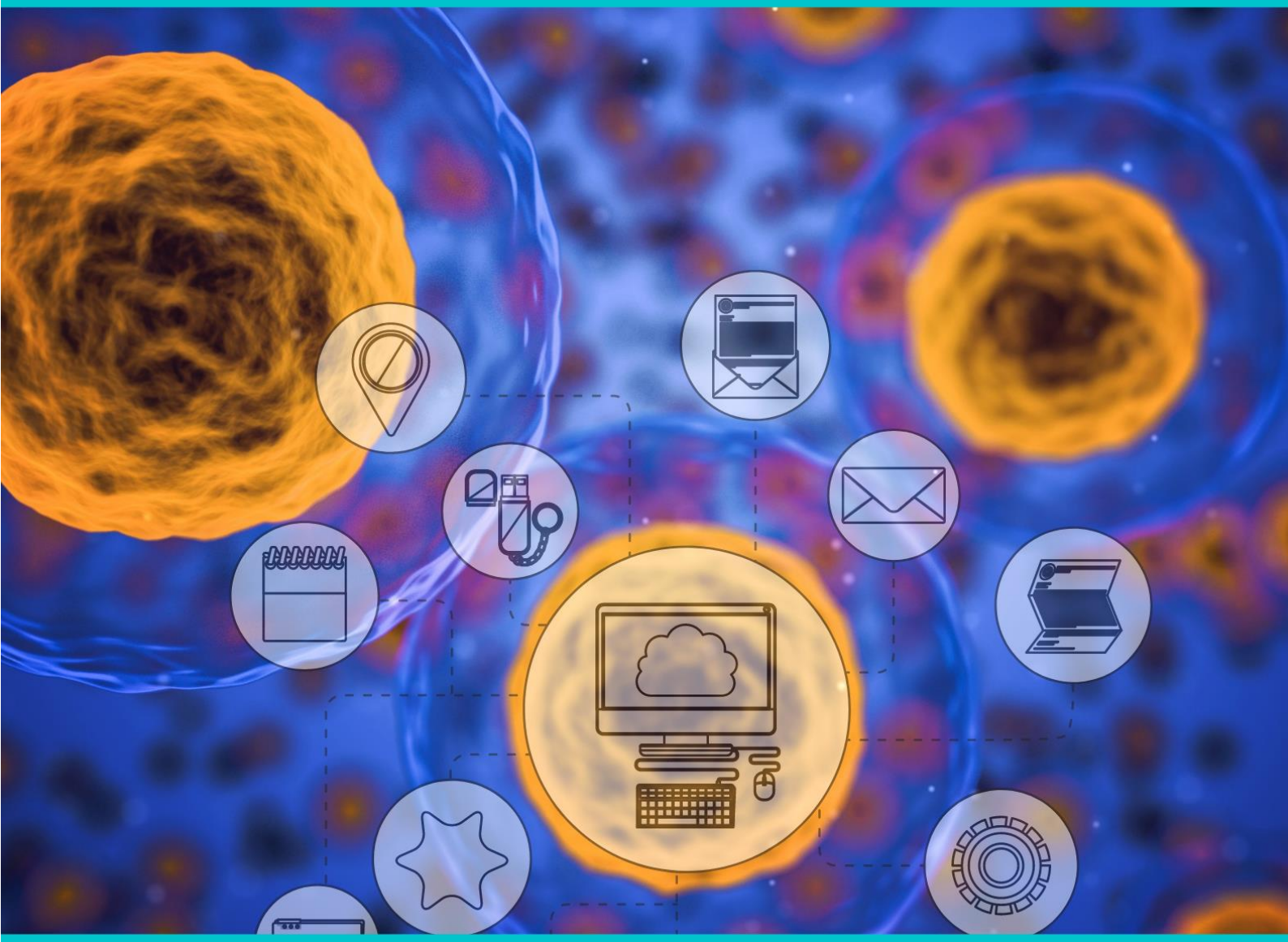
DIEGO LISBOA RIOS | ORG.

ACHADOS DE BIOMATEMÁTICA E A BIOINFORMÁTICA

NA SAÚDE HUMANA



Pantanal Editora
2021

Diego Lisboa Rios
Organizador

ACHADOS DE BIOMATEMÁTICA E A BIOINFORMÁTICA NA SAÚDE HUMANA

Pantanal Editora

2021

Editor Chefe: Prof. Dr. Alan Mario Zuffo
Editores Executivos: Prof. Dr. Jorge González Aguilera
Prof. Dr. Bruno Rodrigues de Oliveira

Diagramação: A editora
Edição de Arte: A editora. Imagens de capa e contra-capa: Canva.com
Revisão: O(s) autor(es), organizador(es) e a editora

Ficha Catalográfica

Dados Internacionais de Catalogação na Publicação (CIP)
(eDOC BRASIL, Belo Horizonte/MG)

A175 Achados de biomatemática e a bioinformática na saúde humana / Organizador Diego Lisboa Rios. – Nova Xavantina, MT: Pantanal, 2021. 85p.

