

**UNIVERSIDADE FEDERAL DE CIÊNCIAS DA SAÚDE DE PORTO ALEGRE
INFORMÁTICA BIOMÉDICA**

FLAVIELLE BLANCO MARQUES

**CLASSIFICAÇÃO DE TEXTO ASSOCIADO À RESISTÊNCIA ANTIMICROBIANA
COM REDES NEURAIS RECORRENTES BIDIRECIONAIS**

**PORTO ALEGRE
2018**

Flavielle Blanco Marques

Classificação de Texto associado à Resistência Antimicrobiana com Redes Neurais Recorrentes
Bidirecionais

Trabalho de Conclusão de Curso de graduação
apresentado ao Curso de Informática Biomédica da
Universidade Federal de Ciências da Saúde de Porto
Alegre, como requisito parcial para a obtenção do
grau de Bacharel em Informática Biomédica.

Orientadora: Prof^a. Dra. Ana Trindade Winck

Porto Alegre
2018

Classificação de Texto associado à Resistência Antimicrobiana com Redes Neurais Recorrentes
Bidirecionais

Trabalho de Conclusão de Curso de graduação
apresentado ao Curso de Informática Biomédica da
Universidade Federal de Ciências da Saúde de Porto
Alegre, como requisito parcial para a obtenção do
grau de Bacharel em Informática Biomédica.

Orientadora: Prof^a. Dra. Ana Trindade Winck

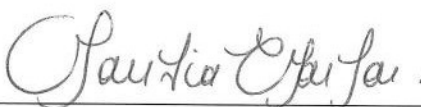
Data de Aprovação: 11/12/2018

Banca Examinadora



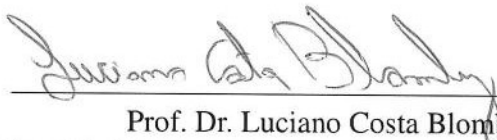
Prof^a. Dra. Ana Trindade Winck

Universidade Federal de Ciências da Saúde de Porto Alegre
Departamento de Ciências Exatas e Sociais Aplicadas
Orientadora



Prof^a. Dra. Claudia Elizabeth Thompson

Universidade Federal de Ciências da Saúde de Porto Alegre
Departamento de Farmacociências
Examinadora



Prof. Dr. Luciano Costa Blomberg

Universidade Federal de Ciências da Saúde de Porto Alegre
Departamento de Ciências Exatas e Sociais Aplicadas
Examinador

AGRADECIMENTOS

É difícil agradecer todas as pessoas que de algum modo, nos momentos tranquilos ou estressantes, fizeram ou fazem parte da minha vida.

Primeiramente, sou grata a todos os professores que contribuíram com a minha trajetória acadêmica, especialmente a professora Dra. Ana Trindade Winck, minha orientadora no Trabalho de Conclusão de Curso. Obrigada pelos conselhos, esclarecimento de dúvidas, pela atenção, paciência e ajuda. Obrigada por me incentivar sempre pela busca do conhecimento e despertar o meu interesse na pesquisa.

Agradeço aos meus pais Nara Ione Blanco Marques e Flavio Medeiros Marques, que sempre me deram apoio, incentivo e amor. Obrigada, Allan Blanco Marques, irmão querido, por ser tão paciente e companheiro. Agradeço ao meu namorado Lincon Lima dos Reis, que jamais me negou apoio, carinho e incentivo. Obrigada, por aguentar tantas crises. Sem você do meu lado esse trabalho não seria possível.

Sou grata também a todos os meus amigos e colegas que tive o prazer de conhecer, especialmente, Eduardo Henrique Paes Pooch e Meiski Mariá Vedovatto, que além de compartilhar a maioria dos momentos e recordações da graduação, hoje, posso chamá-los de colegas de profissão.

Deixo aqui um agradecimento especial aos meus avós, que sempre foram um exemplo de luta e determinação e aos meus tios, especialmente, às minhas tias, Lucinara Blanco e Ana Lucia Blanco Lopes. Eu jamais serei capaz de retribuir todo carinho e amor que recebi de vocês.

Ao meu avô Osvaldo Blanco (*in memoriam*) que não pode estar presente neste momento tão incrível da minha vida. Seus ensinamentos e valores alimentaram minha vida e conduziram meus passos até aqui. Saudades eternas!

Agradeço a minha chefe/orientadora Lissandra Souto Cavalli que me deu a oportunidade de realizar estágio e iniciação científica. Obrigada por me ensinar na prática todo o conhecimento que adquiri durante a minha graduação e por todas as oportunidades de trabalho e crescimento profissional.

Agradeço a minha filha de pelos, Kelly, fiel e companheira, que sempre me ouve e me entende.

Por fim, agradeço a Universidade Federal de Ciências da Saúde de Porto Alegre que me deu a oportunidade de cursar Informática Biomédica e desbravar as fantásticas áreas que a minha profissão carrega.

RESUMO

O surgimento acelerado de microrganismos resistentes coloca em risco a eficácia dos agentes antimicrobianos disponíveis. A falta de informações sobre a resistência que os microorganismos adquirem dificulta o combate aos responsáveis pelas infecções e, consequentemente, a descoberta de possíveis curas. Em busca de fortalecer a base de conhecimento e evidências relacionadas à resistência antimicrobiana, por meio da pesquisa, um método para a classificação de sentenças oriundas de artigos científicos que remetem a termos biológicos de microrganismos ou antimicrobianos foi proposto. Muitos estudos aplicados à aprendizagem de máquina no contexto biológico estão sendo realizados. Em contrapartida, há poucos trabalhos, deste porte, associados à resistência antimicrobiana por meio de textos. Nesse sentido, é apresentado um método de classificação de sentenças de resistência antimicrobiana, com uso de modelos de redes neurais recorrentes bidirecionais, gerados a partir de artigos científicos oriundos de repositórios relacionados às ciências da saúde. Testes e análises do algoritmo construído foram realizadas para encontrar o melhor modelo. A construção de uma ferramenta *web* e integrada a um banco de dados para coleta de novas sentenças foi realizada. A aprendizagem por contexto foi evidenciada por meio de validações com sentenças que eram desconhecidas pela rede.

Palavras-chave: Resistência Antimicrobiana. Agentes Antimicrobianos. Microrganismo. Redes Neurais Recorrentes. Classificação de Texto.

ABSTRACT

The accelerated emergence of resistant microorganisms puts the efficacy of antimicrobial agents available at risk. The lack of information about acquire resistance hinders the combat of organisms responsible for infections and, therefore, to discover possible cures. Seeking to strengthen the knowledge base and evidence related to antimicrobial resistance. We propose a method for the classification of sentences from scientific articles that refer to biological terms of microorganisms or antimicrobials. Many studies applied to machine learning in the biological context are being carried out. Nevertheless, there are few studies of this proportion, associated with antimicrobial resistance through texts. In this sense, a method of classification of sentences of antimicrobial resistance is presented, through bidirectional recurrent neural network models, generated from scientific articles from health science repositories. Tests and analyzes of the algorithm built were performed to find the best model. The construction of a web tool and integrated with a database for the collection of new sentences was accomplished. Contextualized learning was evidenced through validations with sentences that were unknown by the network.

Keywords: Antimicrobial Resistance. Antimicrobial Agents. Microorganism. Recurrent Neural Networks. Text Classification.

LISTA DE FIGURAS

Figura 1 – Tranferência por Conjugação	12
Figura 2 – Tranferência por Transdução	13
Figura 3 – Mecanismos de Resistência	14
Figura 4 – Modelo matemático simples de um neurônio	18
Figura 5 – Estrutura de uma BRNN utilizando três camadas	20
Figura 6 – RNN convencional x RNN LSTM	21

LISTA DE TABELAS

Tabela 1 – Matriz de confusão. Cada linha representa o valor verdadeiro da classe e cada coluna representa o valor predito para a classe.	17
Tabela 2 – Comparação entre os Trabalhos	27

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
BLSTM	<i>Bidirectional Long Short Term Memory</i>
BRNN	Redes Neurais Recorrentes Bidirecionais
CDC	Centro de Controle e Prevenção de Doenças
DNA	<i>Deoxyribonucleic Acid</i>
FN	Falsos Negativos
FP	Falsos Positivos
IA	Inteligência Artificial
IRAS	Infecções Relacionadas à Assistência à Saúde
LSTM	<i>Long Short Term Memory</i>
OMS	Organização Mundial da Saúde
ONU	Organização das Nações Unidas
PLN	Processamento de Linguagem Natural
RNA	Redes Neurais Artificiais
RNN	Redes Neurais Recorrentes
SNC	Sistema Nervoso Central
SVM	<i>Support Vector Machine</i>
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos

SUMÁRIO

1	INTRODUÇÃO	11
1.1	RESISTÊNCIA ANTIMICROBIANA	11
1.2	APRENDIZAGEM DE MÁQUINA	16
1.2.1	Aprendizagem Supervisionada	17
1.2.2	Redes Neurais Artificiais	18
1.2.3	Aprendizagem Profunda	19
1.2.3.1	Redes Neurais Recorrentes	19
1.2.3.1.1	<i>Redes de Memória de Longo Prazo</i>	21
1.3	PROCESSAMENTO DE LINGUAGEM NATURAL	23
1.3.1	Classificação de Texto	23
1.4	TRABALHOS RELACIONADOS	24
2	OBJETIVOS	28
2.1	OBJETIVO GERAL	28
2.2	OBJETIVOS ESPECÍFICOS	28
3	ARTIGO CIENTÍFICO	30
4	DISCUSSÃO	31
5	CONCLUSÃO	33
	REFERÊNCIAS	34
	APÊNDICE A – FERRAMENTA WEB DE COLETA DE SENTENÇAS	39
	ANEXO A – IDENTIFICAÇÃO DE RESISTÊNCIA X DESCOBERTA DO ANTI-BIÓTICO	40

1 INTRODUÇÃO

O desenvolvimento da indústria farmacêutica, aliado aos avanços tecnológicos da área, possibilitou a descoberta e desenvolvimento de medicamentos para o tratamento de infecções. Os medicamentos são agentes antimicrobianos capazes de combater diversos microrganismos que podem ser sintetizados quimicamente ou a partir de outros organismos vivos, esses denominados antibióticos. Todos os antibióticos são agentes antimicrobianos, porém, nem todos os agentes antimicrobianos são antibióticos. Sendo assim, muitas vezes, os termos como antibióticos e agentes antimicrobianos são usados, equivocadamente, como sinônimos (ORGANIZATION et al., 2011).

