RESSALVA

Atendendo solicitação do(a) autor(a), o texto completo deste trabalho será disponibilizado somente a partir de 15/02/2019.

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

José Rafael Pilan

Desenvolvimento de um método de classificação taxonômica de dados de metagenomas

Botucatu, Janeiro de 2017.

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

José Rafael Pilan

Desenvolvimento de um método de classificação taxonômica de dados de metagenomas

Monografia apresentada ao Instituto de Biociências, Campus de Botucatu, UNESP, em preenchimento dos requisitos para a obtenção do título de Mestre no Programa de Pós-Graduação em Biotecnologia.

Área de Concentração: Biotecnologia
Orientador: Prof. Dr. José Luiz Rybarczyk Filho
Co-Orientadora: Prof.ª Dr.ª Agnes Alessandra Sekijima Takeda

Botucatu, Janeiro de 2017.

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Pilan, José Rafael.
Desenvolvimento de um método de classificação taxonômica de dados de metagenomas / José Rafael Pilan. - Botucatu, 2017

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: José Luiz Rybarczyk Filho
Coorientador: Agnes Alessandra Sekijima Takeda
Capes: 90400003

1. Metagenômica. 2. Taxonomia numérica. 3. Metagenoma. 4. Micro-organismos. 5. Código genético.

Palavras-chave: Metagenômica; NGS; Taxonomia.

Agradecimentos

- Ao CNPq por utilizarmos os recursos computacionais referentes aos processos 458810/2013-4 e 473789/2013-2
- Ao Professor Dr. José Luiz Rybarczyk Filho pela orientação e auxilio;
- À Professora Dr.ª Agnes Alessandra Sekijima Takeda pela co-orientação e dedicação;
- Ao Professor Dr. César Martins por disponibilizar o acesso ao *cluster* para este trabalho;
- À toda minha família, principalmente meu pai José Carlos, minha mãe Maria Aparecida e meu irmão César pelo apoio e incentivo que foram fundamentais para a conclusão desse trabalho;
- Aos meus amigos Alex Augusto Biazotti, André Luiz Molan e Carlos Alberto de Oliveira Biagi Junior por toda a ajuda e motivação;
- À minha namorada Ingrid Vidotto de Oliveira por toda paciência e companheirismo;
- Aos demais colegas do departamento de Física e Biofísica e do Instituto de Biotecnologia, professores e funcionários que de alguma forma tenham contribuído para a realização deste trabalho.

Resumo

Na análise de dados metagenômicos temos duas perguntas básicas que podemos fazer: “Quem são?” e “O que estão fazendo?” os microorganismos de uma determinada amostra. Para responder a primeira pergunta utiliza-se a análise taxonômica de microorganismos. Existem diversos *software* que utilizam diferentes metodologias para atingir essa finalidade. Esses métodos são divididos em duas categorias principais: composicional e alinhamento por similaridade. O que diferencia os métodos são principalmente o tempo para ser realizada a análise, poder computacional e eficiência na identificação dos *reads*. Nesse trabalho propomos um novo método por composição que utiliza cinco assinaturas genômicas e suas combinações para identificação dos *reads*: concentração de GC (Guanina/Citocina), entropia de dipletes, entropia de tripletes, entropia de tetrapletes e abundância total de dinucleotídeos. Utilizamos um conjunto de dados referente a 3055 genomas completos de bactérias provenientes do NCBI (*National Center for Biotechnology Information*)que foram fragmentados em dois grupos: teste e controle. Os grupos foram fragmentados em tamanhos de 50-1000pb com partições de tamanho 50pb, buscando se aproximar dos tamanhos de *reads* normalmente gerados pelos equipamentos de sequenciamento de nova geração. O desempenho da metologia foi avaliado por medidas de sensibilidade, especificidade, precisão e média harmônica em comparação aos resultados do grupo teste com o grupo controle. Dentre as combinações analisadas, a concentração de GC apresentou melhor desempenho na identificação dos organismos. Para a comparação do método com os *software* já existentes, prospectamos 233 amostras no EBI (*European Bioinformatics Institute*) do projeto *“A human gut microbial gene catalog established by deep metagenomic sequencing”*, realizamos a análise das amostras com os programas Phymm, Phymmbl e Raiphy e comparamos com os resultados de nossa metodologia. Na comparação, a medida de concentração de GC em conjunto com a medida entropia de dipletes mostrou-se eficiente em comparação as demais atingindo em média 89,5% de identificação dos *reads*.

Abstract

In analyzing metagenomic data we have two basic questions that we can ask: “Who are they?” and “What are doing the microorganisms of a given sample?” . To answer the first question we use the taxonomic analysis of microorganisms. There are several software that use different methodologies to achieve this purpose. These methods are divided into two main categories: compositional and alignment by similarity. What differentiates the methods are mainly the time to perform the analysis, computational power and efficiency in the identification of reads. In this work we propose a new compositional method that uses five genomic signatures and their combinations to identify reads: GC concentration, diplet entropy, triplet entropy, tetraplet entropy and total abundance of dinucleotides. We used a data set of 3055 complete bacterial genomes from the NCBI (National Center for Biotechnology Information) that were fragmented into two groups: test and control. The groups were fragmented in sizes of 50-1000bp with partitions of size 50bp, seeking to approximate the sizes of reads normally generated by the new generation sequencing equipment. The performance of the metology was evaluated by measures of sensitivity, specificity, precision and harmonic mean in comparison to the results of the test group with the control group. Among the combinations analyzed, the GC concentration presented better performance in the identification of organisms. For the comparison of the method with existing software, we prospected 233 samples in the EBI (European Bioinformatics Institute) of the project “A human gut microbial gene established by deep metagenomic sequencing”, we performed the analysis of the samples with the programs Phymm, Phymmbl and Raiphy and compared with the results of our methodology. In the comparison, the GC concentration measure in conjunction with the entropy measurement of diplets proved to be efficient in comparison to the others reaching a mean of 89.5% of the identification of the reads.

Lista de Figuras

1.1 Tecnologia de sequênciamento de Sanger. . . . . p. 2

1.2 Estratégias de imobilização de *templates* utilizada pela Applied Biosystems. . p. 3

1.3 Estratégia de amplificação por fase sólida usada pelos equipamentos da Ilumina. p. 4

1.4 Estratégia de fixação da polimerase no suporte utilizada pela Pacific Biosciences. . . . . p. 5

1.5 Gene 16S rRNA . . . . . p. 7

1.6 16S rRNA . . . . . p. 8

1.7 Etapas de obtenção de dados de metagenoma. . . . . p. 10

1.8 Mapa mundi mostrando a quantidade de programas para identificação taxonômica desenvolvida por cada país. . . . . p. 11

3.1 *Workflow* das etapas do material e métodos . . . . . p. 15

3.2 Exemplo do cálculo de concentração de GC . . . . . p. 17

3.3 Cálculos de entropia . . . . . p. 18

3.4 Cálculo de abundância total de dinucleotídeos . . . . . p. 19

3.5 Filtragem dos dados . . . . . p. 20

3.6 Fluxograma criação grupo controle . . . . . p. 21

3.7 Fluxograma criação grupo controle . . . . . p. 22

3.8 Fluxograma criação banco de dados do grupo controle . . . . . p. 23

3.9 Fluxograma criação grupo teste . . . . . p. 24

3.10 Matriz de Confusão . . . . . p. 27

3.11 Características operantes de sensibilidade e especificidade para classificação taxonômica de fragmentos de 1 kpb das ferramentas RAIphy, TACOA e PhyloPythia. . . . . p. 28

4.1 Relação de tamanho do fragmento com as combinações de medidas S2 ∩ S3 ∩ S4para identificação do gênero *Bartonella.* . . . . . p. 30

4.2 Relação de tamanho do fragmento com as combinações de medidas GC % e S2 para identificação da família *Acetobacteraceae*. . . . . p. 31

4.3 Relação de tamanho do fragmento utilizando a medida GC % para identificação da ordem *Flavobacteriales*. . . . . p. 32

4.4 Relação de tamanho do fragmento utilizando as combinações de medidas S2 e din para identificação da classe *Dehalococcoidia*. . . . . p. 33

4.5 *Boxplots* das medidas estatísticas para o *read* de tamanho 50 pb em comparação com todas as 29 combinações para o gênero *Amycolatopsis* . . . . . p. 35

4.6 *Boxplots* das medidas estatísticas para o *read* de tamanho 50 pb em comparação com todas as 29 combinações para o gênero *Halanaerobium*. . . . . p. 35

