ANDRÉ LUCIANO NADAL

# BUSCA E CARACTERIZAÇÃO *in silico* DE RNAs NÃO CODIFICADORES EM ISOLADOS DE *Bacillus anthracis*, *Bacillus cereus* E *Bacillus thuringiensis* (*Bacillus cereus sensu lato)*

Londrina
2015

**Universidade Estadual de Londrina**

**Instituto Agronômico do Paraná**

**Empresa Brasileira de Pesquisa Agropecuária**

ANDRÉ LUCIANO NADAL

# BUSCA E CARACTERIZAÇÃO *in silico* DE RNAs NÃO CODIFICADORES EM ISOLADOS DE *Bacillus anthracis*, *Bacillus cereus* E *Bacillus thuringiensis* (*Bacillus cereus sensu lato)*

Londrina
2015

ANDRÉ LUCIANO NADAL

# BUSCA E CARACTERIZAÇÃO *in silico* DE RNAs NÃO CODIFICADORES EM ISOLADOS DE *Bacillus anthracis*, *Bacillus cereus* E *Bacillus thuringiensis* (*Bacillus cereus sensu lato*)

Exame de qualificação apresentado ao Programa de Pós–Graduação em Genética e Biologia Molecular, da Universidade Estadual de Londrina, como requisito parcial para a obtenção do título de Mestre.

Orientador: Prof. Dr. Laurival Antonio Vilas-Bôas.
Co-orientador: Prof. Dr. Alexandre Rossi Paschoal.

Londrina
2015

**Catalogação elaborada pela Divisão de Processos Técnicos da Biblioteca Central da Universidade Estadual de Londrina.**

**Dados Internacionais de Catalogação-na-Publicação (CIP)**

N127b Nadal, André Luciano.

Busca e caracterização *in silico* de RNAs não codificadores em isolados de *Bacillus anthracis, Bacillus cereus* e *Bacillus thuringiensis* (*Bacillus cereus sensu lato*) / André Luciano Nadal. – Londrina, 2015.

91 f. : il.

Orientador: Laurival Antonio Vilas-Bôas.

Coorientador: Alexandre Rossi Paschoal.

Dissertação (Mestrado em Genética e Biologia Molecular) – Universidade Estadual de Londrina, Centro de Ciências Biológicas, Programa de Pós-Graduação em Genética e Biologia Molecular, 2015.

Inclui bibliografia.

1. Bacillus (Bactéria) – Teses. 2. Ácido ribonucleico – Teses. 3. Seqüência de nucleotídeos – Teses. 4. Bioinformática – Teses. I. Vilas-Bôas, Laurival Antonio. II. Paschoal, Alexandre Rossi. III. Universidade Estadual de Londrina. Centro de Ciências Biológicas. Programa de Pós-Graduação em Genética e Biologia Molecular. IV. Instituto Agronômico do Paraná. V. EMBRAPA. VI. Título.

CDU 579.852.1

ANDRE LUCIANO NADAL

# BUSCA E CARACTERIZAÇÃO *in silico* DE RNAs NÃO CODIFICADORES EM ISOLADOS DE *Bacillus anthracis*, *Bacillus cereus* E *Bacillus thuringiensis* (*Bacillus cereus sensu lato*)

Exame de qualificação apresentado ao Programa de Pós–Graduação em Genética e Biologia Molecular, da Universidade Estadual de Londrina, como requisito parcial para a obtenção do título de mestre.

**BANCA EXAMINADORA**

_____
Prof. Dr. Laurival Antônio Vilas-Bôas
Universidade Estadual de Londrina – UEL

_____
Prof. Dr. Douglas Silva Domingues
Instituto agronômico do Paraná – IAPAR

_____
Prof. Dr. Fabrício Martins Lopes
Universidade Tecnológica Federal do Paraná – UTFPR

Londrina, 26 de março de 2015.

NADAL, André Luciano; VILAS-BÔAS, Laurival Antônio. **BUSCA E CARACTERIZAÇÃO *in silico* DE RNAs NÃO CODIFICADORES EM ISOLADOS DE *Bacillus anthracis*, *Bacillus cereus* E *Bacillus thuringiensis* (*Bacillus cereus sensu lato*).** 2015. 91f. Dissertação (Programa de Pós–Graduação em Genética e Biologia Molecular) - Universidade Estadual de Londrina, Londrina, 2015.

## RESUMO

O grupo do *Bacillus cereus* reúne microorganismos de grande importância econômica, médica e também em questões de biodefesa, o que se reflete no fato dele conter um grande número de genomas sequenciados intimamente relacionados, como *Bacillus anthracis*, *B. cereus* e *B. thuringiensis.* As ferramentas atuais de bioinformática aplicadas a tal conjunto de informações nos proporcionam grande oportunidade para análises genômicas comparativas completas. Os membros deste grupo tem muito de sua especificidade atribuída aos seus plasmídeos, os quais variam em tamanho e número. Seus cromossomos apresentam um alto nível de sintenia com diferenças limitadas no conteúdo genético, tornando questionáveis as interpretações sobre a especiação dos membros deste grupo. Este trabalho visa contribuir ao esclarecimento sobre a proximidade genética entre estes três constituintes do grupo do *Bacillus cereus sensu lato* anteriormente citados mediante identificação e caracterização de RNAs não codificadores, auxiliando na compreensão taxonômica, no entendimento dos fatores de virulência, na interação patógeno hospedeiro e em interações ecológicas como o comportamento de *B. thuringiensis* no controle de pragas da agricultura e vetores de doenças. Para a identificação dos ncRNAs, foi produzido um conjunto de dados composto por 35 genomas completos dos organismos de interesse, sendo 9 *B. anthracis*, 13 *B. cereus*, 12 *B. thuringiensis* e também 1 *B. weihenstephanensis,* obtidos à partir dos bancos de dados GOLD e NCBI. Tais genomas foram classificados considerando-se a metodologia de montagem após os sequenciamentos, tipo de anotação, manual ou automática e impacto das publicações resultando em 26 escolhidos. Em seguida o material foi processado no software Artemis V.16.0.0 para extração das regiões intergênicas, então submetidas a três métodos diferentes de inferência computacional para identificação de RNAs não codificadores. O primeiro método foi o processamento via Infernal V.1.1 / banco de dados Rfam V.11.0, o segundo foi o processamento via sRNAscanner V.1.9, e finalmente uma análise comparativa com o banco de dados da UTFPR que reúne ncRNAs da literatura, com base no Non-coding RNA Databases Resource (NRDR). Os dados foram então carregados para um banco de dados PostgreSQL V.9.1. Para a caracterização das sequências obtidas, foram criadas tabelas relacionais contendo os dados das 2208 famílias Rfam, agrupando os dados públicos dos sites FTP Rfam e instituto SANGER. Ainda como auxílio à busca e caracterização, os genomas completos foram carregados em banco de dados. Os resultados dos três métodos de descoberta foram pesquisados via consultas diretas no banco de dados (linguagem SQL) mediante agrupamento de regiões contíguas sobrepostas. Foram identificados 181 ncRNA candidatos, distribuidos em 12 famílias exclusivas para um ou outro grupo: *B. anthracis* (2), *B. cereus* (5) e *B. thuringiensis* (5). Posteriormente os candidatos foram caracterizados por espécie e cepa nas 23 famílias identificadas.

**Palavras-chave:** *Bacillus cereus sensu lato*. RNAs não codificadores. sRNA. ncRNA.

NADAL, André Luciano; VILAS-BÔAS, Laurival Antônio. **SEARCH AND *in silico* CHARACTERIZATION OF NON CODING RNAs ON *Bacillus anthracis*, *Bacillus cereus* AND *Bacillus thuringiensis* ISOLATES (CEREUS GROUP)**. 2015. 91p. Dissertação (Programa de Pós–Graduação em Genética e Biologia Molecular) - Universidade Estadual de Londrina, Londrina, 2015.

## ABSTRACT

The *Bacillus cereus* group gathers microorganisms of great economic importance and also at medical and biodefense issues, this is reflected in the fact that it contains a large number of sequenced genomes closely related organisms as *Bacillus anthracis*, *B. cereus* and *B. thuringiensis*. Current bioinformatics tools applied to this set of information provides us great opportunity to complete comparative genomic analyzes. Members of this group has plenty of its specificity attributed to their plasmids, which vary in size and number. Their chromosomes have a high level of synteny with limited differences in genetic content, which makes questionable the interpretations on speciation to the members of this group. This work aims to contribute to the clarification of the genetic proximity between these three constituents of the cereus group by identification and characterization of non-coding RNAs, contributing in taxonomic understanding, the understanding of virulence factors in host pathogen interaction and ecological interactions like the behavior of *B. thuringiensis* in the control of agricultural pests and disease vectors. With the purpose of identifying non-coding RNAs, it was produced a data set consisting of entire genomes of interest organisms, 9 B. anthracis, 13 *B. cereus*, 12 *B. thuringiensis* and also 1 *B. weihenstephanensis*, a total of 35 GenBank format genomes obtained from GOLD and NCBI databases. These genomes were classified considering the assembling methodology, after sequencing process, annotation type, manual or automatic as well the publications impact, resulting in 26 finally selected genomes. Data were then processed using Artemis V.16.0.0 program, for extraction of intergenic regions and submitted to three different methods of computational inference for non-coding RNAs identification. The first method, processing with Infernal V.1.1 / Rfam database V.11.0, the second was processing through sRNAscanner V.1.9, and finally a comparative analysis with the UTFPR database which gathers ncRNA literature based on Non-coding RNA Databases Resource (NRDR). Data from these three analyzes were loaded into a PostgreSQL V.9.1 database. Relational tables were created to the characterization of the obtained sequences. The tables contained all 2208 Rfam families data, grouping public FTP and Rfam institute SANGER sites data. As supporting means to search and characterization, complete genomes to the related organisms were loaded into a data base. Then the results of those three methods of discovery were searched through direct queries on the created database (SQL language) by grouping contiguous overlap regions. Thus, 181 ncRNA candidates were identified and further characterized by species, strain and ncRNA family. 181 ncRNA candidates were identified, distributed in 12 unique families to either group: *B. anthracis* (2), B. cereus (5) and *B. thuringiensis* (5). Later, candidates were characterized by species and strain on 23 identified families.

**Keywords:** *Bacillus cereus sensu lato*. Non-coding RNAs. sRNA. ncRNA.

## LISTA DE ILUSTRAÇÕES

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS, SÍMBOLOS E UNIDADES

| | |
|---|---|
| AFLP | (ingl.) Amplified Fragment Length Polymorphism, polimorfismo de comprimento de fragmentos amplificados |
| Bit | (ingl.) Binary digit, Dígito binário |
| BLAST | (ingl.) Basic Local Alignment Search Tool , ferramenta básica de busca e alinhamento local |
| CDS | (ingl.) Coding sequence - sequência codificadora |
| CM | (ingl.) Covariace model, modelo de covariância, ou CMS, como um perfil de sequência |
| *cry* | Gene codificador da proteina Cry |
| DDL | (ingl.) Data definition language, linguagem de definição de dados (SQL, banco de dados) |
| DER | Diagrama de entidades e relacionamentos (banco de dados) |
| DML | (ingl.) Data manipulation language, linguagem de manipulação de dados (SQL, banco de dados) |
| FTP | (ingl.) File transfer protocol, protocolo de transferência de arquivos |
| GNU | (ingl.) General Public License, licença pública geral, GNU GPL ou simplesmente GPL |
| GOLD | (ingl.) Genomes online database, Banco de dados de genomas 'online' |
| GOLSDTAMP | Identificador único do banco de dados Gold |
| IDE | (ingl.) Integrated development environment, ambiente integrado de desenvolvimento (a ferramenta de desenvolvimento de software, programação) |
| IGNs | IGN ou IGNs, Regiões intergênicas |
| Infernal | (ingl.) INFERence of RNA Alignment, inferência de alinhamento de RNA |
| ITPRO | (ingl.) Professional Information Tecnology (T.I. - Tecnologia da Informação) |
| LUCA | (ingl.) Last Universal Common Ancestor, ultimo ancestral comum universal |
| MLEE | (ingl.) Multilocus Enzyme Electrophoresis, eletroforese de enzimas multilocus |
| MLST | (ingl.) Multilocus sequence typing, tipagem/tipificação multilocus de sequência |
| MLVA | (ingl.) Multi-Locus VNTR Analysis, análise multilocus VNTR, ver VNTR |
| MSA | (ingl.) Multiple sequence alignments, alinhamentos múltiplos de sequência |
| NCBI | (sigla) National Center for Biotechnology Information, centro nacional para informações em biotecnologia |
| NRDR | (ingl.) Non-coding RNA Databases Resource, Bancos de dados de recursos para RNAs não codificadores |
| RNA | (ingl.) Ribonucleic acid, RNA, ácido ribonucleico |
| ncRNA | RNA não-codificante. Sinonímia: npcRNA (non-protein-coding RNA), nmRNA (non-messenger RNA), snmRNA ou sRNA (small non-messenger RNA), fRNA (functional RNA) |
| pb | (ingl.) base pairs, pares de bases |
| PDF | (ingl.) Portable Document Format, formato portátil de documento |
| PFGE | (ingl.) Pulsed Field Gel Electrophoresis, eletroforese em gel de campo |

| | |
|---|---|
| | pulsado |
| pH | Potencial Hidrogeniônico |
| PWM | (ingl.) Positional weight matrix, uma matriz de peso posicional |
| pXO1 e pXO2 | (biol. *Bacillus anthracis*) Plasmídeo toxina pXO1 e pXO2 plasmídeo cápsula |
| RAPD | (ingl.) Random Amplified Polymorphism DNA, Polimorfismo de DNA Amplificado ao Acaso |
| Rep-PCR | (ingl.) Repetitive Sequence-Based PCR, técnica de PCR em sequências palindrômicas extragênicas repetidas |
| Rfam | (sigla - banco de dados) Rna families, famílias de RNA |
| RNAr | (ingl.) Ribossomal ribonucleic acid, RNA ribossômico |
| SANGER | (sigla - Instituto) Instituto de pesquisa genômica |
| SNP | (ingl.) Single nucleotide polymorphisms, polimorfismos de nucleotídeo único |
| SO | Sistema Operacional (ingl. OS) |
| SQL | (ingl.) Structured Query Language,linguagem estruturada de consultas |
| sRNA | smalRna, sRNA |
| TU | (ingl.) TU, TUs, Transcription units, unidades transcricionais |
| UEL | (sigla) Universidade Estadual de Londrina |
| UTFPR | (sigla) Universidade Tecnológica Federal do Paraná |
| VNTRs | (ingl.) Variable number of tandem repeats, número variável de repetições em série |

## FORMATOS DE ARQUIVOS

| | |
|---|---|
| .fasta | (formato de arquivo) formato baseado em texto onde nucleotídeos ou aminoácidos são representados usando códigos de uma única letra |
| .gz | (linux, formato de arquivo) gzip, formato de compressão |
| .tar | (linux, formato de arquivo) arquivo TAR |
| .tar.gz | (linux, formato de arquivo) gzipped tar, usualmente um arquivo tar processado pelo gzip |
| .tgz | (linux, formato de arquivo) arquivo tar processado mediante gzip |
| .zip | (formato de arquivo) formato de compactação de arquivos |
| .asn | (NCBI, formato de arquivo) registro genômico em formato asn |
| .faa | (NCBI, Formato de ARQUIVO) sequências proteicas em formato fasta, arquivo de texto |
| .ffn | (NCBI, Formato de ARQUIVO) segmentos do genoma que codificam para proteínas |
| .fna | (NCBI, Formato de ARQUIVO) sequência fasta do genoma |
| .frn | (NCBI, Formato de ARQUIVO) segmentos do genoma que codificam RNA |
| .gbk | (NCBI, Formato de ARQUIVO) genoma em formato de arquivo genbank |
| .gff | (NCBI, Formato de ARQUIVO) características do genoma |
| .gpff | (NCBI, Formato de ARQUIVO) proteína genbank |
| .ptt | (NCBI, Formato de ARQUIVO) tabela de proteína |
| .rnt | (NCBI, Formato de ARQUIVO) tabela rna |
| .rpt | (NCBI, Formato de ARQUIVO) relatório, sumário |
| .val | (NCBI, Formato de ARQUIVO) arquivo binário (projeto genoma) |

**SUMÁRIO**

## 1 INTRODUÇÃO

O grupo *Bacillus cereus* é composto por seis espécies, *Bacillus cereus, B. thuringiensis, B. anthracis, B. weihenstephanensis, B. mycoides* e *B. pseudomycoides*. Estas espécies são intimamente relacionadas, sendo que as cepas de *B. cereus stricto sensu* (*B. anthracis, B. cereus e B. thuringiensis*) compartilham cromossomos altamente conservados, diferindo sobretudo na presença de plasmídeos contendo genes responsáveis por virulência (KLEE, 2010).

Tal denominação, grupo do *B. cereus,* apesar de não ser taxonômica, denota um agrupamento legítimo das espécies envolvidas baseado na similaridade entre as sequências do gene do *RNAr 16S* destas espécies. Atualmente o gênero *Bacillus*, pode ser dividido em seis grupos, denominados *Bacillus* RNA grupo 1 a grupo 6, sendo que as espécies do grupo do *B. cereus* ficam agrupadas na subdivisão 3 do grupo 1 (STACKEBRANDT & SWIDERSKI, 2002). Apesar desta divisão, a grande similaridade entre as sequências do *loco* do RNAr 16S apresentada pelas linhagens destas espécies é uma das justificativas centrais para discussões taxonômicas deste grupo de bactérias (VILAS-BÔAS et al., 2007).

Entre os organismos investigados neste trabalho, *B. anthracis, B. cereus e B. thuringiensis,* o primeiro é uma bactéria formadora de endósporos que provoca antraz por inalação, por isso tornou-se um assunto muito estudado como resultado de seu uso em um ataque de bioterrorismo nos Estados Unidos em setembro e outubro de 2001 (READ, 2003).

*Bacillus cereus* é onipresente na natureza. Enquanto a maioria dos isolados parece ser inofensivos, alguns estão associados a doenças transmitidas por alimentos, doenças periodontais e outras infecções graves como a Síndrome Diarreica de Aparição Tardia e a Síndrome Emética de Aparição Rápida (BARRETO, 2000). Cepas como *B. cereus* G9241 identificadas a partir de infecções, apontaram *B. cereus* inclusive como agente causador de pneumonia grave em um soldador de Louisiana em 1994 (HAN et al., 2006).

Inúmeras linhagens de *B. cereus* foram reconhecidamente envolvidas em casos de intoxicações gastrointestinais, causadas pela ingestão de alimentos, produzindo diarreias e vômitos (KOTIRANTA et al., 2000). Esta natureza oportunista de *B. cereus* como patógeno está relacionada com sua capacidade de produzir inúmeros fatores de

virulência não específicos, incluindo fosfolipases, hemolisinas e enterotoxinas (DROBNIEWSKI, 1993).

As razões que tornam o *B. cereus* um importante agente de contaminação na indústria de alimentos referem-se a fatores como de que esporos onipresentes, ocasionando problemas de contaminação relacionados a maquinaria de processamento ou nos materiais das embalagens. Tais problemas relacionam-se à característica adesiva dos esporos de diversas cepas, com a posterior formação de biofilmes, além de esporos e células vegetativas poderem apresentar boa tolerância a variados tratamentos de esterilização considerados seguros (HEYNDRICKX E SCHELDEMAN, 2002).

*B. thuringiensis* é a bactéria mais conhecida e estudada entre as linhagens pertencentes ao grupo *B. cereus* (VILAS-BÔAS et al., 2007). A característica entomopatogênica é o que distingue *B. thuringiensis* dos outros bacilos do grupo, e deve-se à formação de proteínas denominadas *Cry*, formadas durante o ciclo de esporulação e liberada na forma de cristal juntamente com um esporo elipsoidal (CRICKMORE et al., 1998). Os cristais proteicos são tóxicos, principalmente para insetos das ordens Lepidoptera, Diptera, Coleoptera, Hymenoptera, Homoptera, Dictyoptera, Orthoptera, Mallophaga, além de nematoides (Strongylida, Tylenchida), protozoários (Diplomonadida) e ácaros (Acari) (FEITELSON et al., 1992; SCHNEPF et al., 1998), não apresentando qualquer dano aos mamíferos, aves, anfíbios ou répteis.

Os primeiros produtos utilizando *B. thuringiensis* no controle de insetos pragas da agricultura e vetores de doenças começaram a ser comercializados em 1938. Em 1950, com a implementação de novas tecnologias como o uso de fermentadores e com a descoberta de novas subespécies de *B. thuringiensis*, houve um aumento na produção e na comercialização dos produtos à base de *B. thuringiensis*, o que ampliou o número de insetos-alvo controlados (ARANTES, VILAS-BÔAS, 2012). Neste trabalho apresentamos os resultados de um pipeline desenvolvido de forma a identificar genes para RNAs não codificantes (ncRNAs), genes que produzem moléculas de RNA funcionais ao invés de codificação de proteínas. Os resultados obtidos, em conjunto com outros estudos acerca da especificidade de hospedeiro e nicho ecológico da bactéria são também características importantes, que devem ser consideradas na definição taxonômica.

# 2 FUNDAMENTAÇÃO TEÓRICA

## 2.1. **Conceitos em Bioinformática**

Em 17 de julho de 2000, o comitê de definição BISTIC (*Biomedical Information Science and Technology Definition Committee*), comissão presidida pelo Dr. Michael Huerta, do Instituto Nacional de Saúde Mental, juntamente com a equipe de *experts* Gregory Downing e Belinda Seto, publicou as seguintes definições de bioinformática e biologia computacional, reconhecendo que nenhuma definição poderia eliminar completamente sobreposições com outras atividades ou impedir variações de interpretação por diferentes indivíduos e organizações, mas propondo um consenso para a comunidade bioinformática internacional (*Biomedical Computation Review, 2000*):

**- Bioinformática:** pesquisa, desenvolvimento ou aplicação de ferramentas computacionais e abordagens para a expansão do uso de dados biológicos, médicos, comportamentais ou de saúde, incluindo métodos para adquirir, armazenar, organizar, arquivar, analisar ou visualizar tais dados (BISTIC, 2000).

**- Biologia Computacional:** desenvolvimento e aplicação de dados analíticos e métodos teóricos, modelagem matemática e técnicas de simulação computacional para o estudo dos sistemas biológicos, comportamentais e sociais (BISTIC, 2000).

### 2.1.1. **O Trabalho *in silico*: Conceito**

A expressão *in silico* é utilizada no âmbito da simulação computacional e áreas correlatas de forma a indicar um processo desempenhado mediante uma simulação computacional. A expressão foi cunhada a partir das expressões latinas *in vivo* e *in vitro*, frequentemente usadas nas Ciências Biológicas, tendo sido utilizada inicialmente em um workshop sobre “Autômatos Celulares: Teoria e Aplicações" (Los Alamos, EUA, 1989). Embora *in silico* não tenha um correlato em Latim, pois a expressão correta seria *in silicio*, a literatura e a comunidade internacional adotaram prontamente o termo *in silico* que hoje aparece inclusive no título de periódicos como o *in silico* Biology, indexado em plataformas como: Pubmed, MEDLINE, Embase, Elsevier BIOBASE, Biological Abstracts, entre outros.

A expressão *in silico* é comumente usada apenas para denotar simulações computacionais que modelam um processo natural ou de laboratório e não para cálculos computacionais genéricos. Mediante a utilização de simulações computadorizadas, objetiva-se poupar tempo, materiais e recursos auxiliando o trabalho de pesquisa, embora exista a necessidade de um maior conhecimento por parte dos proponentes de tal pesquisa, acerca do tema (domínio de estudo) a ser desenvolvido (TRAVASSOS & BARROS, 2003), conforme tópicos listados a seguir:

**- Desempenho:** desafio relacionado ao processamento dos dados, ao uso de infraestruturas e a ambientes para alto desempenho;

- **Identificação de requisitos:** desafio relacionado ao desenvolvimento de novas técnicas de identificação de requisitos;

**- Armazenamento:** desafio relacionado ao armazenamento do alto volume de dados gerados e dos fatores associados à sua manipulação;

**- Colaboração:** desafio relacionado ao impacto da colaboração entre profissionais da T.I. e outros cientistas e ao uso de ferramentas que auxiliem a modelagem colaborativa;

**- Visualização:** desafio relacionado ao desenvolvimento de algoritmos e técnicas de visualização dos resultados gerados.

Travassos & Barros (2003) apresentam uma taxonomia para estudos experimentais em Engenharia de Software, de onde destacamos as duas definições de interesse para este estudo, o conceito de *in virtuo*, representando experimentos onde o ambiente é representado através de modelos computacionais, que descrevem a realidade com qual os participantes interagem diretamente. O outro conceito, *in silico*, descreve onde tanto os participantes quanto o ambiente e objeto de estudo são descritos como modelos computacionais, não havendo (ou reduzindo ao máximo) qualquer tipo de intervenção humana.

### 2.2. **Genômica Comparativa**

Sobrevivência, um imperativo biológico e um dos motores de propulsão da história, materializado na preocupação com os riscos potenciais da utilização de novas tecnologias e fontes de energia, esse foi um dos motivos que levaram o Departamento

de Energia Norte-Americano (DOE) a iniciar um processo que resultaria nos primórdios do Projeto Genoma Humano (HGP) em 1990. Com o intuito de obter uma sequência genômica humana de referência, tal projeto foi concluído em abril de 2003. Como evento de grande vulto, os recursos tecnológicos gerados por ele levariam ao início de muitos outros projetos genoma tanto públicos como em setores privados.

Análises de dados disponíveis em repositórios como o GOLD (Tabela 1) demonstram uma nítida preferência pelo sequenciamento de genomas bacterianos (de menor tamanho que os genomas eucarióticos e, portanto, mais fáceis de serem analisados) e genomas com importância biomédica ou biotecnológica.

**Tabela 1: Projetos registrados no banco de dados GOLD**

| Tipo | Domínios de pesquisa | Nº Projetos |
|---|---|---|
| **Sequenciamento completo de genomas** | - | 46381 |
| **Projetos completos** | - | 25200 |
| **Projetos em andamento** | - | 15802 |
| **Sequenciamento completo** | Bacteria | 10917 (Gênero Bacillus: 225) |
| **Sequenciamento completo** | Eukaria | 4218 |
| **Sequenciamento completo** | Vírus | 427 |
| **Sequenciamento completo** | Archaea | 239 |
| **Total geral de projetos** | - | 49640 |

Fonte: Banco de dados GOLD - Genomes Online Database, Jan. 2015.

