UNIVERSIDADE FEDERAL DO PARANÁ

ELISA TERUMI RUBEL

DESENVOLVIMENTO DE PROCLAT, UMA FERRAMENTA COMPUTACIONAL PARA A CLASSIFICAÇÃO DE PROTEÍNAS: O CASO "DraB" DE *Azospirillum brasilense*

.

CURITIBA

2015

ELISA TERUMI RUBEL

DESENVOLVIMENTO DE PROCLAT, UMA FERRAMENTA COMPUTACIONAL PARA A CLASSIFICAÇÃO DE PROTEÍNAS: O CASO "DraB" DE *Azospirillum brasilense*

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Bioinformática, no curso de Pós-Graduação em Bioinformática, Setor e Educação Profissional e Tecnológica da Universidade Federal do Paraná.

Orientador: Prof. Dr. Fábio de Oliveira Pedrosa
Coorientador: Prof. Dr. Roberto Tadeu Raittz

CURITIBA

2015

R894d

Rubel, Elisa Terumi
Desenvolvimento de PROCLAT, uma ferramenta computacional para a classificação de proteínas : o caso "DraB" de Azospirillum brasilense/ Elisa Terumi Rubel. – Curitiba, 2015.
96 f. : il. color. ; 30 cm.

Dissertação - Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Programa de Pós-graduação em Bioinformática, 2015.

Orientador: Fábio de Oliveira Pedrosa – Co-orientador: Roberto Tadeu Raittz.
Bibliografia: p. 61-63.

1. Bioinformática. 2. Fixação biológica de nitrogênio. 3. Redes neurais (Computação). 4. Proteínas - Classificação. 5. Nitrogenase. I. Universidade Federal do Paraná. II.Pedrosa, Fábio de Oliveira. III. Raittz, Roberto Tadeu . IV. Título.

CDD: 572.60285

TERMO DE APROVAÇÃO

ELISA TERUMI RUBEL

"PROCLAT, ferramenta computacional para a classificação de proteínas: o caso "DRAB" de *Azospirillum brasilense*"

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

Orientador: Prof. Dr. Fábio de Oliveira Pedrosa

Coorientador: Prof. Dr. Roberto Tadeu Raittz

Profª Drª Jeroniza Nunes Marchaukoski
Universidade Federal do Paraná

Prof. Dr. Luciano Fernandes Huergo
Universidade Federal do Paraná

Profª Drª Ana Claudia Bonatto
Universidade Federal do Paraná

Curitiba, 14 de maio de 2015

Para Henrique, Nicole, Erick, Totian, Catian, Fernanda e Paula.
Amo vocês.

AGRADECIMENTOS

A Deus, por todas as bênçãos recebidas, e por ter me conduzido até aqui.

A Universidade Federal do Paraná, ao Programa de Pós-Graduação em Bioinformática e ao Núcleo de Fixação de Nitrogênio.

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

Aos meus orientadores, Professores Fábio de Oliveira Pedrosa e Roberto Tadeu Raittz, pela sabedoria, paciência e confiança, meus grandes Mestres que me guiaram nesses dois anos de trabalho.

A todos professores, à Coordenação e Secretaria do Programa de Pós-Graduação em Bioinformática.

Aos colegas, pela amizade e companheirismo, e pelos trabalhos desenvolvidos em equipe.

A Celepar, Companhia de Tecnologia da Informação e Comunicação do Paraná, empresa onde com muito orgulho trabalho há quase 10 anos, pelo apoio à minha qualificação profissional, liberando algumas horas por semana para minha dedicação ao curso.

Por último, e não menos importante, à minha querida família: papai (Totian), mamãe (Catian) e irmãs (Paula e Fernanda), e em especial ao marido Fernando Henrique Schneider que me apoiou e incentivou, e aos meus filhos Nicole e Erick, ainda a caminho, que sem saber são a minha fonte de inspiração.

Agradeço profundamente, e afirmo que é um privilégio fazer parte do PPG em Bioinformática da UFPR!

Muito obrigada!

"Aprender é a única coisa de que a mente nunca se cansa, nunca tem medo e nunca se arrepende."

Leonardo da Vinci

RESUMO

Azopirillum brasilense é uma bactéria fixadora de nitrogênio que promove o crescimento vegetal e é utilizada como bio-fertilizante na agricultura. Uma vez que o processo demanda um alto gasto de energia, a redução de N_2 para NH_4^+ pela enzima nitrogenase ocorre apenas em condições limitantes de NH_4^+ e O_2. Além disso, a síntese e atividade da nitrogenase são altamente reguladas para evitar o desperdício de energia. Em *A. brasilense,* a atividade da nitrogenase é regulada pelo produto dos genes *draG* e *draT*. O produto do gene *draB*, localizado a jusante do gene *draG* no operon *draTGB*, pode estar envolvido na regulação da atividade da nitrogenase, por um mecanismo ainda desconhecido. Uma análise *in silico* do produto do gene *draB* foi realizada com o objetivo de definir a sua função e provável envolvimento na regulação da fixação de nitrogênio em *A. brasilense*. Neste trabalho, apresentamos uma abordagem envolvendo inteligência artificial/rede neural de aprendizado de máquina para a classificação de proteínas. Esta ferramenta, denominada ProClaT, sugere que o gene *draB* codifica para a proteína NifO associada à nitrogenase. Esta ferramenta permitiu a reclassificação de proteínas homologas, hipotéticas, conservadas hipotéticas ou anotadas como prováveis arsenato redutases, ArsC, depositadas nos bancos de dados biológicos, em NifO. Com os resultados obtidos, uma análise de co-ocorrência dos genes *draB, draT e draG* e outros genes *nif* foi realizada, sugerindo o envolvimento do *draB* (*nifO*) na fixação de nitrogênio, embora sem a definição de uma função específica.

Palavras-chave: Bioinformática, Fixação Biológica de Nitrogênio, Redes Neurais Artificiais, Classificação de Proteínas, Nitrogenase.

ABSTRACT

Azopirillum brasilense is a plant-growth promoting nitrogen-fixing bacteria used as bio-fertilizer in agriculture. Since this process has a high energy demand, the reduction of N_2 to NH_4^+ by nitrogenase occurs only under limiting conditions of NH_4^+ and O_2. Moreover, the synthesis and activity of nitrogenase is highly regulated to prevent energy waste. In *A. brasilense* nitrogenase activity is regulated by the products of *draG* and *draT*. The product of the *draB* gene, located downstream in the *draTGB* operon, may be involved in the regulation of nitrogenase activity by an, as yet, unknown mechanism. A deep *in silico* analysis of the product of *draB* was undertaken aiming at defining its function and probable involvement in the regulation of nitrogenase activity in *A. brasilense*. In this work, we present an artificial inteligence/neural network, machine-learning approach for protein classification. This tool, named ProClaT, suggests that the *draB* gene codes for the nitrogenase associated NifO-like protein. This tool allowed the reclassification of homologous proteins, hypothetical, conserved hypothetical or annoted as putative arsenate reductase, ArsC, deposited in biological databases, into NifO. Based on these results, an analysis of co-ocurrence of *draB*, *draT*, *draG* and the others *nif* genes was performed, suggesting the involvement of *draB* (*nifO*) in nitrogen fixation, without a definition of a specific function.

Keywords: Bioinformatics, Biological Nitrogen Fixation, Artificial Neural Networks, Protein Classification, Nitrogenase.

LISTA DE FIGURAS - ARTIGO CIENTÍFICO

LISTA DE TABELAS - ARTIGO CIENTÍFICO

LISTA DE ABREVIATURAS E SIGLAS

BLAST	Basic Local Alignment Search Tool
DNA	Ácido desoxirribonucléico
ExPASy	Expert Protein Analysis System
FBN	Fixação Biológica de Nitrogênio
FeMoCo	Cofator Ferro Molibdênio
GO	Gene Ontology
HMM	Hidden Markov Model (Modelo Oculto de Markov)
MATLAB	Matrix Laboratory
MLP	MultiLayer Perceptron
NCBI	National Center for Biotechnology Information
NCBI NR	NCBI Non-redundant protein sequences
ORF	Open Reading Frame
PDB	Protein Data Bank
ProClaT	Protein Classifier Tool
RNA	Ácido ribonucleico
Weka	Waikato Environment for Knowledge Analysis

SUMÁRIO

1 INTRODUÇÃO

Azospirillum brasilense é uma bactéria fixadora de nitrogênio, utilizada como biofertilizante na agricultura, promovendo o crescimento vegetal (PEDROSA, 1987). Dotada de uma via metabólica específica para a conversão do nitrogênio gasoso em amônia, o N_2 é fixado sob condições limitantes de NH_4^+ e O_2, através da atividade da enzima nitrogenase, cuja síntese e atividade são reguladas por mecanismos específicos (PICHETH *et al.*, 1999; POSTGATE, 1982 [2]).

Por ser um processo de alto custo energético para a célula, a fixação de nitrogênio só ocorre em condições favoráveis, ou seja, em condições limitantes de NH_4^+ e O_2 (HALBLEIB *et al.*, 2000). Um dos controles da nitrogenase é o pós-traducional (ZUMFT e CASTILLO, 1978) [3], sendo em *A. brasilense*, através do sistema DraG-DraT. A DraT (dinitrogenase ADP-ribosil transferase) é uma enzima que atua no desligamento da nitrogenase pela inativação da dinitrogenase redutase (NifH) em resposta à adição de íons de amônio no meio de cultivo, enquanto a DraG (dinitrogenase redutase glicohidrolase) é a enzima que restaura a atividade da NifH, após o consumo dos ions amônio (HUERGO *et al*., 2012) [4]. Segundo Zhang *et al.* (1992) [5], as enzimas DraT e DraG são codificadas pelos genes *draTG*, que constituem um operon *draTGB* em *A. brasilense*. A jusante do gene *draG* foi sequenciada uma pequena região, similar ao produto do gene *draB* em *Rhodospirillum rubrum* (ZHANG *et al.*, 1992) [5]. Nessa bactéria, estudos sugerem que o gene *draB* regula a atividade das proteínas DraT e DraG (LIANG *et al.,* 1991) [6]. Em *A. brasilense*, o gene *draB* está anotado como codificante para putative arsenate reductase no GenBank e, segundo Zhang *et al*. (2001), apresenta similaridade com os genes *nifO* de *A. vinelandii* e *arsC* de *E. coli*.

O estudo do produto do gene *draB* deverá contribuir para uma melhor compreensão do mecanismo de regulação pós-traducional da enzima nitrogenase, não apenas em *A. brasilense* mas em outros organismos fixadores de nitrogênio.

Por ser ainda uma proteína sem função biológica conhecida, o desenvolvimento de um método *in silico* que permita a análise de seus domínios estruturais em comparação com aqueles de proteínas homólogas presentes nos

bancos de dados poderá sugerir eventuais funções para a proteína DraB de *Azospirillum brasilense* e outros diazotrofos.

O objetivo geral do trabalho, portanto, é desenvolver uma metodologia de Bioinformática para a classificação de proteínas, visando determinar *in silico* a provável função da proteína codificada pelo gene *draB* de *A. brasilense*.

Os objetivos específicos são:

1. Desenvolver abordagem geral para a reclassificação de proteínas, utilizando aprendizado de máquina/ redes neurais artificiais;
2. Realizar análise de co-ocorrência do gene *draB* em relação aos genes *nif* essenciais, *nifHDK* e *nifENB*;
3. Realizar análise de co-ocorrência do gene *draB* em relação aos genes *draT* e *draG*;
4. Realizar análise de vizinhança do gene *draB*;
5. Reclassificar a proteína DraB de *A. brasilense*;
6. Reclassificar proteínas hipotéticas e putativas com a abordagem desenvolvida.

Os resultados deste estudo originaram o artigo científico "***ProClaT, a new bioinformatics approach for in silico protein reclassification: case study of DraB, a putative arsenate reductase protein of Azospirillum brasilense***", apresentado na seção 3 do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 FIXAÇÃO BIOLÓGICA DE NITROGÊNIO

O nitrogênio- elemento fundamental para manutenção da vida- compõe cerca de 80% da atmosfera terrestre, porém em sua forma mais abundante: N_2 (gás dinitrogênio). A forma metabolicamente assimilável pelos organismos é NH_4 (amônio). A Fixação Biológica de Nitrogênio (FBN) é um processo do Ciclo do Nitrogênio (Figura 1), na qual ocorre a redução de N_2 a NH_3 pela enzima nitrogenase, presente nas bactérias diazotróficas, ou fixadoras de nitrogênio. Esta capacidade está presente nos Domínios Archea e Bacteria (POSTGATE, 1982 [2]; HAYNES, 1986).

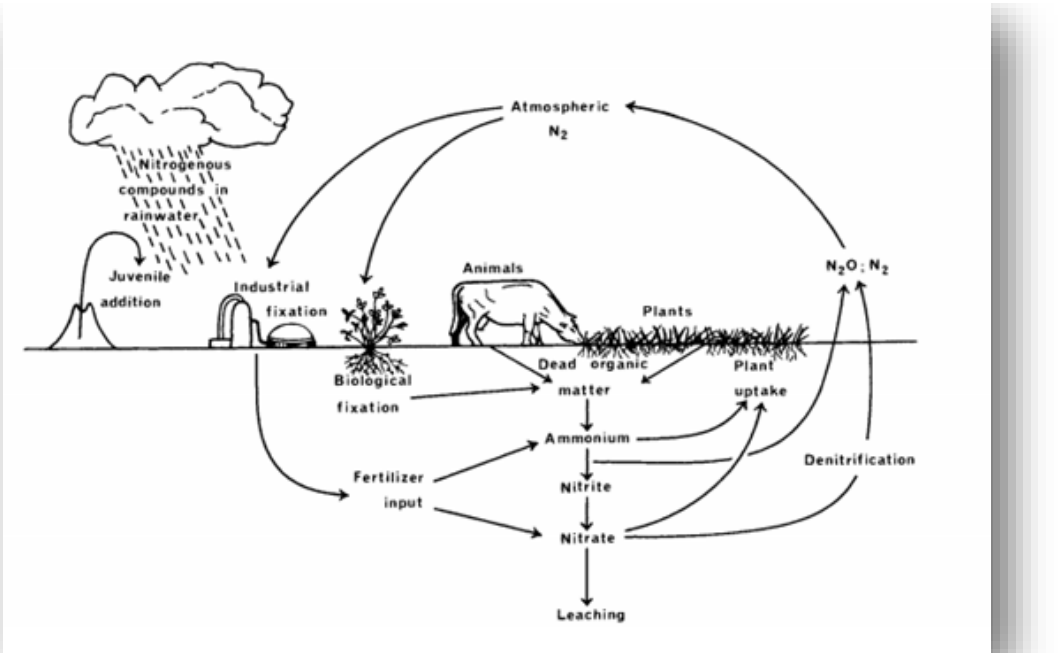


FIGURA 1 – GENERALIZAÇÃO DO CICLO DO NITROGENIO
Fonte: HAYNES (1986)

2.2 IMPORTÂNCIA DA FBN

A FBN tem papel importante numa das principais bases da economia do Brasil:

a agricultura. O nitrogênio é um fator limitante no desenvolvimento de plantas, levando à utilização de fertilizantes químicos que podem alterar o ciclo de nitrogênio, emitir óxidos de nitrogênio na atmosfera, acidificar o solo e poluir rios e lagos. A utilização da FBN, processo natural que faz parte do ciclo do nitrogênio como repositor de nitrogênio na atmosfera, representa uma alternativa barata, limpa e sustentável para a agricultura (PEDROSA, 1987; HUNGRIA *et al*., 2010 [1]).

2.3 *Azospirillum brasilense*

A bactéria *Azospirillum brasilense* (Figura 2) é uma espécie de *Azospirillum*, gênero que naturalmente é encontrado na rizosfera e superfície de raízes de gramíneas de importância agrícola (como milho, trigo, sorgo, cana-de-açúcar e arroz) (ELMERICH e NEWTON, 2007).

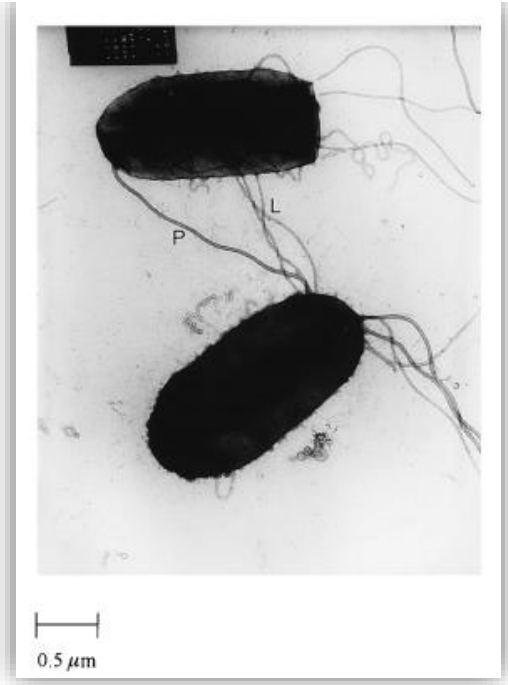


FIGURA 2 – MICROGRAFIA ELETRÔNICA DE TRANSMISSÃO DO TIPO SELVAGEM *A. brasilense* SP7
Fonte: MOENS *et al.* (1995)

As bactérias desse gênero possuem grande potencial como biofertilizantes, influenciando o crescimento da planta, a produtividade da safra e o conteúdo de

nitrogênio do vegetal. Com o uso da *A. brasilense* é possível reduzir 50% do fertilizante nitrogenado mantendo a produtividade, representando no país uma economia de aproximadamente US$ 1,2 bilhões por ano (HUNGRIA *et al.*, 2010) [1].

2.4 CLUSTER *nif*

Nos anos 1980 foi estabelecido o grupo de genes requeridos para a fixação de nitrogênio pela Mo-nitrogenase (genes *nif*), para o organismo *Klebsiella pneumoniae.* Nessa bactéria foram identificados e sequenciados 20 genes *nif*, conforme a Figura 3, numa região de 20 kb do genoma (NEWTON *et al.*, 2007).

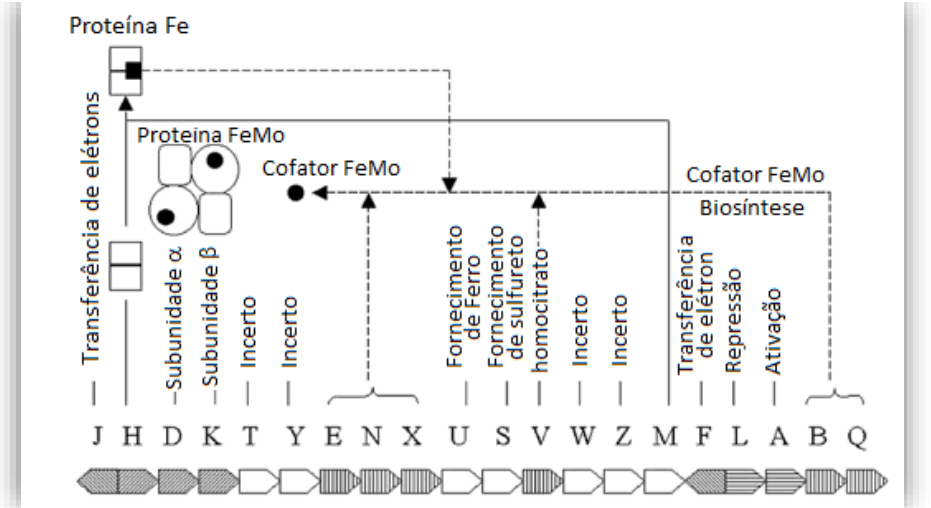


FIGURA 3 – GENES NIF IDENTIFICADOS EM *Klebsiella pneumoniae,* E AS FUNÇÕES DAS SUAS RESPECTIVAS PROTEÍNAS CODIFICADAS. GENES COM MESMO PADRÃO DE LINHAS POSSUEM FUNÇÕES RELACIONADAS.
FONTE: Adaptado de NEWTON *et al.* (2007)

Em 2012, Dos Santos e colaboradores [10] propuseram um novo critério para predição computacional de fixação de nitrogênio em organismos: a presença de 6 genes *nifs*, que codificam para componentes estruturais e biosintéticos: *nifHDK* e *nifENB*. Com esse critério, identificaram 149 potenciais espécies fixadoras, sendo 67 até então não conhecidas como diazotróficas. A determinação dos seis genes *nif* essenciais para a fixação de nitrogênio foi realizada após o estudo da ocorrência desses genes em sistemas caracterizados, na qual foi verificado que quase todos

diazotróficos conhecidos contém esses genes conservados. Apesar de apresentar algumas exceções, a presença desses seis genes *nif* pode ser utilizada em ferramentas *in silico* para a identificação de novos diazotróficos (DOS SANTOS *et al.*, 2012) [10].

2.5 NITROGENASE

A nitrogenase é um complexo enzimático que cataliza o processo de redução do dinitrogênio a amônia, composto por duas proteínas: dinitrogenase redutase (proteína Fe ou NifH) e a dinitrogenase (proteína Fe-Mo ou NifDK). A Figura 4 apresenta a estrutura dessas proteínas. A primeira é produto do gene *nifH* e serve para transportar elétrons à segunda, que por sua vez é produto dos genes *nifDK*. A reação catalisada pela enzima é a seguinte:

$$N_2 + 10H^+ + 8e^- + 16Mg.ATP \dashrightarrow 2NH_4^+ + H_2 + 16Mg.ADP + 16Pi, \quad (1)$$

que representa a redução do dinitrogênio (N_2) a amônio, hidrolizando 16 moléculas de ATP para cada molécula de N_2 fixada (POSTGATE, 1982 [2]; EADY, 1986). Como essa reação possui alto gasto de energia (ATP), esse processo é regulado tanto na atividade da enzima como na expressão dos seus genes.

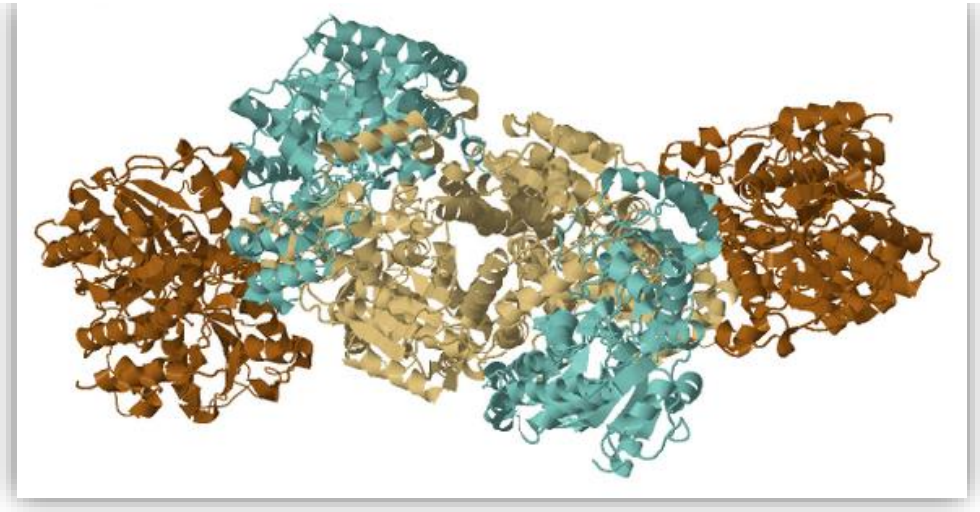


FIGURA 4 – ESTRUTURA TERCIARIA DA NITROGENASE DE *A. VINELANDII*. A MOLECULA FOI CRIADA COM A FERRAMENTA JMOL, UTILIZANDO O ARQUIVO PDB RETIRADO DO RCSB PDB (IDENTIFICAÇÃO DA ESTRUTURA: 4WZB). A ESTRUTURA DE COR MARROM ESCURO REPRESENTA A PROTEINA NifH (OU COMPONENTE II), EM AZUL CLARO A CADEIA ALFA DA PROTEINA FE-MO (NifD), E EM MARRON CLARO NO CENTRO A CADEIA BETA DA PROTEINA FE-MO (NifK).
FONTE: A autora.

2.6 REGULAÇÃO DA NITROGENASE

A regulação da enzima nitrogenase ocorre tanto a nível transcricional quanto a nível pós-traducional. Na regulação transcricional, a síntese da enzima é regulada de acordo com as condições ambientais de oxigênio e amônio (níveis altos reprimem sua transcrição) (DÖBEREINER e PEDROSA, 1987). Já na regulação pós-traducional, é a atividade da nitrogenase que é regulada, sendo inativada com a adição de íons amônio no meio de cultivo, e restaurada quando todo amônio é consumido pela bactéria, num processo conhecido como "*switch-off / switch-on*" (desligamento e religamento da nitrogenase) (ZUMFT e CASTILLO, 1978) [3]. Em *A. brasilense*, a inibição da enzima acontece pela inativação da proteína NifH num processo catalisado pela proteína DraT em resposta à adição de íons amônio ("*switch-off*"), enquanto que a restauração da atividade da NifH ocorre por um processo catalisado pela proteína DraG ("*switch-on*") (HUERGO *et al.*, 2012) [4], como pode ser visto na Figura 5. Esse sistema responde à presença de íons amônio e a mudanças no estado energético da célula (ZHANG *et al.*, 1993).

