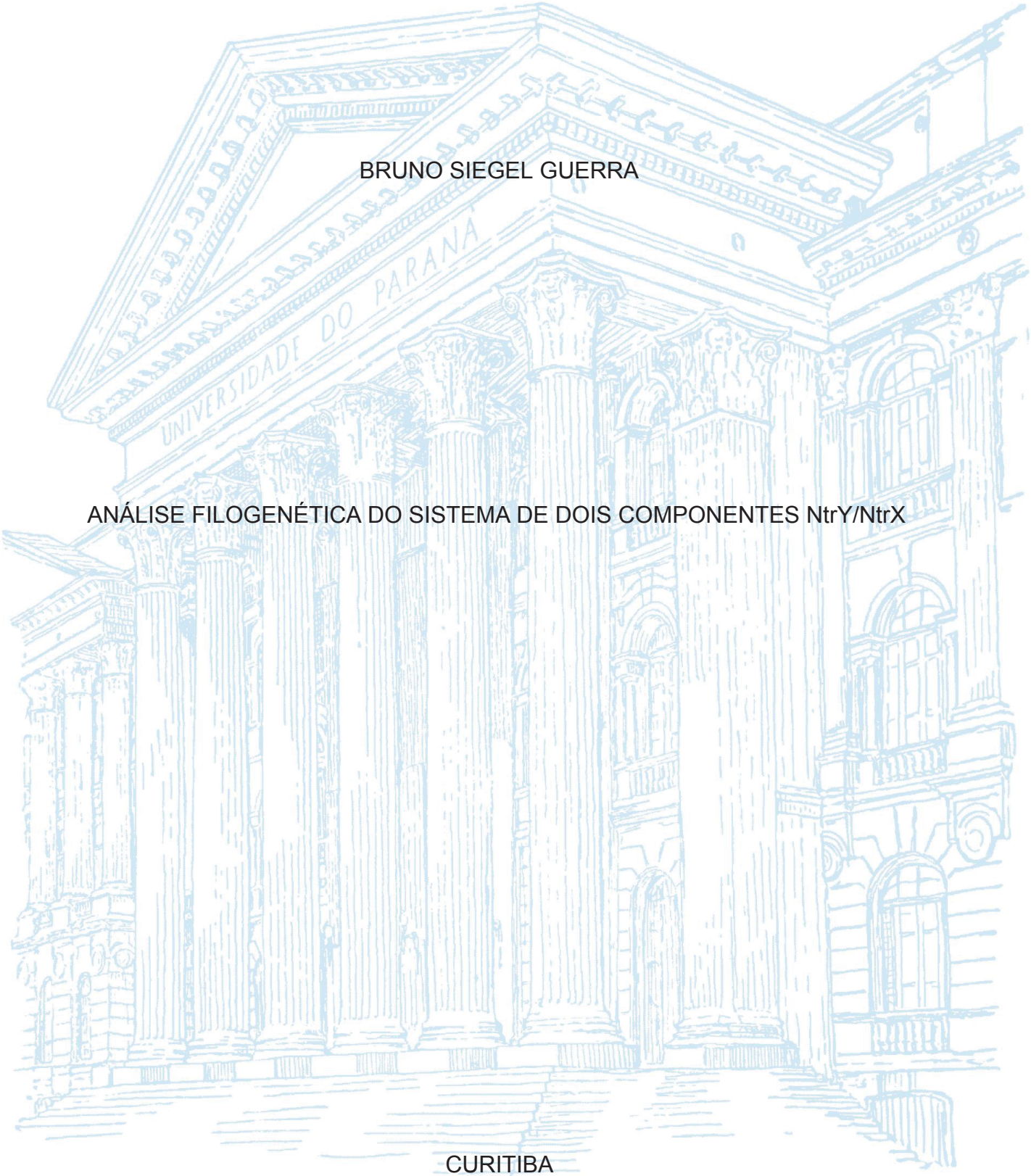
UNIVERSIDADE FEDERAL DO PARANÁ

BRUNO SIEGEL GUERRA



ANÁLISE FILOGENÉTICA DO SISTEMA DE DOIS COMPONENTES NtrY/NtrX

CURITIBA

2018

BRUNO SIEGEL GUERRA

ANÁLISE FILOGENÉTICA DO SISTEMA DE DOIS COMPONENTES NtrY/NtrX

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, área de concentração Bioinformática.

Orientador: Prof. Dr. Emanuel Maltempi de Souza

Co-orientador: Prof. Dr. Leonardo Magalhaes de Souza Cruz

Co-orientadora: Profa. Dra. Leda Chubatsu

CURITIBA

2018

Catalogação na publicação
Sistema de Bibliotecas UFPR
Biblioteca de Educação Profissional e Tecnológica

G934 Guerra, Bruno Siegel
Análise filogenética do sistema de dois componentes NtrY/NtrX / Bruno Siegel Guerra. - Curitiba, 2018.
62 p.: il., tabs, grafs.

Orientador: Emanuel Maltempi de Souza
Co-orientador: Leonardo Magalhaes de Souza Cruz
Co-orientador: Leda Chubatsu
Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Curso de Pós-Graduação em Bioinformática.

1. Filogenia. 2. Genes. 3. Sistema de dois componentes. 4. Bioinformática. I. Cruz, Leonardo Magalhaes de Souza. II. Chubatsu, Leda. III. Título. IV. Universidade Federal do Paraná.

CDD 576.88

Elaboração: Angela Pereira de Farias Mengatto - CRB 9/1002

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA

Pós-Graduação em Bioinformática WWW.BIOINFO.UFPR.BR
E-mail: bioinfo@ufpr.br **Tel: 41 33614906**

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em BIOINFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de BRUNO SIEGEL GUERRA intitulada: **"Analise filogenética do sistema de dois componentes NtrY/NtrX em bactérias"**, após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

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

Curitiba, 29 de março de 2018.

Dr. Emanuel Maltempi de Souza
Presidente
Programa de Pós-graduação em Bioinformática – UFPR
Departamento de Bioquímica - UFPR

Drª. Roseli Wassen
Avaliadora Externa
Departamento de Genética - UFPR

Dr. Dieval Guizelini
Avaliador Interno
Programa de Pós-graduação em Bioinformática - UFPR

Dedico esse trabalho a todas as pessoas que acreditaram que era possível.

AGRADECIMENTOS

Agradeço a Deus por me proporcionar o livre-arbítrio de minhas escolhas e sempre ser meu local de segurança.

Agradeço primeiramente a minha família por todo suporte e apoio durante esse projeto, por acreditarem em mim e por servirem de inspiração em minha vida. Se consegui chegar longe, foi pela solida estrutura que me sustenta.

Agradeço a minha amada pela paciência e palavras de conforto e encorajamento nos momentos mais difíceis. Por me ouvir contar sobre temas que parecem chatos, mas ouvir com toda a atenção e carinho do mundo mesmo não entendendo tudo.

Agradeço ao meu orientador Dr. Emanuel Maltempi de Souza por confiar a mim esse trabalho, espero ter atendido a suas expectativas. Aos meus coorientadores, Dr. Leonardo Magalhaes Cruz e Dra. Leda Chubatsu, por sempre me receberem com atenção e dedicação e me guiarem com a luz de seu conhecimento.

Ao programa de Pós-graduação em Bioinformática da UFPR, assim como o departamento de Bioquímica e Biologia Molecular por tornarem minha segunda casa. Obrigado especial aos professores do PPG-Bioinformática por sempre que precisei eles estavam dispostos a ajudar com seu vasto conhecimento. Obrigado a secretaria do programa, Suzana, por sempre receber as nossas duvidas e respondê-las com o maior carinho do mundo. Aos professores Dieval, Mauro e Roberto por ministrarem excelentes aulas e conversas no programa.

Aos meus queridos amigos, obrigado por todo o apoio e companheirismo de todas as horas. Por me explicarem assuntos complicado e ao mesmo tempo compartilhar tantas risadas.

Obrigado ao meu notebook por não ter desistido em nenhum momento, mesmo com tantas analises, programas e horas de funcionamento. Você foi a ferramenta perfeita para esse trabalho.

“*Intelligence is the ability to adapt to change.*”

Stephen Hawking

RESUMO

As bactérias conseguem viver e se adaptar em diferentes tipos de ambientes; essa característica é essencial para o sucesso da espécie. O mecanismo capaz de perceber e responder os mais diferentes estímulos do meio é o sistema regulador de dois componentes. Esse sistema é composto por duas proteínas, uma HK- histidina kinase que recebe o estimulo do meio e gera uma resposta celular numa segunda proteína chamada de RR- Resposta reguladora. Um exemplo de sistema de dois componentes é o NtrYX, que está amplamente distribuído nas Proteobacterias e que executa as mais variáveis funções. Ao analisarmos os domínios proteicos das proteínas de RR NtrX, ficam evidentes as diferenças entre os táxons: Alpha-Proteobacterias apresentam o domínio AAA+, porém sem o motivo GAFTGA; Já as Beta-Proteobacterias não apresentam o domínio AAA^{+} como a *Herbaspirillum seropedicae* ou esse domínio já não tem o mesmo formato que a das Alpha, como na *Neisseria gonorrhoeae.* Neste trabalho, investigamos com mais detalhes essas diferenças entre os grupos; capturamos o maior número de homólogos possível para o sistema NtrYX; e as relações filogenéticas presentes nesses organismos, a fim de melhor entender o processo de evolução desse sistema. A proteína HK NtrY também foi analisada, para melhor compreender se existem diferenças relevantes em sua estrutura quando comparado entre os organismos.

Palavras-chave: Bioinformática. Sistema de dois componentes. NtrY/X. Evolução. Filogenia.

ABSTRACT

The bacteria can live and adapt to a wide range of ecosystems; this characteristic is essential to the success of the species. The mechanism capable of sensing and respond to the variety of stimulus is the Two-component regular system. Two proteins compose this system, one HK-histidine kinase that sense of chance in the extracellular ambience and generates one intracellular response in a second protein RR-Response regulator. One example of two-component regular system is the two-component NtrY/X, which is vastly distributed between the Proteobacteria and execute several functions. When we analyze with more attention the domain in the protein RR NtrX, its evident the differences within the taxon: Alpha-Proteobacteria have the domain AAA^+ without the GAFTGA motif, essential to sigma-54 interaction. However, for the Beta-Proteobacteria like *Herbaspirillum seropedicae* this domain is completely absent or as for the *Neisseria gonorrhoeae* this AAA^+ doesn't have the same format as for the Alpha. In this work, we investigate with details these differences; we search for the number of homologous sequences for the NtrYX system; and the phylogenetic relationships in these so variable group. The NtrY HK protein was also analyzed, to better understand if they present noticeable differences in the protein domain structure when compared to other bacteria.

Key words: Bioinformatics, Two-component regular system, NtrYX, evolution, phylogenetics.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

HK – Histidina Kinase

RR – Resposta reguladora

SUMÁRIO

1 INTRODUÇÃO

As bactérias conseguem viver e se adaptar em diferentes tipos de ambientes, alguns desses apresentam condições que nem sempre são as ideais para a subsistência das mesmas. Ausência de nutrientes, variações de pH, disponibilidade de oxigênio são alguns exemplos de problemas que podem afetar a sua sobrevivência. A fim de garantir o sucesso biológico nesses meios, as bactérias desenvolveram mecanismos capazes de reagir, responder e se adaptar a essas mudanças. Um desses mecanismos é o sistema regulador de dois componentes, que são capazes de perceber os estímulos do meio e gerar respostas adaptativas (Labes & Finan, 1993; Ninfa & Magasanik, 1986).

O sistema regulador de dois componentes é constituído por um par de reação entre o sensor de histidina kinase (HK) e resposta reguladora (RR) para o estimulo recebido (Stock & Da Re, 2000). Esse sistema está amplamente distribuído dentro do filo das Proteobacterias, porém não apresentam a mesma origem genética e realizam as mais diferentes funções dentro do ecossistema. Um exemplo bem documentado é o sistema NtrB/C, HK do gene NtrB e o RR do gene NtrC, que é responsável pelo controle do metabolismo de nitrogênio em bactérias diazotrópicas (Dixon & Kahn, 2004). Outro sistema conhecido é o componente regulador NtrY/NtrX, que em Beta-Proteobacterias como a *Herbaspirillum seropedicae* também está envolvido no controle da assimilação de nitrato pelo organismo (Bonato *et al*., 2016), origem de uma provável duplicação e diferenciação genica do complexo NtrB/C (Capra *et. al*, 2012). Contudo, apesar desse sistema ter uma possível relação funcional com NtrB/C, ele não apresenta esse sistema em sua vizinhança genética. Já em Alpha-Proteobacterias, como a *Brucella abortus*, esse sistema apresenta relação de vizinhança com complexo NtrB/C e é responsável pela sobrevivência da espécie em ambientes com condições anaeróbicas (Carrica *et al.,* 2012).

O estudo e compreensão das diferenças evolutivas e estruturais do sistema de dois componentes é de extrema importância para melhor predizer sua função e importância biológica. A Bioinformática é uma área de conhecimento capaz de auxiliar de forma decisiva nessa questão. A capacidade de estudar com detalhes os genes responsáveis por essas proteínas, as suas diferenças em seus domínios proteicos e inferir homologia entre as espécies, são poderosas ferramentas ao nosso dispor. Em 2012, CAPRA apresentou os diferentes domínios comuns para os sistemas HK e RR descobertos, essas diferenças foram essenciais para a variabilidade do sistema na natureza. O sistema NtrY/X, mais

especificamente o componente de resposta reguladora NtrX, apresenta diferenças significativas em sua estrutura quando analisamos Alpha e Beta Proteobacterias (Capra & Laub, 2012), sua função ainda é amplamente desconhecida em alguns organismos e sua origem molecular evolutiva também é um mistério.

Este trabalho tem como objetivo compreender o sistema de dois componentes NtrY/X em Proteobacterias, visando melhor descrever as relações evolutivas e funcionais dos grupos, dando enfoque nas espécies hoje já bem descritas pela literatura como a *Herbaspirillum seropedicae*.

Os objetivos específicos são:

- Identificar genes NtrY e NtrX em genomas bacterianos completos depositados em bancos de dados públicos;
- Determinar a composição de domínios funcionais nas proteínas NtrX e NtrY, e relacionar com a taxonomia;
- Determinar a filogenia das proteínas NtrY e NtrX identificadas;

Os resultados desse trabalho originaram o artigo cientifico “NtrY/NtrX Two-Component Regulator System Phylogenetic Study and Insights”, apresentado na seção 3 desse documento.

