UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FABIANA RODRIGUES DE GÓES

GENEFINDER-MG: UM PIPELINE PARA A PREDIÇÃO DE GENES EM DADOS METAGENÔMICOS.

BELÉM – PARÁ

2016

FABIANA RODRIGUES DE GÓES

GENEFINDER-MG: UM PIPELINE PARA A PREDIÇÃO DE GENES EM DADOS METAGENÔMICOS.

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas e Naturais da Universidade Federal do Pará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Sistemas de Computação

Orientador: Prof. Dr. Ronnie Cley de Oliveira Alves

BELÉM – PARÁ

2016

FABIANA RODRIGUES DE GÓES

GENEFINDER-MG: UM PIPELINE PARA A PREDIÇÃO DE GENES EM DADOS METAGENÔMICOS.

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas e Naturais da Universidade Federal do Pará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Sistemas de Computação

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Ronnie Cley de Oliveira Alves (Orientador)
Lab. de Ciência da Computação, Robótica e Microeletrônica de Montpellier – LIRMM

Prof. Dr. Jefferson Magalhães de Morais (Membro Interno)
Universidade Federal do Pará – UFPA

Prof. Dr. Guilherme Corrêa de Oliveira (Membro Externo)
Instituto Tecnológico Vale Desenvolvimento Sustentável – ITVDS

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema de Bibliotecas da UFPA

Góes, Fabiana Rodrigues de, 1992-
Genefinder-mg: um pipeline para a predição de genes em dados metagenômicos. / Fabiana Rodrigues de Góes. - 2016.

Orientador: Ronnie Cley de Oliveira Alves.
Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Ciências Exatas e Naturais, Programa de Pós-Graduação em Ciência da Computação, Belém, 2016.

1. Bioinformática. 2. Matagenômica. 3. Aprendizado de máquina. 4. Predição de genes. 5. Random forest. I. Título.

CDD 22. ed. 570.285

Aos meus pais, aqueles dos quais eu herdei os meus genes.

AGRADECIMENTOS

Primeiramente, agradeço a Deus por toda a proteção e as todas bênçãos concedidas a mim ao longo da minha vida. Em momentos críticos durante a caminhada da pós-graduação busquei força Nele e, assim, consegui seguir em busca dos meus objetivos.

Aos meus amados pais Maria e Góes por terem se dedicado extremamente a mim, desde o meu nascimento. Pais que trabalharam arduamente para que eu sempre dedicasse exclusivamente aos estudos. Mais uma etapa se conclui graças não somente ao meu esforço, mas sim ao de nós três.

À minha família, em especial aos meu tios José, Maria e Arlete, e as minha primas Jéssica e Priscilla pelo amor e apoio destinados a mim ao longo desses anos. À minha família carioca, da qual eu recebi muito carinho, apoio e incentivo.

Aos meus amados amigos Betinha, Diego, Paulinho, Biel, Kamila e Renata pelo amor e fidelidade de longos anos. Aos meus amigos de Nárnia que também me deram muita força em momentos difíceis e encheram o meu coração de alegria. À minha amiga Alessandra por todas as palavras encorajadoras e pelo carinho.

Ao meu namorado e amigo Léo pela compreensão, companheirismo e amor. A querida tia Laudia, por todo o carinho e amizade.

Ao meu orientador Ronnie Alves por ter confiado e estabelecido uma parceria comigo. A ele devo a minha gratidão por tudo o que aprendi, ao longo de três anos, pelo seu comprmetimento e esforço para fazer com que esse trabalho fosse possível. A minha querida professora Regiane Kawasaki por ser uma excelente profissional e ter me recebido de braços abertos quando eu decidi escolher a Bioinformática. Aos professores do Programa de Pós-graduação em Ciência da Computação da Universidade Federal do Pará que contribuíram para a minha formação.

Ao meu amigo Leandro que, ao longo de três anos, compartilhou comigo momentos de altos e baixos e diversas experiências marivilhosas.

Às minhas queridas amigas da computação Clarice, Dani e Lílian, por terem compartilhado momentos de dificuldades, mas principalmente de alegrias. Às minhas doces amigas da Eng. da Computação Mylena e Lua, que me ajudaram em vários momentos.

Aos meus eternos amigos amigos de turma, com os quais eu compartilhei muitas risadas e muitos momentos bons e dos quais eu recebi muito carinho e proteção. Aos meus amigos Lipí e Téo pelo carinho que têm por mim.

As pessoas que desde a minha infância cultivaram um carinho muito grande por mim: Elenice, Maria, Nazaré, Bosco e Margarida "*in memorian*".

A todos aqueles que mesmo não tendo citado o nome foram importantes para o êxito deste trabalho.

“Nunca deixe ninguém dizer que você não pode fazer alguma coisa. Se você tem um sonho, deve correr atrás dele. As pessoas não vencem e dizem que você também não vai vencer, então, se você quer realmente alguma coisa, corra atrás e ponto final.”

(Will Smith)

RESUMO

A Bioinformática, através da sua multidisciplinaridade, continua a proporcionar um grande avanço nas análises e pesquisas desenvolvidas a partir de dados das diversas áreas da Biologia. A Metagenômica é o estudo dos metagenomas, a qual provê um melhor entendimento de como os micro-organismos se relacionam em uma comunidade e como esta contribui para o ambiente no qual habita. Recursos computacionais têm grande importância em projetos deste contexto devido a complexidade e grande quantidade de dados gerados pelas tecnologias de sequênciamento de DNA. Uma forma tradicional de se extrair informações dos (meta)genomas ocorre através da predição de genes. A identificação dos genes possibilita que, a partir dela, outros estudos mais específicos sejam elaborados, sendo assim, uma etapa de grande importância para o estudo dos (meta)genomas. Este trabalho apresenta uma abordagem baseada em aprendizado de máquina supervisionado para o problema de predição de genes em experimentos metagenômicos. O pipeline desenvolvido utiliza uma abordagem de aprendizado *ensemble* para a realização da classificação de regiões codificadoras em sequências de DNA de organismos procariotos. Para observar o desempenho do *pipeline*, avaliações do modelo e comparações com outras quatro ferramentas foram executadas. Os resutados obtidos em todos os testes foram significativos em termos de especificidade a ponto de superar os resultados de todas as ferramentas selecionadas para compor a comparação.

Palavras-chave: Bioinformática. Metagenômica. predição de genes. Aprendizado de Máquina. Random Forest.

ABSTRACT

Bioinformatics, through its multidisciplinarity, continues to provide a great progress in the analyses and research developed from the data of the multiple areas of Biology. The Metagenomics is the study of the metagenomes, that provide a better knowledge about how the microorganisms related to each other in a community and how this community contributes to the environment in which inhabits. Computer resources have a big importance in such projects due to the complexity and the large amount of data generated by the DNA sequencing technologies. One traditional way to extract information from (meta)genomes occurs through gene prediction. The genes identification allows that other more specific studies be elaborated, being a very important step in the (meta)genomes study. This work presents a approach based in machine learning supervised for the problem of gene prediction in metagenomic experiments. The developed pipeline utilizes the ensemble learning approach for the classification of coding regions in DNA sequences of Prokaryote organisms. In order to observe the performance of the pipeline, evaluations of the model and comparisons with four other tools were executed

Keywords: Bioinformatics. Metagenomics. gene prediction. Machine Learning. Random Forest.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
BMP	*Brazilian Microbiome Project*
DNA	*Deoxyribonucleic acid*
KNN	*K-Nearest Neighbor*
NCBI	*National Center for Biotechnology Information*
ORF	*Open Read Frame*
OOB	*Out-of-bag*
PB	Pares de base
RF	*Random Forest*
SVM	*Support Vector Machine*

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTO

Junto com os avanços das pesquisas na área biológica surgiu a necessidade do desenvolvimento de estudos inovadores relacionados a processamento e gerencimaneto dos dados gerados. Na década de 90 a Bioinformática foi definida como uma disciplina independente (OUZOUNIS, 2012), fato influenciado pelo aumento na quantidade de dados produzidos por projetos inovadores e a necessidade do apoio dos recursos computacionais à eles. A Bioinformática, definida como a manipulação computacional e processamento da informação genética, tornou-se uma das áreas mais visíveis da ciência moderna (OUZOUNIS; VALENCIA, 2003).

No final da segunda metade do século 20, a revolução da Biologia Molecular e da Genômica uniu o conhecimento fisiológico com a compreensão da base genética dos microrganismos (HANDELSMAN et al., 2007, p. 13). Contudo, mesmo tendo contribuído para o avanço nos estudos dos seres vivos, as técnicas providas pela Genômica demonstraram com o tempo não fornecer subsídios adequados para a análises de comunidades microbianas. Devido a este fato, houve a necessidade de uma mudança de paradigma sobre o estudo dos microrganismos. A Metagenômica é uma derivação da genômica microbiana convencional, sendo sua principal diferença o desligamento da exigência da obtenção de culturas puras para o sequenciamento (KUNIN et al., 2008). Desta forma, segundo Kumar et al. (2015), ferramentas de caráter metagenômico permitiram o acesso, como nunca visto, a comunidades microbianas naturais, bem como, as suas respectivas potenciais atividades.

Hoje é evidente que os micróbios, como comunidades, são atores fundamentais na manutenção da estabilidade ambiental (HANDELSMAN et al., 2007, p. 12). O advento do estudo dos metagenomas trouxe um considerável avanço para áreas como a Biotecnologia, Biomedicina, Farmacologia, dentre outras; e de acordo com Thomas et al. (2012), o campo da Metegenômica tem sido responsável por grande parte dos avanços nos estudos da ecologia, evolução e diversidade microbiana nos últimos anos. Uma área em expansão é a utilização industrial de microrganismos para a produção de antibióticos, enzimas e outros compostos bioativos. A demanda para a produção comercial de enzimas que são utilizadas em processos industriais de grande escala está crescendo rapidamente (BASHIR et al., 2014), sendo assim a Metagenômica constitui-se como um fator muito importante para esse tipo de pesquisa.

Um dos maiores objetivos de um projeto metagenômico, além de conhecer melhor as características dos organismos que estão presentes em uma comunidade microbiana, é ter domínio da interação destes e capacidade da comunidade como um todo. Um procedimento importante para isto é a realização da predição de genes pois, de acordo com Filippo et al. (2012), é um passo fundamental que permite a anotação e caracterização do potencial funcional da comunidade que está sob investigação. Segundo Mathé et al. (2002), este procedimento pode ser realizado através de duas abordagens: extrínseca (identificação por homologia) ou intrínseca (predição *ab initio*). Porém, Liu et al. (2013) afirma que a identificação de genes é mais complicada em metagenomas do que em genomas isolados pois, neste contexto muitos dos fragmentos gerados são curtos e, sendo assim, em alguns casos não é possível realizar uma boa montagem das sequênicas ou, quando possível, são gerados *contigs* de comprimento pequeno.

Segundo Wooley et al. (2010), embora a Metegenômica seja uma à ciência relativamente recente, nos últimos anos têm-se visto uma explosão de métodos computacionais aplicados à pesquisa baseada nesta área da microbiologia. No contexto da identificação de genes, tanto na predição por homologia como na *ab initio*, este fato não é diferente. O Aprendizado de Máquina (AM), é uma vertente forte na identificação de genes, bem como em outras áreas da Bioinformática, sendo a base da técnica *ab initio*, a qual utiliza métodos computacionais e padrões de sequências para realizar a predição.

De acordo com Domingos (2012), algoritmos de aprendizado de máquina podem descobrir como realizar tarefas importantes pela generalização a partir de exemplos passados, o que é muitas vezes mais viável e de baixo custo em comparação com a programação manual. Além disso, segundo Faceli et al. (2011, p. 3), a expansão da área de Aprendizado de Máquina se intensifica cada vez mais devido a fatores como desenvolvimento de algoritmos mas eficazes e elevada capacidade dos recursos computacionais disponíveis.

Uma variedade de métodos computacionais derivados do AM, têm sido utilizados para a identificação de gene. O estado da arte é caracterizado principalmente pelos Modelos Ocultos de Markov, que são tradicionalmente usados neste contexto, mas atualmente é perceptível a crescente utilização de outros métodos como *Support Vector Machine* (SVM), Redes Neurais, Análise discriminante, dentre outros.

Uma vertente do AM é o aprendizado *ensemble*, a qual consite na elaboração de comitês (*ensemble*) de aprendizado. Esta abordagem é baseada na elaboração de sistemas de classificação constituídos pela combinação de modelos gerados a partir de métodos de AM, como por exemplo o *K-Nearest Neighbor* (KNN). O principal objetivo desta metodologia é tentar melhorar o desempenho de um classificador individual através da indução e combinação de vários classificadores (GALAR et al., 2012). Desta forma, dependendo do problema em questão, técnicas do aprendizado *ensemble* podem agregar ganhos ao processo de predição.

O método *ensemble Random Forest* (RF) tem sido utilizado em problemas da Bioinformática porém, apesar da sua robustez, ele ainda não foi explorado e aplicado no problema de predição de genes em dados metagenômicos. Trabalhos como o de seleção de genes e classificação de dados de microarrays (DÍAZ-URIARTE; ANDRES, 2006), interação entre genes como destacado por (YANG et al., 2010), análise de Pathways (PANG et al., 2006) e predição de microRNA (MENDOZA et al., 2013), demonstram a aplicabilidade do RF, bem como o seu bom desempenho. Strobl et al. (2008) afirma que o RF tem mostrado alta precisão em problemas de predição e são aplicáveis mesmo em problemas com conjunto de dados de alta dimensão e com variáveis altamente correlacionadas, uma situação que muitas vezes ocorre na bioinformática.

1.2 JUSTIFICATIVA

O DNA (*deoxyribonucleic acid*) metagenômico é complexo, pois se trata de conjunto de genomas de vários organismos diferentes, fazendo com que o seu processo de análise seja desafiador (HANDELSMAN et al., 2007). Estas questões refletem em diversas análises, sendo uma delas a predição de genes que, por sua vez, constitui um dos passos fundamentais dos projetos metagenômicos.

A identificação de genes pode ser feita através da busca por homologia, porém, segundo Hoff et al. (2008), a desvantagem desta abordagem é a impossibilidade de encontrar novos genes e, desta forma, em projetos metagenômicos que visam descobrir novos genes e proteínas esta estratégia se torna inadequada. Desta foma, a predição *ab initio* se torna interessante pois, possibilita a identificação de genes ainda não anotados devido não haver a necessidade de comparação entre sequências. Um dos objetivos dos estudos metagenômicos é a identificação das funções potenciais das proteínas presentes nos metagenomas. Sendo assim, uma predição de genes precisa é a base para a descoberta de novos genes bem como a sua anotação funcional correta (HOFF, 2009a).

As dificuldades impostas pela natureza dos dados metagenômicos afetam diretamente o processo de predição que, por sua vez, influencia na anotação das sequências e análises posteriores. Dentre os impedimentos, pode-se destacar: sequências fragmentadas e (ALLALI; ROSE, 2013), fragmentos de tamanhos pequenos (HOFF et al., 2008) e construção de modelos aplicávies a uma grande diversidade de organismos (LIU et al., 2013).

Yok e Rosen (2011) realizaram um estudo em torno da combinação de três ferramentas de predição de genes a fim de obeter melhorias na identificação de genes em dados metagenômicos. Para tal, foram comparadas as predições das ferramentas com as predições obtidas através das combinação entre as mesmas. Os autores afirmam que um consenso (voto da maioria) entre as ferramentas MetaGeneMark (ZHU et al., 2010), Orphelia (HOFF, 2009b) e Metagene Annotator (NOGUCHI et al., 2008) obtiveram melhoria na predição em *reads* de 100, 200, 300 e 400 pares de base (pb), enquanto a interseção das predições da GeneMark e Ophelia foi a melhor estratégia de combinação para *reads* de 500, 600 e 700 pb. Este é um tipo de combinação que pode ser abordado a fim de obter ganhos na qualidade da predição. Ainda assim, combinar diferentes ferramentas/programas é um processo custoso e não trivial, necessitando da configuração e execução cada ferramenta.

Alguns métodos *ensemble*, são particularmente úteis para problemas que lidam com conjuntos de dados de alta dimensão, pois o aumento da precisão da classificação pode ser alcançado pela geração de múltiplos modelos com diferentes subconjuntos de variáveis (YANG et al., 2010). Espera-se, então, com o desenvolvimento de uma ferramenta de predição de genes baseada em *ensemble learning* amenizar os impactos do desafios oriundos dos contexto metagenômico, resultando em um refinamento ainda maior na identificação e anotação de genes.

Com base neste contexto, o presente trabalho tem o objetivo de propor um *pipeline* para identificação de genes, do tipo *ab initio*, em dados metagenômicos fazendo uso do *ensemble Random Forest*.

1.3 OBJETIVOS

Este trabalho tem como objetivo geral a elaboração de um *pipeline* para a predição de genes a partir de dados gerados em experimentos metagenômicos. Para constituir as bases da solução, uma predição de genes do tipo *ab initio* é adotada através de uma abordagem baseada em aprendizado de máquina supervisionado.

Para atender ao objetivo geral, os seguintes objetivos específicos são contemplados:

- Realizar um levantamento bibliográfico relacionado ao tema abordado;
- Determinar a composição do conjunto de dados de treino e de teste;
- Realizar uma análise sobre as variáveis utilizadas na predição de genes e definir o conjunto de variáveis utilizado;

- Analisar as propriedades do *ensemble learning*, que controlam a sua performance, no contexto dos dados selecionados;
- Desenvolver o *pipeline* para a predição de genes;
- Executar o *pipeline* utilizando os dados de teste;
- Realizar uma avaliação de performance do *pipeline* proposto, além de comparar com outras ferramentas de predição de genes de caráter metagenômico.

