UNIVERSIDADE DE SÃO PAULO

FACULDADE DE CIÊNCIAS FARMACÊUTICAS DE RIBEIRÃO PRETO

**Emprego de ferramentas de quimioinformática no estudo do
perfil metabólico de plantas e na desreplicação de matrizes
vegetais**

Tiago Branquinho Oliveira

**Ribeirão Preto
2015**

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE CIÊNCIAS FARMACÊUTICAS DE RIBEIRÃO PRETO

**Emprego de ferramentas de quimioinformática no estudo
do perfil metabólico de plantas e na desreplicação de
matrizes vegetais**

Tese de doutorado apresentada ao Programa
de Pós-Graduação em Ciências Farmacêuti-
cas de Ribeirão Preto para obtenção do Título
de Doutor em Ciências

Área de Concentração: Produtos Naturais e
Sintéticos

Orientado: Tiago Branquinho Oliveira
Orientador: Prof. Dr. Fernando Batista da
Costa

Versão corrigida da Tese de Doutorado apresentada ao Programa de Pós-Graduação em
Ciências Farmacêuticas de Ribeirão Preto no dia 10/09/2015. A versão original encontra-se
disponível na Faculdade de Ciências Farmacêuticas de Ribeirão Preto/USP.

Ribeirão Preto

2015

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Oliveira, Tiago Branquinho

Emprego de ferramentas de quimioinformática no estudo do perfil metabólico de plantas e na desreplicação de matrizes vegetais. Ribeirão Preto, 2015.

128 p.; 30 cm.

Tese de Doutorado, apresentada à Faculdade de Ciências Farmacêuticas de Ribeirão Preto /USP - Área de concentração: Produtos Naturais e Sintéticos.

Orientador: Da Costa, Fernando Batista

1. Quimioinformática. 2. Farmacognosia. 3. Química de Produtos Naturais.

FOLHA DE APROVAÇÃO

Tiago Branquinho Oliveira

Emprego de ferramentas de quimioinformática no estudo do perfil metabólico de plantas e na desreplicação de matrizes vegetais

Tese de doutorado apresentada ao Programa de Pós-Graduação em Ciências Farmacêuticas para obtenção do Título de Doutor em Ciências

Área de Concentração: Produtos Naturais e Sintéticos

Orientado: Tiago Branquinho Oliveira

Orientador: Prof. Dr. Fernando Batista da Costa

Aprovado em:

Banca Examinadora

Prof. Dr. _____

Instituição: _____ Assinatura: _____

Prof. Dr. _____

Instituição: _____ Assinatura: _____

Prof. Dr. _____

Instituição: _____ Assinatura: _____

Prof. Dr. _____

Instituição: _____ Assinatura: _____

Prof. Dr. _____

Instituição: _____ Assinatura: _____

Este trabalho é dedicado a minha esposa, Daniella Silva Prado Branquinho, a qual incentivou-me, impulsionou-me e apoiou-me antes e durante todo o meu doutoramento. Aos meus pais, Elpídio de Oliveira e Angela Branquinho Oliveira, pelo incentivo e esforço despendido em minha educação. À minha irmã Thaís Branquinho Oliveira Fragelli e ao Ricardo Ramos Fragelli por serem exemplos na área da educação. E às minhas sobrinhas Luísa, Vanessa, Ana Clara e Alice.

AGRADECIMENTOS

Eu gostaria de expressar os meus sinceros agradecimentos:

Ao meu orientador, Prof. Dr. Fernando Batista da Costa, pela exemplar atitude profissional como orientador e incentivador de ideias.

Ao meu orientador do doutorado Sanduíche, Prof. Dr. Thomas J. Schmidt, pela supervisão e colaboração.

À Fundação de Amparo à Pesquisa de São Paulo (FAPESP) pela concessão da bolsa de estudos e reserva técnica no Brasil (2011/17860-7) e na Alemanha (2013/14239-5) que possibilitaram o desenvolvimento pleno deste projeto.

Ao Programa de Pós-Graduação em Ciências Farmacêuticas da Faculdade de Ciências Farmacêuticas de Ribeirão Preto - USP, pela infraestrutura oferecida, seus docentes e funcionários.

Ao Patrick Herbeth Guimarães Azevedo, ao Roberto Oliveira Da Silva e ao Silas Yudi Konno de Oliveira pela contribuição técnica, tecnológica e visual do processo de criação e finalização do AsterDB.

Aos que chegaram, passaram e saíram do AsterBioChem, em especial a Daniela Aparecida Chagas de Paula e a Rejane Barbosa de Oliveira.

Aos companheiros do Laboratório de Farmacognosia, pela amizade ao longo destes anos.

Ao Institut für Pharmazeutische Biologie und Phytochemie (IPBP-WWU), em especial ao Mauro Nogueira, Nuria Llurba, Jacqueline Dikti, Sabine Beckert e Marcus Lautenshläger.

“Tudo é uma questão de escala de tempo.”
(Carl Sagan)

RESUMO

Oliveira, T. B. **Emprego de ferramentas de quimioinformática no estudo do perfil metabólico de plantas e na desreplicação de matrizes vegetais**. 2015. 128 f. Tese (Doutorado). Faculdade de Ciências Farmacêuticas da Universidade de São Paulo, Ribeirão Preto, 2015.

Com o surgimento da era computacional com especial aplicação em química, as substâncias de origem naturais puderam ter suas informações armazenadas em bancos de dados. Desta forma, surge a oportunidade de se empregar bancos de dados de produtos naturais e de algumas ferramentas de quimioinformática como os estudos de *Quantitative Structure-Retention Relationship* (QSRR) para acelerar a identificação de substâncias em estudos metabolômicos. Este trabalho propôs o desenvolvimento de três estudos de QSRR, bem como a construção de um banco de dados (AsterDB) com estruturas químicas da família Asteraceae e informações a elas associadas (ex.: ocorrências botânicas e taxonômicas, atividade biológica, informações analíticas etc.) para auxiliar a desreplicação de substâncias em extratos vegetais. O primeiro estudo foi elaborado com 39 lactonas sesquiterpênicas (LST) analisadas em dois diferentes sistemas de solventes (MeOH–H₂O 55:45 e MeCN–H₂O 35:65), três grupos de descritores estruturais (2D-descr, 3D-1conf e 3D-weigh), dois diferentes conjuntos para treino e teste (26:13 e 29:10), quatro algoritmos para seleção de descritores (*best first*, *linear forward* - LFS, *greedy stepwise* e algoritmo genético - GA), três diferentes tamanhos de modelos (quatro, cinco e seis descritores) e dois métodos de modelagem (mínimos quadrados parciais - PLS e redes neurais artificiais - ANN). O segundo foi desenvolvido com 50 substâncias de diferentes classes químicas com intuito de avaliar as diferenças entre substâncias analisadas individualmente e em mistura em três diferentes equipamentos e dois métodos cromatográficos. O terceiro foi elaborado com 2.635 estruturas químicas com um teste externo comum a todos os modelos (25%, n = 656), três métodos de separação para teste e treino (partição baseada na resposta e baseada nos preditores 2D e 3D), três diferentes tamanhos de modelos selecionados por GA e dois métodos de modelagem (MLR e redes neurais *feed-forward* com regularização bayesiana - BRNN). O banco de dados AsterDB foi desenvolvido para ser preenchido de forma gradual e atualmente possui cerca de 2.000 estruturas químicas. O primeiro estudo de QSRR gerou bons modelos capazes de estimar o logaritmo do fator de retenção (*logk*) das LST com $P^2 > 0,81$ para o sistema MeCN–H₂O. O segundo estudo mostrou que não houve diferença estatística entre as substâncias analisadas individualmente e em mistura (p-valor > 0,95) e que a correlação entre os dois métodos cromatográficos e equipamentos utilizados foi reprodutível (R > 0,95). Estas análises mostraram que foi possível desenvolver modelos de QSRR para um método cromatográfico e equipamento e transpô-los para outro equipamento seguindo o uso de substâncias em comum. O terceiro estudo produziu modelos com boa capacidade de predição ($P^2 > 0,81$) utilizando alta amplitude de espaço químico e rigor estatístico. Conclui-se que, estas informações podem ser utilizadas como uma plataforma piloto para análises de dados com objetivo de auxiliar na desreplicação de extratos de plantas em estudos metabolômicos.

Palavras-chaves: Produtos Naturais. Quimioinformática. Banco de Dados. Estimar Tempo de Retenção.

ABSTRACT

Oliveira, T. B. **Application of chemoinformatic tools in the study of plant metabolic profiles and dereplication.** 2015. 128 f. Thesis (Doctoral). School of Pharmaceutical Sciences of Ribeirão Preto - University of São Paulo, Ribeirão Preto, 2015.

After the emergence of the computing era with special application in chemistry, all substances from natural sources might have their information stored in databases. Therefore, the opportunity arises to employ natural product databases and some chemoinformatic tools such as QSRR studies to speed up the identification of substances from metabolomic studies. This paper proposes the development of three QSRR studies as well as the building of a database (AsterDB) with chemical structures from the Asteraceae family and related information (i.e.: botanical and taxonomic occurrences, biological activity, analytical information, etc.) aiming to assist the dereplication of substances in plant extracts. The first study was carried out with 39 sesquiterpene lactones (STLs) analysed using two different solvent systems (MeOH–H₂O 55:45 and MeCN–H₂O 35:65), three groups of structural descriptors (2D-descr, 3D-1conf, and 3D-weight), two different sets for training and testing (26:13 and 29:10), four algorithms for selection of descriptors (best first, LFS, greedy stepwise, and GA), three different model sizes (four, five, and six descriptors) and two modelling methods (PLS and ANN). The second study was developed with 50 compounds of different chemical classification in order to assess the differences between individual and mixed compounds analysed in three different equipments and two chromatographic methods. The third was elaborated with 2,635 chemical structures with a common external test to all models (25%, n = 656), three separation methods for testing- and training-set (based on response and on 2D and 3D predictors partitions), three different sizes of models selected by GA and two modelling methods (MLR and BrNN). The AsterDB database was developed to be populated gradually and currently, it has about 2,000 chemical structures. The first QSRR study generated good models, able to estimate the logarithm of the retention factor (*logk*) of STLs with $P^2 > 0.81$ for the MeCN–H₂O system. The second study showed that there was no statistical difference between the substances analysed individually and mixed (*p-value* > 0.95) and the correlation between the two chromatographic methods and equipments used was reproducible (*R* > 0.95). These analyses showed that it was possible to develop QSRR models for a chromatographic method and equipment and translate them into other equipment following the use of substances in common. The third study produced models with good predictive capacity ($P^2 > 0.81$) using a high range of chemical space and statistical accuracy. In conclusion, this information can be used as a pilot platform for data analysis in order to assist in plant dereplication in metabolomics studies.

Key-words: Natural Products. Chemoinformatics. Cheminformatics. Database. QSRR.

LISTA DE ILUSTRAÇÕES

Figura 1 – Estruturas químicas das 39 lactonas sesquiterpênicas (LST).	35
Figura 2 – Cinquenta substâncias puras do AsterBioChem e IPBP analisadas neste estudo.	41
Figura 3 – Diagrama estrutural do banco de dados AsterDB.	51
Figura 4 – Gráficos cumulativos de participação dos fatores para cada solvente entre os melhores modelos. O eixo horizontal mostra o coeficiente de determinação da validação externa (P^2); o eixo vertical mostra a participação de cada fator (em percentagem). As maiores áreas (barras verticais coloridas) demonstram abordagens mais bem-sucedidas.	66
Figura 5 – Valores experimentais <i>versus</i> calculados de $\log k$ dos modelos MeOH140, MeOH143 e MeOH40 (Tabela 7, página 74). Modelos para o sistema de solventes MeOH–H ₂ O e as 39 LST combinadas nos grupos de treino e teste externo. Nas ampliações (regiões A - D marcadas em retângulos vermelhos) é possível observar a capacidade dos modelos discriminarem entre os epímeros e isômeros geométricos (pares de estruturas A - 10 e 11 , B - 29 e 30 , C - 13 e 14 , e D - 8 e 9).	67
Figura 6 – Valores experimentais <i>versus</i> calculados $\log k$ de MeCN137, MeCN143, MeCN45, MeCN47, MeCN96, e Consensus5 (Tabela 8, página 75). Modelos para o sistema de solventes MeCN–H ₂ O e as 39 LST combinadas dos grupos de treino e teste externo. Nas ampliações (regiões A-D marcadas em retângulos vermelhos) é possível observar a capacidade dos modelos discriminarem entre epímeros e isômeros geométricos (pares de estruturas A - 10 e 11 , B - 29 e 30 , C - 13 e 14 , e D - 8 e 9).	69
Figura 7 – Descritores selecionados nos melhores modelos. As cores indicam os sistemas de solventes (verde = MeOH–H ₂ O e vermelho = MeCN–H ₂ O). A distância de um descritor a partir do centro de cada gráfico se correlaciona com a sua importância (frequência de seleção por diversas técnicas utilizadas) e reflete o seu grau de correlação com o $\log k$	70

Figura 8 – Distribuição do t'_R de 50 substâncias em três diferentes análises. a) Método ABC1 no equipamento UHPLC-MS2 (D: análise individual das substâncias, E: mistura de 10 substâncias e, F: mistura de todas as substâncias disponíveis). b) Método ABC1 no equipamento UHPLC-MS2 (D: análise individual das substâncias, E: mistura de 10 substâncias e, F: mistura de todas as substâncias disponíveis). c) Método IPBP no equipamentos UHPLC-MS2 (G: análise individual das substâncias, H: mistura de 10 substâncias e, I: mistura de todas as substâncias disponíveis). A linha em cada gráfico representa a distribuição normal esperada. Dados: Tabelas 9 (página 77) e 10 (página 80).	85
Figura 9 – Valores de t'_R pelo método ABC1 <i>versus</i> t'_R obtidos pelo método IPBP para as 50 substâncias deste estudo no equipamento UHPLC-MS2 (Tabela 9, página 77).	88
Figura 10 – Valores de t'_R pelo método ABC1 obtidos pelo equipamento UHPLC-MS1 <i>versus</i> t'_R obtidos pelo método IPBP no equipamento UHPLC-MS2 das 43 substâncias deste estudo (Tabela 10, página 80).	89
Figura 11 – Gráfico da matriz de correlação. A parte triangular inferior indica a correlação de Pearson e a parte triangular superior é uma informação gráfica da matriz de correlação. As cores dos círculos indicam os sinais: positivo = azul; negativo = vermelho. O tamanho dos círculos indicam os valores absolutos da correlação. A - cinco descritores selecionados por LFS para os métodos ABC1 e IPBP (LFS5); B - seis descritores selecionados por LFS para os métodos ABC1 e IPBP (LFS6); C - cinco descritores selecionados por GA para os métodos ABC1 e IPBP (GA5); D - cinco descritores selecionados por GA para IPBP (GA5a); E - cinco descritores selecionados por GA para ABC1 (GA5b); F - seis descritores selecionados por GA para IPBP (GA6a); e G - seis descritores selecionados por GA para ABC1 (GA6b).	91
Figura 12 – Valores dos t'_R preditos <i>versus</i> experimentais de 50 substâncias para o método ABC1 calculados por PLS. Círculos mostram os valores das 34 substâncias utilizadas no treino e triângulos os valores estimados para o teste externo (16 substâncias).	93
Figura 13 – Valores dos t'_R preditos <i>versus</i> experimentais de 50 substâncias para o método IPBP calculados por PLS. Círculos mostram os valores das 34 substâncias utilizadas no treino e triângulos os valores estimados para o teste externo (16 substâncias).	94

Figura 14 – Valores do t'_R predito <i>versus</i> experimental de 50 substâncias para o método ABC1 calculados por GA-MLR. Círculos mostram os valores das 34 substâncias utilizadas para treino e triângulos os valores estimados para o teste externo (16 substâncias).	95
Figura 15 – Valores dos t'_R preditos <i>versus</i> experimentais de 50 substâncias para o método IPBP calculados por GA-MLR. Círculos mostram os valores das 34 substâncias utilizadas para treino e triângulo os valores estimados para o teste externo (16 substâncias).	96
Figura 16 – Análise linear discriminante (LDA), revelando os Klekota-Roth <i>Fingerprint Counts</i> (KRFPC) mais importantes para o agrupamento com apenas duas lineares discriminantes (LD1 e LD2).	97
Figura 17 – Distribuição das estruturas químicas pelo $t_R \times 100$ (RRT).	98
Figura 18 – Relação entre os descritores selecionados por diferentes métodos (tempo de retenção - prt, mapa auto organizável do tipo Kohonen com as estruturas 2D - som2D e estruturas 3D - som3D) e tipo (2D-descr, 3D-1conf e 3D-weigh). Descritores, métodos de separação (teste e treino) e tipos de descritores são representados por segmentos do círculo. A cada variável foi atribuída uma cor. Cada descritor recebeu uma cor correspondente à sua classificação e forma de seleção e sua largura indica a quantidade que foi selecionada.	99
Figura 19 – Gráficos cumulativos de participação dos fatores (em percentagem) entre os melhores modelos. O eixo horizontal mostra o coeficiente de determinação no teste externo (P^2); o eixo vertical mostra a participação de cada fator. As maiores áreas (barras verticais coloridas) demonstram as abordagens mais bem-sucedidas.	101
Figura 20 – Exemplos de estruturas presentes no banco de dados da <i>Analyticon Discovery</i> . Dentre as estruturas se encontram epímeros e estruturas com mesma massa molecular (isóbaros).	102
Figura 21 – Representação linear dos valores experimentais <i>versus</i> calculados dos RTTs dos modelos ANA12, ANA42, ANA36, CON1 e CON4. Os melhores modelos para cada tipo de descritor e modelos consensuais com P^2 igual a 0,85 e as 2.635 estruturas combinadas dos grupos de treino, validação externa e teste externo.	104

LISTA DE TABELAS

Tabela 1	– Exemplos de bancos de dados de interesse na área de produtos naturais.	23
Tabela 2	– Alguns exemplos de representação estrutural.	25
Tabela 3	– Classificação geral dos descritores estruturais.	27
Tabela 4	– Programas para calcular descritores estruturais.	28
Tabela 5	– Lista das LST avaliadas e seus logaritmos do fator de retenção nos sistemas MeOH–H ₂ O (55:45), $\log k_1$ e MeCN–H ₂ O (35:65), $\log k_2$.	32
Tabela 6	– Tabela de povoamento padrão da tabela <i>classification</i> .	53
Tabela 7	– Melhores modelos de QSRR para cada variável obtida para o sistema de solvente MeOH–H ₂ O. R ² : coeficiente de determinação; Q ² : coeficiente de determinação para a validação cruzada <i>10-folds</i> ; P ² : coeficiente de determinação para validação externa; MAE: média do erro absoluto; RMSE: raiz quadrada do erro quadrático médio. (Reimpresso e adaptado com permissão de Oliveira et al.(1); <i>copyright 2015 American Chemical Society</i>).	74
Tabela 8	– Melhores modelos de QSRR para cada variável para o sistema de solvente MeCN–H ₂ O. R ² : coeficiente de determinação; Q ² : coeficiente de determinação para a validação cruzada <i>10-folds</i> ; P ² : coeficiente de determinação para validação externa; MAE: média do erro absoluto; RMSE: raiz quadrada do erro quadrático médio. (Reimpresso e adaptado com permissão de Oliveira et al.(1); <i>copyright 2015 American Chemical Society</i>).	75
Tabela 9	– Tempos de retenção ajustado (t'_R) das substâncias analisadas no equipamento UHPLC-MS1.	77
Tabela 10	– Tempos de retenção ajustados (t'_R) das substâncias analisadas no UHPLC-MS2.	80
Tabela 11	– Tempos de retenção ajustados (t'_R) das substâncias analisadas no UPLC-DAD.	83
Tabela 12	– Análise de correlação entre os equipamentos avaliados utilizando o método ABC1.	86
Tabela 13	– Análise de correlação entre os equipamentos avaliados utilizando o método IPBP.	87
Tabela 14	– Modelos obtidos para 50 substâncias de origem natural utilizando descritores 2D e 3D. O conjunto de treino e teste foi de 34 e 16, respectivamente.	92

Tabela 15 – Resumo dos modelos de QSRR obtidos para todo o ambiente estrutural. R^2 : coeficiente de determinação; Q^2 : coeficiente de determinação para a validação externa; P^2 : coeficiente de determinação para teste externo; MAE: média do erro absoluto; RMSE: raiz quadrada do erro quadrático médio (Tabela completa no Apêndice B, página 121).	105
Tabela 16 – Melhores modelos de QSRR obtidos para todo o ambiente estrutural. R^2 : coeficiente de determinação; Q^2 : coeficiente de determinação para a validação externa; P^2 : coeficiente de determinação para teste externo; MAE: média do erro absoluto; RMSE: raiz quadrada do erro quadrático médio.	121

LISTA DE ABREVIATURAS E SIGLAS

t'_R	Tempo de Retenção Ajustado
t_R	Tempo de Retenção
AGC	<i>Automatic Gain Control</i>
AIF	<i>All-Ions Fragmentation</i>
AM1	<i>Austin Model 1</i>
ANN	<i>Artificial Neural Network</i> , ou Redes Neurais Artificiais
AsterDB	<i>Asteraceae Database</i> , ou Banco de dados de Asteraceae
BRNN	<i>Bayesian Regularization for feed-forward Neural Networks</i> , ou redes neurais <i>feed-forward</i> com regularização bayesiana
CASD	<i>Computer-Assisted Synthesis Design</i> , ou Planejamento de Síntese com Auxílio de Computador
CASE	<i>Computer-Assisted Structural Elucidation</i> , ou Elucidação Estrutural com Auxílio de Computador
CASRN	<i>Chemical Abstracts Service Registry Number</i>
CL	Cromatografia Líquida
CLAE ou HPLC	<i>High Performance Liquid Chromatography</i> , ou Cromatografia Líquida de Alta Eficiência
CLUE ou UHPLC	<i>Ultra-High Performance Liquid Chromatography</i> , ou Cromatografia Líquida de Ultra Eficiência
COX	Ciclooxigenase
DAD ou PDA	<i>Photodiode Array/Diode array detectors</i> , ou Detector de Arranjo de Diódos
EC	<i>Enzyme Comission</i>
ESI	<i>Electrospray Ionization</i>
GA	<i>Genetic Algorithm</i> , ou Algoritmo Genético
GC ou CG	<i>Gas Chromatography</i> , ou Cromatografia em fase Gasosa
GUI	<i>Graphical User Interface</i> , ou Interface Gráfica com o Usuário

H-ESI II	<i>Heated Electrospray Ionization</i>
HCA	<i>Hierarchical Cluster Analysis</i> , ou Análise de Agrupamento Hierárquico
IC	<i>Ion-exchange Chromatography</i> , ou Cromatografia de troca iônica
InChI	IUPAC <i>International Chemical Identifier</i>
IR ou IV	<i>Infrared</i> , ou Infravermelho
IUPAC	<i>International Union of Pure and Applied Chemistry</i>
KRFPC	Klekota-Roth <i>Fingerprinter Counts</i>
LDA	<i>Linear Discriminant Analysis</i> , ou Análise Linear Discriminante
LFS	<i>subset Linear Forward Selection</i> , ou Seleção de subconjuntos Linear para a Frente
LMD	<i>Low Mode Dynamics</i>
LOF	<i>Lack Of Fit</i> , ou Falta De Ajuste
LST	Lactonas Sesquiterpênicas, sesquiterpenlactona ou sesquiterpenos lactonizados
m/z	<i>mass-to-charge ratio</i> , ou massa/carga
MAE	<i>Mean Absolute Error</i> , ou Erro Médio Absoluto
MeCN	Acetonitrila
MeOH	Metanol
MLR	<i>Multiple Linear Regression</i> , ou Regressão Linear Múltipla
mol file	<i>Molecular File</i>
MS ou EM	<i>Mass Spectrometry</i> , ou Espectrometria de Massa
ODS	<i>Octadecylsilyl</i> , ou octadecilsilil
OLS	<i>Ordinary Least Squares</i> , ou Mínimos Quadrados Ordinários
PDB	<i>Protein Data Bank</i> , ou Banco de Dados de Proteínas
PLS	<i>Partial Least Squares</i> , ou Mínimos Quadrados Parciais
PTFE	Politetrafluoretileno

QSAR	<i>Quantitative Structure–Activity Relationship</i> , ou Relação Quantitativa da Estrutura-Atividade
QSPR	<i>Quantitative Structure–Property Relationship</i> , ou Relação Quantitativa da Estrutura-Propriedade
QSRR	<i>Quantitative Structure–Retention Relationship</i> , ou Relação Quantitativa da Estrutura-Retenção
RMN ou NMR	<i>Nuclear Magnetic Resonance</i> , ou Ressonância Magnética Nuclear
RMSE	<i>Root Mean Squared Error</i> , ou Raiz Quadrada do Erro Quadrático Médio
RTT	Tempo de Retenção multiplicado por 100
sdf	<i>Structure-Data File</i>
SFC	<i>Supercritical Fluid Chromatography</i> , ou Cromatografia de Fluido Supercrítico
SMILES	<i>Simplified Molecular Input Line Entry Specification</i>
SQL	<i>Structured Query Language</i> , ou Linguagem de Consulta Estruturada
TCMD	<i>Database of Traditional Chinese Medicine</i> , ou Banco de dados da Medicina Tradicional Chinesa
URL	<i>Uniform Resource Locator</i> , ou Localizador Padrão de Recursos
UV	<i>Ultraviolet</i> , ou ultravioleta
Vpp	<i>peak-to-peak Voltage</i> , ou Tensão pico a pico

SUMÁRIO

1	INTRODUÇÃO	20
1.1	Banco de dados	21
1.2	Quimioinformática	22
1.2.1	Representação estrutural	22
1.2.2	Descritores estruturais	24
1.2.3	Quimioinformática e produtos naturais	24
1.2.4	Estimar o tempo de retenção	26
2	OBJETIVOS	30
3	MATERIAIS E MÉTODOS	31
3.1	Coleta de dados	31
3.2	AsterDB	31
3.3	Estudo da retenção cromatográfica de lactonas sesquiterpênicas (LST)	31
3.3.1	Logaritmo do fator de retenção	32
3.3.2	Programas e especificação dos computadores	34
3.3.3	Modelagem molecular e otimização da geometria	36
3.3.4	Geração de descritores	36
3.3.5	Seleção dos descritores	37
3.3.6	Construção dos modelos	38
3.3.7	Validação dos modelos	39
3.3.8	Modelo consensual	40
3.4	Uso de ferramentas de quimioinformática para avaliar a transposição metodológica do tempo de retenção de produtos naturais	40
3.4.1	Substâncias	40
3.4.2	Equipamentos	40
3.4.3	Método	43
3.4.4	QSRR	43
3.4.5	Análises estatísticas	44
3.5	Estudos de QSRR de um banco de dados comercial	44
3.5.1	Padronização das estruturas, modelagem molecular e otimização da geometria	45
3.5.2	Classificação e domínio de aplicação	45
3.5.3	Obtenção das variáveis dependentes	46
3.5.4	Obtenção das variáveis independentes	46

3.5.5	Construção dos modelos	47
3.5.5.1	Treino, validação externa e teste externo	47
3.5.5.2	Seleção dos descritores	47
3.5.5.3	Modelagem	47
3.5.5.4	Avaliação dos modelos	49
4	RESULTADOS E DISCUSSÃO	50
4.1	AsterDB	50
4.1.1	Descrição e conteúdo das tabelas	52
4.1.1.1	Tabela <i>classification</i>	52
4.1.1.2	Tabela <i>structure</i>	59
4.1.1.3	Tabela <i>descriptors</i>	60
4.1.1.4	Tabela <i>analytical</i>	61
4.1.1.5	Tabela <i>activity</i>	61
4.1.1.6	Tabela <i>genera</i>	62
4.1.1.7	Tabela <i>tanonomy</i>	62
4.1.1.8	Tabela <i>botany</i>	63
4.1.1.9	Tabela <i>plantparttable</i>	64
4.1.1.10	Tabela <i>boco</i>	64
4.1.1.11	Tabela <i>structufiles</i>	64
4.2	Estudo da retenção cromatográfica de lactonas sesquiterpênicas (LST)	65
4.3	Uso de ferramentas de quimioinformática para avaliar a transposição metodológica do tempo de retenção de produtos naturais	77
4.4	Estudos de QSRR de um banco de dados comercial	94
5	CONCLUSÕES	107
	REFERÊNCIAS	109
	APÊNDICES	117
	APÊNDICE A – CRIAÇÃO DAS TABELAS DO ASTERDB	118
	APÊNDICE B – MELHORES MODELOS DE QSRR OBTIDOS PARA AS ESTRUTURAS QUÍMICAS CEDIDAS PELA EMPRESA ANALYTICON DISCOVERY	121

ANEXOS	127
ANEXO A – COPYRIGHT	128

1 INTRODUÇÃO

Os estudos das substâncias de baixo peso molecular oriundas de amostras complexas de origem biológica utilizando métodos hifenados cromatográficos-espectrométricos, com objetivo de se obter uma ampla visão do metabolismo de plantas, animais e ou microrganismos, tem gerado controvérsias quanto à sua nomenclatura (2), assim, neste trabalho, usará o termo “estudos metabolômicos”. Estes estudos envolvem várias abordagens, tais como *metabolomics*, *metabonomics*, *metabolic profiling*, *metabolic fingerprinting*, *metabolic footprinting* e *metabolite targeted analysis*. Tais termos podem ser traduzidos respectivamente para português, como se segue:

- metabolômica: identificação e quantificação de todos os metabólitos em um dado sistema biológico (3);
- metabonômica: medição dinâmica e quantitativa da resposta metabólica dos sistemas vivos com o objetivo de avaliar alterações fisiológicas e modificações genéticas (3);
- perfil metabólico: estudo de um grupo pré-definido de metabólitos (aminoácidos, carboidratos, terpenoides, flavonoides, etc.), ou com base na via metabólica (3, 4);
- impressão digital metabólica: análise rápida visando a classificação da amostra por reconhecimento de padrões (3);
- “pegada” metabólica: difere da impressão digital metabólica, pois o objeto de estudo são os metabólitos extracelulares, sejam eles excretados ou não absorvidos (4);
- análise de metabólito alvo: estudo clássico, onde é/são analisado(s) e isolado(s) apenas um ou alguns metabólitos (3).

