REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DO DESENVOLVIMENTO, INDÚSTRIA, COMÉRCIO E SERVIÇOS
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL

CARTA PATENTE Nº BR 102014030893-8

O INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL concede a presente PATENTE DE INVENÇÃO, que outorga ao seu titular a propriedade da invenção caracterizada neste título, em todo o território nacional, garantindo os direitos dela decorrentes, previstos na legislação em vigor.

(21) Número do Depósito: BR 102014030893-8

(22) Data do Depósito: 10/12/2014

(43) Data da Publicação Nacional: 12/07/2016

(51) Classificação Internacional: G16H 10/40.

(54) Título: MÉTODO DE ANÁLISE DE SEQUÊNCIAS GENÔMICAS E SISTEMA

(73) Titular: FUNDAÇÃO UNIVERSIDADE DE CAXIAS DO SUL, Pessoa Jurídica. Endereço: Rua Francisco Getúlio Vargas, nº 1130, Bloco A, Sala 301, Petrópolis, Caxias do Sul, RS, BRASIL(BR), 95070-560, Brasileira

(72) Inventor: RICARDO AUGUSTO MANFREDINI; SÉRGIO ECHEVERRIGARAY LAGUNA; GUNTHER JOHANNES LEWCZUK GERHARDT.

Prazo de Validade: 20 (vinte) anos contados a partir de 10/12/2014, observadas as condições legais

Expedida em: 07/02/2023

Assinado digitalmente por:
Alexandre Dantas Rodrigues
Diretor Substituto de Patentes, Programas de Computador e Topografias de Circuitos Integrados

Relatório Descritivo de Patente de Invenção

MÉTODO DE ANÁLISE DE SEQUÊNCIAS GENÔMICAS E SISTEMA

Campo da Invenção

[0001] A presente invenção descreve um método de análise de sequências genômicas e um sistema para análise de sequências genômicas. A presente invenção situa-se nos campos da Bioinformática, da Biologia Evolutiva e da Engenharia da Computação.

Antecedentes da Invenção

[0002] Atualmente, nota-se que existem inúmeras ferramentas computacionais para o processamento de sequências genômicas. Particularmente, a Bioinformática tem se desenvolvido de modo intenso, tendo em vista o crescente interesse comercial por esse segmento e pelas aplicações das ferramentas computacionais nas áreas da Biologia e da Medicina como, por exemplo, na análise de sequências genômicas como, por exemplo, a análise metagenômica.

[0003] Entretanto, ainda é um desafio do estado da técnica prover soluções para o aprimoramento e melhoramento dos métodos e sistemas de análise de sequências genômicas, de modo a prover melhores resultados (maior acurácia, maior precisão, dentre outros parâmetros), mas mantendo-se a rapidez e sem necessidade de clonagens ou cultivos *in vitro* e permitindo-se a determinação da assinatura do microrganismo identificado pelo método / sistema.

[0004] Na busca pelo estado da técnica em literaturas científica e patentária, foram encontrados os seguintes documentos que tratam sobre o tema:

[0005] O documento "*Triplet Entropy in H1N1 Virus"* (Santos *et al.*, 2011.) revela a aplicação da análise de entropia de triplets (fração GC) e entropia de Shannon em sequências do vírus H1N1, mas não revela, ensina, sugere ou motiva a combinação da análise de entropia de triplets e entropia de Shannon

com outros parâmetros/variáveis não correlacionadas ou em análises metagenômicas.

[0006] O documento "*Network clustering coefficient approach to DNA sequence analysis*" (Gerhardt *et al.*, 2006) revela uma ferramenta alternativa para a análise de sequências de DNA baseada em grafos, mas não se aplica à análises metagenômicas e não revela, ensina, sugere ou motiva a combinação dos quatro parâmetros não correlacionados da presente invenção.

[0007] O documento "*Prediction of probable genes by Fourier analysis of genomic sequences*" (Tiwari et al., 1997) refere-se à análise de periodicidade de triplets (3-bp) com utilização de técnicas de Fourier para reconhecer regiões codificadoras do DNA. Entretanto, revela apenas um dos parâmetros não correlacionados (periodicidade de triplets) e não revela, ensina, sugere ou motiva a combinação dos quatro parâmetros não correlacionados da presente invenção.

[0008] O documento "*Localizing triplet periodicity in DNA and cDNA sequences*" revela resultados específicos da aplicação de transformadas de Fourier em segmentos de DNA para avaliação de periodicidade, porém o referido documento limita-se à avaliação de periodicidade, não se revelando, sugerindo, ensinando ou motivando a combinação de quatro parâmetros não correlacionados como os descritos na presente invenção.

[0009] O documento WO 2007105150 revela um método para gerar espectrograma de DNA gerada a partir de transformada de Fourier para conversão dos nucleotídeos A, T, C e G em representação visual. Assim, o referido documento distingue-se totalmente do presente pedido de patente, por não revelar nenhum dos parâmetros utilizados pela presente invenção.

[0010] Assim, do que se depreende da literatura pesquisada, não foram encontrados documentos antecipando ou sugerindo os ensinamentos da presente invenção, de forma que a solução aqui proposta possui novidade e atividade inventiva frente ao estado da técnica.

[0011] Portanto, o estado da técnica não revela, ensina, sugere ou motiva a combinação de variáveis não correlacionadas para análise de sequências genômicas. Além disso, apenas provê métodos de análise de sequências genômicas baseadas em técnicas baseadas em sequências ribossomais RNA 16 S, com necessidade de cultura *in vitro* e clonagens.

Sumário da Invenção

[0012] Dessa forma, a presente invenção tem por objetivo resolver os problemas constantes no estado da técnica a partir de um método de análise de sequências genômicas que compreender as etapas de:

a) selecionar pelo menos uma sequência genômica para análise; e

b) gerar pelo menos o valor de quatro variáveis não correlacionadas.

[0013] Em um primeiro objeto, a presente invenção refere-se a um método de análise de sequências genômicas compreendendo as etapas de:

a) selecionar pelo menos uma sequência genômica para análise; e

b) gerar pelo menos o valor de quatro variáveis não correlacionadas.

[0014] Em um segundo objeto, a presente invenção refere-se a um sistema de análise de sequências genômicas compreendendo:

a) pelo menos um dispositivo eletrônico para processamento de sequências genômicas, sendo que o dito dispositivo eletrônico executa o método de análise de sequências genômicas conforme definido no presente pedido de patente; e

b) pelo menos um dispositivo eletrônico para armazenamento ou comparação dos dados obtidos pelo dispositivo eletrônico da etapa a).

[0015] Ainda, o conceito inventivo comum a todos os contextos de proteção reivindicados refere-se à combinação de parâmetros distintos não correlacionados e cuja combinação permite uma importante e melhor avaliação de sequências genômicas com aplicação em distintos campos do sequenciamento biológico como, por exemplo, em metagenomas.

[0016] Estes e outros objetos da invenção serão imediatamente valorizados pelos versados na arte e pelas empresas com interesses no

segmento, e serão descritos em detalhes suficientes para sua reprodução na descrição a seguir.

Breve Descrição das Figuras

[0017] Com o intuito de melhor definir e esclarecer o conteúdo do presente pedido de patente, são apresentadas as presente figuras:

[0018] A Figura 1 mostra um exemplo de concretização do método / sistema revelado pela presente invenção. Especificamente, indica-se carregar genoma e a geração dos arquivos tastas/cascas, seguido da obtenção das variáveis não correlacionadas.

[0019] A Figura 2 mostra um exemplo de arquivo “casta” padrão contendo mil (1000) nucleotídeos.

[0020] A Figura 3 mostra um exemplo de dados descritivos armazenados nos arquivos do tipo casta/tasta.

[0021] A Figura 4 mostra os resultados das métricas que foram utilizadas nas análises de sequências genômicas, onde o gráfico **s** demostra a entropia, o gráfico do coeficiente de clusterização.

[0022] A Figura 5 mostra um modelo de dados (no caso, a sequência analisada é hgut1000).

[0023] A Figura 6 mostra 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 1) bem como os dados gerados pelo Processo de Análise (também representado pela Figura 1).

[0024] A figura 7 mostra a um exemplo de execução do processo de análise, neste caso sobre a sequência artificial de *E. coli* e os resultados obtidos.

[0025] A Figura 8 mostra a distribuição das frequências de organismos encontrados.

[0026] A Figura 9 mostra os diversos filos encontrados na análise metagenômica do Mar de Sargasso (Venter *et al.*, 2004), utilizando-se do método revelado pelo presente pedido de patente.

[0027] A Figura 10 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).

[0028] A Figura 11 mostra os mesmos filos encontrados aplicando-se o método / sistema revelado pela presente invenção.

[0029] A Figura 12 mostra os resultados das análises do metagenoma do trato digestivo de um adolescente obeso (Ferrer *et al.*, 2013) utilizando-se o método / sistema revelado pela presente invenção.

[0030] A Figura 13 mostra os resultados das análises do metagenoma do trato digestivo de um adolescente obeso (Ferrer *at al.*, 2013) utilizando-se o método / sistema revelado pela presente invenção.

[0031] A Figura 14 mostra o resultado da análise original (Tringe *et al.*, 2008) do metagenoma, onde sequências genômicas encontradas foram analisadas pelo método de 16S rDNA.

[0032] A Figura 15 mostra os resultados da análise genômica utilizando-se o método / sistema revelado pela presente invenção.

Descrição Detalhada da Invenção

[0033] Em um primeiro objeto, a presente invenção refere-se a um método de análise de sequências genômicas compreendendo as etapas de:

a) selecionar pelo menos uma sequência genômica para análise; e

b) gerar pelo menos o valor de quatro variáveis não correlacionadas.

[0034] Em uma concretização, o método de análise de sequências genômicas compreende as etapas de:

a) selecionar pelo menos uma sequência genômica para análise; e

b) gerar o valor das variáveis não correlacionadas do grupo consistindo de: valor da entropia de Shannon, coeficiente de clusterização de Gerhardt, percentual dos nucleotídeos guanina e citosina, valor da periodicidade de P3, energia livre total de Gibbs e suas combinações.

