

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Willian Apolinário Lopes

**Análise sobre a predominância genética na
anotação de um organismo**

Uberlândia, Brasil

2017

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Willian Apolinário Lopes

**Análise sobre a predominância genética na anotação de
um organismo**

Trabalho de conclusão de curso apresentado
à Faculdade de Computação da Universidade
Federal de Uberlândia, Minas Gerais, como
requisito exigido parcial à obtenção do grau
de Bacharel em Sistemas de Informação.

Orientador: Anderson Rodrigues dos Santos

Universidade Federal de Uberlândia – UFU

Faculdade de Ciência da Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2017

Willian Apolinário Lopes

Análise sobre a predominância genética na anotação de um organismo

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Trabalho aprovado. Uberlândia, Brasil, 02 de Agosto de 2017:

Anderson Rodrigues dos Santos
Orientador

Professor

Professor

Uberlândia, Brasil
2017

Dedico este trabalho a minha família, a qual me deu o apoio necessário que para o mesmo fosse concluído com êxito, e a uma pessoa em especial, Rafaela Nadiele, a qual começou a fazer parte da minha vida sendo minha companheira em tudo, sempre me motivando e acreditando no sucesso deste trabalho.

Agradecimentos

Agradeço primeiramente a Deus por ter me dado mais essa vitória e também ao meu orientador Prof. Anderson Rodrigues, que foi paciente e guiou me em tudo no decorrer deste trabalho.

Resumo

A anotação é a primeira tarefa executada ao término da montagem de um genoma. De posse da sequência completa ou parcial de bases nitrogenadas de um organismo a anotação consiste em nomear funcionalmente cada uma das janelas abertas de leitura ou Open Reading Frame (ORF). A anotação utiliza-se de conhecimento sobre genes de outros organismos, preferencialmente de organismos evolutivamente próximos, para nomear ORFs desconhecidas. A decisão sobre qual gene será o responsável por nomear uma ORF de organismos com sequências gênicas recém-montadas é baseada com maior frequência na similaridade entre sequências proteicas, visto que a similaridade da sequência de DNA dificilmente é conservada mesmo entre organismos evolutivamente próximos. Uma vez anotado um genoma, passamos à fase seguinte no entendimento, por exemplo, dos mecanismos moleculares que levam um organismo a causar uma patologia. Entretanto, essa análise depende de uma anotação funcional de qualidade para que se possa tirar conclusões corretas sobre o modo de vida e infecção de um organismo patogênico. A necessidade de uma anotação de qualidade entre genomas evolutivamente próximos gerou a ferramenta PANNOTATOR, ferramenta que foi melhorada nesse trabalho de conclusão de curso. Na concepção do PANNOTATOR foi utilizada transferência de anotação funcional gênica, possuindo como origem um genoma completo e com qualidade de anotação que fosse evolutivamente próximo de um genoma de destino. Esse genoma foi denominado como mandatório. A transferência de anotação funcional partindo apenas de um genoma de origem mostrou-se mais eficiente do que o padrão antes utilizado de transferências que tinham como origem todos os genomas conhecidos. O motivo é a garantia de continuidade da qualidade de anotação, caso a origem possua tal qualidade.

Nesse trabalho de conclusão de curso o PANNOTATOR foi melhorado para poder incorporar a possibilidade de realizar a anotação funcional de genomas utilizando um segundo genoma como origem ou o genoma opcional de origem. Essa funcionalidade surgiu da demanda de especialistas em anotação ao levantarem a hipótese que um determinado genoma sob estudo estava sobre forte influência evolutiva de outros dois genomas. Para avaliar essa hipótese o PANNOTATOR foi alterado para receber opcionalmente um segundo genoma de origem de anotações funcionais e o PANNOTATOR deveria utilizar ambos para gerar uma anotação em um genoma recém-montado ou genoma de destino. Além disso, foi necessário realizar um teste estatístico para avaliar se determinados intervalos do genoma de destino foram mais influenciados pelo genoma mandatório ou pelo genoma opcional. O teste do Qui quadrado foi utilizado para definir se existia ou não diferença significativa entre as médias de genes anotados por ambos os genomas de origem em intervalos de tamanhos fixos no genoma de destino. Essa análise foi repetida variando-se os intervalos de tamanho no genoma de destino sob análise para 5 mil, 7.5 mil e 10 mil

bases com valores de similaridade entre proteínas (*cutoff*) de 50, 75 e 95% resultando em nove cenários para análises.

Para o caso específico do genoma de destino e seus dois genomas de origem (mandatório e opcional) essa análise não evidenciou regiões de maior predominância do genoma opcional. Esse resultado não corrobora a hipótese levantada pelos especialistas em montagem e análise do *Corinebacterium pseudotuberculosis* que o genoma opcional (linhagem 258) era evolutivamente determinante para o genoma de destino assim como o genoma mandatório (linhagem 316), ambas de uma bactéria que infecta equínos. Entretanto, essa análise proporcionou uma oportunidade de melhoramento do PANNOTATOR que agora possui meios de inferir estatisticamente a importância evolutiva entre dois genomas de origem para um genoma sob estudo. Essa análise pode ser feita considerando como intervalo de influência o genoma de destino como um todo ou intervalos menores customizáveis pelo usuário do PANNOTATOR.

Tal recurso é inédito dentre ferramentas de anotação atuais e abre possibilidades para a elaboração de um artigo revisado por pares em revistas internacionais lançando a versão 2.0 do PANNOTATOR.

Palavras-chave: Pannotator, anotação funcional gênica, bioinformática, Postgres, SQL

Lista de ilustrações

Figura 1 – Representação do Cromossomo e suas variantes (LOPES; ROSSO, 2005)	14
Figura 2 – Representação Gene (MINISTRY, 2015)	15
Figura 3 – Tabela periódica dos aminoácidos (JENKINS, 2015)	16
Figura 4 – Fluxograma Pannotator (SANTOS et al., 2013)	18
Figura 5 – Resultado da análise de acesso à ferramenta.	19
Figura 6 – Esquema relacional do banco de dados	24
Figura 7 – Arquivo EMBL com suas CDS em evidência	25
Figura 8 – Resultado consultas CDS	26
Figura 9 – Resultado Final da anotação com um genoma curado	27
Figura 10 – ExportAll.sql	27
Figura 11 – Cláusula FROM ExportAll.sql	28
Figura 12 – Resultados consulta mais interna ExportAll.sql	28
Figura 13 – Função retorna melhor resultado	29
Figura 14 – Novo ExportAll.sql	30
Figura 15 – Exemplo do resultado final desse trabalho de monografia	31
Figura 16 – Resultado total	31
Figura 17 – Dispersão dos valores anotados	36
Figura 18 – Predominância de anotação do genoma opcional	38

Lista de abreviaturas e siglas

PGH	Projeto Genoma Humano.
CDS	CoDing Segment - Define um segmento de leitura na fita de DNA (codon inicial - codon de parada).
RNA _m	RNA mensageiro.
RMP	Regiões de maior predominância.

Sumário

1	INTRODUÇÃO	11
1.1	Problema	12
1.2	Hipótese	12
1.3	Objetivo	12
1.3.1	Geral	12
1.3.2	Específicos	12
1.4	Justificativa	13
1.5	Organização do trabalho	13
2	REFERENCIAL TEÓRICO	14
2.1	Conceitos Biológicos	14
2.1.1	Cromossomo	14
2.1.2	Gene	14
2.1.3	Aminoácidos	15
2.1.4	Proteína	16
2.2	Banco de Dados	16
2.3	PANNOTATOR	17
2.3.1	Números da Ferramenta	19
2.4	Trabalhos Relacionados	20
2.4.1	Projeto de sequenciamento do genoma Humano - PGH	20
2.4.2	BG7 - Anotação de genoma bacteriano	20
2.4.3	Prokka: Anotação rápida de genoma procarioto	21
2.4.4	eCAMBer: Suporte eficiente para análise comparativa em grande escala de múltiplas cepas bacterianas	21
3	METODOLOGIA	23
3.1	Desenvolvimento	23
3.1.1	Processamento Pannotator	23
3.1.2	ExportAll	27
3.1.3	Função que retorna o melhor resultado	29
3.1.4	Novo ExportAll	30
4	RESULTADOS	31
4.1	Teste Qui-Quadrado X^2	31
4.1.1	Exemplo do uso do teste do qui-quadrado	33
4.2	Detalhamento do Resultado	35

4.3	Dispersão de valores	36
5	CONCLUSÃO	38
	REFERÊNCIAS	40
	 ANEXOS	 42
	ANEXO A – CUTOFF=50, INTERVALO=5000	43
	ANEXO B – CUTOFF=50, INTERVALO=7500	44
	ANEXO C – CUTOFF=50, INTERVALO=10000	45
	ANEXO D – CUTOFF=75, INTERVALO=5000	46
	ANEXO E – CUTOFF=75, INTERVALO=7500	47
	ANEXO F – CUTOFF=75, INTERVALO=10000	48
	ANEXO G – CUTOFF=95, INTERVALO=5000	49
	ANEXO H – CUTOFF=95, INTERVALO=7500	50
	ANEXO I – CUTOFF=95, INTERVALO=10000	51
	ANEXO J – MELHOR RESULTADO	52
	ANEXO K – ANÁLISE DO INTERVALO	54

1 Introdução

Desde que foi proposta uma estrutura molecular para o *DNA* em 1953, por [Watson, Crick et al. \(1953\)](#), pesquisadores não têm medido esforços para compreender a estrutura e o funcionamento da biologia molecular dos seres vivos ([LIMA et al., 2007](#)). A identificação de sequências biológicas e suas funções, em organismos, é denominada predição gênica, ou anotação de genoma.

Genoma é um conjunto de genes de um organismo, ou seja, uma longa sequência de informações genéticas que tem por função a geração de um produto de interesse da célula, como por exemplo: aminoácidos (moléculas orgânicas formadas por átomos de Carbono, Hidrogênio, Oxigênio e Nitrogênio) ([SANTOS, .](#)). Os aminoácidos são responsáveis por formar proteínas com funções específicas, que podem controlar uma característica, como por exemplo, a cor dos cabelos. A partir disto, vê-se a necessidade de desvendar informações por meio da anotação de genoma.

Anotação de genoma é o processo pelo qual são levantadas e rotuladas todas as características relevantes contidas em um genoma. Atualmente tem-se disponíveis servidores de anotação *online* e de livre acesso, portanto, há uma certa facilidade para realizar um sequenciamento e gerar uma nova anotação de um genoma ([TETTELIN et al., 2008](#)). Por exemplo, a ferramenta NBCI fornece o ***Prokaryotic Genomes Automatic Annotation Pipeline*** (PGAAP), ([SEEMANN, 2014](#)), um serviço via e-mail que de acordo com os dados inseridos, retorna o resultado em questão de dias. Uma outra ferramenta, o RAST ([AZIZ et al., 2008](#)), é um servidor web automatizado de anotação bacteriana, o qual identifica proteínas, rRNA e tRNA, atribui funções para os genes e prevê quais subsistemas estão representados no genoma anotado. Esta ferramenta retorna os resultados de 12 a 24 horas após a inserção dos dados.

Junto à estes *pipelines* de predição, existe também o PANNOTATOR *workbench* ([SANTOS et al., 2013](#)), uma ferramenta capaz de produzir uma anotação automática baseando-se em um genoma curado (genoma já anotado e publicado). O PANNOTATOR foi criado para reduzir a carga de trabalho exigida na geração de relatórios e correções das diversas anotações durante um projeto pan-genoma (quando tratados genomas de várias linhagens de uma mesma espécie de organismo). A mesma é composta por uma base de dados relacional em SQL, ferramentas interativas, e uma interface baseada na web. Inicialmente o ambiente de trabalho foi desenvolvido para gerenciar a bactéria *Corynebacterium pseudotuberculosis*, mas posteriormente, foi utilizado para gerir outras espécies bacterianas ([CARNEIRO et al., 2012](#)).

1.1 Problema

Com diversas ferramentas de anotação de genoma disponíveis e de livre acesso, torna-se eficaz a realização do procedimento de anotação (desde que com o devido conhecimento). O site do PANNOTATOR possui o diferencial de ser especializado em pan-genomas, mas ainda carece de algumas funcionalidades existentes em outros sites, como a adequação às necessidades impostas por novas tecnologias.

Uma destas necessidades, por exemplo, é a possibilidade de se anotar automaticamente diversos fragmentos de um genoma a partir de uma única submissão ao site de anotação. Esta necessidade é oriunda do fato que é difícil obter um genoma, por exemplo, de uma bactéria com uma única fita de DNA, sendo mais comum existirem regiões não decifradas do genoma e assim gerando um genoma em diversas partes.

Atualmente, o PANNOTATOR está limitado à entrada de somente um arquivo de Genoma curado, o qual é utilizado para fazer a anotação sobre a nova fita de DNA. Acredita-se que duas fitas de organismos curados podem ser mais apropriadas para a transferência de anotação do que um único organismo, de modo a permitir que a contribuição evolutiva seja quantificada em cada organismo curado sobre uma sequência gênica recentemente montada. A disponibilização de um teste estatístico que mensure o quanto cada genoma fonte participou da anotação de um determinado genoma é uma funcionalidade que seria útil em determinar parentescos evolutivos entre espécies.

1.2 Hipótese

Acredita-se que a anotação com dois genomas curados, em algumas situações, proverá o conhecimento a respeito de influências evolutivas entre organismos bacterianos, bem como anotações gênicas mais sintonizadas com os respectivos organismos de origem.

1.3 Objetivo

1.3.1 Geral

Este trabalho tem como objetivo geral abordar o uso das práticas de anotações automáticas de genomas para validar a hipótese sobre a melhoria dos resultados obtidos no processo de anotação, utilizando-se de dois genomas evolutivamente próximos.

