



MARIENE HELENA DUARTE

**INOVAÇÃO DO MÉTODO MIA-QSAR:
APLICAÇÃO PARA DOENÇAS TROPICAIS
NEGLIGENCIADAS**

**LAVRAS – MG
2016**

MARIENE HELENA DUARTE

**INOVAÇÃO DO MÉTODO MIA-QSAR: APLICAÇÃO PARA DOENÇAS
TROPICAIS NEGLIGENCIADAS**

Tese apresentada ao Departamento de Química da Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Agroquímica, área de concentração em Química/Bioquímica, para a obtenção do título de Doutora.

ORIENTADOR
Dr. MATHEUS PUGGINA DE FREITAS

LAVRAS – MG
2016

**Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca
Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).**

Duarte, Mariene Helena.

Inovação do método MIA-QSAR : aplicação para doenças tropicais negligências / Mariene Helena Duarte. – Lavras : UFLA, 2016.

165 p. : il.

Tese(doutorado)–Universidade Federal de Lavras, 2016.

Orientador(a): Matheus Puggina de Freitas.

Bibliografia.

1. aug-MIA-QSAR. 2. aug-MIA-QSARcolor. 3. PLS. 4. MLR.
I. Universidade Federal de Lavras. II. Título.

MARIENE HELENA DUARTE

**INOVAÇÃO DO MÉTODO MIA-QSAR: APLICAÇÃO PARA DOENÇAS
TROPICAIS NEGLIGENCIADAS
INNOVATION OF THE MIA-QSAR METHOD: APPLICATION TO
NEGLECTED TROPICAL DISEASES**

Tese apresentada ao Departamento de Química da Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Agroquímica, área de concentração em Química/Bioquímica, para a obtenção do título de Doutora.

APROVADA em 20 de abril de 2016.
Dra. Denise Alvarenga Rocha - UFLA
Dra. Daiana Teixeira Mancini - UFLA
Dr. Luiz Gustavo de Lima Guimarães - UFSJ
Dr. Teodorico de Castro Ramalho - UFLA

Dr. Matheus Puggina de Freitas
Orientador

**LAVRAS – MG
2016**

AGRADECIMENTOS

À Universidade Federal de Lavras e ao Departamento de Química, pela oportunidade concedida para realização do doutorado.

À Fundação de Amparo e Pesquisa do Estado de Minas Gerais (Fapemig) pela concessão da bolsa de estudos.

Ao professor Matheus pela orientação, disponibilidade, paciência e valiosa contribuição para o meu crescimento profissional.

Aos professores e amigos do laboratório de Modelagem molecular, em especial Carol, Estella, Daiana, Dani, Geovanna, Josué, Letícia Assis, Letícia Garcia, e Tamiris pela colaboração e amizade.

Ao meu parceiro de pesquisa Stephen Jones pela cooperação e amizade.

Ao meu amor Luís Fernando, pelo incentivo e paciência.

A toda minha família e amigos, em especial meus pais, Geraldo e Aparecida, que foram durante todos esses anos o meu porto seguro e fonte de inspiração para concretização deste trabalho.

RESUMO

A análise multivariada de imagens aplicada à relação quantitativa estrutura-atividade (MIA-QSAR) é uma técnica de QSAR 2D que tem apresentado resultados promissores em relação ao desenvolvimento de novos candidatos a fármacos em razão de sua simplicidade, rapidez e baixo custo. Nesse sentido, o presente estudo tem por objetivo introduzir, aprimorar e consolidar as novas dimensões aug-MIA-QSAR e aug-MIA-QSARcolor, bem como aplicá-las ao estudo de doenças negligenciadas para obtenção de novos compostos-alvo via interpretação químico-biológica dos descritores moleculares. O presente estudo utilizou quatro conjuntos de dados com atividade biológica experimental conhecida contra doença de Chagas, malária, dengue e esquistossomose e realizou um estudo comparativo entre as três abordagens: MIA-QSAR, aug-MIA-QSAR e aug-MIA-QSARcolor. De forma geral, a representação dos átomos por meio de diferentes cores e tamanhos proporcionais ao raio de van der Waals (abordagens aug-MIA) melhorou a capacidade preditiva e interpretativa do método em relação ao sistema tradicional para todos os conjuntos de dados. A utilização de cores proporcionais à eletronegatividade mostrou que os descritores MIA são capazes de identificar propriedades periódicas relevantes para a atividade em estudo. Finalmente, a utilização de cores sólidas, sem efeito de iluminação dos átomos, nos estudos contra malária, dengue e esquistossomose, permitiu uma correta identificação dos átomos através dos valores de pixel (descritores) e, posteriormente, a interpretação química do substituinte em relação à atividade biológica. Pode-se inferir que derivados tiosemicarbazonas e semicarbazonas com um anel tri-substituído no grupo R1 e um grupo triflúor metila na posição R3 em vez de um cloro, resultou em maior atividade antitripanossoma. No estudo de antimaláricos, a atividade das quinolona-4(1*H*)iminas pode ser melhorada quando: 1) nos substituintes em R1 e R2 houver presença de grupos doadores de elétrons, 2) em R3 houver cadeias amino-alquílicas maiores e 3) em R4 houver substituintes com maior volume atômico. No estudo para dengue foi verificado que tetrapeptídeos com aminoácidos pequenos não ramificados nas posições A1 e A4 podem aumentar a afinidade do substrato (K_m) pela proteína NS3 e, nas posições A1 e A2, a taxa de clivagem do substrato (k_{cat}). Por outro lado, aminoácidos ácidos nas posições A2 e A4 foram relacionados com baixa afinidade do substrato a proteína NS3 e, quando presentes em A1, com baixa taxa clivagem do substrato. Por fim, a presença de substituintes metoxila nas posições R1 (ou R2) e R5 no esqueleto dos neolignanas podem favorecer a atividade antiesquistossômica.

Palavras-chave: **aug-MIA-QSAR. aug-MIA-QSARcolor. PLS. MLR.**

ABSTRACT

Multivariate image analysis applied to the quantitative structure-activity relationships (MIA-QSAR) is a 2D QSAR technique that has been presenting promising outcomes for the development of new drug candidates, due to its simplicity, rapidity and low cost. In this way, the present study aims at introducing, consolidating and improving the new dimensions named aug-MIA-QSAR and aug-MIA-QSARcolor, as well as applying them to the study of neglected diseases, in order to obtain new drug targets using chemico-biological interpretation of the MIA molecular descriptors. Four compound data sets with experimental bioactivities against Chagas disease, malaria, dengue and schistosomiasis were evaluated using three approaches: MIA-QSAR, aug-MIA-QSAR and aug-MIA-QSARcolor. In general, representations of atoms as spheres with different colors and sizes proportional to the corresponding van der Waals radii (aug-MIA approaches) improved the predictive ability and interpretability in all data sets. The use of colors proportional to the Pauling's electronegativity showed that MIA descriptors are capable of identifying periodic properties relevant for the studied activity. Finally, solid colors instead of spotlighted atoms allowed a correct identification of atoms by means of pixel values in the studies for malaria, dengue and schistosomiasis, which were, subsequently, useful for the chemical interpretation related to the bioactivity. It can be inferred that semicarbazones and thiosemicarbazones derivative with a tri-substituted ring in R1 group and a trifluoro methyl group in the R 3 position instead of a chlorine antitripanosoma resulted in higher activity. The antimalarial activity of quinolon-4(1*H*)imines can be improved if: 1) R1 and R2 are electron donor groups, 2) R3 has long aminoalkyl chains, and 3) R4 possesses substituents with big atomic volume. In the study for dengue, it was found that tetrapeptides with unbranched small size amino acids in the A1 and A4 positions can increase the substrate affinity (K_m) to the NS3 protein, and when in A1 and A2 positions, the substrate cleavage rate (k_{cat}). On the other hand, acidic amino acids in the A2 and A4 positions were found to be related with low substrate affinity to the NS3 protein and when present in A1, with low substrate cleavage rate. Finally, the presence of methoxy substituents in R1 (or R2) and R5 in the neolignan backbone can favor their antischistosomal activity.

Keywords: aug-MIA-QSAR. aug-MIA-QSARcolor. PLS. MLR.

LISTA DE FIGURAS

Figura 1 -	Etapas do processo de descoberta e planejamento racional de fármacos.....	24
Figura 2 -	Estrutura química de um derivado R-3-amino-pirrolidinas desenhada no programa ChemSketch para construção de um modelo MIA-QSARt.....	35
Figura 3 -	Imagem da estrutura química de um derivado R-3-amino-pirrolidinas desenhada no programa GaussView para construção de um modelo aug-MIA-QSAR.....	36
Figura 4 -	Cores padrão do programa GaussView 5.0.8 para os diferentes átomos.....	37
Figura 5 -	Distribuição dos casos de doença de Chagas reportados à OMS, em nível mundial, 2010-2013.....	59
Figura 6 -	Ciclo de transmissão do <i>T. cruzi</i> em hospedeiros vertebrados e invertebrados.....	61
Figura 7 -	Estrutura química do nifurtimox e benznidazol.....	62
Figura 8 -	Sobreposição das estruturas químicas, segundo a abordagem MIA-QSARt.....	71
Figura 9 -	Sobreposição da série congêneres de compostos que ilustram a subestrutura comum e as porções variáveis dos derivados tiossemicarbazona e semicarbazonas com atividade antitripanossoma para as abordagens aug-MIA-QSAR (A) e aug-MIA-QSARcolor (B).....	73
Figura 10 -	Construção da matriz X a partir do desdobramento do arranjo tridimensional obtido da sobreposição das imagens do conjunto de dados (a abordagem aug-MIA-QSAR foi utilizada como exemplos).....	74
Figura 11 -	Tiossemicarbazonas propostas a partir da combinação das subestruturas dos compostos 1 e 2.....	83
Figura 12 -	Ciclo de vida do <i>Plasmodium</i> no mosquito e no homem.....	92
Figura 13 -	Sobreposição das imagens das estruturas químicas dos derivados de quinolona-4(1 <i>H</i>)-iminas, resultando em um arranjo tridimensional para a abordagem MIA-QSARt.....	98
Figura 14 -	Sobreposição da série congêneres de compostos que ilustram a subestrutura comum e as porções variáveis dos derivados quinolona-4(1 <i>H</i>)-imina com atividade antimalárica para as abordagens aug-MIA-QSAR (A) e aug-MIA-QSARcolor (B).....	100
Figura 15 -	Áreas de ricos de transmissão da dengue no mundo em 2014..	118

Figura 16 -	Sobreposição das imagens dos peptídeos modificados para as abordagens MIA-QSARt (A), aug-MIA-QSAR (B) e aug-MIA-QSARcolor (C).....	126
Figura 17 -	Construção do arranjo tridimensional e matriz X obtida após desdobração das imagens aug-MIA-QSARcolor do conjunto de peptídeos.....	126
Figura 18 -	Distribuição da esquistossomose em todo mundo em 2012.....	146
Figura 19 -	Ciclo de vida do <i>Schistosoma</i> no caramujo e no homem.....	148
Figura 20 -	Sobreposição de neolignanas com atividade antiesquistossomótica para as abordagens MIA-SARt (A), aug-MIA-SAR (B) e aug-MIA-SARcolor (C).....	151

LISTA DE GRÁFICOS

Gráfico 1 -	Número de variáveis latentes <i>versus</i> RMSECV para as abordagens MIA-QSARt (A), aug-MIA-QSAR (B) e aug-MIA-QSARcolor (C).....	77
Gráfico 2 -	Valores de pIC ₅₀ experimentais <i>versus</i> preditos para a série de tiosemicarbazonas e semicarbazonas, segundo a abordagem MIA-QSARt.....	79
Gráfico 3 -	Valores de pIC ₅₀ experimentais <i>versus</i> preditos para a série de tiosemicarbazonas e semicarbazonas, segundo a abordagem aug-MIA-QSAR.....	80
Gráfico 4 -	Valores de pIC ₅₀ experimentais <i>versus</i> preditos para a série de tiosemicarbazonas e semicarbazonas, segundo a abordagem aug-MIA-QSARcolor.....	81
Gráfico 5 -	Resíduos de <i>Student</i> <i>versus</i> <i>leverages</i> para o modelo aug-MIA-QSAR	105
Gráfico 6 -	Valores de pIC ₅₀ experimentais <i>versus</i> preditos utilizando o modelo aug-MIA-QSAR.....	106
Gráfico 7 -	PLS-DA para a série de quinolona-4(1 <i>H</i>)-iminas.....	110
Gráfico 8 -	Pesos das variáveis para as quinolona-4(1 <i>H</i>)-iminas.....	110
Gráfico 9 -	Gráficos de <i>scores</i> para as componentes principais PC1 e PC2 da série de neolignanas, utilizando os descritores MIA-SARt (A), aug-MIA-SAR (B) e aug-MIA-SAR _{color} (C).....	156
Gráfico 10 -	Dendrograma dos modelos MIA-SARt (A), aug-MIA-SAR (B) e aug-MIA-SAR _{color} (C) para as 18 neolignanas.....	158

LISTA DE QUADROS

Quadro 1 -	Série de derivados de tiossemicarbazonas e semicarbazonas utilizada na modelagem QSAR.....	66
Quadro 2 -	Estrutura química e atividade antiplasmódica de derivados quinolona- 4(1 <i>H</i>)-iminas.....	95
Quadro 3 -	Estrutura química e atividade antiesquistossomótica dos derivados neolignanais.....	150

LISTA DE TABELAS

Tabela 1 -	Doenças negligenciadas segundo a OMS e seus agentes etiológicos.....	21
Tabela 2 -	Valores de pixels para representar as cores dos diferentes átomos segundo as abordagens aug-MIA-QSAR e aug-MIA-QSARcolor.....	72
Tabela 3 -	Resultados estatísticos para os modelos MIA-QSAR para a série de tiossemicarbazonas e semicarbazonas com atividade antitripanossômicas.....	78
Tabela 4 -	Valores de pIC ₅₀ preditos para os derivados de tiossemicarbazonas e semicarbazonas (A-E) propostos.....	83
Tabela 5 -	Valores de pixels para os átomos na abordagem aug-MIA-QSAR e valores de pixels e escala de cores para os átomos na abordagem aug-MIA-QSARcolor, segundo a escala de eletronegatividade de Pauling (ϵ).....	99
Tabela 6 -	Coefficiente de correlação entre as variáveis do modelo aug-MIA-QSAR.....	108
Tabela 7 -	Variáveis selecionadas para construção do modelo aug-MIA-QSAR.....	109
Tabela 8 -	Sequência A1'-A4' dos peptídeos modificados com estrutura geral Abz-RRRRXXXX-NY-NH ₂ ^a e valores experimentais obtidos para a constante de Michaelis (K _m) e parâmetro cinético para a constante de clivagem (k _{cat}) para a protease NS3 do DENV2	124
Tabela 9 -	Estatísticas para os modelos MIA-QSAR/PLS obtidos para os parâmetros K _m e k _{cat} dos substratos peptídicos inibidores da protease NS3 do DENV tipo 2.....	129
Tabela 10 -	Parâmetros estatísticos e equações dos modelos MLR para K _m e k _{cat} , para os 54 peptídeos modificados.....	132

Tabela 11 -	Variáveis contidas na abordagem aug-MIA-QSARcolor para o parâmetro K_m , com compostos ordenados do menor para o maior valor de K_m	133
Tabela 12 -	Variáveis contidas no modelo aug-MIA-QSARcolor para o parâmetro k_{cat} , com compostos ordenados do menor para o maior valor de k_{cat}	136
Tabela 13 -	Parâmetros estatísticos para os modelos MIA-SARt, aug-MIA-SAR e aug-MIA-SAR _{color}	159
Tabela 14 -	Variáveis selecionadas para modelos MIA-SARt, aug-MIA-SAR e aug-MIA-SARcolor.....	161

LISTA DE ABREVIATURAS

Bmp	<i>Bitmaps</i>
H	Hidrogênio
B	Boro
C	Carbono
Cl	Cloro
N	Nitrogênio
S	Enxofre
I	Iodo
Br	Bromo
O	Oxigênio
Log	Logaritmo
Px	Pixel
CH ₃	Metil
Me	Metil

LISTA DE SIGLAS

ADME-TOX	Absorção, distribuição, metabolismo, excreção e toxidez
3D	Terceira dimensão
4D	Quarta dimensão
5D	Quinta dimensão
6D	Sexta dimensão
7D	Sétima dimensão
AIDS	<i>Acquired Immunodeficiency Syndrome</i>
ANVISA	Agência Nacional de Vigilância Sanitária

CoMFA	<i>Comparative molecular field analysis</i>
CoMSIA	<i>Comparative molecular similarity indices analysis</i>
CP	Componentes principais
CP1	Componente principal 1
CP2	Componente principal 2
DALY	<i>Disability-Adjusted Life-Years</i>
DNA	<i>Deoxyribonucleic acid</i>
DNDi	<i>Drugs for Neglected Diseases initiative</i>
DTNs	Doenças Tropicais Negligenciadas
DQI	Departamento de Química
EMA	<i>European Medicines Agency</i>
FDA	<i>Food and Drug Administration</i>
GA	<i>Genetic Algorithm</i>
HCA	<i>Hierarchical clusters methods</i>
HQSAR	Holograma QSAR
HTS	<i>High throughput screening</i>
IC	<i>Inhibitory molar concentration</i>
LBVS	<i>Ligand-based virtual screening</i>
LOOCV	<i>Leave-one-out cross-validation</i>
MIA-QSAR	<i>Multivariate image analysis-QSAR</i>
MSF	<i>Medicin Sans frontiers</i>
NDA	<i>New drug application</i>
NCE	<i>New Chemical Entities</i>
OMS	Organização Mundial da Saúde
ONG	Organização Não Governamental
PCA	<i>Principal component analysis</i>
PCR	<i>Principal component regression</i>
PhRMA	<i>Pharmaceutical Research and Manufactures of America</i>
PLS	<i>Partial least square</i>
P&D	Pesquisa e desenvolvimento
QSAR	<i>Quantitative structure-activity relationship</i>
RGB	<i>Red green blue</i>
RMN	Ressonância magnética nuclear
MRL	<i>Multiple linear regression</i>
NTDs	<i>Neglected tropical disease</i>
RMSE	<i>Root mean square error</i>
SAR	<i>Structure-activity relationship</i>
SBVS	<i>Structure-based virtual screening</i>
SVM	<i>Support vector machines</i>
VS	<i>Virtual screening</i>
LV	<i>Latent variable</i>
WHO	<i>World Health Organization</i>

LISTA DE SÍMBOLOS

Log P	Coeficiente de partição octanol/água
r^2	Coeficiente de determinação
q^2	Coeficiente de determinação da validação cruzada
IC ₅₀	50% da concentração inibitória
pIC ₅₀	Logaritmo negativo de IC ₅₀
r^2_{cal}	Coeficiente de determinação da calibração
r^2_{ext}	Coeficiente de determinação da validação externa
r^2_m	Coeficiente de determinação modificado
r^2_p	Coeficiente de determinação de penalização ajustado
r^2_{rand}	Coeficiente de determinação da randomização
Y _i	Valores de pIC ₅₀
\hat{Y}_i	Valor estimado/predito de pIC ₅₀
\bar{y}	Valor médio de pIC ₅₀
K _m	Afinidade do substrato
k _{cat}	Taxa de clivagem

SUMÁRIO

CAPÍTULO 1 - 1 INTRODUÇÃO GERAL.....	15
2 REFERENCIAL TEÓRICO	18
2.1 Doenças infecciosas e parasitárias	18
2.2 Doenças tropicais negligenciadas (NTDs).....	19
2.3 Produção de medicamentos para doenças negligenciadas	22
2.4 Descoberta e desenvolvimento de novos fármacos	23
2.5 Métodos <i>in silico</i> para desenvolvimento de fármacos	27
2.5.1 QSAR.....	29
2.5.1.1 Análise multivariada de imagem aplicada em QSAR.....	33
2.5.1.2 Construção de um modelo QSAR.....	38
REFERÊNCIAS.....	48
CAPÍTULO 2 - MIA-QSAR: Explorando a atividade antitripanossoma de derivados tiossemicarbazonas e semicarbazonas	57
1 INTRODUÇÃO.....	58
1.1 Aspectos gerais da doença de Chagas.....	58
1.2 Tratamento.....	62
1.3 Desenvolvimento de novos medicamentos.....	63
2 METODOLOGIA.....	65
2.1 Construção das imagens e obtenção das matrizes para as abordagens MIA-QSAR.....	65
2.2 Calibração e validação dos modelos	74
3 RESULTADOS E DISCUSSÕES.....	76
3.1 Validação dos modelos MIA-QSAR _t , aug-MIA-QSAR e aug-MIA-QSAR _{color}	76
3.2 Estruturas químicas propostas.....	82
4 CONCLUSÃO.....	84
REFERÊNCIAS	85
CAPÍTULO 3 - Explorando a atividade antimalárica de quinolona-4(1H)-iminas usando MIA-QSAR.....	88
1 INTRODUÇÃO.....	89
1.1 Aspectos gerais da malária.....	89
1.2 Tratamento.....	93
2 METODOLOGIA.....	95
2.1 Obtenção dos descritores MIA.....	95
2.2 Grupo teste e treinamento.....	102
2.3 Construção e validação do modelo MLR.....	102
3 RESULTADOS E DISCUSSÕES.....	103
3.1 Regressão Linear Múltipla	103

3.2 Aspectos sobre a interpretação do modelo.....	107
4 CONCLUSÃO.....	112
REFERÊNCIAS	113
CAPÍTULO 4 - Análise comparativa proteoquimiométrica MIA-QSAR do substrato da protease NS3 do vírus da dengue	
	117
1 INTRODUÇÃO.....	118
1.1 Aspectos gerais da dengue.....	118
1.2 Pesquisas atuais para controle e tratamento da dengue	121
2 METODOLOGIA.....	123
2.1 Construção das matrizes de dados para modelagem	123
2.2 Métodos quimiométricos para modelagem e validação	127
3 RESULTADOS E DISCUSSÕES.....	128
3.1 Análise PLS.....	128
3.2 Regressão Linear Múltipla	131
3.3 Interpretação estrutural do modelo aug-MIA-QSARcolor.....	133
4 CONCLUSÃO.....	140
REFERÊNCIAS.....	141
CAPÍTULO 5 - Modelagem computacional da atividade antiesquistossomótica de derivados neolignanans com base na abordagem MIA-QSAR	
	145
1 INTRODUÇÃO.....	146
2 METODOLOGIA.....	149
2.1 Obtenção do conjunto de dados	149
2.2 Preparação das imagens químicas e obtenção das matrizes de dados.....	151
2.3 Análise de componentes principais e análise de agrupamento hierárquico.....	152
2.4 A análise discriminante por mínimos quadrados parciais (PLS-DA).....	154
3 RESULTADOS E DISCUSSÕES.....	155
3.1 Análise de componentes principais	155
3.2 Análise de agrupamento hierárquico.....	157
3.3 PLS-DA.....	159
3.4 Interpretação do modelo de aug-MIA-SARcolor.....	160
4 CONCLUSÃO.....	162
REFERÊNCIAS	163
CAPÍTULO 6 - CONSIDERAÇÕES FINAIS.....	165

1 INTRODUÇÃO GERAL

O termo doenças negligenciadas é utilizado para referenciar o conjunto de doenças infecciosas e parasitárias que não somente prevalecem nas regiões mais pobres dos países em desenvolvimento, como também contribuem para a manutenção da pobreza, já que representam um forte entrave ao desenvolvimento socioeconômico dos países onde estas doenças são endêmicas (ORGANIZAÇÃO MUNDIAL DE SAÚDE - OMS, 2010; SOUZA, 2010). Segundo a Organização Mundial de Saúde, as doenças negligenciadas afetam 1,4 bilhão de pessoas, e causam mais de 1 milhão mortes a cada ano (OMS, 2015a).

Atualmente, a Organização Mundial de Saúde define as seguintes doenças em situação negligenciada: úlcera de buruli, doença de Chagas, cisticercose, dengue, dracunculíase, equinococose, fasciolíase, tripanossomíase humana africana, leishmaniose, hanseníase, filariose linfática, oncocercose, raiva, esquistossomose, helmintíase transmitida pelo solo, tracoma e boubá (OMS, 2015a). Essas doenças têm como características comuns a alta endemicidade em áreas rurais e urbanas de países em desenvolvimento, não dispõem de tratamentos eficazes e convivem com escassez de recursos para o desenvolvimento de novas profilaxias (OMS, 2010).

Embora nos últimos anos as doenças negligenciadas tenham recebido maior atenção em relação a investimento e pesquisa, o conhecimento produzido não se reverte em avanços terapêuticos como, por exemplo, novos fármacos, métodos diagnósticos e vacinas (HOTEZ, 2007). As razões para esses impasses podem ser atribuídas à escassez de investimento em pesquisa de ponta e ao pequeno interesse da indústria farmacêutica, que não vê nesse tema um mercado tão lucrativo como nas doenças que afetam os países desenvolvidos (SOUZA, 2010).

A escassez de recursos para desenvolvimento de medicamento, a complexidade que envolve as doenças infecciosas e parasitárias, bem como o alto custo relacionado ao desenvolvimento e produção de uma nova entidade terapêutica tem desafiado a comunidade científica em relação à produção de medicamentos para as doenças negligenciadas (PONDER et al., 2014). No entanto, a utilização de recursos computacionais tem favorecido as pesquisas no sentido de reduzir custos, em função da diminuição de testes exploratórios, e de melhorar a compreensão e interpretação de sistemas bioquímicos e alvos terapêuticos (PONDER et al., 2014). Nessa perspectiva, a relação quantitativa entre estrutura química e atividade biológica (QSAR - *Quantitative Structure-Activity Relationships*) tem se mostrado uma importante ferramenta computacional para auxiliar o desenvolvimento de novos fármacos, visto que permite prever propriedades de novas entidades terapêuticas em tempo e custo reduzidos.

Uma técnica tão preditiva quanto os métodos QSAR mais conhecidos, denominada análise multivariada de imagem aplicada em QSAR (MIA-QSAR) (FREITAS; BROWN; MARTINS, 2005), tem apresentado resultados promissores há mais de uma década ao correlacionar pixels de imagens de estruturas químicas (descritores) com atividade biológica e propriedades físicas (FREITAS, 2006). No entanto, desde a sua criação, em 2005, vem-se buscando incorporar informações químicas mais significativas para as imagens estruturais, de forma que os descritores MIA possam se tornar quimicamente mais interpretáveis. Inicialmente o método MIA-QSAR utilizava imagens em preto e branco, produzindo um sistema binário de pixels, não sendo possível diferenciar os átomos e também átomos de ligações químicas. A representação dos átomos por diferentes cores e círculos com tamanho proporcional ao raio de van der Waals (NUNES; FREITAS, 2013) permitiu incluir informação química nos pixels e abriu caminhos para interpretação química dos descritores MIA. Nesse

sentido, o objetivo deste trabalho é aprimorar e consolidar o método augmented-MIA-QSAR (ou, simplesmente, aug-MIA-QSAR) e aug-MIA-QSARcolor, bem como aplicá-lo ao estudo de doenças negligenciadas para obtenção de novos compostos-alvo por meio da interpretação química dos descritores MIA.

2 REFERENCIAL TEÓRICO

2.1 Doenças infecciosas e parasitárias

Ao longo da história da humanidade, diversas doenças infecciosas e parasitárias, tais como a peste negra, gripe espanhola, malária, tuberculose, varíola, cólera e hanseníase, causaram pânico e levaram à morte milhares de pessoas em todo o mundo e, apesar dos expressivos avanços na medicina e no desenvolvimento de medicamentos, muitas doenças infecciosas e parasitárias, que poderiam ser prevenidas e tratadas, ainda hoje estão entre as principais causas de morbidade e mortalidade em países em desenvolvimento na África, América Latina e Ásia (BRASIL, 2004; JONES et al., 2008; VERONESI, 1991).

A alta prevalência dessas enfermidades nas regiões citadas pode ser relacionada a fatores ambientais e biológicos que se apoiam nos níveis elevados de animais selvagens zoonóticos (JONES et al., 2008), mas principalmente devido às precárias condições de saneamento básico, água potável, moradia e degradação ambiental que colaboram para que vírus, bactérias e parasitos sejam facilmente transmitidos para as populações. Adicionalmente, a baixa eficiência da saúde pública, aliada à falta de medicamentos, programas de controle e prevenção contribuem para a manutenção dessas doenças (BRASIL, 2004; GUIDO; ANDRICOPULO; OLIVA, 2010; VERONESI, 1991).

Por estarem relacionadas principalmente às populações de baixa força política e econômica, essas enfermidades têm sido, ao longo dos anos, esquecidas pelas políticas públicas governamentais, instituições de pesquisas e pelas indústrias farmacêuticas, que não veem nessas doenças um mercado tão lucrativo como as doenças que afetam os países desenvolvidos (HOTEZ, 2007). No entanto, o impacto negativo no desenvolvimento socioeconômico dos países

onde as doenças se manifestam, o elevado número de casos e a expansão para regiões antes não afetadas mostram que as doenças infecciosas e parasitárias são um problema mundial e necessitam de atenção urgente (OMS, 2010). Com base nessa preocupante situação, a Fundação Rockefeller criou, em 1977, um programa denominado “As grandes doenças negligenciadas da humanidade” (do inglês *The great neglected diseases Mankind*), para arrecadar fundos para pesquisa em doenças que atingiam milhões de pessoas pobres em todo o mundo, mas que se mostravam carentes de medicamentos eficientes, testes de diagnóstico, segmentação apropriada para terapias e programas e prevenção (SOUZA, 2010).

2.2 Doenças tropicais negligenciadas (NTDs)

As doenças tropicais abrangem todas as doenças que ocorrem nos trópicos. Na prática, o termo é frequentemente utilizado para se referir às doenças infecciosas e parasitárias que se desenvolvem normalmente em climas quentes e úmidos, tais como a malária, doença de Chagas, esquistossomose, tripanossomíase africana e dengue (OMS, 2015b). Na realidade, a ocorrência dessas enfermidades nessas regiões está muito mais relacionada às precárias condições socioeconômicas das populações do que aos fatores climáticos (CAMARGO, 2008; SOUZA, 2010). As características comuns às doenças tropicais e às doenças negligenciadas levaram a adoção do termo doenças tropicais negligenciadas. Assim, o termo NTDs vem sendo utilizado para referenciar o conjunto de doenças que acomete as populações mais pobres e vulneráveis de países em desenvolvimento nas regiões tropicais, mas que contribuem para a perpetuação do ciclo de pobreza e da desigualdade social (BRASIL, 2010; HOTEZ, 2007; SOUZA, 2010), uma vez que prejudicam o desenvolvimento infantil, causam debilidades no longo prazo, invalidez,

desfiguração, incapacidade para o trabalho e, muitas vezes, a morte, resultando em grandes perdas econômicas para suas famílias e para as nações em que vivem (OMS, 2010).

As primeiras doenças definidas como negligenciadas foram: doença do sono, doença de Chagas, malária, leishmaniose, filariose e esquistossomose. Em seguida, incluíram hanseníase, tuberculose, dengue, febre amarela, HIV/AIDS, entre outras (SOUZA, 2010). Fatores como controle, erradicação, emergência ou reemergência, aliados a investimento financeiro determinam a permanência ou exclusão de tais doenças dos programas de prevenção das NTDs (BRASIL, 2004; SOUZA, 2010).

Em função dos altos investimentos financeiros, malária, HIV/AIDS e tuberculose não são mais consideradas negligenciadas pela OMS (SOUZA, 2010). Atualmente, a OMS possui 17 doenças no programa de controle e erradicação de NTDs (Tabela 1). Juntas, essas doenças atingem 149 países e afetam mais de 1,4 bilhão de pessoas (um sexto da população mundial), causando mais de 1 milhão de mortes a cada ano (OMS, 2015a). No entanto, para cada país há um conjunto de doenças prioritárias. No Brasil, as doenças consideradas prioritárias são: malária, tuberculose, dengue, esquistossomose, hanseníase e doença de Chagas (BRASIL, 2010).

O prejuízo causado pelas doenças negligenciadas não pode ser calculado considerando apenas a mortalidade. Em geral, essas enfermidades causam muito mais incapacidade do que mortalidade (HOTEZ et al., 2007). O impacto causado pelas NTDs na saúde coletiva pode ser mais bem compreendido quando medido pelo parâmetro conhecido como DALYs, do inglês *Disability-Adjusted Life-Years*, que é uma estimativa da soma dos anos de vida perdidos por morte prematura e dos anos de vida produtiva perdidos por incapacidade (MURRAY; LOPEZ, 1997). Os impactos mais eminentes apresentados por algumas doenças são: diminuição da capacidade cognitiva da criança (esquistossomose),

Tabela 1: Doenças negligenciadas, segundo a OMS, e seus agentes etiológicos.

DTNs	Agente etiológico
Úlcera de Burili	<i>Mycobacterium ulcerans</i>
Doença de Chagas	<i>Trypanosoma cruzi</i>
Dengue e Chikungunya	<i>Flavivírus</i>
Dracunculíase	<i>Dracunculus medinensis</i>
Equinococose	<i>Echinococcus granulosus</i> ; <i>echinococcus multilocularis</i>
Treponematoses endêmico (Yaws)	<i>Treponema pallidum</i>
Trematodíases alimentar	<i>Trematodes flatworms ou flukes</i>
Trypanossomíase Africana	<i>Trypanosoma brucei</i>
Leishmaniose	<i>Leishmaniose spp.</i>
Hanseníase	<i>Mycobacterium leprae</i>
Filariase linfática	<i>Wuchereria bancrofti</i> ; <i>Brugia malayi</i> ; <i>Brugia timori</i>
Oncocercose	<i>Onchocerca volvulus</i>
Raiva	<i>Lyssavirus genus</i>
Esquistossomose	<i>Schistosoma mansoni</i> , <i>S. intercalatum</i> , <i>S. mekongi</i> , <i>S. japonicum</i> , <i>S. guineenses</i>
Helmitíase transmitida pelo solo	<i>Áscaris lumbricoides</i> ; <i>Trichuris trichiura</i> , <i>Necator americanus</i> ; <i>Ancylostoma duodenale</i>
Teanise/cisticercoses	<i>Teania Solium</i> ; <i>Teania Saginata</i>
Tracoma	<i>Chlamydia trachomatis</i>

Fonte: Adaptado de OMS (2010).

deformidade física, discriminação e marginalização social (hanseníase e filariase linfática), perda da capacidade de trabalho decorrente de problemas no coração, esôfago, estômago e intestinos (doenças de Chagas), lesões na pele, deformações faciais (leishmaniose) e cegueira (oncocercose) (OMS, 2010). Mais informações sobre as características das doenças estudadas neste trabalho estarão expostas nos capítulos subsequentes.

As medidas preventivas e o tratamento para a maioria das NTDs são conhecidos. O controle dos vetores da dengue, esquistossomose, malária e doença de Chagas é visto como uma alternativa indireta de combate dessas doenças. A atenção aos animais domésticos contaminados pode evitar a raiva, leishmaniose, cisticercose e equinococose. As helmitíases transmitidas pelo

solo e as trematodíases alimentares têm, no saneamento básico e na água tratada, forma importante para sua erradicação. Contudo, para doenças como hanseníase, filariose linfática, úlcera de burili, dracunculíase e tracoma o tratamento das pessoas infectadas é a melhor forma de evitar novas infecções (OMS, 2010). Com exceção da dengue e da dracunculíase, as demais doenças possuem algum tipo de terapia; porém, as drogas disponíveis enfrentam pelo menos um desses problemas: alta toxicidade, efeitos adversos, tratamentos longos, resistência, alto custo e a não distribuição para as populações mais necessitadas (OMS, 2010; (WANG; WANG; LIANG, 2012)

2.3 Produção de medicamentos para doenças negligenciadas

Embora nos últimos anos tenha havido um aumento significativo nos investimentos para pesquisas em doenças negligenciadas e a carga global da maioria das doenças tenha diminuído, a condição das pessoas infectas é bastante preocupante, visto que o desenvolvimento de novos medicamentos e vacinas tem apresentado resultados pouco promissores (HOTEZ, 2007). A situação em relação à produção de medicamentos é alarmante. Dos 1556 medicamentos registrados entre 1975 e 2004, apenas 21 (1,3%) eram referentes a doenças negligenciadas, que representam 12% da carga global das doenças (CHIRAC; TORRELE, 2006). Outro estudo mostra que dos 850 novos produtos terapêuticos registrados entre 2000- 2011, apenas 5 (0,6%) foram indicados para doenças negligenciadas, sendo que nenhum deles corresponde a uma entidade química nova ou vacina (PEDRIQUE et al., 2013).

Dada a escassez de recursos para pesquisa e a dificuldade de converter resultados de pesquisas em novos medicamentos, a ONG “Médicos sem Fronteiras” juntamente com o Instituto Pasteur (França), Fiocruz (Brasil), Ministério de Saúde da Malásia e os institutos de pesquisa clínica da Índia e do

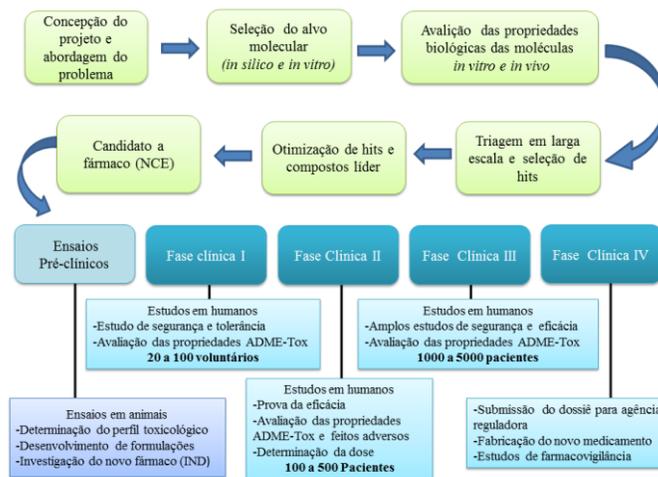
Quênia criaram o programa denominado Iniciativa Medicamentos para Doenças Negligenciadas, do inglês *Drugs for Neglected Diseases initiative* (DNDi). Tal iniciativa tem gerado alguns avanços, tais como a disponibilização de novos tratamentos para malária, doença do sono, leishmaniose e doença de Chagas (VALVERDE, 2015). Contudo, os desafios ainda são enormes, pois o processo de pesquisa e desenvolvimento (P&D) de medicamentos apresenta lacunas que inviabilizam a disponibilização de novos medicamentos às populações afetadas. São elas: (1) os resultados de pesquisa básica são publicados, mas a pesquisa pré-clínica visando desenvolvimento de um novo medicamento não é iniciada; (2) os compostos validados não entram na fase clínica devido às escolhas estratégicas das empresas; (3) os novos medicamentos não chegam aos pacientes (problema de registro, falta de produção, preços muito altos ou produto mais adaptado para as condições locais de utilização) (Médicos Sem Fronteira-MSF, 2001).

2.4- Descoberta e desenvolvimento de novos fármacos

O processo de descoberta e desenvolvimento de fármacos, geralmente, envolve duas fases: (i) fase pré-clínica ou de descoberta e (ii) fase clínica ou de desenvolvimento (GUIDO; ANDRICOPULO; OLIVA; 2010). Na fase pré-clínica, as pesquisas concentram-se na identificação e na otimização de moléculas capazes de representar novas entidades químicas (NCE, do inglês *New Chemical Entities*) (BARREIRO, 2002; GUIDO; ANDRICOPULO; OLIVA, 2010). Na fase clínica, testes verificam a dosagem dos princípios ativos, níveis de toxicidade segura, eficácia e monitoramento de propriedades farmacocinéticas (LIMA, 2007). A Figura 1 mostra um resumo esquemático das duas fases.

