UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI

Faculdade de Ciências Exatas/Departamento de Computação

Fábio Coelho Sampaio

METODOLOGIA DE SUPERFÍCIE DE RESPOSTA E REDE NEURAL ARTIFICIAL NA MODELAGEM EMPÍRICA DE BIOPROCESSOS: Uma revisão sobre diferentes análises comparativas

Fábio Coelho Sampaio

Diamantina

2018

Fábio Coelho Sampaio

METODOLOGIA DE SUPERFÍCIE DE RESPOSTA E REDE NEURAL ARTIFICIAL NA MODELAGEM EMPÍRICA DE BIOPROCESSOS: Uma revisão sobre diferentes análises comparativas

Trabalho de Conclusão de Curso apresentado ao Departamento de Sistemas de Informação da Universidade Federal dos Vales do Jequitinhonha e Mucuri, como requisitos exigidos para a obtenção do título de bacharel em Sistemas de Informação.

Orientador: Prof. Me. Eduardo Pelli

Diamantina

2018

Fábio Coelho Sampaio

METODOLOGIA DE SUPERFÍCIE DE RESPOSTA E REDE NEURAL ARTIFICIAL NA MODELAGEM EMPÍRICA DE BIOPROCESSOS: Uma revisão sobre diferentes análises comparativas

Trabalho de Conclusão de Curso apresentado ao Departamento de Sistemas de Informação da Universidade Federal dos Vales do Jequitinhonha e Mucuri, como requisitos exigidos para a obtenção do título de bacharel em Sistemas de Informação.

Orientador: Prof. Msc. Eduardo Pelli

Data da aprovação: 02/03/2013.

Prof. Me. Eduardo Pelli
Faculdade de Ciências Exatas – UFVJM

Prof. Me. Luís Henrique Silva Rodrigues
Faculdade de Ciências Exatas – UFVJM

Profa. Ana Carolina Rodrigues
Colégio Tiradentes - Diamantina

Diamantina

Dedico este trabalho aos amigos
e Luna, minha cadela companheira da vida.

AGRADECIMENTOS

Agradeço a Deus, por mais essa benção; aos amigos do curso, pelas descobertas, desafios e conquistas; ao professor Eduardo Pelli, pela oportunidade; aos componentes da banca de avaliação, pela atenção e sugestões; aos professores e técnicos do departamento de sistema de informação e, por fim, a todos que tornaram possível a realização desse trabalho.

Muito obrigado!

RESUMO

Os modelos matemáticos são abstrações que usam diferentes linguagens como álgebra, estatística, lógica e/ou algoritmos, e podem ampliar nossa compreensão sobre os sistemas em quase todos os ramos da ciência e da tecnologia. Na otimização de diferentes bioprocessos a modelagem matemática empírica dos dados experimentais pode ser realizada aplicando-se ferramentas como Metodologia de Superfície de Resposta (MSR) e Rede Neural Artificial (RNA). No presente trabalho foi realizada uma revisão sobre análises comparativas entre MSR e RNA aplicadas à modelagem e otimização das condições de cultura para diferentes micro-organismos e processos utilizando diferentes enzimas. No desenvolvimento são apresentadas as sequências metodológicas para a aplicação das ferramentas e os resultados para análises MSR *versus* RNA, considerando os diferentes estudos avaliados e as teorias relacionadas. Por fim, é realizada uma avaliação crítica para as análises comparativas e aplicação de RNA nos experimentos avaliados.

Palavras chave: Micro-organismos. Enzimas. Delineamentos experimentais. Modelagem matemática. Otimização.

ABSTRACT

Mathematical models are abstractions that use different languages and can broaden our understanding of almost all branches of science and technology. In the bioprocess optimization we can perform the empirical mathematical modeling for the experimental data using tools such as Response Surface Methodology (RMS) and Artificial Neural Network (ANN). In the present work a review was carried out on comparative analyses between RSM and ANN applied to the modeling and optimization of the culture conditions for different microorganisms and process using different enzymes. In development are presented the methodological sequences for the application of the tools and the results for RSM versus ANN analyses, considering the different studies evaluated ant the theories related issues. Finally, a critical analysis is carried out for the ANN application in the evaluated experiments.

Keywords: Microorganisms. Enzymes. Experimental Design. Mathematical Modeling. Optimization.

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

MM	Modelos Matemáticos
MSR	Metodologia de Superfície de Resposta
RNA	Rede Neural Artificial
OFATD	*On-Factor-At-a-Time* (Um fator por vez)
DPB	Delineamento Plackett-Burman
FF	Fatorial Fracionado
MT	Método Taguchi
DCC	Delineamento Composto Central
DCCC	Delineamento Composto Central Circunscrito
DCCCF	Delineamento Composto Central de Face Centrada
DCCI	Delineamento Composto Central Inscito
DCCR	Delineamento Composto Central Rotacional
DFC	Delineamento Fatorial Completo
DBB	Delineamento Box-Behnken
PC	Ponto Central
PA	Ponto Axial
ANOVA	Análise de Variância
p-valor	Probabilidade de Significância
H_0	Hipótese Estatística de Nulidade
R^2	Coeficiente de Determinação
R^2_{Aj}	R^2 Ajustado
R^2_P	R^2 da Predição
R	Coeficiente de Correlação
CV	Coeficiente de Variação
RMA	*Ridge Max Analysis* (Análise de Cumes)
AG	Algoritmo Genético

MOEA	*Multiobjective Evolutionary Algorithms Framework*
TR	Conjunto de Treinamento
T	Conjunto de Teste
V	Conjunto Validação
sTR	Subconjunto de Treinamento
sV	Subconjunto de Validação
CV	*Cross Validation* (Validação cruzada)
MLP	*Multi-layers Perceptron*
BP	*Backpropagation*
BBP	*Batch Backpropagation*
QP	Q*uick Propagation*
LM	Levenberg-Marquardt
GC	Gradiente Conjugado
TA	Taxa de Aprendizagem
M	Termo Momento
RMSE	*Root Mean Squared Error*
MSE	*Mean Squared Error*
MRE	*Mean Relative Error*
ARD	*Average relative deviation*
ADD	*Absolute Average Derivation*
RIO	*Rotation Inherit Optimization*

SUMÁRIO

1 INTRODUÇÃO

Os modelos matemáticos (MM) são abstrações que usam diferentes linguagens como álgebra, estatística, lógica e/ou algoritmos, e podem ampliar nossa compreensão sobre os sistemas em quase todos os ramos da ciência e da tecnologia (FLOOD e ISSA, 2010). Eles podem ser classificados em *(i)* mecanísticos (teoricamente derivados) ou *(ii)* empíricos. Um modelo mecanístico é desenvolvido a partir das leis e/ou princípios fundamentais que regem a resposta do sistema (FLOOD e ISSA, 2010). Já um modelo empírico limita-se a representar a relação entre dados amostrais, sendo uma aproximação dos sistemas reais (THOMPSON, 2011), ou seja, ele é desenvolvido a partir de observações da resposta do sistema em investigação e/ou de um análogo desse sistema em diferentes situações (FLOOD e ISSA, 2010).

Nos diferentes ramos da ciência e tecnologia, a modelagem empírica é aplicada para a otimização de diferentes bioprocessos, podendo ser utilizadas, com esse objetivo, as ferramentas *(i)* Metodologia de Superfície de Resposta (MSR) e *(ii)* Rede Neural Artificial (RNA). A MSR é uma coleção de técnicas matemáticas e estatísticas, úteis para modelagem e análise de processos experimentais, que avalia a possível influência da interação entre as variáveis independentes sobre a resposta (MONTOGOMERY, 2001). Comparada à técnica tradicional de análise experimental denominada *On-Factor-At-a-Time*, a MSR permite ainda o ajuste de modelos estatisticamente validados, com menor número de experimentos em diferentes tipos de delineamentos e, consequentemente, um menor tempo para as análises.

Em muitos processos experimentais, quando a relação entre variáveis independentes e dependentes não podem ser descritas por modelos lineares, a MSR não é suficiente para cumprir o objetivo de modelagem/otimização. Para esses casos, a modelagem utilizando RNA pode ser uma alternativa devido à possibilidade de ajustes não lineares (DESAI et al., 2018; WITEK-KROWIAK et al., 2014; VELU; VELAYUTHAM; MANICKKAM, 2016). Trata-se de uma metodologia biologicamente inspirada que utiliza neurônios artificiais para formarem redes de processamento (DESAI et al., 2018; WITEK-KROWIAK et al., 2014) e tem sido aplicada em diferentes áreas como química, finanças, física e ciências biológicas.

Atualmente muitos estudos comparam a capacidade preditiva dos modelos MSR e RNA (MURAKAR e SHASTRI, 2017; PATIL et al., 2017; SAMPAIO et al., 2017; KUMAR; CHHABRA; SHUKLA, 2017). Por sua vez, a modelagem RNA tem maior êxito na maioria dos estudos devido à possibilidade de ajuste lineares e não lineares. Porém, os estudos comparativos geralmente utilizam um pequeno conjunto de entrada-saída, proveniente dos delineamentos experimentais da MSR, para a obtenção dos modelos RNA, mesmo sabendo-se

que a modelagem RNA preconiza a utilização de um grande conjunto de dados.

Assim, a presente revisão teve por objetivo apresentar os resultados das análises comparativas MSR-RNA na modelagem empírica/otimização de condições para cultura de diferentes micro-organismos, bem como para processos utilizando diferentes enzimas, ambos relacionados a diferentes bioprocessos. No desenvolvimento da revisão são apresentadas as duas metodologias em dois tópicos que apresentam todas as estratégias realizadas nos trabalhos revisados. Posteriormente são apresentados os resultados das análises comparativas e, por fim, uma análise crítica para a aplicação de RNA com base nas experiências acadêmicas dos autores e na literatura relacionada.

2 METODOLOGIA DE SUPERFÍCIE DE RESPOSTA (MSR)

Para a aplicação da MSR a dados experimentais, as seguintes etapas devem ser cumpridas: *(i)* Seleção das variáveis independentes e possíveis respostas, *(ii)* Seleção do delineamento experimental, *(iii)* Execução dos experimentos segundo o delineamento, com obtenção dos resultados, (iv) Apresentação de modelo(s) matemático(s) (regressão) que descreve(m) os dados experimentais, *(v)* verificação estatística do modelo utilizando diferentes métricas, com avaliação de gráficos de superfície, *(vi)* determinação das condições (variáveis independentes) para a maximização ou minimização da resposta avaliada e, por fim, *(vii)* validação experimental do modelo e/ou da otimização. Nos próximos itens serão discutidos as etapas citadas em diferentes subtítulos, apresentando informações sobre os estudos que visavam a modelagem/otimização das condições de cultivo para diferentes micro-organismos e modelagem/otimização de diferentes processos que utilizaram enzimas. Os resultados apresentados a seguir são encontrados na Tabela 1. As discussões teóricas apresentadas foram realizadas segundo Snedecor e Cochran (1967), Montgomery (2001) e Rodrigues e Lemma (2009).

2.1 Seleção das variáveis independentes

Segundo Witek-Krowiak et al. (2015) a primeira e mais importante etapa em todo procedimento de modelagem compreende a seleção das variáveis e dos respectivos intervalos de valores. Geralmente essa decisão se baseia em dados da literatura, experiência com outros trabalhos e/ou realização de experimentos de triagem ou seleção de variáveis.