2 REVISÃO DE LITERATURA

2.1 Evolução dos genes

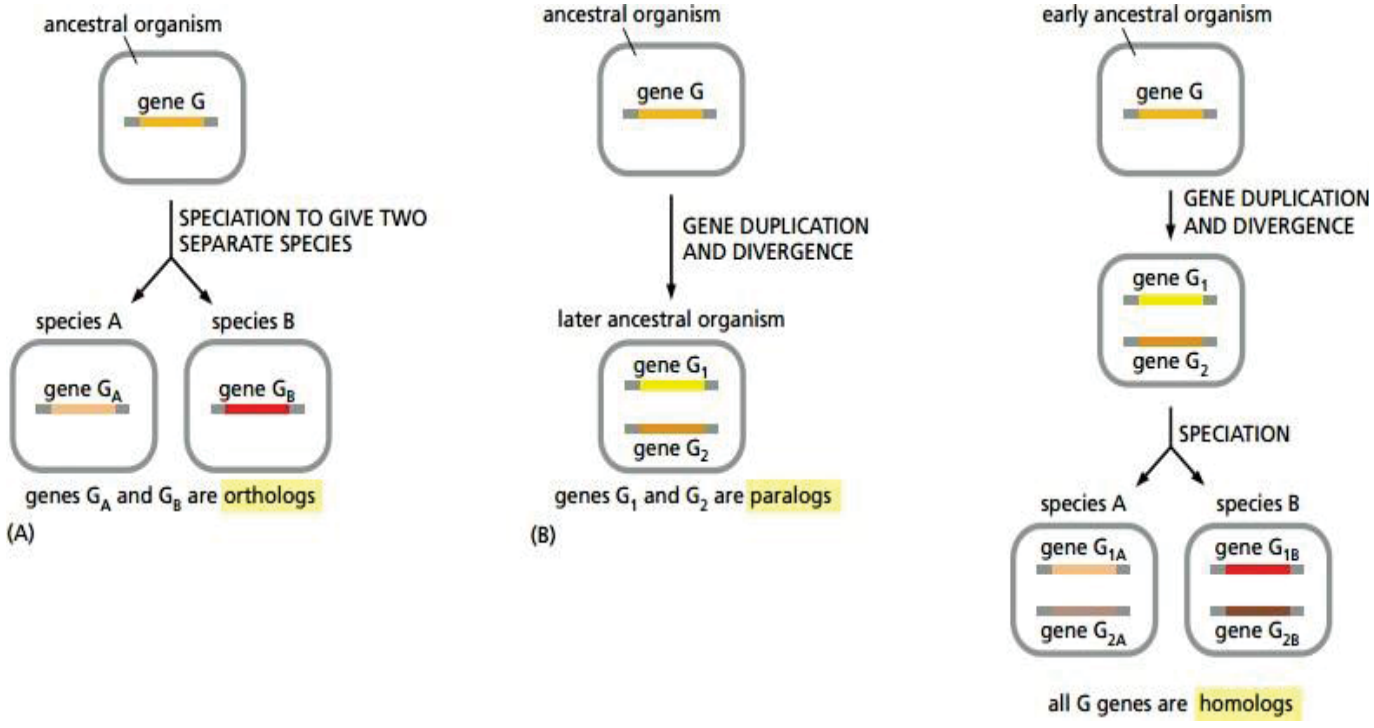
Na natureza, a informação necessária para sobrevivência dos seres vivos está contida em seu material genético, organizada em genes, que é passado de geração em geração a fim de garantir a sobrevivência e a estabilidade daquela espécie. Contudo, na manutenção e cópia da informação genética, ocorrem erros aleatórios, alterando as sequencias genéticas e criando mutações. Em alguns casos, essas mudanças representam uma melhoria na vida desse organismo, tornando assim mais fácil a sobrevivência e reprodução do mesmo, deste modo, através de intermináveis ciclos de mutações e seleções naturais os organismos evoluem e propagam os genes que estão melhores adaptados para aquele ambiente (ALBERTS, 2011). Já as mutações em alguns pontos do código genético prejudicam o organismo, impossibilitando a sobrevivência ou a reprodução deste, dessa forma este gene modificado não se propaga na natureza.

Não existe mecanismo natural capaz de fabricar novas sequencias de DNA de forma completamente randômica, ou seja, a evolução molecular se dá a partir de sequencias de DNA já existente que sofrem diferentes tipos de mutações, como por exemplo:

- Mutações intragênicas – Genes existentes que são modificados por mudanças em suas sequencias de DNA;
- Duplicação gênica – Genes que são duplicados, criando um par de genes inicial que pode futuramente divergir ao longo da evolução;
- Embaralhamento de segmentos – Dois ou mais genes que são quebrados e religados para formar genes híbridos contendo segmentos de DNA originalmente de genes separados;
- Transferência horizontal – DNA que pode ser transferida do genoma de uma célula para outra.

Quando uma região é duplicada, uma dessas cópias estará livre para se diferenciar, através de mutações, e se especializar em funções diferentes da cópia original dentro de uma mesma célula. A divergência genica, repetidas diversas vezes, resultam em família de genes relacionados, seja dentro de uma única célula, ou em espécies diferentes. (ALBERTS, 2011). Genes que são de duas espécies diferentes, mas são derivados do mesmo gene ancestral comum são chamados de ortólogos (Fig-1 A). Já genes que são originados por duplicação genica, mas que apresentam diferente função dentro de um mesmo organismo, são conhecidos como parálogos (Fig-1 B). E os genes que estão relacionados por descendência de qualquer uma das duas maneiras são conhecidos como homólogos. (Fig – 1 C).

Figura 1: O que são genes homólogos



FONTE: Adaptado de ALBERTS (2011).

Nota: A) Genes ortólogos, que são originados do mesmo ancestral comum entre as espécies diferentes. B) Genes parálogos, que são gerados por duplicação e divergência. C) Genes homólogos, que são relacionados evolutivamente.

2.2 Evolução das proteínas

Ao analisar a evolução das proteínas, percebemos diferenças em suas estruturas, funções e domínios. A evolução das proteínas ocorre para cada espécie separadamente, porém proteínas de uma mesma função biológica tendem a ter uma evolução conservada, preservando a função da mesma, mesmo que seja em diferentes espécies. (Fay *et al*. 2013). Regiões que apresentam funções essenciais para o funcionamento das proteínas tendem a evoluir numa taxa mais lenta conservando esses pontos importantes e mantendo um "modelo neutro" para aquela determinada molécula independente da espécie. Portanto o polimorfismo em proteínas chega a ser praticamente constante em várias espécies (AKASHI, 2012).

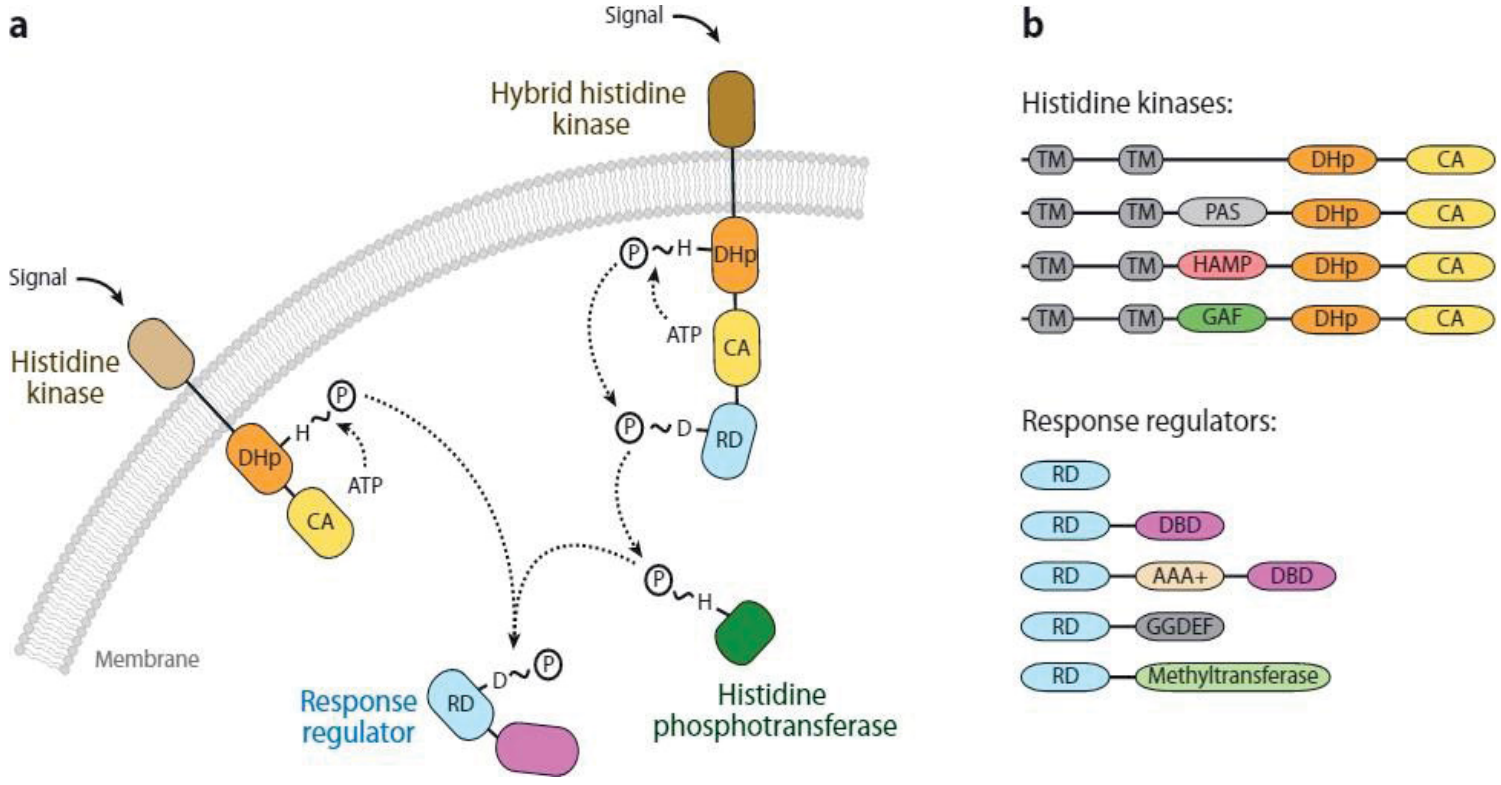
Genes que compartilham uma alta similaridade, identidade e conservação entre suas sequencias de DNA, RNA ou Aminoácidos podem indicar de homologia entre as espécies ou sequencias estudadas. Se esses genes ocorrem numa mesma espécie, os genes e os produtos proteicos são considerados parálogos, contudo, apesar de terem a mesma sequência e estrutura tridimensional ambas apresentam funções diferentes. Já para as proteínas ortólogas que apresentam a mesma função em organismo diferentes, é possível através de analises de genoma predizer qual será o produto oriundo de determinada sequência assim como os domínios e motivos presentes (LEHNINGER, 2011).

Domínios proteicos são regiões compactas e conservadas responsáveis por uma função particular dentro de uma proteína. Domínios semelhantes podem ser encontrados em proteínas com funções diferentes e em diferentes espécies. Por este motivo possibilitam analises para conseguir inferir a relação evolutiva de proteínas e genes assim como identificar um ancestral comum. Da mesma forma que os motivos proteicos auxiliam e muito com essa função. Um motivo é uma pequena região bastante semelhante em alinhamentos de sequencias que representam uma determinada função ou região já conhecida.

2.3 O sistema regulador de dois componentes

O sistema regulador de dois componentes é composto por dois membros conservados: a histidina kinase (HK) que é responsável por perceber os estímulos do meio ambiente como: Luz, oxigênio e pH, através do seu domínio PAS, e ativar a proteína reguladora de resposta (RR) correspondente, que regula a expressão genica na dos alvos bactéria (MANSCHER, 2006.). Ao detectar uma mudança no meio extracelular o domínio CA da HK (Fig.1A) se liga a um ATP e auto-fosforila uma histidina conservada no domínio DHp. O grupo fosforil é então transferido para a região RD do componente de resposta (RR) que catalisa a transferência do grupo fosforil para um resíduo de aspartado presente no composto (SANDERS, 1989; 1992). Com isso inicia um processo de mudança conformacional que ativa a proteína e gera a resposta celular para o sinal recebido pelo HK, normalmente estimulando ou reduzindo a expressão de determinado gene. (CAPRA, 2012).

Figura 2: Funcionamento molecular do sistema de dois componentes



FONTE: CAPRA (2012).

Nota: a) O funcionamento sistema regulador de dois componentes. b) Domínios comuns encontrados por CAPRA ,2012.

2.4 O sistema regulador de dois componentes NtrY/NtrX

Um exemplo de sistema regulador de dois componentes é o sistema NtrY/X onde o componente NtrY é um sensor HK e o NtrX o RR, que foi caracterizado em *Azorhizobium caulinodans* e *Azospirillum brasilense* e estava relacionado com o metabolismo de nitrogênio e nitrato, respectivamente. (Pawlowski *et al*., 1991; Ishida *et al*., 2002). O sistema também foi caracterizado em bactérias patogênicas como a *Brucella* spp. como sendo o mecanismo de sobrevivência necessário para ambientes com pouco oxigênio pois em meios onde a taxa de oxigênio disponível era baixa, houve um aumento na expressão do gene NtrX para essa espécie (CARRICA, 2012).

Nos domínios comuns propostos por CAPRA em 2012, o complexo NtrY é considerado como aquele contendo as seguintes estruturas independente de taxonomia: HAMP family (PF00672), PAS (PF13188), *HisKA* (PF00512) e HATPase_c (PF02518). Já o

NtrX apresenta importantes variações em sua arquitetura quando comparado entre os diferentes filos:

- NtrX com dois domínios: no caso da Beta-Proteobacteria *H. seropedicae* essa proteína apresenta apenas dois domínios, REC response regulator (PF00072) e HTH_8 (PF02954)
- NtrX com três domínios: já para as Alpha-Proteobacteria elas apresentam uma estrutura com três domínios REC response regulator (PF00072), HTH_8 (PF02954) e o domínio AAA^+ (PF00158).
- NtrX com três domínios: Existe uma segunda arquitetura possível para o complexo, como a da Beta-Proteobacteria *Neisseria gonorrhoeae*, que apresenta os mesmos três domínios porém com uma diferença significativa em seu domínio AAA^+ (PF14532)

As diferenças nos domínios AAA^+ entre as espécies, especialmente no motivo GAFTGA, apresentam evidencias interessantes sobre a evolução desse sistema.

2.5 Estado da arte do sistema NtrY/NtrX

Com a evolução de técnicas da bioinformática para realizar analises filogenéticas e identificar relações evolutivas, fica muito clara a necessidade de uma classificação mais precisa para o complexo NtrY/X. Nos últimos anos, trabalhos veem mostrando as diferenças entre os genes NtrY/X principalmente estrutural e como estão organizados nas bactérias:

- Entre 2006 e 2012 os autores: Alves L.; Osaki J.; Guimarães L.F.S e Bonato P; mostram que o sistema NtrY/X em *H. seropedicae* não está relacionada a fixação de nitrogênio, e sim com a assimilação de nitrato e que no sistema NtrX não existe a presença do motivo GAFTGA nem do domínio AAA^{+} que é essencial para interação com a proteína sigma54.

