UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
FACULDADE DE COMPUTAÇÃO
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

ROBERTO BRITO XAVIER JUNIOR

FATORES DE TRANSCRIÇÃO EM CIANOBACTÉRIAS: PREDIÇÃO POR GENÔMICA COMPARATIVA

Belém
2018

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
FACULDADE DE COMPUTAÇÃO
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

ROBERTO BRITO XAVIER JUNIOR

FATORES DE TRANSCRIÇÃO EM CIANOBACTÉRIAS: PREDIÇÃO POR GENÔMICA COMPARATIVA

Trabalho de Conclusão de Curso apresentado como requisito para obtenção do grau de Bacharel em Sistemas de Informação.

Orientadora: Prof.ª Dr.ª Danielle Costa C. Couto
Co-orientador: Msc. Renato Renison Moreira Oliveira

Belém
2018

Roberto Brito Xavier Junior
Fatores de Transcrição em Cianobactérias: Predição por Genômica Comparativa/ Roberto Brito Xavier Junior. – Belém, 2018.
46 p. : il. (algumas color.) ; 30 cm.

Orientadora: Prof.ª Dr.ª Danielle Costa C. Couto
Co-orientador: Msc. Renato Renison Moreira Oliveira

Monografia – Universidade Federal do Pará
Instituto de Ciências Exatas e Naturais
Curso de Bacharelado em Sistemas de Informação, 2018.

1. Cianobactérias. 2. Fatores de Transcrição. 3. Banco de Dados Biológicos. 4. Gênomica Comparativa. I. Título.

ROBERTO BRITO XAVIER JUNIOR

FATORES DE TRANSCRIÇÃO EM CIANOBACTÉRIAS: PREDIÇÃO POR GENÔMICA COMPARATIVA

Trabalho de Conclusão de Curso apresentado como requisito para obtenção do grau de Bacharel em Sistemas de Informação.

Data da Defesa: 01 de março de 2018
Conceito:

Banca Examinadora

Prof.ª Dr.ª Danielle Costa C. Couto
Campus Ananindeua - UFPA
Orientador

Msc. Renato Renison Moreira Oliveira
Faculdade de Computação - UFPA
Membro da Banca

Prof.ª Dr.ª Regiane Silva Kawasaki Frances
Faculdade de Computação - UFPA
Membro da Banca

Prof. Dr. Claudomiro de Sales de Souza Junior
Faculdade de Computação - UFPA
Membro da Banca

Belém
2018

Dedico este trabalho à minha família

AGRADECIMENTOS

Agradeço primeiramente à minha querida avó Juca, que sempre me mostrou o valor e a importância dos estudos e por ter me dado todo o tipo de apoio e ajuda desde o fundamental até à faculdade. Infelizmente ela não se encontra mais entre nós, mas sei que ela ficaria muito feliz em me ver concluindo esta importante etapa da minha vida.

Agradeço aos meus amados pais, Roberto Xavier e Leila Barbosa, por sempre estarem presentes na minha vida, pela ajuda e paciência em momentos de dificuldade, pelo incentivo e dedicação permitindo que eu sempre alcance meus objetivos. Por todo o companheirismo e amizade que me fazem lembrar que nunca estou só, tornando essa jornada mais fácil.

Agradeço à minha namorada, Vanessa Vieira, por ser mais que uma amiga e estar presente na minha vida, em momentos bons e ruins, mas sempre me apoiando.

Agradeço aos amigos que adquiri durante esses anos de curso, Raíssa Lorena, Natanael Medeiros, Rafael Cavalheiro e Caio Messias. Aos amigos que fiz durante o ensino médio e continuaram presentes, Walisson Cardoso, Isaac Cardoso, André Luiz, Gustavo Lúcio, Luiz Junior, Lucas Lima e Kacta Oliveira. Aos que esqueci, mas que estiveram presentes também agradeço.