Formato: PDF
Requisitos de sistema: Adobe Acrobat Reader
Modo de acesso: World Wide Web
ISBN 978-65-88319-54-3
DOI https://doi.org/10.46420/9786588319543

1. Matemática. 2. Informação. 3. Tecnologia. I. Rios, Diego Lisboa. II. Título.

CDD 510

Elaborado por Maurício Amormino Júnior – CRB6/2422

Pantanal Editora
Rua Abaete, 83, Sala B, Centro. CEP: 78690-000. Nova Xavantina – Mato Grosso – Brasil.
Telefone (66) 99682-4165 (Whatsapp).
https://www.editorapantanal.com.br
contato@editorapantanal.com.br

APRESENTAÇÃO

Na era da informação, os avanços tecnológicos ditaram uma nova ordem mundial, determinando novos princípios, alterando antigos conceitos e criando novas áreas do conhecimento humano. Diante disto, as ciências existentes tiveram que se apropriar aos princípios tecnológicos, informatizando-se para poder acompanhar o avanço que a computação propiciou, surgindo a Bioinformática. A palavra "Bioinformática" foi cunhada inicialmente por Pauline Hogeweg em 1979 para estudos de processos de informática em estudos de biologia sistemacional. Há na literatura distintas interpretações sobre a definição de Bioinformática. Na sua definição ampla, a Bioinformática envolve a aplicação de Tecnologias de Informação e de Comunicação (TIC) nas análises de qualquer área da Biologia. De maneira mais restrita, a Bioinformática é a aplicação de informática aos experimentos de Biologia Molecular, ou mais especificamente no manejo da grande quantidade de dados gerados no sequenciamento de DNA, RNA e Genômica, em especial para auxiliar aminoácidos.

Com o desenvolvimento da Bioinformática, suas novas descobertas e os problemas epistemológicos, surge a necessidade de se elaborar modelos matemáticos que possam assumir hipóteses com relação ao fenômeno estudado. São vários os modelos identificados nos problemas computacionais e biológicos, o modelo de Malthus e Verhulst, elaborados para descrever o crescimento de uma população sendo que cada um tem suas próprias limitações em considerar o meio analisado, ambos fazem aproximações na descrição dos fenômenos observados, não chega a ser um fator exato, visto que a biologia é uma ciência que possui componentes complexos.

São vários os processos biológicos que inspiram novos métodos, teorias e técnicas matemáticas. Por exemplo, os algoritmos genéticos que se inspiram nos processos biológicos de seleção, mutação e recombinação, máximos e mínimos de funções de muitas variáveis como as redes neurais que permitem imitar o funcionamento das redes de neurônios. Essa união entre a matemática e as ciências biológicas tem ajudado a desenvolver suas próprias áreas, os sistemas dinâmicos em tempo discreto e em tempo contínuo, probabilidade, estatística e processos estocásticos, equações diferenciais ordinárias e as derivadas parciais, álgebra linear e teoria de grupos são mais exemplos de conteúdos de matemática que foram ganhando espaço através de problemas biológicos.

Uma das grandes dificuldades para o uso da Matemática pelos biólogos, e em menor medida pelos bioinfomatas, é a falta de compreensão entre os praticantes dos dois campos, com frequência, vemos muitos biólogos sem nenhum conhecimento matemático e matemáticos que não têm a mínima ideia do que seja Biologia, fazendo com que a colaboração e interação entre essas duas disciplinas se torne cada vez mais difícil. Profissionais capazes de fazer a ponte entre as duas áreas são raros e altamente valorizados, além da falta de capacitação de ambos, pois é incomum visualizarmos a oferta de disciplina para interação dessas duas áreas nos cursos de formação. Desse modo percebemos que não é comum ver biólogos

utilizando números e fazendo cálculos, nem matemáticos que passam horas admirando a natureza, a distância entre esses dois tipos de disciplinas até existem, mas estão longe de ser distintas, um número cada vez maior de perguntas do mundo biológico está encontrando respostas no universo matemático, fazendo com que a disciplina de Matemática, que era conhecida como um bicho-de-sete-cabeças, pudesse se reinventar e combinar com muitas outras disciplinas, de modo que com essa interdisciplinaridade possa facilitar a aprendizagem de conteúdos matemáticos e biológicos entre outras áreas da educação.

Também se defende que Bioinformática aplica os princípios da Ciência da Informação para interpretar dados biológicos, enquanto a Biologia Computacional aplica os algoritmos matemáticos e computacionais aos experimentos biológicos. Os recursos fundamentais da bioinformática são os programas de computadores e os bancos de dados disponíveis na internet, ação fundamental para a análise de sequências de DNA e proteínas. Esta ferramenta é capaz de promover o aumento da velocidade na análise de sequências de DNAs de diferentes fontes, na comparação de variabilidades e na previsão de resultados de análises.

A bioinformática está sendo utilizada em diversas áreas como, a construção de banco de dados e a mineração de dados com o propósito de tratar esses dados biológicos brutos. A bioinformática se estabeleceu como uma nova área do conhecimento, graças a progressiva necessidade de desenvolver programas computacionais que permitam identificar sequências de genes, predizer a configuração tridimensional de proteínas, distinguir inibidores de enzimas, organizar e relacionar informação biológica, classificar proteínas homólogas, determinar árvores filogenéticas, analisar experimentos de expressão gênica, design de drogas entre outras.

Anteriormente ao surgimento da bioinformática, o sequenciamento de DNA era realizado manualmente, demandando dos sequenciadores muito tempo de trabalho. Além disso, com o tempo, houve um aumento na quantidade de dados, surgindo assim, a necessidade de manter esses dados acessíveis e organizados. Nessa circunstância, a bioinformática foi desenvolvida para atender, num curto espaço de tempo, essa necessidade.

O presente trabalho teve como objetivo contextualizar o uso de algoritmos matemáticos pela bioinformática, explicitando seus conceitos e avanços nas pesquisas dos autores envolvidos, abordando também a importância destes para a saúde.

