Estevão Bombonato Pereira

Estudos de QSAR de derivados de benzilfenildiaminoeter,
promissores para o tratamento da Tripanossomíase Humana Africana
causada pelo *Tripanossoma Brucei Rhondesiense*

Tese apresentada ao Instituto de Química de
São Carlos da Universidade de São Paulo
como parte dos requisitos para obtenção do
título de Doutor em Ciências.

Área de Concentração: Físico-Química

Orientador: Prof^o Dr. Albérico Borges Ferreira
da Silva.

*Feliz o homem que acha sabedoria, e o homem que adquire conhecimento; porque melhor é o lucro que ela dá do que o da prata, e melhor a sua renda do que o ouro mais fino.
(Provérbios 3:13-14)*

Agradecimentos

Primeiramente dou graças a Deus por tudo o que passei, aprendi e vivi nesse período do doutorado e por conseguir concluir essa etapa.

Agradeço aos meus pais, Carlos e Elaine, pois sempre me apoiaram e me encorajaram a estudar desde a infância. Sempre investiram nos meus estudos e deram todo suporte para que pudesse concretizar esse sonho. Também às minhas irmãs, Lídia e Júlia, que sempre me deram apoio, encorajamento e a descontração para aliviar os problemas.

Agradeço à minha esposa Ruhamah, minha grande companheira, que sempre está ao meu lado para me ajudar, me encorajar e não me deixar desistir. Por topar a loucura dos concursos e das mudanças para seguir na carreira de professor e pesquisador.

Agradeço aos parceiros do laboratório, Aline, Régis, Natieli, Amanda, Júlia, Célio, Eriosvaldo, Luiz, Rafael e Guilherme, que sempre contribuíram para enriquecer meu conhecimento e melhorar meu estudo. Também pelos momentos de descontração, para aliviar um pouco a tensão da pesquisa.

Agradeço ao meu orientador Albérico, por ter aceitado me orientar desde o mestrado, acreditar no meu trabalho e prover o que fosse necessário para que eu continuasse o doutorado.

Muito obrigado a todos vocês.

Resumo

A Tripanossomíase Humana Africana, ou doença do sono, é uma doença negligenciada que assola alguns países da África, principalmente na região centro-sul do continente. Essa doença é causada pelo tripanossoma *brucei rhondesiense*, que é transmitido pela mosca silvestre Tsé-Tsé. Há projetos para combate à mosca para evitar a disseminação da doença, mas ainda se faz necessário o estudo de fármacos para combate ao tripanossoma. Com vistas a ajudar nesta busca por fármacos que sejam úteis no combate ao tripanossoma, estudamos um conjunto de derivados de benzofenildiaminoéteres desenvolvidos e testados por Patrick e colaboradores para avaliar quais seriam as características mais importantes para que as moléculas tenham uma melhor atividade contra o tripanossoma.

Ao avaliarmos as propriedades das moléculas e a relação com a atividade biológica, selecionamos como descritores a energia do orbital HOMO, o número de aminas primárias alifáticas ($n\text{NH}_2$), fator hidrofílico (Hy), o índice de primeira ordem da média de carga topológica (JGI1), a autocorrelação de Moran na *lag-1* ponderada pela eletronegatividade de Sanderson (MATS1e), e autocorrelação de Geary na *lag-7* ponderada pela eletronegatividade de Sanderson (GATS7e).

Para o estudo SAR utilizamos os métodos de classificação análise hierárquica de grupos (HCA), análise de componentes principais (PCA), análise dos k-ésimos vizinhos mais próximos (KNN) e SIMCA. Para o estudo QSAR, utilizamos o método de análise de mínimos quadrados parciais (PLS) e redes neurais artificiais (ANN).

Todos os métodos utilizados foram testados e validados, quando cabível. Os resultados obtidos em cada método apresentam uma boa relação entre si, mostrando que o uso de vários métodos em conjunto torna o resultado mais confiável.

Abstract

African Human Trypanosomiasis, or sleeping sickness, is a neglected disease that plagues some African countries, especially in the center-south region of the continent. This disease is caused by the *Trypanosoma Brucei Rhodesiense*, which is transmitted by the wild fly Tse-Tsé. There are projects to combat the fly to avoid the spread of the disease, but it still requires the study of drugs to combat the trypanosoma. In order to aid in this search for drugs that are useful in the fight against trypanosome, we have studied a set of benzyl phenyl ether diamine derivatives developed and tested by Patrick and his group to evaluate the most important characteristics for the molecules to have a better activity against the trypanosome.

When evaluating the properties of the molecules and the relation with the biological activity, we selected the HOMO orbital energy, the number of aliphatic primary amines (nNH₂), hydrophilic factor (Hy), the first order index of the topological load average JGI₁), the Moran autocorrelation at lag-1 weighted by Sanderson's electronegativity (MATS_{1e}), and Geary's autocorrelation at lag-7 weighted by Sanderson's electronegativity (GATS_{7e}).

For the SAR study we used the methods of classification hierarchical cluster analysis (HCA), principal component analysis (PCA), analysis of k-nearest neighbor (KNN) and SIMCA. For the QSAR study we used the method of analysis of partial least squares (PLS) and artificial neural networks (ANN).

All methods used were tested and validated when appropriate. The results obtained in each method show a good relation between them, showing that the use of several methods together makes the result more reliable.

Lista de Figuras

Figura 1. Mapa do número de casos reportados de Tripanossomíase Humana Africana nos países do continente africano.	4
Figura 2. Temos à esquerda o tripanossoma Brucei, causador da doença do sono, e à direita a mosca Glossina ou Tsé-Tsé, transmissora do tripanossoma.	5
Figura 3. Decaimento do número de casos de THA com o passar dos anos devido à organização de ações para combate ao vetor e prevenção da doença.	7
Figura 4. Ilustração da abordagem centroide do método HCA aplicada a um conjunto de amostras hipotéticas.	11
Figura 5. Ilustração da abordagem centroide do método HCA aplicada a um conjunto de amostras hipotéticas.	11
Figura 6. Modelo matemático da estrutura de um perceptron de redes neurais artificiais.	19
Figura 7. Estrutura molecular base, da qual partem todas as moléculas envolvidas no presente estudo.	23
Figura 8. Estrutura dos principais substituintes utilizados nas posições R ₁ e R ₄ das moléculas em estudo e suas abreviações.	23
Figura 9. Gráfico de avaliação da atividade biológica para proposição de grupos.	28
Figura 10. Separação dos grupos obtida pelo método PCA utilizando os 60 moléculas.	30
Figura 11. Ilustração do efeito de ressonância estima-se ocorrer no substituinte Am.	38
Figura 12. Substituinte Am da posição R ₁ da molécula M49, com as cargas de cada átomo.	39
Figura 13. Substituinte AmOH da posição R ₁ da molécula M50, com as cargas de cada átomo.	39
Figura 14. Substituinte AmOMe da posição R ₁ da molécula M51, com as cargas de cada átomo.	40
Figura 15. Análise HCA utilizando a abordagem single para agrupamento das moléculas pelo grau de similaridade.	46
Figura 16. Análise HCA utilizando a abordagem centroide para agrupamento das moléculas pelo grau de similaridade.	47
Figura 17. Análise pelo método HCA utilizando centralização na média para os dados e abordagem single para ligação.	49
Figura 18. Análise pelo método HCA utilizando centralização na média para os dados e abordagem centróide para ligação.	49
Figura 19. Análise do conjunto completo de moléculas em estudo pelo método ANN.	53
Figura 20. Análise do conjunto de treinamento para gerar um modelo de regressão por ANN e ensinar a rede como analisar os dados fornecidos.	53
Figura 21. Análise do grupo de moléculas utilizadas para validação interna do método de regressão gerado por ANN e verificar se a rede aprendeu a avaliar os dados fornecidos.	54
Figura 22. Avaliação do conjunto de teste, moléculas que não foram utilizadas para geração do modelo de regressão por ANN, para avaliar se a rede realmente aprendeu a avaliar os resultados ou apenas decorou os resultados que deveriam ser reportados.	54
Figura 23. Gráfico de Scores obtido utilizando o conjunto de treinamento, onde podemos visualizar a disposição das moléculas de acordo com os dados reportados pelas componentes principais (PC) 1 e 2, e visualizar o agrupamento das moléculas.	57
Figura 24. Gráfico de Loadings, onde é reportada a contribuição de cada descritor para a formação das componentes principais 1 e 2.	58

Figura 25. Gráfico de Scores obtido para o conjunto de testes, onde podemos observar a disposição das moléculas avaliadas pelo modelo PCA gerado pelo conjunto de treinamento.....	59
Figura 26. Gráfico obtido dos Scores do conjunto de treinamento com as PCs 1 e 2, usando centralização na média para tratamento dos dados.	60
Figura 27. Gráfico de Loadings dos descritores para as componentes principais 1 e 2.....	60
Figura 28. Gráfico de Scores obtido pela aplicação do modelo ao conjunto de testes, com tratamento dos dados por centralização na média.....	61
Figura 29. Separação em classes usando processamento de autoescalamento.	63
Figura 30. Predição de classe para o conjunto de teste usando processamento de autoescalamento.	64
Figura 31. Separação das moléculas em classes, de acordo com a votação dos vizinhos mais próximos com processamento centralizado na média.	65
Figura 32. Aplicação do modelo KNN com processamento centralizado na média às moléculas do conjunto de testes.....	66
Figura 33. Gráfico de projeção das classes com método SIMCA com autoescalamento, obtido pelo programa Pirouette.....	68
Figura 34. Gráfico de distância entre as classes, classificadas pelo método SIMCA com autoescalamento.....	69
Figura 35. Gráfico de detecção de outliers das moléculas menos ativas pelo método SIMCA com autoescalamento.....	70
Figura 36. Gráfico de detecção de outliers das moléculas mais ativas pelo método SIMCA com autoescalamento.....	70
Figura 37. Gráfico de distância entre as classes do método SIMCA com autoescalamento aplicado ao conjunto de teste.	71
Figura 38. Gráfico de probabilidade de classificação das moléculas do conjunto de teste em classes.....	71
Figura 39. Gráfico de projeção das classes com método SIMCA com centralização na média, obtido pelo programa Pirouette.....	72
Figura 40. Gráfico de distância entre as classes, classificadas pelo método SIMCA com centralização na média.....	74
Figura 41. Gráfico de detecção de outliers para as moléculas mais ativas pelo método SIMCA com centralização na média.	74
Figura 42. Gráfico de detecção de outliers para as moléculas menos ativas pelo método SIMCA com centralização na média.	75
Figura 43. Gráfico de distância entre as classes pelo método SIMCA com centralização na média, aplicado ao conjunto de teste.....	76
Figura 44. Gráfico da probabilidade de predição de classes para as moléculas do conjunto de testes pelo método SIMCA com centralização na média.	76
Figura 45. Gráfico do modelo PLS gerado utilizando as moléculas do conjunto de treinamento.....	77
Figura 46. Análise das moléculas do conjunto de treinamento utilizadas para obter o modelo PLS para detecção de Outliers por relação entre o Resíduo de Student e a Alavancagem.....	79
Figura 47. Aplicação do modelo de regressão gerado por PLS ao conjunto de teste.....	80
Figura 48. Teste de robustez realizado por meio da retirada de moléculas do conjunto de treinamento utilizado para gerar o modelo PLS.	82
Figura 49. Gráfico de resultados obtidos fazendo a aleatorização do valor de atividade biológica das moléculas do conjunto de treinamento para verificação da validade do modelo.	83

Lista de Tabelas

Tabela 1. Nomeação das moléculas em estudo, seus substituintes e atividade biológica.....	24
Tabela 2. Correlação entre os descritores, e entre descritores e a atividade biológica.	31
Tabela 3. Descritores selecionados para o estudo SAR e QSAR em ordem decrescente de atividade biológica (plC_{50}).	32
Tabela 4. Média dos valores dos descritores para o grupo de moléculas ativas e para o grupo de moléculas menos ativas.	34
Tabela 5. Agrupamento de moléculas similares, separados de acordo com os substituintes das posições R_1 e R_2 para avaliação de influências dos substituintes sobre a atividade biológica e descritores.....	37
Tabela 6. Influência dos substituintes sobre o valor dos descritores em ordem crescente dos valores correspondentes a cada substituinte, e qual valor contribui para que a atividade biológica seja melhor.	44
Tabela 7. Discriminação das moléculas pertencentes a cada grupo formado pelo método HCA.	46
Tabela 8. Agrupamento das moléculas realizado pelo método HCA com tratamento dos dados por centralização na média.	48
Tabela 9. Variância dos descritores após o tratamento por autoescalamento e centralização na média.....	50
Tabela 10. Poder discriminante e poder total de modelagem dos descritores para o modelo SIMCA com autoescalamento gerado utilizando o conjunto de treinamento.	69
Tabela 11. Poder discriminante e poder total de modelagem dos descritores para o modelo SIMCA com centralização na média gerado utilizando o conjunto de treinamento.	73
Tabela 12. Resultados da validação do modelo PLS do conjunto de treinamento.	78
Tabela 13. Resultados da aplicação do modelo PLS no conjunto de teste.	80

Lista de Abreviaturas

ANN – Artificial Neural Network (Redes Neurais Artificiais)

DFT – Density Functional Theory (Teoria do Funcinal de Densidade)

HCA – Hierarchical Clustering Analysis (Análise Hierárquica de Grupos)

IC₅₀ – Inibição do Crescimento em 50%

KNN – K-Nearest Neighborhoods (K-ésimo vizinho)

PCA – Principal Component Analysis (Análise de Componentes Principais)

PC – Principal Component (Componente Principal)

PLS – Partial Least Squares (Regressão por Mínimos Quadrados Parciais)

QSAR – Quantitative Structure/Activity Relationship (Relação Quantitativa Estrutura/Atividade)

SAR – Structure/Activity Relationship (Relação Estrutura/Atividade)

SIMCA – Soft Independent Modeling by Class Analogy

THA – Tripanossomíase Humana Africana

Sumário

1.	Introdução.....	1
1.1.	Doenças Negligenciadas.....	1
1.2.	Tripanossomíase Humana Africana.....	3
1.3.	Métodos Computacionais	8
1.3.1.	Análise de Componentes Principais (<i>Principal Component Analysis – PCA</i>).....	8
1.3.2.	Análise Hierárquica de Grupos (<i>Hierarchical Clusters Analysis – HCA</i>)	9
1.3.3.	Método dos K Vizinhos mais Próximos (<i>K-Nearest Neighbors – KNN</i>).....	12
1.3.4.	SIMCA (<i>Soft Independent Modeling by Class Analogy</i>).....	13
1.3.5.	Mínimos Quadrados Parciais (<i>Partial Least Squares – PLS</i>)	14
1.3.5.1.	Métodos de Validação Interna	15
1.3.5.2.	Detecção de Amostras Atípicas.....	16
1.3.5.3.	Teste de Robustez	17
1.3.5.4.	Teste de Correlação ao Acaso	17
1.3.6.	Redes Neurais Artificiais (<i>Artificial Neural Network – ANN</i>)	17
1.4.	Métodos Teóricos de Cálculo	20
2.	Objetivos	22
3.	Metodologia	23
4.	Resultados e Discussão	28
4.1.	Seleção de descritores	28
4.2.	Influência dos substituintes na atividade biológica	36
4.3.	Análise HCA	45
4.3.1.	Com autoescalamento	45
4.3.2.	Com centralização na média	47
4.4.	Análise ANN.....	51
4.5.	Análise PCA.....	56
4.5.1.	Com autoescalamento	56
4.5.2.	Com centralização na média	59
4.6.	Análise KNN	62
4.6.1.	Com autoescalamento	62
4.6.2.	Com centralização na média	64
4.7.	SIMCA.....	67
4.8.	Com Autoescalamento	67

4.9. Com Centralização na Média	72
4.10. Análise PLS.....	77
4.10.1. Teste de robustez do modelo PLS	81
4.10.2. Teste de correlação ao acaso do modelo PLS	82
5. Conclusão	84
Bibliografia	86

1. Introdução

1.1. Doenças Negligenciadas

A história da humanidade é permeada de vários contos, mitos e lendas sobre formas miraculosas de se obter a juventude eterna, a cura para todas as doenças ou a imortalidade. Na idade média, o principal objetivo dos alquimistas era obter o elixir da vida, algo que pudesse tornar o homem imortal e, de alguma maneira, transmutar qualquer metal em ouro. Isto foi a principal impulsão da ciência na época. Também havia muitos contos sobre a fonte da juventude e a cidade de “El dourado” na América espanhola, que era toda feita de ouro, que levaram muitos navegadores a se aventurarem pelos mares.

Os anos passaram e vemos uma grande corrida hoje para manter padrões de beleza, evitar doenças, melhorar a qualidade de vida de pessoas idosas e diminuir custos e aumentar o lucro. Durante os intervalos dos programas de televisão vemos muitas propagandas de produtos cosméticos, clínicas de estética ou equipamentos que ajudarão a tornar o seu corpo mais saudável e bonito. Para que exista essa gama de produtos e serviços, muito é investido para alcançar esses objetivos.

Então podemos perceber que o foco da sociedade, desde o início, é saúde, beleza e dinheiro. Não que isso seja algo ruim, mas é o que movimenta todo o sistema de vida atual. Os investimentos em pesquisa e desenvolvimento nestas áreas são muito altos e percebemos que têm sido efetivos. A longevidade da população mundial tem aumentado. As mortes por doenças têm diminuído, quando comparado com os tempos da idade média em que uma simples gripe poderia matar. Aos poucos os mitos se tornam reais, as pessoas estão conseguindo manter a aparência jovem por mais tempo e vivendo mais. Só a fórmula mágica para ficar rico é que ainda não foi descoberta.

No entanto, o que por um lado é muito bom para um grupo de pessoas que pode pagar por tratamentos para ter uma boa saúde, por outro lado é desvantajoso para quem não dispõem de recursos para custear estes tratamentos. Pouco investimento há das grandes empresas em medicamentos para o tratamento de doenças endêmicas que afetam

uma população de uma região carente de recursos financeiros. Visto que esta população não teria recursos para custear o tratamento e cobrir todo o gasto empregado no desenvolvimento dos medicamentos. Estas doenças são classificadas como doenças negligenciadas, doenças que nem mesmo são veiculadas nos principais meios de comunicação.

Principalmente no continente africano, na Ásia e nas Américas do sul e central, temos várias doenças endêmicas, ou doenças tropicais, que são negligenciadas pelas grandes empresas farmacêuticas. Algumas destas doenças são a dengue, malária, leishmaniose, doença de chagas, esquistossomose e a doença do sono (Tripanossomíase Humana Africana). Segundo Ricardo Valverde, de toda a carga global de doenças, 11,4% são doenças tropicais e tuberculose, mas, entre os anos de 1954 e 2004, apenas 21 medicamentos foram registrados para o tratamento destas e outros 1556 medicamentos foram registrados para tratar outras doenças (Valverde).