1.3.2 Específicos

1. Alteração da quantidade dos arquivos de entrada do genoma curado, para uso de referência no processo, de um arquivo para dois arquivos.

2. Modificar o código de processamento do PANNOTATOR para que o mesmo possa ser capaz de analisar uma nova fita de DNA, a partir de referências de outras duas já conhecidas.
3. Realizar uma análise comparativa entre os resultados obtidos de uma anotação com apenas um genoma curado e de outra anotação com dois genomas curados, assim será possível dizer se a hipótese levantada é válida para o conjunto de organismos selecionados em particular.

1.4 Justificativa

A necessidade de se identificar a predominância de um organismo evolutivamente próximo no processo de anotação sobre uma sequência gênica recentemente montada. Realizada esta alteração no PANNOTATOR, será disponibilizado ao usuário junto com o resultado de anotação um relatório que mostra se existem regiões do genoma com maior predominância de algum dos genomas curados utilizados como referência.

1.5 Organização do trabalho

Serão utilizadas as seguintes tecnologias no desenvolvimento deste trabalho :

1. Linguagem de Scripts *Bash*, devido as facilidades oferecidas pelo Linux para diversas operações, o motor de processamento do Pannotator é implementado sobre esta linguagem juntamente com algumas funcionalidades em C.
2. Manipulação dos arquivos na base de dados feita através da ferramenta *PostgreSQL Administration Tool*, com comandos em PL/SQL para chamar os processos necessários.

Foi realizado a mudança nas quantidades de entradas do processo de sequenciamento de genoma, gerado um relatório do nível de predominância de cada entrada para a anotação final afim de prover um melhor conhecimento sobre as origens de um genoma, bem como uma melhor qualidade da anotação.

2 Referencial Teórico

2.1 Conceitos Biológicos

2.1.1 Cromossomo

O cromossomo é uma extensa sequência de DNA, a qual é composta por diversos genes e outras sequências de nucleotídeos que exercem funções específicas para as células dos organismos (LOPES; ROSSO, 2005). São compostos também por algumas proteínas chamadas de Histonas, que auxiliam durante a formação de um cromossomo.

Em organismos eucariotos (que possuem núcleo celular envolvido por uma membrana) os cromossomos estão localizados no interior do núcleo celular, como exposto na ??.

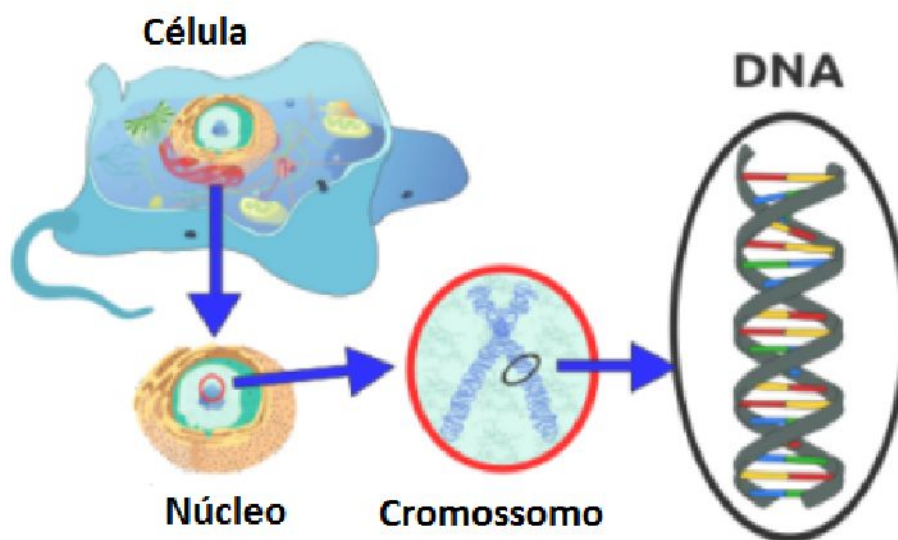


Figura 1 – Representação do Cromossomo e suas variantes (LOPES; ROSSO, 2005)

2.1.2 Gene

Gene é uma estrutura que contém informações genéticas de indivíduos e é transmitida de pais para filhos (MINISTRY, 2015). Estão localizados de forma linear no interior dos cromossomos, e não soltos dentro das células. São compostos por uma sequência específica de ácidos nucleicos responsáveis pelo armazenamento e transmissão da informação genética e por sua tradução, que é expressa pela síntese precisa das proteínas.

Atualmente, gene é um segmento de uma molécula de DNA que contém um código para a produção de alguma substância que tenha interesse para célula, como por exemplo os aminoácidos. Este segmento é composto por sequências codificantes, denominadas de éxons, e não-codificantes que são chamadas de íntrons.

Em organismos eucariotos, a partir de um processo chamado de *Splicing*, realizado sobre o gene, estas sequências são lidas e codificadas, são eliminados os íntrons, e uma nova fita que possui apenas sequências codificantes, os éxons, é gerada. Em organismos procariotos, não existem sequências não-codificantes em genes, sendo mais simples a sua decodificação, leitura e análises. A partir da fita de DNA, o processo de transcrição formará o RNA mensageiro (RNAm) final que será direcionado para o citoplasma para posterior tradução e produção de proteínas. A ?? ilustra como é um gene e onde o mesmo se encontra.

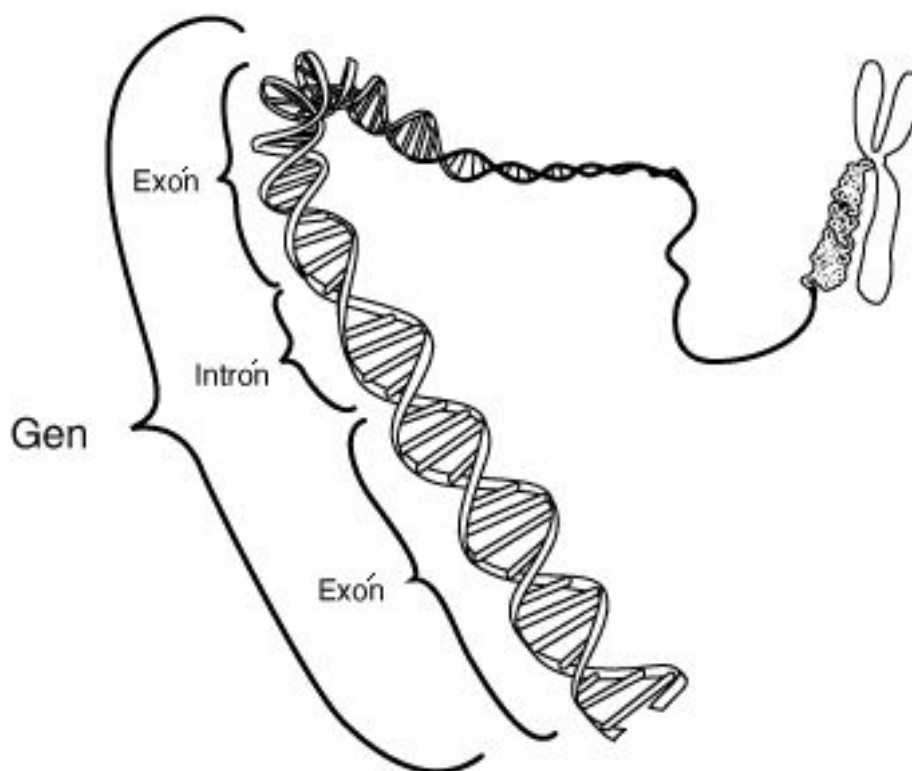


Figura 2 – Representação Gene (MINISTRY, 2015)

2.1.3 Aminoácidos

Uma fita de DNA é composta por sequências de nucleotídeos. Existem quatro nucleotídeos diferentes e cada um contém uma das quatro bases nitrogenadas, abreviadas por A (Adenina), G (Guanina), T (Timina) e C (Citosina), as quais combinadas resultam em uma última análise, as sequências de aminoácidos e proteínas (KLUG et al., 2010). Existem um total 20 aminoácidos diferentes, que combinados entre si sintetizam proteínas.

A [Figura 3](#) mostra quais aminoácidos cada combinação de nucleotídeos pode sintetizar. Percebe-se que nessa tabela não existe a Timina, mas Uracila em seu lugar. Isto ocorre porque aminoácidos são sintetizados a partir de RNAm mensageiro que foi transcrito em Uracila, ao invés da Timina da fita de DNA.

1º Base do Códon	2º Base do Códon				3º Base do Códon
	U	C	A	G	
U	Fenilalanina	Serina	Tirosina	Cisteína	U
	Fenilalanina	Serina	Tirosina	Cisteína	C
	Leucina	Serina	Parada(Stop)	Parada(Stop)	A
	Leucina	Serina	Parada(Stop)	Triptofano	G
C	Leucina	Prolina	Histidina	Arginina	U
	Leucina	Prolina	Histidina	Arginina	C
	Leucina	Prolina	Glutamina	Arginina	A
	Leucina	Prolina	Glutamina	Arginina	G
A	Isoleucina	Treonina	Asparagina	Serina	U
	Isoleucina	Treonina	Asparagina	Serina	C
	Isoleucina	Treonina	Lisina	Arginina	A
	Metionina	Treonina	Lisina	Arginina	G
G	Valina	Alanina	Ac. Aspártico	Glicina	U
	Valina	Alanina	Ac. Aspártico	Glicina	C
	Valina	Alanina	Glutamina	Glicina	A
	Valina	Alanina	Glutamina	Glicina	G

Figura 3 – Tabela periódica dos aminoácidos ([JENKINS, 2015](#))

2.1.4 Proteína

Proteínas são compostos orgânicos de alto peso molecular. São formadas pelo encadeamento de aminoácidos ligados entre si através de uma ligação peptídica (junção entre um grupo amino de um aminoácido e um grupo carboxila de outro aminoácido) ([ALBERTS et al., 2006](#)). Representam cerca de 50 a 80% do peso seco da célula sendo, portanto, o composto orgânico mais abundante de matéria viva.

As proteínas podem ser agrupadas em várias categorias de acordo com a sua função. De uma maneira geral, elas possuem as seguintes funções nos seres vivos: estrutural, enzimática, hormonal, de defesa, nutritiva, coagulação sanguínea e transporte. Uma molécula proteica pode conter desde algumas dezenas até milhares de aminoácidos. A molécula de hemoglobina, por exemplo, é formada por 574 aminoácidos.

2.2 Banco de Dados

O mundo nunca lidou com volumes de dados tão grandes, consequentemente vemos que Bancos de Dados estão cada vez mais ligados às aplicações da Tecnologia de Informação (TI), tornando-se parte fundamental na armazenagem de dados para qualquer tipo de atividades. Bancos de dados são coleções de dados ou registros relacionados, que podem

ser gravados e que possuem um significado implícito, representando aspectos do mundo real (GEREMIA, 2010).

A manipulação de bancos de dados existem softwares gerenciadores de bancos de dados (SGBD). Um SGBD consiste em uma coleção de programas que permitem ao usuário criar e manter um banco de dados. Alguns pontos importantes para seu uso são:

- Facilita os processos de definição, onde ocorre a especificação dos tipos de dados a serem armazenados e suas restrições.
- Otimiza o processo de construção de uma base de dados.
- Armazenamento dos dados em alguma mídia é controlada pelo SGBD.
- Manipula o banco de dados, na qual pode-se incluir várias funções, como por exemplo, a pesquisa para recuperar-se um dado específico, a atualização do banco de dados, a geração de relatórios de dados, e outras atividades de acordo com a necessidade do usuário.
- Permite o acesso de forma concorrente no processo de compartilhamento dos dados entre usuários.

Dentre os SGBD, temos o PostgreSQL (MILANI, 2008), que faz uso da linguagem SQL (*Structured Query Language*), desenvolvida e implementada pelo laboratório de pesquisa da IBM em San Jose no início da década de 70, para realizar tarefas relacionadas à manipulação de base de dados. Inicialmente chamada de SEQUEL (*Structured English Query Language*), foi criada com o intuito de interface entre usuários e o primeiro SGBDR (Sistema Gerenciador de Banco de Dados Relacional) SYSTEM R.

2.3 PANNOTATOR

Este trabalho é uma referência ao PANNOTATOR *workbench*, que foi desenvolvido com intuito de automatizar todo o processo de anotação e está disponível para uso em <http://pannotator.facom.ufu.br>. Esta ferramenta realiza todo o processamento dos arquivos necessários para a anotação com o uso de diferentes softwares. Sua entrada requer dois arquivos que devem ser passados pelo o usuário, a fita de DNA e o genoma curado. A partir deste ponto toda a predição gênica, transferência de dados entre os diferentes software e processo de anotação são realizados automaticamente pela ferramenta, o fluxo é mostrado na Figura 4.

