UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE TECNOLOGIA E GEOCIÊNCIAS

DEPARTAMENTO DE ENGENHARIA BIOMÉDICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA BIOMÉDICA

JOSINEIDE NERI MONTEIRO

IDENTIFICAÇÃO DE BACTÉRIAS DO COMPLEXO *Burkholderia cepacia* ATRAVÉS DE UTILIZAÇÃO DE FERRAMENTAS COMPUTACIONAIS

RECIFE

2017

JOSINEIDE NERI MONTEIRO

IDENTIFICAÇÃO DE BACTÉRIAS DO COMPLEXO *Burkholderia cepacia* ATRAVÉS DE UTILIZAÇÃO DE FERRAMENTAS COMPUTACIONAIS

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Biomédica (PPGEB), da Universidade Federal de Pernambuco (UFPE), como requisito parcial para a obtenção do título de Mestre em Engenharia Biomédica.

Área de concentração: Computação Biomédica

Linha de pesquisa: Inteligência Artificial e Sistemas Inteligentes

ORIENTADOR: Ricardo Yara

RECIFE

2017

Catalogação na fonte

Bibliotecária Margareth Malta, CRB-4 / 1198

M775i Monteiro, Josineide Neri.

Identificação de bactérias do complexo *Burkholderia cepacia* através de utilização de ferramentas computacionais / Josineide Neri Monteiro. - 2017.

75 folhas, il., gráfs., tabs.

Orientador: Prof. Dr. Ricardo Yara.

Dissertação (Mestrado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Engenharia Biomédica, 2017.

Inclui Referências e Apêndices.

1. Engenharia Biomédica. 2. *Burkholderia*. 3. Bioinformática. 4. Alinhamento genético. 5. Algoritmos genéticos. 6. Taxonomia. I. Yara, Ricardo. (Orientador). II. Título.

UFPE

610.28 CDD (22. ed.) BCTG/2017-299

JOSINEIDE NERI MONTEIRO

IDENTIFICAÇÃO DE BACTÉRIAS DO COMPLEXO *Burkholderia Cepacia* ATRAVÉS DE UTILIZAÇÃO DE FERRAMENTAS COMPUTACIONAIS

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia Biomédica e aprovada em sua forma final pelo Orientador e pela Banca Examinadora.

Orientador: _____

Prof. Dr. Ricardo Yara (Doutor pela Universidade de São Paulo – São Paulo , Brasil)

Banca Examinadora:

Prof. Dr. Ricardo Yara, UFPE

Doutor pela Universidade de São Paulo – São Paulo, Brasil

Prof. Dr. Wellington Pinheiro dos Santos, UFPE

Doutor pela Universidade Federal de Campina Grande – Campina Grande, Brasil

Prof. Dr. Otacílio Antunes Santana, UFPE

Doutor pela Universidade de Brasília – Brasília, Brasil

Recife, 08 de setembro de 2016.

Dedico este trabalho a Deus pelas maravilhas que tem feito em minha vida, por me fazer erguer a cabeça mesmo nos momentos mais difíceis da minha vida.

AGRADECIMENTOS

Ao final desse árduo e gratificante trabalho quero expressar minha gratidão àqueles que, de alguma forma, contribuíram para a concretização de uma importante etapa da minha vida profissional.

À Deus, por conhecer todas as minhas dificuldades, aflições, virtudes, defeitos e mesmo assim não desistir de mim.

Aos meus pais pelo dom da vida.

À minha amada mãe pelo apoio, amor e dedicação.

Às minhas queridas irmãs, Janailza e Janicleide por sempre me incentivarem a buscar meus objetivos e se alegrarem com minhas conquistas.

Ao meu sobrinho João Victor pelas alegrias que proporciona na vida da minha família.

À minha tia Leonita pelo apoio desde o período da minha graduação. Serei eternamente grata por tudo o que fez.

À minha grande amiga Lorenna Santos pela amizade, incentivo e apoio durante essa trajetória.

Aos meus amigos Fabiano Pereirae Robson Arruda pelos momentos de aprendizado e descontração vividos ao longo desse período. Tenho certeza que sem vocês não teria tido a mesma alegria e incentivo. Os levarei sempre em meu coração.

Aos meus parceiros de trabalho Ladjane Felix, Vanessa Zaennay e Pedro Felipe pela constante dedicação e apoio durante minha ausência.

Ao meu orientador, professor Ricardo Yara pela constante dedicação e paciência.

Ao professor Wellington Pinheiro pela importante contribuição e direcionamento.

A todos, meu sentimento de gratidão, pois acredito que as pessoas felizes vão sempre lembrar o passado com gratidão e, dessa forma, me alegro com o presente e concluo essa etapa para encarar outros desafios.

“As ideias vieram a mim como vêm a todos nós. A diferença é que levei essas ideias a sério e não deixei ninguém me desencorajar. Eu tinha confiança na minha percepção e não nos dogmas e nas opiniões dos outros e não deixei ninguém me desencorajar e olhe que muitos tentaram, mas a vida não é um concurso de popularidade.”

Jonas Salk, inventor da vacina contra Poliomielite

RESUMO

O gênero *Burkholderia* compreende bactérias gran-negativas, aeróbicas pertencentesà classe β-proteobacteria. Estudos de 16S rDNA revelaram que o gênero *Burkholderia* é composto por bactérias que, apesar de intimamente relacionadas e fenotipicamente muito similares, possuem múltiplas diferenças genéticas, suficientes para permitir subdivisões em espécies ou variantes genômicas, que formam o complexo *B. cepacia*. Dados biológicos, especialmente os de sequenciamento genômico, vêm sendo gerados em ritmo acelerado nas últimas décadas. Com o surgimento da Bioinformática, podemos aplicar técnicas computacionais para manipular dados biológicos. O alinhamento múltiplo de sequências (MAS) é um conjunto de técnicas utilizadas para entender informações biológicas de um conjunto de sequências sendo considerada a tarefa mais comum e mais importante da bioinformática, visto que pode fornecer consideráveis informações sobre estrutura e função de genes. Os algoritmos genéticos (AGs) permitem uma simplificação na formulação e solução de problemas de otimização visto que incorporam uma solução potencial para um problema específico numa estrutura semelhante à de um cromossomo e aplicam operadores de seleção e cruzamento a essas estruturas de forma a preservar informações críticas relativas à solução do problema. O presente trabalho objetivou aplicar técnicas computacionais visando solucionar o problema de alinhamento genético de sequências biológicas de DNA de bactérias do complexo *Burkholderia cepacia*. As sequências analisadas (586) foram obtidas através do banco de dados GenBank do National Center for Biotechnology Information (NCBI). Para alinhamento das sequências, utilizou-se as seguintes ferramentas: Clustal ômega e Kalign. Das ferramentas utilizadas, nenhuma conseguiu gerar dados de boa acurácia. Desse modo, conclui-se que existe a necessidade de desenvolvimento de novos algoritmos/ferramentas de alinhamento genético visando trabalhar com grande quantidade de dados para obtenção de uma otimização. Para o caso de várias sequências, o problema do alinhamento múltiplo é considerado NP-difícil. Desse modo, foi observado que é necessário desenvolver novos algoritmos, para sua resolução em tempo hábil buscando sempre soluções bem aproximadas da solução ótima.

Palavras-chave: *Burkholderia.* Bioinformática. Alinhamento genético. Algoritmos genéticos.Taxonomia.

ABSTRACT

The genus *Burkholderia* comprises gran-negative bacteria, aerobic belonging to β-proteobacteria class. 16S rDNA analyzes have revealed that the genus *Burkholderia* is composed of bacteria which, although closely related and phenotypically very similar, have multiple genetic enough differences to allow subdivisions species or genomic variants that constitute the *B. cepacia complex*. Biological data, especially the genomic sequencing, are being generated at a rapid pace in recent decades. With the emergence of bioinformatics, we can apply computational techniques to manipulate biological data. The multiple sequence alignment (MAS) is a set of techniques used to understand biological information from a set of sequences is considered the most common and most important task of bioinformatics, since it can provide considerable information about the structure and function of genes. AGs allow a simplification in the design and optimization of troubleshooting as incorporate a potential solution to a specific problem in a structure similar to a chromosome and apply selection and crossover operators such critical information to preserve the form of structures for the solution problem. This study aimed to apply computational techniques aimed at solving the genetic alignment problem of biological DNA sequences of bacteria *Burkholderia cepacia complex*. The sequences analyzed (586) were obtained from the GenBank database of the National Center for Biotechnology Information (NCBI). For aligning the sequences, the following tools were used: Clustal omega and Kalign. The tools used, none was able to generate good data accuracy. Thus, it is concluded that there is a need to develop new algorithms / alignment tools genetic targeting working with large amounts of data to obtain an optimization. In the case of multiple sequences, the problem of multiple alignment is considered to be NP-hard. Thus, it was observed that it is necessary to develop new algorithms for its resolution in a timely manner and always seeking approximate solutions of the optimal solution.

Keywords: *Burkholderia.* Bioinformatics. Genetic alignment. Genetic algorithms. Taxonomy.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O gênero *Burkholderia* compreende bactérias Gram-negativas pertencentes à classe β-proteobacteria. São bactérias capazes de metabolizar diferentes fontes orgânicas e isso reflete na capacidade de essas bactérias habitarem vários nichos ecológicos podendo, dessa forma, serem isoladas da água, ambientes hospitalares, solo, rizosfera. Essas bactérias têm sido utilizadas com frequência na biorremediação (MEYER et al., 2001). Desse modo, se reconhece a necessidade em estudar tais microrganismos.

A partir da década de 1980, o uso de ferramentas de genética molecular (hibridização DNA-DNA e sequenciamento do 16S rDNA), aliadasàs técnicas moleculares mais modernas, a exemplo da reação em cadeia da polimerase (PCR), além de sequenciamento de genes específicos para análise de filogenia, levaram à uma modificação e reorganização taxonômica dos gêneros existentes e posterior descrição de novos gêneros (WILLEMS, 2006).

O progressivo avanço das técnicas de biologia molecular associado às mudanças na taxonomia desses microrganismos fazem com que a identificação de novas espécies se torne mais fácil e rápida. No entanto, é fundamental estar atualizado com as novas correntes taxonômicas e atentar para o fato de que novos gêneros e espécies são descritos ou reclassificados de forma constante, visto que se calcula conhecer apenas aproximadamente 12% das espécies bacterianas. Para se definir novas espécies, recomenda-se o uso da “taxonomia polifásica”, a qual integra diversas informações fenotípicas, genotípicas e filogenéticas dos microrganismos em questão, na busca de um consenso (LAJUDIE et al., 1998; VANDAMME et al., 1996).

Através dessas abordagens, uma árvore filogenética de 16S rDNA é a base para construir a classificação das bactérias, e a validação multidimensional é feita examinando-se as características moleculares e fenotípicas dos organismos analisados. A referida metodologia é considerada a abordagem padrão na sistemática bacteriana contemporânea (BOONE & CASTENHOLZ, 2001).

Introduzida por John Holland (HOLLAND, 1975) e popularizados por David Goldberg (GOLDBERG, 1989),Algoritmos Genéticos (AGs) são métodos de

otimização e busca inspirados nos mecanismos de evolução de populações de seres vivos. Os mesmos seguem o princípio da seleção natural e sobrevivência do mais apto, desenvolvido pelo fisiologista Charles Darwin em seu livro “A Origem das Espécies”, em 1859. De acordo com Darwin “quanto melhor um indivíduo se adaptar ao seu meio, maior será sua chance de sobreviver gerando descendentes” (LACERDA; CARVALHO, 1999).

Define-se otimização como a busca da melhor solução para um dado problema. Consiste em tentar várias soluções e utilizar a informação obtida neste processo de forma a encontrar soluções cada vez mais eficazes. Um exemplo básico de otimização é a melhoria da imagem das televisões com antena acoplada ao aparelho. Através do ajuste manual dessa antena, várias soluções são testadas, guiadas pela qualidade de imagem, até a obtenção de uma resposta ótima (LACERDA; CARVALHO, 1999).

AGs são algoritmos probabilísticos que fornecem um mecanismo de busca paralela e adaptativa baseado no princípio de sobrevivência dos mais aptos na reprodução. São algoritmos matemáticos que se inspiram nos mecanismos de evolução natural e recombinação genética. Os conceitos da natureza nos quais os AGs se inspiram são simples. De acordo com a teoria de Darwin, o princípio de seleção faz com que indivíduos mais aptos sejam privilegiados e, dessa forma, com maior probabilidade de gerarem descendentes. Consequentemente, indivíduos com mais descendentes têm mais chance de perpetuarem seus códigos genéticos nas gerações futuras. Tais códigos genéticos constituem a identidade de cada indivíduo sendo representados nos respectivos cromossomos (PACHECO, 1999).

Em seres procariontes a taxonomia pode ser descrita como a ciência que determina a classificação (criação de novas taxas), identificação (alocação de linhagens dentro de espécies conhecidas), além da nomenclatura (VANDAMME et al., 1996) desses organismos. Esta ciência produziu um sistema estável, previsível e altamente informativo que colabora para o avanço de vários ramos da ciência, incluindo não somente a microbiologia, mas também a genômica, ecologia de microrganismos, biotecnologia, evolução, dentre outros (ROSELLÓ-MORA, 2005).

Após o surgimento da taxonomia numérica (SNEATH & SOKAL, 1962) e computacional, dados fenotípicos começaram a ser analisados através de coeficientes numéricos que expressam o grau de similaridade entre linhagens com

o auxílio de ferramentas computacionais. A taxonomia numérica veio proporcionar maior objetividade aos esquemas de classificação microbiana, visto que essa abordagem pressupõe a utilização de um grande número de testes bioquímicos (entre 100 e 200) e uma amostragem diversificada de linhagens, sendo os resultados expressos em percentual (VANDAMME et al.,1996). A aplicação de taxonomia numérica levou a avanços significativos na classificação dos microrganismos, especialmente das bactérias (GOODFELLOW, 2000).

1.1 Motivação e justificativa

Considerando que os testes bioquímicos são inconclusivos pelo fato de os organismos serem fenotipicamente parecidos, aliado ao fato de que a grande maioria dos sistemas comerciais de identificação bacteriana não são capazes de distinguir espécies de forma segura, buscou-se uma abordagem de resolução do problema de identificação taxonômica de bactérias do *complexo Burkholderia cepacia* através da análise do 16S RNAr visto que atualmente é utilizado para diferenciar espécies. Todavia, nem todas as bactérias do *complexo B. cepacia* são diferenciadas por essa abordagem e, dessa forma, surge a necessidade em se utilizar ferramentas computacionais.

1.2 Objetivos

O presente trabalho objetiva fazer um estudo comparativo das diversas técnicas computacionais de alinhamento múltiplo de sequências para identificação de bactérias do *complexo B. cepacia* utilizando sequências 16S RNAr. Os objetivos específicos são:

1.Comparar diferentes ferramentas computacionais na avaliação de alinhamento genético identificando, dessa forma, qual possui maior acurácia;

2. Identificar espécies do *Complexo B. cepacia* através da construção de árvores filogenéticas utilizando algoritmos genéticos;

3. Avaliar a aplicabilidade de AG no estudo taxonômico do *complexo B. cepacia*;

4. Avaliar análise de componentes principais no estudo taxonômico.

1.3 Organização do trabalho

A estrutura deste trabalho está dividida da seguinte maneira: além da parte introdutória, contêm outros quatro capítulos. No capítulo 2 são apresentados conceitos biológicos, bem como taxonomia de bactérias, *complexo B. cepacia* e moléculas 16S e 23S rRNA, um resumo sobre genômica de bactérias e sequenciamento genéticoalém dadescrição de conceitos de Bioinformática, AGs, descrição de métodos de alinhamento Clustal e Kalign, abordagem de técnicas de identificação bacteriana. No capítulo 3, relatamos os materiais e métodos utilizados. O capítulo 4 apresenta os resultados e a discussão. Por fim, uma conclusão sobre o trabalho é apresentada seguida das dificuldades encontradas e sugestões de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 *Burkholderia*–Aspectos gerais

O complexo bacteriano foi inicialmente dividido em cinco espécies (***B. capacia, B. multivorans, B. cenocepacia, B. stabilis, B. vietnamiensis*** (VERMISet al.**,** 2002). Em seguida, foram descritas mais quatro novas espécies baseadas na análise do gene recA: ***B. dolosa, B. ambifaria, B. anthina e B. Pyrrocinia*** (CONYE et al., 2001a). Identificou-se que essas novas espécies classificadas compartilhavam um nível moderado de hibridização DNA-DNA (30-50%), porém uma alta similaridade para genes 16S rRNA (98-99%). Todavia, a identificação dessas espécies ainda é bastante discutida bem como sua classificação pelo fato de ainda não ter sido completamente resolvida (MAHENTHIRALINGAM et al., 2005).

De acordo com perfis de restrição e sequências do gene 16S rDNA, bactérias do gênero *Burkholderia*foram, inicialmente, encontradas em associação com plantas de milho e café (GILLIS et al., 1995). Tem sido dada grande atenção a esse gênero de bactérias, em particular ao *complexo B. capacia*.

Através da análise filogenética do gene *recA*, bem como de sequências de 7 locus (*atpD, gltB, gyrB, recA, lepA, phaC* e *trpB*) foram propostas outras novas espécies: ***B. difusa, B. latens, B. arboris, B. metallica, B. contaminans, B. lata, B. seminalis***(VANLAERE et al., 2008; VANLAERE et al., 2009).

2.1.1 Taxonomia debactérias

O termo taxonomia pode ser estabelecido como a ciência que lida com a classificação (ordenação dos microrganismos de acordo com a similaridade entre eles), identificação (alocação de estirpes desconhecidas dentro de grupos taxonômicos conhecidos compatíveis com suas características) e nomenclatura (nomeação dos grupos de acordo com as regras internacionais descritas pelo International Code of Nomenclature of Bacteria (LAPAGE, 2011; VANDAMME, 1996). Pode-se estabelecer uma classificação significativa utilizando-se processos rigorosos, de forma a evitar erros durante o desenvolvimento de uma pesquisa científica ou durante a reprodução de um produto baseado em culturas microbianas.