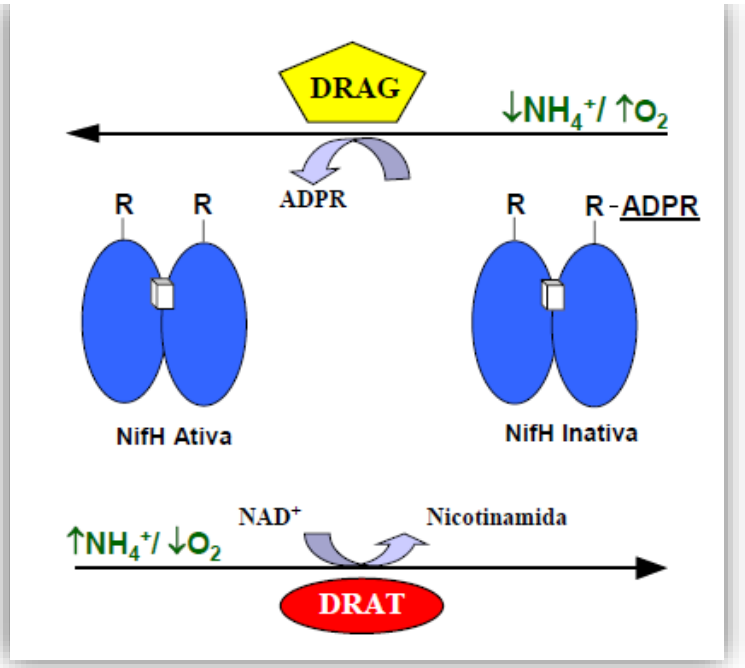


FIGURA 5 – SISTEMA DraT-DraG: INATIVAÇÃO REVERSÍVEL DA PROTEINA NifH POR ADP-RIBOSILAÇÃO (DraT) E SUA REATIVAÇÃO PELA PROTEINA DraG.
FONTE: HUERGO (2006)

2.7 PROTEÍNA DraB

Segundo Zhang *et al.* (1992) [5], em *A. brasilense* as proteínas DraT e DraG são codificadas respectivamente pelos genes *draT* e *draG*, formando um *operon*. A jusante do gene *draG* foi sequenciada uma pequena região cujo primeiros 14 códons formam uma ORF (fase de leitura aberta) similar à ORF presente em *Rhodospirillum rubrum*, conhecida como *draB*. Em *R. rubrum* o produto desse gene parece controlar parcialmente a atividade das enzimas DraT e DraG, visto que mutantes *draB* demonstraram semelhança na regulação da nitrogenase com a estirpe selvagem, mas com desligamento da nitrogenase mais rápido que na estirpe selvagem (LIANG *et al.*, 1991) [6], o que sugere que em *A. brasilense* o mesmo pode ocorrer. De acordo com Zhang *et al.* (2001), além de em *R. Rubrum* e em *A. brasilense*, o gene *draB* também é encontrado a jusante do *draG* em *A. lipoferum*.

Segundo outro artigo publicado por Zhang *et al*. (2001) [7], em um estudo da regulação da atividade da nitrogenase em *Rhodospirillum rubrum*, a proteína

codificada pelo gene *draB* estaria envolvida na regulação da atividade da proteína DraG. A proteína codificada pelo gene *draB* aparentemente é co-transcrita com os genes *draTG* e o gene apresenta pequena similaridade com *nifO* de *A. vinelandii* e alguma similaridade com *arsC* de *E. coli*. Em *A. vinelandii* a proteína NifO pode estar envolvida no metabolismo de molibdénio. Em *E.coli*, o gene *arsC* codifica para enzima arsenate reductase, que catalisa a redução de arseniato em arsenito. Segundo Zhang *et al.* (2011) [7], embora em *R. rubrum* a mutação do *draB* cause a redução da atividade da proteína DraG, na bactéria *Klebsiella pneumoniae* o mesmo não ocorre. Esse resultado foi obtido através de um experimento que verificou que a proteína DraB não está envolvida na regulação da atividade da enzima DraG pelas proteínas PII [1] e GlnK [2] em *K. pneumoniae* (ZHANG *et al.*, 2001) [7].

Embora em *R. rubrum* existem estudos que indicam a função de regulação dessa proteína, uma investigação *in silico* é necessária para uma melhor análise da função da proteína DraB e seu envolvimento na fixação de nitrogênio.

2.8 PROTEÍNA NifO – NITROGENASE-ASSOCIATED PROTEIN

Em *Azotobacter vinelandii*, bactéria fixadora de nitrogênio, a proteína NifO (*nitrogenase-associated protein* ou proteína associada a nitrogenase), codificada pelo gene *nifO* (nome proposto por Quiñones *et al.*, 1993) [8], localiza-se no operon *nifBfdxNnifOQ*.

Através de estudos realizados em laboratório, foi sugerido que a proteína NifO possui papel de regulação da atividade de redução do nitrato (nitrate reductase), visto que mutantes $NifO^-$ não podem fixar nitrogênio sob algumas condições como na presença de baixas concentrações de nitrato (Quiñones *et al.*, 1993 [8]; Gutierrez *et al.*, 1997 [9]).

[1] Proteínas da família PII são proteínas transdutoras dos níveis de nitrogênio intracelular.
[2] Proteína da família PII que controla a atividade inibitória de NifL sobre a proteína NifA em resposta aos níveis de amônio.

2.9 PROTEÍNA ArsC – ARSENATE REDUCTASE

O gene que codifica a proteína ArsC em *Staphylococcus aureus,* plasmídeo pI258, está presente em um operon junto com o *arsR* (que codifica a proteína reguladora do repressor) e *arsB*. A proteína ArsC converte arsenato intracelular para arsenito, tendo sua atividade de redução do arsenato estudada em *Escherichia coli* (JI e SILVER, 1992). Sua filogenia sugere que trata-se de uma enzima evolutivamente antiga (JACKSON *et al*.; 2003).

A família de proteínas ArsC encontra-se nas superfamílias *Thioredoxin_like Superfamily* e LMWPc Superfamily (*Low molecular weight phosphatase Family*), que utiliza glutationa e glutarredoxina ao invés de tioredoxina na redução do arsenato.

2.10 DOMINIOS CONSERVADOS

Os domínios conservados de proteínas permitem a identificação de novos membros de famílias e auxiliam na compreensão do relacionamento entre sequência, estrutura e função. O domínio conservado representa o resultado de um ancestral comum, podendo evidenciar a manutenção de resíduos importantes nos sítios ativos e outras partes funcionais das proteínas. O domínio conservado é utilizado como padrão (*pattern*) para localizar assinaturas de famílias (JONASSEN *et al.*, 1995) [16]. O ExPASy PROSITE do *Swiss Institute of Bioinformatics* (SIB) é um banco de dados de famílias de proteínas e domínios, no qual domínios de proteínas ou proteínas de uma mesma família possuem os mesmos atributos funcionais e são derivadas de um ancestral comum (SIGRIST, 2010) [29].

2.11 CLASSIFICAÇÃO FUNCIONAL DE PROTEINAS

A classificação de proteínas permite a identificação de sua função, que pode ser o transporte de nutrientes, catálise de reações biológicas (como as enzimas), hormônios, entre outras, e com isso caracterizar processos celulares, realizar o

mapeamento em vias metabólicas e compreender o funcionamento do organismo. Para atribuir função a proteínas, existem duas estratégias: a) realização de testes em laboratório, que gera maior confiabilidade, porém demanda mais tempo e recursos, e b) predição usando métodos computacionais (*in silico*), mais adequada ao tratamento de grande volume de dados. A comparação de proteínas é uma operação fundamental desses métodos computacionais, e geralmente é feita através de suas estruturas primárias, ou seja, a sequência de aminoácidos, na qual a função da proteína é inferida pela homologia (LESK, 2008). A sequência de aminoácidos de uma proteína contém informações sobre suas características que permitem a sua classificação em propriedades específicas, como localização, função ou solubilidade. Com isso, houve um aumento no número de ferramentas de bioinformática disponibilizadas para a classificação de proteínas (VAN DER BERG *et al.*, 2014).

Além da similaridade, outros métodos para classificar proteínas são (SIVASHANKARI e SHANMUGHAVEL, 2006):

- Interação proteína-proteína, na qual as funções das proteínas são inferidas com base nos seus parceiros de interação;
- Genômica comparativa, que parte do pressuposto de que as proteínas que funcionam em conjunto numa via metabólica ou num complexo estrutural evoluem juntas;
- Estrutura 3D, na qual são realizadas comparações com base na estrutura da proteína;
- Clusterização, processo de agrupamento dos genes com base em que genes do mesmo conjunto estão envolvidos em funções semelhantes.
- Contexto no genoma, na qual são utilizados métodos pra prever associações funcionais entre proteínas, como interações físicas.
- Outros, como mineração de dados e abordagem Bayesiana.

2.11.1 GO

O *Gene Ontology Consortium* surgiu como uma iniciativa para unificar a representação dos atributos de genes e proteínas de todas as espécies. Através de

um vocabulário dinâmico e controlado que pode ser aplicado a todos genes e seus produtos, possui três ontologias independentes: processos biológicos, função molecular e componente celular. Com uma nomenclatura comum, possibilita a anotação de sequências de genes e proteínas homólogas e a consulta dos mesmos com base em sua biologia compartilhada (ASHBURNER *et al.*, 2000).

2.11.2 Ferramentas para classificação de proteínas

Com base na literatura, algumas ferramentas existentes para predição/ classificação de proteínas são:

1. SPiCE, ferramenta *web* que realiza as três etapas de classificação: extração de características, treinamento e predição da classe, fornecendo gráficos e permitindo parametrização. Para a utilização, é necessário possuir a lista de proteínas e suas respectivas classes (VAN DER BERG *et al.*, 2014).
2. ConFunc, uma abordagem de predição de função de proteínas automática e baseada no GO que utiliza resíduos conservados para gerar perfis de proteínas utilizados para inferir função (WASS, 2008) [24].
3. GOtcha, que utiliza a estrutura GO para combinar os *e-values* dos resultados do BLAST para fazer predições para termos GO individuais, cada um associado com um escore (MARTIN *et al.*, 2004) [30].
4. PFP, abordagem similar ao GOtcha, que realiza predições com base na frequência dos termos GO juntamente com retorno do PSI Blast (HAWKINS *et al.,* 2006) [31].
5. Blast2GO, ferramenta de bioinformática ampla para anotação funcional das sequências de genes e proteínas e de mineração de dados sobre as anotações resultantes, tendo como base o vocabulário GO. Através de um algoritmo elaborado de similaridade, otimiza a transferência de função de sequências homólogas. Possui integração com InterPro, códigos de enzimas (CE), as vias KEGG, grafos acíclicos diretos (DAGs) e GOSlim, sendo uma ferramenta adequada para a investigação genômica devido à sua versatilidade e uso amigável (CONESA e GOTZ, 2008) [25].

6. InterPro, que classifica proteínas em famílias e realiza predição de domínios e sítios importantes, unificando as assinaturas de proteínas de diversos bancos de dados produzindo um poderoso banco de dados integrado. Fazem parte do InterPro Consortium as bases de dados PROSITE, HAMAP, Pfam, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, CATH-Gene3D e PANTHER (THE INTERPRO CONSORTIUM, 2015) [26].
7. PANTHER (*Protein ANalysis THrough Evolutionary Relationships*), sistema de classificação de proteínas (e seus genes) em famílias e subfamílias, função molecular, processo biológico e *pathway* (caminho e relacionamentos entre as interações moleculares). A classificação é feita por algoritmos de bioinformática em dados curados e utiliza HMMs (THOMAS *et al*., 2003) [27].
8. Pfam, coleção de famílias de proteínas, cada uma representada por alinhamentos de sequências e modelos HMMs. A identificação de domínios de proteínas pode fornecer informações da sua função. No banco de dados Pfam-A estão as famílias com curadoria manual, enquanto que no Pfam-B, encontram-se as famílias de menor qualidade, também úteis para identificar regiões funcionalmente conservadas quando não há entradas Pfam-A (FINN *et al.*, 2014) [28].
9. PROSITE, base de dados de famílias de proteínas e seus domínios, na qual proteínas ou domínios proteicos pertencentes a uma família particular podem partilham atributos funcionais, e geralmente são derivados a partir de um antepassado comum. Contém padrões e perfis específicos para mais de mil famílias de proteínas ou domínios, cada uma destas assinaturas com documentação sobre a estrutura e função destas proteínas (SIGRIST, 2010) [29].

2.11.3 Geração da estrutura terciária da proteína

Um arquivo PDB contém dados de coordenadas 3D, informações químicas e descritores qualitativos sobre a estrutura macromolecular, obtidas por métodos como

difração de raios-X. O formato PDB, criado nos anos 1970 como uma representação padrão para a estrutura de moléculas, permite a geração de sua estrutura tridimensional (BERMAN *et al.*, 2007). O banco de dados Protein Data Bank – PDB contém informações curadas sobre as estruturas 3D de proteínas, ácidos nucleicos, e montagens complexas, criando recursos para pesquisa nas áreas biologia molecular, biologia estrutural, biologia computacional, entre outras (RCSB PDB, 2015).

Com os arquivos PDB, as imagens da estrutura 3D das proteínas podem ser alinhadas e geradas em ferramentas como a PyMOL, sistema de visualização molecular que permite a visualização de imagens biomoleculares em 3D. Com mais de 600 configurações e 20 representações, por exemplo *ribbon* e *cartoon*, usadas para identificar estruturas secundárias, a ferramenta PyMOL recebe arquivos PDB como entrada e gera a estrutura terciária da proteína (PyMOL, 2015) [33].

O Expasy SwissModel é um sistema de modelagem de proteínas a partir da sequência de aminoácidos que utiliza homologia para gerar arquivos PDB de proteínas sem correspondentes no banco de dados PDB. Com o SwissModel, a modelagem de proteínas é acessível a todos bioquímicos e biólogos no mundo, sendo a modelagem das estruturas terciaria e quaternária da proteína obtidas através de informação evolutiva, tendo como parâmetro a sua estrutura primária (BIASINI *et al.*, 2014) [32].

2.12 BIOINFORMÁTICA

A Bioinformática, juntamente com a Biologia Computacional, visa resolver problemas biológicos por meio de técnicas computacionais, envolvendo aplicações como alinhamento de sequência, projeto genoma, reconhecimento de padrões e construção de árvores filogenéticas. Com os bancos de dados biológicos públicos, é possível recuperar informações de descrição e classificação de espécies, sequência genética de organismos, comentários gerais e literatura associada (ATTWOOD, 2011).

2.13 MINERAÇÃO DE DADOS

De acordo com Fayyad *et al.* (1996), a mineração de dados, ou *Data Mining* (DM) do inglês, é uma técnica de Informática que tem por objetivo extrair conhecimento (padrão ou característica) de grandes quantidades de dados. Com o aumento de registros de dados como DNA, RNA e proteínas nos bancos de dados biológicos, a mineração de dados é fundamental na busca de novas informações a partir de um grande acervo de dados biológicos. Exemplos de aplicação da mineração de dados são os algoritmos de recuperação de informações, de aprendizagem de dados, redes neurais artificiais e algoritmos genéticos (FAYYAD *et al.*, 1996).

2.13.1 Reconhecimento de padrões

O objetivo principal do Reconhecimento de Padrões (RP) é a classificação de padrões. O RP é a aplicação de técnicas de raciocínio indutivo em cima de um conjunto de dados visando a criação de um modelo que classifique dados não conhecidos, ou seja, o RP é a ciência que descreve ou classifica medidas. A abordagem geralmente utilizada é dividir o problema na extração de características e no módulo de classificação propriamente dito (KASABOV, 1996).

De acordo com Kasabov (1996), alguns exemplos da utilização de RP:

- Análise, segmentação e pré-processamento de imagens;
- Reconhecimento de faces;
- Identificação de impressões digitais;
- Reconhecimento de caracteres;
- Análise de manuscritos;
- Visão computacional;
- Entendimento e reconhecimento de voz;
- Diagnóstico médico;
- Sinais biológicos.

Há técnicas de RP que envolvem desde estatística e matemática até a

Inteligência Artificial (IA), por meio de processo supervisionado de aprendizagem de máquina (como as redes neuronais e conjuntos difusos) (KASABOV, 1996).

2.13.2 Extração de características

Característica é uma medição útil extraída no processo de identificação de padrão, podendo ser simbólica, numérica ou ambas, e ainda, ser variável ou discreta. A extração de características, processo utilizado em aprendizagem de máquina, tem grande relevância no reconhecimento de padrões, evidenciando a necessidade na bioinformática das técnicas de seleção de características (KASABOV, 1996).

A extração das características é a simplificação do problema, ou o resumo, para que possa ser tratado computacionalmente. Após a extração das características, no modelo supervisionado ocorre o aprendizado do sistema, através do treinamento com um conjunto de padrões já classificado. A seguir é realizado o teste, para verificar a capacidade de classificação do sistema, através da taxa de acerto obtida (RAITTZ, 1998).

2.13.3 Redes Neurais Artificiais e MLP

A neurocomputação é área da ciência que lida com métodos e sistemas de processamento de informações usando redes neurais (ou neuronais) artificiais. Uma rede neural artificial é um modelo computacional inspirado no neurônio biológico que consiste no processamento e conexões de neurônios artificiais. As conexões possuem coeficientes (pesos) vinculados, que são a memória do sistema, e constituem a estrutura neuronal e os algoritmos de treino e recordação atrelado com a estrutura (KASABOV, 1996).

A MLP, ou Perceptrons de múltiplas camadas, é uma rede de múltiplas camadas fortemente conectada com conexões *feedfoward*, na qual os neurônios de uma camada estimulam todos os neurônios da camada seguinte. A estrutura é formada por uma camada de entrada, com os sinais de entrada, uma camada de saída, que gera a resposta da rede ao estímulo, e camadas escondidas ou

intermediárias (KASABOV, 1996).

2.14 COEFICIENTE DE CORRELAÇÃO DE PEARSON

O Coeficiente de Correlação de Pearson é uma medida de correlação, originada por Galton ao final do século XIX, que possui o alcance de +1, que significa a correlação perfeita entre duas variáveis aleatórias, -1, que indica a correlação perfeita porém negativa, e 0 indicando a ausência de relação entre as variáveis, ou seja, as duas não dependem linearmente uma da outra (ADLER e PARMRYD, 2010) [22].

2.15 EXPRESSÃO REGULAR

Expressão regular é um método formal de se especificar um padrão de texto, formado por símbolos e caracteres com funções especiais. Uma vez agrupados entre si e com caracteres literais, formam uma sequência que atua como uma regra que indica sucesso se a entrada de dados obedecer exatamente as suas condições (JARGAS, 2005). Segundo Jargas (2015), embora as expressões regulares tenham surgido em 1943, foi em 1968 que foram utilizadas computacionalmente, em um algoritmo de busca do editor de textos dos sistemas Unix (dando origem ao aplicativo grep - "Global Regular Expression Print”).

3 ARTIGO CIENTÍFICO

Esta seção apresenta o artigo científico a ser submetido à BMC *Bioinformatics* como um artigo de metodologia, ou seja, um artigo que apresenta um novo método, exame ou procedimento experimental ou computacional, neste caso, a metodologia de classificação desenvolvida.

A BMC *Bioinformatics* (ISSN: 1471-2105) é um periódico científico de acesso aberto que publica pesquisas na área de Bioinformática e Biologia Computacional, envolvendo todos aspectos do desenvolvimento, testes, novas aplicações computacionais e métodos estatísticos para a modelagem e análise de dados biológicos. Fundada em 2000, faz parte da série de jornais BMC - BioMed Central, publicados no Reino Unido (BMC Bioinformatics, 2015).

O periódico foi selecionado para a submissão do artigo porque, além de apresentar fator de impacto 2,67, e ser indexado em 13 bancos de dados de bibliografia, incluindo o PubMed, possui o escopo mais alinhado com o trabalho desenvolvido. Além disso, os autores mantém os direitos autorais de seus artigos, podendo reproduzir e divulgar seu trabalho [3].

[3] Conforme http://www.biomedcentral.com/about/copyright

ProClaT, a new bioinformatics approach for *in silico* protein reclassification: case study of DraB, a putative arsenate reductase protein of *Azospirillum brasilense*

Elisa T Rubel[1]*, Roberto T Raittz[1]*, Nilson AR Coimbra[1], Michelly AC Gehlen[1], Fabio O Pedrosa[2§]

[1] Laboratory of Bioinformatics, Professional and Technological Education Sector, Federal University of Paraná, Curitiba, PR, Brazil, Rua Dr. Alcides Vieira Arcoverde 1225, Curitiba, Paraná, Brazil.

[2] Department of Biochemistry and Molecular Biology, Federal University of Paraná, Curitiba, PR, Brazil, Av. Cel. Francisco H. dos Santos, s/n, Curitiba, Paraná, Brazil.

* These authors contributed equally to this work

§ Corresponding author

Email addresses:

ETR: terumi@ufpr.br

RTT: raittz@gmail.com

FOP: fpedrosa@ufpr.br

BACKGROUND

Azospirillum brasilense is a diazotrophic organism used as commercial inoculants, since it promotes plant growth [1]. As a nitrogen-fixing bacterium, *A. brasilense* has a specific metabolic pathway for the conversion of gaseous dinitrogen into ammonia. The N_2 is fixed under limiting conditions of NH_4^+ and O_2, through the activity of nitrogenase [2]. A post-translational control of nitrogenase occurs via the DraG-DraT system, in which the DraT enzyme (dinitrogenase reductase ADP-ribosyltransferase) acts in the nitrogenase shutdown by inactivating the NifH (dinitrogenase reductase) in response to the presence of ammonium ions in the environment, while the DraG enzyme (dinitrogenase reductase activating-glycohydrolase) restore the activity of NifH, after ammonium ions consumption [3] [4]. The DraT and DraG enzymes are encoded by the *draTG* genes, forming an operon in *A. brasilense* [5]. Participates in this operon the *draB* gene [5], which its product is annotated as a putative arsenate reductase [GenBank:CCC97498]. No specific function for the *draB* gene product of *Azospirillum brasilense* has been ascribed to date, although in *Rhodospirillum rubrum*, this protein seems to regulate the activity of DraG [6]. The *draB* gene has some similarity to *nifO* of *A. vinelandii* and *arsC* of *E. coli* [7]. The *A. vinelandii* nitrogenase-associated NifO protein, part of operon *nifBfdxNnifOQ*, has a role in regulating the activity of nitrate reductase, whereas mutants $NifO^-$ can not fix nitrogen in the presence of low concentrations of nitrate [8] [9].

To test the hypothesis that the *draB* gene codes for the NifO protein, since DraB protein has no known homologous in Gene Ontology database, we developed a methodology for the classification of hypothetical or little known proteins, named ProClaT - Protein Classifier Tool. Therefore, we could classify the *draB* gene product and identify homologous and related genes. We used a neural network as a machine learning tool, protein sequences to compose the classifier features and the protein conserved pattern to consolidate and label classes. Beyond the *draB* product classification, we also analysed the relationship and co-occurrence of *draB* with other genes related to nitrogen fixation, the minimum *nif* gene set, *nifHDKENB* [10], and with the *draT* and *draG* genes.

METHODS

We present ProClaT, a new machine learning approach to classify proteins based on protein sequence features and conserved domains. ProClaT was used to classify the draB product and to find NifO-like proteins.

Data

ProClaT was applied to 2,773 complete bacterial genomes obtained from the NCBI database [11] via FTP, containing 5,182 GenBank data downloaded in July 2014. The download file size was 78.1 GB.

ProClaT Pattern Recognition sequence-based features

The features used by the pattern recognition model are divided into three categories:

1) Amino acid composition

The relative occurrence of each amino acid residue and its number in each functional group (polar with charge positive, polar with charge negative, apolar and hydrophobic) was calculated by dividing the number of the occurrence by the protein sequence length. The protein sequence length as also used to composed the features.

2) Consensus region alignment scores

The consensus region, whose determination will be described, was used for measuring alignment score of each protein sequence. A self-alignment function and the global and local alignments score sequences using the Needleman-Wunsch algorithm (identity and positive scores) were also used as features.

3) Protein physic-chemical properties

The protein physic-chemical features were the isoelectric point (pI), charge, nominal mass, aromaticity, instability, hydropathy, entropy and energy.

Isoelectric point: The estimated pI for an amino acid sequence were calculated with Matlab and the Bioinformatics Toolbox ™, using the pK values described on http://www.mathworks.com/help/bioinfo/ref/isoelectric.html.

Charge: The estimated charge of the protein sequence for a given pH (default is typical intracellular pH 7.2), calculated by the same Matlab function of the Bioinformatics Toolbox ™ of the pI described above.

Nominal mass: The expected protein nominal mass was calculated by a Matlab function of the Bioinformatics Toolbox ™ by the analyzes of a peptide sequence (http://www.mathworks.com/help/bioinfo/ref/isotopicdist.html).