- Caprica *et. al*, 2012, mostraram que em *Brucella* spp. o sistema NtrY/X está relacionado ao monitoramento do estado redox, conectando o sinal de baixo oxigênio com aceptores alternativos de elétrons como o nitrato. E o sistema NtrX apresenta o domínio AAA+, porém sem o motivo GAFTGA.

- Atack *et. Al*, 2013, propuseram uma análise comparativa entre os genes NtrX de *N. gonorrhoeae* e *R. capsulatus* com todos os genomas bacterianos conhecidos e uma análise filogenética foi realizada. Os resultados mostraram que o gene NtrX pode ser classificado em quatro grupos: NtrX1, NtrX2, NtrX3 e NtrX4. Evidenciando uma a diferença estrutural e filogenética do componente.

2.6 A bioinformática

A bioinformática tem se consolidado como uma poderosa aliada nos estudos de genética e proteínas. Conseguir inferir a relação evolutiva, definir a função e analisar grandes quantidades de dados, tem ajudado a progredir a biologia molecular que conhecemos hoje. A procura por genes ou proteínas homologas em uma coleção de sequencias conhecidas se tornou uma tarefa possível graças a ferramentas de alinhamentos múltiplos como o BLAST e o HMMER que realiza alinhamentos de sequencias, e aos bancos de dados que armazenam as informações necessárias para as anotações genômicas (ATTWOOD, 2011).

Neste projeto foram realizadas três grandes etapas que envolviam conceitos da bioinformática: Coleta de dados disponibilizada em bancos de dados públicos; Controle de qualidade da informação; e finalmente a análise dos dados seguindo metodologias bem estabelecidas na literatura. Em todo o desenvolvimento do trabalho, ferramentas de programação como Python e Perl foram amplamente utilizadas a fim de auxiliar no grande volume de dados encontrado. Programas especializados em analises de filogenia, clusterização e estatística foram essenciais para alcançar os resultados obtidos e confirmar as hipóteses esperadas.

2.6.1 Banco de dados

Novas tecnologias para geração de dados biológicos têm revolucionado a ciência nos últimos anos. Com técnicas de sequenciamento cada vez mais acessíveis e rápidas, o volume de dados gerados tem aumentado de forma exponencial. É esperado que até 2025 o big-data formado pelas ciências genômicas ultrapassará o volume de informação gerado pela astronomia, YouTube e Twitter (Stephens *et al.*, 2015).

Bancos de dados são importantes para que toda essa informação não se perca e que seu acesso seja rápido e preciso. Bancos de dados especializados em proteínas trazem informações já anotadas que descrevem famílias de proteínas, domínios proteicos, sequencias conhecidas e regiões funcionais das sequencias. Utilizando de mecanismos de buscas como alinhamentos múltiplos de sequencias (MSA) ou perfis ocultos de Markov (HMM) os bancos são capazes de recuperar informações de forma precisa e detalhada. Alguns exemplos de bancos amplamente utilizados na bioinformática e nesse projeto são:

- PFAM - Banco de dados público de proteínas que contempla informações curadas de domínios e famílias proteicas. Utiliza metodologias de HMM para estabelecimento e busca de informações (Finn *et al.*, 2014).
- UniProt - Banco de dados público formado por diversos colaboradores como: EBI (European Bioinformatics Institute) e SIB (Swiss Institute of Bioinformatics). Apresenta informações detalhadas sobre as funções biológicas das proteínas e apresenta várias etapas de curadoria e classificação (Bateman *et al*., 2015).
- STRING – Banco de dados que contém informações de interações proteína-proteína a fim de predizer suas relações funcionais e estruturais, assim como a rede biológica em que está inserido (Szklarczyk *et al*., 2015).

2.6.2 Clusterização

Estratégias de clusterização, ou seja, agrupamento automático de sequencias com algum nível de similaridade, são essenciais para analisar, organizar e classificar o grande volume de dados gerados pelas novas tecnologias de sequenciamento. Essa metodologia é aplicada em banco de dados como o UniRef, que disponibiliza diversos tipos de banco baseado na similaridade entre as sequencias lá presentes. Ferramentas como o CD-HIT são utilizadas em cenários onde encontramos grandes quantidades de sequencias, a fim de minimizar efeitos de redundância e facilitar a identificação de padrões (Fu *et al*., 2012).

2.6.3 Mineração de dados

Mineração de dados é o processo automático para recuperação de informações contidas em grandes volumes de dados. Os métodos das tecnologias de data-mining se tornaram aliados importantes para as pesquisas na área de Bioinformática, uma vez que esta apresenta a cada ano um volume maior de informação.

O primeiro passo para uma análise filogenética precisa é a inferência da homologia entre as sequencias de aminoácidos ou nucleotídeos dos organismos de interesse. A procura por essas sequencias pode ser feita atrás da similaridade entre a sequência escolhida e o banco de dados usado no estudo. Ferramentas capazes de calcular a similaridade de sequencias, como o BLAST, utilizam de metodologias de alinhamentos múltiplos de sequencias a fim de responder sobre a possível homologia no grupo de interesse (Altschul et al., 1990). Contudo, a similaridade sozinha não garante homologia uma vez que as sequencias podem ser extremamente parecidas porém não exercer a mesma função biológica. Com essa limitação em mente, outras ferramentas de coleta de dados como o HMMER abordam o problema utilizando de modelos probabilísticos e perfis de substituição de aminoácidos a fim de inferir a homologia das sequencias com mais precisão (Eddy., 2009). Essa técnica é mais custosa computacionalmente do que a apresentada pelo BLAST, porém com o avanço das tecnologias de processamento, se tornou uma poderosa aliada no campo dos estudos filogenéticos pois conseguem capturar grandes volumes de informação não renunciando à qualidade e precisão.

2.6.4 Ferramentas auxiliares

Ao trabalhar com Bioinformática, é uma rotina se deparar com arquivos com grandes quantidades de informação, densos e que precisam ser adequados para o objetivo do projeto. Dependendo da atividade a ser executada nesses arquivos, o tempo e trabalhos necessários podem facilmente se tornar uma realidade impraticável. Portanto, se torna quase que obrigatório que todo profissional da área tenha a sua disposição ferramentas capazes de auxiliar nos trabalhos diários da bioinformática. Ferramentas capazes de executar tarefas variáveis, consumindo recursos de forma responsável e eficaz e dentro de um tempo aceitável tornam o projeto possível e executável. Uma das ferramentas mais exploradas nesse trabalho foi o Python, que é uma linguagem de programação de fácil entendimento e que apresenta uma comunidade ativa que proporciona diferentes tipos de pacotes e scripts para os profissionais da área da Bioinformática: Biopython (Cock *et al*., 2009).

Atividades de coleta de informações, extração de domínios proteicos das sequencias, mudança de cabeçalhos dos arquivos FASTA e filtros aplicados durante todo o processo, só foram possíveis graças a ferramentas auxiliares que exerceram um papel principal dentro desse trabalho.

2.6.5 Filogenia

A filogenia é estudo das relações evolutivas entre organismos. Como a evolução é um processo de ramificação os organismos vão sofrendo especiação, essas relações são definidas a partir dos agrupamentos e analises das sequencias de nucleotídeos ou aminoácidos das populações de interesse. Com as tecnologias de sequenciamento avançando ano após ano, cada vez mais temos explorado técnicas de inferência de função biológica através do conceito de homologia. O estudo da filogenia é capaz de nos indicar a provável função biológica de espécies ou proteínas desconhecidas apenas pela proximidade e similaridade daqueles que temos conhecimento, através dos alinhamentos das sequencias ou probabilidades das matrizes de substituição, definitivamente uma poderosa ferramenta para a biologia molecular. Os principais métodos de inferência filogenética são: parcimônia, métodos de distância e métodos probabilísticos.

A análise de MÁXIMA PARCIMÔNIA baseia-se no conceito de que durante o processo evolucionário, o mais provável acontecimento é aquele que precisou do menor número de mudanças possível, ou seja, o processo conservador (Farris., 1970). Visto que seres biológicos tendem a sempre economizar energia e recursos, analises de MP minimizam os acontecimentos de homoplasia – mudanças na árvore final.

Nos métodos baseados em distância evolutiva o resultado é uma árvore fundamentada em uma matriz de distancias genéticas. A cada duas sequencias a distância é um valor único que representa as diferenças nas posições entre estas sequencias; esse valor recebe o nome de p-distance. Essa informação é então armazenada numa matriz que será necessária para o cálculo das distancias genéticas entre as sequencias com ajuda de um modelo de substituição como o Jukes-Cantor (Felsenstein, 1988).

Os métodos que trabalham com modelos probabilísticos são os mais amplamente utilizados atualmente, fato que só foi possível graças ao avanço da capacidade computacional nos últimos anos. Essas metodologias como MAXIMUM LIKELIHOOD tentam maximizar a probabilidade de observar dentro dos alinhamentos, quais as substituições mais prováveis de ter acontecido, ou seja, a máxima probabilidade dado determinado modelo (Swofford et al., 1996); onde quanto maior a probabilidade de determinada troca ter ocorrido, maior a confiabilidade naquele determinado resultado.

Com diferentes possibilidades para realizar os estudos de filogenia, a escolha de um melhor método se torna uma tarefa difícil. Alguns critérios precisam ser levados em consideração, como: acurácia, reprodutibilidade, tempo de processamento, etc. Uma metodologia que pode auxiliar na decisão da melhor e mais consistente arvore é o suporte estatístico de Bootstrap. Aplicado pela primeira vez em filogenia por Felsenstein em 1985, essa técnica reproduz diversas vezes a metodologia escolhida, e compara com o resultado original, para cada vez que o resultado for o mesmo, esse ramo recebe um valor. Valores altos indicam que aquela parte da arvore tem um suporte estatístico valido, que pode ajudar na decisão de qual melhor metodologia aplicar e testar a hipótese estudada.

2.6.6 Adaptação evolutiva

Durante o processo de evolução, substituições aleatórias de nucleotídeos ocorrem pelas mutações, sejam elas inserções ou deleções, que podem mudar a sequência de aminoácidos de uma proteína, caso essas mutações ocorram em um gene. Essas mudanças podem gerar três possibilidades numa proteína:

- A opção mais frequente é quando a troca de aminoácidos causa uma diminuição na função da proteína, esta substituição é do tipo deletéria e geralmente é eliminada da população por seleção de purificação.
- A opção mais rara é quando a troca de aminoácidos causa uma melhora na função da proteína, esta substituição é do tipo beneficial e é fixada na população por seleção natural ou positiva.
- A opção neutral é quando a troca de aminoácidos não causa nenhuma mudança na proteína, portanto essa é uma substituição neutra, e é fixada na população pelo acaso.

Ao comparar sequencias homologas de DNA, é esperado que o número de mutações silenciosas ou substituições sinônimas (dS), seja maior que o número de mutações positivas/negativas ou substituições não-sinônimas (dN), uma vez que existe a seleção de purificação para retirar as desvantajosas e as silenciosas acumulam numa mesma taxa durante o tempo. Contudo, caso exista uma pressão do meio ambiente que beneficie uma determinada característica ou mudança na proteína, é esperado que o número de substituições não-sinônimas seja maior que as sinônimas (Nei et al., 2000). Ao analisar a relação dN/dS para as substituições de posição, é possível chegar a um valor numérico real, que indica se determinada caraterística sofreu ou não pressão para uma seleção positiva.

3 Artigo científico

Esta seção apresenta o artigo desenvolvido durante a pós-graduação pelo autor, e que será submetido a periódicos científicos como artigo de pesquisa. O periódico escolhido foi Molecular Phylogenetics and Evolution (ISSN: 2329-9002)

Esta revista cientifica, que faz parte do ELSEVIER Publishing Group, apresenta caráter de publicação aberta e é especializada em trabalhos das áreas de Filogenia e Evolução molecular. Foi escolhida como objetivo pelos autores pois se tratar de uma revista com fator de impacto relevante e importante para a área de conhecimento da evolução molecular.

NtrY/NtrX Two-Component Regulator System Phylogenetic Study and Evolutionary Insights

Guerra B.S[a,*] , Souza E.M.[b], Chubatsu L.S. [b], Cruz L.M. [b]

[a]*Laboratório de Bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, Curitiba, Brasil*

[b]*Departamento de Bioquímica e Biologia Molecular, Universidade Federal do Paraná, Curitiba, Brasil*

ABSTRACT

The two-component regulatory system NtrYX is an essential life maintenance tool capable of sensing environmental stimuli and generate an adaptive response. This protein is mainly distributed between the proteobacteria and is involved on the most variable functions within these species. The domain AAA^+ from NtrX response regulator protein present significant structural differences; especially in the GAFTGA motif, responsible for iteration with σ^{54} promoters. NtrY histidine kinase homologous was also verified for structural divergences between the taxon's. In this work, homologous from NtrYX proteins were identified using domain-based search tools; each experimental group was tested using phylogenetic methods. We found relevant differences in domain-structure and taxonomic distribution of the analyzed groups; these results are directly reflected on the phylogeny tree conducted in this work. New insights about this important two-component regulatory system is proposed in this work.

Keywords: Two-component system, Phylogenetics, Molecular Evolution, Proteobacteria

1. INTRODUCTION

The two-component regulator system is an essential adaptation tool for the prokaryotes. This system is capable of sensing environmental stimuli and generating an adaptive response, producing a fast and precise response to a change in the environment. These systems are composed of a sensor histidine kinase (HK) and a response regulator (RR) protein [1]. Where the HK is responsible to sense stimuli, autophosphorylate a histidine residue, and appropriately regulate the signaling pathway, which consequently activates the RR and generates a cellular response.

For the prokaryotes, especially the Proteobacteria phylum, it was made ample use of the Two-component regulator system and it had incontestable relevance. One remarkable example is the NtrB/NtrC system, present on the mostly on the proteobacteria, which is responsible for the regulatory cascade of the nitrogen metabolism that commonly controls the NifA gene expression [2]. Another example of a two-component system is the NtrY/NtrX that is well described and associated with the regulation of nitrogen metabolism in Proteobacteria. This system was observed in Alphaproteobacteria like *Brucela abortus* as capable of responding to oxygen limiting conditions [3]. It was also reported in species like *Azorhizobium caulinodans* [4]; *Azospirillium brasilense* [5]; *Rhodobacter capsulatus* [6] and *Rhizobium tropici* [7], reinforcing its importance. As for the Betaproteobacteria, the system is present in species like *Neisseria gonorrhoeae* and is responsible for activating respiratory enzymes [8]. In addition, in 2016 Bonato and collaborators demonstrated that *Herbaspirillum seropedicae* NtrYX is involved in regulation of the nitrate metabolism and it is related to NtrBC [9].