1.4 METODOLOGIA

Quantos aos procedimentos a pesquisa é considerada bibliográfica. Pois, a fim de fornecer embasamento para a pesquisa como um todo, foi realizada uma pesquisa bibliográfica sobre o contexto da Mategenômica e sobre questões mais específicas sobre a etapa de predição de genes; e sobre *ensemble learning*, mais especificamente o método *Random Forest*, bem como outros pontos relacionados ao Aprendizado de Máquina.

Ainda quanto aos procedimentos a pesquisa pode ser caracterizada como experimental, haja vista que, durante a elaboração do *pipeline* foram executados testes experimentais em relação a parametrização do modelo, as variáveis escolhidas e, por fim, a avaliação da predição realizada pelo modelo gerado.

Quanto aos objetivos a pesquisa é considerada descritiva e explicativa tendo em conta que, propõe um mecanismo cuja sua elaboração e os componentes técnicos envolvidos foram descritos e realiza a análise, e a interpretação, dos resultados dos testes aos quais o mecanismo proposto foi submetido e permitiram avaliar a adequação da solução proposta ao problema abordado.

1.5 CONTRIBUIÇÕES

O presente trabalho apresenta contribuições, as quais são descritas sucintamente a seguir. Duas delas foram publicadas ao longo da pesquisa realizada durante a elaboração desta dissertação e foraneceram embasamento para a mesma.

O artigo (GOÉS et al., 2014a), que foi aceito no *International Workshop on New Frontiers in Mining Complex Patterns/ECML-PKDD*, apresenta uma comparação empírica de desempenho entre quatro algoritmos de classificação do Aprendizado de Máquina aplicados no contexto da predição de genes em dados metagenômicos. Desta forma, apresenta-se um *benchmarking* diferente dos já então realizados no contexto de predição de genes, pois faz uma comparação a nível de métodos computacionais quando aplicados a um mesmo conjunto de dados e variáveis relacionados ao problema em questão ao invés de comparar diretamente as ferramentas. Desta forma, pôde-se analisar a capacidade de predição em nível de métodos computacionais.

O trabalho (GOÉS et al., 2014b), que foi publicado no *Brazilian Symposium on Bioinformatics*, emprega uma abordagem que envolve questões relacionadas ao *Random Forest* como a importância de cada variável utilizada para o modelo RF e como o modelo se comporta com variações no conjunto destas variáveis. Desta forma, contribuindo com uma análise de como um *ensemble learning*, bem como as variáveis selecionadas, se comportam no contexto da predição de genes.

A comparação realizada ao final deste trabalhado possibilitou alguns indicativos de melhorias que podem ser levadas em conta na predição de genes no contexto da Metagenômica. Foi verificado que ocorre a variação significativa entre os valores de sensibilidade, para fragmentos de mesmo tamanho, entre os organismos. Isto aponta que uma análise em torno desta questão pode ser feita, a fim de que as ferramentas consigam ter um uma precisão mais homogênea. Outra observação realizada está relacionda com a composição dos genomas adotados no teste e os resultados das ferramentas, o que indicou uma possível deficiência destas ferramentas em genomas com determinadas caracterísicas estruturais.

Este trabalho também contribuiu para a consolidação do projeto BIOFLOWS [475620/2012-7] que foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico.

1.6 ESTRUTURA DO TRABALHO

Este documento está dividido em mais cinco capítulos que fornecem o embasamento teórico sobre o trabalho, os detalhes sobre a implementação do *pipeline* desenvolvido, as análises dos resultados obtidos e as conclusões obtidas. Os capítulos são os seguintes:

- **Capítulo 2 – A Bioinformática:** Este capítulo apresenta a área da Bioinformática, bem como uma breve introdução sobre sua origem. Também é abordada uma das "ômicas" da Bioinformática, a Metagenômica, e dada ênfase em uma das etapas de um projeto Metagenômico, a predição de genes.

- **Capítulo 3 – O aprendizado de máquina:** Neste capítulo os conceitos de aprendizado de máquina com ênfase no aprendizado supervisionado são apresentados. Questões relacionadas ao aprendizado de máquina aplicado ao contexto da predição de genes serão discutidas juntamente com os trabalhos relacionados nesta área.
- **Capítulo 4 – MetaFindRF - Um pipeline para predição de genes em dados metagenômicos:** Este capítulo apresenta o *pipeline* desenvolvido. Detalhes de sua elaboração e implementação são descritos nesse capítulo, assim como considerações sobre a elaboração do conjunto de dados, as variáveis utilizadas e o método de aprendizado selecionado.
- **Capítulo 5 – Resultados experimentais:** Apresenta uma análise sobre um conjunto de dados de treste experimental e e uma comparação dos resultados do *pipeline* com outras quatro ferramentas.
- **Capítulo 6 – Considerações Finais:** No último capítulo são apresentadas as considerações finais sobre o trabalho desenvolvido, limitações, assim como contribuições e sugestões para futuros trabalhos.

2 A BIOINFORMÁTICA

A Computação e a Biologia Molecular possuem muitos aspectos de caráter comum. De acordo com Setubal (2003), a informação genética também é armazenada de forma digital, sendo escrita em um alfabeto que, em vez de ser binário, é quaternário, composto pelas letras A, T, C, G; e a forma dos genes operarem também é, até certo ponto, digital, já que podem ser "ligados"ou "desligados". Desta forma, torna-se possível realizar o mapeamento da Biologia Molecular para o contexto computacional.

A Bioinformática teve o seu início em um momento de grande geração de dados em pesquisas biológicas, mais precisamente, de caráter genômico - onde técnicas de sequenciamento de DNA ganharam espaço - pois, de acordo com Polanski e Kimmel (2007), a sua origem estava associada a manutenção de bases de dados com conteúdos biológicos e clínicos. Prosdocimi et al. (2002), destaca que nos anos 90 – com o surgimento dos sequenciadores de DNA automáticos – houve a necessidade de se obter recursos computacionais cada vez mais eficientes devido ao crescimento intenso na quantidade de sequências a serem armazenadas. Nesta mesma década, um dos impulsionadores do fortalecimento da necessidade de recursos computacionais foi o Projeto do Genoma Humano que mobilizou pesquisadores de várias partes do mundo em busca do sequenciamento do genoma humano, e trouxe consigo diversos desafios, pois, de acordo com Fuchs (2002), sem poderosos sistemas computacionais, algoritmos sofisticados e capacidade de gestão avançada de dados esse esforço teria sido condenado ao fracasso. Desta forma, como afirmam Polanski e Kimmel (2007), a consagração da Bioinformática emergiu com a conquista do sequenciamento do genoma humano e, subsequentemente, dos genomas de outros organismos.

A maior capacidade de processamento, armazenamento, e o gerenciamento dos diversos tipos de dados biológicos foram avanços que surgiram junto com o progresso da Bioinformática. No entanto, com o fortalecimento de uma nova etapa onde a utilização dela passa a apoiar diretamente o objetivo de expandir a busca de soluções de diversos problemas de processamento e armazenamento de dados da área biológica, o desenvolvimento de ferramentas e métodos agregou também a essa disciplina o poder de fornecer melhor análise e interpretação dos dados armazenados. Desta forma, a Bioinformática se concretizou como uma disciplina, cujo desenvolvimento abrange ciências quantitativas e biológicas. Dentre as ciências geradas da matemática aplicada, algumas são de especial importância para a Bioinformática, a teoria da probabilidade e estatística e algoritmos em ciência da computação.

Com a consagração da Bioinformática como um campo de pesquisa, novos desafios foram traçados, envolvendo problemas da área biológica. A partir dos dados gerados pelas tecnologias de sequenciamento, a necessidade de técnicas e ferramentas para montagem e anotação de sequências de DNA se intensificou (POLANSKI; KIMMEL, 2007). Desta forma, partindo do sequenciamento dos genomas, a aplicação da Bioinformática se aprofunda nas diversas etapas de análise dos mesmos; detectar padrões e estruturas como regiões codificadoras, denominadas também como genes, que por sua vez permitem mapear possíveis proteínas, que agregam função a essas determinadas regiões e compõem um fluxo básico da aplicação da Bioinformática no estudo dos genomas.

Em meio a todo o processo de busca da evolução e do aprimoramento de recursos que possibilitassem uma nova visão do estudo molecular e genético dos organismos, algumas vertentes básicas de análise de genoma surgiram. Além da tradicional Genômica, outras ciências "omicas" como a Metagenômica, Transcriptomica, Proteômica e Metabolômica se estabilizaram enquanto campos de estudo que englobam diversos aspectos biológicos derivados do estudo dos genomas; elas têm revolucionado a percepção do que os genes fazem e de como eles trabalham em conjunto (HANDELSMAN et al., 2007, p. 47).

As "omicas" compartilham com a biologia o desafio de lidar, de forma flexível e eficaz, com quantidades de informações complexas cada vez maiores (SCHNEIDER; ORCHARD, 2011, p. 8). O apoio de algoritmos, métodos e ferramentas – que apoiem a investigação aprofundada dos genomas e que tratem restrições e inconsistências geradas pela grande quantidade de dados advindos dos novos estudos nesta área – também se tornou uma contínua necessidade neste contexto. Segundo Mühlberger et al. (2011, p. 380), etapas de análises incluem o armazenamento de dados, pré-processamento de dados e normalização, a anotação de dados, seguido por análises exploratórias e estatísticas, interpretação funcionais e geração de hipóteses.

2.1 A GENÔMICA

Segundo Kalisky et al. (2011), a era genômica proveu uma ampla gama de tecnologias para a manipulação da grande quantidade de dados biológicos gerados em pesquisas de caráter genético. Com isso, a Genômica tornou-se uma ciência que possibilitou um melhor entendimento dos organismos, em nível de material genético. Junto com a advento da Genômica, a cultura pura tornou-se um parâmetro para estudos dos organismos. Esta técnica caracteriza-se pelo cultivo em laboratório de um organismo em um meio que simule determinadas condições necessárias. De acordo com Teeling e Glöckner (2012), culturas puras permitem o estudo do metabolismo de um organismo isolado e de seu repertório de genes através do sequenciamento do genoma.

Com o sequenciamento do genoma humano, a Genômica pôde expandir um novo rumo em busca do desvendamento do mesmo e, de acordo com Green et al. (2011), contribuir para uma melhor compreensão e melhoria da biologia e da saúde humana, já que a base provida pelo sequenciamento de genomas possibilitou gradualmente a construção de um conhecimento cada vez maior sobre sistemas cada vez mais complexos. É importante destacar também que os esforços destinados aos avanços nos estudos dos genomas – principalmente pela busca do sequenciamento completo do genoma humano – além de prover o aprimoramento de métodos de análise que fornecesse subsídio para tal, possibilitaram a diminuição no preço e o aumento das tecnologias utilizadas para este fim, tendo como consequência a intensificação da Genômica e do sequenciamento das mais variedades de genomas.

De acordo com Handelsman et al. (2007, p. 23), a Genômica também contribuiu para o avanço da microbiologia, pois as técnicas genômicas também forneceram subsídios para uma melhor compreensão da composição genética dos microrganismos e o relacionamento entre os seus habitats. Porém, durante algum tempo, microbiologistas não investigaram diretamente o impacto da aplicação de cultura em laboratório para o estudo dos microrganismos. Segundo Handelsman et al. (2007, p. 25), em 1985, Staley e Konopka revisaram sobre a capacidade de trazer micróbios do ambiente natural para o cultivo em laboratório. Partindo disto, foram gerados avaliações e questionamentos sobre quais impactos existem a partir da principal técnica de análise da Genômica, a cultura pura, na análise dos microrganismos. De acordo com Handelsman (2004), a constatação de que a maioria dos microrganismos não pode ser cultivada facilmente microbiologistas questionaram a sua convicção de que o conhecimento sobre mundo microbiano havia sido dominado. Devido a isto, uma mudança de paradigma surge no estudo de genomas, que será apresentada posteriormente.

2.2 A METAGENÔMICA

A busca pelo desconhecido em termos de microrganismos relacionados à comunidades caracteriza uma das maiores razões para o destaque que a Metagenômica recebeu nos últimos anos; saber onde – em um determinado ambiente –, quem – quais organismos estão presentes vivendo em comunidade – e o quê – que função chave cada organismo desempenha – agrega um importante valor para vários ramos da ciência. Handelsman et al. (2007) afirma que desvendando os segredos das comunidades microbianas, haverá inúmeras maneiras de conhecer e solucionar grandes de desafios na Biomedicina, Agricultura e Gestão Ambiental.

Segundo Wooley et al. (2010), Micróbios não são apenas onipresentes; eles são essenciais para toda a vida, pois são a maior fonte para nutrientes existentes, e os transformadores primários da matéria morta em uma nova forma orgânica disponível para o ambiente. Além disso, eles influenciam no desenvolvimento e saúde de seres vivos e nas transformações químicas para a geração de componentes necessários para a vida no planeta são exemplos de outros papéis básicos que os microrganismos desempenham. Funções como estas são fundamentais para a estabilidade do planeta e dos seres que nele habitam, assim, o estudo dos microrganismos, principalmente em nível de comunidade, é de extrema importância. Funções, como as citadas anteriormente, não são triviais e, de acordo com (HANDELSMAN et al., 2007, p. 1), são realizadas por comunidades complexas, equilibradas e integradas que se adaptam de forma rápida e flexível às mudanças ambientais. Portanto, a conclusão de uma função é o resultado de uma série de etapas que são desempenhadas não somente por microrganismos da mesma espécie, mas por um conjunto de microrganismos de espécies diferentes, que interagem e se comunicam de maneiras complexas para que o objetivo final seja alcançado. Além de contribuir para o desempenho de alguma atividade os microrganismos geralmente necessitam da interação de uns com os outros, e do meio em que vivem, a fim de extrair nutrientes e outros benefícios que permitam a sua sobrevivência.

De acordo com Wooley et al. (2010), embora o paradigma de cultura pura convencional continue a ser importante para a caracterização completa de uma espécie, a sua tradicional análise de caráter individual limita a exploração do mundo microbiano. Neste contexto, houve o surgimento da necessidade da aplicação de um processo que fosse além do caráter individual das pesquisas genômicas, pois, segundo (HANDELSMAN et al., 2007, p. 23), compreender comunidades microbianas exige que as técnicas tradicionais de cultura pura sejam complementadas com novas abordagens. Assim, ocorreu o surgimento de uma vertente de análise dos genomas, sendo voltada para o âmbito das particularidades dos microrganismos em nível de comunidade, portanto o metagenoma.

Com o reconhecimento da necessidade do desenvolvimento de técnicas computacionais que proporcionassem uma melhor análise e compreensão de comunidades microbianas surgiu a Metagenômica. Handelsman et al. (2007, p. 13), a caracteriza como um campo de pesquisa, compreendendo muitas abordagens e métodos relacionados, aplicado ao estudo de comunidades microbianas. Ainda, Thomas et al. (2012) define como a aplicação de um conjunto de tecnologias genômicas e ferramentas de Bioinformática para acessar diretamente o conteúdo genético de comunidades de microrganismos.

Em suas abordagens e métodos a Metagenômica tenta contornar a baixa capacidade de cultivo e diversidade genômica da maioria dos micróbios, o que constituí-se como os maiores obstáculos para avanços em microbiologia clínica e ambiental (HANDELSMAN et al., 2007, p. 3), além de estabelecer hipóteses sobre as interações entre os membros das comunidades microbianas (KUNIN et al., 2008). Desta forma, a Metegenômica possibilita que uma variedade de novos estudos, em diversos tipos de ambientes, seja realizada a fim de tornar possível o uso dos benefícios que o conhecimento de comunidades microbianas pode proporcionar, pois, aplicando técnicas da Bioinformática modeladas para lidar com as condições impostas pelas características das comunidades microbianas, um projeto metagenômico pode, de uma forma mais adequada, realizar estudos sobre metagenomas que gerem resultados mais precisos e condizentes com este contexto.