[0035] Em uma concretização, o método de análise de sequências genômicas compreende as etapas de:

a) selecionar pelo menos uma sequência genômica para análise;

b) gerar o valor da entropia de Shannon;

c) gerar o coeficiente de clusterização de Gerhardt;

d) gerar o percentual dos nucleotídeos guanina e citosina; e

e) gerar o valor da periodicidade de P3.

[0036] Em uma concretização, o método compreende a etapa adicional de armazenar os dados obtidos em qualquer uma das etapas a) a e) em um dispositivo eletrônico de armazenamento de dados.

[0037] Em uma concretização, o armazenamento de dados é em um banco de dados de sequências biológicas.

[0038] Em uma concretização, a etapa de gerar o valor da periodicidade de P3 ocorre por análise espectral.

[0039] Em uma concretização, a análise espectral é a Transformada Discreta de Fourier ou a Transformada Rápida de Fourier.

[0040] Em uma concretização, a sequência genômica é de organismos procariontes.

[0041] Em uma concretização, o método é de análise metagenômica.

[0042] Em uma concretização, a referida análise metagenômica é para a determinação da assinatura do organismo em estudo a partir da combinação de pelo menos quatro variáveis não correlacionadas.

[0043] Em um segundo objeto, a presente invenção refere-se a um sistema de análise de sequências genômicas compreendendo:

[0044] a) pelo menos um dispositivo eletrônico para processamento de sequências genômicas, sendo que o dito dispositivo eletrônico executa o método revelado pela presente invenção; e

[0045] b) pelo menos um dispositivo eletrônico para armazenamento ou comparação dos dados obtidos pelo dispositivo eletrônico da etapa a).

[0046] Em uma concretização, o dispositivo eletrônico da etapa b) é o dispositivo eletrônico para processamento de sequências genômicas da etapa a).

[0047] Em uma concretização, o dispositivo eletrônico é um servidor, um computador portátil, um computador de computação paralela.

[0048] Em uma concretização, o dispositivo eletrônico é dotado de conexão a pelo menos uma base de dados de sequências ou informações genômicas.

Definições de alguns termos e conceitos gerais utilizados no presente pedido de patente

[0049] Variáveis não correlacionadas

[0050] No presente pedido de patente, o termo deve ser entendido como sendo variáveis cujas variações quantitativas independem das outras, ou seja, não existe uma correlação direta entre as variáveis. Exemplos não limitantes de variáveis não correlacionadas são: valor da entropia de Shannon, coeficiente de clusterização de Gerhardt, percentual dos nucleotídeos guanina e citosina, valor da periodicidade de P3, energia livre total de Gibbs e combinações das mesmas.

[0051] Valor da variável não correlacionada

[0052] No presente pedido de patente, o termo deve ser entendido como sendo o valor atribuído (calculado) da variável não correlacionada, podendo ser entendido como um coeficiente, um percentual, ou seja, um dado numérico.

[0053] Sequências genômicas

[0054] No presente pedido de patente, o termo deve ser entendido como qualquer sequência biológica como, por exemplo, mas não se limitando a, sequências inteiras ou parciais de DNA e de RNA (qualquer tipo de RNA, incluindo RNA ribossômico).

[0055] Valor da entropia de Shannon

[0056] No presente pedido de patente, o termo deve ser entendido como uma função de probabilidade $p(x)$ e que é usualmente descrita como:

$$H_s(X) = H(p(x_1), p(x_2), \ldots, p(x_n)),$$

Equação 3.1

[0057] onde, *x* é uma variável discreta aleatória de *X*, em um conjunto discreto *X = {x1, ... ,xn}*, com função de massa de probabilidade *p(x) = Pr(X = x)*. A entropia de Shannon de *X, H(X)*, é definida como:

$$H_s(X) = -\sum p(x) \log_b p(x)$$

Equação 3.2

[0058] Analisando uma sequência de n símbolos, a entropia *(HS(X))* é o somatório negativo das probabilidades de ocorrência de um símbolo *pi* multiplicadas pelo binômio ($log_2 pi$). Para elucidar o conceito seguem-se alguns exemplos.

[0059] 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:

$$-((0.5)(-1) + (0.5)(-1)) = 1$$

Equação 3.3

[0060] 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:

$$-((0.75)(-0.415) + (0.25)(-2)) = 0.81$$

Equação 3.4

[0061] No caso do DNA, considerando o alfabeto quaternário E={4,C,T,G}, e considerando uma sequência aleatória, sem diferenças probabilísticas entre as quatro bases teríamos:

$$-((0.25)(-2) + (0.25)(-2) + (0.25)(-2) + (0.25)(-2)) = 2$$

Equação 3.5

[0062] 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

[0063] 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.

[0064] A teoria da informação de Shannon foi utilizada para calcular a entropia dos 64 possíveis combinações de ACGT num triplet em porções de sequências de DNA de genomas completos de bactérias.

[0065] Coeficiente de clusterização de Gerhardt

[0066] No presente pedido de patente, o termo deve ser entendido como o coeficiente de clusterização conforme definido na equação 5.3 indicada no presente pedido de patente.

[0067] Percentual dos nucleotídeos guanina e citosina

[0068] No presente pedido de patente, o termo deve ser entendido como a porcentagem de nucleotídeos guanina e citosina em relação ao total de bases nitrogenadas (adenina – A, timina – T, citosina – C, guanina – G e, em alguns casos, considera-se também a quantidade de uracila – U).

[0069] Valor da periodicidade de P3

[0070] No presente pedido de patente, o termo deve ser entendido como um valor obtido a partir de função correlação *Fc (n)* que se baseia na equação 6 indicada neste pedido de patente.

[0071] Energia Livre Total de Gibbs

[0072] No presente pedido de patente, o termo deve ser entendido como a energia livre total de Gibbs ($\Delta G°$) que, por exemplo, pode ser calculada na sobreposição da cadeia complementar de DNA para formação da fita dupla do fragmento de DNA, sendo o $\Delta G°$ determinado pela variação da entalpia e da

entropia na região de pareamento baseado no modelo *Nearest-Neighbour* (NN).

[0073] Banco de dados de sequências biológicas

[0074] No presente pedido de patente, o termo deve ser entendido como qualquer tipo de banco de dados que possibilite o armazenamento de sequências biológicas. Exemplos não limitantes de bancos de dados também incluem bancos de dados armazenados “nas nuvens” (cloud computing) e bancos de dados relacionais.

[0075] Análise espectral

[0076] No presente pedido de patente, o termo deve ser entendido como qualquer função matemática que associe tempo em função de frequência podendo ser diversos tipos de transformada de Fourier, como por exemplo, transformada discreta de Fourier, transformada de Fourier de tempo discreto (DTFT), dentre outras.

[0077] Organismos procariontes

[0078] No presente pedido de patente, o termo deve ser entendido como qualquer organismo que não possua material genético delimitado por uma membrana. Incluem-se nesta definição as bactérias, independentemente do gênero e da espécie.

[0079] Dispositivo eletrônico de armazenamento de dados

[0080] No presente pedido de patente, o termo deve ser entendido como qualquer dispositivo eletrônico capaz de armazenar dados/informações biológicas como, por exemplo, de sequências genômicas. Exemplos não limitantes de dispositivos eletrônicos incluem computadores pessoais (PCs), computadores portáteis, máquinas que atuam como servidores, smartphones, máquinas para computação paralela, dentre outros. Além disso, o termo também engloba os dispositivos eletrônicos capazes de armazenarem bancos de dados de informações biológicas como, por exemplo, sequências biológicas. Também compreende os dispositivos eletrônicos que são capazes de realizar

conexão / comunicar-se com outros dispositivos eletrônicos, como por exemplo um servidor.

[0081] Dispositivo eletrônico para processamento de sequências genômicas

[0082] No presente pedido de patente, o termo deve ser entendido como qualquer dispositivo eletrônico que é capaz de processar sequências genômicas. Exemplos não limitantes de dispositivos eletrônicos incluem computadores pessoais (PCs), computadores portáteis, máquinas que atuam como servidores, smartphones, máquinas para computação paralela, dentre outros.

[0083] Dentre as diversas razões técnicas das vantagens apresentadas pela presente invenção, destacam-se: rapidez e baixo custo na obtenção de resultados já que este método não exigem clonagens e nem cultivos *in vitro*. Para identificação dos possíveis organismos realiza-se somente o sequenciamento do DNA total extraído da amostra alvo a ser analisada. Este sequenciamento, hoje, com os sequenciadores modernos é um processo relativamente barato e rápido.

Exemplo 1. Concretização Preferencial

[0084] Os exemplos aqui mostrados têm o intuito somente de exemplificar uma das inúmeras maneiras de se realizar a invenção, contudo sem limitar, o escopo da mesma.

Exemplo I

[0085] Obtenção dos genomas no *GenBank*

[0086] É realizado um FTP (*File Transfer Protocol*) do repositório de sequências genômicas de procariontes do *GenBank*, diretamente do sítio da *National Center for Biotechnology Information* (NCBI), sendo selecionadas todas as sequências que correspondem a genomas completos de organismos.

[0087] Geração dos Arquivos Castas e Tastas

[0088] As sequências genômicas selecionadas no *GenBank* foram divididas em, aproximadamente, 13.500.000 (treze milhões e quinhentos mil)

arquivos de mil bases cada. Este tamanho é para manter similaridades com as sequências geradas por sequenciadores Sangers. O algoritmo responsável por gerar estes arquivos gera dois tipos de arquivos: arquivo com extensão casta, com sequências no sentido do genoma depositado com tamanho de mil nucleotídeos e outro arquivo com extensão tasta, com sequências no sentido inverso ao genoma depositado. Em nenhum dos casos as sequências foram invertidas.

[0089] Armazenamento dos dados dos organismos

[0090] 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 Tabela 1.

