270	-0,049	-0,143	0,170	0,109	-0,434	0,181	0,017	0,214	-0,364	0,130	0,133	1

Os subconjuntos de variáveis submetidos ao processo seletivo foram propostos através de uma combinação de seleção prospectiva e eliminação retrospectiva (*stepwise selection*) limitando-se sistematicamente o número de descritores entre um e quatro. Os resultados desta etapa encontram-se na

Tabela 3-5. A acurácia de classificação aumentou a cada inclusão de descritores atingindo o valor máximo em torno de 86% com quatro variáveis. Os descritores selecionados favorecem o reconhecimento das moléculas ativas (casos de sucesso), pois os valores de especificidade se mantêm maiores que os valores de sensibilidade.

Entretanto, no que diz respeito às moléculas classificadas erroneamente pode-se observar que os problemas de classificação ocorreram de forma persistente com as moléculas 2, 16, 17, 20 e 21 (ver Figura 3-2). A comparação entre as moléculas 6 e 7, 8 e 9, 16 e 17 e 19 e 20 (ver Figura 3-3) sugere que outros descritores eletrônicos deveriam ser incorporados ao conjunto de dados nesta fase do processo de seleção, pois a presença de grupos doadores e retiradores de elétrons causaram diferenças na atividade biológica.

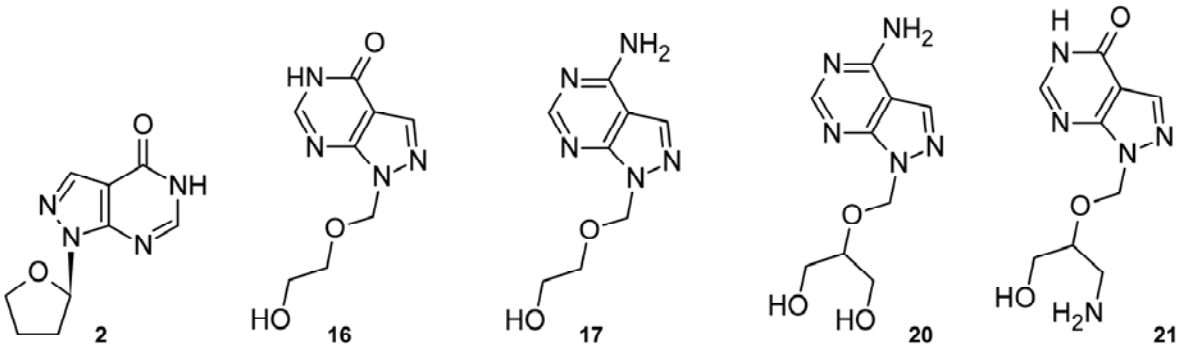


Figura 3-2: Moléculas classificadas erroneamente independente do número de variáveis selecionadas

Tabela 3-5: Eficiência das variáveis selecionadas após segunda redução do conjunto de dados com método baseado em envoltório. As moléculas destacadas correspondem aos casos classificados erroneamente de forma sistemática independente do número de variáveis selecionadas.

Nº Variáveis	Descritores	Acurácia de Classificação (%)	Especificidade	Sensibilidade	AUC†	*Casos Previstos Erroneamente*
1	Mor26v	61,9	0,75	0,44	0,78	*2,6,9,10,16,17,20,21,*
2	Mor26v Mor04m	71,4	0,75	0,67	0,84	*2,8,16,17,20,21*
3	Mor26v Mor04m nH	81,0	0,83	0,78	0,88	*2,17,20,21*
4	*Mor26v* *Mor04m* *nH* *Mor28v*	*85,7*	*0,92*	*0,78*	*0,94*	*16,20,21*

†Área sob a curva ROC.
Para explicação detalhada sobre os índices apresentados nesta tabela consultar seção 2.5.3.

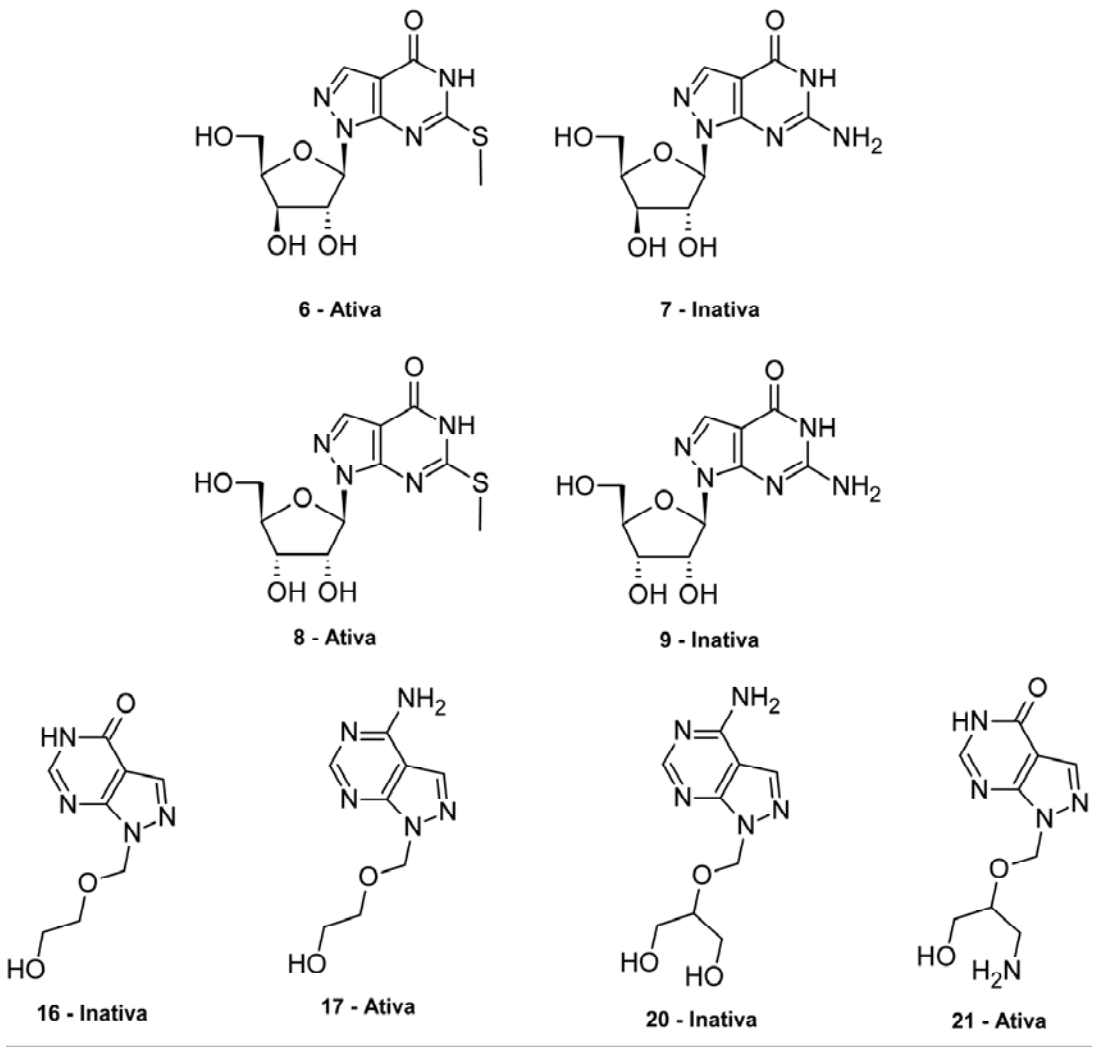


Figura 3-3: Grupos de moléculas similares cuja diferença reside na presença de grupos doadores e retiradores de elétrons.

O processo seletivo foi repetido incluindo oito descritores eletrônicos que foram eliminados durante o filtro com o método UFS, a saber: o calor de formação (ΔH_f), a energia eletrônica (*EE*), a energia de repulsão cerne-cerne (*CCR*), a energia do HOMO (E_{homo}), a energia do LUMO (E_{lumo}), a diferença de energia entre o LUMO e o HOMO (*Gap(Lumo,Homo)*), a diferença de energia entre o HOMO e o HOMO-1, (*Gap(Homo, Homo-1)*) e a energia de solvatação (ΔG_{water}).

Esta estratégia se mostrou eficaz visto que houve uma melhoria na acurácia de classificação (ver Tabela 3-6) ao se considerar descritores eletrônicos entre as variáveis explicativas.

A Tabela 3-7 mostra os valores de correlação entre os descritores eletrônicos incluídos e os descritores selecionados através do método UFS. Este conjunto de dados é um exemplo real onde a combinação de certas variáveis apresenta boa capacidade de previsão, mas estas são excluídas durante o processo automatizado de seleção de variáveis devido à correlação com outras variáveis (ver seção 2.4).

Paralelamente, um método embutido de seleção de variáveis foi aplicado ao conjunto de dados a fim de investigar a possibilidade de criação de regras simples para a classificação dos compostos em ativos e inativos através de uma árvore de classificação. Os resultados desta etapa serão apresentados em uma seção própria mais adiante (ver seção 3.5).

3.5 Proposição do modelo

As combinações de duas e três variáveis apresentaram poder classificatório semelhante (ver Tabela 3-6) e foram utilizadas na proposição de modelos. Os resultados obtidos foram avaliados de acordo com os princípios de Setubal (ver seção 2.1 e 2.5.3) a fim de eleger o modelo mais apropriado para a série de compostos em estudo.

Os modelos encontrados com o auxílio do programa GRETL (módulo *Logit*) para duas e três variáveis foram:

$$Logit(P) = -32{,}318[Mor26v] - 7{,}928[Gap(HOMO, HOMO-1)] + 8{,}794 \qquad (3.1)$$

e

$$Logit(P) = -30{,}901[Mor26v] - 8{,}037[Gap(HOMO, HOMO-1)] - 0{,}002[\Delta H_f] + 8{,}603 \qquad (3.2)$$

Tabela 3-6: Eficiência das variáveis selecionadas com o método baseado em envoltório para o conjunto de dados complementado pelos descritores eletrônicos.

Nº Variáveis[†]	Descritores	Acurácia de Classificação (%)	Especificidade	Sensibilidade	AUC	*Casos Previstos Erroneamente*
1	*Gap(Homo,Homo-1)*	71,4	0,67	0,75	0,82	*2,4,13,15,16,17*
2	*Gap(Homo,Homo-1)* *Mor26v*	90,5	0,89	0,92	0,93	*2,17*
3	*Gap(Homo,Homo-1)* *Mor26v* ΔH_f	*95,2*	*0,90*	*1*	*0.93*	*17*

[†]A inclusão de quatro variáveis não resultou em ganho na acurácia de classificação.

Tabela 3-7: Matriz de correlação para os descritores eletrônicos em relação ao conjunto de descritores selecionados via UFS. Os descritores com correlação superior a 0,6 foram destacados com fundo preto na tabela.

	ΔHf	EE	CCR	Ehomo	Elumo	GAP(Lumo,Homo)	GAP(Homo,Homo-1)	ΔGwater
nH	-0,200	**-0,973**	**0,973**	0,327	0,078	-0,153	-0,270	-0,286
nCOORPh	0,064	-0,060	0,055	-0,326	0,095	0,288	-0,237	0,149
nNH2Ph	-0,061	-0,238	0,241	**0,803**	0,517	-0,119	0,631	-0,262
N-066	0,072	0,083	-0,082	-0,333	0,101	0,297	**-0,718**	-0,031
MATS2e	0,489	0,038	-0,030	0,059	0,096	0,037	-0,079	0,570
PJI3	0,159	-0,144	0,146	0,108	-0,056	-0,115	0,156	0,102
Mor08u	-0,156	0,114	-0,117	-0,151	-0,124	0,001	-0,195	-0,305
Mor24u	-0,665	-0,254	0,242	-0,114	-0,057	0,029	-0,008	-0,562
Mor32u	0,436	-0,062	0,071	-0,193	**0,734**	**0,703**	-0,088	0,271
Mor04m	0,348	0,123	-0,118	-0,228	-0,546	-0,280	-0,281	0,427
Mor17m	0,105	-0,565	0,570	0,189	0,353	0,154	0,049	-0,050
Mor12v	-0,195	-0,117	0,114	-0,092	-0,051	0,020	0,075	-0,123
Mor26v	0,433	0,347	-0,341	0,258	-0,234	-0,353	0,039	0,458
Mor28v	-0,129	0,035	-0,038	0,080	-0,079	-0,115	0,022	0,004
Mor31v	0,126	-0,518	0,523	0,375	-0,074	-0,303	-0,044	0,132
E2u	-0,112	-0,349	0,350	-0,084	0,373	0,348	-0,154	-0,183
ISH	0,215	-0,007	0,011	0,055	-0,310	-0,279	0,305	0,370
HATS8u	-0,349	0,085	-0,096	0,255	-0,276	-0,384	0,244	-0,197
R3m+	-0,316	0,519	-0,527	-0,057	-0,233	-0,146	0,271	-0,221
Dip	-0,492	-0,273	0,266	0,040	0,333	0,236	-0,035	-0,645
ΔHf	1,000	0,297	-0,279	-0,053	0,129	0,136	-0,074	0,912
EE	0,297	1,000	-1,000	-0,299	-0,101	0,116	0,246	0,374
CCR	-0,279	**-1,000**	1,000	0,301	0,107	-0,113	-0,246	-0,359
Ehomo	-0,053	-0,299	0,301	1,000	0,045	-0,619	0,539	-0,203
Elumo	0,129	-0,101	0,107	0,045	1,000	0,757	0,228	-0,118
GAP(Lumo,Homo)	0,136	0,116	-0,113	-0,619	**0,757**	1,000	-0,173	0,040
GAP(Homo,Homo-1)	-0,074	0,246	-0,246	0,539	0,228	-0,173	1,000	-0,097
ΔGwater	**0,912**	0,374	-0,359	-0,203	-0,118	0,040	-0,097	1,000

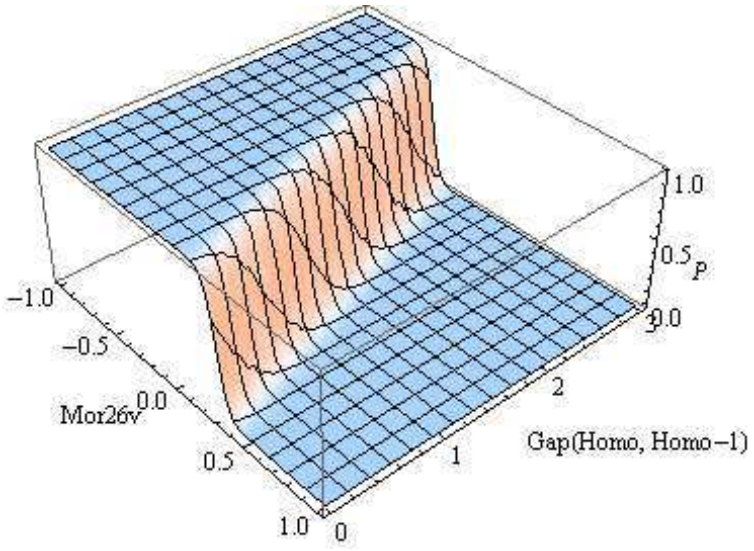
Além dos indicadores da Tabela 3-6, outros índices de confiabilidade destes modelos podem ser encontrados na Tabela 3-8. A análise dos valores calculados para os indicadores -2ln(Verossimilhança), Critério de Informação de Akaike e Critério Bayesiano de Schwarz, indica que a inclusão da terceira variável não representou um ganho real. Portanto, o modelo QSAR final eleito é representado pela Eq. 3.1. O número de variáveis explicativas do modelo final está consistente com a recomendação corrente de usar 10 observações (ou eventos) para cada variável incluída no modelo (PEDUZZI et al., 1996).

Tabela 3-8: Índices de confiabilidade para os modelos com duas e três variáveis.

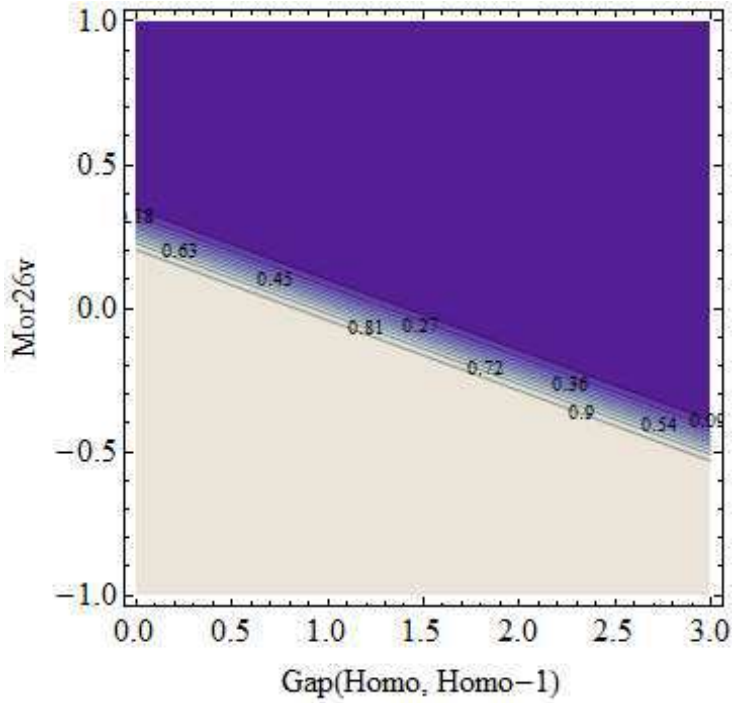
Modelo	-2ln(Verossimilhança)	Pseudo-$R^2_{McFadden}$	AIC[†]	BIC[††]
3.1	-7,586	0,47	21,17	24,31
3.2	-7,568	0,47	23,14	27,31

[†]Critério de Informação de Akaike.
[††]Critério de Bayesiano de Schwarz.
Para explicação detalhada sobre os índices apresentados nesta tabela consultar seção 2.5.3

Ambos descritores aparecem com sinal negativo na equação. Isto significa que a cada decréscimo nos seus valores aumenta a probabilidade de que a molécula seja ativa. A Figura 3-4 mostra a variação dos valores da probabilidade, P, em função dos descritores *Mor26v* e *Gap(Homo, Homo-1)*. Como mostrado na Figura 3-4(b), a região clara corresponde aos valores dos descritores que mantêm a probabilidade acima de 50%, ou seja, as moléculas que apresentam valores de *Mor26v* e *Gap(Homo, Homo-1)* nesta região são previstas como ativas. Na região violeta as moléculas são previstas como inativas. Para uma mesma molécula, as conformações podem apresentar combinações destes descritores na faixa que divide estas regiões. Nesta faixa, pequenas variações nos valores dos descritores são suficientes para que a classificação da conformação mude entre ativa e inativa.



(a)



(b)

Figura 3-4: (a) Valores da probabilidade, P, em função dos descritores *Mor26v* e *Gap(Homo, Homo-1)*. (b) Curvas de nível da superfície de probabilidade. A região clara corresponde aos valores dos descritores que mantêm a probabilidade acima de 50%, ou seja, as moléculas que apresentam valores de *Mor26v* e *Gap(Homo, Homo-1)* nesta região são previstas como ativas. Na região violeta as moléculas são previstas como inativas.

Mor26v é calculado a partir da equação de difração eletrônica, portanto, é um descritor que codifica informações tridimensionais das moléculas. Para esta série de compostos, em particular, fica clara a relação deste descritor com a quiralidade das moléculas após análise da Tabela 3-9. O descritor *Mor26v* assume um valor positivo para o composto 6 ao passo que para o composto 14 este valor é negativo. Entretanto, a única diferença estrutural entre estes compostos é a quiralidade (ver Figura 3-5). O mesmo comportamento para o sinal de *Mor26v* é observado para os pares de compostos 7/15, 10/13, 12/14 (ver Figura 3-5).

Estas diferenças de quiralidade têm influência sobre a atividade destes compostos. Apesar da mudança na quiralidade, os pares de compostos 6/14, 6/8, 12/14, 13/15 são ativos assim como 7/9 e 9/10 são inativos (ver Figura 3-6). Isto demonstra que a quiralidade não é a única propriedade molecular que determina a atividade biológica destas moléculas.