Excetuando-se as análises de metabólito alvo, todas as demais (multi-alvos) necessitam de ferramentas para a detecção com alta sensibilidade e especificidade como importante pré-requisito para a identificação dos metabólitos (5). Além disso, se faz necessário desenvolver procedimentos de triagem com o objetivo de identificar rapidamente e com precisão as substâncias já conhecidas ou para estabelecer o grau de novidade. No entanto, o sucesso de um estudo metabolômico, em especial quando se requer uma análise qualitativa denominada “dereplication”, cuja tradução ainda não foi normalizada para a língua portuguesa (derreplicação (6) ou desreplicação (7, 8)), requer uma elevada qualidade dos dados obtidos, em especial a acuidade e rapidez do processo de identificação de estruturas. Uma maneira de se otimizar tempo e reduzir custos é empregar o uso de bibliotecas de produtos naturais e bancos de dados com informações que auxiliem

neste processo de identificação. Estas informações na forma de substâncias ou de suas respectivas informações virtuais, que tanto podem ser acadêmicas como comerciais, são constituídas de várias substâncias puras isoladas e identificadas associadas a seus respectivos dados espectrométricos e cromatográficos. Logo, os dados destas bibliotecas e bancos de dados podem ser utilizados como auxílio na etapa de desreplicação de matrizes orgânicas complexas (9, 10, 11).

1.1 BANCO DE DADOS

Com o surgimento da era computacional, com especial aplicação em química, as substâncias de origem natural puderam ter suas informações armazenadas em bancos de dados (ou base de dados). De acordo com *National Telecommunications and Information Administration - United States Department of Commerce* (12), define-se como banco de dados um conjunto de dados relacionados a um ou a vários assuntos/objetos de forma que podem ser organizados de tal modo que possam ser consultados por usuários locais ou remotamente.

Programa para bancos de dados são programas com o objetivo de armazenar e organizar dados, assim como existem programas para a edição de texto (Write, Word, Pages, etc.), para a criação de planilhas e gráficos simples (Calc, Excel, Numbers, etc.). Estes integram os sistemas de informação, onde os dados são coletados, processados, organizados e armazenados. A etapa de coleta, tratamento e processamento dos dados normalmente é executada em diferentes interfaces/programas como, por exemplo, *websites*. Porém, o armazenamento e organização desses dados são feitos em um banco de dados. Dentre os principais programas para gestão dos dados há o Oracle Database, SQL Sever, MySQL, PostgreSQL, MongoDB entre outros.

Muitos dados químicos armazenados nos computadores dos laboratórios de pesquisa do Brasil e do mundo possuem pouca ou nenhuma organização estrutural. Poucos estão estruturados na forma de tabelas em planilhas, mas quase nunca utilizando programa específico construído sob a forma de um banco de dados relacional, o qual fornece uma maneira de armazenar grandes quantidades de dados em tabelas que são independentes ou relacionadas uma com a outra (13).

O desenvolvimento de bancos de dados com foco em química pode diferir quanto ao conteúdo, temática (14, 15), quanto ao objeto abordado como, por exemplo, dados etnofarmacológicos e de medicina tradicional (16, 17), informações regionais (18, 19), informações sobre doenças específicas (20), conformações estruturais (21), classes de substâncias específicas (22, 23), espectros de massas (24) e tipo de acesso (pago, *in house* ou livre/aberto/gratuito). Podem disponibilizar informações de substâncias puras e específicas, informações sobre elucidação estrutural e/ou atividade biológica ou dados mais

específicos, inclusive com foco na química de produtos naturais. Há também bancos de dados mistos como, por exemplo, o ChemSpider e o SciFinder (Tabela 1, página 23).

1.2 QUIMIOINFORMÁTICA

Nesta interface entre as bibliotecas de substâncias oriundas de síntese orgânica e/ou produtos naturais e o desenvolvimento de bancos de dados surge uma das aplicabilidades da quimioinformática (do inglês, *chemoinformatics* ou *cheminformatics*), que pode ser definida como a área na fronteira entre a química e a informática que aplica e desenvolve sistemas de informação, algoritmos, técnicas computacionais e métodos estatísticos para resolver problemas de química (25, 26). Seu conceito engloba outra importante área de estudo, como a química computacional a qual é definida como uma disciplina que usa métodos matemáticos para o cálculo das propriedades moleculares e/ou para a simulação do comportamento molecular (27).

Desde os anos 90, várias ferramentas de quimioinformática foram desenvolvidas, tais como cálculos de descritores, algoritmos para investigar a similaridade estrutural, algoritmos de classificação, sistemas de buscas de estruturas químicas e bancos de dados (28). Neste sentido várias áreas do conhecimento utilizam dessas ferramentas, como por exemplo: QSPR (QSAR, QSRR), Quimiometria, Modelagem Molecular, *Virtual Screening*, CASE, CASD, etc. (29).

1.2.1 Representação estrutural

Exemplos referentes a representação molecular podem ser observados na Tabela 2 (página 25), onde verificam-se várias representações da substância tagitinina A (uma LST natural). A representação inicia com o nome comum seguida por meio de uma identidade numérica (CASRN), uma típica entrada de bancos de dados de estruturas químicas (SMILES), a estrutura no formato de *sdf/mol file*, a representação baseada na teoria dos grafos (2D) e geométrica (3D).

SMILES é uma notação em linha (método topográfico usando caracteres imprimíveis) para incorporar e representar estruturas químicas e reações. Sua codificação é relativamente simples, pois os átomos das estruturas são representados pelos seus respectivos símbolos atômicos. Nesta notação, em linhas gerais, os átomos de hidrogênio são omitidos, os átomos vizinhos são representados ao lado de cada átomo, as ligações simples são omitidas, as duplas são representadas por “=” e as triplas por “#”, as ramificações são representadas por parênteses, anéis são representados pela colocação de dígitos nos dois átomos que o fecham e anéis aromáticos são representados por letras minúsculas (30, 31). Com o objetivo de ampliar as informações contidas no SMILES, a IUPAC desenvolveu um identificador para bancos de dados chamado InChI™.

Tabela 1 – Exemplos de bancos de dados de interesse na área de produtos naturais.

Nome	Informação	Acesso	URL
SciFinder	Estruturas químicas, dados correlacionados e artigos	Pago	< www.scifinder.cas.org >
ChemSpider	Estruturas químicas e dados correlacionados	Livre	< www.chemspider.com >
FiehnLib	CG/EM e t_R	Pago	< http://fiehnlab.ucdavis.edu/projects/FiehnLib >
NMRShiftDB	RMN	Livre	< www.nmrshiftdb.nmr.uni-koeln.de >
MeRy-B	RMN, <i>in house</i>	Restrito	< http://services.cbib.u-bordeaux2.fr/MERYB/ >
Protein Data Bank (PDB)	Proteínas e ligantes	Livre	< www.rcsb.org >
SpecInfo	RMN e dados cristalográficos	Parcialmente livre	< www.cds.dl.ac.uk >
KnowItAll	Dados espectrais (RMN, UV, EM, IV)	Pago	< www.knowitallanyware.com >
BindingDB	Atividades biológicas	Livre	< www.bindingdb.org >
Dr. Duke's Phytochemical and Ethnobotanical Databases	Produtos naturais	Livre	< http://www.ars-grin.gov/duke/ >
Phenol-Explorer	Polifenóis de produtos naturais	Livre	< http://phenol-explorer.eu >
Super Natural Database II	Substâncias naturais	Livre	< www.bioinf-applied.charite.de >
TCMD	Produtos naturais	Livre	< www.megabionet.org/tcmid >
TCMD @ Taiwan	Produtos naturais	Livre	< www.tcm.cmu.edu.tw >
NuBe Database	Estruturas de substâncias naturais do Brasil	Livre	< nubbe.iq.unesp.br/portal/nubbedb.html >
Sistemat X	Estruturas de substâncias naturais, classificação e localização	Livre	< www.sistemax.ufpb.br >

Embora o InChITM não seja o método convencional para representar substâncias, ele fornece um meio eficaz de representar a sua identidade. Em resumo, o InChITM é composto por uma série de caracteres obtidos aplicando-se um conjunto de regras para uma estrutura química a fim de proporcionar uma “assinatura” digital única de uma substância (32, 33).

Uma vez que os motores de buscas mais utilizados na Internet não seriam capazes de encontrar as cadeias de caracteres do InChITM devido à sua extensão e uso de caracteres ignorados atualmente por todos os motores de busca, a IUPAC desenvolveu o InChIKey, o qual possui comprimento fixo (27 caracteres) com base no *hash code* (código de dispersão) da InChITM. Ainda assim, o InChIKey herda em um grau limitado as informações do InChITM (34).

1.2.2 Descritores estruturais

Um descritor estrutural é o resultado final de um procedimento lógico e matemático que traduz a informação química codificada dentro de uma representação simbólica da molécula (SMILES, .mol/.sdf, .pdb, etc.) em um valor numérico (descritor teórico), ou o resultado de algum experimento padronizado (descritor experimental). Esta informação pode detalhar e auxiliar a interpretação de alguma propriedade molecular, assim como, ser capaz de tornar-se parte de um modelo para estimar propriedades de moléculas desconhecidas. Em outras palavras, pode-se dizer que um descritor é uma forma matemática de se representar uma estrutura química (35, 36, 37).

O desenvolvimento e interpretação completa dos descritores envolvem conhecimento de álgebra, teoria dos grafos, teoria da informação, química computacional, quimioinformática, reações orgânicas, química quântica, físico-química, etc. Para a sua utilização, requer o conhecimento de estatística e quimiometria. Além disso, a quimioinformática e o uso de programas específicos são muitas vezes inseparáveis companheiros de viagem do pesquisador neste campo (36, 37).

Os descritores podem ser classificados quanto, à dimensão estrutural que podem ser obtidos ou à função. A Tabela 3 (página 27) apresenta de uma forma resumida estas classificações. Os descritores podem ser calculados por diversos programas, como exemplificado na Tabela 4 (página 28).

1.2.3 Quimioinformática e produtos naturais

Em um dos primeiros trabalhos que envolveram Produtos Naturais e Quimioinformática, foi descrita a criação de um banco de dados de estruturas tridimensionais (3D) de substâncias oriundas de plantas da medicina tradicional chinesa e o cômputo de descritores. Essas estruturas foram utilizadas para a busca de inibidores da protease NS3-NS4A do HCV (vírus da hepatite C) (21).

Tabela 2 – Alguns exemplos de representação estrutural.

Representação	Nome
tagitinin A	Nome Comum
6,9-epoxiciclodeca[β]furano, propanoico ácido deriv.	Sinonímia
C ₁₉ H ₂₈ O ₇	Fórmula Molecular
(1 <i>R</i> ,2 <i>S</i> ,4 <i>R</i> ,8 <i>S</i> ,9 <i>R</i> ,11 <i>R</i> ,12 <i>S</i>)-1,12-diidroxí-2,11-dimetil-7-metileno-6-oxo-5,14-dioxatriciclo[9.2.1.0 ^{4,8}]tetradec-9-il,2-metilpropanoato	Nome IUPAC
59979-61-2	CASRN
C[C@H]1C[C@@H]2[C@@H]([C@@H](C[C@@]3([C@H](C[C@]1(O3)O)O)C)OC(=O)C(C)C)C(=C)C(=O)O2	SMILES
1S/C19H28O7/c1-9(2)16(21)25-13-7-18(5)14(20)8-19(23,26-18)10(3)6-12-15(13)11(4)17(22)24-12/h9-10,12-15,20,23H,4,6-8H2,1-3,5H3/t10-,12+,13+,14-,15-,18+,19+/m0/s1	InChI
HREHFPZHVCMOMQ-XNNFIJVS-A-N	InChIKey
<pre> #O0541 011315103320 26 28 0 0 0 0 999 V2000 2.7718 -10.8980 0.0000 C 0 0 1 0 0 0 0 0 0 0 0 0 0 0 2.7718 -11.7224 0.0000 C 0 0 2 0 0 0 0 0 0 0 0 0 0 0 3.3550 -12.3063 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 4.1800 -12.3063 0.0000 C 0 0 2 0 0 0 0 0 0 0 0 0 0 0 4.7632 -11.7224 0.0000 C 0 0 1 0 0 0 0 0 0 0 0 0 0 0 4.7632 -10.8980 0.0000 C 0 0 1 0 0 0 0 0 0 0 0 0 0 0 4.1800 -10.3142 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3.3550 -10.3142 0.0000 C 0 0 2 0 0 0 0 0 0 0 0 0 0 0 4.5553 -13.0407 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5.3699 -12.9124 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5.4977 -12.0977 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2.7718 -9.7310 0.0000 C 0 0 2 0 0 0 0 0 0 0 0 0 0 0 2.1885 -10.3142 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3.5693 -11.1117 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5.9532 -13.4956 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 6.2121 -11.6846 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3.7681 -9.6004 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2.0573 -12.1355 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1.9760 -11.1117 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2.7718 -8.9066 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5.5252 -10.5823 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 6.2397 -10.1699 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 6.2396 -9.3449 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 6.9541 -10.5824 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5.5252 -8.9324 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 6.9541 -8.9325 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 1 0 0 0 0 3 4 1 0 0 0 0 4 5 1 0 0 0 0 5 6 1 0 0 0 0 6 7 1 0 0 0 0 7 8 1 0 0 0 0 9 10 1 0 0 0 0 10 11 1 0 0 0 0 8 12 1 0 0 0 0 12 13 1 0 0 0 0 1 13 1 0 0 0 0 5 11 1 0 0 0 0 4 9 1 6 0 0 0 10 15 2 0 0 0 0 11 16 2 0 0 0 0 8 17 1 6 0 0 0 2 18 1 6 0 0 0 1 19 1 6 0 0 0 12 20 1 6 0 0 0 8 14 1 1 0 0 0 1 14 1 1 0 0 0 2 3 1 0 0 0 0 6 21 1 1 0 0 0 22 23 1 0 0 0 0 22 24 2 0 0 0 0 23 25 1 0 0 0 0 23 26 1 0 0 0 0 21 22 1 0 0 0 0 M END \$\$\$\$ </pre>	sdf/mol file
estrutura 2D (com estereoquímica atribuída)
estrutura 3D

Outro trabalho descreve cerca de 3.000 estruturas tridimensionais (3D) partindo-se de substâncias com atividade anti-inflamatória oriundas de espécies de plantas que haviam sido descritas por Dioscórides na sua obra *De Materia Medica*. Neste trabalho foi realizada uma abordagem etnofarmacológica para a triagem de substâncias químicas de plantas. Porém, todo este processo foi realizado *in silico*, utilizando-se modelos computacionais e empregando-se as enzimas COX I e II virtuais (38).

Estes dois trabalhos desencadearam uma série de outros que unem ambas áreas de estudo na busca de substâncias bioativas (39, 40, 41, 42, 43). Desta forma, pode-se concluir que a elaboração de um banco de dados com estruturas de produtos naturais não garante o aumento de produtividade, mas é necessária para o desenvolvimento de novos conhecimentos. Além da busca por substâncias bioativas, existem outras finalidades, como o auxílio na elucidação estrutural ou a possibilidade de estimar outras propriedades estruturais.

Outro importante foco de convergência é o emprego de bancos de dados de produtos naturais e de algumas ferramentas de quimioinformática com objetivo de identificar substâncias oriundas de estudos metabólicos. Algumas vezes, não é possível identificar rapidamente uma dada substância, seja por ela ainda não ter sido descrita ou pelo fato da biblioteca ser limitada. Logo, são necessárias técnicas adicionais que sejam desenvolvidas com objetivo de otimizar os processos de desreplicação e de elucidação estrutural, em especial, automatizando parte deste processo, no qual mais uma vez se destacam as ferramentas de quimioinformática. O estudo da relação quantitativa entre estrutura química e retenção cromatográfica, ou QSRR, é um exemplo destas técnicas (44, 45, 46, 47).

1.2.4 Estimar o tempo de retenção

QSRR é a correlação estatística entre os parâmetros cromatográficos, hifenizados ou não, e propriedades teóricas das moléculas (descritores), permitindo obter regressões lineares, não lineares, análises de componentes (48) e inteligência artificial (49, 50) para explicar o comportamento de solutos (analitos) em determinados sistemas cromatográficos. A existência destas correlações torna possível estimar fatores ou t_R de substâncias partindo-se de suas estruturas químicas, ou identificar estruturas a partir de índices de retenção (51, 52, 53).

Uma simples busca em sites como Scielo, ScienceDirect, Web of Science e PubMed, utilizando-se a combinação: “*natural products + QSRR*” para os últimos 10 anos até dezembro de 2014, excluindo-se revisões e capítulos de livros foram encontrados 31 resultados e destes, 20 estimaram o tempo de retenção de substâncias voláteis analisadas por métodos de CG, 01 para SFC (54), 02 para IC (55, 56) e 08 para CL (57, 58, 59, 60, 61, 62, 63, 64).

Porém, apenas a partir de 2013 autores como Goryński et al. (60), Oliveira, Da Costa

Tabela 3 – Classificação geral dos descritores estruturais.

Representação Molecular	Função	Exemplo
0D*	Descritores constitucionais	Peso molecular, contagem dos átomos, ligações, número de carbonos, número de oxigênio, ligações rotacionáveis, etc.
1D	Cômputo dos fragmentos e grupos funcionais	Contagem de hidroxilas, ácidos carboxílicos, carbonos primários em um anel, etc.
2D	<i>walk e path counts</i> , índices de conectividade, índice de informações, <i>2D matrix-based</i> , autocorrelações 2D, <i>Burden eigenvalues</i> , descritores P_VSA-like, <i>extended topochemical atom indices</i> , <i>edge adjacency indices</i> , <i>atom-type e-state indices</i> , CATS 2D e 2D atom pairs	<i>Information Content index</i> , ETA alpha, Zagreb index, etc.
3D	Descritores geométricos, <i>3D matrix-based</i> , autocorrelações 3D, RDF, 3D-MoRSE, WHIM, GETAWAY, <i>Randic molecular profiles</i> , <i>3D atom pairs</i> , <i>Grid properties</i> e G-WHIM	Excentricidade molecular, média do potencial eletrostático molecular, etc.
4D*	3D	Média de diferentes confôrmeros
5D*	4D	Obtenção dos confôrmeros através de algum método de ajuste explícito (proteína)
6D*	5D	Obtenção dos confôrmeros através de diferentes solvatações
Outros	Descritores de carga	Total da carga positiva e negativa, polaridade, etc.
Outros	Propriedades moleculares	Índice de saturação, fator de hidrofiliabilidade, coeficiente de partição octanol-água, etc.

* Alguns autores consideram a ausência de dimensão (0D) ou diferentes poses 3D como novas dimensões, não considerando as dimensões físicas. O autor desta tese manteve essa classificação por questões didáticas.

Tabela 4 – Programas para calcular descritores estruturais.

Nome	Desenvolvedor	País	Licença
ADMET Predictor TM	Simulations Plus, Inc.	EUA	Pago
CORINA Symphony	Molecular Networks GmbH	Alemanha	Pago
Pentacle, GRID e VolSurf+	Molecular Discovery	Reino Unido	Pago
CODESSA PRO	CompuDrug International, Inc.	EUA	Pago
Dragon	Talete s.r.l.	Itália	Pago
ISIDA Fragmentor	Laboratoire d'Infochimie, Université de Strasbourg	França	Gratuito
JOelib	Center for Bioinformatics Tübingen, University of Tübingen	Alemanha	Gratuito
JChem e Marvin	ChemAxon Ltd.	Hungria	Pago
MOE	Chemical Computing Group Inc.	Canadá	Pago
Molconn-Z	eduSoft, LC	EUA	Pago
Molgen QSPR	Molgen	Alemanha	Pago
PaDEL-Descriptor	National University of Singapore	Cingapura	Gratuito
PowerMV	National Institute of Statistical Sciences	EUA	Gratuito
PreADMET	Bioinformatics & Molecular Design Research Center	Coréia do Sul	Pago
Sarchitect TM	Strand Life Sciences	Índia	Pago
E-Dragon	Virtual Computational Chemistry Laboratory	Alemanha	Gratuito
Open3DQSAR	Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino	Itália	Gratuito
ADMEWORKS	FQS Poland	Polônia	Pago
ToMoCoMD-CARDD Suite	Environmental and Computational Chemistry Group, Universidad de Cartagena	Espanha	Gratuito
CDK Descriptor	Rajarshi Guha	EUA	Gratuito
ParaSurf TM	CAChe Research	EUA	Pago

e Schmidt (62), Eugster et al. (58) e Oliveira et al. (1) indicaram em seus artigos e resumos a aplicação de estudos de QSRR como uma ferramenta que pode ser aplicada em análises de desreplicação ou de metaboloma. Desta forma, ao se unir os avanços das etapas de purificação, desreplicação, elucidação estrutural e racionalização do processo de seleção do estudo de novas espécies vegetais com as ferramentas de quimioinformática e métodos estatísticos ou mineração de dados, o cronograma para a descoberta de substâncias naturais a partir de matrizes vegetais pode ser reduzido.

Neste contexto, surge a ideia de se organizar os dados de substâncias disponíveis na literatura, isoladas pelo grupo AsterBioChem (<<http://www.asterbiochem.org/>>), análises experimentais e ainda dados de substâncias que nos foram cedidos pela empresa *Analyti-Con Discovery GmbH*, os quais podem servir para a elaboração de modelos de QSRR para se poder prever diferentes propriedades de substâncias com foco na retenção cromatográfica, bem como obter explicações sobre os fenômenos físico-químicos envolvidos nessas propriedades. No âmbito da metabolômica, tais dados podem ainda ser incorporados ao processo de automatização para o estudo de misturas de substâncias e desreplicação de matrizes complexas. Além disso, dados experimentais adicionais a serem obtidos, também podem ser incorporados a essas ferramentas, melhorando assim sua capacidade no âmbito da elucidação estrutural auxiliada por computador.

Logo, o emprego de ferramentas de quimioinformática em produtos naturais é importante para estudos metabolômicos, de desreplicação e de elucidação estrutural, pois os resultados podem ser aplicados como ponto de partida para criar ferramentas definidas pelo usuário e mais refinadas, o que também pode incorporar o gerenciamento de estruturas ou de recuperação de dados. Adicionalmente, o desenvolvimento de um estudo que como este pode auxiliar na restrição de substâncias no processo de identificação e desreplicação em estudos metabolômicos.

2 OBJETIVOS

Esta tese envolve o emprego de diferentes ferramentas de quimioinformática para auxiliar a desreplicação de substâncias naturais em estudos metabolômicos. Seus objetivos específicos são:

1. Construir um banco com dados de t_R de estruturas provenientes de espécies de Asteraceae, assim como de propriedades moleculares (descritores calculados por técnicas de quimioinformática), de informações analíticas e de informações taxonômicas;
2. Investigar propriedades de retenção e prever t_R de substâncias naturais em cromatografia líquida de alta eficiência através de estudos de QSRR;
3. Desenvolver uma plataforma piloto integrada para análise de dados com a finalidade de auxiliar na desreplicação de extratos de plantas em estudos de metaboloma de Asteraceae.

3 MATERIAIS E MÉTODOS

3.1 COLETA DE DADOS

Informações taxonômicas e ocorrências (tribo, subtribo, gênero e espécie), sobre atividade biológica, bem como demais dados da literatura (local de coleta, parte do vegetal estudado, etc.), foram e ainda estão sendo coletados de forma contínua para cada estrutura química de substâncias oriundas da família Asteraceae disponíveis na literatura e/ou isoladas e/ou em investigação pelo grupo AsterBioChem. Tais informações estão sendo utilizadas para povoar o AsterDB, o qual é um banco de dados único, que contém estruturas de terpenoides, flavonoides e derivados do ácido trans-cinâmico, entre outras substâncias, incluindo mais de 1.000 estruturas de LST entre as 3.000 estruturas que compõe o banco até o presente momento.

3.2 ASTERDB

Para simular o ambiente do AsterDB foi utilizado servidor virtual, criado com o VirtualBox 4.3.30 (Oracle, EUA), o qual está hospedado em duas máquinas físicas do Laboratório de Quimioinformática e Farmacognosia da FCFRP-USP;

O servidor virtual possui configuração básica conforme as características a seguir:

- Sistema operacional Debian wheezy GNU/Linux
- Servidor de banco de dados MySQL 5.5.43
- Sevidor Java Tomcat 8
- Sevidor Web Apache 2
- Openjdk 8
- ChemDoodle 7
- Marvin Freeweb

3.3 ESTUDO DA RETENÇÃO CROMATOGRÁFICA DE LAC- TONAS SESQUITERPÊNICAS (LST)

Informações complementares desta seção poderão ser obtidas no link <<http://pubs.acs.org/doi/abs/10.1021/ci500581q>>, e, conforme o Anexo A (página 128), o autor desta tese solicitou o direito de uso das informações, tabelas e figuras contidas no respectivo artigo.

3.3.1 Logaritmo do fator de retenção

O logaritmo do fator de retenção ($\log k$) foi calculado para 39 LST presentes em nossa biblioteca de substâncias puras. A equação (3.1) mostra a forma logarítmica do fator de retenção (k). De acordo com a IUPAC, é expressa pelo tempo que um componente da amostra é retardado pela fase estacionária quando comparado com o tempo da fase móvel ao percorrer a coluna (27).

$$\log k = \log \frac{t_R - t_M}{t_M} \quad (3.1)$$

Na equação 3.1, t_R e t_M são o tempo de retenção total e *hold-up time* (t_R de um componente não retido pela fase estacionária), respectivamente, k é a fator de retenção e $(t_R - t_M)$ é o tempo de retenção ajustado.

As 39 LST usadas neste estudo foram isoladas previamente de espécies da família Asteraceae por membros do AsterBioChem e detalhes sobre respectivas elucidações estruturais foram publicados previamente (65, 66, 67, 68, 69, 70, 71). Este grupo de substâncias compreendem estruturas com seis diferentes tipos de esqueletos de LST (cinco germacrolídeos, três melampolídeos, oito guaianolídeos, 10 furano heliangolídeos, 12 heliangolídeos e um eudesmanolídeo).

O preparo das amostras foi realizado em colaboração com o Prof. Dr. Leonardo Gobbo-Neto (FCFRP-USP), conforme Oliveira et al. ((1)). Para tal, 2 mg de cada LST foram dissolvidas em 200 μL de uma solução MeCN–H₂O 35:65, filtrados em uma membrana PTFE de 0,45 μm e o volume injetado foi 1,0 μL . As LST foram injetadas e analisadas por CLAE sob duas condições isocráticas: MeOH–H₂O (55:45), 1,0 mL min⁻¹ (sistema 1) e MeCN–H₂O (35:65), 1,3 mL min⁻¹ (sistema 2), usando uma ODS C-18 coluna Shimadzu (4,6 x 250 mm, 5 μm). As substâncias foram detectadas com um detector de arranjo de diodos (DAD) a 225 e 265 nm. 2,5-dimetilfenol (DMP) foi usado com padrão interno. Todas as análises foram executadas em um sistema de cromatografia líquida da Shimadzu SCL-10AVP equipado com um detector DAD da Shimadzu SPD-M10AVP DAD, sendo operado com o programa CLASS-VP 5.02.

Cada tempo de retenção (t_R) gerou um $\log k$ em cada sistema de solvente ($\log k_1$ e $\log k_2$ - para os sistemas 1 e 2, respectivamente, conforme dados da Tabela 5, página 32).

Tabela 5 – Lista das LST avaliadas e seus logaritmos do fator de retenção nos sistemas MeOH–H₂O (55:45), $\log k_1$ e MeCN–H₂O (35:65), $\log k_2$.

