UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

Tahila Andrighetti

Ferramenta computacional para identificação de micro-organismos com base em assinaturas genômicas

Brasil

Fevereiro de 2015

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS
CAMPUS DE BOTUCATU

Tahila Andrighetti

FERRAMENTA COMPUTACIONAL PARA IDENTIFICAÇÃO DE MICRO-ORGANISMOS COM BASE EM ASSINATURAS GENÔMICAS

Dissertação de Mestrado apresentada ao Instituto de Biociências, Campus de Botucatu, UNESP, em preenchimento parcial dos requisitos para a obtenção do título de Mestre no Programa de Pós-Graduação em Ciências Biológicas (Genética).

Orientador: Prof. Dr. José Luiz Rybarczyk Filho
Co-Orientador: Prof. Dr. Ney Lemke

Botucatu - SP
2015

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Andrighetti, Tahila.
Ferramenta computacional para identificação de micro-organismos com base em assinaturas genômicas / Tahila Andrighetti. - Botucatu, 2015

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: José Luiz Rybarczyk Filho
Coorientador: Ney Lemke
Capes: 20202008

1. Nucleotídeos. 2. Bioinformática. 3. Entropia. 4. Genoma humano. 5. Seqüenciamento de nucleotídeo.

Palavras-chave: Abundância de dinucleotídeos; Bioinformática; Conteúdo GC; Entropia; Metagenômica.

Agradecimentos

Meus mais sinceros agradecimentos à todos que me ajudaram na elaboração desse trabalho, em especial:

- À Fundação de Amparo à Pesquisa do Estado de São Paulo, pelo apoio financeiro (processo 2013/1517-4);
- Ao Professor Doutor José Luiz Rybarczyk Filho, pela orientação e dedicação;
- Ao Professor Doutor Ney Lemke, pela co-orientação e auxílio;
- Ao Professor Doutor Günther Johannes Lewczuk Gerhardt e à Doutora Agnes Alessandra Sekijima Takeda pelas contribuições e disponibilidade;
- À toda minha família, principalmente meus pais, Eloi e Jaqueline, e minha irmã, Giovana, pelo incentivo, confiança e carinho, que foram fundamentais para o desempenho de meu trabalho;
- Aos meus grandes amigos, principalmente Pedro Rafael Costa, Larissa Cunha, Fernanda Arruda e Rafael Toledo, pela amizade, força e apoio durante a trajetória;
- À equipe do Laboratório de Bioinformática e Biofísica Computacional do Departamento de Física e Biofísica do IBB-Unesp pelo companheirismo, discussões e momentos de descontração.

Resumo

Comunidades microbianas desempenham papéis cruciais em todos ecosistemas da Terra, uma vez que metabolizam compostos essenciais. Essa característica torna importantes alvos de pesquisas em diversas áreas como médica, ambiental, alimentícia e biotecnológica. Entretanto, somente 1% de todas espécies de micro-organismos conhecidos podem ser cultivadas *in vitro*, dificultando o estudo de suas funções e de sua classificação taxonômica. Com o surgimento de novas tecnologias de sequenciamento, o genoma inteiro de micro-organismos de um habitat pode ser experimentalmente extraído, mas em pequenos fragmentos (¡1500 pb), tornando o processamento dos dados um grande desafio. As ferramentas de análise de metagenômica mais utilizadas classificam as sequências por homologia. Entretanto, o tempo computacional aumenta exponencialmente conforme o tamanho dos fragmentos diminuem. Isso mostra uma necessidade evidente de métodos alternativos que possam analisar dados de metagenômica de maneira rápida e precisa. Esse estudo propõe um novo método de identificação de sequências de bactérias que analisa esses dados. Os genomas de 2164 linhagens de bactérias foram obtidos pelo GenBank e fragmentados em grupos de teste e controle. Cada grupo foi aleatóriamente fragmentado em sequências de 64, 128, 256, 512, 1024, 2048 e 4096 pares de base. As medidas de organização de sequências aplicadas nos fragmentos foram: conteúdo GC, abundância de dinucleotídeos e entropias de dipletes, tripletes e tetrapletes. Foram calculados a média e o desvio padrão dos valores das sequências controle para cada espécie, gênero e família de bactéria. Foram feitas combinações de medidas para classificar as sequências em famílias, gêneros e espécies. A performance da metodologia foi determinada por medidas de sensibilidade, especificidade, precição e média harmônica para conjuntos de teste. Os resultados indicaram que o conteúdo GC apresentou a melhor performance dentre as assinaturas investigadas. Dentre as combinações de medidas, a que utiliza conteúdo GC e abundância de dinucleotídeos foi a mais promissora para a classificação de família, gênero e espécie. Por fim, realizamos um estudo de caso usando dados do metagenoma *Pacific Beach Sand*. Comparamos nossos resultados com resultados obtidos usando o *software* MEGAN. Os resultados indicam que há um alto potencial para a metodologia proposta, principalmente para a combinação entre conteúdo GC e abundância de dinucleotídeos.

Abstract

Microbial communities play a crucial role in all ecosystems on Earth since they metabolize essential compounds. Given this relevant role they are investigated in Medicine, Biotechnology, Ecology, Food Sciences among other fields. However, only 1% of all known micro-organisms species can be cultivated *in vitro*. The unravelling of their functions and taxonomic classification demands the development of new approaches. With the advent of new sequencing strategies, the entire genome of microrganisms on a given habitat can be experimentally extracted, but the fragments obtained are small ($<$1500 bps), and the data processing remains a huge challenge. The most used metagenomic analysis tools classify the sequences by homology. However, the computational time grows exponentially as the read length decreases. There is an evident need for alternative methods that can analyze metagenomic data quickly and accurately. This study proposes a new bacteria sequences identification method to be used in metagenomic data. The genomes of 2164 bacterial strains were obtained from the GenBank and distributed into test and control sets. Each group was randomly fragmented into sequences of 64, 128, 256, 512, 1024, 2048, and 4096 base pair. The sequences organization measures applied in the reads were: GC content, dinucleotide abundance and diplets, triplets and tetraplets entropy. The average and standard deviation of the control sequences values of each species, genus and families of bacteria were calculated. Combinations of genomic signatures and entropy were performed allowing classifying bacteria sequences into family, genus and species. The performance of the proposed methodology was determined by measuring sensitivity, specificity, accuracy and harmonic mean for the test set. The results indicated that the GC content presented the best performance among the signatures investigated. We also considered combinations of features, the combination considering GC content and dinucleotide abundance was the most promising for family, genus and species classification. Finally we performed a case study using Pacific Beach Sand metagenome data. We compared our approach with results obtained using MEGAN software. The results indicates a great potential for the proposed methodology specially for the combination of GC content and dinucleotides abundance.

Lista de Figuras

1.1 O gradiente de cores de azul a vermelho representam o intervalo de temperatura de -50 à 150°C, respectivamente. Eucariontes encontram-se distribuídos principalmente a temperaturas intermediárias, enquanto procariontes podem ser encontrados em temperaturas extremas com mais frequência. Modificado de (ROTHSCHILD; MANCINELLI, 2001). . . . . p. 2

1.2 Ilustração das regiões conservadas (representadas em verde) e variáveis (representadas em cinza) do gene 16S rRNA. Modificado de (ALIMETRICS, 2014). . . . . p. 3

1.3 Árvore filogenética demonstrando as classificações dos domínios Bacteria e Archaea baseadas em comparações do gene 16S rRNA. Os domínios Bacteria, Archaea e Eukaryontes são representados em azul, rosa e amarelo, respectivamente. Figura retirada do link http://simple.wikipedia.org/wiki/Evolution. . . . . p. 4

1.4 Etapas de realização da metagenômica. Amostras dos ambientes a serem estudados são obtidas; O material genético dos organismos de interesse é extraído e este é sequenciado. As informações adquiridas são armazenadas e ficam no aguardo de análise. . . . . p. 7

1.5 O DNA obtido do ambiente pode ser analisado de duas formas: i) a partir da informação de somente um gene, sendo o mais utilizado o 16S rRNA, ou ii) a partir da informação de todo o DNA sequenciado, ou metagenoma. . . . . p. 9

3.1 Esquema da distribuição das sequências nos grupos teste e controle. De cada genoma obtido, são selecionadas aleatoriamente sequências de teste e controle com 70 mil pares de base cada. De dentro de cada uma destas, foram selecionadas aleatoriamente 100 sequências para cada um dos tamanhos W (64, 128, 256, 512, 1024, 2048 e 4096 pb). A seleção aleatória das sequências foi feita a partir de um algoritmo desenvolvido em Python que também impede que as sequências selecionadas se sobreponham. . . . . p. 16

3.2 Ilustração do cálculo de conteúdo GC. É contada a quantidade de guanina ($\#_G$) e de citosina ($\#_C$) presentes na sequência. A soma desses valores é dividido pela soma do total da quantidade de todos nucleotídeos na sequência, para calcular o valor de conteúdo GC. . . . . p. 16

3.3 Ilustração do cálculo de abundância de dinucleotideos. É calculada a abundância relativa de cada dinucleotídeo (ρ_{XY}) da sequência utilizando o valor de sua frequência (f_{XY}) e da frequência de cada nucleotídeo (f_X e f_Y) no segmento de DNA. Podemos observar que as frequências de CG e de GC (f_{CG} e f_{GC}) são diferentes, assim como sua abundância relativa (ρ_{CG} e ρ_{GC}). A partir da abundância relativa de todos dinucleotídeos, é feita a média para a obtenção do valor de abundância de dinucleotídeos da sequência completa (Equação 3.3). . . . . p. 17

3.4 Ilustração das janelas aplicadas para as entropias de dipletes, tripletes e tetrapletes. . p. 18

3.5 Ilustração da formação dos grupos de teste. Cada grupo de teste foi formado por 100 sequências positivas (de um mesmo gênero ou família) e 100 sequências negativas (de organismos diferentes daquele dos exemplos positivos). Um dos exemplos ilustrados na figura é a família Rhodothermaceae. Dentre os genomas presentes no banco de dados, 3 deles eram de espécies pertencentes à essa família. Uma vez que para cada genoma há 100 sequências de cada tamanho W, a família Rhodothermaceae possui 300 sequências, a partir das quais os exemplos positivos foram escolhidos aleatoriamente por um algoritmo em Python. Foram gerados 100 grupos de teste para cada família e gênero. . . . . p. 19

3.6 Esquema explicando o nome das combinações utilizadas nas análises das sequências. O primeiro dígito representa quantas medidas estão sendo utilizadas na combinação (2 a 5), e os dois últimos dígitos representa o número de identificação daquela combinação. . . . . p. 20

3.7 Ilustração da utilização das combinações para a seleção de espécies pertencentes àquela sequência. A figura mostra como exemplo a COMB303, que é composta por conteúdo GC, abundância de dinucleotídeos e entropia de tripletes.Um fragmento de DNA é testado para as medidas que pertencem àquela combinação. Os valores de cada medida são comparados, e quando não correspondem às espécies, estas são descartadas. A) Os valores para abundância de dinucleotídeos não correspondem com a Sp 5, somente com as Sp 1, 2, 3 e 4. B) Os valores para conteúdo GC não correspondem com a Sp 2, somente com as Sp 1, 3 e 4. C) Os valores para entropia não correspondem com a Sp 3, somente com as Sp 1 e 4. A sequência seria classificada como Sp 1 ou Sp 4. . . . . p. 21

3.8 Esquema da identificação de uma sequência obtida por metagenômica ($\star$) a partir de duas medidas. A média ($\bullet$) e desvio padrão ($-$) dos valores para cada espécie estão apresentadas no gráfico. Os valores das medidas da sequência incógnita são 0.6 e 0.13, sendo assim, esse fragmento estaria classificado como a Espécie C, pois encontra-se no intervalo de valores correspondentes a ela. . . . . p. 21

3.9 Medidas estatísticas para a avaliação da metodologia. São calculadas a sensibilidade (a), especificidade (b), precisão (c) e média harmônica entre sensibilidade e precisão (d). Calcula-se a sensibilidade a partir dos valores de verdadeiro positivo e falso negativo, a especificidade a partir dos falsos positivos e verdadeiros negativos, a precisão a partir dos verdadeiros positivos e verdadeiros negativos e a média harmônica entre sensibilidade e precisão a partir dos valores de verdadeiros positivos, falsos negativos e verdadeiros negativos. . . . . p. 23

3.10 Esquema ilustrando o critério de seleção para as sequências corretamente identificadas, erroneamente identificadas ou não identificadas do metagenoma *Pacific Beach Sand* a partir das combinações de medidas, utilizando a classificação do *software* MEGAN como referência. . . . . p. 25

4.1 Gráfico de sensibilidade, especificidade, precisão e média harmônica para análises de espécie utilizando somente conteúdo GC. . . . . p. 27

4.2 Gráfico de sensibilidade, especificidade, precisão e média harmônica para a comparação das sequências teste com os valores de controle utilizando abundância de dinucleotídeos. . . . . p. 28

4.3 Gráfico de sensibilidade, especificidade, precisão e média harmônica para análises de espécie utilizando entropias. A entropia de dipletes está representada em verde, a de tripletes em azul e a de tetrapletes em vermelho. . . . . p. 29

4.4 Valores de sensibilidade e especificidade para classificação taxonômica de fragmentos de 1 kpb das ferramentas RAIphy, TACOA e PhyloPythia. Figura publicada em (NALBANTOGLU et al., 2011). . . . . p. 30

4.5 Gráfico ilustrando a quantidade de famílias que cada combinação identifica acima de 60% para todas as medidas estatísticas. . . . . p. 31

4.6 Diagrama ilustrando a quantidade de famílias identificadas (acima de 60% para todas as medidas estatísticas) em comum entre as combinações COMB201, COMB204 e COMB207. . . . . p. 31

4.7 Gráfico de sensibilidade, especificidade, precisão e média harmônica para análises de famílias utilizando COMB201. . . . . p. 33

4.8 Gráfico de sensibilidade, especificidade, precisão e média harmônica para análises de famílias utilizando COMB204. . . . . p. 33

4.9 Gráfico ilustrando a quantidade de gêneros que cada combinação identifica acima de 60% para todas as medidas estatísticas. . . . . p. 34

4.10 Diagrama ilustrando a quantidade de gêneros identificados em comum entre as combinações COMB201, COMB204 e COMB207. . . . . p. 35

4.11 Gráfico de sensibilidade, especificidade, precisão e média harmônica para análises de gêneros utilizando COMB201. O gênero *Vibrio* está representado em azul, o *Salmonella* em verde e o *Escherichia* em vermelho. . . . . p. 36

4.12 Gráfico de sensibilidade, especificidade, precisão e média harmônica para análises de gêneros utilizando COMB204. O gênero *Vibrio* está representado em azul, o *Salmonella* em verde e o *Escherichia* em vermelho. . . . . p. 37

4.13 Gráfico ilustrando a quantidade de espécies que cada combinação identifica acima de 60% para todas as medidas estatísticas. . . . . p. 38

4.14 Gráfico de sensibilidade, especificidade, precisão e média harmônica para COMB204 para espécies, combinação de conteúdo GC e entropia de tetrapletes. p. 39

4.15 Gráfico de sensibilidade, especificidade, precisão e média harmônica para espécies isoladas identificadas por COMB204, combinação de conteúdo GC com entropia de tetrapletes. *Salmonella enterica* está representada em azul, *Escherichia coli* em verde e *Pseudomonas aeruginosa* em vermelho. . . . . p. 40

4.16 Análise do metagenoma *Pacific Beach Sand* pelo *software* MEGAN. . . . . p. 41

4.17 Quantidade de famílias reconhecidas para cada sequência por cada grupo de combinações. Foram selecionadas pelas combinações dentro de um grupo de 208 famílias. . . . . p. 42

4.18 Gráficos de pizza apresentando a porcentagem de sequências classificadas corretamente, classificadas erroneamente, não classificadas em comum com o MEGAN e não classificadas somente pelas combinações para os 4 grupos de combinações, num total de 325 sequências. . . . . p. 43

4.19 Valores de sensibilidade e especificidade para classificação taxonômica de fragmentos de 1 kpb das ferramentas RAIphy, TACOA e PhyloPythia em comparação com os valores da COMB201. Modificado de (NALBANTOGLU et al., 2011). . . . . p. 44

Lista de Tabelas

1.1 Tabela de caracterização dos filos pertencentes ao domínio Bacteria de acordo com (MADIGAN et al., 2010). . . . . p. 4

1.2 Tabela de caracterização dos filos pertencentes ao domínio Bacteria de acordo com (MADIGAN et al., 2010). . . . . p. 5

1.3 Lista de plataformas de sequenciamento de nova geração, o tamanho de seus *reads* e seu custo por GB (MOROZOVA; MARRA, 2008; NEXT GEN SEEK, 2012; THOMAS; GILBERT; MEYER, 2012). . . . . p. 8

1.4 Tabela de ferramentas de bioinformática para análise de metagenômica (MANDE; MOHAMMED; GHOSH, 2012; THOMAS; GILBERT; MEYER, 2012; KIM et al., 2013; LIU et al., 2013). . . . . p. 10

3.1 Tabela de nomenclatura para as combinações de medidas usadas para as análises. . . p. 20

3.2 Tabela esquematizando os critérios de identificação de famílias, gêneros e espécies. . . . . p. 24

4.1 Tabela de famílias em comum para as combinações mostradas na Figura 4.6. . . . . p. 32

4.2 Tabela de gêneros em comum para as combinações mostradas na Figura 4.10. . . . . p. 35

Sumário

1 Introdução p. 1

1.1 Micro-organismos . . . . . p. 1

1.1.1 Comunidades microbianas . . . . . p. 1

1.1.2 Sistemática microbiana . . . . . p. 2

1.2 Metagenômica . . . . . p. 6

1.3 Análise dos dados de metagenômica . . . . . p. 9

1.3.1 Assinaturas genômicas e organização de sequências . . . . . p. 12

1.4 Proposta . . . . . p. 12

2 Objetivo p. 14

2.1 Objetivos específicos . . . . . p. 14

3 Material e Métodos p. 15

3.1 Banco de dados . . . . . p. 15

3.2 Medidas de padrões em sequências . . . . . p. 15

3.2.1 Conteúdo GC (GC%) . . . . . p. 15

3.2.2 Abundância de Dinucleotídeos (din) . . . . . p. 16

3.2.3 N-entropia (S) . . . . . p. 18

3.3 Valores de referência . . . . . p. 18

3.4 Análise estatística . . . . . p. 19

3.4.1 Grupos de teste . . . . . p. 19

3.4.2 Combinações de medidas . . . . . p. 20

3.4.3 Matriz de confusão . . . . . p. 22
3.5 Prova de conceito . . . . . p. 24

4 Resultados e Discussão p. 26
4.1 Avaliação das medidas de padrão de sequência . . . . . p. 26
4.1.1 Conteúdo GC . . . . . p. 26
4.1.2 Abundância de Dinucleotídeos . . . . . p. 27
4.1.3 N-entropias . . . . . p. 28
4.2 Análise de desempenho para famílias . . . . . p. 30
4.2.1 Análise estatística . . . . . p. 32
4.3 Análise de desempenho para gêneros . . . . . p. 34
4.3.1 Análise estatística . . . . . p. 36
4.4 Análise de desempenho para espécies . . . . . p. 38
4.4.1 Análise estatística . . . . . p. 39
4.5 Prova de conceito . . . . . p. 41
4.5.1 Análise pelo MEGAN . . . . . p. 41
4.5.2 Análise pelas combinações de medidas . . . . . p. 42
4.6 Comparação com outras metodologias . . . . . p. 44

5 Conclusões p. 46

Referências Bibliográficas p. 48

1 Introdução

1.1 Micro-organismos

1.1.1 Comunidades microbianas

Os micro-organismos são essenciais para a vida em todos ecossistemas existentes na Terra, pois tornam disponíveis e reciclam elementos como enxofre, carbono, nitrogênio e oxigênio para outros organismos que não os metabolizam (COUNCIL, 2007). Alguns tem habilidade de adaptar-se a condições extremas, como de temperatura, pH, radiação, pressão e salinidade, onde organismos eucariotos não manteriam sua estabilidade molecular (ROTHSCHILD; MANCINELLI, 2001; COUNCIL, 2007; MADIGAN et al., 2010). Além da sua importância ecológica, os procariotos tem sido utilizados para o desenvolvimento de novas tecnologias em diversas áreas, como ambiental, agrícola, médica, farmacêutica e alimentícia, bem como lançando luz sobre questões como a origem da vida e a possibilidade de vida em outros planetas (COUNCIL, 2007; CANGANELLA; WIEGEL, 2011).

Um grupo de micro-organismos de uma mesma espécie que originaram de uma célula em comum e convivem no mesmo ambiente formam uma população. Diversas populações compõem uma comunidade microbiana, onde cada micro-organimo desempenha seu papel em associação com os de outras espécies. Os organismos que constituem as comunidades são aqueles que estão melhores adaptados às condições físicas e químicas do hábitat em que vivem, sendo que essas características são determinadas pelos micro-organismos que integram aquele ambiente. Por exemplo, sub-produtos orgânicos gerados a partir do metabolismo de bactérias de uma espécie podem ser utilizados para a nutrição de uma outra que não o produz (COUNCIL, 2007; MADIGAN et al., 2010). As comunidades bacterianas, juntamente com um complexo dinâmico de plantas e animais, formam um ecossistema. Mesmo em ambientes inóspitos (como em fumarolas, gelo, ambientes ácidos, contaminados ou com pouca água, dentre outros), onde outras formas de vida não sobreviveriam, os micro-organismos estão presentes formando ecossistemas quase ou exclusivamente microbianos (MADIGAN et al., 2010). Um

exemplo dessa característica é mostrado na Figura 1.1, que ilustra uma escala de temperatura e diversos organismos que podem adaptar-se a elas. Podemos observar que os procariotos são os mais adaptados às temperaturas mais extremas, quando comparados aos organismos eucariotos, que são encontrados com mais frequência em temperaturas mais amenas (ROTHSCHILD; MANCINELLI, 2001).