SUMÁRIO

Capítulo I

O uso de softwares com algoritmos matemáticos em análises de metatranscriptoma: o exponencial impacto do *big data* na saúde humana

10.46420/9786588319543cap1

Diego Lisboa Rios[1*]
Silvia de Siqueira Costa[2]
Thiago Araújo Andrade[3]
Paula Margarita Salazar Torres[4]
Lucas Roberto da Silva[5]
Pedro Gontijo Carneiro[6]
Kerley dos Santos Alves [7]
Welington Ribeiro Aquino Marques[8]
João Batista Matos Júnior[9]
Fabyola Antunes Gonçalves Souza[10]

INTRODUÇÃO

Para além do corpo humano, comunidades de microrganismos são encontradas em diversos outros ambientes, como oceano, solo, plantas e outros animais (Li et al., 2017). Cada uma dessas comunidades microbianas tem suas próprias complexidades, diversidades e ecossistemas. Os membros das comunidades interagem entre si e cooperam com seus ambientes ou hospedeiros. Recentemente, com o desenvolvimento de tecnologias de sequenciamento que permitem aos pesquisadores estudar genomas da microbiota (também conhecido como microbioma), se descobriu que a microbiota humana está intimamente relacionada a várias doenças. Uma das áreas mais críticas é a pesquisa do microbioma humano associada a doenças humanas bem conhecidas, como obesidade, doença inflamatória intestinal (DII) e gêmeos magros ou obesos (Chen et al., 2018). Além disso, evidências crescentes indicam que a microbiota

[1] Universidade Federal de Minas Gerais (UFMG), Belo Horizonte - MG.
[2] Universidade José do Rosário Vellano (UNIFENAS), Divinópolis - MG.
[3] Universidade Federal de Minas Gerais (UFMG), Belo Horizonte - MG.
[4] Universidade Federal de São João del-Rei (UFSJ), Divinópolis - MG.
[5] Universidade Federal de São João del-Rei (UFSJ), Divinópolis - MG.
[6] Universidade Federal de Minas Gerais (UFMG), Belo Horizonte - MG.
[7] Universidade Federal de Ouro Preto (UFOP), Ouro Preto - MG.
[8] Universidade Federal de Ouro Preto (UFOP), Ouro Preto - MG.
[9] Instituto Federal do Acre(IFAC), Rio Branco - AC.
[10] Universidade Federal de Ouro Preto (UFOP), Ouro Preto - MG.
[*] Autor(a) correspondente: lisboa.zootec@gmail.com

dentro do corpo humano, especialmente o microbioma intestinal, desempenha um papel fundamental na fisiologia humana (Liu et al., 2016).

Atualmente, vários métodos são aplicados para inferir diferentes níveis de informação sobre o microbioma. A análise *Whole genome sequencing* (WGS) usa informações de todos os genes para identificar taxonomicamente a comunidade microbiana em nível de espécie ou linhagem. Já a análise *shotgun* do transcriptoma completo da amostra, chamado de metatranscriptoma, permite a observação de padrões na expressão gênica e funcional de toda a comunidades microbiana. No entanto, apesar das diversas vantagens em se utilizar esse método, algumas desvantagens são observadas, sendo esta muito carente de *softwares* específicos para análise dos dados (Bushmanova et al., 2019).

Na busca por padrão e banco de dados mais curado, vários estudos de microbioma com base populacional foram criados. O *Human Microbiome Project*, foca no estudo de comunidades microbianas que habitam o corpo humano de indivíduos saudáveis, com ênfase nas áreas nasal, oral, cutânea, gastrointestinal e urogenital. Já o *Interactive Human Microbiome Project* se concentra na compreensão das interações humano-microbioma, por meio de estudos longitudinais que reúnem múltiplos conjuntos de dados ômicos do microbioma humano (Nurk et al., 2017). Além disso, outro projeto chamado Metagenômica do Trato Intestinal Humano (MetaHIT) se baseia na compreensão da relação entre a microbiota intestinal humana e a os benefícios à saúde (Ugarte et al., 2018). Não obstante, o MetaHIT também estuda obesidade e DII. Por fim, o *Earth Microbiome Project* (EMP) se concentra na caracterização da diversidade, distribuição e estrutura dos ecossistemas microbianos em todo o planeta e já reuniu mais de 30.000 amostras de diversos ecossistemas, incluindo humanos, animais, plantas terrestres, marinhas, ambientes reconstruídos e entre outros (Bushmanova et al., 2019). EMP é um dos projetos pioneiros de microbioma a definir alguns protocolos padrão para outros estudos de microbioma.

Com as limitações crescentes na compreensão dos mecanismos de um microbioma individual, e em escala global, além das dificuldades associadas à cultura de espécies microbianas individuais, a metatranscriptômica têm sido usada com mais frequência em estudos recentes (Tarazona et al., 2015). Reconhecendo a importância dos estudos de microbioma para a saúde humana e além, nos esforçamos para fornecer uma revisão abrangente das tecnologias, algoritmos e *softwares* na análise metatranscriptômica existentes, especificamente os métodos de análise de dados. Esperamos que as informações fornecidas ajudem mais pesquisadores a identificar as ferramentas apropriadas para seus estudos de microbioma em seus respectivos projetos.

REVISÃO DE LITERATURA

Análise bioinformática de dados de sequenciamento metatranscriptômico

Em virtude da complexidade do microbioma, o sequenciamento de alto rendimento na forma de *reads* relativamente curtas geralmente são geradas a partir da tecnologia de sequenciamento Illumina. Tem sido mais frequentemente aplicado para estudos de metatranscriptoma, particularmente quando são necessárias várias amostras e cobertura profunda, como nos casos de estudos de expressão gênica diferencial. Uma vez que, *a priori*, a maioria das informações sobre as amostras são desconhecidas, assim como sua composição microbiana, abundância relativa de membros da comunidade, tamanhos do genoma e expressão relativa dentro e entre os genomas, não é trivial encontrar parâmetros experimentais corretos, como profundidade de sequenciamento para metatranscriptômica. Embora o sequenciamento de reads mais longas possa produzir mRNAs completos ou quase completos que podem ajudar a discriminar entre diferentes isoformas (Celaj et al., 2014), e fornecer trechos mais longos de sequência para pesquisas de similaridade e às várias tecnologias de leitura longa atualmente apenas desempenham um papel de apoio que não estão sendo ativamente usadas sozinhas para estudos de metatranscriptoma. Nessa revisão, nos concentramos nas ferramentas e fluxos de trabalho disponíveis para processamento e análise de dados de metatranscriptoma, que se concentram em dados de reads curtas.

