UNIVERSIDADE FEDERAL DO PARANÁ

JULIANA HELENA TIBÃES

APLICAÇÃO DA INTELIGÊNCIA ARTIFICIAL NA ANOTAÇÃO AUTOMÁTICA DE GENOMAS BACTERIANOS

CURITIBA

2012

JULIANA HELENA TIBÃES

APLICAÇÃO DA INTELIGÊNCIA ARTIFICIAL NA ANOTAÇÃO AUTOMÁTICA DE GENOMAS BACTERIANOS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de Mestre em Bioinformática.

Orientador: Prof. Dr. Fábio de Oliveira Pedrosa
Co-orientador: Prof. Dr. Roberto Tadeu Raittz

CURITIBA

2012

TERMO DE APROVAÇÃO

JULIANA HELENA TIBÃES

APLICAÇÃO DA INTELIGÊNCIA ARTIFICIAL NA ANOTAÇÃO AUTOMÁTICA DE GENOMAS BACTERIANOS

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador: Prof. Dr. Fábio de Oliveira Pedrosa

Co-orientador: Prof. Dr. Roberto Tadeu Raittz

Prof Dr João Carlos Marques Magalhães
Universidade Federal do Paraná

Prof. Dr. Emanuel Maltempi de Souza
Universidade Federal do Paraná

Curitiba, 16 de fevereiro de 2012

Dedico esse trabalho ao meu irmão e meus pais,
que sempre deram apoio e incentivo aos meus estudos.

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Fábio de Oliveira Pedrosa, por acreditar no projeto, pelas discussões e apoio.

Ao meu co-orientador, Prof. Dr. Roberto Tadeu Raittz, por todo ensinamento, apoio, incentivo e dedicação durante todos esses anos.

Ao prof. Mtr. Dieval pelos ensinamentos, discussões e auxílio durante todo o percurso da minha graduação até hoje.

Ao prof. Dr. Emanuel pelas discussões, correções e auxílios durante o desenvolvimento do meu projeto.

À Prof. Dr. Jeroniza e à prof. Dr. Berenice por acreditarem no meu trabalho, e também pelo apoio.

Aos demais professores do programa, em especial ao prof. Dr. Adriano, ao prof.Dr. Lucas e ao prof. Dr. Neves pelas agradáveis conversas e auxílio em diversos assuntos.

Aos meus amigos e colegas da bioinformática, em especial para Vanely, Sérgio, Lucas, Alysson, Rafaela, Gustavo, Leviston, Paula, Ricardo e Rodrigo pelo companheirismo nos dias alegres e também nos não tão fáceis.

As secretárias do departamento, Suzana e Léa, pela dedicação e incontáveis ajudas.

Aos meus amigos da graduação, Aline, Angélica, Letícia e Vanely por todo apoio durante os dias conturbados e companheirismos nos dias felizes.

Ao meu irmão Rafael, que me atura todo dia e pelas discussões de informática sem fim.

Aos meus tios, Gina, Lena e Philip, por serem segundos pais durante esses os últimos anos.

Ao meu avô Manuel, que me ensinou inúmeros conhecimentos e nos deixa saudades.

A minha avó Guilhermina, por todo apoio e dedicação durante todos esses anos e por ser quem ela é.

Aos meus amigos Jong, Marília, Dyan, Camila e Fernanda pelas horas agradáveis de diversão.

Ao Fauler por tentar me fazer ter uma nova perspectiva sobre tudo.

Ao Núcleo de Fixação de Nitrogênio do departamento de Bioquímica e Biologia Molecular, da Universidade Federal do Paraná.

Ao Instituto Nacional de Ciência e Tecnologia de Fixação Biológica de Nitrogênio.

Aos órgãos fomentadores: CAPES, CNPq e REUNI.

As todas as formas divinas que de algum modo me trouxeram até aqui.

“A persistência é o menor caminho do êxito.”

Charles Chaplin

RESUMO

O propósito da anotação é identificar sequências de DNA codificadoras de RNAs ou proteínas, esse processo é importante porque atribuem funções moleculares aos produtos gênicos. Para isso, são utilizadas ferramentas computacionais de anotação de genes que usam alinhamentos de sequência de proteína ou de DNA com o propósito de identificar genes homólogos e utilizar as informações de banco de dados de domínio público para inferir a função do gene. Embora sejam técnicas eficientes, elas podem estar sujeitas a erros quando realizada sem curadoria de um perito, em particular quando ocorre inexistência de grau de similaridade significativo de uma sequência comparada com outras sequências ou quando o banco de dados é composto por sequências parciais. Além disso, a taxa de erro de anotação pode ser significativamente aumentada quando a sequência de proteína de consulta é nova, compartilhando nenhuma semelhança com qualquer sequência disponível em bases de dados. Por esses motivos, neste trabalho desenvolveu-se uma ferramenta para verificar anotação de genes em genomas completos de bactérias, o programa *Bioinformatics Tool Based on Bacterial Genomes Comparison* (BOBBLES). Ele realiza a verificação da predição de genes computacionalmente propostos pelo programa *Hybrid-Gene Finder* (HGF). O programa BOBBLES compara a anotação de um genoma de referência completo de bactérias com os genes identificados pelo programa HGF. Este programa utiliza duas abordagens de comparação de sequências, uma utilizando pesquisas de similaridade de sequência através do programa BlastP e a outra utilizando o programa SILA. Ambas as abordagens servem para decidir se as sequências sugeridas pelo programa HGF foram anotadas corretamente. Para testar a ferramenta BOBBLES, utilizou-se um conjunto composto por 14 genomas bacterianos completos. Foram encontrados 365 novos genes e 101 genes com melhor ou similar grau alinhamento em fase de leitura diferente do genoma de referência, resultando em uma porcentagem de acerto de aproximadamente 76 % para esse conjunto de genomas, utilizando o alinhamento das sequências com o programa SILA. Já com o alinhamento realizado pelo programa Blastp obteve-se 529 novos genes. No entanto, o tempo médio estimado de execução do programa BOBBLES tendo em seu algoritmo a ferramenta SILA é de pelo menos cinco vezes mais rápido do que utilizando o programa BlastP. Essa diferença de tempo é justificada pelo fato do programa SILA realizar os alinhamentos das sequências com indexação recursiva em um banco de dados local, o banco de dados de proteínas não redundantes do NCBI, conhecido por NR.

Palavras-chave: Bioinformática. Anotação genômica. Blast. Reanotação genômica.

ABSTRACT

The annotation purpose is to identify DNA sequences coding for proteins or RNAs, this process is important because it gives the molecular function for the genes products. For that, it's used Gene Annotation tools using protein or DNA sequences alignments to identify homologous genes and use information from the public database to infer gene function. Although these are efficient techniques, they can be error-prone when performed without curation of an expert, particularly in cases of similarity sequence with no degree of similarity with other sequences that may be relevant or when the database is composed by partial sequences. In addition, annotation error rate can be significantly increased when it's a new query protein sequence, sharing no similarity with any available sequence in databases. Therefore, this work has developed a tool to verify genes annotation in complete bacterial genomes, the Bioinformatics Tool Based on Bacterial Genomes Comparison program (BOBBLES). It realizes the computationally gene prediction performed by Hybrid-Gene Finder (HGF). The BOBBLES compares a previous complete bacterial genome annotation with the genes identified by HGF program. This program uses two sequence comparison approaches, the first one using the BlastP program, and another approach using the SILA program, to decide whether they were recorded correctly. The BOBBLES was tested using a set composed of 14 complete bacterial genomes. These tests obtained 365 new genes and 101 genes with better or similar alignment in process of reading different from the reference genome, resulting in 76% of correct results for genomes set which used the alignment of sequences with the SILA program. But using the BlastP program, 529 new genes were obtained. However, the estimated average execution time for the BOBBLES program using SILA program was at least five times faster than using the BlastP program. This time difference is justified by the fact that the SILA program performs the alignments of the sequences with recursive indexing into a local database, the NCBI’s non-redundant protein sequence (NR) database.

Keyword: Bioinformatics. Genomic annotation. Blast. Genomic re-annotation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

A – Adenina
BLAST® – *Basic Local Alignment Search Tool*
BLASTP – *Protein* BLAST
BLOSUM – *Blocks Substitution Matrix*
BOBBLES – *Bioinformatics Tool Based on Bacterial Genomes Comparison*
C – Citosina
CDS – (C*oding Sequence*) Sequência de Região Codificante
DDBJ – DNA *Data Bank of Japan*
DNA – Ácido Desoxirribonucleico
EMBL – *European Molecular Biology Laboratory*
ENA – *European Nucleotide Archive*
E-Value – (*Expectation Value*) Valor Provável
G – Guanina
GBK – *Guojia biaozhun kuozhan*
GenBank® – Banco de Dados de Sequência Genética
GFFF – *GenBank Flat File Format*
Glimmer® – *Gene Locator and Interpolated Markov ModelER*
HGF – *Hybrid-Gene Finder*
HSP – *Height-scoring Segment Pair*
IA – Inteligência Artificial
IAC – Inteligência Artificial Cognitiva ou Simbólica
IAS – Inteligência Artificial Conexionista ou Subsimbólica
IMM – Modelo de Markov Interpolado
INREC – Indexação Recursiva
MATLAB® – *Matrix Laboratory*
MLP – *Multilayer Perceptron*
NCBI® – *National Center for Biotechnology Information*
NR – Banco de Dados de Sequências de Proteínas Não Redundantes
ORF – (*Open Reading Frame*) Fase ou quadro de leitura aberta
PAM – (*Percent Accepted Mutation*) Reconhecimento por Porcentagem de Mutação
pb – Pares de Base

RBS – (*Ribosome-Binding Site*) Sítio de Ligação de Ribossomos
ANN – (*Artificial Neural* Network) Rede Neuronal Artificial
rRNA – (*Ribosomal Ribonucleic acid*) Ácido Ribonucleico Ribossomal
SILA – *Sequence-Indexed Local Aligner*
T – Timina
TIGR – *The Institute for Genomic Research*
tRNA – (*Transfer Ribonucleic acid*) Ácido Ribonucleico Transportador
UFPR – Universidade Federal do Paraná

SUMÁRIO

1 INTRODUÇÃO

1.1 GENÔMICA

Genômica é uma área da biologia que estuda os genes de um genoma, incluindo a interação desses genes entre si e com o ambiente. Tem o objetivo de descobrir as funções gênicas e com isso estudar cientificamente problemas, como doenças complexas, por exemplo: cardíacas, asma, diabetes e câncer, outras doenças (*NATIONAL HUMAN GENOME RESEARCH INSTITUTE*, 2010).

Ela pode ser dividida em duas principais vertentes: genômica estrutural e genômica funcional. A primeira caracteriza a natureza física dos genomas completos. Já a segunda, distingue as regiões codificadoras e padrões globais de expressão de genes (GRIFFITHS *et al*, 1999). A caracterização de genomas completos é uma tarefa importante, pois proporciona uma maneira de conseguir uma visão global da arquitetura genética de um organismo, além de prover todos os dados para a descoberta de novos genes, como os envolvidos em doenças. A caracterização de sequências expressas fornece uma representação dos componentes funcionais, os quais são determinantes para a fisiologia celular de um organismo (LIBERMAN, 2004; BROWN *et al*, 2002).

Hoje, existem mais de 2680 sequências completas de DNA de genoma bacteriano e mais de 130 sequências de *archea* no banco de dados genômico de domínio público, os quais são atualizados diariamente (NCBI, 2012). Isso é possível devido aos programas de sequenciamento de genomas que geram grande quantidade de sequências inferidas de aminoácidos. Essas sequências são utilizadas por pesquisadores em análises genômicas. Para a obtenção automatizada de parte dessas análises, em específico a predição de estrutura primária de proteínas ou até mesmo para a obtenção de parceiros de interações é necessário grande esforço computacional, tanto a parte de algoritmo quanto a parte de máquinas computacionais potentes, além de um grupo de pesquisadores para fazer a acurácia de cada informação obtida. Por isso, a caracterização completa de uma proteína em um organismo sempre exigirá investigações experimentais adicionais

nas proteínas purificadas *in vitro* assim como estudos *in vivo* (PETSKO e RINGE, 2003; LIBERMAN, 2004).

1.2 BIOINFORMÁTICA

A partir das últimas décadas o número de informações biológicas expandiu de tal forma começou precisar de mais cuidado em relação ao armazenamento desses conhecimentos. Isso exigiu o aprimoramento de técnicas computacionais para que esses dados pudessem ser mais bem compartilhados pela comunidade científica de forma eficaz e eficiente. Com isso, a Bioinformática, segundo (SETUBAL *et al*, 2004), pode ser definida como o estudo da biologia através de técnicas das ciências da informação. Ela tem o objetivo de permitir a descoberta de novos conhecimentos biológicos, bem como criar uma perspectiva global de princípios unificadores da biologia que podem ser discernidos (NCBI, 2004).

No início, sua principal preocupação era com a criação e manutenção de um banco de dados para armazenamento de informações biológicas, como as sequências de nucleotídeos e aminoácidos. O desenvolvimento desse tipo de banco de dados exigiu questões como uma interface eficiente para que os pesquisadores pudessem ter acesso sem grandes dificuldades, além de permitir a possibilidade de acrescentar ou revisar algum dado contido nesse banco. Essas informações também deveriam ser combinadas para formar uma imagem completa das atividades celulares, com o objetivo de permitir aos pesquisadores acesso a essas informações a fim de que eles pudessem estudar como essas atividades são alteradas de acordo com o objeto de estudo (NCBI, 2004).

Por isso, outro ponto importante na evolução dos estudos na área da Bioinformática foi o desenvolvimento de ferramentas computacionais capazes de realizar análises e interpretações de vários tipos de dados, incluindo as sequências de aminoácidos e nucleotídeos, os domínios de proteínas. Para tal feito, foi necessário desenvolver um conjunto de procedimentos computacionais e métodos estatísticos para avaliar as relações entre os membros de grandes conjuntos de dados, tais como métodos para localizar um gene dentro de uma sequência, prever estrutura e funções de proteína, além de relacionar as famílias de sequências de proteínas (NCBI, 2004).

1.2.1 Sequenciamento e Montagem Genômica

Sequenciamento é um processo utilizado para montagem de sequências de DNA em que o genoma deve ser dividido em fragmentos e estes fragmentos sequenciados precisam ser remontados em uma sequência contínua (SANGER *et al*, 1977). A representação dessa sequência contínua é dada na forma de uma cadeia de letras que correspondem às bases da sequência, A, T, C ou G, correspondendo às bases adenina, timina, citosina e guanina, respectivamente. Na estrutura em dupla fita antiparalela do DNA essas bases são pareadas, sendo A com T e C com G.

O sequenciamento e a montagem genômica podem ser divididos em duas fases: fase experimental e fase de análise, FIGURA 1.1. Na fase experimental, as bactérias são cultivadas e seus DNAs são extraídos e fragmentados, esses fragmentos são clonados; e por fim, ocorre o sequenciamento destes. Na fase de análise computacional, as sequências individuais ou leituras, originadas na fase anterior, são ordenadas e montadas por identidade parcial de suas extremidades; são realizadas estratégias de finalização destas sequencias com o propósito de obter uma única sequência representando o genoma. Então, são buscadas e anotadas genes e regiões de interesse com o intuito de poder servir para novos estudos.

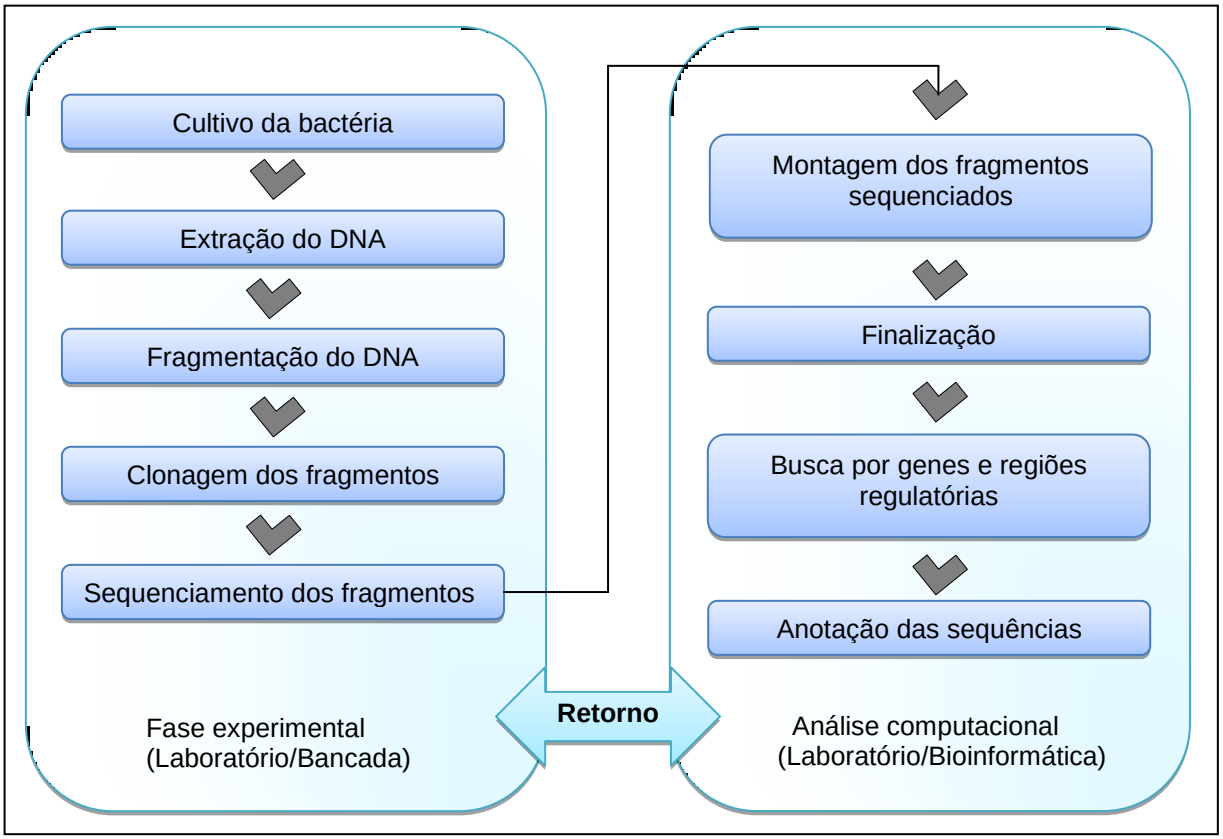


FIGURA 1.1 - ETAPAS DE UM PROCESSO DE SEQUENCIAMENTO E MONTAGEM GENÔMICA
FONTE: Adaptação de GENOMAR. Disponível em: <nfn.genopar.org/nfn/genopar/>. Acesso em: 19/01/2012.

1.2.2 Anotação Genômica

Após os processos de sequenciamento e montagem genômica é necessário identificar as características do genoma, ou seja, quais são genes e sua função provável. Para isso, é realizada a identificação através do processo de anotação genômica. Esse processo evolve a organização de um organismo com o objetivo de extrair informações biológicas úteis sobre a sequência estudada e com isso, por exemplo, identificar genes, elementos funcionais em DNA genômico ou inferências de funções dos genes nos genomas (GIBAS *et al*, 2002; STEIN, 2001; WESTHEAD *et al*, 2002).

1.2.2.1 Etapas do processo de anotação

Pode-se dizer que a anotação de um genoma é um processo dividido em várias etapas e três categorias: anotação de nucleotídeos de nível superior, nível de proteína e nível de processo, como mostra a FIGURA1.2 (STEIN, 2001).

A etapa inicial para a realização da anotação do genoma, FIGURA 1.2(A), consistem em identificar os padrões das sequências, como o reconhecimento dos códons de início (*start*) e final (*stop*) de tradução, regiões codificadoras de proteínas, sítios de ligação de ribossomos (RBSs), regiões reguladoras e promotoras, e outros. Também existem regiões conservadas ou previamente conhecidas que devem ser mapeadas. Por exemplo, ácido ribonucleico transportador (tRNA), ácido ribonucleico ribossomal (rRNA) e elementos repetitivos (STEIN, 2001).

Com as regiões de interesse na sequência identificadas, é necessário abranger o máximo de informações possíveis referentes a essas sequências, FIGURA 1.2(B). Esse processo é feito através de análises com o intuito de obter uma relação completa das proteínas contidas no organismo em estudo, contendo os nomes e funções de cada uma delas. Isso é feito por meio de comparações com genes conhecidos de organismos normalmente próximos taxonomicamente do genoma estudado. Nesse processo de identificação apenas parte do conjunto de genes codifica proteínas com funções conhecidas, os outros genes geralmente codificam proteínas hipotéticas ou conservadas sem função claramente definidas, o que torna esse processo bastante complicado (STEIN, 2001).

Após a definição identificação das proteínas é possível realizar a inferência de suas funções e as relações com processos biológicos, FIGURA 1.2(C). Um conjunto de genes pode ser associado a funções, o que permite agrupar essas proteínas em categorias, como em processamento e armazenamento de informações, processos celulares e sinalização, metabolismo, e outras. Também se podem reconstituir as vias metabólicas, caracterizar sistemas de transporte e secreção, entre outras (STEIN, 2001).

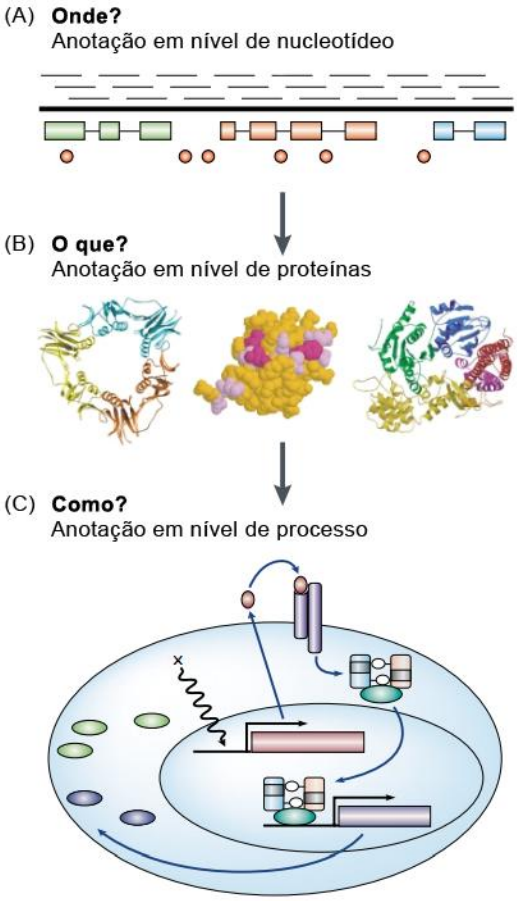


FIGURA 1.2 - ETAPAS DO PROCESSO DE ANOTAÇÃO GENÔMICA
FONTE: Adaptação (STEIN, 2001)

1.2.2.2 Obtenção das informações contidas nos genomas

Uma das subdivisões da anotação genômica é a anotação funcional que é a predição ou determinação das sequências de aminoácidos (LIBERMAN, 2004). Ela designa como e quando estes genes serão expressos e quais as suas interações com outros genes.

Para a obtenção dessas informações contidas nos genomas, existem diversos programas computacionais os quais auxiliam no processo de anotação. “Eles devem ter modelados tanto as anotações externas, armazenadas em fontes de dados públicas externas, quanto as anotações internas, armazenadas no *data warehouse* sob controle do sistema, além de oferecer mecanismos de extensão do modelo para acomodar novas anotações” (LEMOS, 2004).

De acordo com (LEMOS *et al*, 2004), as anotações internas podem ser classificadas em:

- Importada: que é aquela obtida através de banco de dados de domínio público, como o GenBank® (BENSON *et al*, 2004), NodMutDB (MAO *et al*, 2005) e o Swiss-Prot (BOECKMANN *et al*, 2003);
- Automática: corresponde a anotação produzida através de programas analíticos disponíveis na área de Bioinformática, como o Gendb (MEYER *et al*, 2003) e o Blast2go (CONESA *et al*, 2005);
- Manual: é a anotação criada diretamente pelo pesquisador.

As ferramentas mais utilizadas para auxiliar no processo de anotação são aquelas cujo algoritmo procure sequências com similaridade contra sequências de um banco de dados escolhido, que pode ser uma ou um conjunto de sequências determinado ou um banco de dados de domínio público. A similaridade é baseada no alinhamento de sequências que pode ser definido “pela forma de colocar uma sequência sobre a outra, de modo a obter uma correspondência entre cada base” da sequência a ser comparada e cada base da sequência que está comparando. Esse alinhamento pode conter espaços entre as bases das sequências para proibir uma correspondência ou completar posições faltantes (SETUBAL, 2004).

A FIGURA 1.3 ilustra um alinhamento entre duas sequências, onde as barras verticais indicam as posições onde as bases são iguais. Quando não ocorre alinhamento, ou seja, as bases são diferentes ou inexistentes na sequência correspondente, as barras verticais não são apresentadas (SETUBAL, 2004).

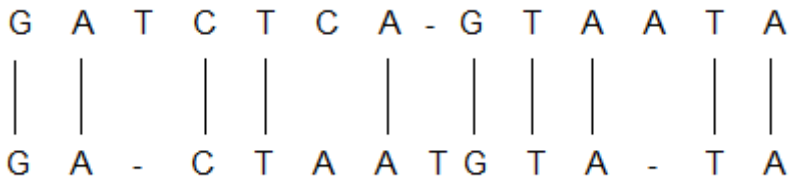


FIGURA 1.3 - ILUSTRALÇÃO DE ALINHAMENTO ENTRE DUAS SEQUÊNCIAS
FONTE: Adaptação (SETUBAL, 2004)

Além disso, existem dois tipos de alinhamento de sequências: alinhamento global e alinhamento local (WESTHEAD *et al*, 2002). No alinhamento global, as duas sequências são conhecidas e todo o seu comprimento deve ser alinhado. Já no alinhamento local, não existe necessidade de estender toda a sequência para o alinhamento. No processo de alinhamento, cada par de sequências recebe uma nota

de pontuação, que pode ser positiva ou negativa. Caso a pontuação negativa exceda o limite estabelecido pelo algoritmo, esse alinhamento pode ser interrompido e reiniciado em outro trecho da sequência. Isso caracteriza um alinhamento parcial, ou seja, cobre apenas um local da sequência.

Esses alinhamentos são feitos através de programas que realizam busca por similaridade entre uma sequência em estudo, alvo ou *query,* com uma ou várias sequências presentes em um banco de dados. O resultado varia de zero até um conjunto de sequências similares à sequência *query,* e cada sequência similar, *hit* ou *subject*, mostrará um valor (*score)* que indicará quanto essa sequência é similar à sequência *query*. Quanto maior esse valor, maior o grau de similaridade. Entre essas ferramentas, as mais utilizadas pelos pesquisadores da área de Bioinformática, são o *Basic Local Alignment Search Tool* (BLAST, que em português significa Ferramenta de Busca Básica de Alinhamento Local) (ALTSCHUL *et al*, 1990), e o FASTA (LIPMAN e PEARSON, 1985). Ambos realizam buscas rápidas em bancos de dados de sequências de nucleotídeos ou de proteínas utilizando tanto o alinhamento global ou o local.

1.2.3 Reanotação genômica

A reanotação genômica consiste em anotar novamente uma proteína utilizando um banco de dados mais recente. Existem vários motivos que levam um genoma a ser reanotado, como o aumento significativo do banco de dados com sequências homólogas, a descoberta experimental da função de novos genes ou até mesmo a utilização de algoritmos de anotação mais eficientes. Durante esse processo, é possível realizar testes e comparações com os meios diferentes de anotação, quais genes foram perdidos e quais foram obtidos através da nova abordagem. A reanotação é uma tarefa importante principalmente em casos onde a anotação original foi baseada em baixo grau de similaridade na comparação das sequências ou quando a anotação do banco de dados era precária (OUZOUNIS e KARP, 2002).

1.2.4 Falhas de predição e anotação genômica

Devido à velocidade com que pesquisadores geram sequências de DNA atualmente, anotar os dados automaticamente tornou-se um desafio computacional, e a análise humana cuidadosa é cada vez mais difícil (LEMOS *et al*, 2004).

A similaridade de sequências possui grande potencial de incorrer em falsos positivos, apesar da anotação funcional automática de genoma ser realizada com grande eficiência, (SANTOS *et al*, 2011; WONG *et al*, 2010; LORENZI *et al*, 2010). Isso acontece porque dado um grau de similaridade suficiente, geralmente assume-se que a funcionalidade do novo gene provavelmente seja a mesma que a dos seus melhores vizinhos resultantes da busca por similaridade. No entanto, esses genes homólogos podem ter sido adquiridos através de similaridade com outras proteínas que não correspondem corretamente, ou seja, "existe possibilidade de ocorrer cadeias de anotações erradas" (LEVY *et al*, 2005), esse processo é conhecido por "*Error Percolation*" (GILKS *et al*, 2002).

Outro problema apontado por (WARREN *et al*, 2010) é a não anotação de vários genes pequenos, ou seja, genes com menos de 100 pares de base (pb), contidos no genoma. Isso ocorre possivelmente porque os programas mais utilizados para a detecção de quadros de leitura aberta; do inglês *open reading frame* (ORFs) que são sequências de DNA que não contém um códon de parada em um determinado quadro de leitura (DEONIER *et al*, 2005), não são capazes de encontrar boa parte esses genes pequenos. Exemplo desses programas são o Glimmer (SALZBERG *et al*, 1998) e o Genemark (SHULAEV *et al*, 2010). Além disso, na pesquisa realizada por (WARREN *et al*, 2010), foram encontrados 1153 genes candidatos faltantes nas anotações de genomas os quais são semelhantes entre si e que não constam em bancos de dados de domínio público, o que implica que essas ORFs, pertencem a famílias de genes ainda não anotadas. Também foram encontrados 38895 ORFs em regiões intergênicas, a maioria com menos de 100 pb, identificadas como genes prováveis pela semelhança com genes anotados.

1.2.5 Glimmer

Gene Locator and Interpolated Markov ModelER (Glimmer) , que em português significa Localizador de Gene e Interpolador do Modelo de Markov ER, é

um sistema para encontrar genes em DNA microbiano, especialmente nos genomas de bactérias, archaea, e vírus. Ele utiliza Modelos de Markov Interpolados (IMM) como uma estrutura para captar dependências entre os nuvleotídeos próximos, em uma sequência de DNA (SALZBERG *et al*, 1998).

O modelo IMM faz previsões com base em um contexto variável, ou seja, uma variável de comprimento variável oligômero (fragmentos curtos de DNA) em uma sequência de DNA. Nesse programa, ele realiza alterações dependendo da composição local da sequência, como resultado, ele é mais flexivel e mais poderoso do que métodos de Markov de fixa ordem para encontrar genes em DNA microbianos. Utilizando essa técnica, ele provou ser capaz de localizar praticamente todos os genes nas sequências testadas, com uma estimativa de mais de 97% dos genes em *Haemophilus influenzae* e *Helicobacter pylori* (SALZBERG *et al*, 1998).