A fase pré-clínica inicia-se com estudos relacionados à fisiopatologia da doença e, posteriormente, com a escolha do alvo terapêutico, o qual passa por estudos químicos e biológicos. O estudo químico consiste em identificar e isolar

Figura 1 - Etapas do processo de descoberta e planejamento reacional de fármacos



Fonte: Adaptado de (DIAS et al., 2013).

novas moléculas, seja a partir da extração de produtos naturais, de sínteses químicas ou de processos biotecnológicos. Nos estudos biológicos, testes *in vitro* e *in vivo* verificam se as moléculas isoladas apresentam algum potencial terapêutico (GONZALEZ, 1999). As moléculas identificadas com atividade biológica são chamadas de *hits*, um termo do inglês que pode traduzido como ligantes bioativos. Após a identificação dos *hits*, processos de otimização são iniciados para que propriedades importantes, tais como afinidade, potência, seletividade, absorção e toxicidade, sejam melhoradas. Os compostos bioativos que apresentarem melhor otimização são eleitos compostos líderes (GUIDO; ANDRICOPULO; OLIVA, 2010; LANG et al., 2011). Nessa fase é testado um grande número de moléculas, de tal forma que as estratégias e triagens

biológicas automatizadas em larga escala (HTS, na sigla inglesa para *high throughput screening*), bem como a triagem virtual (VS, na sigla em inglesa para *virtual traigem*) assistida por computadores (*in silico*), têm se mostrado ferramentas fundamentais para identificar as moléculas mais promissoras a NCEs (HAJDUK; GREER, 2007).

A otimização dos compostos líderes leva à obtenção de NCEs, que ainda são investigadas em relação à sua segurança e toxicidade em modelos animais padronizados. Após a conclusão desta etapa, as NCEs mais promissoras são submetidas à aprovação das agências reguladoras, como a FDA nos Estados Unidos, ANVISA no Brasil e EMEA na Europa, para autorização da realização de testes em humanos (fase clínica) através da licença para investigação de um novo candidato a fármaco (DIAS et al., 2013; DIMASI; HANSEN; GRABOWSKI, 2003; LIMA, 2007).

A fase clínica pode ser dividida em 4 etapas, conforme ilustra a Figura 1 (clínica I; clínica II; clínica III e clínica IV). O objetivo principal da fase I da pesquisa clínica é avaliar a tolerância em seres humanos e determinar a posologia segura. Nessa fase realizam-se estudos farmacocinéticos (absorção, metabolismo e excreção), além de estudos de dosagem e biodisponibilidade. O composto é testado em um número pequeno de indivíduos saudáveis, geralmente 100 ou menos (PhRMA, 2015). Na fase clínica II é verificada a eficácia e a segurança do fármaco (efeitos adversos), além de estudos relacionados à dose e ao regime de tratamento adequado. Esta etapa requer o monitoramento junto a cada paciente e, geralmente, são utilizados de 100 a 500 pacientes. Na fase clínica III, a NCE é testada para um número maior de pacientes (1000 a 5000) para gerar dados estatisticamente relevantes que comprovem a sua eficácia e segurança. É a fase em que a investigação clínica científica do tratamento com o novo fármaco costuma ser mais rigorosa e extensa e, após resultados positivos, uma nova solicitação é enviada à agência reguladora para a aprovação do novo

fármaco (NDA, na sigla inglesa para *New Drug Application*) (DIAS et al., 2013; PhRMA, 2015). Após a aprovação da NDA, o fármaco é lançado no mercado e inicia-se a fase clínica IV, onde testes de segurança, eficácia, efeitos adicionais, adversos e efeitos tardios de morbidade continuam sendo realizados por tempo indeterminado, também conhecido como farmacovigilância (DIMASI; HANSEN; GRABOWSKI, 2003)

Conforme descrito acima, o processo de descoberta e desenvolvimento de fármacos é bastante complexo, demorado e caro, cuja produtividade tem sido questionada em razão da relação inversa ao investimento financeiro (MILNE, 2003). Estima-se que o tempo médio para a descoberta e desenvolvimento do fármaco é de, no mínimo, 10 anos, cujo custo estimado é de US\$1,2 bilhões, podendo chegar a US\$ 2,6 bilhões (PhRMA, 2015). O alto custo desse processo se deve, principalmente, às falhas nas etapas iniciais, que resultam em grande reprovação dos candidatos a fármacos. Um estudo mostra que apenas 5 de 40.000 compostos testados em animais atingem testes em humanos e apenas 1 em cada 5 compostos que atingiram os estudos clínicos é aprovado (KAPETANOVIC, 2008). Mesmo assim, a comercialização de um medicamento não garante lucro, uma vez que apenas três de dez medicamentos comercializados produzem receitas que correspondam ou excedam os custos de desenvolvimento (RESS, 2003).

Os longos prazos para descoberta de NCE, a alta probabilidade de fracasso para os compostos que entram nos testes clínicos, e a imprevisibilidade das vendas, aliados aos altos investimentos e a concorrência com o mercado de genéricos têm levado a indústria farmacêutica a investir em inovações científicas e tecnológicas para reduzir os custos e o tempo de produção, bem como melhorar a qualidade dos produtos e recuperar a rentabilidade das empresas (MUNOS, 2009).

Uma das apostas da indústria farmacêutica para reduzir tempo e os altos custos envolvidos no processo de desenvolvimento de novos fármacos é a utilização da modelagem *in silico* (termo usado para estudos gerados com auxílio de computador) (KAPETANOVIC, 2008).

Os estudos *in silico* têm grande aplicação nos estágios iniciais de P&D para seleção e otimização de moléculas com maior potencial de desenvolvimento (MONTANARI, 2010). Esses modelos permitem a modelagem de grandes números de compostos, direcionando a síntese laboratorial apenas às moléculas mais promissoras. Além disso, permite a visualização tridimensional (3D) do complexo fármaco-receptor e fornece informações sobre os requisitos estruturais essenciais para interações adequadas (BARREIRO et al., 1997). A disponibilização de bancos de dados de compostos e receptores moleculares (enzimas, proteínas, DNA) permite que uma variedade de moléculas seja testada em tempo reduzido e sem grandes investimentos financeiros (CARVALHO et al., 2003).

2.5 Métodos *in silico* para desenvolvimento de fármacos

Os programas computacionais desenvolvidos pela química computacional têm se tornado ferramentas indispensáveis para a pesquisa química e para a indústria farmacêutica. Com auxílio de *softwares* e *hardwares* é possível calcular propriedades termodinâmicas, determinar propriedades estruturais (comprimentos e ângulos de ligação), realizar cálculos de otimização geométrica, análise conformacional, caracterização de estados de transição, estabilidade relativa de isômeros, simulação de interações biológicas entre ligantes e alvos moleculares, dentre outras (MORGON, 2001). A grande aplicabilidade, confiabilidade e praticidade desses métodos é afirmada pelo

crescente número de publicações envolvendo cálculos teóricos a cada ano (KAPETANOVIC, 2008).

No contexto de P&D de fármacos, os métodos computacionais figuram, certamente, entre as ferramentas essenciais nas etapas de identificação de ligantes iniciais, na conversão de ligantes iniciais em compostos líderes e na otimização de compostos líderes para propriedades farmacodinâmicas e farmacocinéticas (DIMITRI, 2011; KAPETANOVIC, 2008). Também conhecidos como triagem virtual, os métodos computacionais seguem, basicamente, duas abordagens: a triagem virtual baseada na estrutura do ligante (LBVS, do inglês *ligand-based virtual screening*) e a triagem virtual baseada na estrutura do receptor (SBVS, do inglês *structure-based virtual screening*) (RODRIGUES, 2012).

A ideia geral da abordagem SBVS é prever compostos de uma base de dados capazes de interagir com o sítio de ligação do alvo-molecular e ordenar estas moléculas de acordo com a sua afinidade pelo sítio receptor, com o intuito de identificar ligantes promissores com potencial atividade farmacológica (RODRIGUES et al., 2012). Essa técnica requer o conhecimento da estrutura tridimensional (3D) do alvo molecular, que frequentemente é obtida por difração de raio-X ou por ressonância magnética nuclear (RMN). Casos em que não há a estrutura cristalográfica do alvo molecular, esta pode ser representada por um modelo construído a partir de proteínas homólogas com alta identidade (FERREIRA; OLIVA; ANDRICOPULO, 2011). O principal método computacional utilizado nesta abordagem é a docagem molecular (do inglês, *molecular docking*) (RODRIGUES et al., 2012).

Quando a estrutura do receptor é desconhecida, somente métodos de planejamento baseados na estrutura do ligante podem ser utilizados. Assim, o objetivo da LBVS é explorar propriedades e características de séries de ligantes que compartilham alguma semelhança estrutural entre si e que tenham atividade

biológica conhecida. O princípio desta abordagem é que se as moléculas compartilham alguma semelhança estrutural entre si, podem também apresentar atividade biológica semelhante (RODRIGUES et al., 2012). As técnicas de LBVS incluem o estudo das relações quantitativas entre a estrutura e atividade (QSAR - *quantitative structure-activity relationships*), uso de impressões digitais moleculares (*molecular fingerprints*), emprego de farmacóforos (*pharmacophores*) e a utilização de inteligência artificial na aprendizagem de máquinas (*machine learning*), tais como as máquinas de vetor suporte (SVM - *support vector machines*) (FERREIRA; OLIVA; ANDRICOPILO, 2011).

2.5.1-QSAR

QSAR é uma ferramenta de destaque no planejamento e desenvolvimento de fármacos (MARTINS; FERREIRA, 2013), visto que, por meio dessa técnica é possível propor alterações estruturais que levem à obtenção de moléculas com propriedades farmacocinéticas e farmacodinâmicas superiores àquelas do conjunto de dados inicialmente disponível, a um custo e tempo menor que os métodos experimentais (HANSCH et al., 2002).

A “relação estrutura-atividade” compreende o estudo dos efeitos que a estrutura química de um composto (ligante) pode causar durante sua interação com o receptor biológico (ARROIO; HONÓRIO; SILVA, 2010). Nesse sentido, um estudo de QSAR busca propor estruturas químicas que melhorem a sua atividade biológica ao interagir com receptores biológicos (HANSCH et al., 2002).

O método QSAR consiste na transformação de informação química de um composto em uma série de descritores numéricos, que representam as características mais relevantes para uma dada propriedade e, posteriormente, estabelecer uma relação quantitativa entre os descritores e valores experimentais

de atividade biológica que se deseja otimizar (NEVESA et al., 1998). Essa abordagem baseia-se no princípio de que as variações da resposta biológica para uma série de moléculas estão relacionadas com as variações nas propriedades físicas, químicas e estruturais de cada molécula (GARKANI-NEJAD; AHMADI-ROUDI, 2010).

A técnica de QSAR começou a ser desenvolvida no século XIX. Contudo, sua sistematização ocorreu somente anos mais tarde com o trabalho de Hansch e Fujita (1964), os quais conseguiram mostrar que a atividade biológica de ácidos benzoicos em organismos vivos correlacionava com os descritores coeficiente de partição octanol/água ($\log P$), constante de Hammett σ e parâmetro de lipofilicidade π . Hansch e Fujita relacionaram o efeito de substituintes na atividade biológica de ácidos benzoicos usando regressão linear múltipla.

A metodologia QSAR desenvolvida até 1980 é conhecida como QSAR clássico, em que a atividade biológica é correlacionada apenas com propriedades físico-químicas (experimentais ou calculadas). Modelos QSAR em que apenas a fórmula molecular é considerada chamam-se QSAR unidimensionais, enquanto modelos QSAR que se baseiam em descritores derivados de representação gráfica da molécula bidimensional são chamados de QSAR 2D. Em um estudo QSAR 1D, os descritores representam os efeitos estéricos, hidrofóbicos e eletrônicos. Os descritores 2D, por sua vez, representam efeitos topológicos, constitucionais, geométricos e químico-quânticos (DAMALE et al., 2014)

Posteriormente a este estudo, a metodologia QSAR ganhou destaque no campo da química medicinal e se aperfeiçoou, principalmente em relação à criação de novos descritores, novos *softwares* e tratamento estatístico (MONTANARI, 2011). Em 1988, Cramer e colaboradores desenvolveram um método QSAR com base na estrutura tridimensional de uma série de moléculas congêneres e, através da análise de confórmeros, levou ao desenvolvimento do

QSAR tridimensional (3D). O método que melhor representa esta inovação é a Análise Comparativa de Campos Moleculares - CoMFA, do inglês *Comparative Molecular Field Analysis*. A introdução desta técnica trouxe importantes avanços para o estudo de QSAR, tais como o efeito de estereoisômeros e enantiômeros sobre a atividade biológica, bem como a criação de vários descritores relacionados à estrutura tridimensional da molécula. Além disso, novas dimensões foram incorporadas à metodologia QSAR, com o QSAR 4D (HOPFINGER et al., 1997), QSAR 5D (VEDANI; DOBLER, 2002), QSAR 6D (VEDANI; BOBLER; LILL, 2005) e QSAR 7D (POLANSKI, 2009).

O QSAR 3D baseia-se na representação das moléculas em sua forma tridimensional, alinhadas dentro de uma caixa tridimensional virtual, cujas intersecções são sondadas por campos estéricos e eletrostáticos, gerando os descritores moleculares (CRAMER et al., 1988). Esse método requer alinhamento conformacional tridimensional dos ligantes e, portanto, há necessidade de similaridade estrutural para sobreposição das estruturas químicas. Um inconveniente do QSAR 3D é que o alinhamento das estruturas do conjunto de dados torna o processo de modelagem lento e tedioso. Outras limitações são o grande número de descritores, a identificação da conformação bioativa, estudo das interações intermoleculares de receptores em locais diferentes e, por fim, os descritores topológicos estudados em QSAR 3D funcionam com mesmo princípio de conectividade que em QSAR 2D (DAMALE et al., 2014).

O método QSAR 3D faz uso de apenas uma conformação (supostamente a conformação bioativa) para gerar modelos. Essa conformação representa a posição média da molécula no sítio de ligação, mas diz pouco sobre o comportamento dinâmico do composto. Diante dessa limitação e da necessidade de melhorar a qualidade das previsões do modelo QSAR 3D, Hopfinger e colaboradores (1997) desenvolveram o QSAR 4D, o qual considera a

dependência temporal em modelos QSAR. Esta metodologia utiliza um perfil de amostragem conformacional para cada ligante e o alinhamento considera tais perfis em vez de uma única conformação. A liberdade conformacional, obtida através de simulação de dinâmica molecular e as melhorias na representação das moléculas, no alinhamento e na estatística, resultaram em melhores previsões para o modelo QSAR (DAMALE et al., 2014). A geometria do receptor incorporada ao método deu origem ao QSAR 4D dependente do receptor, informação que permite que QSAR 4D siga a abordagem SBVS (PAN; TSENG; HOPFINGER, 2003). O método ainda subsidia uma equação mais simples para o desenvolvimento de modelos QSAR 3D, pois reduz o número de descritores (DAMALE et al., 2014).

Com auxílio do software denominado QUASAR, Vedani e Dobler (2002) adicionaram graus de liberdade – a quinta dimensão - para representar a topologia do ligante. O método permite um ajuste induzido da topologia do ligante ao sítio ativo do receptor.

O QSAR 6D permite uma avaliação simultânea de diferentes modelos de solvatação e possibilita uma simulação *in silico* de efeitos adversos provocados pelas drogas e produtos químicos (VEDANI; DOBLER; LILL, 2005). O QSAR 7D foi desenvolvido pela necessidade de inclusão de modelos baseados na dimensão do receptor e na interação virtual entre receptor–ligante (POLANSKI, 2009).

A incorporação de novas dimensões ao método QSAR (formalismo 3D, 4D, 5D, 6D e 7D) tem possibilitado uma melhor simulação da adaptação do sítio ativo de um alvo molecular à topologia do ligante, e vice-versa, condição que constitui um grande avanço na área de química medicinal, uma vez que a simulação de uma condição biológica *in silico* reduz custo e tempo, além de fornecer compostos mais promissores para as etapas subsequentes de P&D de fármacos.

No entanto, o QSAR clássico (formalismo 1D e 2D) não deve ser ignorado, pois os descritores físico-químicos, que representam os índices de conectividade, contagem de átomos, peso molecular, refratividade molar, polarizabilidade, hidrofobicidade, energias de orbitais, etc., não têm se mostrado inferiores a descritores 3D, já que o desempenho de previsão é frequentemente aceitável e os modelos são interpretáveis (BROWN; MARTIN, 1997; ESTRADA; MOLINA; PERDOMO-LOPEZ, 2001). Por fim, a identificação da conformação ativa, escolhidas após varredura conformacional de uma estrutura tridimensional, pode não corresponder à forma bioativa, visto que a menor energia de uma estrutura otimizada não reflete, necessariamente, a conformação bioativa do ligante; nessas condições, gerar descritores a partir das coordenadas tridimensionais de uma molécula livre pode ser uma tarefa árdua e desnecessária (FREITAS; RAMALHO, 2013).

2.5.1.1 Análise multivariada de imagem aplicada ao QSAR

A análise multivariada de imagens aplicada às relações quantitativas entre estrutura e atividade (*multivariate image analysis applied to quantitative structure-activity relationships* ou MIA- QSAR) (FREITAS; BROWN, MARTINS, 2005) é uma abordagem QSAR essencialmente bidimensional, que correlaciona desenhos de estruturas moleculares com a correspondente atividade biológica e/ou propriedades químicas e físico-químicas (CORMANICH; NUNES; FREITAS, 2012; FREITAS; BROWN; MARTINS, 2005; FREITAS, 2006). As principais vantagens dessa abordagem são: ser igualmente preditiva aos métodos QSAR mais sofisticados, tais como CoMFA e CoMSIA; ser uma técnica simples (uso de procedimentos e recursos computacionais amplamente conhecidos); apresentar baixo custo (pode-se utilizar programas de domínio público), que não envolve alinhamento e operações matemáticas complexas; e

proporcionar uma rápida caracterização de estruturas químicas (FREITAS; BROWN; MARTINS, 2005; FREITAS, 2006).

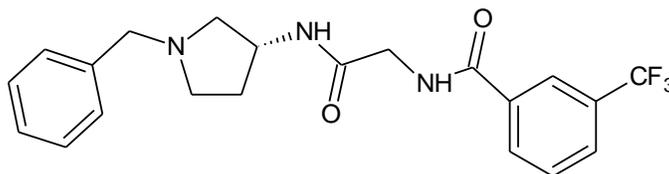
A abordagem MIA-QSAR baseia-se na premissa de que as imagens das estruturas químicas contêm informação relevante sobre o comportamento químico e, portanto, podem ser úteis na descrição das suas propriedades químicas, físicas e biológicas (FREITAS e DUARTE, 2015). Nesse sentido, o método MIA-QSAR utiliza pixels de imagens de estruturas químicas (desenhadas em programas de desenho de moléculas, tais como ChemDraw e ChemSketch) como descritores moleculares (FREITAS, 2006). Dessa forma, modificações estruturais ou alterações nas posições dos substituintes de cada estrutura química, resultam em alterações nas coordenadas dos pixels da imagem, e essas alterações explicam variações nas propriedades moleculares (CORMANICH; NUNES; FREITAS, 2012).

O primeiro modelo MIA-QSAR, atualmente denominado MIA-QSAR tradicional ou MIA-QSARt, foi construído a partir de imagens em preto e branco, formando um sistema binário de pixels. Como as cores podem ser tratadas numericamente pelo sistema RGB de cores, os pixels brancos equivalem ao número 765 (soma da contribuição máxima das cores vermelho, verde e azul, sendo cada canal equivalente a 255), enquanto os pixels pretos (ausência de cor) representados pelo número 0 correspondem a átomos e ligações químicas das moléculas (FREITAS; BROWN; MARTINS, 2005). A Figura 2 mostra a imagem de uma estrutura química desenhada para a abordagem MIA-QSARt.

É notório que o desenho correto da estrutura química na construção de um modelo MIA-QSAR é determinante para obtenção de descritores fiéis às características do conjunto de dados. Diferenciações na estrutura química definem a sequência de pixels e, conseqüentemente, a característica observada. Para tanto, o desenho deve obedecer a critérios como: manter a mesma disposição dos átomos, grupos substituintes e direção da cadeia carbônica para

todas as estruturas e utilizar uma única representação para tipos de substituintes (por exemplo: o grupo metila pode ser apresentado na forma de $-CH_3$ ou $-Me$,

Figura 2 - Estrutura química de um derivado R-3-amino-pirrolidinas desenhada no programa ChemSketch para construção de um modelo MIA-QSARt.



Fonte: Nunes e Freitas (2013).

(por exemplo: o grupo metila pode ser apresentado na forma de $-CH_3$ ou $-Me$, porem é necessário que apenas umas destas representações descrevam os grupos metila do conjunto de dados em estudo para uma dada posição) (GOODARZI; FREITAS; JENSEN, 2009).

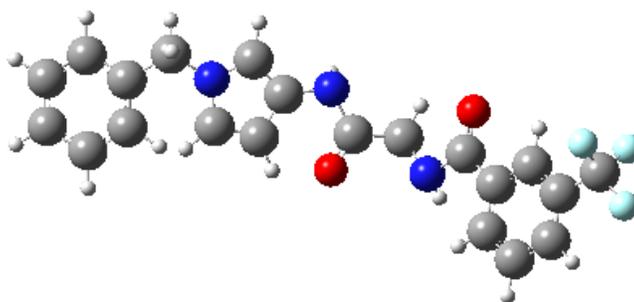
Os descritores MIA-QSAR são capazes de descrever informações químicas importantes, tais como efeitos estéricos (substituintes de moléculas orgânicas ocupando uma grande área no espaço dedicado ao desenho das estruturas) e centros estereogênicos (linhas em cunha ou tracejadas, para representar ligações para frente ou para trás relativas a um carbono quiral) (FREITAS; BROWN; MARTINS, 2005).

A capacidade de correlacionar pixels de imagens de estruturas químicas com atividade biológica pode ser averiguada em vários trabalhos (FREITAS, 2006; GOODARZI; FREITAS, 2010) e também para diferentes propriedades físicas, tais como temperatura de ebulição (GOODARZI; FREITAS, 2008), deslocamento químico (GOODARZI; FREITAS; RAMALHO, 2009) e perfil eletroforético (GOODARZI; FREITAS; 2009).

Apesar de os resultados de predição com a abordagem MIA-QSAR tradicional serem satisfatórios, a utilização de um sistema binário de pixels para representar as imagens das estruturas químicas apresenta uma desvantagem, uma vez que algumas informações contidas na estrutura química, tais como o raio atômico e a natureza dos substituintes, não são consideradas. Esse fato praticamente impossibilita a interpretação dos modelos construídos em termos dos tipos de átomos ou grupos funcionais responsáveis pela variação nas propriedades modeladas, como um passo preliminar à concepção racional de novas estruturas. Essa limitação motivou a incorporação de informações que descrevessem melhor as estruturas químicas. Assim, foi desenvolvida uma nova abordagem, denominada aug-MIA-QSAR (“aug” vem do inglês augmented) (NUNES; FREITAS, 2013).

Na abordagem aug-MIA-QSAR, os diferentes átomos são representados por círculos com tamanhos proporcionais aos respectivos raios de van der Waals, em diferentes cores (Figura 3) (NUNES; FREITAS, 2013).

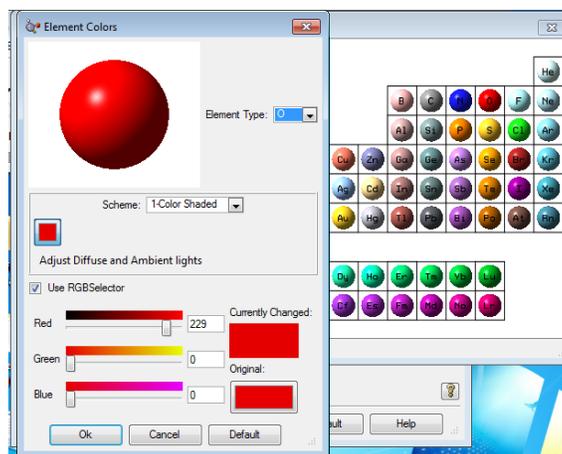
Figura 3 - Imagem da estrutura química de um derivado R-3-amino-pirrolidinas desenhado no programa GaussView para construção de um modelo aug-MIA-QSAR.



Fonte: Nunes e Freitas (2013).

Nesse sistema, enquanto a maior quantidade de pixels em uma área de trabalho relaciona-se com o tamanho molecular, a inclusão de cores (de acordo com o sistema de cores RGB) dá origem a uma variedade de outros pixels, além de 0 (pixels pretos) e 765 (pixels brancos), que permitem diferenciar os átomos e explicar variações nas propriedades atômicas (BARIGYE; FREITAS, 2016). Na abordagem aug-MIA-QSAR, as cores atribuídas para os diferentes tipos de átomos são formadas a partir da contribuição das componentes vermelho, verde e azul. A Figura 4 mostra a tabela de cores padrão do programa GaussView para os diferentes átomos.

Figura 4 - Cores padrão do programa GaussView 5.0.8 para os diferentes átomos



Fonte: Dennington, Keith e Millan (2008).

O aumento na capacidade preditiva da abordagem aug-MIA-QSAR em relação à abordagem MIA-QSAR pode ser verificada em diferentes estudos já realizados (DUARTE, et al., 2015; FREITAS; DUARTE, 2015). A diferenciação dos átomos por tamanho e cores traz informações relevantes sobre propriedades periódicas atômicas, tais como eletronegatividade, massa atômica, volume, etc. Todavia, a metodologia proposta por Nunes e Freitas (2013)

apresenta limitações quanto à interpretação dos descritores gerados, pois o efeito de iluminação dos átomos (efeito 3D) gera valores de pixels diferentes para um mesmo átomo. O uso de cores sólidas é, portanto, recomendado.

O método aug-MIA-QSAR, assim como as demais abordagens QSAR, exibem vantagens e desvantagens ao tentar compreender a relação estrutura/atividade. Portanto, cabe ao pesquisador procurar a técnica que melhor se adapta aos objetivos do estudo. Apesar das diferentes abordagens, a metodologia QSAR apresenta critérios e regras que são comuns a qualquer abordagem e o cumprimento dessas são fundamentais para obter um modelo realístico e confiável.

2.5.1.2 Construção de um modelo QSAR

A obtenção de um modelo de QSAR de boa qualidade depende de muitos fatores, tais como a qualidade dos dados biológicos, a escolha dos descritores e os métodos estatísticos adotados. Nesse sentido, a construção de um modelo QSAR pode ser genericamente dividida em quatro etapas: preparação de dados, análise de dados, construção e validação do modelo (CHERKASOV, 2014):

1 - Preparação dos dados

A preparação dos dados requer o conhecimento prévio do objeto de estudo e a busca de uma série congênere de compostos com similaridade estrutural que apresente atividade e/ou propriedade química experimentalmente conhecida e medidas conforme um mesmo protocolo de ensaio. Além de similaridade estrutural, o perfeito alinhamento das moléculas é um fator determinante para obtenção de um bom modelo. Deve-se, ainda, atentar para

possíveis erros estruturais e experimentais no conjunto de dados, assim como o tamanho do mesmo. É aconselhável que o conjunto de dados tenha, no mínimo, 20 compostos (MONTANARI, 2011).

2 - Análise dos dados

A análise de dados tem por objetivo a seleção de descritores que descrevam as características estruturais e físico-químicas mais relevantes associadas com a atividade biológica. Os descritores moleculares podem ser obtidos por diversos programas computacionais. Esses programas incorporam conceitos teóricos e gráficos que caracterizam a estrutura química dos compostos e, posteriormente, geram uma grande variedade de descritores. Após a obtenção dos descritores, procedimentos de pré-processamento de dados, tais como autoescalamento, padronização, normalização e centrar os dados na média podem ser utilizados quando o conjunto de dados apresenta variáveis de diferentes grandezas e variância (FERREIRA et al., 1999; DAMALE et al., 2014). Por fim, deve-se analisar a uniformidade do conjunto de dados quanto a presença de compostos que apresentem comportamento anômalo aos demais. Também conhecidos como *outlier*, os compostos anômalos podem influenciar a estabilidade do modelo e, portanto, os mesmos devem ser cuidadosamente avaliados e, quando não prejudicar a capacidade de previsão do modelo, devem-se retirá-los do conjunto de dados (GAUDIO; ZANDONADE, 2001). A detecção de compostos *outliers* pode ser avaliada por procedimentos de reconhecimento de padrão, tais como PCA e HCA, ou através da análise de *leverages* e resíduos de *Student* (MARTINS; FERREIRA, 2013).

Com o objetivo de selecionar os melhores descritores para correlacionar com a atividade biológica e, considerando que modelos QSAR normalmente geram milhares de descritores, procedimentos de seleção de variáveis podem ser

necessários para excluir variáveis redundantes ou que contenham informações moleculares não relevantes para descrever o fenômeno sob investigação, tais como aquelas com baixa variância e/ou alta correlação. O processo de seleção consiste em encontrar combinações de k variáveis, dentre as m disponíveis, capazes de produzir modelos matemáticos que descrevam adequadamente os valores observados da atividade biológica (FERREIRA; MONTANARY; GAUDIO, 2002). A seleção de variáveis, além de resolver problemas de alta dimensionalidade e correlação, também traz o benefício da estabilidade do modelo, facilita a interpretação das relações entre os descritores e atividade biológica, e pode ainda tornar o modelo quimicamente mais interpretável (FERREIRA; MONTANARY; GAUDIO, 2002)

A literatura disponibiliza uma variedade de métodos para seleção de variáveis e, segundo Guyon e Elisseeff (2003), esses podem ser divididos em filtros, envoltórios (*wrappers*) e métodos embutidos (*embedded methods*). Os métodos envoltórios e embutidos utilizam algoritmos de aprendizagem em suas definições. Os métodos baseados em filtros, por sua vez, selecionam um subconjunto de variáveis antes da modelagem propriamente dita, ou seja, funcionam como pré-processamento dos dados. Além dessa classificação, os métodos de seleção podem ser classificados em supervisionados e não supervisionados (URIAS et al., 2015). Os métodos supervisionados selecionam variáveis baseados em algum critério pré-definido, tais como valores da variável dependente (Y). Já os métodos não supervisionados trabalham apenas com o conjunto de variáveis independentes (X) (KLEBE; ABRAHAM; MIETZNER, 1994). Dentre os métodos comumente utilizados, o ganho de informação (IG), coeficientes de correlação de Pearson, relação de Fisher, regressão por mínimos quadrados parciais (PLS, do inglês, *partial least squares*) são definidos como supervisionados; já os métodos de análise de agrupamento hierárquico, análise de componente principal (PCA do inglês, *principal component analysis*),

entropia de Shannon, etc. são considerados não supervisionados (URIAS et al., 2015). Dada à diversidade de métodos, cabe ao pesquisador definir aquele que melhor traduz as informações contidas em seu conjunto de dados. Mais informações sobre os métodos utilizados nesta tese estarão descritas nos capítulos subsequentes.

Além dos procedimentos de seleção de variáveis baseados em filtros, envoltórios e métodos embutidos, existem disponíveis técnicas de busca e otimização de variáveis. O algoritmo genético (GA) é uma das técnicas de maior aplicação em estudos de QSAR. O algoritmo genético é inspirado na teoria de Darwin sobre a evolução das espécies, onde processos de cruzamento, seleção natural e mutação genética podem fornecer populações mais evoluídas (FERREIRA; MONTANARY; GAUDIO, 2002). O método algoritmo genético foi proposto por John Holland na década de 1960 (GOLDBERG; HOLLAND, 1988) e revela-se útil para otimização de modelos que descrevam relações entre informações coletadas de um sistema (espectros de substâncias, por exemplo) e valores de propriedades (respostas do sistema), resolução de problemas envolvendo relações, capacidade de planejamento e tomada de decisões e seleção de subconjuntos de dados, etc. (LUCASIUS; KATEMAN, 1994).

O processo GA envolve 4 etapas: iniciação, seleção, cruzamento e mutação. Na etapa de iniciação, as variáveis pertencentes à matriz de dados (população) são agrupadas aleatoriamente, formando os cromossomos. Esse agrupamento é definido por equações de regressão e o valor de R (coeficiente de correlação) e RMSECV (raiz do erro quadrado médio da validação cruzada) são utilizados para definir os melhores cromossomos (seleção). Na etapa de cruzamento, ou *crossover*, as variáveis (genes) dos cromossomos são cruzadas buscando melhores modelos. Posteriormente ao cruzamento, processos de mutação ocorrem para que todas as variáveis tenham chance de participarem do processo. Concluídas essas etapas, o processo é finalizado quando é

preestabelecido um número de gerações ou quando o processo apresenta alguma porcentagem de cópias de um mesmo cromossomo. A qualidade dos modelos, geralmente, é avaliada pelos parâmetros estatísticos: coeficiente de correlação (R), o fator F e o nível de preditividade (Q^2) (FERREIRA; MONTANARY; GAUDIO, 2002; TEÓFILO, 2013).

O método GA envolve algumas limitações quanto a sua utilização. O fato de usar processos aleatórios na etapa de iniciação é impraticável à reprodução do modelo (GOICOCHEA; OLIVIERI, 2003). Variáveis altamente correlacionadas codificadas em pontos adjacentes podem impedir o GA de remover variáveis indesejadas, especialmente quando o *crossover* simples é empregado. Bons modelos não resultam, necessariamente, em uma boa validação externa e grandes conjuntos de dados podem tornar inviável a execução do GA (GOICOCHEA; OLIVIERI, 2003; TEÓFILO, 2013).

3 - Construção do modelo de regressão QSAR

A construção do modelo QSAR envolve a geração de um modelo matemático em que a atividade biológica e os descritores moleculares são correlacionados por meio de, na maioria dos casos, equações de regressão linear. Na grande maioria dos casos, o número de variáveis descritivas é marcadamente superior ao número de variáveis experimentais e, por essa razão, métodos de calibração multivariada devem ser utilizados. A equação 1 expressa a relação entre as propriedades dos compostos investigados, denominadas variáveis independentes ou descritores (\mathbf{X}), com sua variável dependente ou bioatividade (\hat{Y}), por meio de um vetor de regressão (GAUDIO; ZANDONADE, 2001). Os métodos mais empregados em estudos QSAR são a regressão linear múltipla (MLR, do inglês *multiple linear regression*), regressão por componentes

principais (PCR, do inglês, *principal component regression*) e regressão por mínimos quadrados parciais (PLS, do inglês, *partial least squares*).

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_k X_k \dots \quad (1)$$

A regressão linear múltipla (MLR) foi, durante muito tempo, o método de regressão multivariada mais utilizada em estudos QSAR (MARTINS e FERREIRA, 2013). Esse método tem grande aplicabilidade devido à sua simplicidade e facilidade de interpretação dos resultados. No entanto, é limitado no sentido de que o número de variáveis deve ser menor ou igual ao número de compostos, para que a equação da regressão tenha uma única solução (FERREIRA; MONTANARY; GAUDIO, 2002). Outra limitação da MLR é a colinearidade: variáveis altamente correlacionadas podem tornar os coeficientes de regressão instáveis e sem significado. Portanto, é adequado que as variáveis tenham baixos valores de correlação entre si (<0,7) (MARTINS; FERREIRA, 2013). Contudo, a utilização de procedimentos de seleção de variáveis pode reduzir a dimensionalidade das matrizes de dados e resolver o problema de colinearidade. A literatura ainda recomenda que, para evitar correlação por coincidência, apenas uma variável explicativa deva ser incluída no modelo QSAR para cada 5 ou 6 compostos (TOPLISS; COSTELLO, 1972; GAUDIO; ZANDONADE, 2001).

Diferentemente do método de regressão MLR, os métodos PLS e PCR não afetados pela correlação e pela dimensionalidade da matriz de dados. Tais abordagens transformam a matriz de descritores \mathbf{X} em uma nova matriz \mathbf{X}' , projetando as variáveis originais em um espaço de menor dimensão, formado por variáveis não colineares que codificam toda a informação contida na matriz de dados original (MARTINS, 2010). Esta transformação permite lidar com informações irrelevantes, interferentes e colinearidade, bem com a alta

dimensionalidade dos dados (FERREIRA et al., 1999). Isso acontece porque os métodos de PCR e PLS buscam encontrar as combinações lineares das variáveis originais mais relevantes para o modelo. Dessa forma, toda informação irrelevante contida na matriz de dados é descartada, resolvendo o problema de colinearidade e gerando equações de regressão mais objetivas. Além disso, a regressão feita nos *scores* elimina o problema de alta dimensionalidade (GELADI; KOWALSKI, 1986)

No método PCR, as novas variáveis são chamadas de componentes principais e, no método PLS, são denominadas variáveis latentes. As diferentes denominações se devem ao fato de que, na PCR, utiliza-se exclusivamente a variável X para a construção das componentes principais. Na regressão PLS, por sua vez, a construção da componente principal é influenciada pela variável Y (FERREIRA; MONTANARY; GAUDIO, 2002). O número ótimo de variáveis latentes (LV) no modelo é comumente determinado pela avaliação do erro de predição na validação interna cruzada. Essa metodologia é aplicada, pois os métodos de projeção produzem modelos tendenciosos e é necessário evitar o superajuste do modelo (*overfitting*) (MARTINS; FERREIRA, 2013). Embora os métodos PCR e PLS apresentem resultados similares, o método PLS é mais robusto; isso significa que os parâmetros do modelo não se alteram de maneira significativa quando novas amostras são acrescentadas ou retiradas do conjunto de calibração (GELADI; KOWALSKI, 1986; MARTINS; FERREIRA, 2013).

4 - Validação do modelo QSAR

A validação do modelo QSAR é realizada com base em parâmetros de validação interna e externa. Para esse fim, o conjunto de dados é separado em grupo treinamento e teste. Os compostos presentes no grupo treinamento são utilizados na calibração do modelo, enquanto os compostos do grupo teste são

utilizados para validar o modelo QSAR (TROPSHA, 2010). Além de calibrar o modelo, a validação interna é utilizada para determinar o número de componentes principais (na PCR) ou LV (na PLS), além de identificar compostos anômalos. Já a validação externa é utilizada exclusivamente para avaliar a capacidade preditiva e a confiabilidade do modelo (GOLBRAIKH; TROPSHA, 2002). A validação do modelo QSAR requer a utilização de parâmetros estatísticos que certifiquem uma boa calibração, robustez, capacidade preditiva e a confiabilidade do modelo.

O coeficiente de determinação r^2 é um parâmetro que mede a capacidade de calibração do conjunto de dados do modelo e é determinado pela equação 2, onde y_i corresponde ao valor experimental da atividade biológica, \hat{y}_i são os valores da atividade biológica predita e \bar{y} corresponde o valor médio da atividade biológica experimental. Esse parâmetro é considerado aceitável quando apresenta valores superiores a 0,7 (GRAMATICA, 2007).

$$r^2 = 1 - [\Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2] \quad (2)$$

O coeficiente de determinação da validação cruzada (r^2_{cv} ou q^2) é um parâmetro da validação interna importante para se ter uma ideia da capacidade preditiva e da robustez do modelo construído. Esse parâmetro também é determinado pela equação 2. Porém, uma molécula (no caso da validação cruzada *leave-one-out*) ou grupo de moléculas (validação cruzada *leave-n-out*) é retirado do conjunto treinamento; em seguida, realiza-se a regressão e, após esse procedimento, a atividade biológica é prevista para a molécula (ou grupo de moléculas) que permaneceu de fora do modelo de calibração. Este procedimento é repetido até que todas as moléculas tenham ficado de fora do conjunto de calibração pelo menos uma vez. Como resultado desse procedimento, obtém-se o coeficiente de determinação da validação cruzada, r^2_{cv} . Valor de r^2_{cv} superior

a 0,5 é indicativo de um modelo robusto e preditivo (GOLBRAIKH; TROPSHA, 2002). A confiabilidade do modelo também pode ser avaliada pela diferença entre r^2_{cal} e r^2_{cv} . Interpretações confiáveis podem ser atingidas quando a diferença entre esses dois parâmetros não exceder a 0,3, uma diferença maior pode indicar um ajuste forçado, presença de variáveis independentes irrelevantes ou de *outliers* no conjunto de dados (ERIKSSON; JOHANSSON, 1996).