4.7 Identificação do gênero *Anaeromyxobacter* por tamanho de *read* para a medida 15-GC %. . . . . p. 37

4.8 Identificação do gênero *Anaeromyxobacter* por tamanho de *read* para a medida 13-S4. . . . . p. 38

4.9 Identificação do gênero *Sorangium* por tamanho de *read* para a medida 15-GC %. . . . . p. 39

4.10 Identificação do gênero *Sorangium* por tamanho de *read* para a medida 9-S3. p. 40

4.11 Combinações de medidas para identificação de *reads* de tamanho 50 pb para a família *Nocardiaceae* . . . . . p. 43

4.12 *Boxplots* das medidas estatísticas para o *read* de tamanho 50 pb em comparação com todas as 29 combinações para a família *Chlamydiaceae*. . . . . p. 44

4.13 Identificação da família *Microbacteriaceae* por tamanho de *read* para a medida 15-GC %. . . . . p. 45

4.14 Identificação da família *Microbacteriaceae* por tamanho de *read* para a medida 1-din. . . . . p. 46

4.15 *Boxplots* das medidas estatísticas para o *read* de tamanho 50 pb em comparação com todas as 29 combinações para a ordem *Brachyspirales* . . . . . p. 49

4.16 *Boxplots* das medidas estatísticas para o *read* de tamanho 50 pb em comparação com todas as 29 combinações para a ordem *Chlamydiales* . . . . . p. 50

4.17 Identificação da ordem *Frankiales* por tamanho de *read* para a medida 15-GC %. . . . . p. 51

4.18 Identificação da ordem *Frankiales* por tamanho de *read* para a medida 1-din. . p. 52

4.19 *Boxplots* das medidas estatísticas para o *read* de tamanho 50 pb em comparação com todas as 29 combinações para a classe *Chlamydiia* . . . . . p. 55

4.20 *Boxplots* das medidas estatísticas para o *read* de tamanho 50 pb em comparação com todas as 29 combinações para a classe *Deinococci* . . . . . p. 56

4.21 Percentual de gêneros encontrados por *read* nas amostras de metagenoma utilizando a medida GC % . . . . . p. 60

4.22 Percentual de gêneros encontrados por *read* nas amostras de metagenoma utilizando a medida S2 . . . . . p. 61

4.23 Percentual de gêneros encontrados por *read* nas amostras de metagenoma utilizando a combinação de medidas GC % e S2 . . . . . p. 62

4.24 Diagrama de Venn comparando a quantidade de organismos identificados no táxon de gênero por *software* em relação a combinação das medidas 17-GC % S2 para as amostras ERR011172, ERR011190, ERR011230, ERR011259, ERR011277, ERR011305, ERR011308 e ERR011323. . . . . p. 63

Lista de Tabelas

1.1 Tabela comparativa das principais tecnologias de sequenciamento . . . . . p. 6

3.1 Especificações de memória RAM e quantidades de núcleos dos computadores utilizados nas análises. . . . . p. 16

3.2 Tabela de identificadores das combinações de medidas . . . . . p. 25

4.1 Tabela com informações relacionadas a geração de dados durante a análise dos táxons. . . . . p. 29

4.2 Tabela de combinações para o táxon gênero com valores de especificidade, sensibilidade, precisão e média harmônica com valores iguais ou acima de 70 %. . . . . p. 34

4.3 Tabela de gêneros identificados pela metodologia para *reads* de tamanho 50-600 pb. . . . . p. 36

4.4 Tabela de medidas que identificaram organismos do gênero *Anaeromyxobacter* em relação ao tamanho do *read*. . . . . p. 38

4.5 Tabela de medidas que identificaram organismos do gênero *Sorangium* em relação ao tamanho do *read*. . . . . p. 39

4.6 Tabela de combinações e organismos identificados no táxon família para valores de especificidade, sensibilidade, precisão e média harmônica acima de 70 %. . . . . p. 41

4.7 Tabela de organismos identificados pela metodologia para *reads* de tamanho 50-600 pb no táxon família . . . . . p. 42

4.8 Tabela de medidas que identificaram organismos da família *Microbacteriaceae* em relação ao tamanho do *read*. . . . . p. 43

4.9 Tabela de combinações para o táxon ordem com valores de especificidade, sensibilidade, precisão e média harmônica com valores iguais ou acima de 70 %. . . . . p. 47

4.10 Tabela de organismos identificados pela metodologia para *reads* de tamanho 50-600 pb no táxon ordem . . . . . p. 48

4.11 Tabela de medidas que identificaram organismos da ordem *Frankiales* em relação ao tamanho do *read*. . . . . p. 49

4.12 Tabela de combinações para o táxon classe com valores de especificidade, sensibilidade, precisão e média harmônica com valores iguais ou acima de 70 %. . . . . p. 53

4.13 Tabela de organismos identificados pela metodologia para *reads* de tamanho 50-600 pb no táxon classe . . . . . p. 54

4.14 Tabela com o percentual de organismos identificados nas amostras de metagenomas por medida da metodologia . . . . . p. 59

A.1 Tabela comparativa das principais programas de análise . . . . . p. 71

A.2 Tabela comparativa das principais programas de análise (continuação) . . . . p. 72

B.1 Tabela com informações das amostras obtidas no EBI . . . . . p. 73

B.2 Tabela com informações das amostras obtidas no EBI (continuação) . . . . . p. 74

B.3 Tabela com informações das amostras obtidas no EBI (continuação) . . . . . p. 75

Sumário

Resumo . . . . . p. iv

Abstract . . . . . p. v

1 Introdução p. 1

1.1 Comunidades microbianas . . . . . p. 1

1.2 Tecnologias de sequenciamento . . . . . p. 1

1.2.1 Sanger (primeira-geração) . . . . . p. 1

1.2.2 Sequenciamento de nova geração . . . . . p. 3

1.3 Descoberta de microorganismos . . . . . p. 7

1.3.1 Gene 16S rRNA . . . . . p. 7

1.4 Metagenômica . . . . . p. 9

1.5 Análise taxônomica de dados metagenômicos . . . . . p. 9

1.6 Programas de análise taxonômica . . . . . p. 11

2 Objetivos p. 14

2.1 Objetivos Específicos . . . . . p. 14

3 Material e Métodos p. 15

3.1 *Workflow* . . . . . p. 15

3.2 Aquisição de dados . . . . . p. 16

3.2.1 Conjunto de dados para criação do banco de dados . . . . . p. 16

3.2.2 Conjunto de dados para comparação com outras ferramentas . . . . . p. 16

3.3 *Hardware* . . . . . p. 16

3.4 Medidas . . . . . p. 17
3.4.1 Concentração de GC (GC %) . . . . . p. 17
3.4.2 Entropia (S) . . . . . p. 17
3.4.3 Abundância total de dinucleotídeos (din) . . . . . p. 18
3.5 Filtragem dos dados para criação do banco de dados . . . . . p. 19
3.6 Grupo controle . . . . . p. 20
3.7 Grupo teste . . . . . p. 24
3.8 Medidas estatísticas . . . . . p. 25
3.8.1 Matriz de confusão . . . . . p. 25
3.8.2 Sensibilidade . . . . . p. 26
3.8.3 Especificidade . . . . . p. 26
3.8.4 Precisão . . . . . p. 26
3.8.5 Média harmônica entre sensibilidade e especificidade . . . . . p. 26
3.9 Comparação com outras metodologias . . . . . p. 27
3.10 Critério de Corte . . . . . p. 28

4 Resultados e Discussão p. 29
4.1 Processamento das informações . . . . . p. 29
4.2 Avaliação Global . . . . . p. 30
4.2.1 Táxon Gênero . . . . . p. 30
4.2.2 Táxon Família . . . . . p. 31
4.2.3 Táxon Ordem . . . . . p. 32
4.2.4 Táxon Classe . . . . . p. 33
4.3 Aplicação da metodologia no táxon gênero . . . . . p. 34
4.4 Aplicação da metodologia no táxon família . . . . . p. 40
4.5 Aplicação da metodologia no táxon ordem . . . . . p. 47
4.6 Aplicação da metodologia no táxon classe . . . . . p. 53

4.7 Comparação com programas de análise taxonômica. . . . . p. 57

4.8 Dados finais da metodologia . . . . . p. 64

5 Conclusões p. 65

Referências Bibliográficas p. 66

Apêndice A – Tabela comparativa das principais programas de análise p. 71

Apêndice B – Amostras de Metagenoma p. 73

1 Introdução

1.1 Comunidades microbianas

As comunidades microbianas têm sido estudadas por séculos. Partindo do preceito que tanto animais como plantas vivem associados com milhares de espécies de microrganismos diferentes (ROSENBERG et al., 2010), a teoria dos hologenomas considera que a soma da informação genética do indivíduo com a sua microbiota é um fator importante na evolução. Através de estudos bioquímicos e genéticos é possível o entendimento dos processos biológicos que essas comunidades estão envolvidas. A observação desses processos pode ajudar na compreensão de metabolismos mais complexos como os eucariotos (ZILBER-ROSENBERG; ROSENBERG, 2008; ROSENBERG; ZILBER-ROSENBERG, 2011; MADIGAN et al., 2010).