Para um melhor consenso taxonômico podemos utilizar diferentes tipos de dados e informações (fenotípicas, genotípicas e filogenéticas). Esse modelo de estudo integrado é chamado de taxonomia polifásica. As informações fenotípicas são obtidas através de estudos envolvendo a expressão dos genes, como análises de proteínas e suas funções, marcadores quimiotaxonômicos ou outras características que correspondam à expressão final dos genes (GILLIS et al., 1995; VANDAMME et al., 1996). Para informação genotípica, utilizam-se ácidos nucleicos (DNA e RNA), enquanto que a informação fenotípica é derivada de proteínas e suas funções e marcadores quimiotaxonômicos. (VANDAMME et al., 1996).

Para estudos genotípicos, algumas técnicas são utilizadas, todas com base na análise do DNA, como por exemplo: a porcentagem de bases nucleotídicas Guanina + Citosina (G+C), a hibridização DNA-DNA (HDD), análise de polimorfismos por padrões de fragmentos de restrições (RFLP), sequenciamento de genes, entre outros (STACKEBRANDTet al., 2002). Uma espécie bacteriana é definida como um grupo de estirpes genomicamente semelhantes isoladas que compartilham um elevado grau de similaridade em relação às várias características independentes (ROSSELLO-MORA; AMANN, 2001).

Uma das estratégias conhecida para os estudos de taxonomia e filogenia bacteriana consiste na análise conjunta de múltiplos genes (*loci*), os quais apresentam uma taxa de evolução mais rápida quando comparados aos genes ribossomais, mas com um nível de conservação suficiente para conter informações evolutivas fidedignas (STACKEBRANDT et al., 2002). A metodologia de análise a ser utilizada depende do nível de resolução taxonômica que se deseja atingir. Quando o objetivo é classificar em nível de gênero ou espécie, nem sempre é necessária a aplicação de mais de uma técnica. Entretanto, quando se objetiva a descrição de novas espécies, sãoimprescindíveis avaliações fenotípicas, genotípicas e filogenéticas.

Atualmente existe um grande volume de informações de sequências de nucleotídeos e aminoácidos de diversas espécies de microrganismos, que podem ser acessadas em bancos de dados disponíveis na internet, a exemplo do NCBI. Para avanço de técnicas cada vez mais elaboradas de análises de sequências, é necessário o aprimoramento de programas matemáticos, estatísticos e

computacionais, utilizados para a organização e avaliação dos dados. Esse grande volume de dados reflete a importância que estas técnicas, especialmente o sequenciamento do DNA, conquistaram dentro de diferentes ramos da ciência, sobretudo na ciência taxonômica (KLENK; GÖKER, 2010).

2.1.2 Histórico dataxonomia deprocariotos

A partir do século XVI, a classificação de organismos vivos foi um tema de grande interesse para os cientistas que pesquisavam a História Natural na Europa. Lineu propôs um sistema binomial de classificação que é uma das bases da classificação atual dos organismos. Publicado em 1758, a décima edição do *Systema Naturae* de Lineu incluía 5.897 espécies de plantas e animais, os dois reinos nos quais ele dividia os organismos vivos. Durante o século 19, a Taxonomia se tornou uma profissão, resultando em um rápido aumento no número de animais e plantas terrestres conhecidos. Estimativas sugerem que existam pelo menos 6 milhões de espécies de bactérias em solos e oceanos (CURTIS et al., 2002).

O conceito biológico de espécie define as espécies em termos de intercruzamento. Mayr (1963), por exemplo, definiu da seguinte forma: “Espécies são grupos de populações naturais que intercruzam e estão reprodutivamente isoladas de outros grupos desse tipo”. A expressão “reprodutivamente isolada” significa que os membros de uma espécie não intercruzam com membros de outras espécies visto que têm alguns atributos que impedem o intercruzamento. A importância do conceito biológico de espécie deve-se ao fato de que insere a taxonomia das espécies naturais no esquema conceitual da genética de populações.

O conceito de espécie procariótica tem sua própria história e resulta de uma série de melhorias empíricas paralelas ao desenvolvimento das técnicas de análise. Entre os taxonomistas microbianos, há um consenso geral de que o conceito de espécie atualmente em uso é útil, pragmático e universalmente aplicável no mundo procariótico. No entanto, este conceito empiricamente concebido não é abrangido por qualquer um dos, pelo menos, 22 conceitos descritos para eucariotas. A espécie pode ser descrita como "um aglomerado monofilético e genomicamente

coerente de organismos individuais que mostram um alto grau de similaridade geral em muitas características independentes e é diagnosticável por uma propriedade fenotípica discriminativa" (ROSSELLÓ-MORA & AMANN, 2001). A melhor utilização dos conceitos surgiu à partir do uso das seguintes informações: marcadores quimiotaxonômicos, seqüenciamento de rRNA e propriedades de DNA.

Os primeiros sistemas de classificação de procariotos eram baseados apenas em algumas propriedades fenotípicas que eram usadas para agrupar linhagens, a despeito de qualquer afinidade evolutiva verdadeira, e por isso foram tidos como artificiais (BERGEY'S, 1934). Todavia, o principal propósito de um sistema taxonômico utilitário é fornecer classificações que sejam úteis para finalidades científicas e práticas, especialmente a identificação e geração de bases de dados contendo informações relevantes sobre tais organismos fáceis de serem acessadas. Tais classificações devem apresentar como características principais: serem estáveis, objetivas e preditivas.

Estes sistemas refletiam as limitações tecnológicas do referido período. Na prática, sistemas baseados em algumas propriedades morfológicas e comportamentais, levaram a sérios erros de classificação microbiana, nos mais diversos grupos bacterianos (BONNE & CASTENHOLZ, 2001). Tais métodos microbiológicos tradicionais baseados em características fenotípicas, como propriedades morfológicas, fisiológicas e bioquímicas, governaram por décadas a taxonomia microbiana fornecendo informação descritiva para a estruturação de diversas taxas bacterianas.

O surgimento da taxonomia numérica (SNEATH & SOKAL, 1962) aliada à computação, possibilitou que dados fenotípicos começassem a ser analisados por coeficientes numéricos que expressam similaridade entre linhagens com o auxílio de um computador. Desse modo, a taxonomia numérica veio proporcionar maior objetividade aos esquemas de classificação microbiana e a abordagem pressupunha a utilização de um grande número de testes bioquímicos (100 a 200) e uma amostragem grande e diversificada de linhagens, sendo os resultados expressos em porcentagens (VANDAMME et al., 1996). A aplicação de taxonomia numérica levou a avanços significativos na classificação dos microrganismos, principalmente bactérias (GOODFELLOW, 2000).

O constante desenvolvimento nas áreas de química, biologia molecular, estatística e informática fez com que a taxonomia de procariotos sofresse profundas alterações na direção de um sistema que refletisse as relações evolutivas entre os organismos, aproximando a classificação microbiana ao melhor possível da realidade biológica. Além disso, o uso da homologia DNA-DNA associada a uma variedade de características ecológicas e fenotípicas na classificação de microrganismos foi denominada de taxonomia polifásica por Colwell (1970ab).

Colwell propôs a integração da informação do nível molecular ao ecológico para obter identificações e classificações mais precisas e confiáveis. Inicialmente, informações genotípica, fenotípica e filogenética poderiam ser incorporadas na taxonomia polifásica, mas a hibridização de DNA-DNA mostrou ter um papel primordial no delineamento de espécies. A abordagem polifásica da taxonomia tem sido praticada nos últimos 20 anos e pressupõe que as descrições de espécie devem refletir relações filogenéticas, além de serem baseadas em hibridização DNA-DNA do genoma total e fornecer informação genotípica, fenotípica e quimiotaxonômica adicional, dando consistência à espécie definida em termos filogenéticos.

Woese & Fox (1977) publicaram o trabalho seminal sobre o uso de sequências do RNAr 16S para a reconstrução da Árvore da Vida. Posteriormente, demonstrou-se que o RNAr 16S seria extremamente útil na afiliação filogenética de bactérias em espécies, gêneros e famílias (WOESE, 1986). O uso do RNAr 16S foi prontamente incorporado à taxonomia polifásica (STACKEBRANDT & GOEBEL, 1987). O constante desenvolvimento dos métodos de sequenciamento de DNA e o acúmulo da informação desequências em bases de dados públicas de livre acesso têm permitido o sequenciamento comparativo de genes homólogos entre linhagens microbianas sendo considerado procedimento padrão em sistemática microbiana.

A aplicação de conceitos e práticas de taxonomia polifásica apresenta forte embasamento filogenético e teve um efeito significativo na classificação microbiana em todos os níveis da hierarquia taxonômica. Em 1969, a crença na divisão dos seres vivos em cinco reinos, proposta por Whittaker, foi desafiada pelo trabalho de Carl Woese e colaboradores, baseado no sequenciamento comparativo de moléculas de RNAr além da evidência genômica e bioquímica associada. Foi proposta que a classificação dos seres vivos fosse substituída por um esquema baseado em três reinos ou domínios: Bactéria, Archaea e Eucarya, sendo os dois

primeiros microbianos e compostos por células procarióticas. O domínio Eucarya, engloba todos os organismos eucariotos, incluindo os microrganismos fungos e protozoários.

Para classificação microbiana, o uso de sequências de DNAr como ferramenta se deu em estudos de diversidade de microrganismos a partir de amostras ambientais. A utilização de metodologias que independem do isolamento e cultivo de microrganismos levou a uma drástica mudança na perspectiva da diversidade microbiana existente no ambiente.

Por meio das sequências de DNAr 16S diversos grupos de microrganismos nunca antes cultivados puderam ser detectados no ambiente comparando-se sequências depositadas em bases de dados, observou-se que muitas delas pertenciam a organismos filogeneticamente não relacionados às divisões bacterianas já existentes (PACE, 1998). Este impacto na visão da diversidade microbiana pode ser exemplificado pelo número de divisões existentes dentro do domínio bactéria. Em 1987 eram 12 divisões, todas elas descritas com base em organismos cultivados. Em 1998, o número de divisões publicado havia subido para 36 (HUGENHOLTZ et al., 1998), sendo 13 delas divisões candidatas, ou seja, sem representante cultivado e descrição formal.

Um levantamento, publicado em 2003, apontou como 53 o número de divisões dentro do domínio bactéria, sendo que aproximadamente 50% destas não possuem representantes cultivados (RAPPÉ & GIOVANNONI, 2003). Um dos maiores desafios para taxonomistas é o cultivo de representantes destas divisões.

Para identificação de espécies bacterianas, podemos isolar ou coletar um número adequado de estirpes do táxon a ser estudado, e usar todas elas para comparações. Evite, embora às vezes impossível, a descrição de uma espécie baseada em uma única estirpe tendo em vista que isso poderia dificultar a identificação de novos isolados. Além disso, podemos reconhecer os taxa relacionados mais próximos através da análise 16S rRNA e características fenotípicas incluindo, dessa forma, as estirpes relacionados nas análises taxonômicas.

A utilização de valores de 70% de similaridade de DNA como limites absolutos para circunscrever a espécie é aceitável. Devemos considerar que uma única espécie pode consistir em vários grupos genômicos que não

necessariamente têm que ser classificados como espécies diferentes. Isso será possível quando uma propriedade fenotípica que identifica cada um deles é encontrada.

Embora os testes comercialmente disponíveis sejam úteis, as informações recuperadas podem ser insuficientes. O fenótipo não é apenas descrito pelo metabolismo, existem por exemplo, marcadores quimiotaxonómicos que produzem informação importante sobre organismos. Quanto mais exaustivamente o fenótipo for descrito, melhor será a circunscrição.

2.1.3 Taxonomia do Complexo *B. cepacia*

A complexidade taxonômica de organismos *B. cepacia* e a dificuldade de identificação geralmente dificultam estudos que podem estabelecer os papéis desempenhados por essas bactérias bem como o significado patogênico. Esta informação é crucial para propor políticas cientificamente fundamentadas para cada um dos problemas acima mencionados (COENYE et al., 2001).

Burkholder, em 1950, descreveu *Pseudomonas cepacia* como o agente causador da podridão bacteriana da cebola. Em 1992, *P. cepacia* e seis outras espécies pertencentes ao grupo de rRNA II do gênero *Pseudomonas* (*Pseudomonas solanacearum, Pseudomonas pickettii, Pseudomonas gladíolos, Pseudomonas mallei, Pseudomonas pseudomallei*, e *Pseudomonas caryophylli*) (PALLERONI et al., 1973) foram transferidas para o gênero *Burkholderia* (YABUUCHI et al., 1992).

Diversos pesquisadores, a partir de meados dos anos 1990 em diante, observaram que havia uma marcada heterogeneidade entre cepas isoladas de *B. cepacia* a partir de diferentes nichos ecológicos. Estas estirpes foram tentativamente classificadas como *B. cepacia* utilizando uma ampla gama de técnicas. A heterogeneidade entre *B. cepacia* além da problemática da correta identificação e avaliação das técnicas utilizadas mostrou que elas poderiam ser classificadas como: não muito sensível, não muito específica ou nem sensível, nem específico. A diversidade de *B. cepacia* aliada à falta de confiabilidade nos

esquemas de identificação levou Vandamme et al. a proceder um estudo taxonômico polifásico.

Estudos taxonômicos polifásicos posteriores identificaram mais dois membros do complexo *B. cepacia*: *B. cepaciagenomovar VI* presentes em cepas isoladas de pacientes com fibrose cística nos Estados Unidos e Reino Unido. Este organismo pode ser fenotipicamente diferenciado de todos os membros do complexo *B. cepacia* exceto *B. multivorans*. O nome *B. ambifaria* (*B. cepacia* genomovar VII) foi proposto para os isolados a partir de amostras ambientais, clínicas e humanos. *B. ambifaria* também contém várias cepas bem caracterizadas para biocontrole. Além disso, foi recentemente mostrado que a espécie *B. pyrrocinia* também pertence ao complexo *B. cepacia* (PALLERONI et al., 1973).

Geralmente, no complexo *B. cepacia*, representantes de diferentes espécies têm valores de hibridação DNA-DNA entre 30 e 60%, enquanto que os valores obtidos a partir de estirpes pertencentes à mesma espécie são geralmente mais elevadas do que 70%. Valores de ligação DNA-DNA obtido com outras espécies de *Burkholderia* são geralmente abaixo de 30%. Estes valores correspondem a categorias definidas como alto parentesco DNA (70% ou superior) entre estirpes de uma única espécie e parentesco DNA não significativa (30% ou menos). Além disso, as semelhanças entre sequências 16S DNAr obtidas a partir de diferentes membros do complexo *B. cepacia* são mais elevados (97,7%) do que semelhanças entre tais sequências e os de outras espécies de *Burkholderia* (97,0%) (MARTÍNEZ-ROMERO, 1994).

2.1.4 Complexo *B. cepacia*

Complexo *B. cepacia* constitui um grupo de bactérias Gram-negativas não fermentadoras da glicose amplamente encontradas no meio ambiente. A maioria das espécies deste gênero foram descritas inicialmente como fitopatógenos. Todavia, estes microrganismos têm sido identificados com uma frequência cada vez maior como patógenos oportunistas em ambiente hospitalar. A principal patologia associada às infecções causadas por espécies do complexo é a síndrome cepacia, frequente em pacientes acometidos pela fibrose cística, sendo

caracterizado por uma diminuição da função pulmonar, com subsequente bacteremia em muitos casos levando o paciente a óbito (SOUZA et al., 2011).

O complexo é formado por 17 espécies. Apresentam aproximadamente 95% de similaridade genética, de acordo com estudos realizados com o sequenciamento do gene recA. As espécies que fazem parte do complexo são: *B. ambifaria, B. anthina, B. arboris, B. cenocepacia, B. cepacia, B.contaminans, B. diffusa, B. dolosa, B. lata, B. latens, B. metallica, B. multivorans,B. pyrrocinia, B. seminalis, B. stabilis, B. pseudomultivorans* e *B. vietnamiensis* (SOUZA et al., 2011).

As espécies constituintes do complexo apresentam crescimento lento em meios de cultura. Diversas vezes o isolamento a partir de amostras clínicas é dificultado pelo crescimento mais rápido de outros microrganismos que podem estar presentes na amostra. Além disso, a identificação laboratorial a partir de testes bioquímicas manuais ou com sistemas disponíveis comercialmente na maioria dos casos é conflitante, pois algumas espécies bacterianas não estão presentes no banco de dados destes sistemas. Além disso, as técnicas moleculares, apesar da sua alta acurácia, não são amplamente acessíveis aos laboratórios de microbiologia clínica (SHELLY et al., 2000). Diante do exposto, conclui-se que a rápida e confiável identificação desses microrganismos a partir de amostras clínicas constitui um fator de grande importância para introdução da terapia antimicrobiana.

2.1.5 Moléculas 16S e 23S DNAr

As moléculas de DNA 16S e 23S presentes no ribossomo são comumente empregadas na taxonomia de procariotos, pelo fato de serem regiões conservadas e se enquadrarem nos conceitos que definem um marcador filogenético relatado por Piaza et al. (2006). A região 23S é bem maior que a 16S, contendo mais informações genéticas úteis em estudos de filogenia (LUDWING et al., 1992). Todavia, o número de sequências presentes nos bancos de dados é pequeno, limitando a comparação de novas sequências.

A caracterização da sequência do gene ribossomal 16S rDNA tem sido amplamente utilizada em estudos evolucionários, taxonômicos e ecológicos, não

apenas para definir taxas, mas também para detectar quais taxas estão presentes (FOX et al., 1992; OLSEN et al., 1994). A amplificação direta via PCR do 16S sDNA a partir de amostras de solo tornou possível o estudo da biodiversidade microbiana sem a necessidade de cultivar o microrganismo em questão (WARD et al., 1990).

Muitas destas técnicas utilizam definições de agrupamento taxonômico que são a princípio, aleatórias. No entanto, tem se desenvolvido uma nova forma, que hoje é pré-requisito nos estudos de diversidade microbiana, chamada Unidades Taxonômicas Operacioanis (OTUs). Tal definição é cientificamente possível de validar universalmente os grupos taxonômicos. Segundo Yang et al. (2004), quando a diversidade microbiana é inferida a partir de *fingerprints* moleculares ou de informações baseadas em sequências, as OTUs individuais devem ser definidas como espécies em potencial.