Além dos desafios computacionais impostos pelos problemas da área biológica como um todo, a Metagenômica traz consigo peculiaridades e dificuldades devido à natureza dos seus dados. O DNA metagenômico é complexo, pois se trata de conjunto de genomas de vários organismos diferentes, fazendo com que o seu processo de análise seja desafiador (HANDELSMAN et al., 2007, p. 27). Além de contornar as limitações da cultura pura, um projeto metagenômico deve ser adaptado para lidar com desafios impostos pelas características dos dados, que geram uma complexidade computacional muito maior do que em comparação com um projeto genômico, pois, de acordo com Wooley et al. (2010):

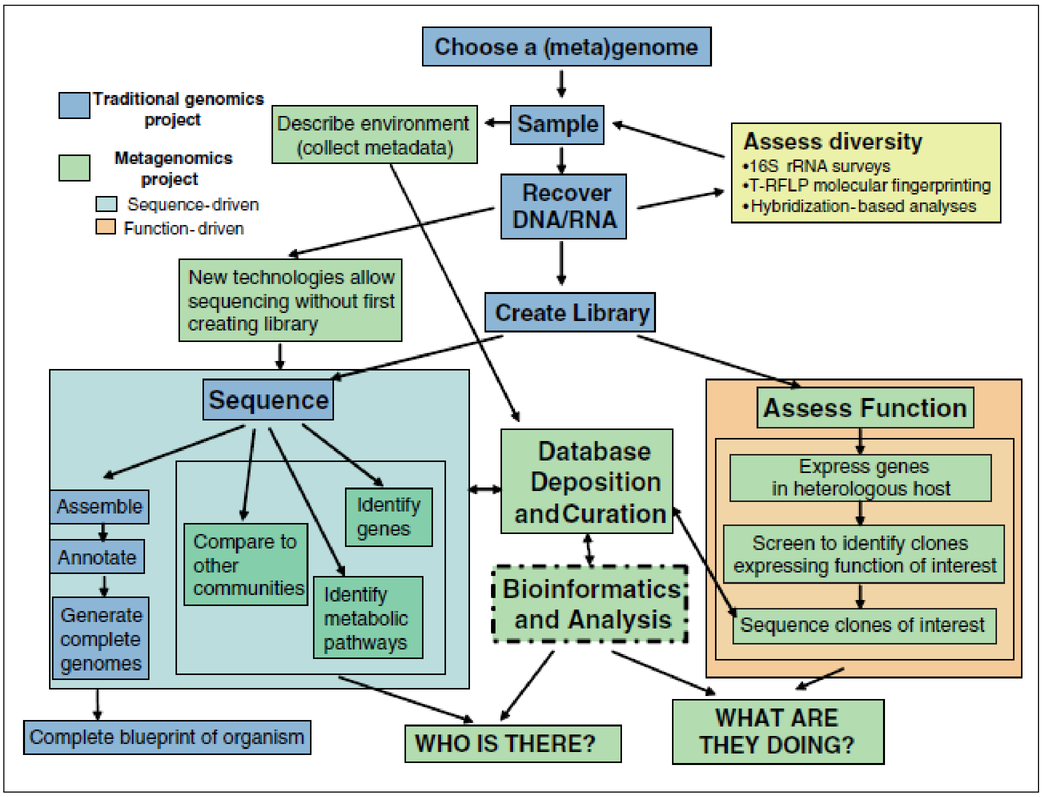
> Análise computacional demonstrou ter um impacto ainda maior em estudos metagenômicos em comparação com projetos genômicos tradicionais, devido não só à grande quantidade de dados metagenômicos, mas também à complexidade introduzida por projetos metagenômicos (por exemplo, montagem de vários genomas simultaneamente é mais desafiador do que a montagem de genomas individuais), e as novas questões que estão sendo levantadas (por exemplo, a interação entre hospedeiro e microrganismo).

2.2.1 O *Workflow* Metagenômico

Os experimentos nas diversas áreas da bioinformática são na maioria dos casos complexos, onde um conjunto de diversas tarefas básicas é desempenhado para solucionar um determinado problema. A fim de se obter um melhor gerenciamento de um projeto científico que possui um fluxo de diversas atividades, ferramentas e métodos, pode ser utilizado um *Workflow* Científico. Zhao e Paschke (2012) afirmam que os Workflows Científicos estão ganhando cada vez mais atenção para dar suporte a experimentos científicos, facilitando os cientistas a realizar o gerenciamento, análise e simulação de dados.

Com a sua concretização como um novo paradigma de estudo dos genomas, a Metagenômica incorpora um *workflow*, que possui claramente como base o *workflow* Genômico, mas que, da mesma forma, possui diferenças muito bem especificadas; diferenças essas que dão suporte para uma melhor compreensão das especificidades apresentadas pelo estudo dos metagenomas. A Figura 1 mostra claramente isto.

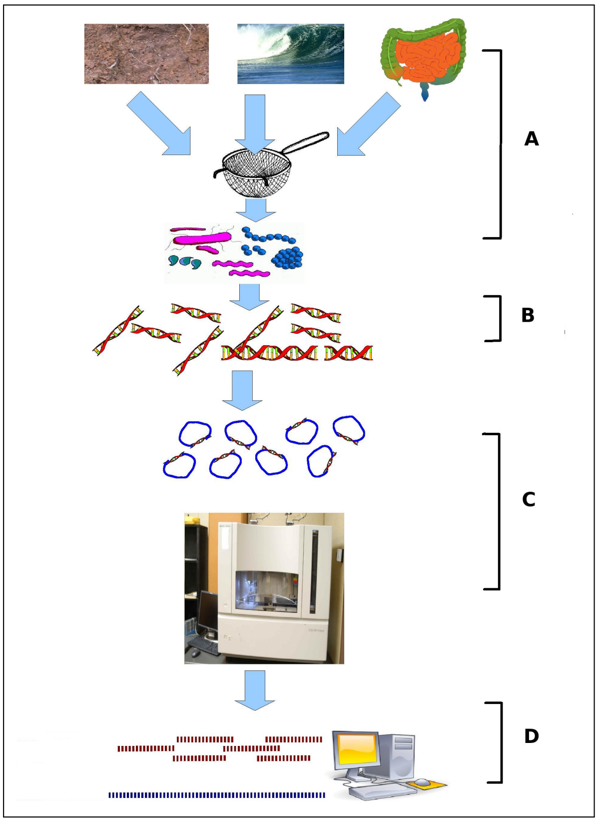
Figura 1 – Etapas de um workflow Genômico em azul e Metaganômico em verde.



Fonte: Adaptado de Handelsman et al. (2007)

Em seu trabalho, Thomas et al. (2012) considera e descreve oito etapas básicas que um típico projeto metagenômico possui. São elas: processamento da amostra, sequenciamento, montagem, binning, anotação, design experimental, análise estatística e armazenamento/compartilhamento de dados. A seguir, será feito um breve detalhamento sobre as cinco primeiras etapas citadas anteriormente. A Figura 2 ilustra com detalhe os processos que serão descritos a seguir. Pode-se perceber que a amostragem da comunidade microbiana e a extração do DNA, diretamente da amostra ambiental, caracteriza-se com a parte inicial de um projeto metagenômico.

Figura 2 – Fluxo de obtenção do DNA metagenômico contendo as etapas iniciais de (A) amostragem e extração (B), e as posteriores de sequenciamento (C) e montagem (D) trabalham com o DNA em formato digitalizado.



Fonte: Adaptado de Wooley et al. (2010)

2.2.1.1 Amostragem

Um *workflow* metagenômico tem como primeira etapa a obtenção da amostragem ambiental. Segundo Wooley et al. (2010), as amostras devem representar a população de onde são tomadas; o problema na ecologia microbiana é que somos incapazes de ver os organismos que estamos tentando capturar na etapa de amostragem. Ou seja, neste processo é fundamental haver cuidado na escolha da técnica para a realização da amostragem, a fim de garantir um tamanho e número de amostras consistentes.

2.2.1.2 Sequenciamento

A etapa de sequenciamento é o primeiro passo para a geração do DNA de forma digitalizada. Segundo Wooley et al. (2010), o sequenciamento de caráter metagenômico é feito de maneira bem parecida com o genômico, porém, o material genômico não vem de um único organismo; mas sim de uma comunidade de micróbios, sendo assim, mais complexo. As tecnologias de sequenciamento tem como base o *Whole genome shotgun sequencing*, na qual o DNA é dividido aleatoriamente em vários pequenos segmentos, que são sequenciados gerando *reads*.

De acordo com Wooley et al. (2010), tecnologias de sequenciamento de segunda geração foram rapidamente ganhando terreno e estão substituindo o sequenciamento Sanger em genomas de tamanhos pequenos e na genômica ambiental. Mas, Thomas et al. (2012) afirma, que o sequenciamento Sanger, ainda é considerado o padrão para o sequenciamento, devido à sua baixa taxa de erro e *reads* de tamanho maior (> 700 pares de base). As novas tecnologias de sequenciamento têm poder de processar um volume de dados maior e de forma mais rápida, contudo, os *reads* gerados são bem menores do que os gerados pelo método Sanger. A etapa do processo de montagem terá uma complexidade computacional maior ao manipular esses *reads*, devido ao seu pequeno tamanho.

2.2.1.3 Montagem

Partindo dos *reads* gerados na etapa de sequenciamento, pode-se então realizar a montagem de sequências mais contíguas, chamadas de *contigs*. Este processo é feito através da ligação de um grande número de *reads* de modo que elas possam reproduzir a ordenação encontrada no DNA do organismo. *Contigs* também podem ser ligados formando sequências maiores e, portanto, ainda mais contínuas, chamadas de scaffolds. Através da montagem dos *contigs* estruturas dos genomas, que influenciam diversos outros tipos de análises, podem ser encontradas de uma melhor forma.

De acordo com Thomas et al. (2012), O desenvolvimento de ferramentas de montagem de *reads* metagenomicos ainda está em uma fase inicial, e a avaliação da precisão deste tipo de procedimento é difícil de ser realizada, pois normalmente não existem referências para que os resultados sejam comparados.

2.2.1.4 *Binning*

Binning o processo no qual sequências de DNA (por exemplo, *contigs*) são agrupadas e classificadas em grupos de genomas de acordo com as suas proximidades, que são determinadas por características das próprias sequências. O fato de que os genomas têm composição conservada de nucleotídeos como determinado conteúdo GC, assim como *codon usage*, é uma forma de classificação utilizada no *Binning*.

2.2.1.5 Anotação

Handelsman et al. (2007, p. 49), caracteriza a anotação como um processo de classificação de genes preditos em famílias de genes conhecidas e bem caracterizadas. Desta forma, a predição de genes constitui a etapa inicial do processo de anotação. Alguns autores descrevem a etapa de predição separadamente da anotação, mas nesta Seção ambas serão englobadas e na Seção 2.2.2 a predição de genes será aprofundada.

Na maioria dos casos, o processo de anotação se inicia após a montagem, fazendo uso dos *contigs*. Como dito na Seção 2.2.1.3, os *contigs*, são sequências mais longas montadas a partir dos *reads*, e devido a isso tendem a carregar as informações estruturais dos genomas de forma mais contínua. Desta forma, Thomas et al. (2012) afirma que em *pipelines* existentes para anotação de genoma o uso deles é preferível. Porém, em alguns projetos metagenômicos é impossível realizar a montagem de todos dos *reads* em *contigs* de forma confiável devido a diversidade presente nas amostras ser muito grande para possibilitar uma alta cobertura no sequenciamento das espécies. Desta forma, a anotação também pode ser feita a partir dos *reads*.

Mas é importante destacar também que algumas etapas de anotação podem ser feitas a partir dos *reads*, fornecidos pelo sequenciamento do DNA.

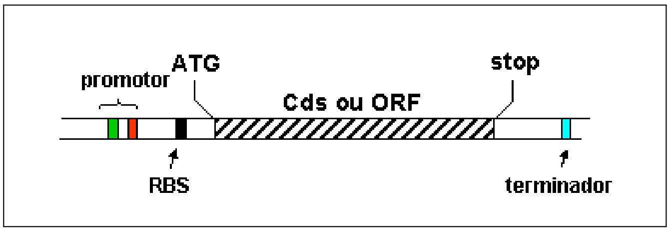
Larrañaga et al. (2006) afirma que se os genes contêm as informações, as proteínas são as responsáveis pelas transfromações destas informações em vida. Sendo assim, a etapa de identificação dos genes – que possuem as informações das proteínas que serão codificadas a partir deles – é a base para os mais diversos tipos de análises em relação ao potencial da comunidade microbiana que está sendo estudada. Com os genes já identificados, pode-se realizar o que, em seu trabalho, Wooley et al. (2010) caracteriza como anotação funcional. Membros de uma comunidade microbiana de determinado ambiente desempenham funções diferentes. Sendo assim, em uma análise metagenômica um dos principais objetivos é conhecer o que os micoorganismos presentes na amostra são capazes de fazer em comunidade. Para isto, em um projeto metagenômico, é realizada a anotação funcional dos genes que foram previamente identificados. Para gerar um estudo funcional da comunidade microbiana no processo de anotação funcional, segundo Wooley et al. (2010), primeiramente é realizada a atribuição biológica e funcional dos genes, e posteriormente a descoberta de genes que constituem redes biológicas, tais como as vias metabólicas. Desta forma, segundo Prakash e Taylor (2012), a análise funcional dos dados metagenômicos desempenha um papel central em tais estudos, fornecendo pistas importantes sobre a diversidade funcional e metabólica das comunidades microbianas.

2.2.2 Predicão de genes

Muitos métodos utilizados na Bioinformática giram em torno da estrutura linear da informação genômica (POLANSKI; KIMMEL, 2007). A identificação de genes se enquadra neste tipo de análise. Segundo Handelsman et al. (2007, p. 49), a predição de genes é importante, pois fornece a base para a determinação do repertório funcional de uma comunidade microbiana e permite a comparação entre diferentes comunidades.

Um gene caracteriza-se como uma região codificadora de proteína ou de RNA. A estrutura de um gene procarioto difere do eucarioto desta forma, fazendo com que o processo de identificação tenha que ser elaborado para um determinado tipo de domínio (Eucarioto ou Procarito), a fim de ser adequado para lidar com as características da estrutura dos organismos deste domínio. A Figura 3 ilustra a estrutura de um gene procarioto, que constitui uma sequência contínua sem a presença de *introns*.

Figura 3 – Estrutura típica de um gene procarioto. Estão destacadas as estruturas: RBS (sítio ligador de ribossomo), ATG (codon de iniciação da síntese protéica), Cds ou ORF, stop (um dos três codons que sinalizam a terminação da tradução), terminador (região terminadora da transcrição).



Fonte: Adaptado de Andrade et al. (2014)

Hoff et al. (2008) afirmam que a maioria dos genes procariotos que codificam proteínas constituem um codon de início (ATG), seguido por um número variável de codons consecutivos que são terminados por um codon de parada (TAG, TAA, TGA); e este arranjo de codons é comumente chamado de *open read frame* (ORF). Um codon constitui-se como uma série de três nucleotídeos. De acordo com (MOUNT; MOUNT, 2001), o método mais simples para encontrar as sequências de DNA que codificam proteínas é a procura pelos *open read frames*.

A busca pelos ORFs codificadores de proteínas pode ser aplicada aos *reads*, gerados a partir do sequenciamento de DNA da amostra metagenômica, ou aos *contigs*, gerados pela etapa de montagem. Reads, independente da tecnologia utilizada, possuem um tamanho menor que os *contigs*. De acordo com Hoff Hoff (2009b, p. 7), em reads metagenomicos, os ORFs frequentemente excedem as extremidades destes fragmentos, mas para estes casos ORFs incompletos podem ser considerados. A chance desta circunstância ocorrer com a utilização dos *contigs* é menor, devido ao fato deste ser uma sequência maior, por outro lado, a montagem dos *contigs* metagenômicos é muito suscetível a sofrer erros, como destacado na Seção 2.2.1.3.

Portanto, a escolha em que tipo de sequência será aplicada a identificação de genes dependerá da característica da amostra metagenômica e da situação dos dados gerados pelas etapas iniciais, como o sequenciamento e a montagem. Portanto, de acordo com (KUNIN et al., 2008), a qualidade da predição de genes depende da qualidade de pré-processamento dos dados. Este pode ser baseado em *standards* ou consórcios, como o *Brazilian Microbiome Project* (BMP), que visam integrar projetos e tecnologias para geração e análise de dados deste âmbito. Porém, Pylro et al. (2014) destaca que um dos desafios é o desenvolvimento/divulgação de normas uniformes para a análise de dados que podessam ser integrados com os consórcios e as normas existentes.

2.2.2.1 Predição de Genes por Homologia

A busca por similaridade ocorre através da comparação de sequências que não possuem os seus genes anotados com sequências disponíveis, juntamente com a sua anotação, em banco de dados. Segundo (WANG et al., 2004), alinhamento local e alinhamento global são dois métodos nos quais as pesquisas por similaridade são baseadas. Apesar da predição de genes por comparação ser categórica nesta área, ela deixa a desejar na predição de genes ainda não anotados. Segundo (HOFF et al., 2008), a pesquisa de homologia se torna uma abordagem inadequada para a predição de gene principalmente nos casos em que estudos metagenômicos visam descobrir novos genes e proteínas. Pois, através da homologia, é possível encontrar apenas genes que já foram anotados, impossibilitando assim, a descoberta de novos genes.

2.2.2.2 Predição de genes *ab initio*

A predição do tipo *ab initio* é classificada por Wang (2004) como uma classe de métodos computacionais para a identificação de genes que consiste na utilização de padrões estruturais de genes como um modelo para detectar genes. Segundo Hoff Hoff (2009b, p. 9), a predição de genes baseada em modelos estatísticos tem a vantagem de proporcionar a descoberta de novos genes com menor custo computacional e sem o pré-requisito de uma elevada conservação destes genes no interior da amostra.

A maioria dos processos de produção de genes utiliza como base a identificação dos ORFs, pois toda região codificadora é um ORF. Porém, se faz necessária a utilização de determinados padrões da sequência para a confirmação que o ORF inicialmente localizado é um possível gene, pois de acordo com Angelova et al. (2010), nem todo ORF é gene.

Segundo (WANG et al., 2004), predições de genes *ab initio* podem utilizar dois tipos de informação: sensores de sinais e sensores de conteúdo. Sinais gênicos ou sensores de sinais baseiam-se em estruturas funcionais características dos genes. Os sensores de conteúdo são padrões característicos de sequências codificadoras de proteína. Na predição *ab initio* geralmente é utilizada mais de uma informação extraída da sequência, isto ocorre devido ao fato de que a identificação da presença de apenas uma informação pode não necessariamente garantir que determinada região seja codificadora de proteína.