When investigating the NtrX protein structure it consists of three domains: an N-terminal REC domain (Response Regulator), a central AAA^+ domain, and a C- terminal helix-turn-helix (HTH). The central AAA^+ domain is capable of regulating σ^{54}promoters thanks to its motif GAFTA, which are also present in NtrC-like systems [10]. However, some representing of Alphaproteobacteria like *Brucela abortus* and *Rhodobacter capsulatus* do not have this essential motif, compromising σ^{54} promoters, and as consequence they regulate $\sigma^{70}$and σ^{32}promoters [11] [12]. For the Betaproteobacteria, like *Herbaspirillum seropedicae*, the entire AAA+ domain is absent; this results in an NtrX protein with only 2-domains wise and no σ^{54}interaction promoters. Is this unique characteristic present in all *β*-Proteobacteria NtrX? Moreover, for the Alphaproteobacteria, is there any difference for NtrX? With the objective of better understanding the evolutionary relationship for the NtrX protein in Proteobacterias, Atack and collaborators, analyzed 534 homologous and classified them into four major groups [13]. Although the research was essential to better understand the role of *Neisseria gonorrhoeae* NtrX protein, the phylogenetic results lacked some NtrX-like proteins. It also did not perform an NtrY comparison, leaving many questions unanswered about the evolution of this important two- component regulator system.

With all the questions still open to this day, in this work we evaluate the evolutional relationship between all homologous of NtrYX protein based on phylogenetic analysis and gene context. Moreover, we try to identify if there is any evidence of positive adaptive evolution within the organism.

2. MATERIALS AND METHODS

2.1. Sequence Retrieval

First, we identified the domains from PFAM database [14] using two NtrX confirmed proteins (*Herbaspirillum seropedicae* Strain SmR1 accession number WP_013232153.1 for NtrX & WP_013232154.1 for NtrY; *Azorhizobium caulinodans* accession number WP_012171607.1; *Neisseria gonorrhoeae* accession number WP_003690124) and retrieve the HMM-profile for each domain found. The search strategy for the NtrYX homologous was based in three major groups:

1. All Two-Domain NtrX candidates: REC response regulator (PF00072) and HTH_8 (PF02954)
2. All Three-Domain NtrX candidates: REC response regulator (PF00072), AAA^{+} Domain (PF00158 or PF14532) and HTH_8 (PF02954);
3. All NtrY candidates: HAMP family (PF00672), PAS domain (PF13188), *HisKA* domain (PF00512) and HATPase_c domain (PF02518).

Amino acid sequences from RefSeq non-redundant protein (Release 85) that presented at least one of the domains were selected using HMMER3 [15]. Sequences that had the Two-domain or Three-domain structure were then chosen. For the NtrY homologues, the same process was applied but with no separation for numbers of domains. However, this methodology was not enough to guarantee that all sequences are truly NtrYX homologues.

2.2. Sequence Filtering

A local Blast was applied to all three-study groups in search for possible e-values and bitscore breakpoints [16]. Using the same three protein sequences as queries from previous step (WP_013232153.1; WP_012171607.1; WP_013232154; WP_003690124) and our retrieval as database, we could evaluate how close related our protein selection was to the confirmed NtrYX proteins. For details about each of the parameters measured (E-value, Bitscore, length and identity) please check Supplementary Material, Fig. As, Bs and Cs. Once filtered, the sequences were clustered by 90% identity using the CD-HIT web tool [17] and one representative of each cluster was selected (Supplementary material, Fig. Ds, Es and Fs). PfamScan were used to confirm that all sequences in the last step had the expected number of domains [18]. The final number of homologous sequences for each of three experimental groups are described in Table 1.

Table 1: Summary of NtrYX homologues

Major groups	HMMER	BLAST	CLUSTER	GLOBAL	SUMMARIZE
Two-Domain	4,430	606	206	176	37
Three-Domain	36,920	3,174	1,046	465	58
NtrY	14,333	2,208	1,070	x	85

2.3. Phylogenetic Analysis

To eliminate redundant information in the phylogenetic analysis, a dataset comprising a lower number of sequences was selected for study, at least one representative of each taxonomic was picked to maintain the original characteristics of the tree. Each domain was extracted separately from the whole protein, based on the annotation from PfamScan [18] output file. This step was only possible with the assist of a Python script. Afterwards, each domain amino-acid sequence was aligned with the MUSCLE software [19] and concatenated with Bio-Edit software respecting the original protein domain order. This process was applied to all three major groups.

For the final phylogenetic trees, the chosen phylogeny inference method was Maximum-Likelihood. All trees were constructed using RAxML software for Linux [20]; the protein substitution model and the estimate distribution parameter were determined with MEGA7 software [21]. Bootstrap values were adjusted to 1000 replicates [22].

2.4. Adaptive Evolution

Adaptive Evolution analyses were performed with HyPhy and Datamonkey phylogenetic software [23] [24]. For the three major groups, all summarize sequences were extracted as nucleotide from the NCBI database. Tests of FEL for detecting amino-acids sites under selection pressure was conducted [25].

3. RESULTS AND DISCUSSION

3.1. Sequence Analysis of the NtrYX Major Groups

The search approach based on Three-different groups was proposed to maximize the probability of collecting the majority of NtrYX homologues in the database. The number of information about protein is still modest, only a few species have been tested to confirm their structure and that they true Two-component regular systems. However, within the confirmed ones, it is possible to detect three main differences between their domains: 1) *Herbaspirillum seropedicae* NtrX-RR has only REC and HTH domains; 2) *Neisseria gonorrhoeae* NtrX-RR has the three domain structure, but with a relevant difference at AAA^{+}domain from the others NtrX-like proteins; 3) *Azorhizobium caulinodans* NtrX protein and others *α*- Proteobacteria have the most conservative three-domain structure. The strategy to collect a considerable number of sequences and apply step-by-step filters for each one of the groups supports the idea that we probably collect an expressive number of homologues.

The HMMER search tool can identify homologous protein sequences based on profile-hidden Markov models for the domains. Such profiles (extracted from Pfam-Database in this work) contain the probability for each possible amino-acid substitution and incorporates it to a probabilistic inference search method. The target was the RefSeq Complete Non-Redundant Protein database. The search, as expected, resulted in big numbers for each one of the domains. Successively two filters were applied to reduce these numbers: Local blast using specific sequences as query and clustering method. The blast search strategy allied with the statistical tool R, resulted in break-points possible to visualize. For all three groups the absolute -log(Evalue) were ordered by size. This allows to select only more suitable cut-points supported by Bitscore, identity and length values. Rule of thumb for e-value homology break-points were not applied in this work, since custom size databases may have different e-value distributions. Although the blast filter presents great insight about the NtrYX homologues and lowers significantly the number of sequences, a clustering method by sequence identity was also tested. Sequence identity threshold of 90% was important to remove redundant information from our data. The number of representatives chosen sequences for each step is shown in Table 1; supplementary material Fig. As, Bs and Cs, present the blast results and break-points; and supplementary material Fig. D, E and F the

All NtrYX sequences were also analyzed by their taxonomic context with aid of GenBank annotation. This Two-component protein is distributed mainly among the Proteobacteria. Most examples from NtrYX are from the *Alphap*roteobacteria class, as expected, the number of sequences with Three-Domains are superior to the Two-Domains. Moreover, the Three-Domain architecture, also presented a few *Betap*roteobacteria, thanks to *Neisseria gonorrhoeae*. Some other phylums had their share of the taxonomic context: Nitrosospira, Firmicutes, Acidobacteria, Gamma and Delta Proteobacteria. For the Two- Domain structure, only *Betap*roteobacteria class was present. This group conferred to be more restricted then the Three-Domain one. The complete summary of the taxonomic context is shown in supplementary material Table. 1s

3.2. Phylogeny of the NtrYX homologues

No phylogenetic tree can be better than the alignment that generates that tree, which is a universal true for all evolutionary scientist. Since our strategy to collect every single of our sequences was based on protein domain, all phylogenetic data were treated as amino-acid sequences. To minimize the influence of bad alignment in our study, every domain from each group was processed separately at MUSCLE 7.0 tool, and concatenated respecting the original form of the protein. Domain-only trees maximizes the potential on studying the differences between these regions and reduces the influence of misalignment in our data.

The survey of the NtrX alignment revealed that REC-Domain and HTH-Domain were highly conserved for every different group or taxonomy. However, the same is not applied to AAA^+ domain. The two variances, one from *Neisseria gonorrhoeae (*PF14532) and another one from *Azorhizobium caulinodans (*PF00158) presented notable differences in their alignments, specifically on the region of the motif GAFTGA, where this both sequences and his homologues have it truncated or either missing Figure 1. Moreover, it remains conserved for other Three-Domains species non-alfa, indicating that these NtrX-like proteins probably didn't lose the σ^{54} interaction.

For the phylogenetic tree reconstruction, the method chosen was Maximum-likelihood. Not only is this method a powerful statistical ally that seeks the tree that maximizes the probability of observing the data, but it also has different models of amino-acid substitution that are tested to guarantee the best log-likelihood. During many years, it considered slow and unpractical for amino-acid. However, with the development of new data processing technologies, it consolidates as one of the most popular methods. All three analyses were conducted under RAxML software; the best parameters for substitution model and evolutionary rates among sites were tested in MEGA7 software. The best scoring model and rates was LG+GAMMA. This result can be found at Supplementary Material Table 2s, 3s and 4s. All trees were supported by Bootstrap method, 1000 itarations, to ensure that our display data is well supported.

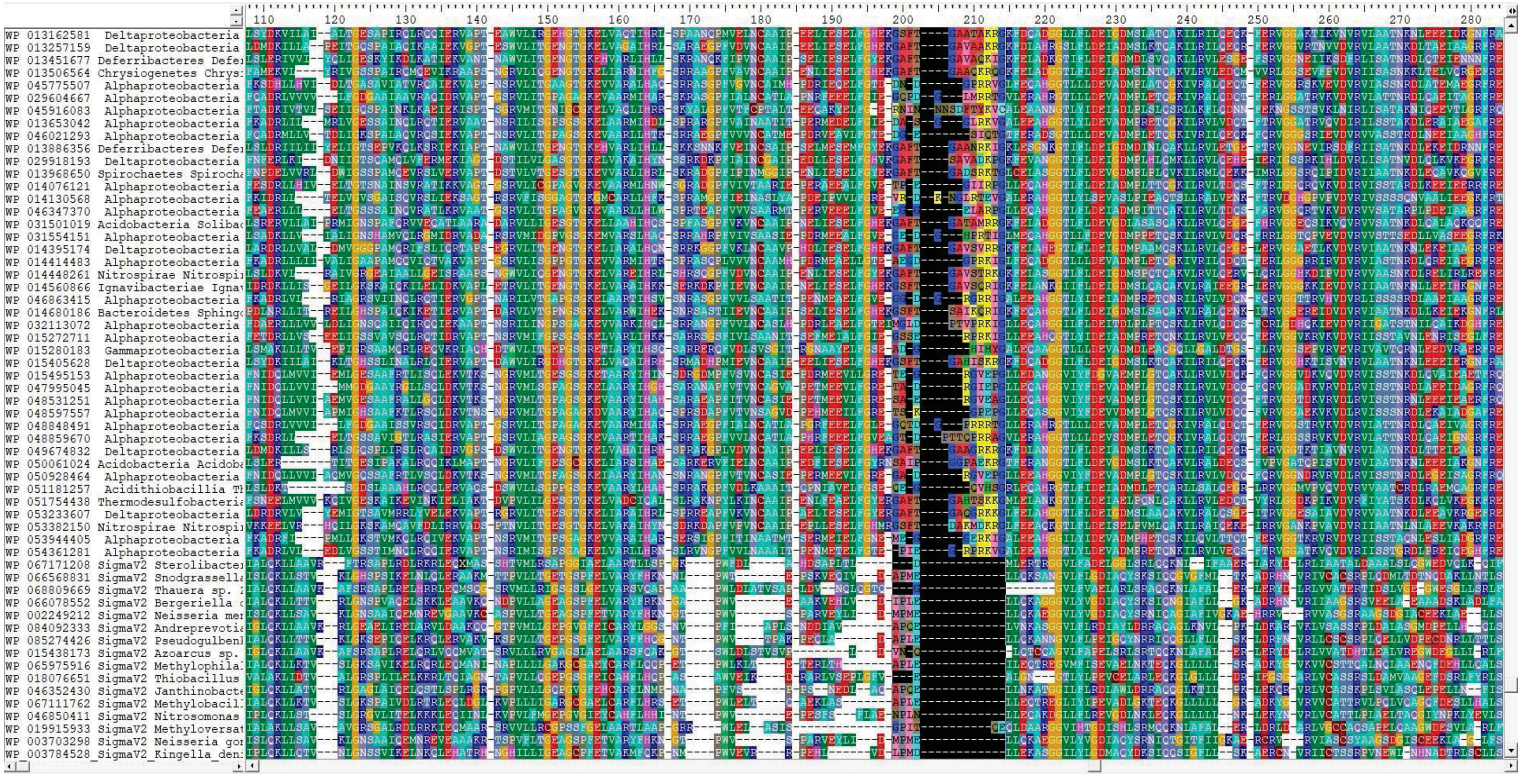


Figure 1: Three-Domain multiple sequence alignment. Region highlighted in black indicates the motif GAFTGA from distinct species in the data. The whole AAA^{+}domain region is shown position 119-280 in the alignment. ID's and Taxonomy on the left. BioEdit software

3.3. Two-Domain Phylogeny

The reconstruction of the phylogenetic tree for the Two-domain group presented some difficulties due to his particularly high sequence conservation and short sequences length. The bootstrap values and short-branched clades support that idea. Two trees were constructed, one containing the GLOBAL number of sequences found after every filter step (Supplementary material Fig. Gs). And another is a summary of this global tree, to present a clearer view of the data Figure 2. These results show that there is no clear or supported partition between the clades. However, within the *H. seropedicae* proximity it is possible to determine some close related taxon's, indicating that these organisms may perform the same biological function. The fact that only *Betap*roteobacteria were found during our survey, corroborate the idea that NtrX with two-domain architecture have correlated genetic origin, since it could be one indicative these proteins don't have and never had the AAA^{+}domain.

Figure 2: Evolutionary relationships of Two-Domain NtrX proteins. Tree was inferred using the Maximum-Likelihood method. The parameters for analysis was LG substitution model, Gamma distribution and fixed base. Optimization Likelihood: -2587.004871. Bootstrap 1000 iterations, numbers indicate the support for each node. In the yellow box is the confirmed NtrX protein from *H. seropedicae.*

3.4. Three-Domain Phylogeny

Three-Domain phylogenetic results infer a more interesting correlation between the groups; the differences in AAA^+ domain is notable Figure 1. The increased information from different variety of the same domain was reflected in the segregation among the clades and supported by high-value bootstrap numbers. The branch lengths are also important to considerer, since the number of genetic change increases for each group. Two trees were constructed, one containing the GLOBAL number of sequences found after every filter step (Supplementary material Fig. Hs). And another is a summarize of this global tree, to present a clearer view of the data Figure 3.