Pré-Processamento

Semelhante a outros conjuntos de dados NGS, uma das primeiras etapas no processamento de dados de RNASeq é executar o Controle de Qualidade (QC) e remover ou aparar leituras falsas ou errôneas. São várias as ferramentas para fazer o QC das reads de RNASeq (Tabela 1).

Tabela 1. Softwares de pré-processamento e de análise da qualidade dos dados**.**

Software	Função	Autor
FastQC	Entre as mais utilizadas atualmente, usadas para *reads* curtas derivadas de sequenciadores Illumina.	(Andrews, 2010)
FaQCs	Entre as mais utilizadas atualmente, usadas para *reads* curtas derivadas de sequenciadores Illumina.	(Lo et al., 2014)
Fastp	Entre as mais utilizadas atualmente, usadas para *reads* curtas derivadas de sequenciadores Illumina.	(Chen et al., 2018)
Trimmomatic	A mais utilizada atualmente, usadas para *reads* curtas derivadas de sequenciadores Illumina.	(Bolger et al., 2014)

SortMeRNA	Remoção de rRNAs.	(Kopylova et al., 2012)*
Barrnap	Remoção de rRNAs.	(Seemann, 2014) *
BMTagger	Pesquisam *kmers* humanos específicos nas *reads*.	(Rotmistrovsky et al., 2011)
-	*Reads* podem ser removidas usando métodos tradicionais de mapeamento que marcam e removem *reads* que mapeadas no genoma humano.	(Li et al., 2017)

* Uma das etapas mais importantes que devem ser levadas em consideração é a remoção ou depleção física dos transcritos de RNA ribossômico altamente abundante (rRNA) das amostras, pois eles geralmente representam mais de 90% de todos os dados gerados no sequenciamento, geralmente removidos usando técnicas moleculares antes do sequenciamento, mas sua abundância nas amostras resulta em uma certa quantidade de contaminação por nas *reads*.

A montagem de novo

As leituras pré-processadas e de alta qualidade agora podem ser montadas em transcrições putativas, *contigs,* usando montadores *de novo*. Dado que a maioria dos microbiomas não é adequado para o alinhamento caracterizado pelos genomas de referência, os montadores *de novo* fornecem uma solução de para a referência, representando segmentos de genoma mais longos e expressos que podem fornecer um conjunto de genes de referência. Isso fornece aos pesquisadores a capacidade de encontrar homólogos de maneira mais direta, estabelecer origem taxonômica e servir como referência para mapeamento e análise de expressão (Quadro 1).

Quadro 1. Montadores *de novo* atuais**.**

Software	**Função**	**Autor**
MEGAHIT	Montadores metagenômicos projetados para lidar com metagenomas complexos que podem compartilhar alguma semelhança de sequência em regiões altamente conservadas, mas podem variar em termos de abundância relativa dentro do microbioma e também podem abrigar variação populacional (nível de deformação).	(Li et al., 2015)*
IDBA-UD		(Peng et al., 2012)*
metaSPAdes		(Nurk et al., 2017)*
Trans-ABySS	Montadores que tentaram resolver os problemas na análise de metatranscriptomas, mas foram originalmente projetados para montar transcriptoma de um único organismo.	(Robertson et al., 2010)**
Trinity		(Grabherr et al., 2011)**
BinPacker		(Liu et al., 2016)**

Oásis		(Schulz et al., 2012)**
SOAPdenovo-Trans		(Xie et al., 2014)**
IDBA-Tran		(Peng et al., 2012)**
rnaSPAdes		(Bushmanova et al., 2019)**
IDBA-MT	Montadores *de novo* projetados especificamente para metatranscriptomas que levam em conta as características únicas dos transcritos e a natureza complexa das comunidades microbianas.	(Leung et al., 2013)***
IDBA-MTP		(Leung et al., 2014)****
Transcript Assembly Graph (TAG		(Ye et al., 2016)*****

* No entanto, a eficácia desses montadores na reconstrução de transcritos que possuem peculiaridades próprias, como íntrons e éxons, diferentes isoformas e RNAs não codificadores mais curtos (ncRNA), raramente são testados; portanto, é recomendável que se use montadores específicos para dados metagenômica em conjuntos de dados de metatranscriptoma.
** Tentaram explicar os problemas no sequenciamento de transcriptomas, mas foram originalmente projetados para montar transcrições de um único organismo. Apesar de seu design para conjuntos de dados transcriptômicos e não metatranscriptômicos, as comparações entre alguns montadores mostraram que, em geral, os montadores testados Oases, Trinity e Metavelvet melhoraram o número de genes anotados nos contigs resultantes, com o montador Trinity superando os demais (Celaj et al., 2014).
*** Baseado no IDBA-UD, usa-se múltiplos valores em um gráfico *de Bruijn*, enquanto considera-se os recursos associados aos mRNAs, como profundidade de sequência desigual e padrões comuns de repetição em diferentes mRNAs, diminuindo assim a taxa de erros de montagem (Leung et al., 2013).
**** Derivado do IDBA-MT para poder montar mRNAs de baixa expressão. Ele usa as informações de sequências de proteínas conhecidas para orientar a montagem, começando com *k-mers* menores para construir sequências de mRNA que são então incluídas com base na sua similaridade (Leung et al., 2014).
*****. Montador que, comparativamente, também usa um gráfico *de Bruijn*, mas para montar o metagenoma correspondente, que é usado como referência para posteriormente mapear as leituras do metatranscriptoma e reconstruir sequências de mRNA. É ineficaz para uso em microbiomas que, além de bactéria, também englobam alguns fungos. Além disso, existe a suposição implícita entre o metagenoma da comunidade para que todos os genes expressos possam ser mapeados (Ye et al., 2016).