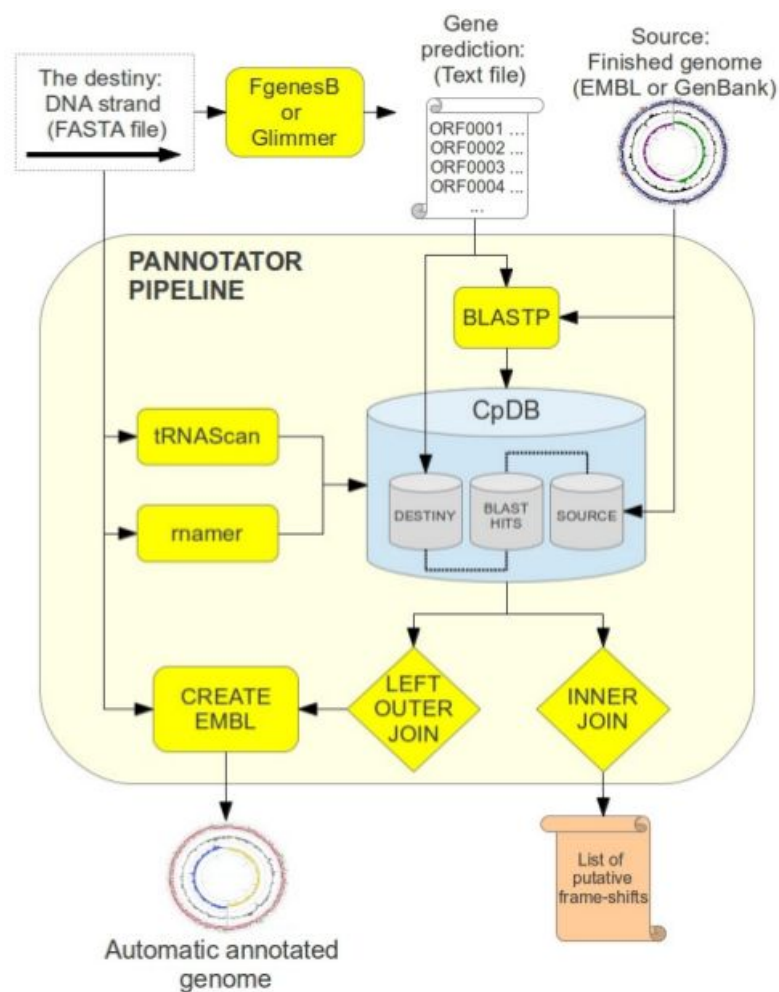


Figura 4 – Fluxograma Pannotator (SANTOS et al., 2013)

Como visto a entrada para o processo aceita apenas um genoma curado (*Finished genome*) como base para anotação, este arquivo é formatado e inserido seus valores na base de dados. O organismo a ser anotado é representado pelo arquivo *DNA strand FASTA FILE*, este é passado para as ferramentas *FgenesB* ou *Glimmer* para ser realizado a predição gênica (identificação das regiões passíveis de codificar algo de interesse para o organismo), após este processo os dados também são formatados e carregados na base de dados. É realizado uma comparação entre os resultados da predição gênica do organismo a ser anotado com as informações do genoma curado, processo feito pela ferramenta *Blast - Basic Local Alignment Search Tool* a qual faz a procura de regiões com similaridade entre sequências. Os resultados do processo de *Blast* também são armazenados na base de dados.

Com as informações necessárias na base de dados são realizadas consultas para mesclar estes dados e retornar um resultado que representa a anotação do organismo desejado. Neste capítulo será descrito todo o processo realizado pelo PANNOTATOR, de forma manual, para que a ferramenta faça à anotação com dois genomas curados (SANTOS et al., 2013).

2.3.1 Números da Ferramenta

Ao realizar-se uma análise (obtida a partir do *Google Analytics*) de acesso à ferramenta, é notável o aumento do uso da mesma a nível global. Uma melhoria no processo de anotação ajudará a destacar a ferramenta em meio a tantas outras. Na [Figura 5](#) é mostrado esta análise.

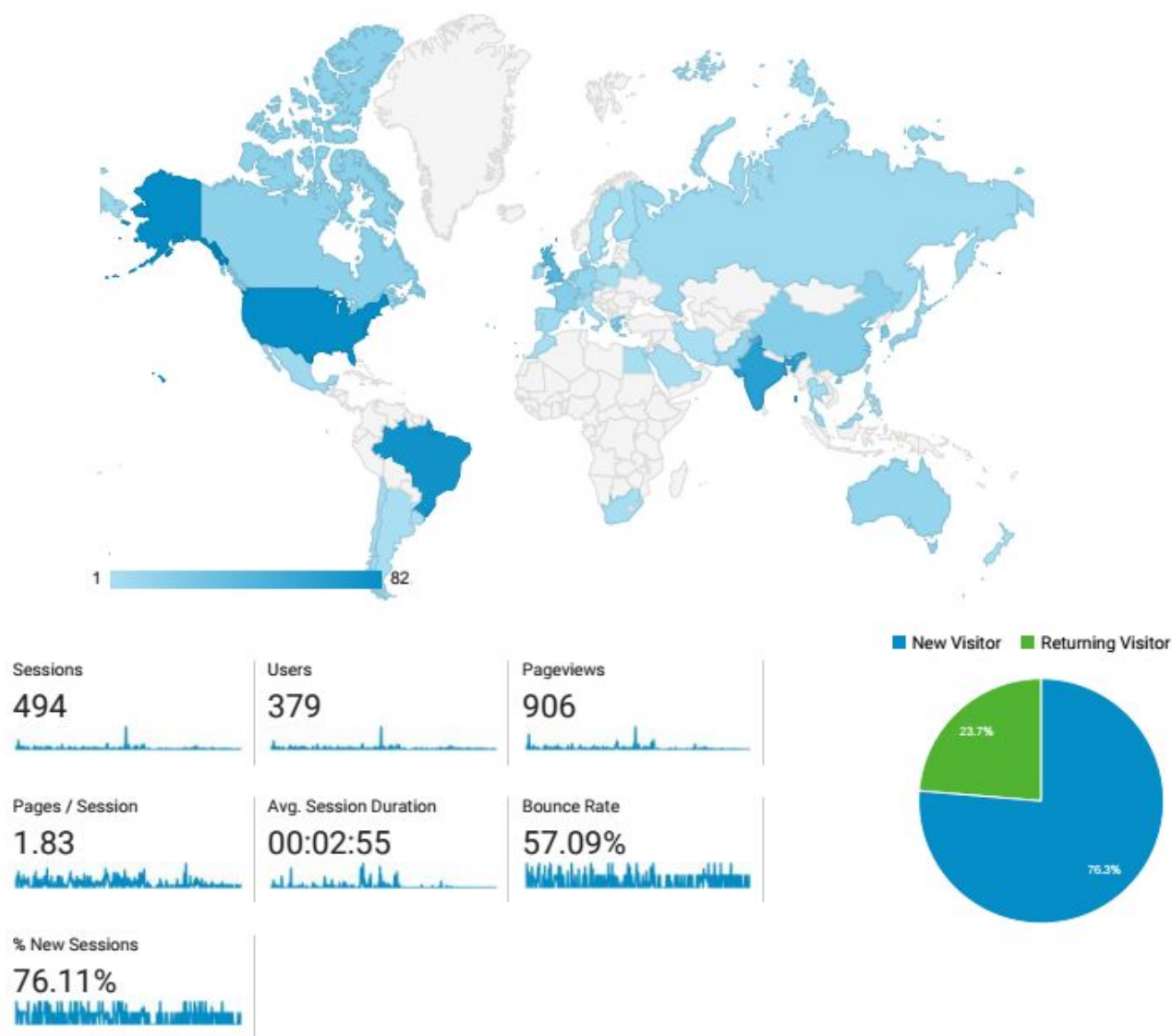


Figura 5 – Resultado da análise de acesso à ferramenta.

Os dados levantados são com base no intervalo de um pouco mais de dois anos, 17 de Julho de 2015 à 23 de Julho de 2017. Na média há um acesso a cada dois dias.

Outro dado que despertou a curiosidade neste trabalho foi de que o artigo [Santos et al. \(2013\)](#) que apresenta a ferramenta para a comunidade acadêmica, tem ganhado mais notoriedade pois, o número de citações ao texto está aumentando, o que contribui para a conclusão da importância de investimento na ferramenta.

2.4 Trabalhos Relacionados

2.4.1 Projeto de sequenciamento do genoma Humano - PGH

Oficialmente fundado em 1990, o Projeto Genoma Humano - PGH foi administrado pelo Departamento de Energia do Instituto Nacional de Saúde dos Estados Unidos por 13 anos. Inicialmente foi planejado para ser finalizado em 15 anos mas com o desenvolvimento acelerado da tecnologia, foi concluído no final de 2003 ([INTERATIVAS et al., 2013](#)). Alguns objetivos do projeto eram:

- Determinar a sequência dos cerca de 3,2 bilhões de pares de bases que compõem o genoma do *Homo sapiens*.
- Anotar e identificar todos os genes humanos.
- Armazenar toda essa informação em bancos de dados.
- Desenvolver ferramentas capazes de analisar a base de dados criada.

Ao término do PGH, foram sequenciados todos os genes, em torno de 25.000, constituindo pouco mais de 20% do material genético total humano ([ALBERTS et al., 2010](#)). Por limitações tecnológicas, cerca de 1% do genoma não pode ser sequenciado por possuir muitas repetições de bases nitrogenadas ([NHGRI et al., 2010](#)). De todos os genes sequenciados, aproximadamente 50% codificam proteínas de função conhecida.

O projeto possibilitou a descoberta de mais de 1800 genes de doenças, assim como facilitou a identificação de outros genes associados a doenças. Além disso, pelo menos 350 produtos biotecnológicos resultantes dos conhecimentos gerados pelo PGH estão sendo testados em laboratórios. As ferramentas desenvolvidas no PGH continuam servindo para levantar genomas de outros organismos importantes.

O PGH teve um grande impacto na biomedicina por ter levantando um grande número de genes associados à doenças importantes, tais como a doença de Huntington, esclerose lateral amiotrófica, tipos neurofibro-matosis 1 e 2, distrofia miotônica, e a síndrome do X frágil. Além de genes que conferem uma predisposição para doenças comuns, tais como câncer de mama, hipertensão, diabetes e doença de Alzheimer ([COLLINS; GALAS et al., 1993](#)).

2.4.2 BG7 - Anotação de genoma bacteriano

O BG7 é um sistema para anotação de genoma bacteriano baseado em uma abordagem própria para a anotação de genomas sequenciados, com tecnologias de sequenciamento para próxima geração. Um sistema versátil e capaz de anotar genes mesmo na

etapa de montagem preliminar do genoma, é especialmente eficaz na detecção de genes inesperados.

Desde as fases iniciais do processo de anotação do gene, o BG7 explora a enorme disponibilidade de sequências de proteínas já anotadas em bancos de dados. O BG7 prevê as ORFs e infere a sua função baseando-se na similaridade da proteína com o amplo conjunto de proteínas de referência, integrando previsão ORF e fases anotação funcional em apenas uma etapa.

O sistema é tolerante a erros durante o processo de sequenciamento e também ao elevado nível de fragmentação do gene, os quais são frequentemente encontrados em genomas que não estão completos (PAREJA-TOBES et al., 2012).

A versão atual do sistema é desenvolvida em Java, faz uso do Amazon Web Services (AWS), recurso de computação em nuvem, mas também pode ser executado localmente em qualquer sistema operacional. O BG7 é uma aplicação rápida, automatizada e escalável que pode lidar com o desafio de analisar as enormes quantidades de genomas que são sequenciados.

Sua capacidade e eficiência foi mostrada em 2011 em um evento na Alemanha, no qual o BG7 foi usado para obter-se as primeiras anotações no sequenciamento do genoma *entero-hemorrágica E.coli*. Este sistema está preparado para lidar com as novas tecnologias no futuro da Bio-informática.

2.4.3 Prokka: Anotação rápida de genoma procarioto

Procariotos são os mais antigos organismos da Terra em termos evolutivos, a maior parte de seu genoma está contida em uma grande molécula de DNA em fita dupla e forma de círculo, geralmente acompanhado de uma ou duas outras pequenas moléculas, também circulares: os plasmídeos.

Para estes tipos de organismos foi desenvolvido então o Prokka, uma ferramenta de software por linha de comando para anotar totalmente um genoma bacteriano em cerca de 10 minutos em um computador *desktop* qualquer (SEEMANN, 2014). Ela gera arquivos de saída compatíveis com os padrões para análise futura ou exibição em programas visualizadores de genoma.

2.4.4 eCAMBer: Suporte eficiente para análise comparativa em grande escala de múltiplas cepas bacterianas

Atualmente, em resultados de anotações de genomas bacteriano, são frequentemente encontradas inconsistências, que frequentemente não são refletidas na discrepância de suas sequências, mas são causadas por genes anotados de forma errada. Diante disto são

necessárias ferramentas para melhorar a consistência e precisão da anotação de genomas bacterianos (WOZNIAK; WONG; TIURYN, 2014).

eCAMBer é uma ferramenta que suporta com eficiência uma análise comparativa de genomas bacterianos dentro da mesma espécie. Foi desenvolvido a partir do Camber, a primeira versão da ferramenta, dimensionado para um conjunto de dados significativamente maior que compreende centenas de genomas bacterianos.

O processo do eCAMBer é dividido em duas fases, primeiro são transferidas anotações de genes entre todos os genomas bacterianos considerados, homologando as famílias de genes e as inconsistências da anotação. Na segunda parte, o eCAMBer procura melhorar a qualidade de anotações resolvendo as inconsistências iniciais do gene e filtrando as famílias de genes decorrentes de erros de anotação propagadas na fase anterior.

Deste modo, o eCAMBer é capaz de identificar de forma eficiente e resolver as inconsistências da anotação entre os genomas bacterianos estreitamente relacionadas. De acordo com seus resultados, ele supera outras ferramentas concorrentes, em termos de tempo e precisão das anotações produzidas na execução. O Software, manual do usuário e demonstrações dos resultados estão disponíveis no site do projeto: <http://bioputer.mimuw.edu.pl/ecamber>.

3 Metodologia

Os arquivos utilizados para experimentos neste trabalho foram:

Genomas curados:

- Mandatório: linhagem 258.
- Opcional: linhagem 316 (enviado por um especialista, o qual diz ser de grande importância).