A adição do descritor *Gap(Homo,Homo-1)* complementa a classificação do modelo visto que as moléculas ativas mostraram diferenças menores entre estes orbitais e as inativas diferenças maiores (ver Tabela 3-9). Barone e colaboradores (1996) observaram que o descritor *Gap(Homo,Homo-1)* estava intimamente relacionado com a atividade carcinogênica de hidrocarbonetos aromáticos policíclicos. A mesma dependência foi encontrada entre hormônios derivados de progesterona (BRAGA et al., 2000). Entretanto, a correlação deste descritor com a atividade biológica não é claramente compreendida.

Para esta série de nucleosídeos, HOMO e HOMO-1 são orbitais π localizados no esqueleto do Alopurinol. A aparência do HOMO das moléculas 8 e 9, por exemplo, são praticamente idênticas. Entretanto, para HOMO-1 observam-se diferenças que estão relacionadas aos efeitos dos substituintes (ver Figura 3-7) sobre a estrutura eletrônica das moléculas.

Tabela 3-9: Valores numéricos para os descritores *Mor26v* e *Gap(Homo,Homo-1)* para o conjunto de trabalho e o conjunto de teste.

Molécula	*Mor26v*	*Gap(Homo,Homo-1)* (eV)	P(A=1)[†] Calculada	Atividade Binária
Conjunto de Trabalho				
1	0,004	0,7568	0,93	1,0
2	0,052	0,8810	0,53	0,0
3	0,139	1,0819	0,01	0,0
4	0,145	0,7168	0,17	0,0
5	0,082	1,0581	0,10	0,0
6	0,059	0,7098	0,78	1,0
7	0,083	1,0250	0,12	0,0
8	0,056	0,7131	0,79	1,0
9	0,038	1,0890	0,26	0,0
10	0,037	1,0710	0,29	0,0
11	-0,002	0,5708	0,99	1,0
12	0,029	0,7151	0,90	1,0
13	-0,015	1,0116	0,78	1,0
14	-0,023	0,6994	0,98	1,0
15	-0,014	1,0226	0,76	1,0
16	0,051	0,9062	0,49	0,0
17	0,111	1,1021	0,03	1,0
18	-0,032	0,5759	0,99	1,0
19	-0,025	0,9083	0,92	1,0
20	0,050	1,0879	0,19	0,0
21	0,106	0,0313	0,99	1,0
Conjunto de Teste				
22	0,000	1,0682	0,58	1,0
23	0,000	0,3273	1,00	1,0
24	0,000	1,3829	0,10	0,0
25	0,000	1,3941	0,09	0,0
26	0,000	1,4164	0,08	1,0
27	0,000	2,0641	0,00	1,0
28	0,000	2,0012	0,00	1,0
29	0,000	1,6613	0,01	1,0
30	0,000	0,8760	0,86	1,0
31	0,000	1,4018	0,09	1,0
32	0,000	0,1121	1,00	1,0
33	0,000	1,2116	0,31	1,0
34	0,000	0,9509	0,78	1,0
35	0,000	1,0209	0,67	1,0

[†]P(A=1) é a probabilidade de que a atividade assuma um valor unitário.

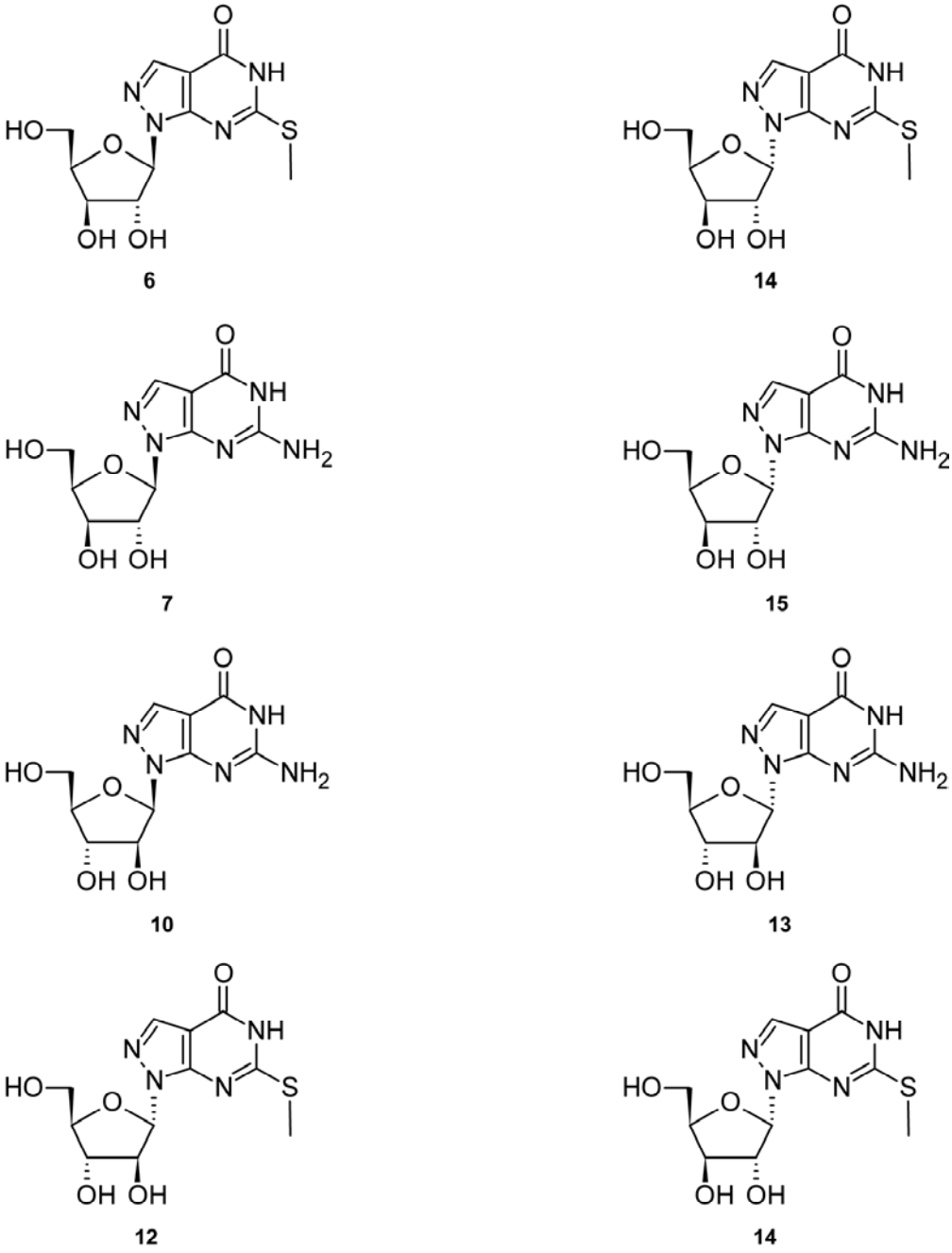


Figura 3-5: Moléculas que diferem entre si pela quiralidade e que apresentam sinais opostos para o descritor *Mor26v*.

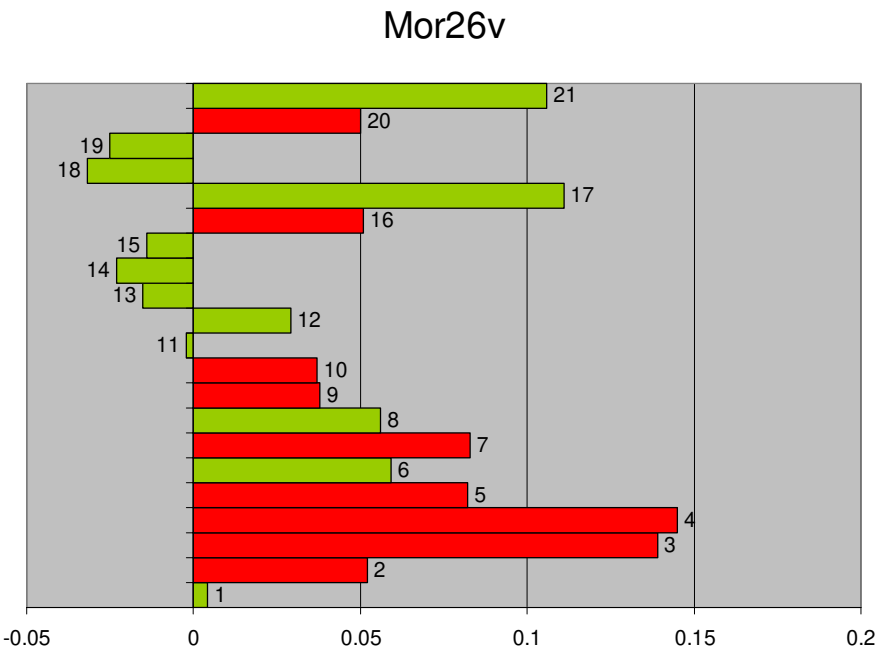


Figura 3-6: Gráfico dos valores de *Mor26v* para o conjunto de trabalho. As moléculas ativas estão representadas em verde e as inativas em vermelho.

A magnitude do *Gap(Homo,Homo-1)* não é afetada pela quiralidade visto que enantiômeros apresentam valores semelhantes como é o caso dos pares 6/14, 6/8, 7/9, 7/15, 9/10, 10/13, 12/14 e 13/15.

O papel dos descritores *Mor26v* e *Gap(Homo,Homo-1)* na Eq. 3.1 é complementar: *Mor26v* distingue conformações e enantiômeros enquanto que *Gap(Homo,Homo-1)* distingue as diferenças na estrutura eletrônica.

Todos os compostos foram classificados adequadamente exceto os compostos 2 e 17. Este resultado corresponde a 90,5% de acurácia de classificação (ver Tabela 3-6) que analisada em conjunto com os outros índices de confiabilidade indica boa qualidade do modelo. Adicionalmente, o modelo foi submetido a validação cruzada (*leave-one-out*) atingindo 85,7% para a acurácia de classificação o que corresponde a um excelente valor para aceitabilidade do modelo.

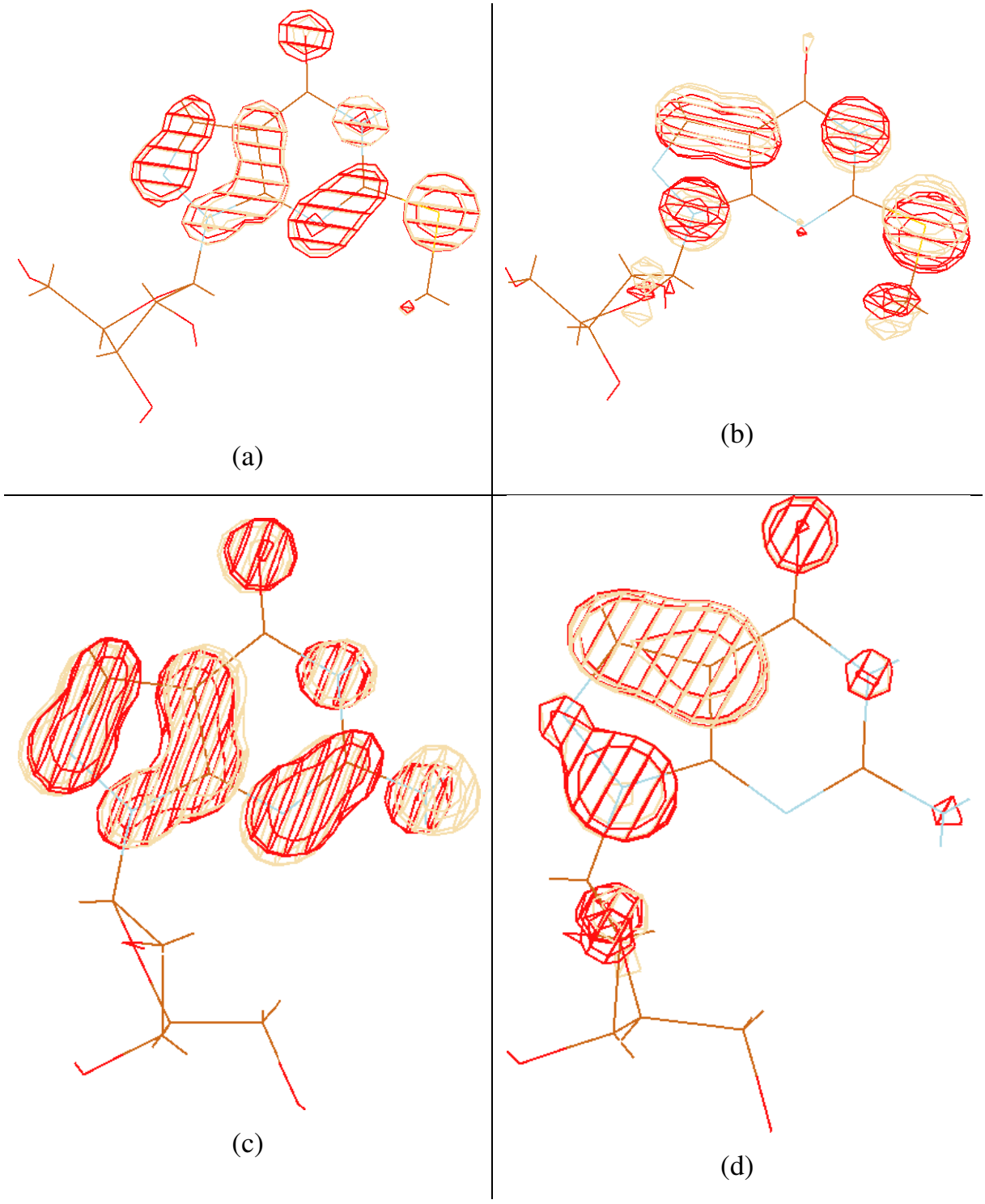


Figura 3-7: Orbitais Moleculares HOMO e HOMO-1. (a) HOMO para a molécula 8. (b) HOMO-1 para a molécula 8. (c) HOMO para a molécula 9. (d) HOMO-1 para a molécula

3.6 Árvore de classificação

A árvore de classificação foi construída com o programa ORANGE (módulo *Classification Tree*) usando o conjunto de descritores selecionados com o método UFS complementado pelos descritores eletrônicos mencionados na seção 3.4.

Este método se baseia no ganho de informação e as variáveis mais relevantes na criação de regras de classificação foram novamente os descritores *Mor26v* e *Gap(Homo, Homo-1)*.

A classificação das moléculas em ativas e inativas segue três regras:

1ª Regra: Se o descritor *Mor26v* < 0,033 então a molécula é ativa;

2ª Regra: Se o descritor *Mor26v* ≥ 0,033 e *Gap(Homo,homo-1)* < 0,715 então a molécula é ativa;

3ª Regra: Se o descritor *Mor26v* ≥ 0,033 e *Gap(Homo,Homo-1)* ≥ 0,715 então a molécula é inativa.

A Figura 3-8 mostra um resumo gráfico destas regras. A acurácia de classificação atingida foi de 95% sendo que apenas a molécula 17 não se enquadra nas regras geradas. Como pode ser observado, moléculas ativas tendem a apresentar valores menores para os descritores *Mor26v* e *Gap(Homo,Homo-1)*. Este resultado está de acordo com a tendência mostrada pelo modelo de regressão logística (ver Figura 3-4). Estas duas abordagens foram concordantes quanto a relevância dos descritores *Mor26v* e *Gap(Homo,Homo-1*) para a atividade leishmanicida de compostos derivados de nucleosídeos.

3.7 Análise das Premissas

O presente modelo de regressão logística foi construído utilizando os confôrmeros de menor energia. Neste caso, assumiu-se que tais confôrmeros são apropriados para modelar a atividade leishmanicida através de variáveis que codificam características relevantes do conjunto de trabalho.

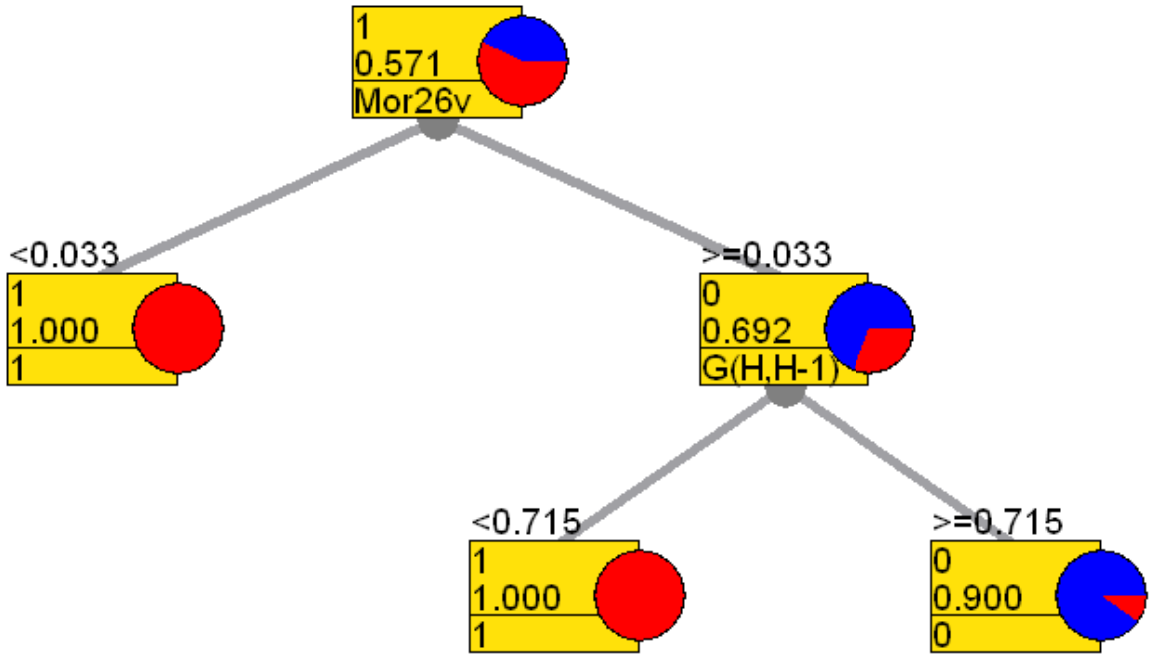


Figura 3-8: Árvore de classificação para a série de nucleosídeos. O topo da árvore é ocupado pelo descritor Mor26v que possui maior poder discriminatório de atividade. O nível seguinte é ocupado pelo descritor Gap(Homo,Homo-1). As regras de classificação estão mostradas acima das caixas.

Para avaliar esta premissa, a Eq. 3.1 foi aplicada a todas as conformações geradas. A substituição numérica dos valores dos descritores de cada conformação na Eq. 3.1 indica se a conformação em questão pertence à categoria ativa ou inativa. A Tabela 3-10 lista a percentagem de conformações ativas e inativas previstas pelo modelo de regressão logística para cada conjunto conformacional.

No caso da molécula 18, por exemplo, há 3589 confôrmeros diferentes na janela de energia de 13 kcal/mol. Todas as conformações desta molécula foram previstas como ativas pelo modelo. Isto porque os valores calculados para *Mor26v* pertencem ao intervalo [-0,15, 0,06] e os valores calculados para o *Gap(Homo, Homo-1)* pertencem ao intervalo [0,47, 0,88] eV. Estes intervalos estão localizados na região onde a probabilidade de uma molécula (ou conformação) ser ativa é superior a 50% conforme mostrado na Figura 3-4b.

O mesmo ocorre com as moléculas 1, 3, 5, 6, 8, 11, 12 e 21 onde todas as conformações pertencem a apenas uma categoria (ativa ou inativa). Tomando a molécula 4 como um outro exemplo, 71,4% do conjunto conformacional correspondente foi classificado como inativo sendo que o confôrmero de menor energia pertence a este grupo. Entretanto, 28,6% das conformações desta molécula foram classificadas como ativas. Para esta molécula, os valores de Mor26v pertencem ao intervalo [0,05, 0,18] e o Gap(Homo, Homo-1) praticamente não varia, pois os valores estão em torno de 0,7 eV. Observando a Figura 3-4b, nota-se que estes valores estão situados em uma região onde pequenas variações nos valores dos descritores influenciam a classificação das conformações.