O fato de estas doenças serem negligenciadas acaba agravando as diferenças sociais. Cabe então ao governo, no Brasil o Ministério da Saúde, suprir esta diferença propondo planos de ação para financiamento e desenvolvimento de novos medicamentos (Ministério da Saúde, Departamento de Ciência e Tecnologia, Secretaria de Ciência, Tecnologia e Insumos Estratégicos, 2010). Também organizações não governamentais sem fins lucrativos têm investido em pesquisa para desenvolvimento de medicamentos e também em novos tratamentos para estas doenças. A iniciativa Medicamentos para Doenças Negligenciadas (DNDi – Drugs for Neglected Diseases initiative) trabalha nesta frente de desenvolvimento e de medicamentos e projetos de prevenção de doenças negligenciadas de maneira local, em parceria também com o projeto Médicos Sem Fronteiras (Drugs for Neglected Diseases initiative).

Ao ver a necessidade de pesquisa para o desenvolvimento de novos medicamentos para tratar doenças negligenciadas, foi escolhido um conjunto de moléculas desenvolvidas para o tratamento da Tripanossomíase Humana Africana. Visto que as condições sociais são mais agravadas que as que presenciamos no Brasil e que várias instituições como a FIOCRUZ e Universidades públicas desenvolvem pesquisas voltadas para endemias presentes no

Brasil, foi escolhido desenvolver um estudo para auxiliar no desenvolvimento de novos medicamentos para a Tripanossomíase Humana Africana.

1.2. Tripanossomíase Humana Africana

A Tripanossomíase Humana Africana (THA) foi combatida fortemente nos anos 50, mas após uma breve queda do número de casos, voltou a crescer. Devido à baixa notificação de casos de THA, ocorre a falsa impressão de que a doença está controlada, mas infelizmente algo mais grave ocorre. Com um sistema de informação precário para a área de saúde, as notificações dos casos e diagnósticos ficam mais difíceis.

As poucas informações passadas às pessoas sobre os riscos da THA e os sintomas causados por ela, fazem com que o diagnóstico seja, na maioria das vezes, feito já na fase avançada da doença. Por falta de informação, algumas pessoas das populações rurais recorrem a curandeiros, tratamento com ervas ou mesmo tratamento espiritual da doença, sem saber do que realmente se trata. Devido a isso, o número de casos fatais da doença é alto, tanto para a crônica como para a aguda.

O número de casos não informados ao sistema de saúde é estimado em 30% do número de casos registrados. Isso se deve à falta de comparecimento das pessoas aos hospitais ou postos de saúde para tratamento, como também o diagnóstico tardio ou confusão da THA com algum estágio da malária.

Estudos indicam que a THA é uma doença endêmica e que o aparecimento em uma região se deve à migração de pessoas ou gado de uma região já contaminada. Assim, foi constatado que a contaminação das cidades ao redor do lago Kyoga, ocorreu pela migração de gado para reprodução e comércio, e de pessoas. (Acup, Bardosh, Picozzi, Waiswa, & Welburn, 2016)

A região com maior incidência de casos de THA é a região da África Oriental e Sudeste da África, como mostrado na figura 1, que reporta o número de casos informados em cada país. Dos países destacados no mapa da figura 1, cinco estão entre os quinze mais pobres do mundo em 2015 e os outros apresentam valores muito baixos de renda per capita. Isso

mostra como a pobreza pode ser um dos fatores que influenciam a propagação de doenças endêmicas (Caleiro, 2015).

Distribuição da Tripanossomíase Humana Africana (*T. b. rhodesiense*)

Figura 1. Mapa do número de casos reportados de Tripanossomíase Humana Africana nos países do continente africano.

*(Figura adaptada de http://gamapserver.who.int/mapLibrary/Files/Maps/HAT_ga_2014.png).

Outra característica da doença é a sua predominância em regiões rurais e suburbanas. Esta limitação existe, pois o transmissor da doença, mosca *Glossina* ou Tsé-Tsé (figura 2), é uma mosca pesada e que não voa longas distâncias. Assim, uma das principais formas de contenção da doença é o confinamento da mosca transmissora em uma região afastada dos centros urbanos e a prevenção contra a multiplicação da população de moscas.

Figura 2. Temos à esquerda o tripanossoma Brucei, causador da doença do sono, e à direita a mosca Glossina ou Tsé-Tsé, transmissora do tripanossoma.

O causador da THA é o *Tripanossoma Brucei* (figura 2), existindo três tipos, o *Tripanossoma Brucei Brucei* (infecta apenas animais), *Tripanossoma Brucei Gambiense* (causa a doença em uma forma mais branda e que pode permanecer meses ou anos assintomática, presente principalmente na região da África Ocidental) e o *Tripanossoma Brucei Rhondosiense* (causador da doença na forma mais rápida e aguda, que se manifesta em poucas semanas e rapidamente invade o sistema nervoso central). O tempo entre a infecção e a morte leva, aproximadamente, 3 anos para T.b. Gambiense e 6 meses para T.b. Rhondosiense (World Health Organization, 2010).

A detecção do tripanossoma é difícil, pois os métodos de detecção tem baixa sensibilidade e a quantidade de tripanossomas em circulação é baixa, mas estando eles presentes em todos os fluidos corporais. Como os parasitas estão presentes em todos os fluidos corporais, algumas crianças já nascem com o parasita, caso a mãe gestante já esteja contaminada (World Health Organization, 2010).

Os sintomas geralmente apresentados na contaminação são a lesão causada pela inoculação, febre intermitente, linfadenopatias, cefaleias, sonolência e coma. Os três últimos sintomas já são um estágio avançado da doença, no qual o tripanossoma já se encontra alojado no sistema nervoso central. A doença também é conhecida como doença do sono devido a um dos sintomas ser a sonolência e também a inversão do relógio do sono, permanecendo a pessoa acordada a noite e dormindo de dia (WORLD HEALTH ORGANIZATION).

Sabendo que os principais métodos de combate à THA é o combate ao agente transmissor e o tratamento das pessoas infectadas pelo tripanossoma, alguns grupos de pesquisa estudam moléculas para combater esta doença (Athri, Wenzler, Ruiz, Brun, &

Boykin, 2006) (Khabnadideh, et al., 2005) (Zubrzycki, 2002) (Patrick, et al., 2013). Quanto maior o número de estudos e desenvolvimento de moléculas para combater o tripanossoma, maiores serão as chances de obter a cura para a THA, pois mesmo que o tripanossoma adquira resistência a um determinado medicamento eficiente para o tratamento, será possível utilizar uma molécula diferente, tão eficaz quanto o primeiro, para combater o parasita.

Os quatro principais medicamentos utilizados no tratamento da doença do sono são a pentamidina, a suramina, o melarsoprol e a eflornitina. Estes são medicamentos perigosos, caros e desconfortáveis para os pacientes, causando efeitos colaterais. Um tratamento com Eflornitina custa, aproximadamente, 600 dólares. Destes quatro, apenas a suramina e o Melarsoprol tem efeito sobre o T.b. Rhondosiense (World Health Organization, 2010).

A OMS, junto com os países que apresentam casos da doença, tem a meta de erradicar a THA até 2030. Uma meta difícil, caso não sejam implementados mais esforços e recursos para melhora do sistema de saúde destes países. Há postos de tratamento, mas como o diagnóstico é demorado, o tratamento também atrasa, agravando o estágio da doença. Percebemos no gráfico da figura 3 um decaimento significativo no número de casos relatados com o passar dos anos, devido a ações de combate ao vetor e prevenção de contágio pela doença. A doença está quase erradicada, mas ainda há muitos casos surgindo, devido a precariedade das condições sanitárias e de veiculação das informações.

Figura 3. Decaimento do número de casos de THA com o passar dos anos devido à organização de ações para combate ao vetor e prevenção da doença.

*Fonte: <http://apps.who.int/gho/data/node.main.A1637?lang=en>.

1.3. Métodos Computacionais

1.3.1. Análise de Componentes Principais (*Principal Component Analysis – PCA*)

A Análise de Componentes Principais foi descrita em 1901 por Pearson e posteriormente Hotelling desenvolveu métodos computacionais de PCA para análise de estruturas de correlação (Kuang Hongyu, 2015). Trata-se de um método fatorial fundamentado na correlação entre as variáveis para criar uma componente principal. A correlação entre as variáveis torna possível a combinação linear das mesmas, com o mínimo de perda de informação, e utilizando um número menor de variáveis, que são chamadas então de fatores ou componentes principais (Principal Components – PC) (J.M. Bouroche, 1982) (Ferreira, 2002).

A primeira componente principal (PC1) descreve a máxima variância do conjunto analisado, e as demais componentes principais descrevem a variância residual do conjunto com relevância decrescente (PC1 > PC2 > PC3 > ...). As PCs são não correlacionadas e ortogonais entre si, ou seja, as informações presentes em uma não estão presentes nas demais (Ferreira, 2002).

Definidas as componentes principais, as amostras são projetadas no novo sistema de eixos, formado pelas PCs. Como cada PC descreve uma quantidade elevada de informações, é possível descrever quase todas as informações das variáveis em poucas PCs. Também não ocorre alteração na relação entre as amostras, mesmo com a transformação no novo sistema de eixos.

Para obter as componentes principais, decompomos a matriz de dados originais fornecidos, X, em duas matrizes, na matriz T de escores e na matriz L de pesos.

$$X = TL = T_A L_A^T + E$$

sabendo que,

$$T_A = [t_1 \dots t_A] \quad e \quad L_A = [l_1 \dots l_A]$$

Ou ainda,

$$[X] = [t_1][l_1^T] + [t_2][l_2^T] + \dots + [t_A][l_A^T] + [E]$$

O número de componentes principais significativas é dado por A , ou seja, o número de PCs necessárias para descrever as informações mais importantes do conjunto de dados original, e E é a matriz de resíduos. A relação entre as amostras é expressa pelos valores dos escores e a relação entre as variáveis é expressa pelos valores dos pesos.

Analisando a equação temos que o vetor t_a da matriz de escores T é constituído das coordenadas de cada amostra na a -ésima nova variável, a PCa, enquanto a coluna l_a da matriz de pesos L contém as informações do quanto cada variável original contribuiu para a formação da PCa. Os pesos variam entre -1 e +1, pois são cosseno do ângulo da PCa com o eixo dos descritores originais. Assim, quanto maiores os valores dos pesos em módulo, mais próxima a PCa será de ser colinear com os descritores originais (Márcia M. C. Ferreira, 2011).

1.3.2. Análise Hierárquica de Grupos (Hierarchical Clusters Analysis – HCA)

Este método tem como objetivo encontrar agrupamentos e estruturas de conjuntos de dados, separando as amostras em grupos de acordo com a similaridade entre elas. O HCA é um método não supervisionado exploratório e multivariado, extremamente útil para propor estruturas de classificação ou reconhecimento de grupos, segundo um critério de homogeneidade (Regazzi, 2001). Este método também pode orientar o uso de métodos supervisionados de reconhecimento de padrão.

Os resultados do HCA são expressos em um dendograma ou árvore hierárquica, de maneira que o comprimento dos ramos da árvore expressa a similaridade entre as amostras. A disposição das amostras no dendograma também é de acordo com a similaridade entre elas, quanto maior a similaridade, mais próximas. Inicialmente é assumido que cada amostra individualmente é um grupo, para o grau de similaridade 1,0, o diminuir o critério de similaridade ocorre o aparecimento de grupos. A variação dos critérios de similaridade resulta em diferentes agrupamentos, e esses critérios variam de acordo com o objetivo do estudo (Malhotra, 2001) (Ferreira, 2002).

O cálculo da distância entre as amostras é dado pela equação (1), onde estamos calculando a distância d_{ab} entre as amostras a e b , num espaço de J dimensões. Sabendo que x_{ij} é elemento da matriz X , correspondente ao j -ésimo descritor do i -ésimo composto. Após o

cálculo são agrupadas as duas amostras mais próximas restando (I-1) grupos. O Cálculo é repetido até que as amostras formem um só grupo.

$$d_{ab} = \sqrt{(x_{a1} - x_{b1})^2 + (x_{a2} - x_{b2})^2 + \dots + (x_{aJ} - x_{bJ})^2}$$

Uma das maneiras de se agrupar as amostras seria utilizar a menor ou maior diferença entre duas amostras de dois grupos diferentes ou a menor distância entre os centroides dos dois grupos. Estas distâncias são transformadas em uma matriz de similaridade S, em que seus elementos são índices de similaridade. Se tomarmos duas amostras a e b, e substituirmos d_{ab} na equação 2, obtermos o índice de similaridade para estas amostras. Sabendo que s_{ab} é elemento da matriz S e $d_{\max}$ é a distância máxima entre duas amostras dos dois conjuntos.

$$s_{ab} = 1,0 - \frac{d_{ab}}{d_{\max}}$$

Caso a e b sejam muito próximos ou iguais, $d_{ab} \approx 0$ então $s_{ab} \approx 1$; se a e b forem muito distantes $d_{ab} \approx d_{\max}$ então $s_{ab} \approx 0$ (Márcia M. C. Ferreira, 2011).

O método HCA trata-se de um método aglomerativo que acrescenta as amostras uma a uma ao grupo, avaliando sempre o grau de similaridade entre a amostra isolada e as amostras do grupo (Valentin, 2000). Duas são as abordagens principais para o agrupamento de amostras, centroide e single. A abordagem centroide foi relatada acima, em que a formação dos grupos é iniciada com as duas amostras mais próximas e depois a amostra mais próxima ao ponto médio entre as duas iniciais, e a aglomeração de mais amostras vai ocorrendo de acordo com a proximidade ao ponto médio do grupo formado. A abordagem single se dá pela distância simples entre duas amostras, iniciando pelo par com menor distância e agrupando a amostra mais próxima a uma das amostras do grupo. As figuras 4 e 5 mostram como seria progressão de cada abordagem agrupando amostras sucessivamente.

Figura 4. Ilustração da abordagem centroide do método HCA aplicada a um conjunto de amostras hipotéticas.

Figura 5. Ilustração da abordagem centroide do método HCA aplicada a um conjunto de amostras hipotéticas.

A abordagem centroide é recomendada quando as amostras do conjunto em estudo apresenta boa separação em grupos, pois ao aplicar o método HCA a separação fica mais

nítida e facilita a interpretação do resultado e diagnósticos de grupos ou subgrupos. Já a abordagem single pode ser preferencialmente utilizada quando as amostras não apresentam boa separação em grupos (Kenneth R. Beeber, 1998).

Não é possível determinar qual é a melhor abordagem a ser utilizada sem conhecer o conjunto de amostras. Assim, é recomendável utilizar ambas e avaliar qual tem o melhor desempenho para o conjunto de amostras em estudo e utilizá-lo (Kenneth R. Beeber, 1998).

1.3.3. Método dos K Vizinhos mais Próximos (*K-Nearest Neighbors – KNN*)

O método KNN não leva em consideração a distribuição da população no espaço amostral. Nele as distâncias entre as amostras presentes no conjunto de treinamento são calculadas no espaço J-dimensional gerado pelos descritores utilizados. A classificação de um composto em uma determinada classe se dá pelos votos dos vizinhos mais próximos. O número de k vizinhos mais próximos dão seus votos cada um votando em sua classe. A classe com o maior número de votos é a classe em que a amostra será classificada, mas caso ocorra um empate nos votos é atribuída a classe com menor distância acumulada (Ferreira, 2002).

Como se trata de um método bem simples, a escolha adequada de k (número de vizinhos) é vital para a boa classificação das amostras. Caso seja escolhido um número muito baixo ou um número muito alto, pode ocorrer uma classificação errônea. Caso k seja muito pequeno, não haverá uma boa representação dos vizinhos mais próximos e caso k seja muito grande, irá privilegiar o conjunto que tiver a maior população, há então de se escolher um número que seja representativo, mas não privilegie a maioria. Pela literatura os melhores valores para k seriam 3, 4 ou 5.

Antes de iniciar a análise é recomendado um pré-tratamento dos dados, para minimizar os efeitos de uma variável dominante. Dois dos métodos recomendados são o autoescalamento e a centralização na média. O autoescalamento consiste em subtrair a média dos valores da variável de cada elemento da coluna e dividir este resultado pelo desvio padrão, como mostrado abaixo. A centralização na média consiste em subtrair a

média da variável do valor para a amostra específica, como também mostrada abaixo (João Paulo A. Martins, 2013).

$$x_{ij(\text{autoescalado})} = \frac{x_{ij} - \bar{x}_j}{s_j};$$

$$x_{ij(\text{cent. med.})} = x_i - \bar{x}_{ij}$$

Para validação interna do modelo, é recomendado que seja utilizado o método de validação cruzada, deixando uma das amostras de cada vez, fora na geração do modelo. Com o modelo validade deve ser procedida a validação externa aplicando o modelo ao conjunto de teste (Márcia M. C. Ferreira, 2011).

1.3.4. SIMCA (Soft Independent Modeling by Class Analogy)

O modelo SIMCA é construído utilizando componentes principais para descrever cada classe. Neste método são formadas hipercaixas utilizando as PCs, e as hipercaixas que têm a dimensões iguais ao número de PCs utilizadas para descrever a classe de amostras. Cada classe é descrita por uma hipercaixa e um número determinado de PCs. Os limites das hipercaixas são definidos por um valor de confiança para determinar se a amostra pertence ou não à classe.

O poder discriminante do modelo está relacionado à não superposição das hipercaixas e à distância entre elas, quanto mais distantes, maior será o poder discriminante do modelo. Se ocorrer sobreposição das hipercaixas, o modelo não obteve uma separação efetiva entre as classes. Podemos também dizer que modelos com uma distância considerável entre as hipercaixas têm maior confiança para o resultado que modelos com baixa distância entre as hipercaixas (Sabin, Ferrão, & Furtado, 2004). A determinação das PCs envolvidas para descrever a classe é feita por validação cruzada.

No método SIMCA a classificação das amostras é realizada comparando a variância residual da amostra com a variância residual média das amostras da classe em que está agrupada. Assim, podemos avaliar a similaridade da amostra à classe e o ajuste do modelo. A amostra só é atribuída à classe com a qual tem alta similaridade. Se a amostra apresentar uma variância residual muito elevada para as classes presentes no modelo de treinamento, ela não será classificada, pois se trata de uma amostra atípica. Caso as hipercaixas das classes utilizadas no treinamento não apresentem uma boa separação, o método SIMCA

pode classificar uma amostra nas duas classes. Isso não ocorreria em outros métodos, como o KNN, que classificariam a amostra em apenas uma das classes (Gemperline, 2006).

1.3.5. Mínimos Quadrados Parciais (*Partial Least Squares – PLS*)

O método PLS é um método supervisionado que faz uso dos valores da atividade biológica das amostras nos cálculos de regressão. Colocando os descritores em uma matriz X e usando a atividade biológica das amostras como um vetor y, os cálculos são realizados de maneira que a matriz T (escores) possa estimar o vetor y.

