NILSON ANTÔNIO DA ROCHA COIMBRA

METODOLOGIA COMPUTACIONAL PARA ESTUDO DE GENES COM VIZINHAÇA CONECTADA: ANÁLISE DO CLUSTER *nif*

CURITIBA
2015

NILSON ANTÔNIO DA ROCHA COIMBRA

METODOLOGIA COMPUTACIONAL PARA ESTUDO DE GENES COM VIZINHAÇA CONECTADA: ANÁLISE DO CLUSTER *nif*

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

Orientador: Prof. Dr. Roberto Tadeu Raittz

Co-orientadora: Profa. Dra. Maria Berenice R. Steffens

CURITIBA
2015

COIMBRA, Nilson Antônio da Rocha
C676 Metodologia computacional para estudo de genes com vizinhança conectada: análise do *cluster nif* / Nilson Antônio da Rocha Coimbra. - Curitiba, 2015.
80 f.: il., tabs, grafs.

Orientador: Prof. Dr. Roberto Tadeu Raittz
Co-orientadora: Profa. Dra. Maria Berenice Reynaud Steffens
Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Curso de Pós-Graduação em Bioinformática.
Inclui Bibliografia.

1. Genética - Processamento de dados. 2. Bioinformática. I. Raittz, Roberto Tadeu. II. Steffens, Maria Berenice Reynaud. III.Título. IV. Universidade Federal do Paraná.

CDD 574.0285ter

TERMO DE APROVAÇÃO

NILSON ANTÔNIO DA ROCHA COIMBRA

"Metodologia computacional para estudo de genes com vizinhança conectada: análise do cluster *nif* "

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. Roberto Tadeu Raittz

Coorientador: Profª Drª Maria Berenice Reynaud Steffens

Dr. Vasco Ariston de Carvalho Azevedo
Universidade Federal de Minas Gerais

Dr. Fábio de Oliveira Pedrosa
Universidade Federal do Paraná

Drª. Jeroniza Nunes Marchaukoski
Universidade Federal do Paraná

Curitiba, 17 de julho de 2015

A Atanagildo e Marileide Coimbra, meus pais.

AGRADECIMENTOS

Em gratidão, gostaria de agradecer à todos que colaboraram para o desenvolvimento deste trabalho e minha formação como Mestre em Bioinformática, especialmente:

Ao meu orientador, professor Dr. Roberto Tadeu Raitzz, e as infindáveis xícaras de café que levaram ao desenvolvimento e amadurecimento dessa linha de pesquisa. Sou hoje, um reflexo do brilhante mestre que me ensinou, não somente, as técnicas para minerar dados, mas sobretudo a “pensar fora da caixa”, de forma expansiva e por aprimorar meus raciocínios lógico e crítico.

À minha co-orientadora, professora Dra. Maria Berenice Reynauld Steffens, pela paciência, dedicação e aos ensinamentos e ensaios bioquímicos *in-silico,* que me possibilitaram aprofundar nos fundamentos de fixação biológica do nitrogênio.

Ao professor Dieval Guizelini, por compartilhar suas idéias, conhecimento e por não medir esforços para o avanço das minhas limitações, quando estas não me permitiram avançar.

À banca examinadora deste trabalho, composta pelos professores: Dr. Fábio Oliveira Pedrosa, Dr. Vasco Ariston Avezedo e Dra. Jeroniza Marchaukoski, por todas as considerações aferidas e análise crítica que levaram ao amadurecimento desse documento.

Aos amigos do Laboratório de Bioinformática: Aniele Leão, Sheyla Trefflich, Vinícius Chagas, Marthin Borba, Rodrigo Langowski e principalmente à turma de 2013: MSc. Elisa Terumi, MSc. Rodrigo Menegazzo, MSc. Flávia Costa, MSc. Ana Bandeira, MSc. Eduardo Langowski, Venício Antunes e especialmente ao MSc. Calebe Elias Brim, meu irmão de orientação, pela boa filosofia compartilhada e as inúmeras vezes que ficamos até tarde estudando dentro do laboratório. Jovem, nossa vontade de ir em frente ainda nos levará longe.

Aos professores e as secretárias do Programa de Pós-graduação em Bioinformática e ao Pós-Doc Dr. Vinícius Weiss, pelos ensinamentos e por compartilhar sua visão sobre bioinformática.

A meus pais, Atangildo e Marileide Coimbra, minhas irmãs, Mayara e Nayana, e as minhas sobrinhas, Vitória e Estela, e aos amigos Manu, Leo, Adolfo, Lucas, Gustavo e os “Negos”, pelo apoio incondicional, crucial durante esse período.

Ao INCT de Fixação Biológica do Nitrogênio, pelo profundo aprendizado e oportunidade de colaborar neste projeto.

À Deus, sobre todas as coisas.

RESUMO

A atividade da nitrogenase é vital para a manutenção da vida terrestre e fonte de conhecimento para diversos campos científicos. Mineração de dados é uma técnica que quando bem executada possuí poder natural poder de interpretação de resultados e aplicada aos estudos moleculares facilitam a identificação de padrões em arranjos genômicos, como os óperons. O presente trabalho apresenta uma metodologia para identificação de genes em vizinhança conectada, que possuem funções correlacionadas, através de um estudo de caso do cluster *nif.* Neste trabalho, foi criada uma matriz de ocorrência organismo-cluster e aplicado técnicas de mineração de dados não supervisionada, identificando padrões no comportamento funcional dos genes *nif.* Também apresentamos as ferramentas RAFTS3GROUPS, para agrupamento de sequencias ortólogas e GASUPERCORR, para visualização de coordenadas bi-dimensional de uma matriz multidimensional. Os agrupamentos criados neste trabalho, identificaram grupos de genes com comportamento funcional similar aos confirmados por análises *in vitro* e acrescentam *insights* úteis para re-utilização desse tipo de abordagem na caracterização de função de proteínas com base na correlação de genes. Para validação da metodologia, apresentamos um estudo de caso dos genes *nifT* e *nifZ* sugerindo a função desses genes com envolvimento em estágios iniciais à mobilização de Fe-S, necessários para a formação do FeMo-co. Também foi realizado um estudo de caso para três proteínas hipotéticas que apresentaram maior número de ocorrência e inferida sua função em atividade de ferredoxinas, envolvidas em estágios intermediários à incorporação do FeMo-co. As ferramentas desenvolvidas neste trabalho podem ser re-aplicadas em estudos análogos, para compreensão do comportamento de dados multidimensionais, quanto à organização da informação biológica.

Palavras chave: mineração de dados; fixação biológica do nitrogênio; biologia de sistemas.

ABSTRACT

Nitrogenase activity is vitally important to maintenance life on earth and knowledge source for many scientific fields. Data mining is powerful technique for results interpretation and when applied through molecular studies provides genomics patterns identification, like operons function structures. This works describes a methodology to identification of function through neighborhood connected gene analysis using cluster as nif study of case. In this works, we built a organism-cluster occurrence matrix and applied unsupervised data mining techniques to identify functional patterns of nif genes. We present RAFTS3GROUPS, a clustering orthologous sequences tool and GASUPERCORR function, an bi-dimensional coordinates visualization from high dimensional matrix. Clustering results identify functional activity gene groups with similar *in vitro* analysis and provides insights to infeer protein function by correlated genes. Metodology validation was conducted from a genetic analysis study case among *nifT* and *nifZ* genes and we suggest function into preliminary stages of Fe-S captation to FeMo-co formation. Also, we analysed three hypotetical protein and suggested like ferredoxin activity putative involved on intermediated stages of FeMo-co incorporation. Tools developed in this works could be re-applied on highdimensional data matrix from other kind of studies, also to understand the genetic information of neighbourhood connected genes.

Keywords: data mining, biological nitrogen fixation; system biology

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

4Fe-4S	Quatro ferro-enxofre
8Fe-7S	Oito ferro sete enxofre
ATP	Adenosina trifosfato
BCOM	Binary co-ocurrencematrix
BLAST	Basic Local Alignment Sequence Tool
BLOSUM	Blocks of Amino Acid Substitution
CDS	Sequencia codificante
CO	Monóxido de carbono
DCBD	Descoberta de conhecimento em base de dados
DNA	Ácido Desoxiribonucléico
FBN	Fixação biológica do nitrogênio
FeMo-co	Co-fator Ferro Molibdênio
H2	Dihidrogênio
HGT	Horizontal gene transfer (Transferência horizonta
ID-ordem	Ordem
kDa	Kilodaltons
KDD	Knowledge Discovery database
Kpb	Kilo pares de base
MoFe	Ferro Molibdênio
N2	Dinitrogênio
NH3	Amônia
NIF	Genes da fixação biológica do nitrogênio
ORF	Open Read Frame
PAM	Point Accepted Mutation
PPG	Programa de pós-graduação
ProClaT	Protein Classifier Tool
RAFTS3	Rapid Alignment Free Tool for Sequence Similari
RNA	Ácido ribonucleico
RNA-SEQ	Técnica de sequenciamento de Cdna em larga es
rRNA	RNA ribossomal
tRNA	RNA transportador
uda	Unidade de distância arbitrária

SUMÁRIO

1 INTRODUÇÃO

Na década corrente, as pesquisas genômicas impulsionam a expansão do conhecimento sobre a organização do material genético dos seres vivos.

O surgimento das plataformas de sequenciamento de nova geração, junto com os primeiros resultados do projeto genoma humano, corroboram para o aumento dos dados biológicos armazenados em base de dados e elucidaram diversas perguntas sobre a organização do material genético na dupla fita. Entretanto, adicionaram novas questões quanto a evolução dos diferentes processos de interações entre proteína-proteína nos organismos (CHUANG, 2010).

Ao longo dos anos, a bioinformática vem contribuindo com a instrumentação para o armazenamento dos dados gerados pelas *ômicas* e com o desenvolvimento de metodologias para a manipulação, análise e compreensão desse tipo de dados (YU, 2005).

Como ciência, a bioinformática analisa a complexidade dos sistemas biológicos sob diferentes perspectivas, da análise à extração de resultados, buscando responder diferentes questões sobre a plasticidade e interações dos mecanismos biológicos.

Bioinformática subdividi-se quanto a sua área de atuação: a bioinformática funcional, voltada à análises de estruturas (primárias, secundárias, terciária, quaternária) e bioinformática estrutural, voltada para análises de interações proteína-proteína, redes complexas e integração de dados dos diferentes processos biológicos, comumente referida como biologia de sistemas.

Biologia de sistemas é uma área emergente e tenta representar o comportamento dos sistemas biológicos de um organismo através da análise integrada dos dados de diferentes *ômicas*. O vasto campo de atuação possibilita a exploração e combinação de conceitos dás áreas de interesse (CHUANG, 2010), levando ao desenvolvimento de novas abordagens para extração de resultados, que podem ser aplicado na identificação de biomarcadores.

1.1 JUSTIFICATIVA

O principal desafio da bioinformática está em analisar e padronizar a informação gerada diariamente, sendo necessário o desenvolvimento de metodologias que facilitem a compreensão do significado biológico existente dentro dados.

Em genomas procariotos, os genes são distribuídos de forma randômica ou em vizinhanças conectadas (operons), possuindo, ou não, atividades correlacionadas à sua função no metabolismo dos organismos (JACOB & MONOD, 1961).

Para a determinação da função exata de uma proteína são necessários inúmeros experimentos *in-vitro* e a evolução das interações entre proteína-proteína, impactam diretamente no funcionamento dos mecanismos biológico, tornando essa uma tarefa árdua de ser analisada (YU, 2005).

O desenvolvimento das tecnologias de sequenciamento de nova geração possibilitou a análise global dos organismos e a descoberta de genes, em regiões adjacentes no genoma, que atuam em um mesmo metabolismo (MEDEMA & FISCHBACH, 2015; YI, 2007).

Das tecnologias existentes, o RNA-seq é utilizado, também, para a identificar genes que atuam em diferentes condições metabólicas. Concomitante com técnicas de clusterização, os resultados dessa combinação é consolidado pela classificação dos genes que foram expressos nas diferentes condições que foram analisados (WANG, 2013; KURUVILLA, 2002; BRAZAM, 2000).

A aplicação de clusterização em dados de RNA-seq demonstra alta confiabilidade nos resultados (WANG, 2013). Extrapolar esse tipo de conceito aos dados de sequenciamento de DNA, permite a criação de metodologias para mineração de genes em vizinhança conectada, objetivando encontrar relações de grupos proteicos com base no seu envolvimento funcional. Assim, podendo inferir a função, principalmente de proteínas hipotéticas, com base na atividade de genes correlacionados.

1.2 OBJETIVOS

1.2.1 OBJETIVO GERAL

Desenvolvimento de uma metodologia *in sílico* para o estudo funcional de genes em vizinhança conectada e análise do cluster *nif.*

1.2.2 OBJETIVOS ESPECÍFICOS

1. Estudar o cluster *nif* dos organismos diazotrófos;
2. Extrair a região genômica contendo o cluster *nif* de diferentes organismos diazotrófos de forma automatizada;
3. Identificar as ORF e anotar o conteúdo genômico;
4. Identificar genes ortólogos aplicando o RAFTS3;
5. Desenvolver rotinas computacionais para execução do método;
6. Descobrir associação entre elementos correlacionados por agrupamento hierárquico
7. Interpretar os resultados e validar da metodologia proposta com base na literatura.

2 REVISÃO BIBLIOGRAFICA

2.1 MINERAÇÃO DE DADOS

Na bioinformática, diante do volume de dados manipulados diariamente, mineração de dados é uma tarefa crucial para análise e interpretação dos dados.

Mineração de dados ou *Data mining* é umas das etapas da descoberta de conhecimento em base de dados (DCBD), do inglês *Knowledge Discovery in Databases* ou KDD e foi desenvolvida para manipulação da informação diante do acumulo de dados armazenado em bancos de dados físicos (GOLDSCHIMIDT & PASSOS, 2005).

A técnica, é um processo analítico e sistemático, projetada para explorar padrões e relacionamentos entre as variáveis armazenadas em bancos de dados. O objetivo é extrair informações e reutilizá-las em processos de tomadas de decisões (DONI, 2004).

Hsu (2006) fez uma analogia à mineração dos dados com a mineração de pedras preciosas onde: *"assim como o garimpeiro revela o ouro existente no solo, o profissional que realiza data mining revela conhecimento ao processar grandes quantidades de dados".*

Mineração de dados é uma técnica multidisciplinar e integra conceitos e aplicações de áreas como estatística, aprendizado de máquina, reconhecimento de padrões e banco de dados (GOLDSCHIMITCH & PASSOS, 2005).

A execução da técnica generaliza-se em três etapas:

1) **Pré-processamento:** captação, tratamento e organização dos dados para a aplicação das técnicas de mineração de dados
2) **Mineração de dados:** aplicação de algoritmos para extração ou classificação de padrões. Sendo a escolha do algoritmo feita de acordo com o objetivo do KDD.
3) **Pós-processamento:** tratamento, interpretação e avaliação dos resultados,

facilitando o entendimento do conhecimento descoberto.

Em todas essas etapas o papel do analista é de fundamental importância para compreensão da informação manipulada e quando aplicados aos dados genômicos, cabe aos profissionais da áreas biológicas e bioinformática analisar os dados à revelar soluções aos diferentes tipos de questões levantadas.

Dentre as inúmeras técnicas de mineração de dados, a clusterização de dados começou a ser exploradas em bioinformática com os sucesso obtidos aos experimentos de sequenciamento de genes expressos em larga escala (RNA-seq). A ideia desse tipo de técnica é agrupar dados com base em caraterísticas comum entre si e no caso do RNA-seq, identificar genes expressos em mesma condição (WANG, 2013 ; KURUVILLA, 2002; BRAZAM, 2000; EISENBERG, 2000).

2.1.1 CLUSTERIZAÇÃO DE DADOS

Clusterização de dados, ou análise de agrupamento, é um grupo de técnicas multivariadas cuja principal finalidade é classificar os objetos em grupos com base em um grupo de características entre si (ZHENG, 2012; HAIR et al, 2009; ZHANG, 2004).

A execução da técnica independe de conhecimento estatístico. Em primeira instância é calculado a similaridade entre os objetos e essa determinará quão semelhantes, ou diferentes, serão esses objetos entre si. (JASKOWIAK, 2014). Os objetos podem ser combinados em grupos pela aplicação de algoritmos hierárquicos, sendo esses aglomerativos ou divisivos, e algoritmos não-hierárquicos.

Nos algoritmos hierárquicos, todos os objetos do grupo são inicialmente considerados como um possível agrupamento e pela medida similaridade, os objetos serão combinados em grupos, reduzindo a quantidade de objetos inicialmente. Esse processo é repetido interativamente de forma que todos os objetos fiquem contidos em apenas um grupo. O resultado final desse tipo de método é exibido na construção de um dendrograma e avaliado pela homogeneidade interna, dentro dos grupos e heterogeneidade externa entre os grupos criados (HAIR et al, 2009).

Os algoritmos não-hierárquicos destinam os objetos em grupos a partir de um número “k” de agrupamentos, especificado previamente, e não há a construção de dendrograma. O agrupamento é iniciado pela escolha de uma semente de agrupamento (k-mers) estipulada pelo analista ou gerada de forma aleatória e seguido a diante pela escolha do tipo de algoritmo de agrupamento, aplicado de forma sequencial, ou paralela ou otimizada (HAIR et al, 2009).

Independente da escolha do algoritmo, a decisão a cerca do número de agrupamentos final é a principal questão do resultado (HANDL, 2005). Apesar de não haver um procedimento padrão para a execução da técnica, quando bem aplicada tem potencial de revelar estruturas dentro dos dados e que não poderiam ser descobertas por outros meios. A análise de agrupamento fornece um método empírico para realizar uma das tarefa natural do ser humano, a classificação (GOLDSCHIDTH & PASSOS, 2005; DONI, 2004).

2.2 ORGANIZAÇÃO DE GENOMAS PROCARIOTOS E HOMOLOGIA DE SEQUÊNCIAS

O gene é a unidade de informação genética que contem as bases de DNA que serão transcritas em uma molécula de RNA. O tamanho de um gene varia de acordo com o tamanho do produto contido na sua sequencia codificante (CDS), e pode ser traduzido tanto em proteínas (cadeia polipeptídica de aminoácidos), quanto em diferentes tipos de RNA (rRNA, tRNA, entre outros) (LEWIN, 2000).

Na estrutura básica de um gene, à montante ou à jusante, são encontradas as sequências reguladoras de função, promotores e terminadores (SNUSTAD E SIMMONS, 2001). O promotor é a região onde a RNA-polimerase se associa ativando a transcrição do gene e fica situado em uma região 5' à montante do gene. Já o terminador, fica situado em uma região 3' do gene e contém a sequência que indicará a dissociação da RNA-polimerase e consequentemente o fim da transcrição do gene (LEWIN, 2000).

Em genomas procariotos, os genes são distribuídos de forma randômica ou

em vizinhanças conectadas, operôns. Os operôns contém um ou mais genes adjacentes que são transcritos em um mesmo mRNA policistrônico, compartilhando mesmo promotor e terminador e atuam em uma ou mais vias metabólicas (JACOB & MONOD, 1961).

Genes de organismos procariotos que possuem similaridade aparente entre as sequências de nucleotídeos e seus resíduos de aminoácidos, genericamente referidos como homólogos. Homologia é a existência de um ancestral comum em um par de estruturas ou genes em diferentes espécies e subdividi-se quanto ao evento evolutivo como especiação, duplicação e transferência horizontal lateral (FITCH, 2000).

