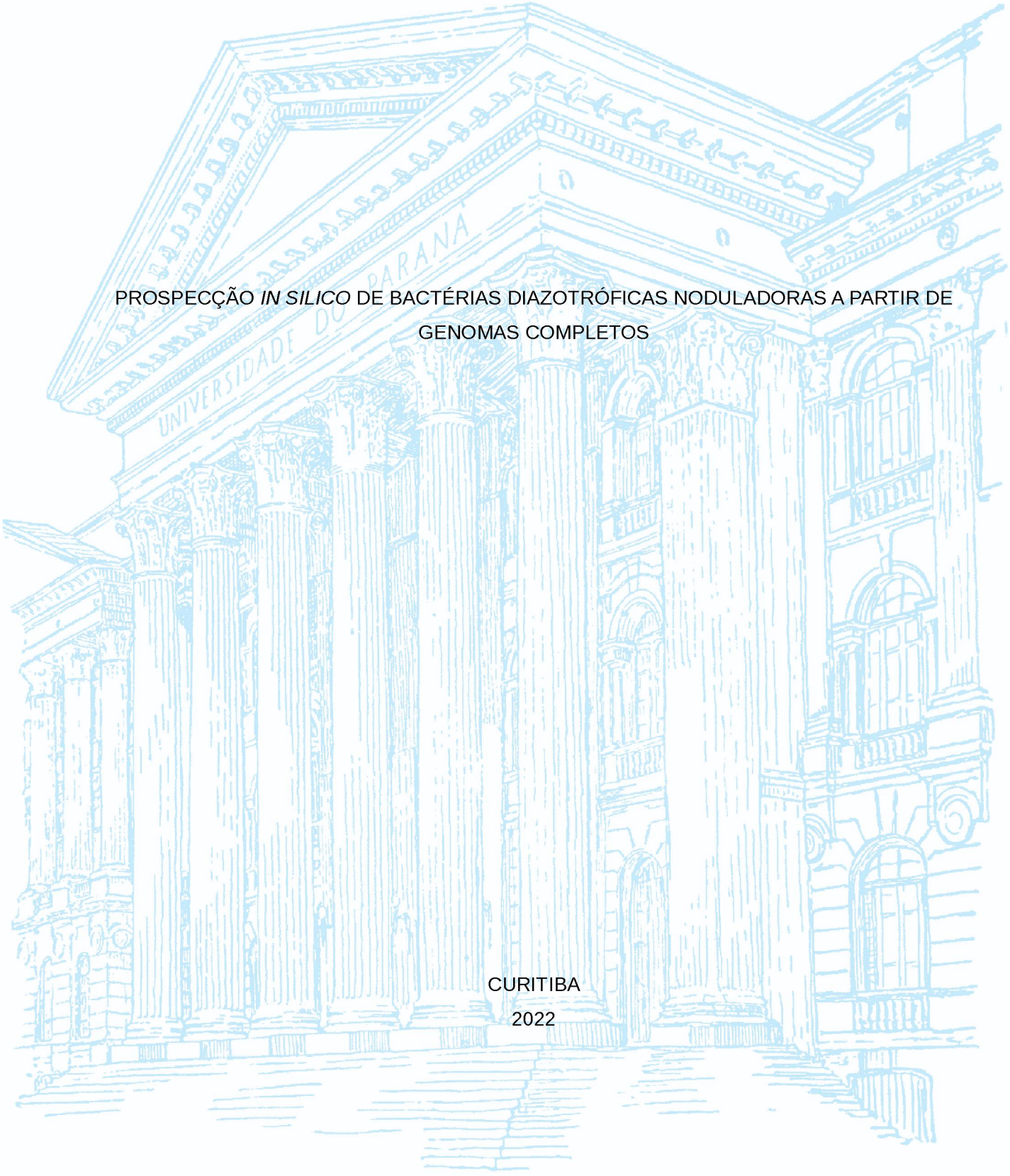
UNIVERSIDADE FEDERAL DO PARANÁ

ARYEL MARLUS REPULA DE OLIVEIRA



PROSPECÇÃO *IN SILICO* DE BACTÉRIAS DIAZOTRÓFICAS NODULADORAS A PARTIR DE GENOMAS COMPLETOS

CURITIBA

2022

ARYEL MARLUS REPULA DE OLIVEIRA

PROSPECÇÃO *IN SILICO* DE BACTÉRIAS DIAZOTRÓFICAS NODULADORAS A PARTIR DE GENOMAS COMPLETOS

Tese apresentada ao Curso de Pós-Graduação em Genética, Setor de Ciências Biológicas, da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Doutor em Genética.

Orientadora: Profa. Dra. Ana Claudia Bonatto

Coorientador: Prof. Dr. Roberto Tadeu Raittz

CURITIBA

2022

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIAS BIOLÓGICAS

Oliveira, Aryel Marlus Repula de
Prospecção *in silico* de bactérias diazotróficas noduladoras a partir de genomas completos / Aryel Marlus Repula de Oliveira. – Curitiba, 2022.
1 recurso on-line : PDF.

Tese (Doutorado) – Universidade Federal do Paraná, Setor de Ciências Biológicas, Programa de Pós-graduação em Genética.
Orientadora: Profa. Dra. Ana Claudia Bonatto.
Coorientador: Prof. Dr. Roberto Tadeu Raittz.

1. Nitrogênio - Fixação. 2. Nodulação. 3. Aprendizado de máquina. 4. Simulação por computador. I. Bonatto, Ana Claudia, 1978. II. Raittz, Roberto Tadeu, 1966-. III. Universidade Federal do Paraná. Setor de Ciências Biológicas. Programa de Pós-graduação em Genética. IV. Título.

Bibliotecária: Giana Mara Seniski Silva CRB-9/1406

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS BIOLÓGICAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO GENÉTICA - 40001016006P1

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação GENÉTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **ARYEL MARLUS REPULA DE OLIVEIRA** intitulada: **Prospecção in silico de bactérias diazotróficas noduladoras a partir de genomas completos.**, sob orientação da Profa. Dra. ANA CLAUDIA BONATTO, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 29 de Novembro de 2022.

Assinatura Eletrônica
30/11/2022 17:31:48.0
ANA CLAUDIA BONATTO
Presidente da Banca Examinadora

Assinatura Eletrônica
06/12/2022 18:30:12.0
DIEVAL GUIZELINI
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
30/11/2022 16:49:54.0
DANILO SIPOLI SANCHES
Avaliador Externo (55002145)

Assinatura Eletrônica
30/11/2022 17:47:00.0
LIZIANE CRISTINA CAMPOS BRUSAMARELLO DOS SANTOS
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
02/12/2022 10:33:11.0
ROBERTO TADEU RAITTZ
Coorientador(a) (UNIVERSIDADE FEDERAL DO PARANÁ)

DEDICATÓRIA

Dedico este trabalho à Ellen e à Suellen, que, mesmo depois de todas as dificuldades enfrentadas pela jornada que decidi seguir, e todas as mudanças que me causaram, continuam ao meu lado como uma família.

AGRADECIMENTOS

Agradeço a Deus pelo discernimento e aprendizado que tive nesta difícil jornada. Um maior conhecimento científico não retirou minha fé.

Agradeço pela paciência que minha família teve nos inúmeros momentos em que eu mesmo não sabia o que isso significava. E pelas tentativas de me resgatar de abismos emocionais que despenquei.

Agradeço minha orientadora Ana por aceitar fazer parte desta jornada, mesmo envolvendo áreas que antes não lhe eram familiares (hoje já lhe são familiares). Espero que eu tenha conseguido contribuir de alguma forma com seu crescimento assim como ela contribuiu com o meu. Poucas pessoas que conheci, tanto nessa jornada quanto em minha vida profissional anterior, possuem mente aberta a ponto de entrar em discussões e trabalhos que estão fora de sua área de conforto. Acredito que esta atitude torna as pessoas melhores do que eram no passado, e sendo uma atitude, minha experiência mostra que é muito mais difícil de se ensinar do que um conhecimento específico. Sua atitude é digna de admiração e exemplo a muitos profissionais.

Agradeço ao meu co-orientador Roberto, que esteve presente em minha jornada acadêmica desde minha graduação em 2008. Hoje, depois de o conhecer melhor durante o mestrado e doutorado, afirmo que seu conhecimento técnico é difícil de ser encontrado tanto dentro quanto fora da academia, e sou grato por ter compartilhado um pouco comigo.

Um curso de pós-graduação é feito por alunos e mestres. Pela minha experiência nesse período, reconheço que os mestres do programa realizam uma excelente tarefa ao guiar os alunos em um caminho científico ao estimular boas discussões. Agradeço aos mestres por isso, acredito que jamais saberão o real impacto que tiveram em seus alunos. Agradeço também aos alunos, com quem tive discussões intensas sobre vários assuntos distintos na sala de convivência. Acredito que esta interação é benéfica para alunos, mestres e para o próprio programa de pós-graduação.

Agradeço também a todos que participaram das diferentes bancas de avaliação deste trabalho, cada um contribuiu significativamente com seu tempo e conhecimento para guiar nosso projeto e melhorá-lo com pontos de vista distintos. Não foram poucas as vezes em que aprendi muito mais em discussões com a banca do que em vários livros, sua experiência e contribuição engrandecem tanto profissionalmente quanto pessoalmente.

“O tempo nos faz esquecer
O que nos trouxe até aqui
Mas eu lembro muito bem
Como se fosse amanhã”
(Armas químicas e poemas, Engenheiros do Hawaii)

RESUMO

A disponibilidade de nitrogênio é essencial para promoção de crescimento vegetal na agricultura. O uso de fertilizantes nitrogenados é amplamente utilizado, mas além de possuir um alto custo financeiro causa danos ao ambiente, sendo a fixação biológica de nitrogênio uma alternativa sustentável. Porém o processo de fixação biológica de nitrogênio é complexo e na maioria das vezes é espécie-específico, dificultando a prospecção para uso em diferentes culturas de plantio. Dentro dos diferentes tipos de fixação de nitrogênio uma das mais eficazes é a simbiótica, em que os organismos formam nódulos especializados para esta função. Para possibilitar a prospecção de diazotrofos noduladores a partir de genomas depositados em bancos de dados, desenvolvemos um modelo de aprendizagem de máquina. O modelo apresentou resultados melhores que os métodos de predições *in silico* atuais e possibilitou a expansão do conjunto dos organismos potencialmente noduladores para exploração e análise biológica.

Palavras-chave: 1. Fixação de nitrogênio 2. Nodulação 3. Machine Learning 4. Prospecção *in silico*

ABSTRACT

Nitrogen availability is essential for promoting plant growth in agriculture. The use of nitrogen fertilizers is widely used, but in addition to having a high financial cost, it causes damage to the environment, and biological nitrogen fixation is a sustainable alternative. However, the process of biological nitrogen fixation is complex and most of the time is species-specific, making it difficult to prospect for use in different planting crops. Among the different types of nitrogen fixation, one of the most effective is the symbiotic one, in which organisms form specialized nodules for this function. To enable the prospection of nodulating diazotrophs from genomes deposited in databases, we developed a machine-learning model. The model presented better results than current *in silico* prediction methods and enabled the expansion of potentially nodulating organisms for exploration and biological analysis.

Keywords: 1. Nitrogen fixing 2. Plant Nodulation 3. Machine Learning 4. Prospecção *in silico*

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS OU SIGLAS

CDSs – Regiões codificadoras do genoma

SUMÁRIO

1 APRESENTAÇÃO

Esta tese está estruturada em formato de artigo e possui três seções principais: A parte I apresenta o contexto, fundamentação teórica e os objetivos deste projeto. A parte II apresenta o artigo, principal produto do projeto. E a parte III é o apêndice, que discute resultados não publicados que possuem potencial para discussão e trabalhos futuros.

Parte I

FUNDAMENTAÇÃO TEÓRICA

2 INTRODUÇÃO

O nitrogênio é o elemento mais utilizado para promover um crescimento vegetal acelerado em produções agrícolas. Disponível principalmente na forma de fertilizantes nitrogenados que são utilizados na maioria da produção de alimentos agrícolas no mundo. Embora os resultados sejam satisfatórios do ponto de vista de promoção de crescimento, estima-se que pelo menos 50% dos nutrientes presentes nesse tipo de fertilizante não são absorvidos pela plantação e são depositados no ambiente em forma de amônia (NH_3), nitrato (NO_3^-) e óxido de nitrato (N_2O), causando aumento dos custos de produção, impactos ambientais e alterações climáticas (Coskun et al., 2017).

Uma alternativa sustentável aos fertilizantes nitrogenados é a fixação biológica de nitrogênio, realizada por um grupo de procariotos chamados de diazotrofos. Estas bactérias são capazes de converter o nitrogênio atmosférico (N_2) em amônia (NH_3), forma que as plantas conseguem absorver. Os diazotrofos incluem organismos aquáticos como cyanobacterias, bactérias de vida livre como *Azotobacter*, bactérias que formam associações como *Azospirillum*, e os noduladores que formam mecanismos especializados durante a simbiose com as plantas hospedeiras, como espécies de *Rhizobium* e *Bradyrhizobium* (Postgate, 1982).

Organismos diazotróficos simbióticos possuem uma interação planta-bactéria complexa, onde nem todos os mecanismos biológicos são conhecidos ou conservados, podendo variar entre diferentes espécies. Essa simbiose não é permanente ou necessária para a sobrevivência das espécies envolvidas, dificultando a prospecção e identificação de novos organismos diazotróficos nos ambientes analisados, pois nem sempre serão visualizados *in vivo*, ainda que eventualmente exista em uma amostra organismos potencialmente diazotróficos. A limitação do número de organismos simbióticos limita também a descoberta dos mecanismos biológicos envolvidos no processo.

2.1 JUSTIFICATIVA

A exploração de organismos diazotróficos simbióticos para promoção de crescimento em produções agrícolas em substituição ao uso de fertilizantes nitrogenados, que impactam o ambiente, é necessária. Porém a limitação de organismos conhecidos e o limitado conhecimento sobre os mecanismos biológicos envolvidos no processo de interação planta-bactéria dificultam a prospecção de potenciais substitutos efetivos para os fertilizantes nitrogenados.

2.2 OBJETIVOS

Este trabalho tem como objetivo melhorar os métodos para prospecção de organismos diazotróficos noduladores atuais e buscar genes potencialmente essenciais no processo de interação planta-bactéria.

2.2.1 Objetivos específicos

– Desenvolver um método para prospecção de organismos diazotróficos noduladores mais eficaz que os atuais;

– Construir ferramentas e bancos de dados que permitam exploração dos genes envolvidos no processo de interação planta-bactéria;

– Identificar genes potencialmente envolvidos no processo de interação planta-bactéria que ainda não são conhecidos.

3 REVISÃO DE LITERATURA

3.1 BACTÉRIAS FIXADORAS DE NITROGÊNIO

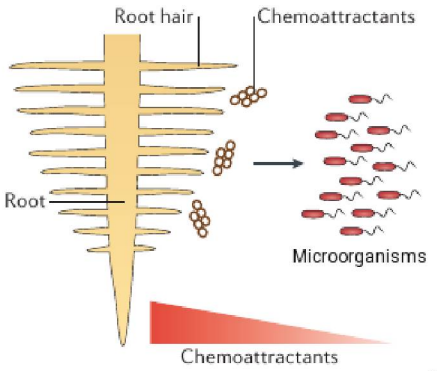
O nitrogênio é um dos principais elementos para promoção de crescimento vegetal. Apesar de ser abundante na atmosfera terrestre, a forma N_2 não pode ser utilizada pela grande maioria dos organismos vegetais. Na agricultura este elemento é disponibilizado para as plantas a partir de fertilizantes nitrogenados ou através da fixação biológica de nitrogênio (Wagner et al., 2011).

A fixação biológica de nitrogênio é realizada por procariotos chamados diazotróficos. Sua utilização é preferível ao de fertilizantes nitrogenados por apresentarem menor custo e menor impacto ambiental (Coskun et al., 2017). Esses procariotos incluem organismos aquáticos como cianobactérias, bactérias de vida-livre como *Azotobacter*, bactérias que criam relações associativas com plantas como *Azospirillum*, e as que formam simbiose e criam nódulos especializados, como *Rhizobium*, *Bradyrizobium, Mesorhizobium* e *Sinorhizobium* entre outros gêneros (Wagner et al., 2011).

Os diferentes organismos e suas distintas formas de associação possuem em comum a capacidade de reduzir o nitrogênio atmosférico à amônia. Esta redução é feita pelo complexo da nitrogenase que é formado a partir dos produtos de vários genes, sendo seis mais encontrados em co-ocorrência e, portanto, considerados um conjunto mínimo necessário: *nifH*, *nifD*, *nifK*, *nifE*, *nifN*, *nifB.* As sequências dos genes *nifHDK* são encontradas de forma mais conservada (dos Santos et al., 2012; Wang et al., 2018).

Os organismos diazotróficos habitam o solo e podem ser atraídos para a região da rizosfera a partir de compostos liberados pela planta (FIGURA 1). A rizosfera é rica em nutrientes e apresenta microrganismos de forma heterogênea (não apenas diazotrofos) e que sofrem constante pressão seletiva. A pressão seletiva é mais intensa quanto mais próximo da planta, como no rizoplano (superfície da raiz) e na endosfera (compartimento interno do córtex da raiz, entre as células da planta). A pressão ocorre tanto entre os microrganismos competindo entre si quanto em resposta a mecanismos de defesa ou seleção da planta. As interações que eventualmente ocorrem entre os organismos e a planta podem ser de promoção de crescimento ou de patogenicidade (Poole et al., 2018).

FIGURA 1 - *GRADIENTE DE ATRAÇÃO DE MICRORGANISMOS*

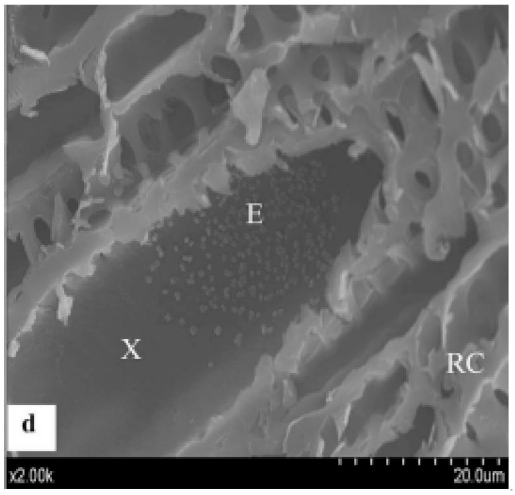


FONTE: Adaptado de Poole et al. (2018).

Quando diazotrofos de vida-livre estão presentes no solo, eles podem disponibilizar formas reduzidas do elemento que são consumidas pela planta, realizando a fixação biológica de nitrogênio sem ter interações diretas com algum hospedeiro. Estes tipos de organismos precisam buscar sua própria fonte de energia e condições ideais para realizar a redução do nitrogênio enquanto competem com outros microrganismos presentes, e por isso a taxa de fixação de nitrogênio é considerada baixa (Wagner et al., 2011).

Diazotrofos associativos por sua vez infectam plantas hospedeiras e as colonizam permanecendo no rizoplano e/ou endosfera entre as células da planta (Figura 2). Diferente dos diazotrofos de vida-livre onde apenas sua presença na rizosfera pode ser suficiente para que a fixação de nitrogênio ocorra, diazotrofos associativos precisam se proteger dos mecanismos de defesa da planta para que sua presença não seja inviabilizada. Para que a associação se estabeleça comportamentos similares aos de organismos patogênicos são observados nesses diazotrofos, embora sua presença não seja maléfica para o hospedeiro. A quantidade de nitrogênio fixada por esses organismos é superior aos de vida-livre, uma vez que as condições para que este processo ocorra são mais favoráveis por haver menor competição com outros microrganismos e as fontes energéticas serem mais abundantes. Esses organismos favorecem o crescimento da planta não só pela habilidade de fixar nitrogênio, mas também por competir e impedir que outros organismos patógenos se estabeleçam na região, algumas espécies ainda secretam fitohormônios reguladores de crescimento que servem como sinalizadores para promoção de crescimento vegetal (Fibach-Paldi et al., 2012). A quantidade de nitrogênio fixado depende de vários fatores como temperatura do solo, capacidade do hospedeiro em criar uma rizosfera com baixa concentração de oxigênio, energia disponível que possa ser consumida pelos diazotrofos, o grau de competição com outros microrganismos presentes, além da própria eficiência do diazotrofo (Wagner et al., 2011).

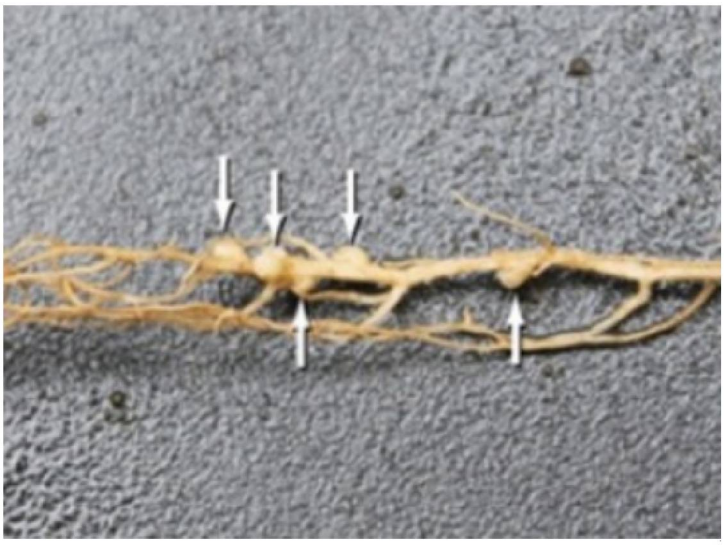
FIGURA 2 - BACTÉRIAS DIAZOTRÓFICAS ENDOFÍTICAS COLONIZAM PLANTAS



Região de raiz de planta inoculada por bactérias endofíticas após 14 dias da infecção. Células bacterianas (E) são encontradas em regiões do xilema (X) e córtex da raiz (RC). FONTE: Adaptado de Li et al. (2016).

Organismos diazotróficos que estabelecem simbiose para fixação de nitrogênio constituem o ambiente mais favorável para que o processo ocorra, mas também o mais complexo. Neste tipo de interação há alterações morfológicas tanto no diazotrofo quanto no hospedeiro, sendo necessário que ambos sejam compatíveis para viabilizar o processo e ocasionando que a simbiose seja espécie-específico em muitos casos. A simbiose estabelece nódulos especializados para fixação de nitrogênio que podem ser observados nas raízes (FIGURA 3). Os organismos diazotróficos se encontram no interior dos nódulos, protegidos pelas células do hospedeiro e, portanto, sem competição com outros microrganismos presentes na rizosfera, além de possuírem fonte de energia disponibilizada pela planta. Em troca os diazotrofos realizam a fixação de nitrogênio necessária para promover o crescimento do hospedeiro (Wagner et al., 2011).

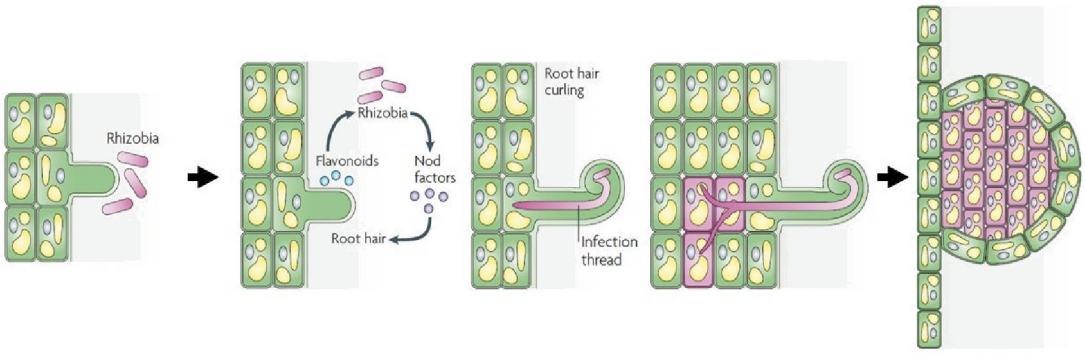
FIGURA 3 - NÓDULOS ESTABELECIDOS PARA FIXAÇÃO DE NITROGÊNIO



FONTE: Raiz de trevo inoculada. Adaptado de Bernhard et al. (2010).

A formação do nódulo é iniciada pelo reconhecimento e atração entre hospedeiro e diazotrofos. As plantas secretam flavonóides (FIGURA 4) na rizosfera que atraem e estimulam os organismos diazotróficos a secretarem fatores Nod. A planta então reage aos fatores e inicia-se o processo a partir de um tubo de infecção (*infection thread*). Os organismos infiltram-se na raiz do hospedeiro e a divisão celular é estimulada até se formar um nódulo que os envolve. As bactérias sofrem alterações morfológicas significativas (formato de bacterióides) e se tornam dependentes da planta para suprimento energético, em troca, realizam a fixação de nitrogênio para o hospedeiro (Deakin et al., 2009).

FIGURA 4 - *PROCESSO DE FORMAÇÃO DO NÓDULO DE FIXAÇÃO DE NITROGÊNIO*

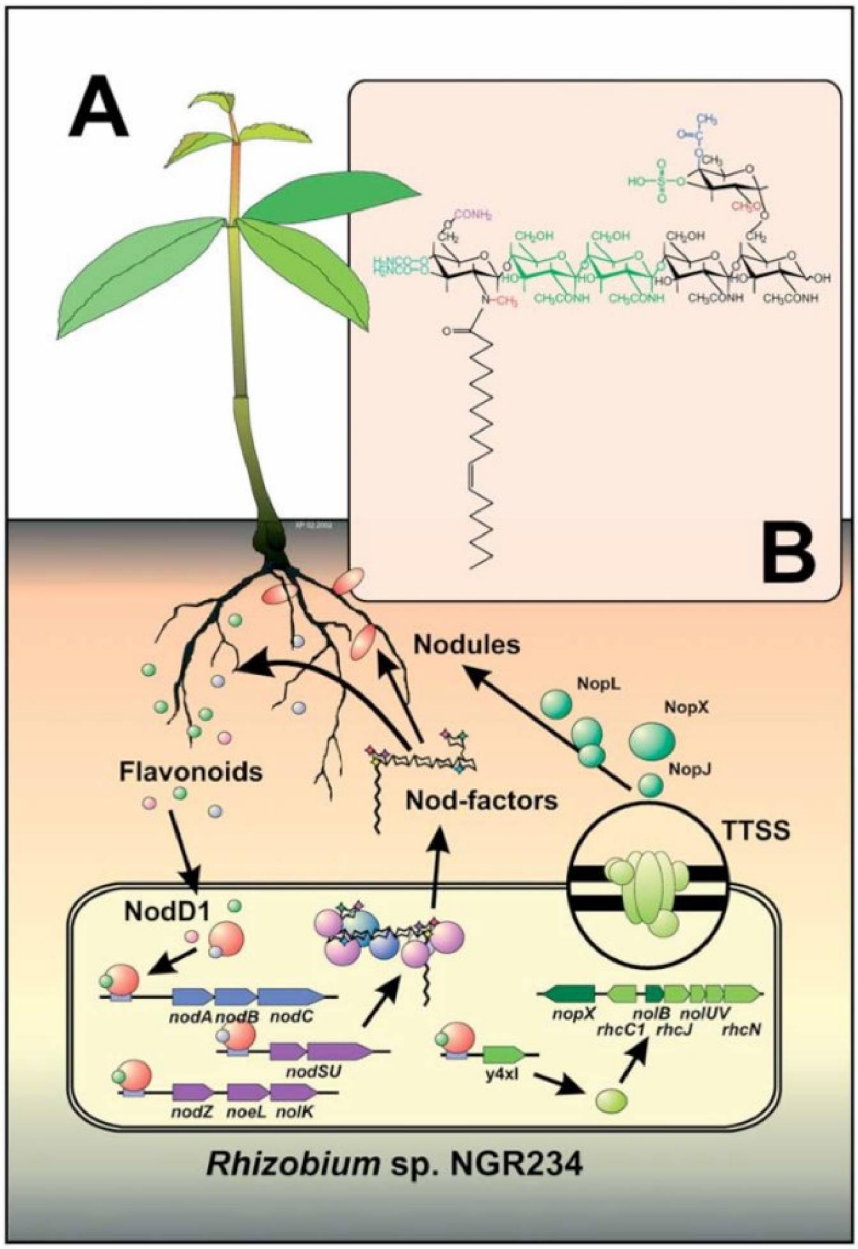


FONTE: Adaptado de Deakin et al. (2009).

É necessário reconhecimento e compatibilidade entre o hospedeiro e o diazotrofo para que a simbiose ocorra. Como por exemplo no processo de nodulação do organismo *Sinorhizobium fredii* NGR234, onde este reconhecimento envolve a liberação de flavonoides pelo hospedeiro, que são reconhecidos pelo diazotrofo e ativam a proteína NodD1, que atua como um regulador transcricional para uma cadeia de genes responsáveis pela produção dos fatores Nod que atuam como sinalizadores para a planta (FIGURA 5). Além de ativar a produção de fatores Nod, o regulador NodD1 promove a ativação do sistema de secreção tipo III da bactéria, que é essencial para inibir o sistema de defesa da planta e viabilizar a formação do nódulo (FIGURA 5, genes em verde destacados com a abreviação TTSS) (Watson et al., 2022).