1.2.6 Banco de Dados Biológicos

Banco de dados biológicos constitui um grande conjunto de dados persistentes, geralmente associado a um programa projetado para atualizar, consultar e recuperar os dados armazenados no sistema. Um banco de dados simples pode ser um arquivo contendo muitos registros, cada um dos quais inclui o mesmo conjunto de informações. Um exemplo disso é um registro associado a um banco de dados de sequências de nucleotídeos que normalmente contém informações como a sequência de entrada, com uma descrição do tipo e do nome do organismo (NCBI, 2004).

Para os pesquisadores poderem se beneficiar desses dados contidos no banco, ele necessita atender os seguintes requisitos:

- Fácil acesso à informação;
- Método eficiente para extrair apenas as informações que o pesquisador precisa.

Um exemplo disso são os vários bancos de dados biológicos contidos no NCBI os quais são ligados através de uma pesquisa única e um sistema de recuperação, o Entrez (GEER *et al*, 2003), que permite ao usuário processar e recuperar as informações específicas a partir de um único banco de dados que está no NCBI. Por exemplo, um cruzamento do banco de dados de Proteínas Entrez com

o de Taxonomia Entrez, permite a um pesquisador encontrar as informações taxonômicas para espécies a partir de uma sequência de proteínas (NCBI, 2004).

1.2.6.1 GenBank®

O GenBank® (BENSON *et al*, 2010) é um banco de dados de sequência de nucleotídeos de domínio público que atualmente é um dos bancos de dados biológicos contidos no NCBI. Ele contém mais de 380 mil organismos nomeados em nível de gênero ou inferior, obtidos principalmente através de observações produzidas por diversos pesquisadores e submetidas em lotes de projetos de sequenciamento em larga escala. Ele realiza a troca diária de dados com o Arquivo Europeu de Nucleotídeos (tradução de *European Nucleotide Archive* (ENA)) (BRUNAK *et al*, 2002) e o Banco de Dados de DNA do Japão (tradução de *DNA Data Bank of Japan* (DDBJ); disponível em http://www.ddbj.nig.ac.jp) para assegurar a cobertura completa de sequências existentes pelo mundo. Além disso, é acessível através do sistema de recuperação Entrez (GEER *et al*, 2003), do NCBI, o qual integra os principais dados de DNA e bancos de dados de sequências de proteína junto com a taxonomia, estrutura do genoma, proteína de mapeamento e informações de domínio e a leitura biomédica através da revista PubMed. Outra vantagem da utilização desse banco é que o BLAST®, um dos programas mais utilizados para alinhamento de sequências, fornece buscas por similaridade de sequências contra banco de dados GenBank®. As atualizações diárias desse banco estão disponíveis em ftp://ftp.ncbi.nih.gov/genbank/ (BENSON *et al*, 2010).

1.2.7 BLAST

Basic Local Alignment Search Tool (BLAST), que em português significa Ferramenta de Busca Básica de Alinhamento Local, é uma ferramenta para encontrar regiões com similaridade entre sequências (ALTSCHUL *et al*, 1990). O programa consiste em fazer comparações entre uma sequência (*query*) de nucleotídeos ou proteínas contra um banco de dados de sequência, utilizando em sua tecnologia o algoritmo Smith-Waterman (SMITH e WATERMAN, 1981) e realiza a significância estatística de cada sequência alinhada. Dessa forma, pode ser

utilizado para inferir relações funcionais e evolutivas entre essas sequências e também auxiliar a identificar membros de famílias de genes (FASSLER e COOPER, 2011).

O algoritmo, FIGURA 1.4, realiza a busca inicial do alinhamento através uma palavra de comprimento “W” em que as pontuações sejam pelo menos “T” quando comparadas com a consulta usando uma matriz de substituição, BLOSUM ou PAM. Quando encontrada essa palavra, a sequência é estendida em qualquer direção na tentativa de gerar um alinhamento com uma pontuação superior ao limiar de “S”. O “T” parâmetro dita a velocidade e sensibilidade da pesquisa (FASSLER e COOPER, 2011).

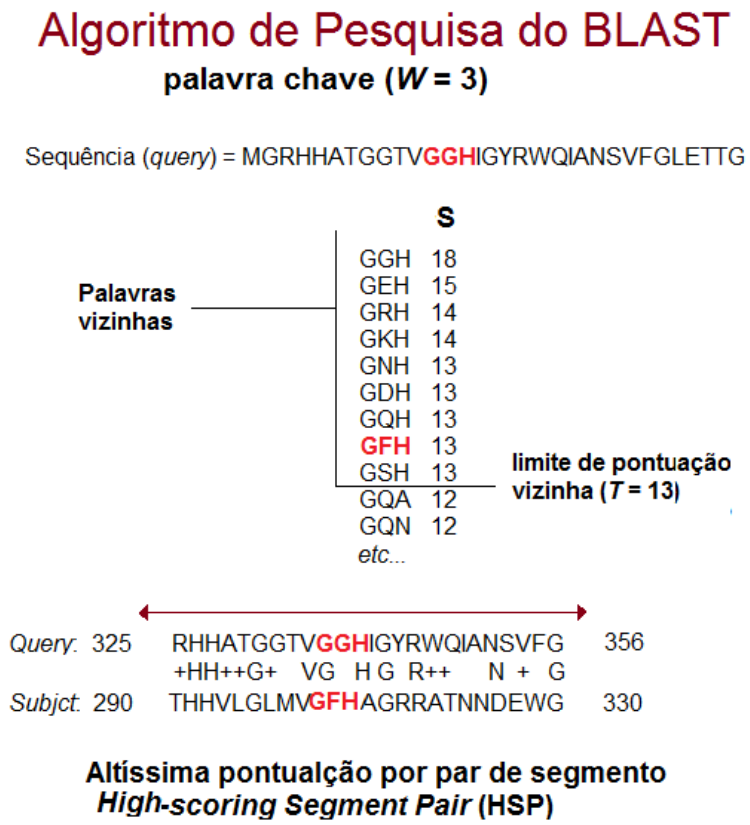
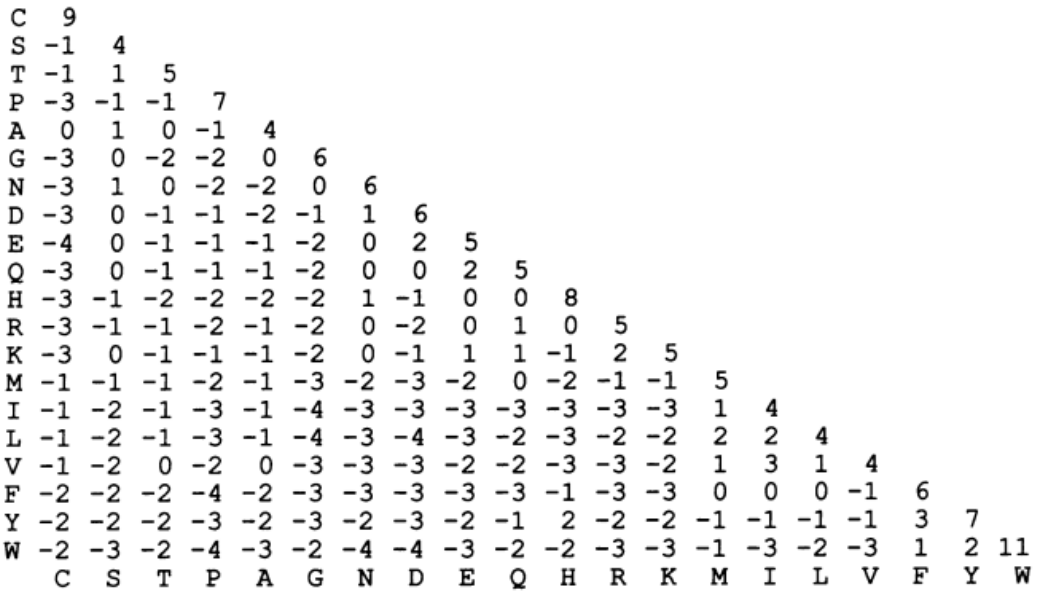


FIGURA 1.4 - ALGORITMO DE PESQUISA DO BLAST
FONTE: Adaptação (FASSLER e COOPER, 2011)

1.2.7.1 Matrizes de substituição

Blocks Substitution Matrix (BLOSUM), que em português significa Matriz de Substituição em Blocos, é uma matriz de pontuação de substituição. Cada posição da matriz possui um valor de pontuação derivados a partir do valor de variação das

frequências de substituições em blocos de alinhamentos locais em proteínas relacionadas. Cada matriz é adaptada a uma determinada distância evolutiva. Por exemplo, a matriz BLOSUM62, FIGURA1.5, a possui o alinhamento com o valor de avaliação (*score*) derivados de sequências as quais partilham não mais do que 62% de identidade. E as sequências com mais de 62% de identidade são representadas por uma única sequência no alinhamento (FASSLER e COOPER, 2011).



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

FIGURA 1.5 - MODELO DE MATRIZ BLOSUM
FONTE: Adaptação (FASSLER e COOPER, 2011)

A *Percent Accepted Mutation* (PAM), que em português significa Reconhecimento por Porcentagem de Mutação, é uma metodologia capaz de identificar a quantidade de mudança evolutiva em uma sequência de proteínas. A matriz PAM(x) constitui-se de uma tabela de procura, em que a pontuação para cada substituição de aminoácido é calculado com base na frequência em que a substituição próxima das proteínas relacionadas sofreu uma quantidade “X” de divergência evolutiva (FASSLER e COOPER, 2011).

Segundo (HENIKOFF *et al*, 1997), as matrizes de BLOSUM possuem melhor desempenho em alinhamentos e buscas de homologia do que as baseadas em mutações aceitas em grupos intimamente relacionados, como a PAM. Pode-se dizer que essas matrizes de segmentos alinhados em blocos, que representam as regiões mais conservadas em proteínas, são mais apropriadas para pesquisas e alinhamentos do que as do tipo PAM.

1.2.7.2 *Bit Score*

O *bit score*, (S'), é derivado da pontuação de alinhamento bruto (S), tendo as propriedades estatísticas do sistema de pontuação em conta. Ou seja, é o valor do escore bruto normalizado. Isso porque essa normalização leva em consideração a escala da matriz de escores utilizada (λ) e a escala do tamanho do espaço de busca (K) (FASSLER e COOPER, 2011; LIBERMAN, 2004). Sendo assim, o *bit score* é dado por:

$$S' = \frac{(\lambda S - \ln K)}{\ln 2}.$$

Segundo Liberman (LIBERMAN, 2004), o *bit score* não possui dependência com o tamanho do banco de dados usado e, por isso, pode ser mais confiável do que o uso do *E-Value* para analisar os resultados da plataforma BLAST®. Ele também sugere um valor em torno de 100 para o limiar de *bit-socre*.

A plataforma BLAST® remoto apresenta cinco faixas de pontuação, FIGURA1.6, divididas por conjuntos de valores de *bit score*, onde cada conjunto é representado por uma cor. São elas:

- Inferior a 40 em preto,
- De 40 até 50 em azul,
- De 50 até 80 em verde,
- De 80 até 200 em rosa,
- E 200 ou superior em vermelho.

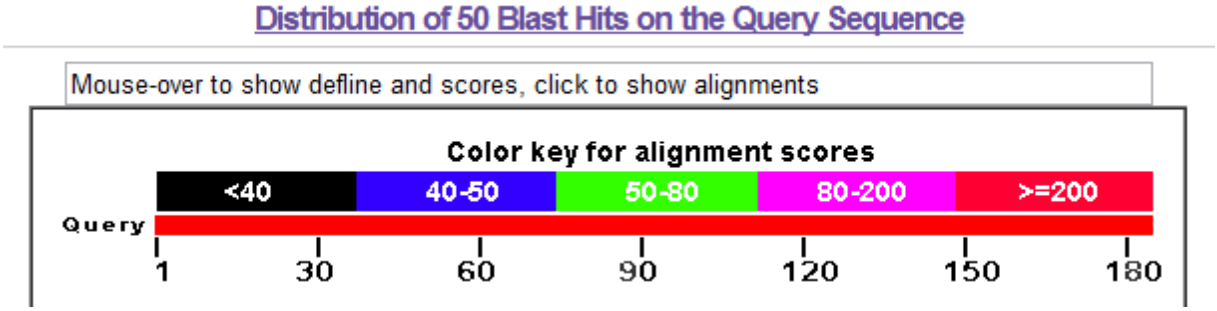


FIGURA 1.6 - CHAVES DE CORES PARA AS PONTUAÇÕES DE ALINHAMENTO DO BLAST®
FONTE: NCBI (<www.ncbi.nlm.nih.gov/blast/Blast.cgi>)

1.2.7.3 E-Value

O *Expectation Value* (E-Value), que em português significa Valor Provável, corresponde à probabilidade de se obter por acaso, com outra sequência aleatória de mesmo tamanho e composição de letras, outro alinhamento de *score* igual ou superior no banco de dados pesquisado. Esse valor é obtido através da seguinte fórmula: $E = mn2^{-S'}$, onde m é o tamanho do banco de dados (quantidade de caracteres), n é o tamanho da sequência de entrada e S' é o valor de *bit score.* Quanto menor esse valor mais significativo é o *score* (FASSLER e COOPER, 2011; LIBERMAN, 2004).

Pode-se dizer que quando uma sequência de entrada alinhada com alguma sequência de um banco de dados cujo E-Value for superior a 1 possivelmente esse alinhamento ocorreu por acaso, ou seja, existe grande indício dessa sequência não possuir valor de representatividade biológico significativo. No entanto, o tamanho da sequência de entrada e o tamanho do banco causam grande impacto no resultado desse cálculo. Por exemplo, quanto menor o tamanho da sequência, maiores são as chances de obter o valor de E-Value próximo ou superior a 1. Por essa razão devem-se observar outros fatores calculados pela plataforma BLAST®, como o *bit score*, percentual de identidade e o percentual de similaridade.

1.3 INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial (IA) é uma área da computação que visa resolver problemas computacionalmente quando a abortagem analítica computacional convencional não é viável.

Ela é dividida em duas abordagens: Inteligência Artificial Simbólica (IAS) ou Cognitiva e Inteligência Artificial Conexonista (IAC) ou Subsimbólica. A primeira tem o objetivo de simular o comportamento da mente humana através de símbolos, que são interpretações de entidades com algum significado dentro da lógica do algoritmo (RUSSELL e NOVIG, 1995; NEWELL e SIMON, 1976). Já a segunda, busca modelar a estrutura cerebral humana simulando o cérebro humano. Neste sistema, o algoritmo é capaz de assimilar, errar e aprender com seus erros (HAYKIN, 2001).

Um dos melhores exemplos de uma IAC é uma Rede Neuronal, ou Neural, Artificial (ANN).

1.3.1 Redes Neuronais Artificiais

Rede Neuronal Artificial (ANN) é uma subárea da IA inspirada no modelo neuronal humano. Nesse modelo, FIGURA 1.7, cada neurônio é representado por uma unidade de processamento e as sinapses correspondem às várias interligações de conexões entre um processamento e outro. De acordo com (HAYKIN, 2001):

> *"Uma rede neural é um processador maciçamente paralelamente distribuído constituído de unidades de processamentos simples que tem a propensão natural para armazenar conhecimento experimental e torna-lo disponível para o uso."*

Uma ANN assemelha-se ao cérebro humano no que diz respeito ao conhecimento que é adquirido pela rede a partir do ambiente de aprendizado e pela conexão entre os neurônios e que são utilizados para armazenar o conhecimento adquirido (HAYKIN, 2001).

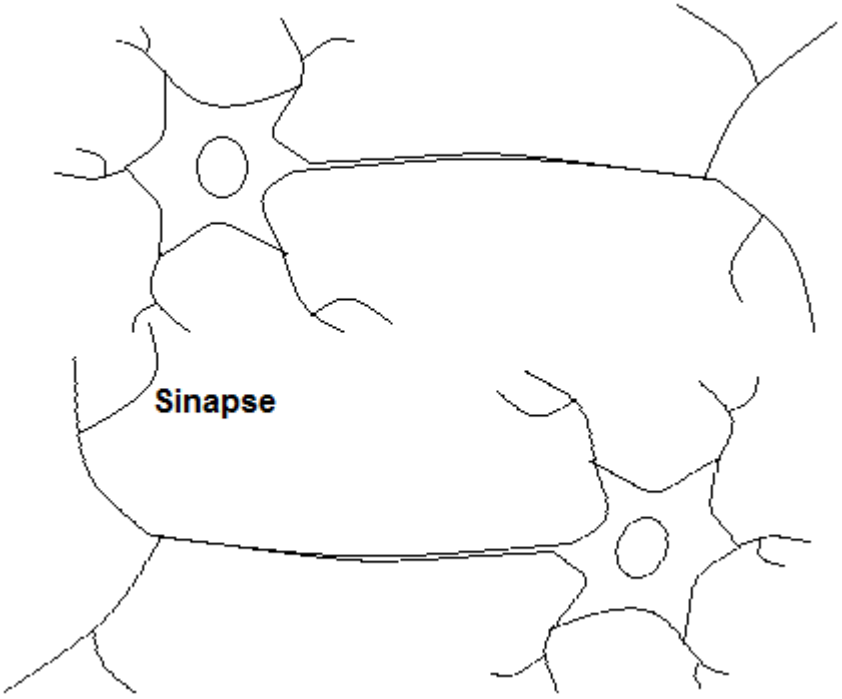


FIGURA 1.7 - REPRESENTAÇÃO DE DOIS NEURÔNIO
FONTE: Adaptação (WASSERMAN, 1989) (2012)

A estrutura computacional de uma ANN é baseada em:

- Estudo do problema;
- Modelos de estruturas e conexões sinápticas;
- Escolha de um algoritmo de aprendizado;
- Construção de um conjunto de treinamento;
- Treinamento da rede;
- Fase de testes;
- Utilização da rede (PAULA, 2000).

Dessa maneira, as ANN possuem a capacidade de tratar sistemas não lineares, ser tolerantes a falhas, adaptáveis a situações diversas, aprender a resolver o problema proposto baseado em modelos, generalização, ou seja, não é necessário saber todos os parâmetros do problema para poder resolvê-lo, e abstração.

1.3.1.1 Neurônio Artificial

Um neurônio artificial, assim como um neurônio biológico, possui um ou mais sinais de entrada e um sinal de saída. O primeiro modelo de uma rede neuronal foi proposta por (MCCULLOCH e PITTS, 1943), FIGURA 1.8. Eles elaboraram um modelo matemático para aproximar do comportamento de um neurônio. Esse modelo possui um dispositivo binário em que os dados passam pelos sinais de entrada, que podem vir de sensores ou de outros neurônios os quais fazem parte da ANN. Essas entradas têm um ganho arbitrário podendo ser excitatórias ou inibitórias. Depois, esses dados são processados e enviados para a saída, com o resultado de pulso ou não pulso, ou seja, ativo ou inativo. Essa saída é determinada de acordo com a soma ponderada das entradas com os respectivos ganhos como fatores de ponderação, excitatórios ou inibitórios. Caso o resultado atinja um determinado limiar, a saída pode ser ativo ou, caso contrário, não ativo.

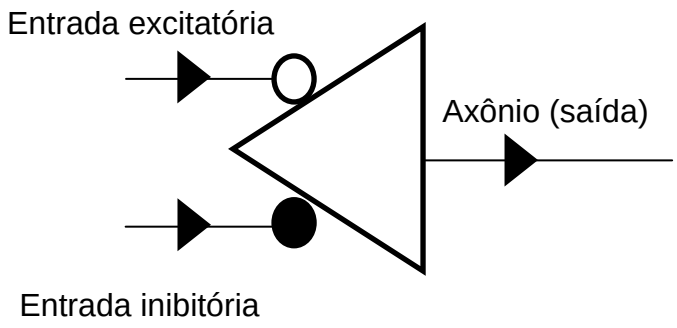


FIGURA 1.8 - MODELO DO NEURÔNIO DE MCCULLOCH
FONTE: Adaptação (PAULA, 2000)

Existem também os pesos que, fazendo uma analogia com um neurônio natural, eles são os representantes das sinapses, ou sinais sinápticos. Cada um deles, no modelo ANN, possui um valor que é alterado em função da intensidade do sinal de entrada, mudando o seu valor representativo para a rede. Esse processo também é conhecido como processo de aprendizagem da ANN. Sendo assim, quanto mais o valor de entrada for estimulado, mais estimulado será o peso correspondente, e quanto mais estimulado for esse peso, mais significante e influente ele será para o resultado do sinal de saída do neurônio. Além disso, os neurônios e os pesos são organizados na rede em forma de camadas, podendo ter uma ou mais camadas. E cada neurônio é conectado a um ou mais neurônios através das sinapses (HAYKIN *et al*, 1999). A FIGURA 1.9 ilustra esse esquema.

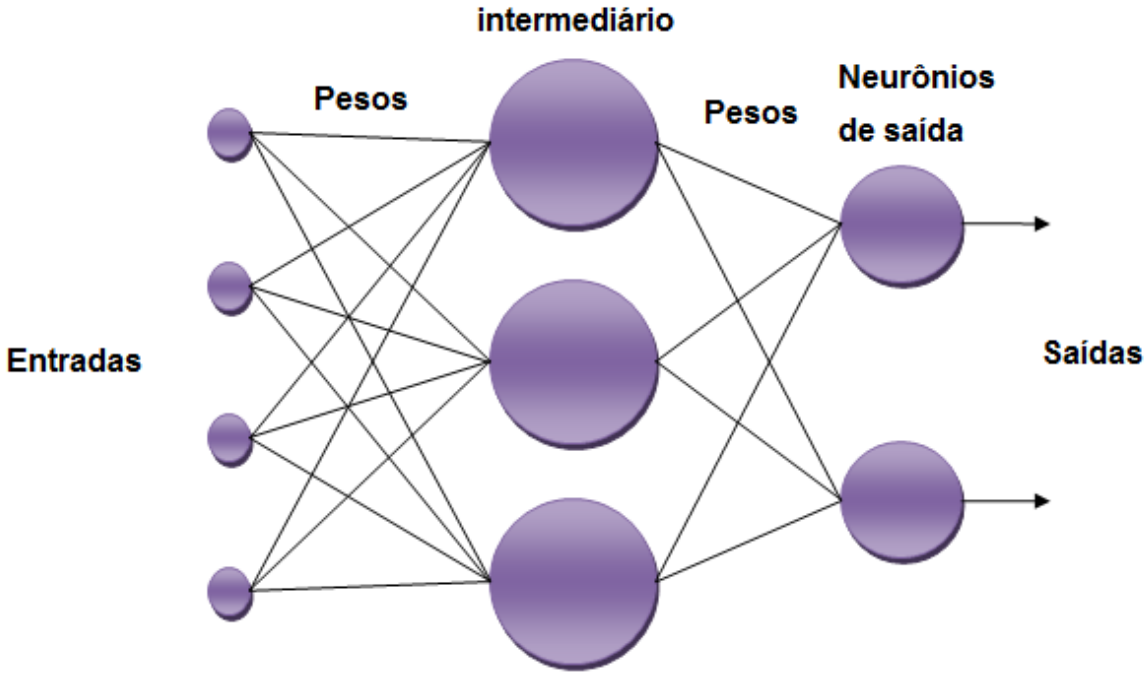


FIGURA 1.9 - MODELO GERAL DE UMA REDE NEURONAL ARTIFICIAL
FONTE: Adaptação (GEHLEN, 2011)

As ANN podem ser classificadas em (i) redes supervisionadas de aprendizado e (ii) redes não supervisionadas de aprendizado. No primeiro, a metodologia, designada por metodologia adaptativa, tenta reduzir do erro da saída. Esse aprendizado é feito com base em informações sobre um problema específico, geralmente através de um conjunto modelos. Enquanto no segundo, a metodologia foca no desenvolvimento de representações internas sem amostras de saídas (ELMASRI e NAVATHE, 2005).

1.4 HYBRID-GENE FINDER

O *Hybrid-Gene Finder* (HGF) realiza buscas de genes em sequências genômicas, ou seja, é capaz de realizar predição gênica. Ele está sendo desenvolvido pelo grupo de pesquisa de Bioinformática da Universidade Federal do Paraná (UFPR), o qual a autora faz parte, orientado pelo professor Dr. Roberto Tadeu Raittz, que é o precursor e o principal desenvolvedor dessa ferramenta.

Essa ferramenta faz uso de técnicas de Inteligência Artificial (IA) para detectar os prováveis *stop códons,* ou códons de parada, de sequências de regiões

codificantes (CDS). E a partir dessa detecção esse programa prevê uma posição para os *start* códons, ou códons de início, aproximando-se da posição real de início do gene (RAITTZ, R.T. dados não publicados, 2011).

Para o desenvolvimento do HGF foi utilizado um conjunto de genomas bacterianos completos a fim de treinar uma Rede Neuronal Artificial (ANN). Com o conjunto de dados treinados, foi utilizado outro conjunto de genomas, os mesmos utilizados na publicação do software Glimmer®, apresentado em 1.2.5, com o objetivo de comparar as duas ferramentas, detectar as variações de resultados e tentar encontrar a causa dessas variações. Nessa comparação foi observada que esta nova ferramenta de detecção de genes consegue detectar praticamente todos os genes que o Glimmer® propõe, além de detectar com eficiência novos genes, destacando-se por encontrar genes pequenos (RAITTZ, R.T. dados não publicados, 2011).

O funcionamento dessa ferramenta é realizado através da inserção do conjunto de dados de treinamento e um arquivo contendo a sequência completa de um genoma no formato FASTA. Os dados são processados e o retorno dele é um arquivo na extensão GBK desse genoma com a marcação dos prováveis CDS nas cores (i) vermelho, (ii) rosa claro e (iii) cinza. Essa diferença de cores serve para classificar o quão provável o gene é verdadeiro, sendo o primeiro com alto grau de chances, o segundo, possui menos chances, e o terceiro com menos chances que o segundo (RAITTZ, R.T. dados não publicados, 2011).

Esse programa está em desenvolvimento e é uma ferramenta promissora para detecção de genes, com diferencial na detecção de genes pequenos. No entanto, ele ainda não está disponível para toda a comunidade científica (RAITTZ, R.T. dados não publicados, 2011).

.

1.4.1 GBK

O *Guojia biaozhun kuozhan* (GBK), que é o chinês para “Regras ou especificações as quais definem as extensões de códigos internos de ideogramas chineses” (ORACLE, 2010; HP, 2012), em Bioinformática é uma extensão utilizada como parte de uma das especificações estabelecidas pelo GenBank® para conter

todo o genoma de um organismo anotado no formato GenBank *Flat File Format* (GFFF), que em português significa Formato de Arquivo Plano GenBank.

1.4.1.1 GFFF

A especificalção GFFF do GenBank é dividida em duas partes: (i) cabeçalho e (ii) sequência de nucleotídeos (GENBANK, 2011).

O primeiro contém os dados do genoma, como nome científico, projeto, citações para os artigos que contém relatos dos dados, lista de autores, título da citação, periódico em que foi publicado, identificação para o PubMed e características do organismo, exemplificado na FIGURA 1.10. Em especial, as características do organismo, também conhecidas por *features* constituem a maior parte do cabeçalho, por exemplo: CDS, gene, rRNA e outras. Isso ocorre poque existirem diversas *features* e cada uma contém diversas informações por *qualifires*. Cada *qualifire* inicia-se na coluna 22 do arquivo com uma “/” seguida por um nome qualificador, como /códon_start, /function e /note; e, se aplicável, um sinal de igual (=), exemplificado na FIGURA 1.11 (GENBANK, 2011).

Já o segundo, refere-se à sequência de nucleotídeos do organismo. Essa sequência é iniciada após a palavra ORIGIN e é relatada na direção 5’ para 3’. Nessa parte do arquivo, existem sessenta bases de nucleotídeos por linha, listadas em grupos de dez seguidas por um espaço em branco e são sempre iniciadas na coluna 11. As colunas de 4 a 9 contêm o número da posição do nucleotídeo referente à coluna 11, exemplificado na FIGURA 1.12 (GENBANK, 2011).

```
LOCUS       NC_010473            4686137 bp    DNA     circular BCT 23-JAN-2012
DEFINITION  Escherichia coli str. K-12 substr. DH10B chromosome, complete
            genome.
ACCESSION   NC_010473
VERSION     NC_010473.1  GI:170079663
DBLINK      Project: 58979
KEYWORDS    .
SOURCE      Escherichia coli str. K-12 substr. DH10B
  ORGANISM  Escherichia coli str. K-12 substr. DH10B
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 4686137)
  AUTHORS   Durfee,T., Nelson,R., Baldwin,S., Plunkett,G. III, Burland,V.,
            Mau,B., Petrosino,J.F., Qin,X., Muzny,D.M., Ayele,M., Gibbs,R.A.,
            Csorgo,B., Posfai,G., Weinstock,G.M. and Blattner,F.R.
  TITLE     The complete genome sequence of Escherichia coli DH10B: insights
            into the biology of a laboratory workhorse
  JOURNAL   J. Bacteriol. 190 (7), 2597-2606 (2008)
   PUBMED   18245285
REFERENCE   2  (bases 1 to 4686137)
  AUTHORS   Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (20-FEB-2008) Department of Genetics and Biotechnology,
            University of Wisconsin, 425G Henry Mall, Madison, WI 53706, USA
REFERENCE   3  (bases 1 to 4686137)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (01-OCT-2007) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from CP000948.
            DH10B and DH10B-T1R are available from Invitrogen Corporation
            (http://www.invitrogen.com).
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
```

FIGURA 1.10 – EXEMPLO DE CABEÇALHO DE ANOTAÇÃO GENÔMICA NO ARQUIVO DA EXTENSÃO GBK
FONTE: (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__DH10B_uid58979/NC_010473.gbk, Acessado em 26/01/2012)

```
     source          1..4686137
                     /organism="Escherichia coli str. K-12 substr. DH10B"
                     /mol_type="genomic DNA"
                     /strain="K-12"
                     /sub_strain="DH10B"
                     /db_xref="taxon:316385"
     gene            190..255
                     /gene="thrL"
                     /locus_tag="ECDH10B_0001"
                     /db_xref="GeneID:6058969"
     CDS             190..255
                     /gene="thrL"
                     /locus_tag="ECDH10B_0001"
                     /note="involved in threonine biosynthesis; controls the
                     expression of the thrLABC operon"
                     /codon_start=1
                     /transl_table=11
                     /product="thr operon leader peptide"
                     /protein_id="YP_001728984.1"
                     /db_xref="GI:170079664"
                     /db_xref="ASAP:AEC-0000073"
                     /db_xref="GeneID:6058969"
                     /translation="MKRISTTITTTITITTGNGAG"
```

FIGURA 1.11 – EXEMPLO DE *FEATURES* E *QUALIFIRES*
FONTE: (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__DH10B_uid58979/NC_010473.gbk, Acessado em 26/01/2012)

```
ORIGIN
        1 agcttttcat tctgactgca acgggcaata tgtctctgtg tggattaaaa aaagagtgtc
       61 tgatagcagc ttctgaactg gttacctgcc gtgagtaaat taaaatttta ttgacttagg
      121 tcactaaata ctttaaccaa tataggcata gcgcacagac agataaaaat tacagagtac
      181 acaacatcca tgaaacgcat tagcaccacc attaccacca ccatcaccat taccacaggt
      241 aacggtgcgg gctgacgcgt acaggaaaca cagaaaaaag cccgcacctg acagtgcggg
      301 ctttttttt cgaccaaagg taacgaggta acaaccatgc gagtgttgaa gttcggcggt
      361 acatcagtgg caaatgcaga acgttttctg cgtgttgccg atattctgga aagcaatgcc
      421 aggcaggggc aggtggccac cgtcctctct gcccccgcca aaatcaccaa ccacctggtg
      481 gcgatgattg aaaaaaccat tagcggccag gatgctttac ccaatatcag cgatgccgaa
      541 cgtatttttg ccgaactttt gacgggactc gccgccgccc agccggggtt cccgctggcg
      601 caattgaaaa ctttcgtcga tcaggaattt gcccaaataa aacatgtcct gcatggcatt
      661 agtttgttgg ggcagtgccc ggatagcatc aacgctgcgc tgatttgccg tggcgagaaa
      721 atgtcgatcg ccattatggc cggcgtatta gaagcgcgcg gtcacaacgt tactgttatc
      781 gatccggtcg aaaaactgct ggcagtgggg cattacctcg aatctaccgt cgatattgct
      841 gagtccaccc gccgtattgc ggcaagccgc attccggctg atcacatggt gctgatggca
      901 ggtttcaccg ccggtaatga aaaaggcgaa ctggtggtgc ttggacgcaa cggttccgac
      961 tactctgctg cggtgctggc tgcctgttta cgcgccgatt gttgcgagat ttggacggac
     1021 gttgacgggg tctatacctg cgacccgcgt caggtgcccg atgcgaggtt gttgaagtcg
     1081 atgtcctacc aggaagcgat ggagctttcc tacttcggcg ctaaagttct tcacccccgc
     1141 accattaccc ccatcgccca gttccagatc ccttgcctga ttaaaaatac cggaaatcct
     1201 caagcaccag gtacgctcat tggtgccagc cgtgatgaag acgaattacc ggtcaagggc
     1261 atttccaatc tgaataacat ggcaatgttc agcgtttctg gtccggggat gaaagggatg
     1321 gtcggcatgg cggcgcgcgt ctttgcagcg atgtcacgcg cccgtatttc cgtggtgctg
     1381 attacgcaat catcttccga atacagcatc agtttctgcg ttccacaaag cgactgtgtg
```

FIGURA 1.12 – EXEMPLO DO SEQUÊNCIA GENÔMICA CONTIDA NA ANOTAÇÃO GENÔMICA EM UM ARQUIVO DA EXTENSÃO GBK
FONTE: (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__DH10B_uid58979/NC_010473.gbk, Acessado em 26/01/2012)
.