A validação externa é considerada o parâmetro mais seguro para atestar o caráter preditivo de um modelo QSAR (GOLBRAIKH; TROPSHA, 2002), uma vez que é utilizada para prever as atividades de compostos que não participaram da construção do modelo. Segundo esse processo, os parâmetros de regressão obtidos do modelo construído com as amostras do grupo de treinamento são utilizados para calcular as atividades biológicas dos compostos do grupo teste. Como os valores experimentais de tais moléculas são conhecidos, é possível comparar os valores experimentais aos preditos por meio de parâmetros estatísticos similares aos da validação cruzada (Equação 3). Porém, nesse caso, \hat{y}_i e y_i passam a ser, respectivamente, o valor real e valor predito de atividade biológica para o composto i do conjunto teste. A literatura recomenda que bons modelos QSAR devam apresentar valores de r^2_{test} superiores a 0,5.

O coeficiente de determinação obtido após a randomização do bloco das variáveis dependentes \mathbf{y} (r^2_{rand}) é um parâmetro utilizado para averiguar se os valores de r^2_{cal} e r^2_{cv} não foram ajustados por acaso, ou seja, se tem uma verdadeira conexão com a variável resposta (\mathbf{y}). Segundo esse procedimento, o bloco das variáveis \mathbf{y} é aleatorizado e um novo modelo QSAR é desenvolvido utilizando a matriz de variáveis independentes originais (\mathbf{X}). O processo é repetido várias vezes, sendo esperado valor r^2_{rand} significativamente menor que

r^2_{cal} . Essa, portanto, é uma técnica a qual assegura que o modelo não foi obtido pelo acaso (RÜCKER; RÜCKER; MERINGER, 2007).

Para certificar que o valor de r^2_{rand} é significativamente inferior ao r^2_{cal} e que o modelo pode ser considerado confiável, Mitra e colaboradores (2010) propuseram a utilização do coeficiente de determinação penalizado e corrigido ${}^c r^2_p$ (Equação 3), o qual avalia se a diferença entre os parâmetros r^2_{cal} e r^2_{rand} é estatisticamente relevante. Valores de ${}^c r^2_p$ superiores a 0,5 são aceitáveis.

$${}^c r^2_p = r \times (r^2_{cal} - r^2_{rand})^{1/2} \quad (3)$$

Durante muito tempo, os parâmetros r^2_{cal} , r^2_{cv} , r^2_{test} e r^2_{rand} foram considerados suficientes para avaliar a capacidade preditiva de modelos QSAR. No entanto, estudos recentes apontam que outros parâmetros são necessários para evidenciar a confiabilidade dos modelos QSAR (GRAMATICA, 2007). Recentemente, Ojah et al. (2011) propuseram o cálculo do coeficiente de determinação modificado r^2_m (equação 4) para comprovar que os resultados da validação externa não somente exibem boa correlação entre os valores experimentais e preditos, mas também que ambos apresentam valores similares. Para que um modelo seja considerado preditivo, r^2_m deve apresentar valores superiores a 0,5.

$$r^2_m = r^2 [1 - (r^2 - r_0^2)^{1/2}] \quad (4)$$

Em que r^2 e r_0^2 correspondem aos valores do coeficiente de correlação ao quadrado entre os valores de atividade observados e preditos para o conjunto teste com e sem intercepto, respectivamente.

REFERÊNCIAS

- ARROIO, A.; HONÓRIO, K. M.; SILVA, A. B. Propriedades químico-quânticas empregadas em estudos das relações estrutura-atividade. **Química Nova**, São Paulo, v. 33, n. 3, p. 694-699, Mar. 2010.
- BARIGYE, S. J.; FREITAS, M. P. Ten Years of the MIA-QSAR Strategy: Historical Development and Applications. **International Journal of Quantitative Structure-Property Relationships (IJQSPR)**, Índia, v. 1, n. 1, p. 64-77, Jan./June 2016.
- BARREIRO, E. J. Estratégia de simplificação molecular no planejamento racional de fármacos: a descoberta de novo agente cardioativo. **Química Nova**, São Paulo, v. 25, n. 6/B, p. 1172-1180, Dez, 2002.
- BARREIRO, E. J. et al. Molecular modeling: a tool for rational drug design in medicinal chemistry. **Química Nova**, São Paulo, v. 20, n. 3, p. 300-310, Mai./June 1997.
- BRASIL, Ministério da Saúde. **Doenças infecciosas e parasitárias: Guia de Bolso** 3 ed., Brasília: Ed. MS, Vol. II, 2004, 334p.
- BRASIL. Ministério da Saúde. Doenças negligenciadas: estratégias do Ministério da Saúde. **Revista Saúde Pública**, São Paulo, v. 44, n. 1, p. 200-202, 2010.
- BROWN, R. D.; MARTIN, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. **Journal of Chemical Information and Computer Sciences**, Washington, v. 37, n. 1, p. 1-9, Jan./Feb. 1997.
- CAMARGO, E. P. Doenças tropicais. **Estudos Avançados**, São Paulo, v. 22, n. 64, Dez. 2008.
- CARVALHO, I. et al. Introdução a modelagem molecular de fármacos no curso experimental de química farmacêutica. **Química Nova**, São Paulo, v. 26, n. 3, p. 428-438, Mar. 2003.
- CHERKASOV, A. et al. QSAR modeling: where have you been? Where are you going to? **Journal of Medicinal Chemistry**, Washington, v. 57, n. 12, p. 4977-5010, June 2014.

CHIRAC, P.; TORREELE, E. Global framework on essential health R&D. **The Lancet**, London, v. 367, n. 9522, p. 1560-1561, May 2006.

CORMANICH, R. A.; NUNES, C. A.; FREITAS, M. P. Chemical drawings correlate to biological properties: MIA-QSAR. **Química Nova**, São Paulo, v. 35, n. 6, p. 1157-1163, June 2012.

CRAMER, R. D.; PATTERSON, D. E.; BUNCE, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. **Journal of the American Chemical Society**, Easton, v. 110, n. 18, p. 5959-5967, Aug. 1988.

DAMALE, M. G. et al. Recent advances in multidimensional QSAR (4D-6D): a critical review. **Mini Reviews in Medicinal Chemistry**, Sharjah, v. 14, n. 1, p. 35-55, Jan. 2014.

DENNINGTON, R. D.; KEITH, T.A; MILLAM, J.M. **GaussView 5.0. 8**; Gaussian Inc., Wallingford, 2008. CD-ROM.

DIAS, L. C. et al. Doenças tropicais negligenciadas: uma nova era de desafios e oportunidades. **Química Nova**, São Paulo, v. 36, n. 10, p. 1552-1556, Dec. 2013.

DIMASI, J. A.; HANSEN, R. W.; GRABOWSKI, H. G. The price of innovation: new estimates of drug development costs. **Journal of Health Economics**, Amsterdam, v. 22, n. 2, p. 151-185, Mar. 2003.

DIMITRI, N. An assessment of R&D productivity in the pharmaceutical industry. **Trends in Pharmacological Sciences**, London, v. 32, n. 12, p. 683-685, Dec. 2011.

DUARTE, M. et al. Computational modelling of the antischistosomal activity for neolignan derivatives based on the MIA-SAR approach. **SAR and QSAR in Environmental Research**, Oxon, v. 26, n. 3, p. 205-216, Mar. 2015.

ERIKSSON, L.; JOHANSSON, E. Multivariate design and modeling in QSAR. **Chemometrics and Intelligent Laboratory Systems**, Berlin, v. 34, n. 1, p. 1-19, Aug. 1996.

ESTRADA, E.; MOLINA, E.; PERDOMO-LÓPEZ, I. Can 3D structural parameters be predicted from 2D (topological) molecular descriptors? **Journal**

of **Chemical Information and Computer Sciences**, Washington, 41, 1015-1021, Jul./Aug. 2001.

FERREIRA, M. et al. Chemometrics I: multivariate calibration, a tutorial. **Química Nova**, São Paulo, v. 22, n. 5, p. 724-731, 1999.

FERREIRA, M. M. C.; MONTANARI, C. A.; GAUDIO, A. C. Variable selection in QSAR. **Química Nova**, São Paulo, v. 25, n. 3, p. 439-448, Sept. 2002.

FERREIRA, R. S.; OLIVA, G.; ANDRICOPULO, A. D. Integração das técnicas de triagem virtual e triagem biológica automatizada em alta escala: oportunidades e desafios em P&D de fármacos. **Química Nova**, São Paulo, v. 34, n. 10, p. 1770-1778, Dec. 2011.

FREITAS, M. P. MIA-QSAR modelling of anti-HIV-1 activities of some 2-amino-6-arylsulfonylbenzotriazoles and their thio and sulfinyl congeners. **Organic & Biomolecular Chemistry**, Cambridge, v. 4, n. 6, p. 1154-1159, Feb. 2006.

FREITAS, M. P.; BROWN, S. D.; MARTINS, J. A. MIA-QSAR: a simple 2D image-based approach for quantitative structure-activity relationship analysis. **Journal of Molecular Structure**, Amsterdam, v. 738, n. 1, p. 149-154, Mar. 2005.

FREITAS, M. P.; DUARTE, M. H. Evolution of Multivariate Image Analysis in QSAR: The Case for. In. **Quantitative Structure-Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment**, Kolkata, p. 84-122, Jan. 2015.

FREITAS, M.; RAMALHO, T. **Princípios de estrutura eletrônica e orbitais em química orgânica**. Lavras: Editora da UFLA, 2013. 129p.

GARKANI-NEJAD, Z.; AHMADI-ROUDI, B. Modeling the antileishmanial activity screening of 5-nitro-2-heterocyclic benzylidene hydrazides using different chemometrics methods. **European Journal of Medicinal Chemistry**, Paris, v. 45, n. 2, p. 719-726, Feb. 2010.

GAUDIO, A. C.; ZANDONADE, E. Proposition, validation and analysis of QSAR models. **Química Nova**, São Paulo, v. 24, n. 5, p. 658-671, Sept./Oct. 2001.

GELADI, P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. **Analytica Chimica Acta**, Amsterdam, v. 185, p. 1-17, July 1986.

GOICOECHEA, H. C.; OLIVIERI, A. C. A new family of genetic algorithms for wavelength interval selection in multivariate analytical spectroscopy. **Journal of Chemometrics**, Malden, v. 17, n. 6, p. 338-345, Aug. 2003.

GOLBRAIKH, A.; TROPSHA, A. Beware of q^2 ! **Journal of Molecular Graphics and Modelling**, New York, v. 20, n. 4, p. 269-276, Jan. 2002.

GOLDBERG, D. E.; HOLLAND, J. H. Genetic algorithms and machine learning. **Machine learning**, Dordrecht, v. 3, n. 2, p. 95-99, Sept. 1988.

GONZALEZ, A. J. V. **A indústria farmacêutica brasileira na década de 90: mudanças na pesquisa & desenvolvimento, na produção de fármacos e de medicamentos**. 1999. 121p. Dissertação (Mestrado em Política científica). Universidade Estadual de Campinas, Instituto de Geociências. Campinas-São Paulo, 1999.

GOODARZI, M.; FREITAS, M. P. MIA-QSAR modelling of activities of a series of AZT analogues: bi-and multilinear PLS regression. **Molecular Simulation**, Oxon, v. 36, n. 4, p. 267-272, Sept. 2010.

GOODARZI, M.; FREITAS, M. P. Prediction of electrophoretic enantioseparation of aromatic amino acids/esters through MIA-QSPR. **Separation and Purification Technology**, Amsterdam, v. 68, n. 3, p. 363-366, Aug. 2009.

GOODARZI, M.; FREITAS, M. P.; JENSEN, R. Feature selection and linear/nonlinear regression methods for the accurate prediction of glycogen synthase kinase -3 beta inhibitory activities. **Journal of Chemistry Information and Modeling**, Washington, v. 49, n. 4, p. 824-832, Apr. 2009.

GOODARZI, M.; FREITAS, M. P.; RAMALHO, T. C. Prediction of ^{13}C chemical shifts in methoxyflavonol derivatives using MIA-QSPR. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, Kidlington, v. 74, n. 2, p. 563-568, Oct. 2009.

GOODARZI, M.; FREITAS, M. P. Predicting Boiling Points of Aliphatic Alcohols through Multivariate Image Analysis Applied to Quantitative Structure- Property Relationships. **The Journal of Physical Chemistry A**, Washington, v. 112, n. 44, p. 11263-11265, Nov. 2008.

GRAMATICA, P. Principles of QSAR models validation: internal and external. **QSAR and Combinatorial Science**, Weinheim, v. 26, n. 5, p. 694, May 2007.

GUIDO, R. V.; ANDRICOPULO, A. D.; OLIVA, G. Planejamento de fármacos, biotecnologia e química medicinal: aplicações em doenças infecciosas. **Estudos Avançados**, São Paulo, v. 24, n. 70, p. 81-98, Set./Dez. 2010.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **The Journal of Machine Learning Research**, Edinburgh, v. 3, p. 1157-1182, Mar. 2003.

HAJDUK, P. J.; GREER, J. A decade of fragment-based drug design: strategic advances and lessons learned. **Nature Reviews Drug Discovery**, London, v. 6, n. 3, p. 211-219, Mar. 2007

HANSCH, C. et al. Chem-bioinformatics: comparative QSAR at the interface between chemistry and biology. **Chemical Reviews**, Washington, v. 102, n. 3, p. 783-812, Mar. 2002.

HANSCH, C.; FUJITA, T. ρ - σ - π Analysis. A method for the correlation of biological activity and chemical structure. **Journal of the American Chemical Society**, Easton, v. 86, n. 8, p. 1616-1626, Nov. 1964.

HOPFINGER, A. et al. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. **Journal of the American Chemical Society**, Washington, v. 119, n. 43, p. 10509-10524, Oct. 1997.

HOTEZ, P. A new voice for the poor. **PLoS Neglected Tropical Disease**, San Francisco, v. 1, n. 1, p. e77, Oct. 2007.

HOTEZ, P. J. et al. Incorporating a rapid-impact package for neglected tropical diseases with programs for HIV/AIDS, tuberculosis, and malaria. **PLoS Medicine**, San Francisco, v. 3, n. 5, p. 576, Sept. 2007.

JONES, K. E. et al. Global trends in emerging infectious diseases. **Nature**, London, v. 451, n. 7181, p. 990-993, Feb. 2008.

KAPETANOVIC, I. Computer-aided drug discovery and development (CADD): in silico-chemico-biological approach. **Chemico-biological Interactions**, Ireland, v. 171, n. 2, p. 165-176, Jan. 2008.

KLEBE, G.; ABRAHAM, U.; MIETZNER, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. **Journal of Medicinal Chemistry**, Washington, v. 37, n. 24, p. 4130-4146, Nov. 1994.

LANG, M. et al. Discovery and hit-to-lead optimization of novel allosteric glucokinase activators. **Bioorganic & Medicinal Chemistry Letters**, Oxford, v. 21, n. 18, p. 5417-5422, Sept. 2011.

LIMA, L. M. Química Medicinal Moderna: desafios e contribuição brasileira. **Química Nova**, São Paulo, v. 30, n. 6, p. 1456, Nov./Dez.2007.

LUCASIUS, C. B.; KATEMAN, G. Understanding and using genetic algorithms Part 2. Representation, configuration and hybridization. **Chemometrics and Intelligent Laboratory Systems**, Berlin, v. 25, n. 2, p. 99-145, Nov.1994.

MARTINS, J. P. A. **Química computacional aplicada a QSAR**. São Paulo: CENAPAD, 2010. Disponível em: <<http://www.cenapad.unicamp.br/serviços/treinamentos/apostilas/apostila-QuimComp.pdf>>. Acesso em: 14/03/2015

MARTINS, J. P. A.; FERREIRA, M. M.C. *QSAR modeling*: um novo pacote computacional *open source* para gerar e validar modelos QSAR. **Química Nova**, São Paulo, v. 36, n. 4, p. 554-560, Mai. 2013.

MÉDICOS SEM FRONTEIRAS- MSF. Fatal imbalance: the crisis in research and development for drugs for neglected diseases. Geneva: MSF, 2001. 21p. Disponível em: <<http://www.dndi.org/2001/media-centre/scientific-articles/scientific-articles-neglected-diseases/fatal-imbalance-the-crisis-in-research-and-development-for-drugs-for-neglected-diseases/>>. Acesso em: 20/03/2015.

MILNE, G. M. Pharmaceutical productivity—the imperative for new paradigms. **Annual Reports in Medicinal Chemistry**, San Diego, v. 38, p. 383-396, Dec. 2003.

MITRA, I.; SAHA, A.; ROY, K.. Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. **Molecular Simulation**, Oxon, v. 36, n. 13, p. 1067-1079, Apr. 2010.

MONTANARI, C. A. **Química Medicinal**: métodos e fundamentos em planejamento de fármacos. São Paulo: EDUSP, 2011. 760p.

MORGON, N. H. Computação em química teórica: informações técnicas. **Química Nova**, São Paulo, v. 24, n. 5, p. 676-682, June 2001.

MUNOS, B. Lessons from 60 years of pharmaceutical innovation. **Nature Reviews Drug Discovery**, London, v. 8, n. 12, p. 959-968, Dec. 2009.

MURRAY, C. J.; LOPEZ, A. D. Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study. **The Lancet**, London, v. 349, n. 9064, p. 1498-1504, May 1997.

NEVES, P. J. D. et al. Top-um programa de cálculo de descritores topológicos para uso em correlações entre estrutura e atividade. **Química Nova**, São Paulo, v. 21, n. 6, p. 709, July 1998.

NUNES, C. A.; FREITAS, M. P. Introducing new dimensions in MIA-QSAR: a case for chemokine receptor inhibitors. **European Journal of Medicinal Chemistry**, Paris, v. 62, p. 297-300, Apr. 2013.

OJHA, P. K. et al. Further exploring r^2 metrics for validation of QSPR models. **Chemometrics and Intelligent Laboratory Systems**, Berlin, v. 107, n. 1, p. 194-205, May 2011.

ORGANIZAÇÃO MUNDIAL DE SAÚDE. **Neglected Tropical diseases**. Disponível em: <http://www.who.int/neglected_diseases/diseases/en/>. Acesso em: 10/01/2015a

ORGANIZAÇÃO MUNDIAL DE SAÚDE. **Tropical diseases**. Disponível em: <http://www.who.int/topics/tropical_diseases/en/,2015b>. Acesso em: 10/01/2015b

ORGANIZAÇÃO MUNDIAL DE SAÚDE. **Working to overcome the global impact of neglected tropical diseases: first WHO report on neglected tropical diseases**. Genebra, 2010. 184p. Disponível em: <http://apps.who.int/iris/bitstream/10665/44440/1/9789241564090_eng.pdf>. Acesso em: 10/01/2015

PAN, D.; TSENG, Y.; HOPFINGER, A. J. Quantitative structure-based design: formalism and application of receptor-dependent RD-4D-QSAR analysis to a set of glucose analogue inhibitors of glycogen phosphorylase. **Journal of Chemical Information and Computer Sciences**, Washington, v. 43, n. 5, p. 1591-1607, Sept./Oct. 2003.

PEDRIQUE, B. et al. The drug and vaccine landscape for neglected diseases (2000–11): a systematic assessment. **The Lancet Global Health**, Oxon, v. 1, n. 6, p. 371-379, Dec. 2013.

PhRMA. Biopharmaceutical Research & development: The process behind new medicines. 2015. Disponível em: < http://www.phrma.org/sites/default/files/pdf/rd_brochure_022307.pdf>. Acesso em 27/10/2015.

POLANSKI, J. Receptor dependent multidimensional QSAR for modeling drug-receptor interactions. **Current Medicinal Chemistry**, Sharjah, v. 16, n. 25, p. 3243-3257, Sept. 2009.

PONDER, E. L. et al. Computational models for neglected diseases: gaps and opportunities. **Pharmaceutical Research**, New York, v. 31, n. 2, p. 271-277, Feb. 2014.

RESS, P. Big pharma learns how to love information technology. **Scientific Computing World**, Cambridge, p.16-18, Jul./Aug. 2003. disponível em: <http://www.scientific-computing.com/features/feature.php?feature_id=111> . Acesso em: 28 /02/2015

RODRIGUES, R. P. et al. Estratégias de Triagem Virtual no Planejamento de Fármacos. **Revista Virtual de Química**, Niterói, v. 4, n. 6, p. 739-776, Dez. 2012.

RÜCKER, C.; RÜCKER, G.; MERINGER, M. γ -Randomization and its variants in QSPR/QSAR. **Journal of Chemical Information and Modeling**, Washington, v. 47, n. 6, p. 2345-2357, Nov./Dec. 2007.

SOUZA, W. D. **Doenças Negligenciadas**. Rio de Janeiro: Acadêmica Brasileira de Ciências, 2010. 56p.

TEÓFILO, R.F. **Métodos Quimiométricos: Uma Visão Geral - Conceitos básicos de quimiometria**, Viçosa: Universidade Federal de Viçosa, Vol. 1, 2013, 118p. Disponível em: <<http://www.deq.ufv.br/area/publicacao/26>>. Acesso em: 05/04/2015

TOPLISS, J. G.; COSTELLO, R. J. Chance correlations in structure-activity studies using multiple regression analysis. **Journal of Medicinal Chemistry**, Washington, v. 15, n. 10, p. 1066-1068, Oct. 1972.

TROPSHA, A. Best practices for QSAR model development, validation, and exploitation. **Molecular Informatics**, Weinheim, v. 29, n. 6-7, p. 476-488, July 2010.

URIAS, R. W. P. et al. IMMAN: free software for information theory-based chemometric analysis. **Molecular Diversity**, Dordrecht, v. 19, n. 2, p. 305-319, May 2015.

VALVERDE, E. **Doenças Negligenciadas**. 2015. Disponível em:
< <http://www.agencia.fiocruz.br/doen%C3%A7as-negligenciadas>>.
Acesso em 20/03/2015

VEDANI, A.; DOBLER, M. 5D-QSAR: the key for simulating induced fit? **Journal of Medicinal Chemistry**, Washington, v. 45, n. 11, p. 2139-2149, May 2002.

VEDANI, A.; DOBLER, M.; LILL, M. A. Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. **Journal of Medicinal Chemistry**, Washington v. 48, n. 11, p. 3700-3703, June 2005.

VERONESI, R. **Doenças infecciosas e parasitárias**. 8ªed., Rio de Janeiro: Ed. G. Koogan, 1991. 1082p.

WANG, W.; WANG, L.; LIANG, Y.-S. Susceptibility or resistance of praziquantel in human schistosomiasis: a review. **Parasitology Research**, New York, v. 111, n. 5, p. 1871-1877, Nov. 2012.

CAPÍTULO 2. MIA-QSAR: EXPLORANDO A ATIVIDADE ANTITRIPANOSSÔMICA DE DERIVADOS TIOSSEMICARBAZONAS E SEMICARBAZONAS

RESUMO

A análise multivariada de imagem aplicada à relação quantitativa estrutura-atividade (MIA-QSAR) é uma ferramenta computacional simples, de baixo custo e de fácil acesso, que vem se aprimorando em relação à análise de dados e ao desenvolvimento de informações mais descritivas desde a sua criação, em 2005. Nesse sentido, o presente estudo tem por objetivo comparar a abordagem MIA-QSARt com as novas dimensões aug-MIA-QSAR e aug-MIA-QSARcolor ao aplicá-las a um conjunto de dados com atividade antitripanossômica. Para a execução desse estudo, um conjunto de dados formado por 53 compostos derivados de tiossemicarbazonas e semicarbazonas com atividade inibitória contra a proteína cruzaina (*T. cruzi* cathepsin L-like) foi modelado para as três abordagens e novos compostos foram propostos utilizando a combinação das subestruturas moleculares dos compostos mais ativos da série. A representação dos átomos como círculos de tamanhos proporcionais ao raio de van der Waals e em diferentes cores melhorou a capacidade preditiva das abordagens aug-MIA-QSAR e aug-MIA-QSARcolor em comparação ao MIA-QSARt, tanto em se tratando de validação externa (r^2_{ext}) quanto aos parâmetros adicionais r^2_{rand} , r^2_m , r^2_p , os quais garantem a confiabilidade e a capacidade preditiva de modelos QSAR. As moléculas propostas apresentaram-se mais ativas que os compostos de maior valor de atividade experimental da série. A interpretação química sobre os requisitos estruturais mostra que um anel fenila trissubstituído contendo grupos e/ou elementos eletronegativos em R1 é promissor enquanto que um grupo triflúor metila na posição 3 é desejável, em vez do cloro.

Palavras-chave: Análise multivariada de imagens. Doenças negligenciadas. *Trypanosoma cruzi*. PLS.

1 INTRODUÇÃO

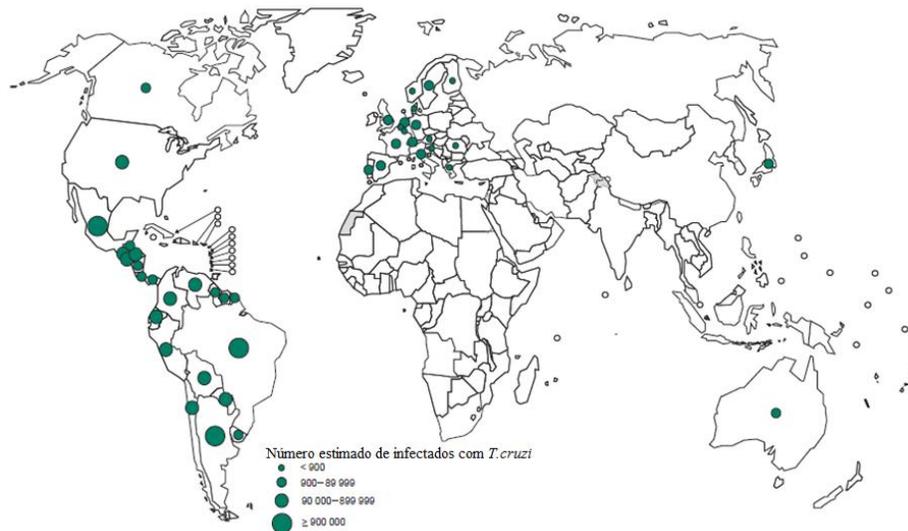
1.1 Aspectos gerais da doença de Chagas

A doença de Chagas é assim denominada em homenagem ao seu descobridor, Carlos Ribeiro Justino das Chagas, médico e pesquisador brasileiro, que em 1909 descreveu uma espécie de protozoário, hemoflagelado, pertencente à Família *Trypanosomatidae*. Na ocasião Chagas verificou que tal protozoário habitava o intestino de insetos e animais domésticos e, ao contaminar o homem, gerava infecções que podiam levar à morte (CHAGAS, 1909). A doença também é conhecida como tripanossomíase americana, em função de o seu agente etiológico *Trypanosoma cruzi* ser encontrado quase que exclusivamente na América Latina (OMS, 2015a).

Segundo a Organização Mundial de Saúde (OMS), a doença de Chagas continua sendo um dos maiores problemas de saúde pública na América Latina, onde causa mais de 7000 mortes por ano. Endêmica em 21 países da América Latina, estima-se que mais de 25 milhões de pessoas estejam em risco de infecção e que mais de 7 milhões de pessoas sejam portadoras da doença em todo o mundo (OMS, 2015b). Desses, cerca de 2 milhões já se encontram na fase crônica, forma mais grave da doença. Estima-se ainda que apenas 0,5% dos infectados receba tratamento e que 23% dos portadores da doença vivem no Brasil (SOUZA, 2010). Apesar de a doença ter distribuição predominante no continente americano, são registrados casos em países da Europa, no Japão e na Austrália por outros mecanismos de transmissão que não seja a forma vetorial (Figura 5) (OMS, 2015b).

O *T. cruzi* se desenvolve em hospedeiro invertebrado (inseto hemíptero e hematófago, dos gêneros *Rhodnius*, *Panstrongylus* e *Triatoma*, conhecidos no Brasil como barbeiro, sendo as espécies *T. infestans*, *T. megistus*, *T. brasiliensis*,

Figura 5 - Distribuição dos casos de doença de Chagas reportados à OMS, em nível mundial, 2010-2013.



Fonte: Adaptado de OMS (2015b)

T. pseudomaculata e *T. sordida* de grande importância epidemiológica), e em hospedeiros vertebrados (BRASIL, 2010; OMS, 2015a)

Aproximadamente, 100 espécies de animais selvagens e domésticos são hospedeiros do *T. cruzi*, sendo potenciais vetores de transmissão da infecção para seres humanos. Dentre estas espécies, estão: macacos, morcegos, roedores, marsupiais, gatos, cachorros, porcos, cabras etc. (COURA; DIAS, 2009).

A transmissão ao homem acontece principalmente por via vetorial, geralmente através do contato com as fezes e urina do inseto vetor (barbeiro) contaminado. Porém, casos de transmissão oral através de alimentos contaminados têm sido relatados pela vigilância epidemiológica, principalmente nos estados da região norte do Brasil (BRASIL, 2015; FERREIRA et al., 2014). A transfusão sanguínea, acidentes de laboratório, transplantes de órgãos, via placentária e manipulação de animais infectados são outras fontes ocasionais de

contaminação (BRASIL, 2015; FERREIRA; BRANQUINHO; CARDARELLI-LEITE, 2014; OMS, 2015b).

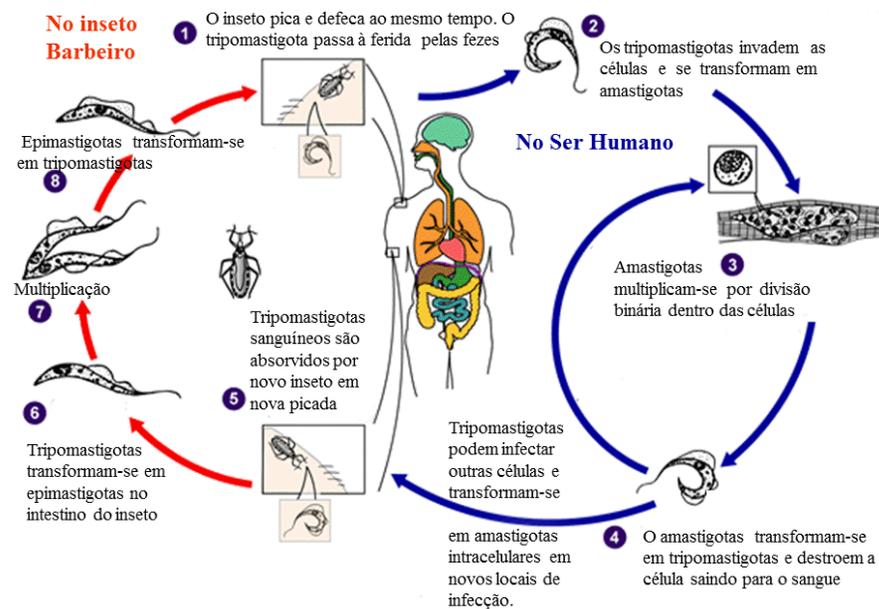
Na transmissão vetorial, o barbeiro contaminado pica o vertebrado, o qual, após se alimentar, defeca no local da picada. O *T. cruzi* presente nas fezes do barbeiro na forma de tripomastigota é assim transmitido ao novo hospedeiro. As membranas mucosas intactas, tais como a conjuntiva, também são portas de entrada para o *T. cruzi*. (BRASIL, 2010). Para que *T. cruzi* seja transmitido a um vertebrado, inicialmente o inseto vetor precisa picar um vertebrado que contenha o parasito na corrente sanguínea. Após repasto sanguíneo, o parasito na forma tripomastigota chega ao intestino do inseto vetor, onde se transforma em epimastigota (multiplicação extracelular), em seguida diferencia-se em tripomastigota metacíclico, que é a forma infectante acumulada nas fezes do inseto (CENTERS FOR DISEASE CONTROL AND PREVENTION-CDC, 2015).

Quando o inseto vetor pica o vertebrado, ele defeca no local e o parasito na forma tripomastigota adentra em células próximas ao local da picada e transforma-se em amastigotas, as quais se multiplicam por divisão binária e se diferenciam em tripomastigotas, sendo depois liberadas na circulação sob a forma tripomastigotas sanguíneos. Na corrente sanguínea, os tripomastigotas não se replicam, porém, ao infectar novas células em vários tecidos, transforma-se em amastigotas dando sequência ao ciclo infeccioso. (CDC, 2015). A Figura 6 ilustra o ciclo de transmissão do *T. cruzi* em hospedeiro vertebrado (homem) e invertebrado (barbeiro).

A doença de Chagas, sob o ponto de vista patológico e clínico, tem uma fase aguda e uma fase crônica. A fase aguda é caracterizada pela presença imediata do parasito no sangue e pode durar de algumas semanas a meses. Essa fase é, para muitos, assintomática ou apresenta sintomas leves e inespecíficos,

passando despercebidos na maioria dos casos. Porém, para indivíduos imunocomprometidos e para crianças pode ser letal.

Figura 6 - Ciclo de transmissão do *T. cruzi* em hospedeiros vertebrados e invertebrados.



Fonte: Adaptado de CDC (2015)

As manifestações gerais são de febre, dores no corpo, fadiga, cefaleia, diarreia, erupções cutâneas, poliadenia, anorexia, vômitos, conjuntive unilateral (sinal de Romaña), complicações cardíacas e hepatoesplenomegalia (BRASIL, 2010).

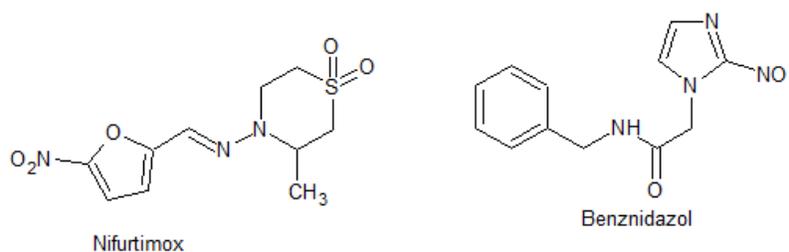
Alguns fatores, tais como dimensão do inóculo do *T. cruzi*, a resposta imunológica do hospedeiro, características das cepas do infectante, a quantidade de tripanosomas inoculados e número de reinfecções, são determinantes para o aparecimento das manifestações clínicas e também para o agravamento da doença (COURA, 2003).

A fase crônica é determinada pelas formas clínicas indeterminada e sintomática. Na forma indeterminada, a doença permanece assintomática indefinidamente, geralmente de 10 a 30 anos, porém com exames sorológicos positivos. Após anos nessa fase, alguns portadores manifestam sintomas da fase crônica, que resultam do comprometimento do sistema cardiovascular, digestivo e, em muitos casos, ambos os sistemas (BRASIL, 2010).

1.2 Tratamento

Apenas dois medicamentos estão disponíveis para tratar a infecção pelo *T. cruzi*: o benznidazol (tratamento de primeira linha na maioria dos países) e o nifurtimox (Figura 7) (BRASIL, 2010). Além da baixa disponibilização dos medicamentos nas áreas mais afetadas, a alta toxicidade e a baixa eficiência desses fármacos são fatores que dificultam o tratamento das pessoas infectadas, colaborando, assim, para a manutenção do ciclo da doença, bem como para a morbidade e mortalidade dos infectados (BRASIL, 2010; COURA; DIAS, 2009).

Figura 7 - Estrutura química do nifurtimox e benznidazol.



Fonte: DIAS et al. (2009).

Introduzidos na década de 1970, ambos são eficazes para tratar a fase aguda, porém apresentam eficiência restrita à fase crônica (CALDAS et al, 2008). O nifurtimox é um composto nitrofurano e apresenta atividade contra as formas tripomastigotas e amastigotas do parasito. A maior eficácia da droga é verificada em pacientes na fase aguda e em menor idade (STREIGER et al., 2004). Além disso, o nifurtimox apresenta vários efeitos adversos, tais como anorexia, alterações psíquicas e gastrointestinais, assim como diarreia e vômitos. Em razão desses problemas, sua comercialização foi interrompida no Brasil pela Agência Nacional de Vigilância Sanitária (ANVISA) na década de 1980 e, posteriormente, Venezuela, Chile e Argentina também interromperam sua comercialização (PEDROSA et al, 2001).

O benznidazol (N-benzil-2-nitroimidazol-1-acetamida) tem por objetivos a eliminação dos parasitos da circulação sanguínea, a prevenção e a redução dos sintomas (CALDAS et al., 2008). A efetividade do benznidazol foi estimada, no Brasil, entre 60 a 90% para os casos agudos e entre 8 a 25% para os casos crônicos, inclusive na forma indeterminada (FERREIRA, 1990). No estudo realizado por Ferreira (1990) o benznidazol mostrou-se mais efetivo que o nifurtimox na fase aguda. Já na fase crônica assintomática, a eficiência de ambos foi baixa, 10% para benznidazol e 6% para nifurtimox. Além da baixa eficiência com o aumento do tempo de contágio, os efeitos adversos da administração desse medicamento, aliado ao longo período de tratamento, conduzem muitos doentes a interrupção do tratamento (OMS, 2015a).

1.3 Desenvolvimento de novos medicamentos

As pesquisas para desenvolvimento de novas drogas contra o *T. cruzi* só ganharam relevância recentemente, a partir dos programas de desenvolvimento de drogas para doenças tropicais negligenciadas. Em relação à extensa lista de

compostos que vêm sendo estudados como potenciais antichagásicos, desde a introdução do nifurtimox e benznidazol, somente alopurinol, itracinazol, fluconazol e posaconazol foram submetidos a ensaios clínicos. No entanto, nenhum deles foi aprovado em razão da baixa eficácia. Por outro lado, avanços expressivos sobre os aspectos biológicos, genéticos e evolucionários do *T. cruzi* vêm sendo alcançados, os quais têm permitido a identificação de diversos alvos macro e micromoleculares, tais como proteinases, biossíntese de esteróis, metabolismo dependente de grupos tióis, via glicolítica, membranas celulares e organelas celulares, entre outros, para o desenvolvimento de novos antichagásicos (SOUZA, 2015).

Um estudo realizado por Doyle e colaboradores (2007) mostra que inibidores da enzima cruzaina são alvos promissores como agentes antitripanossômicos, pois a cruzaina, uma cisteína protease do *T. cruzi*, está envolvida com a fase de acolhimento das infecções das células e participa do processo de replicação e diferenciação do parasito (DU et al., 2002; GREENBAUM et al., 2004). Dentre os possíveis inibidores para esta enzima, os derivados tiossemicarbazonas têm mostrado relevante potencial. As tiossemicarbazonas são bases Schiff, obtidas a partir de uma reação de condensação quimiosseletiva de baixo custo de um aldeído ou cetona com uma tiossemicarbazida (TENÓRIO, 2005). O seu potencial biológico é atribuído principalmente a sua capacidade quelante e de agir como inibidor de enzimas, através da complexação de metais endógenos ou através de reações redox (BERALDO, 2002). Contudo, o mecanismo de ação do relevante potencial inibitório dos derivados tiossemicarbazonas contra a enzima cruzaina ainda não foi elucidado. Nesse contexto, estudos têm buscado a compreensão deste mecanismo e a identificação de substituintes que potencializem a interação entre ligante (tiossemicarbazonas) e receptor (cruzaina).

No contexto das novas tecnologias utilizadas para pesquisa e desenvolvimento de fármacos, o presente estudo tem por objetivo utilizar métodos MIA-QSAR para propor novas estruturas de compostos derivados de tiossemicarbazonas e semicarbazonas com atividade melhorada frente aos existentes. Além disso, os avanços da abordagem MIA-QSAR em relação à incorporação de novas informações químicas aos descritores MIA serão avaliados.