Similarmente, as moléculas 2, 7, 9, 10, 13, 15, 16, 17, 19 e 20 apresentam conformações em ambas as categorias. É razoável pensar que moléculas inativas possam apresentar conformações ativas, porém estas não são favorecidas durante a interação com a macromolécula alvo. O mesmo raciocínio pode ser aplicado às conformações inativas de moléculas ativas (BECKER, 2000).

De acordo com a Tabela 3-10, as moléculas 2 e 17, que não foram classificadas adequadamente, apresentariam conformações mais apropriadas que se enquadram nas respectivas categorias de atividade observadas experimentalmente. Considerações sobre o *ensemble* conformacional são objeto de estudo de análises QSAR-4D (HOPFINGER et al., 1997) as quais consideram como conformação bioativa aquela que otimiza o modelo quantitativo. Apesar de ser possível atingir 100% de acurácia de classificação mudando-se as conformações das moléculas 2 e 17, este tipo de abordagem só seria válida para o conjunto de trabalho, pois para previsões sobre novas moléculas não seria viável adotar um critério objetivo para eleger a conformação.

A conformação de menor energia explica corretamente a atividade de 19 das 21 moléculas do conjunto de trabalho e, na maioria dos casos, a pertence ao

conjunto majoritário (> 50% das conformações). Portanto, pode-se concluir que as conformações de menor energia foram representativas para todo o conjunto conformacional, pois capturaram características moleculares relevantes para a série em estudo.

Tabela 3-10: Percentagem de conformações ativas e inativas estimadas pelo modelo de regressão logística para cada grupo conformacional do conjunto de trabalho. O número total de conformações para cada molécula está listado na última coluna.

Molécula	Conformações Previstas (%)		Atividade Observada	Número de Conformações
	Inativas	Ativas		
1	0	100(*)	Ativa	1
2	42.9	57.1(*)	Inativa	7
3	100(*)	0	Inativa	7
4	71.4(*)	28.6	Inativa	14
5	100(*)	0	Inativa	3
6	0	100(*)	Ativa	58
7	85(*)	15	Inativa	20
8	0	100(*)	Ativa	79
9	81.8(*)	18.2	Inativa	11
10	53.8(*)	46.2	Inativa	13
11	0	100(*)	Ativa	2
12	0	100(*)	Ativa	83
13	17.4	82.6(*)	Ativa	23
14	0	100(*)	Ativa	49
15	62.5	37.5(*)	Ativa	16
16	26.7(*)	73.3	Inativa	60
17	92.2(*)	7.8	Ativa	64
18	0	100(*)	Ativa	3598
19	8.9	91.1(*)	Ativa	213
20	92.4(*)	7.6	Inativa	145
21	0	100(*)	Ativa	233
Total				4699

(*) A conformação de menor energia pertence a este grupo.

3.8 Validação do modelo

O conjunto de teste é formado por quatorze moléculas que compartilham a estrutura nucleosídica e foram enumeradas de 22 a 35 conforme mostrado na Figura 3-9. Elas foram selecionadas a partir de dados na literatura (HASAN et al., 2006; BERMAN et al., 1987; MARR et al., 1984).

As moléculas 22, 23, 26 a 35 exibem atividade leishmanicida. Aciclovir (molécula 24) e Ganciclovir (molécula 25) são inibidores seletivos da enzima timidina cinase do tipo I e são usados no tratamento do vírus *Herpes simplex*. Estes dois compostos não apresentam atividade leishmanicida. De acordo com Birringer et al. (2005) existem dois tipos de timidina cinase: a do tipo I e a do tipo II. A timidina cinase do tipo I é expressa no vírus *Herpes simplex*. A timidina cinase do tipo II é expressa pelo parasito *Leishmania* (MUYOHBWE et al., 1997). Aciclovir e Ganciclovir são inibidores apenas da timidina cinase do tipo I e, por esta razão, estes dois fármacos não apresentam atividade leishmanicida.

Uma premissa muito importante em química medicinal assume que moléculas similares exibem atividades biológicas similares (KUBINYI, 2002). Apesar da similaridade estrutural do Aciclovir e do Ganciclovir com a molécula 22 (ver Figura 3-9), não há atividade biológica similar, aparentemente quebrando um paradigma em análises QSAR.

O mesmo procedimento descrito anteriormente foi repetido para gerar conformações limitando para 1000 conformações, no máximo, por molécula. Então a Eq. 3.1 foi aplicada para prever atividade. Os compostos 22 a 25, 30, 32, 34 e 35 foram classificados corretamente pelo modelo de regressão logística. Entretanto, os compostos 26, 27, 28, 29, 31 e 33 não foram classificados adequadamente, pois foram considerados inativos pelo modelo de regressão logística.

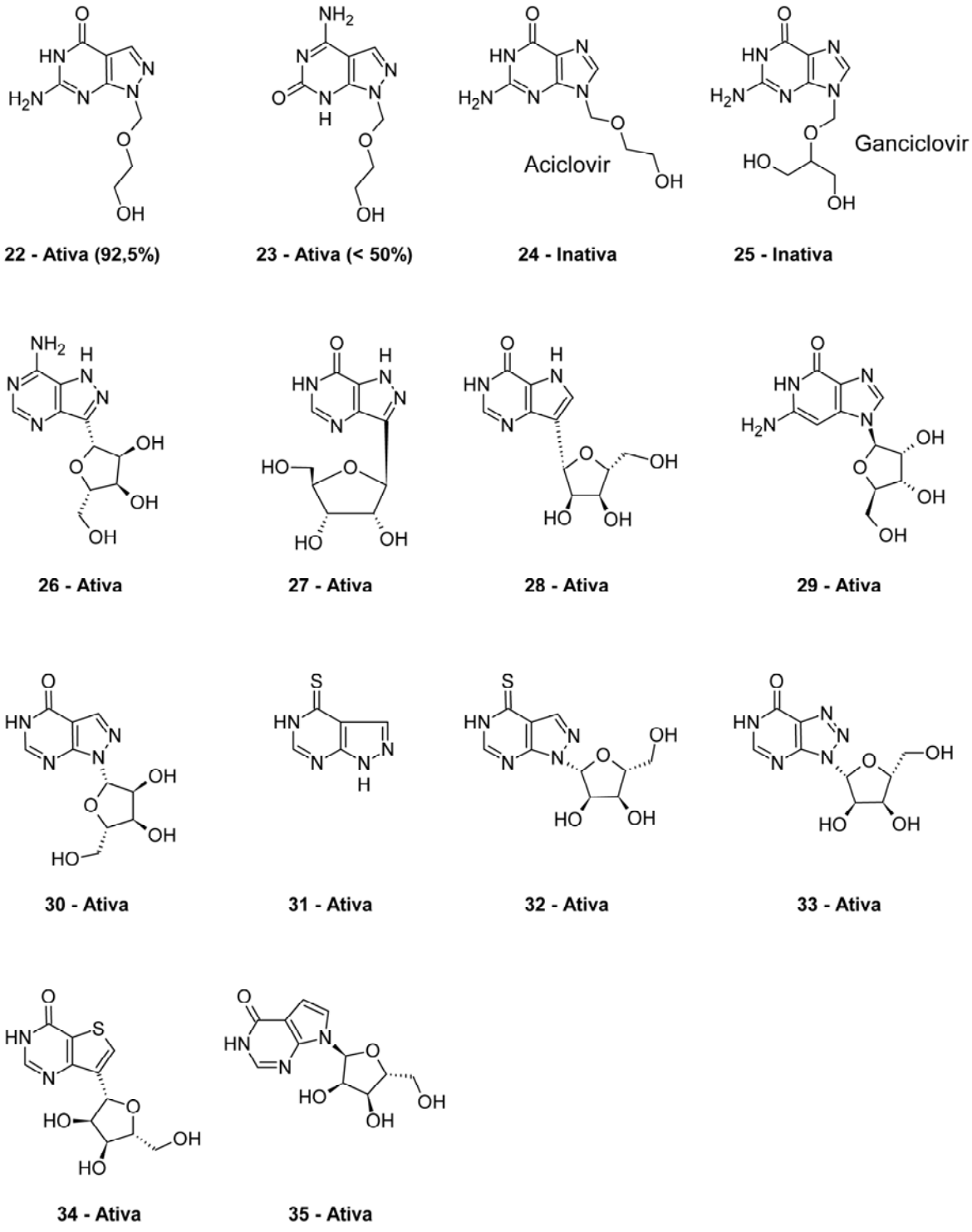


Figura 3-9: Conjunto de teste para a série dos nucleosídeos.

Como nucleosídeos apresentam amplo espectro de atividade, a seleção de compostos por similaridade estrutural para compor o conjunto de teste é uma tarefa difícil mesmo para nucleosídeos com conhecida atividade leishmanicida como os compostos 26, 27, 28, 29, 31 e 33.

Kubinyi (2002) argumenta que similaridade e diversidade química dependem da estrutura tridimensional e das propriedades do sítio ativo da macromolécula alvo e que, portanto, estes conceitos não são facilmente definidos de forma objetiva. O autor sustenta sua argumentação mostrando exemplos de diferentes modos de ação para moléculas percebidas como similares.

Porém, para compreender a razão do desvio de previsão da eq. 3.1 para estes compostos é necessário efetuar uma análise do domínio coberto pelo espaço descritor do modelo. Assim, a análise do domínio QSAR do modelo proposto foi realizada usando distância euclidiana no espaço descritor e análise de componentes principais (PCA) como estágio de pré-processamento. Ambos são módulos integrantes do programa AMBIT DISCOVERY.

Todos os compostos do conjunto de trabalho foram utilizados na determinação do domínio do modelo e a interpolação aplicada no conjunto de teste. Esta análise revelou que os compostos 27 e 28 não pertencem ao domínio do modelo proposto (ver Figura 3-10). Estes compostos estão, portanto, fora do escopo do modelo e devem ser excluídos do conjunto de teste.

O descritor *Mor26v* assumiu valores nulos para todas as moléculas do conjunto de teste e suas respectivas conformações (ver Tabela 3-9), portanto, o poder de discriminação do modelo de regressão logística depende exclusivamente dos valores do *Gap(Homo,Homo-1)*. O valor mínimo encontrado para este descritor entre as moléculas 26, 29, 31 e 33 foi 1,2 eV. Valores acima deste remetem para a região onde as moléculas (ou conformações) são classificadas como inativas (ver Figura 3-4).

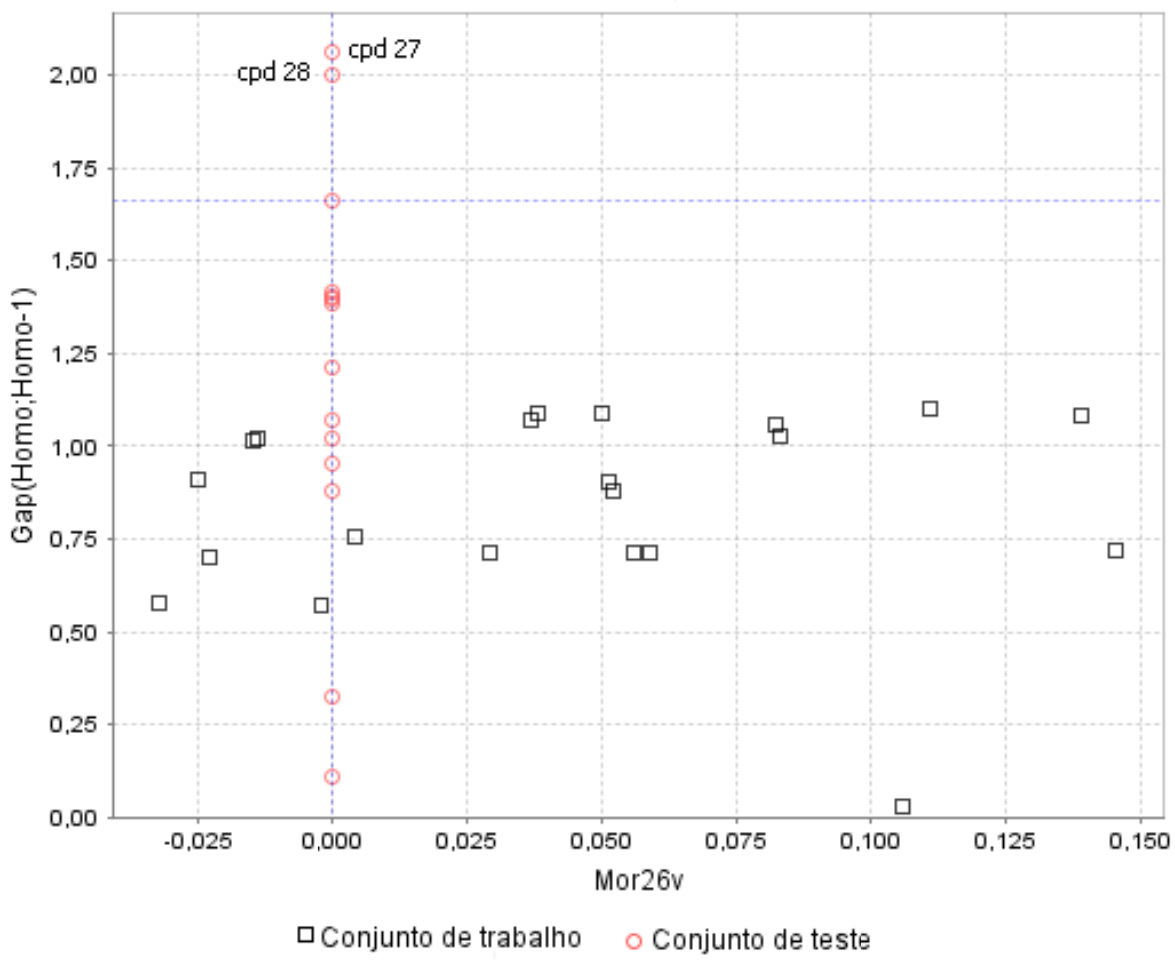


Figura 3-10: Análise do domínio de aplicabilidade do modelo QSAR para a série de nucleosídeos. Os compostos 27 e 28 não pertencem a região de interpolação do modelo.

Seguindo a mesma abordagem feita para o conjunto de trabalho, a Eq. 3.1 foi aplicada para cada conformação do conjunto de teste. Embora as moléculas 27 e 28 não pertençam ao domínio do modelo apenas por completeza. Os resultados podem ser conferidos na Tabela 3-11. No conjunto de teste, as conformações de uma dada molécula foram classificadas em apenas uma categoria (ativa ou inativa). Todas as conformações para Aciclovir (molécula 24) e Ganciclovir (molécula 25) foram classificadas como inativas pelo modelo QSAR de regressão logística. Isto ilustra a capacidade do modelo em discernir isósteros.

Tabela 3-11: Percentagem de conformações ativas e inativas estimadas pelo modelo de regressão logística para cada grupo conformacional do conjunto de teste. O número total de conformações para cada molécula está listado na última coluna.

Molécula	Conformações Previstas (%)		Atividade Observada	Número de Conformações
	Inativas	Ativas		
22	0	100(*)	Ativa	57
23	0	100(*)	Ativa	55
24	100(*)	0	Inativa	29
25	100(*)	0	Inativa	215
26	100(*)	0	Ativa	23
27	100(*)	0	Ativa	35
28	100(*)	0	Ativa	33
29	100(*)	0	Ativa	30
30	0	100(*)	Ativa	22
31	100(*)	0	Ativa	1
32	0	100(*)	Ativa	24
33	100(*)	0	Ativa	22
34	0	100(*)	Ativa	37
35	0	100(*)	Ativa	8
Total				591

(*) A conformação de menor energia pertence a este grupo.

O conjunto de teste originalmente proposto também foi avaliado com as regras derivadas para a árvore de classificação. De acordo com este modelo, todas as moléculas foram previstas como ativas, inclusive o Aciclovir e o Ganciclovir os quais são inativos. A razão para isto é que, em função dos valores nulos de *Mor26v*, a atividade para todas as moléculas foi prevista conforme a primeira regra da árvore de classificação.

O modelo de regressão logística atingiu 58% de acurácia de classificação para o conjunto de teste enquanto que a árvore de classificação atingiu 83%. Apesar do modelo de regressão logística ter apresentado um desempenho menor do que a árvore de classificação no conjunto de teste, aquele foi capaz de discernir corretamente moléculas inativas, ou seja, o modelo de regressão logística apresentou sensibilidade (ver seção 2.5.3).

Capítulo 4

Estudos Sobre Antifúngicos

Introdução

O presente capítulo trata da obtenção do modelo QSAR através de regressão linear múltipla para uma série de antifúngicos com ação leishmanicida *in vitro*. Os compostos pertencentes a esta série apresentam grupos ionizáveis. O estado de ionização que foi utilizado nos cálculos está fundamentado nas interações conhecidas destes compostos com as enzimas alvo que participam na síntese do ergosterol. A descrição do trabalho realizado segue a mesma estrutura apresentada no capítulo anterior, ou seja: a análise conformacional dos compostos pertencentes ao conjunto de trabalho, a geração e seleção de descritores moleculares, a proposição do modelo e a sua validação levando em consideração o respectivo domínio de aplicabilidade.

4.1 Ensaios biológicos

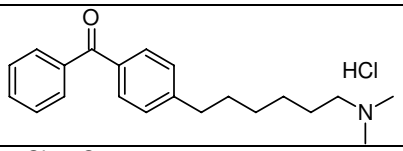
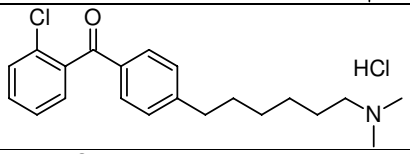
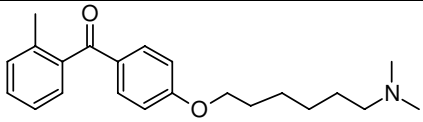
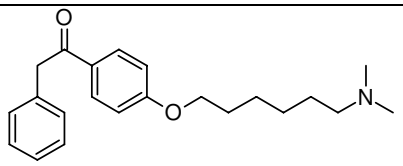
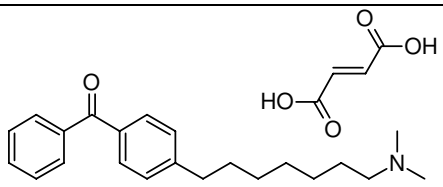
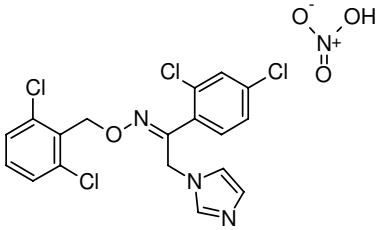
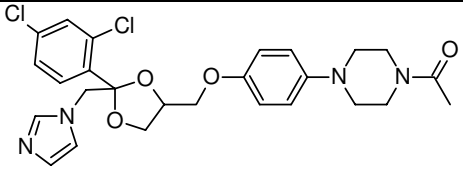
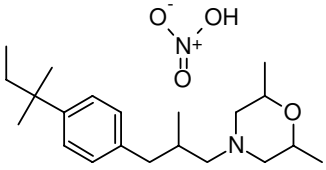
A atividade biológica da série dos antifúngicos mostrada na Tabela 4-1 foi determinada por Gebre-Hiwot e Frommel (1993) *in vitro* em células humanas THP-1 (linhagem monocítica da leucemia humana) infectadas com amastigotas de *Leishmania donovani*. Estes compostos são inibidores da síntese do ergosterol de fungos. A motivação para investigar fármacos fungicidas como potencias agentes leishmanicidas reside na similaridade do metabolismo de lipídios entre fungos e parasitos de *Leishmania*. As células infectadas foram encubadas com os fármacos por seis dias em diferentes concentrações para a determinação da dose efetiva média, ED_{50}. Células de controle, ou seja, células infectadas não submetidas a tratamento foram mantidas para referência. As medidas foram feitas em triplicatas e os valores de ED_{50} determinados a partir da percentagem de inibição das células infectadas em relação às células de controle. O cetoconazol (ver Tabela 4-1) se mostrou praticamente inativo durante os ensaios e o valor de ED_{50} foi estimado por extrapolação.

A citotoxicidade dos compostos foi avaliada através de ensaios MTT (*thiazolyl blue tetrazolium bromide*) após a incubação com células não infectadas por seis dias em diferentes concentrações. Os valores de citotoxicidade foram expressos em termos de dose letal média, LD_{50}. A razão entre LD_{50} e ED_{50} é denominada índice terapêutico, IT. O IT é uma comparação entre a quantidade de um agente terapêutico necessária para causar um efeito terapêutico e a quantidade que causa efeitos tóxicos.