O estado atual da montagem *de novo* para conjuntos de dados metatranscriptômicos ainda está em estágios iniciais. Apenas algumas ferramentas foram desenvolvidas especificamente para a metatranscriptômica, mas sua eficácia em diversos conjuntos de dados não foi testada e seu hardware ou requisitos de memória em uma variedade de complexidades da comunidade e volume de dados também não foram rigorosamente estabelecidos.

Taxonomia dos transcritos

Semelhante ao perfil taxonômico, que é frequentemente realizado com dados metagenômicos de *shotgun*, pode-se usar o mesmo conjunto de ferramentas para realizar atribuições taxonômicas baseadas em *reads* ou *contigs*, a fim de entender quais os organismos estão expressando ativamente em determinada amostra. Um método separado e distinto é focar apenas nos rRNAs para avaliar membros ativos de uma comunidade, embora, como mencionado acima, eles sejam frequentemente removidos (tanto nos protocolos de laboratório quanto no pré-processamento dos dados brutos).

Ferramentas de classificação de taxonomia baseadas em *reads*, como Kraken (Wood e Salzberg, 2014), GOTTCHA (Freitas et al., 2015), MetaPhlan2 (Truong et al., 2015) etc. podem ser usadas para metatranscriptomas (Neves et al., 2017). Como essas ferramentas funcionam em sequências curtas e se baseiam em combinações de nucleotídeos, seu uso efetivo é limitado a microrganismos, mas cujos os membros possuam vizinhos próximos às sequência existentes nos bancos de dados que foram montadas em *contigs* mais longos e possivelmente transcritos completos podem ser usados por várias ferramentas, como Centrífuge (Kim et al., 2016) e Kraken 2 (Wood et al., 2014), para identificar potencialmente uma maior proporção dos membros da comunidade sequenciados.

As atribuições taxonômicas que utilizam *reads* ou regiões de codificação previstas têm um grande número de limitações, incluindo os algoritmos necessários para processar grandes volumes de dados ou acomodar sequências curtas e a escassez de citações nos bancos de dados de referência. Para compor essas questões, está o fato de que a maioria das ferramentas de bioinformática utilizam-se de apenas um subconjunto de genomas disponíveis ou se concentra em certos organismos. Por exemplo, muitas ferramentas não possuem eucariotos como parte de seus bancos de dados. Houveram alguns esforços recentes, como o desenvolvimento de novas ferramentas e melhorias nas ferramentas existentes para incluir genomas eucarióticos em seus bancos de dados, como MetaPhlan2 e Kaiju (Truong et al., 2015), porém sua eficácia na classificação de eucariotos é desconhecida. Além disso, muitas vezes é difícil discernir acerca de baixa abundância de acertos falsos positivos, o que é um problema inato nos estudos de microbioma. A falta de conhecimento sobre a diversidade microbiana geral ou em qualquer sistema biológico também pode limitar a utilização das ferramentas de classificação da taxonomia.

Anotação funcional

Um dos principais objetivos da metatranscriptômica é avaliar a atividade funcional de uma comunidade microbiana. Como os transcritos expressos representam um *proxy* para o fenótipo real, caracterizar a função dos transcritos é uma tarefa fundamental para a metatranscriptômica. A anotação funcional pode ser realizada usando *reads* ou *contigs* montados.

Os perfis funcionais baseados em *reads*, como MetaCLADE (Ugarte et al., 2018), HMM-GRASPx (Zhong et al., 2016) e UProC (Meinicke, 2015) utilizam bancos de dados específicos e exigem frames de leitura abertos previstos como entrada de outras ferramentas, como FragGeneScan (Rho et al., 2010). O MetaCLADE é uma das ferramentas mais recentes e usa um banco de dados que consiste em 2 milhões de modelos probabilísticos derivados de 15.000 domínios da Pfam, portanto centenas de modelos representando um único domínio, para abranger a diversidade de cada domínio na árvore da vida. Uma pesquisa nesse banco de dados resulta em um grande número de ocorrências por leitura que são filtradas com base em redundância, probabilidade e pontuação de *bits* (Ugarte et al., 2018).

De maneira alternativa, a anotação de genes pode ser realizada a partir de *contigs* montados. A anotação de transcritos montados é semelhante às anotações de genomas e metagenomas. A descoberta de genes usando programas como Prodigal (Hyatt et al., 2010) e FragGeneScan (Rho et al., 2010) é seguida por atribuição funcional com base em pesquisas de similaridade usando ferramentas como DIAMOND para pesquisar sua funcionalidade em bancos de dados, como KEGG, NCBI RefSeq, UniProt, entre outros (Buchfink et al., 2015). Outras ferramentas, pipelines e plataformas abrangem uma variedade de utilitários de bioinformática (incluindo busca e anotação de genes), como o Prokka (Seemann, 2014), EDGE Bioinformatics (Li et al., 2017) e MG-RAST (Wilke et al., 2016) que combinam várias pesquisas de similaridade em vários bancos de dados ou podem até unir a montagem, identificação de genes e anotação por meio de pesquisas de similaridade. Depois que as anotações são executadas, as funções enzimáticas também podem ser mapeadas para as vias metabólicas conhecidas, usando ferramentas como MinPath (Ye et al., 2009) ou iPath (Yamada et al., 2011).