A predição computacional de genes de caráter genômico, também possui muitas ferramentas que foram desenvolvidas para a identificação de genes em genomas de organismos procariotos. Porém, ferramentas deste tipo segundo Noguchi et al. (2008), apresentam limitações principalmente porque são projetadas para identificar genes em genomas completos com vários milhões de pares de bases. Porém, esta não é a realidade de dados gerados a partir do sequenciamento e montagem dos genomas de uma comunidade microbiana.

3 APRENDIZADO DE MÁQUINA

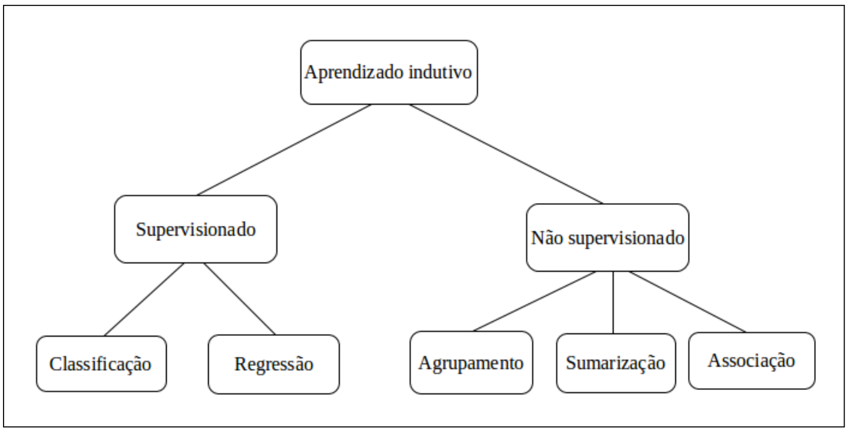
O avanço do desenvolvimento de softwares cada vez mais sofisticados é impulsionado pelo intenso crescimento de volume de dados gerados por diversos domínios — como, por exemplo, da saúde, ambiental e comercial — e pelo surgimento dos mais diversos tipos de problemas a serem solucionados. De acordo com Witten e Frank (2011, p. 25), o Aprendizado de Máquina (AM) é uma tecnologia promissora para a mineração de conhecimento a partir de dados, a qual está tendo um grande reconhecimento da sua importância por parte de muitas pessoas.

Domingos (2012) afirma que algoritmos de aprendizado de máquina podem descobrir como realizar tarefas importantes generalizando a partir de exemplos. Para tal, esses utilizam um viés indutivo, que possibilite a indução de uma hipótese capaz de resolver o problema a partir de um conjunto de dados de uma determinada aplicação. Hand et al. (2001, p. 4) caracteriza um conjunto de dados como, um conjunto de medições obtidas de algum ambiente ou processo; no caso mais simples é constituído por um conjunto de objetos, em que cada um possui um conjunto de valores correspondentes às medições.

Segundo Faceli et al. (2011, p .6), o AM pode ser aplicado a diversas tarefas que podem ser classificadas de acordo com o paradigma de aprendizado a ser adotado, sendo que este pode ser do tipo preditivo ou descritivo. A tarefas de predição, utilizam o conjunto de dados para formular uma hipótese, que possa inferir a classe de um novo objeto em relação aos seus atributos. As tarefas descritivas, caracterizam objetos a partir das propriedades do próprio conjunto de dados. Desta forma, o AM pode ser divido em Aprendizado supervisionado e Aprendizado não supervisionado, sendo estruturado como uma hierarquia, ilustrada na Figura 4, determinada de acordo com os tipos de tarefa do aprendizado.

O aprendizado de máquina não supervisionado se caracteriza por ser de caráter descritivo e pela utilização de técnicas que realizam agrupamento e associações entre grupos. O aprendizado de máquina supervisionado se caracteriza por ser de caráter preditivo, ou seja, pela indução de modelos preditivos. Este tipo de aprendizado pode ser categorizado em classificação e regressão, de acordo com o tipo do valor da classe. De acordo com Han e Kamber (2012, p. 19), enquanto que a classificação prediz classes com valores categóricos (discretos), a regressão modela valores contínuos. Portanto, uma tarefa de predição pode ser aplicada a classes de valores numéricos ou discretos.

Figura 4 – Hierarquia do Aprendizado indutivo.



Fonte: Adaptado de Faceli et al. (2011)

3.1 APRENDIZADO SUPERVISIONADO

A classificação tem como objetivo, em um determinado problema, dividir os objetos em classes de acordo com os padrões extraídos de um modelo de predição. O modelo de predição é gerado a partir de um conjunto de dados de treino que é processado por um método de classificação supervisionado que, segundo Larrañaga et al. (2005), são algoritmos que induzem as regras de classificação dos dados. Após a obtenção de um modelo treinado, um elemento/objeto de um determinado conjunto de dados pode ser classificado de acordo com os valores de suas variáveis, a partir dos conjuntos de regras do modelo.

3.1.1 Conjunto de dados

Em qualquer aplicação direta de algoritmos de AM, há a necessidade de que os dados possuam uma boa qualidade e estejam adequados para o processo de aprendizado, pois dependendo dos problemas e das suas extensões, encontrados nos dados, eles podem prejudicar o processo indutivo. Desta forma, a etapa de pré-processamento dos dados torna-se crucial, pois contribui para o sucesso da aplicação de métodos do aprendizado de máquina. Pois, como Faceli et al. (2011, p. 4) destaca, o objetivo de um algoritmo de AM é aprender, a partir do conjunto de treinamento, um modelo que seja capaz de relacionar os valores dos atributos de entrada com os valores dos atributos de saída. Desta forma, é explícita a importância de um conjunto de dados consistente e balanceado para a geração de um modelo robusto.

Segundo Hand et al. (2001, p. 41), as colunas de uma matriz de dados são caracterizadas como variáveis, e o seu nome é a medição que é representada por cada coluna. As variáveis são os atributos de entrada, dos quais cada objeto terá uma instância. Segundo Domingos (2012), alguns projetos de aprendizagem de máquina são bem sucedidos e alguns falham, e claramente o fator mais importante, que influencia diretamente nessa diferença, é a escolha das variáveis que são utilizadas no processo de aprendizagem. Portanto, investigar o problema em questão a fim de localizar padrões dos quais possam ser extraídas variáveis que o representem bem é uma etapa crucial no processo de aprendizagem.

3.1.2 Avaliação do aprendizado

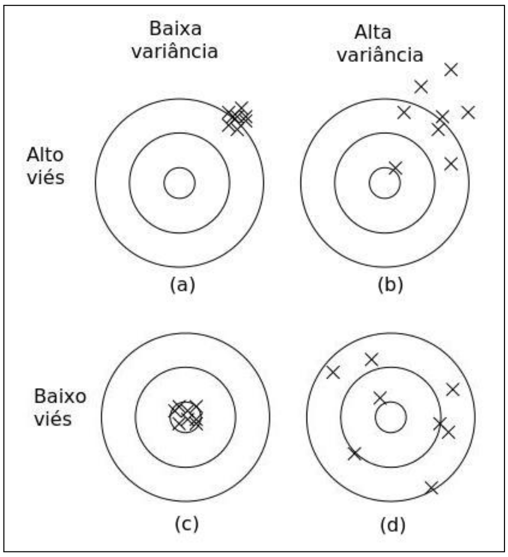
Uma medida básica que pode ser aplicada a fim de averiguar a performance dos classificadores é a taxa de erro. Witten e Frank (2011, p. 148) definem esta medida como a proporção de erros sobre o conjunto total de instancias, medindo assim a performance geral do classificador. Segundo Witten e Frank (2011, p. 152), para prever a performance de um classificador sobre novos dados é necessário acessar a sua taxa de erro sobre um conjunto de dados que não foi aplicado para a formação do classificador, denominado dados de teste. Desta forma, se faz necessária a aplicação da técnica conhecida como *holdout*, que consiste na separação dos dados em um subconjunto de treino - que será utilizado pelo algoritmo de aprendizado para a geração do classificador ou modelo - e um subconjunto de dados de teste.

Witten e Frank (2011, p. 153) afirmam que, cada classe presente no conjunto de dados completo deve ser representada em cerca da mesma proporção no conjunto de dados de treino e de teste. Para tal, se faz necessário que a amostragem para os dados de treino e de teste seja feita de forma aleatória. Neste caso, é possível que seja calculada a taxa de erro em cada iteração, desta forma a taxa de erro global será a média das taxas de erro calculadas.

De acordo com Domingos (2012), essa separação de dados de teste reduz a quantidade disponível para o treinamento. Isto pode ser mitigado fazendo uso do validação cruzada que, de acordo com Witten e Frank (2011, p. 153), constitui uma importante técnica estatística para estimar a performance de um modelo de AM. Na validação cruzada é realizada a divisão aleatória dos dados de treinamento para k subconjuntos, em que um é destinado para teste e o restante para o treinamento. O processo é repetido k vezes, testando o que o classificador aprendeu em cada k iteração. Ao final é realizado o cálculo da média dos resultados dos k testes. A técnica de validação cruzada pode evitar dois grandes problemas que podem ocorrer no Aprendizado de Máquina: *overfitting* e *underfitting*.

Segundo Mitchell et al. (1997) define-se que uma hipótese h sofre *overfitting* em relação aos dados de treinamento se existe outra hipótese h', tal que h erra menos que h' para os dados de treinamento, mas h' erra menos para os dados de teste. Ou seja, existe a possibilidade de um modelo que possui um desempenho muito satisfatório em um conjunto de dados de treinamento não ter precisão satisfatória quando aplicado a novas dados. Por outro lado, a hipótese induzida sofre *Underfitting* quando apresenta um desempenho ruim tanto no conjunto de treino como de teste. Domingos (2012) afirma que maneira de entender essas questões é pela análise do viés e da variância que, respectivamente, se constituem como a tendência de um modelo aprender consistentemente a mesma coisa errada e a tendência de aprender coisas de forma aleatória. A Figura 5 exemplifica os problemas de *overfitting* e *underfitting* em relação ao viés e a variância.

Figura 5 – Análise de viés e variância. Um modelo com *overfitting* (d) modelo com *underfitting* (a).



Fonte: Adaptado de Domingos (2012)

3.2 *ENSEMBLE LEARNING*

Ensemble learning constitui-se como uma vertente do Aprendizado de Máquina que trabalha com a combinação de modelos. Witten e Frank (2011, p. 351) destacam que por intermédio da utilização de *ensembles* é possível transformar um esquema de aprendizado relativamente fraco em um mais robusto. Segundo Han e Kamber (2012) um *ensemble* combina uma série de k modelos (ou classificadores base), M_1, M_2,..., M_k, com o objetivo de criar um modelo combinado mais robusto; e um conjunto de dados, D, é usado para criar k conjuntos de treino, D_1, D_2,..., D_k, onde D_i é usado para gerar o modelo M_i.

Ensembles podem ser gerados por diferentes estratégias. Domingos (2012) elenca as seguintes:

- *Bagging*: gera amostras aleatórias do conjunto de treinamento, aprende um classificador em cada amostra de dados, e combina os resultados dos classificadores através de um esquema de votação.

- *Boosting*: também cria um conjunto de classificadores através de amostragem dos dados de treino, cujos resultados são combiandos por votação (POLIKAR, 2006). Porém, a amostragem é realizada de forma diferente. Cada exemplo no conjunto de treinamento têm pesos, os quais são variados para que cada novo classificador baseie-se nos exemplos dos classificadores anteriormente gerados.

- *Stacking*: as saídas dos classificadores, treinados a partir do cojunto de dados de treino, são utilizadas como as entradas de um clasificador que fará a predição final.

Desta forma, a intenção dos *ensembles* é que, ao combinar a saída de diversos classificadores, este processo resulte em um desempenho melhor do que os desempenhos individuais dos classificadores. Han e Kamber (2012) afirma que um *ensemble* tende a ser mais preciso do que os seus classificadores base. Pois, segundo (POLIKAR, 2006), se cada classificador comete diferentes erros de predição, então uma combinação estratégica destes classificadores pode reduzir o erro total.

3.2.1 O método *Random Forest*

O método *Random Forest* provê uma série de vantagens que estimulam a sua utilização para a solução dos mais diversos tipos de problemas. Segundo Strobl et al. (2008), a identificação de variáveis preditoras relevantes, ao invés de apenas prever a resposta por meio de algum modelo de "caixa-preta", desperta o interesse pelo emprego do RF em muitas aplicações. Esta transparência que é fornecida possibilita a realização de uma série de análises em torno dos parâmetros utilizados para a solução de um determinado problema, permitindo assim, uma melhor adequação deles e possivelmente um melhor resultado.

O método *Random Forest* é constituído por uma combinação de árvores de decisão. Han e Kamber (2012, p. 18) define árvore de decisão como, um fluxograma com a estrutura de uma árvore onde cada nó denota um teste sobre o valor de um atributo, cada braço representa uma saída do teste, e as folhas da árvore representam a classes ou distribuições de classes. De acordo com Qi (2012), a amostragem aleatória e estratégias *ensemble* utilizados no RF permitem que este consiga realizar predições precisas, bem como melhores generalizações.

O RF tem como base a estratégia de *ensemble Bagging* e o algoritmo de árvores de decisão para a geração dos classificadores. Portanto, partindo do princípio do *Bagging*, o conjunto de dados de treino completo é divido — de a acordo com a quantidade de árvores estipuladas — em subconjuntos aleatórios de treinos de tamanhos iguais, os quais são treinados individualmente pelo algoritmo de árvore de decisão, gerando assim um classificador. O resultado é dado pela soma dos votos de cada classificador, sendo que, para um dado objeto, a classe atribuída é a que obtiver a maior quantidade de votos. Desta forma, de acordo com o princípio de *ensemble*, o *Random Forest* constitui o "*strong learner*"gerado a partir da combinação das as árvores de decisão intermediárias que constituem os "*weak learners*". Segundo Qi (2012), ao contrário de árvores de decisão clássicas, não há necessidade de podar árvores no RF uma vez que o esquema de *Bagging* ajuda o RF a minimizar problemas de *Overfitting*.

3.3 O APRENDIZADO DE MÁQUINA NA PREDIÇÃO DE GENES

Segundo Saeys et al. (2007), as ferramentas de predição de genes baseadas em Modelos de Markov foram as pioneiras na identificação de genes e, de acordo com Xiong (2006), os programas mais difundidos no contexto de identificação de genes em procariotos são baseados em Modelos de Markov e Programação Dinâmica. A partir disto, a pesquisa nesta área se intensificou devido a busca por predições mais precisas. Como dito na Seção 2.2.2.2, muitos programas de identificação de genes foram desenvolvidos para predição em dados gerados por experimentos genômicos. Porém, na predição de genes em um projeto metagenômico, assim como as outras etapas, se faz necessário a utilização de recursos computacionais que sejam mais adequados para lidar com as características dos dados de natureza metagenômica. Portanto, programas de predição de genes destinados especificamente para este contexto passaram a ser desenvolvidos.

MetaGene (NOGUCHI et al., 2006) possui uma arquitetura baseada em um esquema de pontuação. Primeiramente, modelos de regressão logística (um para o domínio das Archaeas e outro para Bacterias) são gerados a partir do conteúdo GC e frequências dos codons e dicodons. Além disto, são calculados os log-odds ratio (do portugês, razão de chances) do comprimento e da distância de um codon de inicio ao codon de inicio mais à esquerda dos ORFs. Estas pontuações são então combinadas, levando em conta a orientação do ORF e as distâncias dos ORFs vizinhos para computar as pontuações finais dos ORFs de uma dada sequência de entrada.

GeneMark é um conjunto de programas de predição de genes baseados em Modelos Ocultos de Markov de quinta ordem. Dentre as ferramentas, GeneMark.hmm é voltada especificamente para a identificação de genes em genomas completos de procariotos. O MetaGeneMark (ZHU et al., 2010) é destinado para a predição de genes em sequências de DNA obtidas a partir do sequenciamento de comunidades microbianas e faz uso de um modelo oculto de Markov para representar as dependências entre as frequências de oligonucleotídeos com diferentes comprimentos e conteúdo de GC de uma sequência.

FragGeneScan (RHO et al., 2010) é um programa baseado em modelos ocultos de Markov (HMM) capazes de identificar genes em genomas isolados ou fragmentos de experimentos metagenômicos, utilizando o algoritmo Viterbi para decidir o melhor caminho de estados ocultos que gera o fragmento de nucleotídeos observado.

Prodigal (HYATT et al., 2010) constrói um modelo de gene, através da programação dinâmica, analisando o conteúdo de GC nas três posições de cada codon. Depois de determinar os potenciais genes aplica um filtro através da análise do local de iniciação da tradução (TIS) e comprimento dos ORFs. MetaProdigal (HYATT et al., 2012) é uma versão do Prodigal voltada para a metagenômica, que pode identificar genes em sequências de tamanhos curtos. Esta ferramenta aplica um esquema de clusterização aos dados de treino.

Orphelia (HOFF, 2009b) é uma ferramenta *ORF-based* emprega métodos de Aprendizado de Máquina em duas etapas: a primeira aplica uma análise discriminante linear para a determinação dos valores duas variáveis (codon e dicodon usage) e a segunda utiliza Redes Neurais para a geração do modelo que realiza a predição proprimanete dita. MGC (ALLALI; ROSE, 2013) utiliza como base o Orphelia, e acrescenta duas variáveis (monoaminoacido e diaminoacido *usage*) e a construção de diferentes modelos de acordo com algumas taxas definidas de conteúdo de GC. Estas duas ferramentas caracterizam-se pela alta taxa de especificidade.

