UNIVERSIDADE DE CAXIAS DO SUL

CENTRO DE CIÊNCIAS AGRÁRIAS E BIOLÓGICAS

INSTITUTO DE BIOTECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

Desenvolvimento e Validação Método de Análise de Sequências Genômicas Baseada em Padrões de Entropia, Coeficiente de Clusterização e Periodicidade.

Ricardo Augusto Manfredini

Caxias do Sul

2015

Ricardo Augusto Manfredini

Desenvolvimento e Validação de Método de Análise de Sequências Genômicas Baseada em Padrões de Entropia, Coeficiente de Clusterização e Periodicidade.

Tese apresentada ao Programa de Pós-Graduação em
Biotecnologia da Universidade de Caxias do Sul,
visando a obtenção do grau de Doutor em Biotecnologia.

Orientador: Prof. Dr. Sergio Echeverrigaray

Caxias do Sul

2015

Dados Internacionais de Catalogação na Publicação (CIP)
Universidade de Caxias do Sul
UCS - BICE - Processamento Técnico

M276d Manfredini, Ricardo Augusto, 1966-
Desenvolvimento e validação de método de análise de sequências genômicas baseada em padrões de entropia, coeficiente de clusterização e periodicidade / Ricardo Augusto Manfredini. - 2015.
73 f.: il.; 30 cm

Apresenta bibliografia.
Tese (Doutorado) – Universidade de Caxias do Sul, Programa de Pós-Graduação em Biotecnologia, 2015.
Orientador: Prof. Dr. Sergio Echeverrigaray.

1. Genomas. 2. Microorganismos. 3. Biotecnologia.. I. Título.

CDU 2.ed.: 575.111

Índice para o catálogo sistemático:

1. Genomas	575.111
2. Microorganismos	573.4
3. Biotecnologia	57.08

Catalogação na fonte elaborada pela bibliotecária
Carolina Machado Quadros – CRB 10/2236.

RICARDO AUGUSTO MANFREDINI

Desenvolvimento e Validação de Métodos de Análise de Sequências Genômicas Baseada em Padrões de Entropia, Coeficiente de Clusterização e Periodicidade

Tese apresentada ao Programa de Pós-graduação em Biotecnologia da Universidade de Caxias do Sul, visando à obtenção do título de Doutor em Biotecnologia.

Orientador: Prof. Dr. Sergio Echeverrigaray Laguna

TESE APROVADA EM 27 DE ABRIL DE 2015.

Orientador: Prof. Dr. Sergio Echeverrigaray Laguna

Prof. Dr. Henrique Bunselmeyer Ferreira

Profa. Dra. Ana Paula Longaray Delamare

Profa. Dra. Scheila de Ávila e Silva

Aguardo, equânime, o que não conheço —
Meu futuro e o de tudo.
No fim tudo será silêncio, salvo
Onde o mar banhar nada.
Ricardo Reis,
heterónimo de Fernando Pessoa

A Maria Leonora Zamboni de Almeida

minha companheira e esposa.

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Sergio Echeverrigaray, pelo apoio, contribuições e questionamentos realizados ao longo da realização da tese, que foram fundamentais e imprescindíveis.

Ao Prof. Dr. Günther J. L. Gerhardt, pelas ideias e esclarecimentos matemáticos.

Às Prof. Drª Ana Paula Longaray Delamare e Prof. Dr. Diego Bonatto integrantes da minha banca de qualificação pelo acompanhamento e colaborações pertinentes realizadas.

À Prof. Drª Scheila de Avila e Silva pela força e incentivo.

À Universidade de Caxias do Sul e ao PPG em Biotecnologia pelo apoio ao projeto e ao Núcleo de Pesquisa em Bioinformática.

Ao Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul pela oportunidade de liberação de carga horária para a conclusão desta tese.

Ao Prof. Dr. Aldo J. P. Dillon pelo incentivo e intermináveis discussões políticas.

Aos colegas do laboratório de Microbiologia Aplicada pelo clima descontraído de trabalho e coleguismo nos créditos cursados.

À secretária do PPG, Lucimara Serafini, pela cordialidade e eficiência para tratar de questões burocráticas e presteza na solução destas questões.

CONTEÚDO

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

A	Nucleotídeo de adenina
C	Nucleotídeo de citosina
dNTP	Desoxirribonucleotídeo trifosfatado
ddNTP	Didesoxirribonucleotídeo trisfosfatado
DFT	Transformada Discreta de Fourier
DNA	Ácido desoxirribonucleico
E. coli	*Escherichia coli*
FFT	Transforma Rápida de Fourier
FN	Falsos negativos
FP	Falsos positivos
FT	Transformada de Fourier
FTP	File Transfer Protocol
G	Nucleotídeo de guanina
GWT	Transformada de Gabor-*Wavelet*
IBM	International Business Machine
LRC	Long-range correlations
MGWT	Transformada Modificada de Gabor-*Wavelet*
Pb	Pares de base
RNA	Ácido ribonucleico
RNAm	Ácido ribonucleico mensageiro
RNAp	RNA-polimerase
RpoB	RNA-polimerase B
SGBD	Sistema Gerenciador de Banco de Dados
STFT	Transformada de Tempo-Curto de Fourier
T	Nucleotídeo de timina
VN	Verdadeiros pegativos
VP	Verdadeiros positivos

RESUMO

As sequências genômicas carregam uma ampla gama de informações sobre os organismos que a compõem. Obviamente, devido à grande semelhança destas informações e funções, espera-se que uma determinada sequência possa pertencer a muitos organismos, com probabilidades semelhantes. Entretanto, cada genoma carrega dentro de si certas peculiaridades que podem ser extraídas utilizando as ferramentas adequadas. Neste contexto, este trabalho propõe um processo de análise de sequências genômicas de bactérias, utilizando algumas medidas que são particularmente importantes: a entropia de triples (Sn), a quantificação da periodicidade 3 (P3) em uma sequência, o coeficiente de clusterização (D) e o percentual de GC. O processo aqui proposto nos permite inferir a qual organismo uma determinada sequência genômica pode pertencer, mostrando-se viável a sua utilização em metagenômica. Os resultados neste trabalho demonstram a eficácia deste método. Foram identificados 100% dos organismos presentes nas amostras estudadas (VP). Por outro lado, foi encontrado um grande número de organismos não pertencentes às amostras (FP), o que indica a grande similaridade de determinadas sequências, corroborando com alguns estudos que indicam que o genoma carrega consigo sequências órtologas, comuns a inúmeros organismos.

ABSTRACT

Genomic sequences carry a wide range of information on organism that compose it. Obviously, by reason that great similarity of this information and functions, it is expected that each sequence can belong to many organisms with a similar probability. However, each genome carries within itself certain peculiarities that can be extracted using appropriate tools. In this context , this paper proposes a methodology for the analysis of genomic sequences of bacteria , using some measures that are particularly important : The entropy of triples (Sn) , the quantification of frequency 3 (P3) in a sequence , the clustering coefficient (D) and the percentage of GC . The method proposed here allows us to infer which a particular organism genome sequence may belong, being feasible for use in Metagenomics. The results of this study demonstrate the effectiveness of this method, 100 % of the organisms were identified in the samples studied (VP). On the other hand, a large number of bodies which did not belong samples were found (FP), which indicates the high similarity of certain sequences, corroborating some studies indicate that the genome carries ortholog sequences, common to countless organisms.

1 INTRODUÇÃO

A molécula de DNA carrega a informação básica para a construção e funcionamento de um organismo inteiro (Almeida et al., 2002; Trifonov, 1998). Esta molécula é responsável pela maior parte das funções biológicas, desde a manutenção da informação à mudança da mesma, mas não apresenta função bioquímica em si mesma. Ela é apenas um polímero organizado formado por quatro nucleotídeos que se repetem ao longo da sequência. Existem padrões de repetições embebidos nesta organização nucleotídica compreendendo vários elementos da sequência. As análises destas sequências podem revelar funções e/ou mecanismos utilizados para a manutenção da informação e a organização da mesma (Dodiin et al., 2000). Assim sendo, a análise de correlações e repetições ao longo da sequência de DNA representa um importante auxílio aos estudos experimentais, lembrando que as sequências disponíveis vêm aumentando de forma exponencial nos últimos anos (Anastassiou, 2000).

Entre as diferentes abordagens para análise da informação contida em sequências de DNA, uma é a procura de periodicidades e padrões de repetições. A primeira e mais estudada das repetições é a de 3 pares de bases (3-bp) a qual pode representar uma informação útil na caracterização de regiões codificadoras (Billing et al., 1999; Barral et al., 2000; Almeida et al., 2002). Estas repetições podem estar relacionadas com a formação primária de proteínas

(Trifonov, 1998). Outra periodicidade importante encontrada nas sequências de DNA é a denominada 10-11bp. A função desta periodicidade não é perfeitamente compreendida, mas existem evidências indicando que esta está relacionada com a estrutura de pregas e direção de enrolamento e superenrolamento da molécula de DNA, assim como a disposição de aminoácidos hidrofílicos e hidrofóbicos em estruturas de alfa hélice em diversas proteínas (Trifonov, 1998; Schieg et al., 2004). Considerando que a replicação de DNA depende do estado topológico da molécula, o estudo da prevalência e distribuição de periodicidades ao longo das sequências genômicas é de importância evidente.

Desde a contribuição fundamental de Barabási (Barabási and Albert, 1999), modelos de redes têm sido desenvolvidos para a caracterização de sistemas complexos. Entre outros podemos citar: a internet, as redes sociais, as redes metabólicas, as redes alimentares e as redes linguísticas (Dorogovtsev et al., 2003; Allegrini et al., 2003). Foram propostas metodologias para caracterização de sequências de DNA utilizando análises de redes (Gerhardt et al., 2006). Neste trabalho, os vértices dos gráficos de redes correspondem aos 64 triplets de nucleotídeos e a conexão entre estes vértices são estabelecidas cada vez que dois triplets se encontram contíguos na sequência de DNA. A rede gerada a partir da análise de sequências foi denominada *DNA Sequence Network* (DSN) e corresponde a uma união de métodos estatísticos tradicionais aplicados à teoria de redes no intuito de caracterizar sequências genômicas. A rede incorpora dados relativos à sequência de nucleotídeos, posição de triplets e organização do sistema. Como um todo, o modelo proposto se baseia na medida de coeficiente de agrupamento (clustering coefficient).

Dentro deste contexto, e dando sequência aos trabalhos que o Núcleo de Bioinformática da UCS vem realizando, a presente proposta visa ampliar a utilização de redes para análise de

periodicidades em sequências de DNA incluindo fatores de correção, novas periodicidades, novos atributos de pesos, entre outras, aliando o desenvolvimento de algoritmos com a criação de programas computacionais adequados a análise rápida, eficiente e interativa de periodicidade.

2 OBJETIVOS

2.1 Objetivo geral

Desenvolver ferramentas computacionais para estudo de padrões de sequências genômicas visando sua utilização na identificação dos organismos que esta pode pertencer, bem como sua utilização em estudos de metagenoma.

2.2 Objetivos específicos

Esta tese visa mostrar o quanto podemos descobrir da origem de uma sequência particular, do seu conteúdo GC e organização dos seus triples e como isso se correlaciona com o tamanho da sequência. Embora outras medidas ou "suposições" possam ser feitas sobre os organismos, os dados aqui considerados podem fornecer um pré-filtro a outras técnicas.

- Criar ferramentas computacionais para a caracterização de periodicidades (2pb+1, 3pb) em sequências genômicas.
- Criar ferramentas computacionais por meio de redes para quantificar diferenças genômicas levando ou não em consideração pesos diferenciais baseados na concentração de GC, na posição relativa das bases e por nodos preferenciais.

- Criar ferramentas para quantificação de conjuntos repetitivos de bases em sequências genômicas.
- Desenvolver uma ferramenta de uso simples e universal para análise de periodicidades e redes de genomas.
- Desenvolver softwares para facilitar a análise qualitativa e quantitativas de repetições genômicas.
- Desenvolver softwares para utilização das ferramentas desenvolvidas na análise de genomas e metagenomas.
- Validar o método proposto.

3 REVISÃO BIBLIOGRÁFICA

3.1 Análise de Sequências Genômicas

Houve um expressivo crescimento do número de organismos que tiveram o seu genoma sequenciado, devido aos avanços de técnicas e ferramentas de sequenciamento, usando massivamente a bioinformática (Cohen, 2005). Essas sequências estão depositadas em bases de dados públicas, em especial, no GenBank (Benso et al., 2013). Em alguns anos, muito provavelmente, o sequenciamento de espécies dará lugar ao sequenciamento de indivíduos, fato que gerará uma enorme massa de dados que transcende a análise física do genoma, abrindo um leque enorme para a análise informacional dos mesmos.

A análise mais comum das sequências genômicas é a análise estatística. Inicialmente foram analisadas as frequências dos nucleotídeos, das famílias químicas, as purinas (A e G) e as pirimidinas (C e T), dos códons ou de *motifs* (Vieria, 1999). O passo seguinte foram análises mais elaboradas em relação a distribuições, estatística comparando sequências distintas (Piazza & Lio`, 2005), caracterização de correlações (Herzer et al., 1999) e periodicidades (Hosid et al., 2004).

A concentração de periodicidades de nucleotídeos nas regiões intergênicas de *E. coli* sugere que a periodicidade de dinucleotídeos é uma propriedade típica das regiões intergênicas dos procariontes (Hosid et al., 2004). Os mesmos autores sugerem que esta distribuição da periodicidade dos dinucleotídeos serve a muitas funções biológicas, entre elas, à regulação da transcrição.

Outra periodicidade muito estudada é a de 10-11pb (Schieg & Herze, 2004). Esta periodicidade pode estar associada a curvatura do DNA. Foram observados padrões distintos de periodicidades em arqueas e bactérias, com períodos de cerca de 10pb sendo mais comuns em arqueas e períodos de cerca de 11pb prevalentes em bactérias. Esta diferença é atribuída a uma possível distinção de *supercoiling* do DNA de bactérias e arqueas: os períodos mais curtos do que a média do período helicoidal de DNA ~10,5pb podem levar à formação de superhélices à esquerda correspondentes a *supercoiling* positivo, enquanto que períodos maiores do que 10,5bp promovem superhélices à direita e *supercoiling* negativo. Com base numa análise de padrões de periodicidades no genoma de *E. coli*.

Alguns pesquisadores (Tolstorukov et al., 2005) propuseram um modelo no qual segmentos de DNA curtos dobrados estabilizam o DNA, essas dobras podem ser induzidas por proteínas de ligação do DNA ou pela sequência de periodicidade que dá origem às curvas intrínsecas e a ausência de interações DNA-proteína. Esses estudos realizaram avaliações da periodicidade de forma integral ao longo de todo o cromossoma, não avaliando a variância da periodicidade entre as diferentes regiões cromossômicas. Análises recentes (Mra´zek, 2010) de assinaturas de periodicidade, em uma diversificada coleção de genomas de procariontes, mostraram que esta periodicidade pode variar significativamente entre as diferentes regiões cromossômicas, o que sugere que as considerações de heterogeneidade intracromossômica são importantes para compreender o papel da periodicidade de sequências em genomas procarióticos.

Foi observado que a expressão gênica, ocorre, preferencialmente, em segmentos aperiódicos (Mra´zek, 2010) e quanto menor a frequência de periodicidades, maior são as taxas de transcrição. Também observou-se que as periodicidades não estão relacionadas ao ambiente, estilo de vida (a partir de vida livre ambiental para organismos simbiontes ou agentes patogênicos altamente especializados), morfologia, fisiologia, sugerindo que a periodicidade persistente não é um resultado da adaptação a um determinado ambiente ou a uma característica de um subtipo específico.

Seguramente a periodicidade mais estudada é a de 3pb codificadoras (Billing et al., 1999; Barral et al., 2000; Almeida et al., 2002) e parece estar relacionada com a evolução do

código genético, na origem, criação e organização de aminoácidos e proteínas. Em humanos, várias das doenças hoje conhecidas, têm uma característica em comum a nível molecular: a expansão de *triplets*. Dentro de qualquer das sequências codificantes dos genes atingidos, nos seus *íntrons* ou mesmo nas sequências intergênicas próximas, repetições de *triplets* em tandem são localizadas, com o número de cópias variável, tipicamente de cinco a trinta. Quando a expansão ocorre substancialmente para além do número de cópia normal das repetições, síndromes são desenvolvidas. Surpreendentemente, de dez possíveis tipos diferentes de repetições de *triplets*, um tipo domina, comum à maioria das doenças, o *triplet* GCT (GCU em mRNA).

Segundo Trifonov e Bettecken (Trifonov & Bettecken, 1997), sequências de mRNA são conhecidas por portarem um padrão periódico oculto $(GCU)_n$, que pode ser considerado um remanescente da organização da sequência de mRNA do início da sua evolução, dominada por códons de alanina e os seus derivados de mutações pontuais. Um padrão similar é característica da sequência mestre (consensual) de tRNA, descritas em 1981 por Eigen, Gardiner, Schuster e Winkler-Oswatitsch (Eigen, Gardiner, Schuster & Winkler-Oswatitsch, 1981). A sequência mestre de tRNA representa um dos primeiros mRNAs. A existência de periodicidades anormais de $(GCU)_n$ normalmente estão associadas e doenças genéticas.

3.2 Modelos Computacionais para Análise de Periodicidades

Sequências de DNA possuem um conjunto finito de elementos (A,C,G,T), que possibilita convertê-las em números, Isso permite inúmeros tratamentos probabilístico e matemáticos sobre as mesmas. Nesta seção serão descritos alguns que foram utilizados na elaboração deste trabalho.

3.2.1 Teoria da Informação de Shannom

A teoria da informação é uma formalização matemática do conceito de informação contida em mensagens (Shannon & Weaver,1949). Se uma mensagem é perfeitamente conhecida a priori, a informação nela contida é nula. Entretanto, quanto menos previsível for uma mensagem, maior a quantidade de informação nela contida.

Para compreendermos a Teoria da Informação devemos entender o conceito de entropia. O conceito de entropia foi introduzido por Shannon e Weaver (Shannon & Weaver, 1949), sendo o mais conhecido e empregado aquele que recebeu o seu nome, a Entropia de Shannon. A entropia de Shannon é uma função da probabilidade, *p(x)*, sendo usualmente descrita como:

$$H_s(X) = H(p(x_1), p(x_2), \dots, p(x_n)), \qquad \textbf{Equação 3.1}$$

onde, x é uma variável discreta aleatória de X, em um conjunto discreto $X = \{x_1, \dots, x_n\}$, com função de massa de probabilidade $p(x) = Pr(X = x)$. A entropia de Shannon de X, $H(X)$, é definida como:

$$H_s(X) = -\sum p(x) \log_2 p(x) \qquad \textbf{Equação 3.2}$$

Analisando uma sequência de n símbolos, a entropia ($H_s(X)$) é o somatório negativo das probabilidades de ocorrência de um símbolo p_i multiplicadas pelo binómio ($\log_2 p_i$). Para elucidar o conceito seguem-se alguns exemplos.