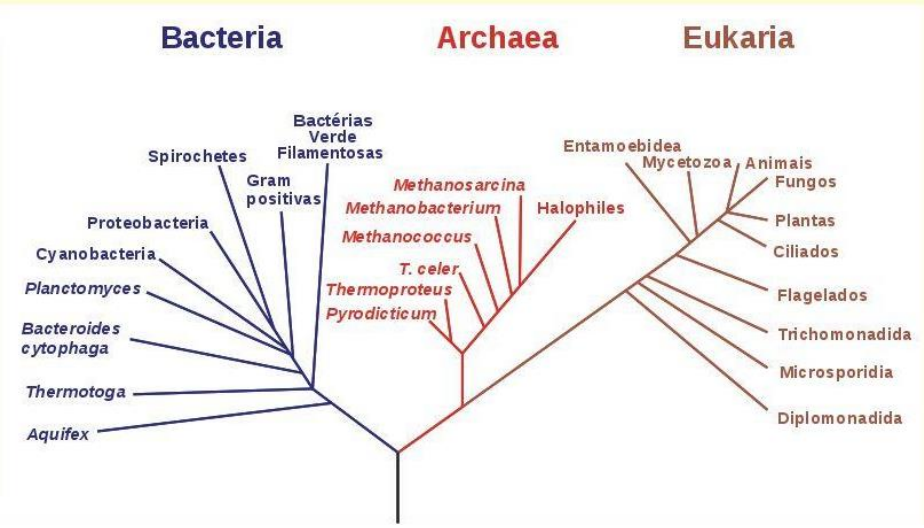
Os métodos que mostraram maior confiabilidade nas análises foram: o sequenciamento, parcial ou total, do gene 16S rDNA, a amplificação do DNA com *primers* específicos pela PCR e a fragmentação do DNA pelas enzimas de restrição, através da técnica de ARDRA (LAGUERRE et al., 2001). Todavia, para uma melhor confiabilidade dos dados, conclui-se que a análise polifásica é mais adequada (DUTTA et al., 2002) sendo usado para o delineamento da taxa em todos os níveis (MURRAY et al., 2012). Os recentes desenvolvimentos na taxonomia polifásica, também chamada de classificação polifásica ou identificação polifásica, constituem um enorme avanço na taxonomia bacteriana moderna (VANDAMME et al., 1996).

2.1.6 Genômica de bactérias

A partir da descoberta de que o ácido desoxirribonucleico (DNA) é responsável por armazenar as informações genéticas,foi iniciada uma busca por uma forma de se obter e decodificar a informação localizada nos cromossomos (MIR, 2004). Um dos grandes desafios da genômica, ciência que estuda a estrutura e funcionamento do material genético de uma espécie, tem sido o sequenciamento rápido de genomas (CHAN, 2005).

Após o surgimento do sequenciamento de DNA desenvolvido por Sanger et al. (1977) começou a haver maior viabilidade em relação ao desenvolvimento de projetos de sequenciamento de genomas. A referida metodologia possibilitou o completo sequenciamento do bacteriófago phi X174. A partir de 1995, com os avanços tecnológicos e utilização dessa metodologia, tornou-se possível fazer o sequenciamento completo de *Haemophilus influenzae* e *Mycoplasma genitalium* por Fleischmann et al., 1995. Após os referidos eventos, houve um grande avanço quando a genômica começou a se aliar a Bioinformática com o intuito de caracterizaros microrganismos presentes na árvore da vida (Figura 1).

Figura 1 - Árvore filogenética baseada no rRNA 16S mostrando que todas as formas de vida são oriundas de um ancestral comum.

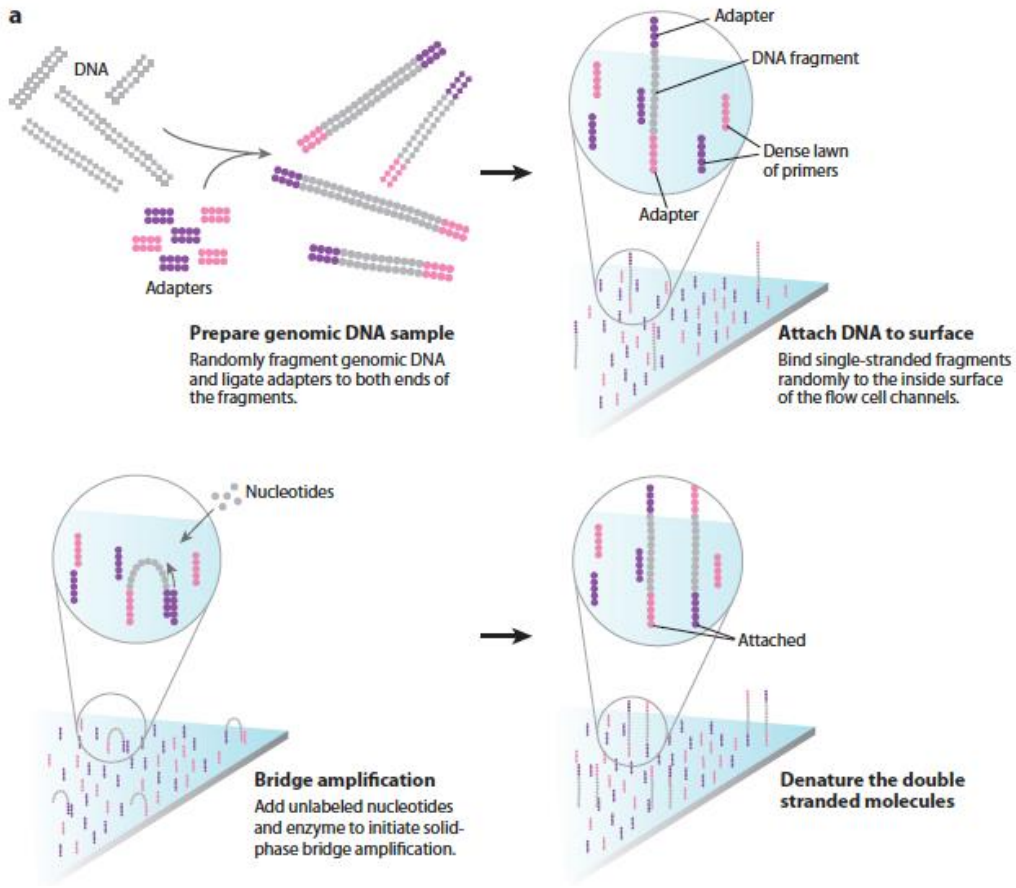


Fonte: Woese et al., 1990

2.2 Sequenciamento genético

O sequenciamento de um gene pode ser definido como um processo através do qual se determina a cadeia de nucleotídeos que o compõe. O fato de um genoma ser extenso para ser sequenciado inteiramente faz com que o mesmo seja dividido em pequenos segmentos, os quais são sequenciados individualmente e, em seguida, ordenados de forma que seja construída uma única sequência a qual corresponderá ao sequenciamento completo do genoma inicial. Esta fragmentação do genoma é comumente realizada através da estratégia *shotgun* (VENTER, 1998) na qual o DNA é submetido a altas taxas de vibração que promovem a quebra da cadeia em vários fragmentos que são geralmente únicos. Após essa etapa, é iniciado o sequenciamento das bases de cada um dos fragmentos através de métodos como o método de Sanger (SANGER; COULSON, 1975) como ilustrado na Figura 2.

Figura 2 - Processo de sequenciamento do Illumina**.**A biblioteca de cadeia dupla é desnaturada para obter DNAs de cadeia única. Estas cadeias simples são dispostas em concentrações muito baixas pelos canais de uma célula de fluxo. Esta "*flow cell*" possui na sua superfície dois tipos de oligonucleotídeos imobilizados complementares aos dois adaptadores, utilizados para produzir a biblioteca de sequenciamento. Estes oligonucleotídeos hibridizam com as moléculas das cadeias das bibliotecas. Por síntese reversa, começando pela zona hibridizada, a nova molécula que está sendo criada encontra-se covalentemente ligada à *flow cell.* Esta nova molécula dobra-se e liga-se a outro oligonucleotídeo complementar ao segundo adaptador que não está ligado à placa, podendo ser usado para sintetizar uma segunda cadeia ligada também covalentemente à placa. Este processo de dobra da molécula e de síntese reversa, chamada de amplificação em ponte é repetido várias vezes e cria aglomerados de milhares de cópias da sequência original, muito próximos na célula de fluxo.



Fonte: Carvalho & Silva, 2010.

O sequenciamento, montagem e anotação do genoma de uma única bactéria, cujo genoma é tipicamente composto por poucos milhões de pares de bases, era uma tarefa difícil (SETUBAL & MEIDANIS, 1997) até o final da década de 1990. Todavia, com o advento dos sequenciadores de alto desempenho desenvolvidos nos últimos anos, tornou-se possível, em um único sequenciamento, a obtenção de grande volume de DNA (SHARON & BANFIELD, 2013). As tecnologias de nova geração começaram a ser comercializadas em 2005 e estão evoluindo constantemente. Elas promovem o sequenciamento de DNA em plataformas capazes de gerar informação sobre milhões de pares de bases em apenas uma corrida.

2.3 Bioinformática

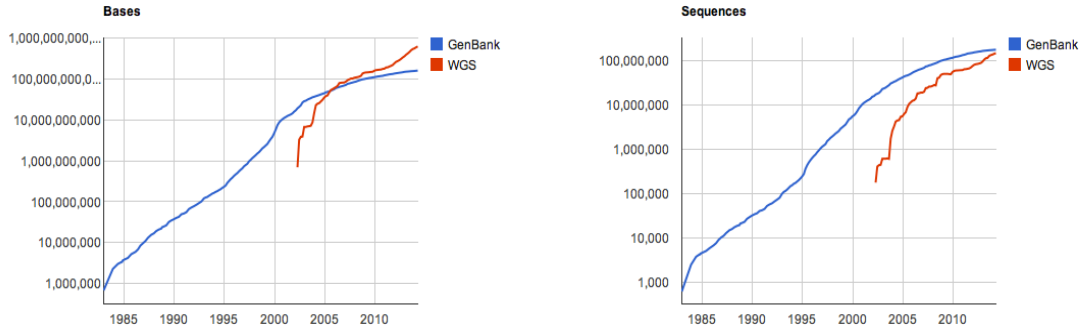
A Bioinformática pode ser definida como a aplicação de técnicas computacionais para manipular dados biológicos (HUGHEY et al., 2001), aplicando técnicas quantitativas e analíticas à modelação de sistemas biológicos. Desenvolve métodos capazes de armazenar e organizar dados biológicos para serem analisados posteriormente, além do desenvolvimento de ferramentas de software capazes de produzir dados de extrema relevância.

Destacam-se como áreas fundamentais da bioinformática: análise de sequências de DNA, análise de expressão genética, análise de regulação da expressão gênica, dentre outras cuja finalidade é estabelecer relações entre os genomas de organismos evolutivamente próximos, visando identificar particularidades, além da possibilidade de se fazer análises filogenéticas.

Segundo Luscombe et al. (2001),a Bioinformática objetiva utilizar informação biológica para preparar, organizar e disponibilizar essa informação para estudos posteriores, facilitando a manipulação e edição de dados através da criação de bancos de dados (a exemplo do NCBI) e redes colaborativas, desenvolvendo ferramentas e recursos capazes de resolverem problemas, além de facilitar a análise desses dados automatizando processos e aumentando a agilidade na obtenção de resultados fidedignos.

De acordo com WEISS (2010), aproximadamente 1000 genomas bacterianos completos estão depositados no GenBank, o banco de dados de Nucleotídeos do NCBI, localizado no National Institutes of Health (NIH). O referido banco de dados armazena informações sobre sequências nucleotídicas de aproximadamente 260.000 espécies (BENSON et al., 2013), dos Institutos de Saúde dos Estados Unidos da América. Bancos de dados similares encontram-se na Europa e no Japão. Abaixo, a Figura 3 representa crescimento do banco de dados GenBank entre 1985 e 2010.

Figura 3 - Crescimento do GenBank. Painel esquerdo crescimento do GenBank em número de bases, painel direito crescimento do GanBank em número de sequências



Fonte: NCBI (2016)

A figura ilustra o número de bases e o número de registros de sequência em cada versão do GenBank. De 1985 até 2010, o número de bases no GenBank dobrou aproximadamente a cada 18 meses. GenBank é o banco de dados de sequência genética NIH, uma coleção anotada de todas as sequências de DNA publicamente disponíveis (Nucleic Acids Research, 2013). O mesmo faz parte da International Nucleotide Sequence Database Collaboration, que compreende o DNA DataBank do Japão (DDBJ), o European Nucleotide Archive (ENA) e o GenBank no NCBI. Essas três organizações trocam dados diariamente.

Uma liberação do GenBank ocorre a cada dois meses e está disponível. As notas da versão atual do GenBank fornecem informações detalhadas sobre o lançamento e as notificações de alterações futuras. As notas de lançamento para versões anteriores do também estão disponíveis. As estatísticas de crescimento do GenBank para as divisões tradicionais do GenBank ea divisão WGS também estão disponíveis.

O banco de dados GenBank foi projetado para fornecer e encorajar o acesso dentro da comunidade científica às informações de sequência de DNA mais atualizadas e abrangentes. Portanto, o NCBI não impõe restrições quanto ao uso ou distribuição dos dados do GenBank. No entanto, alguns autores podem reivindicar patentes, direitos autorais ou outros direitos de propriedade intelectual em toda ou parte dos dados que enviaram (NCBI, 2016).

2.4 Banco dedados do National Center for Biotechnology Information (NCBI)

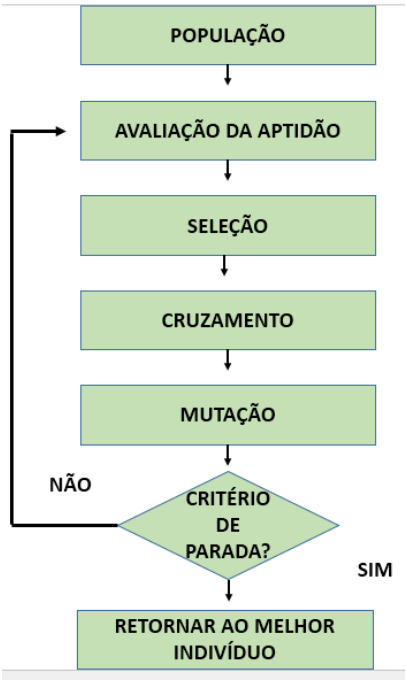
Fundado em 1988, o NCBI é o Centro Nacional de Informação Biotecnológica. O mesmo foi fundado como uma divisão do National Library of Medicine (NLM) no National Institutes of Health (NIH). O site do NCBI contém vários métodos computadorizados de processamento de informações biológicas. NCBI não só realiza pesquisas sobre problemas biomédicos em nível molecular usando matemática e métodos computacionais, mas também fornece inúmeros bancos de dados livres, além de ferramentas de busca moleculares, com ampla documentação de suporte para esses recursos.

2.4.1 Algoritmos genéticos: definição e aplicabilidades

Em meados de 1950 surgiram os primeiros trabalhos relacionados com AGs, quando vários pesquisadores começaram a utilizar sistemas computacionais com o intuito de simular sistemas biológicos. Todavia, o seu desenvolvimento se iniciou de fato a partir de 1970 com uma série de trabalhos publicados por um grupo de pesquisadores da Universidade de Michigan. A partir desse fato surgiram técnicas de soluções de problemas baseados em programação evolutiva, dentro da qual podemos enquadrar os AGs. Apenas recentemente a aplicação dos AGs em problemas de otimização combinatória se tornou um importante tópico de pesquisa (MALAQUIAS, 2006).

Os AGs têm por finalidade simular processos naturais de sobrevivência e reprodução de populações, essenciais em seu processo evolutivo. No processo natural evolutivo, indivíduos de uma mesma população competem entre si, buscando principalmente a sobrevivência, seja através da busca de recursos como alimento, ou visando o processo reprodutivo. Desse modo, indivíduos mais aptos terão um maior número de descendentes, ao contrário dos indivíduos considerados menos aptos. Um dos requisitos para a implementação de um AG é uma população inicial que contenha diversidade suficiente para permitir que o algoritmo combine características e produza novas soluções para o problema em questão. A ideia básica de funcionamento dos AGs é a de tratar as possíveis soluções do problema como indivíduos/espécies de uma referida população que irá evoluir a cada geração (POZO et al., 2000). A Figura 4 demostra a estrutura básica de um AG.

Figura 4 - Estrutura básica de um Algoritmo Genético



Fonte: Pozo et al., 2000

Com referência ao diagrama apresentado na Figura 4, podemos observar que cada iteração do AG corresponde à aplicação de um conjunto de quatro operações básicas: cálculo de aptidão, seleção, cruzamento e mutação. Ao término destas operações cria-se uma nova população, chamada de geração. Desse modo, espera-se que seja representada uma melhor aproximação da solução do problema de otimização que a população anterior. A população inicial é gerada atribuindo-se aleatoriamente valores aos genes de cada cromossomo. A aptidão bruta de um indivíduo da população é medida por uma função de erro, também chamada de função objetivo do problema de otimização. A aptidão bruta é em seguida normalizada (aptidão normalizada), para permitir um melhor controle do processo de seleção. Como critérios de parada do algoritmo em geral são usados a aptidão do melhor indivíduo em conjunto com a limitação do número de gerações. Outros critérios podem envolver, por exemplo, um erro abaixo de um valor especificado pelo projetista para um determinado parâmetro do problema.

Os AGs permitem uma simplificação na formulação e solução de problemas de otimização, visto que incorporam uma solução potencial para um

problema específico numa estrutura semelhante à de um cromossomo e aplicam operadores de seleção e cruzamento a essas estruturas de forma a preservar informações críticas relativas à solução do problema. Normalmente os AGs são vistos como otimizadores de funções, embora a quantidade de problemas para o qual os AGs se aplicam seja bastante abrangente (MALAQUIAS, 2006).

2.4.2 Aplicações de algoritmos genéticos em Bioinformática

Bioinformática é uma área multidisciplinar que utiliza várias técnicas computacionais de matemática aplicada e estatística, visando resolver problemas associados à biologia. Para estudar a evolução e as funções em microbiologia, é necessário comparar moléculas de diferentes espécies. Nessas circunstâncias, as sequências constituem estruturas primitivas que indicam como os aminoácidos se encontram combinados em um gene ou em uma proteína. O alinhamento busca determinar o grau de similaridade entre estas sequências, na sua totalidade ou através de seus fragmentos. Dessa maneira, podemos dizer que um alinhamento é uma forma de organizar sequências de DNA, de RNA ou proteínas, para reconhecer regiões similares indicativas de relações funcionais, estruturais e até mesmo evolucionárias (VIANA; MOURA, 2010).