1.5 SEQUENCE-INDEXED LOCAL ALIGNER

O *Sequence-Indexed Local Aligner (*SILA), que em português significa Indexador de Sequências de Alinhamento Local (VIALLE, 2011), foi desenvolvido pelo Ricardo Assunção Vialle com a orientação do professor Doutor Roberto Tadeu Raittz, que fazem parte do grupo de pesquisa em Bioinformática da Universidade Federal do Paraná, o qual a autora também faz parte.

É um programa capaz de realizar alinhamentos em sequências de DNA contra um banco de dados. Para o desenvolvimento dessa ferramenta foi utilizada uma técnica conhecida por Indexação Recursiva (INREC) (SOUZA, 1999), que é uma técnica usada em reconhecimento de padrões com a finalidade de reduzir a dimensão dos atributos por meio do cálculo de um único índice para o padrão de características. A ferramenta em questão utiliza essa técnica para indexar as sequências de DNA do banco de dados a ser utilizado. Uma vez que esse banco é indexado, o número de comparações realizadas na busca por similaridade de sequências é reduzido, evitando o alinhamento de todas as sequências do banco de dados. Além disso, essa ferramenta utiliza o algoritmo de Smith-Waterman (SMITH e

WATERMAN, 1981) para alinhar as sequências do banco de dados que são representados por índices encontrados na sequência de consulta.

A principal vantagem da utilização desse programa está na velocidade dos alinhamentos. Em um conjunto de testes realizados para a comparação deste programa com o programa BLAST® utilizou 1000 sequências aleatórias do banco de dados NR e as 4146 sequências da *Escherichia coli* K 12. Para a realização dos alinhamentos via BLAST® utilizou-se duas configurações distintas em relação ao tamanho das palavras, uma padrão, com tamanho 3 e outra customizada, com tamanho 7, seguindo os padrões sugeridos por (SHIRYEV S.A *et al* , 2007). Com isso, o tempo de espera para obter os alinhamentos com as 100 sequências aleatórias do NR foi necessário:

- 18406,2592 segundos (5 horas e 6 minutos) seguindo a configuração padrão do BLAST®;
- 12164,1018 segundos (3 horas e 23 minutos) seguindo a configuração customizada do BLAST®;
- 2373,3 (aproximadamente 40 minutos) segundos com o programa SILA.

E para o conjunto de sequências da *Escherichia coli* K 12:

- 65781,534 segundos (18 horas e 16 minutos) seguindo a configuração padrão do BLAST®;
- 41118,259 segundos (18 horas e 16 minutos) seguindo a configuração customizada do BLAST®;
- 6435,0865 segundos (1 hora e 47 minutos) com o programa SILA.

Esse programa está em fase de desenvolvimento, por isso ainda não está disponível para toda a comunidade científica (VIALLE, R.A. dados não publicados, 2011).

1.6 JUSTIFICATIVA

Um estudo feito por (WARREN *et al*, 2010) demonstrou que um conjunto de genes sem anotação provou ser verdadeiro por serem semelhantes entre si, no entanto esses genes não continham indícios nos bancos de dados de domínio público. Uma justificativa para esse resultado foi que um número elevado dos genes anotados utiliza os mesmos programas para detecção de ORFs e tais programas

não são capazes de detectar com eficiência genes pequenos (WARREN *et al* , 2010), uma vez que boa parte dos genes desse resultado continha menos de 100 pares de base (pb).

Além disso, as anotações genômicas mais antigas contidas nos bancos de dados genômicos de domínio público, que utilizaram genes homólogos como base para a validação das sequências, provavelmente estão desatualizadas. Isso ocorre porque esses bancos de dados são atualizados diariamente e novas sequências genômicas são inseridas frequentemente. De modo que, se uma sequência genômica cuja única forma de validação foi homologia de sequências, ela pode não ter sido válida na época em que foi encontrada. No entanto, com a atualização do banco de dados ela pode ser validada utilizando a mesma técnica de validação de sequencias.

Com o desenvolvimento da ferramenta HGF para detecção de genes cujo diferencial é a detecção de genes pequenos tornou-se necessário o desenvolvimento de uma ferramenta capaz de facilitar a comparação da eficiência do HGF em relação aos genes já anotados e, assim, comprovar a eficiência da ferramenta HGF.

1.7 OBJETIVOS

1.7.1 Objetivo Geral

Desenvolver uma ferramenta de bioinformática que visa encontrar automaticamente novos genes em um genoma anotado.

1.7.2 Objetivos Específicos

- Testar e utilizar o *Hybrid-Gene Finder* (HGF) – um programa de predição de sequências regiões codificantes de proteínas (CDS) em procariotos;
- Comparar as predições de CDS do HGF com os genomas anotados disponíveis em bancos de dados públicos internacionais;

- Identificar e validar os novos genes descobertos utilizando alinhamento de sequências com os programas BlastP e SILA;
- Desenvolver uma ferramenta – *Bioinformatics tool based on bacterial genomes comparison* (BOBBLES) - para sistematizar e automatiza as etapas anteriores utilizando o Matlab®;
- Realizar estudos de caso de uma lista de genomas completos pré-selecionados;
- Avaliar o desempenho da ferramenta desenvolvida, comparando o desempenho utilizando os programas BlastP e SILA.

2 MATERIAIS E METODOLOGIA

Nesta sessão serão apresentados os materiais e as estratégias utilizadas para a comparação e validação do resultado do *Hybrid-Gene Finder* (HGF) com os genomas bacterianos completos anotados e depositados no GenBank® através do programa *Bioinformatics Tool Based on Bacterial Genomes Comparison* (BOBBLES).

2.1 HYBRID-GENE FINDER

Neste trabalho, o *Hybrid-Gene Finder* (HGF) foi utilizado para realizar a predição de prováveis genes através da marcação de sequência de região codificante (CDS) em genomas bacterianos completos já anotados e disponíveis no GenBank®.

2.2 BANCO DE DADOS

O banco de dados utilizado para o desenvolvimento e testes de alinhamento das sequências *query* foi o banco de dados de sequências de proteínas não redundantes (NR), localizado no NCBI (http://www.ncbi.nlm.nih.gov).

2.3 BLASTP

As pesquisas de alinhamento das sequências *query* com as sequências alvo do banco de dados NR foram realizadas utilizando a ferramenta Protein Blast (BlastP). Que é um programa pertencente ao *Basic Alignment Search Tool* (BLAST), o qual faz pesquisas por proteínas no banco de dados de sequências de proteínas.

A versão utilizada foi a BLASTP 2.2.26+ e a configuração foi a padrão dessa ferramenta, ou seja:

- Banco de Dados do NCBI (*NCBI Database)*: NR;

- Matriz (*Matrix*): BLOSUM62;
- Número de acessos para manter (*Number of hits to keep*): 500;
- Filtro (*Filter*): nenhum (*none*);
- Serviço Blast (*Blast service*): simples (*plain*).

2.4 SILA

O programa *Sequence-Indexed Local Aligner* (SILA) foi utilizado neste trabalho para a realização dos alinhamentos de sequências, como alternativa ao programa BlastP com o intuito de testar essa ferramenta e conferir se os resultados são semelhantes ao programa BlastP, porém mais rápido.

2.5 MATLAB®

O aplicativo, os *scripts* de desenvolvimento e testes deste projeto foram desenvolvidos utilizando a linguagem Matlab (*Matrix Laboratory*) no ambiente de desenvolvimento Matlab, na versão 2010a.

2.6 ARTEMIS®

O programa Artemis® (RUTHERFORD *et al*, 2000) foi utilizado para auxiliar na comparação manual dos genes, por ser um visualizador de genomas com atalhos para acessar ao programa BlastP. Neste projeto foi utilizada a versão 12.0.

2.7 GENOMAS BACTERIANOS

Foram utilizados 14 genomas bacterianos completos para base para testar a ferramenta BOBBLES, proposta em 1.9.1. Eles foram retirados do banco de dados GenBank® do NCBI e serviram como genomas de referência, são eles: *Bradyrhizobium japonicum* USDA 110, *Burkholderia mallei* SAVP1, *Cyclobacterium*

marinum DSM, *Escherichia coli* K 12 substr DH10B, *Herbaspirillum seropedicae* SmR1, *Methanocaldococcus fervens* AG86, *Ralstonia solanacearum* CFBP2957, *Streptococcus agalactiae* NEM316, *Streptococcus mutans* UA159, *Streptococcus pneumoniae* Hungary19A 6, *Treponema pallidum* Nichols, *Pseudomonas fluorescens* Pf-5, *Thermotoga maritima* MSB8 e *Treponema denticola* ATCC 35405. E esses mesmos genomas foram submetidos ao programa HGF para obter-se uma nova marcação dos genes.

O genoma *Escherichia coli* K 12, foi escolhido por ser um organismo modelo. O genoma *Herbaspirillum seropedicae* SmR1 foi escolhido por ser um organismo seqüenciado, montado e anotado pelo grupo de pesquisa da UFPR e, por isso, existe grande interesse em detectar possíveis falhas de anotação para corrigir a anotação. E os demais genomas foram escolhidos por apresentarem maior índice de variabilidade em testes realizados entre o HGF e o programa Glimmer® (DELCHER, 2007) pelo grupo de pesquisa em Bioinformática da UFPR.

A TABELA 1 exibe a classificação taxonomica de cada um desses genomas. Ela mostra a diversidade taxonômica deles, por conter proteobacteria, bacteioidetes, euryarchaeotes, firmicutes, thermotogales e spirochetes.

TABELA 1 - CLASSIFICAÇÃO TAXONÔMICA DOS GENOMAS BACTERIANOS UTILIZADOS NOS TESTES DO PROGRAMA BOBBLES

GENOMA BACTERIANO	CLASSIFICAÇÃO TAXONÔMICA
Bradyrhizobium japonicum USDA 110	cellular organisms; Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Bradyrhizobiaceae; Bradyrhizobium; Bradyrhizobium japonicum
Burkholderia mallei SAVP1	cellular organisms; Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales; Burkholderiaceae; Burkholderia; pseudomallei group; Burkholderia mallei
Cyclobacterium marinum DSM 745	cellular organisms; Bacteria; Bacteroidetes/Chlorobi group; Bacteroidetes; Cytophagia; Cytophagales; Cyclobacteriaceae; Cyclobacterium; Cyclobacterium marinum
Escherichia coli K 12 substr DH10B	cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia; Escherichia coli; Escherichia coli O17
Herbaspirillum seropedicae SmR1	cellular organisms; Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales; Oxalobacteraceae; Herbaspirillum; Herbaspirillum Seropedicae
Methanocaldococcus fervens AG86	cellular organisms; Archaea; Euryarchaeota; Methanococci; Methanococcales; Methanocaldococcaceae; Methanocaldococcus; Methanocaldococcus fervens
Pseudomonas fluorescens Pf-5	cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas fluorescens group; Pseudomonas fluorescens
Ralstonia solanacearum CFBP2957	cellular organisms; Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales; Burkholderiaceae; Ralstonia; Ralstonia solanacearum
Streptococcus agalactiae NEM316	cellular organisms; Bacteria; Firmicutes; Bacilli; Lactobacillales; Streptococcaceae; Streptococcus; Streptococcus agalactiae; Streptococcus agalactiae serogroup III
Streptococcus mutans UA159	cellular organisms; Bacteria; Firmicutes; Bacilli; Lactobacillales; Streptococcaceae; Streptococcus; Streptococcus mutans
Streptococcus pneumoniae Hungary19A 6	cellular organisms; Bacteria; Firmicutes; Bacilli; Lactobacillales; Streptococcaceae; Streptococcus; Streptococcus pneumoniae
Thermotoga maritima MSB8	cellular organisms; Bacteria; Thermotogae; Thermotogae (class); Thermotogales; Thermotogaceae; Thermotoga; Thermotoga maritime
Treponema denticola ATCC 35405	cellular organisms; Bacteria; Spirochaetes; Spirochaetia; Spirochaetales; Spirochaetaceae; Treponema; Treponema denticola
Treponema pallidum Nichols	cellular organisms; Bacteria; Spirochaetes; Spirochaetia; Spirochaetales; Spirochaetaceae; Treponema; Treponema pallidum; Treponema pallidum subsp. Pallidum

FONTE: (NCBI, 2012)

A TABELA 2 mostra o número de pares de base (pb) em cada genoma e o conteúdo de GC. Como mostra nesta tabela, tanto o tamanho dos genomas quanto ao conteúdo de GC variam de 22190pb até 9105828pb e de 32,2% até 68,4%, respectivamente. Assim, a diversidade tanto em tamanho quanto em conteúdo de GC podem ser testadas para avaliar a capacidade do programa HGF de encontrar os genes e poderem ser conferidos pelo programa BOBBLES.

TABELA 2 – NÚMERO DE pb E CONTEÚDO DE GC DOS GENOMAS BACTERIANOS UTILIZADOS NOS TESTES DO PROGRAMA BOBBLES

GENOMA BACTERIANO	TAMANHO EM pb	CONTEÚDO DE GC (%)
Bradyrhizobium japonicum USDA 110	9105828	64,1
Burkholderia mallei SAVP1	3497479	68,4
Cyclobacterium marinum DSM 745	6221273	38,1
Escherichia coli K 12 substr DH10B	4686137	50,8
Herbaspirillum seropedicae SmR1	5513887	63,4
Methanocaldococcus fervens AG86	22190	32,2
Pseudomonas fluorescens Pf-5	7074893	63,3
Ralstonia solanacearum CFBP2957	3417386	66,5
Streptococcus agalactiae NEM316	2211485	35,6
Streptococcus mutans UA159	2032925	36,8
Streptococcus pneumoniae Hungary19A 6	2245615	39,6
Thermotoga maritima MSB8	1860725	46,2
Treponema denticola ATCC 35405	2843201	37,9
Treponema pallidum Nichols	1138011	52,8

FONTE: (NCBI, 2012)

2.8 BIBLIOTECAS

As bibliotecas utilizadas para o desenvolvimento foram:

- Bibliotecas do Matlab®: Bioinformatics e Matlab;
- Biblioteca desenvolvida pelo laboratório de Bioinformática da Universidade Federal do Paraná (UFPR).

2.9 PERIFÉRICOS

Os periféricos utilizados para o desenvolvimento e testes deste projeto foram:

- Fabricante Dell:
  - Modelo: Studio 1450 (notebook);
  - Processador: Pentium Dual-Core CPU T4400 @2.20GHz 2.20GHz;
  - Memória: 4GB;
  - Sistema Operacional: Windows Seven;

- Tipo de Sistema Operacional: 64*bits*.

- Fabricante Lenovo:
  - Modelo ThinkCentre M90P(desktop);
  - Processador: Intel Core i5 CPU 650 @ 3.20GHz x4;
  - Memória: 8GB;
  - Sistema Operacional: Ubuntu 11.10;
  - Tipo de Sistema Operacional: 64*bits*.

2.10 METODOLOGIA GERAL

A metodologia, esquematizada na FIGURA 2.1, foi divida em duas fases: manual e automatizada. Ambas realizaram a comparação entre um determinado genoma completo de bactéria, um disponibilizado pelo banco de dados GenBank® do NCBI, o qual será chamado de “arquivo GenBank” e o outro produzido através do HGF, o qual será chamado de “arquivo HGF”.

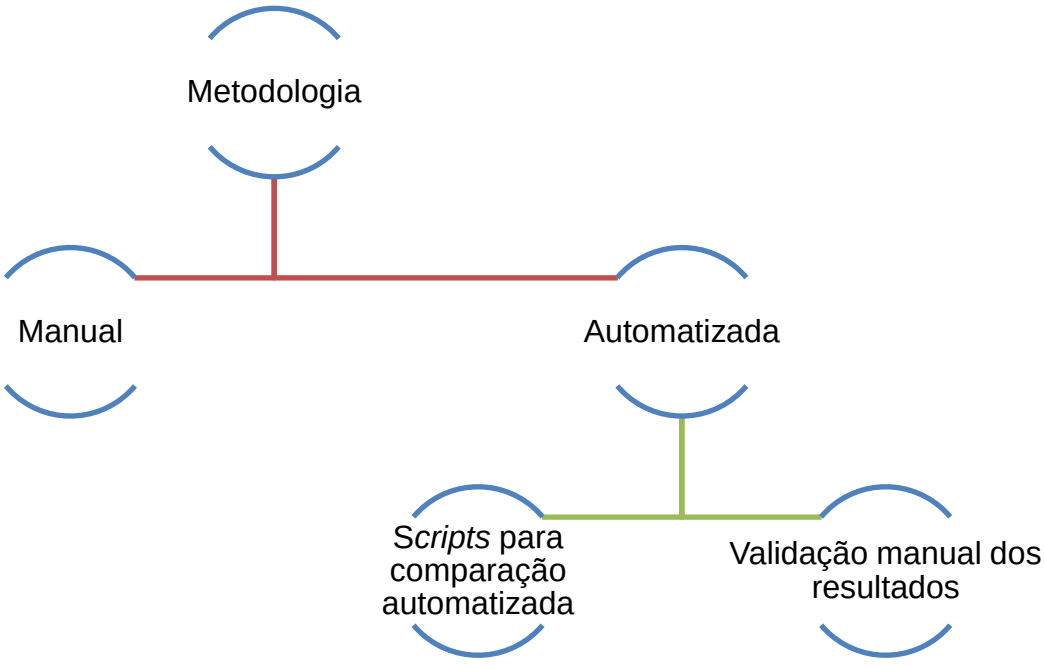


FIGURA 2.1 - MODELO GERAL DA METODOLOGIA DE COMPARAÇÃO DOS GENOMAS
FONTE: O autor (2012)

A fase manual consistiu em comparar alguns genomas do grupo de genomas de teste. Portanto, cada um dos genes alvo foram analisados um a um utilizando o visualizador de genomas Artemis® e o programa BlastP. Dessa forma

puderam-se entender quais as necessidades que a próxima fase, chamada de automatizada, deveria solucionar.

A fase automatizada foi dividida em duas etapas: Testes com *scripts* através do Matlab® e a validação desses testes manualmente. Esses *scripts* foram criados com o objetivo de realizar as mesmas tarefas da fase manual e a etapa de validação consistiu em analisar os resultados obtidos na etapa anterior.

2.10.1 Comparação manual dos genomas

Nessa fase, os arquivos são comparados manualmente utilizando o programa Artemis®. Para isso, ambos os arquivos são abertos em uma mesma camada desse programa. A FIGURA 2.2 mostra os genes do “arquivo GenBank”, em azul, e os genes do “arquivo HGF”, em vermelho, rosa e cinza. Eles estão abertos em uma mesma camada do programa Artemis®, e dispostos nas seis fases de leitura que estão destacadas pelas linhas pretas na lateral esquerda da figura.

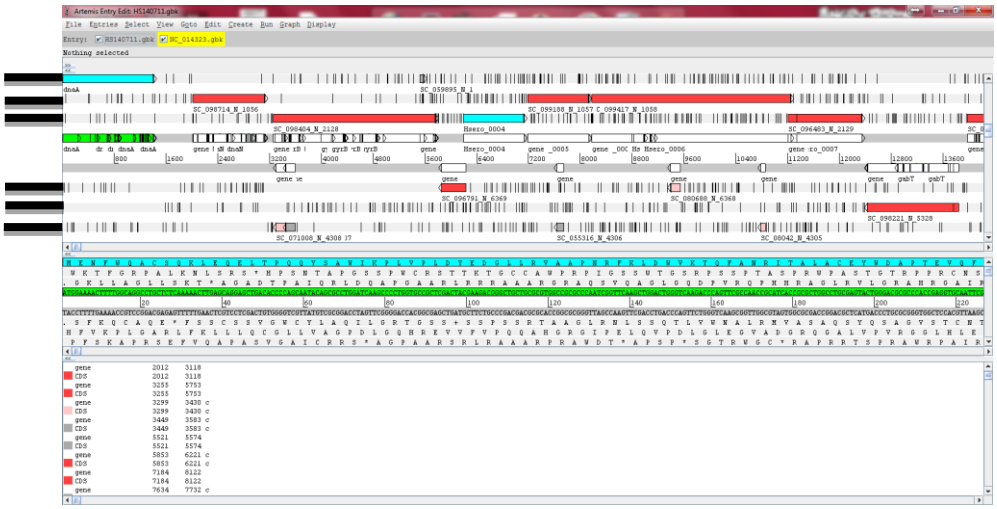


FIGURA 2.2 - PROGRAMA ARTEMIS® MOSTRANDO O “ARQUIVO GENBANK”, EM AZUL, E O “ARQUIVO HGF”, EM VERMELHO, ROSA E CINZA
FONTE: O autor (2012)

Nesse caso foram avaliadas todos os genes novos encontrados pelo programa HGF e todos aqueles cujas fases de leitura do “arquivo HGF” estejam divergendo em uma fase de leitura do “arquivo GenBank”, ou seja quando uma sequência foi anotada em um arquivo numa fase de leitura e pelo outro arquivo em

na fase de leitura diferente. Cada um deles foi submetido ao programa BlastP para encontrar as sequências homólogas. Foram anotadas em uma planilha eletrônica, para posteriormente serem avaliadas por um especialista, as sequências as quais apresentaram resultados cujo valor de *bit score* fosse superior a 80 ou mais do que seis sequências com algum grau de similaridade. Esse valor de corte com o valor de *bit score* superior a 80 foi escolhido por estar presente na faixa de valores de *scores* da plataforma BLAST, mostrado em 1.2.2.2, em que o valor de *bit score* de 100, sugerida por (LIBERMAN, 2004).

2.10.2 Comparação automatizada dos genomas

A segunda fase da metodologia teve o objetivo de automatizar o processo da primeira fase e, para isso, foi dividida em duas etapas. Na primeira, os arquivos foram processados através de *scripts* no Matlab® com o intuito de obter todos os genes candidatos a serem verdadeiros. O resultado dessa etapa foi um arquivo na extensão GBK, também conhecido como formato GFFF ou formato GenBank, contendo todos esses candidatos. E, na segunda etapa, o arquivo gerado na etapa anterior foi validado manualmente utilizando o programa Artemis® para a visualização dos genes e submetidos ao programa BlastP para conferir se o resultado está correto. Dessa forma pode-se analisar a eficiência do *script*.

2.10.2.1 Estratégia de identificação dos genes sobrepostos nos arquivos dos genomas

Na primeira etapa, da segunda fase, cada gene é detectado através das sequências de regiões codificantes (CDS) contidos no arquivo da extensão GBK referente ao genoma bacteriano estudado. Assim, além da localização, é possível saber a orientação deles. Essas localizações são divididas de acordo com a orientação, ou seja, 5’ para 3’ e 3’ para 5’ em cada um dos arquivos, como mostram os esquemas da FIGURA 2.3.

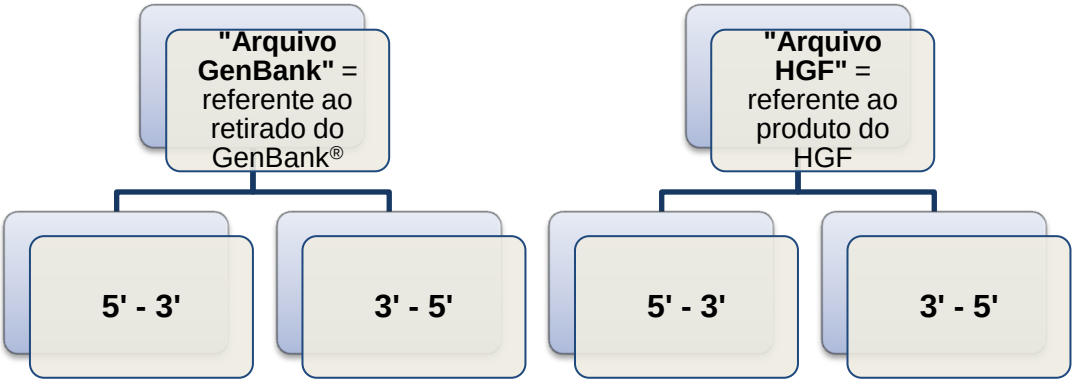


FIGURA 2.3 - ESQUEMAS DA DIVISÃO DOS LOCAIS DE CDS (SEQUÊNCIA DE REGIÃO CODIFICANTE) NOS ARQUIVOS GENBANK E HGF.
FONTE: O autor (2012)

Foram considerados genes idênticos somente os possuíssem localizações análogas de *stop códon*, podendo ou não ter o mesmo *start códon*. Isso porque não é o objetivo do HGF localizar o exato o *start códon* dos genes. No “arquivo HGF”, são retirados esses genes idênticos para não gerar redundância de dados. Em seguida, foram localizados todos os genes com divergência de fase de leitura. Isso acontece quando, em fases de leitura diferentes, existem dois ou mais genes paralelos. Esses genes podem ser menores do que o outro gene nas duas extremidades da sequência genômica, FIGURA 2.4 (A), maiores do que o outro gene nas duas extremidades da sequência genômica, FIGURA 2.4 (B), ou paralela em fase de leitura apenas em uma das pontas do genoma, FIGURA 2.4 (C).

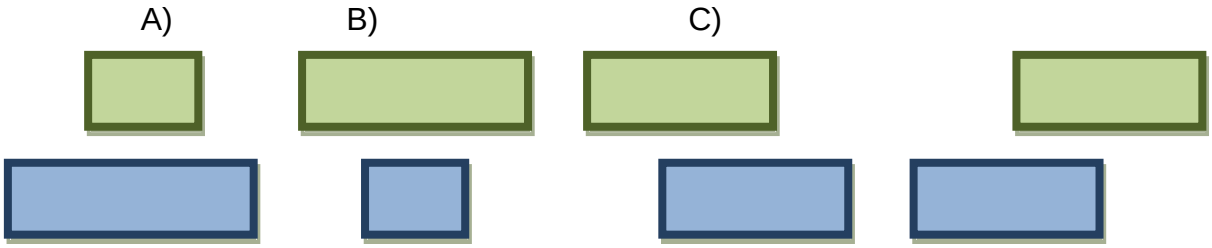


FIGURA 2.4 - TIPOS DE LOCAIS ONDE PODEM OCORRER CONFLITO DE FASE DE LEITURA ENTRE OS ARQUIVOS GENBANK E HGF. A) SEQUÊNCIA MENOR NAS DUAS EXTREMIDADES DA SEQUÊNCIA DO QUE A INFERIOR; B) SEQUÊNCIA MAIOR NAS DUAS EXTREMIDADES DA SEQUÊNCIA DO QUE A INFERIOR; C) PARALELA EM APENAS UMA DAS PONTAS DO GENE

FONTE: O autor (2012)

Portanto, foram comparados:

• Sequências do sentido 5'-3' do “arquivo GenBank” com as sequências do sentido 5'-3' do “arquivo HGF”,

• Sequências do sentido 5'-3' do “arquivo GenBank” com as sequências do sentido 3'-5' do “arquivo HGF”,

• Sequências do sentido 3'-5' do “arquivo GenBank” com as sequências do sentido 5'-3' do “arquivo HGF” e

• Sequências do sentido 3'-5' do “arquivo GenBank” com as sequências do sentido 3'-5 do “arquivo HGF”.

2.10.2.2 Estratégia de análise dos genes sobrepostos realizado pelo programa BOBBLES

Depois de identificadas todas as sequências conflitantes entre os dois arquivos, através do método explicado em 2.10.2.1, cada sequência de gene com divergência em fase de leitura é analisada. Para isso, ambas as sequências concorrentes foram submetidas ao programa BlastP ou ao programa SILA para serem comparados contra o banco de dados NR, FIGURA 2.5.

