Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

LeifDB: um banco de dados para organização de informações provenientes da análise comparativa do genoma da bactéria *Leifsonia xyli*

Pedro Antonio de Castro Bittencourt

JUIZ DE FORA
NOVEMBRO, 2017

LeifDB: um banco de dados para organização de informações provenientes da análise comparativa do genoma da bactéria *Leifsonia xyli*

Pedro Antonio de Castro Bittencourt

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Saul de Castro Leite

Coorientador: Fernanda Nascimento Almeida

JUIZ DE FORA

NOVEMBRO, 2017

LeifDB: um banco de dados para organização de informações provenientes da análise comparativa do genoma da bactéria *Leifsonia xyli*

Pedro Antonio de Castro Bittencourt

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Saul de Castro Leite
Doutor em Modelagem Computacional

Fernanda Nascimento Almeida
Doutora em Bioinformática

Victor Stroele De Andrade Menezes
Doutor em Engenharia de Sistemas e Computação

Wagner Antonio Arbex
Doutor em Engenharia de Sistemas e Computação

JUIZ DE FORA
30 DE NOVEMBRO, 2017

Resumo

A cultura da cana-de-açúcar no Brasil é utilizada principalmente para a produção de açúcar e etanol. Porém existem doenças que acometem a cultura e que preocupam o setor canavieiro quanto a redução de produtividade, dentre elas uma das mais importantes é o raquitismo da soqueira, causada pela bactéria *Leifsonia xyli* subsp. *xyli*. Existem várias espécies que compõem o gênero Leifsonia, sendo que somente uma delas é patogênica a plantas, as demais são bactérias de vida livre. Neste sentido, o objetivo deste trabalho é organizar as informações sobre o genoma completamente sequenciado destas bactérias, quando disponíveis, a fim de estabelecer um ambiente que organize as informações e que auxilie no processo de comparação de sequências e anotação funcional entre as espécies deste gênero, visando aspectos associados a patogenicidade. O banco de dados em construção, chamado de LeifDB, possui atualmente informações obtidas a partir do genoma completamente sequenciado de 11 bactérias (dentre elas: 5 espécies do gênero Clavibacter e 6 espécies do gênero Leifsonia). As informações do genoma de espécies Clavibacter foram adicionadas por se tratar do gênero mais próximo a Leifsonia e da mesma forma apresentando espécies patogênicas a plantas. O banco contém a predição de ORFs dos 11 genomas, obtida pelo programa PROKKA, categorização funcional de acordo com o COG e anotação funcional descrita para Projeto Genoma de bactérias do gênero *Xanthomonas*. O objetivo principal é auxiliar a identificação de grupos de genes associados a patogenicidade. O LeifDB possui um total de 35904 registros que representam os genes associados a 11 bactérias. Deste total foram agrupados e categorizados 87,44% sendo apenas 6,5% genes de categoria funcional indefinida ou hipotética.

Palavras-chave: Banco de dados, Bioinformática, *Leifsonia xyli*, Análise Comparativa, Patogenicidade.

Abstract

Sugarcane culture in Brazil is mainly used for the production of sugar and ethanol. However, there are diseases that affect the production and concern the sugar industry with the reduction of productivity. Among them, one of the most important is the ratoon stunting disease, caused by the bacterium *Leifsonia xyli* subsp. *xyli*. There are several species that compose the genus *Leifsonia*. Only one of them is pathogenic to plants and the others are free living bacteria. The objective of this work is to organize information about the fully sequenced genome of these bacteria, when available, in order to establish an environment that organizes the information and helps the process of sequence comparison and functional annotation among species of this genus, aiming at aspects related to pathogenicity. The database under construction, called LeifDB, currently has information obtained from fully sequenced genome of 11 bacteria (among them: 5 species of the genus *Clavibacter* and 6 species of the genus *Leifsonia*). The genomic information of Clavibacter species was included because Clavibacter is the closest genus to *Leifsonia* and it also contains species that are pathogens to plants. The database contains the ORFs prediction of the 11 genomes obtained by the PROKKA program, functional categorization according to COG and functional annotation described for the Xanthomonas Genome Project. The main goal is to help identify groups of genes associated with pathogenicity. LeifDB has a total of 35904 records representing the genes associated with 11 bacteria. Of this total, 87.44 % were grouped and categorized, with only 6.5 % genes of undefined or hypothetical functional category.

Keywords: Database, Bioinformatics, *Leifsonia xyli*, Comparative analysis, Pathogenicity.

Agradecimentos

Gostaria de agradecer primeiramente a Deus por eu ter conseguido chegar até onde estou hoje, aos meus pais, José Antonio e Marinete, pelo apoio e sustento, aos meus orientadores, Fernanda e Saul, pelo encaminhamento e paciência, aos professores que me ensinaram tanto e a todos que de alguma forma contribuiram nesta minha caminhada.

Gostaria de agradecer também a Claudia Barros Monteiro-Vitorello da Escola Superior de Agricultura Luiz de Queiroz da Universidade de São Paulo (ESALQ/USP) por ter disponibilizado os dados utilizados neste trabalho.

"Seja a mudança que você quer ver no mundo".

Mahatma Gandhi

Conteúdo

Lista de Figuras

Lista de Tabelas

1 Introdução

Atualmente, o Brasil é o país que mais produz cana de açúcar no mundo, sua produção somada com a da Índia, o segundo maior produtor, alcança pouco mais da metade do total produzido mundialmente (G1, 2017). Com a cana de açúcar é feito o açúcar, que está presente na maioria dos países, e também o etanol, que pode ser utilizado como biocombustível ou matéria prima da indústria servindo para a produção de perfumes, tintas, materiais de limpeza entre outros.

Muitos fatores podem atrapalhar ou reduzir a produção da cana, entre eles tem-se as doenças que prejudicam e até acabam com algumas plantações. Uma doença em particular provoca o afinamento do caule da planta e consequentemente há uma diminuição na produção. Essa doença é chamada de raquitismo da soqueira e é provocada por uma bactéria, *Leifsonia xyli* subsp. *xyli*. Análises feitas por Urashima et al. (2017) em algumas plantações da região centro-oeste do Brasil mostrou uma perda econômica anual de US$ 1 milhão somente nas amostras estudadas e ainda revela o uso de material infeccionado em 10% da área cultivada. Assim, aumenta a preocupação dos produtores e medidas precisam ser tomadas para cura e prevenção dessa doença.

Não existe um tratamento eficaz ou cura definitiva para essa doença. Uma solução estudada atualmente é a realização de análises comparativas do genoma completamente sequenciado dessa bactéria com o de outras espécies não patogênicas (ou seja, organismos que estão presentes no hospedeiro mas não manifestam doenças), para identificar as regiões associadas à patogenicidade. A partir deste estudo será possível entender o mecanismo associado a virulência da bactéria e com isso identificar estratégias de prevenção mais adequadas ao combate deste patógeno.

A bactéria *Leifsonia xyli* subsp. *cynodontis* (Lxc) é uma subespécie de *Leifsonia xyli* e sua semelhança com a *Leifsonia xyli* subsp. *xyli* motiva seu estudo. Esta bactéria pode estar presente na cana-de-açúcar, entretanto não causa nenhuma doença. Ela se manifesta no capim de bermuda causando raquitismo e assim prejudicando o crescimento de seu hospedeiro.

Entretanto, as informações relacionadas ao genoma completamente sequenciado dessas bactérias não estão contidos em um ambiente comum, o que dificulta o processo de análise comparativa do genoma de bactérias do gênero *Leifsonia*. Diante disso, devido às ferramentas web disponíveis para comparação de genomas se faz necessário mecanismos automatizados que consigam processar e organizar as informações provenientes de diferentes ferramentas e bancos de dados que tem esta finalidade. É sabido que se esse tipo de análise for feita de maneira manual, os resultados demorariam dias ou meses para serem obtidos e estariam mais suscetíveis a erros do que um processo automático.

Existem diferentes variações das bactérias do gênero *Leifsonia* cujo genomas foram completamente sequenciado e estão disponíveis na internet. Sabe-se que com o auxílio e aplicação das técnicas e métodos da área de Banco de Dados é possível integrar e organizar em um ambiente comum qualquer tipo de dado. Com isso, esta monografia tem como tema principal integrar e organizar em uma plataforma segura as informações sobre o genoma completamente sequenciado de bactérias do gênero *Leifsonia*, em especial, da espécie *Leifsonia xyli* subsp. *xyli* (Lxx) a fim de estabelecer um ambiente que propicie consultas e que auxilie os pesquisadores no processo de comparação do genoma desta bactéria.

Na Bioinformática, a criação de bancos de dados com informações biológicas, os chamados bancos de dados biológicos, permitem o armazenamento, a administração, a extração e a difusão generalizada e sistemática dessa informação, disponibilizando ferramentas computacionais desenvolvidas para atualizar, pesquisar e recolher dados armazenados no sistema. Neste sentido, elas maximizam a quantidade de informações biológicas extraídas a partir das sequências de DNA, o que facilita a comparação dos dados. No caso da análise comparativa de genomas, ela pode ser facilitada quando as informações advindas de fontes distintas estão organizadas em um único ambiente, facilitando e agilizando o acesso aos dados e disponibilizando um método para extrair a informação necessária.

Seguindo esse pensamento, o LeifDB (*Leifsonia xyli* Database) tem como proposta ser um banco de dados relacional que integre e disponibilize de forma organizada as informações relativas à análise comparativa do genoma das bactérias da espécie *Leifsonia*. Foram incluídas nas análises comparativas bactérias do gênero Clavibacter, que

são patogênicas, pertencentes a mesma família da Leifsonia e que possuem características similares quanto aos mecanismos de atuação no hospedeiro.

1.1 Objetivo

O objetivo principal deste trabalho é construir um banco de dados relacional com as informações relativas a comparação do genoma da bactéria Lxx e bactérias do mesmo gênero. Para que seja possível a realização de consultas pelos pesquisadores da área de biologia, uma interface gráfica para acesso aos dados foi desenvolvida. São objetivos específicos deste trabalho:

- Implementação de um banco de dados para integrar e organizar as informações referentes da análise comparativa do genoma de diferentes linhagens da *Leifsonia xyli*.
- Aplicar as categorias funcionais descritas pelo banco de dados COG (*Cluster of Orthologs Group*) e pelo Projeto Genoma de bactérias do gênero *Xanthomonas* aos genes das bactérias *Leifsonia xyli* e demais presentes no banco de dados.
- Propor uma interface gráfica para acesso às informações do banco de dados LeifDB com restrição de acesso.

1.2 Organização do Trabalho

Este trabalho está dividido em capítulos. No capítulo 2 é mostrado o referencial teórico necessário para a realização do trabalho, sendo apresentado Bancos de Dados, a bactéria *Leifsonia xyli* e as ferramentas computacionais utilizadas. O capítulo 3 apresenta como foi feita a estruturação do LeifDB, os organismos que estão armazenados e como se deu a categorização funcional pelo COG e *Xanthomonas*, além de apresentar a importância de uma interface web. No capítulo 4 são mostrados os resultados alcançados e no capítulo 5 é feita uma conclusão sobre o trabalho.

2 Revisão Bibliográfica

Esta revisão se preocupa em apresentar uma descrição sobre bancos de dados do ponto de vista computacional e biológico. Além disso, introduz o organismo que motivou a criação do LeifDB e as ferramentas computacionais que auxiliaram no processo de tratamento dos dados presentes no referido banco de dados.

2.1 Banco de Dados

São considerados bancos de dados qualquer conjunto de dados que sejam organizados, ou seja, existe um planejamento relacionado ao seu armazenamento, confiabilidade e utilização. Alguns exemplos simples de bancos de dados são uma lista de compras ou uma lista de preços de uma loja. A organização de registros é importante para garantir que os dados estejam guardados em local seguro, que estejam corretos e coerentes e para que seu acesso seja fácil em qualquer momento conveniente.