O constante avanço das pesquisas aliado ao crescente número de sequências biológicas cadastradas, tornou necessária a utilização de sistemas gerenciadores de bancos de dados, mais adequados ao gerenciamento de grandes volumes de informações (DOOLITLE apud BILHA et al., 2005). A grande maioria das informações sobre sequências biológicas estão armazenadas em bancos de dados relacionais ou sistemas orientados. Temos como exemplo o GenBank (BENSON apud BILHA et al., 2005), que é um banco de dados público, que contém as informações biológicas e bibliográficas e é produzido pelo NCBI (BILHA et al., 2005).

O alinhamento de sequências biológicas tem como finalidade comparar uma sequência a outra obtendo trechos semelhantes entre as mesmas, podendo, dessa forma, ter várias aplicabilidades. Métodos para determinação de grau de parentesco; métodos para identificação de um

determinado indivíduo, por exemplo, em caso de identificação criminal e métodos para classificação de espécies, podendo ser utilizado para descoberta de um novo organismo (BILHA et al., 2005).

Resultados de alinhamento são utilizados na análise de genomas ou de regiões conservadoras dos genes que sofreram mutações, bem como para construção de árvores filogenéticas (VIANA; MOURA, 2010). Os algoritmos desenvolvidos para alinhamentos buscam a forma que corresponda ao maior grau de similaridade entre as sequências que estão sendo comparadas. As técnicas têm como prioridade minimizar as diferenças entre elas, ou seja, o objetivo principal é buscar um alinhamento ótimo.

No contexto utilizado na teoria da complexidade, este é um problema de otimização chamado de AVS (alinhamento de várias sequências) onde se procura a solução ótima que corresponde à maior similaridade entre as sequências submetidas ao alinhamento. Não são conhecidos algoritmos que resolvam o problema do AVS em tempo rápido, consequentemente ele é classificado como um problema da classe NP-completo (*Non deterministic Polynomial-time complete*). Uma demonstração desta classificação pode ser vista em Wang & Jiang (1994) (VIANA; MOURA, 2010).

Para entender como se processa um alinhamento e como pode ser computado o grau de similaridade, são apresentados alguns algoritmos desenvolvidos para esse fim. O alinhamento global é o tipo mais comum e envolve a comparação de uma extremidade a outra. Após a inclusão dos espaços, as sequências serão alinhadas “uma sobre a outra” permitindo, desse modo, que seja aplicada uma avaliação do grau de similaridade às mesmas. Programas disponíveis em bases de dados públicas, como o CLUSTAL (2010) realizam este tipo de alinhamento. O alinhamento global é frequentemente utilizado para determinar regiões conservadas entre sequências homólogas, ou seja, que retrata a similaridade entre espécies descendentes de um ancestral comum (VIANA; MOURA, 2010).

Atualmente, os algoritmos mais comumente utilizados são os da família BLAST (*Basic Local Alignment Search Tool*) (MEIDANIS apud BILHA et al., 2005), que estão baseados em programação dinâmica (CORMEN apud

BILHA et al., 2005). Na implementação dos referidos, existem alguns parâmetros que variam de acordo com o banco de dados que está sendo pesquisado (sequência de proteínas ou de DNA). O BLAST utiliza como entrada um banco de dados, que nada mais é do que um arquivo texto organizado em um formato chamado FASTA, contendo as sequências com seus respectivos cabeçalhos. Cada cabeçalho possui algumas informações pertinentes à sequência que o segue (BILHA et al., 2005). A seguir, a Figura 5 ilustra um trecho de um arquivo em formato FASTA.

Figura 5 –Exemplo de arquivo em formato FASTA

```
>B.cepaciastrainNBRAJG97gi|189503756|gb|EU734821.1|:59-
147616SribosomalRNAgene,partialsequence
GTCTGGGAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCGCAAGACCAAAGTGGGGGACCT
TCGGGCCTCATGCCATCAGATGTGCCCAGATGGGATTAGCTAGTAGGTGGGGTAACGGCTCACCTAGGCGACGATCCCTAGCTGGTC
TGAGAGGATGACCAGCCACACTGGAACTGAGACACGGTCCAGACTTCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCA
AGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTTGTAAAGTACTTTCAGCGGGGGAGGAAGGCGATAAGGTTAATA
ACCTTGTCGATTGACGTTACCCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAAT
CGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTCTGTCAAGTCGGATGTGAAATCCCCGGGCTCCAACCTGGGAACTGCATTTCGA
AACTGGCAGGCTAGAGTCTTGTAGAGGGGGGTAGTAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGC
GAAGGCGGCCCCCCTGGACAAAGACTGACGCTCAGGTGCGAAAGCGTGGGGGAGCAAACAAGGATTAGATACCCTGGTAGTCCACGC
CGTAAACGATGTCGATTTGGAGGTTGTGCCCTTGAGGCGTGGCTTCCGGAGCTAACGCGTTAAATCGACCGCCTGGGGAGTACGGCC
GCAAGGTTAAAACTCAA
```

Fonte: NCBI (2016)

Para obter o alinhamento ótimo para o par de sequências AAAC e AGC, por exemplo, o propósito básico desses algoritmos é determinar qual o alinhamento ótimo, visto que pode haver mais de um alinhamento (BILHA et al., 2005). Após identificar as possibilidades de alinhamento, podemos calcular o *score* para cada uma delas. Poderíamos executar o mesmo procedimento, para cada uma das possíveis subsequências (de cada uma das sequências originais) restantes. Este método apresenta o problema de gerar um número exponencial, sendo muitas delas redundantes. Desse modo, não há necessidade de calcular mais de uma vez o *score* do alinhamento de duas colunas em duas subsequências. Entretanto, os resultados devem ser guardados de maneira que possam ser consultados de maneira rápida posteriormente, sendo o princípio básico da programação dinâmica. Em geral é utilizada uma matriz para guardar resultados parciais (BILHA et al., 2005).

Para aplicação em biologia, o importante é obter o alinhamento que tenha mais significado biológico. Quando comparamos sequências oriundas de organismos, procura-se verificar a evidência de que eles tiveram um ancestral comum e é consensual que as divergências ocorreram por processos de mutação ou de seleção natural das espécies em questão. O processo de mutação mais simples considera substituição, inserção e deleção de caracteres, e a seleção natural tem a capacidade de potencializar algumas mutações em prejuízo de outras. É importante frisar que eventos como inversões e transposições de bases não são detectados pelos algoritmos tradicionais. As ferramentas existentes geram matrizes de distâncias que são elementos básicos para geração de árvores filogenéticas. Estes algoritmos comparam genes das espécies em estudo, dispondo as pontuações numa tabela de pesos (escores), de modo que bases nitrogenadas iguais têm escore igual a (+2) e diferentes, igual a (–1) indicando a uma penalidade. Para os deslocamentos, o escore atribuído é nulo (VIANA; MOURA, 2010).

Para o caso de várias sequências, o problema do alinhamento múltiplo é NP-difícil. Desse modo, foi observado que é necessário desenvolver novos algoritmos, para sua resolução em tempo hábil buscando sempre soluções bem aproximadas da solução ótima (VIANA; MOURA, 2010).

Os AGs buscam metodologias de otimização de soluções baseados nos mecanismos de seleção e genética naturais. Eles combinam sobrevivência entre estruturas de sequências “saudáveis” com uma estrutura de troca de informação aleatória. Estes algoritmos em questão também se valem de informações históricas para investigar um novo ponto de busca com um esperado resultado melhorado. Além disso, a eficiência e a eficácia necessárias podem ser adquiridas através da adaptação do AG aos sistemas. Caso a metodologia seja realizada de maneira a alcançar altos níveis de adaptação, os sistemas poderão executar funções mais complexas (SILVA, 2005).

O primeiro trabalho a descrever AG foi *Adaptationin Natural and Artificial Systems*. Muitos artigos e dissertações estabeleceram a validade das técnicas em funções de otimização e aplicações de controle. As razões por trás do número crescente de aplicações estão claras. Muitos trabalhos atuais utilizando

AG no problema de alinhamento estão voltados para o multialinhamento. A tentativa de aplicar o AG ao problema de multialinhamento surgiu em 1993 quando Ishikawa publicou um AG híbrido que não tentava otimizar diretamente o alinhamento, mas a ordem na qual as sequências deveriam ser alinhadas utilizando para isso o processo de programação dinâmica. O referido método limita o algoritmo bem como a função objetiva que pode ser usada com programação dinâmica. Todavia, os resultados obtidos daquele modo estavam fomentando o desenvolvimento do uso de AGs em análise de sequências biológicas (SILVA, 2005).

Descrito por Notredame e Higgins, o primeiro AG capaz de trabalhar com sequências numa maneira mais geral foi relatado uns poucos anos depois, imediatamente antes de um trabalho similar por Zhang. Nestes dois AGs, uma população é feita de multialinhamentos completos de sequências e os operadores têm acesso direto as sequências alinhadas: eles inserem e movimentam gaps numa maneira aleatória ou semi-aleatória (SILVA, 2005).

Durante os anos seguintes, pelo menos três novas estratégias de multialinhamento de sequências baseado em algoritmos evolutivos foram introduzidas. Cada população de alinhamentos múltiplos desenvolve-se por seleção, combinação e mutação. A população é feita de alinhamentos e as mutações programas de processamento *strings* (série de caracteres que são processados como uma unidade de informação) que misturam os gaps usando modelos complexos (SILVA, 2005).

Alinhamento múltiplo de sequências (MSA) é considerado um grande problema em biologia computacional. Define-se o problema do MSA como o arranjo de três ou mais sequências de DNA, RNA ou aminoácidos, sobrepostas. Este arranjo é obtido pelos deslocamentos dos elementos destas sequências obtidos pela inserção de espaços vazios ou lacunas (*gaps*). O MAS é uma técnica utilizada para o estudo da função, estrutura e evolução de moléculas biológicas. Dentre as aplicações do MSA podemos citar a análise filogenética (GUSFIELD, 1997).

O MSA é uma extensão do alinhamento por pares, permitindo que três ou mais sequências sejam alinhadas concomitantemente. Uma pequena

similaridade entre pares de sequências alinhadas pode se tornar altamente significativa na presença de outras sequências. Os alinhamentos múltiplos podem revelar semelhanças sutis que os alinhamentos por pares não são capazes de apresentar (SILVA, 2015).

Algoritmos exatos para alinhamento múltiplo têm complexidade exponencial. Uma alternativa que é frequentemente utilizada é o desenvolvimento de heurísticas, que apesar de não garantirem alinhamentos ótimos, podem fornecer respostas rápidas e razoavelmente boas (BILHA et al., 2005).

Uma das suposições na descoberta de padrões em sequências biológicas é que as regiões conservadas na evolução são importantes do ponto de vista funcional. Assim sendo, é natural usar relações filogenéticas conhecidas entre as sequências para guiar a busca de padrões. Para encontrar um elemento regulatório, em vez de usar regiões regulatórias de vários genes correlacionados da mesma espécie, podem ser usados regiões regulatórias do mesmo gene de várias espécies relacionadas. Assumindo que a árvore evolutiva destas espécies é conhecida, é possível tentar descobrir um padrão pequeno melhor conservado na evolução. Este método é utilizado em Lemos et al., 2003. Todavia, alguns problemas associados à descoberta de padrões são NP-difíceis, ou seja, não existe um algoritmo com tempo polinomial que o resolva. A classe NP pode ser vista informalmente, como a classe dos problemas de decisão para os quais a verificação de que uma solução estimada para uma dada entrada satisfaz todos os requerimentos do problema, pode ser checada rapidamente. Portanto, um problema é NP-difícil se ele é pelo menos tão difícil de resolver quanto qualquer problema em NP (LEMOS et al., 2003).Trata-se de um problema para o qual ainda não é conhecido um algoritmo que o resolva em tempo satisfatório, dificultando com isso a utilização de métodos exatos. Tal fato justifica o emprego de técnicas heurísticas e metaheurísticas (EVEN e SHAMIR, 1976).

Como o algoritmo básico de alinhamento possui complexidade quadrática, vários outros métodos alternativos (heurística) foram desenvolvidos para obter menor tempo de execução, visto que à medida que o volume de dados a ser analisado torna-se maior, o tempo de execução se torna crítico. Existem

vários algoritmos de alinhamento. Podemos classifica-los em famílias, como por exemplo, os da família FAST e os da família BLAST. As famílias representam métodos que são largamente utilizados por pesquisadores da área (MEIDIANIS apud BILHA et al., 2005).

Para solução de problemas deMAS, a utilização de Algoritmos Evolucionários (AEs) tem-se apresentado com relativa frequência no estado da arte. Em 1997, Zhang e Wong (1997) apresentaram uma solução baseada no alinhamento exato de colunas, alcançando bons resultados. No entanto, a solução limita-se ao tratamento de sequências com alto grau de similaridade. Em 2002, Thomsen et al. (2002) apresentaram uma solução que utilizava como entrada alinhamentos já executados pelo Clustal (HIGGINS et al., 1992), visando melhorar resultados anteriormente alcançados. Meshoul et al. (2005) propuseram, em 2005 e 2006, algoritmos que combinam conceitos de computação quântica e algoritmos evolucionários com o intuito de obter maior reprodutibilidade. Gondro e Kinghorn (2007), propuseram um AG tendo dois operadores para crossover, um para combinações horizontais e outro para combinações verticais, e quatro operadores de mutação, todos operando sobre lacunas do alinhamento, garantindo resultados superiores quando comparados ao Clustal W (THOMPSON et al., 1994). Masulli et al. (2010) apresentaram uma solução onde uma matriz com pesos posicionais (MPP) representava um indivíduo da população, representando a probabilidade de uma determinada posição da sequência ser associada a uma coluna do alinhamento (SILVA, 2015).

2.5 ALINHAMENTO MÚLTIPLO DE SEQUÊNCIAS

2.5.1 Clustal

Clustal é um programa de MAS utilizado para alinhamento tanto de DNA quanto de proteínas que tem por finalidade calcular as melhores correspondências para sequências e alinhá-las de forma que suas semelhanças e divergências possam ser observadas. O método foi descrito por Higgins & Sharp em 1988, sendo

projetado para ser eficiente em computadores pessoais da época. Os autores visaram combinar técnicas para o uso eficiente de memória através de algoritmos de programação dinâmica descritos por Myers & Miller, 1988, com a estratégia de alinhamento desenvolvida por Feng & Doolittle, 1987 e Taylor, 1988. De um modo geral, o MSA é construído de forma progressiva utilizando uma série de alinhamento entre duas sequências (HIGGINS & SHARP, 1988; LARKIN et al., 2007; THOMPSON, et al., 1994). Dessa maneira, em uma análise de múltiplas sequências pelo CLUSTAL, as sequências com alto grau de similaridade poderão ser posicionadas de forma mais próxima do que sequências com menor grau de similaridade.

2.5.2 Clustal W

CLUSTAL W éuma ferramenta que dispõe de um ambiente com possibilidade de realizar a leitura de arquivos em diferentes formatos gerando alinhamentos com pouca perda de qualidade em tempo cabível para uma quantidade significativa de dados. Alguns algoritmos de alinhamento múltiplo utilizam cada vez mais heurísticas baseadas em informações biológicas para guiar o processo de MAS e as versões mais atuais dos algoritmos denominadas CLUSTAL W (LARKIN et al., 2007; THOMPSON, et al., 1994) e CLUSTAL OMEGA (SIEVERS; HIGGINS, 2014) podem apresentar um desempenho superior quando comparado às primeiras versões, devido ao acúmulo de conhecimento inserido no software (OGATA, 2007).

2.5.3 Kalign

Kalign é um método de alinhamento rápido e robusto. É especialmente bem adequado para a tarefa cada vez mais importante de alinhar uma grande quantidade de sequências. O algoritmo Kalign segue uma estratégia análoga ao método progressivo padrão para alinhamento de sequências (FENG & DOOLITTLE, 1987). Através dessa ferramenta, são calculadas distâncias entre pares, uma árvore filogenética é construída de modo que sequências/perfis estejam alinhados na ordem dada pela árvore. Em contraste com os métodos existentes, o

algoritmo Wu-Manber (WU S & MANBER, 1992) é utilizado no cálculo da distância e, opcionalmente, na programação dinâmica usada para alinhar os perfis.

2.6 Técnicas de identificação bacteriana

Descrita por Saiki et al. (1985), PCR permite a amplificação de pequenos seguimentos do DNA. A técnica ganhou impulso nos últimos anos e está sendo usada na detecção e identificação de microrganismos em ambientes naturais. Através da referida técnica se obtém (*in vitro*), várias cópias de um seguimento de DNA, previamente conhecido. Para executar a amplificação de certa sequência de DNA é necessário que se faça a extração do DNA, seguida de sua amplificação (utilizado PCR) com algum *primer* (oligonucleotídeo) ou iniciador em um termociclador. Em genética, a PCR pode ser utilizada para identificar e quantificar a variabilidade genética. Esta técnica oferece vantagens por ser rápida e multifuncional, possibilitando que um grande número de genótipos possa ser caracterizado em curto intervalo de tempo (YAMAOKA-YANO & VALARINI, 1998; FUNGARO & VIEIRA, 1998).

Para estudos de diversidade de microrganismos, a técnica é utilizada através de várias metodologias a exemplo da análise de restrição do DNAr amplificado ARDRA (análise de restrição do DNA ribossomal amplificado), polimorfismo dos espaçadores do DNA ribossômico, DGGE (eletroforese em gel por gradiente de desnaturação), TGGE (eletroforese em gel por gradiente de temperatura), PCR de sequências repetitivas de DNA (rep-PCR) e AFLP (STRALIOTTO & RUMJANEK, 1999).

O método ARDRA é uma metodologia que consiste em análises combinadas de sequências de rDNA amplificadas por PCR e digeridas com enzimas de restrição de corte frequente (sítios de 4 pb) gerando padrões de RFLP. Essa técnica foi inicialmente utilizada por Laguerre et al., (1994). A topologia das árvores filogenéticas obtidas por mapeamento dos sítios de restrição e por alinhamento de sequências apresentou-se bem relacionada, mostrando que o método é uma ferramenta poderosa para se obter uma estimativa rápida de relações filogenéticas (LINDSTRÖM et al., 1998).