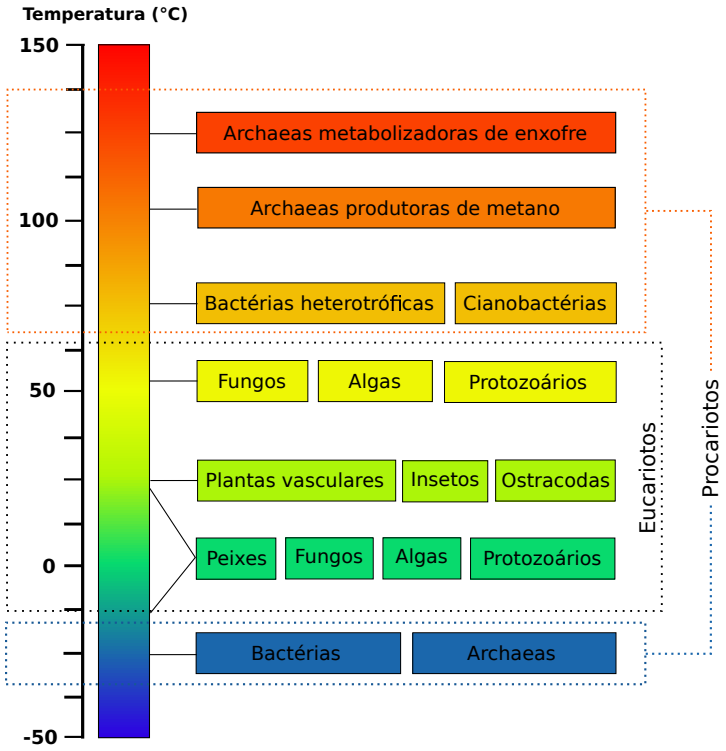


Figura 1.1: O gradiente de cores de azul a vermelho representam o intervalo de temperatura de -50 à 150°C, respectivamente. Eucariontes encontram-se distribuídos principalmente a temperaturas intermediárias, enquanto procariontes podem ser encontrados em temperaturas extremas com mais frequência. Modificado de (ROTHSCHILD; MANCINELLI, 2001).

1.1.2 Sistemática microbiana

Sistemática é o estudo da diversidade de organismos, sua classificação biológica e suas relações, fornecendo ferramentas para o estudo de aspectos históricos da evolução (SCHUH, 2000; GOODFELLOW, 2000; MADIGAN et al., 2010). Nessa ciência, os organismos são primeiramente classificados em espécies de acordo com diferentes características, podendo ser elas morfológicas, fisiológicas, químicas, filogenéticas e genotípicas. Estas são posteriormente comparadas e categorizadas em grupos hierárquicos superiores de acordo com relações evolutivas (SCHUH, 2000; AMORIM, 2002). Taxonomia é uma disciplina inclusa na sistemática. Implica na classificação de táxons de acordo com suas relações filogenéticas (AMORIM, 2002; MADIGAN et al., 2010).

A taxonomia bacteriana era tradicionalmente fundamentada em comparações fenotípicas, como características morfológicas, metabólicas, químicas e fisiológicas da célula. Com o aumento da informação genética, traços filogenéticos e genômicos também passaram a ser considerados na classificação taxonômica dos organismos. As análises filogenéticas agrupam os micro-organismos em um quadro evolucionário, e as genotípicas avaliam a comparação entre traços genotípicos das sequências de cada bactéria (MADIGAN et al., 2010).

O conceito biológico de espécies as definem como "grupos de organismos de uma população que são reprodutivamente isolados", entretanto essa definição é válida somente para organismos que se reproduzem sexualmente. Uma conotação mais aplicável aos procariotos seria: "grupo de linhagens que dividem as mesmas características principais e diferem em uma ou mais propriedades significantes de outros grupos de linhagens; definidas filogenéticamente como monofiléticas, grupo exclusivo baseado na sequência do DNA". Espécies similares são agrupadas em gêneros, gêneros em famílias, seguido de ordem, classe, filo e domínio (MADIGAN et al., 2010).

A técnica de identificação genotípica mais utilizada é a comparação da sequência do gene 16S rRNA da bactéria de interesse com as espécies já identificadas (CLARRIDGE, 2004). Esse gene possui regiões hiperconservadas e variáveis intercaladas ao longo de sua sequência (Figura 1.2). As regiões conservadas são quase idênticas dentre os organismos, enquanto as outras variam proporcionalmente à proximidade filogenética entre as espécies. O alto índice de conservação de algumas regiões permitiu o desenvolvimento de *primers* universais que possibilitam o isolamento dos genes da amostra. As regiões mutáveis são usadas como parâmetros de identificação dos organismos. Os genes são isolados, amplificados a partir de PCR, e depois sequenciados (NIKOLAKI; TSIAMIS, 2013). Para dois micro-organismos serem considerados da mesma espécie, deve haver mais de 99% de identidade entre as sequências de 16S rRNA. Entre gêneros, a identidade deve estar entre 97 e 99%. Para a classificação de uma espécie como sendo potencialmente nova, o nível de identidade deve ser menor do que 97% comparado às espécies já existentes (DRANCOURT; BERGER; RAOULT, 2004).

Figura 1.2: Ilustração das regiões conservadas (representadas em verde) e variáveis (representadas em cinza) do gene 16S rRNA. Modificado de (ALIMETRICS, 2014).

Essa metodologia tem sido usada como a principal para a classificação dos organismos.

Baseado em comparações do 16S rRNA, os seres-vivos foram divididos entre três domínios: Eukarya, Archaea e Bacteria (Figura 1.3), sendo que os dois últimos são compostos por micro-organismos procariotos (WOESE; KANDLER; WHEELIS, 1990).

O domínio Bacteria é dividido em filos. Alguns dos principais filos são conhecidos pelo estudo das bactérias em culturas, enquanto outros foram identificados pelo 16S rRNA de comunidades em seus habitats naturais. (Tabela 1.1 e Tabela 1.2) (MADIGAN et al., 2010).

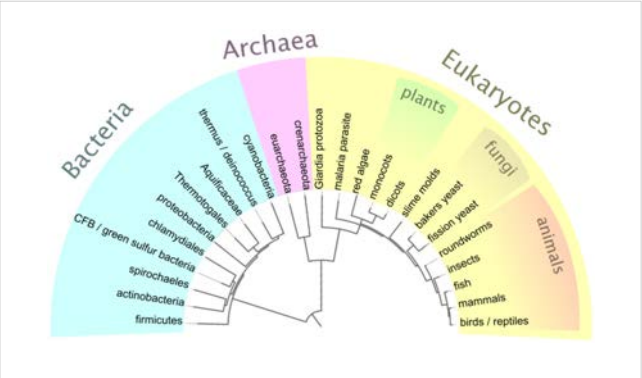


Figura 1.3: Árvore filogenética demonstrando as classificações dos domínios Bacteria e Archaea baseadas em comparações do gene 16S rRNA. Os domínios Bacteria, Archaea e Eukaryontes são representados em azul, rosa e amarelo, respectivamente. Figura retirada do link http://simple.wikipedia.org/wiki/Evolution.

Tabela 1.1: Tabela de caracterização dos filos pertencentes ao domínio Bacteria de acordo com (MADIGAN et al., 2010).

Filo	**Principais Características**
Acidobacteria	Acidofílicas; Comuns em solos ácidos e outros habitats; Gram-negativos;
Actinobacteria	Gram-positivas; Formato de bastões a filamentos; Primariamente aeróbicas; Habitam solos e plantas.
Bacterias verdes sulfurosas	Fototróficas utilizadoras de sulfido de hidrogênio (H_2S); Vivem em ambientes aquáticos anóxicos e sulfídicos (como fumarolas); Podem ter formato de bastão, esfera e oval; Termofílicas.
Chlamydia	Gram-negativas; Causadoras de doenças em humanos e outros animais.
Cyanobacteria	Filo morfologicamente e ecologicamente diversificado; Fototróficas e produtoras de oxigênio, sendo os primeiros organismos com esse tipo de metabolismo na Terra; Responsáveis pela mais alta conversão de gás carbônico em oxigênio que ocorre em nosso planeta; Parede celular é similar à das gram-negativas.

Tabela 1.2: Tabela de caracterização dos filos pertencentes ao domínio Bacteria de acordo com (MADIGAN et al., 2010).

Filo	**Principais Características**
Cytophaga	Gram-negativas; Terrestres e aquáticas; Formato de esferas longas e finas;
Firmicutes	Grande grupo das bactérias gram-positivas; Bactérias de ácido lático; Formadoras de esporos: • Forma de bastão (exceto o gênero *Sporosacina*); • Encontradas no solo; • Não formadoras de esporos: • Forma de bastão e coccus.
Flavobacteria	Aeróbias obrigatórias a anaeróbias obrigatórias; Podem ser encontradas em ambientes aquáticos, plantas, trato intestinal de humanos e outros animais, etc.; Podem ser patógenos associados com infecções intestinais.
Mollicutes	Gram-positivas e sem parede celular; Alguns dos menores organismos conhecidos; Possuem esteróis e lopoglicanos que auxiliam na resistência e estabilidade da célula; Comumente habitam solo e plantas.
Planctomycetes	Morfologicamente únicas (diferente compartimentarização da célula); Primariamente aquáticas; Parede rica em cisteína e prolina;
Prochlorophytas	Fototróficas produtoras de oxigênio; Algumas são simbiontes de invertebrados marinhos, sendo essas de formato esférico; Outras habitam lagos de água fresca. Filamentosas.
Proteobacteria	Gram-negativas; Maior filo de bactérias; Grande diversidade morfológica, fisiológica e metabólica; Vivem em habitats diversos; Divididas em seis classes: Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria, Deltaproteobacteria, Epsilon- proteobacteria e Zetaproteobacteria.
Spirochetes	Sua morfologia é preditora de filogenia; Gram-negativas; Formato de e-spiral; Encontrados em ambiente aquático e em animais; Algumas causam doenças, incluindo sífilis; Possui flagelo para sua motilidade.
Verrumicriobia	Podem formar apêndices citoplasmáticos chamados *prosthecae*; Aeróbicas obrigatórias ou facultativas; Habitam tanto ambientes aquáticos marinhos e de água doce quanto solos florestais e agrícolas

1.2 Metagenômica

Metagenômica é o campo da genômica que estuda os genomas de comunidades microbianas presentes em um determinado habitat, baseado no DNA extraído de um ambiente sem a necessidade de cultivo dos micro-organismos (MARCO, 2011). Esse tipo de análise permitiu a superação de um dos principais obstáculos que pesquisadores tem encontrado no estudo de comunidades microbianas: mais de 99% dos micro-organismos não podem ser cultivados (HANDELSMAN, 2004).

A partir de então, a caracterização de diversos ambientes tem sido realizada. Naviaux e colaboradores coletaram amostras de areia de 14 praias ao longo do oceano pacífico e, dentre os mais de 3 mil genes encontrados, 99% eram novos (NAVIAUX et al., 2005). Alguns estudos realizados em diferentes ambientes apontam que não somente bactérias, como inferia-se anteriormente, mas também archaeas são responsáveis pela oxidação de amônia. A partir de pesquisas de metagenômica encontrou-se o gene de proteínas amônia mono-oxigenases perto de genes marcadores de archaeas, e concluiu-se que elas são as principais oxidadoras de amônia em ambientes terrestres e aquáticos (NICOL; SCHLEPER, 2006; BEMAN; POPP; FRANCIS, 2008; MOSIER; FRANCIS, 2011). Análises evolutivas foram realizadas a partir de dados obtidos em amostras metagenômicas de ambientes marinhos a partir de proteínas quinases (KANNAN et al., 2007). Análises de metagenoma do solo de florestas tropicais revelaram espécies de bactérias lignocelulósicas promissoras, que podem vir a ser utilizadas para a produção de biocombustível (WOO et al., 2014).

Além de estudos ambientais, a metagenômica tem mostrado potencial também em áreas humanas. Um exemplo é o Projeto Microbioma Humano (*Human Microbiome Project*), financiado pelo NIH (*National Institutes of Health*). Esse projeto tem como objetivo sequenciar o metagenoma de partes do corpo como cavidade gastro-intestinal, olhos, pele, vias aéreas, trato urogenital e sangue, para esclarecer o papel do microbioma na saúde e desenvolver novas ferramentas e dados que possam ser utilizados posteriormente em prol de outras pesquisas (PETTERSSON; LUNDEBERG; AHMADIAN, 2009). Além desse projeto, outros relacionados à saúde tem sido desenvolvidos como o de Ridaura e colaboradores, que mostraram a influência da microbiota intestinal na obesidade (RIDAURA et al., 2013). Outros estudos demonstraram que a importância da composição microbiológica do intestino interfere no rendimento de energia, aquisição de nutrientes e outras vias metabólicas do organismo, estando desequilibrada pode induzir o desenvolvimento de doenças como diabetes tipo 2 e obesidade (DEVARAJ; HEMARAJATA; VERSALOVIC, 2013). Também foi observada a importância da microbiota da pele, que auxilia na proteção e imunidade do corpo (CHEN; TSAO, 2013). Em vista disso, o en-

tendimento da microbiota do corpo humano é fundamental para o desenvolvimento de fármacos e novos métodos de diagnósticos e prevenção de doenças.

Esses exemplos mostram a habilidade que a metagenômica tem de lançar luz sobre muitas questões ainda desconhecidas pela comunidade científica. Questões não somente de conhecimentos de base, mas também conhecimentos aplicáveis a diversos campos, potencializando a exploração de novas proteínas, genes ou micro-organismos de importância biotecnológica, médica e ambiental.

Para a execução da metagenômica, primeiramente o DNA é extraído diretamente de todos micro-organismos de um determinado ambiente e posteriormente sequenciado (COUNCIL, 2007) (Figura 1.4).

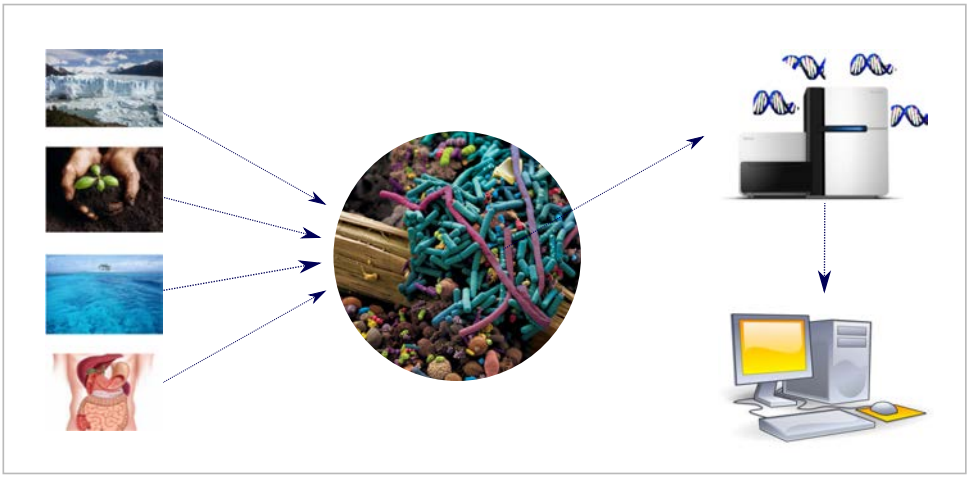


Figura 1.4: Etapas de realização da metagenômica. Amostras dos ambientes a serem estudados são obtidas; O material genético dos organismos de interesse é extraído e este é sequenciado. As informações adquiridas são armazenadas e ficam no aguardo de análise.

Embora nos últimos 10 anos as plataformas de sequenciamento de nova geração (SNG) tenham permitido o sequenciamento de genomas inteiros com muito mais rapidez e custos menores do que a técnica de sequenciamento de Sanger, esta ainda é bastante utilizada por causa do seu baixo índice de erros e do tamanho de seus fragmentos (*reads*), que são maiores do que 700 pb, facilitando as análises dos metagenomas. Entretanto, o custo por gigabase chega a U$ 400 mil, e sua execução é mais complexa do que a das tecnologias de nova geração. (SHENDURE; JI, 2008; THOMAS; GILBERT; MEYER, 2012; NIKOLAKI; TSIAMIS, 2013).

Por conseguinte, o sequenciamento de Sanger tem sido gradativamente substituído por plataformas de SNG, cujo princípio consiste no sequenciamento massivo de moléculas de DNA em uma plataforma. Dentre as mais utilizadas para a metagenômica estão tecnologias desenvolvidas por empresas como Life Technologies (SOLiD), Roche (pirosequenciamento 454) e Illumina (MiSeq e HiSeq). A tecnologia da Roche produz os *reads* de maior tamanho dentre

as SNG, sendo eles de tamanho 400 a 700 pb, embora a maiores custos (U$ 20 mil) e menor rendimento (80 a 120Mb) do que as outras. As plataformas da Illumina realizam os sequenciamentos de menor custo dentre as SNG, sendo de U$ 50 por gigabyte, com *reads* de 100 a 150 pb e corridas com rendimento de 1 Gb. A SOLiD, da Life Technologies produz a maior quantidade de dados de sequenciamento, podendo chegar a 3Gb, entretanto lê os menores tamanhos de fragmentos, de 35 a 75 pb, por U$ 130 por gigabyte. (SCHOLZ; LO; CHAIN, 2012; NIKOLAKI; TSIAMIS, 2013) (Tabela 1.3).

Tabela 1.3: Lista de plataformas de sequenciamento de nova geração, o tamanho de seus *reads* e seu custo por GB (MOROZOVA; MARRA, 2008; NEXT GEN SEEK, 2012; THOMAS; GILBERT; MEYER, 2012).

Tecnologia de sequenciamento	**Tamanho dos *reads***	**Custo por GB**	**Rendimento por corrida**
Sanger	$>$ 700 pb	U$ 400 000	96 Kb
454 / Roche	400 - 700 pb	U$ 20 000	80-120 Mb
Illumina	100 - 150 pb	U$ 50	1 Gb
Life Technologies / SOLiD	35 - 75 pb	U$ 130	1-3 Gb

Depois de sequenciados, os dados brutos obtidos devem passar pela etapa de "controle de qualidade", pois podem conter material de baixa qualidade e sem informação como sequências artefatos, duplicadas, contaminantes e até os próprios adaptadores utilizados no sequenciamento que devem ser removidos para não interferir nos resultados. A metodologia utilizada dependerá do tipo de plataforma utilizada (COUNCIL, 2007; HUNTER et al., 2014).

A análise do metagenoma podem ser feita a partir da sequência de nucleotídeos ou pela determinação das proteínas expressadas a partir daqueles genes (análise funcional). O estudo da sequência nucleotídica fornece conhecimento sobre a composição taxonômica e as interações entre os membros das comunidades microbianas, relações filogenéticas e suas identidades. Essa classificação pode ser feita tanto diretamente em seus fragmentos quanto a partir do genoma montado. As análises funcionais oferecem dados sobre diferentes papéis das proteínas que podem ser encontradas no meio, como produção de vitaminas ou resistência a antibióticos. (Figura 1.5) (COUNCIL, 2007; RAES; FOERSTNER; BORK, 2007).

Embora os custos do sequenciamento de genomas inteiros tenham sido consideravelmente reduzidos, a maioria dos dados obtidos é caracterizada a partir de genes marcadores como o 16S rRNA (NIKOLAKI; TSIAMIS, 2013). Entretanto, essa metodologia é limitada pelo pequeno tamanho dos fragmentos obtidos pelos sequenciamentos de nova geração, dificuldades de acesso às unidades taxonômicas operacionais e baixa resolução do gene dentre espécies próximas, o que impede a classificação taxonômica de chegar a níveis tão restritos. Quando utilizado todo o material genético para a análise das comunidades microbianas, é possível estudar o ambiente de

maneira mais robusta, sem a necessidade de clonar e amplificar genes específicos (PORETSKY et al., 2014). Quão correto é considerar "metagenômica pura" uma metodologia que utiliza somente um dos genes e despreza o resto do material genético? Esposito e colaboradores (ESPOSITO; KIRSCHBERG, 2014) discutem essa questão em seu artigo, e sugerem que as metodologias que utilizam apenas um gene sejam chamadas de "metagenética" pelo fato de que a palavra "metagenômica" remeta à análise do metagenoma como um todo (ESPOSITO; KIRSCHBERG, 2014). Embora o gene 16S rRNA ainda seja o mais utilizado para a caracterização taxonômica dos ambientes, a metagenômica vem substituindo o uso dessa metodologia.

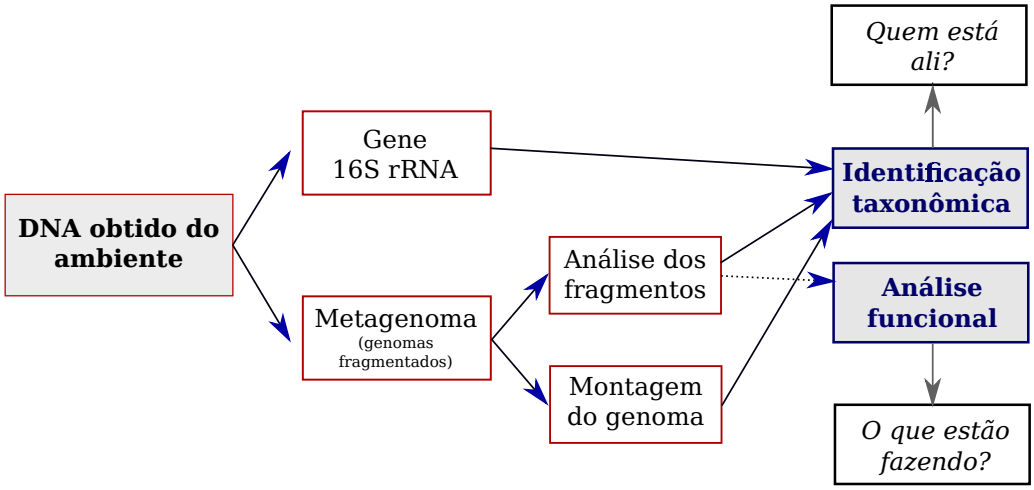


Figura 1.5: O DNA obtido do ambiente pode ser analisado de duas formas: i) a partir da informação de somente um gene, sendo o mais utilizado o 16S rRNA, ou ii) a partir da informação de todo o DNA sequenciado, ou metagenoma.