Substância	Nome	$\log k_1$	$\log k_2$
1	eriofertina	0,502	0,502

Continua na próxima página

Tabela 5 – Continuação da página anterior

Substância	Nome	$\log k_1$	$\log k_2$
2	ovatifolina	0,746	0,883
3	tagitinina A	0,536	0,436
4	viguilenina	0,780	0,660
5	budleina B	0,472	0,373
6	(5 <i>R</i> ,6 <i>R</i> ,7 <i>R</i> ,8 <i>R</i>)-3-oxo-4 β ,10 α -diidroxi-8 β -metilpropanoiloxi-guaia-11(13)-en-6 α ,12-olídeo	0,668	0,494
7	zaluzanina C	0,975	0,722
8	15-desoxibudleina A	1,134	1,270
9	tiglato de atripliciolídeo	1,105	1,233
10	budleina A	0,682	0,564
11	tiglato de budleina A	0,626	0,516
12	erioflorina	0,623	0,636
13	heliangina	0,460	0,778
14	leptocarpina	0,557	0,839
15	tagitinina E	0,642	0,639
16	tagitinina F	0,885	1,216
17	1 β ,10 α -epoxi-3 β -hidroxi-8 β -(2' <i>R</i> ,3' <i>R</i> -epoxiangeloiloxi)-germacra-4,11(13)-dien-6 α ,12-olídeo	0,446	0,406
18	1 β ,10 α -epoxi-3 β -acetoxi-8 β -(2' <i>R</i> ,3' <i>R</i> -epoxiangeloiloxi)-germacra-4,11(13)-dien-6 α ,12-olídeo	0,669	0,842
19	1 β ,10 α -epoxi-3 β -(2-metilbutanoiloxi)-8 β -(2' <i>R</i> ,3' <i>R</i> -epoxiangeloiloxi)-germacra-4,11(13)-dien-6 α ,12-olídeo	1,088	1,423
20	1 β ,10 α -epoxi-3 β -isovaleroiloxi-8 β -(2' <i>R</i> ,3' <i>R</i> -epoxiangeloiloxi)-germacra-4,11(13)-dien-6 α ,12-olídeo	1,347	1,574
21	1 β ,10 α -epoxi-3 β -(2-metilpropanoiloxi)-8 β -(2' <i>R</i> ,3' <i>R</i> -epoxiangeloiloxi)-germacra-4,11(13)-dien-6 α ,12-olídeo	0,832	1,271
22	1 β ,3 α -dimetoxi-3 β ,10 β -epoxi-8 β -metilpropanoiloxi-germacra-4,11(13)-dien-6 α ,12-olídeo	1,189	1,258
23	acetato de heliangina	1,074	1,267

Continua na próxima página

Tabela 5 – Continuação da página anterior

Substância	Nome	$\log k_1$	$\log k_2$
24	2 β -metoxi-3 α -hidroxi-3,10 β -epoxi-8 β -metilpropanoiloxi-germacra-11(13)-en-6,12-olídeo	1,309	1,233
25	1 β ,2 α -epoxitagitinina C	0,718	0,730
26	angelato de acantospermolídeo	1,214	1,451
27	2-oxo-8 β -metacriloiloxi-guaia-3,10(14),11(13)-trien-6 α ,12-olídeo	1,054	0,998
28	2-oxo-8 β -metacriloiloxi-guaia-1(10),3,11(13)-trien-6 α ,12-olídeo	1,381	1,160
29	2-oxo-8 β -metacriloiloxi-10 α -hidroxi-guaia-3,11(13)-dien-6 α ,12-olídeo	0,770	0,816
30	2-oxo-8 β -metacriloiloxi-10 β -hidroxi-guaia-3,11(13)-dien-6 α ,12-olídeo	0,609	0,427
31	2-oxo-8 β -tigloiloxi-guaia-1(10),3,11(13)-trien-6 α ,12-olídeo	1,357	1,376
32	2 α ,3 β -dihidroxi-8 β -metacriloiloxi-germacra-1(10),4,11(13)-trien-6 α ,12-olídeo	0,755	0,435
33	partenolídeo	1,076	1,154
34	niveusina A	0,691	0,376
35	tagitinina C	0,855	0,787
36	1 α -metacriloiloxi-2 β ,8 α -dihidroxi-3 α ,4 α -epoxi-eudesm-11(13)-en-6 α ,12-olídeo	0,558	0,328
37	enhidrina	0,831	1,177
38	uvedalina	1,188	1,437
39	2-oxo-8 β -epoxiangeloiloxi-guaia-1(10),3,11(13)-trien-6 α ,12-olídeo	0,785	0,903

3.3.2 Programas e especificação dos computadores

Os cálculos estatísticos foram executados com auxílio de um 2.7 GHz Intel®Pentium Dual Core E5400 com 3 GB de RAM com sistema operacional OpenSuse 13.2 (*openSUSE Project*, Alemanha). MarvinSketch 15.7.13.0 (ChemAxon, Hungria) foi utilizado para desenhar as estruturas moleculares (72). Weka 3.6.12 (Universidade de Waikato, Nova Zelândia)

Figura 1 – Estruturas químicas das 39 lactonas sesquiterpênicas (LST).

Fonte: Reimpresso e adaptado com permissão de Oliveira et al.(1); copyright 2015 American Chemical Society.

(73) e R 3.2.1 (*R Foundation*, Áustria) (74) foram empregados tanto para a seleção de descritores quanto para gerar e validar as modelagens utilizando ANN. A seleção de descritores com GA, análises conformacionais, cálculo dos descritores 3D e modelos de regressão por PLS foram efetuados com o MOE 2010.11 (*Chemical Computing Group*, Canadá) (75) em um Intel® Core 2 Duo com 2 GB de RAM instalado sobre o sistema operacional Microsoft Windows XP®.

3.3.3 Modelagem molecular e otimização da geometria

As estruturas 2D das 39 LST foram inicialmente pré-minimizadas usando o gerador de estruturas 3D CORINA (*Molecular Networks*, Alemanha) (76). As geometrias resultantes foram refinadas por meio da busca conformacional usando o LMD (*low mode dynamics*), campo de força MMFF94x e configuração padrão do MOE.

3.3.4 Geração de descritores

As estruturas 2D das moléculas foram utilizadas para cálculo dos seguintes descritores estruturais: descritores globais e coeficiente de autocorrelação 2D utilizando propriedade de polarizabilidade dos pares de átomos, calculados com o Adriana.Code 2.2.6 (*Molecular Networks*, Alemanha) (77); propriedades moleculares, índices de informação e descritores constitucionais calculados com o Dragon 5.5 (Talet, Itália) (78); cômputo dos grupos funcionais, estados eletrotópicos do tipo de átomo, índice de conectividade, complexidade de fragmentos, cômputo das ligações rotacionáveis, propriedades moleculares, complexidade das estruturas e descritores baseados na matriz 2D foram calculados usando PaDEL 2.21 (Universidade Nacional de Singapura, Singapura) (79). Este conjunto de descritores (655) doravante serão chamados de “2D-descr”.

Descritores obtidos com base em coordenadas das matrizes cartesianas de estruturas 3D minimizadas, tais como a área de superfície, descritores de volume e forma dependentes da carga foram calculados utilizando MOE. O conjunto de 123 descritores 3D calculados a partir de um único confômero de menor energia encontrado durante a busca conformacional será chamado de “3D-1conf” e os mesmos 123 descritores 3D calculados a partir de um conjunto conformacional ponderados pela distribuição de Boltzmann foram denominados “3D-weigh” (80). Este último grupo de descritores foi calculado utilizando um máximo de 10 confômeros diferentes para cada molécula, para o qual todos os descritores foram calculados e ponderados de acordo com o método AM1 usando a equação de Boltzmann (81):

$$p_i = \frac{e^{\frac{-\Delta E_i}{RT}}}{\sum_{i=1}^t e^{\frac{-\Delta E_i}{RT}}} \quad (3.2)$$

onde a probabilidade p de encontrar uma estrutura em um estado particular i em um equilíbrio de t estados (conformações) a uma dada temperatura absoluta T está relacionada para a diferença de energia do estado i para o mínimo global de energia ΔE e R é a constante universal dos gases.

3.3.5 Seleção dos descritores

Como qualquer estudo de QSPR, a seleção de descritores estruturais é o passo mais importante que afeta diretamente a qualidade dos modelos (82). Os três grupos de descritores previamente descritos (2D-descr, 3D-1conf e 3D-weigh) foram tratados com o pacote Caret v. 5.15-87 (83) usando as funções *nearZeroVar* e *findCorrelation*. Na primeira função, todos os descritores com variância igual ou próxima a zero foram excluídos. Na segunda função, todos os descritores foram avaliados quanto à alta correlação entre eles. Neste caso corre-se o risco de criar modelos com descritores redundantes, portanto quando ocorrem um par de descritores com alta correlação entre si exclui-se um deles. Neste trabalho, usando-se essas duas funções e o limite máximo de correlação de 0,90, foram eliminados 565 descritores do grupo 2D-descr, 62 do grupo 3D-1conf e 60 do grupo 3D-weigh.

Em sequência, quatro diferentes algoritmos foram utilizados para selecionar as combinações de descritores mais relevantes para os modelos de QSRR: (i) *Best First*, (ii) *Greedy Stepwise*, (iii) *Linear Forward* e (iv) GA. Os primeiros três métodos (calculados no Weka) avaliam o valor de um subconjunto de descritores considerando a capacidade preditiva individual de cada descritor, juntamente com o grau de redundância entre eles e com validação cruzada *10-folds* (84). O quarto método, GA, foi calculado com auxílio do programa MOE. Esses quatro algoritmos foram utilizados para reduzir o número de descritores para um tamanho apropriado (quatro, cinco e seis descritores) utilizando diferentes critérios de seleção. Tal número de descritores é condizente com o número de estruturas químicas para criar um modelo parcimonioso.

A seleção *best first* (i) usa um algoritmo que inicia com um único descritor aleatório, do qual sua correlação é comparada com o $\log k$; então, ele encontra uma melhor solução com buscas para trás (*backward*) e para frente (*forward*) e, sequencialmente, adiciona um descritor de cada vez. A seleção *greedy stepwise* (ii) inicia a partir de um ponto arbitrário na matriz de descritores e executa uma busca para frente adicionando grupos de descritores. Ele para quando a adição de todos os descritores restantes gera apenas reduções na correlação com o $\log k$. O *linear forward selection* (iii) é uma extensão do método *best first*, o mesmo leva um número restrito (*fixed-set*) de n descritores. Todos estes três algoritmos foram avaliados ao considerar a capacidade de previsão individual de cada um juntamente com o grau de redundância entre eles por validação cruzada *10-folds*.

O método de seleção GA otimiza diretamente uma família ou população de combi-

nações de descritores por meio de seu desempenho em MLR. Neste procedimento, cada descritor dentro de uma combinação particular representa um “gene” dentro do modelo. Os genes individuais podem ser trocados aleatoriamente (“mutação”) para novos descritores criando assim novos modelos. Troca de grupos de genes (ou seja, mais de um descritor) entre os modelos já existentes representa eventos de “*crossing over*”. Evoluindo uma população de tais modelos ao longo de um grande número de ciclos de “mutações” e “*crossing overs*” sob consequente re-avaliação após cada ciclo e eliminação de modelos com fraco desempenho, por uma questão de manter os modelos com melhor desempenho, conduz automaticamente a uma otimização da população no que se refere ao parâmetro escolhido para representar o desempenho. O critério de desempenho aqui escolhido foi a diminuição da LOF no MLR dos descritores em relação ao valor-alvo, $\log k$. O tamanho da população escolhida para uma otimização de execução foi de 100 modelos. O critério de término utilizado foi ou uma LOF predefinido que foi atingido ou, um número máximo de 1000 ciclos tinham sido realizados. Após o término, toda a população que foi submetida a uma validação cruzada *leave-one-out* e as combinações de descritores com o maior coeficiente de determinação da validação cruzada foram escolhidos para posterior avaliação.

O objetivo desta etapa foi selecionar três diferentes grupos de quatro, cinco e seis descritores, usando-se quatro diferentes métodos de seleção (i, ii, iii e iv).

3.3.6 Construção dos modelos

As estruturas das LTS foram ordenadas sequencialmente do menor para o maior $\log k$ e dois grupos foram selecionados para teste e treino. Os grupos de treino e teste foram selecionados de tal maneira que a distribuição global das amostras foi preservada sempre que possível. No primeiro grupo, 26 estruturas foram selecionadas para construção do modelo e 13 como conjunto de teste, no segundo grupo 29 estruturas foram usadas como treinamento e 10 como conjunto de teste. Os conjuntos de teste foram utilizados para validar os modelos (ver seção 3.3.7, página 39).

Em seguida, os modelos QSRR foram construídos usando os dois conjuntos de treino diferentes. Os descritores selecionados e os $\log k$ foram correlacionados por PLS e um modelo não-linear, neste caso, ANN.

Em PLS, um método linear, em primeiro lugar, os vetores foram gerados pela sequência de Krylov $\{b, Sb, S^2b, \dots, S^{A-1}b\}$; em seguida, uma matriz de ponderação $V_A = (v_1, v_2, \dots, V_A)$ foi construída por ortogonalização de Gram-Schmidt, onde os v_i são os vetores com n colunas, e A , o grau de ajuste do PLS, o qual é um número inteiro, inferior ou igual a n . Então, para obter os A^{th} coeficientes PLS, foram resolvidos por (85, 75):

$$a = V_A(V_A^T S V_A)^{-1} V_A^T b \quad (3.3)$$

Um algoritmo para determinar sucessivos ajustes PLS será encerrado no A^{th} passo quando $\|v_{A+1}\| = 0$. Pode-se, no entanto, querer utilizar um vetor de regressão de PLS de baixa ordem para ajustar um modelo linear. Se o modelo continua até $A = n$, a matrix de peso é quadrática e completamente ranqueada. O resultado da regressão é o mínimo quadrado ordinário, também conhecido como OLS.

Em ANN, uma rede do tipo *multilayer perceptron* com o algoritmo de determinação dos pesos chamado *back-propagation* foi utilizada no Weka. Em cada nó (ou neurônio) uma função sigmoide foi utilizada e as equações para este modelo foram definidas (86):

$$f(x) = \frac{1}{1 + e^{-x}} \quad (3.4)$$

Para aplicar o procedimento gradiente descendente, a função de erro é:

$$E = \frac{1}{2}(y - f(x))^2 \quad (3.5)$$

onde $f(x)$ é a rede obtida do neurônio de saída e y é o $\log k$ (normalizado entre 0 e 1). O peso para cada neurônio é calculado por um *perceptron* com uma camada oculta (*hidden layer*) e por derivação usa-se a equação a seguir:

$$\frac{\partial E}{\partial w_{ij}} = (f(x) - y)f'(x)w_i f'(x_i)a_j \quad (3.6)$$

onde $f'(x)$ simboliza a derivada (∂) de $f(x)$, w_i e w_{ij} são os pesos e a_j um vetor. Um novo peso é calculado para cada instância de treinamento, as alterações estão associadas a valores regulados durante a afinação (*tuning*) da rede neural como momento (α , *momentum*), taxa de aprendizagem (η), tempo de treinamento (k), e número de camada oculta. O peso da instância anterior é multiplicado pela taxa de aprendizagem, e o resultado é depois subtraído do valor próximo de w_{ij} . Devido a este mecanismo de propagação de erros, esta versão da estratégia de gradiente descendente é chamado de *backpropagation* (86).

3.3.7 Validação dos modelos

No presente estudo, validação cruzada *10-folds* foi realizada para avaliar a robustez e validade dos modelos, bem como a sua previsibilidade interna. Conjuntos de teste com estruturas (13 e 10) que não foram utilizadas no desenvolvimento do modelo QSRR (ver seção 3.3.6, página 38) foram então usados para a validação externa. Posteriormente, a qualidade do ajuste dos modelos foi avaliada usando os seguintes parâmetros: coeficiente de correlação ao quadrado para o modelo ajustado (grupo treino), R^2 , o quadrado do coeficiente de correlação da validação cruzada (grupo treino), Q^2 , o quadrado do coeficiente

de correlação para o conjunto de teste, P^2 , conforme equação 3.7; o MAE (equação 3.8) e RMSE (equação 3.9):

$$R^2 = Q^2 = P^2 = \left(\frac{\frac{\sum_{i=1} (p_i - \bar{p})(a_i - \bar{a})}{n-1}}{\sqrt{\left(\frac{\sum_{i=1} (p_i - \bar{p})^2}{n-1} \right) \left(\frac{\sum_{i=1} (a_i - \bar{a})^2}{n-1} \right)}} \right)^2 \quad (3.7)$$

$$MAE = \frac{|p_1 - a_1| + \dots + |p_n - a_n|}{n} \quad (3.8)$$

$$RMSE = \sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n}} \quad (3.9)$$

onde $p_1, p_2, \dots, p_n$ são os valores estimados e $a_1, a_2, \dots, a_n$ os valores experimentais, e n o número de estruturas em análise.

3.3.8 Modelo consensual

A partir dos melhores modelos de cada sistema cromatográfico, os modelos consensuais 1 e 2 foram construídos para o sistema MeOH–H₂O e os modelos 3 e 4 para o sistema MeCN–H₂O (cada par contendo 29 e 26 estruturas como conjunto de treino, respectivamente). O modelo consensual 5 foi construído para o sistema MeCN–H₂O utilizando 26 estruturas como treino, levando-se em conta os melhores desempenhos estatísticos dos descritores 2D assim como a separação dos isômeros que os descritores 3D fornecem.

3.4 USO DE FERRAMENTAS DE QUIMIOINFORMÁTICA PARA AVALIAR A TRANSPOSIÇÃO METODOLÓGICA DO TEMPO DE RETENÇÃO DE PRODUTOS NATURAIS

3.4.1 Substâncias

50 substâncias puras oriundas das bibliotecas de substâncias do grupo de pesquisa AsterBioChem e do *Institute of Pharmaceutical Biology and Phytochemistry, WWU-Münster* (IPBP) foram utilizadas neste estudo (Figura 2, página 41). As substâncias foram analisadas em três diferentes equipamentos e dois métodos. As substâncias isoladas foram injetadas separadamente e em mistura (grupos de 10 e um grupo de 50).

3.4.2 Equipamentos

UHPLC-MS1: as separações cromatográficas foram realizadas num sistema de cromatografia líquida Acella (*Thermo Scientific*, EUA). As substâncias eluídas foram detectadas

Figura 2 – Cinquenta substâncias puras do AsterBioChem e IPBP analisadas neste estudo.

utilizando um UV Accela PDA (*Thermo Scientific*) ao longo de um intervalo de comprimento de onda de 200-400 nm e um espectrômetro de massas Exactive Plus Orbitrap (*Thermo Scientific*) equipado com uma sonda de *electrospray* (H-ESI II) e operado em modos de ionização negativa e positiva usando a varredura completa sem fragmentação e aquisição em um intervalo de massas de 120-1200 Da, usando as seguintes configurações do equipamento: nebulizador de nitrogênio de 8 bar; 0,5 mL/min, 300 °C, tensão de pulverização de 3600 V no modo de ionização negativa e 3200 V no modo de ionização positiva, gás de bainha a 30 unidades arbitrárias; temperatura do capilar em 300 °C e temperatura do aquecedor a 300 °C; *S-Lens* com 55 unidades arbitrárias. Controle automático de ganho (AGC) teve valor-alvo fixado em cargas de 3×10^6 e o tempo máximo de injeção foi ajustado para 250 ms. A resolução foi ajustada a 70.000 para cada módulo (m/z 200), com exatidão da massa < 5 ppm e tempo de varredura de 2,5 scans/s. Os dados foram adquiridos no modo de perfil. O sistema foi controlado pelos programas Xcalibur e Tune (*Thermo Scientific*).

UHPLC-MS2: as separações cromatográficas foram realizadas num sistema de cromatografia líquida Ultimate 3000 RS (Dionex, EUA). As substâncias eluídas foram detectadas utilizando um Ultimate DAD-3000 RS (Dionex, EUA) ao longo de um intervalo de comprimento de onda de 200-400 nm e um espectrômetro de massas por tempo de voo micrOTOF-QII (Bruker Daltonics, Alemanha) equipado com uma fonte de ionização por *eletrospray* Apollo no modo positivo a 5 Hz ao longo de um intervalo de massas de 50 - 1000 m/z , usando as seguintes configurações do equipamento: nebulização de gás nitrogênio a 5 bar; gás de secagem, 9 L/min, 220 °C; tensão capilar de 4500 V; *end plate offset* a -500 V; tempo de transferência de 70 μ s; o gás de colisão foi o nitrogênio; as configurações de energia de colisão e colisão de RF foram combinadas para cada espectro único de 1000 *summations* como se segue: 250 *summations* com 20% de energia de colisão de base e 130 Vpp + 250 *summations* com 100% de energia de colisão de base e 500 Vpp + 250 *summations* com 20% de energia de colisão de base e 130 Vpp + 250 *summations* com 100% de energia de colisão de base e 500 Vpp. A energia de colisão de base foi 50 eV para íons precursores com um m/z menor que 500 e, em seguida, interpolados linearmente frente ao m/z , até um máximo de 70 eV para íons precursores com um m/z de até 1000. A calibração interna (modo HPC) foi realizada para cada análise utilizando o espectro de massa de uma solução 10 mM de formiato de sódio em 50% de isopropanol, que foi infundida durante o reequilíbrio do cromatógrafo líquido usando uma válvula de desvio do circuito contendo 20 μ L de calibrante.

UPLC-DAD: as separações cromatográficas foram realizadas num sistema de cromatografia líquida Acquity (Waters, EUA). As substâncias eluídas foram detectadas utilizando um detector PDA e λ (Waters, EUA).

3.4.3 Método

AsterBioChem 1 (ABC1): a fase estacionária foi uma coluna ACE3 C18-300 (3,0 x 150 mm, 3,0 μm - ACE, Reino Unido) a 30 °C com uma pré-coluna C18 e um gradiente binário (A: água com 0,1% de ácido fórmico; B: acetonitrila com 0,1% de ácido fórmico) a 0,3 mL/min: 0 a 5 min: eluição isocrática com 5 % B; 5,0 a 50,0 min: aumento linear de 5 de % B a 100% de B; 50 a 55 min: eluição isocrática com 100% B; 55 a 65 min: redução linear de 100% a 5% de B. O volume de injeção foi de 2 μL a 0,1mg/mL.

AsterBioChem 2 (ABC2): a fase estacionária foi uma coluna Kinetex XB-C18 100 Å (2,1 x 150 mm, 1,7 μm - Phenomenex, EUA) a 30 °C com uma pré-coluna C18 e um gradiente binário (A: água com 0,1% de ácido fórmico; B: acetonitrila com 0,1% de ácido fórmico) a 0,3 mL/min: 0 a 5,0 min: eluição isocrática com 5% de B; 5 a 50 min: aumento linear de 5% de B a 100% de B; 50 a 55 min: eluição isocrática com 100% de B; 55 a 65 min: redução linear de 100% a 5% de B. O volume de injeção foi de 2 μL a 0,1mg/mL.

IPBP: a fase estacionária foi uma coluna Acclaim RSLC 120 C18 (2,1 x 100 mm, 2.2 μm - Dionex, EUA) a 40 °C com um gradiente binário (A: água com 0,1% de ácido fórmico; B: acetonitrila com 0,1% de ácido fórmico) a 0,8 mL/min: de 0 a 9,5 min: aumento linear de 5 % B a 100% B; 12,5 to 12,6 min: redução linear de 100% B a 5% B; 12,6 a 15 min: gradiente isocrático de 5% B. O volume de injeção foi de 2 μL a 0,1mg/mL.

3.4.4 QSRR

As estruturas químicas 2D foram desenhadas usando o MarvinSketch e as coordenadas 3D foram geradas utilizando o MOE. Cada estrutura foi minimizada pelo método semi empírico AM1. O ambiente que se encontra uma molécula pode interferir com a sua conformação e essas conformações disponíveis para uma molécula podem afetar a atividade, a reatividade e outras propriedades. Dessa forma, para resolver este problema foi realizada uma busca de conformação. Nesta etapa, 20 confôrmeros foram selecionados com 4 Kcal de intervalo de energia utilizando o método sistemático.

As matrizes das coordenadas 2D das estruturas foram utilizadas para calcular os descritores 1D e 2D com auxílio dos programas MOE e PaDEL. Esse conjunto de descritores (173 e 2315 descritores obtidos pelo MOE e PaDEL, respectivamente) serão chamados doravante por “2D-MOE” e “2D-Padel”. Os descritores obtidos com base nas coordenadas das matrizes cartesianas dos diferentes confôrmeros 3D, tais como área de superfície, volume, forma e conformação bem como, descritores dependentes da carga eletrônica foram calculados utilizando o MOE. O conjunto de descritores 3D (n = 149) foram calculados a partir de um conjunto conformacional utilizando a equação 3.2 e será chamado de “3D-weigh”.

Os três conjuntos de descritores previamente descritos (2D-MOE, 2D-Padel e 3D-

weigh) foram pré-tratados com o pacote Caret 6.0-34 (83) utilizando as funções *nearZeroVar* e *findCorrelation* (suas funções são explicadas na seção 3.3.5, página 37). Nesta etapa, aplicando estes dois algoritmos e limiar de 0,75 com a função *findCorrelation*, 147 descritores foram eliminados do grupo 2D-MOE, 2186 do grupo 2D-Padel e 101 do grupo 3D-weigh.

Em seguida foram fundidos todos os grupos de descritores e as estruturas foram dissociadas numa formação para modelagem ($n = 34$) e um conjunto de teste ($n = 16$), independentemente das suas estruturas químicas, mas de tal modo que ambos os conjuntos abrangessem uniformemente toda a gama de t'_R .

Dois algoritmos de seleção de descritores foram utilizados para selecionar as combinações de descritores mais relevantes para os modelos de QSRR: (i) LFS e (ii) GA. Ambos os métodos de seleção de subconjuntos foram avaliados pelos seus respectivos desempenhos frente a uma MLR. Estes algoritmos foram também utilizados para encontrar subconjuntos de cinco e seis descritores.

Modelos computacionais (ex.: MLR, PLS, ANN e árvores de decisão) para estimar o t'_R foram construídos por métodos de aprendizagem de máquina com auxílio dos programas Weka, R e MOE.

Análises quimiométricas e ferramentas de quimioinformática foram utilizadas para avaliar as diferenças entre as análises.

3.4.5 Análises estatísticas

O teste de Shapiro foi utilizado para se conhecer a normalidade dos dados. E o teste Kruskal-Wallis (quando os dados não estão distribuídos em uma curva normal) foi utilizado para as comparações do t'_R .

3.5 ESTUDOS DE QSRR DE UM BANCO DE DADOS COMERCIAL

A empresa *Analyticon Discovery GmbH* <<http://www.ac-discovery.com>> (Potsdam, Alemanha) disponibilizou 2949 estruturas de diferentes origens (vegetais e microrganismos), diferentes classes químicas e seus respectivos t_R em CL (mediana, média, desvio padrão e número de análises) para que se fossem construídos modelos de QSRR. A empresa atua na área de pesquisa e desenvolvimento de produtos fabricados com produtos naturais, nas áreas farmacêuticas, de alimentos e cosméticos. Ela possui diferentes coleções de produtos naturais (MEGx) e substâncias semi-sintéticas (NATx) derivadas de produtos naturais. Atualmente sua biblioteca possui mais de 25000 substâncias.

3.5.1 Padronização das estruturas, modelagem molecular e otimização da geometria

Uma das dificuldades de se realizar um estudo com banco de dados desenvolvido por terceiros é obter estruturas confiáveis e seus dados associados (ex.: número de eluições, t_R , média do t_R , etc.). Nesse estudo, utilizamos o princípio de se trabalhar com *Big Data* em quimioinformática, ou seja, o uso de ferramentas robustas para auxiliar na redução de perda de informação.

Inicialmente utilizou-se o programa *Structure Checker* 6.2 (Chemaxon, Hungria) com o objetivo de verificar a correta configuração estereoquímica, erros nas ligações químicas, presença de solventes, cátions, ânions, erros em ângulos e distâncias de ligação das 2.949 estruturas disponibilizadas no banco de dados MEGx. Nesse procedimento 47 estruturas foram removidas do estudo.

Em uma segunda etapa, nas 2902 estruturas restantes adicionou-se hidrogênio, aromatizou e reconstruiu-se todas as estruturas 2D (mantendo as respectivas estereoquímicas) com o programa *Standardizer* 6.2 (Chemaxon, Hungria).

As coordenadas 3D foram geradas utilizando o MOE. Cada estrutura foi minimizada pelo método semi empírico AM1. Foi efetuada uma busca de 11 confôrmers com 7 kcal de intervalo de energia utilizando o método LMD.

Após essa etapa houve a reunião das estruturas com suas informações de tempo de retenção, obtendo no final 2635 estruturas com seus respectivos tempos de retenção.

3.5.2 Classificação e domínio de aplicação

Eugster et al.(58) demonstraram que com apenas 35 descritores moleculares foi possível separar estruturas de forma similar às suas classes fitoquímicas. Conforme esse pressuposto e devido à necessidade de se realizar uma classificação automatizada para grandes dados, usou-se 4.860 subestruturas geradas por fragmentação de cada estrutura do tipo KRFPC. O programa Padel gerou uma tabela que descreveu a ausência (zero) e quando presente, quantas vezes que cada fragmento estava na estrutura (1, 2,... , n) (87). Desta forma, obteve-se um *fingerprint* completo de cada estrutura para auxiliar na classificação. A matriz completa obtida foi utilizada nos passos subsequentes.

Com o objetivo de se obter o domínio de aplicação e quantidade de grupos químicos das estruturas calculou-se como medida de dissimilaridade a distância euclidiana das estruturas dada pela equação 3.10.

$$d(x, y) = \left(\sum_{j=1}^d (x_j + y_j)^2 \right)^{1/2} \quad (3.10)$$

onde d é a distância n -dimensional entre os vetores x e y .

Em um segundo passo utilizou-se o algoritmo de agrupamento *k-means* para reunir as estruturas semelhantes. O algoritmo começa com uma estimativa inicial para os centros dos grupos e cada estrutura é alocada no grupo ao qual ela está mais próxima. Os centros dos grupos são então atualizados e todo o processo é repetido até que os centros dos grupos já não se movam (88).

A definição ideal para o número de grupos se deu pelo algoritmo *C-index*, calculado pela seguinte equação:

$$Cindex = \frac{S_w - S_{min}}{S_{max} - S_{min}}, S_{min} \neq S_{max}, Cindex \in (0, 1) \quad (3.11)$$

onde S_{min} é a soma das menores distâncias entre todos os pares de pontos N_w em todo o conjunto de dados (existem N_t pares) e S_{max} é a soma das maiores distâncias entre todos os pares de pontos N_w em todo o conjunto de dados. O menor valor para o *Cindex* indica o valor ideal para o número de grupos (88).