A síntese dos fatores Nod necessita de uma cadeia de genes no processo, mas três são encontrados em co-ocorrência na maioria dos organismos e são considerados essenciais (embora existam exceções): *nodA*, *nodB* e *nodC*. Estes genes codificam para as enzimas que sintetizam o núcleo principal dos fatores Nod (Perret et al., 2000). Os genes envolvidos posteriormente ao estabelecimento do núcleo dos fatores Nod e o início do processo de nodulação apresentam variabilidade entre as espécies, tornando-se um dos fatores limitantes para prospecção de novos organismos dessa categoria, pois muitos deles estão relacionados à particularidade de interações espécie-específicas e não são universais para a simbiose. A identificação experimental requer condições específicas para cada espécie tornando difícil tanto a prospecção quanto a identificação de genes universais, pois há relativamente poucos organismos explorados (dos Santos et al., 2012). Este tipo de associação é o mais eficiente, porém mais complexo para se formar e prospectar, em contraste com os diazotrofos de vida livre por exemplo que não requerem nenhuma especificidade mas possuem menor taxa de fixação de nitrogênio (Stephen et al., 2011; Wang, 2019; Mergaet, 2020; Wang, 2018; Dixon, 2004).

FIGURA 5 - *ESTÍMULOS BÁSICOS PARA FORMAÇÃO DE NÓDULOS*

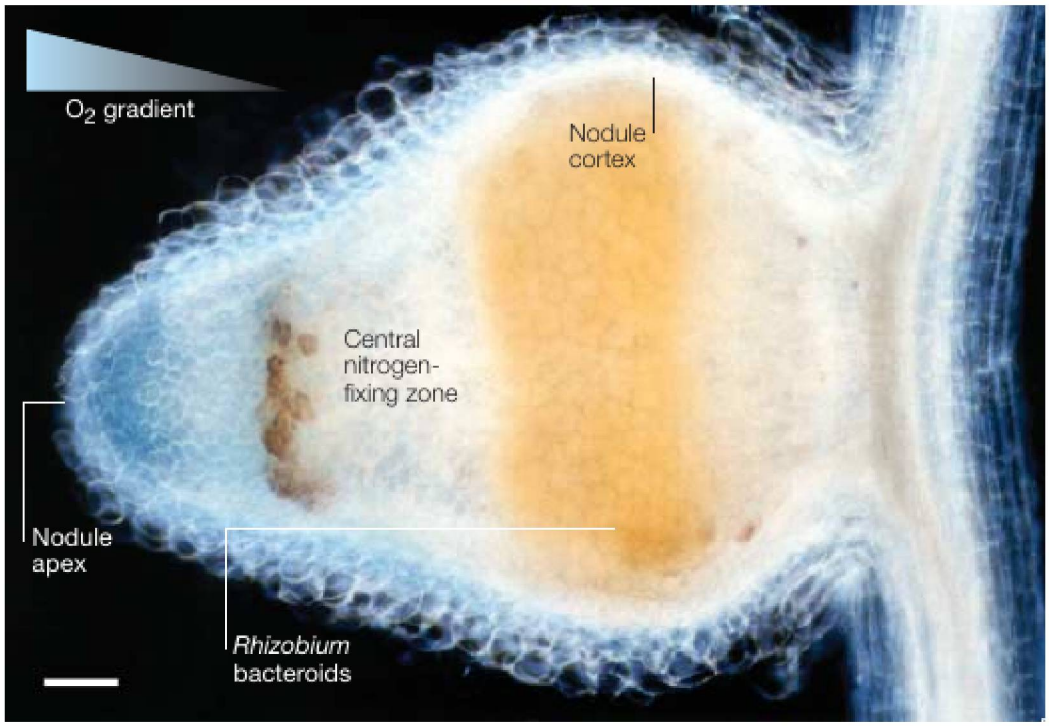


Processo de nodulação do diazotrofo *Sinorhizobium fredii* NGR234. Iniciado pelo estímulo recebido pela bactéria a partir dos flavonoides secretados pela planta, que utiliza o ativador transcricional NodD1 para promover a criação dos fatores Nod (produzidos por enzimas codificadas pelos genes nod), que serão reconhecidos pela planta. A proteína NodD1 também ativa o sistema de secreção tipo 3 (TTSS) que promoverá a secreção de proteínas necessárias para a nodulação. FONTE: Adaptado de Watson et al. (2022).

A atividade do complexo da nitrogenase é sensível à presença de oxigênio e por isso organismos de vida-livre e associativos exigem um cenário favorável que geralmente depende da capacidade do hospedeiro em manter a rizosfera com baixa concentração de oxigênio (Wagner et al., 2011). Já nos organismos diazotróficos simbióticos, o gradiente ideal de oxigênio é criado para favorecer o processo de

fixação biológica de nitrogênio e por isso é um mecanismo em geral mais eficaz (além de não promover competição entre diazotróficos e outros microrganismos). Este gradiente de oxigênio é facilitado na zona central pela concentração da proteína legmoglobina, que possui afinidade com as moléculas de oxigênio e as sequestram do ambiente. As bactérias podem então realizar a fixação de nitrogênio em um ambiente microaeróbico com alta concentração de nitrogênio, sem que a nitrogenase perca eficácia ou que a respiração seja comprometida (Dixon et al., 2004).

FIGURA 6 - *GRADIENTE DE OXIGÊNIO EM UM NÓDULO*



FONTE: Adaptado de Dixon et al. (*2004*).

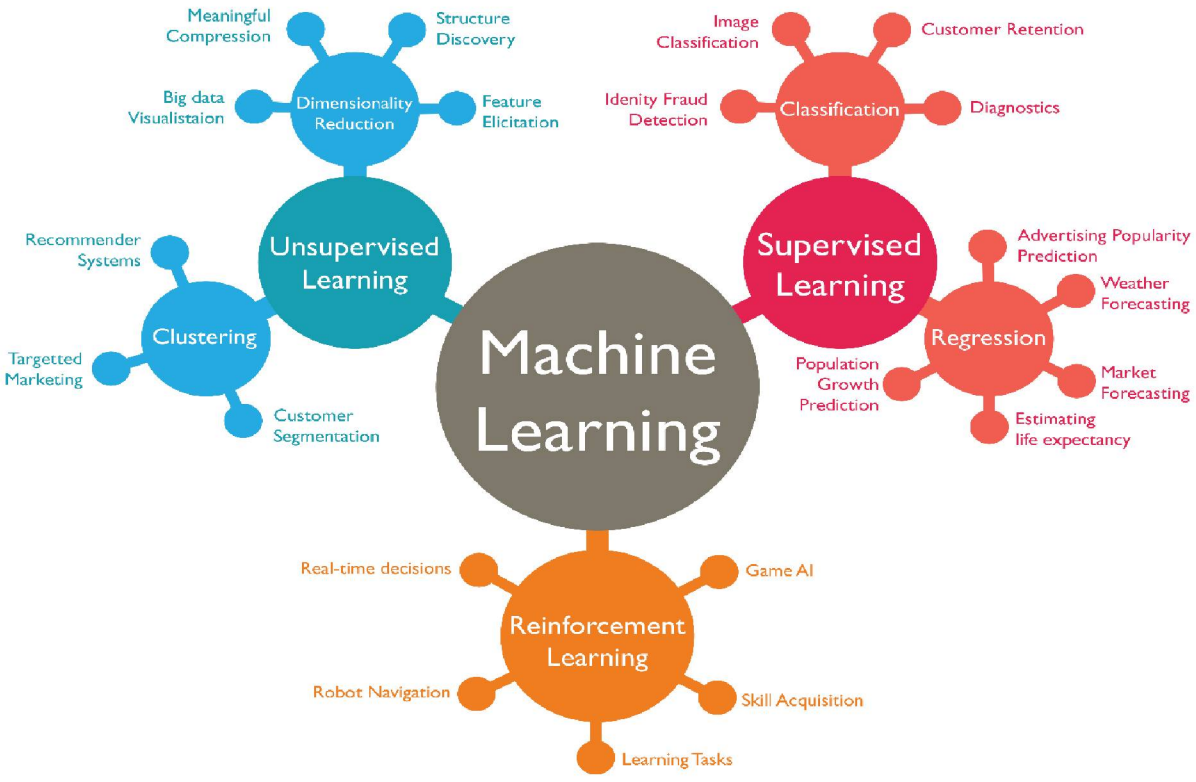
3.2 APRENDIZAGEM DE MÁQUINA (*MACHINE LEARNING*)

Aprendizagem de máquina (*machine learning*) refere-se a um conjunto de diferentes métodos capazes de gerar modelos para realizar predições ou identificar grupos semelhantes em dados. Estes métodos buscam imitar ou aproximar a capacidade humana em reconhecer padrões e aplicá-lo em escalas onde seria inviável para uma pessoa realizar análises, ou ainda quando se pretende automatizar o processo de análise para otimizar tempo e/ou obter padronização (Greener et al., 2022).

Cenários onde há grande quantidade de dados e/ou são muito complexos (por exemplo, muitos atributos para cada dado individual) a aprendizagem de máquina é útil para auxiliar a realizar análises. Os dados biológicos/genéticos frequentemente apresentam estas duas características e estão crescendo em quantidade e complexidade nas últimas décadas, e embora o uso de aprendizagem de máquina seja aplicado neles a décadas, somente nos últimos anos é que estão sendo explorados mais objetivamente, aplicando-se os métodos corretos para os cenários específicos (Greener et al., 2022).

Essencialmente há três categorias principais em aprendizagem de máquina: modelos supervisionados, modelos não-supervisionados e por reforço, e cada um possui diferentes métodos para criá-los. A figura 7 apresenta uma visão geral e alguns exemplos de aplicações. Neste trabalho exploramos modelos supervisionados (tanto regressão quanto classificação) e modelos não-supervisionados (clusterização e redução de dimensionalidade).

FIGURA 7 – Taxonomia das áreas de *machine learning* e exemplos



Taxonomia dos métodos de *machine learning* e exemplos de aplicações. FONTE: https://wordstream-files-prod.s3.amazonaws.com/s3fs-public/machine-learning.png

Os modelos supervisionados são utilizados quando há elementos conhecidos em um conjunto de dados, que geralmente foram classificados por especialistas humanos e obtidos através de validação ou experimentação, e deseja-se realizar prospecções no conjunto de dados atual ou em dados que serão obtidos futuramente, com objetivo de reconhecer novos elementos com as classificações determinadas (Greener et al., 2022).

No caso de modelos não-supervisionados não há uma classificação previamente determinada por especialistas humanos, e o objetivo é criar modelos capazes de explorar e descobrir padrões nos dados brutos disponíveis. Algumas vezes utiliza-se um modelo semi-supervisionado, onde alguns elementos conhecidos são adicionados ao conjunto desconhecido para aprimorar a performance do modelo (Greener et al., 2022).

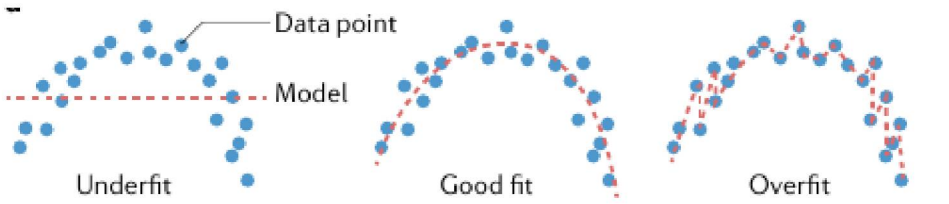
Mesmo escolhendo a categoria de aprendizagem de máquina correta para um objetivo, os resultados geralmente variam consideravelmente entre métodos e configurações aplicadas. Em geral as saídas (classificações supervisionadas ou não) dos modelos de aprendizagem de máquina nunca são ideais e divergem em algum grau do que se conhece das classificações verdadeiras (atribuídas por humanos). Esta divergência entre as classificações obtidas de um modelo e das classificações ideais é medida por algumas métricas que serão abordadas em outro capítulo deste documento, mas são definidas como ‘funções de perda’ (*loss functions*) ou ‘funções de custo’ (*cost functions*) (Greener et al., 2022).

As divergências entre as classificações do modelo de aprendizagem de máquina e as classificações reais definidas por especialistas humanos são utilizadas como referência para realizar melhoramentos. Todos os métodos supervisionados e não-supervisionados possuem diferentes parâmetros de configuração (hyperparametros) que alteram significativamente o desempenho do modelo criado, e as ‘funções de perda’ ou de ‘custo’ são utilizadas como guia para a escolha dos hyperparametros que melhor aderem aos tipos de dados (Greener et al., 2022).

Neste contexto, os modelos precisam ser ajustados em uma fase chamada de ‘treinamento’ e ‘otimização’ antes de serem utilizados para realizar predições. Porém há um problema em otimizar demais o modelo para aderir perfeitamente às classificações previamente conhecidas, eventualmente isso pode levar a um caso chamado de *Overfitting*, onde o modelo apresenta um resultado ótimo nos dados utilizados durante o treinamento, mas apresenta um resultado fraco em dados

novos, como observado na FIGURA 7. Nota-se que o modelo passa por todos os exemplos conhecidos e acaba ‘decorando’ esses pontos, portanto somente pontos extremamente próximos dos conhecidos seriam classificados corretamente. Como os problemas em geral são complexos isso é dificilmente aplicável em dados reais e o modelo apresentará resultados ruins. Um outro problema também relacionado é a falta de aderência durante a fase de treinamento, que pode levar ao chamado *Underfitting* (FIGURA 8), gerando um modelo que não consegue generalizar porque falha ao capturar as associações das classes. Isso pode ocorrer devido à pouca utilização de parâmetros e/ou o processo de treinamento incompleto. O que se deseja é o mais próximo do ‘*Good fit*’ demonstrado na FIGURA 8, um modelo que consegue definir as classes próximas à sua função matemática, sem ‘decorar’ todos os pontos, mas ajustando de uma forma mais adequada para acertar a maioria das classificações. Além disso, possuir poder de generalização mais elevado, resultando em melhores resultados em dados novos apresentados ao modelo treinado.

FIGURA 8 - *OVERFITTING, UNDERFITTING E GOOD FIT*

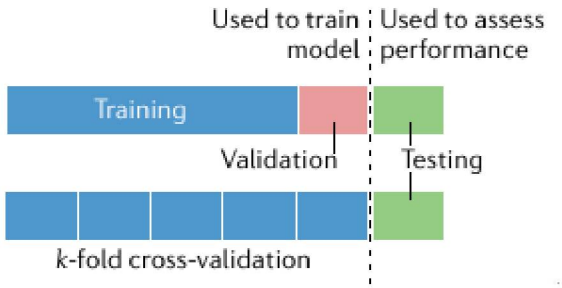


FONTE: Adaptado de Greener et al. (2022).

Ainda no processo de treinamento há o conceito de *bias* (viés) e variância a ser observado. O modelo pode tender a favorecer alguma solução particular em detrimento de outras de acordo com os dados e parâmetros apresentados. O modelo pode preferir, por exemplo, uma classe que está super-representada ou beneficiada por parâmetros. Ter um *bias* baixo é desejado porém isso geralmente eleva a variância, que é o quanto o modelo varia entre diferentes conjuntos ou sub-conjuntos de treinamento. Esse *trade-off* entre *bias* e variância ocorre por características dos algoritmos de aprendizagem de máquina, geralmente há um conflito entre a parametrização: ao se diminuir *bias*, observa-se uma maior variância e vice-versa. Controlar esse *trade-off* entre *bias* e variância é um dos pontos principais para se evitar *Overfitting* e *Underfitting* (Greener et al., 2022).

Durante a etapa de treinamento e otimização do modelo de aprendizagem de máquina supervisionado é necessário ter um conjunto de dados disponível para realizar a validação do modelo. Idealmente há um conjunto que será utilizado durante a fase de treinamento e validação, e um outro conjunto distinto que será utilizado para a fase de testes do modelo. Geralmente utiliza-se uma técnica chamada de *cross-validation* durante a fase de treinamento e validação, onde os dados são repartidos e passam pelo processo de treinamento e validação '*k*' vezes (ou *k-fold)* como observado na FIGURA 9, sendo que as divisões mais utilizadas são *k-fold* de dez (*10-fold*) e cinco (*5-fold*) (Greener et al., 2022).

FIGURA 9 - *DADOS DE TREINAMENTO E VALIDAÇÃO*



FONTE: Adaptado de Greener et al. (2022).

3.3 CRIAÇÃO E OTIMIZAÇÃO DE MODELOS DE APRENDIZAGEM DE MÁQUINA

O desenvolvimento de um modelo de aprendizagem de máquina requer as etapas: definição do conjunto de treinamento (e/ou de exploração no caso de aprendizagem não-supervisionada), definição dos atributos (preditores), escolha do algoritmo de aprendizagem de máquina alinhado ao objetivo, e a validação e otimização do algoritmo/modelo.

A definição de um conjunto de treinamento válido é um dos principais passos no processo de criação de um modelo de aprendizagem de máquina supervisionado. Os representantes das classes alvo devem ser os mais confiáveis possível pois serão as referências para a construção de todo o modelo, preferencialmente (e geralmente nos melhores modelos) são conferidos e definidos manualmente. A quantidade de representantes disponíveis também pode afetar a escolha do algoritmo mais adequado e também a qualidade do modelo final (Greener et al., 2022). No caso de aprendizagem não-supervisionada as classes são desconhecidas e, portanto, o conjunto será explorado e descoberto com auxílio do próprio modelo.

Especificamente em dados biológicos alguns problemas podem apresentar uma quantidade grande de representantes disponíveis enquanto outros podem apresentar quantidades muito pequenas. Por exemplo os bancos de dados públicos GenBank e UniProt apresentam uma quantidade de sequências proteicas abundante em volume e variedade, embora informações relativas à interação entre essas proteínas sejam muito mais difíceis de serem encontradas por se tratarem de um problema muito mais complexo envolvendo a mesma fonte – sequências proteicas. Esta característica encontrada em dados biológicos faz com que a exploração utilizando diferentes técnicas de aprendizagem de máquina seja justificada, pois elas conseguem entregar resultados consistentes dentro de um contexto complexo como este (Greener et al., 2022).

Conjuntos de dados com grande ou pequena (o valor exato depende do problema abordado) quantidade de representantes de cada classe podem impactar diretamente na qualidade e viabilidade do modelo de aprendizagem de máquina. E embora seja desejado uma quantidade suficiente de representantes das classes para a criação do modelo, nem sempre isso é possível devido a limitações da área de aplicação (Greener et al., 2022). Quando as classes possuem um número de representantes que não são aproximadamente iguais, o conjunto de treinamento está desbalanceado e isso pode afetar o modelo de aprendizagem de máquina. Classes super-representadas podem atrapalhar o treinamento e mascarar as métricas de validação como acurácia (Chawla et al., 2002). Para mitigar problemas de desbalanceamento de classes podemos utilizar algumas técnicas como a *Synthetic Minority Over-sampling Technique* (SMOTE). Esta técnica possui variações em implementação, porém todas consistem em criar dados sintéticos para as classes sub-representadas em um conjunto para que estejam igualmente balanceadas. O funcionamento basicamente consiste em conectar os pontos (representantes de uma classe) e seus ‘vizinhos’, e sortear pontos entre eles com algum grau de aleatoriedade (a depender de parâmetros ou diferentes implementações), estes novos pontos sintéticos se tornam novos representantes das classes sub-representadas, desta forma há grande probabilidade de serem próximas dos pontos reais, enquanto se mantém distintas dos já previamente definidos. Esta técnica pode viabilizar o uso de aprendizagem de máquina em situações em que os conjuntos são desbalanceados (Chawla et al., 2022).

Os conjuntos de treinamento também podem ser afetados por *outliers* presentes, que podem ser representantes de uma classe atípicos e muito diferentes dos demais. A presença de *outliers* pode introduzir viés ao modelo, onde, por exemplo, se cria uma especificidade para *outliers* e perde em generalidade, reduzindo a capacidade de reconhecimento de grande parte das classes canônicas. Algumas técnicas podem ser aplicadas para remoção dos *outliers* de um conjunto de treinamento, desde as estatísticas tradicionais ou a *Isolation Forest*, que é utilizado quando o conjunto apresenta muitos atributos para cada representante, comum no contexto de aprendizagem de máquina (Liu et al., 2008).

Um outro passo para a criação de um modelo de aprendizagem de máquina é a definição e seleção de atributos (ou preditores). Esta etapa consiste em extrair características dos representantes das classes que possam ser utilizadas para reconhecer padrões. Embora muitos atributos possam ser extraídos dos representantes, alguns deles podem não ter relevância para a classificação e sua presença pode apenas gerar ruído e prejudicar o desempenho do modelo. Portanto explorar quais atributos podem ter mais relevância na classificação e testá-los para verificar a performance do modelo a ser criado é um passo importante no processo de desenvolvimento. É possível explorar os atributos utilizando técnicas estatísticas como teste t de *student* ou correlação e ranqueá-los a partir de seu P-Value por exemplo, ou utilizar mecanismos como SVM-RFE, que escolhe quais atributos são mais relevantes para o treinamento de uma *Support Vector Machine* (SVM) (Guyon et al., 2002), ou, especificamente para dados biológicos, sigFeatures, um método que combina estatística e SVM para identificar os atributos mais relevantes em dados genéticos (Das et al., 2020).

A escolha do algoritmo para criação do modelo também é uma etapa que muitas vezes demanda tempo de processamento, pois diferentes problemas podem apresentar diferentes performances em cada algoritmo. Inevitavelmente vários algoritmos serão testados para averiguar qual apresenta o melhor resultado dentro do contexto do problema, muitas vezes esse processo é feito por força bruta – testando e verificando todos os possíveis. O que se aconselha nesse processo é verificar se é justificável o uso de aprendizagem de máquina dentro do contexto, pois se for identificado que um algoritmo tradicional e menos complexo puder resolver o problema (que não seja de aprendizagem de máquina), ele deverá ser utilizado, evitando uma complexidade desnecessária. Apesar de parecer intuitivo este fato,

alguns casos de aplicações desnecessárias de aprendizagem de máquina já foram observados no contexto biológico (Greener et al., 2022).

Definido o algoritmo a ser utilizado é necessário ajustar os parâmetros (*hyper parameters*) pois, todos os métodos de aprendizagem de máquina possuem ao menos alguns parâmetros a serem informados que afetarão o desempenho do modelo final. Em geral, a quantidade de possibilidades de combinações de *hyper parameters* pode até tender ao infinito (números sem limite máximo ou mínimo por exemplo). Como em muitos casos testar todas as possibilidades é impossível, há algumas técnicas para tentar predizer os parâmetros que apresentam o melhor resultado de treinamento, sendo a Otimização Baysiana (Mockus et al., 1978) a mais utilizada por não ser baseada em força bruta (testar uma quantidade aleatória ou sequencial por exemplo, sem heurística), e geralmente obtém uma performance melhor que outros métodos para encontrar o valor mais próximo de um ótimo global (Snoek et al., 2012).

3.4 CLUSTERIZAÇÃO EM DADOS GENÉTICOS

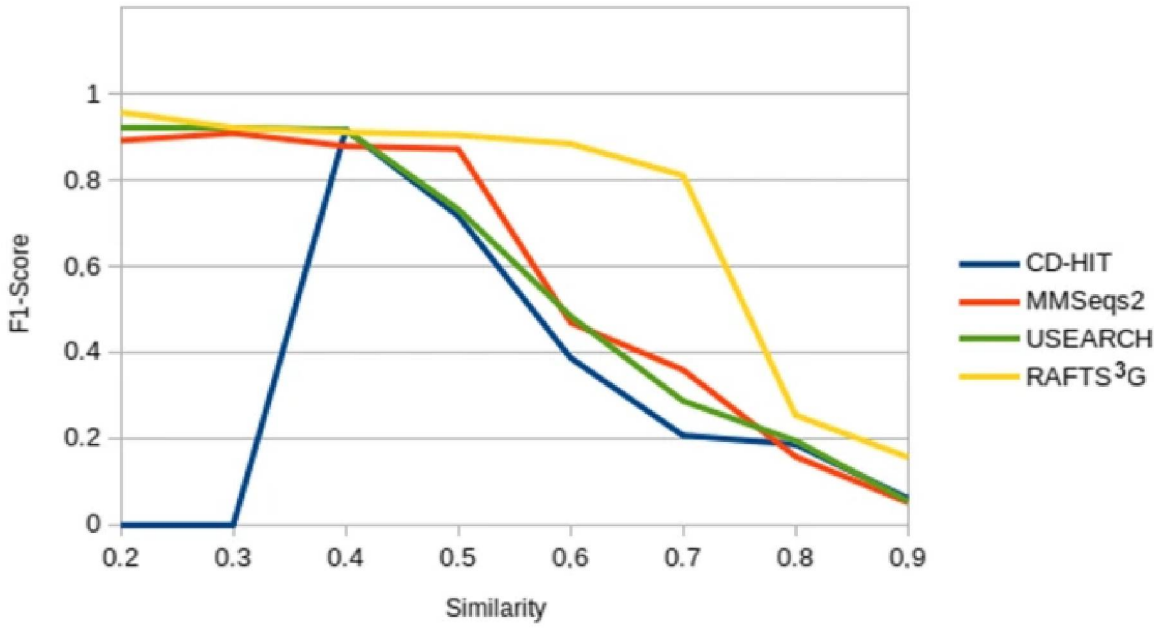
A clusterização é categorizada como um método de aprendizagem não-supervisionada, é utilizada quando a classificação de um conjunto é completamente ou parcialmente desconhecida. Os métodos agrupam instâncias de um conjunto utilizando critérios de similaridade, permitindo explorar e identificar classes de interesse (Xu et al., 2010).

Os dados genéticos são em sua maioria armazenados na forma de sequências de nucleotídeos ou aminoácidos, e alguns bancos de dados públicos como UniProt (Consortium, 2016) já disponibilizam dados agrupados por clusterização, permitindo que o pesquisador observe uma sequência de interesse dentro de um grupo (contexto). A clusterização é um dos principais métodos para realizar inferências de homologia entre sequências de aminoácidos (Rottger et al., 2013). O agrupamento de sequências similares permite a identificação de domínios conservados por exemplo, que são um dos critérios para inferir homólogos.

Existem ferramentas especializadas para clusterização de sequências de aminoácidos e que apresentam um resultado satisfatório para agrupar homólogos. Mas, assim como nos métodos tradicionais, possuem parâmetros informados manualmente pelo pesquisador que alteram a qualidade do resultado final. O maior

desafio em relação aos métodos não-supervisionados é a definição destes parâmetros, já que a maior parte dos grupos são desconhecidos previamente. Como a definição dos parâmetros acaba sendo feita arbitrariamente em dados genéticos e biomédicos (Rottger et al., 2013), a escolha de ferramentas que apresentem um resultado superior em diferentes configurações em um conjunto de difícil inferência como homólogos remotos, tende a ser a melhor estratégia. Conforme observado na FIGURA 10, a ferramenta RAFTS3G apresentou maior qualidade (F1-Score) nas clusterizações na base de dados ASTRAL/SCOPe em diferentes escolhas de parâmetros de similaridade. Esta base de dados contém sequências de proteínas homólogas que foram agrupadas experimentalmente, elas possuem similaridade em alinhamento menor que 0.5 mesmo possuindo atividade biológica semelhante, sugerindo que este é um valor para o critério que pode ser utilizado para agrupar homólogos remotos com maior sucesso conforme observado na figura.

FIGURA 10 - *FERRAMENTAS DE CLUSTERIZAÇÃO PARA SEQUÊNCIAS DE AMINOÁCIDOS*



FONTE: Adaptado de Nichio et al. (2019).

PARTE II

Artigo

4 ARTIGO

Artigo a ser submetido à revista IEEE/ACM Transactions on Computational Biology and Bioinformatics, fator de impacto 3.702.

Prospecção de diazotrofos noduladores *in silico*

RESUMO

A disponibilidade de nitrogênio é essencial para promover o crescimento vegetal na agricultura. O emprego de bactérias fixadoras de nitrogênio é uma alternativa sustentável ao uso de fertilizantes industriais que apresentam alto custo e causam danos ao ambiente. Entre os diazotrofos, o grupo dos rizóbios é caracterizado por formar nódulos nas raízes das plantas, uma associação que requer compatibilidade molecular. A identificação de novos organismos diazotróficos noduladores geralmente envolve um longo processo de isolamento de caracterização. A análise *in silico* de genomas bacterianos pode predizer novos candidatos a noduladores e fornecer informações sobre o processo de reconhecimento bactéria-planta hospedeira. O modelo SVM NodProspect foi desenvolvido para realizar predições *in silico* a partir do proteoma de organismos. Em um estudo de caso com genomas completos de alfa e betaproteobactérias, NodProspect forneceu uma lista de 303 organismos de uma variedade de gêneros, dos quais 133 foram confirmados como noduladores a partir de dados da literatura. As bactérias prospectadas foram clusterizadas e formaram quatro grupos caracterizados pela presença dos principais gêneros de noduladores, *Sinorhizobium*, *Bradyrhizobium*, *Rhizobium* e *Mesorhizobium*. Todos os demais gêneros foram agrupados com o gênero *Bradyrhizobium*. Em outra análise de clusters foram identificados 10 clusters proteicos com maior correlação com organismos noduladores e a presença destas proteínas foi avaliada em todos os 303 organismos. Os resultados obtidos sugerem que o modelo NodProspect é capaz de prospectar organismos noduladores *in silico* a partir de genomas completos fornecendo candidatos para ensaios de nodulação e promoção de crescimento vegetal. Além disso, outros organismos prospectados podem fornecer informações sobre grupos de genes envolvidos com o processo de nodulação.

