UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE
DEPARTAMENTO DE FARMÁCIA

REGINALDO MATHEUS GOIS MASCARENHAS

AVALIAÇÃO DE TOXICIDADE GERAL EM CONSTITUINTES ALIMENTARES UTILIZANDO FERRAMENTAS *IN SILICO*

SÃO CRISTÓVÃO – SE
2019

REGINALDO MATHEUS GOIS MASCARENHAS

AVALIAÇÃO DE TOXICIDADE GERAL EM CONSTITUINTES ALIMENTARES UTILIZANDO FERRAMENTAS *IN SILICO*

Trabalho de Conclusão de Curso apresentado ao Departamento de Farmácia da Universidade Federal de Sergipe – DFA/UFS, para obtenção do título de Bacharel em Farmácia. Orientado pelo Prof. Dr. Tiago Branquinho Oliveira.

São Cristóvão – SE
2019

REGINALDO MATHEUS GOIS MASCARENHAS

AVALIAÇÃO DE TOXICIDADE GERAL EM CONSTITUINTES ALIMENTARES UTILIZANDO FERRAMENTAS *IN SILICO*

Trabalho de Conclusão de Curso apresentado ao Departamento de Farmácia da Universidade Federal de Sergipe – DFA/UFS, para obtenção do título de Bacharel em Farmácia. Orientado pelo Prof. Dr. Tiago Branquinho Oliveira.

Aprovado em: _____de_____de_____.

BANCA EXAMINADORA:

_____

Prof. Dr. Tiago Branquinho Oliveira – DFA/UFS

(Orientador)

_____

Prof. Dr. Marcelo Cavalcante Duarte – DFA/UFS

(1º Examinador)

_____

Ma. Iraê Oliveira Moura – DFA/UFS

(2º Examinador)

Acima de tudo, agradeço a Deus por ter me dado força e saúde para que eu chegasse até esse momento. Dedico esta, bem como todas as minhas conquistas, aos meus pais Reginaldo e Maria Elena, a minha irmã Alice e aos meus dois sobrinhos. À todos aqueles que de alguma forma estiveram envolvidos com o meu trajeto acadêmico e pessoal.

SUMÁRIO

1. Introdução 8

1.1Quimioinformática **8**

1.2 Toxicidade Alimentar… **8**

1.3 *Quantitative Structure-Activity Relationship(QSAR)* **9**

1.4. Ferramentas e Software **9**

2. Objetivo 11

3. Materiais e métodos 11

3.1 Coleta de Dados **11**

3.2 Seleção de Estruturas Químicas **11**

3.3 Cálculo das propriedades teóricas(Descritores)… **11**

3.4 Desenvolvimento do Modelo **12**

4. Resultados e discussão..... 12

4.1 Acurácia **12**

4.2 Sensibilidade **13**

4.3 Especificidade **13**

4.4 Eficiência **13**

4.5 Coeficienteφ(phi) **13**

5. Conclusão 15

6. Referencias bibliográficas 15

7. Anexo 17

7.1 Anexo – Diretos Autorais **17**

Revista Virtual de Química

ISSN 1984-6835

Artigo

Avaliação de toxicidade geral em constituintes alimentares utilizando ferramentas *in silico*

Mascarenhas, R. M. G.; Lima, C. A.; Rodrigues, R. P.; Kitagawa, R. R.; Faraoni, A. S.; Oliveira, T. B.*

Rev. Virtual Quim., **2019**, *11* (2), 543-553. Data de publicação na Web: 24 de abril de 2019

http://rvq.sbq.org.br

Evaluation of general toxicity in food constituents using *in silico* tools

Abstract: Preliminary knowledge about the toxicity of new substances for food use may contribute to the rapid selection of useful and increasingly safe substances. For this purpose, a Quantitative Structure-Toxicity Relationship (QSTR) model was developed with 139,395 structures obtained in three different lists of toxic (US EPA DSSTox) and non-toxic (FEMA GRAS ™ and FDA GRAS) substances. The 2D coordinates were obtained, standardized and checked, and a total of 4,860 fingerprints fragments defined by Klekota and Roth were calculated for each substance and used as independent variables. The data were processed in order to remove highly correlated variables and fragments close to zero variance, reducing fragments to 166. Dependent variables consisted of a binary classification, where zero corresponds to non-toxic whereas 1 corresponds to toxic. The classification models were created with decision tree using the J48 algorithm and random tree. The models (training, cross-validation and external validation) were evaluated based on their predictive performance. The best selected model was the random tree to obtain the best values external validation (accuracy = 0.9658, sensitivity = 0.9798, specificity = 0.5495, efficiency = 0.7640 and phi coefficient = 0.4941). The developed of a QSTR model can be used to predict the toxicity of novel food additives, manufacturing technology adjuvants and nutraceuticals.

Keywords: Food additives; toxicity; chemoinformatics.