4.2 Considerações sobre o estado de ionização

A tendência de ionização de um composto é expressa pelo valor da constante de ionização, K_a, e é uma função da acidez ou basicidade de grupos presentes no

Tabela 4-1: Série de antifúngicos com atividade leishmanicida avaliada *in vitro* (Gebre-Hiwot e Frommel, 1993). Os valores mostrados correspondem a dose efetiva média, ED50, a dose letal média, LD50, e o índice terapêutico, IT.

	Nº	Molécula	Estrutura	ED_{50} (μM)	LD_{50} (μM)	IT
Inibidores de 2,3 oxidoesqualeno ciclase	1	RO 43-3815		1,7	7,8	4,6
	2	RO 43-5955		2,37	13,1	5,5
	3	RO 43-8208		10	29,5	3,0
	4	RO 42-6589		6,22	19,7	3,2
	5	RO 43-0688		1,54	10,2	6,6
Inibidores de 14-α–desmetilase	6	Oxiconazol		6,69	13,4	2,0
	7	Cetoconazol		>300	17,7	<0,1
Inibidor de Δ−14 –redutase e Δ-8- Δ−7-isomerase	8	Amorolfina		4,19	38,4	9,2

composto. Comumente, esta propriedade é expressa pelo pK_a que é definido como $-\log K_a$. O estado de ionização de uma molécula depende do pK_a e do pH do meio com o qual ela interage. O percentual de ionização relaciona estas grandezas através da expressão:

$$\%ionização = \frac{100}{1+10^{x(pH-pK_a)}} \tag{4.1}$$

onde x = -1 se o composto é acido ou 1 se for básico.

Os maiores efeitos da ionização são refletidos sobre a solubilidade dos compostos e a permeabilidade através das membranas celulares. Os efeitos da ionização são opostos nestas propriedades. Moléculas ionizadas são mais solúveis e menos permeáveis, pois a forma neutra é considerada predominante durante a permeação por difusão passiva (KERNS e DI, 2008).

Entretanto, se a molécula interage com um sítio ionizado do receptor, a forma ionizada da molécula favorecerá a interação. Por outro lado, a permeabilidade através das membranas celulares é desfavorecida. Dependendo do pH do meio e do pK_a da molécula, haverá um equilíbrio entre a forma ionizada e a neutra que permitirá a permeação através das membranas até alcançar o receptor onde a forma ionizada será beneficiada. Este equilíbrio poderia ocorrer indefinidamente até que todas as moléculas fossem absorvidas, porém as moléculas sofrem metabolismo e excreção antes de cruzar as membranas e alcançarem o receptor (SILVERMAN, 2004).

Dependendo do sistema utilizado para obtenção dos dados biológicos (enzimático, celular, *in vivo*, etc.), o estado de ionização das moléculas interfere nos valores obtidos para a atividade biológica. Em estudos CoMFA (*Comparative Molecular Field Analysis*), por exemplo, recomenda-se que a atividade dos compostos seja corrigida pelo pK_a nos casos onde a ionização influência a

atividade biológica ou é importante para ela a fim de evitar superestimação (ou subestimação) dos valores de atividade (FOLKERS, MERZ e ROGNAN, 2000).

A série de moléculas em estudo apresenta grupos ionizáveis conforme mostrado na Tabela 4-2. Os valores de pK_a foram calculados com o programa MARVIN e comparados com os valores encontrados na literatura. Apenas a previsão feita para amorolfina não produziu resultados satisfatórios.

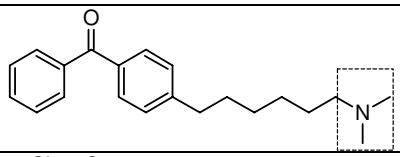
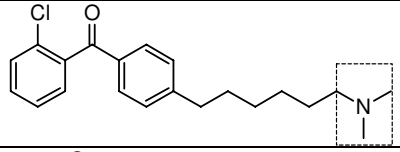
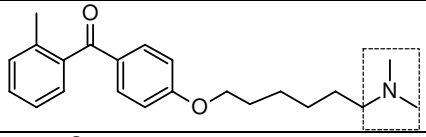
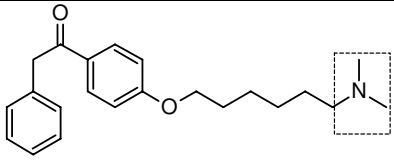
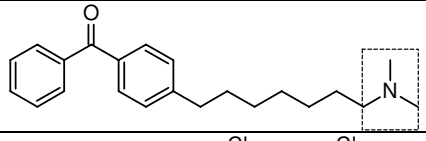
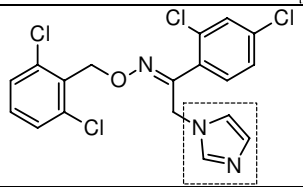
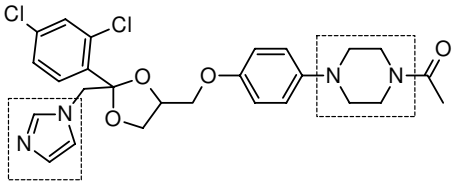
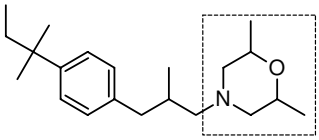
O pH do meio utilizado para os ensaios *in vitro* conduzidos por Gebre-Hiwot e Frommel (1993) para estas moléculas não foi explicitamente mencionado. Como os experimentos foram realizados com células humanas THP-1 assume-se como premissa que o pH do meio seja fisiológico, ou seja, em torno de 7.4. Portanto, nestas condições os compostos enumerados de 1 a 5 na Tabela 4-2, estão predominantemente ionizados de acordo com a Eq. 4.1. Os demais compostos se apresentam na forma neutra.

Resta analisar se a ionização destes compostos é importante para a interação com os respectivos receptores. As classes de compostos em estudo são fungicidas que apresentam o mesmo modo de ação: a inibição da síntese do ergosterol que é o principal esterol encontrado nas membranas dos fungos. O mecanismo de ação destes compostos é distinto e ocorre pela inibição de enzimas que atuam em etapas consecutivas na biossíntese do ergosterol conforme ilustrado na Figura 4-1.

As estruturas tridimensionais destas enzimas não foram elucidadas para *Leishmania donovani*. Entretanto, as discussões sobre as interações com as respectivas enzimas são válidas para todos os organismos eucariontes.

A enzima 2,3-oxidoesqualeno ciclase converte epóxido esqualeno em lanosterol através de uma reação de ciclização que envolve a formação de carbocátions na porção final do esqueleto esteroidal. O sítio ativo desta enzima acomoda tanto a porção lipofílica do esqualeno quanto o carbocátion intermediário.

Tabela 4-2: Série de antifúngicos com ação leishmanicida *in vitro*. Os fragmentos destacados nas moléculas correspondem a partes ionizáveis. Os valores de pK_a foram calculados teoricamente e comparados com valores experimentais da literatura.

Nº	Estrutura	pK_a		
		Calc.	Exp.	Referência
1		9,8	9,8	HALL, 1957
2		9,8	9,8	HALL, 1957
3		9,8	9,8	HALL, 1957
4		9,8	9,8	HALL, 1957
5		9,8	9,8	HALL, 1957
6		6,7	-	Coerente com a faixa relatada por BEGGS, 1991
7		Piperazina 4,0	3,0	BEGGS, 1991
		Imidazol 6,7	6,5	BEGGS, 1991
8		8,5	6,6	DRUG INFORMATION SYSTEM

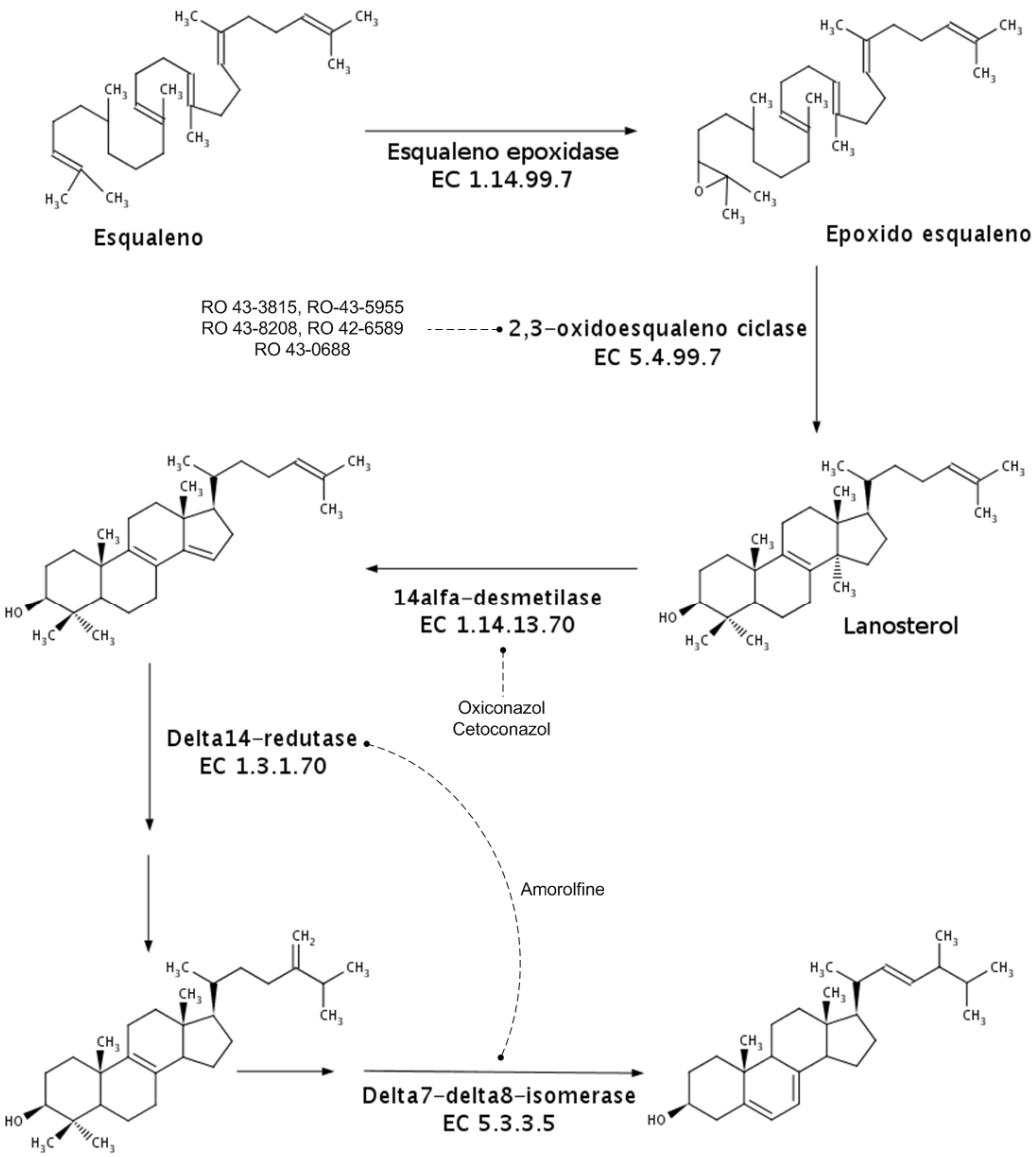


Figura 4-1: Esquema simplificado da biossíntese do ergosterol com as enzimas e os respectivos inibidores em estudo no presente trabalho. Os números EC se referem a classificação das enzimas conforme sistema internacional. Extraído e adaptado de THOMAS, 2003.

A região do sítio onde ocorre a protonação do esqualeno é formada por resíduos de ácido aspártico que são responsáveis por estabilizar o complexo devido à carga negativa. Inibidores desta enzima mimetizam estas interações com a presença de

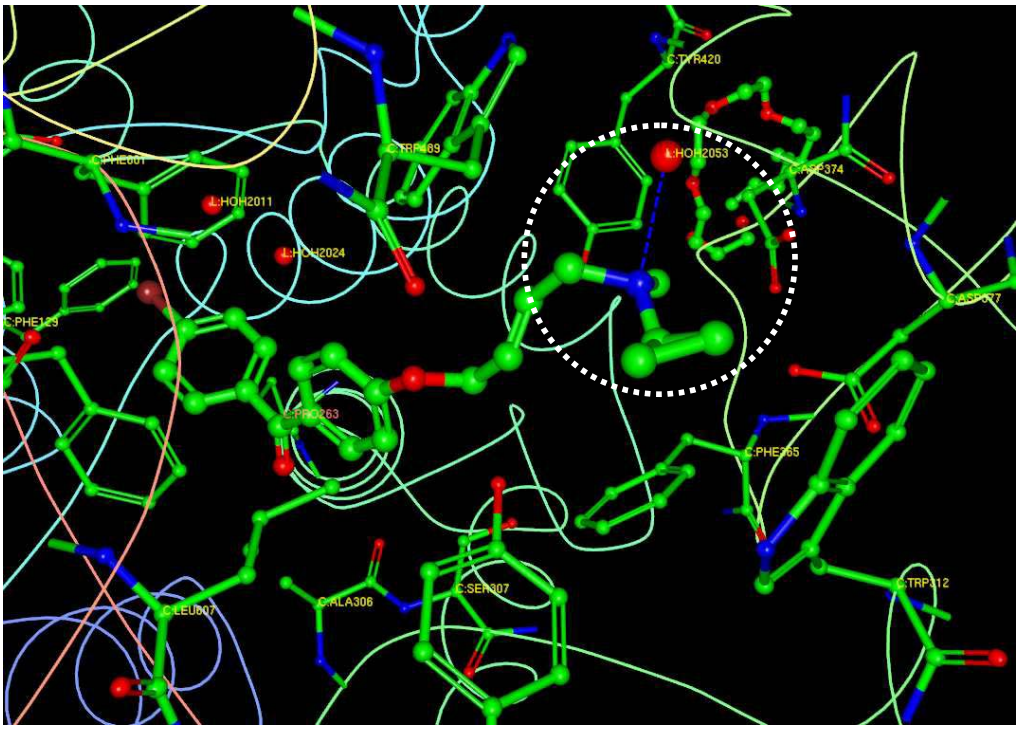


Figura 4-2: Sítio ativo da enzima esqualeno-hopeno ciclase co-cristalizada com um inibidor da enzima oxidoesqualeno humana (código de acesso no Protein Data Bank 1H36). A esqualeno-hopeno ciclase apresenta arquitetura do sítio ativo similar à oxidoesqualeno ciclase. A amina terciária (destacada no círculo tracejado) interage na forma ionizada com uma molécula de água que coordena a interação com os resíduos ASP374, ASP376, ASP377.

grupos ionizáveis em suas estruturas, tais como aminas terciárias, cuja estabilização com os resíduos de ácido aspártico ocorre com o auxílio de uma molécula de água (LENHART et al., 2003) conforme mostrado na Figura 4-2. Portanto, as aminas terciárias presentes nos compostos 1 a 5 (ver Tabela 4-2) simulam o estado de transição do complexo enzima-substrato quando estão na forma ionizada. Logo, esta é a forma favorecida na interação destes compostos com o receptor.

A enzima lanosterol 14α-desmetilase é uma enzima pertencente à superfamília de citocromos P450 e também é conhecida por CYP51. Imidazóis como cetoconazol e oxiconazol bem como triazóis tais como itraconazol e fluconazol são exemplos de inibidores desta enzima. A inibição da atividade desta enzima por estas classes de compostos é atribuída à coordenação de um dos átomos

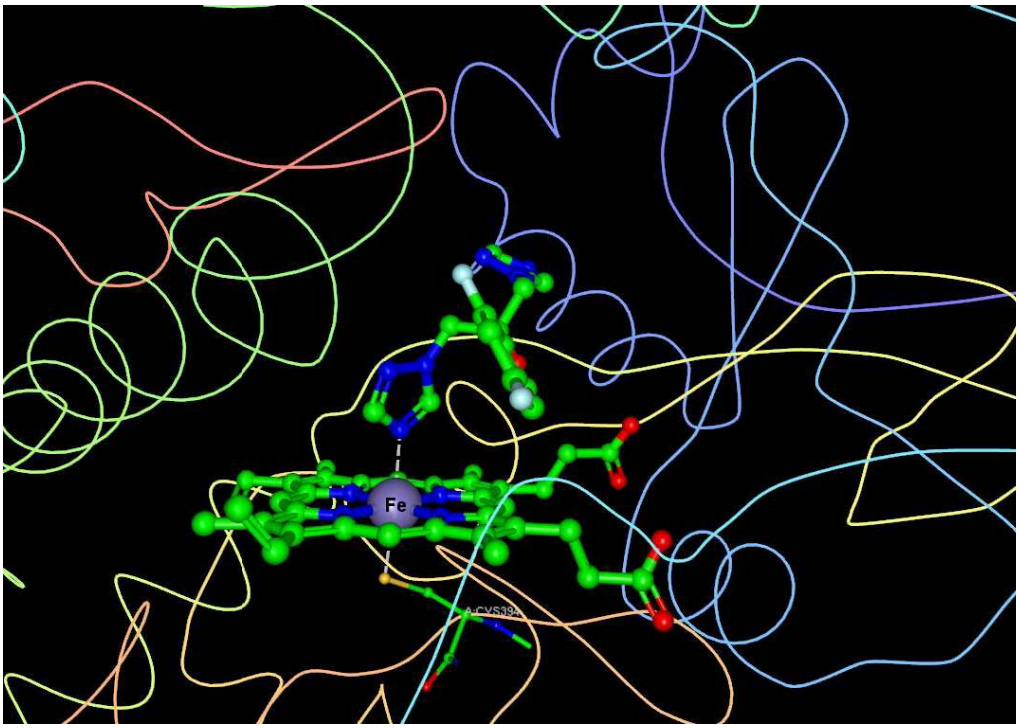


Figura 4-3: Sítio ativo da enzima lanosterol 14α-desmetilase da Mycobacterium tuberculosis co-cristalizada com o inibidor fluconazol (código de acesso no Protein Data 1EA1). O anel triazol interage na forma neutra com o átomo de Ferro do grupo HEME.

de nitrogênio do anel azol com o átomo de ferro do grupo heme como pode ser observado na Figura 4-3. A interação com esta enzima só é possível na forma não ionizada.

As enzimas Δ-14-esterol redutase e Δ-7-Δ-8-isomerase não possuem estrutura tridimensional elucidada até o presente momento. Portanto, não se pode afirmar preferência da amorolfina pelo estado ionizado. Como esta molécula se apresenta na forma neutra em pH fisiológico, este será o estado adotado.

4.3 Análise conformacional

A análise conformacional foi realizada de duas maneiras distintas com o programa OMEGA. Na primeira abordagem, todos os compostos foram considerados na forma neutra. Na segunda abordagem, a amina terciária dos

compostos enumerados de 1 a 5 na Tabela 4-2 foi considerada na forma protonada que é a forma preferencial de interação com a enzima alvo. Os demais compostos foram mantidos na forma neutra. A janela de energia considerada durante as análises foi de 13 kcal/mol em relação mínimo global encontrado. Todas as conformações foram otimizadas com o campo de força MMFF corrigido com o termo de solvatação de Sheffield. O critério de corte estabelecido para eliminação de conformações duplicadas foi ajustado de uma abordagem para outra em função do número excessivo de conformações geradas devido à flexibilidade das moléculas. Na primeira abordagem, foi utilizado um RMSD ≤ 0,8Å e que resultou em 23062 conformações. Na segunda abordagem, este critério foi elevado para RMSD ≤ 1,6Å e o número de conformações geradas foram de 2232.

O passo seguinte foi a reordenação por energia sem otimização de geometria de acordo com os valores obtidos por meio de cálculos semi-empíricos *single point* com hamiltoniano PM3 tanto no vácuo quanto em meio aquoso através do modelo para solvente SM5.42R. Todos os cálculos foram realizados com o programa AMSOL 7.1.

Os resultados obtidos para a primeira abordagem da análise conformacional serão utilizados em momento oportuno com o propósito de avaliar a influência da ionização sobre os descritores utilizados na proposição do modelo. Por ora, os resultados apresentados estarão voltados para a segunda abordagem.