1.3 Análise dos dados de metagenômica

O progresso das tecnologias de sequenciamento promoveram aquisição de grande quantidade de dados, tanto de genomas únicos quanto de dados de metagenômica. Consequentemente, há a necessidade de armazenamento dessas informações em bases de dados, bem como desenvolvimento de ferramentas para análise. Essa etapa tem sido considerada a mais desafiadora para os pesquisadores, pois utilizam ferramentas computacionais pouco acessíveis, insuficientemente precisas e que demandam muito tempo de análise (RAES; FOERSTNER; BORK, 2007; HUNTER et al., 2011; SCHOLZ; LO; CHAIN, 2012). Além disso, as plataformas de sequenciamento retornam sequências fragmentadas, portanto um metagenoma possui frações de DNA de diversos organismos. Esses fragmentos variam de 700 a 400 pares de base, se obtidos pela plataforma de sequenciamento Roche 454, entre 100 e 150 pb pelas tecnologias da Illumina e entre 50 e 75 pb se retornados pela SOLiD (SCHOLZ; LO; CHAIN, 2012).

A montagem do genoma é muitas vezes requerida para análises funcionais para que se possa observar sua estrutura, e para a utilização da maioria das ferramentas de bioinformática. Entretanto, devido ao tamanho das segmentos de DNA obtidos por sequenciamento de nova geração (KIM et al., 2013), o custo computacional é muito alto, é necessária muita informação das sequências para que a montagem seja precisa (HUTCHISONIII, 2007; RAES; FOERSTNER; BORK, 2007; SCHOLZ; LO; CHAIN, 2012; LIU et al., 2013). Portanto, caracterizar taxonomicamente os micro-organismos de um meio a partir dos fragmentos separados passa a ser uma alternativa para decifrar a estrutura de uma comunidade microbiótica (MANDE; MOHAMMED; GHOSH, 2012) de forma mais eficiente, levando-se em consideração o custo financeiro e computacional (DESAI et al., 2012). Isso demonstra a importância de ferramentas que identifiquem fragmentos com tamanhos a partir de 50 pb, o menor tamanho de *reads* obtido pelas plataformas de sequenciamento utilizadas pela metagenômica (THOMAS; GILBERT; MEYER, 2012).

Para identificar fragmentos de DNA de metagenômica sem a necessidade de montar o genoma, existem alguns *softwares* disponíveis, que são classificados de acordo com a metodologia de análise que utilizam (similaridade ou composição de sequência). As ferramentas de similaridade utilizam primeiramente algoritmos de alinhamento como BLAST (*Basic Local Alignment Search Tool*) para alinhar os fragmentos (*reads*) com os genomas disponíveis no banco de dados. Posteriormente, o algoritmo realiza análises dos dados obtidos e avalia as sequências de acordo com similaridade para inferir a taxonomia do fragmento não identificado (THOMAS; GILBERT; MEYER, 2012; MANDE; MOHAMMED; GHOSH, 2012). As ferramentas de análise filogenética e taxonômica mais utilizadas são MEGAN, SmashCommunity e Kraken (Tabela 1.4) (HUSON et al., 2007; ARUMUGAM et al., 2010; WOOD; SALZBERG, 2014).

Tabela 1.4: Tabela de ferramentas de bioinformática para análise de metagenômica (MANDE; MOHAMMED; GHOSH, 2012; THOMAS; GILBERT; MEYER, 2012; KIM et al., 2013; LIU et al., 2013).

Alinhamento de sequências	MG-RAST / CAMERA, MEGAN, SOrt-ITEMS, DiScRiBinATE, ProVIDE, MARTA, MetaPhyler, CARMA, AMPHORA, TreeMap, TreePhyler, Pplacer, Papara, SmashCommunity, IMG/M, KRAKEN
Composição de sequências	PHYLOPYTHA, NBC Classifier, TACOA, Phymm, RAIPhy, INDUS, ClaMS, S-GSOM, PCAHIER, Orphelia, Glimmer, MetaTISA, MetaGUN, TETRA, CD-hit, MetaCV
Híbridas	SPHINX, PhymmBL, MetaCluster, RITA

As ferramentas de composição de sequência baseiam-se no padrão de disposição dos nucleotídeos, levando-se em consideração a conservação entre as espécies ao longo da evolução. Conteúdo GC, padrões de utilização de códons ou de distribuição de oligonucleotídeos (*k-mers*)

são exemplos de medidas utilizadas. Os algoritmos utilizam essas propriedades para comparar a sequência não identificada com outras ou com modelos presentes em bancos de dados, e posteriormente estimar a classificação taxonômica da sequência (THOMAS; GILBERT; MEYER, 2012; MANDE; MOHAMMED; GHOSH, 2012). O *software* Phylopythia utiliza aprendizado de máquina por *Support Vector Machines* (SVM) para a identificação dos fragmentos de acordo com a composição de oligonucleotídeos de vários tamanhos (MCHARDY et al., 2007). Phymm também utiliza esses parâmetros, mas construindo um modelo de Markov interpolado. Os fragmentos são comparados com esses modelos e, a partir de uma ferramenta Bayesiana, classifica-se a sequência (BRADY; SALZBERG, 2009). ClaMs é um *software* que gera modelos de sequências referência utilizando grafos e propriedades de cadeias markovianas. As assinaturas pré-existentes são comparadas com as já classificadas para identificar o segmento (PATI et al., 2011). INDUS representa cada genoma na forma de vetores que deduzem os modelos de frequência de tetranucleotídeos em fragmentos individuais com tamanhos de aproximadamente 1.000 pares de base. Utiliza a distância composicional entre o fragmento e a sequência padrão, identificando o mais similar a nível taxonômico (MOHAMMED et al., 2011). TACOA aplica o método de "*k-nearest neighbor (k-NN)*", analisando padrões de oligonucleotídeos vizinhos, combinado com a "função de suavização de *kernel*"para a identificação taxonômica das sequências de metagenômica (DIAZ et al., 2009). RAIphy é um algoritmo que emprega padrões de abundância de oligonucleotídeos de sequências cuja classificação já é conhecida como referência para comparar os valores das não identificadas (NALBANTOGLU et al., 2011).

Além dos métodos descritos, há outros algoritmos, como RITA, SPHINX e PhymmBL, os quais utilizam ambas metodologias. São denominados métodos híbridos. Realizam primeiramente uma pré-seleção de táxons possíveis com algoritmos de composição de nucleotídeos, e posteriormente executam o alinhamento para uma análise mais apurada. A associação desses dois tipos de ferramentas utiliza a vantagem das ferramentas de composição de sequências: o baixo tempo computacional, juntamente com a vantagem das ferramentas de alinhamento: a classificação mais específica dos fragmentos (MANDE; MOHAMMED; GHOSH, 2012; KIM et al., 2013).

Embora uma grande variedade de algoritmos para a classificação dos metagenomas encontre-se disponível, ainda existem exigências a serem cumpridas quanto sua efetividade. Programas que utilizam alinhamento de sequências as classificam em táxons mais específicos, entretanto, seu custo computacional cresce com a diminuição dos tamanhos dos fragmentos a serem analisados. Algoritmos baseados na composição da sequência são mais rápidos, mas devido a quantidade limitada de genomas sequenciados há uma maior dificuldade em detectar organismos distantemente relacionados. Portanto, a demanda de algoritmos que compensem leitura de

pequenos fragmentos com rapidez e acurácia continua em alta (HUTCHISONIII, 2007; LIU et al., 2013).

1.3.1 Assinaturas genômicas e organização de sequências

Uma estratégia que pode vir a ser utilizada para o desenvolvimento de ferramentas para caracterização taxonômica dos metagenomas é a busca por padrões de organização de sequências e assinaturas genômicas. De acordo com o ambiente em que o organismo vive, ele estará exposto a diferentes demandas de nucleotídeos, aminoácidos e outros nutrientes e diferentes condições físicas e climáticas, portanto necessita adaptar-se de modo a otimizar o seu rendimento metabólico e estabilidade nesse meio. Por conseguinte, sua sequência de DNA passa a evoluir de maneira a suprir essas necessidades e diferentes organismos passam a desenvolver características distintas, desenvolvendo assinaturas genômicas e padrões de organização de sequência próprios (CAMPBELL; MRAZEK; KARLIN, 1999).

Por exemplo, organismos que vivem em ambientes mais quentes possuem uma maior quantidade dos nucleotídeos guanina e citosina na composição de seu DNA (TIWARI et al., 1997; CAMPBELL; MRAZEK; KARLIN, 1999; LI; SAYOOD, 2005; WANG et al., 2005; PASSEL et al., 2006), pois a interação entre adenina e timina ocorre por duas ligações de hidrogênio enquanto guanina e citosina apresentam três ligações de hidrogênio, o que confere maior estabilidade à molécula de DNA. Alguns vírus que infectam humanos, como o influenza H1N1, apresentam supressão de guanina e citosina como tentativa de assemelhar-se mais ao corpo humano, e assim driblar mais facilmente o sistema imunológico (GREENBAUM et al., 2008; GERHARDT et al., 2013).

A organização dos nucleotídeos ao longo da molécula também passa a variar conforme as diferenças entre a maquinaria de replicação e reparação das células e estrutura do DNA, como a repetição de nucleotídeos a cada três pares de base observadas nas bactérias para manter a estrutura helicoidal da sua molécula de DNA (GERHARDT; CORSO, 2006; SANTOS; RYBARCZYK-FILHO; GERHARDT, 2011; GERHARDT et al., 2013).

1.4 Proposta

Levando em conta o crescimento da utilização da metagenômica nas pesquisas científicas e os obstáculos enfrentados na análise dos dados obtidos, propomos o desenvolvimento de uma ferramenta computacional que utiliza assinaturas genômicas e medidas de organização

de sequências para a análise taxonômica dos fragmentos de DNA presentes nos metagenomas.

A utilização desses parâmetros como classificadores taxonômicos podem vir a reduzir consideravelmente o tempo computacional das análises, uma vez que essas características são facilmente mesuráveis em sequências disponíveis em bancos de dados. Sua comparação entre famílias, gêneros e espécies pode fornecer informações importantes sobre o padrão observado em cada uma e, portanto, sua utilização na identificação das sequências pode ser uma alternativa mais precisa do que as ferramentas composicionais disponíveis.

2 Objetivo

O presente trabalho tem como objetivo desenvolver um método de identificação de sequências de DNA de bactérias para aplicar no estudo de metagenomas.

2.1 Objetivos específicos

- Definir as assinaturas genômicas necessárias para identificação de micro-organismos;
- Verificar a aplicação do conceito de entropia para identificação de micro-organismos;
- Avaliar as combinações de assinaturas genômicas e entropias necessárias para identificar bactérias nos taxons família, gênero e espécie;
- Analisar o comportamento das combinações na aplicação de um metagenoma real em comparação com outra metodologia.

3 *Material e Métodos*

3.1 Banco de dados

As sequências de DNA de bactérias analisadas foram obtidas do GenBank (março de 2014) (ACLAND et al., 2013). Foi realizado o *download* de genomas completos de bactérias. Um algoritmo composto em Python selecionou as que possuíam somente ATCG em sua composição e que possuíam mais de 210000 nucleotídeos. No total, foram obtidas 2164 sequências, pertencentes a 208 famílias e 611 gêneros, com potencial para a inclusão de novas sequências para serem utilizadas de referência.

Cada genoma completo foi dividido em dois grupos (teste e controle) de 70 mil pares de base (kpb) escolhidos aleatoriamente a partir de um algoritmo desenvolvido na linguagem Python. Esse procedimento certifica que em ambos grupos não ocorram tendências nas posições. As sequências de teste e controle de cada bactéria foram então divididas em subgrupos de 100 fragmentos de cada tamanho W (64, 128, 256, 512, 1024, 2048 e 4096 pares de base), também selecionados aleatoriamente por um algoritmo em Python. A obtenção de sequências de diversos tamanhos permite analisar a influência do tamanho dos *reads* na metodologia. O algoritmo de seleção aleatória das sequências foi desenvolvido de maneira que as sequências selecionadas não se sobrepusessem (Figura 3.1).

3.2 Medidas de padrões em sequências

3.2.1 Conteúdo GC (GC%)

O conteúdo GC (Equação 3.1) tem sido utilizado como uma espécie de "impressão digital" dos organismos (GERHARDT; CORSO, 2006), pois a interação entre guanina-citosina (GC) é mais estável do que a de timina-adenosina (AT). A interação entre guanina e citosina nas moléculas de DNA de uma fita e outra é feita por três ligações de hidrogênio, enquanto que entre adenina e timina há somente duas. Portanto, quanto maior for o conteúdo GC mais estável será

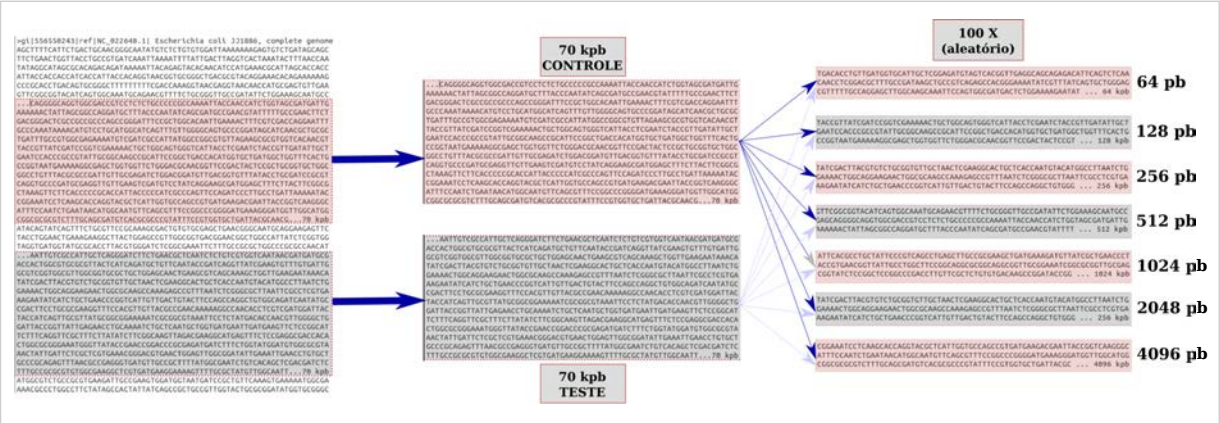


Figura 3.1: Esquema da distribuição das sequências nos grupos teste e controle. De cada genoma obtido, são selecionadas aleatoriamente sequências de teste e controle com 70 mil pares de base cada. De dentro de cada uma destas, foram selecionadas aleatoriamente 100 sequências para cada um dos tamanhos W (64, 128, 256, 512, 1024, 2048 e 4096 pb). A seleção aleatória das sequências foi feita a partir de um algoritmo desenvolvido em Python que também impede que as sequências selecionadas se sobreponham.

a sequência (YAKOVCHUK; PROTOZANOVA; FRANK-KAMENETSKII, 2006). O cálculo de conteúdo GC de uma sequência é dado pela equação 3.1 e é ilustrada pela figura 3.2:

$$GC\% = \frac{G+C}{A+C+G+T} \tag{3.1}$$

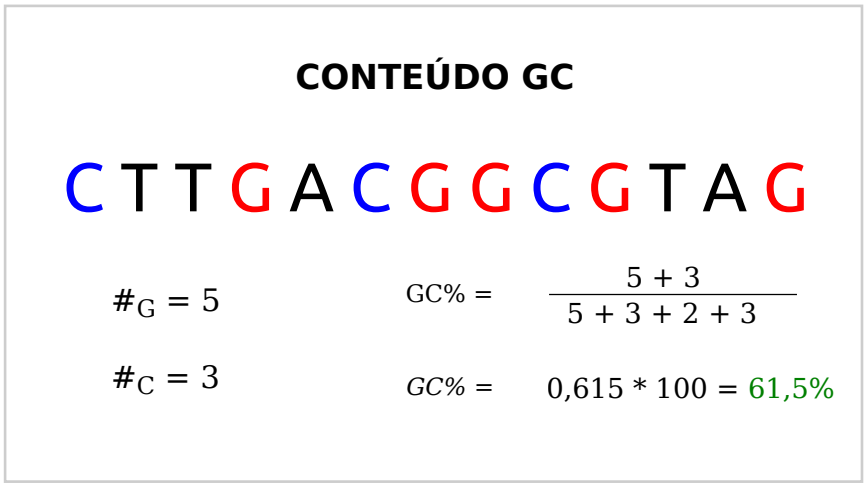


Figura 3.2: Ilustração do cálculo de conteúdo GC. É contada a quantidade de guanina ($\#_G$) e de citosina ($\#_C$) presentes na sequência. A soma desses valores é dividido pela soma do total da quantidade de todos nucleotídeos na sequência, para calcular o valor de conteúdo GC.

3.2.2 Abundância de Dinucleotídeos (din)

A frequência de dinucleotídeos tem grande influência sobre a estrutura da molécula de DNA, podendo interferir na sua interação com proteínas, principalmente as que atuam no pro-

cesso de replicação e tradução (KARLIN; BURGE, 1995). Estes processos são vitais para a célula, portanto sofrem grande pressão de seleção, o que passa a acentuar a diferença dos padrões entre as espécies e entre os organismos de diferentes ambientes (DUTTA; PAUL, 2012), aumentando sua eficiência como parâmetro de identificação de sequências.

A abundância relativa de dinucleotídeos é a medida mais utilizada para calcular frequência. É representada pela equação 3.2 (KARLIN; BURGE, 1995; KARLIN, 1998):

$$\rho_{XY} = \frac{f_{XY}}{f_X f_Y} \tag{3.2}$$

onde f_{XY} representa a frequência do dinucleotídeo XY na sequência, f_X representa a frequência do nucleotídeo X e f_Y a frequência do nucleotídeo Y. Essa equação resulta na diferença entre a frequência observada do dinucleotídeo XY na sequência e sua frequência esperada em uma sequência aleatoria que possua a mesma quantidade dos nucleotídeos X e Y (KARLIN; BURGE, 1995; KARLIN, 1998; DUTTA; PAUL, 2012).

O valor da abundância de dinucleotídeos é obtido pela média da abundância relativa de cada dinucleotídeo, como representado na equação 3.3 e na figura 3.3.

$$din = \frac{1}{16}\sum_X\sum_Y \frac{f_{XY}}{f_X f_Y},\ X \in \{\mathrm{A,C,T,G}\},\ Y \in \{\mathrm{A,C,T,G}\} \tag{3.3}$$

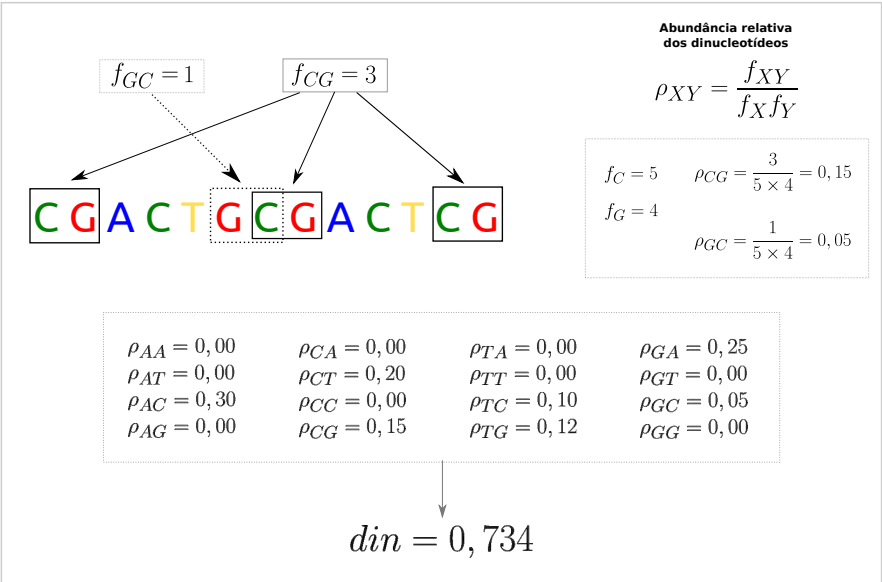


Figura 3.3: Ilustração do cálculo de abundância de dinucleotideos. É calculada a abundância relativa de cada dinucleotídeo (ρ_{XY}) da sequência utilizando o valor de sua frequência (f_{XY}) e da frequência de cada nucleotídeo (f_X e f_Y) no segmento de DNA. Podemos observar que as frequências de CG e de GC (f_{CG} e f_{GC}) são diferentes, assim como sua abundância relativa (ρ_{CG} e ρ_{GC}). A partir da abundância relativa de todos dinucleotídeos, é feita a média para a obtenção do valor de abundância de dinucleotídeos da sequência completa (Equação 3.3).

3.2.3 N-entropia (S)

A palavra *entropia* conota desordem ou incerteza. A métrica de entropia de um sistema é valor da desordem ou incerteza inerente neste sistema (HANKERSON; JOHNSON; HARRIS, 1998). Neste caso a entropia representa o padrão de desordem na distribuição de oligonucleotídeos em uma sequência. Neste trabalho estamos aplicando medidas de entropia de dipletes (S_2), tripletes (S_3) e tetrapletes (S_4).