Existem diversos métodos para o gerenciamento de dados. Os sistemas de planilhas eletrônicas são muito utilizados para organização dos dados por sua facilidade na visualização e manipulação dos dados. Tais sistemas contam também com muitas opções extras como: criação de gráficos, automatização de cálculos e o fácil e rápido aprendizado das operações básica. No entanto, não há mecanismos de segurança confiáveis, é difícil manter a consistência dos dados, redundância dos registros ocorrem com mais facilidade e são difíceis de resolver além da integração com outros sistemas de dados ser um processo difícil e trabalhoso.

Os arquivos estruturados aparecem como uma opção no gerenciamento de dados. São reconhecidos em diversos sistemas e contam com um padrão para sua escrita. Na área da Bioinformática os arquivos FASTA são utilizados para armazenar dados de sequências de aminoácidos ou nucleotídeos. Seu padrão para guardar registros obedece a seguinte regra: a primeira linha é um comentário com algumas informações sobre a sequência, começando com o símbolo $>$ e as linhas seguintes representam a sequência propriamente

dita, a Figura 2.1 mostra um exemplo deste tipo de arquivo. Outro exemplo de arquivo estruturado são os arquivos CSV. Estes tipos de arquivo formam um padrão para armazenar tabelas com as colunas sendo delimitadas por vírgula ou ponto e vírgula. Embora esses arquivos sigam padrões e sejam reconhecidos por diversos sistemas a busca por um determinado registro não é simples e demanda algum esforço manual, além de sofrerem com os mesmos problemas de segurança das planilhas.

```
>gi|21229480| DNA polymerase III beta chain
MRFTLQREAFLKPLAQVVNVVERRQTLPVLANLLVQVNNGQLSLTGTDLEVEMISRTMVEDAQDGETTIPARKLFDILRA
LPDGSRVTVSQTGDKVTVQAGRSRFTLATLPANDFPSVDEVEATERVAVPEAGLKELMERTAFAMAQQDVRYYLNGLLFD
LRDGLLRCVATDGHRLALCETELEKSGSAKRQIIVPRKGVTELLRLLEAADRDVELELGRSHIRVKRGDVTFTSKLIDGR
FPDYEAVIPIGADREVKVDREALRASLQRAAILSNEKYRGVRVEVSPGQLKISAHNPEQEEAQEEIEADTKVDDLAIGFN
VNYLLDALSALRDEHVVIQLRDANSSALVREASSEKSRHVVMPLRL
>gi|21229481| DNA replication and repair RecF protein
MHVARLSIHRLRRFEAVEFHPASTLNLLTGDNGAGKTSVLEALHVMAYGRSFRGRVRDGLIRQGGQDLEIFVEWRERAGD
STERTRRAGLRHSGQEWTGRLDGEDVAQLGSLCAALAVVTFEPGSHVLISGGGEPRRRFLDWGLFHVEPDFLALWRRYAR
ALKQRNALLKQGAQPQMLDAWDHELAESGETLTSRRLQYLERLQERLVPVATAIAPSLGLSALTFAPGWRRHEVSLADAL
LLARERDRQNGYTSQGPHRADWAPLFDALPGKDALSRGQAKLTALACLLAQAEDFAHERGEWPIMALDDLGSELDRHHQA
RVIQRLASAPAQVLITATELPPGLADAGKTLRRFHVEHGQLVPTTAAD
>gi|21229482| DNA gyrase subunit B
MTDEQTTPPTPNGTYDSSKITVLRGLEAVRKRPGMYIGDVHDGTGLHHMVFEVVDNSVDEALAGHADDIVVKIHVDGSVA
VSDNGRGVPVDIHKEEGVSAAEVILTVLHAGGKFDDNSYKVSGGLHGVGVSVVNALSEHLWLDIWRDGFHYQQEYALGEP
QYPLKQLEASTKRGTTLRFKPAVEIFSDVEFHYDILARRLRELSFLNSGVKIALIDERGEGRRDDFHYEGGIRSFVEHLA
QLKTPLHPNVISVTGEHNGIVVDVALQWTDAYQETMYCFTNNIPQKDGGTHLAGFRGALTRVLSNYIEQNGIAKQAKITL
TGDDMREGMIAVLSVKVPDPSFSSQTKEKLVSSDVRPAVENAFGARLQEFLQENPNEAKAITGKIVDAARAREAARKARD
LTRRKGALDIAGLPGKLADCQEKDPALSELFIVEGDSAGGSAKQGRNRKNQAVLPLRGKILNVERARFDRMLASDQVGTL
ITALGTGIGRDEYNPDKLRYHRIILMTDADVDGSHIRTLLLTFFYRQMPELIERGYIYIGLPPLYKLKQGKSELYLKDDA
ALNAYLASSAVEGAALIPASDEPPITGEALEKLLLLFAGAKEAIARNAHRYDPALLTALIDLPPLDVVQLQAEGDVHPTL
DALQAVLNRGTLGTARYHLRFDPATDSAAASLVSVRKHMGEEFTQVLPMGAFESGELRPLREVALALHGLVREGAQILRG
NKSHPITSFAQAQAWLLEEAKRGRQVQRFKGLGEMNAEQLWETTVNPDTRRLLQVRIEDAVAADQIFSTLMGDVVEPRRD
FIEDNALKVSNLDI
>gi|21229483| conserved hypothetical protein
MTMSAVLPPSPAPVSVPGPPSLRSAVLGFCIDLLIAIGLLLLLSVAGFAVWGFLRSMGEVQAVRAQGGSPSPAAIMAAIG
QPGVMVQLLIALVSTATPAVLLYVWRRRATPAEQATSRAAIRRRSTWGWIAAVAAGVFMLSNLVSVLASALGIKPVPTNL
PLMEEAIKQWPLALVVFAVAIAPAYEELLFRRVLFGRLLAAGRPWLGVVLSSLTFALVHEVPGISGNGVVAIAQLWLVYG
GMGAAFAWLYWRTGTLWAPILAHGINNATALAALYFFGLG
```

Figura 2.1: Exemplo do conteúdo de um arquivo FASTA.

Os bancos de dados orientados a documentos estão presentes em aplicações voltadas em geral para a Internet e utilizam seu próprio tipo de arquivo estruturado. Essa alternativa oferece facilidade de uso e os gerenciadores que utilizam esse modelo apresentam melhor desempenho em operações como consultas simples, inserções e deleção de dados (BOICEA; RADULESCU; AGAPIN, 2012).

Um dos meios mais populares para a manipulação de dados atualmente é o banco de dados relacional que consiste em um sistema capaz de armazenar dados para fins de consultas e alterações (DATE, 2004). Isto é, as informações inseridas no banco de dados podem ser alteradas e consultadas por seus usuários em qualquer momento, sempre que

houver necessidade. Esse método cobre a maior parte das implementações que existem no mercado atualmente, contando com muitas ferramentas para seu uso integrado. Utilizando esse modo de gerenciar dados pode-se eliminar muitos problemas descritos nas outras abordagens e ter vários benefícios. As principais vantagens são: maior segurança sobre os dados; consistência dos dados; a possibilidade de reduzir redundância durante a fase de projeto; e a maior interoperabilidade com outras plataformas e aplicações.

Os bancos de dados relacionais oferecem controle de acesso em nível de usuário ou grupos de usuários sendo flexível a configuração sobre as permissões. Os dados são armazenados em estruturas especiais que impedem a visualização direta de seus arquivos, como em arquivos texto comum. Por seu armazenamento ser em arquivos binários, além da segurança, promove otimização nas consultas realizadas, oferecendo um melhor desempenho. A grande maioria dos gerenciadores de banco de dados adota a linguagem SQL para manipulação dos dados, o que torna fácil a migração entre plataformas gerenciadoras dos dados e facilita a integração com outros bancos de dados relacionais.

Durante a fase de modelagem de dados, para demonstrar aos usuários como os dados serão estruturados sem explicitar como será a implementação dos mesmos, é utilizado um modelo conceitual de alto nível, chamado de modelo entidade-relacionamento. Nesse modelo tem-se a entidade como objeto básico para representar de forma abstrata algo que exista no mundo real, fisicamente ou conceitualmente. Cada entidade conta com características que a definem, as quais são chamadas de atributos. Assim, um registro no banco de dados nesse modelo seria os valores designados aos atributos de uma entidade. Esse modelo demonstra as relações que diferentes entidades têm entre si, abstraindo as relações existentes no domínio dos dados representados. A Figura 2.2 apresenta um exemplo de diagrama de entidade e relacionamento aplicado a um banco de funcionários de uma empresa. Nesta figura, o losango representa o relacionamento entre duas entidades, representadas pelos retângulos, e os atributos das entidades são representados pelas elipses. Nesse diagrama tem-se que um empregado é lotado em um departamento e um departamento pode ter vários empregados, esta relação é representada pelo número 1 e a letra N nas retas que ligam o relacionamento às entidades.

O modelo relacional conta com uma estrutura de dados uniforme, onde as enti-

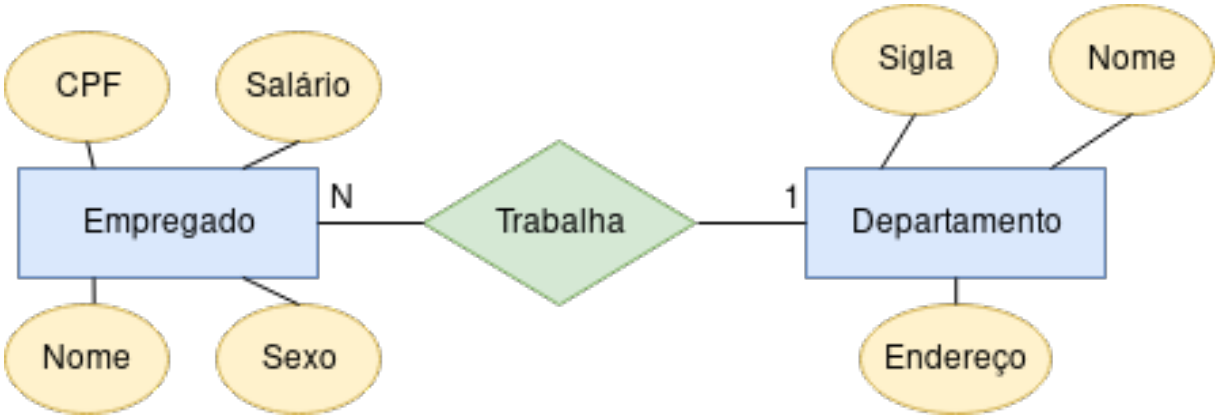

Figura 2.2: Exemplo de um modelo de entidade-relacionamento

dades e suas relações se transformam em tabelas ou arquivos de registros. Cada coluna dessas tabelas representa um atributo, assim uma linha é um registro com os valores dos atributos. Essa simplicidade na representação dos dados e sua formalidade foram os principais motivos para o modelo relacional ser o mais difundido e utilizado atualmente.

Para facilitar e implementação destes bancos de dados foram criados os sistemas gerenciadores de bancos de dados (SGBD). Eles tem a finalidade de prover uma interface simples para a manipulação dos registros armazenados nos bancos de dados. Além disso, alguns SGBDs também oferecem métodos para exportação e importação de dados. Os SGBDs buscam facilitar as operações mais comuns como: consulta, inserção e deleção de registros. Ao utilizar estes sistemas, o usuário não precisa ter um conhecimento aprofundado na linguagem SQL, sendo que muitas operações podem ser feitas pela interface gráfica.