These results help to elucidate some questions about this two-component regular system. The *α* and some *β* proteobacteria have this GAFTGA motif truncated or either missing, but not at the same speed. For the *Alphap*roteobacteria, that has his neighborhood related to NtrB/C genes, this AAA^+ domain seems to be closer to GAFTGA-present ones. Since they have almost the same length, this could indicate one gene duplication and specialization process. Moreover, this difference indicates the specialization in NtrX function for this group, since without GAFTGA motif, the σ^{54}interaction is lost. When we focus on the *Betap*roteobacteria, his neighborhood is not related to NtrB/C genes and the AAA^+ domain is truncated by approximately 50 AA the branch lengths for the group support this fact. Furthermore, why in bacteria like *H. seropedicae* this domain is completely absent and *N. gonorrhoeae* there is remains of it while maintaining the same function as others NtrX proteins? The hypothesis of gene duplication for the *β*- Proteobacteria still remains unclear.

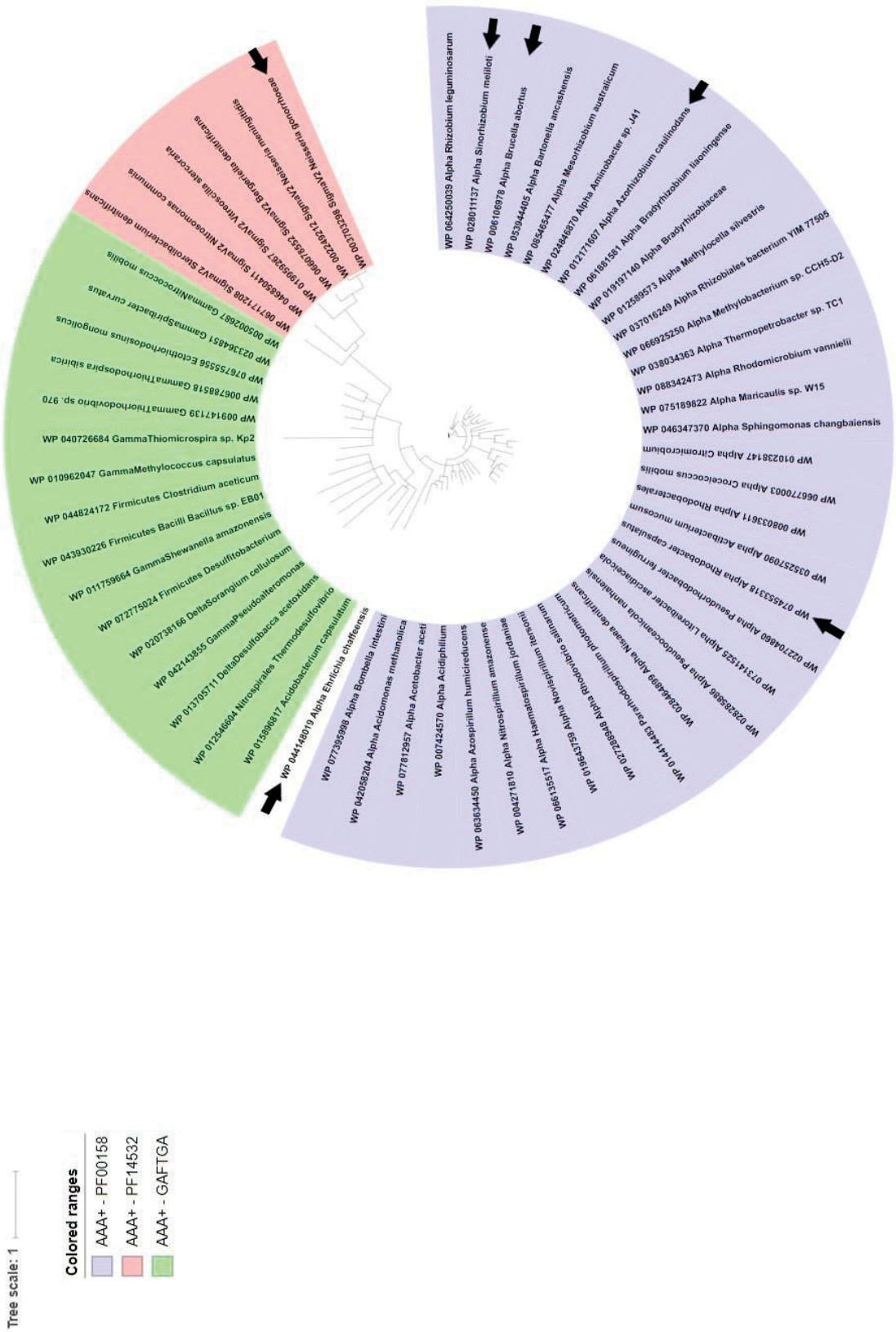


Figure 3: Evolutionary relationships of Three-Domain NtrX proteins. Tree was inferred using the Maximum- Likelihood method. The parameters for analysis was LG substitution model, Gamma distribution and fixed base. ML Optimization Likelihood: -20509.561964. Bootstrap 1000 replicates, only branches with <50% support is collapsed. Black arrows indicate NtrX confirmed proteins. Colored ranges indicate the possible groups. iTOL online software was used to reproduce the tree. [26]

3.5. NtrY Phylogeny

The NtrYX is a two-component regular system, which means that every NtrX-RR should have a NtrY-HK. The selected sequences for analyses were the NtrY-HK protein from the Two-domain and Three-domain summarize trees. This method guarantees that every NtrX- like protein found in this work has his correspondent NtrY-like. The domains HAMP family (PF00672), *HisKA* domain (PF00512) and HATPase_c domain (PF02518) from NtrY structure were highly conserved between all taxons in this work. However, for the PAS domain a variation in domain length was noticed, reflecting in the quality of the multiple sequence alignment. One tree was constructed with all summarize sequences from previous analysis, to present a clearer view of the data Figure 4.

The NtrY ML-Tree presents a small branch length for most of the sequences. However, one group formed from all types of taxon contained in this work has a notable genetic distance from others organism. This could be a result from the PAS domain differences. Two big clusters were formed by *α* and *β* Proteobacteria NtrY sequences, this may be related to the differences between both NtrYX neighborhood and their origins. One remarkable occurrence was the fact that not every group was clearly defined, where some taxons got mixed in the tree. This indicates, along with the conservative characteristics of NtrY domains, that the Histidine Kinase receptor from NtrYX two-component system is more conserved than his response regulator pair.

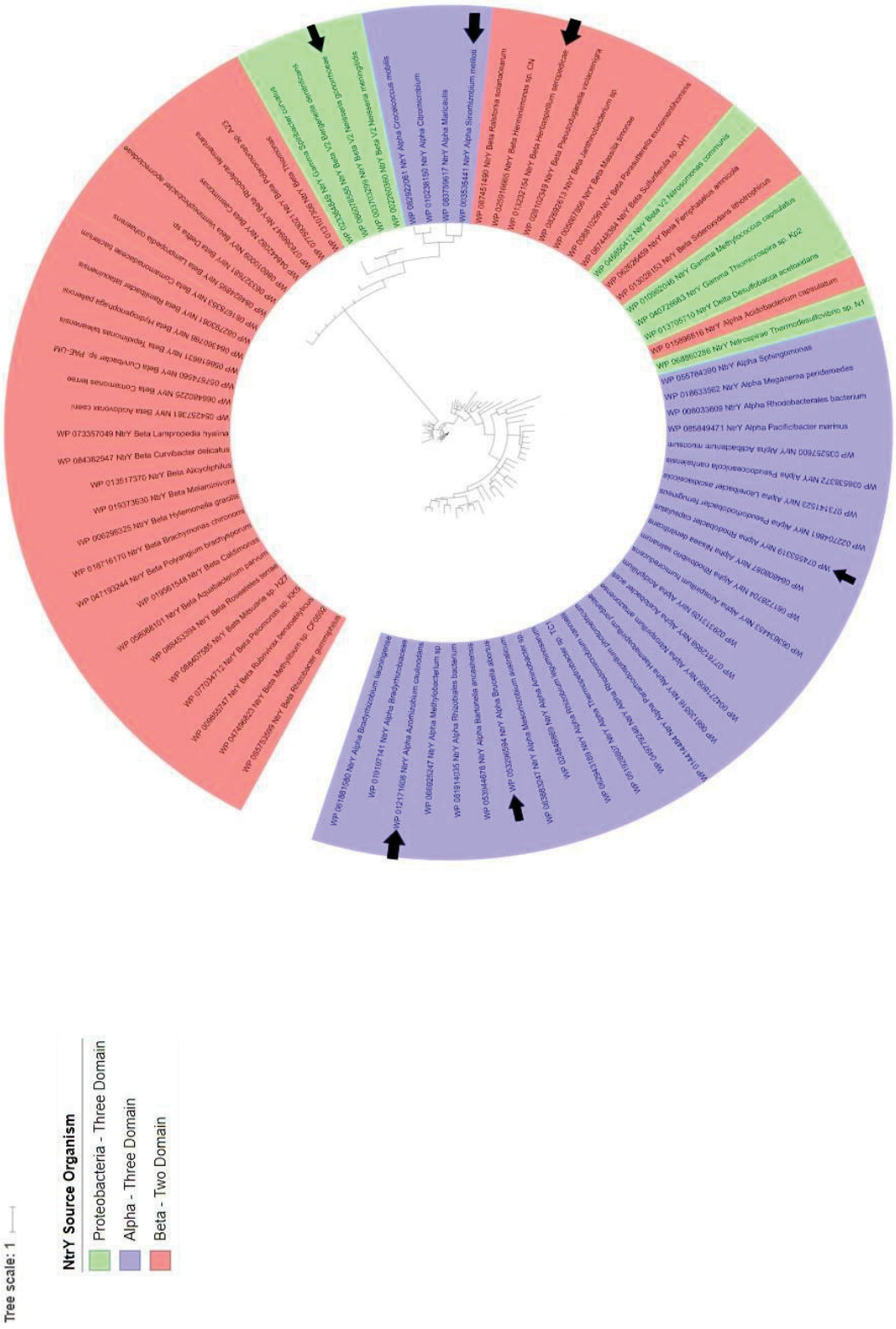


Figure 4: Evolutionary relationships of NtrY proteins. Tree was inferred using the Maximum-Likelihood method. The parameters for analyses was LG substitution model, Gamma distribution and empirical base frequencies. ML Optimization Likelihood: -33986.544579. Bootstrap 1000 replicates, only branches with <50% support is collapsed. Black arrows indicate NtrY confirmed proteins. Colored ranges indicate the possible groups. iTOL online software was used to reproduce the tree. [26]

3.6. Adaptive Evolution

To test the hypothesis that the differences between the amino-acid sequences from the two major NtrX groups were under action of adaptive evolution pressure, tests for this method were applied. FEL or Fixed Effects Likelihood is a tool from HyPhy program capable of using a maximum-likelihood approach to infer the dN/dS substitution rates for amino-acid sequences exploring his coding nucleotide alignment. It was expected that, for the Three- Domain group, pressure was occurring. Nucleotide coding sequences for each of NtrX major group sequences were extracted from NCBI data-base, CODON-alignments were applied using MUSCLE software, and small corrections were applied manually. Surprisingly, the results for the Three studying groups (Two-domain; Three-domain; Two+Three Domain) were purifying selection at most nucleotide sites in Table 2.

Table 2: Estimates of evolutionary divergence between the two NtrX major groups

Groups	Diversifying selection	Purifying selection	ω
Two-Domain	7	176	0.197
Three-Domain	1	425	0.145
Two+Three-Domain	1	412	0.180

Note: The number of significant base substitutions per site from analysis between sequences is shown. The ω indicates the dN/dS ratio normalized by statistical parameters. Value of $\omega > 1$ indicates diversifying selection; value $\omega < 1$ indicates purifying selection.

4. CONCLUSION

The NtrYX two-component regular system is an essential mechanism that allows bacteria to better interact with ecosystems. The complete understand of his function, origin and evolutive directions is still unknown. However, this work contributes elucidating that the NtrYX system has significant differences between their domain structures; the taxonomic distribution and exclusivity for some groups suggest that the origin of this gene may not have been the same for this protein. Moreover, the AAA^+ domain especially the GAFTGA motif has shown to be the most variable structure for the NtrX-RR protein; in the same matter the PAS domain for the NtrY-HK protein. The process of homologous identification resulted an expressive number of sequences and species, this can be helpful to the scientific community.

Although this work has contributed to better understanding the NtrYX evolution process; analysis of gene neighborhood and a more detailed adaptive evolution method could support the main hypothesis of gene duplication and specialization for the NtrYX complex.

ACKNOWLEDGEMENTS

Grants from the Brazilian National Research Council (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) supported this work. B.S. Guerra received scholarship from CAPES.

REFERENCES

[1] A. M. Stock, V. L. Robinson, P. N. G. and, Two-component signal transduction, Annual Review of Biochemistry 69 (1) (2000) 183–215.
URL 10.1146/annurev.biochem.69.1.183;https://doi.org/10.1146/annurev.biochem.69.1.183

[2] R. Dixon, D. Kahn, Genetic regulation of biological nitrogen fixation, Nature Rev. Microbiol. 2 (2004) 621–31.

[3] M. Carrica, I. Fernandez, M. Marti, G. Paris, F. A. Goldbaum, The ntry/x two-component system of brucella spp. acts as a redox sensor and regulates the expression of nitrogen respiration enzymes, Molecular microbiology 85 (2012) 39–50.

[4] K. G. P. K. (Max-Planck-Institut fuer Zuechtungsforschung, U. Klosse, F. de Bruijn, Characterization of a novel azorhizobium caulinodans ors571 two-component regulatory system, ntry/ntrx, involved in nitrogen fixation and metabolism (1991).

[5] M.L., M. Assump A§ A, H. Machado, E. Benelli, E. Souza, F. I. Pedrosa, Identification and characterization of the two-component ntry/ntrx regulatory system in azospirillum brasilense, Brazilian Journal of Medical and Biological Research 35 (2002) 651–661.
URL http://www.scielo.br/scielo.php?script=sci_arttext\&pid=S0100-879X2002000600004\&nrm= iso

[6] J. Gregor, T. Zeller, A. Balzer, K. Haberzettl, G. Klug, Bacterial regulatory networks include direct contact of response regulator proteins: Interaction of rega and ntrx in rhodobacter capsulatus, Journal of molecular microbiology and biotechnology 13 (2007) 126–39.

[7] J. Nogales, R. Campos, H. BenAbdelkhalek, J. Olivares, C. Lluch, J. Sanjuan, Rhizobium tropici genes involved in free-living salt tolerance are required for the establishment of efficient nitrogen-fixing symbiosis with phaseolus vulgaris, Molecular plant-microbe interactions : MPMI 15 (2002) 225–32.

[8] J. M. Atack, Y. N. Srikhanta, K. Y. Djoko, J. P. Welch, N. H. M. Hasri, C. T. Steichen, R. N. V. Hoven, S. M. Grimmond, D. S. M. P. Othman, U. Kappler, M. A. Apicella, M. P. Jennings, J. L. Edwards, A. G. McEwan, Characterization of an ntrx mutant of neisseria gonorrhoeae reveals a response regulator that controls expression of respiratory enzymes in oxidase-positive proteobacteria, Journal of Bacteriology 195 (11) (2013) 2632–2641.