Genoma a ser anotado: *Corynebacterium pseudotuberculosis*

- Linhagem MB20: uma bactéria com membrana única (gram-positiva), intracelular facultativo (coloniza as células, incluindo macrófagos do sistema de defesa) criando abscessos purulentos na carne e na pele de animais infectados. Causa perdas econômicas devido à inutilização de partes significativas de animais de corte. É problema em grandes rebanhos em países de clima tropical. No Brasil acomete mais cabras e ovelhas. Porém, infecta desde cavalos (EUA) até gado (Israel) em outros países.

3.1 Desenvolvimento

Nesta seção será especificado todo o processamento realizado pelo Pannotator, serão usadas algumas nomenclaturas para facilitar o entendimento do processo de anotação. Adotou-se o nome de "**origem**" para os arquivos dos genomas curados, ou seja, os arquivos que já foram anotados e já sabe o que cada um de seus genes sintetizam. Adotou-se também o nome de "**destino**" para a fita de DNA, a qual será anotada a partir dos dois arquivos curados inseridos.

3.1.1 Processamento Pannotator

Para exemplificar o fluxo de processos automatizados pela ferramenta, a anotação foi realizada manualmente, utilizando-se como parâmetro de *origem* o arquivo curado **NC016932.gbk**, e uma fita de DNA **MB20.fasta** do organismo *Corynebacterium pseudotuberculosis* - *Cp*, como *destino*.

Primeiramente é criada uma base de dados no PostgreSQL, a qual é responsável pelo armazenamento dos dados obtidos em todo o processo. Com a base criada, é inserido o esquema relacional para criar toda a estrutura de tabelas e seus relacionamentos.

A base em si possui um total de 11 tabelas para representar um genoma bacteriano. Dentre todas estas, as tabelas mais significativas neste trabalho são: *gene* (que contém as informações do genoma a ser anotado), *curated* (guarda os dados dos arquivos curados passados como referência) e *blast* (representa uma relação que cria um link entre gene e curated por meio de análises de similaridade proteica).

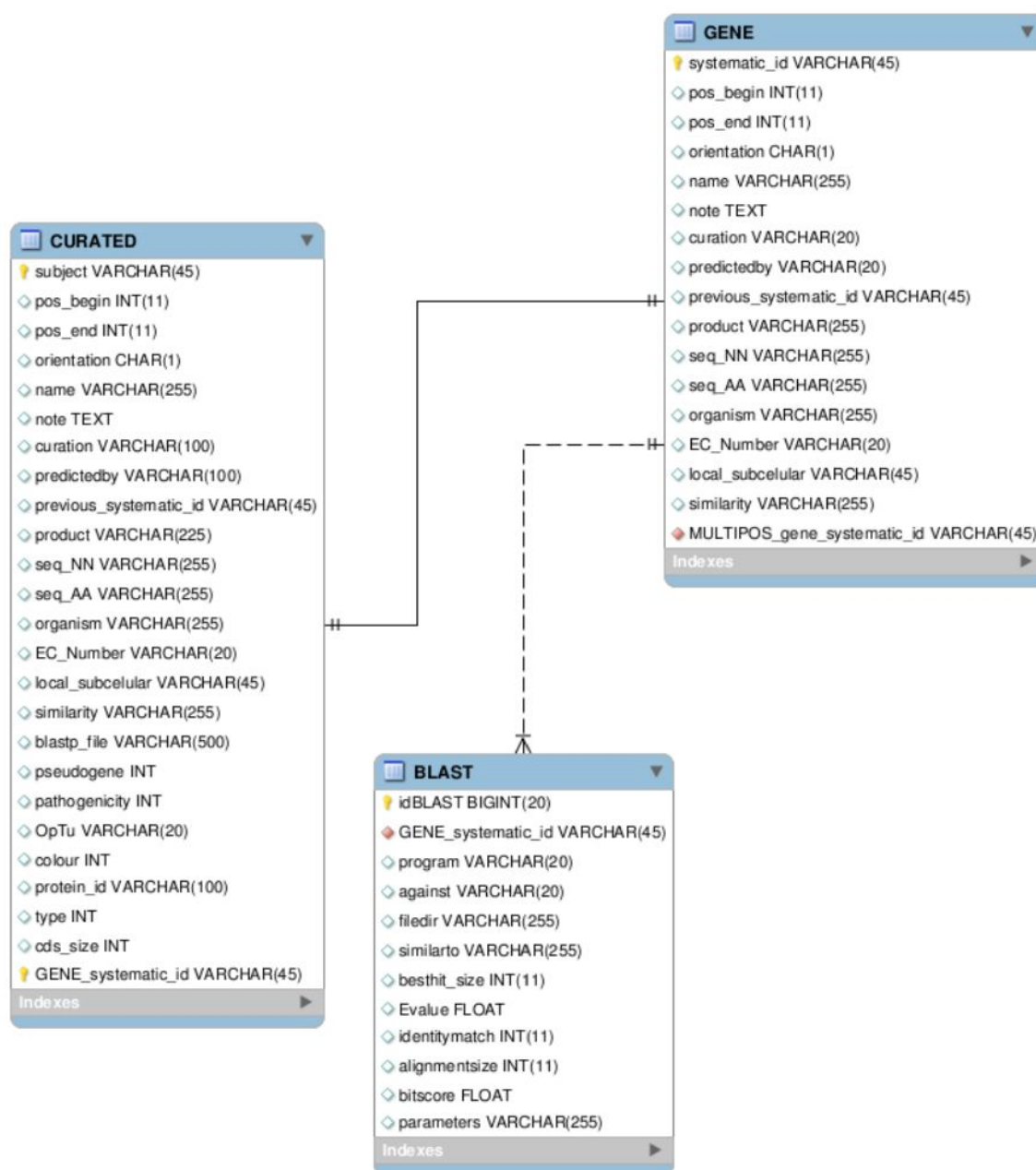


Figura 6 – Esquema relacional do banco de dados

Com a base de dados pronta, é realizada a predição gênica do arquivo de destino com a ferramenta Glimmer (DELCHER et al., 1999), um sistema para detecção de genes em DNA microbiano, especialmente os genomas de bactérias. A ferramenta está disponível para download através do site <https://ccb.jhu.edu/software/glimmer/>. Esta aplicação re-

cebe uma fita de DNA como entrada, e realiza a predição de acordo com o organismo mais próximo (do organismo que gerou a fita de DNA como entrada) conhecido, a ferramenta possui uma lista de organismos da qual deve ser selecionado um como referência, nos experimentos deste trabalho é utilizado o organismo *Corynebacterium diphtheriae* NCTC 13129. Ao fim deste processo, tem-se um arquivo que será formatado para informar as regiões que são codificantes, denominadas de CDS, na fita de *destino*.

É necessário gerar um arquivo **.embl** a partir da fita de DNA, para este processo é utilizado o programa *Artemis Genome Browser* (CARVER et al., 2008), uma ferramenta que facilita a manipulação dos arquivos de um genoma. Com o arquivo **.embl** criado, incorpora-se a ele o arquivo que contém as CDS da fita de *destino*, processo que também é realizado pelo Artemis. Assim é possível analisar visualmente as regiões codificantes na fita de DNA, como é mostrado na Figura 7.

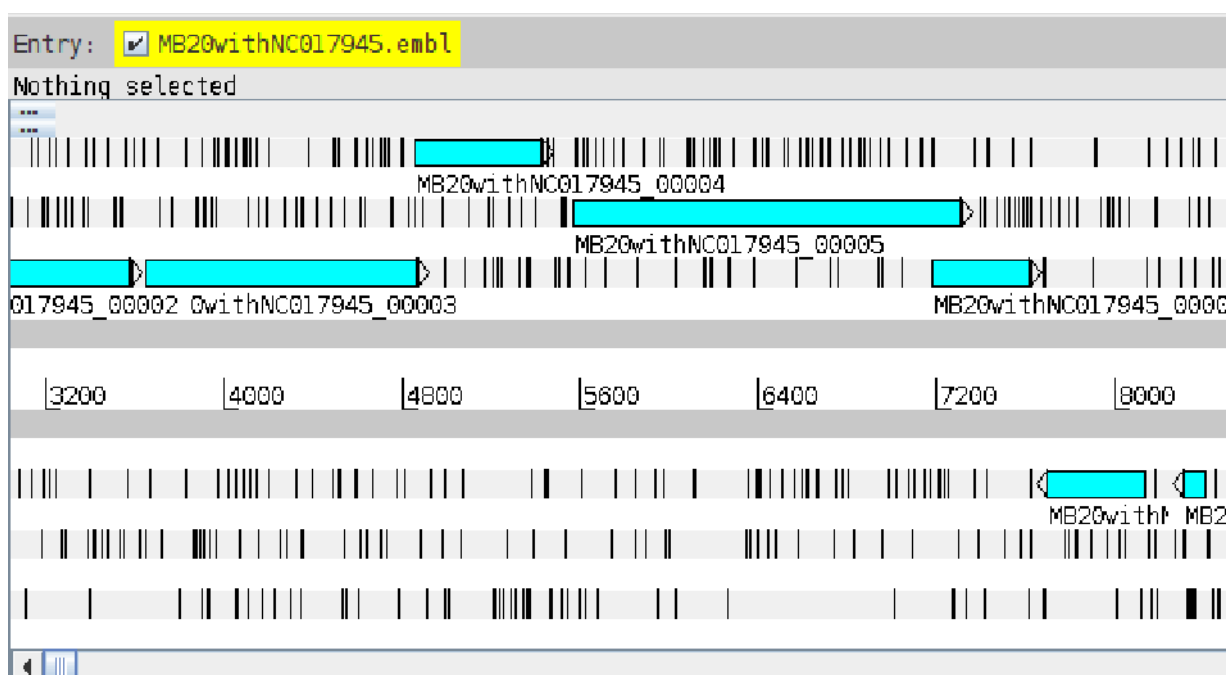


Figura 7 – Arquivo EMBL com suas CDS em evidência

Na Figura 7 vê-se o arquivo **.embl** da fita de *destino* com suas regiões codificantes nomeadas e em destaque, além disso é possível ver em cada CDS quais são os códons inicial e final da região.

Como cada região codificante encontrada no arquivo *destino* é capaz de sintetizar algum aminoácido, necessita-se da realização de uma comparação com às CDS (obtidas pelo mesmo processo) dos arquivos de *origem*.

Neste passo é utilizada a ferramenta *Blast - Basic Local Alignment Search Tool*, que faz a procura de regiões com similaridade entre sequências. O programa comparará as sequências de proteínas entre as CDS inseridas e calculará a significância estatística

dos resultados (TATUSOVA; MADDEN, 1999).

O arquivo resultante do processo *Blast* é utilizado para dar carga na tabela *blasthist* na base de dados. Estes dados serão incorporados na extração do resultado final de todo o processo.

Necessita-se de extrair-se alguns dados dos arquivos de *origem* e estes precisam ser formatados, isto é feito para dar carga na tabela *curated*, na qual são guardadas as informações dos arquivos de *origem*. Estes dados são obtidos por um executável que gera como resultado arquivos *.sql*, os quais são executados na base de dados para popular a tabela. Este mesmo processo é repetido para o arquivo *.embl* obtido da fita de DNA *destino*, os dados obtidos são usados para popular desta vez a tabela *gene*, esta contém dados do organismo que será anotado.

Com as tabelas devidamente populadas, executa-se uma consulta sql na qual realiza-se um **join**, entre as tabelas *gene*, *blasthist* e *curated*, para saber qual produto é sintetizado em cada região codificante do organismo de destino. Incorpora-se o resultado gerado ao arquivo *.embl* da fita de DNA inserida para anotação através do *Artemis*. Este resultado é mostrado na Figura 8.

```

1 FT      CDS      complement(1..1812)
2 FT      /locus_tag="organismMB20_00001"
3 FT      /gene="dnaA"
4 FT      /product="Chromosomal replication initiation protein"
5 FT      /colour=3
6 FT      /blastp_file="blastp/Cp19.embl.seq.00001.out"
7 FT      CDS      complement(2397..3584)
8 FT      /locus_tag="organismMB20_00002"
9 FT      /gene="dnaN"
10 FT     /product="DNA polymerase III subunit beta"
11 FT     /colour=7
12 FT     /blastp_file="blastp/Cp19.embl.seq.00002.out"

```

Figura 8 – Resultado consultas CDS

Algumas cores são utilizadas para exemplificar melhor cada região processada. A cor verde para as sequências de genes que alinham 100% (indicadas na tag *colour* com valor igual a 7). A cor amarela (*colour* igual a 3) para representar as regiões onde o percentual está entre um valor mínimo determinado e menor do que 100%. Os demais valores abaixo deste mínimo são preenchidos com a cor vermelha (*colour* igual a 2).

Após incorporar-se o resultado da consulta à fita de DNA, o sistema exibe cada CDS com suas respectivas informações, como mostrado na Figura 9.

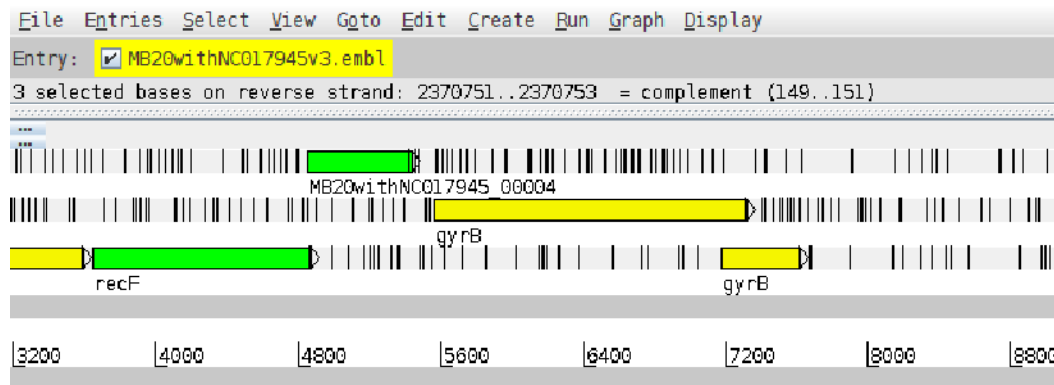


Figura 9 – Resultado Final da anotação com um genoma curado

3.1.2 ExportAll

Para validar a hipótese de que a entrada da anotação com dois genomas curados proverá um melhor resultado, o fluxo da seção anterior sofreu alterações. Foram inseridos nas tabelas *curated* e *blasthits*, os dados referentes a outro genoma curado. Assim, o banco de dados está com duas informações de origem para uma só fita de DNA como destino, ou seja, para cada CDS identificada existem duas opções possíveis.