Agradeço à minha querida professora e orientadora Danielle Costa C. Couto, por ter me aceitado como orientando para a realização deste trabalho, e pela paciência e dedicação para me instruir e guiar ao longo do desenvolvimento deste trabalho. Ao meu grande amigo e co-orientador Renato Oliveira, por aceitar me co-orientar, e ter me ajudado imensamente na realização deste trabalho, além de compartilhar ao máximo seus conhecimentos de Bioinformática.

Agradeço a todos os professores que fizeram parte da minha educação acadêmica, em especial à professora Regiane Kawasaki por me dar boas vinda na área de Bioinformática e me apresentar à minha professora e orientadora.

“Ouse conquistar a sí mesmo.”
(Friedrich Nietzsche)

RESUMO

O avanço da biologia molecular nas últimas décadas permitiu que grandes quantidades de dados tornarem-se disponíveis, possibilitando a criação de bancos de dados e ferramentas de análise, com o objetivo de sequenciar genomas e conhecer seus genes. Os fatores de transcrição são um grupo de genes, e estudá-los é de grande importância para a investigação da história evolutiva dos organismos e suas características. As cianobactérias são um antigo grupo de bactérias e ainda pouco se sabe sobre seus fatores de transcrição. Neste trabalho, utilizou-se uma base composta por 52 genomas completos de cianobactérias obtidos no NCBI, e uma base de 1288 fatores de transcrição putativos identificados a partir de 21 genomas de cianobactérias completamente sequenciados disponíveis no banco cTFbase. Várias técnicas foram aplicadas para implementar um *pipelin*e automático e abrangente usando uma combinação de softwares como AUGUSTUS e HMMER a fim de descobrir novos fatores de transcrição, por meio da predição por genômica comparativa, em cianobactérias que estão publicadas em bancos de dados públicos, mas não estão anotados, ajudando na análise evolutiva dos organismos e suas características. O *pipeline* obteve resultados importantes na descoberta de fatores de transcrição em novos genomas de cianobactérias, e também na identificação de novos fatores de transcrição em 8 cianobactérias que não foram identificados pelo cTFbase.

Palavras-chave: Cianobactérias. Fatores de Transcrição. Banco de Dados Biológicos. Genômica Comparativa.

ABSTRACT

The advancement of molecular biology in the last decades is available, allowing the creation of databases and analysis tools, with the objective of sequencing genomes and knowing their genes. Transcription factors are a group of genes, and studying them is of great importance for research into the evolutionary history of organisms and their characteristics. As cyanobacteria are an ancient group of bacteria and still little know their transcription factors. In this work, we used a base composed of 52 complete cyanobacterial genomes obtained from non-NCBI, and a base of 1288 putative transcription factors identified from 21 fully sequenced cyanobacteria genomes available without cTFbase database. Several techniques applied to implement an automatic and comprehensive pipelin using a combination of software such as AUGUSTUS and HMMER to discover new transcription factors through prediction by comparative genomics in cyanobacteria that are published in databases public, but are not annotated, helping in the evolutionary analysis of organisms and their characteristics. The pipeline obtained important results in the discovery of transcription factors in new cyanobacteria genomes and in the identification of new transcription factors in 8 cyanobacteria that were not identified by cTFbase.

Keywords: Cyanobacteria. Transcription Factors. Biological Database. Comparative genomics.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

BLAST	Basic Local Alignment Tool (Ferramenta de Alinhamento Básico Local)
DBD	DNA Binding Domain
DDBJ	DNA Data Bank of Japan (Banco de Dados de DNA do Japão)
DNA	Deoxyribonucleic Acid (Ácido Desoxirribonucleico)
EBI	European Bioinformatics Institute (Instituto Europeu de Bioinformática)
EMBL	European Molecular Biology Laboratory (Laboratório Europeu de Biologia Molecular)
FASTA	Fast Alignment Tool (Ferramenta de Alinhamento Rápido - Formato utilizado para armazenar sequências de bases e de aminoácidos em arquivo texto)
GenBank	Banco de dados público do National Center for Biological Information, do Instituto de Saúde dos Estados Unidos da América.
HMM	Hidden Markov Model (Modelo Oculto de Markov)
INSDC	International Nucleotide Sequence Database Colaboration (Colaboração Internacional de Base de Dados de Sequências de Nucleotídeos)
mRNA	RNA mensageiro
NCBI	National Center for Biotechnology Information (Centro Nacional de Informações em Biotecnologia)
NGS	Next-Generation Sequencing (Tecnologias de sequenciamento de nova geração)
ORF	Open Reading Frames (Sequência codificadora de proteína)
RNA	Ribonucleic Acid (Ácido Ribonucléico)
TF	Transcription Factor (Fator de Transcrição)
TI	Tecnologia da Informação
TRN	Transcriptional Regulatory Network
UFPA	Universidade Federal do Pará
UTR	Untranstaled Regions (Regiões não traduzidas)