Tabela 1. Exemplo de dados descritivos armazenados em arquivos castas/tastas (adaptado do arquivo original casta/tasta)

Locus	Nome	Sequência	Taxid	Espécie	Gênero
NC014750	Marivirga tractuosa DSM 4126 uid60837	0	643867	*Marivirga tractuosa*	Marivirga
NC014497	Candidatas Zinderia insecticola CARI uid52459	6976200	522306	*Accumulibacter phosphatis* UW-1	Candidatus Accumulibacter phosphatis Hesselmann et...
NC017033	Frateuria aurantia DSM 6220 uid81775	2609400	81475	*Acetobacter aurantium* (sic) Kond and Ameyama 1958	Frateuria

Tabela 1. Continuação da Tabela 1

Locus	Família	Ordem	Classe	Filo
NC014750	Flammeovirgaceae	Cytophagales	Cytophagia	Bacteroidetes
NC014497	Candidatus Accumulibacter Hesselmann et al. 1999	Undassified Betaproteobacteria	Betaproteobacteria	Proteobacteria
NC017033	Lysobacteraceae classe	Xanthomonadales	Gammaproteobacteria	Proteobacteria

[0091] Conversão dos arquivos de extensão casta e de extensão tasta

[0092] Os arquivos com extensão casta ou com extensão tasta (neste pedido de patente também referidos simplesmente como arquivos castas e tastas, respectivamente) são convertidos da seguinte forma: nucleotídeos citosina são convertidos para 0 (zero), nucleotídeos timina são convertidos para 1 (um), nucleotídeos guanina são convertidos para 4 (quatro) e nucleotídeos adenina são convertidos para 3 (três). Esta atividade prepara os arquivos para os cálculos que foram executados nas próximas etapas.

[0093] Geração da clusterização (ou Gera Clusterização)

[0094] Numa perspectiva matemática (Gerhardt *et al*., 2006), 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. O número de vértices da rede é *N* e o número de ligações de rede é *L*. A sequência de DNA (DSN) 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. Consideram-se aqui janelas de tamanho L definidas ao longo da sequência.

[0095] A matriz adjacente *m* é um gráfico de uma matriz quadrada de tamanho *N*. O elemento *mi*,*j* é 1 (um), se o vértice *i* está ligado ao vértice *j* e 0 (zero) caso contrário. Com o objetivo de quantificar a organização hierárquica destes, elegem-se gráficos, é definido o coeficiente de clusterização.

[0096] 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:

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

Equação 1

[0097] Para encontrar *ci* é normalizado *Li* ao número máximo possível de ligações entre os vértices *ki* ligado ao vértice *i*, isso significa:

$$c_i - \frac{2\mathcal{L}_i}{k_i(k_i - 1)}$$

Equação 2

[0098] Quando *ki* = 0 ou 1, definimos *ci* = 0. Finalmente *C*, o coeficiente de clusterização para o gráfico é obtido através de uma média:

$$C = \frac{1}{L} \sum_{i-1}^{L} c_i$$

Equação 3

[0099] O coeficiente de clusterização de cada um dos arquivos casta e tasta é obtido pela Equação 3, onde L é o tamanho da janela da sequência de DNA utilizada. Neste trabalho foram utilizadas janelas de 250 (duzentos e cinquenta) nucleotídeos conforme Gerhardt et al. (2006).

[0100] Geração da entropia (ou Gera Entropia)

[0101] Os arquivos castas e fastas são a entrada desta etapa, gerando-se a entropia de prevalência dos triplets de cada um destes arquivos. A entropia é definida pela clássica entropia de Shannon como:

$$S = -\sum_{i=1}^{64} P_n log P_n \,,$$

Equação 4

[0102] 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 4 dessa forma é, naturalmente, influenciada pelo conteúdo GC em si. Assim, pode-se definir uma entropia normalizada como:

$$Sn = S - S_{GC}(rand),$$

Equação 5

[0103] onde S_{GC}(rand) é a entropia calculada por uma sequência aleatória com conteúdo GC. Este procedimento garante que a Equação 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 a degeneração natural dos aminoácidos.

[0104] Geração da Periodicidade (ou Gera Periodicidade)

[0105] Na etapa de verificar e calcular/gerar a periodicidade, transforma-se 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, foi utilizada uma função correlação *Fc (n)* do tipo

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

Equação 6

e 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.

[0106] Usando a Equação 6, é possível observar as variações das repetições que aparecem na sequência genômica em análise. As oscilações que ocorrem na Equação 6 representam repetições em sequências. Para que se entenda a contribuição de cada período, utiliza-se qualquer tipo de análise espectral, sem restrições. Neste problema específico, uma transformada de Fourier pode resolver as frequências envolvidas. A definição clássica da transformada de Fourier é indicada na Equação 7 a seguir:

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

Equação 7

[0107] 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 da sequência analisada em termos de periodicidade.

[0108] Armazenamento dos resultados (ou Armazena Resultados)

[0109] A Figura 3 exemplifica os valores de clusterização (d), entropia (s), periodicidade (p3) e o percentual de GC (%gc) de cada um dos 562.781 arquivos CASTAS e FASTAS de *E. coli*. Observa-se que os valores, na sua maioria, mantém-se constantes em cada fragmento do genoma analisado.

[0110] O somatório dos valores encontrados (clusterização, entropia e periodicidade) está indicado na Figura 3, e a média aritmética, a mediana e o desvio padrão para cada um destes valores (clusterização, entropia e periodicidade) são calculadas também. Para cada genoma completo analisado

é armazenada a tupla (uma linha formada por uma lista ordenada de colunas representa um registro no banco de dados.):

<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:

[0111] **locus** é a codificação da entrada do organismo no GenBank;

[0112] **pares de base** é o total de nucleotídeos do genoma;

[0113] **sequência** é um número inteiro;

[0114] **s_meio** é valor médio da entropia de Shannon calculada;

[0115] **s_desvio** é o desvio padrão da entropia;

[0116] **s_mediana** é a mediana da entropia;

[0117] **d_medio** é o valor médio do coeficientede clusterização;

[0118] **d_desvio** é o desvio padrão do coeficiente de clusterização;

[0119] **d_mediana** é a mediana do coeficiente de clusterização;

[0120] **p3_medio** é o valor da média da periodicidade;

[0121] **p3_desvio** é o desvio padrão da periodicidade;

[0122] **p3_mediana** é o mediana da periodicidade;

[0123] **%GC** é a média dos percentuais de nucleotídeos GC;

[0124] **Data_inclusão** é a data que o organismo foi inserido na base;

[0125] **tipo** identifica se é uma genoma, tasta ou casta

[0126] **Desvio_%GC** é o desvio padrão do %GC.

Tabela 2. 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

[0127] A Tabela 2 exemplifica as quatro métricas (s é a entropia, d é o coeficiente de clusterização, p3 é a periodicidade e Perc GC é o percentual GC) com como seus respectivos desvios padrões. Para efeitos de visualização foram ocultadas as colunas s_mediana, d_mediana, p3_mediana, data_inclusão e tipo. As colunas desvios (s_desvio, d_desvio, p3_desvio e desvio_gc) são utilizadas como limites inferiores e superiores conforme descrito no método (Equação 9).

[0128] A figura 4 mostra os resultados das métricas que foram 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.*

[0129] Observa-se, analisando-se os gráficos da figura 3, que os organismos classificados pela sua taxinomia possuem valores muito próximos nas quatro métricas (s,d, p3 e %gc) quando agrupados pela sua taxinomia.

[0130] Base de Dados

[0131] A base de dados, implementada em MySql, contém neste exemplo somente três tabelas: a tabela Organismo para armazenar os organismos e sua taxonomia; a tabela Metadados para armazenar os metadados dos organismos que permitem a assinatura dos mesmos e a tabela Análise para armazenar as análises realizadas e que possui a mesma estrutura da tabela Metadados.

[0132] A Figura 6 demonstra uma consulta simples à base de dados da tabela análise onde são armazenados os dados gerados pelo processo gerada pelo Processo Criação de Metadados (Figura 1) bem como os dados gerados pelo Processo de Análise (também representado pela Figura 1). 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 é hgut1000, sequência metagenômica esta que foi analisada no item “Comparativo com Metagenomas Conhecidos” indicado neste pedido de patente.

[0133] O Processo de Análise

[0134] A atividade inicial desse processo (Figura 1) consiste na seleção do genoma/metagenoma alvo, sendo verificado a integridade do arquivo do padrão fasta, 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.

[0135] As atividades Gera Casta/Tasta, Converte, Gera Clusterização, Gera Entropia, Gera Periodicidade e Armazena Análise são iguais às descritas acima neste pedido de patente.

[0136] A atividade de Armazenar Analise, inclui no banco de dados, para cada arquivo CASTA ou TASTA, a tupla

[0137] *<locus,pares de base, smedio, sdesvio, smediana, dmedio, ddesvio, dmediana, p3_medio, p3_desvio, p3_mediana, %GC, datainclusão, tipo, desvio %GC>.*

[0138] Na etapa Seleciona Organismos com Similaridade, através de instruções SQL *(Strutured Query Language)* (Beaulieu 2009), é realizado o produto cartesiano, representado pela Equação 8

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

Equação 8

[0139] 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 AmM, gera três milhões e novecentos e quarenta e dois conjuntos, representado pela tupla.

[0140] *<<locus,pares de base, s_medio, sdesvio, smediana, d_medio, ddesvio, dmediana, p3_medio, p3_desvio, p3_mediana, %GC, datainclusão, tipo, desvio %GC>, <locus,pares de base, smedio, s desvio, s mediana, d medio, d desvio, d mediana, p3_medio, p3_desvio, p3_mediana, %GC, data inclusão, tipo, desvio %GC>>*

[0141] 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

[0142] *A.smedio > = M.smedio - M.sdesvio* E

[0143] *A.smedio < = M.s medio + M.s desvio* E

[0144] *A.dmedio > = M.dmedio -M.d desvio* E

[0145] *A.d medio < = M.d medio + M.d desvio* E **Equação 9**

[0146] *A.p3_medio > = M.p3_medio - M.p3_desvio* E

[0147] *A.p3_medio < = M.p3_medio + M.p3_desvio* E

[0148] *A. %GC >= M. %GC - M.desvio%GC* E

[0149] *A.%GC <= M.%GC + M.desvio%GC*

[0150] Pela Equação 9 acima, a entropia média das sequências Castas e Tastas *(A.s medio*) deve estar dentro do intervalo da entropia média do organismo *(M.smedio)* mais e menos seu desvio padrão *(M.s desvio),* isto é:

$$M.s_{medio} — M.s\_desvio > A.s\_medio < M.s\_medio + M.s\_desvio\ .$$

[0151] O coeficiente de clusterização médio das sequências Castas e Tastas *(A.smedio*) 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 > A. d_medio < M. d_medio + M. s_desvio

[0152] 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:

[0153] *M.p3_medio — M.p3_desvio > A.p3_medio < M.p3_medio + M.p3_desvio* .