Resumo

O conhecimento preliminar sobre a toxicidade de novas substâncias para uso alimentar pode contribuir com a rápida seleção de substâncias úteis e cada vez mais seguras. Com esse objetivo, um modelo de QSTR (*Quantitative Structure-Toxicity Relationship*) foi desenvolvido com 139.395 estruturas obtidas a partir de três diferentes listas de substâncias tóxicas (US EPA DSSTox) e atóxicas (FEMA GRAS™ e FDA GRAS). As coordenadas 2D foram obtidas, padronizadas e checadas, resultando em um total de 4.860 fragmentos dos *fingerprints* definidos por Klekota e Roth, que foram calculados para cada substância, sendo utilizados como variáveis independentes. Os dados foram processados com o objetivo de eliminar as variáveis altamente correlacionadas e os fragmentos com variância próxima a zero, reduzindo o número de fragmentos a 166. As variáveis dependentes consistiram na classificação 0 (atóxicos)/1(tóxicos). Os modelos de classificação foram criados com árvore de decisão usando o algoritmo J48 e árvore aleatória. Já os modelos treino, validação cruzada e validação externa, foram avaliados com base no seu desempenho de previsão. O melhor modelo selecionado foi a árvore aleatória, por obter os melhores valores para validação externa (acurácia = 0,9658; sensibilidade = 0,9798; especificidade = 0,5495; eficiência = 0,7640 e coeficiente phi = 0,4941). O modelo de QSTR desenvolvido pode ser utilizado para prever a toxicidade de novos aditivos alimentares, coadjuvantes de tecnologia de fabricação e nutracêuticos.

Palavras-chave: Aditivos alimentares; toxicidade; quimioinformática.

* Universidade Federal de Sergipe, Campus Cidade Univ. Prof. José Aloísio de Campos, Departamento de Farmácia, Av. Marechal Rondon, s/n, Jd. Rosa Elze, CEP 49100-000, São Cristóvão-SE, Brasil.

tiago.branquinho@ufs.br

DOI: 10.21577/1984-6835.20190041

Volume 11, Número 2 Março-Abril 2019

Revista Virtual de Química
ISSN 1984-6835

Avaliação de toxicidade geral em constituintes alimentares utilizando ferramentas *in silico*

Reginaldo M. G. Mascarenhas,[a] Camila A. de Lima,[a] Ricardo P. Rodrigues,[b] Rodrigo R. Kitagawa,[b] Aurelia S. Faraoni,[a] Tiago B. Oliveira[a,*]

[a] Universidade Federal de Sergipe, Campus Cidade Univ. Prof. José Aloísio de Campos, Departamento de Farmácia, Av. Marechal Rondon, s/n, Jd. Rosa Elze, CEP 49100-000, São Cristóvão-SE, Brasil.

[b] Universidade Federal do Espírito Santo, Programa de Pós-Graduação em Ciências Farmacêuticas, Departamento de Ciências Farmacêuticas, Av. Marechal Campos 1468, CEP 29043-900, Vitória-ES, Brasil.

* tiago.branquinho@ufs.br

Recebido em 20 de abril de 2019. Aceito para publicação em 20 de abril de 2019

1. Introdução

1.1. Quimioinformática
1.2. Toxicidade alimentar
1.3. *Quantitative Structure-Activity Relationship* (QSAR)
1.4. Ferramentas e software

2. Objetivo

3. Materiais e Métodos

3.1. Coleta de dados
3.2. Seleção de estruturas químicas
3.3. Cálculo das propriedades teóricas (Descritores)
3.4. Desenvolvimento do modelo

4. Resultados e Discussão

4.1. Acurácia
4.2. Sensibilidade
4.3. Especificidade
4.4. Eficiência
4.5. Coeficiente ϕ (phi)

5. Conclusão

1. Introdução

1.1. Quimioinformática

Com o desenvolvimento de novas tecnologias na era computacional, houve um aumento da quantidade de informação armazenada e processada, com especial aplicação em química. Esta nova tendência abriu as portas para um campo de pesquisa da quimioinformática, assim denominado pela primeira vez por Frank Brown em 1998. Segundo[1], pode ser definido 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.[2]

Durante anos o campo da quimioinformática evoluiu bastante, deixando de ser uma ciência apenas de aspectos práticos e técnicas de representação, manipulação e processamento de estruturas químicas até adquirir notoriedade com seu papel primordial na atualidade: exploração de bases de dados químicas e descoberta de novos compostos com atividade e/ou propriedades desejadas.[3] Ao explorar bases de dados nos deparamos com uma infinidade de informações que podem auxiliar na compreensão do comportamento de determinado grupo de estruturas químicas, sendo assim possível construir modelos computacionais que possam ser utilizados para predizer a atividade de substâncias que apresentem poucos dados experimentais, ou seja, dados de estudos *in vitro* e *in vivo*.[4-5]