MetaGUN (LIU et al., 2013) é uma ferramenta *ORF-based* que trabalha com uma estratégia constituída por três etapas. Primeiramente é realizada a classificação dos fragmentos de entrada em grupos filogenéticos através de um método *k-mer* com base em um classificador bayesiano, em seguida os ORFs extraídos das sequências, sendo que estes são representados por um vetor de variáveis (perfil de densidade de entropia (EDP) de *codon usage*, local de iniciação da tradução (TIS) e tamanho da sequência), são classificados através do modelo treinado pelo método SVM e, por fim, os TIS de todos os genes preditos são realocados a fim de obter anotações de TIS de alta qualidade.

4 GENEFINDER-MG - UM PIPELINE PARA PREDIÇÃO DE GENES EM DADOS METAGENÔMICOS

Para embasar a escolha do método de aprendizado, utilizado no *pipeline* desenvolvido, no trabalho de Goés et al. (2014a) foi realizada uma comparação empírica do desempenho entre diferentes estratégias de classificação aplicadas no contexto da predição de genes, voltada para dados metagenômicos. Para o estudo, foram selecionados quatro métodos de classificação, os quais empregam distintas estratégias de aprendizado e possuem maneiras particulares de generalizar o espaço de busca: *K-Nearest Neighbor*, *Support Vector Machine*, Redes Neurais e *Random Forest*. Ao final do estudo, o método *Random Forest* mostou, no geral, melhor desempenho de acordo com as métricas utilizadas. A Tabela 1 e a Tabela 2 mostram os resultados Acurácia e Kappa obtidos no trabalho.

Tabela 1 – Comparação da performance dos classificadores de acordo com a medida de Acurácia. As células destacadas mostram os melhores resultados.

Espécies	RF	ANN	KNN	SVML
Ferroplasma acidarmanus	**0.9173**	0.8702	0.8302	0.785
Leptospirillum ferriphilum	**0.9156**	0.9097	0.8854	0.8835
Leptospirillum ferrooxidans	**0.9263**	0.9143	0.8888	0.8767
Sulfobacillus acidophilus	**0.9383**	0.9235	0.8913	0.8947
Thermoplasmatales archaeon BRNA	**0.9577**	0.9175	0.8875	0.9175

Fonte: Adaptada de Goés et al. (2014a)

Tabela 2 – Comparação da performance dos classificadores de acordo com a medida de Kappa. As células destacadas mostram os melhores resultados.

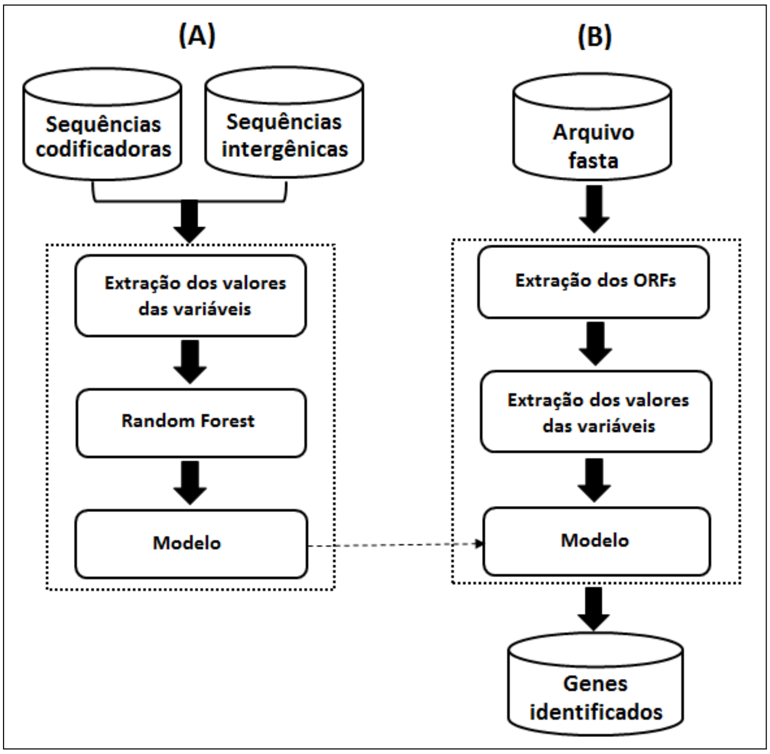
Espécies	RF	ANN	KNN	SVML
Ferroplasma acidarmanus	**0.8275**	0.7298	0.6182	0.5317
Leptospirillum ferriphilum	0.8256	**0.9097**	0.7599	0.7565
Leptospirillum ferrooxidans	**0.8472**	0.8213	0.7666	0.741
Sulfobacillus acidophilus	**0.8741**	0.8434	0.7746	0.7834
Thermoplasmatales archaeon BRNA	**0.9089**	0.8243	0.7577	0.737

Fonte: Adaptada de Goés et al. (2014a)

O tipo de predição implementada neste trabalho é a *ab initio*, pois, como destacado na Seção 2.2.2, esta abordagem se torna mais adequada, do que a busca por homologia, para a identificação de genes no contexto da Metagenômica, principalmente quando se trata da sua aplicação em sequências ainda não conhecidas e anotadas. Desta forma, a composição do *pipeline* baseia-se no Aprendizado de Máquina Supervisionado. Sendo assim, o processo de predição de genes constitui-se como uma abordagem de classificação binária, pois os valores das classes adotadas são discretos (gene e não gene), e, para tal, utiliza do método *Random Forest* para a geração do modelo classificador. A Figura 6 ilustra a arquitetura do *pipeline* desenvolvido que é composta por duas etapas, as quais são constituídas por um fluxo específico de ações.

A primeira etapa (A) é responsável pela geração do modelo classificador, ou seja, pelo aprendizado. O modo de aprendizado do *pipeline* é não incremental (*batch*); em outras palavras, para atualizar o modelo, um novo aprendizado deve ser efetuado. Primeiramente, foram selecionadas sequências que são codificadoras e sequências intergênicas, que, respectivamente, deram origem aos exemplos positivos e negativos utilizados no treinamento. Em seguida, a extração dos valores das variáveis de cada sequência é realizada para o conjunto de treino a ser montado. Por fim, a partir deste conjunto, que contém os valores das variáveis extaídas das sequências, o método *Random Forest* é aplicado e um modelo classificador é gerado.

Figura 6 – *Pipeline* GeneFinder-MG.



Fonte: Elaborada pelo autor

Com a primeira etapa executada, pode-se então realizar a predição de genes propriamente dita (B). Sequências de DNA completas e incompletas - como por exemplo *reads* e contigs gerados, respectivamente, pela etapa de sequênciamento e montagem - no formato fasta podem ser submetidas ao *pipeline*. Após a importação das sequências, o módulo de extração de ORFs é executado, e, desta forma, as as sequências candidatas a genes são obtidas. Tendo em mãos os ORFs extraídos, ocorre a extração dos valores das variáveis de cada ORF. Por fim, o modelo gerado na primeira etapa (A) é utilizado para realizar a classificação dos ORFs em gene e não gene.

4.1 ELABORAÇÃO DO CONJUNTO DE DADOS DE TREINO

Para dar origem ao conjunto de dados de treino, foram selecionados 32 organismos procariotos a partir dos 131 utilizados no trabalho de Hoff et al. (2008) e os 8 organismos utilizados no trabalho de Goés et al. (2014a), dos quais 35 são pertecentes do domínio das Bacterias e 5 das Archaeas. Todos os organismos utilizados, apresentados no Apêndice A, foram obtidos a partir da base de dados RefSeq (PRUITT et al., 2007) que é denominado como um banco de dados de sequências de referência do *National Center for Biotechnology Information* (NCBI).

Para a obtenção das regiões codificadoras (exemplos positivos), foram utilizados os arquivos contendo as anotações dos genomas, de cada um dos microorganismos escolhidos. Para a extração das regiões intergênicas, que correspondem aos exemplos negativos, foi implementado um algoritmo que realiza a extração destas regiões, a partir do genoma completo e tendo como base as localizações das regiões codificadoras (gênicas). Para a composição do conjunto de dados de treino, foram consideradas as regiões codificadoras e intergênicas das duas fitas de DNA.

Tomando como base que toda região codificadora é um ORF, como descrito na Seção 2, foi executado um novo processamento nas sequências correspondentes as regiões intergênicas. A partir de cada uma delas, foi realizada a extração de todos os seus respectivos ORFs. Desta forma, o conjunto de dados de treino final é composto por ORFs positivos e ORFs negativos, que correspondem respectivamente as regiões codificadoras (genes) anotadas, e aos ORFs extraídos das regiões intergênicas.

Após a obtenção dos ORFs positivos e negativos, foram selecionados para a composição do conjunto de treino apenas as sequências cujos valores de tamanho são maiores ou iguais a 60 pb, que, de acordo com Liu et al. (2013), correspondem tamanhos rasoáveis para a sequência carregar informações significantes. Por fim, para cada organismo foram selecionadas 410 ORFs positivos e 410 ORFs negativos, formando um conjunto de dados composto por 32.800

sequências.

4.2 VARIÁVEIS UTILIZADAS

Como discutido nas seções anteriores, o processo de aprendizado supervisionado depende crucialmente do seu conjunto de dados. Além de possuir objetos consistentes, estes devem ser constituídos por variáveis que carreguem definição significante para o problema em questão. Portanto, a de seleção das variáveis, que é utilizada para caracterizar possíveis genes, constitui uma etapa de extrema importância para o *pipeline* de predição.

Segundo Ermolaeva et al. (2001), muitos genes em genomas bacterianos são organizados em estruturas chamadas de operons, que podem ser definidas como um série de genes que são transcritos em uma única molécula de RNA. Desta forma, sensores de sinais, descritos na Seção 2, que se fazem presentes em regiões codificadoras podem estar presentes somente no início de um conjunto consecutivo de genes ao invés de estarem localizados no início de cada gene. Mathé et al. (2002) afirma que os genes de organismos procariotos podem muitas vezes se sobrepor uns aos outros e, devido a isto, as estruturas de início de tradução tornam-se difíceis de identificar corretamente. Baseado nestes detalhes, para elaboração deste trabalho, foram adotadas apenas padrões do tipo sensor de conteúdo.

No *pipeline* de predição de genes em questão, o conjunto de treinamento é composto por um atributo de saída, representando a classe, e seis atributos de entrada: conteúdo de GC, conteúdo de GC na primeira posição de cada codon, conteúdo na segunda posição de cada codon, conteúdo de GC na terceira posição de cada codon, tamanho e variância da frequência do codon *usage* da sequência. A definição dos atributos de entrada utilizados foi feita com base na realização de um estudo sobre os padrões das sequências mais importantes para a caracterização de genes, bem como os mais utilizados nas ferramentas existentes. O Quadro 1 apresenta a relação entre seis variáveis, amplamente utilizadas no contexto de identificação de regiões codificadoras de proteína, e ferramentas de predição de genes em dados metagêmicos descritas na seção 3.3.

Quadro 1 – Relação entre seis variáveis e sete programas de predição de genes.

Ferramenta	Conteúdo GC	Tamanho	Codon *usage*	Dicodon *usage*	TIS	Aminoacido *usage*
MetaGene	x		x	x		
MetaGeneMark	x					
FragGeneScan			x			
Prodigal		x			x	
MetaProdigal		x			x	
Orphelia	x	x	x	x		
MGC	x	x	x	x	x	x
MetaGUN		x	x		x	

Fonte: Elaborada pelo autor

O conteúdo de GC corresponde à quantidade de Guanina e Citosina presente uma sequência de DNA e constitui um dos padrões mais tradicionais utilizados para predição de genes, bem como para outros tipos de análises. (FICKETT, 1982) afirma que sequências codificadoras têm, em média, conteúdo de GC mais alto do que as sequências não codificadoras. Por este motivo, o conteúdo de GC foi escolhido para compor o conjunto de variáveis do *pipeline*. Além do conteúdo de GC geral, foram selecionados como variáveis o conteúdo de GC na primeira, na segunda e na terceira posição dos codons, a fim de se observar o impacto desses atributos no processo de predição.

O tamanho de uma dada sequência é utilizado como variável, baseando-se na diferença de tamanho característico que as sequências gênicas e intergênicas (regiões não codificadoras de proteínas) possuem, pois, segundo (MATHÉ et al., 2002), os ORFs das regiões intergênicas caracterizam-se por serem menores em comparação às regiões gênicas.

Segundo Hoff et al. (2008), possivelmente as mais importantes variáveis utilizadas para a discriminação entre regiões codificadoras e não codificadoras tenham sido derivadas a partir do codon *usage*. Este é caracterizado como a frequência com que os vários codons alternativos são usados para a codificação de aminoácidos em uma sequência de DNA. Devido ao fato de que existem 4 nucleotídeos, e cada códon é formado por três nucleotídeos, existem 64 codons diferentes que podem compor as sequências de DNA (61 codons que codificam os aminoácidos mais 3 codons de parada). Uma das variáveis do *pipeline* é composta pela extração de um vetor com as 64 frequências da utilização de cada codons, em uma dada sequência, e do cálculo da variância dos 61 codons que codificam aminoácidos, já que genes não possuem codons de parada em sua extensão.

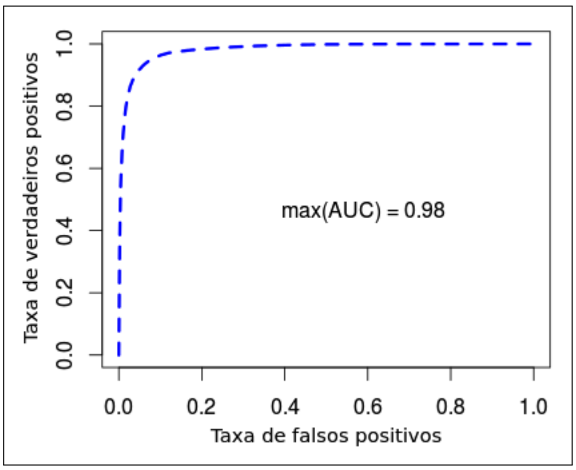
Os valores de conteúdo de GC, conteúdo de GC na primeira, segunda, e terceira posição e o vetor de frequências dos 64 codons foram obtidos a partir da utilização de funções fornecidas pelo pacote Seqinr (SEQINR. . . , 2003) da linguagem R.

4.3 GERAÇÃO E ANÁLISE DO MODELO

O suporte para a implementação da etapa de aprendizado de máquina foi dado pelo pacote Caret (KUHN, 2010), fornecido pela linguagem R. Através dele, fez-se possível a utilização do método *Random Forest* e a parametrização para a geração do modelo, como o número de árvores utilizadas, e a técnica de avaliação utilizada. Para a avaliação, durante o treinamento, foi utilizada a estratégia de validação cruzada de 10-fold com 3 repetições.

A curva ROC (*Receiver Operator Characteristic Curve*) é uma forma de realizar avaliação de classificadores em problemas binário, da qual o gráfico é plotado de acordo com a taxa de verdadeiros positivos (eixo y) e falsos positivos (eixo x). A partir da curva ROC, pode-se extrair a medida AUC (*Area Under the Curve*) que produz valores entre 0 e 1, e, em um caso de comparação entre modelos, esta medida tem a vantagem de fornecer um valor único que mede a capacidade geral de discriminação dos classificadores. Faceli et al. (2011) afirma que, aconselha-se calcular o AUC em um procedimento de validação cruzada, obtendo a média e desvio padrão de seus valores para diferentes partições dos dados. Desta forma, usufruindo do esquema de validação cruzada, utilizado durante o treinamento, foi plotada a curva ROC assim como o valor máximo de AUC, que são mostrados na Figura 7.

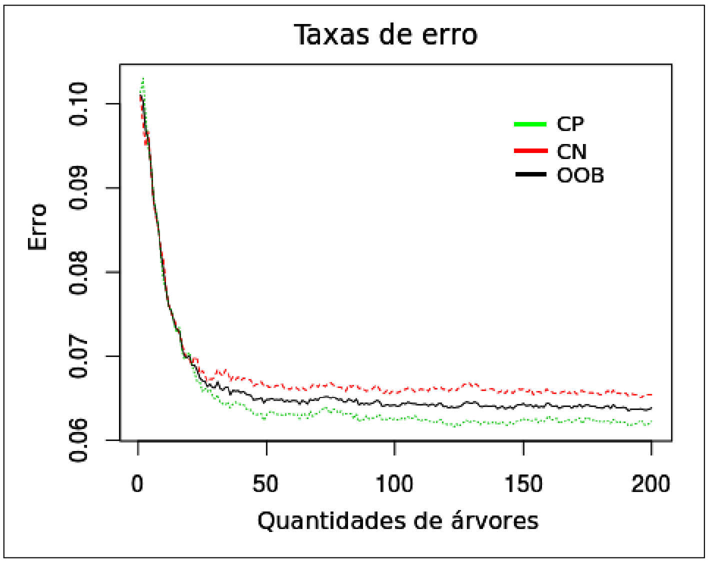
Figura 7 – Curva ROC e o valor máximo da medida AUC do classificador utilizado no *pipeline*.