No início, os microrganismos foram classificados utilizando como base parâmetros como características morfológicas ou a seleção de perfis bioquímicos (BULLOCK; ASLANZADEH, 2013). Segundo pesquisas (SCHLOSS; HANDELSMAN, 2005; FIERER; JACKSON, 2006; MOREIRA, 2015), cerca de 99 % da microbiota existente na biosfera não é cultivável em laboratório utilizando técnicas de isolamento. Devido a falta de conhecimento das necessidades nutricionais destes organismos ou mesmo da dependência entre eles para o seu crescimento (LEWIS et al., 2010). Esses fatores corroboram com a limitação do entendimento da fisiologia, genética e comunidade desses microrganismos.

1.2 Tecnologias de sequenciamento

1.2.1 Sanger (primeira-geração)

Sanger e colaboradores (SANGER; NICKLEN; COULSON, 1977) criaram um método de sequenciamento que utiliza dideoxinucleotídeos como inibidores de terminação de cadeia de DNA, denominado “primeira-geração” (METZKER, 2010) (figura 1.1). Nesse método de sequenciamento manual, ocorre a síntese de uma fita complementar a sequência de DNA de

interesse, utilizando DNA polimerase, um iniciador (primer), além de uma mistura de deoxinucleotídeos (dNTPs) e dideóxinucleotídeos (ddNTPs). Durante a síntese, várias cópias da fita complementar são sintetizadas com a adição de dNTPs. Porém, quando os ddNTPs são incorporados, ocorre a interrupção da inclusão de novos nucleotídeos. Isso pode ocorrer em qualquer etapa da síntese, gerando fragmentos de diferentes tamanhos. Após essa etapa, as amostras são aplicadas em gel de poliacrilamida, onde ocorre a separação dos fragmentos de DNA. A visualização dessas bandas indicará a sequência de nucleotídeos da fita complementar e, por complementaridade de bases é obtida a sequência de interesse. Posteriormente o método passou a ser automático, com a utilização de computadores para leitura do gel e processamento das sequências e a utilização de fluorocromos para os nucleotídeos (HEATHER; CHAIN, 2016).

Este método permitiu novos tipos de análises nas quais o isolamento de materiais genéticos de um único organismo tem sido gradualmente substituído pelo sequenciamento de vários organismos simultaneamente. Utilizando esse método foi possível, por exemplo, sequenciar o genoma humano (METZKER, 2010).

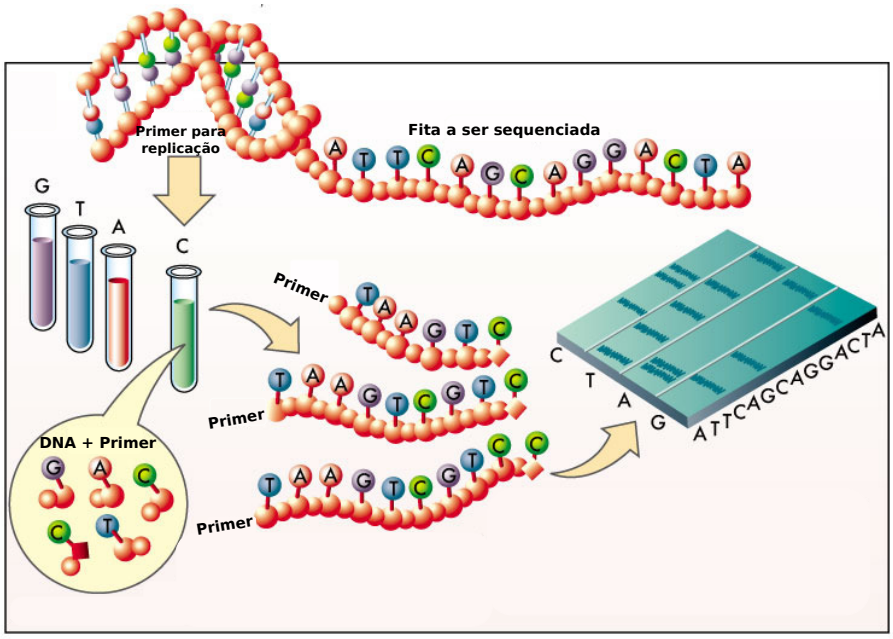

Figura 1.1: Ilustração do funcionamento da tecnologia de sequênciamento de Sanger. À 95°C a fita dupla se desnatura em duas fitas simples. Essa fita simples encontra-se em um tubo contendo *primer*, nucleotídeos e DNA polimerase. A aproximadamente 50°C ocorre o anelamento do *primer* e com 72°C a DNA polimerase entra em ação realização a extensão da fita complementar. Quando os nucleotídeos modificados são incorporados a fita de DNA ocorre o bloqueio da adição de novos nucleotídeos, gerando assim fragmentos de tamanhos diferentes. Ao colocar no gel essas fitas migram por causa da sua massa e carga. No sequenciamento automatizado os ddNTPs (dideoxinucleotídeos) são fluorescentes e utilizando um sistema de laser e detector é possível identificar esses nucleotídeos no capilar. Adaptado de (WINNICK, 2004).

1.2.2 Sequenciamento de nova geração

Os novos métodos de sequenciamento que utilizam estratégias como preparação de fitas molde, sequenciamento e imagem, alinhamento de genoma e métodos de montagem são chamadas de sequenciamentos de nova geração, em inglês *next-generation sequencing (NGS)*. Cada fabricante de equipamentos de NGS utiliza métodos diferentes para a criação dos *templates* (modelos), sequenciamento e captura de imagem. Essas diferenças refletem diretamente na qualidade e no custo dos dados gerados. (METZKER, 2010).

Entre as principais fabricantes de equipamento de sequenciamento temos: Applied Biosystems, Ilumina, Pacific Biosciences, Ion Torrent, Oxford Nanopore e a SOLiD. Para exemplificar essas tecnologias veremos o funcionamento da metodologia aplicada pelas empresas Applied Biosystems, Ilumina, Pacific Biosciences.

1.2.2.1 Applied Biosystems

O método de preparação do *template* onde são fixados *primers* individuais distribuídos espacialmente sobre um suporte sólido é utilizado pela Applied Biosystems (figura 1.2). Nessa técnica os tubos possuem uma mistura de *primers*, *templates*, dNTPs e polimerase. Ocorre uma reação de amplificação por PCR, que é quebrada por emulsão ocorrendo assim a separação do *template*. As esferas são fixadas em uma lamina de vidro.

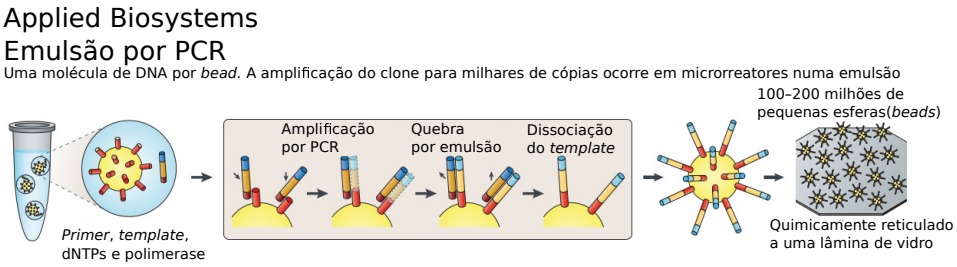

Figura 1.2: Estratégias de imobilização de *templates* utilizada pela Applied Biosystems. Para essa técnica ocorre a adição e ligação dos adaptadores e seleção dos fragmentos com adaptadores. Em uma segunda etapa acontece a ligação da fita simples de DNA às esferas e em seguida a amplificação do DNA por PCR, por último temos a eliminação das gotículas. Com a presença de reagentes (sulfurilase) é gerada luz que é captada pelo equipamento. Adaptado de (METZKER, 2010).

1.2.2.2 Ilumina

A Ilumina utiliza um método de fixação de uma única molécula de DNA por *cluster* (figura 1.3). Nessa técnica são criados agrupamentos utilizando fragmentos ou *mate-paired templates* em uma lâmina de vidro. Cada agrupamento possui *primers* que durante a fase de amplificação criarão pontes entre os segmentos *foward* e *reverse* da fita de DNA.