[9] P. Bonato, L. Alves, J. H. Osaki, L. Rigo, F. Pedrosa, E. Souza, N. Zhang, J. Schumacher, M. Buck,
R. Wassem, L. Chubatsu, The ntry/ntrx two-component system is involved in controlling nitrate assimilation in herbaspirillum seropedicae strain smr1, The FEBS journal 283.

[10] M. Bush, R. Dixon, The role of bacterial enhancer binding proteins as specialized activators of σ54-dependent transcription, Microbiology and Molecular Biology Reviews 76 (3) (2012) 497–529.
URL 10.1128/MMBR.00006-12;http://mmbr.asm.org/content/76/3/497.abstract

[11] W. C. Bowman, R. G. Kranz, A bacterial atp-dependent, enhancer binding protein that activates the housekeeping rna polymerase, Genes & Development 12 (12) (1998) 1884–1893.
URL 10.1101/gad.12.12.1884;http://genesdev.cshlp.org/content/12/12/1884.abstract

[12] I. Fernandez, L. Otero, S. Klinke, M. Carrica, F. A. Goldbaum, Snapshots of conformational changes shed light into the ntrx receiver domain signal transduction mechanism, Journal of Molecular Biology 427.

[13] J. Atack, Y. Srikhanta, K. Y. Djoko, J. P. Welch, N. Hasri, C. T. Steichen, R. N. V. Hoven, S. Grimmond, D. S. M. P. Othman, U. Kappler, M. A. Apicella, M. P. Jennings, J. L. Edwards, A. Mcewan, Characterization of an ntrx mutant of neisseria gonorrhoeae reveals a response regulator that controls expression of respiratory enzymes in oxidase-positive proteobacteria, Journal of bacteriology 195.

[14] R. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. L. Sonnhammer, A. Bateman, Pfam: clans, web tools and services, Nucleic acids research 34 (2006) D247–51.

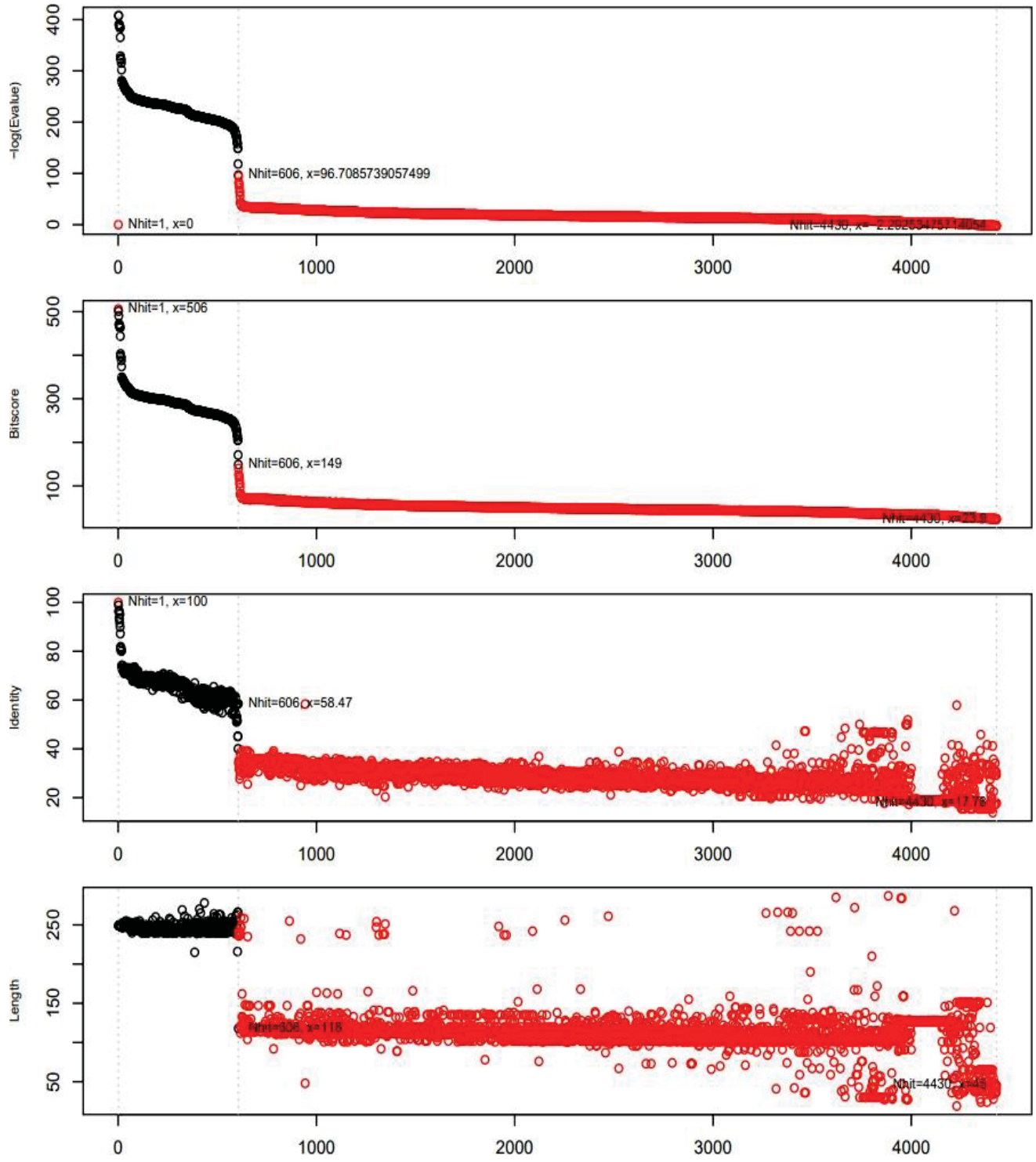
[15] R. E. SEAN, A new generation of homology search tools based on probabilistic inference.

[16] S. Altschul, T. L. Madden, A. Shaffer, J. Zhang, Z. Zhang, Gapped blast and psi-blast:a new generation of protein database search programs, Nucl. Acids. Res. 25 (1996) 3389–3402.

[17] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide

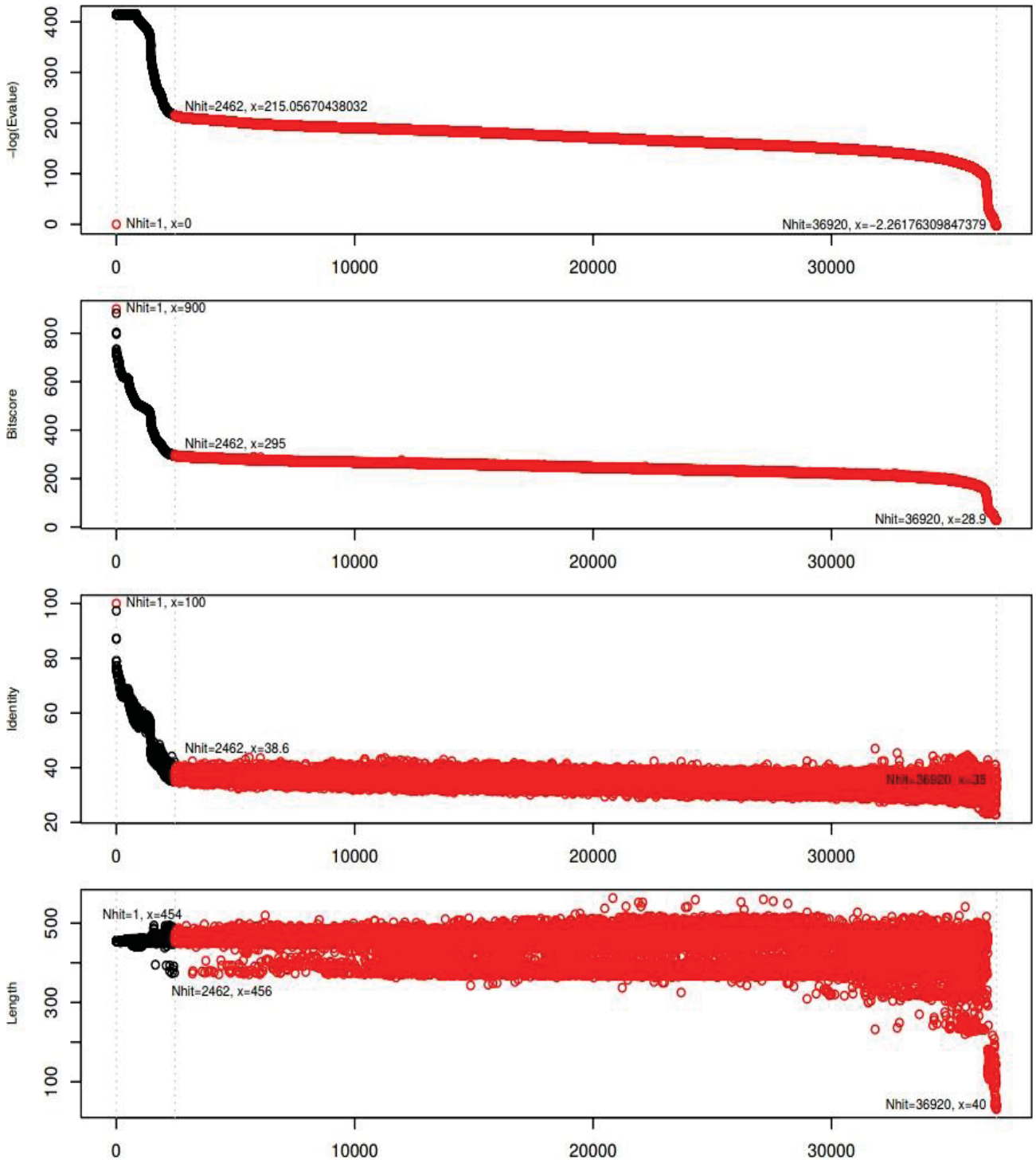
sequences, Bioinformatics (Oxford, England) 22 (2006) 1658–9.
[18] J. Mistry, A. Bateman, R. Finn, Predicting active site residue annotations in the pfam database, BMC bioinformatics 8 (2007) 298.
[19] R. C. Edgar, Muscle: Multiple sequence alignment with high accuracy and high throughput, Nucleic acids research 32 (2004) 1792–7.
[20] A. Stamatakis, Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics (Oxford, England) 30.
[21] S. Kumar, G. Stecher, K. Tamura, Mega7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets (2016).
[22] J. Felsenstein, Confidence limits on phylogenies: An approach using the bootstrap, Evolution 39 (1985) 783–791.
[23] S. L. K. Pond, S. D. W. Frost, S. V. Muse, Hyphy: hypothesis testing using phylogenies, Bioinformatics 21 (5) (2005) 676–679.
URL 10.1093/bioinformatics/bti079;+http://dx.doi.org/10.1093/bioinformatics/bti079
[24] W. Delport, A. F. Y. Poon, S. D. W. Frost, S. L. K. Pond, Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology, Bioinformatics 26 (19) (2010) 2455–2457.
URL 10.1093/bioinformatics/btq429;+http://dx.doi.org/10.1093/bioinformatics/btq429
[25] S. L. K. Pond, S. D. W. Frost, Not so different after all: A comparison of methods for detecting amino acid sites under selection, Molecular Biology and Evolution 22 (5) (2005) 1208–1222.
URL 10.1093/molbev/msi105;+http://dx.doi.org/10.1093/molbev/msi105
[26] I. Letunic, P. Bork, Interactive tree of life (itol) v3: An online tool for the display and annotation of phylogenetic and other trees, Nucleic Acids Research 44 (2016) gkw290.

4 MATERIAL SUPLEMENTAR



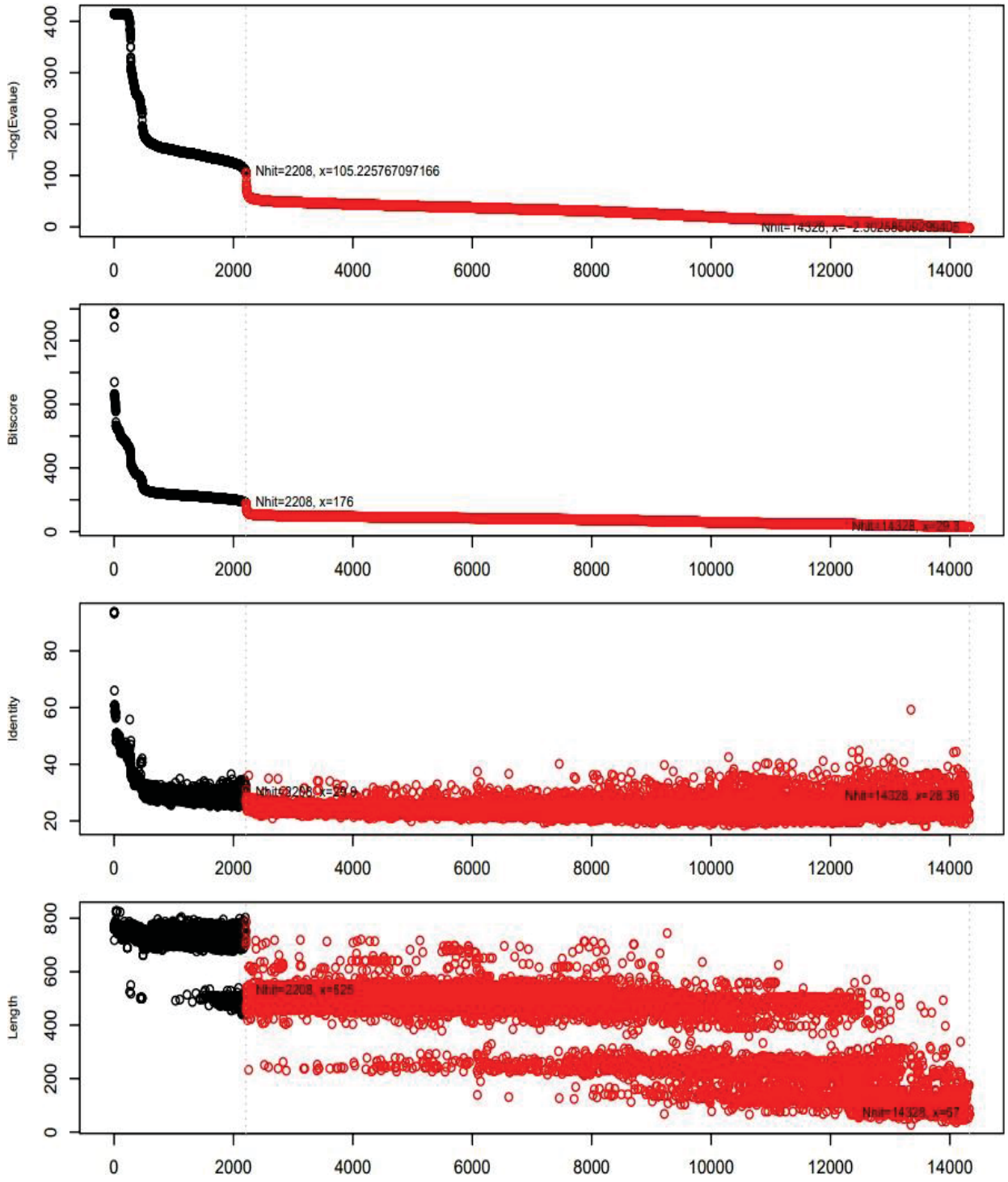
FONTE: O Autor (2018).