Os métodos de validação do modelo para avaliar a possibilidade de predição são leave-N-out e validação cruzada. Um bom modelo tem como característica apresentar baixos valores de PRESS e de SEP com o menor número de variáveis latentes (Ferreira, 2002).

No esquema abaixo fica evidente a participação da atividade biológica no processo do cálculo para obtenção das variáveis latentes e então o modelo final. No primeiro cálculo é obtida a primeira variável latente, as demais são obtidas com a repetição do processo utilizando as matrizes já obtidas (Márcia M. C. Ferreira, 2011).

$$\left\{ \begin{array}{l} \text{Bloco } X \left\{ \begin{array}{ll} X = yw^T + E_X & (X^T = wy^T) \\ \hat{w} = X^T(y^T y)^{-1}y & (\text{pesos}) \\ \hat{w} = \frac{\hat{w}}{\sqrt{\hat{w}^T \hat{w}}} & (\text{normalização}) \\ t = X\hat{w} & (\text{escores}) \\ l = X^T t (t^T t)^{-1} & (\text{loadings}) \end{array} \right. \\ \text{Bloco } y \left\{ \begin{array}{l} \text{Relacionando os vetores } t \text{ e } y \\ q = (t^T t)^{-1} t^T y \end{array} \right. \\ \text{Atualização de } X \text{ e } y \left\{ \begin{array}{l} E = X - tl^T \\ f = y - qt \\ X = E \\ y = f \end{array} \right. \end{array} \right.$$

O vetor de regressão $\hat{b}$ pode ser obtido por meio da equação abaixo:

$$\hat{b} = W(L^T W)^{-1} q^T$$

1.3.5.1. Métodos de Validação Interna

Os métodos de validação são utilizados para verificar a eficiência dos modelos de regressão. Para a validação do método, a separação em dois grupos, sendo um para geração do modelo e outro para a validação externa, ou aplicação do modelo. Para que este conjunto de validação, ou conjunto de testes, tenha uma aceitabilidade maior, é recomendado o uso de 30% do número de compostos utilizado para a geração do modelo (Márcia M. C. Ferreira, 2011) .

Para validar o modelo, é necessário testá-lo utilizando apenas os compostos empregados na confecção do modelo, também chamado de conjunto de treinamento. Um método é a validação cruzada com *leave-one-out*, onde uma das amostras é deixada de fora na geração do modelo e é estimada sua atividade biológica. Cada um dos compostos do conjunto de treinamento é excluído uma única vez, então são calculados os respectivos erros de previsão e a soma do quadrado dos erros de previsão, PRESS. A equação (2) representa o cálculo de PRESS, sendo y_{iexp} a atividade biológica experimental e $\hat{y}_{ival}$ a atividade gerada pelo modelo sem o composto i .

$$PRESS = \sum (y_{iexp} - \hat{y}_{ival})^2 \quad (2)$$

Quando menor a diferença entre o valor experimental e o gerado pelo modelo, maior é a capacidade preditiva do modelo em relacionar os descritores com a atividade biológica. Logo, quanto menor o valor de PRESS, melhor será o modelo.

Outro parâmetro que pode ser utilizado para avaliar a validade do modelo é o erro padrão de validação, ou SEP_{val} . Ele consiste na avaliação da capacidade de predição do modelo, por meio do valor de PRESS e o número de amostras do conjunto treino (I), como mostra a equação (3).

$$SEP_{val} = \sqrt{\frac{PRESS}{I}} \quad (3)$$

Há também o erro padrão de calibração, SEP_{cal} , que é calculado quando todas os compostos do conjunto treino são utilizados na geração do modelo. O SEP_{cal} leva em consideração o $\hat{y}_{ical}$, (valor estimado da atividade biológica, obtida utilizando todos os

compostos no modelo) e os graus de liberdades do modelo, dado pelo número de variáveis latentes (A), ou fatores, e o número de compostos usado no modelo, como mostra a equação (4).

$$SEP_{cal} = \sqrt{\frac{\sum(y_{iexp} - \hat{y}_{ical})^2}{I-A-1}} \quad (4)$$

Utilizando os valores de Q^2 e R^2 é possível avaliar a qualidade da predição e do ajuste do modelo. Sendo Q^2 a variância explicada na validação cruzada ou na previsão do modelo, e R^2 a variância explicada no ajuste do modelo. Podem Q^2 e R^2 ser calculados pelas equações (5) e (6):

$$Q^2 = 1 - \frac{\sum(y_{iexp} - \hat{y}_{ival})^2}{\sum(y_{iexp} - y_{med})^2} \quad (5)$$

$$R^2 = 1 - \frac{\sum(y_{iexp} - \hat{y}_{ical})^2}{\sum(y_{iexp} - y_{med})^2} \quad (6)$$

sendo y_{med} o valor médio da atividade biológica experimental dos compostos. Os valores de Q^2 e R^2 remetem a um bom modelo quando $Q^2 > 0,5$ e $R^2 > 0,6$ (Márcia M. C. Ferreira, 2011).

1.3.5.2. Detecção de Amostras Atípicas

A determinação de amostras atípicas pode ser feita utilizando o resíduo de *Student* e a alavancagem das amostras. O resíduo de *Student* consiste no resíduo padronizado de cada uma das amostras do conjunto de treinamento, calculado durante a validação cruzada do modelo. A alavancagem é a medida da influência de cada uma das amostras no modelo. A melhor maneira de utilizar essas duas informações é em um gráfico com resíduo de *Student* pela alavancagem, com cada amostra sendo um ponto.

Para que uma amostra seja considerada típica, ela deve ter resíduo de *Student* abaixo de 2,0 e alavancagem menor que $3A/I$, sendo A o número de fatores e I o número de amostras no conjunto de treinamento. Caso uma amostra apresente baixa alavancagem e elevado resíduo de *Student*, é provável que tenha ocorrido algum problema na determinação experimental de sua atividade biológica. As demais amostras que forem identificadas com elevada alavancagem e elevado resíduo de *Student*, devem ser avaliadas. As amostras

identificadas e avaliadas como atípicas devem ser retiradas do modelo (Márcia M. C. Ferreira, 2011) .

1.3.5.3. Teste de Robustez

Esse teste consiste em avaliar a robustez do modelo, se ele sofre poucas variações com a retirada de mais compostos do conjunto treino. Para isso são gerados vários modelos com validação cruzada *leave-N-out*, variando o N de 1 até 20 ou 30% do número de compostos presentes no conjunto de treino.

A avaliação é mediante o valor de Q^2 de cada um dos modelos gerados. Se os modelos gerados tiverem valores de $Q^2 > 0,5$ e cada modelo tiver um desvio de Q^2 variando entre 0,05 e -0,05, o modelo pode ser considerado robusto (Márcia M. C. Ferreira, 2011).

1.3.5.4. Teste de Correlação ao Acaso

Para verificação de correlação ao acaso no modelo, é utilizada a aleatorização do vetor y , pois é possível que descritores artificiais estejam bem correlacionados com a atividade biológica que gerem um bom modelo, mas sem nenhum significado químico. Esse teste faz uma permutação aleatória da atividade biológica dos compostos do conjunto de treino e mantendo a disposição dos descritores inalterada. Assim, é esperado que o resultado gerado pelo novo modelo seja o pior possível, mas se forem gerados bons modelos, é sinal de que há correlação ao acaso e o modelo inicial não tem significado químico relevante.

Pela literatura, o uso de 10 aleatorizações do vetor y seriam o suficiente para demonstrar se no modelo há ou não correlação ao acaso. Os valores de Q^2 e R^2 obtidos nos modelos randomizados são comparados com o modelo inicialmente proposto. Para um modelo sem correlação ao acaso o esperado é que os valores aleatorizados de Q^2 e R^2 sejam inferiores a 0,3 e 0,4, respectivamente (Márcia M. C. Ferreira, 2011).

1.3.6. Redes Neurais Artificiais (Artificial Neural Network – ANN)

Em todo o sistema nervoso a principal célula responsável por receber e transmitir informações é o neurônio. Para processar e transmitir as informações, os neurônios utilizam

vários processos bioquímicos, como trocas iônicas, pulsos elétricos e liberação de neurotransmissores. Um único neurônio pode receber informações de vários outros neurônios e irá compilar as informações de entrada em uma informação de saída que irá gerar um resultado. Este ponto em que ocorre a interação entre os neurônios para transmissão de informações é chamado axônio, nele ocorrem os processos bioquímicos.

Nos seres humanos, de maneira simplificada, ocorre um estímulo nos dendritos do neurônio e a informação é transmitida até o centro do neurônio, onde as informações são analisadas e é tomada uma decisão. Esta decisão tomada é transmitida pelo axônio para o local onde deve ser tomada a ação. Caso a informação que chegue ao neurônio não seja relevante, ele permanece inerte. A comunicação entre os neurônios ocorre por meio de sinapses, onde os impulsos nervosos são transmitidos de um neurônio para o outro. O cérebro humano possui por volta de 10^{11} neurônios e um número de sinapses maior que 10^{14} , formando uma imensa rede de troca e processamento de informações. O método de Redes Neurais Artificiais é um sistema de processamento de dados que opera de maneira análoga ao funcionamento dos neurônios do cérebro humano. (Cerqueira, Andrade, Poppi, & Mello, 2001).

Para simular o funcionamento do sistema nervoso, de maneira que artificialmente seja possível aprender a processar uma informação e reportar um resultado, são utilizadas redes neurais artificiais. Um modelo matemático é desenvolvido para que um computador processe as informações de entrada, utilizando várias funções matemáticas como neurônios, e reportar um resultado coerente. A rede neural artificial funciona como o sistema nervoso biológico. Primeiro ela é treinada, aprendendo a associar determinadas informações de entrada a um determinado resultado. Para um bom treinamento é necessário um grande conjunto de informações de entrada e respostas, e o conjunto deve contemplar todas as possíveis combinações de informações e o resultado que ocorrerá. Quanto mais informações são utilizadas na confecção do treinamento da rede neural, mais robusta ela se torna e mais confiável é seu resultado.

As redes neurais tem como arquitetura mais simples o perceptron (Figura 6), o qual é capaz de classificar padrões linearmente separáveis, por meio de uma única camada de

neurônios com pesos e bias ajustáveis. O perceptron funciona com um algoritmo da seguinte forma:

- Cada padrão de treinamento, que são os dados de entrada, são recebidos como x_i e os resultados, a saída da rede, é calculada e reportada como y .
- O resultado desejado para o padrão inserido, ou saída desejada, é dado por d_i , e o erro entre a saída desejada e a saída da rede é reportado por $e_i = d_i - y_i$.
- Para atualizar os pesos que conectam os dados de entradas a cada saída da rede e ao bias do neurônio, são utilizadas as seguintes regras:

$$w_i(t+1) = w_i(t) + \alpha \cdot e_i \cdot x_i;$$

$$b(t+1) = b(t) + \alpha \cdot e_i;$$

Figura 6. Modelo matemático da estrutura de um perceptron de redes neurais artificiais.

A saída pode ser descrita pela função $y = f(u) = f(\sum_{j=1}^M w_j x_j + b_k)$ (Von Zuben & Attux, 2010).

O processo de aprendizagem da figura 6 pode ser descrito, simplificadaamente, nas seguintes etapas:

1. O ambiente gera estímulos que ativam a rede neural, que seriam os valores de entrada X_j ;
2. Os parâmetros da rede são alterados (os pesos, W_j , e o bias, b_k) devido ao estímulo recebido;
3. A rede neural responde aos estímulos (y), conforme as alterações que ocorreram em sua estrutura.

Assim, temos que os algoritmos envolvidos no processo de aprendizagem não são estruturas fixas, mas variam de maneira a se ajustar aos pesos e remeter uma resposta adequada ao estímulo recebido (Pimentel, 2005).

1.4. Métodos Teóricos de Cálculo

Para calcular as propriedades utilizadas na análise quimiométrica, é necessário fazer uma modelagem da estrutura das moléculas sob estudo. Para esta modelagem são utilizados métodos da mecânica quântica, que podem ser *ab initio*, semi-empíricos ou *DFT (Density Functional Theory)*. Estes métodos podem ser utilizados para calcular propriedades eletrônicas das moléculas, além de fornecer a estrutura molecular espacial para que as propriedades físicas referentes a ela sejam calculadas.

A principal diferença entre estes métodos são as próprias características que os definem e não estão presentes nos outros métodos. Os métodos *ab initio* tem por característica a resolução da equação de *Schrödinger* utilizando apenas aproximações matemáticas e geralmente são cálculos com um custo computacional mais elevado. Os métodos semiempíricos, resolvem a equação de *Schrödinger* utilizando parâmetros obtidos de dados experimentais, tem um custo computacional geralmente menor que o dos outros métodos. Os métodos da *DFT* resolvem a equação de *Schrödinger* descrevendo os elétrons através de um funcional de densidade eletrônica, em vez de descreve-los individualmente como nos outros métodos. Os métodos *DFT* apresentam um custo computacional moderado e geram bons resultados quanto à caracterização da estrutura eletrônica das moléculas (Morgon & Coutinho, 2007).

2. Objetivos

O presente estudo visa avaliar o conjunto de 60 moléculas derivadas de benzo-fenil-éter com substituintes diamínicos, publicado no artigo “*Synthesis and antiprotozoal activities of benzyl phenyl ether diamidine derivatives*” (Patrick, et al., 2013), de maneira a relacionar as propriedades estruturais e eletrônicas das moléculas com sua atividade biológica e desenvolver modelos quimiométricos de predição de atividade biológica e classificação de moléculas análogas.

Os objetivos específicos consistem em:

- Obter e selecionar descritores eletrônicos e moleculares para as moléculas em estudo;
- Proposição e validação de um modelo QSAR utilizando o método ANN, e seleção dos grupos de treinamento e teste;
- Avaliar a influência do processamento dos dados por autoescalamento e centralização na média nos modelos QSAR e SAR;
- Proposição de modelos SAR utilizando o método HCA e avaliação das abordagens single e centroide;
- Proposição e validação de um modelo SAR utilizando o método PCA;
- Proposição e validação de um modelo SAR utilizando o método KNN;
- Proposição e validação de um modelo SAR utilizando o método SIMCA;
- Proposição e validação de um modelo QSAR utilizando o método PLS.

3. Metodologia

A estrutura base das moléculas envolvidas no presente estudo está representada na figura 7. Os principais substituintes, utilizados nas posições R₁ e R₄ das moléculas, estão representados na figura 8, com suas respectivas abreviações. A configuração estrutural de cada molécula utilizado neste estudo está descrita na tabela 1, juntamente com o respectivo valor de 50% inibição de crescimento (IC₅₀) do *Tripanossoma Brucei Rhondesiense* em µM e a nomenclatura utilizada para identificar cada uma das moléculas.

Figura 7. Estrutura molecular base, da qual partem todas as moléculas envolvidas no presente estudo.

Figura 8. Estrutura dos principais substituintes utilizados nas posições R₁ e R₄ das moléculas em estudo e suas abreviações.

Tabela 1. Nomeação das moléculas em estudo, seus substituintes e atividade biológica.

Moléculas	R ₁	R ₂	R ₃	R ₄	Tripanossoma Brucei
					Rhondesiense
					IC ₅₀ (μM)
M01	4-Am	H	H	H	1,84
M02	4-Am	I	H	4-NO ₂	1,9
M03	4-Am	I	I	3-NO ₂	2,66
M04	3-Am	H	H	H	9,32
M05	3-Am	H	H	3-NO ₂	0,481
M06	H	H	H	4-Am	3,71
M07	4-NO ₂	H	H	4-Am	4,41
M08	4-Cl	H	H	4-Am	2,71
M09	4-iPr	H	H	4-Am	1,29
M10	4-OMe	H	H	4-Am	3,95
M11	4-NH ₂	H	H	4-Am	3
M12	3-NO ₂	H	H	4-Am	2,12
M13	3-NO ₂	H	H	3-Am	7,28
M14	4-Am	H	H	4-Am	0,091
M15	4-iPrAm	H	H	4-iPrAm	1,35
M18	4-Am	Cl	H	4-Am	0,022
M19	4-Am	Br	H	4-Am	0,016
M20	4-AmOH	Br	H	4-AmOH	24
M21	4-AmOMe	Br	H	4-AmOMe	18,8
M22	4-Am	I	H	4-Am	0,017
M23	4-AmOH	I	H	4-AmOH	25,8
M24	4-AmOMe	I	H	4-AmOMe	50,4
M25	4-Am	Cl	Cl	4-Am	0,024
M26	4-Am	Br	Br	4-Am	0,017
M27	4-Am	I	I	4-Am	0,122
M28	4-Am	OMe	OMe	4-Am	0,11
M29	4-AmOH	OMe	OMe	4-AmOH	161
M30	4-AmOMe	OMe	OMe	4-AmOMe	129
M31	4-Am	H	H	3-Am	0,024

M32	4-Am	NO ₂	H	3-Am	0,034
M33	4-Am	Cl	H	3-Am	0,007
M34	4-AmOH	Cl	H	3-AmOH	49,2
M35	4-AmOMe	Cl	H	3-AmOMe	26,1
M36	4-Am	Br	H	3-Am	0,005
M37	4-AmOH	Br	H	3-AmOH	14,7
M38	4-AmOMe	Br	H	3-AmOMe	79,1
M39	4-Am	I	H	3-Am	0,005
M40	4-AmOH	I	H	3-AmOH	50,2
M41	4-AmOMe	I	H	3-AmOMe	34,7
M42	4-Am	OMe	H	3-Am	0,019
M43	4-Am	Cl	Cl	3-Am	0,01
M44	4-AmOH	Cl	Cl	3-AmOH	26,4
M45	4-AmOMe	Cl	Cl	3-AmOMe	60,1
M46	4-Am	Br	Br	3-Am	0,004
M47	4-AmOH	Br	Br	3-AmOH	6,89
M48	4-AmOMe	Br	Br	3-AmOMe	69,7
M49	4-Am	I	I	3-Am	0,003
M50	4-AmOH	I	I	3-AmOH	13,6
M51	4-AmOMe	I	I	3-AmOMe	26,3
M52	4-Am	Br	OMe	3-Am	0,005
M53	4-AmOH	Br	OMe	3-AmOH	87,9
M54	4-AmOMe	Br	OMe	3-AmOMe	81,5
M55	4-Am	I	OMe	3-Am	0,003
M56	4-AmOH	I	OMe	3-AmOH	38,7
M57	4-AmOMe	I	OCH ₃	3-AmOMe	75,1
M58	4-Am	OMe	OMe	3-Am	0,183
M59	4-AmOH	OMe	OMe	3-AmOH	157
M60	4-AmOMe	OMe	OMe	3-AmOMe	109
M61	3-Am	H	H	4-Am	0,028
M62	3-Am	H	H	3-Am	0,097

Das estruturas descritas na tabela 1 e figuras 7 e 8, foram modeladas cada uma das moléculas e submetidas a cálculos mecânico quânticos de minimização de energia. Estes

cálculos de minimização de energia foram realizados utilizando o programa Gaussian09 e GaussView05 (FRISCH, et al., 2009), em ambiente Linux e Windows, respectivamente. Para realizar estes cálculos foi utilizada a Teoria do Funcional de Densidade (DFT - *Density Functional Theory*) com funcional de densidade híbrido B3LYP e função de base 6-31+G(d,p).