Para selecionar as variáveis e/ou seus respectivos intervalos dos valores, Mohamed et al. (2013), Schubert et al. (2015) e Patil et al. (2016, 2017) utilizaram os resultados de experimentos preliminares e/ou dados da literatura. Haque et al. (2016a) utilizaram informações obtidas a partir da aplicação da MSR em trabalho anterior (HAQUE et al., 2016b) para definir o tamanho dos *bead* e o tempo para lise/extração da enzima colesterol oxidase de células de uma *Escherichia coli* recombinante. Já Zafar et al. (2012), além de dados da literatura, realizaram um fatorial 4 x 3, combinando 4 possíveis fontes de nitrogênio e 3 fontes de carbono, em um único nível, para avaliar a produção de plástico biodegradável por *Azohydromonas lata*. Posteriormente, eles ainda realizaram um delineamento *On-Factor-At-a-Time* (OFATD) variando a percentagem de melaço de cana, o qual contém as fontes de carbono avaliadas no estudo (glicose, frutose e sacarose). Meena et al. (2014a), Pathak et al. (2015) e Kumar, Chhabra e Shukla (2017) também utilizaram o OFATD na avaliação das

Tabela 1 – MSR na modelagem/otimização das condições de cultura para micro-organismos e de utilização de enzimas para diferentes processos

Produção/Processo	**Micro-organismo/Enzima**	**Delineamento**[1]	**Tratamentos**[2]	**Software**[3]	**Referência**
Biomassa	*B. bifidum/L. acidophilus*[4]	DOFAT/DBB	6/54 (6PC)	*Minitab*	Meena et al. (2014)
Biomassa	*Kluyveromyces lactis*	DFC	29 (3PC)+3V	*Design Expert*	Sampaio et al. (2016)
Biomassa/Lipídio	*Tetraselmis* sp.	DCCC	19 (5PC+6 PA)	*Design Expert*	Mohamed et al. (2013)
Biomassa/Etanol	*Saccharomyces cerevisiae*	DCCC	20 (6PC+6PA))	*Design Expert*	Esfahanian et al. (2013)
Esporulação	*Monascus purpurens*	DCCC	41 (5PC+14PA)	*Design Expert*	Ajdari et al. (2012)
L-Glutaminase	*Bacillus cereus*	DCCC	54 (9PC+12 PA)	*Minitab*	Singh et al. (2013)
L-Asparaginase	*Aspergillus terreus*	DCCFC	32 (6 PC)	Nd	Baskar e Sahadevan (2012)
Xilanase	*Thermomyces lanuginosus*	DOFAT/DCCI	3/30 (3PC)	*Design Expert/MATLAB*	Kumar et al. (2017)
Lipase	*Escherichia coli*[5]	DPB/DCCC	32 (8PC+8PA)	*Statistica*	Nelofer et al. (2012)
PoliAGlacturonase	*T. frigidphilosprodundus*[6]	MT/DCCC	57/30 (6PC+8PA)	*Minitab/Design Expert*	Rekha et al. (2013)
Oxidase	*Streptomyces* sp.	OFAT/DCCC	8/36 (10PC +10PA)	*MATLAB*	Pathak et al. (2015)
Ácido Cítrico	*Aspergillus niger*	DBB	62 (Nd)	Nd	Kana et al. (2012)
Xilitol	*Debayomyces hansenii*	DFC	30 (4PC)	*Design Expert/MOEA*	Sampaio et al. (2017)
Goma Xantana	*Xanthomonas campestris*	DPB/DCCC	20/52 (10PC+10PA)	*Minitab/Design Exper*	Velu et al. (2016)
Goma	*Paenibacillus polymyxa*	DBB	54 (6PC)	*Design Expert*	Rafigh et al. (2014)
Plástico Biodegradável	*Azohydromonas lata*	DOFAT/DBB	1/17 (5PC)	*Design Expert/MATLAB*	Zafar et al. (2012)
Hidrogênio	*Enterobacter* spp.	DPB/DCCC	12/20 (4PC+6PA)	*Design Expert*	Karthic et al. (2013)
Iodine (I_2)[7]	Lacase	DCCI	36 (12PC)	Nd	Schubert et al., 2015
2-Etilhexil Ferrulat [8]	Lipase	DBB	15 (3PC)	Nd	Huang et al. (2016)
2-Fenietil Acetato [9]	Lipase	DCCC	27 (3PC+8PA)	*SAS*	Kuo et al. (2014)
Resveratrol [10]	Poligalacturonase	DCCC	27 (3PC+9PA)	Nd	Lin et al. (2016)
Hidrólise de resíduo[11]	Alcalase	DOFAT/DCCC	4/29 (5PC+8PA)	*Design Expert*	Zhang et al. (2016)

Cont. Tabela 1

Hidrólise de amido	α-Amilase+Glucoamilase	DBB	17 (5PC)	*Design Expert*	Olusola et al. (2014)
Extração de pectina	Protopectinase	DBB	24 (Nd)	*Minitab*	Murakar e Shastri (2017)
Purificação enzimática	Xilanase	DOFAT/DCCC	Nd/29 (3PC+10PA)	*Modde*	Rahimpour et al. (2016)
Extração de enzima	Arginina deaminase	DCCFC	30 (6PC)	Nd	Patil et al. (2017)
Extração de enzima	Colesterol oxidase	MSR/DCCC	Nd/30 (6PC+8PA)	Nd	Haque et al. (2016)
Extração de enzima	Arginina deaminase	DCCFC	20 (6PC)	*Design Expert*	Patil et al. (2016)

[1]Delienamento experimental, onde: DOFAT, Delinemanto o*ne-factor-at-a-Time;* DBB, delineamento Box-Behnken; DFC; delineamento fatorial completo; DCCC, delineamento composto central circunscrito; DCCFC, delineamento composto central de face centrada; DCCI, delineamento composto central inscrito; DPB; delineamento Plackett-Burman e MT, método Taguchi. [2]Tratamentos realizados segundo o delineamento, onde: PC, ponto central, PA, ponto axial e Nd, não definido. [3]Software utilizado para aplicar a MSR e/ou otimização, onde: Nd, não definido. [4]Mistura probiótica de *Bifidobacterium bifidum*+*Lactobacillus acidophilus*. [5]Espécie de *E. coli* Recombinante. [6]Espécie *Thalassospira frigidphilosprodundus*. [7]Antimicrobiano. [8]Antioxidante. [9]Aroma. [10]Fitoquímico. [11]Resíduo de camarão.

condições de cultura para a produção de biomassa probiótica, oxidase de *Streptomyces* sp. e xilanase de *Thermomyces lanuginosus*, respectivamente. De forma semelhante, em estudos de purificação de xilanase e hidrólise enzimática de resíduos (alcalase), Rahimpour, Hatti-Kaul e Mamo (2016) e Zhang et al. (2016), respectivamente, também utilizaram a mesma metodologia. Nesse delineamento, a influência individual de cada variável sobre uma ou mais respostas é determinada variando o valor de uma delas e fixando o das demais. Porém, como desvantagem, são realizados muitos experimentos, sendo necessário maior tempo para avaliação e, consequentemente, alto custo (WITEK-KROWIAK et al., 2015).

Outro delineamento utilizado com objetivo de triagem de variáveis é o *Plackett-Burman* (DPB) (PLACKETT e BURMAN, 1946 *apud* WITEK-KROWIAK et al., 2015). Ele permite avaliar as variáveis independentes e definir os efeitos das mesmas sobre a(s) resposta(as) avaliada(s), para que outros delineamentos finais com número reduzido de variáveis independentes possam ser utilizados. Dessa forma, reduz-se o tempo de análise e os custos relacionados. Dentre os estudos avaliados, Velu, Velayutham e Manickkam (2016), para 15 variáveis em diferentes níveis, escolhidas segundo dados da literatura e experiência do grupo de pesquisa, avaliaram a produção de goma (xantana) por *Xanthomonas campestris* utilizando um DPB. Do resultado dessa análise foram definidas cinco variáveis (concentração de glicose, peptona, KH_2PO_4, $(NH)_4SO_4$ e $FeCl_3.6H_2O$) que, posteriormente, compuseram um delineamento composto central (CCD). Utilizando os resultados de um PBD realizado previamente (NELOFER et al., 2010), Nelofer et al. (2012) selecionaram quatro variáveis significantes (glicose, NaCl, temperatura e tempo de indução) para otimizar as condições de cultura para produção de uma lipase de *Escherichia coli* recombinante.

Outra estratégia que pode ser utilizada é a realização de experimentos *steepest ascent*/*descent* após o DPB. Nela, os valores das variáveis na direção ascendente ou descendente são avaliados em novos experimentos, tentando melhorar o valor da resposta. Karthic et al. (2013), após realizar um DPB para sete variáveis independentes e definirem três delas para continuidade dos estudos, ainda utilizaram a técnica *steepest ascent* para definir o intervalo de valores das concentrações de xilose (ascendendo de 8 até 18 g/L) e de peptona (de 2 até 7 g/L), além do valor do pH (de 6,0 até 8,5) do meio, avaliando a produção de hidrogênio por uma espécie de *Enterobacter*. Um total de novos seis tratamentos foram realizados, combinando os menores valores de cada variável avaliada (8g/L de xilose+2,0 g/L de peptona+pH6,0) e assim, sucessivamente, até combinar os maiores valores (18g/L de xilose+7,0 g/L de peptona+pH8,5).

Utilização de Fatorial Fracionado (FF) também pode ser uma alternativa para a seleção ou triagem de variáveis independentes. Objetivando a produção de biomassa probiótica

(*Bifidobacterium bifidus+Lactobacillus acidophilus*), Meena et al. (2014a) avaliaram a influência de seis variáveis independentes utilizando um FF. Já Rekha et al. (2013), avaliando condições de cultura para o cultivo da bactéria *Thalassospira frigidphilosprofundus* para produção de poligalacturonase, aplicaram o método Taguchi (MT) (*L27 Orthogonalarray* – L_{27}OA) que utiliza fatoriais fracionados e arranjos ortogonais. Foram avaliadas a influência de seis variáveis em 27 experimentos, definindo-se quatro variáveis para continuidade dos estudos.

Os delineamentos PB e FF são indicados quando muitas variáveis independentes podem influenciar a(s) resposta(s) (*Screning Design*) e/ou quando se tem experimentos desconhecidos, ou seja, com nenhuma informação prévia para direcionamento. Segundo Rodrigues e Lemma (2005), para até oito variáveis independentes, tanto DPB quanto o FF podem ser utilizados. Já a partir de nove variáveis independentes o DPB torna-se a melhor alternativa por ser mais flexível quanto ao número de experimentos. Para essas duas metodologias são necessários no mínimo quatro ensaios a mais que o número de variáveis independentes e pelo menos três repetições na condição do ponto central (RODRIGUES e LEMMA, 2005).

Para os resultados do DPB é impossível ajustar um modelo de segunda ordem, ou seja, não é possível obter informações sobre as interações das variáveis, restringindo-se aos efeitos principais das mesmas (WITEK-KROWIAK et al., 2015). No FF podem ser obtidas informações sobre os efeitos principais e as interações de primeira ordem (RODRIGUES e LEMMA, 2005).

2.2 Delineamentos experimentais

Geralmente o delineamento composto central (DCC) é o mais utilizado dentre as diferentes possibilidades. Ele é definido como um delineamento que inclui um fatorial 2^k (k fatores), o ponto central que será executado com repetição para estimar o erro puro e os pontos axiais que determinam os termos quadráticos. O DCC pode ser do tipo *(i)* Composto Central Circunscrito (DCCC), *(ii)* Composto central de Face Centrada (DCCFC) ou *(iii)* Composto Central Inscrito (DCCI).