KEYWORDS: Nodulação, fixação de nitrogênio, *machine learning*.

INTRODUÇÃO

O nitrogênio é um elemento crítico para o crescimento e desenvolvimento das plantas, mas sua disponibilidade no solo pode não ser suficiente. As culturas são frequentemente suplementadas com fertilizantes químicos nitrogenados, aumentando a produtividade, mas levando a problemas ambientais como a eutrofização de sistemas terrestres e aquáticos (Gruber; Galloway, 2008). Uma alternativa sustentável é a fixação biológica de nitrogênio, processo natural realizado por bactérias diazotróficas que reduzem o nitrogênio atmosférico a amônia (Timmusk et al., 2017). Essa reação é catalisada pela nitrogenase, um complexo metaloenzimático formado por duas proteínas, dinitrogenase e dinitrogenase redutase, codificadas pelos genes *nifDK* e *nifH*, respectivamente (Dean et al., 1993). Embora outros genes *nif* sejam necessários para a montagem e funcionamento da nitrogenase, e o conjunto desses genes varie

entre as espécies, a análise *in silico* sugeriu que a maioria dos diazotrofos tem *nifHDK* e *nifENB* como um conjunto mínimo de genes necessários para a fixação de nitrogênio (Dos Santos et al., 2012; Rubio e Ludden, 2008).

Os diazotrofos podem ser encontrados no solo ou associados a plantas. Bactérias diazotróficas associativas e endofíticas estão localizadas na superfície radicular ou colonizando os espaços intercelulares e sistemas vasculares das plantas hospedeiras (Carvalho et al., 2014). As bactérias noduladoras, por sua vez, estabelecem uma relação simbiótica com o hospedeiro, levando à formação de nódulos nas raízes da planta (Poole et al., 2018). Estas bactérias, denominadas coletivamente de rizóbios, incluem gêneros das classes Alphaproteobacteria (ex.: *Azorhizobium*, *Bradyrhizobium*, *Rhizobium*, *Mesorhizobium* e *Sinorhizobium*) e Betaproteobacteria (ex.: *Burkholderia* e *Cupriavidus*) e podem nodular um grande número de espécies da família Fabaceae (Gyaneshwar et al., 2011).

O processo de nodulação depende da compatibilidade entre leguminosa e rizóbio, definida por uma troca molecular. As bactérias são atraídas pelos flavonoides liberados pelas plantas, e a maioria dos rizóbios responde produzindo e secretando fatores Nod (lipoquitooligossacarídeos). O reconhecimento dos fatores Nod pela planta permite que os rizóbios invadam as células corticais da raiz, onde podem se diferenciar em bacteroides, a forma fixadora de nitrogênio dos rizóbios (Deakin; Broughton, 2009). Todo esse processo depende de várias etapas de produção e reconhecimento de moléculas, e essa complexa conversa molecular envolve os produtos de muitos genes bacterianos. Em geral, os flavonoides ativam uma proteína NodD, um ativador da transcrição que inicia uma cascata envolvendo vários genes, incluindo outros reguladores da transcrição. Entre os primeiros genes expressos estão *nodA*, *nodB* e *nodC*, cujos produtos são responsáveis pela síntese do *core* dos fatores Nod. Posteriormente, os produtos de outros genes *nod* e genes *nol* e *noe* adicionam substituintes como ácidos graxos e carboidratos ao *core* (PERRET et al., 2000). A maioria dos rizóbios possui *nodABC* e *nodD*, entretanto, a ocorrência de genes que adicionam grupos substituintes aos fatores Nod varia muito entre os organismos. Além disso, o reconhecimento rizóbio-hospedeiro também é afetado por outras moléculas apresentadas pela bactéria. Polissacarídeos extracelulares podem atuar na fixação à raiz e supressão da resposta imune do hospedeiro, e sistemas de secreção como T3SS ou T4SS podem entregar proteínas efetoras no citoplasma da planta (DOWNIE, 2010). Todas essas possibilidades fazem com que os conjuntos de genes simbióticos variem muito entre as espécies ou mesmo estirpes.

Embora muitos grupos de bactérias noduladoras sejam conhecidos, e alguns deles bem estudados, a prospecção de novos organismos com essa capacidade é um desafio. A simbiose não é permanente, ou seja, os nódulos podem ocorrer durante um determinado período. A rizosfera contém um grande número de organismos, e a discriminação de qual poderia ser um potencial candidato à simbiose de fixação de nitrogênio pode não ser muito simples. Nem todos os organismos diazotróficos conhecidos contêm alta homologia com os genes centrais conhecidos, e alguns organismos não diazotróficos podem ter alguns dos genes centrais.

Neste trabalho, os genes noduladores mais comuns foram usados para prospectar diazotrofos simbiontes, mas apenas organismos atuais bem conhecidos foram encontrados. Para melhorar este

processo foi desenvolvido NodProspect, um modelo SVM de aprendizado de máquina, para prospectar bactérias diazotróficas simbiontes usando genomas completos. Este modelo foi usado para prospectar organismos noduladores em Alpha e Betaproteobacteria disponíveis no banco de dados NCBI. O resultado foi uma lista de 303 estirpes de diversos gêneros incluindo os principais gêneros de rizóbios (*Bradyrhizobium*, *Rhizobium*, *Sinorhizobium* e *Mesorhizobium*). As bactérias prospectadas foram validadas através de dados da literatura e de submissão do genoma. Entre as bactérias prospectadas não confirmadas como noduladoras, muitas são boas candidatas para testes de promoção de crescimento vegetal. A partir de uma análise de clusters com todas as proteínas dos organismos diazotróficos noduladores, foram encontradas dez proteínas com maior correlação com estirpes noduladoras, incluindo proteínas Nif e Nod.

MATERIAIS E MÉTODOS

FONTES DE DADOS

Para a prospecção de bactérias noduladoras foram utilizados genomas completos e sequências de aminoácidos obtidos a partir de bancos de dados públicos e organizados em três bases de dados. A base **Nod/Nif** contém 58.729 sequências de aminoácidos anotadas como NodA, NodB, NodC, NifH, NifD e NifK obtidas dos bancos de dados Swiss-Prot e TrEMBL em outubro de 2020. A base **Bactérias CDSs** foi obtida a partir do NCBI Database Assembly e contém os arquivos de 51.544.181 sequências de amino ácidos no formato fasta, referentes à todas as bactérias que estavam registradas no banco de dados com genoma completo em novembro de 2020. A partir do NCBI Database Assembly também foram obtidos os arquivos com genomas completos de 3.072 bactérias das classes Alphaproteobacteria e Betaproteobacteria disponíveis em dezembro de 2020. Estes genomas formam a base de dados **Genomas Alfa/Beta**.

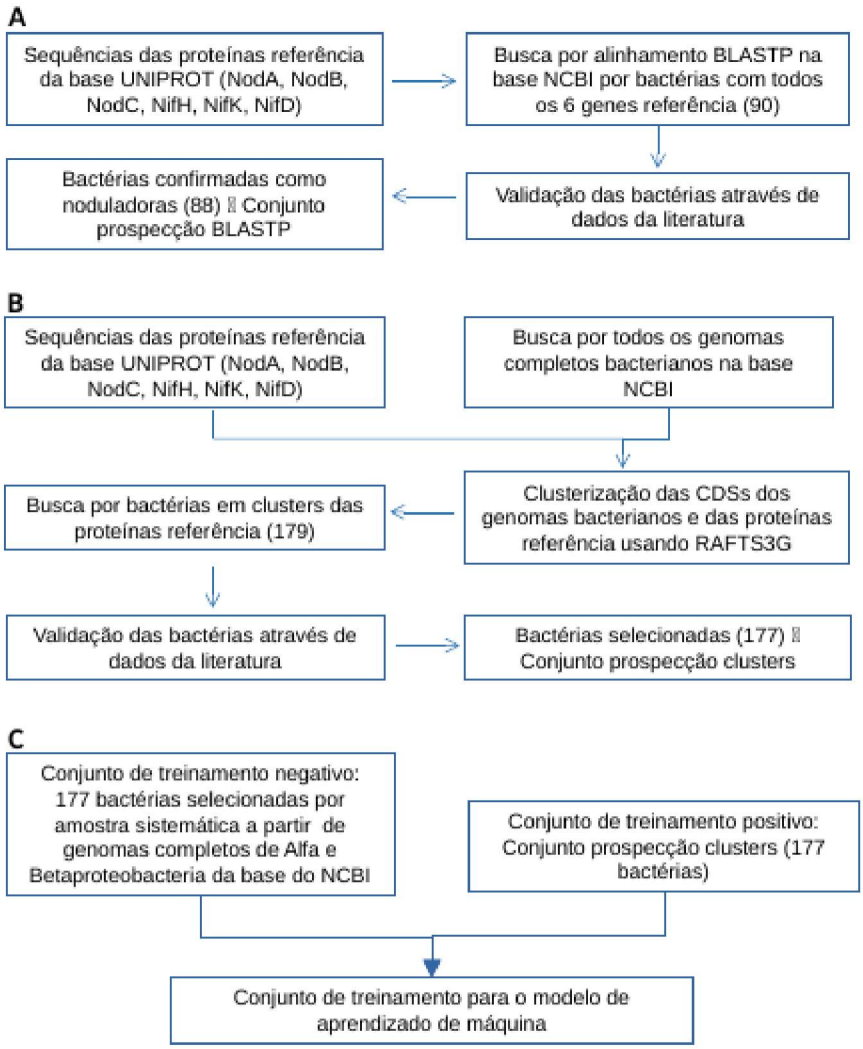
PROSPECÇÃO USANDO FERRAMENTAS DE ALINHAMENTO E CLUSTERIZAÇÃO

Bactérias diazotróficas noduladoras foram buscadas utilizando a ferramenta BLASTP. Para isso foi realizada a busca de proteínas com similaridade em alinhamento local com as sequências das proteínas definidas como referência (NodABC e NifHDK) na base de dados Bactérias CDSs. Foram selecionados como prováveis noduladores os organismos que apresentaram as sequências das proteínas NodABC e NifHDK simultaneamente (*expected threshold*: 0.5 e *word size*: 6). Esta busca retornou 90 organismos, dos quais 88 foram confirmados como organismos diazotróficos noduladores com base em dados da literatura. Este conjunto foi denominado Conjunto Prospecção BLASTP (Figura 1a).

A prospecção de bactérias noduladoras também foi realizada através de uma análise de clusters. Para isso todas as sequências de aminoácidos de regiões codificadoras (CDSs) da base Genomas Alfa/Beta e as sequências da base Nod/Nif foram agrupados. A clusterização foi realizada utilizando a ferramenta RAFTS3G (NICHIO et al., 2019) (critério de similaridade 0.5). As bactérias que

apresentaram sequências nos mesmos clusters que as proteínas referência NodABC e NifHDK simultaneamente foram consideradas potenciais noduladoras. Esta análise retornou 179 organismos, dos quais 124 foram confirmados em literatura como bactérias diazotróficas noduladoras, incluindo os 88 organismos do Conjunto Prospecção BLASTP e dois confirmados como não noduladores. O conjunto de 177 organismos foi definido como Conjunto Prospecção Clusters (Figura 1b).

FIGURA 1. Obtenção de conjuntos de dados



Fluxo de obtenção dos conjuntos de dados. A, Conjunto Prospecção BLASTP; B, Conjunto Prospecção Clusters; C, Conjunto de treinamento aprendizado de máquina.

DESENVOLVIMENTO DO MODELO DE APRENDIZADO DE MÁQUINA

Um modelo supervisionado de aprendizado de máquina foi criado para realizar a prospecção de bactérias diazotróficas noduladoras a partir de genomas completos. O conjunto de 177 organismos prospectados pela análise de clusters (Conjunto Prospecção Clusters) foi utilizado como conjunto positivo de treinamento. Para o conjunto de bactérias não noduladoras (negativos) foram selecionados 177

organismos no conjunto de dados da base Genomas Alfa/Beta (após a remoção dos 177 do conjunto positivo) utilizando o método estatístico amostra sistemática. Estes dois conjuntos foram definidos como o Conjunto de treinamento aprendizado de máquina (Figura 1c).

O desenvolvimento do Modelo de predição ocorreu em três passos: a extração de atributos para reconhecimento de padrão, a escolha do algoritmo de aprendizagem de máquina, e a otimização modelo escolhido. Como conjunto de atributos para reconhecimento de padrão do modelo foi utilizada uma representação vetorial do proteoma dos organismos da base Genomas Alfa/Beta. A representação vetorial das sequências foi feita utilizando a ferramenta SweeP com projeção de tamanho 1369 (DE PIERRI et al., 2020). Para a escolha do modelo de aprendizado de máquina foi realizada uma comparação de performance entre diferentes kernel de SVM, *Random Forest*, KNN, regressão logística, árvores de decisão, e redes neuronais artificiais que estão disponíveis no MATLAB *Classification Learner* (2017b). O modelo "Gaussian SVM" apresentou a melhor performance em *cross-validation* (10 fold). O modelo Gaussian SVM foi otimizado através de engenharia de atributos e ajuste de parâmetros. A validação do modelo foi realizada em *cross-validation* (10 fold) considerando as métricas F1-Score, *Precision*, *Recall*, variância em cross-*validation* e *bias* (Tabela suplementar 1).

Para minimizar o impacto das variações em relação ao conjunto negativo (não noduladores) foi realizado o treinamento, validação e otimização de cem modelos SVM variando a escolha dos organismos do conjunto negativo de forma exclusiva e aleatória. O conjunto de positivos utilizado foi de 183 organismos, formado pelo Conjunto de treinamento descrito anteriormente (177) e seis organismos prospectados pelo modelo SVM inicial. O ensemble criado com os cem modelos treinados foi denominado NodProspect. Foram considerados para análise os organismos com score ≥ 0.5.

VALIDAÇÃO E ANÁLISE DAS BACTÉRIAS PROSPECTADAS

As bactérias noduladoras prospectadas entre as classes alfa e betaproteobacteria utilizando o modelo NodProspect foram validadas através da busca de dados biológicos na literatura e nas informações de submissão de genomas no NCBI.

A análise correlação de proteínas de alfa e betaproteobacterias com o processo de nodulação foi realizada a partir da clusterização de todas as proteínas (12.399.511 sequências) da base Genomas Alfa/Beta. As sequências foram clusterizadas utilizando a ferramenta RAFTS3G com critério de similaridade de alinhamento local em 0.5. Os clusters foram correlacionados com os 183 organismos utilizados para o treinamento dos modelos *ensembled* e ranqueados segundo o valor de correlação de Pearson. As sequências das proteínas dos 100 clusters com maior correlação foram vetorizados com a ferramenta SweeP (projeção de tamanho 1369) e clusterizados em quatro grupos utilizando o algoritmo k-means utilizando o quadrado da distância euclidiana (MATLAB 2017b). Os quatro clusters foram representados em uma análise de componente principal (PCA - *Principal Component Analysis*) considerando os dois principais componentes.

Uma árvore filogenética foi criada a partir dos genomas completos vetorizados dos 3072 organismos da Alpha Beta Genomes Database considerando as distâncias euclidianas entre os vetores e o algoritmo *neighbor-joining* (MATLAB 2017b). Os organismos prospectados foram marcados na árvore conforme o seu grupo obtido pelo k-means.

Os dez clusters com maior correlação com o conjunto de treinamento foram utilizados como referência para a geração de *heat maps* com todos os 303 organismos prospectados por NodProspect. A presença ou ausência das proteínas dos clusters foi analisada em cada um dos quatro clusters descritos anteriormente.

RESULTADOS

PROSPECÇÃO DE BACTÉRIAS NODULADORAS UTILIZANDO FERRAMENTAS DE ALINHAMENTO E CLUSTERIZAÇÃO

Para prospectar potenciais bactérias diazotróficas noduladoras a partir de genomas depositados em bancos de dados, foram utilizados inicialmente dois métodos baseados na presença dos produtos de genes envolvidos com os processos de fixação de nitrogênio (*nifHDK*) e nodulação (*nodABC*).

No primeiro método a ferramenta BLASTP foi utilizada para a busca de organismos. As sequências das proteínas NodABC e NifHDK obtidas dos bancos de dados Swiss-Prot e TrEMBL (base Nod/Nif) foram utilizadas como *query* e a base Bactérias CDSs como banco de dados de busca. Esta análise retornou uma lista de 90 organismos que apresentam os genes que codificam para as seis proteínas utilizadas na busca, dos quais 88 foram confirmados como noduladores a partir de dados da literatura.

No segundo método, a busca de organismos noduladores foi realizada através da análise de clusters formados a partir das sequências das proteínas NodABC e NifHDK. A ferramenta RAFTS3G foi utilizada para a clusterização das sequências de aminoácidos preditas como regiões codificadoras (CDSs) da base de dados Genomas Alfa/Beta e as sequências das proteínas NodABC e NifHDK da base Nod/Nif. Esta análise resultou em uma lista de 179 organismos possíveis noduladores, incluindo os 88 organismos prospectados usando BLASTP. Através de dados da literatura, 124 bactérias foram confirmadas como noduladoras. Outras 53 bactérias prospectadas foram isoladas de nódulos, no entanto, não foram encontrados na literatura resultados que confirmem sua capacidade de nodular. Nestes casos, as bactérias não foram consideradas como noduladoras confirmadas, uma vez que os nódulos podem apresentar mais de uma espécie bacteriana e algumas destas são “não-rizóbios associados a nódulos” (MARTÍNEZ-HIDALGO and HIRSCH, 2017).

Os resultados obtidos a partir dos dois métodos apresentados sugerem que a utilização das sequências das proteínas NodABC e NifHDK como critério de busca pode ser eficiente para prospectar noduladores. Além disso, a prospecção através da análise de clusters resultou em uma lista com o dobro de organismos em relação à estratégia de alinhamento das sequências. No entanto, esta lista contém

basicamente noduladores conhecidos e gêneros bem caracterizados como simbióticos. Também deve se considerar que, apesar dos genes *nodABC* serem amplamente distribuídos entre os noduladores, sua presença não é obrigatória, uma vez que foram identificados simbiontes que não possuem estes genes.

PROSPECÇÃO DE BACTÉRIAS NODULADORAS UTILIZANDO UM MODELO DE APRENDIZADO DE MÁQUINA

Para melhorar a eficiência da prospecção de possíveis noduladores foi então desenvolvido um método baseado em aprendizagem de máquina. Esse modelo utiliza um algoritmo SVM para classificar as bactérias e prospectar noduladores a partir do conjunto total de sequências proteicas presentes nos organismos. Como conjunto de treinamento positivo foi utilizada a lista de 177 organismos gerada pela análise de clusters. Em uma análise de performance em *cross-validation*, o modelo SVM selecionado apresentou valores das métricas F1-Score de 0.93, Precisão de 0.93, Sensibilidade de 0.92, variância em *cross-validation* de 0.000521 e bias de 0.0674. Para realizar a prospecção o modelo requer arquivos no formato *Genomic GenBank format* (.gbff) com o genoma completo dos organismos a serem analisados e utiliza o proteoma para realizar a classificação. Como saída, o modelo indica se o organismo é um potencial diazotrofo nodulador ou não.

O modelo SVM foi então utilizado em um estudo de prospecção a partir do conjunto de genomas de Alpha e Betaproteobactérias da base Genomas Alfa/Beta. Como resultado foi obtida uma lista de 320 potenciais organismos diazotróficos noduladores. Para verificar se o treinamento não foi favorecido pelo conjunto de negativos utilizado, o treinamento foi repetido 100 vezes, variando o conjunto de negativos de forma aleatória, e criando 100 modelos ensemble. Esta análise resultou em uma lista de 303 organismos, sendo 275 em comum com a lista obtida a partir do modelo inicial. Este resultado sugere que os modelos são estáveis, já que a maioria dos organismos prospectados foi identificada da mesma forma a partir dos diferentes modelos usados. O modelo ensemble de prospecção foi denominado NodProspect.

AS BACTÉRIAS PROSPECTADAS SÃO UM GRUPO DIVERSIFICADO

A lista de bactérias prospectadas gerada a partir dos modelos ensemble foi validada manualmente através da busca de dados na literatura e informações de submissão de genomas (Tabela suplementar 2).

A partir desta análise, 131 bactérias foram confirmadas como noduladoras (grupo a), incluindo várias espécies dos gêneros *Bradyrhizobium*, *Mesorhizobium*, *Rhizobium* e *Sinorhizobium* (alphaproteobacteria) e *Paraburkholderia* (betaproteobacteria). Muitas destas bactérias estabelecem simbiose com plantas de interesse agrícola como *Glycine max* (soja), *Phaseolus vulgaris* (feijão comum) and *Vigna unguiculata* (feijão-caupi).

Entre os organismos confirmados como noduladores estão duas espécies de *Bradyrhizobium* que não possuem os genes *nod* normalmente encontrados em outros rizóbios simbióticos. *Bradyrhizobium*

sp. ORS 278, que apresenta apenas *nodB*, e *B. oligotrophicum* S58, que não possui o conjunto *nodABC*, formam nódulos efetivos em espécies de *Aeschynomene* (GIRAUD et al., 2007; OKUBO et al., 2013). A classificação destas espécies como noduladores sugere que o modelo NodProspect utiliza padrões biológicos mais complexos do que presença ou ausência dos genes *nif* e *nod*.

Entre as bactérias que não foram confirmadas como noduladoras, ou seja, não foram encontrados dados de testes de nodulação, estão 58 que foram isoladas de nódulos de diferentes hospedeiros (grupo b). Destas, 55 pertencem aos gêneros *Bradyrhizobium*, *Mesorhizobium*, *Rhizobium* e *Sinorhizobium* que possuem a maioria das espécies noduladoras conhecidas. Muitos destes organismos foram isolados e sequenciados, mas não possuem nenhum estudo quanto a nodulação, fixação de nitrogênio, ou promoção do crescimento vegetal. A classificação destes organismos como noduladores por NodProspect sugere que estas bactérias são boas candidatas para testes de nodulação.

Doze bactérias listadas são consideradas não-noduladoras, ou seja, bactérias que falharam em testes de nodulação (grupo d). Destas, nove foram isoladas de nódulos e variam quanto a presença de genes *nif* e *nod*. As estirpes de *Bradyrhizobium symbiodeficiens* não possuem os genes clássicos *nod* ou *nif* (Bromfield et al., 2020); *Bradyrhizobium cosmicum* S23321 não possui os genes *nod*, mas possui genes para fixação de nitrogênio, incluindo *nifHDK* e *fixABC* (Wasai-Hara et al., 2020). Por outro lado, *R. leguminosarum* Norway, que possui os genes *nodABC* e *nifHDK* além de outros genes *nif* e *nod*, induz apenas nódulos não efetivos em várias espécies de *Lotus* (LIANG et al., 2018).

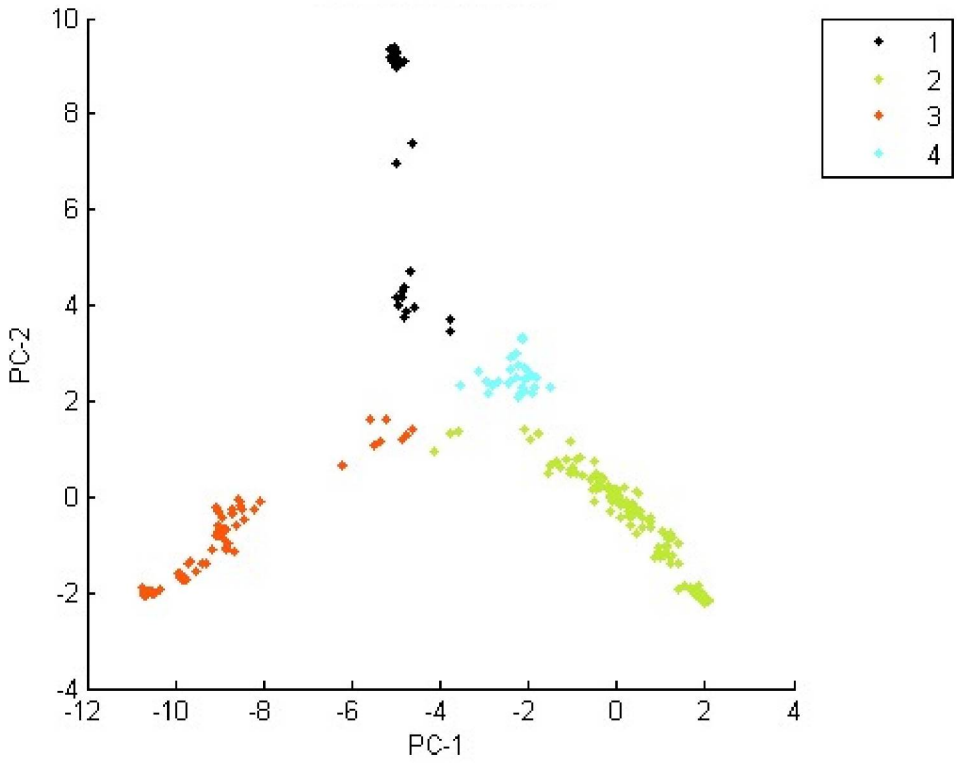
As demais bactérias não confirmadas como noduladoras foram isoladas de diversos ambientes ou não tem informação sobre a fonte de obtenção (grupo c). Este grupo é bastante heterogêneo e inclui, por exemplo, espécies representantes de gêneros que possuem espécies noduladoras como *Bradyrhizobium*, *Mesorhizobium*, *Rhizobium*, *Sinorhizobium, Neorhizobium, Burkholderia* e *Cupriavidus*; espécies PGPB (Plant growth-promoting bacteria) não noduladoras que se associam às plantas hospedeiras de outras formas, como algumas espécies do gênero *Azospirillum*, que permanecem na rizosfera; espécies isoladas de ambientes bastante diversos daqueles onde normalmente se encontram os noduladores e seus hospedeiros, como a bactéria termofílica *Chelatococcus daeguensis*, o isolado de *permafrost Caballeronia* SBC1 e o isolado de amostra clínica *Paracoccus yeei*. A variedade de gêneros prospectados com os diazotrofos noduladores sugere que parte do conjunto proteico que caracteriza este grupo de simbiontes é compartilhado por bactérias de diversos estilos de vida.

CLUSTERS DAS PROTEÍNAS COM MAIOR CORRELAÇÃO COM NODULAÇÃO

Para explorar as proteínas com maior correlação com o processo de nodulação, as proteínas obtidas dos genomas de Alpha e Betaproteobactérias da base Genomas Alfa/Beta foram clusterizadas e os clusters de proteínas com maior correlação com os organismos da lista de treinamento foram selecionados. As sequências das proteínas dos 100 clusters com maior correlação foram vetorizadas e clusterizadas. A análise de componente principal dos 4 clusters obtidos está representada na Figura 2 e os organismos presentes em cada grupo estão indicados na Tabela suplementar 2.

Os 4 principais gêneros de noduladores foram separados nos 4 grupos. O grupo 1 apresenta apenas espécies dos gêneros *Sinorhizobium* e *Ensifer*, sendo 32 noduladores confirmados e 3 isolados de nódulos; o grupo 3 apresenta espécies do gênero *Rhizobium*, sendo 37 noduladores confirmados, 29 isolados de nódulos e um não nodulador *R. leguminosarum* Norway (induz apenas nódulos não efetivos) (LIANG et al., 2018); o grupo 4 apresenta espécies de *Mesorhizobium*, sendo 29 noduladores confirmados, 4 isolados de nódulos e 1 sem informações sobre nodulação. Já o grupo 2 é bastante heterogêneo, e apresenta as espécies do gênero *Bradyrhizobium* e os demais gêneros prospectados, incluindo Betaproteobacteria. Com exceção de uma não-noduladora e uma bactéria sem informação sobre a nodulação, todas as demais classificadas como “c” e “d” estão neste grupo.