Eventos de especiação geram ortólogos (SONNHAMMER & KOONIN, 2002). Ortólogos são genes de diferentes espécies que originaram por descendência vertical de um único gene ancestral comum e sempre, mas nem sempre, tem a mesma função. Pela quantidade de genes ortólogos em genomas de diferentes espécies é possível inferir o grau evolutivo entre essas espécies (LEWIN, 2000).

Eventos de duplicação geram parálogos (FITCH, 1970) que em um primeiro estágio geram cópias idênticas do gene envolvido e dependendo das pressões seletivas podem ocorrer duas situações: 1) uma cópia do gene duplicado supre a demanda do produto original e a outra cópia pode divergir em sequência dando origem a um gene parálogo, codificando um produto relacionado com funcionalidade distinta, porém não idêntico ao original, ou 2) a divergência progressiva associada a falta de pressão seletiva levando a manutenção de uma das cópias não funcional do gene caracterizado-o como um pseudo-gene (ZAHA, 2014).

A transferência horizontal gênica (HGT) é reconhecida como um processo marcante na evolução de organismos procariotos (BROWN, 2003) e mediada por elementos genéticos móveis capazes de levar ao surgimento de genes ortólogos (KUZNIAR et al., 2008). Neste caso duas ou mais espécies podem adquirir um mesmo gene de forma independente a partir de eventos de HGT da espécie de origem do gene. Os genes ortólogos adquiridos dessa forma são chamados de genes xenólogos (ZAHA, 2014).

O grau de conservação para que duas sequencias sejam consideradas homólogas é quantitativo e arbitrário. Na literatura, homologia entre sequências é

comumente considerada com mínimo de 30% de identidade e 60% de similaridade, usando as matrizes de alinhamento PAM ou BLOSUM, implementadas nos pacotes do programa BLAST. Genes com graus de identidade próximos a 100% de identidade são considerados cópias, pois se assume que codificam produtos sem distinção funcional.

2.3 FIXAÇÃO BIOLÓGICA DO NITROGÊNIO E OS DIAZOTRÓFOS

Na natureza, a fixação de nitrogênio e realizada em três formas: por fenômenos físicos, como relâmpagos e faíscas; por métodos químicos, utilizados no processo industrial Haber-Bosch na produção de fertilizantes nitrogenados pela síntese de amoníaco e pela forma biológica, denominada fixação biológica do nitrogênio (FBN) promovida pela ação dos microrganismos chamados diazotrófos presentes no solo ou/em nódulos de raízes de leguminosas, cujo o produto final a amônia (POSTGATE, 1982).

É estimado que o processo de fixação biológica do nitrogênio contribua com 60% em todo o nitrogênio fixado no planeta, dimensionando um alto impacto na atividade agro-ecológica mundial, pois trata-se de uma alternativa natural aos fertilizantes nitrogenados que causam danos ao meio ambiente como a poluição ambiental, eutrofização de rios e lagos, acidificação do solo e bem como reduz a emissão de dióxidos de carbonos decorrente do processo de fabricação dos mesmos (SHIRIDAR, 2012).

Os diazotrófos captam o nitrogênio atmosférico (N_2) e o reduzem a formas assimiláveis de amônia (NH_3) através da catalise pela enzima nitrogenase, um complexo metaloproteico altamente conservado, codificado por genes localizado no cromossomo ou em plasmídeos (PRAKASH; SCHUPEROORT & NUTI, 1981). A capacidade de fixar o nitrogênio, encontrado em abundância na atmosfera na forma inerte do gás dinitrogênio, somente é encontrada nos organismos, dos domínios *Bacteria* e *Archeae* (YOUNG, 1992).

Os diazotrófos se adaptaram a diferentes ambientes, através de simbiose ou

vida livre. No meio ambiente, esses organismos são encontrados no solo, em amplo espectro de temperatura; em ambientes aquáticos de água-doce ou salobra; e em ambientes intracelulares, associados às gramíneas pela capacidade endofítica, apresentada por algumas espécies (POSTGATE, 1982).

2.4 ESTRUTURA E MONTAGEM DA NITROGENASE

A reação da nitrogenase (1) requer um alto custo energético (NUNES, 2003). A formação de uma molécula de dihidrogênio (H_2), está associada à redução de uma molécula de dinitrogênio (N_2) e duas moléculas de ATP (adenosina trifosfato), por elétrons transferidos.

$$N_2 + 8e^- + 16\ MgATP + 8H^+ \rightarrow 2NH_3 + H_2 + 16MgADP + 16P^- \quad (1)$$

Na ausência de dinitrogênio, a nitrogenase é capaz de reduzir prótons a dihidrogênio e esta capacidade pode estar relacionada a processos e vias que ocorrem paralelamente a transferência de elétrons (HOFFMAN et al, 2014).

A nitrogenase é capaz de reduzir substratos similares ao N_2 (SPATZAL, 2014), que apresentem duplas ou triplas ligações, sendo a única enzima que consegue quebrar as ligações triplas do gás dinitrogênio e reduzir monóxido de carbono (CO) a hidrocarbonetos sob a mesma reação (RIBBE, 2013).

A nitrogenase (Figura 01) em si, é um complexo redox-ativo hidrolisante de ATP, composto por duas metallo-proteinas: a dinitrogenase e a dinitrogenase redutase. Esta enzima é extremamente sensível ao oxigênio que a inativa irrevessivelmente (HOFFMAN et al, 2014).

A Dinitrogenase, ou proteína MoFe (Ferro-Molibdênio) ou componente I, é um sítio de redução de substrato necessário para a redução do dinitrogênio composto por um tetrâmetro de ~240kDA com duas subunidades não idênticas α2β2: a

subunidade α, produto do gene nifD e subunidade β, produto dos gene *nif*K (HOFFMAN et al, 2014).

As duas subunidades são compostas por dois grupos metálicos, o P-cluster e o FeMo-co (co-fator Ferro Molibdênio). P-cluster está localizado na interface de cada sub-unidade e acredita-se que esse agrupamento seja um estado intermediário à transferência de elétrons do componente I ao componente II. FeMo-co é localizado na parte interna da proteína MoFe e possui estrutura única, não identificada em nenhuma outra metalloproteina (DOS SANTOS, 2004).

Cada agrupamento é composto por sítios de ligação de homocitrato e redução do substrato, necessário para a quebra da tripla ligação do nitrogênio gasoso (BODY, 2013). P-cluster e FeMo-co possuem estrutura similar, porém o P-cluster, pode apresentar forma rômbica 4Fe-4S ou cúbica 8Fe-7S (DOS SANTOS, 2004).

A dinitrogenase redutase, proteína Ferro ou componente II, é um dímero de ~70 kDa, produto do gene *nif*H, contém um agrupamento metálico 4Fe-4S, entre os monômero e os resíduos de cisteína, e dois sítios ligantes de ATP, na interface da subunidade. A proteína Ferro é o doador de elétrons obrigatório para a proteína MoFe. Os elétrons são transferidos do agrupamento metálico 4Fe-4S da proteína Fe para o P-cluster da proteína MoFe e encadeados na formação do FeMo-co para a redução do substrato. A função da proteína Ferro é descrita tanto na formação do FeMo-co quanto para a maturação da proteína Ferro (RUBIO & LUDDEN, 2008; RUBIO & LUDDEN, 2005).

As principais discussões sobre a FBN referem-se aos diferentes mecanismos envolvidos para a formação do(s) co-fator(res) necessários para a ativação da nitrogenase (BODY et al, 2013, BODY et al, 2012).

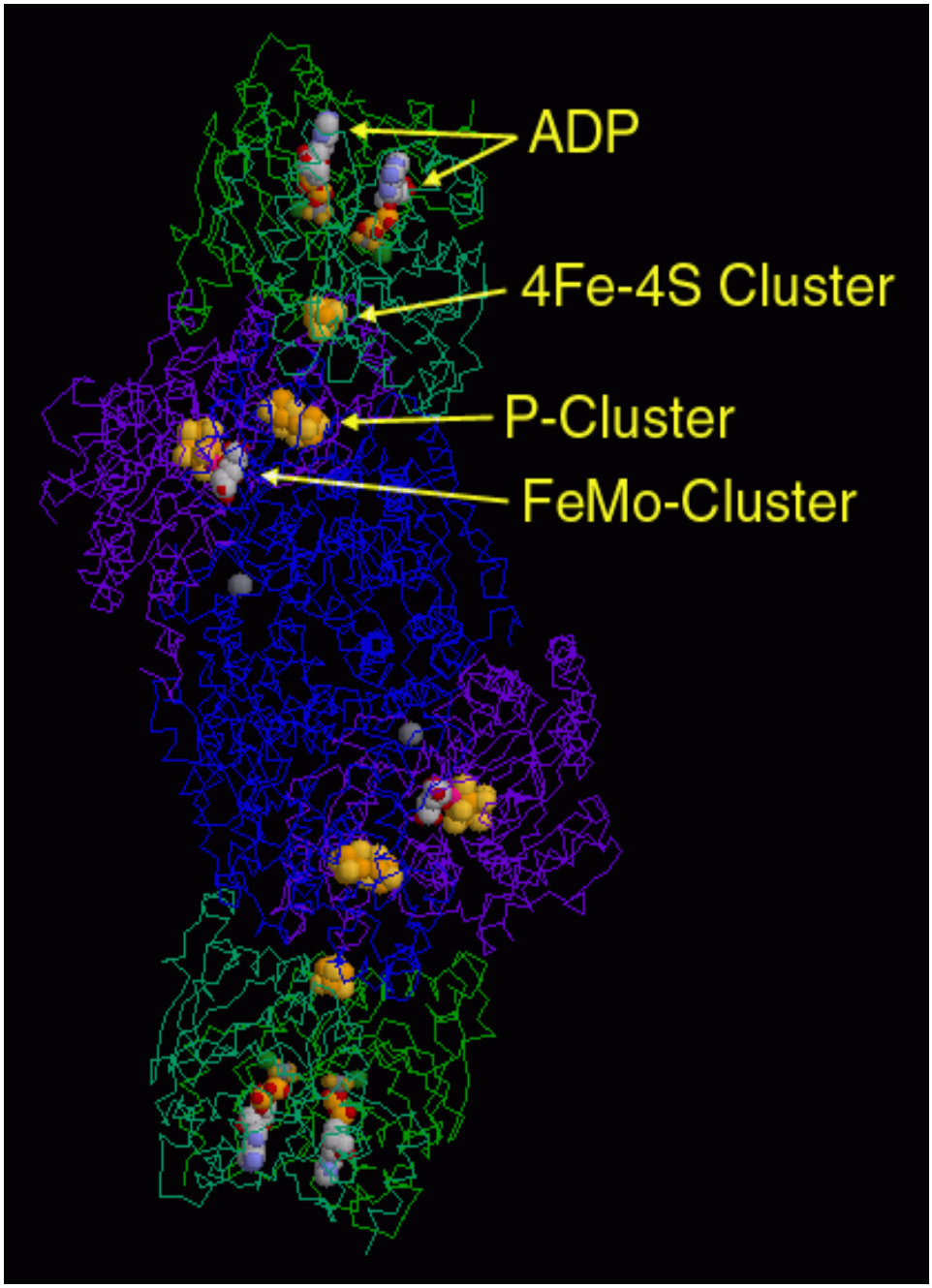


FIGURA 1- ENZIMA NITROGENASE. Em verde, a dinitrogenase, gene nifH e o ligante metálico 4Fe-S dependente de ATP, doador de elétrons para o P-cluster e FeMo-co. Em azul e roxo, as subunidades (α e β), genes nifD e nifK e o FeMo-co e P-cluster, sitios de redução do substrato.

FONTE: PDB Data Bank (2015)

2.5 PERFIL GÊNICO DO CLUSTER NIF

Os genes responsáveis pela formação da nitrogenase são denominados *nif* e foram inicialmente descritos em um conjunto de 19 genes dispostos em uma região de 24 Kbp (kilo pares de base) organizados em 8 operôns no genoma de *Klebsiella pneumoniae,* envolvidos na biossíntese e transcrição da nitrogenase (MERRICK, 1992a).

Análises bioquímicas já consolidadas, descrevem o gene *nifH* como a proteína dinitrogenase reductase; o gene *nifD*, como dinitrogenase ferro-molibdênio sub-unidade alfa e o gene *nifK* como dinitrogenase ferro-molibdênio sub-unidade beta. Também é descrita a função dos genes *nifE*, *nifN*, *nifB*, *nifX*, *nifQ*, *nifV*, *nifH* envolvidos em diferentes níveis na biossíntese e inserção do FeMo-co na dinitrogenase. O gene *nifY*, codifica uma chaperona e junto com o *nifW* auxiliam na maturação do complexo MoFe (RUBIO E LUDDEN, 2005).

Os genes *nifS* e *nifU* são necessários para a mobilização de ferro e enxofre para a ligação do substrato ao co-fator da dinitrogenase. O gene *nifM* foi caracterizado como necessário para a maturação da proteína Fe, e alguns organismos utilizam o maquinário *nifA-nifL* como reguladores positivo e negativo que controlam a atividade da nitrogenase durante variações da concentração de nitrogênio e oxigênio do meio (DIXON, 2004).

A atividade inibitória de *nifL* sobre o ativador de transcrição *nifA*, é mediada por proteínas da classe PII, transdutoras dos níveis de ions de amônio (HUERGO, 2006).

Em *K. pneumoniae* o gene *nifF* codifica uma flavodoxina, descrita como doadora fisiológica de elétrons para o cluster metálico presente no componente I da nitrogenase. E o gene *nifJ* codifica a proteína piruvato de flavodoxina, uma proteína da classe das ferredoxinas presente na família oxidoredutase, com atividade de transferência de elétrons para o gene *nif*F (TEMME, 2012).

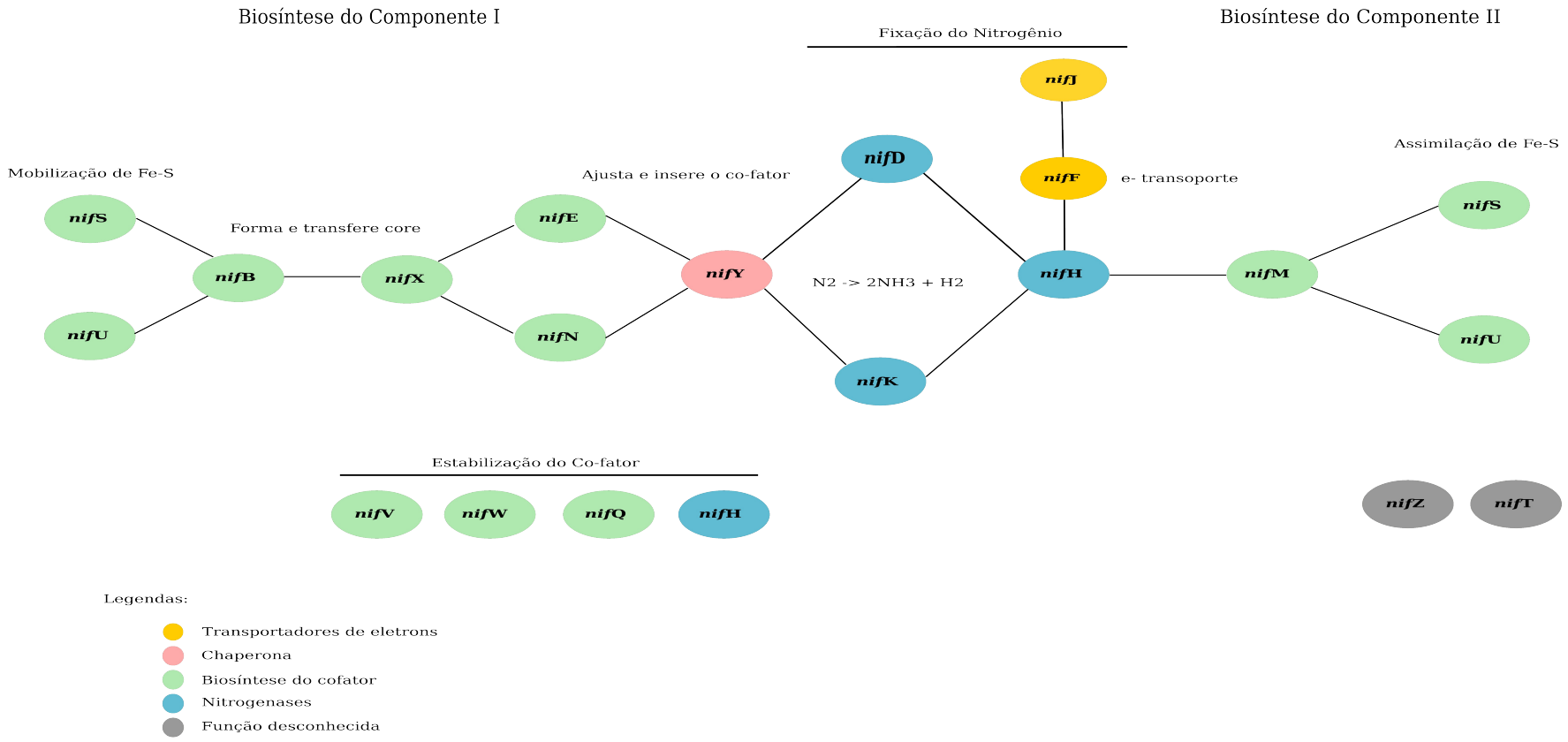


FIGURA 2 - MAPA DE INTERAÇÃO DOS GENES DA FIXAÇÃO BIOLÓGICA DO NITROGÊNIO. Em verde, os *nif* caracterizados para a formação do co-fator necessário para atividade da nitrogenase. Em amarelo, os genes caracterizados como transportadores de elétrons obrigatório para a dinitrogenase redutase, *nifH*. Em azul os genes caracterizados para a formação estrutural da enzima. Em cinza, genes nif com função desconhecida.

FONTE: O autor (2015)

3 MATERIAS E MÉTODOS

Esta sessão destina-se a apresentação dos materiais aplicados e explanação do fluxo metodológico desenvolvido para a proposta deste trabalho.

Ao longo da existência do PPG em Bioinformática da Universidade Federal do Paraná, diversas ferramentas voltadas à análises genômicas foram desenvolvidas sob a orientação do Prof. Roberto Raittz, chefe e mentor do Laboratório de Inteligência Artificial aplicada à Bioinformática, localizado no Setor de Educação Profissional e Tecnológica da UFPR. As ferramentas aplicadas nesse trabalho são frutos das dissertações desenvolvidas no laboratório e objetivam a mineração de dados de dados genômico. Para a obtenção dos objetivos traçados no trabalho, exploramos o uso das ferramentas criadas em laboratório e construída a metodologia aplicada no estudo do cluster *nif*.

Um pacote de *scritps* foi desenvolvido em linguagem MATLAB e seus códigos-fonte encontram-se nos anexos dessa dissertação.

3.1 MATERIAIS

3.1.1 ORGANISMOS

Os organismos utilizados nesse trabalho foram adquiridos do trabalho de Dos Santos e colaboradores (2012), no qual os autores propõem um modelo de classificação *in-sílico* para organismos diazotrófos, baseando-se na identificação de um conjunto mínimo de seis genes: os *nifHDK* (responsáveis pela formação estrutural da enzima) e os *nifENB* (responsáveis pela formação de parte da atividade

catalítica). No trabalho, os autores confirmam oitenta organismos pela identificação dos genes minimos e classificam 67 outros organismos como possíveis diazotrófos.

Neste trabalho, utilizamos os oitenta organismos confirmados pelos autores.