Para obtenção dos descritores moleculares será utilizado o programa e-Dragon. Os descritores eletrônicos foram obtidos diretamente pelos cálculos mecânico-quânticos de minimização de energia feitos com o Gaussian 09.

A correlação entre as propriedades e a atividade biológica foi calculada utilizando o programa Excel do pacote Office. Também a estimativa inicial de quantas classes poderiam haver no conjunto em estudo foi feito utilizando o Excel, observando a formação de grupos unindo as moléculas com atividade biológica mais próxima. Esta avaliação foi feita aplicando uma função $-\log(x)$ na atividade biológica das moléculas e plotando um gráfico de pIC_{50} x pIC_{50} .

Os modelos QSAR serão obtidos utilizando os métodos exploratórios (não supervisionados), métodos de classificação (supervisionados) e métodos de regressão. Dos métodos exploratórios utilizaremos as técnicas de Análise de Componentes Principais (*Principal Component Analysis* - PCA) e Análise Hierárquica de Grupos (*Hierarchical Cluster Analysis* - HCA). Em todos os métodos foram comparadas as abordagens de pré-tratamento dos dados por autoescalamento e centralização na média e para o método HCA foram utilizadas também duas abordagens de ligação, centroide e single.

Dentre os métodos de classificação supervisionados foram utilizados o Reconhecimento de Padrões do k-ésimo Vizinho mais Próximo (*k-Nearest Neighbors* - KNN) e o método SIMCA (*Soft Independent Modeling by Class Analogy*). Este método foi utilizado com duas abordagens de tratamento de dados, centralização na média e autoescalamento.

Dos métodos de regressão foi utilizado o de Regressão por Mínimos Quadrados Parciais (*Partial Least Squares* - PLS). Os modelos QSAR serão gerados utilizando o programa Pirouette 3.10.

O modelo de redes neurais artificiais foi feito utilizando os programa MATLAB(R2011a) com função de transferência tansig e 17 neurônios para o processamento dos dados, 10

moléculas foram separados para validação interna do modelo, 12 para o conjunto de teste e 38 para o conjunto de treinamento. A separação das moléculas para o conjunto de treinamento, validação e teste são feitas automaticamente pelo programa, de acordo com a programação do script.

O script utilizado nas redes neurais foi desenvolvido por Aline Alves de Oliveira, durante seu doutorado junto ao grupo de pesquisa. Ele realiza a aleatorização das moléculas e as agrupa para formar o conjunto de treino, o de validação e o de testes. Após isso, ele utiliza o conjunto de treinamento e validação para desenvolver o modelo de redes neurais, utiliza o conjunto de validação para realizar a validação interna e posteriormente o conjunto de teste para realizar a validação externa. Este processo é repetido 10 mil vezes e são registrados apenas os modelos que apresentaram performance dentro dos critérios estabelecidos. Estes modelos registrados ficam disponíveis em arquivo de texto para avaliação do usuário e seleção do modelo mais adequado.

O modelo PLS foi gerado utilizando os mesmos conjuntos de treinamento e teste obtido pelo método de redes neurais, sem utilizar o conjunto de validação interna das redes neurais (Lipinski, Oliveira, Honorio, Oliveira, & da Silva, 2018).

4. Resultados e Discussão

4.1. Seleção de descritores

Inicialmente foi avaliado o conjunto de moléculas em estudo quanto à formação de grupos e a separação entre os mesmos. Como expressa o gráfico na figura 9, podemos perceber a formação de dois grupos principais, os ativos, em vermelho, e os menos ativos, em verde (visto que nenhum das moléculas é inativa, optamos pela denominação “menos ativos”). O ponto que está indicado como M5, não foi pré-classificado devido à sua distância entre os dois grupos. Mas inicialmente sabemos que há dois potenciais grupos principais, devido aos valores de atividade biológica das moléculas estarem separados por uma diferença considerável.

Figura 9. Gráfico de avaliação da atividade biológica para proposição de grupos.

Após ser observada a separação entre dois grupos principais, foi dada continuidade ao estudo, realizando a modelagem das moléculas e os cálculos de minimização de energia para obter a estrutura mais próxima do real. Embora esta etapa pareça simples, é de fundamental importância escolher um bom conjunto de bases e funcional para a realização do cálculo. A estrutura das moléculas foi avaliada quanto à quebra de ligações ou obtenção de

frequências negativas, quando em alguma dessas era observada na estrutura final do cálculo, esse era realizado novamente após algumas modificações em ângulos de torção ou rotação, dos substituintes da molécula.

Com a modelagem pronta e sem apresentar nenhuma incoerência, as estruturas foram submetidas ao programa e-Dragon e gerados os descritores. As propriedades foram obtidas do banco de dados do e-Dragon e conjuntamente foram obtidas propriedades diretamente dos cálculos de minimização de energia. No total foram geradas 1194 propriedades relacionadas às moléculas do conjunto em estudo. Levar em consideração todas estas propriedades pelos métodos quimiométricos é impraticável, por isso foi feita uma pré-seleção desprezando as propriedades que apresentavam correlação com a atividade biológica inferior a 6,0, restando apenas 54 propriedades. Este valor foi adotado como referência, pois descritores com correlação com a atividade biológica inferior a esta não oferecem uma quantidade relevante de informações e quantidade de propriedades restantes é suficiente para o prosseguimento dos estudos (Márcia M. C. Ferreira, 2011) (Kiralj & Ferreira, 2009).

As 54 propriedades restantes foram avaliadas inicialmente quanto à duplicidade e relevância de suas informações. Informações redundantes ou que não condizem com a estrutura das moléculas em estudo foram retiradas, diminuindo um pouco mais o conjunto de propriedades. Com as restantes foi aplicado o método PCA, com os dados autoescalados, para verificar se ocorreria a separação das moléculas em dois grupos principais, como visto na figura 8.

O processo de seleção das propriedades nesta etapa era avaliar os valores de Loading das propriedades para as componentes principais 1 e 2, as que apresentavam valores próximos a zero em ambas as PCs eram retiradas e um novo cálculo PCA era realizado gerando um novo modelo. As propriedades que apresentavam baixa correlação com a atividade biológica também eram retiradas e um novo modelo era gerado. Seguindo este procedimento, era avaliada a melhora ou piora no modelo e se as contribuições de outras propriedades aumentavam ou diminuía. Quando ocorria melhora no modelo de separação a propriedade retirada era desprezada para a continuidade do estudo, mas quando o

modelo não apresentava melhora, a propriedade era reintegrada ao conjunto e outra era retirada.

O objetivo inicial era a segregação das moléculas em dois grupos principais, como visto na figura 9. O que reporta a figura 10 é o resultado obtido com os 6 propriedades selecionadas pelo procedimento relatado acima. Vemos que os dois grupos principais foram separados adequadamente e que o grupo principal das moléculas menos ativas se dividiu em 3 subgrupos. Passamos então a nomear as propriedades selecionadas como descritores, pois apresentam informações relevantes para descrever o conjunto de moléculas em estudo.

Figura 10. Separação dos grupos obtida pelo método PCA utilizando os 60 moléculas.

As propriedades selecionadas como descritores foram, a energia do orbital HOMO, o número de aminas primárias alifáticas ($n\text{NH}_2$), fator hidrofílico (Hy), o índice de primeira ordem da média de carga topológica (JGI1), a autocorrelação de Moran na *lag-1* ponderada pela eletronegatividade de Sanderson (MATS1e), e autocorrelação de Geary na *lag-7* ponderada pela eletronegatividade de Sanderson (GATS7e). As autocorrelações de Moran e Geary são métodos estatísticos para analisar uma distribuição em um determinado espaço (MORAN, 1950) (Geary, 1954).

Os métodos de Moran e Geary são ferramentas estatísticas utilizadas, inicialmente, para analisar uma região dividindo-a em partes e avaliar a contribuição de cada parte para o todo.

Estas ferramentas foram empregadas inicialmente em mapeamentos de rendimentos econômicos em países da Europa ou regiões comerciais. Posteriormente foi aplicado para avaliação da estrutura de moléculas como se fossem mapas, cada região da molécula colaborando com determinada intensidade. No caso em que estamos analisando, a contribuição das regiões da molécula é referente à eletronegatividade de Sanderson.

Os valores de correlação destes descritores com a atividade biológica e entre eles estão expressos na tabela 2, e nela vemos que todos os descritores apresentam valores relevantes de correlação com a atividade biológica. Peculiarmente também observamos uma elevada correlação entre $n\text{NH}_2$ e Hy, o que poderia indicar duplicidade de informação, mas ao avaliarmos esses os dois descritores em conjunto, percebermos que não há a duplicidade de informações. Sabemos que a presença de ligações N-H contribui para o aumento de Hy, devido às ligações de hidrogênio que podem ser formadas, mas há outros fatores que também podem influir sobre Hy. A presença de grupos retiradores, ou doadores, de elétrons fortes no anel aromático pode aumentar a polaridade da molécula e torna-la mais hidrofílica. Isso denota que $n\text{NH}_2$ contribui para Hy, mas não é o único fator a influenciar Hy, por isso podemos manter os dois descritores no modelo.

Tabela 2. Correlação entre os descritores, e entre descritores e a atividade biológica.

	pIC₅₀	HOMO	nNH₂	Hy	MATS1e	JGI1	GATS7e
pIC ₅₀	1						
HOMO	-0,68009	1					
nNH ₂	0,944941	-0,68381	1				
Hy	0,894173	-0,57249	0,903589	1			
MATS1e	-0,75426	0,233427	-0,7352	-0,84078	1		
JGI1	0,710234	-0,66347	0,619099	0,722985	-0,47315	1	
GATS7e	0,688189	-0,45598	0,646895	0,684071	-0,62039	0,60883	1

Na tabela 3, vemos que as moléculas que apresenta orbital HOMO com menor energia tendem a ser ativas, na maioria dos casos. Calculamos a média dos valores de energia para o orbital HOMO das moléculas ativas e das menos ativas (tabela 4), obtendo para as moléculas ativas o valor de -0,24003 e para as menos ativas tem valor de -0,22399. Esta diferença reforça o que observamos analisando individualmente cada molécula de cada grupo, ativas e menos ativas. Pelos princípios da regra de Hund temos que, os orbitais moleculares são preenchidos do nível de energia mais baixo ao mais alto, logo os níveis de baixa energia são

mais facilmente preenchidos, sendo também mais estáveis. Assim, percebemos que as moléculas que apresentam orbitais HOMO mais estáveis energeticamente tendem a ser ativas, e as que apresentam orbitais HOMO menos estáveis tendem a ser menos ativas.

Peculiarmente o grupo de moléculas menos ativas G3 apresenta em média energia do HOMO menos que o grupo das moléculas ativas, mas apresenta atividade biológica inferior. Isso nos mostra que a energia do orbital HOMO não é um fator isolado a contribuir para atividade biológica, mas deve estar combinado a outros fatores. No entanto, mantemos a observação de que, preferencialmente, valores baixos de energia do orbital HOMO contribuem positivamente para a atividade biológica.

A importância do orbital HOMO nos chamou a atenção para também analisar a diferença energética HOMO-LUMO e avaliar a relação com a atividade biológica. Tal relação é muito baixa, indicando que transições eletrônicas entre os dois orbitais não tem relevância para a atividade biológica das moléculas.

Tabela 3. Descritores selecionados para o estudo SAR e QSAR em ordem decrescente de atividade biológica (pIC₅₀).

Molécula	Atividade	pIC ₅₀	Descritores					
			Homo	nNH ₂	Hy	MATS1e	JGI1	GATS7e
M49	Ativo	8,522878745	-0,24842	2	1,923	-0,237	0,262	2,821
M55	Ativo	8,522878745	-0,24216	2	1,873	-0,182	0,25	2,639
M46	Ativo	8,397940009	-0,24789	2	1,923	-0,233	0,262	3,028
M36	Ativo	8,301029996	-0,24282	2	1,931	-0,222	0,25	2,875
M39	Ativo	8,301029996	-0,24105	2	1,931	-0,224	0,25	2,773
M52	Ativo	8,301029996	-0,24304	2	1,873	-0,186	0,25	2,671
M33	Ativo	8,15490196	-0,24268	2	1,931	-0,206	0,25	2,919
M43	Ativo	8,000000000	-0,24821	2	1,923	-0,204	0,262	3,094
M19	Ativo	7,795880017	-0,23989	2	1,931	-0,222	0,25	2,555
M22	Ativo	7,769551079	-0,2349	2	1,931	-0,224	0,25	2,376
M26	Ativo	7,769551079	-0,2475	2	1,923	-0,233	0,262	2,767
M42	Ativo	7,721246399	-0,2249	2	1,879	-0,173	0,238	2,606
M18	Ativo	7,657577319	-0,23754	2	1,931	-0,206	0,25	2,679
M25	Ativo	7,619788758	-0,24852	2	1,923	-0,204	0,262	2,955
M31	Ativo	7,619788758	-0,23457	2	1,938	-0,212	0,237	2,729
M61	Ativo	7,552841969	-0,23054	2	1,938	-0,212	0,237	1,39
M32	Ativo	7,468521083	-0,2519	2	1,916	0,024	0,25	3,179
M14	Ativo	7,040958608	-0,23457	2	1,938	-0,212	0,237	2,336
M62	Ativo	7,013228266	-0,2356	2	1,938	-0,212	0,237	1,817
M28	Ativo	6,958607315	-0,22786	2	1,827	-0,156	0,239	2,209

M27	Ativo	6,913640169	-0,24407	2	1,923	-0,237	0,262	2,419
M58	Ativo	6,73754891	-0,23209	2	1,827	-0,156	0,239	2,307
M5	M.A.	6,317854924	-0,24302	1	0,417	0,123	0,237	0,92
M9	M.A.	5,88941029	-0,22392	1	0,996	-0,133	0,25	1,629
M15	M.A.	5,869666232	-0,22781	0	0,218	-0,073	0,26	1,4
M1	M.A.	5,735182177	-0,23054	1	0,362	-0,113	0,188	2,21
M2	M.A.	5,721246399	-0,24703	1	0,431	0,12	0,25	2,076
M12	M.A.	5,673664139	-0,25329	1	0,417	0,123	0,237	1,854
M3	M.A.	5,575118363	-0,25854	1	0,444	0,117	0,262	1,633
M8	M.A.	5,567030709	-0,23012	1	0,383	-0,116	0,235	1,571
M11	M.A.	5,522878745	-0,19592	1	2,007	-0,185	0,235	1,376
M6	M.A.	5,43062609	-0,22698	1	0,362	-0,113	0,188	1,105
M10	M.A.	5,403402904	-0,20654	1	0,349	-0,092	0,222	2,734
M7	M.A.	5,355561411	-0,25607	1	0,417	0,123	0,237	1,12
M47	M.A.	5,161780778	-0,22307	0	0,425	-0,053	0,239	2,124
M13	M.A.	5,137868621	-0,25336	1	0,417	0,123	0,237	2,019
M4	M.A.	5,030584088	-0,2292	1	0,362	-0,113	0,188	1,01
M50	M.A.	4,866461092	-0,2177	0	0,425	-0,023	0,239	1,745
M37	M.A.	4,832682665	-0,22289	0	0,413	-0,032	0,227	1,903
M21	M.A.	4,725842151	-0,21593	0	-0,672	0,073	0,208	1,115
M20	M.A.	4,619788758	-0,22321	0	0,413	-0,032	0,227	1,447
M23	M.A.	4,588380294	-0,2227	0	0,413	-0,016	0,227	1,206
M35	M.A.	4,583359493	-0,21517	0	-0,672	0,06	0,208	1,609
M51	M.A.	4,580044252	-0,21472	0	-0,646	0,085	0,22	1,383
M44	M.A.	4,578396073	-0,22313	0	0,425	-0,063	0,239	2,404
M41	M.A.	4,459670525	-0,21631	0	-0,672	0,09	0,208	1,369
M56	M.A.	4,412289035	-0,22037	0	0,397	-0,037	0,229	2,06
M34	M.A.	4,308034897	-0,22135	0	0,413	-0,038	0,227	2,066
M40	M.A.	4,299296283	-0,22246	0	0,413	-0,016	0,227	1,711
M24	M.A.	4,297569464	-0,21551	0	-0,672	0,09	0,208	0,958
M45	M.A.	4,221125528	-0,21602	0	-0,646	0,032	0,22	1,829
M48	M.A.	4,156767222	-0,21583	0	-0,646	0,052	0,22	1,629
M57	M.A.	4,124360063	-0,21337	0	-0,659	0,054	0,212	1,732
M38	M.A.	4,101823517	-0,21669	0	-0,672	0,073	0,208	1,494
M54	M.A.	4,088842391	-0,21367	0	-0,659	0,04	0,212	1,8
M53	M.A.	4,056011125	-0,22055	0	0,397	-0,05	0,229	2,161
M60	M.A.	3,962573502	-0,21161	0	-0,671	0,032	0,204	1,799
M30	M.A.	3,88941029	-0,21174	0	-0,671	0,032	0,204	1,628
M59	M.A.	3,804100348	-0,21802	0	0,37	-0,047	0,22	2,038
M29	M.A.	3,793174124	-0,21712	0	0,37	-0,047	0,22	1,86

Tabela 4. Média dos valores dos descritores para o grupo de moléculas ativas e para o grupo de moléculas menos ativas.

Descritores	Ativo	Menos ativos	G1	G2	G3
HOMO	-0,24003	-0,22399	-0.22118	-0.21471	-0.25189
nNH ₂	2,000	0,342	0.35	0	1
Hy	1,912	0,118	0.49565	-0.66317	0.423833
MATS1e	-0,197	0,001	-0.0696	0.059417	0.1215
JGI1	0,249	0,224	0.2258	0.211	0.243333
GATS7e	2,597	1,677	1.788	1.52875	1.603667

Ao analisarmos a tabela 4, percebemos que as moléculas ativas apresentam em média 2 grupos de amins primárias alifáticas e as menos ativas apresentam em média menos que uma, logo podem apresentar uma ou nenhuma. Para ter uma visão mais ampla e confiável, analisamos também a tabela 3 e vemos que todas as moléculas do grupo das ativas apresentam amins primárias alifáticas, e as moléculas do grupo das menos ativas, a maioria, não apresentam amins primárias alifáticas. Nisto vemos que a presença de amins primárias alifáticas tem elevada relação com a atividade biológica das moléculas e que, provavelmente, esses grupos de amins primárias alifáticas sejam importantes para executar interações do tipo ligação de hidrogênio ou pontos eletrofílicos em que possa haver interação com o par de elétrons livre do nitrogênio. Comparando também as moléculas percebemos que a presença de grupos de amins primárias alifáticas também contribui para que o orbital HOMO da molécula tenha uma energia menor, pois todas as moléculas que tem esses grupos tem energia do HOMO menor que as moléculas que não tem.