Um exemplo de DCCC é o delineamento denominado Composto Central Rotacional (DCCR). Para planejá-lo, alguns pressupostos devem ser atendidos: *(i)* realizar 2^K tratamentos fatoriais, *(ii)* no mínimo 3 tratamentos no ponto central e *(ii)* adicionar 2 x k pontos axiais (PA). Os PA adicionados no DCCR são do tipo $\pm\alpha$, onde $\alpha=(2^k)^{1/4}$.

Considerando os PA apresentados anteriormente, quando eles são os valores extremos das variáveis, ou seja, ±α=±1, o delineamento já é do tipo DCCFC. Por fim, quando os pontos axiais do DCC são <±1, o delineamento é do tipo DCCI.

A Tabela 1 apresenta os principais delineamentos utilizados para aplicação MSR na otimização de condições de cultura de micro-organismos e produção/processos envolvendo diferentes enzimas. Além dos delineamentos utilizados para a triagem de variáveis (OFATD, PPD, FFD e TM) e dos citados anteriormente (DCCC, DCCI e DCCFC), tem-se ainda a utilização dos delineamentos denominados Fatorial Completo (DFC) e Box-Behnken (DBB).

A maioria dos estudos aqui avaliados utilizou os delineamentos do tipo DCCC. Já Kumar, Chhabra e Shukla (2017), otimizando condições de cultivo para produção de xilanase por *T. lanuginosus*, e Schubert et al. (2015), avaliando a produção de antimicrobiano Iodine (I_2) utilizando a enzima lacase, realizaram um DCCI. Delineamentos do tipo DCCFC foram utilizados por Baskar e Sahadevan (2012) na produção de L-asparaginase por *Aspergillus terreus* e por Patil et al. (2016, 2017) na extração da enzima arginina deaminase de *Pseudomonas putida*.

Os DFC utilizados nos trabalhos de Sampaio et al. (2016, 2017) compreenderam a realização de 3^K tratamentos, adicionados das repetições no ponto central, para avaliar a produção de biomassa de *Kluyveromyces lactis* e de xilitol por *Debaryomyces hansenii*, respectivamente. Considerando três fatores, um total de 27 tratamentos, com duas (SAMPAIO et al., 2016) e três repetições (SAMPAIO et al., 2017) no ponto central foram realizados. Vale ressaltar que a utilização de três níveis nesse tipo de fatorial permite ajuste de modelos de segunda ordem.

Já o DBB (BOX e BEHNKEN, 1960 *apud* WITEK-KROWIAK et al., 2015) é um fatorial de três níveis incompletos, onde os tratamentos combinam três valores de três variáveis, sendo que dois deles estarão nos níveis +1 e/ou -1 e um no nível 0. Esse tipo de delineamento foi utilizado nos estudos de Meena et al. (2014a), Kana et al. (2012), Rafigh et al. (2014) e Zafar et al. (2012) que avaliaram as condições de cultura para produção de biomassa probiótica, de ácido cítrico por *Aspergillus niger*, de goma por *Paenibacillus polymyxa* e de plástico biodegradável por *Azohydromonas lata*, respectivamente. Olusola, Akindele e Abiodun (2014), Huang et al. (2016) e Murakar e Shastri (2017) também utilizaram o DBB avaliando a hidrólise enzimática de amido por ação de α-amilase+glucoamilase, a produção de um antioxidante por ação da enzima lipase e a extração enzimática de pectina, respectivamente.

Os delineamentos utilizados nos trabalhos foram realizados com no mínimo três tratamentos no ponto central (PC), chegando-se ao máximo de 12 PC como apresentado por

Schubert et al. (2015), que avaliaram a produção de um antimicrobiano. Para os delineamentos DCCC foram utilizados no mínimo seis pontos axiais (PA) até o máximo de 14 PA como descrito por Ajdari et al. (2012), que definiram as condições de cultivo para a esporulação do fungo *Monascus purpurens*. Essas variações devem-se ao número de variáveis avaliadas, aos delineamentos propostos e/ou decisões dos pesquisadores.

Alguns trabalhos descrevem sobre replicação de tratamentos e/ou repetição dos delineamentos, apresentando apenas as médias das respostas. Os tratamentos foram realizados em duplicata nos estudos realizados por Kuo et al. (2014), Haque et al. (2016a), Huang et al. (2016) e Lin et al. (2016). Já Zafar et al. (2012), Singh et al. (2013), Rafigh et al. (2014) e Schubert et al. (2015) realizaram os tratamentos em triplicata, sendo que os últimos autores realizaram uma repetição total do delineamento. A replicação e/ou repetição são estratégias experimentais que dão maior confiabilidade aos resultados experimentais coletados e, consequentemente, contribuem para um modelo de regressão fidedigno à realidade experimental.

A ordem de execução dos tratamentos experimentais deve ser randômica para evitar o viés (SINGH et al., 2013; KUO et al., 2014; HUANG et al., 2016; LIN et al., 2016; ZHANG et al., 2016; RAHIMPOUR; HATTI-KAUL; MAMO, 2016). Porém, Singh et al. (2013) apresentaram pontos centrais com o mesmo valor de resposta ao avaliarem condições de cultura para produção de L-glutaminase por *Bacillus cereus*.

Por fim, dentre os estudos, Mohamed et al (2013) e Esafahanian et al. (2013), cujos delineamentos experimentais foram do tipo DCCC, avaliaram as condições de cultura para otimizar a produção de biomassa/lipídio por *Tetraselmis* sp. e biomassa/etanol por *Saccharomyces cerevisiae*, respectivamente. Já Kuo et al. (2014) e Lin et al. (2016), também utilizando um DCCC, avaliaram a produção de aroma utilizando uma lipase e a extração do fitoquímico resveratrol por ação da poligalacturonase, respectivamente.

2.3 Análise estatística da regressão e gráficos superfície de resposta

Após a realização dos delineamentos e a partir dos resultados dos tratamentos, um modelo de regressão é proposto para cada resposta avaliada, geralmente utilizando os softwares estatísticos citados no item 2.5. A partir do modelo de regressão os seguintes passos são cumpridos, em sequência: *(i)* Avaliar a significância estatística da regressão, *(ii)* Avaliar a significância da falta de ajuste, (*iii)* Avaliar os coeficientes ou termos da regressão, eliminando os não significativos ou mantendo os não significativos devido à hierarquia, (*iv)*

Avaliar as superfícies geradas a partir da regressão avaliada estatisticamente e, por fim, *(v)* Realizar a otimização.

A maioria das regressões ajustadas aos problemas apresentados foram modelos polinomiais de segunda ordem. Já Ajdari et al. (2012), após avaliarem os possíveis ajustes com os modelos polinomiais dos tipos linear, de dois fatores, quadrático e cúbico, observaram insignificância estatística para os mesmos na análise de variância (ANOVA). Assim, eles utilizaram a metodologia denominada *backward (i)* partindo do modelo cúbico, *(ii)* eliminando os termos com maiores valores de *p-valor* um a um, respeitando a hierarquia, e, por fim, *(iii)* realizando a análise estatística da regressão obtida. Seguindo essa sequência iterativa, estes autores chegaram a um modelo final que foi definido como quase quadrático.

Na ANOVA, o teste estatístico deve revelar o modelo como significativo enquanto a falta de ajuste deve ser não significativa, para que, assim, a modelagem continue sendo avaliada. Porém, um modelo significativo com falta de ajuste significativa pode ainda ser avaliado pelo coeficiente de determinação (R^2). Caso o valor dessa métrica seja próximo a 1, ainda é possível utilizar o modelo para a etapa de otimização, mesmo com falta de ajuste.

O próximo passo é a avaliação estatística dos coeficientes (termos) da regressão. Para esse passo, Baskar e Sahadevan (2012), Nelofer et al. (2012) e Rekha et al. (2013) utilizaram o teste t de Student associado ao *p-valor*. Já os demais estudos utilizaram o teste F associado ao *p-valor* para essa finalidade. Esse último teste também foi utilizado para avaliar a significância da regressão e da falta de ajuste, segundo a análise de variância (ANOVA), em todos os estudos avaliados. Não existe uma explicação para a escolha do teste estatístico (t *Student* ou F), sendo ela uma decisão do pesquisador.

Por sua vez, o *p-valor* (probabilidade de significância ou nível descritivo) indica a possibilidade da hipótese estatística de nulidade (H_0) ser rejeitada no teste estatístico. Se *p-valor*$<\alpha$ (nível de confiança pré-definido), H_0 é rejeitada, significando, por exemplo, que o modelo ou um coeficiente da regressão avaliado são significativos. Por outro lado, se *p-valor*$>\alpha$, H_0 é aceita e, por exemplo, a falta de ajuste é considerada não significativa.

Uma métrica bastante utilizada para avaliar modelos MSR é o coeficiente de determinação da regressão (R^2). Esta métrica determina a variação da resposta avaliada que é explicada pela regressão, ou seja, uma medida do grau de ajustamento da regressão aos dados experimentais. Além do R^2, as seguintes métricas também podem ser utilizadas: *(i)* R^2 ajustado (R^2_{Aj}) (NELOFER et al., 2012; KANA et al., 2012; ZAFAR et al., 2012; ESFAHANIAN et al., 2013; MOHAMED et al., 2013; REKHA et al., 2013; RAFIGH et al., 2014; ZHANG et al., 2016; PATIL et al.; 2017), *(ii)* R^2 da predição (R^2_P) (ESFAHANIAN et al., 2013; REKHA et al., 2013; RAFIGH et al., 2014), *(iii)* coeficiente de correlação (R)

(KANA et al., 2012; PATHAK et al., 2015; PATIL et al., 2017) e, por fim, *(iv)* coeficiente de variação (CV) (KUMAR; CHHABRA; SHUKLA, 2017).

A determinação dos três tipos de R^2 (R^2, R^2_{Aj} e R^2_P) é importante, principalmente para avaliar um possível *overfitting* do modelo MSR. O R^2_{Aj} é um R^2 modificado que considera o número de coeficientes na regressão (CR) para o seu cálculo. A adição de CR pode ou não aumentar o poder de explicação do R^2_{Aj}. Enquanto essa adição, por si só, aumenta o valor de R^2, ela pode diminuir o valor de R^2_{Aj} caso o poder de explicação da regressão não aumente. Assim, o valor de R^2_{Aj} será menor ou igual a R^2, sendo que a adição de coeficientes no modelo pode aumentar ou diminuir o seu valor. Por sua vez, R^2_P, quando avaliado, demonstra a capacidade de uma regressão predizer a resposta para observações desconhecidas. Para seu cálculo, um tratamento de entrada é retirado do conjunto de dados e utilizado para testar a capacidade de predição do modelo.

A métrica R é uma medida da intensidade da relação linear entre as variáveis. Na MSR ela revela sobre a aceitabilidade da correlação entre os valores preditos e os valores experimentais, sendo melhor a correlação quando o seu valor é o mais próximo de +1. Por sua vez, CV é definido como a razão entre o erro padrão da estimativa e a média dos valores experimentais. Ele é uma medida de reprodutibilidade e repetitividade dos modelos (CHEN; BUI; KRZYZAK, 2010) e indica a precisão e confiabilidade dos experimentos. Geralmente, quanto menor (<10%), maior será a confiabilidade dos experimentos realizados (SNEDECOR e COCHRAN, 1967).