Com o aumento do poder computacional e da quantidade de dados disponível, o desenvolvimento de técnicas mais robustas foram necessárias, como é o caso das redes neurais artificiais, que utilizam métodos capazes de aprender. A aprendizagem profunda ou *deep learning* utiliza diversas camadas de processamento para resolver problemas mais complexos. As redes neurais recorrentes são utilizadas para aprendizagem profunda de dados sequenciais, por exemplo, dados textuais (SUTSKEVER et al., 2011).

O uso de aprendizagem profunda, especificamente redes neurais recorrentes, pode ser utilizado para a classificação de texto em diferentes contextos. Classificar texto levando em consideração o contexto biológico e a resistência antimicrobiana pode auxiliar na descoberta de novos contextos de resistência e futuramente identificar novos agentes etiológicos. Agentes etiológicos são microrganismos e toxinas microbianas que causam doenças (KNUDSEN et al., 1993) e que podem ser fatores desencadeantes para resistência antimicrobiana.

1.1 RESISTÊNCIA ANTIMICROBIANA

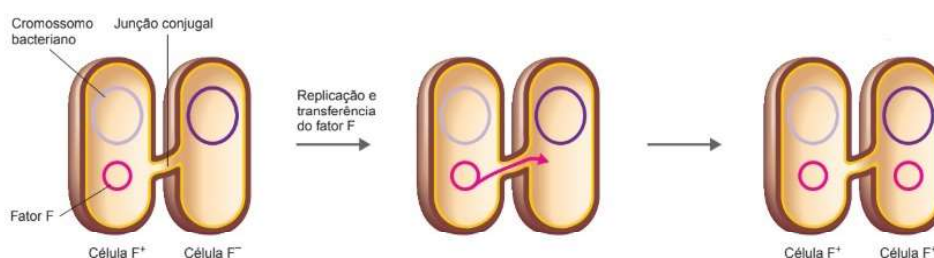
A capacidade de determinado agente causar uma doença que supera as defesas do hospedeiro é chamado de patogenicidade e o grau de patogenicidade do microrganismo é determinado virulência. A exposição de microrganismos a um novo antimicrobiano tende a uma taxa de mortalidade elevada causando a morte da maioria dos microrganismos. Porém, aqueles que sobrevivem, geralmente, possuem características genéticas que possibilitam a adaptação e a sobrevivência (TORTORA et al., 2016).

A resistência aos antimicrobianos está diretamente relacionada à existência de genes contidos no microrganismo que codificam diferentes mecanismos bioquímicos que impedem a ação dos medicamentos (TAVARES, 2000). Além disso, cada vez mais, os microrganismos

compartilham essa resistência uns com os outros, dificultando a manutenção e o desenvolvimento de novos fármacos. A identificação de resistência e o desenvolvimento de novos medicamentos andam juntos, o que dificulta que novos tratamentos sejam disponibilizados. O Anexo A ilustra essa dificuldade, ao apresentar uma linha temporal acerca da identificação da resistência versus a descoberta do antibiótico.

A resistência antimicrobiana ocorre por meio de mutações aleatórias que são espalhadas horizontalmente entre os microrganismos. Os processos de transferência horizontal podem ocorrer por conjugação ou transdução. A conjugação utiliza plasmídeos (fragmento de DNA circular independente do cromossomo celular). Algumas características do processo diferem a conjugação da transdução, como a célula doadora conter o plasmídeo enquanto a célula receptora, em sua maioria, não, além do contato célula-célula exemplificado na Figura 1 (TORTORA et al., 2016).

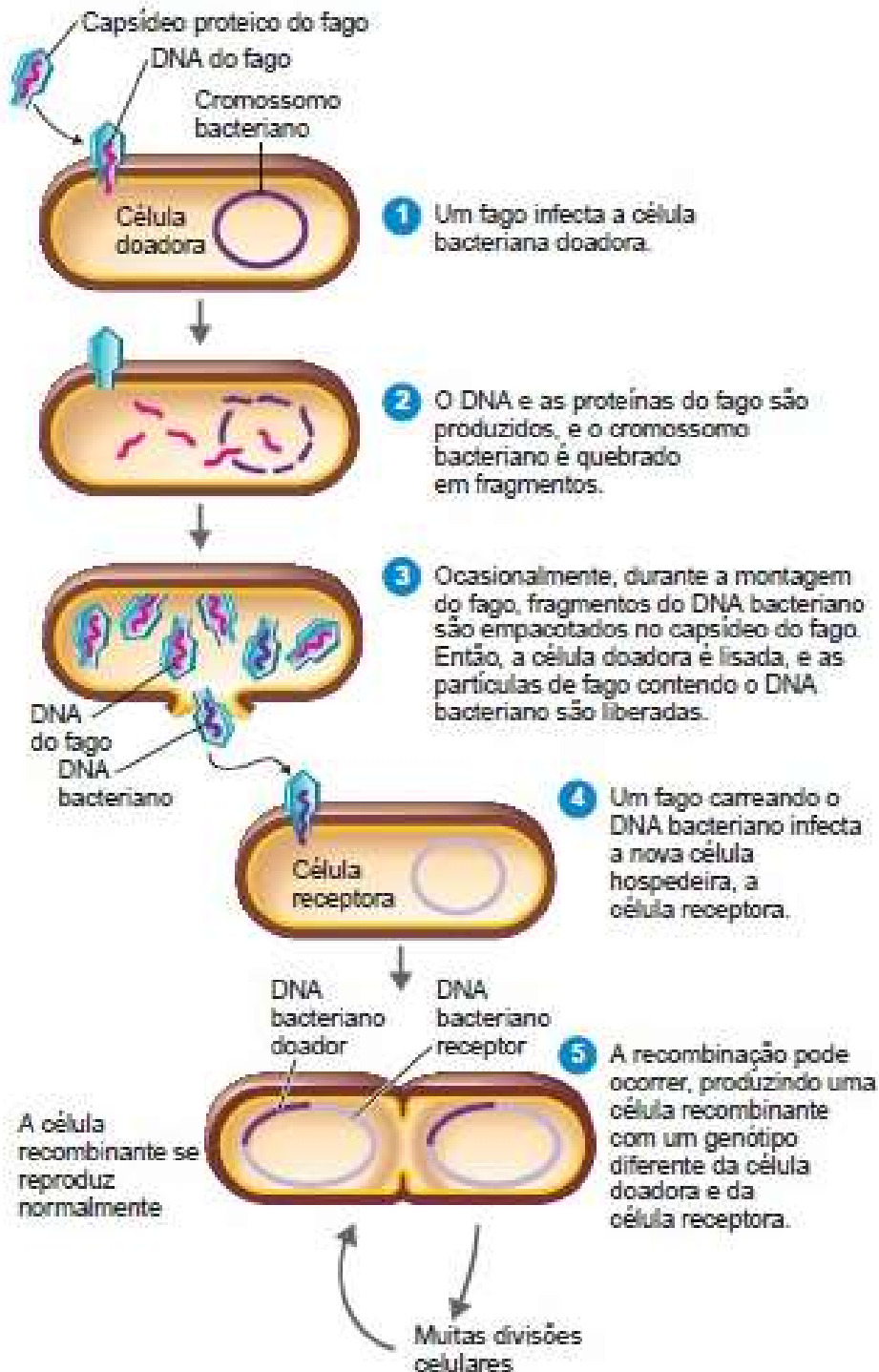
Figura 1 – Transferência por Conjugação



Fonte: (TORTORA et al., 2016)

Na transdução o material genético é transferido por meio de um bacteriófago ou fago (vírus que infecta uma bactéria). O fago infecta a bactéria que produz o DNA e as proteínas do vírus. Em algumas situações, o conteúdo da bactéria é carregado no capsídeo do fago que é liberado para infectar outra bactéria. A bactéria receptora ao ser infectada inclui o material genético em sua estrutura, resultando em uma bactéria com 9 genótipos diferentes. (TORTORA et al., 2016). Esse conceito está ilustrado na Figura 2.