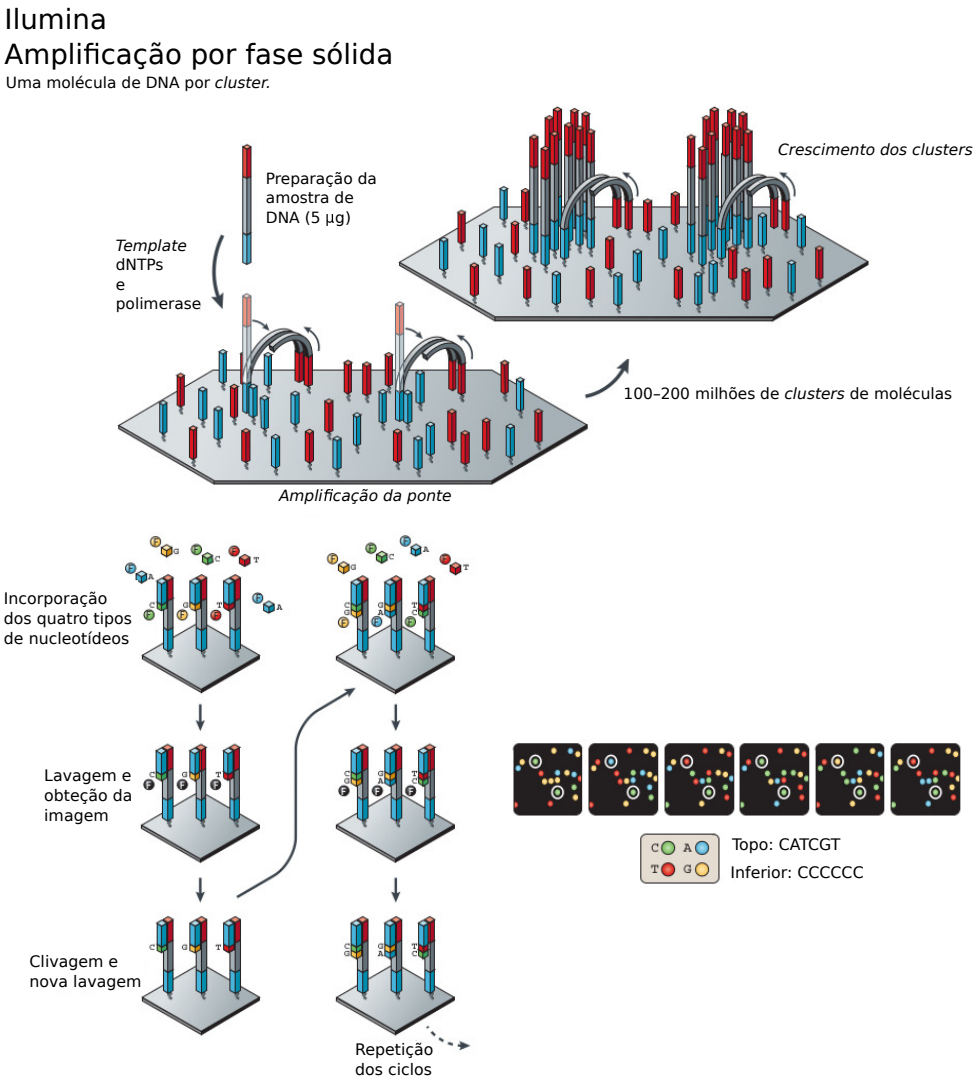

Figura 1.3: Estratégia de amplificação por fase sólida usada pelos equipamentos da Ilumina. Em cada *cluster* temos a ligação entre os segmentos da fita de DNA *foward* e *reverse*. Com isso os nucleotídeos são incorporados para a criação da fita de DNA complementar, ocorre uma lavagem para retirar os nucleotídeos não incorporados, é obtida a imagem pelo equipamento, é realizada a clivagem para retirar os fluorocromos (responsável pela emissão de luz). Esse ciclo é repetido até a obtenção de toda a sequência de nucleotídeos da fita complementar. Adaptado de (METZKER, 2010).

1.2.2.3 Pacific Biosciences

O método utilizado pela Pacific Biosciences realiza a detecção dos fragmentos incorporados em tempo real e é considerado de terceira geração. Nesse método a polimerase é fixada a superfície sólida, a detecção dos nucleotídeos incorporados ocorre em tempo real, com isso é possível conseguir tamanhos de *reads* maiores do que as outras metodologias (figura 1.4)(LIU et al., 2012b; METZKER, 2010).

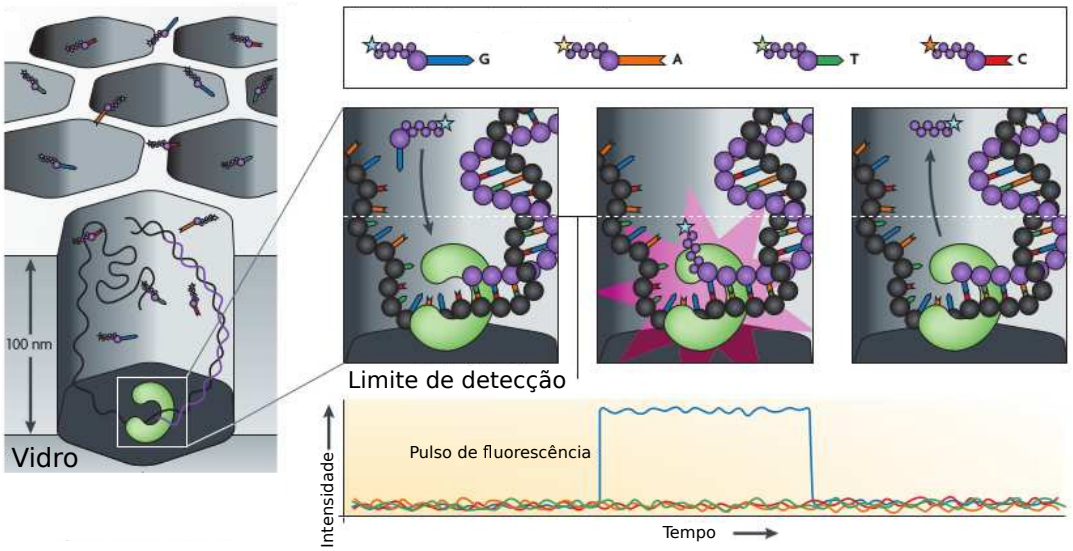

Figura 1.4: Estratégia de fixação da polimerase no suporte utilizada pela Pacific Biosciences. A polimerase é fixada em uma superfície solida. O equipamento possui um limite de detecção para captura da imagem, com isso consegue verificar a intensidade em relação ao tempo do pulso de fluorescência quando um nucleotídeo é incorporado. Após a captura da imagem ocorre a clivagem do fluorocromo. Adaptado de (METZKER, 2010).

1.2.2.4 Captura das imagens, alinhamento e montagem para NGS

Para a fase de captura das imagens é realizado um consenso entre os sinais observados quando os nucleotídeos são ligados a *templates* idênticos. (METZKER, 2010). Os arquivos resultantes da fase de captura de imagem possuem os *reads* que são os fragmentos do DNA obtidos pelos equipamentos de NGS (SHARPTON, 2014).

Alinhamento e montagem são as etapas mais demoradas e custosas computacionalmente que são realizadas após os *reads* terem sido obtidos e dependem do propósito do sequenciamento. Os *reads* podem ser alinhados a um sequência de referência conhecida, com isso conseguimos comparar a genomas já conhecidos. Podemos realizar a montagem *de novo* que não utiliza genomas de referência diminuindo assim a possibilidade de resultado tendencioso.

(PENG et al., 2011).

Com o advento das tecnologias de sequenciamento de nova geração é possível atender a necessidade cada vez maior por informações genômicas rápidas, precisas e de baixo custo. A geração de grande quantidade de dados é uma vantagem em relação aos métodos convencionais. A possibilidade de sequenciar o genoma completo de vários tipos de organismos permitiu a comparação em larga escala assim como realizar estudos evolutivos que antes não eram possíveis (METZKER, 2010; CHUN; RAINEY, 2014).

A tabela 1.1 possui informações comparativas em relação entre as principais tecnologias de sequenciamento em uso atualmente, o tamanho de *reads* gerados por essas tecnologias, o tempo de corrida, rendimento por corrida (Gpb/corrida) e custo por GB (U$$).