FIGURE 2. Análise de componente principal (PCA) das 303 bactérias prospectadas



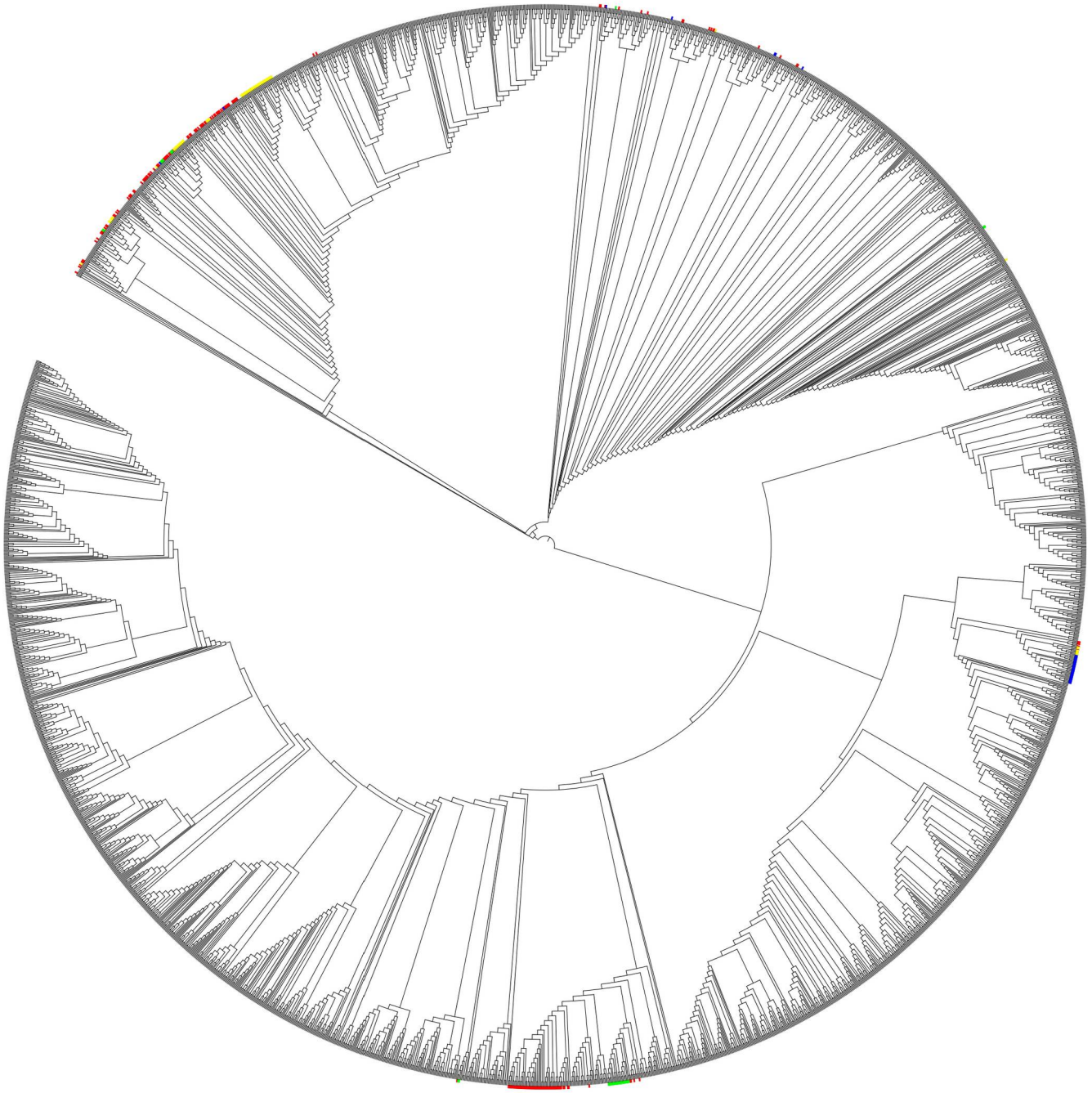
As bactérias prospectadas foram clusterizadas utilizando as sequências das proteínas dos 100 clusters com maior correlação com os organismos da lista de treinamento.

Uma análise filogenética foi realizada baseada nas sequências de aminoácidos preditas para os genomas de alfa e betaproteobactérias a partir da base Genomas Alfa/Beta (Figura 3). A maior parte dos 303 organismos prospectados se encontram em seis grandes blocos e apresentam uma distribuição semelhante aos 4 grupos da figura 1. As bactérias dos grupos 1, 3 e 4 formam um bloco cada, as bactérias do grupo 2 formam dois grandes blocos e o sexto bloco é formado por bactérias de todos os grupos. Vale

notar que a árvore filogenética foi construída a partir de todos os CDSs das bactérias, enquanto os quatro grupos da PCA foram obtidos a partir das sequências dos 100 clusters da análise de correlação.

Analisando as bactérias do grupo 2, o grupo mais heterogêneo, a maior parte das espécies de *Bradyrhizobium*, estão em um único ramo, incluindo as espécies não-noduladoras. Os demais gêneros que compõem o grupo 2 estão, em sua maioria, nos mesmos ramos ou em ramos próximos a outras bactérias prospectadas. Este resultado sugere novamente a existência de um padrão de sequências codificadoras nos diversos organismos prospectados.

FIGURE 3. Árvore filogenética de alfa e betaproteobactérias da base Genomas Alfa/Beta



As bactérias prospectadas estão marcadas de acordo com o cluster: Azul, cluster 1; vermelho, cluster 2; amarelo, cluster 3; verde, cluster 4. A árvore foi construída a partir dos CDSs dos genomas

completos vetorizados das bactérias. A seta indica o início dos organismos listados na Figura suplementar 1.

Para explorar as proteínas com maior correlação com o processo de nodulação, os dez clusters de proteínas com maior correlação com os organismos da lista de treinamento foram selecionados. As proteínas representativas destes clusters estão listadas na Tabela 1. Seis proteínas Nif e Nod estão nos clusters com maior correlação com as bactérias noduladoras: as proteínas NodA, NodB e NodC (síntese de fatores Nod), NodD (ativação transcricional dos genes *nod*), NifA (ativação transcricional dos genes *nif*) e NifN (síntese do cofator metálico da nitrogenase FeMoco). Completam a lista uma proteína hipotética, uma hidroxi-ácido desidrogenase, uma proteína com domínio SRPBCC e uma proteína com domínio adenilato/guanilato ciclase. As três últimas pertencem a grandes famílias de enzimas e, não há relato na literatura de função relacionada com a nodulação.

A presença das proteínas dos 10 clusters foi verificada em cada um dos 303 organismos prospectados e analisada nos 4 grupos do k-means (Figura Suplementar 1). Nos grupos 1 (*Sinorhizobium* e *Ensifer*) e 3 (*Rhizobium*), apenas dois organismos não apresentaram as proteínas Nif e Nod dos 10 clusters principais, estando provavelmente em outros clusters com proteínas homólogas. No grupo 4 (*Mesorhizobium*), a maioria dos organismos não apresenta as proteínas dos clusters 6 e 7. Já no grupo 2, noventa organismos não apresentaram nenhum dos 10 genes dos clusters e outros 26 apresentaram apenas 1 ou 2 genes. A grande maioria destes organismos (0, 1 ou 2 genes) pertencem a gêneros que não possuem espécies noduladoras conhecidas. Também estão incluídos neste grupo alguns não-noduladores como as estirpes de *B. symbiodeficiens, B. cosmicum* S23321 e *B. amphicarpaeae* 39S1MB. Os noduladores *Bradyrhizobium* sp. ORS 278 e *B. oligotrophicum* S58, que não possuem os genes *nod* clássicos, apresentaram apenas as sequências dos clusters NifA e NifN. Outra característica do grupo 2 é que apenas 8 organismos apresentaram os genes 8 e/ou 10. Apesar das proteínas dos clusters 6, 7, 8 e 10 não terem uma relação conhecida com a nodulação, a sua presença em tantos organismos confirmados como noduladores sugere que essas proteínas podem atuar neste processo. Além disso, a ausência das proteínas dos clusters 8 e 10 em todos os *Bradyrhizobium* do grupo 2 e das proteínas dos clusters 6 e 7 na maioria dos *Mesorhizobium* do grupo 4, sugere que estas proteínas possam ter um papel específico em diferentes gêneros de noduladores.

TABELA 1. Proteínas com maior correlação com organismos noduladores.

Cluster	Proteína	Correlação
1	NodA	0.9554
2	NodB	0.9521
3	NodC	0.9521
4	NodD	0.9493
5	NifA	0.7903
6	SRPBCC domain-containing protein	0.7670
7	Adenylate/guanylate cyclase domain-containing protein	0.7670
8	Hypothetical protein	0.7629
9	NifN	0.7622
10	Hydroxyacid dehydrogenase	0.7601

Os clusters estão ordenados pelo valor da Correlação de Pearson. A identificação de cada cluster foi feita pela sequência codificadora representativa do cluster.

DISCUSSÃO

A pesquisa de organismos diazotróficos noduladores é essencial para o desenvolvimento de uma agricultura sustentável. Antes do sequenciamento de nova geração, os novos isolados eram extensivamente caracterizados quanto a suas habilidades para promoção do crescimento vegetal e apenas alguns tinham seu genoma seria sequenciado. Atualmente, as metodologias em uso permitem que as novas bactérias isoladas tenham seu genoma sequenciado e caracterizado antes de quaisquer outros estudos. Assim, um número crescente de novas espécies ou estirpes bacterianas tem seus genomas depositados em bancos de dados, mas com pouco ou nenhum conhecimento de suas características. Muitas destas bactérias podem ter potencial aplicação como PGPB, seja como fixadores de nitrogênio, produtores de fitohormônios ou antagonistas de fitopatógenos. A prospecção de bactérias de interesse agrícola através da busca em genomas depositados em bancos pode acelerar o processo de conhecimento de novas PGPB.

O processo de nodulação envolve vários conjuntos de genes e estes conjuntos variam entre as estirpes bacterianas. A presença ou ausência de alguns genes, determina, por exemplo, quais espécies ou variedades de plantas podem ser noduladas por uma bactéria. A prospecção de organismos através dos métodos que se baseiam apenas nas informações já conhecidas sobre o processo biológico se demonstraram limitados. A busca de diazotrofos noduladores através de estratégias de alinhamento e clusterização resultou em conjuntos formados basicamente por noduladores já conhecidos. O modelo NodProspect, um modelo supervisionado de aprendizado de máquina, foi desenvolvido para prospectar possíveis noduladores a partir do proteoma de genomas completos depositados no NCBI. A prospecção de noduladores a partir dos genomas completos de 3072 alfa e betaproteobacteria resultou em um

conjunto de 303 organismos, sendo 133 noduladores já conhecidos. Este grupo inclui duas estirpes noduladoras que não contém os genes clássicos *nodABC*, sugerindo que NodProspect tem flexibilidade para reconhecer simbiontes mesmo quando estes genes canônicos estejam ausentes. Entre as bactérias não confirmadas, muitas são de gêneros que contem espécies noduladoras e são excelentes candidatos a ensaios de nodulação. Também foram prospectados vários gêneros que nunca foram descritos como possíveis noduladores, incluindo espécies de ambientes pouco favoráveis a simbiose com plantas. Análises de clusters entre os 303 organismos prospectados sugerem que as bactérias destes gêneros têm mais relação com as espécies de *Bradyrhizobium* do que com outros gêneros de noduladores. Algumas características importantes do gênero *Bradyrhizobium* são o grande número de espécies (estimado em 800) o grande tamanho dos genomas (a maioria entre 7 e 10 Mbp) e, a hipótese de que a nodulação surgiu em bradyrhizobia (Ormeño-Orrillo and Martínez-Romero, 2019). Além disso, este é o único gênero que contém bactérias fotossintéticas e rizóbios que não necessitam de fatores Nod para a nodulação (GIRAUD et al., 2007; OKUBO et al., 2013). Assim, a diversidade e amplitude do gênero *Bradyrhizobium* pode ter possibilitado o agrupamento deste gênero com tantos gêneros diversos presentes na nossa análise.

Em outra análise foram clusterizados todas as proteínas de genomas completos vetorizados de alfa e betaproteobacterias disponíveis no NCBI. A partir desse conjunto foram selecionados os dez clusters com maior correlação com as bactérias noduladoras do conjunto de treinamento. Seis destes clusters correspondem a proteínas Nod e Nif e estão presentes na maioria das 303 bactérias prospectadas. Entre os organismos que não apresentam proteínas destes clusters, estão algumas não noduladoras confirmadas e gêneros que parecem não ter relação com a nodulação. Alguns organismos noduladores também não apresentaram algumas ou todas as proteínas dos 10 clusters principais. Nestes casos, as proteínas Nif e Nod destes organismos foram agrupadas em outros clusters homólogos. Um exemplo é a presença dos genes *nif* em espécies de *Bradyrhizobium*. Em um estudo recente, Tao e colaboradores (2021) descreveram uma análise dos genes *nif* de mais de 300 espécies de *Bradyrhizobium* e observaram a divisão destes genes em dois clados principais, um composto principalmente por estirpes simbióticas e outro por estirpes fotossintéticas e de vida livre além de algumas simbióticas. Entre as bactérias prospectadas em nossa análise, com exceção de uma estirpe, todas as espécies noduladoras confirmadas de *Bradyrhizobium* apresentam as proteínas NifA e NifN dos clusters 5 e 9, respectivamente. Por outro lado, as espécies isoladas de solo *B*. sp. KBS0725 e *B*. sp. KBS0727, não apresentam as proteínas destes clusters, que são compartilhadas com os noduladores, e as proteínas Nif destas estirpes provavelmente foram agrupadas com as Nif de outros fixadores de vida livre.

CONCLUSÕES

O modelo de SVM NodProspect demonstrou dados estatísticos de performance consistentes, além de flexibilidade e robustez no estudo de caso utilizando o conjunto de Alpha e Betaproteobacteria. Os resultados obtidos por este novo método expandem as possibilidades de estudos futuros de genômica

comparativa para identificar os mecanismos biológicos envolvidos no processo de nodulação, assim como a exploração *in vivo* de potenciais organismos a serem utilizados para promover crescimento vegetal.

Sheet1

Model	Features	Data set	CV Accuracy	F1-Score	Precision	Recall	TP	FN	FP	TN	CV Bias (error)	CV Variance
Gaussian optimized	Sweep Not Scaled	Balanced	92.7000%	0.9257	0.9419	0.9101	91.0112%	8.9888%	5.6180%	94.3820%	0.0730	0.000014
Gaussian optimized	Sweep Scaled	Balanced	93.2600%	0.9322	0.9375	0.9270	92.6966%	7.3034%	6.1798%	93.8202%	0.0674	0.000521
Gaussian default	Sweep Scaled	Balanced	92.4200%	0.9226	0.9415	0.9045	90.4494%	9.5506%	5.6180%	94.3820%	0.0758	0.002200
Linear optimized	Sweep Scaled	Balanced	92.1300%	0.9213	0.9213	0.9213	92.1348%	7.8652%	7.8652%	92.1348%	0.0787	0.002400
Gaussian optimized	Sweep SVM-RFE Not Scaled	Balanced	92.1300%	0.9209	0.9261	0.9157	91.5730%	8.4270%	7.3034%	92.6966%	0.0787	0.001700
Gaussian optimized	Sweep SVM-RFE Scaled	Balanced	93.5400%	0.9348	0.9429	0.9270	92.6966%	7.3034%	5.6180%	94.3820%	0.0646	0.001200
Gaussian optimized	Sweep P-Value<=0.3 Not Scaled	Balanced	93.8200%	0.9375	0.9483	0.9270	92.6966%	7.3040%	5.0562%	94.9438%	0.0618	0.002400
Gaussian optimized	Sweep P-Value<=0.3 Scaled	Balanced	93.5400%	0.9345	0.9480	0.9213	92.1348%	7.8652%	5.0562%	94.9438%	0.0646	0.002300
Gaussian optimized	Sweep P-Value<=0.2 Not Scaled	Balanced	93.2600%	0.9318	0.9425	0.9213	92.1348%	7.8652%	5.6180%	94.3820%	0.0674	0.002100
Gaussian optimized	Sweep P-Value<=0.2 Scaled	Balanced	93.2600%	0.9314	0.9477	0.9157	91.5730%	8.4270%	5.0562%	94.9438%	0.0674	0.001600
Gaussian optimized	Sweep SigFeatures Bioconductor NotS	Balanced	93.2600%	0.9318	0.9425	0.9213	92.1348%	7.8652%	5.6180%	94.3820%	0.0674	0.001900
Gaussian optimized	Sweep SigFeatures Bioconductor Scale	Balanced	91.2900%	0.9122	0.9200	0.9045	90.4494%	9.5506%	7.8652%	92.1348%	0.0871	0.000930
Gaussian optimized	Sweep Scaled	Balanced – outliers removed	92.7500%	0.9272	0.9309	0.9235	92.3529%	7.6471%	6.8571%	93.1429%	0.0725	0.000756
Gaussian optimized	Sweep Not Scaled	Balanced – SMOTE2x	92.7000%	0.9270	0.9270	0.9270	92.6966%	7.3034%	7.3034%	92.6966%	0.0730	0.002200
Gaussian optimized	Sweep Scaled	Balanced – SMOTE2x	92.9800%	0.9288	0.9422	0.9157	91.5730%	8.4270%	5.6180%	94.3820%	0.0702	0.001700
Gaussian optimized	Sweep Not Scaled	ImBalanced – 1x Pos 2x Neg	95.1500%	0.9322	0.9815	0.8876	88.7640%	11.2360%	1.6760%	98.3240%	0.0485	0.001000
Gaussian optimized	Sweep Not Scaled	ImBalanced – 1x Pos 3x Neg	94.6800%	0.9079	0.9764	0.8483	84.8315%	15.1685%	2.0522%	97.9478%	0.0532	0.000562
Gaussian optimized	Sweep Not Scaled	ImBalanced – 1x Pos 4x Neg	95.7400%	0.9025	0.9867	0.8315	83.1461%	16.8539%	1.1204%	98.8796%	0.0426	0.000692
Gaussian optimized	Sweep Not Scaled	ImBalanced – 1x Pos 5x Neg	96.5400%	0.9097	0.9882	0.8427	84.2697%	15.7303%	1.0090%	98.9910%	0.0346	0.000177
Gaussian optimized	Sweep Not Scaled	ImBalanced – 1x Pos 6x Neg	96.8700%	0.9096	0.9879	0.8427	84.2697%	15.7303%	1.0280%	98.9720%	0.0313	0.000135
Gaussian optimized	Sweep Scaled	ImBalanced – 1x Pos 2x Neg	95.5200%	0.9385	0.9817	0.8989	89.8876%	10.1124%	1.6760%	98.3240%	0.0448	0.000858
Gaussian optimized	Sweep Scaled	ImBalanced – 1x Pos 3x Neg	94.8200%	0.9088	0.9785	0.8483	84.8315%	15.1685%	1.8657%	98.1343%	0.0518	0.000400
Gaussian optimized	Sweep Scaled	ImBalanced – 1x Pos 4x Neg	95.6300%	0.9018	0.9851	0.8315	83.1461%	16.8539%	1.2605%	98.7395%	0.0437	0.000488
Gaussian optimized	Sweep Scaled	ImBalanced – 1x Pos 5x Neg	96.2600%	0.8999	0.9880	0.8263	82.8543%	17.4157%	1.0090%	98.9910%	0.0374	0.000117
Gaussian optimized	Sweep Scaled	ImBalanced – 1x Pos 6x Neg	96.6300%	0.8966	0.9887	0.8202	82.0225%	17.9775%	0.9346%	99.0654%	0.0337	0.000125

Informations:	
Pos	Negative trainning set
Neg	Positive trainning set
TP	True positive
FN	False negative
FP	False positive
TN	True negative
CV	Cross-validation
Sweep	Amino acid sequence represented as numerical vector https://doi.org/10.1038/s41598-019-55627-4
Scaled	Standardizes the Sweep vectors using their corresponding weighted means and weighted standard deviations https://www.mathworks.com/help/stats/fitcsvm.html#bt7oo83-5
SVM-RFE	SVM Recursive Feature Elimination (RFE) applied to Sweep vectors https://doi.org/10.1023/A:1012487302797
SigFeatures Bioconductor	Significant feature selection method https://doi.org/10.3389/fgene.2020.00247
SMOTE2x	Synthetic Minority Over-sampling Technique (SMOTE) applied to double the number of positive class samples https://doi.org/10.1613/jair.953
Outliers removed	Outliers detected using Isolation Forest method https://doi.org/10.1109/ICDM.2008.17
Balanced	Data set created using the same number of positives and negatives organisms. Positives was made by validated organisms, and negatives by systematic sample from the Alpha Beta bacteria population
P-Value	Sweep vectors features selected by p-values, calculated using t-statistic, considering the two classes (nodulating and non nodulating)
Optimized	SVM Kernel optimized by bayesian optimization https://arxiv.org/abs/1206.2944
Default	Matlab default settings provided by the classification app in machine learning toolbox

Page 1

Tabela suplementar 1 – Performance dos métodos de aprendizagem de máquina.

Tabela suplementar 2 – Validação dos organismos prospectados.

Ensemble Score	Bacteria - Genome accession ID	Confirmed nodulation ability	Host examples
0.53	Achromobacter xylosoxidans A8 AccessionID NC 014640 1	a	*Cicer arietinum*
0.69	Achromobacter xylosoxidans AccessionID NZ CP045222 1	c	
0.62	Agrobacterium larrymoorei AccessionID NZ CP039691 1	c	
0.78	Agrobacterium radiobacter K84 AccessionID NC 011985 1	c	
0.78	Agrobacterium sp 33MFTa1 1 AccessionID NZ CP036358 1	c	
0.94	Agrobacterium sp H13 3 AccessionID NC 015508 1	c	
0.87	Agrobacterium sp RAC06 AccessionID NZ CP016499 1	c	
0.95	Agrobacterium tumefaciens Agrobacterium radiobacter Rhizobiu	c	
0.88	Agrobacterium tumefaciens Agrobacterium radiobacter Rhizobiu	c	
0.57	Agrobacterium vitis AccessionID NZ AP023279 1	c	
0.57	Agrobacterium vitis S4 AccessionID NC 011989 1	c	
0.88	Aminobacter sp MDW 2 AccessionID NZ CP060197 1	c	
0.83	Ancylobacter pratisalsi AccessionID NZ CP048630 1	c	
0.89	Aromatoleum aromaticum EbN1 AccessionID NC 006513 1	c	
0.87	Asticcacaulis excentricus AccessionID NZ AP018827 1	c	
1	Azorhizobium caulinodans ORS 571 AccessionID NC 009937 1	a	*Sesbania rostrata*
0.58	Azospirillum brasilense AccessionID NZ CP032339 1	c	
0.51	Azospirillum brasilense AccessionID NZ CP033318 1	c	
0.78	Azospirillum oryzae AccessionID NZ CP054619 1	c	
0.86	Blastomonas fulva AccessionID NZ CP020083 1	c	
0.77	Bosea sp ANAM02 AccessionID NZ AP022848 1	c	
0.94	Bosea sp F3 2 AccessionID NZ CP042331 1	c	
0.91	Bosea vaviloviae AccessionID NZ CP017147 1	b	
0.96	Bradyrhizobium amphicarpaeae AccessionID NZ CP029426 1	d	
1	Bradyrhizobium arachidis AccessionID NZ CP030050 1	a	*Arachis hypogaea* and *Lablab purpureus*
0.97	Bradyrhizobium cosmicum AccessionID NC 017082 1	d	
1	Bradyrhizobium diazoefficiens AccessionID NZ AP014685 1	a	*Glycine max*
1	Bradyrhizobium diazoefficiens AccessionID NZ AP022638 1	b	
1	Bradyrhizobium diazoefficiens AccessionID NZ AP022639 1	b	
1	Bradyrhizobium diazoefficiens AccessionID NZ AP022640 1	b	

1	Bradyrhizobium diazoefficiens AccessionID NZ AP022641 1	b	
1	Bradyrhizobium diazoefficiens AccessionID NZ CP013127 1	a	*Glycine max*
1	Bradyrhizobium diazoefficiens AccessionID NZ CP029603 2	b	
1	Bradyrhizobium diazoefficiens AccessionID NZ CP032617 1	a	soybean
1	Bradyrhizobium diazoefficiens AccessionID NZ CP050064 1	b	
1	Bradyrhizobium diazoefficiens AccessionID NZ CP055233 1	a	soybean
1	Bradyrhizobium diazoefficiens USDA 110 AccessionID NC 004463 1	a	*Glycine max*
1	Bradyrhizobium diazoefficiens USDA 110 AccessionID NZ CP011360	a	*Glycine max*
1	Bradyrhizobium elkanii USDA 61 AccessionID NZ AP013103 1	a	*Glycine max*
1	Bradyrhizobium genosp B AccessionID NZ CP061379 1	a	*Bossiaea ensata*
1	Bradyrhizobium genosp L AccessionID NZ CP061378 1	a	*Hardenbergia violacea*
1	Bradyrhizobium guangdongense AccessionID NZ CP030051 1	a	*Arachis hypogaea*
1	Bradyrhizobium guangdongense AccessionID NZ CP030057 1	a	*Arachis hypogaea*
1	Bradyrhizobium guangxiense AccessionID NZ CP022219 1	a	*Lablab purpureus* and *Aeschynomene indica*
1	Bradyrhizobium guangzhouense AccessionID NZ CP030053 1	a	*Arachis hypogaeaand* and *Lablab purpureus*
1	Bradyrhizobium icense AccessionID NZ CP016428 1	a	*Phaseolus lunatus*
1	Bradyrhizobium japonicum AccessionID NZ CP010313 1	a	*Glycine max*
1	Bradyrhizobium japonicum AccessionID NZ CP017637 1	a	*Glycine max*
1	Bradyrhizobium japonicum AccessionID NZ CP058354 1	b	
1	Bradyrhizobium japonicum USDA 6 AccessionID NC 017249 1	a	*Glycine max*
1	Bradyrhizobium oligotrophicum S58 AccessionID NC 020453 1	a	*Aeschynomene indica*
1	Bradyrhizobium ottawaense AccessionID NZ CP029425 1	a	*Glycine max*
1	Bradyrhizobium paxllaeri AccessionID NZ CP042968 1	a	*Phaseolus lunatus* and *Vigna unguiculata*
0.98	Bradyrhizobium sp 1 2017 AccessionID NZ CP050022 1	d	
1	Bradyrhizobium sp 6 2017 AccessionID NZ CP049289 1	b	
0.93	Bradyrhizobium sp AccessionID NZ LN901633 1	d	
1	Bradyrhizobium sp CCBAU 051011 AccessionID NZ CP022222 1	b	
1	Bradyrhizobium sp CCBAU 21365 AccessionID NZ CP030036 1	b	
1	Bradyrhizobium sp CCBAU 51753 AccessionID NZ CP030037 1	b	
1	Bradyrhizobium sp CCBAU 51765 AccessionID NZ CP030038 1	b	
1	Bradyrhizobium sp CCBAU 53338 AccessionID NZ CP030048 1	b	
1	Bradyrhizobium sp CCBAU 53340 AccessionID NZ CP030055 1	b	
1	Bradyrhizobium sp CCBAU 53351 AccessionID NZ CP030059 1	b	

1 Bradyrhizobium sp CCBAU 53421 AccessionID NZ CP030047 1	b	
1 Bradyrhizobium sp CCGE LA001 AccessionID NZ CP013949 1	b	
0.86 Bradyrhizobium sp KBS0725 AccessionID NZ CP042175 1	c	
0.86 Bradyrhizobium sp KBS0727 AccessionID NZ CP042176 1	c	
1 Bradyrhizobium sp LCT2 AccessionID NZ CP034432 1	b	
1 Bradyrhizobium sp ORS 278 AccessionID NC 009445 1	a	*Aeschynomene indica*
1 Bradyrhizobium sp ORS 285 AccessionID NZ LT859959 1	a	*Aeschynomene afraspera* and *A. indica*
0.99 Bradyrhizobium sp SG09 AccessionID NZ AP021854 1	c	
0.97 Bradyrhizobium sp TM102 AccessionID NZ AP021855 1	d	
0.83 Bradyrhizobium symbiodeficiens AccessionID NZ CP029427 1	d	
0.86 Bradyrhizobium symbiodeficiens AccessionID NZ CP041090 1	d	
0.85 Bradyrhizobium symbiodeficiens AccessionID NZ CP050065 1	d	
0.84 Bradyrhizobium symbiodeficiens AccessionID NZ CP050066 1	d	
1 Bradyrhizobium vignae AccessionID NZ LS398110 1	a	*Vigna unguiculata*, *Arachis hypogaea* and *Lablab purpu*
1 Bradyrhizobium zhanjiangense AccessionID NZ CP022221 1	a	*Arachis hypogaea*
0.62 Burkholderia sp JP2 270 AccessionID NZ CP029824 1	c	
0.65 Burkholderia sp THE68 AccessionID NZ AP022315 1	c	
0.91 Caballeronia sp SBC1 AccessionID NZ CP049156 1	c	
0.8 Caulobacter rhizosphaerae AccessionID NZ CP048815 1	c	
0.69 Caulobacter sp K31 AccessionID NC 010338 1	c	
0.87 Chelatococcus daeguensis AccessionID NZ CP018095 1	c	
0.52 Chelatococcus sp CO 6 AccessionID NZ CP012398 1	c	
0.85 Ciceribacter thiooxidans AccessionID NZ CP059896 1	c	
0.56 Cupriavidus malaysiensis AccessionID NZ CP017754 1	c	
0.92 Cupriavidus necator N 1 AccessionID NC 015726 1	c	
0.57 Cupriavidus oxalaticus AccessionID NZ CP032518 1	c	
0.9 Cupriavidus pauculus AccessionID NZ CP033969 1	c	
0.77 Cupriavidus pinatubonensis JMP134 AccessionID NC 007347 1	c	
0.69 Cupriavidus sp USMAHM13 AccessionID NZ CP017751 1	c	
0.73 Delftia lacustris AccessionID NZ CP065748 1	c	
0.96 Ensifer alkalisoli AccessionID NZ CP034909 1	a	*Sesbania cannabina*
0.9 Ensifer mexicanus AccessionID NZ CP041238 1	a	*Acacia cochliacantha* and *Phaseolus vulgaris*
0.95 Ensifer sojae CCBAU 05684 AccessionID NZ CP023067 1	a	*Glycine max*, *G. soja* and *Vigna unguiculata*