3.1.2 MATLAB - MATRIX LABORATORY

O MATLAB é uma linguagem de programação alto nível e ambiente interativo, para computação numérica, visualização e análise de dados. A sintaxe fácil da linguagem permite a criação de ferramentas e a construção de funções para explorar e gerar soluções de forma rápida e fácil para diferentes áreas.

O software é utilizado em diversos campos científicos: para tratamento de imagem e som, engenharia aeroespacial e, em decorrência da crescente demanda dos últimos anos, a bioinformática (MATHWORKS, 2015).

O software foi utilizado para desenvolvimento e análise dos dados durante todo este trabalho.

3.1.3 JGBPARSER

O JGBParser é um analisador para arquivos texto em formato GenBank (*.gbk*). O software foi desenvolvido e implementado em JAVA, e utiliza técnicas otimizadas para busca em banco de dados visando a identificação de termos existentes arquivos de anotação.

A ferramenta separa em características, os identificadores existentes nos arquivos texto, permitindo a análise de forma parcial, e/ou incremental e/ou por sessões pré-definida existentes nos arquivos de anotação (GUIZELINNI, 2010).

Para a utilização do software no MATLAB foi desenvolvida a função

LoadJGBParser.m (ANEXO I – Load JGBParser.m). O software está disponível sob a licença GNU no seguinte endereço eletrônico: http://sourceforge.net/projects/jgbparser/

3.1.4 SILA

O SILA é um sistema de anotação genômica de alto desempenho, desenvolvido em linguagem MATLAB e utiliza abordagem heurística para avaliar a similaridade entre sequências. O SILA combina o resultado dos preditores de ORF(Open Read Frame): Prodigal - *Prokaryotic Dynamic Programming Genefinding Algorithm* (HYATT, 2010) e HGF - *Hybrid Gene Finder*, comparando as sequencias de aminoácidos das ORF preditas, em três bancos de dados de sequências de proteínas: o NCBI Non-Redundant, PFAM e COG, ambos para anotação de produto. O SILA demonstra resultados comparáveis ao estado da arte na anotação genômica: altas taxas de acerto na predição ORF e identificação de genes aos bancos de dados (VIALLE, 2013).

O serviço de anotação foi desenvolvido por Ricardo Vialle (2013) e utilizado nesse trabalho na sua versão local, na etapa para anotação de produto nos das sequencias dos arquivos de anotação (ANEXO II e ANEXO III). Para acesso público, o software encontra-se no endereço eletrônico: (http://200.236.3.34/SILA/login.jsp).

3.1.5 RAFTS3GROUPS

O RAFTS3GROUPS (ANEXO IV) é uma aplicação da ferramenta RAFTS3 (VIALLE, 2013) para agrupamento de sequências homólogas. O *script* foi desenvolvido em conjunto ao orientador deste trabalho e realiza o agrupamento de

sequencias pelo valor de similaridade "*self-score*" computados pelo software RAFTS3.

O *script* recebe de entrada um arquivo multifasta, contendo o conjunto de proteínas a serem agrupados. Para criação dos *clusters* ao arquivo de entrada é criado um banco de dados e feita a consulta através do RAFTS3. Cada proteína no arquivo de entrada é avaliada inicialmente como um possível grupo, e o resultado da consulta ao banco avalia a similaridade entre as proteínas pelo valor de *self-score*. Os grupos de ortólogos são formados com base na verificação do número de ocorrências reconhecidas pelo mínimo de similaridade avaliada (*self-score* = 0.5).

Para interpretação dos resultados gerados pelo RAFTS3GROUP foi criada desenvolvida a função *agruparafts.m* (ANEXO V – Função Agruparafts.m), que faz a verificação dos grupos com mínimo de duas sequencias, exportando-os em arquivos indexados em formato fasta. É também exportado um arquivo tabular, contendo a ordem de indexação dos grupos criados, o produto anotado pelo SILA e o número de ocorrências em cada grupo.

3.1.6 ProClaT - Protein Classifier Tool

O ProClaT (TERUMI et al, 2015 – No prelo) é uma ferramenta de classificação de proteínas. O software utiliza uma sequencia *consensus* e características físico-químicas, para classificação de proteínas através de aprendizado de máquina e redes neurais artificiais. A ferramenta foi aplicada para o reconhecimento dos nif*HDKENB* posteriormente ao reconhecimento de ortólogos usando o RAFTS3GROUP. A ferramenta está disponível público no endereço eletrônico: http://sourceforge.net/projects/proclat/

3.1.7 PROSITE

O PROSITE é um banco de dados que consiste em uma documentação com a descrição de domínios, famílias e sítios funcionais de proteínas. O banco é complementado por uma coletânea de regras (Prorule), que adiciona informações sobre a função e estruturas crítica dos aminoácidos. A versão mais recente do banco, datada no dia 03 de março de 2015, possui 1718 sequencias documentadas, 1308 padrões de domínio e 1109 perfis de proteínas e 1108 ProRules (SIGRIST et al, 2012).

O banco foi utilizado para enriquecer as informações dos grupos gerados pelo RAFTS3GROUPS afim de identificar grupos que compartilham mesmo domínio funcional.

3.1.8 PFAM

PFAM é um banco de dados de família de proteínas. Cada família dentro do PFAM possui um modelo estatístico baseado em cadeias de Markov que é treinado usando alinhamento de sequencias. O PFAM é mantido pelo Instituto Europeu de Bioinformática EMBL-EBI, e até a escrita desse trabalho possuía informações sobre 14.831 famílias de proteínas (FINN et al, 2014).

O banco foi utilizado para enriquecer as informações dos grupos gerados pelo RAFTS3GROUPS afim de identificar grupos que compartilham mesma família proteica.

3.1.9 VETORES E MATRIZES

Vetor algébrico, é um arranjo unidimensional (1D) ordenando *n* números reais. Vetores são infinitamente representados e computacionalmente dependem da quantidade de blocos de memórias disponíveis na memória do computador.

Matriz é um arranjo bi-dimensional (2D) dos números reais. Matrizes são usualmente compostas por linhas e colunas, onde cada linha e cada coluna podem ser descritas como vetores.

Neste trabalho, nós criamos uma matriz de ocorrência (ANEXO VI) para armazenar os dados de proteínas existentes no conjunto de ortólogos, remanescentes após a etapa de enriquecimento de informação, dentro dos organismos analisados.

3.1.10 CLUSTERIZAÇÃO HIERARQUICA NÃO SUPERVISIONADA - CLUSTERGRAM.m

O MATLAB dispõe de uma biblioteca bastante abrangente de funções matemáticas, geração de gráficos e manipulação de dados que auxiliam o trabalho do analista. A *toolbox* de bioinformática contém uma série de funções específicas e aplicativos para dados NGS. A *toolbox* também é composta por algoritmos estatísticos, analises de microarray e ontologia de sequencias.

Dentre as funções da *toolbox,* a função c*lustergram.m* faz a análise hierárquica de um conjuntos de dados existentes em uma matriz. Em um primeiro passo, é calculado a similaridade (*pdist*) entre pares de objetos usando as medidas: distância euclidiana, correlação de Pearson, distância cityblock, distância mahalanobis entre outras.

Uma vez calculado a similaridade entre os objetos é feito uma série de sucessivos agrupamentos (linkage), onde os elementos serão agregados de acordo com a similaridade calculada no passo anterior.

Como resultado da função *clustergram.m,* os grupos são representados por um diagrama bi-dimensional combinados o dendrograma do agrupamento e um

mapa de calor (*heatmap*) entre os valores de similaridade entre os os objetos. O *heatmap* gerado pela função é avaliado pela força da similaridade entre os dados agrupados. Quanto mais intenso ao vermelho, mais os dados tendem a ser relacionados.

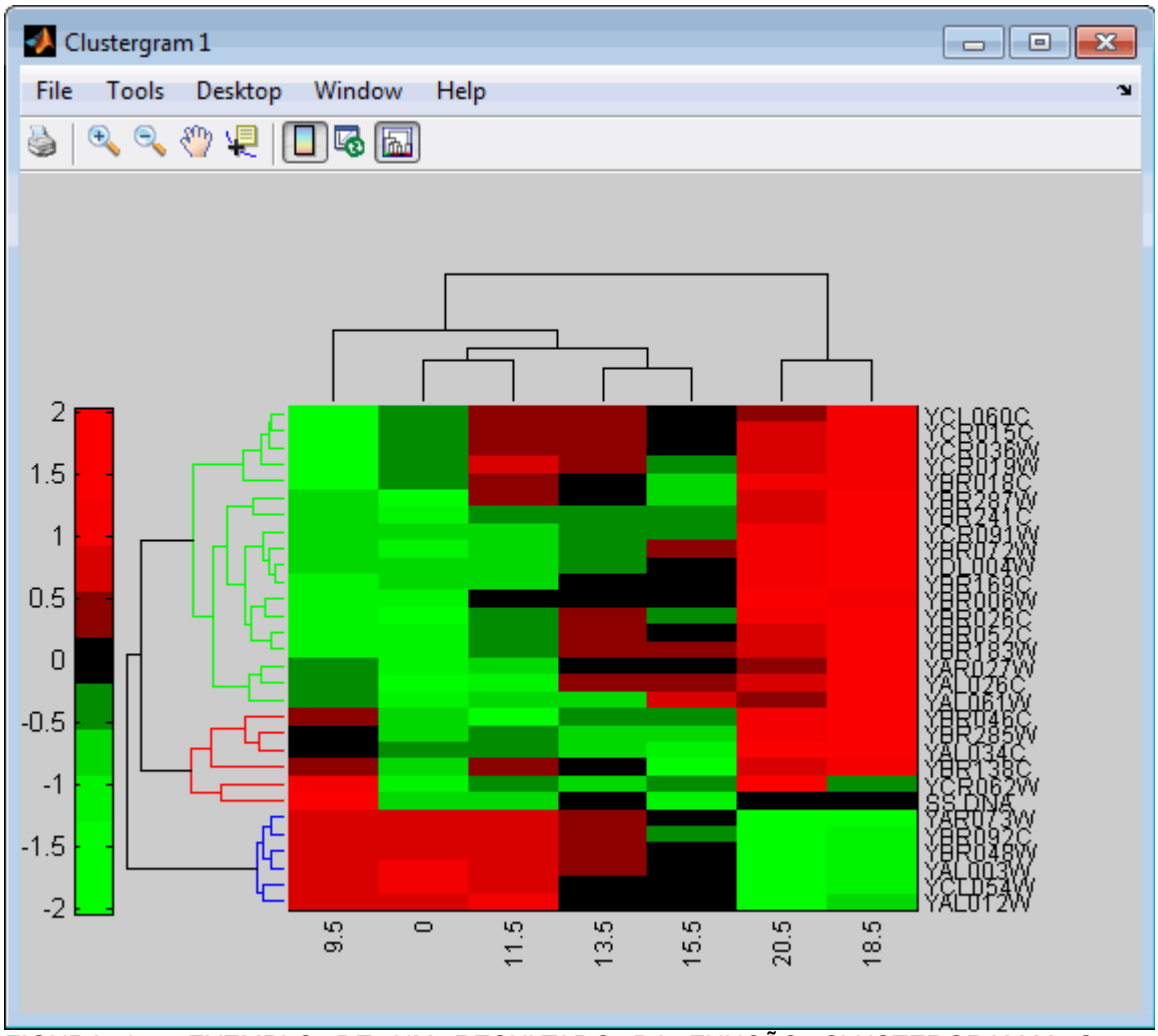


FIGURA 3 - EXEMPLO DE UM RESULTADO DA FUNÇÃO CLUSTERGRAM.M Genes agrupados de acordo com a similaridade calculada pela distância de similaridade. Em vermelho, os genes mais correlacionados.

FONTE: MATHWORKS (2015)

3.2 MÉTODO

Para atender os objetivos do trabalho, desenvolvemos a metodologia (Figura 04) deste trabalho, dividida em quatro etapas: 1) Seleção, 2) Limpeza, 3) Enriquecimento de Informação e 4) Codificação. Essas etapas são necessárias pois preparam os dados de forma sistemática à descoberta de grupos com funções correlacionadas pela formação dos agrupamentos.

Os dados utilizados nesse trabalho foram adquiridos de Dos Santos (2012), onde a autora propõe um grupo de seis genes mínimo para classificação de um organismo diazotrófos, utilizando 80 organismos referência da FBN.

1) Seleção: Os arquivos dos 80 genomas do trabalho de Dos Santos (2012) foram coletados do NCBI Genbank e armazenados em um pasta. A manipulação das estruturas contidas nos arquivos *.gbk* foram manipulados pela aplicação do JGBParser sendo úteis para extração de informações relevantes como Taxonomia e Número de acesso e identificação do gene âncora, *nifH*.

2) Limpeza: Em primeira instância, é feita a identificação o gene âncora, com base na *tag* de anotação “gene” ou “locus_tag”, para secção da região de interesse a ser explorada. O script *extractnifH.m* (ANEXO II), foi escrito para identificar a *tag* de anotação gene *nifH* e seccionar uma região de 25 kp à montante e à justante da posição inicial e final do gene. Casos que não foram identificados através da tag de anotação do gene nif foi utilizado o script *extractlocustag.m* (ANEXO III), onde a região genômica é extraída com base no “locus_tag” específico.

Após seccionada o fragmento de sequencia de 50 kpb é feito o reconhecimento de ORF e anotação automática dos produtos usando a plataforma de anotação automática SILA.

Uma vez obtido todos as sequencias dos arquivos anotados dos 80 genomas, esses arquivos são concatenados em um único multifasta e submetido ao reconhecimento de ortólogos aplicando o RAFTS3GROUPS (ANEXO IV).

3) Enriquecimento de Informação: Aos grupos de ortólogos criados com o RAFTS3GROUPS é realizado a etapa de enriquecimento de informação para reagrupar grupos que possuem mesmo domínio conservado e topologia de

superfamílias através de consultas manuais aos bancos de dados PROSITE e PFAM.

4) Codificação: Esta etapa visa dispor os dados gerados em uma matriz de ocorrência, designada organismo-cluster. É realizada a re-identificação de cada sequência de cada proteína existente nos grupos de ortólogos em cada um dos organismo com a busca RAFTS3 e *self-score* 1. Ao final, possuímos uma matriz contendo os genes identificados como ortólogos identificados nos 80 organismos analisados.

Após a codificação dos dados e criação da matriz de ocorrência, aplicam-se métodos de clusterização para a criação dos agrupamentos e feita a avaliação dos resultados com descoberta de associação entre os elementos da matriz.

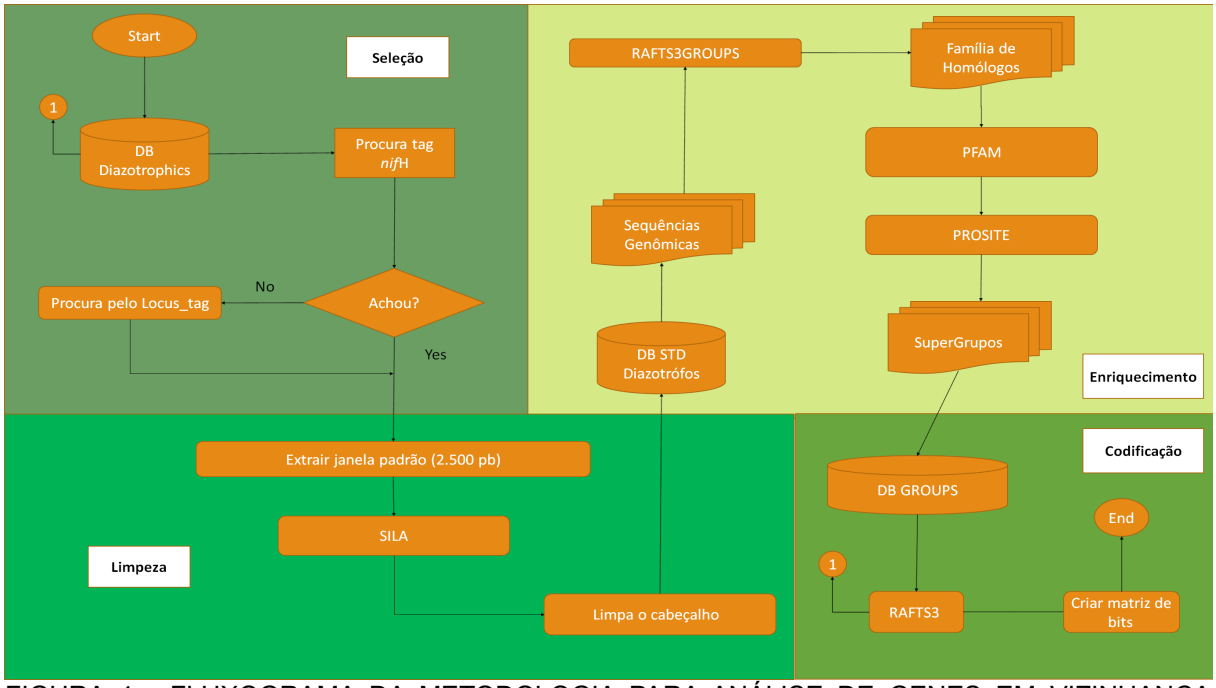


FIGURA 4 - FLUXOGRAMA DA METODOLOGIA PARA ANÁLISE DE GENES EM VIZINHANÇA CONECTADA. Blocos representam as etapas aplicadas neste trabalho, seleção, limpeza, enriquecimento e codificação.

FONTE: O autor (2015)

4 RESULTADOS E DISCUSSÃO

Os resultados apresentados neste trabalho foram obtidos através da metodologia desenvolvida para a mineração genômica (Figura 04). Com exceção da etapa de enriquecimento de informação, toda a metodologia foi integrada em rotinas MATLAB para facilitar a manipulação dos dados. Os códigos-fonte dos *scripts* encontram-se disponíveis nos anexos da dissertação e foram comentados de forma à facilitar a re-aplicação desse tipo de trabalho.

A metologia foi desenvolvida em um computador de configuração simples, visando a aplicação desse tipo de análise independente de infra-estrutura computacional robusta, um dos principais desafios da bioinformática nos dias atuais.

Os ensaios *in-silico* mostraram-se válidos através das hipóteses que serão abordadas adiante, sendo posteriormente necessária confirmação e validação *in-vitro* desses resultados. Ao conduzir esse tipo de estudo, outros trabalhos foram encontrados aplicando mineração de dados em diversos campos como marketing, engenharia agronômica, ciências farmacêuticas e etc (BOYD, 2013; MACEDO, 2010).

A associação estrutural de um gene está relacionada à essencialidade da sua função na manutenção do gene em operons conservados. Esse tipo de questão possibilita a exploração do conceito de ortologia de sequencias e a reconstrução da árvore dos domínios da vida (DELSUC, 2005). Entretanto, delimitamos o foco deste estudo em interpretar os resultados da clusterização aplicada à dados genômico para compreender os significados dos mecanismos biológicos através da correlação de genes em vizinhança.

4.1 SELEÇÃO DA REGIÃO GENÔMICA DE INTERESSE

Analisamos oitenta organismos diazotrófos descritos inicialmente no trabalho de Dos Santos e Colaboradores (2012). A partir da localização do gene *nif*H (dinitrogenase redutase ou proteina Fe do complexo nitrogenase), delimitamos e extraímos um fragmento de DNA com 25 Kpb (vinte e cinco mil pares de base) à jusante e à montante da posição inicial e final do gene *nifH.* A identificação deste gene foi feita usando o JGBParser a partir da tag de anotação "/gene=nifH". Organismos que não foram obtidos em primeira instância foram identificados pela tag "/locus_tag" descrita no trabalho de Dos Santos e colaboradores (2012).