SUMÁRIO

1 INTRODUÇÃO

Com o avanço da biologia molecular nas últimas décadas, grandes quantidades de dados se tornaram disponíveis, promovendo a criação de bancos de dados e ferramentas de análise, permitindo a construção de modelos mais amplos, capazes de lidar com aspectos e fenômenos biológicos até então inacessíveis (Verli 2014).

A genômica é uma subdivisão do campo da genética, gerada pela união da biologia clássica e da biologia molecular, com o objetivo de sequenciar e conhecer os genes, as interações gênicas, os elementos genéticos e as estruturas dos genomas (Klug et al. 2010). O que torna a genômica diferente de outras pesquisas biológicas é o fato dela utilizar informações em larga escala, assim como o uso de computadores potentes para pesquisar características ou analisar simultaneamente os padrões de expressões de dezenas de milhares de genes (Watson et al. 2009). A genômica ainda possui várias subdivisões, entre as principais estão a genômica funcional, responsável por descrever as atividades dos genes e proteínas, e a genômica comparativa, que busca conhecer as relações e a homologia entre as sequências genéticas.

Em cada organismo vivo, mecanismos de desenvolvimento, morfológicos e fisiológicos, como aqueles que permitem aclimatação a mudanças ambientais, são o resultado da modulação da expressão do genoma. Um nível desta modulação está relacionado à expressão gênica, em que os fatores de transcrição (*Transcription factors* - TF) estão entre os principais atores (Heydarizadeh et al. 2014).

Desde o primeiro estudo sobre a identificação de TFs em quatro genomas de archeas (Aravind et al. 1999), o aumento do número de genomas sequenciados facilitou a identificação de TFs através de estudos *in silico* (Wu et al. 2007; Rayko et al. 2010). Tais dados taxonomicamente diversos permitem análises comparativas entre diferentes espécies ou linhagens (Wu et al. 2007; Rayko et al. 2010; Lang et al. 2010) e compreensão dos aspectos evolutivos através de TFs (Charoensawan et al. 2010).

Neste estudo, realizamos a primeira identificação de TFs com comparação de genomas completos em cianobactérias utilizando um *pipeline* otimizado e automatizado. Este *pipeline* de análise combina pesquisa por semelhanças de TFs conhecidos e domínios de proteínas usando um banco de dados contendo TFs de cianobactérias já identificados na literatura.

1.1 Motivação

A enorme acumulação de dados de sequenciamento de genomas fez com que muitos pesquisadores se voltassem para abordagens *in silico* com o objetivo de avaliar rapidamente o potencial natural de um organismo recém-sequenciado. É nesse contexto que várias ferramentas de bioinformática foram desenvolvidas (Micallef et al. 2014).

Além disso, ferramentas de bioinformática também podem realizar predição de genes, por intermédio da busca por domínios essenciais e/ou genes repórteres, diminuindo principalmente custos de pesquisas com a mesma busca sendo realizada com organismos vivos em laboratório.

As cianobactérias são um antigo grupo de bactérias gram negativas com forte variação no tamanho do genoma variando de cerca de 1,6 a 9,1 Mb e ainda pouco se sabe sobre seus fatores de transcrição. Portanto, construir um *pipeline* e um banco de dados para classificar e analisar todos os resultados de TFs putativos em genomas de cianobactérias, seguido de análise comparativa do genoma, é muito importante.