0.9 Haematospirillum jordaniae AccessionID NZ CP014525 1	c	
0.65 Indioceanicola profundi AccessionID NZ CP030126 1	c	
1 Iodobacter sp H11R3 AccessionID NZ CP034433 1	c	
0.71 Ketogulonicigenium vulgare AccessionID NZ CP016592 1	c	
0.73 Labrenzia sp THAF82 AccessionID NZ CP045354 1	c	
0.75 Labrys sp KNU 23 AccessionID NZ CP043489 1	c	
0.78 Leisingera methylohalidivorans DSM 14336 AccessionID NC 023135	c	
0.91 Litoreibacter sp LN3S51 AccessionID NZ CP042261 1	c	
0.51 Magnetospirillum sp XM 1 AccessionID NZ LN997848 1	c	
0.78 Maritalea myrionectae AccessionID NZ CP021330 1	c	
0.54 Marivivens sp JLT3646 AccessionID NZ CP018572 1	c	
0.59 Martelella mediterranea DSM 17316 AccessionID NZ CP020330 1	c	
0.86 Martelella sp NC18 AccessionID NZ CP054858 1	c	
0.93 Martelella sp NC20 AccessionID NZ CP054861 1	c	
0.85 Massilia putida AccessionID NZ CP019038 1	c	
1 Mesorhizobium amorphae CCNWGS0123 AccessionID NZ CP015318	a	*Robinia pseudoacacia*
1 Mesorhizobium australicum WSM2073 AccessionID NC 019973 1	a	*Biserrula pelecinus* and *Astragalus membranaceus*
1 Mesorhizobium ciceri AccessionID NZ CP015062 1	a	*Cicer arietinum*
1 Mesorhizobium ciceri biovar biserrulae AccessionID NZ CP015064 1	a	*Biserrula pelecinus*
1 Mesorhizobium ciceri biovar biserrulae WSM1271 AccessionID NC	a	*Biserrula pelecinus* and *Astragalus membranaceus*
1 Mesorhizobium erdmanii AccessionID NZ CP033361 1	a	*Lotus corniculatus* and *L. pedunculatus*
1 Mesorhizobium huakuii AccessionID NZ CP050296 1	a	*Oxytropis kamtschatica*
1 Mesorhizobium japonicum AccessionID NZ CP051773 1	a	*Lotus corniculatus*
1 Mesorhizobium japonicum AccessionID NZ CP052769 1	a	*Lotus corniculatus*
1 Mesorhizobium japonicum AccessionID NZ CP052770 1	a	*Lotus japonicum*
1 Mesorhizobium japonicum MAFF 303099 AccessionID NC 002678 2	a	*Lotus japonicum*
1 Mesorhizobium japonicum R7A AccessionID NZ CP033366 1	a	*Lotus corniculatus*
1 Mesorhizobium japonicum R7A AccessionID NZ CP051772 1	a	*Lotus corniculatus*
0.97 Mesorhizobium jarvisii AccessionID NZ CP033507 1	a	*Lotus corniculatus* and *L. pedunculatus*
1 Mesorhizobium loti AccessionID NZ CP033334 1	a	*Lotus pedunculatus*
0.97 Mesorhizobium loti AccessionID NZ CP033368 1	a	*Lotus corniculatus*
1 Mesorhizobium loti AccessionID NZ CP050293 1	a	*Oxytropis kamtschatica*
1 Mesorhizobium loti NZP2037 AccessionID NZ CP016079 1	a	*Lotus divaricatus*

1 Mesorhizobium loti R88b AccessionID NZ CP033367 1	a	*Lotus corniculatus*
1 Mesorhizobium opportunistum WSM2075 AccessionID NC 015675	a	*Lotus peregrinus and Macroptilium atropurpureum*
0.91 Mesorhizobium sp 8 AccessionID NZ CP040914 1	c	
1 Mesorhizobium sp AA22 AccessionID NZ CP048406 1	b	
1 Mesorhizobium sp M1B F Ca ET 045 04 1 1 AccessionID NZ CP0344	a	*Cicer arietinum*
1 Mesorhizobium sp M1D F Ca ET 043 01 1 1 AccessionID NZ CP0344	a	*Cicer arietinum*
1 Mesorhizobium sp M1E F Ca ET 045 02 1 1 AccessionID NZ CP0344	a	*Cicer arietinum*
1 Mesorhizobium sp M2A F Ca ET 043 02 1 1 AccessionID NZ CP0344	a	*Cicer arietinum*
1 Mesorhizobium sp M2A F Ca ET 043 05 1 1 AccessionID NZ CP0344	a	*Cicer arietinum*
1 Mesorhizobium sp M2A F Ca ET 046 03 2 1 AccessionID NZ CP0344	a	*Cicer arietinum*
1 Mesorhizobium sp M3A F Ca ET 080 04 2 1 AccessionID NZ CP0344	a	*Cicer arietinum*
1 Mesorhizobium sp M9A F Ca ET 002 03 1 2 AccessionID NZ CP0344	a	*Cicer arietinum*
0.97 Mesorhizobium sp NZP2077 AccessionID NZ CP033362 1	c	
1 Mesorhizobium sp NZP2077 AccessionID NZ CP051293 1	b	
1 Mesorhizobium sp NZP2234 AccessionID NZ CP033364 1	b	
1 Mesorhizobium sp NZP2298 AccessionID NZ CP033365 1	b	
0.95 Mesorhizobium sp Pch S AccessionID NZ CP029562 1	c	
1 Mesorhizobium sp WSM1497 AccessionID NZ CP021070 1	a	*Biserrula pelecinus*
0.55 Methylosinus trichosporium OB3b AccessionID NZ CP023737 1	c	
1 Neorhizobium galegae bv officinalis bv officinalis str HAMBI 1141	a	*Galega officinalis*
1 Neorhizobium galegae bv orientalis str HAMBI 540 AccessionID NZ	a	*Galega orientalis*
0.87 Neorhizobium sp NCHU2750 AccessionID NZ CP030827 1	c	
0.65 Neorhizobium sp SOG26 AccessionID NZ CP025512 1	c	
0.61 Nitratireductor sp OM 1 AccessionID NZ CP029208 1	c	
0.65 Nitratireductor sp SY7 AccessionID NZ CP042301 2	c	
0.55 Nitrobacter hamburgensis X14 AccessionID NC 007964 1	c	
0.5 Nitrospirillum amazonense CBAmc AccessionID NZ CP022110 1	c	
0.75 Novosphingobium aromaticivorans DSM 12444 AccessionID NC 007	c	
0.54 Ochrobactrum anthropi AccessionID NZ CP044970 1	c	
0.89 Pannonibacter phragmitetus AccessionID NZ CP013068 1	c	
1 Paraburkholderia atlantica AccessionID NC 014117 1	b	
0.77 Paraburkholderia caledonica AccessionID NZ CP024905 1	c	
0.62 Paraburkholderia caribensis AccessionID NZ CP013102 1	c	

0.69 Paraburkholderia caribensis AccessionID NZ CP026101 1	c	
1 Paraburkholderia phenoliruptrix BR3459a AccessionID NC 018695 1	a	*Mimosa flocculosa*
1 Paraburkholderia phymatum STM815 AccessionID NC 010622 1	a	*Phaseolus vulgaris*
0.75 Paraburkholderia sp 7Q K02 AccessionID NZ CP046909 1	c	
0.95 Paraburkholderia sp PGU19 AccessionID NZ AP023179 1	c	
1 Paraburkholderia sprentiae WSM5005 AccessionID NZ CP017561 1	a	*Lebeckia sepiaria* and *Lebeckia ambigua*
0.57 Paraburkholderia terrae AccessionID NZ CP026111 1	c	
0.84 Paracoccus aminophilus JCM 7686 AccessionID NC 022041 1	c	
0.87 Paracoccus aminovorans AccessionID NZ LN832559 1	c	
0.91 Paracoccus denitrificans AccessionID NZ CP035090 1	c	
0.97 Paracoccus kondratievae AccessionID NZ CP045072 1	c	
0.9 Paracoccus pantotrophus AccessionID NZ CP044426 1	c	
0.52 Paracoccus yeei AccessionID NZ CP020442 2	c	
0.92 Paracoccus yeei AccessionID NZ CP044081 1	c	
0.81 Phenylobacterium zucineum HLK1 AccessionID NC 011144 1	c	
0.93 Phyllobacterium sp 628 AccessionID NZ CP050301 1	d	
1 Phyllobacterium zundukense AccessionID NZ CP017940 1	b	
1 Rhizobium acidisoli AccessionID NZ CP034998 1	b	
1 Rhizobium esperanzae AccessionID NZ CP013500 1	a	*Phaseolus vulgaris*
1 Rhizobium etli 8C 3 AccessionID NZ CP017241 1	a	*Phaseolus vulgaris*
1 Rhizobium etli AccessionID NZ CP020906 1	a	*Phaseolus vulgaris*
1 Rhizobium etli bv mimosae str Mim1 AccessionID NC 021905 1	a	*Phaseolus vulgaris*
1 Rhizobium etli bv phaseoli str IE4803 AccessionID NZ CP007641 1	a	*Phaseolus vulgaris*
1 Rhizobium etli CFN 42 AccessionID NC 007761 1	a	*Phaseolus vulgaris*
1 Rhizobium etli CIAT 652 AccessionID NC 010994 1	a	*Phaseolus vulgaris*
1 Rhizobium gallicum AccessionID NZ CP017101 1	a	*Phaseolus vulgaris*
1 Rhizobium grahamii AccessionID NZ CP043498 1	b	
0.92 Rhizobium hidalgonense AccessionID NZ CP054027 1	b	
1 Rhizobium indicum AccessionID NZ CP054021 1	b	
1 Rhizobium indicum AccessionID NZ CP054031 1	b	
1 Rhizobium jaguaris AccessionID NZ CP032694 1	a	*Calliandra grandiflora*
1 Rhizobium leguminosarum AccessionID NZ CP016286 1	b	
1 Rhizobium leguminosarum AccessionID NZ CP018228 1	b	

0.96 Rhizobium leguminosarum AccessionID NZ CP025012 1	d	
1 Rhizobium leguminosarum AccessionID NZ CP030760 1	a	*Trifolium pratense*
1 Rhizobium leguminosarum bv trifolii AccessionID NZ CP050080 1	b	
1 Rhizobium leguminosarum bv trifolii AccessionID NZ CP050085 1	b	
1 Rhizobium leguminosarum bv trifolii AccessionID NZ CP050091 1	b	
1 Rhizobium leguminosarum bv trifolii AccessionID NZ CP050097 1	b	
1 Rhizobium leguminosarum bv trifolii AccessionID NZ CP050103 1	b	
1 Rhizobium leguminosarum bv trifolii AccessionID NZ CP050108 1	b	
1 Rhizobium leguminosarum bv trifolii AccessionID NZ CP053439 1	a	*Trifolium repens*
1 Rhizobium leguminosarum bv trifolii CB782 AccessionID NZ CP007(	a	*Trifolium semipilosum*
1 Rhizobium leguminosarum bv trifolii TA1 AccessionID NZ CP053205	a	*Trifolium spumosum*
1 Rhizobium leguminosarum bv trifolii WSM1325 AccessionID NC 01	a	*Trifolium subterraneum*
1 Rhizobium leguminosarum bv trifolii WSM1689 AccessionID NZ CP	a	*Trifolium uniflorum*
1 Rhizobium leguminosarum bv trifolii WSM2304 AccessionID NC 01	a	*Trifolium polymorphum*
1 Rhizobium leguminosarum bv viciae 248 AccessionID NZ CP048280	a	*Vicia sativa* and *V. hirsuta*
1 Rhizobium leguminosarum bv viciae 3841 AccessionID NC 008380 1	a	*Pisum sativum* and *Vicia cracca*
0.57 Rhizobium leguminosarum bv viciae AccessionID NZ CP022665 1	b	
1 Rhizobium leguminosarum bv viciae AccessionID NZ CP025509 1	a	*Pisum sativum* L. cv. Frisson
1 Rhizobium phaseoli AccessionID NZ CP013522 1	b	
1 Rhizobium phaseoli AccessionID NZ CP013527 1	b	
1 Rhizobium phaseoli AccessionID NZ CP013532 1	b	
1 Rhizobium phaseoli AccessionID NZ CP013537 1	b	
1 Rhizobium phaseoli AccessionID NZ CP013542 1	b	
1 Rhizobium phaseoli AccessionID NZ CP013547 1	b	
1 Rhizobium phaseoli AccessionID NZ CP013552 1	b	
1 Rhizobium phaseoli AccessionID NZ CP013557 1	b	
1 Rhizobium phaseoli AccessionID NZ CP013563 1	b	
1 Rhizobium phaseoli AccessionID NZ CP013568 1	b	
1 Rhizobium phaseoli AccessionID NZ CP013574 1	b	
1 Rhizobium phaseoli AccessionID NZ CP013580 1	b	
1 Rhizobium phaseoli AccessionID NZ CP013585 1	b	
1 Rhizobium phaseoli AccessionID NZ CP064931 1	b	
1 Rhizobium phaseoli Brasil 5 AccessionID NZ CP020896 1	a	*Phaseolus vulgaris*

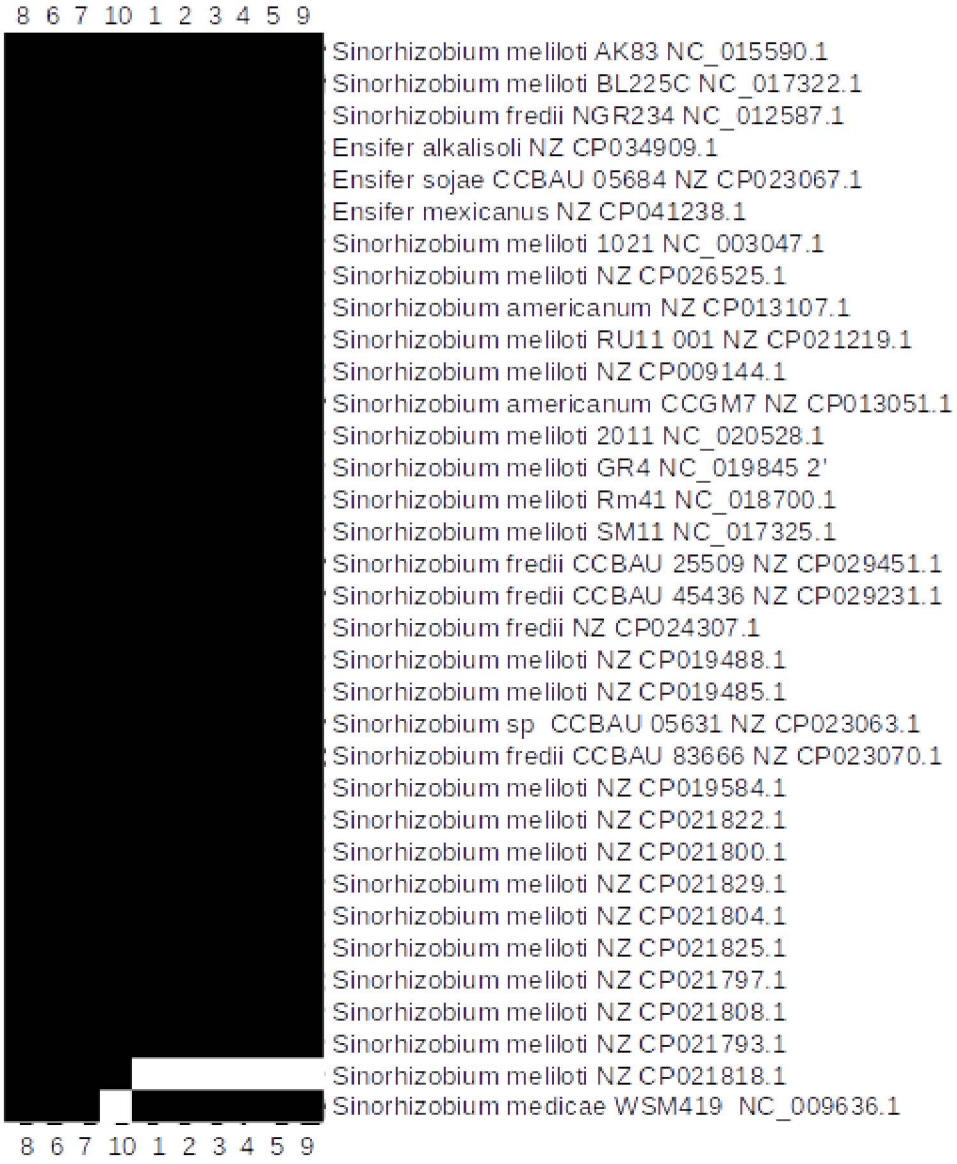
0.91	Rhizobium pseudoryzae AccessionID NZ CP049241 1	c	
1	Rhizobium pusense AccessionID NC 022535 1	a	*Sesbania cannabina*
0.67	Rhizobium pusense AccessionID NZ CP053856 1	c	
1	Rhizobium sp 007 AccessionID NZ CP064187 1	b	
1	Rhizobium sp 11515TR AccessionID NZ CP022998 1	a	*Centrosema pubescens*, *Phaseolus vulgaris* and *Vigna ur*
0.85	Rhizobium sp ACO 34A AccessionID NZ CP021371 1	c	
1	Rhizobium sp CCGE531 AccessionID NZ CP032684 1	a	*Phaseolus vulgaris* and *Leucaena leucocephala*
1	Rhizobium sp CCGE532 AccessionID NZ CP032689 1	a	*Phaseolus vulgaris* and *Leucaena leucocephala*
1	Rhizobium sp CIAT894 AccessionID NZ CP020947 1	a	*Phaseolus vulgaris*
1	Rhizobium sp IE4771 AccessionID NZ CP006986 1	a	*Phaseolus vulgaris*
1	Rhizobium sp Kim5 AccessionID NZ CP021124 1	a	*Phaseolus vulgaris*
1	Rhizobium sp N113 AccessionID NZ CP013517 1	a	*Phaseolus vulgaris*
1	Rhizobium sp N1314 AccessionID NZ CP013511 1	a	*Phaseolus vulgaris*
1	Rhizobium sp N1341 AccessionID NZ CP013505 1	a	*Phaseolus vulgaris*
1	Rhizobium sp N324 AccessionID NZ CP013630 1	a	*Phaseolus vulgaris*
1	Rhizobium sp N621 AccessionID NZ CP013495 1	a	*Phaseolus vulgaris*
1	Rhizobium sp N6212 AccessionID NZ CP013490 1	a	*Phaseolus vulgaris*
1	Rhizobium sp N731 AccessionID NZ CP013601 1	a	*Phaseolus vulgaris*
1	Rhizobium sp N741 AccessionID NZ CP013595 1	a	*Phaseolus vulgaris*
1	Rhizobium sp N871 AccessionID NZ CP013590 1	a	*Phaseolus vulgaris*
0.89	Rhizobium sp NIBRBAC000502774 AccessionID NZ CP041204 1	c	
1	Rhizobium sp NXC14 AccessionID NZ CP021030 1	a	*Phaseolus vulgaris*
1	Rhizobium sp NXC24 AccessionID NZ CP024311 1	b	
0.84	Rhizobium sp S41 AccessionID NZ CP016320 1	c	
1	Rhizobium sp TAL182 AccessionID NZ CP021024 1	a	*Phaseolus vulgaris*
0.83	Rhizobium sp WL3 AccessionID NZ CP042826 1	c	
0.73	Rhizobium sp Y9 AccessionID NZ CP017999 1	c	
1	Rhizobium tropici CIAT 899 AccessionID NC 020059 1	a	*Acacia nilotica* and *Phaseolus vulgaris*
0.93	Rhodobacter blasticus AccessionID NZ CP020470 1	c	
0.62	Rhodopseudomonas palustris Rhodopseudomonas rutila Accessio	c	
0.56	Rhodopseudomonas palustris Rhodopseudomonas rutila Accessio	c	
0.56	Rhodopseudomonas palustris Rhodopseudomonas rutila Accessio	c	
0.93	Rhodopseudomonas palustris BisA53 AccessionID NC 008435 1	c	

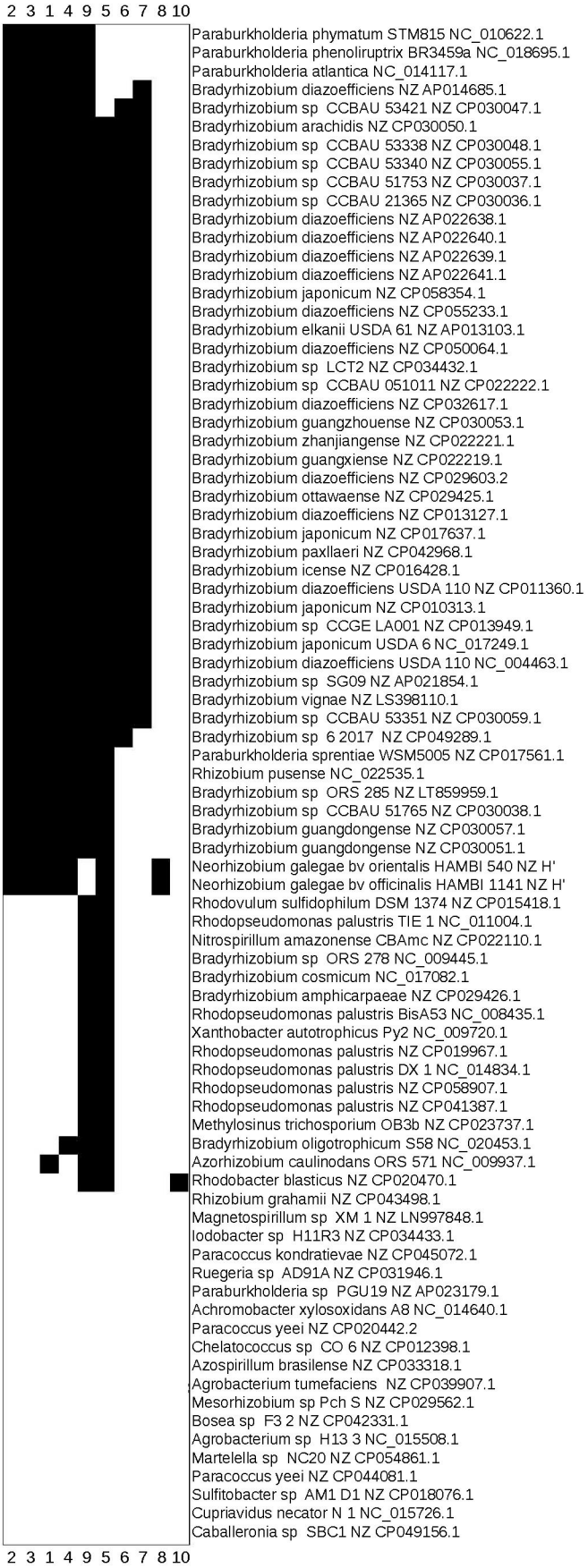
0.61 Rhodopseudomonas palustris DX 1 AccessionID NC 014834 1	c	
0.53 Rhodopseudomonas palustris TIE 1 AccessionID NC 011004 1	c	
0.53 Rhodovulum sulfidophilum DSM 1374 AccessionID NZ CP015418 1	c	
0.97 Ruegeria sp AD91A AccessionID NZ CP031946 1	c	
1 Sinorhizobium americanum AccessionID NZ CP013107 1	a	*Leucaena leucocephala* and *Phaseolus vulgaris*
1 Sinorhizobium americanum CCGM7 AccessionID NZ CP013051 1	a	*Phaseolus vulgaris* and *Medicago truncatula*
1 Sinorhizobium fredii AccessionID NZ CP024307 1	a	*Glycine max*
1 Sinorhizobium fredii CCBAU 25509 AccessionID NZ CP029451 1	a	*Glycine soja*
1 Sinorhizobium fredii CCBAU 45436 AccessionID NZ CP029231 1	a	*Glycine soja*
1 Sinorhizobium fredii CCBAU 83666 AccessionID NZ CP023070 1	b	
1 Sinorhizobium fredii NGR234 AccessionID NC 012587 1	a	*Phaseolus vulgaris* - 112 legume genera
1 Sinorhizobium medicae WSM419 Ensifer medicae WSM419 Access	a	*Marchatia polymorpha* and *M. arabica*
1 Sinorhizobium meliloti 1021 AccessionID NC 003047 1	a	*Medicago sativa*
1 Sinorhizobium meliloti 2011 AccessionID NC 020528 1	a	*Medicago sativa*
1 Sinorhizobium meliloti AccessionID NZ CP009144 1	b	
1 Sinorhizobium meliloti AccessionID NZ CP019485 1	a	*Medicago sativa*
1 Sinorhizobium meliloti AccessionID NZ CP019488 1	a	*Medicago sativa*
1 Sinorhizobium meliloti AccessionID NZ CP019584 1	a	*Medicago truncatula*
1 Sinorhizobium meliloti AccessionID NZ CP021793 1	a	*Medicago sativa*
1 Sinorhizobium meliloti AccessionID NZ CP021797 1	b	
1 Sinorhizobium meliloti AccessionID NZ CP021800 1	a	*Medicago truncatula*
1 Sinorhizobium meliloti AccessionID NZ CP021804 1	a	*Medicago sativa*
1 Sinorhizobium meliloti AccessionID NZ CP021808 1	a	*Medicago sativa*
1 Sinorhizobium meliloti AccessionID NZ CP021818 1	a	*Medicago truncatula*
1 Sinorhizobium meliloti AccessionID NZ CP021822 1	a	*Medicago sativa*
1 Sinorhizobium meliloti AccessionID NZ CP021825 1	a	*Medicago sativa*
1 Sinorhizobium meliloti AccessionID NZ CP021829 1	a	*Medicago sativa*
1 Sinorhizobium meliloti AccessionID NZ CP026525 1	a	*Medicago sativa*
0.97 Sinorhizobium meliloti AK83 Ensifer meliloti AK83 AccessionID NC	a	*Medicago sativa*
1 Sinorhizobium meliloti BL225C Ensifer meliloti BL225C AccessionI	a	*Medicago sativa*
1 Sinorhizobium meliloti GR4 AccessionID NC 019845 2	a	*Medicago sativa*
1 Sinorhizobium meliloti Rm41 AccessionID NC 018700 1	a	*Medicago truncatula*
1 Sinorhizobium meliloti RU11 001 AccessionID NZ CP021219 1	a	*Medicago sativa*

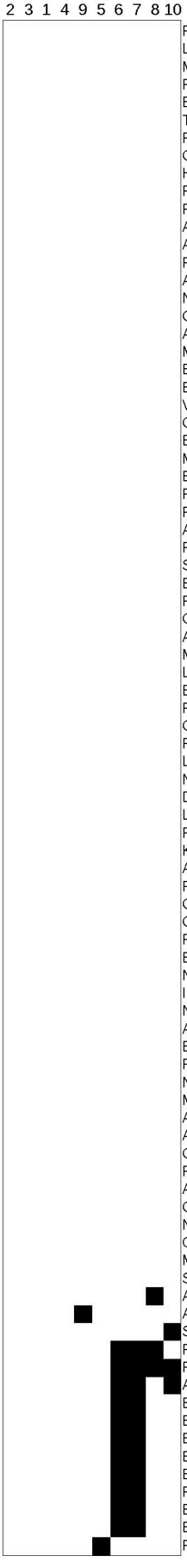
1 Sinorhizobium meliloti SM11 AccessionID NC 017325 1	a	*Medicago sativa*
1 Sinorhizobium sp CCBAU 05631 AccessionID NZ CP023063 1	a	*Glycine max*
0.7 Sinorhizobium sp RAC02 AccessionID NZ CP016450 1	c	
0.53 Sphingobium yanoikuyae AccessionID NZ CP020925 1	c	
0.83 Sphingosinicella sp BN140058 AccessionID NZ CP035501 1	c	
0.92 Sulfitobacter sp AM1 D1 AccessionID NZ CP018076 1	c	
0.9 Tardiphaga robiniae AccessionID NZ CP050292 1	d	
0.86 Variovorax sp PMC12 AccessionID NZ CP027773 1	c	
0.87 Xanthobacter autotrophicus Py2 AccessionID NC 009720 1	c	

a, Nodulation ability was confirmed in nodulation tests.
b, Isolated from nodules however nodulation ability was not tested.
c, No data found about nodulation.
d, The strain failed in nodulation tests.