Ao analisarmos o Hy também pela tabela 4, percebemos uma grande discrepância entre as moléculas ativas e as menos ativas, demonstrando que as ativas têm alta hidrofiliidade e as menos ativas apresentam baixa hidrofiliidade. Isso realmente é observado ao analisarmos as moléculas individualmente na tabela 3, onde percebemos que todas as moléculas ativas apresentam valores elevados para Hy e as menos ativas apresentam baixos valores para Hy, algumas sendo até hidrofóbicas, com Hy negativo. Como ao fazer média são somados todos os valores, tanto os positivos quanto os negativos, a média apresentada na tabela 4 é abaixo dos valores apresentados pelas moléculas menos ativas que não são hidrofóbicas, podendo gerar equívocos. Separamos as moléculas hidrofóbicas e percebemos que estas são as moléculas que forma o grupo G2, que foi observado na análise PCA (Figura

8) e que as demais moléculas menos ativas não tem um caráter hidrofóbico tão pronunciado.

Visto que o grupo G2, pela análise PCA encontra-se mais afastados do grupo das moléculas ativas que os grupos G1 e G3, e que ele é o único grupo que apresenta moléculas hidrofóbicas, percebemos que o fator hidrofílico realmente é um descritor importante. Essa análise nos mostra que uma alta hidrofiliabilidade é uma característica importante que a molécula apresente elevada atividade biológica.

Os descritores MATS1e, GATS7e e JGI1, são propriedades de autocorrelação 2D as duas primeiras e a última é topológica de índice de carga. Os três estão relacionados à distribuição de carga na molécula, mostrando a importância dos átomos que têm efeito de atração eletrônica para si ou de doação. As moléculas em estudo apresentam grupos retiradores ou halogênios em sua estrutura, mostrando que a presença deste efeito retirador contribui para a atividade biológica.

O descritor MATS1e, pela média reportada na tabela 4, mostra que há uma grande diferença entre as moléculas ativas e as menos ativas, mas ao vermos a média para cada um dos grupos percebemos que o grupo G1 contribui com valores negativos e os outros dois grupos com valores positivos. Assim, a discrepância entre as moléculas dos grupos G2 e G3 com as moléculas ativas é muito maior do que o apresentado pela média geral. Analisando a tabela 3, percebemos também que todas as moléculas ativas, com exceção da M32, apresentam valores negativos e que das menos ativas, apenas as moléculas do grupo G1 apresentam valores negativos. Retornando novamente à análise PCA (figura 8), percebemos que dos três grupos de moléculas menos ativas, G1 é o mais próximo das moléculas ativas. Esses dados nos levam a ver valores negativos para MATS1e colaboram para que a molécula tenha uma atividade biológica mais alta, indicando também que é importante que uma determinada região da molécula tenha uma maior densidade de carga eletrônica.

No descritor JGI1 não percebemos tamanha discrepância entre os grupos como nos demais descritores, mas há leves tendências em cada grupo que podem ser usadas para separá-los (Tabela 4). Percebemos que as mais ativas apresentam, geralmente, valores mais elevados que as demais moléculas, levando-nos a perceber a importância de que a molécula tenha uma região ligeiramente deficiente em elétrons para ter uma boa atividade biológica.

O descritor GATS7e reporta uma boa separação entre as moléculas ativas e as menos ativas, mas não há uma diferença relevante entre os três grupos de moléculas menos ativas, como visto na tabela 4. Sabendo que na tabela 4 o valor a que nos referimos é a média, cabe analisar os resultados de cada molécula na tabela 3. Percebemos então que a maioria das moléculas ativas apresentam valores elevados para GATS7e, sendo apenas duas com valores menor que 2. As moléculas menos ativas apresentam, majoritariamente, valores abaixo de 2, sendo apenas oito as que apresentam valores acima deste. Isto demonstra a importância da eletronegatividade da região determinada pelo descritor para a atividade biológica da molécula.

Ao avaliarmos individualmente os descritores, quanto à sua relação com a atividade biológica e o que foi reportado pela análise PCA, vemos a validade das informações. Caso os descritores não demonstrassem uma contribuição real, algo que não pudesse ser relacionado com a atividade biológica, uma nova seleção de descritores deveria ser feita. Fazendo essa análise podemos evitar a combinação de dados aleatórios para obtenção de resultados aparentemente satisfatórios. Sendo os descritores aprovados na seleção, podemos iniciar a etapa de desenvolvimento de modelos quimiométricos.

4.2. Influência dos substituintes na atividade biológica

Na tabela 5 temos algumas moléculas selecionadas por serem similares e as agrupamos de acordo com os substituintes nas posições R_1 e R_2 para analisar a influência dos substituintes sobre os descritores e a atividade biológica. Foram selecionadas moléculas similares para que possamos observar nas pequenas mudanças estruturais, relações com a atividade biológica e descritores. Assim, poderemos ter uma melhor visão de qual seria a região de ação dos descritores, principalmente MATS1e e GATS7e, que são provenientes de ferramentas estatísticas de mapeamento.

Tabela 5. Agrupamento de moléculas similares, separados de acordo com os substituintes das posições R₁ e R₂ para avaliação de influências dos substituintes sobre a atividade biológica e descritores.

Moléculas	R ₁	R ₂	R ₃	R ₄	pIC ₅₀	Homo	nNH ₂	Hy	MATS1e	JGI1	GATS7e
M49	4-Am	I	I	3-Am	8.52288	-0.24842	2	1.923	-0.237	0.262	2.821
M46	4-Am	Br	Br	3-Am	8.39794	-0.24789	2	1.923	-0.233	0.262	3.028
M43	4-Am	Cl	Cl	3-Am	8.00000	-0.24821	2	1.923	-0.204	0.262	3.094
M36	4-Am	Br	H	3-Am	8.30103	-0.24282	2	1.931	-0.222	0.25	2.875
M39	4-Am	I	H	3-Am	8.30103	-0.24105	2	1.931	-0.224	0.25	2.773
M33	4-Am	Cl	H	3-Am	8.15490	-0.24268	2	1.931	-0.206	0.25	2.919
M55	4-Am	I	OMe	3-Am	8.52288	-0.24216	2	1.873	-0.182	0.25	2.639
M52	4-Am	Br	OMe	3-Am	8.30103	-0.24304	2	1.873	-0.186	0.25	2.671
M26	4-Am	Br	Br	4-Am	7.76955	-0.2475	2	1.923	-0.233	0.262	2.767
M25	4-Am	Cl	Cl	4-Am	7.61979	-0.24852	2	1.923	-0.204	0.262	2.955
M27	4-Am	I	I	4-Am	6.91364	-0.24407	2	1.923	-0.237	0.262	2.419
M19	4-Am	Br	H	4-Am	7.79588	-0.23989	2	1.931	-0.222	0.25	2.555
M22	4-Am	I	H	4-Am	7.76955	-0.2349	2	1.931	-0.224	0.25	2.376
M18	4-Am	Cl	H	4-Am	7.65758	-0.23754	2	1.931	-0.206	0.25	2.679
M47	4-AmOH	Br	Br	3-AmOH	5.16178	-0.22307	0	0.425	-0.053	0.239	2.124
M50	4-AmOH	I	I	3-AmOH	4.86646	-0.2177	0	0.425	-0.023	0.239	1.745
M44	4-AmOH	Cl	Cl	3-AmOH	4.57840	-0.22313	0	0.425	-0.063	0.239	2.404
M37	4-AmOH	Br	H	3-AmOH	4.83268	-0.22289	0	0.413	-0.032	0.227	1.903
M34	4-AmOH	Cl	H	3-AmOH	4.30803	-0.22135	0	0.413	-0.038	0.227	2.066
M40	4-AmOH	I	H	3-AmOH	4.29930	-0.22246	0	0.413	-0.016	0.227	1.711
M56	4-AmOH	I	OMe	3-AmOH	4.41229	-0.22037	0	0.397	-0.037	0.229	2.06
M53	4-AmOH	Br	OMe	3-AmOH	4.05601	-0.22055	0	0.397	-0.05	0.229	2.161
M51	4-AmOMe	I	I	3-AmOMe	4.58004	-0.21472	0	-0.646	0.085	0.22	1.383
M45	4-AmOMe	Cl	Cl	3-AmOMe	4.22113	-0.21602	0	-0.646	0.032	0.22	1.829
M48	4-AmOMe	Br	Br	3-AmOMe	4.15677	-0.21583	0	-0.646	0.052	0.22	1.629
M35	4-AmOMe	Cl	H	3-AmOMe	4.58336	-0.21517	0	-0.672	0.06	0.208	1.609
M41	4-AmOMe	I	H	3-AmOMe	4.45967	-0.21631	0	-0.672	0.09	0.208	1.369
M38	4-AmOMe	Br	H	3-AmOMe	4.10182	-0.21669	0	-0.672	0.073	0.208	1.494
M57	4-AmOMe	I	OMe	3-AmOMe	4.12436	-0.21337	0	-0.659	0.054	0.212	1.732
M54	4-AmOMe	Br	OMe	3-AmOMe	4.08884	-0.21367	0	-0.659	0.04	0.212	1.8

O descritor que apresenta discrepância mais nítida ao compararmos as moléculas é nNH₂, que reporta a presença de amins primárias alifáticas apenas para o substituinte Am. Ao analisar os outros substituintes, percebemos que AmOH e AmOMe também apresentam amins primárias alifáticas. Um provável motivo para que ao avaliar a molécula o programa e-Dragon não tenha atribuído uma amina primária alifática a AmOH e AmOMe, seja um efeito de ressonância.

Quando analisamos os substituintes sob a premissa do efeito de ressonância, se considerarmos que o par de elétrons livres da amina passe a se deslocar para o carbono ligado a ela, isso poderia ser entendido com uma dupla ligação pelo programa e-Dragon e as aminas que apresentam este efeito não seriam computadas. Como apenas Am apresenta duas aminas, uma liga ao carbono por ligação simples e a outra por ligação dupla, com a ressonância sempre haverá uma realizando apenas ligações simples e o átomo de hidrogênio entre elas acaba sendo compartilhado por ambas (figura 11). Em AmOH e AmOMe há um oxigênio presente, e sabido que este é muito eletronegativo, ele atrai para si a densidade eletrônica e acaba aumentando o caráter de dupla ligação entre a amina e o carbono.

Figura 11. Ilustração do efeito de ressonância estima-se ocorrer no substituinte Am.

Para avaliarmos se as suposições sobre o efeito de ressonância e a atração eletrônica ocasionada pela presença de oxigênio na molécula são válidas, vamos analisar os substituintes da posição R₁ das moléculas M49, M50 e M51. Nas figuras 12, 13 e 14, vemos que a carga do carbono em que os átomos de nitrogênio estão ligados diminui o valor da carga de Am para AmOH e para AmOMe. Isso mostra que a presença do átomo de oxigênio atrai elétrons por efeito indutivo para aquela região da molécula, tornando aquele carbono mais eletronegativo. Também percebemos uma pequena variação na carga do hidrogênio da amina primária, os mais positivos são mais ácidos e estão mais disponíveis para interações, e os com menor valor estão menos disponíveis.

O mesmo efeito também é observado em outros trios de moléculas análogas, como M46, M47 e M48. Neste trio é observado o mesmo efeito de carga, mas com a peculiaridade de que a variação entre as cargas do carbono de M47 e de M48 são muito pequenas, após a

terceira casa decimal. Isso provavelmente é um efeito advindo dos substituintes presentes em R₂ e R₃ que acabam tornando o comportamento observado menos intenso.

Um fato a ser denotado, é que a presença de amins primárias alifáticas contribui significativamente para melhora da atividade biológica. Isso nos mostra que a presença dessas amins pode estar vinculada a algum processo importante de ação da molécula, seja a interação com sítio ativo, ou algum processo intermediário que anteceda esta interação.

Figura 12. Substituinte Am da posição R₁ da molécula M49, com as cargas de cada átomo.

Figura 13. Substituinte AmOH da posição R₁ da molécula M50, com as cargas de cada átomo.

Figura 14. Substituinte AmOMe da posição R_1 da molécula M51, com as cargas de cada átomo.

Tendo em vista que grupos amina interferem na hidrofiliicidade de uma molécula, devido às ligações de hidrogênio que podem realizar, daremos continuidade à avaliação da relação entre os substituintes e os descritores, pelo descritor H_y .

Primeira hipótese de combinação a ser trabalhada é manter constantes os substituintes em R_1 e R_4 e variar apenas os haletos presentes em R_2 e R_3 . Fazendo isso, notamos que em todos os casos, não há nenhuma variação em H_y . A identidade do haleto presente na molécula, se Cl, Br ou I, não exercer nenhuma influência sobre H_y . No entanto, quando a identidade dos haletos é alterada, observamos uma no valor da atividade biológica, mostrando que a identidade dos haletos tem relação com a atividade biológica e pode ser um fator importante em outros dos descritores selecionados.

Visto que a identidade dos haletos não influencia H_y , vamos variar o substituinte da posição R_3 , permutando o haleto por H ou OMe, e mantendo a proposição da primeira hipótese. Neste caso observamos variações na atividade biológica, mostrando que existe uma influência sobre H_y . Para o substituinte Am em R_1 e R_4 , vemos que o H_y aumenta quando permutamos o haleto por H e diminui quando o permutamos por OMe. No caso do substituinte AmOH, percebemos que o valor de H_y diminui ao permutarmos o haleto por H e diminui um pouco mais quando o permutamos por OMe. Com o substituinte AmOMe observamos que utilizando H em R_3 , o valor é menor que com o haleto e que utilizando OMe o valor de H_y fica entre o valor com o haleto e com o H. Assim, percebemos que o efeito em

Hy da variação do substituinte de R_3 está condicionado a qual substituinte está presente nas posições R_1 e R_4 .

Uma segunda hipótese a ser trabalhada é manter R_2 e R_3 constantes e variar R_1 e R_4 . Executando esta hipótese percebemos uma variação mais expressiva em Hy do que a observada na primeira hipótese. Quando o substituinte em R_1 e R_4 é Am, independente da posição, 3 ou 4 em R_4 , o valor de Hy é elevado; quando é AmOH, o valor de Hy corresponde a quase um quinto do apresentado pela molécula similar com Am; e quando o substituinte é AmOMe, o valor de Hy é negativo. Conciliando estas observações com as que foram feitas para o descritor nNH_2 com relação aos substituintes Am, AmOH e AmOMe, vemos que a disponibilidade para realizar interações do tipo ligação de hidrogênio influencia diretamente na hidrofobicidade da molécula. Há outros fatores que colaboram para que o Hy aumente ou diminua, mas vemos que a ação de forças intermoleculares é o principal.

O Descritor JGI1 tem um comportamento similar ao Hy. Geralmente apresenta valores mais elevados quando a molécula é composta de dois haletos, nas posições R_2 e R_3 , e valores menores quando apresenta um haleto e um hidrogênio nas respectivas posições. Também percebemos que a variação dos substituintes nas posições R_1 e R_4 tem maior influência sobre JGI1, sendo os maiores valores observados quando o substituinte é Am, os valores médios com AmOH e os menores valores com AmOMe.

Os três descritores, nNH_2 , Hy e JGI1, nos mostram que os substituintes das posições R_1 e R_4 tem fundamental importância para a atividade biológica da molécula e que os substituintes em R_2 e R_3 tem uma contribuição não tão significativa para estes descritores quanto parece ter para a atividade biológica. Assim, utilizar apenas estes descritores poderia realizar uma boa separação em grupos, qualitativamente, mas fica evidente ainda ser necessária a contribuição dos demais descritores para uma melhor classificação das moléculas.

Ao analisarmos o descritor de energia do HOMO, percebemos a mesma tendência observada para os descritores anteriores quando aos substituintes R_1 e R_4 . No entanto, percebemos que a diminuição da energia do HOMO é o que caracteriza as moléculas com maior atividade biológica, na maioria dos casos. Os substituintes Am tem o HOMO com menor energia que AmOH, que tem menor energia que AmOMe. Neste descritor

percebemos uma particularidade que não foi constatada nos descritores anteriores, a diferença entre os substituintes em R₄, das posições 4 e 3. Para as moléculas com substituintes 3-Am em R₄, constatamos uma energia para o HOMO menor que para as moléculas com 4-Am, e a atividade biológica para 3-Am maior que para 4 Am.

Em relação à variação dos substituintes em R₂ e R₃, percebemos que há variação da energia do HOMO entre as moléculas de acordo com os haletos que as compõem, variação esta que também não havia sido observada nos descritores anteriores. Vemos então que a identidade dos haletos afeta a energia do HOMO e também a atividade biológica das moléculas. Na maioria dos casos, as moléculas com dois haletos apresenta menor energia para o HOMO e maior atividade biológica. O haleto que, geralmente, apresenta menor energia para o HOMO é o Cl, mas a molécula em que ele está presente não é a de melhor atividade biológica das moléculas em comparação. O que reforça a ideia de que o resultado reportado por um grupo de descritores para classificação das moléculas é mais confiável que o reportado por um descritor isolado.

A energia do HOMO indica sua estabilidade, quanto menor a energia, mais estável é o orbital. Assim, percebemos que as moléculas com melhor atividade biológica são as mais estáveis. Isso indica a probabilidade de que haja algum processo de interação durante o percurso até o sítio ativo que acabe retendo moléculas com orbitais menos estáveis.

Para o descritor MATS1e também percebemos que os substituintes em R₁ e R₄ são os que exercem maior influência e os presentes nas posições R₂ e R₃, uma influência menor. Vemos que o substituinte Am é o que apresenta menor valor, seguido por AmOH e o maior valor é apresentado quando o substituinte é AmOMe. Quando avaliamos a variação em R₂ e R₃, percebemos que o menor valor auxilia na melhora da atividade biológica, mas não é uma regra. Há casos em que a molécula apresenta menor valor para MATS1e, mas não é a de maior atividade. Logo esse descritor deve estar associado a outra propriedade que também influencia sobre a atividade biológica.

Algo interessante é que geralmente nas moléculas mais ativas dos grupos que estamos comparando, apresentam haletos maiores, com nuvem eletrônica maior, no caso I e Br. É provável que as moléculas com Cl não tenham atividade biológica tão elevada quanto as demais, pois a nuvem eletrônica deste haleto não é tão deformável. As nuvens eletrônicas

do I e Br são mais maleáveis e volumosas, o que talvez auxilie em processos relacionados à atividade biológica ou estabilização de carga na molécula internamente.