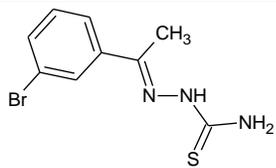
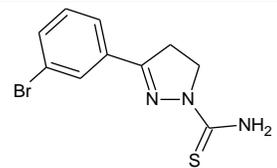
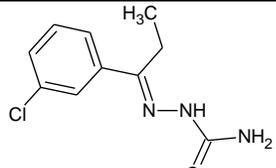
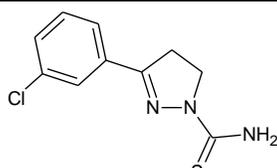
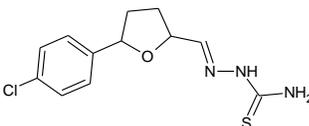
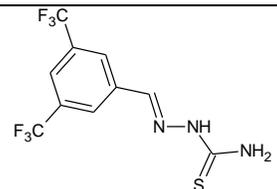
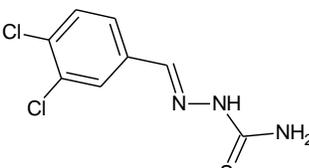
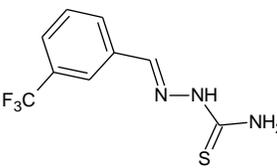
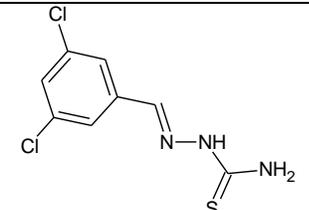
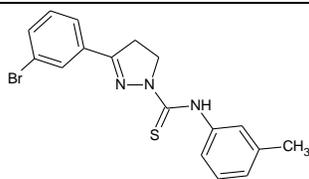
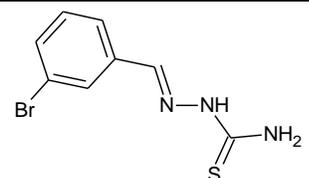
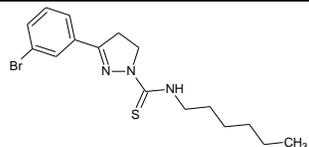
2 METODOLOGIA

Inicialmente, um conjunto de dados constituído de 53 compostos com valores experimentais de pIC_{50} [$-\log IC_{50}$ (IC_{50} é a metade da concentração inibitória máxima, em mol L^{-1} , referente à eficácia contra a proteína recombinante purificada catepsina L-*T. cruzi* (cruzaína)] (Quadro 1) (DU et al, 2002; LOZANO et al., 2012), foi desenhado segundo as abordagens MIA-QSARt, aug-MIA-QSAR e aug-MIA-QSARcolor. Após a obtenção das matrizes de dados, o modelo QSAR foi gerado e validado e, por fim, novas estruturas foram propostas com bases nos melhores valores de pIC_{50} .

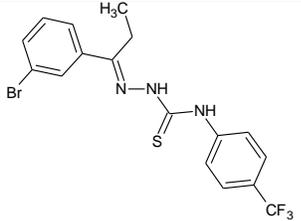
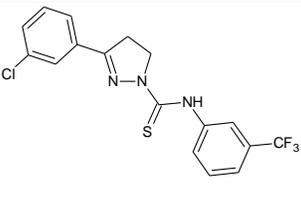
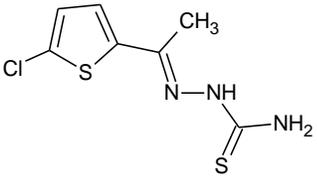
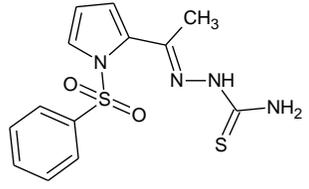
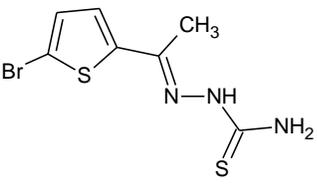
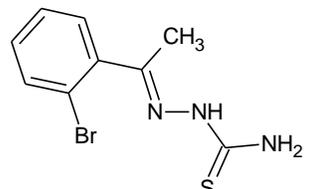
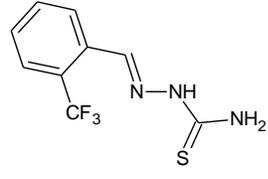
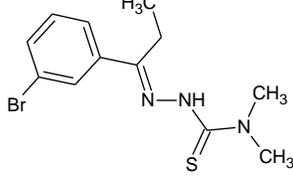
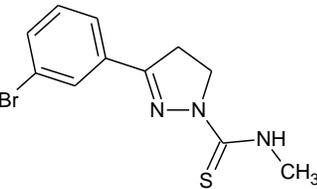
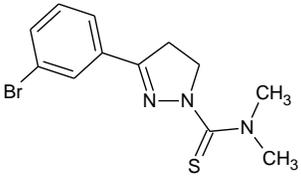
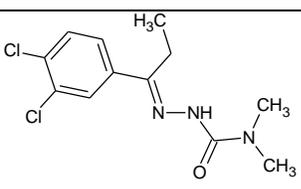
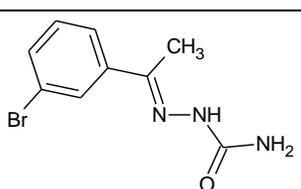
2.1 Construção das imagens e obtenção das matrizes para as abordagens MIA-QSAR

Para obtenção das imagens referentes à abordagem MIA-QSAR, os 53 compostos foram sistematicamente desenhados utilizando o programa ChemSketch (ADVANCED CHEMISTRY DEVELOPMENT - ACD, 2009). Cada estrutura química foi transferida para uma área de trabalho com dimensões previamente estabelecidas (400x400px) do aplicativo *Paint Microsoft Windows*, para que todas as imagens tivessem a mesma dimensão. Em seguida, um pixel

Quadro 1 (Continua)

Cp	Estrutura	pIC ₅₀	Cp	Estrutura	pIC ₅₀
11		6,70	12		6,70
13		6,66	14		6,64
15		6,52	16		6,49
17		6,34	18		6,32
19		6,25	20		6,00
21		5,89	22		5,85

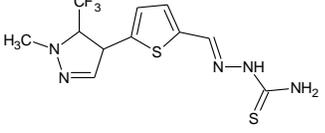
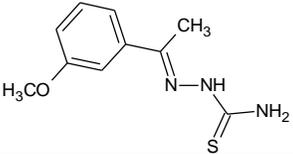
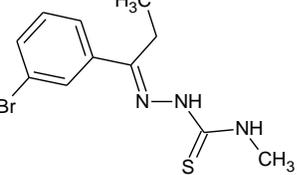
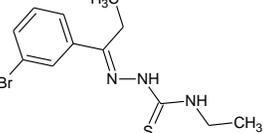
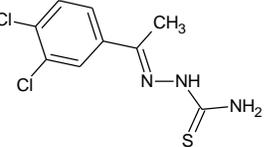
Quadro 1(Continua)

Cp	Estrutura	pIC ₅₀	Cp	Estrutura	pIC ₅₀
23		5,80	24		5,77
25		5,72	26		5,55
27		5,42	28		5,41
29		5,00	30		5,00
31		5,00	32		5,00
33		5,00	34		5,00

Quadro 1(Continua)

Cp	Estrutura	pIC ₅₀	Cp	Estrutura	pIC ₅₀
35		5,00	36		5,00
37		5,00	38		5,00
39		5,00	40		5,00
41		7,22	42		7,10
43		7,00	44		6,57
4		6,55	46		6,55
47		6,25	48		5,55

Quadro 1(Conclusão)

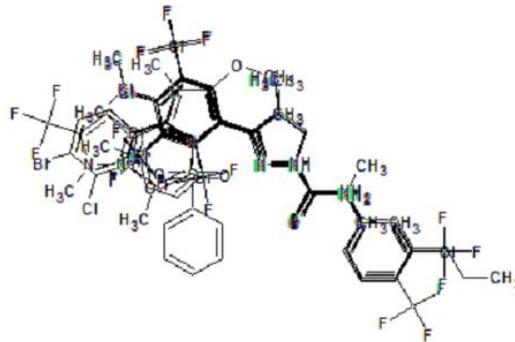
Cp	Estrutura	pIC ₅₀	Cp	Estrutura	pIC ₅₀
49		5,15	50		5,00
51		5,00	52		5,00
53					5,00

Fonte: Adaptado de Lozano et al. (2012).

química foi movida, de tal forma que o pixel escolhido para fins de alinhamento 2D fosse ajustado na coordenada determinada. Em seguida, cada estrutura foi salva como *bitmaps*. Essa etapa corresponde ao alinhamento bidimensional (2D) e é feita manualmente (com auxílio do *mouse*); portanto, requer precisão manual, mas, contudo, é mais simples que o tratamento realizado para alinhamento em tridimensional. O alinhamento das imagens resulta na sobreposição das estruturas químicas, formando um arranjo tridimensional em que as partes comuns entre as estruturas da série congênere (esqueleto básico) estejam congruentes (Figura 8).

Para a obtenção das imagens para as abordagem aug-MIA-QSAR, os 53 compostos foram desenhados no programa GaussView 5.0 (DENNINGTON; KEITH; MILLAN, 2008). Nesse programa, os átomos foram representados por círculos com tamanho proporcional ao raio de van der Waals e cores aleatórias foram utilizadas para diferenciá-los. Com objetivo de tornar os descritores aug-MIA-QSAR ainda mais

Figura 8 - Sobreposição das estruturas químicas segundo a abordagem MIA-QSARt.



Fonte: Do autor (2016).

interpretáveis quimicamente, uma nova extensão dessa abordagem foi implementada, denominada aug-MIA-QSARcolor, na qual as cores são cuidadosamente definidas para se correlacionarem numericamente com propriedades periódicas de interesse químico, físico, físico-químico e biológico, tais como eletronegatividade, hidrofobicidade atômica, polarizabilidade, refração atômica, raio covalente. Neste trabalho, as cores foram selecionadas para serem proporcionais à escala de eletronegatividade de Pauling. Considerando que as cores fornecidas pelo software a cada átomo são formadas pela resultante da combinação das cores vermelho, verde e azul (sistema RGB de cores), sendo que a contribuição de cada um desses canais varia de 0 a 255, a cor de cada átomo foi definida aleatoriamente, porém atribuída de tal forma que o valor total do pixel fosse proporcional à eletronegatividade de cada átomo. Assim, baixos valores de eletronegatividade foram relacionados com cores que contribuem com um valor pequeno de pixel. Por exemplo: o átomo de hidrogênio tem eletronegatividade igual a 2,1 sua cor foi definida para um valor de pixel igual 400, já para o átomo de flúor, eletronegatividade igual a 4,0, a cor foi definida para apresentar um alto valor de pixel (750). A Tabela 2 apresenta

os valores de pixel para diferentes átomos segundo as abordagens aug-MIA-QSAR e aug-MIA-QSARcolor

Tabela 2 - Valores de pixels para representar as cores dos diferentes átomos segundo as abordagens aug-MIA-QSAR e aug-MIA-QSARcolor

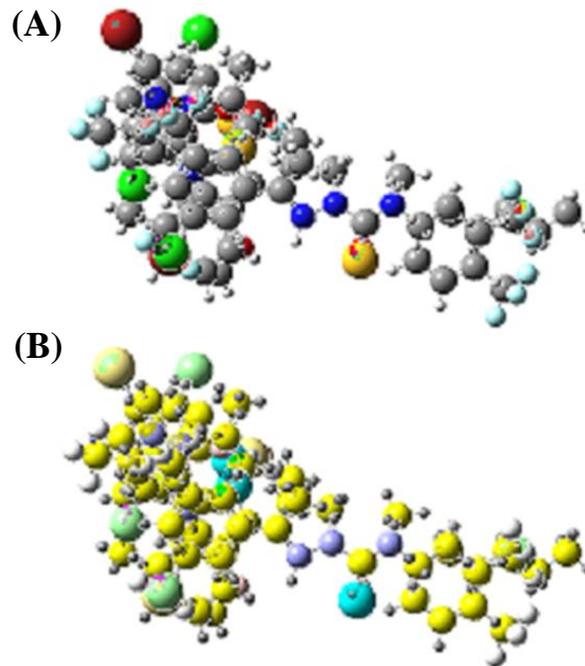
Átomo	aug-MIA-QSAR		^c aug-MIA-QSARcolor			
	E ^a	Pixel ^b	Pixel	Vermelho	Verde	Azul
H	2,1	612	400	134	133	133
C	2,5	426	480	240	240	0
O	3,5	279	670	255	207	207
N	3,0	229	575	160	160	255
F	4,0	688	750	250	250	250
Cl	3,0	493	575	160	255	160
Br	2,8	289	535	255	140	140
I	2,5	231	480	240	0	240
S	2,5	294	480	0	240	240

^aEletronegatividade de Pauling; ^bValor de pixel correspondente à cor padrão dos átomos no GaussView, resultante das somas das componentes vermelho, verde e azul; ^cValor de pixel proporcional à eletronegatividade de Pauling.

Fonte: Do autor (2016).

O passo seguinte ao desenho das estruturas é o alinhamento. Diferentemente do processo manual utilizado na abordagem MIA-QSARt, o alinhamento das imagens desenhadas para as abordagens aug-MIA-QSAR e aug-MIA-QSARcolor é realizado pelo simples fato de, ao desenhar as estruturas em uma janela de tamanho definida no programa GaussView e com o cuidado de manter a mesma coordenada para todas as estruturas e copiar as imagens para o aplicativo *Paint Microsoft Windows*, as imagens são automaticamente alinhadas em uma janela de tamanho definido pelo aplicativo *Microsoft Paint*, utilizando como referência o tamanho da janela proveniente do programa GaussView. Por meio desse procedimento, obtém-se a sobreposição das estruturas químicas (formando um arranjo tridimensional) em que as partes comuns das estruturas da série congênere (esqueleto básico) estejam congruentes (Figura 9).

Figura 9 - Sobreposição da série congênere de compostos que ilustram a subestrutura comum e as porções variáveis dos derivados tiossemicarbazonas e semicarbazonas com atividade antitripanossoma para as abordagens aug-MIA-QSAR (A) e aug-MIA-QSARcolor (B).

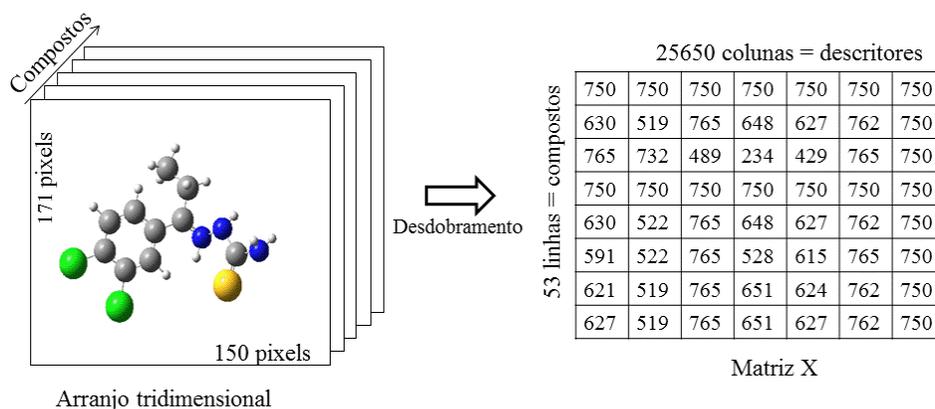


Fonte: Do autor (2016).

Uma vez que as imagens são constituídas por pixels, esses podem ser convertidos em números, de acordo com o sistema de composição de cores RGB. A transformação das imagens em matrizes numéricas foi realizada usando o programa Chemoface (NUNES et al., 2012). Em seguida, cada imagem formada por $n \times m$ pixels é desdobrada para um vetor linha de dimensão $1 \times (n \times m)$. Assim, cada linha da matriz \mathbf{X} representa uma estrutura química. Após o procedimento, as matrizes para abordagens MIA-QSARt, aug-MIA-QSAR e aug-MIA-QSARcolor passaram a ter as seguintes dimensões 53×160000 ,

53×25650 e 53×25650, respectivamente. Subsequentemente, colunas com variância nula (espaços brancos comuns para todas as imagens e partes congruentes das estruturas químicas) foram eliminadas das matrizes de dados para reduzir a dimensionalidade dos dados e, conseqüentemente, os custos computacionais. As novas matrizes foram reduzidas para as seguintes dimensões: 53×7671 para MIA-QSARt, 53×7088 para aug-MIA-QSAR e aug-MIA-QSARcolor. A Figura 10 ilustra a construção do arranjo tridimensional e a conversão das imagens em matrizes numéricas.

Figura 10: Construção da matriz **X** a partir do desdobramento do arranjo tridimensional obtido da sobreposição das imagens do conjunto de dados (a abordagem aug-MIA-QSAR foi utilizada como exemplo).



Fonte: Adaptado de Freitas (2006).

2.2 Calibração e validação dos modelos

Após obtenção das matrizes de dados, o conjunto de compostos foi dividido aleatoriamente em grupo treinamento (43 compostos) e grupo teste (10 compostos). O conjunto teste formado pelos compostos 5, 12, 19, 23, 27, 31, 36, 42, 46 e 50 (cerca de 20% de toda a série) foi escolhido de tal forma que os compostos com baixa, média e alta atividade fossem distribuídos de forma

homogênea. Em seguida, modelos QSAR foram gerados a partir de regressão por PLS da matriz \mathbf{X} (descritores) com o bloco \mathbf{y} (bioatividades). A qualidade dos modelos foi avaliada pelos parâmetros estatísticos coeficientes de determinação para calibração (r^2_{cal}), validação cruzada *leave-one-out* (r^2_{cv}), raiz quadrada do erro da calibração e validação cruzada (RMSE_c e RMSE_{cv}, respectivamente). Estes parâmetros, sozinhos, não garantem a confiabilidade do modelo. Portanto, adicionalmente, foi realizado o teste da validação externa (r^2_{ext}) e *y-randomization* (r^2_{rand}), bem como a raiz quadrada do erro da validação externa (RMSEP). A validação externa é considerada o parâmetro mais importante para garantir se um modelo QSAR é confiável (GOLBRAIKH; TROPSHA, 2002), enquanto que o teste *y-randomization* avalia se o bom resultado de calibração não ocorreu por obra do acaso ou por *overfitting*. Tal teste é efetuado sobre o conjunto de treinamento e uma calibração é realizada utilizando o bloco \mathbf{y} aleatorizado e a matriz \mathbf{X} intacta. Uma correlação pior em comparação com o modelo de calibração real é esperado se os descritores realmente codificam as propriedades biológicas. Outros critérios adicionais também foram utilizados para garantir a confiabilidade e a sensibilidade dos modelos QSAR. Para avaliar a diferença estatística entre r^2 e r^2_{rand} foi utilizado o coeficiente de determinação penalizado corrigido ($^C r^2_p$) (MITRA, SAHA e ROY, 2010). Para detectar a proximidade entre a atividade observada e atividade esperada, o coeficiente de determinação modificado foi calculado para o grupo teste (r^2_m) (MITRA, SAHA e ROY, 2010; OJHA et al., 2011).

3 RESULTADOS E DISCUSSÕES

3.1 Validação dos modelos MIA-QSARt, aug-MIA-QSAR e aug-MIA-QSARcolor

Uma vez que um grande número de descritores (em geral milhares) é gerado nas matrizes de dados em MIA-QSAR, a regressão por mínimos quadrados parciais é, muitas vezes, utilizada no passo de calibração, exigindo, portanto, a escolha de um número ótimo de variáveis latentes (PLS componentes) para prosseguir com a construção do modelo QSAR. Isso pode ser alcançado através da análise dos erros quadrados médios da validação cruzada *leave-one-out* (RMSECV) em função do número de variáveis latentes. Um número apropriado de variáveis latentes para utilizar em um modelo é aquele em que o RMSECV é minimizado ou que não varia significativamente. O número ótimo de variáveis latentes para as três abordagens pode ser obtido do Gráfico 1.

A utilização de variáveis latentes mais do que o necessário pode levar a superajuste, em que as informações desnecessárias/indesejáveis são calibradas, fornecendo bons coeficientes de correlação para a calibração dos compostos do conjunto de treino, mas não é um modelo preditivo de confiança para as amostras externas. Apesar de os modelos aug-MIA-QSAR e aug-MIA-QSARcolor apresentarem número ótimo de variável latente menor que o MIA-QSARt (7LV, 5LV e 9LV, respectivamente), os modelos QSAR para as três abordagens foram construídos com 9 LV, com a finalidade de reproduzir as condições aplicadas na abordagem MIA-QSARt, para título de comparação das mesmas. A Tabela 3 apresenta os resultados estatísticos para as três abordagens MIA-QSAR.

Gráfico 1 - Número de variáveis latentes *versus* RMSECV para as abordagens MIA-QSARt (A), aug-MIA-QSAR (B) e aug-MIA-QSARcolor (C).

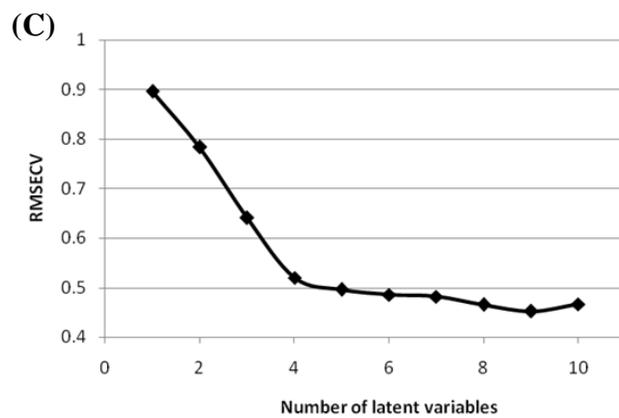
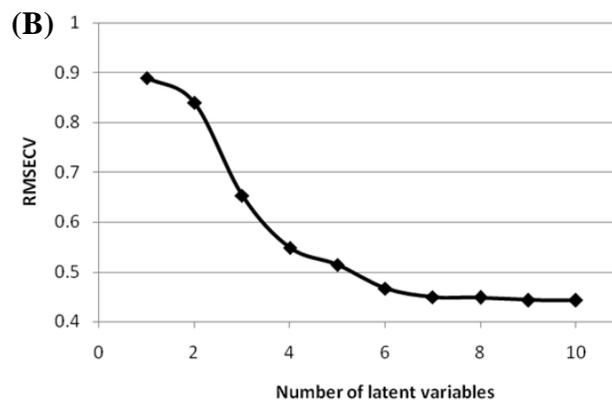
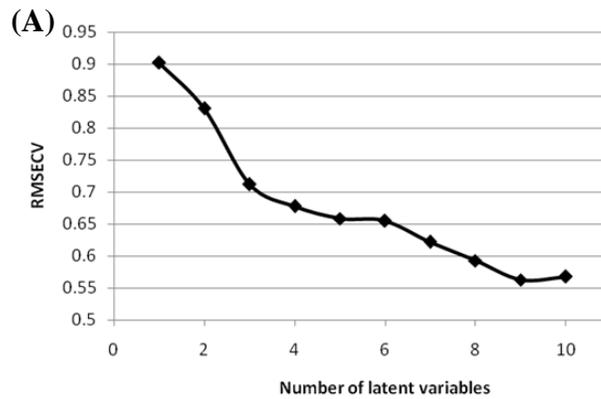


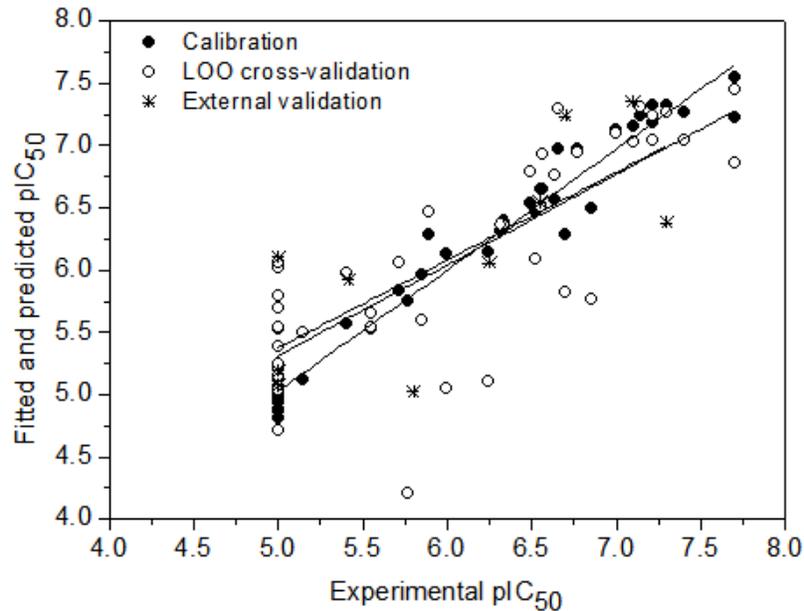
Tabela 3 - Resultados estatísticos para os modelos MIA-QSAR para a série de tiossemicarbazonas e semicarbazonas com atividade antitripanossômica.

Parâmetros	MIA-QSARt	aug-MIA-QSAR	aug-MIA-QSARcolor
r^2_{cal}	0.967	0.966	0.966
RMSEC	0.162	0.166	0.164
r^2_{cv}	0.624	0.763	0.756
RMSECV	0.563	0.443	0.452
r^2_{test}	0.569	0.654	0.714
RMSEP	0.579	0.512	0.480
r^2_m	0.382	0.481	0.594
$r^2_{rand}^a$	0.736	0.702	0.696
${}^c r^2_p$	0.473	0.504	0.511

O modelo MIA-QSARt encontrado parece ter um ajuste forçado, em razão da boa correlação obtida no passo de calibração ($r^2_{cal} = 0,967$, RMSEC = 0,162) não ser significativamente diferente dos resultados obtidos da regressão PLS entre a matriz **X** intacta (matriz de descritores) e os dados do bloco **y** aleatorizados, cujo r^2_{rand} foi 0,736 (média de 10 repetições). A semelhança estatística entre r^2_{cal} e r^2_{rand} pode ser avaliada pelo parâmetro ${}^c r^2_p$ acima mencionado, o qual foi encontrado ser inferior a 0,5 (0,473). Além disso, apesar de fornecer correlações aceitáveis na validação cruzada (*leave-one-out*) e validação externa (Tabela 3), o baixo valor de r^2_m obtido para o conjunto teste (0,382) indica que os valores previstos absolutos de pIC_{50} não são significativamente congruentes com os dados experimentais correspondentes. A razão para esse desvio deve-se a problemas com a inclinação e/ou intercepto da linha ótima esperada para gráfico de pIC_{50} experimental *versus* predito (Gráfico 2).

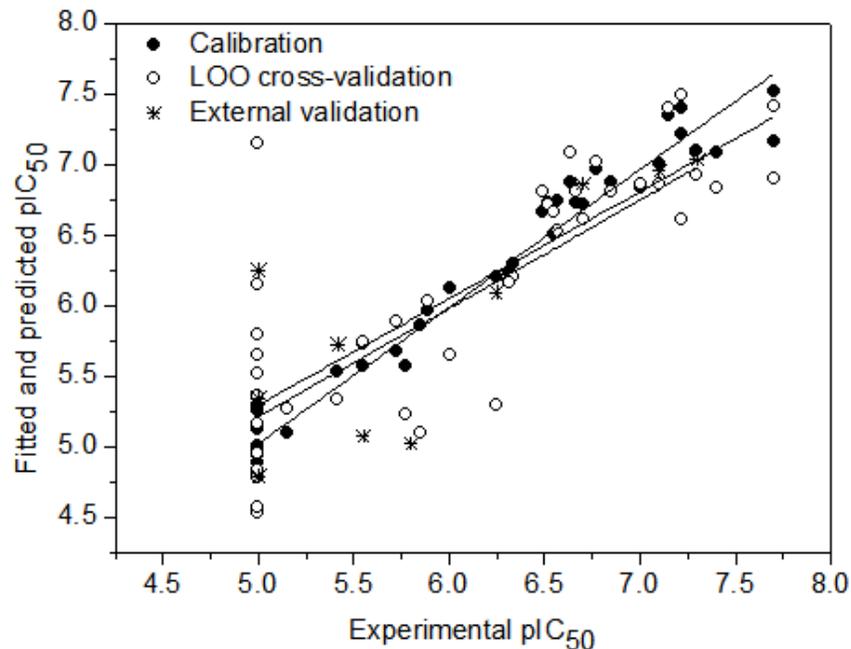
A abordagem aug-MIA-QSAR apresentou uma calibração com alta correlação entre valores de pIC_{50} experimentais e preditos. O modelo não apresentou ajuste forçado, já que o valor ${}^c r^2_p$ foi superior a 0,5 (Tabela 3). Além disso, a validação cruzada *leave-one-out* foi melhorada

Gráfico 2 - Valores de pIC_{50} experimentais *versus* preditos para a série de tiossemicarbazonas e semicarbazonas, segundo a abordagem MIA-QSARt.



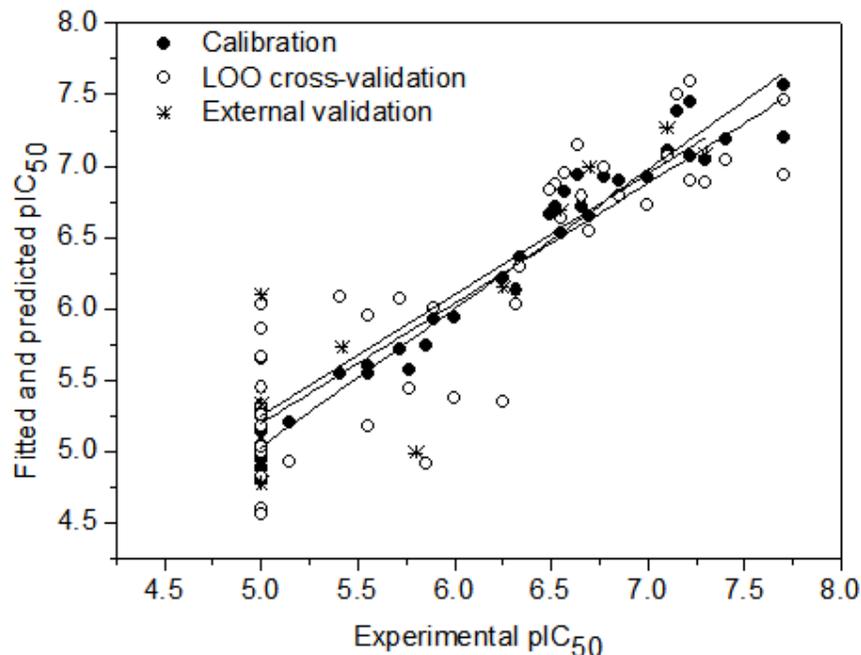
em relação ao MIA-QSAR, com $r^2_{cv} = 0,763$, o que é bastante satisfatório. No entanto, a validação externa tem sido apontada como a única maneira de estabelecer um modelo QSAR confiável (GOLBRAIKH e TROPSHA, 2002) e, apesar da boa correlação entre os valores experimentais e preditos para este conjunto de compostos ($r^2_{test} = 0,654$), o valor de r^2_m foi inferior a 0,5 ($r^2_m = 0,481$). Nesse sentido, apesar de ter apresentado progressos em relação à abordagem MIA-QSARt, o modelo aug-MIA-QSAR também requer melhorias para certificar que as previsões para os novos candidatos a fármacos sejam realmente confiáveis. As curvas de correlação para o modelo aug-MIA-QSAR são apresentadas no Gráfico 3.

Gráfico 3 - Valores de pIC_{50} experimentais *versus* preditos para a série de tiossemicarbazonas e semicarbazonas, segundo a abordagem aug-MIA-QSAR.



O resultado de calibração para a abordagem aug-MIA-QSARcolor (Tabela 3) manteve-se estatisticamente significativo, com baixo RMSEC, alto valor de r^2 e valor de $^c r^2_p$ superior a 0,5. Contudo, a principal vantagem desse modelo está relacionada à sua validação, pois os resultados de validação cruzada (*leave-one-out*) sugerem uma previsibilidade elevada ($r^2_{cv} = 0,756$), o que é confirmado pela validação externa, em que r^2_{test} foi melhorado para 0,714, demonstrando que existe uma boa correlação entre os valores de pIC_{50} experimentais e previstos e que ambos apresentam valores similares. As correlações para o modelo aug-MIA-QSARcolor são ilustradas no Gráfico 4.

Gráfico 4 - Valores de pIC_{50} experimentais *versus* preditos para a série de tiossemicarbazonas e semicarbazonas, segundo a abordagem aug-MIA-QSARcolor.



O modelo aug-MIA-QSARcolor satisfaz os parâmetros de validação mais importantes para atestar a confiabilidade de um modelo QSAR, o que sugere que as propriedades químicas que correlacionam com atividades biológicas são devidamente codificadas pelos descritores aug-MIA-QSARcolor e que as interações dipolares desempenham um papel fundamental para o mecanismo de ação dessa série de compostos. O bom desempenho dos descritores aug-MIA-QSAR mostra que a abordagem pode ser utilizada para prever a atividade de novas drogas. Uma análise QSAR baseada nos descritores físico-químicos para a mesma série de compostos tem sido realizada, mas usando um diferente conjunto de compostos de teste. O melhor modelo PLS utilizando 8 descritores selecionados resultou em $r^2 = 0,85$, $q^2 = 0,78$ e teste $r^2 = 0,55$ (LOZANO et al., 2012). Assim, os modelos baseados em MIA são

consistentes com o QSAR clássico, utilizando descritores físico-químicos. No entanto, uma comparação mais profunda não pode ser realizada por falta de validações adicionais no trabalho citado.

3.2 Estruturas químicas propostas

Uma estratégia útil para a obtenção de novos compostos que contenham um determinado esqueleto básico é combinar as subestruturas moleculares de compostos altamente ativos de uma série congênere. Isso pode ser conseguido usando o conjunto de dados do Quadro 1. Este procedimento pode resultar em uma substância altamente ativa com propriedades melhoradas, tais como uma menor toxicidade e uma elevada atividade contra as cepas resistentes à quimioterapia disponível. Por exemplo, os compostos **1** e **2** do Quadro 1 são os mais ativos da série, mas eles diferem um do outro pelos substituintes R1 do anel benzênico [(3,4-Cl) em **1** e (3,5-CF₃) em **2**]. Assim, novos compostos semelhantes podem ser concebidos com alterações em R1. As seguintes alterações foram propostas no anel benzênico: **A**: 3-CF₃, 4-Cl; **B**: 4-Cl, 5-CF₃; **C**: 3,5-CF₃, 4-Cl; **D**: 3,4-Cl, 5-CF₃; **E**: 3-Cl, 5-CF₃. Todas essas possibilidades têm os grupos Cl e CF₃ calibrados nas respectivas posições do anel. Assim, os compostos **A-E** da Figura 11 foram propostos e suas bioatividades foram estimadas utilizando os parâmetros de regressão obtidos para os modelos MIA-QSARt, aug-MIA-QSAR e aug-MIA-QSARcolor.

De acordo com o modelo MIA-QSARt, os compostos **A** e **C** seriam os mais ativos dentre as propostas (Tabela 4), mas não mais ativos do que os compostos disponíveis experimentalmente **1** (7,70) e **2** (7,70). Nos modelos aug-MIA-QSAR e aug-MIA-QSARcolor, os compostos **C** e **D** apresentam maiores valores de atividade e superior ao valor dos compostos mais ativos da série (1 e 2). Como estes dois modelos mostram mais preditivos que o método MIA-

QSARt, as predições para os compostos **C** e **D** podem ser consideradas confiáveis e realísticas.

Figura 11 - Tiossemicarbazonas propostas a partir da combinação das subestruturas dos compostos **1** e **2**.

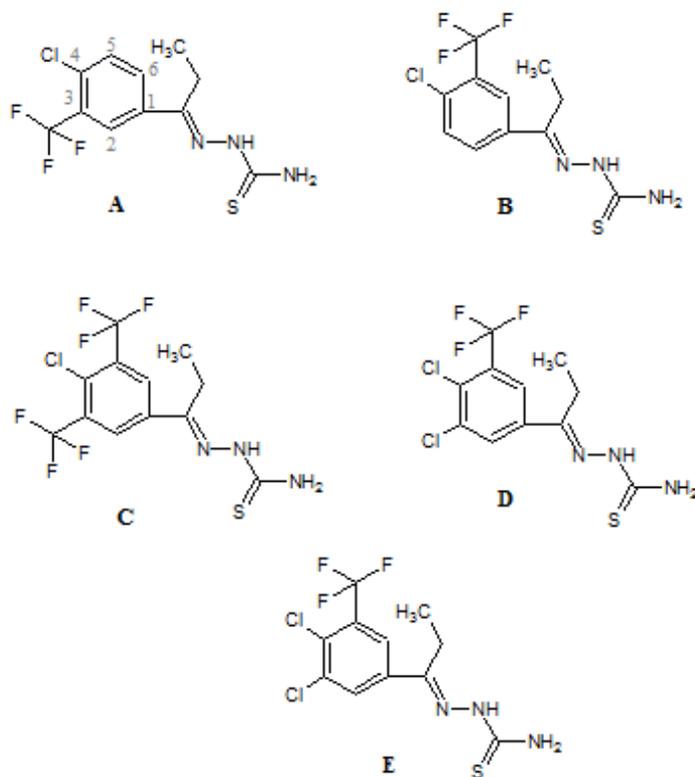


Tabela 4 - Valores de pIC_{50} preditos para os derivados de tiossemicarbazonas e semicarbazonas (**A-E**) propostos.

Composto	MIA-QSARt	aug-MIA-QSAR	aug-MIA-QSAR _{color}
A	7.37	7.55	7.48
B	6.73	7.38	7.44
C	7.60	7.98	8.01
D	7.07	7.76	7.76
E	6.95	7.30	7.31

Consequentemente, esses compostos propostos podem direcionar à síntese de novas drogas antitripanossoma. Além disso, esses resultados permitem a interpretação química sobre os requisitos estruturais de novos compostos com potencial antitripanossoma: tiosemicarbazonas contendo um anel fenila tri-substituído no grupo R1 são promissores, enquanto que o cloro na posição 3 é preferido, em vez do grupo triflúor-metila.

4 CONCLUSÃO

A representação molecular da abordagem MIA-QSAR foi melhorada com a introdução de círculos com diferentes tamanhos e cores correspondentes aos átomos nas moléculas. Os métodos aug-MIA-QSAR e aug-MIA-QSARcolor mostraram-se mais preditivos para a atividade antitripanossoma dos derivados de tiosemicarbazonas e semicarbazonas do que a aproximação MIA-QSARt. A incorporação de cores proporcionais à eletronegatividade de Pauling foi codificada pelo modelo aug-MIA-QSARcolor e resultou em uma maior capacidade preditiva quando comparado aos demais modelos baseados em descritores MIA.

A combinação dos compostos mais ativos da série resultou em estruturas com grande potencial antitripanossoma, decorrente da presença de um anel fenila trisubstituído no grupo R1 e um grupo triflúor-metila na posição 3 em vez de um cloro.

REFERÊNCIAS

ADVANCED CHEMISTRY DEVELOPMENT-ACD. **ACD/ChemSketch - version 12.01**. Toronto, Advanced Chemistry Development, 2009. Disponível em: <http://alainsheratan.com/chemsk_tl2.pdf>. Acesso em: 10/ 02/2013.

BERALDO, H. Semicarbazonas e tiosemicarbazonas: o amplo perfil farmacológico e usos clínicos. **Química Nova**, São Paulo, v. 27, n. 3, p. 461-471, June 2004.

BRASIL. Ministério da Saúde. **Doença de Chagas aguda no Brasil: série histórica de 2000 a 2013**. Boletim epidemiológico. vol.46 , nº 21, 2015. Disponível em: <<http://portalsaude.saude.gov.br/index.php/o-ministerio/principal/leia-mais-o-ministerio/197-secretaria-svs/11955-boletins-epidemiologicos-arquivos>>. Acesso: 15/02/2015

BRASIL. Ministério da Saúde. **Doenças infecciosas e parasitárias: guia de bolso**. 8 ed., Brasília: Ed. MS, 2010, 444p.

CALDAS, I. S. et al. Benznidazole therapy during acute phase of Chagas disease reduces parasite load but does not prevent chronic cardiac lesions. **Parasitology Research**, New York, v. 103, n. 2, p. 413-421, July 2008.