Considerando os lançamentos de uma moeda "justa", as probabilidades para cara ou coroa seriam as mesmas (50%), assim pela aplicação da fórmula de Shannon, tem-se:

$$-\left((0{,}5)(-1) + (0{,}5)(-1) \right) = 1 \qquad \textbf{Equação 3.3}$$

Considerando uma moeda "viciada", na qual as ocorrências de cara fossem caracterizadas por uma probabilidade de 75% e a face da coroa ocorresse com 25% de probabilidade, então a entropia seria mais reduzida:

$$-\left((0{,}75)(-0{,}415) + (0{,}25)(-2) \right) = 0.81 \qquad \textbf{Equação 3.4}$$

No caso do DNA, considerando o alfabeto quaternário $\Sigma=\{A,C,T,G\}$, e considerando uma sequência aleatória, sem diferenças probabilísticas entre as quatro bases teríamos:

$$-\left((0{,}25)(-2) + (0{,}25)(-2) + (0{,}25)(-2) + (0{,}25)(-2) \right) = 2 \qquad \textbf{Equação 3.5}$$

Considerando que a probabilidade de ocorrência de A ou T é de 90%, sendo de apenas 10% a probabilidade de C ou G, então a entropia seria reduzida para:

- (2(0,45)(-1,15) + 2(0,05)(-4,32)) = 1.47 **Equação 3.6**

Nas sequências naturais de DNA, as probabilidades de ocorrência das bases são muito similares, as diferenças entre elas, geralmente não ultrapassam os 5%, tendo-se, assim, elevado grau de entropia associado.

No âmbito deste trabalho, a teoria da informação de Shannom será utilizada para calcular a entropia dos 64 possíveis combinações de ACGT num triplet em porções de sequências de DNA de genomas completos de bactérias.

3.2.2 A Transformada de Fourier

A Transformada de Fourier (**FT**) além da matemática, também é muito utilizada na física, química e biologia, como ferramenta essencial na análise espectral, teoria da probabilidade, processamento de sinal, criptografia, geografia sísmica, entre outras áreas do conhecimento. Trata-se, em linhas gerais, de uma transformação integral que expande uma função real em termos harmônicos, isto é, na base senoidal ***{1, cos(x), sin(x)}***.

A transformada de Fourier permite que representemos sinais que não são periódicos ("aperiódicos"). Um sinal aperiódico pode ser visto como um sinal periódico com um período infinito.

A transformada de Fourier de uma função f(x) é definida como:

$$\mathcal{F}[f(x)] \equiv F(w_x) = \int_{-\infty}^{\infty} f(x) e^{-2i\pi w_x x} dx$$

Equação 3.7

e a transformada inversa, que recupera a função original é definida como:

$$\mathcal{F}^{-1}[F(w_x)] \equiv f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(w_x) e^{2i\pi w_x x} dw_x$$

Equação 3.8

Onde ω_x é a frequência angular, e $i \equiv\sqrt{-1}$, Para cada frequência ω_x, integramos a função $f(x)$ sobre todos os valores da coordenada *x*. Se o valor da integral for grande para esta

frequência, então o sinal tem um componente significativo nesta frequência, isto é, uma parte significativa deste sinal é composto por esta frequência.

Podemos também definir:

$$\mathcal{F}[f(x)] \equiv F(\nu) = \int_{-\infty}^{\infty} f(x) e^{-i\nu x} dx$$

Equação 3.9

e a transformada inversa, que recupera a função original é definida como :

$$\mathcal{F}^{-1}[F(\nu)] \equiv f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\nu) e^{i\nu x} d\nu$$

Equação 3.10

onde $v = 2\pi\omega_x$ é a frequência linear.

Mas deve-se lembrar que a equação I é válida se não houver pontos de descontinuidade. Caso contrário, substitui-se *f(x)* por $\frac{f(x)+f(-x)}{2}$.

3.2.2.1 Transformada Discreta de Fourier

Em casos práticos, normalmente a avaliação da transformada de Fourier não é feita utilizando-se os procedimentos analíticos, muitas vezes, não dispõe-se de uma expressão analítica para a função que se deseja analisar o espectro.

A Transformada Discreta de Fourier (DFT) é muito usada no estudo do espectro de sinais e é determinada numericamente com o auxílio do computador digital.

Considerando-se N amostras do sinal no domínio do tempo, denotadas *f(k), k=0,1,2,...,N-1*, a DFT é dada por um conjunto de *N* amostras do sinal no domínio da frequência, denotadas por *F(n), n= 0,1,2,...,N-1* e definidas por

$$F(n) \equiv \frac{1}{N} \sum_{k=0}^{N-1} f(k)\, e^{-j2\pi nk/N}$$

Equação 3.11

Diz-se então que *f(k)* ↔ *F(n)* formam um par transformado e a reobtenção do sinal no domínio do temporal pode ser feita usando a transformada inversa discreta de Fourier:

$$f(k) = \sum_{n=0}^{N-1} F(n)\, e^{j2\pi kn/N}$$

Equação 3.12

3.2.2.2 Transformada Rápida de Fourier (FFT)

O custo computacional para o cálculo da DFT é extremamente alto, o número de operações realizadas no cálculo da DFT através da definição é proporcional à N^2, isto é, para cada dos N valores de u, a expansão de $F(u)$ requer N multiplicações complexas de $x(n)$ por W^{ux} além de $(N-1)$ adições dos resultados. Exemplificando, para um N de grandeza 8.192 seriam necessárias 671.088.964 multiplicações complexas, tabela 3.1.

Visando diminuir o alto custo computacional do cálculo da DFT, Cooley (IBM) em colaboração com Tukey (Bell Labs) (Cooley & Turkey, 1965) publicaram a transformada rápida de Fourier, a FFT. Um método altamente eficiente de reagrupar os cálculos dos coeficientes de uma DFT.

No algoritmo, que implementa o método da FFT, alguns dos termos podem ser computados uma vez e armazenados para serem usados em operações futuras. Logo tais multiplicações de $x(n)$ não são consideradas na implementação. Algumas decomposições apropriadas na equação da DFT, na seção anterior definida, podem tornar o número de multiplicações e adições proporcionais a $N.Log_2N$. A tabela 3.1 compara os custos computacionais da DFT x FFT. A FFT será utilizada neste trabalho para o cálculo da periodicidade dos triplets.

Tabela 3.1 – DFTxFFT

N	N^2 (DFT)	$N.log_2$	*N* (FFT) Vantagem
2	4	2	2
4	16	8	2
8	64	24	2,67
16	256	64	4
32	1024	160	6,4
64	4096	384	10,67
128	16384	896	18,29
256	65536	2048	32
512	262144	4068	56,89
1024	1048576	10240	102,4
2048	4194304	22528	186,18
4096	16777216	49512	341,33
8192	671088964	106496	630,15

3.2.3 Transformada *Wavelet*

A transformada *wavelets* nos possibilita uma análise tempo-escala de um sinal, representando-o em termos de sinais simples construídas por translações e dilatações de uma *wavelet* de análise ψ (Chan, 1995). . A transformada em *wavelets* contínua de um sinal u é calculada a partir do produto interno do sinal com o conjugado complexo da *wavelet* de análise $\psi : R \to C$.

$$U(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} u(t)\psi^*\left(\frac{t-b}{a}\right)dt,$$

Equação 3. 13

com $a, b \epsilon R. a > 0$, é o coeficiente de escala que permite a compressão ou expansão da função *wavelet*, sendo inversamente relacionada à frequência, *b* é o coeficiente que permite a sua translação através do eixo do tempo (ou posição), e $\psi\ (t)$ é a função base *wavelet*. O fator $\frac{1}{\sqrt{a}}$ é usado para normalização da energia.

Para sinais discretos, a transformada em *wavelets* do sinal u, é definida como

$$U(b, a) = 2^{-\frac{a}{2}} \sum_{k} u(k)\psi(2^{-a}k - b),$$

Equação 3.14

onde a e b são valores discretos, e a função discreta ψ pode ser tomada como uma versão amostrada da contraparte contínua. A *wavelet* de análise ψ é geralmente escolhida para ser bem localizada em tempo e frequência. Esta função pode ser real ou complexa, resultando também em uma transformada real ou complexa. Na análise do sinal, nenhuma escala é privilegiada, pois a mesma função ψ é utilizada em diversas escalas. Portanto, a transformada mantém uma resposta fortemente dependente da função ψ (Chan, 1995).

A transformada em *wavelets* permite calcular os coeficientes dos espectros de frequência para cada posição de um sinal específico, isto é, a energia em função de posição e frequência. Os espectros de frequência derivados da transformada de Fourier (Seção 3.3), mantêm energia como função de frequência, onde toda a informação espacial é oculta ou perdida.

3.2.3.1 Transformada Modificada de Gabor-*Wavelet*

Uma transformada multiescalar de um sinal de u pode ser definida com

$$U\ (b, a) = \int u(x)\, \psi(x, b, b)dx,$$

Equação 3.15

onde *a* > 0 é o parâmetro de escala, *b* é o parâmetro tempo (ou posições) e indica a função de análise (Meno-Chaco et al., 2008).

Diferentes funções de análise podem ser adotadas para essa transformada geral. Em particular, a transformada de Fourier de Tempo-Curto (STFT) usando uma janela Gaussiana (também conhecida como transformação Gabor) é definida como (Costa and Cesar Jr, 2000).

$$\psi_{STFT}\,(x,b,a) = \; e^{-\frac{(x-b)^2}{2}}\, e^{ja(x-b)}, \qquad \textbf{Equação 3.16}$$

onde $j = \sqrt{-1}$. A função de análise na transformada de *wavelet* usa uma Gabor-*wavelet* (também chamado de Morlet-*wavelet*) é definida como (Chan, 1995):

$$\psi_{GWT}\,(x,b,a) = \; e^{-\frac{(x-b)^2}{2}}\, e^{j\omega_0\left(\frac{x-b}{a}\right)}, \qquad \textbf{Equação 3.17}$$

onde ω_0 é a frequência básica de ψ_{GWT} . A função de Gabor e a transforma Gabor-*Wavelet* são comumente adotadas para realizar a análise de frequência local, porque eles estão bem localizadas no domínio do tempo e frequência.

A função de Gabor e a transformada Gabor-*Wavelet* analisam sinais em diferentes frequências e escalas, mas de um forma dependente, ou seja, a faixa de frequência depende do conteúdo na escala de trabalho. Isto não é desejável para a análise de DNA, como as regiões codificantes são caracterizadas por uma periodicidade fixa, que podem ocorrer em diferentes escalas. Nenhuma destas transformações é adequado para tal tarefa. Uma modificação (Meno-Chaco et al., 2008) da função de análise de Gabor para analisar um sinal em uma dada frequência fixa (frequência da exponencial complexa constante), e de escala variável. A Equação-3.17 é modificada de modo que o parâmetro de escala *a* é usado para manter a frequência da exponencial complexa constante, enquanto varia o desvio padrão da Gaussiana (isto é, a escala *a*).

$$\psi_{MGWT}\,(x,b,a) = \; e^{-\frac{(x-b)^2}{2a^2}}\, e^{j\omega_0(x-b)}. \qquad \textbf{Equação 3.18}$$

Portanto, o MGWT é definido como uma função *a* de *b* e *a* como

$$U\,(b,a) = \int u(x)\, e^{-\frac{(x-b)^2}{2a^2}}\, e^{j\omega_0(x-b)}. \qquad \textbf{Equação 3.19}$$

3.2.4 As Transformadas de Fourier e Wavelet na Análise de Sequências de DNA

Na literatura inúmeros trabalhos utilizam as transformadas de Fourier e Wavelet para análises de sequências de DNA (Tiwari tt al., 1997; Audit et al., 2004; Meno-Chaco et al., 2008; Wang and Stein, 2010; Raghavan et al., 2011).

A utilização da Transformada de Fourier (Tiwari et al., 1997) para analisar a periodicidade de 3 pb no arranjo de nucleotídeos evidencia um pico acentuado em frequência f - 1/3 no espectro de Fourier. Da extensa análise espectral de sequências de DNA de comprimento total de mais de 5,5 milhões de pares de bases a partir de uma ampla variedade ou organismos (incluindo o humem), examinando-se separadamente e codificando sequências não-codificantes, descobrimos que a altura relativa do pico em f = 1/3, no espectro de Fourier é um bom discriminador potencial de codificação. Este recurso foi utilizado para detectar regiões prováveis de sequências codificantes no DNA, examinando a relação sinal-para-ruído no local do pico dentro uma janela deslizante.

Audit et al. (2014) utilizaram a transformada de *Wavelet* para realizar análises comparativa do DNA e dos perfis de flexão correspondentes, gerados com tabelas de curvatura, baseadas em dados de posicionamento dos nucleossomos. Esta exploração pela ótica da chamada “transformada microscópia de *wavelet*" revela uma escala característica de 100-200 pb que separa dois regimes de diferentes de correlações de longa distância (LRC). Estes estudos focam na existência de LRC no regime de pequena escala (200 pb). Análise de genomas de plantas, animais, fungos, protistas, bactérias e arqueas revelam que este regime está especificamente relacionado com a presença de nucleossomos. De fato, em pequena escala LRC são observados em genomas eucarióticos e em menor grau em genomas de arqueas, em contraste com a sua ausência em genomas eubacterianas. Do mesmo modo, este regime é observado em eucariotos, mas não em genomas bacterianos e de DNA viral. Há uma exceção para genomas de Poxvirus, os vírus de DNA de animais que não se replicam no núcleo da célula e não apresentam LRC em pequena escala. Além disso, LRC de pequena escala não são detectados nos genomas de todos os vírus de RNA examinados, com uma exceção, no caso de retrovírus. No seu conjunto, estes resultados

sugerem fortemente que a pequena escala são uma assinatura de LRC da estrutura nucleossômica. Finalmente, possíveis interpretações de que esta pequena escala de LRC em termos dos mecanismos que governam o posicionamento, a estabilidade e a dinâmica dos nucleossomos ao longo da cadeia de DNA.

A utilização da transformada descrita pela Equação-3.18 propõe um método para a identificação de regiões codificadoras de proteínas (Meno-Chaco et al., 2008). A transformada pode ser utilizada para identificar quaisquer regiões periódicas. O método permite o uso de múltiplas escalas, analisando tanto as pequenas regiões codificadoras com pequenas escalas e regiões codificadoras maiores, com grandes escalas. A principal vantagem deste método é, portanto, a sua robustez para dimensionar variações sobre a análise de sequências de DNA. Outra vantagem é a flexibilidade e forma da representação gráfica da periodicidade de 3pb encontrada localmente em regiões codificadoras. Esta capacidade de visualização é útil para explorar o significado biológico das regiões com periodicidade de 3pb. Pode também ser explorada para a detecção limite da região. Os resultados sugerem que a utilização da transformada modifica de de Garnor-*Wavelet* permite uma identificação mais precisa das regiões codificadoras curtas, comprado com outros métodos que utilizam transformadas de Fourier.

Raghavan et al. (2011) utilizaram duas abordagens, transformadas de Fourier e *Wavelet*, de decomposição de sinal que foram aplicados à sequências de DNA, com vista a investigação de padrões de dinucleotídeos CG. As análises utilizando Fourier destacam frequências características, presentes nas ilhas CpG e em regiões não codificantes, mas não forneceram qualquer informação sobre a localização destas regiões. Análises utilizando Wavelet deram mais detalhes sobre as frequências, mas também a sua localização ao longo da sequência de DNA.

Na elaboração deste trabalho foi utilizada a FFT para a obtenção da periodicidade 3bp (P3).

3.3 Sequenciamento de DNA

Entende-se por sequenciamento de DNA qualquer método que determina a ordem dos nucleotídeos em uma amostra. O primeiro método popular de sequenciamento do DNA foi o de terminação de cadeia de Sanger, publicado em 1975. Em 1986 foi lançado o primeiro sequenciador automático de DNA, o ABI 370, e em 1998, o primeiro sequenciador de eletroforese capilar, o ABI 3700. Com a automação, foi possível realizar grandes projetos de sequenciamento, como do genoma humano e do camundongo entre outros.

Nesta seção serão descritos alguns sequenciadores automáticos disponíveis no mercado.

3.3.1 Método de Sanger

Atualmente muito dos sequenciadores comercialmente disponíveis são baseados no método desenvolvido em 1975 por Frederick Sanger (Sanger & Coulson, 1975) conhecido como o método de rescisão de Dideoxy ou, simplesmente, Método de Sanger.

Nesse método, primeiramente deve-se separar a fita dupla de DNA em fitas simples; a amostra de fitas simples são separadas em quatro e em cada uma delas são adicionados os quatros desoxirribonucleotídeo trifosfatado (dNTP) e apenas um didesoxirribonucleotídeo trisfosfatado (ddNTP), além do primer, marcado radioativemente; a síntese da fita complementar ao molde ocorre a partir do primer marcado, com a ação enzima DNA-polimerase. Os ddNTP diferem dos dNTP pela ausência do grupo 3’-OH, ou seja, quando incorporados à fita há a interrupção de sua síntese. Pequenas quantidades de ddNTP são adicionadas e, aleatoriamente, a síntese será interrompida com o ddNTP de sua respectiva amostra, formando fragmentos de tamanhos diferentes, como demonstra a Figura 3.1. Esses fragmentos podem ser separados e analisados através de eletroforese, com posterior revelação radiológica.