Aromaticity: The aromaticity value of a protein was calculated according to Lobry [12] (the relative frequency of Phe+Trp+Tyr).

Instability: The protein instability index was calculated according to Guruprasad *et al* [13], used to test a protein for stability (which any value above 40 means the protein is unstable, or has a short half life).

Hydropathy or GRAVY (Grand Average of Hydropathy) Index: The protein GRAVY was calculated according to Kyte and Doolittle methodology [14]. This index indicates the solubility of a protein, where a positive GRAVY value corresponds to a hydrophobic protein and a negative GRAVY value corresponds to a hydrophilic protein. The GRAVY value of a peptide/protein is calculated by adding the values of hydropathy of each amino acid, divided by the number of residues found in sequence.

Entropy and Energy: In this context, the descriptors Energy and Entropy represent, respectively, the degree of uniformity and disorder of each protein sequences. Co-occurrence matrices 3x3 were generated from amino acids based on the sequences, and for each entry, the sequence was read from the right to the left and stored in a 3x3 amino acids arrangement. Based on this list, the combinations in pairs were analyzed one by one, and in case of co-occurrence, the count and recording of data was updated. This calculation were based on the Haralick methodology [15] called "matrix of co-occurrence", developed for the description of textures images

based on second-order statistics.

The Aromaticity, Instability and Hydropathy were calculated using the package Biopython. The features extraction is part of the tool. Table 1 shows the summary of the three feature categories, including the number of features generated and the functions used to extract them.

Table 1 – Features of the ProClaT Pattern Recognition model

Feature Category	**Number of features**	**Function (matlab or python)**
AA composition *		
AA composition *	20	aacount (sequence)
AA functional property *	5	codon2aa (sequence)
Protein length	1	length (sequence)
Scores alignment with consensus region		
Self align with consensus region	1	selfalign (sequence,CSeq)
Global alignment score with consensus region	2	getIdentity (sequence,CSeq,'G')
Local alignment score with consensus region	2	getIdentity (sequence,CSeq,'L')
Protein physico-chemical properties		
pI	1	isoelectric (sequence)
Charge	1	isoelectric (sequence)
Nominal mass	1	isotopicdist (sequence)
Aromaticity	1	ProtParam.ProteinAnalysis (seq).aromaticity() (python)
Instability	1	ProtParam.ProteinAnalysis (seq).instability_index() (python)
Hydropathy	1	ProtParam.ProteinAnalysis (seq).gravy() (python)
Entropy	1	function developed in python
Energy	1	function developed in python

*AA: amino acid

ProClaT algorithm

ProClaT development algorithm flow can be seen in Figure 1.

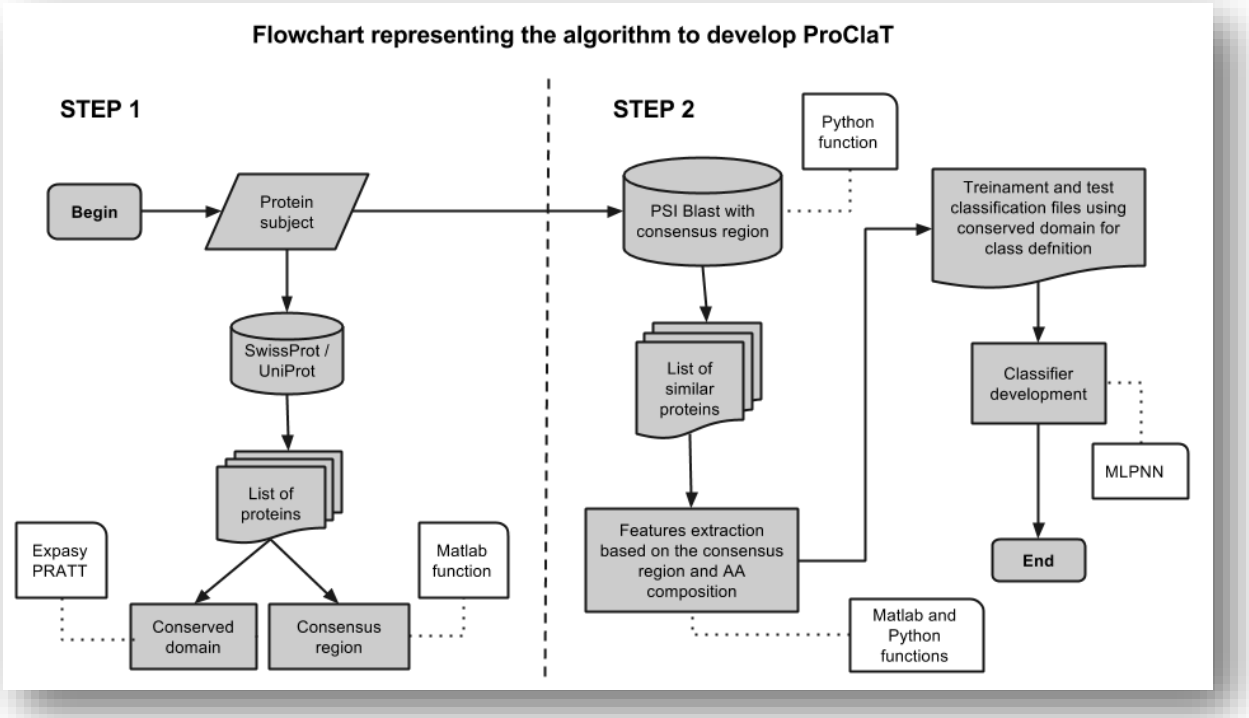


Figure 1 - Flowchart representing the algorithm to develop ProClaT. In the first step, the domain conserved protein and the consensus region are generated. In the second step, a search is performed in NCBI NR with the generated region consensus as query. With the list of similar proteins, the features are extracted and the classifier is trained and tested.

The protein conserved domain and consensus region were determined using the curated sequences protein deposited in the SwissProt database. Since there are no reviewed NifO proteins in the SwissProt database, the NifO proteins deposited in the Uniprot database were used. To generate the protein conserved domain, we used the Expasy PRATT tool [16]. This conserved domain may be a common ancestor consequence with the evolutionary pressure to maintain important residue in the active site and other relatively important parts of the protein and are useful to identify new family members [16]. The conserved NifO domain generated by PRATT (P-X-L-I-R-R-P-L-[ILM]) originates a regular expression, shown in Figure 2.

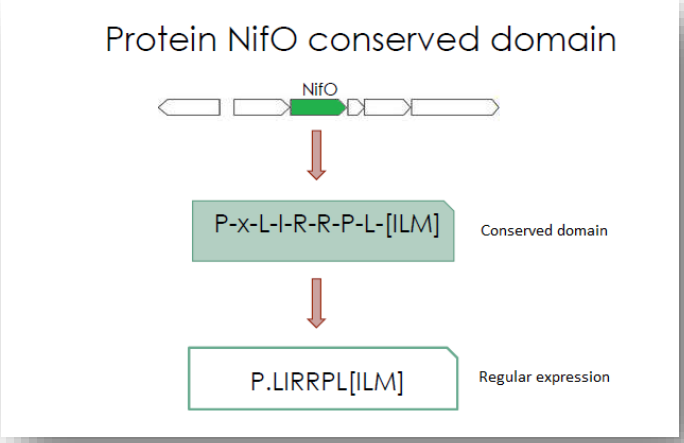


Figure 2 - Conserved domain of NifO-like proteins generated with Expasy PRATT tool after the refinement phase, and the regular expression correspondent. Considering that the number of coded amino acids residues in proteins is 20, the probability of this amino acids sequence occur randomically is $1.1719*10^{-10}$.

The consensus region (Figure 3) was used as a query in a PSI-Blast search in the NR NCBI proteins library, returning 5,000 hits of similar proteins using the Blast default values. The regular expression allowed the identification of proteins among the 5,000 that have the conserved domain. These proteins were submitted to the feature extractor and were used to create the classifier treinament and test files, as the Label 1 class ("TRUE to NifO"). To compose the Label 0 class ("FALSE to NifO"), were used the proteins with the lowest similarity levels that does not have the conserved domain.

Figure 3 - NifO-like consensus region, generated from the multiple alignment of the NifO proteins.

ProClaT was parameterized in order to classify the NifHDK, NifENB, DraT and DraG proteins. Instead of a single TRUE/FALSE classifier, its returns 1 for NifH, 2 for

NifD, 3 for NifK, 4 for NifE, 5 for NifN and 6 for NifB. For DraT and DraG, it returns 1 and 2 respectively.

ProClaT only ranks candidate proteins, with at least 0.2 of identity calculated by a self-alignment function. This function returns the average of the global alignment of two sequences using the Needleman-Wunsch algorithm:

$$selfalign = \frac{\frac{globalAlign(seq1,seq2)}{globalAlign(seq1,seq1)} + \frac{globalAlign(seq1,seq2)}{globalAlign(seq2,seq2)}}{2} \quad (1)$$

Implementation

As shown in Table 2, ProClaT was developed in the programming language Matlab ®, which also worked as Integrated Development Environment (IDE), using the Bioinformatics Toolbox ™. Some feature extractions were performed in Python using the Biopython package [17].

Table 2 - Version of softwares

Software	Version	Application
Matlab	r2012B (8.0.0.783)	Functions to get the conserved domain, features extraction and create the classifier.
Python	3.4.2	Functions to performe PSI-Blast and features extraction.
Expasy PRATT	2.1	Generate the protein conserved domains
Weka	3.6.12	Test of the classifiers algorithms

The ProClaT algorithm for supervised classification chosen was the Multilayer Perceptron Neural Network (MLPNN), a feed-forward back-propagation machine learning method [18]. MLPNN returned the best results, according to the Weka data mining software [19], as shown in Table 3. In this case, the implementation without the cross-validation technique showed better results. For the algorithm selection, were considered the best algorithms according to the Top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) presented in December 2006 in Hong Kong [20].

Table 3 – Correctly classified proteins by Weka algorithms

Algorithm	Options	Correctly Classified Instances without cross-validation	Correctly Classified Instances with cross-validation
Multilayer Perceptron	-L 0.3 –M 0.2 –N 500 –V 0 –S 0 –E 20 –H a	99.61%	99.41%
Simple Cart	-S 1 –M 2.0 –N 5 –C 1.0	99.09%	99.22%
Nnge	-G 5 –I 5	99.09%	99.02%
J48	-C 0.25 –M 2	98.96%	98.71%
Ada BoostM1	-P 100 –S 1 –I 0 –W weka.classifiers.trees. DecisionStump	32.51%	33.35%
Naive Bayes	-	99.22%	98.90%

Using the default parameters proposed by Weka, the neural network training and test files were submitted to the six algorithms above. MLPNN showed the best number of correctly classified proteins.

For the *nifO* neighbourhood analysis, we identified the *nifO* neighboring genes in a five window genes upstream and downstream using ProClaT.

RESULTS AND DISCUSSION

ProClaT was used to classify the *draB* gene product into NifO-like protein and to find homologous proteins. Additionally, confirming its general applicability, it was applied in the classification of NifHDK, NifENB, DraT and DraG, allowing the analysis of co-occurrence of these proteins with NifO in completed bacterial genomes.

Using ProClaT, the *Azospirillum brasilense* DraB protein was classified as a NifO-like protein. With the classifier, we found 82 NifO-like proteins belonging to 76 complete bacterial genomes, representing 56 bacterial species. Figure 4 shows the annotation of the 82 NifO-like proteins revealed using ProClaT.

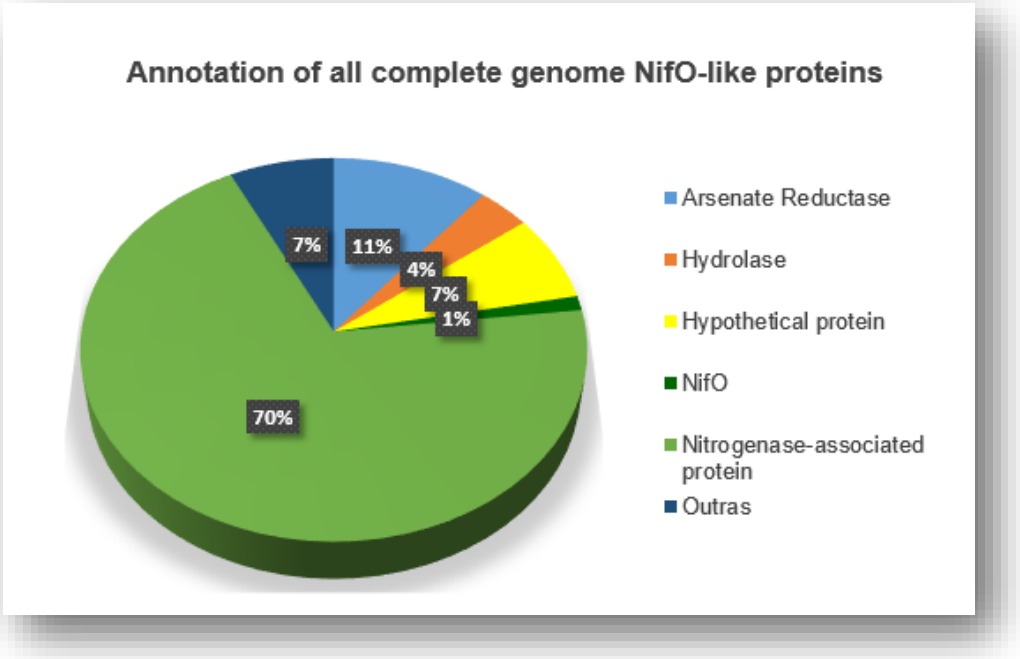


Figure 4 – Annotation of all complete genome NifO-like proteins. Of the 82 proteins classified as NifO-like with ProClaT, most corresponds to nitrogenase-associated protein, NifO. The proteins annotated as arsenate reductase, hypothetical and others, totaling 21 proteins, might be re-classified as NifO-like, also.

The product of the *PST1305* gene of *Pseudomonas stutzeri* A1501, classified as NifO-like with ProClaT, was suggested to participate in biological nitrogen fixation, probably involved in electron transport or in an oxygen protection mechanism for nitrogenase [21]. Fan *et al* [21] considered this gene product to be required for optimal nitrogenase activity of *Pseudomonas stutzeri* A1501.

Moreover, the *A. vinelandii* NifO protein was also classified as NifO-like, as expected. Laboratory tests suggests that this protein has a role on ammonium repression of the nitrite-nitrate (*nasAB*) assimilatory operon of *Azotobacter vinelandii*. Considering that the *nifO* gene is present in the molybdenum (Mo) metabolism operon in *A.vinelandii*, and that nitrogenase and nitrate reductase contain Mo cofactors, NifO may be involved in regulating the conversion of Mo into the nitrogenase FeMoco, rendering Mo inaccessible to the synthesis of the nitrate reductase cofactor [9].

The Additional file 1 lists all bacteria species containing at least five essencial *nif* genes, and the presence of *nifHDK*, *nifENB*, *nifO*, *draT* and *draG* genes, according to ProClaT. Of the 80 bacterial species (or 119 strains) that have the six essential *nif*

genes, 42 (or 61 strains) have also the *nifO,* including *Acidithiobacillus ferrivorans*, *Bradyrhizobium japonicum*, *Burkholderia xenovorans*, *Magnetospirillum magneticum*, *Pseudomonas stutzeri* and *Rhodospirillum rubrum. H*owever, 41 bacterial species (or 58 strains) have no *nifO*-like genes, incluing *Herbaspirillum seropedicae*, *Klebsiella oxytoca, Enterobacter sp* and *Burkholderia phenoliruptrix*.

All the genes coding for NifO-like identified belong to bacteria having at least three of the essential *nif* genes. Figure 5 shows the number of genes coding for NifO-like in the presence of some gene coding for a essential Nif protein in complete genomes, analyzing the bacterial species.

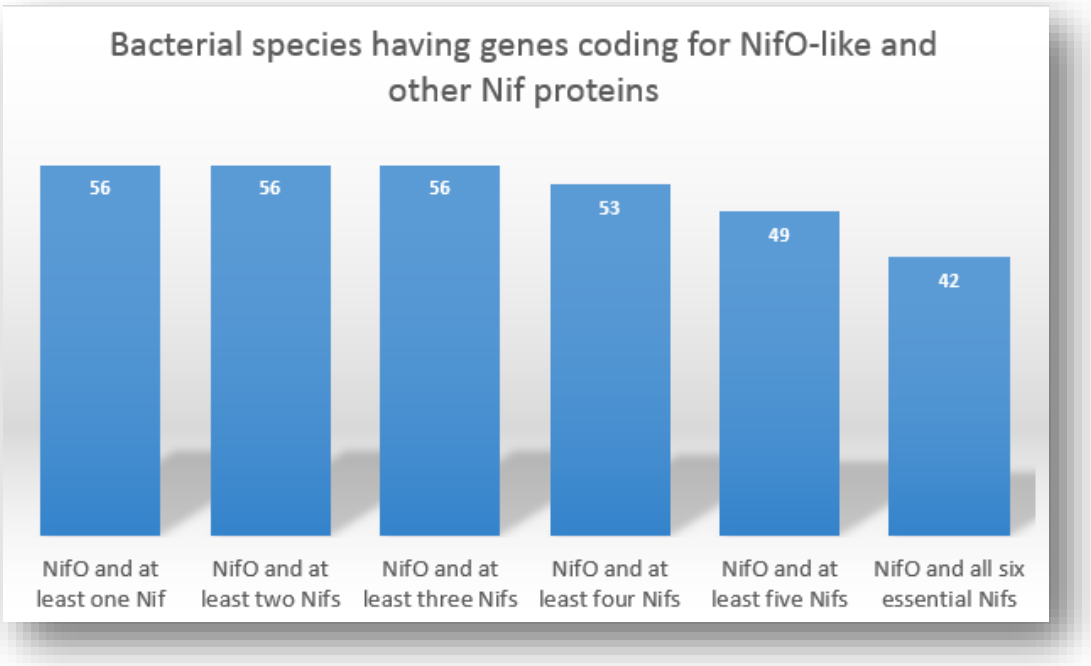


Figure 5 – Bacterial species containg gene coding for NifO-like and to some Nif proteins. ProClaT identified 56 bacterial species containing genes coding for nifO- like. All belong to a genome that contain at least three genes coding for a essential Nif protein. 53 species contain at least 4 *nif* genes, 49 contain at least 5 *nif* genes and 42 contain all the 6 essential *nif* genes.

Figure 6 shows the number of gene groups found in the complete genome with ProClaT, analyzing the bacterial species.

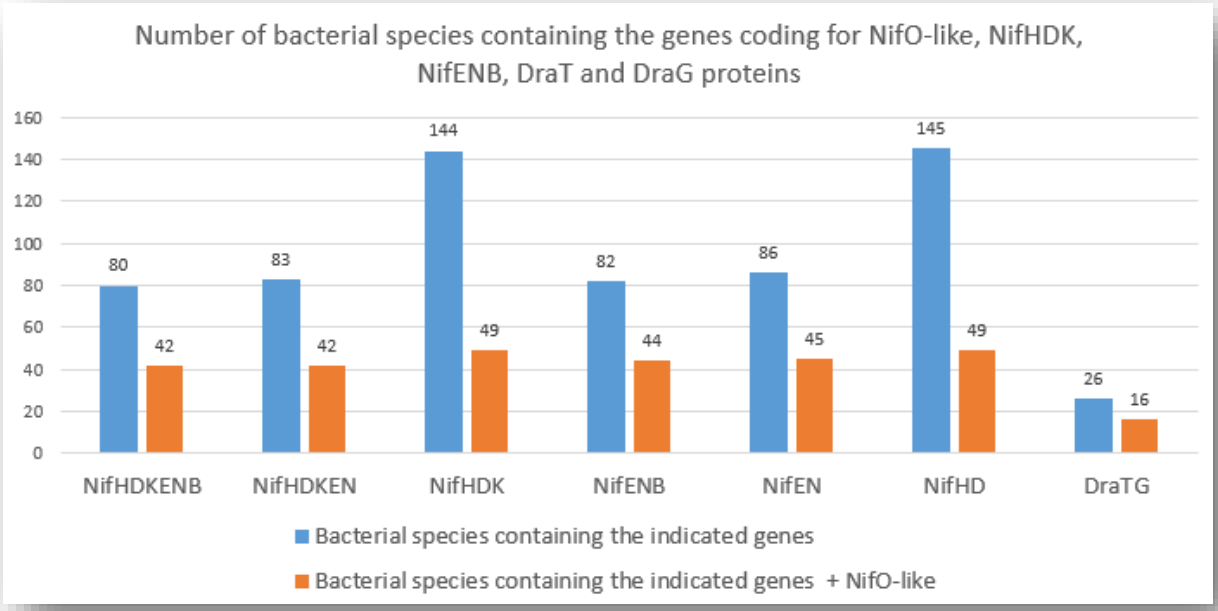


Figure 6 – Bacterial species containing gene groups with the presence of *nifO*. In blue, the number of species of bacterial complete genomes containing the genes indicated below, and in red, the number of the species containing these genes in addition with the gene coding for NifO- like.

In Additional file 2 are the same information but analyzing the bacterial strains. Interestingly, the species *Azospirillum sp.*, *Azospirillum brasilense*, *Azospirillum lipoferum* and *Azotobacter vinelandii* have two genes coding for NifO-like protein, in the same genome, according to ProClaT. Moreover, no genes coding for NifO-like were found in plasmids.

The co-occurrence of the genes coding for NifO-like, NifHDK-like, NifENB-like, DraT-like and DraG-like proteins in the genomes was determined, using the Pearson Correlation Coefficient. Figure 7 shows this correlation for the complete bacterial genomes analysed.

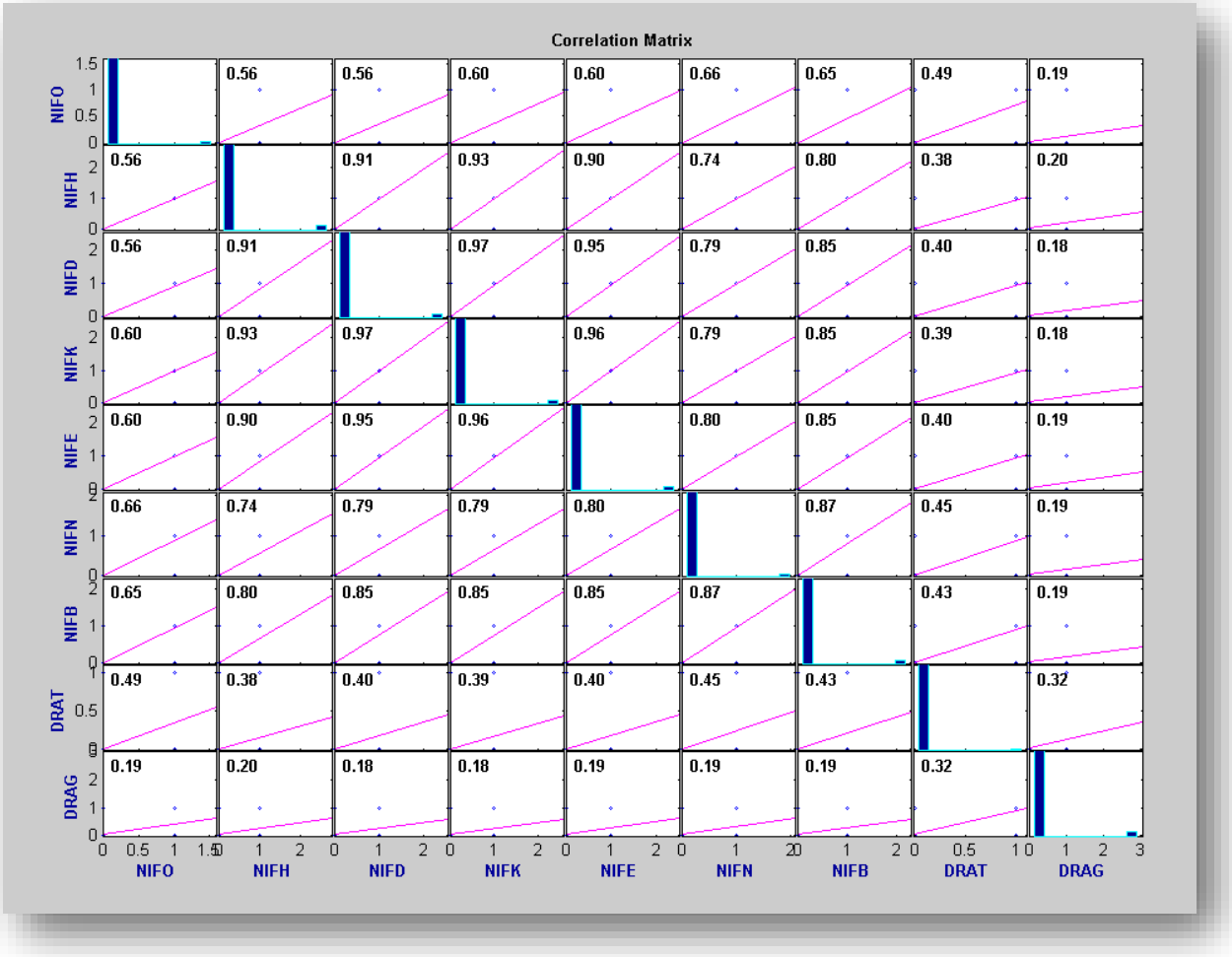


Figure 7 - Pearson Correlation Coefficient of the co-occurrence of *nifO*, *nifH, nifD, nifK, nifE, nifN, nifB* and *draT* and *draG* genes in complete bacterial genomes. The Pearson Correlation Coefficient is a well-established measure of correlation whose has range of +1 (perfect correlation) to -1 (perfect but negative correlation), in which 0 is the absence of a relationship [22]. The highest *p-value* found was $6.7*10^{-39}$, indicating that all pairs of variables have correlation significantly different from zero. Image generated by Matlab.