Uma das informações que podemos extrair utilizando ferramentas de quimioinformática é a classificação toxicológica das substâncias através do cálculo de alguns descritores (parâmetro adotado para calcular e comparar informações contidas nas substâncias em estudo, ex: massa molecular, coeficiente de partição octanol-água, hidrofobicidade, etc.). De forma clássica e mais usual, estes descritores podem ser classificados de 3 maneiras quanto às suas dimensões. Unidimensionais (1D), são descritores que se baseiam em propriedades físico-químicas e da fórmula molecular (ex., massa molecular, refratividade molar, logP, entre outros); bidimensionais (2D), que descrevem propriedades de uma representação de 2 dimensões (ex., número de átomos, número de ligações, índices de conectividade, entre outros); e tridimensionais (3D), que dependem da conformação 3D das moléculas (ex., volume de Van der Waals, área de superfície acessível ao solvente, entre outros).[6] Utiliza-se desses descritores para avaliação do potencial de toxicidade de substâncias, principalmente aquelas encontradas em alimentos, visto que o uso seguro de substâncias na alimentação é uma condição indispensável para o seu consumo. Adicionalmente, compreender as características que conferem a uma substância uma propriedade, seja ela uma propriedade nutracêutica, de flavor ou de toxicidade.

1.2. Toxicidade alimentar

Os alimentos podem ser definidos como misturas complexas de substâncias, em sua maioria água, lipídios, carboidratos, proteínas, vitaminas e ácidos nucleicos, que por serem absorvidas e possuírem utilidade na manutenção e do metabolismo humano são consideradas nutrientes[7].

Na composição dos alimentos também é possível encontrar substâncias sem valor nutricional, que podem, por exemplo, conferir características como sabor, cor e odor, ainda que não possuam funções no corpo humano. A preocupação surge com aquelas que podem fazer interações e de alguma forma se tornem prejudiciais ao organismo, classificadas como substâncias tóxicas.[7]

As substâncias tóxicas presentes nos alimentos podem ter diferentes fontes, podendo ser constitutivos do próprio alimento ou fruto de contaminação. Uma causa comum de contaminação são fungos. Estima-se que mais de 100 tipos de fungos

produtores de micotoxinas possam ser encontrados em cultivos agrícolas, podendo conter substâncias que causam pouca toxicidade ou até mesmo as que possam estar associadas com efeitos carcinogênicos, hepatotóxicos e mutagênicos, como as mais de 20 substâncias já descritas na literatura[8]. No Brasil, dentre estas toxinas, a mais conhecida é a aflatoxina do tipo B1, presente no feijão, amendoim, entre outros produtos alimentares. De fácil acesso na mesa dos brasileiros, essas toxinas estão principalmente relacionadas com câncer hepático.[9] Outro fator contaminante que gera preocupação é o agrotóxico. Em estudos realizados com trabalhadores e animais expostos cronicamente a agrotóxicos foram encontrados efeitos adversos, tais como alterações no sistema neurológico, reprodutor, imune, além de alterações metabólicas.[10]

A ocorrência de casos clínicos relacionados com intoxicação alimentar é bastante comum, podendo apresentar desde consequências leves até casos mais graves, levando à morte do indivíduo. Nos Estados Unidos, cerca de 6,5 milhões de casos estão relacionados a infecções e 9.000 a óbitos decorrentes de enfermidades transmitidas por alimentos a cada ano. No Brasil ocorrem cerca de 6.300 óbitos/ano, desencadeados por doenças transmitida por alimentos.[11]

1.3. *Quantitative structure-activity relationship* (QSAR)

A química medicinal é uma ciência que trabalha no desenvolvimento, otimização e busca de novos fármacos através de multidisciplinaridade de áreas como a química, farmácia, medicina, biologia e bioinformática, sendo possível o estudo de novas entidades químicas com propriedades terapêuticas. Baseada no conhecimento específico que estas áreas podem oferecer, é possível identificar substâncias com potencial de atividade para determinado alvo biológico apenas apoiando-se em mecanismos de ação já conhecidos de substâncias endógenas ou da própria estrutura do alvo investigado.[12-13]

Comumente, a química medicinal se utiliza de *Quantitative Structure-Activity Relationship* (QSAR) para estabelecer uma relação matemática entre propriedades físico-químicas e a atividade biológica de uma determinada substância. Desde o estudo de Hammett na década de 1930 esse campo de estudo vem se aprimorando, principalmente com o auxílio de recursos computacionais, que além de agilizarem os cálculos tornaram possíveis novas abordagens, onde não só parâmetros físico-químicos testados experimentalmente ou calculados são utilizados, mas também o estudo das estruturas moleculares em três dimensões levando em conta, por exemplo, efeitos estéricos e eletrostáticos.[12-14]

Além da ampla utilização desse estudo pela indústria farmacêutica visando à descoberta de novos fármacos, o QSAR também se mostra uma ferramenta útil para a identificação de substâncias tóxicas, o *Quantitative Structure-Toxicity Relationship* (QSTR), na criação de medicamentos mais seguros e/ou como alternativa para diminuir ou até mesmo substituir testes clínicos.[15]

Utilizando os conceitos abordados busca-se estabelecer uma relação entre características físico-químicas, atividade tóxica e modelos moleculares usando recursos computacionais como algoritmos, descritores moleculares e parâmetros estatísticos.