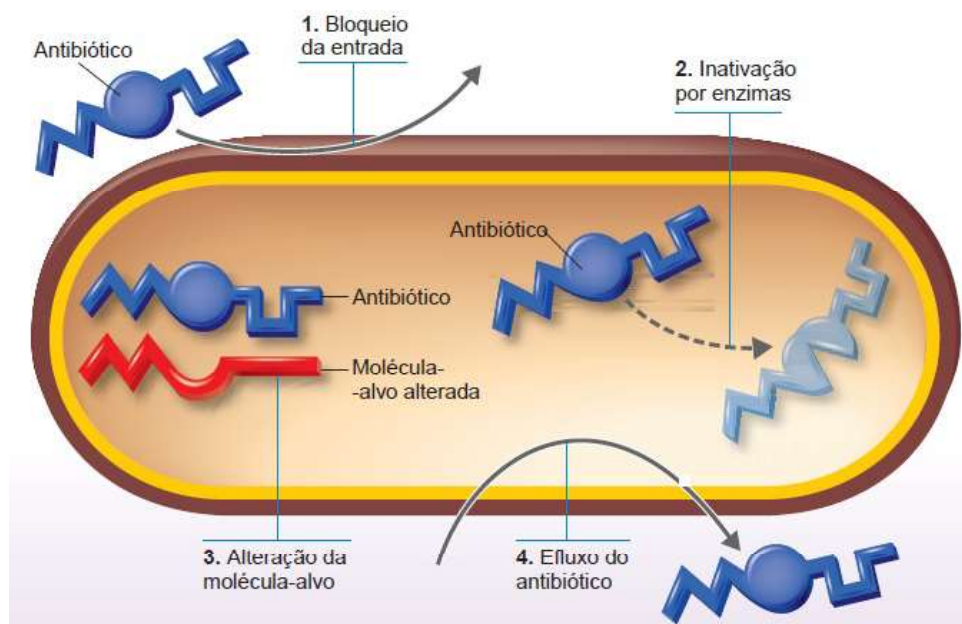
Figura 2 – Transferência por Transdução



Fonte: (TORTORA et al., 2016)

Há alguns mecanismos que tornam as bactérias resistentes, tais como: i) destruição ou inativação enzimática da droga; ii) prevenção da entrada no sítio-alvo; iii) alterações no sítio-alvo da droga; e iv) efluxo rápido (ejeção) do antibiótico, explanados pela Figura 3 (TORTORA et al., 2016).

Figura 3 – Mecanismos de Resistência



Fonte: (TORTORA et al., 2016)

A utilização descontrolada ou inadequada dos antimicrobianos pode acarretar em microrganismos cada vez mais resistentes (HOUSE, 2014). Em 2013, o Centro de Controle e Prevenção de Doenças (CDC) estimou que, pelo menos, 2 milhões de doenças e 23 mil mortes foram causadas por bactérias resistentes à antibióticos, apenas nos Estados Unidos. As infecções causadas por microrganismos resistentes são de difícil tratamento e geram altos custos para os setores públicos e privados, sendo uma ameaça mundial (ANVISA, 2017a). As bactérias ao se alojarem no hospedeiro, podem ser transmitidas para outros indivíduos, favorecendo a disseminação de infecções (ANVISA, 2017b; ECDC, 2018).

Existem algumas formas de contaminação, sendo que uma delas é por meio de infecções hospitalares e/ou relacionadas à assistência à saúde (IRAS), podendo se manifestar durante a internação ou após a alta, levando muitas vezes ao óbito. Conforme o CDC, em um dia cerca de 1 em cada 25 pacientes hospitalizados tem, pelo menos, uma infecção relacionada à assistência à saúde (CDC, 2018).

Em 2015, na 68ª Assembleia Mundial Saúde, foi aprovado o Plano de Ação Global que determinou cinco objetivos a serem alcançados para combater a resistência antimicrobiana (CDC, 2018; ANVISA, 2017b; WHO, 2017):

- Melhorar a conscientização e compreensão do que é a resistência antimicrobiana, por meio da comunicação eficaz, educação e formação;

- Fortalecer a base de conhecimento e evidências por meio da vigilância e pesquisa;
- Diminuir a incidência de infecções por meio de medidas de saneamento, higiene e prevenção;
- Otimizar o uso de antimicrobianos na saúde humana e animal;
- Desenvolver um plano econômico para investimentos sustentáveis em todos os países e aumentar o investimento em novos medicamentos, ferramentas de diagnóstico e vacinas.

O Plano de Ação Global destaca a necessidade da prevenção das infecções de forma segura e acessível. A Organização Mundial da Saúde (OMS) em parceria com os países-membros da Organização das Nações Unidas (ONU), buscam incentivar a criação de planos de controle e combate à resistência antimicrobiana (ANVISA, 2017b). A fim de contornar os problemas econômicos, o apoio técnico, o desenvolvimento de medicamentos e outras formas de intervenção são de extrema importância (WHO, 2017).

As infecções são causadas por meio de diferentes microrganismos que são estranhos ao hospedeiro e que causam doenças, na qual, muitas são provocadas por bactérias. As infecções bacterianas são comuns, a maioria não deixa sequelas, mas muitas podem provocar lesões irreversíveis, recorrência ou colocar em risco a vida do ser humano, seja a curto ou longo prazo. Além disso, a resistência antimicrobiana favorece que as infecções sejam cada vez mais difíceis de tratar (SCULLY, 2014).

Entre as diversas infecções que causam graves problemas de saúde estão as infecções do Sistema Nervoso Central (SNC), que possuem uma baixa taxa de identificação do agente etiológico e serviram como tema motivador do presente estudo. Em algumas dessas infecções, em mais de 50% dos casos, nos Estados Unidos, não há descrição etiológica (GLASER et al., 2006). Algumas infecções do SNC são comuns na população, como por exemplo meningites e encefalites, porém é estimado que os agentes etiológicos sejam reconhecidos em apenas um terço dessas doenças, inclusive considerando o melhor cenário laboratorial (ROTBART, 2000).

Apesar das infecções do SNC ocasionarem graves sequelas nos indivíduos (ERDEM et al., 2017), há muitos casos de infecções ainda mais difíceis de serem tratadas como, por exemplo, aquelas que acometem pacientes que utilizam implantes neurocirúrgicos, pacientes imunodeprimidos ou pacientes infectados após submetidos a uma cirurgia cerebral. Em alguns desses casos, bactérias resistentes já foram encontradas (BEEK et al., 2016). A falta de informações dificulta o combate a organismos responsáveis pelas infecções e, conseqüentemente, a

descoberta de possíveis curas. Nessas situações é necessário buscar, por meio da pesquisa, a identificação dos agentes etiológicos a fim de construir um protocolo que considere a resistência aos antimicrobianos (TSIOUTIS et al., 2017).

Há diversas pesquisas sendo realizadas com cepas de diferentes microrganismos que apresentam resistência (LEAVIS et al., 2007; YIN et al., 2017; SI-TUAN et al., 2017). A publicação de artigos científicos que abordam esse tema cresce cada vez mais. Sendo assim, a dificuldade de acompanhar as informações relevantes sobre o assunto se torna maior. Como apoio a esses estudos, algumas técnicas computacionais estão sendo utilizadas para a construção de bases de dados de resistência antimicrobiana (MCARTHUR et al., 2013; LIU; POP, 2008; WANG et al., 2015a; WAGHU et al., 2015).

1.2 APRENDIZAGEM DE MÁQUINA

A Inteligência Artificial (IA) é um dos campos mais recentes em ciências e engenharia (NORVIG; RUSSELL, 2014). A IA abrange diferentes subáreas, relacionadas a processos de pensamento e raciocínio ou a processos comportamentais, podendo ser comparada ao pensamento humano. O uso da inteligência artificial, na atualidade, está mais relacionado à aplicação de técnicas já existentes do que na proposta de novas teorias.

A aprendizagem de máquina (AM) é um ramo da inteligência artificial que busca resolver problemas computacionalmente utilizando algoritmos. Algoritmo é definido como uma sequência de instruções que devem ser seguidas para transformar dados de entrada em dados de saída (ALPAYDIN, 2014). A aprendizagem de máquina permite que os sistemas aprendam e se adaptem (MICHALSKI et al., 2013), por meio de dados de exemplo ou experiências passadas (ALPAYDIN, 2014).

O aprendizado de máquina pode ser dividido em aprendizagem supervisionada, aprendizagem por reforço e aprendizagem não supervisionada. Na aprendizagem não supervisionada, o agente aprende padrões na entrada sem que seja fornecido um feedback, onde a tarefa mais comum desse tipo de aprendizagem é o agrupamento. Na aprendizagem por reforço, se utilizam os conceitos de recompensas ou punições. Na aprendizagem supervisionada, o agente observa um conjunto de exemplos de entrada e saída, realizando um mapeamento da entrada para a saída (NORVIG; RUSSELL, 2014).

1.2.1 Aprendizagem Supervisionada

Algoritmos de aprendizagem supervisionada são algoritmos que aprendem por meio de um conjunto de treinamento que possuem dados de entrada e saída (GOODFELLOW et al., 2016a). A associação entre entrada e saída é definida por uma função (modelo de classificação). Quando a saída é categórica, o problema é definido como classificação, quando é verificado se determinado atributo pertence ou não a uma classe, chamamos de classificação binária (NORVIG; RUSSELL, 2014). Algumas técnicas de aprendizagem supervisionada são muito utilizadas, como é o caso das árvores de decisão, classificadores baseados em regra, redes neurais artificiais (RNA) e máquina de vetores de suporte (SVM).

A aprendizagem supervisionada busca dentro de um conjunto de hipóteses aquela que possui o melhor desempenho, mesmo com dados que não sejam de treinamento. Para medir a precisão da hipótese, utiliza-se um conjunto de teste que se difere do conjunto de treinamento. O desempenho do modelo de classificação utiliza os resultados classificados corretamente e incorretamente no conjunto de teste, denominado matriz de confusão, exemplificada na Tabela 1 (TAN et al., 2006).

Tabela 1 – Matriz de confusão. Cada linha representa o valor verdadeiro da classe e cada coluna representa o valor predito para a classe.