Tabela 1.1: Tabela comparativa das principais tecnologias de sequenciamento

Tecnologia de sequenciamento	Tamanho dos reads	Tempo de corrida	Rendimento por corrida (Gpb/corrida)	Custo por Gb (US$)
Applied Biosystems 3730 (capillary)	650	2h	$9{,}6\text{x}10^{-5}$	2.307.692,31
Illumina MiSeq	50 - 600	4h - 56 h	0,30 - 15,00	102,00 -1.866,67
Illumina NextSeq 500	75 - 300	11h - 29 h	19,50 - 120	35,33 -52,82
Illumina HiSeq 2500	50 - 500	10h - 6 dias	15 - 500	28,82 -95,33
Illumina HiSeq 4000	50 - 300	1 dia - 3,5 dias	125 - 750	20,53 -48,08
Illumina HiSeq X - Five	300	$\leq$ 3 dias	900	7,08 -10,63
Ion Torrent - PGM	400	4 hrs - 7h	0,22 - 2,20	397,27 -2.154,55
Ion Torrent - S5	200 - 400	2,5 h - 4h	2 - 16	79,69 -476,50
Oxford Nanopore	10000	Não especificado	6 - 15500	6,14 -150,00
Pacific Biosciences	10000 - 12000	$\leq$ 6h	0,66 - 3,85	181,82 -303,03
SOLiD - 5500	700 - 1410	8 dias	77 - 155,10	67,72 -79,23

Fonte: Adaptado de Glenn, TC. 2011. Field Guide to Next Generation DNA Sequencers. Molecular Ecology. Disponível em: `http://www.molecularecologist.com/next-gen-table-2-2016/`. Acesso em: 29 jul. 2016.

1.3 Descoberta de microorganismos

1.3.1 Gene 16S rRNA

Woese e colaboradores (WOESE; FOX, 1977) apresentaram uma proposta de utilizar uma parte conservada do código genético, os genes ribossomais, para classificação taxonômica de bactérias, sendo que o gene mais utilizado para esse tipo de análise é o 16S rRNA (TORTOLI, 2003) que é ilustrado na figura 1.5.

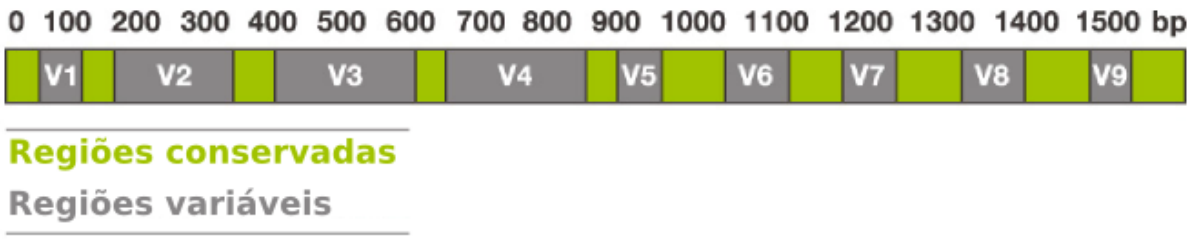

Figura 1.5: Gene 16S rRNA. Em verde é ilustrada as regiões conservadas e em cinza as regiões variáveis do gene. Adaptado de (ALIMETRICS,).

O gene 16S rRNA possui nove regiões variáveis (V1 - V9) que são ladeadas por regiões conservadas. Essas regiões variáveis possuem polimorfismo que são utilizadas para caracterizar os diferentes tipos de microorganismos (CHAKRAVORTY et al., 2007). Quando um microorganismo é comparado às espécies já existentes, e o valor de identidade obtido entre as sequências de 16S rRNA (figura 1.6) é de 99 % o microorganismo é caracterizado como sendo da mesma espécie. Para valores entre 97 % e 99 % é considerado como do mesmo gênero. Quando a identidade é menor do que 97 % o microorganismo é considerado como uma espécie potencialmente nova. (DRANCOURT; BERGER; RAOULT, 2004)

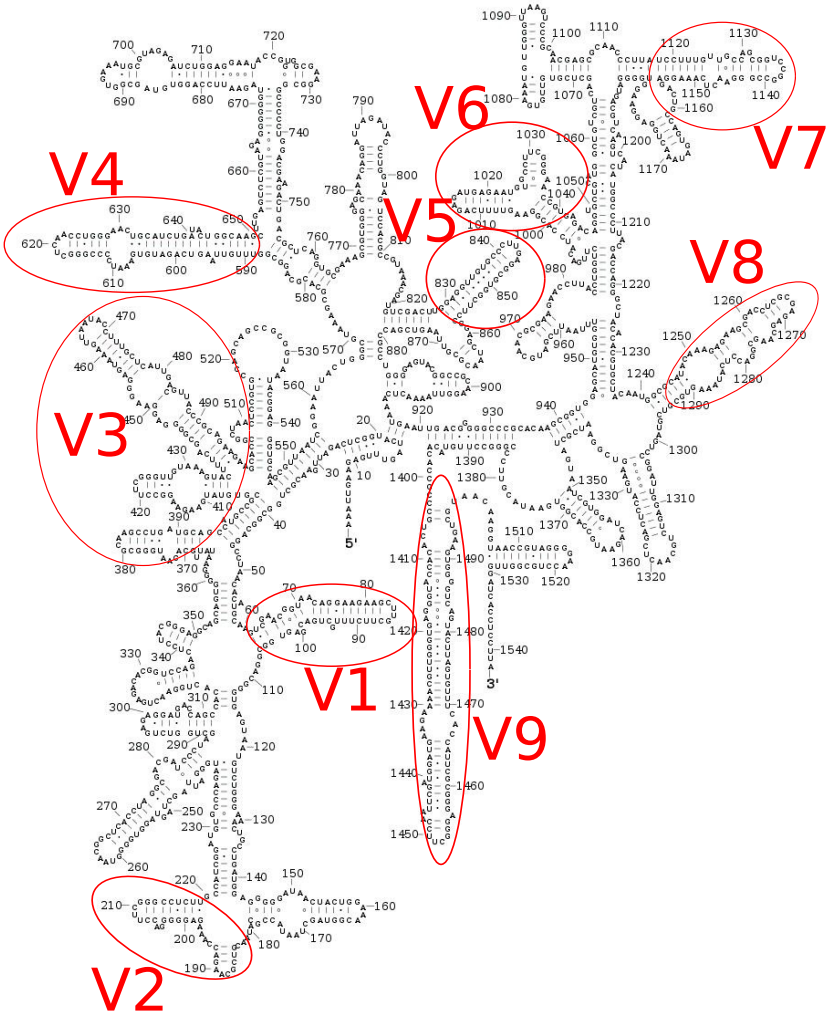

Figura 1.6: Estrutura do 16S rRNA. Na figura estão circuladas as 9 regiões variáveis do 16S rRNA. Essas regiões quando comparadas entre os microorganismos auxiliam na identificação de microorganismos de uma mesma espécie. Adaptado de (CASE et al., 2007)

Entre as razões para a utilização do gene 16S rRNA estão: esse gene pode ser encontrado em praticamente todas as bactérias, como não há alteração na função do gene 16S as análises das sequências nas regiões variáveis podem ser utilizadas como medidas em relação a evolução e o tamanho do gene é suficientemente grande para propósitos relacionados a computação (JANDA; ABBOTT, 2007). Na técnica de aplicação por reação de PCR o gene 16S rRNA é amplificado por reação de PCR utilizando *primers* específicos, clonado e sequenciado (KLINDWORTH et al., 2012). A utilização do método de Sanger em conjunto com a utilização do gene 16S rRNA abriram novas possibilidades revolucionando assim o estudo e classificação dos microrganismos (ESCOBAR-ZEPEDA; LEÓN; SANCHEZ-FLORES, 2015). Com isso, atualmente a diversidade microbiana pode ser caracterizada através de análises dependentes ou independentes de cultivo.

1.4 Metagenômica

Em 1998, Jo Handelsman (HANDELSMAN et al., 1998) utilizou pela primeira vez o termo metagenômica, uma técnica que possibilita a análise de genomas em um nível maior de complexidade sem a necessidade de cultivo prévio. Com a metagenômica, conseguimos obter informação da capacidade metabólica e funcional da comunidade microbiana através da análise do DNA obtido diretamente da amostra ambiental ou orgânica. Essa técnica possibilita responder as perguntas biológicas: “quem são?” e “o que estão fazendo?” os microrganismos de uma determinada amostra (HANDELSMAN, 2004).