CENTERS FOR DIASEASE CONTROL AND PREVENTION- CDC. American Trypanosomiasis (also known as Chagas Disease). 2015. Disponível em: < <http://www.cdc.gov/parasites/chagas/biology.html> >. Acesso em: 20/09/2015.

CHAGAS, C. Nova tripanozomiaze humana: estudos sobre a morfologia e o ciclo evolutivo do Schizotrypanum cruzi n. gen., n. sp., agente etiológico de nova entidade morbida do homem. **Memórias do Instituto Oswaldo Cruz**, Rio de Janeiro, v. 1, n. 2, p. 159-218, Ago. 1909.

COURA, J. R. Tripanosomose, doença de Chagas. **Ciência e Cultura**, São Paulo, v. 55, n. 1, p. 30-33, Jan. 2003.

COURA, J. R.; DIAS, J. C. P. Epidemiology, control and surveillance of Chagas disease: 100 years after its discovery. **Memórias do Instituto Oswaldo Cruz**, Rio de Janeiro, v. 104, p. 31-40, July 2009.

DIAS, L. C. et al. Quimioterapia da doença de Chagas: estado da arte e perspectivas no desenvolvimento de novos fármacos. **Química Nova**, São Paulo, v. 32, n. 9, p. 2444-2457, 2009

GENNINGTON, R. D.; KEITH, T.A; MILLAM, J.M. **GaussView 5.0. 8**; Gaussian Inc., Wallingford, 2008. CD-ROM.

DOYLE, P. S. et al. A cysteine protease inhibitor cures Chagas' disease in an immunodeficient-mouse model of infection. **Antimicrobial Agents and Chemotherapy**, Washington, v. 51, n. 11, p. 3932-3939, Nov. 2007.

DU, X. et al. Synthesis and structure-activity relationship study of potent trypanocidal thio semicarbazone inhibitors of the trypanosomal cysteine protease cruzain. **Journal of Medicinal Chemistry**, Washington, v. 45, n. 13, p. 2695-2707, May 2002.

FERREIRA, H. D. O. Tratamento da forma indeterminada da doença de Chagas com nifurtimox e benzonidazol. **Revista da Sociedade Brasileira de Medicina Tropical**, Uberaba, v. 23, n. 4, p. 209-211, Out./Dez. 1990.

FERREIRA, R. T. B.; BRANQUINHO, M. R.; LEITE, C. P. Transmissão oral da doença de Chagas pelo consumo de açaí: um desafio para a Vigilância Sanitária. **Vigilância Sanitária em Debate**, Rio de Janeiro, v. 2, n. 4, p. 4-11, Mai. 2014.

FREITAS, M. P. MIA-QSAR modelling of anti-HIV-1 activities of some 2-amino-6-arylsulfonylbenzotrioles and their thio and sulfinyl congeners. **Organic & Biomolecular Chemistry**, Cambridge, v. 4, n. 6, p. 1154-1159, Feb. 2006.

GEENBAUM, D. C. et al. Synthesis and structure-activity relationships of parasiticidal thiosemicarbazone cysteine protease inhibitors against *Plasmodium falciparum*, *Trypanosoma brucei*, and *Trypanosoma cruzi*. **Journal of Medicinal Chemistry**, Washington, v. 47, n. 12, p. 3212-3219, June 2004.

GOLBRAIKH, A.; TROPSHA, A. Beware of q²! **Journal of Molecular Graphics and Modelling**, New York, v. 20, n. 4, p. 269-276, Jan. 2002.

LOZANO, B. H. et.al. Molecular features for antitrypanosomal activity of thiosemicarbazones revealed by OPS-PLS QSAR studies. **Medicinal Chemistry**, Sharjah, v. 8, n. 6, p. 1045-1056, Nov. 2012.

MITRA, I.; SAHA, A.; ROY, K. Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. **Molecular Simulation**, Oxon, v. 36, n. 13, p. 1067-1079, Oct. 2010.

NUNES, C. A. et al. Chemoface: a novel free user-friendly interface for chemometrics. **Journal of the Brazilian Chemical Society**, Campinas, v. 23, n. 11, p. 2003-2010, Nov. 2012.

OJHA, P. K. et al. Further exploring r^2_m metrics for validation of QSPR models. **Chemometrics and Intelligent Laboratory Systems**, Berlin, v. 107, n. 1, p. 194-205, May 2011.

ORGANIZAÇÃO MUNDIAL DE SAÚDE-OMS. Chagas disease (American trypanosomiasis). 2015a. Disponível em: < <http://www.who.int/mediacentre/factsheets/fs340/en/>>. Acesso em 15/10/2015

ORGANIZAÇÃO MUNDIAL DE SAÚDE- OMS. **Investing to overcome the global impact of neglected tropical diseases: third WHO report on neglected diseases**. Geneva: Organização Mundial de Saúde, 2015b,191p. Disponível em: <<http://apps.who.int/bookorders/anglais/detart1.jsp?codlan=1&codcol=15&codcch=3853>>. Acesso em: 15/06/15

PEDROSA, R. C. et al. Time-dependent oxidative stress caused by benzimidazole. **Redox Report**, Leeds, v. 6, n. 4, p. 265-270, Aug. 2001.

SOUZA, J. B. T. Alvos moleculares promissores para o desenvolvimento de novos fármacos antichagásicos. **Augustus**, Rio de Janeiro, v. 19, n. 38, p. 9-28, Jul./Dez. 2014.

SOUZA, W. D. **Doenças negligenciadas**. Rio de Janeiro: A. Brasileira de Ciências, 2010. 56 p.

STREIGER, M. L. et al. Estudo longitudinal e quimioterapia específica em crianças, com doença de Chagas crônica, residentes em área de baixa endemicidade da República Argentina. **Revista da Sociedade Brasileira de Medicina Tropical**, Uberaba, v. 37, n. 5, p. 365-375, Oct. 2004.

TENÓRIO, R. P. et al. Tiossemicarbazonas: métodos de obtenção, aplicações sintéticas e importância biológica. **Química Nova**, São Paulo, v. 28, n. 6, p. 1030-1037, Dez. 2005.

CAPÍTULO 3: EXPLORANDO A ATIVIDADE ANTIMALÁRICA DE QUINOLONA-4(1*H*)-IMINAS USANDO MIA-QSAR

RESUMO

Uma série de derivados quinolona-4(1*H*)-iminas foi recentemente descoberta como antimalárica, visando tanto estágios exoeritrocitários quanto estágios eritrocitários de desenvolvimento do parasito, que correspondem à fase de sintomas clínicos. Dotadas de estabilidade química e metabólica, as quinolona-4(1*H*)-iminas se apresentam como promissores antimaláricos de duplo estágio. Três versões do método de análise multivariada de imagens aplicada à relação quantitativa estrutura-atividade (MIA-QSAR), ou seja, MIA-QSAR tradicional, aug-MIA-QSAR e aug-MIA-QSAR codificado por cores (aug-MIA-QSARcolor), foram aplicadas para modelar as atividades antimaláricas dessa série de compostos. Os vários modelos de regressão linear indicaram que o método aug-MIA-QSAR é mais preditivo e confiável ($r^2 = 0.8079$, $r^2_{cv} = 0,6647$, $r^2_m = 0,9111$) que os modelos MIA-QSARt ($r^2 = 0.7682$, $r^2_{cv} = 0,6055$ e $r^2_m = 0,2564$) e aug-MIA-QSARcolor ($r^2 = 0.7399$, $r^2_{cv} = 0,6304$ e $r^2_m = 0,6641$). Os descritores aug-MIA-QSAR selecionados foram utilizados para reconhecimento de padrões usando análise discriminante por mínimos quadrados parciais (PLS-DA), a fim de separar os compostos com baixa, moderada e alta bioatividade. A interpretação dos descritores aug-MIA-QSAR sugere que a presença de grupos doadores de elétrons nas posições R1 e R2, cadeias aminoalquílicas maiores na posição R3 e substituintes com maior volume atômico na posição R4 contribuem para o aumento da atividade antiplasmodica dos derivados quinolona-4(1*H*)-iminas.

Palavras-chave: Malária. Análise multivariada de imagem. Regressão linear múltipla. QSAR. PLS-DA.

1 INTRODUÇÃO

1.1 Aspectos gerais da malária

A malária é uma doença parasitária causada por protozoários do gênero *Plasmodium*, cujo vetor é a fêmea do mosquito *Anopheles* infectado (BRASIL, 2014). Endêmica em muitas regiões de clima tropical e subtropical, a malária é considerada um problema de saúde pública mundial, afetando principalmente regiões da África. Estimativas apontam que existem mais de 214 milhões de infecções em todo o mundo. Somente no ano de 2015, dados apontam que cerca de 438.000 mil mortes tenham ocorrido em razão da doença (OMS, 2015). Enquanto as estatísticas sugerem que uma redução de 37% na incidência de malária foi alcançada em todo o mundo no período de 2000 a 2012, cerca de 1.300 jovens perdem suas vidas a cada dia, principalmente crianças menores de 5 anos (OMS, 2013).

No Brasil, a malária ainda é uma das principais endemias parasitárias e, por essa razão, está entre as doenças brasileiras consideradas negligenciadas, uma vez que sua ocorrência e permanência estão relacionadas a questões sociais, tais como precárias condições de moradias e falta de saúde pública (BRASIL, 2014).

A área endêmica da doença no Brasil compreende a região amazônica brasileira, incluindo os estados do Acre, Amazonas, Amapá, Pará, Rondônia, Roraima, Tocantins, Mato Grosso e Maranhão. Essa região é responsável por 99% dos casos autóctones do país. Fora da região amazônica, mais de 80% dos casos registrados são importados dos estados pertencentes à área endêmica brasileira ou de outros países amazônicos, continente africano e Paraguai. Desde 2000, tem havido uma redução de mais de 50% no número de casos de malária no Brasil (BRASIL, 2014). Um estudo realizado por Lapouble e colaboradores

(2015) mostra que, em 2012, foram registrados 241.806 casos de malária, representando uma redução de 60,1% em relação a 2005 e de 9,1% em relação a 2011. Embora o número de casos e óbitos por malária tenha reduzido, a vigilância com a doença precisa ser efetiva, já que a transmissão persiste e novos surtos podem acontecer a qualquer momento, principalmente porque as regiões da Venezuela que fazem fronteira com o Brasil têm apresentado índices crescentes de infecções (PAN AMERICAN HEALTH ORGANIZATION - PAHO, 2013).

A malária é causada por cinco diferentes tipos de *Plasmodium spp*: *Plasmodium falciparum*, *P. vivax*, *P. malariae*, *P. ovale* e *P. knowlesi*, dentre os quais o *P. falciparum* é o que apresenta maiores taxas de complicações e mortalidade (MURRAY et al., 2012). No que se refere à distribuição geográfica, o *P. vivax* é a espécie mais amplamente distribuída no mundo, com maior ocorrência na Ásia Central, seguida do Sudeste da Ásia, Américas e África, respectivamente (GETHING et al., 2012). No Brasil, somente os três primeiros estão presentes, sendo o *P. vivax* e o *P. falciparum* as espécies predominantes (BRASIL, 2014).

O ciclo de transmissão da malária humana tem apenas dois hospedeiros, o homem e a fêmea do mosquito anofelino. Porém, o efeito patogênico ocorre apenas para o homem (CDC, 2015).

Os mosquitos anofelinos são assim chamados por pertencerem ao gênero *Anopheles*. Esse gênero compreende cerca de 400 espécies, sendo a espécie *Anopheles darlingi* a maior transmissora da doença, inclusive no Brasil. Essa espécie normalmente se desenvolve em águas de baixo fluxo, profundas, límpidas, sombreadas e com pouco aporte de matéria orgânica e sais. Entretanto, em situações de alta densidade, o *A. darlingi* acaba ocupando vários outros tipos de criadouro, incluindo pequenas coleções hídricas e criadouros temporários (BRASIL, 2014).

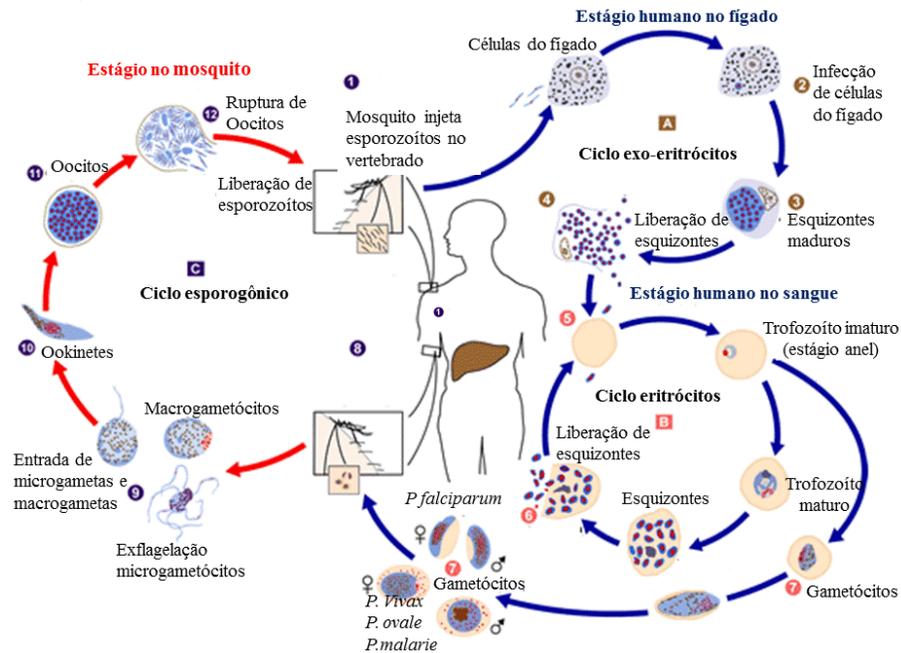
Além da transmissão por meio da picada da fêmea do mosquito do gênero *Anopheles* infectados com o *Plasmodium*, a doença também pode ser adquirida por meio do contato direto com o sangue de uma pessoa infectada (como, por exemplo, em transfusões sanguíneas ou transplante de órgãos, via placentária ou ainda por compartilhamento de seringas contaminadas). Contudo, essas formas são bastante raras (BRASIL, 2004).

Para que ocorra a transmissão do patógeno ao homem, inicialmente, o mosquito vetor tem que ser infectado por uma pessoa portadora do parasito. Ao alimentar-se de sangue de uma pessoa contaminada, os gametócitos circulantes no hospedeiro vertebrado são transportados ao sistema digestivo do mosquito vetor, onde amadurecem. Os gametócitos machos e fêmeas fundem-se e formam um zigoto que se desenvolve em oocineto tetraploide móvel, que pode penetrar as paredes do intestino médio e, após meiose, forma os oocistos. Os oocistos, por sua vez, aumentam de tamanho e rompem as células, liberando os esporozoítos, que migram para as glândulas salivares do mosquito, onde ficam disponíveis para infectar novos vertebrados (CDC, 2015; DRAKELEY et al., 2006).

Ao inocular a forma infectante do parasito (esporozoítos) no hospedeiro vertebrado, no caso o homem, através da saliva, os parasitos, por meio da corrente sanguínea, chegam rapidamente ao fígado, onde penetram nos hepatócitos e multiplicam por reprodução assexuada (ciclo exo-eritrocítico), originando milhares de merozoítos hepáticos, que são liberados para as células sanguíneas. Nas células sanguíneas, inicia-se o ciclo assexuado eritrocitário, onde os merozoítos transformam-se em trofozoítos, os quais, ao atingir a fase adulta, transformam-se em esquizontes, os quais originarão outros merozoítos capazes de infectar novos eritrócitos (DRAKELEY et al., 2006; BAER et al., 2007). Ainda na fase inicial trofozoítos, alguns parasitos se diferenciam em estágios sexuais eritrocitários (gametócitos), forma disponível para novas

infecções em invertebrados (CDC, 2015). A Figura 12 retrata o ciclo do parasito no mosquito vetor e no homem.

Figura 12 - Ciclo de vida do *Plasmodium* no mosquito e no homem.



Fonte: Adaptado de CDC (2012)

Os sintomas clínicos associados à malária aparecem no estágio em que o parasito se encontra no sangue. Nessa fase, as principais manifestações clínicas são: febre, associada ou não a calafrios, tremores, suores intensos, dor de cabeça, dores no corpo, anemia em graus variáveis, vômitos, diarreia, dor abdominal, falta de apetite, tonteira e sensação de cansaço (BRASIL, 2004). A febre na malária corresponde ao momento em que as hemácias estão se rompendo. Após a fase inicial, a febre assume um caráter intermitente, dependente do tempo de duração dos ciclos eritrocíticos de cada espécie de plasmódio: 48 horas para *P. falciparum* e *P. vivax* (malária terçã), e 72 horas para *P. malariae* (malária

quartã) (BRASIL, 2004). A forma grave da doença aparece quando ocorre falência de órgãos ou graves anomalias no sangue ou no metabolismo do paciente. As manifestações desse estágio incluem forte cefaleia, hipertermia, vômitos, convulsões (malária cerebral), insuficiência renal aguda, edema pulmonar agudo, hipoglicemia, disfunção hepática, hemólise intravascular aguda maciça e choque, que pode levar a óbito em média 10% dos casos (CDC, 2015; BRASIL, 2004).

1.2 Tratamento

Existem alguns medicamentos disponíveis para o tratamento da malária, tais como primaquina, quinino, cloroquina, mefloquina, artemisina, artesunato, entre outros. No Brasil, o primaquina e a cloroquina são indicados como tratamento de primeira linha para o *P. vivax*, enquanto o primaquina e o quinino são indicados para *P. falciparum*, e a mefloquina é um tratamento de segunda linha (BRASIL, 2004). A indicação de tais medicamentos depende do tipo de *Plasmodium*, da fase de infecção e também devido aos casos de resistência em determinada região. Os principais obstáculos ao tratamento da malária são os efeitos colaterais dos medicamentos, que ocasionam interrupção do tratamento, a não disponibilização dos medicamentos nas áreas mais remotas e a baixa eficácia para atingir as fases de infecções no sangue e fígado (BRASIL, 2004; OMS, 2011).

Atualmente, o controle e erradicação da malária tem dependido tanto dos medicamentos disponíveis como da eliminação do vetor (ZHANG et al., 2012). Ambos são sérios problemas para os governos, pois as regiões endêmicas são áreas de reprodução favoráveis para os mosquitos, quer devido a condições climáticas ou pelos métodos inadequados de controle de vetores. Por outro lado, a eficácia reduzida de medicamentos antipalúdicos decorrente do aumento da

resistência, tem sido motivo de grande preocupação, principalmente porque terapias à base de artemeter-lumefantrina artemisinina (ACTs), artesunato-mefloquina e artesunatosulfadoxina-pirimetamina, recentemente empregadas no tratamento da doença, já apresentaram relatos de resistências em pelo menos uma região de ocorrência da malária (OMS, 2015). Portanto, para lidar com as taxas de resistência em desenvolvimento rápido, incansáveis esforços estão em curso para desenvolver novas entidades moleculares ou modificar as já existentes para serem capazes de lidar com o parasito em diferentes fases do seu ciclo de vida.

Estudos recentes sobre quinolona-4(1H)-iminas (LOPES et al., 2004; RESSURREIÇÃO et al., 2013; RODRIGUES et al., 2013) revelaram que esses compostos possuem atividade antiplasmódica de dupla-fase, ou seja, para as fases de infecção do parasito no sangue e no fígado. No entanto, mais estudos são necessários para obter um maior entendimento sobre o efeito dos diferentes ligantes (substituintes) na dupla atividade antiplasmódica, nos modos de interação da proteína alvo-ligante e para preencher as lacunas de conhecimento sobre o mecanismo de ação para essa classe de compostos. Juntos, esses aspectos fornecem conhecimentos sobre as possíveis modificações estruturais que produzem compostos mais potentes e de largo espectro. Métodos computacionais têm, cada vez mais, encontrado importante aplicabilidade no fornecimento de respostas a essas interrogações na descoberta de medicamentos. As vantagens e papel dessas estratégias alternativas não podem ser subestimados e, nos dias atuais, representam um componente-chave do *pipeline* de descoberta de novas drogas.

Este capítulo tem por objetivo avaliar a relação quantitativa estrutura-atividade (QSAR) de derivados de quinolona-4(1H)-iminas e elucidar a influência subjacente dos diferentes substituintes na sua atividade antiplasmódica, a partir de uma perspectiva de análise multivariada de imagens

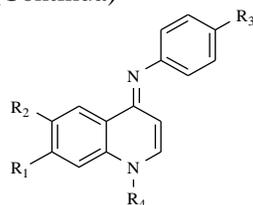
(MIA). Levando em consideração que mais duas abordagens foram incorporadas ao método MIA-QSAR, a primeira tarefa é comparar o seu desempenho nos estudos de correlação com os derivados de quinolona-4(1*H*)-imina. Posteriormente, são feitos esforços para interpretar o melhor modelo obtido, a fim de revelar informações codificadas em termos do efeito dos substituintes na atividade antiplasmódica.

2 METODOLOGIA

2.1 Obtenção dos descritores MIA

No presente estudo, uma série de 23 compostos derivados de quinolona-4(1*H*)-iminas com atividade antimalárica foi obtida da literatura (RESSURREIÇÃO *et al.*, 2013) (Quadro 2).

Quadro 2 - Estrutura química e atividade antiplasmódica de derivados quinolona-4(1*H*)-iminas. (Continua)

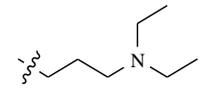
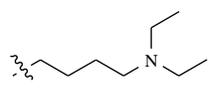
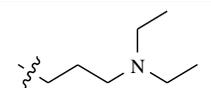
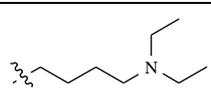
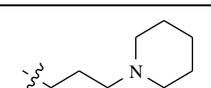
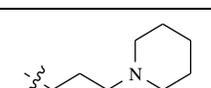


^a Cp	R ₁	R ₂	R ₃	R ₄	IC ₅₀	pIC ₅₀
1*	Cl	H	Ph		498	6,30
2	Cl	H	Ph		111	6,95
3	Cl	H	Ph		144	6,84

Quadro 2 (Continua)

Cp	R ₁	R ₂	R ₃	R ₄	IC ₅₀	pIC ₅₀
4*	Cl	H	Cl		259	6,59
5	Cl	H	Br		74	7,13
6	Cl	H	H		273	6,56
7	Cl	H	Me		331	6,48
8*	Cl	H	F		369	6,43
9	Cl	H	OPh		160	6,80
10	CF ₃	H	Ph		525	6,28
11	CF ₃	H	Ph		226	6,65
12	CF ₃	H	Ph		360	6,44
13		H	Ph		216	6,67
14*		H	Ph		55,4	7,26
15		H	Ph		357	6,45
16		H	Ph		53,7	7,27

Quadro 2 (Conclusão)

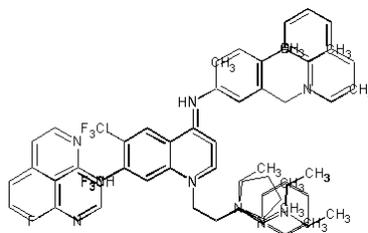
Cp	R ₁	R ₂	R ₃	R ₄	IC ₅₀	pIC ₅₀
17*	Cl	H	Ph		151	6,82
18	Cl	H	Ph		67,6	7,17
19	CF ₃	H	Ph		144	6,84
20	CF ₃	H	Ph		50,3	7,30
21	H	Cl	Ph		99,7	7,00
22	H	CF ₃	Ph		631	6,20

^a Compostos com asterisco (*) foram selecionados para o grupo teste. ^b Valores de IC₅₀ em nM para o estágio do parasito no sangue. Fonte: Ressurreição et al. (2013).

Nota-se que, a partir da base de dados original, o primaquina (PQ) foi excluído por não ter o valor de IC₅₀ relatado, ao passo que o composto identificado pelo número 3 também foi excluído em razão de o seu valor de IC₅₀ ter sido obtido a partir de uma fonte diferente e a homogeneidade das condições experimentais não pode ser verificada. Os valores da variável resposta (IC₅₀) foram transformados em pIC₅₀ para lidar com a assimetria dos pontos de dados.

O primeiro passo para a obtenção dos descritores MIA-QSARt foi o desenho das estruturas químicas no programa *ChemSketch* (ACD, 2009). O processo foi realizado conforme descrito no capítulo 2. Porém, a dimensão da área de trabalho utilizada foi 400×350 pixels e o pixel comum a todas estruturas foi alinhado na coordenada definida em 170×190 px. Este procedimento resultou na sobreposição das imagens das estruturas químicas, conforme exemplificado na Figura 13.

Figura13 - Sobreposição das imagens das estruturas químicas dos derivados quinolona-4(1H)-iminas, resultando em um arranjo tridimensional para a abordagem MIA-QSARt.



Fonte: Do autor (2016).

As imagens das estruturas químicas para a abordagem aug-MIA-QSAR e aug-MIA-QSARcolor foram construídas no programa GaussView (DENNINGTON, KEITEH, MILLAN, 2008). Na abordagem aug-MIA-QSARcolor, as cores foram selecionadas para serem proporcionais à eletronegatividade de Pauling, enquanto que na abordagem aug-MIA-QSAR as cores seguiram o padrão do programa (Tabela 5). O grande diferencial do presente trabalho em relação aos já publicados utilizando a abordagem aug-MIA-QSAR é que, ao utilizar cores sólidas, sem efeito de iluminação dos átomos (*spotlight*), os círculos não apresentam gradiente de cores e isso faz com que apenas um valor de pixel represente um determinado átomo. Tal modificação permite a correta identificação de um átomo em uma matriz de dados e, por conseguinte, a interpretação das informações químicas contidas nos pixels (descritores moleculares). Outro fator relevante é que os valores de pixels para a abordagem aug-MIA-QSARcolor passaram a ter relação direta com a eletronegatividade. Por exemplo, o átomo de hidrogênio possui eletronegatividade igual a 2,1. Para determinar o seu valor de pixel, o valor de eletronegatividade (2,1) foi multiplicado por um fator (100), para que as cores tivessem correlação de 1, resultando no valor 210 (Tabela 5).

Tabela 5 - Valores de pixels para os átomos na abordagem aug-MIA-QSAR e valores de pixels e escala de cores para os átomos na abordagem aug-MIA-QSARcolor, segundo a escala de eletronegatividade de Pauling (ϵ).

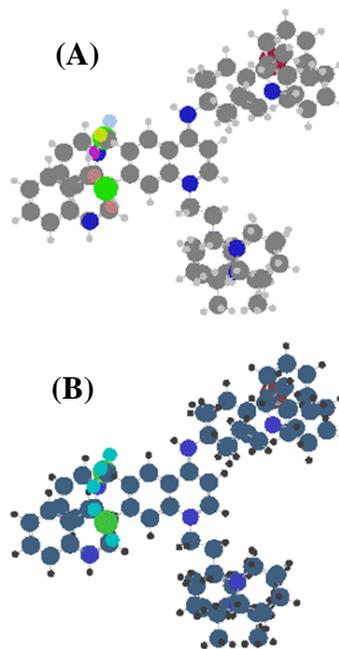
Átomos	ϵ	aug-MIA-QSAR		aug-MIA-QSARcolor		
		Pixel	Pixel	Vermelho	Verde	Azul
H	2.1	612	210	70	70	70
C	2.5	426	250	50	100	100
N	3.0	279	300	50	50	200
O	3.5	229	350	250	50	50
F	4.0	688	400	400	0	0
S	2.5	493	250	125	125	0
Cl	3.0	289	300	50	200	50
Br	2.8	231	280	130	75	75
I	2.5	294	250	125	0	125

Fonte: Do autor (2016).

Após o desenho de cada estrutura, a imagem foi transferida para uma janela do aplicativo *Paint* da *Microsoft Windows* e salvas como *bitmaps*. Nesse processo, as imagens foram automaticamente alinhadas em uma janela de tamanho definido em 342×300 px, obtendo, assim, a sobreposição das estruturas químicas e, conseqüentemente, o arranjo tridimensional, em que as partes comuns entre as estruturas da série congênere (esqueleto básico) mostram-se congruentes (Figura 14).

No programa Chemoface (NUNES et al., 2012) as imagens alinhadas foram desdobradas em matrizes numéricas com as seguintes dimensões 23×140000 para MIA-QSARt e 23×102600 para aug-MIA-QSAR e aug-MIA-QSARcolor. Subseqüentemente, colunas com variância nula (espaços brancos comuns para todas as imagens e partes congruentes das estruturas químicas) foram eliminadas para reduzir a dimensionalidade das matrizes de dados e

Figura 14 - Sobreposição da série congênere de compostos que ilustram a subestrutura comum e as porções variáveis dos derivados quinolona-4(1*H*)-imina com atividade antimalárica para as abordagens aug-MIA-QSAR (A) e aug-MIA-QSARcolor (B).



Fonte: Do autor (2016).

agilizar os cálculos, resultando em matrizes de tamanho 23×4588 para MIA-QSAR e 23×11586 para aug-MIA-QSAR e aug-MIA-QSARcolor.

As abordagens MIA-QSAR produzem um número extremamente grande de descritores em relação ao número de compostos, e a abordagem clássica, quando lida com tais matrizes de dados, utiliza técnicas estatísticas que implicam em redução da dimensionalidade, tais como mínimos quadrados parciais (PLS) e regressão por componentes principais (PCR). Em princípio, esses métodos consistem na transformação das variáveis originais em algumas variáveis latentes (ou componentes), que são projeções ortogonais que codificam

toda a informação contida na matriz de dados original (WATERBEEND et al., 2008). São essas variáveis que, posteriormente, serão utilizadas no processo de construção do modelo. No entanto, esses procedimentos não são vantajosos em termos de recuperar a informação original (isto é, variáveis) das projeções ortogonais. Portanto, se o objetivo é explorar a contribuição das diferentes características, esse objetivo se torna complexo e desafiante. Adicionalmente, ao utilizar PLS e PCR assume-se que algumas variáveis são irrelevantes para a propriedade modelada e, assim, o risco de usar muito menos variância das variáveis independentes para explicar as variáveis dependentes pode ser cometido. Por isso, nesta parte da tese, explorou-se o método de regressão linear múltipla como técnica estatística para a calibração.

Dado o espaço de alta dimensionalidade encontrado nas matrizes de dados geradas pelos métodos MIA-QSAR, foram empregados procedimentos de pré-filtragem para produzir um subconjunto com menos variáveis usando as seguintes estratégias: em primeiro lugar, procedeu-se com a medida da entropia de Shannon (SE), que é um filtro à base de classificação não supervisionada, aplicado às variáveis utilizando um regime de 22 intervalos discretos (GODDEN; STAHURA; BAJORATH, 2000; BARIGYE et al., 2013). As variáveis com menos do que 5% do máximo SE ($SE_{MAX} = \text{Log}22$) foram descartadas. Este procedimento foi realizado no programa IMMAN (URIAS, et al. 2015). No entanto, a medida SE é um filtro à base relevância, isto é, não leva em consideração a possível redundância das variáveis, que é conhecida por ter efeitos prejudiciais sobre o poder preditivo de modelos MLR devido à multicolinearidade das variáveis. Nesse sentido, os conjuntos de variáveis reduzidas foram ainda filtrados através do coeficiente de correlação (x/x), em que para cada conjunto de variáveis com um coeficiente de correlação 0,99, uma variável é retida. Ao final desse processo, as matrizes de dados foram

comprimidas para um total de 145, 154 e 168 variáveis para as abordagens MIA-QSARt, aug-MIA-QSAR e aug-MIA-QSARcolor, respectivamente.

2.2 Grupo teste e treinamento

O conjunto de dados foi dividido em grupo teste e treinamento utilizando o método de análise de *cluster*, com base nas variáveis obtidas para a abordagem MIA-QSARt (sistema binário), que foi considerada como estratégia padrão para comparação. Inicialmente, a análise de agrupamento hierárquico foi realizada utilizando o algoritmo de Ward (WALLER *et al.*, 2008) e a distância euclidiana ao quadrado (MINGOTI, 2005) como a regra de fusão e medida de dissimilaridade, respectivamente, para determinar o número ideal de *clusters* para o conjunto de dados estudado. A distância que corresponde ao aumento mais acentuado no cronograma amálgama foi identificada e utilizada para determinar o número ótimo de aglomerados do dendrograma, que foi determinado como sendo igual a cinco. Subsequentemente, o conjunto de dados foi dividido usando análise de agrupamento K-média para um valor de K igual a cinco. Para cada grupo, os compostos foram selecionados aleatoriamente para constituir os conjuntos de treinamento e teste, que foram formados por 17 (78%) e 5 (22%) compostos, respectivamente.

2.3 Construção e validação do modelo MLR

Os modelos MLR foram construídos usando o subconjunto de variáveis obtidas após os procedimentos de redução de dimensionalidade de acordo com as estratégias MIA-QSARt, aug-MIA-QSAR e aug-MIA-QSARcolor, respectivamente, e o melhor modelo foi selecionado em cada caso, levando em consideração o parâmetro de validação cruzada *leave-one-out* (r^2_{cv}) como

função de otimização. Ao final desse processo, cada modelo foi selecionado com 4 variáveis (descritores). Esse procedimento foi realizado no programa MobiDigs (TODESCHINI et al., 2003) e utilizou o algoritmo genético (GA) para otimização. Posteriormente, os modelos foram validados utilizando 10 procedimentos Y-randomização (r^2 y-rand, ${}^c r^2$ p e RMSE y-rand) e validação externa [r^2 pred, e RMSEp], respectivamente, com o último executado no conjunto de validação de compostos anteriormente não utilizados na construção do modelo. O primeiro teste avalia a possibilidade de correlações fortuitas, enquanto o segundo avalia o verdadeiro poder preditivo do modelo. O coeficiente de determinação modificado r^2 m, também foi utilizado para verificar se os resultados da validação externa exibem boa correlação entre os valores experimentais e preditos e se ambos são similares.

3 RESULTADOS E DISCUSSÕES

3.1 Regressão Linear Múltipla

Sob uma perspectiva prática, bons modelos QSAR devem ser robustos (avaliados pelo procedimento de validação cruzada), possuírem bom poder preditivo (avaliado pelo desempenho na validação externa) e não propensos a correlações casuais (medidas pela estratégia de Y-randomização). Além disso, o domínio de aplicação do modelo (AD) deve ser claramente delimitado, de modo a determinar o espaço de aplicabilidade para o modelo. As equações dos modelos para as abordagens MIA-QSARt, aug-MIA-QSAR e aug-MIA-QSARcolor, bem como os parâmetros estatísticos correspondentes são mostrados abaixo pelas Equações 1, 2 e 3, respectivamente:

Equação 1 (MIA-QSARt):

$pIC_{50} = 8,0170 - 0,2946 \times X_{885} - 0,4846 \times X_{1432} - 0,8453 \times X_{2343} - 0,4034 \times X_{330}$
 $N_{train}=17, N_{test}=5, r^2_{cal} = 0,7682, RMSEc = 0,1586, r^2_{cv} = 0,6055, RMSEcv = 0,2092, r^2_{y-rand} = 0,2160, RMSE_{y-rand} = 0,2899, r^2_{p(y-rand)} = 0,5708, r^2_{pred} = 0,7973, r^2_m (test) = 0,2564, RMSEp = 0,1865.$

Equação 2 (aug-MIA-QSAR):

$pIC_{50} = 7,7081 - 0,0010 \times X_{1422} + 0,0025 \times X_{2069} - 0,0030 \times X_{3659} - 0,0011 \times X_{5845}$
 $N_{train}=17, N_{test}=5, r^2_{cal} = 0,8079, RMSEc = 0,1444, r^2_{cv} = 0,6647, RMSEcv = 0,1931, r^2_{y-rand} = 0,2518, RMSE_{y-rand} = 0,2840, r^2_{p(y-rand)} = 0,6025, r^2_{pred} = 0,9691, r^2_m (test) = 0,9111, RMSEp = 0,1353.$

Equação 3 (aug-MIA-QSARcolor):

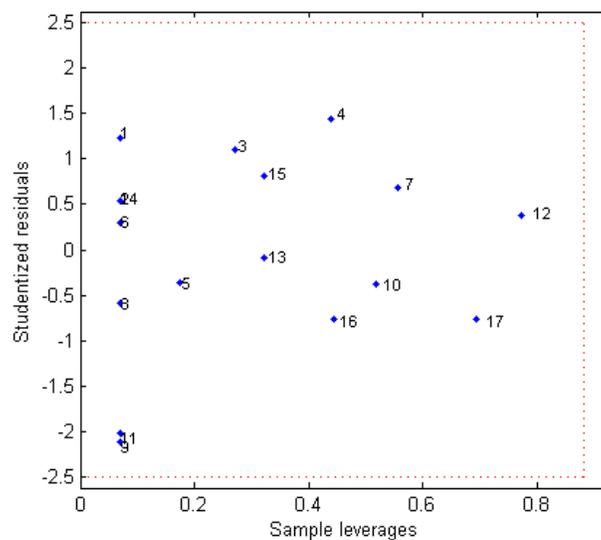
$pIC_{50} = 2,2831 + 0,0021 \times X_{2024} + 0,0034 \times X_{2568} - 0,0029 \times X_{2752} + 0,1010 \times X_{6111}$
 $N_{train}=17, N_{test}=5, r^2_{cal} = 0,7399, RMSEc = 0,1680, r^2_{cv} = 0,6304, RMSEcv = 0,2081, r^2_{y-rand} = 0,1323, RMSE_{y-rand} = 0,3062, r^2_{p(y-rand)} = 0,5767, r^2_{pred} = 0,8404, r^2_m (test) = 0,6641, RMSEp = 0,1452.$

Em geral, os três modelos apresentaram bom desempenho, com capacidade preditiva externa robusta e não propensos à correlação ao acaso, requisitos importantes para garantir a qualidade dos modelos QSAR (GOLBRAIKH; TROPSHA, 2002). Na comparação das três estratégias MIA-QSAR pode-se observar que a aproximação aug-MIA-QSAR apresentou parâmetros estatísticos com rendimentos médios superiores ao MIA-QSARt e aug-MIA-QSARcolor. Particularmente, o modelo aug-MIA-QSAR produziu valores de $r^2_{cv} = 0,6647$ e $r^2_{pred} = 0,9691$ superiores ao aug-MIA-QSARcolor ($r^2_{cv} = 0,6304$ e $r^2_{pred} = 0,8404$) e MIA-QSAR ($r^2_{cv} = 0,6055$ e $r^2_{test} = 0,7973$), demonstrando que o modelo aug-MIA-QSAR possui maior robustez e poder preditivo. Adicionalmente, o modelo aug-MIA-QSAR é menos influenciado por correlação fortuita. Além disso, o valor de r^2_m , que é um parâmetro que certifica se os valores experimentais e preditos são congruentes (MITRA; SAHA; ROY, 2010; OJHA et al., 2011; ROY et al., 2012), foi maior para aug-MIA-QSAR

($r^2_m = 0,9111$), seguido de aug-MIA-QSARcolor ($r^2_m = 0,6641$), mas menor do que o valor recomendado (0,5) para o MIA-QSARt ($r^2_m = 0,2564$).