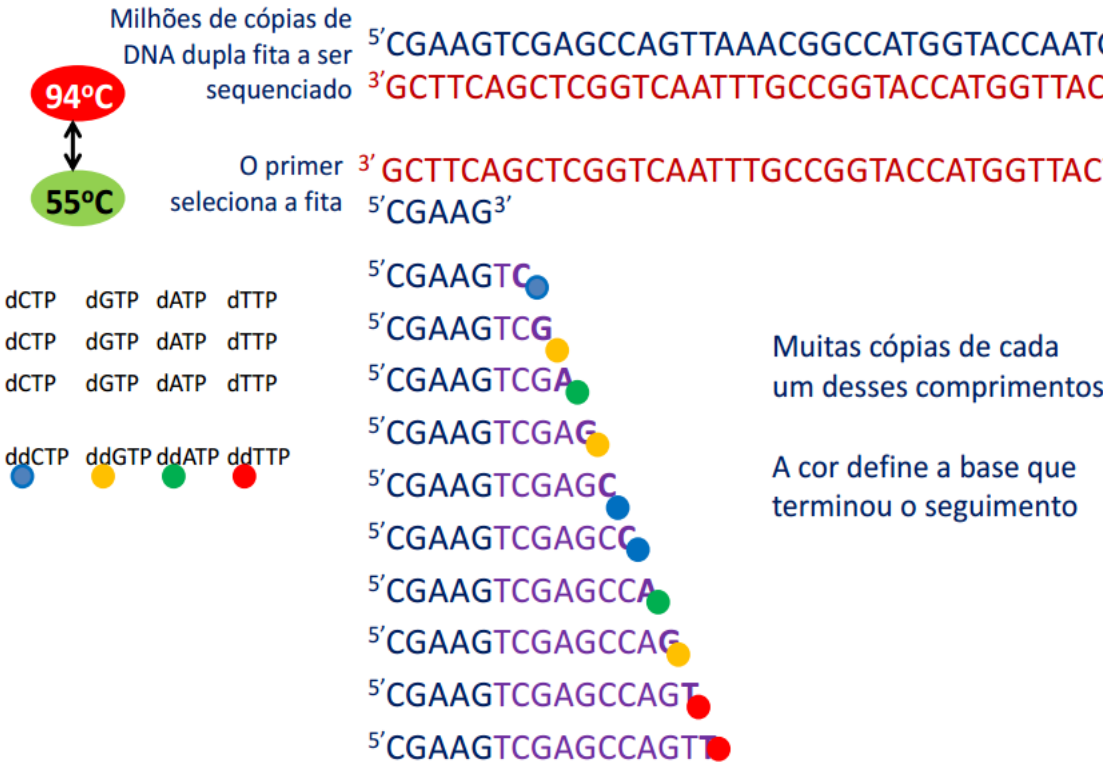


Figura 3.1 - Método de Sanger (Sanger & Coulson, 1975)

Os sequenciadores comerciais atuais, que utilizam o método de Sanger, usam marcadores fluorescente nos diferente ddNTP que, quando excitados com laser, emitem fluorescência com cores distintas de cada base, não necessitando assim das quatro reações paralelas do método original. Estes equipamentos ainda utilizam a eletroforese capilar.

Na eletroforese capilar não há a necessidade da preparação do gel de poliacrilamida, pois utiliza o processo de eletroinjeção para analisar as moléculas. Quando os fragmentos do DNA atingem a janela de detecção mostrado na Figura 3.2, um laser excita as moléculas de corante, cuja fluorescência emitida é capturada pela câmara CCD. Um software interpreta os resultados, atribuindo uma cor específica para cada base marcada. Por convenção, A fica como verde, C como azul, G como preto e T como vermelho, como demonstra a Figura 3.3.

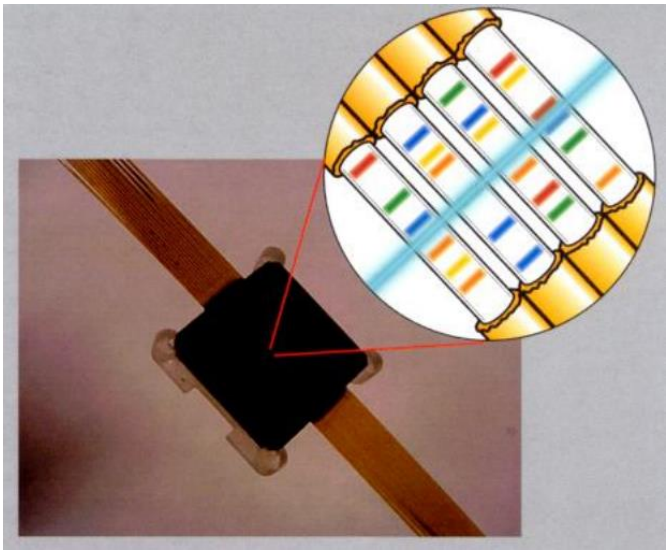


Figura 3.2 - Eletroforese Capilar

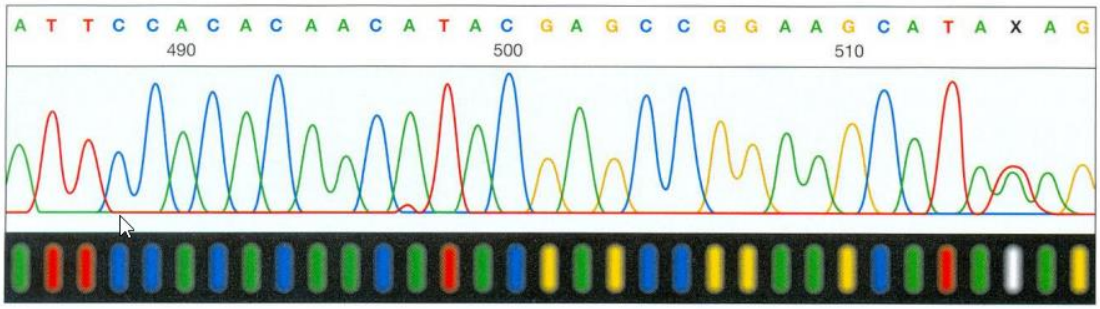


Figura 3.3 - Convenção de Cores dos Nucleotídios

Os sequwnciadores automáticos, como o 3730*xl* DNA Analyzer (https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catID=601642), possui tecnologia baseada no método de sequenciamento de Sanger, realizado através de eletroforese capilar, podendo gerar sequências de boa qualidade próximas a 1 kb de tamanho, com 98,5% de acurácia. O sistema está equipado com 96 capilares de 50 cm de extensão e pode comportar amostras contidas em 10 placas de 96 poços empilhadas em seu compartimento de alimentação. São utilizados 3 modos de performance (Fast, Long e XL-extended) com duração

de tempo de corrida variando de 1 a 3 h por placa que geram leituras de 400-900pb com alta confiança. O sistema utiliza o polímero POP-7 durante a eletroforese capilar. O sistema óptico deste equipamento é constituído por câmera CCD, juntamente com vários filtros.

Neste trabalho, para efeitos de validação da metodologia, serão utilizadas sequencias de nucleotídeos semelhantes a aquelas produzidas pelos sequenciadores que utilizam o método de Sanger.

3.3.2 Método de Pirossequenciamento

Este método é um dos novos processos de sequenciamento que surgiram nos últimos anos, conhecidos com NGS, em inglês *Next Generation Sequencing*. Este método é baseado no princípio da síntese de sequências, baseando-se na detecção de pirofosfatos liberados na incorporação de nucleotídeos. Este método foi desenvolvido por pesquisadores do *Royal Institute od Technology* de Estocolmo, Suécia em 1996 (Ronaghi et al., 1996).

A sequência de DNA desejada pode ser determinada pela luz que é emitida na incorporação do próximo nucleotídio complementar, considerando que apenas um dos quatro possíveis nucleotídeos A, T, C ou G são adicionados e disponíveis ao mesmo, de modo que apenas um nucleotídeo pode ser incorporada no molde de cadeia única (o que é a sequência a ser determinada). É a intensidade da luz que determina se há mais do que uma destas "letras" consecutivas. O nucleotídeo anterior (um de quatro possíveis dNTP) é degradado antes do nucleotídio seguinte ser adicionado para síntese, permitindo revelar os nucleotídeos incorporados na sequência, através da intensidade da luz resultante, como exemplifica a Figura 3.4. Este processo é repetido com cada uma dos quatro letras até que a sequência de DNA de cadeia simples seja determinado.

O sequenciador comercial, GS FLX+ System da empresa suíça Roche (http://454.com/seewhatspossible/#performance), estende um pouco a limitação de 850pb dos primeiros sequenciadores deste método. Permite ler sequências acima de 1100 pb, com um desempenho maior que 99% de acurácia, segundo o fabricante.

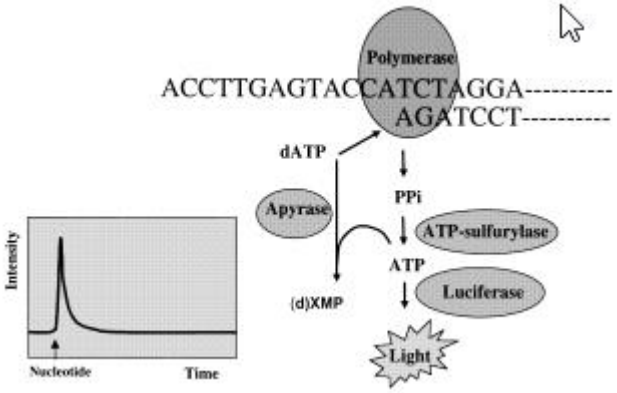


Figura 3.4 - Pirosequenciamento

3.3.3 Método de Ion Torrent

Estes método também faz partes dos NGS, tendo como objetivo acelerar e baixar o custo do processo de sequenciamento. Seus processos se baseiam no processamento paralelo massivo de fragmento de DNA, tendo capacidade de ler até bilhões de fragmentos ao mesmo tempo. Quase a totalidade dos métodos de sequenciamento possuem as etapas de preparo da amostra, amplificação da biblioteca e sequenciamento.

No preparo das amostras, o DNA é fragmentado por um processo químico, mecânico ou enzimático, Figura3.5-a. Cada um desses fragmentos é chamado de *template*. Independentemente do método de fragmentação escolhido, é importante que ele quebre o DNA de maneira aleatória, de forma que todo o genoma seja coberto de maneira mais uniforme possível. Na sequência, adaptadores e sequências artificias conhecidas são incorporadas ao *template*. Preparando-se, assim, a biblioteca que será amplificada.

A etapa de amplificação de bibliotecas objetiva gerar milhares de cópias de cada fragmento de DNA produzido anteriormente (Figura3.5-b). O objetivo é aumentar a fonte de íons, no caso do Ion Torrent, que será detectado na etapa de sequenciamento (Figura3.5-c).

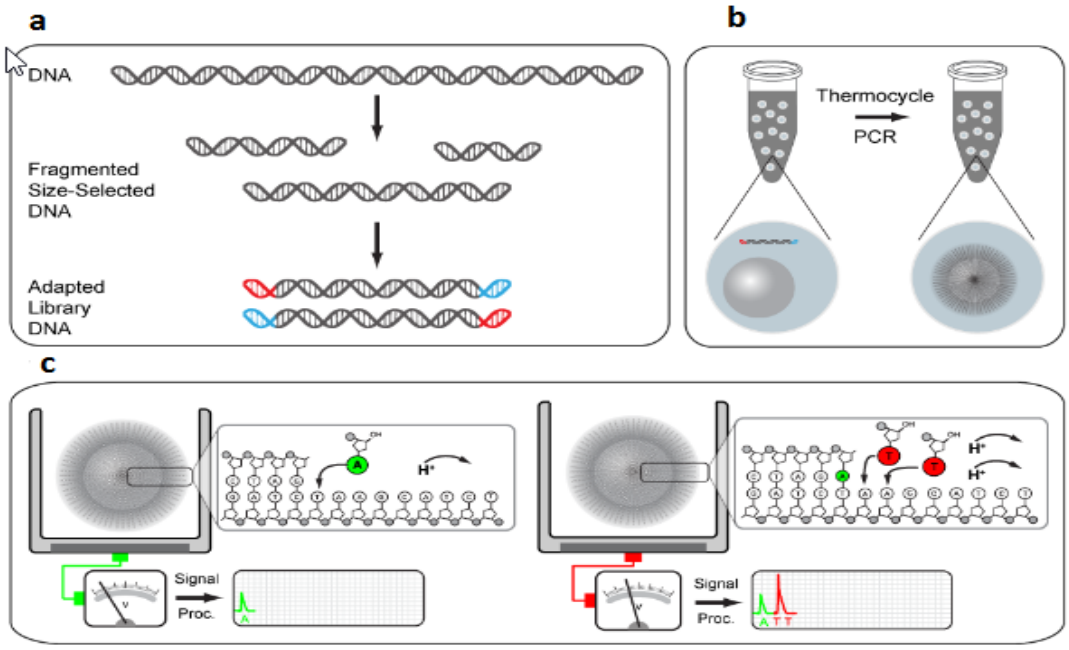


Figura 3.5 - Método de Sequenciamento (Rothberh et al., 2011)

No sequenciamento pelo método de Ion Torrent (Rothberh et al., 2011), ao contrário da maioria dos sequenciadores que utiliza DNA-polimerase para gerar fitas complementares aos *templates*, bases marcadas e câmeras de detecção, a detecção é feita diretamente, já que a reação de polimerização gera naturalmente um H^{+3}, ou seja um próton que altera pH do meio. Essa alteração do pH é detectada por um transistor ISFET (Bergveld, 1970) e convertida num sinal elétrico.

Na implementação do sequenciador baseado neste método (Rothberh et al., 2011), foi desenvolvida uma arquitetura onde milhões de sensores ISFET simples são colocados num circuito integrado do tipo CMOS, como mostra a figura 3.6. O circuito integrado é composto de um grande *array* de elementos de sensores, cada um com uma única porta flutuante ligada a um ISFET subjacente (Figura 3.7-a). Para a sequência de confinamento que dependem de 3,5 mícrons de diâmetro assim formado pela adição de uma camada dieléctrica de 3 mícrons de espessura sobre a placa eletrônica e gravando o sensor sobre ele, figura 3.7-b. Uma camada de

óxido de tântalo é adicionada para prover sensibilidade aos prótons. As leituras em alta velocidade realizadas pelo sistema eletrônico de semicondutores integrados com a matriz de sensor, figura 3.7-c. O sensor e a eletrônica embarcada fornecem uma tradução direta do evento de incorporação para um sinal eletrônico. Ao contrário da tecnologia de sequenciamento baseado em luz, não são usados os elementos do *array* para coletar fótons e formar uma imagem maior para detectar a incorporação de uma base. Em vez disso, cada sensor é usado para monitorar de forma independente e diretamente os íons de hidrogênio liberados durante a incorporação de nucleotídeos.

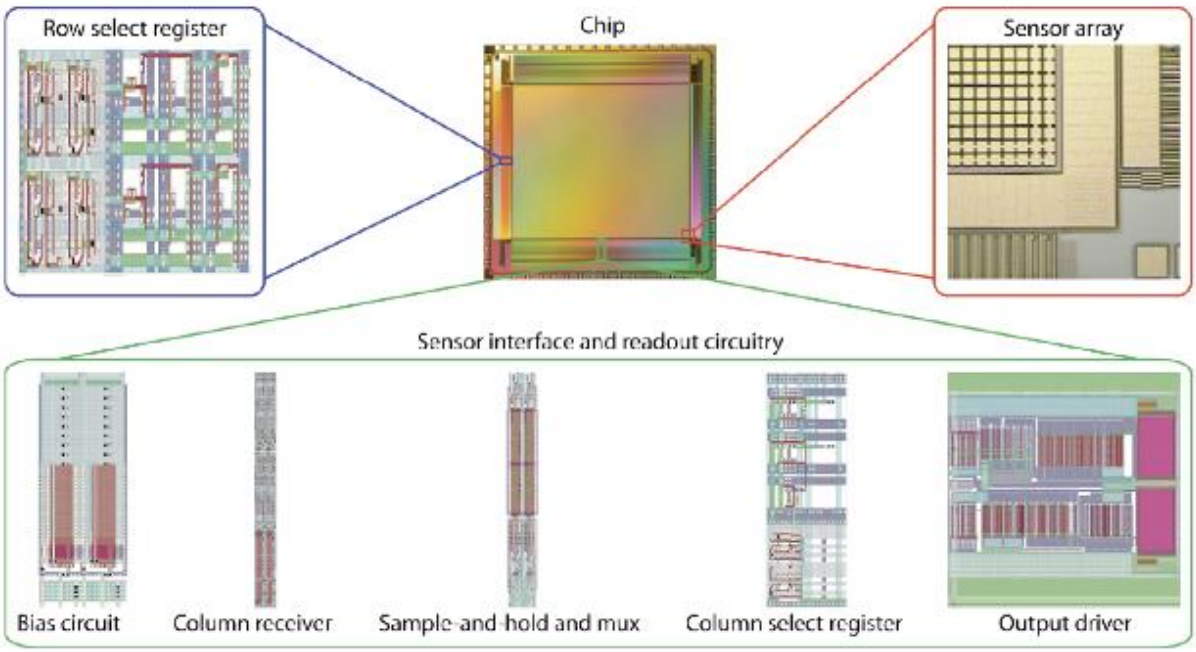


Figura 3.6 - Blocos Funcionais de um *Ion Chip* (Rothberh et al., 2011)

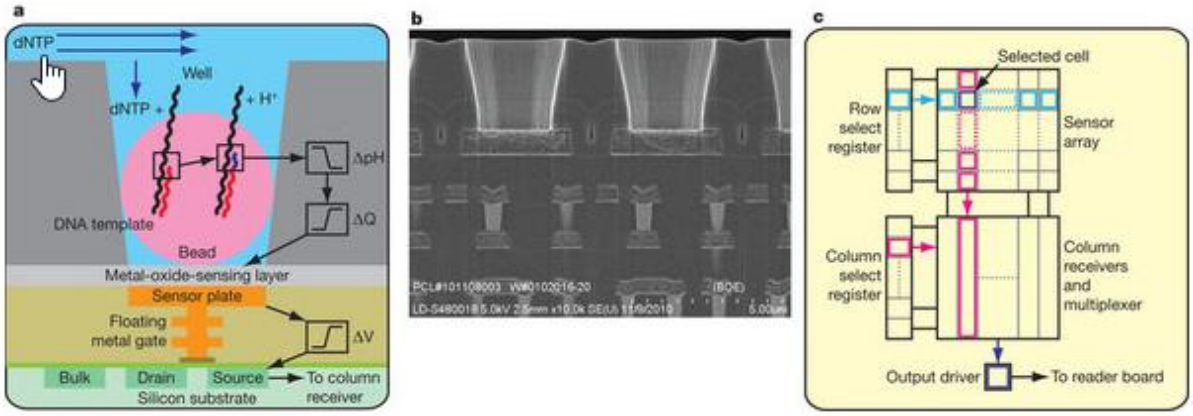


Figura 3.7 - Sensor de Ions (Rothberh et al., 2011)

Por esse método são sequenciadas no máximo 400 pb por ciclo, com uma acurácia de 99,9%. Os sequenciadores atuais que utilizam o método de Sanger, por exemplo, obtém por ciclo um maior número de pares de bases, o que é esperado para sequenciamentos de genomas grandes. A grande vantagem dos sequenciadores *Ion Torrents* é o baixo custo para a obtenção das sequência resultantes. Além disso, por utilizarem microprocessadores de DNA, estão sujeitos a lei de Moore (Moore, 1965) e, portanto, sofreram um aumento exponencial de desempenho associados a um custo cada vez menor.