Supplementary material, Fig. As – Blast results for NtrX *H. seropedicae*. The selected break-points are in black. Nhit indicates the number of selected and last sequences position. X indicates the value for each Nhit hit. Adapted from Bio3D R Package. e-value = 1,00E-042.



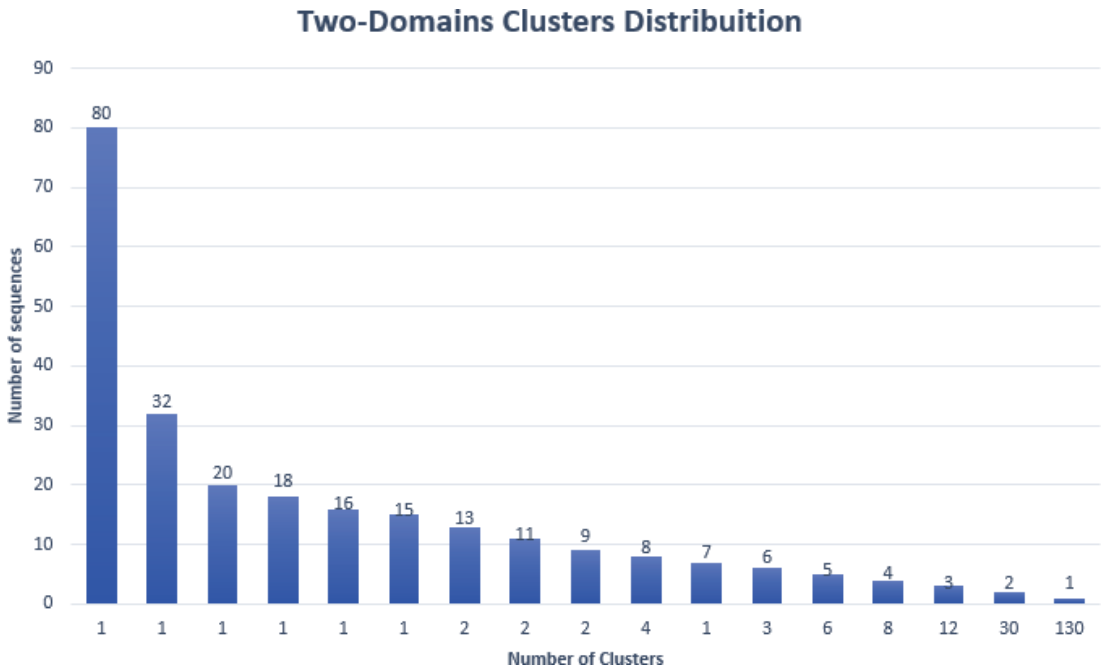
FONTE: O Autor (2018).

Supplementary material, Fig. Bs – Blast results for NtrX *Azorhizobium caulinodans*. The selected break-points are in black. Nhit indicates the number of selected and last sequences position. X indicates the value for each Nhit score. Adapted from Bio3D R Package. e-value = 4,00E-094.



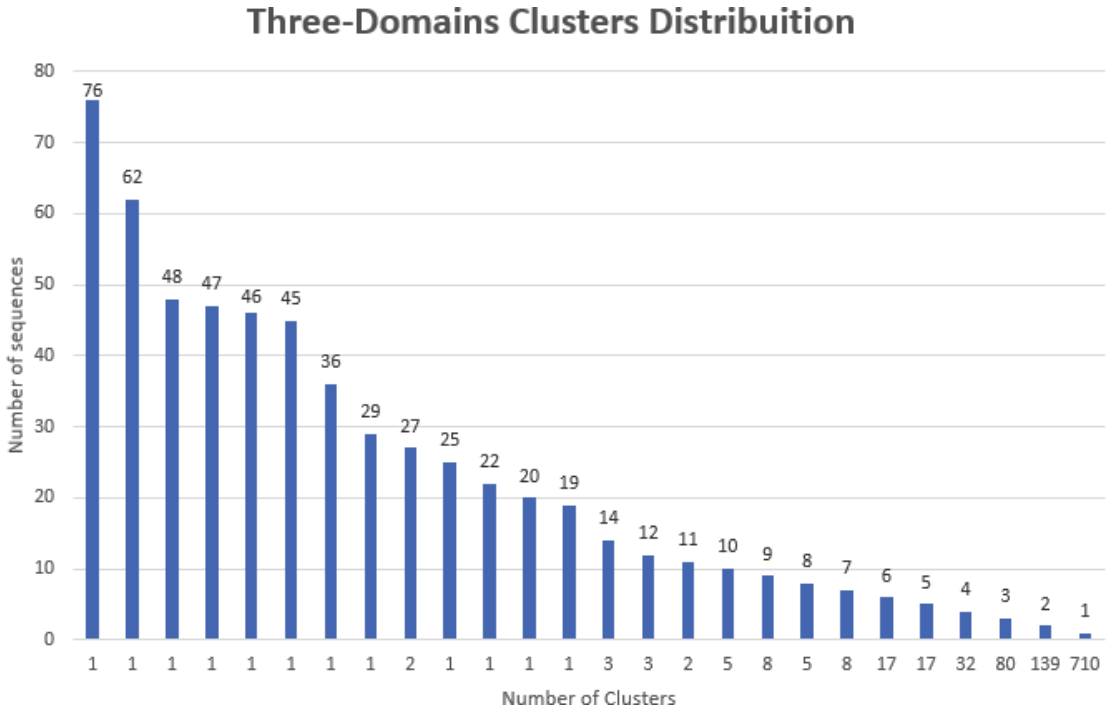
FONTE: O Autor (2018).

Supplementary material, Fig. Cs – Blast results for NtrY *H. seropedicae*. The selected break-points are in black. Nhit indicates the number of selected and last sequences position. X indicates the value for each Nhit score. Adapted from Bio3D R Package.



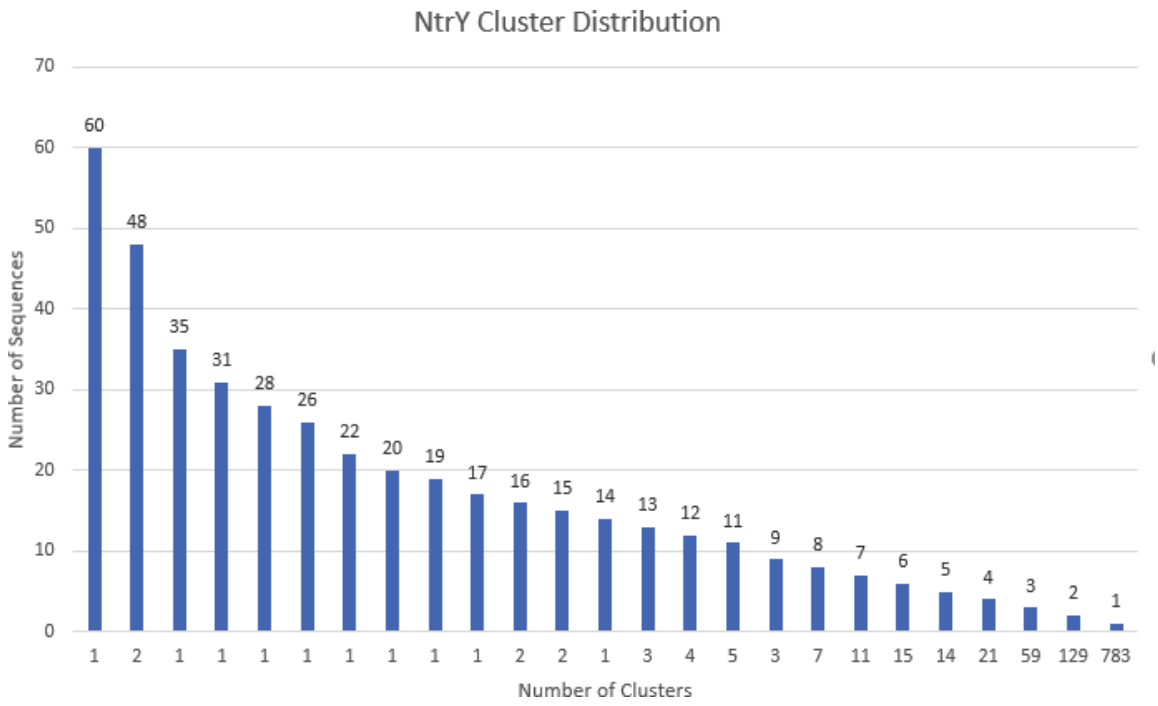
FONTE: O Autor (2018).

Supplementary material, Fig. Ds – Cluster distribution for Two-Domain group. Number of sequences x Number of Clusters. The values on top of each bar indicates the number of sequences for each cluster size.



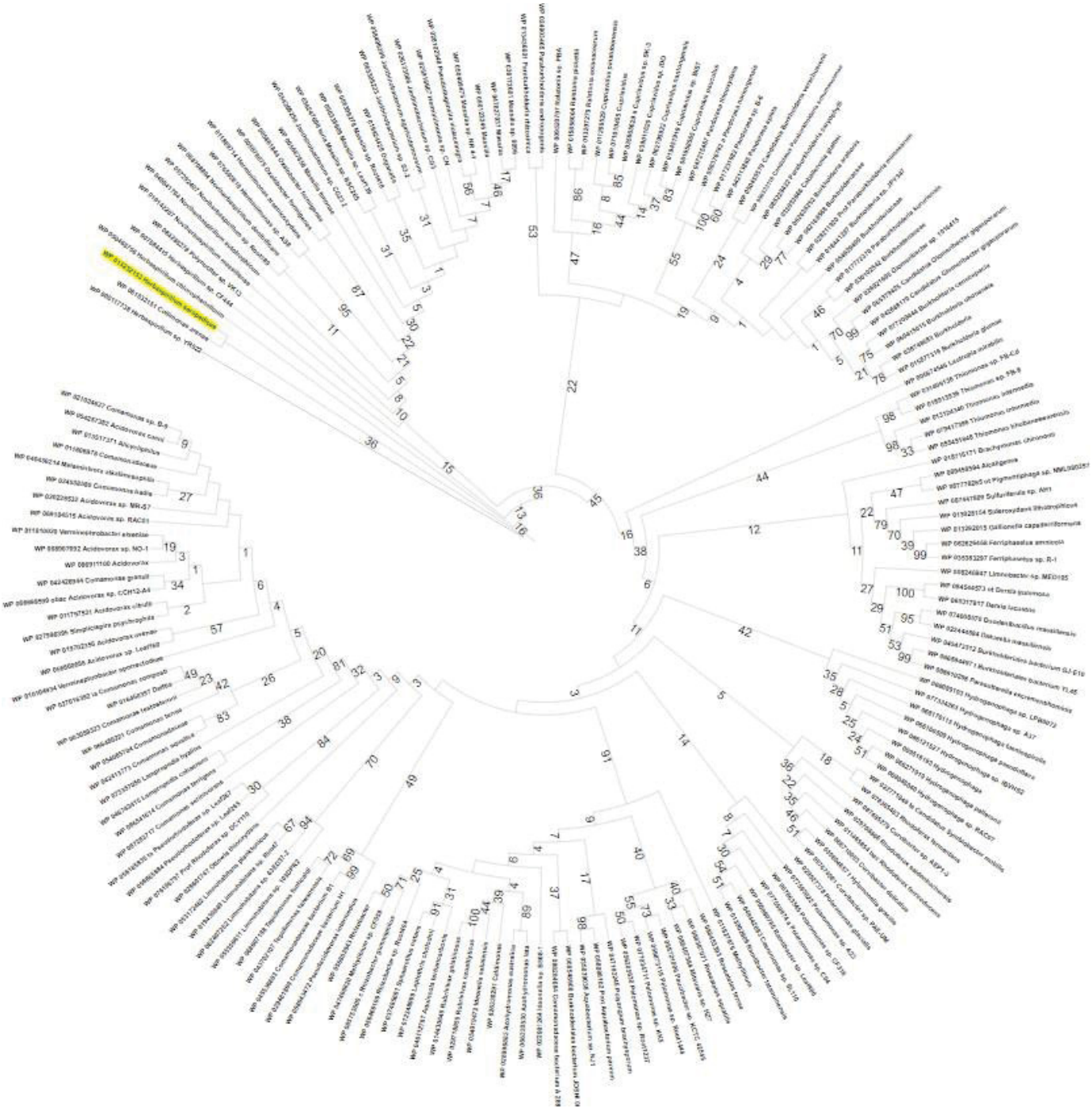
FONTE: O Autor (2018).

Supplementary material, Fig. Es – Cluster distribution for Three-Domain group. Number of sequences x Number of Clusters. The values on top of each bar indicates the number of sequences for each cluster size.



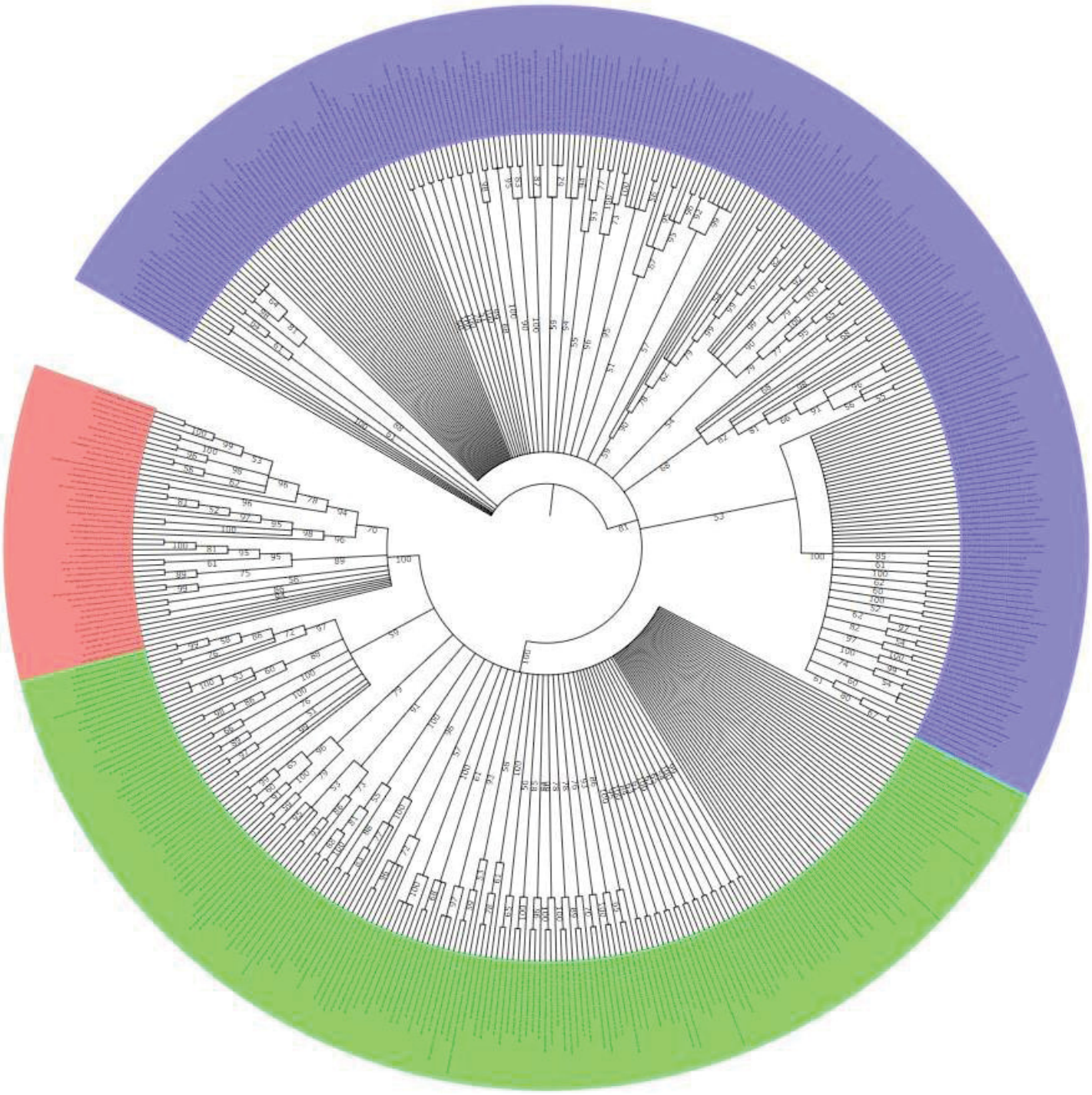
FONTE: O Autor (2018).