Após avaliação estatística do modelo, as relações matemáticas podem ser representadas graficamente, seja na forma de gráficos de contorno (BASKAR e SAHADEVAN, 2012; KARTHIC et al., 2013; SINGH et al., 2013; RAFIGH et al., 2014; SAMPAIO et al., 2017) e/ou em gráficos 3D (NELOFER et al., 2012; AJDARI et al., 2012; ESFAHANIAN et al., 2012; KANA et al., 2012; MOHAMED et al., 2013; RAFIGH et al., 2014; VELU; VELAYUTHAM; MANICKKAM, 2016; KUMAR; CHHABRA; SHUKLA, 2017). A geração dos gráficos e a interpretação das superfícies geradas permite avaliar a curvatura das superfícies de resposta e, assim, definir se a região ótima está ou não dentro dos limites das variáveis independentes avaliadas.

2.4 Otimização e validação experimental

Para a otimização, quando apenas uma resposta é avaliada, podem ser seguidos os seguintes passos: *(i)* cálculo da primeira derivada, *(ii)* cálculo da segunda derivada, para avaliar a existência de ponto de sela e, caso os resultados das derivadas anteriores seja zero,

(iii) uma avaliação gráfica para definição de extremos locais (WITEK-KROWIAK et al., 2015). Já no caso de múltiplas respostas pode ser utilizada a otimização da função de *desirability* por métodos numéricos, sendo que maiores informações são descritas por Witek-Krowiak et al. (2015).

As metodologias de otimização do modelo MSR dos diferentes trabalhos avaliados podem ser resumidas em: *(i)* otimização numérica/gráfica utilizando o software selecionado (maioria dos trabalhos), *(ii)* avaliação aplicando a metodologia *Ridge Max Analysis* (RMA) – análise de cumes máximos (KUO et al., 2014; HUANG et al., 2016), *(iii)* utilização de um algoritmo genético (AG) (ZAFAR et al., 2012; PATHAK et al., 2015; KUMAR; CHHABRA; SHUKLA, 2017; SAMPAIO et al., 2017) e, por fim, *(iv)* avaliação dos gráficos de superfície gerados por s*takeholders* (ESFAHANIAN et al., 2012; RAFIGH et al., 2014; VELU; VELAYUTHAM; MANICKKAM, 2016).

O método RMA permite encontrar o máximo absoluto dentro da superfície do modelo matemático, pesquisando em esferas concêntricas de raios variados e que estão centralizadas pelo tratamento que representa o ponto central (x_1=0, x_2=0…, x_k=0). Já o AG, segundo Sampaio et al. (2017) pode ser utilizado para a otimização multi-objetivo, quando se tem várias respostas para mesmas entradas, obtendo um conjunto de soluções não dominadas no fronte de Pareto. Trata-se de um algoritmo baseado em genética de populações, que realiza a otimização em quatro etapas que imitam os processos da evolução natural: *(i)* Inicialização de populações de soluções, conhecidas como cromossomos, *(ii)* Cálculo do ajuste com base na função objetiva, *(iii)* Seleção dos melhores cromossomos e *(iv)* Propagação genética de cromossomos parentais selecionados usando operadores genéticos (crossover e mutação) para criar uma nova população cromossômica (MICHALEWICZ, 1992; DESAI et al., 2008). Este ciclo é repetido até a convergência para o Fronte de Pareto, que apresenta as soluções não dominadas.

Após definir as entradas para uma saída ótima é necessário realizar a validação experimental, ou seja, avaliar a capacidade de predição do modelo de regressão proposto para a entrada-saída ótima. A maioria dos trabalhos realiza a validação experimental em triplicata, obtendo boa capacidade preditiva para os modelos propostos. Já Rahimpour, Hatti-Kaul e Mamo (2016) e Sampaio et al. (2016) não apresentaram os resultados sobre a validação experimental. Os últimos autores dispensaram a necessidade da validação ao considerarem que a entrada predita como ótima foi praticamente um dos pontos do delineamento.

2.5 Softwares utilizados

Os programas mais utilizados para geração dos delineamentos e/ou análises segundo a MSR são: *(i) Design Expert* (Stat-Ease, Inc.)*, (ii) Minitab* (Minitab Inc.), *(iii) Matlab* (MathWorks)*, (iv) Statistica* (StatSoft)*, (v) SAS* (SAS Institute Inc.) *e (vi) modde* (Umeå).

Alguns estudos utilizaram mais de um software para cumprir diferentes objetivos. Rekha et al. (2013) utilizaram o software Minitab para a metodologia de Taguchi (TM) e o *Design Expert* para avaliar o DCCC. Por sua vez, Velu, Velayutham e Manickkam (2016) utilizaram o software *Design Expert* apenas para geração dos gráficos 3D. Já Zafar et al. (2012) e Kumar, Chhabra e Shukla (2017) utilizaram o software Matlab na etapa de otimização do modelo MSR, enquanto Sampaio et al. (2017) utilizaram o Framework MOEA (*Multiobjective Evolutionary Algorithms Framework*) com esse mesmo objetivo. Baskar e Sahadevan (2012) utilizaram o MATLAB para a otimização do modelo MSR, mas não ficou claro no trabalho se esse mesmo software foi utilizado para as demais análises.

3 REDE NEURAL ARTIFICIAL (RNA)

Resumidamente, para a obtenção de um modelo RNA, as seguintes etapas devem ser cumpridas: *(i)* seleção dos dados para a obtenção da RNA, *(ii)* normalização e divisão dos dados em conjuntos, *(iii)* treinamento e teste para obtenção da RNA, com definição de diferentes parâmetros (seleção da arquitetura da rede, algoritmo de aprendizagem, função de transferência, taxa de aprendizagem, dentre outros) e, por fim, *(iv)* avaliação do modelo RNA (validação experimental, representação gráfica, otimização, dentre outros). Nos próximos itens serão discutidas essas etapas em diferentes subtítulos, apresentando informações sobre os mesmos estudos onde foram aplicadas a MSR. Todos os resultados discutidos a seguir são apresentados na Tabela 2. As discussões teóricas apresentadas foram realizadas segundo Jang, Sun e Mizutani (1997), Braga, Carvalho e Ludemir (200) e Haykin (2002).

3.1 Dados de entrada

Uma RNA aprende "por exemplo" utilizando um conjunto de entrada-saída. A maioria dos trabalhos avaliados utilizou os dados dos delineamentos experimentais da MSR para obtenção do modelo RNA, com algumas exceções. Além dos tratamentos dos delineamentos, Mohamed et al. (2013) e Sampaio et al. (2017) utilizaram um total de 10 e 3 tratamentos experimentais desconhecidos, respectivamente, para realização da etapa de validação. Por sua vez, Rekha et al. (2013) utilizaram também os dados experimentais do TM, totalizando 57 tratamentos. Já Zhang et al. (2016) utilizaram os dados experimentais obtidos nos delineamentos OFATD e CCCD (50 tratamentos). De forma única entre os trabalhos avaliados, Meena et al. (2014a) utilizaram resultados experimentais prévios do cultivo de *Bifidobacterium bifidum* (MEENA et al., 2011) para obtenção de uma RNA que descreveu o crescimento do consórcio probiótico *B. bidium*+*Lactobacillus acidophilus* (38 tratamentos).

Em relação à utilização de entradas repetidas para o treinamento da RNA, sabe-se que a utilização de replicatas não melhoraram a capacidade preditiva do modelo. Assim, para os tratamentos no ponto central dos delineamentos, apenas um valor único médio foi utilizado em alguns estudos (AJDARI et al., 2012; SINGH et al., 2013; SAMPAIO et al., 2016; SAMPAIO et al., 2017).

O maior número de exemplos (62 dados de entrada-saída) utilizados para obtenção de uma RNA foi utilizado por Kana et al. (2012), enquanto o menor número (15 dados de entrada-saída) foi utilizado na modelagem realizada por Huang et al. (2016). Entretanto, o

Tabela 2 – RNA na modelagem/otimização das condições de cultura para micro-organismos e de utilização de enzimas para diferentes processos

Produção/Processo	**Micro-organismo/Enzima**	**Dados**[1]	**Método**[2]	**Topologia**[3]	**Software**[4]	**Referência**
Biomassa	*B. bifidum* /*L. acidophilus*[5]	29 (Nd)	BP/AG	6-11-1	*MATLAB*	Meena et al. (2014)
Biomassa	*Kluyveromyces lactis*	30 (27TR+27CV+3V)	BP/AG	3-default-1	*WEKA/MOEA*	Sampaio et al. (2016)
Biomassa/Lipídio	*Tetraselmis* sp.	29 (15TR+4T+10V)	LM/RIO	3-10-1	*NeuralPower*	Mohamed et al. (2013)
Biomassa/Etanol	*Saccharomyces cerevisiae*	20 (70%TR/T+30%V)	LM/-	3-6-1(1)	*MATLAB*	Esfahanian et al. (2013)
Esporulação	*Monascus purpurens*	37 (33TR+4T)	BP/Nd	7-15-1	*NeuralPower*	Ajdari et al. (2012)
L-Glutaminase	*Bacillus cereus*	45 (33TR+6T+6V)	LM/Nd	6-3-1	*Neurosolutions*	Singh et al. (2013)
L-Asparaginase	*Aspergillus terreus*	32 (25TR +7T)	BP/AG	5-4-1	*NeuralPower*	Baskar e Sahadevan (2012)
Xilanase	*Thermomyces lanuginosus*	30 (70%TR+30%T/V)	Nd/AG	4- Nd[8]-1	*MATLAB*	Kumar et al. (2017)
Lipase	*Escherichia coli*[6]	32 (16TR+8T+8V)	BP-CG/-	4-9-1	*STATISTICA*	Nelofer et al. (2012)
Poligalacturonase	*T. frigidphilosprodundus*[7]	57 (Nd)	BP/AG	4-10-10-10-1	Nd	Rekha et al. (2013)
Oxidase	*Streptomyces* sp.	36 (24TR+6T+6V)	LM/AG	5-15-1	*MATLAB*	Pathak et al. (2015)
Ácido Cítrico	*Aspergillus niger*	62 (52TR+10T)	LM/AG	7-5-1	*easyNN*	Kana et al. (2012)
Xilitol	*Debayomyces hansenii*	27 (27TR+27CV)	BP/AG	3-default-1[9]	*WEKA/MOEA*	Sampaio et al. (2017)
Goma Xantana	*Xanthomonas campestris*	52 (26TR+26T)	BP/Nd	5-Nd-Nd-1	Nd	Velu et al. (2016)
Goma	*Paenibacillus polymyxa*	54 (70%TR+30%T/V)	BP-CV/-	6-13-1	*MATLAB*	Rafigh et al. (2014)
Plástico Biodegradável	*Azohydromonas lata*	Nd	BP/AG	3-4-1	*MATLAB*	Zafar et al. (2012)
Hidrogênio	*Enterobacter* spp.	20 (80%TR+20%T)	LM/Nr	3-8-1	Nd	Karthic et al. (2013)
Iodine (I_2)[10]	Lacase	36 (Nd)	LM/AG	4-2-1	*MATLAB*	Schubert et al., 2015
2-Etilhexil Ferrulat[11]	Lipase	15 (Nd)	LM/RMA	3-6-1	*NeuralPower*	Huang et al. (2016)
2-Fenietil Acetato[12]	Lipase	27 (Nd)	BBP/RMA	4-6-1	*NeuralPower*	Kuo et al. (2014)