A comparação entre as conformações mais estáveis obtidas no vácuo e em solução aquosa encontra-se na Tabela 4-3. Nota-se o mesmo perfil encontrado para a série dos nucleosídeos, ou seja, as moléculas mais estáveis no vácuo não são necessariamente as mais estáveis em solução. As conformações mais favorecidas foram aquelas com maiores valores para o momento dipolar.

Tabela 4-3: Comparação entre as conformações mais estáveis obtidas em solução e no vácuo para a série dos antifúngicos. Os valores de energia foram calculados utilizando a forma ionizada para as moléculas de 1 a 5 e a forma neutra para as demais.

Nº	Molécula	Conformação N°		$\Delta H_{Solução}$ (kcal/mol)	$\Delta H_{Vácuo}$ (kcal/mol)	RMSD (Å) Sol.Aq./Vac.
		Solução Aquosa	Vácuo			
1	RO 43-3815	6	222	87,58	148,62	3,24
2	RO 43-5955	2	207	83,10	148,11	1,65
3	RO 43-8208	26	111	50,63	116,65	1,60
4	RO 42-6589	111	58	50,74	119,37	1,71
5	RO 43-0688	3	268	82,88	146,30	2,46
6	Oxiconazol	149	20	85,89	95,09	2,90
7	Cetoconazol	65	1	-58,57	-36,99	2,27
8	Amorolfina	41	2	-63,02	-59,92	1,88

4.4 Cálculo dos descritores moleculares

Tomando-se como base a experiência obtida com a série dos nucleosídeos, foram considerados na modelagem dos dados descritores eletrônicos provenientes dos cálculos quânticos e descritores correlacionados com propriedades farmacocinéticas.

Ao todo foram calculados 14 descritores, a saber: o calor de formação (ΔH_f), a energia eletrônica (*EE*), a energia de repulsão cerne-cerne (*CCR*), a energia do HOMO (E_{homo}), a energia do LUMO (E_{lumo}), a diferença de energia entre o LUMO e o HOMO (*GAP(Lumo,Homo)*), a diferença de energia entre o HOMO e o HOMO-1, (*GAP(Homo,Homo-1)*), o momento dipolar total (*Dip*), a energia de solvatação (ΔG_{water}), o peso molecular (*MW*), o coeficiente de distribuição no *pH* 7,4 ($logD_{7.4}$), a refratividade molar (*MR*), o número total de aceptores e doadores de pontes de hidrogênio (*H-Bond*) e a área da superfície polar (*TPSA*).

Os nove primeiros descritores foram provenientes dos cálculos quânticos mencionados na seção 4.2 considerando os efeitos de solvatação. Os descritores

MW, *MR* e *H-Bond* foram calculados com o programa DRAGON ao passo que $logD_{7.4}$ e *TPSA* foram calculados com os programas MARVIN e MOLINSPIRATION respectivamente.

Os últimos cinco descritores citados acima estão relacionados com o tamanho, com a forma molecular, com a capacidade de estabelecer pontes de hidrogênio, com a lipofilicidade e com a área da superfície polar. Estes parâmetros são comumente utilizados para prever absorção de moléculas (ERTL et al., 2000). A refratividade molar é um descritor abrangente, pois está correlacionado com a lipofilicidade, com o volume molar, volume estérico e a polarizabilidade. Quanto maior a parte polar de uma molécula maior é o valor de *MR* (KUBINYI, 1993). O $logD_{7.4}$ é uma correção aplicada ao coeficiente de partição octanol-água, *logP*, que leva em consideração a ionização das moléculas. Este descritor é definido por:

$$\log D = \log P - \log(1 + 10^{x(pH - pK_a)}) \tag{4.2}$$

onde x=-1 se o composto é acido ou 1 se for básico.

Após o cálculo do coeficiente de correlação de Pearson para este conjunto de variáveis, verificou-se que muitos pares de variáveis estão correlacionados entre si apresentando correlação acima de 0,6 (ver Tabela 4-4).

4.5 Seleção dos descritores

Como o composto cetoconazol se mostrou inativo e o valor de atividade para ED_{50} foi obtido por extrapolação, este dado foi removido do conjunto de dados durante o procedimento de seleção de variáveis e proposição do modelo.

O processo de seleção de variáveis utilizou um método supervisionado baseado em envoltório onde o algoritmo de aprendizado foi a própria função de regressão linear. As variáveis foram avaliadas através de busca sistemática por

Tabela 4-4: Matriz de correlação para os descritores calculados. Os descritores com coeficiente de correlação superior a 0,6 foram destacados com fundo preto.

	ΔHf	EE	CCR	Ehomo	Elumo	GAP(Lumo,Homo)	GAP(Homo,Homo-1)	Dip	ΔGwater	TPSA	logD7.4	MW	MR	H-bond
ΔHf	1	0,862	0,894	0,909	0,977	0,715	0,772	0,962	0,932	0,536	0,139	0,305	0,47	0,209
EE	0,862	1	0,997	0,952	0,839	0,314	0,902	0,705	0,651	0,248	0,052	0,18	0,043	0,018
CCR	0,894	0,997	1	0,956	0,865	0,371	0,889	0,752	0,701	0,288	0,016	0,193	0,098	0,008
Ehomo	0,909	0,952	0,956	1	0,924	0,431	0,961	0,772	0,711	0,383	0,154	0,326	0,254	0,118
Elumo	0,977	0,839	0,865	0,924	1	0,744	0,808	0,918	0,88	0,496	0,05	0,313	0,455	0,181
GAP(Lumo,Homo)	0,715	0,314	0,371	0,431	0,744	1	0,226	0,815	0,834	0,5	0,386	0,169	0,629	0,219
GAP(Homo,Homo-1)	0,772	0,902	0,889	0,961	0,808	0,226	1	0,597	0,534	0,313	0,295	0,319	0,113	0,114
Dip	0,962	0,705	0,752	0,772	0,918	0,815	0,597	1	0,991	0,639	0,269	0,323	0,63	0,312
ΔGwater	0,932	0,651	0,701	0,711	0,88	0,834	0,534	0,991	1	0,657	0,357	0,276	0,609	0,333
TPSA	0,536	0,248	0,288	0,383	0,496	0,5	0,313	0,639	0,657	1	0,174	0,81	0,717	0,929
logD7.4	0,139	0,052	0,016	0,154	0,05	0,386	0,295	0,269	0,357	0,174	1	0,688	0,046	0,38
MW	0,305	0,18	0,193	0,326	0,313	0,169	0,319	0,323	0,276	0,81	0,688	1	0,614	0,873
MR	0,47	0,043	0,098	0,254	0,455	0,629	0,113	0,63	0,609	0,717	0,046	0,614	1	0,601
H-bond	0,209	0,018	0,008	0,118	0,181	0,219	0,114	0,312	0,333	0,929	0,38	0,873	0,601	1

modelos de uma e duas variáveis visto que o conjunto de trabalho é composto por apenas sete moléculas. O programa utilizado nesta etapa foi o BUILDQSAR.

4.6 Proposição do modelo

O melhor modelo obtido correlaciona a atividade leishmanicida dos antifúngicos estudados com os descritores energia eletrônica e a área da superfície polar. O modelo obtido é descrito por:

$$\log(\frac{1}{ED_{50}}) = 0{,}0001(\pm 0{,}0000)[EE] - 0{,}0285(\pm 0{,}0111)[TPSA] + 3{,}3448 \quad (4.3)$$

$$n = 7; \quad r = 0{,}976; \quad s = 0{,}084; \quad F = 40{,}83; \quad Q^2 = 0{,}866; \quad SPRESS = 0{,}142$$

onde n é o número de compostos, r é o coeficiente de correlação, F é o coeficiente de Fisher, Q^2 é coeficiente de correlação da validação cruzada e $SPRESS$ é a soma quadrática dos desvios de previsão.

Os índices de confiabilidade são bons e há uma boa correlação (~98%) entre os valores calculados e observados para $\log(1/ED_{50})$ (ver Figura 4-4). A Tabela 4-5 mostra os valores numéricos para os descritores EE e TPSA bem como os resíduos do modelo. A avaliação do ajuste do modelo também pode ser acompanhada pela Tabela 4-6 que apresenta os dados de análise de variância.

O modelo prevê aumento da atividade leishmanicida com o aumento do valor da energia eletrônica e com a redução da área da superfície polar dos compostos. A energia eletrônica, por sua vez, aumenta negativamente com o número de elétrons. Estas observações são consistentes com os dados obtidos para esta série de compostos onde foi encontrada uma correlação negativa de 98,6% entre a energia eletrônica e o número de elétrons (ver Figura 4-5).

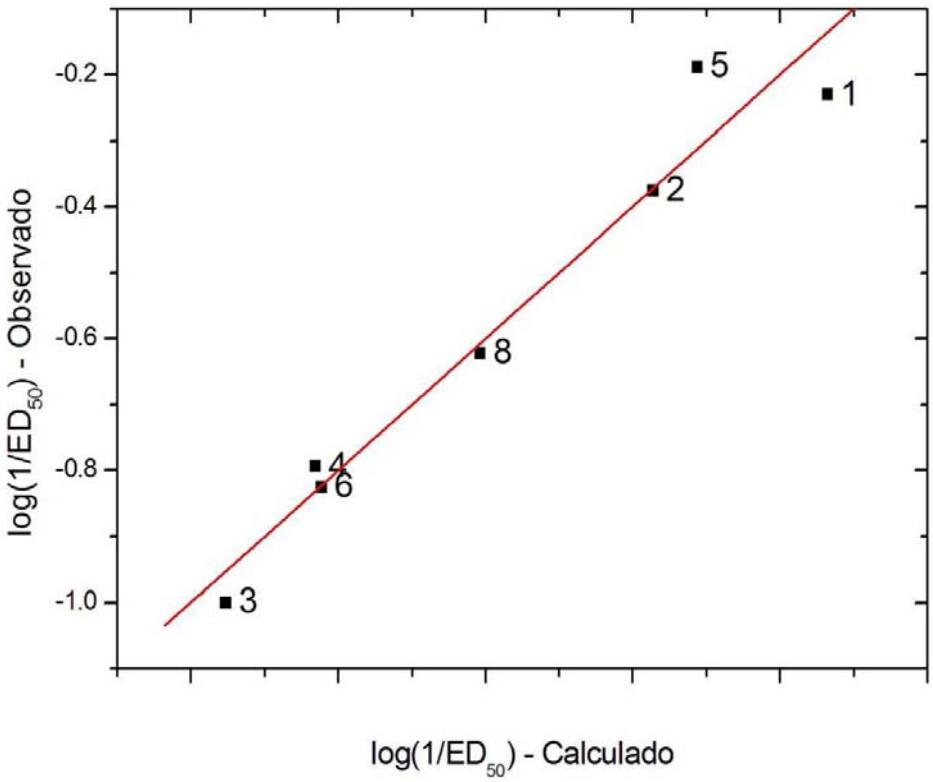


Figura 4-4: Correlação entre os valores observados e previstos pelo modelo para log(1/*ED50*).

Tabela 4-5: Valores observados e calculados para log(1/ED$_{50}$) e valores residuais utilizando a Eq. 4.3. As últimas duas colunas correspondem aos valores numéricos calculados para os descritores *EE* e *TPSA*.

N°	Molécula	Atividade (log 1/*ED$_{50}$*)		Resíduos	*EE*	*TPSA*
		Observada	Prevista		(eV)	(Å^2)
1	RO 43-3815	-0,230	-0,136	-0,094	-25237	22
2	RO 43-5955	-0,375	-0,372	-0,003	-27308	22
3	RO 43-8208	-1,000	-0,953	-0,047	-30110	31
4	RO 42-6589	-0,794	-0,831	0,037	-29036	31
5	RO 43-0688	-0,188	-0,312	0,124	-26782	22
6	Oxiconazol	-0,825	-0,823	-0,003	-26782	39
8	Amorolfina	-0,622	-0,607	-0,015	-28163	12

Tabela 4-6: Tabela ANOVA para o modelo de regressão linear múltipla representado pela Eq. 4.3.

Fonte de Variação	N° Graus de Liberdade	Soma dos Quadrados	Média da Soma dos Quadrados	F	P
Modelo	2	0,574	0,287	40,83	0,002
Resíduo	4	0,028	0,007		
Total	6	0,602			

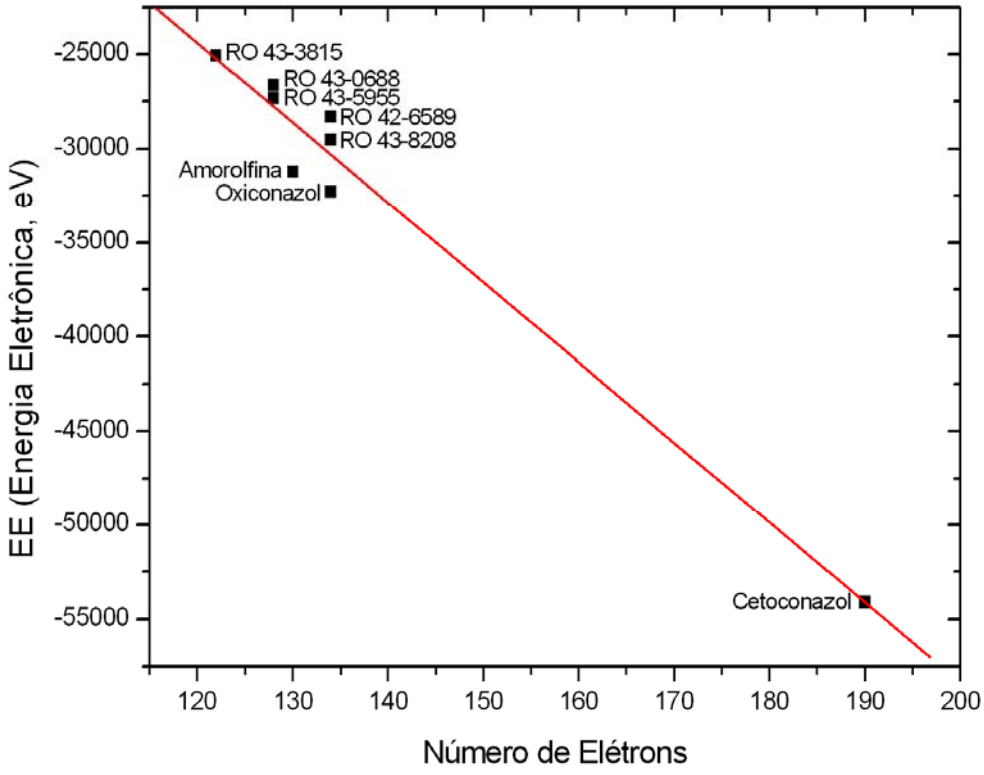


Figura 4-5: Correlação entre a energia eletrônica e o número de elétrons.

A tendência expressa pelo modelo através destes dois descritores explica características que são comuns no que diz respeito ao mecanismo de reconhecimento molecular das enzimas 2,3-oxidoesqualeno ciclase e lanosterol 14α-desmetilase, pois os substratos naturais bem como os inibidores destas enzimas são ancorados no sítio catalítico através de interações lipofílicas principalmente

com resíduos aromáticos tais como fenilalanina (Phe), tirosina (Tyr) e triptofano (Trp) cujas cadeias laterais são ricas em elétrons (ver Figura 4-6 e Figura 4-7). Portanto, compostos com maior afinidade são aqueles que apresentam certa deficiência eletrônica (GOLDMAN et al., 1996; RUGE et al., 2005). A redução na atividade em função do aumento da área da superfície polar dos compostos também corrobora com esta conclusão, pois o aumento deste parâmetro desfavorece as interações hidrofóbicas. Além disso, a área da superfície polar também inclui efeitos de absorção *in vivo*. Quanto maior o valor deste descritor menor é a probabilidade de absorção (ERTL et al., 2000).

É provável que estas observações possam ser ampliadas para outras enzimas da via biossintética do ergosterol tais como Δ-14-esterol redutase e Δ-7-Δ-8-isomerase visto que os substratos destas enzimas são extremamente similares entre si (ver Figura 4-1) e, portanto, o mesmo princípio de reconhecimento molecular seria empregado por todas.

Assim, o modelo expresso pela Eq. 4.3 trata dos requerimentos gerais para o reconhecimento de ligantes que atuam como inibidores da síntese do ergosterol, também conhecidos por EBIs (*Ergosterol Biosynthesis Inhibitors*).

Finalmente, conforme mencionado na seção 4.3, a análise conformacional e o cálculo dos descritores foram realizados tanto para a forma ionizada quanto para a forma neutra dos compostos 1 a 5. É possível avaliar a influência da ionização sobre os descritores *EE* e *TPSA* utilizados na proposição do modelo comparando-se os valores obtidos em ambas as abordagens. Como pode ser observado na Tabela 4-7, a protonação da amina terciária destes compostos não alterou significativamente o valor destes descritores e a Eq. 4.3 permanece válida também para forma neutra dos compostos.

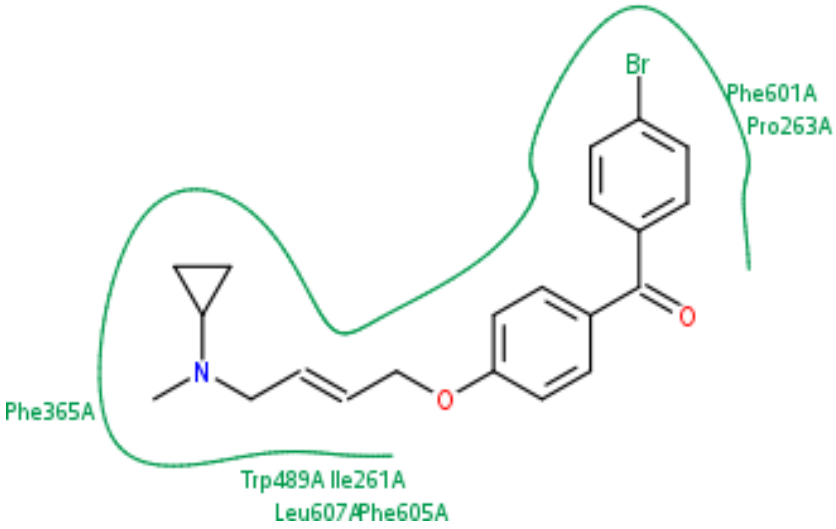


Figura 4-6: Modelo esquemático dos resíduos do sítio ativo envolvidos no ancoramento de um inibidor da enzima da oxidoesqualeno ciclase humana (código de acesso no Protein Data Bank 1H36). O diagrama foi feito com o software POSEVIEWWEB.

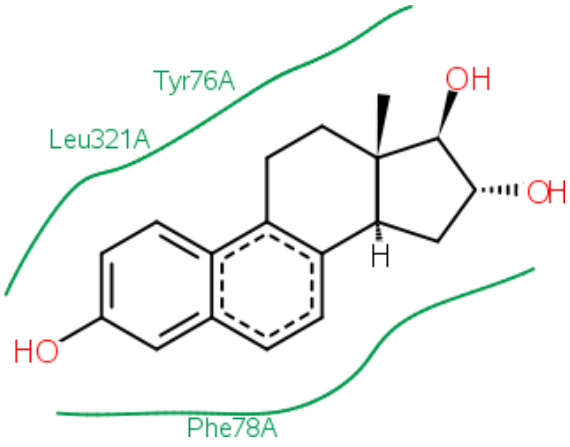


Figura 4-7: Modelo esquemático dos resíduos envolvidos na interação do estriol com a enzima lanosterol 14α-desmetilase (código de acesso no Protein Data Bank 1X8V). O diagrama foi feito com o software POSEVIEWWEB.

Tabela 4-7: Valores para os descritores EE e TPSA calculados para os compostos considerando a forma ionizada e neutra.

N°	Composto	Forma Ionizada		Forma Neutra	
		EE (eV)	TPSA ($Å^2$)	EE (eV)	TPSA ($Å^2$)
1	RO 43-3815	-25237	22	-25053	20
2	RO 43-5955	-27308	22	-27310	20
3	RO 43-8208	-30110	31	-29521	30
4	RO 42-6589	-29036	31	-28308	30
5	RO 43-0688	-26782	22	-26621	20

4.7 Validação do modelo

O conjunto de teste utilizado na validação do modelo é formado pelo cetoconazol que foi inicialmente excluído do conjunto de dados e pelos antifúngicos mostrados na Figura 4-8 para os quais se conhece a ação contra leishmaniose visceral. Colakoglu et al (2006 a e b) relataram casos de cura com uso combinado de fluconazol e alopurinol. Itraconazol e posaconazol não são efetivos no tratamento da leishmaniose visceral de acordo com os trabalhos de El-Hassan (2001) e Al-Abdely e colaboradores respectivamente.