Supplementary material, Fig. Fs – Cluster distribution for NtrY group. Number of sequences x Number of Clusters. The values on top of each bar indicates the number of sequences for each cluster size.



FONTE: O Autor (2018).

Supplementary material, Fig. Gs – Evolutionary relationships of Two-Domain NtrX proteins GLOBAL tree. Tree was inferred using the Maximum-Likelihood method. The parameters for analysis was LG substitution model, Gamma distribution and fixed base. ML Optimization Likelihood: -5394.37849. Bootstrap 1000 replicates. iTOL online software was used to reproduce the tree. [26]



FONTE: O Autor (2018).

Supplementary material, Fig. Hs – Evolutionary relationships of Three-Domain NtrX proteins GLOBAL tree. Tree was inferred using the Maximum-Likelihood method. The parameters for analysis was LG substitution model, Gamma distribution and fixed base. ML Optimization Likelihood: -135950.152829. Bootstrap 1000 replicates, only branches with <50% support is collapsed. iTOL online software was used to reproduce the tree. [26]

Supplementary material Table 1s – Summary of taxonomic distribution

GROUPS	HITS	BETA	ALPHA	GAMMA	DELTA	FIRMICUTES	NITROSPIRA
TWO-DOMAIN	**606**	**606**	**X**	**X**	**X**	**X**	**X**
THREE-DOMAIN	**3174**	**45**	**1539**	**173**	**307**	**76**	**34**

FONTE: O Autor (2018).

Supplementary material Table 2s – Two-Domain Substitution model test

Table. Maximum Likelihood fits of 56 different amino acid substitution model

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)
LG+G+I	351	15946.135	13135.759	-6211.395	0.38	0.64
LG+G	350	15994.827	13192.427	-6240.760	n/a	0.33
JTT+G+I	351	16050.555	13240.179	-6263.605	0.38	0.65
WAG+G+I	351	16086.581	13276.206	-6281.618	0.31	0.49
JTT+G	350	16091.484	13289.084	-6289.089	n/a	0.34
WAG+G	350	16107.371	13304.971	-6297.032	n/a	0.32
rtREV+G+I	351	16123.830	13313.454	-6300.243	0.35	0.52
Dayhoff+G+I	351	16129.739	13319.363	-6303.197	0.26	0.39
Dayhoff+G	350	16139.528	13337.128	-6313.111	n/a	0.32
rtREV+G	350	16176.277	13373.876	-6331.485	n/a	0.32
LG+G+I+F	370	16222.384	13260.514	-6254.158	0.38	0.62

FONTE: O Autor (2018).

Note: Models with the lowest BIC scores indicates the more suitable algorithm to use. AICc value, Maximum Likelihood value (lnL), and number of parameters are considered. Non-uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution (+G) with 5 rate categories and by assuming that a certain fraction of sites is evolutionarily invariable (+I). MEGA7 [21].

Supplementary material Table 3s – Three-Domain

Table. Maximum Likelihood fits of 56 different amino acid substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)
LG+G	114	31722.958	30864.212	-15317.164	n/a	0.98
LG+G+I	115	31726.079	30859.817	-15313.950	0.03	1.19
LG+G+F	133	31917.937	30916.434	-15323.935	n/a	0.96
LG+G+I+F	134	31921.461	30912.447	-15320.922	0.03	1.18
WAG+G+I	115	32013.272	31147.009	-15457.546	0.04	1.49
WAG+G	114	32019.600	31160.854	-15465.485	n/a	1.18
rtREV+G+F	133	32053.731	31052.228	-15391.832	n/a	0.95
rtREV+G+I+F	134	32058.781	31049.767	-15389.582	0.03	1.15
rtREV+G	114	32105.097	31246.351	-15508.234	n/a	1.01
rtREV+G+I	115	32109.614	31243.352	-15505.717	0.03	1.21
JTT+G	114	32162.912	31304.166	-15537.141	n/a	1.01

FONTE: O Autor (2018).

Note: Models with the lowest BIC scores indicates the more suitable algorithm to use. AICc value, Maximum Likelihood value (lnL), and number of parameters are considered. Non-uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution (+G) with 5 rate categories and by assuming that a certain fraction of sites is evolutionarily invariable (+I). MEGA7 [21].

Supplementary material Table 4S – NtrY

Table. Maximum Likelihood fits of 56 different amino acid substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)
LG+G	168	24179.121	22956.228	-11307.505	n/a	1.01
LG+G+I	169	24186.811	22956.671	-11306.695	0.02	1.06
LG+G+F	187	24272.459	22911.927	-11265.727	n/a	0.95
LG+G+I+F	188	24281.080	22913.308	-11265.382	0.01	0.98
WAG+G	168	24300.751	23077.859	-11368.320	n/a	1.16
WAG+G+I	169	24309.338	23079.198	-11367.958	0.01	1.22
JTT+G	168	24318.406	23095.513	-11377.147	n/a	1.11
JTT+G+I	169	24326.191	23096.051	-11376.385	0.02	1.17
rtREV+G+F	187	24361.561	23001.030	-11310.278	n/a	0.97
rtREV+G+I+F	188	24370.497	23002.725	-11310.091	0.01	0.99
WAG+G+F	187	24404.540	23044.008	-11331.767	n/a	1.13
WAG+G+I+F	188	24412.735	23044.963	-11331.210	0.01	1.18

FONTE: O Autor (2018).

Note: Models with the lowest BIC scores indicates the more suitable algorithm to use. AICc value, Maximum Likelihood value (lnL), and number of parameters are considered. Non-uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution (+G) with 5 rate categories and by assuming that a certain fraction of sites is evolutionarily invariable (+I). MEGA7 [21].

5 CONCLUSÃO

O sistema de dois componentes NtrYX é essencial para a adaptação e sobrevivência das bactérias nos mais diferentes ambientes. Compreender sua função, sua origem e suas limitações é essencial para melhor entender todo o seu funcionamento biológico. Muito de sua história ainda é desconhecido. Contudo, este trabalho ajudou a visualizar as diferenças significativas do sistema NtrYX entre as bactérias. A distribuição taxonômica dos grupos sugere que a origem desse tão importante mecanismo talvez não seja o mesmo para os diferentes filos. O domínio AAA^+ que apresentou uma forma tão variável, especialmente no motivo GAFTGA para as NtrX, indica que as funções biológicas dentro dessas espécies também são variáveis. Já ao analisar a proteína NtrY, as diferenças não ficaram tão marcantes, o que indica que a função de receber o sinal do meio seja bem parecido para todo o grupo. Este trabalho também conseguiu coletar um número expressivo de homólogos do sistema NtrYX, isto pode auxiliar futuras pesquisas sobre o sistema.

Apesar deste trabalho ter contribuído de forma tão significativa ao melhor entendimento do sistema de dois componentes, algumas analises ainda são essenciais para responder perguntas que ficaram ainda mais relevantes. Uma verificação de vizinhança genética pode trazer mais detalhes sobre a origem desse sistema; assim como uma análise com outras metodologias para confirmar a pressão evolutiva pode indicar se essa proteína sofreu influência do meio para o domínio AAA^+.

REFERÊNCIAS

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W., & LIPMAN, D. J. Basic local alignment search tool. **Journal of Molecular Biology,** *215*(3), 403–10. 1990.

ALBERTS, B. ET AL. **Biologia Molecular da Célula**. 5 eds. *Porto Alegre: Artmed*, 2010.

ATACK ET. AL. Characterization of an *ntrX* Mutant of *Neisseria gonorrhoeae* Reveals a Response Regulator That Controls Expression of Respiratory Enzymes in Oxidase-Positive Proteobacteria**. Journal of Bacteriolog**, 2013.

AKASHI, H.; OSADA N.; TOMOKO O.; Weak Selection and Protein Evolution**.** **Genetics September** *1*, *vol 192*, 2012.

ATTWOOD, T. K. G. A E. N.-E., E, B.-R., T.K. ATTWOOD, A. G., 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**, 1–36. 2011.

BATEMAN, A., MARTIN, M. J., O'DONOVAN, C., MAGRANE, M., APWEILER, R., ALPI, E., ZHANG, J. UniProt: A hub for protein information. **Nucleic Acids Research**, *43*(D1), D204–D212. 2015.

BONATO, PALOMA & ALVES, LYSANGELA & H OSAKI, JULIANA & RIGO, LIU & PEDROSA, FABIO & SOUZA, EMANUEL & ZHANG, NAN & SCHUMACHER, JÖRG & BUCK, MARTIN & WASSEM, ROSELI & CHUBATSU, LEDA. The NtrY/NtrX two-component system is involved in controlling nitrate assimilation in Herbaspirillum seropedicae strain SmR1. **The FEBS journal**. 283. 10.1111/febs.13897. 2016.

CAPRA, E. J., PERCHUK, B. S., SKERKER, J. M., & LAUB, M. T. Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. **Cell**, *150*(1), 222–232. 2012.

CAPRA ET. AL. Evolution of two-component signal transduction systems. **Annual Review of Microbiology**, 2012.

COCK PA, ANTAO T, CHANG JT, CHAPMAN BA, COX CJ, DALKE A, FRIEDBERG I, HAMELRYCK T, KAUFF F, WILCZYNSKI B AND DE HOON MJL Biopython: freely available Python tools for computational molecular biology and bioinformatics. **Bioinformatics**, 25, 1422-1423. 2009.

CARRICA ET. AL. The NtrY/X two-component system of Brucella spp. acts as a redox sensor and regulates the expression of nitrogen respiration enzymes. **Blackwell Publishing Ltd, Molecular Microbiology**, 2012.

DIXON, R., & KAHN, D. Genetic regulation of biological nitrogen fixation. **Nature Reviews Microbiology**, *2*(8), 621–631. 2004.

EDDY, SEAN. A new generation of homology search tools based on probabilistic inference. Genome informatics. **International conference on genome informatics**. 23. 205-11. 2009.

FARRIS, J. S. Methods for computing Wagner trees. **Systematic Zoology** 19, 83-92. 1970.

FAY ET AL. Sequence divergence, functional constraint, and selection in protein evolution. **Annu. Rev. Genom. Hum. Genet**, 2003.

FELSENSTEIN, J. Phylogenies from Molecular Sequences: Inference and Reliability. **Annual Review of Genetics**, *22*(1), 521–565. 1988.

FINN, R. D., BATEMAN, A., CLEMENTS, J., COGGILL, P., EBERHARDT, R. Y., EDDY, S. R., PUNTA, M. Pfam: The protein families database. **Nucleic Acids Research**, *42*(D1), 281–288. 2014.

FU, L., NIU, B., ZHU, Z., WU, S., & LI, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. **Bioinformatics**, *28*(23), 3150–3152. 2012.

ISHIDA ET. AL Identification and characterization of the two-component NtrY/NtrX regulatory system in Azospirillum brasilense. **Braz J Med Biol**, 2002.

LABES, M., & FINAN, T. M. Negative regulation of sigma 54-dependent dctA expression by the transcriptional activator DctD. **Journal of Bacteriology**, *175*(9), 2674–2681. 1993.

MASCHER ET AL. Stimulus perception in bacterial signal-transducing histidine kinases. **Microbiology and Molecular Biology Reviews.** 2006.

NINFA, A. J., & MAGASANIK, B. Covalent modification of the glnG product, NRI, by the glnL product, NRII, regulates the transcription of the glnALG operon in Escherichia coli. **Proceedings of the National Academy of Sciences of the United States of America**, *83*(16), 5909–13.1986.

NEI M & KUMAR S Molecular Evolution and Phylogenetics. **Oxford University Press**, New York. 2000.

SANDERS ET. AL. Phosphorylation site of NtrC, a protein phosphatase whose covalent intermediate activates transcription. **Journal of Bacteriology**, 1992.

SANDERS ET. AL. Identification of the site of phosphorylation of the chemotaxis response regulator protein, CheY. **The Journal of Biological Chemistry**, 1989.

STOCK, J., & DA RE, S. Signal transduction: Response regulators on and off. **Current Biology**, *10*(11), 420–424. 2000.

STEPHENS, Z. D., LEE, S. Y., FAGHRI, F., CAMPBELL, R. H., ZHAI, C., EFRON, M. J., ROBINSON, G. E. Big data: Astronomical or genomical? **PLoS Biology**, *13*(7), 1–11. 2015.

SWOFFORD DL, OLSEN GJ, WADDELL PJ & HILLIS DM. Phylogenetic Inference. In Hiillis DM, Moritz D, and Mable BK, editors, **Molecular Systematics**, pp. 407-514. (**1996**).

SZKLARCZYK, D., FRANCESCHINI, A., WYDER, S., FORSLUND, K., HELLER, D., HUERTA-CEPAS, J.VON MERING, C. STRING v10: Protein-protein interaction networks, integrated over the tree of life. **Nucleic Acids Research**, *43*(D1), D447–D452. 2015.

LEHNINGER, A.L.; NELSON, D.L.; COX, M.M. **Princípios de Bioquímica**. 4. ed. *São Paulo: Sarvier*, 2011.

PAWLOWSKI ET AL. Characterization of a novel Azorhizobium caulinodans ORS 571 two-component regulatory system, NtrY/NtrX, involved in nitrogen fixation and metabolism. **Mol Gen Genet**, 1991.