Todo o processo de classificação e definição da quantidade ótima de grupos foi calculado com auxílio do pacote Nbclust 3.0 para o R.

3.5.3 Obtenção das variáveis dependentes

Os tempos de retenção utilizados neste estudo foram obtidos utilizando-se como fase estacionária uma coluna C-8 e como fase móvel um gradiente MeOH/MeCN/H₂O e tampão formiato de amônio (HCOO⁻NH₄⁺) com pH $\approx$ 3,0. Os valores obtidos foram multiplicado por um fator ($RRT = 100t_R$).

Quando as substâncias possuíam vários t_R , optou-se por utilizar a mediana por ser uma medida estatística robusta quando comparada com a média.

3.5.4 Obtenção das variáveis independentes

As estruturas 2D das estruturas padronizadas (seção 3.5.1, página 45) foram utilizadas para cálculo dos descritores estruturais usando o Dragon (78), PaDEL (79) e o MOE (75). Cada matriz gerada por cada programa foi tratada individualmente com o pacote Caret para o R (83) usando as funções *nearZeroVar* e *findCorrelation*. Ao final, as matrizes foram unificadas e novamente utilizou a função *findCorrelation* com o objetivo de excluir descritores com correlação maior que 0,90. Este conjunto de descritores ($n = 454$) doravante será chamado de “2D-descr”.

As coordenadas das matrizes cartesianas das estruturas 3D a partir de um único conformero de menor energia (seção 3.5.1, página 45) foram utilizadas para cálculo dos descritores estruturais usando o Dragon (78), PaDEL (79) e MOE (75). Seguiu-se os mes-

mos pre-tratamentos dos descritores 2D. Este conjunto de descritores ($n = 303$) doravante será chamado de “3D-1conf”. Os 271 descritores 3D calculados a partir de um conjunto conformacional tiveram suas médias ponderadas tendo como peso a distribuição de Boltzmann e foram denominados como “3D-weight”. Este último grupo de descritores foi calculado utilizando a equação 3.2, página 36.

3.5.5 Construção dos modelos

3.5.5.1 Treino, validação externa e teste externo

Em um primeiro momento, separou-se as estruturas químicas de forma estratificada e randômica, com o objetivo de se ter uma amostra representativa quanto aos RRT para um **teste externo** (25%, $n = 656$ estruturas) comum a todos os modelos criados.

Com as demais estruturas ($n = 1979$) fez-se uma nova separação para **treino/modelagem** e **validação externa**. Neste caso separou-se três conjuntos para treino e validação externa seguindo dois modelos teóricos.

- i Partição simples baseada na resposta: foi utilizada a função *createDataPartition* do pacote Caret para criar divisões aleatórias estratificadas do conjunto de dados. Neste caso, 75% ($n = 1487$) dos dados foram usados para a formação do modelo e o restante (25%, $n = 492$) utilizados para avaliar o desempenho do modelo (validação externa). A função cria as separações aleatórias dentro do intervalo de RRT de modo que a distribuição total do RRT foi preservada tão bem quanto possível. Esta separação foi codificada pela sigla “prt”.
- ii Partição baseada nos preditores: duas outras separações foram obtidas através de mapas auto organizáveis do tipo Kohonen. A primeira utilizou-se a matriz 2D-descr (“som2D”, treino = 1486) e a segunda 3D-1conf (“som3D”, treino = 1486). Após qualquer um desses agrupamentos e o uso da função *createDataPartition*, obteve-se separações aleatórias dentro de cada agrupamento, de modo que a distribuição total de cada grupo foi preservada tão bem quanto possível.

3.5.5.2 Seleção dos descritores

Para cada grupo de treino foram feitas seleções de quatro, cinco e seis descritores através de GA (seção 3.3.5, página 37).

3.5.5.3 Modelagem

Após a seleção dos descritores, os dados foram modelados utilizando MLR e BRNN.

A seleção destes descritores se baseou nos grupos de descritores (quatro, cinco ou seis) com menor RMSE e/ou maior coeficiente de determinação na validação cruzada do

tipo *leave one out* (Q^2). Em seguida ajustou-se o modelo para MLR utilizando *least squares regression*:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (3.12)$$

onde k são os descritores moleculares selecionados e $\hat{y}$ o valor estimado de RTT. Os valores $\beta_0, \beta_1, \dots, \beta_k$ foram estimados a partir da minimização da soma dos quadrados dos resíduos (SSE):

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.13)$$

onde n é o número de estruturas utilizadas no treino, y o valor experimental de RRT e, $\hat{y}$ o valor estimado de RRT.

O algoritmo BRNN ajusta uma rede neural de duas camadas tal como descrito por MacKay(89) e Foresee e Hagan(90):

$$\hat{y} = \sum_{k=1}^s (w_k g_k(b_k + \sum_{j=1}^p x_{ij} \beta_j^{[k]})) + e_i, i = 1, \dots, n \quad (3.14)$$

onde n é o número de estruturas utilizadas no treino, y o valor experimental de RRT, $\hat{y}$ o valor estimado de RRT, s o número de neurônios, w_k o peso do k -enésimo neurônio ($k = 1, \dots, s$), b_k um viés para o k -enésimo neurônio ($k = 1, \dots, s$) e $\beta_j^{[k]}$ o peso para a j -enésima entrada da rede ($j = 1, \dots, s$). Ele utiliza o algoritmo de Nguyen e Widrow(91) para atribuir pesos iniciais (g_k):

$$g_k(x) = \frac{\exp(2x) - 1}{\exp(2x) + 1} \quad (3.15)$$

e o algoritmo de Gauss-Newton para executar a otimização. Desta forma, o algoritmo se ajusta evitando o super ajuste (*overfitting*) (92):

$$F = \beta E_D + \alpha E_W \quad (3.16)$$

onde $E_D = SSE$, ou seja, a soma dos quadrados dos resíduos; E_W é a soma dos quadrados dos parâmetros da rede (pesos e viés); $\beta = \frac{1}{2\sigma_e^2}$; e $\alpha = \frac{1}{2\sigma_\theta^2}$, onde σ_θ^2 é um parâmetro de dispersão para pesos e viés.

O valor de s de cada modelo foi selecionado levando em consideração a quantidade máxima de descritores de cada modelo e o menor RMSE médio de uma validação cruzada *10-folds* repetidas 10 vezes.

Os melhores modelos finais foram selecionados pelos valores da validação externa e a comparação de predição de novas estruturas não presentes na construção dos modelos (teste externo). Em seguida, decidiu-se construir modelos consensuais.

3.5.5.4 Avaliação dos modelos

A capacidade de estimar novos RTT e a significância estatística dos modelos desenvolvidos neste estudo foram avaliadas utilizando parâmetros estatísticos, tais como R^2 (R^2 - treino, Q^2 - validação externa e P^2 - teste externo), MAE e RMSE.

Primeiramente foi computado o quadrado do coeficiente de correlação de acordo com a equação 3.7, página 40. Com o objetivo de mensurar o erro entre o valor estimado e o RTT, calculou-se o MAE usando a equação 3.8, página 40. Adicionalmente ao MAE, também foi calculado o RMSE, o qual é frequentemente usado como uma medida estatística das diferenças entre os valores previstos pelo modelo e os valores efetivamente observados. RMSE é uma medida da capacidade de estimativa do modelo (equação 3.9, página 40).

4 RESULTADOS E DISCUSSÃO

4.1 ASTERDB

O banco de dados AsterDB foi desenvolvido para ser preenchido de forma gradual pelos usuários. Atualmente contém arquivos no formato `.sdf` e `.mol` e possui cerca de 2000 estruturas químicas. O diagrama estrutural do AsterDB pode ser observado na Figura 3, página 51.

Figura 3 – Diagrama estrutural do banco de dados AsterDB.

4.1.1 Descrição e conteúdo das tabelas

Nas próximas sessões serão detalhados os conteúdos das várias tabelas do banco de dados AsterDB (Figura 3, página 51). Como pretende-se futuramente disponibilizar o banco na internet, sua estrutura foi redigida no idioma inglês.

4.1.1.1 Tabela *classification*

Descrição

Nesta tabela encontram-se os dados relacionados ao sistema de classificação de produtos naturais, o nível 0 (zero) indicará a classificação mais próxima da estrutura e ela deverá ser um valor único. Essa é uma tabela fixa, alterada apenas pelos administradores do AsterDB.

Conteúdo

As informações contidas nessa tabela estão apresentadas na Tabela 6, página 53.

Tabela 6 – Tabela de povoamento padrão da tabela *classification*.

Level 0 (lv0)	Level 1 (lv1)	Level 2 (lv2)	Level 3 (lv3)	Level 4 (lv4)
1.1 alcohols	1 alcohols and phenols	1.1 alcohols		
1.2 phenols	1 alcohols and phenols	1.2 phenols		
2.1 aldehydes	2 aldehydes and keto- nes	2.1 aldehydes		
2.2 ketones	2 aldehydes and keto- nes	2.2 ketones		
3.1 chalcones and dihydro- chalcones	3 flavonoids	3.1 chalcones and dihy- drochalcones		
3.2 isoflavans	3 flavonoids	3.2 isoflavans		
3.3 flavanones	3 flavonoids	3.3 flavanones		
3.4 flavonols	3 flavonoids	3.4 flavonols		
3.5 flavones	3 flavonoids	3.5 flavones		
3.6 isoflavones	3 flavonoids	3.6 isoflavones		
3.7 anthocyanidins	3 flavonoids	3.7 anthocyanidins		
4. glucosinolates	4 glucosinolates			
5.1.1 monoterpenes	5 isoprenoids	5.1 terpenoid-type	5.1.1 monoterpenes	
5.1.2 iridoids	5 isoprenoids	5.1 terpenoid-type	5.1.2 iridoids	

Continua na próxima página

Tabela 6 – Continuação da página anterior

Level 0 (lv0)	Level 1 (lv1)	Level 2 (lv2)	Level 3 (lv3)	Level 4 (lv4)
5.1.3.1 sesquiterpenes lactones	5 isoprenoids	5.1 terpenoid-type	5.1.3 sesquiterpenes	5.1.3.1 sesquiterpenes lactones
5.1.3.2 non-lactone sesquiterpenes	5 isoprenoids	5.1 terpenoid-type	5.1.3 sesquiterpenes	5.1.3.2 non-lactone sesquiterpenes
5.1.4 diterpenes	5 isoprenoids	5.1 terpenoid-type	5.1.4 diterpenes	
5.1.5 limonoids	5 isoprenoids	5.1 terpenoid-type	5.1.5 limonoids	
5.1.6 triterpenes	5 isoprenoids	5.1 terpenoid-type	5.1.6 triterpenes	
5.1.6.1 triterpenes saponins	5 isoprenoids	5.1 terpenoid-type	5.1.6 triterpenes	5.1.6.1 triterpenes saponins
5.1.7 tetraterpenes	5 isoprenoids	5.1 terpenoid-type	5.1.7 tetraterpenes	
5.2.1 sterols	5 isoprenoids	5.2 steroid-type	5.2.1 sterols	
5.2.2 ketosterones	5 isoprenoids	5.2 steroid-type	5.2.2 ketosterones	
5.2.3 cardenolides and bufadienolides	5 isoprenoids	5.2 steroid-type	5.2.3 cardenolides and bufadienolides	
5.2.4 steroid sapogenins	5 isoprenoids	5.2 steroid-type	5.2.4 steroid sapogenins	
5.2.5 steroid saponins	5 isoprenoids	5.2 steroid-type	5.2.5 steroid saponins	

Continua na próxima página

Tabela 6 – Continuação da página anterior

Level 0 (lv0)	Level 1 (lv1)	Level 2 (lv2)	Level 3 (lv3)	Level 4 (lv4)
6.1 amines	6 nitrogen containing compounds	6.1 amines		
6.2 amides	6 nitrogen containing compounds	6.2 amides		
6.3 amino acids	6 nitrogen containing compounds	6.3 amino acids		
6.4 nitrophenanthrene carboxylic acids	6 nitrogen containing compounds	6.4 nitrophenanthrene carboxylic acids		
6.5 cyanoglycosides	6 nitrogen containing compounds	6.5 cyanoglycosides		
6.6 nucleosides	6 nitrogen containing compounds	6.6 nucleosides		
6.7.1 pyridines	6 nitrogen containing compounds	6.7 alkaloids	6.7.1 pyridines	
6.7.2 pyrazines	6 nitrogen containing compounds	6.7 alkaloids	6.7.2 pyrazines	
6.7.3 piperidines	6 nitrogen containing compounds	6.7 alkaloids	6.7.3 piperidines	

Continua na próxima página

Tabela 6 – Continuação da página anterior

Level 0 (lv0)	Level 1 (lv1)	Level 2 (lv2)	Level 3 (lv3)	Level 4 (lv4)
6.7.4 pyrrolidines	6 nitrogen containing compounds	6.7 alkaloids	6.7.4 pyrrolidines	
6.7.5 quinolizidines	6 nitrogen containing compounds	6.7 alkaloids	6.7.5 quinolizidines	
6.7.6 tropanes	6 nitrogen containing compounds	6.7 alkaloids	6.7.6 tropanes	
6.7.7 pyrrolizidines	6 nitrogen containing compounds	6.7 alkaloids	6.7.7 pyrrolizidines	
6.7.8 quinolines	6 nitrogen containing compounds	6.7 alkaloids	6.7.8 quinolines	
6.7.9 isoquinolines	6 nitrogen containing compounds	6.7 alkaloids	6.7.9 isoquinolines	
6.7.10 phenanthridines	6 nitrogen containing compounds	6.7 alkaloids	6.7.10 phenanthridines	
6.7.11 quinazolines	6 nitrogen containing compounds	6.7 alkaloids	6.7.11 quinazolines	
6.7.12 indols	6 nitrogen containing compounds	6.7 alkaloids	6.7.12 indols	

Continua na próxima página

Tabela 6 – Continuação da página anterior

Level 0 (lv0)	Level 1 (lv1)	Level 2 (lv2)	Level 3 (lv3)	Level 4 (lv4)
6.7.13 beta-carbolines	6 nitrogen containing compounds	6.7 alkaloids	6.7.13 beta-carbolines	
6.7.14 sesquiterpene alkaloids	6 nitrogen containing compounds	6.7 alkaloids	6.7.14 sesquiterpene alkaloids	
6.7.15 diterpene alkaloids	6 nitrogen containing compounds	6.7 alkaloids	6.7.15 diterpene alkaloids	
6.7.16 steroidalkaloids	6 nitrogen containing compounds	6.7 alkaloids	6.7.16 steroidalkaloids	
6.7.17 others alkaloids	6 nitrogen containing compounds	6.7 alkaloids	6.7.17 others alkaloids	
7.1 lactones	7 organic acids and esters	7.1 lactones		
7.2 benzoic acid derivatives	7 organic acids and esters	7.2 benzoic acid derivatives		
8.1 stilbenes	8 phenylpropanes	8.1 stilbenes		
8.2 coumarins	8 phenylpropanes	8.2 coumarins		
8.3 lignans	8 phenylpropanes	8.3 lignans		

Continua na próxima página

Tabela 6 – Continuação da página anterior

Level 0 (lv0)	Level 1 (lv1)	Level 2 (lv2)	Level 3 (lv3)	Level 4 (lv4)
8.4 chromones and xanthenes	8 phenylpropanes	8.4 chromones and xanthenes		
8.5 cinnamic acid derivatives	8 phenylpropanes	8.5 cinnamic acid derivatives		
8.6 others phenylpropanes	8 phenylpropanes	8.6 others phenylpropanes		
9.1 hydroquinones	9 quinones	9.1 hydroquinones		
9.2 benzoquinones	9 quinones	9.2 benzoquinones		
9.3 naphthoquinones	9 quinones	9.3 naphthoquinones		
9.4 anthraquinones	9 quinones	9.4 anthraquinones		
10.1 catechins	10 tannins	10.1 catechins		
10.2 gallotannins	10 tannins	10.2 gallotannins		
10.3 ellagitannins	10 tannins	10.3 ellagitannins		

4.1.1.2 Tabela *structure*

Descrição

Nesta tabela estão os valores referentes às estruturas químicas.

Conteúdo

casrn: chave primária; neste campo é preenchido pelo *CAS Registry Number*. O número CAS de uma estrutura química é um número com um registro único no banco de dados do *Chemical Abstracts Service*, uma divisão da *American Chemical Society*. Este número de registro é gerado para cada estrutura descrita na literatura. O *CAS Registry Number* é composto por três partes: a primeira pode conter até seis algarismos, a segunda até dois dígitos e a terceira é um código verificador.

Ex.: 59481-48-0.

iupac: nomenclatura IUPAC; este campo será preenchido com uma *string* contendo símbolos e sem espaços (como, por exemplo: -, /, (), [], etc.).

Ex.: (2*Z*,4*R*,8*R*,9*R*,11*R*)-2-(Hydroxymethyl)-11-methyl-7-methylene-6,12-dioxo-5,14-dioxatricyclo[9.2.1.0^{4,8}]tetradeca-1(13),2-dien-9-yl (2*E*)-2-methyl-2-butenolate.

usualn: nome comum de uma substância, nomes sempre iniciando com letra minúscula.

Ex. budlein A.

sysn: este campo é preenchido com o nome sistemático da substância; o mesmo é dado de forma sistemática a um grupo único de substância química, contendo símbolos e letras do alfabeto grego.

Ex.: 1-one-3 β ,10 β -epoxy-8 β -angeloyloxy-15-hydroxy-germacra-2,4,11(13)-trien-6 α ,12-olide.

inchikey: a InChIKey complementa a informação anterior e é dividida em três blocos: o primeiro bloco (14 letras - BMVJFNLJSZHNNS) codifica o esqueleto molecular, o segundo (10 letras - BVPXTQGSSA) codifica tautomeria, estereoquímica, isótopos e ligações e o último bloco (1 letra - N) codifica a versão da classificação.

Ex.: BMVJFNLJSZHNNS-BVPXTQGSSA-N

inchi: a InChITM é um identificador não proprietário para estruturas químicas que pode ser usado em fontes de dados impressos e eletrônicos e que permite anexar vários dados à estrutura. As camadas na cadeia InChI são separadas pelo caractere “/” seguido de uma letra minúscula (exceto para a primeira camada, a fórmula química). As camadas InChI são: “Número da versão InChI”/ “Fórmula Molecular” /c “conectividade dos átomos, excluindo hidrogênios” /h “conectividade dos átomos de hidrogênio” /b “ligações duplas” .

Ex.: InChI=1S/C20H22O7/c1-5-10(2)18(23)26-15-8-20(4)16(22)7-13(27-20)12(9-21)6-14-17(15)11(3)19(24)25-14/h5-7,14-15,17,21H,3,8-9H2,1-2, 4H3/b10-5+,12-6-/t14-,15-,17+,20-/m1/s1

smiles: SMILES é uma notação em linha (método topográfico usando caracteres imprimíveis) para incorporar e representar estruturas químicas e reações. Sua codificação se torna simples, pois os átomos são representados pelos seus símbolos atômicos. Nesta notação, os átomos de hidrogênio são omitidos, os átomos vizinhos são representados ao lado de cada átomo, as ligações simples são omitidas, as duplas são representadas por “=” e as triplas por “#”, as ramificações são representadas por parênteses, anéis são representados pela colocação de dígitos nos dois átomos que o fecham e anéis aromáticos são representados por letras minúsculas, a estereoquímica próxima a ligações duplas são especificadas usando “\” ou “/” e a configuração do carbono por “@” e “@@”.

Ex.: C/C=C(\C)/C(=O)O[C@@H]1C[C@@]2(C(=O)C=C(O2)/C(=C\[C@@H]3[C@@H]1C(=C)C(=O)O3)/CO)C.

reference: referências bibliográficas dos dados inseridos. Este campo nunca deverá permanecer vazio. O estilo utilizado é o seguinte: nome do periódico, (ano), volume:fascículo:primeira página.

Ex.: *Phytochemistry*, (1976), 15:4:525.

class: classificação conforme level 0 (lvl0) da tabela *classification* (seção 4.1.1.1, página 52).

4.1.1.3 Tabela *descriptors*

Descrição

Nesta tabela estão os valores referentes aos descritores calculados para as estruturas presentes na tabela *structure* (seção 4.1.1.2, página 59). Essa tabela é calculada diretamente pelo programa da ChemAxon.

Conteúdo

casrn: número CAS presente na tabela *structure* (ver seção 4.1.1.2, página 59).

mwmono: massa molecular monoisotópica, informação necessária para auxiliar na desreplicação de extratos de Asteraceae empregando-se espectrometria de massas.

nhdonn: *number of donor atoms H bonds* (contagem de átomos doadores de ligação de hidrogênio, número inteiro).

nhacc: *number of acceptor atoms H bonds* (contagem de átomos aceptores de ligação de hidrogênio, número inteiro).

alogp: *ghose-crippen octanol water partition coefficient* (coeficiente de partição octanol-água de Ghose-Crippen).

mlogp: *moriguchi octanol water partition coefficient* (coeficiente de partição octanol-água de Moriguchi).

logs: logS (solubilidade em água).

logd: logD (coeficiente de distribuição).

Pode-se acrescentar mais descritores nesse ponto, uma vez que são calculados através do aplicativo para web da ChemAxon. Ex.: Lipinski/Drogability.

4.1.1.4 Tabela *analytical*

Descrição

Nesta tabela estão os valores referentes às informações analíticas obtidas pelos membros do grupo para as estruturas químicas da tabela *structures* (seção 4.1.1.2, página 59). Pode-se acrescentar novas informações nesta tabela, pois cada membro possui seu método específico de obtenção de tempo de retenção.

Conteúdo

casrn: número CAS presente na tabela *structure* (seção 4.1.1.2, página 59).

trmeoh: valor numérico (método isocrático em CLAE com MeOH–H₂O).

trmecn: valor numérico (método isocrático em CLAE com MeCN–H₂O).

traster: valor numérico, obtido pelo método AsterBioChem. Método atualmente padronizado e validado para padronizar e acelerar o processo de desreplicação de extratos vegetais.

4.1.1.5 Tabela *activity*

Descrição

Nesta tabela estão os valores referentes às informações sobre atividades biológicas obtidas pelos membros do grupo para as substâncias correlacionadas com as respectivas estruturas químicas da tabela *structures* (seção 4.1.1.2, página 59). Atualmente pode-se incluir informações *in house* e informações da literatura.

Conteúdo

casrn: número CAS presente na tabela *structure* (seção 4.1.1.2, página 59).

target: nome da enzima/célula/parasita/bactéria estudada, pode ser preenchida por dados do laboratório (*in house*) ou dados da literatura, mas sempre com preferência a atividades realizadas pelo grupo.

Ex.: COX-1, COX-2, 5-LOX, etc.

ecn: é um esquema de classificação numérica para as enzimas baseado nas reações químicas que catalisam. Cada código de enzimas consiste nas duas letras EC seguidas por 4 números separados por pontos.

organism: nome do organismo de origem da enzima.

Ex.: *Mus musculus*, *Ratus norvegicus*, etc.

bioactivity_type: tipo de atividade biológica informada. Ex.: IC₅₀, EC₅₀, EC₁₀₀, inibição etc.

method: uma rápida descrição da análise ou nome do kit enzimático.

value: valor numérico da atividade cadastrada.

unit: unidade dos resultados. Sempre que possível transformar os resultados de IC₅₀ para $\mu\text{M/mL}$. Ex.: % para inibição e $\mu\text{M/mL}$ para IC₅₀.

reference: referência bibliográfica dos dados de atividade inseridos. Este campo nunca deve permanecer vazio. O formato a ser utilizado será: nome do periódico, (ano), volume:fascículo:primeira página.

Ex.: *Phytochemistry*, (1976), 15:4:525.

4.1.1.6 Tabela *genera*

Descrição

Nesta tabela estão os nomes científicos de todos os gêneros (atualmente 1.539) da família Asteraceae, preenchido previamente pelos administradores do banco de dados AsterDB.

Conteúdo

genus: Ex.: *Artemisia*, *Baccharis*, *Chresta*, *Dahlia*, *Eremanthus*, *Faxonia*, *Goyazianthus*, *Helianthus*, *Irwinia*, *Joseanthus*, *Koanophyllon*, *Lactuca*, *Mikania*, *Nothobaccharis*, *Oxyphyllum*, *Pycnocephalum*, *Russowia*, *Smallanthus*, *Tithonia*, *Voluntaria*, *Werneria*., *Xanthocephalum*, *Yongia*, *Zyzxia*.

4.1.1.7 Tabela *taxonomy*

Descrição

Nesta tabela estão os valores referentes às informações taxonômicas das plantas estudadas pelo grupo AsterBioChem. Essa tabela não tem relação com as anteriores.

Conteúdo

family: esse banco de dados foi desenvolvido com o objetivo de incluir apenas a família Asteraceae, porém essa coluna foi projetada para futuras ampliações.

Ex.: Asteraceae

subfamily: É uma categoria taxonômica que está entre a família e a tribo.

Ex.: Barnadesioideae

tribe: Faz parte da chave de classificação botânica, localizada entre a sub-família e o gênero.

Ex.: Heliantheae

subtribe: Faz parte da chave de classificação botânica, localizada entre a tribo e o gênero.

Ex.: Espeletiinae

genus: Esta coluna deve ser utilizada para um sistema de busca no GUI; para isso, se faz necessário o preenchimento do nome do gênero em uma coluna separada. Está relacionada à tabela *genera* (seção 4.1.1.6, página 62).

Ex.: *Helianthus*

species: Essa coluna é a chave primária desta tabela e deverá estar composta de gênero e da espécie.

Ex.: *Helianthus annuus*

reference: referência bibliográfica dos dados taxonômicos inseridos. Este campo nunca deverá permanecer vazio. O formato a ser utilizado é: nome do periódico, (ano), volume:fascículo:primeira página.

Ex.: *Phytochemistry*, (1976), 15:4:525.

4.1.1.8 Tabela *botany*

Descrição

Nesta tabela estão os valores referentes às informações botânicas das plantas estudadas pelo grupo AsterBioChem. Nesta tabela podem existir várias repetições de entrada (mas, pelo menos 1 por voucher).

Conteúdo

species: gênero e espécies presentes na tabela *taxonomy* (seção 4.1.1.7, página 62).

country: país onde a planta foi coletada.

locality: cidade ou local descritivo onde a planta foi coletada.

Ex.: Serra Dourada, Chapada Diamantina, etc.

habitat: o habitat é o lugar onde se encontrou a espécie.

Ex.: lake, mountain, etc.

soil: tipos de solo.

Ex.: clay, loam, clay-loam, silt-loam, etc.

vegetation_type: tipo de vegetação.

Ex.: tundra, taiga, temperate broadleaf and mixed forest, temperate steppe, subtropical moist forest, mediterranean, monsoon forest, desert, xeric shrubland, dry steppe,

semi-desert, grass savanna, tree savanna, tropical and subtropical dry forest, tropical rainforest, alpine tundra, montane forest, etc.

voucher: identificação científica do material em foco necessita de adequada documentação com espécimes-testemunha (“vouchers”), que devem estar depositados em herbários.

gps: coordenadas de GPS.

4.1.1.9 Tabela *plantparttable*

Descrição

Nesta tabela estão os valores referentes à cada parte da planta estudada. Essa é uma tabela fixa, alterada apenas pelos administradores do AsterDB.

Conteúdo

plantpart: valores únicos a serem cadastrados (*inflorescences, stems, fruits, seeds, barks, stem barks, leaves, aerial parts, underground parts, others*)

4.1.1.10 Tabela *boco*

Descrição

Nesta tabela se faz a união das informações botânicas às informações estruturais.

Conteúdo

species: gênero e espécies presentes na tabela *taxonomy* (seção 4.1.1.7, página 62).

casrn: número CAS presente na tabela *strutcture* (seção 4.1.1.2, página 59).

plantpart: parte da planta estudada presente na tabela *plantparttable* (seção 4.1.1.9, página 64).

locality: localidade, presente na tabela *botany* (seção 4.1.1.8, página 63).

gps: coordenadas de GPS, presente na tabela *botany* (seção 4.1.1.8, página 63).

4.1.1.11 Tabela *structufiles*

Descrição

Nesta tabela estão inseridos os arquivos das informações analíticas.

Conteúdo

casrn: número CAS presente na tabela *structure* (seção 4.1.1.2, página 59).

nmrfile: arquivos originais oriundos de espectros de RMN.

msfile: arquivos originais oriundos de espectros de EM.

jcampfile: arquivos formato jcamp para espectros de RMN e EM.

uv: arquivo de espectros de UV

4.2 ESTUDO DA RETENÇÃO CROMATOGRÁFICA DE LAC- TONAS SESQUITERPÊNICAS (LST)

Em resumo, seis fatores foram avaliados nesta parte do estudo:

- dois diferentes sistemas de solventes (MeOH–H₂O 55:45 e MeCN–H₂O 35:65);
- três grupos de descritores estruturais (2D-descr, 3D-1conf e 3D-weight);
- dois diferentes conjuntos de treino e teste (26:13 e 29:10);
- quatro algoritmos para seleção de descritores (*best-first*, *linear forward* e *greedy stepwise*, GA);
- três diferentes tamanhos de modelos (quatro, cinco e seis descritores);
- dois métodos de modelagem (PLS e ANN).