A multidisciplinaridade de áreas e notadamente a bioinformática e a biologia computacional possibilitaram à comunidade científica inúmeras aplicações em genômica (TRAVASSOS & BARROS, 2003). Conceituada como a obtenção e análise de sequências genômicas completas de inúmeros organismos, os trabalhos em genômica somados às tecnologias de alto desempenho na geração e análise de dados como a transcriptômica e a proteômica, tem permitido à comunidade científica o uso de abordagens sistêmicas (abrangentes), no estudo da estrutura, organização e evolução de genomas, expressão diferencial de genes e proteínas, predição e classificação funcional de genes (CATANHO; DEGRAVE; MIRANDA, 2008, p. 20).

Uma das novas abordagens é a genômica comparativa que consiste na análise e comparação do material genético de diferentes espécies ou cepas, com o propósito de estudar a estrutura, organização e evolução dos genomas juntamente com as funções dos genes e regiões não codificantes, um dos objetivos deste trabalho.

### 2.2.1. **Comparações: Regiões Não Codificantes**

A regulação transcricional é um mecanismo adaptativo importante em procariotos. Seus elementos chave são as proteínas regulatórias e os sinais regulatórios localizados em regiões extra gênicas. Desta forma, o trabalho de comparação de regiões não codificantes em genomas de diferentes espécies procarióticas têm sido muito útil para a identificação e caracterização de segmentos genômicos regulatórios (PAREJA et al., 2006), auxiliando no entendimento dos circuitos genéticos de regulação transcricional nestes organismos. Tais abordagens levam em consideração a suposição de que as regiões funcionalmente importantes sofrem a pressão seletiva e, tendendo portanto a evoluir a taxas menores comparativamente às regiões sem papel funcional (WEI et al., 2002).

## 2.3. **Características Gerais e Taxonomia do Grupo *Bacillus cereus***

Segundo ROH et al 2007, *B. thuringiensis* é um patógeno de insetos, *B. cereus* é uma bactéria conhecida principalmente por causar intoxicação alimentar, além de diarreia e vômitos podendo também causar infecções mais graves (DROBNIEWSKI, 1993). *B. anthracis*, o agente etiológico do antraz, encontra-se em todo o mundo e é capaz de infectar virtualmente todos os mamíferos. É uma questão de debate se essas bactérias representam três espécies distintas ou subespécies de *B. cereus sensu lato* (DAFFONCHIO, 2000 e HELGASON, 2000).

As características de cada espécie, bem como o perfil de patogenicidade são determinados por genes frequentemente presentes em plasmídeos (como o gene *cry1Ac* em *B. thuringiensis*). Com relação aos produtos dos genes, podem ser considerados toxinas e cápsula de *B. anthracis*, as proteínas cristal inseticidas de *B. thuringiensis*, e a síntese de cereulide em variedades eméticas de *B. cereus*. No entanto, outros fatores de virulência como fatores hemolíticos, motilidade, e resistência a antibióticos são codificados por genes presentes geralmente nos cromossomos (KLEE, 2010).

*B. anthracis* é um clado altamente monofilético, sendo que seus isolados são diferenciados por meio da identificação de polimorfismos de nucleotídeo único (SNP) e um número variável de repetições em série (VNTRs) (VAN ERT et al., 2007; KEIM et al., 2000). Uma exotoxina tripartite e uma cápsula de ácido poli-D-glutâmico, expresso como o principal determinante de virulência, são essenciais para a patogenicidade completa e são codificados por genes presentes nos plasmídeos pXO1 e pXO2 (CHUN et al., 2012). O agente patogênico é capaz de causar edema e morte celular pelo tríplice complexo de toxina: o antígeno protetor, o fator de edema e o fator letal (MOCK & MIGNOT, 2003). A produção de uma cápsula de ácido poli glutâmico permite ao organismo escapar do sistema imune (FOUET & MESNAGE, 2002).

Fatores de virulência são codificados no plasmídeo toxina pXO1 (MURAWSKA et al., 2013), e no plasmídeo cápsula, pXO2 (MAKINO et al., 1989). Embora as sequências de pXO1 e em menor extensão da pXO2 estejam largamente distribuídas entre as cepas do grupo *B. cereus*, a presença de plasmídeos que contêm os genes da toxina e da cápsula ocorre apenas raramente (PANNUCCI et al., 2002).

*B. cereus*, *B. thuringiensis* e *B. anthracis* são organismos que se destacam devido aos diferentes impactos que apresentam na atividade humana. *B. thuringiensis* é amplamente utilizado no controle de insetos considerados como pragas agrícolas e vetores de doenças (SCHNEPF et al., 1998). O *B. cereus* pode ser considerado como um patógeno oportunista devido a seu envolvimento em eventos de contaminação alimentar (KOTIRANTA et al., 2000). O *B. anthracis* por sua vez é um patógeno de animais agindo inclusive sobre a espécie humana sendo conhecido por sua capacidade de produzir uma toxina letal, também utilizada como arma biológica (JERNIGAN et al., 2002).

A semelhança morfológica entre esporos e células vegetativas em *B. thuringiensis*, *B. anthracis* e *B. cereus* pode ser verificada ao microscópio óptico mas foram as diferenças fenotípicas entre as três variedades fez com que fossem classificados inicialmente como espécies distintas. Alguns genes responsáveis pelas características fenotípicas divergentes de *B. thuringiensis* e *B. anthracis* estão presentes em grandes plasmídeos que podem ser transferidos para outras linhagens do grupo *B. cereus* através de mecanismos de transferência horizontal de genes como a conjugação, transdução e a transformação, e também rearranjos genéticos promovidos por elementos móveis de DNA (VILAS-BÔAS et al., 2007).

A dificuldade de resolução taxonômica entre as linhagens de *B. cereus* e *B. thuringiensis* deve-se à extensiva similaridade de seus genomas e também ao grau de polimorfismo genético. Carlson et al. 1994, em uma das primeiras investigações quanto às relações taxonômicas em questão, utilizou eletroforese em gel de campo pulsado (*Pulsed Field Gel Electrophoresis*, PFGE), e também eletroforese de aloenzimas (*Multilocus Enzyme Electrophoresis*, MLEE).

Como foi observado um alto grau de variabilidade genética interespecífico e intraespecífico e não sendo possível a divisão em duas espécies, foi sugerido o agrupamento de *B. cereus* e *B. thuringiensis* como uma única espécie. A análise de sequências de genes cromossômicos e mesmo outros estudos usando MLEE revelaram um alto grau de similaridade genética indicando a falta de diferenciação entre *B. thuringiensis* e *B. cereus* (HELGASON et al., 1998, 2000).

Por outro lado, autores como Ticknor, 2001, afirmam que no sequenciamento de DNAr 16S, resultados de polimorfismo de comprimento de fragmentos amplificados (AFLP) e MLEE, evidenciaram o alto grau de polimorfismo entre *B. cereus* e *B. thuringiensis* apontando que para uma caracterização destas espécies seria insuficiente a simples análise de um número limitado de linhagens.

As análises utilizando-se de outras ferramentas genéticas mostram resultados controversos quanto à taxonomia de *B. cereus*, *B. thuringiensis* e *B. anthracis*, com trabalhos que ora distinguem as três espécies de *Bacillus* como unidades taxonômicas distintas, ora como linhagens de uma mesma espécie. Desta forma, tem se mostrado relevante, acrescentar aos estudos genéticos, estudos relacionados à ecologia das linhagens do grupo *B. cereus*. A soma destas duas áreas pode, de forma elucidativa, contribuir para a definição taxonômica deste grupo (FAZION;VILAS-BÔAS, 2012).

### 2.3.1. ***Bacillus anthracis***

*Bacillus anthracis* é uma bactéria formadora de endósporos que provoca antraz por inalação (READ, 2003). Os membros desta espécie não são móveis, e são todos caracterizados pela presença de quatro profagos e uma mutação sem sentido no gene regulador *plcR* (KLEE et al., 2010)[1].

[1] Este estudo apresenta variadas tabelas com análise comparativa de características entre as espécies e árvore filogenética, tal material apresenta relevância para o presente estudo devendo ser revisado posteriormente.

O antraz, infecção bacteriana muitas vezes fatal, ocorre quando endósporos de *B. anthracis* entram no corpo através de lesões na pele, por inalação ou ingestão. Além de sua capacidade de causar antraz, este agente tornou-se famoso como uma arma biológica devido ao seu uso em um ataque de bioterrorismo nos Estados Unidos em setembro e outubro de 2001. Tal agente apresenta endósporos muito resistentes ao meio ambiente. Durante o curso da doença, endósporos são tomados por macrófagos alveolares e germinam em compartimentos fagolisossomais e, ao escaparem do macrófago, as células vegetativas acabam eventualmente infectando a corrente sanguínea (DIXON, 1999).

Estudos anteriores sugeriram que o *B. anthracis ames ancestor*, o isolado Ames original contendo o plasmídeo virulento apresenta-se como referência ideal, servindo como modelo para o estudo de outras cepas classificadas como *B. anthracis* (RAVEL et al., 2008). A primeira sequência descrita foi da estirpe virulenta conhecida como *Ames ancestor* (ancestral Ames, nomenclatura criada para distinção quanto a seus descendentes). Este bacilo foi isolado em 1981 a partir de um cadáver de novilha Beefmaster de 14 meses de idade em Sarita no Texas.

### 2.3.2. ***Bacillus cereus***

*Bacillus cereus* é onipresente na natureza, enquanto a maioria dos isolados parecem ser inofensivos, alguns estão associados a doenças transmitidas por alimentos, doenças periodontais e outras infecções mais graves. Variedades como *B. cereus* G9241 identificadas a partir de infecções o apontaram inclusive como agente causador de pneumonia grave em um soldador de Louisiana em 1994 (HAN et al., 2006).

*B. cereus* é uma bactéria Gram-positiva, formadora de esporos, móvel pela presença de flagelos peritríquios, aeróbia, anaeróbia facultativa, mesófila. Devido a sua ampla distribuição no ambiente, como o solo e a vegetação, também é considerada saprófita de solo, sendo encontrada também em alimentos de origem vegetal, animal e produtos lácteos (GRANUM et al., 1993).

Inúmeras linhagens de *B. cereus*, foram reconhecidamente envolvidas em casos de intoxicação gastrointestinal, causados pela ingestão de alimentos contaminados, produzindo diarreias e vômitos (KOTIRANTA et al., 2000). Esta natureza oportunista de *B. cereus* como patógeno é relacionada com sua capacidade de produzir inúmeros fatores de virulência não específicos, incluindo fosfolipases, hemolisinas e enterotoxinas (DROBNIEWSKI, 1993).

As razões que tornam o *B. cereus* um importante agente de contaminação na indústria de alimentos referem-se a vários fatores, como de que seus esporos estão em todos os lugares, o que torna praticamente impossível impedir sua presença em alimentos crus e em ingredientes. Contaminações ocorrem inclusive devido a deficiências na maquinaria de processamento ocasionada pelo desgaste ou inadequação, ou ainda deficiências nos materiais das embalagens. Mesmo processos como a pasteurização são eficientes para matar células vegetativas, mas não esporos. Outro fator relevante, esporos de diversas cepas de *B. cereus* têm características adesivas, o que facilita sua fixação às superfícies de encanamentos e equipamentos de processamentos com a posterior formação de biofilmes. E ainda, esporos e células vegetativas podem apresentar boa tolerância às condições como tratamentos por baixas temperaturas e baixo pH, comumente considerados métodos seguros de esterilização (HEYNDRICKX E SCHELDEMAN, 2002).

### 2.3.3. ***Bacillus thuringiensis***

*B. thuringiensis* é a bactéria mais conhecida e estudada entre as linhagens pertencentes ao grupo *B. cereus* (VILAS-BÔAS et al., 2007). A característica entomopatogênica é o que distingue *B. thuringiensis* dos demais bacilos do grupo, como *B. cereus* e deve-se à produção de proteínas denominadas *Cry*, formadas em seu ciclo de esporulação, normalmente junto com um esporo elipsoidal. Tais proteínas se juntam e cristalizam ainda dentro da célula vegetativa, formando cristais proteicos (CRICKMORE et al., 1998). Os cristais proteicos são tóxicos, principalmente para insetos das ordens Lepidoptera, Diptera, Coleoptera, Hymenoptera, Homoptera, Dictyoptera, Orthoptera, Mallophaga, além de nematoides (Strongylida, Tylenchida), protozoários (Diplomonadida) e ácaros (Acari) (FEITELSON et al., 1992; SCHNEPF et al., 1998), não apresentando qualquer ação aos mamíferos, aves, anfíbios ou répteis.

As proteínas *Cry*, inicialmente na forma de protoxinas são solubilizadas em monômeros e, através de proteases presentes no intestino do inseto alvo, são clivadas e passam para sua forma ativa. Nesta forma liga-se a receptores específicos presentes na porção apical das microvilosidades do intestino médio do alvo, levando o inseto à morte devido à formação de poros e posterior lise das células epiteliais (SCHNEPF et al., 1998).

Os primeiros produtos utilizando *B. thuringiensis* no controle de insetos pragas da agricultura e vetores de doenças começaram a ser comercializados em 1938. Em

1950, com a implementação de novas tecnologias como o uso de fermentadores e com a descoberta de novas subespécies de *B. thuringiensis*, houve um aumento na produção e comercialização dos produtos à base de *B. thuringiensis*, o que ampliou o número de insetos-alvo controlados. Sobre a utilização comercial temos que:

> A principal característica que diferencia os produtos à base de *B. thuringiensis* de outros produtos comercialmente disponíveis é a atividade entomopatogênica restrita, ou seja, a maioria dos produtos à base de *B. thuringiensis* apresenta espectro de ação somente a um ou poucos insetos. Além disso, os produtos são inócuos a outros animais, além de não contaminar o ambiente e não favorecer a rápida seleção de insetos resistentes, o que acontece frequentemente quando se utiliza produtos agroquímicos sintéticos (ARANTES, VILAS-BÔAS, 2002)

Considerando a utilização de bioinseticidas, pode-se dizer que a utilização de produtos à base de *B. thuringiensis* permite integrar estratégias de manejo no controle de insetos praga, preservando seus inimigos naturais.

### 2.4. **Genômica do Grupo Cereus**

O termo "Grupo do *Bacillus cereus*" não é um termo taxonômico, mas tem sido comumente utilizado e refere-se a um grupo de bactérias cuja relação taxonômica vem sendo amplamente discutida. Composto por seis espécies, tem como uma importante característica a taxonomia controversa de seus componentes. Na literatura pode-se encontrar diversas metodologias que já foram empregadas com intuito de realizar a caracterização molecular dessas espécies com graus variados de sucesso. Entre essas metodologias destacam-se aquelas baseadas em PCR (Rep-PCR, RAPD, Box-PCR) e as baseadas em sequenciamento de DNA (MLVA, MLST, sequenciamento do genoma total).

Duas metodologias para tipagem de cepas do grupo do *Bacillus cereus sensu lato* se encontram atualmente bastante difundidas: o Rep-PCR e o MLST. Estas metodologias, que se baseiam na análise de diferentes regiões do genoma, vem se destacando como sendo as principais ferramentas utilizadas em estudos filogenéticos dessas espécies.

O Rep-PCR que tem como alvo uma sequência Bc-Rep 26pb altamente conservada e específica para as espécies *B. cereus, B. anthracis* e *B. thuringiensis* mostrou-se capaz de gerar padrões de bandas característicos dos diferentes sorotipos de *B. thuringiensis* e fraca amplificação com *B. subtilis* e vem demonstrando ser útil na tipagem dessas espécies.

O MLST tem sido amplamente utilizado como o principal método de tipagem para analisar as relações genéticas das espécies que compõem o grupo do *B. cereus sensu lato*. Informações geradas por esta metodologia vêm permitindo uma visão integrada da estrutura genômica e populacional das espécies do grupo. Em suma, muitos estudos têm sido realizados na estrutura populacional, com ênfase particular na ocorrência e distribuição de tipos patogênicos. No entanto, a evolução e a estrutura populacional das seis espécies que compõem o grupo *B. cereus sensu lato* permanecem indefinidas (VIVONI, 2011).

## 2.5. **RNAs Não Codificadores**

### 2.5.1. **Histórico**

O primeiro ncRNA caracterizado foi uma alanina tRNA, encontrada na levedura de padeiro (*Saccharomyces cerevisiae*), tendo sua estrutura publicada em 1965. Para produzir uma amostra purificada, HOLLEY et al. utilizaram 140 kg de levedura (fermento comercial para pão) de forma a obter 1g de tRNA purificado para análise (HOLLEY et al.,1965).

Em uma segunda etapa foi realizado um trabalho de cromatografia e identificação dos terminais 5' e 3', de forma a organizar os fragmentos para o estabelecimento da sequencia do RNA. Tal trabalho garantiu  o Prêmio Nobel em Fisiologia e Medicina (1968) a Robert W. Holley, Har Gobind Khorana e Marshall W. Nirenberg, "por sua interpretação do código genético e sua função na síntese proteica" (Nobel Media Online, Abr. 2014).

Entre as estruturas originalmente propostas para tal RNAt (HOLLEY et al.,1965), a estrutura em "trevo" foi sugerida independentemente em várias publicações seguintes (MADISON et al.,1966; ZACHAU et al.; 1966, CRAMER et al., 1968; DUDOCKS et al., 1969). Contudo, a estrutura secundária de folha de trevo, finalizada mediante análise de cristalografia de raios-X, seria realizada somente em 1974, por dois grupos independentes (KIM et al., 1973, LADNER et al., 1975).

### 2.5.2. **Conceitos**

Um RNA não-codificante (ncRNA) é definido como qualquer molécula de RNA que não é traduzida em proteína. Embora o termo pequeno RNA (small RNA ou sRNA) ainda seja usado para bactérias, alguns ncRNA são muito grandes. Sinônimo de uso

menos frequente são npcRNA (non-protein-coding RNA), nmRNA (non-messenger RNA), snmRNA (small non-messenger RNA), fRNA (functional RNA). Os RNAs não codificantes pertencem a vários grupos, estando envolvidos em muitos processos celulares. Estes variam de ncRNAs de importância central que são conservados em todas ou na maior parte da vida celular através de ncRNAs mais transitórios específicos para uma ou algumas espécies estreitamente relacionadas. Os ncRNAs mais conservados são considerados fósseis moleculares ou relíquias do LUCA (Last Universal Common Ancestor) e do primitivo mundo de RNA (JEFFARES et al., 1998).

Genes para RNAs não codificantes (ncRNAs) produzem moléculas de RNA funcionais ao invés de proteínas. No entanto, quase todos os meios de identificação de genes supõem que estes codificam proteínas, e devido a isto, mesmo com a utilização de sequências genômicas completas os genes codificadores para ncRNA muitas vezes não foram detectados. Recentemente, várias sistemáticas diferentes têm identificado um número surpreendentemente grande de novos genes ncRNA. Os RNAs não codificantes parecem ser particularmente abundantes em funções que requerem o reconhecimento altamente específico de ácido nucleico, assim como na regulação pós-transcricional da expressão gênica ou modificações na orientação do RNA (EDDY, 2001).

A análise computacional de sequências dos genomas, que foi revolucionária para a análise de genes e proteínas, também deve ser capaz de responder a questões sobre o número e diversidade de genes de RNAs não codificantes. No entanto, RNAs não codificantes apresentaram à genômica computacional um novo conjunto de desafios (EDDY, 2002).

## 2.6. **Estratégias para a obtenção de sequências de ncRNAs**

### 2.6.1. **Extração das Regiões Intergênicas: Software Artemis**

O Artemis é um visualizador de sequências de DNA e também uma ferramenta de anotação que permite a verificação de características das sequências bem como dos resultados das análises dentro do contexto da sequência e de seus seis quadros de tradução (seis frames). Sua função de extração de regiões intergênicas pode ser utilizada como passo inicial para metodologias de identificação de ncRNAs. O Artemis V.16.0.0 é codificado em Java e processa sequências de formato EMBL ou GenBank e tabelas, podendo trabalhar com sequências de todos os tamanhos sem limitação, tendo sido

publicado como um software livre sob os termos da GNU General Public License (Manual Artemis: Genome Research Limited, 2014).

### 2.6.2. **Banco de Dados Rfam e o Pacote de Software Infernal**

O Rfam, sigla para RNA Families (NAWROCKI et al., 2003) é um banco de dados contendo informações sobre famílias de ncRNAs e outros elementos de RNA estruturados. É um banco de dados de anotações, de acesso livre, mantido pelo Instituto Wellcome Trust Sanger e colaboradores. Os ncRNAs, ao contrário das proteínas, muitas vezes possuem estruturas secundárias muito semelhantes, contudo sem similaridades quanto à sequência primária. O Rfam divide os ncRNAs em famílias com base na evolução de um ancestral comum, de forma que, partindo de alinhamentos múltiplos de sequências destas famílias, pode-se inferir informações sobre a sua estrutura e função (GRIFFITHS-JONES et al., 2003).

A interface no site Rfam permite aos usuários pesquisar ncRNAs por palavra-chave, nome de família, ou genoma, bem como a busca por sequências de ncRNA ou número de acesso EMBL. A base de dados de informação também está disponível para o *download*, instalação e utilização usando o pacote de software INFERNAL "INFERence of RNA ALignment" (NAWROCKI & EDDY, 2014) , que pode ser usado com o Rfam para anotação de sequências homólogas a ncRNAs conhecidos, incluindo genomas completos (GRIFFITHS-JONES et al., 2005). Infernal é um software desenvolvido para a pesquisa de estruturas de RNA e de similaridades entre sequências a partir de bancos de dados de sequências de DNA. É uma implementação para um caso especial de gramáticas de perfil estocástico (gramáticas livres de contexto) chamadas modelos de covariância (CMs). Um CM é como um perfil de sequência, mas ele pontua uma combinação entre o consenso de sequência e o consenso da estrutura secundária do RNA. Portanto em muitos casos, é mais capaz de identificar RNAs homólogos que conservam mais sua estrutura secundária do que a sua sequência primária (NAWROCKI & EDDY, 2013).

### 2.6.3. **Suíte sRNAscanner**

Classificado como um dos vários métodos de inferência computacional, o sRNAscanner é uma ferramenta para detecção de unidades transcricionais (TUs) específicas de pequenos RNAs (smalRna, sRNA) em arquivos de sequenciamento de genomas completos de bactérias. Inicialmente ele utiliza uma matriz de peso posicional

(PWM, positional weight matrix - Hertz and Stormo, 1999) para identificar os sinais transcricionais intergênicos "órfãos", posteriormente são identificados e analisados os sinais mais significativos usando um algoritmo eficiente de coordenadas, para identificar unidades de transcrição intergênicas. As unidades transcricionais identificadas são então diferenciadas em não-codantes e codantes com base na presença de sítios de ligação do ribossomo (RBS, ribosome binding sites) e códons de iniciação (SRIDHAR *et al.,* 2010). Exemplificando: uma matriz de peso posicional (PWM) descreve dados positivos de treinamento em sequências de DNA, tais dados são usados posteriormente para facilitar a identificação eficiente de sRNA em regiões intergênicas do DNA bacteriano. Assim é possível utilizar a ferramenta "PWM_create", componente da suíte sRNAscanner, para a construção de uma PWM descritiva de sinais transcricionais específicos para sRNA, experimentalmente definidos (Ex: boxes promotores -35 e -10 e sinais terminadores independentes de rho) (SRIDHAR, 2009).

### 2.6.4. **Non-coding RNA Databases Resource: NRDR**

O NRDR é uma ferramenta desenvolvida para a recuperação de informações à partir de bancos de dados de ncRNAs, fruto de um trabalho em conjunto entre os grupos de Bioinformática da USP, Instituto de Matemática e Estatística IME, Instituto de Química e Equipe do Laboratório de Bioinformática da UTFPR, Campus de Cornélio Procópio (PASCHOAL et al., 2012). Tal ferramenta, disponibilizada via web, foi desenvolvida de forma a possibilitar uma integração, incluindo em seu mecanismo de buscas o crescente número de bases de dados de ncRNAs. Este trabalho agrupa atualmente 102 bancos de dados públicos, utilizando quatro categorizações para classificar estes bancos e assim auxiliar os pesquisadores na busca das informações necessárias às suas pesquisas. Estas categorias são: família de RNA, fonte e conteúdo das informações e mecanismos de busca disponíveis. Seu banco de dados é atualizado a cada seis meses mediante processo de curadoria manual dirigida pela literatura.

## 2.7. **Anotação Gênica**

A anotação gênica é um processo que inclui vários procedimentos sequenciais como: a anotação em nível de nucleotídeos, anotação em nível de proteínas e anotação em nível de processo. Na anotação em nível de nucleotídeos procura-se encontrar a localização física das sequências de DNA e descobrir onde estão genes, RNAs, elementos repetitivos, etc. No processo de anotação em nível proteico procura-se

descobrir a provável função dos genes, identificando quais são aqueles que determinado organismo possui e quais não possui. A anotação em nível de processo, objetiva identificar vias e processos nos quais diferentes genes interagem, montando uma anotação funcional eficiente (STEIN, 2001).

# 3 OBJETIVOS

## 3.1. GERAL

Localizar e caracterizar RNAs não codificadores a partir de bancos de dados públicos de sequências, em isolados de *Bacillus anthracis, B. cereus* e *B. thuringiensis*;

## 3.2. ESPECÍFICOS

- Mediante técnicas de bioinformática, efetuar a identificação posicional dos segmentos de RNAs não codificadores nos genomas públicos de *B. thuringiensis, B. anthracis e B. cereus*;

- Possibilitar uma melhor compreensão do papel dos ncRNAs no funcionamento dos genes em processos biológicos sobretudo aqueles relacionados ao controle biológico e fatores de virulência;

### 4.1.1. Softwares: Aplicativos e Versões Utilizadas

Foram empregados softwares de uso livre para análise, processamento e alinhamento de sequências além dos softwares desenvolvidos durante as pesquisas, para tarefas específicas como os testes prévios de extração de ncRNA via coordenadas e cargas do banco de dados PostgreSQL a partir do conjunto de dados específico produzido neste trabalho, funcionalidades não disponíveis nas comunidades de software de bioinformática (Tabela 2).