Em *K. pneumoniae,* organismo modelo da FBN, os 19 *nif* foram identificados em uma janela de 24 kbp (ARNOLD et al, 1988).

No decorrer deste trabalho, conduzimos testes primários em fragmentos de 20 kpb e 40 kpb, até obter um fragmento de sequencia com número de bases suficientes para abranger o conteúdo dos genes *nif* de todos os oitenta organismos analisados. A divergência do tamanho do cluster *nif* varia de organismo para organismo e reflete diretamente pressões seletivas ocorridas durante a evolução desses organismos (BOYD, 2013).

Os fragmentos extraídos dos 80 organismos foram submetidos ao reconhecimento de Open Read Frame (ORF) e anotação padronizada dos produtos usando a plataforma de anotação SILA, na versão local (VIALLE, 2013).

O número de genes presentes no fragmento de sequência extraído de cada organismo é descrito na tabela 01.

TABELA 01 – NÚMERO DE GENES EXISTENTE DENTRO DE CADA FRAGMENTO EXTRAÍDO DOS OITENTA ORGANISMOS UTILIZADOS NESSE TRABALHO

Organismo	Número de ORF's
Acidithiobacillus ferrooxidans ATCC 23270	59
Allochromatium vinosum DSM 180 chromosome	33
Anabaena variabilis ATCC 29413 chromosome	37
Arcobacter nitrofigilis DSM 7299 chromosome	39
Azoarcus sp. BH72 chromosome	57
Azorhizobium caulinodans ORS 571 chromosome	66
Azospirillum sp. B510 chromosome	53
Azotobacter vinelandii DJ chromosome	53
Beijerinckia indica subsp. indica ATCC 9039 chromosome	60
Bradyrhizobium japonicum USDA 110 chromosome	81
Bradyrhizobium sp. BTAi1 chromosome	69
Burkholderia phymatum STM815 plasmid pBPHY02	56
Burkholderia sp. CCGE1002 plasmid pBC201	69
Burkholderia vietnamiensis G4 chromosome 3	67
Burkholderia xenovorans LB400 chromosome 2	60
Chlorobaculum parvum NCIB 8327 chromosome	58
Chlorobium limicola DSM 245 chromosome	57
Chlorobium phaeobacteroides BS1 chromosome	57
Chlorobium tepidum TLS chromosome	42
Clostridium acetobutylicum ATCC 824 chromosome	49
Clostridium beijerinckii NCIMB 8052 chromosome	44
Clostridium kluyveri DSM 555 chromosome	42
Cupriavidus taiwanensis plasmid pRALTA	55
Cyanobacterium UCYN-A	53
Cyanothece sp. ATCC 51142 chromosome circular	66
Dehalococcoides ethenogenes 195	58
Desulfitobacterium hafniense DCB-2 chromosome	52
Desulfotomaculum ruminis DSM 2154 chromosome	58
Desulfovibrio vulgaris subsp. vulgaris DP4 plasmid pDVUL01	43
Frankia alni ACN14a chromosome	56
Frankia sp. CcI3 chromosome	56
Geobacter metallireducens GS-15 chromosome	64
Geobacter sulfurreducens PCA chromosome	66
Geobacter uraniireducens Rf4 chromosome	47
Gluconacetobacter diazotrophicus PAl 5 chromosome	53
Halorhodospira halophila SL1 chromosome	49
Heliobacterium modesticaldum Ice1 chromosome	41
Herbaspirillum seropedicae SmR1 chromosome	60
Klebsiella pneumoniae 342 chromosome	50
Klebsiella variicola At-22 chromosome	48
Magnetospirillum magneticum AMB-1 chromosome	54
Mesorhizobium ciceri biovar biserrulae WSM1271 chromosome	63
Mesorhizobium loti MAFF303099 chromosome	58
Mesorhizobium opportunistum WSM2075 chromosome	63
Methanobacterium sp. AL-21 chromosome	55
Methanococcus aeolicus Nankai-3 chromosome	48
Methanococcus maripaludis C5 chromosome	46
Methanosarcina acetivorans C2A chromosome	48
Methanosarcina barkeri str. Fusaro chromosome	59
Methanosarcina mazei Go1 chromosome	48
Methanothermobacter thermautotrophicus str. Delta H chromosome	50
Methylobacterium nodulans ORS 2060 chromosome	65
Methylobacterium sp. 4-46 chromosome	50
Methylocella silvestris BL2 chromosome	61
Methylococcus capsulatus str. Bath chromosome	60
Methylomonas methanica MC09 chromosome	49
Nostoc azollae 0708	84
Nostoc punctiforme PCC 73102 chromosome	48
Nostoc sp. PCC 7120 chromosome	52
Pantoea sp. At-9b plasmid pPAT9B03	55
Pelobacter propionicus DSM 2379 chromosome	50
Polaromonas naphthalenivorans CJ2 chromosome	67
Prosthecochloris aestuarii DSM 271 chromosome	55
Pseudomonas stutzeri A1501 chromosome	63
Rhizobium etli CFN 42 symbiotic plasmid p42d	64
Rhizobium leguminosarum bv. trifolii WSM1325 plasmid pR132501	61
Rhizobium leguminosarum bv. viciae 3841 plasmid pRL10	73
Rhodobacter capsulatus SB 1003 chromosome	47
Rhodobacter sphaeroides ATCC 17029 chromosome 1	53
Rhodomicrobium vannielii ATCC 17100 chromosome	49
Rhodopseudomonas palustris CGA009 chromosome	56
Rhodospirillum centenum SW chromosome	47
Rhodospirillum rubrum ATCC 11170 chromosome	39
Sinorhizobium fredii NGR234 plasmid pNGR234a	67
Sinorhizobium medicae WSM419 plasmid pSMED02	74
Sinorhizobium meliloti 1021 plasmid pSymA	64
Synechococcus sp. JA-2-3Ba(2-13) chromosome	57
Teredinibacter turnerae T7901 chromosome	59
Trichodesmium erythraeum IMS101 chromosome	42
Xanthobacter autotrophicus Py2 chromosome	66
Total	4452

FONTE: O autor (2015)

A análise do primeiro resultado obtido levou ao aumento progressivo da janela até o tamanho final de 25 kbp, possibilitando englobar o conteúdo integral dos 19 genes *nif* que compõem o *cluster* de *K. pneumoniae*. Os primeiros programas testados necessitaram de uma infraestrutura computacional robusta, em razão da divergência de sequencias do conjunto de entrada. O reconhecimento de ortólogos do conjunto de entrada, contendo 4.452 proteínas, foi realizado em 72 horas. Diante deste extenso tempo de processamento e complexidade computacional necessário para o agrupamento de ortólogos, fomos motivados a desenvolver uma aplicação do *script* RAFTS3 (VIALLE, 2013) para diminuir o tempo gasto nesta etapa do processo.

4.2 RAFTS3GROUPS UMA FERRAMENTA PARA RECONHECIMENTO DE GENES ORTÓLOGOS

A ferramenta RAFTS3 (VIALLE, 2013), desenvolvida no Laboratório de Inteligência Artificial aplicada à Bioinformática do PPG em Bioinformática, é uma alternativa ao programa BLAST e possui o objetivo de identificar a similaridade entre sequências usando técnicas livre de alinhamento.

A maior diferença entre os programas está na redução da complexidade da informação e avaliação dos *scores* que retornam das análises de similaridades entre sequências.

Enquanto o programa BLAST utiliza matrizes PAM, e BLOSUM e conduz o algoritmo através de alinhamento par-a-par das sequências, a heurística de busca de similaridade do RAFTS3 independe de alinhamento.

O RAFTS3 e inicia a busca de similaridade do conjunto de dados de proteínas pela criação de uma tabela *hash* de k-*mers* e uma matriz binária de aminácidos co-ocorrentes (BCOM) e avalia a similaridade pela comparação de todos os possíveis *k-mers* da sequência com as BCOM, resultando em um ranking de similaridade de sequências.

Nesse trabalho, desenvolvemos o RAFTS3GROUPS, uma aplicação da ferramenta RAFTS3 para o reconhecimento de ortologia de sequências. O RAFTS3GROUP utiliza heurística todos-contra-todos e determina o grau de ortologia pelo valor *self-score,* resultado da similaridade da sequência *query* na base de dados formatada para o RAFTS3.

O reconhecimento de ortólogos nos métodos tradicionais, implementando as matrizes PAM e BLOSUM, retornam valores de identidade e similaridade, e são avaliados pelo mínimo de 30% de similaridade e 60% de identidade entre as sequências.

Para identificação de ortólogos usando o RAFTS3GROUPS usamos o valor de *self-score* em 0.5 análogo a 50% de similaridade implementando as matrizes PAM e BLOSUM. Contudo, esse valor pode ser conduzido, com parcimônia, até o mínimo 0.3, maiores testes são necessários.

O RAFTS3GROUPS recebeu de entrada 4.452 sequências e gerou 423 grupos de ortólogos, contendo um total de 1.851 sequências, com suporte mínimo de duas sequências em cada grupo. Os grupos foram armazenados em formato fasta e submetidos à anotação de produtos, os resultados foram armazenados em tabelas descritivas.

A aplicação do RAFTS3GROUPS no conjunto de 4.452 proteínas, foi executada em menos de 20 minutos em um computador com configuração simples (Core I5, 2.5 Ghz e 12 Gb RAM). O mesmo conjunto de dados foi submetido a um notebook (AMD A10 2.8 Ghz e 4GB de RAM) e executado em 30 minutos, apresentando-se como uma ferramenta de alto rendimento e baixo custo computacional para o reconhecimento de ortólogos.

4.3 ANÁLISE DE ENRIQUECIMENTO DE INFORMAÇÃO

A escolha do valor de 0.5 para o *self-score* caracterizou ortólogos com

provável função biológica em pequenos grupos com duas três sequências. Para validar os grupos de ortólogos gerados pelo RAFTS3GROUPS realizamos uma análise de enriquecimento de informação através de dois bancos de dados: PROSITE, para identificar domínio funcional e PFAM para identificar a topologia na estrutura secundária. Os grupos com mesmo resultados de PROSITE e PFAM foram re-agrupados em um novo conjunto contendo os resultados compartilhados.

Essa etapa possibilitou a identificação 122 grupos falsos-positivos que possuíam mesma função biológica e foram caracterizados em grupos diferentes, provavelmente pela restrição utilizada no RAFTS3GROUPS (*Self-score*=0.5). Essa análise reduziu o conjunto inicial de 423 grupos de ortólogos à 301 grupos (FIGURA 05).

4.4 DETERMINAÇÃO CUTOFF MÍNIMO

Em cada grupo de ortólogo, determinamos um *cutoff* de três sequencias, com suporte mínimo de três organismos diferentes. Os grupos de ortólogos que não atenderam ao *cutoff* foram excluídos da análise. E essa etapa visou eliminar grupos com baixa ocorrência, o que sinaliza uma atividade não relacionada ao metabolismo da FBN. Essa etapa reduziu o conjunto de dados de 301 grupos à 117 grupos de ortólogos (FIGURA 05).

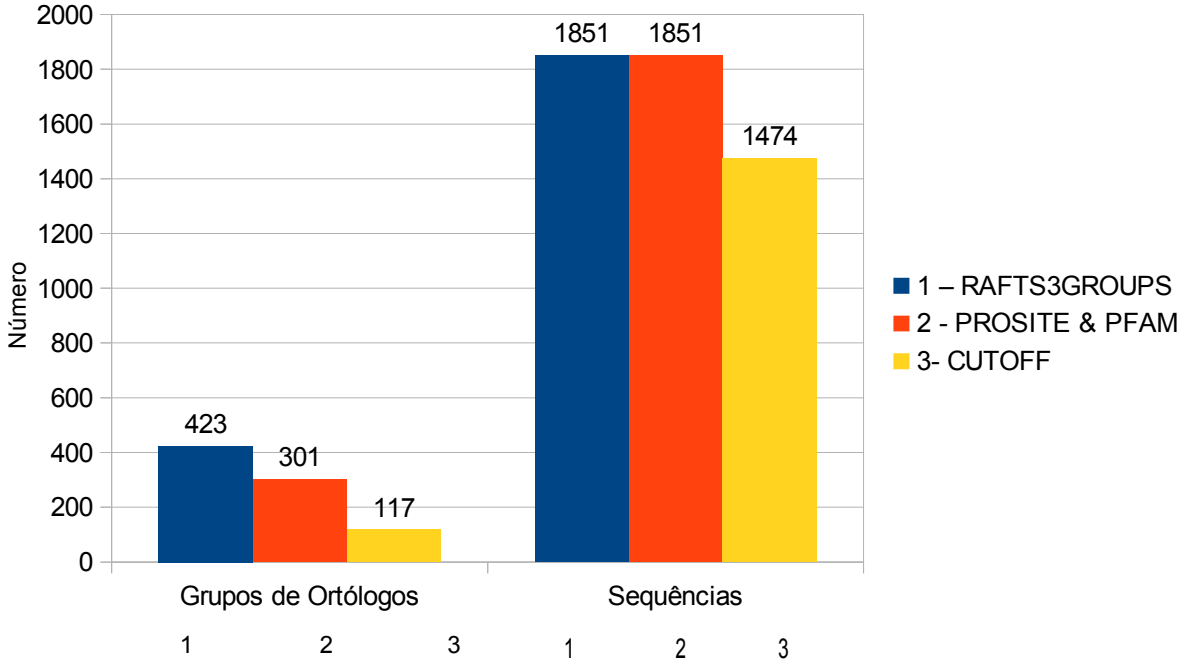


FIGURA 5 - NÚMERO DE GRUPOS DE ORTÓLOGOS E SEQUÊNCIAS EXISTENTES NO CONJUNTO DE DADOS GERADOS DURANTE CADA ETAPA EXECUTADA NA METODOLOGIA. 1) Quantidade de grupos de ortólogos reconhecidos pelo RAFTS3GROUPS. 2) Quantidade de grupos de ortólogos após o enriquecimento de informação com PROSITE e PFAM. 3) Quantidade de grupos de ortólogos após o CUTOFF, três sequencias em três organismos distintos.

FONTE: O autor (2015)

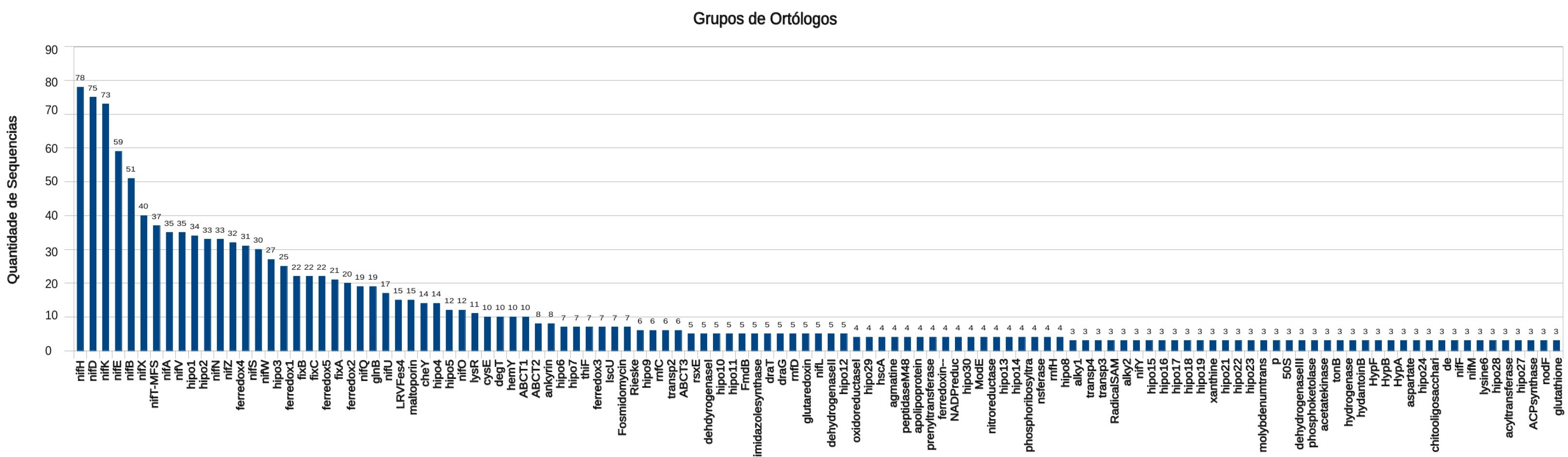


FIGURA 6 - GRUPOS DE ORTÓLOGOS APÓS O CUTOFF. 117 grupos de ortólogos e suas quantidades de sequências em cada grupo. Os seis genes mínimo da fixação do nitrogênio (nifHDKENB) apresentaram maiores quantidade de sequencias.

FONTE: O autor (2015)

4.5 CLASSIFICAÇÃO DOS *NIF* USANDO PROCLAT

Durante a execução da metodologia, avaliamos o conjunto mínimo dos seis genes *nif* (DOS SANTOS, 2012). Os genes *nif* possuem alta similaridade entre os genes *nifE* e *nifN*, decorrentes de duplicação gênica de um ancestral comum envolvendo a participação dos genes *nif*D e *nif*K (BOYD,2013).

Para distinguir sequencias agrupadas erroneamente, nós aplicamos o ProClaT, um classificador de proteínas que utiliza aprendizado de máquina para o reconhecimento de padrões de sequências genômicas, em cada um dos seis genes *nif*. O conjunto inicial de sequencias do grupo *nifN* gerados com o RAFTS3GROUP agrupou apenas cinco sequencias correspondentes ao gene *nifN*. Quando avaliados durante a análise de enriquecimento de informação, usando PFAM e PROSITE, nenhum outro grupo de sequencias referentes a *nifN* foi reconhecido, pelo fato de possuir mesma atividade dos genes *nifEDK*. Somente quando aplicados ao ProClaT foram reconhecidas 9 sequências de *nifN* agrupadas com o grupo *nifE*. As demais sequência de *nifN*, foram dispersadas em pequenos grupos de duas ou três sequências, durante o agrupamento de ortológos com RAFTS3GROUP e com a classificação do ProClaT foram inseridas no seu devido grupo.

A literatura descreve a localização adjacente dos *nifEN* aos *nifDK* (DOS SANTOS, 2012), explicando a baixa ocorrência destes genes no conjunto de dados final. Os resultados dessa avaliação encontram-se na tabela 02.

TABELA 02 – AVALIAÇÃO DOS GRUPOS DE GENES MÍNIMO DA FIXAÇÃO BIOLÓGICA DO NITROGÊNIO DE ACORDO COM A QUANTIDADE DE SEQUENCIA EM CADA ETAPA ANALISADA.

	Total	*nifH*	*nifD*	*nifK*	*nifE*	*nifN*	*nifB*
1 - RAFTS3GROUPS	423	79	54	55	74	5	41
2 - PFAM &PROSITE	301	79	77	76	74	5	53
3 - CUTOFF & PROCLAT	117	79	79	76	65	33	53

Fonte: O autor (2015)

4.6 CONSTRUÇÃO DA MATRIZ DE OCORRÊNCIA

Para a descoberta de associação de genes em vizinhança foi necessário a criação de uma estrutura a dispor os dados gerados à aplicação dos algoritmos de mineração de dados. Nós utilizamos o mesmo conceito da matriz de expressão dos experimentos de RNA-seq, onde em cada coluna da matriz contém a informação de um gene avaliado e em cada coluna o nível de expressão calculado.