		Valor Predito	
		Positivo	Negativo
Valor Verdadeiro	Positivo	VP	FN
	Negativo	FP	VN

Fonte: Autor

Outra métrica utilizada é a acurácia, representada pela Equação 1, que indica o quanto o algoritmo foi capaz de diferenciar as classes. O cálculo é realizado por meio do número de verdadeiros positivos (VP) e verdadeiros negativos (VN) dividido pelo total de predições.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \quad (1)$$

Além da acurácia, outros parâmetros são utilizados para identificar que o algoritmo está realizando predições corretamente, como a sensibilidade e a especificidade. A Sensibilidade é identificada a partir da quantidade de VP dividido pela soma de VP e FN, conforme mostra a

equação 2. Já a especificidade é identificada pela quantidade de VN dividida pela soma de VN e FP, como representada na equação 3 respectivamente (BARATLOO et al., 2015).

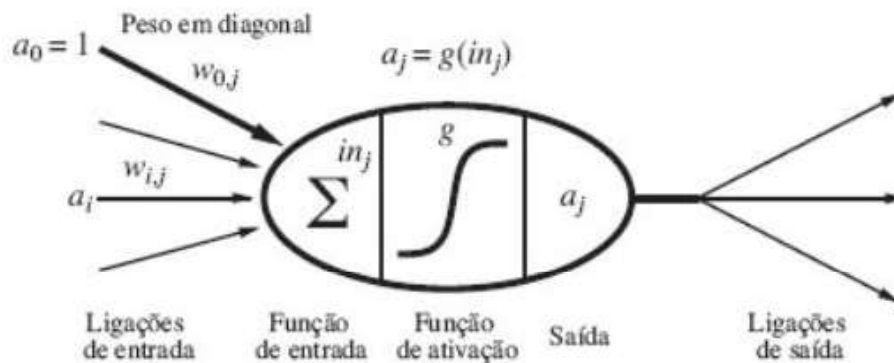
$$Sensibilidade = \frac{VP}{VP + FN} \quad (2)$$

$$Especificidade = \frac{VN}{VN + FP} \quad (3)$$

1.2.2 Redes Neurais Artificiais

A capacidade do cérebro humano em processar informações serviu de inspiração para a investigação e criação das redes neurais artificiais (NORVIG; RUSSELL, 2014). Uma rede neural artificial é um modelo matemático não-linear baseado no funcionamento dos neurônios, podendo ser utilizado como um classificador (HAYKIN, 1994). O primeiro modelo artificial de um neurônio biológico foi desenvolvido por McCulloch e Pitts (1943), representado em forma de desenho por Norvig e Russell (2014) (Figura 4).

Figura 4 – Modelo matemático simples de um neurônio



Fonte: (MCCULLOCH; PITTS, 1943; NORVIG; RUSSELL, 2014)

O modelo é representado por um vetor de entradas que recebe pesos. A soma ponderada das entradas é submetida a uma função de ativação que determina se o valor é maior que o limiar do neurônio, propagando o resultado para as ligações de saída (NORVIG; RUSSELL, 2014).

Há duas formas distintas de conectar os neurônios. A primeira usa uma alimentação para frente (*feedforward*) em que a informação flui por meio da função avaliada e cálculos intermediários são realizados para definir a saída. Por exemplo, determinado nó h recebe como entrada a saída de um nó x e propaga o resultado para um nó y (GOODFELLOW et al., 2016b).

A segunda ocorre quando as saídas do modelo são realimentadas, ou seja, quando as redes *feedforward* são ampliadas para permitir conexões de *feedback*, tornando o processo dinâmico. Essas redes são conhecidas como redes neurais recorrentes (RNNs) (NORVIG; RUSSELL, 2014).

1.2.3 Aprendizagem Profunda

A aprendizagem profunda, também conhecida como *Deep Learning*, permite que modelos computacionais com múltiplas camadas de processamento aprendam as representações de dados por meio de múltiplos níveis de abstração. As técnicas convencionais de aprendizado de máquina não possuem capacidade de processar dados naturais em sua forma bruta (LECUN et al., 2015).

A arquitetura de uma *Deep Learning* pode ser do tipo *feedforward* ou recorrente. O fluxo de uma rede *feedforward* é sempre para frente não tendo retorno da saída de um neurônio para ele mesmo. Já as redes recorrentes processam um elemento por vez na sequência de entrada e armazenam informações sobre o histórico de todos os elementos antigos da sequência em um vetor de estado, podendo aplicar a retropropagação para o treinamento (LECUN et al., 2015).

A retropropagação ou *backpropagation* calcula o erro das saídas, por meio de uma função gradiente. Essa função é aplicada várias vezes para todos os módulos, a partir da saída na parte superior em que a predição é produzida, até a parte inferior, onde a entrada é alimentada (LECUN et al., 2015).

1.2.3.1 Redes Neurais Recorrentes

Quando o conceito de retropropagação foi utilizado, a aplicação mais interessante foi para o treinamento de redes neurais recorrentes (RNNs) (LECUN et al., 2015). A compreensão de sequências, como frases, pelo ser humano é realizada, não somente utilizando a palavra lida no momento, mas com a compreensão das palavras lidas anteriormente. Ao se ler algo, o ser humano associa aquilo que já está em sua memória. As redes neurais convencionais não deixam claro como utilizar eventos passados para prever eventos futuros. Sendo assim, as RNNs abordam esse tipo de situação.

As RNNs são um conjunto de redes voltadas para o processamento de dados sequenciais de diferentes domínios como música (BOULANGER-LEWANDOWSKI et al., 2012; ECK; SCHMIDHUBER, 2002), captura em movimento (SUTSKEVER et al., 2009) e texto (SUTSKEVER et al., 2011). As RNNs podem ser treinadas utilizando sequências de dados reais, seguindo um passo de cada vez para prever o que virá depois. Utilizando previsões probabilísticas, novas

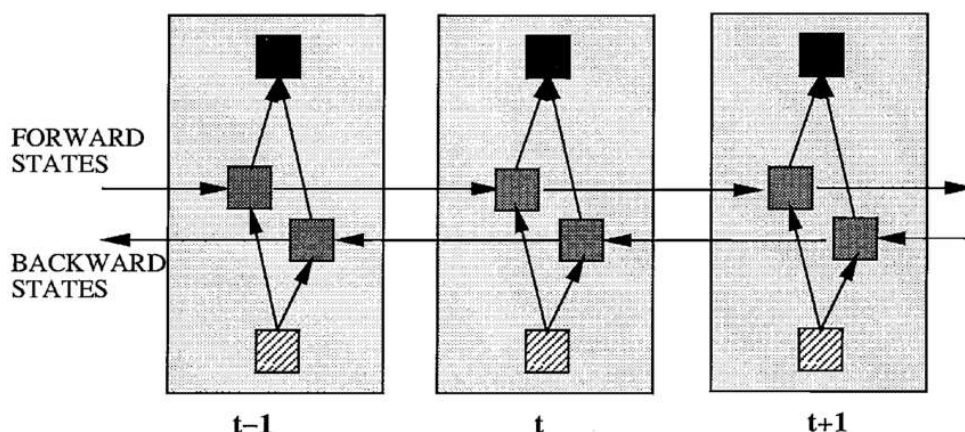
sequências podem ser geradas a partir da distribuição de saída e usadas como entrada para o próximo degrau (GRAVES, 2013). Sendo assim, as redes recorrentes, ao contrário de redes *feedforward*, produzem modelos dinâmicos, realizando classificações precisas e dependentes do contexto dos dados de exemplo.

Em teoria, uma rede neural recorrente é suficiente para gerar sequências de complexidade arbitrária, porém na prática, uma RNN padrão é incapaz de armazenar informações passadas a medida que a sequência aumenta, pois quanto maior a sequência mais difícil de construir uma RNN que consiga aprender a conectar informações (HOCHREITER et al., 2001). Sendo assim, redes neurais com um estado oculto de alta dimensionalidade não-linear, que permita lembrar e processar informações passadas, são necessárias (SUTSKEVER et al., 2011).

As redes recorrentes, assim como outros tipos de redes neurais, utilizam representações internas para realizar a interpolação de alta dimensão entre os exemplos de treinamento (GRAVES, 2013). Além disso, os gradientes da rede possuem baixo custo computacional com a retropropagação ao longo do tempo (SUTSKEVER et al., 2011). Contudo, as RNNs sofrem com problemas de gradiente de fuga. O gradiente de fuga ocorre quando os pesos computados nos estágios iniciais perdem influência conforme o número de iterações aumenta, tornando a rede sensível a novas entradas. Sendo assim, a variância de contexto que as RNNs convencionais conseguem captar, a medida que o tamanho da sequência aumenta, se torna altamente limitado.

As RNNs podem ser construídas utilizando uma arquitetura bidirecional, sendo denominadas de redes neurais recorrentes bidirecionais (BRNNs) (GRAVES; SCHMIDHUBER, 2005; GREFF et al., 2017). Uma BRNN pode ser treinada usando todas as informações de entrada, do passado e do futuro, disponíveis de um tempo específico (SCHUSTER; PALIWAL, 1997). Uma parte do estado do neurônio é responsável pelas direções positivas e negativas, *forward states* e *backward states*. As saídas dos estados não são conectadas entre si (Figura 5).