Ao analisar a relação entre o descritor GATS7e, a atividade biológica e as estruturas das moléculas vemos uma relação semelhante à que foi vista nos descritores anteriores. Neste pequeno conjunto que tomamos para analisar, observamos que as moléculas com maior atividade biológica apresentam valores elevados para GATS7e. A principal segregação foi vista quando separamos os grupos pelos substituintes R₁ e R₄, onde vemos que as moléculas com substituintes 3-Am em R₄ apresentam maiores valores de GATS7e do que a molécula correspondente com 4-Am em R₄, e também maior atividade biológica. As moléculas com substituinte Am são as que tem maior valor para GATS7e, posteriormente temos as moléculas com substituintes AmOH e as com menor valor para GATS7e as moléculas com substituinte AmOMe.

Com relação aos substituintes em R₂ e R₃, percebemos que a variação destes, bem com a identidade dos haletos, influi sobre os valores de GATS7e, mas não apresenta uma relação direta com a atividade biológica. Entre os haletos percebemos que os que apresentam maior atividade biológica são Br e I, embora seja a presença de Cl na molécula que contribua para que as moléculas tenham valores maiores de GATS7e.

A tabela 6 abaixo resume as observações feitas com relação à influência dos substituintes sobre o valor dos descritores e relação com a atividade biológica. Foram organizados os substituintes de acordo com os valores que as moléculas apresentam para cada descritor quando estes estão presentes, em ordem crescente. Também foi colocada uma coluna para evidenciar qual característica para o valor do descritor contribui para melhora da atividade biológica da molécula, se quanto maior melhor ou quanto menor melhor. Assim podemos analisar quais as características que contribuem para que as moléculas tenham uma atividade biológica maior e quais as regiões que mais influenciam para isto.

Pela tabela 6, os substituintes que teriam maior contribuição para melhora da atividade biológica seriam, para o HOMO, os substituintes 3-Am e 2 haletos; Para nNH₂, 4-Am ou 3-Am, sem influência dos substituintes em R₂ e R₃; para Hy, 4-Am ou 3-Am, e 1 haleto; para JGI1, 4-Am ou 3-Am, e 2 haletos; para MATS1e, 3-Am ou 4-Am, e 2 haletos; e para GATS7e,

3-Am e 2 haletos. Tendo essas características como as melhores para cada descritor, a configuração mais indicada para que uma molécula tenha boa atividade biológica seria, 3-Am e 2 haletos.

Tabela 6. Influência dos substituintes sobre o valor dos descritores em ordem crescente dos valores correspondentes a cada substituinte, e qual valor contribui para que a atividade biológica seja melhor.

Ordem crescente dos valores dos descritores para os substituintes			
	Para R ₁ e R ₄	Para R ₂ e R ₃	Melhor atividade biológica
HOMO	3-Am < 4-Am < 3-AmOH < 3-AmOMe	2 haletos < 1 haleto < OMe	Menor
nNH₂	3-AmOMe = 3-AmOH < 4-Am = 3-Am	Não tem influência	Maior
Hy	3-AmOMe < 3-AmOH < 4-Am = 3-Am	OMe < 2 haleto < 1 haletos	Maior
JGI1	3-AmOMe < 3-AmOH < 4-Am = 3-Am	1 haleto < OMe < 2 haletos	Maior
MATS1e	3-Am = 4-Am < 3-AmOH < 3-AmOMe	2 haletos < 1 haleto < OMe	Menor
GATS7e	3-AmOMe < 3-AmOH < 4-Am < 3-Am	OMe < 1 haleto < 2 haletos	Maior

4.3. Análise HCA

4.3.1. Com autoescalamento

O método HCA foi utilizado para agrupar as moléculas de acordo com suas similaridades. Utilizando os descritores como parâmetro para avaliar a similaridade entre as moléculas, as moléculas são agrupadas conforme o grau de similaridade entre elas. Em um grau de similaridade zero, todas as moléculas fazem parte de um único grupo, mas à medida em que o grau de similaridade aumenta, ocorre a separação em mais grupos. Como mostram as figuras 15 e 16, o método HCA foi eficiente para agrupar as moléculas, fornecendo um resultado muito parecido com o do método PCA.

Adotamos o autoescalamento dos dados e dois tipos de abordagem para agrupamento das moléculas, centroide e single. Comparando as duas abordagens, percebemos que ambas conseguem separar as moléculas em grupos de maneira análoga ao método PCA, mas observamos que a abordagem centroide é mais eficiente em separar as moléculas em dois grupos, ativas e menos ativas. A abordagem centroide consegue separar as moléculas ativas das menos ativas já com um grau de similaridade menor que a abordagem single. Em ambas as abordagens pode-se visualizar a separação entre os quatro grupos (Ativas, G1, G2 e G3), quando adotado o grau de similaridade 0,5. A abordagem single consegue separar os grupos os grupos utilizando um grau de similaridade um pouco maior do que a abordagem centroide, mas a abordagem centroide tem uma separação mais nítida entre menos ativas e ativas.

A determinação de qual abordagem é mais vantajosa varia de acordo com a finalidade do emprego do método e o grau de similaridade com que se quer avaliar as moléculas. A abordagem centroide é melhor para uma separação entre moléculas ativas e menos ativas, já a abordagem single é um pouco melhor para separar os grupos de moléculas menos ativas com um grau mais elevado de similaridade. No caso do nosso estudo, a separação entre os grupos é mais fácil de ser visualizada utilizando a abordagem centroide. Na tabela 7 estão os grupos e as respectivas moléculas que os compõem.

Tabela 7. Discriminação das moléculas pertencentes a cada grupo formado pelo método HCA.

Grupo	Moléculas
Ativos	M49, M55, M46, M36, M52, M39, M33, M43, M19, M26, M22, M42, M18, M31, M25, M61, M32, M14, M62, M28, M27, M58
G1	M9, M15, M1, M8, M11, M6, M10, M47, M4, M50, M37, M20, M23, M44, M56, M34, M40, M53, M59, M29
G2	M21, M35, M51, M41, M24, M45, M48, M57, M38, M54, M60, M30
G3	M5, M2, M3, M12, M7, M13

*As moléculas discrepantes estão inclusas no grupo G1.

Figura 15. Análise HCA utilizando a abordagem single para agrupamento das moléculas pelo grau de similaridade.

Figura 16. Análise HCA utilizando a abordagem centroide para agrupamento das moléculas pelo grau de similaridade.

O método HCA é utilizado apenas para classificação das moléculas em grupos de acordo com o grau de similaridade, não podendo ser empregado para predição de classificação. Por isso não há como fazer uma validação do modelo. Para avaliar se o modelo é válido, fizemos a análise da estrutura das moléculas que compõem cada grupo e avaliamos se são similares ou não. No decorrer de nosso estudo também realizaremos uma análise utilizando o método PCA, a correspondência entre os dois métodos poderá dizer se o método HCA é válido ou não. Apesar do método HCA utilizar 100% das informações fornecidas pelos descritores e o método PCA utilizar as PCs, que representam apenas uma porcentagem da informação dos descritores, se os modelos forem válidos haverá uma correspondência entre eles (Kenneth R. Beeber, 1998).

4.3.2. Com centralização na média

Com o tratamento dos dados por centralização na média, temos os resultados que estão reportados nas Figuras 17 e 18 e os agrupamentos das moléculas estão na tabela 8. Como no método utilizando autoescalamamento dos dados, temos a separação em grupo mais ativo e os três subgrupos menos ativos nas duas abordagens de ligação, centroide e single. Na

abordagem single o grupo de moléculas ativas inclui apenas as que realmente são ativas (Figura 17), mas na abordagem centroide o grupo de moléculas ativas incluiu também a molécula M11 (Figura 18). A molécula M11 deveria estar segregada no subgrupo menos ativo G1, como reportado pelas outras três abordagens. A separação dos grupos se dá com similaridade em 0,418 para o método single e 0,539 para o método centroide. Apesar dos grupos serem separados com maior similaridade com a abordagem centroide, a separação é melhor com a abordagem single.

As moléculas denominadas com as mais discrepantes de todo o conjunto em estudo (M9, M15, M1, M8, M6 e M4), foram agrupadas juntamente com as moléculas do subgrupo menos ativo G3, mas na abordagem com autoescalamiento essas moléculas não foram agrupadas adequadamente. Os agrupamentos são melhores visualizados no método com centralização na média do que com autoescalamiento.

Tabela 8. Agrupamento das moléculas realizado pelo método HCA com tratamento dos dados por centralização na média.

Grupo	Moléculas
Ativos	M49, M55, M46, M36, M52, M39, M33, M43, M19, M26, M22, M42, M18, M31, M25, M61, M32, M14, M62, M28, M27, M58
G1	M15, M8, M11, M47, M50, M37, M20, M23, M44, M56, M34, M40, M53, M59, M29
G2	M21, M35, M51, M41, M24, M45, M48, M57, M38, M54, M60, M30
G3	M5, M9, M1, M2, M12, M3, M8, M11, M6, M10, M7, M13, M4

Figura 17. Análise pelo método HCA utilizando centralização na média para os dados e abordagem single para ligação.

Figura 18. Análise pelo método HCA utilizando centralização na média para os dados e abordagem centróide para ligação.

A diferença observada entre os resultados com centralização na média e autoescalamento está relacionada à variância do conjunto de dados após o tratamento. Na tabela 9 podemos perceber a diferença entre a variância dos descritores após o tratamento de cada método. Com a centralização na média percebemos que a variância nos descritores é menor que com autoescalamento, mas a variância entre os descritores com autoescalamento é mais homogênea que com centralização na média. Assim percebemos que pelo autoescalamento torna menor a influência de um único descritor sobre o modelo, mas a centralização na média permite que um ou mais descritores tenham maior influência sobre o modelo (João Paulo A. Martins, 2013).

Tabela 9. Variância dos descritores após o tratamento por autoescalamento e centralização na média.

Centrado na média	Variância	Autoescalado	Variância
GATS7e	0.000217	GATS7e	0.977786
Homo	0.758865	Homo	0.993628
Hy	0.934492	Hy	0.987305
JGI1	0.015025	JGI1	0.988229
MATS1e	0.000361	MATS1e	0.969144
nNH2	0.327158	nNH2	0.980346

4.4. Análise ANN

Para determinar quais moléculas comporiam o conjunto de treinamento e o conjunto de testes foi desenvolvido um modelo utilizando redes neurais artificiais. A doutora Aline Alves de Oliveira escreveu um script que realiza várias redes neurais em sequência, utilizando agrupamentos aleatórios de moléculas para os conjuntos de teste, treinamento e validação, e reporta apenas os modelos que apresentam MSE (Mean Square Error – Erro quadrático médio) abaixo do limite estabelecido (0,009). No script é predefinido o número de neurônios, as funções utilizadas para tratar as informações de entrada e a quantidade de moléculas, em porcentagem, que irão compor os conjuntos de treinamento, teste e validação.

Ao iniciar o processo de desenvolvimento de modelos, o script já foi configurado para reportar apenas os apresentassem MSE menor que 0,009, e cada modelo como arquivo individual. Para facilitar a avaliação dos modelos, o script também reporta um arquivo de texto com o nome do modelo gerado e os valores de MSE e R^2 do modelo com todas as moléculas, com apenas o conjunto de treinamento, apenas o conjunto de teste e apenas o conjunto de validação. Gerados os modelos, começamos a avalia-los quanto ao MSE, dando preferência aos que apresentam menor valor, e depois avaliamos os valores de R^2 obtidos, sendo mais interessantes os que apresentassem valores mais próximos de 1. No entanto, cabe ressaltar que avaliar apenas o valor de R^2 pode nos levar a escolher um modelo de maneira equivocada, visto que ele reporta apenas o ajuste dos pontos à curva que os descreve. É fundamental avaliar a distribuição gráfica dos pontos, a inclinação da curva e se os pontos selecionados para o conjunto de teste e validação tem boa representação das amostras selecionadas para o estudo. Em alguns dos modelos obtidos observamos que quase todos os pontos do conjunto de validação ou teste pertenciam ao grupo de moléculas menos ativas, tendenciando o modelo para este grupo e nos dando poucas informações sobre o grupo de moléculas ativas. Estes modelos foram descartados do estudo.

O modelo que apresentou melhores comportamento foi obtido utilizando 17 neurônios para o tratamento dos dados e função tansig, com 70% das moléculas para o conjunto treinamento, 20% para o teste e 10% para validação.

O conjunto todo foi tratado utilizando redes neurais e percebemos uma boa assertividade, comparando os resultados obtidos pelo modelo com os resultados experimentais, como expresso na figura 19. O valor de R^2 , 0,9456, é satisfatório, demonstrando um bom ajuste das moléculas à linha de tendência. Também percebemos que a inclinação da curva que descreve a distribuição das moléculas parece próxima de 45° , demonstrando que os resultados obtidos pelo modelo são próximos do obtido experimentalmente.

O modelo foi desenvolvido utilizando o conjunto de treinamento e reportou bons resultados, como observado na figura 20. O valor de R^2 , 0,9337, demonstra um bom ajuste à linha de tendência, e a inclinação, que os valores calculados são próximos aos valores normalizados reportados experimentalmente.

Para avaliar o modelo é realizada a validação interna (figura 21), que chamamos de validação, e a validação externa (figura 22), que chamamos de teste. O grupo de validação é um grupo moléculas que utilizado na confecção do modelo e o grupo de teste é um conjunto de moléculas que não foi utilizado na confecção do modelo. Os valores de R^2 obtidos na validação interna e externa são maiores que 0,95 e a inclinação da curva também é satisfatória, demonstrando que a rede gerada consegue reportar valores próximos ao valor experimental normalizado. O bom desempenho na validação demonstra que o modelo consegue avaliar o conjunto de dados fornecido e reportar um valor que corresponde ao esperado, o valor experimental normalizado, mas isso ainda não comprova que a rede aprendeu a fazer o tratamento dos dados de entrada. Existe a possibilidade de que o modelo gerado tenha apenas “decorado” os valores de saída correspondentes aos de entrada. Para avaliar se o modelo criado realmente aprendeu a tratar adequadamente o conjunto de dados fornecidos, é feita a validação externa utilizando o conjunto de teste. O bom desempenho do conjunto de teste comprova que o modelo realmente aprendeu a tratar os dados de entrada e reporta um resultado próximo ao valor experimental normalizado.

Figura 19. Análise do conjunto completo de moléculas em estudo pelo método ANN.

Figura 20. Análise do conjunto de treinamento para gerar um modelo de regressão por ANN e ensinar a rede como analisar os dados fornecidos.

Figura 21. Análise do grupo de moléculas utilizadas para validação interna do método de regressão gerado por ANN e verificar se a rede aprendeu a avaliar os dados fornecidos.

Figura 22. Avaliação do conjunto de teste, moléculas que não foram utilizadas para geração do modelo de regressão por ANN, para avaliar se a rede realmente aprendeu a avaliar os resultados ou apenas decorou os resultados que deveriam ser reportados.

A análise ANN não foi utilizada somente para desenvolver um modelo que possa prever o valor da atividade biológica de acordo com as informações de entrada fornecidas, mas também para separar as moléculas em estudo em dois conjuntos, o de treinamento e o de testes. Estes dois conjuntos serão utilizados para validação dos modelos a serem confeccionados utilizando outros métodos. O conjunto de treinamento será utilizado para elaboração do modelo e o conjunto de teste para avaliar se o modelo relaciona adequadamente as informações fornecidas pelos descritores com a atividade biológica. As moléculas selecionadas para o conjunto de teste são: M54, M24, M56, M44, M21, M4, M28, M62, M26, M43, M33 e M36.

4.5. Análise PCA

4.5.1. Com autoescalamento

Obtidos os conjuntos de treinamento e teste pelo método ANN, podemos analisar o conjunto de dados com o método PCA. Na figura 23 temos a distribuição das moléculas utilizando as PCs 1 e 2. Ao analisarmos o gráfico, percebemos a separação do conjunto em 4 grupos, um sendo composto pela moléculas mais ativas e os outros três pelas moléculas menos ativas do conjunto. Esta separação entre ativos e menos ativos, já satisfaz nossa pretensão inicial, mas intentamos ainda extrair o máximo de informações sobre o conjunto de moléculas em estudo e averiguar a profundidade alcançada pelo modelo.

Ao analisarmos os grupos formados, percebemos que o grupo das moléculas mais ativas é formado apenas pelas moléculas que tem o substituinte Am nas posições R₁ e R₄. O grupo menos ativo G3 é formado por moléculas que tenha o substituinte NO₂ na posição R₁ ou R₄. O grupo menos ativo G2 é formado pelas moléculas que tenha nas posições R₁ e R₄ o substituinte AmOMe. Já o grupo menos ativo G1 é formado pelas moléculas que tenham em R₁ e R₄ o substituinte AmOH e também as demais moléculas com outros substituintes nestas posições, diferentes dos já citados. No grupo dos menos ativos G1, as moléculas que tem substituintes diferentes de AmOH estão mais afastadas do centro deste grupo.

Estas separações demonstram que os descritores selecionados representam adequadamente o conjunto de moléculas em estudos e as PCs 1 e 2 conseguiram extrair boa parte da informação contida neles. Apesar de observarmos no grupo menos ativo G1 moléculas não tão similares estruturalmente, vemos que os valores dos descritores são próximos. Se houvessem mais moléculas similares a estas do grupo menos ativo G1, com substituintes diferentes de AmOH, é provável que ocorreria a segregação de um novo grupo de moléculas menos ativas.

As PCs 1 e 2 representam 84,47% da variação de dados presente nos 6 descritores utilizados para obter o modelo. A PC 1 tem maior parte da representação dos dados, com 66,21%, e a PC 2 representa 17,26% dos dados fornecidos pelos descritores. Apesar da representação não ser tão elevada quanto gostaríamos, maior que 90%, ainda assim, foi

suficiente para classificar satisfatoriamente as moléculas em grupos (Kenneth R. Beeber, 1998).

Figura 23. Gráfico de Scores obtido utilizando o conjunto de treinamento, onde podemos visualizar a disposição das moléculas de acordo com os dados reportados pelas componentes principais (PC) 1 e 2, e visualizar o agrupamento das moléculas.

Retomando os resultados apresentados pelo método HCA na figura 16 ou figura 17, e comparando-os com o apresentado pelo método PCA na figura 23, vemos que há uma similaridade entre eles. Foram formados 4 grupos em ambos os métodos e as moléculas que compõem cada grupo são as mesmas nos dois métodos. Esta correspondência de resultados valida o método HCA empregado para classificação das moléculas em grupos.

A figura 24 mostra a contribuição de cada descritor para as PCs 1 e 2. O descritor HOMO tem maior contribuição na PC2 do que os demais descritores e Hy tem a maior contribuição para a PC1. Os descritores HOMO e MATS1e são os que mais contribuem para a separação vertical das moléculas e os demais descritores contribuem majoritariamente para a separação horizontal das moléculas.

Podemos perceber que para separar as moléculas em ativas e menos ativas, os descritores de maior influência são $n\text{NH}_2$, Hy, GATS7e e JGI1, e para a separação entre os subgrupos dos menos ativos, os principais descritores são HOMO e MATS1e. Assim, caso tenhamos a intenção de apenas separar as moléculas entre ativas e menos ativas, sem muito detalhamento, e utilizando menos descritores, isso pode ser feito utilizando apenas os descritores $n\text{NH}_2$, Hy, GATS7e e JGI1. Como queremos extrair o máximo de informações possíveis sobre o conjunto de dados e os modelos quimiométricos, uma abordagem mais simplista não é de interesse.