[0154] 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 :

[0155] *M.p3_medio — M.p3_desvio > A.p3_medio < M.p3_medio + M.p3_desvio* .

[0156] A instrução SQL exemplo abaixo, implementa a Equação 5.9. Esta mesma instrução foi utilizada para obtenção de dados de controle.

[0157] *SELECT DISTINCT o.nome, count(*)*

[0158] *FROM analise AS a, metadados AS m, organismo AS o*

[0159] *WHERE a.tipo = "ET3"*

[0160] *AND a.p3_medio > m.p3_medio - m.p3_desvio*

[0161] *AND a.p3_medio < m.p3_medio + m.p3_desvio*

[0162] *AND a.s medio > m.s medio - m.s desvio*

[0163] *AND a.s medio < m.s medio + m.s desvio*

[0164] *AND a.d medio > m.d medio - m.d desvio*

[0165] *AND a.d medio < m.d medio + m.d desvio*

[0166] *AND a.percGC > m.percGC - m.desviopercGC*

[0167] *AND a.percGC < m.percGC + m.desviopercGC*

[0168] *AND m.tipo = "medcastas1000"*

[0169] *AND o.locus = m.locus*

[0170] *GROUP BY o.nome*

[0171] *ORDER BY count(*) DESC*

[0172] O resultado da instrução SQL acima é exemplificado na Tabela 3 a seguir:

Tabela 3. Resultados das instruções SQL

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

[0173] Análises Realizadas

[0174] Para validação da metodologia desenvolvida foram realizadas as seguintes análise:

- Análise de uma sequência artificial de *E. coli*, gerada pelo MetaSim (Richter at 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.

[0175] Em todas estas análises, foram executados processos, como mostra a Figura 5, semelhantes ao da obtenção dos metadados definido na Figura 1.

[0176] Sequência artificial de *E. Coli*

[0177] A sequência gerada utilizando o software de simulação MetaSim, gerando sequências padrões as geradas por sequenciadores baseados no Método de Sanger (Sanger and 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.

[0178] A execução do Processo de Análise sobre a sequência artificial de *E. coli* apresentou os seguintes resultados, indicados na Figura 7:

- A sequência "artificial" de *E. coli* versus a base de metadados (Item 5.1.8) gerou um produto cartesiano AnM corresponde 39.122.379 tuplas, das quais 826.774 tuplas atenderam os critérios estabelecidos na Equação 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.* Cabe ressaltar que o próximo organismo *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.*

[0179] Esses resultados mostram que o organismo, a *E. coli,* presente no metagenoma alvo e claramente identificado pelo método da presente

invenção. Os demais organismos encontrados, podem ser classificados como Falsos Positivo (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 3.

[0180] Metagenoma de Sequência Genômica Artificial

[0181] Foi gerado 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.

[0182] A execução do Processo de Análise Metagenômico (Figura 6) sobre o metagenoma artificial apresentou os seguintes resultados:

- A sequência artificialmente gerada versus a base de metadados (Item 5.1.8) gerou um produto cartesiano *A**a**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 selecionados;
- O organismo de maior similaridade nessas 5.554 tuplas foi a *anaplasma,* com uma frequência de 494, como mostra a Figura 8;

- O segundo organismo de similaridade foi a *yersinia,* com uma frequência de 390;

[0183] 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 no item “Base de Dados” do presente pedido de patente, que quanto maior a frequência de um organismo nas tuplas selecionadas maior a probabilidade dele ser um Verdadeiro Positivo (VP).

[0184] Metagenoma de Trinta Gêneros

[0185] Foi gerado um metagenoma de trinta organismos de trinta gêneros diversos, utilizando o MetaSim. A Tabela 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 *AmM* correspondem 30.938.163 tuplas;
- Executando a seleção proposta na Equação 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 encontrados;
- A análise resultou em 457 gêneros, ou seja, 427 gêneros Falso Positivos (FP), mas destes 316 com baixa frequência, inferior a 2000;
- Onze dos 30 gêneros do metagenoma tiveram baixa frequência, inferior a 2000.

[0186] O resultado desta análise é que a metodologia 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.

[0187] Comparativo com Metagenomas Conhecidos

[0188] Foram obtidos as sequências utilizadas para elaboração do metagenoma do Mar de Sargaço (Venter at al., 2004), metagenoma de organismos encontrados do intestino de um adolescente obeso (Ferrer at al., 2013) e metagenoma de organismos encontrados em tubulações de ar condicionado de um *shopping center* localizado em Singapura (Tringe at al., 2008).

[0189] Nestas três sequências foram executados, individualmente, a metodologia implementada pelo processo/método descrito em etapas gerais pela Figura 1.

[0190] Em todas as análises 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.

[0191] A figura 10 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 at al., 2004).

[0192] Já a Figura 11 mostra os mesmos filos encontrados utilizando-se a metodologia proposta conforme indicado no item “Análises Realizadas” neste pedido de patente. Pode-se observar uma distribuição semelhante das quantidades dos filos identificado nas duas abordagens, com exceção das *Crenarchaeotas,* em que o método revelado pelo presente pedido de patente identificou mais segmentos deste filo do que a análise original, conforme já indica a Figura 11.

[0193] As figuras 12 e 13 mostram os resultados das análises do metagenoma do trato digestivo de um adolescente obeso (Ferrer at al., 2013) utilizando-se a metodologia aqui proposta. Em comparativo aos organismos encontrados utilizando-se 16S rDNA (Ferrer at al., 2013), a análise por esse método encontrou uma quantidade muito maior de *Firmicutes,* aproximadamente 93% comparando-os com *Actinibactérias, Proteobactérias* e *Bacteiroidetes.* Análise realizada usando a metodologia aqui proposta, como demonstram as figuras 12 e 13, a quantidade de sequências de *Firmicutes* comparadas com as de *Actinibactérias*, *Proteobactérias* e *Bacteiroidetes,* foi de 3%.

[0194] As figuras 14 e 15 mostram os resultados das análises do metagenoma extraído de dutos de ar condicionado de um *shoppingcenter* localizado em Singapura. A figura 14 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 15 mostra os resultados da análise utilizando-se o método descrita no item “Análises Realizadas” neste pedido de patente. Pode-se observar, pela análise demonstrada na figura 15, que foram encontradas 100% dos organismos encontrados na análise da figura 14, e como ocorreram nas análise demonstradas nas Figuras 11, 12 e 13, não coincidem os percentuais de organismos encontrados nas suas respectivas análises originais (Venter at al.,

2004 (Ferrer at al., 2013) (Tringe at al., 2008).

[0195] Os versados na arte valorizarão os conhecimentos aqui apresentados e poderão reproduzir a invenção nas modalidades apresentadas e em outras variantes, abrangidas no escopo das reivindicações anexas.

Reivindicações

1. Método de análise de sequências genômicas **caracterizado** por compreender as etapas de:

a) selecionar pelo menos uma sequência genômica para análise; e

b) gerar pelo menos o valor de quatro variáveis não correlacionadas selecionadas do grupo consistindo de: valor da entropia de Shannon, coeficiente de clusterização de Gerhardt, percentual dos nucleotídeos guanina e citosina, valor da periodicidade de P3, energia livre total de Gibbs e suas combinações.

2. Método, de acordo com a reivindicação 1, **caracterizado** por compreender a etapa adicional de armazenar os dados obtidos em qualquer uma das etapas a) a b) em um dispositivo eletrônico de armazenamento de dados.

3. Método, de acordo com a reivindicação 2, **caracterizado** pelo armazenamento de dados ser um banco de dados de sequências biológicas.

4. Método, de acordo com a reivindicação 1, **caracterizado** pelo valor da periodicidade de P3 ser obtido por análise espectral.

5. Método, de acordo com a reivindicação 4, **caracterizado** pela análise espectral ser a Transformada Discreta de Fourier ou a Transformada Rápida de Fourier.

6. Método, de acordo com qualquer uma das reivindicações 1 a 4, **caracterizado** pela sequência genômica ser de organismos procariontes.

7. Sistema de análise de sequências genômicas **caracterizado** por compreender:

a) pelo menos um dispositivo eletrônico para processamento de sequências genômicas, sendo que o dito dispositivo eletrônico executa o método conforme definido em qualquer uma das reivindicações 1 a 6; e

b) pelo menos um dispositivo eletrônico para armazenamento ou comparação dos dados obtidos pelo dispositivo eletrônico da etapa a).

8. Sistema, de acordo com a reivindicação 7, **caracterizado** pelo dispositivo eletrônico da etapa b) ser o dispositivo eletrônico para processamento de sequências genômicas da etapa a).