Para o cálculo da entropia, é necessário primeiramente estimar a probabilidade de cada oligonucleotídeo possível na sequência (P_n). Para a obtenção desse valor, conta-se a quantidade de vezes que cada oligonucleotídeo em uma janela deslizante aparece na sequência e divide-se esse número pela quantidade total de nucleotídeos do segmento (Figura 3.4) (GERHARDT; CORSO, 2006; SANTOS; RYBARCZYK-FILHO; GERHARDT, 2011).

Figura 3.4: Ilustração das janelas aplicadas para as entropias de dipletes, tripletes e tetrapletes.

O cálculo de entropia é realizado a partir da equação da entropia de Shannon (Equação 3.4) (SHANNON, 1948):

$$S = -\frac{1}{\ln N} \sum_{n=1}^{N} P_n \log P_n, \tag{3.4}$$

P_n simboliza a probabilidade do e-nésimo oligonucleotídeo presente em uma sequência de tamanho W (64 pb, 128 pb, 256 pb, 512 pb, 1024 pb, 2048 pb e 4096 pb), N representa a quantidade de oligonucleotídeos possíveis, sendo 16 para dinucleotídeos, 64 para trinucleotídeos e 256 para tetranucleotídeos.

3.3 Valores de referência

Cada genoma selecionado foi fragmentado aleatoriamente em grupos de 100 sequências de tamanhos W (64 pb, 128 pb, 256 pb, 512 pb, 1024 pb, 2048 pb e 4096 pb), tanto para teste quanto para controle. Para cada grupo, foram calculadas a partir de algoritmos desenvolvidos em Python as medidas citadas acima e a partir destas, calculadas a média e desvio padrão dos valores de controle.

Isso possibilitou a geração de uma tabela para cada sequência com largura W. Nesta tabela estavam apresentados os valores de referência de média e desvio padrão das assinaturas genômicas (conteúdo GC, abundância de dinucleotídeos e entropia de dipletes, tripletes e tetrapletes) para cada cepa. Assim, cada micro-organismo foi representado por um vetor de posição $\vec{R}_{(B,W)} = (GC, din, S_2, S_3, S4)$ e um vetor de desvio padrão $\vec{\sigma}_{(B,W)} = (\sigma_{GC}, \sigma_{din}, \sigma_{S_2}, \sigma_{S_3}, \sigma_{S_4})$.

3.4 Análise estatística

3.4.1 Grupos de teste

Foram gerados 100 grupos de teste de cada tamanho W para cada gênero e família que possuíam mais de uma espécie (resultando em 133 famílias e 179 gêneros), e 1 grupo de teste para cada espécie. Esses grupos eram formados por 100 sequências "positivas"(pertencentes a somente uma família ou gênero) e 100 sequências "negativas"(não pertencentes ao mesmo organismo dos exemplos positivos). Essas sequências foram selecionadas aleatóriamente a partir de um algoritmo desenvolvido em Python que certificava-se que não se repetissem dentro de um grupo. Esse algoritmo também assegurava-se de que os gêneros ou famílias não se repetissem dentro dos exemplos negativos dos grupos de teste. Os valores das assinaturas genômicas e entropias de cada sequência dos grupos de teste foram comparados com os valores de referência do organismo a quem pertenciam os exemplos positivos de cada grupo (Fig. 3.5).

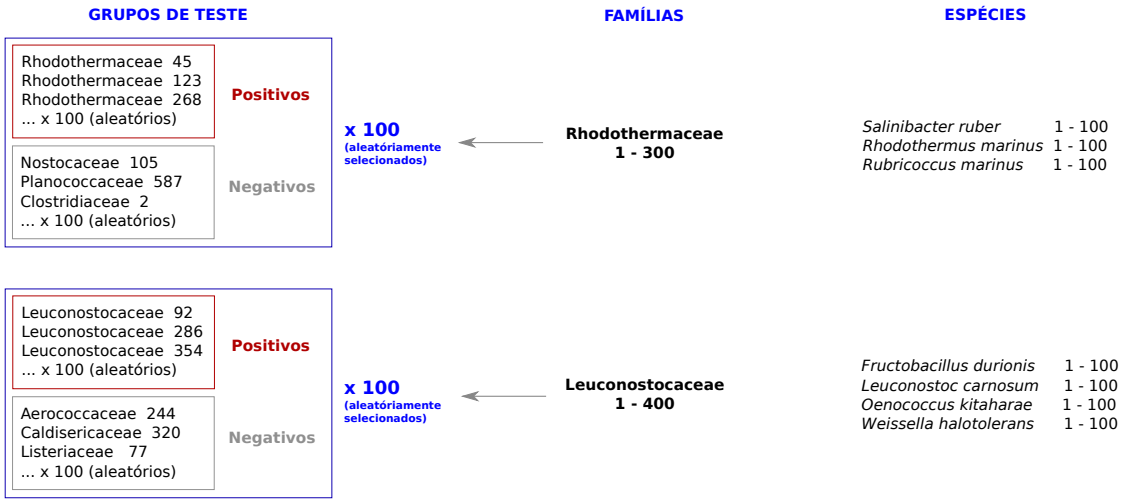


Figura 3.5: Ilustração da formação dos grupos de teste. Cada grupo de teste foi formado por 100 sequências positivas (de um mesmo gênero ou família) e 100 sequências negativas (de organismos diferentes daquele dos exemplos positivos). Um dos exemplos ilustrados na figura é a família Rhodothermaceae. Dentre os genomas presentes no banco de dados, 3 deles eram de espécies pertencentes à essa família. Uma vez que para cada genoma há 100 sequências de cada tamanho W, a família Rhodothermaceae possui 300 sequências, a partir das quais os exemplos positivos foram escolhidos aleatoriamente por um algoritmo em Python. Foram gerados 100 grupos de teste para cada família e gênero.

3.4.2 Combinações de medidas

As combinações foram nomeadas conforme a quantidade de medidas envolvida (2 a 5, no primeiro dígito) e uma identificação numérica da combinação em questão (Figura 3.6).
Foram geradas 26 combinações de medidas que determinavam quais delas seriam utilizadas para identificar a sequência teste (Tabela 3.1), pelas quais as sequências passaram como uma espécie de "filtro" para a seleção de famílias, gêneros ou espécies a que pertence aquele fragmento de DNA (Figura 3.7). Cada grupo de teste foi analisado com cada uma dessas combinações (como ilustrado na figura 3.8), e os valores estatísticos foram gerados para cada uma com o intuito comparar a eficiência de cada combinação.

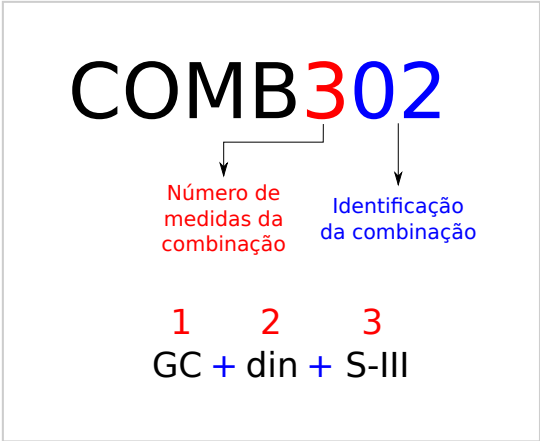


Figura 3.6: Esquema explicando o nome das combinações utilizadas nas análises das sequências. O primeiro dígito representa quantas medidas estão sendo utilizadas na combinação (2 a 5), e os dois últimos dígitos representa o número de identificação daquela combinação.

Tabela 3.1: Tabela de nomenclatura para as combinações de medidas usadas para as análises.

Nome	**Combinações de medidas**			**Nome**	**Combinações de medidas**				
COMB201	din	GC		COMB304	GC	S-II	S-III		
COMB202	GC	S-II		COMB305	GC	S-II	S-IV		
COMB203	GC	S-III		COMB306	GC	S-III	S-IV		
COMB204	GC	S-IV		COMB307	din	S-II	S-III		
COMB205	din	S-II		COMB308	din	S-II	S-IV		
COMB206	din	S-III		COMB309	din	S-III	S-IV		
COMB207	din	S-IV		COMB310	S-II	S-III	S-IV		
COMB208	S-II	S-III		COMB401	din	GC	S-II	S-III	
COMB209	S-II	S-IV		COMB402	din	GC	S-II	S-IV	
COMB210	S-III	S-IV		COMB403	din	GC	S-III	S-IV	
COMB301	din	GC	S-II	COMB404	GC	S-II	S-III	S-IV	
COMB302	din	GC	S-III	COMB405	din	S-II	S-III	S-IV	
COMB303	din	GC	S-IV	COMB501	din	GC	S-II	S-III	S-IV

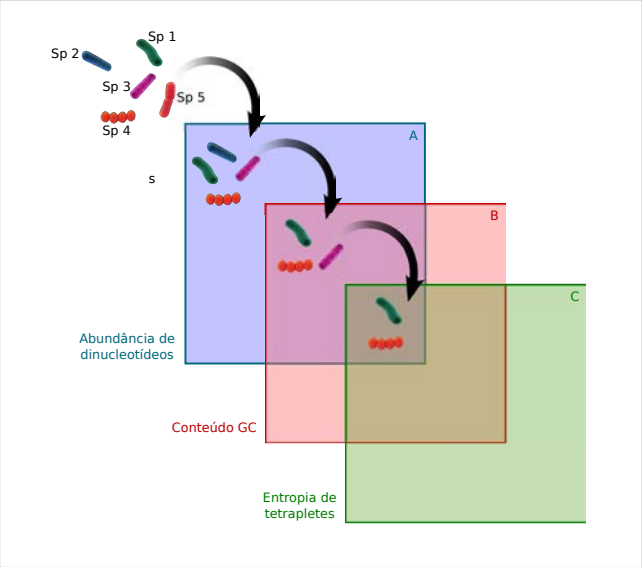


Figura 3.7: Ilustração da utilização das combinações para a seleção de espécies pertencentes àquela sequência. A figura mostra como exemplo a COMB303, que é composta por conteúdo GC, abundância de dinucleotídeos e entropia de tripletes.Um fragmento de DNA é testado para as medidas que pertencem àquela combinação. Os valores de cada medida são comparados, e quando não correspondem às espécies, estas são descartadas. A) Os valores para abundância de dinucleotídeos não correspondem com a Sp 5, somente com as Sp 1, 2, 3 e 4. B) Os valores para conteúdo GC não correspondem com a Sp 2, somente com as Sp 1, 3 e 4. C) Os valores para entropia não correspondem com a Sp 3, somente com as Sp 1 e 4. A sequência seria classificada como Sp 1 ou Sp 4.

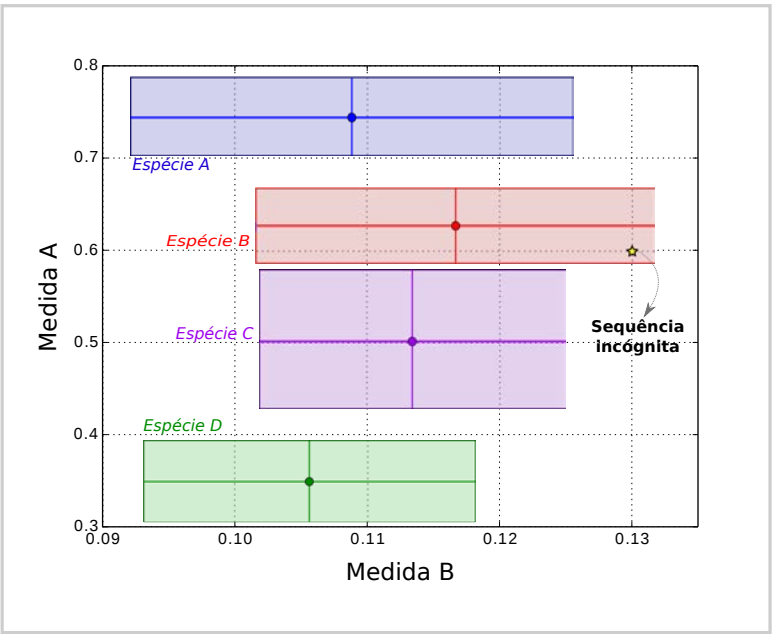


Figura 3.8: Esquema da identificação de uma sequência obtida por metagenômica ($\star$) a partir de duas medidas. A média ($\bullet$) e desvio padrão ($-$) dos valores para cada espécie estão apresentadas no gráfico. Os valores das medidas da sequência incógnita são 0.6 e 0.13, sendo assim, esse fragmento estaria classificado como a Espécie C, pois encontra-se no intervalo de valores correspondentes a ela.

3.4.3 Matriz de confusão

Observando a tabela gerada do grupo controle ($\vec{R}_{(B,W)}$) foi possível examinar quantos e quais organismos estão dentro do desvio padrão ($\vec{\sigma}_{(B,W)}$) de cada parâmetro. A partir de então, foi possível computar a quantidade de: i) ocorrências em que as sequências positivas foram registradas como sendo o organismo de referência correto (verdadeiro positivo - VP); ii) ocorrências em que as sequências positivas foram computadas como não sendo a espécie de referência (falsos negativos - FN); iii) ocorrências em que as sequências negativas foram identificadas como sendo a espécie de referência (falsos positivos - FP); iv) ocorrências em que as sequências negativas foram indicadas como não sendo a espécie de referência (verdadeiros negativos - VN) (Fig 3.9).

A partir dos valores descritos acima, foram calculados alguns parâmetros para a verificação estatística da metodologia (POWERS, 2007), tais como:

1. Sensibilidade, ou índice de verdadeiros positivos (Fig. 3.9(a)). Representa a proporção dos casos realmente positivos que são classificados como tal. É representada por

$$Sens_W = \frac{VP}{VP+FN}; \tag{3.5}$$

2. Especificidade, ou índice de verdadeiros negativos (Fig. 3.9(b)). Calcula a proporção de casos realmente negativos que são corretamente preditos como negativos. É representada por

$$Espec_W = \frac{VN}{FP+VN}; \tag{3.6}$$

3. Precisão, ou índice de predição de acertos (Fig. 3.9(c)). Caracteriza a proporção de casos preditos corretamente como positivos, representada por

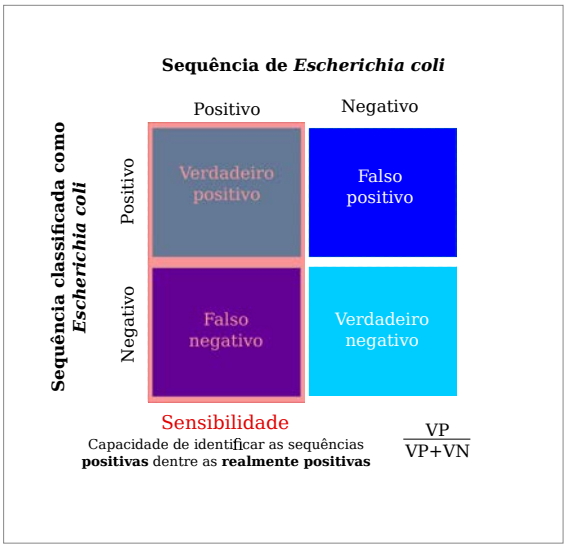
$$Prec_W = \frac{VP}{VP+FP}; \tag{3.7}$$

4. Média harmônica da precisão e sensibilidade (Fig. 3.9(d)). Calculada a partir da equação

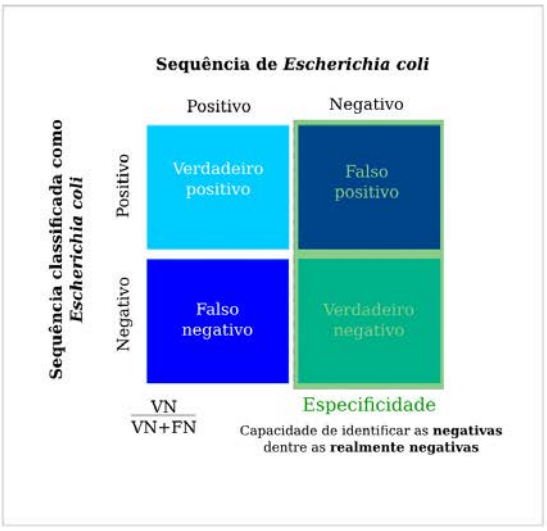
$$Harm_W = \frac{2VP}{2VP+FP+FN}; \tag{3.8}$$

Para cada combinação, foram observados os valores das médias dos grupos de teste de famílias, gêneros e espécies. Índices de acertos próximos ou abaixo de 0,5 (50%) são considerados aleatoriamente obtidos.

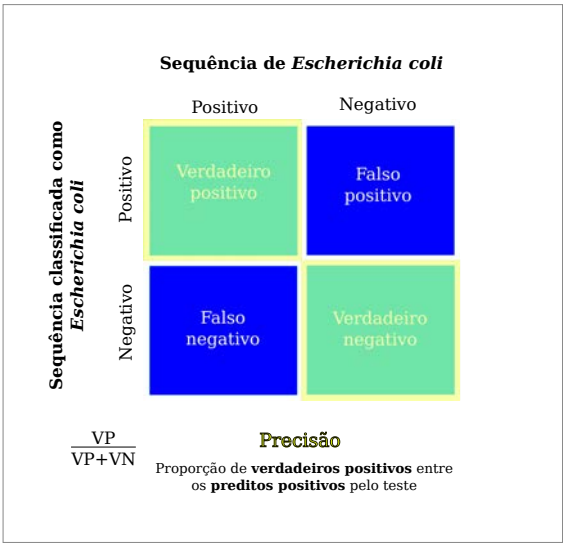
São considerados valores ideiais aqueles que passarem de 0,6 (60%). Portanto, famílias, gêneros ou espécies que apresentaram todos as medidas estatísticas citados acima de 0.6 (60%)



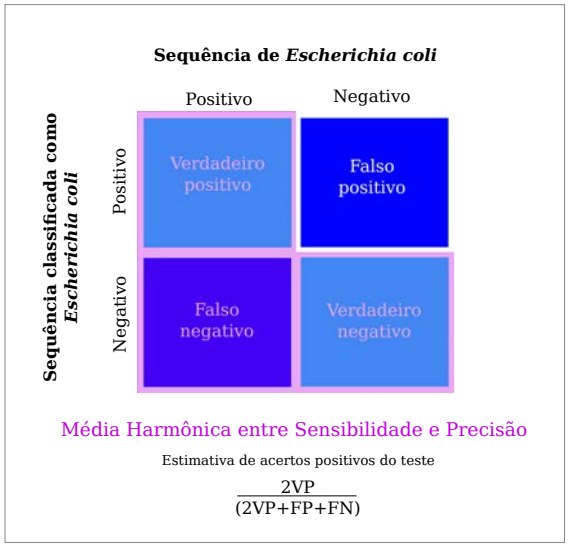
(a) Sensibilidade



(b) Especificidade



(c) Precisão



(d) Média Harmônica

Figura 3.9: Medidas estatísticas para a avaliação da metodologia. São calculadas a sensibilidade (a), especificidade (b), precisão (c) e média harmônica entre sensibilidade e precisão (d). Calcula-se a sensibilidade a partir dos valores de verdadeiro positivo e falso negativo, a especificidade a partir dos falsos positivos e verdadeiros negativos, a precisão a partir dos verdadeiros positivos e verdadeiros negativos e a média harmônica entre sensibilidade e precisão a partir dos valores de verdadeiros positivos, falsos negativos e verdadeiros negativos.

considerados como identificados pelas combinações testadas (Tabela 3.2).

Tabela 3.2: Tabela esquematizando os critérios de identificação de famílias, gêneros e espécies.

Familia	Sens.	Espec.	Prec.	MH	ID
Fam. 1	0.62	0.73	0.70	0.66	**IDENTIFICADO**
Fam. 2	0.47	0.81	0.72	0.57	**NÃO IDENTIFICADO**
Fam. 3	0.35	0.14	0.29	0.48	**NÃO IDENTIFICADO**

3.5 Prova de conceito

Utilizamos o metagenoma da areia das praias do oceano Pacífico (*Pacific Beach Sand*) (NAVIAUX et al., 2005) (disponível no portal CAMERA (CAMERA, 2014)) para a comparação de nossa metodologia com a do software MEGAN (HUSON et al., 2007) . Esse metagenoma possui 4981 sequências de DNA com, em média, 1203 pb cada um. Primeiramente foi utilizado BlastX com configurações padrão (ALTSCHUL et al., 1990) para comparar os fragmentos de DNA do metagenoma com o banco de dados NCBI-NR ("não redundante") (BENSON et al., 2012). O arquivo resultante foi utilizado como entrada para o MEGAN com os parâmetros padrão do programa.

Após a análise com o MEGAN, foram calculadas as entropias, conteúdo GC e abundância de dinucleotídeos dos fragmentos do metagenoma em questão. Os valores obtidos foram comparados com os valores controle das 208 famílias e 611 gêneros (para sequências de 1024 pb) de acordo com as combinações de medidas apresentadas na Tabela 3.1. As famílias e gêneros selecionados compunham um vetor de possibilidades de táxons aos quais aquele fragmento de DNA tenha a probabilidade de ser classificado.

Verificou-se no vetor de cada sequência se a família e gênero atribuído a ela pelo MEGAN estava presente. Em caso positivo, aquela sequência era considerada como corretamente identificada pela combinação. As sequências cujo vetor não era composto de nenhuma família ou gênero eram declaradas como não classificadas (Figura 3.10).