O valor do método do ARDRA está na sua rapidez e habilidade para avaliar diferenças sutis entre grupos filogenéticos, possibilitando análises em vários níveis taxonômicos, inclusive em estudos de evolução, gerando novos marcadores para estudos de genética de populações (JORGENSEN & CLUSTER, 1989). Esta técnica utiliza enzimas de restrição para fragmentar o DNA em diferentes comprimentos, evidenciando o polimorfismo no comprimento dos fragmentos obtidos. Para identificar os polimorfismos, é necessário que as sequências de nucleotídeos nas fitas de DNA dos organismos sejam distintas (YAMAOKA-YANO & VALARINI, 1998).

Os métodos de REP - PCR (Reação em cadeia da polimerase de sequências palindrômicas extragênicas repetitivas), ERIC - PCR (Reação em cadeia da polimerase de sequências de DNA entre sequências intergênicas consensuais repetitivas de enterobactérias), baseiam-se na amplificação de sequências repetitivas (rep-elements) no genoma bacteriano. Quando um desses elementos repetitivos é detectado dentro de uma distância amplificável durante a PCR, um produto de PCR de tamanho característico é gerado, de modo que o genoma possa gerar padrão de polimorfismo (*fingerprinting*) em um gel (VERSALOVIC et al., 1991). O método é uma poderosa ferramenta para estudar a diversidade genética intraespecífica em nível de estirpe, fornecendo uma análise complementar à prévia caracterização utilizando outras metodologias.

Para avaliar diversidade genética de populações microbianas BOX – PCR vem sendo muito utilizada visto que reúne diversas vantagens, uma vez que é uma técnica rápida, de fácil execução e altamente discriminatória para espécies ao gerar resultados que representam bem as análises baseadas na homologia do DNA-DNA.

Os geneticistas têm utilizado poderosos recursos em estudos de biologia de populações e ecologia, em especial os marcadores genéticos, visando conhecer a estrutura dessas populações. O RAPD (Amplificação Aleatória de Polimorfismos DNA) é um dos marcadores moleculares derivados da técnica de PCR, que gera fragmentos únicos de DNA com apenas um oligonucleotídeo de sequência aleatória (WILLIAMS et al., 1990). A técnica é uma das mais populares variações da PCR e apresenta vantagens em relação a outros métodos, pois requer pequena

quantidade de DNA, além de não necessitar de informações sobre a sequência de nucleotídeos do genoma sendo capaz de revelar alto grau de marcas polimórficas, sendo um método rápido que processa grande número de microrganismos ao mesmo tempo (YAMAOKA-YANO; VALARINI, 1998).

Outra ferramenta de uso crescente na prática de identificação microbiana, em especial bactérias, é a amplificação de regiões específicas do genoma e posterior sequenciamento de bases. A identificação é determinada pela comparação das sequênciasobtidas com a de outros organismos disponíveis no banco de dados do NCBI.

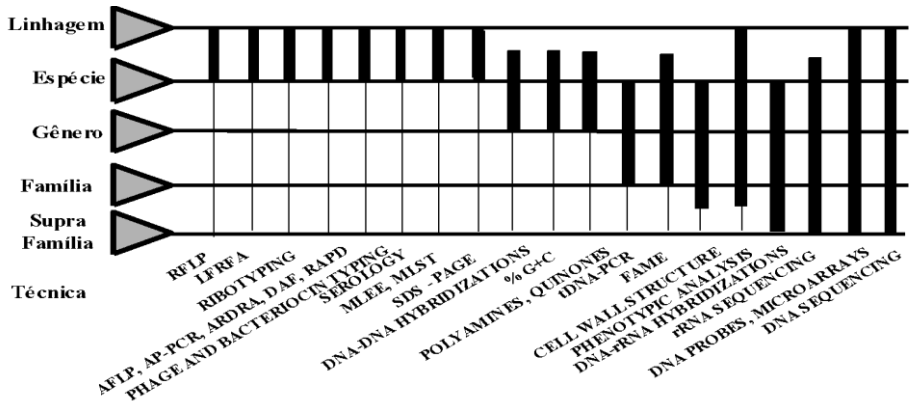
2.6.1 Princípios, estratégias e técnicas

A taxonomia polifásica é um consenso entre sistematas (VANDAMME et al., 1996). A mesma integra dados fenotípicos, quimiotaxômicos, moleculares e genômicos visando representar a biodiversidade em seus diversos níveis (Figura 7). Na década passada, diversas técnicas genômicas baseadas em padrões de banda ou códigos de barra (fingerprints) foram aplicadas (DIJKSHOORN et al., 2005).Vários estudos mostraram uma alta correlação entre a similaridade de padrões de AFLP e de hibridização DNA-DNA para diversos grupos taxonômicos modelo, incluindo *Burkholderia* (COENYE et al., 2000). Por este motivo, o AFLP foi sugerido como uma alternativa para as hibridizações de DNA (STACKEBRANDT et al., 2002; THOMPSON et al., 2002).

O fato de a técnica de AFLP ser rápida, altamente discriminatória, e dos resultados poderem ser acumulados em bases de dados locais, não torna a comparação de padrões de AFLP, gerados em diferentes laboratórios,menos difícil, comprometendo tremendamente a criação de bancos de dados públicos para a identificação de procariotos.

Figura 6 - Discriminação de diversas técnicas empregadas na taxonomia polifásica

Fonte:



Adaptação de Vandamme et al., 1996

A não-portabilidade de dados fenotípicos, como por exemplo, perfis de ácidos graxos, de proteínas e moleculares, por exemplo, AFLP, resulta na concentração do conhecimento taxonômico sobre diferentes grupos de procariotos em poucos laboratórios internacionais de referência. Esta tendência leva a uma demora na categorização dos grupos, uma vez que diferentes taxonomistas utilizam diferentes ferramentas para estudarem os mesmos grupos taxonômicos. Além disso, o emprego destas técnicas requer a inclusão de linhagens de referência em cada novo estudo, encarecendo as técnicas.

O uso de Multi Locus Sequence Typing (MLST) tem ampliado a visão sobre a biodiversidade, bem como da evolução de bactérias (COHAN, 2002). A metodologia contemporânea consiste no sequenciamento e análise de fragmentos de genes conservados (essenciais na manutenção do ciclo celular) espaçados ao longo do genoma bacteriano (MAIDEN et al., 1998). A principal vantagem desta técnica é que a diferença entre linhagens é indexada diretamente nas sequências de DNA. O fato de estes genes evoluírem lentamente os torna ideais para estudos de longo termo em epidemiologia e identificação.

Utilizando dados de MLST pode-se calcular a contribuição de mutação e recombinação na evolução de complexos clonais dentro de uma dada espécie de bactérias (FEIL et al., 2003). Para elaboração de vacinas mais eficientes para microrganismos patogênicos ou para tomada de medidas epidemiológicas, podem ser auxiliadas por informações dessa natureza. Este tipo de metodologia abre uma

nova possibilidade para integrar o conhecimento sobre a biodiversidade das diferentes regiões brasileiras.

2.6.2 Análise de ácidos graxos

A relativa simplicidade, associada ao alto grau de automaçãoe baixos custos para análise de ácidos graxos de proteínas totais vem tornando-a uma técnica valiosa para a identificação rápida a nível laboratorial. No entanto, Vandamme et al. (1996) relataram a falha na análise de ácidos graxos de proteínas totais para distinguir entre as cinco primeiras espécies conhecidas do complexo *B. cepacia*.E dados mais recentes confirmaram esta conclusão. Foi também demonstrado que com a análise de ácidos graxos não é possível diferenciar os membros do complexo *B. cepacia* (WILSHER et al., 1999).

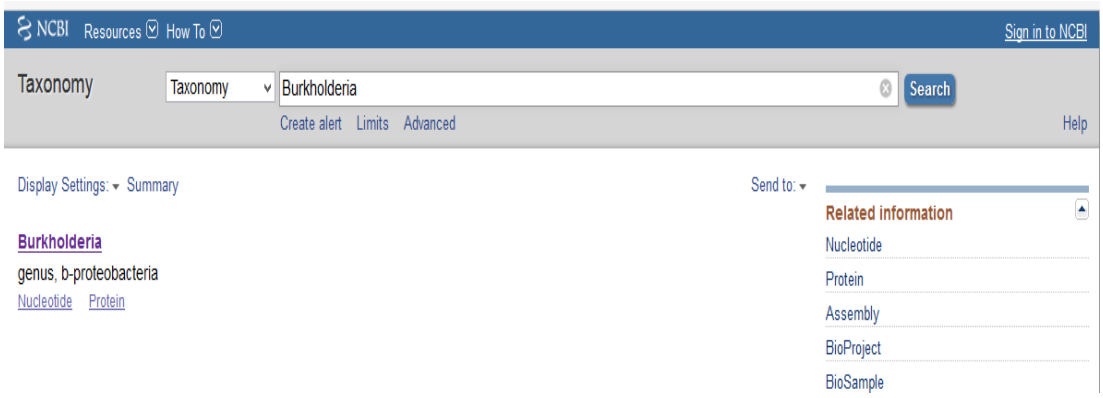
A técnica apresenta como principal vantagem a existência de uma base de dados comercial para a identificação dos isolados que permite a rápida separação de organismos do complexo *B. cepacia* e organismos relacionados tanto de outros gram-negativos não fermentadores (como *P. aeruginosa* e *S. maltophilia*) e de *Enterobacteriaceae*. A técnica também pode ser usada para atribuir isolados que não podem ser classificados por outros métodos de triagem para uma grande linhagem filogenética.

3 MATERIAIS E MÉTODOS

3.1 Retiradas de sequências do NCBI

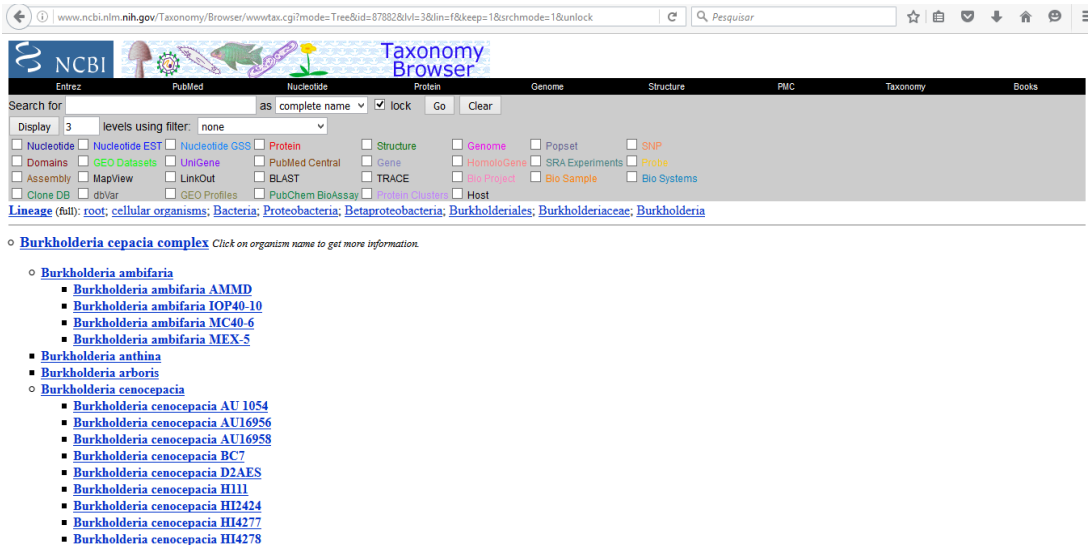
As sequências analisadas foram obtidas através do banco de dados GenBank do NCBI, através da página inicial do NCBI no endereço www.ncbi.nlm.nih.gov, como demostra asFiguras 7 e 8.

Figura 7 -Seleção de sequências para alinhamento



Fonte: http://www.ncbi.nlm.nih.gov/taxonomy/?term=Burkholderia (2016)

Figura 8 -Seleção de sequências do complexo *B. cepacia* para alinhamento



Fonte: http://www.ncbi.nlm.nih.gov/taxonomy/?term=Burkholderia (2016)

3.2 Seleção de sequências

Os dados foram organizados de modo a ficarem uniformes, tais como retirada de sequências que não correspondiam ao gene selecionado (ex.: 23S rDNA); retirada de sequências “sp.” (linhagens cujas espécies não haviam sido definidas); e retirada de sequências como menos de 700 pb.

3.3 Análise das sequências

As sequências genômicas utilizadas no presente estudo foram fornecidas em um formato de arquivo baseado em texto que representa as sequências de nucleotídeos utilizando códigos na forma de letras (A, G, C, T). Esse formato citado é denominado FASTA e sua estrutura permite a manipulação e análise das sequências através de ferramentas de bioinformática. A sequência no respectivo formato começa com uma identificação singular seguida por linhas de dados da sequência de DNA. Em cada linha de identificação possui o símbolo “>” na primeira coluna, e a sequência de dados começa na próxima linha, como demonstra a Figura 9.

Figura 9 - Formato de arquivo FASTA

```
>gi|189503756|gb|EU734821.1|:59-1476 Burkholderia cepacia strain NBRAJG97 16S ribosomal
RNA gene, partial sequence
GTCTGGGAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCGCAAGACCAAAGTGGGGGACCT
TCGGGCCTCATGCCATCAGATGTGCCCAGATGGGATTAGCTAGTAGGTGGGGTAACGGCTCACCTAGGCGACGATCCCTAGCTGGTC
TGAGAGGATGACCAGCCACACTGGAACTGAGACACGGTCCAGACTTCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCA
AGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTTGTAAAGTACTTTCAGCGGGGGAGGAAGGCGATAAGGTTAATA
ACCTTGTCGATTGACGTTACCCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAAT
CGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTCTGTCAAGTCGGATGTGAAATCCCCGGGCTCCAACCTGGGAACTGCATTTCGA
AACTGGCAGGCTAGAGTCTTGTAGAGGGGGGTAGTAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGC
GAAGGCGGCCCCCCTGGACAAAGACTGACGCTCAGGTGCGAAAGCGTGGGGGAGCAAACAAGGATTAGATACCCTGGTAGTCCACGC
CGTAAACGATGTCGATTTGGAGGTTGTGCCCTTGAGGCGTGGCTTCCGGAGCTAACGCGTTAAATCGACCGCCTGGGGAGTACGGCC
GCAAGGTTAAAACTCAA
>gi|361073087|gb|JN903379.1|:1-1407 Burkholderia cepacia strain N8 16S ribosomal RNA
gene, partial sequence
GCCTAGGAATCTGCCTGGTAGTGGGGGACAACGTTTCGAAAGGAACGCTAATACCGCATACGTCCTACGGGAGAAAGCAGGGGACCT
TCGGGCCTTGCGCTATCAGATGAGCCTAGGTCGGATTAGCTAGTTGGTGAGGTAATGGCTCACCAAGGCGACGATCCGTAACTGGTC
TGAGAGGATGATCAGTCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAA
AGCCTGATCCAGCCATGCCGCGTGTGTGAAGAAGGTCTTCGGATTGTAAAGCACTTTAAGTTGGGAGGAAGGGCAGTAAATTAATAC
TTTGCTGTTTTGACGTTACCGACAGAATAAGCACCGGCTAACTCTGTGCCAGCAGCCGCGGTAATACAGAGGGTGCAAGCGTTAATC
GGAATTACTGGGCGTAAAGCGCGCGTAGGTGGTTCGTTAAGTTGGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATCCAAAAC
TGGCGAGCTAGAGTATGGTAGAGGGTGGTGGAATTTCCTGTGTAGCGGTGAAATGCGTAGATATAGGAAGGAACACCAGTGGCGAAG
GCGACCACCTGGACTGATACTGACACTGAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAAC
GATGTCAACTAGCCGTTGGGAGCCTTGAGCTCTTAGTGGCGCAGCTAACGCATTAAGTTGACCGCCTGGGGAGTACGGCCGCAAGGT
TAAAACTCAA
```

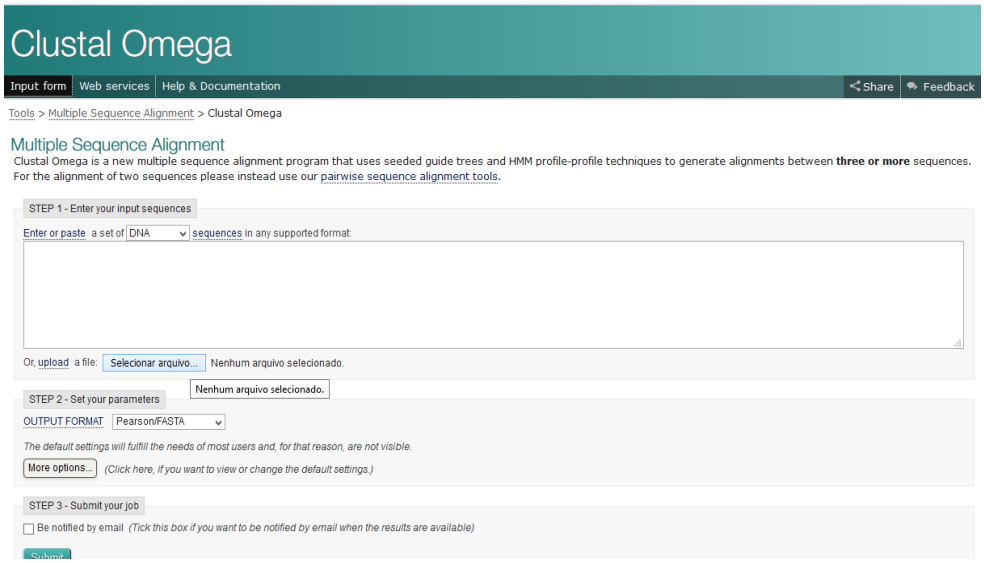
Fonte: NCBI (2016)

3.4 Alinhamento prévio da ssequências utilizando o clustal ômega e observando através do BioEdit

O alinhamento múltiplo realizado pelo CLUSTAL (THOMPSON et al., 1994) tem por objetivo inferir a similaridade entre os nucleotídeos que constituem os genes das bactérias do gênero *Burkholderia*, objeto deste estudo. Este processo é importante para a identificação das regiões alinhadas entre várias espécies. A execução do MLS foi realizada através do *Web services* desenvolvidos pelo Instituto Europeu de Bioinformática (EMBL-EBI) e disponíveis em http://www.ebi.ac.uk/Tools/msa/clustalo/, como demonstra a Figura 10. Para alinhamento utilizamos os seguintes passos:

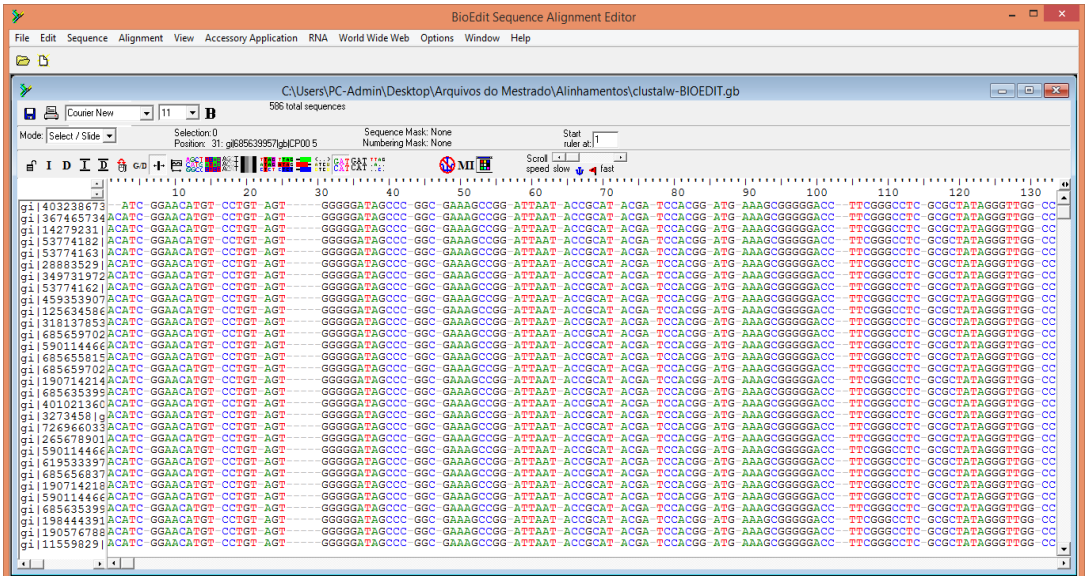
1-Selecionar as sequências;
2- Copiar e colar no clustal ômega;
3- Selecionar a opção “DNA”;
4- Output/format Person/FASTA;
5- Submeter;
6- Após o fim do processamento clicar em “dowloand”;
7- Após o arquivo aparecer na tela, pedir para salvar no formato FASTA que posteriormente será utilizado no Bioedit;
8- Abrir no BioEdit, clicar em “file” e depois “open”;
9- Visualizar as sequências alinhadas no Bioedit, como demostra a Figura 11.