Para resolver este problema, a consulta que retorna o resultado final, a qual realiza a junção entre as três tabelas (gene, curated e blasthits) sofreu modificação, o código da consulta original está exposto na [Figura 10](#):

```
select
'FT  gene                complement(' || gene.pos_begin || '..' || gene.pos_end || ')',
'FT                      /locus_tag=" || gene.systematic_id || "',
'FT                      /gene=" || curated.name || "',
'FT                      /product=" || curated.product || "',
'FT  CDS                  complement(' || gene.pos_begin || '..' || gene.pos_end || ')',
'FT                      /gene=" || curated.name || "',
'FT                      /product=" || curated.product || "',
'FT                      /similarity="Similar to ' || similarito.organismo || ', ' || similarito.produto ||
' (' || similarito.tamanho_subject || ' aa), e-value: ' || similarito.evalue ||
', ' || similarito.percentual || ' id in ' || similarito.tamanho_query || ' aa" ',
'FT                      /colour=" || decidecolor(gene.pseudogene, besthits.identity, besthits.size,
gene.cds_size, :p_cutoff),
'FT                      /note="Operon(FgenesB): ' || gene."OpTu" || ', Local_subcelular(Surfg): ' ||
gene.local_subcelular || '"
from gene LEFT OUTER JOIN (select distinct *
                           from gene JOIN blasthits ON (gene.systematic_id = blasthits.query)
                           where blasthits.identity > (:p_cutoff*1.0/100.0) and
                           blasthits.size > (gene.cds_size*(p_cutoff*1.0/100.0))) as besthits
                           ON (gene.systematic_id = besthits.query)
                           LEFT OUTER JOIN similarito ON (gene.systematic_id = similarito.query)
                           LEFT OUTER JOIN curated USING (subject)
where gene.orientation='-'
```

Figura 10 – ExportAll.sql

O resultado desta consulta é agregado pela cláusula **UNION**, a outra chamada

desta mesma consulta, mas com a condição **where gene.orientation** igual a: "+" dentro da consulta principal. Esta orientação indica para a busca, se os resultados são da fita reversa "-" ou direta "+".

O produto final da consulta é um arquivo no formato aceito pelo programa Artemis, com os respectivos dados inseridos na base de dados. Importa-se o arquivo resultante na fita de DNA destino para analisar-se os resultados.

Como agora para cada CDS na fita destino temos duas tuplas de resultados, é necessário que a consulta determine qual tupla possui o melhor resultado e, desta forma, garantir que não haja múltiplas anotações para um gene no arquivo final. A parte da consulta responsável por gerar a tabela de resultados está na cláusula **FROM**. O **SELECT** interno realiza a junção entre as tabelas gene e blasthits, gerando como resultado uma tabela temporária nomeada de **besthits**, como mostrado na Figura 11.

```
from gene LEFT OUTER JOIN (select distinct * from gene JOIN blasthits ON (gene.systematic_id = blasthits.query)
                             where blasthits.identity > 60 and blasthits.size > (gene.cds_size*0.60))
                             as besthits ON (gene.systematic_id = besthits.query)
LEFT OUTER JOIN similarto ON (gene.systematic_id = similarto.query)
LEFT OUTER JOIN curated USING (subject)
where gene.orientation='-'
```

Figura 11 – Cláusula **FROM** ExportAll.sql

Ao executar-se apenas esta consulta, obtém-se como resultados os dados a serem formatos pela consulta externa para o padrão do programa Artemis. Para demonstrar estes resultados, a Figura 12 exibe a saída de execução da consulta mais interna.

Data Output		Explain	Messages	History		
	systematic_id character varying(45)	pos_begin integer	pos_end integer	orientation character(1)	name character varying(255)	
1	organismMB20 00001	1	1812	+		
2	organismMB20 00001	1	1812	+		
3	organismMB20 00002	2397	3584	+		
4	organismMB20 00002	2397	3584	+		
5	organismMB20 00003	3651	4865	+		
6	organismMB20 00003	3651	4865	+		
7	organismMB20 00004	4855	5427	+		
8	organismMB20 00004	4855	5427	+		

Figura 12 – Resultados consulta mais interna ExportAll.sql

Como visto na Figura 12, para cada intervalo (determinado pelas colunas **pos_begin** e **pos_end**) temos duas tuplas de resultados. Necessita-se do retorno de apenas uma dessas duas tuplas, portanto, fez-se necessária a criação de uma função na base de dados da

aplicação para realizar a execução desta consulta, e assim, retornar o melhor resultado. Assim, o trabalho de unir as tabelas fica a cargo desta função, de maneira que a consulta principal fique por conta apenas da estruturação do resultado para o modelo de arquivo aceito pelo programa Artemis.

3.1.3 Função que retorna o melhor resultado

Baseando-se no problema exposto na [Figura 12](#) esta nova função deve retornar um dentre todos os resultados obtidos para uma determinada região, a qual é inserida como parâmetro para o processamento. Esta região é indicada na consulta através do atributo **systematic_id** da tabela *gene*, na qual se encontram os dados do organismo a ser anotado.

Diante dos resultados de cada região, necessita-se da identificação do melhor resultado. Para isto, avalia-se três atributos importantes contidos nas tuplas de resultados, *tamanho do alinhamento*, *identidade* (valor que representa quanto o intervalo correspondeu no processo de *Blast*) e *e-value* (valor de um teste estatístico, quanto menor melhor). Realizando-se uma ordenação por estes três valores, considera-se que a primeira tupla será aquela de interesse, então, retorna-se a primeira tupla para que esta seja formatada no padrão do arquivo de saída.

O conteúdo da função é mostrado na [Figura 13](#):

```
select into gene_curado, l_pseudogene, l_besthit identity, l_besthit_size, l_gene_cds_size
blasthits.subject, gene.pseudogene, blasthits.identity, blasthits.size, gene.cds_size
from gene JOIN blasthits ON (gene.systematic_id = blasthits.query)
where blasthits.identity > (p_cutoff*1.0/100.0) and
blasthits.size > (gene.cds_size*(p_cutoff*1.0/100.0)) and
blasthits.query = p_gene_systematic_id
order by gene.cds_size desc, blasthits.identity desc, blasthits.evalue limit 1;
```

Figura 13 – Função retorna melhor resultado

A última condição da cláusula **where** garante que a busca será feita apenas para a região inserida como parâmetro, indicada pelo atributo **p_gene_systematic_id**. A cláusula **order by** realiza a ordenação dos campos-chaves citados, para retornar o melhor resultado e a condição **limit 1** garante o retorno de apenas uma tupla de resultado.

A função recebe dois parâmetros: **p_cutoff** e **colorize**. O primeiro parâmetro corresponde ao valor mínimo determinado no momento da execução do programa, citado na seção de processamento do Pannotator. Este valor indica uma porcentagem sobre os resultados de *blasthit* na anotação de cada CDS e é inserido pelo usuário. O segundo parâmetro é utilizado como um sinalizador para indicar quando a função deve ou não, executar o processamento que retorna a cor da região anotada.

3.1.4 Novo ExportAll

Com os requisitos prontos, modificou-se a consulta **ExportAll.sql** mostrada na Figura 10, para contemplar estas mudanças, inseriu-se todo o trabalho de processamento feito anteriormente na mesma, para a função criada, denominada **annotation** (descrita anteriormente). Como exposto na Figura 14, esta nova versão fica responsável apenas por retornar os resultados no formato aceito pelo programa Artemis.

```
select
'FT  gene                complement(' || gene.pos_begin || '..' || gene.pos_end || '),
'FT                      /locus_tag="" || gene.systematic_id || "",
                        annotation(gene.systematic_id, :p_cutoff, FALSE),
'FT  CDS                complement(' || gene.pos_begin || '..' || gene.pos_end || '),
'FT                      /similarity="Similar to ' || similarito.organismo || ', ' ||
                        similarito.produto || ' (' || similarito.tamanho_subject ||
                        ' aa), e-value: ' || similarito.evalue || ', ' || similarito.percentual
                        || ' id in ' || similarito.tamanho_query || ' aa" ',
'FT                      /note="Operon(FgenesB): ' || gene."OpTu" || ', Local_subcelular(Surfg):
                        ' || gene.local_subcelular || '",
                        annotation(gene.systematic_id, :p_cutoff, TRUE)
from gene LEFT OUTER JOIN similarito ON (gene.systematic_id = similarito.query)
where gene.orientation='-.'
UNION
select
'FT  gene                ' || gene.pos_begin || '..' || gene.pos_end,
'FT                      /locus_tag="" || gene.systematic_id || "",
                        annotation(gene.systematic_id, :p_cutoff, FALSE),
'FT  CDS                ' || gene.pos_begin || '..' || gene.pos_end,
'FT                      /similarity="Similar to ' || similarito.organismo || ', ' ||
                        similarito.produto || ' (' || similarito.tamanho_subject ||
                        ' aa), e-value: ' || similarito.evalue || ', ' || similarito.percentual
                        || ' id in ' || similarito.tamanho_query || ' aa" ',
'FT                      /note="Operon(FgenesB): ' || gene."OpTu" || ', Local_subcelular(Surfg):
                        || gene.local_subcelular || '",
                        annotation(gene.systematic_id, :p_cutoff, TRUE)
from gene LEFT OUTER JOIN similarito ON (gene.systematic_id = similarito.query)
where gene.orientation='+'

```

Figura 14 – Novo ExportAll.sql

Diante desta mudança, faz-se necessária uma análise dos resultados obtidos com essa nova estrutura do processamento de anotação, com isto, é possível a realização de um teste de hipótese e sua validação se o resultado apresenta-se satisfatório. Além disso, é possível analisar-se o quanto o organismo é afetado pelos dois genomas curados utilizados (em percentual), sabendo-se que o ambos fazem parte da mesma família bacteriana.

4 Resultados

Para o caso analisado neste trabalho, a anotação final foi realizada baseando-se em dois genomas curados, definidos como mandatório e opcional. Criou-se uma tabela na base de dados para representar os resultados finais da anotação, facilitando a coleta de dados. Esta tabela é um resumo do resultado da consulta **ExportAll.sql** exibida anteriormente. Na [Figura 15](#) há um exemplo das informações contidas na tabela.

	systematic_id character varying(45)	orientation character(1)	pos_begin integer	pos_end integer	curated_name text
1	MB20genome 00001	+	1	1812	Cp258
2	MB20genome 00002	+	2397	3584	Cp316
3	MB20genome 00003	+	3705	4865	Cp316
4	MB20genome 00004	+	4855	5427	Cp258
5	MB20genome 00005	+	5567	7309	Cp258

Figura 15 – Exemplo do resultado final desse trabalho de monografia

Na [Figura 15](#) são exibidas as anotações automáticas de cinco genes, as informações associadas a estes são: identificador sequencial de um possível gene, a orientação, posição inicial e final na fita de DNA e o nome do genoma que deu origem à anotação destes possíveis genes. A provável função dos genes foi omitida.

Incorporou-se o resultado da anotação obtido a partir dos dois genomas curados à fita de DNA inserida no processo. No resultado final, de 2033 anotações realizadas, 1549 basearam-se no genoma mandatório e apenas 484 anotações foram feitas a partir do genoma opcional. A partir disto, conclui-se que, de uma forma geral, as anotações do genoma mandatório predominaram sobre as do genoma opcional.

	Genoma Curado	Quantidade anotada
1	Não anotado	344
2	Cp258	1549
3	Cp316	484

Figura 16 – Resultado total

4.1 Teste Qui-Quadrado X^2

O teste estatístico do qui-quadrado X^2 é utilizado para testar hipóteses, avaliar se existe associação entre duas variáveis qualitativas ([VIEIRA, 2015](#)). Neste trabalho, será

avaliado se um genoma possuiu predominância sobre o outro na anotação do genoma destino e quais foram essas áreas de predominância.

Para a realização do teste de hipótese via teste qui-quadrado, é necessário basear-se em duas hipóteses, sendo estas: a hipótese de nulidade, chamada de H_0 , a qual refere-se ao o que ocorre no cenário atual e também a hipótese H_1 , a qual foi formulada baseando-se no estudo do objetivo deste trabalho.

Para este trabalho, a hipótese a priori H_0 supõe que:

- Há regiões do genoma sob análise que possuem maior predominância de anotações de algum dos dois genomas curados, mandatório ou opcional.

A hipótese H_1 , formulada com base no objetivo deste trabalho, supõe que:

- Não há nas regiões sob análise diferença significativa na quantidade de genes anotados por nenhum dos genomas utilizados.

Para aplicação do teste de qui-quadrado, é necessário calcular as frequências das hipóteses. Com base nos dois arquivos curados inseridos e na fita de DNA, será realizado um levantamento em toda a anotação dos resultados dos genomas curados obrigatório e opcional.

É necessário estabelecer o nível de significância, que por definição é:

- A probabilidade de rejeitar-se H_0 , sendo H_0 verdadeira.