1.4. Ferramentas e software

Tendo em vista o crescimento tecnológico observado nas últimas décadas seguido de um exponencial aumento na capacidade de armazenamento de informação, se observou, em contrapartida, dificuldade no processamento desta informação, e diante do baixo aproveitamento do conhecimento obtido dessas informações, surgiu a

necessidade de criar ferramentas para agrupar e processar a quantidade de dados gerados.

Um exemplo de ferramenta é a mineração de dados ou *data mining*, que vem sendo bastante empregado para auxiliar e aperfeiçoar novas descobertas. O conceito de mineração de dados é basicamente a transformação de grandes quantidades de dados em padrões e regras significativos, consistentes e sistemáticas entre as variáveis, através do uso de algoritmos e descritores pré-selecionados para o estudo em questão, utilizando o conhecimento em estatística e com auxílio de softwares como, por exemplo, o *Waikato Environment for Knowledge Analysis* (Weka), podendo assim transformar uma grande quantidade de dados em resultados menos complexos, e direcionado ao problema proposto. Portanto, a mineração de dados tenta substituir muita desinformação (na forma de dados espalhados) em informações úteis.[16]

Sua prioridade é o uso dos modelos obtidos para prever um comportamento futuro, melhorar seu estudo ou apenas explicar coisas, que caso contrário, sem a mineração desses dados não seriam possíveis de se explicar. Estes modelos podem confirmar o que já pensávamos, ou ainda melhor, podem achar coisas novas em nossos dados que nem sabíamos que existiam. A mineração de dados é dividida em: definição do problema; pré-processamento dos dados; mineração (análise) dos dados; interpretação dos resultados.

Criado por pesquisadores da Universidade de Waikato (Nova Zelândia) o WEKA foi implementado pela primeira vez em 1997. O programa consiste em um pacote de software de livre acesso baseado em java, que surgiu através da necessidade de unificar os dados de trabalho e facilitar o estudo. Hoje em dia, o WEKA é mundialmente reconhecido como um sistema de exploração de dados e aprendizagem de máquinas.[17]

O WEKA suporta uma grande quantidade de dados em um formato de arquivo próprio, os algoritmos expressos pelo WEKA, calculam e organizam os descritores e o software é capaz de apresentar os resultados em forma de diagrama de árvore.

As principais vantagens de utilizar o método de árvore é por elas "tomarem decisões" levando em consideração aqueles atributos considerados mais relevantes.[18] Por esse motivo é considerado uma das melhores escolhas para o estudo em questão, por ser prático e fácil de interpretar os modelos mais comuns são obtidos em forma de diagrama de árvore. A definição mais empregada para o diagrama de árvore é a representação gráfica que consiste no apoio à tomada de decisão das alternativas geradas a partir de uma pergunta inicial. Uma das grandes vantagens de um modelo em árvores é a possibilidade de transformação/ decomposição de um problema complexo em diversos subproblemas mais simples. De uma forma recursiva, os novos subproblemas identificados voltam a ser decompostos em subproblemas ainda mais simples, como é o caso da árvore de decisão.[19]

O algoritmo de Árvore Aleatória ou *Random Tree* funciona exatamente como o Algoritmo J48 que é empregada na Árvore de Decisão, com uma exceção: para cada divisão, apenas um subconjunto aleatório de atributos é selecionado.

Tendo em vista a necessidade de conhecer a toxicidade das substâncias presente em alimentos, a utilização de ferramentas *in silico* torna-se imprescindível no processo, uma vez que o auxílio deste sistema vai otimizar o estudo sobre toxicidade. O presente artigo emprega o campo da quimioinformática aplicado à área de alimentos, denominado por alguns autores como "Foodinformatics".[20]

2. Objetivo

O objetivo deste trabalho foi empregar ferramentas computacionais, de quimioinformática e mineração de dados em dados oriundos de estudos sobre toxicidades para comparar substâncias consideradas tóxicas com substâncias consideradas não tóxicas, e assim determinar se o modelo criado tem a capacidade ou não de predizer por meios de cálculos estatísticos, sua toxidade.

3. Materiais e Métodos

3.1. Coleta de dados

A coleta de dados foi realizada em duas etapas:

- Realizou-se o levantamento de informações toxicológicas das substâncias encontradas no banco de dados da *United States Environmental Protection Agency* (EPA) (https://www.epa.gov), n = 183.925 e pré-definidas numericamente com valor 1 para sua atividade toxicológica.[21]

- Em contrapartida, houve o levantamento de informações de substâncias de uso seguro na alimentação através de dados da agência regulatória, *Food and Drug Administration* (FDA, http://www.fda.gov), n = 364 e da *Flavor and Extract Manufacturers Association* (FEMA, https://www.femaflavor.org/), n = 2.500 e pré-definidas com valor 0 para sua atividade toxicológica.[22-23]