**Tabela 2: Ferramentas utilizadas para o Pipeline desenvolvido neste trabalho**

| Ferramenta | Versão | Classificação | Função Utilizada | Ambiente Utilizado (S.O.) |
|---|---|---|---|---|
| **Artemis** | V.16.0.0 | Aplicativo | Visualização de sequências, Extração de IGNs | Windows 7 /Linux Ubuntu 14.04 LTS 64Bits |
| **Infernal** | V.1.1 | Aplicativo | Pesquisa de estruturas de RNA a partir bancos de sequências de DNA | Linux Ubuntu Server 64Bits |
| **Rfam** | V.1.1 | Banco de dados de ncRNA | Classificação de famílias | Linux Ubuntu 64Bits |
| **sRNAscanner** | V.1.9 | Aplicativo | Detecção de TUs | Linux Ubuntu 14.04 LTS 64Bits |
| **Sequence Ripper** | V.1, 2 e 3 | Aplicativo Java | Extração de IGNs | Windows 7 64Bits |
| **Database Loader** | V.1 | Aplicativo Java | Carga de dados | Windows 7 64Bits |
| **PostgreSQL** | V.9.1 | SGBDOR | Persistência de dados | Windows 7 64Bits |
| **PgAdmin III** | V.1.14.3 | Gerenciador de banco de dados | Administração do PostgreSQL | Windows 7 64Bits |
| **SQL Power Architect** | V.1.0.7 | Designer de Banco de dados | Criação do DER | Windows 7 64Bits |
| **Eclipse** | Indigo, Kepler | IDE | Desenvolvimento de software | Windows 7 64Bits |

Fonte: Pipeline do Laboratório de Bioinformática CCB-UEL, 2014.

### 4.1.2. O Servidor: Hardware e Software

O principal equipamento utilizado para o processamento e armazenamento de informações, a máquina servidora ou servidor, na qual foram realizadas as análises propostas, utilizou o Sistema Operacional (SO) Ubuntu 12.04 Server, e SO Windows

Server 2012 com licença de avaliação, para Junho 2013. O hardware utilizado foi um servidor HP ProLiant ML350p Gen8.

### 4.1.3. **Pipeline Desenvolvido**

A seguir é apresentado um diagrama exemplificando os procedimentos, aplicativos, bancos de dados utilizados e etapas de processamento (o pipeline) para as tarefas propostas (Figura 1):

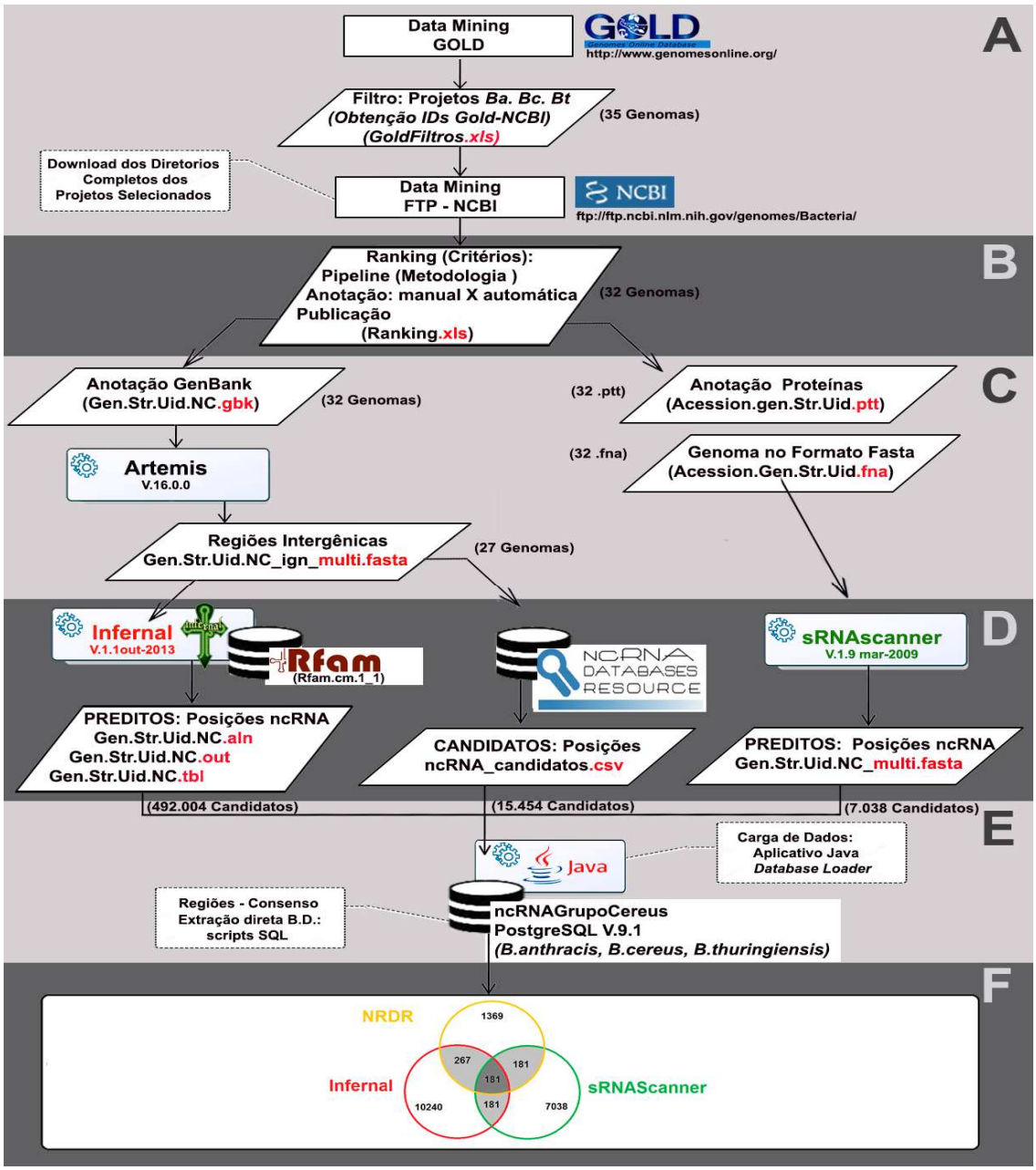


Figura 1: Etapas do Pipeline, (A) Mineração de dados; (B) Escolha do material; (C) Extração das regiões intergênicas; (D) Inferência computacional mediante três estratégias diferentes: Infernal, NRDR e sRNAscanner; (E) Compilação dos dados em um B.D. próprio; (F) Extração das informações de identificação, caracterização e agrupamento por organismo, espécie/cepa e família Rfam.

### 4.1.1. **Mineração de Dados: Linhagens utilizadas**

Devido ao grande volume de dados empregados em genômica, é comum referir-se ao processo de seleção e pesquisa de dados como mineração de dados ou *data mining*. Contudo a utilização de tal terminologia na informática refere-se ao processo de investigação de grandes quantidades de dados em busca de padrões consistentes, regras de associação, de forma a evidenciar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados. Neste trabalho o termo mineração de dados foi empregado no sentido biológico de busca e seleção.

Assim, os genomas completos dos organismos escolhidos, representantes das espécies *B. cereus*, *B. thuringiensis* e *B. anthracis* foram obtidos mediante busca nos bancos de dados Gold (Genomes Online Database) e FTP do NCBI (*National Center for Biotechnology Information*). O banco de dados GOLD foi inicialmente utilizado por proporcionar a identificação de dados referentes aos projetos de sequenciamento completos. Nesta análise preliminar, visando a identificação de genomas de referência, os projetos listados foram classificados identificando-se: a metodologia de trabalho empregada (o pipeline de cada trabalho), uma breve descrição sobre o foco dos trabalhos e suas publicações bem como o fato da citação recorrente entre tais trabalhos, sendo alguns identificados como sendo as primeiras descrições de cada genoma. Dentro desta classificação, o tipo de anotação foi um dos itens que receberam um status de maior relevância, considerando-se anotações manuais ou com algum nível de curadoria de conteúdo preferíveis aos processos de anotação completamente automatizados. A classificação realizada possibilitou o estabelecimento de um *ranking* entre os projetos pesquisados, de forma a esclarecer possíveis dúvidas sobre a relação de confiabilidade entre os projetos nas etapas posteriores. Tal *ranking* estabelece uma classificação numérica onde o menor valor significa maior importância dentro de cada grupo. Uma vez identificados, foram selecionados 35 genomas completos de acesso público entre os 27.988 disponíveis até agosto de 2013, sendo 9 de *B. anthracis*, 13 de *B. cereus*, 12 de *B. thuringiensis* e 1 de *B. weihenstephanensis*.

Como resultado da mineração de dados para classificação dos genomas foi construída uma tabela interativa com links para os artigos publicados online e extensa descrição dos projetos, contendo os seguintes itens de identificação: identificador Gold (GOLDSTAMP), identificação de ncRNA no projeto, linhagem, posição dos genes, link

alternativo (no caso de links Gold quebrados), GOLDSTAMP antigo, nome do organismo, classificação sugerida, primeiro link para o artigo online, segundo link para referências complementares, terceiro link para correções do artigo, quarto link para o documento PDF e outros formatos, publicação, referência anterior, tipo de anotação gênica, metodologia utilizada, domínio, classificação, identificador do táxon no NCBI, reino, filo, classe, ordem, família, gênero, espécie, identificador do projeto NCBI, nome do projeto no NCBI, coleção ou cultura, tipo de projeto, status do projeto, status de sequenciamento, disponibilidade, centro de sequenciamento, financiamento e o nome do pesquisador para contato.

Durante o processo de seleção de materiais para o presente estudo, foi criada uma versão em meio eletrônico da tabela contendo os links para os artigos que embasaram as publicações dos variados projetos de sequenciamento completo dos genomas selecionados. Após análise foi criada uma tabela resumida, uma versão para impressão da tabela em meio eletrônico, contendo a classificação de 32 genomas para estudo (Tabela 3):

**Tabela 3: Análise comparativa de 32 projetos selecionados após mineração dos dados de interesse nos bancos de dados públicos GOLD e FTP-NCBI**

| GOLDSTAMP | NCBI PROJECT ID | Identificação de ncRNA | Organismo Linhagem | Ranking* | Sobre a Publicação | Referência Anterior | Tipo de Anotação | Metodologia |
|---|---|---|---|---|---|---|---|---|
| Gc00993 | 33543 | Sim | *B. anthracis A0248* | --- | --- | --- | --- | --- |
| Gc00136 | 309 | Sim | *B. anthracis Ames* | 1 | Boa documentação: Revista Nature V.423, maio 2003; | Primeira descrição; | Sequência editada manualmente (TIGR) e adicional PCR; | Aponta Metodol. para outra Referência. Reações de sequenciamento para fechar lacunas, melhorar a cobertura, e resolver as ambiguidades de sequência. Cobertura de regiões 'gap' (SP1772) por sequenciamento assistido por transposon (New England Biolabs PGPS Transposon Kit) e mapeamento das inserções de transposons antes da montagem; |
| Gc00189 | 10784 | Não | *B. anthracis Ames Ancestor A2084* | 3 | Selecionado para suprir possível carência de material com a amostra *B. anthracis CI;* | Primeira descrição de genoma completo; | --- | Montagem do genoma: Celera assembler; Electropherograma no NCBI; |
| Gc00984 | 31329 | Sim | *B. anthracis CDC 684* | --- | --- | --- | --- | --- |
| Gc01313 | 36309 | Sim | *B. anthracis CI* | 2 | *B.cereus biovar anthracis*: considerar *B. cereus;* | --- | CDS: cura manual, verif. por comp. com as bases de dados públicas: SwissProt, GenBank, ProDom, COG, e Prosite utilizando o software de anotação ERGO; | CDS e ORFs previstos com YACOP, usando: Glimmer ORF-finders, Critica e Z-curve; |
| Gc02203 | 49361 | Não | *B. anthracis H9401* | 4 | --- | --- | Anotação via Uniref90, nos bancos de dados: NCBI nr, COG e KEGG; | Sequenciamento: 454 GS-FLX (454 Life Sciences Corp. Basel, Suiça). Análise: utilizando GLIMMER, tRNAscan-SE e RNAmmer. Comparações na estrutura genômica foram realizadas com ACT e MUMmer; |
| Gc00195 | 10878 | Não | *B. anthracis Sterne* | | --- | --- | --- | --- |
| Gc00975 | 31307 | Sim | *B. cereus 03BB102* | 2 | Boa documentação: Utilização estudos filogenia; | --- | --- | ANÁLISES: fenotípica, antigênicas e análise da sequência de cepas paga; |
| Gc00897 | 17715 | Sim | *B. cereus AH187 (F4810/72)* | 6 | Artigo pobre e antigo 1987, material pobre, só lista a sequência; | --- | --- | --- |
| Gc00917 | 17711 | | *B. cereus AH820* | | --- | --- | --- | --- |
| Gc00173 | 74 | | *B. cereus ATCC 10987* | | --- | --- | --- | --- |
| Gc00135 | 384 | Não | *B. cereus ATCC 14579* | 4 | Infecção de *B. thuringiensis* pelo fago Bam35; | --- | --- | --- |
| Gc00903 | 17731 | --- | *B. cereus B4264* | --- | --- | --- | --- | --- |
| Gc00617 | 13624 | Não | *B. cereus cytotoxis NVH 391-98* | 5 | O artigo pesquisa toxinas (Cytotoxin K - CytK), não tem o foco deste trabalho; | --- | --- | --- |
| Gc00215 | 12468 | Não | *B. cereus E33L (ZK)* | 1 | --- | --- | Anotação e predição mediante Glimmer. Identificação de tRNAs mediante tRNAScan-SE; | Análise e alinhamento: MUMmer2, ACT Sanger Software e Pipmaker. Lista de ortólogos: script perl que determina melhores hits bidirecionais; Para identificar elementos IS em B. thuringiensis 97-27 e B. cereus e E33L comparação com elementos IS presentes em outros membros do grupo B. cereus, todo os elementos IS usados como motifs em BLAST X Linhagens B. anthracis (Ames, A2012, e Sterne), e B. cereus (E33L, ATCC, 14579); |
| Gc02061 | 15716 | --- | *B. cereus F837/76* | 3 | --- | Artigo remete ao Trabalho do Al Hakan como o mais similar encontrado; | --- | Genoma completo: shotgun usando a tecnologia Sanger dideoxy. Predição de genes, CDS e atribuição de funções: suite anotação RAST. Putativa para profagos indutíveis não detectadas: ferramenta Prophinder, da base de dados ACLAME; |
| Gc02329 | 171769 | --- | *B. cereus FRI-35* | --- | --- | --- | --- | --- |
| Gc00914 | 17733 | Sim | *B. cereus G9842* | --- | --- | --- | --- | --- |
| Gc02105 | 18977 | --- | *B. cereus NC7401* | --- | --- | --- | --- | --- |
| Gc00895 | 16220 | --- | *B. cereus Q1* | --- | --- | --- | --- | --- |
| Gc00463 | 18255 | Não | *B. thuringiensis Al Hakam* | 1 | Indicado para Estudos Evolutivos: relações entre organismos do grupo do *B. cereus;* | Artigo Principal da Série de Estudos | Anotação Manual | Predição genes: Glimmer; Phred/Phrap/Consed: montagem de seqs.e qualidade; Shotgun sequencing: montagem de reads com phrap paralelo (High Performance Software, LLC); Correções de montagem: transposon bombing (Epicenter Biotechnologies) of bridging clones. Fechamento de Gaps entre contigs: edição no Consed; |
| Gc00196 | 10877 | Não | *B. thuringiensis sv konkukian 97-27* | 4 | Artigo com excelente documentação; | --- | Anotação Manual | Predição de genes: Glimmer; Sanger; Definição de Genes e classes funcionais adicionadas manualmente por um grupo de anotadores utilizando Blast; |
| Gc01669 | 60447 | Não | *B. thuringiensis sv. finitimus YBT-020* | 2 | Artigo considerado indicado para Estudos Evolutivos; | Remete aos trabalhos Al Hakan, Konkukian,BMB171; | --- | Sequenciamento: Illumina SolexaGA; Montagem de Contigs: SOAPdenovo; |
| Gc01699 | 43737 | Não | *B. thuringiensis sv chinensis CT-43* | 5 | --- | Remete a BMB171; | --- | --- |
| Gc01286 | 43631 | Não | *B. thuringiensis BMB171* | 3 | --- | --- | --- | Sequenciamento: Pyrosequenciamento 454 GS-FLX; Montagem: Newbler 454 Life Sciences; Relacionamento entre contigs: PCR multiplex, shotgun ABI 3770; Qualidade na montagem: Phred Phrap Consed; |
| Gc02506 | 185468 | Não | *B. thuringiensis sv. kurstaki HD73* | 6 | Possui referencia não Publicada; | --- | --- | Montagem: newbler v. 2.3; Sequenciamento: Sanger dideoxy sequencing; 454; Illumina; |
| --- | --- | --- | *(Novo) B. thuringiensis Bt407* | 7 | Possui referencia não Publicada. Não Submetido a revisão final NCBI; | --- | --- | Sanger dideoxy sequencing; 454; |
| --- | --- | --- | *(Novo) B. thuringiensis HD 789* | 11 | Variedade usada para mapeamento e análise de SNP; | --- | --- | Programa: NCBI Pipeline |
| --- | --- | --- | *(Novo) B. thuringiensis MC28* | 8 | Possui referência não Publicada; | --- | --- | Montagem: SOAPdenovo v. 1.05; LaserGene v. 6.0; Vector v. 9.0; Sequenciamento: Sanger dideoxy sequencing; Illumina; |
| --- | --- | --- | *(Novo) B. thuringiensis serovar IS5056* | 9 | Possui referência não Publicada. Não Submetido a revisão final NCBI; | --- | --- | Montagem: GS Data Analysis Software v. 2.8; Sequenciamento: 454; Illumina; |
| Gc02483 | 29717 | --- | *B. thuringiensis Bt407* | 12 | Repetido - ver acima; | --- | Anotação: Genomas Microb. Integrados pipeline (IMG), dados curados mediante comparações BLAST X Swiss-Prot/trEMBL; | Montagem: Newbler v. v2.6; Sequenciamento: Sanger dideoxy sequencing; 454; |
| Gc02477 | 171845 | Não | *B. thuringiensis HD-771* | 10 | Possui referencia não Publicada. Não Submetido a revisão final NCBI; | --- | Anotação Automática: Pipeline Group; | Programa: NCBI Pipeline; |

*Ranking: Classificação de interesse do material para este estudo

Fonte: Ranking Bancos de dados GOLD e NCBI, 2013

### 4.1.2. **Extração das Regiões Intergênicas: Artemis**

Para otimizar as etapas posteriores de descoberta de ncRNAs, o material resultante da etapa de mineração de dados anterior foi submetido ao processamento pelo Artemis (V.16.0.0) para extração das regiões intergênicas. Cada linhagem em estudo foi processada gerando um arquivo em formato multifasta contendo as sequências intergênicas identificadas. Entre os genomas pesquisados para a etapa de testes do pipeline, o Artemis apresentou restrições no processamento de alguns itens devido a problemas de anotação em tais arquivos (para detalhes, ver protocolo de trabalho).

Também após a análise proposta e classificação dos materiais, decidiu-se pela não inclusão dos itens abaixo no pipeline de processamento, devido a dúvidas quanto à sua classificação e também devido a não apresentarem pares para comparação, como o caso da existência de um único genoma de *B. weihenstephanensis* sequenciado. Os genomas assim classificados foram então considerados dispensáveis para a análise proposta. Incertezas foram geradas devido a uma modificação na padronização dos identificadores de projetos entre os banco de dados GOLD e o FTP do NCBI, durante o período onde as análises já haviam sido iniciadas. Tal modificação ocorreu entre 23 de março de 2013 e 13 de janeiro de 2014:

- Bacillus_anthracis_A16_uid40303: identificador modificado;
- Bacillus_anthracis_A16R_uid40353: identificador modificado;
- Bacillus_thuringiensis_YBT_1518_uid63189: identificador modificado;
- Bacillus_weihenstephanensis_KBAB4_uid13623: único genoma para a espécie;

Foi utilizada a funcionalidade "Intergenic Features" gerando arquivos fasta com as sequências intergênicas para cada linhagem em estudo de forma a preparar o conjunto de dados para as próximas etapas do pipeline (Figura 20).

### 4.1.1. **Métodos de Inferência Computacional: Identificação de ncRNAs**

Os procedimentos de inferência de elementos como ncRNAs mediante utilização de ferramentas computacionais específicas são também chamados de métodos de descoberta computacional. Os dados das regiões intergênicas identificadas pelo Artemis nos 26 genomas selecionados foram gerados e salvos em arquivos multifasta e então submetidos aos três métodos diferentes de inferência computacional para identificação de RNAs não codificadores (ncRNAs). O primeiro método foi o

processamento via Infernal V.1.1 / banco de dados Rfam V.11.0, o segundo foi o processamento via sRNAscanner V.1.9, e finalmente uma análise comparativa com os dados produzidos no pipeline da UTFPR (Candidatos, tabela nrdr_candidatos com base na ferramenta de busca, agrupamento e recuperação de informações em bases de dados de ncRNA, o *Non-coding RNA Databases Resource* (NRDR). Os resultados dos três métodos adotados foram combinados de forma a obter-se 6 estratégias de busca (leia-se, ">" Contém):

(1) infernal>nrdr; (2) infernal>srna; (3) srna>infernal; (4) srna>nrdr; (5) nrdr>infernal; (6) NRDR>srna;

#### *4.1.1.1. Rfam e Infernal*

Nesta etapa de inferência computacional, os arquivos contendo as sequências intergênicas (26 entradas) foram submetidos localmente (servidor HP ProLiant) ao processamento do software Infernal V.1.1 e pesquisados contra o banco de dados Rfam V.1.1., tendo sido gerados 26 arquivos de saída ou resultados (outputs).

#### *4.1.1.2. sRNAscanner*

Como requisito do programa sRNAscanner, esta etapa também necessitou de execução em ambiente Linux. As análises propostas foram executadas no servidor do laboratório de Bioinformática. A execução de rotinas corretivas de processamento, quando necessárias, foram realizadas em um laptop Acer Aspire com ambiente Windows em dual boot com Linux.

O sRNAscanner recebeu como entradas para cada genoma, os arquivos correlatos provenientes de seus próprios projetos de sequenciamento (originários no FTP-NCBI), não necessitando utilizar as sequências intergênicas produzidas pelo Artemis.

#### *4.1.1.3. PIPELINE NRDR*

Os arquivos multifasta contendo as regiões intergênicas identificadas na etapa de processamento pelo Artemis, foram submetidas ao pipeline de identificação de ncRNAs da UTFPR. Os dados das sequências intergênicas foram utilizados como parâmetros com base na ferramenta Non-coding RNA Databases Resource (NRDR) para gerar um relatório de dados de anotação, conforme o formato a seguir:

**(Formato) >**Genero_especie_cepa_projeto-ID_NCBI-misc_feature misc_feature undefined product posicao_inicial:posicao_final fita_tamanho

**(Exemplo)>**Bacillus_anthracis_A0248_uid59385-NC_012659-misc_feature misc_feature undefined product 106165:106396 forward_232

Foram encontradas 15.454 ocorrências de anotações para os genomas selecionados, tendo sido selecionados os melhores resultados (1416), posteriormente filtrados com base nas identidades e coberturas das consultas (queries). Os filtros finais de seleção restringiram a amostra a 1369 resultados.

#### 4.1.1.4. Carga de Dados

Devido ao volume de dados produzido, foi necessário o desenvolvimento de um aplicativo para efetuar a inserção de dados inicial no banco. Testes de carga manual foram realizados utilizando a interface do PgAdminIII no PostgreSQL V.9.1. Nestes testes foram executadas cargas de arquivos de texto (*.csv) , formatados para tratamento com linguagem SQL. Após a formatação final produziu-se um aplicativo java automatizando a carga de dados geral. A execução foi realizada em um Laptop Acer Aspire 4745-7494, tendo sido utilizados os Sistemas Operacionais Windows Seven Ultimate - service pack 1 64Bits em dual boot com Ubuntu 14.04 LTS 64Bits. Após esta primeira etapa de carga, os dados foram tratados conforme os procedimentos descritos nas etapas de 3 a 5 do item 1.484.

### 4.1.2. **Banco de Dados Final: PostgreSQL**

A versão de banco de dados utilizada foi o PostgreSQL V.9.1. , no qual foi criado o banco de nome grupo_cereus que recebeu as cargas dos dados provindos dos métodos de processamento e descoberta executados nas etapas anteriores, conforme exemplificado na Figura 2, onde é apresentado o diagrama de entidades e relacionamentos (DER) produzido para a confecção do modelo físico do banco de dados criado. Para atender a demanda de geração de relatórios, a tabela organismos recebeu os dados de nomenclatura das cepas de forma a facilitar a extração de relatórios.