3.4 Metagenômica

O termo "metagenoma" foi utilizado inicialmente para descrever a sequência genômica de comunidades inteiras e é relativamente novo (Covacci et al., 1997). Já o termo "metagenômica" foi usado pela primeira vez por Handelsman no artigo "*Molecular biology provides access to the chemistry of unknown soil microbes: a new frontier for natural products*" de 1998 (Sharma et al., 2012). O termo metagenômica referencia a ideia de que um conjunto de genes sequenciados a partir do ambiente pode ser analisado de uma maneira análoga à de um estudo genoma único. O crescimento do interesse em genética ambiental, resultou numa maior utilização de metagenômica para descrever qualquer sequenciamento do material genético de amostras ambientais (ou seja, sem cultura), até mesmo trabalhos que incidem sobre um organismo ou gene. Recentemente, Kevin Chen e Lior Pachter (pesquisadores da Universidade da Califórnia , Berkeley) definiram metagenômica como "a aplicação de técnicas modernas de genômica para o estudo de comunidades de microrganismos

diretamente em seu habitat natural, não sendo necessário isolar e cultivar em laboratório as espécies individuais (Chen & Pacther, 2005).

Segundo Handelsman (2007) existem duas abordagens metodológicas paralelas para metegenoma. Na primeira abordagem, conhecida como *sequence-driven metagenomics* (orientada à sequência), o DNA a partir do ambiente de interesse é sequenciado e sujeito à análise computacional, Figura 3.8. As sequências metagenômicas são comparadas com sequências depositadas em bases de dados disponíveis publicamente, como o GENBANK . Os genes são, então, agrupados em grupos de funções similares previstas, e a distribuição das várias funções e tipos de proteínas que sintetizam estas funções podem ser avaliadas.

Na segunda abordagem, conhecida como *metagenomics-driven* (orientada à função), o DNA extraído a partir do meio ambiente é captado e armazenado também num hospedeiro substituto, mas em vez de sequenciar, as janelas de fragmentos capturados de DNA , ou clones, para uma certa função, figura 3.8. A função deve estar ausente no hospedeiro substituto para que aquisição da função possa ser atribuída ao DNA metagenômico.

Estas abordagens têm vantagens e limitações, e ambas são eficazes em nos informar sobre a diversidade do mundo microbiano. A abordagem orientada à sequência é limitada pelo conhecimento existente: se uma sequência metagenômica não se parece com uma sequência conhecida depositada nos bancos de dados, pouco pode ser definido sobre a sequência. A análise orientada à função pode identificar os genes que não estão relacionadas a qualquer coisa previamente estudada porque os genes são identificados pela sua função expressa, em vez de pela sequência, mas a desvantagem é que a maioria dos genes de organismos em comunidades selvagens não podem ser facilmente expressados por um determinado hospedeiro substituto. Portanto, as duas abordagens são complementares e devem ser realizadas em paralelo.

Além de fornecer ferramentas para estudar a maioria dos microrganismos não cultiváveis, metagenômica apresenta uma nova maneira de pensar. A Genômica trouxe consigo uma nova abordagem para a genética humana que se baseia no conhecimento de todos os genes do organismo e a compreensão de algumas das interações entre os genes e seus produtos. Genes podem ser abordados como nós em uma rede que modula as atividades e sucesso do hospedeiro. Da mesma forma a metagenômica apresenta uma visão em nível do sistema de

comunidades microbianas. Em vez de estudar organismos individuais ou funções individuais, a metagenômica examina a interação de organismos e genes em uma comunidade.

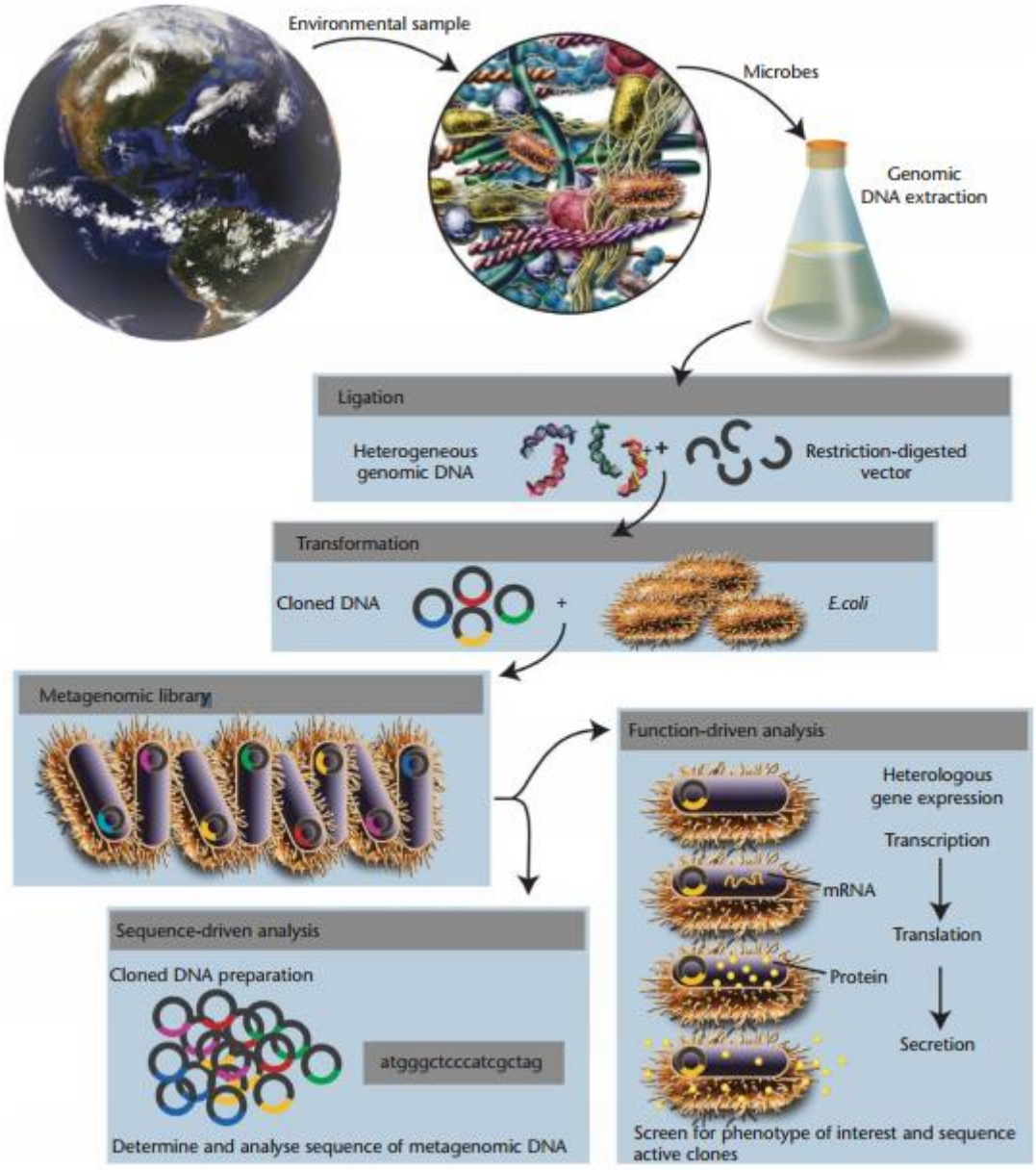


Figura 3.8 – Construção e Análise de Bibliotecas de Metagenomas (Handelsman, Jo, 2007)

4 MATERIAS E MÉTODOS

4.1 Organismos Estudados

Para obtenção dos dados que serão utilizados no método de análise (Capítulo 5), proposto por essa tese, foram selecionados somente organismos procariontes. Utilizou-se 1970 organismos que possuíam sequências de DNA com genomas completos disponíveis até então, totalizando 1580 espécies, disponíveis na maior base pública de sequências genéticas, o GenBank (BENSON et al., 2013). As informações e sequências genômicas estão disponíveis no endereço de internet: http://www.ncbi.nlm.nih.gov.

4.2 Ferramentas

Na construção das ferramentas utilizadas para a análise genômicas e metagenômica, na confecção da base de dados, armazenamento e recuperação de dados utilizou-se software livre (http://opensource.org) ou de uso livre para pesquisas acadêmicas. São eles:

- **MySql:** é um sistema de gerenciamento de banco de dados (SGBD), disponível em http://dev.mysql.com/downloads/. Atualmente é o terceiro SGBD mais utilizado no mundo, com mais de dez milhões de usuários. É um Software Livre com base na GPL, entretanto, se o programa que acessar o Mysql não for GPL, uma licença comercial deverá ser adquirida (http://docs.oracle.com/cd/E19957-01/mysql-refman-6.0/index.html#legalnotice)

- **MetaSIM** (Richter et al., 2008)**:** software simulador de metagenomas. O software pode ser utilizado para gerar conjuntos de leituras sintéticas que refletem a composição taxonômica diversificada de conjuntos típicos de dados metagenômicos. Com uma base de dados de genomas, é possível desenharmos um metagenoma, especificando o número de genomas presentes em diferentes níveis da taxonomia NCBI, gerando metagenomas que simulam sequências geradas pelas tecnologias de sequenciamento. Os conjuntos de dados resultantes podem ser usados como cenários de testes padronizados, para o planejamento de projetos de sequenciamento, para elaboração de benchmarking ou softwares metagenômicos que realizam alinhamento de sequências biológicas.

Possui um banco de dados de sequências genômicas, permitindo gerar conjuntos sintéticos baseado em modelos adaptáveis de erro de sequenciamento Sanger (por exemplo, para a química, Roche do 454 e Illumina), permite que o usuário configure parâmetros de abundância para cada organismo modelando composições taxonômicas específicas, fornece uma amostra da população para gerar sequências evolutivas baseados em genomas de origem e de uma dada árvore evolutiva. Ainda pode ser controlado através de uma interface gráfica ou em modo de linha de comando.

- **C** (Kernighan & Ritchie, 1983)**:** linguagem de programação utilizada para a desenvolvimento do software de elaboração da base metagenômica e análises dos metagenomas utilizados. Atualmente é a linguagem de programação mais utilizada no mundo (Programming Language Popurity, 2013), foi escolhida para este desenvolvimento principalmente por questões de desempenho, onde o mesmo era fator a ser considerado em cada nova construção da base de dados. Exemplificando, a construção da base de dados, numa estação de trabalho padrão, consumia vinte e dois dias de processamento.

- **R:** software para manipulação e análise de dados. Este software permite a realização de análise estatística, cálculo de FFT, entre outras funções. Disponível em http://www.r-project.org.

- **iTOL:** *Interative Tree of Life* (Letunic & Bork, 2006) é uma ferramenta online para a visualização e manipulação de árvores filogenéticas. Ela fornece um *layout* de árvore circular, o que o torna fácil de visualizar árvores de médio porte (até vários milhares de folhas). As

árvores podem ser exportados para diversos formatos gráficos, tanto bitmap ou vetorial. Existem várias árvores pré-geradas, disponíveis para visualização, incluindo a principal árvore da vida, descrita em Ciccarelli, et al. (2006). A ferramenta nos permite carregar e exibir árvores de organismos gerados pela ferramenta de análise metagenômica desenvolvida nesta tese, utilizando a página 'upload de dados ", ou através da conta pessoal e usuário.

5 RESULTADOS E DISCUSSÃO

Neste capítulo serão detalhados os métodos de criação de metadados (Figura 5.1) e análise (Figura 5.8). Também serão descritas as análises realizadas para a validação destes métodos e os resultados obtidos.

5.1 Processo para a Criação dos Metadados

Entende-se por **metadados**, a assinatura de cada organismo contendo as métricas que serão utilizadas pelo método proposto, clusterização (d), entropia (s), periodicidade (p3) e o percentual de GC (%gc) que serão detalhas nas próximas seções, para as análises das sequências genômicas.

Este processo (Figura 5.1) está dividido em oito macro atividades. São elas:

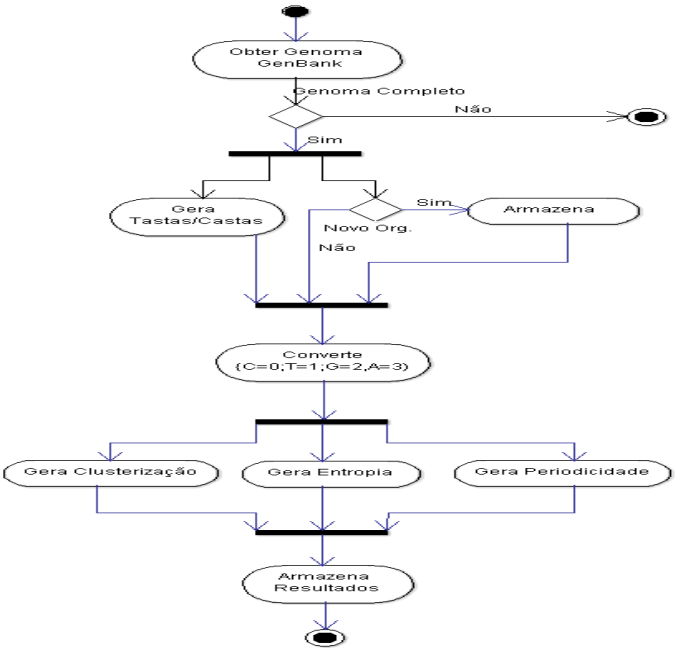


Figura 5.1 Fluxo do Processo de Criação dos Metadados

5.1.1 Obter genomas no GenBank

É realizado um FTP (*file transfer trotocol*) do repositório de sequências genômicas de procariontes do GenBank, diretamente do sítio da NCBI. São descartadas todas as sequências que não correspondem a genomas completos de organismos, como de plasmídeos, por exemplo.

5.1.2 Gerar Arquivos Castas e Tastas

As sequências genômicas anteriormente selecionas foram divididas em, aproximadamente, treze milhões e quinhentos mil arquivos de mil bases cada. Este tamanho é para manter similaridades com as sequências geradas por sequenciadores comerciais. O algoritmo responsável por gerar estes arquivos gera dois tipos de arquivos. Um com extensão **casta** com sequências *no sentido do genoma depositado* com tamanho de mil nucleotídeos, e outro, com extensão **tasta** com sequências *no sentido inverso ao genoma depositado*. Em nenhum dos casos as sequências foram invertidas.

A figura 5.2 mostra um arquivo **casta** padrão contendo mil nucleotídios.

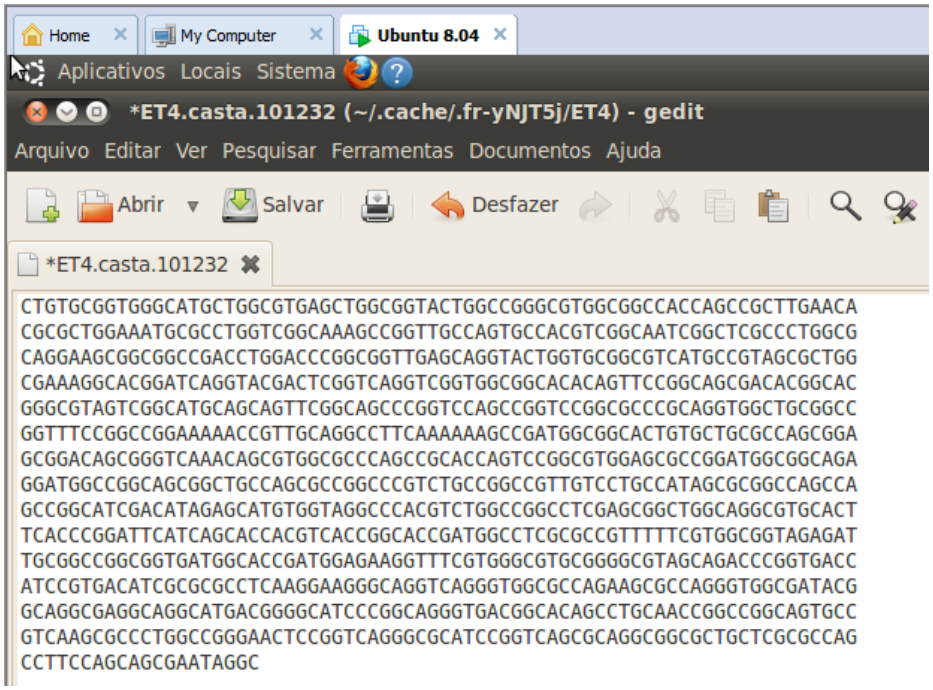


Figura 5.2 – Exemplo de Arquivo Casta Contendo 1000 Nucleotídios

5.1.3 Armazenar Organismo

Se os arquivos castas/tastas pertenciam a um organismo ainda não analisado, a descrição e a taxonomia do organismo foram armazenadas. Os dados descritivos armazenados incluíram: linhagem/isolado, espécie, gênero, família, ordem, classe e filo. Um exemplo dos dados descritivos armazenados são apresentados na Figura 5.3. Detalhes da estrutura da base de dados serão descritos na seção 5.4.

locus	nome	sequencia	tax_id	especie	genero	familia	ordem	classe	filo
NC_014750	Marivirga tractuosa DSM 4126 uid60837	0	643867	Marivirga tractuosa	Marivirga	Flammeovirgaceae	Cytophagales	Cytophagia	Bacteroidetes
NC_014497	Candidatus Zinderia insecticola CARI uid52459	6976200	522306	Accumulibacter phosphatis UW-1	Candidatus Accumulibacter phosphatis Hesselmann et...	Candidatus Accumulibacter Hesselmann et al. 1999	unclassified Betaproteobacteria	Betaproteobacteria	Proteobacteria ...
NC_017033	Frateuria aurantia DSM 6220 uid81775	2609400	81475	Acetobacter aurantium (sic) Kond and Ameyama 1958	Frateuria	Lysobacteraceae	Xanthomonadales	Gammaproteobacteria	Proteobacteria ...

Figura 5.3 – Exemplo dos Dados dos Organizamos Armazenados

5.1.4 Converter Arquivos

Os arquivos castas e tastas são convertidos da seguinte forma: nucleotídeos **C**itosina são convertidos para zero, nucleotídeos **T**imina são convertidos para um, nucleotídeos **G**uanina são convertidos para quatro, e finalmente, nucleotídeos **A**denina são convertidos para três. Esta atividade prepara os arquivos para os cálculos que serão executados nas próximas atividades.