2.1.1 Bancos de Dados Biológicos

A geração em massa de dados vêm acontecendo há algum tempo, Bell, Hey e Szalay (2009) utilizam o termo “dilúvio de dados” para que seja compreendido o cenário que têm se estabelecido em diversos ramos de estudo, não somente na grande área da informática. E para conseguir dar sentido a esses dados, ou seja, transformar o dado em uma informação útil para determinada atividade de pesquisa ou conhecimento, precisamos manipulá-los corretamente, de maneira que possamos aproveitar todo seu potencial.

Na área da Bioinformática, o cenário se repete, há uma grande e crescente quantidade de dados gerados, em especial pelos projetos de sequenciamento. Com a capacidade

de processamento sempre aumentando o resultado é a produção de dados em escala cada vez maior. Assim o desenvolvimento de novos métodos para a análise de dados biológicos se fez necessário, bem como meios para armazenamento e manipulação destes dados, para facilitar consultas e modificações. Desta forma, bancos de dados são muito empregados e desempenham um papel importante no auxílio de pesquisas biológicas como uma fonte de recursos (ZOU et al., 2015).

Os bancos de dados biológicos são caracterizados por seu conteúdo de caráter biológico, podem conter informações sobre sequências de DNA, proteínas, informações relacionadas a doenças, histórico médico, etc. Alguns bancos de dados não armazenam somente essas informações, eles buscam oferecer o maior número de informações sobre determinados conjuntos de dados, fornecendo informações bibliográficas sobre as sequências e anotações biológicas, por exemplo. Esses bancos de dados são importantes e diversas pesquisas são realizadas com base nas informações disponibilizadas nesses bancos de dados.

O maior objetivo dos bancos de dados biológicos é oferecer um sistema que facilite o armazenamento de um grande volume de dados além de prover uma estrutura que consiga organizar e recuperar informações em qualquer instante que o usuário necessite. Além disso, esses bancos de dados buscam prover APIs para que outras bases de dados possam realizar a troca e integração de informações de maneira automatizada (ZOU et al., 2015).

Segundo dados de Rigden, Fernández-Suárez e Galperin (2015) os recursos disponibilizados pelos três maiores centros de bioinformática mundial, Centro Nacional para Informação de Biotecnologia dos Estados Unidos (NCBI[1]), Instituto Europeu de Bioinformática (EMBL-EBI[2]) e Instituto Suiço para Bioinformática (SIB[3]), somam ao todo 1685 bancos de dados de caráter biológico ao final de 2015. Esses dados são atualizados constantemente para proporcionar fontes de dados mais seguras e confiáveis aos pesquisadores.

Os bancos de dados biológicos são desenvolvidos com propósitos diferentes entre

[1]https://www.ncbi.nlm.nih.gov/
[2]https://www.ebi.ac.uk/
[3]https://www.sib.swiss/

si e mesmo aqueles que tenham a mesma finalidade diferem dos outros por alguns aspectos de implementação. Outras diferenças podem ocorrer, desde a maneira que os dados foram obtidos até a definição de qual estrutura de dados será utilizada para o armazenamento e recuperação destes dados.

Zou et al. (2015) fala sobre essas diferenças entre os milhares de bancos de dados existentes. Muitos são destinados a oferecer dados relacionados a assuntos específicos como um grupo de organismos que causam determinada doença ou ainda que armazenem somente sequências de proteínas de uma determinada espécie ou gênero. Por outro lado, existem aqueles que não possuem um propósito específico claro, podendo conter sequências tanto de proteínas quanto de nucleotídeos e de vários organismos não necessariamente relacionados. Segundo dados de (**??**) publicados na revista *Nucleic Acids Research* (NAR), até a submissão da última edição em Janeiro deste ano, foram relatados 54 novos bancos de dados e a atualização de outros 98. Devido ao grande número de bancos de dados a revista NAR criou uma classificação de acordo com a linha de pesquisa do banco: (i) armazenar todas as sequências; (ii) sequências genômicas; (iii) estruturas e sequências de proteínas; (iv) vias metabólicas; e (v) com finalidade específica. O LeifDB encaixa-se na última classificação por se tratar de um banco de dados que tem como objetivo tratar informações relacionadas a bactéria *Leifsonia xyli*. A seguir há alguns exemplos dos bancos de dados biológicos públicos mais conhecidos:

- GenBank (BILOFSKY et al., 1986): da categoria (i) é um dos maiores bancos de dados públicos existentes. Contém todas as sequências de DNA públicas anotadas. Além de seus próprios dados conta com informações vindas do European Nucleotide Archive(ENA) e do DAN DataBank of Japan(DDBJ). Uma nova versão é lançada a cada dois meses com atualizações. Seu banco de dados conta com quase 260 mil espécies formalmente descritas sendo sua maioria de submissões proeminentes de laboratórios isolados e projetos para sequenciamento de espécies em larga escala.

- FlyBase (GRAMATES et al., 2017): da categoria (v) é o principal banco de dados que contém informações genéticas e genômicas sobre a família *Drosophilidae* que abrange moscas geralmente pequenas. Esta base de dados permite realizar buscas sobre DNA e sequências de proteínas além de contar com ontologias para realizar

buscas de dados funcionais. É mantido por um grupo de pesquisadores das Universidades de Havard (USA), Cambridge(UK), Indiana(USA) e New Mexico(USA).

- Omnione: da categoria (ii) é um banco de dados associado ao sistema CMR (*Comprehensive Microbial Resource*) que disponibiliza algumas informações sobre DNA, propriedades químicas das proteínas, grupo taxonômico a que o organismo pertence e ligações com outras fontes de dados públicas (PETERSON et al., 2001).

- PBD: da categoria (iii), neste banco de dados estão contidas informações sobre proteínas de diferentes organismos, com direcionamento em suas estruturas tridimensionais. Ele ainda oferece informações gerais e específicas assim como os métodos utilizados para determinação das mesmas (BERMAN et al., 2000).

2.2 A bactéria *Leifsonia xyli*

O gênero *Leifsonia* abrange sete subespécies de bactérias, que são encontradas em diferentes nichos, tais como: em plantas, no solo e em ambiente aquático (Tabela 2.1). Entre estas bactérias se destacam a *Leifsonia xyli* subsp. *xyli* (Lxx) e *Leifsonia xyli* subsp. *cynodontis* (Lxc). Estas bactérias são os agentes causadores de doenças em plantas, sendo consideradas bactérias fitopatogênicas.

A bactéria Lxx é restrita ao xilema, mede 0,25-0,50μm por 1-4μm, é corineforme[4], reta e ligeiramente curva, aeróbia, imóvel e Gram-positiva[5], apresentando colônias circulares, é extremamente fastidiosa em seus requerimentos nutricionais e de difícil detecção (DIAS et al., 2016). A Lxx é uma bactéria endofítica (habitam o interior de plantas) obrigatória da cana-de-açúcar (MONTEIRO-VITORELLO et al., 2004). Seu genoma possui 2.4 Mb e seu conteúdo de G+C é 67,7 mol % de conteúdo de GC (bases nitrogenadas Guianina e Citosina) e 307 pseudogenes[6], que sugere um processo de decaimento genético (MONTEIRO-VITORELLO et al., 2004). Por outro lado, quando comparado com outras bactérias fitopatogênicas a Lxx possui um quantidade baixa de genes relacionados com a

[4]pode causar infecções oportunistas no hospedeiro imunocomprometido

[5]bactérias com parede celular menos resistente a substâncias que tentem entrar em seu interior, oposto às Gram-negativas

[6]sequência genômica similar a um gene mas que não se expressa, ou seja, não exerce uma função

Tabela 2.1: Bactérias do gênero *Leifsonia*.

Leifsonia antarctica
Leifsonia aquatica
Leifsonia aquatica ATCC 14665
Leifsonia aquatica H1aii
Leifsonia bigeumensis
Leifsonia kafniensis
Leifsonia lichenia
Leifsonia naganoensis
Leifsonia pindariensis
Leifsonia poae
Leifsonia psychrotolerans
Leifsonia rubra
Leifsonia rubra CMS 76R
Leifsonia shinshuensis
Leifsonia soli
Leifsonia xyli
Leifsonia xyli subsp. cynodontis
Leifsonia xyli subsp. cynodontis DSM 46306
Leifsonia xyli subsp. xyli
Leifsonia xyli subsp. xyli str. CTCB07
Candidatus Leifsonia sp. Metals-1
Candidatus Leifsonia sp. Metals-2
Leifsonia sp.

Fonte: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=110932

patogenicidade (DIAS et al., 2016).

O sequenciamento da bactéria Lxx trouxe importantes informações a cerca de genes e enzimas que podem estar envolvidos na síntese do processo de inibição do crescimento em plantas. Tais genes poderiam ser responsáveis pelo principal sintoma da doença, o raquitismo (DIAS et al., 2016).

As doenças que acometem a plantação de cana-de-açúcar representam um dos fatores responsáveis pelo decréscimo da sua produtividade em todo o mundo, e o raquitismo-da-soqueira, causado pela Lxx é uma das mais importantes, podendo ocasionar perdas significativas, estimadas em até 21% (MARCUZ et al., 2009). Tal doença afeta o crescimento do caule da planta, diminuindo consideravelmente sua produção. Atualmente não há cura para o raquitismo que afeta a cana-de-açúcar. No entanto, existem métodos alternativos para seu controle, como por exemplo, o tratamento térmico (JUNIOR, 2006).

A anotação dos genes desta bactéria em nível de nucleotídeos foi feita por Monteiro-Vitorello et al. (2004) e consistiu na identificação onde cada gene se encontra na sequência genômica completa da bactéria. Uma vez que a localização dos genes foi realizada, o próximo passo é identificar quais as funções que esses genes desempenham no organismo. Ou seja, fazer a anotação funcional das proteínas associadas a estes genes e, com isso identificar o que causa a patogenicidade da bactéria. Esta informação pode ser usada para desenvolver métodos de controle destas pragas.

A bactéria Lxc assim como a Lxx também é capaz de colonizar o xilema da cana-de-açúcar, sem, no entanto, causar doença. Esta bactéria possui alta relação filogenética com a Lxx, embora apresentem diferenças quanto à doença e hospedeiro. A Lxc causa raquitismo na *Cynodon dactylon*, uma planta daninha que se espalha por plantações de milho, arroz e cana-de-açúcar, também conhecida como: capim de burro, capim de bermuda ou grama seda.

A anotação de seus genes seguiu a anotação da Lxx e foi feita por Monteiro-Vitorello et al. (2013). Mesmo a Lxc prejudicando uma planta daninha, seu estudo é interessante pelas duas bactérias serem da mesma espécie, *Leifsonia xyli*.

2.3 Ferramentas computacionais

Nesta seção serão apresentadas as ferramentas computacionais utilizadas para a construção do banco de dados para as bactérias da espécie *Leifsonia xyli* (LeifDB [*Leifsonia* Database]) quanto ao tratamento dos dados. Primeiramente são expostas as ferramentas para alinhamento de sequências, em seguida são abordados os métodos para o agrupamento destas sequências e por fim o método de categorização funcional escolhido para categorizar os genes dos organismos armazenados no LeifDB.

2.3.1 Alinhamento de sequências

Um dos métodos mais utilizados atualmente para dar início ao processo de anotação de genes de um organismo é por meio do alinhamento de sequências (Figura 2.3). O alinhamento de sequências é uma técnica que organiza sequências de DNA, RNA ou proteínas,

a fim de identificar regiões similares entre elas. Ocorre a comparação par a par de uma sequência requerida com um conjunto de sequências sobre o qual há informações já consolidadas. Esse conjunto de sequências geralmente estão armazenadas em banco de dados disponibilizados por grandes instituições que se comprometem em realizar a manutenção desses dados com constantes atualizações sobre novas sequências e atualizações sobre as sequências já existentes. O maior sistema de informação atualmente é o NCBI (*National Center for Biotechnology Information*), que mantém vários bancos de dados com finalidades distintas, integram dados de diferentes fontes e disponibilizam ferramentas computacionais que auxiliam no processo de busca e análise. Este sistema de informação é mantido pelo governo dos Estados Unidos.