Ou seja, foi encontrado um valor de qui-quadrado, este valor favorece a hipótese H_1 , assim pode-se afirmar que H_1 está correta. Há possibilidade de se aceitar H_1 como a hipótese verdadeira erroneamente, ou seja, no caso de H_0 ser verdadeira, este erro é denominado nível de significância.

Este nível de significância é estabelecido para o estudo, e é representando por um valor de X%, no qual X varia de 0 a 100, para se ter um resultado mais seguro, indica-se que este nível seja estabelecido como o menor valor possível, assim, menores são as chances de se cometer um erro.

Além do grau de significância, é necessário calcular-se o número de graus de liberdade baseando-se na amostra do trabalho, este valor é calculado como $n - 1$, onde n é o número de elementos da amostra.

Com os dados prontos, pode-se aplicar o teste qui-quadrado com intuito de se aceitar ou rejeitar a hipótese H_0 , pois H_1 é consequência de sucesso ou falha de H_0 . A fita de DNA será percorrida com base neste intervalo, e para cada RMP será realizado o teste

de qui-quadrado, assim, será possível saber se em algum intervalo houve predominância do genoma opcional.

O valor de qui-quadrado é estabelecido pela seguinte fórmula:

$$X^2 = \sum_{i=1}^r \frac{(O_i - E_i)^2}{E_i}$$

Onde:

- O_i : frequência observada, valores com base no estudo.
- E_i : frequência esperada, valores que correspondem a H_0 .

Com o valor de qui-quadrado encontrado, é realizado uma comparação com o valor de tabela do qui-quadrado baseado nos graus de significância e liberdade do problema (tabela disponível em livros de bioestatística, como [Vieira \(2015\)](#) por exemplo) para aceitar uma das conclusões:

- Se X^2 encontrado for maior ou igual a X^2 da tabela, rejeita-se H_0 .
- Se X^2 encontrado for menor que X^2 da tabela, não rejeita-se H_0 .

4.1.1 Exemplo do uso do teste do qui-quadrado

Hipoteticamente, é preciso determinar se o tipo sanguíneo de um grupo de indivíduos independe de sua origem. Com base no objetivo do estudo, foram formuladas as hipóteses H_0 e H_1 .

- H_0 : tipo sanguíneo independe da origem.
- H_1 : depende da origem.

Com as hipóteses levantadas, foi colhida uma amostra da população onde estudou-se um grupo de 1136 pessoas, destas, 316 da nacionalidade A e 820 pessoas não pertencentes a nacionalidade A, com base na hipótese H_0 de que o tipo sanguíneo independe da origem do indivíduo.

O estudo foi realizado com base na tipagem sanguínea (estudo feito com base no tipo sanguíneo e fator Rh, positivo ou negativo) que um indivíduo possui ([ARONE; PHILIPPI, 1994](#)). Com o valores observados obtidos foi montado a seguinte tabela:

Com os valores observados prontos (O_i da equação do qui-quadrado), foram calculados os valores esperados para este estudo utilizando-se o método matemático denominado de regra de três. Com base no valor total de indivíduos observados de 1136, foram

	Rh^+	Rh^-	
Nacionalidade A	289	27	316
Não nacionalidade A	737	83	820
Total			1136

Tabela 1 – Exemplo tabela de contingência

encontradas 1026 pessoas Rh^+ , então para 316 pessoas da nacionalidade A esperava-se encontrar o valor de E_1 .

Logo, tem se:

$$1136 \text{ pessoas} \rightarrow 1026 \text{ } Rh^+$$

$$316 \text{ pessoas de A} \rightarrow E_1$$

$$E_1 = \frac{316 \cdot 1026}{1136} = 285,40$$

A [Tabela 1](#) apresenta que a quantidade de indivíduos da nacionalidade A com Rh^+ é de 289, mas é necessário calcular-se o valor esperado (baseando-se na hipótese H_0) como pedido na equação de qui-quadrado anteriormente exibida, e este valor deve ser calculado também para todos os outros valores encontrados na tabela (A Rh^- , Não A Rh^+ e Não A Rh^-).

Aplicando-se a mesma lógica para os demais valores, temos que:

- E_2 (A Rh^-): 30,59
- E_3 (Não A Rh^+): 79,40
- E_4 (Não A Rh^-): 740,59

Com os valores esperados encontrados, aplica-se a fórmula de qui-quadrado, substituindo-se os valores tem-se que:

$$X^2 = \frac{(289 - 285,40)^2}{285,40} + \frac{(27 - 30,59)^2}{30,59} + \frac{(83 - 79,40)^2}{79,40} + \frac{(737 - 740,59)^2}{740,59} = 0,646$$

Com o valor encontrado de $X^2 = 0,646$, estabelecendo-se um valor de significância de $x = 5\%$ e com grau de liberdade igual a 1, procurando-se o valor de X^2 para estes valores obtém-se $X^2 = 3,84$, observa-se que o valor obtido neste exemplo foi menor que o valor determinado na tabela, $X^2 = 0,646 < X^2 = 3,84$, com isso pode-se concluir que H_0 não pode ser rejeitada, ou seja, H_0 é a hipótese verdadeira, o tipo sanguíneo independe da origem, logo H_1 estava incorreta.

4.2 Detalhamento do Resultado

De modo geral, a predominância na anotação foi do genoma mandatório, realizou-se uma análise por intervalos no resultado final, ou seja, foram determinadas algumas regiões gênicas. Estas regiões são intervalos de tamanhos empiricamente determinados na fita de DNA, levantou-se nesta região a quantidade de anotações dos genomas mandatório e opcional e também, as regiões não anotadas. Baseando-se nestas informações calculou-se e verificou-se via teste qui-quadrado, se em alguma destas regiões houve a predominância de anotações baseadas no genoma opcional sobre anotações baseadas no genoma mandatório.

Por convenção foram definidos os seguintes tamanhos de intervalos na fita de DNA:

- 5000
- 7500
- 10000

Intervalos maiores que 10000 para este caso não alteraram significativamente os resultados. Para cada um dos intervalos acima, as anotações foram realizadas inserindo-se o parâmetro **p_cutoff** com três valores diferentes (citado na seção de processamento do PANNOTATOR), os quais foram (50%, 75% e 95%). Esperou-se que com valores diferentes sobre a porcentagem dos resultados de blasthit em diferentes intervalos, houvesse a possibilidade de aparição de alguma predominância das anotações do genoma opcional sobre o genoma mandatório.

Realizou-se a mesma anotação nove vezes (três intervalos com três valores de porcentagens diferentes), e para cada uma destas execuções, gerou-se um relatório que exibe se em algum intervalo houve a predominância esperada do genoma opcional. Utilizando-se o teste qui-quadrado, analisou-se o intervalo e a presença ou não predominância (conforme explicado na seção do método qui-quadrado, quando o resultado for menor do que 1, considera a hipótese validada). Na seção anexo, são exibidos todos os intervalos nos quais houve a predominância do genoma opcional sobre o mandatório. Na tabela 2, tem-se uma parte dos resultados para a anotação realizada com parâmetro **p_cutoff** igual a 75% e o intervalo de 7500 bases.

Intervalo		Requerido	Opcional	Total	Qui Quadrado	Predominância Opcional
202501	210001	2	3	7	0.67	TRUE
360001	367501	3	5	8	0.60	TRUE
592501	600001	2	4	6	0.50	TRUE

Tabela 2 – Intervalos onde houve predominância do genoma opcional

4.3 Dispersão de valores

Com base nos resultados encontrados, para as nove anotações realizadas foi gerado um gráfico *box plot* (gráfico de plotagem) para mostrar como foi a dispersão dos valores para as anotações realizadas.

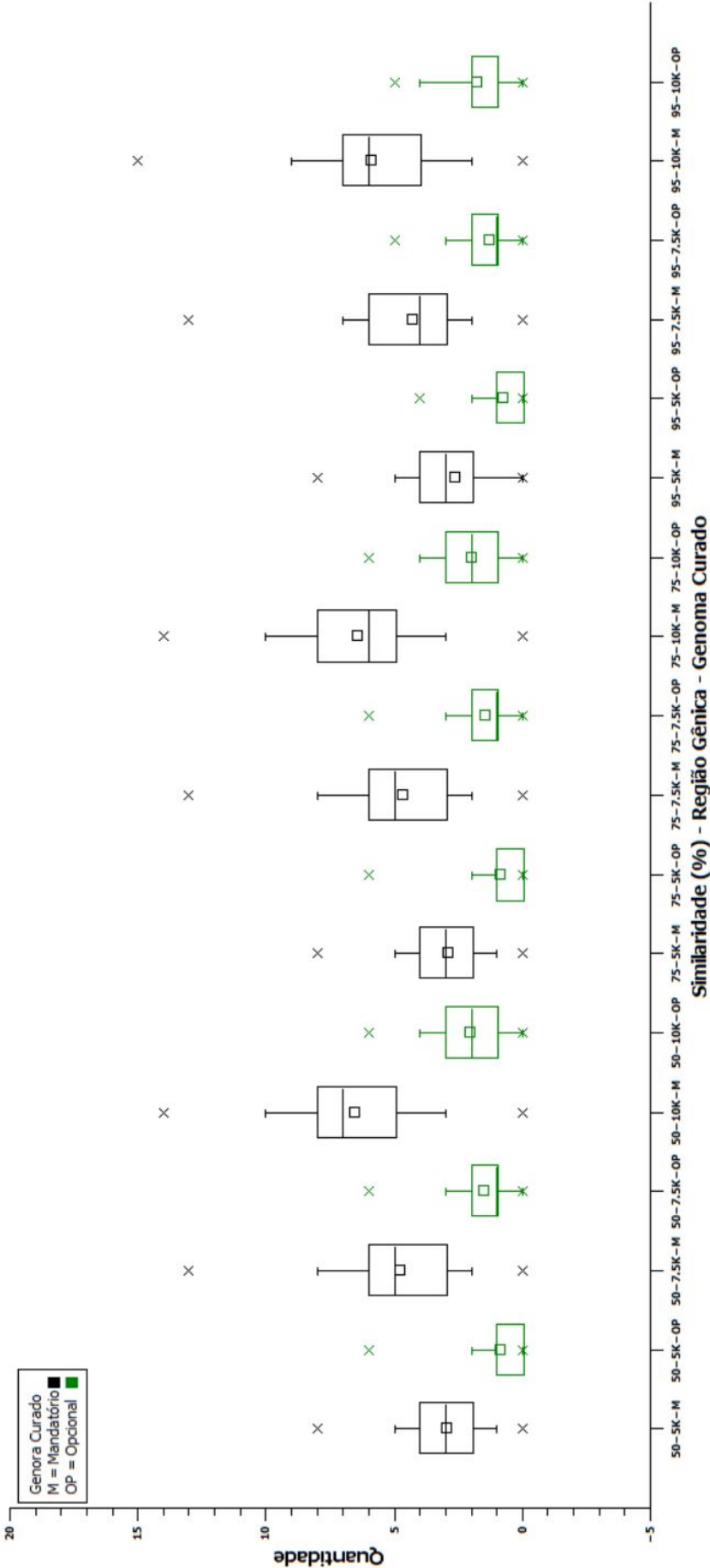


Figura 17 – Dispersão dos valores anotados

A Figura 17 exibe um gráfico, ao analisa-lo em pares, da esquerda para a direita, as caixas na cor preta representam a anotação para o genoma mandatório e as caixas na cor verde, a anotação do genoma opcional. Com isto, conclui-se que independente dos valores observados (intervalo gênico e percentual de **p__cutoff**) houve a predominância de anotação do genoma mandatório.

5 Conclusão

O genoma curado opcional utilizado para este trabalho foi analisado por um especialista, o qual informou o mesmo como sendo o segundo genoma mais próximo no organismo anotado além do genoma curado mandatório.

Foi visto que existe a possibilidade de que pequenas regiões gênicas com maior predominância evolutiva do genoma menos significativo também venham a ser identificadas nesse processo de anotação, utilizando mais de um genoma evolutivamente próximo. Como no exemplo a seguir, considerando um intervalo gênico de tamanho 5000, tem-se que o genoma opcional (denotado por CP_316) teve uma quantidade de anotação maior em relação a do genoma mandatório.