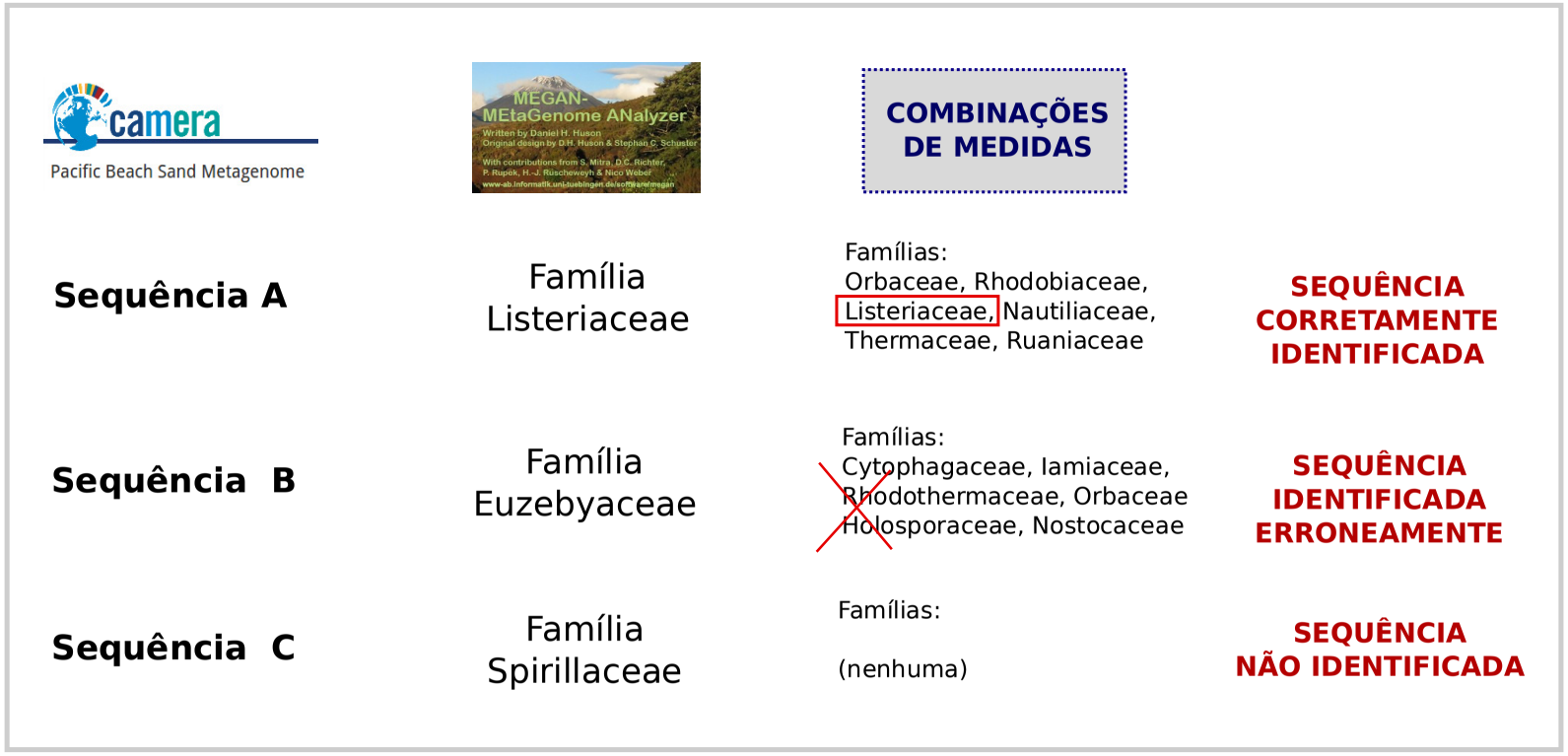


Figura 3.10: Esquema ilustrando o critério de seleção para as sequências corretamente identificadas, erroneamente identificadas ou não identificadas do metagenoma *Pacific Beach Sand* a partir das combinações de medidas, utilizando a classificação do *software* MEGAN como referência.

4 Resultados e Discussão

4.1 Avaliação das medidas de padrão de sequência

Para analisar o comportamento de cada medida separadamente na identificação de sequências, foram feitos testes de identificação de táxons utilizando as medidas de conteúdo GC, abundância de dinucleotídeo e entropias de dipletes, tripletes e tetrapletes isoladamente.

A partir dessas comparações foram gerados valores de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Com base nesses valores foram calculadas a sensibilidade, especificidade, precisão e média harmônica para cada família.

4.1.1 Conteúdo GC

A partir da identificação das sequências somente com conteúdo GC (Figura 4.1), podemos observar que o valor médio de todas medidas estatísticas na identificação de famílias estão em torno ou acima de 70%. Isso mostra que o conteúdo GC utilizado isoladamente pode ser considerado uma medida promissora para a identificação tanto de verdadeiros positivos quanto de verdadeiros negativos de um conjunto de sequências não classificadas. Obervamos que a utilização do conteúdo GC é eficaz para a identificação de todos tamanhos de fragmentos utilizados na análise, evidenciando sua aplicabilidade nas sequências obtidas pelas diversas tecnologias de sequenciamento disponíveis.

O conteúdo GC é uma medida comprovadamente influenciada por condições ambientais, tamanho da sequência e maquinaria de replicação e reparação (FOERSTNER et al., 2005; GARCIA et al., 2008). Esses fatores evidenciam a diferença composicional das sequências de cada espécie, tornando o conteúdo GC um bom parâmentro para a diferenciação das espécies.

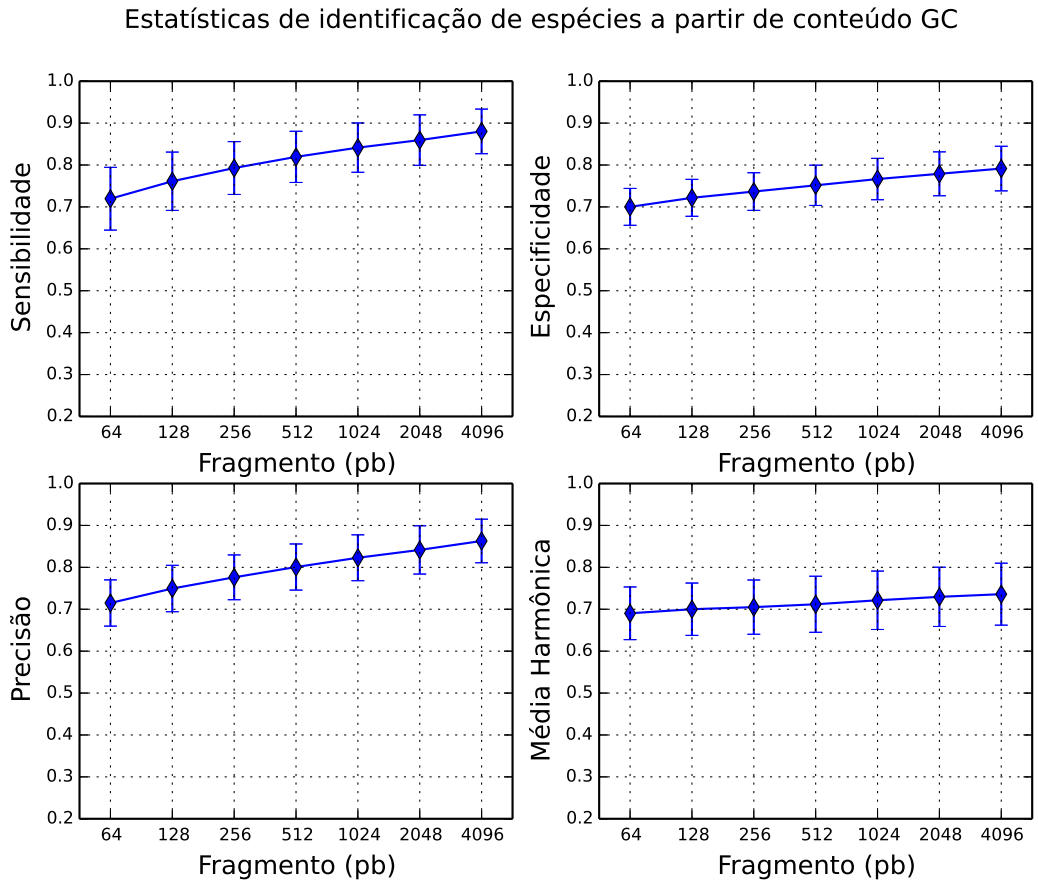


Figura 4.1: Gráfico de sensibilidade, especificidade, precisão e média harmônica para análises de espécie utilizando somente conteúdo GC.

4.1.2 Abundância de Dinucleotídeos

Podemos observar no gráfico 4.2 que a medida de abundância de dinucleotídeos apresenta altos valores de especificidade. Os valores passam de, em média, 62% para fragmentos de 64 pb e aumentam juntamente com o número dos pares de base, atingindo um valor médio de 70% para fragmentos de 4096 pb.

Entretanto, apresentam baixos valores de sensibilidade - que chegam até 22% para os fragmentos de 64 pb. Para as sequências de 4096 pb, a média da sensibilidade passa de 60%, mas o valor de seu desvio padrão está abaixo de 50%. Isso indica que, utilizada isoladamente, a abundância de dinucleotídeos pode ser um bom parâmetro de identificação de verdadeiros negativos, mas mostra um baixo potencial na identificação de verdadeiros positivos.

Seus valores de especificidade foram o fator chave para a inclusão da medida de abundância de dinucleotídeos na análise de combinações de medidas.

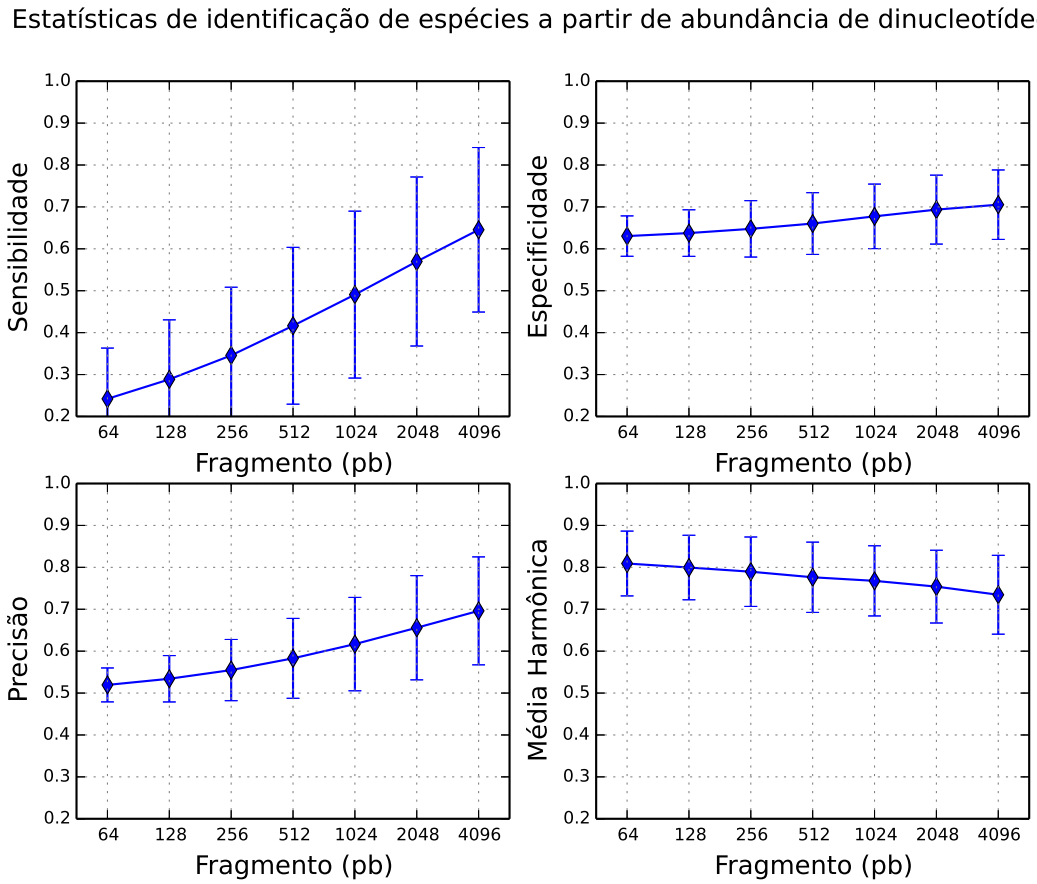


Figura 4.2: Gráfico de sensibilidade, especificidade, precisão e média harmônica para a comparação das sequências teste com os valores de controle utilizando abundância de dinucleotídeos.

Os dinucleotídeos apresentam papéis cruciais na estrutura do DNA, como torção da hélice, curvatura e flexibilidade do DNA, o que os faz serem selecionados evolutivamente (KARLIN; BURGE, 1995; KARLIN, 1998). Isso torna a abundância de dinucleotídeos um parâmetro interessante para incluir nas análises de identificação de sequências metagenômicas.

4.1.3 N-entropias

Para avaliar o desempenho de cada entropia separadamente, foram feitas análises utilizando somente essas medidas para identificar as famílias das sequências de teste.

Quando utilizadas cada entropia isoladamente, podemos obervar no gráfico que as três apresentam valores aproximados. Para a sensibilidade, o valor médio das entropias para o tamanho 4096 pb foi de 60%, entretanto o desvio padrão chega a valores abaixo de 50%. Além disso, os valores das entropias chegam a 30% de sensibilidade para as entropias de dipletes e tripletes, indicando que a utilização das entropias isoladamente pode não ser um bom parâmetro para identificar a quais famílias os fragmentos de DNA podem pertencer.

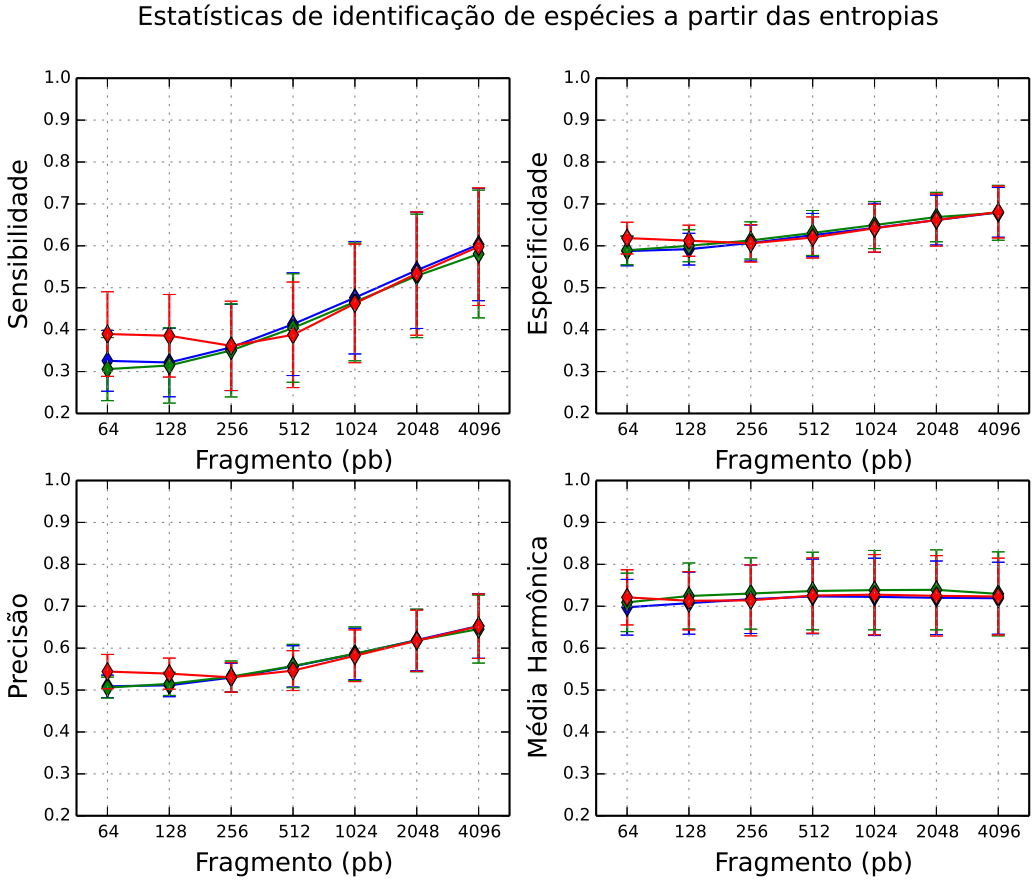


Figura 4.3: Gráfico de sensibilidade, especificidade, precisão e média harmônica para análises de espécie utilizando entropias. A entropia de dipletes está representada em verde, a de tripletes em azul e a de tetrapletes em vermelho.

Entretanto, a média dos valores de especificidade localizam-se em torno de 60%, indicando que as entropias podem apresentar um bom potencial para identificar a que famílias as sequências metagenômicas não pertencem. Por esse fator, as entropias foram utilizadas para serem analisadas juntamente com outras medidas nas combinações.

Embora as entropias apresentem pouca diferença significativa entre elas, há um pequeno aumento na sensibilidade, especificidade e precisão para os tamanhos 64 e 128 pb na curva da entropia de tetrapletes. Estas, apresentam um valor um pouco mais elevado, atingindo em torno de 10% a mais de sensibilidade para 64 e 128pb do que as outras entropias. Isso pode ser uma indicação de um maior potencial da entropia de tetrapletes para a análise de sequências menores, embora esses resultados tenham sido obtidos para as medidas de entropia isoladamente.

4.2 Análise de desempenho para famílias

A partir dos grupos de teste, foi avaliada a quantidade de famílias que cada combinação pôde identificar corretamente. Foram consideradas como identificadas famílias cuja combinação atingiu valores de todas medidas estatísticas (sensibilidade, especificidade, precisão e média harmônica) acima de 60%. Esse valor foi considerado como corte ideal porque corresponde ao desempenho médio dos algoritmos disponíveis para identificação dos fragmentos, como ilustrado na Figura 4.4.

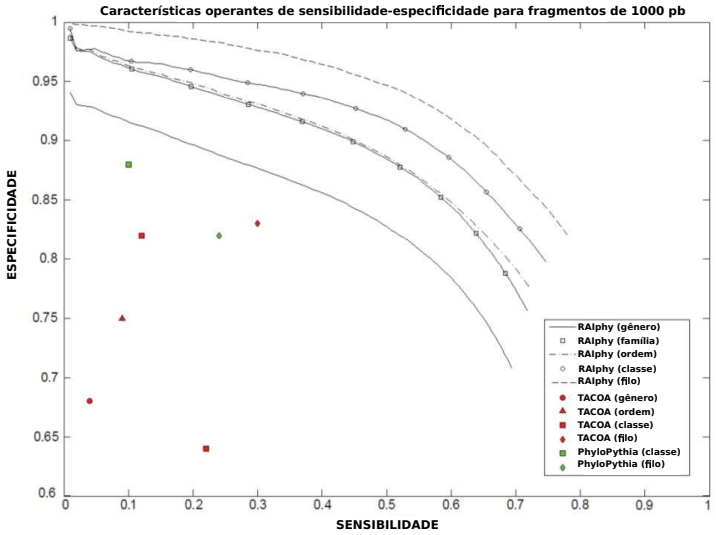


Figura 4.4: Valores de sensibilidade e especificidade para classificação taxonômica de fragmentos de 1 kpb das ferramentas RAIphy, TACOA e PhyloPythia. Figura publicada em (NALBANTOGLU et al., 2011).

A quantidade de famílias que cada combinação pôde identificar corretamente, para as sequências de tamanho 64 pb, está representada na Figura 4.5.

Observamos que a COMB201 (combinação entre abundância de dinucleotídeos e conteúdo GC) identificou corretamente 68 das 133 famílias, totalizando 51%, seguido da COMB204 (combinação entre conteúdo GC e entropia de tetrapletes), que identificou 20 das 133 (15%), e da combinação COMB207, composta por abundância de dinucleotídeos e entropia de tetrapletes, que identificou 5% das famílias, ou seja, 6 famílias das 133.

Das famílias identificadas pelas combinações COMB201, COMB204 e COMB207, 1 delas foi identificada pelas três combinações. 16 famílias foram identificadas em comum pelas combinações COMB201 e COMB204, 3 identificadas pelas COMB201 e COMB207 e nenhuma pelas COMB204 e COMB209. 48 famílias foram identificada somente pela COMB201, 3 famílias pela COMB204 e 2 pela COMB207 (Figura 4.6). A Tabela 4.1 apresenta as famílias identificadas em comum e individualmente pelas combinações.

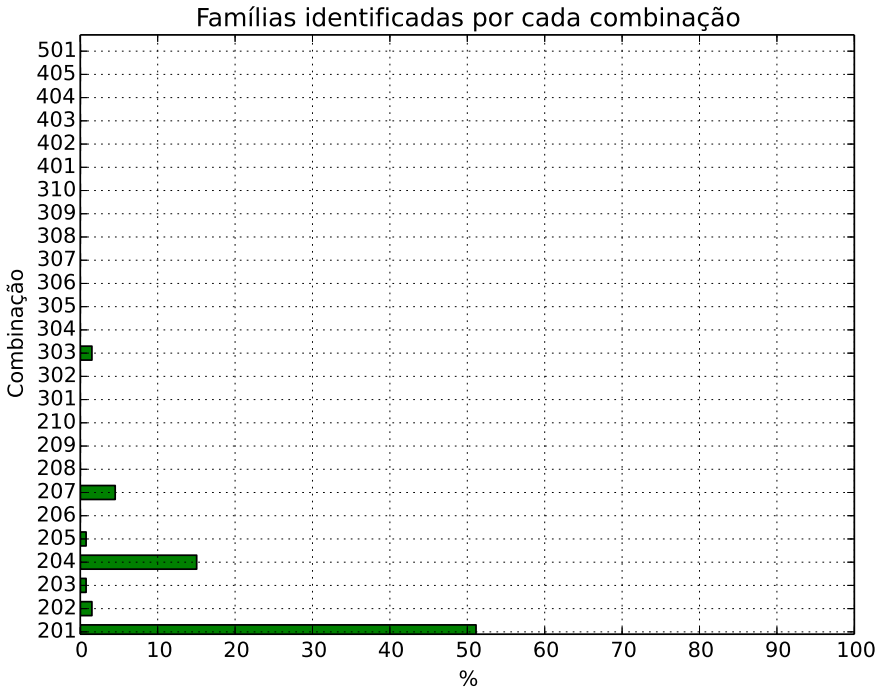


Figura 4.5: Gráfico ilustrando a quantidade de famílias que cada combinação identifica acima de 60% para todas as medidas estatísticas.