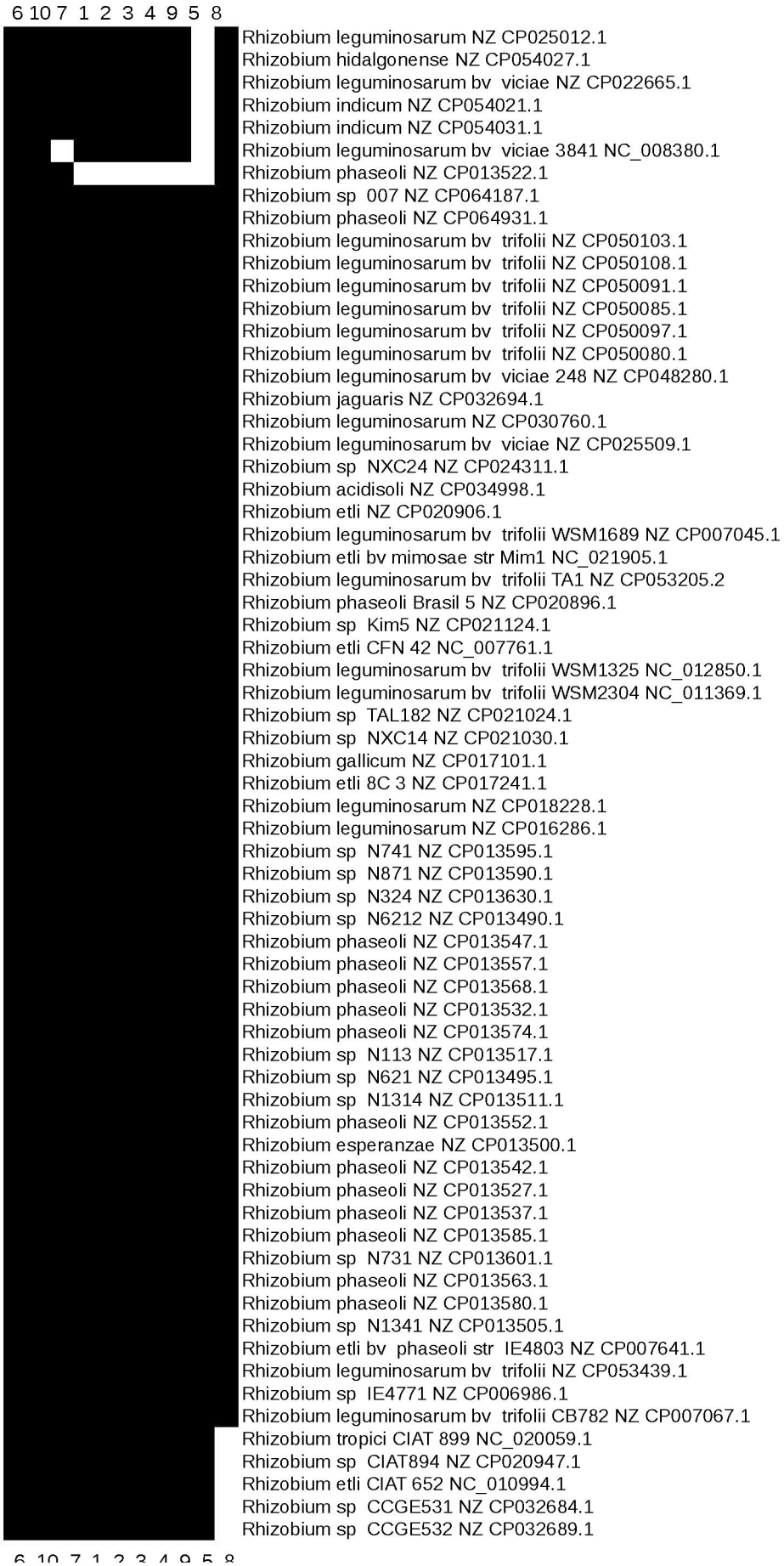
Figura suplementar 1 – Presença das proteínas dos 10 clusters de maior correlação com a nodulação nas bactérias prospectadas (em cada um dos quatro clusters).

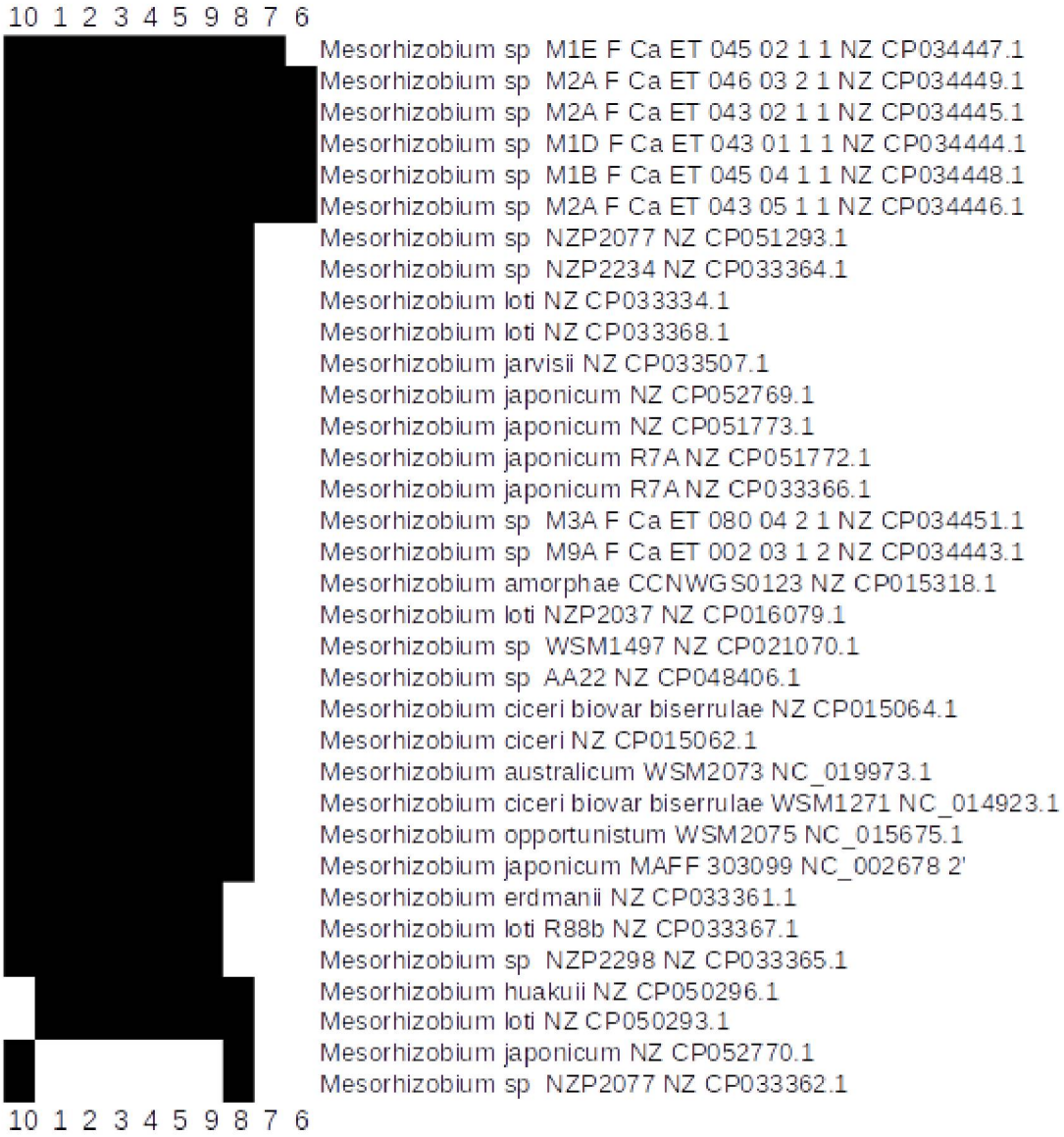



2 3 1 4 9 5 6 7 8 10
Paraburkholderia phymatum STM815 NC_010622.1
Paraburkholderia phenoliruptrix BR3459a NC_018695.1
Paraburkholderia atlantica NC_014117.1
Bradyrhizobium diazoefficiens NZ AP014685.1
Bradyrhizobium sp CCBAU 53421 NZ CP030047.1
Bradyrhizobium arachidis NZ CP030050.1
Bradyrhizobium sp CCBAU 53338 NZ CP030048.1
Bradyrhizobium sp CCBAU 53340 NZ CP030055.1
Bradyrhizobium sp CCBAU 51753 NZ CP030037.1
Bradyrhizobium sp CCBAU 21365 NZ CP030036.1
Bradyrhizobium diazoefficiens NZ AP022638.1
Bradyrhizobium diazoefficiens NZ AP022640.1
Bradyrhizobium diazoefficiens NZ AP022639.1
Bradyrhizobium diazoefficiens NZ AP022641.1
Bradyrhizobium japonicum NZ CP058354.1
Bradyrhizobium diazoefficiens NZ CP055233.1
Bradyrhizobium elkanii USDA 61 NZ AP013103.1
Bradyrhizobium diazoefficiens NZ CP050064.1
Bradyrhizobium sp LCT2 NZ CP034432.1
Bradyrhizobium sp CCBAU 051011 NZ CP022222.1
Bradyrhizobium diazoefficiens NZ CP032617.1
Bradyrhizobium guangzhouense NZ CP030053.1
Bradyrhizobium zhanjiangense NZ CP022221.1
Bradyrhizobium guangxiense NZ CP022219.1
Bradyrhizobium diazoefficiens NZ CP029603.2
Bradyrhizobium ottawaense NZ CP029425.1
Bradyrhizobium diazoefficiens NZ CP013127.1
Bradyrhizobium japonicum NZ CP017637.1
Bradyrhizobium paxllaeri NZ CP042968.1
Bradyrhizobium icense NZ CP016428.1
Bradyrhizobium diazoefficiens USDA 110 NZ CP011360.1
Bradyrhizobium japonicum NZ CP010313.1
Bradyrhizobium sp CCGE LA001 NZ CP013949.1
Bradyrhizobium japonicum USDA 6 NC_017249.1
Bradyrhizobium diazoefficiens USDA 110 NC_004463.1
Bradyrhizobium sp SG09 NZ AP021854.1
Bradyrhizobium vignae NZ LS398110.1
Bradyrhizobium sp CCBAU 53351 NZ CP030059.1
Bradyrhizobium sp 6 2017 NZ CP049289.1
Paraburkholderia sprentiae WSM5005 NZ CP017561.1
Rhizobium pusense NC_022535.1
Bradyrhizobium sp ORS 285 NZ LT859959.1
Bradyrhizobium sp CCBAU 51765 NZ CP030038.1
Bradyrhizobium guangdongense NZ CP030057.1
Bradyrhizobium guangdongense NZ CP030051.1
Neorhizobium galegae bv orientalis HAMBI 540 NZ H'
Neorhizobium galegae bv officinalis HAMBI 1141 NZ H'
Rhodovulum sulfidophilum DSM 1374 NZ CP015418.1
Rhodopseudomonas palustris TIE 1 NC_011004.1
Nitrospirillum amazonense CBAmc NZ CP022110.1
Bradyrhizobium sp ORS 278 NC_009445.1
Bradyrhizobium cosmicum NC_017082.1
Bradyrhizobium amphicarpaeae NZ CP029426.1
Rhodopseudomonas palustris BisA53 NC_008435.1
Xanthobacter autotrophicus Py2 NC_009720.1
Rhodopseudomonas palustris NZ CP019967.1
Rhodopseudomonas palustris DX 1 NC_014834.1
Rhodopseudomonas palustris NZ CP058907.1
Rhodopseudomonas palustris NZ CP041387.1
Methylosinus trichosporium OB3b NZ CP023737.1
Bradyrhizobium oligotrophicum S58 NC_020453.1
Azorhizobium caulinodans ORS 571 NC_009937.1
Rhodobacter blasticus NZ CP020470.1
Rhizobium grahamii NZ CP043498.1
Magnetospirillum sp XM 1 NZ LN997848.1
Iodobacter sp H11R3 NZ CP034433.1
Paracoccus kondratievae NZ CP045072.1
Ruegeria sp AD91A NZ CP031946.1
Paraburkholderia sp PGU19 NZ AP023179.1
Achromobacter xylosoxidans A8 NC_014640.1
Paracoccus yeei NZ CP020442.2
Chelatococcus sp CO 6 NZ CP012398.1
Azospirillum brasilense NZ CP033318.1
Agrobacterium tumefaciens NZ CP039907.1
Mesorhizobium sp Pch S NZ CP029562.1
Bosea sp F3 2 NZ CP042331.1
Agrobacterium sp H13 3 NC_015508.1
Martelella sp NC20 NZ CP054861.1
Paracoccus yeei NZ CP044081.1
Sulfitobacter sp AM1 D1 NZ CP018076.1
Cupriavidus necator N 1 NC_015726.1
Caballeronia sp SBC1 NZ CP049156.1
2 3 1 4 9 5 6 7 8 10



Rhizobium pseudoryzae NZ CP049241.1
Litoreibacter sp LN3S51 NZ CP042261.1
Mesorhizobium sp 8 NZ CP040914.1
Paracoccus denitrificans NZ CP035090.1
Bosea vaviloviae NZ CP017147.1
Tardiphaga robiniae NZ CP050292.1
Paracoccus pantotrophus NZ CP044426.1
Cupriavidus pauculus NZ CP033969.1
Haematospirillum jordaniae NZ CP014525.1
Rhizobium sp NIBRBAC000502774 NZ CP041204.1
Pannonibacter phragmitetus NZ CP013068.1
Aromatoleum aromaticum EbN1 NC_006513.1
Agrobacterium tumefaciens NZ CP061003.1
Paracoccus aminovorans NZ LN832559.1
Asticcacaulis excentricus NZ AP018827.1
Neorhizobium sp NCHU2750 NZ CP030827.1
Chelatococcus daeguensis NZ CP018095.1
Agrobacterium sp RAC06 NZ CP016499.1
Martelella sp NC18 NZ CP054858.1
Bradyrhizobium symbiodeficiens NZ CP041090.1
Blastomonas fulva NZ CP020083.1
Variovorax sp PMC12 NZ CP027773.1
Ciceribacter thiooxidans NZ CP059896.1
Bradyrhizobium symbiodeficiens NZ CP050065.1
Massilia putida NZ CP019038.1
Bradyrhizobium symbiodeficiens NZ CP050066.1
Rhizobium sp S41 NZ CP016320.1
Paracoccus aminophilus JCM 7686 NC_022041.1
Ancylobacter pratisalsi NZ CP048630.1
Rhizobium sp WL3 NZ CP042826.1
Sphingosinicella sp BN140058 NZ CP035501.1
Bradyrhizobium symbiodeficiens NZ CP029427.1
Phenylobacterium zucineum HLK1 NC_011144.1
Caulobacter rhizosphaerae NZ CP048815.1
Agrobacterium sp 33MFTa1 1 NZ CP036358.1
Maritalea myrionectae NZ CP021330.1
Leisingera methylohalidivorans DSM 14336 NC_023135.1
Bosea sp ANAM02 NZ AP022848.1
Paraburkholderia caledonica NZ CP024905.1
Cupriavidus pinatubonensis JMP134 NC_007347.1
Paraburkholderia sp 7Q K02 NZ CP046909.1
Labrys sp KNU 23 NZ CP043489.1
Novosphingobium aromaticivorans DSM 12444 NC_007794.1
Delftia lacustris NZ CP065748.1
Labrenzia sp THAF82 NZ CP045354.1
Rhizobium sp Y9 NZ CP017999.1
Ketogulonicigenium vulgare NZ CP016592.1
Achromobacter xylosoxidans NZ CP045222.1
Paraburkholderia caribensis NZ CP026101.1
Cupriavidus sp USMAHM13 NZ CP017751.1
Caulobacter sp K31 NC_010338.1
Rhizobium pusense NZ CP053856.1
Burkholderia sp THE68 NZ AP022315.1
Nitratireductor sp SY7 NZ CP042301.2
Indioceanicola profundi NZ CP030126.1
Neorhizobium sp SOG26 NZ CP025512.1
Agrobacterium larrymoorei NZ CP039691.1
Burkholderia sp JP2 270 NZ CP029824.1
Paraburkholderia caribensis NZ CP013102.1
Nitratireductor sp OM 1 NZ CP029208.1
Martelella mediterranea DSM 17316 NZ CP020330.1
Azospirillum brasilense NZ CP032339.1
Agrobacterium vitis NZ AP023279.1
Cupriavidus oxalaticus NZ CP032518.1
Paraburkholderia terrae NZ CP026111.1
Agrobacterium vitis S4 NC_011989.1
Cupriavidus malaysiensis NZ CP017754.1
Nitrobacter hamburgensis X14 NC_007964.1
Ochrobactrum anthropi NZ CP044970.1
Marivivens sp JLT3646 NZ CP018572.1
Sphingobium yanoikuyae NZ CP020925.1
Aminobacter sp MDW 2 NZ CP060197.1
Azospirillum oryzae NZ CP054619.1
Sinorhizobium sp RAC02 NZ CP016450.1
Phyllobacterium zundukense NZ CP017940.1
Rhizobium sp 11515TR NZ CP022998.1
Agrobacterium radiobacter K84 NC_011985.1
Bradyrhizobium genosp B NZ CP061379.1
Bradyrhizobium genosp L NZ CP061378.1
Bradyrhizobium sp 1 2017 NZ CP050022.1
Bradyrhizobium sp TM102 NZ AP021855.1
Bradyrhizobium sp NZ LN901633.1
Phyllobacterium sp 628 NZ CP050301.1
Bradyrhizobium sp KBS0725 NZ CP042175.1
Bradyrhizobium sp KBS0727 NZ CP042176.1
Rhizobium sp ACO 34A NZ CP021371.1


6 10 7 1 2 3 4 9 5 8
Rhizobium leguminosarum NZ CP025012.1
Rhizobium hidalgonense NZ CP054027.1
Rhizobium leguminosarum bv viciae NZ CP022665.1
Rhizobium indicum NZ CP054021.1
Rhizobium indicum NZ CP054031.1
Rhizobium leguminosarum bv viciae 3841 NC_008380.1
Rhizobium phaseoli NZ CP013522.1
Rhizobium sp 007 NZ CP064187.1
Rhizobium phaseoli NZ CP064931.1
Rhizobium leguminosarum bv trifolii NZ CP050103.1
Rhizobium leguminosarum bv trifolii NZ CP050108.1
Rhizobium leguminosarum bv trifolii NZ CP050091.1
Rhizobium leguminosarum bv trifolii NZ CP050085.1
Rhizobium leguminosarum bv trifolii NZ CP050097.1
Rhizobium leguminosarum bv trifolii NZ CP050080.1
Rhizobium leguminosarum bv viciae 248 NZ CP048280.1
Rhizobium jaguaris NZ CP032694.1
Rhizobium leguminosarum NZ CP030760.1
Rhizobium leguminosarum bv viciae NZ CP025509.1
Rhizobium sp NXC24 NZ CP024311.1
Rhizobium acidisoli NZ CP034998.1
Rhizobium etli NZ CP020906.1
Rhizobium leguminosarum bv trifolii WSM1689 NZ CP007045.1
Rhizobium etli bv mimosae str Mim1 NC_021905.1
Rhizobium leguminosarum bv trifolii TA1 NZ CP053205.2
Rhizobium phaseoli Brasil 5 NZ CP020896.1
Rhizobium sp Kim5 NZ CP021124.1
Rhizobium etli CFN 42 NC_007761.1
Rhizobium leguminosarum bv trifolii WSM1325 NC_012850.1
Rhizobium leguminosarum bv trifolii WSM2304 NC_011369.1
Rhizobium sp TAL182 NZ CP021024.1
Rhizobium sp NXC14 NZ CP021030.1
Rhizobium gallicum NZ CP017101.1
Rhizobium etli 8C 3 NZ CP017241.1
Rhizobium leguminosarum NZ CP018228.1
Rhizobium leguminosarum NZ CP016286.1
Rhizobium sp N741 NZ CP013595.1
Rhizobium sp N871 NZ CP013590.1
Rhizobium sp N324 NZ CP013630.1
Rhizobium sp N6212 NZ CP013490.1
Rhizobium phaseoli NZ CP013547.1
Rhizobium phaseoli NZ CP013557.1
Rhizobium phaseoli NZ CP013568.1
Rhizobium phaseoli NZ CP013532.1
Rhizobium phaseoli NZ CP013574.1
Rhizobium sp N113 NZ CP013517.1
Rhizobium sp N621 NZ CP013495.1
Rhizobium sp N1314 NZ CP013511.1
Rhizobium phaseoli NZ CP013552.1
Rhizobium esperanzae NZ CP013500.1
Rhizobium phaseoli NZ CP013542.1
Rhizobium phaseoli NZ CP013527.1
Rhizobium phaseoli NZ CP013537.1
Rhizobium phaseoli NZ CP013585.1
Rhizobium sp N731 NZ CP013601.1
Rhizobium phaseoli NZ CP013563.1
Rhizobium phaseoli NZ CP013580.1
Rhizobium sp N1341 NZ CP013505.1
Rhizobium etli bv phaseoli str IE4803 NZ CP007641.1
Rhizobium leguminosarum bv trifolii NZ CP053439.1
Rhizobium sp IE4771 NZ CP006986.1
Rhizobium leguminosarum bv trifolii CB782 NZ CP007067.1
Rhizobium tropici CIAT 899 NC_020059.1
Rhizobium sp CIAT894 NZ CP020947.1
Rhizobium etli CIAT 652 NC_010994.1
Rhizobium sp CCGE531 NZ CP032684.1
Rhizobium sp CCGE532 NZ CP032689.1
6 10 7 1 2 3 4 9 5 8



Os números representam os 10 clusters com maior correlação com os organismos noduladores: 1, NodA; 2, NodB; 3, NodC; 4, NodD; 5, NifA; 6, SRPBCC domain-containing protein; 7, Adenylate/guanylate cyclase domain-containing protein; 8, hypothetical protein; 9, NifN; 10, Hydroxyacid dehydrogenase. A, cluster 1; B, cluster 2; C, cluster 3; D, cluster 4.

5 CONSIDERAÇÕES FINAIS

A prospecção de novos organismos diazotróficos noduladores é necessária tanto para a exploração como alternativa aos fertilizantes químicos, quanto para compreensão dos mecanismos envolvidos nesta relação simbiótica. A construção de uma estrutura que altera morfologicamente organismos tão distintos como uma planta e uma bactéria, e estabelece uma relação com uma especificidade complexa para viabilizar a fixação de nitrogênio é intrigante tanto do ponto de vista de processo quanto evolutivo. As várias etapas envolvidas em identificar experimentalmente estes organismos dificultam as possibilidades de exploração para se conhecer os mecanismos fundamentais e consequentemente as oportunidades de maximizar o processo de promoção do crescimento vegetal.

O método atualmente utilizado para prospecção de diazotrofos, que é utilizando um conjunto mínimo de genes essenciais, se demonstrou ineficaz para encontrar novos organismos potenciais em nossos resultados. Este comportamento é esperado pois ao se utilizar apenas os mecanismos biológicos conhecidos, dificilmente encontraremos algo distinto e desconhecido nos resultados. E o processo biológico possui muitas variáveis e especificidades ainda desconhecidas. Aplicando um método de análise de clusters com a ferramenta RAFTS3G (em alternativa ao método padrão de busca via BLASTp), porém os resultados ainda trouxeram basicamente organismos já conhecidos.

Foi desenvolvido então um método de reconhecimento de padrões (NodProspect) que utiliza o proteoma dos organismos, sem induzir diretamente à necessidade da presença dos conjuntos mínimos já estabelecidos. Este modelo foi capaz de reconhecer organismos já conhecidos que utilizam processos alternativos, e ainda prospectar novos potenciais organismos para exploração. O fato de que organismos que utilizam métodos alternativos terem sido prospectados pelo NodProspect sugere que ainda há muitos elementos envolvidos que podem ser considerados como essenciais, mas ainda não foram descobertos por experimentos em laboratório. A exploração a partir dos rankings de correlação de clusters de genes mais relevantes pode auxiliar nessa identificação em projetos futuros, onde a experimentação de genes altamente correlacionados, mas ainda desconhecidos, sugerem um potencial produto necessário para o sucesso da simbiose.

REFERÊNCIAS

BASSI, D.; MENOSSI, M.; MATTIELLO, L. Nitrogen supply influences photosynthesis establishment along the sugarcane leaf. **Nature Scientifics Reports**, 8, 2327, 2018. DOI: 10.1038/s41598-018-20653-1.

BERNHARD, A. The Nitrogen Cycle: Processes, Players, and Human Impact. **Nature Education Knowledge** 3(10):25, 2010.

BROMFIELD, E. S. P.; CLOUTIER, S.; NGUYEN, H. D. T. Description and complete genome sequences of Bradyrhizobium symbiodeficiens sp. nov., a non-symbiotic bacterium associated with legumes native to Canada. **International journal of systematic and evolutionary microbiology**, 70(1), 442–449, 2020. DOI: 10.1099/ijsem.0.003772.

CARVALHO, T. L. G. BALSEMÃO-PIRES, E. SARAIVA, R. M. FERREIRA P. C. G. HEMERLY A. S. Nitrogen signalling in plant interactions with associative and endophytic diazotrophic bacteria. **Journal of Experimental Botany**, Vol. 65, pp. 5631–5642, 2014. DOI: 10.1093/jxb/eru319.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic minority over-sampling technique. **Journal of artificial intelligence research**, 16, 321-357, 2002.

CONSORTIUM, T. U. Uniprot: the universal protein knowledgebase. **Nucleic Acids Research**,

COSKUN, D.;BRITTO, D. T.; SHI, w.; KRONZUCKER, H. J. Nitrogen transformations in modern agriculture and the role of biological nitrification inhibition. **Nature Plants**, v. 3, 17074, Jun. 2017. DOI: 10.1038/nplants.2017.74.

DAS, P.; ROYCHOWDHURY, A.; DAS, S.; ROYCHOWDHURY, S.; TRIPATHY, S. sigFeature: Novel Significant Feature Selection Method for Classification of Gene Expression Data Using Support Vector Machine and t Statistic. **Frontiers in Genetics**, v. 11, 2020. DOI: 10.3389/fgene.2020.00247.

DEAKIN, W. J.; BROUGHTON, W. J. Symbiotic use of pathogenic strategies: rhizobial protein secretion systems. **Nature reviews microbiology**, 7, 312-320, 2009. DOI: 10.1038/nrmicro2091.

DEAN, R. D.; BOLIN, J. T.; ZHENG, L. Nitrogenase metalloclusters: structure, organization and synthesis. **Journal of bacteriology**, v. 175 (21), p. 6737-6744, 1993. DOI: 10.1128/jb.175.21.6737-6744.1993.

DIXON, R.; KAHN, D. Genetic regulation of biological nitrogen fixation. **Nature reviews microbiology**, v. 2, 2004. DOI: 10.1038/nrmicro954.

DOS SANTOS, P. C.; FANG, Z.; MASON, S. W.; SETUBAL, J. C.; DIXON, R. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. **BMC Genomics**, 13:162, Maio 2012. DOI: 10.1186/1471-2164-13-162.

DOWNIE, J. A. The roles of extracellular proteins, polysaccharides and signals in the interactions of rhizobia with legume roots. **FEMS Microbiology Reviews**, 34, 150–170, 2010. DOI: 10.1111/j.1574-6976.2009.00205.x.
estimation for finding clusters of homologous proteins—tracing actinobacterial pathogenicity

FIBACH-PALDI, S.; BURDMAN, S.; OKON, Y. Key physiological properties contributing to rhizosphere adaptation and plant growth promotion abilities of Azospirillum brasilense. **FEMS Microbiology letters**, 326, 99-108, 2012. DOI: 10.1111/j.1574-6968.2011.02407.x.

GIRAUD, E.; MOULIN, L.; VALLENET, D.; BARBE, V.; CYTRYN, E.; AVARRE, J. C.; JAUBERT, M.; SIMON, D.; CARTIEAUX, F.; PRIN, Y.; BENA, G.; HANNIBAL, L.; FARDOUX, J.; KOJADINOVIC, M.; VUILLET, L.; LAJUS, A.; CRUVEILLER, S.; ROUY, Z.; MANGENOT, S.; SEGURENS, B.; DOSSAT, C.; FRANCK, W. L.; CHANG, W. S.; SAUNDERS, E.; BRUCE, D.; RICHARDSON, P.; NORMAND, P.; DREYFUS, B.; PIGNOL, D; STACEY, G.; EMERICH, D.; VERMÉGLIO, A.; MÉDIGUE, C.; SADOWSKY, M. Legumes symbioses: absence of Nod genes in photosynthetic bradyrhizobia. **Science**, 1;316(5829):1307-12, 2007 doi: 10.1126/science.1139548. PMID: 17540897.
GREENER, J. G.; KANDATHIL, S. M.; MOFFAT, L.; JONES, D. T. A guide to machine learning for biologists. **Nature reviews molecular cell biology**, 23, 40-55, 2022. DOI: 10.1038/s41580-021-00407-0.