Fonte: os autores

Um parâmetro que é relacionado com o desempenho do modelo é o número de árvores que a "*forest*" conterá. A fim de determinar a quantidade de árvores que seriam utilizadas para elaboração do modelo foi realizado um teste prévio através da geração de um modelo com 500 árvores, a partir do qual foi possível perceber que com 200 árvores as taxas de erro se estabilizaram. Desta forma, foi escolhida a solução mais simples (modelo de 200 árvores) em detrimento de outras mais complexas (modelos com mais de 200 árvores). A Figura 8 ilustra a variação da taxa de erro do modelo (para a classe negativa, positiva e *out-of-bag* (OOB)) de acordo com o aumento do número de árvores; pode-se perceber que as taxas diminuem à medida que o número de árvores no *ensemble* aumenta. A taxa de erro OOB, que do modelo em questão tem o valor de 6.38%, é estimada pelo próprio algoritmo *Random Forest* da seguinte forma: de cada subconjunto (amostra) de dados, selecionados a partir do conjunto de treino geral, dois terços é destinado ao crescimento da árvore (*bag*) e um terço é reservado para teste da árvore (*out-of-bag*), através do qual é estimada a taxa de erro durante o crescimento dela.

Figura 8 – Taxas de erro para o modelo treinado. Em vermelho a taxa de erro para a classe negativa (CN), em preto para OOB e em verde para classe positiva (CP).



Fonte: os autores

A performance dos resultados foi medida, no decorrer do treinamento por meio do esquema de validação cruzada, através das métricas de Acurácia e Kappa. Ambas as métricas utilizadas na análise do modelo são computadas tendo como base a matriz de confusão. Esta matriz quantifica o número de objetos dos dados de teste classificados como verdadeiro positivo (VP), falso positivo (FP), verdadeiro negativo (VN) e falso negativo (FN). Kappa é uma medida estatística de aceitação entre os dados reais e dados preditos, ou seja uma medida de concordância, dada pela fórmula:

$$Kappa = \frac{P(a) - P(e)}{1 - Pr(e)} \tag{4.1}$$

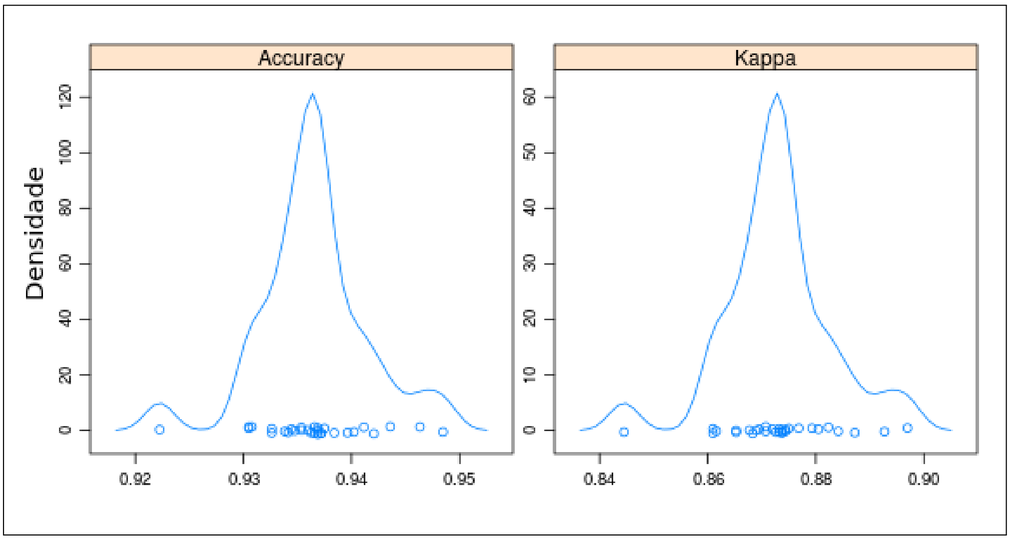
onde P(a) é a observação relativa de aceitação entre os avaliadores e P(e) é hipoteticamente a probabilidade de acerto.

A Acurácia é a proporção do número total de predições corretas, dada pela fórmula:

$$ACC = \frac{VP + VN}{P + N} \tag{4.2}$$

A utilização de validação cruzada possibilitou a análise do comportamento da distribuição dos resultados de Acurácia e Kappa de acordo com as três repetições de 10-fold. Pode-se perceber que os valores de Acurácia se concentram entre 0.93 e 0.94 e os de Kappa entre 0.86 e 0.88, como ilustra a Figura 9.

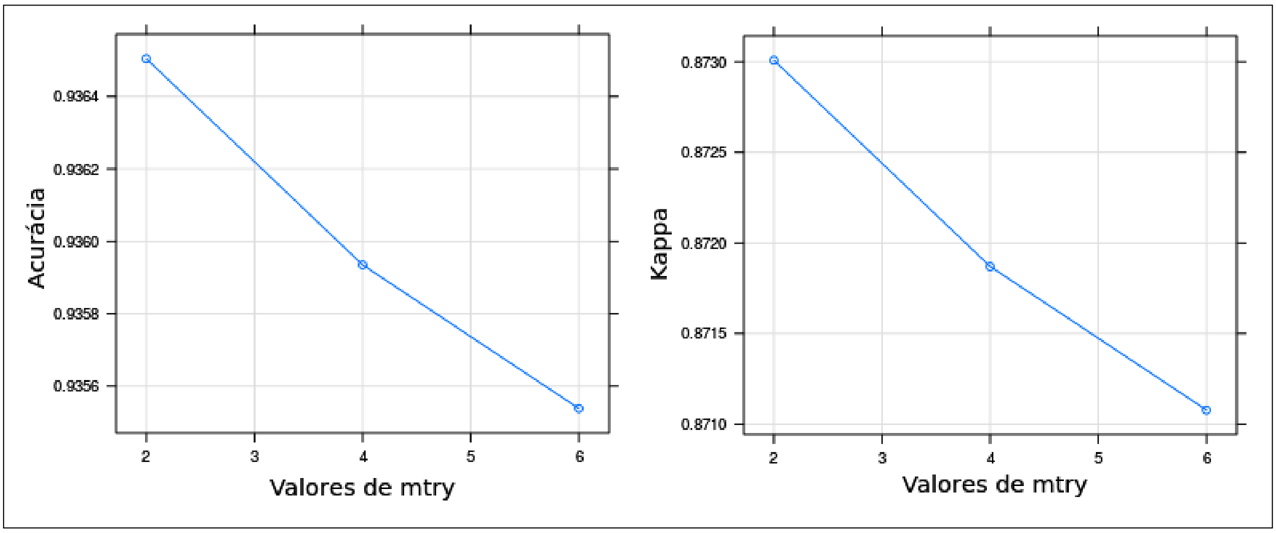
Figura 9 – Distribuição dos valores de Acurácia e Kappa obtidos a partir da validação cruzada de 10-fold com três repetições.



Fonte: os autores

Segundo Mendoza et al. (2013), o *Random Forest* é conhecido por ser sensível ao parâmetro mtry (denominação utilizada pelo pacote *Random Forest* (LIAW; WIENER, 2002) da linguagem R), que corresponde ao número de variáveis selecionadas aleatoriamente, do conjunto de atributos, para serem utilizadas nos nós da árvore. Desta forma, foi observada a variação dos Valores de Acurácia e Kappa de acordo com o número de variáveis, selecionadas aleatoriamente em cada nó, a ser considerado para o processo de divisão (ramificação) de todas as árvores construídas. A diferença entre resultados obtidos através das variações do parâmetro mtry empregadas, durante o processo de treinamento, não foi significativa, porém o valor final do mtry foi definido como 2 pois alcançou os maiores valores de Aurácia e Kappa, como pode ser visto na Figura 10.

Figura 10 – Valores de Acurácia e Kappa de acordo com a variação (2, 4 e 6) do parâmetro mtry.



Fonte: os autores

Segundo Liu et al. (2013), fragmentos de sequências presentes em metagenomas, em muitos casos, podem ser originados de diversas espécies e, devido a isto, um dos maiores desafios é como treinar modelos estatísticos que podessam capturar corretamente características de sequências de diferentes genomas de origem. Sendo assim, partindo desta questão, foi executado um teste no qual a mesma preparação de dados descrita na Seção 4.1 foi aplicada a genomas de organismos - que não estão contidos conjunto de dados de treinos - apresentados na Tabela 2. A avaliação foi realizada em termos das métricas de Sensibilidade e Especificidade. A Sensibilidade é a proporção de verdadeiros positivos, ou seja, avalia a capacidade do modelo classificar uma sequência como gene dado que realmente ele é gene, reprensetada pela fórmula:

$$SEN = \frac{VP}{VP + FN} \tag{4.3}$$

A Especificidade é a proporção de verdadeiros negativos, isto é, avalia a capacidade do modelo predizer uma sequência como não gene dado que ele realmente é não gene, reprensetada pela fórmula:

$$ESPEC = \frac{VN}{VN + FP} \tag{4.4}$$

Desta forma, é possível observar como o modelo se comporta quando aplicado a dados extraídos de genomas de organismos que não foram utilizados na etapa de aprendizado. Os valores estão dispostos na Tabela 3 e pode-se, então, perceber que os valores de sensibilidade variam de 0.93 a 0.96 e os de especificidade variam 0.92 a 0.98, de dependendo o organismo de teste.

Tabela 3 – Valores de sensibilidade e Especificidade.

Espécies	Sensibilidade	Especificidade
Archaeoglobus fulgidus	0.94	0.94
Natronomonas pharaonis	0.93	0.95
Buchnera aphidicola	0.95	0.98
Burkholderia pseudomallei	0.94	0.96
Bacillus subtilis	0.95	0.93
Corynebacterium jeikeium	0.95	0.96
Chlorobium tepidum	0.94	0.93
Escherichia col	0.93	0.92
Helicobacter pylori	0.94	0.96
Pseudomonas aeruginosa	0.96	0.98
Prochlorococcus marinus	0.93	0.92
Wolbachia endosymbiont	0.94	0.93

Fonte: Elaborada pelo autor

5 RESULTADOS EXPERIMENTAIS

A partir do teste inicial, apresentado na Seção 4, implementado com o objetivo de analizar a capacidade de classificação do modelo quando aplicado dados extraídos de organismos não presentes nos dados de treino, se faz, então, necessario avaliar o desempenho do classificador quando aplicado em um cenário de teste que simula a fragmentação das sequências que são submetidas para a predição de genes. Esta questão, intregra-se aos desafios presententes em experimentos metagenômicos.

Desta forma, para avaliar a performance estratégia de aprendizado de máquina desenvolvida, testamos o *pipeline* em genomas artificialmente fragmentados. Nas próximas seções, serão apresentados a elababoração do cenário de teste, que inclui uma comparação com outras quatro ferramentas, e os resultados de desempenho da predição de genes em fragmentos de sequências de vários tamanhos.

5.1 ELABORAÇÃO DOS DADOS DE TESTE

O teste experimental é baseado em um conjunto de dados constituído por fragmentos artificiais de DNA obtidos a partir dos genomas de dois organismos pertencentes ao domínio das archaeas e dez organismos pertencentes ao domínio das bacterias, aprensentados no Quadro 2. É importante destacar que estes organismos não não integram o conjunto de dados de treino e que a escolha destes foi baseada nos trabalhos de Noguchi et al. (2006), Hoff et al. (2008) e Liu et al. (2013).

Para a geração dos *reads*, foi utilizado o simulador de sequenciamento MetaSim (RICHTER et al., 2008) que simula o sequenciamento das plataformas 454, Sanger e Illumina, além de gerar *reads* sem erros advindos de sequenciamentos. Para esta análise experimental, foram utilizados apenas fragmentos livres de erros como nos trabalhos de Hoff et al. (2008) e Liu et al. (2013). Desta forma, as medidas de desempenho foram calculadas a partir das comparações das predições das ferramentas com os genes anotados presentes nos fragmentos. Para tal, as localizações das regiões codificadoras foram utilizadas para a contabilização das predições corretas e incorretas.

Quadro 2 – Espécies microbianas que foram utilizadas para a avaliação experimental e os seus respectivos números de acesso RefSeq. As espécies destacadas com "*"são archaeas, enquanto demais pertencem ao domínio das bactérias.

Espécies	RefSeq Acc.
*Archaeoglobus fulgidus**	NC_000917
*Natronomonas pharaonis**	NC_007426
Buchnera aphidicola	NC_002528
Burkholderia pseudomallei	NC_006350
Bacillus subtilis	NC_000964
Corynebacterium jeikeium	NC_007164
Chlorobium tepidum	NC_002932
Escherichia coli	NC_000913
Helicobacter pylori	NC_000921
Pseudomonas aeruginosa	NC_002516
Prochlorococcus marinus	NC_007577
Wolbachia endosymbiont	NC_006833

Fonte: Elaborada pelo autor

Para cada organismo presente no Quadro 2 foram gerados aleatoriamente 1000 fragmentos de comprimentos de 300, 500, 900 e 1200 pb. Esta variação de tamanho foi aplicada a fim de serem analisados os desempenhos de cada ferramenta de predição de gene, selecionadas para o teste, de acordo com a manipulação de dados de diferentes fragmentos. Desta forma, pode-se observar o impacto desta variação na capacidade de classificação de cada ferramenta e, por fim, comparar os pontos fortes e fracos de cada uma.

A Figura 11 ilustra um dos *reads*, de comprimento de 700 pb, gerado pelo simulador a partir do genoma do organismo *Buchnera aphidicola*. Com base nesta imagem é possível perceber uma das dificuldades da realização da predição de genes em *reads*: a fragmentação das regiões codificadoras. Este *read* cobre uma região intergênica (em vermelho) e duas partes de duas regiões codificadoras (em verde) que excedem as extremidades do fragmento. Desta forma as ferramentas de predição devem possuir a capacidade de lidar com a identificação de genes em fragmentos que não contém a informação completa destas regiões.

Como dito na Seção 2, a predição de genes pode ser feita em contigs gerados na etapa de montagem, porém em alguns experimentos metagenômicos este processo pode não ser viável e a identificação passa a ser realizada a partir de *reads*. Então, por este motivo, para tentar abranger esse dois casos foram empregados fragmentos curtos (300 pb) e mais mais extensos (1200 pb), como no trabalho de Liu et al. (2013), a fim de simular as duas situações nas quais esse tipo de predição de genes pode ser aplicada.

Figura 11 – Um *read* do organismo *Buchnera aphidicola*. Estão destacadas de verde dois trechos de duas regiões codificadoras e em vermelho uma região intergênica.

```
>r8.1 |SOURCES={GI=15616630,fw,227196-227896}|ERRORS={}|SOURCE_1=
"Buchnera aphidicola str. APS (Acyrthosiphon pisum) chromosome"
(3086b174e1033d1f2184874a52d7dadb69d14915)

ATAATAAAAGTAGTTGCATATGCATTAGAAAAATTCCCTATTTTTAACAGTTCACTAAATATTAA
TAATAAAAAAATTATTTTAAAAAAATATATTAATATCGGGTTTGCAATAGATGTAAATAACGATT
TATTTGTTCCTGTTTTAAAAGACGTCAATAAAAAAAAATATTAAACAATTATCTTCTGAATTAATA
TTGCTATCAGAAAAAGCACGAACAAGAAAATTAAATATTGAAGATATGACAGGAGGCTGTTTTAC
AATATCTAATTTAGGAGGTATTGGAGGAAGCTGGTTTTCACCAATTATCAATTCACCGGAAGTAG
CAATTCTTGGTATTTCAAAATCTCAGATAAAACCGTCATGGAATGGAAAAGAATTTATTCCTTCT
TTAATGTTACCATTATCTTTATCTTATGATCATCGTGTAATAAATGGTGCTTATGCAGCGCGTTT
TATTACATTTATTAGTAGAGTTTTATCAGATATGCATTTTTTAATTATGTAGTTTTCGCTATTTC
ACATTACTTAAATTTTTTTTAAGAGGTCTTAATGAATAAAAAAATTTATACACAAGTAGTAGTTA
TTGGATCAGGTCCAGCAGGTTACTCTGCAGCTTTTCGTTGTGCAGATCTAGGTTTAGATACTGTC
TTAATAGAACGTTATGATAAATTAGGAGGTGTTTGTTTAAATGTTGGTTG
```

Fonte: Elaborada pelo autor

5.2 BENCHMARKING

A capacidade de detectar genes foi medida como sensibilidade (SEN). Para a medição da sensibilidade da predição de genes, VP (verdadeiro positivo) denota as sobreposições corretas, ou seja com genes, e FN (falso negativo) indica genes que não foram identificados. Apenas as sobreprosições de pelo menos 60 pb com genes anotados foram contabilizadas como VP.

$$SEN = \frac{VP}{VP + FN} * 100 \tag{5.1}$$

As especificidades (ESPEC) das predições foram calculadas em relação aos genes preditos que não correspondem a qualquer gene na anotação, ou seja como FP (falsos positivos). Esta definição de especificidade difere da tradicional utilizada no Aprendizado de Máquina que, por sua vez, mede a taxa de acerto da classe negativa. Porém, neste contexto, se torna interessante realizar esta medida em relação aos falsos positivos, ou seja medir quanto as ferramentas classificam uma região intergênica como gene. Então, assim como nos trabalhos de Noguchi et al. (2006), Hoff et al. (2008), Rho et al. (2010) e Liu et al. (2013) a medida de especificidade é dada pela seguinte fórmula:

$$ESPEC = \frac{VP}{VP + FP} * 100 \tag{5.2}$$

Para avaliar o desempenho do *pipeline* elaborado, testamos o *pipeline* desenvolvido

(GeneFinder-MG), FragGeneScan, MetaGene, Prodigal e Orphelia em genomas artificialmente fragmentados. Sendo assim, apresentamos os resultados dos desempenhos das predições em fragmentos de vários comprimentos. Os resultados finais de cada ferramenta, para cada tamanho de fragmento, foram obtidos através da média dos valores de sensibilidade (Tabela 4) e especificidade (Tabela 5) de cada organismo empregado no teste.