As bibliotecas genômicas para utilização na metagenômica são criadas utilizando análise baseada na sequência ou na função (HANDELSMAN, 2004). A técnica consiste na obtenção do DNA de uma amostra ambiental adquirida de forma direta ou indireta. Na extração direta ocorre a lise celular direto do solo. Após a lise, o DNA é extraído e purificado para posterior análise. Na extração indireta as células bacterianas são extraídas da matriz do solo e depois lisadas para a subsequente extração do DNA, purificação e análise. O método direto de extração de DNA é o mais utilizado. Após a extração, os fragmentos de DNA são ligados a vetores, montando vetores de clonagem com capacidade de replicação, denominados DNA recombinante. Por último, essas moléculas de DNA são inseridas em uma bactéria hospedeira cultivável em laboratório, originando uma biblioteca metagenômica, em que os clones recombinantes com parte do DNA presente em um determinado ambiente ou amostra podem ser multiplicados e mantidos por muitos anos (DELMONT et al., 2011; HANDELSMAN, 2004). A análise baseada em sequência pode ser utilizada por ferramentas computacionais para as posteriores etapas de análise (Figura 1.7). A metagenômica possui grande importância no setor de biotecnologia, podendo atender a grande demanda das indústrias e da medicina em busca de enzimas mais eficientes. Isso é possível através da descoberta de novos microrganismos ou do desenvolvimento de processos de base biotecnológica pois ela permite que seja acessada uma fonte de conhecimento ainda não explorada de diversidade biológica e molecular (MOREIRA, 2015).

1.5 Análise taxônomica de dados metagenômicos

Com a diminuição do custo e o aumento da quantidade de dados gerados através da metagenômica as análises foram divididas em etapas, facilitando assim a descoberta de novos organismos, a montagem correta do genoma ou o mapeamento das sequências. Essas etapas podem ser realizadas isoladas ou combinadamente para a análise do metagenoma, e são classificadas em: controle de qualidade, montagem, detecção de genes, anotação de genes, classificação ta-

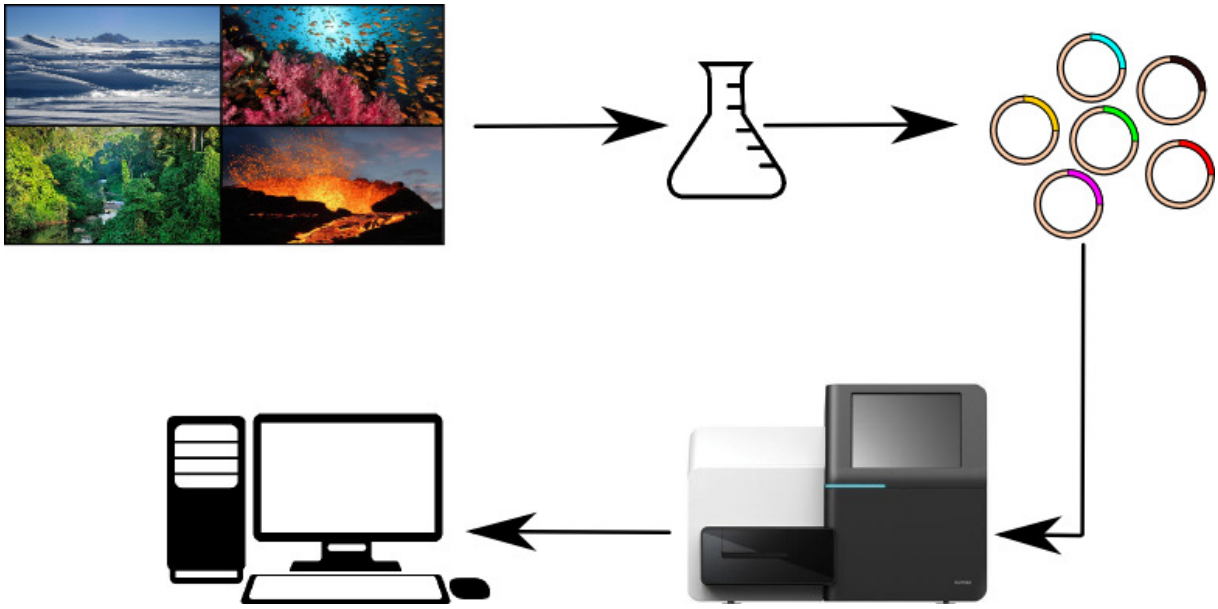

Figura 1.7: Etapas de obtenção de dados de metagenoma: a amostra de interesse é obtida para o estudo, realiza-se a extração do material genético que são inseridas em vetores para a criação de bibliotecas metagenômica. Esse vetores são sequenciados utilizando equipamentos de NGS. A informação adquirida é armazenada para posterior análise.

xonômica (*binning*), análise comparativa integrada e gerenciamento de dados (LADOUKAKIS; KOLISIS; CHATZIIOANNOU, 2015).

O processo denominado *binning* ou classificação taxonômica permite a análise dos dados metagenômicos sem a necessidade de utilização de genes marcadores filogenéticos, como o gene 16S rRNA, utilizado na identificação taxonômica de bactérias que apesar de possuir alto nível de precisão é encontrado somente em cerca de 1 % das sequências obtidas da amostra do metagenoma (MOREIRA, 2015).

Está análise de dados metagenômica pode ser baseada na similaridade entre as sequências do metagenoma com sequências já classificadas e depositadas em banco de dados de referência como em programas baseados em BLAST (*Basic local alignment search tool*), que possuem um alto nível de precisão mas que tornam o processo lento e custoso computacionalmente ou na composição dos nucleotídeos (LIU et al., 2012a) que utiliza cálculos de medidas como: uso de códon, conteúdo GC e frequência de oligonucleotídio (HIGASHI et al., 2012). Existem vários programas híbridos que mesclam os dois métodos.

1.6 Programas de análise taxonômica

Segundo a base de dados OMICtools (HENRY et al., 2014) até agosto de 2016 existiam disponíveis para *download* 66 programas de classificação taxonômica sendo desses: 48 baseados em técnicas de alinhamento, 11 baseados em composição e 7 programas que mesclam as duas técnicas. Como podemos ver na figura 1.8 esses programas foram desenvolvidos entre América do Norte, Europa e Ásia.

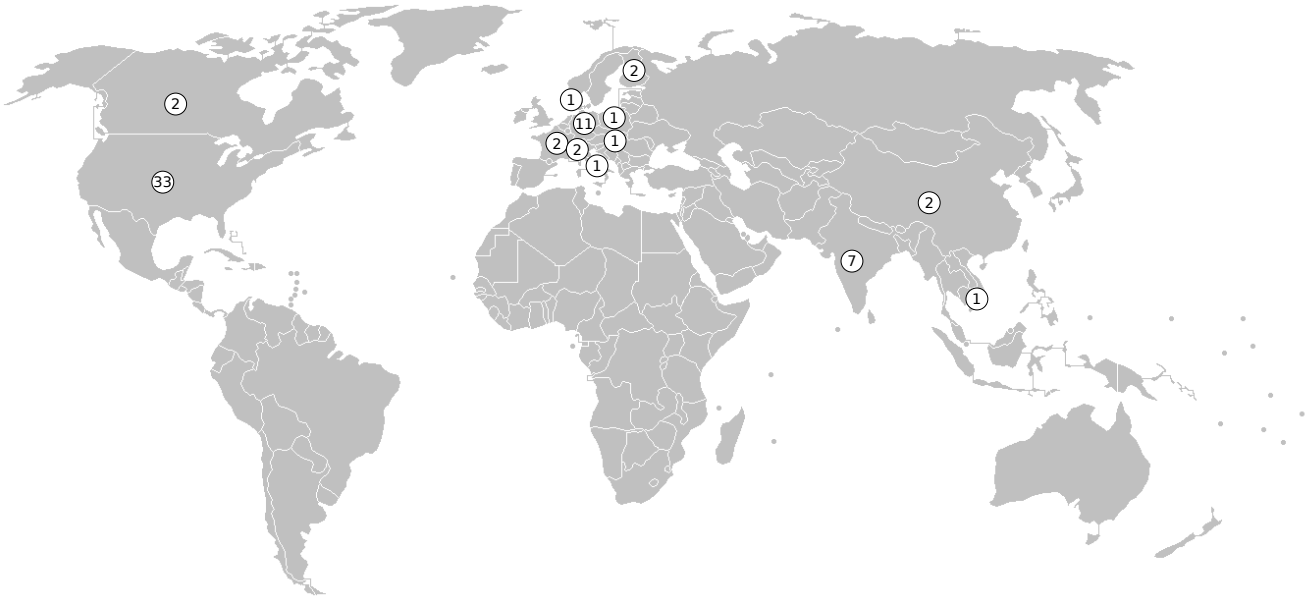

Figura 1.8: Mapa mundi mostrando a distribuição de programas para identificação taxonômica desenvolvida por cada país. Estados Unidos, Europa e Asia são os lugares com maior desenvolvimento de programas de análise taxonômica.