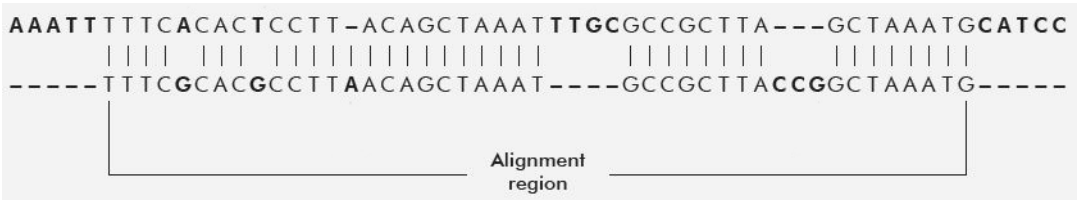

Figura 2.3: Exemplo de alinhamento entre duas sequências.

Ferramentas computacionais são necessárias para a realização de alinhamento de sequências. Como já mencionado, o alinhamento é feito par a par e a comparação é feita com banco de dados a fim de se encontrar a melhor combinação possível. No entanto, o problema está no tamanho dos bancos de dados. Eles armazenam milhões de registros de sequências e a realização de uma comparação entre as sequências e os bancos de dados demanda muito tempo se for feita manualmente.

Uma das ferramentas mais conhecidas, capaz de realizar alinhamento de sequências de nucleotídeos e aminoácidos, é o algoritmo BLAST, *Basic Local Alignment Search Tool*, desenvolvido por Altschul et al. (1990). O BLAST é um algoritmo heurístico[7] que realiza o alinhamento procurando um número pequeno de caracteres similar entre as sequências e aos poucos aumenta a quantidade de caracteres até encontrar um limite.

O NCBI oferece uma plataforma na web para executar o BLAST, o qual permite consultar todos os bancos de dados lá armazenados, sendo possível realizar alterações nos parâmetros do algoritmo, como por exemplo, aumentar o tamanho da *query* inicial e o

[7]algoritmo que não garante encontrar uma solução ótima mas obtém bons resultados em um período de tempo aceitável

número máximo de sequências alinhadas que serão retornadas ao usuário. Este programa conta também com diferentes tipos de alinhamento como apresentado na Tabela 2.2.

Tabela 2.2: Diferentes tipos de BLAST

Programa	Sequência de Entrada	Banco de Dados	Formato da sequência comparada
BLASTn	Nucleotídeos	Nucleotídeos	Nucleotídeos
BLASTp	Proteínas	Proteínas	Proteínas
BLASTx	Nucleotídeos	Proteínas	Proteínas
tBLASTn	Proteínas	Nucleotídeos	Proteínas
tBLASTtx	Nucleotídeos	Nucleotídeos	Proteínas

Outro modo de executar o BLAST sem depender da plataforma do NCBI é executando o programa localmente sem acesso a Internet. O NCBI disponibiliza um pacote do programa.

2.3.2 Métodos de agrupamento

No processo de anotação dos genes de um organismo os conceitos de genes ortólogos e parálogos ajudam na identificação de suas funções. Ortólogos são genes que possuem ancestral em comum mas divergiram durante a evolução por especiação. Ou seja, originaram espécies diferentes. Os parálogos são genes, geralmente da mesma espécie mas que foram originados por um evento de duplicação gênica. A Figura 2.4 apresenta um exemplo de genes parálogos e ortólogos, os pares de genes 1a do organismo 1 e 1a do organismo 2 são ortólogos, enquanto os genes 1a e 1b são parálogos. Esses conceitos tem origem no campo da sistemática molecular de Fitch (1970). Por causa de sua ancestralidade os genes ortólogos tendem a ter funções parecidas ou exatamente iguais e os parálogos, funções que se diferenciam. Tomando este conceitos, genes anotados, aqueles que possuem suas funções definidas, e não anotados podem ser submetidos a formações de grupos com base nas funções conhecidas de genes que tem funcionalidades conhecidas e bem caracterizadas e assim os outros poderão ter suas funções reconhecidas ao serem agrupados com os demais.

O algoritmo orthoMCL (LI; STOECKERT; ROOS, 2003) utiliza um conjunto de ferramentas a fim fornecer um método automatizado para a identificação dos grupos de ortólogos presentes em um conjunto inicial de sequências. Ele executa um BLAST de proteínas (BLASTp) de todas as sequências contra todas e utiliza o resultado para separar

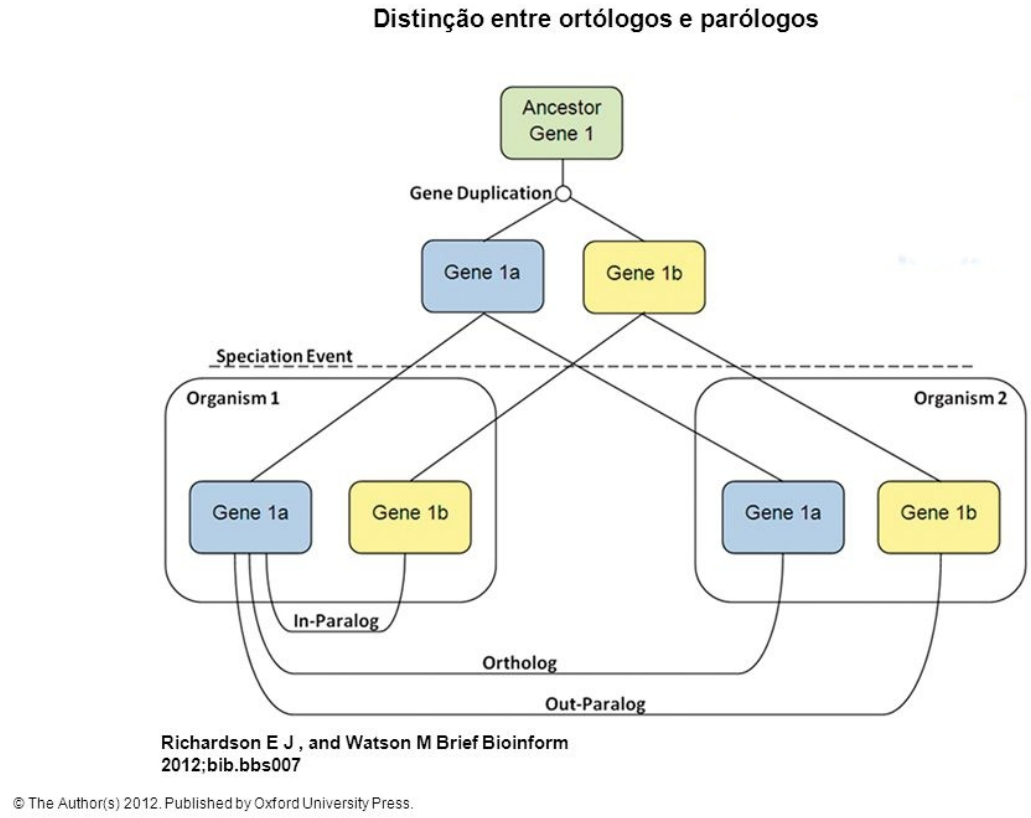

Figura 2.4: Exemplo de genes parálogos e ortólogos.

em pares os genes, atribuindo um valor para cada par referente ao nível de similaridade[8]. Em seguida, esses pares são fornecidos ao algoritmo *Markov Cluster* (MCL) (DONGEN, 2001). Este algoritmo utiliza probabilidade e teoria dos grafos para formar os melhores grupos com base em sua similaridade. Ele simula caminhadas aleatórias em um grafo utilizando matrizes de Markov para determinar a probabilidade de transição entre os nós (LI; STOECKERT; ROOS, 2003). Enright, Dongen e Ouzounis (2002) demonstraram a eficiência do algoritmo de Markov em termos de desempenho e confiabilidade ao executá-lo em um grande conjunto de proteínas. Após este passo, o orthoMCL gera grupos de proteínas onde cada grupo é formado por ortólogos ou parálogos que sejam de no mínimo duas espécies. A Figura 2.5 apresenta o fluxo de execução do programa orthoMCL.

[8]similaridade entre duas sequências é quando elas se parecem em nível paptídico ou nucleico

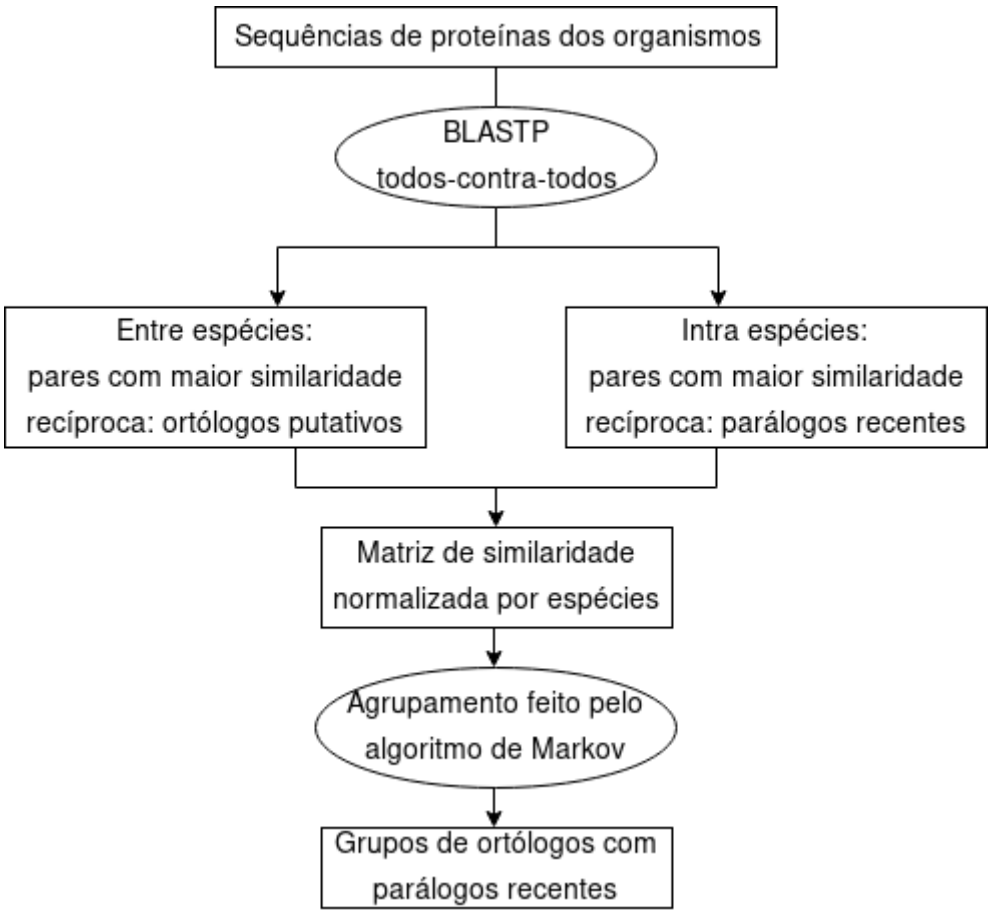

Figura 2.5: Fluxo do algoritmo OrthoMCL adaptado de Li, Stoeckert e Roos (2003).

2.3.3 Categorização Funcional dos Genes

Uma categoria funcional, ou classe funcional, se refere ao propósito do gene no organismo, abrangendo processo biológico, função molecular ou componente celular (ALMEIDA, 2007). Então ao atribuir categorias aos genes ocorre a categorização funcional. Quando o agrupamento dos genes é realizado, os grupos que apresentarem pelo menos um gene devidamente categorizado pode ter seus membros categorizados na mesma categoria funcional, estendendo assim a categorização individual para o grupo.