There is a likely co-occurrence correlation of *nifO* and other *nif* genes. *NifO* gene also has correlation with the *draT* and *draG*, though smaller. Among the *nifs,* the correlation is obviously high. The low correlation of *draB* and *nif* genes with *draG* is explained by the fact of DraG-like proteins are necessarily not related to nitrogen fixation, since they are found in organisms unable to fix nitrogen such as *Escherichia coli* ([GenBank:WP_001634525]) and *Salmonella enterica* ([GenBank:WP_010835]). The Pearson correlation coefficient of *nifO* co-occurrence with all the six *nif* genes is 0.6350, and with the presence of both *draT* and *draG* genes is 0.4544 (as can be seen in Additional file 2, item 2.3).

The analysis of neighbourhood genes, in a five window genes upstream and downstream, showed that *nifO* is regularly located close to at least one *nif* gene, as weel as with *draT* or *draG*. Table 4 shows the number of the *nif* genes present in the *nifO* neighbourhood.

Table 4 – Genes present in the *nifO* neighborhood

Gene	Absolute number of occurrences of the genes in the *nifO* neighborhood
nifH	24
nifD	12
nifK	8
nifE	1
nifN	0
nifB	19
draT	11
draG	13

The number of genes present in the *nifO* neighborhood, in a five window gene upstream and downstream.

ProClaT comparison and validation

Table 5 compares the NifO-like proteins predicted by ProClaT with those predicted by cut-off score, conserved domain and both cut-off score and conserved domain.

Table 5 – Sensitivity and specificity of protein prediction methods

Method	**TP** [1]	**TN** [2]	**FP** [3]	**FN** [4]	**Calculated sensitivity (%)**	**Calculated specificity (%)**
1. Cut-off score (> 30% local identity and > 50% positive)	229	2704	0	67	**77.36**	**100**
2. Conserved domain	231	2704	10	55	**80.77**	**99.63**
3. Conserved domain with cutoff score	219	2704	0	77	**73.99**	**100**
4. ProClaT	289	2704	0	7	**97.64**	**100**

1 TP: True Positive
2 TN: True Negative
3 FP: False Positive
4 FN: False Negative

A PSI-Blast was performed on the NCBI NR proteins library, using the consensus region of NifO as input query. It returned 3,000 hits of similar proteins, which 296 are NifO-like, after curation. All these proteins were submitted to the above methods. ProClaT showed the best sensitivity.

We also tested ProClaT sucess rate applying it on a list of known Nif proteins, obtained from the SwissProt database. Table 6 shows its hit rate.

Table 6 - ProClaT hit rate

Protein	Class	Quantity of curated proteins	ProClaT Hits	Calculated hit rate
NifH	1	92	91	98.91 %
NifD	2	23	22	95.65 %
NifK	3	17	16	94.12 %
NifE	4	14	14	100 %
NifN	5	10	10	100 %
NifB	6	13	12	92.31 %

A search was performed in the SwissProt protein database by the proteins name NifHDK and NifENB, cured manually. Each found protein was applied to ProClaT, and the hit rate was calculated. The average of sucess rate was 96.83%.

DraB classification with existents prediction proteins tools

Additional file 3 contains in detail DraB prediction results using some existing Bioinformatics tools.

Since *A. brasilense* DraB protein has no homologous in GO database, as BLAST performed with AmiGO web tool [23], the functional classification services based on GO terms were not specific or reliable. ConFunc [24] predicted for the DraB protein terms GO: 0008794 (ontology: molecular function, description: arsenate reductase glutaredoxin activity) with probability 0.667, and GO: 0006351 (ontology: biological process, description: transcription, DNA- templated) with probability 0.306. With Blast2GO [25], the terms suggested to the DraB protein are GO: 0055114 (ontology: biological process, description: oxidation-reduction process) and GO: 0016491 (ontology: molecular function, description: oxidoreductase activity), although

based on only two homologous proteins. Other tools not based on the term GO predicted that DraB can be part of the families Arsenate reductase-like (InterPro[26] and PANTHER[27]), Thioredoxin-like fold (InterPro[26], Pfam[28] and PROSITE[29]) and Nitrogenase-associated protein (InterPro[26]). The protein prediction methods based on its tertiary structure are either not recommended in this case, since there are no models of tertiary structure of DraB / NifO homologous obtained via experiments laboratory in protein structure databases.

CONCLUSIONS

This *in silico* study suggests that *draB* gene codes for a NifO-like protein. There is evidence that NifO is involved in nitrogen fixation, probably involved in electron transport or in an oxygen protection mechanism for nitrogenase [21]. We found four bacterial species containing two *nifO*-like genes, and all the *nifO* genes in chromosomes.

All the genes coding for NifO-like found with ProClaT belong to a bacteria having at least three of the six essential *nif* genes, *nifHDK* and *nifENB* [10]. With a correlation analysis of co-occurrence of these genes in complete bacterial genomes, we observed that the *nifO/draB* gene has a higher correlation coeficient with the essential *nif* genes than with *draT* and *draG*.

Analysis of neighbourhood revealed that *nifO* may have both *nif* and *draT* and *draG* genes as neighbours, but no clear pattern was identified.

Of the 80 bacterial species containg the six essential nif genes, 42 contain also the *nifO*, however, 41 bacterial species have no nifO-like genes, which suggests that *nifO* is not necessarily essential for the nitrogen fixation by nitrogenase.

ProClaT found nine genes annotated as arsenate reductase, six as hypotheticals and six with variable names in complete bacterial genomes. This suggests that these gene product should be reclassified as NifO-like.

This tool was developed to this case, since the existentes methods and tools for protein prediction were not useful for the *A. brasilense* DraB protein. The prediction using only conserved domain and identity/similarity (cut-off) also were not the most

appropriate, as seen in the comparative table between the methods.

ProClaT was tested with curated Nif proteins and showed average hit rate of 96.83% in classifying known Nif proteins, confirming that it can be useful in the (re)classification of other proteins.

AVAILABILITY AND REQUIREMENTS

Project name: ProClaT.
Project home page: http://www.bioinfo.ufpr.br/proclat/.
Operating system(s): Platform independente.
Programming language: Matlab (R2012b) and Python 3.4.
Other requirements: MathWorks Bioinformatics Toolbox ™ and Biopython.
License: GNU GPL v3.
The data sets supporting the results of this article are available in the ProClaT SourceForge repository.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

FP and RTT proposed the concept, validated the results and revised the manuscript. The methodology, implementation and results achievement was developed by ETR and RTT, under the supervision of FP. NARC and MACG provided technical assistance and developed some functions. All authors contributed to and approved the manuscript.

ACKNOWLEDGEMENTS

We thank R.A. Vialle, C.E. Brim, V. Weiss for technical assistance. We thank the Graduate Program in Bioinformatics of Federal University of Paraná and the National Science and Technology Institutes of Biological Nitrogen Fixation (INCT).

REFERENCES

1. Hungria M, Campo RJ, Souza EM, Pedrosa FO. Inoculation with selected strains of Azospirillum brasilense and *A. lipoferum* improves yields of maize and wheat in Brazil. Plant Soil. 2010; Vol 331.

2. Postgate JF. The fundamentals of nitrogen fixation. Cambridge Univ. Press. 1982; p. 252.

3. Zumft WG, Castillo, F. Regulatory properties of the nitrogenase from *Rhodopseudomonas palustris*. Arch Microbiol. 1978; vol 117, p 53-60.

4. Huergo LF, Pedrosa FO, Muller-Santos M, Chubatsu LS, Monteiro RA, Merrick M, Souza EM. PII signal transduction proteins: pivotal players in post-translational control of nitrogenase activity. Microbiology. 2012; p. 176-190.

5. Zhang Y, Burris RH, Roberts GP. Cloning, sequencing, mutagenesis, and functional characterization of *draT* and *draG* genes from *Azospirillum brasilense*. Journal of Bacteriology. 1992; p. 3364–3369.

6. Liang J, Nielsen GM, Lies DP, Burris RH, Roberts GP, Ludden PW. Mutations in the *draT* and *draG* Genes of *Rhodospirillum rubrum* Result in Loss of Regulation of Nitrogenase by Reversible ADP-Ribosylation. Journal of Bacteriology. 1991; Vol 173, p. 6903 – 6909.

7. Zhang Y, Pohlmann EL, Halbleib CM, Ludden PW, Roberts GP. Effect of PII and Its Homolog GlnK on Reversible ADP-Ribosylation of Dinitrogenase Reductase by Heterologous Expression of the Rhodospirillum rubrum Dinitrogenase Reductase ADP-Ribosyl Transferase-Dinitrogenase Reductase-Activating Glycohydrolase Regulatory System in *Klebsiella pneumonia*. Journal of Bacteriology. 2001; p. 1610–

1620.

8. Quiñones FR, Bosh R, Imperial J. Expression of the *nifBfdxNnifOQ* Region of *Azotobacter vinelandii* and Its Role in Nitrogenase Activity. Journal of Bacteriology. 1993; p. 2926–2935.

9. Gutierrez JC, Santero E, Tortolero M. Ammonium repression of the nitrite-nitrate (*nasAB*) assimilatory operon of *Azotobacter vinelandii* is enhanced in mutants expressing the *nifO* gene at high levels. Mol Gen Genet. 1997; p. 172-179.

10. Dos Santos PC, Fang Z, Mason SW, Setubal JC, Dixon R. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. BMC Genomics. 2012;13:162.

11. NCBI GenBank FTP. ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Bacteria (2015). Accessed 19 Apr 2015.

12. Lobry JR, Gautier C. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Res. 1994; 3174-80.

13. Guruprasad K, Reddy BV, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng. 1990; 155-61.

14. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol.1982; 105-132.

15. Haralick RM. Statistical and structural approaches to texture. Proc IEEE. 1979; 67:786-804.

16. Jonassen I, Collins JF, Higgins DG. Finding flexible patterns in unaligned protein sequences. Protein Science. 1995; 4:1587-1595.

17. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–1423. doi: 10.1093/bioinformatics/btp163

18. Jain AK, Duin RPW, Mao J: Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 2000, 22(1):4–37.

19. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsl. 2009; 11:10–18.

doi: 10.1145/1656274.1656278.

20. Wu X, Kumar V, Quinlan JR, Ghosh J, Motoda QYH, Mclachlan GJ, Ng A, Liu B, Yu PS, Zhou Z, Steinbach M, Hand DJ, Steinberg D. Top 10 algorithms in data mining. Knowl Inf Syst. 2008;14:1–37. DOI 10.1007/s10115-007-0114-2

21. Fan H, Yan Y, Li Y, Ping S, Zhang W, Chen M, Lin M, Lu W. Analysis of a new nitrogen fixation gene in Pseudomonas stutzeri A1501. Acta Microbiologica Sinica. 2009; p. 580-584.

22. Adler J, Parmryd I. Quantifying Colocalization by Correlation: The Pearson Correlation Coefficient is Superior to the Mander's Overlap Coefficient. Wiley InterScience, 2010.

23. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, AmiGO Hub, Web Presence Working Group. AmiGO: online access to ontology and annotation data. Bioinformatics. Jan 2009;25(2):288-9.

24. Wass M N, Sternberg J E. ConFunc - functional annotation in the twilight zone. Bioinformatics. 2008; 798-806.

25. Conesa A, Gotz S. Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. International Journal of Plant Genomics. 2008; 619832.

26. The InterPro Consortium. InterPro: An integrated documentation resource for protein families, domains and functional sites. Brief Bioinform. 2002; 225-35.

27. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. Genome Res. 2003; 2129–2141.

28. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hethweington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. The Pfam protein families database. Nucleic Acids Research. 2014; 42:D222-D230.

29. Sigrist CJA, Cerutti L, De Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 2010; 161-6 , 2010.

30. Martin DM, Berriman M, Barton GJ. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics. 2004; 5:178.

31. Hawkins T, Luban S, Kihara D. Enhanced automated function prediction using

distantly related sequences and contextual association by PFP. Protein Science. 2006; 15:1550–1556.

32. Biasini N, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucl. Acids Res. 2014; 42 (W1): W252-W258.

33. PyMOL. A molecular visualization system. http://www.pymol.org (2015). Accessed 05 Apr 2015.

4 CONSIDERAÇÕES FINAIS

Com o trabalho, foi desenvolvida uma metodologia para classificação de proteínas denominada ProClaT – *Protein Classifer Tool*, com o objetivo de classificar a proteína codificada pelo gene *draB* e buscar proteínas homólogas, uma vez que os métodos e ferramentas disponíveis não atenderam ao objetivo. Como a proteína DraB não possui homólogos no banco de dados do GO, conforme o BLAST realizado com a ferramenta AmiGO, os serviços de classificação funcional baseados no GO não se mostraram eficientes ou confiáveis. ConFunc previu para proteína DraB os termos GO:0008794 (ontologia: função molecular), cuja descrição é *Arsenate reductase (glutaredoxin) activity*, com 0.667 de probabilidade, e GO:0006351 (ontologia: processo biológico), cuja descrição é *transcription, DNA-templated*, com 0.306 de probabilidade. Com Blast2GO, os termos sugeridos para a proteína DraB são GO:0055114 (ontologia: processo biológico), cuja descrição é processo de oxidação-redução e GO:0016491 (ontologia: função molecular), cuja descrição é atividade de *oxireductase*, porém baseado na anotação de apenas duas proteínas. A ferramenta SPiCE não foi utilizada neste estudo de caso, uma vez que necessita dos grupos de proteínas já consolidados (proteínas rotuladas). Outras ferramentas não baseadas no termo GO predisseram que DraB pode encontrar-se nas famílias *Arsenate reductase-like* (InterPro e PANTHER), *Thioredoxin-like fold* (InterPro, Pfam e PROSITE) e *Nitrogenase-associated protein* (InterPro). Os métodos de predição de proteínas baseados na sua estrutura terciária tampouco são recomendados neste caso, pois até a presente data não há registros de modelos da estrutura terciária de proteínas homólogas a DraB/NifO obtidos via experimentos em laboratório nos bancos de dados de estrutura de proteínas. As predições utilizando apenas domínio conservado e/ou nota de corte também apresentaram falhas, como visto na tabela comparativa entre os métodos. ProClaT obteve a melhor sensibilidade.

ProClaT é uma abordagem de aprendizagem de máquina que utiliza redes neuronais (no caso, a MLP) cujas características do modelo de reconhecimento de padões são provenientes da sequência primária das proteínas (composição de aminoácidos, escore de alinhamento com região consenso e propriedades físico-

químicas das proteínas). ProClaT fez a predição do produto do *draB* de *A. brasilense* como NifO-like. A análise da correlação de co-ocorrência permitiu verificar que o gene *draB/nifO* possui relação com os genes *nif*, *draT* e *draG.* Com a análise da vizinhança do gene *draB* e homólogos, numa janela de cinco genes *upstream* e *downstream*, verificamos que o gene *draB/nifO* possui com certa frequência em sua vizinhança genes *nif*, *draT* e *draG*, embora não tenha sido possível perceber algum padrão de *operon*. Embora todos os genes *draB/nifO* encontrados pertencem a uma espécie bactéria que contém pelo menos 3 genes *nif*, nem todas bactérias fixadoras de nitrogênio necessariamente possuem o gene *draB/nifO*, como a *H. seropedicae*.

A aplicação de ProClaT na base de dados NCBI NR sugere a reclassificação de 143 proteínas hipotéticas, 42 arsenate reductase (ArsC) e três hidrolases em nitrogenase associated protein NifO-*like*, totalizando 188 proteínas. Analisando os genomas completos de bactérias, nove proteínas arsenate reductase (ArsC), seis hipotéticas, três hidrolases e seis proteínas com outros nomes poderiam ser reclassificadas para NifO-*like*, totalizando 24 proteínas. Dos genes enquadrados como NifO-*like* com ProClaT, além do NifO de *A. vinelandii*, outro que possui estudos que sugerem seu envolvimento na fixação de nitrogênio é o gene *PST1305* de *Pseudomonas stutzeri* A1501 (FAN *et al.*, 2009) [21]. Seu produto, estudado em 2009, sugere que o gene está envolvido no transporte de elétrons ou no mecanismo de proteção de oxigênio da nitrogenase, considerado necessário para uma atividade otimizada da nitrogenase de *Pseudomonas stutzeri* A1501.

O domínio conservado gerado pela ferramenta Expasy PRATT atuou como um ponto de partida eficaz para selecionar proteínas para treinar o classificador. As propriedades das proteínas utilizadas foram satisfatórias como características do modelo de reconhecimento de padrões.

Embora a ferramenta ProClaT tenha sido projetada e construída visando este estudo de caso específico, sua arquitetura genérica permite a sua utilização em outras situações. Como exemplo, foram criados classificadores para as proteínas Nif essenciais e para as proteínas DraG e DraT, usando a mesma metodologia. Esta metodologia de classificação de proteínas se demonstrou útil neste estudo de caso e pode ser aplicado a outras proteínas.

5 DOCUMENTAÇÃO

5.1 LOGO

A logo de ProClaT – *Protein Classifier Tool* foi elaborada pela autora juntamente com uma equipe especializada em identidade visual de produtos. Como pode ser visto na Figura 6, a logo representa uma rede neural para classificação de informações biológicas, no caso, proteínas.

FIGURA 6 – LOGO DA FERRAMENTA PROCLAT.
FONTE: A autora

5.2 PACOTES

A arquitetura geral do sistema pode ser dividida em três pacotes, conforme Figura 6.

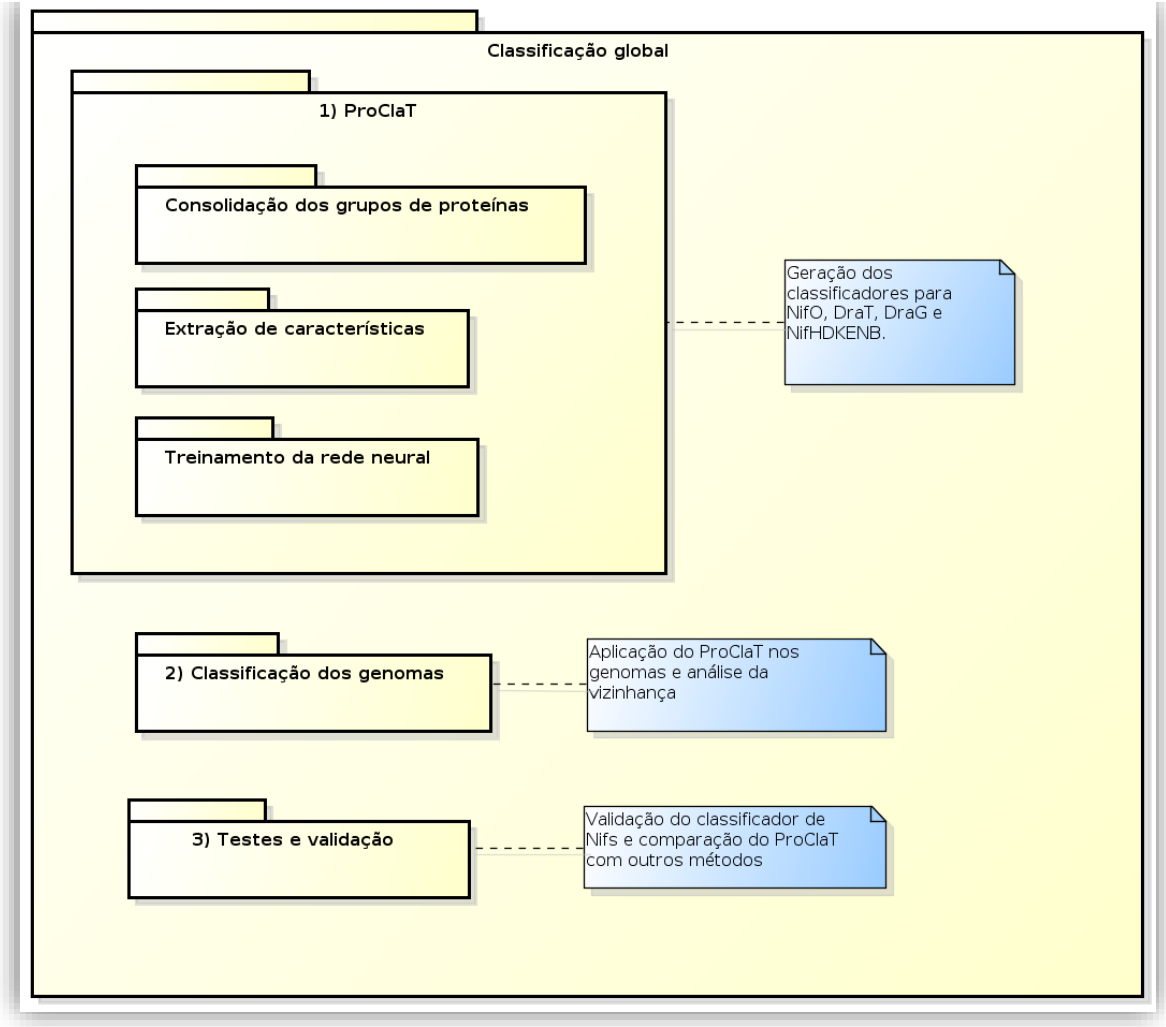


FIGURA 7 - DIAGRAMA DE PACOTES CONTENDO OS MÓDULOS DA CLASSIFICAÇÃO GLOBAL.
FONTE: A autora

5.3 FUNCIONALIDADES

As funcionalidades gerais do sistema podem ser vistas no Diagrama de Casos de Uso (Figura 7). Cada caso de uso corresponde a uma funcionalidade, disponível ao usuário que pode ser um bioinformata, biólogo, bioquímico ou pesquisador/cientista. O caso de uso “Gerar Classificador” refere-se a geração do ProClaT, que aciona os casos “Buscar Registros”, “Gerar Região Consenso”, “Gerar Expressão Regular” e “Extrair caracteristicas”. “Classificar Proteína” descreve a aplicação do ProClaT na classificação de uma proteína enviada pelo usuário (*query*).

O caso de uso “Classificar Genomas” permite a classificação de proteínas de genomas completos, e permite a análise de vizinhança, a geração de correlação de co-ocorrência e gravação dos dados para posterior análise.

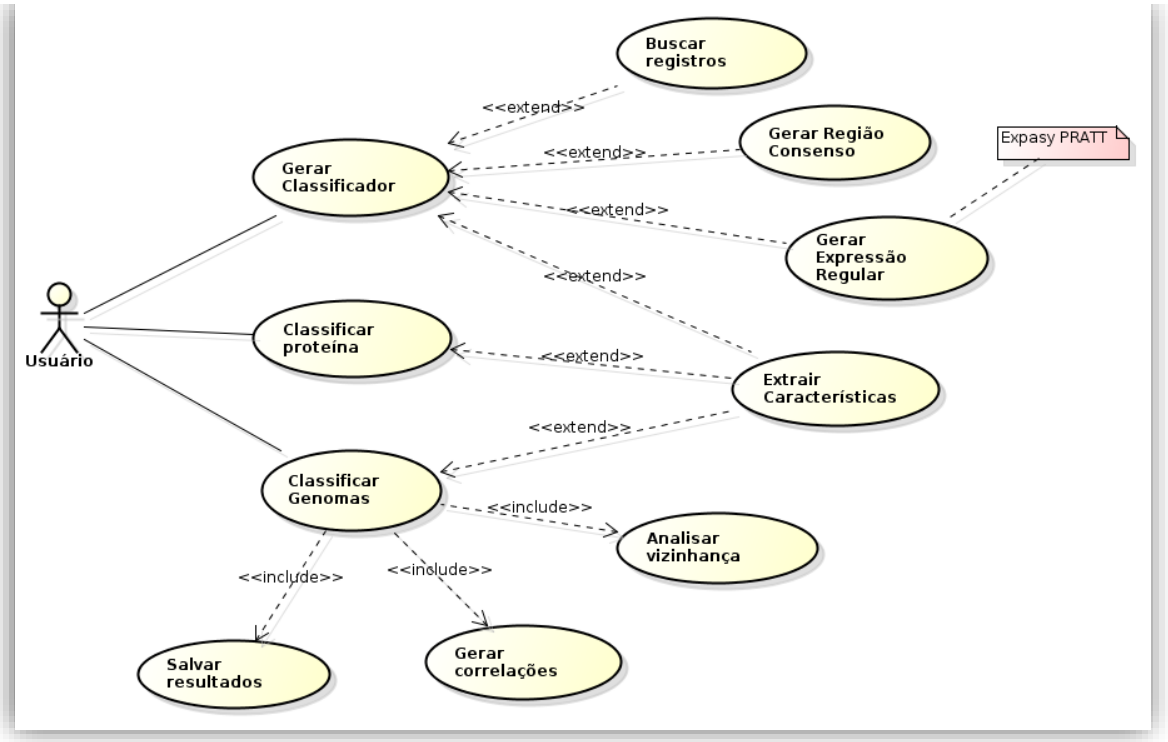


FIGURA 8 - DIAGRAMA DE PACOTES CONTENDO AS FUNCIONALIDADES DESENVOLVIDAS.
FONTE: A autora

5.4 DISPONIBILIZAÇÃO E REQUISITOS

Os scripts para geração e aplicação do classificador ProClaT estão disponíveis no endereço: https://sourceforge.net/projects/proclat/.