FIGURAS

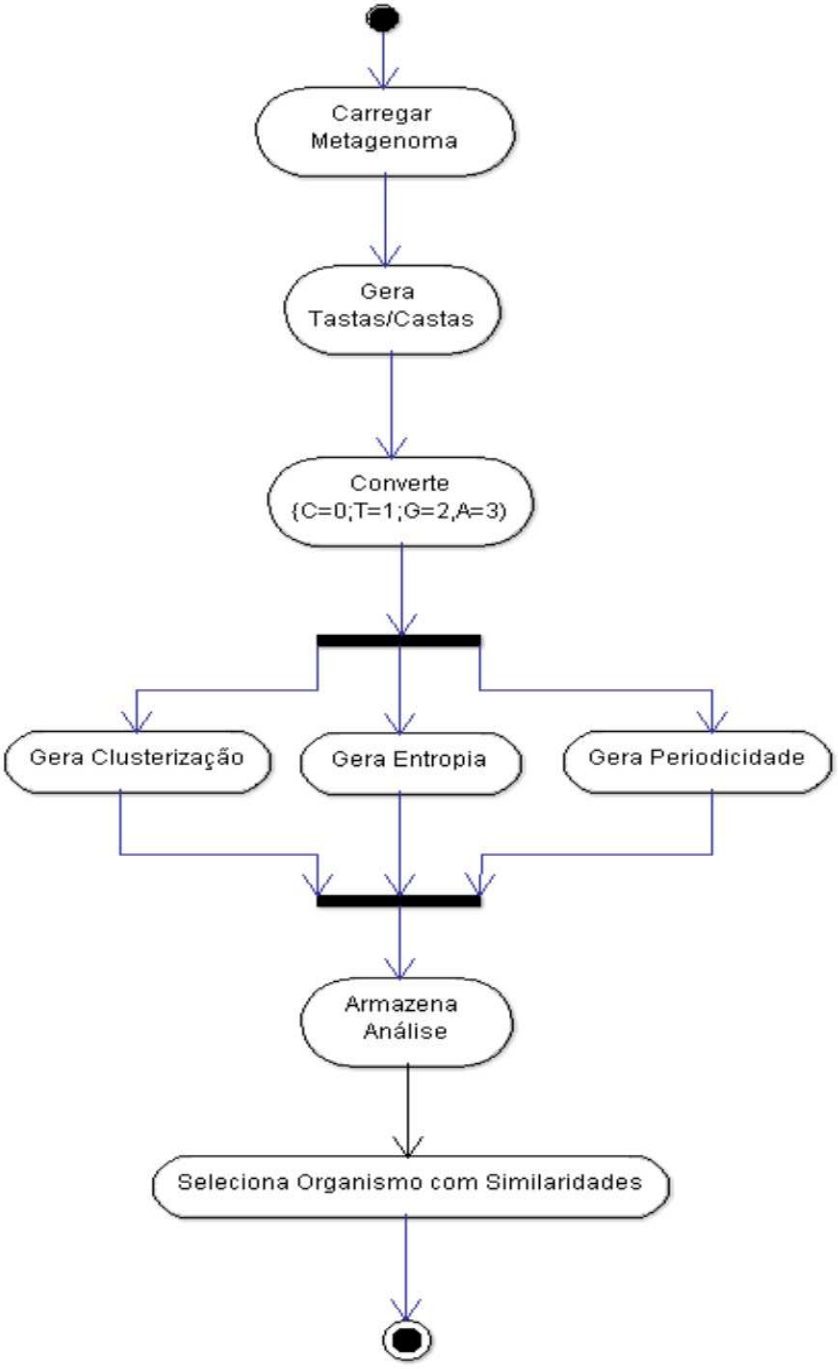


Figura 1

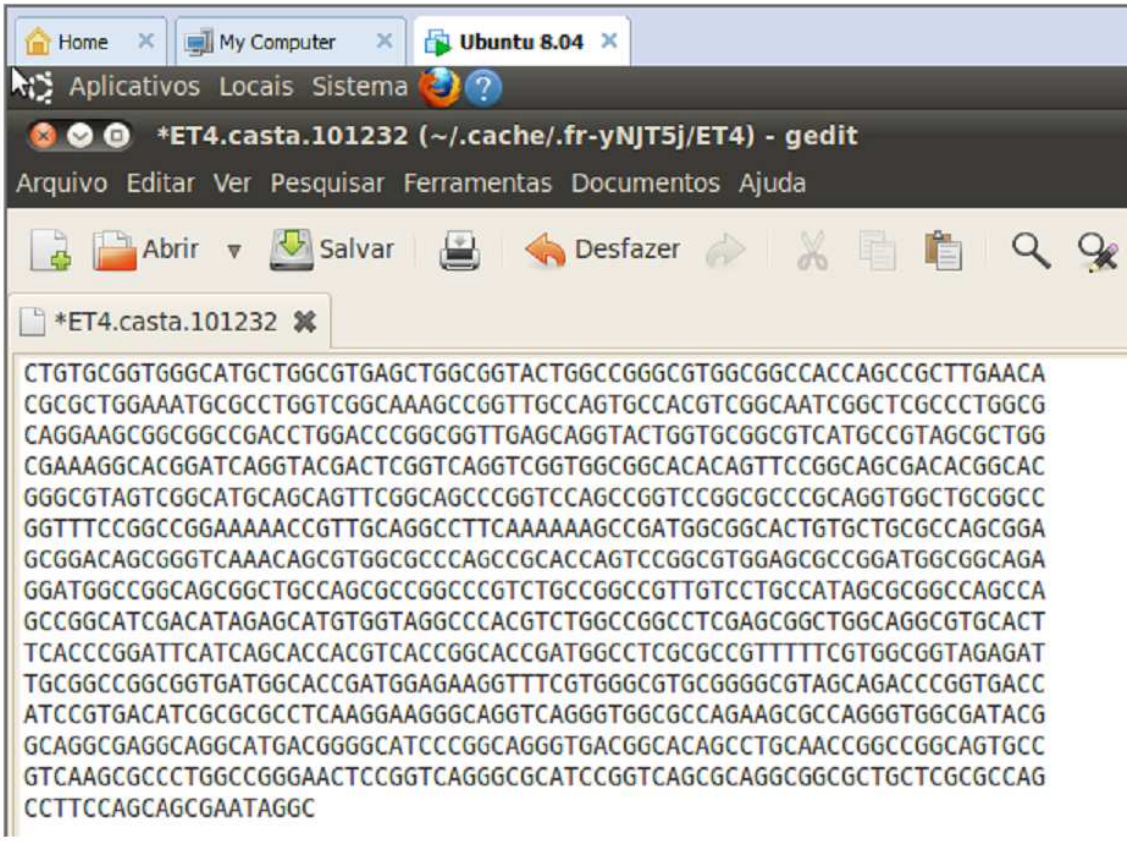

Home
My Computer
Ubuntu 8.04
Aplicativos Locais Sistema
*ET4.casta.101232 (~/.cache/.fr-yNJT5j/ET4) - gedit
Arquivo Editar Ver Pesquisar Ferramentas Documentos Ajuda
Abrir
Salvar
Desfazer
*ET4.casta.101232
CTGTGCGGTGGGCATGCTGGCGTGAGCTGGCGGTACTGGCCGGGCGTGGCGGCCACCAGCCGCTTGAACA
CGCGCTGGAAATGCGCCTGGTCGGCAAAGCCGGTTGCCAGTGCCACGTCGGCAATCGGCTCGCCCTGGCG
CAGGAAGCGGCGGCCGACCTGGACCCGGCGGTTGAGCAGGTACTGGTGCGGCGTCATGCCGTAGCGCTGG
CGAAAGGCACGGATCAGGTACGACTCGGTCAGGTCGGTGGCGGCACACAGTTCCGGCAGCGACACGGCAC
GGGCGTAGTCGGCATGCAGCAGTTCGGCAGCCCGGTCCAGCCGGTCCGGCGCCCGCAGGTGGCTGCGGCC
GGTTTCCGGCCGGAAAAACCGTTGCAGGCCTTCAAAAAAGCCGATGGCGGCACTGTGCTGCGCCAGCGGA
GCGGACAGCGGGTCAAACAGCGTGGCGCCCAGCCGCACCAGTCCGGCGTGGAGCGCCGGATGGCGGCAGA
GGATGGCCGGCAGCGGCTGCCAGCGCCGGCCCGTCTGCCGGCCGTTGTCCTGCCATAGCGCGGCCAGCCA
GCCGGCATCGACATAGAGCATGTGGTAGGCCCACGTCTGGCCGGCCTCGAGCGGCTGGCAGGCGTGCACT
TCACCCGGATTCATCAGCACCACGTCACCGGCACCGATGGCCTCGCGCCGTTTTTCGTGGCGGTAGAGAT
TGCGGCCGGCGGTGATGGCACCGATGGAGAAGGTTTCGTGGGCGTGCGGGGCGTAGCAGACCCGGTGACC
ATCCGTGACATCGCGCGCCTCAAGGAAGGGCAGGTCAGGGTGGCGCCAGAAGCGCCAGGGTGGCGATACG
GCAGGCGAGGCAGGCATGACGGGGCATCCCGGCAGGGTGACGGCACAGCCTGCAACCGGCCGGCAGTGCC
GTCAAGCGCCCTGGCCGGGAACTCCGGTCAGGGCGCATCCGGTCAGCGCAGGCGGCGCTGCTCGCGCCAG
CCTTCCAGCAGCGAATAGGC