5.1.5 Gerar Clusterização

Numa perspectiva matemática um gráfico *G* é determinado por dois conjuntos *G = G (V, K)*, onde *V* é um conjunto de vértices e *K* um conjunto de links (Günther et al., 2006). O número de vértices da rede é *N* e o número de ligações de rede é *L*. A sequência de DNA é definido como: vértices são triplets de nucleotídeos e uma ligação entre dois triplets é estabelecida se dois triplets são justapostos em algum lugar na sequência de DNA. Consideramos aqui janelas de tamanho *L* definidas ao longo da sequência.

A matriz adjacente *m* é um gráfico de uma matriz quadrada de tamanho *N*. O elemento *mi, j* é 1, se o vértice *i* está ligado o vértice j e 0 caso contrário. Com o objetivo de quantificar a organização hierárquica destes elegemos gráficos, é definido o coeficiente de clusterização.

Este parâmetro é uma quantidade global, obtida a partir de um número local *ci* que mede a fracção de pares de vizinhos de um nó, que também são vizinhos um do outro. Suponha-se que o vértice *i* está ligado ao conjunto de vértices ç. As ligações são contadas dentro deste subgrafo da seguinte forma:

$$\mathcal{L} - \sum_{j-1}^{L} m_{i,j} \left[\sum_{k \in ç} m_{i,j} \right] \quad \textbf{Equação 5.1}$$

Para encontrar *ci* é normalizado *Li* ao número máximo possível de ligações entre os vértices *ki* ligado ao vértice *i*, isso significa:

$$c_i - \frac{2\mathcal{L}_i}{k_i(k_i - 1)} \qquad \textbf{Equação 5.2}$$

Quando *ki* = 0 ou 1, definimos $c_i = 0$. Finalmente *C*, o coeficiente de clusterização para o gráfico é obtido através de uma média:

$$C = \frac{1}{L}\sum_{i-1}^{L} c_i \qquad \textbf{Equação 5.3}$$

O coeficiente de clusterização de cada um dos arquivo casta e tasta é obtido pela equação 5.3, onde *L* é o tamanho da janela da sequência de DNA utilizada. Neste trabalho foram utilizadas janelas de duzentos e cinquenta nucleotídeos conforme Gerhardt et al. (2006).

5.1.6 Gerar Entropia

A entrada desta atividade são os arquivos castas e tastas, é gerando a entropia de prevalência dos triplets de cada um destes arquivos. A entropia é definida pela clássica entropia de Shannom como:

$$S = -\sum_{i=1}^{64} P_n log P_n \,, \qquad \textbf{Equação 5.4}$$

onde *Pn* é a probabilidade do nth triplet em uma determinada sequência de tamanho W. Se o tamanho da janela não é um múltiplo de três, usamos uma condição de contorno periódica.

Equação (5.4) dessa forma é, naturalmente, influenciada pelo conteúdo GC em si. Assim, podemos definir uma entropia normalizada como:

$$Sn = S - S_{GC}(rand), \qquad \text{Equação 5.5}$$

onde S_{GC} *rand ()* é a entropia calculada por uma sequência aleatória com conteúdo GC. Este procedimento garante que a equação 5.5, irá medir informações sobre a organização do triplet não correlacionadas com conteúdo GC e pode ser considerada como uma medida de correlação com AT complexos e apresentam assimetria CG em sequência e evita levar em conta o degeneração natural dos aminoácidos.

5.1.7 Gerar Periodicidade

No estudo das repetições de DNA existem algumas técnica importantes para transformar uma sequência de letras de quatro possibilidades {ATCG} em uma sequência de números que preserva as características de frequência relevantes do conjunto preliminar de letras. Inicialmente, usamos uma função correlação *Fc (n)* do tipo

$$Fc(n) = \sum_{I=0}^{W} \frac{\delta_{i,i+1+n}}{L-i} \qquad \textbf{Equação 5.6}$$

que realiza a contagem das correlações presentes na sequência. δ é uma função delta de Kronecker (gap) entre dois nucleotídeos na posição i e i + 1 + n.

Usando a equação 5.6 acima, observamos as variações das repetições que aparecem na sequência. As oscilações que ocorrem na equação 5.6 representam repetições em sequências. Para descobrir a contribuição de cada período, podemos usar qualquer tipo de análise espectral. Neste problema específico, uma transformada de Fourier pode resolver as frequências envolvidas. Sua definição clássica é:

$$P\{F_c(n)\}(\omega) = \left| \int dn F_c(n) e^{i\omega n} \right|^2, \qquad \textbf{Equação 5.7}$$

onde ω representa a frequência. Para cada sequência foi coletado o valor de P (ω = 1 = 3) (P3) e utilizada como uma medida de caracterização.

5.1.8 Armazenar Resultados

A figura 5.4 demonstra os valores de clusterização (d), entropia (s) , periodicidade (p3) e o percentual de GC (%gc) obtidos de cada um dos 562.781 arquivos **CASTAS** e **FASTAS** de

E.coli. Observa-se que os valores, na sua maioria, mantêm-se constantes em cada segmento do genoma analisado.

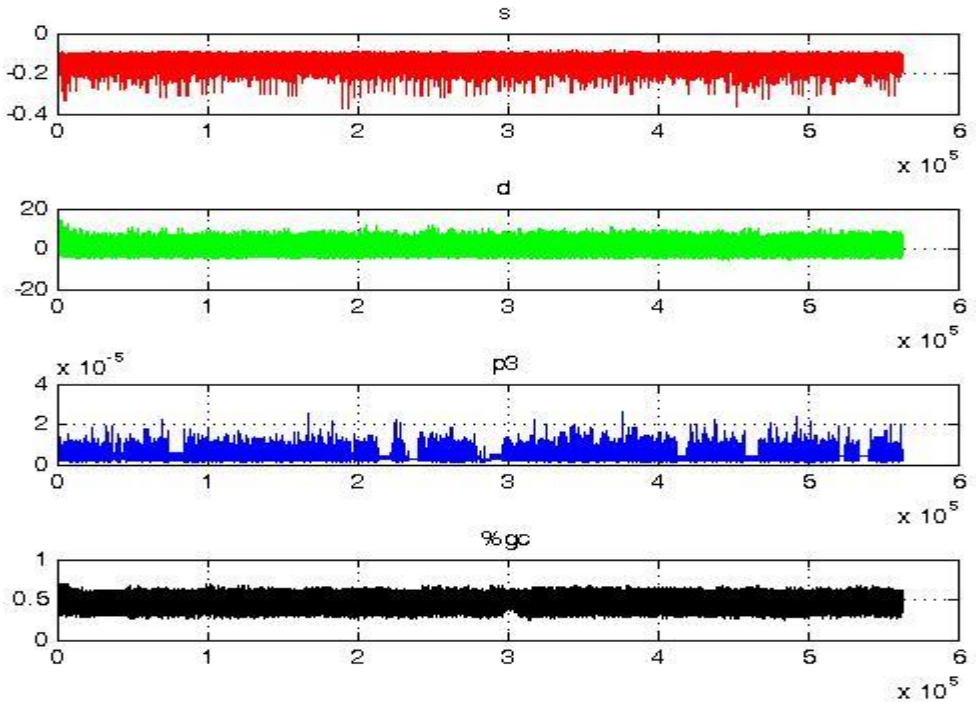


Figura 5.4 - Resultados de S,D,P3 e %GC para cada segmento de 1000 bases de *E.coli*

O somatório dos valores encontrados (clusterização, entropia e periodicidade), exemplificado na figura 5,4, é calculada a média aritmética, a mediana e o desvio padrão para cada um destes valores. Para cada genoma completo analisado é armazenada a tupla[1]:

<locus,pares de base, sequencia, s_medio, s_desvio, s_mediana, d_medio, d_desvio, d_mediana, p3_medio, p3_desvio, p3_mediana, %GC, data_inclusão, tipo, desvio %GC>

Onde:

- **locus** é a codificação da entrada do organismo no GenBank;
- **pares de base** é o total de nucleotídeos do genoma;
- **sequência** é um número inteiro;
- **s_meio** é valor médio da entropia de Shannon calculada;
- **s_desvio** é o desvio padrão da entropia;

[1] Linha formada por uma lista ordenada de colunas representa um registro no banco de dados.

- **s_mediana** é a mediana da entropia;
- **d_medio** é o valor médio do coeficiente de clusterização;
- **d_desvio** é o desvio padrão do coeficiente de clusterização;

d_mediana é a mediana do coeficiente de clusterização;

- **p3_medio** é o valor da média da periodicidade;
- **p3_desvio** é o desvio padrão da periodicidade;
- **p3_mediana** é o mediana da periodicidade;
- **%GC** é a média dos percentuais de nucleotídeos GC;
- **Data_inclusão** é a data que o organismo foi inserido na base;
- **tipo** identifica se é uma genoma, tasta ou casta
- **Desvio_%GC** é o desvio padrão do %GC.

A Tabela 5.1, exemplifica as quatro métricas (s, a entropia; d, o coeficiente de clusterização; p3, a periodicidade e o % de GC) com como seus respectivos desvios padrão. Para efeitos de visualização foram ocultadas as colunas s_mediana, d_mediana, p3_mediana, data_inclusão e tipo.

Tabela 5.1 - Exemplo de Análise de Organismo Armazenada

Locus	**s_medio**	s_desvio	**d_medio**	d_desvio	**p3_medio**	p3_desvio	**percGC**	desvio_gc
NC_019702	-0,13064	0,013683	-1,348859	1,2048196	4,18E-06	1,12E-07	0,4061953	0,0300263
NC_013791	-0,133912	0,015532	-1,105623	1,2473736	2,76E-06	7,48E-08	0,4026769	0,0322474
NC_013162	-0,136165	0,01461	-1,10277	1,2657501	3,36E-06	1,39E-07	0,3958782	0,0396186
NC_011729	-0,136088	0,016854	-1,049473	1,357752	4,72E-06	1,30E-07	0,3860876	0,0407744
NC_019689	-0,130685	0,013859	-1,047938	1,3711347	3,85E-06	1,09E-07	0,4519283	0,0442737
NC_019683	-0,134495	0,019329	-1,022962	1,3744358	4,15E-06	1,15E-07	0,4386651	0,0349902

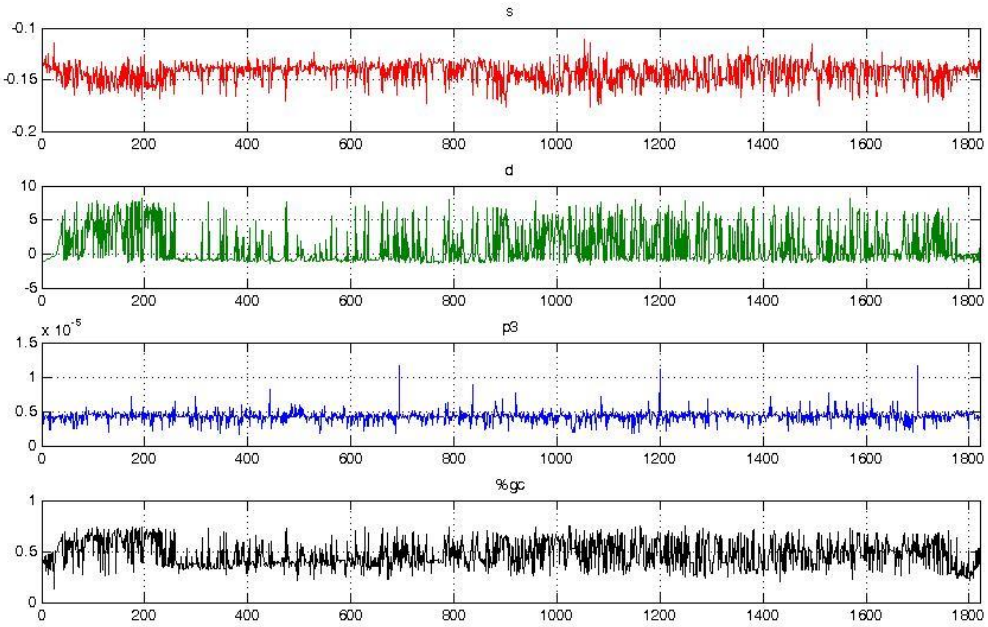


Figura 5.5 - Resultados da Entropia (s), Clusterização (d), Periodicidade (p3) e Percentual de GC (%gc) por taxionomia de todos os 1970 organismo que compõem a base de metadados

A figura 5.5 mostra os resultados das métricas que serão utilizadas nas análises de sequências genômicas. Onde o gráfico s demostra a entropia, o gráfico d o coeficiente de clusterização, o gráfico p3 a periodicidade e finalmente //o gráfico %gc o percentual de GC. O eixo x, comum aos 4 gráficos, representa os genomas dos organismos procariontes analisados que compõem o banco de dados, onde 1, o primeiro valor de x é um organismo do gênero *Propionibacterineae*, da família *Actinomycetales*, da ordem *Actinobacteridae*, da classe *Actinobacteria* e do filo *Actinobacteria* .O maior valor de x é 1821 sendo um organismo do gênero *Petrotoga*, da família *Thermotogaceae*, da ordem *Thermotogales*, da classe *Togobacteria* e do filo *Thermotogae*.

Observa-se, analisando-se os gráficos da Figura 5.5, que quando os organismos são classificados pela sua taxionomia, possuem valores muito próximos nas quatro métricas (s,d, p3 e %gc).

5.2 Base de Dados

A base de dados demonstrada na Figura 5.6, implementada em MySql, é simples, contendo somente três tabelas. A tabela Organismo para armazenar os organismos e sua taxonomia. A tabela Metadados para armazenar os metadados dos organismos que nos servem de assinatura dos mesmos e a tabela Análise, que possui a mesma estrutura da tabela Metadados, para armazenar as análises realizadas.

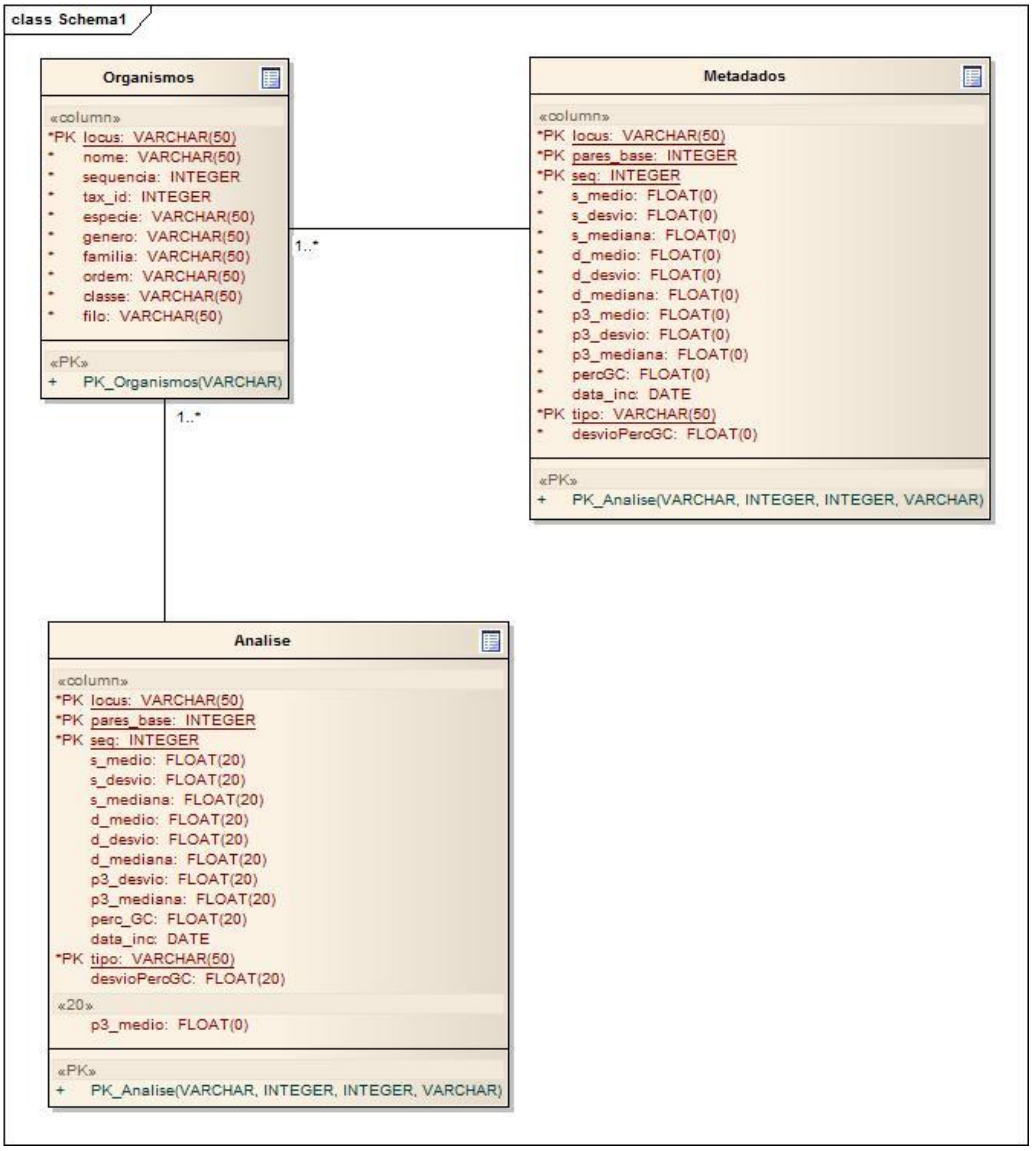
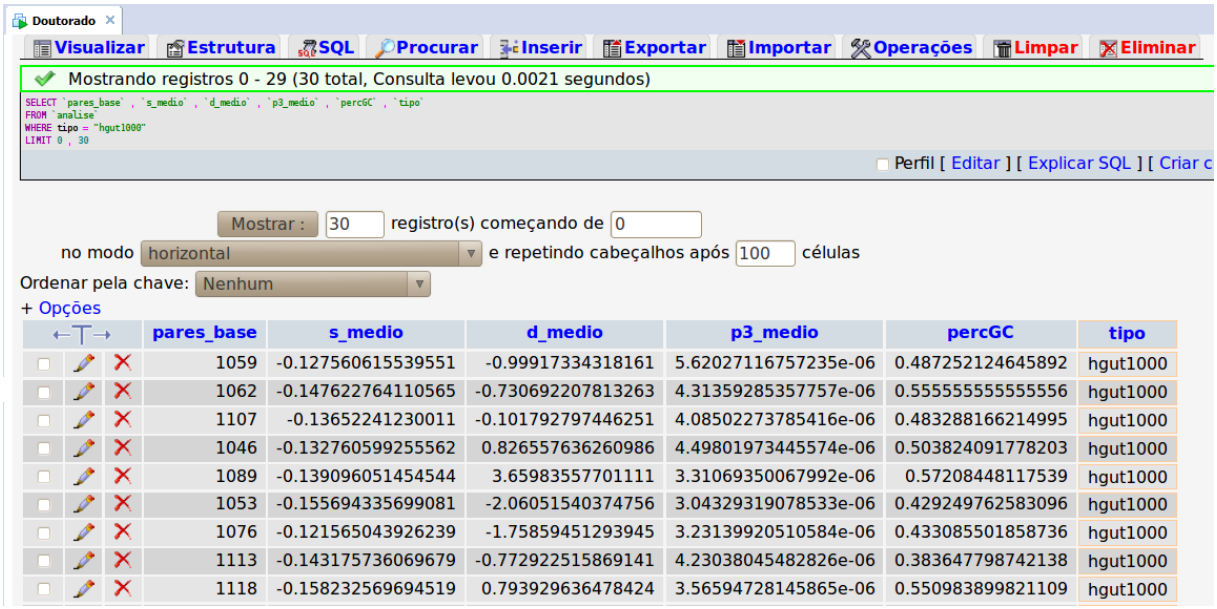