Cont. Tabela 2

Resveratrol[13]	Poligalacturonase	27 (Nd)	BBP/-	4-15-1	*NeuralPower*	Lin et al. (2016)
Hidrólise de resíduos[14]	Alcalase	50 (70%TR+30%T/V)	BP/-	4-11-1	*MATLAB*	Zhang et al. (2016)
Hidrólise de amido	α-amilase+Glucoamilase	17 (14TR+3T)	QP/AG	3-15(16)-15(16)-1	*NeuralPower*	Olusola et al. (2014)
Extração de pectina	Protopectinase	24(75-80%TR+25-20%T)	Nd/-	2-10-10-10-1	*elite-ANN*	Murakar e Shastri (2017)
Purificação enzimática	Xilanase	29 (Nd)	LM/-	5-7(9, 6, 6)-1	*MATLAB*	Rahimpour et al. (2016)
Extração de enzima	Arginina Deaminase	30 (Nd) + 6V	LM/-	4-6-3	*MATLAB*	Patil et al. (2017)
Extração de enzima	Colesterol Oxidase	30 (20TR+5T+5V)	LM/AG	4-Nd-1	*MATLAB*	Haque et al. (2016)
Extração de enzima	Arginina Deaminase	20 (Nd)+8V	LM/-	3-3(3,3)-1	*MATLAB*	Patil et al. (2016)

[1]Dados utilizados para obtenção da RNA, onde: TR, conjunto de treinamento; T, conjunto de teste; V, conjunto de validação; CV, conjunto para *cross validation* e Nd, não definido. [2]Método de treinamento da RNA/Otimização, onde: BP, *backpropagation;* AG, algoritmo genético; LM, algoritmo Levenberg-Marquardt*;* RIO, r*otation inherit optimization*; GC, algoritmo gradiente conjugado; RMA, r*idge max analysis;* BBP, *Batch Backpropagation;* QP, *Quick Propagation*, (-), não realizou otimização da RNA e Nd, não definido. [3]Topologia da RNA, onde: primeiro/último número, número de neurônios na camada de *input* e *output*, respectivamente; números do meio, número de camadas escondidas e de neurônios das mesmas; número entre parênteses, número de neurônios para outras topologias apresentadas; *default,* padrão definido pelo software e Nd, não definido. [4]Software utilizados para treinamento e otimização da RNA, onde: Nd, não definido. [4]Software utilizados para treinamento e otimização da RNA. [5]Mistura probiótica de *Bifidobacterium bifidum*+*Lactobacillus* acidophilus. [6]Espécie de *E. coli* Recombinante. [7]Espécie *Thalassospira frigidphilosprodundus*. [8]Utilizaram 10 camadas escondidas, mas não apresentam o número de neurônios. [9]Foram modeladas 4 repostas que apresentam essa mesma topologia. [10]Antimicrobiano. [11]Antioxidante. [12]Aroma. [13]Fitoquímico. [14]Resíduo de camarão.

maior valor citado ainda é pequeno diante do tamanho dos conjuntos de dados geralmente utilizados para obtenção de uma boa modelagem RNA.

3.2 Normalização e divisão dos dados de entrada-saída

A normalização é uma etapa importante porque as funções de transferência/ativação trabalham na região saturada da curva quando os valores do conjunto entrada-saída são de altas grandezas. Segundo Chojaczyk et al. (2015) essa normalização é importante porque as diferenças nas grandezas dos valores resultarão em uma diferença insignificante na saída da função, o que dificulta o processo de treinamento. Dentre os trabalhos avaliados, duas possibilidades foram *(i) a* normalização para valores entre 0,1-0,9 (ZAFAR et al., 2012; KARTHIC et al., 2013) e *(ii)* para valores entre 0 e 1 (ESFAHANIAN et al., 2013; RAHIMPOUR; HATTI-KAUL; MAMO, 2016). A normalização dos dados de entrada-saída também foi realizada nos estudos de Rafigh et al. (2014), Haque et al. (2016a), Patil et al. (2016), Rahimpour, Hatti-Kaul e Mamo (2016), Zhang et al. (2016) e Patil et al. (2017).

Para a maioria dos trabalhos, os dados de entrada selecionados foram divididos aleatoriamente em diferentes grupos, sendo eles: *(i)* conjunto de treinamento (TR), *(ii)* teste (T) e *(iii)* validação (V). A Tabela 2 apresenta a divisão dos tratamentos entre os grupos, sendo que os grupos de TR+T geralmente compreendem mais de 70% dos dados. Porém, nem todos os estudos apresentaram essas informações, além disso, alguns não descrevem sobre conjunto de validação.

Uma alternativa para o treinamento de uma RNA é a utilização da validação cruzada (*Cross validation*, CV). Nela, o conjunto de dados é dividido em conjunto de TR e conjunto de V. Por usa vez, o conjunto de TR é dividido em subconjunto de Treinamento (sTR), utilizado para selecionar o modelo RNA, e subconjunto de Validação (sV), usado para validar o modelo. Sampaio et al. (2016, 2017) utilizaram o método de CV denominado *k-fold*, mais precisamente o *10-fold cross validation* (CV). Neste método, o conjunto de dados de entrada (n) é dividido em 10 subconjuntos (10 fold), sendo 9 utilizados para treinamento (TR) e 1 para validação (V). Posteriormente, o subconjunto V é utilizado para TR e outro subconjunto, antes classificado como TR, é utilizado como subconjunto V. Essa etapa é repetida iterativamente até que todos os conjuntos tenham sido TR e V em algum ciclo. Dos dois estudos que utilizaram CV, apenas Sampaio et al. (2017) realizaram uma etapa de validação utilizando um grupo desconhecido de 3 tratamentos. Vale ressaltar que a VC é uma estratégia alternativa quando se tem pequeno conjunto de dados entrada-saída, já que todo o conjunto de

TR é utilizado como subconjunto de TR e V, melhorando o cálculo dos erros relacionados e evitando, assim, a possibilidade de *overfitting/overtraining*.

3.3 Treinamento e teste

Após a seleção, normalização e divisão do conjunto dos dados, a próxima etapa nos estudos avaliados foi a realização do treinamento que seguiu o paradigma supervisionado. Nessa etapa, o objetivo é convergir uma rede, ou seja, a rede aprenderá sobre o conjunto entrada-saída pela definição dos parâmetros livres estáveis (pesos sinápticos e níveis de *bias*) que levem à minimização do erro, num processo iterativo (CHOJACZYK et al., 2015).

Por sua vez, o erro é calculado comparando a saída da rede com a saída real. Caso o(s) critério(s) de convergência (parada) de treinamento, baseado no erro calculado, não seja(m) atendido(s), é utilizado um método matemático (algoritmo de treinamento) para atualizar os valores dos parâmetros livres. Caso contrário, a próxima etapa, ou seja, a de teste será realizada.

No teste o objetivo é avaliar a capacidade de generalização da rede, calculando também o erro. Uma rede apresenta boa capacidade de generalização quando o mapeamento de entrada/saída computado é, pelo menos, aproximadamente correto para o teste. Assim, o treinamento será interrompido quando os erros (de treinamento e teste) atingirem o menor valor, aproximadamente constante, e que não aumente numa próxima iteração do treinamento. Para determinação do erro aceitável é utilizado o método de gradiente de erro, sobre o qual, maiores informações podem ser obtidas em Haykin (2002).

Geralmente os parâmetros para obtenção da RNA são selecionados por tentativa e erro, o que pode dificultar o ajuste dos mesmos e, consequentemente, a obtenção de uma boa modelagem. Nos próximos subitens será apresentada uma sequência de boas práticas para o cumprimento da etapa de treinamento e teste, baseada nos estudos avaliados.

3.3.1 Arquitetura ou topologia

A arquitetura utilizada em todos os estudos avaliados foi do tipo *Multi-layers Perceptron* (MLP) *feedforward*. Segundo Chojaczyk et al. (2015) essa é a arquitetura mais popular para processos de modelagem utilizando RNA. Uma rede MLP é composta por um conjunto de unidades sensoriais chamados nós (ou neurônios) de fonte que compõem a camada de entrada (*input layer*), uma ou mais camadas ocultas ou intermediárias (*hidden layer*) de nós computacionais e uma camada de saída (*output layer*) de nós computacionais

(HAYKIN et al., 2002). Estas camadas estão conectadas (conexões sinápticas) e o fluxo de entrada do conjunto de dados é sempre da entrada para saída (*Feedforward)*, ou seja, no sentido positivo, sem ciclos (HAYKIN et al., 2002).

A maioria dos estudos apresentados utilizou apenas uma camada de entrada, uma escondida e uma de saída de neurônios (Tabela 2). Já Olusola, Akindele e Abiodun (2014) e Velu, Velayutham e Manickkam (2016) utilizaram duas camadas escondidas. Por sua vez, Rekha et al. (2013), Murakar e Shastri (2017) e Kumar, Chhabra e Shukla (2017) utilizaram três e 10 camadas escondidas, respectivamente.

Para definir a melhor topologia de rede é possível ainda variar o número de neurônios da camada escondida. Aumentando esse número pode-se aumentar a capacidade de mapeamento não linear da rede, já que são estes neurônios que capacitam as redes a extraírem progressivamente as características mais significativas das entradas, ou seja, representar o comportamento de não linearidade para os dados de entrada-saída. Porém, sempre existe a possibilidade de *overfitting*/*overtraining* ou *underfitting/undertraining* associado à utilização de um grande e pequeno número de neurônios na camada escondida, respectivamente.

Para definir o número de neurônios na camada escondida, Baskar e Sahadevan (2012) utilizaram duas regras: *(i)* o número de neurônio deve ser um valor entre o número de entradas e de saídas utilizados e *(ii)* deve ser 2/3 do tamanho da camada de entradas somado ao tamanho da camada de saída. Entretanto, existem ainda outras regras definidas nos livros de referência (HAYKIN et al., 2002).

Os neurônios devem ser adicionados um a um, com avaliação da topologia obtida após cada acréscimo (MOHAMED et al., 2013). Dentre os estudos, o número de neurônios avaliados variou de *(i)* 3 a 6 neurônios (HUANG et al., 2016), *(ii)* 4 a 15 (KUO et al., 2014), *(iii)* 1 a 20 (OLUSOLA; AKINDELE; ABIODUN, 2014; RAFIGH et al., 2014), *(iv)* 1 a 10 (RAHIMPOUR; HATTI-KAUL; MAMO, 2016; PATIL et al., 2017), *(v)* 1 a 3 (PATIL et al., 2016), *(vi)* 5 a 30 (MOHAMED et al., 2013) e *(vii)* 1 a 6 (ZAFAR et al., 2012). Como alternativa, Sampaio et al. (2016, 2017) utilizaram um número *default* de neurônios para a camada escondida segundo o software WEKA (*Waikato Environment for Knowledge Analysis*, *University of Waikato*). Já o número de neurônios da camada de entrada e saída estão relacionados ao número de variáveis independentes e resposta, respectivamente.

Para as melhores topologias dos estudos (Tabela 2), o maior e menor número de neurônios utilizados foi dois e 16 para as topologias descritas por Schubert et al. (2015) e Olusola, Akindele e Abiodun (2014), respectivamente. Vale ressaltar que Esfahanin et al. (2013), Olusoloa et al. (2014), Rahimpour, Hatti-Kaul e Mamo (2016) e Patil et al. (2016)

avaliaram mais de uma resposta, gerando de 2 a 3 topologias, ou seja, uma topologia para cada resposta.