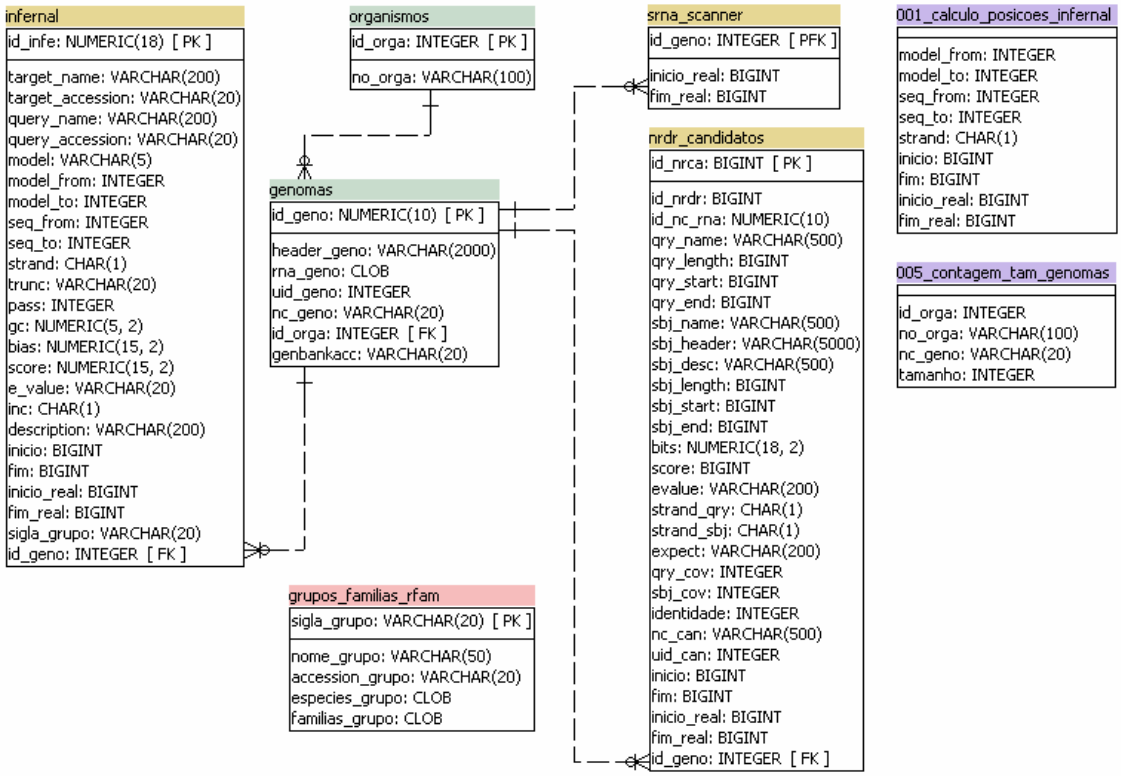


**Figura 2:** Diagrama de entidades e relacionamentos para o banco de dados do projeto. As tabelas físicas infernal, srna_scanner e nrdr_candidatos armazenam os dados gerados pelas estratégias de identificação de ncRNAs, respectivamente à partir dos programas: Infernal, sRNAscanner e NRDR (identificados à partir da mineração de dados no banco de ncRNA da UTFPR (Dr. Paschoal – dados não publicados). A tabela física grupos_familias_rfam compõe a caracterização das sequências obtidas em termos de famílias de ncRNA. Foram criadas tabelas relacionais descritivas para as 2.208 famílias Rfam, à partir dos dados públicos disponibilizados pelos mantenedores, o site FTP Rfam e o instituto SANGER. As tabelas numeradas (alto à direita) representam visões (views), são tabelas virtuais criadas dinamicamente à partir do banco de dados quando um acesso à visão (um relatório) é solicitado.

Os candidatos resultantes de cada estratégia foram selecionados e agrupados mediante seleção de regiões contíguas sobrepostas, para investigação através da aplicação dos seguintes critérios ou filtros de qualidade:

**Infernal:** Somente candidatos que atingiram o limiar mínimo de inclusão do Infernal (significando um E-value de 0.01), e que tenham também apresentado um score maior ou igual a 25 (score, pontuação da comparação target / query), representado pela expressão:

**( (**inc = !**)** **E** score >=25**)**

**NRDR:** Foi utilizada outra estratégia combinada tendo um mínimo de 90% de cobertura da query ou do subject aliados a uma identidade de 90%, que podemos representar pela expressão:

**( (**Cobertura Qry >= 90 **OU** Cobertura Sbj >= 90 **) E** Identidade >= 90**)**

A Figura 3: Demonstração do agrupamento realizado. Da esquerda para a direita: script SQL para seleção dos casos de sobreposição de regiões identificadas, demonstração dos casos correspondentes. demonstra a composição dos resultados para o agrupamento inicial e construção do Diagrama de Venn com os totais. Para economia de recursos computacionais, ao invés de comparar as sequências, os critérios de busca SQL utilizados foram utilizados de forma a identificar cada correspondência positiva (match) como 1 sendo o inverso (mismatches) identificado como 0.

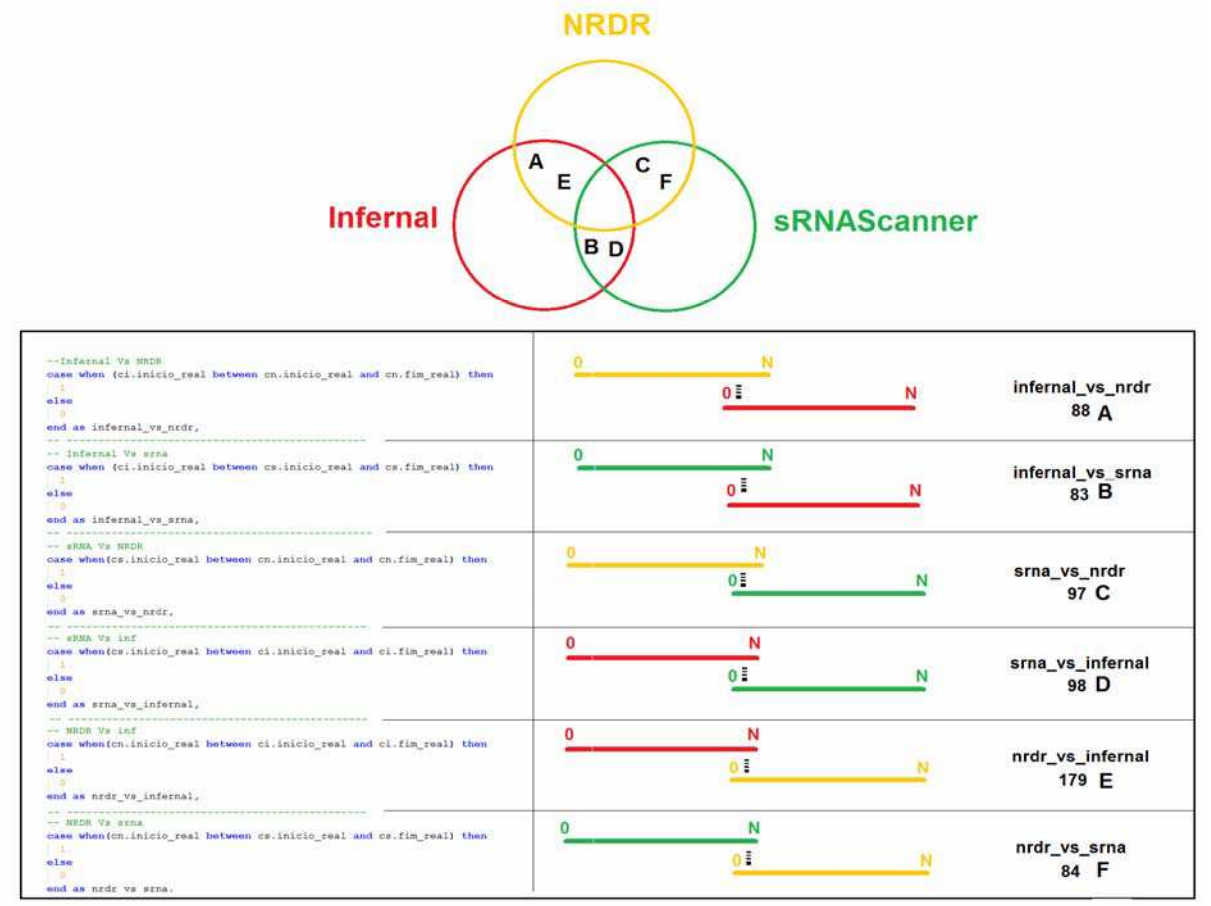


**Figura 3: Demonstração do agrupamento realizado. Da esquerda para a direita: script SQL para seleção dos casos de sobreposição de regiões identificadas, demonstração dos casos correspondentes.**

Uma vez obtida a interseção entre os resultados identificados pelos pares das estratégias, os dados receberam o seguinte procedimento de agrupamento para as 3 estratégias, obtendo os totais conforme Figura 4: Demonstração do agrupamento final. Da esquerda para a direita: script SQL para seleção dos casos de sobreposição de regiões identificadas, demonstração dos casos correspondentes.:

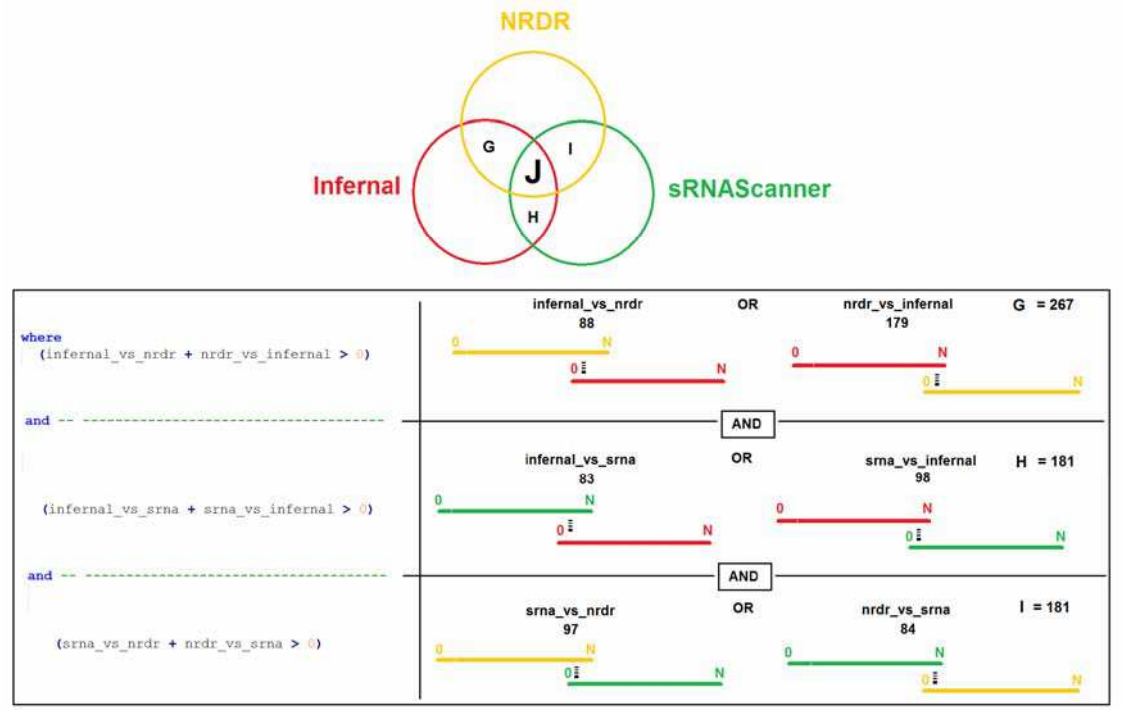


**Figura 4: Demonstração do agrupamento final. Da esquerda para a direita: script SQL para seleção dos casos de sobreposição de regiões identificadas, demonstração dos casos correspondentes.**

A descrição das etapas de operações de definição (DDL) e manipulação de dados (DML) realizadas no banco de dados PostgreSQL, para a criação, carga e formatação dos dados do banco *grupo_cereus* podem ser verificadas no item Protocolo (Anexos).

## 5 RESULTADOS E DISCUSSÃO

A aplicação dos métodos de inferência computacional, sobre os arquivos das linhagens selecionadas provenientes do FTP do NCBI permitiu a criação de um banco de dados para a análise genômica comparativa, visando a detecção de possíveis ncRNA candidatos. O número de registros e descrições das tabelas do banco de dados produzido é apresentado na Tabela 4:

**Tabela 4: Entidades componentes do banco de dados deste trabalho**

| Nomenclatura | Descrição | Nº Registros |
|---|---|---|
| **organismos** | Tabela de identificação de cepas | 26 |
| **nrdr_candidatos** | Tabela de ncRNA contendo os melhores candidatos selecionados no Laboratório da UTFPR | 1369 |
| **genomas** | Tabela de genomas completos | 26 |
| **infernal** | Tabela de ncRNA candidatos identificados pelo Infernal/Rfam com base ... | 492004 |
| **srna_scanner** | Tabela de ncRNA candidatos identificados pelo sRNAscanner com base ... | 7038 |
| **familias_rfam_ftp*** | Tabela de famílias Rfam contendo as espécies por família | 383004 |
| **familias_rfa_web*** | Tabela de famílias Rfam contendo suas descrições | 2208 |
| **grupos_familias_rfam** | Tabela de agrupamentos arbitrários deste projeto (siglas Rfam) | 534 |

Fonte: Banco de dados Grupo_cereus, 2014 (nomes das entidades em minúsculas segue o padrão do banco de dados). *Representam entidades básicas, não apresentadas no DER.

Embora algumas linhagens pré-selecionadas tenham sido excluídas da análise devido a questões de processamento nos programas de predição ou por questões de classificação (*B. cereus biovar anthracis* CI uid50615, *B. cereus* FRI 35, *B. cereus* Q1, *B. thuringiensis* HD 771, *B. thuringiensis* HD 789, *B. anthracis* A16, *B. anthracis* A16R, *B. thuringiensis* YBT 1518, *B. weihenstephanensis* KBAB4), o número de linhagens restantes pôde garantir a representatividade de organismos das três espécies selecionadas para as análises, já que 26 cepas pertencentes ao grupo do *B. cereus sensu lato* foram analisadas (Tabela 5).

**Tabela 5: Identificadores das cepas selecionadas após mineração de dados**

| Cepa | ID* NCBI | ID* |
|---|---|---|
| ***Bacillus anthracis* str. A0248** | NC_012659 | 1 |
| ***Bacillus anthracis* str. 'Ames Ancestor'** | NC_007530 | 2 |
| ***Bacillus anthracis* str. Ames** | NC_003997 | 3 |
| ***Bacillus anthracis* str. CDC 684** | NC_012581 | 4 |
| ***Bacillus anthracis* str. H9401** | NC_017729 | 5 |
| ***Bacillus anthracis* str. Sterne** | NC_005945 | 6 |
| ***Bacillus cereus* 03BB102** | NC_012472 | 7 |
| ***Bacillus cereus* AH187** | NC_011658 | 8 |
| ***Bacillus cereus* AH820** | NC_011773 | 9 |
| ***Bacillus cereus* ATCC 10987** | NC_003909 | 10 |
| ***Bacillus cereus* ATCC 14579** | NC_004722 | 11 |
| ***Bacillus cereus* B4264** | NC_011725 | 12 |
| ***Bacillus cereus* E33L** | NC_006274 | 13 |
| ***Bacillus cereus* F837/76** | NC_016779 | 14 |
| ***Bacillus cereus* G9842** | NC_011772 | 15 |
| ***Bacillus cereus* NC7401** | NC_016771 | 16 |
| ***Bacillus cereus* Q1** | NC_011969 | 17 |
| ***Bacillus thuringiensis* str. Al Hakam** | NC_008600 | 18 |
| ***Bacillus thuringiensis* BMB171** | NC_014171 | 19 |
| ***Bacillus thuringiensis* serovar chinensis CT-43** | NC_017208 | 20 |
| ***Bacillus thuringiensis* serovar finitimus YBT-020** | NC_017200 | 21 |
| ***Bacillus thuringiensis* serovar konkukian str. 97-27** | NC_005957 | 22 |
| ***Bacillus thuringiensis* serovar kurstaki str. HD73** | NC_020238 | 23 |
| ***Bacillus thuringiensis* Bt407** | NC_018877 | 24 |
| ***Bacillus thuringiensis* MC28** | NC_018693 | 25 |
| ***Bacillus thuringiensis* serovar *thuringiensis* str. IS5056** | NC_020376 | 26 |

* Identificadores das cepas nos bancos de dados, NCBI e Grupo_cereus (este trabalho). Fonte: Banco de dados grupo_cereus, 2014;

### 5.1. **Busca: Total de Candidatos Identificados**

A análise dos resultados combinados dos três métodos de inferência computacional e a intersecção das seis estratégias de agrupamento dos candidatos identificados por cada método, permitiu a conclusão do processo de busca e identificação dos ncRNAs candidatos por posição nos respectivos genomas das cepas analisadas. O número total de candidatos identificados por estratégia bem como os candidatos indicados pela junção de duas ou três estratégias é exibido conforme o diagrama de Venn (Figura 5).

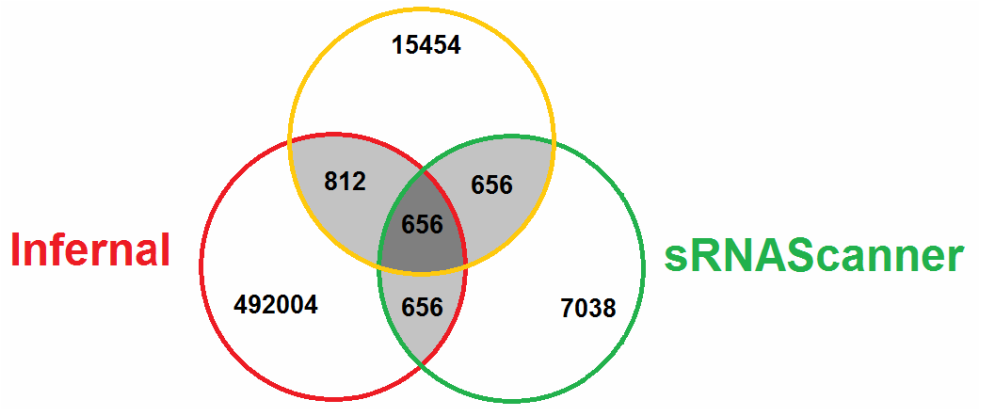


**Figura 5: Número total de ncRNA candidatos por método de descoberta (inferência).**

Os totais de candidatos identificados pelas estratégias, Infernal, sRNAScanner e NRDR foram, respectivamente: 492004, 7038, 15454. O total de candidatos identificados pela intersecção entre procedimentos é exibido ao centro do diagrama e sua composição será detalhada adiante (Figura 6). Foi obtido um total de 656 candidatos identificados pelas intersecções entre as estratégias Infernal vs sRNAScanner, sRNAScanner vs NRDR e a união entre as três estratégias.

Este resultado inicial é relativo ao processo de trabalho utilizado por cada estratégia computacional, indicando uma eficiência de identificação muito próxima, não devendo ser utilizado para sugerir a proximidade entre os organismos investigados.

Para finalizar a identificação e também preparar o conjunto de dados para a caracterização na próxima etapa, além da aplicação dos novos critérios, foram excluídos os candidatos repetidos. O diagrama de Venn produzido após aplicação dos novos critérios é exibido pela Figura 6, na qual foram deixados para análise somente os ncRNA candidatos com estrutura primária não repetida, que chamaremos de sequência primária.

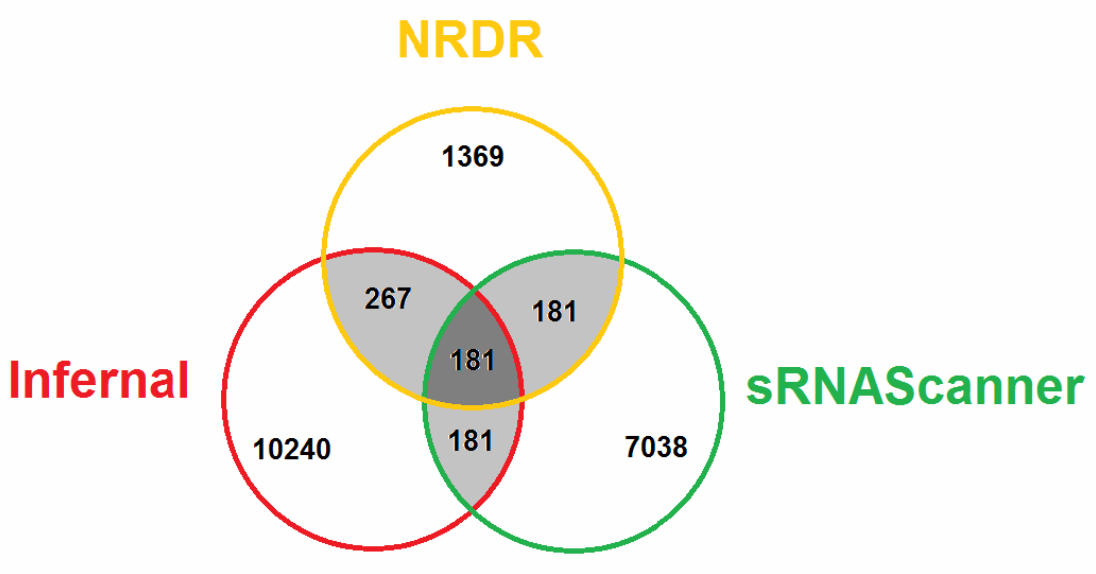


**Figura 6: Número final de ncRNA candidatos selecionados e exclusivos.**

O total geral de candidatos identificados mediante a utilização das três estratégias de inferência computacional foi de 18.647. Os diferentes métodos de descoberta empregados pelas estratégias utilizadas, refletiram-se nas diferenças entre o número de identificações das estratégias do Infernal, sRNAScanner e NRDR, respectivamente: 10.240, 7.038, 1.369. O resultado final para as três estratégias revela 181 sequências de ncRNA candidatos (Figura 6), e os totais obtidos pela interseção de duas ou mais estratégias evidenciaram os seguintes números de candidatos:

**(1)** Infernal vs NRDR: 88;

**(2)** Infernal vs sRNAScanner: 83;

**(3)** sRNAScanner vs Infernal: 98;

**(4)** sRNAScanner vs NRDR: 97;

**(5)** NRDR vs Infernal: 179;

**(6)** NRDR vs sRNAScanner: 84;

Após a identificação, foi encontrada uma distribuição desigual no número de ncRNA candidatos para as 26 cepas de *B. cereus sensu lato*. Quando as 181 sequências de ncRNA candidatas são separadas por espécie, temos 34 sequências para *B. anthracis* (Tabela 6), para 68 sequências para *B. cereus* (

Tabela 7) e 79 sequências para *B. thuringiensis* (Tabela 8) .

Os números totais de candidatos identificados, relacionados pela espécie e família onde foram localizados também são apresentados na Figura 18. Desta forma, a estratégia mais estringente (restritiva) para a identificação das sequencias, foi a realizada na etapa de identificação de candidatos quando se utilizou a comparação baseada no NRDR, obtendo-se 1369 candidatos quando comparado com o INFERNAL que identificou 10240 e o sRNAScanner com 7038 (Figura 6). A maior estringência da classificação feita com base no NRDR pode ser explicada pelos filtros aplicados, embora tal ferramenta possa proporcionar uma busca em um total de 102 bancos de dados diferentes (PASCHOAL *et.al.*, 2012).

**Tabela 6: Número de ncRNA candidatos distribuídos por cepa em *Bacillus anthracis*, com identificação positiva nas três estratégias de inferência computacional**

| ID* | Cepa | Nº Candidatos** |
|---|---|---|
| **1** | *Bacillus anthracis* str. A0248 | 8 |
| **2** | *Bacillus anthracis* str. 'Ames Ancestor' | 3 |
| **3** | *Bacillus anthracis* str. Ames | 8 |
| **4** | *Bacillus anthracis* str. CDC 684 | 6 |
| **5** | *Bacillus anthracis* str. H9401 | 8 |
| **6** | *Bacillus anthracis* str. Sterne | 1 |

* Identificadores das cepas no banco de dados, ** Sequências primárias exclusivas;
Fonte: Banco de dados grupo_cereus, 2014;

**Tabela 7: Número de ncRNA candidatos distribuídos por cepa em *Bacillus cereus*, com identificação positiva nas três estratégias de inferência computacional**

| ID* | Cepa | Nº Candidatos** |
|---|---|---|
| **7** | *Bacillus cereus* 03BB102 | 5 |
| **8** | *Bacillus cereus* AH187 | 10 |
| **9** | *Bacillus cereus* AH820 | 10 |
| **10** | *Bacillus cereus* ATCC 10987 | 8 |
| **11** | *Bacillus cereus* ATCC 14579 | 2 |
| **12** | *Bacillus cereus* B4264 | 4 |
| **13** | *Bacillus cereus* E33L | 6 |
| **14** | *Bacillus cereus* F837/76 | 9 |
| **15** | *Bacillus cereus* G9842 | 8 |
| **16** | *Bacillus cereus* NC7401 | 4 |
| **17** | *Bacillus cereus* Q1 | 2 |

* Identificadores das cepas no banco de dados, ** Sequências primárias exclusivas;
Fonte: Banco de dados grupo_cereus, 2014;

**Tabela 8: Número de ncRNA candidatos distribuídos por cepa em *Bacillus thuringiensis*, com identificação positiva nas três estratégias de inferência computacional**

| ID* | Cepa | Nº Candidatos** |
|---|---|---|
| **18** | *Bacillus thuringiensis* str. Al Hakam | 4 |
| **19** | *Bacillus thuringiensis* BMB171 | 15 |
| **20** | *Bacillus thuringiensis* serovar chinensis CT-43 | 3 |
| **21** | *Bacillus thuringiensis* serovar finitimus YBT-020 | 7 |
| **22** | *Bacillus thuringiensis* serovar konkukian str. 97-27 | 9 |
| **23** | *Bacillus thuringiensis* serovar kurstaki str. HD73 | 12 |
| **24** | *Bacillus thuringiensis* Bt407 | 7 |
| **25** | *Bacillus thuringiensis* MC28 | 11 |
| **26** | *Bacillus thuringiensis* serovar *thuringiensis* str. IS5056 | 11 |

* Identificadores das cepas no banco de dados, ** Sequências primárias exclusivas;
Fonte: Banco de dados grupo_cereus, 2014;

A ocorrência diferencial de ncRNA candidatos putativos, como destacado por EDDY, 2002, é um fato biológico que sugere uma rede de diversidade quanto aos genes de RNAs não codificantes.

A relação entre o estudo das sequências cromossômicas e plasmidiais pode ser estabelecida em estudos como ZHENG et.al, 2015, onde é sugerido que em *B. cereus sensu lato* os plasmídeos são vetores de genes cromossômicos redundantes.

A distribuição desigual demonstrada (TabelasTabela 6,

Tabela 7 e Tabela 8) bem como a relação entre o estudo das sequências plasmidiais e cromossômicas, podem estar reafirmando o estudo de ncRNAs, entre outros elementos, como uma das ferramentas corretas em questões de resolução taxonômica no grupo do *B. cereus*, onde as características e o perfil de patogenicidade são determinados por genes frequentemente presentes em plasmídeos (item 2.3) e outros fatores de virulência são codificados por genes presentes geralmente nos cromossomos (KLEE, 2010).

Os maiores números de ocorrência de ncRNA candidatos foram encontrados em cepas de *B. thuringiensis*, sendo a maior ocorrência na cepa BMB171, com 15 candidatos identificados. Tal resultado vai de encontro ao trabalho de HE et al., 2010, onde esta cepa foi considerada como a mais indicada para pesquisas de regulação e expressão gênica, especialmente com relação à produção de cristais proteicos e estudo da rede metabólica, tendo sido utilizada para produção de uma cepa mutante acristalífera de mesmo nome.