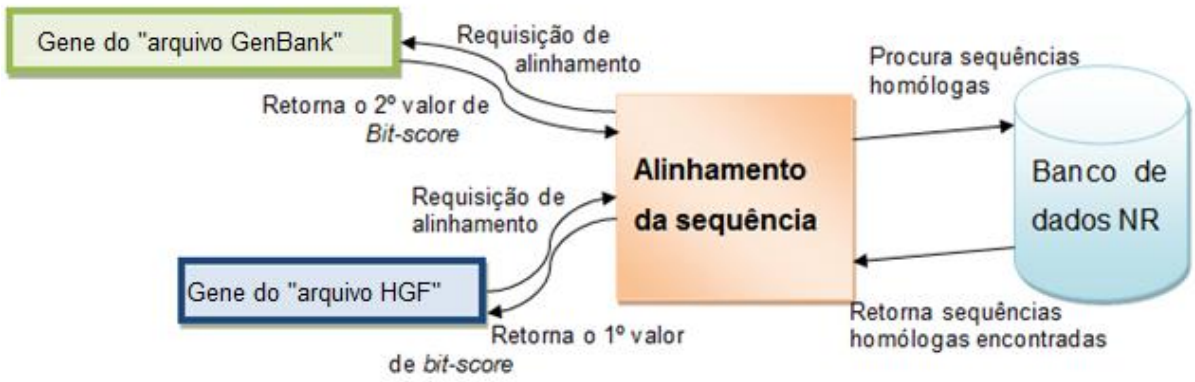


FIGURA 2.5 - REPRESENTAÇÃO DO ALGORITMO DE COMPARAÇÃO DOS GENES CONCORRENTES

A avaliação das sequencias de genes com divergência em fase de leitura consistiu em comparar o segundo maior valor de *bit score* de retorno do “arquivo GenBank” com o primeiro maior valor de *bit score* do “arquivo HGF”. Optou-se descartar o maior valor do “arquivo GenBank” porque ele consiste no resultado do

alinhamento dessa sequência contra ela mesma, podendo causar inconsistência com as demais respostas, o que não ocorre com outro arquivo. Com esses resultados, o algoritmo verifica qual deles é maior. Se o valor do “arquivo HGF” fosse superior ele o manteria na lista com os candidatos a genes verdadeiros, senão essa sequência seria descartada. Quando a sequência do “arquivo HGF” cujo valor de *bit score* fosse superior a 80, independente se o valor do “arquivo GenBank”, elas também seriam inseridas nessa lista de candidatos a genes verdadeiros. Dessa forma, foi possível encontrar os genes cujas sobreposições ocorressem apenas nas pontas dos genes, o que não seria possível se fosse atribuído um valor de corte único porque as sequências possuem tamanhos diferentes e uma quantidade de bases pode ser significativa em determinadas sequências e em outra não. Outro motivo para isso foi uma segunda avaliação manual da qualidade de ambas as sequências.

2.10.2.3 Estratégia de análise dos novos genes contidos no “arquivo HGF” realizado pelo programa BOBBLES

Além dos genes concorrentes, o “arquivo HGF” apresentou novos genes, ou seja, genes que estão presentes no “arquivo HGF” que não estão contidos no “arquivo GenBank”. A FIGURA 2.6 mostra um exemplo desses genes encontrados pelo HGF e vistos através do programa Artemis®, eles são os que estão contornados em preto na imagem. O “arquivo GenBank” também apresentou genes os quais não continham no outro genoma. No entanto, o foco dessa pesquisa é identificar novos genes e não validar novamente os que já estavam anotados, por isso eles não foram avaliados.

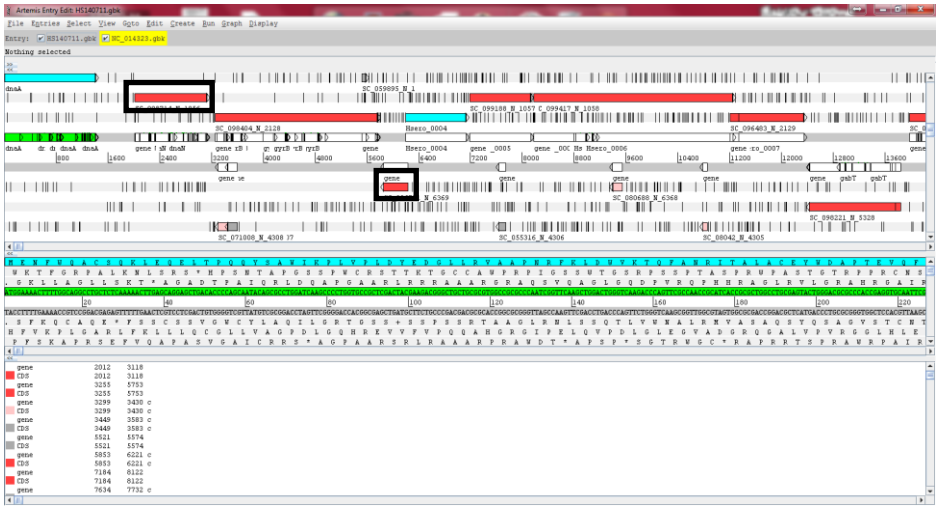


FIGURA 2.6 – EXEMPLO DE NOVOS GENES ENCONTRADOS PELO PROGRAMA HGF E VISUALIZADO ATRAVÉS DO PROGRAMA ARTEMIS®
FONTE: O autor (2012)

Para identificar esses novos genes foram retirados do “arquivo HGF” todos os casos com conflito de fase de leitura e todos os genes classificados como idênticos ao “arquivo GenBank”, FIGURA 2.7. Ou seja, a função dessa detecção foi constituída através:

- Entrada dos dois arquivos do genoma, na extensão GBK;
- Entrada de todas as sequências do “arquivo GenBank” e “arquivo HGF”, explicado em 2.10.2.1;
- Obtenção das sequências alvo através do cálculo: $R = A - B$, onde R representa todas as sequências novas contidas no “arquivo HGF”, A representa todas ORFs contidas no “arquivo HGF” e B o cálculo:

$$B = \text{ORFs com divergência} + \text{ORFs consideradas idênticas};$$

- Alinhamento das sequências alvo, R, contra o banco de dados NR;
- Caso o valor de *bit score* for maior do que 80 a posição da sequência será anotada em uma lista de sequências novas prováveis verdadeiras.

Dessa forma, todas as sequências com chances de serem novos genes foram encontradas. Optou-se por descartar aquelas cujo valor fosse inferior a 80, assim somente aquelas com grandes chances de serem verdadeiras foram gravadas no arquivo de genes para serem avaliados na etapa de validação manual dos resultados.

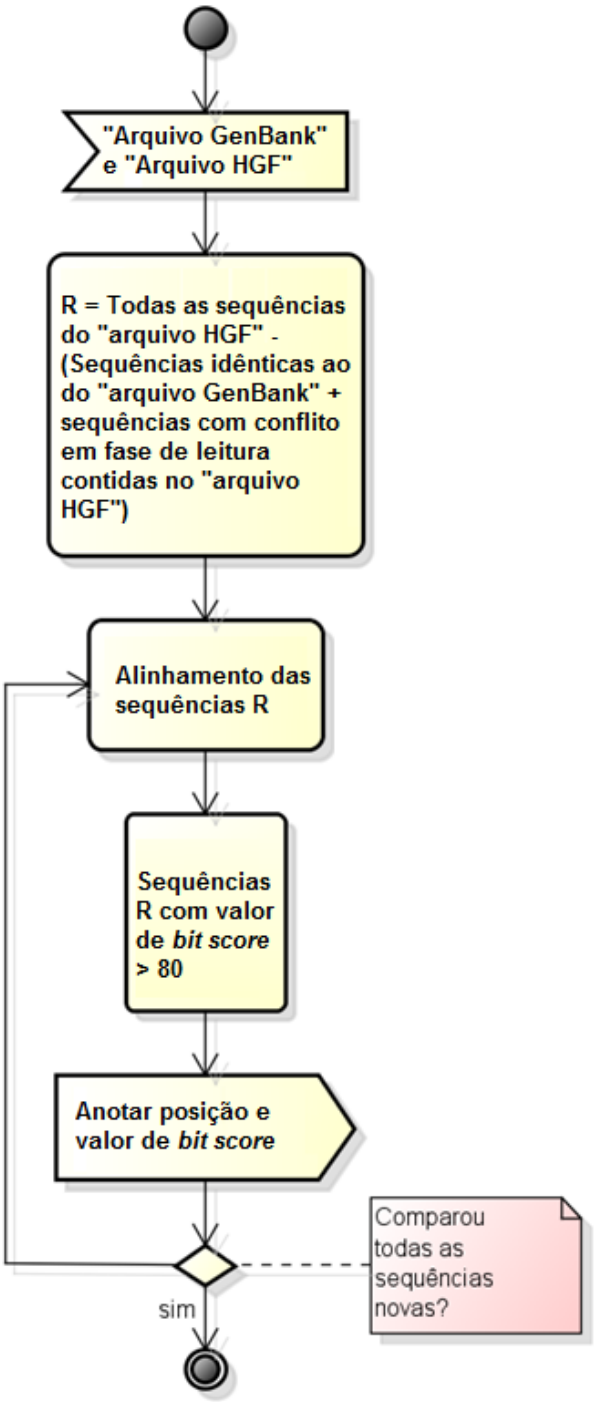


FIGURA 2.7 – FLUXOGRAMA DE IDENTIFICAÇÃO DE NOVOS GENES CONTIDOS NO “ARQUIVO HGF”
FONTE: O autor (2012)

2.10.2.4 Alinhamento das sequências através do programa BlastP

O alinhamento das sequências, tanto nas sequências concorrentes quanto nas novas, foram um ponto chave no desenvolvimento do programa BOBBLES. Por

isso, foram realizados vários testes utilizando o programa BlastP, tanto o executável local quanto por acesso remoto.

Foram realizados testes do funcionamento do programa BlastP local tanto no sistema operacional Seven, do Windows®, quanto no Ubuntu 10.11, distribuição Linux. Em ambos os casos foram testados a usabilidade através do terminal com chamada da função no Matlab®, Shell no Ubuntu e Prompt de Comando no Windows®, e através das funções da biblioteca Bioinformatics do Matlab®. Em todos os casos foi necessário obter a última versão da base de dados de sequências de proteínas não redundantes (NR) e montar o banco de dados, que via terminal foi utilizado o seguinte comando:

```
$makeblastdb –in nr –dbtype prot
```

Onde, `nr` é o nome do banco de dados e `prot` é o tipo de banco de dados, no caso banco de dados de proteínas. Para montar o banco dentro da plataforma Matlab® foi utilizada a seguinte função (RAITTZ, R.T. dados não publicados, 2011):

```
criaBDparaBlastp(nr);
```

Onde, `nr` é o arquivo contendo as sequências que estarão no banco de dados. Para a execução do alinhamento de cada sequência executada localmente através de terminal Shell foi elaborado um *script* contendo a seguinte função:

```
./bin/blastp -query $* -db ./db/nr -outfmt 6 | awk 'NR==1 {print $12 }' > result.txt
```

Onde, `./bin/blastp` é a chamada para executar o programa BlastP e os parâmetros na frente dele serão utilizados por ele, `$*` é o arquivo com a sequência para o alinhamento, `./db/nr` indica o local e o nome do banco de dados, `–outfmt 6` indicando a escolha do formato tabular de saída dos resultados do alinhamento, (MORGULIS, 2008), e `| awk 'NR==1 {print $12 }' > result.txt` para adicionar o primeiro valor de *bit score* no arquivo `result.txt` E para o alinhamento dentro da plataforma Matlab® foi utilizada a seguinte função (RAITTZ, R.T. dados não publicados, 2011):

```
blastpseqbd(seq,nr);
```

Onde, `seq` é a sequência que irá para alinhamento contra o banco de dados `nr`. Para a utilização do BlastP remoto, ou seja, fazendo solicitação via internet para o BLAST® online[1], na plataforma Matlab® foi utilizada as seguintes funções da biblioteca Bioinformatics do Matlab®:

```
arqBlast = nt2aa(setGene);
RID = blastncbi(arqBlast, 'blastp');
blast = getblast(RID, 'WAITTIME', 2);
```

Onde `arqBlast` recebe a sequência do gene em formato de nucleotídeos e a transforma em sequência aminoácido, `RID` envia essa sequência para o BLAST® para realizar o alinhamento da sequência contra o banco de dados NR utilizando o programa BlastP e, `blast` recebe o retorno do BLAST® contendo o resultado do alinhamento. Na versão final do BOBBLES foi utilizado essa opção de alinhamento das sequências por apresentar melhor desempenho em relação ao programa BlastP executado localmente, utilizando os periféricos descritos em 2.9.

2.10.2.5 Alinhamento das sequências através do programa SILA

O programa SILA realiza o alinhamento das sequências de forma local, ou seja, não existe necessidade de se estabelecer conexão com outro computador ou na internet. Para executá-lo foi necessário carregar o arquivo `dadosindice.mat` que é um arquivo resultado de uma rede neuronal previamente treinada contendo os índices necessários para a execução desse programa. No Matlab® esse arquivo é carregado com o comando:

```
load dadosindice.mat
```

[1] Link de acesso do BLAST® online para executar o programa BlastP: http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&BLAST_SPEC=&LINK_LOC=blasttab&LAST_PAGE=blastn

Para a execução do alinhamento da sequência foram utilizadas as seguintes funções pertencentes à biblioteca Bioinformatics do Matlab® e ao programa SILA, respectivamente:

```
arqBlast = nt2aa(setGene);
blastAuxRef =  getalignindx(arqBlast, dadosindice, 1);
```

Onde, `arqBlast` recebe a sequência do gene (`setGene`) em formato de nucleotídeos e a transforma em sequência aminoácido, e `blastAuxRef` recebe o resultado do alinhamento dessa sequência.

2.10.2.6 Estratégia de execução do programa BOBBLES para encontrar os genes alvo

Depois de concluídas as duas formas de pesquisa dos genes, tanto as sequências concorrentes do “arquivo HGF” com o “arquivo GenBank” quanto as sequências novas contidas no “arquivo HGF” foram elaboradas funções para validação dessas sequências em uma interface amigável para facilitar a utilização do BOBBLES. Essa interface possui as opções de alinhamento das sequências utilizando o programa BlastP ou o programa SILA. Tanto essas funções quanto a interface foram elaboradas utilizando a plataforma Matlab®.

A FIGURA 2.8 mostra a visão geral do funcionamento do programa BOBBLES, onde:

1. São inseridos os arquivos a serem comparados, ou seja, o arquivo do genoma marcado pelo HGF e o do genoma de referência, ambos na extensão GBK;
2. Escolher entre realizar a comparação utilizando o alinhamento das sequências pelo programa SILA ou pelo programa BlastP;
3. Processos das sequências, utilizando o programa SILA ou o programa BlastP;
4. Arquivo na extensão GBK contendo os potenciais novos genes.

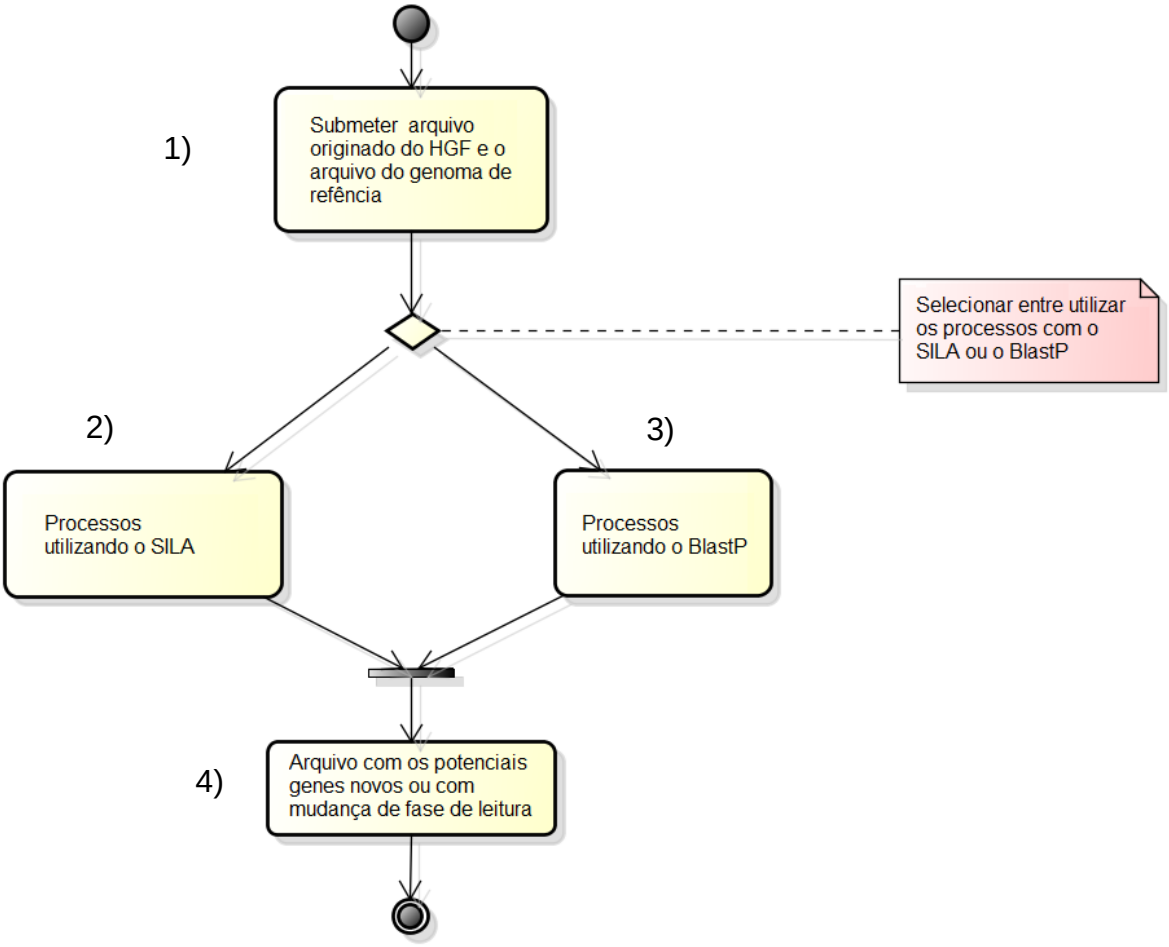


FIGURA 2.8 - VISÃO GERAL DO ALGORITMO DO PROGRAMA BOBBLES
FONTE: O autor (2012)

Esses processos, utilizando o programa SILA ou o programa BlastP, são executados de forma semelhante, diferenciando-se apenas no programa utilizado para o alinhamento das sequências alvo, ou seja as que estão com divergência de fase de leitura e as novas.

2.10.2.7 Interface e execução do programa BOBBLES para encontrar os genes alvo

Optou-se pelo desenvolvimento de uma interface amigável ao usuário, que no caso é o pesquisador o qual irá comparar dois arquivos de genomas para descobrir onde existem divergência em fase de leitura e novos genes. Essa interface, FIGURA 2.9, possui o mínimo de botões possíveis e não é necessário entrar em mais de uma tela do programa para executá-lo. Na parte superior do

programa estão os botões referentes à aplicação (*Aplication*) e os referentes ao contato e manual de instrução, contidos no botão *Help*. Abaixo está o botão de atalho para o programa HGF, FIGURA 2.10, seguido dos botões de inserção: do “arquivo GenBank” (primeiro botão *Search*), do “arquivo HGF” (segundo botão *Search*) e do nome do arquivo de saída (terceiro botão *Search*). No lado esquerdo de cada botão *Search* existe um campo que mostrará o caminho da localização do arquivo selecionado. Abaixo, estão os botões SILA e BLAST, ao usuário escolher o botão SILA o programa será executado e o alinhamento das sequências para validação dos dados será feito utilizando o programa SILA. Se for escolhido o botão BLAST o programa será executado e o alinhamento das sequências para validação dos dados será feito utilizando o programa BlastP. O próximo item desta figura é o campo *Status* o qual serve para mostrar o que está ocorrendo no programa seguido pelo botão *Exit* que serve para fechar o programa.

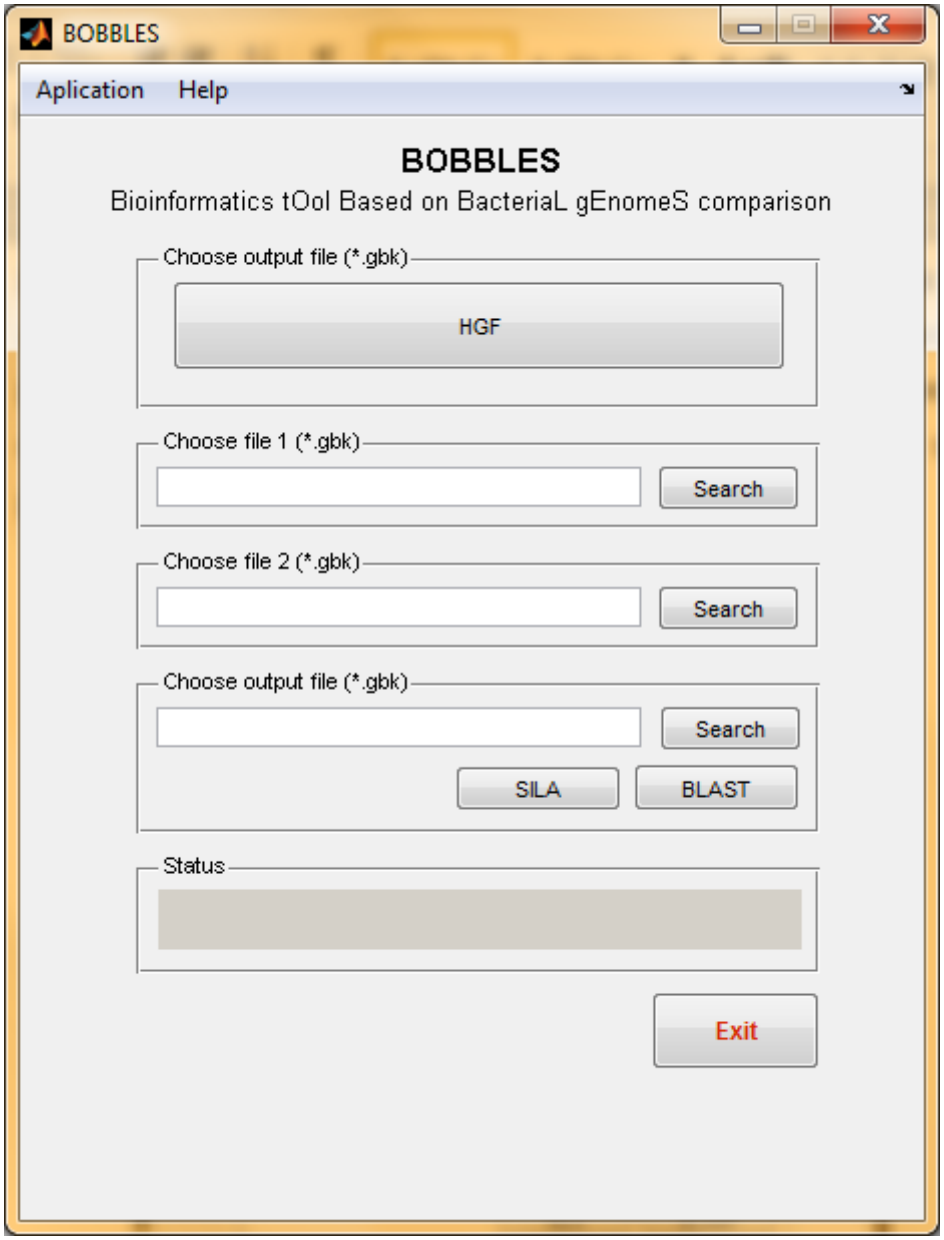


FIGURA 2.9 - INTERFACE DO PROGRAMA BOBBLES
FONTE: O autor (2012)

Além disso, todos os itens foram escritos na língua inglesa para facilitar o entendimento das suas funções por qualquer pesquisador, independente da nacionalidade. Outro ponto importante é que ele segue os mesmos padrões de interface do programa HGF, FIGURA 2.10, fazendo com que o usuário não precise entender mais de um tipo de interface.

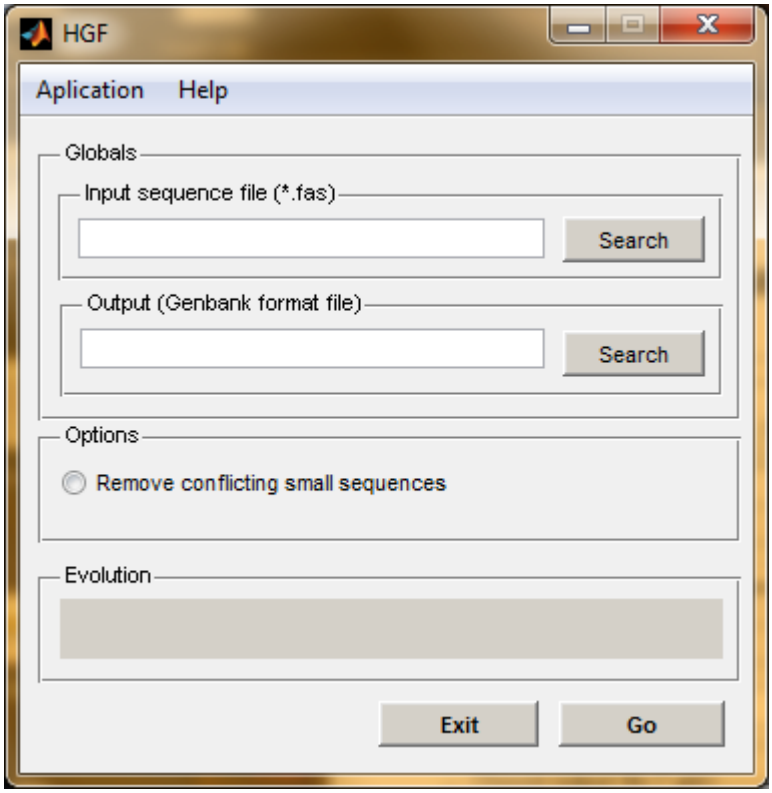


FIGURA 2.10 - INTERFACE DO PROGRAMA HGF
FONTE: O autor (2012)

Após a execução do programa, é gerado um arquivo na extensão GBK, no formato GFFF, FIGURA 2.11, com nome e local escolhidos pelo usuário e uma mensagem de término é exibida no campo *Status*.

```
     gene            8445188..8445574
     CDS             8445188..8445574
                     /gene="SC_-897_N_206"
                     /color=6
     gene            9041530..9041760
     CDS             9041530..9041760
                     /gene="SC_-145_N_207"
                     /color=6
     gene            9103397..9103639
     CDS             9103397..9103639
                     /gene="SC_-132_N_208"
                     /color=6
     gene            complement(5742815..5743459)
     CDS             complement(5742815..5743459)
                     /gene="SC_-142_N_209"
                     /color=6
     gene            complement(7285423..7285656)
     CDS             complement(7285423..7285656)
                     /gene="SC_-124_N_210"
                     /color=6
ORIGIN
        1 TTAATTAATA GTCTTTGACT GCAATACTGG GCGATATGAT CCGGAAGCGC CGATAGCGGC
       61 GCCCGCTTGG CTTCTTCCCG TACTCGCGCT CCCACGTCGA AGTGTCAATC CGCACATCCG
      121 TCATGTTCGC TGATCGCCTC AAAGATTACA ACCTTGCCCT AGCGACCGTG CTTCAGAGCG
      181 TCAATTCCTT CGAGCTGGTC GGGGTTGGGC TCGTCCTTGA TGTTCCAACG GATATCGCAA
      241 GCCGCCGAAT CTTCTCCAAT CGAACAAGCG CCTCTTGGAG TAACGAGGCT CGGAATTTGC
      301 AAGTCGTCGG CGCGGGGTGG CGCTGCTCAG CCGTCCTTCG GAAATCCCTC CGTTCAATGC
      361 AGAAACACCA TAGCGGCGGC ACGCCTGACA GCTCTTTATT TCCGTCAGCT ACAAATTCTT
      421 GTAGTTGACC CGCGCGGTGT CCTCCGGGCT GAATCGCGAT CAACCCCACA GGAGAGCAAA
```

FIGURA 2.11 - EXEMPLO DO ARQUIVO DE SAÍDA DO PROGRAMA BOBBLES VISTO NO FORMATO TEXTO
FONTE: O autor (2012)

2.10.2.8 Validação manual dos resultados obtidos pelo programa BOBBLES

Depois de obtidos todas as sequências com divergência e novas no “arquivo HGF” com potencial a serem verdadeiras através de um arquivo na extensão GBK, foi necessário conferir todas as sequências e descobrir qual delas era ou não falso-positivo. Essa análise foi feita manualmente utilizando programa Artemis®. Assim como na fase de comparação manual dos genomas, em 2.10.2, cada um dos arquivos foi aberto em uma mesma aba desse programa. A diferença dessa etapa esteve na adição do arquivo contendo somente os genes do “arquivo HGF” com potencial a serem verdadeiras.

Essa foi uma etapa importante no processo de validação do programa BOBBLES, pois permite a validação de um novo gene e verifica se as nomenclaturas e a localizações exatas desses genes foram mantidas.

Os genes marcados pelo BOBBLES são mostrados no Artemis® na cor magenta, A FIGURA 2.12, distinguindo-se das marcadas pelo programa HGF, que são as em vermelho, rosa claro e cinza, e as do genoma de referência, em azul.

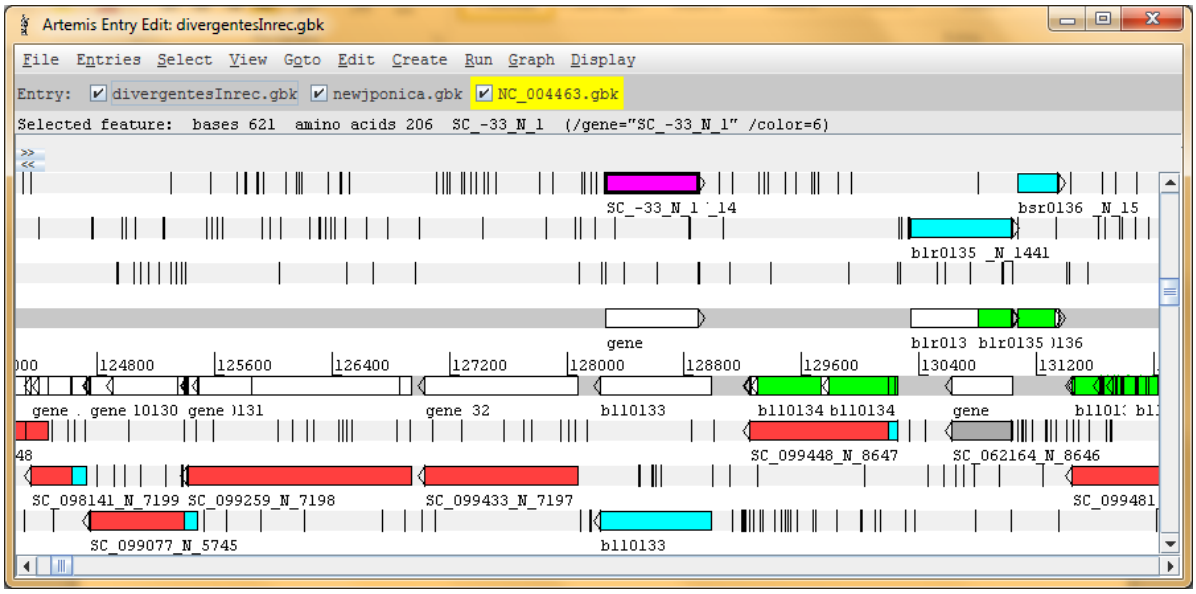


FIGURA 2.12 - ARQUIVOS DOS GENOMAS MOSTRADOS NO ARTEMIS PARA CONFERÊNCIA DOS GENES
FONTE: O autor (2012)

Além disso, para os casos em que a validação das sequências por alinhamento delas realizadas pelo programa SILA foi atribuído um cálculo de porcentagem de acerto:

$$Porcentagem\ de\ acerto = \frac{Número\ total\ de\ genes\ verdadeiros * 100}{Número\ total\ de\ genes}$$

Onde, $Número\ total\ de\ genes\ verdadeiros$ corresponde ao número total de genes novos somados ao número total de sequências de genes divergência em fase de leitura com valor de *bit score* superior a 80 ou, no caso das sequências de genes divergêntes em fase de leitura, as sequências com valor de *bit score* superior ou próximo ao valor de *bit score* do gene de referêcia. O $Número\ total\ de\ genes$ corresponde ao número total de genes novos somados ao número total de genes

concorrentes marcados pelo programa HGF e identificados pelo programa BOBBLES. Esse cálculo serviu para mostrar a confiabilidade do alinhamento das sequências através do programa SILA comparando-a com o programa BlastP.