Figura 10 - Execução do MLS realizada através do Web servisse



Fonte: http://www.ebi.ac.uk/Tools/msa/clustalo/(2016)

Figura 11 - Visualização do alinhamento no BioEdit



Fonte: Elaborado pela autora (2016)

3.5 Representação do alinhamento de sequências

Representamos uma sequência de DNA em formato texto, de modo que cada base pode ser exibida por um caractere. Dessa forma, podemos obter as seguintes interpretações: A (Adenina), C (Citosina), G (Guanina) e T (timina). O código oficial para essa representação de DNA é mantido pela IUPAC e inclui também códigos para identificar bases ambíguas, ou seja, aqueles casos em que não se sabe ao certo a base correta, mas se sabe que deve ser um C ou T, ou algo similar, como demonstra a Tabela 1.

Tabela 1 - Código IUPAC utilizado para representar o DNA

A	Adenina
C	Citosina
G	Guanina
T (ou U)	Timina (ou Uracila)
R	A ou G
Y	C ou T
S	G ou C
W	A ou T
K	G ou T
M	A ou C
B	C ou G ou T
D	A ou G ou T
H	A ou C ou T
V	A ou C ou G
N	qualquer base
. ou -	gap

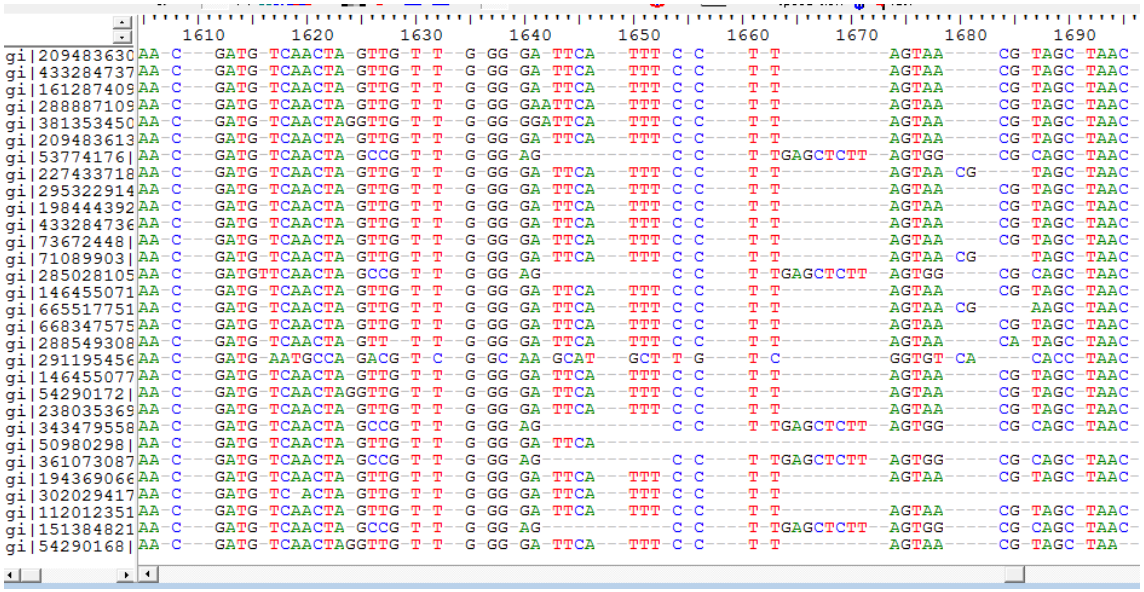
Fonte: IUPAC (1982)

BioEdit é um editor de alinhamento de sequências biológicas. Através dessa ferramenta, podemos melhor visualizar o alinhamento das sequências. Uma interface intuitiva com múltiplos recursos faz o alinhamento e a manipulação de sequências de forma prática e fácil no computador. Várias sequências de manipulação facilitam um ambiente de trabalho que permite análise e manipulação de sequências.

3.6 Eliminação de sequências atípicas

Após o alinhamento, foi observado que a maior parte das sequências formaram um “bloco”. As sequências que não alinharam nesse grupo(Figura 12) foram avaliadas individualmente.As sequências que não alinharam foram reanalisadas utilizando-se a sequência inverso complementar.As sequências que, ainda assim, não alinharam ao maior grupo foram, dessa forma, eliminadas.

Figura 12 - Blocos considerados fora do alinhamento



Fonte: Autora (2016)

Após a retirada das sequências que não foram alinhadas, o maior grupo foi uniformizado (editado) de modo a manterem uma média de 700 pb.Dessa forma, retirou-se os “resíduos” do início e fim de cada sequência.Após o processo de retirada desses “resíduos” (Figura 13) as sequências foram alinhadas pelas seguintes ferramentas: kalign e clustal ômega.Em seguida, foi gerado um cladograma de todas as espécies previamente alinhadas

Figura 13 - Regiões de bordas que foram retiradas

Fonte: Autora (2016)

4 RESULTADOS E DISCUSSÃO

As sequências genéticas (nucleotídeos) de organismos do complexo *B. capacia* foram primeiramente alinhadas utilizando programas (softwares) específicos para alinhamento genético. Após o alinhamento, os respectivos cladogramas foram comparados de acordo com os agrupamentos formados. Foram obtidas 720 sequências pelo NCBI, das quais foram selecionadas 586. Após o alinhamento e edição, obteve-se sequências de, em média, 700pb. As sequências genômicas utilizadas no presente estudo foram fornecidas em um formato de arquivo baseado em texto, que representa as sequências de nucleotídeos utilizando códigos na forma de letras (A, G, C, T). Esse formato citado é denominado FASTA e sua estrutura permite a manipulação e análise das sequências através de ferramentas de bioinformática.

4.1 Quantidade de sequências por espécies

As sequências separadas em grupos de acordo com cada espécie foram distribuídas de modo a serem avaliadas separadamente. A seguir, a Tabela 2 ilustra a quantidade de sequências de cada espécie utilizada na análise.Foram escolhidos grupos com posição incerta ou duvidosa em sistemas de classificação/identificação. Algumas precauções na escolha do grupo devem ser observadas: os táxons devem estar bem circunscritos e delimitados em relação a outros,e o grupo deve ser abrangente o suficiente para conter todas as relações mais próximas. A seleção inicial do grupo deve ser questionada para evitar a tendência de seguir o sistema de classificação passado.

Tabela 2 - Grupos de espécies

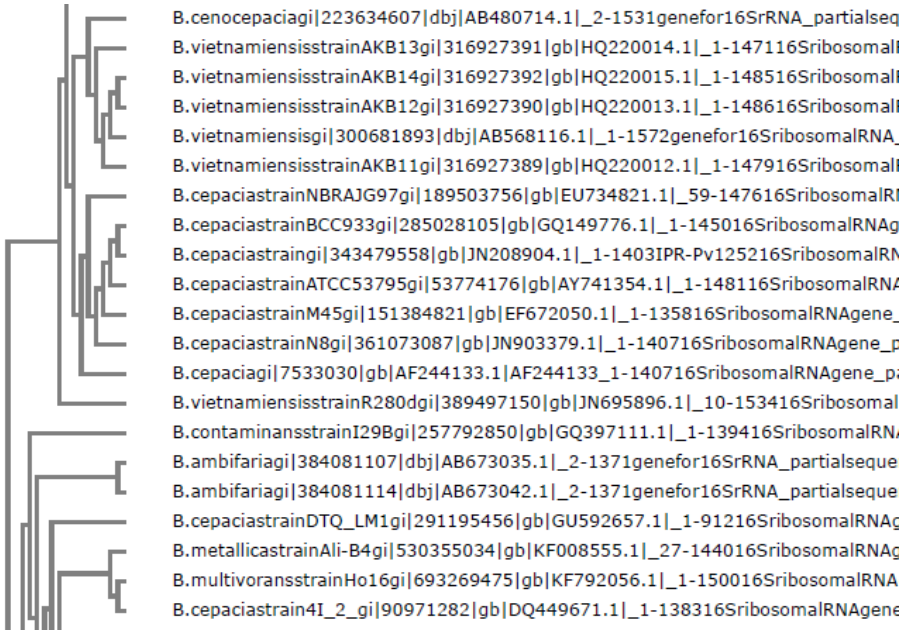
ESPÉCIES	QUANTIDADE DE SEQUÊNCIAS	DESCRIÇÃO DA ESPÉCIE
B. cepacia	243	Importante em pacientes acometidos pela fibrose cística, além de causar infecções pulmonares em pacientes com doença granulomatosa crônica (
B. cenocepacia	85	Sua infecção é particularmente problemática, uma vez que este organismo tem altos níveis de resistência aos antibióticos, tornando difícil de erradicar. As infecções crônicas resultantes estão associadas a declínios severos na função pulmonar e taxas de mortalidade aumentadas (HOLDEN et al., 2008).
B. ambifaria	46	Podem ser utilizadas para fins de controlo biológico, todavia, causam infecções em seres humanos. Pode ser diferenciada dos outros membros do complexo de B. cepacia por meio de AFLP fingerprinting, análise de ácidos graxos, testes bioquímicos e um novo Ensaio de PCR baseado no gene recA desenvolvido (COENYE et al., 2001)
B. vietnamiensis	54	
B. contaminans	2	Seu nome refere-se ao fato de ser considerada contaminante, poluente, referindo-se aometagenoma que foi recuperado do mar de Sargazos, mas que provavelmente representou umcontaminante da amostra. Isolados do CBC cresceu muito mal na água do mar, sugerindo que o oceano aberto não é um habitat natural de espécies do complexo. São bacilos gram-negativos não esporulados(MAHENTHIRALINGAM et al., 2006).
B. lata	5	Espécie do CBC bastante comum no mundo, representadas por células gram-negativas, aeróbicas, não esporuladas. Todas as linhagens conhecidas crescem em ágar MacConkey. O crescimento é observado em 30 à 37 ºC, mas não a 42 ºC (exceto R-18628). Algumas estirpespigmentadas são amarelas ou amarelo-roxo (VANLAERE et al., 2009).
B. multivorans	48	O nome *B. multivorans*foi proposto para uma destas espécies genômicas, que anteriormente era referida como *B. cepacia genomovar*. O crescimento é observado a 37 ° C e 42 ° C; alguns estirpes crescem à temperatura ambiente. Até o momento, nenhuma cepa pigmentada foi detectada. Apresenta crescimento em agar MacConkey e ágar citrato Simmons (VANDAMME et al., 1997).
B. pyrrocinia	23	Baseado na avaliação de hibridização DNA-DNA e sequenciamento 16S rDNA, B. pyrrocinia classifica-se como uma bactéria de solo. Foi descrita na década de 1960 e revelou altos níveis de similaridade para bactérias do CBC (VANDAMME et al., 2002).
B. seminalis	10	Relativa à semente, referente à superfície da semente de arroz, a partir da qual foram isoladas várias estirpes. São bactérias gram-negativas, aeróbicas, não-esporuladas. A maioria das cepas são pigmentadas em amarelo. Não foi observada hemólise (VANLEARE et al., 2009)
B. anthina	12	A análise do recA por PCR fornece um meio simples para identificar este organismo.Usando rRNA baseado em ensaios de PCR, estirpes de *B. anthina* poderiam ser distinguidas*de B. multivorans, B.*

		vietnamiensis, mas não a partir dos outros membros do complexo (VANDAMME et al., 2002).
B. dolosa	7	Bactérias gram-negativas, pequenas, móveis, em forma de vara. As características bioquímicas são descritas por Coenye et al. (2001a). O meio seletivo é incapaz de utilizar ácido azelaico, triptamina ou salicina. Comparado com outras bactérias do CBC, essas estirpes geram 16S RFLP e RFLP recA e podem ser identificados utilizando um ensaio de PCR baseado em rDNA 16S específico. Isolados foram obtidos a partir do ambiente e do escarro de pacientes com fibrose cística (VERMIS et al., 2004).
B. metallica	11	Sua nomenclatura deve-se ao fato de suas colônias apresentarem um brilho metálico. As estirpes não são hemolíticas ea maioria apresenta coloraçãoamarela e crescem em agar MacConkey (VANLAERE et al., 2009).
B. difusa	7	Estirpes conhecidas crescem em ágar MacConkey. Algumas cepas podem crescer na BCSA,Meio alcalino ou ácido. O crescimento é observado em 30 e 37 ºC. Não foram detectadas cepas pigmentadas e hemólise também não foi observada. A estirpe do tipo AU1075T foi recuperada do escarro de umpaciente com fibrose nos EUA em 1999 (VANLAERE et al., 2009).
B. latens	1	Seu nome deve-se ao estudo taxonômico ter demorado um certo tempo (latente). São bactérias gram-negativas, aeróbias, não esporuladas que apresentam aspecto mucóide, Estirpes conhecidas crescem em ágar MacConkey. As estirpes crescem O crescimento é observado em 30, 37 e 42 ºC; Alguns isolados produzem uma difusão, semelhante à melanina, pigmento castanho a 37 e 42 ºC em agar Luria-Bertani. Não foram detectadas estirpes pigmentadas ou hemólise (VANLEARE et al., 2009).
B. arboris	8	Recebeu esse nome devido ao fato de ter sido isolada de uma floresta, na Filadélfia. Estirpes conhecidas crescem em ágar MacConkey. As estirpes podem crescer em BCSA e transformar o meio alcalino em ácido. O crescimento é observado à 30 e 37 ºC; Apenas algumas estirpes são capazes de crescer a 42 ºC. Estirpes R-13059 e R-20536 produzem um pigmento roxo após dias de cultivo. Seis das treze estirpes conhecidas mostram b-hemólise, uma característica não comumente observada entre as espécies do CBC (VANDAMME et al., 2002).
B.pseudomultivorans	2	*Pseudomultivorans*, o falso (*Burkholderia*) multivorans, referindo-se ao fato de que os isolados desta espécie são muito semelhantes a isolados de*B. multivorans* (PEETERS et al., 2013).

Fonte: Autora (2016)

A Figura 14 ilustra o dendograma gerado pelo Kalign das sequências submetidas à análise onde demonstra que o mesmo também não consegue agrupar as mesmas espécies no mesmo ramo e, dessa forma, identifica as mesmas espécies como sendo espécies diferentes.

Figura 14 -Dendograma gerado pelo kaling demonstrando que o mesmo não agrupa as mesmas espécies no mesmo ramo



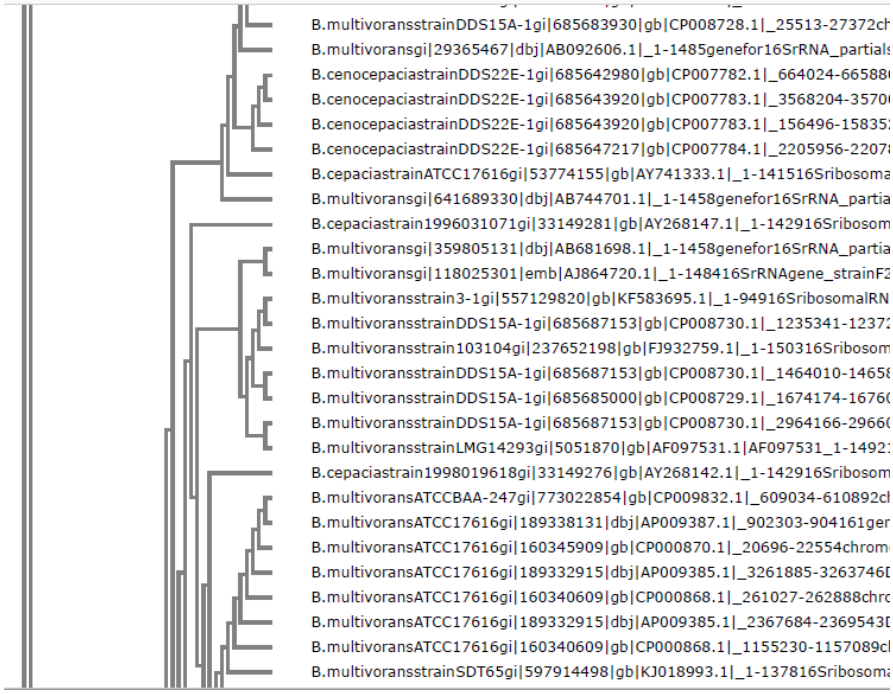
Fonte: Autora (2016)

Desde o estabelecimento dos princípios fundamentais da teoria da evolução por Darwin, um dos maiores objetivos das ciências biológicas é a determinação da história de vida de seus descendentes (RADFORD, 1986) e um cladograma pode ser utilizado como base para um sistema de classificação (NELSON& PLATINK, 1981),como demonstra a Figura 15 onde há um grupo contendo oito espécies de *B. ambifaria*. Os cladogramas gerados pelo Kaling não reproduziram agrupamentos de espécies previamente identificadas como sendo oriundas de um ancestral comum. A Figura 16 ilustra um grupo de *B. multivorans* contendo espécies onde os agrupamentos também não foram considerados relevantes do ponto de vista de identificação à nível taxonômico.