A única família identificada em comum pelas três combinações foi a Bdellovibrionaceae. Ela pertence ao filo Proteobacteria, classe Deltaproteobacteria. As Proteobacterias são de um dos maiores e mais bem descritos filos de bactérias, a qual pertencem a maioria das bactérias de importância médica, industrial e agrícola (MADIGAN et al., 2010). Portanto, tende a ser bastante estudado e caracterizado, o que pode a vir facilitar na identificação dos fragmentos que pertencem a esse grupo.

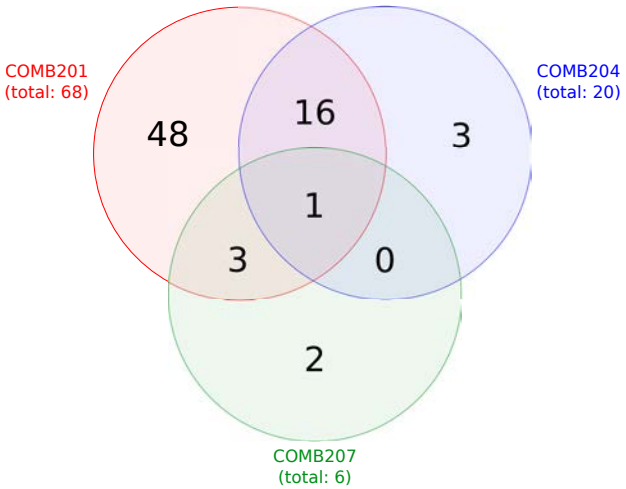


Figura 4.6: Diagrama ilustrando a quantidade de famílias identificadas (acima de 60% para todas as medidas estatísticas) em comum entre as combinações COMB201, COMB204 e COMB207.

Tabela 4.1: Tabela de famílias em comum para as combinações mostradas na Figura 4.6.

Somente COMB201	Acetobacteraceae, Acidithiobacillaceae, Aeromonadaceae, Alcanivoracaceae, Anaplasmataceae, Aquificaceae, Beijerinckiaceae, Bifidobacteriaceae, Brucellaceae, Burkholderiaceae, Campylobacteraceae, Caulobacteraceae, Chlamydiaceae, Chlorobiaceae, Comamonadaceae, Cyclobacteriaceae, Cytophagaceae, Desulfobacteraceae, Enterococcaceae, Eubacteriaceae, Flavobacteriaceae, Geobacteraceae, Helicobacteraceae, Lachnospiraceae, Leptospiraceae, Leuconostocaceae, Moraxellaceae, Mycobacteriaceae, Mycoplasmataceae, Myxococcaceae, Neisseriaceae, Nocardiaceae, Pasteurellaceae, Peptococcaceae, Porphyromonadaceae, Propionibacteriaceae, Pseudomonadaceae, Rhodobacteraceae, Rhodocyclaceae, Ruminococcaceae, Saprospiraceae, Shewanellaceae, Sphingobacteriaceae, Staphylococcaceae, Streptococcaceae, Streptomycetaceae, Synergistaceae, Vibrionaceae
Somente COMB204	Dictyoglomaceae, Micromonosporaceae, Rhizobiaceae
Somente COMB207	Alteromonadaceae, Oceanospirillaceae
COMB201 & COMB204	Acidaminococcaceae, Bartonellaceae, Clostridiaceae, Cryomorphaceae, Cystobacteraceae, Deferribacteraceae, Entomoplasmataceae, Frankiaceae, Halobacteroidaceae, Methylobacteriaceae, Nocardiopsaceae, Polyangiaceae, Pseudonocardiaceae, Rickettsiaceae, Xanthobacteraceae, Xanthomonadaceae
COMB204 & COMB207	Dehalococcoidaceae, Methylococcaceae, Nitrosomonadaceae
COMB201 & COMB204 & COMB207	Bdellovibrionaceae

4.2.1 Análise estatística

As figuras 4.7 e 4.8 mostram os gráficos das medidas de sensibilidade, especificidade, precisão e média harmônica para as melhores combinações segundo a figura 4.5: COMB201 (conteúdo GC e abundância de dinucleotídeos) e COMB204 (conteúdo GC e entropia de tetrapletes), respectivamente.

As figuras mostram que a COMB201 é uma medida classificadora de verdadeiros positivos mais eficaz do que a COMB204, pois apresenta valores de sensibilidade em torno de 60% para todos os tamanhos de fragmento, enquanto os valores da outra encontram-se mais próximos de 50%. Além disso, os valores da média harmônica entre sensibilidade e precisão da COMB201 também são maiores do que da COMB204, embora a precisão tenha apresentado valores quase idênticos.

De acordo com os valores de especificidade, as duas combinações apresentaram um ótimo potencial para a indicação de famílias a quem as sequências não pertencem, pois apresentam valores maiores do que 70% para as sequências de tamanho 64 pb e passam de 85% para 4096 pb. Embora as duas tenham valores similares entre si, a COMB204 apresenta um potencial ligeiramente maior, pois seus valores são sutilmente mais elevados.

Entretanto, a COMB201 pode ser considerada um melhor classificador da família das sequências, pois valores ótimos para todas medidas estatísticas, enquanto a COMB204 apre-

senta sensibilidade mais próxima a 50%.

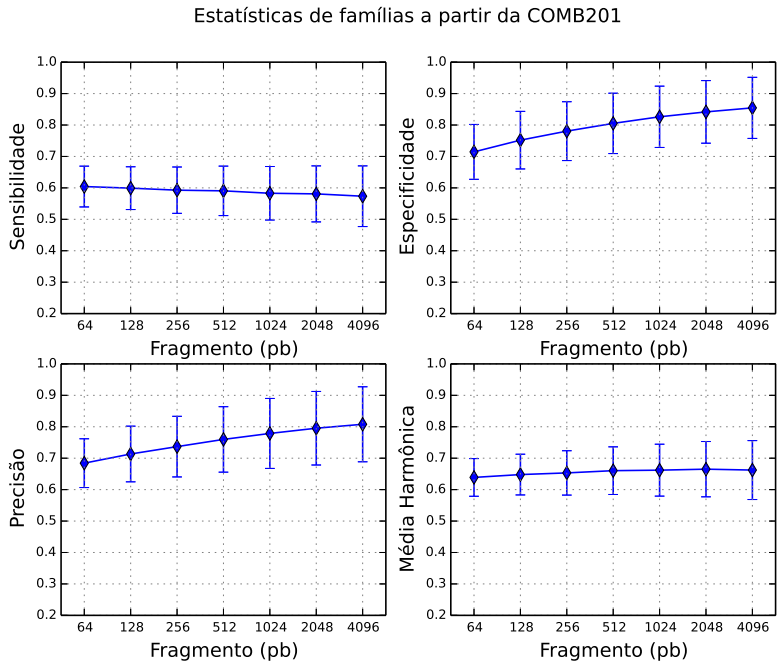


Figura 4.7: Gráfico de sensibilidade, especificidade, precisão e média harmônica para análises de famílias utilizando COMB201.

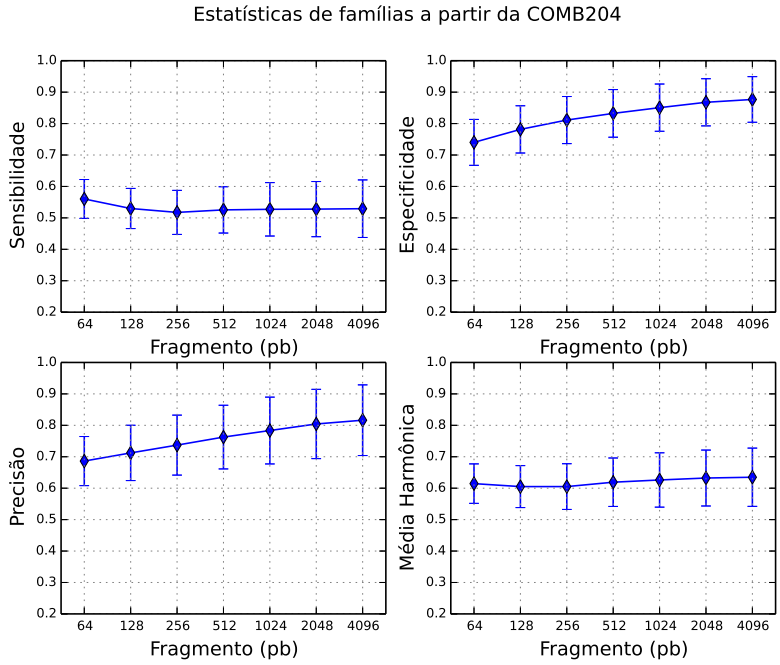


Figura 4.8: Gráfico de sensibilidade, especificidade, precisão e média harmônica para análises de famílias utilizando COMB204.

4.3 Análise de desempenho para gêneros

A partir dos grupos de teste, foram contados para quantos gêneros cada combinação obteve todos valores estatísticos maiores do que 60%. Esses resultados estão representadas na figura 4.9, apresentando a porcentagem de gêneros que cada combinação identificou corretamente.

As combinações que apresentaram mais acertos entre os gêneros foram a COMB201, que inclui abundância de dinucleotídeos e conteúdo GC, e a COMB204, composta por conteúdo GC e entropia de tetrapletes.

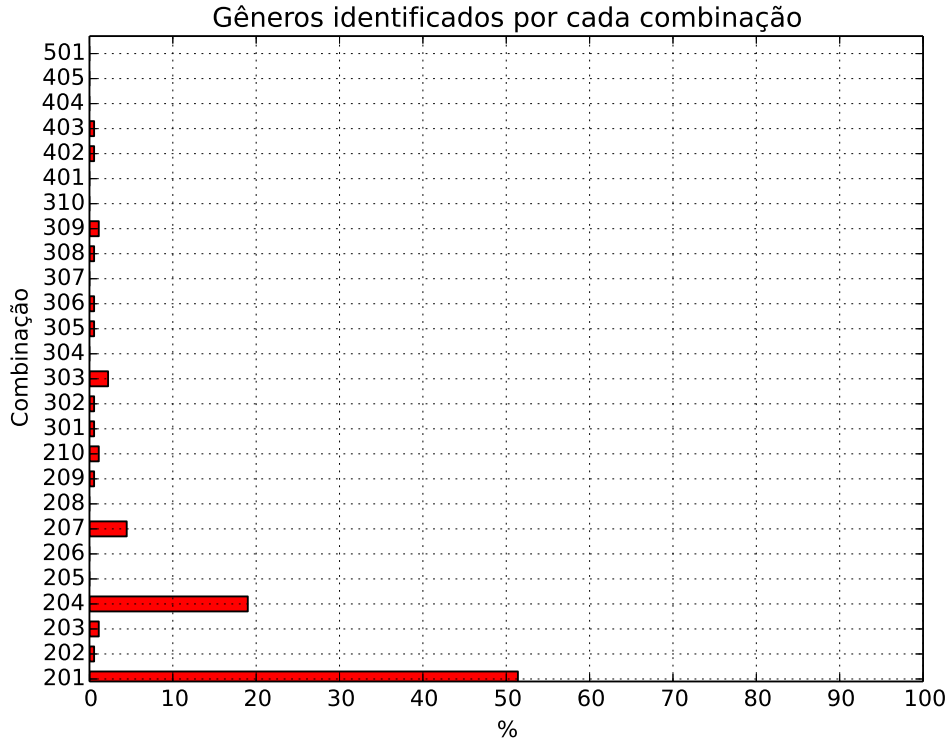


Figura 4.9: Gráfico ilustrando a quantidade de gêneros que cada combinação identifica acima de 60% para todas as medidas estatísticas.

Observamos que 2 gêneros foram identificados em comum pelas três medidas: *Anaeromyxobacter* e *Bdellovibrio*. 24 foram identificados pelas combinações COMB201 e COMB204, 3 pela COMB201 e COMB207 e nenhum pelas COMB204 e COMB207. Os gêneros identificados em comum e individualmente pelas combinações estão apresentados na Tabela 4.2.

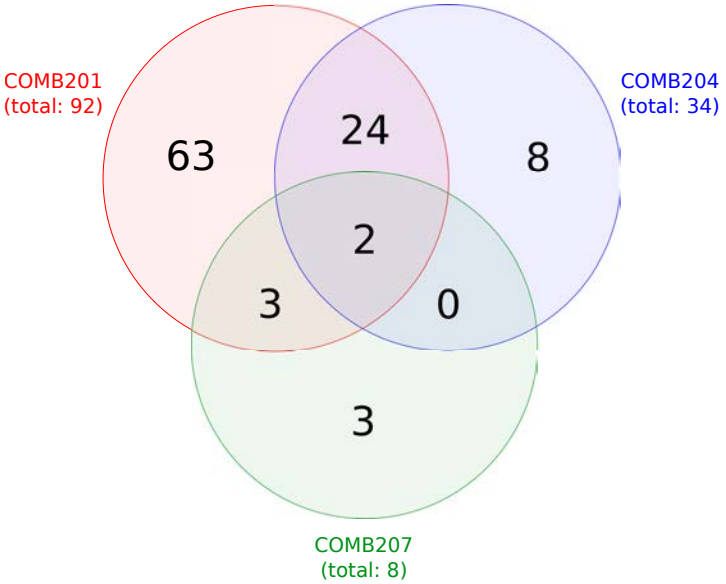


Figura 4.10: Diagrama ilustrando a quantidade de gêneros identificados em comum entre as combinações COMB201, COMB204 e COMB207.

Tabela 4.2: Tabela de gêneros em comum para as combinações mostradas na Figura 4.10.

Somente COMB201	*Acetobacter*, *Acidithiobacillus*, *Acidovorax*, *Acinetobacter*, *Actinobacillus*, *Aeromonas*, *Agrobacterium*, *Amycolatopsis*, *Azotobacter*, *Brucella*, *Burkholderia*, *Campylobacter*, *Caulobacter*, *Chlamydia*, *Chlamydophila*, *Chlorobium*, *Clostridium*, *Cronobacter*, *Dehalococcoides*, *Desulfotomaculum*, *Desulfovibrio*, *Dickeya*, *Edwardsiella*, *Enterobacter*, *Erwinia*, *Eubacterium*, *Fibrobacter*, *Flavobacterium*, *Geobacillus* , *Geobacter*, *Gluconobacter*, *Gordonia*, *Helicobacter*, *Hydrogenobacter*, *Klebsiella*, *Lactococcus*, *Leptospira*, *Leptospirillum*, *Mannheimia*, *Marinobacter*, *Meiothermus*, *Myxococcus*, *Neorickettsia*, *Pasteurella*, *Pectobacterium*, *Photorhabdus*, *Polynucleobacter*, *Proteus*, *Pseudomonas*, *Psychrobacter*, *Riemerella*, *Rothia*, *Shewanella*, *Shigella*, *Staphylococcus*, *Streptomyces*, *Thermotoga*, *Xylella*, *Yersinia*
Somente COMB204	*Acidaminococcus*, *Arcobacter*, *Cellulomonas*, *Dictyoglomus*, *Novosphingobium*, *Salinibacter*, *Sulfurimonas*, *Taylorella*
Somente COMB207	*Alteromonas, Nitrosococcus, Nitrosomonas*
COMB201 & COMB204	*Actinoplanes*, *Alcanivorax*, *Alkaliphilus*, *Bartonella*, *Buchnera*, *Enterococcus*, *Escherichia*, *Fervidobacterium*, *Frankia*, *Haemophilus*, *Mesoplasma*, *Paracoccus*, *Pedobacter*, *Planctomyces*, *Porphyromonas*,*Rickettsia*
COMB201 & COMB207	*Marinomonas*, *Octadecabacter*,*Pantoea*
COMB201 & COMB204 & COMB207	*Anaeromyxobacter*,*Bdellovibrio*

4.3.1 Análise estatística

Foram selecionados os gêneros *Escherichia*, *Salmonella* e *Vibrio* para observar a sensibilidade, especificidade, precisão e média harmônica das COMB201 (Figura 4.11) e COMB204 (Figura 4.12). Esses gêneros fazem parte da classe Gammaproteobacteria, das Proteobactérias e são bem caracterizados por serem causadores de gastroenterites em humanos (MADIGAN et al., 2010).

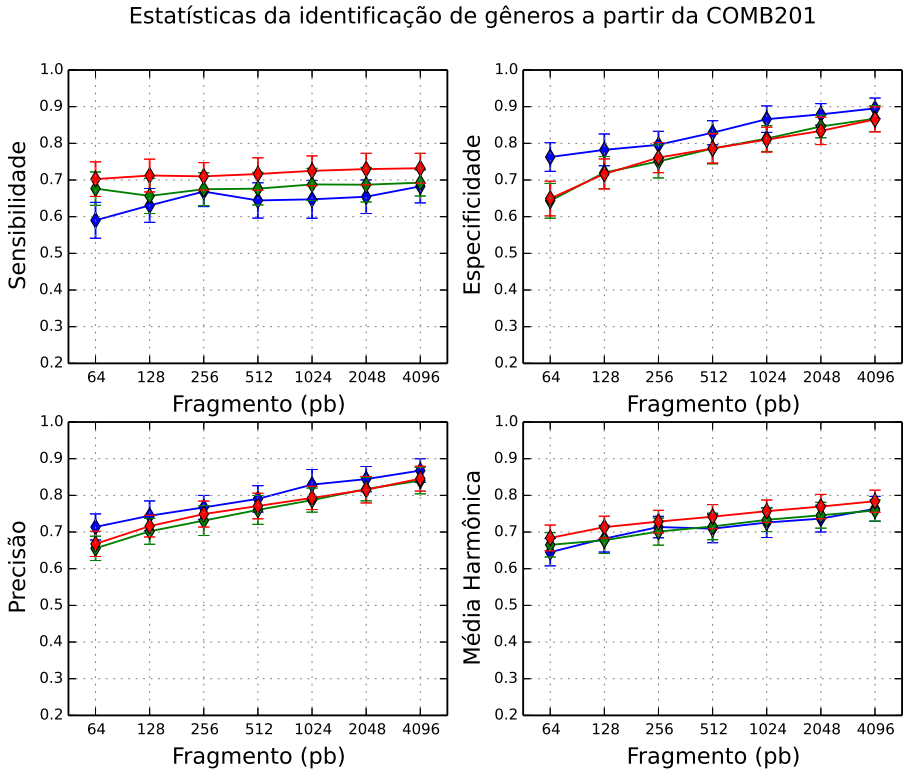


Figura 4.11: Gráfico de sensibilidade, especificidade, precisão e média harmônica para análises de gêneros utilizando COMB201. O gênero *Vibrio* está representado em azul, o *Salmonella* em verde e o *Escherichia* em vermelho.

Podemos observar para a COMB201 (Figura 4.11) que, embora haja pequenas variações entre os gêneros, os valores de sensibilidade são constantes para todos os tamanhos de sequência. O gênero *Vibrio* apresentou uma variação no valor de sensibilidade, com 59% para fragmentos de 64 pb, aumentando para 68% para os de 256 pb, diminuindo para 64% para as sequências de 512 pb, e voltando a subir gradativamente até atingir 69% para os segmentos de 4096 pb. *Escherichia* manteve os valores de sensibilidade constante em 70% para todos tamanhos de sequência e *Salmonella* em 68%.

A precisão apresentou valores mais baixos para fragmentos de 64 pb, sendo ele de 68% a 70%. Conforme o tamanho dos fragmentos aumenta, observa-se que a precisão varia entre 84%

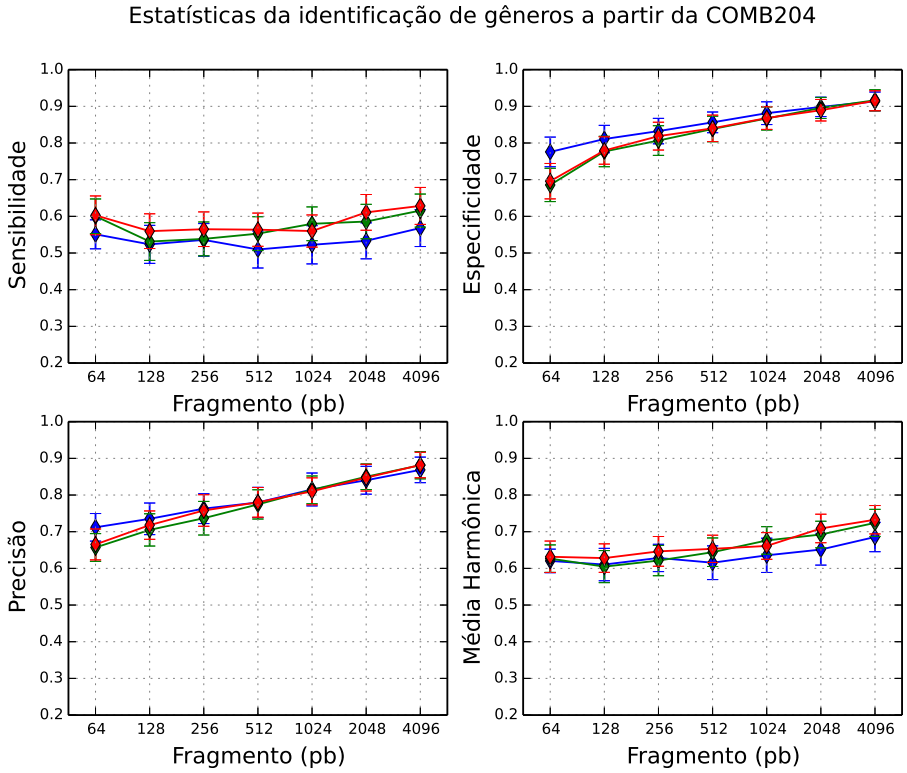


Figura 4.12: Gráfico de sensibilidade, especificidade, precisão e média harmônica para análises de gêneros utilizando COMB204. O gênero *Vibrio* está representado em azul, o *Salmonella* em verde e o *Escherichia* em vermelho.

e 88%.