Adicionalmente, outros antifúngicos (ver Figura 4-9) foram avaliados com o intuito de efetuar previsões de atividade leishmanicida. Os compostos do conjunto de teste e de previsão são inibidores da enzima 14α-desmetilase e, portanto, interagem na forma neutra.

Os mesmos procedimentos mencionados anteriormente para análise conformacional e geração de descritores foram adotados para estes compostos. A análise do domínio QSAR do modelo proposto foi realizada utilizando todos os compostos do conjunto de trabalho com o programa AMBIT DISCOVERY usando

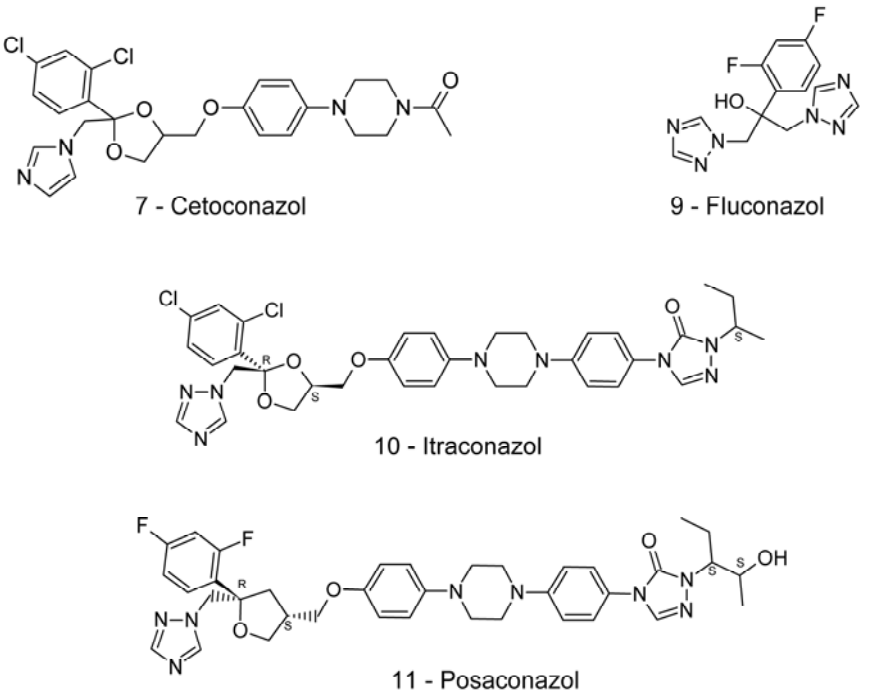


Figura 4-8: Conjunto de teste para a série dos antifúngicos.

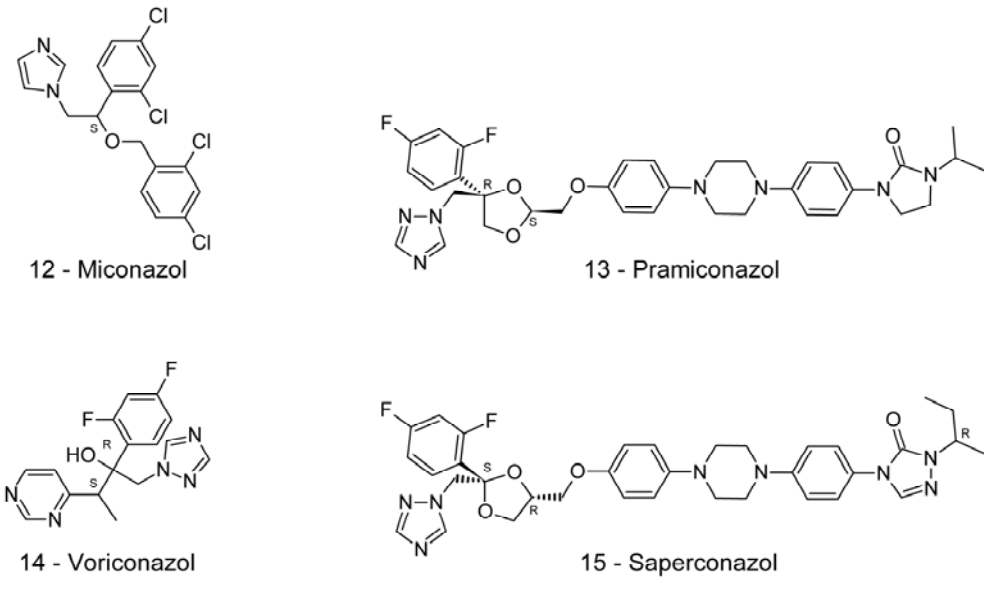


Figura 4-9: Conjunto de antifúngicos utilizados para previsão.

distância euclidiana no espaço descritor e análise de componentes principais (PCA) como estágio de pré-processamento. O composto 12 (miconazol) encontra-se no limiar do domínio estabelecido para interpolação. Os demais compostos estão fora do domínio do modelo visto que os valores de *EE* estão abaixo do limite inferior estabelecido e os valores de TPSA são elevados (ver Figura 4-10). Portanto, é difícil assegurar os valores de previsão para estes compostos. A situação ideal seria testá-los experimentalmente *in vitro* a fim de obter os respectivos valores de ED_{50} e incluí-los no modelo atual para ampliar a região de interpolação.

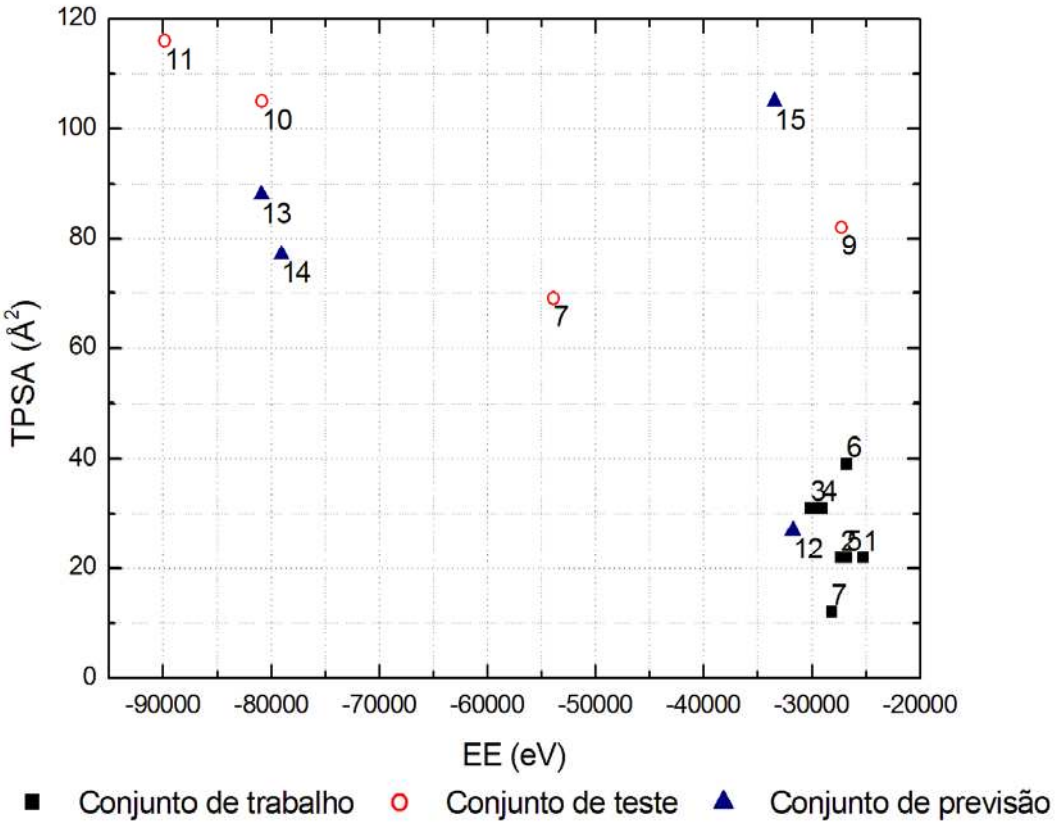


Figura 4-10: Análise do domínio de aplicabilidade do modelo QSAR para o conjunto de teste e para o conjunto de previsão.

Tabela 4-8: Valores calculados de ED_{50} para o conjunto de teste. Moléculas estimadas como inativas foram representadas com ED_{50} > 300μM.

Nº	Composto	ED_{50} estimado (μM)	*EE* (eV)	*TPSA* (Å²)	Referência
7	Cetoconazol	> 300	-53865	69	HIWOT e FROMMEL, 1993
9	Fluconazol	51	-27222	82	COLAKOGLU, 2006 a e b
10	Itraconazol	> 300	-80860	105	EL-HASSAN, 2001
11	Posaconazol	> 300	-89824	116	AL-ABDELY, 1999

A Tabela 4-8 mostra os valores previstos para ED_{50} para o cetoconazol e os compostos do conjunto de teste. Embora os compostos não estejam dentro dos limites de previsão do modelo, observam-se previsões coerentes para o cetoconazol e todos os compostos do conjunto de teste.

Finalmente, de acordo com os resultados apresentados na Tabela 4-9, o miconazol seria o único antifúngico com potencial leishmanicida dentro do conjunto de previsão estudado. Para confirmar estas previsões são necessários estudos experimentais.

Tabela 4-9: Valores calculados de ED_{50} para o conjunto de previsão. Moléculas estimadas como inativas foram representadas com ED_{50} > 300μM.

Nº	Composto	ED_{50} estimado (μM)	*EE* (eV)	*TPSA* (Å²)
12	Miconazol	4	-31734	27
13	Pramiconazol	> 300	-80899	88
14	Saperconazol	> 300	-79074	77
15	Voriconazol	> 300	-33374	105

Capítulo 5

Conclusões gerais e sugestão de novos compostos

Introdução

Este capítulo apresenta as conclusões obtidas para os estudos QSAR aplicados para as séries de compostos nucleosídeos e antifúngicos.

Novos compostos são sugeridos abrindo a possibilidade de novos projetos de pesquisa a partir deste trabalho.

5.1 Conclusões gerais

O principal objetivo deste trabalho era disponibilizar modelos QSAR para triagem *in silico* de novos compostos com ação leishmanicida. Os modelos propostos para ambas as séries se enquadram na categoria de modelos QSAR globais. Ambas as séries são compostas por moléculas muito flexíveis e requereram um tratamento diferenciado durante a análise conformacional através de um sistema baseado em regras.

Os dados de atividade biológica para a série dos nucleosídeos foram expressos em termos de percentagem de inibição do crescimento de formas amastigotas de *Leishmania donovani* em uma dose fixa. Experimentos conduzidos desta forma são típicos na prospecção de protótipos. Entretanto, percentagem de inibição não é a forma apropriada para análises QSAR convencionais e, portanto, foi necessário efetuar as análises através de regressão logística. Modelos deste tipo fornecem a probabilidade de que um evento ocorra. No presente trabalho, significa fornecer a probabilidade de que um composto derivado de nucleosídeo apresente ação leishmanicida. O mecanismo de ação para esta série não foi elucidado, portanto, o modelo QSAR resultante foi proposto com base no efeito dos compostos e não tem o propósito de explorar aspectos mecanísticos.

Tanto o modelo de regressão logística quanto a árvore de classificação apontaram os descritores *Mor26v* e o *Gap(Homo,Homo-1)* como relevantes para a explicação da atividade leishmanicida destes compostos. Estes descritores estão associados respectivamente às variações estruturais como quiralidade e eletrônicas em função dos efeitos ocasionados por substituintes doadores e retiradores de elétrons.

O modelo de regressão logística atingiu 90,5% para acurácia de classificação para o conjunto de trabalho sendo que apenas os compostos 2 e 17 não foram

classificados adequadamente. Para o conjunto de teste, a acurácia de classificação foi de 58% após a análise do domínio de aplicabilidade do modelo. A razão para esta queda no desempenho classificatório se deve ao fato de que os compostos 26, 29, 31 e 33 foram classificados incorretamente. A combinação de valores encontrados dos descritores *Mor26v* e *Gap(Homo,Homo-1)* para estes compostos encontra-se na região de transição na classificação de atividade onde a probabilidade sofre variações abruptas. Previsões feitas fora da região de transição tendem a ser mais precisas tanto que o modelo foi capaz de discernir isósteros inativos, como o aciclovir e ganciclovir, do seu análogo ativo composto 22.

O modelo para árvore de classificação alcançou 95% para acurácia de classificação para o conjunto de trabalho e 86% para o conjunto de teste. Neste caso, apenas o composto 17 do conjunto de trabalho e os compostos aciclovir e ganciclovir do conjunto de teste foram classificados incorretamente. As regras geradas pela árvore de classificação são extremamente simples e fáceis de programar computacionalmente durante triagens virtuais de compostos. Portanto, sugere-se o uso inicial desta metodologia para seleção inicial de compostos e posteriormente a aplicação do modelo de regressão logística.

A série dos antifúngicos possui moléculas contendo grupos ionizáveis. O estado de ionização usado para os cálculos dos descritores foi selecionado de acordo com o pK_a no *pH* fisiológico e com estado que é favorecido nas interações com os respectivos receptores. Para esta série foi utilizado um modelo de regressão linear múltipla onde a energia eletrônica e a área da superfície polar dos compostos se mostraram importantes para a atividade leishmanicida da série. Os valores previstos exibem aproximadamente 98% de correlação com os valores experimentais. Os compostos do conjunto de trabalho apresentam o mesmo modo de ação, ou seja, são inibidores da síntese do ergosterol, porém, por mecanismos distintos. O modelo obtido incorpora os requerimentos gerais para o

reconhecimento molecular de enzimas que atuam por esta via bioquímica. Outro aspecto interessante é que o modelo obtido é independente do estado de ionização, isto é, a mesma equação é válida também para a forma neutra dos compostos. Apesar dos compostos do conjunto de teste estar fora do domínio do modelo, os valores de atividade previstos estão de acordo com a literatura. Adicionalmente, o modelo foi utilizado para fazer previsões de atividade leishmanicida para o miconazol, pramiconazol, saperconazol e voriconazol para os quais não há relato, até o presente momento, de aplicação na terapia contra leishmaniose. De acordo com o modelo, o miconazol seria o único antifúngico com potencial leishmanicida.

5.2 Sugestão de novos compostos

Conforme mencionado no Capítulo 1, o desenvolvimento de pró-fármacos é uma abordagem utilizada para contornar barreiras ao desenvolvimento como permeabilidade, solubilidade, metabolismo, estabilidade, transporte e segurança farmacêutica (KERNS e DI, 2008).

A esterificação de fármacos é uma estratégia utilizada para aumentar a permeabilidade da difusão passiva. A conversão dos pró-fármacos para a forma ativa é mediada normalmente por esterases nestes casos.

O composto 16 da série dos nucleosídeos é inativo. Por outro lado, o composto 18, que é obtido a partir da esterificação do composto 16, é ativo (ver Figura 5-1). Uma hipótese para explicar esta diferença de atividade é que o composto 18 se comporta como um pró-farmaco do composto 16. Os dados de atividade leishmanicida foram obtidos *in vivo* e, provavelmente, a inatividade do composto 16 se deve à baixa permeabilidade devido ao caráter mais polar desta molécula levando a uma rápida eliminação do organismo. O composto 18 sob a forma de éster é mais lipofílico e, portanto, torna-se mais biodisponível.

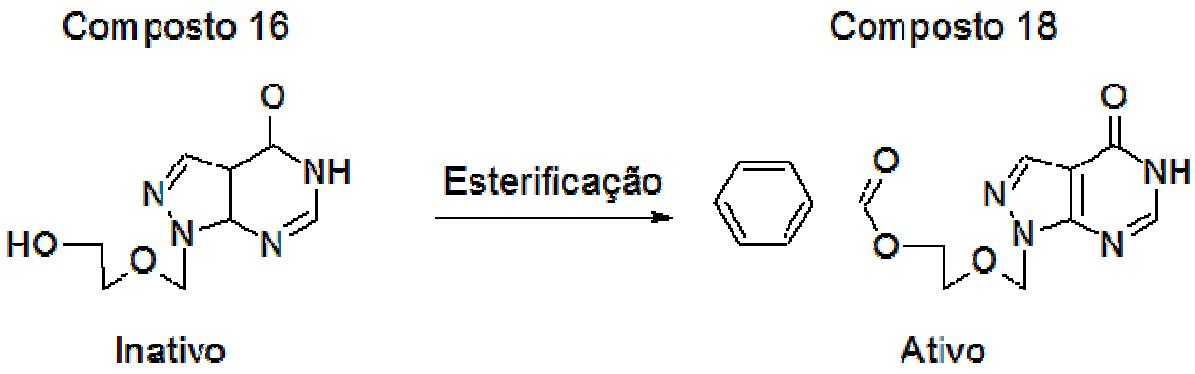


Figura 5-1:Hipótese sobre o perfil de atividade dos compostos 16 e 18. O composto 18 se comporta como um pró-fármaco do composto 16.

Também se pode observar que os pares de compostos 16 e 19 guardam o mesmo conceitual estrutural que o aciclovir e o ganciclovir (ver Figura 5-2). Estes exemplos evidenciam que a abordagem via pró-fármacos ou mimetizando os conceitos aplicados para antivirais conhecidos é um caminho a ser seguido. Estas observações são inéditas para esta série de compostos com ação leishmanicida.

As esterificações comumente aplicadas a fármacos contendo grupos hidroxilas com o objetivo de formar pró-fármacos são (KERNS e DI, 2008; WERMUTH, 2003; CHIN e FERREIRA, 1999):

- Ésteres alquílicos: o balanço entre lipofilicidade/hidrofilicidade pode ser alcançado através da regulação do tamanho e insaturações nas cadeias alquílicas;
- Ésteres utilizados como grupos de proteção para hidroxilas: pivaloato, benzoato, carbamatos, etc;
- Ésteres derivados de aminoácidos;

Propostas de compostos que exemplificam o uso destes conceitos estão mostradas na Figura 5-3.