Dada todas as possíveis combinações usando esses seis fatores, 288 diferentes modelos foram gerados e comparados entre si. Os 40 melhores modelos foram selecionados, seis do sistema MeOH–H₂O e 36 do MeCN–H₂O. A seleção ocorreu seguindo os seguintes critérios: P^2 maior que 0,60 e $R^2 > Q^2 > P^2$.

A Figura 4 (página 66) indica, para cada sistema de solvente, a frequência relativa dos vários fatores investigados (% , eixo *y*) em função da qualidade preditiva dos modelos resultantes (validação externa P^2 , eixo *x*). Dentre os fatores utilizados, o melhor desempenho para o sistema MeOH–H₂O ($P^2 = 0,76$) foi obtido com ANN, enquanto três modelos com $P^2 = 0,91$ foram obtidos para sistema MeCN–H₂O, dois com ANN e um utilizando PLS.

O GA foi o melhor método de seleção de variáveis para o conjunto de estruturas e em ambos os sistemas de solventes. Isso pode ser explicado por duas razões principais: primeiro, o critério utilizado no GA para a seleção do melhor número de variáveis (quatro, cinco e seis descritores) foi a LOF em uma MLR para cada conjunto de descritores (conforme seção 3.3.5, página 37), segundo, os outros métodos de seleção utilizaram a capacidade individual de cada descritor; este último, de fato, reduz o tempo de máquina para a seleção quando usado com grandes volumes de dados, porém não obteve os melhores resultados.

Figura 4 – Gráficos cumulativos de participação dos fatores para cada solvente entre os melhores modelos. O eixo horizontal mostra o coeficiente de determinação da validação externa (P^2); o eixo vertical mostra a participação de cada fator (em percentagem). As maiores áreas (barras verticais coloridas) demonstram abordagens mais bem-sucedidas.

Fonte: Reimpresso e adaptado com permissão de [Oliveira et al.\(1\)](#); copyright 2015 American Chemical Society

Ainda na Figura 4 (página 66), considerando-se os três conjuntos de descritores (2D-descr, 3D-1conf e 3D-weight), não foi possível observar diferenças entre os dois sistemas de solventes. Para MeOH–H₂O e MeCN–H₂O o uso de descritores com as estruturas 2D gerou um maior número de modelos viáveis. Isso pode transparecer uma relativa vantagem, pois não há necessidade de gerar as coordenadas 3D, por conseguinte, reduz-se uma etapa bastante trabalhosa na construção dos modelos. No entanto, os 2D-descr não foram capazes de separar epímeros (por exemplo, os pares de estruturas **29** e **30**, **29** e **30**) e isômeros (por exemplo, os pares de estruturas **8** e **9**, **10** e **11**, e **13** e **14**), todos eles apresentando isomerismo em seus ésteres de cadeia lateral (Figura 1, página 35). Porém, os descritores 2D são capazes de explicar interações mais simples entre as substâncias.

Em dois modelos do sistema MeOH–H₂O, os $\log k$ do furano-heliangolídeo **24** (Figura 5, página 67) foram um dos piores preditos, com erro absoluto de 0,426 e 0,390 nos modelos MeOH140 e MeOH40, respectivamente. O modelo MeOH40 (2D) foi capaz de estimar corretamente o $\log k$ de todos os melampolídeos. Entre os heliangolídeos, em 23% (estruturas **11**, **12**, **20**, **21**, e **24**), o valor do erro absoluto foi maior que 0,13.

Ressalta-se que há um maior número de modelos para o sistema MeCN–H₂O (Figura 6, página 69) com resultados dentro dos critérios adotados no início desta seção quando comparado com os modelos para o sistema MeOH–H₂O (Figura 5, página 67). É, portanto, uma questão interessante, uma vez que os dados obtidos em MeOH–H₂O

Figura 5 – Valores experimentais *versus* calculados de $\log k$ dos modelos MeOH140, MeOH143 e MeOH40 (Tabela 7, página 74). Modelos para o sistema de solventes MeOH–H₂O e as 39 LST combinadas nos grupos de treino e teste externo. Nas ampliações (regiões A - D marcadas em retângulos vermelhos) é possível observar a capacidade dos modelos discriminarem entre os epímeros e isômeros geométricos (pares de estruturas A - **10** e **11**, B - **29** e **30**, C - **13** e **14**, e D - **8** e **9**).

Fonte: Reimpresso e adaptado com permissão de Oliveira et al.(1); copyright 2015 American Chemical Society

também são muito mais difíceis de descrever em um modelo de QSRR do que aqueles obtidos em MeCN–H₂O. A resposta pode estar no fato de que o metanol (um solvente prótico) e a água interagem entre si por ligações de hidrogênio para formar espécies que são diferentes daquelas em que apenas água pura ou metanol são levadas em consideração. Por isso, o sistema poderia, de fato, ser considerado um sistema de solvente ternário (o metanol, água e o metanol-água), tal como sugerido por Bosch et al.(93). Esta condição peculiar, na verdade, gera uma interação muito mais complexa entre os analitos, os dois solventes e a fase estacionária em comparação quando o MeCN (solvente aprótico) e a água são utilizados como fase móvel (94).

Para efeitos comparativos, todos os descritores selecionados pelos diferentes algoritmos de seleção foram analisados em gráficos do tipo radar (Figura 7, página 70). É importante salientar que os descritores de maior intensidade nos gráficos estão presentes nos melhores modelos. Neste gráfico, é possível avaliar os descritores selecionados para os sistemas de solventes e pelos tipos de descritores.

Independentemente da dimensionalidade dos descritores, os fatores de maior importância para a correlação entre eles e os $\log k$ das estruturas foram aqueles que trazem a informação quanto à hidrofobicidade, hidrofiliidade e a presença de grupos polares nas estruturas com suas interações intra- e extramoleculares.

Pode-se observar que os descritores 2D (Figura 7, página 70) que foram selecionados com maior frequência para o sistema de solvente MeOH–H₂O foram ALOGP2 (quadrado do coeficiente de partição *n*-octanol-água de Ghose-Crippen), SIC3 (conteúdo de informação estrutural - simetria de vizinhança de 3ª ordem(95)), maxsssCH [máximo do estado eletro-topológico (*e-state*) dos átomos: >CH- (96, 97)], maxHBint7 (estado máximo eletro-topológico de força para as ligações potenciais de hidrogênio de comprimento 7 (96, 97)), e BIC4 (conteúdo de informação de ligação - simetria de vizinhança de 4ª ordem (36, 37)). Os descritores 2D que foram selecionados com maior frequência para o sistema de solvente MeCN–H₂O também foram o ALOGP2, maxHBint7 e BIC4, destacando-se o descritor RBF (fração de ligações rotacionáveis).

Como mostra a Figura 7 (página 70), o conjunto de descritores 3D-1conf selecionados para o sistema MeOH–H₂O foram: vsurf ID7 (distância entre o centro de massa de uma estrutura e o centro das regiões hidrofóbicas em torno dele a -1,4 e (98, 99)), vsurf D1 (região hidrofóbica da estrutura e é definida quando uma sonda seca está interagindo com um alvo da estrutura a -1,2 e (98, 99)), npr1 (razão do momento de inércia principal normalizada - $pmi1/pmi2$ (75)), glob (globularidade), e ASAP7 (área de superfície acessível fracionada devido a átomos no intervalo de carga parcial > 0,3 e (100)). Os descritores 3D-1conf selecionados para o sistema de solvente MeCN–H₂O foram: vsurf D1, vsurf CW3, CW2 e CW1 (fator de capacidade das estruturas, calculados nos níveis de energia de -1,0, -0,5 e -0,2 kcal/mol, respectivamente (98, 99)), e npr1.

Figura 6 – Valores experimentais *versus* calculados $\log k$ de MeCN137, MeCN143, MeCN45, MeCN47, MeCN96, e Consensus5 (Tabela 8, página 75). Modelos para o sistema de solventes MeCN–H₂O e as 39 LST combinadas dos grupos de treino e teste externo. Nas ampliações (regiões A-D marcadas em retângulos vermelhos) é possível observar a capacidade dos modelos discriminarem entre epímeros e isômeros geométricos (pares de estruturas A - 10 e 11, B - 29 e 30, C - 13 e 14, e D - 8 e 9).

Fonte: Reimpresso e adaptado com permissão de Oliveira et al.(1); copyright 2015 American Chemical Society

Figura 7 – Descritores selecionados nos melhores modelos. As cores indicam os sistemas de solventes (verde = MeOH–H₂O e vermelho = MeCN–H₂O). A distância de um descritor a partir do centro de cada gráfico se correlaciona com a sua importância (frequência de seleção por diversas técnicas utilizadas) e reflete o seu grau de correlação com o log *k*.

Fonte: Reimpresso e adaptado com permissão de [Oliveira et al.\(1\)](#); *copyright 2015 American Chemical Society*

Finalmente, os descritores 3D-weight selecionados foram o vsurf Wp6 (volume polar da estrutura, calculado ao nível energético de $-4,0 \text{ kcal/mol}$ (98, 99)), vsurf DD12 (distância de contato do vsurf DDmin), vsurf D1, vsurf CW2, ASAP1 (fração da área superficial acessível por átomos com carga parcial entre 0 e 0,05 e (100)), ASAN2 (fração da área superficial acessível por átomos com carga parcial entre -0,05 e -0,1 e (100)) para os sistema de solvente MeOH–H₂O e vsurf R (dimensão da superfície rugosa molecular - rugosidade) (99)), vsurf CW2 e CW1, glob, ASAP5 (fração da área superficial acessível por átomos com carga parcial entre 0,2 e 0,25 e (100)) para o sistema de solvente MeCN–H₂O.

A classificação dos melhores modelos foi obtida por validação externa, ou seja, a capacidade do modelo em estimar o $\log k$ de um grupo de estruturas que não fez parte da construção do mesmo. Em outras palavras, os melhores modelos foram selecionados com base no maior P^2 , menor MAE e RMSE.

Os melhores modelos para o sistema de solvente MeOH–H₂O apresentados na Tabela 7 (página 74) apresentam parte dos descritores da Figura 7 (página 70). Todavia, entre os modelos selecionados observa-se o acréscimo de alguns descritores que podem conter importantes informações para auxiliar a compreensão da relação estrutura-retenção, tais como, Eta Epsilon 5 (cômputo da medida dos átomos eletronegativos (96, 97)) e nR03 (número de anéis com 3 membros), que foram selecionados pelo GA para os modelo MeOH40; std dim2 (dimensão padrão 2 - a raiz quadrada do segundo maior *eigenvalue* da matriz de covariância das coordenadas atômicas; uma dimensão padrão é equivalente ao desvio-padrão ao longo de um eixo do componente principal), vsurf CW5 e vsurf CW6 (fator de capacidade das moléculas, estes são calculados nos níveis de energia de -3,0, -4,0 kcal/mol , respectivamente (98, 99)) foram selecionados por GA para o modelo MeOH140.

A Tabela 8 (página 75) apresenta os descritores observados na Figura 7 (página 70) e também observa-se o acréscimo de novos descritores, tais como os descritores IC2 (índice de informação de conteúdo - simetria de vizinhança de 2ª ordem (95)), SHBint2, SHBint7 e SHBint8 (produto do nível *e-state* para átomos aceptores de elétrons e o nível *He-state* de átomos de hidrogênio no doador separados por 2, 7 e 8 ligações químicas) (101)) e minsCH3 (valor mínimo do *e-state* para o fragmento –CH₃) nos modelos MeCN47 e MeCN45. O modelo MeCN96 com os descritores ASAN7, ASAP1 e ASAP4 (fração da área superficial acessível por átomos com cargas parciais nos intervalos: <-0,30, 0 a 0,05 e 0,15 a 0,2 e, respectivamente (100)), pmi1 (primeiro elemento diagonal do tensor oblíquo do momento de inércia) (75)) e vsurf ID8 (distância entre o centro de massa de uma estrutura e o centro de uma região hidrofóbica próxima a -1,6 e (98, 99)). O modelo MeCN137 acrescenta os descritores ASAN6 (fração da área superficial acessível por átomos com carga parcial entre -0,25 a -0,3 e (100)), ASAP5, pmi1, vsurf HB1 (descreve a capacidade de ligação de H de um alvo molecular obtidos com uma sonda polar (98, 99)) e vsurf ID8.

Os modelos para o sistema de solvente MeOH–H₂O (Tabela 7, página 74) apresentaram correlações aceitáveis com um $R^2 = 0,94-0,74$ no treino, $Q^2 = 0,91-0,69$ na validação cruzada e $P^2=0,76-0,65$ na validação externa. A RMSE para o modelo MeOH140 ($P^2=0,76$) foi 0,10 para o treino; 0,13 para a validação cruzada e 0,19 para a validação externa .

Os modelos para o sistema de solvente MeCN–H₂O (Tabela 8, página 75) apresentaram maior correlação estatística com um $R^2 = 0,96-0,92$ para o treino, $Q^2 = 0,93-0,86$ para validação cruzada e $P^2 = 0,91-0,81$ da validação externa. A RMSE para o ANN, GA e 2D-descr foi de 0,08 para o treino, 0,10 para a validação cruzada e 0,11 para a validação externa.

Os descritores 3D descritos acima relacionam a influência da tridimensionalidade no $\log k$ para os modelos construídos. Adicionalmente, as interações dependem da área hidrófila da estrutura e concentração de grupos polares expostos. Por exemplo, como esperado, nenhum dos modelos construídos com os descritores 2D-descr foi capaz de diferenciar corretamente os isômeros geométricos com os grupos angelato e tiglatato pelo $\log k$. O modelo MeCN143 (Tabela 8, página 75), o qual foi construído com os descritores 3D-weight, foi o único capaz de estimar a sequência correta dos epímeros (estruturas **29** e **30**) e os pares de isômeros geométricos **8** e **9**, **10** e **11** e, **13** e **14**.

Embora o desempenho estatístico de cada um dos melhores modelos individuais (Tabela 7, página 74 e 8, página 75) são comparáveis entre si, é difícil julgar qual modelo é o melhor e qual deve ser escolhido para estimar o $\log k$ de novas LST. Além do desempenho estatístico, neste estudo, a separação e a correta sequência de epímeros e isômeros geométricos estimados são características importantes para a escolha do(s) melhor(es) modelo(s). Muitos autores recomendam o uso de modelos consensuais em estudos de QSAR com o objetivo de aumentar a estabilidade e capacidade de generalização das previsões. Em vários casos esses modelos consensuais normalmente tiveram maior poder de previsão externa em comparação com qualquer modelo individual utilizado no modelo de consenso, uma vez que os erros anulam-se mutuamente.

Por esta razão, seguindo uma estratégia que foi provada ser bem sucedida em estudos anteriores (102, 103, 104, 105), neste estudo vários modelos consensuais foram desenvolvidos pela média dos modelos individuais, considerando os dois conjuntos de treino e teste e com o foco em diferenciar os estereoisômeros. Embora os modelos consensuais para o sistema MeOH–H₂O apresentaram uma melhora no desempenho estatístico (Tabela 7, página 74) apenas um modelo para o sistema MeCN–H₂O (Consensus5) foi capaz de separar os isômeros geométricos e epímeros (este modelo contém os descritores 3D-weight) e também houve uma melhora no desempenho (Tabela 8, página 75).

Nesta etapa foi possível construir bons modelos que foram capazes de estimar o $\log k$ de estruturas complexas de terpenoides. Todos os modelos construídos possuíam informações relacionadas a carga e volume, independente da dimensionalidade estudada

(2D, 3D-1conf ou 3D-weight). No entanto, é importante enfatizar que o uso dos descritores Q_frASAs e dos descritores calculados a partir de um conjunto de conformações estruturais e aplicando-se uma média ponderada de Boltzmann está sendo descrito pela primeira vez em estudos de QSRR.

É recomendável o uso do modelo Consensus5 no sistema MeCN–H₂O para estudos de QSRR posteriores utilizando LST, desde que o conjunto de dados (maior ou diferente) inclua estruturas com razoável grau de complexidade ou de diferentes classes. A limitação deste método é que o mesmo pode ser utilizado apenas para LST. No entanto, ele pode ser ajustado acrescentando-se novas estruturas ao modelo de treino, gerando uma melhoria no modelo de predição.

Tabela 7 – Melhores modelos de QSRR para cada variável obtida para o sistema de solvente MeOH–H₂O. R²: coeficiente de determinação; Q²: coeficiente de determinação para a validação cruzada 10-folds; P²: coeficiente de determinação para validação externa; MAE: média do erro absoluto; RMSE: raiz quadrada do erro quadrático médio. (Reimpresso e adaptado com permissão de [Oliveira et al.\(1\)](#); *copyright 2015 American Chemical Society*).

Modelos	Descritores	Tipo de des- critores	Seleção	Modelagem	Treino	R ²	MAE	RMSE	Q ²	MAE	RMSE	P ²	MAE	RMSE
MeOH140	ASAN2, ASAP1, std dim2, vsurf CW5, vsurf CW6	3D-weight	GA	ANN	29	0,88	0,08	0,10	0,77	0,11	0,13	0,76	0,15	0,19
MeOH40	ALOGP ² , Eta Epsilon 5, maxssCH, nR03	2D-descr	GA	ANN	29	0,94	0,06	0,07	0,91	0,07	0,08	0,75	0,11	0,16
MeOH143	AM1 dipole, pmi1, vsurf CW5, vsurf CW6, vsurf EDmin1, vsurf IW5	3D-weight	GA	ANN	26	0,92	0,06	0,08	0,79	0,10	0,13	0,65	0,13	0,16
MeOH15	ALOGP ² , SIC3, BIC4, maxHBint7	2D-descr	GS/LF	ANN	26	0,74	0,09	0,12	0,71	0,10	0,13	0,65	0,13	0,16
MeOH31	ALOGP ² , SIC3, BIC4, maxHBint7, maxssO	2D-descr	LF	ANN	26	0,74	0,09	0,12	0,69	0,11	0,14	0,65	0,13	0,16
Consensus1	-	-	-	-	29	0,92	0,06	0,08	-	-	-	0,77	0,13	0,17
Consensus2	-	-	-	-	26	0,90	0,07	0,09	-	-	-	0,74	0,12	0,14

Tabela 8 – Melhores modelos de QSRR para cada variável para o sistema de solvente MeCN–H₂O. R²: coeficiente de determinação; Q²: coeficiente de determinação para a validação cruzada 10-folds; P²: coeficiente de determinação para validação externa; MAE: média do erro absoluto; RMSE: raiz quadrada do erro quadrático médio. (Reimpresso e adaptado com permissão de [Oliveira et al.\(1\)](#); copyright 2015 American Chemical Society).

Modelos	Descritores	Tipo de des- critores	Seleção	Modelagem	Treino	R ²	MAE	RMSE	Q ²	MAE	RMSE	P ²	MAE	RMSE
MeCN47	ALOGP ² , IC2, SH-Bint2, SHBint7, SHBint8, minsCH3	2D-descr	GA	ANN	26	0,96	0,06	0,08	0,93	0,09	0,10	0,91	0,08	0,11
MeCN45	ALOGP ² , IC2, SH-Bint2, SHBint7, SHBint8, minsCH3	2D-descr	GA	PLS	26	0,96	0,07	0,08	0,92	0,09	0,10	0,91	0,08	0,11
MeCN96	ASAN7, ASAP1, ASAP4, pmi1, vsurf CW3, vsurf ID8	3D-1conf	GA	ANN	29	0,92	0,08	0,11	0,91	0,08	0,11	0,91	0,10	0,12
MeCN41	ALOGP ² , IC2, SH-Bint2, SHBint7, SHBint8	2D-descr	GA	PLS	26	0,94	0,07	0,09	0,90	0,10	0,12	0,90	0,09	0,12
MeCN91	ASAN6, ASAP4, pmi1, vsurf CW3, vsurf ID8	3D-1conf	GA	ANN	26	0,92	0,08	0,10	0,89	0,10	0,12	0,88	0,10	0,13
MeCN137	ASAN6, ASAP5, pmi1, vsurf HB1, vsurf ID8	3D-weigh	GA	PLS	26	0,93	0,08	0,10	0,89	0,10	0,12	0,86	0,18	0,25

Continua na próxima página

Tabela 8 – Continuação da página anterior

Modelos	Descritores	Tipo de des- critores	Seleção	Modelagem	Treino	R ²	MAE	RMSE	Q ²	MAE	RMSE	P ²	MAE	RMSE
MeCN37	ALOGP ² , IC2, SH- Bint2, SHBint7	2D-descr	GA	PLS	26	0,92	0,09	0,10	0,86	0,12	0,14	0,86	0,10	0,14
MeCN133	ASAP5, pmi1, vsurf HB1, vsurf ID8	3D-weight	GA	PLS	26	0,92	0,09	0,10	0,90	0,10	0,12	0,85	0,18	0,26
MeCN95	ASAN7, ASAP1, ASAP4, pmi1, vsurf CW3, vsurf ID8	3D-1conf	GA	ANN	26	0,93	0,08	0,10	0,88	0,11	0,13	0,84	0,11	0,15
MeCN87	ASAP4, pmi1, vsurf CW3, vsurf ID8	3D-1conf	GA	ANN	26	0,92	0,09	0,11	0,86	0,12	0,14	0,83	0,11	0,16
MeCN143	glob, vsurf CW1, vsurf G, vsurf ID7, vsurf ID8, vsurf Wp6	3D-weight	GA	ANN	26	0,95	0,08	0,10	0,90	0,10	0,12	0,81	0,14	0,17
Consensus3	-	-	-	-	29	0,94	0,09	0,12	-	-	-	0,96	0,08	0,09
Consensus4	-	-	-	-	26	0,97	0,06	0,07	-	-	-	0,87	0,10	0,13
Consensus5	-	-	-	-	26	0,97	0,05	0,06	-	-	-	0,94	0,07	0,10

4.3 USO DE FERRAMENTAS DE QUIMIOINFORMÁTICA PARA AVALIAR A TRANSPOSIÇÃO METODOLÓGICA DO TEMPO DE RETENÇÃO DE PRODUTOS NATURAIS

Os tempos de retenção ajustados (t'_R) das substâncias que foram obtidas em diferentes combinações podem ser observados nas Tabelas 9 (página 77), 10 (página 80) e 11 (página 83).

Inicialmente as análises descritivas foram efetuadas com o objetivo de avaliar a existência de diferença significativa entre as análises. Desta forma, o teste de Shapiro-Wilk foi realizado com o t'_R nos valores observados na Tabela 9 (página 77) para avaliar a distribuição (normalidade) dentro de cada grupo. A significância foi estabelecida em um nível alfa de 0,05. As análises estatísticas e análise visual dos histogramas (Figura 8, página 85) mostraram que as amostras não seguem a distribuição normal. Assim, foi necessária a utilização de ferramentas estatísticas não paramétricas para as análises seguintes.

Tabela 9 – Tempos de retenção ajustado (t'_R) das substâncias analisadas no equipamento UHPLC-MS1.

ID	Substância	AsterBioChem 1		
		t'_R puro	t'_R pool de 10	t'_R pool de 50
1	ácido quínico	00,00	00,00	00,00
2	ácido cinâmico	21,98	22,08	22,08
3	ácido <i>trans</i> -4-hidroxicinâmico	-	-	-
4	ácido cumárico	15,02	15,02	15,01
5	ácido cafeico	12,58	12,63	12,62
6	ácido ferúlico	15,74	15,83	15,81
7	ácido 3,4-di- <i>O</i> -cafeoilquínico	16,80	16,85	16,84
8	ácido 4,5-di- <i>O</i> -cafeoilquínico	16,78	16,85	16,84
9	ácido clorogênico	11,17	11,22	11,21
10	ácido 3,5-di- <i>O</i> -cafeoilquínico	16,17	16,34	16,34
11	crisina	28,01	28,09	28,08
12	luteolina	20,09	20,27	20,27
13	apigenina-7- <i>O</i> -glicosídeo	16,53	16,61	16,61
14	galangina	28,76	28,85	28,85
15	kaempferol	22,96	23,01	23,01
16	quercetina	20,39	20,47	20,48
17	miricetina	17,72	17,78	17,77

Continua na próxima página

Tabela 9 – Continuação da página anterior

ID	Substância	t'_R puro	t'_R pool de 10	t'_R pool de 50
18	7',3',4'-trimetoxi-5-hidroxiflavona	32,22	32,29	32,29
19	santina	30,29	30,40	30,40
20	isoquercitrina	15,02	15,02	15,10
21	rutina	14,38	14,45	14,44
22	catequina	11,27	11,39	11,37
23	naringenina	22,66	22,75	22,74
24	hesperidina	16,55	16,64	16,61
25	uridina	00,00	00,00	00,00
26	guanosina	00,00	00,00	00,00
27	timidina	-	-	-
28	ácido protocatecuico	08,50	08,64	08,62
29	7-hidroxi-cumarina	15,72	15,81	15,81
30	escopoletina	15,61	15,71	15,70
31	2-metóxi-1,4-naftoquinona	21,62	21,73	21,72
32	galato de (-)epigallocatequina	12,84	12,92	12,91
33	6-O-isobutirato de trilobolídeo	27,07	27,26	27,27
34	santonina	-	-	-
35	8-hidroxisantonia	31,79	31,92	31,90
36	enidrina	27,09	27,38	27,27
37	uvedalina	29,62	29,70	29,74
38	ambrosina	22,98	23,05	23,18
39	damsina	19,00	19,16	19,22
40	peruvina	-	-	-
41	mexicanina I	-	-	-
42	acetato de diidrohelenalina	-	-	-
43	budleína A	-	21,73	21,72
44	15-desoxigoiazensolídeo	28,47	28,57	28,56
45	goiazensolídeo	21,88	21,96	22,08
46	15-acetoxieremantolídeo B	29,66	29,73	29,74
47	15-hidroxieremantolídeo B	24,56	24,65	24,64
48	tagitinina A	20,69	20,91	20,87
49	arglabina	-	-	-
50	cnicina	18,99	19,08	19,08

Continua na próxima página

Tabela 9 – Continuação da página anterior

ID	Substância	t'_R puro	t'_R pool de 10	t'_R pool de 50
Padrão Externo				
43	budleína A	21,63	-	-
43	budleína A	21,57	-	-
43	budleína A	21,56	-	-
43	budleína A	21,59	-	-

Não há diferença significativa entre os t'_R quando analisados individualmente ou em mistura (*pool*):

(Kruskal-Wallis qui-quadrado = 0,0777, df = 2, p-valor = 0,9619)

Tabela 10 – Tempos de retenção ajustados (t'_R) das substâncias analisadas no UHPLC-MS2.

ID	Substância	IPBP				AsterBioChem 1			
		t'_R puro	t'_R pool de 10	t'_R pool de 50	t'_R puro	t'_R pool de 10	t'_R pool de 50	t'_R pool de 10	t'_R pool de 50
1	ácido quínico	0,00	0,00	0,00	00,00	00,00	00,00	00,00	00,00
2	ácido cinâmico	4,12	4,13	4,11	20,96	20,95	20,95	20,99	20,99
3	ácido <i>trans</i> -4-hidroxicinâmico	2,66	2,67	2,67	13,77	13,90	13,92	13,92	13,92
4	ácido cumárico	2,67	2,67	2,67	13,96	13,89	13,92	13,92	13,92
5	ácido cafeico	2,16	2,16	2,16	11,22	11,30	11,32	11,32	11,32
6	ácido ferúlico	2,92	2,91	2,91	14,95	14,90	14,93	14,93	14,93
7	ácido 3,4-di- <i>O</i> -cateoilquínico	3,31	3,30	3,30	16,81	16,86	16,88	16,88	16,88
8	ácido 4,5-di- <i>O</i> -cateoilquínico	3,31	3,30	3,30	16,86	16,88	16,88	16,88	16,88
9	ácido clorogênico	1,98	1,98	1,97	10,12	10,10	10,12	10,12	10,12
10	ácido 3,5-di- <i>O</i> -cateoilquínico	3,17	3,17	3,17	16,22	16,20	16,21	16,21	16,21
11	crisina	5,48	5,49	5,49	26,71	26,68	26,69	26,69	26,69
12	luteolina	3,93	3,94	3,94	19,51	19,62	19,68	19,68	19,68
13	apigenina-7- <i>O</i> -glicosídeo	3,26	3,28	3,27	16,55	16,63	16,72	16,72	16,72
14	galangina	5,64	5,64	5,64	27,48	27,54	27,55	27,55	27,55
15	kaempferol	4,45	4,48	4,44	22,00	22,08	22,02	22,02	22,02
16	quercetina	3,94	3,97	3,94	19,69	19,77	19,76	19,76	19,76
17	miricetina	3,41	3,41	3,41	17,18	17,23	17,26	17,26	17,26
18	7',3',4'-trimetoxi-5-hidroxiflavona	6,43	6,42	6,42	30,59	30,60	30,55	30,55	30,55
19	santina	6,03	6,04	6,02	28,93	28,93	28,99	28,99	28,99

Continua na próxima página

Tabela 10 – Continuação da página anterior

ID	Substância	t'_R puro	t'_R pool de 10	t'_R pool de 50	t'_R puro	t'_R pool de 10	t'_R pool de 50
20	isoquercitrina	2,96	2,95	2,95	15,12	15,10	15,13
21	rutina	2,84	2,85	2,85	14,58	14,62	14,64
22	catequina	1,97	1,98	1,97	10,21	10,15	10,16
23	naringenina	4,35	4,35	4,35	21,69	21,74	21,79
24	hesperidina	3,29	3,30	3,30	16,79	16,86	16,88
25	uridina	0,00	0,00	0,00	00,00	00,00	00,00
26	guanosina	0,00	0,00	0,00	00,00	00,00	00,00
27	timidina	0,42	0,41	0,42	01,63	01,68	01,59
28	ácido protocatecuico	0,90	0,81	0,88	03,99	03,95	03,96
29	7-hidroxi-cumarina	2,83	2,84	2,82	14,51	14,46	14,46
30	escopoletina	2,90	2,89	2,91	14,67	14,69	14,64
31	2-metoxi-1,4-naftoquinona	4,07	4,06	4,06	20,08	20,04	20,05
32	galato de (-)epigallocatequina	2,42	2,40	2,41	12,49	12,55	12,56
33	6-O-isobutirato de trilobolideo	5,37	5,36	5,37	26,06	26,17	26,17
34	santonina	-	4,53	4,53	-	22,16	21,76
35	8-hidroxisantonina	6,22	6,20	6,20	30,00	30,02	30,00
36	enidrina	5,29	5,29	5,29	25,87	25,94	25,93
37	uvedalina	5,83	5,82	5,83	28,22	28,31	28,32
38	ambrosina	4,39	4,40	4,41	21,67	21,72	21,79
39	damsina	3,64	3,66	3,65	18,21	18,26	18,33
40	peruvina	3,92	3,92	3,94	19,40	19,42	19,53

Continua na próxima página

Tabela 10 – Continuação da página anterior

ID	Substância	t'_R puro	t'_R pool de 10	t'_R pool de 50	t'_R puro	t'_R pool de 10	t'_R pool de 50
41	mexicanina I	3,47	3,47	3,47	17,29	17,35	17,42
42	acetato de diidrohelenalina	4,41	4,41	4,41	21,67	21,72	21,76
43	budleína A	4,21	4,22	4,22	20,78	20,79	20,94
44	15-desoxigoiazensolideo	5,56	5,57	5,55	26,95	27,08	27,08
45	goiazensolideo	4,13	4,23	4,24	21,18	21,17	21,18
46	15-acetoxieremantolideo B	5,839	5,829	5,831	28,32	28,39	28,41
47	15-hidroxieremantolideo B	4,83	4,82	4,82	23,81	23,85	23,84
48	tagitinina A	4,07	4,06	4,07	20,19	20,31	20,20
49	arglabina	5,87	5,85	5,85	28,15	28,27	28,32
50	cnicina	3,78	3,78	3,78	18,90	18,87	18,90
Padrão Externo							
34	santonina	4,54	-	-	22,22	-	-
34	santonina	4,54	-	-	22,20	-	-
34	santonina	4,54	-	-	22,20	-	-
34	santonina	4,53	-	-	22,06	-	-
34	santonina	4,52	-	-	22,12	-	-
34	santonina	4,52	-	-	22,22	-	-
34	santonina	4,53	-	-	22,19	-	-

Não pode ser observada diferença significativa entre os métodos IPBP e AsterBioChem 1 t'_R quando as substâncias foram analisadas individualmente ou em mistura (pool): Kruskal-Wallis qui-quadrado = 0,0019, df = 2, p-valor = 0,999 e Kruskal-Wallis qui-quadrado = 0,0242, df = 2, p-valor = 0,988, respectivamente.