3.3.2 Algoritmo de treinamento/aprendizado

Na etapa de treinamento ou aprendizagem, enquanto alguns estudos avaliam a influência do algoritmo utilizado variando-o (AJDARI et al., 2012; MOHAMED et al., 2013; KUO et al., 2014; OLUSOLA; AKINDELE; ABIODUN, 2014; LIN et al., 2016; HUANG et al., 2016; PATIL et al., 2017), a maioria utiliza um único algoritmo. Vale ressaltar que os algoritmos diferem entre si principalmente pelo modo de modificação dos parâmetros livres da rede.

Dentre os algoritmos avaliados e/ou utilizados, destacam-se: *(i) backpropagation* (BP), *(ii)* b*atch backpropagation* (BBP), *(iii)* q*uick propagation (QP)*, *(iv)* Levenberg-Marquardt (LM), *(v)* Algoritmo Genético, *(vi) BFGS Quasi-Newton, (vii)* gradiente conjugado (GC) (*Powell/Beale Restarts*, *Fletcher-Powell*, *Polak-Ribiére* e *scaled back propagation*) e *(viii) Incrimental Learning* (IL).

O algoritmo BP merece destaque pela frequência de utilização para as melhores topologias apresentadas nos estudos. O BP é o algoritmo que se ajusta bem para a modelagem de uma grande variedade de problemas e adequa-se computacionalmente às arquiteturas MLP. Ele atua em dois passos distintos: *(i) forward* e *(ii) backward*. Enquanto no primeiro passo a informação flui da camada de entrada até a camada de saída, utilizando pesos e funções de ativação/transferência, no segundo, o fluxo faz o caminho inverso, ou seja, da camada de saída até a camada de entrada, realizando ajustes dos parâmetros livres. Esse ajuste é feito de forma sequencial, ou seja, após a apresentação de cada exemplo de treinamento. Durante o passo para frente os parâmetros livres são fixos, enquanto que no passo para trás eles são ajustados de acordo com a regra de correção de erro, utilizando o gradiente descendente para encontrar o mínimo local para uma função de erro.

Por sua vez, LM é baseado no método Gauss-Newton e utiliza uma função para correção dos pesos exclusiva que utiliza a matriz jacobina de pesos e o vetor de erros (JANG; SUN; MIZUTANI, 1997). Para o LM a atualização dos parâmetros livres considera tanto o gradiente descendente para o erro como a curvatura da superfície gerada nesse gradiente. Outros dois algoritmos utilizados foram o GC descendente (NELOFER et al., 2012) e o QP (OLUSOLA; AKINDELE; ABIODUN, 2014). Enquanto o primeiro utiliza uma taxa de aprendizagem não fixa que varia a cada iteração, o segundo tolera altas taxas de aprendizado.

Por sua vez, Kuo et al. (2014) e Lin et al. (2016) utilizaram o BBP, que difere do BP apenas pela atualização dos pesos que ocorre após a apresentação de todo o conjunto de treinamento.

3.3.3 Função de ativação/transferência

Alguns estudos avaliaram diferentes funções para as camadas escondida e/ou de saída (AJDARI et al., 2012; KANA et al., 2012; MOHAMED et al., 2013; KUO et al., 2014; LIN et al., 2016; OLUSOLA; AKINDELE; ABIODUN, 2014). Dentre as funções não lineares destacam-se três: *(i) S*igmóide (ESFAHANIAN et al., 2013; KARTHIC et al., 2013; SINGH et al., 2013; PATHAK et al., 2015; HUANG et al., 2016; ZHANG et al., 2016; KUMAR; CHHABRA; SHUKLA, 2017), *(ii)* Tangente Hiperbólica (BASKAR e SAHADEVAN, 2012; RAFIGH et al., 2014; SCHUBERT et al., 2015; PATIL et al., 2016; SAMPAIO et al., 2016; VELU; VELAYUTHAM; MANICKKAM, 2016; PATIL et al., 2017) e *(iii)* Gaussiana. Já a linearidade para um modelo RNA foi avaliada nos trabalhos de Ajdari et al. (2012), Mohamed et al., (2013), Olusola, Akindele e Abiodun (2014) utilizando as seguintes funções: *(i) lin*ear, *(ii) Theshold linear* e *(iii) Bipolar linear*.

Para a camada escondida e saída, a maioria dos trabalhos utilizou a tangente hiperbólica e a função linear, respectivamente. Já Olusola, Akindele e Abiodun (2014) e Lin et al. (2016), utilizarem as funções sigmóide/tangente hiperbólica (2 modelos RNA) e sigmóide para a camada de saída, respectivamente. Enquanto a função sigmóide apresenta valores de ativação dentro do intervalo (0,1), a função tangente hiperbólica assume valores positivos e negativos no intervalo (-1,1).

3.3.4 Taxa de aprendizagem, momento e número de épocas/iterações

Quando a taxa de aprendizagem (TA) tem um valor pequeno, tem-se menor variação dos parâmetros livres de uma iteração para outra e, consequentemente a aprendizagem é lenta. Já para um grande valor da TA essa variação é grande, com aprendizado acelerado e a geração de uma rede instável devido à possível não detecção dos mínimos locais na superfície de erro. Alternativamente, pode-se incluir o termo momento (M), que além de ajudar no treinamento, contribui para que o processo de aprendizagem não termine em um mínimo local para o erro.

Baskar e Sahadevan (2012), Kana et al. (2012) e Ajdari et al. (2012) apresentaram os valores de TA (0,6 e 0,15 e 0,01, respectivamente) e M (0,8) utilizados para a obtenção da RNA. Já Sampaio et al. (2016) descreveram apenas que o valor da TA foi igual a 0,01. Alternativamente, Esfahanian et al. (2013) e Mohamed et al. (2013) relataram esses e outros

parâmetros para obtenção da RNA como *default* segundo os respectivos software utilizados. Vale ressaltar que geralmente esses parâmetros se ajustam quando o número de iterações aumenta, sendo que os valores iniciais mais utilizados para eles encontram-se entre 0 e 1.

Época é a apresentação completa do conjunto de treinamento, compreendendo um passo *forward* e um *backward*. Uma rede é treinada de época em época, até que os pesos sinápticos e os níveis de bias se estabilizam e o erro atinja um valor mínimo. Vale ressaltar que a ordem de apresentação dos tratamentos, em cada época, deve ser aleatória para dar caráter estocástico (variações aleatórias) à busca do menor valor de erro. Já o termo iteração é definido como o número de vezes que todo o conjunto de entrada-saída passa pelos passos *forward*/*backward* segundo o *batch size*. Esse último termo apresentado é um número pré-definido para modelagem que pode ser menor ou igual ao número total de exemplos para treinamento.

Dentre os estudos, Haque et al. (2016a) obtiveram uma RNA eficiente segundo os seus critérios, após seis iterações em duas épocas. Ou seja, o conjunto de todos os exemplos de treinamento passou duas vezes pelos passos *forward*/*backward*, sendo que o *batch size* determinou três iterações para cada época, ou seja, três iterações para que o conjunto total passasse uma vez pelo passo *forward*/*backward*. Para os estudos de Esfahanian et al. (2013), Rahimpour, Hatti-Kaul e Mamo (2016) e Sampaio et al. (2016) o número de épocas foi de 100, 1000 e 5000, respectivamente. Lin et al. (2016), Olusola, Akindele e Abiodun (2014) e Kana et al. (2012) descrevem 10.000, 100.000 e 57,000 iterações, respectivamente. Patil et al. (2017) realizaram 1000 iterações ao avaliar seis funções de transferência e até 10 neurônios na camada escondida.

3.3.5 Critérios de parada

Geralmente, para demonstrar que a rede convergiu, é necessário que existam critérios para encerrar essa operação. Para os estudos que explicitaram essa informação, a decisão quanto à parada utilizou um ou mais dos seguintes critérios relacionados às métricas matemáticas: *(i)* maximizar o R, que deve ser o mais próximo de 1,0 (AJDARI et al., 2012; MOHAMED et al., 2013; KUMAR; CHHABRA; SHUKLA, 2017), *(ii)* maximizar o valor de R^2, que deve ser o mais próximo de 1,0 (AJDARI et al., 2012; ESFAHANIAN et al., 2013; MOHAMED et al., 2013; KUO et al., 2014; OLUSOLA; AKINDELE; ABIODUN, 2014; HUANG et al., 2016; SAMPAIO et al, 2016; VELU; VELAYUTHAM; MANICKKAM, 2016; KUMAR; CHHABRA; SHUKLA, 2017), *(iii)* minimizar o valor de RMSE (*Root Mean Squared Error*), que deve ser mais próximo de zero (AJDARI et al., 2012; KUO et al., 2014;

OLUSOLA; AKINDELE; ABIODUN, 2014; HUANG et al., 2016) e/ou *(iv)* minimizar o valor de MSE (*Mean Squared Error*), que deve ser mais próximo de zero (KANA et al., 2012; ESFAHANIAN et al., 2013; SINGH et al., 2013; SCHUBERT et al., 2015; SAMPAIO et al, 2016; LIN et al., 2016).

Na definição da RNA ideal, que represente o conjunto entrada-saída avaliado, várias arquiteturas são obtidas. Por exemplo, Nelofer et al. (2012) e Mohamed et al. (2013) descrevem que avaliaram 60 e 300 arquiteturas para definirem a melhor RNA, respectivamente. Outra estratégia utilizada para a obtenção de um modelo melhor ajustado é a repetição das topologias. Kuo et al. (2014) e Zhang et al. (2016) realizaram 5 e 10 repetições das topologias para iniciação randômica dos pesos, o que é importante para evitar tendências no treinamento e, consequentemente, uma escolha errada da arquitetura.

3.3.6 Métricas

Métricas são funções matemáticas que são utilizadas para medir a capacidade de erro\acerto para os modelos e podem ser utilizadas no treinamento/teste (número de camadas escondidas e neurônios). Nos diferentes estudos, a decisão quanto ao número de camadas escondidas e/ou neurônios da mesma foi baseada em uma ou mais das seguintes métricas: *(i)* R^2, *(ii)* MSE, *(iii)* RMSE, *(iv)* MRE (Mean Relative Error), e/ou *(v)* ARD (*Average relative deviation*) (ESFAHANIAN et al., 2013; KARTHIC et al., 2013; OLUSOLA; AKINDELE; ABIODUN, 2014; SCHUBERT et al., 2015; HUANG et al., 2016; RAHIMPOUR; HATTI-KAUL; MAMO, 2016; ZHANG et al., 2016). Para as etapas de treinamento e teste, bem como para a validação (próximo item), foram calculados as métricas citadas anteriormente, excetuando MRE e ARD, e incluindo *(vi)* R e/ou *(vii)* ADD (*Absolute Average Derivation*) (AJDARI et al., 2012; HAQUE et al., 2016a; ZHANG et al., 2016; KUMAR; CHHABRA; SHUKLA, 2017). Alguns estudos utilizaram ainda as métricas aqui citadas nos critérios de parada, mesmo que não explicitando essa informação.

3.4 Avaliação do modelo RNA

3.4.1 Etapa de validação com dados desconhecidos

A etapa de teste, apresentada anteriormente, utilizando um conjunto desconhecido, diferente do conjunto de treinamento, serve para avaliar a capacidade de generalização da rede. Porém, ela ainda é utilizada para a tomada de decisão quanto ao melhor treinamento.

Assim, é necessária uma etapa extra de validação para comprovar a capacidade de generalização a partir dados verdadeiramente desconhecidos, que será realizada apenas após o treinamento/teste.