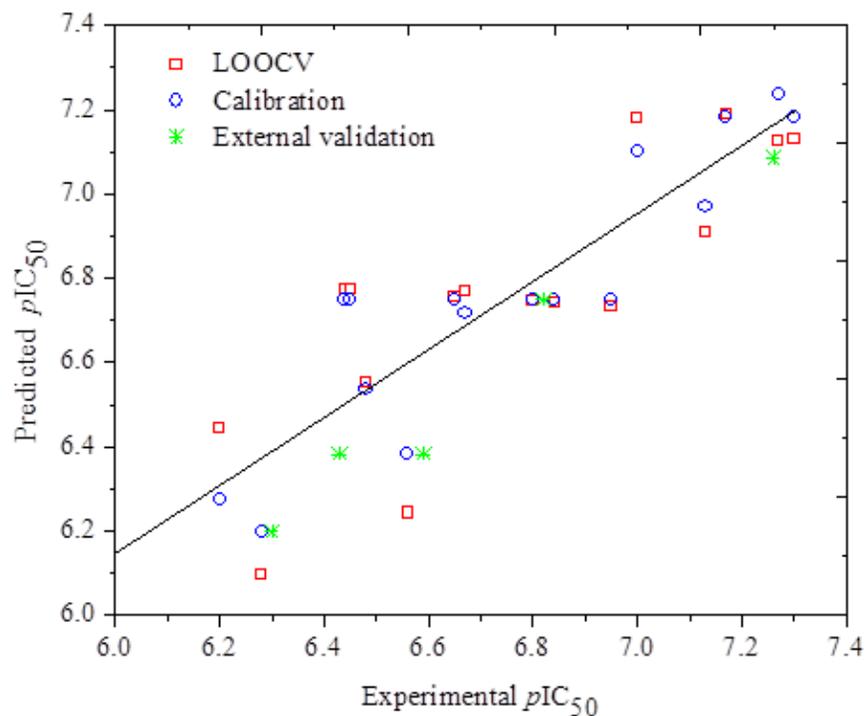
De acordo com o gráfico de William (Gráfico 5) para o modelo aug-MIA-QSAR, tanto os compostos de treinamento quanto os de teste encontram-se definidos no domínio de aplicabilidade do modelo, o que demonstra a sua aplicabilidade. Além disso, o gráfico de dispersão dos valores experimentais contra preditos (Gráfico 6) mostra que os compostos da série de treinamento e teste possuem apenas pequenos desvios da linha de ajuste.

Gráfico 5 - Resíduo de *Student* versus *leverages* para o modelo aug-MIA-QSAR.



A partir de uma perspectiva crítica, o desempenho superior do modelo aug-MIA-QSAR sugere uma maior relevância do tamanho dos substituintes (como uma propriedade) para a atividade antiplasmódica das quinilona-4(1*H*)-iminas do que a eletronegatividade. Diferentes estudos em relação às propriedades atômicas têm sugerido a existência de uma forte correlação entre

Gráfico 6: Valores de pIC_{50} experimentais *versus* preditos utilizando o modelo aug-MIA-QSAR.



volume de van der Waals e comportamento inespecífico de moléculas, tal como a hidrofobicidade (COWLEY, 2011; RODRIGUES, 2013).

De fato, estudos prévios sobre a atividade antimalárica de quinolona-4(1*H*)-iminas têm revelado que interações hidrofóbicas com a proteína-alvo do parasito desempenham um papel crucial na sua atividade antiplasmódica (YÉPEZ et al., 2000; RODRIGUES et al., 2013). Estudos experimentais sobre a influência das posições dos diferentes substituintes no esqueleto dos quinolona-4(1*H*)-imina com atividade antiplasmódica revelaram que existe uma boa relação linear entre os valores Hammett π (constantes hidrofóbica) para diferentes substituintes e a correspondente atividade contra o parasito (RESSURREIÇÃO et al., 2013). Por outro lado, uma fraca correlação foi

encontrada com a constante de Hammett σ (constantes eletrônicas) dos substituintes das quinolons-4(1H)-iminas e a sua atividade antiplasmódica (RESSURREIÇÃO et al., 2013). Essas descobertas podem explicar o desempenho superior do modelo construído de acordo com a abordagem aug-MIA-QSAR, em comparação com os modelos aug-MIA-QSARcolor (baseado na eletronegatividade) e MIA-QSARt. No entanto, é interessante notar que a incorporação de informação química atômica ao método binário MIA-QSAR tradicional produz modelos com maior robustez e capacidade de previsão.

3.2 Aspectos sobre a interpretação do modelo

A interpretação em QSAR desempenha um papel importante no processo de descoberta de drogas, uma vez que permite uma maior visão sobre as propriedades estruturais ou parte de grupos funcionais responsáveis por aumentar (ou diminuir) a potência, bem como sobre outros aspectos que afetam a qualidade de medicamentos, tais como a toxicidade, biodisponibilidade, reações adversas, etc. Por conseguinte, é de interesse conhecer a contribuição das diferentes variáveis para o modelo que explica a atividade antiplasmódica dos compostos estudados, bem como as características dessas variáveis e as informações que elas codificam. Nesse sentido, os compostos do conjunto de dados foi dividido de acordo com sua atividade antiplasmódica (pIC_{50}) em 3 grupos, isto é, compostos de baixa (**22, 10, 1, 8, 12, 15, 7, 6, 4, 11, 13**), média (**9, 17, 3, 19, 2, 21, 5, 18**) e alta (**14, 16, 20**) atividade, a fim de avaliar a importância das variáveis do modelo nos diferentes grupos. Essa iniciativa é, portanto, traduzida em um problema de reconhecimento de padrão, em que se procura classificar os compostos de acordo com os três grupos. O gráfico de *scores* oferece uma ferramenta valiosa para estudar a influência das variáveis na atividade dos compostos químicos. No entanto, esse gráfico não é possível de se

obter por meio de análise discriminante linear (LDA), já que este não transforma as variáveis em projeções ortogonais em um hiperplano. No entanto, sabe-se que quando as variáveis em um modelo MLR não são correlacionadas, os métodos MLR e PLS produzem resultados idênticos (WOLD; SJÖSTRÖM; ERIKSSON, 2001). Portanto, se as variáveis no modelo aug-MIA-QSAR não estão correlacionadas, a PLS-DA pode ser utilizada para identificar as variáveis importantes para as diferentes classes. Posteriormente, as variáveis originais do modelo MLR foram analisadas para obter a informação codificada, a fim de racionalizar a sua importância para determinadas subestruturas ou porções de grupo funcionais sobre a bioatividade dos compostos. A Tabela 6 mostra os coeficientes de correlação para as quatro variáveis utilizadas na equação MLR.

Tabela 6: Coeficientes de correlação para as variáveis do modelo aug-MIA-QSAR.

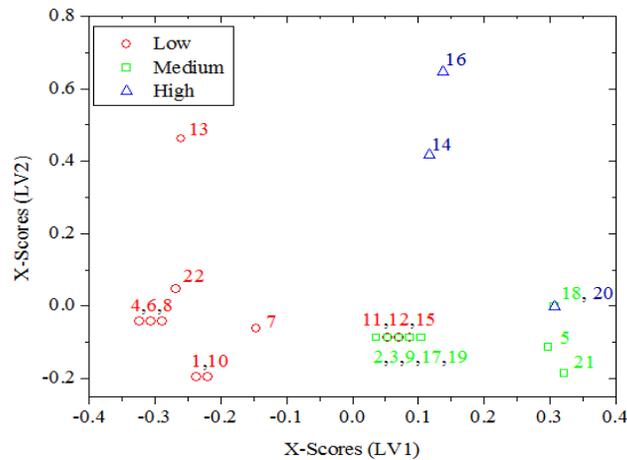
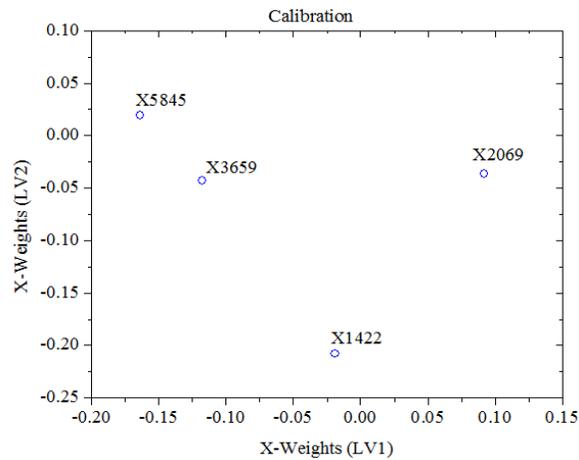
	X1422	X2069	X3659	X5845
X1422	1.00			
X2069	-0.02	1.00		
X3659	0.02	0.00	1.00	
X5845	-0.10	0.02	-0.02	1.00

A Tabela 7 contém as variáveis selecionadas para a construção do modelo aug-MIA-QSAR e segundo uma análise destas variáveis foi possível identificar as orientações aproximadas de cada pixel (variável) selecionado. A variável **X1422** localiza-se na região do substituinte R1, **X2069** em R2, **X3659** em R4 e **X5845** em R3.

Tabela 7 - Variáveis selecionadas para construção do modelo aug-MIA-QSAR.

Cp	X1422 (em R1)	X2069 (em R2)	X3659 (em R4)	X5845 (em R3)	pIC ₅₀
1	750	612	612	426	6.3
2	750	612	426	426	6.95
3	750	612	426	426	6.84
4	750	612	426	750	6.59
5	750	612	426	231	7.13
6	750	612	426	750	6.56
7	750	612	426	612	6.48
8	750	612	426	750	6.43
9	750	612	426	426	6.8
10	750	612	612	426	6.28
11	750	612	426	426	6.65
12	750	612	426	426	6.44
13	426	612	426	750	6.67
14	426	612	426	426	7.26
15	750	612	426	426	6.45
16	279	612	426	426	7.27
17	750	612	426	426	6.82
18	750	612	279	426	7.17
19	750	612	426	426	6.84
20	750	612	279	426	7.3
21	750	750	426	426	7
22	750	426	426	426	6.2

De acordo com os gráficos de *scores* e *loadings* (Gráficos 7 e 8), baixos valores para a variável **X1422** separam dois compostos de elevada (**14** e **16**) e um de baixa bioatividade (**13**) dos demais compostos da série. Logo, substituintes em R1 parecem não ser decisivos na classificação de compostos com baixa, média e alta atividade biológica, apesar de estudos prévios afirmarem que a maior atividade antiplasmódica é favorecida por substituintes

Gráfico 7 - PLS-DA para a série de quinolona-4(1*H*)-iminas.Gráfico 8 - Pesos das variáveis para a série de quinolona-4(1*H*)-iminas

doadores de elétrons em C-7 (R1) do anel quinolônico, enquanto substituintes retiradores de elétrons teriam efeito deletério (RESSURREIÇÃO et al., 2013; RODRIGUES et al., 2013).

Por outro lado, as variáveis **X2069**, **X3659** e **X5845** separam os compostos com média/elevada atividade biológica (*scores* positivos em LV1)

dos compostos com baixa bioatividade (*scores* negativos em LV1). Uma análise conjunta com os dados da Tabela 7 indica que, de forma geral, baixos valores em **X3659** e **X5845** (átomos de carbono indicando cadeia alquílica longa em R3 e R4, nitrogênio em R4 e bromo em R3) tendem a aumentar a atividade antimalárica dos compostos, enquanto a ausência de substituintes (765) ou grupos pequenos (612, hidrogênio) nessas posições acarreta diminuição na atividade biológica.

Um exame da variável **X5845** revela que os compostos com substituintes pequenos em R3, ou seja, Me (composto **7**), F (composto **8**), H (composto **6** e **13**) e Cl (composto **4**) são classificados com baixa atividade [identificado pelo pixel do hidrogênio metílico (612) e o restante com espaços em branco (765), devido ao seu tamanho muito pequeno], enquanto que o composto **5** pertence ao grupo de atividade média com átomos de maior tamanho Br (231). Uma comparação da atividade dos compostos **4**, **5**, **7** e **8** com os mesmos substituintes para R1, R2 e R4, mas diferindo apenas nos substituintes da posição R3, ou seja, Cl, Br, Me, e F, respectivamente, mostra a seguinte tendência na bioatividade: F>Me>Cl>Br. Essa descoberta sugere que um aumento na lipofilicidade para R3 resulta, geralmente, em um aumento na atividade antiplasmódica. Uma inferência semelhante foi apurada em estudos anteriores (RESSURREIÇÃO et al., 2013; RODRIGUES et al., 2013; ZHANG et al., 2012;)

Para exemplificar a tendência encontrada a partir da análise da variável **X3659**, a classificação dos compostos com ligantes *N*-dietílicos (composto **1** e **10**) pertencentes ao grupo de baixa atividade e dos compostos com cadeia de 4 átomos de carbono **18** e **20** pertencentes ao grupo de atividade média/alta, sugere que um aumento no comprimento da cadeia alquílica favorece um aumento na potência antiplasmódica. Esse resultado é consistente com o que é

reportado na literatura para R4 (RESSURREIÇÃO et al., 2013; RODRIGUES et al., 2013; ZHANG et al., 2012;).

A variável **X2069** varia apenas para os compostos **21** (765, média bioatividade) e **22** (426, baixa bioatividade) em relação aos demais (612), sugerindo que baixos valores em R2, tais como para halogênios e nitrogênio, diminuem a atividade antimalárica. De fato, tem sido reportado que grupos retiradores de elétrons em C6 (R2) diminuem a potência antiplasmódica de compostos quinolônicos (RESSURREIÇÃO et al., 2013).

4 CONCLUSÃO

A utilidade da abordagem MIA-QSAR na codificação de informações químicas relevantes para a bioatividade de quinolona-4(1*H*)-iminas foi demonstrada. O desempenho superior do modelo aug-MIA-QSAR sugere maior relevância do volume dos substituintes para a atividade antimalárica de compostos químicos estudados do que efeitos eletrostáticos e dipolares. Uma interpretação da informação codificada pelo modelo aug-MIA-QSAR produz as seguintes inferências quando se refere às características dos substituintes que aumentam atividade antiplasmódica: 1) R1 e R2: grupos doadores de elétrons são aconselhados para esta posição para maior atividade; 2) R3: cadeias aminoalquílicas maiores são favorecidas para essa posição, embora cuidados devam ser tomados para não influenciar significativamente o equilíbrio entre as propriedades biológicas e físico-químicas; 3) R4: substituintes com maior volume atômico são benéficos para essa posição. Esses resultados devem fornecer orientações importantes sobre as modificações estruturais apropriadas no esqueleto de quinolona-4(1*H*)-iminas, resultando em compostos antiplasmódicos mais eficientes, bem como a descoberta de novos esqueletos base que mais eficazes contra o *plasmodium*.

REFERÊNCIAS

- ADVANCED CHEMISTRY DEVELOPMENT. **ACD/ChemSketch - version 12.01**. Toronto, Advanced Chemistry Development, 2009. Disponível em: <http://alainsheratan.com/chemsk_t12.pdf>. Acesso em: 10 /02/2013.
- BAER, K. et al. Release of hepatic Plasmodium yoelii merozoites into the pulmonary microvasculature. **PLoS Pathogens**, San Francisco, v. 3, n. e171, ed. 11, p. 1651-1668, Nov. 2007.
- BARIGYE, S. J. et al. Relations frequency hypermatrices in mutual, conditional, and joint entropy-based information indices. **Journal of Computational Chemistry**, Hoboken, v. 34, n. 4, p. 259-274, Feb. 2013.
- BRASIL, Ministério da Saúde. **Doenças infecciosas e parasitárias: Guia de Bolso**. 3 ed., Brasília: Ed. MS, Vol. II, 2004, 334p.
- BRASIL. Ministério da Saúde. **Guia de Vigilância em Saúde**. Brasília, 2014, 812 p. Disponível em:< <http://portalsaude.saude.gov.br/images/pdf/2014/novembro/27/guia-vigilancia-saude-linkado-27-11-14.pdf>> Acesso em: 15/11/2105
- CENTERS FOR DIASEASE CONTROL AND PREVENTION -CDC. Malaria. 2015. Disponível em: < <http://www.cdc.gov/malaria/about/biology/index.html> >. Acesso em: 15/11/2015.
- COWLEY, R. et al. The development of quinolone esters as novel antimalarial agents targeting the Plasmodium falciparum bc 1 protein complex. **MedChemComm**, Cambridge, v. 3, n. 1, p. 39-44, Sep.2011.
- DENNINGTON, R. D.; KEITH, T.A; MILLAM, J.M. **GaussView 5.0. 8**; Gaussian Inc., Wallingford, 2008. CD-ROM.
- DRAKELEY, C. et al. The epidemiology of Plasmodium falciparum gametocytes: weapons of mass dispersion. **Trends in parasitology**, Oxford,v. 22, n. 9, p. 424-430, Sep. 2006.
- GETHING, P. W. et al. A long neglected world malaria map: Plasmodium vivax endemicity in 2010. **PLoS Neglected Tropical Disease**, San Francisco, v. 6, n. 9, p. e1814, Sep. 2012.

GODDEN, J. W.; STAHURA, F. L.; BAJORATH, J. Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. **Journal of Chemical Information and Computer Sciences**, Washington, v. 40, n. 3, p. 796-800, Apr. 2000.

GOLBRAIKH, A.; TROPSHA, A. Beware of q^2 ! **Journal of Molecular Graphics and Modelling**, New York, v. 20, n. 4, p. 269-276, Jan. 2002.

LAPOUBLE, O. M. M.; SANTELLI, A. C. F.S.; MUNIZ-JUNQUEIRA, M. I. Epidemiological situation of malaria in the Brazilian Amazon region, 2003 to 2012. **Revista Panamericana de Salud Pública**, Washington, v. 38, n. 4, p. 300-306, Oct. 2015.

LOPES, F. et al. Amidomethylation of amodiaquine: antimalarial N-Mannich base derivatives. **Tetrahedron Letters**, Oxford, v. 45, n. 41, p. 7663-7666, Oct. 2004.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte, MG: Ed. UFMG, 2005. 295 p.

MITRA, I.; SAHA, A.; ROY, K. Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. **Molecular Simulation**, Oxon, v. 36, n. 13, p. 1067-1079, Oct. 2010.

MURRAY, C. J. et al. Global malaria mortality between 1980 and 2010: a systematic analysis. **The Lancet**, New York, v. 379, n. 9814, p. 413-431, Feb. 2012.

NUNES, C. A. et al. Chemoface: a novel free user-friendly interface for chemometrics. **Journal of the Brazilian Chemical Society**, Campinas, v. 23, n. 11, p. 2003-2010, Nov. 2012.

OJHA, P. K. et al. Further exploring r^2_m metrics for validation of QSPR models. **Chemometrics and Intelligent Laboratory Systems**, Berlin, v. 107, n. 1, p. 194-205, May 2011.

ORGANIZAÇÃO MUNDIAL DE SAÚDE- OMS. **Investing to overcome the global impact of neglected tropical diseases: third WHO report on neglected diseases**. Geneva: Organização Mundial de Saúde, 2015, 191p. Disponível em: <<http://apps.who.int/bookorders/anglais/detart1.jsp?codlan=>

1&codcol=15&codcch=3853>. Acesso em: 15/06/15

ORGANIZAÇÃO MUNDIAL DE SAÚDE-OMS. **World Malaria Report**. Gene Geneva: Organização Mundial de Saúde. 2013, 280 pg. Disponível em:<http://www.who.int/malaria/publications/world_malaria_report_2013/en/>. Acesso em 15/10/2015

PAN AMERICAN HEALTH ORGANIZATION- PAHO. **Malaria in the americas**. 2013. Disponível em: <http://www.who.int/malaria/publications/world_malaria_report_2013/en/> . Acesso 15/10/2015

RESSURREIÇÃO, A. S. et al. Structural Optimization of Quinolona-4 (1 H)-imines as Dual-Stage Antimalarials: Toward Increased Potency and Metabolic Stability. **Journal of Medicinal Chemistry**, Washington, v. 56, n. 19, p. 7679-7690, Oct. 2013.

RODRIGUES, T. et al. Quinolin-4(1H)-imines are Potent Antiplasmodial Drugs Targeting the Liver Stage of Malaria. **Journal of Medicinal Chemistry**, Washington, v. 56, n. 11, p. 4811-4815, Jun. 2013.

ROY, K. et al. Comparative studies on some metrics for external validation of QSPR models. **Journal of Chemical Information and Modeling**, Washington, v. 52, n. 2, p. 396-408, Dec. 2012.

TODESCHINI, R. et al. MobyDigs: software for regression and classification models by genetic algorithms. In: LEARDI, R. **Nature-Inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks**. Italy: Elsevier, 2003, Vol. 23, p. 141-167.

URIAS, R. W. P. et al. IMMAN: free software for information theory-based chemometric analysis. **Molecular Diversity**, Dordrecht, v. 19, n. 2, p. 305-319, May. 2015.

WALLER, N. G. et al. A comparison of the classification capabilities of the 1-dimensional kohonen neural network with two partitioning and three hierarchical cluster analysis algorithms. **Psychometrika**, New York, v. 63, n. 1, p. 5-22, Mar.1998.

WATERBEEND W. et al. **Chemometric methods in molecular design**. Cambridge: John Wiley & Sons , Vol. 2, Jul. 2008, 379p.

WOLD, S.; SJÖSTRÖM, M.I; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics and intelligent laboratory systems**, Berlin, v. 58, n. 2, p. 109-130, Oct. 2001.

YÉPEZ, M. C. et al. Factores asociados con el incumplimiento del tratamiento antipalúdico en pacientes ecuatorianos. **Revista Cubana de Medicina Tropical**, Ciudad de la Habana , v. 52, n. 2, p. 81-89, Aug. 2000.

ZHANG, Y. et al. Lead optimization of 3-carboxyl-4 (1 H)-quinolonaes to deliver orally bioavailable antimalarials. **Journal of Medicinal Chemistry**, Washington, v. 55, n. 9, p. 4205-4219, May 2012

CAPÍTULO 4: ANÁLISE COMPARATIVA PROTEOQUIMIOMÉTRICA MIA-QSAR DO SUBSTRATO DA PROTEASE NS3 DO VÍRUS DA DENGUE

RESUMO

A Análise Multivariada de Imagens aplicada à Relação Quantitativa Estrutura-Atividade (MIA-QSAR) demonstrou ser uma ferramenta promissora na modelagem de bioatividades de compostos químicos. Este capítulo fornece uma análise comparativa das abordagens MIA-QSARt, aug-MIA-QSAR e aug-MIA-QSARcolor na previsão da constante de Michaelis (k_m) e na taxa de clivagem do substrato (k_{cat}) de peptídeos modificados contra o vírus da dengue tipo 2 (DENV 2). O método aug-MIA-QSARcolor demonstrou poder preditivo superior em comparação com os modelos MIA-QSARt e aug-MIA-QSAR, sugerindo que a incorporação de esquemas de cores para discriminar os diferentes tipos de átomos da estrutura química e a modificação do tamanho atômico para ser proporcional ao raio van der Waals melhora a capacidade preditiva do método MIA-QSAR. A interpretação estrutural dos modelos aug-MIA-QSARcolor revela que aminoácidos pequenos não ramificados nas posições A1 e A4 estão associados com maior afinidade de ligação do substrato e a uma maior taxa de clivagem do substrato (k_{cat}) quando presentes nas posições A1 e A2, enquanto que os aminoácidos ácidos nas posições A2 e A4 estão relacionados com baixa afinidade do substrato e baixa taxa de clivagem do substrato quando na posição A1. Esses resultados fornecem informações químicas pertinentes sobre as características estruturais críticas no planejamento de novos compostos-alvo contra a DENV.

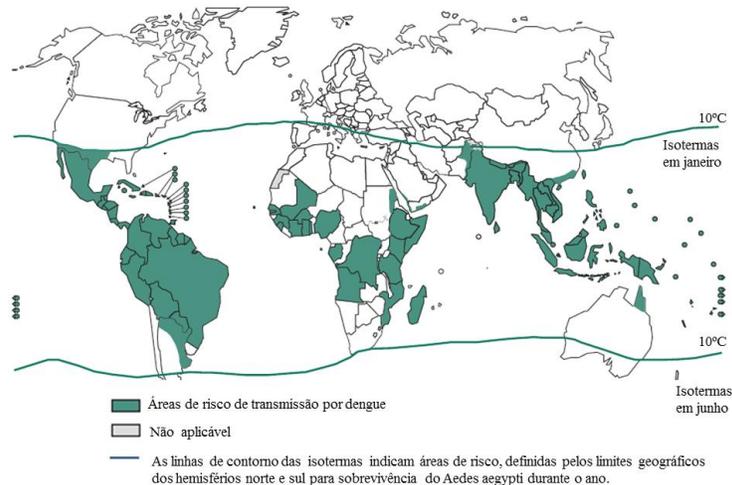
Palavras-chave: aug-MIA-QSARcolor. Protease NS3. Regressão linear múltipla. PLS.

1 INTRODUÇÃO

1.1 Aspectos gerais da dengue

A dengue é uma doença febril aguda de evolução geralmente benigna que afeta as regiões tropicais e subtropicais em todo o mundo (Figura 15). Estimativas recentes indicam que 3,9 bilhões de pessoas em 128 países estão em risco de infecções por dengue (BRADY et al., 2012). Em 2013, foi reportado que 390 milhões de infecções por dengue tenham ocorrido no ano de 2010 (BHATT et al., 2013). A expansão geográfica e o aumento do número de casos tornaram a doença um sério problema de saúde pública em todo o mundo (GUZMAN e HARRIS, 2015; OMS, 2015a).

Figura 15: Áreas de risco de transmissão da dengue no mundo em 2014.



Fonte: OMS (2015b).

A dengue é considerada uma doença reemergente em vários lugares do mundo, tais como Paquistão, Arábia Saudita, Iêmen, Djibouti, Somália, Sudão e

em países da América (OMS,2015b). A transmissão do vírus da dengue foi interrompida em grande parte das regiões das Américas nas décadas de 1950 e 1960, mas retornou na década de 1970 por falhas na vigilância epidemiológica e pelas mudanças sociais e ambientais propiciadas pela urbanização acelerada dessa época (BRASIL, 2010). As regiões do Caribe, América Central e do Sul agora estão em um estado hiperendêmico, com transmissão autóctone em quase todos os países (OMS, 2015b). No Brasil, o ano de 2015 registrou 1.649.008. Sendo confirmados 863 óbitos por dengue, o que representa um aumento de 82,5% em comparação com o mesmo período de 2014, quando foram confirmados 473 óbitos. A região sudeste concentra 65,2% dos óbitos do país e o estado de São Paulo o maior número de registros (BRASIL, 2016).

A dengue tem uma ampla variedade de manifestações clínicas graves e não graves, que vão desde ausência de sintomas e febres simples ao quadro de hemorragias e a síndrome do choque da dengue (GUBLER, 1998). A dengue clássica (CD) inicia-se com febre alta, seguida de cefaleia, prostração, náuseas, anorexia, astenia, vômitos, dor abdominal generalizada (principalmente em crianças), etc. Esses sintomas duram de 5 a 7 dias, mas a debilidade física pode permanecer por várias semanas. Na febre hemorrágica da dengue (FHD) e na síndrome de choque hemorrágico (SCD) os sintomas iniciais são os mesmos da dengue clássica. Porém, entre o 3º e 4º dia, o quadro se agrava com hemorragias espontâneas, diminuição da pressão arterial, derrames cavitários, cianose e diminuição brusca da temperatura. A principal característica fisiopatológica associada ao grau de severidade da FHD é o extravasamento do plasma, que se manifesta por meio de valores crescentes do hematócrito e hemoconcentração, situação que pode levar ao SCD. O choque é decorrente do aumento de permeabilidade vascular, seguida de hemoconcentração e falência circulatória, de duração curta pode levar ao óbito em 12 a 24 horas (BRASIL, 2010).

Até o presente momento, o tratamento para conter as infecções causadas pelo vírus da dengue (DENV) tem sido basicamente através de medidas paliativas que controlam os sintomas da doença (KALAYANAROOJ, 2011). Assim, a prevenção e redução das infecções por dengue têm dependido inteiramente da interrupção do contato humano-vetor. Esses métodos não têm se mostrado satisfatórios, em razão da dependência de questões sociais, tais como saneamento básico, participação coletiva das comunidades e de políticas públicas de saúde para conter os criadouros do mosquito ou promover o uso de repelentes não específicos.

A transmissão da dengue ocorre, principalmente, pela picada do mosquito fêmea da espécie *Aedes aegypti* e, em menor extensão, por *Aedes albopictus* contaminados com o Arbovírus pertencente à família *Flaviviridae* e gênero *Flavivirus*. O processo de transmissão inicia-se quando a fêmea do mosquito passa a se alimentar de sangue para a maturação de seus ovos. Nessa fase, a fêmea pode ser contaminada pelo vírus da dengue a partir de hospedeiros vertebrados. Após a infecção, o vírus fica encubado por um período de 8 a 12 dias. Em seguida, multiplica-se no intestino médio do vetor e infecta outros tecidos, chegando às glândulas salivares e aos ovos presentes no seu aparelho reprodutor. O mosquito vive, em média, 45 dias e, nesse período, pode contaminar até 300 pessoas. Além do vírus da dengue, o mosquito pode transmitir outras arboviroses, tais como o chikungunya e o zica vírus (OMS, 2015b). A diversidade de reservatórios de água para deposição dos ovos pela fêmea do mosquito, principalmente em áreas urbanas, tem contribuído substancialmente para o aumento do número de vetores e, conseqüentemente, o número de infecções transmitidas pelo mosquito.

O vírus da dengue compreende uma partícula esférica com diâmetro de 40-50 nm, com um envelope de lipopolissacarídeo. O genoma é constituído por uma fita de RNA, com aproximadamente 11.000 nucleotídeos, formando uma

única grelha de leitura aberta que codifica uma única poliproteína precursora 5'-CPRM(M)-E-NS1-NS2A-NS2B-NS3-NS4A-NS4B-NS5-3'. Essa poliproteína viral é clivada por proteases em 10 produtos gênicos: três proteínas estruturais (capsídeo, membrana precursora e envelope), essenciais para a entrada, montagem e a maturação viral, e sete proteínas não estruturais (NS1, NS2A, NS2B, NS3, NS4A, NS4B, NS5), responsáveis por realizar a replicação viral (MASON et al., 1987; LANCIOTTI et al., 1994).

Atualmente, são conhecidos quatro sorotipos do vírus da dengue: DENV1, DENV2, DENV3 e DENV4, os quais estão distribuídos em todas as regiões do mundo onde a doença é endêmica (DINU et al., 2015). Os quatro sorotipos são caracterizados por diferentes propriedades antigênicas (MASON et al., 1987) que causam a dengue, a febre hemorrágica da dengue e a síndrome do choque da dengue (LANCIOTTI et al., 1994). A infecção por um deles dá proteção permanente para o mesmo sorotipo, mas imunidade parcial e temporária contra os outros (GUSMAN; HARRIS, 2015). Essas características virais têm dificultado o desenvolvimento de medicamentos eficazes e vacinas para combater a doença. Além disso, a falta de um modelo animal para testar a eficácia da vacina tem dificultado as avaliações experimentais *in vivo*, uma vez que ratos e macacos não apresentam patologia semelhante aos seres humanos, quando infectados com o vírus (GIBBONS; VAUGHN, 2002).

1.2 Pesquisas atuais para controle e tratamento da dengue

As pesquisas para busca de um tratamento para a dengue envolvem estudos que buscam o controle do mosquito vetor, o desenvolvimento de vacinas e a produção de medicamentos antivirais. Em relação ao controle do vetor, estudos têm seguido com estratégias que visam reduzir a população do mosquito *A. aegypti*, tais como a utilização de peixes predadores dos ovos da fêmea do

mosquito depositados em grandes recipientes de água (PAMPLONA et al., 2009) e a infecção da população de mosquitos com bactérias do gênero *wolbachia*, que faz com que os mosquitos se tornem parcialmente resistentes ao vírus da dengue (MCMENIMAN et al., 2009). Há também ensaios com indivíduos machos do *A. aegypti* geneticamente modificados que, após cruzarem com as fêmeas, transmitem um gene letal (VALDEZ et al., 2011).

No contexto do desenvolvimento de vacinas e produção de antivirais, a compreensão dos alvos moleculares essenciais dentro do ciclo biológico do vírus tem sido fundamental para o desenvolvimento de tais terapias. Dentre os alvos considerados estratégicos para o desenvolvimento de fármacos, destacam-se a proteína de membrana (prM), proteína de envelope (E) (KAMPMANN et al., 2009; ZHANG et al., 2004) e a proteína do capsídeo (C), NS5, NS3 e, principalmente, NS2B-NS3pro, uma vez que estão associadas a processos essenciais do DENV, tais como o processo de infecção (entrada viral), maturação e replicação viral (GODÓI et al., 2014).

A proteína multifuncional NS3 é um importante alvo molecular para o desenvolvimento de terapias específicas para a prevenção da dengue (LESCAR et al., 2008). Caracterizada por ser uma proteína conservada para os quatro sorotipos da dengue, a NS3 é composta por 618 aminoácidos e desempenha papel na clivagem da poliproteína viral e é essencial para o processamento de replicação e maturação viral (FALGOUT et al., 1991; LUO et al., 2008). No entanto, as NS3 requer a proteína proteolítica NS2B como cofator para a sua ativação e estabilidade estrutural (LESCAR et al., 2008). A proteína NS3 contém dois domínios que possuem várias atividades enzimáticas: um domínio serina protease localizado dentro dos resíduos de aminoácidos N-terminais (BAZAN; FLETTERICK, 1989; FALGOUT et al., 1991) e um RNA helicase e NTPase na região C-terminal (ASSENBERG et al., 2009).

O conhecimento sobre um alvo molecular tem gerado avanços expressivos no entendimento de um processo fisiopatológico e, conseqüentemente, no desenvolvimento de moléculas bioativas mais específicas (GUIDO; ANDRICOPULO; OLIVA, 2010). A caracterização de alvos moleculares tem se beneficiado de técnicas de cristalografia e ressonância magnética nuclear para determinação de suas estruturas tridimensionais, as quais fornecem informações valiosas sobre suas características estruturais e moleculares, tal como sítios de ligação. Por vezes, técnicas de modelagem molecular e *screening* virtual são capazes de simular interações entre alvos moleculares e estruturas bioativas, ou simplesmente prever estruturas químicas mais adequadas a tais alvos moleculares.

Nesse contexto, este capítulo apresenta uma análise comparativa da estratégia de análise multivariada de imagens aplicada à relação quantitativa estrutura-atividade (MIA-QSAR) na predição da atividade de substratos peptídicos com atividade contra o DENV, utilizando os métodos de regressão por mínimos quadrados parciais (PLS) e regressão linear múltipla (MLR). Além disso, uma interpretação do melhor modelo construído é realizada com o objetivo de obter maior conhecimento sobre as características estruturais que favoreçam a atividade dos peptídeos contra o DENV, contribuindo, assim, para o planejamento de novos compostos eficazes contra essa infecção.

2 METODOLOGIA

2.1 Construção das matrizes de dados para modelagem MIA-QSAR

Um conjunto de dados de 54 peptídeos modificados, conhecidos como substratos da protease NS3 do DENV tipo 2, cuja constante de Michaelis (K_m) e o parâmetro cinético para a constante de taxa de clivagem do substrato (K_{cat})

foram determinados experimentalmente, foi obtido da literatura (PRUSIS *et al.*, 2008; SILLA *et al.*, 2011) (Tabela 8).

Tabela 8. Sequência A1'-A4' dos peptídeos modificados com estrutura geral Abz-RRRRXXXX-NY-NH₂^a e valores experimentais obtidos para a constante de Michaelis (K_m) e taxa de clivagem (k_{cat}) para a protease NS3 do DENV2. (Continua)

Cp	Peptídeos	K _m ^a	k _{cat} ^b	Cp	Peptídeos	K _m ^a	k _{cat} ^b
1*	APCN ^c	0,73	-0,62	28	STPH	1,24	-0,03
2*	HHGN	0,64	-1,1	29	GLGF	0,64	-0,05
3	DDGN	1,9	-1,15	30	DTSF	1,18	-1,22
4	SNSN	0,8	0,23	31	HPTF	1,09	-1,1
5	NWTN	0,78	-0,52	32	ADTF	1,43	-0,68
6	GTVN	0,81	-0,48	33	NSVF	1,04	-0,54
7	AGPN	0,62	-0,39	34	SAPF	0,73	0,3
8	SHCD	1,06	-0,13	35*	ATCL	0,86	-0,66
9	APGD	1,29	-1,4	36	NGGL	0,86	-1
10	HWSD	1,17	-1,15	37	HDSL	1,27	-0,66
11	GGTD	1,11	-0,37	38*	GSSL	1,02	-0,32
12	AAVD	1,26	-0,12	39	SLTL	0,7	0,25
13	NDVD	1,64	-1	40	DHTL	1,46	-1,15
14	NLPD	1,22	-0,66	41	DNPL	1,24	-1,15
15	HNCC	0,94	-0,85	42	HTG-	0,82	-1,1
16	NACC	0,51	-0,55	43	DLS-	1,29	-0,85
17	SWGc	0,46	-0,41	44	NHT-	0,78	-0,77
18	AHSC	0,75	-0,59	45*	AWV-	0,35	-0,57
19*	DSTC	1,11	-1,15	46	GPP-	0,71	-0,77
20	SGVC	0,63	-0,01	47	SAGM	0,54	0,27
21	GDPC	1,12	-0,74	48*	SWPL	0,74	-0,2

Tabela 12 (Conclusão)

Cp	Peptídeos	K_m^a	k_{cat}^b	Cp	Peptídeos	K_m^a	k_{cat}^b
22*	DGCH	0,85	-1,05	49*	AGVL	0,34	-0,47
23	ANGHH	0,85	-1,05	50	GTGN	0,48	-0,54
24*	NPSH	0,71	-1,1	51*	SAG-	0,56	0,4
25	GATH	0,67	-0,03	52	AAGM	0,39	-0,26
26	HLVH	1,15	-0,85	53	SAAM	0,52	0,19
27	HSPH	0,76	-1,15	54	SAGA	0,49	0,21

^aAbz: ácido o-aminobenzoico, nY: 3-nitrotyrosina, ^{a,b}expresso em escala logarítmica, valores originais de k_m e k_{cat} em μM e min^{-1} , respectivamente. ^cpeptídeos representados pela letra inicial de cada aminoácido; *Grupo teste. Fonte: Adaptado de Prusis et al. (2008).

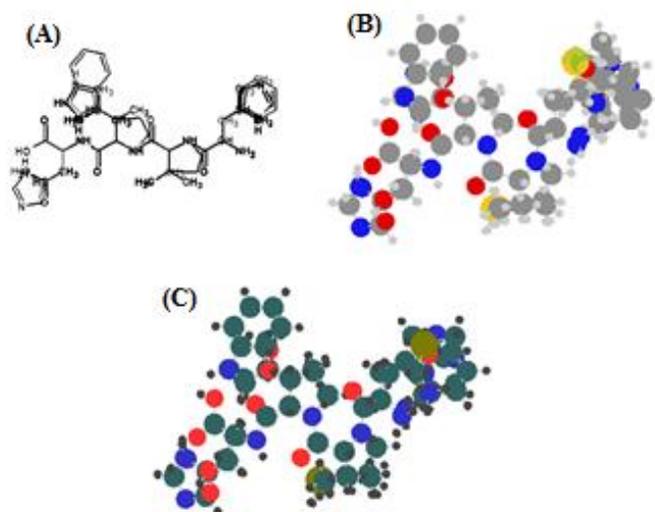
As imagens em preto e branco dos peptídeos modificados (ou seja, para a abordagem MIA-QSARt) foram preparadas usando o programa ChemSketch (ACD, 2009) e as imagens coloridas das abordagens aug-MIA-QSAR e aug-MIA-QSARcolor foram obtidas usando o programa GaussView (Version 5.0.8) (DENNINGTON, KEITH e MILLAN, 2008). É importante ressaltar que, para a construção das imagens segundo as abordagens aug-MIA-QSAR e aug-MIA-QSARcolor, foram utilizadas cores sólidas, sem efeito de iluminação dos átomos (*spotlight*).

Após desenho, cada estrutura química foi transferida para um espaço de trabalho no aplicativo *Paint* do *Microsoft Windows* de dimensão pré-definida (ou seja, 500×300 pixels para a abordagem MIA-QSARt e 420×300 pixels para a aug-MIA-QSAR e aug-MIA-QSARcolor) e cada imagem foi salva como um arquivo de *bitmaps* e alinhadas. O processo para obtenção das estruturas e o alinhamentos das moléculas foi realizado conforme descrito no capítulo 2. A Figura 16 ilustra o conjunto de dados alinhado para as três abordagens.