Figura 2

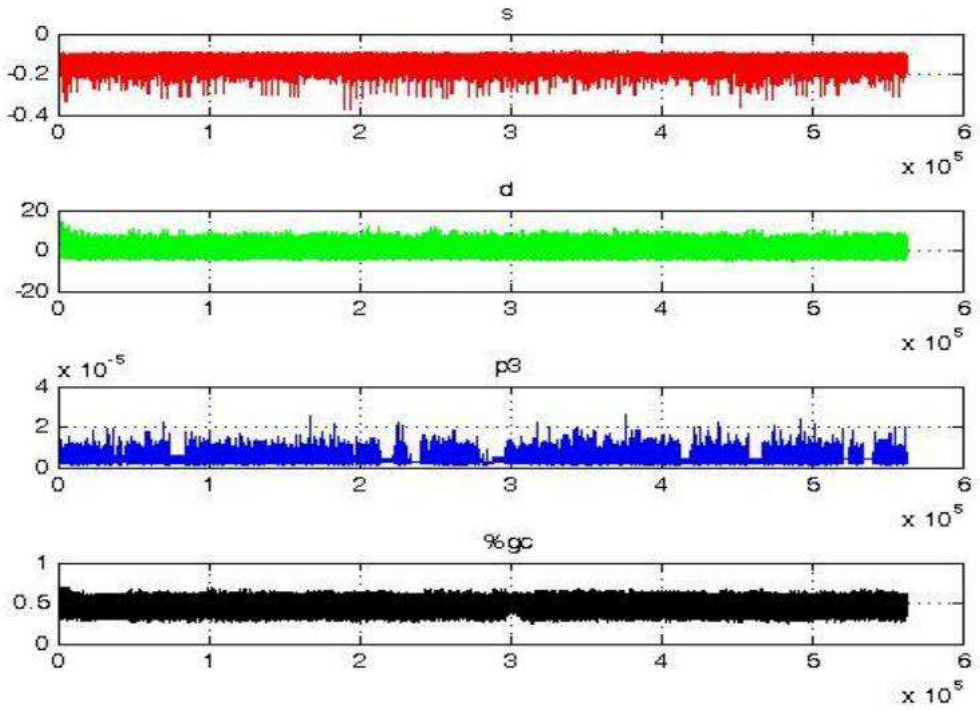


Figura 3

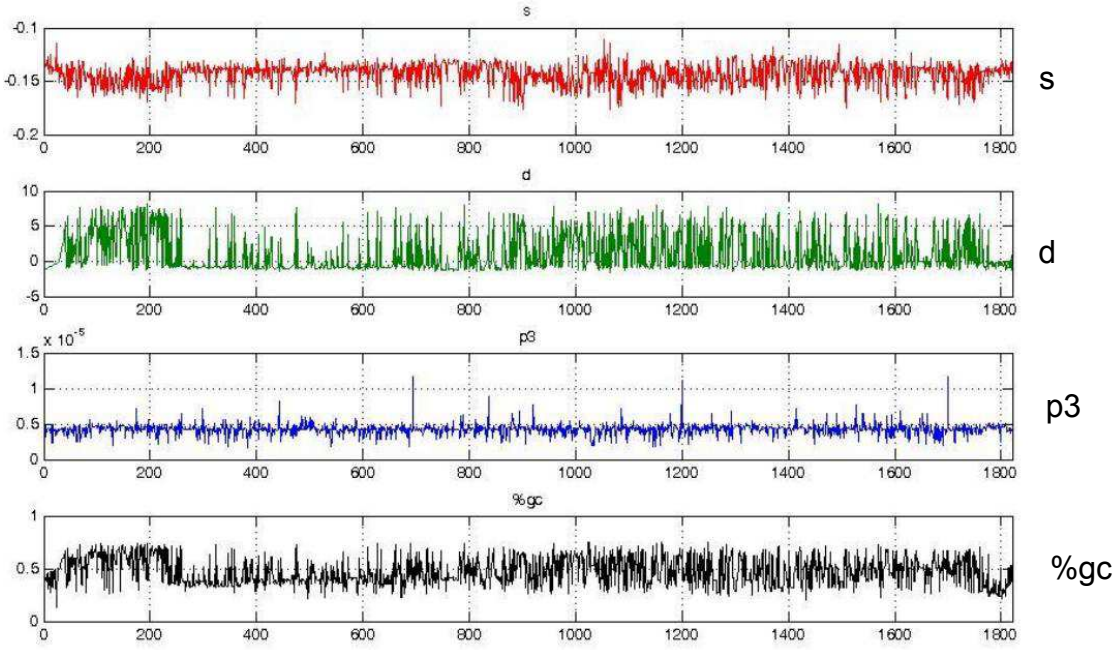


Figura 4

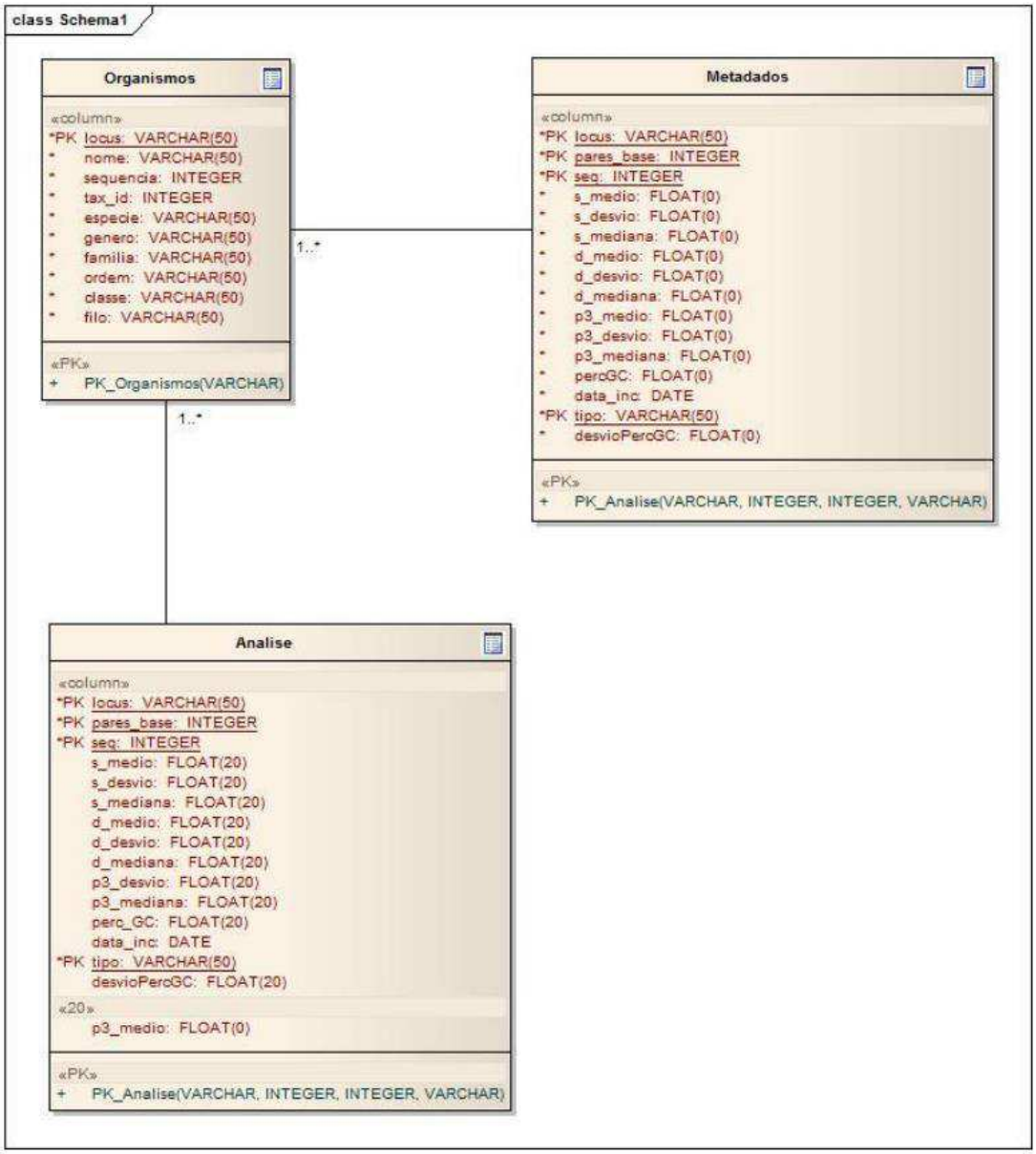
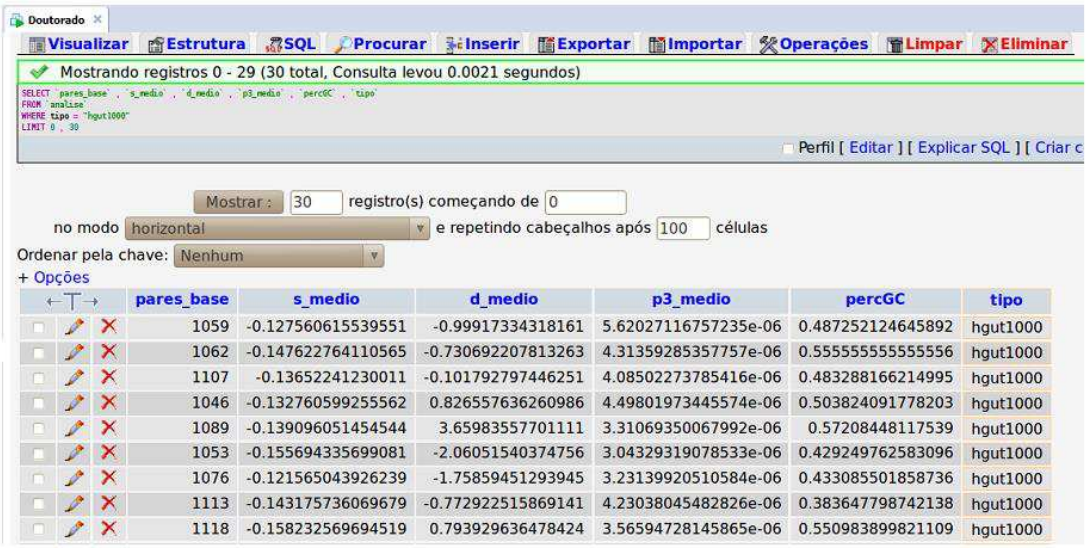

class Schema1
Organismos
«column»
*PK locus: VARCHAR(50)
* nome: VARCHAR(50)
* sequencia: INTEGER
* tax_id: INTEGER
* especie: VARCHAR(50)
* genero: VARCHAR(50)
* familia: VARCHAR(50)
* ordem: VARCHAR(50)
* classe: VARCHAR(50)
* filo: VARCHAR(50)
«PK»
+ PK_Organismos(VARCHAR)
1..*
1..*
Metadados
«column»
*PK locus: VARCHAR(50)
*PK pares_base: INTEGER
*PK seq: INTEGER
* s_medio: FLOAT(0)
* s_desvio: FLOAT(0)
* s_mediana: FLOAT(0)
* d_medio: FLOAT(0)
* d_desvio: FLOAT(0)
* d_mediana: FLOAT(0)
* p3_medio: FLOAT(0)
* p3_desvio: FLOAT(0)
* p3_mediana: FLOAT(0)
* percGC: FLOAT(0)
* data_inc: DATE
*PK tipo: VARCHAR(50)
* desvioPercGC: FLOAT(0)
«PK»
+ PK_Analise(VARCHAR, INTEGER, INTEGER, VARCHAR)
Analise
«column»
*PK locus: VARCHAR(50)
*PK pares_base: INTEGER
*PK seq: INTEGER
s_medio: FLOAT(20)
s_desvio: FLOAT(20)
s_mediana: FLOAT(20)
d_medio: FLOAT(20)
d_desvio: FLOAT(20)
d_mediana: FLOAT(20)
p3_desvio: FLOAT(20)
p3_mediana: FLOAT(20)
perc_GC: FLOAT(20)
data_inc: DATE
*PK tipo: VARCHAR(50)
desvioPercGC: FLOAT(20)
«20»
p3_medio: FLOAT(0)
«PK»
+ PK_Analise(VARCHAR, INTEGER, INTEGER, VARCHAR)