Figura 5 – Estrutura de uma BRNN utilizando três camadas



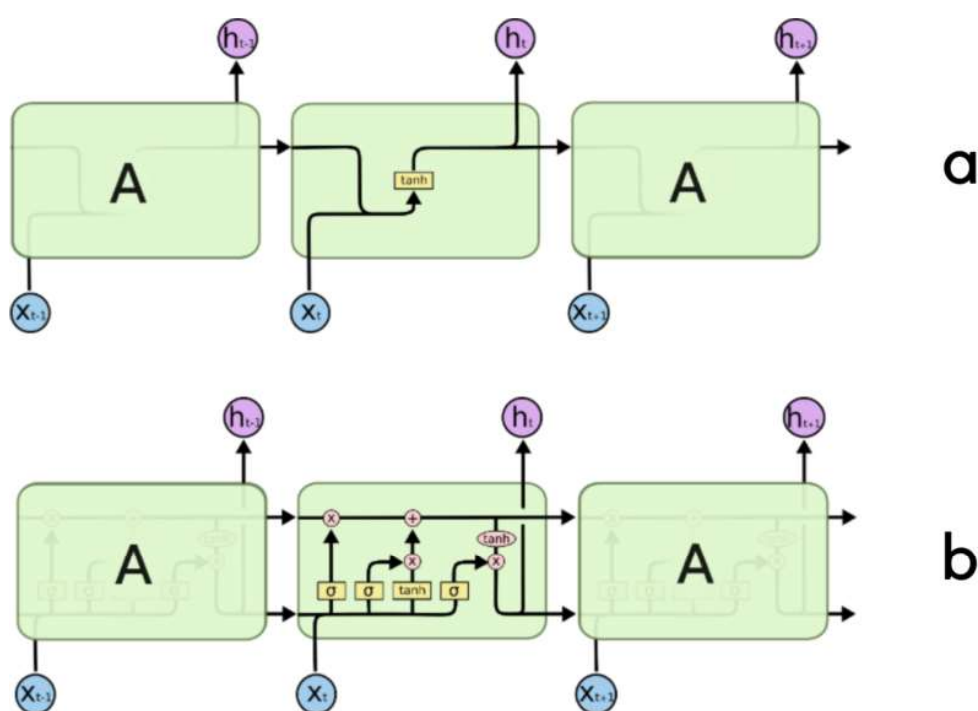
Fonte: (SCHUSTER; PALIWAL, 1997)

O diferencial de uma rede BRNN é o *backward states*, que se removido, transforma a rede em uma RNN simples unidirecional. Ao utilizar informações do passado e do futuro como entrada é possível minimizar a função objetivo sem a necessidade de atrasos para a inserção de novas informações (SCHUSTER; PALIWAL, 1997).

1.2.3.1.1 Redes de Memória de Longo Prazo

As redes de memória de longo prazo (LSTMs) foram propostas com o intuito de resolver o problema do gradiente de fuga (HOCHREITER; SCHMIDHUBER, 1997), separando a memória e a representação de saída, onde cada unidade de memória atual é dependente linearmente da unidade de memória anterior (RAVURI; STOLCKE, 2015). Recordar a informação por longos tempos é, praticamente, o comportamento padrão desse tipo de rede. Todas as RNNs possuem módulos repetitivos, porém em RNNs convencionais o módulo de repetição é estruturado de forma simples, contendo apenas uma camada, ao contrário de redes LSTMs, que possuem quatro camadas de interação (Figura 6) (OLAH, 2015).

Figura 6 – RNN convencional x RNN LSTM



Fonte: (OLAH, 2015)

O estado da célula é representado pela linha horizontal na Figura 6b, e é o componente fundamental para que as informações fluam. A primeira etapa decide quais informações devem ser descartadas do estado da célula que é determinado pela camada sigmóide ou camada de entrada. A camada sigmóide produz números entre zero (descartar completamente) e um (manter completamente). A segunda etapa define quais informações devem ser armazenadas no estado da célula; para que isso ocorra, a camada sigmóide decide quais valores devem ser atualizados e em seguida a camada **tanh** (tangente hiperbólica) cria um vetor de valores com possíveis candidatos que podem ser adicionados ao estado. A terceira etapa consiste no que será enviado para a saída. O resultado é baseado no estado da célula em que uma camada sigmóide decide quais dados devem ser produzidos, inserindo a informação por meio de uma função **tanh** (determina valores entre -1 e 1) e multiplica pela saída da camada sigmóide, reproduzindo apenas o que é de interesse e que deve ser passado adiante (HOCHREITER; SCHMIDHUBER, 1997). A saída da camada oculta LSTM pode ser calculada por meio da função composta descrita na Equação 4 (GRAVES, 2013).

$$\begin{aligned}
 i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\
 f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
 c_t &= f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
 o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\
 h_t &= o_t \tanh(c_t)
 \end{aligned} \tag{4}$$

em que σ é a função sigmóide e i, f, o e c são, respectivamente, os vetores de entrada, esquecimento, saída e ativação da célula (GRAVES, 2013). A mudança dos estados ocultos nas BRNNs por módulos de memória das LSTMs resulta na estrutura das LSTMs bidirecionais (RNN - BLSTM) (WANG et al., 2015b). Neste trabalho, foi utilizado para a camada oculta a estrutura das RNN-BLSTM. A saída da camada oculta serviu de entrada para a camada de saída.

A célula de memória atua como um acumulador ou um neurônio, pois possui uma conexão com ela mesma e copia seu próprio estado real. Essa auto-conexão é multiplicada por outra célula que aprende a decidir quando limpar o conteúdo da memória. Sendo assim, as redes LSTM mostram-se mais eficazes que as RNNs convencionais (LECUN et al., 2015).

1.3 PROCESSAMENTO DE LINGUAGEM NATURAL

Dados não estruturados, especialmente textos, imagens e vídeos, contêm uma grande quantidade de informações. O Processamento de Linguagem Natural (PLN) é um campo da ciência da computação e inteligência artificial com fundamento na linguística para analisar e representar textos que ocorrem naturalmente em um ou mais níveis de análise linguística. As aplicações de PLN, geralmente, são usadas para i) recuperação de informação a partir de textos; ii) tradução automática; iii) interpretação de textos; e iv) inferências a partir de textos (LIDDY, 2001).

Do ponto de vista linguístico, as pesquisas em PLN são distribuídas em níveis de análise fonético ou fonológico, morfológico, sintático, semântico e pragmático. Cada nível é considerado a partir do propósito da aplicação, por exemplo, a preocupação com a análise fonológica, regularmente, não é necessária para aplicações voltadas a textos científicos. Em contrapartida, considerar as análises sintáticas e semânticas são necessárias nas aplicações de classificação de texto (PERNA et al., 2010). O reconhecimento e a definição de sentenças agregadas a classificação das palavras possibilita o desenvolvimento de aplicações de interação entre máquinas e linguagens naturais.

1.3.1 Classificação de Texto

A abundância de documentos disponibilizados em formato digital e a necessidade de encontrar informações relevantes são desafiadoras para os seres humanos. A classificação de texto é a ação de dividir um conjunto de documentos ou sentenças de entrada em duas ou mais classes, onde cada sentença pode pertencer a uma ou mais classes (SEBASTIANI, 2002).

A compreensão sequencial pelo ser humano é realizada, não somente utilizando uma palavra, mas com inferências contextuais. Algumas técnicas de IA são utilizadas para essas classificações automáticas, das quais abordam PLN e RNNs (LECUN et al., 2015).

O pré-processamento e ajuste dos dados são essenciais e fazem parte da etapa de conversão na qual traduz os dados para uma linguagem compreendida por máquinas (JUNIOR, 2007). A tokenização é uma operação utilizada para detectar os limites das palavras (*tokens*) por meio da análise morfológica. A extração de características (*features*) ocorre por meio da vetorização de *tokens*, na qual as palavras são transformadas em inteiros. O mapeamento das palavras para um espaço vetorial é conhecido por incorporação de palavras (do inglês *word embedding*) (PENNINGTON et al., 2014). O coeficiente desse *token* é calculado a partir da contagem das palavras. A contagem de palavras é realizada por meio do cálculo de TF-IDF (do

inglês *term frequency–inverse document frequency*) descrita pela equação 5:

$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right) \quad (5)$$

em que tf é o número de ocorrências de termos i no documento j , multiplicado pelo $\log$ do número total de documentos N , dividido pelo número de documentos que apresentam i (ROBERTSON, 2004).

Caracterizar sentenças entre outras em um conjunto, usualmente, chamado de Corpus, é uma tarefa relativamente trivial. A importância da palavra de uma sentença relacionada a uma coleção de sentenças, calculada pelo TF-IDF, agregado a *word embeddings* permite codificar o significado das palavras para espaços vetoriais que são compreendidos por algoritmos de aprendizado de máquina (SENEL et al., 2018).

1.4 TRABALHOS RELACIONADOS

Alguns estudos têm se preocupado em aplicar técnicas de inteligência artificial para problemas da biologia computacional. Dentre esses, esta seção apresenta cinco trabalhos que se relacionam de alguma forma com esta pesquisa.

O estudo "*Automatic document classification of biological literature*" descreve a classificação de documentos em contexto biológico usando um corpus de *Caenorhabditis elegans* e apoiado por uma ferramenta de marcação desenvolvida anteriormente. Os autores utilizaram SVM e clusterização baseada em frase para classificar e categorizar os documentos. Para a devida classificação, especialistas criaram regras baseadas na realidade para que os resultados pudessem ser obtidos de forma eficiente. O desenvolvimento de uma plataforma *web* possibilitou a visualização e procura por documentos pertencentes a conceitos específicos (CHEN et al., 2006).