A média harmônica entre as duas medidas citadas acima demonstrou que, em um conjunto de 100 fragmentos de DNA, a COMB201 pode identificar corretamente de 66 a 69 sequências de 64 pb dos gêneros analisados, e 79 a 89 para segmentos de 4096pb.

A especificidade para *Salmonella* e *Escherichia* aumenta de 64%, para sequências de 64 pb até 89% para as de tamanho 4096 pb, enquanto *Vibrio* apresenta valores de 78% a 90%.

Para a COMB204 (Figura 4.12), os valores de sensibilidade estão em um intervalo de 55% a 60% para segmentos de 64 pb e 58% a 62% para os de tamanho 4096 pb. Os valores de precisão aumentam com o crescimento do tamanho da sequência, passando, em média, de 69% para 89%. A média harmônica entre essas medidas mostra que 62 fragmentos de DNA (tamanho 64 pb) entre 100 dos gêneros *Escherichia*, *Vibrio* e *Salmonella* são atribuídos corretamente a eles.

A especificidade apresentou um aumento entre as sequências de 64 a 4096 pb, passando de 75% para os segmentos de tamanho 64 pb até 91% para os mais longos, de 4096 pb.

Dentre os gêneros avaliados, os que mais tiveram dificuldade de serem identificados foram

Bifidobacterium, *Pseudomonas* e *Bacillus*.

Bifidobacterium (filo Actinobacteria), por ser um grupo há muito tempo sendo cultivado *in vitro*, fora de seu habitat natural - o intestino humano - perdeu muitas características adaptativas que poderiam ser importantes para a sua identificação (LEE; O'SULLIVAN, 2010; LUKJANCENKO; USSERY; WASSENAAR, 2012), isso pode ter sido um fator que dificultou a classificação a partir da busca nos padrões das sequências, pois pode tornar sua sequência mais similar à de outros microrganismos, principalmente outros probióticos que também podem vir a ser cultivados em meios similares.

Os gêneros *Bacillus* (filo Firmicutes) e *Pseudomonas* (filo Proteobacteria, classe Gammaproteobacteria) possuem o genoma extremamente heterogêneo, o que pode ter dificultado na sua identificação (ASH et al., 1991; XU; COTE, 2003; HILARIO; BUCKLEY; YOUNG, 2004),

4.4 Análise de desempenho para espécies

Foram utilizados os resultados dos grupos de teste para avaliar a quantidade de sequências teste que cada combinação pôde identificar corretamente. Foram consideradas como identificadas cepas cuja combinação atingiu valores de todas medidas estatísticas (sensibilidade, especificidade, precisão e média harmônica) acima de 60% (Figura 4.13).

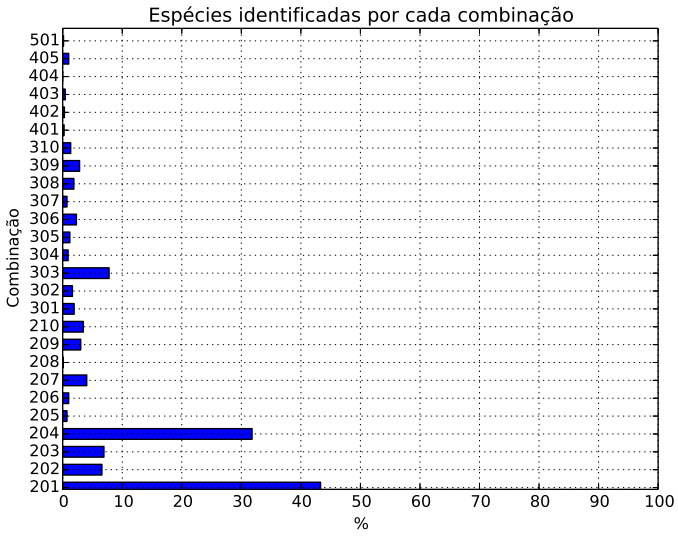


Figura 4.13: Gráfico ilustrando a quantidade de espécies que cada combinação identifica acima de 60% para todas as medidas estatísticas.

As combinações que mais identificam espécies foram a COMB201, que inclui conteúdo GC e abundância de dinucleotídeos, que identificaram mais de 40% das espécies, seguido da

COMB204, composta por conteúdo GC e entropia de tetrapletes.

4.4.1 Análise estatística

Os valores para sensibilidade, precisão e média harmônica apresentados pela COMB204 (Figura 4.14) foram os melhores comparados à outras combinações. Os valores de sensibilidade variam de 79%, para fragmentos de 64 pb, a 91% para segmentos de tamanho 4096. A precisão aumenta conforme o tamanho dos fragmentos, no intervalo de 70% a 89%. A média harmônica oscila entre 50% e 59%. Os valores de especificidade foram de 61% para fragmentos entre 64 e 256 pb, aumentando gradativamente para 69% para fragmentos de 4096%, conforme o aumento do tamanho dos fragmentos.

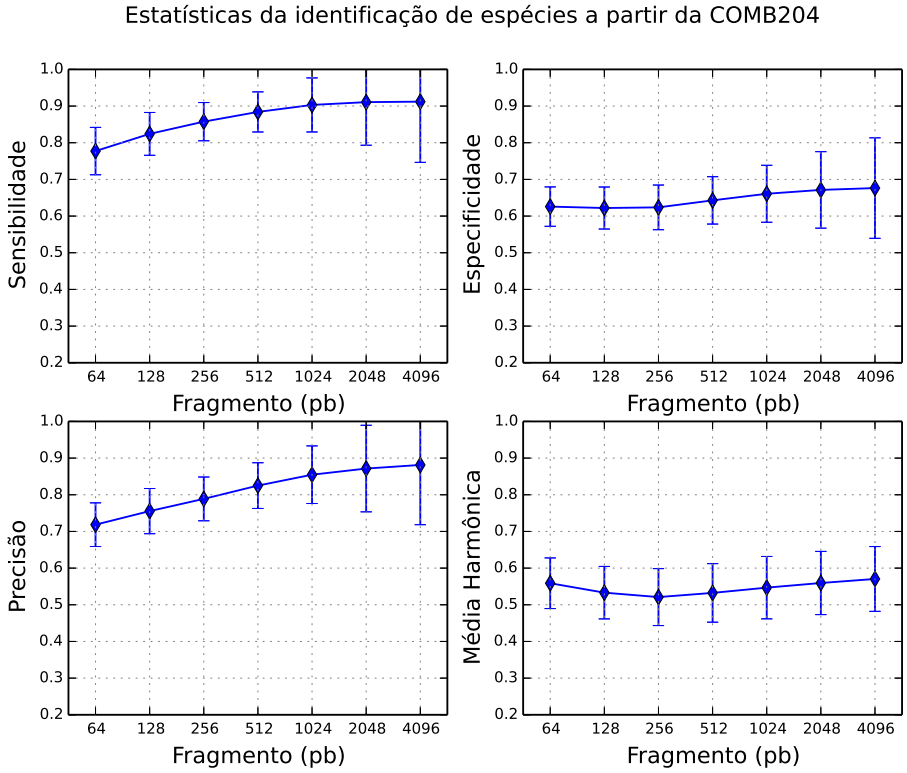


Figura 4.14: Gráfico de sensibilidade, especificidade, precisão e média harmônica para COMB204 para espécies, combinação de conteúdo GC e entropia de tetrapletes.

Testes para avaliar a eficiência das combinações para espécies distintas também foram realizadas. A espécie que apresentou maior valor de sensibilidade e média harmônica para a COMB204 (Figura 4.15) foi *Salmonella enterica*, para 512 pb, com 67% e 72%, respectivamente. Os valores de especificidade e precisão mais elevados foram de *Escherichia coli*, para fragmentos de 2048 pb, sendo eles 95% e 90%, respectivamente.

E. coli e *Pseudomonas aeruginosa* apresentaram valores de sensibilidade entre 49% e 65%.

Os valores de especificidade de *P. aeruginosa* e *S. enterica* ficaram entre 71% e 95%. As suas precisões variaram entre 70% para os tamanhos menores até 90% para 2048 pb. A média harmônica variou entre 60% e 71% para essas espécies.

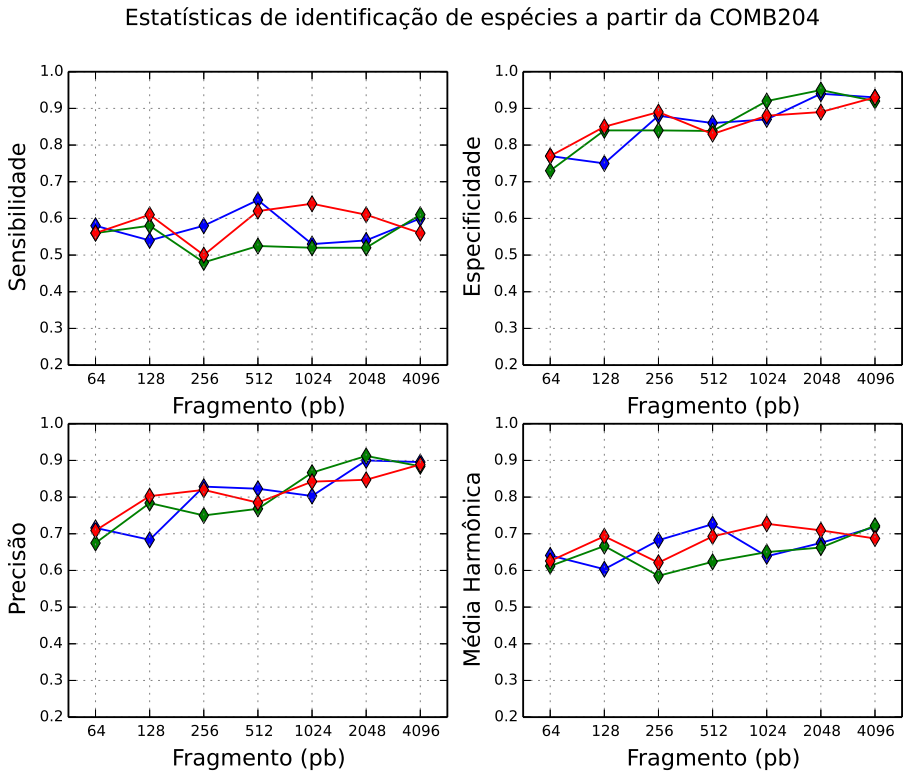


Figura 4.15: Gráfico de sensibilidade, especificidade, precisão e média harmônica para espécies isoladas identificadas por COMB204, combinação de conteúdo GC com entropia de tetraplets. *Salmonella enterica* está representada em azul, *Escherichia coli* em verde e *Pseudomonas aeruginosa* em vermelho.

Os resultados obtidos indicam que as combinações podem ser melhores para algumas espécies do que para outras. Fatores como a classificação taxonômica das espécies e a sua heterogeneidade genética também podem influenciar na sua identificação. Por exemplo, os genomas do gênero *Pseudomonas* são bastante heterogêneos (HILARIO; BUCKLEY; YOUNG, 2004) , além de apresentarem alta taxa de recombinação intragenômica causada por transferência lateral, o que pode vir a interferir nos padrões característicos do táxon (BODILIS et al., 2011) dificultando sua identificação. O genoma do gênero *Salmonella* pode assemelhar-se muito ao de outros como *Escherichia*, podendo ser um fator que torna a classificação confusa (MADIGAN et al., 2010). A *Escherichia coli*, por sua vez, é um micro-organismo muito bem estudado, por ser utilizado como modelo em pesquisas e por sua importância médica (MADIGAN et al., 2010), sendo bem caracterizado e, assim, melhor classificado.

4.5 Prova de conceito

O metagenoma da areia da praia do Pacífico foi classificado por famílias pelas combinações de medidas da tabela 3.1 e comparado com os resultados obtidos pelo *software* MEGAN para avaliar sua eficiência.

4.5.1 Análise pelo MEGAN

O alinhamento do metagenoma *Pacific Beach Sand* pelo BlastX gerou um arquivo .xml de 3,4 GB, que foi utilizado como entrada no MEGAN. O programa reconheceu 325 sequências das 4981 presentes no metagenoma. Essa análise obteve 5% (16 sequências) classificadas como não pertencendo ao domínio das bactérias (classficadas como Archaea, Eukarya ou vírus); 27,3% (89 sequências de 325) que não puderam ser classificadas (sequências cuja classificação não chegou ao nível de família ou que não parearam com outras na análise do BlastX) e 220 sequências, totalizando 67,7%, classificadas como pertencendo a alguma família de bactérias (Figura 4.16).

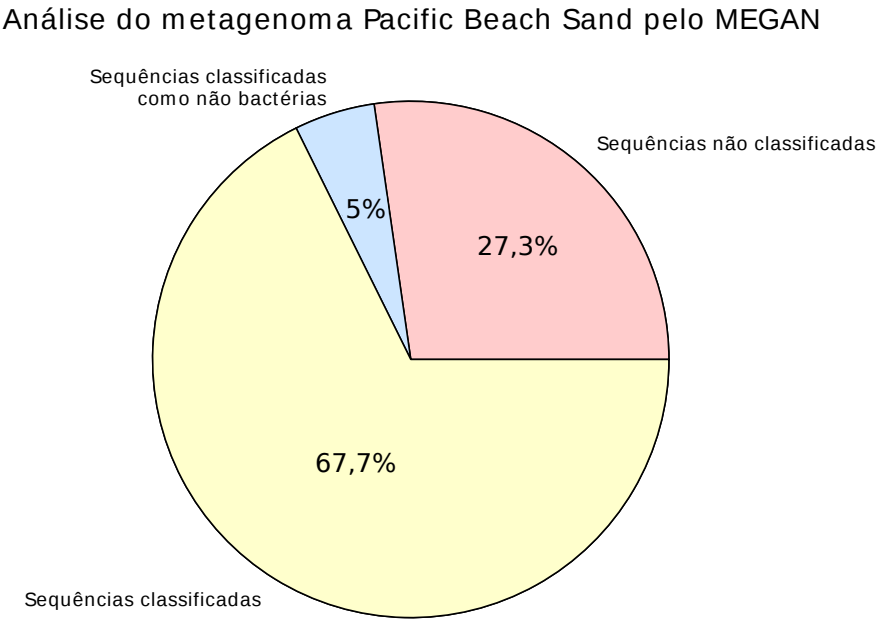


Figura 4.16: Análise do metagenoma *Pacific Beach Sand* pelo *software* MEGAN.

4.5.2 Análise pelas combinações de medidas

A partir dos resultados obtidos pelas combinações de medidas, elas foram divididas em 4 grupos que apresentaram o mesmo resultado entre elas:

- **Grupo 1:** as que não possuem entropia (COMB201);
- **Grupo 2:** as que possuem entropia de dipletes como última medida (COMB202, COMB205 e COMB301);
- **Grupo 3:** as que possuem entropia de tripletes como última medida (COMB203, COMB206, COMB208, COMB302, COMB304, COMB307 e COMB401) ;
- **Grupo 4:** as que possuem entropia de tetrapletes como última medida da combinação (COMB204, COMB207, COMB209, COMB210, COMB303, COMB305, COMB306, COMB308, COMB309, COMB310, COMB402, COMB403, COMB404, COMB405, COMB501).

Como ilustrado na figura 3.10, cada combinação retornou para cada sequência do metagenoma um vetor de famílias a quem ela pode pertencer. A quantidade de famílias presentes nos vetores variou entre as sequências e entre os grupos de combinações. As sequências que retornaram 0 famílias foram consideradas como *não classificadas* pela combinação. A variação da quantidade de famílias retornadas por cada grupo para as sequências classificadas está ilustrado na Figura 4.17.

A quantidade de famílias reconhecidas para cada sequência classificada foram de 25 a 57 para o grupo 1, de 10 a 188 para o grupo 2, de 31 a 187 para o grupo 3 e de 7 a 194 o grupo 4.

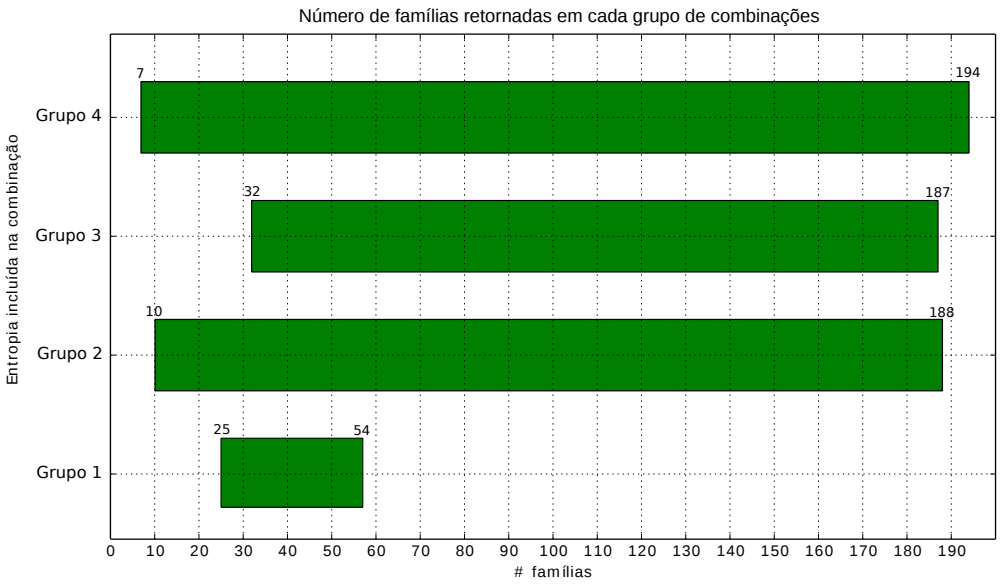


Figura 4.17: Quantidade de famílias reconhecidas para cada sequência por cada grupo de combinações. Foram selecionadas pelas combinações dentro de um grupo de 208 famílias.

Observamos quantas das 325 sequências reconhecidas pelo MEGAN foram corretamente identificadas pelos grupos de combinações. Foi considerada como a "família correta" para cada sequência aquela atribuída pelo MEGAN. Portanto, as sequências consideradas como *corretamente identificadas* foram aquelas cuja família considerada pelo MEGAN estava presente no vetor retornado pela combinação.

Foram observadas também quantas sequências não foram classificadas. Consideramos como *não classificadas* as sequências cujas combinações não retornaram nenhuma possibilidade de família a quem possam pertencer. Comparamos quantas sequências não identificadas pelos grupos também não foram classificadas pelo MEGAN. A partir dessa comparação é possível inferir a parcela de sequências que não foram identificadas por peculiaridades da nossa metodologia, no caso daquelas não identificadas somente pelas combinações, e a de sequências que não foram identificadas por apresentar características próprias que dificultem a análise independente da metodologia, no caso das sequências não classificadas em comum com o MEGAN.

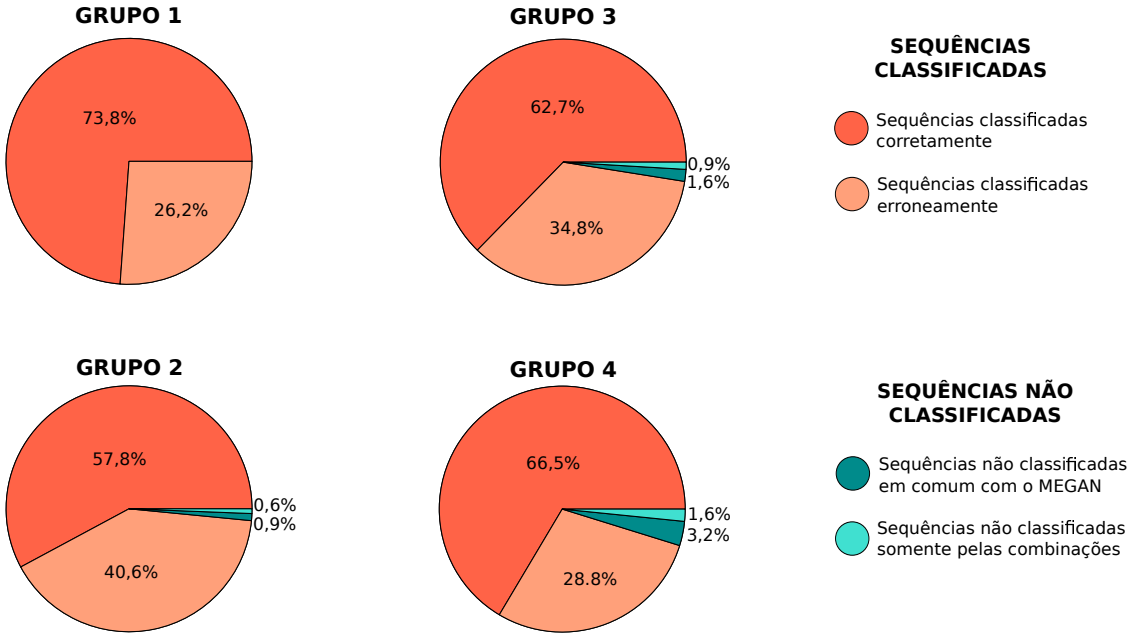


Figura 4.18: Gráficos de pizza apresentando a porcentagem de sequências classificadas corretamente, classificadas erroneamente, não classificadas em comum com o MEGAN e não classificadas somente pelas combinações para os 4 grupos de combinações, num total de 325 sequências.