Em outro grupo, os organismos com o segundo maior número de ncRNA candidatos identificados em *Bacillus cereus stricto sensu*, foram as cepas *B. cereus* F837/76 e *B. cereus* AH187 e AH820, respectivamente com contagem de 9 e 10 candidatos encontrados pelas três estratégias utilizadas.

Análises de sequências de nucleotídeos (segmentos parciais) e análises genômicas completas evidenciam que a evolução bacteriana tem frequentemente ocorrido por fluxo horizontal entre diversas espécies e gêneros. A transferência gênica se dá mediante transformação, transdução ou conjugação entre diferentes organismos, em uma ampla variedade de nichos no ambiente. Este conhecimento é necessário para o entendimento da evolução plasmidial e sua ecologia (DAVISON, 2002).

Plasmídeos têm desempenhado papéis importantes na evolução dos genomas bacterianos, com grande impacto sobre o metabolismo da célula hospedeira (a própria bactéria). Foi sugerido por ZHENG et al., 2015, que para todo o grupo *B. cereus*, genes adaptativos são preservados em ambos os plasmídeos e cromossomos. Contudo, em uma única célula, genes homólogos em plasmídeos e no cromossomo são controlados por diferentes reguladores, reduzindo o custo biológico para a manutenção de genes redundantes (ZHENG et al., 2015).

Um dos objetivos secundários deste trabalho, com a união das três estratégias de inferência computacional empregadas, aponta um crescente número de ncRNA candidatos putativos, partindo de *B. anthracis* (em média 5) passando por *B. cereus* (em média 6) até chegar a *B.thuringiensis* (em média 8).

A distribuição dos 181 ncRNA candidatos obtidos no presente trabalho bem como seu registro em banco de dados, associados à implicação de trabalhos sobre a relação entre plasmídeos e cromossomos, como em ZHENG et al., 2015, e HE et al., 2010, nos traz uma nova perspectiva. Seria justificável uma segunda etapa de busca e caracterização, onde o pipeline utilizado neste trabalho seja executado tendo como objeto de estudo a relação entre plasmídeos e cromossomos para os organismos aqui investigados.

Com relação ao banco de dados gerado durante este trabalho, o cruzamento dos resultados de identificação e caracterização com novas estratégias enfatizando os plasmídeos, certamente agregaria maior compreensão tanto sobre a as distinções do grupo, quanto sobre questões específicas como a característica entomopatogênica de *B. thuringiensis*, como mencionado no item 2.3.2 (FEITELSON et al., 1992; SCHNEPF et al., 1998).

5.2. **Caracterização: Famílias de RNAs**

### 5.2.1. **Distribuição de Candidatos Por Família**

Além da identificação posicional dos candidatos nos genomas dos organismos estudados, as sequências finais foram submetidas ao Rfam de forma a serem caracterizadas quanto às famílias de RNAs a que pertenciam. O resultado da busca e processamento "batch" no website Rfam foi de 422 registros, sendo que os 181 candidatos submetidos foram classificados em 23 famílias de RNA. Uma breve descrição para as famílias encontradas é exibida na Tabela 9.

**Tabela 9: Listagem de Famílias de RNAs atribuídas aos 181 candidatos provenientes da intersecção das três estratégias de inferência computacional empregadas.**

| Nome | Rfam Acession | Descrição* |
|---|---|---|
| **5S_rRNA** | RF00001 | 5S ribosomal RNA |
| Bacteria_large_SRP | RF01854 | Bacterial large signal recognition particle RNA |
| Bacteria_small_SRP | RF00169 | Bacterial small signal recognition particle RNA |
| c-di-GMP-I | RF01051 | Cyclic di-GMP-I riboswitch |
| Cobalamin | RF00174 | Cobalamin riboswitch |
| Glycine | RF00504 | Glycine riboswitch |
| Intron_gpII | RF00557 | Group II catalytic intron |
| L10_leader | RF00558 | Ribosomal protein L10 leader |
| L20_leader | RF00168 | Ribosomal protein L20 leader |
| Lysine | RF01749 | Lysine riboswitch |
| Pan | RF00522 | pan motif |
| PreQ1 | RF01054 | preQ1-II (pre queuosine) riboswitch |
| Purine | RF00167 | Purine riboswitch |
| PyrR | RF00515 | PyrR binding site |
| SAM | RF00162 | SAM riboswitch (S box leader) |
| **SSU_rRNA_archaea** | RF01959 | Small subunit ribosomal RNA |
| **SSU_rRNA_bacteria** | RF00177 | Small subunit ribosomal RNA |
| **SSU_rRNA_eukarya** | RF01960 | Small subunit ribosomal RNA |
| **SSU_rRNA_microsporidia** | RF02542 | Small subunit ribosomal RNA |
| T-box | RF00230 | T-box leader |
| TPP | RF00059 | TPP riboswitch (THI element) |
| ydaO-yuaA | RF00379 | ydaO/yuaA leader |
| ykoK | RF00380 | ykoK leader |

*Foi mantida a descrição original do Rfam. Fonte: Banco de dados Grupo_cereus, 2015.

São consideradas significativas as pontuações acima de 20 bits, significando que a busca executada retornou uma correspondência significativa com o modelo de

covariância correspondente (Rfam References 12.0, Jul 2014). Para os candidatos selecionados, todos os resultados das buscas (comparações) no Rfam apresentaram pontuação (score) acima de 53 bits, exceto um dos candidatos (Família: PreQ1, Acession: RF00522, Oid do candidato: 1603396), que apresentou score de 49.8 .

Uma verificação da eficácia dos métodos utilizados foi a existência de candidatos caracterizados como pertencentes às famílias de RNA ribossômico 5S e também de subunidades pequenas de RNA ribossômico, elementos conhecidos e anotados, cuja identificação positiva é esperada em trabalhos como este (Tabelas Tabela 10, Tabela 11 Tabela 12).

A caracterização dos candidatos quanto às famílias Rfam e sua identificação por espécie e cepa são apresentados na Tabela 13.

**Tabela 10: Famílias Rfam para os candidatos identificados em *B. anthracis***

| Família Rfam | Candidatos* |
|---|---|
| 5S_rRNA | 2 |
| Lysine | 6 |
| PyrR | 1 |
| SAM | 1 |
| SSU_rRNA_archaea | 11 |
| SSU_rRNA_bacteria | 11 |
| SSU_rRNA_eukarya | 11 |
| SSU_rRNA_microsporidia | 11 |
| T-box | 12 |

* Identificação positiva para a família nos candidatos. Fonte: banco de dados grupo_cereus 2015.

**Tabela 11: Famílias Rfam para os candidatos identificados em *B. cereus***

| Família Rfam | Candidatos* |
|---|---|
| **5S_rRNA** | 5 |
| c-di-GMP-I | 3 |
| Cobalamin | 1 |
| Intron_gpII | 2 |
| L10_leader | 3 |
| L20_leader | 1 |
| pan | 2 |
| Purine | 1 |
| SAM | 3 |
| **SSU_rRNA_archaea** | 25 |
| **SSU_rRNA_bacteria** | 25 |
| **SSU_rRNA_eukarya** | 25 |
| **SSU_rRNA_microsporidia** | 25 |
| T-box | 20 |
| TPP | 1 |
| ydaO-yuaA | 1 |

* Identificação positiva para a família nos candidatos. Fonte: banco de dados grupo_cereus 2015.

**Tabela 12: Famílias Rfam para os candidatos identificados em *B. thuringiensis***

| Família Rfam | Candidatos* |
|---|---|
| **5S_rRNA** | 1 |
| Bacteria_large_SRP | 1 |
| Bacteria_small_SRP | 1 |
| c-di-GMP-I | 2 |
| Glycine | 1 |
| pan | 1 |
| PreQ1 | 1 |
| Purine | 1 |
| SAM | 4 |
| **SSU_rRNA_archaea** | 44 |
| **SSU_rRNA_bacteria** | 44 |
| **SSU_rRNA_eukarya** | 44 |
| **SSU_rRNA_microsporidia** | 44 |
| T-box | 16 |
| TPP | 6 |
| ykoK | 2 |

* Identificação positiva para a família nos candidatos. Fonte: banco de dados grupo_cereus 2015.

Além das famílias 5S e SSU rRNA, os dados obtidos registraram a identificação das famílias, SAM e T-Box (Figura 18), nas três espécies em questão, *B.anthracis, B.cereus* e *B. thuringiensis.* Ainda ressaltamos uma observação conforme orientação da ferramenta (Rfam References 12.0, Jul 2014): os resultados obtidos também dependem da existência de modelos de comparação atualizados do próprio Rfam. Além disto, a comparação entre as estratégias pode ficar prejudicada caso a estrutura dos elementos de algumas famílias, possivelmente ainda não descritos, apresentem variação entre os organismos estudados (GRUNDY, HENKIN, 1998).

O r*iboswitch SAM* (*S-box leader, SAM-I riboswitch*) é encontrado a montante em genes que codificam para proteínas envolvidas na biossíntese de metionina ou cisteína em bactérias Gram-positivas como o *B. cereus*. Estudos em *Bacillus subtilis* revelaram dois riboswitches atuando no controle do término da transcrição, embora acredite-se que muitos *Riboswitches SAM* possam regular a expressão no nível da tradução.

O elemento T-box está envolvido na regulação de genes de tradução associada, geralmente em bactérias Gram-positivas. Tipicamente, são encontrados a montante dos genes da aminoacil-tRNA sintetase e também de alguns genes biossintéticos de aminoácidos. Descarregado, RNAt atua na anti terminação da transcrição de genes da família T-box.

### 5.2.2. **Famílias Rfam com exclusividade entre *B. cereus* e *B. thuringiensis***

Também foram identificadas quatro famílias em comum somente para *B.cereus* e *B. thuringiensis,* os riboswitches TPP, Purine, e c-di-GMP-I, e o pan RNA (RNA motif). Dentre estas pan RNA e c-di-GMP-I (GEMM motif) foram elementos preditos através da aplicação de técnicas de bioinformática, sendo pan RNA relacionado à regulação da síntese de pantotenato (WEINBERG et al., 2010). O riboswitch TPP é relacionado a mecanismos de regulação de expressão gênica, sendo descrito como a forma ativa da tiamina (vitamina B1). O riboswitch c-di-GMP-I (di-GMP-I cíclico) é descrito como um segundo mensageiro para vários processos microbianos envolvendo virulência, motilidade e a formação de biofilme (SUDARSAN et al., 2008). O riboswitch Purine é relacionado ao à regulação da quantidade de purina, participando do metabolismo desta, bem como em sua absorção pela membrana (MANDAL et al., 2003).

Com base nas considerações anteriores e tendo em vista os critérios de tratamento aplicados ao conjunto de dados produzido e frente a problemática taxonômica envolvendo as espécies estudadas, enfatizaremos as famílias com identificação exclusiva para os organismos estudados:

- *Bacillus anthracis*: *Lysine, PyrR*;
- *Bacillus cereus*: *Cobalamin, Intron_gpII, L10 e L20_leader, ydaO-yuaA*;
- *Bacillus thuringiensis*: *Bacteria - Large and Small SRP, Glycine, PreQ1, ykoK*;

O grupo Rfam, enquanto instituição, coordena o registro de anotações das famílias disponibilizadas em seu website mediante um grupo criado na Wikipedia (Rfam References 12.0, Jul 2014). Desta forma as descrições e representações estruturais das famílias (lembrando, exclusivas para cada espécie) encontradas foram obtidas diretamente no Website Rfam, e são apresentadas a seguir.

### 5.2.3. **Famílias Rfam Com Exclusividade em *B. anthracis, B. cereus* e *B. thuringiensis*: Descrição e Estrutura Secundária**

Família: Lysine (RF00168)

Descrição: *Lysine riboswitch*[2]

*O Riboswitch de lisina* é um elemento metabólico de ligação de RNA encontrado dentro de certos RNAs mensageiros que servem como um sensor de precisão para o aminoácido lisina. O rearranjo alostérico (alteração na estrutura terciária ou quaternária) da estrutura do mRNA é mediada por sua ligação, resultando na modulação da expressão do gene (MANDAL et al., 2003).

Este *riboswitch* é encontrado em um certo número de genes envolvidos no metabolismo de lisina, incluindo LysC (SUDARSAN et al., 2003, RODIONOV et al., 2003). Riboswitches de lisina também foram identificados de forma independente tendo

[2] *Riboswitches,* (em livre tradução: interruptor ribossômico). Um riboswitch é uma ribozima que cliva a si mesma na presença de concentrações suficientes de seu metabolito. São domínios de ligação de metabólitos dentro de certos mRNAs que servem como sensores de precisão para seus alvos específicos, ou dito de outra forma, receptores para metabólitos específicos. Apresentam-se como domínios complexos enovelados de RNA.

sido denominados L-box (GRUNDY, LEHMAN, HENKIN, 2003). Sua estrutura secundária é exibida na Figura 7.

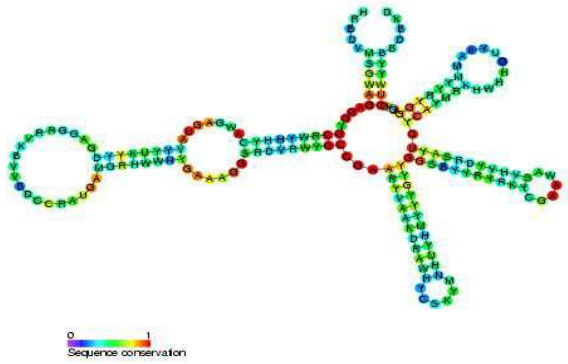


**Figura 7: Estrutura secundária predita e sítios conservados no *Riboswitch de lisina*. Fonte: Wikimedia Commons .**

Família: PyrR (RF00515)

Descrição: *PyrR binding site*

O *Sítio de Ligação PyrR* é um elemento de RNA que se encontra a montante (*upstream*) de uma variedade de genes envolvidos no transporte e na biossíntese de pirimidina. A estrutura do RNA permite a ligação da proteína PyrR que regula a biossíntese de pirimidina em *Bacillus subtilis*. Quando tal proteína é ligada, é formado um grampo terminador (*hairpin*) a jusante (*downstream*), reprimindo a transcrição de genes da biossíntese (BONNER et al., 2001) (Figura 8).

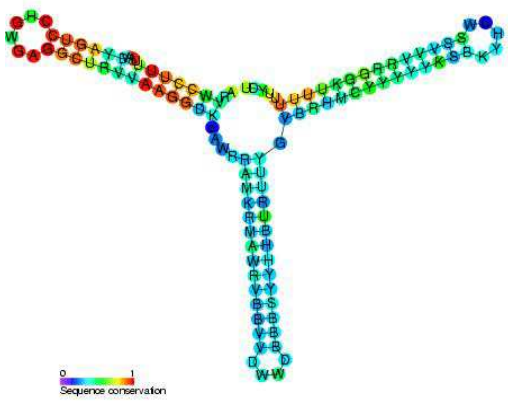


**Figura 8: Estrutura secundária predita e sítios conservados para o *Sítio de Ligação PyrR*. Fonte: Wikimedia Commons**

Família: Cobalamin (RF00174)

Descrição: *Cobalamin riboswitch*

O *Riboswitch de Cobalamina* é um elemento regulatório que atua em CIS (atuação local), amplamente distribuído em regiões 5' não traduzidas de vitamina B12 (Cobalamina) relacionados a genes em bactérias (NAHVI, 2002) . O rearranjo alostérico da estrutura do mRNA é mediado pela ligação do ligando (ligante), e isto resulta na modulação da expressão de genes ou a tradução de mRNA para a obtenção de uma proteína. A cobalamina na forma de adenosilcobalamina (Ado-CBL) é conhecida por reprimir a expressão de proteínas para a biossíntese de vitamina B12 por meio de um mecanismo de regulação pós-transcricional que envolve a ligação direta de Ado-CBL a (terminais) 5' UTRs em genes relevantes, evitando a ligação do ribossomo e a tradução desses genes (Figura 9).

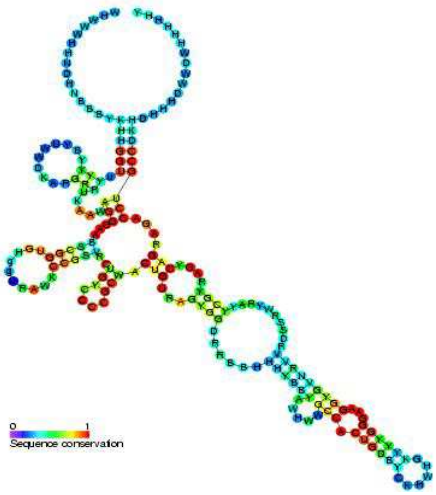


**Figura 9: Estrutura secundária predita e sítios conservados no *Riboswitch de Cobalamina*. Fonte: Wikimedia Commons.**

Família: Intron_gpII (RF00029)

Descrição: *Group II catalytic intron*

*Íntrons catalíticos do grupo II* são uma grande classe de ribozimas auto-catalíticas, também classificadas como elementos genéticos móveis encontrados dentro de genes nos três grandes domínios biológicos (*Bacteria, Plantae, Fungi*). A atividade de ribozima (por exemplo, auto-splicing) pode ocorrer sob condições de alta salinidade in vitro. No entanto, o auxílio de proteínas é necessário para o splicing in vivo. Em contraste com os introns do grupo I, neste grupo a excisão ocorre na ausência de GTP e envolve a formação de um laço, com uma ramificação um resíduo que se assemelha

fortemente que com a ramificação encontrada nos laços formados durante o splicing de pre-mRNA nuclear. Supõe-se que o splicing de pré-mRNA pode ter evoluído a partir de íntrons grupo II, devido ao mecanismo catalítico semelhante bem como devido à semelhança estrutural da subestrutura do Domínio V com o prolongamento U6/U2 snRNA. Em termos evolutivos, acredita-se que estes íntrons são os ancestrais do próprio spliceossomo (ZAHA et al., 2014). Mais de 25% dos genomas de bactérias sequenciados apresentam íntrons do grupo II, geralmente pouco numerosos e apresentando-se como retroelementos ativos. Finalmente, a sua capacidade de mobilização sítio-específica para novos locais do DNA tem sido explorada como uma ferramenta para a biotecnologia.

Íntrons catalíticos grupo II são encontrados em rRNA, tRNA e mRNA de organelas (cloroplastos e mitocôndrias), em plantas, fungos e protistas, e também em mRNA em bactérias. Eles são grandes ribozimas com auto-splicing e têm seis domínios estruturais (usualmente designados dI a dVI). O modelo apresentado (Figura 10) bem como seu alinhamento, representa apenas os domínios V e VI.

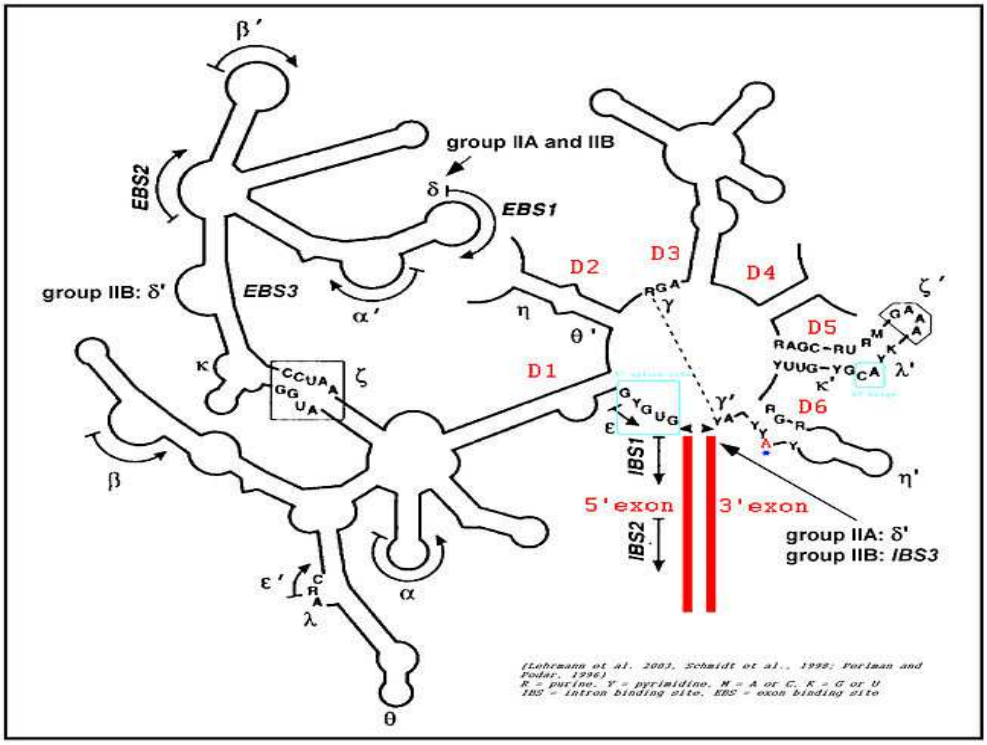


**Figura 10: Estrutura de *Íntron Grupo II*. Fonte: Wikimedia Commons.**

Família: L10_leader (RF00557)

Descrição: *Ribosomal protein L10 leader*

Esta família representa uma estrutura autoregulatória putativa para uma proteína ribossômica líder encontrada em *B. subtilis* e outras bactérias Gram-positivas de baixo percentual de GC. Tal estrutura está localizada em regiões não traduzidas 5' dos mRNAs que codificam proteínas ribossomais L10 e L12 (rplJ-rplL).

Uma estrutura de terminador de transcrição independente de Rho, provavelmente envolvida na regulação está incluída na extremidade 3' (ZENGEL & LINDAHL, 1994) (Figura 11).

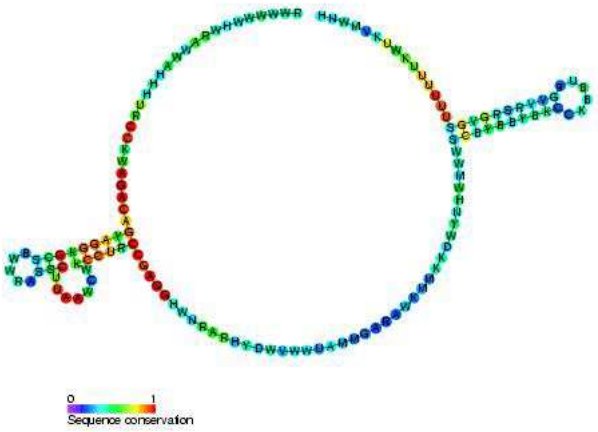


**Figura 11: Estrutura secundária predita e sítios conservados na *Proteína Ribossômica Líder L10*. Fonte: Wikimedia Commons.**

Família: L20_leader (RF00558)

Descrição: *Ribosomal protein L20 leader*

Esta família representa uma estrutura auto regulatória putativa para uma proteína ribossômica líder encontrada em *B. subtilis* e outras bactérias Gram-positivas de baixo percentual de GC. Tal estrutura está localizada em regiões não traduzidas 5' dos mRNAs que codificam fatores de iniciação 3 seguidos por proteínas ribossomais L35 e L20 (infC-rpmI-rplT). Uma estrutura de terminador de transcrição independente de Rho, provavelmente envolvida na regulação, sendo incluída na extremidade 3' (ZENGUEL & LINDAHL, 1994) (Figura 12).

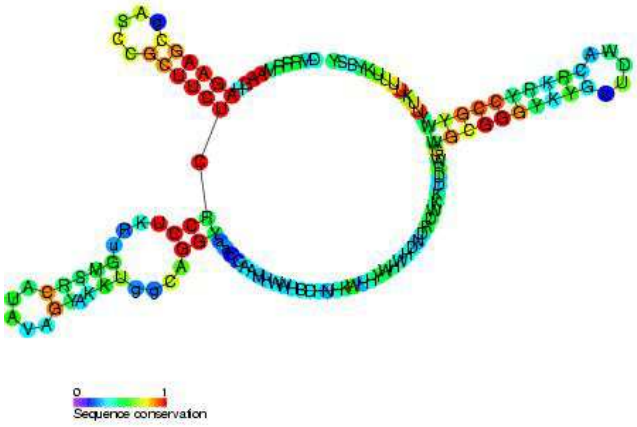


**Figura 12: Estrutura secundária predita e sítios conservados na *Proteína Ribossômica Líder L20*. Fonte: Wikimedia Commons.**

Família: ydaO-yuaA (RF00379)[3]

Descrição: *ydaO/yuaA leader*

Tal elemento representa uma estrutura conservada de RNA encontrada a montante dos genes *ydaO e yuaA*. Acredita-se que tal elemento possa ser acionado em situações de choque osmótico, levando à ativação de ydaO, um gene predito, transportador de aminoácidos, e também ativando membros do operon yuaA-yubG, o qual codifica para os transportadores KtrA e KtrB K+ (BARRICK et al., 2004). Trabalhos posteriores verificaram que o elemento ydaO tem distribuição generalizada e está associado a um conjunto diversificado de genes os quais ele controla de maneira dependente da sequência e conformação estrutural (BLOCK, HAMMOND, BREAKER, 2010) (Figura 13).

[3] Esta família não possui uma estrutura oficial registrada no banco de dados Rfam, contudo as referências encaminham a trabalhos independentes e são aqui exibidas.

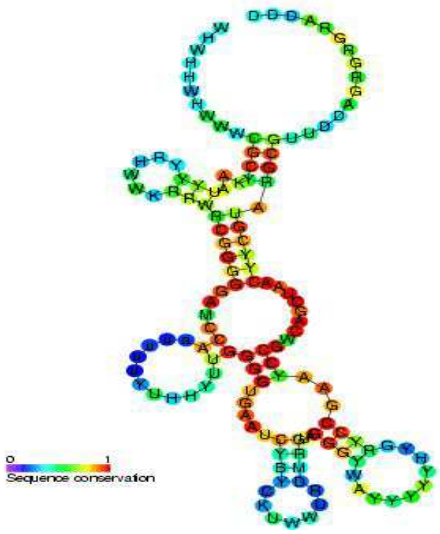


**Figura 13: Estrutura secundária predita e sítios conservados do elemento *ydaO/yuaA leader*. Obs: estrutura não oficial. Fonte: Wikimedia Commons**

Família: Bacteria_small_SRP[4] (RF00169) e Bacteria_large_SRP (RF01854)

Descrição: *Bacterial small signal recognition particle RNA* e *Bacterial large signal recognition particle RNA*

O domínio de reconhecimento de sinal do RNA, também conhecido como 7SL, 6S, ffs, ou RNA 4.5S, é o componente RNA do complexo ribonucleoproteico da partícula de reconhecimento de sinal (SRP). O SRP é uma ribonucleoproteína universalmente conservada que direciona o tráfego de proteínas no interior da célula e permitindo que sejam secretadas. Os SRP RNAs, em conjunto com uma ou mais proteínas SRP contribuem para a ligação e libertação do peptídeo de sinal. O RNA e os componentes proteicos deste complexo são altamente conservados, mas variam entre os diferentes reinos. Na maioria das bactérias, o SRP é composto por uma molécula de RNA (4.5S) e a proteína Ffh (um homólogo da proteína SRP54 eucariótica). Algumas bactérias Gram-positivas (por exemplo, *B. subtilis*) tem um SRP RNA similar ao eucariota, que inclui um domínio Alu (REGALIA, ROSENBLAD, SAMUELSSON, 2002).