Para a porcentagem média de acerto do programa SILA foi utilizado o seguinte cálculo:

$$Porcentagem\ média\ de\ acerto = \frac{\sum Porcentagem\ de\ acerto}{\text{Número de genomas utilizados}}$$

Onde, $\sum Porcentagem\ de\ acerto$ corresponde à somatória da porcentagem dos genomas utilizados dividida pelo número de geomas utilizados, $\text{Número de genomas utilizados}$, resulta na porcentagem média de acerto, $Porcentagem\ média\ de\ acerto$.

Também foi aplicado o cálculo de coeficiente de correlação de Pearson (ρ) para medir a relação das $Porcentagem\ de\ acerto$ com o conteúdo de GC e $Porcentagem\ de\ acerto$ com o tamanho dos genomas em pb. Calculado na planilha eletrônica Excel através do comando:

=correl(matrizA;matrizB)

Onde, *=correl* é a chamada para o cálculo de ρ, *matrizA* corresponde aos valores de cada $Porcentagem\ de\ acerto$, (para ambos os cálculos) e matrizB é a seleção de conteúdo de GC ou o tamanho do genoma em pb.

A interpretação do valor de ρ pode ser representada por:

- (0,0 > | ρ | > 0,19) indica correlação muito fraca;
- (0,20 > | ρ | > 0,39) indica correlação fraca;
- (0,40 > | ρ | > 0,69) indica correlação moderada;
- (0,70 > | ρ | > 0,89) indica correlação forte;
- (0,90 > | ρ | > 1,0) indica correlação muito forte.

3 RESULTADOS E DISCUSSÃO

Assim como na sessão 2.10, os resultados foram obtidos em duas fases diferentes: manual e automatizada. Na fase automatizada, os resultados foram obtidos utilizando programa BlastP e também o programa SILA. Os resultados obtidos através do programa SILA foram conferidos manualmente utilizando a ferramenta BlastP através do visualizador de genomas Artemis®.

3.1 COMPARAÇÃO MANUAL DO GENOMA

Esta fase foi uma etapa experimental com o intuito de entender a problemática e elaborar um algoritmo capaz de auxiliar nessa atividade. Por essa razão nem todos os genomas bacterianos listados em 2.7 foram analisados.

A TABELA 3 mostra o resultado da comparação manual feita em uma amostra de genomas completos bacterianos cujos genes foram preditos pelo programa HGF e comparados com o genoma de referência disponível no banco de dados GenBank®.

TABELA 3 - NÚMERO DE NOVOS GENES VERDADEIROS

GENOMA BACTERIANO	NÚMERO DE NOVOS GENES		NÚMERO DE SEQUÊNCIAS COM DIVERGENCIA EM FASE DE LEITURA	
	POSITIVO	FALSO POSITIVO	POSITIVO	FALSO POSITIVO
Escherichia coli K 12 substr DH10B	256	00	03	00
Herbaspirillum seropedicae SmR1	22	02	16	00
Ralstonia solanacearum CFBP2957	19	00	00	00
Rhizobium leguminosarum viciae 3841	55	00	25	00
TOTAL	352	02	44	00

FONTE: O autor (2012)

Notou-se que os números de novos genes e de genes com divergência de fase de leitura em relação ao genoma de referência não contemplam um padrão numérico, isso ocorre porque cada genoma de referência foi anotado de uma maneira diferente e por grupos de pesquisa diferentes. Foi observado que três dos quatro genomas avaliados não apresentaram números de novos genes

proporcionais aos números de sequências de regiões codificantes (CDS). No entanto, a quantidade total de novos genes foi considerada suficiente para justificar a continuidade deste trabalho com a automatização deste processo.

3.2 COMPARAÇÃO AUTOMATIZADA DO GENOMA

A comparação automatizada dos genomas bacterianos foi realizada pelo programa BOBBLES que em sua interface tem as opções de alinhamento das sequências utilizando o programa BlastP ou utilizando o programa SILA. O resultado do alinhamento das sequências foi utilizado como validador dessas sequências, tanto as novas quanto aquelas que apresentaram divergência em fase de leitura em relação ao gene contido no genoma de referência. Foram realizados testes utilizando ambos os programas de alinhamento para posteriormente analisar os resultados, principalemente os resultados obtidos através do programa SILA.

3.2.1 Conjunto de dados da pesquisa

A TABELA 4 mostra o conjunto de genomas de referência comparado com os genomas obtidos através do programa HGF. Pode se observar que para esse conjunto o número mínimo de genes com o mesmo *stop* códon contido na anotação do genoma de referência e no arquivo gerado pelo programa HGF foi de 820 genes e o máximo 7321 genes. No entanto, não é possível fazer uma estimativa de acertos utilizando o número de genes contidos no arquivo de referência com o número de genes obtidos através do programa HGF porque cada um dos genomas utilizados para esta pesquisa foi originado através de uma metodologia diferente.

TABELA 4 – GENES DO ARQUIVO DE REFERÊNCIA DO GENOMA BACTERIANO COMPARADO COM OS GENES OBTIDOS ATRAVÉS DO PROGRAMA HGF

GENOMA BACTERIANO	NÚMERO DE GENES NO ARQUIVO DE REFERÊNCIA	NÚMERO DE GENES NO ARQUIVO HGF	NÚMERO DE GENES COM O MESMO *STOP* CÓDON
Bradyrhizobium japonicum USDA 110	8317	8667	7321
Burkholderia mallei SAVP1	1734	1394	1201
Cyclobacterium marinum DSM 745	4998	6276	4733
Escherichia coli K 12 substr DH10B	4127	5380	3940
Herbaspirillum seropedicae SmR1	4735	6369	4296
Methanocaldococcus fervens AG86	1545	2335	1494
Pseudomonas fluorescens Pf-5	6107	6638	5875
Streptococcus agalactiae NEM316	2094	2554	2046
Streptococcus mutans UA159	1960	2379	1866
Streptococcus pneumoniae Hungary19A 6	2155	2730	2019
Thermotoga maritima MSB8	1854	2063	1716
Treponema denticola ATCC 35405	2767	3002	2527
Treponema pallidum Nichols	1034	955	820

FONTE: O autor (2012)

A TABELA 5 mostra o conjunto de genes avaliados nesta pesquisa, conforme apresentado em 2.10.2. Nota-se que o conjunto mínimo de genes para serem avaliados pelos programas SILA e BlastP através da ferramenta BOBBLES varia de 193 genes até 2073 genes. Isso justifica a variação de tempo de execução do programa BOBBLES.

TABELA 5 – NÚMERO DE GENES AVALIADOS PELO PROGRAMA BOBBLES

GENOMA BACTERIANO	NÚMERO DE GENES NO ARQUIVO HGF	NÚMERO DE GENES COM O MESMO *STOP* CÓDON	NÚMERO DE GENES AVALIADOS
Bradyrhizobium japonicum USDA 110	8667	7321	1346
Burkholderia mallei SAVP1	1394	1201	193
Cyclobacterium marinum DSM 745	6276	4733	1543
Escherichia coli K 12 substr DH10B	5380	3940	1440
Herbaspirillum seropedicae SmR1	6369	4296	2073
Methanocaldococcus fervens AG86	2335	1494	841
Pseudomonas fluorescens Pf-5	6638	5875	763
Streptococcus agalactiae NEM316	2554	2046	508
Streptococcus mutans UA159	2379	1866	513
Streptococcus pneumoniae Hungary19A 6	2730	2019	711
Thermotoga maritima MSB8	2063	1716	347
Treponema denticola ATCC 35405	3002	2527	475
Treponema pallidum Nichols	955	820	135

FONTE: O autor (2012)

3.2.2 Comparação automatizada utilizando o programa SILA

Os testes utilizando o programa *Sequence-Indexed Local Aligner* (SILA), explicado em 1.6, foram aplicados para os seguintes genomas: *Bradyrhizobium japonicum* USDA 110, *Cyclobacterium marinum* DSM, *Escherichia coli* K 12 substr DH10B, *Herbaspirillum seropedicae* SmR1, *Ralstonia solanacearum* CFBP2957, *Streptococcus agalactiae* NEM316, *Streptococcus mutans* UA159, *Streptococcus pneumoniae* Hungary19A 6, *Treponema pallidum* Nichols, *Pseudomonas fluorescens* Pf-5, *Thermotoga maritima* MSB8 e *Treponema denticola* ATCC 35405.

A TABELA 6 mostra os genomas, obtidos através do programa HGF, com os seus respectivos números totais de genes novos e com divergência em fase de leitura em relação ao genoma de referência. E também a soma de todos os genes novos e os com divergência em fase de leitura, mostrado na última linha da tabela.

TABELA 6 – GENES ENCONTRADOS PELO PROGRAMA HGF E VALIDADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA

GENOMA BACTERIANO	TOTAL	NOVOS GENES	GENES COM DIVERGENCIA EM FASE DE LEITURA
Bradyrhizobium japonicum USDA 110	65	25	40
Burkholderia mallei SAVP1	33	11	22
Cyclobacterium marinum DSM 745	26	12	14
Escherichia coli K 12 substr DH10B	154	138	16
Herbaspirillum seropedicae SmR1	44	03	41
Methanocaldococcus fervens AG86	16	08	08
Pseudomonas fluorescens Pf-5	59	31	28
Ralstonia solanacearum CFBP2957	36	22	14
Streptococcus agalactiae NEM316	30	21	09
Streptococcus mutans UA159	19	07	12
Streptococcus pneumoniae Hungary19A 6	102	89	13
Thermotoga maritima MSB8	28	12	16
Treponema denticola ATCC 35405	21	06	15
Treponema pallidum Nichols	19	01	18
TOTAL	652	386	266

FONTE: O autor (2012)

Os novos genes encontrados pelo programa HGF e avaliados pelo programa BOBBLES, utilizando o alinhamento das sequências através do programa SILA, foram novamente avaliados manualmente utilizando o programa BlastP remoto. Isso gerou os genes classificados em (i) positivo e (ii) falso positivo. O primeiro são aqueles que obtiveram valor de *bit score* superior a 80 pelo programa BlastP. Já o segundo, corresponde a aqueles que não conseguiram atingir esse valor. Foram observados, na TABELA 7, que os falsos positivos correspondem a um número inferior em relação aos genes classificados como positivo. Isso acontece porque o programa SILA possui um algoritmo diferente do programa BlastP. No entanto, essa diferença não se apresentou significativa. Além disso, foi notado que quase todos os genes, classificados como falso positivo, apresentaram pouca diferença no valor da linha de corte ou quase nenhuma ou nenhuma similaridade com o banco de dados NR. Nos casos em que houve pouca diferença entre o limite da linha de corte os genes classificados pelo programa SILA também estavam pouco acima da linha de corte estabelecida para ele, que era o mesmo valor 80 que foi designado para o programa BlastP. Já os genes que obtiveram quase nenhuma ou nenhuma similaridade com o banco de dados NR apresentaram em sua maioria a

classificação deles pelo programa HGF de muito provável, e isso pode significar genes novos ainda não disponíveis no banco de dados.

TABELA 7 – NOVOS GENES ENCONTRADOS PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E CONFERIDAS ATRAVÉS DO PROGRAMA BLASTP

GENOMA BACTERIANO	TOTAL	NOVOS GENES	
		POSITIVO	FALSO POSITIVO
Bradyrhizobium japonicum USDA 110	26	22	04
Burkholderia mallei SAVP1	11	09	02
Cyclobacterium marinum DSM 745	12	11	01
Escherichia coli K 12 substr DH10B	138	134	04
Herbaspirillum seropedicae SmR1	03	02	01
Methanocaldococcus fervens AG86	08	08	00
Ralstonia solanacearum CFBP2957	22	22	00
Streptococcus agalactiae NEM316	21	18	03
Streptococcus mutans UA159	07	07	00
Streptococcus pneumoniae Hungary19A 6	89	85	04
Treponema pallidum Nichols	01	01	00
Pseudomonas fluorescens Pf-5	31	30	01
Thermotoga maritima MSB8	12	11	01
Treponema denticola ATCC 35405	06	06	00
TOTAL	386	349	19

FONTE: O autor (2012)

A TABELA 8 mostra os genes encontrados pelo programa HGF e com divergência de fase de leitura, avaliados pelo programa BOBBLES utilizando o alinhamento das sequências através do programa SILA foram novamente avaliados manualmente utilizando o programa BlastP remoto. Isso gerou os genes classificados em (i) positivo, (ii) falso positivo, (iii) neutro e (iv) genes com valor de *bit score* menor que o obtido pelo gene concorrente do genoma de referência e com valor de *bit score* superior a 80. (i) são referentes a aqueles que apresentaram valor de *bit score* superior ao do gene concorrente, independente de possuir ou não valor de *bit score* superior a 80. (ii) refere-se aos genes com divergência e não obtiveram valor de *bit score* superior a 80 e também não apresentaram esse valor maior do que o segundo valor do gene concorrente. Já o (iii), corresponde aos genes cujos valores de *bit score* apresentaram pouca ou nenhuma diferença entre os genes sobrepostos. E o (iv) apresenta todos os genes com valor de *bit score* menor que o obtido pelo gene concorrente do genoma de referência, mas com valor de *bit score* superior a 80. Esta é uma característica do programa BOBBLES, para os genes com

valor de *bit score* superior a 80 pudessem ser avaliados posteriormente, uma vez que existem outros fatores que podem ser levados em consideração na hora de anotar um gene. Apesar da quantidade de genes falso positivo ser elevada em relação aos outros itens, deve-se levar em consideração que o algoritmo do programa SILA é diferente do algoritmo do programa BlastP e também que parte dos genes avaliados com divergência não apresentaram valor de *bit score* superior a 80.

TABELA 8 - GENES ENCONTRADOS PELO PROGRAMA HGF COM DIVERGÊNCIA DE FASE DE LEITURA E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIA ATRAVÉS DO PROGRAMA SILA E CONFERIDAS ATRAVÉS DO PROGRAMA BLASTP

GENOMA BACTERIANO	TOTAL	GENES COM DIVERGÊNCIA EM FASE DE LEITURA			
		POSITIVO	FALSO POSITIVO	NEUTRO	VALOR DE *BIT SCORE* < X* > 80
Bradyrhizobium japonicum USDA 110	40	16	14	04	06
Burkholderia mallei SAVP1	22	09	07	00	06
Cyclobacterium marinum DSM 745	14	01	10	03	00
Escherichia coli K 12 substr DH10B	16	04	04	00	08
Herbaspirillum seropedicae SmR1	41	13	21	06	01
Methanocaldococcus fervens AG86	08	00	07	01	00
Pseudomonas fluorescens Pf-5	28	11	11	02	04
Ralstonia solanacearum CFBP2957	14	02	11	00	01
Streptococcus agalactiae NEM316	09	02	07	00	00
Streptococcus mutans UA159	12	05	06	01	00
Streptococcus pneumoniae Hungary19A 6	13	05	05	00	03
Thermotoga maritima MSB8	16	04	12	00	00
Treponema denticola ATCC 35405	15	04	09	02	00
Treponema pallidum Nichols	18	06	11	00	01
TOTAL	266	82	135	19	30

X* gene do genoma de referência
FONTE: O autor (2012)

A TABELA 9 mostra as quantidades totais de genes encontrados por genoma estudado e a quantidade desses genes que foram classificados como positivo, assim, pôde-se obter a porcentagem de acerto mostrada na última coluna dessa tabela. O valor médio de acerto foi de 76,073%.

TABELA 9 - NÚMEROS TOTAIS DE GENES ENCONTRADOS E QUANTOS DELES SÃO VERDADEIROS E A PORCENTAGEM DE ACERTO POR GENOMA BACTERIANO

GENOMA BACTERIANO	NÚMEROS DE GENES ENCONTRADOS	NÚMEROS DE GENES POSITIVOS	PORCENTAGEM APROXIMADA DE ACERTO (%)
Bradyrhizobium japonicum USDA 110	65	47	72
Burkholderia mallei SAVP1	33	24	73
Cyclobacterium marinum DSM 745	26	15	58
Escherichia coli K 12 substr DH10B	154	146	95
Herbaspirillum seropedicae SmR1	44	22	50
Methanocaldococcus fervens AG86	16	09	56
Pseudomonas fluorescens Pf-5	59	47	80
Ralstonia solanacearum CFBP2957	36	25	49
Streptococcus agalactiae NEM316	30	20	67
Streptococcus mutans UA159	19	13	68
Streptococcus pneumoniae Hungary19A 6	102	93	91
Thermotoga maritima MSB8	28	15	53
Treponema denticola ATCC 35405	21	12	57
Treponema pallidum Nichols	19	08	42
TOTAL	652	496	76

FONTE: O autor (2012)

A TABELA 10 mostra a comparação da porcentagem de acerto do programa BOBBLES utilizando o alinhamento das sequências com o programa SILA com o conteúdo de GC contido em cada genoma do grupo de teste. Foi observado que todos os genomas na faixa de 70% de acerto continham a faixa de 60% do conteúdo de GC no genoma, porém nessa mesma faixa de conteúdo de GC apresentou um genoma cuja porcentagem de acerto foi de 49,44%, a *Ralstonia solanacearum* CFBP2957. Na faixa de 30 % do conteúdo de GC no genoma, nota-se que apresentaram a maioria das porcentagens de acerto de 57% e 68%, menos o genoma *Streptococcus pneumoniae* Hungary19A 6 que apresentou 91,176% de acerto. Nas faixas de 40% e 50% do conteúdo de GC os resultados foram discrepantes, variando de 42,105% a 94,805% de acerto.

O coeficiente de variação aplicado para os dados desta tabela resultou em aproximadamente 0,05, provando a correlação muito fraca para esses dados.

TABELA 10 - COMPARAÇÃO DA PORCENTAGEM DE ACERTO COM O CONTEÚDO DE GC

GENOMA BACTERIANO	PORCENTAGEM APROXIMADA DE ACERTO (%)	CONTEÚDO DE GC (%)
Bradyrhizobium japonicum USDA 110	72	64,1
Burkholderia mallei SAVP1	73	68,4
Cyclobacterium marinum DSM 745	58	38,1
Escherichia coli K 12 substr DH10B	95	50,8
Herbaspirillum seropedicae SmR1	50	63,4
Methanocaldococcus fervens AG86	56	32,2
Pseudomonas fluorescens Pf-5	80	63,3
Ralstonia solanacearum CFBP2957	49	66,5
Streptococcus agalactiae NEM316	67	35,6
Streptococcus mutans UA159	68	36,8
Streptococcus pneumoniae Hungary19A 6	91	39,6
Thermotoga maritima MSB8	53	46,2
Treponema denticola ATCC 35405	57	37,9
Treponema pallidum Nichols	42	52,8

FONTE: O autor (2012) e (NCBI, 2012)

A TABELA 11 exibe a comparação da porcentagem de acerto do programa BOBBLES utilizando o alinhamento das sequências com o programa SILA com o número de pares de base (pb) do arquivo do genoma. Notou-se a porcentagem de acerto não possui relação com o número de pb do arquivo do genoma, o que caracteriza que o programa BOBBLES quanto os periféricos utilizados não apresentaram problemas em relação a capacidade computacional para realizar a comparação dos arquivos.

O coeficiente de variação aplicado para os dados desta tabela resultou em aproximadamente 0,28, provando a correlação fraca para esses dados.

TABELA 11 - COMPARAÇÃO DA PORCENTAGEM DE ACERTO COM O NÚMERO DE pb DO GENOMA

GENOMA BACTERIANO	PORCENTAGEM APROXIMADA DE ACERTO (%)	TAMANHO EM pb
Bradyrhizobium japonicum USDA 110	72	9105828
Burkholderia mallei SAVP1	73	3497479
Cyclobacterium marinum DSM 745	58	6221273
Escherichia coli K 12 substr DH10B	95	4686137
Herbaspirillum seropedicae SmR1	50	5513887
Methanocaldococcus fervens AG86	56	22190
Pseudomonas fluorescens Pf-5	80	7074893
Ralstonia solanacearum CFBP2957	49	3417386
Streptococcus agalactiae NEM316	67	2211485
Streptococcus mutans UA159	68	2032925
Streptococcus pneumoniae Hungary19A 6	91	2245615
Thermotoga maritima MSB8	53	1860725
Treponema denticola ATCC 35405	57	2843201
Treponema pallidum Nichols	42	1138011

FONTE: O autor (2012) e (NCBI, 2012)

A TABELA 12 apresenta a comparação da porcentagem de acerto do programa BOBBLES utilizando o alinhamento das sequências com o programa SILA com a grupo taxonômico dos genomas utilizados para teste. Foi observado que os genomas cujos grupos taxonômicos são proteobacteria, enterobacteria e firmicutes apresentaram mais genomas com os maiores valores de porcentagem de acerto. Enquanto os genomas cujos grupos taxonômicos são bacterioidetes, euryarchaeotes, thermotogales e spirochetes mais genomas com os menores valores de porcentagem de acerto. Isso pode caracterizar um indício dos genomas em que o programa SILA obtém melhores resultados.

TABELA 12 - COMPARAÇÃO DA PORCENTAGEM DE ACERTO COM O GRUPO TAXONÔMICO DO GENOMA

GENOMA BACTERIANO	PORCENTAGEM APROXIMADA DE ACERTO (%)	GRUPO TAXONÔMICO
Bradyrhizobium japonicum USDA 110	72	Proteobacteria
Burkholderia mallei SAVP1	73	Proteobacteria
Cyclobacterium marinum DSM 745	58	Bacterioidetes
Escherichia coli K 12 substr DH10B	95	Proteobacteria
Herbaspirillum seropedicae SmR1	50	Proteobacteria
Methanocaldococcus fervens AG86	56	Euryarchaeota
Pseudomonas fluorescens Pf-5	80	Proteobacteria
Ralstonia solanacearum CFBP2957	49	Proteobacteria
Streptococcus agalactiae NEM316	67	Firmicutes
Streptococcus mutans UA159	68	Firmicutes
Streptococcus pneumoniae Hungary19A 6	91	Firmicutes
Thermotoga maritima MSB8	53	Thermotogae
Treponema denticola ATCC 35405	57	Spirochetes
Treponema pallidum Nichols	42	Spirochetes

FONTE: O autor (2012) e (NCBI, 2012)

A TABELA 13 exibe os novos genes encontrados pelo programa HGF e avaliados pelo programa BOBBLES utilizando o alinhamento das sequências com o programa SILA comparado com o alinhamento das sequências com o programa BlastP. Foi observado que a maioria dos casos não apresentam muita diferença no número de genes utilizando os programas de alinhamento de sequência diferentes.

TABELA 13 – NOVOS GENES ENCONTRADOS PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA COMPARADOS COM PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA BLASTP

GENOMA BACTERIANO	NOVOS GENES	
	SILA	BLASTP
Bradyrhizobium japonicum USDA 110	22	61
Burkholderia mallei SAVP1	09	06
Cyclobacterium marinum DSM 745	11	30
Escherichia coli K 12 substr DH10B	134	122
Herbaspirillum seropedicae SmR1	02	06
Methanocaldococcus fervens AG86	08	12
Pseudomonas fluorescens Pf-5	30	51
Ralstonia solanacearum CFBP2957	22	19
Streptococcus agalactiae NEM316	18	15
Streptococcus mutans UA159	07	07
Streptococcus pneumoniae Hungary19A 6	85	79
Thermotoga maritima MSB8	11	21
Treponema denticola ATCC 35405	06	09
Treponema pallidum Nichols	01	08
TOTAL	366	529

FONTE: O autor (2012)

Os anexos de 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25 e 27 mostram todos os genes novos marcados pelo programa HGF e avaliados pelo programa BOBBLES utilizando o alinhamento de sequência pelo programa SILA desses 14 genomas avaliados. Foram observados que praticamente todos os genes de referência obtidos por esse processo pertencem a organismos do mesmo gênero e inclusive alguns desses com organismos da mesma espécie. Sendo os anexos 1, 3, 5, 7, 9, 13 e 15 apresentam genes pertencentes ao grupo taxonômico Proteobacteria; 5 apresenta genes pertencentes ao grupo taxonômico Bacterioidetes; 11 apresenta genes pertencentes ao grupo taxonômico Euryarchaeota; 17, 19, e 21 apresentam genes pertencentes ao grupo taxonômico Firmicutes; 23 apresenta genes pertencentes ao grupo taxonômico Thermotogae; 25, e 27 apresentam genes pertencentes ao grupo taxonômico Spirochaetes.

Já os anexos 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26 e 28 mostram todos os genes com divergência em fase de leitura marcados pelo programa HGF e avaliados pelo programa BOBBLES utilizando o alinhamento de sequência através do programa SILA desses 14 genomas avaliados. Estes anexos mostram que existe mais discrepância em relação aos gêneros e espécies obtidos através do

alinhamento das sequências. E os anexos de 29 até o 41 mostram todos os novos genes encontrados pelo programa BOBBLES utilizando o alinhamento de sequências pelo programa BlastP. Em todos os anexos são apresentados a localização dos gene e o valor de *bit score* apresentado pelo alinhamento utilizando o programa BlastP, dessa forma é possível notar quais desses genes possuem ou não grau de similaridade significativo para ser considerado um gene verdadeiro.

3.3 AVALIAÇÃO DE DESEMPENHO DO PROGRAMA BOBBLES

A primeira versão do BOBBLES utilizava o programa BlastP local com o banco de dados NR e todas as suas funções eram feitas na linguagem Matlab®. O banco de dados NR demorou quatro horas para ser montado utilizando as funções descritas em 2.10.2.4. Nessa versão, foram realizados testes com as sequências do *Herbaspirillum seropedicae* SmR1 e cada consulta ao banco de dados demorou de seis a 10 minutos, por essa razão foi descartada essa versão.

A partir disso foram realizados testes com esse mesmo genoma utilizando o alinhamento das sequências pelo programa BlastP local via terminal e remotamente. Observou-se que o acesso do Matlab® para alinhamentos de sequências via terminal e remotamente não apresentavam diferença significativa de tempo para consulta das sequências. Apesar dos alinhamentos feitos diretamente no terminal demorarem frações de segundo, via termal Shell do Linux, quando esses *scripts* eram acessados via plataforma Matlab® eles perdiam em desempenho, pois o tempo de espera para a realização dos alinhamentos aumentava significativamente, passando a ser tão demorado quanto em acesso via acesso remoto, o qual variou de cinco segundos até dois minutos por alinhamento. Além disso, o acesso local de alinhamento via terminal Windows® nos computadores utilizados demoravam de quatro até 10 minutos para realizar cada alinhamento de sequência contra o banco de dados NR.

Por essa razão a versão final do programa BOBBLES possui a opção de alinhamento via BlastP remoto. Mesmo que a utilização do programa BlastP remoto tenha se apresentado diretamente afetada de acordo com a qualidade da conexão com a internet, ou seja, se o programa BOBBLES fosse desconectado da internet ele não conseguia obter os resultados corretos. Entretanto, cada alinhamento

demorou de cinco segundos até dois minutos para ser executado, sendo que quanto mais sequências precisassem ser alinhadas mais era o tempo necessário para alinhar cada uma delas. Isso aconteceu porque o programa BlastP remoto impõe tempo de espera maior quando as solicitações de alinhamento vindo de um mesmo local são muitas.

Outro fator é que o programa SILA foi utilizado como uma opção de alinhamento nas sequências por ser pelo menos cinco vezes mais rápido que o programa BLAST e isso fez com que reduzisse o tempo de execução do programa BOBBLES, que foram de acordo com os testes apresentados em 1.7. Mesmo os resultados da comparação dos alinhamentos deste com o programa BLASTP não serem idênticos, o programa SILA ainda não se encontra em sua versão final, o que torna essa opção para a utilização do BOBBLES promissora.

4 CONCLUSÕES

- A ferramenta, *Bioinformatics tool based on bacterial genomes comparison*, BOBBLES, foi testada em 14 genomas bacterianos de diferentes tamanhos e grupos taxonômicos e mostrou-se eficiente na detecção de novos genes e genes erroneamente localizados.
- A ferramenta HGF mostrou-se eficiente na detecção de novos genes não apresentando tendência relevante em relação ao grupo taxonômico, conteúdo de GC e tamanho do genoma bacteriado.
- A maioria dos novos genes encontrados pelo programa HGF e analisados pelo programa BOBBLES pertencem ao mesmo gênero do organismo avaliado.
- A execução do programa BOBBLES utilizando alinhamento de sequências com o programa SILA foi pelo menos cinco vezes a execução mais rápida do que utilizando o alinhamento com o programa BLASTP.
- O programa SILA é uma ferramenta promissora para realizar alinhamento de sequências e pode ser utilizada em conjunto com o programa BLASTP.

5 PERSPECTIVAS FUTURAS

Como perspectiva futura, é sugerida a elaboração de novas técnicas utilizando outros algoritmos para comparar os genomas a fim de diminuir a taxa de erro e o tempo de execução do programa BOBBLES, mesmo ele tendo apresentado bom desempenho de execução. Essa técnica pode consistir em utilizar o programa SILA e o programa BLASTP em conjunto. Sendo, o programa SILA para a verificação dos genes mais prováveis por homologia com outros genes e os outros genes serem submetidos ao programa BLASTP.

Além disso, neste trabalho não foi contemplado o estudo dos genes não localizados pelo programa HGF contidos nos genomas de referência. Esse estudo pode verificar se esses genes possuem homologia com outros genes e identificar aqueles que podem ter sido anotados erroneamente. E, assim, identificar novos padrões para identificação de genes e, se possível, contemplá-la em uma nova versão do programa HGF.

REFERÊNCIAS

ALMEIDA, A.C.B. de. **BIOANOT: um sistema multi-agentes para notificação de (re) anotações de sequências em bancos de dados genômicos**. Rio de Janeiro: Curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia, 2006.

ALTSCHUL, S.F.; GISH, W.; MILLER, W.; MYERS, E.W.; LIPMAN, D.J. **Basic Local Alignment Search Tool**. Bethesda: National Center for Biotechnology Information. Jornal of Molecular Biology, vol 215, páginas 403 até 410, 1990.

BENSON, D.A.; KARSCH-MIZRACHI, I.; LIPMAN, D.; OSTELL, J.; WHEELER, D. **Genbank: update**. Bethesda: Nucleic Acids Research, volume 32, Database Issue D23 – D26, Oxford University Press, 2004.

BENSON, D.A.; KARSCH-MIZRACHI, I.; LIPMAN, D.; OSTELL, J.; SAYERS, E.W. **Genbank**. Bethesda: Nucleic Acids Research, volume 39, Database Issue D32 – D37, DOI:10.1093/nar/gkq107, 2010.

BOECKMANN, B.; BAIROCH, A.; APWEILER, R.; BLATTER, M.C.; ESTREICHER, A.; GASTEIGER, E.; MARTIN, M.J.; MICHOUD, K.; O'DONOVAN, C.; PHAN, I.; PILBOUT, S.; SCHNEIDER, M. **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**, Geneva: Nucleic Acids Research, volume 31, páginas 365 até 370, 2003.