3.2. Seleção de estruturas químicas

Para aquelas substâncias cuja informação sobre sua estequioquímica completa era ausente, optou-se por pesquisar em base de dados digitais como Pubchem (https://pubchem.ncbi.nlm.nih.gov/) e ChemSpider (http://www.chemspider.com/), assim pode se obter todas as informações estequiométricas das estruturas, através do número CAS, que é o número utilizado pelo *Chemical Abstracts Service* para confirmar que a estrutura única.[24-25]

Todas as estruturas foram validadas com o auxílio da base de dados SciFinder.[26] Como pré-tratamento das estruturas químicas provenientes dos bancos de dados, removeu-se os isótopos, sais e misturas, estruturas duplicadas, inválidas e ao final, a adição de hidrogênio e normalização das estruturas, utilizando o software *Standardizer* 16.2.29 e *Structure Checker* 16.2.29 – Chemaxon, Hungria.[27]

Essas substâncias foram classificadas como tendo o "uso seguro" também denominado *Generally Recognized as Safe* (GRAS), que é a designação da FDA para que um produto químico ou substância adicionada ao alimento seja considerado seguro, com essa informação podemos partir do princípio que a substância presente nos alimentos não apresentam atividades toxicológicas.[28]

As estruturas coletadas foram utilizadas para construir modelos como variáveis dependentes (y) e farão parte de um Banco de Dados do Laboratório de Alimentos e Bebidas da UFS para futuras pesquisas.

3.3. Cálculo das propriedades teóricas (descritores)

Uma redução do total dos descritores do tipo *Klekota-Roth fingerprint* (KRFP) previamente calculados com o programa PaDel[29] para a obtenção de um resultado mais consistente, devido ao grande número de substâncias coletadas, dividindo-se em dez subgrupos diferentes, estes subgrupos foram refinados através do software R, com o pacote Caret v.6.0 usando as funções *nearZeroVar* e *findCorrelation*. Na primeira função, todos os descritores que apresentaram variância igual ou bem próximo a zero foram eliminados. Na

segunda função, todos os descritores restantes foram avaliados quanto à alta correlação entre eles, esta função determina quando um par de descritores com alta correlação entre si, exclui-se um deles para evitar possíveis erros.[30]

3.4. Desenvolvimento do modelo

Com o auxílio do software WEKA dois modelos utilizando diferentes algoritmos (árvore de decisão e árvore aleatória) foram desenvolvidos com os KRFP restantes, após serem filtrados. E, com o intuito de avaliar o poder de predição e escolha do melhor modelo foi efetuada a validação cruzada 10-*folds* e validação externa com 66 % (n=139.395) das estruturas selecionadas aleatoriamente para treino e 33 % (n=47.394) para o teste externo.

Optou-se a utilização de parâmetros estatísticos: acurácia, sensibilidade, eficiência, especificidade e coeficiente ϕ (*phi*) e por classificação: verdadeiro positivo (VP), falso positivo (FP), verdadeiro negativo (VN) e falso negativo (FN).

Definiu-se que, VP seria usado para substâncias tóxicas, calculadas como tóxicas; FP, substâncias não tóxicas, calculadas como tóxicas; VN, substâncias não tóxicas, calculadas como não tóxicas; e, FN, substâncias tóxicas, calculadas como não tóxicas.

4. Resultados e Discussão

Na coleta dos dados, foram obtidas 205.959 substâncias concatenadas dos bancos de dados da FDA, FEMA e EPA, onde se excluíram 19.170 por não apresentarem informações suficientes quanto suas estruturas químicas e/ou constituintes majoritários.

Para a estruturação de um modelo prévio de predição da toxicidade calculou-se o KRFP (um conjunto de 4860 subestruturas químicas que podem auxiliar na interpretação de atividades biológicas) para as 186.789 estruturas químicas.

Dos 4860 KRFP obtidos inicialmente, foram filtrados usando as duas funções do R (*nearZeroVar* e f*indCorrelation*), resultando em 166 restantes.

Das 166 subestruturas de KRFP, com os parâmetros obtidos pelo modelo, tornou-se possível a comparação entre os dados estatísticos (acurácia, sensibilidade, eficiência, especificidade e coeficiente ϕ (phi)) e a classificação: verdadeiro positivo (VP), falso positivo (FP), verdadeiro negativo (VN) e falso negativo (FN).

O resultado obtido a partir do cálculo de validação externa do modelo da árvore de decisão apresentou uma avaliação correta de 97,08 % das substâncias (n=46.009) e 2,92 % de avaliação incorreta das substâncias (n=1.385) (Tabela 1). A validação externa para o modelo de árvore aleatória os resultados obtidos foram: 96,58 % de substâncias avaliadas corretamente (n=45.773) e 3,42 % de substâncias avaliadas incorretamente (n=1.621) (Tabela 2).