Nas buscas pelo BLAST, as sequências podem ter múltiplas combinações. Programas como o Kraken (WOOD; SALZBERG, 2014), MG-RAST (KEEGAN; GLASS; MEYER, 2015) e MEGAN (HUSON et al., 2007) utilizam uma aproximação chamada "Menor Ancestral Comum" (*Lowest Common Ancestor* (LCA)) (HUSON et al., 2007) que atribui a sequência ao maior nível taxonômico que pode ser classificado como pertencendo ao LCA reduzindo assim o resultado de falsos positivos. Porém essa metodologia possui seu viés. O MEGAN, por exemplo, pode fornecer resultados falsos negativos reduzindo a sensibilidade do *binning* quando descarta sequências que não satisfaçam cortes pré-definidos pelo usuário. Uma vez que o tamanho do genoma está relacionado ao número de leituras em amostras metagenômicas, o Kraken tenta resolver esse problema utilizando um banco de dados com informações de todos os k-mers (frequência de oligonucleotídeos) e dos LCAs de todos os organismos que possuam aquele k-mer exato. Programas como o CARMA (KRAUSE et al., 2008) e o SOrt-ITEMS

(HAQUE et al., 2009) são baseados em BLAST mas não utilizam o LCA. O CARMA realiza a classificação filogenética dos *reads* baseados em domínios conservados do banco de dados de proteínas Pfam. O SOrt-ITEMS adota uma abordagem exaustiva para julgar primeiro a qualidade do alinhamento do *read* com as sequências homólogas na base de dados para encontrar um nível adequado na árvore taxonômica em que o *read* possa ser atribuído. O algoritmo usa então a abordagem de ortologia para identificar sequências homólogas que mostram ortologia significativa, ou seja, semelhança de reciprocidade com o *read* da amostra.

Os programas que utilizam o método de composição de sequência baseiam-se no padrão de disposição dos nucleotídeos, levando em consideração a conservação entre as espécies ao longo da evolução. A ferramenta Phylopythia (MCHARDY et al., 2007) utiliza aprendizado de máquina por *Support Vector Machine* (SVM) junto com marcadores filogenéticos para a identificação dos *reads* de acordo com a composição de oligonucleotídeos de vários tamanhos. O RAIphy (NALBANTOGLU et al., 2011) utiliza um modelo de índice relativo de abundância (RAI), em que um valor de índice é calculado para cada *k-mer*, com base nas estatísticas super e sub-abundância recolhidos a partir do táxon. INDUS (MOHAMMED et al., 2011a) verifica a distância entre as frequências dos vetores de tetranucleotídeos. Sequedex (BERENDZEN et al., 2012) usa um *k-mer* de tamanho 10 para realizar a identificação e localiza *strings* de amino ácidos de tamanho 10 que compartilham ao menos dois genomas de referência de gêneros distintos de bactérias. MetaCV (LIU et al., 2012a) transforma o nucleotídeo em peptídeos de 6 *frames* que é decomposto em oligopeptídeos de tamanho fixo. Estes são utilizados para a classificação taxonômica quando comparado a um banco de dados de proteínas. FOCUS (SILVA et al., 2014) usa a frequência de oligonucleotídeos (k-mers) do genoma completo e calcula o *set* de abundância otimizado do organismos usando o método de mínimos quadrados não negativos. TAC-ELM (RASHEED; RANGWALA, 2012) utiliza oligonucleotídeos e conteúdo GC para criar uma rede neural. *Large scale metagenomics* (VERVIER et al., 2016) usa algoritmos de aprendizagem de máquina para realizar a classificação taxonômica.

Existem programas que utilizam uma combinação da metodologia de alinhamento com o método de composição de sequências, esses métodos são denominados métodos híbridos. O Binpairs (DUTTA et al., 2014) usa uma estratégia da tecnologia do Ilumina para emparelhamento das informações de *reads* pequenos para aumentar a acurácia do *binning* em conjuntos de dados metagenômicos. PhymmBL (BRADY; SALZBERG, 2009) realiza o BLAST junto a modelos interpolados de Markov para caracterizar oligonucleotídeos de tamanho variável típicos de grupos filogenéticos. RITA (MACDONALD; PARKS; BEIKO, 2012) primeiro utiliza um classificador Naive Bayers e depois três algoritmos de BLAST (D-BLASTN, BLASTN e BLASTX). SeMeta (LE; TRAN; TRAN, 2016) separa os *reads* em *clusters* e depois atribui

cada *cluster* ao táxon que melhor se enquadra utilizando como base a similaridade entre os *reads* e a base de dados de referência. SPHINX (MOHAMMED et al., 2011b) possui três etapas: a primeira é composicional e identifica um pequeno *subset* de sequências na base de dados de referência que é similar em composição ao *read* a ser identificado; a segunda utiliza BLAST para verificar o alinhamento do *read* contra o *subset*; a terceira analisa o resultado do BLAST e classifica o *read*. TWARIT (REDDY; MOHAMMED; MANDE, 2012) faz a classificação em duas etapas: a primeira etapa procura sequências iguais às já existentes no banco de dados de referência; a segunda etapa verifica em um banco de dados de composição a distância entre vetores de frequência de tetranucleotídeos. Treephyler (SCHREIBER et al., 2010) baseia-se nas predições do PFAM junto à perfis de modelo oculto de Markov pré calculados da família do PFAM e utiliza uma árvore filogenética de referência para atribuir os táxons aos *reads*.

5 *Conclusões*

Com a análise das combinações de medidas para identificação de microorganismos nos táxons gênero, família, ordem e classe, verificamos que a medida predominante foi a concentração de GC. Contudo, identificamos que para alguns tamanhos de fragmentos outras medidas como abundância total de dinucleotídeos, entropia de tripletes e entropia de tetrapletes conseguiram resultados acima de 70% de identificação.

Na aplicação da metodologia e comparação com as metodologias já existentes Raiphy, Phymmbl e Phymm nas amostras de metagenomas obtidas do projeto *"A human gut microbial gene catalog established by deep metagenomic sequencing"* houveram poucos *reads* identificamos pelas três metodologias como sendo do mesmo gênero. Ao aplicarmos as medidas do método que estamos propondo a combinação das medidas GC e S2 obtiveram os melhores resultados com identificações acima de 89 % e conseguindo diminuir a quantidade de possíveis gêneros candidatos por *read*. O tempo gasto para as análises foi menor em comparação com os programas já existentes.

Mesmo com a diminuição da quantidade de gêneros candidatos ainda consideramos alta a quantidade de candidatos. Uma forma de solucionar esse problema seria utilizando uma abordagem híbrida entre o método composicional e o método por alinhamento. O método composicional faria um filtro entre os possíveis gêneros e o método por alinhamento selecionaria entre os possíveis resultados de forma mais específica. Com isso haveria uma diminuição do tempo computacional necessário para a análise pois a metodologia que estamos propondo diminuiria consideravelmente a busca pela informação no banco de dados para realização do alinhamento.

Referências Bibliográficas

ALIMETRICS. *DNA Sequence Analysis - The only species-specific approach for novel bacteria*. `http://www.alimetrics.net/en/index.php/dna-sequence-analysis`. [Online; accessed 19-July-2008].

BERENDZEN, J. et al. Rapid phylogenetic and functional classification of short genomic fragments with signature peptides. *BMC research notes*, BioMed Central Ltd, v. 5, n. 1, p. 460, 2012.

BOHLIN, J. et al. Analysis of intra-genomic gc content homogeneity within prokaryotes. *BMC genomics*, v. 11, p. 464, 2010. ISSN 1471-2164.

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.

BULLOCK, N. O.; ASLANZADEH, J. Biochemical profile-based microbial identification systems. In: *Advanced Techniques in Diagnostic Microbiology*. [S.l.]: Springer, 2013. p. 87–121.

CASE, R. J. et al. Use of 16s rrna and rpob genes as molecular markers for microbial ecology studies. *Applied and environmental microbiology*, Am Soc Microbiol, v. 73, n. 1, p. 278–288, 2007.

CHAKRAVORTY, S. et al. A detailed analysis of 16s ribosomal rna gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods*, v. 69, p. 330–339, May 2007. ISSN 0167-7012.

CHUN, J.; RAINEY, F. A. Integrating genomics into the taxonomy and systematics of the bacteria and archaea. *International journal of systematic and evolutionary microbiology*, Microbiology Society, v. 64, n. 2, p. 316–324, 2014.

DELMONT, T. O. et al. Metagenomic comparison of direct and indirect soil dna extraction approaches. *Journal of microbiological methods*, Elsevier, v. 86, n. 3, p. 397–400, 2011.