BROWN, T.A.; **Genomes**. New York: Wiley-Liss, 2ºed., 2002.

BRUNAK, S.; DANCHIN, A.; HATTORI, M.; NAKAMURA, H.; SHINOZAKI, K.; MATISE, T.; PREUSS, D. **Nucleotide Sequence Database Policies**. Science 298 (5597): 1333, 2002.

CHARNIAK, E.; MCDERMOTT, D. **Introduction to Artificial Intelligence**. Addison: Wesley, Reading, MA, 1985.

CONESA, A.; GÖTZ, S.; GARCÍA-GÓMEZ, J.M.; TEROL, J.; TALÓN, M.; ROBLES, M. **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. Valencia: Centro de Genomica. Bioinformatics

Applications Notes. Volume 21 nº 18, páginas 3674 até 3676. DOI:10.1093/bioinformatics/bti610, 2005.

DELCHER, A.L.; BRATKE, K.A.; POWERS, E.C.; SALZBERG, S.L. **Identifying bacterial genes and endosymbiont DNA with Glimmer**. Bioinformatics volume 23, nº 6, páginas 673 até 679, 2007.

DEONIER, R.C.; TAVARÉ, S.; WATERMAN, M.S. **Computational Genome Analysis: an introduction**. New York: Springer-Verlag. ISBN 0387987851, 2005.

ELMASRI, E.; NAVATHE, S. **Sistemas de Bancos de Dados**. São Paulo: Addison Wesley, Pearson, 4ª ed., 2005.

FASSLER, J.; COOPER, P. **BLAST Glossary**. Disponível em: <http://www.ncbi.nlm.nih.gov/books/NBK62051/>. Último acesso: 19/01/2012. Última atualização: 2011.

GEER, R.C.;SAYERS, E.W. **Entrez: Making use of its power**. Briefings in Bioinformatics. Henry Stewart Publications 1467-5463. volume 4, nº 2, páginas 179 até 184. Junho de 2003.

GEHLEN, M.A.C. **Estudo e levantamento dos genes nif publicados no NCBI usando conceitos de mineração de dados e Inteligência Artificial**. Curitiba: Universidade Federal do Paraná, 2011.

GENBANK. **Genetic Sequence Data Bank**. Disponível em: <ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>. Último acesso: 26/01/2012. Última atualização: 2011.

GIBAS, C.; JAMBECK, P. **Desenvolvendo Bioinformática**. São Paulo: Editora Campus, 2002.

GILKS, W.R.; AUDIT, B.; DE ANGELIS, D.; TSOKA, S.; OUZOUNIS, C.A. **Modeling the percolation of annotation errors in a database of protein sequences**. Bioinformatics, volume 12, páginas 1641 até 1649, 2002.

GRIFFITHS, A.J.F.; MILLER,, J.H.; GELBART, W.M.; LEWOTIN, R.C. **Modern genetic analysis**. New York: W. H. Freeman & Co, 1999.

HAYKIN, S. **Neural Networks – A Compreensive Foundation**. New Jersey: Prentice-Hall, 2nd edition, 1999.

HAYKIN, S. **Redes Neurais Princípios e Prática.** Tradução de: Paulo Martins Engel. Porto Alegre: Bookman, 2001.

HENIKOFF, S. *et al*. **Gene families: the taxonomy of protein paralogs and chimeras**. Science, volume 278, páginas 609 até 614, 1997.

HP. **GBK (5)**. Disponível em <http://h30097.www3.hp.com/docs/base_doc/DOCUMENTATION/V51_HTML/MAN/MAN5/0020____.HTM>. Acessado em: 26/01/2012.

LATIMER, K. **Improving the detection of orthologs by altering and comparing various methods in bioinformatics**. Waterloo: Wilfrid Laurier University, 2007.

LEMOS, M. **Workflow para Bioinformática**. Rio de Janeiro: Programa de Pós-graduação em Informática da PUC-Rio, 2004.

LEVY, E.D.; OUZOUNIS, C.A.; GILKS, W.R.; AUDIT, B. **Probabilistic annotation of protein sequences based on functional classifications**. Cambridge: BMC Bioinformatics. DOI: 10.1186/1471-2105-6-302, 2005.

LIBERMAN, F. **Análise dos fatores determinantes para a qualidade da anotação genômica automática**. Brasília: Programa de Pós- Graduação "*Strictu Sensu*" em Biotecnologia e Ciências Genômicas da Universidade Católica de Brasília, 2004

LIPMAN, D.J.; PEARSON, W.R. **Rapid and sensitive protein similarity searches**. Science 22, volume 227 nº 4693 páginas 1435 até 1441. DOI: 10.1126/science.2983426, 1985.

LORENZI, H.A.; PUIU, D.; MILLER, J.R.; BRINKAC, L.M.; AMEDEO, P.; *et al*. **New assembly, reannotation and analysis of the Entamoeba histolyca genome reveal new genomic features and protein content formation**. PLoS Negl Trop Dis., 4:e716, 2010.

MAO, C.; QIU, J.; WANG, C.; CHARLES, T.C.; SOBRAL, B.W.S. **NodMutDB: a database for genes and mutants involved in symbiosis**. Virginia: Virginia Bioinformatics Institute of Virginia Polytechnic and State University, Blacksburg. Volume 21 nº 12, páginas 2927 até 2929. DOI:10.1093/bioinformatics/bti427, 2005

MCCULLOCH, W.S.; PITTS, W. **A logical calculus of the ideas immanent in nervous activity**. Bulletin of Mathematical Biophysics, vol. 5, páginas 115 até 133, 1943.

MEYER, F.; GOESMANN, A.; MCHARDY, A.C.; BARTELS, D.; BEKEL, T.; CLAUSEN, J.; KALINOWSKI, J.; LINKE, B.; RUPP, O.; GIEGERICH R.; PUÈHLER, A. **GenDBÐan open source genome annotation system for prokaryote genomes**. Bielefeld: Center for Genome Research. Nucleic Acids Research, volume 31, nº 8, páginas 2187 até 2195. DOI: 10.1093/nar/gkg312, 2003.

MORGULIS, A. **Database indexing for production MegaBLAST searches**. USA: National Center for Biotechnology Information, National Institutes of Health, Department of Health and Human. Volume 24, nº 16, páginas 1757 até 1764. DOI:10.1093/bioinformatics/btn322, 2008

NATIONAL HUMAN GENOME RESEARCH INSTITUTE. **Frequently asked questions about genetic and genomic science**. Disponível em: < http://www.genome.gov/190 16904>. Último acesso: 19/01/2012. Última atualização: 2010.

NCBI. **Just the Facts: A Basic Introduction to the Science Underlying NCBI Resources**. Disponível em: <ncbi.nlm.nih.gov/About/primer/bioinformatics.html>. Último acesso: 19/01/2012. Última atualização: 2004.

NCBI. **Browse Genomes**. Disponível em: < http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html>. Último acesso: 19/01/2012. Última atualização: 2012.

NEWELL, A.; SIMON, H.A. **Computer Science as Empirical Inquiry: Symbols and Search**. Communications of the ACM, volume 19, nº 3, páginas 113 até 126, DOI:10.1145/360018.360022, 1976.

ORACLE. GB18030-2000 – **The New Chinese National Standart**. Disponível em: <http://developers.sun.com/dev/gadc/technicalpublications/articles/gb18030.html>. Último acesso: 26/01/2012. Última atualização: 2010.

OUZOUNIS, C.A.; KARP, P.D. **The past, present and future of genome-wide re-annotation**. Cambridge: Genome Biology, volume 3, páginas 1 até 6, 2002.

PAULA, M.G. **Reconhecimento de palavras faladas utilizando Redes Neurais Artificiais**. Pelotas: Universidade Federal de Pelotas, 2000.

PETSKO, G.A.; RINGE, D. **Protein structure and function**. Waltham: New Science Press Ltd, 2003.

RAUBER, T.W. **Redes Neurais Artificiais**. Vitória: Universidade Federal do Espírito Santo: 2006.

RUSSELL S.; NORVIG P. **Artficial Intelligence: A Modern Approach**. Prentice-Hall, Saddle River, NJ, 1995.

RUTHERFORD, K.; PARKHILL, J.; CROOK, J.; HORSNELL, T.; RICE, P.; RAJANDREAM, M.A.; BARRELL, B. **Artemis: sequence visualization and annotation**. England: Bioinformatics, volume16 nº 10, páginas 944 até 945, 2000.

SANGER, F.; NICKLEN, S.; COULSON, A.R. **DNA Sequencing with Chain-Terminating Inhibitors**. Cambridge: Medical Research Council Laboratory of Molecular Biology. PNAS, volume 74, nº 12, páginas 5463 até 5467, 1977.

SALZBERG, S.L; DELCHER, A.L.; KASIF, S.; WHITE, O. **Microbial gene identification using interpolated Markov models**. Chicago: Nucleic Acids Research, volume 26, nº 2, páginas 544 até 548, 1998.

SANTOS, A.; AZEVEDO, V.; SCHNEIDER, M.P.; SILVA, A.C DA.; MIYOSHI A.; BORÉM, A. **Manual prático-teórico: sequenciamento, montagem e anotação de genomas bacterianos**. Belo Horizonte: Suprema. ISBN: 978.85.60249-83-4, páginas 91 até 109, 2011.

SETUBAL, J.C.; *et al*. **Genômica**. São Paulo: Editora Atheneu, páginas 107 até 118, 2004.

Shiryev S.A.; PAPADOPOULOS J.S.; SCHÄFFER A.A.; AGARWALA R. **Improved BLAST searches using longer words for protein seeding**. Bioinformatics. Volume 23, nº 21, páginas 2949 até 2951, 2007.

SHULAEV, V.; SARGENT, D.J.; CROWHURST, R.N.; MOCKLER, T.C.; *et al*. **The genome of woodland strawberry (Fragaria vesca)**. Nature Genetics, Nature America, Inc., 2010.

SMITH, T.F.; WATERMAN, M.S. **Identification of Common Molecular Subsequences**. London: Jornal of Molecular Biology, volume147, páginas 195 até197, 1981.

SOUZA, J. A. **Reconhecimento de padrões usando indexação recursiva**. Florianópolis: Universidade Federal de Santa Catarina, 1999.

STEIN, L. **Genome annotation: from sequence to biology**. New York: Nature Reviews Genetics, volume 2, páginas 493 até 503, 2001.

VIALLE, R.A.; SOUZA, E.M.; PEDROSA, F.O.; MARCHAUKOSKI, J.N.; STEFFENS, M.B.R.; GUIZELINI, D.; TIBÃES J.H.; SOUZA, V.; RAITTZ, R.T. **Recursive indexing (INREC) applied to sequence similarity searches in large databases**. Florianopolis: 7th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (AB3C) and 3rd International Conference of the IberoAmerican Society of Bioinformatics (SolBio), 2011.

WARREN, A.S; ARCHULETA, J.; FENG, W.; SETUBAL, J.C. **Missing genes in the annotation of prokaryotic genomes**. BMC Bioinformatics 2010, 11:131. DOI:10.1186/1471-2105-11-131, 2010.

WASSERMAN, P.D. **Neural Computing: Theory and Practice**. New York: Van Nostrand Reinhold, 1989.

WESTHEAD, D.R.; PARISH, J.H.; TWYMAN, R.M**. Instant Notes: Bioinformatics**. Cambridge: BIOS Scientific Publishers Limited, 2002.

WONG, W.; MAURER-STROH, S.; EISENHABER, F. **More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology**. PLoS Comput Biol.,6:e1000867, 2010.

APÊNDICES

APÊNDICE 1 – GENES NOVOS ENCONTRADOS NO GENOMA *Bradyrhizobium japonicum* USDA 110 PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE* DO BLAST**	**GENOMA DE REFERÊNCIA**
539055..539441	207	hypothetical protein BJ6T_04720 [*Bradyrhizobium japonicum* USDA 6]
1960604..1960807	136	ID204 [*Bradyrhizobium japonicum*]
1989091..1989420	207	hypothetical protein bll1960 [*Bradyrhizobium japonicum* USDA 110]
2023135..2023596	293	ID344 [*Bradyrhizobium japonicum*]
2069369..2069698	207	hypothetical protein BJ6T_88720 [*Bradyrhizobium japonicum* USDA 6]
2364690..2365223	288	hypothetical protein BJ6T_76610 [*Bradyrhizobium japonicum* USDA 6]
2453313..2453525	115	hypothetical protein BJ6T_75630 [*Bradyrhizobium japonicum* USDA 6]
2677987..2678196	140	hypothetical protein BJ6T_73800 [*Bradyrhizobium japonicum* USDA 6]
2769549..2769740	120	ypothetical protein BJ6T_72860 [*Bradyrhizobium japonicum* USDA 6]
2848756..2848950	117	hypothetical protein BJ6T_72050 [*Bradyrhizobium japonicum* USDA 6]
3073497..3073937	36,2	hypothetical protein PaTRP_07219 [*Paracoccus sp.* TRP]
4250641..4250967	198	hypothetical protein BJ6T_60230 [*Bradyrhizobium japonicum* USDA 6]
4332929..4333168	156	hypothetical protein BJ6T_59350 [*Bradyrhizobium japonicum* USDA 6]
4634660..4635250	365	hypothetical protein BJ6T_56010 [*Bradyrhizobium japonicum* USDA 6]
4673447..4673761	34,5	erminase small subunit [*Proteus mirabilis* ATCC 29906]
4859967..4860221	156	hypothetical protein BJ6T_54070 [*Bradyrhizobium japonicum* USDA 6]
5185998..5186216	144	hypothetical protein BJ6T_50420 [*Bradyrhizobium japonicum* USDA 6]
5458668..5458871	116	hypothetical protein BJ6T_47610 [*Bradyrhizobium japonicum* USDA 6]
5724635..5724883	104	hypothetical protein blr4764 [*Bradyrhizobium japonicum* USDA 110]
complement(5742815..5743459)	317	transposase [*Bradyrhizobium japonicum* USDA 110]
complement(5774731..5775075)	66	hypothetical protein BJ6T_20380 [*Bradyrhizobium japonicum* USDA 6]
complement(6038905..6039423)	276	hypothetical protein BJ6T_15570 [*Bradyrhizobium japonicum* USDA 6]

7761226..7761519	170	hypothetical protein BJ6T_23640 [*Bradyrhizobium japonicum* USDA 6]
complement(8176664..8176810)	85,5	hypothetical protein BJ6T_18640 [*Bradyrhizobium japonicum* USDA 6]
8392906..8393514	40	cob(I)yrinic acid a,c-diamide adenosyltransferase [*Frankia sp.* EAN1pec]

APÊNDICE 2 – GENES COM SOBREPOSIÇÃO DE FASE DE LEITURA NO GENÔMA *Bradyrhizobium japonicum* USDA 110 ENCONTRADOS PELO PROGRAMA HGF E VALIDADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE* DO BLAST**	**GENOMA DE REFERÊNCIA**
128266..128886	305	hypothetical protein BJ6T_01470 [*Bradyrhizobium japonicum* USDA 6]
complement(529569..529919)	35,8	thymidine phosphorylase [*Verrucosispora maris* AB-18-032]
1691443..1691619	32,7	hypothetical protein MCAG_03089 [*Micromonospora* sp. ATCC 39149]
1700608..1700907	202	hypothetical protein BJ6T_82900 [*Bradyrhizobium japonicum* USDA 6]
1818439..1818723	147	hypothetical protein Nham_1349 [*Nitrobacter hamburgensis* X14]
1827991..1828356	40	phosphopantethiene--protein transferase domain [*Desulfuromonas acetoxidans* DSM 684]
1924411..1924770	248	hypothetical protein BJ6T_80570 [*Bradyrhizobium japonicum* USDA 6]
2105350..2105646	36,2	neuregulin 3 variant 9 [*Homo sapiens*]
2140851..2141612	124	hypothetical protein F11_12550 [*Rhodospirillum rubrum* F11]
complement(2332004..2332108)	29	hypothetical protein amb0401 [*Magnetospirillum magneticum* AMB-1]
2347755..2348552	38,2	hypothetical protein Caur_0152 [*Chloroflexus aurantiacus* J-10-fl]
2744061..2744900	429	hypothetical protein BJ6T_73100 [*Bradyrhizobium japonicum* USDA 6]
2838536..2838811	119	hypothetical protein BJ6T_72130 [*Bradyrhizobium japonicum* USDA 6]
2912528..2912860	35	hypothetical protein [*Plasmodium vivax* Sal-1]
3009820..3009996	35	hypothetical protein PPSIR1_29835 [*Plesiocystis pacifica* SIR-1]
3311734..3311967	152	hypothetical protein BJ6T_67670 [*Bradyrhizobium japonicum* USDA 6]
4077397..4077705	179	hypothetical protein BJ6T_61590 [*Bradyrhizobium japonicum* USDA 6]
4179580..4180041	233	hypothetical protein BJ6T_61110 [*Bradyrhizobium japonicum* USDA 6]

4288920..4289195	36,6	hypothetical protein AURANDRAFT_61023 [*Aureococcus anophagefferens*]
4402944..4403873	519	hypothetical protein BJ6T_58710 [*Bradyrhizobium japonicum* USDA 6]
4593997..4594254	35,4	hypothetical protein Rvan_2791 [*Rhodomicrobium vannielii* ATCC]
complement(4610199..4610477)	36	glycosyl hydrolase, family 25 [*Prevotella veroralis* F0319]
5013752..5014453	308	unknown [*Bradyrhizobium japonicum*]
5220355..5220828	160	hypothetical protein BJ6T_50090 [*Bradyrhizobium japonicum* USDA 6]
5336430..5336750	33,5	hypothetical protein Gobo1_18286 [*Gluconacetobacter oboediens*]
5548820..5549206	35,8	predicted protein [*Streptomyces* sp. C]
5551384..5551629	57	acyltransferase [*Ahrensia sp*. R2A130]
5663729..5663935	32,7	hypothetical protein CGSSp14BS69_02514 [*Streptococcus pneumonia*]
5701286..5701741	37,4	hypothetical protein Acid345_0226 [*Candidatus Koribacter versatilis* Ellin345]
6608812..6609108	36	histidine ammonia-lyase [*Taylorella asinigenitalis* MCE3]
6957236..6957547	34,5	hypothetical protein [*Entamoeba dispar* SAW760]
7035168..7035983	145	6-pyruvoyl-tetrahydropterin synthase [*Rothia mucilaginosa* DY-18]
7346107..7346655	139	Collagen triple helix repeat [*Rhodopseudomonas palustris* TIE-1]
7348696..7349049	220	hypothetical protein BJ6T_27450 [*Bradyrhizobium japonicum* USDA 6]
7584988..7586121	703	hypothetical protein BJ6T_24880 [*Bradyrhizobium japonicum* USDA 6]
8084086..8084448	102	hypothetical protein RPE_1043 [*Rhodopseudomonas palustris* BisA53]
8177312..8178403	189	hypothetical protein BJ6T_18610 [*Bradyrhizobium japonicum* USDA 6]
8775567..8776439	41,2	hypothetical protein RSPO_m00435 [*Ralstonia solanacearum* Po82]
8827927..8828355	105	hypothetical protein blr7656 [*Bradyrhizobium japonicum* USDA 110]
8857656..8857967	34,3	hypothetical protein Cpin_0029 [*Chitinophaga pinensis* DSM 2588]
8982661..8983215	378	hypothetical protein BJ6T_87830 [*Bradyrhizobium japonicum* USDA 6]
9072171..9072449	33,1	hypothetical protein [*Arabidopsis thaliana*]

APÊNDICE 3 – GENES NOVOS ENCONTRADOS NO GENOMA *Burkholderia mallei* SAVP1 PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR *BIT-SOCRE* DO SILA**	**VALOR *BIT-SCORE* DO BLAST**
4051.. 4324	80,33333	174
120353.. 121583	893,6667	802
208762.. 209932	865,3333	806
264988.. 265129	98,33333	94
688284.. 688461	133,6667	119
800849.. 801089	183	169
1245846.. 1246050	149	138
1376855.. 1376984	91	84
1410339.. 1410537	137,3333	43,9
complement(153663.. 158316)	86,66667	3096
complement(1646557.. 1646821)	143	53,3

APÊNDICE 4 – GENES COM SOBREPOSIÇÃO EM FASE DE LEITURA NO GENÔMA *Burkholderia mallei* SAVP1 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR *BIT-SOCRE* DO SILA**	**VALOR *BIT-SCORE* DO BLAST**
491647.. 491851	155,6667	133
686892.. 687123	127,6667	120
798044.. 799712	813	828
848871.. 849123	101,6667	95,9
1279435.. 1281925	1801,333	1585

123224.. 123476	191,3333	166
242401.. 242506	79	67,4
316031.. 317033	106	102
439726.. 441187	1028	955
491647.. 491851	155,6667	132
553687.. 553975	208	186
848871.. 849123	101,6667	95
868462.. 868624	22	31,6
918830.. 921131	51	57,8
1095064.. 1095298	21	59,7
1279435.. 1281925	1801,333	1585
1340227.. 1341649	52,33333	39,7
1464128.. 1464893	562,6667	521
1610523.. 1610643	63,66667	47,8
complement(516639.. 517455)	112	3,5
complement(1332624.. 1333059)	137,6667	160
complement(979442.. 979877)	266,6667	289

APÊNDICE 5 – GENES NOVOS ENCONTRADOS NO GENOMA *Cyclobacterium marinum* DSM 745 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR *BIT-SCORE* DO BLAST**	**GENOMA DE REFERÊNCIA**
584003..584656	213	cytochrome-c peroxidase [*Dyadobacter fermentans* DSM 18053]
677806..678099	58,9	putative transcriptional regulator [*Mucilaginibacter paludis* DSM 18603]
1774383..1775456	247	hypothetical protein Cycma_0840 [*Cyclobacterium marinum* DSM 745]
1887891..1888427	365	hypothetical protein Cycma_1596 [*Cyclobacterium marinum* DSM 745]

2544801..2545067	175	excinuclease ABC subunit C [*Cyclobacterium marinum* DSM 745]
2568783..2568986	97,1	Orn/DAP/Arg decarboxylase 2 [*Cyclobacterium marinum* DSM 745]
3891860..3892063	122	hypothetical protein Cycma_3309 [*Cyclobacterium marinum* DSM 745]
4058637..4059221	397	transposase IS4 family protein [*Cyclobacterium marinum* DSM 745]
4388583..4389935	548	iduronate-2-sulfatase [*Maribacter sp*. HTCC2170]
4535294..4536886	774	TonB-dependent receptor plug [*Dyadobacter fermentans* DSM 18053]
4942550..4943686	795	putative transposase [*Cyclobacterium marinum* DSM 745]
4943595..4943903	152	putative transposase [*Cyclobacterium marinum* DSM 745]

APÊNDICE 6 – GENES COM SOBREPOSIÇÃO EM FASE DE LEITURA NO GENÔMA *Cyclobacterium marinum* DSM 745 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR *BIT-SCORE* DO BLAST**	**GENOMA DE REFERÊNCIA**
9091..9279	31,2	unnamed protein product [*Tetraodon nigroviridis*]
268704..268889	Sem resultado	
299511..299708	33,1	hypothetical protein KAOT1_18542 [*Kordia algicida* OT-1]
981840..982007	34,3	conserved hypothetical protein [*Capsaspora owczarzaki* ATCC 30864]
2020868..2021290	36,6	5-methyltetrahydropteroyltriglutamate--homocysteine S-methyltransferase [*Megasphaera sp*. UPII 199-6]
complement(3492300..3492452)	31,2	DHHA1 domain protein [*Megasphaera sp*. UPII 199-6]
complement(3801439..3801603)	32,6	FHA domain-containing protein [*Cyanothece sp.* PCC 8802]
4130424..4130771	Sem resultado	
4155821..4155982	Sem resultado	
4156357..4156635	35,8	hypothetical protein ECA0083 [*Pectobacterium atrosepticum* SCRI1043]
complement(4608532..4608825)	36,2	unnamed protein product [*Oikopleura dioica*]
5455076..5455384	35,8	hypothetical protein MICPUN_50172 [*Micromonas sp.* RCC299]

5479762..5479935	33,2	hypothetical protein SYNPCC7002_A2695 [*Synechococcus sp*. PCC 7002]
5860001..5860084	Sem resultado	

APÊNDICE 7 – GENES NOVOS ENCONTRADOS NO GENOMA *Escherichia coli* str. K-12 substr. DH10B ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR *BIT-SCORE* DO BLAST**	**GENOMA DE REFERÊNCIA**
58474..59125	448	conserved hypothetical protein [*Escherichia coli* DH1]
62450..63136	470	fructose repressor [*Escherichia coli* RN587/1]
196792..196905	80,1	hypothetical protein ECH7EC4486_2746 [*Escherichia coli* O157:H7 str. EC4486]
213294..213482	125	putative transposase [*Escherichia coli* XH140A]
224428..224931	348	MbhA [*Escherichia coli*]
227571..227837	183	hypothetical protein E4_12150 [*Escherichia* sp. 4_1_40B]
246190..247320	773	RNA-directed DNA polymerase [*Escherichia coli* DH1]
262073..262594	353	transposase protein A [*Escherichia coli*]
262837..263442	422	putative transposase insK for insertion sequence element IS150 [Shigella flexneri 2a str. 2457T]
265284..265412	90,1	CP4-57 prophage; predicted protein [*Escherichia coli* str. K-12 substr. MG1655]
323343..323450	69,3	hypothetical protein EcE24377A_0387 [*Escherichia coli* E24377A]
328806..330266	965	outer membrane autotransporter barrel domain protein [*Escherichia coli* MS 146-1]
331777..332973	789	hypothetical protein [*Escherichia coli* str. K-12 substr. MG1655]
390467..390610	87,4	hypothetical protein EcSMS35_0473 [*Escherichia coli* SMS-3-5]
467428..467685	176	hypothetical protein ECSE_0527 [*Escherichia coli* SE11]
505030..505242	140	putative replication protein [*Escherichia coli* XH140A]
520089..520652	385	putative tail fiber assembly protein [*Escherichia coli* DH1]
521430..521615	127	hypothetical protein ECO7815_15338 [*Escherichia coli* O55:H7 str. 3256-97 TW 07815]
596586..596813	150	hypothetical protein [*Escherichia coli*]

596810..597373	386	predicted amidase [*Escherichia coli* O103:H2 str. 12009]
633349..633912	385	putative tail fiber assembly protein [*Escherichia coli* DH1]
634690..634875	127	hypothetical protein ECO7815_15338 [*Escherichia coli* O55:H7 str. 3256-97 TW 07815]
709846..710073	150	hypothetical protein [*Escherichia coli*]
710070..710633	386	amidase [*Escherichia coli* O103:H2 str. 12009]
788919..789776	581	H repeat-associated protein in rhsC-phrB intergenic region [*Escherichia coli* STEC_EH250]
1132088..1132297	139	hypothetical protein ECs5411 [*Escherichia coli* O157:H7 str. Sakai]
1134605..1135336	482	probable peroxidase b1017 - *Escherichia coli* (strain K-12)
1145563..1145691	89,4	hypothetical protein HMPREF9540_04559 [*Escherichia coli* MS 115-1]
1146027..1146686	410	diguanylate cyclase domain protein [*Escherichia coli* MS 116-1]
1146728..1147093	244	IS2 ORF1 [*Shigella sonnei* Ss046]
1147165..1147956	557	IS2 orfB [*Shigella boydii* ATCC 9905]
1148083..1148721	440	diguanylate cyclase domain protein [*Escherichia coli* STEC_S1191]
1259065..1260585	993	conserved hypothetical protein [*Escherichia coli* MS 187-1]
1260793..1261686	620	putative part of putative ATP-binding component of a transport system [*Escherichia coli* H736]
1263084..1263371	160	RecName: Full=Putative uncharacterized protein ycgI
1279516..1279665	101	hypothetical protein EschWDRAFT_0296 [*Escherichia coli* W]
1317108..1317326	151	hypothetical protein Ec53638_4583 [*Escherichia coli* 53638]
1339573..1340076	348	IS1 transposase B [*Escherichia coli* str. K-12 substr. MG1655]
1358341..1358550	130	hypothetical protein SentesT_29900 [*Salmonella enterica* subsp. enterica serovar Typhi str. M223]
1468241..1469209	666	predicted sugar transporter subunit [*Escherichia coli* str. K12 substr. W3110]
1513568..1513708	96,7	hypothetical protein ECDH1ME8569_1311 [*Escherichia coli* DH1]
1521799..1522815	708	IS5 transposase and trans-activator [*Escherichia coli* str. K-12 substr. MG1655]
1554011..1556569	1659	conserved hypothetical protein [*Escherichia coli* DH1]
1559477..1562632	2082	EntS/YbdA MFS transporter [*Escherichia coli* XH140A]
complement(1578580..1578957)	254	glyceraldehyde 3-phosphate dehydrogenase protein [*Escherichia coli* MS 116-1]
1616509..1618557	1414	RecName: Full=Putative protein rhsE

1619469..1619951	330	transposase [*Escherichia coli* TA007]
1620070..1620195	85,9	hypothetical protein ECSE_0238 [*Escherichia coli* SE11]
1739158..1739682	Sem resultado	
1740755..1741693	612	predicted defective integrase [*Escherichia coli* str. K12 substr. W3110]
1800881..1801081	132	hypothetical protein ECSE_1756 [*Escherichia coli* SE11]
1809629..1809856	149	putative inner membrane protein [*Escherichia coli* BW2952]
1826227..1826451	155	predicted protein [*Escherichia coli* B354]
1826439..1827464	712	DNA-binding transcriptional repressor PurR [*Escherichia coli* O157:H7 str. EDL933]
1864182..1865777	1084	fused putative acetyl-CoA:acetoacetyl-CoA transferase: alpha subunit/beta subunit [*Escherichia coli* BW2952]
1892419..1893588	800	hypothetical protein b1721 - *Escherichia coli* (strain K-12)
1983400..1984761	943	para-aminobenzoate synthase subunit I [*Escherichia coli* ATCC 8739]
complement(2065938..2066453)	324	ranscriptional activator FlhC [*Escherichia coli* O104:H4 str. LB226692]
2123053..2123568	345	outer membrane porin truncated homolog b1964 precursor [similarity]
2123886..2124275	265	outer membrane porin truncated homolog b1966 [similarity] - *Escherichia coli* (strain K-12)
2128510..2129514	688	oxidoreductase, molybdopterin binding [*Escherichia coli* 101-1]
2159267..2159536	168	predicted disrupted hemin or colicin receptor [*Escherichia coli* str. K12 substr. W3110]
2159863..2160243	251	putative GTP-binding protein [*Escherichia coli* DH1]
2285504..2286328	568	WGR domain protein [*Escherichia coli* MS 116-1]
2286785..2288377	1075	molR_2 protein - *Escherichia coli* (strain K-12)
2288607..2289299	475	WGR domain-containing protein [*Escherichia coli* DH1]
2311020..2311178	106	conserved hypothetical protein [*Escherichia coli* H736]
2446826..2447011	122	conserved hypothetical protein [*Escherichia coli* MS 116-1]
2470618..2471547	621	RNase BN, tRNA processing enzyme [*Escherichia coli* str. K-12 substr. MG1655]
2471775..2472278	348	IS1 transposase B [*Escherichia coli* str. K-12 substr. MG1655]
2517593..2517754	94,4	hypothetical protein HMPREF9345_03164 [*Escherichia coli* MS 107-1]
2574161..2577754	2485	hybrid sensory histidine kinase in two-component regulatory system with EvgA [*Escherichia coli* BW2952]
2710768..2710965	95,1	conserved hypothetical protein [*Escherichia coli* MS 153-1]