[4] SRP: (Ingl. *signal recognition particle RNA*) Partícula ou domínio RNA de reconhecimento de sinal.

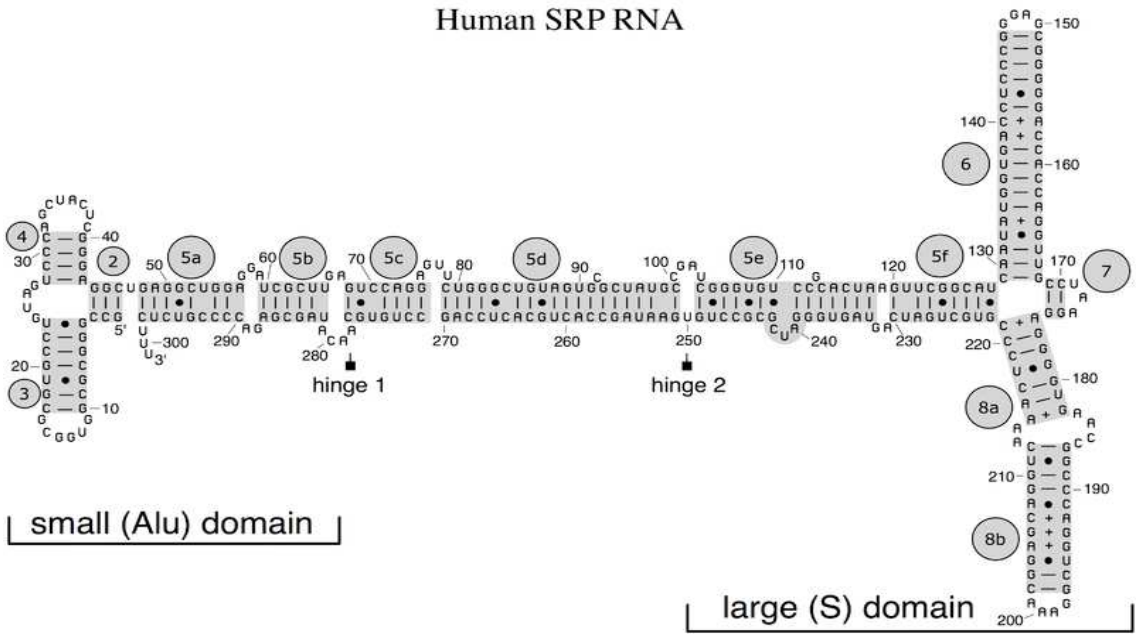


**Figura 14: Estrutura secundária do RNA SRP humano. As Hélices estão numeradas de 2 a 8. As seções helicoidais (em cinza) são nomeadas com letras minúsculas. Os resíduos são numerados em incrementos de dez. Estão indicadas as extremidades 5' e 3'. Em destaque: as duas "dobradiças" e os domínios RNA SRP, o domínio menor (Alu) e o grande (S, específico) .**

Família: Glycine (RF00504)

Descrição: *Glycine riboswitch*

O *Riboswitch Bacteriano de Glicina* é um elemento do RNA que pode ligar o aminoácido glicina. Riboswitches de glicina geralmente consistem de dois domínios de aptâmeros de ligação do metabolito com estruturas semelhantes em tandem. Os aptâmeros ligam a glicina cooperativamente, para regular a expressão de genes a jusante. Em *Bacillus subtilis*, este riboswitch é encontrado a montante do operon gcvT, o qual controla a degradação da glicina. Acredita-se que, quando há glicina em excesso, ele irá ligar-se a ambos os aptâmeros para ativar tais genes e facilitar a degradação da glicina (MANDAL et al., 2004) . Uma versão truncada do Riboswitch de Glicina apresenta curvas sigmoidais de ligação com coeficientes de Hill superiores a um, o que levou à idéia de cooperatividade positiva entre os dois domínios de aptâmeros (MANDAL et al., 2004, KWON & STROBEL, 2008). Um estudo de 2012 mostra que a ligação cooperativa não ocorre na troca com sua extremidade líder 5', mas o propósito do interruptor duplo de aptâmero ainda é incerto (SHERMAN et al., 2012) (Figura 15).

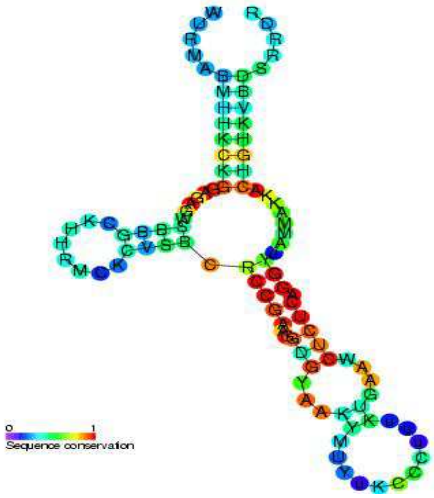


**Figura 15: Estrutura secundária predita e sítios conservados no *Riboswitch de Glicina*. Fonte: Wikimedia Commons.**

Família: PreQ1 (RF00522)

Descrição: *PreQ1 riboswitch*

O Riboswitch PreQ1-I é um elemento identificado em bactérias, que atua em CIS. Tal elemento regula a expressão de genes envolvidos na biossíntese do nucleosídeo[5] queuosina[6] a partir de GTP (ROTH et al., 2007). Este elemento de RNA, conhecido como um riboswitch, liga a preQ1 (pré-queuosina1), intermediária na via de queuosina. Sua função como riboswitch foi caracterizada em *Bacillus subtilis*, em que o riboswitch está localizado no líder do operon ykvJKLM (queCDEF), o qual codifica quatro genes necessários para a produção de queuosina (READER et al., 2004). Considera-se que neste organismo, a ligação de PreQ1 com o aptâmero do riboswitch, induz a terminação prematura da transcrição na fita líder, exercendo um controle negativo na expressão destes genes. O Riboswitch preQ1 se distingue por seu aptâmero, incomumente pequeno, se comparado com outros riboswitches (KLEIN; EDWARDS; FERRÉ-D´AMARÉ, 2009) (Figura 16).

[5] Um nucleosídeo é um nucleotídeo sem o grupamento fosfato, apenas o açúcar e sua base nitrogenada;
[6] Queuosine é um nucleosídeo modificado presente em certos tRNAs em *Bacteria* e *Eucaria.*

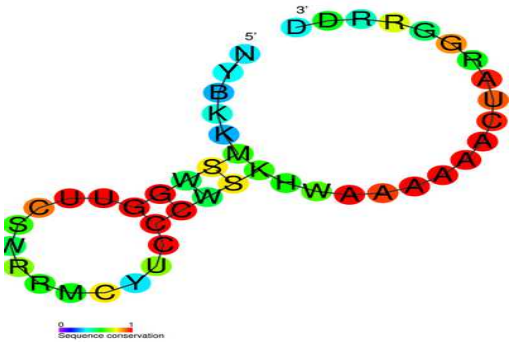

**Figura 16: Estrutura secundária predita e sítios conservados no *Riboswitch PreQ1*. Fonte: Wikimedia Commons.**

Família: ykoK (RF00380)

Descrição: *ykoK leader*

O *líder ykoK* ou *M-box* é uma estrutura do RNA para detecção de Mg2+-, que controla a expressão de proteínas transportadoras de íons de magnésio em bactérias. Ele é uma estrutura distinta de resposta do RNA ao elemento Magnésio. O *líder ykoK* foi originalmente descrito como uma sequência conservada com uma função riboswitch em potencial, encontrada a montante do gene de *B. subtilis ykoK*, e também de genes relacionados com as funções de outras bactérias (BARRICK et al., 2004). Exemplos da estrutura conservada de RNA M-box ocorrem a montante de cada uma das três grandes famílias de transportadores de Mg2+ (CorA, MgtE e MgtA/MgtB) em várias espécies de bactérias (DANN et al., 2007). A estrutura molecular do (exemplo) M-box a montante do gene de *B. subtilis ykoK* inclui seis ligações de íons Mg2+. Estudos bioquímicos indicaram que este RNA M-Box se compacta na presença de Mg2+ e outros íons bivalentes. Este processo de dobramento parece interromper uma estrutura antiterminador, e assim permitir a formação de uma estrutura de terminação de transcrição. Como esperado a partir deste modelo, as células de *B. subtilis* reprimem a expressão de um *gene repórter*[7] a jusante, quando cultivadas na presença de Mg2+. Portanto, o M-box parece funcionar como um "interruptor genético de desligamento", o que é importante para a manutenção da homeostase de Mg2+ em bactérias (Figura 17).

[7] Gene frequentemente de origem procariótica, que produz um produto facilmente detectado em células eucarióticas, sendo utilizado como marcador para determinar a atividade de outro gene, com o qual seu DNA foi intimamente ligado ou combinado.

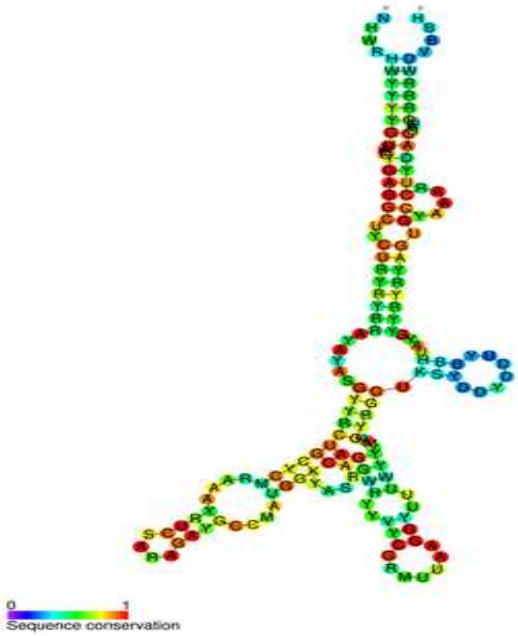


**Figura 17: Estrutura secundária predita e sítios conservados em *ykoK*. Fonte: Wikimedia Commons.**

## 6 CONCLUSÃO

Em uma análise geral sobre as características das famílias obtidas para *B. anthracis, B. cereus e B. thuringiensis*, um resultado esperado é evidenciado para os candidatos identificados. Tal resultado deve-se às características dos softwares utilizados, como a busca de sinais transcricionais específicos e sequências terminadoras (SRIDHAR, 2009) como utilizados pela ferramenta sRNAscanner, que em conjunto, identificou várias famílias envolvidas em processos de regulação gênica.

Ainda conforme afirmado por EDDY, 2001, os ncRNA candidatos foram classificados também em famílias relacionadas com funções que requerem o reconhecimento específico de ácido nucleico, assim como a regulação pós-transcricional da expressão gênica ou modificações na orientação do RNA.

Um diagrama é proposto na Figura 18 Anexo A, reunindo algumas das principais características encontradas e relacionando-as por espécie para uma exibição mais concisa, de forma similar à construção de um mapa mental. As famílias Rfam atribuídas aos 181 ncRNA candidatos obtidos, são assim distribuídas por espécie segundo uma divisão por cores (famílias exclusivas). Em vermelho temos os dados referentes a *B. anthracis*, sendo o verde e o azul respectivamente para *B. cereus* e *B. thuringiensis*. As famílias em comum para os três grupos podem ser visualizadas no círculo maior, ao centro da figura, enquanto que a família TPP (canto inferior direito) é compartilhada somente por dois grupos, *B. cereus* e *B. thuringiensis*. Ainda seguindo o mesmo padrão de cores, os números maiores, ao lado da representação da estrutura secundária para cada família indicam o número de candidatos nesta identificados.

Os totais para as famílias compartilhadas são exibidos na representação das "Placas de Petri" para cada espécie, e equivalem aos totais obtidos nas Tabela 10, Tabela 11 e Tabela 12 sendo a classificação de candidatos encontrados por família e cepa é demonstrada na Tabela 13.

Os resultados desta etapa de trabalho *in silico* permitem algumas sugestões para o trabalho em bancada. Com a identificação posicional de ncRNA candidatos por cepa e sua separação em famílias, seriam indicadas pesquisas envolvendo os mecanismos de ação dos riboswitches identificados em cada organismo (Figura 18).

Para cada grupo identificado, sugere-se uma linha de experimentos envolvendo questões como: degradação de glicina (*B. thuringiensis* MC28), modulação de lisina (*B.*

*anthracis*, cepas Ames Ancestor, Ames, CDC 684, H9401, Sterne), tiamina (*B. cereus B4264*; *B. thuringiensis BMB171* cepas *kurstaki str. HD73, Bt407* e *IS5056*) bem como a biossíntese de pirimidina (*B. anthracis* CDC 684), cobalamina (*B. cereus* 03BB102) e queosina (*B. thuringiensis* MC28). Também experimentos sob condições de alta salinidade *in vitro*, relacionados à inibição ou inativação gênica envolvendo ribozimas (*B. cereus* AH187).

**REFERÊNCIAS**


**1** ARANTES, O. M. N.; VILAS-BÔAS, L. A.; VILAS-BÔAS, G. F. L. T. *Bacillus thuringiensis*: estratégias no controle biológico. In: SERAFINE, L.A; BARROS, N. M; AZEVEDO, J. L. (Org.). Biotecnologia: avanços na agricultura e na agroindústria. Caxias do Sul: Agropecuária. p. 269-293, 2002.

**2** AUGER. S.; GALLERON. N.; SÉGURENS. B.; DOSSAT. C.; BOLOTIN. A.; WINCKER. P.; SOROKIN. A. Complete genome sequence of the highly hemolytic strain *Bacillus cereus* F837/76. **Journal of Bacteriology,** Washington DC, v. 194, n. 6, p. 1-5, mar. 2012. PMCID: PMC3294841.

**3** BONNER, E.R.; D'ELIA, J.N.; BILLIPS, B.K.; SWITZER, R.L. Molecular recognition of *pyr mRNA* by the *Bacillus subtilis* attenuation regulatory protein *PyrR*. **Nucleic Acids Research,** Oxford, v. 29, n. 23, p. 4851–4865, set. 2001. PMID11726695.

**4** BURGE. S. W.; DAUB. J.; EBERHARDT. J.; TATE. L.; NAWROCKI. E. P.; EDDY. S. R.; GARDNER. P. P.; BATEMAN. A. Rfam 11.0: 10 years of RNA families. **Nucleic Acids Research,** London, v. 41, n/d, p. 226-232, jan. 2013. PMC3531072

**5** CARLSON, T. N.; GILLIES, R. R; PERRY, E. M. A method to make use of thermal infrared temperature and NDVI measurements to infer soil water content and fractional vegetation cover. **Remote Sensing Reviews,** Pennsylvania, v. 52, n. 1-2, p. 161-173, out. 1994. DOI101080

**6** CATANHO, M.; DEGRAVE, W.; MIRANDA, A.B. Análise Comparativa de Genomas Procarióticos. **Biotecnologia Ciência & Desenvolvimento**, Brasília, Ano X, n. 37, p. 20-29, 2008. BIO377b

**7** CHALLACOMBE. J.F.; ALTHERR. M.R.; XIE. G.; BHOTIKA. S.S.; BROWN. N., et al. The Complete Genome Sequence of *Bacillus thuringiensis* Al Hakam.

**Journal of Bacteriology,** Washington DC, v. 189, n. 9, p. 3680-3681, mar. 2007. PMID17337577. PMC1855882.

**8** CHUN. J.; HONG. K.; CHA. S.H.; CHO. M.; LEE. K.J; JEONG. D.H.; YOO. C.K.; RHIE. G.E. Complete Genome Sequence of *Bacillus anthracis* H9401, an Isolate from a Korean Patient with Anthrax. **Journal of Bacteriology,** Washington DC, v. 194, n. 15, p. 4116–4117, ago. 2012. doi: 10.1128/JB.00159-12, 2012. PMC3416559

**9** CLUSTAL. Website - Clustal, Multiple Sequence Alignment. Science Foundation Ireland (SFI), 2008 - 2012. Clustal Help. Disponível em: <http://www.clustal.org/download/clustalx_help.html>. Acesso em 15, jul 2013.

**10** CRAMER, F.; DOEPNER, H.; HAAR, F. VD.; SCHLIMME, E.; SEIDEL, H. On the conformation of transfer RNA. **Proceedings of the National Academy of Sciences,** Washington DC, v. 61, n. 4, p. 1384–1391, dez. 1968. PMC225267.

**11** CRICKMORE, N.; ZEIGLER, D.R.; FEITELSON, J.; SCHNEPF, E.; VAN RIE, J.; LERECUS, D.; BAUM, J. and DEAN, D.H. Revision of the nomenclature for the *Bacillus thuringiensis* pesticidal crystal proteins. **Microbiology and Molecular Biology Reviews,** Washington DC, v. 62, n. 3, p. 807-813, set. 1998. PMC98935 .

**12** DAFFONCHIO. D.; CHERIF. A.; BORIN. S. Homoduplex and heteroduplex polymorphisms of the amplified ribosomal 16S–23S internal transcribed spacers describe genetic relationships in the ‘‘*Bacillus cereus* group’’. **Applied and Environmental Microbiology,** Washington DC, v. 66, n. 12, p. 5460–5468, dez. 2000. PMID11097928

**13** DAVISON. J. Genetic Exchange between Bacteria in the Environment. Institut National de la Recherche Agronomique, **Plasmid,** New York, v. 42, n. 2, p. 73-91. Disponível online em 26, mar. 2002. DOI101006

**14** DIXON. T. C.; MESELSON. M.; GUILLEMIN. J; HANNA. P. C.; Anthrax. **The New England Journal of Medicine,** Waltham, v. 9, n. 341, p. 815-826, set. 1999. PMID10477781.

**15** DOENÇAS TRANSMITIDAS POR ALIMENTOS. Secretaria Municipal de Saúde - RJ. ***Doenças transmitidas por alimentos: Bacillus cereus*** (Boletim Técnico, n º 06, Junho, 2000). Disponível em: < http://www6.ensp.fiocruz.br/visa/files/bol6.pdf >. Acesso em: 18 08 2014.

**16** DROBNIEWSKI. F.A.; *Bacillus cereus* and related species. **Clinical Microbiology Reviews,** Washington DC, v. 6, n. 4, p. 324–338, out. 1993. PMC358292.

**17** DUDOCK, B.S.; KATZ, G.; TAYLOR, E.K.; HOLLEY, R.W.; Primary structure of wheat germ phenylalanine transfer RNA. **Proceedings of the National Academy of Sciences,** Washington DC**,** v. 62, n. 3, p. 941–945, mar. 1969. DOI11073.

**18** EDDY. S. R. Computational Genomics of Noncoding RNA Genes. **Cell Press,** Cambridge, v. 109, s/n, p. 137–140, abr. 2002. CS3742004.

**19** EDDY. S. R. Non–coding RNA genes and the modern RNA world. **Nature Reviews Genetics,** Southampton, v. 2, s/n, p. 919-929, dez. 2001. DOI101038.

**20** FAGERLUND. A.; BRILLARD. J.; FÜRST1. R.; GUINEBRETIÈRE. M.; GRANUM. P. E.; Toxin production in a rare and genetically remote cluster of strains of the *Bacillus cereus* group. **BMC Microbiology,** London, v. 7, n. 43, p. 1471-2180, maio. 2007. PMC1888693.

**21** FAZION, F. A. P. EFEITO DO GENE *cry* SOBRE O COMPORTAMENTO DE LINHAGENS DE *Bacillus cereus* E *Bacillus thuringiensis* EM LARVAS DE *Anticarsia gemmatalis*. 2012. 71f. Dissertação (Mestrado em Genética e Biologia Molecular) – Centro de Ciências Biológicas, Universidade Estadual de Londrina, Londrina.

**22** FEITELSON, J. S.; PAYNE, J.; KIM, L. *Bacillus thuringiensis*: insects and beyond. **Nature Biotechnology,** New York, v. 10, s/n, p. 271-275, n.a. 1992. NPG.

**23** FOUET. A.; MESNAGE. S.; *Bacillus anthracis* cell envelope components. **Current Topics in Microbiology Immunology,** New York, v. 271, s/n, p. 87–113, mes. 2002. PMID12224525.

**24** GRANUM, P. E.; BRYNESTAD, S.; KRAMER, J. M. Analysis of enterotoxin production by *Bacillus cereus* from dairy products, food poisoning incidents and non-gastrointestinal infections. **International Journal of Food Microbiology,** Amsterdam, v. 17, n. 4, p. 269–279, fev. 1993. PMID8466800.

**25** GRIFFITHS-JONES. S.; BATEMAN. A.; MARSHALL. M.; KHANNA. A.; EDDY. S. R. Rfam: an RNA family database. **Nucleic Acids Research,** London, v. 31, n. 1, p. 439-441, jan. 2003. PMC165453.

**26** GRIFFITHS-JONES. S.; MOXON. S.; MARSHALL. M.; KHANNA. A.; EDDY. S. R.; BATEMAN. A. Rfam: annotating non-coding RNAs in complete genomes. **Nucleic Acids Research,** London, v. 33, s/n, p. 121-124, jan. 2005. PMC540035.

**27** GRUNDY, F. J.; HENKIN, T. M. The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria. **Molecular Microbiology,** Boston, v. 30, n. 4, p. 1365-2958, nov. 1998. DOI101046. PMID10094622.

**28** GRUNDY, F.J.; LEHMAN, S.C.; HENKIN, T.M. The L box regulon: Lysine sensing by leader RNAs of bacterial lysine biosynthesis genes. **Proceedings of the National Academy of Sciences,** Washington DC, v. 100, n. 21, p. 12057–12062, out. 2003. PMID14523230.

**29** GUAN. P.; AI. P.; DAI. X.; ZHANG. J.; XU. L.; ZHU. J.; LI. Q.; DENG. Q.; LI. S.; WANG. S.; LIU. H.; WANG. L.; LI. P.; ZHENG. A. Complete genome sequence of *Bacillus thuringiensis* serovar Sichuansis strain MC28. **Journal of Bacteriology,** Washington DC, v. 194, n. 24, p. 6975, dez. 2012. PMID23209229. PMC3510617 .

**30** HAN.C. S.; XIE. G.; CHALLACOMBE. J. F.; ALTHERR. M. R.; BHOTIKA. S. S, et al. Pathogenomic Sequence Analysis of *Bacillus cereus* and *Bacillus thuringiensis* Isolates Closely Related to *Bacillus anthracis*. **Journal of Bacteriology,** Washington DC, v. 188, n. 9, p. 3382-3390, maio. 2006. PMC1447445 .

**31** HAHN. M. W.; WRAY. G. A. The g-value paradox. **Evolution & Development,** Carolina do Norte, v. 4, n. 2, p. 73-75, mes. 2002. PDEMWH3.

**32** HE. J.; WANG. J.; YIN. W.; SHAO. X.; ZHENG. H.; LI. M.; ZHAO. Y.; SUN. M.; WANG. S.; YU. Z. Complete genome sequence of *Bacillus thuringiensis* subsp. chinensis strain CT-43. **Journal of Bacteriology,** Washington DC, v. 193, n. 13, p. 3407-3408, jul. Ano. 2011. PMID1551307. PMC3133296.

**33** HE. J.; SHAO. X.; ZHENG. H.; LI. M.; WANG. J.; ZHANG. Q.; LI. L.; LIU. Z.; SUN. M.; WANG. S.; YU. Z. Complete genome sequence *of Bacillus thuringiensis* mutant strain BMB171. **Journal of Bacteriology,** Washington DC, v. 192, n.15, p. 4074-4075, ago. 2010. PMID20525827. PMC2916366.

**34** HELGASON, E., D. A.; CAUGANT, M. M.; LECADET, Y.; CHEN, J.; MAHILLON, A.; LO¨VGREN, I.; HEGNA, K.; KVALOY, A. B.; KOLSTØ. Genetic diversity of *Bacillus cereus / Bacillus thuringiensis* isolates from natural sources. **Current Microbiology,** New York, v. 37, n. 2, p. 80-87, ago. 1998. PMID9662607 .

**35** HELGASON. E.; ØKSTAD. O. A.; CAUGANT. D. A.; JOHANSEN. H. A.; FOUET. A. ET AL. *Bacillus anthracis, Bacillus cereus*, and *Bacillus thuringiensis* - one species on the basis of genetic evidence. **Applied and Environmental Microbiology,** Washington DC, v. 66, n. 6, p. 2627-2630, jun. 2000. PMID10831447 .

**36** HEYNDRICKX, M.; SCHELDEMAN, P. Bacilli associated with spoilage in dairy products and ther food. In Applications and Systematics of Bacillus and Relatives,

Edited by Berkeley, R. C. W.; Heyndrickx, M.; Logan, N. A.; De Vos, P. **Blackwell Science,** Oxford, v. 6, n/d, p. 64-82, abr. 2008. DOI101002

**37** HOFFMASTER. A.R.; HILL. K.K.; GEE. J.; MARSTON. C.K.; DE. B.K. Characterization of *Bacillus cereus* Isolates Associated with Fatal Pneumonias: Strains Are Closely Related to *Bacillus anthracis* and Harbor *B. anthracis* Virulence Genes. **Journal of Clinical Microbiology,** Washington, v. 44, n. 9, p. 3352-3360, set. 2006. DOI101128.

**38** HOLLEY, R.W.; APGAR, J.; EVERETT, G.A. et al. STRUCTURE OF A RIBONUCLEIC ACID**. Science,** New York, v. 147, n. 3664, p. 1462-1465, mar. 1965. PMID14263761.