Almeida (2007) utilizou as categorias funcionais definidas para o projeto genoma das bactérias das espécies *Xanthomonas axonopodis* pv. *citri* e *Xanhomonas campestris* pv. *campestris* (SILVA et al., 2002) para classificar os genes que compuseram seu banco de dados. De forma automatizada, os genes foram categorizados herdando a classificação funcional dos genes do genoma das *Xanthomonas* que estivessem no mesmo grupo. Para o caso de grupos sem genes classificados, seus membros, consequentemente continuaram sem atribuição da categorização funcional.

A escolha para utilizar a classificação de acordo com as bactérias do gênero *Xanthomonas* se deve ao genoma destes organismos terem sido manualmente curados, aumentando a confiabilidade dos dados, sua abrangência na classificação e por ser uma bactéria associada a plantas com mais de cinco mil genes categorizados incluindo categorias ligadas a virulência, patogenicidade e adaptação ao hospedeiro. A Tabela 2.3 apresenta todas as categorias do projeto *Xanhomonas*.

Tabela 2.3: Categorização funcional de bactérias do gênero *Xanthomonas*

I. Intermediary metabolism
A. Degradation
1. Degradation of polysaccharides and oligosaccharides 2. Degradation of small molecules 3. Degradation of lipids
B. Central Intermediary metabolism
1. Amino sugars 2. Entner-Douderogg 3. Gluconeogenesis 4. Glyoxylate 5. Miscellaneous glucose metabolism 6. Non-oxidative branch, pentose pathway 7. Nucleotide hydrolysis 8. Nucleotide interconversions 9. Phosphorus compounds 10. Pool, multipurpose conversions 11. Sugar-nucleotide biosynthesis, conversions 12. Sulfur metabolism
C. Energy metabolism
1. Aerobic respiration 2. Anaerobic respiration and fermentation

3. Electron transport
4. Glycolysis
5. Oxidative branch, pentose pathway
6. Pyruvate dehydrogenase
7. TCA cylce
8. ATP-proton motive force interconversion
D. Regulatory functions
1. Two component system
2. Activators-Repressors
3. Kinases-Phosphatases
4. Sigma factors and other regulatory components
5. Not used
II. Biosynthesis of small molecules
A. Amino acids biosynthesis
1. Glutamate family/nitrogen assimilation
2. Aspartate family, pyruvate family
3. Glycine-serine family/sulfur metabolism
4. Aromatic amino acid family
5. Histidine
B. Nucleotides Biosynthesis
1. Purine ribonucleotides
2. Pyrimidine ribonucleotides
3. 2'-Deoxyribonucleotides
4. Salvage of nucleosides and nucleotides
C. Sugars and sugar nucleotides biosynthesis
D. Cofactors, prosthetic groups, carries biosynthesis
1. Biotin
2. Folic acid
3. Lipoate
4. Molybdopterin

5. Pantothenate 6. Pyridoxine 7. Pyridine nucleotides 8. Thiamin 9. Riboflavin 10. Thioredoxin, glutaredoxin, glutathione 11. Menaquinone, ubiquinone 12. Heme, porphyrin 13. Biontin corboxyl carrier protein (BCCP) 14. Cobalamin 15. Enterochelin 16. Biopterin 17. Others
E. Fatty acid and phosphatidic acid biosynthesis
F. Polyamines biosynthesis
III. Macromolecule metabolism
A. DNA metabolism
1. Replication 2. Structural DNA binding proteins 3. Recombination 4. Repair 5. Restriction, modification
B. RNA metabolism
1. Ribosomal and stable RNAs 2. Ribosomal proteins 3. Ribosomes - maturation and modification 4. Aminoacyl tRNA synthetases, tRNA modification 5. RNA synthesis, modification, DNA transcription 6. RNA degradation
C. Protein metabolism

1. Translation and modification
2. Chaperones
3. Protein degradation
D. Other macromolecules metabolism
1. Polysaccharides
2. Phospholipids
3. Lipoprotein
IV. Cell structure
A. Membrane components
1. Inner membrane
2. Outer membrane constituents
B. Murein sacculus, peptidoglycan
C. Surface polysaccharides, lipopolysaccharides, and antigens
D. Surface structures
V. Cellular processes
A. Transport
1. Amino acids, amines
2. Anions
3. Carbohydrates, organic acids, alcohols
4. Cations
5. Nucleosides, purines, pyrimidines
6. Protein, peptide secretion
7. Other
B. Cell division
C. Chemotaxis and mobility
D. Osmotic adaptation
E. Cell killing
VI. Mobile genetic elements
A. Phage-related functions and prophages

B. Plasmid-related functions
C. Transposon- and intron-related functions
VII. Pathogenicity, virulence and adaptation
A. Avirulence
B. Hypersensitive response and pathogenicity
C. Toxin production and detoxification
D. Host cell wall degradation
E. Exopolysaccharides
F. Surface proteins
G. Adaptation, atypical conditions
H. Other
VIII. Hypothetical
A. Conserved hypothetical proteins
B. Hypothetical proteins (includes no hits or only low score hits)
C. Xanthomonas conserved hypothetical
IX. ORFs with undefined category

O COG (TATUSOV; KOONIN; LIPMAN, 1997) é um banco de dados que armazena as informaçes sobre grupos de genes de organismos unicelulares baseando-se nos conceitos de parálogos e ortólogos. Assim um método automático para a divisão dos genes foi criado, para preencher esse banco de dados, onde um grupo de ortólogos formado apresentasse no mínimo três organismos diferentes. Criado originalmente em 1997, o COG recebeu duas atualizações em seu banco de dados, uma em 2003 (TATUSOV et al., 2003) e a mais recente que aconteceu em 2014 (GALPERIN et al., 2014). Hoje, seu banco de dados conta com 711 organismos e 4631 grupos de ortólogos classificados entre as 26 categorias estabelecidas, cada uma recebendo uma letra do alfabeto como forma de identificação, a Tabela 2.4 apresenta essas categorias.

Assim, o banco de dados COG e a classificação funcional de acordo com o projeto *Xanthomonas* são formas adequadas para a categorização funcional dos genes da *Leifsonia*. O banco COG possui um banco de dados que abrange muitos organismos, o que facilita

Tabela 2.4: Categorias do COG

	Information storage and processing
A	RNA processing and modification
B	Chromatin structure and dynamics
J	Translation, ribosomal structure and biogenesis
K	Transcription
L	Replication, recombination and repair
	Cellular processes
D	Cell cycle control, cell division, chromosome partitioning
M	Cell wall/membrane/envelope biogenesis
N	Cell motility
O	Posttranslational modification, protein turnover, chaperones
P	Inorganic ion transport and metabolism
T	Signal transduction mechanisms
U	Intracellular trafficking, secretion, and vesicular transport
V	Defense mechanisms
X	Mobilome: prophages, transposons
W	Extracellular structures
Y	Nuclear structure
Z	Cytoskeleton
	Metabolism
C	Energy production and conversion
G	Carbohydrate transport and metabolism
E	Amino acid transport and metabolism
F	Nucleotide transport and metabolism
H	Coenzyme transport and metabolism
I	Lipid transport and metabolism
Q	Secondary metabolites biosynthesis, transport and catabolism
	Poorly characterized
R	General function prediction only
S	Function unknown

a identificação e o agrupamento de sequências similares, e o projeto *Xanthomonas* trata de um organismo patogênico, manualmente curado e ligado a plantas.

3 Metodologia

Neste capítulo será apresentado o banco de dados LeifDB e como ele foi estruturado para comportar as informações armazenadas. Há também a apresentação de outros organismos incluídos no banco de dados além da *Leifsonia xyli*, logo em seguida, a categorização funcional dos genes segundo as categorias do COG e das *Xanthomonas* e por fim uma idealização da interface gráfica num ambiente web para acesso e consulta ao LeifDB.

3.1 Organismos no LeifDB

O LeifDB contém informações sobre as sequências das duas subespécies conhecidas de *Leifsonia xyli*. Além disso, estão presentes também bactérias do gênero *Clavibacter* totalizando onze organismos no banco.

As bactérias *Clavibacter* presentes no banco de dados são patogênicas e se hospedam em plantas causando doenças graves que levam a morte do seu hospedeiro. Elas foram escolhidas para integrarem o LeifDB devido à semelhança com as *Leifsonia* e por terem alguns de seus genes categorizados, além de também serem completamente sequenciadas. Evtushenko et al. (2000) mostra a renomeação da *Leifsonia xyli* para a nomenclatura atual, uma vez que eram classificadas como *Clavibacter*.

Na Tabela 3.1 se encontram os organismos incluídos no banco que são patogênicos, ou seja, que causam algum malefício ao seu hospedeiro. Esta tabela também mostra a doença causada pela bactéria assim como seu hospedeiro e a sua referência bibliográfica.

3.2 Estrutura do Banco de Dados

O sistema gerenciador de banco de dados escolhido para implementar o LeifDB foi o MariaDB (BARTHOLOMEW, 2015), um banco de dados criado a partir de uma implementação do MySQL, um dos mais famosos bancos de dados gratuitos. Este SGBD foi escolhido por apresentar simplicidade de uso, ter sua licença *open source* e não exigir

Tabela 3.1: Bactérias Disponibilizadas no Banco.

Bactéria	Hospedeiro	Doença causada	Referência
Leifsonia xyli subsp. xyli str. CTCB07	Cana-de-açúcar	Raquitismo da soqueira	Monteiro-Vitorello et al. (2004)
Leifsonia xyli subsp. cynodontis	Grama	Raquitismo	Monteiro-Vitorello et al. (2013)
Clavibacter michiganensis subsp. sepedonicus	Batata	Podridão anelar	Bentley et al. (2008)
Clavibacter michiganensis subsp. michiganensis NCPPB 382	Tomate	Cancro-bacteriano	Gartemann et al. (2008)
Clavibacter michiganensis subsp. nebraskensis NCPPB 2581	Milho	Seca de Goss	Gartemann et al. (2011)
Clavibacter michiganensis CF11	Tomate	Cancro-bacteriano	Du et al. (2015)
Clavibacter cf. michiganensis LMG 26808	Tomate	Cancro-bacteriano	Załuga et al. (2014)

muitos recursos de hardware.

Pensando no modelo relacional, foram criadas tabelas e seus atributos para as principais abstrações do domínio que seriam armazenadas. O diagrama entidade relacionamento está presente na Figura 3.1. A principal tabela é a que armazena as informações sobre os genes, chamada de *Gene*. Ela guarda informações sobre a posição de um gene no genoma, sua anotação, além da sequência de aminoácidos, a qual espécie ele pertence e dois identificadores, são eles: um identificador único do LeifDB e um identificador gerado pelo software Prokka (SEEMANN, 2014). O software Prokka foi utilizado para a anotação do genoma, ele separa os genes atribuindo identificadores únicos a eles, sua saída é um arquivo FASTA com todos os genes de uma espécie que é utilizado como entrada para os programas: BLAST e orthoMCL.