2819464..2819601	93,6	conserved hypothetical protein, partial [*Escherichia coli* TA206]
2837674..2838540	579	hypothetical protein Z3905 [*Escherichia coli* O157:H7 str. EDL933]
2845946..2847187	853	P4-57 prophage; integrase [*Escherichia coli* str. K-12 substr. MG1655]
2876184..2876516	225	conserved hypothetical protein [*Escherichia coli* MS 196-1]
2876535..2877221	475	conserved hypothetical protein [*Escherichia coli* MS 116-1]
2877393..2878025	441	hypothetical protein EC_CP1639_04 [*Enterobacteria phage* CP-1639]
2881060..2882508	983	succinate-semialdehyde dehydrogenase I, NADP-dependent [*Escherichia coli* str. K-12 substr. MG1655]
2889921..2890262	228	hypothetical protein Z3972 [*Escherichia coli* O157:H7 str. EDL933]
2894485..2895624	760	glycine betaine/L-proline transport ATP binding subunit [*Escherichia coli* MS 107-1]
2895847..2896350	348	IS1 transposase B [*Escherichia coli* str. K-12 substr. MG1655]
2898880..2899146	175	predicted transporter [*Escherichia coli* str. K12 substr. W3110]
2899140..2900057	601	putative transport protein [*Escherichia coli* UMN026]
2901334..2901864	361	transcriptional repressor MprA [*Escherichia coli* O157:H7 str. Sakai]
2909820..2909936	80,1	hypothetical protein EcSMS35_2819 [*Escherichia coli* SMS-3-5]
2954157..2954795	437	predicted class II aldolase [*Escherichia coli* str. K-12 substr. MG1655]
2996121..2996261	91,7	hypothetical protein EcHS_A2919 [*Escherichia coli* HS]
2996662..2997147	332	ygcG [*Escherichia coli* E1520]
3061229..3061375	98,6	ypothetical protein NRG857_13983 [*Escherichia coli* O83:H1 str. NRG 857C]
3085831..3086307	328	putative peptidoglycan-binding-like protein [*Escherichia coli* IAI1]
3090617..3090754	94	hypothetical protein ECSTECDG1313_3894 [*Escherichia coli* STEC_DG131-3]
3152742..3154886	1476	methylmalonyl-CoA mutase, large subunit [*Escherichia coli* DH1]
3156694..3158172	1015	succinate CoA transferase [*Escherichia coli* DH1]
3170651..3171667	708	IS5 transposase and trans-activator [*Escherichia coli* str. K-12 substr. MG1655]
3174834..3174977	96,7	hypothetical protein HMPREF9536_04465 [*Escherichia coli* MS 84-1]
3181375..3182769	939	sugar transporter [*Escherichia coli* 'BL21-Gold(DE3)pLysS AG']
3183438..3184145	494	deoxyribonuclease I [*Escherichia coli* 'BL21-Gold(DE3)pLysS AG']
3198757..3199506	484	transport of nucleosides [Shigella dysenteriae Sd197]

3199576..3200784	833	transposase [Plasmid R100]
3200781..3203747	2054	conserved domain protein [*Escherichia coli* MS 116-1]
3206327..3207169	570	putative general secretion pathway protein L [Salmonella enterica subsp. enterica]
3212036..3213244	832	transposase [Plasmid R100]
3463594..3464721	778	conserved protein [*Escherichia coli* str. K-12 substr. MG1655]
3523049..3523186	93,6	conserved hypothetical protein, partial [*Escherichia coli* TA206]
3638934..3639812	605	predicted transposase [*Escherichia coli* str. K-12 substr. MG1655]
3719640..3719900	176	conserved hypothetical protein [Escherichia sp. 1_1_43]
3749299..3749481	124	hypothetical protein EcE24377A_3991 [*Escherichia coli* E24377A]
3765137..3765256	83,2	conserved hypothetical protein [Escherichia sp. 1_1_43]
3792226..3793797	1075	cellulose production protein [*Escherichia coli* str. K-12 substr. MG1655]
3814784..3814948	Sem resultado	
3850142..3850354	142	hypothetical protein SFV_3946 [Shigella flexneri 5 str. 8401]
3857783..3861916	2828	rhsA element core protein RshA [*Escherichia coli* str. K-12 substr. MG1655]
3865065..3865280	146	hypothetical protein ECs4472 [*Escherichia coli* O157:H7 str. Sakai]
3935683..3935817	Sem resultado	
3964966..3966048	753	putative protein CbrA [*Escherichia coli* MS 116-1]
4033843..4034772	524	transcriptional repressor RbsR [*Escherichia coli* UTI89]
4047503..4048486	668	acetolactate synthase II [*Escherichia coli* DH1]
4048638..4049147	353	acetolactate synthase II, large subunit, C-ter fragment, truncated protein [*Escherichia coli* DH1]
4050421..4052271	1268	dihydroxyacid dehydratase [*Escherichia coli* str. K-12 substr. MG1655]
4088174..4090642	1714	adenylate cyclase [*Escherichia coli* str. K-12 substr. MG1655]
4108199..4108918	487	putative inner membrane protein [*Escherichia coli* BW2952]
4166580..4167083	348	IS1 transposase B [*Escherichia coli* str. K-12 substr. MG1655]
4176518..4176676	86,2	no significant matches domain protein [*Escherichia coli* G58-1]
4190600..4190731	88,6	hypothetical protein HMPREF9346_03596 [*Escherichia coli* MS 119-7]
4373190..4374776	1095	predicted cyclic-di-GMP phosphodiesterase [*Escherichia coli* str. K-12 substr. MG1655]
4387425..4387928	348	IS1 transposase B [*Escherichia coli* str. K-12 substr. MG1655]

complement(4421046..4421666)	416	phnE [*Escherichia coli* str. K-12 substr. MG1655]
4535140..4535880	509	PAPS (adenosine 3'-phosphate 5'-phosphosulfate) 3'(2'),5'-bisphosphate nucleotidase [*Escherichia coli* str. K-12 substr. MG1655]
4551136..4551657	353	transposase protein A [*Escherichia coli*]
4551900..4552505	422	putative transposase insK for insertion sequence element IS150 [Shigella flexneri 2a str. 2457T]
4559384..4559686	204	cytochrome b(562) [*Escherichia coli* O157:H7 str. Sakai]
4601763..4601906	99,4	hypothetical protein ECIAI39_4746 [*Escherichia coli* IAI39]
4606737..4606940	132	predicted protein [*Escherichia coli* FVEC1412]
4618846..4619055	142	conserved hypothetical protein [Escherichia sp. 3_2_53FAA]
4625662..4625787	86,3	hypothetical protein HMPREF9348_05336 [*Escherichia coli* MS 145-7]
4634846..4635862	702	putative frameshift suppressor; KpLE2 phage-like element [*Escherichia coli* UMN026]

APÊNDICE 8 – GENES COM SOBREPOSIÇÃO EM FASE DE LEITURA NO GENÔMA *Escherichia coli* str. K-12 substr. DH10B ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR *BIT-SCORE* DO BLAST**	**GENOMA DE REFERÊNCIA**
16892..17032	69,7	ypothetical protein ECP_0016 [*Escherichia coli* 536]
109045..109305	86,3	putative transcriptional regulator [*Salmonella enterica* subsp. enterica serovar Typhi str. E98-2068]
236652..236759	Sem resultado	
596156..596404	166	CrcB [*Escherichia coli*]
709416..709664	166	CrcB [*Escherichia coli*]
complement(894478..894837)	43,1	glycoside hydrolase family 2 sugar binding [*Mucilaginibacter paludis* DSM 18603]
1852604..1852855	32,7	PREDICTED: uncharacterized protein LOC100811133 [*Glycine max*]
complement(1950559..1950729)	103	hypothetical protein SD1617_5381 [*Shigella dysenteriae* 1617]
2076476..2076814	229	conserved hypothetical protein [*Shigella sonnei* Ss046]
2369113..2369616	350	RecName: Full=Putative uncharacterized protein BicB

3376468..3376869	258	RecName: Full=Putative N-acetylgalactosamine permease IIC component 2; AltName: Full=EIIC-Aga'; AltName: Full=PTS system N-acetylgalactosamine-specific EIIC component 2
complement(3401728..3401943)	32,7	hypothetical protein Shew185_3298 [*Shewanella baltica* OS185]
3532285..3533661	925	potassium transporter peripheral membrane component [*Escherichia coli* O157:H7 str. EDL933]
3573734..3574276	316	GTG start codon, orf159 [*Escherichia coli*]
3680172..3680327	105	hypothetical protein ECIAI1_3589 [*Escherichia coli* IAI1]
3698249..3698521	38,1	hypothetical protein BuboB_13967 [*Burkholderia ubonensis* Bu]

APÊNDICE 9 – GENES NOVOS ENCONTRADOS NO GENOMA *Herbaspirillum seropedicae* SmR1 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR *BIT-SCORE* DO BLAST**	**GENOMA DE REFERÊNCIA**
complement(1236479..1236688)	33,9	phosphorylcholine phosphatase [*Verticillium albo-atrum* VaMs.102]
1997398..1997772	154	IISP family preprotein translocase auxillary membrane component [*Herminiimonas arsenicoxydans*]
4485213..4485821	407	0S ribosomal subunit protein L25 [*Herbaspirillum seropedicae*]

APÊNDICE 10 – GENES COM SOBREPOSIÇÃO EM FASE DE LEITURA NO GENÔMA *Herbaspirillum seropedicae* SmR1 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR *BIT-SCORE* DO BLAST**	**GENOMA DE REFERÊNCIA**
complement(43614..43865)	34,3	hypothetical protein MGG_14565 [*Magnaporthe oryzae* 70-15]
80743..80823	Sem Resultado	
203037..203966	48,5	hypothetical protein BCAL0353 [*Burkholderia cenocepacia* J2315]
complement(203117..203611)	Sem Resultado	

complement(219133..220704)	544	hypothetical protein Hsero_0205 [*Herbaspirillum seropedicae* SmR1]
612920..613150	33,9	conserved hypothetical protein [*Prevotella copri* DSM 18205]
840881..841798	Sem Resultado	
complement(1290436..1290693)	Sem Resultado	
1478929..1479987	520	hypothetical protein mma_1009 [*Janthinobacterium sp*. Marseille]
1532010..1532381	45,8	hypothetical protein BURPSS13_C0160 [*Burkholderia pseudomallei* S13]
1711358..1711651	33,9	onserved hypothetical protein [*Veillonella sp*. oral taxon 158 str. F0412]
1784256..1784774	36,6	diguanylate cyclase/phosphodiesterase [*Rhodopseudomonas palustris* HaA2]
1820970..1821320	34,7	hypothetical protein CLOSTASPAR_03270 [*Clostridium asparagiforme* DSM 15981]
complement(1876707..1877051)	54,7	hypothetical protein BamIOP4010DRAFT_5511 [*Burkholderia ambifaria* IOP40-10]
2028471..2028923	177	outer membrane lipoprotein [*Collimonas fungivorans* Ter331]
2029786..2029971	Sem Resultado	
complement(2051368..2051697)	36,2	hypothetical telomeric SfiI fragment 20 protein 3 [*Theileria parva*]
2277064..2277924	147	hypothetical protein Mmol_1216 [*Methylotenera mobilis* JLW8]
complement(2336626..2336976)	60,5	hypothetical protein SFxv_4749 [*Shigella flexneri* 2002017]
2388380..2388796	35,8	PREDICTED: RING finger protein 213 [*Bos taurus*]
2404986..2406029	415	hypothetical protein Daro_1870 [*Dechloromonas aromatica* RCB]
2557882..2558586	471	conserved hypothetical protein [*Ricinus communis*]
2579755..2580219	Sem Resultado	
3051166..3051795	91,3	hypothetical protein BamMEX5DRAFT_6990 [*Burkholderia ambifaria* MEX-5]
3238800..3239123	36,6	hypothetical protein CHLNCDRAFT_57702 [*Chlorella variabilis*]
3263078..3263302	Sem Resultado	
3368343..3368837	Sem Resultado	
3568644..3569210	203	hypothetical protein Hsero_3114 [*Herbaspirillum seropedicae* SmR1]
3671480..3672160	256	hypothetical protein IMCC9480_3649 [*Oxalobacteraceae bacterium* IMCC9480]
3838985..3839071	Sem Resultado	
4123783..4124112	35,8	ABC transporter, ATP-binding protein [*Prevotella disiens* FB035-09AN]
4159862..4160560	71,6	hypothetical protein Plav_2239 [Parvibaculum lavamentivorans DS-1]

4289947..4290501	34,7	hypothetical protein OB2597_13493 [*Oceanicola batsensis* HTCC2597]
4301706..4302164	39,7	hypothetical protein [*Podospora anserina* S mat+]
4336842..4337318	37,4	hypothetical protein SORBIDRAFT_02g021220 [*Sorghum bicolor*]
complement(4452686..4452910)	31,6	hypothetical protein [*Paramecium tetraurelia* strain d4-2]
4498107..4498427	35	hypothetical protein SINV_03763 [*Solenopsis invicta*]
4715676..4717007	602	hypothetical protein PFLU2000 [*Pseudomonas fluorescens* SBW25]
complement(4823138..4823518)	34,3	hypothetical protein TcasGA2_TC008372 [*Tribolium castaneum*]
5036223..5036411	33,9	hypothetical protein glr1449 [*Gloeobacter violaceus* PCC 7421]
5095532..5095843	39,3	hypothetical protein CRE_05854 [*Caenorhabditis remanei*]

APÊNDICE 11 – GENES NOVOS ENCONTRADOS NO GENOMA *Methanocaldococcus fervens* AG86 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR *BIT-SOCRE* DO SILA**	**VALOR *BIT-SCORE* DO BLAST**
159220.. 159721	246	294
158904.. 159176	105	135
358592.. 359135	166	276
528342.. 528513	107,6667	107
complement(982133.. 982993)	91,33333	92,8
complement(982972.. 984927)	398	392
complement(984927.. 986099)	258	246
complement(986224.. 986910)	192	188

APÊNDICE 12 – GENES COM SOBREPOSIÇÃO EM FASE DE LEITURA NO GENÔMA *Methanocaldococcus fervens* AG86 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR *BIT-SOCRE* DO SILA**	**VALOR *BIT-SCORE* DO BLAST**
324713.. 325016	17,33333	41,6
433501.. 433831	20,66667	Sem resultado
763725.. 763881	20,66667	Sem resultado
874848.. 874980	17,66667	32
1090988.. 1091315	27	38,1
complement(414250.. 414358)	88,66667	Sem resultado
complement(990977.. 991049)	21,66667	Sem resultado
complement(1293701.. 1293776)	60	Sem resultado

APÊNDICE 13 – GENES NOVOS ENCONTRADOS NO GENOMA *Pseudomonas fluorescens* pf 5 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR *BIT-SCORE* DO BLAST**	**GENOMA DE REFERÊNCIA**
82818..83108	190	cytochrome c [Pseudomonas brassicacearum subsp. brassicacearum NFM421]
210719..212467	877	von Willebrand factor, type A [*Pseudomonas fluorescens* Pf0-1]
764081..765394	587	precorrin-3B synthase [Pseudomonas brassicacearum subsp. brassicacearum NFM421]
934135..934989	445	Helix-turn-helix domain-containing protein [*Pseudomonas putida* BIRD-1]
935147..935521	242	DNA-binding protein [*Pseudomonas fluorescens* Pf-5]
971334..971810	318	RecName: Full=Ribosome maturation factor rimP
1304462..1305304	580	peptide chain release factor 2 [*Pseudomonas fluorescens* Pf-5]
1751313..1751525	111	transposase IS4 family protein [*Pseudomonas syringae* pv. japonica str. M301072PT]

1888933..1890144	653	cytochrome c-type biogenesis protein [*Pseudomonas fluorescens* Pf0-1]
1933365..1933841	272	phenylacetic acid degradation-like protein [*Pseudomonas fluorescens* Pf0-1]
1933838..1934290	262	hypothetical protein PFLU1829 [*Pseudomonas fluorescens* SBW25]
2243668..2244222	246	ultraviolet light resistance protein RulA [*Pseudomonas fluorescens* Pf-5]
2295008..2297608	956	motility protein [Pseudomonas brassicacearum subsp. brassicacearum NFM421]
2345493..2346008	73,9	hypothetical protein PFLU3398 [*Pseudomonas fluorescens* SBW25]
2473176..2473346	94,7	RelE/ParE family plasmid stabilization system protein [*Pseudomonas aeruginosa* PA7]
2618904..2619257	92,8	hypothetical protein PFLU2998 [*Pseudomonas fluorescens* SBW25]
2975872..2976165	142	hypothetical protein PFL_1012 [*Pseudomonas fluorescens* Pf-5]
2992710..2994809	1120	acyl-CoA synthetase [*Pseudomonas putida* KT2440]
3981679..3981948	92	site-specific recombinase [*Pseudomonas entomophila* L48]
4672344..4672838	143	LysE family transporter [*Cupriavidus metallidurans* CH34]
4810622..4810750	88,6	cell wall-associated hydrolase [*Vibrio cholerae* B33]
4829256..4831094	984	unnamed protein product [*Pseudomonas aeruginosa* LESB58]
4912198..4912419	144	hypothetical protein Pfl01_3963 [*Pseudomonas fluorescens* Pf0-1]
5022472..5022957	321	short chain dehydrogenase [*Pseudomonas fluorescens* Pf0-1]
5439350..5439766	684	hypothetical protein PputGB1_2994 [*Pseudomonas putida* GB-1]
5650742..5652934	1256	hypothetical protein PSEBR_a4519 [*Pseudomonas brassicacearum* subsp. brassicacearum NFM421]
5731166..5731420	131	prophage CP4-57 regulatory [*Pseudomonas putida* S16]
5732252..5732536	143	hypothetical protein PFWH6_0199 [*Pseudomonas fluorescens* WH6]
6009574..6009702	88,6	cell wall-associated hydrolase [*Vibrio cholerae* B33]
6277164..6278216	355	diguanylate cyclase [*Pseudomonas entomophila* L48]
6383394..6383522	88,6	cell wall-associated hydrolase [*Vibrio cholerae* B33]

APÊNDICE 14 – GENES COM SOBREPOSIÇÃO EM FASE DE LEITURA NO GENÔMA *Pseudomonas fluorescens* pf 5 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR *BIT-SCORE* DO BLAST**	**GENOMA DE REFERÊNCIA**
99941..101506	81,3	putative D-methionine ABC transporter, ATP-binding protein [*Streptomyces sp*. Tu6071]
547476..548792	40,4	hypothetical protein MGG_07681 [*Magnaporthe oryzae* 70-15]
1253944..1254684	258	thioesterase superfamily protein [*Pseudomonas fluorescens* Pf0-1]
1432267..1432665	57,8	hypothetical protein MEALZ_0705 [*Methylomicrobium alcaliphilum*]
1752646..1753101	140	hypothetical protein PFL_1557 [*Pseudomonas fluorescens* Pf-5]
complement(2578603..2579052)	137	hypothetical protein PFL_2019 [*Pseudomonas fluorescens* Pf-5]
2585899..2586201	36,2	peptidase, putative [*Pseudoalteromonas tunicata* D2]
2732164..2732769	37,4	PREDICTED: protein SHORT-ROOT-like [*Glycine max*]
2771993..2772385	94	hypothetical protein PSEEN0542 [*Pseudomonas entomophila* L48]
2815012..2815545	54,7	hypothetical protein CATMIT_01619 [*Catenibacterium mitsuokai* DSM 15897]
3242465..3243046	234	hypothetical protein Pfl01_2555 [*Pseudomonas fluorescens* Pf0-1]
3793240..3793899	45,4	ypothetical protein PSYCIT7_35427 [*Pseudomonas syringae* Cit 7]
3952545..3952970	37	NADH:flavin oxidoreductase [*Marinobacter sp*. MnI7-9]
3955594..3956022	119	hypothetical protein Psefu_2852 [*Pseudomonas fulva* 12-X]
4133440..4133922	47,4	hypothetical protein PFL_4188 [*Pseudomonas fluorescens* Pf-5]
complement(4287027..4287371)	36,6	PDR-like ABC transporter, putative, expressed [*Oryza sativa* Japonica Group]
4365352..4365819	58	hypothetical protein BpseB_41036 [*Burkholderia pseudomallei* B7210]
4378055..4378561	124	hypothetical protein CTS44_06238 [*Comamonas testosteroni* S44]
4792402..4792521	Sem resultado	
complement(4823298..4823558)	107	lipoprotein [*Pseudomonas fluorescens* Pf0-1]
5770949..5771485	36,2	hypothetical protein TGME49_062450 [*Toxoplasma gondii* ME49]
6088870..6089901	117	hypothetical protein PSYPI_40379 [*Pseudomonas syringae* pv. pisi str. 1704B]
6095031..6095375	38,9	Hypothetical protein [*Corynebacterium glutamicum* ATCC 13032]

6101804..6102160	36,6	Transporter, MFS superfamily [*Bacillus thuringiensis* serovar israelensis ATCC 35646]
6300978..6302588	73,9	hypothetical protein c0526 [*Escherichia coli* CFT073]
6361683..6362258	105	hypothetical Protein PANA_3622 [*Pantoea ananatis* LMG 20103]
6515347..6515832	40,4	psbO [*Microcystis aeruginosa* PCC 7806]
6878418..6879056	80,9	exonuclease III [*Xanthomonas oryzae* pv. oryzae KACC10331]

APÊNDICE 15 – GENES NOVOS ENCONTRADOS NO GENOMA *Ralstonia solanacearum* CFBP2957 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR *BIT-SCORE* DO BLAST**
42589..43170	395
44432..44626	116
81144..81578	277
81608..82198	378
421079..421231	86,7
950478..950732	134
964390..964827	214
1144920..1145333	265
1145197..1145595	216
1239900..1240256	218
1512696..1513061	211
1537057..1537272	139
1858471..1858998	268
1879692..1880591	441
2459617..2459805	115
2560884..2561420	196

2580977..2581729	503
2581823..2582062	168
2904153..2904413	119
3146741..3146956	624
3247221..3249294	1323
3249557..3249892	181

APÊNDICE 16 – GENES COM SOBREPOSIÇÃO EM FASE DE LEITURA NO GENÔMA *Ralstonia solanacearum* CFBP2957 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR *BIT-SCORE* DO BLAST**
complement(48927..49181)	33,5
198556..199062	35,4
268006..268464	37,7
complement(821094..821336)	34,3
860871..861311	35,8
1343700..1344083	37,4
1857103..1857240	85
2195072..2195611	Sem resultado
2560600..2560983	213
2695630..2696694	38,9
2753033..2753332	35,4
2929874..2930899	43,1
3272299..3272502	31,6
3410552..3410854	35,4

APÊNDICE 17 – GENES NOVOS ENCONTRADOS NO GENOMA *Streptococcus agalactiae* NEM316 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE* DO BLAST**	**GENÔMA DE REFERÊNCIA**
247762..247935	109	putative transporter [*Streptococcus agalactiae* 515]
457984..458118	87	hypothetical protein SAG0401 [*Streptococcus agalactiae* 2603V/R]
507764..508303	367	DNA integration/recombination/invertion protein [*Streptococcus pyogenes* MGAS6180]
614158..614415	177	transposase [*Streptococcus agalactiae* 18RS21]
614541..614750	145	ISPsy9, transposase OrfB [*Streptococcus agalactiae* H36B]
622249..622518	134	conserved hypothetical protein [*Streptococcus agalactiae* 18RS21]
652566..653135	386	sortase family protein [*Streptococcus agalactiae* 2603V/R]
656398..656517	78,6	hypothetical protein HMPREF9171_1434 [*Streptococcus agalactiae* ATCC 13813]
673222..673632	266	putative permease [*Streptococcus agalactiae* 515]
857993..858151	99	hypothetical protein SAG0813 [*Streptococcus agalactiae* 2603V/R]
909831..910067	152	hypothetical protein SAK_0988 [*Streptococcus agalactiae* A909]
953122..953265	89	conserved hypothetical protein [*Streptococcus agalactiae* 515]
1003514..1003867	242	nisin-resistance protein, putative [*Streptococcus agalactiae* CJB111]
1346913..1347293	258	IS861, transposase OrfB [*Streptococcus agalactiae* COH1]
1347337..1347597	182	transposase OrfB, IS3 family, truncation [*Streptococcus agalactiae* CJB111]
1357291..1358052	523	ISSag4, transposase orfB [*Streptococcus agalactiae* A909]
1624421..1624597	219	hypothetical protein SAG1492 [*Streptococcus agalactiae* 2603V/R]
1861064..1861198	84,3	hypothetical protein HMPREF9171_0481 [*Streptococcus agalactiae* ATCC 13813]
1900520..1900639	79,3	hypothetical protein HMPREF9171_0252 [*Streptococcus agalactiae* ATCC 13813]
2007798..2007947	78,2	hypothetical protein SAG1798 [*Streptococcus agalactiae* 2603V/R]
2150474..2150974	276	phage integrase family domain-containing protein [*Streptococcus anginosus* 1_2_62CV]

APÊNDICE 18 – GENES COM SOBREPOSIÇÃO EM FASE DE LEITURA NO GENÔMA *Streptococcus agalactiae* NEM316 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE* DO BLAST**	**GENOMA DE REFERÊNCIA**
197055..197138	Sem resultado	
778934..779104	33,5	hypothetical protein HMPREF9467_04030 [Clostridium clostridioforme 2_1_49FAA]
1100403..1100549	32	branched-chain amino acid transport ATP-binding protein [Renibacterium salmoninarum ATCC 33209]
1173291..1173557	32,3	hypothetical protein Phep_2013 [Pedobacter heparinus DSM 2366]
1506201..1506368	32,3	hypothetical protein GOPIP_019_00080 [Gordonia polyisoprenivorans NBRC 16320]
1589100..1590638	422	hypothetical protein SAL_1538 [Streptococcus agalactiae 515]
1694564..1694719	30,8	Hypothetical protein GL50581_2446 [Giardia intestinalis ATCC 50581]
1709123..1709440	203	hypothetical protein SAN_1714 [Streptococcus agalactiae COH1]
1868887..1869042	Sem resultado	

APÊNDICE 19 – GENES NOVOS ENCONTRADOS NO GENOMA *Streptococcus mutans* UA159 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE* DO BLAST**	**GENÔMA DE REFERÊNCIA**
858919..859035	77,4	acetyltransferase [*Streptococcus mutans* NN2025]
1176307..1176609	147	hypothetical protein STRMA_1081 [*Streptococcus macacae* NCTC 11558]
1308023..1308691	464	transposase, ISSmu1 [*Streptococcus mutans* UA159]
1373026..1373340	215	hypothetical protein SmuNN2025_0661 [*Streptococcus mutans* NN2025]
1791293..1791631	224	hypothetical protein SmuNN2025_0242 [*Streptococcus mutans* NN2025]
1915262..1915399	89,4	hypothetical protein SmuNN2025_1787 [*Streptococcus mutans* NN2025]
1977973..1978122	100	50S ribosomal protein L33 [*Streptococcus thermophilus* LMG 18311]

APÊNDICE 20 – GENES COM SOBREPOSIÇÃO EM FASE DE LEITURA NO GENÔMA *Streptococcus mutans* UA159 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE* DO BLAST**	**GENÔMA DE REFERÊNCIA**
596428..596739	34,7	hypothetical protein Shewmr7_0132 [*Shewanella* sp. MR-7]
741927..742055	70,1	hypothetical protein SmuNN2025_1216 [*Streptococcus mutans* NN2025]
complement(883032..883331)	Sem resultado	
931808..931987	31,6	CDP-diacylglycerol synthase, putative [*Ixodes scapularis*]
complement(1030495..1030653)	Sem resultado	
1171256..1171435	31,6	hypothetical protein BRAFLDRAFT_87520 [*Branchiostoma floridae*]
1253188..1253409	Sem resultado	
1304084..1304179	88,2	transposase [*Streptococcus mutans* UA159]
1539999..1540253	36,2	PREDICTED: WD repeat-containing protein 75-like [*Anolis carolinensis*]
complement(1781174..1781269)	Sem resultado	
1790557..1790733	168	hypothetical protein SmuNN2025_0256 [*Streptococcus mutans* NN2025]
1791293..1791631	35	hypothetical protein TRIVIDRAFT_178072 [*Trichoderma virens* Gv29-8]

APÊNDICE 21 – GENES NOVOS ENCONTRADOS NO GENOMA *Streptococcus pneumoniae* Hungary19A 6 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE* DO BLAST**	**GENÔMA DE REFERÊNCIA**
26900..27061	107	hypothetical protein SPAP_0029 [*Streptococcus pneumoniae* AP200]
98058..98573	270	lytic amidase [*Streptococcus pneumoniae* GA11426]

98699..99013	211	lytic amidase (N-acetylmuramoyl-L-alanine amidase) [*Streptococcus pneumoniae* CDC3059-06]
157122..157661	372	hypothetical protein spr0072 [*Streptococcus pneumoniae* R6]
157728..158138	280	hypothetical protein spr0074 [*Streptococcus pneumoniae* R6]
219764..220144	242	hypothetical protein CGSSp9BS68_03268 [*Streptococcus pneumoniae* SP9-BS68]
277068..277211	98,6	hypothetical protein spr0181 [*Streptococcus pneumoniae* R6]
complement(301034..301174)	87,4	iron(III) ABC transporter, permease protein [*Streptococcus pneumoniae* SP9-BS68]
323553..323717	94,4	ypothetical protein SpneCMD_03115 [*Streptococcus pneumoniae* str. Canada MDR_19F]
420048..420359	214	Transposase, uncharacterized, truncation [*Streptococcus pneumoniae* SP18-BS74]
420628..420879	147	hypothetical protein HMPREF0837_11111 [*Streptococcus pneumoniae* TCH8431/19A]
420950..421108	105	transposase-like protein [*Streptococcus pneumoniae* GA19077]
442677..442871	127	hypothetical protein SPAR74_0357 [*Streptococcus pneumoniae* GA41688]
450588..450764	122	hypothetical protein SPAR31_0417 [*Streptococcus pneumoniae* GA13494]
451198..451842	403	choline binding protein G, truncation [*Streptococcus pneumoniae* R6]
451935..452063	85,5	choline binding protein G [*Streptococcus pneumoniae* R6]
495269..495418	106	glycosyl transferase, family 8 [*Streptococcus pneumoniae* GA11184]
complement(528483..529010)	348	transposase [*Streptococcus pneumoniae* 670-6B]
578591..578947	242	hypothetical protein CGSSp6BS73_05270 [*Streptococcus pneumoniae* SP6-BS73]
590829..590990	92	hypothetical protein [*Streptococcus pneumoniae*]
594091..594324	152	hypothetical protein SP_0546 [*Streptococcus pneumoniae* TIGR4]
614698..615843	784	hypothetical protein SPCG_0538 [*Streptococcus pneumoniae* CGSP14]
616059..616601	371	hypothetical protein SPCG_0540 [*Streptococcus pneumoniae* CGSP14]
616817..617257	298	hypothetical protein SPCG_0541 [*Streptococcus pneumoniae* CGSP14]
674275..674400	85,5	hypothetical protein HMPREF0837_10925 [*Streptococcus pneumoniae* TCH8431/19A]
685082..685252	75,9	Glycero-transferase [*Streptococcus pneumoniae* INV200]
686432..686644	143	hypothetical protein SPAR91_0825 [*Streptococcus pneumoniae* GA47283]
686787..686972	79	hypothetical protein CGSSp6BS73_11426 [*Streptococcus pneumoniae* SP6-BS73]
689746..689937	122	hypothetical protein CGSSp6BS73_02535 [*Streptococcus pneumoniae* SP6-BS73]