Para criar e executar ProClaT, são necessários os seguintes *softwares*:

- Matlab e Bioinformatics Toolbox;
- Python e pacote Biopython [4];
- Navegador (sem restrição) e acesso à internet.

[4] Obter do endereço: http://biopython.org/wiki/Download

5.5 PASSOS PARA UTILIZAÇÃO

Após o *download* dos scripts, o diretório ProClaT/Matlab deve ser adicionado ao PATH do Matlab, conforme Figura 8.

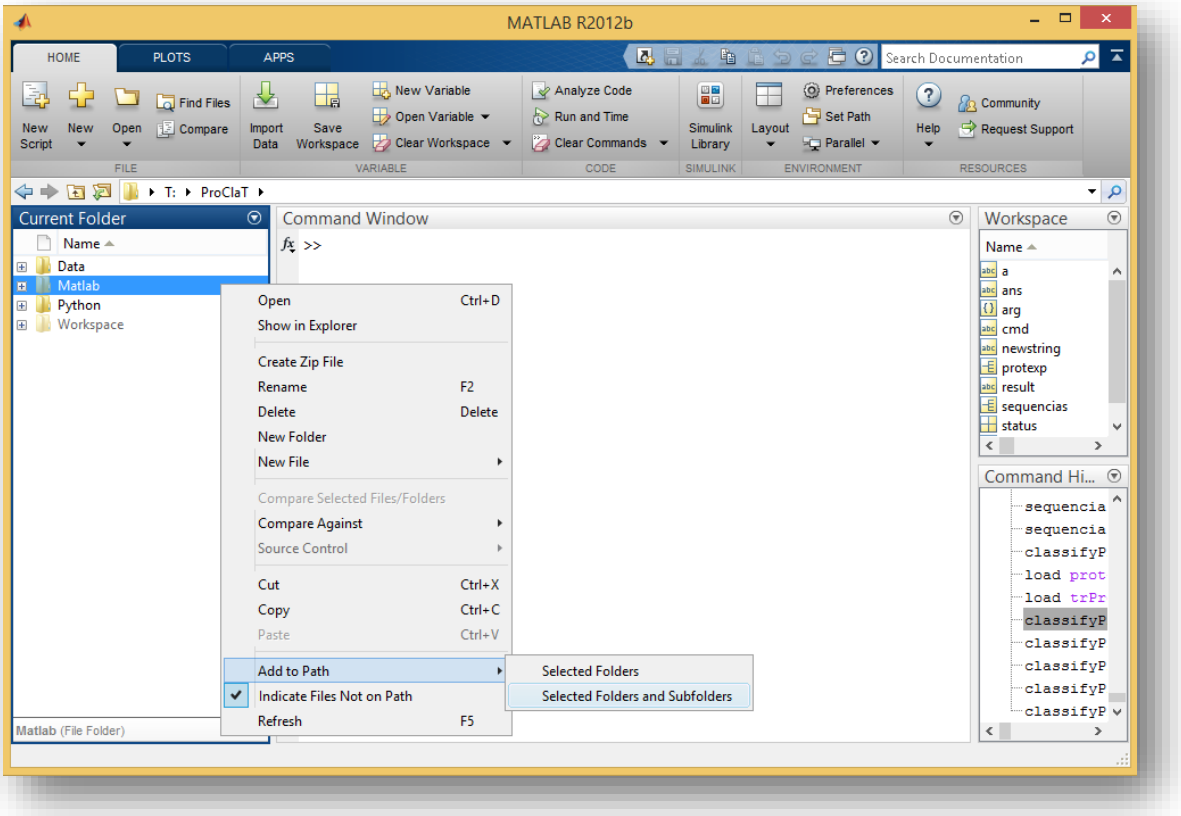


FIGURA 9 – ADIÇÃO DA PASTA CONTENDO OS *SCRIPTS* DO PROCLAT NO *PATH* DO MATLAB.
FONTE: A autora

A seguir, executar as funções na seguinte ordem:

>> ***createProClaT(1,0,5000,'nifO')*** *% a TRUE/FALSE classifier*

% esta função irá criar o classificador, de acordo com os parâmetros informados.

% Neste exemplo, está sendo criado o classificador para NifO, conforme o utilizado no trabalho. Os parâmetros de entrada são:

% número de classes, neste caso, apenas 1 (NifO)

% indicativo se sistema deve procurar proteínas curadas (1) ou não curadas

(0) no SwissProt
% numero de registros retornados do Blast, para compor os grupos consolidados
% nome da(s) proteína(s), neste caso, apenas NifO.
SequenciasSwissProt.fasta file created. Please generate the conserved domain via PRATT (http://web.expasy.org/pratt/) and inform:
>> **'P-x-L-I-R-R-P-L-[ILM]'** % informar domínio conservado gerado no PRATT
% nome da(s) proteína(s), neste caso, apenas NifO.
Performing search homologous proteins nifO through blast, returning 5000 hits. This option can be time consuming.
Blasting with python
Generating consolidated group of the protein nifO
Creating the classifier...
Classifier successfuly created! It will return 1 to TRUE to nifO and 0 to FALSE (not nifO).

O classificador ProClaT foi gerado com sucesso. Para aplicar o classificador, primeiro é necessário carregar os classificador e dados da proteínas, que o sistema salvou. Acesse o diretório ProClaT/Data e execute os comandos:

>> ***load protexp*** % rede neural gerada
>> ***load trProt*** % arquivo com informações da(s) proteína(s), como domínio conservado e região consenso

O diretório ProclaT/Workspace é pasta de trabalho onde o usuário pode trabalhar com suas sequencias. Acesse esse diretório e execute os comandos abaixo:

>> **sequencias = fastaread('sequencias.fasta');** % onde sequencias.fasta é o arquivo contendo as sequências a serem classificadas
>> **classifyProtein(sequencias, protexp, trProt)** % executando o classificador. Retorna um vetor com o resultado da classificação.

Classifing protein 1

Classifing protein 2

Classifing protein 3

Classifing protein 4

Classifing protein 5

mret =

1 1 0 0 0

% neste exemplo, as duas primeiras proteínas foram classificadas com TRUE, e as três ultimas, como FALSE

REFERÊNCIAS COMPLEMENTARES

ASHBURNER, M.; BALL, C. A.; BLAKSE, J. A.; BOTSTEIN, D.; BUTLER, H.; CHERRY, M.; DAVIS, A. P.; DOLINSKI, K.; DWIGHT, S. S.; EPPIG, J. T.; HARRIS, M. A.; HILL, D. P.; ISSEL-TARVER, L.; KASARSKIS, A.; LEWIS, S.; MATESE, J. C.; RICHARDSON, J. E.; RINGWALD, M.; RUBIN, G. M.; SHERLOCK, G. Gene Ontology: tool for the unification of biology. **Nature Genetics.** 25, 25 – 29, 2000.

ATTWOOD, T. K.; GISEL, A.; ERIKSSON, N-E; BONGCAM-RUDLOFF, E. Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective. **Bioinformatics – Trends and Methodologies**. 2011.

BERMAN, H.; HENRICK, K.; NAKAMURA, H.; MARKLEY, J. L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. **Nucleic Acids Res**. 35, D301–D303. 2007.

BMC Bioinformatics. Disponível em: <http://www.biomedcentral.com/bmcbioinformatics>. Acesso em: 18/04/2015.

DÖBEREINER, J.; PEDROSA, F. O. Nitrogen-fixing bacteria in nonleguminous crop plant. **Science Tech Publishers**. Springer-Verlag, 1987.

EADY, R. R. Enzimology in free-living diazotrophs. In: BROUGHTON, W. J.; PUHLER,S. (Ed). **Nitrogen Fixation**, v. 4, p. 1-49, 1986.

ELMERICH, C.; NEWTON, W. E. **Associative and Endophytic Nitrogen-fixing Bacteria and Cyanobacterial Associations**. Springer, 2007.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **American Association for Artificial Intelligence**. 1996.

HALBLEIB, C. M.; ZHANG, Y.; ROBERTS, G. P.; LUDDEN, P. W. Effects of Perturbations of the Nitrogenase Electron Transfer Chain on Reversible ADP-Ribosylation of Nitrogenase Fe Protein in *Klebsiella pneumoniae* Strains Bearing the *Rhodospirillum rubrum dra* Operon. **Journal of Bacteriology**. Vol 182, p. 3681–3687, 2000.

HAYNES, R. J. **MINERAL NITROGEN IN THE PLANT-SOIL SYSTEM**. Chapter 1 - Origin, Distribution, and Cycling of Nitrogen in Terrestrial Ecosystems. University of Wisconsin. 1986.

HUERGO, L. F. Regulação do metabolismo de nitrogênio em *Azospirillum brasilense*. Curitiba, 2006. 170 f. Teste (Doutorado em Bioquímica) - Setor de Ciências Biológicas, Universidade Federal do Paraná.

JACKSON, C. R.; DUGAS, S. L. Phylogenetic analysis of bacterial and archaeal arsC gene sequences suggests an ancient, common origin for arsenate reductase. **BMC Evol Biol**. 2003; 3: 18.

JARGAS, A.M. **Expressões Regulares: Guia de Consulta Rápida**. Disponível em: <http://guia-er.sourceforge.net/index.html>. Último acesso: 21/05/2015.

KASABOV, N. K. **Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering**. Massachusetts Institute of Technology. USA, 1996.

LESTK, A. M. **Introdução à Bioinformática**. Porto Alegre: Artmed, 2008.

MOENS, S.; MICHIELS, K.; KEIJERS, V.; VAN LEUVEN, F.; VANDERLEYDEN, J. Cloning, Sequencing, and Phenotypic Analysis of laf1, Encoding the Flagellin of the Lateral Flagella of *Azospirillum brasilense* Sp7. **Journal Of Bacteriology**, Vol. 177, p. 5419 -5426, 1995.

NEWTON, W.E.; FERGUNSON, S.J.; BOTHE, H. **Biology of the Nitrogen Cycle**. Cap. 8 – Physiology, Biochemistry, and Molecular Biology of Nitrogen Fixation. Elsevier, Amsterdam, 2007.

PEDROSA, F. O. Fixação biológica de nitrogênio: fértil idéia. **Ciência Hoje**, v. 6, p. 12-13, 1987.

PICHETH, C. M. T. F.; SOUZA, E. M.; RIGO, L.U.; FUNAYAMA S.; PEDROSA, F. O. Regulation of *Azospirillum brasilense nifA* gene expression by ammonium and oxygen. **FEMS Microbiology Letters**, p. 281-288, 1999.

RAITTZ, R. T.; Souza, J. A.; Dandolini, G. A.; Pacheco, R. C. S.; Martins, A.; Gauthier, F. A.; Barcia, M. FAN: learning by means of free associative neurons. In: **IEEE World Congress on Computational Intelligence**. Anchorage, AK, USA. p 425-430, 1998.

RCSB PDB. **RCSB PDB Protejn Data Bank.** Disponível em: <http://www.rcsb.org/pdb/home/home.do>. Último acesso: 21/05/2015.

SIVASHANKARI, S.; SHANMUGHAVEL, P. Functional annotation of hypothetical proteins – A review.2006. **Bioinformation, an open access forum**. Biomedical Informatics Publishing Group. 1(8): 335 -338, 2006.

VAN DER BERG, B. A.; REINDERS, M. J. T.; ROUBOS J. A.; RIDDER, D. SPiCE: a web-based tool for sequence-based protein classification and exploration. **BMC Bioinformatics**, 2014, 15:93

ZHANG, Y.; BURRIS, R. H.; LUDDEN, P. W., ROBERTS G. P. Posttranslational regulation of nitrogenase activity by anaerobiosis and ammonium in *Azospirillum brasilense*. **Journal of Bacteriology**. Vol 175, p. 6781–6788, 1993.

ZHANG, Y.; KIM, K.; LUDDEN, P.W.; ROBERTS, G.P. Isolation and characterization of *draT* mutants that have altered regulatory properties of dinitrogenase reductase ADPribosyltransferase in *Rhodospirillum rubrum*. **Microbiology**, Vol 147, p. 193–202, 2001.

APÊNDICES

APÊNDICE 1 - LIST OF BACTERIAL SPECIES HAVING AT LEAST 5 GENES *nif* AND THE PRESENCE OF THE GENES *nif*, *nifO*, *draT* AND *draG*

In the list below are all bacterial species that contains at least five essential *nif* genes according to ProClaT, analyzing the complete genomes of bacteria. The columns indicate the presence of *nifHDK*, *nifENB*, *nifO*, *draT* and *draG* genes.

Bacterial Specie x gene presence	*nifH*	*nifD*	*nifK*	*nifE*	*nifN*	*nifB*	*nifO*	*draT*	*draG*
Acidithiobacillus ferrivorans	x	x	x	x	x	x	x	x	x
Acidithiobacillus ferrooxidans	x	x	x	x	x	x	x	x	x
Allochromatium vinosum	x	x	x	x	x	x		x	x
Anabaena cylindrica	x	x	x	x	x	x	x		x
Anabaena sp.	x	x	x	x	x	x	x		
Anaeromyxobacter sp.	x	x	x	x		x		x	x
Arcobacter sp.	x	x	x	x		x	x		
Azoarcus sp.	x	x	x	x	x	x	x	x	
Azoarcus sp.	x	x	x	x	x	x		x	x
Azorhizobium caulinodans	x	x	x	x	x	x	x		
Azospirillum brasilense	x	x	x	x	x	x	x	x	x
Azospirillum lipoferum	x	x	x	x	x	x	x	x	x
Azospirillum sp.	x	x	x	x	x	x	x	x	x
***Azotobacter vinelandii* CA**	x	x	x	x	x	x	x		
Beijerinckia indica	x	x	x	x	x	x	x		
Bradyrhizobium diazoefficiens	x	x	x	x	x	x	x		
Bradyrhizobium japonicum	x	x	x	x	x	x	x		
Bradyrhizobium oligotrophicum	x	x	x	x	x	x	x		
Bradyrhizobium sp.	x	x	x	x	x	x	x		
Burkholderia phenoliruptrix	x	x	x	x	x	x			
Burkholderia phymatum	x	x	x	x	x	x			
Burkholderia sp.	x	x	x	x	x	x	x		
Burkholderia vietnamiensis	x	x	x	x	x	x	x		
Burkholderia xenovorans	x	x	x	x	x	x	x		
Calothrix sp.	x	x	x	x	x	x	x		x
Candidatus Accumulibacter phosphatis clade	x	x	x	x	x	x	x	x	x
Chlorobaculum parvum	x	x	x	x		x			
Chlorobium luteolum	x	x	x	x		x			
Chlorobium phaeobacteroides	x	x	x	x		x			
Chlorobium tepidum	x	x	x	x		x			
Chroococcidiopsis thermalis	x	x	x	x		x			
Clostridium kluyveri	x	x	x	x		x			
Clostridium saccharoperbutylacetonicum	x	x	x	x		x			
Cupriavidus taiwanensis	x	x	x	x	x	x			
***Cyanobacterium* UCYN-A**	x	x	x	x	x	x	x		
Cyanothece sp.	x	x	x	x	x	x	x		
Cylindrospermum stagnale	x		x	x	x	x	x		x
Dechloromonas aromatica	x	x	x	x	x	x	x	x	x

(continues)

Bacterial Specie x gene presence	*nifH*	*nifD*	*nifK*	*nifE*	*nifN*	*nifB*	*nifO*	*draT*	*draG*
***Dechlorosoma suillum* PS**	x	x	x	x	x	x	x	x	x
Denitrovibrio acetiphilus	x	x	x	x		x		x	x
Desulfovibrio vulgaris	x	x	x	x		x			
Dickeya dadantii	x	x	x	x	x	x			
Enterobacter sp.	x	x	x	x	x	x			x
Ethanoligenens harbinense	x	x	x	x		x			
Frankia alni	x	x	x	x		x			
Frankia sp.	x	x	x	x	x				
Frankia symbiont	x	x	x	x	x	x			
Gluconacetobacter diazotrophicus	x	x	x	x	x	x			
Halorhodospira halophila	x	x	x	x	x				
Halothece sp.	x	x	x	x	x	x			
Heliobacterium modesticaldum	x	x	x	x	x				
Herbaspirillum seropedicae	x	x	x	x	x	x			
Hyphomicrobium sp.	x	x	x	x	x	x	x		
Klebsiella oxytoca	x	x	x	x	x	x			x
Leptospirillum ferrooxidans	x	x	x	x	x	x	x		
Leptothrix cholodnii	x	x	x	x	x	x			
Magnetococcus marinus	x	x	x	x	x	x	x	x	x
Magnetospirillum magneticum	x	x	x	x	x	x	x	x	x
Mesorhizobium australicum	x	x	x	x	x	x			
Mesorhizobium ciceri biovar biserrulae	x	x	x	x	x	x			
Mesorhizobium loti	x	x	x	x	x	x			
Mesorhizobium opportunistum	x	x	x	x	x	x			
Methylacidiphilum infernorum	x	x	x	x	x	x			
Methylobacterium nodulans	x	x	x	x	x	x			
Methylobacterium sp.	x	x	x	x	x	x			
Methylococcus capsulatus	x	x	x	x	x	x	x	x	
Methylocystis sp.	x	x	x	x	x	x	x		
Methylomonas methanica	x	x	x	x	x	x	x	x	x
Microcoleus sp.	x	x	x	x	x	x			x
Nostoc punctiforme	x		x	x	x	x	x		x
Nostoc sp.	x	x	x	x	x	x	x		x
Paenibacillus polymyxa	x	x	x	x	x	x			x
Paenibacillus terrae	x	x	x	x	x	x			x
Paludibacter propionicigenes	x	x	x	x		x	x		
Pantoea sp.	x	x	x	x	x	x			
Pectobacterium atrosepticum	x	x	x	x	x	x			
Pelodictyon phaeoclathratiforme	x	x	x	x		x	x		
Pleurocapsa sp.	x	x	x	x	x	x	x		
Polaromonas naphthalenivorans	x	x	x	x	x	x	x		
Pseudomonas stutzeri	x	x	x	x	x	x	x		
Rahnella aquatilis	x	x	x	x	x	x			
Rhizobium etli	x	x	x	x	x	x			
Rhizobium leguminosarum	x	x	x	x	x	x			
Rhizobium tropici	x	x	x	x	x	x			
Rhodobacter capsulatus	x	x	x	x	x	x		x	x
Rhodobacter sphaeroides	x	x	x	x	x	x			
Rhodomicrobium vannielii	x	x	x	x	x	x	x		
Rhodopseudomonas palustris	x	x	x	x	x	x	x	x	x
***Rhodospirillum centenum* SW**	x	x	x		x	x			
Rhodospirillum photometricum	x	x	x	x	x	x		x	x
Rhodospirillum rubrum	x	x	x	x	x	x	x	x	x
Sideroxydans lithotrophicus	x	x	x	x	x	x	x	x	x

(continues)

Bacterial Specie x gene presence	*nifH*	*nifD*	*nifK*	*nifE*	*nifN*	*nifB*	*nifO*	*draT*	*draG*
Sinorhizobium fredii	x	x	x	x	x	x			
Sinorhizobium medicae	x	x	x	x	x	x			
Sinorhizobium meliloti	x	x	x	x	x	x			
Spirochaeta smaragdinae	x	x	x	x		x			x
Sulfuricurvum kujiense	x	x	x	x		x	x		
Sulfurospirillum barnesii	x	x	x	x	x	x			
Synechococcus sp.	x	x	x	x	x	x			
Syntrophobacter fumaroxidans	x	x	x	x		x			
Syntrophobotulus glycolicus	x	x	x	x		x			
Teredinibacter turnerae	x	x	x	x	x	x	x	x	x
Thermoanaerobacterium thermosaccharolyticum	x	x	x	x		x			
Thioflavicoccus mobilis	x	x	x	x	x	x		x	x
Tolumonas auensis	x	x	x	x	x	x		x	x
Treponema primitia	x	x	x	x		x			
Trichodesmium erythraeum	x	x	x	x		x	x		
Xanthobacter autotrophicus	x	x	x	x	x	x	x		
Zymomonas mobilis	x	x	x	x	x	x	x		

APÊNDICE 2 - ADDITIONAL RESULTS

This section provides additional information about the proteins classification with ProClaT.

2.1 NCBI NR PROTEINS CLASSIFICATION USING PROCLAT

Figure 1 shows how the NCBI NR proteins classified as NifO-like are currently annoted.

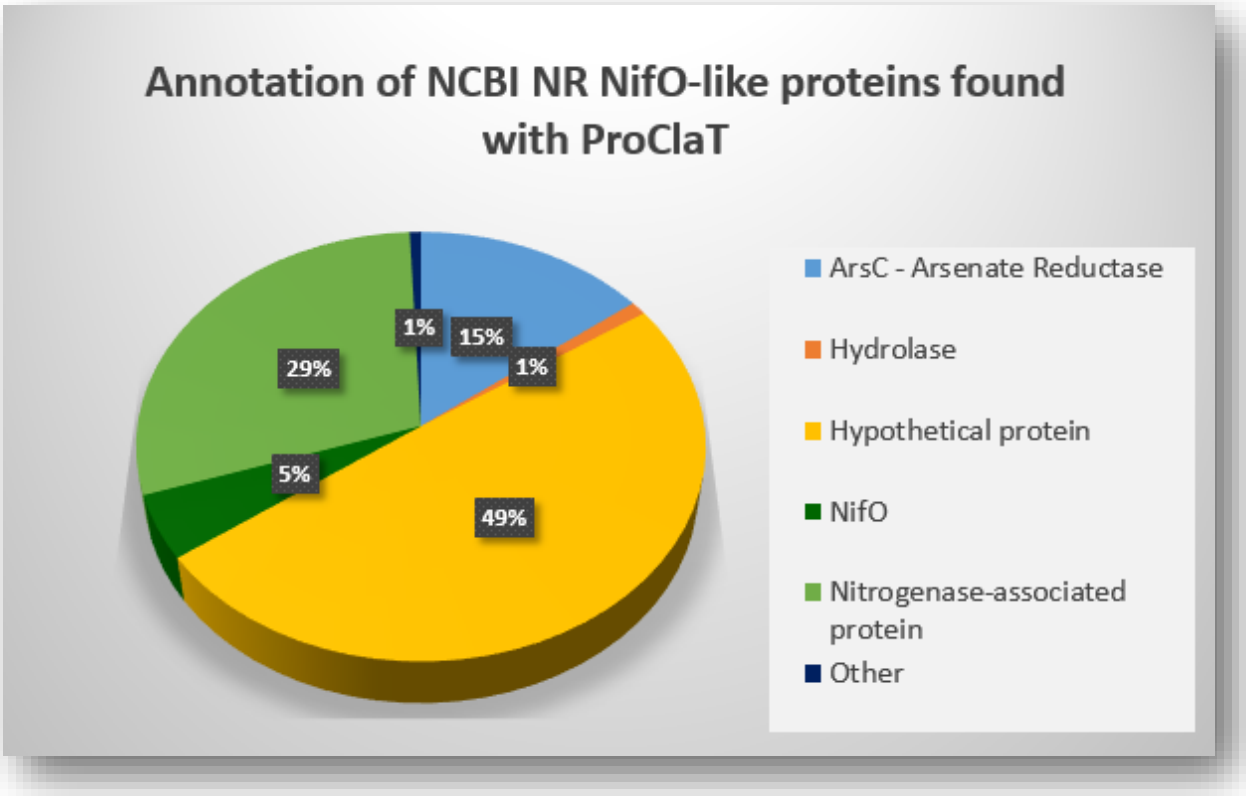


FIGURE 1 – ANNOTATION OF PROTEINS CLASSIFIED AS NifO-LIKE WITH PROCLAT. WERE SELECTED 5,000 PROTEINS BY A PSI-BLAST IN NCBI NR USING NifO CONSENSUS REGION AS QUERY. 289 NifO-LIKE PROTEINS WERE FOUND, ACCORDING TO PROCLAT. ALMOST HALF OF THEM ARE ANNOTATED AS HYPOTHETICAL PROTEIN.