Podemos observar que a combinação que não apresentava entropia em sua composição apresentou o maior número de sequências classificadas, bem como a maior porcentagem de identificações corretas comparado às outras categorias. Dentre os grupos de combinações com entropia, a que apresentou maior taxa de sequências não classificadas foi o grupo 4 - com 4,8%, contra 1,5% e 2,5% dos grupos 2 e 3. Embora tenha mais sequências sem famílias atribuídas, o grupo 4 foi o que apresentou maior índice de sequências corretamente identificadas

dentre as classificadas: 66,5%, enquanto o grupo 2 reconheceu a família correta de 57,8% e o grupo 3, 62,7%. Antagonicamente, o grupo 2 apresentou a menor proporção de sequências não identificadas (1,5%) e também de sequências atribuídas às famílias corretas (57,8%).

Esses resultados podem indicar um maior potencial da entropia de tetrapletes dentre as outras na identificação das famílias de sequências presentes em metagenomas. Entretanto, a combinação sem entropias detectou famílias em todas as sequências, com um maior índice de acerto, e retornando um número mais restrito de famílias selecionadas (como observado na Figura 4.17).

4.6 Comparação com outras metodologias

Comparamos os valores de sensibilidade e especificidade para a identificação de gêneros e famílias da COMB201 com os valores das ferramentas composicionais RAIphy, PhyloPythia e TACOA que foram apresentados na Figura 4.4.

A partir da Figura 4.19, observamos que os valores de sensibilidade obtidos na identificação taxonômica de fragmentos de 1 kpb são superiores aos valores das ferramentas TACOA e PyloPythia. Para famílias, os valores de sensibilidade e especificidade encontram-se próximos aos valores de RAIphy. Para gêneros, os valores da COMB201 ultrapassam os de RAIphy. Esses

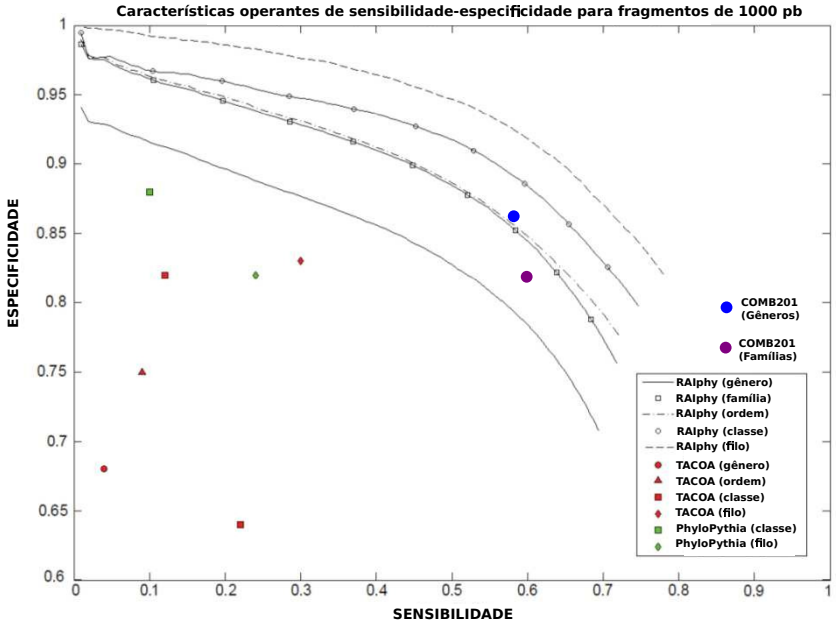


Figura 4.19: Valores de sensibilidade e especificidade para classificação taxonômica de fragmentos de 1 kpb das ferramentas RAIphy, TACOA e PhyloPythia em comparação com os valores da COMB201. Modificado de (NALBANTOGLU et al., 2011).

resultados indicam que a COMB201 apresenta potencial similar ou maior do que as ferramentas

apresentadas para a identificação de famílias e gêneros de sequências de 1000 pb.

5 *Conclusões*

A partir das análises de medidas de padrão de organização de sequências, foi possível observar dentre as assinaturas genômicas que o conteúdo GC foi o que apresentou a melhor eficiência, seus resultados foram acima de 70%, nos diversos tamanhos de fragmentos analisados. A abundância de dinucleotídeos apresentou melhores resultados para especificidade em fragmentos maiores e, dessa maneira foi mantido nas análises de combinações de parâmetros.

A entropia aplicada em dipletes, tripletes e tetrapletes apresentou resultados similares entre elas. Valores de especificidade, precisão e média harmônica foram superiores a 50%, o que demonstra que a entropia pode ser um potencial parâmetro a ser utilizado na classificação de fragmentos. No entanto, valores de sensibilidade indicam que essa medida não pode ser utilizada isoladamente para identificação das famílias de fragmentos de DNA de bactérias.

Combinações de assinaturas genômicas e entropias foram realizadas almejando a identificação das bactérias nos táxons família, gênero e espécie. Dentre as combinações, a que apresentou conteúdo GC associada à abundância de dinucleotídeos foi a mais promissora para família, gênero e espécies.

Na análise do metagenoma *Pacific Beach Sand* pelas combinações em comparação ao *software* MEGAN, a COMB201 também apresentou os resultados mais promissores para famílias, corroborando com os outros dados: selecionou mais precisamente as famílias de cada sequência, retornando um número menor de opções dentre as quais a família correta estava presente numa taxa maior do que as combinações com entropias.

Com os resultados observados pode-se inferir que a utilização das assinaturas conteúdo GC com abundância de dinucleotídeos é uma abordagem promissora para identificação de fragmentos. As entropias apresentaram altos valores de especificidade, precisão e média harmônica quando utilizadas isoladamente e em fragmentos de maior tamanho, o que indica um potencial na classificação de sequências maiores, entretanto necessita de mais estudos para que auxilie na identificação quando adicionada nas combinações que incluem conteúdo GC e abundância de dinucleotídeos.

A partir dessa metodologia, obtemos um vetor de possíveis táxons a quem as sequências podem pertencer. Uma perspectiva para esse trabalho, é utilizar os resultados obtidos pelo método como uma pré-filtragem de famílias, gêneros e espécies, posteriormente aplicando ferramentas como alinhamento de sequências para uma classificação mais específica. Isso diminuiria consideravelmente o tempo da análise taxonômica, pois ao invés de utilizar todos genomas dos bancos de dados como referência para o alinhamento, seriam utilizados somente aqueles indicados pela nossa ferramenta.

Referências Bibliográficas

ACLAND, A. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, v. 41, n. Database issue, p. D8–D20, Jan 2013.

ALIMETRICS. *DNA Sequence Analysis - The only species-specific approach for novel bacteria*. http://www.alimetrics.net/en/index.php/dna-sequence-analysis, 2014. Acessado Jan 2015. Disponível em: <http://www.alimetrics.net/en/index.php/dna-sequence-analysis>.

ALTSCHUL, S. F. et al. Basic local alignment search tool. *Journal of molecular biology*, Elsevier, v. 215, n. 3, p. 403–410, 1990.

AMORIM, D. S. *Fundamentos de Sistemática Filogenética*. Editora Holos, 2002.

ARUMUGAM, M. et al. Smashcommunity: a metagenomic annotation and analysis tool. *Bioinformatics*, v. 26, n. 23, p. 2977–2978, Dec 2010.

ASH, C. et al. Phylogenetic heterogeneity of the genus bacillus revealed by comparative analysis of small-subunit-ribosomal rna sequences. *Letters in Applied Microbiology*, v. 13, p. 202–206, 1991.

BEMAN, J. M.; POPP, B. N.; FRANCIS, C. A. Molecular and biogeochemical evidence for ammonia oxidation by marine crenarchaeota in the gulf of california. *ISME J*, v. 2, n. 4, p. 429–441, Apr 2008. Disponível em: <http://dx.doi.org/10.1038/ismej.2007.118>.

BENSON, D. A. et al. Genbank. *Nucleic acids research*, Oxford Univ Press, p. gks1195, 2012.

BODILIS, J. et al. A long-branch attraction artifact reveals an adaptive radiation in pseudomonas. *Mol Biol Evol*, v. 28, n. 10, p. 2723–2726, Oct 2011. Disponível em: <http://dx.doi.org/10.1093/molbev/msr099>.

BRADY, A.; SALZBERG, S. L. Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nature methods*, Nature Publishing Group, v. 6, n. 9, p. 673–676, 2009.

CAMERA. *Pacific Beach Sand Metagenome - CAMERA Portal*. 2014. Disponível em: <http://camera.crbs.ucsd.edu/projects/>.

CAMPBELL, A.; MRAZEK, J.; KARLIN, S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proceedings of the National Academy of Sciences*, v. 96, p. 9184–9189, 1999.

CANGANELLA, F.; WIEGEL, J. Extremophiles: from abyssal to terrestrial ecosystems and possibly beyond. *Naturwissenschaften*, v. 98, n. 4, p. 253–279, Apr 2011.

CHEN, Y. E.; TSAO, H. The skin microbiome: Current perspectives and future challenges. *J Am Acad Dermatol*, v. 69, p. 143–155, 2013.

CLARRIDGE, J. E. Impact of 16s rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev*, v. 17, n. 4, p. 840–62, table of contents, Oct 2004. Disponível em: <http://dx.doi.org/10.1128/CMR.17.4.840-862.2004>.

COUNCIL, N. R. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. The National Academies Press, 2007. ISBN 978-0-309-10676-4. Disponível em: <http://www.nap.edu/catalog/11902/the-new-science-of-metagenomics-revealing-the-secrets-of-our>.

DESAI, N. et al. From genomics to metagenomics. *Curr Opin Biotechnol*, v. 23, n. 1, p. 72–76, Feb 2012. Disponível em: <http://dx.doi.org/10.1016/j.copbio.2011.12.017>.

DEVARAJ, S.; HEMARAJATA, P.; VERSALOVIC, J. The human gut microbiome and body metabolism: implications for obesity and diabetes. *Clin Chem*, v. 59, p. 617–628, 2013.

DIAZ, N. N. et al. Tacoa–taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC bioinformatics*, BioMed Central Ltd, v. 10, n. 1, p. 56, 2009.

DRANCOURT, M.; BERGER, P.; RAOULT, D. Systematic 16s rRNA gene sequencing of atypical clinical isolates identified 27 new bacterial species associated with humans. *J Clin Microbiol*, v. 42, n. 5, p. 2197–2202, May 2004.

DUTTA, C.; PAUL, S. Microbial lifestyle and genome signatures. *Curr Genomics*, v. 13, n. 2, p. 153–162, Apr 2012.

ESPOSITO, A.; KIRSCHBERG, M. How many 16s-based studies should be included in a metagenomic conference? It may be a matter of etymology. *FEMS Microbiology Letters*, v. 351, p. 145–146, 2014.

FOERSTNER, K. U. et al. Environments shape the nucleotide composition of genomes. *EMBO Rep*, v. 6, n. 12, p. 1208–1213, Dec 2005.

GARCIA, J. A. L. et al. Ecophysiological significance of scale-dependent patterns in prokaryotic genomes unveiled by a combination of statistic and genometric analyses. *Genomics*, v. 91, n. 6, p. 538–543, Jun 2008. Disponível em: <http://dx.doi.org/10.1016/j.ygeno.2008.03.001>.

GERHARDT, G. J. L. et al. Triplet entropy analysis of hemagglutinin and neuraminidase sequences measures influenza virus phylodynamics. *Gene*, v. 528, n. 2, p. 277–281, Oct 2013. Disponível em: <http://dx.doi.org/10.1016/j.gene.2013.06.060>.

GERHARDT, N. L. G. J.; CORSO, G. Network clustering coefficient approach to DNA sequence analysis. *Chaos, Solitons and Fractals*, v. 28, p. 1037–1045, 2006.

GOODFELLOW, M. Microbial systematics: Background and uses. *Applied Microbial Systematics*, p. 1–18, 2000.

GREENBAUM, B. D. et al. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathog*, v. 4, n. 6, p. e1000079, Jun 2008. Disponível em: <http://dx.doi.org/10.1371/journal.ppat.1000079>.

HANDELSMAN, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*, v. 68, n. 4, p. 669–685, Dec 2004. Disponível em: <http://dx.doi.org/10.1128/MMBR.68.4.669-685.2004>.

HANKERSON, D.; JOHNSON, P. D.; HARRIS, G. A. *Introduction to Information Theory and Data Compression*. 1st. ed. Boca Raton, FL, USA: CRC Press, Inc., 1998. ISBN 0849339855.

HILARIO, E.; BUCKLEY, T. R.; YOUNG, J. M. Improved resolution on the phylogenetic relationships among pseudomonas by the combined analysis of atp d, car a, rec a and 16s rRNA. *Antonie Van Leeuwenhoek*, v. 86, n. 1, p. 51–64, Jul 2004. Disponível em: <http://dx.doi.org/10.1023/B:ANTO.0000024910.57117.16>.

HUNTER, C. et al. Metagenomic analysis: the challenge of the data bonanza. *Briefings in Bioinformatics*, v. 13, p. 743–746, 2011.

HUNTER, S. et al. Ebi metagenomics - a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res*, v. 42, n. Database issue, p. D600–D606, Jan 2014. Disponível em: <http://dx.doi.org/10.1093/nar/gkt961>.

HUSON, D. H. et al. Megan analysis of metagenomic data. *Genome Res*, v. 17, n. 3, p. 377–386, Mar 2007. Disponível em: <http://dx.doi.org/10.1101/gr.5969107>.

HUTCHISONIII, C. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research*, v. 35, n. 18, p. 6227–6237, 2007. Disponível em: <http://nar.oxfordjournals.org/content/35/18/6227.full.pdf>.

KANNAN, N. et al. Structural and functional diversity of the microbial kinome. *PLoS Biol*, v. 5, n. 3, p. e17, Mar 2007. Disponível em: <http://dx.doi.org/10.1371/journal.pbio.0050017>.

KARLIN, S. Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol*, v. 1, n. 5, p. 598–610, Oct 1998.

KARLIN, S.; BURGE, C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*, v. 11, n. 7, p. 283–290, Jul 1995.

KIM, M. et al. Analytical tools and databases for metagenomics in the next-generation sequencing era. *Genomics Inform*, v. 11, n. 3, p. 102–113, Sep 2013. Disponível em: <http://dx.doi.org/10.5808/GI.2013.11.3.102>.

LEE, J.-H.; O'SULLIVAN, D. J. Genomic insights into bifidobacteria. *Microbiol Mol Biol Rev*, v. 74, n. 3, p. 378–416, Sep 2010. Disponível em: <http://dx.doi.org/10.1128/MMBR.00004-10>.

LI, J.; SAYOOD, K. A genome signature based on markov modeling. *Conf Proc IEEE Eng Med Biol Soc*, v. 3, p. 2832–2835, 2005.

LIU, J. et al. Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. *Nucleic Acids Res*, v. 41, n. 1, p. e3, Jan 2013. Disponível em: <http://dx.doi.org/10.1093/nar/gks828>.

LUKJANCENKO, O.; USSERY, D. W.; WASSENAAR, T. M. Comparative genomics of bifidobacterium, lactobacillus and related probiotic genera. *Microb Ecol*, v. 63, n. 3, p. 651–673, Apr 2012. Disponível em: <http://dx.doi.org/10.1007/s00248-011-9948-y>.

MADIGAN, M. T. et al. *Brock Biology of Microorganisms (13th Edition)*. 13. ed. Benjamin Cummings, 2010. Hardcover. ISBN 032164963X. Disponível em: <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/032164963X>.

MANDE, S. S.; MOHAMMED, M. H.; GHOSH, T. S. Classification of metagenomic sequences: methods and challenges. *Brief Bioinform*, v. 13, n. 6, p. 669–681, Nov 2012. Disponível em: <http://dx.doi.org/10.1093/bib/bbs054>.

MARCO, D. *Metagenomics: Current Innovations and Future Trends*. Caister Academic Press, 2011. ISBN 9781904455875. Disponível em: <http://books.google.com.br/books?id=YGc1P0RiAOgC>.

MCHARDY, A. C. et al. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods*, v. 4, n. 1, p. 63–72, Jan 2007. Disponível em: <http://dx.doi.org/10.1038/nmeth976>.

MOHAMMED, M. H. et al. Indus-a composition-based approach for rapid and accurate taxonomic classification of metagenomic sequences. *BMC genomics*, BioMed Central Ltd, v. 12, n. Suppl 3, p. S4, 2011.

MOROZOVA, O.; MARRA, M. A. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, Elsevier, v. 92, n. 5, p. 255–264, 2008.

MOSIER, A. C.; FRANCIS, C. A. Determining the distribution of marine and coastal ammonia-oxidizing archaea and bacteria using a quantitative approach. *Methods Enzymol*, v. 486, p. 205–221, 2011. Disponível em: <http://dx.doi.org/10.1016/B978-0-12-381294-0.00009-2>.

NALBANTOGLU, O. U. et al. Raiphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC bioinformatics*, BioMed Central Ltd, v. 12, n. 1, p. 41, 2011.

NAVIAUX, R. K. et al. Sand DNA: a genetic library of life at the water's edge. *Marine ecology. Progress series*, Inter-Research, v. 301, p. 9–22, 2005.

NEXT GEN SEEK. *Comparing Price and Tech. Specs. of Illumina MiSeq, Ion Torrent PGM, 454 GS Junior, and PacBio RS*. http://nextgenseek.com/2012/08/comparing-price-and-tech-specs-of-illumina-miseq-ion-torrent-pgm-454-gs-junior-and-pacbio-rs/, Agosto 2012. Acessado 13 Nov 2014. Disponível em: <http://nextgenseek.com/2012/08/comparing-price-and-tech-specs-of-illumina-miseq-ion-torrent-pgm-454-gs-junior-and-pacbio-rs/>.

NICOL, G. W.; SCHLEPER, C. Ammonia-oxidising crenarchaeota: important players in the nitrogen cycle? *Trends Microbiol*, v. 14, n. 5, p. 207–212, May 2006. Disponível em: <http://dx.doi.org/10.1016/j.tim.2006.03.004>.

NIKOLAKI, S.; TSIAMIS, G. Microbial diversity in the era of omic technologies. *Biomed Res Int*, v. 2013, p. 958719, 2013. Disponível em: <http://dx.doi.org/10.1155/2013/958719>.

PASSEL, M. W. J. van et al. The reach of the genome signature in prokaryotes. *BMC Evol Biol*, v. 6, p. 84, 2006.

PATI, A. et al. Clams: a classifier for metagenomic sequences. *Standards in genomic sciences*, Genomic Standards Consortium, v. 5, n. 2, p. 248, 2011.

PETTERSSON, E.; LUNDEBERG, J.; AHMADIAN, A. Generations of sequencing technologies. *Genomics*, v. 93, n. 2, p. 105–111, Feb 2009.

PORETSKY, R. et al. Strengths and limitations of 16s rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PloS one*, Public Library of Science, v. 9, n. 4, p. e93827, 2014.

POWERS, D. M. W. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, v. 2, n. 1, p. 37–63, 2007.

RAES, J.; FOERSTNER, K. U.; BORK, P. Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol*, v. 10, n. 5, p. 490–498, Oct 2007.

RIDAURA, V. K. et al. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science*, v. 341, n. 6150, p. 1241214, Sep 2013. Disponível em: <http://dx.doi.org/10.1126/science.1241214>.

ROTHSCHILD, L. J.; MANCINELLI, R. L. Life in extreme environments. *Nature*, v. 409, n. 6823, p. 1092–1101, Feb 2001. Disponível em: <http://dx.doi.org/10.1038/35059215>.

SANTOS, L. dos; RYBARCZYK-FILHO; GERHARDT, G. J. L. Triplet entropy in H1N1 virus. *Tendências em Matemática Aplicada e Computacional*, v. 12, p. 253–261, 2011.

SCHOLZ, M. B.; LO, C.-C.; CHAIN, P. S. G. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr Opin Biotechnol*, v. 23, n. 1, p. 9–15, Feb 2012.

SCHUH, R. *Biological Systematics: Principles and Applications*. Cornell University Press, 2000. ISBN 9780801436758. Disponível em: <http://books.google.com.br/books?id=Hn3Pe6UgPGoC>.

SHANNON, C. E. A mathematical theory of communication. *Bell System Technical Journal*, v. 27, n. 3, p. 379–423, 1948.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. *Nature biotechnology*, Nature Publishing Group, v. 26, n. 10, p. 1135–1145, 2008.

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

TIWARI, S. et al. Prediction of probable genes by fourier analysis of genomic sequences. *Comput Appl Biosci*, v. 13, n. 3, p. 263–270, Jun 1997.

WANG, Y. et al. The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene*, v. 346, p. 173–185, Feb 2005. Disponível em: <http://dx.doi.org/10.1016/j.gene.2004.10.021>.

WOESE, C. R.; KANDLER, O.; WHEELIS, M. L. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proc Natl Acad Sci U S A*, v. 87, n. 12, p. 4576–4579, Jun 1990.

WOO, H. L. et al. Enzyme activities of aerobic lignocellulolytic bacteria isolated from wet tropical forest soils. *Syst Appl Microbiol*, v. 37, n. 1, p. 60–67, Feb 2014. Disponível em: <http://dx.doi.org/10.1016/j.syapm.2013.10.001>.

WOOD, D. E.; SALZBERG, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*, v. 15, n. 3, p. R46, 2014. Disponível em: <http://dx.doi.org/10.1186/gb-2014-15-3-r46>.

XU, D.; COTE, J.-C. Phylogenetic relationships between bacillus species and related genera inferred from comparison of 3' end 16s rRNA and 5' end 16s-23s its nucleotide sequences. *Int J Syst Evol Microbiol*, v. 53, n. Pt 3, p. 695–704, May 2003.

YAKOVCHUK, P.; PROTOZANOVA, E.; FRANK-KAMENETSKII, M. D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res*, v. 34, n. 2, p. 564–574, 2006.