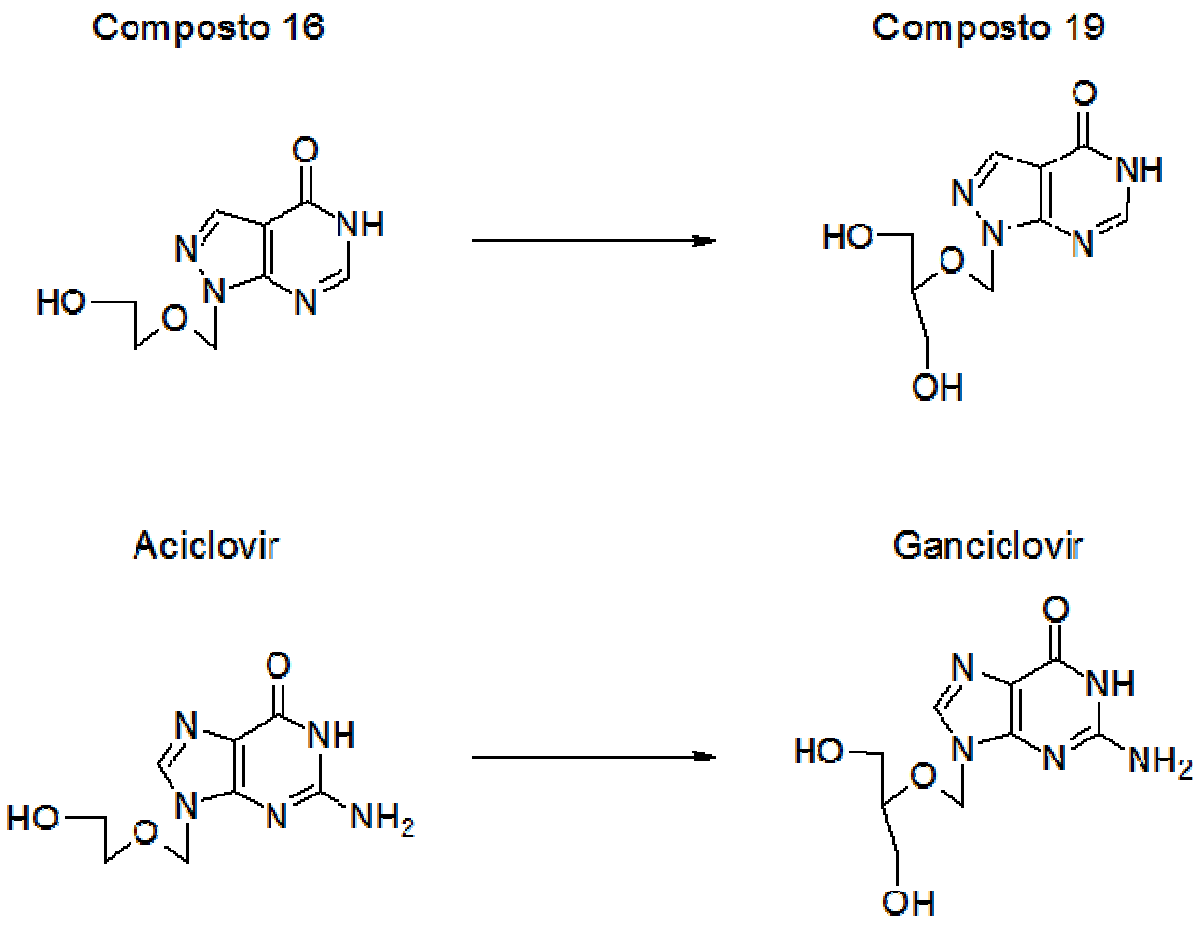


Figura 5-2: Conceitual estrutural dos compostos 16 e 19 é similar ao aciclovir e ganciclovir.

De Clercq e Field (2006) apresentaram uma revisão sobre o espectro de atuação de pró-farmacos de nucleosídeos em termos de atividade biológica. Estes são utilizados no tratamento de cânceres, vírus, bactérias e parasitas e as modificações propostas durante o planejamento racional destes nucleosídeos também podem ser utilizadas como referencial para expandir os compostos da série dos nucleosídeos pirazolo-pirimidínicos do presente trabalho.

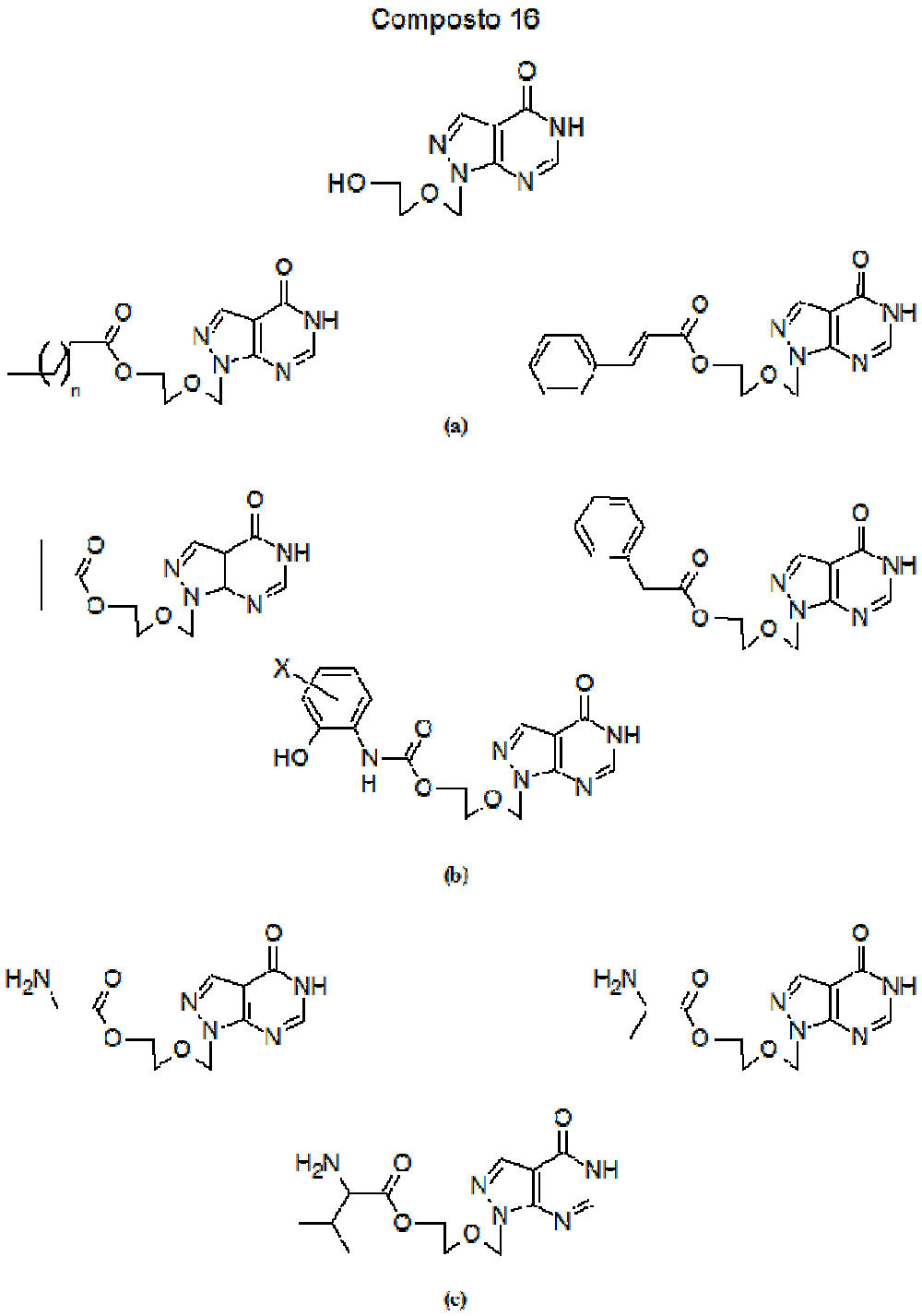


Figura 5-3: Propostas de pró-fármacos a partir da esterificação do composto 16 da série dos nucleosídeos. (a) Exemplos de ésteres alquílicos. (b) Exemplos de ésteres utilizados como grupo proteção. (c) Exemplos de ésteres dos aminoácidos glicina, alanina e valina

Exemplos de modificações que podem ser aplicadas são:

- Miméticos do penciclovir;
- Miméticos do famciclovir: a exemplo do que ocorre com o famciclovir espera-se a remoção dos dois grupos acetil e a oxidação do nucleosídeo pirazolo-pirimidínico;
- Substituições sistemáticas com flúor no açúcar e no anel pirazolo-pirimidínico.

A Figura 5-4 apresenta exemplos de modificações que podem ser aplicadas ao composto 19. A Figura 5-5 apresenta exemplos de substituições sistemáticas com flúor incluindo todas as possíveis combinações de enantiômeros.

Ao todo, as proposições de novos protótipos somam pelo menos cinquenta novos compostos a serem explorados do ponto de vista sintético, patentário e em termos de aferição da atividade biológica.

5.3 Terapias multicomponentes com antifúngicos

O desenvolvimento de novos antifúngicos é uma área em constante desenvolvimento. O potencial terapêutico desta classe de compostos no tratamento da Leishmaniose abre novas possibilidades de investigação para avaliar as previsões feitas no presente trabalho como também a expansão do conjunto de trabalho de moléculas a fim de aumentar o domínio de aplicabilidade do modelo.

Atualmente, há linhas de pesquisa que seguem estudando os efeitos e benefícios de terapias multicomponentes (JIA et al., 2009) a exemplo do que ocorre com o coquetel da AIDS. Conforme mencionado no Capítulo 1, existem relatos de combinações bem sucedidas entre antifúngicos e alopurinol no tratamento de Leishmaniose visceral. Estudos futuros sobre combinações sinergísticas entre compostos nucleosídeos e antifúngicos — incluindo também os compostos

estudados e propostos neste trabalho — que potencializem os efeitos terapêuticos possibilitarão novas abordagens para o tratamento da Leishmaniose visceral.

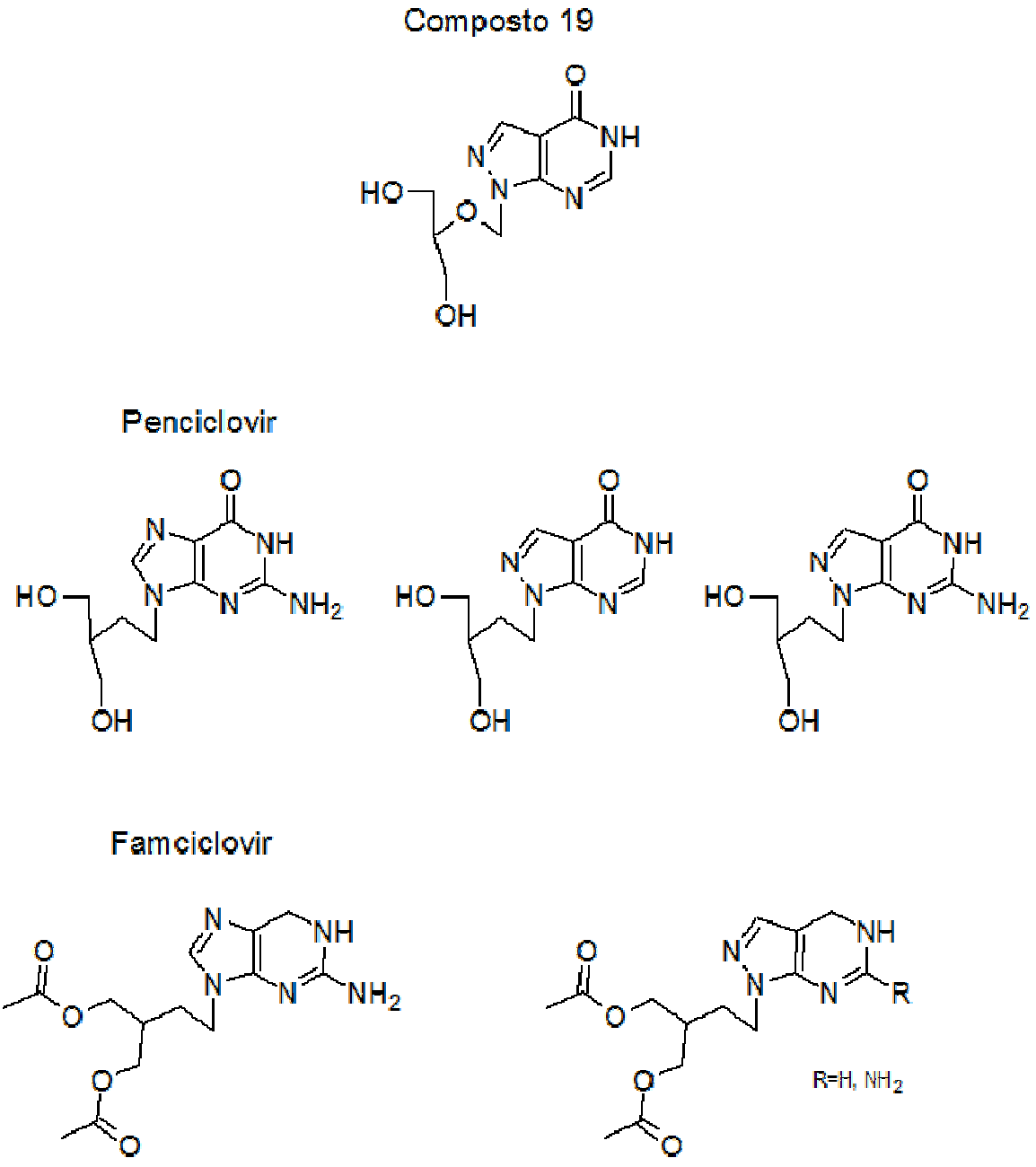


Figura 5-4: Exemplos de modificações que podem ser aplicadas ao composto 19 da série dos nucleosídeos.

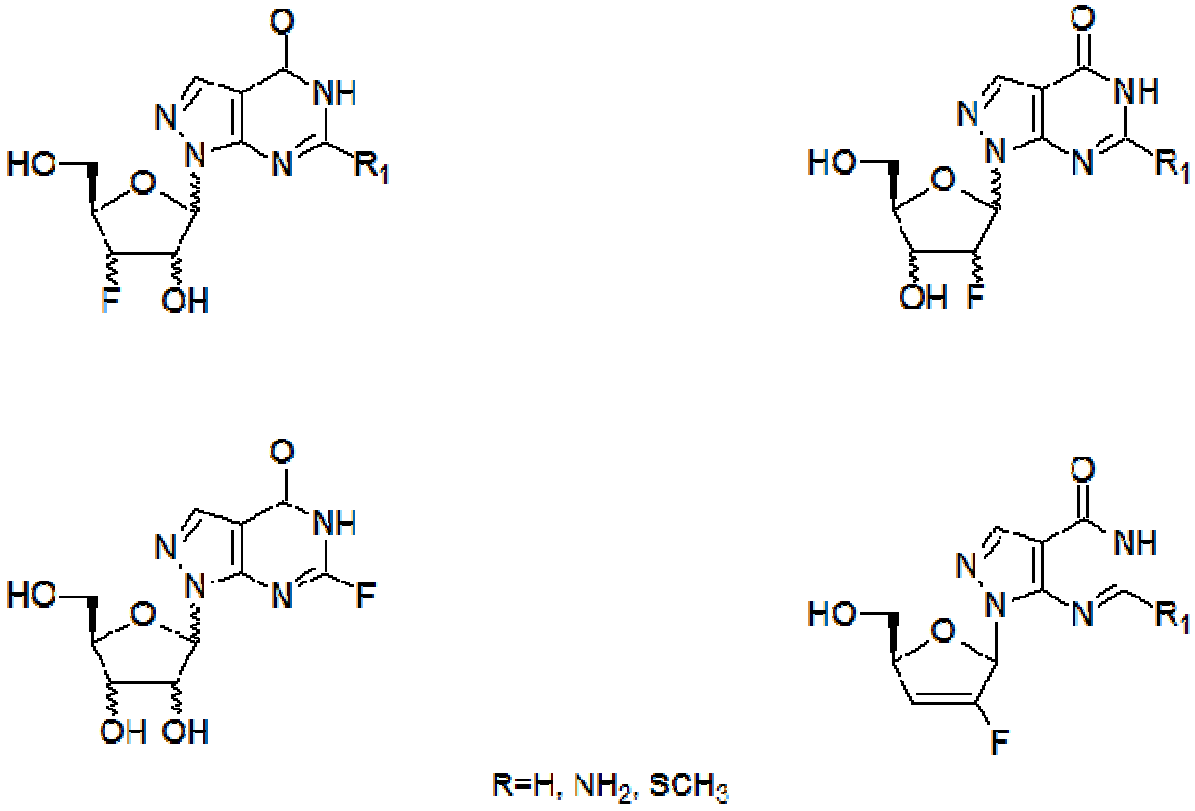


Figura 5-5: Propostas de substituições sistemáticas com flúor para a série dos nucleosídeos incluindo os possíveis enantiômeros

Referências

AL-ABDELY, H. M. et al. Efficacy of the triazole SCH 56592 against Leishmania amazonensis and Leishmania donovani in experimental murine cutaneous and visceral leishmaniasis. *Antimicrob. Agents Chemother.*, v. 43, n. 12, p. 2910-2914, 1999.

ANVISA, Como a Anvisa vê o uso off label de medicamentos: 2005. Seção Registro de Medicamentos. Disponível em: <http://www.anvisa.gov.br/medicamentos/registro/registro_offlabel.htm> Acesso em: 26.out.2007.

AMBIT DISCOVERY, versão 0.04. Disponível em: <http://ambit.acad.bg/>. Acesso em: 15.nov.2007.

AMSOL, versão 7.1; Hawkins, G. D.; Giesen, G. D.; Lynch, G. C.; Chambers, C. C.; Rossi, I..; Storer, J. W.; Li, J.; Zhu, T.; Rinaldi, D. A.; Liotard, D.; Cramer, C. J.; Truhlar, D. G.: University of Minnesota, Minneapolis (2004).

BAGLEY, S. C.; WHITE, H.; GOLOMB, B. A., Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. *J. Clin. Epidemiol.*, v. 54, p. 979-985, 2001.

BALANÃ-FOUCE, R. et al. The pharmacology of leishmaniasis. *Gen. Pharmac.*, v. 30, n. 4, p. 435-443, 1998.

BARONE, P. M. V. S.; CAMILO JR., A.; GALVÃO, D. S. Theoretical approach to identify carcinogenic activity of polycyclic aromatic hydrocarbons. *Phys. Rev. Lett.*,v. 77, p. 1186-1189, 1996.

BASHFORD, D. et al. Generalized Born Models of Macromolecular Solvation Effects. *Annu. Rev. Phys. Chem.*, v. 51, p. 129-152, 2000.

BASSET, D. et al. Visceral leishmaniasis in organ transplant recipients: 11 new cases and a review of the literature. *Microbes and Iinfection*, v. 7, n. 13, p. 1370-1375, 2005.

BAYSAL, C.; MEIROVITCH, H. New conformational search method based on local torsionaldeformations for cyclic molecules, loops in proteins, and dense polymer systems. *J. Chem. Phys.*, v. 105, n. 17, p. 7868-7871, 1996.

BAYSAL, C.; MEIROVITCH, H. Efficiency of the local torsional deformations method for identifying the stable structures of cyclic molecules. *J. Phys. Chem.*, v. 101, p. 2185-2191, 1997.

BECKER, O. M. et al. Flexibility, Conformation Spaces, and Bioactivity. *J. Phys. Chem. B*, v. 104, n. 9, p. 2123-2135, 2000.

BENFENATI, E. Predicting toxicity through computers: a changing world. *Chemistry Central Journal*, v. 1, n. 32, 2007

BEGGS, W. H. Protonation of ketoconazole in relation to fungistatic activity. *Mycopathologia*, v. 116, p. 3-4, 1991.

BERMAN, J. D et al. Activity of purine analogs against *Leishmania donovani* in vivo. *Antimicrob. Agents Chemother.*, v. 31, n. 1, p. 111-113, 1987.

BEUSEN, D. D. et al. Systematic search in conformational analysis. *J. Mol. Struct.*, v. 370, p. 157-171, 1996.

BHAKUNI, D. S. et al. Studies on nucleosides: Part XX – Synthesis and antileishmanial activity of 4,6-substituted pyrazolo[3,4-*d*]pyrimidine nucleosides. *Indian J. of Chem.*, v. 28B, May, p.403-409, 1989.

BHAKUNI, D. S. et al. Studies in nucleosides: Part XXV – Synthesis of alicyclic and cyclic nucleosides of 4(5*H*)-oxopyrazolo[3,4-*d*]pyrimidine and 4-aminopyrazolo[3,4-*d*]pyrimidine and their antileishmanial activity. *Indian J. of Chem.*, v. 29B, January, p. 40-46, 1990.

BIRRINGER, M. S. et al. Structure of a type II thymidine kinase with bound dTTP. *FEBS Lett.*, v. 579, p. 1376-1382, 2005.

BOSTRÖM, J. Reproducing the conformations of protein-bound ligands: A critical evaluation of several popular conformational searching tools. *J. Comp.-Aid. Mol. Design*, v. 15, n.12, p. 1137-1152, 2001.

BOSTRÖM, J; GREENWOOD, J. R.; GOTTFRIES, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graph. Model.*, v. 21, p. 449-462, 2003.

BRAGA, R. S.; VENDRAME, R.; GALVÃO, D. S. Structure-activity relashionship studies of 17 alpha-acetoxyprogesterone hormones. *J .Chem. Inf. Comput. Sci.*, v.40, n. 6, p. 1377-1385, 2000.

CALLAHAN, H. L. et al. An axenic amastigote system for drug screening. *Antimicrob. Agents Chemother.*, v. 41, n. 4, p. 818-822, 1997.

CEARÁ, Secretaria da Saúde do Estado do Ceará. Boletim Epidemiológico - Leishamaniose Visceral. Ceará. 2007. p. 1-7.