Figura 5.6 - Modelo de Dados Da Base da Dados Utilizada para Armazenas os Metadados e as Análises de Sequências Genômicas Realizadas



pares_base	s_medio	d_medio	p3_medio	percGC	tipo
1059	-0.127560615539551	-0.99917334318161	5.62027116757235e-06	0.487252124645892	hgut1000
1062	-0.147622764110565	-0.730692207813263	4.31359285357757e-06	0.555555555555556	hgut1000
1107	-0.13652241230011	-0.101792797446251	4.08502273785416e-06	0.483288166214995	hgut1000
1046	-0.132760599255562	0.826557636260986	4.49801973445574e-06	0.503824091778203	hgut1000
1089	-0.139096051454544	3.65983557701111	3.31069350067992e-06	0.57208448117539	hgut1000
1053	-0.155694335699081	-2.06051540374756	3.04329319078533e-06	0.429249762583096	hgut1000
1076	-0.121565043926239	-1.75859451293945	3.23139920510584e-06	0.433085501858736	hgut1000
1113	-0.143175736069679	-0.772922515869141	4.23038045482826e-06	0.383647798742138	hgut1000
1118	-0.158232569694519	0.793929636478424	3.56594728145865e-06	0.550983899821109	hgut1000

Figura 5.7 – Exemplo de Dados Obtidos Numa Consulta à Base de Dados

A Figura 5.7, demonstra uma consulta simples a base de dados, da tabela análise onde são armazenados os dados gerados pelo processo gerada pelo Processo Criação de Metadados (Figura 5.1) bem como os dados gerados pelo Processo de Análise (Figura 5.8). O que difere estes dados é tão somente o valor continho na coluna **tipo**, onde para os metadados seu conteúdo é **medcastas1000** e para sequências a serem analisadas. o conteúdo da coluna tipo, é o nome dado à sequência a ser analisada, no exemplo da figura 5.6 é **hgut1000**, sequência metagenômica que será analisada na seção 5.3.4 Comparativo com Metagenomas Conhecidos".

5.3 O Processo de Análise

A atividade inicial desse processo (Figura 5.8) consiste na seleção do genoma/metagenoma-alvo, sendo verificada a integridade do arquivo do padrão fasta, que possui somente letras que representam os nucleotídeos (A, C, G e T) e linhas de comentários iniciadas por >, e se necessário adequações ao referido padrão serão realizadas.

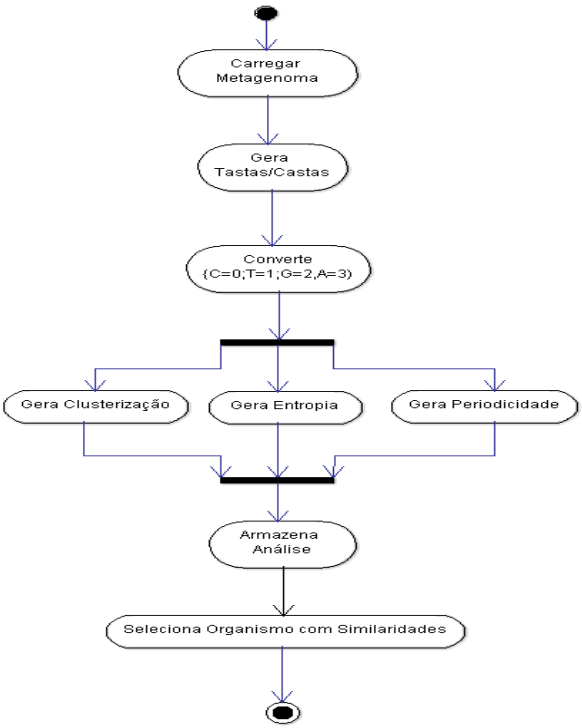


Figura 5.8 - Processo de Análise

As atividades Gera Casta/Tasta, Converte, Gera Clusterização, Gera Entropia, Gera Periodicidade e Armazena Análise são iguais as descritas na seção 5.1 deste capítulo.

A atividade de Armazenar Analise, inclui no banco de dados, para cada arquivo CASTA ou TASTA, a tupla

<locus,pares de base, s_medio, s_desvio, s_mediana, d_medio, d_desvio, d_mediana, p3_medio, p3_desvio, p3_mediana, %GC, data_inclusão, tipo, desvio %GC>,

sendo idêntica à descrita na seção 5.1.8.

Na etapa Seleciona Organismos com Similaridade, através de instruções SQL (*Strutured Query Language)* (Beaulieu, 2009), é realizado o produto cartesiano, representado pela Equação 5.8

$$A \bowtie M = \{(a,m) \mid a \in A \ \wedge \ m \in M\}$$ **Equação 5.8**

Onde *A* representa o conjunto de todas as análises dos arquivos Casta e Tasta da sequência alvo, *M* representa o conjunto de todos os metadados, ou assinatura, dos organismos que compõem a base de dados de assinaturas. Para exemplificar: se uma sequência (*A*) possui um milhão de pares de base, ele irá geral mil arquivos Tasta e mil arquivos Casta. Considerando que a base de dados (*M*) possui mil novecentos e setenta e um organismos, então $A \bowtie M$, gera três milhões e novecentos e quarenta e dois conjuntos, representado pela tupla.

<<locus,pares de base, s_medio, s_desvio, s_mediana, d_medio, d_desvio, d_mediana, p3_medio, p3_desvio, p3_mediana, %GC, data_inclusão, tipo, desvio %GC>, <locus,pares de base, s_medio, s_desvio, s_mediana, d_medio, d_desvio, d_mediana, p3_medio, p3_desvio, p3_mediana, %GC, data_inclusão, tipo, desvio %GC>>

Para melhor entendimento, o subconjunto proveniente de *A* está em vermelho e o subconjunto de *M* em azul. Como critério de seleção de possível organismo pertencente ao genoma alvo foram selecionadas aquelas tuplas onde

A.s_medio >= M.s_medio – M.s_desvio E
A.s_medio <= M.s_medio + M.s_desvio E
A.d_medio >= M.d_medio – M.d_desvio E
A.d_medio <= M.d_medio + M.d_desvio E **Equação 5.9**
A.p3_medio >= M.p3_medio – M.p3_desvio E
A.p3_medio <= M.p3_medio + M.p3_desvio E
A.%GC >= M.%GC – M.desvio%GC E
A.%GC <= M.%GC + M.desvio%GC

Pela Equação 5.9 a entropia média das sequências Castas e Tastas (*A.s_medio*) deve estar dentro do intervalo da entropia média do organismo (*M.s_medio*) mais e menos seu desvio padrão (*M.s_desvio*), isto é

$$M.s_{medio} - M.s\_desvio \ \geq A.s\_medio \ \leq \ M.s\_medio \ + M.s\_desvio \ .$$

O coeficiente de clusterização médio das sequências Castas e Tastas (*A.s_medio*) deve estar dentro do intervalo do coeficiente médio do organismo (*M.d_medio*) mais e menos seu desvio padrão (*M.d_desvio*), isto é :

$$M.d\_medio - M.d\_desvio \ \geq A.d\_medio \ \leq \ M.d\_medio \ + M.s\_desvio \ .$$

A periodicidade média das sequências Castas e Tastas (*A.p3_medio*) deve estar dentro do intervalo da periodicidade média do organismo (*M.p3_medio*) mais e menos seu desvio padrão (*M.s_desvio*), isto é:

$$M.p3\_medio - M.p3\_desvio \geq A.p3\_medio \leq M.p3\_medio + M.p3\_desvio .$$

O percentual de GC médio das sequências Castas e Tastas (*A.p3_medio*) deve estar dentro do intervalo da periodicidade média do organismo (*M.p3_medio*) mais e menos seu desvio padrão (*M.s_desvio*), isto é :

$$M.p3\_medio - M.p3\_desvio \geq A.p3\_medio \leq M.p3\_medio + M.p3\_desvio .$$

A instrução SQL exemplo abaixo, implementa a Equação 5.9. Esta mesma instrução foi utilizada para obtenção de dados de controle.

```
SELECT  DISTINCT  o.nome, count( * )
FROM analise AS a, metadados AS m, organismo AS o
WHERE a.tipo = "ET3"
AND a.p3_medio > m.p3_medio - m.p3_desvio
AND a.p3_medio < m.p3_medio + m.p3_desvio
AND a.s_medio > m.s_medio - m.s_desvio
AND a.s_medio < m.s_medio + m.s_desvio
AND a.d_medio > m.d_medio - m.d_desvio
AND a.d_medio < m.d_medio + m.d_desvio
AND a.percGC > m.percGC - m.desviopercGC
AND a.percGC < m.percGC + m.desviopercGC
AND m.tipo = "medcastas1000"
AND o.locus = m.locus
GROUP BY o.nome
ORDER BY count( * ) DESC
```

O resultado da instrução SQL acima é exemplificado na figura 5.9.

nome	count(*)	genero
Escherichia coli MG1655	62219	Escherchia
Bacillus leprae Hansen 1880	19788	Corynebacterineae
Neisseria gonorrhoeae NCCP11945	18856	Gonococcus Lindau 1898
Lactobacillus fermentum IFO 3956	11594	Lactobacillus
Shigella flexneri 2002017	10707	Shigella
Yersinia pseudotuberculosis PB1/+	10420	Yersinia
Shewanella baltica OS155	9792	Shewanella
Porphyromonas gingivalis ATCC 33277	7987	Porphyromonadaceae
Geobacillus kaustophilus HTA426	7000	Geobacillus
Acetobacter pasteurianus IFO 3283-01	6907	Acetimonas Orla-Jensen 1909

Figura 5.9 - Resultado da Instrução SQL

5.3 Análise Realizadas

Para validação do método desenvolvido foram realizadas as seguintes análise:

- Análise de uma sequência artificial de *E. coli*, gerada pelo MetaSim (Richter et al., 2008) contendo somente aleatórios de *E. coli*;
- Análise de uma sequência genômica artificial contendo somente nucleotídeos aleatórios randomicamente gerados;
- Análise de um metagenoma artificial, gerado pelo MetaSim, contendo organismos de trinta gêneros conhecidos;
- Análise de metagenomas conhecidos.

Em todas estas análises foram executados processos, como mostra a Figura 5.5, semelhantes ao da obtenção dos metadados definido na Figura 5.1.

5.3.1 Sequência artificial de *E. coli*

A sequência gerada utilizando o software de simulação MetaSim, gerando sequências-padrão as geradas por sequenciadores baseados no método de Sanger (Sanger & Coulson, 1975). Com a mesma acurácia, inserções, exclusões e deslocamentos de bases. O arquivo fasta desta simulação contém aproximadamente 9.700.000 bases, divididas em sequências de aproximadamente 1000 bases, gerando 9.700 arquivos.

A execução do processo de análise (Figura 5.10) sobre a sequência artificial de *E. coli* apresentou os seguintes resultados:

- A sequência "artificial" de *E. coli* versus a base de metadados (Item 5.1.8) gerou um produto cartesiano $A \bowtie M$ corresponde 39.122.379 tuplas, das quais 826.774 tuplas atenderam os critérios estabelecidos na Equação 5.9, onde o valores de s (entropia) da base de dados são maiores que o valor s da sequência em análise, menos o desvio padrão e menores que o valor de s mais o desvio padrão, os valores de d (coeficiente de clusterização) da base de dados são maiores que o valor d da sequência em análise, menos o desvio padrão e menores que o valor de d mais o desvio padrão, os valores de p3 (periodicidade) da base de dados são maiores que o valor p3 da sequência em análise menos o desvio padrão e menores que o valor de p3 mais o desvio padrão e os valores do %gc (percentual de GC) da base de dados são maiores que o valor %gc da sequência em análise menos o desvio padrão e menores que o valor de %gc mais o desvio padrão;
- Destas 826.774 tuplas, 218.460 (26,4%) corresponderam a tuplas de *E. coli*. O segundo organismo de maior frequência identificado, o *Corynebacterium* apresentou apenas 75.374 tuplas relacionadas, correspondente a 9,1% das tuplas. Além disso, quatro das cinco espécies com maior representatividade corresponderá a *Enterobacterias* e dos dez organismos com maior número de tuplas relacionadas, sete (incluída *E. coli*) pertencem ao grupo das *Proteobacterias.*

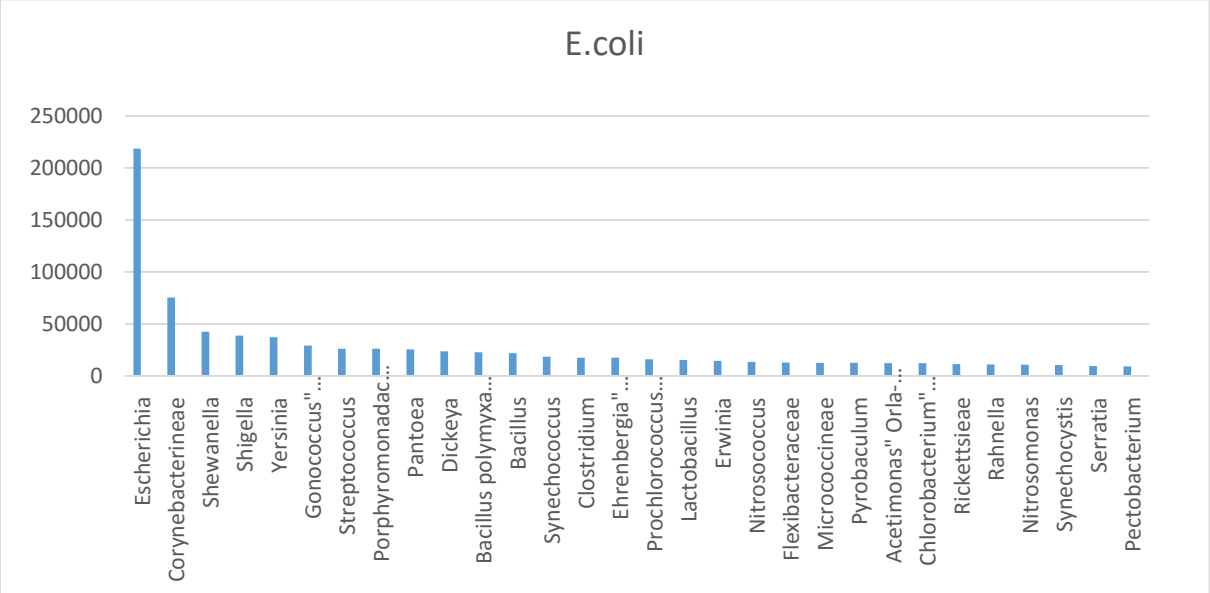


Figura 5.10 - Distribuição das frequências de organismo identificados utilizando-se o processo de análise proposta no seção 5.3

Esses resultados mostram que o organismo, a *E. coli*, presente na sequência-alvo foi claramente identificado pelo método proposto nesta tese. Os demais organismos identificados, podem ser classificados como falsos fositivo (FP), e isso deve-se fundamentalmente a similaridades encontradas em certas regiões do genoma que são muito comuns. Ainda pode-se concluir que os segmentos não identificados foram aqueles que obtiveram valores de s, p3, d ou %GC muito acima ou a baixo dos valores médios de E.Coli conforme demostrados na figura 5.4.

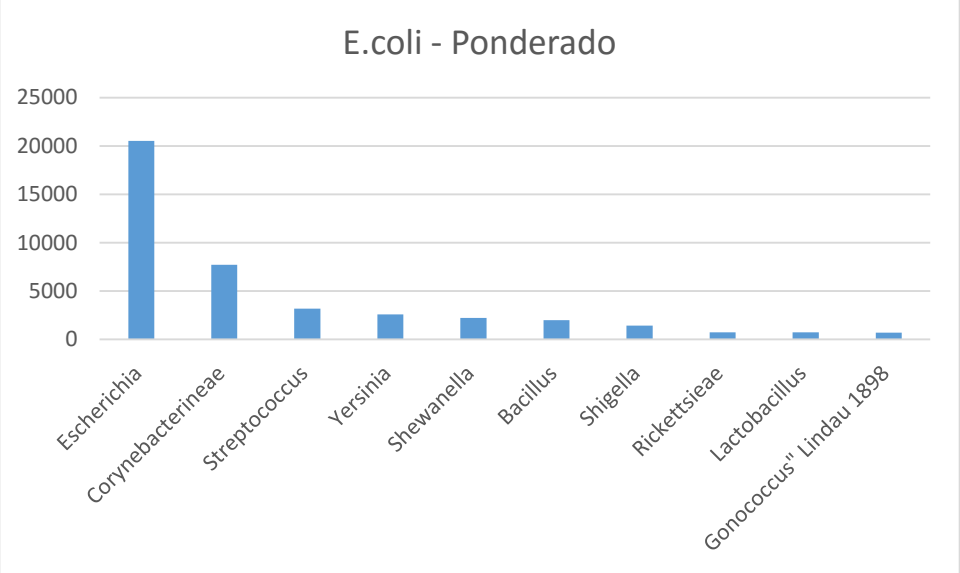


Figura 5.11 – Distribuição das frequências de organismo identificados ponderada

A Figura 5.9 demostra a análise dos possíveis organismos identificados no metagenoma alvo, ponderada pela quantidade de sequências destes organismos depositadas no Genbank. Pode-se observar que os resultados da análise são quase idênticos, com pequenas variações de frequências intermediárias.

5.3.2 Metagenoma de Sequência Genômica Artificial

Foi gerada uma sequência genômica contendo 9.999.220 de nucleotídeos de forma aleatória contendo 2.499.700 guaninas, 2.499.724 timinas, 2.499.879 citosinas e 2.499.833 adeninas.