Tabela 11 – Tempos de retenção ajustados (t'_R) das substâncias analisadas no UPLC-DAD.

ID	Substância	IPBP	AsterBioChem 1	AsterBioChem 2
		t'_R pool de 10	t'_R pool de 10	t'_R pool de 10
1	ácido quínico	0,00	00,00	00,00
2	ácido cinâmico	3,26	19,75	17,42
3	ácido <i>trans</i> -4-hidroxicinâmico	1,87	12,88	11,49
4	ácido cumárico	1,88	12,88	11,48
5	ácido cafeico	1,45	10,54	09,19
6	ácido ferúlico	2,10	13,82	15,21
7	ácido 3,4-di- <i>O</i> -cafeoilquínico	2,35	15,67	15,21
8	ácido 4,5-di- <i>O</i> -cafeoilquínico	2,49	15,72	12,59
9	ácido clorogênico	1,28	09,58	08,87
10	ácido 3,5-di- <i>O</i> -cafeoilquínico	2,35	15,02	14,57
11	crisina	4,63	25,58	24,00
12	luteolina	3,10	18,49	17,71
13	apigenina-7- <i>O</i> -glicosídeo	2,45	15,46	14,72
14	galangina	4,78	26,36	24,68
15	kaempferol	3,60	20,89	19,93
16	quercetina	3,10	18,59	17,72
17	miricetina	2,57	16,10	15,37
18	7',3',4'-trimetoxi-5-hidroxiflavona	5,56	29,40	27,98
19	santina	5,17	27,74	26,30
20	isoquercitrina	2,12	13,95	13,44
21	rutina	2,03	14,68	13,06
22	catequina	1,26	09,57	08,73
23	naringenina	3,48	20,57	19,31
24	hesperidina	2,49	15,67	14,81
25	uridina	0,00	00,00	00,00
26	guanosina	0,00	00,00	00,00
27	timidina	0,47	01,94	1,54
28	ácido protocatecuico	0,79	04,66	03,06
29	7-hidroxi-cumarina	2,03	13,43	11,94
30	escopoletina	2,10	13,57	12,46
31	2-metoxi-1,4-naftoquinona	3,22	18,89	17,07
32	galato de (-)epigallocatequina	1,60	11,45	10,98

Continua na próxima página

Tabela 11 – Continuação da página anterior

ID	Substância	t'_R pool de 10	t'_R pool de 10	t'_R pool de 10
33	6-O-isobutirato de trilobolídeo	4,55	26,47	23,58
34	santonina	3,69	20,99	19,59
35	8-hidroxisantonina	-*	27,06	29,64
36	enidrina	4,44	24,66	23,36
37	uvedalina	4,97	27,06	25,71
38	ambrosina	3,31	20,54	17,93
39	damsina	2,84	17,11	15,78
40	peruvina	3,10	16,60	16,98
41	mexicanina I	2,65	16,21	14,84
42	acetato de diidrohelenalina	3,57	20,54	19,03
43	budleína A	3,40	19,67	18,29
44	15-desoxigoiazensolídeo	4,70	25,78	24,22
45	goiazensolídeo	3,41	19,93	18,45
46	15-acetoxieremantolídeo B	4,99	27,17	25,67
47	15-hidroxieremantolídeo B	4,00	22,62	21,04
48	tagitinina A	3,22	17,75	17,71
49	arglabina	4,99	26,99	25,11
50	cnicina	2,98	17,75	16,71
Padrão Externo (puro)				
34	santonina	3,69	20,96	19,81
34	santonina	3,67	20,98	19,82
34	santonina	3,65	20,97	19,79

* Não foi possível identificar essa substância por UV.

Diferença entre as substâncias analisadas individualmente e em misturas

Uma vez que a suposição de normalidade foi violada, o teste de Kruskal-Wallis foi selecionado para a avaliação da igualdade entre as três análises. Se o p-valor é maior que 0,05 e o valor do teste estatístico é menor do que o qui-quadrado tabulado, a hipótese nula não é rejeitada, sendo que a hipótese nula foi “as médias dos três grupos são estatisticamente iguais”.

O resultado do teste de Kruskal-Wallis demonstrou que não houve uma diferença significativa no t'_R quando analisados individualmente ou por grupos entre os t'_R das substâncias avaliadas nos métodos ABC1 nos equipamentos UHPLC -MS1, IPBP no equipamento

Figura 8 – Distribuição do t'_R de 50 substâncias em três diferentes análises. a) Método ABC1 no equipamento UHPLC-MS2 (D: análise individual das substâncias, E: mistura de 10 substâncias e, F: mistura de todas as substâncias disponíveis). b) Método ABC1 no equipamento UHPLC-MS2 (D: análise individual das substâncias, E: mistura de 10 substâncias e, F: mistura de todas as substâncias disponíveis). c) Método IPBP no equipamentos UHPLC-MS2 (G: análise individual das substâncias, H: mistura de 10 substâncias e, I: mistura de todas as substâncias disponíveis). A linha em cada gráfico representa a distribuição normal esperada. Dados: Tabelas 9 (página 77) e 10 (página 80).

(a) Método ABC1 no equipamento UHPLC-MS1 (A: análise individual das substâncias, B: mistura de 10 substâncias e, C: mistura de todas as substâncias disponíveis)

(b) Método ABC1 no equipamento UHPLC-MS2 (D: análise individual das substâncias, E: mistura de 10 substâncias e, F: mistura de todas as substâncias disponíveis)

(c) Método IPBP no equipamento UHPLC-MS2 (G: análise individual das substâncias, H: mistura de 10 substâncias e, I: mistura de todas as substâncias disponíveis)

Fonte: Oliveira, T.B. (2015)

UHPLC-MS2, e ABC1 no equipamento UHPLC-MS2 (ver nota de rodapé das Tabelas 9, página 77 e 10, página 80).

Assim, utilizou-se a média desses dados nas análises subsequentes.

Diferença entre equipamentos

O resultado do teste de Kruskal-Wallis não revelou diferença significativa entre os t'_R no método ABC1 quando realizado nos equipamentos UHPLC-MS1, UHPLC-MS2 e UPLC-DAD (qui-quadrado de Kruskal-Wallis = 2,1845, df = 2, p-valor = 0,3355).

Além do teste de Kruskal-Wallis, o coeficiente de correlação (R) foi calculado entre as substâncias analisadas nos diferentes equipamentos. Observou-se que existe uma associação linear positiva entre todas as substâncias estudadas, independentemente do equipamento utilizado ($0,995 < R < 0,997$, Tabela 12, página 86). Para cada aumento de 1 minuto no t'_R em um dos três equipamentos, espera-se o aumento médio de 0,923 até 1,072 do t'_R , quando utilizado o método ABC1.

Tabela 12 – Análise de correlação entre os equipamentos avaliados utilizando o método ABC1.

Interação	R*	Intercepto	$\sigma^\dagger$	coeficiente angular	$\sigma^\ddagger$
UHPLC-MS1 x UHPLC-MS2	0,995	0,474	0,319	1,023	0,016
UHPLC-MS2 x UHPLC-MS1	0,995	-0,275 [§]	0,316	0,968	0,015
UHPLC-MS1 x UPLC-DAD	0,995	0,662	0,328	1,072	0,017
UPLC-DAD x UHPLC-MS1	0,995	-0,425	0,312	0,923	0,015
UHPLC-MS2 x UPLC-DAD	0,997	0,211	0,201	1,048	0,011
UPLC-DAD x UHPLC-MS2	0,997	-0,114 [§]	0,193	0,950	0,010

* coeficiente de correlação; † erro padrão do intercepto; ‡ erro padrão do coeficiente angular; [§] não significativo.

Para o método IPBP o teste de Kruskal-Wallis mostrou uma leve diferença significativa entre t'_R quando as eluições foram realizadas nos equipamentos UHPLC-MS2 e UPLC-DAD (qui-quadrado de Kruskal-Wallis = 7,4325, gl = 1, p-valor = 0,006406). Uma vez que o método IPBP é rápido (15 minutos), qualquer ligeira alteração pode causar alteração nos resultados obtidos nos diferentes equipamentos.

No entanto, observou-se que existe uma relação linear positiva entre todas as substâncias estudadas, independentemente do equipamento utilizado ($R = 0,993$, Tabela 13, página 87). Para cada aumento de 1 minuto no t'_R em um equipamento, espera-se um aumento médio de 0,875 até 1,127 t'_R quando o método IPBP é usado.

Portanto, mesmo com a forte correlação (R) observada no método IPBP entre os dois equipamentos houve diferença significativa observada pelo teste de Kruskal-Wallis,

Tabela 13 – Análise de correlação entre os equipamentos avaliados utilizando o método IPBP.

Interação	R*	Intercepto	$\sigma^{\dagger}$	coeficiente angular	$\sigma^{\ddagger}$
UHPLC-MS2 x UPLC-DAD	0,993	0,383	0,062	1,127	0,019
UPLC-DAD x UHPLC-MS2	0,993	-0,295	0,060	0,875	0,015

* coeficiente de correlação; $\dagger$ erro padrão do intercepto; $\ddagger$ erro padrão do coeficiente angular.

acredita-se que isto ocorreu devido as variações entre os equipamentos dos t'_R ter média de 0,747 minutos (mediana = 0,828) e essa diferença de quase 1 minuto é significativa quando um método rápido é utilizado por ele ser mais sensível quanto as variações de tempo .

Diferença entre métodos

Com a finalidade de descrever a relação entre os diferentes métodos e para estimar o tempo de retenção de uma substância analisada em apenas um dos métodos, foi realizada uma análise por regressão linear. Nesta etapa foram utilizados apenas os dados obtidos pelos equipamentos acoplados a um espectrômetro de massas, uma vez que este método é mais preciso para estudos metabolômicos que incluem desreplicação.

Figura 9 (página 88) mostra o gráfico e a equação de regressão linear dos t'_R das 50 substâncias deste estudo no equipamento UHPLC-MS2. Obviamente, há uma estreita correlação entre ambos os métodos (ABC1 e IPBP) e os dados representam baixa dispersão em torno da linha de tendência com o coeficiente angular e intercepto de 4,8 e 0,611, respectivamente. Em comparação com o gráfico da Figura 10 (página 89), com exceção do ácido protocatecuico (substância 28), o modelo também foi capaz de obter uma correlação linear entre os dois métodos.

Estudo de QSRR

As substâncias foram divididas em um grupo de treino ($n = 34$) e um conjunto de teste ($n = 16$, Figura 2 (página 41), estruturas: **4, 6, 7, 8, 15, 16, 21, 27, 30, 37, 38, 40, 42, 45, 47 e 48**). A fim de garantir uma cobertura ótima do espaço químico para ambos os subgrupos, esta divisão foi baseada numa análise por HCA com base no cálculo da distância Euclidiana (equação 3.10, página 45) dos *fingerprints* PubChem. Esses *fingerprints* são calculados de forma que para cada estrutura química é construído um vetor com informações binárias (1/0 bits). Cada *bit* representa uma determinação booleana; em outras palavras, um teste para a presença ou ausência de um determinado fragmento; por exemplo, a presença de um átomo de carbono, o tipo de um anel aromático, o ambiente atômico, etc. A diversidade estrutural foi calculada com base nas estruturas 2D. As 36 estruturas foram selecionadas usando *complete-linkage clustering*, onde pelo menos uma estrutura foi selecionada visualmente a partir de cada conjunto.

Figura 9 – Valores de t'_R pelo método ABC1 *versus* t'_R obtidos pelo método IPBP para as 50 substâncias deste estudo no equipamento UHPLC-MS2 (Tabela 9, página 77).

Fonte: Oliveira, T.B. (2015).

De acordo com as avaliações anteriores, pôde-se assumir uma correlação entre os métodos, por meio de regressão linear. Assim, os estudos de QSRR foram aplicados aos métodos IPBP e ABC1, os quais foram realizados no equipamento UHPLC-MS2, uma vez que em tal equipamento teve disponível o maior número de substâncias.

Nesse estudo optou-se por usar duas estratégias de modelagem. Primeiro, criou-se dois modelos utilizando duas ferramentas que carregam em si a capacidade de redução de variáveis e são robustas quanto à presença de variáveis correlacionadas: PLS como uma ferramenta de análise multivariada e árvore de decisão como ferramenta de mineração de dados.

Todos os descritores 2D e 3D foram calculados e unidos como variáveis independentes (2.649 descritores) a fim de construir um modelo de PLS e uma árvore de decisão com o algoritmo M5P. Na primeira etapa, todos os descritores com valores zero de variância e variação próxima a zero foram excluídos totalizando 1.116 descritores. O modelo PLS foi definido com 4 PLS-componentes (variáveis latentes) após análise de validação cruzada e a árvore de decisão buscando-se minimizar o número de instâncias. Os dados estatísticos podem ser observados na Tabela 14 (página 92).

A segunda estratégia foi avaliar métodos diferentes de seleção de variáveis, LFS e GA, com a mesma ferramenta de modelagem (MLR).

Figura 10 – Valores de t'_R pelo método ABC1 obtidos pelo equipamento UHPLC-MS1 *versus* t'_R obtidos pelo método IPBP no equipamento UHPLC-MS2 das 43 substâncias deste estudo (Tabela 10, página 80).

Fonte: Oliveira, T.B. (2015).

Cada grupo de descritores [2D-MOE (175 descritores), 2D-Padel (2.327 descritores) e 3D-weigh (147 descritores)] foi pré-tratado separadamente antes de se prosseguir para o passo de seleção dos descritores. No primeiro passo, todos os descritores com valores zero de variância e variância próximo a zero foram excluídos. No segundo passo foram excluídos os descritores com um elevado grau de correlação de pares. No final foram obtidos 26 (2D-MOE), 139 (2D-Padel) e 48 (3D-weigh) descritores. Estes 213 descritores foram unidos em uma única matriz para se prosseguir com as seleções por LFS e GA.

Para a seleção do subconjunto de cinco descritores utilizando-se o método LFS houve uma concordância nos resultados para os métodos cromatográficos IPBP e ABC1, uma vez que foi selecionado o mesmo conjunto de descritores: a ICM, logS, PEOE VSA FPNEG, vsurf CW2 e vsurf EWmin2 (LFS5, Figura 11, A, página 91) e para a seleção do subconjunto de seis descritores o mesmo conjunto de descritores acima citado foi selecionado acrescentado-se o descritor vsurf HL2 para ambos métodos cromatográficos (LFS6, Figura 11, B, página 91).

O método GA apresentou uma concordância apenas para a seleção do subconjunto de cinco descritores e quando observado os dois melhores modelos selecionados (ranqueados conforme seção 3.3.5, página 37). Assim, para esta seleção foram construídos modelos com os dois melhores conjuntos de descritores selecionados (GA5, Figura 11, C;

GA5a, Figura 11, **D**; GA5b, Figura 11, **E**, página 91). No método GA para a seleção do subconjunto de seis descritores não houve acordo entre as 10 melhores seleções, por isso foi selecionado apenas o melhor modelo obtido pela validação cruzada e menor RMSE (GA6a e GA6b, Figura 11, **F** e **G**, página 91).

Figura 11 – Gráfico da matriz de correlação. A parte triangular inferior indica a correlação de Pearson e a parte triangular superior é uma informação gráfica da matriz de correlação. As cores dos círculos indicam os sinais: positivo = azul; negativo = vermelho. O tamanho dos círculos indicam os valores absolutos da correlação. A - cinco descritores selecionados por LFS para os métodos ABC1 e IPBP (LFS5); B - seis descritores selecionados por LFS para os métodos ABC1 e IPBP (LFS6); C - cinco descritores selecionados por GA para os métodos ABC1 e IPBP (GA5); D - cinco descritores selecionados por GA para IPBP (GA5a); E - cinco descritores selecionados por GA para ABC1 (GA5b); F - seis descritores selecionados por GA para IPBP (GA6a); e G - seis descritores selecionados por GA para ABC1 (GA6b).

Fonte: Oliveira, T.B. (2015).

Tabela 14 – Modelos obtidos para 50 substâncias de origem natural utilizando descritores 2D e 3D. O conjunto de treino e teste foi de 34 e 16, respectivamente.

Seleção - Algoritmo	Fatores	UHPLC-MS2									
		ABC1					IPBP				
		Training set		Test set		R ²	Training set		Test set		P ²
		MAE	RMSE	MAE	RMSE		MAE	RMSE	MAE	RMSE	
PLS-PLS1	4 PLS-Componentes	1,49	2,12	1,61	1,75	0,94	0,31	0,45	0,36	0,40	
Árvore de Decisão-MP5T	-	2,02	2,68	2,57	3,09	0,90	0,45	0,55	0,52	0,60	
LFS5-MLR	5 descritores	2,23	3,11	2,28	2,80	0,86	0,46	0,66	0,52	0,63	
LFS6-MLR	6 descritores	2,22	3,10	2,29	2,82	0,86	0,46	0,66	0,52	0,63	
GA5-MLR	5 descritores	1,69	2,63	1,85	2,22	0,90	0,37	0,57	0,42	0,51	
GA5a-MLR	5 descritores	-	-	-	-	0,90	0,44	0,57	0,67	0,88	
GA5b-MLR	5 descritores	2,16	2,95	2,26	2,64	-	-	-	-	-	
GA6a-MLR	5 descritores	-	-	-	-	0,90	0,43	0,57	0,52	0,64	
GA6b-MLR	6 descritores	1,56	1,89	3,42	4,25	-	-	-	-	-	

Para avaliar os modelos com maior capacidade de predição foram construídos gráficos para relacionar os valores reais e os valores preditos para os melhores modelos (Figuras 12 - página 93, 13 - página 94, 14 - página 95 e 15 - página 96).

Figura 12 – Valores dos t'_R preditos *versus* experimentais de 50 substâncias para o método ABC1 calculados por PLS. Círculos mostram os valores das 34 substâncias utilizadas no treino e triângulos os valores estimados para o teste externo (16 substâncias).

Fonte: Oliveira, T.B. (2015).

É importante salientar que em um processo de desreplicação de matrizes vegetais as substâncias consideradas como “piores casos” são aquelas com mesma massa molecular. Assim, com o objetivo de analisar a capacidade preditiva dos melhores modelos, decidiu-se por comparar as substâncias com a mesma massa molecular.

As substâncias **3** e **4** não foram separadas pelos seus t'_R experimentais; logo não se poderia esperar diferenças do tempo de predição das mesmas. Fato similar ocorre quando se comparam as substâncias **7** e **8**. Tanto os modelos construídos por PLS quanto GA-MLR estimaram t'_R similares. Entretanto quando se compara a substância **10** com as substâncias **7** e **8** apenas o modelo GA-MLR foi capaz de prever a sequência correta de separação. Os pares de substâncias **12** e **15**; **34** e **38** foram separadas na sequência correta pelos modelos de PLS e GA-MLR.

Nesta seção, a avaliação de diferentes métodos cromatográficos, equipamentos e algoritmos para seleção de variáveis e ferramentas de modelagem permitiu propor bons modelos de QSRR em cromatografia líquida de fase reversa com substâncias de diferentes classes químicas. Não houve diferença estatística entre os valores de t'_R das substâncias analisadas individualmente e aquelas analisadas em mistura. A correlação entre os dois

Figura 13 – Valores dos t'_R preditos *versus* experimentais de 50 substâncias para o método IPBP calculados por PLS. Círculos mostram os valores das 34 substâncias utilizadas no treino e triângulos os valores estimados para o teste externo (16 substâncias).

Fonte: Oliveira, T.B. (2015).

métodos e equipamentos foi reprodutível. Estas análises mostraram que é possível desenvolver modelos de QSRR para um método cromatográfico e equipamento e transpô-los para outro equipamento seguindo o uso de substâncias em comum.

O melhor modelo para prever t'_R foi desenvolvido usando PLS, uma vez que utiliza a informação de todos os descritores em apenas 4 PLS-componentes. O modelo de QSRR resultante pode ser usado para prever os dados de retenção para uma ampla gama de produtos naturais, com razoável precisão. Após os devidos acertos, especialmente com respeito à possibilidade de transferência entre diferentes sistemas cromatográficos, esse método de criação de modelos de QSRR pode-se tornar uma ferramenta valiosa para a desreplicação de metabólitos secundários em extratos de plantas.

4.4 ESTUDOS DE QSRR DE UM BANCO DE DADOS COMERCIAL

As estruturas ($n = 2.635$) foram classificadas conforme descrito na seção 3.5.2 (página 45) em 12 grupos/*clusters* a saber: **a** ($n = 123$), **b** ($n = 78$), **c** ($n = 136$), **d** ($n = 168$), **e** ($n = 40$), **f** ($n = 200$), **g** ($n = 849$), **h** ($n = 128$), **i** ($n = 438$), **j** ($n = 92$), **k** ($n = 303$) e **l** ($n = 80$).

Para auxiliar a compreensão do espaço químico utilizado neste estudo e quais

Figura 14 – Valores do t'_R predito *versus* experimental de 50 substâncias para o método ABC1 calculados por GA-MLR. Círculos mostram os valores das 34 substâncias utilizadas para treino e triângulos os valores estimados para o teste externo (16 substâncias).

Fonte: Oliveira, T.B. (2015).

informações estruturais foram mais importantes na separação dos grupos, efetuou-se uma LDA (Figura 16, página 97) e com dois eixos discriminantes (LD1 e LD2) foi possível explicar 86,6% dos agrupamentos.

Um dos fragmentos mais importante para a separação foi o KRFPC3706 (código smiles = CCCCC), uma vez que o grupo **b** possui estruturas com valores superiores a 92 unidades, o grupo **a** com valor entre 92 e 15 unidades, e os grupos **d**, **e** e **f** possuem estruturas com valor superior a 15 unidades e os grupos **g**, **i** e **l** com valor igual ou inferior a 15 unidades.

Dentro do agrupamento com KRFPC3706 >15 unidades, o fragmento KRFPC3712 (código smiles = CCCCC) com valor superior a 63 ficam os grupos **a** e **b** e os demais grupos (**d**, **e** e **f**) possuem valores menores ou iguais a 63 unidades. Dentro do grupo **e**, cerca de 90% das estruturas possuem menos de 35 fragmentos KRFPC3719 (código smiles = CCCCCC).

Os fragmentos KRFPC3659 (código smiles = CCC(C)C) são limitantes para os grupos **i** e **g** e ambos possuem no máximo seis desses fragmentos. Os grupos **c**, **j**, **k** e **l** possuem contagem máxima igual a 21.

As 250 estruturas classificadas no grupo **g** não possuem o fragmento KRFPC3640 (código smiles = CCC). As demais possuem contagem máxima de 11 unidades. Nenhuma

Figura 15 – Valores dos t'_R preditos *versus* experimentais de 50 substâncias para o método IPBP calculados por GA-MLR. Círculos mostram os valores das 34 substâncias utilizadas para treino e triângulo os valores estimados para o teste externo (16 substâncias).

Fonte: Oliveira, T.B. (2015).

estrutura agrupada em **i** possui o fragmento KRFPC3657 (código smiles = CCC(C)(C)C), 99,1% das estruturas agrupadas como **g**, 98,9% agrupadas como **j** e 90% das agrupadas como **c**, **k** e **l** também não possuem o fragmento KRFPC3657. O fragmento KRFPC3660 (código smiles = CCC(C)CC) inexistente em 97,8% das estruturas classificadas no grupo **g**. Para o fragmento KRFPC3369 (código smiles = CC(C)C), 99,2% do grupo **g** e 99,9% do grupo **i** não possuem ou possuem no máximo 2 fragmentos.

O fragmento KRFPC3692 (código smiles = CCCC), assim como os demais fragmentos carbônicos sem ramificação, auxiliam na separação entre os grupos quanto ao tamanho das estruturas. É interessante observar que, mesmo com a elevada diversidade química a classificação se deu principalmente devido a fragmentos simples contendo apenas átomos de carbono.

Vale ressaltar que o fragmento KRFPC3956 (código smiles = CO) não esteve presente em apenas 123 estruturas utilizadas neste estudo. O fragmento KRFPC4080 (código smiles = N) está presente em todas as estruturas do grupo **l** e em nenhum dos grupos **b** e **e**. Considerando os demais grupos (**a**, **c**, **d**, **f**, **g**, **h**, **i**, **j** e **k**), o fragmento KRFPC4080 está presente de 1,1% (uma estrutura) no grupo **j** a 18,1% no grupo **k** (55 estruturas).

Figura 16 – Análise linear discriminante (LDA), revelando os Klekota-Roth *Fingerprinter Counts* (KRFPC) mais importantes para o agrupamento com apenas duas lineares discriminantes (LD1 e LD2).

Fonte: Oliveira, T.B. (2015).

A distribuição das estruturas em relação aos RRT está entre 200 e 3800 min x 100, com desvio padrão médio de 13,43. Na Figura 17 (página 98) pode-se observar que existe uma distribuição estrutural em todas as faixas de tempo.

Na criação dos modelos quatro diferentes fatores foram avaliados, a saber:

- um teste externo comum a todos os modelos (25%) e separação entre teste e treino;
- três métodos de separação para teste e treino (preservando a distribuição do RTT (prt) e baseados nos preditores (2D - som2D e 3D - som3D);
- três diferentes tamanhos de modelos selecionados por GA (quatro, cinco e seis descritores);
- dois métodos de modelagem (MLR e BRNN).

O GA foi otimizado conforme descrito na seção 3.3.5 (página 37). O critério de desempenho aqui escolhido foi a diminuição do RMSE e aumento do coeficiente de de-

Figura 17 – Distribuição das estruturas químicas pelo $t_R \times 100$ (RRT).

Fonte: Oliveira, T.B. (2015).

terminação da validação cruzada *leave-one-out* no MLR dos descritores em relação ao valor-alvo, RTT.

As combinações dos mesmos descritores utilizados no MLR foram escolhidas para posterior construção dos modelos utilizando-se o BRNN. O seleção do número de neurônios e os parâmetros de ajuste da rede foram efetuadas baseando-se na menor RMSE de uma validação cruzada *10-folds* repetidos por 10 vezes. Os descritores selecionados e os parâmetros estatísticos dos modelos criados podem ser observados no Apêndice B (página 121) e na Figura 18 (página 99).

Dentre os 2D-descr, os mais selecionados foram o ETA dEpsilon D (medida da contribuição de átomos doadores de ligação de hidrogênio (97)), log S (solubilidade em água - mol/L (106)) e MLFER S (dipolaridade/polarizabilidade combinados) (107)) e o descritor BIC0 (conteúdo de informação de ligação - simetria de vizinhança de ordem 0) (36, 37)). Os três primeiros descritores (ETA dEpsilon D, log S e MLFER S) estão relacionados às informações que podem descrever a interação entre soluto e solvente e o BIC0 não está diretamente relacionado a simetria, uma vez que é de ordem 0, entretanto

Figura 18 – Relação entre os descritores selecionados por diferentes métodos (tempo de retenção - prt, mapa auto organizável do tipo Kohonen com as estruturas 2D - som2D e estruturas 3D - som3D) e tipo (2D-descr, 3D-1conf e 3D-weigh). Descritores, métodos de separação (teste e treino) e tipos de descritores são representados por segmentos do círculo. A cada variável foi atribuída uma cor. Cada descritor recebeu uma cor correspondente à sua classificação e forma de seleção e sua largura indica a quantidade que foi selecionada.