Nós criamos uma matriz de ocorrência a partir dos 117 grupos de ortólogos re-identificados nos 80 genomas analisados (FIGURA 06). A matriz foi construída pela identificação da proteína de cada grupos de ortólogos nos conjunto de ORFS dos fragmentos extraídos dos organismos.

Na matriz criada, cada linha representa um dos oitenta organismos analisados e em cada coluna um dos 117 grupos de ortólogos. As informação em linha foram ordenadas pela taxonomia dos organismos (tabela 03) e as colunas em ordem decrescente da quantidade de sequencias presente em cada grupo (tabela 04). A matriz foi populada com base na presença da proteína nos organismos. Cada ponto azul na matriz da figura 06 representa a presença da proteína dentro do organismo.

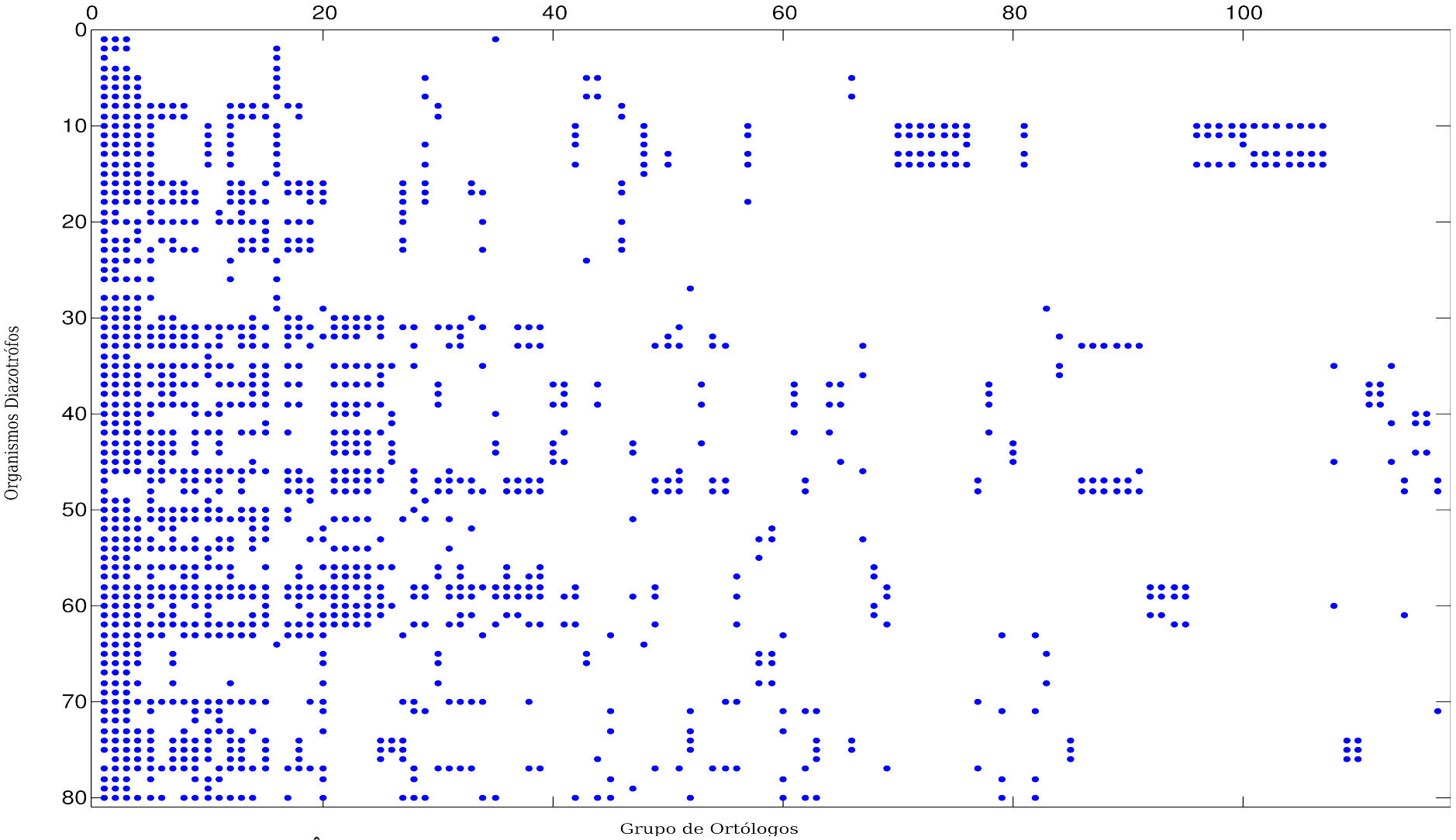


FIGURA 7 - MATRIZ DE OCORRÊNCIA ORGANISMOS-CLUSTERS. Organismos diazotrófos dispostos em linhas e grupos de ortólogos em colunas. Pontos azuis representam a ocorrência da proteína do grupo de ortólogo, em linha, no organismos, em coluna.

FONTE: O autor (2015)

TABELA 03 – TABELA DOS DIAZOTRÓFOS DISPOSTOS NA MATRIZ E NÚMERO DE OCORRÊNCIAS

Ordem	Filo	Organismo	Ocorrências
1	Euryarchaeota	Methanobacterium sp	4
2	Euryarchaeota	Methanothermobacter thermautotrophicus str	4
3	Euryarchaeota	Methanococcus aeolicus Nankai-3 chromosome	2
4	Euryarchaeota	Methanococcus maripaludis C5 chromosome	4
5	Euryarchaeota	Methanosarcina acetivorans C2A chromosome	9
6	Euryarchaeota	Methanosarcina barkeri str	5
7	Euryarchaeota	Methanosarcina mazei Go1 chromosome	9
8	Actinobacteria	Frankia alni ACN14a chromosome	16
9	Actinobacteria	Frankia sp	15
10	Chlorobi	Chlorobaculum parvum NCIB 8327 chromosome	31
11	Chlorobi	Chlorobium tepidum TLS chromosome	24
12	Chlorobi	Chlorobium limicola DSM 245 chromosome	13
13	Chlorobi	Chlorobium phaeobacteroides BS1 chromosome	25
14	Chlorobi	Prosthecochloris aestuarii DSM 271 chromosome	32
15	Chloroflexi	Dehalococcoides ethenogenes 195	7
16	Cyanobacteria	Cyanobacterium UCYN-A	19
17	Cyanobacteria	Cyanothece sp	20
18	Cyanobacteria	Synechococcus sp	18
19	Cyanobacteria	Anabaena variabilis ATCC 29413 chromosome	6
20	Cyanobacteria	Nostoc azollae 0708	20
21	Cyanobacteria	Nostoc punctiforme PCC 73102 chromosome	3
22	Cyanobacteria	Nostoc sp	14
23	Cyanobacteria	Trichodesmium erythraeum IMS101 chromosome	16
24	Firmicutes	Clostridium acetobutylicum ATCC 824 chromosome	8
25	Firmicutes	Clostridium beijerinckii NCIMB 8052 chromosome	2
26	Firmicutes	Clostridium kluyveri DSM 555 chromosome	7
27	Firmicutes	Desulfitobacterium hafniense DCB-2 chromosome	1
28	Firmicutes	Desulfotomaculum ruminis DSM 2154 chromosome	6
29	Firmicutes	Heliobacterium modesticaldum Ice1 chromosome	7
30	Proteobacteria	Beijerinckia indica subsp	15
31	Proteobacteria	Methylocella silvestris BL2 chromosome	33
32	Proteobacteria	Bradyrhizobium japonicum USDA 110 chromosome	23
33	Proteobacteria	Bradyrhizobium sp	34
34	Proteobacteria	Rhodomicrobium vannielii ATCC 17100 chromosome	4
35	Proteobacteria	Methylobacterium nodulans ORS 2060 chromosome	27
36	Proteobacteria	Methylobacterium sp	12
37	Proteobacteria	Mesorhizobium ciceri biovar biserrulae WSM1271 chromosome	32
38	Proteobacteria	Mesorhizobium loti MAFF303099 chromosome	14
39	Proteobacteria	Mesorhizobium opportunistum WSM2075 chromosome	32
40	Proteobacteria	Rhizobium leguminosarum bv	15
41	Proteobacteria	Rhizobium leguminosarum bv	9
42	Proteobacteria	Sinorhizobium fredii NGR234 plasmid pNGR234a	23
43	Proteobacteria	Sinorhizobium medicae WSM419 plasmid pSMED02	19
44	Proteobacteria	Sinorhizobium meliloti 1021 plasmid pSymA	20
45	Proteobacteria	Rhizobium etli CFN 42 symbiotic plasmid p42d	13
46	Proteobacteria	Rhodopseudomonas palustris CGA009 chromosome	27
47	Proteobacteria	Azorhizobium caulinodans ORS 571 chromosome	39
48	Proteobacteria	Xanthobacter autotrophicus Py2 chromosome	37
49	Proteobacteria	Rhodobacter capsulatus SB 1003 chromosome	7
50	Proteobacteria	Rhodobacter sphaeroides ATCC 17029 chromosome 1	15
51	Proteobacteria	Gluconacetobacter diazotrophicus PAl 5 chromosome	24
52	Proteobacteria	Azospirillum sp	11
53	Proteobacteria	Magnetospirillum magneticum AMB-1 chromosome	17
54	Proteobacteria	Rhodospirillum centenum SW chromosome	18
55	Proteobacteria	Rhodospirillum rubrum ATCC 11170 chromosome	5
56	Proteobacteria	Burkholderia phymatum STM815 plasmid pBPHY02	26
57	Proteobacteria	Burkholderia sp	19
58	Proteobacteria	Burkholderia vietnamiensis G4 chromosome 3	42
59	Proteobacteria	Burkholderia xenovorans LB400 chromosome 2	43
60	Proteobacteria	Cupriavidus taiwanensis plasmid pRALTA	18
61	Proteobacteria	Polaromonas naphthalenivorans CJ2 chromosome	24
62	Proteobacteria	Herbaspirillum seropedicae SmR1 chromosome	37
63	Proteobacteria	Azoarcus sp	23
64	Proteobacteria	Desulfovibrio vulgaris subsp	6
65	Proteobacteria	Geobacter metallireducens GS-15 chromosome	11
66	Proteobacteria	Geobacter sulfurreducens PCA chromosome	10
67	Proteobacteria	Geobacter uraniireducens Rf4 chromosome	3
68	Proteobacteria	Pelobacter propionicus DSM 2379 chromosome	11
69	Proteobacteria	Arcobacter nitrofigilis DSM 7299 chromosome	3
70	Proteobacteria	Acidithiobacillus ferrooxidans ATCC 23270	27
71	Proteobacteria	Teredinibacter turnerae T7901 chromosome	18
72	Proteobacteria	Allochromatium vinosum DSM 180 chromosome	5
73	Proteobacteria	Halorhodospira halophila SL1 chromosome	14
74	Proteobacteria	Klebsiella pneumoniae 342 chromosome	22
75	Proteobacteria	Klebsiella variicola At-22 chromosome	22
76	Proteobacteria	Pantoea sp	19
77	Proteobacteria	Methylococcus capsulatus str	33
78	Proteobacteria	Methylomonas methanica MC09 chromosome	13
79	Proteobacteria	Azotobacter vinelandii DJ chromosome	5
80	Proteobacteria	Pseudomonas stutzeri A1501 chromosome	29

Quantidade de sequencias ortólogas re-identificadas no conjunto de dados originais. A coluna ID corresponde a posição do organismo na matriz de ocorrência. A coluna Filo corresponde ao filo do organismos. A coluna organismos estão os organismos e localização da nitrogenase identificada. Na coluna ocorrência estão a quantidade de ortólogos re-identificados em cada organismo.

Fonte: O autor (2015)

TABELA 04 – TABELA DE OCORRENCIA DE ORTÓLOGOS DISPOSTOS NA MATRIZ E O NÚMERO DE OCORRÊNCIAS.

ID	Gene/ORF	Número
1	nifH	79
2	nifD	79
3	nifK	76
4	nifE	65
5	nifB	53
6	Hipotetica 1	45
7	nifX	43
8	nifZ	42
9	nifT	39
10	nifA	38
11	Ferredoxin 4	35
12	nifV	35
13	nifS	34
14	Hipotetica 2	33
15	nifN	33
16	glnB	32
17	Hipotetica 3	29
18	nifW	27
19	Ferredoxin 2	25
20	Ferredoxin 1	24
21	fixB	22
22	fixC	22
23	Ferredoxin 5	21
24	fixA	20
25	nifQ	19
26	LysR family transcriptional regulator	17
27	nifU	15
28	LRV Fes4 cluster domain	15
29	maltoporin	14
30	cheY	14
31	Hipotetica 4	12
32	Hipotetica 5	12
33	nifO	11
34	serine O-acetyltransferase	10
35	ABC transporter ATP-binding protein	10
36	ankyrin	10
37	Hipotetica 6	10
38	degT	10
39	hemY	10
40	Transposase 2	9
41	Hipotetica 7	8
42	molybdate ABC transporter substrate-binding protein	8
43	antibiotic ABC transporter ATP-binding protein	8
44	Fosmidomycin resistance protein	8
45	electron transport complex RsxE subunit	8
46	ThiF	7
47	succinate-semialdehyde dehdyrogenase	7
48	Ferredoxin 3	7
49	IscU	7
50	Rieske	7
51	Hipoteitca 9	6
52	electron transporter RnfC	6
53	Hipotetica 8	6
54	Hipotetica 10	5
55	Hipotetica 11	5
56	FmdB	5
57	imidazole glycerol phosphate synthase	5
58	draT	5
59	draG	5
60	electron transporter RnfD	5
61	MULTISPECIES: short-chain dehydrogenase	5
62	monothiol glutaredoxin	5
63	nifL	5
64	alkyl hydroperoxide reductase	5
65	Transposase 4	4
66	Hipotetica 12	4
67	oxidoreductase	4
68	Hipotetica 29	4
69	HscA	4
70	agmatine deiminase	4
71	peptidase M48	4
72	apolipoprotein acyltransferase	4
73	1.4-dihydroxy-2-naphthoate prenyltransferase	4
74	ferredoxin–NADP reductase	4
75	Hipotetica 30	4
76	ModE	4
77	nitroreductase	4
78	Hipotetica 13	4
79	Hipotetica 14	4
80	Transposase 3	4
81	hypoxanthine phosphoribosyltransferase	4
82	protein RnfH	4
83	Radical SAM domain protein	3
84	alkyl hydroperoxide reductase	3
85	nifY	3
86	Hipotetica 15	3
87	Hipotetica 16	3
88	Hipotetica 17	3
89	Hipotetica 18	3
90	Hipotetica 19	3
91	xanthine dehydrogenase	3
92	Hipotetica 21	3
93	Hipotetica 22	3
94	Hipotetica 23	3
95	molybdenum transporter	3
96	50S ribosomal protein L27	3
97	malate dehydrogenase	3
98	phosphoketolase	3
99	acetate kinase	3
100	TonB-denpendent receptor	3
101	hydrogenase	3
102	hydantoin utilization protein B	3
103	hydrogenase maturation protein HypF	3
104	hydrogenase nickel incorporation protein HypB	3
105	hydrogenase nickel incorporation protein HypA	3
106	aspartate carbamoyltransferase	3
107	hipotetica 24	3
108	chitooligosaccharide deacetylase	3
109	nifF	3
110	nifM	3
111	L-lysine 6-monooxygenase	3
112	hypothetical protein	3
113	acyltransferase	3
114	Hipotetica 27	3
115	beta-ACP synthase	3
116	nodulation protein NodF	3
117	glutathione S-transferase	3

Fonte: O autor (2015)

4.7 SELEÇÃO DA DISTANCIA APROPRIADA PARA CLUSTERIZAÇÃO NÃO SUPERVISIONADA

As escolhas dos métodos de distancia e agrupamento reflete diretamente na interpretação dos dados (JASKOWIAK, 2014). Os agrupamentos foram gerados com a função *clustergram.m,* nativa da toolbox de bioinformática do MATLAB.

Para escolha da melhor medida de agrupamento nós avaliamos cinco medidas (distancias) de similaridade e seis métodos de ligação pelo coeficiente cofenético. As medidas de similaridade avaliadas foram: distância Euclidiana, distância Cityblock (Manhattan) , Correlação de Pearson, distância de Hamming e distância de Spearman. E os métodos de ligação: Average Linkage (AL), Centroid (CD), Complete Linkage (CL), Median (MD), Single Linkage (SL) e Ward (WD).

Foi realizada a comparação de todos as medidas de similaridade com os métodos de ligação. A avaliação desse resultados foi calculada pelo coeficiente cofenético, resultado do coeficiente de Pearson entre a matriz de similaridade e os agrupamentos geradas. O coeficiente cofenético estima força agrupamentos. Os resultados do coeficiente cofenético para as medidas avaliadas encontram-se na tabela 05. O melhor resultado do coeficiente cofenético é o que aproxima de 1, porém valores acima de 0.6 são considerados aceitáveis e viáveis de aplicação.

Para a construção da figura 08 utilizamos a medida similaridade correlação de Pearson e o método de ligação Ward, pois apresentaram o melhor resultado visual e evidenciaram os diferentes agrupamentos criados na clusterização. Os agrupamentos foram destacados no dendrograma pela da distância mínima de 1.1 resultando em 9 agrupamentos de ortólogos (FIGURA 09).