**39** JERNIGAN, D. B.; RAGHUNATHAN, P. L.; BELL, B. P.; BRECHNER, R.; BRESNITZ, E.A.; BUTLER, J. C. Investigation of bioterrorism-related anthrax, United States, 2001: epidemiologic findings. **Emerging Infectious Diseases,** Atlanta, v. 8, n. 10, p. 1019-1028, out. 2002. PMID12396909.

**40** JEFFARES. D. C; POOLE. A. M.; PENNY. D. Relics from the RNA world. **Journal of Molecular Evolution,** New York, v. 46, n. 1, p. 18–36, jan. 1998. PMID 9419222.

**41** KIM, S.H.; QUIGLEY, G.J.; SUDDATH, F.L. et al. Three-dimensional structure of yeast phenylalanine transfer RNA: folding of the polynucleotide chain. **Science,** New York, v. 179, n. 4010, p. 285-288, jan. 1973. PMID4566654.

**42** LADNER, J.E.; JACK, A.; ROBERTUS, J.D. et al. Structure of yeast phenylalanine transfer RNA at 2.5 A resolution. **Proceedings of the National Academy of Sciences,** Washington DC, v. 72, n. 11, p. 4414–4418, nov. 1975. DOI101073 PMC388732. PMID1105583.

**43** LALLOO, R.; MAHARAJH, D.; GÖRGENS, J.; GARDINER, N. A downstream process for production of a viable and stable Bacillus cereus aquaculture biological

agent. **Applied Microbiology and Biotechnology,** New York, v. 86, n. 2, p. 499-508, mar. 2010. PMID19921182.

**44** MADISON, J.T.; EVERETT, G.A.; KUNG, H. Nucleotide sequence of a yeast tyrosine transfer RNA. **Science,** New York, v. 153, n. 3735, p. 531–534, jul. 1966. PMID5938777.

**45** MILLER. K. National Institute of Mental Health (NIH) Working Definition of Bioinformatics and Computational Biology, Biomedical Information Science and Technology Definition Committee (BISTIC). **Biomedical Computation Review 2000**. Disponível em: < http://www.bisti.nih.gov/docs/CompuBioDef.pdf>. Acesso em: 15 fev. 2015.

**46** KLEIN, D. J.; EDWARDS, T. E.; FERRÉ-D'AMARÉ, A.R. Cocrystal structure of a *class I preQ1 riboswitch* reveals a pseudoknot recognizing an essential hypermodified nucleobase. **Nature Structural & Molecular Biology,** New York, v. 16, n. 3, p. 343–344, mar. 2009. PMC2657927.

**47** KEIM. P.; PRICE. L. B.; KLEVYTSKA. A. M.; SMITH. K. L.; SCHUPP. J. M. et al. Multiple locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. **Journal of Bacteriology,** Washington DC, v. 182, n. 10, p. 2928–2936, maio. 2000. PMC102004.

**48** KLEE. R.S.; BRZUSZKIEWICZ. E.B.; NATTERMANN. H.; BRÜGGEMANN. H.; DUPKE. S., et al. The Genome of a Bacillus Isolate Causing Anthrax in Chimpanzees Combines Chromosomal Properties of *B. cereus* with *B. anthracis* Virulence Plasmids. **PLoS ONE - Public Library Of Science,** San Francisco, v. 5, n. 7, p. e10986, jul. 2010. PMC2901330.

**49** KOTIRANTA, A. Epidemiology and pathogenesis of *Bacillus cereus* infections. **Microbes and Infection (Pasteur Institute)**, Paris, v. 2, n. 2, p. 189-198, fev. 2000. PMID10742691.

**50** KWON, M.; STROBEL, S.A. Chemical basis of glycine riboswitch cooperativity. **RNA,** Cold Spring Harbor NY, v. 16, n. 11, p. 2291, nov. 2010. 1ª Publicação: DOI101261. PMC2151043. Artigo corrigido em: PMC2957066.

**51** LEE, J.; ROY. P. Complete sequence of the NS1 gene (M6 RNA) of US bluetongue virus serotype 10. **Nucleic Acids Research,** London, v. 15, n. 17, p. 7207, jul. 1987. PRJNA58753. PMC306231.

**52** LEWIN, B. Genes. In:_____. O conteúdo do genoma. São Paulo: Artmed, 2009. cap.4, p.60-61.

**53** LIU, G.; SONG. L.; SHU. C.; WANG. P.; DENG. C.; PENG. Q.; LERECLUS. D.; WANG. X.; HUANG. D.; ZHANG. J.; SONG. F. Complete Genome Sequence of *Bacillus thuringiensis* subsp. kurstaki Strain HD73. **Genome Announcements,** Washington DC, v. 1, n. 2, p. 80, mar. 2013. PMID89188. PMC3622971.

**54** MAKINO. S.; UCHIDA. I.; TERAKADO. N.; SASAKAWA. C.; YOSHIKAWA. M. Molecular characterization and protein analysis of the cap region, which is essential for encapsulation in *Bacillus anthracis*. **Journal of Bacteriology,** Washington DC, v. 171, n. 2, p. 722–730, fev. 1989. PMC209657.

**55** MANDAL, M.; BOESE, B.; BARRICK, J.E.; WINKLER, W.C.; BREAKER, R.R. Riboswitches Control Fundamental Biochemical Pathways in *Bacillus subtilis* and Other Bacteria. **Cell Press,** Cambridge, v. 113, n. 5, p. 577–586, maio. 2003. DOI101016. PMID12787499.

**56** MANDAL, M.; LEE, M.; BARRICK, J.E.; WEINBERG, Z.; EMILSSON, G.M.; RUZZO, W.L.; BREAKER, R.R. A glycine-dependent riboswitch that uses cooperative binding to control gene expression. **Science,** New York, v. 306, n. 5694, p. 275–279, out. 2004. DOI101126. PMID15472076.

**57** MARCONI, M. A.; LAKATOS. E. M. **Fundamentos de metodologia científica.** 5. Ed. São Paulo: Atlas, 2003.

**58** MLA STYLE. The Nobel Prize in Physiology or Medicine 1968. Nobelprize.org. Nobel Media AB 2014. Web. 15 Dez 2014. Disponível em: < http://www.nobelprize.org/nobel_prizes/medicine/laureates/1968/index.html>. Acesso em: 15 fev. 2015.

**59** MOCK, M.; MIGNOT. T.; Anthrax toxins and the host: a story of intimacy. **Cell Microbiology,** Oxford, v. 5, n. 1, p. 15-23, jan. 2003. DOI101046.

**60** MURAWSKA. E.; FIEDORUK. K.; BIDESHI. D. K.; SWIECICKA. I. Complete Genome Sequence of *Bacillus thuringiensis subsp. thuringiensis* Strain IS5056, an Isolate Highly Toxic to Trichoplusia ni. **Genome Announcements,** Washington DC, v. 1, n. 2, p. 108-113, mar. 2013. DOI101128. PMC3622978.

**61** NAHVI, A. S.; SUDARSAN, N.; EBERT, M. S.; ZOU, X.; BROWN, K. L.; BREAKER, R. R. Genetic control by a metabolite binding mRNA. **Chemistry & Biology,** London, v. 9, n. 9, p. 1043–1049, set. 2002. DOI101016.

**62** NAWROCKI. E.P.; BURGE. S. W.; BATEMAN. A.; DAUB. J.; EBERHARDT. R. Y.; EDDY. S. R.; FLODEN. E. W.; GARDNER. P. P.; JONES. T. A.; TATE.; J.; FINN. R. D. Rfam 12.0: *updates to the RNA families database*. **Nucleic Acids Research,** London, v. 43, n/d, p. 130-137, nov. 2014. DOI101093. PMID25392425.

**63** NAWROCKI. E. P.; EDDY. S. R. Infernal 1.1: 100-fold faster RNA homology searches , **BMC Bioinformatics,** London, v. 15;29, n. 22, p. 2933-2935, nov. 2013. PMID24008419.

**64** NCBI. Website - National Center for Biotechnology Information. National Library of Medicine (NLM), 2009. Blast Help Manual. Disponível em: <http://www.ncbi.nlm.nih.gov/books/NBK1762/>. Acesso em: 14 jul. 2013. Bookshelf ID: NBK1762.

**65** NONCODE. Website - Integrative annotation of long noncoding RNAs. Banco de dados para todos os tipos de ncRNAs exceto tRNAs e rRNAs. Repositório de sequências públicas, perfis de expressão, IncRNAs e funções potenciais, relacionado

ao banco de dados NPInter-ncRNA-Protein Interaction Database. 2012. Disponível em: <http://www.noncode.org/NONCODERv3/guide.htm > . Acesso em: 8 ago. 2013. NONCODERv3.

**66** OKINAKA. R.; CLOUD. K.; HAMPTON. O.; HOFFMASTER. A.; HILL. K.; et al. Sequence, assembly and analysis of pX01 and pX02. **Journal of Applied Microbiology,** Oxford, v. 87, n. 2, p. 261-262, ago. 1999. PMID10475962.

**67** PAGANI, I.; LIOLIOS, K.; JANSSON, J. et al. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata". **Nucleic Acids Research,** London, v. 40, n/d, p. 571-579, jan. 2012. PMC3245063

**68** PANNUCCI. J.; OKINAKA. R.T.; SABIN. R.; KUSKE. C.R.; *Bacillus anthracis* pXO1 plasmid sequence conservation among closely related bacterial species. **Journal of Bacteriology,** Washington DC, v. 184, n. 1, p. 134–141, jan. 2002. PMC134754

**69** PANNUCCI. J.; OKINAKA. R.T.; WILLIAMS. E.; SABIN. R.; TICKNOR. L. O. et al. DNA sequence conservation between the *Bacillus anthracis* pXO2 plasmid and genomic sequence from closely related bacteria. **BMC Genomics,** London, v. 3, n. 34, p. 1471-2164, dez. 2002. DOI101186.

**70** PAREJA, E.; PAREJA-TOBES, P.; MANRIQUE, M.; PAREJA-TOBES, E.; BONAL, J.; TOBES, R. ExtraTrain: a database of Extragenic regions and Transcriptional information in prokaryotic organisms. **BMC Microbiology,** London, v. 15, n. 6, p. 29, mar. 2006. PMID16539733.

**71** PASCHOAL. A. R.; MARACAJA-COUTINHO. V.; Setubal. J.C.; SIMÕES. Z. L. P.; ALMEIDA. S. V.; DURHAM. A. M. Non-coding transcription characterization and annotation: A guide and web resource for non-coding RNA Databases. **RNA Biology,** Georgetown, v. 9, n. 3, p. 274–282, mar. 2012. PMID22336709.

**72** RASKO. D. A; ALTHERR. M. R.; HAN. C. S.; RAVEL. J. Genomics of the *Bacillus cereus* group of organisms. **Microbiology Reviews,** Amsterdam, v. 29, n. 2, p. 303-329, abr. 2005. PMID5808746.

**73** RAVEL. J.; JIANG. L.; STANLEY. S.T.; WILSON. MR.; DECKER. R.S.; READ. T.D.; WORSHAM. P.; KEIM. P.S.; SALZBERG. S.L.; FRASER-LIGGETT. C.M.; RASKO. D.A. The complete genome sequence of *Bacillus anthracis* Ames "Ancestor". **Journal of Bacteriology,** Washington DC, v. 191, n. 1, p. 445-446, jan. 2009. PMID18952800.

**74** READ. T. D.; PETERSON. S. N.; TOURASSE. N.; BAILLIE. L. W.; PAULSEN. I. T.; KOLSTØ. A. B., et al. The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. **Nature,** London, v. 423, n. 6935, p. 81-86, maio. 2003. PMID12721629.

**75** REGALIA, M.; ROSENBLAD, M.A. & SAMUELSSON, T. Prediction of signal recognition particle RNA genes. **Nucleic Acids Research,** London, v. 30, n. 15, p. 3368–3377, ago. 2002. DOI101093. PMID12140321.

**76** RICIETO, A. P. S. IDENTIFICAÇÃO E SEQUENCIAMENTO DE GENES *cry11* DE ISOLADOS BRASILEIROS DE *Bacillus thuringiensis***.** 2012. 60f. Dissertação (Mestrado em Genética e Biologia Molecular) – Centro de Ciências Biológicas, Universidade Estadual de Londrina, Londrina.

**77** ROH. J.Y.; CHOI. J.Y.; LI. M.S.; JIN. B. R.; JE. Y. H. *Bacillus thuringiensis* as a specific, safe, and effective tool for insect pest control. **World Journal of Microbiology and Biotechnology,** Oxford, v. 17, n. 4, p. 547–559, abr. 2007. PMID18051264.

**78** RUTHERFORD. K.; PARKHILL. J.; CROOK. J.; HORSNELL. T.; RICE. P.; RAJANDREAM. MA.; BARREL. B. ARTEMIS: SEQUENCE VISUALIZATION AND ANNOTATION. **Bioinformatics**, Oxford, v. 16, n. 10, p. 944-945, maio. 2000. PMID11120685.

**79** SCHNEPF, E.; CRICKMORE, N.; VAN RIE, J.; LERECLUS, D.; BAUM, J.; FEITELSON, J.; ZEIGLER, D. R.; DEAN, D. H. *B. thuringiensis* and its pesticidal crystal proteins. **Microbiology and Molecular Biology Reviews**, Washington DC, v. 62, n. 3, p. 775-806, set. 1998. PMID9729609.

**80** SHEPPARD. A.E.; POEHLEIN. A.; ROSENSTIEL. P.; LIESEGANG. H.; SCHULENBURG. H. Complete Genome Sequence of *Bacillus thuringiensis* Strain 407 Cry-. **Genome Announcements,** Washington DC, v. 1, n. 1, p. 158, jan. 2013. DOI101128. PMID23405326.

**81** SHERMAN, E. M.; ESQUIAQUI, J.; ELSAYED, G.; YE, J.-D. An energetically beneficial leader-linker interaction abolishes ligand-binding cooperativity in glycine riboswitches. **RNA,** Cold Spring Harbor NY, v. 18, n. 3, p. 496–507, mar. 2012. PMID22279151.

**82** SIMPÓSIO DE CONTROLE BIOLÓGICO, VIVONI. A. M. 12., 2011. São Paulo. Anais eletrônicos... Disponível em: < http://seb.org.br/eventos/SINCONBIOL2011 /palestras.html>. Acesso em: 10 jul. 2013. SINCOBIOL35927 .

**83** SRIDHAR. J.; NARMADA. S. R.; SABARINATHAN. R.; OU. H-Y.; DENG. Z.; SEKAR. K.; RAFI. Z. A.; RAJAKUMAR. K. sRNAscanner: A Computational Tool for Intergenic Small RNA Detection in Bacterial Genomes. **PLoS ONE - Public Library Of Science,** San Francisco, v. 5 n. 8, p. 11970, set. 2010. PMC2916834.

**84** STACKEBRANDT, E.; SWIDERSKI, J. From phylogeny to systematics: the dissection of the genus *Bacillus*. In: Berkeley, R.; Heyndrickx, M.; Logan, N.; De Vos, P. **Applications and systematics of *Bacillus* and relatives.** 1.ed. New York: Wiley Online, 2008. p. 8-22. DOI101002.

**85** STEIN. L. Genome annotation: from sequence to biology. **Nature Reviews Genetics,** Southampton, v. 2, n. 7, p. 493-503, jul. 2001. PMID11433356.

**86** STRÖMSTEN. N. J.; BENSON. S. D.; BURNETT. R. M.; BAMFORD.D. H.; BAMFORD. J. K .H. **Journal of Bacteriology,** Washington DC, v. 185, n. 23, p. 6985–6989, dez. 2003. PMC262720.

**87** SUDARSAN, N; WICKISER, J.K.; NAKAMURA, S.; EBERT, M.S.; BREAKER, R.R. An mRNA structure in bacteria that controls gene expression by binding lysine. **Genes & Development,** Cold Spring Harbor NY, v. 17, n. 21, p. 2688–2697, nov. 2003. PMID14597663.

**88** TICKNOR, L. O.; KOLSTØ, A.; HILL, K. K.; KEIM, P.; LAKER, M. T.; TONKS, M.; JACKSON, P. J. Fluorescent Amplified Fragment Length Polymorphism Analysis of Norwegian *Bacillus cereus* and *Bacillus thuringiensis* soil isolates. **Applied and Environmental Microbiology,** Washington DC, v. 67, n. 10, p. 4863-4873, out. 2001. DOI101128.

**89** TRAVASSOS, G. H., BARROS, M. O., 2003. Contributions of In Virtuo and In Silico Experiments for the Future of Empirical Studies in Software Engineering. In: Proceedings of 2nd Workshop Series on Empirical Software Engineering. Rome: RESEARCH GATE, 2004. p. 117–130. RESG2004.

**90** VAN ERT. M. N.; EASTERDAY. W. R.; HUYNH. L. Y.; OKINAKA. R. T.; HUGH-JONES. M. E. et al. *Global genetic population structure of Bacillus anthracis*. **PLoS ONE - Public Library Of Science,** San Francisco, v. 2, n. 5, p. 461, maio. 2007. DOI101371.

**91** VILAS-BÔAS, G. T.; PERUCA, A. P. S.; ARANTES, O. M. N. Biology and taxonomy of *Bacillus cereus*, *Bacillus anthracis* and *Bacillus thuringiensis*. **Canadian Journal Microbiology,** Ottawa, v. 53, n. 6, p. 673-687, jun. 2007. PMID17668027.

**92** WEI L.; LIU Y, DUBCHAK I, SHON J, PARK J. Comparative genomics approaches to study organism similarities and differences. **Journal of Biomedical Informatics,** San Diego, v. 35, n. 2, p. 142-150, abr. 2002. PMID12474427.

**93** WEINBERG, Z.; WANG, J.X.; BOGUE, J. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea and their metagenomes. **Genome Biology,** Oxford, v. 11, n. 3, p. 31, mar. 2010. PMID20230605.

**94** ZACHAU, H.G.; DÜTTING, D.; FELDMANN, H.; MELCHERS, F.; KARAU, W. Serine specific transfer ribonucleic acids. XIV. Comparison of nucleotide sequences and secondary structure models. In: Symposia on quantitative biology**,** 31., 1966, local. **Anais**… Cold Spring Harbor: CSHLPRESS, 1966. p. 417–424. PMID5237198.

**95** ZAHA, A.; FERREIRA, H.B.; PASSAGLIA, L. M. **Biologia Molecular Básica.** 5. Ed. Artmed, 2014 - 407 pg.

**96** ZENGEL, J.M.; LINDAHL, L. Diverse mechanisms for regulating ribosomal protein synthesis in *Escherichia coli*. **Progress in Nucleic Acid Research and Molecular Biology,** New York, v. 47, n. --, p. 331–370, set. 1994. DOI101016.

**97** ZHU. Y.; SHANG. H.; ZHU. Q.; JI. F.; WANG. P.; FU. J.; DENG. Y.; XU. C.; YE. W.; ZHENG. J.; ZHU. L.; RUAN. L.; PENG. D.; SUN. M.; Complete genome sequence of *Bacillus thuringiensis* serovar finitimus strain YBT-020. **Journal of Bacteriology,** Washington DC, v. 193, n. 9, p. 2379-2380, maio. 2011. PMC3133068.

## PERSPECTIVAS

O objetivo central deste trabalho foi a identificação e caracterização do conjunto de RNAs não codificadores como forma de contribuir para o entendimento do funcionamento de elementos como fatores de virulência e genes *cry*, fundamentais para esclarecimentos sobre o processo de patogenicidade frente aos insetos alvo. Contudo além do objetivo principal, outra meta desejável deste trabalho, com o emprego de extensa metodologia *in silico* e com o consequente aprimoramento das técnicas utilizadas por nossa equipe, uma proposta futura seria a disponibilização de uma ferramenta online mediante utilização de um gerenciador de conteúdos ou um site de buscas para o banco de dados criado, de forma a contribuir com pesquisas externas e também atrair a colaboração de outras universidades e centros de pesquisa.

Também com o propósito de dar continuidade ao trabalho aqui apresentado, propõe-se a expansão das análises desenvolvidas em direção à inclusão de outros organismos que não exclusivamente representantes do grupo do *B. cereus*, ampliando a base de dados de RNAs não codificadores e permitindo a criação de um banco de dados mais abrangente, no interesse desta instituição e de suas parcerias. A distribuição dos ncRNA candidatos obtidos neste trabalho, associados à implicação de trabalhos sobre a relação entre plasmídeos e cromossomos poderia justificar uma segunda etapa de aplicação do pipeline aqui utilizado, de forma a identificar e caracterizar os plasmídeos dos organismos utilizados, para inferência de homologia quanto ao material cromossômico.

**AGRADECIMENTOS**

Somos gratos a toda a equipe de Bioinformática da UTFPR, Campus Cornélio Procópio, notadamente na pessoa do Dr. Alexandre Rossi Paschoal do Grupo de Pesquisa em Bioinformática e Reconhecimento de Padrões, por nos auxiliar fornecendo informações provenientes de dados ainda não publicados, tendo gentilmente realizando o processamento e comparação de nossos dados de sequências intergênicas em seus bancos de dados de ncRNA, utilizados para compor o banco de dados gerador das análises aqui apresentadas. Este estudo também contou com a parceria do Dr. Douglas Silva Domingues, do Instituto agronômico do Paraná, Laboratório de Biotecnologia Vegetal (LBI) / Melhoramento Genética Vegetal. E finalmente aos companheiros de trabalho do Laboratório de Bioinformática do Departamento de Biologia Geral do CCB-UEL, nossos sinceros agradecimentos.

**APÊNDICES**

**APÊNDICE A**

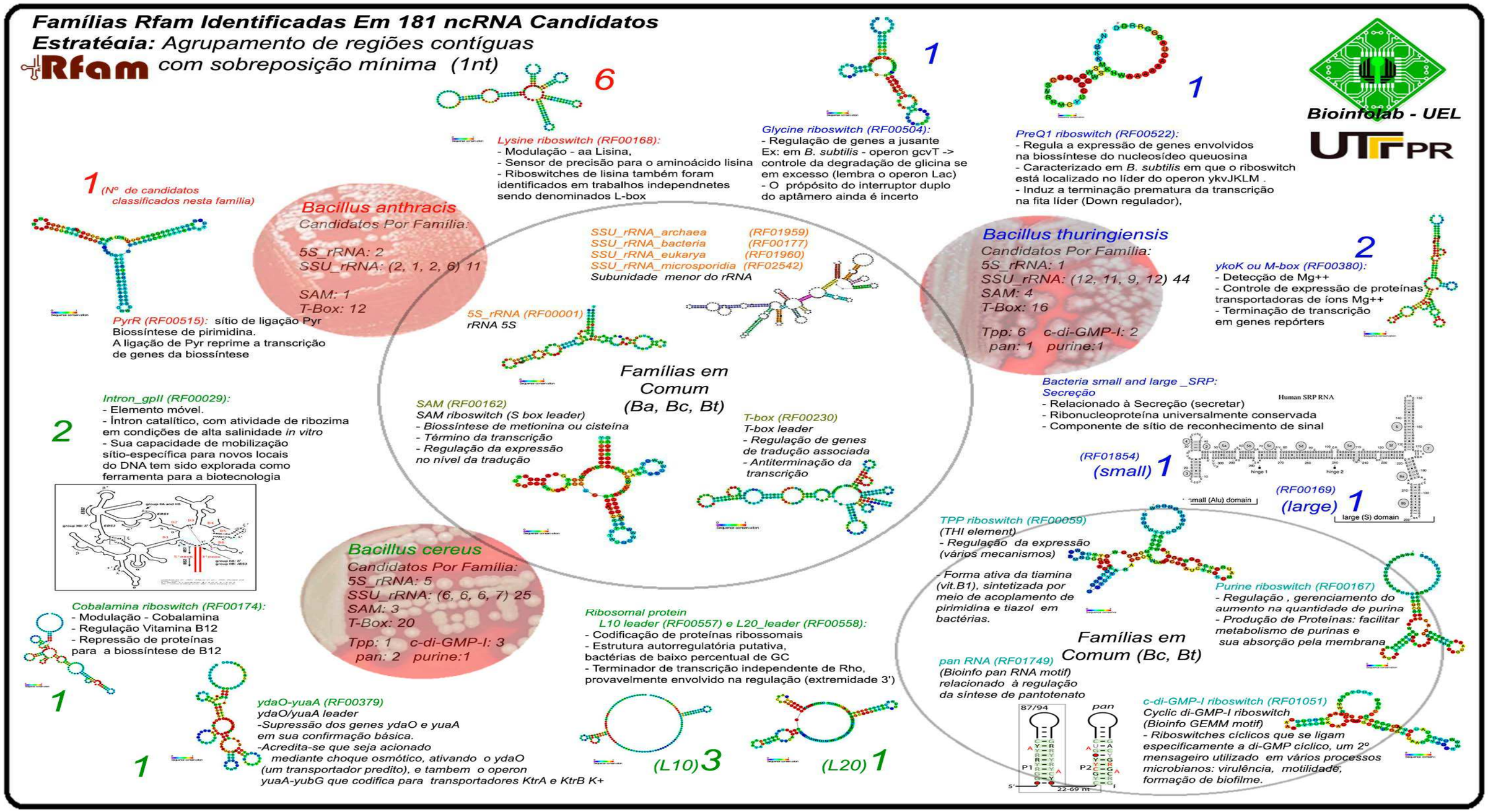


**Figura 18: Famílias Rfam atribuídas aos 181 ncRNA candidatos obtidos, distribuídas por espécie. Legenda por cores (famílias exclusivas), em vermelho referem-se a *B. anthracis*, sendo o verde e o azul respectivamente para *B. cereus* e *B. thuringiensis.* As famílias em comum para os três grupos podem ser visualizadas no círculo maior (ao centro), enquanto que a família *TPP* (canto inferior direito) é compartilhada somente por dois grupos. Utilizando o mesmo padrão de cores, os números maiores, ao lado da representação da estrutura secundária para cada família indicam o número de candidatos identificados em cada família. Os totais para as famílias compartilhadas são exibidos na representação da Placa de Petri, para cada espécie.**

**APÊNDICE B**

Tabela 13: Número de ncRNA candidatos (preditos) agrupados por espécie, cepa, famílias Rfam e características do genoma.