Fonte: Oliveira, T.B. (2015).

relaciona-se às informações sobre ligações e átomos (complexidade da molécula). Todos esses quatro descritores foram selecionados pelos três métodos de separação entre teste e treino (prt, som2D e som3D) e estão presentes nos melhores modelos.

Outro descritor que aparece, porém em menor quantidade, mas está presente entre os diferentes métodos de seleção e entre os melhores modelos, (Tabela 15, página 105) é o AATS4s (média ponderada da autocorrelação de Broto-Moreau na distância 4 cujo peso é o l-state). Descritores menos selecionados como BCUT SlogP 0 (*Burden-CAS-University of Texas eigenvalues*, calculados a partir dos valores próprios de uma matriz de adjacência modificado usando contribuição atômica para logP e usando o método de SlogP de Wildman e Crippen (36, 37)) e Uindex (Balaban U index (36, 37)) trouxeram a contribuição de ampliar o domínio químico dos modelos (Figura 18, página 99).

Quatro descritores também se destacaram dentre os 3D, uma vez que foram selecionados independente de sua origem (3D-1conf e 3D-weight) e método de seleção. Estes descritores foram: TDB1u (autocorrelação baseada distância topológica 3D - intervalo 1 / não ponderada), TDB2i (autocorrelação baseada distância topológica 3D - intervalo 2 / ponderado pelo primeiro potencial de ionização), vsurf CW2 (fator de capacidade das estruturas e é calculado no nível de energia de $-0,5 \text{ kcal/mol}$ (98, 99)) e vsurf D7 (região hidrofóbica da estrutura e é definida quando uma sonda seca está interagindo com um alvo da estrutura a $1,4 \text{ kcal/mol}$ (98, 99)). O descritor Mor11e (décimo primeiro sinal do código de representação molecular da estrutura conforme a difração de elétrons/ ponderado pela eletronegatividade atômica de Sanderson (108, 109, 110)) foi selecionado em todos os métodos de seleção para os descritores 3D-weight e o ASAP2 (fração da área superficial acessível por átomos com carga parcial no intervalo de 0,05 a 0,10 e (100)) para 3D-1conf. Os descritores dipole (momento dipolar) e dipole AM1 (momento dipolar calculado usando o AM1 Halmiltoniano) contribuíram para ampliar o domínio químico dos modelos que utilizaram os descritores 3D-weight e o descritor vsurf Wp5 (volume polar da estrutura e é calculado ao nível energético de $-3,0 \text{ kcal/mol}$ (98, 99)) para os modelos que utilizaram os descritores 3D-1conf (Figura 18, página 99).

A Figura 19 (página 101) indica a frequência relativa dos vários fatores investigados (% , eixo y) em função do qualidade preditiva dos modelos em uma amostragem comum (teste externo P^2 , eixo x). Dentre os fatores utilizados, o melhor desempenho foi obtido com BRNN ($P^2 = 0,86$). Considerando os três conjuntos de descritores (2D-descr, 3D-1conf e 3D-weight), o uso de descritores 2D gerou os melhores modelos. Adicionar mais descritores aos modelos (iniciando com 4 e finalizando com 6 descritores) também trouxe uma melhoria dos resultados que podem ser observados na região com P^2 acima de 0,81.

Os melhores modelos foram ordenados pelo melhor P^2 (coeficiente de correlação do teste externo), uma vez que essa avaliação é a mais indicada para representar como um modelo se comporta ao estimar o RTT de estruturas que não fizeram parte da construção

Figura 19 – Gráficos cumulativos de participação dos fatores (em porcentagem) entre os melhores modelos. O eixo horizontal mostra o coeficiente de determinação no teste externo (P^2); o eixo vertical mostra a participação de cada fator. As maiores áreas (barras verticais coloridas) demonstram as abordagens mais bem-sucedidas.

Fonte: Oliveira, T.B. (2015).

do modelo. Sequencialmente, modelos consensuais foram construídos iniciando com os modelos com seis descritores (ANA12, ANA18, ANA06, ANA42, ANA48, ANA54, ANA36, ANA24, ANA30), para esses modelos, devido às diferentes formas para seleção de teste e treino, a avaliação estatística foi apenas com as estruturas presentes no teste externo. Após busca exaustiva de combinação, obteve-se o modelo CON1 (ANA12, ANA18 e ANA42) com o melhor P^2 (0,85) e contendo descritores 3D com o objetivo de manter possíveis diferenças entre enantiômeros. Outros modelos contendo o mesmo método de seleção para teste e treino também foram construídos: CON2 (ANA12, ANA30 e ANA48), CON3 (ANA16, ANA18, ANA36 e ANA54) e CON4 (ANA04, ANA06, ANA24 e ANA42). Na Tabela 15 (página 105) e na Figura 21 (página 104) são apresentados um resumo dos melhores modelos para cada variável e os modelos consensuais.

Após a seleção dos melhores modelos e construção dos modelos consensuais foi necessário avaliar alguns parâmetros, como a separação de isóbaros. Mas, primeiramente, deve-se compreender que para estruturas com a mesma massa molecular e variação de RTT inferior ao MAE e a RMSE não se deve esperar que os modelos estimem a sequência

Figura 20 – Exemplos de estruturas presentes no banco de dados da *Analyticon Discovery*. Dentre as estruturas se encontram epímeros e estruturas com mesma massa molecular (isóbaros).

Fonte: Oliveira, T.B. (2015).

correta entre elas, porém, se faz necessário verificar esses critérios para o uso real de tais modelos.

O primeiro parâmetro foi avaliar se os modelos foram capazes de separar epímeros e isômeros geométricos. Para esse fim, foram usados os pares de estruturas selecionadas aleatoriamente na Figura 20 (página 102). Os pares de epímeros **109 e 510/ 2454 e 2488** possuem uma variação do RTT menor que o desvio padrão médio (13,44) das substâncias analisadas mais de uma vez. Portanto, não se espera que os modelos separem essas substâncias. Assim, todos os modelos que utilizaram os 2D-descr se enquadram neste critério, assim como o modelo CON1. O par de epímeros **1196 e 1197** possui uma pequena diferença entre seus RTT (Δ RTT = 60) e nenhum dos modelos foi capaz de estimar a sequência correta entre as duas estruturas. Para os pares **23 e 2434**, **27 e 42**, **43 e 303**, **47 e 1723**, **74 e 75**, **109 e 510**, **89 e 144**, **1374 e 1377**, apenas o modelo ANA54 foi capaz de estimar a correta sequência. Desta forma, acredita-se que o modelo ANA54 deva ser mais promissor para a separação desses pares.

O segundo parâmetro foi avaliar se os modelos são capazes de separar o tempo das estruturas com a mesma massa molecular. Esse fator é importante no momento da desreplicação de uma matriz vegetal (Figura 20, página 102).

Os derivados do ácido cinâmico **27**, **28**, **42**, **87** e a substância fenólica **445** foram classificadas no grupo i. O modelo ANA48 foi capaz de estimar a sequência correta das substâncias. Os derivados do ácido cinâmico foram estimados na sequência correta pelo modelo ANA54, mas a substância **445** não foi estimada na sequência correta. Por fim, se for

avaliado apenas o aspecto estatístico, como a menor variação do RTT real e RTT estimado, e excluindo o enantiômero **42** o melhor modelo é o CON4.

As lignanas **47** e **1723**, a isoflavona **1064** e o derivado do ácido cinâmico **1334** foram classificados no grupo **g**. O modelo ANA54 estimou a sequência correta para as substâncias **47**, **1723** e **1334**, mas não para a substância **1064**.

As flavonas **109**, **338**, **510** e **1335** foram classificadas no grupo **i**. Os modelos ANA04, ANA06, ANA12, ANA16, ANA18 e CON3 estimaram na sequência correta as estruturas **109**, **510** e **1335**, mas não a estrutura **338**.

O sesquiterpeno **1160**, classificado no grupo **k**, os fenilpropanois **329**, **2454** e **2488**, classificados no grupo **g**, foram estimados na sequência correta pelo modelo ANA24.

Mesmo fazendo parte do *applicability domain* as substâncias **2513**, **2586** e **2587** (classificados no grupo **k**) e **2515** (classificado no grupo **f**) tiveram valores atípicos (*outliers*) estimados. Este fato pode ter ocorrido devido a estas substâncias apresentarem apenas um valor de RTT (apenas 1 eluição).

Após as avaliações dos epímeros, isômeros geométricos, isóbaros e valores atípicos pode-se observar que não foi possível gerar um único modelo geral capaz de prever todas as possíveis situações de um espaço químico, mesmo sendo ele composto apenas de substâncias oriundas de produtos naturais. Todavia, o desenvolvimento deste estudo auxilia na redução de substâncias no processo de identificação e desreplicação em estudos metabolômicos.

Figura 21 – Representação linear dos valores experimentais *versus* calculados dos RTTs dos modelos ANA12, ANA42, ANA36, CON1 e CON4. Os melhores modelos para cada tipo de descritor e modelos consensuais com P^2 igual a 0,85 e as 2.635 estruturas combinadas dos grupos de treino, validação externa e teste externo.

Fonte: Oliveira, T.B. (2015).

Tabela 15 – Resumo dos modelos de QSRR obtidos para todo o ambiente estrutural. R²: coeficiente de determinação; Q²: coeficiente de determinação para a validação externa; P²: coeficiente de determinação para teste externo; MAE: média do erro absoluto; RMSE: raiz quadrada do erro quadrático médio (Tabela completa no Apêndice B, página 121).

Modelos	Descritores	Tipos de descritores	Modelagem	Treino	R ²	MAE	RMSE	Q ²	MAE	RMSE	P ²	MAE	RMSE
ANA12	AATS4s, BCUT SLOGP 0, BIC0, ETA dEpsilon D, ML-FER S, logS	2D-descr	BRNN	som2D	0,86	177,28	230,85	0,82	185,03	268,93	0,84	183,39	247,38
ANA18	AATS4s, BIC0, ETA dEpsilon D, MLFER S, Uindex, logS	2D-descr	BRNN	som3D	0,86	171,58	226,57	0,81	196,05	279,08	0,84	182,50	248,13
ANA04	AATS4s, BIC0, ETA dEpsilon D, MLFER S, logS	2D-descr	BRNN	prt	0,84	183,94	247,59	0,84	194,41	259,10	0,83	192,06	256,21
ANA06	BCUT SLOGP 0, BIC0, ETA dEpsilon D, GATS6p, ML-FER S, logS	2D-descr	BRNN	prt	0,85	184,85	244,70	0,84	193,68	260,12	0,81	192,41	264,80
ANA42	Mor11e, TDB1u, TDB2i, dipole, vsurf CW2, vsurf D7	3D-weight	BRNN	prt	0,82	193,00	262,80	0,85	191,97	251,98	0,78	209,01	286,75
ANA48	Mor11e, TDB1u, TDB2i, vsurf CW2, vsurf D7, dipole	3D-weight	BRNN	som2D	0,84	191,18	251,49	0,78	200,82	296,20	0,78	207,41	288,76
ANA54	Mor11e, TDB1u, TDB2i, vsurf CW2, vsurf D7, AM1 dipole	3D-weight	BRNN	som3D	0,84	187,30	247,95	0,78	209,94	299,26	0,77	212,68	294,37
ANA36	ASAP2, TDB1u, TDB2i, vsurf CW2, vsurf D7, vsurf Wp5	3D-1conf	BRNN	som3D	0,82	203,29	265,79	0,76	224,26	313,27	0,77	218,28	294,46

Continua na próxima página

Tabela 15 – Continuação da página anterior

Modelos	Descritores	Tipos de descritores		Modelagem	Treino	R ²	MAE	RMSE	Q ²	MAE	RMSE	P ²	MAE	RMSE
CON1	-	-	-	-	-	-	-	-	-	-	-	0,85	168,33	235,32
CON2	-	-	-	-	-	0,87	169,85	224,47	0,82	183,31	266,09	0,83	177,52	251,85
CON3	-	-	-	-	-	0,88	163,85	217,04	0,83	181,50	266,31	0,84	172,69	243,29
CON4	-	-	-	-	-	0,86	172,27	230,77	0,87	175,81	229,61	0,85	175,61	241,78

5 CONCLUSÕES

Este estudo reúne características distintas pelo fato de combinar as áreas de farmacognosia e química de produtos naturais com a quimioinformática envolvendo o desenvolvimento de dados, modelos computacionais e aplicações práticas.

O banco de dados AsterDB foi desenvolvido e está sendo continuamente povoado por todos os membros do grupo. Atualmente possui arquivos `.sdf` e `.mol` e existem cerca de 2000 estruturas com informações associadas oriundas de Asteraceae, além de ser o único que agrega informações botânicas, químicas e biológicas com uma interface *web* em construção em conjunto com o laboratório de quimioinformática da UFPB.

Neste trabalho foi possível construir bons modelos ($P^2 = 0,94$) que são capazes de estimar o $\log k$ de estruturas complexas de terpenoides presentes no AsterDB, no caso 39 LST e reportando pela primeira vez o uso de modelos consensuais, um conjunto conformacional ponderados pela distribuição de Boltzmann e descritores Q_frASAs utilizados em estudos de QSRR.

Também observou-se que para algumas substâncias naturais da biblioteca não houve diferença estatística entre as substâncias analisadas individualmente e em mistura ($p\text{-valor} > 0,95$) e que a correlação entre os dois métodos cromatográficos e equipamentos utilizados foi reprodutível ($R > 0,95$). Estas análises mostraram que foi possível desenvolver modelos de QSRR para um método cromatográfico e equipamento e transpô-los para outro equipamento seguindo o uso de substâncias em comum.

Para as estruturas químicas e dados cedidos pela empresa *AnalytiCon Discovery* um estudo de QSRR foi capaz de produzir modelos com boa capacidade de predição ($P^2 > 0,81$) utilizando alta amplitude de espaço químico e rigor estatístico. Apenas pelo fato de utilizar diferentes classes químicas este estudo pode ser considerado inédito, uma vez que os estudos de QSRR geralmente empregam estruturas congênicas

Conclui-se que, estas informações podem ser utilizadas como uma plataforma piloto para análises de dados com objetivo de auxiliar na desreplicação de extratos de plantas em estudos metabolômicos em pelo menos duas situações concretas: (i) para restringir uma pesquisa em um banco de dados e acelerar a identificação de substâncias desconhecidas em extratos de plantas, ao comparar as massas moleculares, UV e seus parâmetros de tempo de retenção obtidos por análises de CL/DAD/EM e (ii) para localizar substâncias conhecidas e desejáveis em extratos de plantas após obter-se o parâmetro de tempo de retenção estimado de suas respectivas estruturas químicas e pesquisas em bancos de dados combinado com suas massas moleculares. Mesmo assim, ainda se faz necessário a supervisão humana e avaliação de dados, incluindo dados de massas, UV e tempo de

retenção.

REFERÊNCIAS

- 1 OLIVEIRA, T. B.; GOBBO-NETO, L.; SCHMIDT, T. J.; DA COSTA, F. B. Study of chromatographic retention of natural terpenoids by chemoinformatic tools. *J. Chem. Inf. Model.*, v. 55, n. 1, p. 26–38, 2015.
- 2 THEODORIDIS, G. A.; GIKA, H. G.; WANT, E. J.; WILSON, I. D. Liquid chromatography-mass spectrometry based global metabolite profiling: a review. *Anal. Chim. Acta*, v. 711, p. 07–16, 2012.
- 3 GARCÍA-PÉREZ, I.; VALLEJO, M.; GARCÍA, A.; LEGIDO-QUIGLEY, C.; BARBAS, C. Metabolic fingerprinting with capillary electrophoresis. *J. Chromatogr. A*, v. 1204, n. 2, p. 130–139, 2008.
- 4 KELL, D. B.; BROWN, M.; DAVEY, H. M.; DUNN, W. B.; SPASIC, I.; OLIVER, S. G. Metabolic footprinting and systems biology: the medium is the message. *Nat. Rev. Microbiol.*, v. 3, n. 7, p. 557–565, 2005.
- 5 KIND, T.; WOHLGEMUTH, G.; LEE, D. Y.; LU, Y.; PALAZOGLU, M.; SHAHBAZ, S.; FIEHN, O. FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal. Chem.*, v. 81, n. 24, p. 10038–10048, 2009.
- 6 MONTANARI, C. A.; BOLZANI, V. S. Planejamento racional de fármacos baseado em produtos naturais. *Quim. Nova*, v. 24, n. 1, p. 105–111, 2001.
- 7 CARNEVALE NETO, F.; SIQUITELLI, C. D.; PILON, A. C.; SILVA, D. H. S.; BOLZANI, V. S.; CASTRO-GAMBOA, I. Dereplication of phenolic derivatives of *Qualea grandiflora* and *Qualea cordata* (Vochysiaceae) using liquid chromatography coupled with ESI-QToF-MS/MS. *J. Braz. Chem. Soc.*, v. 24, n. 5, p. 758–764, 2013.
- 8 DA SILVA, F. M. A.; KOOLEN, H. H. F.; DE ALMEIDA, R. A.; DE SOUZA, A. D. L.; PINHEIRO, M. L. B. Desreplicação de alcalóides aporfinicos e oxoaporfinicos de *Unonopsis guatteroides* por ESI-IT-MS. *Quim. Nova*, v. 35, n. 5, p. 944–947, 2012.
- 9 BUTLER, M. S.; FONTAINE, F.; COOPER, M. A. Natural product libraries: assembly, maintenance, and screening. *Planta Med.*, v. 80, n. 14, p. 1161–1170, 2014.
- 10 LAM, K. S. New aspects of natural products in drug discovery. *Trends Microbiol.*, v. 15, n. 6, p. 279–289, 2007.
- 11 WELSCH, M. E.; SNYDER, S. A.; STOCKWELL, B. R. Privileged scaffolds for library design and drug discovery. *Curr. Opin. Chem. Biol.*, Elsevier, v. 14, n. 3, p. 347–361, 2010.
- 12 UNITED STATES DEPARTMENT OF COMERCE. *National telecommunications and information administration*. 1996. Disponível em: <http://www.its.bldrdoc.gov/fs-1037/dir-010/_1404.htm>.
- 13 O'DONNELL, T. J. *Design and use of relational databases in chemistry*. 1. ed. London: CRC, 2008.

- 14 NINGTHOUJAM, S. S.; TALUKDAR, A. D.; POTSANGBAM, K. S.; CHOUDHURY, M. D. Challenges in developing medicinal plant databases for sharing ethnopharmacological knowledge. *J. Ethnopharmacol.*, Elsevier, v. 141, n. 1, p. 9–32, 2012.
- 15 LOUB, W. D.; FARNSWORTH, N. R.; SOEJARTO, D. D.; QUINN, M. L. NAPRALERT: computer handling of natural product research data. *J. Chem. Inf. Comput. Sci.*, v. 25, p. 99–103, 1985.
- 16 CHEN, C. Y. TCM Database@Taiwan: the world's largest traditional chinese medicine database for drug screening *in silico*. *PLoS One*, v. 6, n. 1, p. e15939, 2011.
- 17 NAEEM, S.; HYLANDS, P.; BARLOW, D. Construction of an Indonesian herbal constituents database and its use in random forest modelling in a search for inhibitors of aldose reductase. *Bioorg. Med. Chem.*, Elsevier, v. 20, n. 3, p. 1251–1258, 2012.
- 18 MANHA, E. M.; SILVA, M. C.; ALVES, M. G. C.; ALMEIDA, M. B.; BRANDAO, M. G. L. PLANT: A bibliographic database about medicinal plants. *Braz. J. Pharmacogn.*, v. 18, n. 4, p. 614–617, 2008.
- 19 VALLI, M.; DOS SANTOS, R. N.; FIGUEIRA, L. D.; NAKAJIMA, C. H.; CASTRO-GAMBOA, I.; ANDRICOPULO, A. D.; BOLZANI, V. S. Development of a natural products database from the biodiversity of Brazil. *J. Nat. Prod.*, v. 76, n. 3, p. 439–444, 2013.
- 20 VOIGT, J. H.; BIENFAIT, B.; WANG, S.; NICKLAUS, M. C. Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.*, v. 41, n. 3, p. 702–712, 2001.
- 21 QIAO, X.; HOU, T.; ZHANG, W.; GUO, S.; XU, X. A 3D structure database of components from chinese traditional medicinal herbs. *J. Chem. Inf. Comput. Sci.*, v. 42, n. 3, p. 481–489, 2002.
- 22 KUMARI, S.; PUNDHIR, S.; PRIYA, P.; JEENA, G.; PUNETHA, A.; CHAWLA, K.; FIRDOS JAFAREE, Z.; MONDAL, S.; YADAV, G. EssOilDB: a database of essential oils reflecting terpene composition and variability in the plant kingdom. *Database (Oxford)*, v. 2014, p. 1–12, 2014.
- 23 ROTHWELL, J. A.; URPI-SARDA, M.; NEZ, M. B.; KNOX, C.; LLORACH, R.; EISNER, R.; CRUZ, J.; NEVEU, V.; WISHART, D.; MANACH, C.; ANDRES-LACUEVA, C.; SCALBERT, A. Phenol-Explorer 2.0: a major update of the phenol-explorer database integrating data on polyphenol metabolism and pharmacokinetics in humans and experimental animals. *Database (Oxford)*, v. 2012, p. bas031, 2012.
- 24 SAWADA, Y.; NAKABAYASHI, R.; YAMADA, Y.; SUZUKI, M.; SATO, M.; SAKATA, A.; AKIYAMA, K.; SAKURAI, T.; MATSUDA, F.; AOKI, T.; HIRAI, M. Y.; SAITO, K. RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database. *Phytochemistry*, Elsevier, v. 82, p. 38–45, 2012.
- 25 GASTEIGER, J.; FUNATSU, K. Chemoinformatics – an important scientific discipline. *J. Comput. Chem., Japan*, v. 5, n. 2, p. 53–58, 2006.
- 26 SOUSA, J. A. Químio-informática: conteúdos que urge ensinar. *Bol. Soc. Port. de Quím.*, v. 84, p. 55–59, 2002.

- 27 IUPAC. *Compendium of chemical terminology (the gold book)*. 2. ed. Research Triangle Park, NC: IUPAC, 2009.
- 28 XU, J.; HAGLER, A. Chemoinformatics and drug discovery. *Molecules*, v. 7, n. 8, p. 566–600, 2002.
- 29 GASTEIGER, J. Chemoinformatics: a new field with a long tradition. *Anal. Bioanal. Chem.*, v. 384, n. 1, p. 57–64, 2006.
- 30 WEININGER, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, v. 28, n. 1, p. 31–36, 1988.
- 31 WEININGER, D.; WEININGER, A.; WEININGER, J. L. SMILES. 2. algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.*, v. 29, n. 2, p. 97–101, 1989.
- 32 IUPAC. *The IUPAC international chemical identifier (InChI)*. 2015. Disponível em: <www.iupac.org/inchi>.
- 33 LEIGH, J. InChIs and registry numbers. *Chem. Int.*, v. 34, n. 6, p. 23, 2012.
- 34 INCHI-TRUST. *Technical FAQ*. 2015. Disponível em: <<http://www.inchi-trust.org/technical-faq/#2.7>>.
- 35 GASTEIGER, J.; ENGEL, T. *Chemoinformatics: a textbook*. Weinheim: WILEY-VCH Verlag, 2003. 649 p.
- 36 TODESCHINI, R.; CONSONNI, V. *Handbook of molecular descriptors, methods and principles in medicinal chemistry, vol. 11*. Weinheim: WILEY-VCH Verlag, 2000. 667 p.
- 37 TODESCHINI, R.; CONSONNI, V. *Molecular Descriptors for Chemoinformatics*. 2. ed. Weinheim: Wiley-VCH Verlag, 2009. v. 41. 967 p. (Methods and Principles in Medicinal Chemistry, v. 41).
- 38 ROLLINGER, J. M.; HAUPT, S.; STUPPNER, H.; LANGER, T. Combining ethnopharmacology and virtual screening for lead structure discovery: COX-inhibitors as application example. *J. Chem. Inf. Comput. Sci.*, v. 44, n. 2, p. 480–488, 2004.
- 39 HONG-FANG, J.; XUE-JUAN, L.; HONG-YU, Z. Natural products and drug discovery. Can thousands of years of ancient medical knowledge lead us to new and powerful drug combinations in the fight against cancer and dementia? *EMBO Rep.*, v. 10, p. 194–200, 2009.
- 40 KONG, D. X.; LI, X. J.; TANG, G. Y.; ZHANG, H. Y. How many traditional chinese medicine components have been recognized by modern western medicine? A chemoinformatic analysis and implications for finding multicomponent drugs. *ChemMedChem*, v. 3, p. 233–236, 2008.
- 41 LACHANCE, H.; WETZEL, S.; KUMAR, K.; WALDMANN, H. Charting, navigating, and populating natural product chemical space for drug discovery. v. 55, p. 5989–6001, 2012.
- 42 SINGH, N.; GUHA, R.; GIULIANOTTI, M. A.; PINILLA, C.; HOUGHTEN, R. A.; MEDINA-FRANCO, J. L. Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J. Chem. Inf. Model.*, v. 49, p. 1010–1024, 2009.

- 43 YONGYE, A. B.; WADDELL, J.; MEDINA-FRANCO, J. L. Molecular scaffold analysis of natural products databases in the public domain. *Chem. Biol. Drug. Des.*, v. 80, p. 717–724, 2012.
- 44 JIN, G.; XUE, X.; ZHANG, F.; ZHANG, X.; XU, Q.; JIN, Y.; LIANG, X. Prediction of retention times and peak shape parameters of unknown compounds in traditional chinese medicine under gradient conditions by ultra performance liquid chromatography. *Anal. Chim. Acta*, v. 628, n. 1, p. 95–103, 2008.
- 45 LIAO, L. M.; ZHU, J.; LI, J. F.; LEI, G. D. QSRR study on the components of *Styrax Japonicus* Sieb flowers using improved molecular electronegativity-distance vector (I-MEDV). *Chinese J. Struct. Chem.*, v. 30, p. 105–110, 2011.
- 46 NOORIZADEH, H. Investigation of retention behaviors of essential oils by using QSRR. *J. Chinese Chem. Soc.*, v. 57, n. 5A, p. 982–991, 2010.
- 47 PUT, R.; HEYDEN, Y. V. Review on modelling aspects in reversed-phase liquid chromatographic quantitative structure-retention relationships. *Anal. Chim. Acta*, v. 602, n. 2, p. 164–172, 2007.
- 48 NIKITAS, P.; PAPPA-LOUISI, A.; TSOUMACHIDES, S.; JOUYBAN, A. A principal component analysis approach for developing retention models in liquid chromatography. *J. Chromatogr. A*, Elsevier, v. 1251, p. 134–140, 2012.
- 49 KARIMI, H.; NOORIZADEH, H.; FARMANY, A. A QSRR modeling of hazardous psychoactive designer drugs using GA-PLS and L-M ANN. *ISRN Chromatogr.*, v. 2012, p. 1–9, 2012.
- 50 NOORIZADEH, H.; NOORIZADEH, M. QSRR-based estimation of the retention time of opiate and sedative drugs by comprehensive two-dimensional gas chromatography. *Med. Chem. Res.*, v. 21, n. 8, p. 1997–2005, 2011.
- 51 HÉBERGER, K. Quantitative structure-(chromatographic) retention relationships. *J. Chromatogr. A*, v. 1158, n. 1-2, p. 273–305, 2007.
- 52 KIND, T.; FIEHN, O. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal. Rev.*, v. 2, n. 1-4, p. 23–60, 2010.
- 53 KUMARI, S.; STEVENS, D.; KIND, T.; DENKERT, C.; FIEHN, O. Applying *in-silico* retention index and mass spectra matching for identification of unknown metabolites in accurate mass GC-TOF mass spectrometry. *Anal. Chem.*, v. 83, n. 15, p. 5895–902, 2011.
- 54 WEST, C.; GUENEGOU, G.; ZHANG, Y.; MORIN-ALLORY, L. Insights into chiral recognition mechanisms in supercritical fluid chromatography. II. Factors contributing to enantiomer separation on tris-(3,5-dimethylphenylcarbamate) of amylose and cellulose stationary phases. *J. Chromatogr. A*, Elsevier, v. 1218, n. 15, p. 2033–2057, 2011.
- 55 UKIĆ, S.; NOVAK, M.; ZUVELA, P.; AVDALOVIĆ, N.; LIU, Y.; BUSZEWSKI, B.; BOLANCA, T. Development of gradient retention model in ion chromatography. part I: conventional QSRR approach. *Chromatogr.*, v. 77, p. 985–996, 2014.
- 56 UKIĆ, S.; NOVAK, M.; ZUVELA, P.; AVDALOVIĆ, N.; LIU, Y.; BUSZEWSKI, B.; BOLANCA, T. Development of gradient retention model in ion chromatography. part II: conventional QSRR approach. *Chromatogr.*, v. 77, p. 997–1007, 2014.