TABELA 05 – AVALIAÇAO DO COEFICIENTE COFENÉTICO ENTRE AS MEDIDAS DE DISTÂNCIA E OS MÉTODOS DE LIGAÇÃO

	Medidas de Similaridade				
Método de Ligação	Cityblock	Pearson	Euclidiana	Hamming	Spearman
Average	0.85**	0.81**	0.90*	0.85*	0.81*
Centroid	0.87**	0.78**	0.88**	0.87*	0.78*
Complete	0.77**	0.57*	0.83*	0.78*	0.57**
Median	0.72*	0.75*	0.76**	0.77*	0.75**
Single	0.82**	0.65**	0.86*	0.82*	0.65**
Ward	0.66*	0.58**	0.69**	0.67**	0.58**

*Agrupamento consistente **Agrupamento viável, cabível de prudência

Fonte: O autor (2015)

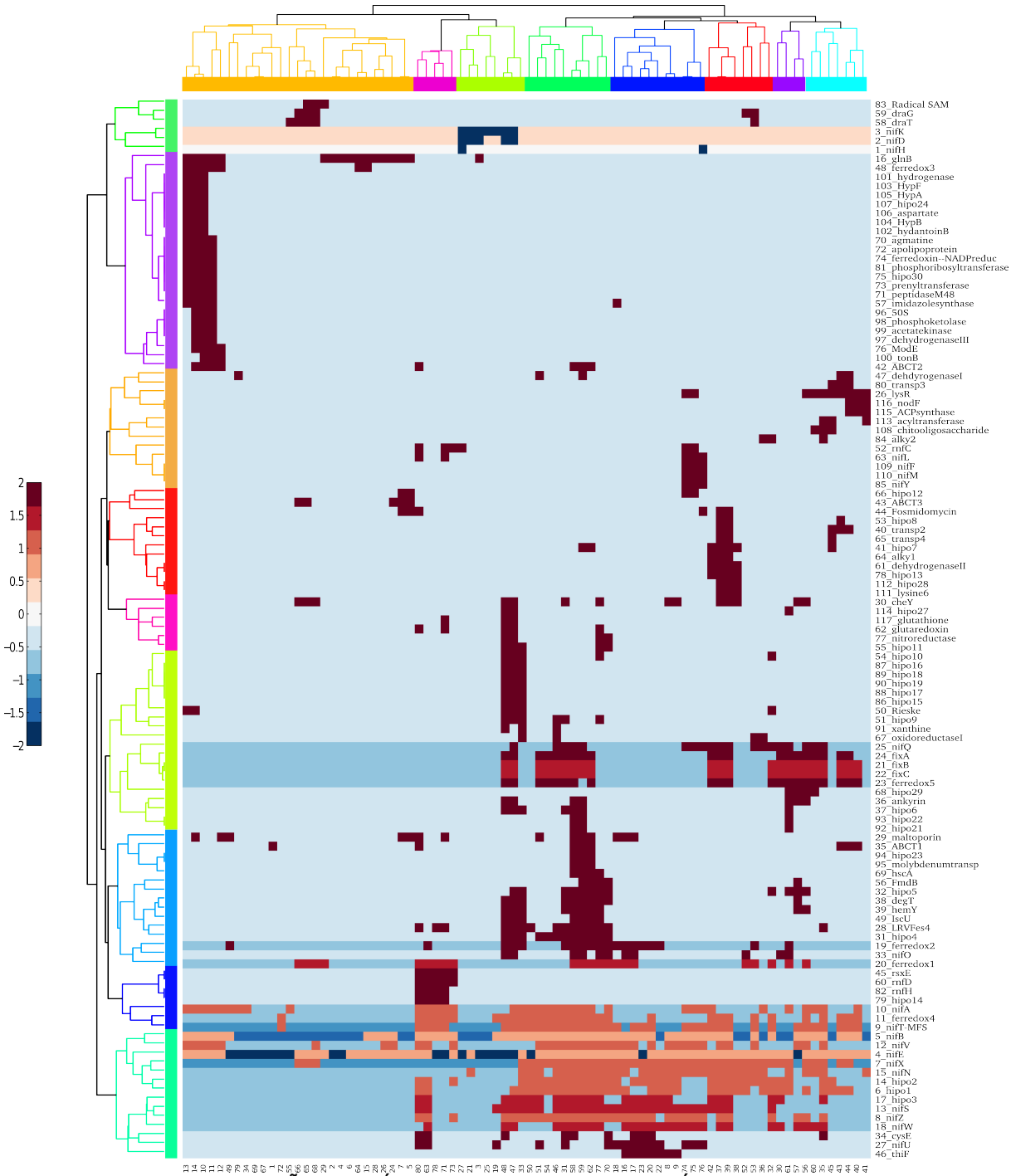


FIGURA 8 - CLUSTERIZAÇÃO HIERÁRQUICA DOS GENES ORTÓLOGOS DE 80 ORGANISMOS DA FIXAÇÃO BIOLÓGICA DO NITROGÊNIO. A figura foi feita com a função clustergram da toolbox de bioinformática do MATLAB. O agrupamento gerado com a medida de similaridade correlação de Pearson e o método de ligação de Ward. Os agrupamentos gerados foram destacados a partir da distância mínima de 1.1 evidenciando 9 grupos de ortólogos em 8 grupos de organismos. Mapa de calor avaliado pelo nível de similaridade entre os genes, variando na escala de 2 a -2. Genes correlacionados foram agrupados de acordo com a similaridade entre os organismos que possuem o mesmo gene.

Fonte: O autor (2015)

4.8 INTERPRETAÇÃO DA CLUSTERIZAÇÃO

O uso de clusterização de dados aos estudos biológicos foram inicialmente aplicados para extração de resultados em dados de expressão gênica (experimentos de RNA-seq). Os resultados brutos desse tipo de experimento é similar ao apresentado na figura 06, reportados através de valores de expressão detectados situações onde o gene é induzido à estresses e/ou controle.

Tanto nas matrizes de expressão gênica, quanto na matriz de ocorrência apresentada nesse trabalho, os resultados podem ser extraídos pela comparação entre as informações contidas em linhas ou colunas, e para a análise de genes em vizinhança conectada, nós focamos a aplicação de clusterização tomando base a informação contida nas colunas, onde estão as informações dos ortólogos identificados nos organismos.

Os agrupamentos são construídos pelos elementos com maiores correlação, formando grupos. A clusterização da figura 07 apresenta dois dendrogramas, um referente aos grupos de genes ortólogos e o outro referente aos organismos, com base no seu conteúdo compartilhado. Independente da escolha dos métodos de agrupamentos escolhido, o dendrograma da figura 08 indica a conservação da relação entre os genes ortólogos analisados.

Para clusterização de dados existem uma gama de algoritmos disponíveis, entretanto, em bioinformática é cada vez mais comum encontrar a aplicação de algoritmos hierárquicos inerentes a esse tipo de problema diante do aumento exponencial do volume de dados armazenados nos bancos de dados biológicos (BRAZMA, 2000).

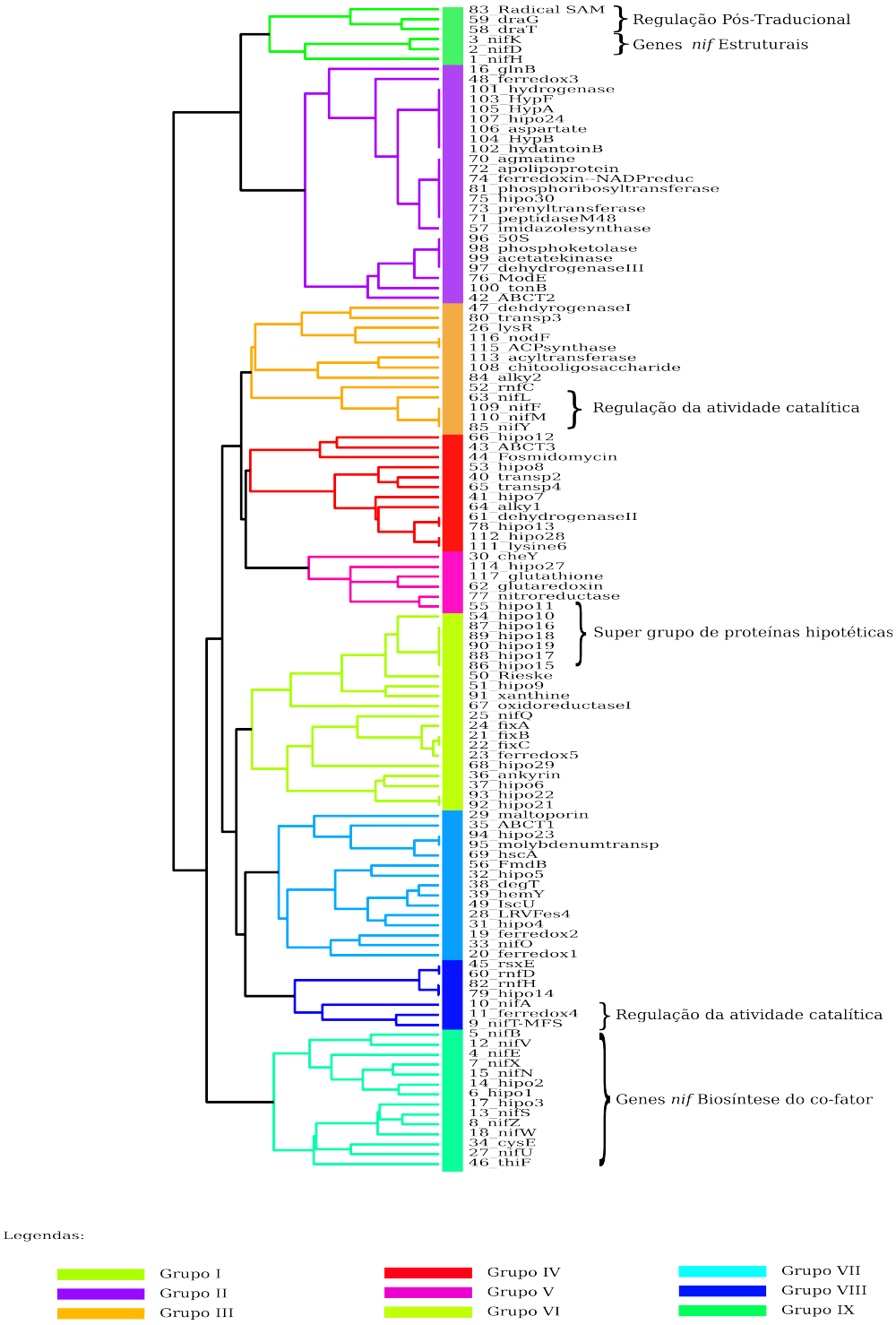


FIGURA 9 - DENDROGRAMA DA CLUSTERIZAÇÃO HIERÁRQUICA DOS GRUPOS DE ORTÓLOGOS. Dendrograma extraído da figura 08 para melhor visualização. Agrupamentos destacados a partir da distância mínima de 1.1 evidenciando 9 grandes grupos de genes. Grupos identificados de acordo com o envolvimento funcional dos seus genes.

FONTE O autor (2015)

O dendrograma dos grupos ortólogos na figura 09, gerou agrupamentos dos *nif* de acordo com as suas diferentes atividades necessárias para a ação da nitrogenase. E nos motivou a busca de novas informações para o estudo de genes em vizinhança conectada.

A validação da metodologia foi feita pela confirmação na literatura dos elementos existente nos agrupamentos.

O grupo I, agrupou os genes envolvidos em processos de montagem da enzima nitrogenase e regulação pós-traducionais (HUERGO, 2006).

O grupo III, agrupou quatro genes responsáveis pela regulação da atividade catalítica da enzima. O dendrograma do grupo III subdividive os genes de regulação catalítica com os demais genes, sendo interpretados em possíveis envolvimentos durante esse processo. Maiores investigações são necessárias.

No grupo IV foi o que mais apresentou o agrupamento de proteínas hipotéticas junto com alguns genes de interesse a prospecção ao processo de fixação biológica do nitrogênio, como a ankyrina.

No grupo VIII remanesceram alguns genes envolvidos no processo de regulação da atividade catalítica.

O grupo IX foi o grupo que mais apresentou consistência, agrupando os genes envolvidos na biosíntese dos co-fatores necessários para a atividade da enzima nitrogenase. Da tabela 04 selecionamos os genes *nif*Z, *nif*T e o grupo de proteínas hipotéticas 1, 2 e 3 que tiveram a maior ocorrência de ortólogos para uma inspeção mais profunda.

Os nifZ e nifT ainda não possuem função exata descrita na literatura, e a partir da pela análise dos agrupamentos em hipótese seria possível inferir a função com base nos grupos de genes correlacionados.

Para visualização pontual de cada elemento dos grupos de interesse, nós desenvolvemos um método para a visualização de correlação em dados de uma matriz multidimensional.

4.9 VISUALIZAÇÃO BI-DIMENSIONAL DE CORRELAÇÃO – A FUNÇÃO GASUPERCORR

A função *gasupercorr.m* foi desenvolvida de forma à expandir a análise dos agrupamentos da figura 08 através dos seus elementos co-relacionados.

A função busca coordenadas bi-dimensionais para pontos que representam os elementos considerados na matriz de correlação, de forma que, a matriz de distâncias dadas pela equação (2) seja respeitada da melhor forma possível. A otimização do gráfico é através de um algoritmo genético, que ajustará os genes correlacionados na melhor configuração otimizada.

$$d_a(x,y) = 1 - r(x,y) \qquad (2)$$

onde :

x, linha do elemento na matriz;

y, coluna do elemento na matriz e,

r, correlação de pearson.

Os grupos correlacionados ao grupo de interesse são ajustados no gráfico através pelo algoritmo genético, que busca grupos com a menor significância (*p*-value) e um mínima valor de correlação (r).

4.8 OS ESTUDOS DE CASO

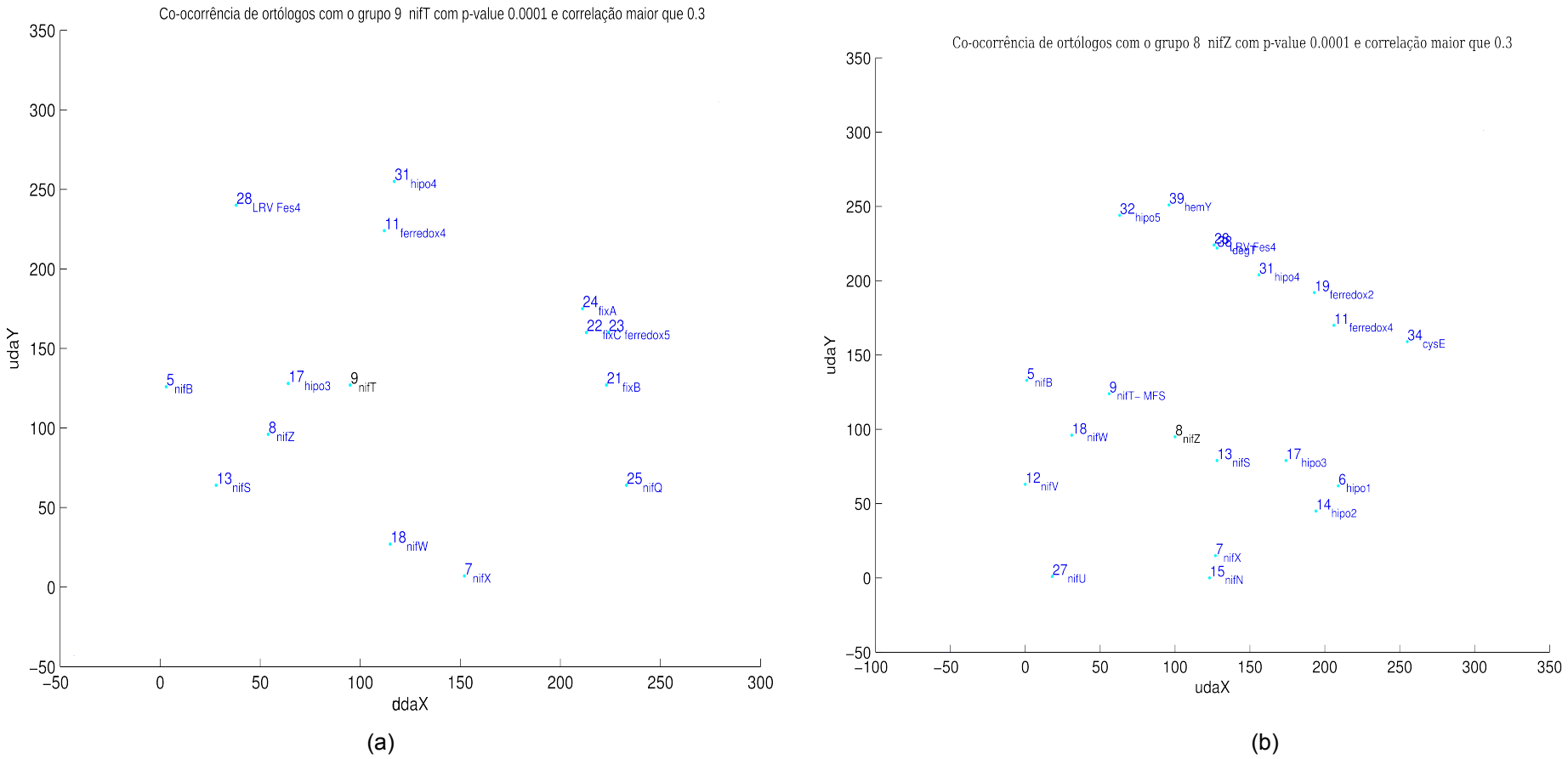
4.8.1 OS *nif*T-*nif*Z

Durante a revisão bibliográfica não foi encontrado muitas informações a respeito da função do gene *nif*T-*nif*Z, apenas que eles estão envolvidos em processos relativos maturaçao da proteína *nif*H indeterminado ao nível de expressão e regulação da atividade da enzima.

No agrupamento de ortológos, as sequencias referentes ao gene *nif*T formaram grupos junto com sequências correspondente ao gene fixU e proteínas hipotéticas. O Mapa de correlação da figura 10-a mostra proximidade com do gene nifT com os genes *nif*Z, com os *nif*S e *nif*B. Os *nif*S e *nif*B são descritos em etapas na mobilização de Fe-S e formação do núcleo do co-fator. As buscas nos bancos de dados biológicos PDB, InterProScan e Pfam apenas reportam a homologia junto com os genes *fix*U, que também é sem função definida. Na busca dentro da literatura foi encontrado três referencias com os termos *nif*T e *nif*Z, a primeira datada de 1989 (MERRICK, 1989) e a segunda datada de 1996 (SIMON, 1996), ambas relatam a uma baixa pertubação nos níveis de expressão de nitrogênio fixado e indicam que a ação do gene *nif*T e *nif*Z estão envolvidos no processo de formação ou acumulo da dinitrogenase ativa porém não sendo essenciais para o processo biológico da fixação do nitrogênio, no genoma de *K. pneumoniae.* A terceira referência (STRICKER et al, 1997) relata o *nif*T em um segundo óperon *nif* juntamente com os *nif*V e *nifZ*. O produto do gene n*if*V é conhecido que por catalizar a síntese de homocitrato, parte essencial do FeMo-co, porém quando analisados os níveis de nitrogênio fixado os autores do trabalho também chegaram a mesma conclusão de Simon e Merrick.

A partir da análise dos mapas de correlações dos genes *nif*T (FIGURA 10 - a)

e *nif*Z (FIGURA 10 - b) gerados pela função *gasupercorr,* tanto o mapa do *nif*T quanto o mapa do *nif*Z mostram alguns genes ocorridos em comum como no caso do gene nifS que codifica uma cisteina desulforase, necessário para a biossintese do agrupamento 4Fe-4S da dinitrogenase redutase (*nif*H) e *nif*B, identificado como essencial na biossíntese do FeMo-co, dessa forma, apesar de ser descrito pela literatura que o comportamento desses genes não altera o processo de fixação do nitrogênio, pelo mapa de correlações é possível sugerir que a ação dos genes *nif*T e *nif*Z pode ocorrer em etapas anteriores a captação de Ferro e Enxofre ou como via alternativa para a ação do nifB. A inferência demonstrada pelo mapa de correlações só pode ser confirmada mediante à análises *in vitro.*



(a) (b)

FIGURA 10 – MAPA DE CORRELAÇÕES GASUPERCORR. A figura (a) apresenta o mapa de correlações com os ortólogos com o grupo *nif*T. A figura (b) apresenta o mapa de correlações de ortólogos com o grupo *nif*Z. Genes de interesse destacados em preto nos mapas. As maiores correlações podem ser vistas pela menor distância ao gene de interesse.

Fonte: O autor (2015)

4.8.2 AS PROTEÍNAS HIPOTÉTICAS

Para determinação da função das proteínas hipotéticas, nós dividimos a investigação em 4 passos: 1) Análise de sequencias dos ortólogos, 2) Análise do dendrograma; 3) Análise do mapa de correlações e 4) Busca na literatura.

Para hipotética 1, foi identificado uma alta ocorrência de produtos ortólogos anotados como ferredoxinas, dando primeiro indício da função da proteína. No dendrograma da figura 10, o grupo hipo1 agrupou-se junto com os genes *nif* responsáveis pela biossíntese do co-fator. Pelo mapa das correlações hipo1 (FIGURA 12 - a) aparece perto de *nif*X e *nif*N, caracterizando a ação desse grupo envolvido na formação do FeMo-co. A formação do FeMo-co é descrita extensamente na literatura pela ação de um mínimo de oito genes juntamente com a maturação da nitrogenase. A hipotética 01 teve 45 ocorrências de sequências homólogos em 34 organismos diferentes. A maior parte das proteínas envolvidas na formação do FeMo-co são proteínas Fe-S. A incidências de ferredoxinas nós permitem sugerir a ação dessas proteínas durante a participação da formação do FeMo-co.