DRANCOURT, M.; BERGER, P.; RAOULT, D. Systematic 16s rrna gene sequencing of atypical clinical isolates identified 27 new bacterial species associated with humans. *Journal of clinical microbiology*, v. 42, p. 2197–2202, May 2004. ISSN 0095-1137.

DUTTA, A. et al. Binpairs: Utilization of illumina paired-end information for improving efficiency of taxonomic binning of metagenomic sequences. *PloS one*, Public Library of Science, v. 9, n. 12, p. e114814, 2014.

DUTTA, C.; PAUL, S. Microbial lifestyle and genome signatures. *Current genomics*, Bentham Science Publishers, v. 13, n. 2, p. 153–162, 2012.

ESCOBAR-ZEPEDA, A.; LEÓN, A. V.-P. de; SANCHEZ-FLORES, A. The road to metagenomics: from microbiology to dna sequencing technologies and bioinformatics. *Frontiers in genetics*, Frontiers Media SA, v. 6, 2015.

FIERER, N.; JACKSON, R. B. The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America*, National Acad Sciences, v. 103, n. 3, p. 626–631, 2006.

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.

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

HANDELSMAN, J. et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, Elsevier, v. 5, n. 10, p. R245–R249, 1998.

HAQUE, M. M. et al. Sort-items: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, Oxford Univ Press, v. 25, n. 14, p. 1722–1730, 2009.

HEATHER, J. M.; CHAIN, B. The sequence of sesequence: The history of sequencing dna. *Genomics*, v. 107, n. 1, p. 1–8, January 2016.

HENRY, V. J. et al. Omictools: an informative directory for multi-omic data analysis. *Database*, Oxford University Press, v. 2014, p. bau069, 2014.

HIGASHI, S. et al. Analysis of composition-based metagenomic classification. *BMC genomics*, BioMed Central, v. 13, n. 5, p. 1, 2012.

HUSON, D. H. et al. Megan analysis of metagenomic data. *Genome research*, Cold Spring Harbor Lab, v. 17, n. 3, p. 377–386, 2007.

JANDA, J. M.; ABBOTT, S. L. 16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, Am Soc Microbiol, v. 45, n. 9, p. 2761–2764, 2007.

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.

KEEGAN, K. P.; GLASS, E. M.; MEYER, F. Mg-rast, a metagenomics service for analysis of microbial community structure and function. *Methods in molecular biology (Clifton, NJ)*, Springer, v. 1399, p. 207–233, 2015.

KLINDWORTH, A. et al. Evaluation of general 16s ribosomal rna gene pcr primers for classical and next-generation sequencing-based diversity studies. *Nucleic acids research*, Oxford Univ Press, p. gks808, 2012.

KRAUSE, L. et al. Phylogenetic classification of short environmental dna fragments. *Nucleic acids research*, Oxford Univ Press, v. 36, n. 7, p. 2230–2239, 2008.

LADOUKAKIS, E.; KOLISIS, F. N.; CHATZIIOANNOU, A. A. Integrative workflows for metagenomic analysis. *Multi-omic Data Integration*, Frontiers Media SA, p. 115, 2015.

LE, V. V.; TRAN, L. V.; TRAN, H. V. A novel semi-supervised algorithm for the taxonomic assignment of metagenomic reads. *BMC bioinformatics*, BioMed Central, v. 17, n. 1, p. 1, 2016.

LEWIS, K. et al. Uncultured microorganisms as a source of secondary metabolites. *The Journal of antibiotics*, Nature Publishing Group, v. 63, n. 8, p. 468–476, 2010.

LIU, J. et al. Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. *Nucleic acids research*, Oxford Univ Press, p. gks828, 2012.

LIU, L. et al. Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, v. 2012, p. 251364, 2012. ISSN 1110-7251.

MACDONALD, N. J.; PARKS, D. H.; BEIKO, R. G. Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic acids research*, Oxford Univ Press, p. gks335, 2012.

MADIGAN, M. T. et al. *Brock Biology of Microorganisms 13th edition*. [S.l.]: Benjamin Cummings, 2010.

MCHARDY, A. C. et al. Accurate phylogenetic classification of variable-length dna fragments. *Nature methods*, Nature Publishing Group, v. 4, n. 1, p. 63–72, 2007.

METZKER, M. L. Sequencing technologies - the next generation. *Nature reviews genetics*, Nature Publishing Group, v. 11, n. 1, p. 31–46, 2010.

MOHAMMED, M. H. et al. Indus-a composition-based approach for rapid and accurate taxonomic classification of metagenomic sequences. *BMC genomics*, BioMed Central, v. 12, n. 3, p. 1, 2011.

MOHAMMED, M. H. et al. Sphinx - an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics*, Oxford Univ Press, v. 27, n. 1, p. 22–30, 2011.

MOREIRA, L. M. *Ciencias genomicas : fundamentos e aplicacoes*. Sociedade Brasileira de Genetica, 2015. Disponível em: <https://drive.google.com/file/d-/0Bwb0CldyiHPwV3IxRGdyYklIdGc/view>.

MUTO, A.; OSAWA, S. The guanine and cytosine content of genomic dna and bacterial evolution. *Proceedings of the National Academy of Sciences of the United States of America*, v. 84, p. 166–169, Jan 1987. ISSN 0027-8424.

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

PENG, Y. et al. Meta-idba: a de novo assembler for metagenomic data. *Bioinformatics (Oxford, England)*, v. 27, p. i94–101, Jul 2011. ISSN 1367-4811.

QIN, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, v. 464, n. 7285, p. 59–65, Mar 2010. Disponível em: <http://dx.doi.org/10.1038-/nature08821>.

RASHEED, Z.; RANGWALA, H. Metagenomic taxonomic classification using extreme learning machines. *Journal of bioinformatics and computational biology*, World Scientific, v. 10, n. 05, p. 1250015, 2012.

REDDY, R. M.; MOHAMMED, M. H.; MANDE, S. S. Twarit: an extremely rapid and efficient approach for phylogenetic classification of metagenomic sequences. *Gene*, Elsevier, v. 505, n. 2, p. 259–265, 2012.

ROSENBERG, E. et al. The evolution of animals and plants via symbiosis with microorganisms. *Environmental microbiology reports*, Wiley Online Library, v. 2, n. 4, p. 500–506, 2010.

ROSENBERG, E.; ZILBER-ROSENBERG, I. Symbiosis and development: the hologenome concept. *Birth Defects Research Part C: Embryo Today: Reviews*, Wiley Online Library, v. 93, n. 1, p. 56–66, 2011.

SANGER, F.; NICKLEN, S.; COULSON, A. R. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 74, n. 12, p. 5463–5467, 1977.

SANTOS, L. dos; RYBARCZYK-FILHO; GERHARDT, G. J. L. Triplet entropy in h1n1 virus. *TendÃªncias em MatemÃ¡tica Aplicada e Computacional*, v. 12, p. 253–261, 2011.

SCHLOSS, P. D.; HANDELSMAN, J. Metagenomics for studying unculturable microorganisms: cutting the gordian knot. *Genome biology*, BioMed Central, v. 6, n. 8, p. 1, 2005.

SCHREIBER, F. et al. Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics*, Oxford Univ Press, v. 26, n. 7, p. 960–961, 2010.

SHARPTON, T. J. An introduction to the analysis of shotgun metagenomic data. *Frontiers in plant science*, v. 5, p. 209, 2014. ISSN 1664-462X.

SILVA, G. G. Z. et al. Focus: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ*, PeerJ Inc., v. 2, p. e425, 2014.

TORTOLI, E. Impact of genotypic studies on mycobacterial taxonomy: the new mycobacteria of the 1990s. *Clinical microbiology reviews*, Am Soc Microbiol, v. 16, n. 2, p. 319–354, 2003.

VERVIER, K. et al. Large-scale machine learning for metagenomics sequence classification. *Bioinformatics*, Oxford Univ Press, v. 32, n. 7, p. 1023–1032, 2016.

WINNICK, E. Dna sequencing industry sets its sights on the future: new technologies may duke it out with upgrades of the current method. *The Scientist*, Scientist Inc., v. 18, n. 18, p. 44–48, 2004.

WOESE, C. R.; FOX, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 74, n. 11, p. 5088–5090, 1977.

WOOD, D. E.; SALZBERG, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, BioMed Central, v. 15, n. 3, p. 1, 2014.

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.

ZILBER-ROSENBERG, I.; ROSENBERG, E. Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. *FEMS microbiology reviews*, The Oxford University Press, v. 32, n. 5, p. 723–735, 2008.