GRUBER, N.; GALLOWAY, J.N. An Earth-system perspective of the global nitrogen cycle. **Nature**. Vol 451, p 293 – 296, 2008. DOI: 10.1038/nature06592.
GUO, Y.; MATSUOKA, Y.; MIURA, T.; NISHIZAWA, T.; OHTA, H.; & NARISAWA, K. Complete genome sequence of Agrobacterium pusense VsBac-Y9, a bacterial symbiont of the dark septate endophytic fungus Veronaeopsis simplex Y34 with potential for improving fungal colonization in roots. **Journal of biotechnology**, 284, 31–36, 2018. DOI: 10.1016/j.jbiotec.2018.07.045.

GUYON, I.; WESTON, J.; BARNHILL, S. Gene selection for cancer classification using support vector machines. **Machine learning**, 46, 389-422, 2002. DOI: 10.1023/A:1012487302797.

GYANESHWAR, P.; HIRSCH, A. M.; MOULIN, L.; CHEN, W. M.; ELLIOTT, G. N.; BONTEMPS, C.; ESTRADA-DE LOS SANTOS, P.; GROSS, E.; DOS REIS, F. B.; SPRENT, J. I.; YOUNG, J. P.; JAMES, E. K. Legume-nodulating betaproteobacteria: diversity, host range, and future prospects. **Molecular plant-microbe interactions : MPMI**, 24(11), 1276–1288, 2011. DOI: 10.1094/MPMI-06-11-0172

LI, X.; GENG, X.; XIE, R.; FU, L.; JIANG, J.; GAO, L.; SUN, J. The endophytic bacteria isolated from elephant grass (Pennisetum purpureum Schumach) promote plant growth and enhance salt tolerance of Hybrid Pennisetum. **Biotechnology for biofuels**, 9:190, 2016. DOI: 10.1186/s13068-016-0592-0.

LIANG, J.; HOFFRICHTER, A.; BRACHMANN, A.; MARIN, M. Complete genome of Rhizobium leguminosarum Norway, an ineffective Lotus micro-symbiont. **Environmental Microbiome**, 13, 36, 2018. DOI: 10.1186/s40793-018-0336-9.
lifestyles. **Bioinformatics**, v. 29, p. No 2, 215–222, 2013. DOI: 10.1093/bioinformatics/bts653.

LIU, F. T.; TING, K. M.; ZHOU, Z. Isolation Forest. **Eighth IEEE International Conference on Data Mining**, pp. 413-422, 2008. DOI: 10.1109/ICDM.2008.17.

MARTÍNEZ-HIDALGO, P.; HIRSCH, A. M. The nodule microbiome: N2-fixing rhizobia do not live alone. **Phytobiomes**, v 1, p. 70-82, 2017. DOI: 10.1094/PBIOMES-12-16-0019-RVW.

MERGAERT, P.; KERESZT, A.; KONDOROSI, E. Gene Expression in Nitrogen-Fixing Symbiotic Nodule Cells in Medicago truncatula and Other Nodulating Plants. **The plant cell**, v. 32, 1, 42-68, 2020. DOI: 10.1105/tpc.19.00494.

MOCKUS, J.;TIESIS, V.; ZILINSKAS, A. The application of Bayesian methods for seeking the extremum. **Towards Global Optimization**, 2:117–129, 1978.

NICHIO, B. T. L.; OLIVEIRA, A. M. R.; PIERRI, C. R.; SANTOS, L. G. C.; LEJAMBRE, A. Q.; VIALLE, R. A.; COIMBRA, N. A. R.; GUIZELINI, D.; MARCHAUKOSKI, J. N.; PEDROSA, F. O.; RAITTZ, R. T. RAFTS3G: an efficient and versatile clustering software to analyses in large protein datasets. **BMC Bioinformatics** 20, 392, 2019. DOI: 10.1186/s12859-019-2973-4.

OKUBO, T.; FUKUSHIMA, S.; ITAKURA, M.; OSHIMA, K.; LONGTONGLANG, A.; TEAUMROONG, N.; MITSUI, H.; HATTORI, M.; HATTORI, R.; HATTORI, T.; MINAMISAWA, K. Genome analysis suggests that the soil oligotrophic bacterium Agromonas oligotrophica (Bradyrhizobiumoligotrophicum) is a nitrogen-fixing symbiont of Aeschynomene indica. **Applied and environmental microbiology**, 79(8), 2542–2551, 2013. DOI: 10.1128/AEM.00009-13.

PERRET, X.; STAEHELIN, C.; BROUGHTON, W. J. Molecular basis of symbiotic promiscuity. **microbiology and molecular biology reviews**, v. 64, 180-201, 2000. DOI: 10.1128/MMBR.64.1.180-201.2000.

PIERRI, C. R.; VOYCEIK, R.; MATTOS, L. G. C. S.; KULIK, M. G.; CAMARGO, J. O.; OLIVEIRA, A. M. R.; NICHIO, B. T. L.; MARCHAUKOSKI, J. N.; FILHO, A. C. S.; GUIZELINI, D.; ORTEGA, J. M.; PEDROSA, F. O.; RAITTZ, R. T. SWeeP: representing large biological sequences datasets in compact vectors. **Nature scientific reports**, 10, 91, 2020. DOI: 10.1038/s41598-019-55627-4.

POOLE, P.; RAMACHANDRAN, V.; TERPOLILLI, J. Rhizobia: from saprophytes to endosymbionts. **Nature reviews microbiology**, v. 16, 291-303, Maio 2018. DOI: 10.1038/nrmicro.2017.171.
REVIEWS IN BIOMEDICAL ENGINEERING, v. 3, p. 1937–3333, 2010. DOI: 10.1109/RBME.2010.2083647.

ROTTGER R.; KALAGHATGI, P. S. P. S. S. C. A. V. W. T.; BAUMBACH, J. Density parameter
RUBIO, L.M; LUDDEN P.W. Biosynthesis of the Iron-Molybdenum Cofactor of Nitrogenase. **Annu. Rev. Microbiol,**V. 62, p. 93–111, 2008. DOI: 10.1146/annurev.micro.62.081307.162737

SNOEK, J.; LAROCHELLE, H.; ADAMS, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. **Statistics machine learning**, 1206, 2944, 2012. DOI:10.48550/arXiv.1206.2944

TIMMUSK, S.; BEHERS, L.; MUTHONI, J.; MURAYA, A.; ARONSSON, A. C. Perspectives and Challenges of Microbial Application for Crop Improvement. **Frontiers in plant science**, Volume 8, Article 49, 2017. DOI: 10.3389/fpls.2017.00049.

UNAY, J.; & PERRET, X. A Minimal Genetic Passkey to Unlock Many Legume Doors to Root Nodulation by Rhizobia. **Genes**, 11(5), 521, 2020. DOI: 10.3390/genes11050521.
VOL. 45, p. D158–D169, 2016.

WAGNER, S. C. Biological Nitrogen Fixation. **Nature Education Knowledge**, 3(10):15, 2011.

WANG, L.; SUN, Z.; SU, C.; WANG, Y.; YAN, Q.; CHEN, J.; OTT, T.; LI, X. A GmNINa-miR172c-NNC1 Regulatory Network Coordinates the Nodulation and Autoregulation of Nodulation Pathways in Soybean. **Molecular plant**, v. 12, 9, 1211-1226, 2019. DOI: 10.1016/j.molp.2019.06.002.

WANG, Q.; LIU, J.; ZHU, H. Genetic and Molecular Mechanisms Underlying Symbiotic Specificity in Legume-Rhizobium Interactions. **Frontiers in plant science**, 9:313, 2018. DOI: 10.3389/fpls.2018.00313.

WASAI-HARA, S.; MINAMISAWA, K.; CLOUTIER, S.; & BROMFIELD, E. S. P. Strains of Bradyrhizobium cosmicum sp. nov., isolated from contrasting habitats in Japan and Canada possess photosynthesis gene clusters with the hallmark of genomic islands. **International journal of systematic and evolutionary microbiology**, 70(9), 5063–5074, 2020. DOI: 10.1099/ijsem.0.004380

WATSON, L. J.; RAHMANI, T. P. D. A glimpse of the Nitrogen-Fixing Wheat possibility. **Journal of natural sciences and mathematics research**, v. 8, 1-9, 2022. DOI: 10.21580/jnsmr.2022.8.1.11766

XU, R.; WUNSCH, D. C. Clustering algorithms in biomedical research: A review. **IEEE**

YAN J.; YAN, H.; LIU, L. X.; CHEN, W. F.; ZHANG, X. X. ; VERÁSTEGUI-VALDÉS, M. M.; WANG, E.T.; HAN, X. Z. Rhizobium hidalgonense sp. nov., a nodule endophytic bacterium of Phaseolus vulgaris in acid soil. **Archives of microbiology**, Jan;199(1):97-104, 2017. DOI: 10.1007/s00203-016-1281-x.

APÊNDICE 1 – ANÁLISE DE ORGANISMOS PROSPECTADOS

Realizamos a prospecção utilizando o NodProspect e o NodProspectEnsemble em todas as bactérias disponíveis no NCBI. A diferença de construção entre os dois está em que o NodProspect é um modelo SVM tradicional treinado e otimizado massivamente em um conjunto de negativos extraídos utilizando amostra sistemática, e o NodProspectEnsemble são 100 SVM treinadas e otimizadas em conjuntos negativos extraídos de forma aleatória.

O NodProspect prospectou 320 organismos e foi capaz de identificar três bactérias que possuem capacidade de nodulação observada, porém não possuem o conjunto mínimo de genes conhecidos: os organismos *Bradyrhizobium sp ORS278* e *Bradyrhizobium sp BTAil*, que apresentam apenas o *nodB*, e o *Bradyrhizobium oligotrophium S58*, que não apresenta o conjunto *nodABC*. Os três organismos formam nódulos efetivos em espécies de Aechynomene (Girald et al., 2017; Okubo et al., 2013). Uma outra característica observada na prospecção é que os organismos *R. hidalgonense JKLM 19E* e *R. Leguminosarum Norway*, que não criam nódulos efetivos mas que foram prospectados pelo método de análise de clusters, não foram prospectados pelo NodProspect, mesmo quando o último apresenta os genes *nodABC* e *nifHDK* (Bromfield et al., 2020; YAN et al., 2017; LIANG et al., 2018), demonstrando que o modelo criado não utiliza a presença ou ausência desses seis genes como parâmetro discriminante, e sugerindo que o processo envolve muitos outros genes não tão explorados para realizar a simbiose.

O NodProspectEnsemble prospectou 303 organismos, sendo que o *Bradyrhizobium oligotrophium S58*, nodulador que não apresenta o conjunto *nodABC*, foi prospectado também assim como no NodProspect, porém os outros dois exóticos não foram prospectados. O ensemble também não foi capaz de reconhecer como falsos positivos os organismos *R. hidalgonense JKLM 19E* e *R. Leguminosarum Norway*, prospectando-os de forma incorreta. Porém esta prospecção desses falsos positivos não pode ser interpretada como um resultado de todo ruim, pois eles apresentam genes de noduladores e o primeiro foi isolado de nódulo, o que indica que ele possivelmente apresenta mecanismos que não o impediram de estar presente no nódulo, mesmo que não realizando fixação de nitrogênio.

Comparativamente, o NodProspect apresentou dois organismos confirmados que não foram prospectados pelo NodProspectEnsemble: *Methylobacterium sp. 4-46* e Bradyrhizobium *sp BTAil*. E o NodProspectEnsemble obteve um organismo extraído de nódulo, mas sem testes de capacidade de fixação que não foi prospectado pelo NodProspect: *Rhizobium hidalgonense*. O NodProspect encontrou 43 organismos extraídos

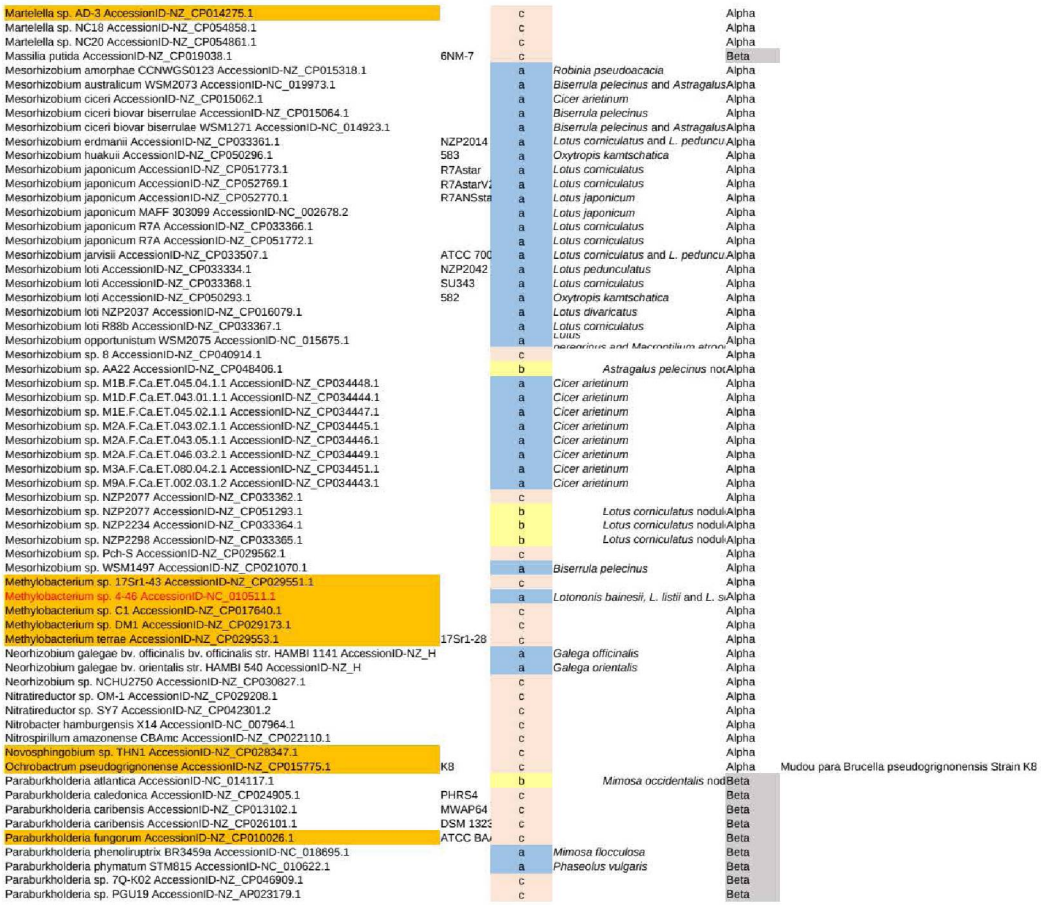
de nódulos que não foram prospectados pelo ensemble, enquanto o ensemble encontrou 24 organismos extraídos de nódulos que não foram prospectados pelo modelo único; estes organismos extraídos de nódulos ainda não possuem atividade de fixação validadas. O ensemble também apresentou 3 não-noduladores confirmados que o modelo tradicional não prospectou.

Acreditamos que as duas abordagens, um modelo SVM tradicional e um modelo em formato de ensemble, podem ser utilizados de forma complementar para realizar prospecção de organismos diazotrofos noduladores, pois como esta classe de organismos oferece um conjunto de treinamento relativamente pequeno, realizar explorações unindo os resultados obtidos por diferentes métodos eventualmente pode identificar diazotrofos de interesse, desde que sejam feitas análises mais detalhadas dos genomas e características conhecidas.

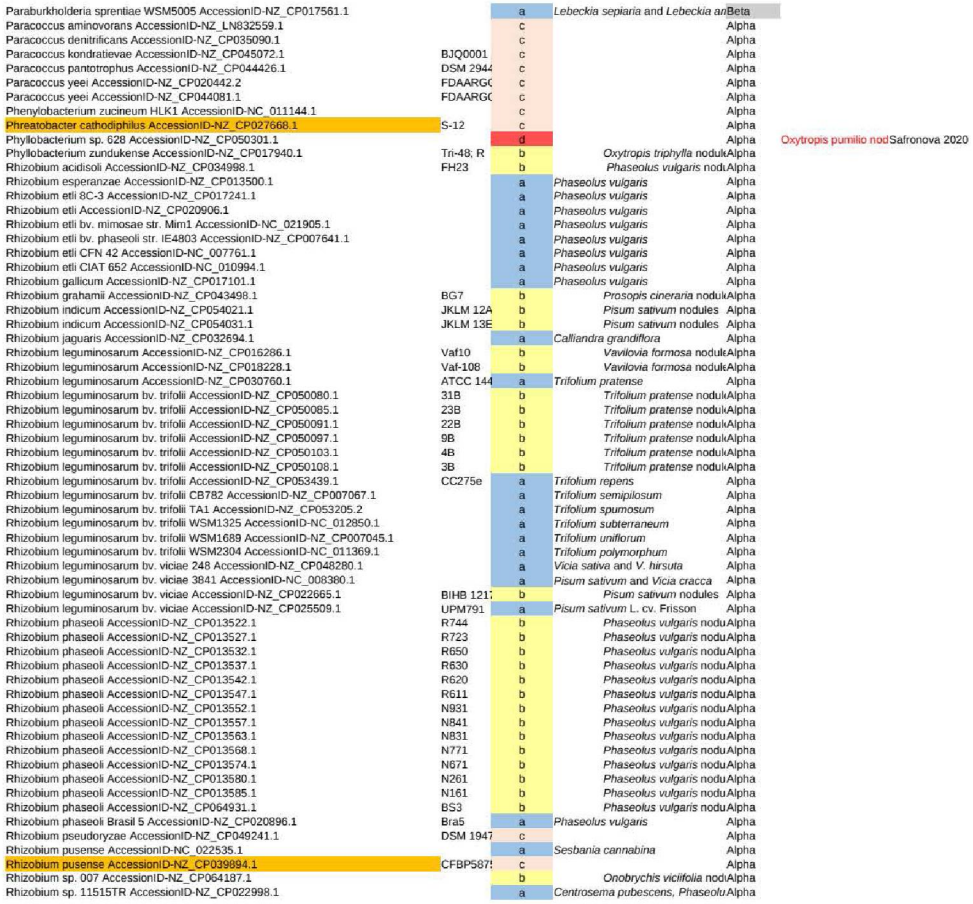
Tabela A1 – Prospecções realizadas no conjunto *Alpha Beta Genomes*

Bacteria - Genome accession ID	Strain	Confirmed nodulation ability	st example source of strain isolatio	alfa ou beta		
Achromobacter xylosoxidans A8 AccessionID-NC_014640.1		a	*Cicer arietinum and Glycine max*	Beta		
Achromobacter xylosoxidans DN002 AccessionID-NZ_CP045222.1		c		Beta		
Agrobacterium larrymoorei CFBP5473 AccessionID-NZ_CP039691.1		c		Alpha		
Agrobacterium rhizogenes K599 AccessionID-NZ_CP019701.2		c		Alpha		
Agrobacterium sp. 33MFTa1.1 AccessionID-NZ_CP036358.1		c		Alpha		
Agrobacterium sp. H13-3 AccessionID-NC_015508.1		c		Alpha		
Agrobacterium sp. RAC06 AccessionID-NZ_CP016499.1		c		Alpha		
Agrobacterium tumefaciens (Agrobacterium radiobacter = Rhizobium radiobacter) Acc	Ach5	c		Alpha		
Agrobacterium tumefaciens (Agrobacterium radiobacter = Rhizobium radiobacter) Acc	15955	c		Alpha		
Agrobacterium tumefaciens (Agrobacterium radiobacter = Rhizobium radiobacter) Acc	A6	c		Alpha		
Agrobacterium tumefaciens (Agrobacterium radiobacter = Rhizobium radiobacter) Acc	CFBP662	c		Alpha		
Agrobacterium tumefaciens (Agrobacterium radiobacter = Rhizobium radiobacter) Acc	BIM B-13	c		Alpha		
Agrobacterium vitis AccessionID-NZ_AP023268.1	VAR03-1	c		Alpha		
Agrobacterium vitis AccessionID-NZ_AP023279.1	VAT03-9	c		Alpha		
Agrobacterium vitis S4 AccessionID-NC_011989.1		c		Alpha		
Alcaligenes faecalis AccessionID-NZ_CP032521.1	J481	c		Beta		
Aminobacter sp. MDW-2 AccessionID-NZ_CP060197.1	MDW-2	c		Alpha		
Aromatoleum aromaticum EbN1 AccessionID-NC_006513.1		c		Beta		
Asticcacaulis excentricus M6 AccessionID-NZ_AP018827.1		c		Alpha		
Azorhizobium caulinodans ORS 571 AccessionID-NC_009937.1		a	*Sesbania rostrata*	Alpha		
Azospira sp. I09 AccessionID-NZ_AP021844.1		c		Beta		
Azospirillum brasilense AccessionID-NZ_CP007793.1	Az39	c		Alpha		
Azospirillum brasilense AccessionID-NZ_CP032339.1	MTCC403	c		Alpha		
Azospirillum lipoferum 4B AccessionID-NC_016622.1		c		Alpha		
Azospirillum oryzae AccessionID-NZ_CP054619.1	KACC 144	c		Alpha		
Azospirillum ramasamyi AccessionID-NZ_CP029829.1	M2T2B2	c		Alpha		
Bosea sp. F3-2 AccessionID-NZ_CP042331.1		c		Alpha		
Bosea vaviloviae AccessionID-NZ_CP017147.1	Vaf18	b	*Vavilovia formosa* nodule	Alpha		
Bradyrhizobium amphicarpaeae AccessionID-NZ_CP029426.1	39S1MB	d		Alpha	soybean nodules	Bromfield 2019
Bradyrhizobium arachidis CCBAU 051107 AccessionID-NZ_CP030050.1		a	*Arachis hypogaea* and *Lablab purp*	Alpha		
Bradyrhizobium diazoefficiens AccessionID-NZ_AP014685.1		a	*Glycine max*	Alpha		
Bradyrhizobium diazoefficiens AccessionID-NZ_AP022638.1		b	soybean nodules	Alpha		
Bradyrhizobium diazoefficiens AccessionID-NZ_AP022639.1		b	soybean nodules	Alpha		
Bradyrhizobium diazoefficiens AccessionID-NZ_AP022640.1		b	soybean nodules	Alpha		
Bradyrhizobium diazoefficiens AccessionID-NZ_AP022641.1		b	soybean nodules	Alpha		
Bradyrhizobium diazoefficiens AccessionID-NZ_CP013127.1		a	*Glycine max*	Alpha		
Bradyrhizobium diazoefficiens AccessionID-NZ_CP029603.2		b	soybean nodules	Alpha		
Bradyrhizobium diazoefficiens AccessionID-NZ_CP032617.1		a	*soybeans*	Alpha		
Bradyrhizobium diazoefficiens AccessionID-NZ_CP050064.1		b	soybean nodules	Alpha		
Bradyrhizobium diazoefficiens AccessionID-NZ_CP055233.1		a	*soybean*	Alpha		
Bradyrhizobium diazoefficiens USDA 110 AccessionID-NC_004463.1		a	*Glycine max*	Alpha		
Bradyrhizobium diazoefficiens USDA 110 AccessionID-NZ_CP011360.1		a	*Glycine max*	Alpha		
Bradyrhizobium elkanii USDA 61 AccessionID-NZ_AP013103.1		a	*Glycine max*	Alpha		
Bradyrhizobium genosp. B AccessionID-NZ_CP061379.1	BDV5040	a	*Bossiaea ensata*	Alpha		
Bradyrhizobium genosp. L AccessionID-NZ_CP061378.1	BDV5419	a	*Hardenbergia violacea*	Alpha		
Bradyrhizobium guangdongense AccessionID-NZ_CP030051.1	CCBAU 5	a	*Arachis hypogaea*	Alpha		
Bradyrhizobium guangdongense AccessionID-NZ_CP030057.1	CCBAU 5	a	*Arachis hypogaea*	Alpha		
Bradyrhizobium guangxiense AccessionID-NZ_CP022219.1	CCBAU 5	a	*Lablab purpureus* and *Aeschynom*	Alpha		
Bradyrhizobium guangzhouense AccessionID-NZ_CP030053.1	CCBAU 5	a	*Arachis hypogaeaand* and *Lablab*	Alpha		
Bradyrhizobium icense AccessionID-NZ_CP016428.1		a	*Phaseolus lunatus L*	Alpha		
Bradyrhizobium japonicum AccessionID-NZ_CP010313.1	E109	a	*Glycine max*	Alpha		
Bradyrhizobium japonicum AccessionID-NZ_CP017637.1	J5	a	*Glycine max*	Alpha		
Bradyrhizobium japonicum AccessionID-NZ_CP058354.1	5038	b	soybean nodules	Alpha		
Bradyrhizobium japonicum USDA 6 AccessionID-NC_017249.1		a	*Glycine max*	Alpha		
Bradyrhizobium oligotrophicum S58 AccessionID-NC_020453.1		a	*Aeschynomene indica*	Alpha		
Bradyrhizobium ottawaense AccessionID-NZ_CP029425.1		a	*Glycine max*	Alpha		
Bradyrhizobium paxllaeri AccessionID-NZ_CP042968.1	LMTR 21	a	*Phaseolus lunatus* and *Vigna ungu*	Alpha		
Bradyrhizobium sp. 1(2017) 63S1MB AccessionID-NZ_CP050022.1		d		Alpha	soybean nodules	Bromfield 2017
Bradyrhizobium sp. 6(2017) AccessionID-NZ_CP049289.1	1S3	b	soybean nodules	Alpha		
Bradyrhizobium sp. BF49 AccessionID-NZ_LN901633.1		d		Alpha	free living	Jones 2016

Bradyrhizobium sp. BTAi1		a	*Aeschynomene sensitiva*	Alpha			
Bradyrhizobium sp. CCBAU 051011 AccessionID-NZ_CP022222.1		b	*Arachis hypogaea* nodul	Alpha			
Bradyrhizobium sp. CCBAU 21365 AccessionID-NZ_CP030036.1		b	*Arachis hypogaea* nodul	Alpha			
Bradyrhizobium sp. CCBAU 51753 AccessionID-NZ_CP030037.1		b	*Arachis hypogaea* nodul	Alpha			
Bradyrhizobium sp. CCBAU 51765 AccessionID-NZ_CP030038.1		b	*Arachis hypogaea* nodul	Alpha			
Bradyrhizobium sp. CCBAU 53338 AccessionID-NZ_CP030048.1		b	*Arachis hypogaea* nodul	Alpha			
Bradyrhizobium sp. CCBAU 53340 AccessionID-NZ_CP030055.1		b	*Arachis hypogaea* nodul	Alpha			
Bradyrhizobium sp. CCBAU 53351 AccessionID-NZ_CP030059.1		b	*Arachis hypogaea* nodul	Alpha			
Bradyrhizobium sp. CCBAU 53421 AccessionID-NZ_CP030047.1		b	*Arachis hypogaea* nodul	Alpha			
Bradyrhizobium sp. CCGE-LA001 AccessionID-NZ_CP013949.1		b	*Phaseolus microcarpus* t	Alpha			
Bradyrhizobium sp. KBS0725 AccessionID-NZ_CP042175.1		c		Alpha			
Bradyrhizobium sp. KBS0727 AccessionID-NZ_CP042176.1		c		Alpha			
Bradyrhizobium sp. LCT2 AccessionID-NZ_CP034432.1		b	*Lespedeza cuneata* nodu	Alpha			
Bradyrhizobium sp. ORS 278 AccessionID-NC_009445.1		a	*Aeschynomene indica*	Alpha			
Bradyrhizobium sp. ORS 285 AccessionID-NZ_LT859959.1		a	*Aeschynomene afraspera* and *A. i*	Alpha			
Bradyrhizobium sp. SG09 AccessionID-NZ_AP021854.1		c		Alpha			
Bradyrhizobium sp. TM102 AccessionID-NZ_AP021855.1		d		Alpha	isolado de raiz de sorgo		
Bradyrhizobium symbiodeficiens AccessionID-NZ_CP029427.1	85S1MB	d		Alpha	soybean nodules	Bromfield 2020	
Bradyrhizobium symbiodeficiens AccessionID-NZ_CP041090.1	65S1MB	d		Alpha	soybean nodules	Bromfield 2020	
Bradyrhizobium symbiodeficiens AccessionID-NZ_CP050065.1	141S2	d		Alpha	soybean nodules	Bromfield 2020	
Bradyrhizobium symbiodeficiens AccessionID-NZ_CP050066.1	101S1MB	d		Alpha	soybean nodules	Bromfield 2020	
Bradyrhizobium vignae AccessionID-NZ_LS398110.1		a	*Vigna unguiculata*, *Arachis hypoga*	Alpha			
Bradyrhizobium zhanjiangense AccessionID-NZ_CP022221.1	CCBAU 5	a	*Arachis hypogaea*	Alpha			
Burkholderia cenocepacia AccessionID-NZ_CP015036.1	895	c		Beta			
Burkholderia gladioli AccessionID-NZ_CP033430.1	Co14	c		Beta			
Burkholderia humptydooensis AccessionID-NZ_CP065686.1	FDAARG	c		Beta			
Burkholderia multivorans AccessionID-NZ_CP020397.1	FDAARG	c		Beta			
Burkholderia multivorans ATCC BAA-247 AccessionID-NZ_CP009832.1		c		Beta			
Burkholderia sp. 2002721687 AccessionID-NZ_CP009549.1		c		Beta			
Burkholderia sp. Bp5365 AccessionID-NZ_CP013380.1	MSMB43	c		Beta			
Burkholderia sp. JP2-270 AccessionID-NZ_CP029824.1		c		Beta			
Burkholderia sp. RPE67 AccessionID-NZ_AP014576.1		c		Beta			
Burkholderia thailandensis 34 AccessionID-NZ_CP010017.1		c		Beta			
Burkholderia ubonensis AccessionID-NZ_CP013414.1	MSMB203	c		Beta			
Caballeronia sp. SBC1 AccessionID-NZ_CP049156.1		c		Beta			
Caulobacter sp. K31 AccessionID-NC_010338.1		c		Alpha	Registro removido do NCBI		
Chelatococcus daeguensis TAD1 AccessionID-NZ_CP018095.1		c		Alpha			
Chelatococcus sp. CO-6 AccessionID-NZ_CP012398.1		c		Alpha			
Ciceribacter thiooxidans AccessionID-NZ_CP059896.1	F43B	c		Alpha			
Cupriavidus malaysiensis AccessionID-NZ_CP017754.1	USMAA1	c		Beta			
Cupriavidus necator (Ralstonia eutropha) AccessionID-NZ_CP017757.2	NH9	c		Beta			
Cupriavidus necator N-1 AccessionID-NC_015726.1		c		Beta			
Cupriavidus oxalaticus AccessionID-NZ_CP032518.1	T2	c		Beta			
Delftia lacustris AccessionID-NZ_CP065748.1	FDAARG	c		Beta			
Dinoroseobacter shibae DFL 12 = DSM 16493 AccessionID-NC_009952.1		c		Alpha			
Ensifer alkalisoli AccessionID-NZ_CP034909.1	YIC4027	a	*Sesbania cannabina*	Alpha			
Ensifer mexicanus AccessionID-NZ_CP041238.1	ITTG R7	a	*Acacia cochliacantha* and *Phaseol*	Alpha			
Ensifer sojae CCBAU 05684 AccessionID-NZ_CP023067.1		a	*Glycine max*, *G. soja* and *Vigna un*	Alpha			
Epibacterium mobile AccessionID-NZ_LR027553.1		c		Alpha			
Haematospirillum jordaniae AccessionID-NZ_CP014525.1	H5569	c		Alpha			
Indioceanicola profundi AccessionID-NZ_CP030126.1	SCSIO 08	c		Alpha			
Iodobacter sp. H11R3 AccessionID-NZ_CP034433.1		c		Beta			
Janthinobacterium agaricidamnosum AccessionID-NZ_CP033019.1	BHSEK	c		Beta			
Ketogulonicigenium vulgare AccessionID-NZ_CP016592.1	SKV	c		Alpha			
Labrenzia sp. THAF82 AccessionID-NZ_CP045354.1		c		Alpha			
Labrys sp. KNU-23 AccessionID-NZ_CP043489.1		c		Alpha			
Leisingera methylohalidivorans DSM 14336 AccessionID-NC_023135.1		c		Alpha			
Leisingera sp. NJS201 AccessionID-NZ_CP038234.1		c		Alpha			
Litoreibacter sp. LN3S51 AccessionID-NZ_CP042261.1		c		Alpha			
Maritalea myrionectae AccessionID-NZ_CP021330.1	HL2708#5	c		Alpha			
Marivivens sp. JLT3646 AccessionID-NZ_CP018572.1		c		Alpha			
Martelella endophytica AccessionID-NZ_CP010803.1	YC6887	c		Alpha			
Martelella mediterranea DSM 17316 AccessionID-NZ_CP020330.1	MACL11	c		Alpha			