Figura 5



Doutorado

Visualizar | Estrutura | SQL | Procurar | Inserir | Exportar | Importar | Operações | Limpar | Eliminar

Mostrando registros 0 - 29 (30 total, Consulta levou 0.0021 segundos)

```
SELECT `pares_base` , `s_medio` , `d_medio` , `p3_medio` , `percGC` , `tipo`
FROM `analise`
WHERE tipo = "hgut1000"
LIMIT 0 , 30
```

Perfil [Editar] [Explicar SQL] [Criar c

Mostrar : 30 registro(s) começando de 0

no modo horizontal e repetindo cabeçalhos após 100 células

Ordenar pela chave: Nenhum

+ Opções

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 6

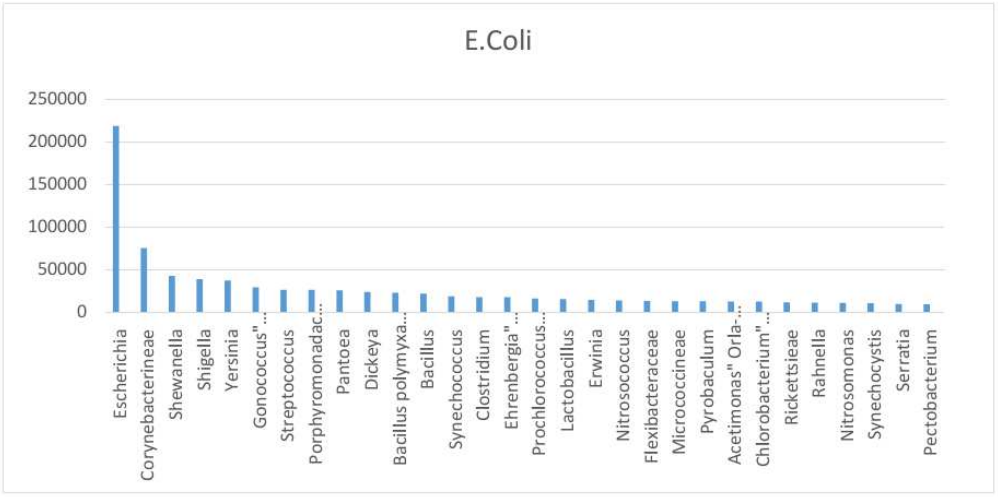


Figura 7

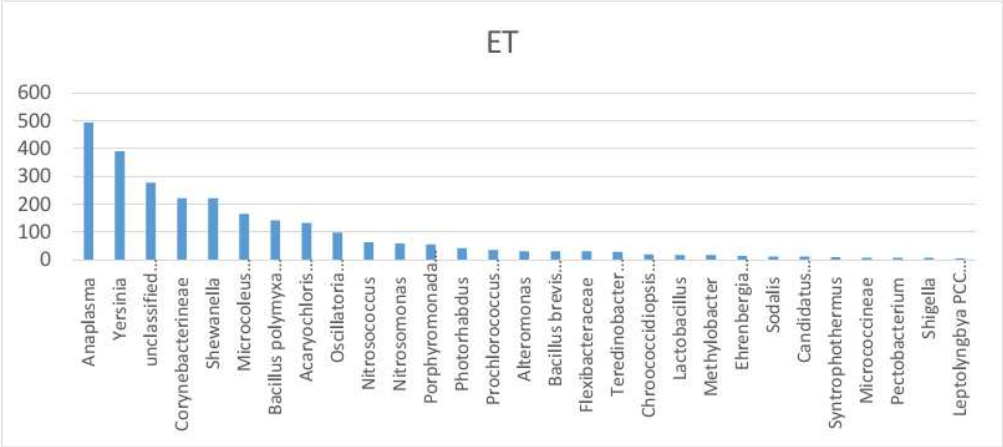


Figura 8

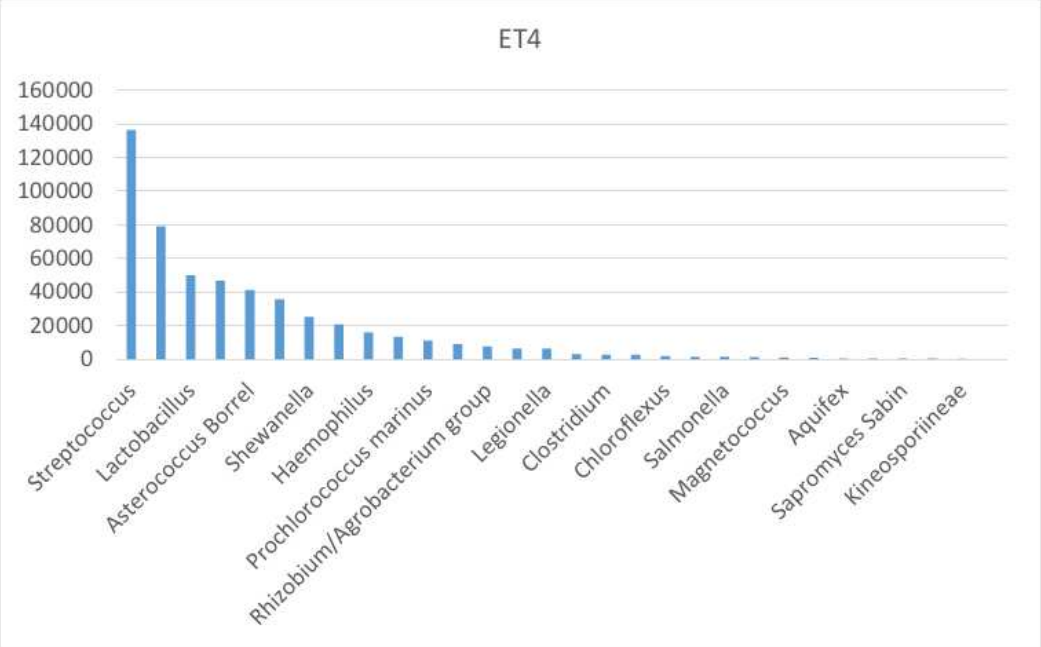


Figura 9

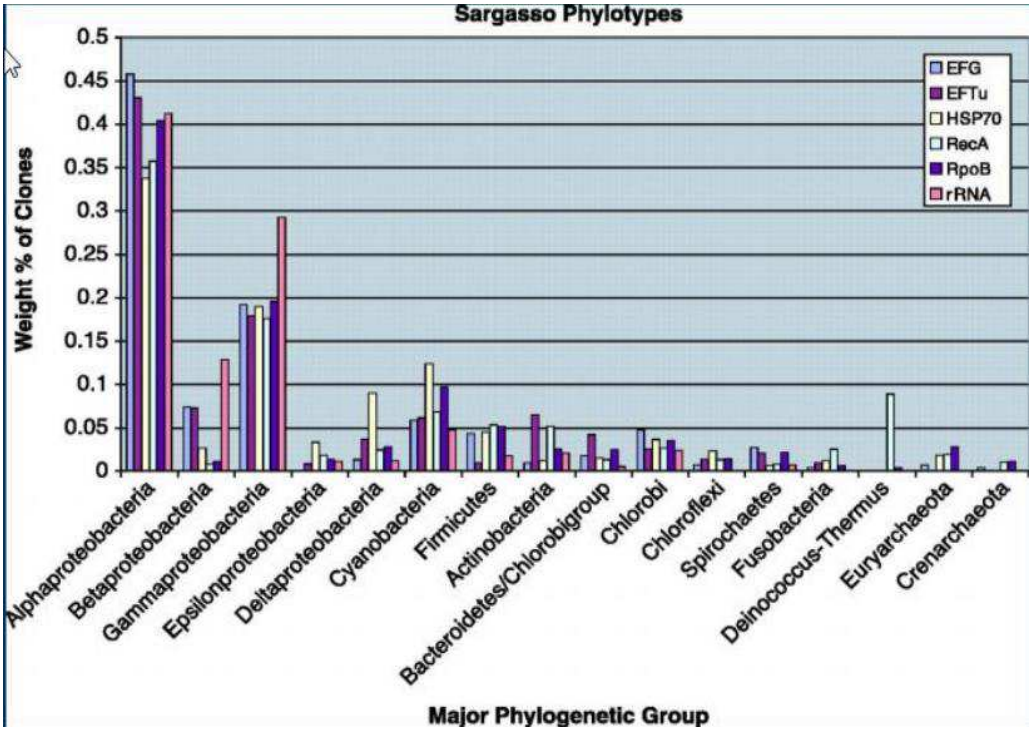


Figura 10

Figura 11

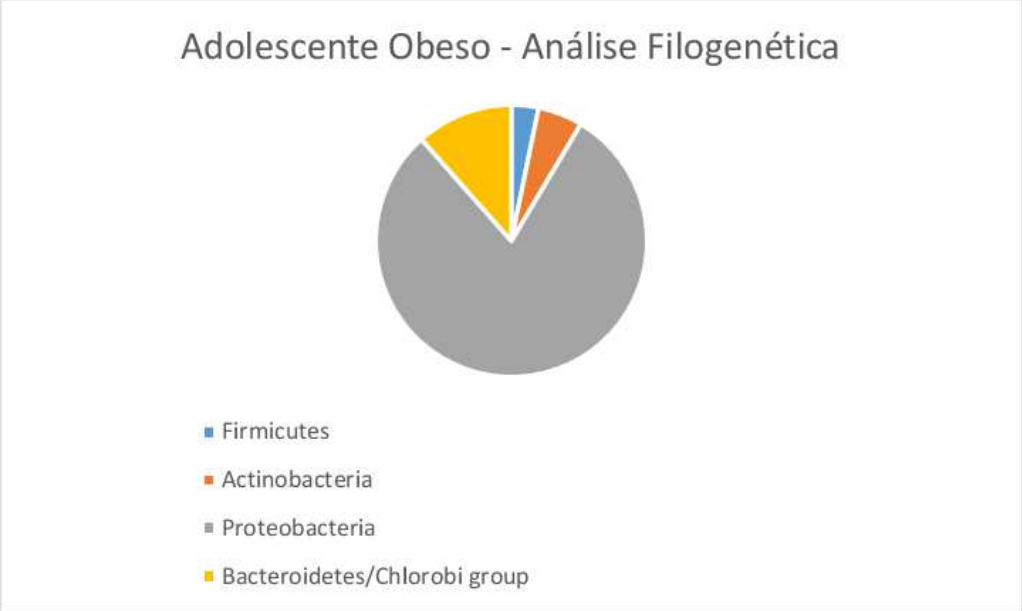


Figura 12

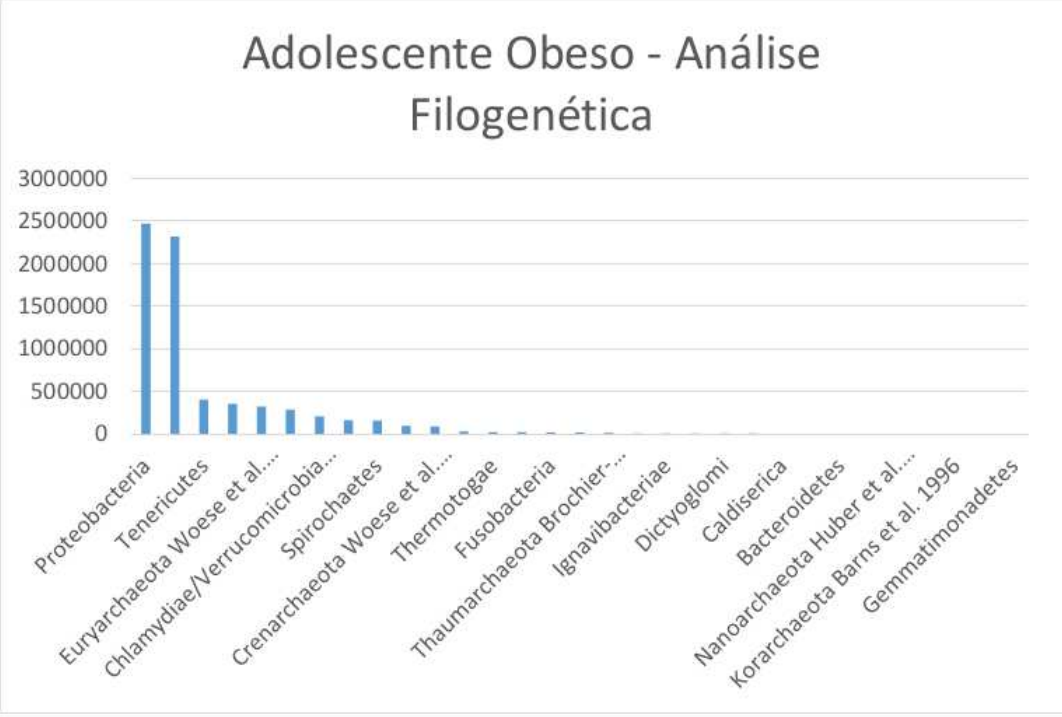