Através da Tabela 4 pode-se perceber que a ferramenta FragGene teve perda nos valores de sensibilidade conforme o aumento do tamanho dos fragmentos. MetaGene teve resultados mais estáveis em relação a variação dos tamanhos dos fragmentos. Orphelia e GeneFinder-MG apresentaram crescimento no desempenho conforme o aumento dos tamanhos dos fragmentos e Prodigal teve performance semelhante, porém teve perda no valor sensibilidade em *reads* de 900 pb.

Tabela 4 – Valores de sensibilidade de cada ferramenta. O desempenho foi medido em fragmentos de 300, 500, 700, 900, 1200 pb gerados aleatoriamente a partir de cada genoma teste. Cada célula da tabela é composta pela média dos resultados de cada organismo de teste.

Tamanho	FragGene	MetaGene	Prodigal	Orphelia	GeneFinder-MG
300pb	93.1	90.6	51.9	88.7	39.8
500pb	92.6	92.8	56.7	89.8	64.6
700pb	91.6	92.0	64.3	89.9	72.6
900pb	91.2	91.4	60.8	90.0	76.4
1200pb	90.6	90.5	64.5	90.5	78.8

Fonte: Elaborada pelo autor

A Tabela 5 mostra que Orphelia obteve as maiores taxas de especificidades em todos os tamanhos de fragmento, seguida por GeneFinder-MG, MetaGene e FragGeneScan. Isto que significa que, no geral, de 7% a 10%, de acordo com os casos de variação de tamanho, Orpheia classifica ORFs pertencentes a regiões intergênicas como ORFs codificadores (possuem sobreposição com uma região anotada do seu respectivo genoma).

Tabela 5 – Valores de especificidade de cada ferramenta. O desempenho foi medido em fragmentos de 300, 500, 700, 900, 1200 pb gerados aleatoriamente a partir de cada genoma teste. Cada célula da tabela é composta pela média dos resultados de cada organismo de teste.

Tamanho	FragGene	MetaGene	Prodigal	Orphelia	GeneFinder-MG
300pb	90.4	87.1	54.1	93.8	90.5
500pb	89.0	89.9	57.0	93.2	90.6
700pb	87.8	89.5	63.8	92.2	89.7
900pb	87.5	89.3	60.1	91.7	88.8
1200pb	86.4	88.8	63.8	90.8	88.0

Fonte: Elaborada pelo autor

Observando as predições das ferramentas juntamente com as coordenadas das regiões anotadas dos genomas, percebeu-se que algumas sequências preditas, pelas ferramentas, como genes possuem sobreposição com uma região anotada, porém ultrapassam o limite dessa região e, desta forma, se sobrepõem também a uma região intergênica. Então, novos resultados foram gerados em relação a especificidade das ferramentas, desta vez as predições que incluem uma gerião codificadora e uma intergênica são contabilizadas como um VP e um FP, ao invés de serem contabilizadas apenas como VP. A Tabela 6 apresenta os novos valores de especificidade para cada uma das ferramentas e, a pertir dela, pode-se perceber que houve uma queda significativa em relação aos valores presentes na Tabela 5. Isto indica que, em média, 20% das predições corretas não são "exatas", ou seja, não possuem sobreposição apenas em regiões codificadoras. Pode-se perceber que, nesta análise o *pipeline* GeneFinder-MG mostrou realizar uma predição mais "exata"em comparação com as predições das outras ferramentas.

Tabela 6 – Valores de especificidade de cada ferramenta. Os trechos de regiões intergênicas presentes em predições corretas, foram contabilizados como FP. Cada célula da tabela é composta pela média dos resultados de cada organismo de teste.

Tamanho	FragGene	MetaGene	Prodigal	Orphelia	GeneFinder-MG
300pb	76.6	74.3	46.7	79.6	85.0
500pb	72.2	73.4	47.4	75.2	80.4
700pb	69.5	71.2	51.6	72.7	77.0
900pb	67.7	69.5	48.0	70.9	73.8
1200pb	65.8	67.7	49.0	68.9	71.2

Fonte: Elaborada pelo autor

Apesar do *pipeline* não superar as outras ferramentas em termos de sensibilidade, os resultados de especificidade do *pipeline* GeneFinder-MG são expressivos, perdendo apenas para a Orphelia na análise da tabela 5 e superando todas as ferramentas na análise da tabela 6. Desta forma, conclui-se que o preditor tende a acertar quando classifica uma sequência como gene e, portanto, possuir menos ocorrências de falsos positivos.

A Tabela 7 aprenta os valores de sensibilidade de cada ferrementa, quando aplicadas a fragmentos de 700 pb, para cada um dos organismos utilizados no teste. Através dela é possível observar que, mesmo em fragmentos do mesmo tramanho, as ferramentas tem valores de sensibilidade diferentes para cada organismo. As variações dos valores de cada ferramenta são de :

- FragGeneScan de 6,4% (88.3% a 94.7%);
- MetaGene de 8.0% (86.8% a 94.8%);
- Prodigal de 20.1% (55.4% a 75.5%);
- Orphelia de 13,3% (80.7% a 94.0%);
- GeneFinder-MG de 18.7% (63.3% a 82.0%).

Tabela 7 – Valores de sensibilidade, para cada organismo, em fragmentos de tamanho de 700 pb

Espécies	FragGene	MetaGene	Prodigal	Orphelia	GeneFinder-MG
Archaeoglobus fulgidus	88.8	93.1	75.5	90.4	74.9
Natronomonas pharaonis	90.1	93.7	62.3	90.8	70.2
Buchnera aphidicola	94.0	94.8	59.6	92.0	75.2
Burkholderia pseudomallei	93.3	93.2	58.6	91.2	79.5
Bacillus subtilis	94.3	92.6	68.2	91.8	73.1
Corynebacterium jeikeium	91.1	91.2	68.4	89.7	72.6
Chlorobium tepidum	88.9	88.6	59.5	87.8	75.0
Escherichia coli	92.6	92.2	68.9	90.8	68.9
Helicobacter pylori	94.7	94.0	60.7	94.0	73.0
Pseudomonas aeruginosa	93.0	94.0	60.0	92.7	82.0
Prochlorococcus marinus	88.9	90.1	74.3	84.6	63.3
Wolbachia endosymbiont	88.3	86.8	55.4	80.7	64.3

Fonte: Elaborada pelo autor

A partir dos resultados da Tabela 7, foi realizada uma análise sobre a composição dos genomas adotados - em termos de densidade gênica, tamanho médio de gene e conteúdo de GC - a fim de investigar um possível causa para a variação, entre os organismos, dos vaores de sensibilidade. A densidade gênica dos doze organismos foi caulcuda através da razão entre a quantidade de genes e o número de pares de bases do genoma que, por sua vez, está na unidade megabases (Mb), ou seja, em termos de um milhão de pares de bases.

Os desempenhos mais baixos das ferramentas, em relação a sensibilidade (Tabela 7), são em organismos cuja a densidade gênica é maior e, por tanto, possuem um tamanho médio dos genes menor, como pode ser observado na Tabela 8. Sendo assim, esta análise indica uma possível deficiência das ferramentas em genomas com essas caracterísicas estruturais.

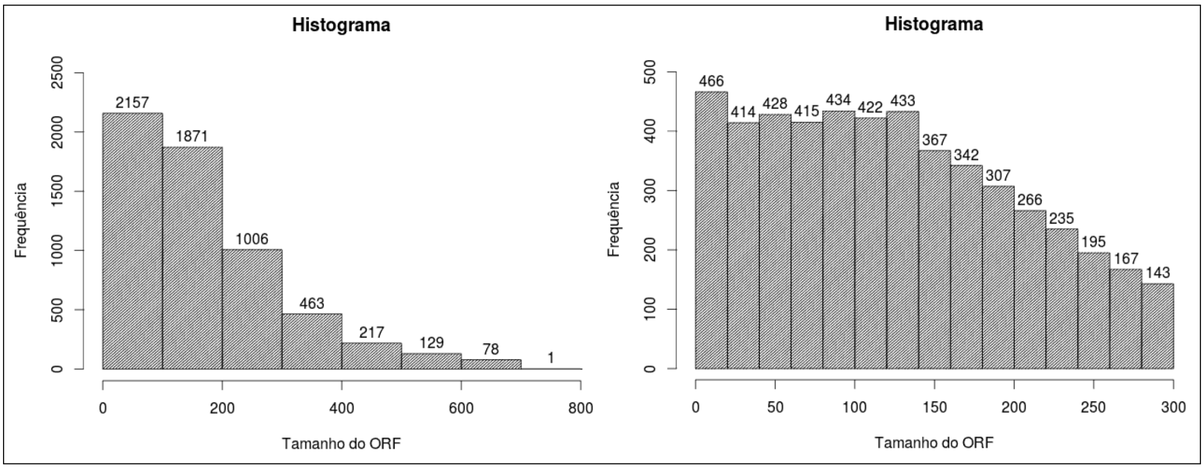
Tabela 8 – Valores de densidade gênica (genes/Mb) e tamanhos médios dos genes de cada organismo utilizado no teste.

Espécies	Densidade gênica	Tamanho médio do gene
Archaeoglobus fulgidus	1104.5	834.5
Natronomonas pharaonis	990.3	899.3
Buchnera aphidicola	880.3	986.9
Burkholderia pseudomallei	834.0	997.6
Bacillus subtilis	990.4	883.7
Corynebacterium jeikeium	847.5	1047.3
Chlorobium tepidum	1041.8	843.8
Escherichia coli	930.5	936.9
Helicobacter pylori	866.4	987.4
Pseudomonas aeruginosa	889.5	1002.6
Prochlorococcus marinus	1065.4	837
Wolbachia endosymbiont	875.9	885.5

Fonte: Elaborada pelo autor

Partindo dos resultados de sensibilidade do *pipeline* (GeneFinder-MG) desenvolvido - que não foram superiores aos das ferramentas FragGeneScan, MetaGene e Orphelia - uma análise foi feita a fim de averiguar qual ponto pode estar influenciando negativamente no seu desempenho. Para isto, foram registrados todos os ORFs, extraídos dos 12 organismos utilizados na comparação, que foram classificados como ORFs intergênicos mas que, na verdade, pertencem a regiões codificadoras presentes nos *reads*. Através de um histograma (Figura 12), pode-se observar que a maior concetração de classificação incorreta é composta por ORFs de tamanhos pequenos, com destaque para os tamanhos de 60 pb a 150 pb. Desta forma, pode-se concluir que o preditor do *pipeline* não consegue ter uma boa generalização em casos de ORFs que estão nesta faixa de tamanho. Em relação aos ORFs de tamanhos a partir de 200 pb, que possuem uma concetração menor no histograma, pode-se estabelecer a hipótese de que o preditor não consiga generalizar em alguns casos de ORFs extraídos de regiões codificadoras (que também constituem-se como ORFs), ou seja, em subORFs de regiões codificadoras o modelo classificador não consegue generalizar tão bem a ponto de classificá-los como parte de uma região codificadora.

Figura 12 – Histograma dos tamanhos dos ORFs extraídos de regiões dos *reads* que são codificadoras, mas que foram classificados como ORFs não codificadores.



Fonte: Elaborada pelo autor

6 CONCLUSÕES

Neste Capítulo são apresentadas, de forma geral, as principais conclusões obtidas com o desenvolvimento deste trabalho. Inicialmente, apresentam-se as considerações finais relacionadas à proposta na qual este trabalho é fundamentado. Em seguida, apresentam-se as limitações identificadas no decorrer da elebaração e conclusão da ferramenta. Por fim, são indicados os trabalhos futuros que poderão ser desenvolvidos a partir do que foi realizado por esta pesquisa.

6.1 CONSIDERAÇÕES FINAIS

A Bioinformática surge para apoiar o desenvolvimento de técnicas que proporcionem a solução de diversos tipos de problemas da área biológica. Uma das áreas que é sustentada pelo avanço na elaboração dessas técnicas é a Metagenômica que se tornou essencial para uma ampla gama de aplicações, devido ao ganho provido pelo estudo e melhor conhecimento de comunidades microbianas.

A predição de genes se caracteriza como uma análise importante para o entendimento de diversos problemas biológicos, sendo esta fortemente auxiliada por recursos computacionais. A predição de genes do tipo *ab initio* é embasada nas técnicas e métodos computacionais do Aprendizado de Máquina e a predição por homologia utiliza algoritmos voltados para a comparação entre sequências. Constantemente são desenvolvidas novas ferramentas de predição de gene, devido a busca por um processo de identificação de genes cada vez mais preciso. A predição de genes *ab initio* possui como estado da arte o uso de Modelos de Markov porém, analizando os trabalhos presentes na literatura, é possível perceber uma mudança neste cenário no qual novos métodos vêm sendo aplicados, como, por exemplo, Redes Neurais e *Suporte Vector Machine*, dentre outros.

Baseado neste contexto, esta dissertação propôs-se um novo *pipeline* para predição de genes em dados metagenômicos por meio da utilização de um método *ensemble learning*. Desta forma, este trabalho apresentou o *pipeline* elaborado, bem como uma análise sobre o preditor utilizado e execução de testes experimentais jutamente com uma comparação com outras quatro ferramentas de predição de genes deste âmbito. Esta solução computacional é baseada no Aprendizado de Máquina através do uso do método computacional *Random Forest* e de um conjunto de variáveis que correspondem a padrões extraídos de sequências de DNA.

O *pipeline* apresentou um resultado muito significativo quando aplicado a regiões codificadoras completas. Porém, no teste experimental realizado, no qual o *pipeline* foi aplicado a *reads* (sequências mais fragmentadas) simulados, o desempenho em termos de sensibilidade ficou abaixo do esperado quando comparado com outras quatro ferramentas selecionadas. Por outro lado, o desempenho em nível de especifidade constituí-se como o ponto positivo do *pipeline*.

6.2 LIMITAÇÕES

Apesar do esforço empregado na elaboração desta pesquisa, é certo que muito ainda pode ser realizado para o seu aperfeiçomento. Para que o *pipeline* alcance um melhor grau de competição em nível de sensibilidade com as ferramentas existentes, indentificou-se a necessidade de refinamento na predição em fragmentos pequenos, como os 300 pb, e para tal é indipensável que a classificação de ORFs menores (60 a 150 pb) seja mais precisa. Assim como, na predição em fragmentos na faixa de tamanho de 200 a 300 pb que, em alguns casos, não se obteve corretude.

6.3 MELHORIAS E TRABALHOS FUTUROS

O cenário atual permitiu o desenvolvimento completo da ferramenta proposta, jutamente com o módulo de *benchmarking*. Desta forma, deixa-se em aberto apenas melhorias futuras, que poderão ser aplicadas a qualquer módulo do *pipeline*, para a geração de novos resultados.

6.3.1 Melhorias

Ao longo do texto deste trabalho foi enfatizada a importância dos dados de treino para a capacidade de generalização do modelo gerado. A partir dos resultados experimentais observou-se que o preditor não conseguiu generalizar de forma satisfatória em alguns subORFs de regiões codificadoras e, sendo assim, pode-se interpretar que, de uma forma geral, estas sequências não carregam informações suficientes para o modelo consiga classificá-las como partes de regiões codificadoras. Desta forma, é possível perceber a necessidade de testes a fim da obtenção de um nova estratégia para a construção dos dados de treino na qual possívelmente envolverá ORFs extraídos de regiões codificadoras.

Um dos prontos críticos do processo de aprendizado do modelo é a escolha do conjunto de variáveis. Pretende-se, então, testar o impacto de outras variáveis no processo de geração e análise experimental do preditor. Pois, apenas o acréscimo de uma nova variável pode mudar a capacidade de generalização do modelo e gerar novos resultados e análises.

Além de implementar as melhorias citadas acima, um ponto interessante de ser verificado é a variação dos valores de sensibilidade, para fragmentos de mesmo tamanho, entre os organismos. Pensando nisso, uma análise em torno desta questão pode ser feita, a fim de que a ferramenta consiga ter um uma precisão mais homogênea. Fato que, se consumado, será uma nova contribuição para o contexto da predição de genes em dados metagenômicos.

6.3.2 Trabalhos Futuros

Metagenômica é uma abordagem poderosa que pode ser utilizada para descrever o potencial genético dos microorganismos presentes num dado ambiente. No entanto, ele tem uma função muito limitada em revelar a sua actividade ou expressão do gene (WANG et al., 2015). Enquanto a Metagenômica tenta compreender o repertório genético de uma comunidade microbiana, a Metatranscriptômica captura as atividades da comunidade através de informações da expressão gênica envolvida. Porém, além disto, para um estudo comparativo, os dados advindos da Metatranscriptômica devem ser combinados com os dados resultantes de análises da Metagenômica a fim de compreender se a abundância dos transcritos reflete em mudanças na composição da comunidade (BIKEL et al., 2015).

Tendo em vista as vantagens que a Metatranscriptoma traz para a análise de comunidades microbianas pretende-se agregar esta pesquisa de predicação de genes em dados Metagenômicos, através da ultização do método *Random Forest*, mais uma etapa: a identificação de genes em RNA transcritos.