Para hipo2, todos as sequências do agrupamento de ortólogos foram anotadas com o termo *hypothetical protein.* No dendrograma da figura 10, esse grupo de ortólogos ficou no ramo de hipo1 juntamente com os genes de biossíntese do co-fator. Pelo mapa das correlações (FIGURA 12 - b) de hipo2, aparece sobreposta a hipo1 pela alta correlação entre esses genes e próximas aos *nif*X, *nif*N *nif*Z, caracterizado neste trabalho em etapas antes de *nif*S e *nif*U. Hipo2 teve ocorrência em 33 organismos diferentes, e informações sobre os genes correlacionados a essa proteína podem ajudar a elucidar essa a informação a respeito da funcionalidade dessa proteína.

Para hipo3, na análise de produto das sequencias ortólogas foi identificado anotações de produtos dos genes ErpA, HesB, HemY e termos *hypothetical protein*. Na literatura, esses três tipos de produtos são descritos como necessários para biossíntese dos co-fatores Fe-S. Em *E. coli* ErpA foi

descrito como uma proteina Fe-S A-type, importante para biogênese das proteínas Fe-S e respiração celular (LOUISEAU et al, 2007). No mapa de correlações de hipo3 (FIGURA 11) o comportamento dos ortólogos demonstraram um *insight* sobre a ação de hipo3 necessária para a formação do cluster metálico para nifS e posterior inserção ao FeMo-co.

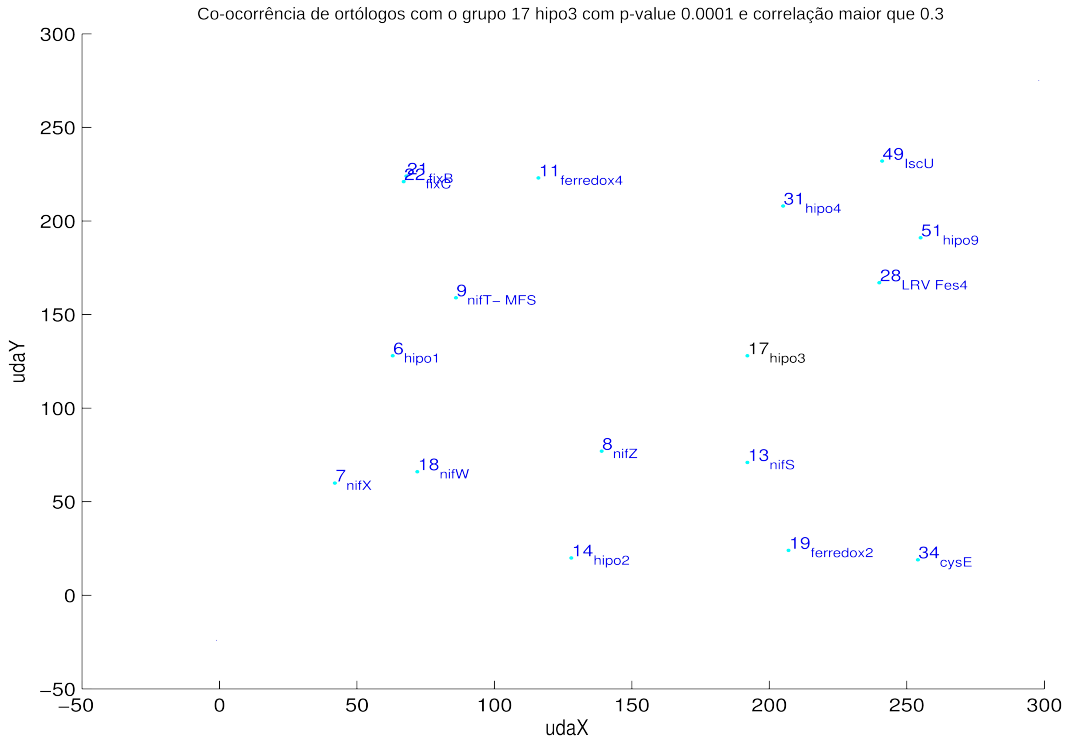
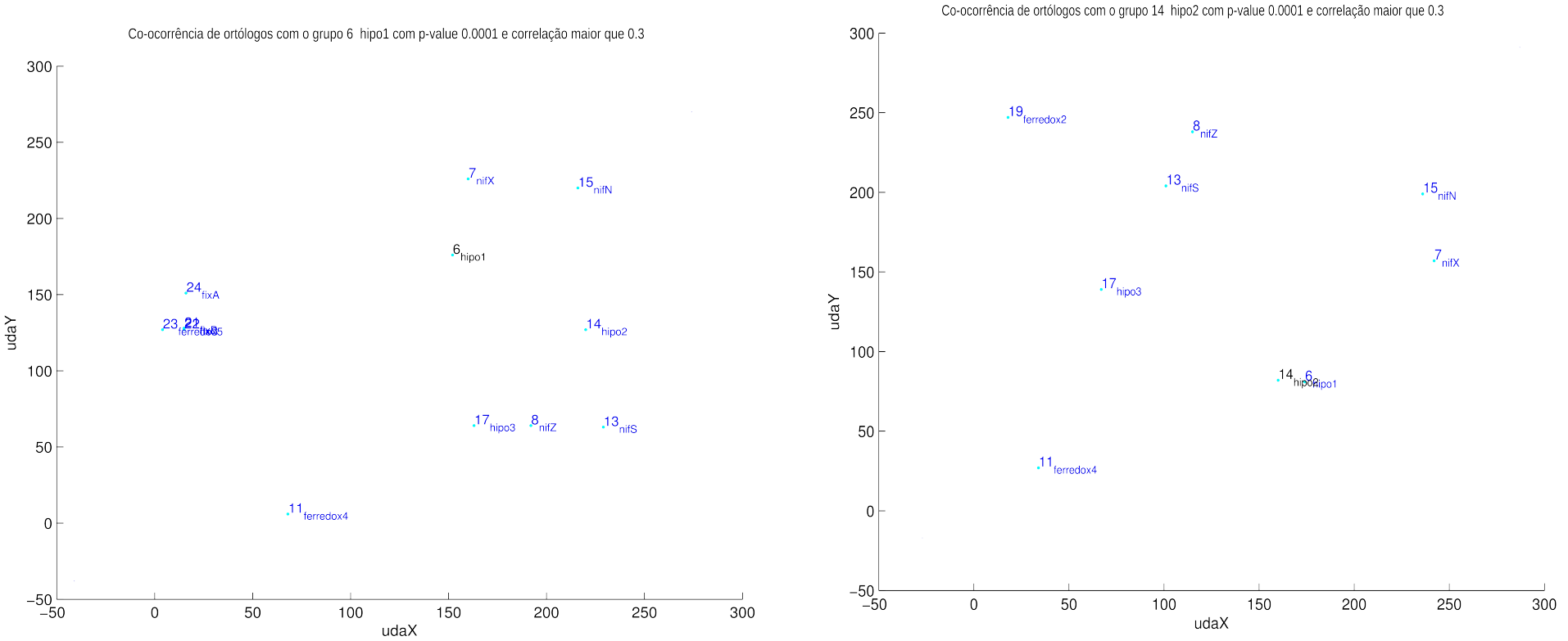


FIGURA 11- CORRELAÇÕES DE ORTÓLOGOS COM O GRUPO HIPO3.

Fonte: O autor (2015)



(a) (b)

FIGURA 12 – CORRELAÇÕES DOS GRUPOS HIPO1 E HIPO2. Hipo1 teve maior ocorrências de homólogos com proteínas ferredoxinas, a figura (a) a mostra as correlações com o grupo hipo1 sugerindo o envolvimento dessa proteína na estabilização e/ou inserção do co-fator. Hipo2 (b) no mapa de correlação aparece sobreposta a hipo1, demonstrando alta correlação, supondo que essa proteína também pode ser uma ferredoxina e estar estágios antes da formação do co-fator ou na mobilização de Fe-S.

Fonte: O autor (2015**)**

5 CONCLUSÕES

A bioinformática é uma área em constante evolução e o desenvolvimento de novas metodologias auxiliam na melhor compreensão de mecanismos biológicos.

A evolução das técnicas moleculares e de sequenciamento genômico geram dados além da capacidade humana de processamento. O DNA contém todo código funcional de um organismo ao longo da sua cadeia e a exploração dos seus genes é fonte de conhecimento para o avanço da biologia moderna.

Neste trabalho, a partir da identificação de homólogos em uma janela de 50 kb da nitrogenase dependente de molibdênio e pela re-identificação dos organismos de origem, criamos uma matriz de ocorrência que facilitou a aplicação de técnicas de análise de *clusters* e permitindo a inferência de função de proteínas pela correlação de interação proteína-proteína.

Esse trabalho foi embasado na aplicação de técnicas de clusterização de informação biológica descritos nos trabalhos de Zhang (2014), Hyotylainen (2014), Sloutsky (2012), Treagen (2012), Carvalho (2012), Zheng (2012), Gomide e colaboradores (2012), Yi (2007), Nowak (2007), Meyer (2002), Volfosvsk (2001) e Donna (1995), e por *insights* obtidos em experimentos de RNA-seq retratados por Wang (2013), Kuruvilla (2002), Brazam (2000) e Eisenberg (2000), e nos cálculos de inferência evolutiva nos micro-organismos nos estudos publicados por Lang (2013), Boyd (2013), Boyd (2011), Chuang (2010), Gehlenborg (2010), Keedweell (2005), Desluc (2005) e Dedysh (2004).

Durante o desenvolvimento do trabalho foi desenvolvida em conjunto com o orientador do mesmo a ferramenta para agrupamento de ortólogos RAFTS3GROUPS que utiliza abordagem heurística, livre de alinhamento.

A criação da matriz de ocorrência permitiu a aplicação de algoritmos de clusterização não supervisionada e os resultados agruparam genes biologicamente relacionados. A explicação para agrupamento dos elementos está na quantidade de sequencias ortólogos reconhecidas dentro dos organismos. E para melhor

entendimento das correlações entre os genes foi desenvolvida a função *gasupercorr* e realizada a análise dos genes *nifT* e *nifZ*, que não possuem função descrita na literatura, bem como a análise das proteínas hipotéticas que tiveram o maior numero de ortólogos.

Os genes *nifT* e *nifZ* foram sugeridos por esse trabalho em estágios anterior a mobilização de Ferro-Enxofre, necessário para a compactação do FeMo-co, ou como vias alternativas para a formação do FeMo-co e as hipotéticas 1, 2 e 3 são sugeridas com atividades de proteínas Fe-S(ferredoxinas) e envolvidas em estágios intermediários a mobilização e incorporação do FeMo-co.

Mineração de dados é uma tecnologia relativamente nova, com início nos anos 90, e assim como a bioinformática está em constante desenvolvimento. Os questionamentos a respeito dos resultados podem ser levantados porém não interferem nos *insights* criados. Maiores informações permitiram observar e extrair padrões para facilitar a compreensão da informação biológica por trás dos dados.

A metodologia apresentada foi integrada em um pacote de algoritmos e os códigos-fonte integram os anexos deste documento. A exceção das buscas manuais aos bancos de dados PFAM e PROSITE, toda a metodologia integra o pacote desenvolvido. A automatização dessa etapa facilitará a re-aplicação da metodologia em genomas completos ou em regiões genômicas de interesse, de forma independente e fácil para a descoberta de associação de genes em um grande conjunto de dados.

6 REFERÊNCIAS BIBLIOGRÁFICAS

ARNOLD, W.; KHIPP, W; PRIEFER, U. B.; PUMP, A.; PUHLER, A.; **Nucleotide Sequence of 24,206 base-pair DNA fragment carring the entire nitrogen fixation gene clusters of K. pneumoniae**. J. Mol. Bio, 203, v.3, p715-738, 1988.

BOYD, E. S.; PETERS, J. W. **New insights into the evolutionary history of biological nitrogen fixation**. Front Microbiol, v. 4, n. 201, p. 1 – 12, 2013.

BOYD, E. S.; ANBAR, A. D.; MILLER, S.; et al. **A late methanogen origin for molybdenum-dependent nitrogenase.** Geobiology, v.9, p. 221 – 232, 2012.

BOYD, E. S.; HAMILTON, T. L.; PETERS, J. W. **An alternative path for the evolution of biological nitrogen fixation**. Front Microbiol, v. 2, n. 205, p. 1-11, 2011.

BRAZAM, A.; VILO, J. **Gene expression data analysis**. FEBS Lett, v.480, p. 17 – 24, 2000.

BROWN, J. **Ancient horizontal gene transfe**r. Nature Reviews Genetics, v. 4, p. 121-132, 2003.

CARVALHO, M. O.; LORETO, E. L. S. **Methods for detection of horizontal transfer of transposable elements in complete genomes**. Genetics Mol Biol, v. 35, n. 4, p. 1078 – 1084, 2012.

CHUANG, H. Y.; HOFREE, M.; IDEKER, T. **A decade of Systems Biology.** The Annual Review of Cell and Developmental Biology, v. 26 p. 44, 2010.

DEDYSH, S. N.; RICKE, P.; LIESACK, W. **NifH and NifD phylogenies: an evolutionary basis for understanding nitrogen fixation capabilities of methanotrophic bacteria**. Microbiol, v. 150, p. 1301 – 1312, 2004.

DELSUC, F.; BRINKMAN, H.; PHILIPPE, H. **Phylogenomics and the reconstruction of the tree of life**. Nature Reviews Genetics. v. 6, 2005.

DIXON, R.; KAHN, D; **Genetic Regulation of Biological Nitrogen Fixatio**n. Nature Reviews Microbiology, v. 2, p. 621-631, 2004.

DONNA, M. J.; MULLIGAN. M. E. **Characterization of a nitrogen-f ixation (nif) gene cluster from Anabaena azollae las hows that closely related cyanobacteria have highly variable but structured intergenic regions**. Microbiology, v. 141, p 2235-2244, 1995.

DONI, M. **Análise de Cluster: Métodos Hierárquicos e de Particionamento**. Trabalho de Graduação (Bacharelado em Sistemas de Informação) – Faculdade de Computação e Informática, Universidade Presbiteriana Mackenzie, 2004.

DOS SANTOS, P; FANG, Z; MASON, S; SETUBAL, J; DIXON, R; **Distribution Of Nitrogen Fixation And Nitrogenase-Like Sequences Amongst Microbial Genomes**. BMC Genomics, v.13, p. 162, 2012.

DOS SANTOS, P.; SIMITH, D. A.; FRAZZON, J.; CASH, L. V; F. VALERIE, L. C.; JOHNSON, M.; DEAN, R. D. **Iron-Sulfur Cluster Assembly: NifU-Directed Activation of the Nitrogenase Fe Protein**. Journal of Biological Chemistry, 279, p. 19705-190711, 2004

EISENBERG, D; MARCOTTE, M, E; XENARIOS, I; YEATES, T; **Protein function in the post-genome era**. Nature v, 405, p. 823-826, 2000.

FAYYAD, U. M. **Data mining and knowledge discovery: Making sense out of data**. IEEE Expert, v. 11, p. 20-25, 1996

FINN R. D.; BATEMAN, A.; CLEMENTS, J.; COGGILL, P.; EBERHARDT, R. Y.; EDDY, S. R.; HEGER, A.; HETHERINGTON, K.; HOLM, L.; MISTRY, J.; SONNHAMMER, E. L .L.; TATE, J.**The PFAM protein familly database**. Nucleic Acids Research 2014.

FITCH, W. M. **Distinguishing Homologous from Analogous Proteins.** Syst Biol, v. 19, n. 2, p. 99–113, 1970 .

FITCH, W. M. **Homology**. Trends Genet, v. 16, n. 5, p. 227–231, 2000.

GOLDSCHIMIDT, R.; PASSOS, E. **Data mining: Um guia prático**. Rio de Janeiro. Campos, 2005.

GOMIDE, J.; MELO-MINARDI, R.; SANTOS, M. A.; et al. **Using linear algebra for protein structural comparison and classification**. Genetics Mol Biol, v. 32, n. 3, p. 645 – 651, 2009.

GEHLENBORG, N.; O'DONOGHUE, S. I.; BALIGA, N.; GOESMANN, A.; et al. **Visualization of omics data for systems biology**. Nat Rev, v. 7, n. 3s, p. 56 – 68, 2010.

GUIZELINI, D. **Banco de dados biológico no modelo relacional para mineração de dados em genomas completos de procariotos disponibilizados pelo NCBI GenBank** .Dissertação (Mestrado em Bioinformática) – Setor de Educação Profissional e Técnológica, Universidade Federal do Paraná, Curitiba, 2012.

HAIR, JR. J. et al. **Análise Multivariada de dados**; tradução Adonai Schulp Sant'Anna – 6. ed. - Porto Alegre: Bookman, 2009.

HANDL, J.; JOSHUA, K; KELL, D. **Computational cluster validation in post-genomic data analysis**. Bioinformatic Oxford Journal, v.21, p. 3201-3212, 2005.

HAYNES, R. **Physiological ecology: Mineral Nitrogen in the plant-soil System**- Academic Press. 1986.

HOFFMAN, B. M.; LUKOYANOV, D.; YANG, Z.Y.; DEAN, D. R.; SEEFELDT, L. C. **Mechanism of nitrogen fixation by nitrogenase: The next stage**. Chem Rev, v. 114, p. 4041-4062, 2014.

HYOTYLAINEN, T.; ORESIC, M. **Systems biology strategies to study lipidomes in health disease.** Prog Lipid Res, v. 55, p. 43 – 60, 2014.

HSU, H. **Advanced Data Mining Technologies in Bioinformatics**. Idea Group Publishing, 329P. 2006.

HUERGO, L. F. **Regulação do metabolism de nitrogênio em Azospirillum brasilense.** 187 f. Tese (Doutorado em Ciências – Bioquímica) – Setor de Ciências Bioquímica, Universidade Federal do Paraná, Curitiba, 2006 .

HYATT, D.; CHEN G. L.; LOCASCIO P. F.; LAND M. L.; LARIMER F. W.; HAUSER L. J. **Prodigal: prokaryotic gene recognition and translation initiation site identification**. BMC Bioinformatics. V.8, 2010

JASKOWIAK, A. P.; CAMPELLO, J. G. B. R.; COSTA, G. I.; **On the selection of appropriate distance for gene expression data clustering.** BMC Bioinformatics, v. 15. 2014.

JACOB, F.; MONOD, J. **Genetic regulatory mechanism in the synthesis of proteins.** Journal of Molecular Biology, v.3, p.318-356, 1961.

KEEDWELL, E.; NARAYANA, A. **Intelligent Bioinformatics: The application of artificial intelligence techniques to bioinformatics problems**. John Wiley & Sons Ltd, England, 2005.

KUZNIAR, A. et al. **The quest for orthologs: finding the corresponding gene acrross genomes**. Trends Genet, v. 24, n. 11, p. 539–551, 2008.

KURUVILLA, F. G.; PARK, P. J.; SCHREIBER, S. L. **Vector algebra in the analysis of genome-wide expression data**. Gen Biol, v. 3, n. 3,p. 1 – 11, 2002.

LANG, M. J.; DARLING, E. A.; EISEN, A. J; **Phylogeny of Bacterial and Archaeal Genomes Using Conserved Genes: Supertrees and Supermatrice**s. PLOS ONE. v. 8, n. 4, 2013.