2.2 PROTEINS CLASSIFICATION USING PROCLAT IN COMPLETE BACTERIAL STRAINS

Figure 2 shows the number of genes group found in the complete genome with ProClaT, analyzing all bacterial strains.

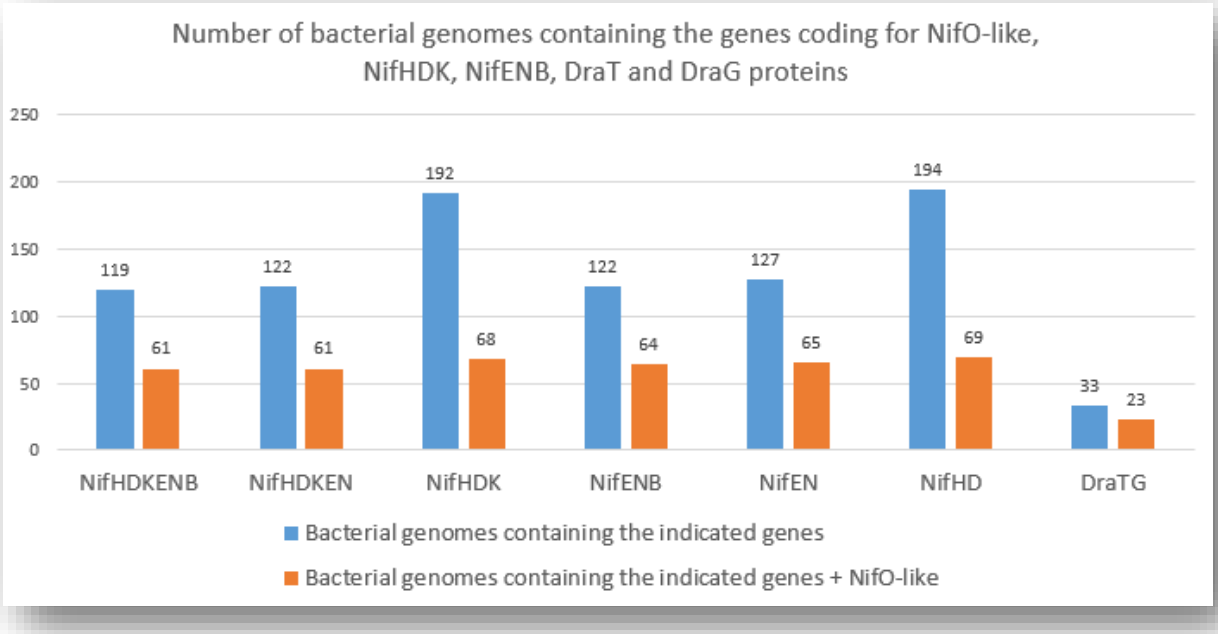


FIGURE 2 – IN BLUE, THE NUMBER OF ALL STRAINS OF BACTERIAL COMPLETE GENOMES CONTAINING THE GENES INDICATED BELOW, AND IN RED, THE NUMBER OF THE GENOMES CONTAINING THESE GENES IN ADDITION WITH THE GENE CODING FOR NifO- LIKE. DATA GENERATED WITH PROCLAT APPLICATION.

Figure 3 shows the number of genes coding for NifO-like in the presence of some gene coding for a essential Nif protein in complete genomes, analyzing bacterial strains.

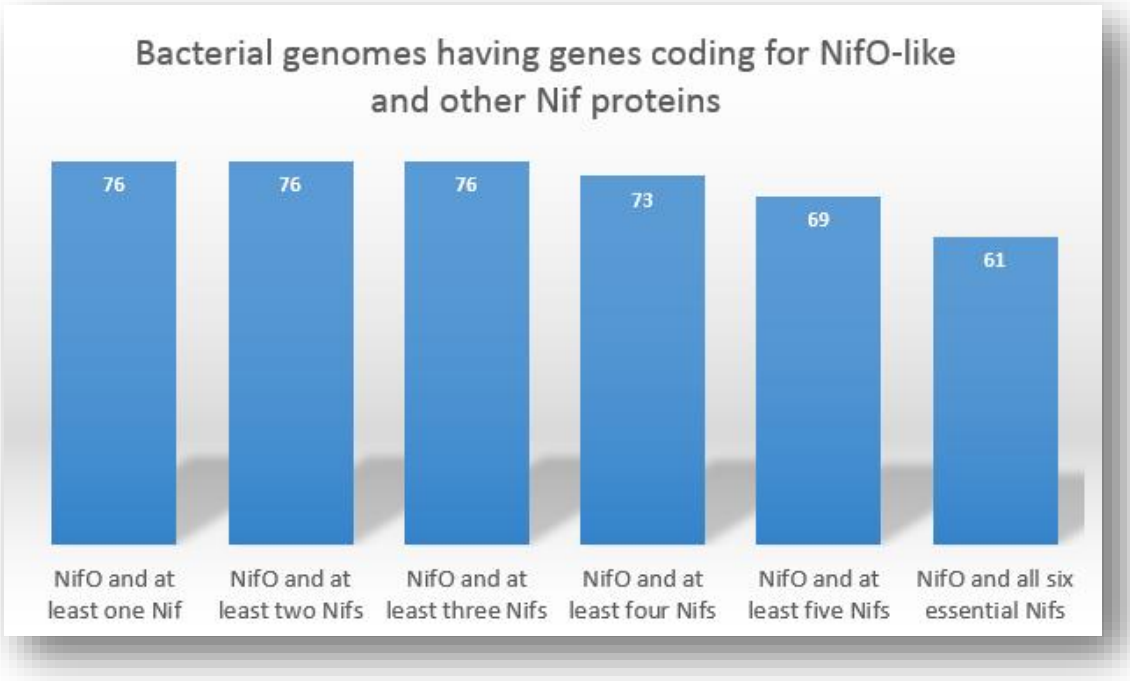


FIGURE 3 - PROCLAT IDENTIFIED 76 BACTERIAL STRAINS CONTAINING GENES CODING FOR NifO- LIKE. ALL BELONG TO A GENOME THAT CONTAIN AT LEAST THREE GENES CODING FOR A ESSENTIAL Nif PROTEIN. OF THE 73 STRAINS THAT CONTAIN AT LEAST 4 *nif* GENES, 69 CONTAIN AT LEAST 5 *nif* GENES AND 61 CONTAIN ALL THE 6 ESSENTIAL *nif* GENES.

2.3 CO-OCCURRENCE CORRELATION COEFFICIENTS

With the classification process in the complete bacterial genomes with the ProClaT tool, we found the genomes containing the genes coding for NifO-like, NifHDK, NifENB, DraT and DraG. Thus, the co-occurring correlations of the genes coding for these proteins were generated using the Pearson coefficient. Table 1 shows the coefficient correlation between the gene coding for NifO-like with all six essential *nif* genes, while Table 2 shows the correlation with *draTG* genes.

TABLE 1 - PEARSON'S CORRELATION COEFFICIENT OF CO-OCCURRENCE OF GENES CODING FOR NifO-LIKE WITH SIX ESSENTIAL *nif* GENES

Gene	***nifHDKENB***	***nifHDK***	***nifENB***
nifO	0,6350	0,5539	0,6586

TABLE 2 - PEARSON'S CORRELATION COEFFICIENT OF CO-OCCURRENCE OF GENES CODING FOR NifO-LIKE WITH *draT* AND *draG* GENES.

Gene	***draTG***	***draT***	***draG***
nifO	0,4544	0,4923	0,1901

APÊNDICE 3 – CLASSIFICATION OF *A. brasilense* DraB PROTEIN WITH EXISTING BIOINFORMATICS TOOLS

The *A. brasilense* DraB protein was subjected to some existing Bioinformatics tools to predict function / Gene Onthology (GO) terms.

3.1 AmiGO

Searching for homologous proteins, it was performed a BLAST in GO database using DraB as query with AmiGO [23] (collection of web tools for searching and browsing in the GO database), version 1.8. No results were found.

3.2 Blast2GO

The *A. brasilense* DraB protein was submitted to Blast2GO [25] version 3.0.9 PRO in 04/06/2015. With the mapping process, seen in Figure 1, two GO descriptors have been selected for the DraB protein: GO: 0055114 (ontology: biological process, description: oxidation-reduction process) and GO: 0016491 (ontology: molecular function, description: oxidoreductase activity).

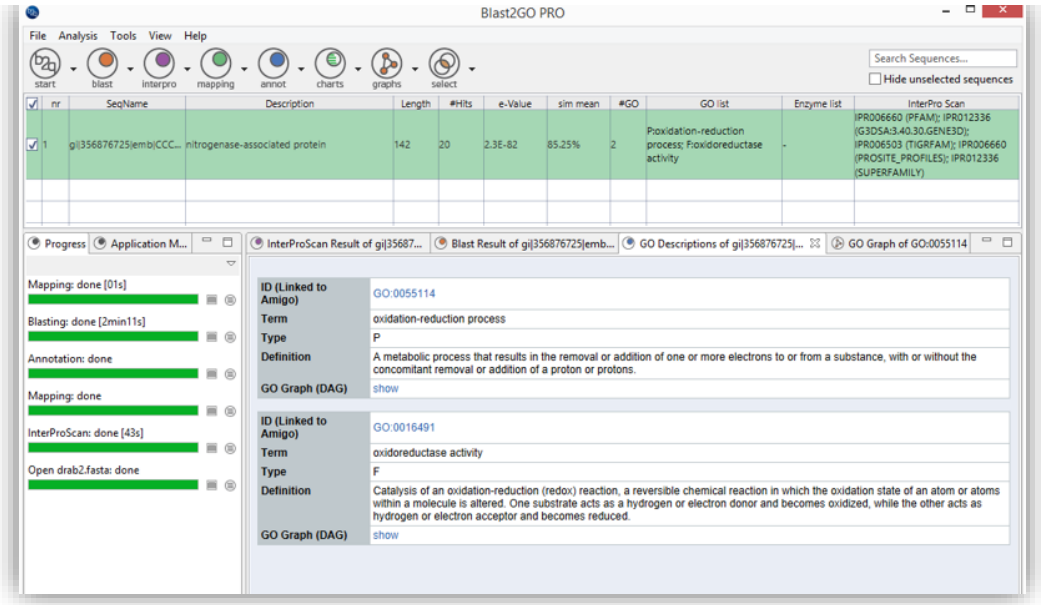


FIGURE 1 – RESULT OF MAPPING PROCESS OF DraB PROTEIN, PERFORMED IN BLAST2GO.

In Figure 2 are the graphics mapping of the biological process and molecular function suggested to DraB protein by Blast2GO tool. The results were supported by two hits, ie, only two proteins similar to query have GO description.

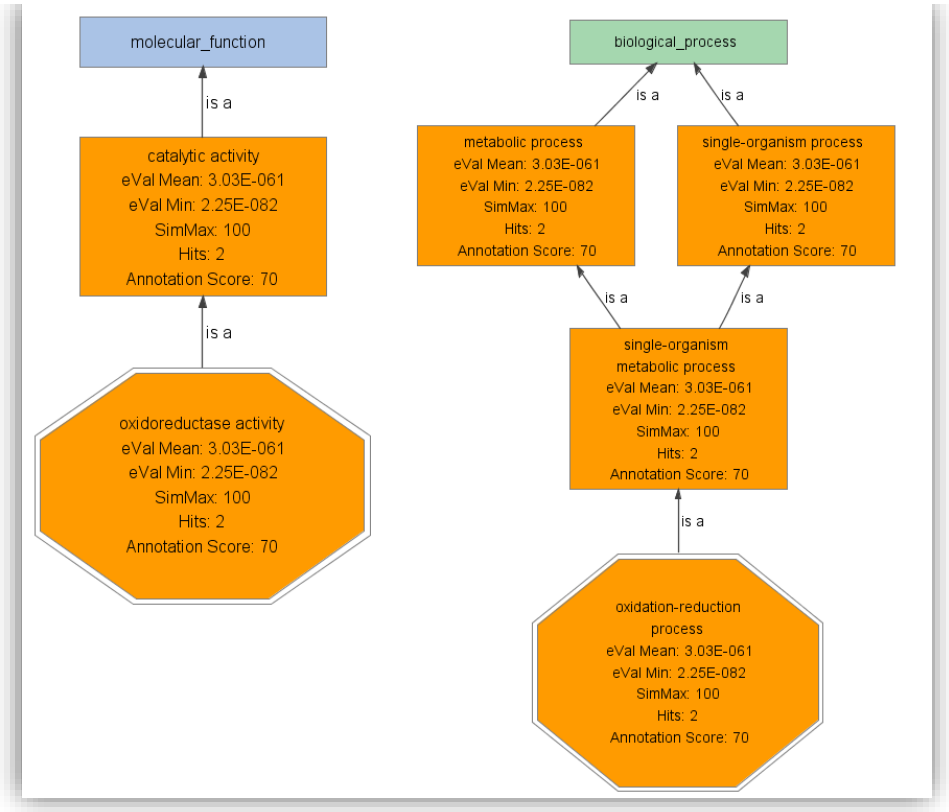
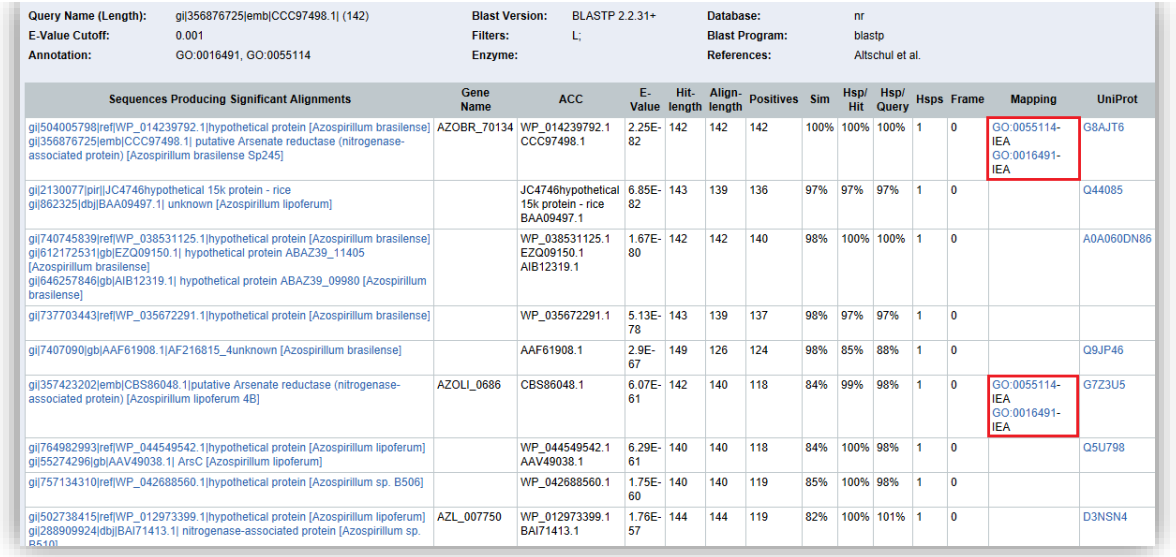


FIGURE 2 – GRAPHICS OF BIOLOGICAL PROCESS AND MOLECULAR FUNCTION, RESPECTIVELY, SUGGESTED FOR DraB PROTEIN BY BLAST2GO.

Figure 3 shows the homologous proteins that the tool used to infer the molecular function and biological process for the input protein.



Query Name (Length):	gi|356876725|emb|CCC97498.1| (142)	**Blast Version:**	BLASTP 2.2.31+	**Database:**	nr							
E-Value Cutoff:	0.001	**Filters:**	L;	**Blast Program:**	blastp								
Annotation:	GO:0016491, GO:0055114	**Enzyme:**		**References:**	Altschul et al.								

Sequences Producing Significant Alignments	Gene Name	ACC	E-Value	Hit. length	Align-length	Positives	Sim	Hsp/Hit	Hsp/Query	Hsps	Frame	Mapping	UniProt
gi|504005798|ref|WP_014239792.1|hypothetical protein [Azospirillum brasilense] gi|356876725|emb|CCC97498.1| putative Arsenate reductase (nitrogenase-associated protein) [Azospirillum brasilense Sp245]	AZOBR_70134	WP_014239792.1 CCC97498.1	2.25E-82	142	142	142	100%	100%	100%	1	0	GO:0055114-IEA GO:0016491-IEA	G8AJT6
gi|2130077|pir|JC4746hypothetical 15k protein - rice gi|862325|dbj|BAA09497.1| unknown [Azospirillum lipoferum]		JC4746hypothetical 15k protein - rice BAA09497.1	6.85E-82	143	139	136	97%	97%	97%	1	0		Q44085
gi|740745839|ref|WP_038531125.1|hypothetical protein [Azospirillum brasilense] gi|612172531|gb|EZQ09150.1| hypothetical protein ABAZ39_11405 [Azospirillum brasilense] gi|646257846|gb|AIB12319.1| hypothetical protein ABAZ39_09980 [Azospirillum brasilense]		WP_038531125.1 EZQ09150.1 AIB12319.1	1.67E-80	142	142	140	98%	100%	100%	1	0		A0A060DN86
gi|737703443|ref|WP_035672291.1|hypothetical protein [Azospirillum brasilense]		WP_035672291.1	5.13E-78	143	139	137	98%	97%	97%	1	0		
gi|7407090|gb|AAF61908.1|AF216815_4unknown [Azospirillum brasilense]		AAF61908.1	2.9E-67	149	126	124	98%	85%	88%	1	0		Q9JP46
gi|357423202|emb|CBS86048.1|putative Arsenate reductase (nitrogenase-associated protein) [Azospirillum lipoferum 4B]	AZOLI_0686	CBS86048.1	6.07E-61	142	140	118	84%	99%	98%	1	0	GO:0055114-IEA GO:0016491-IEA	G7Z3U5
gi|764982993|ref|WP_044549542.1|hypothetical protein [Azospirillum lipoferum] gi|55274296|gb|AAV49038.1| ArsC [Azospirillum lipoferum]		WP_044549542.1 AAV49038.1	6.29E-61	140	140	118	84%	100%	98%	1	0		Q5U798
gi|757134310|ref|WP_042688560.1|hypothetical protein [Azospirillum sp. B506]		WP_042688560.1	1.75E-60	140	140	119	85%	100%	98%	1	0		
gi|502738415|ref|WP_012973399.1|hypothetical protein [Azospirillum lipoferum] gi|288909924|dbj|BAI71413.1| nitrogenase-associated protein [Azospirillum sp. B510]	AZL_007750	WP_012973399.1 BAI71413.1	1.76E-57	144	144	119	82%	100%	101%	1	0		D3NSN4

FIGURE 3 – RESULT OF NCBI NR BLAST PERFORMED IN BLAST2GO, BEFORE THE MAPPING PROCESS. MARKED IN RED, THE SEQUENCES USED IN PREDICTION OF THE GO VALUE FOR DraB PROTEIN.

3.3 ConFunc

The DraB protein also was subjected to ConFunc [24] on 04/04/2015 under the Job Unique Identifier dc7d735a2a8da9e1. The results can be seen in Table 1.

TABLE 1 – PREDICTION OF MOLECULAR FUNCTION AND BIOLOGICAL PROCESS GO TERMS FOR THE DraB PROTEIN BY CONFUNC, USING COMBINED METHODS.

GO Term	Description	SVM	Probability
Prediction of molecular function – combined methods			
GO:0008794	Arsenate reductase (glutaredoxin) activity	10	0.667
GO:0005488	Binding	10	0.348
GO:0016209	Antioxidante activity	9	0.294
Prediction of the biological process - combined methods			
GO:0006351	Transcription, DNA-templated	10	0.306
GO:0045892	Negative regulation of transcription, DNA-templated	10	0.306
GO:0050896	Response to stimulus	9	0.301

THE FIRST COLUMNS DISPLAY THE GO TERMS AND THEIR CORRESPONDENT DESCRIPTIONS, THE THIRD COLUMN SHOWS THE NUMBER OF SVM THAT HAVE PREDICTED THIS TERM (OF 10), AND THE LAST COLUMN INDICATES THE CONFIDENCE OF THE PREDICTED TERM.

According to ConFunc, DraB protein has 66.7% probability of having arsenate reductase activity, 34.8% of binding, and 29.4% antioxidant activity. Regarding the cellular processes, DraB has 30.6% probability of being related to transcription, DNA-templated or negative regulation of transcription, DNA-templated and 30.1% of response to stimulus.

The results of ConFunc individual analysis are catalytic activity (GO: 0003824) and oxidoreductase activity (GO: 0016491), both with *z-score* of 7.7508 for function prediction; and oxidation-reduction process (GO: 0055114), metabolism (GO: 0044710) and process only body (GO: 0044699), all having *z-score* 8.0931; being *z-score* calculated for the significance of the result.

3.4 GOtcha

The *A. brasilense* DraB protein was submitted to the GOtcha [30] on 04/04/2015, under the job 1428170538-14922. The results can be seen in Table 2.

TABLE 2 – GO TERMS PREDICTION FOR CELLULAR COMPONENT, MOLECULAR FUNCTION AND BIOLOGICAL PROCESS FOR DraB PROTEIN WITH GOtcha TOOL

GO Term	Description	Confidence	Confidence level
Cellular componente			
GO:0005575	Cellular componente	0.2	Low
Molecular function			
GO:0003674	Molecular function	1	Low
Biological process			
GO:0008150	Biological process	1	Low
GO:0008152	Metabolism process	0.4	Low

With GOtcha, could not perform the prediction of DraB protein because the returned GO terms are not specific and the confidence level is low.

3.5 PFP

The DraB protein was subjected to PFP [31] on 04/04/2015, under the job id

37765. Results can be seen in Table 3.

TABLE 3 – GO TERMS PREDICTION FOR CELLULAR COMPONENT, MOLECULAR FUNCTION AND BIOLOGICAL PROCESS FOR DraB PROTEIN WITH PFP TOOL

GO Term	Description	Confidence	Confidence level
Cellular componente			
GO:0005737	Cytoplasm	121.98	Low
GO:0005622	Intracellular	113.70	Low
GO:0044424	Intracellular part	111.72	Low
Molecular function			
GO:0000166	Nucleotide binding	49.67	Very low
GO:0005488	Binding	38.77	Very low
GO:0005524	ATP binding	37.04	Very low
Biological process			
GO:0006351	Transcription-dependent DNA	88.75	Very low
GO:0032774	RNA biosynthesis Process	87.26	Very low
GO:0008152	Metabolism process	68.35	Very low

FONTE: A autora

With PFP, could not perform the prediction of DraB protein because the returned GO terms are not specific and the confidence level is low or very low.

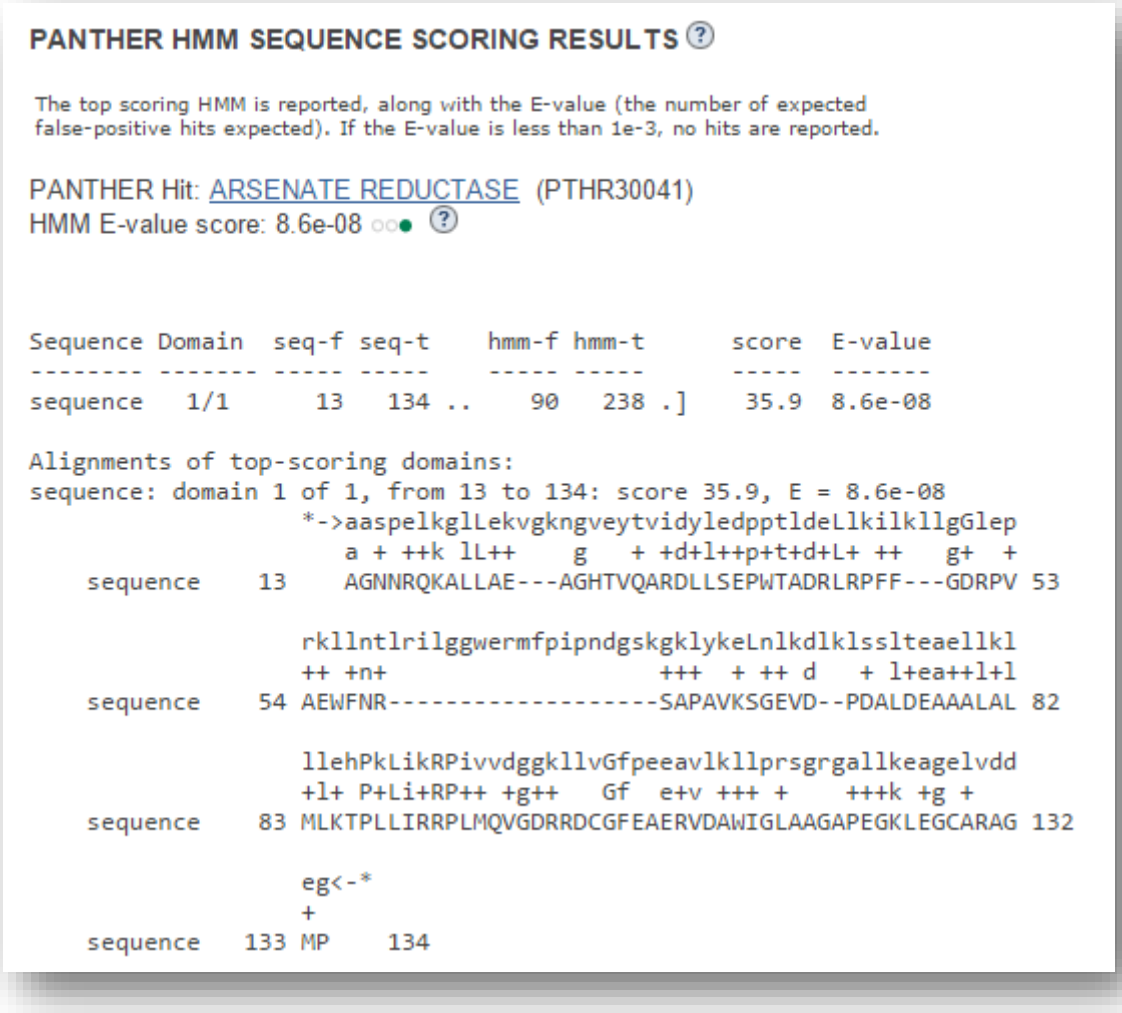
3.6 InterPro

With InterPro [26] 51.0 web tool, using the default values as parameters, were found to DraB the Thioredoxin-like fold domain (IPR012336), and families arsenate reductase-like (IPR006660) by applications/databases Pfam and PrositeProfiles and Nitrogenase -associated protein (IPR006503) by application/ database TIGRFAM. The results can be seen in Figure 4. The prediction of GO terms returned no results.