Dentre os estudos avaliados, Mohamed et al. (2013) e Sampaio et al. (2017) realizaram validação com experimentos desconhecidos utilizando dez e três experimentos, respectivamente. Já Ajdari et al. (2012) consideraram quatro testes experimentais para avaliar a predição das entradas ótimas para a resposta como teste de validação. Ou seja, validando a otimização eles consideraram que estavam validando a capacidade de generalização da rede.

3.4.2 Contribuição das variáveis e otimização

A definição da importância das variáveis independentes para a resposta contribuem para o entendimento sobre a influência das mesmas sobre a resposta. Ajdari et al.(2012), Baskar e Sahadevan (2012), Kana et al. (2012), Nelofer et al. (2012), Mohamed et al. (2013) e Zhang et al. (2016) definiram em seus estudos a importância da contribuição de cada entrada nas respectivas respostas avaliadas. Uma possibilidade de equação para cumprir esse objetivo é apresentada no trabalho de Garson (1991).

Para a otimização, a maioria dos trabalhos utilizou um algoritmo genético, respeitando os limites de variação das variáveis independentes utilizados no MSR para a busca. De forma alternativa, Patil et al. (2016 e 2017) apenas avaliaram o(s) valor(s) da(s) resposta(s) dos tratamentos de entrada e/ou validação com dados desconhecidos para definirem os tratamentos ótimos. Já Nelofer et al. (2012) avaliaram os gráficos 3D gerados para o modelo RNA, realizando apenas uma otimização gráfica. Ainda, Kuo et al. (2014) e Huang et al. (2016) utilizaram a metodologia RMA (*Ridge Max Analysis*) que foi descrita na otimização do modelo MSR (item 2.4). Mohamed et al. (2013) utilizaram um método denominado *Rotation Inherit Optimization* (RIO), que segundo os autores, é um algoritmo evolucionário muito semelhante ao algoritmo genético ou enxame de partículas, porém com uma convergência mais rápida, dispensando a definição dos parâmetros pelo experimentador, excetuando-se o tamanho da população.

Nem sempre a otimização do modelo RNA foi realizada nos estudos avaliados. Singh et al. (2013), Meena et al. (2014a), Rafigh et al. (2014), Rahimpour, Hatti-Kaul e Mamo (2016) e Sampaio et al. (2016) não apresentaram resultados para otimização do modelo RNA. Já Esfahanian et al. (2012), Karthic et al.(2013), Velu, Velayutham e Manickkam (2016) e Zhang et al. (2016) apenas avaliaram e validaram experimentalmente a predição do modelo RNA para as entradas otimizadas pelo modelo MSR.

Para a validação das informações de entrada-saída da otimização, geralmente foram realizados experimentos em triplicata. Por sua vez, a capacidade preditiva dessas otimizações foi avaliada por análises do tipo valores da(s) resposta(s) predita(s) *versus* valores experimentais e/ou utilizando muitas das métricas já citadas anteriormente. Como alternativa, Schubert et al. (2015) e Sampaio et al. (2017) validaram as informações da otimização da RNA utilizando resultados de trabalhos publicados na literatura.

3.5 Software utilizados para modelagem e otimização

Para obtenção e análises das RNA, incluindo a otimização, os software mais utilizados foram o *Matlab* (MathWorks) e o *neuralPower* (CPC-X Software). Poucos estudos utilizaram os softwares *WEKA* (*University of Waikato*), *Statistica* (StatSoft), *easyNN* (Neural Planner Software) e *elite-ANN*. Para a otimização do modelo RNA, Sampaio et al. (2016, 2017) foram os únicos a utilizarem o *Framework* MOEA.

4 COMPARAÇÃO ENTRE MSR E RNA

Para comprar as duas metodologias de modelagem, grande parte dos estudos apresentados utilizaram um ou mais dos seguintes critérios: *(i)* Capacidade preditiva para os dados experimentais do delineamento (AJDARI et al., 2012; BASKAR e SAHADEVAN, 2012; NELOFER et al., 2012; ESFAHANIAN et al., 2013; KARTHIC et al., 2013; MOHAMED et al., 2013; REKHA et al., 2013; SINGH et al., 2013; PATHAK et al., 2015; VELU; VELAYUTHAM; MANICKKAM, 2016; HUANG et al., 2016; PATIL et al., 2016; ZHANG et al., 2016; PATIL et al., 2017; SAMPAIO et al., 2017), *(ii)* Capacidade preditiva para dados experimentais desconhecidos (RAFIGH et al., 2014; OLUSOLA; AKINDELE; ABIODUN, 2014; HUANG et al., 2016; LIN et al., 2016; PATIL et al., 2016; SAMPAIO et al., 2017), *(iii)* Capacidade preditiva para validação experimental relacionada à otimização (BASKAR e SAHADEVAN, 2012; KANA et al., 2012; NELOFER et al., 2012; ZAFAR et al., 2012; ESFAHANIAN et al., 2013; KARTHIC et al., 2013; MOHAMED et al., 2013; KUO et al., 2014; PATHAK et al., 2015; HUANG et al., 2016; KUMAR; CHHABRA; SHUKLA, 2017), *(iv)* Avaliação de *stakeholders* (HAQUE et al., 2016a) *e (v)* Avaliação econômica para os dados da otimização (HAQUE et al., 2016a).

Para realizar a comparação segundo o critério *(i)*, Huang et al. (2016*)* e Sampaio et al. (2016, 2017) avaliaram graficamente a variação do resíduo calculado entre o predito e os dados experimentais de entrada. Considerando o critério *(ii)*, Lian et al. (2016) e Huang et al. (2016) utilizaram 12 e 5 experimentos desconhecidos, respectivamente. Segundo o critério *(iii)*, Kumar, Chhabra e Shukla (2017) concluíram que a metodologia híbrida MSR-AG retornou um maior valor da resposta quando comparada à otimização numérica do modelo MSR utilizando o software *Design Expert*. Por fim, segundo os critérios *(iv)* e *(v)*, Haque et al. (2016a) avaliaram que alguns termos (variáveis independentes) não significativos estatisticamente após avaliação do modelo MSR, seriam fundamentais para o processo de lise celular para extração de colesterol oxidase, considerando a experiência com essa prática. Posteriormente, estes autores realizaram uma análise econômica para o processo e concluíram que a otimização RNA-AG resultou em um aumento de 3,7 vezes na quantidade de enzima por dólar americano/dia quando comparada a otimização do modelo MSR. Para cumprimento do critério *(iii)* o ideal seria a utilização das mesmas metodologias de otimização para os dois modelos. Dentre os estudos, Zafar et al. (2012), Kumar, Chhabra e Shukla (2017) e Sampaio et al. (2017) utilizaram uma metodologia híbrida MSR-AG e RNA-AG para a etapa de otimização.

Para a maioria das comparações há sugestão de que RNA foi mais robusta quando comparada a MSR. Já alguns trabalhos, com diferentes objetivos experimentais, apresentaram outras conclusões: *(i)* Mohamed et al. (2013) e Ajdari et al. (2012) concluíram que, apesar da maior precisão da metodologia RNA, elas se complementam na interpretação dos resultados, *(ii)* Meena et al. (2014a), Zhang et al. (2016), Kumar, Chhabra e Shukla (2017) e Murakar e Shastri (2017) concluíram que as duas metodologias foram mutuamente competentes, *(iii)* Patil et al. (2016) concluíram que a melhor metodologia de modelagem variou segundo a resposta avaliada para um único problema experimental e, por fim, *(iv)* Sampaio et al. (2016, 217) descrevem que MSR foi superior a RNA, apesar dos valores das métricas para a última modelagem serem melhores. Vale ressaltar que, de forma única em seu estudo comparativo, Meena et al. (2014a) obtiveram um modelo RNA a partir de informações experimentais para a bactéria *B. bifidum* (MEENA et al., 2011) e o comparam ao modelo MSR obtido para *L. acidophillus* (MEENA et al., 2014b), avaliando a capacidade de predição desses modelos para experimentos de cultivo do consórcio *B. bifidum+L. acidophillus* (MEENA et al., 2014a).

Para a comparação das duas metodologias são utilizadas uma ou mais das métricas citadas anteriormente na etapa de treinamento/teste, excetuando-se o R e incluindo *(i)* R^2_{Aj}, *(ii)* SEP (*Standard Error of Prediction*), *(iii)* MRPE (*Mean Relative Percentage Error), (iv)* MAPE (*Mean Absolute Percentage Error), (v)* MPE *(Mean Percentage Error), (vi)* RPD (*Relative Percente Derivation*), *(vii)* PRE (*Proportion of the relative error*), *(viii)* B_f (Bias Factor) e/ou *(ix)* A_f (*Accuracy Factor*). Segundo Schubert et al. (2015), algumas métricas que são calculadas segunda a razão entre do valor observado e predito não permitem uma boa avaliação quando, por exemplo, é obtido um valor experimental diferente de zero para um tratamento e o modelo prediz uma resposta igual ou próxima à zero, ou vice-versa. Assim, o cálculo do PRE é uma alternativa já que para calculá-lo é providenciado um valor igual a menos 1 (-1) quando valores experimentais e/ou preditivos são zero ou próximos de zero.

5 MODELAGEM RNA PARA PEQUENO CONJUNTO DE DADOS

A modelagem RNA tem se mostrado eficiente em aplicações em várias áreas como química, finanças, física e ciências biológicas, quando modelos mecanísticos não podem ser ajustados e, principalmente, ao descrever relações não lineares entre variáveis independentes e resposta. Porém, alguns desafios surgem diante das aplicações já realizadas.

Sabe-se que a modelagem RNA depende da qualidade e quantidade de dados disponíveis, geralmente sendo necessário um grande conjunto de dados. Entretanto, os trabalhos avaliados utilizaram um pequeno conjunto de entrada-saída. Quando a rede aprende um número pequeno de exemplos corre-se o risco da apresentação excessiva com memorização dos dados de treinamento, o que leva a rede à perda da capacidade de generalizar com outros exemplos similares. Entretanto, mesmo utilizando pequenos conjuntos de dados, alguns dos trabalhos avaliados realizaram a importante etapa de validação com dados desconhecidos, obtendo resultados idênticos ou melhores que os obtidos com a modelagem MSR. Mas, de fato, a modelagem/otimização utilizando uma RNA treinada com um pequeno conjunto de entrada-saída é confiável em suas predições?

A maioria dos estudos demonstrou que a otimização do modelo RNA previu valores maiores para as respostas, que foram validados experimentalmente. Porém, caso um número bem maior de exemplos entrada-saída, para experimentos realizados nas mesmas condições, fosse utilizado, a otimização desse modelo RNA retornaria um valor ainda melhor? Realizando uma pesquisa, até onde sabemos, nenhum resultado foi encontrado na literatura que pudesse responder essa questão. Assim, faz-se necessária a realização de experimentos comparativos entre modelos RNA obtidos a partir de pequenos e grandes números de exemplos, para diferentes áreas experimentais.

Por fim, uma das vantagens de utilização de um modelo RNA é a possibilidade de extrapolação do aprendizado. As RNA possuem habilidade de produzir respostas para conjuntos entrada-saída desconhecidos, o que as tornaria capazes de extrapolar e/ou interpolar conhecimento. Porém, as modelagens RNA realizadas não foram avaliadas quanto à capacidade de extrapolação, ou seja, os estudos se concentraram apenas na interpolação para encontrar as regiões de ótimos resultados. Sendo assim, outro questionamento pode ser apresentado: Será que uma RNA treinada a partir de um conjunto grande de exemplos permitirá uma extrapolação confiável?