Principalmente porque estudar os fatores de transcrição, que são alguns dos principais atores da expressão gênica, é de grande interesse para a investigação da história evolutiva dos organismos através de características específicas da linhagem (Thiriet-Rupert et al. 2016).

1.2 Objetivos

1.2.1 Objetivo Geral

Realizar a identificação de novos fatores de transcrição em genomas completos de cianobactérias.

1.2.2 Objetivos Específicos

- Implementar um *pipeline* automático e abrangente usando uma combinação de softwares como HMMER, Augustus e Pfam;
- Identificar e classificar os TFs dos genomas do NCBI;
- Comparar de famílias de TFs entre linhagens de cianobactérias;

1.3 Organização do TCC

Além desta introdução previamente apresentada, as seções deste trabalho estão distribuídas da seguinte forma: a Seção 2 apresenta a fundamentação teórica, descrevendo as áreas de conhecimento que compõem o presente trabalho; A Seção 3 intitulada como materiais e métodos possui os procedimentos metodológicos aplicados neste trabalho. Já a Seção 4 apresenta os resultados e discussões sobre o trabalho desenvolvido, baseado na metodologia aplicada e, por fim, as conclusões deste trabalho são apresentadas na Seção 5.

2 REFERENCIAL TEÓRICO

2.1 Genética

Os primeiros conceitos da genética foram definidos por Gregor Mendel por meio de seus estudos e resultados obtidos em seu experimento com ervilhas de jardim, na qual é responsável por explicar o funcionamento da hereditariedade, ou seja, Mendel (1865) concluiu que cada característica da ervilha é controlada por um par de "fatores", que chamamos agora de genes, que realizam a transmissão de traços de uma geração para outra.

Ao decorrer dos anos, a teoria de Mendel foi testada diversas vezes por geneticistas e suas pesquisas, que encontraram traços de hereditariedade. Segundo Klug et al. (2012), em alguns casos, existia alguma diferença ao trabalho de Mendel com ervilhas, porém sua teoria poderia sempre ser aplicada, além de continuar sendo a base para explicar como as características são passadas de geração a geração em diversos organismos, incluindo o homem.

O DNA (ácido desoxirribonucleico) foi descoberto por Johann F. Miescher, em sua pesquisa que consistia em isolar uma substância que ele chamou de "nucleína", a partir do núcleo das células brancas do sangue (Dahm 2004). Segundo Griffiths et al. (2008), resultado de várias pesquisas sobre hereditariedade, foram descobertas informações sobre várias características situadas nos genes, chegando à conclusão que o DNA é o material genético.

Em 1953, Watson e Crick propuseram o modelo de dupla hélice do DNA que levou a uma compreensão detalhada sobre o funcionamento da hereditariedade. Segundo Watson (1968), uma molécula de DNA é feita de dois filamentos torcidos enrolados um no outro em forma de dupla hélice. Cada um dos filamentos é um arcabouço composto de cópias repetidas de um açúcar, chamado desoxirribose, e fosfato, projetando-se ao longo do arcabouço uma base nucleotídica.

Existem quatro tipos diferentes de bases nos nucleotídeos de DNA: adenina (A), timina (T), guanina (G) e citosina (C). No modelo em dupla hélice, o arcabouço açúcar-fosfato de cada filamento fica por fora da hélice, enquanto cada base nucleotídica está para dentro formando pares com o filamento oposto, adenina com timina, e guanina com citosina. Como resultado, a sequência dos dois filamentos de uma determinada dupla hélice apresentam uma relação de complementaridade e a sequência de qualquer fita de DNA define exatamente a sequência de sua fita complementar (Griffiths 2008; Watson et al. 2006).

O Dogma Central da Biologia, determinado por Crick (1970), é fundamental na explicação de como ocorre o fluxo de informações do código genético, contido no DNA. Resumidamente, as moléculas de DNA sofrem replicação, ou seja, o DNA é o molde para sua própria replicação, gerando cópias idênticas delas mesmas. No processo de transcrição, ocorre a síntese de RNA (ácido ribonucleico) a partir de um molde de DNA. No processo de tradução, o RNA mensageiro formado no processo de transcrição se associa com os ribossomos, onde ocorre a síntese de

proteínas.