Figura 15 **-** Dendograma gerado pelo kaling demonstrando espécies em diferentes ramos

Fonte: Autora (2016)

Figura 16 **-** Dendograma gerado pelo kaling demonstrando espécies *B. multivorans* em diferentes ramos



Fonte: Autora (2016)

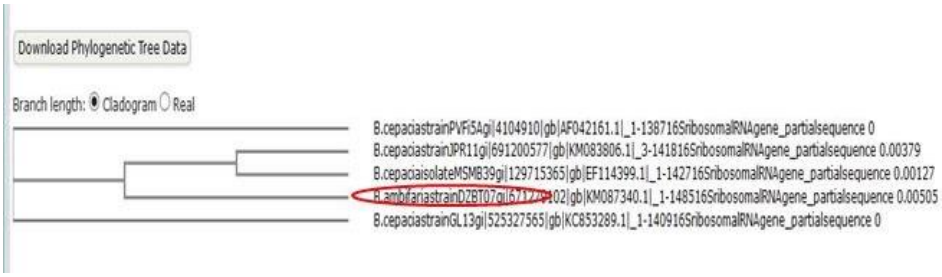
O clustal ômega, uma das ferramentas utilizadas para alinhamento global múltiplo de sequências e geração de cladograma também não conseguiu gerar dados de boa acurácia, como podemos visualizar na Figura 17.

Figura 17 **-** Dendograma gerado pelo Clustal Ômega demonstrando que o mesmo também não consegue agrupar espécies

Fonte: Autora (2016)

O dendograma da Figura 18 demonstra que, ao analisar grupos da mesma espécie (*B. cepacia*) e inserir uma espécie distinta (*B. ambifaria*), a distinta se agrupa à uma das espécies de *B. cepacia*.

Figura 18 **-** Dendograma gerado pelo Kalign demonstrando que o mesmo não agrupa espécies ao se inserir uma única espécie distinta num grupo de espécies

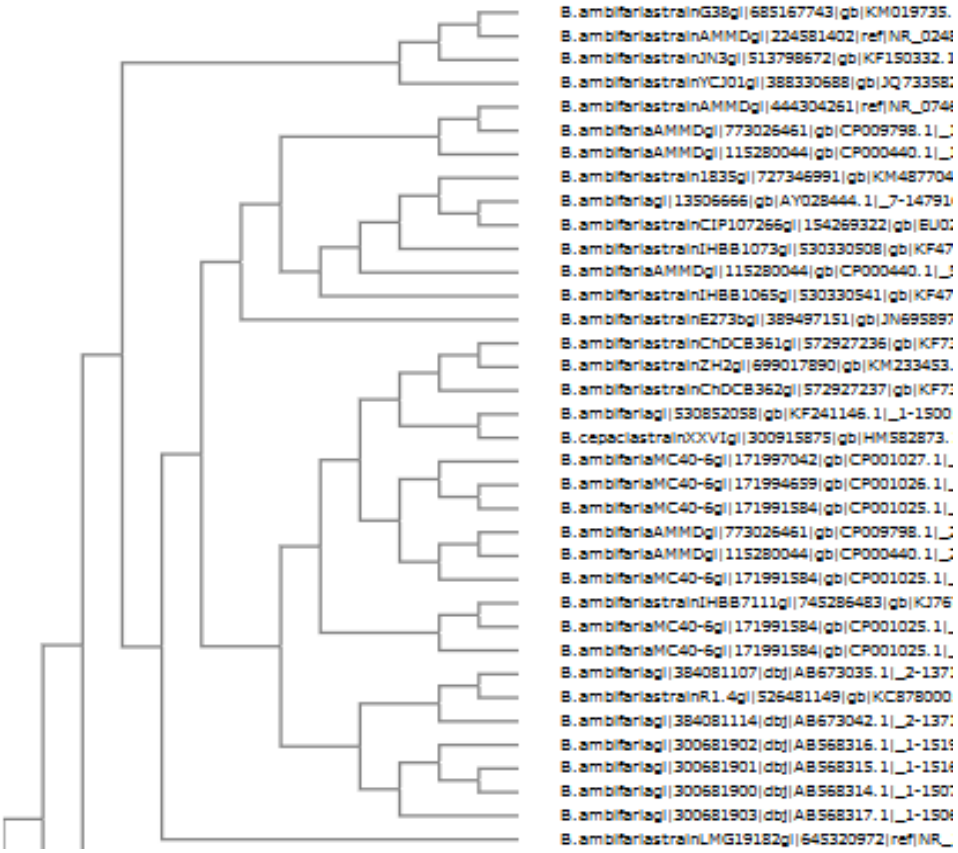


Fonte: Autora (2016)

A seguir, a ilustração da Figura 19 demonstra que, ao colocarmos um grupo contendo espécies de *B. ambifaria* com apenas uma espécie de *B. cepacia*, os

dados também são conflitantes, informando que a ferramenta (no que diz respeito ao *Complexo B. cepacia*) não resolveria o problema de identificação à nível taxonômico apesar de, atualmente, ser utilizado na identificação de alguns microrganismos, em particular, bactérias, de acordo com a literatura.

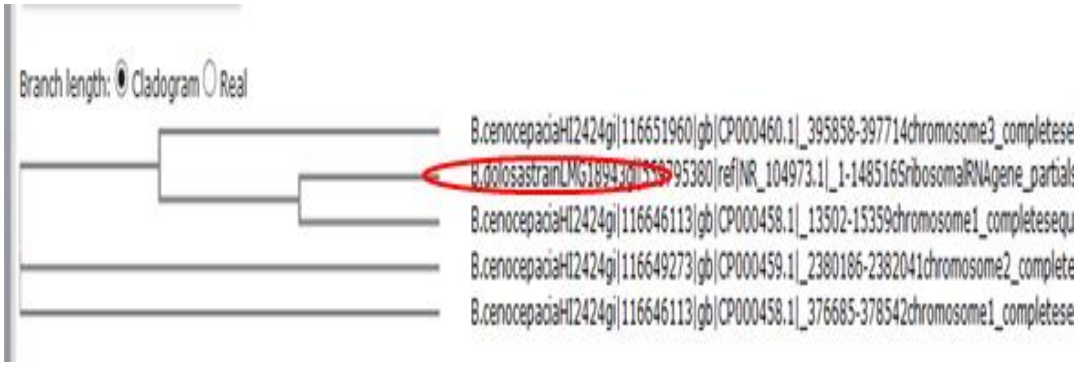
Figura 19 **-** Dendograma gerado pelo Kalign de um grupo de *B. ambifaria* e apenas uma *B. cepacia,* onde é demonstrando que a espécie distinta se agrupa ao maior grupo



Fonte: Autora (2016)

Quando uma espécie *B. dolosa* foi acrescentada a um grupo contendo quatro *B. cenocepacia* e submetidas à análise, o mesmo fato se repetiu, como podemos identificar na Figura 20.

Figura 20 **-** Dendograma gerado pelo Clustal Ômega de um grupo de *B. cenocepacia* e apenas uma *B. dolosa,* onde é demonstrado que a espécie distinta se agrupa ao maior grupo



Fonte: Autora (2016)

Outra alternativa à submissão de sequências foi a avaliação do grupo contendo uma espécie de cada sequência (17 espécies) com adição de uma espécie:*B. pyrrocinia*. Desse modo, teríamos apenas a espécie *B. pyrrocinia* se repetindo, visando gerar um cladograma onde essas duas sequências repetidas ficassem o mais aproximadas possível, fato que não foi identificado, como podemos avaliar de acordo com a Figura 21.

Figura 21 **-** Dendograma gerado pelo Kalign de um grupo de cada uma das 17 espécies e uma espécie *repetida (B. pyrrocinia)* demonstrando que a "espécie que se repete" não se agrupa no mesmo ramo

Fonte: Autora (2016)

Identificação bacteriana do *complexo B. cepacia* representa um problema de ordem complexa a ser resolvido por sistemas computacionais, principalmente se tratando de amostras com grande quantidade de sequências, como exposto no referido trabalho, pois o número de árvores possíveis é gigantesco. Para resolver esse problema, pesquisadores têm aplicado métodos computacionais especiais que exploram diversas possibilidades de maneira mais eficiente, de modo a se chegar a uma solução mais aproximada a real. Todavia, esses *softwares*, ao invés de construir todas as árvores possíveis para posteriormente decidir qual delas é a melhor, procuram encontrar padrões lógicos de maneira heurística, visando desenvolver estratégias de exploração que se concentra a maior parte da pesquisa baseada na aproximação para o referido problema.Com a tentativa de analisar grupos de mesma espécie e inserção de espécies distintas, gerou-se outros cladogramas como demonstra a Figura 22.

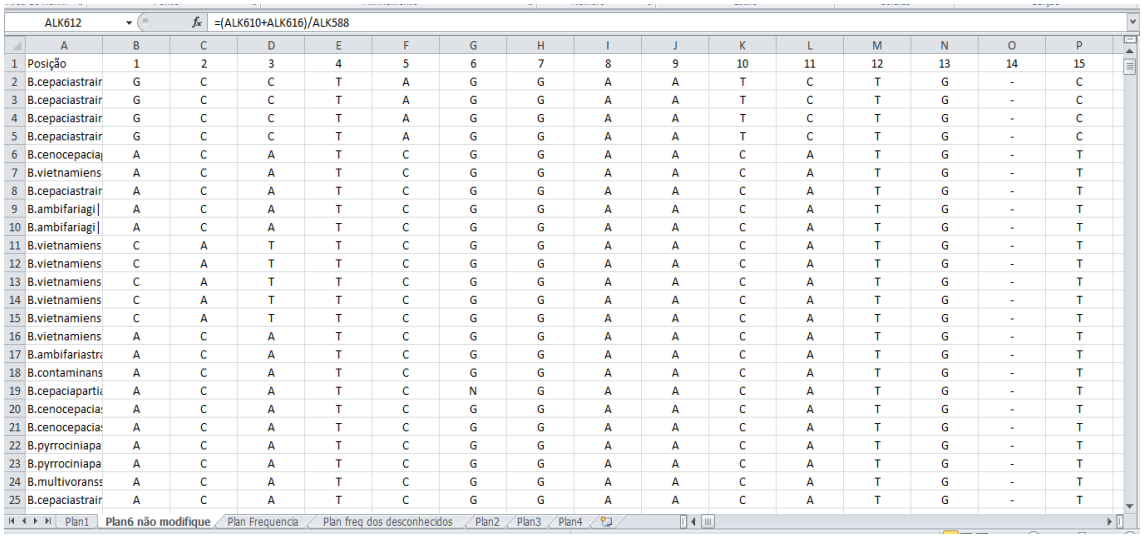
Figura 22 **-** Dendograma gerado pelo Kalign de um grupo de dois grupos de *B. cepacia* onde, no primeiro, acrescentou-se uma B. difusa e, no segundo, uma *B.arboris.* Em ambos os grupos não houve agrupamentos de espécies.

Fonte: Autora (2016)

4.2 Análise de Componentes Principais

Foi criada uma tabela de frequência do Excel contendo 586 linhas (total de sequências) e 1.080 colunas (média de bases por sequência) com o intuito de identificar a frequência de cada variável como demonstra a Figura 23. Foram definidas 7 variáveis: A, T, G, C, N e outros.

Figura 23 - Planilha de frequência



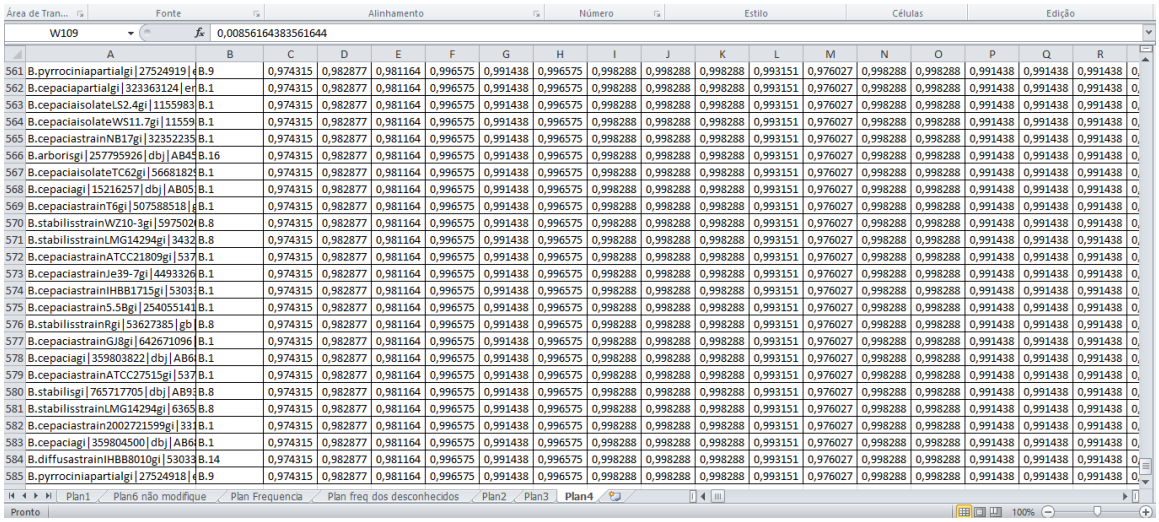
Fonte: Autora (2016)

4.3 Substituição da planilha de variáveis por frequências numéricas

Após a transferência das variáveis alinhadas do BioEdit para o Excel fizemos a substituição por frequências numéricas (Figura 23) com o intuito de identificar as frequências mais similares quando comparadas a cada grupo para posteriormente submeter o material à análise pelo *Statistical Package for Social Sciencefor Windows* (SPSS).O SPSS é um *software* utilizado para análise estatística de matrizes de dados, em um ambiente amigável. Para isso, utiliza-se menus e janelas de diálogo, que permite realizar cálculos complexos além de visualizar resultados de forma simples e autoexplicativa. A ferramenta é capaz de transformar dados em informações importantes, capazes de reduzircustos aumentando dessa forma a lucratividade, gerando gráficos de dispersões e

distribuições que podem ser usados em análises de correlação entre variáveis. A primeira versão data de 1968 e, a mais recente é a SPSS for Windows 7 (2014). O aspecto inicial do editor é apresentado na figura a seguir (Figura 24). Podemos encontrar: o Data View (Data Editor), em que as colunas são as variáveis e as linhas os casos (ou indivíduos). As células podem conter valores numéricos ou alfanuméricos, mas não podem conter fórmulas. O intuito da geração desse banco de dados seria avaliar, através da Análise de Componentes Principais (ACP) se as mesmas espécies formariam grupos distintos ou similares e, estudar os grupos que por ventura ficassem “fora dos padrões estabelecidos” de maneira particular.

Figura 24 **-** Planilha de frequências numéricas



Fonte: Autora (2016)

Após a substituição das variáveis pelas frequências, observou-se que 266 colunas apresentavam “variáveis que não variavam” e, dessa forma, foram retiradas, visto que o interesse seria realizar uma análise de componentes principais visando identificar as colunas mais discrepantes. Todavia, para isso, haveria necessidade em se trabalhar com sequências divergentes.

5 CONCLUSÕES

Ao término desse trabalho podemos concluir que as ferramentas computacionais utilizadas (Clustal Ômega e Kalign) na referida abordagem não foram suficientes para identificação bacteriana do *Complexo B. cepacia* bem como a região estudada (16S). Todavia, houve a possibilidade em se fazer alinhamento múltiplo das sequências estudadas, além da comparação de diferentes ferramentas computacionais na avaliação do mesmo. Além do exposto, identificou-se a necessidade em definir critérios de edição das sequências estudadas visto que as mesmas não foram aplicadas da forma como foram retiradas da base de dados do NCBI aliada à necessidade de desenvolvimento de novas ferramentas capazes de solucionar o problema de identificação bacteriana do complexo *B. cepacia*, visando avaliar a aplicabilidade dos AGs no estudo taxonômico.

5.1 Contribuições do trabalho

Esse trabalho contribuiu para verificar o alcance das ferramentas que utilizam a região 16S para a identificação taxonômica de espécies do *Complexo B. cepacia* demonstrando a necessidade de desenvolvimento de novas ferramentas que sejam capazes de trabalhar com uma grande quantidade de sequências bem como fazer a distinção entre as referidas espécies do complexo em questão visando à melhoria da identificação no que se refere ao tempo e confiabilidade de resultados.

5.2 Dificuldades encontradas

Para a realização do referido trabalho, foram encontradas as seguintes dificuldades: seleção de sequências a serem trabalhadas visto que cada sequência apresenta suas particularidades tais como tamanho, bases indefinidas, GAPs, entre outros; edição das sequências selecionadas para estudo visando trabalhar com informações mais uniformes para eliminar eventuais interferentes; eliminação de sequências consideradas atípicas

definindo quais seriam as mesmas; encontrar *softwares* capazes de processar grande quantidade de dados em tempo hábil; agrupar espécies bem como separar as distintas através da produção de árvores filogenéticas para que possamos obter dados confiáves e fidedignos.