| Gênero | | *BACILLUS* | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grupo | | Grupo do *Bacillus cereus*, constituído por seis espécies: *Bacillus cereus*, considerada a espécie sensu stricto e característica do grupo, *Bacillus anthracis*, *Bacillus thuringiensis*, *Bacillus mycoides*, *Bacillus pseudomycoides* e *Bacillus weihenstephanensis*. | | | | | | | | | | | | | | | | | | | | | | | |
| *spp* | | *Bacillus anthracis:* ncRNA Preditos | | | | | | *Bacillus Cereus:* ncRNA Preditos | | | | | | | | | | | *Bacillus Thuringiensis:* ncRNA Preditos | | | | | | | | |
| Característica | | Membros desta espécie não são móveis e são todos caracterizados pela presença de quatro profagos e uma mutação sem sentido no gene regulador plcR (KLEE et al., 2010) | | | | | | Espécie onipresente in natura. Capacidade de produzir inúmeros fatores de virulência não específicos, incluindo fosfolipases, hemolisinas e enterotoxinas (HAN et al., 2006) | | | | | | | | | | | Característica entomopatogênica, deve-se à formação de proteínas denominadas Cry, formadas em seu ciclo de esporulação, juntamente com um esporo elipsoidal. Tais proteínas se juntam e cristalizam ainda dentro da célula vegetativa, formando cristais proteicos (CRICKMORE et al., 1998) | | | | | | | | |
| ID Banco de dados grupo_cereus | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
| Cepa (strain) | | *A0248* | *'Ames Ancestor'* | *Ames* | *CDC 684* | *H9401* | *Sterne* | *03BB102* | *AH187* | *AH820* | *ATCC 10987* | *ATCC 14579* | *B4264* | *E33L* | *F837/76* | *G9842* | *NC7401* | *Q1* | *Al Hakam* | *BMB171* | *serovar chinensis CT-43* | *serovar finitimus YBT-020* | *serovar konkukian str. 97-27* | *serovar kurstaki str. HD73* | *Bt407* | *MC28* | *IS5056* |
| Tamanho dos Genomas (p | | 5227419 | 5227419 | 5227293 | 5230115 | 5218947 | 5228663 | 5269628 | 5269030 | 5302683 | 5224283 | 5411809 | 5419036 | 5300915 | 5222906 | 5387334 | 5221581 | 5214195 | 5257091 | 5330088 | 5486830 | 5355490 | 5237682 | 5646799 | 5500501 | 5414494 | 5491935 |
| % GC no Genoma | | 35,38 | 35,38 | 35,38 | 35,38 | 35,37 | 35,38 | 35,44 | 35,59 | 35,41 | 35,59 | 35,28 | 35,30 | 35,35 | 35,37 | 35,26 | 35,59 | 35,56 | 35,43 | 35,29 | 35,38 | 35,54 | 35,41 | 35,28 | 35,41 | 35,41 | 35,38 |
| UID_NNNNN | | **59385** | **58083** | **57909** | **59303** | **162021** | **58091** | **59299** | **58753** | **58751** | **57673** | **57975** | **58757** | **58103** | **83611** | **58759** | **82815** | **58529** | **58795** | **49135** | **158151** | **158875** | **58089** | **189188** | **177931** | **176369** | **190186** |
| NC_NNNNNN | | **NC_012659** | **NC_007530** | **NC_003997** | **NC_012581** | **NC_017729** | **NC_005945** | **NC_012472** | **NC_011658** | **NC_011773** | **NC_003909** | **NC_004722** | **NC_011725** | **NC_006274** | **NC_016779** | **NC_011772** | **NC_016771** | **NC_011969** | **NC_008600** | **NC_014171** | **NC_017208** | **NC_017200** | **NC_005957** | **NC_020238** | **NC_018877** | **NC_018693** | **NC_020376** |
| Sigla Família | Total | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- |
| 5S_rRNA | 8 | 1 | | | | 1 | | | | 1 | 1 | | | 1 | | 1 | 1 | | 1 | | | | | | | | |
| Bacteria_small_SRP | 1 | | | | | | | | | | | | | | | | | | | | | | | | | 1 | |
| c-di-GMP-I | 5 | | | | | | | | | | | 2 | 1 | | | | | | | 1 | | | | 1 | | | |
| Cobalamin | 1 | | | | | | | 1 | | | | | | | | | | | | | | | | | | | |
| Glycine | 1 | | | | | | | | | | | | | | | | | | | | | | | | | 1 | |
| Intron_gpII | 2 | | | | | | | | 2 | | | | | | | | | | | | | | | | | | |
| L10_leader | 3 | | | | | | | | 1 | 1 | 1 | | | | | | | | | | | | | | | | |
| L20_leader | 1 | | | | | | | | | 1 | | | | | | | | | | | | | | | | | |
| Lysine | 6 | | 1 | 1 | 1 | 2 | 1 | | | | | | | | | | | | | | | | | | | | |
| pan | 3 | | | | | | | | | 1 | | | | | 1 | | | | | | | | 1 | | | | |
| PreQ1 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | 1 | |
| Purine | 2 | | | | | | | | | | 1 | | | | | | | | | | | | | 1 | | | |
| PyrR | 1 | | | | 1 | | | | | | | | | | | | | | | | | | | | | | |
| SAM | 8 | 1 | | | | | | | 1 | | 1 | | | 1 | | | | | | 1 | 1 | | | | | 1 | 1 |
| SSU_rRNA_archaea | 20 | | | | 1 | 1 | | 2 | 1 | 1 | | | | | 1 | 1 | | | | 2 | | 2 | 1 | 1 | 1 | 1 | 4 |
| SSU_rRNA_bacteria | 18 | | | | | 1 | | | 1 | | 2 | | 1 | 2 | | | | | 2 | 3 | | 1 | 1 | 2 | | 1 | 1 |
| SSU_rRNA_eukarya | 17 | | | | 1 | 1 | | | | 2 | | | 1 | | 1 | | 1 | 1 | | 4 | 1 | | | 1 | 2 | | 1 |
| SSU_rRNA_microsporidia | 25 | 2 | 1 | | 2 | 1 | | | | 1 | 2 | | | | 1 | 1 | 1 | 1 | | 3 | 1 | 1 | | 2 | 3 | | 2 |
| T-box | 47 | 4 | 1 | 6 | | 1 | | 2 | 3 | 2 | | | | 2 | 5 | 5 | 1 | | 1 | | | 3 | 4 | 2 | 2 | 3 | |
| TPP | 7 | | | | | | | | | | | | 1 | | | | | | | 1 | | | | 2 | 1 | | 2 |
| ydaO-yuaA | 1 | | | | | | | | 1 | | | | | | | | | | | | | | | | | | |
| ykoK | 2 | | | | | | | | | | | | | | | | | | | | | | | | | 2 | |
| Total Cepa | 180** | 8 | 3 | 7 | 6 | 8 | 1 | 5 | 10 | 10 | 8 | 2 | 4 | 6 | 9 | 8 | 4 | 2 | 4 | 15 | 3 | 7 | 9 | 12 | 7 | 11 | 11 |

* Total de candidatos por família; ** *B. anthracis Ames* - 1 Família não identificada; Fonte: banco de dados grupo_cereus 2015

Fonte: Pipeline do Laboratório de Bioinformática CCB-UEL, 2015, Total 23 famílias Rfam - Banco de dados grupo_cereus. Famílias - Rfam V11.0.

## APÊNDICE C

# 1 PROTOCOLO DE TRABALHO PARA LOCALIZAÇÃO E CARACTERIZAÇÃO DE RNAS NÃO CODIFICADORES;

### 1.1. O Servidor: Hardware e Software

Ferramentas utilizadas no Laboratório de Bioinformática do CCB-UEL:

**Software:** o servidor onde foram realizadas as análises propostas utilizou Sistema Operacional (SO): Ubuntu 12.04 Server e SO: Windows Server 2012 com licença de avaliação, para Junho 2013.

**Hardware, especificações do Servidor HP ProLiant ML350p Gen8:**

- Chassi: Torre;
- CPU: 2 x 2GHz Xeon E5-2650;
- Memória: 16GB DDR3 RDIMM com expansão até 384GB (768GB com LR-DIMMs);
- Storage (armazenamento): 2 x 600GB HP SAS 10K hot-swap SFF hard disks (max.24);
- RAID: Smart Array P420i with 2GB FBWC/capacitor;
- Array support: RAID0, 1, 10, 5, 50, 6, 60;
- Expansão: 8 x PCI-e Gen3, 1 x PCI-e Gen2;
- Rede: 4 x Gigabit;
- Energia: 2 x 750W abastecimento hot-plug;
- Gerenciamento remoto: HP iLO4[8] Standard com 10/100;

### 1.2. Extração das Regiões Intergênicas: Software Artemis

Utilizar a funcionalidade "Intergenic Features" (em livre tradução: características, recursos intergênicos).

Execução: Menu Principal / Create Menu.

Ação: geração de arquivos fasta com as sequências intergênicas para cada linhagem em estudo (Figura 19). Tal funcionalidade processa os arquivos de formato

[8] iLO: Os servidores HP possuem o serviço de gerenciamento iLO (HP Integrated Lights-Out) inclusive para acesso via smartfones, em verificação a viabilidade financeira de utilização deste, caso contrário serão utilizadas tecnologias de acesso livres como SSH (Linux) ou TeamViewer (Máquinas virtuais no mesmo servidor).

padrão de anotação (neste caso GenBank, *.gbk) gerando a análise para as regiões entre CDS anotados nas sequências processadas, identificando tais regiões como "misc_feature" (características, traços diversos).

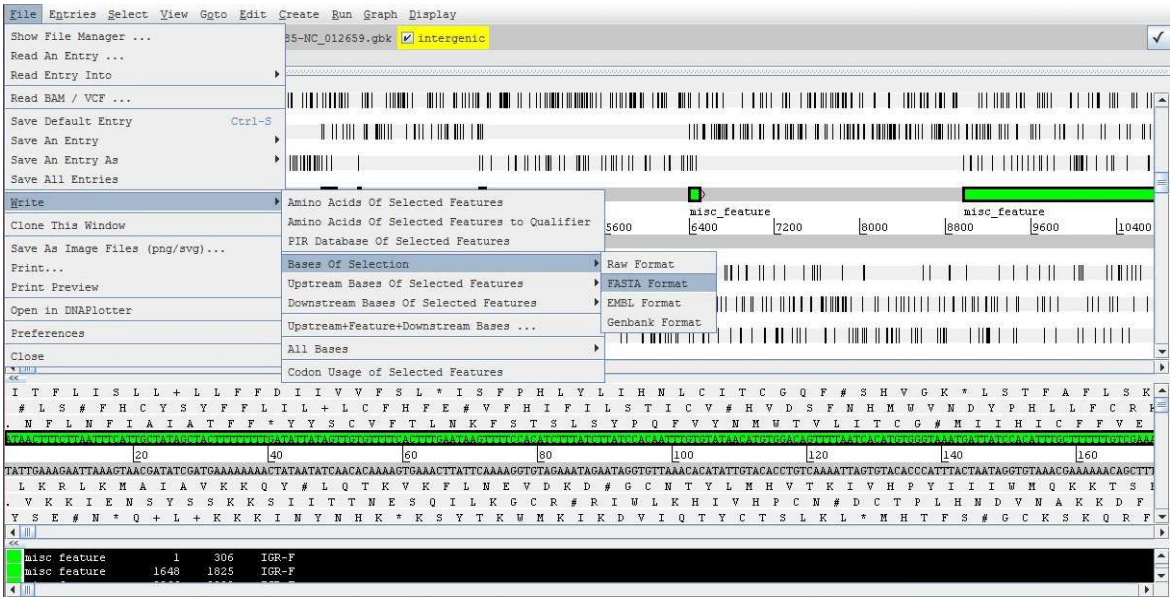


**Figura 19: Geração do arquivo fasta com sequências intergênicas**

Entre os genomas pesquisados para a etapa de testes do pipeline, o Software Artemis (versão utilizada 16.0.0) apresentou restrições no processamento de alguns itens, listados a seguir. A listagem abaixo mantém a nomenclatura de processamento dos arquivos, notadamente não formatados para a notação científica, também mantivemos o texto original do log de processamento:

- Bacillus_cereus_biovar_anthracis_CI_uid50615-NC_014335: processado com restrições[9];
- Bacillus_cereus_FRI_35_uid173403-NC_018491: não processado;
- Bacillus_cereus_Q1_uid58529-NC_011969: processado com restrições;
- Bacillus_thuringiensis_HD_771_uid173374-NC_018500: não processado;
- Bacillus_thuringiensis_HD_789_uid173860-NC_018508: não processado;

### 1.3. **Métodos de Inferência Computacional: Identificação de ncRNAs**

[9] Exemplos de “warnings” - logs de processamento, restrições, avisos: (1) BAM & VCF not visible, (2) source can't have biovar as a qualifier, (3) tRNA can't have codon_recognized as a qualifier, (4) internal stop codon found, (5) no valid stop codon found.

- **Primeiro método:** processamento via Infernal V.1.1 / banco de dados Rfam V.11.0,
- **Segundo método:** análise comparativa com o banco de dados da UTFPR que reúne ncRNAs conhecidos e anotados com base em pesquisa no Non-coding RNA Databases Resource (NRDR).
- **Terceiro método:** processamento via sRNAscanner V.1.9. Atenção: esta etapa não utiliza as regiões intergênicas, ver especificações no item 1.3.3 (Suíte sRNAScanner);
- **Entrada de dados/Input:** os dados das regiões intergênicas dos 26 genomas selecionados, identificadas pelo Artemis foram gerados e salvos em arquivos multifasta (Figura 20).
- **Procedimento:** submeter os arquivos multifasta a dois métodos de inferência computacional para identificação de ncRNAs.

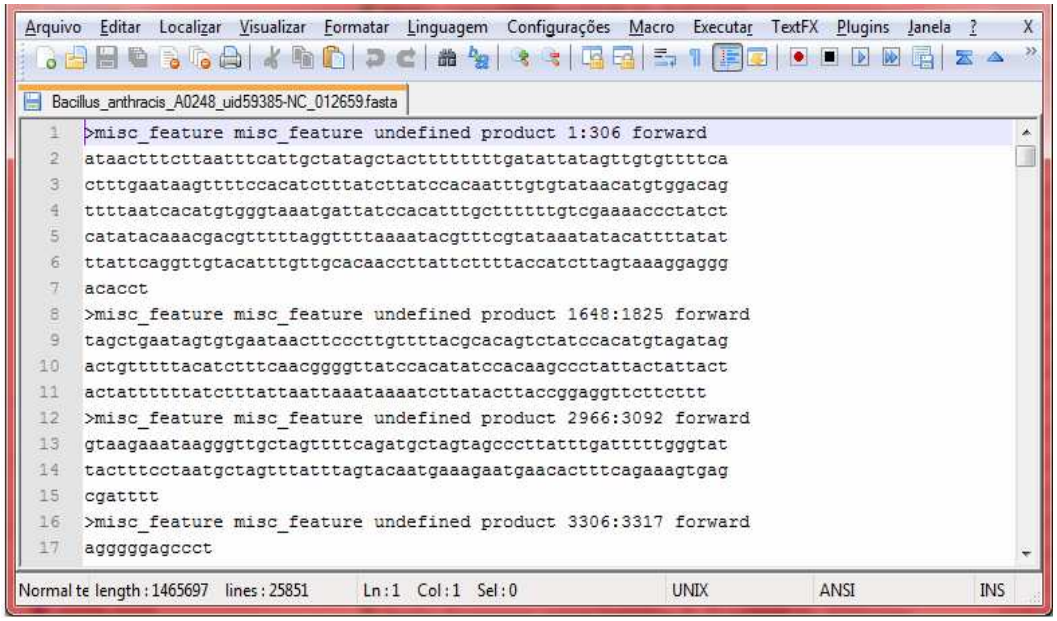


**Figura 20: Arquivo multifasta de sequências intergênicas (Artemis V.16.0.0)**

### 1.3.1. Banco de Dados Rfam e o Pacote de Software Infernal

**Método de descoberta de ncRNAs:** Rfam X Infernal

**Banco de dados Rfam (RNA Families):** informações sobre famílias de ncRNAs com base na evolução de um ancestral comum possibilitando inferir informações sobre a sua estrutura e função;

Acesso: livre e base de dados disponível para download e instalação local (utilização do pacote de software INFERNAL)

Instituição mantenedora: Instituto Wellcome Trust Sanger e colaboradores.

Funcionalidades da interface: pesquisa de ncRNAs por palavra-chave, nome de família, ou genoma, sequências de ncRNA, número de acesso EMBL.

Infernal (Inference of RNA Alignment): anotação de genomas completos ou sequências homólogas a ncRNAs conhecidos. Pesquisa de estruturas de RNA e similaridades entre sequências a partir de bancos de dados de sequências de DNA.

1.3.2. **Etapa de inferência computacional:**

- **Entrada de dados/Input:** 26 arquivos contendo as sequências intergênicas
- **Ambiente/Máquina:** processamento local - servidor HP ProLiant ML350p Gen8, sistema operacional Ubuntu Server 64Bits;
- **Software utilizado:** software Infernal V.1.1
- **Banco de dados para pesquisa e versão:** Rfam V.1.1.;

- **Saída de dados/Output:** 26 arquivos, conforme os parâmetros nas linhas de comando (ambiente Linux) do exemplo local abaixo (demonstração da aplicação dos parâmetros):

**Sintaxe:**

***usuario@maquina:****/diretório#programa_de_busca <parâmetros> ...*
***-o*** *Arquivo_saida.out <multifasta>*
***-A*** *Arquivo_saida.aln <multifasta>*
***--tblout*** *Arquivo_saida.tbl <tabela>*
*Banco.versao*
*Arquivo_entrada.fasta &*

**Parâmetros utilizados:**

***root@HP_server:****/../artemis_ign_fasta#cmsearch*
***-o*** *Bacillus_anthracis_A0248_uid59385-NC_012659****.out***
***-A*** *Bacillus_anthracis_A0248_uid59385-NC_012659****.aln***
***--tblout*** *Bacillus_anthracis_A0248_uid59385-NC_012659****.tbl***

***Rfam.cm.1_1***

*Bacillus_anthracis_A0248_uid59385-NC_012659.fasta* ***&***

1.3.3. **Suíte sRNAscanner**

- **Método de descoberta de ncRNAs:** sRNAscanner.
- **Requisito do programa sRNAscanner:** execução em ambiente Linux.
- **Execução:** Servidor HP ProLiant ML350p Gen8 (Infolab). Sistema operacional Ubuntu Server 64Bits.
- **Rotinas corretivas (quando necessário):** Laptop Acer Aspire 4745-7494, Sistemas Operacionais Windows Seven Ultimate - service pack 1 64Bits em dual boot com Ubuntu 14.04 LTS 64Bits, conforme tutorial abaixo (**Tabela 14** e Figura 21).

- **Execução do sRNAscanner para processamento dos genomas escolhidos:**

O sRNAscanner recebeu como entradas para cada genoma, os respectivos arquivos renomeados como:

Acession.Gen.Str.Uid.ptt: Formato NCBI para tabelas de proteínas;

Acession.Gen.Str.Uid.fna: Formato NCBI para sequências fasta de genomas;

Desta forma note-se que, devido ao processo de identificação para ncRNAs utilizado pelo sRNAscanner, o processamento nesta etapa não utilizou as sequências intergênicas produzidas pelo Artemis.

- **Entrada de dados:** obter somente os arquivos *.fna e *.ptt (Diretório/repositório: 010-FTP-NCBI) ;

  **Obs:** escolher arquivos identificados como “complete genome”;

- **Descrição dos formatos NCBI:**

  *.gbk: genome in genbank file format;

  *.fna: genome fasta sequence;

  *.ptt: protein table;

- **Diretórios:** Utilizar um diretório para cada execucão do sRNAscanner, ou criar vários Input.data, como[10]:

  **Input1.data**

  **Input2.data**

  **Input3.data**

- Cada Input.data deve apontar para seu conjunto de arquivos específico (*.fna e *.ptt); Pode-se reunir todos os arquivos fna e ptt na mesma pasta, acrescentando-se o nome padrao dos gbks, ex:

Tabela 14: sRNAscanner - Exemplo de preparação dos arquivos para processamento

| Nomenclatura original | Nomenclatura utilizada como padrão neste projeto |
|---|---|
| NC_008600.gbk | Bacillus_thuringiensis_Al_Hakam_uid58795-NC_008600.gbk |
| NC_008600.fna | Bacillus_thuringiensis_Al_Hakam_uid58795-NC_008600.fna |
| NC_008600.ptt | Bacillus_thuringiensis_Al_Hakam_uid58795-NC_008600.ptt |

Fonte: Pipeline do Laboratório de Bioinformática CCB-UEL, 2014.

- **Estrutura de pastas:**

***usuario@maquina:***~*Diretorio_Base*\ <*arquivos de configuração e executáveis*>

***usuario@maquina:***~*Diretorio_Base*\ *output_files* \ <*arquivos de processamento - resultados*>

***usuario@maquina:***~*Diretorio_Base*\ *tmp* \ <*arquivos temporários de execução, testes, matrizes*>

- Para evitar sobrescrever os dados, os resultados da pasta output_files foram removidos a cada nova geração de arquivos[11];

Cada pasta de execução pode ser renomeada conforme o organismo, ex:
\output_files_Bacillus_thuringiensis_Al_Hakam_uid58795-NC_008600

- O SRNAscanner foi executado no diretório de processamento conforme a linha de comando exemplificativa a seguir (Figura 21):

***andre@andre-Aspire-4745:***~/Documentos/062-sRNA processamento/sRNAscanner_Ubuntu10$ *./sRNAscanner_Ubuntu10.exec Input.data*

[10] Otimização: utilizar a nomenclatura listada no exemplo para o caso de implementação de uma estrutura de repetição (iteração ou loop) como um "for".

[11] O processamento de um genoma de 5Mb/5.000Kb tem tempo de execução ~ 60 minutos.

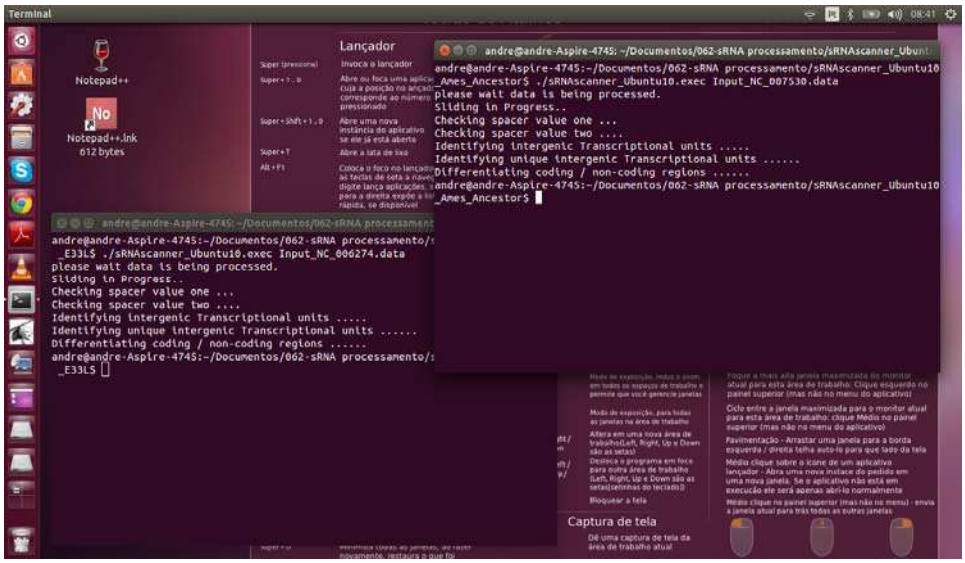


**Figura 21: sRNAscanner - processamento de duas *strains* simultaneamente**

1.4. **Banco de Dados Final: PostgreSQL**

Banco de dados e versão: PostgreSQL V.9.1;

Nomenclatura do banco de dados criado: Grupo_cereus (dados provindos dos métodos de processamento e descoberta, Figura 2);

Foi utilizada a nomenclatura das tabelas criadas – Ver Material Suplementar no DVD: Banco de Dados\Scripts SQL):

- **Etapa 1:** Instruções DDL para criação do banco de dados;
  (Material suplementar: _001-criacao_banco.sql)
- **Etapa 2:** Carga das tabelas primárias do banco de dados;
  a. InfernalRfam:
  b. (Material suplementar: _Tabular output formats - Descricao Colunas.txt)
  c. sRNAscanner: 26 itens (multifasta);
  d. Genomas: 26 genomas (*.fna);
  e. bd_NRDR: 1369 candidatos carregados, selecionadas em um total de 16328;
- **Etapa 3:** Instruções DML para adequação do banco;
  (Material suplementar: _002-correcoes.sql)
  **Operações:**
  Cálculo de posições de início e fim reais no infernal, (tabelas: genomas X infernal);
  Cálculo de posições de início e fim reais no nrdr, (tabelas: genomas X nrdr);
  Verificações para fita reverse (“-”);
  **Testes:**

Seleção de regiões coincidentes;
Selecionar onde infernal está contido no srna;
Selecionar onde infernal contem o srna;
Selecionar onde infernal = srna;
Selecionar srna onde não existe coincidência com infernal;
Selecionar infernal onde não existe coincidência com srna;
Seleção das sequências ncRNAs identificadas;
Seleção das regiões coincidentes com união dos métodos de descoberta;
Geração das tabelas web e ftp para a criação da tabela geral de famílias Rfam;
**Correções:** caracteres especiais não tratados após carga no banco de dados;
**Criação das tabelas de visão:** relatórios;

- **Etapa 4:** Instruções DML para preparação das análises;
  (Material suplementar: _003-atualizacoes.sql)
  Tratamento das famílias Rfam: inserção das descrições e espécies por família;
  Indexação da tabela de famílias para otimização das buscas;
  Geração do agrupamento final: 534 famílias;
  Inserção das famílias na tabela de grupos;
  Indexação: geração do nc_gen sem a versão;
  Verificações: existência de dois accessions para genomas diferentes;
  Após updates: seleciona todos os infernais que contem srna;
  Após updates: seleciona todos os infernais que contem nrdr;

- **Etapa 5:** Instruções DML para extração de dados e geração dos relatórios;
  (Material suplementar: _004-montagem_totalizadora_cabecalhos_transposta.sql)
  Montagem da tabela totalizadora: cabeçalho - grupo, espécie, id, cepa, tamanho e percentual de GC nos genomas, identificadores UID e NC;

Montagem da tabela totalizadora: corpo de dados - totalizadores dos métodos de descoberta (3) por estratégia (6);