No trabalho "*Long Short-Term Memory Recurrent Neural Networks for Antibacterial Peptide Identification*" (YOU MANS et al., 2017) redes LSTM bidirecionais foram implementadas empregando características individuais dos aminoácidos dentro de cada sequência a fim de classificar peptídeos como antibacterianos ou não antibacterianos. Na primeira linha de defesa do sistema imune, contra patógenos microbianos, encontramos esses peptídeos (PASUPULETI et al., 2012; DESLOUCHES et al., 2013) que podem ser usados para a descoberta de uma nova classe de antibióticos (PASUPULETI et al., 2012). Além disso, outros estudos utilizam técnicas de aprendizado de máquina para a identificação de peptídeos por meio de dados genômicos

(CHERKASOV et al., 2008; NAGARAJAN et al., 2018), porém a necessidade de especialistas também é de extrema necessidade.

Em "*Long short-term memory RNN for biomedical named entity recognition*" é proposto uma arquitetura de BLSTM baseada em incorporação de palavras e representação de caracteres utilizando campos aleatórios condicionais. A abordagem proposta gera modelos com incorporações de palavras sem e com pré-treinados utilizando BioCreative II e JNLPBM. A abordagem com contexto mais generérico possibilita que não exista etapas manuais. Porém, aplicações mais específicas e pouco mapeadas precisam considerar etapas manuais (LYU et al., 2017).

O estudo "*Identifying Antimicrobial Peptides using Word Embedding with Deep Recurrent Neural Networks*" usa incorporações de sequências de proteínas aplicadas na representação de bacteriocinas. Bacteriocinas são produtos peptídicos antimicrobianos produzidos pelas bactérias. A descoberta de novas bacteriocinas possibilita a exploração de novos possíveis fármacos. A predição de bacteriocinas sem técnicas de similaridade nas sequências foi proposta por meio de RNN e SVM. A abordagem previu seis bacteriocinas ainda desconhecidas em *Lactobacillus*. A representação das palavras/sequências por meio de incorporações preserva a informação em sequência o que possibilita resolver problemas de classificação de proteína, nas quais a homologia de sequências não pode ser usada (HAMID; FRIEDBERG, 2018).

O estudo "*Deep learning of mutation-gene-drug relations from the literature*" apresenta dois métodos que utilizam artigos do PubMed para auxiliar na extração e curadoria nas relações entre genes, mutações e fármacos. O desenvolvimento de uma ferramenta de pesquisa de entidades biomédicas foi proposta acoplada a recursos de treinamento e classificação. Adicionalmente, incorporações de palavras usando redes neurais convolucionais foram contruídas empregando resumos de artigos científicos e o Google Notícias. Dessa forma, a extração de relações entre mutação-gene e mutação-droga foi realizada. Um banco de dados das relações mutação-gene-droga encontra-se foi elaborado no intuito de servir como recurso futuro para pesquisadores de medicina de precisão (LEE et al., 2018).

Os trabalhos acima citados estão diretamente relacionados a assuntos biológicos, porém poucos trabalhos deste porte estão associados à resistência antimicrobiana. O uso de *word embeddings* próprios são apresentados na maioria dos trabalhos, o que é considerável visto que os métodos computacionais propostos aprendem o significado, o contexto e as relações dos dados por meio de valores numéricos que representam as palavras.

Alguns estudos utilizam redes neurais do tipo BLSTM, pois a alta complexidade de trabalhar com estruturas sequenciais podem ser resolvidas com essa técnica. Infelizmente, poucos

trabalhos disponibilizam, por meio de ferramentas *web*, o melhor modelo gerado o que interfere na disseminação do conhecimento e da pesquisa. Estudos complexos que geram informações devem propor estratégias que armazenam os novos dados a fim de melhorar a performance do trabalho atual ou propor novos estudos. Neste sentido, poucos estudos apresentam a construção de um banco de dados.

Para melhor visualização das comparações aqui realizadas, a Tabela 2 esquematiza, brevemente, as características deste estudo comparado a estudos já publicados.

Tabela 2 – Comparação entre os Trabalhos

DOI	Autores	Resistência Antimicrobiana	Incorporação de palavras próprias	BLSTM	Ferramenta web	Banco de Dados
-	MARQUES; WINCK 2018	X	X	X	X	X
10.1186/1471-2105-7-370	CHEN et al. 2006	-	X	-	X	-
10.1109/BIBM.2017.8217697	YOU MANS et al. 2017	X	-	X	-	-
10.1186/s12859-017-1868-5	LYU et al. 2017	-	X	X	-	-
10.1101/255505	HAMID; FRIEDBERG 2018	-	X	-	-	-
10.1186/s12859-018-2029-1	LEE et al. 2018	-	X	-	X	X

Fonte: Autor

2 OBJETIVOS

2.1 OBJETIVO GERAL

Este trabalho tem como principal objetivo apresentar uma metodologia de classificação de textos científicos, a partir de contexto, utilizando um algoritmo rede neural de aprendizagem profunda, para identificar sentenças que apresentem microrganismos ou antimicrobianos associados à resistência antimicrobiana.

Este propósito vai em direção à uma abordagem que propicie uma análise ampla e automática da literatura biomédica para coleta e catalogação dos agentes já conhecidos na comunidade científica relacionados à resistência antimicrobiana. Com essa catalogação, espera-se contribuir para futuras investigações acerca do assunto, ao permitir que futuros experimentos sejam realizados com o propósito de prever microrganismos resistentes.

Dessa forma, espera-se que o resultado deste trabalho traga, futuramente, benefícios aos pacientes, suas famílias e a diminuição de custos hospitalares com internações e tratamentos, muitas vezes, ineficazes. A proposta deste trabalho vai ao encontro do segundo objetivo proposto pelo Plano de Ação Global de combate à resistência antimicrobiana, ao promover a ampliação do alcance e democratização do conhecimento científico acerca dos agentes etiológicos e antimicrobianos relacionados a casos de resistência.

2.2 OBJETIVOS ESPECÍFICOS

Para atender ao objetivo principal, define-se os seguintes objetivos específicos:

- Elaborar um conjunto de palavras chave para utilizar como busca de artigos relevantes em repositórios científicos da área;
- Construir e classificar um conjunto de dados com frases de resistência antimicrobiana a partir dos artigos científicos relevantes;
- Implementar os algoritmos e treinar as redes com o conjunto de dados construído e analisar os resultados de treinamento para gerar o modelo mais apropriado;
- Construir uma plataforma *web* para capturar novos exemplos de texto relacionadas à resistência antimicrobiana, bem como agentes etiológicos por meio da integração do modelo gerado;

- Construir uma base de dados com frases que contenham agentes de infecção e resistência antimicrobiana por meio da plataforma *web*.

3 ARTIGO CIENTÍFICO

Nesta seção é apresentado um artigo científico, de título homônimo à esta monografia, desenvolvido considerando os moldes de submissão do periódico *BMC Bioinformatics*. Este periódico conta, atualmente, com fator de impacto 2.213.

As autoras declaram que o artigo ainda não foi submetido ao periódico. Entretanto, desejam submetê-lo após tradução e revisão.

Classificação de Texto associado à Resistência Antimicrobiana com Redes Neurais Recorrentes Bidirecionais

Autores

Flavielle Blanco Marques (1)

Ana Trindade Winck (1)

Afiliação

1. Informática Biomédica, Universidade Federal de Ciências da Saúde de Porto Alegre

E-mail de Correspondência

flaviellebm@gmail.com

anawinck@gmail.com

4 DISCUSSÃO

Os casos de resistência antimicrobiana, que tem levado a óbito vários pacientes, têm provocado alerta para o uso exarcebado e descontrolado de antibióticos. Diversos são os tipos infecções que estão relacionados à resistência antimicrobiana. Entretanto, aquelas relacionadas ao Sistema Nervoso Central apresentaram uma motivação para o desenvolvimento deste trabalho.

Entende-se que é possível contribuir para o diagnóstico das infecções de difícil tratamento do SNC, de maneira automatizada. O projeto intitulado "Diagnóstico de Infecções do Sistema Nervoso Central e Desenvolvimento de Novas Estratégias Terapêuticas"(THOMPSON, 2017) busca desenvolver protocolos experimentais a partir de sequenciamento de DNA para identificação dos patógenos presentes nos pacientes com infecções de difícil controle. Para tanto, busca-se, com o referido projeto, ir em direção à medicina personalizada, ao desenvolver algoritmos de aprendizado de máquina capazes de prever o tratamento antimicrobiano que possa ser mais efetivo para um dado paciente, considerando o patógeno identificado.

Como discutido nesta monografia, algoritmos de aprendizado preditivo aprendem a partir de experiências passadas. Nesse sentido, é necessário conhecer toda a gama cientificamente conhecida de organismos de resistência antimicrobiana, bem como os antimicrobianos associados à resistência, para que novos tratamentos possam ser preditos.

Frente a esse cenário, entende-se como fundamental realizar uma catalogação dos patógenos relacionados à resistência e para isto entende-se que as ações iniciais devam ir em busca dos agentes que são conhecidos pela comunidade científica e relatados na literatura. Assim, este trabalho buscou contribuir com a implementação de redes neurais sobre análise de textos científicos, para a identificação de sentenças que remetam a agentes de resistência. Foi apresentada uma metodologia que engloba desde a seleção de artigos científicos relevantes a casos de resistência para criação de um *dataset*, passando pela implementação e testes de algoritmos de redes neurais para análise desses textos, até a criação de uma plataforma *web* que permite classificar uma sentença, podendo incluí-la em uma base de dados.