- 57 AKBAR, J.; IQBAL, S.; BATOOL, F.; KARIM, A.; CHAN, K. W. Predicting retention times of naturally occurring phenolic compounds in reversed-phase liquid chromatography: a quantitative structure-retention relationship (QSRR) Approach. *Int. J. Mol. Sci.*, v. 13, n. 11, p. 15387–15400, 2012.
- 58 EUGSTER, P. J.; BOCCARD, J.; DEBRUS, B.; BRÉANT, L.; WOLFENDER, J. L.; MARTEL, S.; CARRUPT, P. A. Retention time prediction for dereplication of natural products (C_xH_yO_z) in LC–MS metabolite profiling. *Phytochemistry*, Elsevier, 2014.
- 59 GHAVAMI, R.; RASOULI, Z. Investigation of retention behavior of anthraquinoids in RP-HPLC on 17 different C18 stationary phases by means of quantitative structure retention relationships. *Med. Chem. Res.*, v. 22, n. 6, p. 2677–2691, 2012.
- 60 GORYŃSKI, K.; BOJKO, B.; NOWACZYK, A.; BUCIŃSKI, A.; PAWLISZYN, J.; KALISZAN, R. Quantitative structure-retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: endogenous metabolites and banned compounds. *Anal. Chim. Acta*, v. 797, p. 13–19, 2013.
- 61 JI, C.; LI, Y.; SU, L.; ZHANG, X.; CHEN, X. Quantitative structure-retention relationships for mycotoxins and fungal metabolites in LC-MS/MS. *J. Sep. Sci.*, v. 32, p. 3967–3979, 2009.
- 62 OLIVEIRA, T. B.; DA COSTA, F. B.; SCHMIDT, T. J. A quantitative structure-retention relationship (QSRR) model to aid plant metabolomic studies. In: DIAS, A.; SILVA, P. (Ed.). *Book of Abstracts: 62nd international congress and annual meeting of the society of medicinal plant and natural product research*. Guimarães: Planta Med., 2014. p. P1M10.
- 63 STEFOVA, M.; STAFILOV, T.; KULEVANOVA, S.; STEFKOV, G.; BANKOVA, V. S. QSRR of flavones: evaluation of substituent contributions to RP HPLC Retention. *J. Liq. Chromatogr. Relat. Technol.*, v. 30, n. January 2015, p. 1035–1049, 2007.
- 64 YANG, X.; WANG, J.; LUO, J.; KONG, L. One-step large-scale preparative isolation of isoquinoline alkaloids from *Rhizoma coptidis chinensis* by polyamide column chromatography and their quantitative structure-retention relationship analysis. *J. Liq. Chromatogr. Relat. Technol.*, v. 35, n. 13, p. 1842–1852, 2012.
- 65 DA COSTA, F. B.; SCHORR, K.; ARAKAWA, N. S.; SCHILLING, E. E.; SPRING, O. Intraspecific variation in the chemistry of glandular trichomes of two brazilian *Viguiera* species (Heliantheae; Asteraceae). *J. Braz. Chem. Soc.*, v. 12, n. 3, p. 403–407, 2001.
- 66 SCHORR, K.; GARCIA-PINERES, A. J.; SIEDLE, B.; MERFORT, I.; DA COSTA, F. B. Guaianolides from *Viguiera gardneri* inhibit the transcription factor NF-kappaB. *Phytochemistry*, v. 60, n. 7, p. 733–740, 2002.
- 67 SPRING, O.; ZIPPER, R.; CONRAD, J.; VOGLER, B.; KLAIBER, I.; DA COSTA, F. B. Sesquiterpene lactones from glandular trichomes of *Viguiera radula* (Heliantheae; Asteraceae). *Phytochemistry*, v. 62, n. 8, p. 1185–1189, 2003.
- 68 STEFANI, R.; EBERLIN, M. N.; TOMAZELA, D. M.; DA COSTA, F. B. Eudesmanolides from *Dimerostemma vestitum*. *J. Nat. Prod.*, v. 66, n. 3, p. 401–403, 2003.
- 69 SCHORR, K.; MERFORT, I.; DA COSTA, F. B. A novel dimeric melampolide and further terpenoids from *Smallanthus sonchifolius* (Asteraceae) and the inhibition of the transcription factor NF-kB. *Nat. Prod. Commun.*, v. 2, n. 4, p. 367–374, 2007.

- 70 AMBRÓSIO, S. R.; OKI, Y.; HELENO, V. C. G.; CHAVES, J. S.; NASCIMENTO, P. G. B. D.; LICHSTON, J. E.; CONSTANTINO, M. G.; VARANDA, E. M.; DA COSTA, F. B. Constituents of glandular trichomes of *Tithonia diversifolia*: relationships to herbivory and antifeedant activity. *Phytochemistry*, v. 69, n. 10, p. 2052–2060, 2008.
- 71 CHAVES, J. S.; DA COSTA, F. B. A proposal for the quality control of *Tanacetum parthenium* (feverfew) and its hydroalcoholic extract. *Rev. Bras. Farmacogn.*, v. 18, n. 3, p. 360–366, 2008.
- 72 CHEMAXON LTD. *Marvin*. Budapest, Hungary: ChemAxon, 2012. Disponível em: <<http://www.chemaxon.com/>>.
- 73 HALL, M.; FRANK, E.; HOLMES, G.; BERNHARD, P.; REUTEMANN, P.; WITTEN, I. H.; PFAHRINGER, B. The WEKA data mining software: an update. *ACM SIGKDD*, v. 11, n. 1, p. 10–18, 2009.
- 74 R CORE TEAM. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2013. Disponível em: <<http://www.r-project.org/>>.
- 75 CHEMICAL COMPUTING GROUP INC. *MOE: Molecular Operating Environment*. Montreal, Quebec, Canada: Chemical Computing Group, 2010. Disponível em: <<http://www.chemcomp.com/index.htm>>.
- 76 MOLECULAR NETWORKS GMBH. *CORINA (COoRdINAtes)*. Erlangen, Germany: Molecular Networks, 2011. Disponível em: <<http://www.molecular-networks.com>>.
- 77 MOLECULAR NETWORKS GMBH. *Adriana.Code - calculation of molecular descriptors*. Erlangen, Germany: Molecular Networks, 2011. Disponível em: <<http://www.molecular-networks.com/>>.
- 78 TALETE SRL. *Dragon*. Milano, Italy: Talete, 2007. Disponível em: <<http://www.talete.mi.it/index.htm>>.
- 79 YAP, C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.*, v. 32, n. 7, p. 1466–1474, 2011.
- 80 CHEN, I. J.; FOLOPPE, N. Conformational sampling of druglike molecules with MOE and catalyst: implications for pharmacophore modeling and virtual screening. *J. Chem. Inf. Model.*, England, v. 48, p. 1773–1791, 2008.
- 81 WIDOM, B. *The Boltzmann distribution law and statistical thermodynamics*. 1. ed. Cambridge: Cambridge University, 2002. 182 p.
- 82 YAN, J.; CAO, D. S.; GUO, F. Q.; ZHANG, L. X.; HE, M.; HUANG, J. H.; XU, Q. S.; LIANG, Y. Z. Comparison of quantitative structure-retention relationship models on four stationary phases with different polarity for a diverse set of flavor compounds. *J. Chromatogr. A*, Elsevier, v. 1223, p. 118–125, 2012.
- 83 KUHN, M. Building predictive models in R using the caret package. *J. Stat. Softw.*, v. 28, p. 1–26, 2008.
- 84 HALL, M. A.; SMITH, L. A. Feature subset selection: a correlation based filter approach. *Prog. Connec.-Based Inf. Syst.*, Nova Zelândia, p. 855–858, 1998.

- 85 HELLAND, I. S. On the structure of partial least squares regression. *Commun. Stat.-Simul. C.*, v. 17, n. 2, p. 581–607, 1988.
- 86 WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: practical machine learning tools and techniques*. 3. ed. Amsterdam: Elsevier, 2008. 629 p.
- 87 KLEKOTA, J.; ROTH, F. P. Chemical substructures that enrich for biological activity. *Bioinformatics*, v. 24, n. 21, p. 2518–2525, 2008.
- 88 GHAZZALI, N. NbClust : an R package for determining the relevant number of clusters in a data set. *J. Stat. Softw.*, v. 61, n. 6, 2014.
- 89 MACKAY, D. J. C. Bayesian interpolation. *Neural Comput.*, v. 4, n. 3, p. 415–447, 1992.
- 90 FORESEE, F. D.; HAGAN, M. T. Gauss-Newton approximation to bayesian regularization. In: *Proceedings of the 1997 international joint conference on neural networks*. EUA: IEEE Xplore, 1997. p. 1930–1935.
- 91 NGUYEN, D.; WIDROW, B. Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In: *International joint conference on neural networks*. EUA: IEEE Xplore, 1990. p. 21–26.
- 92 VAZQUEZ, A. I.; BATES, D. M.; ROSA, G. J. M.; GIANOLA, D.; WEIGEL, K. A. Technical note: an R package for fitting generalized linear mixed models in animal breeding. *J. Anim. Sci.*, v. 88, n. 2, p. 497–504, 2010.
- 93 BOSCH, E.; BOU, P.; ALLEMANN, H.; ROSÉS, M. Retention of ionizable compounds on HPLC. pH scale in methanol-water and the pK and pH values of buffers. *Anal. Chem.*, v. 68, n. 20, p. 3651–3657, 1996.
- 94 SNYDER, L. R.; KIRKLAND, J. J.; DOLAN, J. W. *Introduction to modern liquid chromatography*. 3. ed. New York: WILEY-VCH Verlag, 2009.
- 95 BASAK, S. C.; HARRISS, D. K.; MAGNUSON, V. R. Comparative study of lipophilicity versus topological molecular descriptors in biological correlations. *J. Pharm. Sci.*, v. 73, n. 4, p. 429–437, 1984.
- 96 ROY, K.; GHOSH, G. Introduction of extended topochemical atom (ETA) indices in the valence electron Mobile (VEM) environment as tools for QSAR/QSPR studies. *Internet Electron. J. Mol. Des.*, v. 2, n. 9, p. 599–620, 2003.
- 97 ROY, K.; KABIR, H. QSPR with extended topochemical atom (ETA) indices: Modeling of critical micelle concentration of non-ionic surfactants. *Chem. Eng. Sci.*, Elsevier, v. 73, p. 86–98, 2012.
- 98 MOORTHY, N. S. H. N.; RAMOS, M. J.; FERNANDES, P. A. Analysis of van der Waals surface area properties for human ether-a-go-go-related gene blocking activity: computational study on structurally diverse compounds. *SAR QSAR Environ. Res.*, v. 23, n. 5-6, p. 521–536, 2012.
- 99 CRUCIANI, G.; CRIVORI, P.; CARRUPT, P. A.; TESTA, B. Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. *J. Mol. Struct.: THEOCHEM*, v. 503, n. 1-2, p. 17–30, 2000.

- 100 SCHMIDT, T. J.; HEILMANN, J. Quantitative structure-cytotoxicity relationships of sesquiterpene lactones derived from partial charge (q)-based fractional accessible surface area descriptors (Q_frASAs). *Quant. Struct.-Act. Relat.*, v. 21, n. 3, p. 276–287, 2002.
- 101 ZHIVKOVA, Z.; DOYTCHINOVA, I. Quantitative structure–plasma protein binding relationships of acidic drugs. *J. Pharm. Sci.*, v. 101, n. 12, p. 4627–4641, 2012.
- 102 GRAMATICA, P.; GIANI, E.; PAPA, E. Statistical external validation and consensus modeling: a QSPR case study for Koc prediction. *J. Mol. Graphics Modell.*, v. 25, n. 6, p. 755–766, 2007.
- 103 ZHU, H.; TROPSHA, A.; FOURCHES, D.; VARNEK, A.; PAPA, E.; GRAMATICA, P.; OBERG, T.; DAO, P.; CHERKASOV, A.; TETKO, I. V. Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.*, v. 48, n. 4, p. 766–784, 2008.
- 104 ZHU, H.; MARTIN, T. M.; YE, L.; SEDYKH, A.; YOUNG, D. M.; TROPSHA, A. Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. *Chem. Res. Toxicol.*, v. 22, n. 12, p. 1913–1921, 2009.
- 105 HÉBERGER, K.; SKRBIĆ, B. Ranking and similarity for quantitative structure-retention relationship models in predicting Lee retention indices of polycyclic aromatic hydrocarbons. *Anal. Chim. Acta.*, v. 716, p. 92–100, 2012.
- 106 HOU, T. J.; XIA, K.; ZHANG, W.; XU, X. J. ADME evaluation in drug discovery. 4. prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.*, v. 44, n. 1, p. 266–275, 2004.
- 107 PLATTS, J. A.; BUTINA, D.; ABRAHAM, M. H.; HERSEY, A. Estimation of molecular linear free energy relation descriptors using a group contribution approach. p. 835–845, 1999.
- 108 GASTEIGER, J.; SADOWSKI, J. Chemical information in 3D space. *J. Chem. Inf. Comput. Sci.*, v. 36, n. 96, p. 1030–1037, 1996.
- 109 SCHUUR, J. H.; SELZER, P.; GASTEIGER, J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.*, v. 36, n. 2, p. 334–344, 1996.
- 110 SCHUUR, J.; GASTEIGER, J. Infrared spectra simulation of substituted benzene derivatives on the basis of a 3D structure representation. *Anal. Chem.*, v. 69, n. 13, p. 2398–2405, 1997.

Apêndices

APÊNDICE A – CRIAÇÃO DAS TABELAS DO ASTERDB

```
CREATE TABLE classification (  
  lvl0 varchar(100) UNIQUE, -- Last low level classification  
  lvl1 varchar(100), -- first high level classification  
  lvl2 varchar(100),  
  lvl3 varchar(100),  
  lvl4 varchar(100),  
  lvl5 varchar(100)  
);
```

```
CREATE TABLE structure (  
  casrn varchar(12) PRIMARY KEY, -- CAS Registry Number  
  iupac varchar(1000), -- Name IUPAC  
  usualn varchar(1000), -- Usual name  
  sysn varchar(1000), -- Systematic name  
  inchikey varchar(29), --InChIKey  
  inchi varchar(1000), --InChI  
  smiles varchar(1000), -- Smiles format  
  reference text NOT NULL, --reference  
  class varchar(1000) REFERENCES classification(lvl0) --Last level classification  
);
```

```
CREATE TABLE descriptors (  
  casrn varchar(12) REFERENCES structure(casrn),  
  mwmono numeric, -- molecular weigh monoisotopic  
  nhdonn integer, -- number of donor atoms H bonds  
  nhacc integer, -- number of acceptor atoms H bonds  
  alogp numeric, -- ghose-crippen octanol water partition coeff  
  mlogp numeric, -- moriguchi octanol water partition coeff  
  logs numeric, -- logS  
  logd numeric -- logD  
);
```

```
CREATE TABLE analytical (  
  casrn varchar(12) REFERENCES structure(casrn),
```

```
trmeoh numeric, -- isocratic method jcim article
trmecn numeric, -- isocratic method jcim article
tr1hb numeric, -- gradient method paula-chagas brazil
tr1hg numeric, -- gradient method paula-chagas germany
traster numeric -- AsterBioChem final method
);
```

```
CREATE TABLE activity (
casrn varchar(12) REFERENCES structure(casrn),
enzyme text, -- (COX-1, COX-2, 5-LOX, etc.)
organism text, -- (mus musculus, rattus novergicus, etc.)
bioactivity\_type varchar(20), -- IC50, Inhibition, etc.
method text, -- assay description or kit
value numeric, --assay results
unit varchar(10), -- unit assay results (% , mg/mL, etc)
reference text NOT NULL
);
```

```
CREATE TABLE taxomomic (
specie text PRIMARY KEY, -- genera and specie
genera text, -- only genera
family text,
tribe text,
subtribe text,
reference text NOT NULL
);
```

```
CREATE TABLE botanic (
specie text REFERENCES taxomomic(specie), -- genera and specie
country varchar(80),
city varchar(80),
voucher text,
gps point
);
```

```
CREATE TABLE partplanttable (
partplant text UNIQUE
);
```

```
CREATE TABLE boco (  
  specie text REFERENCES taxonomic(specie), -- genera and specie  
  casrn varchar(12) REFERENCES structure(casrn),  
  partplant text REFERENCES partplanttable (partplant),  
  gps point,  
  reference text NOT NULL  
);
```

APÊNDICE B – MELHORES MODELOS DE QSRR OBTIDOS PARA AS ESTRUTURAS QUÍMICAS CEDIDAS PELA EMPRESA ANALYTICON DISCOVERY

Tabela 16 – Melhores modelos de QSRR obtidos para todo o ambiente estrutural. R²: coeficiente de determinação; Q²: coeficiente de determinação para a validação externa; P²: coeficiente de determinação para teste externo; MAE: média do erro absoluto; RMSE: raiz quadrada do erro quadrático médio.

Modelos	Descritores	Tipo de descritores	Modelagem	Treino	R ²	MAE	RMSE	Q ²	MAE	RMSE	P ²	MAE	RMSE
ANA01	ETA BetaP s, ETA dEpsilon D, SHBint4, logS	2D-descr	MLR	prt	0,79	213,69	286,73	0,82	214,39	277,24	0,78	216,88	290,65
ANA02	ETA BetaP s, ETA dEpsilon D, SHBint4, logS	2D-descr	BRNN	prt	0,82	195,27	261,81	0,84	196,89	255,03	0,81	198,88	265,68
ANA03	AATS4s, BIC0, ETA dEpsilon D, MLFER S, logS	2D-descr	MLR	prt	0,80	207,04	274,40	0,84	196,32	258,05	0,79	214,80	281,90
ANA04	AATS4s, BIC0, ETA dEpsilon D, MLFER S, logS	2D-descr	BRNN	prt	0,84	183,94	247,59	0,84	194,41	259,10	0,83	192,06	256,21
ANA05	BCUT SLOGP 0, BIC0, ETA dEpsilon D, GATS6p, MLFER S, logS	2D-descr	MLR	prt	0,81	204,70	267,83	0,83	203,11	265,32	0,79	211,97	283,03
ANA06	BCUT SLOGP 0, BIC0, ETA dEpsilon D, GATS6p, MLFER S, logS	2D-descr	BRNN	prt	0,85	184,85	244,70	0,84	193,68	260,12	0,81	192,41	264,80
ANA07	ETA BetaP s, ETA dEpsilon D, SHBint4, logS	2D-descr	MLR	som2D	0,80	213,85	279,62	0,78	211,21	298,07	0,78	217,08	291,45

Continua na próxima página

Tabela 16 – Continuação da página anterior

Modelos	Descritores	Tipo de descritores	Modelagem	Treino	R ²	MAE	RMSE	Q ²	MAE	RMSE	P ²	MAE	RMSE
ANA08	ETA BetaP s, ETA dEpsilon D, SHBint4, logS	2D-descr	BRNN	som2D	0,83	197,35	255,53	0,81	191,98	273,95	0,82	197,49	264,84
ANA09	BCUT SLOGP 0, BIC0, ETA dEpsilon D, MLFER S, logS	2D-descr	MLR	som2D	0,82	204,93	265,22	0,78	208,46	292,88	0,79	210,45	284,58
ANA10	BCUT SLOGP 0, BIC0, ETA dEpsilon D, MLFER S, logS	2D-descr	BRNN	som2D	0,85	187,33	243,23	0,80	196,62	280,07	0,83	192,86	257,14
ANA11	AATS4s, BCUT SLOGP 0, BIC0, ETA dEpsilon D, MLFER S, logS	2D-descr	MLR	som2D	0,83	197,15	256,37	0,79	197,13	284,85	0,80	203,45	275,57
ANA12	AATS4s, BCUT SLOGP 0, BIC0, ETA dEpsilon D, MLFER S, logS	2D-descr	BRNN	som2D	0,86	177,28	230,85	0,82	185,03	268,93	0,84	183,39	247,38
ANA13	ETA BetaP s, ETA dEpsilon D, SHBint4, logS	2D-descr	MLR	som3D	0,81	208,98	271,41	0,75	226,71	320,32	0,78	216,75	291,28
ANA14	ETA BetaP s, ETA dEpsilon D, SHBint4, logS	2D-descr	BRNN	som3D	0,84	192,05	249,34	0,79	209,15	291,39	0,82	196,14	264,24
ANA15	ETA BetaP s, ETA dEpsilon D, SHBint4, logS, nCconj	2D-descr	MLR	som3D	0,82	203,64	264,85	0,75	227,27	318,45	0,78	212,02	288,18
ANA16	ETA BetaP s, ETA dEpsilon D, SHBint4, logS, nCconj	2D-descr	BRNN	som3D	0,85	183,45	240,58	0,80	204,10	288,48	0,82	192,87	259,99
ANA17	AATS4s, BIC0, ETA dEpsilon D, MLFER S, Uindex, logS	2D-descr	MLR	som3D	0,83	194,10	254,38	0,79	208,89	292,76	0,80	207,76	277,64

Continua na próxima página

Tabela 16 – Continuação da página anterior

Modelos	Descritores	Tipo de descritores	Modelagem	Treino	R ²	MAE	RMSE	Q ²	MAE	RMSE	P ²	MAE	RMSE
ANA18	AATS4s, BIC0, ETA dEpsi- lon D, MLFER S, Uindex, logS	2D-descr	BRNN	som3D	0,86	171,58	226,57	0,81	196,05	279,08	0,84	182,50	248,13
ANA19	TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-1conf	MLR	prt	0,72	251,23	331,44	0,74	248,30	330,10	0,72	249,38	327,76
ANA20	TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-1conf	BRNN	prt	0,75	231,43	307,16	0,77	229,54	307,09	0,73	242,42	319,59
ANA21	ASAP2, TDB1p, TDB2i, vsurf CW2, vsurf D7	3D-1conf	MLR	prt	0,73	245,68	321,81	0,76	239,73	315,85	0,72	248,11	329,04
ANA22	ASAP2, TDB1p, TDB2i, vsurf CW2, vsurf D7	3D-1conf	BRNN	prt	0,77	225,70	297,66	0,78	224,81	307,23	0,73	242,85	324,08
ANA23	ASAP2, RDF10s, TDB1u, TDB2i, vsurf CW1, vsurf D7	3D-1conf	MLR	prt	0,75	237,59	308,18	0,78	233,16	305,48	0,73	242,58	322,46
ANA24	ASAP2, RDF10s, TDB1u, TDB2i, vsurf CW1, vsurf D7	3D-1conf	BRNN	prt	0,80	210,69	275,05	0,81	212,81	278,95	0,77	219,09	297,23
ANA25	TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-1conf	MLR	som2D	0,73	250,60	327,38	0,70	253,26	341,21	0,72	249,98	328,35
ANA26	TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-1conf	BRNN	som2D	0,76	231,46	303,82	0,73	238,87	327,04	0,74	242,07	318,33
ANA27	ASAP2, TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-1conf	MLR	som2D	0,76	238,82	309,96	0,72	245,19	332,09	0,73	242,16	321,48
ANA28	ASAP2, TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-1conf	BRNN	som2D	0,80	211,18	278,92	0,73	229,66	326,75	0,75	231,58	310,85

Continua na próxima página

Tabela 16 – Continuação da página anterior

Modelos	Descritores	Tipo de descritores	Modelagem	Treino	R ²	MAE	RMSE	Q ²	MAE	RMSE	P ²	MAE	RMSE
ANA29	ASAP2, TDB1u, TDB2i, vsurf CW2, vsurf D7, vsurf Wp5	3D-1conf	MLR	som2D	0,77	231,66	300,68	0,73	237,62	324,31	0,74	236,47	313,03
ANA30	ASAP2, TDB1u, TDB2i, vsurf CW2, vsurf D7, vsurf Wp5	3D-1conf	BRNN	som2D	0,82	202,20	264,95	0,75	226,57	315,92	0,77	220,23	299,59
ANA31	TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-1conf	MLR	som3D	0,74	246,45	317,60	0,67	264,26	368,07	0,72	249,67	328,06
ANA32	TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-1conf	BRNN	som3D	0,77	228,47	296,95	0,71	247,70	345,67	0,74	240,33	317,07
ANA33	J3D, TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-1conf	MLR	som3D	0,77	230,80	300,75	0,69	254,30	358,05	0,74	237,72	313,35
ANA34	J3D, TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-1conf	BRNN	som3D	0,79	218,69	284,12	0,73	236,56	331,37	0,76	228,43	300,48
ANA35	ASAP2, TDB1u, TDB2i, vsurf CW2, vsurf D7, vsurf Wp5	3D-1conf	MLR	som3D	0,78	227,49	293,20	0,71	249,71	344,42	0,75	235,08	311,81
ANA36	ASAP2, TDB1u, TDB2i, vsurf CW2, vsurf D7, vsurf Wp5	3D-1conf	BRNN	som3D	0,82	203,29	265,79	0,76	224,26	313,27	0,77	218,28	294,46
ANA37	TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-weight	MLR	prt	0,74	239,53	317,46	0,77	235,21	308,39	0,74	237,66	317,00
ANA38	TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-weight	BRNN	prt	0,76	225,30	300,85	0,80	223,56	287,19	0,75	231,47	308,09

Continua na próxima página

Tabela 16 – Continuação da página anterior

Modelos	Descritores	Tipo de descritores	Modelagem	Treino	R ²	MAE	RMSE	Q ²	MAE	RMSE	P ²	MAE	RMSE
ANA39	Mor11e, TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-weight	MLR	prt	0,76	226,75	301,41	0,81	216,63	284,04	0,76	227,80	303,93
ANA40	Mor11e, TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-weight	BRNN	prt	0,80	204,53	276,37	0,84	191,84	257,75	0,77	218,88	295,52
ANA41	Mor11e, TDB1u, TDB2i, dipole, vsurf CW2, vsurf D7	3D-weight	MLR	prt	0,78	217,91	291,64	0,82	213,54	274,04	0,77	217,05	299,10
ANA42	Mor11e, TDB1u, TDB2i, dipole, vsurf CW2, vsurf D7	3D-weight	BRNN	prt	0,82	193,00	262,80	0,85	191,97	251,98	0,78	209,01	286,75
ANA43	Mor11e, TDB2i, vsurf CW2, vsurf D7	3D-weight	MLR	som2D	0,75	237,73	311,51	0,73	230,64	326,95	0,73	242,21	319,41
ANA44	Mor11e, TDB2i, vsurf CW2, vsurf D7	3D-weight	BRNN	som2D	0,79	219,96	289,05	0,77	211,91	302,06	0,75	233,66	310,40
ANA45	Mor11e, TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-weight	MLR	som2D	0,78	227,23	295,40	0,77	216,42	301,61	0,76	228,33	304,83
ANA46	Mor11e, TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-weight	BRNN	som2D	0,82	199,01	263,58	0,78	203,61	291,84	0,78	215,27	293,00
ANA47	Mor11e, TDB1u, TDB2i, vsurf CW2, vsurf D7, dipole	3D-weight	MLR	som2D	0,80	218,13	282,64	0,77	214,67	301,61	0,76	217,34	300,97
ANA48	Mor11e, TDB1u, TDB2i, vsurf CW2, vsurf D7, dipole	3D-weight	BRNN	som2D	0,84	191,18	251,49	0,78	200,82	296,20	0,78	207,41	288,76
ANA49	Mor11e, TDB2i, vsurf CW2, vsurf D7	3D-weight	MLR	som3D	0,77	227,80	299,30	0,69	260,37	360,01	0,73	241,97	319,11

Continua na próxima página

Tabela 16 – Continuação da página anterior

Modelos	Descritores	Tipo de descritores	Modelagem	Treino	R ²	MAE	RMSE	Q ²	MAE	RMSE	P ²	MAE	RMSE
ANA50	Mor11e, TDB2i, vsurf CW2, vsurf D7	3D-weight	BRNN	som3D	0,80	212,50	280,63	0,73	240,33	336,07	0,75	234,67	310,31
ANA51	Mor11e, TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-weight	MLR	som3D	0,79	219,51	284,91	0,73	238,35	331,21	0,76	227,30	304,08
ANA52	Mor11e, TDB1u, TDB2i, vsurf CW2, vsurf D7	3D-weight	BRNN	som3D	0,83	196,65	258,92	0,77	215,84	306,05	0,77	217,95	295,18
ANA53	Mor11e, TDB1u, TDB2i, vsurf CW2, vsurf D7, AM1 dipole	3D-weight	MLR	som3D	0,80	214,35	278,64	0,74	233,72	325,09	0,76	220,62	301,06
ANA54	Mor11e, TDB1u, TDB2i, vsurf CW2, vsurf D7, AM1 dipole	3D-weight	BRNN	som3D	0,84	187,30	247,95	0,78	209,94	299,26	0,77	212,68	294,37

Anexos

ANEXO A – COPYRIGHT

Rightslink® by Copyright Clearance Center

03/02/15 13:09

RightsLink®

Home

Create Account

Help

ACS Publications
 Most Trusted. Most Cited. Most Read.
Title:

Study of Chromatographic Retention of Natural Terpenoids by Chemoinformatic Tools

Author:

Tiago B. Oliveira, Leonardo Gobbo-Neto, Thomas J. Schmidt, et al

Publication:

Journal of Chemical Information and Modeling

Publisher:

American Chemical Society

Date:

Jan 1, 2015

Copyright © 2015, American Chemical Society

LOGIN

If you're a **copyright.com** user, you can login to RightsLink using your copyright.com credentials. Already a **RightsLink** user or want to [learn more?](#)

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

BACK

CLOSE WINDOW

Copyright © 2015 [Copyright Clearance Center, Inc.](#) All Rights Reserved. [Privacy statement.](#)
 Comments? We would like to hear from you. E-mail us at customer@copyright.com