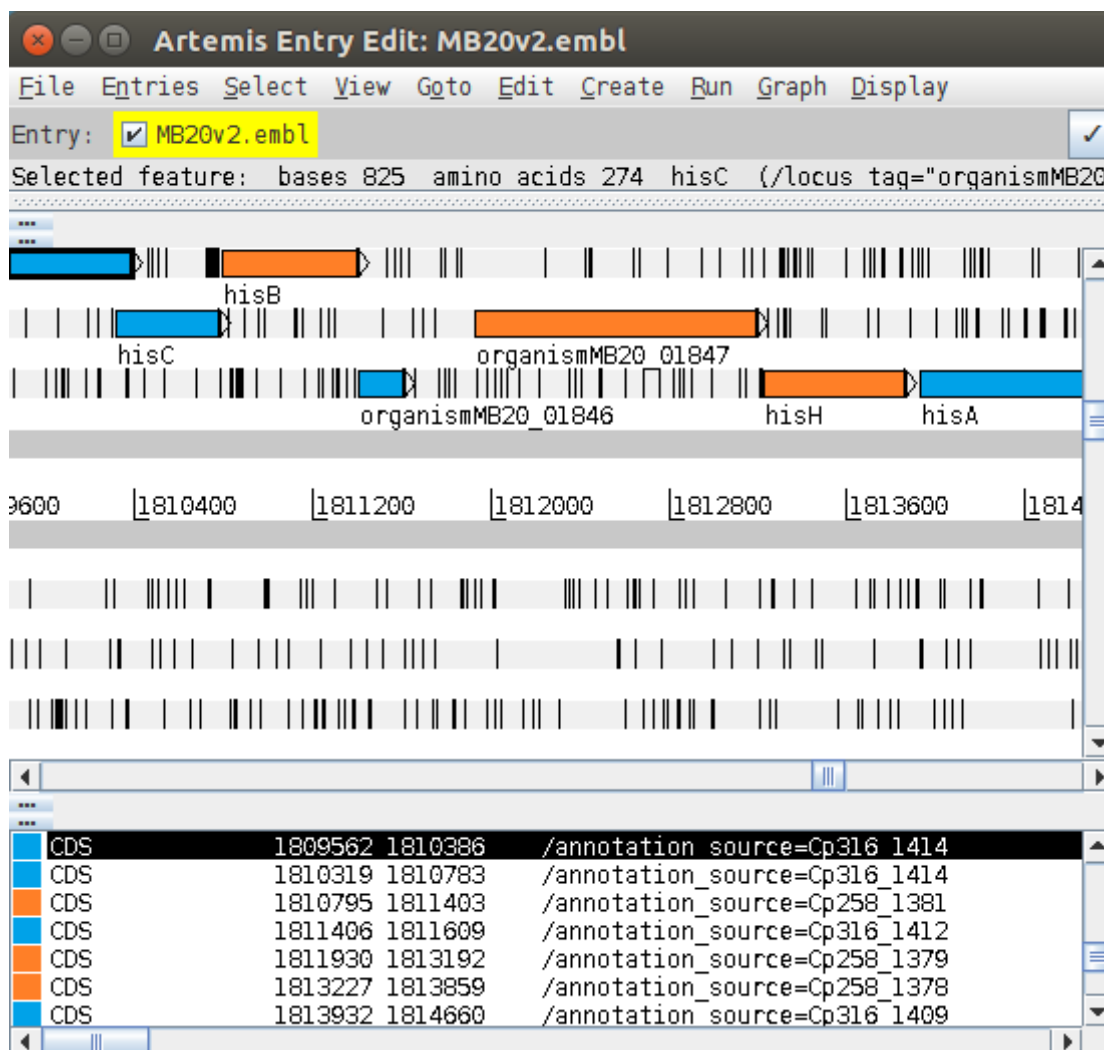


Figura 18 – Predominância de anotação do genoma opcional

Conclui-se assim que os resultados obtidos corroboram a opinião do especialista sobre a maior proximidade evolutiva do genoma mandatório para o genoma alvo. Entretanto, em situações em que um usuário do PANNOTATOR não possua informações sobre o parentesco evolutivo entre genomas, este trabalho poderá munir um pesquisador com dados sobre o parentesco evolutivo entre linhagens de organismos procariotos.

Em posse desses dados, um usuário do PANNOTATOR poderá conseguir uma anotação mais próxima da realidade, sem necessariamente necessitar de um curador especialista sobre o organismo trabalhado.

Referências

- ALBERTS, B. et al. *Fundamentos da biologia celular*. [S.l.]: Artmed Editora, 2006. Citado na página 16.
- ALBERTS, B. et al. *Biologia molecular da célula*. [S.l.]: Artmed, 2010. Citado na página 20.
- ARONE, E.; PHILIPPI, M. D. S. *Introducao a Enfermagem Medico - Cirurgica*. [S.l.]: SENAC, 1994. (Apontamentos Saude). ISBN 9788585578305. Citado na página 33.
- AZIZ, R. K. et al. The rast server: rapid annotations using subsystems technology. *BMC genomics*, BioMed Central Ltd, v. 9, n. 1, p. 75, 2008. Citado na página 11.
- CARNEIRO, A. R. et al. Genome sequence of *exiguobacterium antarcticum* b7, isolated from a biofilm in ginger lake, king george island, antarctica. *Journal of bacteriology*, Am Soc Microbiol, v. 194, n. 23, p. 6689–6690, 2012. Citado na página 11.
- CARVER, T. et al. Artemis and act: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, Oxford University Press, v. 24, n. 23, p. 2672–2676, 2008. Citado na página 25.
- COLLINS, F.; GALAS, D. et al. A new five-year plan for the us human genome project. *SCIENCE-NEW YORK THEN WASHINGTON-*, Citeseer, v. 262, p. 43–43, 1993. Citado na página 20.
- DELCHER, A. L. et al. Improved microbial gene identification with glimmer. *Nucleic acids research*, Oxford University Press, v. 27, n. 23, p. 4636–4641, 1999. Citado na página 24.
- GEREMIA, J. Tutorial de introdução a banco de dados. 2010. Citado na página 17.
- INTERATIVAS, F. R. S. et al. Projeto genoma humano. Minuto Biomedicina, 2013. Citado na página 20.
- JENKINS, J. *Genetic Code – Introduction to Biology*. 2015. Acessado: 2016-05-26. Disponível em: <<http://www.academyofone.org/genetic-code-introduction-to-biology-8-5/>>. Citado 2 vezes nas páginas 7 e 16.
- KLUG, W. S. et al. *Conceitos de genética*. [S.l.]: Artmed, 2010. Citado na página 15.
- LIMA, R. S. et al. Bioagents: Um sistema multiagente para anotação manual em projetos de sequenciamento de genomas. 2007. Citado na página 11.
- LOPES, S.; ROSSO, S. *Biologia: volume único*. São Paulo: Saraiva, v. 1, 2005. Citado 2 vezes nas páginas 7 e 14.
- MILANI, A. *PostgreSQL-Guia do Programador*. [S.l.]: Novatec Editora, 2008. Citado na página 17.
- MINISTRY, C. C. A. e R. *Gene definition*. 2015. Acessado: 2016-05-26. Disponível em: <<https://carm.org/dictionary-gene>>. Citado 3 vezes nas páginas 7, 14 e 15.

NHGRI, N. H. G. R. I. et al. The human genome project completion: Frequently asked questions. National Human Genome Research Institute (NHGRI), 2010. Citado na página 20.

PAREJA-TOBES, P. et al. Bg7: a new approach for bacterial genome annotation designed for next generation sequencing data. *PLoS One*, Public Library of Science, v. 7, n. 11, p. e49239, 2012. Citado na página 21.

SANTOS, A. et al. Pannotator: an automated tool for annotation of pan-genomes. *Genet. Mol. Res*, v. 12, p. 2982–2989, 2013. Citado 4 vezes nas páginas 7, 11, 18 e 19.

SANTOS, V. S. dos. *O que é genoma?* Acessado: 2017-06-20. Disponível em: <<http://brasilecola.uol.com.br/o-que-e/biologia/o-que-e-genoma.htm>>. Citado na página 11.

SEEMANN, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, Oxford Univ Press, p. btu153, 2014. Citado 2 vezes nas páginas 11 e 21.

TATUSOVA, T. A.; MADDEN, T. L. Blast 2 sequences, a new tool for comparing protein and nucleotide sequences. *FEMS microbiology letters*, The Oxford University Press, v. 174, n. 2, p. 247–250, 1999. Citado na página 26.

TETTELIN, H. et al. Comparative genomics: the bacterial pan-genome. *Current opinion in microbiology*, Elsevier, v. 11, n. 5, p. 472–477, 2008. Citado na página 11.

VIEIRA, S. *Introdução à bioestatística*. [S.l.]: Elsevier Brasil, 2015. Citado 2 vezes nas páginas 31 e 33.

WATSON, J. D.; CRICK, F. H. et al. Molecular structure of nucleic acids. *Nature*, v. 171, n. 4356, p. 737–738, 1953. Citado na página 11.

WOZNIAK, M.; WONG, L.; TIURYN, J. ecamber: efficient support for large-scale comparative analysis of multiple bacterial strains. *BMC bioinformatics*, BioMed Central Ltd, v. 15, n. 1, p. 65, 2014. Citado na página 22.

Anexos

ANEXO A – cutoff=50, Intervalo=5000

Intervalo		Requerido	Opcional	Total	Qui Quadrado	Predominância Opcional
1	5001	0	2	2	0.00	TRUE
20001	25001	2	3	7	0.67	TRUE
35001	40001	1	2	3	0.50	TRUE
80001	85001	1	2	3	0.50	TRUE
100001	105001	0	3	3	0.00	TRUE
140001	145001	0	3	3	0.00	TRUE
200001	205001	0	1	3	0.00	TRUE
285001	290001	2	3	6	0.67	TRUE
320001	325001	1	2	3	0.50	TRUE
330001	335001	2	3	5	0.67	TRUE
350001	355001	0	1	1	0.00	TRUE
360001	365001	2	4	6	0.50	TRUE
495001	500001	2	3	5	0.67	TRUE
515001	520001	0	2	2	0.00	TRUE
525001	530001	1	2	3	0.50	TRUE
595001	600001	1	3	4	0.33	TRUE
675001	680001	1	2	9	0.50	TRUE
850001	855001	1	4	5	0.25	TRUE
870001	875001	1	2	3	0.50	TRUE
900001	905001	1	2	3	0.50	TRUE
915001	920001	2	3	5	0.67	TRUE
920001	925001	0	2	3	0.00	TRUE
980001	985001	1	2	3	0.50	TRUE
985001	990001	2	3	5	0.67	TRUE
1030001	1035001	1	3	5	0.33	TRUE
1090001	1095001	0	3	3	0.00	TRUE
1170001	1175001	1	2	3	0.50	TRUE
1210001	1215001	1	3	4	0.33	TRUE
1255001	1260001	1	2	4	0.50	TRUE
1340001	1345001	0	6	6	0.00	TRUE
1580001	1585001	1	2	4	0.50	TRUE
1590001	1595001	0	2	2	0.00	TRUE
1600001	1605001	2	3	6	0.67	TRUE
1620001	1625001	1	2	3	0.50	TRUE
1625001	1630001	1	2	3	0.50	TRUE
1725001	1730001	1	3	4	0.33	TRUE
1730001	1735001	1	4	5	0.25	TRUE
1780001	1785001	0	2	2	0.00	TRUE
1785001	1790001	2	3	5	0.67	TRUE
1925001	1930001	2	3	5	0.67	TRUE
1960001	1965001	0	1	1	0.00	TRUE
2085001	2090001	1	3	4	0.33	TRUE
2275001	2280001	0	2	2	0.00	TRUE
2285001	2290001	1	2	3	0.50	TRUE

ANEXO B – **cutoff**=50, Intervalo=7500

Intervalo		Requerido	Opcional	Total	Qui Quadrado	Predominância Opcional
202501	210001	2	3	7	0.67	TRUE
360001	367501	3	5	8	0.60	TRUE
592501	600001	2	4	6	0.50	TRUE
675001	682501	1	2	13	0.50	TRUE
720001	727501	2	3	7	0.67	TRUE
727501	735001	1	2	3	0.50	TRUE
907501	915001	2	3	6	0.67	TRUE
915001	922501	3	5	8	0.60	TRUE
982501	990001	2	6	8	0.33	TRUE
1027501	1035001	2	5	8	0.40	TRUE
1342501	1350001	2	3	5	0.67	TRUE
1620001	1627501	2	3	5	0.67	TRUE
1725001	1732501	2	6	8	0.33	TRUE
1747501	1755001	2	4	6	0.50	TRUE
1860001	1867501	2	4	7	0.50	TRUE
1957501	1965001	2	3	5	0.67	TRUE
2085001	2092501	2	4	6	0.50	TRUE
2190001	2197501	1	2	3	0.50	TRUE

ANEXO C – **cutoff**=50, Intervalo=10000

Intervalo		Requerido	Opcional	Total	Qui Quadrado	Predominância Opcional
140001	150001	3	5	9	0.60	TRUE
200001	210001	2	4	8	0.50	TRUE
590001	600001	3	4	7	0.75	TRUE
910001	920001	4	6	10	0.67	TRUE
920001	930001	2	3	6	0.67	TRUE
980001	990001	3	6	9	0.50	TRUE
1340001	1350001	2	6	8	0.33	TRUE
1620001	1630001	3	4	7	0.75	TRUE
1780001	1790001	3	5	8	0.60	TRUE
2190001	2200001	2	3	5	0.67	TRUE
2270001	2280001	3	4	7	0.75	TRUE

ANEXO D – cutoff=75, Intervalo=5000

Intervalo		Requerido	Opcional	Total	Qui Quadrado	Predominância Opcional
1	5001	0	2	2	0.00	TRUE
20001	25001	2	3	7	0.67	TRUE
35001	40001	1	2	3	0.50	TRUE
80001	85001	1	2	3	0.50	TRUE
100001	105001	0	3	3	0.00	TRUE
140001	145001	0	3	3	0.00	TRUE
200001	205001	0	1	3	0.00	TRUE
285001	290001	2	3	6	0.67	TRUE
320001	325001	1	2	3	0.50	TRUE
330001	335001	2	3	5	0.67	TRUE
350001	355001	0	1	1	0.00	TRUE
360001	365001	2	4	6	0.50	TRUE
495001	500001	2	3	5	0.67	TRUE
515001	520001	0	2	2	0.00	TRUE
525001	530001	1	2	3	0.50	TRUE
595001	600001	1	3	4	0.33	TRUE
675001	680001	1	2	9	0.50	TRUE
850001	855001	1	4	5	0.25	TRUE
870001	875001	1	2	3	0.50	TRUE
900001	905001	1	2	3	0.50	TRUE
915001	920001	2	3	5	0.67	TRUE
920001	925001	0	2	3	0.00	TRUE
980001	985001	1	2	3	0.50	TRUE
985001	990001	2	3	5	0.67	TRUE
1030001	1035001	1	3	5	0.33	TRUE
1090001	1095001	0	3	3	0.00	TRUE
1170001	1175001	1	2	3	0.50	TRUE
1210001	1215001	1	3	4	0.33	TRUE
1255001	1260001	1	2	4	0.50	TRUE
1340001	1345001	0	6	6	0.00	TRUE
1580001	1585001	1	2	4	0.50	TRUE
1590001	1595001	0	2	2	0.00	TRUE
1600001	1605001	2	3	6	0.67	TRUE
1620001	1625001	1	2	3	0.50	TRUE
1625001	1630001	1	2	3	0.50	TRUE
1725001	1730001	1	3	4	0.33	TRUE
1730001	1735001	1	4	5	0.25	TRUE
1780001	1785001	0	2	2	0.00	TRUE
1785001	1790001	2	3	5	0.67	TRUE
1925001	1930001	2	3	5	0.67	TRUE
1960001	1965001	0	1	1	0.00	TRUE
2085001	2090001	1	3	4	0.33	TRUE
2275001	2280001	0	2	2	0.00	TRUE
2285001	2290001	1	2	3	0.50	TRUE