Análise de expressão diferencial

Além da descrição simples de quem são os membros ativos e quais genes estão sendo expressos em um único momento ou amostra, estão os estudos de expressão diferencial de genes, nos quais a metatranscriptômica pode ser usada para comparar diferentes condições, parâmetros ambientais e seus efeitos na comunidade. Existem muitas ferramentas originalmente desenvolvidas para uso de genomas únicos que possam ser aproveitadas para estudos de expressão gênica diferencial metatranscriptômica. Essas ferramentas requerem como dados a abundância de entrada por gene (ou transcrito) e por amostra (representando a expressão sob uma condição específica ou um momento específico). A abundância pode ser obtida de várias maneiras, mas geralmente envolvem algumas formas de alinhamentos/mapeamentos das *reads* para um genoma de referência, um conjunto de referência ou um conjunto de genes de referência.

Os programas EdgeR (Robinson et al., 2010), DeSeq2 (Love et al., 2014) e LIMMA (Ritchie et al., 2015) são pacotes do R que são frequentemente usados, juntamente com as informações de abundância, para identificar genes que são estatisticamente diferencialmente expressos entre várias amostras. Da

mesma forma, ferramentas como Análise do Conjunto de Genes (GAGE) podem ser usadas para identificar caminhos enriquecidos em uma condição sobre outra (Li et al., 2017). Como as amostras de metatranscriptômica replicadas não são triviais comparadas aos estudos transcriptômicos com organismos isolados, métodos não paramétricos como a implementação no NOISeq (Tarazona et al., 2015) também devem ser considerados.

Existem peculiaridades nas análises metatranscriptômicas que tornam as análises de expressão diferencial bastante desafiadoras, principalmente o resultado do sequenciamento de uma grande diversidade de transcritos (de uma ampla variedade de organismos). Problemas como genes compartilhados entre organismos intimamente relacionados e variação na composição taxonômica dos transcritos podem resultar em avaliação incorreta dos perfis de expressão gênica. Recentemente, foi proposto um método de normalização que pode minimizar a influência da diversidade taxonômica na amostra, normalizando os dados de contagem com base na composição taxonômica em diferentes amostras, porém esse método também é influenciado pela representação em bancos de dados taxonômicos (Klingenberg et al., 2017).

Pipelines – Fluxos de trabalho disponíveis para análise metatranscriptômica

A análise de um conjunto de dados de metatranscriptoma está repleta de etapas bioinformáticas com muitas opções de ferramentas para qualquer etapa. Quais etapas e ferramentas devem ser selecionadas são frequentemente escolhidas pelos objetivos do experimento, cujos os detalhes podem crescer em complexidade com base nas especificidades do estudo. No entanto, existem fluxos de trabalho em bioinformática, chamados de *pipelines* ou *workflow*, que visam otimizar parte dessa complexidade conectando várias ferramentas individuais que transformam *reads* brutas de sequenciamento e as processam, fornecendo arquivos de dados que representam os resultados das saídas, caracterizando identidades taxonômicas, genes funcionais e transcritos diferencialmente expressos. Os detalhes de oito fluxos de trabalho disponíveis, seus recursos e as ferramentas específicas de bioinformática que eles usam podem ser encontrados como um resumo na Tabela 2.

Tabela 2. *Pipelines* de metatranscriptômica e seus recursos.

		Baseado em *Reads*					Baseado em montagem		
		MetaTrans	**COMAN**	**FMAP**	**SAMSA2**	**HUMAnN2**	**SqueezeMeta**	**IMP**	**MOSCA**
Processamento	**QC**	✓	✓	✓	✓	×	✓	✓	✓
	Remoção das reads do hospedeiro	×	×	✓	×	×	×	✓	×
	Remove rRNA	✓	✓	×	✓	×	✓	✓	✓
Montagem *de novo*		×	×	×	×	×	✓	✓	✓
Binning		×	×	×	×	×	✓	✓	×
Perfil Taxônomico	***Reads***	✓	✓	×	✓	✓	×	×	×
	Contigs	×	×	×	×	×	✓	✓	✓
Anotação Funional	***Reads***	✓	✓	✓	✓	✓	×	×	×
	Contigs	×	×	×	×	×	✓	✓	✓
Analise *Pathway*		✓	✓	✓	×	✓	✓	✓	×
Precisa de Metagenoma		×	×	×	×	×	×	✓	×
Reporta resumos		×	×	×	×	×	×	✓	×
Interface *Web*		×	✓	×	×	×	×	×	×
Comparação de múltiplas amostras		✓	✓	✓	✓	✓	✓	×	✓

Expressão Diferencial	✓	✓	✓	✓	×	×	×	✓
Docker	×	×	×	×	✓	×	✓	✓
Conda	×	×	×	×	✓	×	✓	×
Suporta *reads* longas	×	×	×	×	×	✓	×	×
Repositório de código público.	✓	×	✓	✓	✓	✓	✓	✓

Em comparação às análises baseadas em leitura, os fluxos de trabalho baseados em montagem abrigam uma etapa analítica extra. As análises são reunidas primeiramente em *contigs* maiores, ajudando a reduzir o tamanho dos dados que precisam ser processados para análises posteriores e aumentando a quantidade contínua do comprimento das transcrições expressas.

Como um dos principais objetivos das análises de metatranscriptoma é obter uma quantificação relativa da expressão gênica, todos os fluxos de trabalho baseados em leitura e em montagem fornecem alguma forma de cobertura por gene ou métrica de abundância (por exemplo, contagem bruta por gene ou número de leituras por *kb* por milhão de leituras sequenciadas). Esses valores de abundância podem usar de ferramentas adicionais para comparar a expressão relativa de genes entre condições de crescimento ou durante experimentos ao longo do tempo, cujo o objetivo é frequentemente ajudar a entender quais genes e caminhos podem ser importantes para um fenótipo específico em estudo. Para esses tipos de estudos, muitas vezes são necessárias experiências com replicação para obter resultados estatisticamente significativos, portanto, as comparações relativas de abundância de genes costumam ser uma comparação entre muitas amostras diferentes que incluem várias repetições biológicas.