Martelella sp. AD-3 AccessionID-NZ_CP014275.1		c		Alpha	
Martelella sp. NC18 AccessionID-NZ_CP054858.1		c		Alpha	
Martelella sp. NC20 AccessionID-NZ_CP054861.1		c		Alpha	
Massilia putida AccessionID-NZ_CP019038.1	6NM-7	c		Beta	
Mesorhizobium amorphae CCNWGS0123 AccessionID-NZ_CP015318.1		a	*Robinia pseudoacacia*	Alpha	
Mesorhizobium australicum WSM2073 AccessionID-NC_019973.1		a	*Biserrula pelecinus* and *Astragalus*	Alpha	
Mesorhizobium ciceri AccessionID-NZ_CP015062.1		a	*Cicer arietinum*	Alpha	
Mesorhizobium ciceri biovar biserrulae AccessionID-NZ_CP015064.1		a	*Biserrula pelecinus*	Alpha	
Mesorhizobium ciceri biovar biserrulae WSM1271 AccessionID-NC_014923.1		a	*Biserrula pelecinus* and *Astragalus*	Alpha	
Mesorhizobium erdmanii AccessionID-NZ_CP033361.1	NZP2014	a	*Lotus corniculatus* and *L. peduncu*	Alpha	
Mesorhizobium huakuii AccessionID-NZ_CP050296.1	583	a	*Oxytropis kamtschatica*	Alpha	
Mesorhizobium japonicum AccessionID-NZ_CP051773.1	R7Astar	a	*Lotus corniculatus*	Alpha	
Mesorhizobium japonicum AccessionID-NZ_CP052769.1	R7AstarV	a	*Lotus corniculatus*	Alpha	
Mesorhizobium japonicum AccessionID-NZ_CP052770.1	R7ANSsta	a	*Lotus japonicum*	Alpha	
Mesorhizobium japonicum MAFF 303099 AccessionID-NC_002678.2		a	*Lotus japonicum*	Alpha	
Mesorhizobium japonicum R7A AccessionID-NZ_CP033366.1		a	*Lotus corniculatus*	Alpha	
Mesorhizobium japonicum R7A AccessionID-NZ_CP051772.1		a	*Lotus corniculatus*	Alpha	
Mesorhizobium jarvisii AccessionID-NZ_CP033507.1	ATCC 700	a	*Lotus corniculatus* and *L. peduncu*	Alpha	
Mesorhizobium loti AccessionID-NZ_CP033334.1	NZP2042	a	*Lotus pedunculatus*	Alpha	
Mesorhizobium loti AccessionID-NZ_CP033368.1	SU343	a	*Lotus corniculatus*	Alpha	
Mesorhizobium loti AccessionID-NZ_CP050293.1	582	a	*Oxytropis kamtschatica*	Alpha	
Mesorhizobium loti NZP2037 AccessionID-NZ_CP016079.1		a	*Lotus divaricatus*	Alpha	
Mesorhizobium loti R88b AccessionID-NZ_CP033367.1		a	*Lotus corniculatus*	Alpha	
Mesorhizobium opportunistum WSM2075 AccessionID-NC_015675.1		a	*Lotus peregrinus* and *Macroptilium atropu*	Alpha	
Mesorhizobium sp. 8 AccessionID-NZ_CP040914.1		c		Alpha	
Mesorhizobium sp. AA22 AccessionID-NZ_CP048406.1		b	*Astragalus pelecinus* nod	Alpha	
Mesorhizobium sp. M1B.F.Ca.ET.045.04.1.1 AccessionID-NZ_CP034448.1		a	*Cicer arietinum*	Alpha	
Mesorhizobium sp. M1D.F.Ca.ET.043.01.1.1 AccessionID-NZ_CP034444.1		a	*Cicer arietinum*	Alpha	
Mesorhizobium sp. M1E.F.Ca.ET.045.02.1.1 AccessionID-NZ_CP034447.1		a	*Cicer arietinum*	Alpha	
Mesorhizobium sp. M2A.F.Ca.ET.043.02.1.1 AccessionID-NZ_CP034445.1		a	*Cicer arietinum*	Alpha	
Mesorhizobium sp. M2A.F.Ca.ET.043.05.1.1 AccessionID-NZ_CP034446.1		a	*Cicer arietinum*	Alpha	
Mesorhizobium sp. M2A.F.Ca.ET.046.03.2.1 AccessionID-NZ_CP034449.1		a	*Cicer arietinum*	Alpha	
Mesorhizobium sp. M3A.F.Ca.ET.080.04.2.1 AccessionID-NZ_CP034451.1		a	*Cicer arietinum*	Alpha	
Mesorhizobium sp. M9A.F.Ca.ET.002.03.1.2 AccessionID-NZ_CP034443.1		a	*Cicer arietinum*	Alpha	
Mesorhizobium sp. NZP2077 AccessionID-NZ_CP033362.1		c		Alpha	
Mesorhizobium sp. NZP2077 AccessionID-NZ_CP051293.1		b	*Lotus corniculatus* nodul	Alpha	
Mesorhizobium sp. NZP2234 AccessionID-NZ_CP033364.1		b	*Lotus corniculatus* nodul	Alpha	
Mesorhizobium sp. NZP2298 AccessionID-NZ_CP033365.1		b	*Lotus corniculatus* nodul	Alpha	
Mesorhizobium sp. Pch-S AccessionID-NZ_CP029562.1		c		Alpha	
Mesorhizobium sp. WSM1497 AccessionID-NZ_CP021070.1		a	*Biserrula pelecinus*	Alpha	
Methylobacterium sp. 17Sr1-43 AccessionID-NZ_CP029551.1		c		Alpha	
Methylobacterium sp. 4-46 AccessionID-NC_010511.1		a	*Lotononis bainesii*, *L. listii* and *L. su*	Alpha	
Methylobacterium sp. C1 AccessionID-NZ_CP017640.1		c		Alpha	
Methylobacterium sp. DM1 AccessionID-NZ_CP029173.1		c		Alpha	
Methylobacterium terrae AccessionID-NZ_CP029553.1	17Sr1-28	c		Alpha	
Neorhizobium galegae bv. officinalis bv. officinalis str. HAMBI 1141 AccessionID-NZ_H		a	*Galega officinalis*	Alpha	
Neorhizobium galegae bv. orientalis str. HAMBI 540 AccessionID-NZ_H		a	*Galega orientalis*	Alpha	
Neorhizobium sp. NCHU2750 AccessionID-NZ_CP030827.1		c		Alpha	
Nitratireductor sp. OM-1 AccessionID-NZ_CP029208.1		c		Alpha	
Nitratireductor sp. SY7 AccessionID-NZ_CP042301.2		c		Alpha	
Nitrobacter hamburgensis X14 AccessionID-NC_007964.1		c		Alpha	
Nitrospirillum amazonense CBAmc AccessionID-NZ_CP022110.1		c		Alpha	
Novosphingobium sp. THN1 AccessionID-NZ_CP028347.1		c		Alpha	
Ochrobactrum pseudogrignonense AccessionID-NZ_CP015775.1	K8	c		Alpha	Mudou para Brucella pseudogrignonensis Strain K8
Paraburkholderia atlantica AccessionID-NC_014117.1		b	*Mimosa occidentalis* nod	Beta	
Paraburkholderia caledonica AccessionID-NZ_CP024905.1	PHRS4	c		Beta	
Paraburkholderia caribensis AccessionID-NZ_CP013102.1	MWAP64	c		Beta	
Paraburkholderia caribensis AccessionID-NZ_CP026101.1	DSM 1323	c		Beta	
Paraburkholderia fungorum AccessionID-NZ_CP010026.1	ATCC BAA	c		Beta	
Paraburkholderia phenoliruptrix BR3459a AccessionID-NC_018695.1		a	*Mimosa flocculosa*	Beta	
Paraburkholderia phymatum STM815 AccessionID-NC_010622.1		a	*Phaseolus vulgaris*	Beta	
Paraburkholderia sp. 7Q-K02 AccessionID-NZ_CP046909.1		c		Beta	
Paraburkholderia sp. PGU19 AccessionID-NZ_AP023179.1		c		Beta	



Paraburkholderia sprentiae WSM5005 AccessionID-NZ_CP017561.1		a	*Lebeckia sepiaria* and *Lebeckia an*	Beta	
Paracoccus aminovorans AccessionID-NZ_LN832559.1		c		Alpha	
Paracoccus denitrificans AccessionID-NZ_CP035090.1		c		Alpha	
Paracoccus kondratievae AccessionID-NZ_CP045072.1	BJQ0001	c		Alpha	
Paracoccus pantotrophus AccessionID-NZ_CP044426.1	DSM 2944	c		Alpha	
Paracoccus yeei AccessionID-NZ_CP020442.2	FDAARG(	c		Alpha	
Paracoccus yeei AccessionID-NZ_CP044081.1	FDAARG(	c		Alpha	
Phenylobacterium zucineum HLK1 AccessionID-NC_011144.1		c		Alpha	
Phreatobacter cathodiphilus AccessionID-NZ_CP027668.1	S-12	c		Alpha	
Phyllobacterium sp. 628 AccessionID-NZ_CP050301.1		d		Alpha	Oxytropis pumilio nod Safronova 2020
Phyllobacterium zundukense AccessionID-NZ_CP017940.1	Tri-48; R	b	*Oxytropis triphylla* nodul	Alpha	
Rhizobium acidisoli AccessionID-NZ_CP034998.1	FH23	b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium esperanzae AccessionID-NZ_CP013500.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium etli 8C-3 AccessionID-NZ_CP017241.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium etli AccessionID-NZ_CP020906.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium etli bv. mimosae str. Mim1 AccessionID-NC_021905.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium etli bv. phaseoli str. IE4803 AccessionID-NZ_CP007641.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium etli CFN 42 AccessionID-NC_007761.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium etli CIAT 652 AccessionID-NC_010994.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium gallicum AccessionID-NZ_CP017101.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium grahamii AccessionID-NZ_CP043498.1	BG7	b	*Prosopis cineraria* nodul	Alpha	
Rhizobium indicum AccessionID-NZ_CP054021.1	JKLM 12A	b	*Pisum sativum* nodules	Alpha	
Rhizobium indicum AccessionID-NZ_CP054031.1	JKLM 13E	b	*Pisum sativum* nodules	Alpha	
Rhizobium jaguaris AccessionID-NZ_CP032694.1		a	*Calliandra grandiflora*	Alpha	
Rhizobium leguminosarum AccessionID-NZ_CP016286.1	Vaf10	b	*Vavilovia formosa* nodule	Alpha	
Rhizobium leguminosarum AccessionID-NZ_CP018228.1	Vaf-108	b	*Vavilovia formosa* nodule	Alpha	
Rhizobium leguminosarum AccessionID-NZ_CP030760.1	ATCC 144	a	*Trifolium pratense*	Alpha	
Rhizobium leguminosarum bv. trifolii AccessionID-NZ_CP050080.1	31B	b	*Trifolium pratense* nodul	Alpha	
Rhizobium leguminosarum bv. trifolii AccessionID-NZ_CP050085.1	23B	b	*Trifolium pratense* nodul	Alpha	
Rhizobium leguminosarum bv. trifolii AccessionID-NZ_CP050091.1	22B	b	*Trifolium pratense* nodul	Alpha	
Rhizobium leguminosarum bv. trifolii AccessionID-NZ_CP050097.1	9B	b	*Trifolium pratense* nodul	Alpha	
Rhizobium leguminosarum bv. trifolii AccessionID-NZ_CP050103.1	4B	b	*Trifolium pratense* nodul	Alpha	
Rhizobium leguminosarum bv. trifolii AccessionID-NZ_CP050108.1	3B	b	*Trifolium pratense* nodul	Alpha	
Rhizobium leguminosarum bv. trifolii AccessionID-NZ_CP053439.1	CC275e	a	*Trifolium repens*	Alpha	
Rhizobium leguminosarum bv. trifolii CB782 AccessionID-NZ_CP007067.1		a	*Trifolium semipilosum*	Alpha	
Rhizobium leguminosarum bv. trifolii TA1 AccessionID-NZ_CP053205.2		a	*Trifolium spumosum*	Alpha	
Rhizobium leguminosarum bv. trifolii WSM1325 AccessionID-NC_012850.1		a	*Trifolium subterraneum*	Alpha	
Rhizobium leguminosarum bv. trifolii WSM1689 AccessionID-NZ_CP007045.1		a	*Trifolium uniflorum*	Alpha	
Rhizobium leguminosarum bv. trifolii WSM2304 AccessionID-NC_011369.1		a	*Trifolium polymorphum*	Alpha	
Rhizobium leguminosarum bv. viciae 248 AccessionID-NZ_CP048280.1		a	*Vicia sativa* and *V. hirsuta*	Alpha	
Rhizobium leguminosarum bv. viciae 3841 AccessionID-NC_008380.1		a	*Pisum sativum* and *Vicia cracca*	Alpha	
Rhizobium leguminosarum bv. viciae AccessionID-NZ_CP022665.1	BIHB 121	b	*Pisum sativum* nodules	Alpha	
Rhizobium leguminosarum bv. viciae AccessionID-NZ_CP025509.1	UPM791	a	*Pisum sativum* L. cv. Frisson	Alpha	
Rhizobium phaseoli AccessionID-NZ_CP013522.1	R744	b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium phaseoli AccessionID-NZ_CP013527.1	R723	b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium phaseoli AccessionID-NZ_CP013532.1	R650	b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium phaseoli AccessionID-NZ_CP013537.1	R630	b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium phaseoli AccessionID-NZ_CP013542.1	R620	b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium phaseoli AccessionID-NZ_CP013547.1	R611	b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium phaseoli AccessionID-NZ_CP013552.1	N931	b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium phaseoli AccessionID-NZ_CP013557.1	N841	b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium phaseoli AccessionID-NZ_CP013563.1	N831	b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium phaseoli AccessionID-NZ_CP013568.1	N771	b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium phaseoli AccessionID-NZ_CP013574.1	N671	b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium phaseoli AccessionID-NZ_CP013580.1	N261	b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium phaseoli AccessionID-NZ_CP013585.1	N161	b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium phaseoli AccessionID-NZ_CP064931.1	BS3	b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium phaseoli Brasil 5 AccessionID-NZ_CP020896.1	Bra5	a	*Phaseolus vulgaris*	Alpha	
Rhizobium pseudoryzae AccessionID-NZ_CP049241.1	DSM 1947	c		Alpha	
Rhizobium pusense AccessionID-NC_022535.1		a	*Sesbania cannabina*	Alpha	
Rhizobium pusense AccessionID-NZ_CP039894.1	CFBP587	c		Alpha	
Rhizobium sp. 007 AccessionID-NZ_CP064187.1		b	*Onobrychis viciifolia* nod	Alpha	
Rhizobium sp. 11515TR AccessionID-NZ_CP022998.1		a	*Centrosema pubescens*, *Phaseolu*	Alpha	

Rhizobium sp. ACO-34A AccessionID-NZ_CP021371.1		c		Alpha	
Rhizobium sp. CCGE531 AccessionID-NZ_CP032684.1		a	*Phaseolus vulgaris* and *Leucaena*	Alpha	
Rhizobium sp. CCGE532 AccessionID-NZ_CP032689.1		a	*Phaseolus vulgaris* and *Leucaena*	Alpha	
Rhizobium sp. CIAT894 AccessionID-NZ_CP020947.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium sp. IE4771 AccessionID-NZ_CP006986.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium sp. Khangiran2 AccessionID-NZ_LR723668.1		c		Alpha	
Rhizobium sp. Kim5 AccessionID-NZ_CP021124.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium sp. N113 AccessionID-NZ_CP013517.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium sp. N1314 AccessionID-NZ_CP013511.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium sp. N1341 AccessionID-NZ_CP013505.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium sp. N324 AccessionID-NZ_CP013630.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium sp. N621 AccessionID-NZ_CP013495.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium sp. N6212 AccessionID-NZ_CP013490.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium sp. N731 AccessionID-NZ_CP013601.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium sp. N741 AccessionID-NZ_CP013595.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium sp. N871 AccessionID-NZ_CP013590.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium sp. NIBRBAC000502774 AccessionID-NZ_CP041204.1		c		Alpha	
Rhizobium sp. NXC14 AccessionID-NZ_CP021030.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium sp. NXC24 AccessionID-NZ_CP024311.1		b	*Phaseolus vulgaris* nodu	Alpha	
Rhizobium sp. S41 AccessionID-NZ_CP016320.1		c		Alpha	
Rhizobium sp. TAL182 AccessionID-NZ_CP021024.1		a	*Phaseolus vulgaris*	Alpha	
Rhizobium tropici CIAT 899 AccessionID-NC_020059.1		a	*Acacia nilotica* and *Phaseolus vulg*	Alpha	
Rhodobacter blasticus AccessionID-NZ_CP020470.1	28/5	c		Alpha	mudou para Fuscovulum blasticum strain 28/5
Rhodopseudomonas palustris (Rhodopseudomonas rutila) AccessionID-NZ_CP019967	YSC3	c		Alpha	
Rhodopseudomonas palustris (Rhodopseudomonas rutila) AccessionID-NZ_CP058907	CGMCC 1	c		Alpha	
Rhodopseudomonas palustris BisA53 AccessionID-NC_008435.1		c		Alpha	
Rhodopseudomonas palustris DX-1 AccessionID-NC_014834.1		c		Alpha	
Rhodovulum sulfidophilum AccessionID-NZ_CP015421.1	SNK001	c		Alpha	
Roseovarius sp. AK1035 AccessionID-NZ_CP030999.1		c		Alpha	
Ruegeria sp. AD91A AccessionID-NZ_CP031946.1		c		Alpha	
Sinorhizobium americanum AccessionID-NZ_CP013107.1		a	*Leucaena leucocephala* and *Phasu*	Alpha	
Sinorhizobium americanum CCGM7 AccessionID-NZ_CP013051.1		a	*Phaseolus vulgaris* and *Medicago*	Alpha	
Sinorhizobium fredii AccessionID-NZ_CP024307.1		a	*Glycine max*	Alpha	
Sinorhizobium fredii CCBAU 25509 AccessionID-NZ_CP029451.1		a	*Glycine soja*	Alpha	
Sinorhizobium fredii CCBAU 45436 AccessionID-NZ_CP029231.1		a	*Glycine soja*	Alpha	
Sinorhizobium fredii CCBAU 83666 AccessionID-NZ_CP023070.1		b	*Glycine max* nodules	Alpha	
Sinorhizobium fredii NGR234 AccessionID-NC_012587.1		a	*Phaseolus vulgaris* - 112 legume g	Alpha	
Sinorhizobium medicae WSM419 (Ensifer medicae WSM419) AccessionID-NC_009636.1		a	*Marchatia polymorpha* and *M. ara*	Alpha	
Sinorhizobium meliloti 1021 AccessionID-NC_003047.1		a	*Medicago sativa*	Alpha	
Sinorhizobium meliloti 2011 AccessionID-NC_020528.1		a	*Medicago sativa*	Alpha	
Sinorhizobium meliloti AccessionID-NZ_CP009144.1	RMO17	b	*Medicago orbicularis* noc	Alpha	
Sinorhizobium meliloti AccessionID-NZ_CP019485.1	B401	a	*Medicago sativa*	Alpha	
Sinorhizobium meliloti AccessionID-NZ_CP019488.1	B399	a	*Medicago sativa*	Alpha	
Sinorhizobium meliloti AccessionID-NZ_CP019584.1	CCMM B5	a	*Medicago truncatula*	Alpha	
Sinorhizobium meliloti AccessionID-NZ_CP021793.1	USDA115	a	*Medicago sativa*	Alpha	
Sinorhizobium meliloti AccessionID-NZ_CP021797.1	USDA110	b	*Medicago sativa*	Alpha	
Sinorhizobium meliloti AccessionID-NZ_CP021800.1	USDA102	a	*Medicago truncatula*	Alpha	
Sinorhizobium meliloti AccessionID-NZ_CP021804.1	T073	a	*Medicago sativa*	Alpha	
Sinorhizobium meliloti AccessionID-NZ_CP021808.1	Rm41	a	*Medicago sativa*	Alpha	
Sinorhizobium meliloti AccessionID-NZ_CP021818.1	M162	a	*Medicago truncatula*	Alpha	
Sinorhizobium meliloti AccessionID-NZ_CP021822.1	KH46	a	*Medicago sativa*	Alpha	
Sinorhizobium meliloti AccessionID-NZ_CP021825.1	KH35c	a	*Medicago sativa*	Alpha	
Sinorhizobium meliloti AccessionID-NZ_CP021829.1	HM006	a	*Medicago sativa*	Alpha	
Sinorhizobium meliloti AccessionID-NZ_CP026525.1	AK21	a	*Medicago sativa*	Alpha	
Sinorhizobium meliloti AK83 (Ensifer meliloti AK83) AccessionID-NC_015590.1		a	*Medicago sativa*	Alpha	
Sinorhizobium meliloti BL225C (Ensifer meliloti BL225C) AccessionID-NC_017322.1		a	*Medicago sativa*	Alpha	
Sinorhizobium meliloti GR4 AccessionID-NC_019845.2		a	*Medicago sativa*	Alpha	
Sinorhizobium meliloti Rm41 AccessionID-NC_018700.1		a	*Medicago trunculata*	Alpha	
Sinorhizobium meliloti RU11/001 AccessionID-NZ_CP021219.1		a	*Medicago sativa*	Alpha	
Sinorhizobium meliloti SM11 AccessionID-NC_017325.1		a	*Medicago sativa*	Alpha	
Sinorhizobium sp. CCBAU 05631 AccessionID-NZ_CP023063.1		a	*Glycine max*	Alpha	
Sinorhizobium sp. RAC02 AccessionID-NZ_CP016450.1		c		Alpha	
Sphingobium sp. RAC03 AccessionID-NZ_CP016456.1		c		Alpha	

Sphingobium yanoikuyae AccessionID-NZ_CP020925.1	SHJ	c	Alpha
Sphingomonas panacis AccessionID-NZ_CP014168.1	DCY99	c	Alpha
Sphingopyxis sp. 113P3 AccessionID-NZ_CP009452.1		c	Alpha
Sphingosinicella sp. BN140058 AccessionID-NZ_CP035501.1		c	Alpha
Sphingosinithalassobacter sp. zrk23 AccessionID-NZ_CP049109.1		c	Alpha
Sulfitobacter sp. AM1-D1 AccessionID-NZ_CP018076.1		c	Alpha
Varibacter gotjawalensis AccessionID-NZ_AP014946.1	GJW-30	c	Alpha
Xanthobacter autotrophicus Py2 AccessionID-NC_009720.1		c	Alpha

a, Nodulation ability was confirmed in nodulation tests.
b, Isolated from nodules however nodulation ability was not tested.
c, No data found about nodulation.
d, The strain failed in nodulation tests.

320 predições realizadas pelo modelo SVM único, em laranja estão as que não foram preditas pelo SVM *Ensemble*.

APÊNDICE 2 – ESCOLHA DE ALGORITMOS DE APRENDIZAGEM SUPERVISIONADA

Para construir o modelo de aprendizagem supervisionada que foi chamado de NodProspect, um dos passos iniciais foi a escolha do algoritmo mais promissor para aplicar ao nosso contexto. Para isso testamos os algoritmos apresentados na Tabela A2 utilizando a base de dados de treinamento criada, utilizando *cross-validation* 10 *fold* para verificar a performance. A SVM com o *Gaussian kernel* foi selecionada por apresentar a melhor performance neste teste inicial e foi otimizada por otimização bayesiana posteriormente.

Tabela A2 – Validação dos algoritmos de aprendizagem supervisionada

Algoritmo	*Cross-validation accuracy*	PCA
Fine tree	77.2%	Não
Medium tree	77.2%	Não
Coarse tree	75.5%	Não
Linear discriminant	89.9%	Não
Logistic regression	72.8%	Não
SVM Linear	89.9%	Não
SVM Gaussian	91.0%	Não
SVM Cubic	90.2%	Não
SVM Fine Gaussian	68.8%	Não
SVM Medium Gaussian	90.4%	Não
SVM Coarse Gaussian	88.2%	Não
KNN Fine	85.7%	Não
KNN Medium	78.1%	Não
KNN Coarse	69.7%	Não
KNN Cosine	89.0%	Não
KNN Cubic	77.2%	Não
KNN Weighted	80.6%	Não

Ensemble boosted trees	57.3%	Não
Ensemble bagged trees	87.1%	Não
Ensemble subspace discriminant	75.0%	Não
Ensemble subspace KNN	84.8%	Não
Ensemble RUSBossted trees	59.3%	Não
Fine tree	87.6%	95% da variância
Medium tree	87.6%	95% da variância
Coarse tree	85.1%	95% da variância
Linear discriminant	89.6%	95% da variância
Logistic regression	88.2%	95% da variância
SVM Linear	90.2%	95% da variância
SVM Gaussian	89.6%	95% da variância
SVM Cubic	88.2%	95% da variância
SVM Fine Gaussian	88.5%	95% da variância
SVM Medium Gaussian	79.2%	95% da variância
SVM Coarse Gaussian	75.6%	95% da variância
KNN Fine	78.7%	95% da variância
KNN Medium	60.1%	95% da variância
KNN Coarse	50.0%	95% da variância
KNN Cosine	88.8%	95% da variância
KNN Cubic	61.8%	95% da variância
KNN Weighted	70.2%	95% da variância
Ensemble boosted trees	53.1%	95% da variância
Ensemble bagged trees	90.2%	95% da variância
Ensemble subspace discriminant	89.6%	95% da variância
Ensemble subspace KNN	81.7%	95% da variância

Ensemble RUSBoosted trees	53.1%	95% da variância
MLP	85.1%	Não