Figura 13

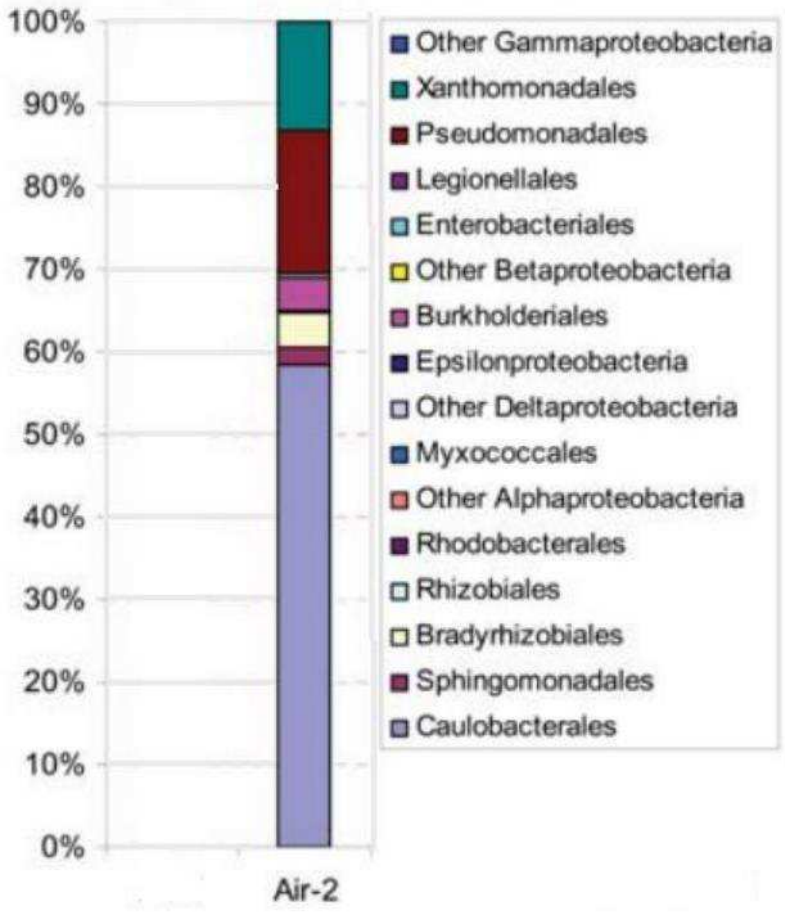

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Air-2
Other Gammaproteobacteria
Xanthomonadales
Pseudomonadales
Legionellales
Enterobacteriales
Other Betaproteobacteria
Burkholderiales
Epsilonproteobacteria
Other Deltaproteobacteria
Myxococcales
Other Alphaproteobacteria
Rhodobacterales
Rhizobiales
Bradyrhizobiales
Sphingomonadales
Caulobacterales

Figura 14

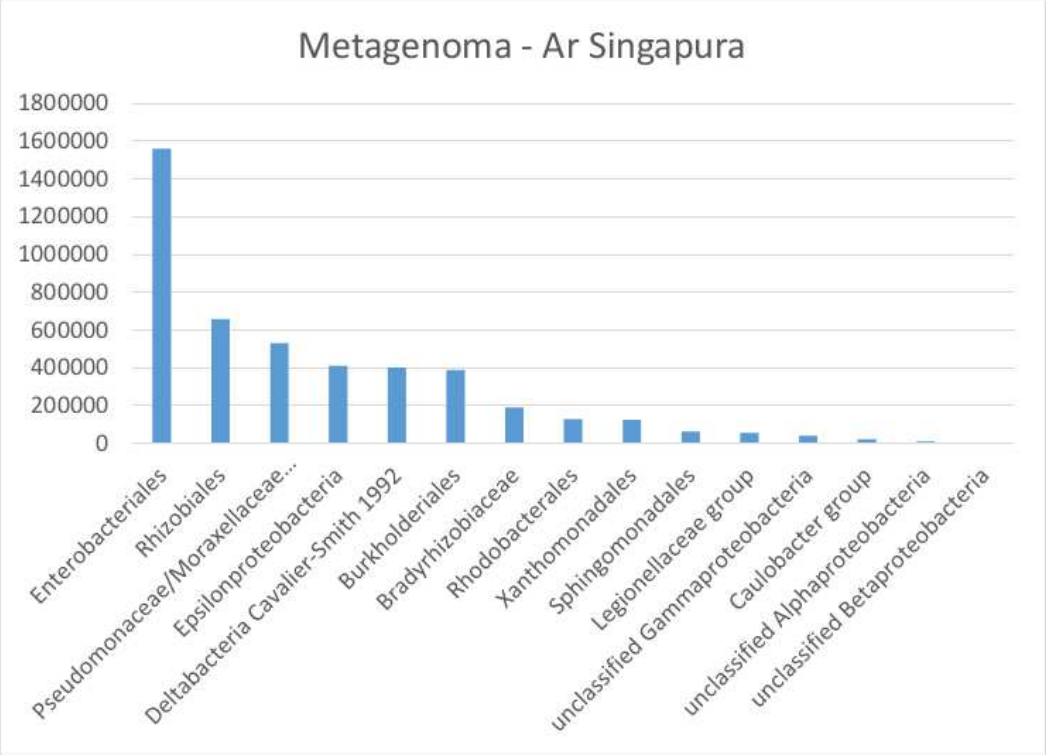

Metagenoma - Ar Singapura
1800000
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
Enterobacteriales
Rhizobiales
Pseudomonaceae/Moraxellaceae...
Epsilonproteobacteria
Deltabacteria Cavalier-Smith 1992
Burkholderiales
Bradyrhizobiaceae
Rhodobacterales
Xanthomonadales
Sphingomonadales
Legionellaceae group
unclassified Gammaproteobacteria
Caulobacter group
unclassified Alphaproteobacteria
unclassified Betaproteobacteria

Figura 15