698336..698944	412	hypothetical protein SPCG_0607 [*Streptococcus pneumoniae* CGSP14]
700842..701312	306	putative membrane protein [*Streptococcus pneumoniae* GA41410]
701300..702301	666	putative membrane protein [*Streptococcus pneumoniae* GA41410]
731288..731821	326	hypothetical protein SPN23F_06190 [*Streptococcus pneumoniae* ATCC 700669]
731826..733307	991	hypothetical protein spr0601 [*Streptococcus pneumoniae* R6]
749085..749654	387	lactate oxidase [*Streptococcus pneumoniae* GA41688]
767638..768249	398	phosphosugar-binding transcriptional regulator, RpiR family protein [*Streptococcus pneumoniae* SP19-BS75]
780453..780776	220	transposase [*Streptococcus pneumoniae* GA07228]
780859..781257	268	mobile genetic element [*Streptococcus pneumoniae* JJA]
845512..845898	231	Transposase [*Streptococcus pneumoniae* SP18-BS74]
845876..846058	128	transposase family protein [*Streptococcus pneumoniae* TCH8431/19A]
849195..849611	291	IS630-Spn1, transposase Orf1 [*Streptococcus pneumoniae* CGSP14]
874001..874213	138	hypothetical protein SPAR87_0356 [*Streptococcus pneumoniae* GA47033]
complement(887309..887836)	351	transposase [*Streptococcus pneumoniae* GA04375]
complement(931820..931987)	65,1	peptidase S24-like family protein [*Streptococcus pneumoniae* GA41410]
941519..941629	76,6	hypothetical protein CGSSp6BS73_05940 [*Streptococcus pneumoniae* SP6-BS73]
complement(948435..948926)	309	transposase [*Streptococcus pneumoniae* Taiwan19F-14]
1009705..1010070	256	degenerative transposase [*Streptococcus pneumoniae* R6]
1021980..1022183	113	histidine triad protein [*Streptococcus mitis* SK1080]
1042447..1042680	158	Type I restriction-modification system methylation subunit [*Streptococcus pneumoniae* SP19-BS75]
1059052..1059333	193	hypothetical protein CGSSp14BS69_02901 [*Streptococcus pneumoniae* SP14-BS69]
1059438..1059632	128	IS1381, transposase OrfA [*Streptococcus pneumoniae* SP6-BS73]
complement(1074743..1075270)	350	Transposase, orf 2 [*Streptococcus pneumoniae* INV104]
1081244..1081915	449	mobile genetic element [*Streptococcus pneumoniae* Taiwan19F-14]
1082015..1082329	217	transposase (IS4 family) [*Streptococcus pneumoniae* CDC1087-00]
1119248..1119619	248	phosphopyruvate hydratase [*Streptococcus pneumoniae* SP18-BS74]
1119768..1119893	84,3	Phosphoserine phosphatase, truncation [*Streptococcus pneumoniae* SP9-BS68]

1123630..1123890	139	hypothetical protein SPAP_0788 [*Streptococcus pneumoniae* AP200]
1162320..1163963	1127	SNF2 family protein [*Streptococcus sanguinis* SK49]
1288884..1289582	481	LICD family protein [*Streptococcus pneumoniae* GA44511]
1442983..1443195	137	hypothetical protein HMPREF0837_11116 [*Streptococcus pneumoniae* TCH8431/19A]
1443245..1443427	122	Type II restriction endonuclease, uncharacterized, truncation [*Streptococcus pneumoniae* R6]
1586691..1587302	389	hypothetical protein SP70585_1652 [*Streptococcus pneumoniae* 70585]
1587473..1587745	175	hypothetical protein SPAR22_1643 [*Streptococcus pneumoniae* GA11304]
1664485..1664787	175	3-ketoacyl-(Acyl-carrier-protein) reductase [*Streptococcus pneumoniae* GA47388]
1678452..1678763	204	transcriptional activator, Rgg/GadR/MutR family protein [*Streptococcus pneumoniae* SP23-BS72]
1696537..1696848	206	periplasmic binding s and sugar binding domain of the LacI family protein [*Streptococcus pneumoniae* GA44500]
1721268..1721711	283	dicarboxylate/amino acid:cation (Na+ or H+) symporter (DAACS) family protein [*Streptococcus pneumoniae* SP9-BS68]
1761144..1761569	287	transcriptional regulator [*Streptococcus pneumoniae* CGSP14]
1842891..1843592	482	NAD-dependent epimerase/dehydratase family protein [*Streptococcus pneumoniae* SP19-BS75]
1843719..1843850	84,2	UDP-glucose 4-epimerase [*Streptococcus mitis* NCTC 12261]
1859201..1859719	356	ribonuclease III [*Streptococcus pneumoniae* GA41301]
1859710..1860003	196	transposase [*Streptococcus pneumoniae* TCH8431/19A]
1860000..1860527	350	transposase [*Streptococcus pneumoniae* GA47373]
1878246..1878764	354	ribonuclease III [*Streptococcus pneumoniae* GA41437]
1878755..1879513	491	IS1167, transposase [*Streptococcus pneumoniae* SP18-BS74]
1889947..1890465	354	transposase [*Streptococcus pneumoniae* 670-6B]
1890456..1890749	196	transposase [*Streptococcus pneumoniae* TCH8431/19A]
1890746..1891264	335	transposase [*Streptococcus pneumoniae* GA17227]
1915098..1915487	266	putative IS1381 transposase [*Streptococcus pneumoniae* GA44511]
1994137..1994397	161	hypothetical protein SPT_2011 [*Streptococcus pneumoniae* Taiwan19F-14]
2028016..2029311	850	putative IS1167 transposase [Streptococcus pneumoniae]
2046221..2046439	128	hypothetical protein smi_0187 [Streptococcus mitis B6]

2055222..2055467	166	hypothetical protein CGSSp3BS71_05184 [*Streptococcus pneumoniae* SP3-BS71]
2086245..2087183	143	hypothetical protein CGSSp23BS72_01442 [*Streptococcus pneumoniae* SP23-BS72]
2100046..2100498	312	ypothetical protein CGSSp6BS73_12286 [*Streptococcus pneumoniae* SP6-BS73]
2133917..2134114	125	IS1381, transposase OrfA [*Streptococcus pneumoniae* SP6-BS73]
2134059..2134493	290	mobile genetic element [*Streptococcus pneumoniae* str. Canada MDR_19F]
2134529..2134765	165	transposase (IS4 family) [*Streptococcus pneumoniae* CDC1087-00]
complement(2223998..2224807)	550	ABC transporter, ATP-binding protein [*Streptococcus pneumoniae* SP3-BS71]

APÊNDICE 22 – GENES COM SOBREPOSIÇÃO EM FASE DE LEITURA NO GENÔMA *Streptococcus pneumoniae* Hungary19A 6 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE* DO BLAST**	**GENÔMA DE REFERÊNCIA**
200755..200922	112	hypothetical protein CGSSp6BS73_10106 [*Streptococcus pneumoniae* SP6-BS73]
241673..241840	31,2	protein of unknown function DUF88 [*Desulfovibrio fructosovorans* JJ]
362704..362874	116	hypothetical protein SPAR123_0281 [*Streptococcus pneumoniae* 4027-06]
391442..391813	82,4	OrfC [Streptococcus pneumoniae]
460432..460665	154	hypothetical protein CGSSp23BS72_07066 [*Streptococcus pneumoniae* SP23-BS72]
464746..465042	37,4	strongly similar to elongation factor Ts (EF-Ts) [Candidatus Kuenenia stuttgartiensis]
509314..509499	81,6	hypothetical protein CGSSp11BS70_09170 [*Streptococcus pneumoniae* SP11-BS70]
946258..946407	93,2	hypothetical protein HMPREF0837_11572 [*Streptococcus pneumoniae* TCH8431/19A]
966278..966427	41,2	hypothetical protein HMPREF1042_0479 [Streptococcus constellatus subsp. pharyngis SK1060]
complement(1214835..1215041)	39,9	hypothetical protein BJ6T_82530 [Bradyrhizobium japonicum USDA 6]
1481198..1481521	35	type IV pilus assembly PilZ [Sulfurospirillum deleyianum DSM 6946]
1810887..1811117	44,3	hypothetical protein SpneCM_00672 [*Streptococcus pneumoniae* str. Canada MDR_19A]
complement(1965312..1965503)	118	hypothetical protein SpneC19_09726 [*Streptococcus pneumoniae* CCRI 1974M2]

APÊNDICE 23 – GENES NOVOS ENCONTRADOS NO GENOMA *Thermotoga maritima* MSB8 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE* DO BLAST**	**GENÔMA DE REFERÊNCIA**
291472..291750	179	hypothetical protein Tpet_0647 [*Thermotoga petrophila* RKU-1]
291764..292783	696	extracellular solute-binding protein family 1 [*Thermotoga maritima* MSB8]
399656..399880	135	alkylhydroperoxidase like protein, AhpD family [*Thermotoga maritime* MSB8]
409601..409858	45,1	hypothetical protein CTN_0086 [Thermotoga neapolitana DSM 4359]
701680..703326	1077	flagellar hook-basal body protein [*Thermotoga maritima*MSB8]
703313..703549	160	flagellar hook-basal body protein [*Thermotoga maritima* MSB8]
895085..896029	615	ATPase AAA-2 domain protein [*Thermotoga maritima* MSB8]
896068..897462	917	ATPase AAA-2 domain protein [*Thermotoga maritima* MSB8]
1183460..1185085	1097	alpha-glucan phosphorylase [*Thermotoga petrophila* RKU-1]
1337505..1338665	785	Xenobiotic-transporting ATPase [*Thermotoga marítima* MSB8]
1349623..1349742	162	hypothetical protein Tpet_1440 [*Thermotoga petrophila* RKU-1]
1735874..1736434	377	Pyruvate/ketoisovalerate oxidoreductase [*Thermotoga marítima* MSB8]

APÊNDICE 24 – GENES COM SOBREPOSIÇÃO EM FASE DE LEITURA NO GENÔMA *Thermotoga maritima* MSB8 ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE* DO BLAST**	**GENÔMA DE REFERÊNCIA**
43828..44376	360	hypothetical protein ThemaDRAFT_1453 [*Thermotoga maritima* MSB8]
complement(346968..347735)	40	ABC transporter, multidrug resistance associated protein [*Chlamydomonas reinhardtii*]

473731..473958	32	AGAP009030-PA [*Anopheles gambiae* str. PEST]
497657..498232	40,4	hypothetical protein AURANDRAFT_37344 [*Aureococcus anophagefferens*]
623454..623735	181	hypothetical protein ThemaDRAFT_0709 [*Thermotoga maritima* MSB8]
667904..668422	36,2	Protein C54D10.4 [*Caenorhabditis elegans*]
668398..668964	37,4	predicted protein [*Physcomitrella patens* subsp. patens]
768503..768751	34,7	DHH family protein [*Clostridium botulinum* E1 str. 'BoNT E Beluga']
1019818..1020255	34,7	PREDICTED: major facilitator superfamily domain-containing protein 6-like [*Apis mellifera*]
1138411..1138704	33,1	hypothetical protein NCU05190 [*Neurospora crassa* OR74A]
1242181..1242708	67	137aa long hypothetical protein [*Pyrococcus horikoshii* OT3]
1317244..1317504	34,7	oligopeptidase A [*Endoriftia persephone* 'Hot96_1+Hot96_2']
1349623..1349742	Sem resultado	
1483609..1483971	35	ypothetical protein BATDEDRAFT_92259 [*Batrachochytrium dendrobatidis* JAM81]
1598433..1598681	32,7	aminotransferase, DegT/DnrJ/EryC1-family protein [*Desulfobacterium autotrophicum* HRM2]
1668641..1668910	34,4	TPA_inf: HDC02996 [*Drosophila melanogaster*]

APÊNDICE 25 – GENES NOVOS ENCONTRADOS NO GENOMA *Treponema pallidum* Nichols ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE* DO BLAST**	**GENÔMA DE REFERÊNCIA**
624365..625738	127	hypothetical protein TP0409 [Treponema pallidum subsp. pallidum str. Nichols]

APÊNDICE 26 – GENES COM SOBREPOSIÇÃO EM FASE DE LEITURA NO GENÔMA *Treponema pallidum* Nichols ENCONTRADOS PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE* DO BLAST**	**GENÔMA DE REFERÊNCIA**
6832..7074	159	putative ABC antibiotics transporter [*Treponema pallidum* subsp. pallidum str. Chicago]
134772..135653	40,4	DNA (cytosine-5-)-methyltransferase [*Bacteroides ovatus* SD CMC 3f]
151011..151853	40,4	DNA (cytosine-5-)-methyltransferase [*Bacteroides ovatus* SD CMC 3f]
155702..156379	265	hypothetical protein TPCCA_0133b [*Treponema paraluiscuniculi* Cuniculi A]
171783..172025	33,5	TonB-dependent receptor [*Brenneria sp*. EniD312]
196333..196635	34,3	hypothetical protein HMPREF9093_00469 [*Fusobacterium sp*. oral taxon 370 str. F0437]
429611..429853	166	putative ectonucleoside triphosphate diphosphohydrolase 6 [*Treponema pallidum* subsp. pallidum str. Chicago]
435074..435460	261	reprotein translocase subunit YajC [*Treponema paraluiscuniculi* Cuniculi A]
797592..797996	35	DNA-directed RNA polymerase II subunit two [*Leucogyrophana olivascens*]
915955..916515	36,6	PREDICTED: tRNA (uracil-5-)-methyltransferase homolog-B-like [*Oreochromis niloticus*]
934095..934304	143	hypothetical protein TPCCA_0856 [*Treponema paraluiscuniculi* Cuniculi A]
965519..966076	Sem resultado	
979939..980205	31,6	hypothetical protein MSWAN_1123 [*Methanobacterium sp*. SWAN-1]
1000781..1001605	560	onserved hypothetical protein [*Treponema pallidum* subsp. pallidum str. Chicago]
1024398..1024598	33,1	radical SAM domain-containing protein [*Syntrophobacter fumaroxidans* MPOB]
1036722..1037015	202	hypothetical protein TPCCA_0954a [*Treponema paraluiscuniculi* Cuniculi A]
1116680..1117093	36,2	sulfate permease family protein [*Porphyromonas endodontalis* ATCC 35406]
complement(1117139..1117276)	Sem resultado	

APÊNDICE 27 – GENES NOVOS ENCONTRADOS NO GENOMA *Treponema denticola* ATCC 35405 PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE* DO BLAST**	**GENÔMA DE REFERÊNCIA**
721437..721832	270	single-stranded DNA-binding protein [*Treponema denticola* F0402]
802042..804342	1497	serine protease Treponema denticola [*Treponema denticola* F0402]
884374..884595	147	hypothetical protein HMPREF9353_02308 [*Treponema denticola* F0402]
1008835..1009599	372	binding-protein-dependent transport system inner membrane component [*Treponema denticola* F0402]
1046385..1048073	734	ATP-dependent protease LA [*Treponema vincentii* ATCC 35580]
2093898..2094086	139	DNA mismatch repair protein [*Treponema denticola* F0402]

APÊNDICE 28 – GENES COM SOBREPOSIÇÃO EM FASE DE LEITURA NO GENÔMA *Treponema denticola* ATCC 35405 PELO PROGRAMA HGF E LOCALIZADOS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA SILA E VALIDADOS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE* DO BLAST**	**GENÔMA DE REFERÊNCIA**
275101..275427	35	glutamine--scyllo-inositol transaminase [Desulfomicrobium baculatum DSM 4028]
327930..328055	31,6	hypothetical protein TTHERM_00242550 [Tetrahymena thermophila]
559000..559074	sem blast	
complement(566782..566982)	32	XRE family transcriptional regulator [Burkholderia phytofirmans PsJN]
874843..875076	146	hypothetical protein HMPREF9353_02295 [Treponema denticola F0402]
974262..974354	sem blast	
complement(1195932..1196264)	35	poly-beta-hydroxybutyrate polymerase [Aeromonas hydrophila subsp. hydrophila ATCC 7966]
1452986..1453267	33,1	hypothetical protein HMPREF0988_01647 [Lachnospiraceae bacterium 1_4_56FAA]
1762389..1762679	37	FMN-binding domain-containing protein [Syntrophobacter fumaroxidans MPOB]
complement(1765377..1765460)	sem blast	

1847615..1847776	31,2	hypothetical protein SINV_04347 [Solenopsis invicta]
1875278..1875532	33,9	PREDICTED: similar to CG3168 CG3168-PA, partial [Hydra magnipapillata] Length=238
1886826..1887026	102	hypothetical protein HMPREF9353_01922 [Treponema denticola F0402]
complement(1918196..1918432)	34,7	F5/8 type C domain containing protein [Trichomonas vaginalis G3]
1940485..1940808	36,2	glycoside hydrolase clan GH-D [Paenibacillus sp. JDR-2]

APÊNDICE 29 – GENES NOVOS ENCONTRADOS NO GENOMA *Bradyrhizobium japonicum* USDA 110 PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE***
529253..529463	131
539055..539442	178
757372..757786	145
910182..910461	102
920756..921083	194
1201046..1201274	94,4
1446729..1446978	162
1796642..1797083	207
1798029..1798245	143
1844194..1844428	169
1855414..1855636	153
1989091..1989421	169
2023135..2023597	242
2069369..2069699	191
2237838..2238027	126
2337047..2337455	144

2379763..2380855	443
2632677..2632923	120
2677987..2678197	142
2769549..2769741	119
2786486..2786720	142
2821629..2821860	110
2848756..2848951	117
3445455..3445611	94,4
3574480..3574690	94,4
4187490..4187790	177
4250641..4250968	196
4280510..4280747	144
4285353..4285602	148
4332929..4333169	150
4413348..4413657	108
4634660..4635251	360
4639098..4639356	159
4677230..4677527	162
4859967..4860222	152
4865222..4865426	120
4949639..4950008	236
5014290..5014614	129
5022206..5022452	99
5044580..5044784	94
5044923..5045085	90,1
5145209..5145422	92,4
5185998..5186217	147
5509132..5509477	189

5724635..5724884	105
complement(5742815..5743460)	144
6330373..6330562	118
6352804..6353041	141
6750731..6750932	130
6753971..6754193	103
6787721..6787970	161
complement(7285423..7285657)	126
7342573..7342825	124
7761226..7761520	168
8168312..8168537	107
8168567..8168798	125
8365523..8366030	253
8405720..8406245	150
8445188..8445575	91,7
9041530..9041761	147
9103397..9103640	134

APÊNDICE 30 – GENES NOVOS ENCONTRADOS NO GENOMA *Burkholderia mallei* SAVP1 PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE***
4051..4324	178
208762..209932	763
264988..265129	92,8
800849..801089	169

1245846..1246050	135
complement(1668465..1668948)	94

APÊNDICE 31 – GENES NOVOS ENCONTRADOS NO GENOMA *Escherichia coli K 12* substr DH10B PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE***
58474..59125	399
213294..213483	122
227571..227838	155
246190..247321	736
262073..262595	343
262837..263443	411
331777..332974	767
467428..467686	176
505030..505243	137
520089..520653	379
521430..521616	126
596586..596814	90,5
596810..597374	356
633349..633913	379
634690..634876	126
709846..710074	90,5
710070..710634	356
788919..789777	572
1132088..1132298	137

1134605..1135337	417
1145563..1145692	90,5
1146027..1146687	346
1146728..1147094	237
1148083..1148722	424
1259065..1260586	979
1260793..1261687	591
1263084..1263372	163
1279516..1279666	99
1317108..1317327	155
1339573..1340077	349
1351842..1352037	90,1
1358341..1358551	127
1468241..1469210	631
1513568..1513709	95,9
1521799..1522816	687
1619469..1619952	267
1739158..1739683	197
1740755..1741694	585
1800881..1801082	133
1809629..1809857	130
1826227..1826452	158
1864182..1865778	1054
1892419..1893589	736
1983400..1984762	913
2123053..2123569	287
2123886..2124276	271
2128510..2129515	689

2159267..2159537	169
2159863..2160244	247
2285504..2286329	506
2286785..2288378	990
2288607..2289300	474
2311020..2311179	105
2446826..2447012	123
2470618..2471548	613
2471775..2472279	349
2517593..2517755	99
2710768..2710966	95,1
2819464..2819602	96,3
2837674..2838541	440
2845946..2847188	820
2876184..2876517	226
2876535..2877222	474
2877393..2878026	388
2881060..2882509	954
2889921..2890263	224
2894485..2895625	724
2895847..2896351	349
2898880..2899147	169
2899140..2900058	489
2901334..2901865	353
2954157..2954796	394
2996121..2996262	94,7
2996662..2997148	174
3085831..3086308	303

3090617..3090755	95,1
3170651..3171668	687
3174834..3174978	95,5
3181375..3182770	887
3183438..3184146	474
3198757..3199507	477
3199576..3200785	818
3200781..3203748	1986
3206327..3207170	526
3212036..3213245	819
3376468..3376870	170
3463594..3464722	721
3523049..3523187	96,3
3532285..3533662	884
3638934..3639813	585
3680172..3680328	105
3719640..3719901	173
3749299..3749482	128
3850142..3850355	139
3857783..3861917	2724
3865065..3865281	122
4033843..4034773	503
4047503..4048487	597
4050421..4052272	1220
4088174..4090643	1623
4166580..4167084	349
4190600..4190732	91,7
4373190..4374777	1006

4386964..4387198	148
4387425..4387929	349
4535140..4535881	469
4551136..4551658	343
4551900..4552506	411
4559384..4559687	198
4601763..4601907	99
4606737..4606941	135
4618846..4619056	145
4634846..4635863	674
complement(894478..894838)	139
complement (1578580..1578958)	91,7
complement (2067250..2068267)	111
complement (2913272..2914349)	99,4
complement (3127076..3127958)	93,6
complement (3958104..3959472)	114
complement (3973312..3975727)	93,6
complement (4368768..4371591)	102
complement (4421046..4421667)	97,1

APÊNDICE 32 – GENES NOVOS ENCONTRADOS NO GENOMA *Herbaspirillum seropedicae* SmR1 PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE***
523229..523409	100
3015403..3015601	109

3869895..3870360	107
4485213..4485822	370
complement(920849..921080)	99,4
complement(4992102..4992252)	91,7

APÊNDICE 33 – GENES NOVOS ENCONTRADOS NO GENOMA *Methanocaldococcus fervens* AG86 PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE***
157779..157986	105
158476..158740	99
158904..159177	95,5
159220..159721	293
358592..359135	286
398800..399202	205
528342..528513	105
1049502..1049982	273
1283792..1283981	112
complement(391331..391870)	355
complement(982972..984480)	293
complement(992328..992528)	97,1

APÊNDICE 34 – GENES NOVOS ENCONTRADOS NO GENOMA *Pseudomonas fluorescens* Pf-5 PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE***
82818..83109	193
206003..206585	123
206557..206944	108
210719..212468	796
934135..934990	432
935147..935522	231
971334..971811	266
980650..981211	159
1304462..1305305	527
1751313..1751526	109
1933365..1933842	268
1933838..1934291	253
1988240..1988591	137
2295008..2297609	721
2341690..2341924	126
2343240..2344092	267
2473176..2473347	99,8
2620440..2620812	103
2975872..2976166	105
2992710..2994810	998
3135628..3140014	802
3140224..3141334	438
3460629..3461034	97,8
3520227..3520932	328

3680941..3681349	166
3981099..3981510	214
3981679..3981949	91,7
4393528..4393771	113
4672344..4672839	144
4810622..4810751	94
4912198..4912420	107
5022472..5022958	286
5439350..5439767	175
5457455..5457977	164
5649811..5651101	467
5650742..5652935	1231
5688771..5689491	107
5731166..5731421	128
5731375..5731723	169
5732252..5732537	141
5737961..5738777	121
5940872..5941712	353
6009574..6009703	94
6083278..6083854	144
6277164..6278217	311
6279331..6280414	163
6383394..6383523	94
complement(4287027..4287372)	91,7
complement(4823298..4823559)	114
complement(6831056..6831482)	129
complement(6931758..6932352)	266

APÊNDICE 35 – GENES NOVOS ENCONTRADOS NO GENOMA *Ralstonia solanacearum* CFBP2957 PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE***	**GENE DE REFERÊNCIA**
42589..43170	396	transposase, IS256 family [*Ralstonia solanacearum* CFBP2957]
44432..44626	116	Hypothetical Protein RRSL_04194 [*Ralstonia solanacearum* UW551]
81144..81578	277	hypothetical protein RSIPO_00004 [*Ralstonia solanacearum* IPO1609]
81608..82198	378	hydrolase protein [*Ralstonia solanacearum* MolK2]
complement(868788..869180)	242	DNA-3-methyladenine glycosylase II [*Ralstonia solanacearum* UW551]
complement(870580..871107)	314	Hypothetical Protein RRSL_01228 [*Ralstonia solanacearum* UW551]
950478..950732	174	conserved hypothetical protein [*Ralstonia solanacearum* Po82]
964390..964827	214	conserved hypothetical protein [*Ralstonia syzygii* R24]
1144920..1145333	265	transposase [*Ralstonia solanacearum* CFBP2957]
1145197..1145595	216	isrso16-transposase orfb protein [*Ralstonia sp*. 5_7_47FAA]
1239900..1240256	218	glycosyl transferase, family 2; protein [*Ralstonia solanacearum* MolK2]
complement(1305330..1305587)	34,7	FAT domain-containing protein [*Glomerella graminicola* M1.001]
1512696..1513061	211	hypothetical protein RSc1491 [*Ralstonia solanacearum* GMI1000]
1537057..1537272	139	Hypothetical Protein RRSL_02186 [*Ralstonia solanacearum* UW551]
1858471..1858998	268	fad dependent oxidoreductase protein [*Ralstonia solanacearum* Po82]
2459617..2459805	115	hypothetical protein RRSL_04418 [*Ralstonia solanacearum* UW551]
2580977..2581729	503	Hypothetical cytosolic protein [*Ralstonia solanacearum* UW551]
3146741..3146956	62,4	protein of unknown function duf1328 [*Ralstonia solanacearum* MolK2]
3249557..3249892	181	hypothetical protein RS06029 [*Ralstonia solanacearum* GMI1000]

APÊNDICE 36 – GENES NOVOS ENCONTRADOS NO GENOMA *Streptococcus agalactiae* NEM316 PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE***
69256..69424	103
507764..508304	337
614158..614416	176
614541..614751	145
652566..653136	341
673222..673633	256
857993..858152	101
909831..910068	155
953122..953266	90,1
1003514..1003868	233
1346913..1347294	263
1347337..1347598	182
1357291..1358053	490
1624421..1624598	124
2150474..2150975	269

APÊNDICE 37 – GENES NOVOS ENCONTRADOS NO GENOMA *Streptococcus mutans* UA159 PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE***
complement(883032..883332)	93,6
923185..923320	90,5

1176307..1176610	142
1308023..1308692	449
1373026..1373341	237
complement(1734295..1734487)	126
1977973..1978123	101

APÊNDICE 38 – GENES NOVOS ENCONTRADOS NO GENOMA *Streptococcus pneumoniae* Hungary19A 6 PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE***
98058..98574	268
98699..99014	232
157122..157662	329
157728..158139	271
219764..220145	252
277068..277212	99,4
323553..323718	101
420048..420360	189
420628..420880	142
420950..421109	109
complement(421237..422191)	107
442677..442872	127
450588..450765	122
451198..451843	395
451935..452064	92,4
495269..495419	105

529058..529388	121
578591..578948	240
594091..594325	153
614698..615844	766
616059..616602	334
616817..617258	287
674275..674401	90,9
686432..686645	139
689743..689938	128
700842..701313	198
701300..702302	643
731288..731822	267
731826..733308	844
748519..749182	351
749085..749655	381
767638..768250	381
780453..780777	213
845876..846059	129
849195..849612	290
874001..874214	142
887690..887900	92
1009705..1010071	256
1021980..1022184	111
1042447..1042681	163
1059438..1059633	124
1081244..1081916	427
1082015..1082330	209
1119248..1119620	244

1123630..1123891	95,1
1157725..1162330	2798
1162320..1163964	1081
complement(1214835..1215042)	129
1288884..1289583	482
1442983..1443196	146
1443245..1443428	124
1586691..1587303	388
1587473..1587746	177
1664485..1664788	187
1677893..1678475	390
1678452..1678764	201
1696537..1696849	202
1721268..1721712	229
1761144..1761570	281
1842891..1843593	469
1859201..1859720	352
1859710..1860004	191
1860000..1860528	318
1878246..1878765	350
1878755..1879514	485
1889947..1890466	349
1890456..1890750	191
1890746..1891265	339
1915098..1915488	251
complement(1965312..1965504)	132
1994137..1994398	120
2028016..2029312	837

2055222..2055468	169
2086245..2087184	243
2100046..2100499	229
2133917..2134115	125
2134059..2134494	276
2134529..2134766	159
complement(2223998..2224808)	98,2

APÊNDICE 39 – GENES NOVOS ENCONTRADOS NO GENOMA *Thermotoga maritima* MSB8 PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE***
241493..241718	113
264809..265193	214
265947..266634	414
291472..291751	144
291764..292784	696
344045..344306	127
359486..359663	117
396013..396163	92
397237..397867	355
397878..398208	219
399656..399881	103
703313..703550	154
895085..896030	567
896068..897463	780

1024676..1025624	523
1182714..1183452	444
1183460..1185086	1108
1309337..1310105	470
1337505..1338666	752
1361809..1362058	168
1735874..1736435	333

APÊNDICE 40 – GENES NOVOS ENCONTRADOS NO GENOMA *Treponema denticola* ATCC PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE***
721437..721833	243
884374..884596	151
1008835.,1009600	333
1045669..1046401	414
1046385..1048074	699
1084846..1085185	102
2092294..2093242	618
2093898..2094087	132
complement(566782..566983)	99

APÊNDICE 41 – GENES NOVOS ENCONTRADOS NO GENOMA *Treponema pallidum* Nichols PELO PROGRAMA HGF E VALIDADAS ATRAVÉS DO PROGRAMA BOBBLES UTILIZANDO O ALINHAMENTO DE SEQUÊNCIAS PELO PROGRAMA BLASTP

LOCAL (CDS)	**VALOR DE *BIT-SCORE***
221129..221750	377
221860..222025	105
623682..624330	395
624365..625739	820
complement(331225..332320)	115
complement(577970..578171)	121
complement(729643..729820)	91,7
complement(945872..946820)	204