CSERMELY, P.; Ágoston, V.; Pongor, S. The efficiency of multi-targets drugs: the network approach might help drug design. *Trends in Pharmacological Sciences*, v. 26, p. 178-182, 2005.

CHANG, G.; GUIDA, W. C.; STILL, W. C. An internal coordinate Monte Carlo method for searching conformational space. *J. Am. Chem. Soc.*, v. 111, p. 4379-4386, 1989.

CHIN, C. M.; FERREIRA, E. I. O processo de latenciação no planejamento de fármacos. *Química Nova*, v. 22, n. 1, p. 75-84, 1999.

COLAKOGLU, M. et al. Successful treatment of visceral leishmaniasis with fluconazole and allopurinol in a patient with renal failure. *Scandinavian Journal of Infectious Diseases*, v. 38, n. 2, p. 152-154, 2006.

COLAKOGLU, M. et al. Successful treatment of visceral leishmaniasis with fluconazole and allopurinol in a patient with renal failure. *Scandinavian Journal of Infectious Diseases*, v. 38, n. 3, p. 208-210, 2006.

COLUCI, V. R. et al. Identifying relevant molecular descriptors related to carcinogenic activity of polycyclic aromatic hydrocarbons (PAHs) using pattern recognition methods. *J. Chem. Inf. Comput. Sci.*, v. 42, n. 6, p. 1479-1489, 2002.

CONSONNI, V.; TODESCHINI, R. Methods and Principles of Medicinal Chemistry. In: Mannhold, R., Kubibyi, H. e Timmerman, H. *Handbook of molecular descriptors*, Wiley-VCH publishers, 2000.

CONSONNI, V.; TODESCHINI, R.; PAVAN, M., Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *J. Chem. Inf. Comput. Sci.*, v. 42, p. 682-692, 2002.

CONSONNI, V.; TODESCHINI, R.; PAVAN, M., Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies. *J. Chem. Inf. Comput. Sci.*, v. 42, p. 693-705, 2002.

CROFT, S. L.; YARDLEY, V. Chemotherapy of leishmaniasis. *Curr. Pharm. Des.*, v. 8, n. 4, p. 319-342, 2002.

CROFT, S. L.; SEIFERT, K.; YARDLEY, V. Current scenario of drug development for leishmaniasis. *Indian J. Med. Res.*, v. 123, p. 399-410, 2006.

CRONIN, M. T. D. et al. Structure-based classification of antibacterial activity. *J. Chem. Inf. Comput. Sci.*, v. 42, p. 869-878, 2002.

CRONIN, M. T. D.; SCHULTZ, T. W., Pitfalls in QSAR. *J. Mol. Struct. (Theochem)*, v. 622, p. 39-51, 2003.

CRUM-BROWN, A.; FRASER, T.R. On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the ammonium bases, derived from Strychia, Brucia, Thebaia, Codeia, Morphia and Nicotia. *Trans. Roy. Soc. Edinburgh*, v. 25, p. 151-203, 1868.

DE CLERCQ, E. Antiviral drug discovery and development: where chemistry meets with biomedicine. *Antiviral Research*, v. 67, p. 56-75, 2005.

DE CLERQ, E. D.; FIELD, H. J. Antiviral prodrugs – the development of successful prodrug strategies for antiviral chemotherapy. *Br. J. Pharmacol.*, v. 147, p. 1-11, 2006.

DE OLIVEIRA, D. B.; GAUDIO, A. C. BuildQSAR: A new computer program for QSAR analysis. *Quant. Struct.-Act. Relat.*, v. 19, n. 6, p. 599-601, 2001.

DRAGON, versão 3.0; Milano Chemometrics and QSAR Research Group: Dept. of Environmental Sciences, University of Milano-Bicocca, Plaza della Scienza, 1 - 20126 Milano, Italy.

DRUG INFORMATION SYSTEM. Banco de dados sobre fármacos. Disponível em: < http://www.druginfosys.com/index.aspx>. Acesso: 30.Jun.09.

EGAN, W. J.; ZLOKARNIC, G.; GROOTENHUIS, P. D. J. In silico prediction of drug safety: despite progress there is abundant room for improvement. *DDT:Tecnologies*, v. 1, n. 4, p. 381- 387, 2004.

EL-HASSAN, A. M. Leishmaniasis in Sudan 4. Post kala-azar dermal leishmaniasis. *Trans Roy Soc Trop Med Hyg*, v. 95, p. 59-76, 2001.

ERIKSSON, L. et al. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health. Perspect.*, v. 11, n. 10, p. 1361-1375, 2003.

ERTL, P.; ROHDE, B.; SELZER, P. Fast calculation of molecular polar surface area as a sum od fragment based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.*, v. 43, n. 20, p. 3714-3717, 2000.

FOLKERS, G.; MERZ, A.; ROGNAN, D. CoMFA: scope and limitations. In: KUBINYI, H. 3D QSAR in drug design: Theory methods and applications. Holanda: Kluwer Academic Publishers, 2000. p. 583-617.

GÁLVEZ, J. et al. Charge indexes. New topological descriptors, *J. Chem. Inf. Comput. Sci.*, v. 34, n. 3, 1994.

GASTEIGER, J. et al. Chemical information in 3D space. *J. Chem. Inf. Comput. Sci.*, v. 36, p. 1030-1037, 1996.

GAUDIO, Anderson Coser. Estudo teórico das relações entre estrutura química e atividade biológica de antagonistas do cálcio da classe 1,4-diidropiridinas. 1992. 141f. Tese (Mestrado em Química). Universidade Estadual de Campinas – UNICAMP-IQ, Campinas, 1992.

GAUDIO, A. C.; Zandonade, E. Proposição, validação e análise dos modelos que correlacionam estrutura química e atividade biológica. *Quim. Nova*, v. 24, n. 5, p. 658-671, 2001.

GAUDIO, A. C.; Ferreira, M. M. C.; Montanari, C. A.. Seleção de variáveis em QSAR. *Quim. Nova*, v. 25, n. 3, p. 439-448, 2002.

GEBRE-HIWOT, A.; FROMMEL, D. The in-vitro anti-leishmanial activity of inhibitors of ergosterol biosynthesis. *J. Antimicrob. Chemother.*, v. 32, p. 837-842, 1993.

GHOSE, A. K. et al. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationship. 4. Additional parameters for hidrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.*, v. 29, n. 3, p. 163-172, 1989.

GONTIJO, C. M. F.; MELO, M. N. Leishmaniose visceral no Brasil: Quadro atual, desafios e perspectivas. *Rev. Bras. Epidemiol*., v. 7, n. 3, p. 338-349, 2004.

GOLDMAN, R. C. et al. Inhibition of 2,3-oxidosqualene-lanosterol cyclase in Candida albicans by pyridinium ion-based inhibitors. *Antimicrob. Agents Chemother.*, v. 40, n. 4, p. 1044-1047, 1996.

GOODMAN, J. M.; STILL, W. C. An unbounded systematic search of conformational space. *J. Comput. Chem.*, v. 12, n. 9, p. 1110-1117, 1991.

GOTO, H.; OSAWA, E. Approaches to the global minimum problem. *J. Mol. Struct.*, v. 285, p. 157-168, 1993.

GRETL, versão 1.6.0; Econometrics Software. Disponível em <http://gretl.sourceforge.net/>. Acesso em: 19.nov.2007.

GUARNIERI, F.; WILSON, S.R. Simulated annealing of rings using an exact ring closure algorithm. *Tetrahedron*, v. 48, n. 21, p. 4271-4282, 1992.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, v.3, p. 1157-1182, 2003.

HAJDUK, P. J. et al. Privileged molecules for protein binding identified from NMR-Based Screening. *J. Med. Chem.*, v. 43, p. 3443-3447, 2000.

HALL, H. K. Jr. Correlation of the base strengths of amines. *J. Am. Chem. Soc.*, v. 79, n. 20, p. 5441-5444, 1957.

HALIM, M. A. et al. Successful treatment of visceral leishmaniasis with allopurinol plus ketoconazole in a renal-transplant recipient after the occurrence of pancreatitis due to stibogluconate. *Clinical Infectious Diseases*, v. 16, n. 3, p. 397-399, 1999.

HARMS, G.; FELDMEIER, H. Review: HIV infection and tropical parasitic diseases – deleterious interactions in both directions? *Trop. Med. Int. Health*, v. 7, n. 6, p. 479-488, 2002.

HASAN, A. et al. Acyclic pyrazolo[3,4-*d*]pyrimidine nucleoside as potential leishmaniostatic agent. *Nucleosides, Nucleotides and Nucleics Acids*, v. 25, p. 55-60, 2006.

HERWALDT, B. L. Leishmaniasis. *The lancet*, v. 354, p. 1191-1199, 1999.

HOPFINGER et al. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.*, v. 119, n. 43, p. 10509-10524, 1997.

HOPKINS, A. L. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, v. 4, n. 11, p. 682-690, 2008.

HOWARD, A. E.; KOLLMAN, P. A. An analysis of current methodologies for conformational searching of complex molecules. *J. Med. Chem.*, v. 31, n. 9, p. 1669-1675, 1988.

JAWORSKA, J.; NIKOLOVA-JELIAZKOVA, N.; ALDENBERG, T., Qsar applicability domain estimation by projection of the training set descriptor space: a review. *Altern. Lab. Anim.*, v. 33, n. 5, p. 445-459, 2005.

JI, H. et al. A three-dimensional model of lanosterol 14a-demethylase of Candida albicans and its interaction with azole antifungals. *J. Med. Chem.*, v. 43, n. 13, p. 2493-2505.

JIA, J. et al. Mechanisms of drug combinations: interaction and network perspectives. *Nature Reviews Drug Discovery*, v. 8, p. 111-128, 2009.

JUDSON, R.S. et al. Conformational searching methods for small molecules. II. Genetic algorithm approach. *J. Comput. Chem.*, v. 14, n. 11, p. 1407-1414, 1993.

KIER, L. et al. An index of electrotopological state for atoms in molecules. *J. Math. Chem.*, v. 7, n. 1, p. 229-241, 1991.

KERNS, E. H.; DI, L. *Drug-like properties: concepts, structure, design and methods: from ADME to toxicity optimization*, Califórnia: Academic Press, 2008.

KLEBE, G.; MIETZNER, T. A fast and efficient method to generate biologically relevant conformations. *J. Computer-Aided Mol. Design*, v.8, p.583-606, 1994.

KLUCIK, J. et al. Targacept active conformation search: a new method for predicting the conformation of a ligand bound to its protein target. *J. Med. Chem.*, v. 47, n. 27, p. 6831-6839, 2004.

KOLOSSVARY, I.; GUIDA, W. C. Torsional flexing: conformational searching of cyclic molecules in biased internal coordinate space. *J. Comput. Chem.*, v. 14, n. 6, p. 691-698, 1993.

KUBINYI, H. Statistical Methods. In: Mannhold, R., Krogsgaard-Larsen, P. e Timmerman, H. *QSAR: Hansch analysis and related approaches*, VCH Publishers, New York, NY (USA) 1993.

KUBINYI, H., Chemical similarity and biological activities, *J. Braz. Chem. Soc.*, v. 13, n. 6, p. 717-726, 2002.

KUYUCU, N. et al. Successful treatment of visceral leishmaniasis with allopurinol plus ketoconazole in an infant who developed pancreatitis caused by meglumine antimoniate. *Pediatric Infectious Disease Journal*, v. 20, n. 4, p. 455-457, 2001.

LEACH, A. R. *Molecular modeling principles and applications*, Addison Wesley Longman Limited Publishers, England, (UK) 1996.

LENHART, A. et al. Binding structures and potencies of oxidosqualene cyclase inhibitors with the homologous squalene-hopene cyclase. J. Med. Chem., v. 46, n. 11, p. 2083-2092, 2003.

LI, Y. et al. 4D-Fingerprint categorical QSAR models for Skin sensitization based on the classification of local lymph node assay measures. *Chem. Res. Toxicol.*, v.20, p. 114-128, 2007.

LLORENTE, S. et al. Therapy of visceral leishmaniasis in renal transplant recipients intolerant to pentavalent antimonials. *Transplantation*, v. 70, n. 5, p. 800-801, 2000.

LUKES, J. et al. Evolutionary and geographical history of the Leishmania donovani complex with revision of current taxonomy. *PNAS*, v. 104, n. 22, p. 9375-9380, 2007.

MARR, J. J. et al. Biological action of inosine analogs in *Leishmania* and *Trypanossoma* spp. *Agents Chemother.*, v. 25, n. 2, p. 292-295, 1984.

MARTIN, Y. C.; KOFRON, J. L.; TRAPHAGEN, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, v. 45, p. 4350-4358, 2002.

MARVIN versão 5.2; Chemaxon software. Disponível em: <http://www.chemaxon.com/marvin/sketch/index.jsp>. Acesso em: 29/jun/09.

MCGARRAH, D. B.; JUDSON, R.S. Analysis of the genetic algorithm method of molecular conformation determination. *J. Comput. Chem.*, v. 14, n. 11, 1385-1395, 1993.

MIKOLAJCZYK, R. T.; DISILVESTO, A.; ZHANG, J. Evaluation of logistic regression reporting in current obstetrics and gynecology literature. *Obstetrics & Gynecology*, v. 111, n. 2, p. 413-419.

MOLINSPIRATION; Molinspiration software.
Disponível em: <http://www.molinspiration.com/cgi-bin/properties >. Acesso em: 10/jul/09.

MUYOHBWE, M.; QUELLETTE, M.; PARADOPOULOU, B. Selective killing of *Leishmania* amastigotes expressing a thymidine cinase suicide gene.
Exp. Parasitol., v. 85, n. 1, p. 35-42, 1997.

NGO, J.T.; KARPLUS, M. Pseudosystematic conformational search. Application to cycloheptadecane. *J. Am. Chem. Soc.*, v. 119, n. 24, p. 5657-5667, 1997.

OMEGA, versão 1.8.0; OpenEye Scientific Software: 9 Bisbee Court, Suite D, Santa Fe, NM 87508.

ORANGE, versão 0.9.62; Demsar, J.; Zupan, B.; Leban, G. Orange: From Experimental Machine Learning to Interactive Data Mining, White Paper 2004 (http://www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana.

OROZCO, M. et al. Tautomerism of xanthine oxidase substrates hypoxanthine and allopurinol. *J. Org. Chem.*, v. 61, p. 5964-5971, 1996.

PEDUZZI, P. et al. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.*, v. 49, n. 12, p. 1373-1379, 1996.

PEINSHOFF, C. E.; DIXON, J. S. Improvements to the distance geometry algorithm for conformational sampling of cyclic structures. *J. Comput. Chem.*, v. 13, n. 5, p. 565-569, 1992.

PEROLA, E.; CHARIFSON, P. S. Conformational Analysis of Drug-Like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization upon Binding. *J. Med. Chem.*, v. 47, p. 2499-2510, 2004.

PLIEGO JR., J. R. Modelos contínuos do solvente: Fundamentos. *Quim. Nova*, v. 29, n. 3, p. 535-542, 2006.

PODUST, L.M.; POULOS, T. L. e WATERMAN, M. R. Crystal structure of cytochrome P450 14α-sterol demethylase (CYP51) from Mycobacterium tuberculosis in complex with azole inhibitors. *Proc. Natl. Acad. Sci*, v. 98, n. 6, p. 3068-3073, 2001.

POSEVIEWWEB – A web interface to PoseView. Disponível em: <http://poseview.zbh.uni-hamburg.de/~poseview/poseview.php>. Acesso em: 27/07/09.

RANDIC, M. Molecular shape profiles. *J. Chem. Inf. Comput. Sci.*, v. 35, n. 3, p. 373-382, 1995.

ROLLAND, C. et al. G-Protein- Coupled Receptor affinity prediction based on the use of a profiling dataset: QSAR design, synthesis, and experimental validation. *J. Med. Chem.*, v. 48, n. 21, p. 6563-6574, 2005.

RÜCKER, C. et al. Counts of all walks as atomic and molecular descriptors. *J. Chem. Inf. Comput. Sci.*, v. 33, n. 5, 683-695, 1993.

RUGE, E.; KORTING, H. C.; BORELLI, C. Current state of three-dimensional characterization of antifungal targets and its use for molecular modelling in drug design.

SAUNDERS, M. et al. Conformations of cycloheptadecane. A comparison of methods for conformational searching. *J. Am. Chem. Soc.*, v. 112, p. 1419-1427, 1990.

SCHMIDT, C. F. Computational Approach to the Study of Thinking. Disponível em http://www.rci.rutgers.edu/~cfs/305_html/Computation/comptoc.html. Acesso em: 07 Jul. 2008.

SCHROETER, T. S. et al. Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *J. Comput. Aided Mol. Des.*, v. 21, p. 485-498, 2007.

SCHUUR, J. H.; SELZER, P.; GASTEIGER, J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.*, v. 36, p. 334-344, 1996.

SINGH, R. K.; PANDEY, H. P.; SUNDAR, S. Visceral leishmaniasis (kala-azar): Challenges ahead. *Indian J. Med. Res.*, v.123, p. 331-344, 2006.

SILVERMAN, R. B. *The organic chemistry of drug design and drug action*, 2ª Edição, Califórnia: Elsevier Academic Press, 2004.

SPELLMEYER, D. C. et al. Conformational analysis using distance geometry methods. *J. Mol. Graphics Mod.*, v. 15, p. 18-36, 1997.

SUNDAR, S.; CHATTERJEE, M. Visceral leishmaniasis – current therapeutic modalities. *Indian J. Med. Res.*, v. 123, p. 345-352, 2006.

THOMAS, G. Química Medicinal Uma Introdução. In:____Membranas biológicas. 1ª edição. Rio de Janeiro, 2003, p. 105-140.

TRUHLAR, D.G. et al. New class IV charge model for extracting accurate partial charges from wave function. *J. Phys. Chem. A*, v. 102, n. 10, p. 1820-1831, 1998.

TRUHLAR, D.G. et al. Extension of the plataform of applicability of the SM5.42R universal solvation model. *Theor. Chem. Acc.*, v. 103, p. 9-63, 1999.

VIGHI, M. The use of QSARs for heterogeneous chemical substances: meaning, predictive capability, and practical applications. *Biotherapy*, v. 11, p. 97-104, 1998.

WANG, C. Efficient algorithm for conformational search of macrocyclic molecules. *J. Comput. Chem.*, v. 18, n. 2, p. 277-289, 1997.

WANG, C. C.; WANG, A. L. Introduction to antiparasitic chemotherapy. In: KATZUNG, B. G. (Org.). *Basic & Clinical Pharmacology*, 10ª edição. McGraw-Hill, (USA) 2007, p. 842-844.

WEINBERG, N.; WOLFE, S. A comprehensive approach to the conformational analysis of cyclic compounds. *J. Am. Chem. Soc.*, v. 116, p. 9860-9868, 1994.

WERMUTH, C. G. Design prodrugs and bioprecursors. In: WERMUTH, C. G. (EDITOR). *The practice of medicinal chemistry*, 2ª edição, Elsevier Academic Press, 2003, p. 561-585.

WHO, Report on Global Surveillance of Epidemic-prone Infectious Diseases: 2000. Dispnível em: <http://www.who.int/csr/resources/publications/surveillance/WHO_CDS_CSR_ISR_2000_1/en/>. Acesso em: 9.out.2007.

WHO, The World Health Report 2004 – changing history. Disponível em: http://www.who.int/whr/2004/en/. Acesso em: 9.out.2007.

WHO, Leishmaniasis and Hiv co-infection: 2007. Seção Burden of disease. Disponível em: <http://www.who.int/leishmaniasis/burden/hiv_coinfection/burden_hiv_coinfection/en/index.html>. Acesso em: 9.out.2007.

WHITLEY, D. C.; FORD, M.G.; LIVINGSTONE, D. J. Unsupervised forward selection: a method for eliminating redundant variables. *J. Chem. Inf. Comput. Sci.*, v. 40, n. 5, p.1160-1168, 2000.

YOUNG, D. C. *Computational Chemistry: A practical guide for applying techniques to real-world problems*, John Wiley & Sons, INC, NY (USA) 2001.

XIA, X. et al. Classification of kinase inhibitors using a bayesian model. *J. Med. Chem.*, v. 47, n. 18, p.4463-4470, 2004.