5.3 Trabalhos futuros

A identificação pela região 16S de bactérias do *Complexo Burkholderia cepacia* não demonstrou bons resultados através da utilização de ferramentas computacionais. Uma sugestão seria o desenvolvimento de novas ferramentas capazes de trabalhar com uma grande quantidade de sequências com o intuito de obter uma melhor separação dos grupos de espécies, capazes de identificar diferenças existentes entre os grupos. Outra sugestão na identificação desse **complexo** bacteriano seria a utilização de outras regiões gênicas mais informativas.

REFERÊNCIAS

BENSON D., et al. **Genbank.**Núcleic Acids Research, 2000.

_____. **Nucleic Acids Res**. (Database issue). Jan, 2013. Acesso em: 26 fev. 2016.

BERGEY, D. H. N. B., et al. **Bergey's manual of determinative bacteriology**. **Baltimore: Williams and Wilkins, 1934.**

BILHA E., et al. Algoritmos de alinhamento de sequências moleculares. **Revista de Informática Aplicada** – Imes Universidade, Ano 1, n.1, p. 13-20, 2005.

BOONE, D. R.; CASTENHOLZ, R. W. **Bergey´s Manual of Systematic Bacteriology**. 2 ed., Volume One, Springer-Verlag, USA, 2001.

BURKHOLDER, W. H.**Sour skin, a bacterial rot of onion bulbs.** Phytopathology 40, p.115–117, 1950.

CHAN, E. Y. **Advances in sequencing technology. Mutation Research**, v. 573, p. 13-40, 2005.

COHAN, F. M. **What are bacterial species?.**Annual Reviews in Microbiology, v. 56, n. 1, p. 457-487, 2002.

COLWELL, R. R. **Polyphasic taxonomy of the genus Vibrio:** numerical taxonomy of Vibrio cholerae, Vibrio parahaemolyticus, and related Vibrio species. J Bacteriol104, p. 410-433, 1970a.

_____. **Polyphasic taxonomy of bacteria**. In Culture Collections of Microorganisms. H. Iizuka & T. Hasegawa. (eds.). Tokyo: University of Tokyo Press, 1970b. p. 421-436.

CORMEN, T. H.; LEISERSON, C. E.; RIVEST, R. L. **Introduction to Algorithms**. MIT Press, 1990.

DIJKSHOORN, L., et al. **Prevalence of Acinetobacter.** 2005.

DOOLITLE, R. F. **Molecular Evolution:** Computer Analisys of Protein and Nucleic Acid Sequences. Methods in Enzimology. Academic Press, 1990.

DUTTA, T. K., et al. **Partial replacement of concentrate mixture with *Leucaena leucocephala* leaves in pelleted feed of goats.** Indian J. Anim. P. 820-822, 2002.

FENG, D.-F.; DOOLITTLE, R. F. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. **Journal of Molecular Evolution**, v. 25, n. 4, p. 351-360, 1987. Acessado em 22/03/2016.

FLEISCHMANN, R. D. et al. Whole-genome random sequencing and assembly of Haemophilus influenzae. **Rd. Science**. 28 jul, p. 496-512, 1995.

FUNGARO, M. H. P.; VIEIRA, M. L. C**.** Aplicações da PCR em Ecologia Molecular. In: MELO, I. S., de AZEVEDO, J. L. (Ed.). **Ecologia Microbiana**. Jaguariúna: Embrapa, CNPMA, 1998. p. 205-227.

GILLIS, M. et al. Polyphasic taxonomy in the genus Burkholderia leading to an emended description of the genus and proposition of Burkholderia vietnamiensis sp. nov. for N2-fixing isolates from rice in Vietnam.**International Journal of SystematicBacteriology**, v. 45, n. 2, p. 274-289, 1995.

GOLDBERG, D. E. (1989). **Genetic algorithms in search, optimization, and machine learning**. Addison-Wesley, 1989.

GONDRO, C.; KINGHORN, B. **A simple genetic algorithm for multiple sequence alignment**. Genet. Mol. Res, [S.l.], v.6, n.4, p. 964-982, 2007.

GOODFELLOW, M. **Microbial systematics:** background and uses. In: Applied Microbial Systematics Priest, F.G. & Goodfellow, M. (Eds.). Kluwer Academic Publishers, Dordrecht. 2000. Acesso em 19 jun. 2016.

GOUGH, J. et al. **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure**. 2 nov, p. 903-19, 2001.

GRIMONT, A. et al. Report of the ad hoc committee for there-evaluation of the species definition in bacteriology. **International Journal of Systematic and Evolutionary Microbiology**, p. 1043–1047, 2002.

GUSFIELD, Dan.**Algorithms on Strings, Trees and Sequences:** Computer Science and Computational Biology. Cambridge: Cambridge University Press, 28 mai, 1997.

HIGGINS, D. G.; BLEASBY, A. J.; FUCHS, R. CLUSTAL V: improved software for multiple sequence alignment. Computer applications in the biosciences: **CABIOS** [S.l.], v.8, n.2, p. 189-191, 1992.

HIGGINS, D. G.; SHARP, P. M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. **Gene**, v. 73, n. 1, p. 237-244, 1988. Acesso em 22 mai. 2016.

HOLLAND, J. H. **Adaptation in Natural and Artificial Systems**. MIT Press, 1975.

JORGENSEN, R. A.; CLUSTER, P. D. **Modes and tempos in the evolution of nuclear ribossomal DNA: new characters for evolutionary studies and new markers for genetic and population studies.** Ann. Mo. Bot. Gard., p. 1238-1247, 1989.

KLENK, H. P.; GÖKER, M. En route to a genome-based classification of Archaea and Bacteria?.**Systematic and Applied Microbiology**, v. 33, n. 4, p. 175-182, jun. 2010.

LACERDA, E. G. M.; CARVALHO, A. C. P. L. F. Introdução aos Algoritmos Genéticos. In: SBC'99 - Congresso Nacional da Sociedade Brasileira de Computação 19.,**Anais...**Rio de Janeiro, jul, v. 2, p. 51–126, 1999.

LAGUERRE, G. et al. Classification of rhizobia based on nodC and nifH gene analysis reveals a close phylogenetic relationship among Phaseolus vulgaris symbionts. **Microbiology**, v. 147, n. 4, p. 981-993, 2001.
_____. Rapid identification of rhizobia by restriction fragment length polymorphism analysis of PCR-amplified 16S rRNA genes. **Applied and Environmental Microbiology**, v. 60, n. 1, p. 56-63, 1994.

LAJUDIE, P. et al. Characterization of tropical tree rhizobia and description of Mesorhizobium plurifarium**. International Journal of Systematic Bacteriology**, v. 48, p. 369-382, 1998.

LAPAGE, S. et al. International code of nomenclature of bacteria. **International Journal of Systematic Evolutionary Microbioly**, v. 61, p. 6-7, 2011.

LARKIN, M. A. et al. Clustal W and Clustal X version 2.0. **Bioinformatics**, v. 23, n. 21, p. 2947-2948, 2007. Acesso em 10 mai. 2016.

LEMOS, M.; ARAGÃO, M. V. S. P.; CASANOVA, M. A. **Padrões em Biossequências**. PUC-Rio. Rio de Janeiro, 2003.

LINDSTROM, J. M. et al. Antibody to acetylcholine receptor in myasthenia gravis Prevalence, clinical correlates, and diagnostic value. **Neurology**, v. 51, n. 4, p. 933-933-a, 1998.

MAHENTHIRALINGAM, E.; BALDWIN, A.; DOWSON, C. G.*Burkholderia cepacia* complex bacteria: opportunistic pathogens with important natural biology.**Journal of Applied Microbiology.** 2005. Volume 104, Issue, p. 1539–1551, 6 jun. 2008

MAIDEN, M. C. et al.**Multilocus sequence typing:** a portable approach to the identification of clones within populations of pathogenic microorganisms. 17 mar. 1998.

MALAQUIAS, N. G. L. **Uso dos Algoritmos Genéticos para a Otimização de Rotas de Distribuição**. Dissertação (Mestrado em Ciências) – Pós- Graduação em Engenharia Elétrica, Universidade Federal de Uberlândia, Uberlândia, 2006.

MASULLI, F.; PETERSON, L.; TAGLIAFERRI, R. **Computational Intelligence Methods for Bioinformatics:** 6th international meeting, cibb 2009, Genoa, Italy, october 15-17, 2009, revised selected papers. [S.l.]: Springer, 2010.

MEIDANIS, J.; SETÚBAL, J. C. **Uma Introdução à Biologia Computacional**. Escola de Computação. Recife, 1994.

MESHOUL, S.; LAYEB, A.; BATOUCHE, M.A Quantum Evolutionary Algorithm for Effective Multiple Sequence Alignment. In: BENTO, C.; CARDOSO, A.; DIAS, G. (Eds.). **Progress in Artificial Intelligence**. [S.l.]: Springer Berlin Heidelberg, p. 260-271, 2005.

MEYER, C. et al. Transcription of mutS- and mutL-homologous genes during meiosis in Saccharomyces cerevisiae and identification of a regulatory cis-element for meiotic induction of MSH2.**Mol Genet Genomics,** p. 826-36, 2001.

MIR, L. **Genômica**. São Paulo: Ed. Atheneu, 2004.

MURRAY C. J. et al. **Lancet**. Dez, 2012.

MYERS, E. W.; MILLER, W. Optimal alignments in linear space. **Computer applications in the biosciences**, v. 4, n. 1, p. 11-17, 1988. Disponível em: < http://www.cs.cornell.edu/Courses/cs628/2004fa/secure/papers/Myers_Miller_optimal _alignments_in_linear_space_CABIOS88.pdf>. Acesso em: 19 abr. 2016.

LUSCOMBE, N. M.; GREENBAUM, D.; GERSTEIN, M. **What is Bioinformatics?** A Proposed Definition and Overview of the Field. New Haven, USA: Department of Molecular Biophysics and Biochemistry Yale University, 2001.

NORNAM, R.; PACE, P. H.;BRETT, M. G.**Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity**, v. 180, n. 18, p. 4765–4774, 1998.

OGATA, A. K. O. **Multialinhamento de sequências biológicas utilizando algoritmos genéticos**. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional). São Carlos: Universidade de São Paulo, 2006. Acesso em: 10 abr. 2016.

PACHECO, M. A. C. **Algoritmos Genéticos:** Princípios e Aplicações. Departamento de Engenharia Elétrica, PUC-Rio, Rio de Janeiro, 1999.

PALLERONI, N. J. et al.**Nucleic acid homologies in the genus Pseudomonas**. Int. J. Syst. Bacteriol, 1973.

PETER, H. G. et al. Acid pH tolerance in strains of Rhizobium and Bradyrhizobium, and initial studies on the basis for acid tolerance of Rhizobium tropici. **Journal Canadian Journal of Microbiology**, v. 40, p. 198-207, 1994.

PIAZZA, Gregory; GOLDHABER, S. Z. Acute pulmonary embolism part I: epidemiology and diagnosis. **Circulation**, v. 114, n. 2, p. e28-e32, 2006.

POZO, A. et. al. **Computação Evolutiva**. Grupo de Pesquisa em Computação Evolutiva. Departamento de Informática. Universidade Federal do Paraná. Apostila, 2000. 61p.

RAPPÉ, M. S.; GIOVANNONI, S. J. **Annu Rev Microbiol**. p. 369-94, 2003.

ROSSELLO-MORA, R.; AMANN, R. The species concept for prokaryotes. **FEMS Microbiology Reviews**, v. 25, n. 1, p. 39-67, 2001.

SAIKI, R. K. et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. **Science**. p. 1350-4, 1985.

SANGER, F.; COULSON, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. **Journal of Molecular Biology**, v. 94, n. 3, p. 441–448, 1975. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/1100841>. Acesso em: 16 fev. 2016.

SETUBAL, J. C.; MEIDANIS, J. Introduction to Computational Molecular Biology.. Boston, EUA: PWS Publishing Company, 1997. Acesso em: 28 fev. 2016.

SHARON, I.; BANFIELD, J. F. Genomes from metagenomics. Science, p. 1057–1058, 2013. Acesso em 28 fev. 2016.

SIEVERS, F.; HIGGINS, D. G. Clustal Omega, accurate alignment of very large numbers of sequences. **Methods in Molecular Biology** 1079, p. 105-116, 2014. Acesso em: 10 jun. 2016.

SILVA JR, A. L. V. da. **Uma abordagem de alinhamento múltiplo de sequência utilizando evolução diferencial**. Dissertação (Mestrado em

Engenharia Biomédica) – Programa de Pós-Graduação em Engenharia Biomédica, Universidade Federal de Pernambuco, Recife, 2015.

SILVA, P. E. M. da. **Alinhamento de Sequências Biológicas Utilizando Algoritmo Genético e Processamento Distribuído**. Dissertação (Mestrado em Engenharia Elétrica) – Departamento de Engenharia Elétrica, Universidade Estadual de Londrina, Londrina, 2005.

SKACKEBRANDT, F. et al. **Progressive sequence alignment as a prerequisite to correct phylogenetic trees.**p. 351-360, 1987.

SNEATH, P. H.; SOKAL, R. R. Numerical taxonomy. **Nature**, v.193, p.855-860, 1962.

STACKEBRANDT, E. et al. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. **International Journal of systematic andevolutionary microbiology**, v. 52, n. 3, p. 1043-1047, 2002.

STRALIOTTO, R.; RUMJANEK, N. G. Biodiversidade do rizóbio que nodula o feijoeiro (Phaseolus vulgaris L.) e os principais fatores que afetam a simbiose. **Embrapa Agrobiologia**, 51p., 1999.

TAYLOR, W. R. A flexible method to align large numbers of biological sequences. **Journal of Molecular Evolution**, v. 28, n. 1-2, p. 161-169, 1988. Acesso em: 22 mai. 2016.

THOMAS, P. et al. **Estimating prokaryotic diversity and its limits**. Edited by Robert May. Oxford, United Kingdom: University of Oxford, 2001.

THOMPSON, J. D.; HIGGINS, D. G.; GIBSON, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Research**, v. 22, n. 22, p. 4673-4680, 1994. Acesso em: 23 mai. 2016.

THOMPSON, J. D.; HIGGINS, D. G.; GIBSON, T. J. **CLUSTAL W:** improving the sensitivity of progressive multiple sequence alignment throught sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, [S.l.], v.22, n.22, p. 4673-4680, 1994.

THOMSEN, R.; FOGEL, G.; KRINK, T. **A Clustal alignment improver using evolutionary algorithms**. In: EVOLUTIONARY COMPUTATION, 2002. CEC'02. Proceedings of the 2002 congress on. Anais... [S.l.: s..n.], 2002, v.1, p. 121-126.

TSUTSUI, S.; FUJIMOTO, Y.**Forking genetic algorithms with blocking and shrinking modes**. In: Proceedings INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS, 5, p.206-213, 1993.

VANDAMME, P. et al. Polyphasic taxonomy, a consensus approach to bacterial systematics**. Microbiological reviews**, v. 60, n. 2, p. 407-438, 1996. Acesso em: 18 jun. 2016.

VANLAERE, Elke et al. Burkholderia latens sp. nov., Burkholderia diffusa sp. nov., Burkholderia arboris sp. nov., Burkholderia seminalis sp. nov. and Burkholderia metallica sp. nov., novel species within the Burkholderia cepacia complex. **International Journal of Systematic and Evolutionary Microbiology**, v. 58, n. 7, p. 1580-1590, 2008.

_____. Taxon K, a complex within the Burkholderia cepacia complex, comprises at least two novel species, Burkholderia contaminans sp. nov. and Burkholderia lata sp. nov. **International journal of systematic and evolutionary microbiology**, v. 59, n. 1, p. 102-111, 2009.

VENTER, J. C. GENOMICS: Shotgun Sequencing of the Human Genome. **Science,** v. 280, n. 5369, p. 1540–1542, 5 jun. 1998. Disponível em: <http://www.sciencemag.org/cgi/doi/10.1126/science.280.5369.1540>. Acesso em: 24 fev. 2016.

VIANA, G. V. R.; MOURA, H. A. S. **Algoritmos para Alinhamento de Sequências**. Revista Científica da Faculdade Lourenço Filho, v.7, n.1, p. 67-83, 2010.

WANG, L.;JIANG, T. J. On the complexity of multiple sequence alignment. **Comput Biol**.p .337-48, 1994.

WARD, D. M.; WELLER, R.; BATESON, M. M. **16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature.** p. 63-5, 3 mai. 1990.

WEISS, M. S. et al. Citations in supplementary material. **Acta Cryst**, p. 1269–1270, 2010. Acesso em: 27 fev. 2016.

WILLEMS, Anne. **The taxonomy of rhizobia:** an overviewlant and Soil. p. 3-14. 2006.

WILLIAMS, J. G. K. et al. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. **Nucleic Acids Research**, Oxford, v.18, n.22, p. 6531-6535, 1990.

WILSHER, A.; HODGES, E. **Computer generated control panel for a computer monitor**. U.S.A, 2 Mar. 1999.

WOESE, C. R.; FOX, G. E. **Phylogenetic structure of the prokaryotic domain:** the primary kingdoms. Proc. Natl. Acad. Sci., U.S.A., p. 5088–5090, 1977.

WOESE, C. R.; OLSEN, G. J. **Archaebacterial phylogeny:** perspectives on the urkingdoms. Syst Appl Microbiol.p. 161–177, 1986.

WU, S.; MANBER, U. Fast Text Searching Allowing Errors. **Communications of the ACM,** p. 83-91, 1992.

YABUUCHI, E. et al. Proposal of Burkholderia gen. nov; andtransfer of seven species of the Pseudomonas homology group II to the newgenus, with the type species Burkholderia cepacia. **Microbiol. Immunol**, 1251–1275, nov. 1992.

YAMAOKA-YANO, D. M.; VALARINI, P. J. Métodos de identificação de Bactérias. In: Melo, I.S.; Azevedo, J.L. (ed.). **Ecologia Microbiana**. EMBRAPA - CNPMA, Jaguariúna, p. 369-419, 1998.

YANG, Jian et al. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. **IEEE transactions on pattern analysis and machine intelligence**, v. 26, n. 1, p. 131-137, 2004.

ZANG, C.; WONG, A. K**. A genetic algarithm for multiple molecular sequence alignment**. Computer applications in the biosciences: CABIOS, [S.l], v.13, n.6, p. 565-581, 1997.