FIGURE 4 – RESULT OF SEARCH FOR DOMAINS AND FAMILIES OF DraB PROTEIN WITH INTERPRO.

3.7 PANTHER

With the PANTHER tool [27] we performed a search in the database PANTHER by the sequence of DraB protein, and the HMM model returned was arsenate reductase (PTHR30041). The protein class in database or its pathway were not returned. The results were based on 48 genes, including *Escherichia coli* ArsC. The *e-value* of $8.6e^{-08}$ presents the "distantly related" indication, i.e., the DraB protein is evolutionarily related with the model protein found by the tool, but its function may have diverged. The results can be seen in Figure 5. The prediction of GO terms returned no results.



PANTHER HMM SEQUENCE SCORING RESULTS

The top scoring HMM is reported, along with the E-value (the number of expected false-positive hits expected). If the E-value is less than 1e-3, no hits are reported.

PANTHER Hit: ARSENATE REDUCTASE (PTHR30041)
HMM E-value score: 8.6e-08

```
Sequence Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
sequence   1/1      13   134 ..    90   238 .]    35.9  8.6e-08

Alignments of top-scoring domains:
sequence: domain 1 of 1, from 13 to 134: score 35.9, E = 8.6e-08
                   *->aaspelkglLekvgkngveytvidyledpptldeLlkilkllgGlep
                      a + ++k lL++    g   + +d+l++p+t+d+L+ ++   g+  +
    sequence    13    AGNNRQKALLAE---AGHTVQARDLLSEPWTADRLRPFF---GDRPV 53

                   rkllntlrilggwermfpipndgskgklykeLnlkdlklsslteaellkl
                   ++ +n+                   +++  + ++ d   + l+ea++l+l
    sequence    54 AEWFNR------------------SAPAVKSGEVD--PDALDEAAALAL 82

                   llehPkLikRPivvdggkllvGfpeeavlkllprsgrgallkeagelvdd
                   +l+ P+Li+RP++ +g++   Gf  e+v +++ +    +++k +g +
    sequence    83 MLKTPLLIRRPLMQVGDRRDCGFEAERVDAWIGLAAGAPEGKLEGCARAG 132

                   eg<-*
                   +
    sequence   133 MP    134
```

FIGURE 5 – RESULT OF SEARCH FOR HMM CORRESPONDENT MODEL TO DraB PROTEIN WITH PANTHER.

3.8 Pfam

With EMBL-EBI Pfam web tool [28], the DraB protein was classified in ArsC family (PF03960 - or IPR006660), related to glutaredoxinas, part of the Thioredoxin family with *e-value* of 9.3 e^{-12}, same result returned by InterPro tool.

3.9 PROSITE

According to Expasy PROSITE [29], the DraB protein belongs to the ArsC family (Thioredoxin family), same result obtained by InterPro and Pfam tools, but results based on only one hit.

3.10 EXPASY SwissModel

In order to compare the structure of *A. brasilense* DraB protein, *E.coli* ArsC and *A. vinelandii* NifO, PDB files containing information of protein tertiary structures were generated with Expasy SwissModel tool [32] as well as scripts (in Python) to perform the alignment of the proteins and generation of protein structure image with PyMOL tool [33].

Figures 6 and 7 show the possible structures of DraB x NifO and DraB x ArsC proteins, respectively, generated with SwissModel (PDB) and aligned by PyMOL.

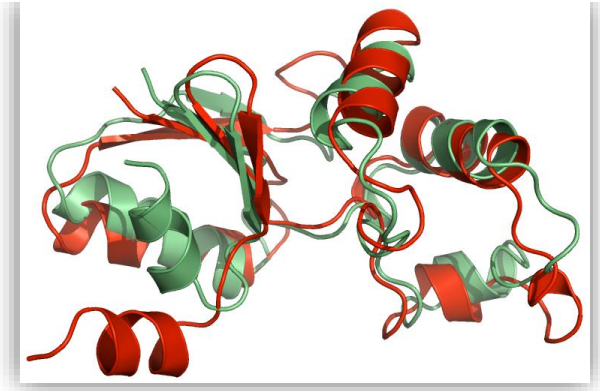


FIGURA 6 – ALIGNMENT AND OVERLAP OF POSSIBLE 3D STRUCTURES OF PROTEINS *A. brasilense* DraB (RED) AND *A. vinelandii* NifO (GREEN) GENERATED WITH SWISSMODEL AND PyMOL TOOLS.

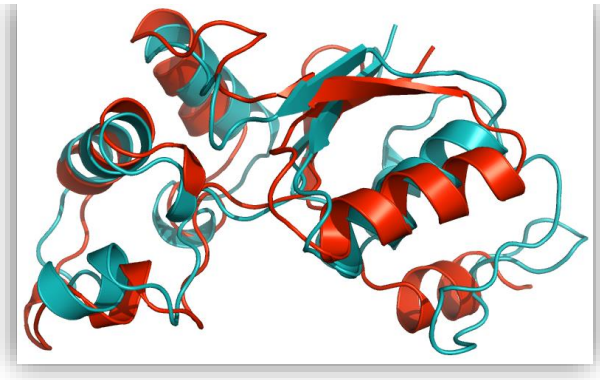


FIGURA 7 – ALIGNMENT AND OVERLAP OF POSSIBLE 3D STRUCTURES OF PROTEINS *A. brasilense* DraB (RED) AND *E. coli* ArsC (BLUE) GENERATED WITH SWISSMODEL AND PyMOL TOOLS.

Figure 8 shows a comparison of the possible structures of proteins DraB, NifO and ArsC at two different angles.

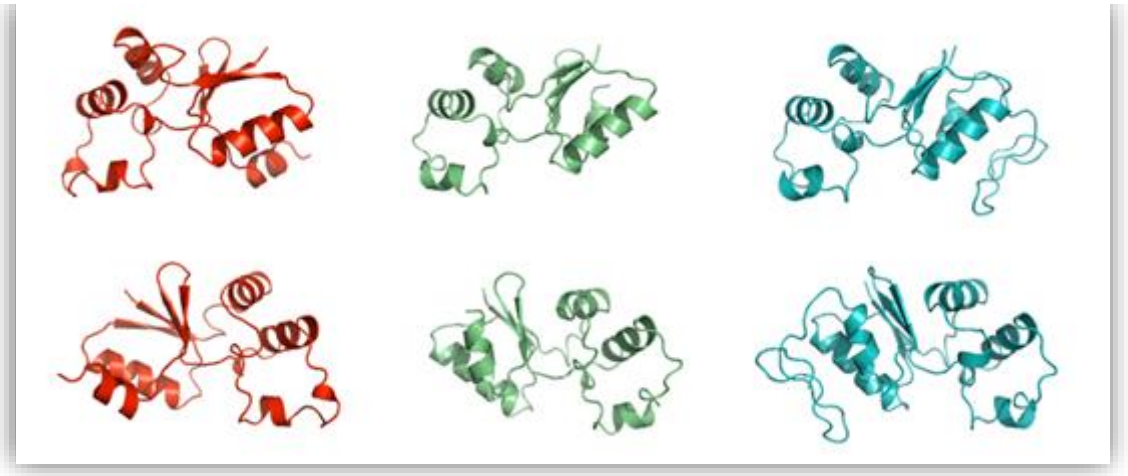


FIGURA 8 – ALIGNMENT OF POSSIBLE 3D STRUCTURES OF PROTEINS *A. brasilense* DraB (RED), *A. vinelandii* NifO (GREEN) AND *E. coli* ArsC (BLUE) AT TWO DIFFERENT ANGLES, GENERATED WITH SWISSMODEL AND PyMOL TOOLS.

Although it appears a structural similarity of DraB with both NifO and ArsC, the generated structures via theoretical modeling (non-experimentally) may contain errors, especially in automatic modeling, where there is no human intervention during building model [32]. The structure of DraB protein was generated based on the template "2kok.1.A" - arsenate reductase (obtained by the NMR method) with cover 0.80 and 28.07% of identity. The structure of NifO protein was generated based on the template "3f0i.1.A", also an arsenate reductase (obtained by X-ray) with 0.76 coverage and 25.89% of identity. The ArsC structure was based on the template "1j9b.1.A", arsenate reductase (obtained by X-ray), with a cover 1 and 100% of identity. The 3D structures of the three proteins were assembled based on arsenate reductase, since the SwissModel tool did not found closest templates for DraB and NifO.

APÊNDICE 4 – MATERIAIS E MÉTODOS – INFORMAÇÕES COMPLEMENTARES

Nesta seção são apresentadas informações complementares sobre os materiais e métodos utilizados no trabalho.

4.1 MÁQUINAS DE TRABALHO

- Computador LENOVO Intel Core i5 CPU 650 3.2 GHz – 64 bits – Memória RAM 1.7 GB. Sistema operacional Ubunto 12.04 LTS e Windows 7 (alocado no laboratório UFPR - Bioinfo).
- HP Pavilion dm1 Notebook PC, Sistema Operacional Windows 7, Processador AMD 1.60 GHz, Memória RAM de 3 GB (propriedade da aluna).
- Computador Intel ® Core ™ i5-4460 CPU 3.20 GHz – 64 bits - Memória RAM 16 GB. Sistema operacional Windows 8 (propriedade da aluna).

4.2 PROTEÍNAS DE REFERÊNCIA

Número de acesso (GI [5]) das proteínas utilizadas:

- DraB de *A. brasilense*: 356876725;
- DraB de *R. rubrum*: 83575257;
- NifO de *A. vinelandii*: 502038043;
- DraT de *A. brasilense*: 356876723;
- DraG de *A. brasilense*: 356876724;
- ArsC de *E. coli*: 693150965.

4.3 PROCLAT

4.3.1 Escolha do algoritmo de classificação

[5] O número GI é uma série de dígitos atribuídos consecutivamente a cada sequência registrada pelo NCBI, usado como identificador do registro.

O algoritmo de classificação escolhido para o ProClaT foi o MLP. A escolha foi feita através do emprego dos arquivos de treinamento e teste em seis algoritmos, conforme consta no artigo. A ferramenta *KnowledgeFlow* do Weka foi utilizada para realizar as comparações, conforme as Figuras 1 e 2.

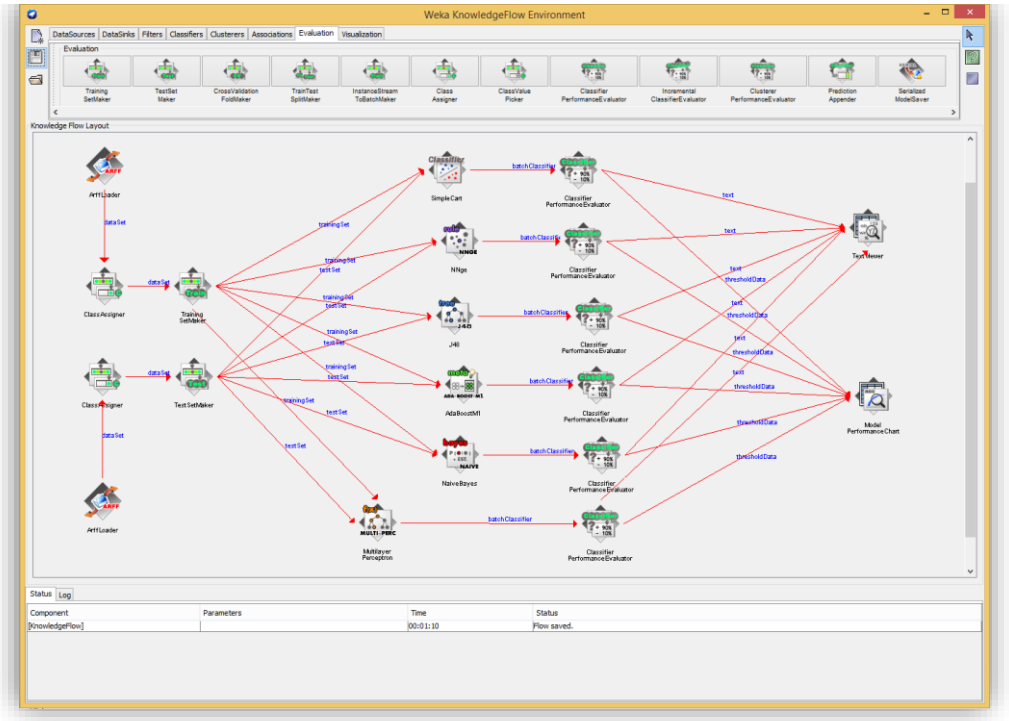


FIGURA 1 – WORKFLOW PARA APLICAÇÃO DOS ARQUIVOS DE TREINAMENTO E TESTE EM VÁRIOS ALGORITMOS DE CLASSIFICAÇÃO, SEM A OPÇÃO *CROSS-VALIDATION*, GERADO NA FERRAMENTA *KNOWLEDGEFLOW* DO WEKA .
FONTE: A autora.

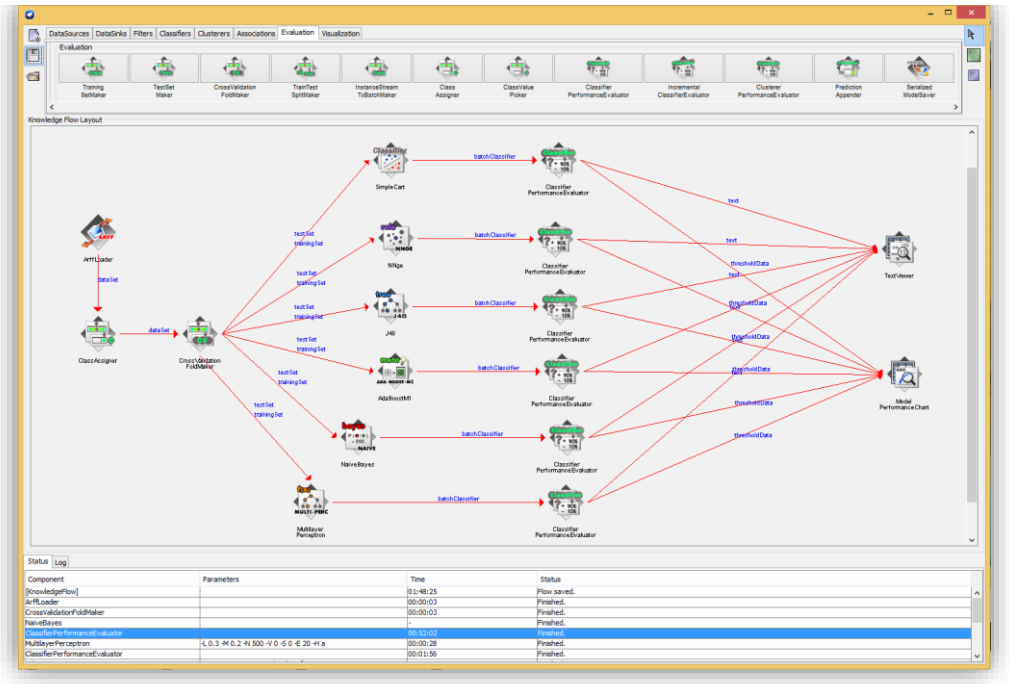


FIGURA 2 – WORKFLOW PARA APLICAÇÃO DOS ARQUIVOS DE TREINAMENTO E TESTE EM VÁRIOS ALGORITMOS DE CLASSIFICAÇÃO, COM A OPÇÃO *CROSS-VALIDATION*, GERADO NA FERRAMENTA *KNOWLEDGEFLOW* DO WEKA .
FONTE: A autora.

A Figura 3 mostra o treinamento da rede neural MLP realizado no Matlab.

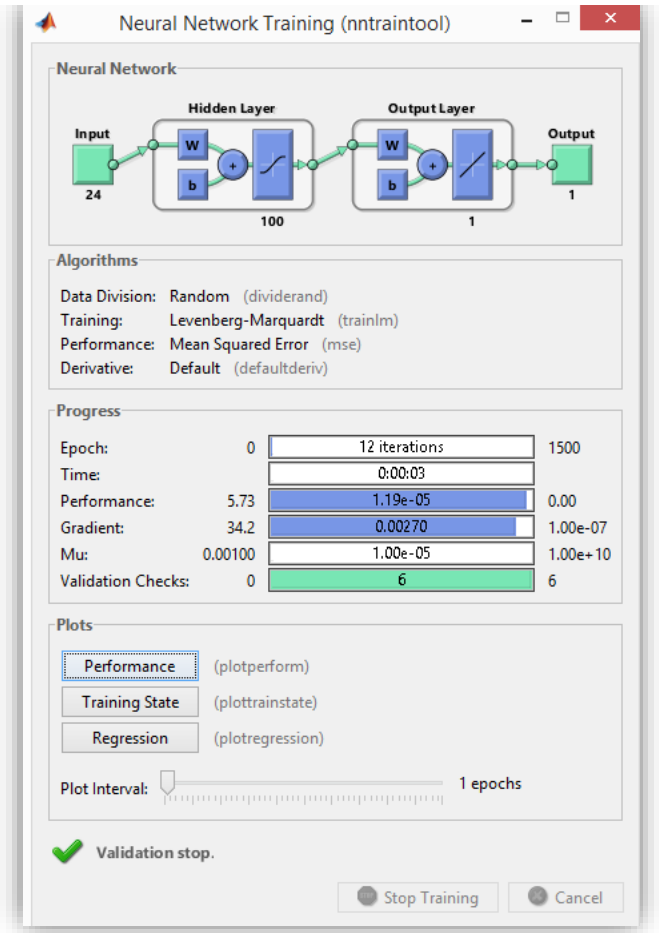


FIGURA 3 – EXECUÇÃO DO TREINAMENTO DA REDE NEURAL MLP REALIZADO NO MATLAB. FONTE: A autora.

4.3.2 Tempo de processamento

O tempo de processamento da classificação de todos os genomas bacterianos foi de 136,11 horas, ou 5,67 dias ininterruptos. Considerando que foram 5.182 genomas classificados, presentes nas 2.773 estirpes bacterianas, a média foi de 1,57 minutos por genoma.

4.3.3 Informações técnicas sobre os classificadores gerados

A seguir, são apresentados a URL contendo o endereço para obtenção das proteínas utilizadas para gerar os domínios conservados do banco de dados UniProt/SwissProt, a região consenso gerada, o domínio conservado gerado pelo Expasy PRATT e os valores de desempenho de cada classificador. Para as proteínas

Nif, as proteínas com sequência menor que 50 aminoácidos foram descartadas. Para a DraG e DraT, proteínas com sequência maior que 400 aminoácidos foram descartadas. ProClaT utiliza o domínio conservado após a fase de refinamento (aplicação de um algoritmo nos padrões encontrados durante a fase de pesquisa inicial, na qual símbolos mais ambíguos podem ser adicionados [6]). Para compor a classe "TRUE", as proteínas devem possuir o domínio conservado e possuir escore de alinhamento global maior que 0.2 com a região consenso gerada.

4.3.3.1. CLASSIFICADOR DE NifO-*LIKE*

Para gerar o domínio conservado e região consenso de NifO-*like*, foram utilizadas 14 proteínas, retornadas da URL:
http://www.uniprot.org/uniprot/?query=name%3Anifo&format=xml

A região consenso gerada com essas proteínas foi:

MHWFWAFRRQGLLSHHFSGYQGCRHADASTVPSVPPGSTGDDAARLRMKGTRSMANVVFYEKPGCRNNTKQ
KNLLLAAGHNLEERNLLTEPWKPENLRPFFGDLPVNEWFNPSAPRIKSGEVIPDNLNPEQALELMIADPLLIRRPLI
HVDGRRRVGFDPEKIDAWIGLNPDDPVTEDLETCPRSHEEQGCTHDNVHTHHKCACKER

O Quadro 1 mostra o domínio conservado gerado para proteína NifO-*like*.

Fase	***Fitness***	***Hits (seqs)***	***Pattern***	**Expressão Regular**
Antes refinamento	29.1904	14(14)	P-x-L-I-R-R-P-L	P.LIRRPL
Após refinamento	32.0793	14(14)	P-x-L-I-R-R-P-L-ILM]	P.LIRRPL[ILM]

QUADRO 1 – DOMÍNIO CONSERVADO DA PROTEÍNA NifO-*LIKE* GERADO COM A FERRAMENTA EXPASY PRATT E A EXPRESSÃO REGULAR CORRESPONDENTE.
FONTE: A autora

O Quadro 2 mostra o desempenho do classificador MLP para proteina NifO-*like*.

Acertos	**Matriz confusão**

[6] Conforme documentação: http://web.expasy.org/pratt/pratt_doc.html

99.28	71 1 0 67

QUADRO 2 – DESEMPENHO DO CLASSIFICADOR PARA PROTEÍNA NifO-*LIKE*.
FONTE: A autora

O número total de registros de treinamento é 323 e de teste, 139.

Na Figura 4, encontra-se a árvore filogenética com 40 proteínas aleatórias utilizadas no arquivo de treinamento do classificador, sendo metade pertencente a cada classe (1 e 0). É possível perceber que a classe 0 (em vermelho) deriva de um mesmo ancestral comum.