Após alinhamento, as imagens foram transferidas para o software Chemoface (NUNES et al., 2012) e desdobradas em matrizes numéricas com dimensões de 54x 150000 para MIA-QSARt e 54x126000 para aug-MIA-QSAR

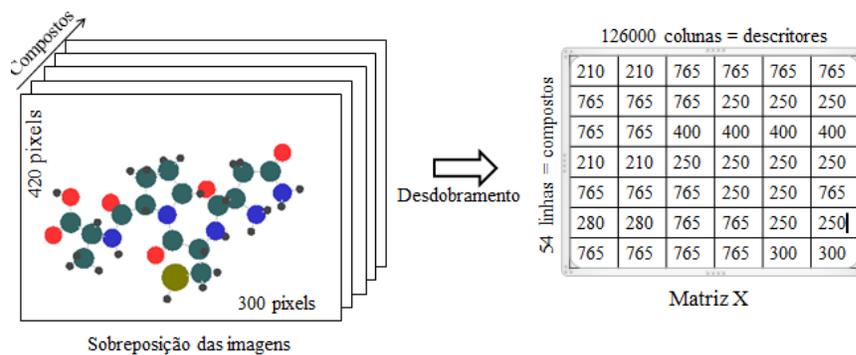
e aug-MIA-QSARcolor. A Figura 17 ilustra o alinhamento e o desdobramento das imagens em matrizes numéricas.

Figura 16 - Sobreposição das imagens dos peptídeos para as abordagens MIA-QSARt (A), aug-MIA-QSAR (B) e aug-MIA-QSARcolor (C).



Fonte: Do autor (2016).

Figura 17: Construção do arranjo tridimensional e matriz X obtida após desdobramento das imagens aug-MIA-QSARcolor do conjunto de peptídeos.



Fonte: Do autor (2016).

Posteriormente, as colunas da matriz X que apresentavam variância 0 (espaços em branco comuns para todas as imagens em branco comuns para todas as imagens e partes congruentes das estruturas químicas) foram removidas, restando matrizes de menor dimensionalidade 54×4723 para MIA-QSAR e 54×22756 para aug-MIA-QSAR e aug-MIA-QSARcolor.

2.2 Métodos quimiométricos para modelagem e validação

Dado o grande número de descritores gerados pelas abordagens MIA-QSAR, a técnica PLS foi empregada como método de construção do modelo primário. Para esta análise de regressão, o conjunto de dados de peptídeos modificados foi dividido em séries de treinamento e teste, contendo 43 (80%) e 11 (20%) compostos, respectivamente, tal como no estudo anterior (SILLA *et al.*, 2011).

A qualidade dos modelos PLS foi avaliada utilizando os parâmetros: coeficiente de determinação para a calibração (r^2) e validação cruzada *leave-one-out* (r^2_{cv}), bem como a raiz do erro quadrado médio de validação cruzada (RMSECV). Além disso, os modelos foram vigorosamente validados usando a validação externa (r^2_{pred}) e Y-randomização (avaliado estatisticamente por RMSErand e r^2_{rand}), os quais são procedimentos utilizados para atestar o poder preditivo dos modelos e a ausência de correlação ao acaso, respectivamente. Dada à necessidade da realização de procedimentos de validação ainda mais vigorosos na modelagem QSAR, parâmetros adicionais como r^2_m e r^2_p foram utilizados para tornar os resultados de modelagem ainda mais confiáveis (CHIRICO; GRAMATICA, 2012). O coeficiente de determinação modificado (r^2_m) foi utilizado para detectar a proximidade entre a atividade observada e esperada (OJHA *et al.*, 2011; ROY *et al.*, 2013), e o coeficiente de determinação

penalizado e corrigido (r^2_p), utilizado para avaliar a diferença entre r^2 e r^2_{rand} (MITRA; SAHA; ROY, 2010).

Embora PLS seja uma técnica preferida quando se lida com matrizes de dados com numerosas variáveis, o fato de que as variáveis originais são convertidas em projeções ortogonais torna a interpretação da propriedade modelada complexa, uma vez que as variáveis originais ficam "mascaradas" nas variáveis latentes. Portanto, a fim de buscar uma interpretação mais simples para as abordagens MIA-QSAR, também foi utilizada a técnica MLR. Para lidar com a elevada dimensionalidade das matrizes de dados MIA-QSAR foi aplicada uma técnica de pré-filtragem com base em entropia de Shannon (SE), onde as variáveis com menos de 15% da entropia máxima ($SE_{MAX} = \text{Log}54$) foram eliminadas (URIAS *et al.*, 2015). As matrizes ainda foram filtradas com base no coeficiente de correlação (x/x), onde, para variáveis com coeficientes de correlação (x/x) $\geq 0,98$, uma variável foi retida. Posteriormente, modelos baseados em MLR foram construídos usando os conjuntos filtrados de variáveis e os modelos com os maiores valores de r^2 e r^2_{cv} para cada abordagem MIA-QSAR foram retidos, segundo os critérios de otimização e busca da técnica de algoritmo genético (GOLDBERG; HOLLAND, 1988). Os modelos obtidos foram, posteriormente, validados utilizando as técnicas de validação externa e Y-randomização.

3 RESULTADOS E DISCUSSÕES

3.1 Análise PLS

Os modelos PLS para o parâmetro cinético K_m foram construídos usando 3 variáveis latentes (LVs) para a abordagem MIA-QSARt e 2 LVs para as abordagens aug-MIA-QSAR e aug-MIA-QSARcolor, enquanto que para o

parâmetro cinético k_{cat} foram consideradas 6 LVs para a abordagem MIA-QSARt e 5 LVs para as abordagens aug-MIA-QSAR e aug-MIA-QSARcolor. O número de LVs foi determinado com base na redução do erro quadrático médio da validação cruzada *leave-one-out* (RMSECV). A Tabela 9 mostra os parâmetros estatísticos para os modelos PLS obtidos de acordo com as abordagens MIA-QSARt, aug-MIA-QSAR e aug-MIA-QSARcolor.

Tabela 9 - Estatísticas para os modelos MIA-QSAR/PLS obtidos para os parâmetros K_m e k_{cat} dos substratos peptídicos inibidores da protease NS3 do DENV tipo 2.

Parâmetros	MIA-QSARt		aug-MIA-QSAR		aug-MIA-QSARcolor	
	k_{cat}	K_m	k_{cat}	K_m	k_{cat}	K_m
LV	6	3	5	2	5	2
r^2_{cal}	0,97	0,66	0,89	0,44	0,90	0,44
r^2_{cv}	0,52	0,09	0,58	0,14	0,51	0,12
RMSEcv	0,08	0,21	0,16	0,22	0,15	0,35
r^2_{rand}	0,90	0,62	0,62	0,30	0,66	0,37
RMSErand	0,15	0,23	0,30	0,30	0,28	0,29
$^c r^2_p$	0,26	0,16	0,49	0,25	0,46	0,17
r^2_{pred}	0,78	0,05	0,70	0,14	0,59	0,01
RMSEp	0,23	0,31	0,25	0,3	0,35	0,33
r^2_m	0,67	0,00	0,58	0,00	0,57	0,00

De um modo geral, os modelos construídos mostraram boa capacidade de previsão para a taxa de clivagem do substrato (k_{cat}), com valores de $r^2 > 0,7$ e r^2_{cv} , r^2_{pred} e $r^2_m > 0,5$ (Tabela 9). Na comparação das três abordagens, o método MIA-QSARt mostrou melhores resultados para calibração e validação externa [$r^2 = 0,97$, $r^2_{pred} = 0,78$ ($r^2_m = 0,67$)] quando comparado às abordagens aug-MIA-QSAR [$r^2 = 0,89$, $r^2_{pred} = 0,70$ ($r^2_m = 0,58$)] e aug-MIA-QSARcolor [$r^2 = 0,90$, $r^2_{pred} = 0,59$ ($r^2_m = 0,57$)]. Entretanto, o alto valor de r^2_{rand} para o MIA-QSARt (0,90) comparado com as abordagens aug-MIA-QSARcolor (0,66)

e aug-MIA-QSAR (0,62), sugere que o modelo MIA-QSARt é mais propenso à correlação ao acaso e pode, assim, ser considerado não confiável. Isso pode ser avaliado pelo valor de r^2_p , que foi inferior para o MIA-QSARt (0,26) se comparado às abordagens aug-MIA-QSAR (0,49) e aug-MIA-QSARcolor (0,46), embora nenhuma abordagem tenha atingido o nível recomendado, isto é $r^2_p \geq 0,50$ (MITRA, SAHA e ROY, 2010). Nesse sentido, o modelo baseado na abordagem aug-MIA-QSAR pode ser considerado mais confiável para previsão do que as abordagens aug-MIA-QSARcolor e MIA-QSARt, pois mostrou melhor desempenho nos ensaios de aleatorização e bons resultados de validação externa.

Para os modelos MIA-QSAR/PLS para a constante de Michaelis (K_m) dos substratos peptídicos para a protease NS3(Tabela 9), observa-se que as abordagens MIA-QSARt, aug-MIA-QSAR e aug-MIA-QSARcolor não produziram resultados satisfatórios, uma vez que os valores para r^2_{cv} , r^2_{pred} , r^2_p e r^2_m foram menores do que 0,5 e o r^2 menor que 0,7 (MITRA, SAHA, ROY, 2010; OJHA et al., 2011). Esse resultado sugere que a técnica PLS não é adequada para a modelagem do parâmetro K_m para o conjunto de dados dos peptídeos modificados considerados, ou que os descritores MIA não são capazes de capturar a tendência nos valores de K_m , ou que os dados de K_m experimentais não estão corretos/precisos. Nesse sentido, também cabe mencionar que foi observado baixo desempenho para modelagem de K_m para essa classe de compostos em estudos anteriores na literatura (PRUSIS et al., 2008; SILLA et al., 2011). Adiante, por meio de comparação com os resultados MLR, verificar-se-á que a primeira hipótese (inabilidade do PLS em modelar K_m) é a mais plausível.

3.2 Regressão Linear Múltipla

Seis modelos MLR foram construídos para os parâmetros K_m e k_{cat} . A Tabela 10 mostra as equações (1-6) para os modelos construídos, bem como os parâmetros estatísticos correspondentes, usando as abordagens MIA-QSARt, aug-MIA-QSAR e aug-MIA-QSARcolor.

Como pode ser observado na Tabela 10, os modelos para os parâmetros K_m e k_{cat} apresentam bons parâmetros estatísticos. Na comparação do desempenho dos modelos K_m para as diferentes abordagens MIA-QSAR (verificar Equações 1, 2 e 3), observa-se que, apesar de os parâmetros de calibração e de validação cruzada serem superiores para a abordagem MIA-QSARt [r^2 (0,77), r^2_{cv} (0,68), ${}^c r^2_p$ (0,67)] perante as abordagens aug-MIA-QSAR [r^2 (0,64), r^2_{cv} (0,49), ${}^c r^2_p$ (0,58)] e aug-MIA-QSARcolor [r^2 (0,71), r^2_{cv} (0,60), ${}^c r^2_p$ (0,64)], o modelo baseado em aug-MIA-QSARcolor possui maior robustez e capacidade de previsão, uma vez que produz os maiores valores para os parâmetros de validação externa [isto é, r^2_{pred} (0,64) e r^2_m (0,55)] comparados com as abordagens MIA-QSARt [r^2_{pred} (0,24), r^2_m (0,00)] e aug-MIA-QSAR [r^2_{pred} (0,58) e r^2_m (0,53)].

Quanto aos modelos para previsão de k_{cat} (Equações 4, 5 e 6), observa-se que, enquanto um desempenho comparável é observado para as três abordagens quanto aos procedimentos de calibração e validação interna, isto é, MIA-QSARt [r^2_{cal} (0,90), r^2_{cv} (0,87)], aug-MIA-QSAR [r^2_{cal} (0,87), r^2_{cv} (0,82)] e aug-MIA-QSARcolor [r^2_{cal} (0,88), r^2_{cv} (0,83)], um desempenho superior é observado para a abordagem aug-MIA-QSARcolor para o procedimento de validação externa [r^2_{pred} (0,92), r^2_m (0,80)] relativo às abordagens MIA-QSARt [r^2_{pred} (0,66), r^2_m (0,66)] e aug-MIA-QSAR [r^2_{pred} (0,76), r^2_m (0,60)].

O poder preditivo superior para a abordagem aug-MIA-QSPRcolor sugere que a incorporação de informações químicas mais significativas para as

Tabela 10: Parâmetros estatísticos e equações dos modelos MLR para K_m e k_{cat} , para os 54 peptídeos modificados.

Métodos	N [†]	r^2_{cal}	r^2_{cv}	RMSE _{cv}	r^2_{rand}	r^2_p	r^2_{pred}	r^2_m	Modelos	Eq
Constante de Michaelis (K_m)										
MIA-QSARt	54	0,77	0,68	0,18	0,18	0,67	0,24	0,00	$K_m = 1,3702 - 0,0005X482 - 0,0004X874 - 0,0004X19020 - 0,0004X2243 - 0,0005X3869 - 0,0004X4451$	(1)
aug-MIA-QSAR	54	0,64	0,49	0,26	0,12	0,58	0,58	0,53	$K_m = 2,1299 - 0,0021X1377 + 0,0007X1704 - 0,0008X4256 - 0,017X6809 + 0,0025X18185 - 0,0008X19371$	(2)
aug-MIA-QSARcolor	54	0,71	0,60	0,21	0,13	0,64	0,64	0,55	$K_m = 0,7346 + 0,0013X1377 - 0,0016X1704 + 0,0010X4256 - 0,005X6809 + 0,0006X18185 - 0,0010X19371$	(3)
Constante de Clivagem (k_{cat})										
MIA-QSARt	54	0,90	0,87	0,18	0,13	0,83	0,66	0,66	$k_{cat} = 0,3369 - 0,0015X482 + 0,0008X874 - 0,0004X19020 + 0,0006X224 - 0,0002X3869 - 0,0005X4451$	(4)
aug-MIA-QSAR	54	0,87	0,82	0,21	0,11	0,81	0,76	0,60	$k_{cat} = -1,2412 + 0,0015X803 - 0,0011X1843 + 0,0011X2988 - 0,0020X7062 + 0,0023X7655 - 0,0006X11236$	(5)
aug-MIA-QSARcolor	54	0,88	0,81	0,21	0,11	0,82	0,92	0,80	$k_{cat} = -1,7105 + 0,0009X309 + 0,0013X2090 - 0,012X2126 + 0,0005X2498 + 0,0006X3892 - 0,0008X7320$	(6)

[†]N_{treinamento} = 43 and N_{teste} = 11

as imagens, especialmente a modificação das dimensões atômicas para alcançar a proporcionalidade ao raio de van der Waals e o uso de esquemas de cores cuidadosamente definidos para correlacionar com a eletronegatividade atômica, melhora o desempenho do método MIA-QSAR. Adicionalmente, além de contribuir para a melhoria do poder de previsão, os esquemas de cores abre caminho para a interpretação química dos modelos construídos, uma vez que os átomos estão identificados pelos seus valores de pixel nas variáveis. Nesse sentido, foi realizada a interpretação das informações químicas codificadas pela abordagem aug-MIA-QSARcolor para os parâmetros K_m e k_{cat} , que consequentemente contribui para o desenvolvimento de substratos mais promissores com uma maior afinidade pela proteína NS3 do DENV-02.

3.3 Interpretação estrutural do modelo aug-MIA-QSARcolor

Uma análise das variáveis contidas no modelo aug-MIA-QSARcolor para o parâmetro K_m (Tabela 11) revela que as variáveis **X1377** e **X1704** correspondem a coordenadas de pixel para as cadeias laterais A1 de aminoácidos, as variáveis **X4256** e **X6809** codificam informação sobre as cadeias de aminoácidos A2 e a variáveis **X18185** e **X19371** em grupos R no aminoácido A4 (note que cada variável corresponde a um único pixel de coordenadas na imagem multivariada).

Tabela 11 - Variáveis contidas na abordagem aug-MIA-QSARcolor para o parâmetro K_m , com compostos ordenados do menor para o maior valor de K_m

Cp	X1377	X1704	X4256	X6809	X18185	X19371	Km
49	765	765	765	765	765	765	0.34
45	765	765	250	250	765	765	0.35
52	765	765	765	765	765	765	0.39
17	765	612	250	250	765	765	0.46

Tabela 11 (Continua)

Cp	X1377	X1704	X4256	X6809	X18185	X19371	Km
50	765	765	765	210	765	765	0.48
54	765	765	765	765	765	765	0.49
16	300	250	765	765	765	765	0.51
53	765	765	765	765	765	765	0.52
47	765	765	765	765	765	765	0.54
51	765	765	765	765	765	765	0.56
7	765	765	765	765	210	210	0.62
20	765	765	765	765	765	765	0.63
2	300	250	612	612	210	210	0.64
29	765	765	765	612	765	612	0.64
25	765	765	765	765	612	765	0.67
39	765	765	765	612	765	765	0.7
24	300	250	765	765	612	765	0.71
46	765	765	765	765	765	765	0.71
1	765	765	765	765	210	210	0.73
34	765	612	765	765	765	612	0.73
48	765	612	250	250	765	765	0.74
18	765	765	765	612	765	765	0.75
27	300	250	765	210	612	765	0.76
5	300	250	250	250	210	210	0.78
44	300	250	612	612	765	765	0.78
4	765	765	765	350	210	210	0.8
6	765	765	765	210	765	765	0.81
22	300	250	765	765	765	765	0.85
23	765	765	765	350	612	765	0.85
35	765	765	765	210	765	765	0.86
36	300	250	765	765	765	765	0.86
15	300	250	765	350	765	765	0.94
38	765	765	765	210	765	765	1.02
33	300	250	765	210	765	612	1.04
8	765	765	612	612	765	210	1.06
31	300	250	765	765	765	612	1.09
11	765	765	765	765	765	210	1.11
19	765	250	765	765	765	765	1.11
21	765	765	765	350	765	765	1.12
26	300	250	765	612	612	765	1.15

Tabela 11 (Conclusão)

Cp	X1377	X1704	X4256	X6809	X18185	X19371	K_m
10	300	250	250	250	765	210	1.17
30	350	250	765	210	765	612	1.18
14	300	250	765	612	210	210	1.22
28	765	612	765	210	612	765	1.24
41	350	250	765	350	765	765	1.24
12	765	765	765	765	765	210	1.26
37	300	250	765	350	765	765	1.27
9	765	765	765	765	765	210	1.29
43	765	250	765	612	765	765	1.29
32	765	765	765	350	765	612	1.43
40	765	250	612	612	765	765	1.46
13	300	250	765	350	765	210	1.64
42	300	250	765	210	765	765	1.82
3	765	250	765	350	210	210	1.9

A ausência de variáveis correspondentes para as coordenadas de pixel para as cadeias laterais A3 sugere pouca contribuição de aminoácidos nesta posição para o parâmetro cinético K_m , corroborando os resultados obtidos num estudo anterior realizado por Prusis *et al.* (2008).

Uma análise intuitiva revela que espaço em branco (valores de pixel 765) para as variáveis **X1377** e **X1704** são indicativos de aminoácidos de grupos R pequenos na posição A1 (como exemplo, alanina, serina e glicina), os quais, quando concomitantes com grupos R de tamanho semelhante em A4 [identificados por espaços em branco (765) para os substituintes **X18185** e **X19371**, (por exemplo, formando parte de glicina, alanina, leucina e metionina)], geram baixos valores de K_m (isto é, elevada atividade). Em contraste, quando pequenos grupos R em A1 aparecem concomitantes com um grupo carboxila [identificados por pixels 765 (espaço em branco) e 210 (H) para **X18185** e **X19371**, respectivamente, ou seja, hidrogênio que faz parte do grupo carboxila] contendo cadeias laterais em A4, um efeito oposto é observado (ou seja, está relacionado com altos valores de K_m). Ainda mais interessante, independentemente dos grupos R em A1 e A4, quando a cadeia

lateral A2 compreende um grupo carboxila [identificado por pixels 765 (espaço em branco) e 350 (O) para variáveis ou posições de pixel **X4256** e **X6809**, respectivamente], os valores de K_m aumentam.

Note-se que os pixels nas posições **X4256** e **X6809** (ou **X18185** e **X19371**) ditam que o grupo carboxila é separado aproximadamente por um carbono de distância da cadeia do peptídeo, como é no ácido aspártico. Esse resultado é consistente com os relatados por Prusis e colaboradores (2008), em que a baixa afinidade para o substrato foi atribuída à presença de unidades de ácido aspártico nos substratos peptídicos. Por outro lado, um grupo R do sistema bicíclico para o aminoácido A2 [identificado por átomos de carbono (250) na posição de pixel **X4256** e **X6809**] está relacionado com baixos valores de K_m (elevada afinidade para o substrato), com um efeito sinérgico das cadeias laterais A1 e A4. Essa é uma descoberta interessante, de acordo com um estudo anterior no qual o triptofano (derivado de indol) foi citado como o aminoácido mais favorável para esta posição devido à alta afinidade pelo substrato (PRUSIS et al., 2008).

Uma avaliação cuidadosa das variáveis que constituem o modelo MLR para a taxa de clivagem do substrato (Tabela 12) indica que as variáveis **X309**, **X2090**, **X2126** e **X2498** correspondem a posições de pixel para o grupo R dos aminoácidos em A1, enquanto que as variáveis **X6660** e **X7320** codificam informações sobre as cadeias laterais dos aminoácidos em A2.

Tabela 12: Variáveis contidas no modelo aug-MIA-QSARcolor para o parâmetro k_{cat} , com compostos ordenados do menor para o maior valor de k_{cat}

Cp	X309	X2090	X2126	X2498	X6660	X7320	k_{cat}
9	765	765	765	210	765	765	-1.4
30	210	350	250	250	765	350	-1.22
3	210	350	250	765	350	250	-1.15
10	250	250	250	250	250	250	-1.15
19	210	350	250	765	765	350	-1.15

Tabela 12 (Continua)

Cp	X309	X2090	X2126	X2498	X6660	X7320	k_{cat}
27	250	250	250	250	765	350	-1.15
40	210	350	250	250	612	250	-1.15
41	210	350	250	250	350	250	-1.15
2	250	250	250	250	612	250	-1.1
24	612	350	250	250	765	765	-1.1
31	250	250	250	765	765	765	-1.1
42	250	250	250	765	765	350	-1.1
22	612	350	250	250	765	765	-1.05
23	765	765	765	210	350	250	-1.05
13	612	350	250	250	350	250	-1
36	612	350	250	250	765	765	-1
15	250	612	250	250	350	250	-0.85
26	250	250	250	765	765	250	-0.85
43	210	350	250	250	765	250	-0.85
44	612	350	250	250	612	250	-0.77
46	765	765	765	765	765	765	-0.77
21	765	765	765	765	350	250	-0.74
32	765	765	765	765	350	250	-0.68
14	765	350	250	250	765	250	-0.66
35	765	765	765	210	765	350	-0.66
37	250	612	250	250	350	250	-0.66
1	765	765	765	765	765	765	-0.62
18	765	765	765	210	612	250	-0.59
45	765	765	210	210	250	250	-0.57
16	612	350	250	250	765	210	-0.55
33	612	350	250	250	765	350	-0.54
50	765	765	765	765	765	350	-0.54
5	765	350	250	250	250	250	-0.52
6	765	765	765	765	765	350	-0.48
49	765	765	765	765	765	765	-0.47
17	765	765	350	350	250	250	-0.41
7	765	765	210	210	765	765	-0.39
11	765	765	765	765	765	765	-0.37
38	765	765	765	765	765	350	-0.32
52	765	765	765	765	765	210	-0.26

Tabela 12 (Conclusão)

Cp	X309	X2090	X2126	X2498	X6660	X7320	k_{cat}
48	765	765	350	350	250	250	-0.2
8	765	765	350	350	612	250	-0.13
12	765	765	765	765	765	210	-0.12
29	765	765	765	765	765	250	-0.05
25	765	765	765	765	765	210	-0.03
28	765	765	350	350	765	350	-0.03
20	765	765	350	765	765	765	-0.01
53	765	765	350	765	765	210	0.19
54	765	765	350	765	765	210	0.21
4	765	765	350	765	350	250	0.23
39	765	765	350	350	765	250	0.25
47	765	765	350	765	765	210	0.27
34	765	765	350	350	765	210	0.3
51	765	765	350	765	765	210	0.4

Dado que nenhuma variável no modelo de aug-MIA-QSPRcolor corresponde aos grupos R nas posições A3 e A4, pode-se inferir que essas não têm qualquer influência sobre a taxa de clivagem do substrato e, assim, essas posições podem ser importantes para a modificação de outras propriedades, tais como fatores físico-químicos e farmacocinéticos sem alterar significativamente a taxa de clivagem.

Esse resultado é consistente com um estudo anterior sobre inibidores aldeídos tetrapeptídeos (LI et al., 2005), no qual foi relatado que interações na cadeia lateral de aminoácidos na posição A2 são mais importantes do que nas cadeias laterais em A1, A3 e A4, e que modificações nas posições A3 e A4 são toleradas sem redução significativa da taxa de clivagem, com este último (A4) contribuindo pouco para a ligação com a enzima.

Cadeias laterais de tamanho pequeno em A1 [identificadas pelos pixels de valores 765 (**X309**), 765 (**X2090**), 350 (**X2126**) e 765 (**X2498**), correspondente a um grupo hidroxila separado por uma distância de um carbono do esqueleto peptídico, como na serina, e valores de pixel 765 (espaço em branco) para todas as quatro posições de pixel indicando grupo R

muito pequeno nesta posição, tais como o hidrogênio e um grupo metila, como na alanina e glicina] estão associadas a valores elevados de k_{cat} . Essas descobertas são consistentes com estudos anteriores em que foi relatado que pequenos aminoácidos não ramificados na posição A1 contribuem para o aumento da taxa de clivagem dos substratos, já que esses favorecem as interações enzima-substrato (CHAMBERS, et al., 1990; LI et al., 2005). Note-se que quando o substrato tetrapeptídico contém aminoácidos pequenos não ramificados em A1 e um grupo metila para o aminoácido A2 [identificado por pixel de valores 765 (espaço em branco) e 210 (H) para as variáveis **X6660** e **X7320**, respectivamente], maiores valores de taxa de clivagem são obtidos, enquanto que o oposto é observado quando o aminoácido em A2 contém uma cadeia lateral cíclica formando parte da espinha dorsal [identificado por espaços em branco para as variáveis **X6660** e **X7320**], como na prolina. Os demais substituintes do grupo R para A2 não mostram uma tendência em particular, uma vez que estão espalhados por toda a gama de valores de k_{cat} . Isso sugere que A2 permite uma variada gama de aminoácidos, corroborando os resultados de um estudo anterior de LI et al. (2005). Por outro lado, o grupo amida [identificado por valores de pixel 612 (**X309**), 350 (**X2090**), 250 (**X2126**) e 250 (**X2498**)], imidazol [representado por o pixel átomo de carbono (250) para as quatro variáveis] ou carboxila [identificado por o pixel de valores 210 (**X309**), 350 (**X2090**), 250 (**X2126**) e 250 (**X2498**)], como cadeias laterais de aminoácidos em A1, estão relacionados com baixos valores de K_{cat} e, assim, asparagina, histidina e ácido aspártico são considerados inadequados para esta posição. No seu conjunto, esses resultados fornecem informações valiosas sobre as características-chave estruturais no desenvolvimento de novos tetrapeptídeos com maior atividade contra o DENV 02.

4 CONCLUSÃO

Desde o início da abordagem MIA-QSAR, em 2005, mudanças têm ocorrido visando incorporar informações químicas mais significativas para as imagens estruturais. O uso de esquemas de cores para distinguir diferentes tipos de átomos e a modificação de tamanhos atômicos de acordo com o raio de van der Waals melhorou o desempenho dos modelos MIA-QSAR e abriu caminho para a sua interpretação estrutural. No presente trabalho, a abordagem aug-MIA-QSARcolor produziu modelos baseados em MLR mais preditivos para os parâmetros cinéticos K_m e k_{cat} do que as abordagens MIA-QSAR tradicional e aug-MIA-QSAR. As variáveis contidas no modelo aug-MIA-QSARcolor destacou a importância dos aminoácidos pequenos e ramificados nas posições A1 e A4 no aumento da afinidade para o substrato à proteína NS3 e na taxa de clivagem do substrato em A1 e A2. Por outro lado, aminoácidos ácidos nas posições A2 e A4 foram relacionados com baixa afinidade do substrato para a proteína NS3 e baixos valores de clivagem do substrato na posição A1. Esses resultados fornecem informação química útil para o planejamento de compostos-alvo contra o DENV.

REFERÊNCIAS

- ADVANCED CHEMISTRY DEVELOPMENT. **ACD/ChemSketch - version 12.01**. Toronto, Advanced Chemistry Development, 2009.
Disponível em: <http://alainsheratan.com/chemsk_t12.pdf>. Acesso em: 10/02/2013.
- ASSENBERG, R. et al. Crystal structure of a novel conformational state of the flavivirus NS3 protein: implications for polyprotein processing and viral replication. **Journal of Virology**, Washington, v. 83, n. 24, p. 12895-12906, Dec. 2009.
- BAZAN, J. F.; FLETTERICK, R. J. Detection of a trypsin-like serine protease domain in flaviviruses and pestviruses. **Virology**, San Diego, v. 171, n. 2, p. 637-639, Aug. 1989.
- BHATT, S. et al. The global distribution and burden of dengue. **Nature**, London, v. 496, n. 7446, p. 504-507, Apr.2013.
- BRADY, O. J. et al. Refining the Global Spatial Limits of Dengue Virus Transmission by Evidence-Based Consensus. **PLoS Neglected Tropical Disease**, San Francisco, v. 6, n. 8, Aug. 2012.
- BRASIL. Ministério da saúde. **Boletim Epidemiológico: Monitoramento dos casos de dengue, febre de chikungunya e febre pelo vírus Zika até a Semana Epidemiológica 52, 2015**: Secretaria de Vigilância em Saúde. Vol.47, n. 20, 2016. Disponível em: <<http://portalsaude.saude.gov.br/index.php/oministerio/principal/leia-mais-o-ministerio/197-secretaria-svs/11955-boletins-epidemiologicos-arquivos>>. Acesso em: 20/02/2016
- BRASIL. Ministério da Saúde. **Doenças infecciosas e parasitárias**: Guia de bolso. 8 ed. Brasília: Ed. MS, 2010, 444p.
- CHAMBERS, T. J. et al. Evidence that the N-terminal domain of nonstructural protein NS3 from yellow fever virus is a serine protease responsible for site-specific cleavages in the viral polyprotein. **Proceedings of the National Academy Sciences of the United States America**, Washington, v. 87, n. 22, p. 8898-8902, Nov. 1990
- CHIRICO, N.; GRAMATICA, P. Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. **Journal Chemical Information and Modeling**, Washington, v. 52, n. 8, p. 2044-2058, June 2012.

DENNINGTON, R. D.; KEITH, T.A; MILLAM, J.M. **GaussView 5.0. 8**; Gaussian Inc., Wallingford, 2008. CD-ROM.

DINU, S. et al. Molecular epidemiology of dengue fever cases imported into Romania between 2008 and 2013. **Travel Medicine and Infectious Disease**, Oxford, Vol. 13, n. 1, p. 69-73, Jan-Feb.2015

FALGOUT, B. et al. Both nonstructural proteins NS2B and NS3 are required for the proteolytic processing of dengue virus nonstructural proteins. **Journal of Virology**, Washington, v. 65, n. 5, p. 2467-2475, May 1991.

GIBBONS, R. V.; VAUGHN, D. W. Dengue: an escalating problem. **British Medical Journal**, London, v. 324, n. 7353, p. 1563-1566, June 2002.

GODÓI, I. P. et al. NS2B-NS3pro como Alvo Molecular para o Desenvolvimento de Fármacos contra Dengue. **BBR-Biochemistry and Biotechnology Reports**, Londrina, v. 3, n. 2, p. 16-30, Jul./Dez. 2014.

GOLDBERG, D. E.; HOLLAND, J. H. Genetic algorithms and machine learning. **Machine learning**, Dordrecht, v. 3, n. 2, p. 95-99, Oct. 1988.

GUBLER, D. J. Dengue and dengue hemorrhagic fever. **Clinical Microbiology Reviews**, Washington, v. 11, n. 3, p. 480-496, Jul.1998.

GUIDO, R. V.; ANDRICOPULO, A. D.; OLIVA, G. Planejamento de fármacos, biotecnologia e química medicinal: aplicações em doenças infecciosas. **Estudos Avançados**, São Paulo, v. 24, n. 70, p. 81-98, Set./Dez. 2010.

GUZMAN, M. G.; HARRIS, E. Dengue. **The Lancet**, New York, v. 385, n. 9966, p. 453-465, Jan. 2015.

KALAYANAROOJ, S. Clinical manifestations and management of dengue /DHF/DSS. **Tropical Medicine and Health**, Nagasakiv, v. 39, n. 4 Suppl, p. 83, Dec. 2011.

KAMPMANN, T. et al. In silico screening of small molecule libraries using the dengue virus envelope E protein has identified compounds with antiviral activity against multiple flaviviruses. **Antiviral research**, Amesterdan, v. 84, n. 3, p. 234-241, Dec. 2009.

LANCIOTTI, R. S. et al. Molecular evolution and epidemiology of dengue-3 viruses. **Journal of General Virology**, Berks, v. 75, n. 1, p. 65-76, Jan. 1994.

LESCAR, J. et al. Towards the design of antiviral inhibitors against flaviviruses: The case for the multifunctional NS3 protein from Dengue

virus as a target. **Antiviral Research**, Amsterdam, v. 80, n. 2, p. 94-101, Nov. 2008.

LI, J. et al. Functional profiling of recombinant NS3 proteases from all four serotypes of dengue virus using tetrapeptide and octapeptide substrate libraries. **Journal Biological Chemistry**, Bethesda, v. 280, n. 31, p. 28766-28774, Aug. 2005.

LUO, D. et al. Insights into RNA unwinding and ATP hydrolysis by the flavivirus NS3 protein. **EMBO Journal**, Hoboken, v. 27, n. 23, p. 3209-3219, Dec. 2008.

MCMENIMAN, C. J. et al. Stable introduction of a life-shortening Wolbachia infection into the mosquito *Aedes aegypti*. **Science**, Washington, v. 323, n. 5910, p. 141-144, Jan. 2009.

MASON, P. W. et al. Sequence of the dengue-1 virus genome in the region encoding the three structural proteins and the major nonstructural protein NS1. **Virology**, San Diego, v. 161, n. 1, p. 262-7, Nov. 1987.

MITRA, I.; SAHA, A.; ROY, K. Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. **Molecular Simulation**, Oxon, v. 36, n. 13, p. 1067-1079, Nov. 2010.

NUNES, C. A. et al. Chemoface: a novel free user-friendly interface for chemometrics. **Journal of the Brazilian Chemical Society**, Campinas, v. 23, n. 11, p. 2003-2010, Nov. 2012.

OJHA, P. K. et al. Further exploring r m 2 metrics for validation of QSPR models. **Chemometrics and Intelligent Laboratory Systems**, Amsterdam, v. 107, n. 1, p. 194-205, May 2011.

ORGANIZAÇÃO MUNDIAL DE SAÚDE-OMS. **Dengue and severe dengue**. Disponível em: <<http://www.who.int/mediacentre/factsheets/fs117/en/>>. Acessado em 15/05/2015a.

ORGANIZAÇÃO MUNDIAL DE SAÚDE- OMS. **Investing to overcome the global impact of neglected tropical diseases: third WHO report on neglected diseases**. Geneva: Organização Mundial de Saúde, 2015b, 191p. Disponível em: <<http://apps.who.int/bookorders/anglais/detart1.jsp?codlan=1&codcol=15&codcch=3853>>. Acesso em: 15/06/15

PAMPLONA, L. D. G. C. et al. Reduced oviposition of *Aedes aegypti* gravid females in domestic containers with predatory fish. **Tropical**

Medicine & International Health, Hoboken, v. 14, n. 11, p. 1347-1350, Nov. 2009.

PRUSIS, P. et al. Proteochemometrics analysis of substrate interactions with dengue virus NS3 proteases. **Bioorganic Medicinal Chemistry**, Oxford, v. 16, n. 20, p. 9369-9377, Oct. 2008

ROY, K. et al. Some case studies on application of “ r^2_m ” metrics for judging quality of quantitative structure–activity relationship predictions: emphasis on scaling of response data. **Journal Computational Chemistry**, Hoboken, v. 34, n. 12, p. 1071-1082, May 2013.

SILLA, J. M. et al. MIA-QSPR and effect of variable selection on the modeling of kinetic parameters related to activities of modified peptides against dengue type 2. **Chemometrics and Intelligent Laboratory Systems**, Amsterdam, v. 108, n. 2, p. 146-149, Oct. 2011.

URIAS, R. W. P. et al. IMMAN: free software for information theory-based chemometric analysis. **Molecular Diversity**, Dordrecht, v. 19, n. 2, p. 305-319, May.2015.

VALDEZ, M. R. W. et al. Genetic elimination of dengue vector mosquitoes. **Proceedings of the National Academy Sciences of the United States America**, Washington, v. 108, n. 12, p. 4772-4775, Mar. 2011.

ZHANG, Y. et al. Conformational changes of the flavivirus E glycoprotein. **Structure**, Cambridge, v. 12, n. 9, p. 1607-1618, Sept. 2004.

CAPÍTULO 5: MODELAGEM COMPUTACIONAL DA ATIVIDADE ANTIESQUISTOSSOMÓTICA DE DERIVADOS NEOLIGNANAS COM BASE NA ABORDAGEM MIA-SAR

RESUMO

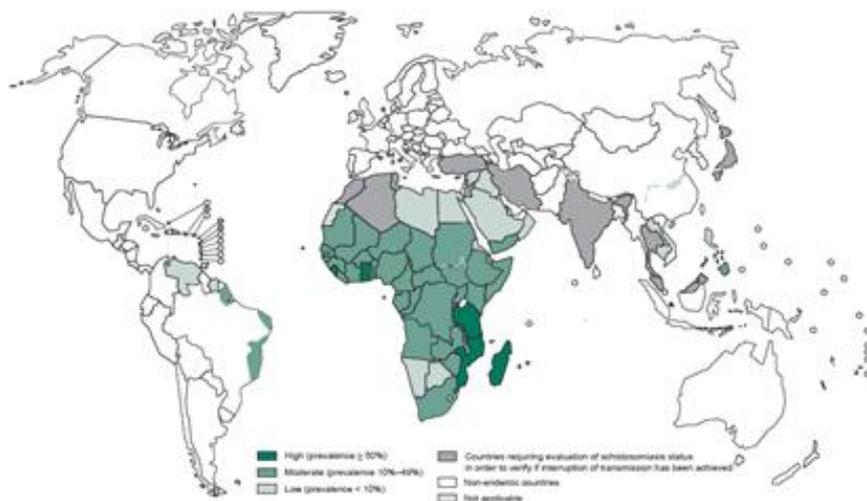
Modelos teóricos para explorar a atividade antiesquistossomótica de um conjunto de dados de 18 neolignanas sintéticas foram construídos utilizando a abordagem de análise multivariada de imagens aplicada a relações estrutura-atividade (MIA-SAR). Os modelos obtidos foram validados usando a acurácia da validação cruzada *leave-one-out* (Acc), validação externa e procedimentos de Y-randomização, resultando em uma classificação correta superior a 80%, 70% e 60%, respectivamente. Adicionalmente, foi feita uma comparação dos modelos obtidos a partir de imagens binárias (preto e branco) e coloridas; as cores (valores de pixel) foram selecionadas para corresponder às propriedades químicas. Observou-se que os modelos obtidos a partir de imagens coloridas com os valores de pixel correspondentes à eletronegatividade (conhecida como abordagem *aug-MIA-SARcolor*) produziram parâmetros estatísticos superiores aos obtidos a partir de imagens binárias (MIA-SAR) e de imagens coloridas aleatoriamente (em que átomos são coloridos de acordo com o padrão de cores do programa utilizado). Interpretação mecanística da influência de diferentes substituintes na atividade antiesquistossomótica revelou que substituintes metoxila nas posições R1 (ou R2) e R5 do esqueleto básico das neolignanas são indispensáveis para a atividade antiesquistossomótica. Os resultados obtidos fornecem conhecimento das possíveis modificações estruturais para produzir novos compostos derivados de neolignanas com atividade antiesquistossomótica.