REFERÊNCIAS

ALLALI, A. E.; ROSE, J. R. Mgc: a metagenomic gene caller. **BMC bioinformatics**, BioMed Central, v. 14, n. 9, p. 1, 2013.

ANDRADE, P.; QUEIROZ, R.; KIDO, **Genética e Biologia Molecular - Portal de informação em genética e biologia molecular e áreas afins**. 2014. Disponível em: <https://www.ufpe.br/biolmol/Genetica-Medicina/genes-estrutura_e_organizacao.htm>.

ANGELOVA, M.; KALAJDZISKI, S.; KOCAREV, L. Computational methods for gene finding in prokaryotes. **ICT Innovations**, Citeseer, p. 11–20, 2010.

BASHIR, Y.; SINGH, S. P.; KONWAR, B. K. Metagenomics: An application based perspective. **Chinese Journal of Biology**, Hindawi Publishing Corporation, v. 2014, 2014.

BIKEL, S.; VALDEZ-LARA, A.; CORNEJO-GRANADOS, F.; RICO, K.; CANIZALES-QUINTEROS, S.; SOBERÓN, X.; POZO-YAUNER, L. D.; OCHOA-LEYVA, A. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. **Computational and structural biotechnology journal**, Elsevier, v. 13, p. 390–401, 2015.

DÍAZ-URIARTE, R.; ANDRES, S. A. D. Gene selection and classification of microarray data using random forest. **BMC bioinformatics**, BioMed Central Ltd, v. 7, n. 1, p. 3, 2006.

DOMINGOS, P. A few useful things to know about machine learning. **Communications of the ACM**, ACM, v. 55, n. 10, p. 78–87, 2012.

ERMOLAEVA, M. D.; WHITE, O.; SALZBERG, S. L. Prediction of operons in microbial genomes. **Nucleic acids research**, Oxford Univ Press, v. 29, n. 5, p. 1216–1221, 2001.

FACELI, K.; LORENA, C.; GAMA, J.; CARVALHO, A. **Inteligência artificial: uma abordagem de aprendizado de máquina**. [S.l.]: Grupo Gen-LTC, 2011.

FICKETT, J. W. Recognition of protein coding regions in dna sequences. **Nucleic acids research**, Oxford Univ Press, v. 10, n. 17, p. 5303–5318, 1982.

FILIPPO, C. D.; RAMAZZOTTI, M.; FONTANA, P.; CAVALIERI, D. Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. **Briefings in bioinformatics**, Oxford Univ Press, v. 13, n. 6, p. 696–710, 2012.

FUCHS, R. From sequence to biology: the impact on bioinformatics. **Bioinformatics**, Oxford Univ Press, v. 18, n. 4, p. 505–506, 2002.

GALAR, M.; FERNANDEZ, A.; BARRENECHEA, E.; BUSTINCE, H.; HERRERA, F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. **Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on**, IEEE, v. 42, n. 4, p. 463–484, 2012.

GOÉS, F.; ALVES, R.; CORRÊA, L.; CHAPARRO, C.; THOM, L. A comparison of classification methods for gene prediction in metagenomics. In: **Proceedings of the 3rd Workshop on New Frontiers in Mining Complex Patterns**. Nancy: [s.n.], 2014. p. 136–147.

GOÉS, F.; ALVES, R.; CORRÊA, L.; CHAPARRO, C.; THOM, L. Towards an ensemble learning strategy for metagenomic gene prediction. In: **Advances in Bioinformatics and Computational Biology**. [S.l.]: Springer, 2014. p. 17–24.

GREEN, E. D.; GUYER, M. S.; INSTITUTE, N. H. G. R. et al. Charting a course for genomic medicine from base pairs to bedside. **Nature**, Nature Publishing Group, v. 470, n. 7333, p. 204–213, 2011.

HAN, J.; KAMBER, M. **Data mining: concepts and techniques**. [S.l.]: Morgan Kaufmann San Francisco, Calif, USA, 2012.

HAND, D. J.; MANNILA, H.; SMYTH, P. **Principles of data mining**. [S.l.]: MIT press, 2001.

HANDELSMAN, J. Metagenomics: application of genomics to uncultured microorganisms. **Microbiology and molecular biology reviews**, Am Soc Microbiol, v. 68, n. 4, p. 669–685, 2004.

HANDELSMAN, J.; TIEDJE, J.; ALVAREZ-COHEN, L.; ASHBURNER, M.; CANN, I.; DELONG, E.; DOOLITTLE, W. F.; FRASER-LIGGETT, C.; GODZIK, A.; GORDON, J. et al. The new science of metagenomics: Revealing the secrets of our microbial planet. **Nat Res Council Report**, v. 13, 2007.

HOFF, K. J. The effect of sequencing errors on metagenomic gene prediction. **Bmc Genomics**, BioMed Central, v. 10, n. 1, p. 1, 2009.

HOFF, K. J. **Gene prediction in metagenomic sequencing reads**. Tese (Doutorado) — Georg August University Göttingen, 2009.

HOFF, K. J.; TECH, M.; LINGNER, T.; DANIEL, R.; MORGENSTERN, B.; MEINICKE, P. Gene prediction in metagenomic fragments: a large scale machine learning approach. **BMC bioinformatics**, BioMed Central Ltd, v. 9, n. 1, p. 217, 2008.

HYATT, D.; CHEN, G.-L.; LOCASCIO, P. F.; LAND, M. L.; LARIMER, F. W.; HAUSER, L. J. Prodigal: prokaryotic gene recognition and translation initiation site identification. **BMC bioinformatics**, BioMed Central, v. 11, n. 1, p. 1, 2010.

HYATT, D.; LOCASCIO, P. F.; HAUSER, L. J.; UBERBACHER, E. C. Gene and translation initiation site prediction in metagenomic sequences. **Bioinformatics**, Oxford Univ Press, v. 28, n. 17, p. 2223–2230, 2012.

KALISKY, T.; BLAINEY, P.; QUAKE, S. R. Genomic analysis at the single-cell level. **Annual review of genetics**, Howard Hughes Medical Institute, v. 45, 2011.

KUHN, M. The caret package homepage. **URL http://caret. r-forge. r-project. org**, 2010.

KUMAR, S.; KRISHNANI, K. K.; BHUSHAN, B.; BRAHMANE, M. P. Metagenomics: Retrospect and prospects in high throughput age. **Biotechnology research international**, Hindawi Publishing Corporation, v. 2015, 2015.

KUNIN, V.; COPELAND, A.; LAPIDUS, A.; MAVROMATIS, K.; HUGENHOLTZ, P. A bioinformatician's guide to metagenomics. **Microbiology and Molecular Biology Reviews**, Am Soc Microbiol, v. 72, n. 4, p. 557–578, 2008.

LARRAÑAGA, P.; CALVO, B.; SANTANA, R.; BIELZA, C.; GALDIANO, J.; INZA, I.; LOZANO, J. A.; ARMAÑANZAS, R.; SANTAFÉ, G.; PÉREZ, A. et al. Machine learning in bioinformatics. **Briefings in bioinformatics**, Oxford Univ Press, v. 7, n. 1, p. 86–112, 2006.

LIAW, A.; WIENER, M. Classification and regression by randomforest. **R News**, v. 2, n. 3, p. 18–22, 2002. Disponível em: <http://CRAN.R-project.org/doc/Rnews/>.

LIU, Y.; GUO, J.; HU, G.; ZHU, H. Gene prediction in metagenomic fragments based on the svm algorithm. **BMC bioinformatics**, BioMed Central Ltd, v. 14, n. Suppl 5, p. S12, 2013.

MATHÉ, C.; SAGOT, M.-F.; SCHIEX, T.; ROUZE, P. Current methods of gene prediction, their strengths and weaknesses. **Nucleic acids research**, Oxford Univ Press, v. 30, n. 19, p. 4103–4117, 2002.

MENDOZA, M. R.; FONSECA, G. C. da; LOSS-MORAIS, G.; ALVES, R.; MARGIS, R.; BAZZAN, A. Rfmirtarget: predicting human microrna target genes with a random forest classifier. **PloS one**, v. 8, n. 7, p. e70153, 2013.

MITCHELL, T. M. et al. **Machine learning. WCB**. [S.l.]: McGraw-Hill Boston, MA:, 1997.

MOUNT, D. W.; MOUNT, D. W. **Bioinformatics: sequence and genome analysis**. [S.l.]: Cold spring harbor laboratory press New York:, 2001. v. 2.

MÜHLBERGER, I.; WILFLINGSEDER, J.; BERNTHALER, A.; FECHETE, R.; LUKAS, A.; PERCO, P. Computational analysis workflows for omics data interpretation. In: **Bioinformatics for Omics Data**. [S.l.]: Springer, 2011. p. 379–397.

NOGUCHI, H.; PARK, J.; TAKAGI, T. Metagene: prokaryotic gene finding from environmental genome shotgun sequences. **Nucleic acids research**, Oxford Univ Press, v. 34, n. 19, p. 5623–5630, 2006.

NOGUCHI, H.; TANIGUCHI, T.; ITOH, T. Metageneannotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. **DNA research**, Kazusa DNA Resh Ins, v. 15, n. 6, p. 387–396, 2008.

OUZOUNIS, C. A. Rise and demise of bioinformatics? promise and progress. **PLoS Comput Biol**, Public Library of Science, v. 8, n. 4, p. e1002487, 2012.

OUZOUNIS, C. A.; VALENCIA, A. Early bioinformatics: the birth of a discipline—a personal view. **Bioinformatics**, Oxford Univ Press, v. 19, n. 17, p. 2176–2190, 2003.

PANG, H.; LIN, A.; HOLFORD, M.; ENERSON, B. E.; LU, B.; LAWTON, M. P.; FLOYD, E.; ZHAO, H. Pathway analysis using random forests classification and regression. **Bioinformatics**, Oxford Univ Press, v. 22, n. 16, p. 2028–2036, 2006.

POLANSKI, A.; KIMMEL, M. Bioinformatics. **Bioinformatics**, Springer, p. 155–185, 2007.

POLIKAR, R. Ensemble based systems in decision making. **Circuits and systems magazine, IEEE**, IEEE, v. 6, n. 3, p. 21–45, 2006.

PRAKASH, T.; TAYLOR, T. D. Functional assignment of metagenomic data: challenges and applications. **Briefings in bioinformatics**, Oxford Univ Press, v. 13, n. 6, p. 711–727, 2012.

PROSDOCIMI, F.; COUTINHO, G.; NINNECW, E.; SILVA, A. F.; REIS, A. N. dos; MARTINS, A. C.; SANTOS, A. C. F. dos; JÚNIOR, A. N.; FILHO, F. C. Bioinformática: manual do usuário. **Biotecnologia Ciência & Desenvolvimento**, v. 29, p. 12–25, 2002.

PRUITT, K. D.; TATUSOVA, T.; MAGLOTT, D. R. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. **Nucleic acids research**, Oxford Univ Press, v. 35, n. suppl 1, p. D61–D65, 2007.

PYLRO, V. S.; ROESCH, L. F. W.; MORAIS, D. K.; CLARK, I. M.; HIRSCH, P. R.; TÓTOLA, M. R. Data analysis for 16s microbial profiling from different benchtop sequencing platforms. **Journal of microbiological methods**, Elsevier, v. 107, p. 30–37, 2014.

QI, Y. Random forest for bioinformatics. In: **Ensemble machine learning**. [S.l.]: Springer, 2012. p. 307–323.

RHO, M.; TANG, H.; YE, Y. Fraggenescan: predicting genes in short and error-prone reads. **Nucleic acids research**, Oxford Univ Press, p. gkq747, 2010.

RICHTER, D. C.; OTT, F.; AUCH, A. F.; SCHMID, R.; HUSON, D. H. et al. Metasim: a sequencing simulator for genomics and metagenomics. **PloS one**, Public Library of Science, v. 3, n. 10, p. e3373, 2008.

SAEYS, Y.; ROUZÉ, P.; PEER, Y. Van de. In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi and protists. **Bioinformatics**, Oxford Univ Press, v. 23, n. 4, p. 414–420, 2007.

SCHNEIDER, M. V.; ORCHARD, S. Omics technologies, data and bioinformatics principles. In: **Bioinformatics for Omics Data**. [S.l.]: Springer, 2011. p. 3–30.

SEQINR: Biological Sequences Retrieval and Analysis. Disponível em:< https://cran.r-project.org/web/packages/seqinr/index.html >. v. 10, 2003.

SETUBAL, J. C. **A origem e o sentido da bioinformática. Disponível em:< http://www. comciencia. br/reportagens/bioinformatica/bio10. shtml>. Acesso em 15 Dezembro de 2015**, v. 10, 2003.

STROBL, C.; BOULESTEIX, A.-L.; KNEIB, T.; AUGUSTIN, T.; ZEILEIS, A. Conditional variable importance for random forests. **BMC bioinformatics**, BioMed Central Ltd, v. 9, n. 1, p. 307, 2008.

TEELING, H.; GLÖCKNER, F. O. Current opportunities and challenges in microbial metagenome analysis—a bioinformatic perspective. **Briefings in bioinformatics**, Oxford Univ Press, p. bbs039, 2012.

THOMAS, T.; GILBERT, J.; MEYER, F. Metagenomics-a guide from sampling to data analysis. **Microb Inform Exp**, v. 2, n. 3, p. 1–12, 2012.

WANG, W.-L.; XU, S.-Y.; REN, Z.-G.; TAO, L.; JIANG, J.-W.; ZHENG, S.-S. Application of metagenomics in the human gut microbiome. **World journal of gastroenterology: WJG**, Baishideng Publishing Group Inc, v. 21, n. 3, p. 803–814, 2015.

WANG, Z.; CHEN, Y.; LI, Y. et al. A brief review of computational gene prediction methods. **Genomics Proteomics Bioinformatics**, v. 2, n. 4, p. 216–221, 2004.

WITTEN, I. H.; FRANK, E. **Data Mining: Practical machine learning tools and techniques**. [S.l.]: Morgan Kaufmann, 2011.

WOOLEY, J. C.; GODZIK, A.; FRIEDBERG, I. A primer on metagenomics. **PLoS Comput Biol**, v. 6, n. 2, p. e1000667, 2010.

XIONG, J. **Essential bioinformatics**. [S.l.]: Cambridge University Press, 2006.

YANG, P.; YANG, Y. H.; ZHOU, B. B.; ZOMAYA, A. Y. A review of ensemble methods in bioinformatics. **Current Bioinformatics**, Bentham Science Publishers, v. 5, n. 4, p. 296–308, 2010.

YOK, N. G.; ROSEN, G. L. Combining gene prediction methods to improve metagenomic gene annotation. **BMC bioinformatics**, BioMed Central, v. 12, n. 1, p. 1, 2011.

ZHAO, Z.; PASCHKE, A. A survey on semantic scientific workflow. **Semantic Web Journal, IOS press**, p. 1–5, 2012.

ZHU, W.; LOMSADZE, A.; BORODOVSKY, M. Ab initio gene identification in metagenomic sequences. **Nucleic acids research**, Oxford Univ Press, v. 38, n. 12, p. e132–e132, 2010.

APÊNDICES

APÊNDICE A – Orgamismos do cojunto de dados de treino

Organismos	NCBI Reference Sequence - RefSeq
Acidimicrobium ferrooxidans DSM	NC_013124
Acidithiobacillus caldus SM-1	NC_015850
Acidithiobacillus ferrivorans SS3	NC_015942
Acidithiobacillus ferrooxidans ATCC 53993	NC_011206
Acinetobacter sp ADP1	NC_005966
Aeropyrum pernix *	NC_000854
Bdellovibrio bacteriovorus	NC_005363
Bordetella bronchiseptica	NC_002927
Brucella abortus 9-941	NC_006932
Campylobacter jejuni RM1221	NC_003912
Candidatus Blochmannia floridanus	NC_005061
Candidatus Nitrospira defluvii	NC_014355
Dechloromonas aromatica RCB	NC_007298
Dehalococcoides ethenogenes 195	NC_002936
Enterococcus faecalis V583	NC_004668
Francisella tularensis tularensis	NC_006570
Frankia CcI3	NC_007777
Gloeobacter violaceus	NC_005125
Gluconobacter oxydans 621H	NC_006677
Haemophilus ducreyi 35000HP	NC_002940
Hahella chejuensis KCTC 2396	NC_007645
Idiomarina loihiensis L2TR	NC_006512
Lactobacillus acidophilus NCFM	NC_006814
Lactococcus lactis	NC_002662
Mesoplasma florum L1	NC_006055
Mesorhizobium loti	NC_002678
Methanopyrus kandleri *	NC_003551
Photobacterium profundum SS9	NC_006370
Photorhabdus luminescens	NC_005126
Porphyromonas gingivalis W83	NC_002950
Staphylococcus aureus RF122	NC_007622
Streptococcus agalactiae 2603	NC_004116
Streptomyces avermitilis	NC_003155
Thermoanaerobacter tengcongensis	NC_003869
Thermobifida fusca YX	NC_007333
Thermococcus kodakaraensis KOD1*	NC_006624
Thermodesulfovibrio yellowstonii DSM 11347	NC_011296
Thermoplasma acidophilum*	NC_002578
Thermoplasma volcanium GSS1*	NC_002689
Thermus thermophilus HB27	NC_005835

Tabela 9 – O símbolo “*” destaca as *Archaeas*.