ANEXO E – **cutoff**=75, Intervalo=7500

Intervalo		Requerido	Opcional	Total	Qui Quadrado	Predominância Opcional
202501	210001	2	3	7	0.67	TRUE
360001	367501	3	5	8	0.60	TRUE
592501	600001	2	4	6	0.50	TRUE
675001	682501	1	2	13	0.50	TRUE
720001	727501	2	3	7	0.67	TRUE
727501	735001	1	2	3	0.50	TRUE
907501	915001	2	3	6	0.67	TRUE
915001	922501	3	5	8	0.60	TRUE
982501	990001	2	6	8	0.33	TRUE
1027501	1035001	2	5	8	0.40	TRUE
1342501	1350001	2	3	5	0.67	TRUE
1620001	1627501	2	3	5	0.67	TRUE
1725001	1732501	2	6	8	0.33	TRUE
1747501	1755001	2	4	6	0.50	TRUE
1860001	1867501	2	4	7	0.50	TRUE
1957501	1965001	2	3	5	0.67	TRUE
2085001	2092501	2	4	6	0.50	TRUE
2190001	2197501	1	2	3	0.50	TRUE

ANEXO F – **cutoff**=75, Intervalo=10000

Intervalo		Requerido	Opcional	Total	Qui Quadrado	Predominância Opcional
140001	150001	3	5	9	0.60	TRUE
200001	210001	2	4	8	0.50	TRUE
590001	600001	3	4	7	0.75	TRUE
910001	920001	4	6	10	0.67	TRUE
920001	930001	2	3	6	0.67	TRUE
980001	990001	3	6	9	0.50	TRUE
1340001	1350001	2	6	8	0.33	TRUE
1620001	1630001	3	4	7	0.75	TRUE
1780001	1790001	3	5	8	0.60	TRUE
2190001	2200001	2	3	5	0.67	TRUE
2270001	2280001	3	4	7	0.75	TRUE

ANEXO G – cutoff=95, Intervalo=5000

Intervalo		Requerido	Opcional	Total	Qui Quadrado	Predominância Opcional
1	5001	0	2	2	0.00	TRUE
20001	25001	2	3	7	0.67	TRUE
35001	40001	1	2	3	0.50	TRUE
80001	85001	1	2	3	0.50	TRUE
100001	105001	0	3	3	0.00	TRUE
140001	145001	0	3	3	0.00	TRUE
200001	205001	0	1	3	0.00	TRUE
285001	290001	2	3	6	0.67	TRUE
320001	325001	1	2	3	0.50	TRUE
330001	335001	2	3	5	0.67	TRUE
350001	355001	0	1	1	0.00	TRUE
360001	365001	2	4	6	0.50	TRUE
495001	500001	2	3	5	0.67	TRUE
515001	520001	0	2	2	0.00	TRUE
525001	530001	1	2	3	0.50	TRUE
595001	600001	1	3	4	0.33	TRUE
675001	680001	1	2	9	0.50	TRUE
850001	855001	1	4	5	0.25	TRUE
870001	875001	1	2	3	0.50	TRUE
900001	905001	1	2	3	0.50	TRUE
915001	920001	2	3	5	0.67	TRUE
920001	925001	0	2	3	0.00	TRUE
980001	985001	1	2	3	0.50	TRUE
985001	990001	2	3	5	0.67	TRUE
1030001	1035001	1	3	5	0.33	TRUE
1090001	1095001	0	3	3	0.00	TRUE
1170001	1175001	1	2	3	0.50	TRUE
1210001	1215001	1	3	4	0.33	TRUE
1255001	1260001	1	2	4	0.50	TRUE
1340001	1345001	0	6	6	0.00	TRUE
1580001	1585001	1	2	4	0.50	TRUE
1590001	1595001	0	2	2	0.00	TRUE
1600001	1605001	2	3	6	0.67	TRUE
1620001	1625001	1	2	3	0.50	TRUE
1625001	1630001	1	2	3	0.50	TRUE
1725001	1730001	1	3	4	0.33	TRUE
1730001	1735001	1	4	5	0.25	TRUE
1780001	1785001	0	2	2	0.00	TRUE
1785001	1790001	2	3	5	0.67	TRUE
1925001	1930001	2	3	5	0.67	TRUE
1960001	1965001	0	1	1	0.00	TRUE
2085001	2090001	1	3	4	0.33	TRUE
2275001	2280001	0	2	2	0.00	TRUE
2285001	2290001	1	2	3	0.50	TRUE

ANEXO H – cutoff=95, Intervalo=7500

Intervalo		Requerido	Opcional	Total	Qui Quadrado	Predominância Opcional
202501	210001	2	3	7	0.67	TRUE
360001	367501	3	5	8	0.60	TRUE
592501	600001	2	4	6	0.50	TRUE
675001	682501	1	2	13	0.50	TRUE
720001	727501	2	3	7	0.67	TRUE
727501	735001	1	2	3	0.50	TRUE
907501	915001	2	3	6	0.67	TRUE
915001	922501	3	5	8	0.60	TRUE
982501	990001	2	6	8	0.33	TRUE
1027501	1035001	2	5	8	0.40	TRUE
1342501	1350001	2	3	5	0.67	TRUE
1620001	1627501	2	3	5	0.67	TRUE
1725001	1732501	2	6	8	0.33	TRUE
1747501	1755001	2	4	6	0.50	TRUE
1860001	1867501	2	4	7	0.50	TRUE
1957501	1965001	2	3	5	0.67	TRUE
2085001	2092501	2	4	6	0.50	TRUE
2190001	2197501	1	2	3	0.50	TRUE

ANEXO I – **cutoff**=95, Intervalo=10000

Intervalo		Requerido	Opcional	Total	Qui Quadrado	Predominância Opcional
140001	150001	3	5	9	0.60	TRUE
200001	210001	2	4	8	0.50	TRUE
590001	600001	3	4	7	0.75	TRUE
910001	920001	4	6	10	0.67	TRUE
920001	930001	2	3	6	0.67	TRUE
980001	990001	3	6	9	0.50	TRUE
1340001	1350001	2	6	8	0.33	TRUE
1620001	1630001	3	4	7	0.75	TRUE
1780001	1790001	3	5	8	0.60	TRUE
2190001	2200001	2	3	5	0.67	TRUE
2270001	2280001	3	4	7	0.75	TRUE

ANEXO J – Melhor Resultado

Para o correto funcionamento do novo comando **ExportAll.sql**, criar a função **annotation** na base de dados.

```

1 CREATE OR REPLACE FUNCTION annotation(p_gene_systematic_id character
    varying, p_cutoff integer, colorize boolean)
2 RETURNS text AS
3 $BODY$
4 DECLARE
5     gene_curado text;
6     nome text;
7     produto text;
8     cor text;
9     nome_produto text;
10    l_pseudogene integer;
11    l_besthit_identity numeric;
12    l_besthit_size integer;
13    l_gene_cds_size integer;
14 BEGIN
15     nome_produto = '';
16     select into gene_curado, l_pseudogene, l_besthit_identity, l_besthit_size
        ,l_gene_cds_size
17     blasthits.subject, gene.pseudogene, blasthits.identity, blasthits.size,
        gene.cds_size
18     from gene JOIN blasthits ON (gene.systematic_id = blasthits.query)
19     where blasthits.identity > p_cutoff and
20     blasthits.size > (gene.cds_size*(p_cutoff*1.0/100.0)) and
21     blasthits.query = p_gene_systematic_id
22     order by gene.cds_size desc, blasthits.identity desc, blasthits.evaluate
        limit 1;
23
24     if gene_curado is not null then
25         select into nome, produto "name", product from curated where subject =
            gene_curado;
26         if produto is not null then
27             nome_produto = 'FT' || '/gene="' || nome || '"';FT
                || '/product="' || produto || '"';FT
                || '/annotation_source="' || gene_curado || '"';';
28         end if;--produto
29     end if;--gene_curado
30
31
32     if colorize then
33         select into cor decidecolor(l_pseudogene, l_besthit_identity,

```

```
        l_besthit_size,l_gene_cds_size, p_cutoff);
34     nome_produto = nome_produto || 'FT'           /colour=' || cor
        || ';' ;
35 end if;---colorize
36
37     return nome_produto;
38 END;
```

ANEXO K – Análise do intervalo

Levantamento dos resultados para análise da significância de anotação.

Criar a função **retorna_linha**:

```

1 CREATE OR REPLACE FUNCTION retorna_linha(p_gene_systematic_id character
    varying, p_cutoff numeric)
2 RETURNS text AS
3 $BODY$
4 DECLARE
5     nome_gene text;
6 BEGIN
7     select into nome_gene blasthits.subject
8         from gene JOIN blasthits ON (gene.systematic_id = blasthits.query)
9         where blasthits.identity > (p_cutoff*1.0/100.0) and
10             blasthits.size > (gene.cds_size*(p_cutoff*1.0/100.0)) and
11             blasthits.query = p_gene_systematic_id
12     order by gene.cds_size desc, blasthits.identity desc, blasthits.evaluate
13         limit 1;
14     return nome_gene;
15 END;
```

Gerar a view **Resultados** para auxiliar na coleta dos dados:

```

1 create view Resultados as
2     select systematic_id, orientation, pos_begin, pos_end,
3     substring(retorna_linha(systematic_id,70) from 1 for position('_' in
4         retorna_linha(systematic_id,70)) - 1) as curated_name
5     from gene;
```

Após criada a view **Resultados** (exemplo na [Figura 15](#)), inserir a seguinte função na base de dados para que seja feita a análise dos intervalos de acordo com os parâmetros recebidos pela função (**intervalo_ini**, **intervalo_fim**).

```

1 CREATE OR REPLACE FUNCTION analise_resultados(IN intervalo_ini bigint, IN
    intervalo_fim bigint)
2 RETURNS TABLE(x double precision, y double precision, z double precision,
    w double precision) AS
3 $BODY$
4 DECLARE
5     totalIntervalo float;
```

```

6   curated_A text;
7   curated_B text;
8   total_curated_A float;
9   total_curated_B float;
10  P_A_anotou float;
11  P_A_nanotou float;
12  P_B_anotou float;
13  P_B_nanotou float;
14  total float;
15  quiQuadrado float;
16  percentDiferenca float;
17 BEGIN
18  — Identificar os dois genomas curados passados para anotação
19  select distinct curated_name into curated_A
20  from Resultados
21  where curated_name <> ''
22  limit 1;
23
24  select distinct curated_name into curated_B
25  from Resultados
26  where curated_name <> ''
27  limit 1 offset 1;
28
29  —Quantidade anotada do genoma curated A
30  select count(curated_name) into total_curated_A from Resultados
31  where curated_name = curated_A AND
32  pos_begin > intervalo_ini AND pos_end < intervalo_fim;
33
34  —Quantidade do genoma curated B
35  select count(curated_name) into total_curated_B from Resultados
36  where curated_name = curated_B AND
37  pos_begin > intervalo_ini AND pos_end < intervalo_fim;
38
39  —Quantidade total de genomas no intervalo
40  select count(*) into totalIntervalo from Resultados
41  where pos_begin > intervalo_ini AND pos_end < intervalo_fim;
42
43  —Total
44  select (total_curated_A + (totalIntervalo - total_curated_A)) + (
45  total_curated_B + (totalIntervalo - total_curated_B)) into total;
46
47  —Pcp258_Anotou
48  select (totalIntervalo / total) * (total_curated_A + total_curated_B)
49  into P_A_anotou;
50
51  —Pcp258_No_Anotou
52  select (totalIntervalo / total) * ((totalIntervalo - total_curated_A) +

```

```

        (totalIntervalo - total_curated_B)) into P_A_nanotou;
51
52  --Pcp316_Anotou
53  select (totalIntervalo * (total_curated_A + total_curated_B)) / total
        into P_B_anotou;
54
55  --Pcp316_No_Anotou
56  select (totalIntervalo * ((totalIntervalo - total_curated_A) + (
        totalIntervalo - total_curated_B))) / total into P_B_nanotou;
57
58  --Calculo quiQuadrado
59  select (((total_curated_A - P_A_anotou)^2) / P_A_anotou) +
60  (((((totalIntervalo - total_curated_A) - P_A_nanotou)^2) / P_A_nanotou)
        +
61  (((total_curated_B - P_B_anotou)^2) / P_B_anotou) +
62  (((((totalIntervalo - total_curated_B) - P_B_nanotou)^2) / P_B_nanotou)
        into quiQuadrado;
63
64  select ((total_curated_A / (total_curated_A + (totalIntervalo -
        total_curated_A))) / (total_curated_B / (total_curated_B + (
        totalIntervalo - total_curated_B)))) into percentDiferenca;
65
66  return query select total_curated_A, total_curated_B, totalIntervalo,
        percentDiferenca;
67 END

```

Executando a função acima com o intervalo desejado, os resultados obtidos são as tabelas exibidas nos anexos A a I.