2.2 Bioinformática

Após a descoberta de Watson e Crick, surgiu uma nova ciência, a Biologia Molecular, responsável pelo estudo da estrutura e função do código genético e das proteínas. Em seguida, surgiram métodos e máquinas de sequenciamento, com o objetivo de obter a sequência completa do DNA dos organismos, chamado de genoma. Dessa forma surgiu também a Bioinformática, que está em constante crescimento, principalmente, porque a enorme quantidade de dados biológicos produzidos hoje por meio desses métodos e sequenciadores já contabilizam milhares de genomas completos montados e publicados em bancos de dados online.

Segundo Gibas e Jambeck (2001), a Bioinformática utiliza a Tecnologia da Informação (TI) para auxiliar o gerenciamento de dados biológicos. A grande quantidade de dados biológicos produzidos exponencialmente a serem armazenados em bancos de dados, exige a necessidade de avanços em Hardware, Software e/ou técnicas de análise de dados cada vez mais eficientes. Com o desenvolvimento de computadores mais poderosos e mais baratos, e com o avanço dos projetos genômicos, podemos abordar problemas mais complexos, e com os bancos de dados públicos é possível armazenar, analisar e visualizar os dados biológicos.

De acordo com Prosdocimi (2002), a bioinformática é uma ciência que envolve diversas linhas de conhecimento - a engenharia de softwares, a matemática, a estatística, a ciência da computação e a biologia molecular, e tem como objetivo principal o estudo e a análise dos dados biológicos que são obtidos através das sequências DNA e proteínas.

Segundo Lesk (2008), a bioinformática é uma ciência aplicada, que surgiu do uso da capacidade de processamento e armazenamento de dados dos computadores para utilizar programas para fazer inferências a partir de dados biológicos obtidos em bancos de dados públicos ou locais, com o objetivo de fazer conexões entre eles e para derivar predições importantes e relevantes.

Resultado da grande quantidade de dados biológicos produzidos atualmente, que devem ser armazenados, além da grande necessidade de gerenciar, analisar e visualizar tais informações biológicas, o uso de banco de dados são de extrema importância no contexto da Bioinformática. Dessa forma, houve grande investimento na construção de bancos de dados públicos que alavancaram o sucesso dos projetos genomas, que armazenam e disponibilizam sequências de DNA de milhares de organismos para serem analisados.

Dentre os principais bancos de dados públicos está o INSDC (*INTERNATIONAL NUCLEOTIDE SEQUENCE DATABASE*) que disponibiliza uma grande quantidade de sequências e é a união de três bancos de dados parceiros que compartilham informações entre si, o GenBank (EUA), onde possui sequências de DNA de aproximadamente 279 mil organismos (GOLD,

2017), o DDBJ (DNA *Data Bank of Japan*), e EMBL (EMBL *NUCLEOTIDE SEQUENCE DATABASE*).

Vários Banco de Dados Públicos e Online, como o GenBank, DDBJ e EMBL, armazenam genes individuais, genomas completos, RNAs, anotações, sequências expressas, cDNAs e sequências sintéticas (Benson et al. 2013; Espindola 2010; Pevsner 2009).

Segundo Espindola et al. (2010), os dados biológicos gerados pelo sequenciamento dos genomas transformaram a biologia, que permitiu o avanço da bioinformática, expandindo em novas ciências ômicas, como a genômica, transcriptômica e a proteômica. A genômica tem como objetivo principal estudar a expressão e a interação dos genes. Já a transcriptômica tem como o objetivo isolar e caracterizar o RNA, e a proteômica de caracterizar as proteínas.

2.3 Genômica

A genômica é a ciência que permite entender o funcionamento do código genético através da análise dos genes e de suas funções. O sequenciamento, que permite obter ao final do processo o genoma completo dos organismos, e têm o objetivo de conhecer a informação genética contida na estrutura é realizado na genômica (Moreira 2015).