A disponibilidade de vários fluxos de trabalho permite que os usuários escolham o que é mais apropriado para analisar seu metatranscriptoma. Embora os usuários devam selecionar fluxos de trabalho idealmente com base na capacidade/funcionalidade e qualidade dos algoritmos/ferramentas utilizados, considerações adicionais podem incluir os requisitos de recursos computacionais que variam entre os fluxos de trabalho e a frequência de manutenção ou desenvolvimento ativo do código-fonte, que podem sofrer modificações frequentes à medida que novos avanços, ferramentas ou métodos continuam sendo desenvolvidos. A Tabela 3 é uma compilação desses fluxos de trabalho disponíveis e pode ser usada como um guia em potencial para escolher um fluxo de trabalho com base em fatores importantes para abordar

as perguntas de qualquer pesquisador. Por exemplo, se a análise de expressão diferencial é o objetivo de um estudo, a lista de fluxos de trabalho para escolher é limitada a cinco.

CONSIDERAÇÕES FINAIS

Atualmente, os estudos metatranscriptômicos são realizados principalmente com tecnologia de *reads* curtas (ou seja, Illumina), exigindo um grande número de ferramentas analíticas para ajudar em todos os aspectos da análise de dados. Nesta revisão, destacamos alguns dos principais métodos de análise de dados de metatranscriptômica, o que impacta diretamente no conhecimento humano da influência do microbioma na saúde humana. Algumas das ferramentas específicas de bioinformática usadas para realizar essas análises e alguns fluxos de trabalhos metatranscriptômicos mais complexos combinam várias dessas ferramentas para abordar questões biológicas com esforço mínimo de usuário menos experientes. Conquanto, para tornar as coisas mais complexas, novos avanços em tecnologias de sequenciamento continuarão impulsionando o desenvolvimento de novas ferramentas e fluxos de trabalho. Um *pipeline* ou *workflow* padrão, uma estrutura, deve ser aceita para comparar novas ferramentas, o que ajudaria no progresso de trabalho e possivelmente se uniria a fluxos de trabalho mais precisos e apropriados. Apesar de alguns dos problemas com a metatranscriptômica como método, o desenvolvimento contínuo de novas ferramentas e algoritmos para análise de dados metatranscriptômicos, juntamente com nossa crescente compreensão dos desafios apresentados por esses conjuntos de dados, fica claro que a próxima geração de ferramentas de metatranscriptômica são uma grande promessa para impulsionar e facilitar nossa compreensão da fração biologicamente ativa de microbiomas das vias relevantes envolvidas.

REFERÊNCIAS BIBLIOGRÁFICAS

Andrews S (2010). FastQC: a quality control tool for high throughput sequence data.

Bolger AM et al. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Buchfink B et al. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12: 59–60.

Bushmanova E et al. (2019). rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience 8.

Celaj A et al. (2014). Comparison of assembly algorithms for improving rate of metatranscriptomic functional annotation. Microbiome 2: 39. doi: 10.1186/2049-2618-2-39

Chen S et al. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34: i884–i890.

Freitas TAK et al. (2015). Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. Nucleic Acids Res. 43: e69.

Grabherr MG et al. et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–652.

Kim D et al. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 26: 1721–1729.

Klingenberg H et al. (2017). How to normalize metatranscriptomic count data for differential expression analysis. PeerJ 5: e3859.

Kopylova E et al. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28: 3211–3217.

Leung HCM et al. (2014). IDBA-MTP: a hybrid metatranscriptomic assembler based on protein information. Res. Comput. Mol. Biol. 22(5). doi: 10.1007/978-3-319-05269-4_12

Leung HCM et al. (2013). IDBA-MT: de novo assembler for metatranscriptomic data generated from next-generation sequencing technology. J. Comput. Biol. 20: 540–550.

Li D et al. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31: 1674–1676.

Li PE et al. (2017). Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. Nucleic Acids Res. 45: 67–80.

Liu J et al. (2016). BinPacker: packing-based de novo transcriptome assembly from RNA-seq data. PLoS Comput. Biol. 12: e1004772.

Lo C-C et al. (2014). Rapid evaluation and quality control of next generation sequencing data with FaQCs. BMC Bioinf. 15: 366.

Love MI et al. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15: 550.

Meinicke P (2015). UProC: tools for ultra-fast protein domain classification. Bioinformatics 31: 1382–1388. d

Neves ALA et al. (2017). Enhancing the resolution of rumen microbial classification from metatranscriptomic data using Kraken and Mothur. Front. Microbiol. 8: 2445.

Nurk S et al. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27: 824–834.

Peng Y et al. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28: 1420–1428.

Rho M et al. (2010). FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 38: e191.

Ritchie ME et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43: e47. doi: 10.1093/nar/gkv007

Robertson G et al. (2010). De novo assembly and analysis of RNA-seq data. Nat. Methods 7: 909–912.

Robinson MD et al. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140.

Rotmistrovsky K et al. (2011). BMTagger: best match tagger for removing human reads from metagenomics datasets.

Schulz MH et al. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28: 1086–1092.

Seemann T (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics 30: 2068–2069.

Tarazona S et al. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. 43: e140.

Truong DT et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods 12: 902–903.

Ugarte A et al. (2018). A multi-source domain annotation pipeline for quantitative metagenomic and metatranscriptomic functional profiling. Microbiome 6: 149.

Wilke A et al. (2016). The MG-RAST metagenomics database and portal in 2015. Nucleic Acids Res. 44: D590–D594.

Wood DE et al. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15: R46.

Xie Y et al. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30: 1660–1666.

Yamada T et al. (2011). iPath2.0: interactive pathway explorer. Nucleic Acids Res. 39: W412–W415.

Ye Y et al. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS Comput. Biol. 5: e1000465.