Estudos semelhantes por descobertas de toxicidades já foram propostos, como foi feito por Braga e colaboradores quando desenvolveram um servidor online baseado em QSAR para a predição de toxicidade cardíaca (bloqueio da hERG).[31] Em 2014, no desafio Tox21, vários modelos de previsão de toxicidade foram desenvolvidos, demonstrando que estes estudos são adequados para a implantação em pesquisas toxicológicas de ponta.[32]

4.1. Acurácia

O valor para a acurácia variou entre 0,9658 e 0,9823. Esse parâmetro nos mostra a proporção total de predições corretas, sem levar em consideração as substâncias tóxicas e não tóxicas. Esta medida é altamente suscetível e não deve ser levada como um fator determinante para o resultado final do conjunto de dados, pois pode facilmente

induzir a uma conclusão errada sobre o desempenho do sistema.[33]

Este parâmetro para o treino, validação cruzada e validação externa referente ao modelo de classificação utilizando a árvore de decisão foram 0,9756, 0,9711 e 0,9708, respectivamente, e para o modelo de classificação utilizando a árvore aleatória foram 0,9823, 0,9667 e 0,9658, respectivamente. Os valores de acurácia dos modelos não foi fator de diferenciação para selecionar o melhor método de classificação.

4.2. Sensibilidade

Os valores para a sensibilidade variaram entre 0,9798 e 0,9955. Esse parâmetro determina a proporção de substâncias que realmente são tóxicas (VP), ou seja, a capacidade do sistema em prever corretamente a toxicidade para substâncias que realmente são tóxicas.[34]

Este parâmetro para o treino, validação cruzada e validação externa referente ao modelo de classificação utilizando a árvore de decisão foram 0,9955, 0,9931 e 0,9936, respectivamente, e para o modelo de classificação utilizando a árvore aleatória foram 0,9955, 0,9798 e 0,9798, respectivamente. A sensibilidade dos modelos não foi fator de diferenciação para selecionar o melhor método de classificação, pois os valores não diferem significativamente.

4.3. Especificidade

A especificidade apresentou uma variação de valores entre: 0,2890 e 0,7615. Ao contrário da sensibilidade, a especificidade determina a proporção de substâncias que não são tóxicas (VN), ou seja, a capacidade do sistema em predizer corretamente a ausência da toxicidade para substâncias que realmente não são tóxicas.[35]

Este parâmetro para o treino, validação cruzada e validação externa referente ao modelo de classificação utilizando a árvore de decisão foram 0,3895, 0,3202 e 0,2893, respectivamente, e para o modelo de classificação utilizando a árvore aleatória foram 0,7615, 0,5798 e 0,5495, respectivamente. A diferença entre os valores de especificidade mostra que o modelo de árvore aleatória demonstrou ser mais específico.

4.4. Eficiência

Os valores para eficiência variaram entre 0,6415 e 0,8756. Esse parâmetro consiste na média aritmética da sensibilidade e especificidade. Isto é, geralmente, quando um método é muito sensível a positivos, tende a gerar muitos falsos positivos, e vice-versa. Assim, um método de decisão perfeito (100 % de sensibilidade e 100 % de especificidade) raramente é alcançado, e este método oferece um balanço entre ambos.[35]

Este parâmetro para o treino, validação cruzada e validação externa referente ao modelo de classificação utilizando a árvore de decisão foram 0,6925, 0,6566 e 0,6415, respectivamente, e para o modelo de classificação utilizando a árvore aleatória foram 0,8756, 0,7798 e 0,7640, respectivamente. A diferença entre os valores mostra que o modelo de árvore aleatória demonstrou ser mais eficiente.

4.5. Coeficiente de correlação de Matthews (coeficiente ϕ (*phi*))

Os valores para coeficiente de correlação de Matthews (coeficiente ϕ (*phi*), variaram entre: 0,4052 e 0,7287. Esse parâmetro é uma medida de qualidade de duas classificações binárias que pode ser usada com classes que possuem tamanhos bastante diferentes. Retorna um valor entre (-1) e (+1), em que um coeficiente de (+1) representa uma predição

perfeita, (0) representa uma predição aleatória média, e (-1) uma predição inversa. Esta medida estatística é equivalente ao coeficiente *phi*, e tenta, assim como a eficiência, resumir a qualidade da tabela de contingência em um único valor numérico passível de ser comparado.[33]

Este parâmetro para o treino, validação cruzada e validação externa referente ao modelo de classificação utilizando a árvore de decisão foram 0,5279, 0,4290 e 0,4052, respectivamente, e para o modelo de classificação utilizando a árvore aleatória foram 0,7287, 0,5172 e 0,4941, respectivamente. Os resultados para os valores de coeficiente ϕ (phi) dos modelos apresentados mostram que a árvore aleatória tem o melhor método de classificação.