Um dos projetos genômicos que definitivamente impulsionaram o ramo, foi o projeto do sequenciamento do genoma humano, que buscava identificar todos os genes do DNA humano, determinar as sequências de base, armazenar as informações em banco de dados e desenvolver ferramentas para análise dos dados. O projeto aprimorou métodos e técnicas de análise para a bioinformática, ajudando em outros projetos genômicos, além de ter contribuído para diversas áreas da ciência e da saúde humana, que inclusive permitiu descobrir a causa de inúmeras doenças (Espindola 2010; Moreira 2015; Rodríguez-Ezpeleta 2011).

A grande explosão de dados biológicos sendo produzidos diariamente por meio do sequenciamento dos genomas de diversos organismos são armazenados em grandes bancos de dados públicos. O objetivo agora é analisar, interpretar e identificar os significados dessas sequências, ou seja, mudar o foco do DNA e RNA para as proteínas, estudar a expressão dos genes codificados pelo genoma desses organismos. Esta é a era pós-genômica (Prosdocimi 2002).

2.4 Análise de Dados Genômicos

2.4.1 Alinhamento de Sequências

São técnicas de comparação entre duas ou mais sequências biológicas, que buscam séries de caracteres individuais que se encontram na mesma ordem nas sequências analisadas. Existem vários programas online, como o BLAST e HMMER, que são capazes de alinhar centenas de sequências em poucos minutos. Geralmente, as moléculas consideradas por estes programas,

sejam elas formadas por nucleotídeos (DNA ou RNA) ou aminoácidos (proteínas), são polímeros representados por uma série de caracteres, e a comparação entre as moléculas depende apenas da comparação entre as respectivas letras. Apesar da aparente facilidade do processo, a análise de similaridade das sequências é uma tarefa complexa e uma etapa decisiva para grande parte dos métodos de bioinformática que fazem uso de sequências biológicas (Verli 2014).

Durante o alinhamento, o algoritmo utilizado deve buscar a melhor correspondência para as sequências, representadas por caracteres, permitindo a criação de espaços entre estes caracteres, chamados *gaps*, para que todas as sequências tenham o mesmo comprimento. Isto possibilita verificar a similaridade entre as sequências. O objetivo principal do algoritmo é minimizar as diferenças entre as sequências, buscando um alinhamento ótimo. A Figura 1 representa a qualidade de um alinhamento que é determinada pela soma dos pontos obtidos por cada caractere pareado, chamado de *match*, menos as penalidades pela introdução de *gaps* e caracteres não pareados, chamado de *mismatch* (Moreira 2015; Verli 2014).

Figura 1 – Alinhamento de duas sequências de proteínas.

Fonte: Prosdocimi (2002)

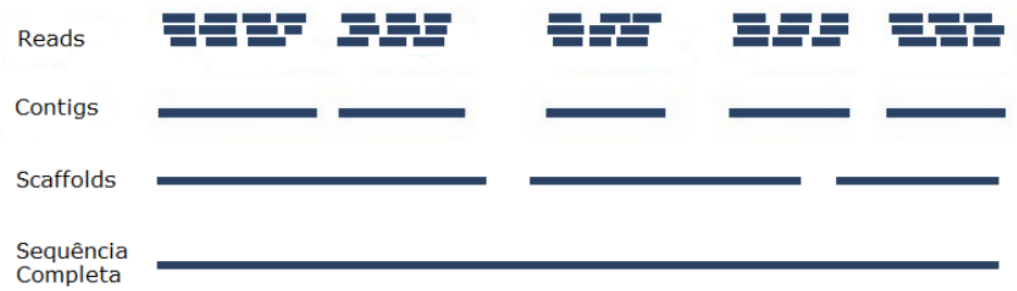
2.4.2 Sequenciamento e Montagem

Segundo Moreira (2015), o sequenciamento é uma técnica utilizada com o objetivo de obter e identificar uma sequência completa de DNA ou RNA, buscando conhecer a informação genética contida na estrutura.