Palavras-chave: *Schistosoma mansoni*, PLS-DA, PCA, HCA.

1 INTRODUÇÃO

A esquistossomose, também conhecida como bilharziose ou “barriga d’água”, é uma doença causada por helmintos do gênero *Schistosoma* (BRASIL, 2004). Endêmica em 78 países de regiões tropicais e subtropicais, a doença afeta, principalmente, comunidades pobres sem água potável e saneamento adequado. Estima-se que mais de 700 milhões de pessoas vivam em áreas de risco de infecção, número que corresponde a 10% da população mundial. O número de pessoas infectadas em 2012 foi de 249 milhões, das quais 45% eram crianças em idade escolar (5 a 14 anos) (OMS; 2014). Atualmente, 52 países apresentam transmissão suficientemente efetiva que justifique a quimioterapia preventiva, destes 90% estão situados no continente africano. Contudo, moderadas taxas de transmissão continuam em países como Brasil, Suriname, Venezuela, Filipinas, Camboja e Laos (OMS, 2015) (Figura 18).

Figura 18: Distribuição da esquistossomose em todo mundo em 2012



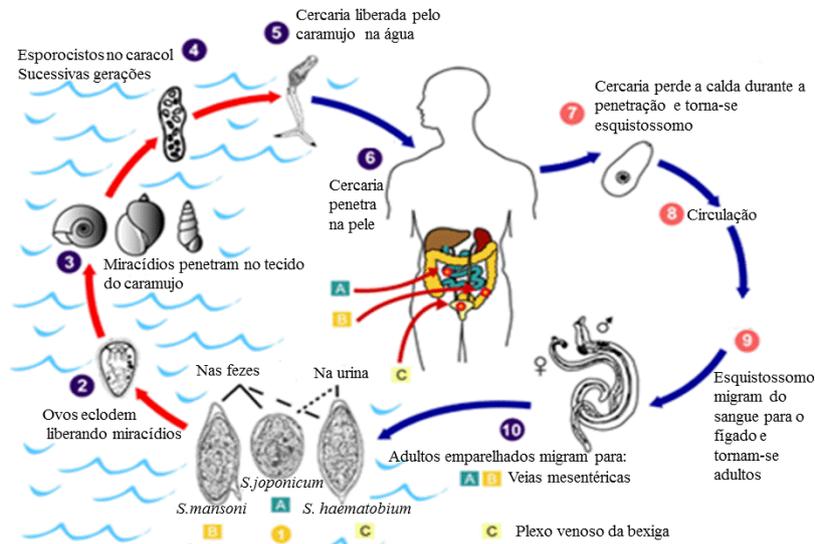
Fonte: (OMS, 2015)

Seis espécies de schistosomas são responsáveis por infecções em humanos: *S. haematobium*, *S. japonicum*, *S. mansoni*, *S. intercalatum*, *S. mekongi* e *S. guineensis* (BRASIL, 2004; OMS, 2015). A maioria dos casos da doença é resultante de infecções por *S. haematobium* (que causa a esquistossomose urogenital) e *S. mansoni* (causador da esquistossomose intestinal). Os *haematobium* estão basicamente distribuídos na África e Oriente Médio, enquanto que o *S. mansoni* pode ser encontrado na África, Oriente Médio, Brasil, Venezuela, Suriname e Ilhas Caribenhas. (CDC, 2002; OMS, 2015).

A principal fonte de contaminação do parasito ao homem é através de água doce estagnada, em lagos, represas e açudes que contenham as larvas (cercárias) do *Schistosoma* (BRASIL, 2004). Geralmente, a contaminação desses locais se faz por meio do hospedeiro vetor e através de fezes de vertebrados contaminados (OMS, 2015). Os ovos do *S. mansoni* são eliminados nas fezes do hospedeiro infectado. Na água eclodem, liberando uma larva ciliada denominada miracídio, que infecta os caramujos do gênero *Biomphalaria*: *B. glabrata*, *B. tenagophila* e *B. straminea*. Após 4 a 6 semanas, abandonam o caramujo, na forma de cercária, ficando livres nas águas naturais (Figura 19). Essas possuem considerável resistência, podendo permanecer por dias na água, até encontrar um hospedeiro vertebrado. O contato humano com águas infectadas pelas cercárias é a maneira pela qual o indivíduo adquire a esquistossomose (BRASIL, 2004; OMS, 2013).

As manifestações clínicas da doença dependem do estágio de evolução do parasito no hospedeiro e do tipo de parasito. A doença é dividida em fase aguda e crônica. A fase aguda, geralmente, ocorre em um período de 3 a 7 semanas após exposição ao parasito, sendo caracterizada por febre, anorexia, dor abdominal, cefaleia e vômito. Após 6 meses de infecção, a doença pode ser considerada crônica, podendo apresentar-se na forma de diária recorrente, aumento ou contração do fígado devido à fibrose, esplenomegalia, ascite, varizes de esôfago, hematêmase, anemia, desnutrição e hiperesplenismo (BRASIL, 2004).

Figura 19 : Ciclo de vida dos *Schistosoma* no caramujo e no homem.



Fonte: (CDC, 2002)

O principal medicamento utilizado no tratamento da esquistossomose é o praziquantel. Essa droga atua contra todas as espécies de *Schistosoma* (CIOLI; PICA-MATTOCCIA, 2003). No entanto, apesar de demonstrar resultados satisfatórios e baixo custo, sua disponibilidade ainda é limitada. Dados da OMS de 2014 mostram que o número de pessoas tratadas por esquistossomose foi de 61.635.963. Isto representa 20,7% da cobertura global da população que necessitava de terapia preventiva e apenas 34,6% das 123.329.536 crianças em idade escolar (5 a 14 anos) que necessitavam de tratamento (OMS, 2016). Outro fator relevante é que o praziquantel vem sendo utilizado pelos programas de controle e prevenção ao *S. mansoni* há mais de 40 anos e os riscos de resistência do parasito são eminentes (WANG, WANG; LIANG, 2012). Nesse sentido, pesquisas têm sido realizadas visando buscar drogas mais eficientes ao tratamento de esquistossomose, visto que o praziquantel não impede a reincidência da doença (MCMANUS; LOUKAS, 2008), é inativo contra schistossomas jovens e possui efeito limitado em pacientes com lesões avançadas no fígado e baço (CHATTERJEE et al., 2002; SABAH et al., 1986.).

Um estudo iniciado na década de 1970 aponta que neolignanas extraídas de folhas da planta *Virola surinamensis* apresentavam atividade biológica contra microbactérias (SOUZA et al., 2011), fungos (ZACCHINO et al., 1999) e parasitos, tais como *Leishmania* spp. (SANTOS, 1991, BARATA et al., 2000), *Trypanosoma cruzi* (NOCITO et al., 2007) e *S. mansoni* (SANTOS, 1991). Extrato das folhas e, posteriormente, substâncias isoladas mostraram alta atividade contra a penetração de cercárias do *Schistosoma mansoni* em camundongos (SANTOS, 1991). Deste então, análogos e derivados dessas substâncias têm sido estudados visando compostos potencialmente mais promissores ao combate da esquistossomose.

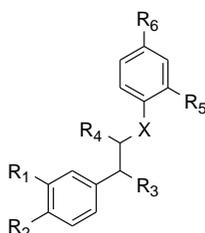
Tendo em vista a necessidade de conhecer melhor as propriedades e mecanismo de ação desses compostos, técnicas computacionais para planejamento racional de fármacos têm sido grandes aliadas nas pesquisas, uma vez que reduz tempo, custo e projeta compostos mais promissores ao desenvolvimento de uma nova droga. Estudos já realizados sugerem que propriedades eletrônicas são importantes para a atividade antiesquistossomótica (ALVES et al., 2002). Nesse sentido, o objetivo deste trabalho é mostrar como as metodologias MIA-QSAR podem contribuir para a interpretação e elaboração de neolignanas mais promissoras ao tratamento da esquistossomose.

2 METODOLOGIA

2.1 Obtenção do conjunto de dados

Um conjunto de dados constituído de 18 neolignanas com atividade relatada contra a penetração da cercária do *S. mansoni* em ratos foi explorado (Quadro 3) (ALVES et al., 2002). As neolignanas são dímeros formados a partir do acoplamento oxidativo de fenóis alílicos e propenílicos que ocorrem nas plantas da família *myristicaceae*. As neolignanas

Quadro 3 - Estrutura química e atividade antiesquistossomótica dos derivados neolignanais



Cp	Ativ.	R1	R2	R3	R4	R5	R6	X
1	Inativo	H	H	O	H	H	H	O
2	Inativo	H	H	O	CH ₃	H	H	O
3	Inativo	H	H	O	CH ₃	OCH ₃	H	O
4	Inativo	H	H	O	CH ₃	OCH ₃	CH ₂ CHCH ₂	O
5*	Ativo	OCH ₃	OCH ₃	O	CH ₃	OCH ₃	CHCHCH ₃	O
6	Ativo	OCH ₃	OCH ₃	O	CH ₃	OCH ₃	CH ₂ CHCH ₂	O
7	Inativo	H	H	OH	CH ₃	OCH ₃	H	O
8	Ativo	OCH ₃	OCH ₃	OH	CH ₃	OCH ₃	CH ₂ CHCH ₂	O
9	Ativo	OCH ₃	OCH ₃	OH	CH ₃	OCH ₃	CHCHCH ₃	O
10*	Inativo	H	H	O	CH ₃	OCH ₃	H	S
11	Inativo			O	CH ₃	OCH ₃	H	S
12	Inativo	OCH ₃	OCH ₃	O	CH ₃	H	H	S
13	Inativo	OCH ₃	OCH ₃	O	CH ₃	H	CH ₃	S
14	Inativo	H	H	O	H	H	H	NH
15	Inativo	OCH ₃	OCH ₃	O	CH ₃	H	H	NH
16*	Inativo	OCH ₃	OCH ₃	O	CH ₃	CH ₃	H	NH
17	Ativo	OCH ₃	OCH ₃	O	CH ₃	OCH ₃	H	NH
18	Inativo	OCH ₃	OCH ₃	O	CH ₃	H	H	NCH ₃

*Compostos selecionados para grupo teste. Fonte: Adaptado de Alves et al. (2002)

que ocorrem nas plantas da família *myristicaceae*. As neolignanais empregadas neste trabalho são as do tipo “8.O.4”, com uma ponte de oxigênio entre os carbonos 8 e 4 (APERS.; VLIETINCK; PIETERS, 2003), onde cinco compostos foram relatados como sendo ativos (5, 6, 8, 9, 17) e 13 inativos (1, 2, 3, 4, 7, 10, 11, 12, 13, 14, 15, 16, 18) ao *S. mansoni*. Esses compostos foram modelados segundo as abordagens MIA-SAR, aug-MIA-SAR e aug-MIA-SARcolor, e avaliados pelas técnicas de reconhecimento de padrão: análise de componentes principais (PCA), análise de agrupamento

Após o alinhamento das imagens das 18 estruturas químicas, a estrutura de imagem multivariada que se seguiu foi desdobrada para obter uma matriz de dados bidimensional. Subsequentemente, as colunas (variáveis) com variação nula foram eliminadas, produzindo matrizes com dimensões de: 18×1149 para MIA-SARt e 18×5349 para as abordagens aug-MIA-SAR e aug-MIA-SARcolor. A conversão das imagens em matrizes de dados e a redução da dimensionalidade foram realizadas no programa Chemoface (NUNES et al., 2012).

2.3 Análise de componentes principais e análise de agrupamento hierárquico

Análise de componentes principais (PCA) é um método não supervisionado de análise de dados destinado à estratificação de informações contidas numa matriz de dados em variáveis latentes (conhecido como componentes principais), que são de natureza ortogonal. O primeiro fator explica a maior variância do conjunto de dados e cada componente sucessiva procura maximizar a variância não explicada pelo fator anterior e, assim, obtém uma estrutura de dados de menor dimensão. Neste relatório, a PCA foi realizada para analisar as diferenças/semelhanças intrínsecas entre as estruturas químicas, utilizando como exemplo o gráfico de *scores*. O critério do autovalor ($\lambda > 1,00$) foi utilizado para determinar o número de componentes principais importantes.

Como um estudo complementar, a análise de agrupamento hierárquico (HCA) foi realizada para adquirir maior compreensão da estrutura subjacente da matriz do conjunto de dados. Note que, embora a HCA também pertença aos métodos de análise de dados não supervisionados, possui duas grandes diferenças do PCA. Primeiro, a HCA emprega métricas de distância (por exemplo, a distância euclidiana) como medidas de similaridade e/ou dissimilaridade e, portanto, não implica uma suposição *a priori* de uma relação linear entre as variáveis/casos, enquanto

que os componentes principais são derivados a partir de matriz de correlação/covariância, o que sugere uma relação linear. Segundo, uma vez que os algoritmos de amalgamação não procuram maximizar a variância do conjunto de dados em um determinado agrupamento, mas sim maximizar a distância intergrupo, HCA origina uma estrutura mais abrangente, que permite a detecção da tendência natural da estrutura de dados, bem como a detecção de comportamentos anômalos.

2.4 A análise discriminante por mínimos quadrados parciais (PLS-DA)

A PLS-DA é uma ferramenta útil em tarefas de classificação quando se lida com matrizes de dados de alta dimensionalidade e com variáveis altamente correlacionadas (CHEVALLIER et al., 2006). Esse método se baseia na transformação das variáveis originais em variáveis latentes, as quais são linearmente independentes. As variáveis latentes são uma combinação linear das variáveis originais. No entanto, um estudo anterior demonstrou que, quando há poucas variáveis originais na matriz de alta dimensionalidade, que são de fato relevantes para a propriedade modelada, o PLS é mais propenso a erros do tipo I (omitindo características estruturais importantes para o bloco y). Portanto, levando em conta que o método MIA-SAR se baseia no princípio de que a classificação de compostos como ativos/inativos é uma função de porções químicas em posições particulares do esqueleto base de uma série congênere de compostos, o risco de erros do tipo I é, em grande parte, possível. Nesse sentido, procedimentos de redução de dimensão foram realizados nas matrizes de dados originais, utilizando a informação de parâmetro teórico "ganho de informação" (IG). Em seguida, a PLS-DA foi aplicada para redução da matriz de dados das abordagens MIA-SARt, aug-MIA-SAR e aug-MIA-SARcolor, usando 14 (77,8%) compostos na série treinamento e 4 (22,2%) na série de testes, escolhido aleatoriamente.

A qualidade dos modelos de classificação obtidos foi avaliada considerando os parâmetros: classificação correta (acurácia) na previsão de

[Acc (p)]; validação cruzada *leave-one-out* [Acc (CVloo)]; y-randomização (média de 10 repetições); e validação externa [Acc (ext)]. No procedimento CVloo é retirada uma amostra ao acaso do conjunto de dados e um modelo é desenvolvido com as (n - 1) amostras restantes. Em seguida, uma predição da amostra excluída é realizada. Este processo é continuado até que os modelos *n* validados são obtidos. A precisão do modelo de classificação é, então, determinada como a média de desempenho para os *n* modelos validados. Uma limitação desse procedimento de validação cruzada é que ele, sozinho, não fornece critérios que garantam o poder preditivo do modelo estimado e, portanto, a validação de um conjunto de compostos não empregados na construção de modelos (isto é, a validação externa) é necessário. Tendo em mente que poucos pontos de dados são empregados no presente estudo, o conjunto de dados foi dividido 5 vezes em conjuntos de treinamento e teste e a precisão média foi, então, relatada.

Uma das principais limitações da PLS-DA como um método de classificação estatística é a interpretabilidade, uma vez que as variáveis originais são transformadas em variáveis latentes para maximizar a comunalidade dos blocos de variáveis independentes e dependentes, assim, 'mascarando' as informações contidas nas variáveis originais. No entanto, sabe-se que a PLS-DA e a análise discriminante linear (LDA) fornecem o mesmo resultado quando o número de variáveis latentes considerado em PLS-DA é igual ao número de variáveis originais. Por conseguinte, a fim de produzir modelos mais interpretáveis, pequenos subconjuntos de variáveis ortogonais foram extraídos a partir das matrizes de dados reduzidas com base em um algoritmo de seleção de variáveis por passos, em que correlação com a classificação de variáveis é maximizada, enquanto a multicolinearidade entre as variáveis independentes é minimizada. Prevê-se que os modelos construídos sobre este subconjunto de variáveis ortogonais devem fornecer um maior conhecimento sobre a influência de diferentes substituintes na atividade antiesquistossomótica.

3 RESULTADOS E DISCUSSÕES

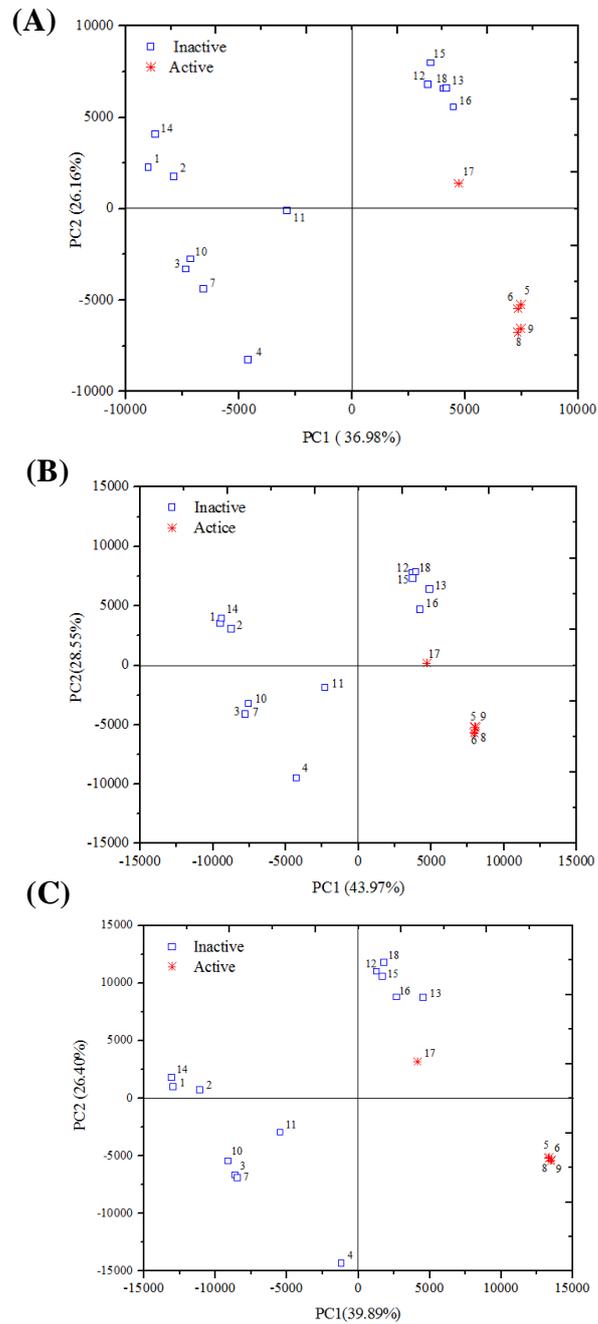
3.1 Análise de componentes principais

O gráfico de *scores* para as duas primeiras componentes principais (PC1 e PC2) para as estratégias MIA-SARt (Gráfico 9A), aug-MIA-SAR (Gráfico 9B) e aug-MIA-SARcolor (Gráfico 9C) explicam 63,14%, 72,52% e 66,29% da variância dos dados, respectivamente.

Como podem ser observados, com a exceção do composto 17, os três gráficos de *scores* mostram a delimitação das classes de dados das neolignanas ativas e inativas (ou seja, os compostos ativos agrupam-se). A diferença-chave entre a molécula 17 e o conjunto de compostos ativos (C1) encontra-se nas posições dos substituintes X (O para NH) e R6 (grupo propenila para H), enquanto os compostos inativos (C2) (12, 13, 15, 16, 18) diferem na sua totalidade da molécula 17 somente nos substituintes da posição R5 (nenhum possui um grupo metoxila).

Nota-se uma maior proximidade da molécula 17 à classe C2, em vez de C1, o que sugere uma maior semelhança estrutural com a primeira. Essa tendência produz três inferências importantes: (1) substituição isostérica em substituinte da posição X parece não ter nenhum efeito sobre a atividade antiesquistossomótica, isto é, o substituinte X desloca a molécula 17 de C1, mas a atividade é mantida; (2) o grupo metoxila na posição substituinte R5 é importante para a atividade antiesquistossomótica, ou seja, enquanto todos os compostos inativos no conjunto C2 têm grupos metoxila em R1 e R2, sua única dissimilaridade global com a molécula 17 encontra-se com um grupo metoxi adicional em R5 para o último, enquanto que todos os membros de C1 possuem o substituinte metoxila em R5; (3) substituição do grupo propenila em R6 por H não tem efeito deletério sobre a atividade.

Gráfico 9 - Gráficos de *scores* para as componentes principais PC1 e PC2 da série de neolignanas, utilizando os descritores MIA-SARt (A), aug-MIA-SAR (B) e aug-MIA-SAR_{color} (C).



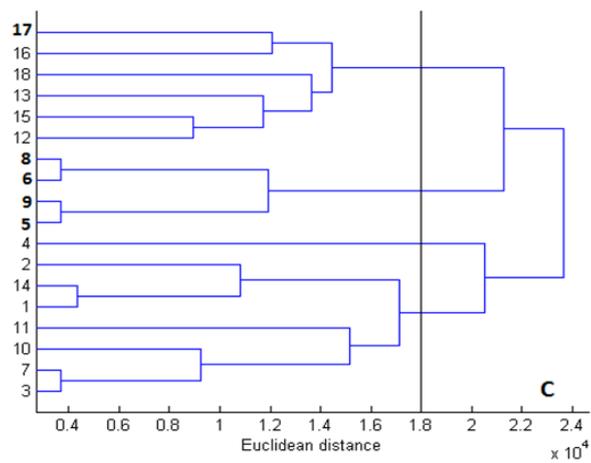
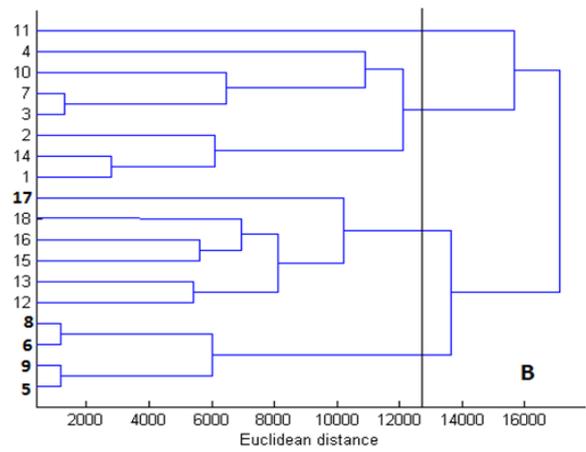
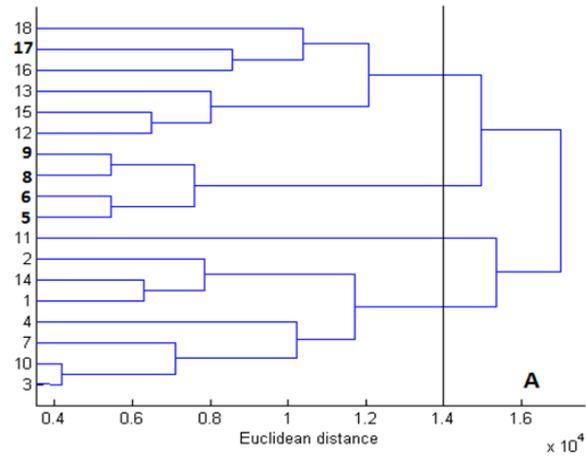
É importante notar que uma vez que PC1 e PC2 representam mais de 63% da variância em todos os casos, as variáveis originais (> 1000) parecem ser altamente correlacionadas. Isso é, na verdade, uma característica comum de matrizes de dados obtidas com o método MIA-SAR. Como consequência, os resultados HCA devem ser analisados com mais precaução.

3.2 Análise de agrupamento hierárquico (HCA)

O Gráfico 10 mostra os dendrogramas obtidos com a HCA de acordo com as estratégias MIA-SARt (Gráfico 10A), aug-MIA-SAR (Gráfico 10B), e aug-MIA-SARcolor (Gráfico 10C). Considerando-se as distâncias de corte delimitadas pelas linhas verticais nos dendrogramas, três aglomerados são obtidos para cada uma das estratégias MIA-SAR.

Uma análise desses aglomerados revela que, com exceção do composto 17, os compostos ativos se agrupam nos dendrogramas para as abordagens aug-MIA-SARcolor, MIA-SARt e aug-MIA-SAR de forma similar à PCA. Além disso, os dendrogramas das três abordagens sugerem característica estrutural atípica para o composto 11, que é um resultado lógico, visto que possui um único substituinte (grupo 1,3-dioxolano) nas posições R1 e R2. No entanto, o dendrograma para a abordagem aug-MIA-SARcolor (Gráfico 10C) indica comportamento atípico para o composto 4, o que pode ser atribuído à presença combinada do grupo propenila em R6 e átomos de hidrogênio em R1 e R2, dado que compostos com substituintes H em R1, R2 e R6 não exibem um comportamento atípico, nem aqueles com um grupo propenila em R6 e grupos metoxila em R1 e R2. Os resultados obtidos com HCA, em geral, corroboram com as conclusões obtidas na PCA sobre as características estruturais necessárias para a atividade antiesquistossomótica das neolignanas.

Gráfico 10 - Dendrograma dos modelos MIA-SARt (A), aug-MIA-SAR (B) e aug-MIA-SAR_{color} (C) para as 18 neolignanas.



3.3 PLS-DA

A Tabela 13 mostra os parâmetros estatísticos para os modelos MIA-SARt, aug-MIA-SAR e aug-MIA-SARcolor construídos usando a matriz original reduzida e os subconjuntos de descritores obtidos com o filtro IG e procedimentos de seleção de variáveis *stepwise*.

Tabela 13: Parâmetros estatísticos para os modelos MIA-SARt, aug-MIA-SAR e aug-MIA-SAR_{color}

Parâmetros (%)*	MIA-SARt	aug-MIA-SAR	aug-MIA-SARcolor
Acc(p)	100± 0,00	100± 0,00	100± 0,00
Acc(Y-rand)	69,57±8,35	70,54±8,83	69,43±8,16
Acc(CVloo)	94,29±5,35	96,43±3,57	95,71±3,50
Acc(ext)	90,00±12,25	87,50±12,50	90,00±12,50
Seleção de variáveis <i>stepwise</i>			
Acc(p)	94,64±3,09	89,29±3,57	100± 0,00
Acc(Y-rand)	71,43±0,00	68,04±2,11	70,00±0,71
Acc(CVloo)	94,64±3,09	82,14±6,19	100± 0,00
Acc(ext)	93,75±12,35	75,00±0,00	100± 0,00

*Valores médios.

Como pode ser observado, os modelos obtidos, geralmente, possuem bom comportamento estatístico, evidenciado pelas altas porcentagens de classificação correta (precisão) na previsão de [Acc (p)], validação cruzada interna [Acc (CVloo)] e de validação externa [Acc (ext)]. Adicionalmente, os baixos valores Y-randomização sugerem que os modelos obtidos não são propensos ao fenômeno de correlação ao acaso.

Em relação aos modelos obtidos usando o subconjunto filtrado por IG e procedimento de seleção de variáveis *stepwise*, esses aparentam ser comparáveis, apesar de uma melhora na precisão para o conjunto de validação externa ser observada com a estratégia aug-MIA-SARcolor (isto é, de 75% a 100%). Se as três abordagens MIA-SAR são comparadas, nota-se que os melhores resultados são obtidos com a estratégia aug-MIA-SARcolor, enquanto comportamento comparável é observado entre as abordagens MIA-

SARt e aug-MIA-SAR. O desempenho superior da abordagem aug-MIA-SARcolor indica a relevância da eletronegatividade para determinar a atividade antiesquistossomótica (pois os valores dos pixels dos átomos coloridos são proporcionais à eletronegatividade). Na verdade, isso está de acordo com relatos experimentais anteriores, que sugeriram a importância de propriedades eletrônicas na atividade antiesquistossomótica de neolignanas (Alves *et al.*, 2002).

Tendo em vista o melhor desempenho obtido com o modelo aug-MIA-SARcolor, foi realizada uma análise exploratória das informações codificadas pelas variáveis em termos da influência de diferentes átomos/grupos funcionais sobre a atividade biológica modelada.

3.4 Interpretação do modelo de aug-MIA-SARcolor

A incorporação de esquemas de cores para a abordagem MIA-SAR permite a identificação direta de átomos representados nas variáveis e uma maior compreensão de sua influência sobre a bioatividade modelada. A Tabela 14 mostra as variáveis contidas nos modelos finais obtidos de acordo com as abordagens MIA-SARt, aug-MIA-SAR e aug-MIA-SARcolor, respectivamente.

O aug-MIA-SARcolor é construído a partir de duas variáveis ortogonais, obtido pela seleção de variáveis *stepwise*, com cada variável representando um pixel em uma posição específica da imagem multivariada. Como pode ser observado, a variável **X78** possui valores de pixel que correspondem a espaços em branco (750) para as moléculas (1-4, 7, 10-11, 14) e valores de pixels correspondentes a ligações químicas (612) para o restante das moléculas (5-6, 8-9, 12-13, 15-18), o que sugere que os espaços em branco correspondem a um pixel na vizinhança de substituintes pequenos sobreposto com um pixel para ligação vinculada ao esqueleto básico de um substituinte grande.

Tabela 14 - Variáveis selecionadas para modelos MIA-SARt, aug-MIA-SAR e aug-MIA-SARcolor.

Cp	MIA-QSAR		aug-MIA-QSAR				aug-MIA-QSAR _{color}	
	X1	X422	X596	X3520	X4996	X5237	X78	X4392
1	765	0	765	750	750	750	750	210
2	765	0	765	750	750	750	750	210
3	765	0	765	750	426	612	750	350
4	765	0	0	426	426	612	750	350
5	0	765	0	426	426	612	612	350
6	0	765	0	426	426	750	612	350
7	765	765	765	750	426	612	750	350
8	0	765	0	426	426	612	612	350
9	0	765	0	426	426	612	612	350
10	765	0	765	750	426	612	750	350
11	765	0	765	750	426	750	750	350
12	0	0	765	750	750	750	612	210
13	0	0	765	750	750	750	612	210
14	765	0	765	750	750	750	750	210
15	0	0	765	750	750	750	612	210
16	0	765	765	750	612	750	612	250
17	0	765	765	750	426	750	612	350
18	0	765	765	750	750	750	612	210

Uma cuidadosa análise desse padrão revela que os pixels da variável **X78** correspondem ao substituinte R1/R2 (note que R1 e R2 são idênticos), onde os substituintes de menor e maior dimensão são hidrogênio e os grupos metoxila, respectivamente (ver também Quadro 3). Curiosamente, todos os compostos sem grupos metoxila para essa variável são inativos. Note-se que o inverso não é verdadeiro, isto é, nem todos os compostos com grupo metoxila nessa posição são ativos. Para obter informações complementares nessa linha, a segunda variável, **X4392**, é analisada. Essa variável contém valores de pixel para hidrogênio (210) nas moléculas 1-2, 12-15, 18, valores de pixel para oxigênio (350) nas moléculas 3-11, 17, e um valor de pixel para o átomo de carbono (250) para a molécula 16. Um exame desse padrão indica que os pixels da variável **X4392** correspondem aos substituintes da posição R5. Este resultado sugere que a atividade antiesquistossomótica é determinada, principalmente, por substituintes nas posições R1 (ou R2) e R5, onde os grupos metoxila em R1 (ou R2) e R5 produzem compostos ativos,

enquanto que a substituição de qualquer um dos dois elimina a atividade antiesquistossomótica. Em outras palavras, não é imperativo, para neolignanas, possuir grupos metoxila em ambas as posições R1 e R2 para garantir a bioatividade pretendida. Enquanto estudos experimentais anteriores têm enfatizado a importância de grupos metoxila no anel B para atividade biológica das neolignanas (SANTOS, 1991), o descarte de um dos grupos metoxila representa uma novidade do presente estudo.

4 CONCLUSÃO

Este trabalho apresenta modelos não supervisionados (PCA e HCA) e supervisionados (PLS-DA) para avaliar a atividade antiesquistossomótica de derivados de neolignanas com base na abordagem MIA-SAR. O estudo revela que a incorporação de esquemas de cores nas imagens das estruturas moleculares produz um desempenho superior ao de imagens binárias (preto e branco). Além disso, a interpretação da influência de diferentes substituintes na atividade antiesquistossomótica destaca a importância de substituintes metoxila nas posições R1 (ou R2) e R5 do esqueleto das neolignanas na determinação da sua bioatividade. Os resultados obtidos fornecem conhecimento crítico sobre as possíveis modificações estruturais que podem ser realizadas para produzir novos compostos derivados de neolignanas com atividade antiesquistossomótica.

REFERÊNCIAS

ADVANCED CHEMISTRY DEVELOPMENT. **ACD/ChemSketch - version 12.01**. Toronto, Advanced Chemistry Development, 2009.

Disponível em: <http://alainsheratan.com/chemsk_tl2.pdf>. Acesso em: 10/02/2013.

ALVES, C. N. et al. A structure-activity relationship (SAR) study of neolignan compounds with anti-schistosomiasis activity. **Journal of the Brazilian Chemical Society**, São Paulo, v. 13, n. 3, p. 300-307, May 2002.

APERS, S.; VLIETINCK, A.; PIETERS, L. Lignans and neolignans as lead compounds. **Phytochemistry Rev**, Dordrecht, v. 2, n. 3, p. 201-217, Jan. 2003.

BARATA, L. E. et al. Anti-leishmanial activity of neolignans from *Virola* species and synthetic analogues. **Phytochemistry**, Oxford, v. 55, n. 6, p. 589-595, Nov. 2000.

BRASIL, Ministério da Saúde. **Doenças infecciosas e parasitárias: Guia de Bolso**: 3 ed. Brasília, Ed. MS, Vol. II, 2004, 334p.

CENTERS FOR DISEASE CONTROL AND PREVENTION-CDC. **Parasites - Schistosomiasis**. 2002. Disponível em: <<http://www.cdc.gov/parasites/schistosomiasis/disease.html>>. Acessado em 15/10/2015

CHEVALLIER, S. et al. Application of PLS-DA in multivariate image analysis. **Journal. Chemometrics**, Hoboken, v. 20, n. 5, p. 221-229, Nov. 2006.

CHATTERJEE, S. et al. Does the neuropeptide somatostatin have therapeutic potential against schistosomiasis? **Trends in Parasitology**, Oxford, v. 18, n. 7, p. 295-298, July 2002.

CIOLI, D.; PICA-MATTOCCIA, L. Praziquantel. **Parasitology Research**, New York, v. 90, n. 1, p. S3-S9, Nov. 2003.

SOUZA, A. O. et al. Structure activity relationship, acute toxicity and cytotoxicity of antimycobacterial neolignan analogues. **Journal of Pharmacy and Pharmacology**, Malden, v. 63, n. 7, p. 936-942, July 2011.

DENNINGTON, R. D.; KEITH, T.A; MILLAM, J.M. **GaussView 5.0. 8**; Gaussian Inc., Wallingford, 2008. CD -ROM.

MCMANUS, D. P.; LOUKAS, A. Current status of vaccines for schistosomiasis. **Clinical Microbiology Reviews**, Washington, v. 21, n. 1, p. 225-42, Jan 2008.

NOCITO, I. et al. Activity of 8. O. 4'-neolignans against *Trypanosoma cruzi*. **Parasitology Research**, New York, v. 101, n. 5, p. 1453-1457, Nov. 2007.

NUNES, C. A. et al. Chemoface: a novel free user-friendly interface for chemometrics. **Journal of the Brazilian Chemical Society**, São Paulo, v. 23, p. 2003-2010, Nov. 2012.

ORGANIZAÇÃO MUNDIAL DE SAÚDE-OMS. Schistosomiasis: number of people receiving preventive chemotherapy in 2012. **Weekly epidemiological record**, Geneva, v. 89, p. 21–28, Jan. 2014. Disponível em: <<http://www.who.int/wer/2014/wer8902.pdf?ua=1>>. Acesso em: 17/02/2015

ORGANIZAÇÃO MUNDIAL DE SAÚDE- OMS. **Investing to overcome the global impact of neglected tropical diseases: third WHO report on neglected diseases**. Geneva: Organização Mundial de Saúde, 2015, 191p. Disponível em: <<http://apps.who.int/bookorders/anglais/detart1.jsp?codlan=1&codcol=15&codcch=3853>>. Acesso em: 15/06/15

ORGANIZAÇÃO MUNDIAL DE SAÚDE-OMS. **Schistosomiasis: number of people treated worldwide in 2014**. Geneva, Vol. 91, n.5, p. 53-60, Feb. 2016. Disponível em: <<http://www.who.int/wer/2016/wer9105.pdf>>. Acesso em: 15/03/2016.

SABAH, A. A. et al. *Schistosoma mansoni*: chemotherapy of infections of different ages. **Experimental parasitology**, San Diego, v. 61, n. 3, p. 294-303, June 1986.

SANTOS, L. D. S. **Síntese e atividade biológica de neolignanas 8.O.' derivados e compostos correlacionados**. 1991. 419p. Tese (Doutorado em ciências). Instituto de Química, Universidade Estadual de Campinas, Campinas, 1991.

WANG, W.; WANG, L.; LIANG, Y.-S. Susceptibility or resistance of praziquantel in human schistosomiasis: a review. **Parasitology Research**, New York, v. 111, n. 5, p. 1871-1877, Nov. 2012.

ZACCHINO, S. A. et al. In vitro evaluation of antifungal properties of phenylpropanoids and related compounds acting against dermatophytes. **Journal of Natural Products**, Washington, v. 62, n. 10, p. 1353-1357, Oct. 1999.

CAPÍTULO 6: CONSIDERAÇÕES FINAIS

Este trabalho contribuiu para que o método MIA-QSAR se elevasse a um patamar não imaginado quando do seu surgimento, há mais de 10 anos: o patamar da interpretabilidade, que era um problema recorrentemente apontado por revisores e avaliadores dos trabalhos que usavam o método MIA-QSAR.

O uso de círculos para representar átomos, com tamanhos proporcionais aos respectivos raios de van der Waals e cores sólidas com valores de pixels proporcionais a alguma propriedade, tal como eletronegatividade, permitiu caracterizar quimicamente o aspecto estrutural que determinava a tendência nas atividades biológicas do conjunto de dados em estudo. Contribuindo assim, para o desenvolvimento de novos compostos com maior potencial pretendido.

Os compostos escolhidos estiveram relacionados com doenças tropicais (dengue, doença de Chagas, malária e esquistossomose), pelo fato de que estas representam um grande entrave socioeconômico aos países em desenvolvimento e, ainda assim, recebem pouca atenção da "big pharma" e das políticas públicas governamentais para o desenvolvimento de tratamentos mais eficientes e acessíveis. Assim, o uso de metodologias como o MIA-QSAR na pesquisa e desenvolvimento de novos medicamentos pode resultar em um processo mais rápido e barato.

Novas estruturas químicas potencialmente ativas contra dengue, doença de Chagas, malária e esquistossomose puderam ser propostas a partir da racionalização das subestruturas responsáveis por um aumento ou diminuição da atividade graças às implementações, aprimoramentos e inovações no método MIA-QSAR outrora desenvolvido. Contudo, novas análises em relação aos parâmetros farmacológicos e efeitos toxicológicos podem ser realizados para posterior síntese e avaliação em *in vivo*.