Tabela 1. Matriz de confusão da árvore de decisão (*Decision Tree*) do treino e validação cruzada (n=139.395) e validação externa (n=47.394)

		VALORES REAIS					
Valores previstos		Tóxicos (treino)	Não tóxicos (treino)	Tóxicos (validação cruzada)	Não tóxicos (validação cruzada)	Tóxicos (validação externa)	Não tóxicos (validação externa)
	Tóxicos	134.228 [a]	2.782 [b]	133.906 [a]	3.098 [b]	45.564 [a]	1.093 [b]
	Não tóxicos	610 [c]	1.775 [d]	932 [c]	1.459 [d]	292 [c]	445 [d]

(a) substâncias tóxicas, calculadas como tóxicas (VP), (b) substâncias não tóxicas, calculadas como tóxicas (FP), (c) substâncias tóxicas, calculadas como não tóxicas (FN), substâncias não tóxicas, calculadas como não tóxicas (VN)

Tabela 2. Matriz de confusão da árvore aleatória (*Random Tree*) do treino e validação cruzada (n=139.395) e validação externa (n=47.394)

		VALORES REAIS					
Valores previstos		Tóxicos (treino)	Não tóxicos (treino)	Tóxicos (validação cruzada)	Não tóxicos (validação cruzada)	Tóxicos (validação externa)	Não tóxicos (validação externa)
	Tóxicos	133.454 [a]	1.087 [b]	132.114 [a]	1.915 [b]	44.928 [a]	693 [b]
	Não tóxicos	1.384 [c]	3.470 [d]	2.724 [c]	2.642 [d]	928 [c]	845 [d]

(a) substâncias tóxicas, calculadas como tóxicas (VP), (b) substâncias não tóxicas, calculadas como tóxicas (FP), (c) substâncias tóxicas, calculadas como não tóxicas (FN), substâncias não tóxicas, calculadas como não tóxicas (VN)

5. Conclusão

Através da utilização de ferramentas *in silico* pode-se desenvolver um modelo de fácil interpretação, acessível e sem custo financeiro. Os resultados obtidos por este estudo de toxicidade em alimentos apresentaram previsões precisas e sensíveis, o que se torna algo importante quando se trata de previsão de qualquer que seja a toxicidade. Concluiu-se que a sensibilidade para o modelo de validação externa da árvore de decisão ficou expresso em 0,9936, enquanto a árvore aleatória apresentou 0,9800 para aquelas substâncias que verdadeiramente apresentavam toxicidade, e a especificidade para a árvore de decisão ficou com valor abaixo da árvore aleatória apresentando 0,2893, 0,5494 respectivamente, o que evidencia que o modelo possui falsos negativos, que podem ser explicados com outros estudos, tal como a dose x resposta em experimentos *in vivo*.

A eficiência para a árvore de decisão obtido pelo cálculo do modelo foi 0,6415 e os encontrados no modelo de árvore aleatória, 0,7646. Baseado nos dados apresentado pelos dois modelos utilizados, conclui-se que o modelo com o melhor desempenho para predizer toxicidade de uma substância foi o modelo desenvolvido com árvore aleatória, pois os parâmetros observados e comparado entre eles, demonstram que este, obteve um melhor desempenho em relação ao modelo de árvore de decisão.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), do Programa Institucional de Bolsas de Iniciação Científica da UFS (PIBIC-COPES/UFS) e da Fundação de Apoio à Pesquisa e à Inovação Tecnológica do Estado de Sergipe (FAPITEC/SE).

Referências Bibliográficas

[1] Sousa, J. A. Quimio-informatica: conteudos que urge ensinar. *Boletim da Sociedade Portuguesa de Química* **2002**, *84*, 55.

[2] Gasteiger, J. A.; Funatsu, K. Chemoinformatics – An Important Scientific Discipline. *Journal of Computer Chemistry* **2006**, *5*, 53. [Link]

[3] Alves, V.; Braga, R C.; Muratov, E. N.; C. H. Andrade. Quimioinformática: uma introdução. *Química Nova* **2017**, *41*, 202. [CrossRef]

[4] Gasteiger, J. Chemoinformatics: Achievements and Challenges, a Personal View. *Molecules* **2016**, *21*, *151. [CrossRef]*

[5] *Fourches, D. Em Application of Computational Techniques in pharmacy and Medicine*; Gorb, L.; Kuz'min, V.; Muratov, E. eds; Springer Netherlands: Dordrecht, 2014, cap. 16

[6] Xue, L; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Combinatorial chemistry & high throughput screening* **2000**, *3*, p. 363. [PubMed]

[7] ANVISA, Decreto-lei nº 986, de 21 de outubro de 1969. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto-lei/Del0986.htm> Acesso em: 14 Março 2019.

[8] As micotoxinas. *Revista Food Ingredients Brasil* **2009**, *7*, 32. [Link]

[9] Oliveira, C. A. F.; Germano, P. M. L. Aflatoxinas: conceitos sobre mecanismos de toxicidade e seu envolvimento na etiologia do câncer hepático celular. *Revista de Saúde Pública* **1997**, *31*, 417.[CrossRef]

[10] Cantarutti, T. F. P.; *Dissertação de Mestrado*, Universidade Federal do Paraná, Brasil, 2005.[Link]

[11] Almeida, C. F.; Araújo, E. S.; Soares, Y. C.; Diniz, R. L. C.; Fook, S. M. L.; Vieira, K. V. M. Perfil epidemiológico das intoxicações alimentares notificadas no Centro de Atendimento Toxicológico de Campina

Grande, Paraíba. *Revista Brasileira de Epidemiologia* **2008**, *11*, 191. [Link]

[12] Thomas, G.; *Quimica Medicinal: Uma Introdução*. Guanabara Koogan: Rio de Janeiro, 2003.