LEWIN, B; **Genes VII**. Cambridge, Massachussetts, 2000.

LOISEAU, L; GEREZ., C; BEKKER, M.; CHOUDENS, S.; PY, B; SANAKIS, Y.; MATTOS, J.; FONTECAVE, M.; BARRAS, F. **ErpA, an iron-sulfer (Fe-S) protein of the A-type essential for respiratory metabolism in Escherichia coli.** Proceedings of the National Academy of Sciences, v. 104, n. 34, p. 13626-13631, 2007.

MACEDO, D. **Extração de conhecimento através de mineração de dados**. Revista de Engenharia e Tecnologia. v.2, n.2, 2010.

MEDEMA H. M.; FISHCBACH A. M. **Computational approaches to natural product discovery**. Nature Chemical Biology. v. 11, n. 1, p 639-648, 2015.

MEYER, A. S. **Comparação de coeficientes de similaridade usados em análises de agrupamentos com dados de marcadores moleculares dominantes.** 118 f. Dissertação (Mestrado em Agronomia) – Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2002.

MERRICK, J. M.; EDWARDS, A. R. **Nitrogen Control Microbiological Reviews**. v, 59. n.4, p. 0146-0749, 1995.

NUNES, S. F.; RAIMONDI, C. A.; NIEDWIESKI, C. A.; **Fixação de Nitrogenio: Estrutura, função e modelagem bioinorgânica das nitrogenase**. Química Nova. v. 26, n. 6, p. 872-879, 2003.

NOWAK, G. **Complementary hierarchical clustering**. Bioestatistics, v. 9, n.3, p. 467 – 483, 2007.

PAUL, W.; MERRICK, M. **The roles of the nifW, nifZ and nifM genes of Klebsiella pneumoniae in nitrogenase biosynthesis**. European Journal of Biochemistry, v. 178, p. 675-682, 1989.

POSTGATE, R. J. **The fundamentals of nitrogen fixation.** 1 ed. Cambrigde. Cambrigde University Press, 1982.

POZA-CARRIÓN, C.; JIMÉNEZ-VICENTE, E.; NAVARRO-RODRÍGUEZ, M.; ECHAVARRI-ERASUN, C.; RUBIO, L.M. **Kinetics of nif gene expression in a nitrogen-fixing bacterium**. J Bac, v. 196, n. 3, p. 595-603, 2014.

PRAKASH, R. K.; SCHILPEROORT, R. A.;NUTI, M. P. **Large plasmids of fast-growing rhizobia: homology studies and location of structural nitrogen fixation (nif) genes.** J. Bacteriol.145:1129–1136.

RIBBE, W. M.; HU, Y.; HODGSON, O. K.; HEDMAN, B. **Biosynthesis of Nitrogenase Metalloclusters**. Chemical Review. v. 114, p. 4063−4080, 2014.

RUBIO, L. M.; LUDDEN, P. W. **Biosynthesis of the Iron-Molybdenum cofactor of nitrogenase**. Annual Review Microbiology, v. 62, p. 93-111, 2008.

RUBIO, L. M.; LUDDEN, P. W. **Maturation of nitrogenase: a biochemical puzzle.** J Bac, v. 187, n.2, p. 405 – 414, 2005.

SIGRIST C. J. A.; DE CASTRO E.; CERUTTI, L.; CUCHE, B. A.; HULO N.; BRIDGE A.; BOUGUELERET, L.; **New and continuing developments at PROSITE** Nucleic Acids Res. 2012.

SHRIDHAR, S. B. **Review: Nitrogen Fixing Microorganisms**. International Journal of Microbiological Research, v. 3, p. 42-52, 2012.

SIMON, H.; HOMER, M.; ROBERTS, G. **Pertubation of nifT Expression in Klebsiella pneumoniae Has limited effect on nitrogen fixation**. Journal of Bacteriology, v. 178,n. 10, p 2975-2977, 1996.

SNUSTAD, P.; SIMMONS, M. **Fundamentos de Genética**. 2. ed. Rio de Janeiro: Editora Guanabara, 2001.

SLOUTSKY, R.; JIMENEZ, N.; SWAMIDASS, S. J.; NAEGLE, K. M. A**ccounting for noise when clustering biological data**. Brief Bioinform, v. 14, n. 4, p. 423 – 436, 2012.

SONNHAMMER, E. L. L.; KOONIN, E. V. **Orthology, paralogy and proposed classification for paralog subtypes**. Trends Genet., v. 18, n. 12, p. 619–620, 2002.

SPATZAL, T.; PEREZ, A. K.; EINSLE, O.; HOWARD, B. J.; REES, C. D.; **Ligand binding to the FeMo-cofactor: Structures of CO-bound and reactivated nitrogenase.** Science. v. 345, n. 6204, 2014.

STRICKER, O.; MASEPOHL, B.; KLIPP, W.; BOHME, H.; **Identification and Characterization of the nifV-nifZ-nitT genes region from the filamentous cyanobacterium Anabaena sp. Strain PCC 7120**. Journal of Bacteriology, v. 179, n. 9, p. 2930-2937, 1997.

TEMME, K.; ZHAO, D.; VOIGT, C. **Refactoring the nitrogen fixation gene cluster from Klebsiella oxytoca**. Proceedings of the National Academy of Sciences, v. 109, n. 18, p. 7085-7090, 2012.

VIALLE, A. R. **SILA – Ferramenta de alto desempenho para anotação automática genômica.** Dissertação (Mestrado em Bioinformática) – Setor de Educação Profissional e Técnológica, Universidade Federal do Paraná, Curitiba, 2013.

VOLFOVSKY, N.; HAAS, J. B.; SALZBERG, L. S. A. **Clustering method for repeat analysis in DNA sequences**. Genome Biologyl, v. 2, n. 8, 2001.

YI, G.; SZE, S.; THON, M. R. **Identifying clusters of functionally related genes in genomes**. Bioinformatics, v. 23, n. 9, p. 1053 – 1060, 2007.

YOUNG, J. P. W.; **Phylogenetic Classification of Nitrogen-fixing organism. Biological Nitrogen Fixation**. New York, London. Chapan and Hall, 1. p 43-86, 1992.

YU, U.; LEE, H, S.; YOUNG, J, K.; KIM, S. **Bioinformatics in the Post-genome Era**. Journal of Biochemistry and Molecular Biology, v. 37, n. 1. p. 75-82, 2004.

WANG, N. W.; WANG, Y.; HAO, H.; et al. **A bi-Poisson model for clustering gene expression. profiles by RNA-seq**. Brief Bioinform, v. 15. n. 4, p. 534 – 541, 2013.

ZAHA, A.; FERREIRA, H.; PASSAGLIA L. **Biologia Molecular Básica**. 5. ed. Porto Alegre; Artmed 2014.

ZHANG, Y.; HORVATH, S.; OPHOFF, R.; TELESCA, D. **Comparison of clustering methods for time course genomic data: applications to aging effects**. Cornell University Library, v. 1, p. 1 – 23, 2014.

ZHENG, Y. **Clustering methods in data mining with its application in high education**. IPCSIT, v. 43, p. 1 – 7, 2012.

ANEXOS

ANEXO I – Função loadJGBParser.m – Carrega o JGBParser para execução no ambiente MATLAB.

```
function mret = loadJBGParser(file)
%   Função carregar o JGBParser
%   Copyright (C) 2015  Federal University Of Parana
%   Laboratório de Bioinformática - SEPT
%   Pós-graduação em Bioinformática
%   Universidade Federal do Paraná
%   Rua Dr. Alcides Vieira Arcoverde, 1225, Jardim das Américas
%   Curitiba - PR
%   CEP 81520-260
%   Brasil
%
%   Nilson Coimbra
%   nilson.coimbra@ufpr.com


  import br.ufpr.*;

    parser = br.ufpr.bioinfo.genbank.parser.GenBankParser;

    mret =  parser.processFile(file);

end
```

ANEXO II – Função extractnifH.m - Execução automática da seleção de dados

```
function mret = extractnifH(arqgbk,path,dbstruct)
%   Função para extração automática uma janela cromossômica com base na identificação do gene nifH. A janela
%   cromossômica é anotada usando o SILA.
%   Copyright (C) 2015  Federal University Of Parana
%
%   Laboratório de Bioinformática - SEPT
%   Pós-graduação em Bioinformática
%   Universidade Federal do Paraná
%   Rua Dr. Alcides Vieira Arcoverde, 1225, Jardim das Américas
%   Curitiba - PR
%   CEP 81520-260
%   Brasil
%
%   Usage: extractnifH <arquivo.gbk> <output_path> <dbstruct_sila>
%
%   Nilson Coimbra
%   nilson.coimbra@ufpr.com

global var;
var = 0;

%% Parser Genbank file
gbp = genBankParser(arqgbk);
nfeatures = gbp.getFeatures.size;
disp('Searching nifH')

%% Cria ID para o Arquivo
id = char(gbp.getDefinition);

%% Cria Ambiente de Operação
if(~path)
   output = pwd;
else
   output = path;
end

%% Código Principal
for i=0:nfeatures-1
   if (gbp.getFeatures.get(i).getKey.equals('CDS'))
      tamquali = gbp.getFeatures.get(i).getQualifiers.size;
      for j=0:tamquali-1
         try
            if(gbp.getFeatures.get(i).getQualifiers.get(j).getKey.equals('gene') &
gbp.getFeatures.get(i).getQualifiers.get(j).getValue.equals('nifH'))
               disp('nifH encontrado')
               minor = gbp.getFeatures.get(i).getLocation.getMinor;
               disp(['Posição de Início: ' int2str(minor)]);
               major = gbp.getFeatures.get(i).getLocation.getMajor;
               disp(['Posiçãoo de Fim: ' int2str(major)]);
               seq = char(gbp.getOrigin);
               disp(length(seq))
               seq = seq(minor-25000:major+25000);
               header = char(arqgbk);
               fastawrite([output '/' id '_50K_Sequence.fasta'], header, seq);
               fas = fastaread([output '/' id '_50K_Sequence.fasta']);
               annot = sila(dbstruct, fas);
               printgbk2(annot, [output '/' id '_SILA.gbk']);
               %mret = seq;
               fileid = fopen([output '/' id '_nifH_report.find'],'w');
               fprintf(fileid,['Organismo\tPosição Inicial\tPosiÃ§Ã£o Final\t\n']);
               rs = [id '\t' int2str(minor) '\t' int2str(major)];
               fprintf(fileid,[rs '\n']);
               fclose(fileid);
               var = 1;
```

```
            end

            files(i).name
        catch E

        end

    end

  end
end

if var == 1;
else
   fileid2 = fopen([output '/' id '_nifH_report.notfind'],'w');
   fprintf(fileid2,['Organismo\n']);
   fprintf(fileid2, [char(arqgbk) '\n']);
   fclose(fileid2);
end
end
```

ANEXO III – Função extractlocustag.m para execução de extração da janela padrão com base na identificação do locus_tag.

```
function mret = extractlocustag(arqgbk,locus_tag,path,dbstruct)
%   1 - Função de extração de uma janela de sequência com base na identificação do valor da tag locus_tag no arquivo %
gbk
%   Copyright (C) 2014  Federal University Of Parana
%
%   Laboratório de Bioinformática - SEPT
%   Pós-graduação em Bioinformática
%   Universidade Federal do Paraná
%   Rua Dr. Alcides Vieira Arcoverde, 1225, Jardim das Américas
%   Curitiba - PR
%   CEP 81520-260
%   Brasil
%
%   Usage: extractnifH <arquivo.gbk> <var_locus_tag> <output_path> <dbstruct_sila>
%
%   Nilson Coimbra
%   nilson.coimbra@ufpr.com

gbp = genBankParser(arqgbk);
locus = locus_tag;
nfeatures = gbp.getFeatures.size;
%% Cria ID para o Arquivo
id = char(gbp.getDefinition);

%% Cria Ambiente de OperaÃ§Ã£o
if(~path)
   output = pwd;
else
   output = path;
end
%%


% Código Principal
for i=0:nfeatures-1
   if (gbp.getFeatures.get(i).getKey.equals('CDS'))
       tamquali = gbp.getFeatures.get(i).getQualifiers.size;
       for j=0:tamquali-1
          if(gbp.getFeatures.get(i).getQualifiers.get(j).getKey.equals('locus_tag') &
gbp.getFeatures.get(i).getQualifiers.get(j).getValue.equals(locus))
              disp(['Encontrado'])
              minor = int64(gbp.getFeatures.get(i).getLocation.getMinor);
              disp(['Posiçãoo de Início: ' minor]);
              major = int64(gbp.getFeatures.get(i).getLocation.getMajor);
              disp(['Posição de Fim: ' major]);
              seq = char(gbp.getOrigin);
              disp(length(seq))
              seq = seq(minor-25000:major+25000);
              fastawrite([ output '/' id '_' locus '_50K_Sequence.fasta'], seq);
              fas = fastaread([output '/' id '_' locus '_50K_Sequence.fasta']);
              annot = sila(dbstruct, fas);
              printgbk2(annot, [ output '/' id '_' locus '_SILA.gbk']);
              mret = seq;
          end

       end
   end
end


end
```

ANEXO IV – Função rafts3groups.m

```
function mret = rafts3group(fsall,tipo)

%   Agrupa sequencias de uma variavel fasta fsall com corte de selfscore 0.5
%
%   tipo = 1; %Sequencia em NT
%   tipo = 2; %Sequencia em aa
%   Copyright (C) 2014. Federal University of Parana
%
%   Laboratório de Bioinformática - SEPT
%   Pós-graduação em Bioinformática
%   Universidade Federal do Paraná
%   Rua Dr. Alcides Vieira Arcoverde, 1225, Jardim das Américas
%   Curitiba - PR
%   CEP 81520-260
%   Brasil
%
%   Usage: rafts3groups <fasta_file > <var>
%
%   Roberto Raittz
%   raitzz@gmail.com

tic;
if tipo==1
    dball = formatdb(nt2aafas(fsall)); %Formata Rafts3
    xfas = nt2aafas(fsall);
else
    dball = formatdb(fsall); %Formata Rafts3
    xfas = fsall;
end
n = length(xfas);
%
grps = zeros(n,1);
cont = 1;
for i=1:n
    if mod(i,100)==0
        disp([i cont]);
    end
    if ~grps(i)
        q = rafts3(xfas(i),dball,50);
        u = q.scores;
        igr = u(u(:,2)>0.5,1:2);
        lgrs = grps(igr(:,1));
        ugp = unique(lgrs(find(lgrs)))';
        %disp(ugp);
        if sum(ugp)==0
            grps(igr(:,1)) = cont;
            cont = cont+1;
        else
            grps(igr(:,1)) = mode(ugp);
        end
    end
end
%
z = contaocorr(grps);
zu = z(z(:,2)>1,:);
[xx ii] = sort(zu(:,2),'descend');
toc;
mret.igrp = grps;
mret.contall = z;
mret.contg2 = zu(ii,:);
```

ANEXO V – Função agruparafts.m Agrupa um arquivo multifasta com 0.5 de self-score para RAFTS3

```
function mret = agruparafst(file)
%   1 - Função para agrupamento de sequencias homólogas usando o RAFTS3
%   Copyright (C) 2014. Federal University of Parana
%
%   Laboratório de Bioinformática - SEPT
%   Pós-graduação em Bioinformática
%   Universidade Federal do Paraná
%   Rua Dr. Alcides Vieira Arcoverde, 1225, Jardim das Américas
%   Curitiba - PR
%   CEP 81520-260
%   Brasil
%
%   Usage: agruparafts <fasta_file >
%
%   Nilson Coimbra
%   nilson.coimbra@ufpr.com

%carrega dbstruct do SILA
load dbstruct
system('mkdir Clusters')
output = [pwd '/Clusters'];
allf = fastaread(file);
rftgroups = rafts3group(allf,2);
save rftsgroups.mat, rftgroups

%índices das sequencias agrupadas
posis = (int64(rftgroups.igrp));
fileid = fopen([output '/Orthologs_Clustered_report.tab'],'w');
fprintf(fileid,['Cluster \tProduto\tQuantidade\n']);

%verifica os grupos com mais de duas sequencias
for i=1:length(rftgroups.contg2)
    % agrupa as sequencias com mesmo índice
    clu = allf(posis(:)==(rftgroups.contg2(i)));
    % anota sequencias do grupo identificado e adiciona informação no report
    qtd = int2str(length(clu));
    anota = sila(dbstruct,clu);
    name = anota.annotation(1).annotatedFasta.Header;
    barra = strfind(name,'|');
    colchete = strfind(name,'[');
    name = name(barra(4)+1:colchete(1)-1);
    rs = [int2str(i) '\t ' name  '\t' qtd];
   %formata escrita
    fprintf(fileid,[rs '\n' ]);


    % Formata o cabeçalho do grupo identificado
    for j=1:length(clu)
        clu(j).Header = ['>Cluster_' int2str(i) '_' clu(j).Header];
    end

    fastawrite([ output '/Cluster_' int2str(i) '.fasta'], clu);
end
fclose(fileid);
cd([pwd '/Clusters'])
system('cat Cluster* > allClusters.fasta');
end
```

ANEXO VI – Função criamatrizbinária.m

```
function mret = criamatrizbinaria(path_organism, all_homologs_xfas)
%   1 - Função para criação da matriz binária com self-score  0.5 usando o RAFTS3
%   Copyright (C) 2014. Federal University Of Parana
%
%   Laboratório de Bioinformática - SEPT
%   Pós-graduação em Bioinformática
%   Universidade Federal do Paraná
%   Rua Dr. Alcides Vieira Arcoverde, 1225, Jardim das Américas
%   Curitiba - PR
%   CEP 81520-260
%   Brasil
%
%   Usage: criamatrizbinaria <path_files_orgs_sequences_aa > <rafts3groups_fastas >
%   Nilson Coimbra
%   nilson.coimbra@ufpr.com

orgpath = path_organism;
gpfas = fastaread(all_homologs_xfas);
%Lista todos os arquivos dentro do
xfiles = dir([orgpath '/*.fasta']);
n = length(xfiles);
% Cria matriz populada com zeros pela quantidade de organismos e grupo de famílias homólogas
vect = zeros(n,length(gpfas));
dball = formatdb(gpfas);
fileID = fopen('Report_Matriz.tab','w');
fprintf(fileID,['POSIÇÃO NO VETOR\t ORGANISMO\t QUANTIDADE DE SEQUENCIAS\tCLUSTER ID\n']);
for i=1:n
   %criabd para bicho i
   disp(['Criando BD para ' xfiles(i).name])
   xfile = fastaread([orgpath '/' xfiles(i).name]);
   m = length(xfile);
   name =  char(xfiles(i).name);
   virgula = split('.',name);
   namerep = replace(char(virgula(1)), ' ', '_');
   fprintf(fileID,[ int2str(i) '\t' namerep '\t' int2str(m) ]);
   %
   for j=1:m
      %Procura ocorrencia grupos entre os genes de org-i
      q = rafts3(xfile(j),dball,50);
      if q.self~=-1
         z = struct2cell(q.fas(find(q.self>0.5)));                %procura hits com mais de 50% de similaridade
         C = z(1,:);
         iok = contaocorr(cell2mat(cellfun(@clusternumb, C, 'UniformOutput', false)));
         if ~isempty(iok)
            is_in = iok(1,1);
            vect(i,is_in) = 1;
            fprintf(fileID,[ int2str(is_in) '\t']);
         end
      end
   end
   fprintf(fileID,['\n']);
end
end
```