Figura 24. Gráfico de Loadings, onde é reportada a contribuição de cada descritor para a formação das componentes principais 1 e 2.

Ao aplicarmos o modelo desenvolvido utilizando o conjunto de treinamento às moléculas do conjunto de testes, obtemos uma separação em dois grupos, o das mais ativas (vermelho) e o das menos ativas (azul). O gráfico da figura 25 mostra que houve boa separação entre os dois grupos ao classificar as moléculas. O grupo das moléculas menos ativas não possibilita observar a formação de subgrupos, pelo baixo número de moléculas utilizadas no teste. Caso o modelo seja aplicado em um grupo maior de moléculas, é provável que seja observado a formação do grupo mais ativos e o grupo menos ativo com os três subgrupos.

Figura 25. Gráfico de Scores obtido para o conjunto de testes, onde podemos observar a disposição das moléculas avaliadas pelo modelo PCA gerado pelo conjunto de treinamento.

4.5.2. Com centralização na média

A abordagem usando centralização na média para tratar os dados reporta resultados similares aos com a abordagem usando autoescalamento. No entanto, percebemos que a separação dos subgrupos é mais nítida na abordagem com centralização na média (Figura 26). A representatividade deste modelo, utilizando as PCs 1 e 2 é de 95,24% de toda a informação fornecida pelos descritores. Como a variância da abordagem com centralização na média é menor que a da autoescalado, temos que as moléculas apresentam valores mais próximos, assim a dispersão é menor e as moléculas dos grupos ficam mais próximas.

Figura 26. Gráfico obtido dos Scores do conjunto de treinamento com as PCs 1 e 2, usando centralização na média para tratamento dos dados.

Ao analisar o gráfico da figura 27, percebemos que os descritores Homo, JGI1 e MATS1e tem baixa contribuição para as PCs 1 e 2, sendo a contribuição majoritária dos descritores GATS7e, nNH2 e Hy. Na abordagem por autoescalamento (figura 24) todos os descritores tem alguma contribuição para as PCs, demonstrando que a abordagem utilizada para o tratamento dos dados influencia seriamente nos resultados a serem obtidos.

Figura 27. Gráfico de Loadings dos descritores para as componentes principais 1 e 2.

O modelo gerado pelo método PCA com centralização na média foi aplicado ao conjunto de teste para validá-lo. O resultado obtido está demonstrado na Figura 28, onde podemos ver a separação das moléculas em dois grupos, mais ativos e menos ativos. A segregação dos subgrupos dos menos ativos não é possível de ser visualizada, devido à baixa quantidade de moléculas no conjunto de testes, mas a separação entre as mais ativas e as menos ativas já é suficiente para validade do modelo.

Figura 28. Gráfico de Scores obtido pela aplicação do modelo ao conjunto de testes, com tratamento dos dados por centralização na média.

4.6. Análise KNN

4.6.1. Com autoescalamiento

Pelo método KNN podemos separar as moléculas em classes usando os votos de seus vizinhos mais próximos. Utilizamos duas abordagens, uma processando os dados por autoescalamiento e outra utilizando centralização na média. Geralmente o autoescalamiento é mais recomendado, como disseram João Paulo e Márcia (João Paulo A. Martins, 2013), mas a centralização na média também pode ser utilizada e fornecer bons resultados.

Nossa primeira abordagem foi utilizando o processamento dos dados por autoescalamiento. Como podemos observar na Figura 26, houve uma boa classificação das moléculas em classes, estando as moléculas mais ativas localizados no quadrante superior esquerdo e os menos ativas no quadrante inferior direito. Nos chama a atenção que as moléculas M11, M61 e M32 estão deslocadas da classe em que deveriam estar alocadas. As moléculas M11 e M32 estão no quadrante superior direito e a M61 no quadrante inferior direito. Esse deslocamento pode ter ocorrido devido às três moléculas terem valor de atividade biológica próxima à faixa limite que separa as moléculas em mais ativas e menos ativas. Como este método não avalia a atividade biológica, mas as propriedades que selecionamos, temos uma indicação de que as características destas moléculas tornam difíceis suas classificações. Provavelmente há um número parecido de votos para classificá-las nos dois grupos.

Para avaliar se o número de vizinhos altera a distribuição das moléculas nas classes, o modelo foi gerado utilizando votações de 1 a 10 vizinhos. Independente do número de vizinhos votantes, a classificação das moléculas era sempre a mesma, como apresentado no gráfico da figura 29. Isto mostra que, entre as moléculas com valor de atividade biológica próximo há uma boa similaridade, utilizando os descritores selecionados.

O número de vizinhos recomendado pelo programa sempre foi 1, independente de quantos vizinhos fossem configurados para participarem da votação. Assim vemos que há uma boa separação entre os dois grupos principais, molécula ativas e inativas, visto não serem necessários muitos vizinhos para a votação (Kenneth R. Beeber, 1998). Também percebemos que o método KNN com processamento de autoescalamiento para este

conjunto de dados, não consegue classificar as moléculas em grupos como visto nos métodos PCA e HCA, mas apenas em dois principais. Isso nos leva a considerar que o uso destes três métodos em conjunto seria mais interessante que utilizá-los separadamente, pois suas informações se complementam.

Figura 29. Separação em classes usando processamento de autoescalamento.

Para avaliar o modelo gerado pelo método KNN com processamento de autoescalamento, o método foi aplicado ao conjunto de teste. O resultado está expresso no gráfico da Figura 30, onde observamos a classificação das moléculas como previsto, exceto por M28. A molécula M28 ficou um pouco abaixo do eixo de separação, mas não é classificado como menos ativo. Isso ocorre, pois esta molécula está na transição de classificação de mais ativo para menos ativo. Este resultado demonstra a validade do modelo gerado por KNN, sendo capaz de classificar adequadamente um conjunto de moléculas similares às utilizadas no desenvolvimento do modelo, mas externas a este.

Figura 30. Predição de classe para o conjunto de teste usando processamento de autoescalamento.

4.6.2. Com centralização na média

A segunda abordagem foi utilizando o método KNN com processamento de centralização na média. Na Figura 31 podemos perceber que o modelo teve dificuldade em classificar apenas as moléculas M11 e M61, provavelmente pelos mesmos motivos que a abordagem com autoescalamento, mas a molécula M32 foi classificada adequadamente. Devemos ressaltar também que a M61 foi classificada como menos ativa na abordagem com autoescalamento, mas na abordagem com centralização na média ela não é classificada em nenhuma das duas. Isso pode ser visto como uma melhora se partirmos do ponto que o modelo a classificou em uma classe intermediária em vez de classifica-la na classe errada.

Algo muito importante que aparece na Figura 31 são os grupos G1, G2 e G3, observados nos métodos anteriores, PCA e HCA. Isso mostra que essa abordagem do KNN com centralização na média consegue, além de classificar as moléculas em duas classes principais (mais ativos e menos ativos), separar os subgrupos presentes na classe dos menos ativos. A

abordagem utilizando o autoescalamiento, no entanto, consegue apenas a classificação das moléculas em duas classes, sem a separação dos subgrupos da classe dos menos ativos.

Sobre as duas abordagens, sabemos que no processamento de autoescalamiento, de cada elemento da variável é subtraída a média e o resultado é dividido pelo desvio padrão e no processamento por centralização na média, do valor de cada elemento da variável é apenas subtraído o valor da média. Como o desvio padrão das variáveis é menor que 1, quando realizamos o autoescalamiento o valor resultante é maior do que o valor da diferença entre a variável e a média. Pelo que observamos, a divisão pelo desvio padrão desfavorece a formação dos subgrupos, e isso deve ocorrer, pois algumas moléculas do subgrupo devem apresentar desvios padrões diferentes das demais moléculas do subgrupo e a divisão pelo desvio padrão potencializa essa diferença, aproximando a molécula de outro subgrupo.

Figura 31. Separação das moléculas em classes, de acordo com a votação dos vizinhos mais próximos com processamento centralizado na média.

Quando observamos a Figura 32, percebemos a separação em duas classes, mais ativos e menos ativos, e também a separação dos subgrupos menos ativos. Isso demonstra que o modelo consegue avaliar as moléculas do conjunto de teste de maneira satisfatória e reforça a validade do modelo.

Figura 32. Aplicação do modelo KNN com processamento centralizado na média às moléculas do conjunto de testes.

4.7. SIMCA

O método SIMCA foi aplicado ao conjunto de treinamento da mesma maneira que os métodos anteriormente apresentados, utilizando duas abordagens diferentes para o pré-tratamento dos dados.

4.8. Com Autoescalamento

Os resultados apresentados pelo método SIMCA com autoescalamento são parecidos com os reportados pelos métodos PCA e KNN. Como podemos ver no gráfico de projeções (Figura 33), as moléculas estão dispostas de forma razoavelmente dispersa, contidas em hipercaixas com volumes grandes. A exceção são as moléculas M9, M10, M11, M32 e M61, que não estão contidas dentro das hipercaixas nas quais deveriam estar. Isso mostra que, para essas moléculas em questão, o resultado não tem boa confiabilidade. Ao analisarmos a estrutura dessas moléculas, não percebemos relações que evidenciassem o motivo delas não estarem contidas nas hipercaixas. No método KNN também foi observada a mesma dificuldade do método para classificar as moléculas M11, M32 e M61, isso pode indicar que deve haver outras informações a serem consideradas sobre as estruturas das moléculas que melhorem a classificação do conjunto em estudo.

Figura 33. Gráfico de projeção das classes com método SIMCA com autoescalamento, obtido pelo programa Pirouette.

Ao observarmos o gráfico de projeções da figura 33, a separação entre as classes parece razoável, mas comparando com o gráfico de distância entre as classes (Figura 34), percebemos que não há uma boa separação entre elas. Assim, pela proximidade entre as classes, é possível que ocorram erros de classificação, principalmente para as moléculas do grupo menos ativo que se encontram no quadrante inferior esquerdo do gráfico da figura 33.

Na tabela 10, temos o poder discriminante e o poder total de modelagem dos descritores em nosso modelo SIMCA. O descritor de energia do orbital HOMO apresentou maior poder discriminante e poder total de modelagem que os demais descritores, indicando que as contribuições desse descritor para o modelo são muito importantes. Como o descritor HOMO é o principal contribuinte para a classificação das moléculas no modelo, é possível que a proximidade entre as classes e a alocação das moléculas M9, M10, M11, M32 e M61 estejam relacionadas aos valores de energia do orbital HOMO das moléculas.

Tabela 10. Poder discriminante e poder total de modelagem dos descritores para o modelo SIMCA com autoescalamiento gerado utilizando o conjunto de treinamento.

Descritores	Poder Discriminante	Poder Total de Modelagem
Homo	3.889518	0.678887
nNH ₂	1.240615	0.277302
Hy	1.030425	0.516272
MATS1e	1.929076	0.634916
JGI1	1.389867	0.447745
GATS7e	1.882483	0.405314

Figura 34. Gráfico de distância entre as classes, classificado pelo método SIMCA com autoescalamiento.

Nas Figuras 35 e 36, temos os gráficos de detecção de outliers e neles foi constatado que as moléculas M3, M6, M11, M32 e M11 são prováveis outliers. Para averiguar se essas moléculas são ou não outliers, foi confeccionado um novo modelo SIMCA com autoescalamiento sem essas moléculas no conjunto de treinamento. Não foi apresentada significativa melhora nos resultados sem as moléculas detectadas como prováveis outliers que justificasse a exclusão permanente delas e as confirmasse como outliers. As classes continuaram muito próximas, com sobreposição de algumas moléculas e no gráfico de projeções ainda haviam moléculas fora das hipercaixas em que deveriam estar contidas.

Figura 35. Gráfico de detecção de outliers das moléculas menos ativas pelo método SIMCA com autoescalamiento.

Figura 36. Gráfico de detecção de outliers das moléculas mais ativas pelo método SIMCA com autoescalamiento.

Para verificar a validade do modelo gerado, o aplicamos ao conjunto de testes. Para o conjunto de testes a distância entre as classes foi melhor que a apresentada pelo conjunto de treinamento (Figura 37). A distância não é muito elevada, mas não há sobreposição dos grupos, o que é satisfatório para a classificação do conjunto de teste.

Na figura 38 temos a probabilidade de predição das classes das moléculas, onde observamos a segregação das moléculas em duas classes, as mais ativas à direita e as menos ativas à esquerda. Apesar do conjunto de treinamento apresentar uma sobreposição entre as classes, a aplicação ao conjunto de teste foi satisfatória.

Figura 37. Gráfico de distância entre as classes do método SIMCA com autoescalamento aplicado ao conjunto de teste.

Figura 38. Gráfico de probabilidade de classificação das moléculas do conjunto de teste em classes.

4.9. Com Centralização na Média

Ao analisarmos o gráfico de projeções do modelo obtido pelo método SIMCA com centralização na média para o conjunto de treinamento (Figura 39), percebemos que a hipercaixa do grupo de moléculas classificadas como mais ativas tem volume menor que o mesmo grupo para o modelo SIMCA com autoescalamiento. As moléculas que ficaram fora das hipercaixas nesse modelo foram M5, M10, M11 e M61, indicando que o modelo não conseguiu classifica-las com tanta confiabilidade quanto às demais. Observamos a formação de alguns subgrupos dentro do grupo de moléculas menos ativas, fato não observado com o método usando autoescalamiento, no qual as moléculas estão mais dispersas dentro da hipercaixa do grupo menos ativo.

Figura 39. Gráfico de projeção das classes com método SIMCA com centralização na média, obtido pelo programa Pirouette.

Quando comparamos o resultado da tabela 11 com a tabela 10, vemos uma grande discrepância entre os valores de poder discriminante e poder total de modelagem para as duas abordagens, autoescalamiento e centralização na média. A principal diferença é a

ordem dos descritores como contribuintes principais para os poderes discriminante e total de modelagem, pois com autoescalamiento a energia do orbital HOMO é o descritor com maior contribuição e $n\text{NH}_2$ e Hy tem menor contribuição, e com a centralização na média $n\text{NH}_2$ e Hy são os contribuintes principais com valores muito maiores que a energia do orbital HOMO, que passa a ser um dos descritores de menor contribuição. Logo, temos que a escolha da abordagem para o pré-tratamento das moléculas leva a modelos consideravelmente distintos.

Tabela 11. Poder discriminante e poder total de modelagem dos descritores para o modelo SIMCA com centralização na média gerado utilizando o conjunto de treinamento.

Descritores	Poder Discriminante	Poder Total de Modelagem
Homo	3.258397	0.256481
$n\text{NH}_2$	636.058655	0.967126
Hy	635.547485	0.974680
MATS1e	236.734818	0.965904
JGI1	2.503122	0.087275
GATS7e	72.002251	0.999591

A separação de classes nos mostra se a confiabilidade do modelo para classificar as moléculas é adequada, e na figura 40 temos o gráfico de separação entre as classes. Nesse gráfico vemos que as moléculas estão separadas em dois grupos e sem sobreposição de classes. A distância entre as classes não é muito elevada, mas é suficiente para que tenhamos confiabilidade no resultado da classificação dos grupos. A baixa distância entre as classes pode estar relacionada à alta similaridade entre as moléculas, visto que em algumas há apenas a diferenciação por um substituinte.

As figuras 41 e 42 nos mostram quais são os possíveis outliers para o modelo gerado. Para verificarmos se essas moléculas são realmente outliers, desenvolvemos um novo modelo excluindo-as do conjunto de treinamento. O novo modelo sem os possíveis outliers não apresentou melhora ou resultados que confirmassem que as moléculas em questão são outliers. O que podemos constatar é que são moléculas com valores de atividade biológica próximo ao valor de separação entre os grupos de moléculas mais ativas e menos ativas.

Figura 40. Gráfico de distância entre as classes, classificadas pelo método SIMCA com centralização na média.

Figura 41. Gráfico de detecção de outliers para as moléculas mais ativas pelo método SIMCA com centralização na média.

Figura 42. Gráfico de detecção de outliers para as moléculas menos ativas pelo método SIMCA com centralização na média.

Para verificar a validade do modelo gerado, aplicamos este às moléculas do conjunto de testes e o resultado está reportado nas figuras 43 e 44. Na figura 43 temos o gráfico de distância entre as classes, onde não vemos uma sobreposição nítida entre as classes. No entanto, um ponto a ser destacado é que a molécula M28 está localizada após a linha de corte entre as duas classes e na região da classe das moléculas menos ativas. Avaliando a molécula M28, vemos que o valor de sua atividade biológica é bem próximo ao valor de separação entre as duas classes, logo suas características estruturais possam levar o modelo a não conseguir classificá-la adequadamente.

Na figura 44 temos o gráfico de probabilidade de predição das classes das moléculas e nele vemos a classificação satisfatória de todas as moléculas, exceto da molécula M28, que foi classificada como menos ativa sendo ela mais ativa. Podemos notar que a localização da molécula M28 é muito próxima à linha limite entre as duas classes, o que pode indicar que o resultado não é tão confiável como os das demais moléculas mais ativas que estão mais distantes da linha limite. Como o modelo falhou em classificar a molécula M28, não podemos atestar sua validade, apesar de ter predito satisfatoriamente a classe das outras moléculas do conjunto de testes. Em contrapartida o modelo SIMCA com

Distância entre as classes

The scatter plot displays the distance between two classes. The Y-axis represents 'Classe 1' and the X-axis represents 'Classe 2'. A vertical line at x=0 is labeled 'M28'. The legend indicates that blue diamonds represent 'Mais Ativas' and red squares represent 'Menos Ativas'.

Classe 2	Classe 1	Grupo
0,00	0,020	Mais Ativas
0,00	0,018	Mais Ativas
0,00	0,017	Mais Ativas
0,00	0,012	Mais Ativas
0,00	0,017	M28
0,26	0,018	Menos Ativas
0,32	0,007	Menos Ativas
0,32	0,003	Menos Ativas
0,34	0,004	Menos Ativas
0,35	0,006	Menos Ativas

Probabilidade de predição de classe

Classe 1

Classe 2

◆ Mais Ativas

■ Menos Ativas

M28

76

4.10. Análise PLS

Com o método PLS buscamos obter um modelo que, utilizando as informações fornecidas pelos descritores, reporte o valor da atividade biológica de cada molécula o mais próximo possível do valor real. Os outros métodos são qualitativos, voltados à formação de grupos e qualificação das moléculas, mas o método PLS busca, por regressão, fornecer valores numéricos para a atividade biológica de cada molécula.