[13] Alves, V.; *Dissertação de Mestrado*, Universidade Federal de Goiás, Brasil. 2014. [Link]

[14] Montanari, C. A.; Pilli, R. A. *Journal of the Brazilian Chemical Society*, **2002**, *13* [CrossRef]

[15] Guido, R. V. C.; Andricopulo, A. D.; Oliva, G. Planejamento de fármacos, biotecnologia e química medicinal: aplicações em doenças infecciosas. *Estudos Avançados* **2010**, *24*, 81. [CrossRef]

[16] Oliveira, C.; da Silva, J. C.; *Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas*, Universidade Federal de Goiás, 2009. [Link]

[17] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* **2009**, *11*, 10. [Link]

[18] Quinlan, J R. Induction of Decision Trees. *Machine Learning* **1986**, *1*, 81. [CrossRef]

[19] Gama, J. Functional Trees. *Machine Learning* **2004**, *55*, 219. [CrossRef]

[20] Martinez-Mayorga, K.; Medina-Franco, J. L.; *Foodinformatics: Applications of Chemical Information to Food Chemistry*, Springer International: Nova Iorque, 2014.

[21] EPA. United States Environmental Protection Agency | US EPA. Disponível em: <https://www.epa.gov/>. Acesso em: 15 Março 2019.

[22] FDA. Administration, U.S. Food and Drug. Disponível em: <https://www.fda.gov/>. Acesso em: 16 Março 2019.

[23] FEMA. Association, Flavor & Extract Manufacturers. Disponível em: <https://www.femaflavor.org/>. Acesso em: 15 Março 2019.

[24] Pubchem. National Center for Biotechnology Information, U.S. National Library of Medicine. Disponível em: <https://pubchem.ncbi.nlm.nih.gov/>. Acesso em: 15 Março 2019.

[25] ChemSpider, Search and share chemistry. Disponível em: <http://www.chemspider.com/>. Acesso em: 20 Março 2019.

[26] Scifinder. American Chemical Society, 2017. Disponível em: <https://www.cas.org/>. Acesso em: 14 Março 2019.

[27] Csizmadia, F. Jchem: java applets and modules supporting chemical database handling from web browsers. *Journal of chemical information and computer sciences* **2000**, 323.[PubMed]

[28] Burdock, G. A.; Carabin, I. G. Generally recognized as safe (GRAS): history and description. *Toxicology Letters* **2004**, *150*, 3. [Link]

[29] Yap, C. W. PaDEL-descriptor: An open source software to calculatemolecular descriptors and fingerprints. *Journal of Computational Chemistry* **2010**, *32*, 1466. [CrossRef]

[30] Klekota, J.; Roth, F. P. Chemical substructures that enrich for biological activity. *Bioinformatics* **2008**, *24*, 2518. [CrossRef]

[31] Braga, R. C.; Alves, V. M.; Silva, M. F. B.; Muratov, E.; Fourches, D.; Lião, L. M.; Tropsha, A.; Andrade, C. H. Pred-hERG: A Novel web-Accessible Computational Tool for Predicting Cardiac Toxicity. *Molecular informatics* **2015**, *34*, 698. [CrossRef]

[32] Unterthiner, T.; Mayr, A.; Klambauer, G.; Hochreiter, S. Toxicity Prediction using Deep Learning. *arXiv* **2015**, 1503.01445. [Link]

[33] Lopes, B.; Ramos, I. C.; Ribeiro, G.; Correa, R.; Valbon, B.; Luz, A.; Salomão, M.; Lyra, J. M.; Ambrósio Junior, R. Biostatistics: fundamental concepts and practical applications. *Revista Brasileira de Oftalmologia* **2014**, *73*, 16. [CrossRef]

[34] Witten, I. H.; Eibe, F.; Hall, M. A.; *Data mining: pratical machine learning tools and techenique*, 3rd ed. USA, Burlington: Morgan Kaufmann,2011. [Link]

[35] Kawamura, T. Interpretação de um teste sob a visão epidemiológica: eficiência de um teste. *Arquivos Brasileiros de Cardiologia* **2002**, *79*, 437. [CrossRef]

ANEXO – DIREITOS AUTORAIS

Revista Virtual de Química

ISSN 1984-6835

Niterói, 3 de junho de 2019

De David Rodrigues da Rocha (editor-coordenador)
Para: **Reginaldo M. G. Mascarenhas**

Autorizo que o autor Reginaldo M. G. Mascarenhas utilize todas as informações contidas em seu manuscrito intitulado "Avaliação de toxicidade geral em constituintes alimentares utilizando ferramentas *in silico*" publicado na Revista Virtual de Química em seu volume 11, número 2, páginas 543-553 (DOI: 10.21577/1984-6835.20190041)

Cordialmente,

Prof. David Rodrigues da Rocha
Editor Coordenador da Revista Virtual de Química