São armazenadas também as informações de todos os grupos formados pelo orthoMCL com os genes de *Leifsonia* e *Clavibacter*, como: o número de membros, a maior sequência presente, dois identificadores, um do LeifDB e outro do orthoMCL e outras informações de interesse. Como um gene pode ser classificado em mais de um grupo e um grupo pode ter mais de um gene foi modelado uma relação de muitos para muitos

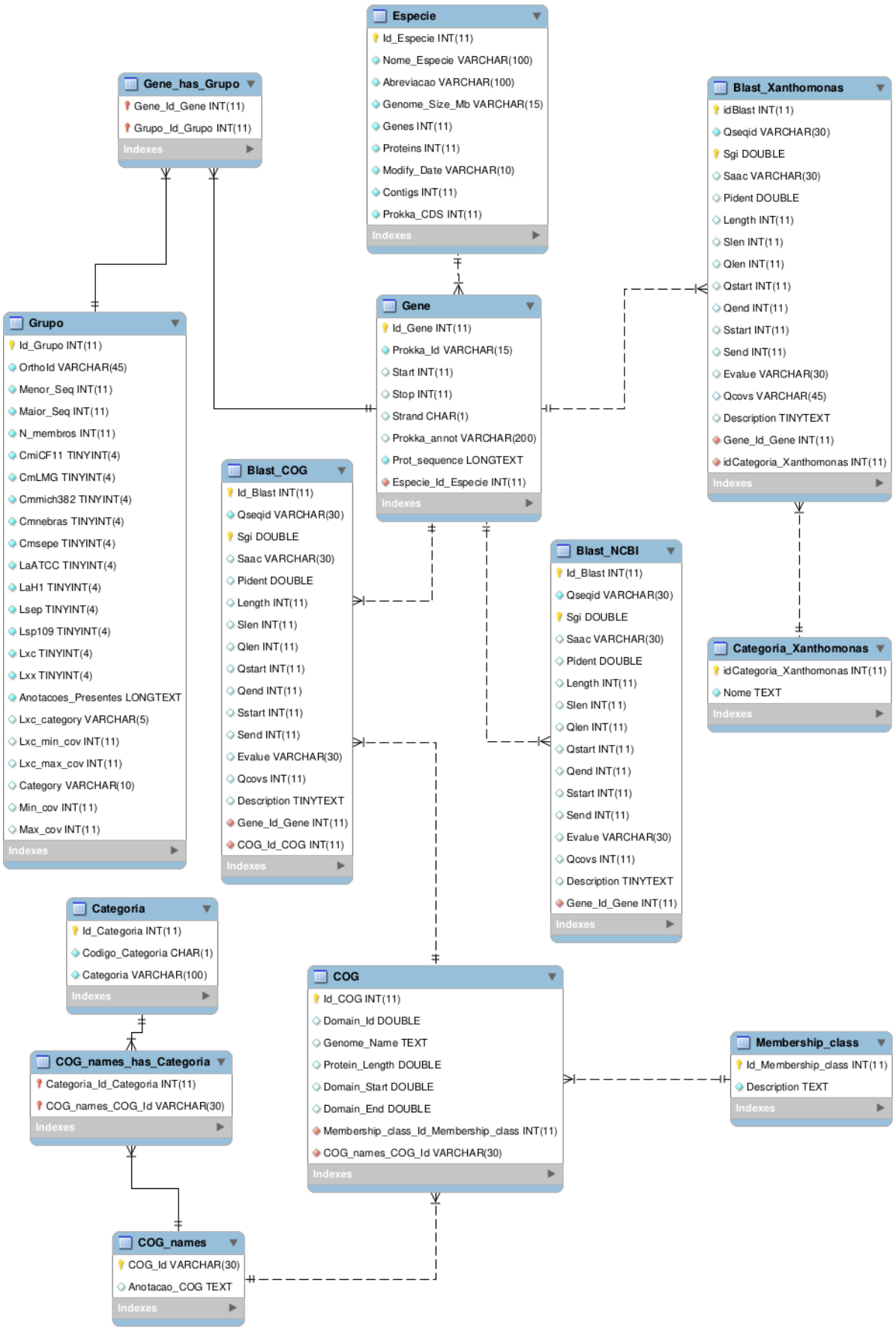

Figura 3.1: Diagrama Entidade-Relacionamento do LeifDB.

entre as tabelas de gene e grupo.

Foram criadas três tabelas separadas para armazenar os dados do BLAST dos genes contra o COG, as *Xanthomonas* e o banco de dados do NCBI. Essas tabelas armazenam qual foi o melhor resultado do BLAST e através delas podemos obter a categorização funcional dos genes.

Alguns dos dados que compõem o banco de dados foram disponibilizados em formato de planilha eletrônica sendo necessário um tratamento antes da inserção no banco de dados. Além disso, pela grande quantidade de dados a ser inserida, programas que automatizassem essa tarefa também precisaram ser criados. Ambos, tratamento dos dados e automatização das inserções, foram desenvolvidos em linguagem Perl e Shell Script, sendo essas as melhores linguagens para tratamento deste tipo de dados. Perl tem uma curva de aprendizado rápida e utiliza de diversas otimizações para trabalhar com cadeia de caractéres, as *strings*, e textos, o que ajuda ao manipular grandes sequências genômicas.

3.3 Classificação COG e *Xanthomonas*

Todas as informações sobre o COG, incluindo os grupos formados e as sequências de genes do organismos, estão disponíveis publicamente através do portal do NCBI. Deste banco de dados foram extraídas as categorias do COG para serem inseridas no LeifDB, assim como os grupos formados.

Para categorizar os genes das bactérias *Leifsonia xyli* subsp. *xyli* e *Leifsonia xyli* subsp. *cynodontis* segundo o banco COG foi realizado o BLAST das proteínas com as informações extraídas sobre os grupos e as categorias. Somente o melhor resultado foi aproveitado para ser armazenado. Com isso, tem-se registrado, no LeifDB, o gene do COG similar ao gene a ser categorizado e por inferência este gene recebe a mesma categoria atribuída ao gene do COG.

3.4 Interface Web

Para tornar os dados contidos no banco acessíveis para pesquisadores interessados, uma interface web para acesso ao banco de dados foi esquematizada. Através desta plataforma

os usuários podem ter acesso aos dados sem precisar ter conhecimento aprofundado sobre linguagem de consulta à bancos de dados, como o SQL, ou conhecimento sobre a estrutura do banco. A proposta é permitir que os usuários possam fazer buscas por palavras-chave em relação aos genes e navegar pelas informações relacionadas, como os grupos e categorias funcionais, através de links disponíveis nas páginas. Além disso, ligações externas a outros bancos de dados poderão ser vistas para que o usuário obtenha outras informações não armazenadas, como por exemplo, buscar artigos relacionados no banco de dados Pubmed[9]. Para ter maior controle dos usuários do LeifDB, o acesso remoto será restrito a usuários previamente cadastrados.

[9]https://www.ncbi.nlm.nih.gov/pubmed/

4 Resultados e Discussão

Esta seção é iniciada ilustrando resultados provenientes de um estudo de caso, em que uma palavra-chave é usada para fazer uma busca no banco de dados. Este processo ilustra os resultados obtidos através da interface web proposta. Em seguida, estudos estatísticos sobre o conteúdo do LeifDB e sobre a classificação funcional automatizada são apresentados.

4.1 Visualização Web

Para a demonstração foi feita uma busca utilizando a interface gráfica prototipada para demonstrar como seriam os resultados e a aparência do LeifDB em um ambiente web. A prototipação utilizou tecnologias de simples implementação como HTML, CSS, Bootstrap e Perl.

O primeiro contato com o sistema é feito por uma página de autenticação onde são requeridos um login (e-mail) e uma senha para ter acesso aos dados. Há também nesta parte a opção de solicitar o acesso ao LeifDB e de recuperar seu acesso caso tenha esquecido a senha. Na Figura 4.1 é demonstrado o protótipo da tela de acesso.

Figura 4.1: Página de autenticação.

Uma vez que o usuário foi autenticado, ele possui acesso ao sistema sendo direcionado para o mecanismo de busca ao banco de dados (Figura 4.2). Nele existem três opções para a busca, inserindo o identificador do gene segundo o software Prokka, escrevendo parte da anotação de um gene e selecionando qual organismo será consultado com as características anteriores.

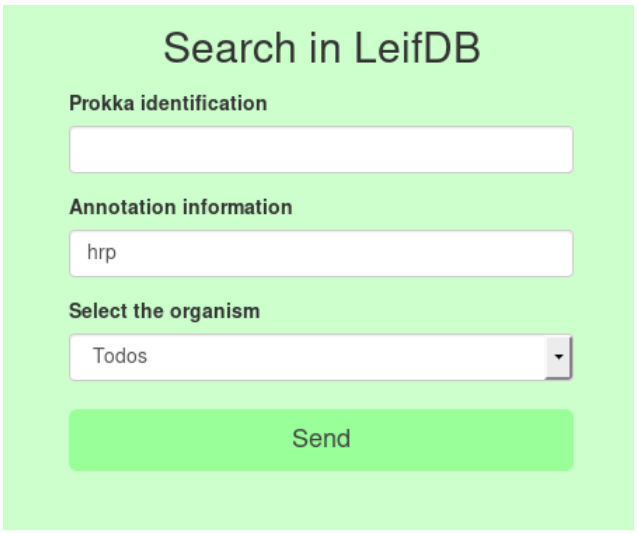

Figura 4.2: Exemplo de busca que contenha a palavra-chave 'hrp' na 'Annotation information'.

Para esta busca não foi inserido nenhum identificador e procurou-se pelos genes de todos os organismos que tivessem em sua anotação "hrp", uma sigla para *hypersensitive response and pathogenicity*. Foi escolhida essa anotação para a pesquisa uma vez que os genes anotados com essa sigla estão envolvidos com a formação de um sistema de secreção que é essencial para a virulência de bactérias que infectam plantas, animais e humanos. Este sistema é utilizado pelo organismo para injetar proteínas efetoras dentro das células hospedeiras a fim de sobrepor as defesas de seu hospedeiro (HUECK, 1998; GALÁN; COLLMER, 1999). Portanto, os grupos de genes reultantes tendem a ter essa mesma função.

Na Figura 4.3 são apresentados alguns dos genes que foram retornados após a busca. Nesta etapa aparecem algumas informações sobre os genes como seu identificador, a anotação completa, a qual organismo e grupo o gene pertence e sua classificação funcional de acordo com o COG e *Xanthomonas*.

Pode-se observar que as categorias referente ao COG dos genes retornados são parecidas e os genes são todos do mesmo grupo, o que faz sentido uma vez que todos os

Search Results

Prokka Id	Annotation	Organism	Group	COG	Xanthomonas
Lxx_00943	ATP-dependent RNA helicase HrpB	Leifsonia xyli subsp. xyli str. CTCB07	orth1346	L	-
Lxx_01958	ATP-dependent RNA helicase HrpA	Leifsonia xyli subsp. xyli str. CTCB07	orth1925	J	-
Cmsepe_01696	ATP-dependent RNA helicase HrpB	Clavibacter michiganensis subsp. sepedonicus	orth1346	L	-
Cmsepe_02365	ATP-dependent RNA helicase HrpB	Clavibacter michiganensis subsp. sepedonicus	orth1925	J	-
Cmmich382_01769	ATP-dependent RNA helicase HrpB	Clavibacter michiganensis subsp. michiganensis NCPPB 382	orth1346	L	-

Figura 4.3: Resultados da busca.

genes desempenham a mesma função. Aqui os genes não foram classificados segundo as *Xanthomonas*.

Há a possibilidade de se obter mais detalhes sobre um gene específico apenas clicando no identificador do gene que desejar e na página da Figura 4.4 será exibida um perfil completo do gene com todas as informações disponíveis no LeifDB. Essa visualização é importante pois mostra o nome da categoria COG e a sequência de proteínas daquele gene.

Nesta etapa pode-se navegar até o grupo que contém o gene clicando no nome do grupo que ele pertence. Assim todos os membros pertencentes aquele grupo são exibidos, como mostrado na Figura 4.5. Esta página demonstra apenas um resumo dos genes, sendo possível também clicar em um identificador de gene para obter mais informações. Percebe-se que este agrupamento faz sentido, uma vez que as anotações dos genes são iguais e a categorização do COG é exatamente a mesma para todos.