Neste trabalho, diversos experimentos foram realizados com o propósito de encontrar o melhor modelo e configuração de rede. A busca por um modelo que suportasse e tivesse bom desempenho na classificação de novas sentenças foi gerado. Testes mostraram que o suporte a um modelo de incorporação de palavras pré treinado tem melhores resultados, assim como, valores de dimensionalidade semelhantes ao tamanho das sentenças contidas no conjunto de dados.

O conjunto de dados criado e rotulado neste trabalho possui poucas instâncias se comparado ao poder computacional que as redes neurais recorrentes bidirecionais podem suportar. O número baixo de instâncias e a rotulação apenas por um ponto de vista pode causar um viés no próprio conjunto de dados e explicar tantos testes que apresentaram *overfitting*. Apesar desses contrapontos, uma configuração adaptável e aceitável foi atingida.

A inserção de sentenças desconhecidas e a classificação correta pelo modelo pode evidenciar o aprendizado por contexto. Sentenças com a presença de microrganismos e que evidenciavam relação à resistência antimicrobiana, ser prejudicial ou ruim, em sua maioria, foram classificadas corretamente. Sentenças que apresentavam microrganismos com termos biológicos específicos tiveram desempenho reduzido ou foram classificadas erradas.

Para resolver problemas com siglas e termos biológicos, a construção de um modelo de incorporação de palavras específico para resistência antimicrobiana poderá melhorar os resultados, consideravelmente. Além disso, novas categorias de classificação podem ser criadas no conjunto de dados, possibilitando a separação de microorganismos e agentes antimicrobianos.

O modelo gerado foi incorporado a uma ferramenta *web*, cuja tela é ilustrada no Apêndice A. A ferramenta é integrada ao banco de dados IRRA que permite salvar as novas sentenças classificadas corretamente pelo modelo. Essa permitirá, futuramente, que mais dados sejam usados para o treinamento da rede.

Além disso, pesquisas futuras, correlacionando, resistência antimicrobiana e extração de entidades nomeadas poderão ser realizadas. O refinamento do método aqui criado poderá ser realizado com o auxílio das novas sentenças. A descoberta de novos padrões ainda desconhecidos poderá servir para novos estudos.

Encontrar trabalhos relacionados e que façam, exatamente, todas as etapas que o método aqui proposto faz, foi difícil, o que permite que novos estudos sejam realizados. A falta de trabalhos que realizam as mesmas coisas dificulta a comparação com outros métodos. Porém, instiga a necessidade de novas técnicas nessa área.

As incertezas e complexidades dos mecanismos e a diversidade de fatores que levam os microrganismos a apresentar resistência dificultam na descoberta de conhecimento, porém, neste estudo, apesar dos resultados promissores, sabemos da carência e dificuldade de mecanismos de identificação de agentes etiológicos que ainda são desconhecidos.

5 CONCLUSÃO

A resistência antimicrobiana é um problema à saúde pública mundial (HUTTNER et al., 2013). Dado esse problema, encontramos estudos que buscam identificar os fatores que causam a resistência aos antimicrobianos e, conseqüentemente, geram custos elevados na área da saúde. Pesquisas que envolvam técnicas de aprendizado de máquina estão cada vez mais frequentes, porém, estudos que abordam estratégias aplicadas à resistência antimicrobiana não foram encontrados com tanta facilidade na literatura.

Neste trabalho, a estratégia de classificar resistência antimicrobiana por meio do contexto utilizando sentenças científicas foi proposta. Técnicas de aprendizado de máquina foram executadas, mais especificamente, redes neurais recorrentes bidirecionais. Vários experimentos foram realizados para alcançar o melhor modelo.

Com o propósito de alcançar o melhor modelo de classificação, foram realizados testes que envolveram mudanças de parâmetros e estratégias como o uso de um modelo pré treinado de incorporação de palavras. O desempenho da rede sobre sentenças desconhecidas foi observado e a escolha do melhor modelo foi obtida considerando a maior sensibilidade e o menor *overfitting*.

Acreditamos que os resultados sejam promissores e levaram a uma ferramenta que permite a inserção de novas sentenças, visando trabalhos futuros que fortaleçam a base de conhecimento e evidências sobre resistência antimicrobiana.

Assim, como trabalhos futuros, propõe-se:

- Aprimorar a plataforma *web* e a disponibilidade do banco de dados para promover a relação de acesso público dos microorganismos com antimicrobianos;
- Fazer uso dos dados coletados por esse trabalho para construir um Banco de Dados de Perfil de Resistência;
- Desenvolver métodos de PLN que possam extrair e reconhecer Entidades Nomeadas relacionadas à resistência antimicrobiana, inserindo esse método na ferramenta desenvolvida;
- e
- Desenvolver algoritmos de aprendizado de máquina para, com o suporte das bases de dados construídas, reconhecer novos padrões de resistência antimicrobiana.

REFERÊNCIAS

- AGÊNCIA NACIONAL DE VIGILÂNCIA SANITÁRIA. **Anvisa**: Plano nacional para a prevenção e o controle da resistência microbiana nos serviços de saúde. Brasília, 2017a.
- AGÊNCIA NACIONAL DE VIGILÂNCIA SANITÁRIA. **Anvisa**: Diretriz nacional para elaboração de programa de gerenciamento do uso de antimicrobianos em serviços de saúde. Brasília, 2017b.
- ALPAYDIN, E. **Introduction to machine learning**. [S.l.]: MIT press, 2014.
- BARATLOO, A.; HOSSEINI, M.; NEGIDA, A.; ASHAL, G. E. Part 1: simple definition and calculation of accuracy, sensitivity and specificity. **EMERGENCY**, 2015.
- BEEK, D. Van de; CABELLOS, C.; DZUPOVA, O.; ESPOSITO, S.; KLEIN, M.; KLOEK, A.; LEIB, S.; MOURVILLIER, B.; OSTERGAARD, C.; PAGLIANO, P. et al. Escmid guideline: diagnosis and treatment of acute bacterial meningitis. **Clinical microbiology and infection**, Elsevier, v. 22, p. S37–S62, 2016.
- BOULANGER-LEWANDOWSKI, N.; BENGIO, Y.; VINCENT, P. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. **arXiv preprint arXiv:1206.6392**, 2012.
- CDC. **HAI Data**. [S.l.], 2018. Disponível em: <<https://www.cdc.gov/hai/surveillance/index.html>>. Acesso em: 20 mar. 2018.
- CHEN, D.; MÜLLER, H.-M.; STERNBERG, P. W. Automatic document classification of biological literature. **BMC bioinformatics**, BioMed Central, v. 7, n. 1, p. 370, 2006.
- CHERKASOV, A.; HILPERT, K.; JENSSEN, H.; FJELL, C. D.; WALDBROOK, M.; MULLALY, S. C.; VOLKMER, R.; HANCOCK, R. E. Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs. **ACS chemical biology**, ACS Publications, v. 4, n. 1, p. 65–74, 2008.
- DESLOUCHES, B.; STECKBECK, J. D.; CRAIGO, J. K.; DOI, Y.; MIETZNER, T. A.; MONTELARO, R. C. Rational design of engineered cationic antimicrobial peptides consisting exclusively of arginine and tryptophan: WreCAP activity against multidrug-resistant pathogens. **Antimicrobial agents and chemotherapy**, Am Soc Microbiol, p. AAC–02218, 2013.
- ECDC. **Factsheet for experts - Antimicrobial resistance**. [S.l.], 2018. Disponível em: <<https://ecdc.europa.eu/en/antimicrobial-resistance/facts/factsheets/experts>>. Acesso em: 21 mar. 2018.
- ECK, D.; SCHMIDHUBER, J. A first look at music composition using lstm recurrent neural networks. **Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale**, v. 103, 2002.
- ERDEM, H.; INAN, A.; GUVEN, E.; HARGREAVES, S.; LARSEN, L.; SHEHATA, G.; PERNICOVA, E.; KHAN, E.; BASTAKOVA, L.; NAMANI, S. et al. The burden and epidemiology of community-acquired central nervous system infections: a multinational study. **European Journal of Clinical Microbiology & Infectious Diseases**, Springer, v. 36, n. 9, p. 1595–1611, 2017.