Ye Y et al. (2016). Utilizing de Bruijn graph of metagenome assembly for metatranscriptome analysis. Bioinformatics 32: 1001–1008.

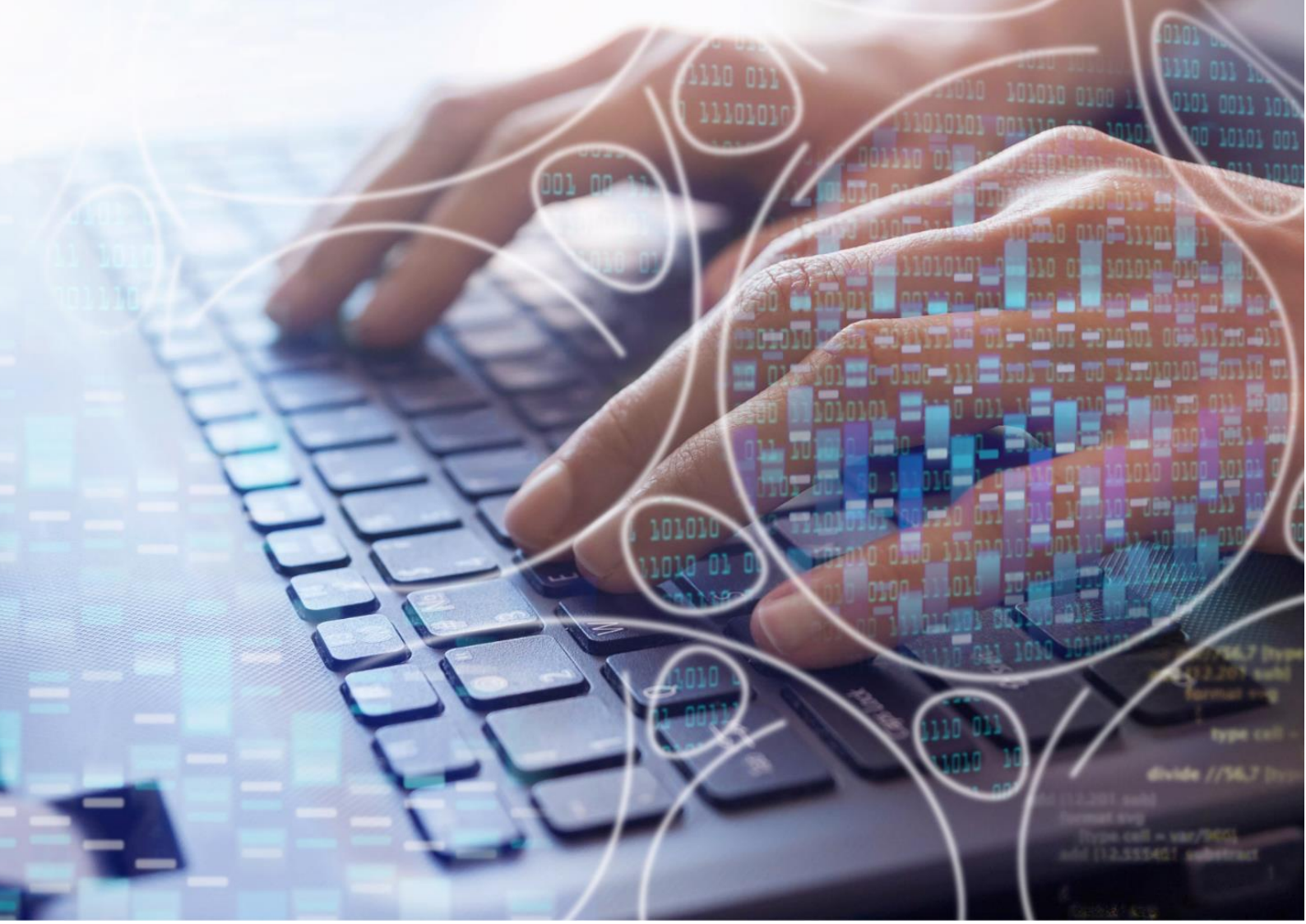
Zhong C et al. (2016). Metagenome and metatranscriptome analyses using protein family profiles. PLoS Comput. Biol. 12: e1004991.

ÍNDICE REMISSIVO

SOBRE O ORGANIZADOR

Diego Lisboa Rios

Graduado em Ciências Biológicas, Matemática, Análise e Desenvolvimento de Sistemas e Zootecnia, com formação complementar em Administração e Marketing, pela Universidade Federal de Minas Gerais - UFMG. Pós-graduação *lato sensu* em Gestão e Docência no Ensino Superior pela Faculdade de Teologia e Ciências Humanas - FATECH. Mestrado em Biologia Molecular no curso de Pós-Graduação em Ciências Farmacêuticas pela Universidade Federal de São João del Rei - UFSJ, onde desenvolveu projetos relacionados a síntese translesão do DNA, expressão e purificação de proteínas recombinantes. Doutorado pela Pós-Graduação em Genética, Área de Concentração: Genética Molecular, de Microrganismos e Biotecnologia, na Universidade Federal de Minas Gerais - UFMG, onde desenvolveu projeto de metatranscriptoma das bebidas de kefir de água e leite do Brasil e Argentina. Hoje residente de pós-doutorado na Pós-Graduação em Microbiologia da UFMG. Fluente na língua inglesa, possui conhecimento avançado em informática, Big Data, programação na linguagem Python, R e MySQL.



ISBN 978-658831954-3

Pantanal Editora
Rua Abaete, 83, Sala B, Centro. CEP: 78690-000
Nova Xavantina – Mato Grosso – Brasil
Telefone (66) 99682-4165 (Whatsapp)
https://www.editorapantanal.com.br
contato@editorapantanal.com.br