4.2 Análise do Sistema LeifDB

O banco de dados biológicos LeifDB conta um total de 35904 registros que representam os genes associados aos 11 organismos (Tabela 4.1) armazenados atualmente. Os métodos

Gene: Lxx_00943

Prokka Id	Lxx_00943
Annotation information	ATP-dependent RNA helicase HrpB
Organism	Leifsonia xyli subsp. xyli str. CTCB07
Group	orth1346
COG Classification	L - Replication, recombination and repair
Xanthomonas Classification	-
Sequence	MTDQQQLSPAERFAASHQRARQPLLETFLTGLGFDLDPFQREACTCLENGRSVLVAAPTGAGKTIVAEFAVFLAMRQANA KVFYTTPMKALSNQKFQEFQDTYGPESVGLLTGDTNINSHARIVVMTTEVLRNMLYADSDLLGDLAYVVMDEVHYLADRF RGAVWEEVIIHLPPAVRMVSLSATVSNAEEFGDWLQAVRGDTDVVVSEERPVPLEQHILMRSKLIDLFDSSGLAAANRVN PELVQMARSGGRVLSSRQRRDIGRYHSRGGRPDSFRMNRAEIVRLLDEHNLLPAIFFLFSRNGCDAAVRQTLRAGVRLTE QRERDDIRSIVEERCRTLMDEDLAVLGYWEWLEGLEHGVAAHHAGMLPAFKEVVEELFRRKLVKVVFATETLALGINMPA RTVVLEKLEKFNGESRVPITPGEYTQLTGRAGRRGIDVEGNSVIQWEDGLDPQSVASLASRRSYPLNSSFRPTYNMAVNL IDQFGRQRTREILESSFAQFQADRAVVDLARKVRQQEESLAGYEKAMTCHLGDFREYSGVRRELTDLERKGQRLDSASRA DRDRRQRQLTELRKRMREHPCHRCSDREQHARWAERWWKLKRETDLLSAQIQSRTGAVAKVFDRVSDVLDELGYLVVEDG VTKLTVHGRTLKRIYGERDLLVAECLRRGTWKELDAPSLAAMACALVFEPRRDDGLGHDRALPRGAFLPALDKTTDLWAR LDDRERENRLPGSEPPSTALALAMHQWARGSGLDAVLREADMAAGDFVRWTKQTIDLLDQLSLVAQGNLGRTARQALEAI RRGIVAYSSVA

Figura 4.4: Visualização das informações de um gene.

Group: orth1346

Prokka Id	Annotation	Organism	COG	Xanthomonas
Lxx_00943	ATP-dependent RNA helicase HrpB	Leifsonia xyli subsp. xyli str. CTCB07	L	-
Cmsepe_01696	ATP-dependent RNA helicase HrpB	Clavibacter michiganensis subsp. sepedonicus	L	-
Cmmich382_01769	ATP-dependent RNA helicase HrpB	Clavibacter michiganensis subsp. michiganensis NCPPB 382	L	-
Cmnebras_01657	ATP-dependent RNA helicase HrpB	Clavibacter michiganensis subsp. nebraskensis NCPPB 2581	L	-
CmiCF11_00665	ATP-dependent RNA helicase HrpB	Clavibacter michiganensis CF11	L	-
CmLMG_00615	ATP-dependent RNA helicase HrpB	Clavibacter cf. michiganensis LMG 26808	L	-

Figura 4.5: Informações sobre um grupo.

de agrupamento resultaram na formação de 5387 grupos que abrangem 31396 (87,44%) genes armazenados (Figura 4.6).

A categorização funcional segundo o projeto *Xanthomonas* classificou 724 genes

Tabela 4.1: Organismos

Clavibacter michiganensis subsp. sepedonicus
Clavibacter michiganensis subsp. michiganensis NCPPB 382
Clavibacter michiganensis subsp. nebraskensis NCPPB 2581
Clavibacter michiganensis CF11
Clavibacter cf. michiganensis LMG 26808
Leifsonia aquatica H1aii
Leifsonia aquatica ATCC 14665
Leifsonia rubra CMS 76R
Leifsonia xyli subsp. xyli str. CTCB07
Leifsonia xyli subsp. cynodontis DSM 46306
Leifsonia sp. 109

da bactéria *Leifsonia xyli* subsp. *xyli* (Figura 4.7) e 736 genes da *Leifsonia xyli* subsp. *cynodontis* (Figura 4.8). Os genes das categorias VIII (Hipotético[10]) e IX (ORFs com Categorias Indefinidas) não possuem uma função definida ainda ou apenas há especulações sobre seu propósito, assim essa classificação mostra-se uma boa prática para determinar funções dos genes uma vez que aproximadamente 6,5% destes genes foram classificados nestas categorias.

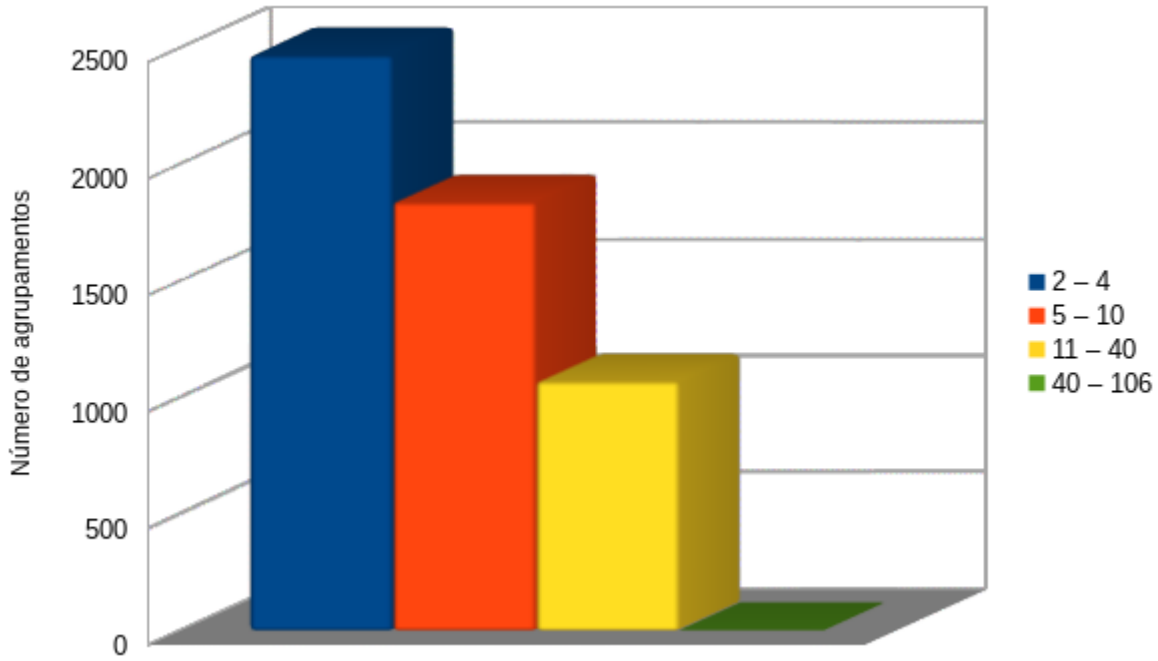

Figura 4.6: Número de agrupamentos divididos em seções de acordo com o número de membros. A menor faixa conta apenas com 3 membros.

A categorização funcional de acordo com o banco de dados COG classificou 2238

[10]imagina-se que determinada região do DNA ou RNA encontrada seja um gene

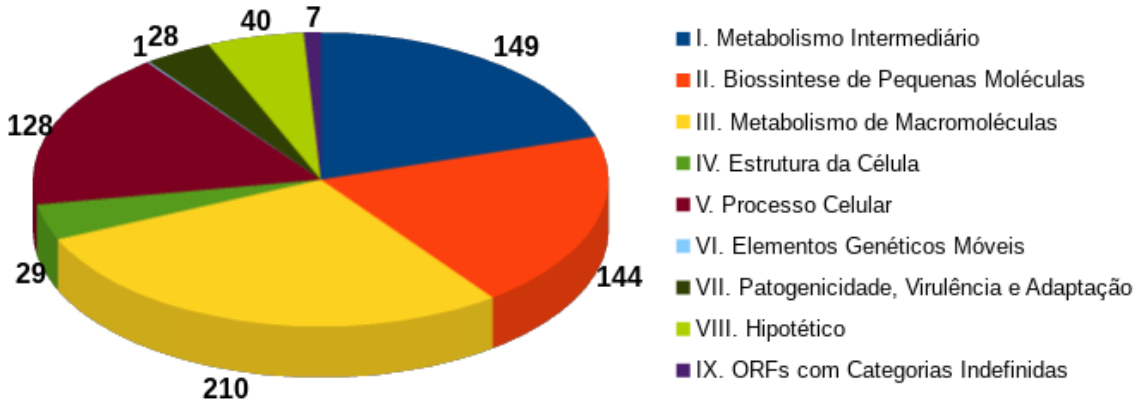

Figura 4.7: Número de genes em cada categoria das *Xanthomonas* da *Leifsonia xyli* subsp. *cynodontis*.

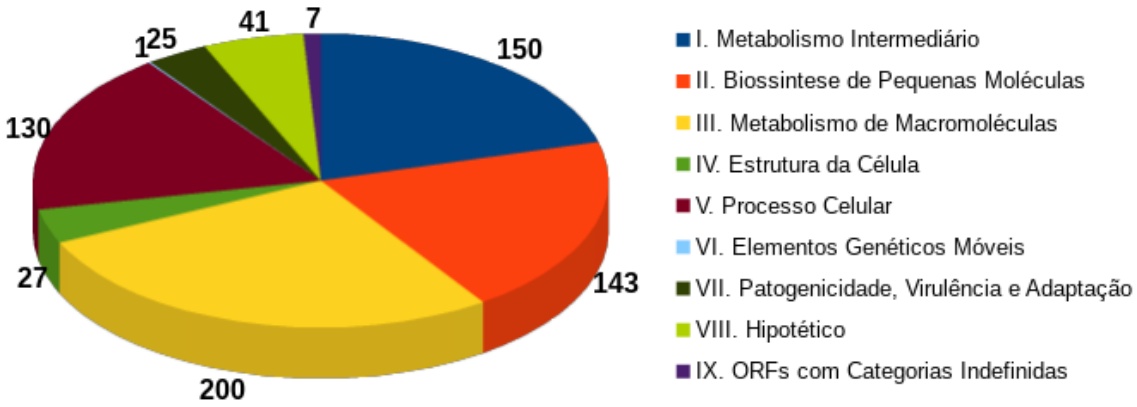

Figura 4.8: Número de genes em cada categoria das *Xanthomonas* da *Leifsonia xyli* subsp. *xyli*.

genes da bactéria *Leifsonia xyli* subsp. *cynodontis* (Figura 4.9) e 2214 genes da bactéria *Leifsonia xyli* subsp. *xyli* (Figura 4.10). Observa-se pouca variação na quantidade de genes por categoria e quais categorias têm mais genes, o que já era de se esperar uma vez que as duas bactérias são da mesma espécie. As categorias B (Chromatin structure and dynamics) e Y (Nuclear structure) não obtiveram nenhum gene classificado como sendo de suas respectivas categorias.