A execução do processo de análise (Figura 5.7) sobre a sequência-alvo artificial apresentou os seguintes resultados:

- A sequência artificialmente gerada versus a base de metadados (Item 5.1.8) gerou um produto cartesiano $A \bowtie M$ corresponde 35.188.263 tuplas, das quais 5.554 tuplas atenderam os critérios estabelecidos na Equação 5.9 onde o valores de s (entropia) da base de dados são maiores que o valor s da sequência em análise menos o desvio padrão e menores que o valor de s mais o desvio padrão, os valores de d (coeficiente de clusterização) da base de dados são maiores que o valor d da sequência em análise menos o desvio padrão e menores que o valor de d mais o desvio padrão, os valores de p3 (periodicidade) da base de dados são maiores que o valor p3 da sequência em análise menos o desvio padrão e menores que o valor de p3 mais o desvio padrão e os valores do %gc (percentual de GC) da base de dados são maiores que o valor %gc da sequência em análise menos o desvio padrão e menores que o valor de %gc mais o desvio padrão;
- 124 organismos distintos foram identificados;
- O organismo de maior similaridade nessas 5.554 tuplas foi a *Anaplasma*, com uma frequência de 494, como mostra a Figura 5.10;
- O segundo organismo de similaridade foi a Y*ersinia sp.*, com uma frequência de 390;

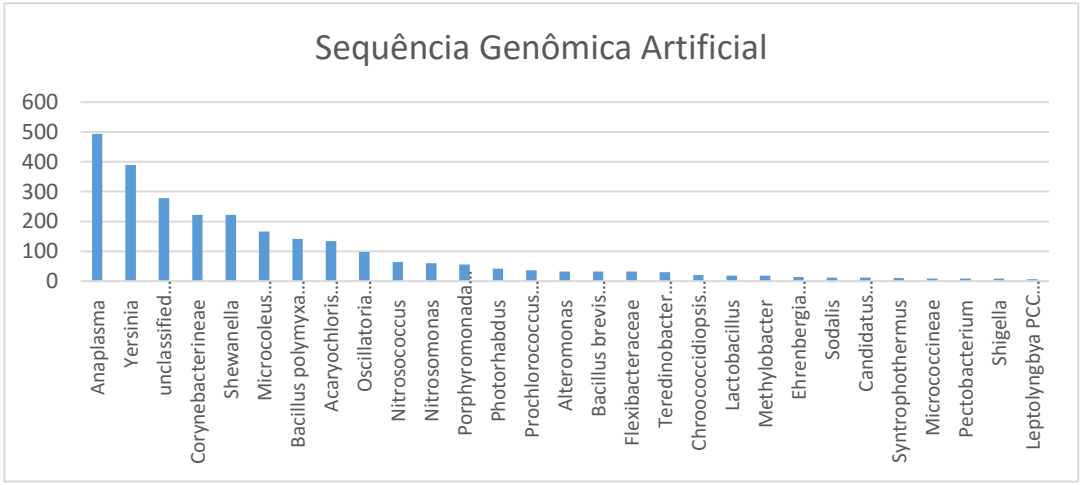


Figura 5.12 - Distribuição das frequências de organismos identificados aplicando-se o processo de análise sobre a sequência artificial

Os cento e vinte e quatro organismos obtidos desta análise são todos Falso Positivos, o metagenoma é artificial. Mas o que chama a atenção é a baixa frequência o que nos leva a concluir com esta análise e a anterior, apresentada na seção 5.2, que quanto maior a frequência

de um organismo nas tuplas selecionadas maior a probabilidade dele ser um Verdadeiro Positivo (VP).

5.3.3 Metagenoma de Trinta Spp. De Gêneros Diferente

Foi gerado um metagenoma de trinta organismos de trinta gêneros diversos, utilizando o MetaSim (Richter et al., 2008). A Tabela 5.2 lista os gêneros presentes no metagenoma. O metagenoma gerado contém 8.936.500 nucleotídeos. Para visualizarmos o comportamento da ferramenta de análise, na presença de múltiplos organismos, selecionamos aproximadamente 297.000 nucleotídeos por gênero. Os resultados desta análise foi o seguinte:

- O produto cartesiano $A \bowtie M$ correspondem 30.938.163 tuplas;
- Executando a seleção proposta na Equação 5.9, somente alterado o *tipo* que representa o metagenoma em análise, por ET4, foram obtidos 1.742.592 tuplas de 457 gêneros distintos, onde o valores de s (entropia) da base de dados são maiores que o valor s da sequência em análise menos o desvio padrão e menores que o valor de s mais o desvio padrão, os valores de d (coeficiente de clusterização) da base de dados são maiores que o valor d da sequência em análise menos o desvio padrão e menores que o valor de d mais o desvio padrão, os valores de p3 (periodicidade) da base de dados são maiores que o valor p3 da sequência em análise menos o desvio padrão e menores que o valor de p3 mais o desvio padrão e os valores do %gc (percentual de GC) da base de dados são maiores que o valor %gc da sequência em análise menos o desvio padrão e menores que o valor de %gc mais o desvio padrão;
- Não foram gerados Falso Negativo (FN), ou seja, todos os gêneros de organismos presentes na amostra metagenômica foram identificados;
- A análise resultou em 457 gêneros, ou seja, 427 gêneros falso fositivos (FP), mas destes 316 com baixa frequência, inferior a 2000;
- Onze dos 30 gêneros do metagenoma tiveram baixa frequência, inferior a 2000.

Tabela 5.2 - Metagenoma de 30 spp. de gêneros distintos

Gênero	Frequência Encontrada
Streptococcus	136516
Escherichia	79158
Lactobacillus	50220
Aurococcus	46868
Asterococcus Borrel	41362
Listerella	35782
Shewanella	25502
Acinetobacter	20922

Haemophilus	16164
Yersinia	13538
Prochlorococcus marinus	11242
Bartonia	9258
Rhizobium/Agrobacterium group	7670
Methanococcus	6546
Legionella	6526
Cystobacterineae	3176
Clostridium	2948
Acidivorax	2922
Chloroflexus	2134
Marinomonas	1742
Salmonella	1652
Aeropyrum	1532
Magnetococcus	1298
Gloeobacter violaceus	1090
Aquifex	770
Microcystis aeruginosa	668
Sapromyces Sabin	668
Jannaschia	622
Kineosporiineae	412
Laribacter	362

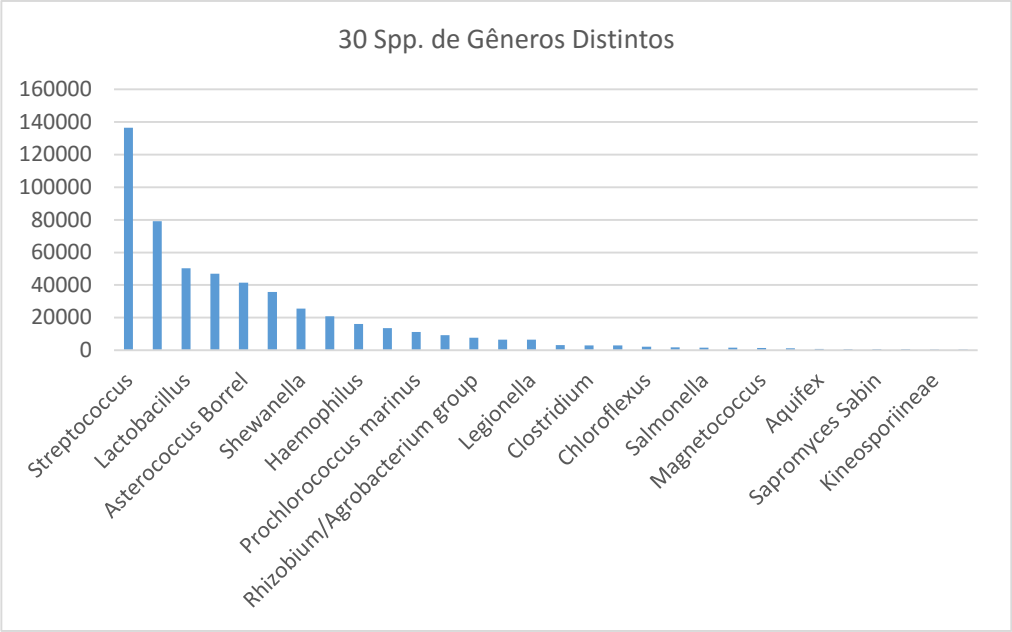


Figura 5.13 - - Distribuição das frequências de gêneros identificados aplicando-se o processo de análise sobre os 30 spp. de gêneros distintos

O resultado desta análise é que o metodo de análise metagenômica desenvolvida identificou 100% dos gêneros de organismos presentes na amostra metagenômica. Também é importante salientar que mesmo com um grande número de FP a grande maioria deles tiveram uma frequência baixa, inferior a 2000 .

5.3.4 Comparativo com Metagenomas Conhecidos

Foram obtidas as sequências utilizadas para elaboração do metagenoma do Mar de Sargaço (Venter et al., 2004), metagenoma de organismos identificados do intestino de um adolescente obeso (Ferrer et al., 2013) e metagenoma de organismos identificados em tubulações de ar condicionado de um *shopping center* localizado em Singapura (Tringe et al., 2008). Nestas três sequências foram executados, individualmente, o método implementado pelo processo da Figura 5.8, descrito na seção 5.3.

Em todas as análise não foram identificados FN, ou seja, 100% dos organismos identificados nos metagenomas originais foram identificados pelo método aqui proposto. Outro dado significante, que aproximadamente 90% dos organismos identificados nos metagenomas originais, estão entre os cinquenta de maior frequência.

A Figura 5.14 mostra os diversos filos encontrados, utilizando-se métodos distintos de identificação de organismos (EFG, EFTU, HSP70, RecA, RpoB e rRNA), na análise metagenômica do Mar de Sargasso (Venter et al., 2004). Já a figura 5.13 mostra os mesmos filos identificados utilizando-se o método proposto na seção 5.3. Pode-se observar uma distribuição semelhante das quantidades dos filos identificado nas duas abordagens, com exceção das *Crenarchaeotas*, onde o método aqui proposto identificou mais segmentos deste filo que a análise original.

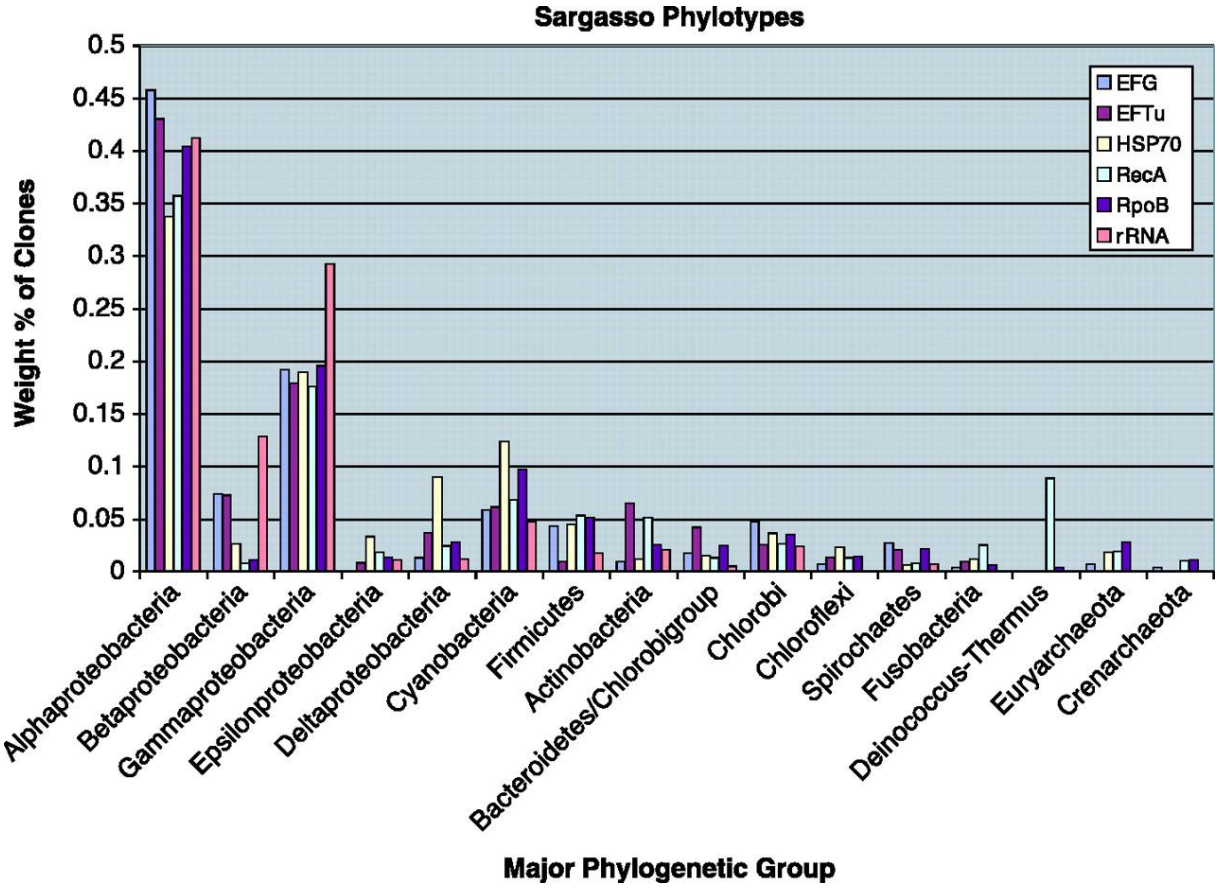


Figura 5.14 - Mar de Sargasso (Venter et al., 2004)

Figura 5.15 - Distribuição das frequências de organismos identificados aplicando-se o processo de análise sobre o metagenoma do mar de Sargasso

As figuras 5.16 e 5.17 mostram os resultados das análises do metagenoma do trato digestivo de um adolescente obeso (Ferrer et al., 2013) utilizando-se o método aqui proposto. Em comparativo aos organismos identificados utilizando-se 16S rDNA (Ferrer et al., 2013), a análise por esse método identificou uma quantidade muito maior de *Firmicutes*, aproximadamente 93% comparando-os com *Actinobactérias*, *Proteobactérias* e *Bacteiroidetes.* Análise realizada usando o método aqui proposto, como demonstra as figuras 5.15 e 5.16, a quantidade de sequências de *Firmicutes* comparadas com as de *Actinibactérias*, *Proteobactérias* e *Bacteiroidetes*, foi de 3%.

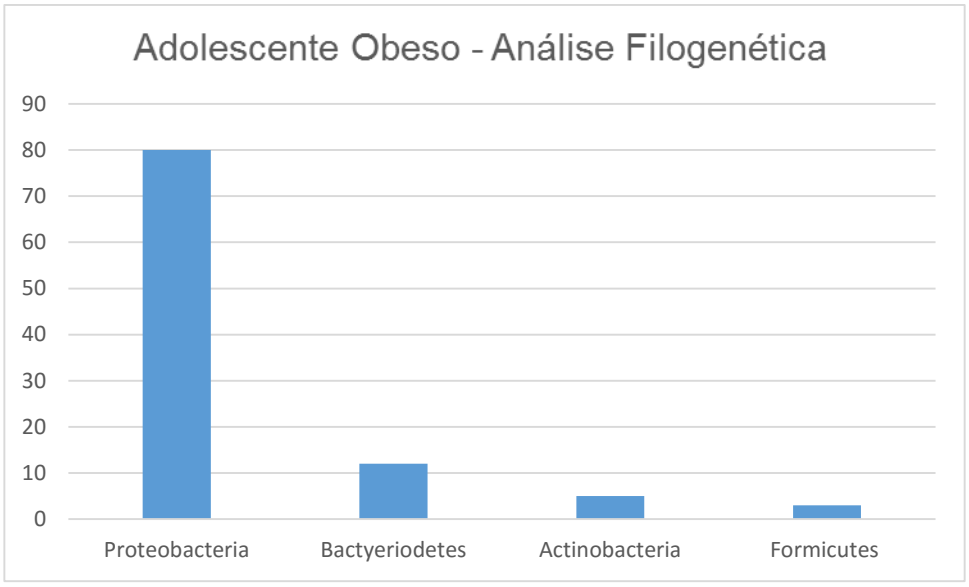


Figura 5.16 - Adolescente Obeso - Filo Comparativos

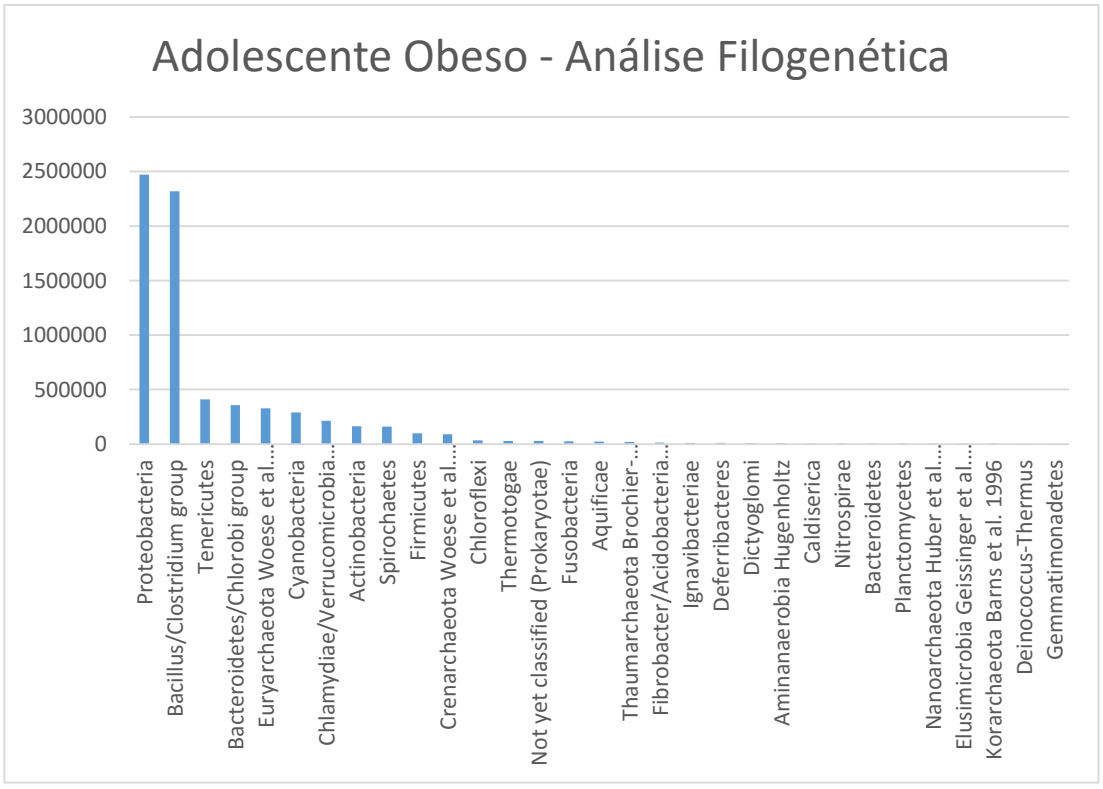


Figura 5.17 - Adolescente Obeso (Todos os Filos)

As figuras 5.18 e 5.19 mostram os resultados das análises do metagenoma extraído de dutos de ar condicionado de um *shopping center* localizado em Singapura. A figura 5.18 mostra o resultado da análise original (Tringe al., 2008) do metagenoma, onde sequências genômicas encontradas foram analisadas pelo método de 16S rDNA. Já a figura 5.19 mostra os resultados da análise utilizando-se o método descrita na seção 5.3. Pode-se observar, pela análise demonstrada na figura 5.19, que foram encontradas 100% dos organismos identificados na análise da figura 5.18, e como ocorreram nas análises demonstradas nas figura 5.13, 5.15 e 5.17, não coincidem os percentuais de organismos identificados nas suas respectivas análises originais (Venter et al., 2004; Ferrer et al., 2013; Tringe et al., 2008).