6 CONCLUSÃO

Esta revisão realizou um levantamento sobre artigos recentes que avaliaram a modelagem/otimização das condições de cultura para diferentes micro-organismos, bem como a modelagem/otimização de processos utilizando diferentes enzimas, aplicando as ferramentas MSR e RNA. A partir dos resultados dos estudos, do levantamento da literatura pertinente e, por fim, da experiência dos autores, foi possível apresentar um guia de boas práticas para as metodologias de modelagem citadas. Da análise comparativa foi possível concluir que para a maioria dos trabalhos a modelagem/otimização utilizando RNA foi superior à MSR. Porém, ainda é necessária a realização de estudos, nas diversas áreas de pesquisas, que comparem a modelagem RNA utilizando pequenos e grandes conjuntos de entrada-saída para as etapas de treinamento, teste, validação e otimização.

REFERÊNCIAS

AJDARI, Z. et al. A statistical modeling study by response surface methodology and artificial neural networks on medium optimization for *Monascus purpureus* FTC5391 sporulation **Minerva Biotecnologica**, v. 24, p. 71-81, 2012.

BASKAR, G.; SAHADEVAN, R. Optimization of culture conditions and bench-scale production of L-Aasparaginase by submerged fermentation of *Aspergillus terreus* MTCC 1782. **Journal of Microbiology and Biotechnology**, v. 22, n. 7, p. 923-929, 2012.

BOX, G. E. P.; BEHNKEN, D. W. Some new three level designs for the study of quantitative variables. **Technometrics**, v. 2, p. 455-475, 1960.

BRAGA, A. P.; DE CARVALHO, A. P. L. F.; LUDEMIR, T. B. **Redes neurais artificiais teoria e aplicações**. Rio de Janeiro: LTC, 2000. 262 p.

CHEN G.; BUI, T. D.; KRZYZAK, A. Denoising of three dimensional data cube using bivariate wavelet shrinking. In: CAMPILHO A.; KAMEL M. (Eds). **Image analysis and recognition**. ICIAR 2010. Lecture notes in computer science, vol 6111. Berlin, Heidelberg: Springer, 2010.

CHOJACZYK, A. A. et al. Review and application of artificial neural networks models in reliability analysis of steel structures. **Structural Safety**, v. 52, p. 78-89, 2015.

DESAI, K. M. et al. Comparison of artificial neural network (ANN) and response surface methodology (RSM) in fermentation media optimization: case study of fermentative production of scleroglucan. **Biochemical Engineering Journal**, v. 41, n. 3, p. 266-273, 2008.

ESFAHANIAN, M. et al. Modeling and optimization of ethanol fermentation using *Saccharomyces cerevisiae*: response surface methodology and artificial neural network. **Chemical Industry & Chemical Engineering Quarterly**, v. 19, n. 2, p. 241-252, 2013.

FLOOD, I.; ISSA, R. A. Empirical modeling methodologies for construction. **Journal of Construction Engineering and Management**, v. 136, n. 1, p. 36-48, 2010.

GARSON, D. G. **Interpreting neural network connection weights**. AI. Expert, v. 6, p. 47-51, 1991.

HAQUE, S. et al. (2016). Modeling and optimization of a continuous bead milling process for bacterial cell lysis using response surface methodology. **RSC AdvANCES**, v. 6 , p. 16348-16357, 2016b.

HAQUE, S. et al. Artificial intelligence vs. statistical modeling and optimization of continuous bead milling process for bacterial cell lysis. **Frontiers in Microbiology**, v. 7, Article 1852, 2016a.

HAYKIN, S. **Redes neurais: princípios e prática.** 2. ed. Porto Alegre: Bookman, 2002. 900 p.

HUANG, K-C. et al. Highly efficient synthesis of an emerging lipophilic antioxidant: 2-ethylhexyl ferulate. **Molecules**, v. 21, n. 4, 478, 2016.

JANG, J. R. S.; SUN, C. T.; MIZUTANI, E. **Neuro fuzzy e soft computing**. New Jersey: Prentice Hall., 1997. 129 p.

KANA, E. B. g. et al. Comparative evaluation of artificial neural network coupled genetic algorithm and response surface methodology for modeling and optimization of citric acid production by *Aspergillus Niger* MCBN297. **Chemical Engineering Transactions**, v. 27, p. 397-402, 2012.

KARTHIC, P. et al. Optimization of biohydrogen production by *Enterobacter* species using artificial neural network and response surface methodology. **Journal of Renewable and Sustainable Energy**, v. 5, 033104, 2013.

KUMAR, V.; CHHABRA, D.; SHUKLA, P. Xylanase production from *Thermomyces lanuginosus* VAPS-24 using low cost agro-industrial residues via hybrid optimization tools and its potential use for saccharification. **Bioresource Technology**, v. 243, p. 1009-1019, 2017.

KUO, C-H. et al. Response surface methodology and artificial neural network optimized synthesis of enzymatic 2-phenylethyl acetate in a solvent-free system. **Biocatalysis and Agricultural Biotechnology**, v. 3, p. 1-6, 2014.

LIN, J-A. et al. A novel enzyme-assisted ultrasonic approach for highly efficient extraction of resveratrol from *Polygonum cuspidatum*. **Ultrasonics Sonochemistry**, v. 32, p. 258-264, 2016.

MEENA et al. Growth characteristics modeling of *Bifidobacterium bifidum* using RSM and ANN. **Brazilian Archives of Biology and Technology**, v. 54, p. 1357-1366, 2011.

MEENA et al. Growth characteristics modeling of *Lactobacillus acidophillus* using RSM and ANN. **Brazilian Archives of Biology and Technology**, v. 57, p. 15-22, 2014b.

MEENA, G. S. et al. Growth characteristics modeling of mixed culture of *Bifidobacterium bifidum* and *Lactobacillus acidophilus* using response surface methodology and artificial neural network. **Brazilian Archives of Biology and Technology**, v. 57, n. 6, p. 962-970, 2014a.

MICHALEWICZ, Z. **Genetic Algorithms+Data Structures=Evolution Programs**. Berlin: Springer-Verlag, 1992.

MOHAMED, M. S. et al. Comparative analyses of response surface methodology and artificial neural network on medium optimization for *Tetraselmis* sp. FTC209 grown under mixotrophic condition. **The Scientific World Journal**, v. 2013, ID 948940, 2013.

MONTGOMERY, D. C. **Design and analysis of experiments**. 5. ed. New York: John Wiley and Sons, 2001.

MURARKAR, K.; SHASTRI, P. Artificial neural network (ANN) and response surface methodology (RSM) of extraction of pectin from sweet lemon peels by microbial protopectinase. **International Journal of Recent Trends in Engineering & Research (IJRTER)**, v. 3, n. 10, p. 11, p. 202-206, 2017.

NELOFER R. et al. Sequential optimization of production of a thermostable and organic

solvent tolerant lipase by recombinant *Escherichia coli*. **Annals of Microbiology**, v. 6, n. 3, p. 535-544, 2011.

NELOFER, R. et al. Comparison of the estimation capabilities of response surface methodology and artificial neural network for the optimization of recombinant lipase production by *E. coli* BL21. **Journal of Industrial Microbiology & Biotechnology**, v. 39, p. 243-254, 2012.

OLUSOLA, A.; AKINDELE, O.; ABIODUN, O. Comparative studies of response surface methodology (RSM) and artificial neural network (ANN) predictive capabilities on enzymatic hydrolysis optimization of sweet potato starch. **International Journal of Advanced Research**, v. 2, n. 10, p. 849-860, 2014.

PATHAK, L. et al. Artificial intelligence versus statistical modeling and optimization of cholesterol oxidase production by using *Streptomyces* sp.. **PLoS ONE**, v. 10, n. 9, e0137268, 2017.

PATIL, M. D. et al. Disruption of *Pseudomonas putida* by high pressure homogenization: a comparison of the predictive capacity of three process models for the efficient release of arginine deiminase. **AMB Express**, v. 6, 84, 2016.

PATIL, M. D. et al. Ultrasonic disruption of *Pseudomonas putida* for the release of arginine deiminase: Kinetics and predictive models. **Bioresource Technology**, v. 233, p. 74-83, 2017.

PLACKETT, R. L.; BURMAN, J. P. The design of optimum multifactorial experiments. **Biometrika,** v. 33, p. 305-325, 1946.

RAFIGH, S. M. et al. Optimization of culture medium and modeling of curdlan production from *Paenibacillus polymyxa* by RSM and ANN International. **Journal of Biological Macromolecules**, v. 70, p. 463-473, 2014.

RAHIMPOUR, F.; HATTI-KAUL, R.; MAMO, G. Response surface methodology and artificial neural network modelling of an aqueous two-phase system for purification of a recombinant alkaline active xylanase. **Process Biochemistry**, v. 51, p. 452-462, 2016.

REKHA, V. P. B. et al. Optimization of polygalacturonase production from a newly isolated *Thalassospira frigidphilosprofundus* to use in pectin hydrolysis: statistical approach. **BioMed Research International**, v. 2013, ID 750187, 2013.

RODRIGUES, M. I.; LEMMA, A. F. **Planejamento de experimentos e otimização de processos**. 2. ed. Campinas: Casa do Espírito Amigo Fraternidade Fé e Amor, 2009. 358p.

SAMPAIO, F. C. et al. Batch growth of *Kluyveromyces lactis* cells from deproteinized whey: response surface methodology versus Artificial neural network—genetic algorithm approach. **Biochemical Engineering Journal**, v. 109, p. 305-311, 2016.

SAMPAIO, F. C. et al. Comparison of response surface methodology and artificial neural network for modeling xylose-to-xylitol bioconversion. **Chemical Engineering and Technology**, v. 40, n. 1, p. 122-129, 2017.

SCHUBERT, M. et al. Prediction and optimization of the laccase-mediated synthesis of the antimicrobial compound iodine (I_2). **Journal of Biotechnology**, v. 193, p. 134-136, 2015.

SINGH, P. et al. Optimization of cultural conditions using response surface methodology versus artificial neural network and modeling of L-glutaminase production by *Bacillus cereus* MTCC 1305. **Bioresource Technology**, v. 137, p. 261-269, 2013.

SNEDECOR, G. W.; COCHRAN, W. G. **Statistical Methods**. 6. ed. Ames: Iowa State University Press, 1967.

THOMPSON, J. R. **Empirical Model Building: Data, Models, and Reality**. 2. ed. New Jersey: John Wiley & Sons, 2011. 464 p.

VELU, S.; VELAYUTHAM, V.; MANICKKAM, S. Optimization of fermentation media for xanthan gum production from *Xanthomonas campestris* using response surface methodology and artificial neural network techniques. **Indian Journal of Chemical Technology**, v. 23, p. 353-361, 2016.

WITEK-KROWIAK, A. et al. Application of response surface methodology and artificial neural network methods in modelling and optimization of biosorption process. **Bioresource**

Technology, v. 160, p. 150–160, 2014.

ZAFAR, M. et al. Artificial intelligence based modeling and optimization of poly(3-hydroxybutyrate-co-3-hydroxyvalerate) production process by using *Azohydromonas lata* MTCC 2311 from cane molasses supplemented with volatile fatty acids: A genetic algorithm paradigm. **Bioresource Technology**, v. 104, p. 631-641, 2012.

ZHANG, K. et al. Modeling and optimization of Newfoundland shrimp waste hydrolysis for microbial growth using response surface methodology and artificial neural networks. **Marine Pollution Bulletin**, v. 109, p. 245-252, 2016.