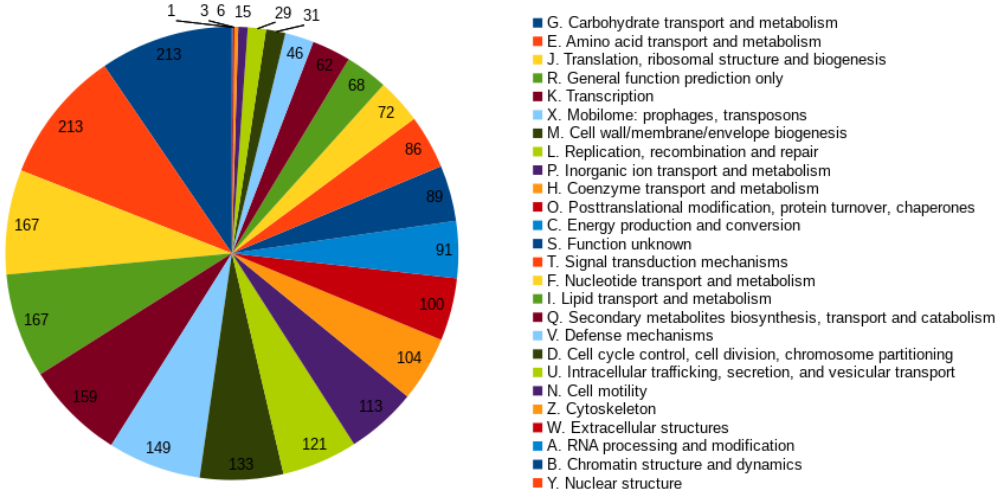

Figura 4.9: Número de genes da *Leifsonia xyli* subsp. *cynodontis* em cada categoria do COG.

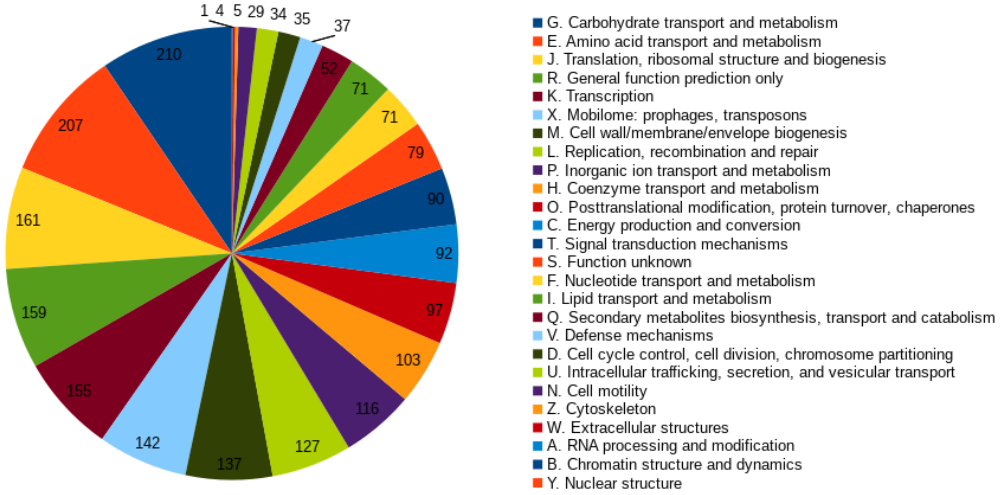

Figura 4.10: Número de genes da *Leifsonia xyli* subsp. *xyli* em cada categoria do COG.

5 Conclusão

Pode-se observar que o banco de dados LeifDB preenche uma lacuna existente em relação aos bancos de dados biológicos que tratam sobre a bactéria *Leifsonia xyli* auxiliando pesquisadores no combate ao raquitismo-da-soqueira nas plantações de cana-de-açúcar ao prover em um ambiente organizado informações referentes aos seus genes. Pesquisas feitas sobre esses dados podem auxiliar no entendimento da maquinaria funcional do genoma e podem ser aliadas no desenvolvimento de métodos preventivos. O LeifDB conta com 35904 genes em seus registros, com a formação de grupos abrangendo 87,44% destes genes. Desta forma, o LeifDB tem potencial para se tornar uma importante ferramenta para o estudo deste e de outros agentes etiológicos causadores de doenças.

Em uma perspectiva futura um sistema de informação pode ser feito para oferecer mais funcionalidades do que as presentes hoje no LeifDB, como por exemplo análises comparativas em tempo real. Uma outra opção é a criação de uma API, *Application Programming Interface*, para que haja a troca de informações entre sistemas computacionais com diferentes bancos de dados de uma maneira automatizada.

Bibliografia

ALMEIDA, F. N. *Implementação de um Banco de Dados de Proteomas de Bactérias Associadas a Plantas: Probacter*. Dissertação (Mestrado), 2007.

ALTSCHUL, S. F. et al. Basic local alignment search tool. *Journal of molecular biology*, Elsevier, v. 215, n. 3, p. 403–410, 1990.

BARTHOLOMEW, D. *Getting started with MariaDB*. [S.l.]: Packt Publishing Ltd, 2015.

BELL, G.; HEY, T.; SZALAY, A. Beyond the data deluge. *Science Magazine*, 2009.

BENTLEY, S. D. et al. Genome of the actinomycete plant pathogen clavibacter michiganensis subsp. sepedonicus suggests recent niche adaptation. *Journal of bacteriology*, Am Soc Microbiol, v. 190, n. 6, p. 2150–2160, 2008.

BERMAN, H. M. et al. The protein data bank. *Nucleic acids research*, Oxford Univ Press, v. 28, n. 1, p. 235–242, 2000.

BILOFSKY, H. S. et al. The genbank genetic sequence databank. *Nucleic acids research*, Oxford Univ Press, v. 14, n. 1, p. 1–4, 1986.

BOICEA, A.; RADULESCU, F.; AGAPIN, L. I. Mongodb vs oracle–database comparison. In: IEEE. *Emerging Intelligent Data and Web Technologies (EIDWT), 2012 Third International Conference on*. [S.l.], 2012. p. 330–335.

DATE, C. J. *Introdução a sistemas de bancos de dados*. [S.l.]: Elsevier Brasil, 2004.

DIAS, V. D. et al. Detecção com técnicas moleculares de leifsonia xyli subsp. xyli e xanthomonas albilineans em cana-deaçúcar. Universidade Federal de Goiás, 2016.

DONGEN, S. M. V. *Graph clustering by flow simulation*. Tese (Doutorado), 2001.

DU, Y. et al. Draft genome sequence of the cellulolytic bacterium clavibacter sp. cf11, a strain producing cold-active cellulase. *Genome announcements*, Am Soc Microbiol, v. 3, n. 1, p. e01304–14, 2015.

ENRIGHT, A. J.; DONGEN, S. V.; OUZOUNIS, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, Oxford University Press, v. 30, n. 7, p. 1575–1584, 2002.

EVTUSHENKO, L. I. et al. Leifsonia poae gen. nov., sp. nov., isolated from nematode galls on poa annua, and reclassification of'corynebacterium aquaticum'leifson 1962 as leifsonia aquatica (ex leifson 1962) gen. nov., nom. rev., comb. nov. and clavibacter xyli davis et al. 1984 with two subspecies as leifsonia xyli (davis et al. 1984) gen. nov., comb. nov. *International journal of systematic and evolutionary microbiology*, Microbiology Society, v. 50, n. 1, p. 371–380, 2000.

FITCH, W. M. Distinguishing homologous from analogous proteins. *Systematic zoology*, Society of Systematic Zoology, v. 19, n. 2, p. 99–113, 1970.

G1. *Produção de cana no Brasil aumenta em 2017*. 2017. Disponível em: ⟨https://g1.globo.com/economia/agronegocios/agro-a-industria-riqueza-do-brasil/noticia/producao-de-cana-no-brasil-aumenta-em-2017.ghtml⟩.

GALÁN, J. E.; COLLMER, A. Type iii secretion machines: bacterial devices for protein delivery into host cells. *Science*, American Association for the Advancement of Science, v. 284, n. 5418, p. 1322–1328, 1999.

GALPERIN, M. Y.; FERNÁNDEZ-SUÁREZ, X. M.; RIGDEN, D. J. The 24th annual nucleic acids research database issue: a look back and upcoming changes. *Nucleic acids research*, Oxford University Press, v. 45, n. D1, p. D1–D11, 2017.

GALPERIN, M. Y. et al. Expanded microbial genome coverage and improved protein family annotation in the cog database. *Nucleic acids research*, Oxford University Press, v. 43, n. D1, p. D261–D269, 2014.

GARTEMANN, K. et al. Clavibacter michiganensis subsp. nebraskensis ncppb 2581 complete genome. *Unpublished*, 2011.

GARTEMANN, K.-H. et al. The genome sequence of the tomato-pathogenic actinomycete clavibacter michiganensis subsp. michiganensis ncppb382 reveals a large island involved in pathogenicity. *Journal of bacteriology*, Am Soc Microbiol, v. 190, n. 6, p. 2138–2149, 2008.

GRAMATES, L. S. et al. Flybase at 25: looking to the future. *Nucleic acids research*, Oxford University Press, v. 45, n. D1, p. D663–D671, 2017.

HUECK, C. J. Type iii protein secretion systems in bacterial pathogens of animals and plants. *Microbiology and molecular biology reviews*, Am Soc Microbiol, v. 62, n. 2, p. 379–433, 1998.

JUNIOR, J. D. B. C. *EFEITO DO TRATAMENTO TÉRMICO E DA INOCULAÇÃO DE BACTÉRIAS ENDOFÍTICAS NO CONTROLE DO RAQUITISMO DA SOQUEIRA DA CANA-DE-AÇÚCAR*. Tese — Centro de Ciências e Tecnologias Agropecuárias da Universidade Estadual do Norte Fluminense Darcy Ribeiro, Julho 2006.

LI, L.; STOECKERT, C. J.; ROOS, D. S. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome research*, Cold Spring Harbor Lab, v. 13, n. 9, p. 2178–2189, 2003.

MARCUZ, F. S. et al. Levantamento da incidência de leifsonia xyli subsp. xyli em plantios de cana-de-açúcar do noroeste do paraná. *Ciênc. agrotec.,(Impr.)*, v. 33, n. spe, p. 1935–1939, 2009.

MONTEIRO-VITORELLO, C. B. et al. The genome sequence of the gram-positive sugarcane pathogen leifsonia xyli subsp. xyli. *Molecular plant-microbe interactions*, Am Phytopath Society, v. 17, n. 8, p. 827–836, 2004.

MONTEIRO-VITORELLO, C. B. et al. Complete genome sequence of leifsonia xyli subsp. cynodontis strain dsm46306, a gram-positive bacterial pathogen of grasses. *Genome announcements*, Am Soc Microbiol, v. 1, n. 6, p. e00915–13, 2013.

PETERSON, J. D. et al. The comprehensive microbial resource. *Nucleic acids research*, Oxford Univ Press, v. 29, n. 1, p. 123–125, 2001.

RIGDEN, D. J.; FERNÁNDEZ-SUÁREZ, X. M.; GALPERIN, M. Y. The 2016 database issue of nucleic acids research and an updated molecular biology database collection. *Nucleic acids research*, Oxford University Press, v. 44, n. D1, p. D1–D6, 2015.

SEEMANN, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, Oxford University Press, v. 30, n. 14, p. 2068–2069, 2014.

SILVA, A. R. da et al. Comparison of the genomes of two xanthomonas pathogens with differing host specificities. *Nature*, Nature Publishing Group, v. 417, n. 6887, p. 459–463, 2002.

TATUSOV, R. L. et al. The cog database: an updated version includes eukaryotes. *BMC bioinformatics*, BioMed Central, v. 4, n. 1, p. 41, 2003.

TATUSOV, R. L.; KOONIN, E. V.; LIPMAN, D. J. A genomic perspective on protein families. *Science*, American Association for the Advancement of Science, v. 278, n. 5338, p. 631–637, 1997.

URASHIMA, A. et al. Prevalence and severity of ratoon stunt in commercial brazilian sugarcane fields. *Plant Disease*, Am Phytopath Society, v. 101, n. 5, p. 815–821, 2017.

ZAŁUGA, J. et al. Comparative genome analysis of pathogenic and non-pathogenic clavibacter strains reveals adaptations to their lifestyle. *BMC genomics*, BioMed Central Ltd, v. 15, n. 1, p. 392, 2014.

ZOU, D. et al. Biological databases for human research. *Genomics, proteomics & bioinformatics*, Elsevier, v. 13, n. 1, p. 55–63, 2015.