Ao analisarmos a figura 45, percebemos uma dispersão dos pontos próximos à linha de tendência, demonstrando que o modelo pode ser descrito por uma equação de 1º grau. Percebemos também uma separação entre os dois grupos, as moléculas mais ativas acima de 6,5 e as menos ativas abaixo de 6,5. Algumas moléculas parecem estar muito distantes da linha de tendência do modelo e existe a possibilidade de que sejam outliers, mas para esta avaliação será empregada a verificação exposta na figura 46.

Figura 45. Gráfico do modelo PLS gerado utilizando as moléculas do conjunto de treinamento.

Os resultados expressos na tabela 12 facilitam a análise da figura 45, pois estes são os resultados da validação do modelo PLS. O que chama mais atenção na tabela são os valores elevados de Q^2 e R^2 , pois demonstram que o modelo consegue calcular adequadamente os valores da atividade biológica das moléculas e depois prevê-los com boa precisão. A literatura relata que bons modelos de regressão para QSAR devem ter $R^2 > 0,6$ e $Q^2 > 0,5$ (Márcia M. C. Ferreira, 2011). Também temos que PRESS, SEP_{val} e SEP_{cal} indicam a

intensidade dos erros, tanto de predição como calibração, e quanto menores estes erros, melhor será o modelo.

Um efeito chamado “super ajuste” não evidenciado neste modelo. Este efeito ocorre quando é obtido um valor muito elevado de R^2 e Q^2 para o modelo, geralmente valores próximos a 1,0. Quando ocorre o “super ajuste” é provável que exista algum erro ou anomalia no conjunto de dados utilizados para desenvolvimento do modelo.

O modelo PLS gerado apresenta valores elevados de R^2 e Q^2 , e baixos valores para os erros, indicando que a regressão obtida avalia adequadamente o conjunto de moléculas em estudo. Podemos também utilizar os erros, na tabela 12, como parâmetros para avaliar a qualidade dos descritores presentes no modelo. Os erros são razoavelmente baixos, não indicando o comprometimento do modelo ou alguma anomalia. Assim, pelos erros, percebemos que os descritores selecionados para o modelo tem boa correlação com a atividade biológica.

Tabela 12. Resultados da validação do modelo PLS do conjunto de treinamento.

Propriedades Calculadas	Resultados obtidos
Q^2	0.866287
R^2	0.880725
PRESS	14.69567
SEP _{val}	0.553317
SEP _{cal}	0.565444

Para verificar se há outliers no conjunto é realizada a análise de resíduos das moléculas. A figura 46 mostra o gráfico de resíduos pela alavancagem, onde temos que a distribuição das moléculas indica se são prováveis outliers. As moléculas dentro do quadrante principal ($y=[-2; 2]$ e $x=[0,0; 0,042]$) não são outliers, os que estiverem à direita do quadrante principal, mas ainda dentro do limite de resíduos ($y=[-2; 2]$), tem poucas chances de serem outliers. As moléculas fora das regiões relatadas tem uma maior chance de serem outliers.

Para avaliar se as moléculas são ou não outliers, é necessário retirá-las, uma de cada vez, do conjunto de treinamento e realizar a confecção de um novo modelo de regressão. Caso o novo modelo de regressão tenha resultados melhores que o modelo inicial, a molécula é um outlier e pode ser retirado do conjunto, mas se o novo modelo não apresentar melhora, a

molécula não é um outlier. A retirada de outliers é necessária, pois estes contribuem para que o modelo haja de maneira errônea, por isso quando retirados o modelo apresenta uma melhora nítida na avaliação dos dados e geração de resultados.

Do conjunto em estudo as moléculas mais propensas a serem outliers são M5, M59, M53, M46, M25, M27, M41 e M49. O teste para avaliar se as moléculas são outliers foi realizado e nenhuma delas foi considerada outlier. O modelo não apresentou melhora quando as moléculas citadas foram retiradas. Percebemos então que a contribuição destas moléculas para o modelo é válida, mas fica também evidenciado que o modelo apresenta dificuldades em tratar esse grupo de moléculas.

Avaliando as 8 moléculas suspeitas de serem outliers, não foi percebida nenhuma característica destoante das demais moléculas do conjunto em estudo. O que se percebe que é das 8 moléculas, 6 apresentam o substituinte R_2 igual a R_3 , mas isso não é significativo, visto que muitas outras moléculas com esta característica e que não foram evidenciadas como prováveis outliers.

Figura 46. Análise das moléculas do conjunto de treinamento utilizadas para obter o modelo PLS para detecção de Outliers por relação entre o Resíduo de Student e a Alavancagem.

Para verificar a validade do modelo, se realmente ele é capaz de avaliar as informações fornecidas pelos descritores e reportar o valor correto da atividade biológica de uma molécula, é empregada uma validação externa. Nesta validação externa utilizamos um conjunto de moléculas, chamado conjunto de testes, que não participaram da geração do modelo. Os resultados desta validação estão apresentados na figura 47 e na tabela 13.

Analisando a figura 47 juntamente com a tabela 13, podemos perceber que o modelo predisse adequadamente os valores de atividade biológica para as moléculas do conjunto de teste, visto que há um bom ajuste das moléculas à linha de tendência. Os valores de erro são baixos e o Q^2 é satisfatório, não sendo evidenciado “super ajuste” para este grupo de moléculas. Estas informações demonstram que o modelo gerado é válido.

Figura 47. Aplicação do modelo de regressão gerado por PLS ao conjunto de teste.

Tabela 13. Resultados da aplicação do modelo PLS no conjunto de teste.

Propriedades Calculadas	Resultados obtidos
Q^2	0.893533
PRESS	3.462688
SEP	0.537175

4.10.1. Teste de robustez do modelo PLS

No teste de robustez avaliamos o modelo quanto à retirada de moléculas do conjunto de treinamento. Podemos observar na figura 48 que o valor de Q^2 varia com a retirada das moléculas, tendo sido retiradas 20 moléculas do modelo correspondendo a até 42% do modelo, aproximadamente. Para obter um resultado mais confiável, as retiradas foram feitas em quintuplicata e reportada a média com o respectivo desvio.

O modelo gerado é mais robusto até a retirada de 10 moléculas, com a retirada de mais de 10 moléculas percebemos que o modelo oscila mais. Conforme o número de moléculas retiradas aumenta, o desvio também aumenta, demonstrando que a retirada se torna cada vez mais impactante. No entanto o modelo não chega a ter variações críticas. O valor de Q^2 oscila dentro dos limites reportados pela literatura para que o modelo seja tido como robusto.

A oscilação do valor de Q^2 do modelo oscila na faixa de 0,86 a 0,91, aproximadamente, considerando também os desvios. Os limites de oscilação são de 0,05 a -0,05 do valor de Q^2 , sendo a diferença entre os limites máximo e mínimo 0,10. Como a diferença entre 0,86 e 0,91 é menor que 0,10, podemos dizer que o modelo PLS gerado é robusto quanto à retirada de molécula.

A robustez do modelo evidencia o peso da contribuição das moléculas para gerar o modelo. Caso o modelo não seja robusto, isso indica que poucas moléculas tem uma contribuição alta no modelo e as demais uma contribuição reduzida, demonstrando que o modelo não representa o montante de moléculas, mas apenas algumas. Um modelo robusto é um modelo em que a contribuição de cada molécula para o modelo é significativa e assim o modelo representa todo o montante de moléculas do conjunto, ou quase todo.

A tendência de queda no valor de Q^2 quando mais moléculas são retiradas evidencia que a qualidade do modelo decai quando tiramos mais moléculas. Isso reforça o que foi discutido acima, mostrando que as moléculas tem contribuição significativa para o modelo.

Figura 48. Teste de robustez realizado por meio da retirada de moléculas do conjunto de treinamento utilizado para gerar o modelo PLS.

4.10.2. Teste de correlação ao acaso do modelo PLS

Para verificar se o modelo realmente representa o conjunto de dados gerado e não foi uma mera combinação de dados ao acaso, é realizado o teste de correlação ao acaso, dado pela aleatorização de y . Neste teste o valor atividade biológica é permutado entre as moléculas, aleatoriamente, e gerado um novo modelo. Caso o valor de Q^2 seja menor que 0,3 e de R^2 menor que 0,4 o modelo é verdadeiro. A princípio foi pensado que seria possível o surgimento de algum modelo acima dos limites estipulados pela literatura, pois algumas moléculas têm características estruturais muito similares e também valores de atividade biológica muito próximos. Felizmente esta hipótese falhou, como podemos ver ao analisarmos o gráfico da figura 49.

O gráfico da figura 49 foi obtido com os resultados dos modelos de 30 aleatorização de y realizadas e um sem aleatorização, para termos o modelo real expresso no gráfico. Ao analisarmos o gráfico, percebemos que todos os resultados das aleatorizações de y estão abaixo de 0,2 para Q^2 e abaixo de 0,3 para R^2 , demonstrando que não há um resultado com permutação aleatória da atividade biológica das moléculas que coloque em dúvida a veracidade do modelo. Isto refuta a possibilidade de que o modelo tenha sido gerado ao

acaso e demonstra que uma relação de particularidade entre os descritores selecionados e a atividade biológica da respectiva molécula.

5. Conclusão

Após a avaliação das características estruturais das moléculas e os valores reportados pelos descritores, percebemos que o substituinte Am, nas posições R₁ e R₄, é o que mais contribui para que as moléculas tenham uma elevada atividade biológica. Percebemos então que nestas posições a disponibilidade para realização de interações do tipo ligação de hidrogênio é um fator muito importante para que a molécula tenha atividade biológica elevada. As posições dos substituintes para R₁ e R₄ favorecem a atividade biológica se forem 4 e 3, respectivamente, demonstrando que pode haver um impedimento estérico em alguma etapa do processo até que a molécula esteja ligada ao sítio ativo, ou mesmo no sítio ativo.

A variação dos substituintes nas posições R₂ e R₃ não são tão influentes quanto em R₁ e R₄, mostrando que os substituintes de R₂ e R₃ atuam secundariamente, intensificando os efeitos dos substituintes das posições R₁ e R₄. Assim, precisamos saber quais os substituintes de R₁ e R₄ para que se possa dizer qual será o efeito dos substituintes de R₂ e R₃ sobre a atividade biológica. Devido a isso, até o momento, não há como afirmar qual configuração para R₂ e R₃ seria a mais significativa para que a atividade biológica fosse maior. O que podemos dizer é que na maioria dos casos, a presença de 2 haletos na molécula tem apresentado melhores valores de atividade biológica.

Os resultados obtidos pelos métodos quimiométricos utilizados confirma as considerações feita observando a relação dos substituintes com a atividade biológica. Os métodos HCA, PCA, KNN, SIMCA, PLS E ANN, apresentaram resultados satisfatórios e válidos, exceto pela abordagem SIMCA com centralização na média. Os resultados gerados por cada método são semelhantes e satisfatórios, visto que em todos foi possível a separação das classes mais ativa e menos ativa, assim um acaba por corroborar o outro.

Quanto à escolha do método de pré-tratamento dos dados, vimos claramente no método SIMCA que cada abordagem leva a um modelo diferente, apesar dos resultados muito próximos. Nos demais métodos qualitativos também percebemos a influência das abordagens de pré-tratamento, algumas em menor e outras e maior grau. Logo, a escolha da

abordagem para o pré-tratamento dos dados influencia relevantemente a geração dos modelos e os resultados gerados.

No método HCA percebemos a influência das abordagens de ligação single e centroide, sendo necessário avaliar qual a mais adequada para o estudo, pois a abordagem de pré-tratamento também influencia o resultado das duas abordagens de ligação.

Os métodos quantitativos, PLS e ANN, apresentaram resultados interessantes. Ambos apresentaram valores de ajuste das moléculas elevados, baixos resíduos e boa predição da atividade biológica das moléculas. Apesar de o método PLS utilizar um processamento linear e o ANN um processamento não linear, ambos foram satisfatório para o tratamento do conjunto em estudo. Os resultados de validação interna e externa do modelo ANN foram satisfatórios, com valor elevado de R^2 e baixos resíduos, demonstrando que a rede não decorou como processar os resultados, mas realmente aprendeu a fazê-lo. Os resultados obtidos pelo modelo PLS foram aprovados na validação externa e nos testes de robustez e obtenção de resultado ao acaso, estando dentro dos limites estabelecidos pela literatura.

Por fim vemos que o uso de vários métodos quimiométricos diferentes para análise do conjunto em estudo aumenta a confiabilidade dos resultados obtidos. Assim, os modelos gerados são válidos para testar novas moléculas para o tratamento da THA causada pelo tripanossoma brucei rhodesiense.

Bibliografia

- Acup, C., Bardosh, K. L., Picozzi, K., Waiswa, C., & Welburn, S. C. (2016). Factors influencing passive surveillance for T. b. rhodesiense human african trypanosomiasis in Uganda. *Acta Tropica*, 10.
- Athri, P., Wenzler, T., Ruiz, P., Brun, R., & Boykin, D. W. (2006). 3D QSAR on a library of heterocyclic diamidine derivatives. *Bioorganic & Medicinal Chemistry*, 3144–3152.
- Caleiro, J. P. (18 de julho de 2015). *Economia*. Acesso em 6 de novembro de 2017, disponível em Exame: <https://exame.abril.com.br/economia/os-15-paises-mais-pobres-do-mundo/>
- Cerqueira, E. O., Andrade, J. C., Poppi, R. J., & Mello, C. (2001). REDES NEURAIS E SUAS APLICAÇÕES EM CALIBRAÇÃO MULTIVARIADA. *Química Nova*, 864-873.
- Drugs for Neglected Diseases initiative. (s.d.). *Drugs for Neglected Diseases initiative*. Acesso em 5 de Novembro de 2017, disponível em Drugs for Neglected Diseases initiative: <http://www.dndial.org/>
- Ferreira, M. M. (2002). Multivariate QSAR. *Journal Brazilian Chemical Society*, 742-753.
- FRISCH, M. J., B., T. G., E., S. G., R., R. M., SCALMANI, G., BARONE, V., et al. (2009). Gaussian 09.
- Geary, R. C. (1954). The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician*, 115-127+129-146.
- Gemperline, P. (2006). *Practical Guide to Chemometrics*. Boca Raton: Taylor & Francis Group.
- J.M. Bouroche, G. S. (1982). *Análise de dados*. Rio de Janeiro: Zahar.
- João Paulo A. Martins, M. M. (12 de Março de 2013). QSAR Modeling: Um novo pacote computacional open source para gerar e validar modelos QSAR. *Química Nova*, pp. 554-560.
- Kenneth R. Beeber, R. J. (1998). *Chemometrics: A Practical Guide*. Wiley-Interscience.
- Khabnadideh, S., Pez, D., Musso, A., Brun, R., Pérez, L. M., González-Pacanowska, D., et al. (2005). Design, synthesis and evaluation of 2,4-diaminoquinazolines as inhibitors of trypanosomal and leishmanial dihydrofolate reductase. *Bioorganic & Medicinal Chemistry*, 2637–2649.
- Kiralj, R., & Ferreira, M. M. (2009). Basic Validation Procedures for Regression Models in QSAR and QSPR Studies: Theory and Application. *Journal Brazilian Chemical Society*, pp. 770 - 787.
- Kuang Hongyu, V. L. (2015). Análise de Componentes Principais: resumo teórico, aplicação e interpretação. *Engineering and Science*, pp. 83-90.
- Lipinski, C. F., Oliveira, A. A., Honorio, K. M., Oliveira, P. R., & da Silva, A. B. (22 de Fevereiro de 2018). A molecular modeling study of combretastatin-like chalcones as anticancer agents using PLS, ANN and consensus models. *Structural Chemistry*.
- Malhotra, N. K. (2001). *Pesquisa de Marketing: uma orientação aplicada*. Porto Alegre: Bookman.

- Márcia M. C. Ferreira, R. K. (2011). Métodos quimiométricos em relações quantitativas estrutura-atividade (QSAR). In: C. A. Montanari, *Química Medicinal: Métodos e Fundamentos em Planejamento de Fármacos* (pp. 387-452). São Paulo: Editora da Universidade de São Paulo.
- Ministério da Saúde, Departamento de Ciência e Tecnologia, Secretaria de Ciência, Tecnologia e Insumos Estratégicos. (2010). Doenças negligenciadas: estratégias do Ministério da Saúde. *Revista de Saúde Pública*, 200-202.
- Montanari, C. A. (2011). *QUÍMICA MEDICINAL: Métodos e Fundamentos em Planejamento de Fármacos*. São Paulo: Editora da Universidade de São Paulo.
- MORAN, P. A. (1950). NOTES ON CONTINUOUS STOCHASTIC PHENOMENA. *Biometrika*, 17-23.
- Morgon, N. H., & Coutinho, K. (2007). *Métodos de Química Teórica e Modelagem Molecular*. São Paulo : Livraria da Física.
- Patrick, D. A., Bakunov, S. A., Bakunova, S. M., Jones, S. K., Wenzler, T., Barszcz, T., et al. (2013). Synthesis and antiprotozoal activities of benzyl phenyl ether diamidine derivatives. *European Journal of Medicinal Chemistry*, 310-324.
- Pimentel, W. R. (2005). *TESE - APLICAÇÃO DE REDES NEURAIS ARTIFICIAIS E DE QUIMIOMETRIA NA MODELAGEM DO PROCESSO DE CRAQUEAMENTO CATALÍTICO FLUIDO*. Campinas: Faculdade de Engenharia Química - UNICAMP.
- Regazzi, A. J. (2001). *INF 766 - Análise multivariada*. Viçosa: Universidade Federal de Viçosa.
- Sabin, J. G., Ferrão, M. F., & Furtado, J. C. (Setembro de 2004). Análise multivariada aplicada na identificação de fármacos antidepressivos. Parte II: Análise por componentes principais (PCA) e o método de classificação SIMCA. *Revista Brasileira de Ciências Farmacêuticas*, 387-396.
- Valentin, J. L. (2000). *Ecologia numérica: uma introdução a análise multivariada de dados ecológicos*. Rio de Janeiro: Interciência.
- Valverde, R. (s.d.). *Doenças Negligenciadas*. Acesso em 5 de Novembro de 2017, disponível em Fundação Oswaldo Cruz: <https://agencia.fiocruz.br/doen%C3%A7as-negligenciadas>
- Von Zuben, F. J., & Attux, R. R. (s.d.). *Redes Neurais Artificiais – Parte I*.
- World Health Organization. (2010). *First WHO report on neglected tropical diseases: working to overcome the global impact of neglected tropical diseases*. World Health Organization.
- WORLD HEALTH ORGANIZATION. (s.d.). *Trypanosomiasis, African*. Acesso em 30 de junho de 2015, disponível em WORLD HEALTH ORGANIZATION: http://www.who.int/topics/trypanosomiasis_african/en/
- Zubrzycki, I. Z. (2002). Homology Modeling and Molecular Dynamics Study of NAD-Dependent Glycerol-3-Phosphate Dehydrogenase from *Trypanosoma brucei rhodesiense*, a Potential Target Enzyme for Anti-Sleeping Sickness Drug Development. *Biophysical Journal*, 2906–2915.