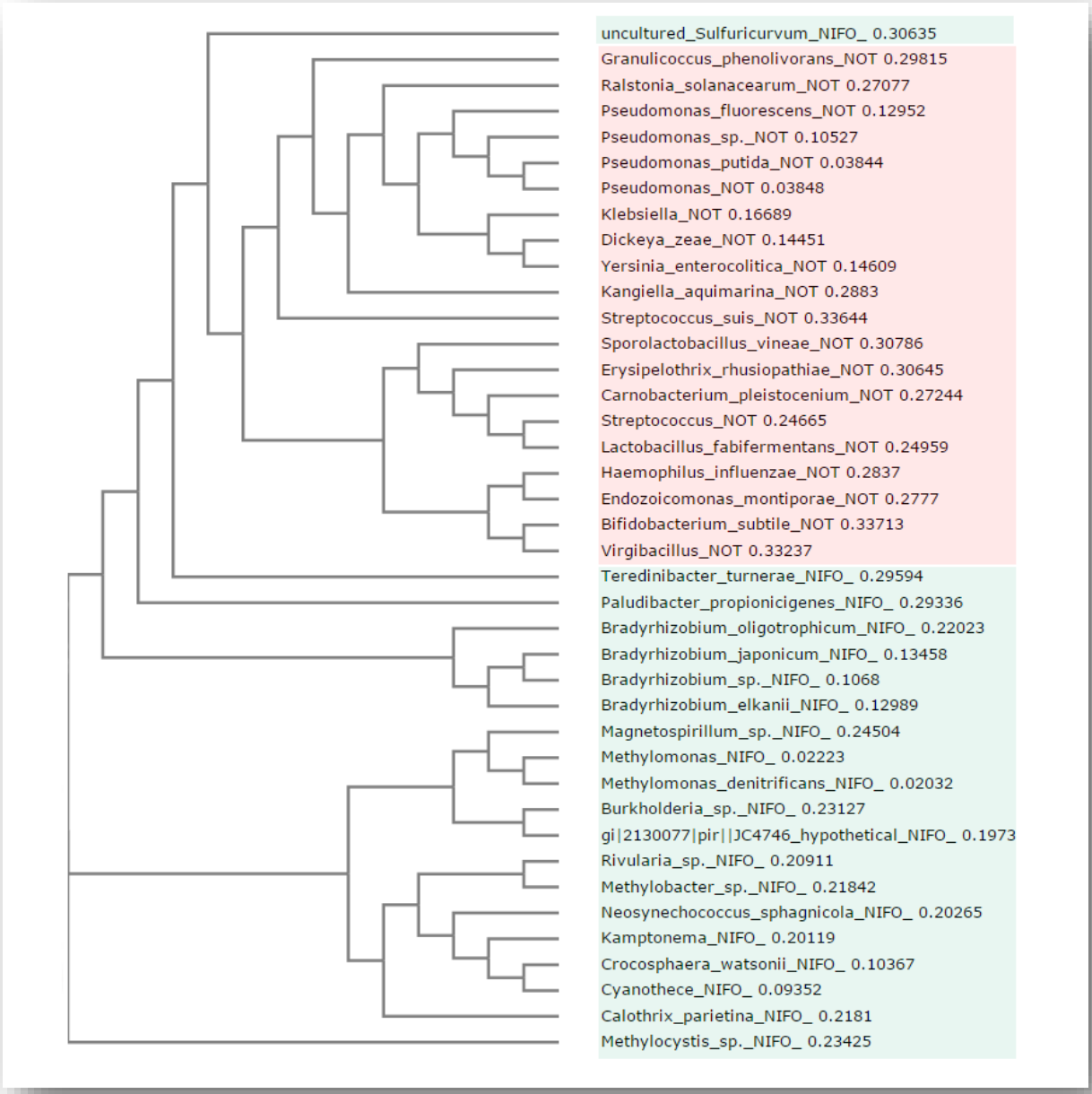


FIGURA 4 – ÁRVORE FILOGENÉTICA DE PROTEÍNAS UTILIZADAS NO TREINO DA REDE. PARA MONTAR A ÁRVORE, FORAM SELECIONADAS 40 PROTEÍNAS ALEATÓRIAS UTILIZADAS PARA TREINAR O CLASSIFICADOR, SENDO 20 PERTENCENTES A CLASSE 1 (“NIFO”) REPRESENTADAS EM VERDE, E OUTRAS 20 A CLASSE 0 (“NOT NIFO”), REPRESENTADAS EM VERMELHO. O ALINHAMENTO DAS PROTEÍNAS FOI REALIZADO COM A FERRAMENTA EMBL-EBI CLUSTAL OMEGA E A ÁRVORE, GERADA COM EMBL-EBI CLUSTALW2 – PHYLOGENY.
FONTE: A autora

4.3.3.2. CLASSIFICADOR DE NifH-*LIKE*

Para gerar o domínio conservado e região consenso da NifH, foram utilizadas

proteínas, retornadas da URL:
http://www.uniprot.org/uniprot/?query=reviewed%3Ayes+AND+%28name%3Afemo+or+name%3Anitrogenase%29+AND+gene_exact%3ANifH&format=xml

A região consenso gerada com essas proteínas foi:

MSRNSQGFLTTIRIIVTSATDQPSNFIHSLSNKRDENLRQIAFYGKGGIGKSTTSQNTVAALAEEMGQKIMIVGCDP
KADSTRLILHGKAQDTVLDLAAEEGSVEDLELEDVMKTGYGGFNPEQPSPSGGWVKCVESGGPEPGVGCAGRG
VITSINFLEENGAYDDEDDLDYVFYDVLGDVVCGGFAMPIREGKAQEIYIVCSGEMMAMYAANNICKGILKYAHRV
SGGVRLGGLICNSRKVDREQELIEALAKRLGTQMIHFVPRDNIVQHAELRRMTVIEYDPDSKQADEYRQLAKKIHN
NDMFGVIPTPITMDELEELLMEFGIMDDEDESIVGIAAKEAAAAATRAAAAAA

O Quadro 3 mostra o domínio conservado gerado para proteína NifH.

Fase	***Fitness***	***Hits (seqs)***	***Pattern***	**Expressão Regular**
Antes refinamento	20.3503	166(92)	G-x(2)-G-x-G-x(2)-T-x(2,3)-N	G.{2}G.G. {2}T.{2,3}N
Após refinamento	47.0221	93(92)	L-x(0,2)-G-x(0,1)-D-[PV]-x-[AC]-[DG]-[GS]-x(2)-[LM]-x-[ILMV]-x(4)-[AQ]-x-[DET]-[IV]-x(2)-[LV]-[ACTV]-[AST]	L.{0,2}G. {0,1}D[PV].[AC][DG] [GS].{2}[LM] .[ILMV].{4}[AQ]. [DET][IV].{2} [LV][ACTV][AST]

QUADRO 3 – DOMÍNIO CONSERVADO DA PROTEÍNA NifH GERADO COM A FERRAMENTA EXPASY PRATT E A EXPRESSÃO REGULAR CORRESPONDENTE.
FONTE: A autora

O Quadro 4 mostra o desempenho do classificador MLP para proteinas Nif.

Acertos	**Matriz confusão**
99.2198	127 0 0 0 0 0 0 127 0 0 0 0 0 1 131 0 0 0 0 0 0 123 3 0 0 0 0 2 132 0 0 0 0 0 0 123

QUADRO 4 – DESEMPENHO DO CLASSIFICADOR PARA AS PROTEÍNAS Nif.
FONTE: A autora

O número total de registros de treinamento é 1.794 e de teste, 769.

4.3.3.3. CLASSIFICADOR DE NifD-*LIKE*

Para gerar o domínio conservado e região consenso da NifD, foram utilizadas proteínas, retornadas da URL:
http://www.uniprot.org/uniprot/?query=reviewed%3Ayes+AND+%28name%3Afemo+or+name%3Anitrogenase%29+AND+gene_exact%3ANifD&format=xml

A região consenso gerada com essas proteínas foi:

MSFSMTDSDAEIPARKIPDNKELIQEVLKAYPEKAAKRRAKHLNVHDEGKPDCGCPQHRCEMKCVKSNIKSIPGV
MTVRGCAYAGSKGVVWGPIKDMIHISHGPVGCGYYSWGGRRNYYVGTTGVDSFVFPIENFNLTMQFTSDFQEKD
IVFGGDKKLKKIIDEIEELFPLNNGISIQSECPIGLIGDDIEAVARKKSKEIGGKPVVPVRCEGFRGVSQSLGHHIAND
AIRDWIFGKADPEKKDNKPKFEPTPYDVAIIGDYNIGGDAWSSRILLEEMGLRVIAQWSGDGTLNEMEWTPKAKLN
LIHCYRSMNYISRHMEEKYGIPWMEYNFFGPTKIAESLRKIAAYFDDPKIKEKAEKVIAKYQPQVDAVIAKYRPRLE
GKTVMLYVGGLGKRPRHVIPAYEDLGMEVVGTGYEFGHNDDYQRTTVIPTIKIDADSKNIPEITVTPDEQKYRVVIP
EDKVEELKKAGVPLSSYGGFQHYVKDGTLIYDDVNGYEFEEFVEKLKPDLVGSGIKEEKYIFQKMGVPFRQMHSW
DYSGPYHGYDGFAIFARADMDMAINNPVWKLLKAPWKKASSESKAKVDAAE

O Quadro 5 mostra o domínio conservado gerado para proteína NifD.

Fase	***Fitness***	***Hits (seqs)***	***Pattern***	**Expressão Regular**
Antes refinamento	66.7208	23(23)	N-x(4)-P-G-x(2)-T-x(3)-C-A-x-A-G-x(2)-G-V-x(5)-K-D-x(5)-H-x-P-x-G-C	N.{4}PG.{2}T.{3}CA.AG .{2}GV.{5}KD.{5}H.P. GC
Após refinamento	114.0485	23(23)	N-x(2)-[ST]-x-P-G-x-[ILM]-T-x(2)-[GP]-C-A-[FY]-A-G-[ACSV]-x-G-V-[IV]-x-[GS]-[AP]-[IV]-K-D-x-[AILV]-x-[IMV]-x-H-[GS]-P-[AIV]-G-C-[AGST]-x(2)-[AGST]-x-[AGS]-[EGQST]	N.{2}[ST].PG.[ILM]T.{2}[GP]CA[FY]AG [ACSV].GV[IV].[GS][AP] [IV]KD.[AILV].[IMV].H[GS]P[AIV]GC[AGST].{2}[AGST].[AGS] [EGQST]

QUADRO 5 – DOMÍNIO CONSERVADO DA PROTEÍNA NifD GERADO COM A FERRAMENTA EXPASY PRATT E A EXPRESSÃO REGULAR CORRESPONDENTE.
FONTE: A autora

4.3.3.4. CLASSIFICADOR DE NifK-*LIKE*

Para gerar o domínio conservado e região consenso da NifK, foram utilizadas proteínas, retornadas da URL:
http://www.uniprot.org/uniprot/?query=reviewed%3Ayes+AND+%28name%3Afemo+

or+name%3Anitrogenase%29+AND+gene_exact%3ANifK&format=xml

A região consenso gerada com essas proteínas foi:

MSMSPHHTSQNADKVLDHYDLFRQPEYQELFANKKKMFEKGPHPPEEVQRVAEWTKTWEYREKNFAREALTVN
PAKACQPLGAMFAALGFEGTLPFVHGSQGCVAYFRSHFNRHFKEPASCVSSSMTEDAAVFGGLNNMIDGLQNS
YALYKPKMIAVCTTCMAEVIGDDLNAFIKNAKEEGSVPQDFPVPNKIKIVFAHTPSFVGSHITGYDNMMKGILDHFW
EGKGGTNPKLERKPNGKINIIPGFDGYTVGNMREIKRMLSLMGVDYTILSDTSDVFDSPLRPSKADGEFRMYDGG
TPLEDMKDAINAKATISLQQYCTEKTAKFIKKHWQQPKVSLNYPIGVKGTDEFLMALSRISGKPIPEELEDERGRLV
DAMADISHAWLHGKKFAIYGDPDLCMGMARFLLELGAEPVHVLCGNGNKKWQEEMKAILAASPYGEDANVWPG
KDLWHLRSLLFTEPVDFMIGNSYGKYLQRDTLHKGKEFGIPLIRIGFPIFDRHHHHRYPVWGYQGALNLLNWIVNTI
LDEMDRNTNIMGKTDYSFDLVRAAL

O Quadro 6 mostra o domínio conservado gerado para proteína NifK.

Fase	***Fitness***	***Hits (seqs)***	***Pattern***	**Expressão Regular**
Antes refinamento	45.8706	17(17)	P-x(3)-G-S-Q-G-C-x(4)-R-x(4)-R-H-x(2)-E-P	P.{3}GSQGC.{4}R.{4}RH .{2}EP
Após refinamento	93.9813	17(17)	P-x-[SV]-x-G-S-Q-G-C-[CSTV]-[AST]-[FY]-x-R-x(2)-[FL]-[ANST]-R-H-[FY]-[KR]-E-P-x(3)-[ASV]-[STV]-[ADST]-S-[FLM]-x-E-x-[AP]-[AS]	P.[SV].GSQGC[CSTV][AST][FY].R.{2}[FL][ANST]RH[FY][KR]EP.{3}[ASV][STV] [ADST]S[FLM].E.[AP][AS]

QUADRO 6 – DOMÍNIO CONSERVADO DA PROTEÍNA NifK GERADO COM A FERRAMENTA EXPASY PRATT E A EXPRESSÃO REGULAR CORRESPONDENTE.
FONTE: A autora

3.3.3.5. CLASSIFICADOR DE NifE-*LIKE*

Para gerar o domínio conservado e região consenso da NifE, foram utilizadas proteínas, retornadas da URL:
http://www.uniprot.org/uniprot/?query=reviewed%3Ayes+AND+%28name%3Afemo+or+name%3Anitrogenase%29+AND+gene_exact%3ANifE&format=xml

A região consenso gerada com essas proteínas foi:

MKHNTDNCNRKMQDGNNDDDFDIEYQIPNSISLKAKIQEIFDEPGCEHNRSKDDDGRPKCCSQSLPPGATQGGC
AFDGARVVLMPITDAAHLVHGPIGCAGNSWDNRGSASSGPELYRTGFTTDLSEKDVVFGHGEKKLFKAIREINEA
YHPPAIFVYSTCVTALIGDDIDAVCKAAQEKFGTPVIPVNSPGFAGFSKNLGNKLAGDALLDHVIGTREPEDEHEGS

EFPPATPYDINLIGEFNIAGEFWQVKPLLDKLGIRVLACITGDARYAEIASAHRAKLNMMVCSKAMINLARKMEERY
GIPYFEGSFYGIDDTSESLRQIAEFLVKQGGDEDLLKRTEALIAEEEAKAWKALEPYRPRLKGKRVLLYTGGVFKS
WSIVSAFQDLGMEVVGTGTKKSTPEDKERIRELMGDDTIMFDDANPRELYQMLKEYKADIMISGGRNQYIALKAGI
PWCDINQERHHPYAGYQGMIEFAREIDQAIHSPIWEQVRKPAPWDCGPARQEQMSSSQPDHDSTSIAESFRRAR
NICVCNRVDLGTIEDAISVHGLRSVAAVREHTNAAGGCCQGRIEDMLMSEPDAHRFGER

O Quadro 7 mostra o domínio conservado gerado para proteína NifE.

Fase	***Fitness***	***Hits (seqs)***	***Pattern***	**Expressão Regular**
Antes refinamento	49.0406	14(14)	C-x(3)-G-A-x(3)-L-x-P-x(3,4)-A-x(0,1)-H-x(2)-H-x(4)-C-x(4)-W-x(2)-R-x(3)-S	C.{3}GA.{3}L.P.{3,4}A.{0,1}H.{2}H.{4}C.{4}W.{2}R.{3}S
Após refinamento	94.8287	14(14)	C-[ASV]-[FY]-[CDG]-G-A-x(2)-[ASTV]-L-x-P-x(3,4)-A-x(0,1)-H-[IL]-[IV]-H-[AG]-[PS]-[AILV]-[AG]-C-x-[AGS]-x-[GST]-W-[DGN]-x-R-[GS]-[AST]-x-S-[ST]	C[ASV][FY][CDG]GA.{2}[ASTV]L.P.{3,4}A.{0,1}H[IL][IV]H[AG][PS][AILV][AG]C.[AGS].[GST]W[DGN].R[GS][AST].S[ST]

QUADRO 7 – DOMÍNIO CONSERVADO DA PROTEÍNA NifE GERADO COM A FERRAMENTA EXPASY PRATT E A EXPRESSÃO REGULAR CORRESPONDENTE.
FONTE: A autora

4.3.3.6. CLASSIFICADOR DE NifN-*LIKE*

Para gerar o domínio conservado e região consenso da NifN, foram utilizadas proteínas, retornadas da URL: http://www.uniprot.org/uniprot/?query=reviewed%3Ayes+AND+%28name%3Afemo+or+name%3Anitrogenase%29+AND+gene_exact%3ANifN&format=xml

A região consenso gerada com essas proteínas foi:

MARILRPNKAAAVNPLKSSQPLGAALAFLGIEGAMPLFHGSQGCTAFALVLFVRHFREAIPLQTTAMDEVSTILGGA
DHIEQAILNIKKRAKPKIIGICSTGLTETRGDDIAGYLRDIRQKHAPELKGTPIVFVNTPDFDGAMQDGWAKAVEAMV
EQWVPRPQQAPRSRKATLIEAITRPGEQTRRPRQVNILPGCHLTPGDIEELRDMVESFGLRPIILPDLSGSLDGHLP
DGRWSPTTYGGTSIEEIRELGRSAYCIAIGREHMRGAAEILQDRTGVPYRVFQRLTGLEAVDRFIQLLSEISGNRCD
HHFPIVQPVPAKYRRQRAQLQDAMLDGHFHFRGGKKIAIAAEPDLLYQLATFLTSMGAQIVAAVTTTGQSPILEKIP
VEQVQIGDLEDLEDLGTKGAFARAAHADLLITHSHGRQAAERLGIPLYRVGFPIFDRGLGSGHRCTVGYRGTRDLI
FEIANIIQAHHHAPTPEQTDAWRKPEGFDHHVGHRNHSTGPPTAPG

O Quadro 8 mostra o domínio conservado gerado para proteína NifN.

Fase	*Fitness*	*Hits (seqs)*	*Pattern*	Expressão Regular
Antes refinamento	35.5305	10(10)	F-G-L-x-P-x(3)-P-D-x(1,3)-S-x(1,3)-D-G	FGL.P.{3}PD.{1,3}S.{1,3}DG
Após refinamento	131.9818	10(10)	P-[LV]-[FLMV]-H-G-[AS]-Q-G-C-[ST]-[AS]-F-[AG]-x-[TV]-x(2)-[IV]-[QR]-[DH]-F-[HKR]-[DE]-[APST]-[IV]-P-L-[AQS]-[ST]-T-A-M-[DNST]-[DEPQ]-x-[AST]-x(2)-[LM]-G-[AG]-x-[ADEG]-x-[ILV]-x(2)-A-[ILV]	P[LV][FLMV]HG[AS]QGC[ST][AS]F[AG].[TV].{2}[IV][QR][DH]F[HKR][DE][APST][IV]PL[AQS][ST]TAM[DNST][DEPQ].[AST].{2}[LM]G[AG].[ADEG].[ILV].{2}A[ILV]

QUADRO 8 – DOMÍNIO CONSERVADO DA PROTEÍNA NifN GERADO COM A FERRAMENTA EXPASY PRATT E A EXPRESSÃO REGULAR CORRESPONDENTE.
FONTE: A autora

4.3.3.7. CLASSIFICADOR DE NifB-*LIKE*

Para gerar o domínio conservado e região consenso da NifB, foram utilizadas proteínas, retornadas da URL:
http://www.uniprot.org/uniprot/?query=reviewed%3Ayes+AND+%28name%3Afemo+or+name%3Anitrogenase%29+AND+gene_exact%3ANifB&format=xml

A região consenso gerada com essas proteínas foi:

MTSCMIFGIQDIREPPTTGSSVTEHTQCAASSGGCGSSCSSSSEPSDMDPEIWEKIKNHPCYSEEAHHHFARMHV
AVAPACNIQCNYCNRKYDCANESRPGVVSEKLTPEQAVRKVIAVANEIPQMSVLGIAGPGDPCANWKKTFRTFELI
AEQIPDIKLCLSTNGLALPDHVDRLADMNVDHVTITINMVDPEIGAKIYPWIFYNHRRYTGVEAARILHERQMEGLE
MLTERGILCKVNSVMIPGINDEHLVEVNKMVKERGAFLHNIMPLISAPEHGTHFGLTGQRGPSAQELKALQDRCEQ
DDSGNMNMMRHCRQCRADAVGLLGEDRGQEFTLDKIPEMEPEYDAAKRQAYHEHVERERDHKAAKEKAQIAT
GGRSPAAASAESAASNPSILVAVATKGGGRINQHFGHAKEFQVYEVSQSGVRFVGHRRVDQYCQGGWGCDPQ
EEEATLDNIIRALKDCDAVLCAKIGDCPKEKLMQAGIQPVDAYAHDYIEKAVMAFYRQWLGSEPPAEIHHLPRGDP
PRWPGDYISVQSTQATA

O Quadro 9 mostra o domínio conservado gerado para proteína NifB.

Fase	*Fitness*	*Hits (seqs)*	*Pattern*	Expressão Regular
Antes refinamento	120.9315	13(13)	H-x(3)-A-R-M-H-x(2)-V-A-x(2)-C-N-x-Q-C-x(2)-C-N-R-K-x-D-C-x-N-E-S-	H.{3}ARMH.{2}VA.{2}CN.QC.{2}CNRK.DC.NESRPGV.S.{2}LTP.{2}A

			R-P-G-V-x-S-x(2)-L-T-P-x(2)-A	
Após refinamento	159.6465	13(13)	H-x(2)-[FY]-A-R-M-H-[LV]-[APS]-V-A-[PS]-[AG]-C-N-[IL]-Q-C-x-[FY]-C-N-R-K-[FY]-D-C-[AST]-N-E-S-R-P-G-V-x(0,1)-S-x(2)-L-T-P-[DE]-[DEQ]-A-[ALV]-x-[KR]	H.{2}[FY]ARMH[LV] [APS]VA[PS][AG] CN[IL]QC.[FY]CNRK [FY]DC[AST] NESRPGV.{0,1}S.{2}LTP[DE] [DEQ]A[ALV].[KR]

QUADRO 9 – DOMÍNIO CONSERVADO DA PROTEÍNA NifB GERADO COM A FERRAMENTA EXPASY PRATT E A EXPRESSÃO REGULAR CORRESPONDENTE.
FONTE: A autora

4.3.3.8. CLASSIFICADOR DE DraT-*LIKE*

Para gerar o domínio conservado e região consenso da DraT, foram utilizadas proteínas, retornadas da URL: http://www.uniprot.org/uniprot/?query=name%3A%22dinitrogen+reductase%22+%22adp+d+ribosyltransferase%22+gene_exact%3Adrat&format=xml

A região consenso gerada com essas proteínas foi:

MSDASDDSGRRPARAPELDPTTPRAARWILCAGYNARHFPGPAGGGIMAPMNDAGQRPRRGIGHSTNLCGLPP
WILASRHFNDHPQPLHIQGVREMNPSLFEMLDQAPDAEEAGEVFQDYMSAMFGLDPEQQQAHGASAPGARRRF
RSSYLRLLRGWGFDSNGPEGAVLKGWVESRFGLVPTFHKEPIGRIHSPAWMTYVEERMSGRFHNNAIWSQLDLL
YEFCQWELRRFWPDAPMFPGQRHLTLYRGVNDFDEHQIVERLGKGKGRLREAVIRLNNLVSFSSDRDRAGCFG
DTILEVRVPLSKILFFNDLLPSHPLKGEGEYLVIGGEYRVRMVMCSYL

O Quadro 10 mostra o domínio conservado gerado para proteína DraT.

Fase	***Fitness***	***Hits (seqs)***	***Pattern***	**Expressão Regular**
Antes refinamento	16.1802	29(29)	A-x(2)-K-x(5)-R-x(0,1)-G	A.{2}K.{5}R.{0,1}
Após refinamento	52.7477	29(29)	L-x(4)-G-x(3,4)-D-[AEST]-[EGNQ]-[AGNSV]-x(2)-[AGL]-[AV]-[GIV]-[LMV]-x-[AGI]-x-[AV]-[AEQ]-[GS]-x-[FIMV]-[GV]-[GIL]-x(4)-[DH]	L.{4}G.{3,4}D[AEST] [EGNQ][AGNSV].{2} [AGL][AV][GIV][LMV].[AGI]. [AV][AEQ][GS].[FIMV] [GV][GIL].{4}[DH]

QUADRO 10 – DOMÍNIO CONSERVADO DA PROTEÍNA DRAT GERADO COM A FERRAMENTA EXPASY PRATT E A EXPRESSÃO REGULAR CORRESPONDENTE.
FONTE: A autora

O Quadro 11 mostra o desempenho do classificador MLP para proteinas DraT e DraG.

Acertos	Matriz confusão
97.4138	61 0 3 52

QUADRO 11 – DESEMPENHO DO CLASSIFICADOR PARA AS PROTEÍNAS DraT E DraG.
FONTE: A autora

O número total de registros de treinamento é 268 e de teste, 116.

4.3.3.9. CLASSIFICADOR DE DraG-*LIKE*

Para gerar o domínio conservado e região consenso da DraG, foram utilizadas proteínas, retornadas da URL:
http://www.uniprot.org/uniprot/?query=name%3AADP-ribosyl-%5Bdinitrogen+reductase%5D+glycohydrolase+AND+gene_exact%3Adrag&format=xml

A região consenso gerada com essas proteínas foi:

MDMDLMNRMNALDKYVFKYRRYLPISNPSVQRRMQYTYDNKPLESPSNSDVQKQWDLTPSCNISLAESRDRAL
GALLGLAVGDALGTTVEFMPRDEIKARYGYLRDMTGGGWFRNAACYLKPGEWTDDTSMALCLAESLLEKGGED
DICDRNRICEWFAHWYNSSPGICFDIGNTCRRALERYMTGGSMWAGNNHSAHMKDPQDAGNGAIMRMAPVALF
FYNDPDKMEKFHYAKEQSRITHGHPESIDACLMFARMLMHLINGSNKQEAFKPAKELFDQYYDIKGFNFSASLSA
RLELPQEIPRFMRINEHEYFQFDEYQIRSSGYVVDTLEAAMWCFWNTDNFRDAILQAVNLGDDADTVGAIAGQLA
GAYYGYDGIPQEWLDKLDKKERIETMAQALYLLAPAQKMEEQNEGGDAGNGEDTMS

O Quadro 12 mostra o domínio conservado gerado para proteína DraG.

Fase	*Fitness*	*Hits (seqs)*	*Pattern*	Expressão Regular
Antes refinamento	15.1802	24(18)	A-x(3)-A-x(0,2)-L-x(0,1)-G	A.{3}A.{0,2}L.{0,1}G
Após refinamento	22.7123	18(18)	G-x(3)-G-x(0,1)-D-[AL]-[ILMV]-[GP]-x-[AGNPT]	G.{3}G.{0,1}D[AL][ILMV][GP].[AGNPT]

QUADRO 12 – DOMÍNIO CONSERVADO DA PROTEÍNA DRAG GERADO COM A FERRAMENTA

EXPASY PRATT E A EXPRESSÃO REGULAR CORRESPONDENTE.
FONTE: A autora