- GLASER, C.; HONARMAND, S.; ANDERSON, L.; SCHNURR, D.; FORGHANI, B.; COSSEN, C.; SCHUSTER, F.; CHRISTIE, L.; TUREEN, J. Beyond viruses: clinical profiles and etiologies associated with encephalitis. **Clinical Infectious Diseases**, The University of Chicago Press, v. 43, n. 12, p. 1565–1577, 2006.
- GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. <<http://www.deeplearningbook.org>>.
- GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.; BENGIO, Y. **Deep learning**. [S.l.]: MIT press Cambridge, 2016. v. 1.
- GRAVES, A. Generating sequences with recurrent neural networks. **arXiv preprint arXiv:1308.0850**, 2013.
- GRAVES, A.; SCHMIDHUBER, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. **Neural Networks**, Elsevier, v. 18, n. 5-6, p. 602–610, 2005.
- GREFF, K.; SRIVASTAVA, R. K.; KOUTNÍK, J.; STEUNEBRINK, B. R.; SCHMIDHUBER, J. Lstm: A search space odyssey. **IEEE transactions on neural networks and learning systems**, IEEE, v. 28, n. 10, p. 2222–2232, 2017.
- HAMID, M. N.; FRIEDBERG, I. Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. **bioRxiv**, Cold Spring Harbor Laboratory, p. 255505, 2018.
- HAYKIN, S. **Neural networks: a comprehensive foundation**. [S.l.]: Prentice Hall PTR, 1994.
- HOCHREITER, S.; BENGIO, Y.; FRASCONI, P.; SCHMIDHUBER, J. et al. **Gradient flow in recurrent nets: the difficulty of learning long-term dependencies**. [S.l.]: A field guide to dynamical recurrent neural networks. IEEE Press, 2001.
- HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.
- HOUSE, W. National strategy for combating antibiotic resistant bacteria. **The White House, Washington, DC**. https://www.whitehouse.gov/sites/default/files/docs/carb_national_strategy.pdf, 2014.
- HUTTNER, A.; HARBARTH, S.; CARLET, J.; COSGROVE, S.; GOOSSENS, H.; HOLMES, A.; JARLIER, V.; VOSS, A.; PITTET, D. Antimicrobial resistance: a global view from the 2013 world healthcare-associated infections forum. **Antimicrobial resistance and infection control**, BioMed Central, v. 2, n. 1, p. 31, 2013.
- JUNIOR, J. R. C. Desenvolvimento de uma metodologia para mineração de textos. **Departamento de Engenharia Elétrica, Pontífica Universidade Católica do Rio de Janeiro**, 2007.
- KNUDSEN, R. C.; SLAZYK, W. E.; RICHMOND, J. Y.; HANNON, W. H. Guidelines for the shipment of dried blood spot specimens. **Infant Screening**, v. 16, n. 1, 1993.
- LEAVIS, H. L.; WILLEMS, R. J.; WAMEL, W. J. V.; SCHUREN, F. H.; CASPERS, M. P.; BONTEN, M. J. Insertion sequence–driven diversification creates a globally dispersed emerging multiresistant subspecies of *e. faecium*. **PLoS pathogens**, Public Library of Science, v. 3, n. 1, p. e7, 2007.

- LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group, v. 521, n. 7553, p. 436, 2015.
- LEE, K.; KIM, B.; CHOI, Y.; KIM, S.; SHIN, W.; LEE, S.; PARK, S.; KIM, S.; TAN, A. C.; KANG, J. Deep learning of mutation-gene-drug relations from the literature. **BMC bioinformatics**, BioMed Central, v. 19, n. 1, p. 21, 2018.
- LIDDY, E. D. Natural language processing. 2001.
- LIU, B.; POP, M. Ardb—antibiotic resistance genes database. **Nucleic acids research**, Oxford University Press, v. 37, n. suppl_1, p. D443–D447, 2008.
- LYU, C.; CHEN, B.; REN, Y.; JI, D. Long short-term memory rnn for biomedical named entity recognition. **BMC bioinformatics**, BioMed Central, v. 18, n. 1, p. 462, 2017.
- MCARTHUR, A. G.; WAGLECHNER, N.; NIZAM, F.; YAN, A.; AZAD, M. A.; BAYLAY, A. J.; BHULLAR, K.; CANOVA, M. J.; PASCALE, G. D.; EJIM, L. et al. The comprehensive antibiotic resistance database. **Antimicrobial agents and chemotherapy**, Am Soc Microbiol, p. AAC-00419, 2013.
- MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, Springer, v. 5, n. 4, p. 115–133, 1943.
- MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. **Machine learning: An artificial intelligence approach**. [S.l.]: Springer Science & Business Media, 2013.
- NAGARAJAN, D.; NAGARAJAN, T.; ROY, N.; KULKARNI, O.; RAVICHANDRAN, S.; MISHRA, M.; CHAKRAVORTTY, D.; CHANDRA, N. Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria. **Journal of Biological Chemistry**, ASBMB, v. 293, n. 10, p. 3492–3509, 2018.
- NORVIG, P.; RUSSELL, S. **Inteligência Artificial: Tradução da 3a Edição**. [S.l.]: Elsevier Brasil, 2014. v. 1.
- OLAH, C. Understanding lstm networks. **GITHUB blog, posted on August**, v. 27, p. 2015, 2015.
- ORGANIZATION, W. H. et al. **Frequently asked questions on antimicrobial resistance**. [S.l.], 2011.
- PASUPULETI, M.; SCHMIDTCHEN, A.; MALMSTEN, M. Antimicrobial peptides: key components of the innate immune system. **Critical reviews in biotechnology**, Taylor & Francis, v. 32, n. 2, p. 143–171, 2012.
- PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: Global vectors for word representation. In: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543.
- PERNA, C. L.; DELGADO, H. K.; FINATTO, M. J. **Linguagens Especializadas em Corpora: modos de dizer e interfaces de pesquisa**. [S.l.]: EDIPUCRS, 2010.
- RAVURI, S.; STOLCKE, A. Recurrent neural network and lstm models for lexical utterance classification. In: **Sixteenth Annual Conference of the International Speech Communication Association**. [S.l.: s.n.], 2015.

ROBERTSON, S. Understanding inverse document frequency: on theoretical arguments for idf. **Journal of documentation**, Emerald Group Publishing Limited, v. 60, n. 5, p. 503–520, 2004.

ROTBART, H. A. Viral meningitis. In: COPYRIGHT© 2000 BY THIEME MEDICAL PUBLISHERS, INC., 333 SEVENTH AVENUE, NEW YORK, NY 10001, USA. TEL.:+ 1 (212) 584-4662. **Seminars in neurology**. [S.l.], 2000. v. 20, n. 03, p. 277–292.

SCHUSTER, M.; PALIWAL, K. K. Bidirectional recurrent neural networks. **IEEE Transactions on Signal Processing**, IEEE, v. 45, n. 11, p. 2673–2681, 1997.

SCULLY, C. **Scully's Medical Problems in Dentistry E-Book**. [S.l.]: Elsevier Health Sciences, 2014.

SEBASTIANI, F. Machine learning in automated text categorization. **ACM computing surveys (CSUR)**, ACM, v. 34, n. 1, p. 1–47, 2002.

SENEL, L. K.; UTLU, I.; YUCESAY, V.; KOC, A.; CUKUR, T. Semantic structure and interpretability of word embeddings. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, IEEE, 2018.

SI-TUAN, N.; NGOC, H. M.; HANG, P. T. T.; NGUYEN, C.; VAN, P. H.; HUONG, N. T. New eight genes identified at the clinical multidrug-resistant acinetobacter baumannii dms06669 strain in a vietnam hospital. **Annals of clinical microbiology and antimicrobials**, BioMed Central, v. 16, n. 1, p. 74, 2017.

SUTSKEVER, I.; HINTON, G. E.; TAYLOR, G. W. The recurrent temporal restricted boltzmann machine. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2009. p. 1601–1608.

SUTSKEVER, I.; MARTENS, J.; HINTON, G. E. Generating text with recurrent neural networks. In: **Proceedings of the 28th International Conference on Machine Learning (ICML-11)**. [S.l.: s.n.], 2011. p. 1017–1024.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. Introduction to data mining. 2006.

THOMPSON, C. E. **Diagnóstico de Infecções do Sistema Nervoso Central e Desenvolvimento de Novas Estratégias Terapêuticas**. 2017.

TORTORA, G. J.; CASE, C. L.; FUNKE, B. R. **Microbiologia-12ª Edição**. [S.l.]: Artmed Editora, 2016.

TSIOUTIS, C.; KARAGEORGOS, S. A.; STRATAKOU, S.; SOUNDOULOUNAKI, S.; KARABETSOS, D. A.; KOUYENTAKIS, G.; GIKAS, A. Clinical characteristics, microbiology and outcomes of external ventricular drainage-associated infections: The importance of active treatment. **Journal of Clinical Neuroscience**, Elsevier, v. 42, p. 54–58, 2017.

WAGHU, F. H.; BARAI, R. S.; GURUNG, P.; IDICULA-THOMAS, S. Campr3: a database on sequences, structures and signatures of antimicrobial peptides. **Nucleic acids research**, Oxford University Press, v. 44, n. D1, p. D1094–D1097, 2015.

WANG, G.; LI, X.; WANG, Z. Apd3: the antimicrobial peptide database as a tool for research and education. **Nucleic acids research**, Oxford University Press, v. 44, n. D1, p. D1087–D1093, 2015.

WANG, P.; QIAN, Y.; SOONG, F. K.; HE, L.; ZHAO, H. A unified tagging solution: Bidirectional lstm recurrent neural network with word embedding. **arXiv preprint arXiv:1511.00215**, 2015.

WHO. **Global action plan on antimicrobial resistance. 2015.** 2017.

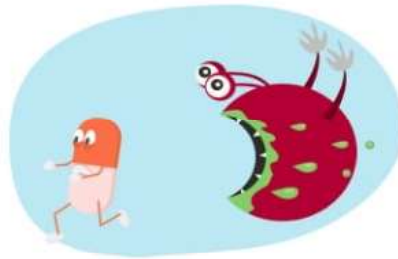
YIN, W.; LI, H.; SHEN, Y.; LIU, Z.; WANG, S.; SHEN, Z.; ZHANG, R.; WALSH, T. R.; SHEN, J.; WANG, Y. Novel plasmid-mediated colistin resistance gene mcr-3 in escherichia coli. **MBio**, Am Soc Microbiol, v. 8, n. 3, p. e00543–17, 2017.

YOUMANS, M.; SPAINHOUR, C.; QIU, P. Long short-term memory recurrent neural networks for antibacterial peptide identification. In: IEEE. **Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on.** [S.l.], 2017. p. 498–502.

APÊNDICE A – Ferramenta *web* de coleta de sentenças

Home

Ferramenta de Coleta de Sentenças de Resistência Antimicrobiana



Buscar

Desenvolvido por: Flavielle B. Marques

ANEXO A – Identificação de Resistência x Descoberta do Antibiótico