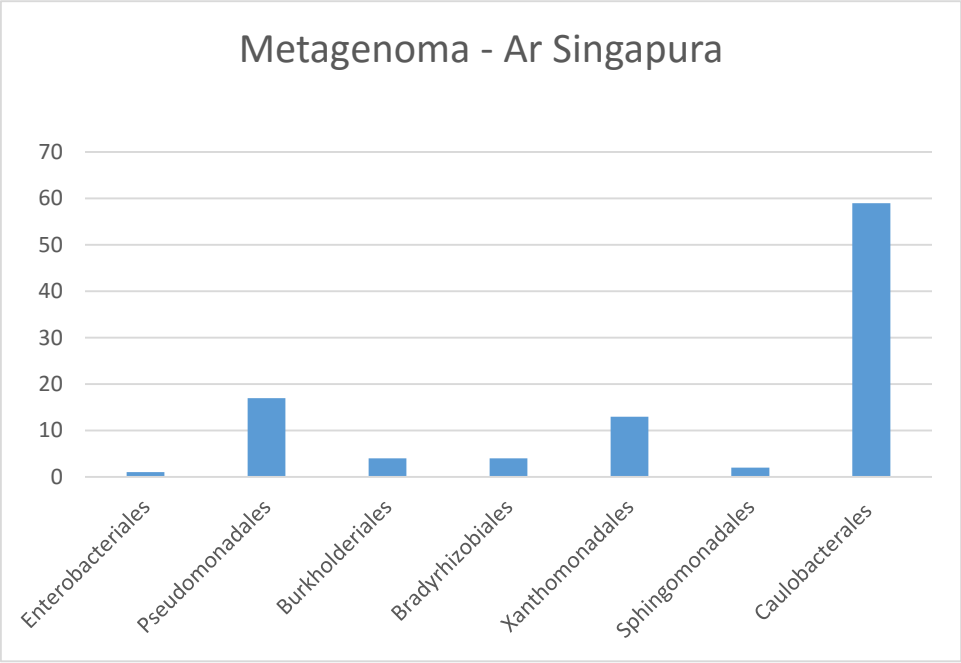


Figura 5.18 - Metagenoma ar condicionado em Singapura (Tringe al., 2008).

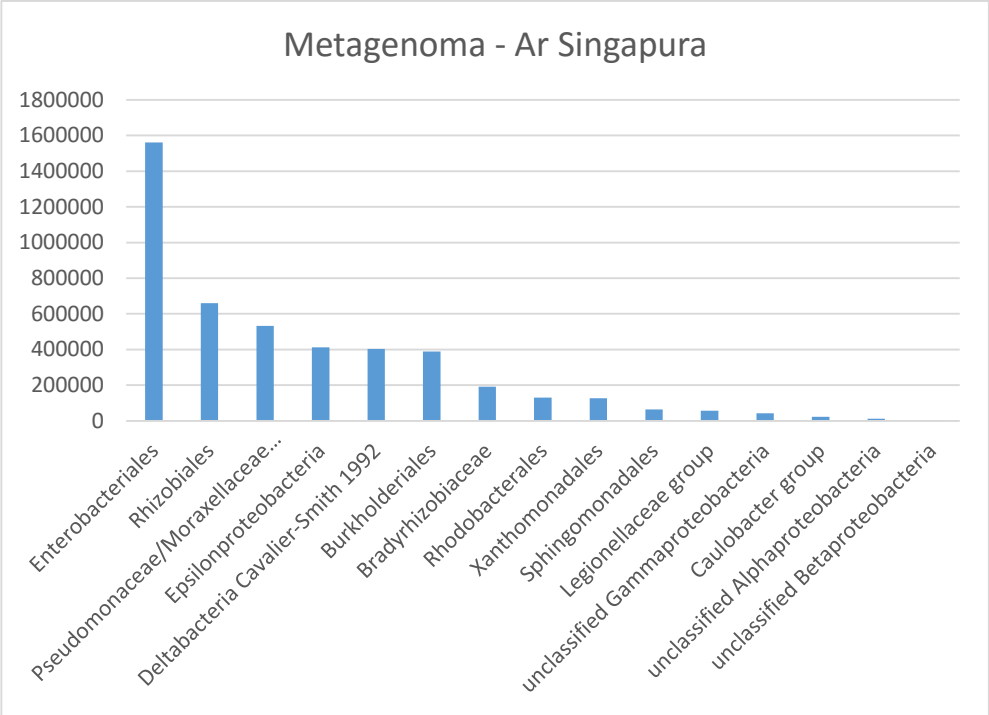


Figura 5.19 - Metagenoma ar condicionado em Singapura

6 CONCLUSÃO

Esta trabalho gerou um pedido de patente no Instituto Nacional de Propriedade Intelectual, depositada em 10/12/2014 sob número BR 10 2014 030893-8.

O método desenvolvido neste trabalho e suas ferramentas auxiliares, mostrou-se rápido e eficaz de análise de sequências genômicas, buscando identificar os organismos nelas presentes, independente do percentual dos mesmos.

Além disso, comparando a outros métodos existentes na literatura, verificou-se resultados semelhantes, o que indica a eficiência da tese desenvolvida, com ganhos significativos de tempo na obtenção dos resultados.

O método desenvolvido pode ser utilizada com outros, de forma complementar, ou ainda, como filtro para limitar os organismos a serem analisados, reduzindo desta forma o número de falsos positivos.

REFERÊNCIAS BIBLIOGRÁFICAS

Allegrini P., Grigolini P., Palatella L. (2003) Intermittency and scale-free networks: A dynamical model for human language complexity. http://lanl.arxiv.org/PS_cache/cond-mat/pdf/0310/0310648.pdf

Almeida, J., Vinga, S (2002). Universal sequence map (USM) of arbitrary discrete sequences. *BMC Bioinformatics*

Anastassiou, D.(2000) Frequency-domain analysis of biomolecular sequences. *Bioinformatics* 16(12): 1073.

Audit, B., Vaillant, C., Arnéodo, A., D'aubenton-carafa, Y. and Thermes, C. (2004). Wavelet Analysis of DNA Bending Profiles reveals Structural Constraints on the Evolution of Genomic Sequences. *Journal of Biological Physics* **30:** 33–81.

Baldi, P., **Brunak S.** (2001). Bioinformatics The Machine Lerning Approch. *Masschusetts Institute of Technology.*

Barral, J., Hasny, A., Jimenez, J., Marcano, A.(2000) Nonlinear modeling technique for the analysis of DNA chains. *Physical Review E* 61(2): 1812.

Barabási A.L., Albert R. (1999) Emergence of Scaling in Random Networks. *Science* 286:509–511.

Beaulieu, A. (2009). Learning SQL, 2º edition. **:** O'Reilly Media, Inc. Disponível em: http://dl.acm.org/citation.cfm?id=1611255&dl=ACM&coll=DL&CFID=367236585&CFTOKEN=72321548. Acessado em abril de 2013.

Benson, D., Karsch-Mizraschi, I., Lipman, D. (2013). GebBank. Nucleic Acids Res. 2013 Jan;41(Database issue):D36-42. doi: 10.1093/nar/gks1195. Epub 2012 Nov 27.

Bergveld, P. (1970) *Development of an ion-sensitive solid-state device for neurophysiological measurements.* IEEE Trans. Biomed. Eng. 17, pp. 70–71.

Billings, S. A., Coca, D. (1999)Discrete wavelet models for identification and qualitative analysis of chaotic systems. *International Journal of Bifurcation and Chaos* 9(7):1263.

Boggess, A. and Narcowich, F. (2001) First Course in Wavelets with Fourier Analysis (Hardcover). *Prentice Hall; 1 edition.*

Costa, R. M. and Cesar Jr, L. F. (2000) *Shape analysis and classification: Theory and practice*, CRC Press, Inc.

Chan, Y. T. (1995) *Wavelet basics*. **Kluwe**r Academic Publishers.

Chen, K. and Pachter, L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comp Biol 1 (2): 24.*

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. (2006) *Toward automatic reconstruction of a highly resolved tree of life.* Science. Mar 3;311(5765):1283-7.

Cooley, J. W. and Tukey, J. W. (1965). An Algorithm for the Machine Calculation of Complex Fourier Series. Math. Computat., 19, 297–301.

Covacci, A., Kennedy, GC, Cormack, B., Rappouli, R., Falkow, S. (1997). From microbial genomics to meta-genomics. *Drug Dev.* Res. 41:180-192

Cirne, W., Santos-Neto. E. (2006) Grids Computacionais: da Computação de Alto Desempenho a Serviços sob Demanda.

Cohen, J., Computer Science and Bioinformatics,. *Communications of the ACM*, vol. 48, nº3(3), 2005.

Cover, T. M., Thomas, J.A., (1991*)* Elements of Information Theory. NY: *John Wiley & Sons, Inc.*

Dodiin, G., Vandergheynst, P., Levoir, P., Cordier, C., Marcout, L (2000). Fourier and wavelet transform analysis, a tool for visualizing regular pattern in DNA sequences. *Journal of Theoretical Biology* 206:323.

Dorogovtsev S.N., Mendes J.F.F. (2003) *Evolution of Networks: From Biological nets to the Internet and WWW*. Oxford University Press, Oxford.

Eigen M., Gardiner W., Schuster P., Winkler-Oswatitsch R. (1981). The origin of genetic information. *Sci. Am.* **244** (4): 88–92, 96, et passim.

Ferrer M., et al. (2013) Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure. Environmental Microbiology **15**(1), 211–226

Ficket, J., Tung, C. (1992) Assessment of protein coding measures, *Nucleic Accids Research 20(24*): 6441-6450

Gan, G., Ma, C. and Wu, J. (2007) Data Clustering, Teory, Algotithms, and Aplications. *ASA-SIAM – American Statistical Association and Society for Industrial and Applied Mathematics, Philadelphia*.

Gerhardt G.J.L., Lemke N., Corso G. (2006) Network Clustering Coefficient approach to DNA sequence analysis. *Chaos, Solitons and Fractals* **28:**1037–1045.

Guharay, S., Hunt , B. R., Yorke, J. A., White, O. R. (2000) Correlations in DNA sequences across the three domains of life. *Physica D, 146*, p.388-396.

Handelsman, Jo. (2007) Metagenomics and Microbial Communities. eLS. John Wiley & Sons Ltd, Chichester. Diponível em http://www.els.net. Acessado em Outubro 2013.

Herzel, H., Große, I. (1995) Meassuring correlations ins symbol sequences. *Physica A, 216*, p.518-542.

Herzel, H., Weiss, O., Trifonov, E. (1999). 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics* 15 (3):187-193.

Howard, Eves. (2002) História da Matemática. Trad. Hygino H. Domingues. Ed. da Unicamp. Campinas.

Hosid, S., Trifonov, E., Bolshoy, A. (2004) Sequence periodicity of *Escherichia Coli* is concentrated in intergenic regions. *BMC Molecular Biology* 5:14.

Katsaloulis, P.,Theoharis, T., Provata, A. (2002) Statistical distributions of oligonucleotide combinations: applications in human chromosomes 21 and 22.. *Physica A, 316*, p.380-396.

Kernighan, Brian W., Ritchie, Dennis M. (1983) *C Programming Languagem* Second Edition. . AT&T Bell Laboratories. Murray Hill, New Jersey. Prentice Hall

Lavine, B. (1992) Clustering and Classification of Analytical Data. *Published by John Wiley & Sons, Inc., Chichester.*

Letunic I, Bork P. (2007) *Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation.* Bioinformatics. 2007 Jan 1;23(1):127-8. Epub 2006 Oct 18.

Oliveira, J.V. and Pedricz, W. (2007) Advances in Fuzzy Clustering and its Applications. *John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester*.

Osake, M., Gohare, K., Ishii, S., Kishida, H., Hayyakawa, H. (1999) Symbolic string and spatial $1/f$ spectra. *Physica A, 316*, p.142-154.

Martinoto, André Luis ; Jornada, F. H. ; Cassol, Luciano A. ; Gava, Vanius ; Dorneles, Ricardo Vargas ; Perotonni, C. A (2008). Generation of Continuous Random Networks by Simulated Annealing. In: *23rd Annual ACM Symposium on Applied Computing, 2008, Fortaleza. Special Track on Advances in Computer Simulation. v. 1. p. 48-49.*

Mena-Chalco, J. P., Carrer, H., Zana, Y., Cesar Jr, R.M. (2008) *Identification of Protein Coding Regions Using the Modified Gabor-Wavelet Transform*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 5, no. 2,.

Moore, Gordon E. (1965) *Cramming More Components Onto Integrated Circuits*. Electronics, Volume 38, Number 8, April 19, 1965.

Mra´zek, J. (2010) Comparative Analysis of Sequence Periodicity among Prokaryotic Genomes Points to Differences in Nucleoid Structure and a Relationship to Gene Expression. *Journal of Bacreriology*, July 2010, p. 3763–3772

Mitchel, T. (1997) Machine Learning. *McGraw-Hill Science/Engineering/Math.*

Mitra, S.and Acharya, S. (2003) Data Mining Multimedia, Soft Computing and Bioinformatics. *Published by John Wiley & Sons, Inc., Hoboken, New Jersey.*

Piazza, F., Lio`, P. (2005). Statistical analysis of simple repeats in the human genome. Physica A. vol. 347, pp. 472-488

Programming Language Popurity. (2013). Disponível em http://www.langpop.com/. Acessada em abril de 2013.

Raghavan, K., Ruskin, H.J. and Perrin, D. (2011). Computational Analysis of Epigenetic Information in Human DNA Sequences. *International Conference on Bioscience, Biochemistry and Bioinformatics IPCBEE vol.5.*

Richter DC, Ott F, Auch AF, Schmid R, Huson DH (2008) *MetaSim—A Sequencing Simulator for Genomics and Metagenomics*. PLoS ONE 3(10): e3373.doi:10.1371/journal.pone.0003373. Disponível em: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0003373. Acessada em março de 2013.

Ronaghi M., Karamohamed S., Pettersson B., Uhlén M., Nyrén P. (1996) *Real-Time DNA Sequencing Using Detection of Pyrophosphate Release*. Analytical Biochemistry, Volume 242, Issue 1, 1 November 1996, pp 84-89

Rothberg J., Hinz W., Rearick T.M., et al. (2011). *An integrated semiconductor device enabling non-optical genome sequencing.* Nature 475, pp 348–352. Disponível em: http://www.nature.com/nature/journal/v475/n7356/full/nature10242.html. Acessada em agosto de 2013.

Sanger, F.; Coulson, A.R. (1975), A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase, Journal of Molecular Biology **94** (3): 441–448

Schieg, P., Herze, H. (2004) Periodicities of 10–11 bp as Indicators of the Supercoiled State of Genomic DNA. *Journal of Molecular Biology*, 891-901.

Shannon, C. E.,Weaver,W.(1949) The Mathematical Theory of Communication. *Univeristy of Illinois Press*, 1949.

Sharma, S. Rana, S. and Singh, R. (2012). A Short Note – Metagenomics. *International Journal of Biomedical Research, Vol 3 Nº 4.*

Stain, A., Bina, M. (1999) À signed encoded in vertebrate DNA that influences nucleosome position and alignment. *Nucleid Acids Research* 23 (3):848-853.

Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S. and Ramaswamy, R. (1997). Prediction of probable genes by Fourier analysis of genomic sequences. *Cabios – Oxford Univercity Press. Vol. 13 no. 3, pp. 263-270*

Trifonov, E. (1990) Making sense of the human genome, *Structure & Methods Adenine Press*, vol. 1, pp. 69-77.

Trifonov, E, Bettecken, T. (1997). Sequence fossus, triplet expansion, and reconstruction od earliest códons. *Na International Jornal on Gene and Genomas.* Gene(205):1-6.

Trifonov, E. (1998) 3-, 10.5-, 200-, 400-base periodicities in genome sequences *Physica A, 249*, p.511-516.

Tringe SG, Zhang T, Liu X, Yu Y, Lee WH, et al. (2008) The Airborne Metagenome in an Indoor Urban Environment. *PLoS ONE 3(4): e1862.doi:10.1371/journal.pone.0001862*.

Disponível em: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0001862. Acessada em agosto de 2013.

Tolstorukov, M. Y., K. M. Virnik, S. Adhya, and V. B. Zhurkin. (2005). A-tract clusters may facilitate DNA packaging in bacterial nucleoid. *Nucleic Acids Res.* 33:3907–3918.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W (2004). *Environmental genome shotgun sequencing of the Sargasso Sea*. Science 304: 66-74.

Vieira, M.S.(1999) Statistics of DNA sequences: a low-frequency analysis, *Phys. Rev. E*, vol. 60(5), pp. 5932–5937, 1999.

Wang, Liya , Sten, Lincoln D. (2010). Localizing triplet periodicity in DNA and cDNA. *BMC Bioinformatics, 11:550*

Zhang, Chun-Ting. (1997). A Symmetrical Theory of DNA Sequences and Its Applications. *J*. T*theor*. *Biol*. 187, 297-306.

Zeyaullan, Md., Kamil, M.R., Ilam, B.,Atif, Benkhayal, F.A., Nehal, M., Rizvi, M.A, Ali, A. (2008). *Metagenomics – Na advanced approach for non-cultivable micro-organisms.* Biotecnology and Molecular Biology Reviews Vol.4 (3), pp. 049-954.