Existem diferentes metodologias de sequenciamento, como o método de Sanger e o *Next-Generation Sequencing* (NGS), onde ambos utilizam a técnica de Shotgun, na qual o DNA a ser sequenciado é quebrado aleatoriamente em pequenos fragmentos, as *reads*, como podemos observar na Figura 2. O segundo método é capaz de obter uma maior quantidade de *reads* e de forma mais rápida, porém as *reads* geradas são menores que as obtidas pelo método de Sanger. Nos dois métodos, as *reads* passam por um controle de qualidade para que possam ser utilizadas na montagem do genoma (Rodríguez-Ezpeleta 2011; Verli 2014).

No processo de montagem do genoma, cada *read* obtida no processo de sequenciamento é alinhada entre si em busca de regiões de identidade ou sobreposição, para formar fragmentos contíguos (*contigs*), ou seja, *contigs* são a união de duas ou mais *reads*. Como mostrado na Figura 2, os contigs também irão se agrupar formando sequências mais contínuas, chamadas *scaffolds*.

Figura 2 – Sequenciamento de Shotgun.



Fonte: Adaptada de Adams (2008)

Para isso foram desenvolvidas várias ferramentas (*assemblers*) que permitem a montagem do genoma, como SPAdes, Velvet, Megahit, dentre outros (Moreira 2015; Rodríguez-Ezpeleta 2011).

2.4.3 Anotação Gênica

O processo de anotação de genomas tem como objetivo principal descobrir os genes e suas respectivas funções presentes na sequência de DNA obtida a partir dos processos de sequenciamento e montagem.

Segundo Verli (2014), existem mecanismos de delimitação (conjunto de protocolos e fluxos de trabalho) da sequência genômica, capaz identificar genes e predizer a sua função com base na similaridade com sequências conservadas.

Existe uma grande diferença na estrutura de genes procariotos e eucariotos. Genes procariotos codificantes de proteínas são colineares com seus produtos gênicos, dessa forma o mecanismo de delimitação será definido por um códon de início e um códon de término, chamado de *Open Reading Frame* (ORF). Já os genes eucariotos codificantes de proteína são caracterizados pela presença de sequências intervenientes ou íntrons. Algoritmos como o GenScan (Burge e Karlin 1998) realizam buscas por ORFs, identificadas por um códon de início e um códon de término, que correspondem a sequências possivelmente codificadoras.

De acordo com Moreira (2015), a anotação é feita através da comparação das ORFs ou dos transcritos com genes que estão depositados e armazenados em bancos de dados públicos, como o GenBank, utilizando alguma ferramenta para realizar a comparação entre as sequências, por exemplo, o uso do BLAST. A anotação pode ser automática, na qual requer o uso de programasc omo Glimmer (Delcher et al. 2007), GeneMark (Ter-Hovhannisyan et al. 2008), BLAST2GO (Götz 2008) e tRNAscan (Lowe 2016) que são capazes de anotar o genoma de um organismo de uma vez só.

A anotação também pode ser manual, que é mais demorada, por anotar cada gene

separadamente, porém ocorre uma curadoria, ou seja, o processo é feito de maneira mais cuidadosa tornando-a mais confiável.

2.5 Cianobactérias

As cianobactérias são um antigo grupo de bactérias que surgiram aproximadamente três bilhões de anos atrás, com tamanhos de genomas que variam de 1 a 10 Mb (Walter et al. 2017). São as principais fontes de oxigênio, nitrogênio e carbono da natureza, sendo a fotossíntese oxigenada responsável por criar a atmosfera da Terra rica em oxigênio, estimulando assim a evolução das espécies (Leao et al. 2017).

As cianobactérias são microrganismos procariotos fotossintéticos, ou seja, utilizam a luz solar como sua principal fonte de energia e estão presentes na grande maioria dos ecossistemas do nosso planeta. As cianobactérias ocupam diversos nichos ecológicos, muitas espécies vivem em água salgada e água doce, enquanto outras pertencem a ecossistemas terrestres (Kopf e Hess 2015). Dessa forma, as cianobactérias são consideradas organismos modelo para o estudo da fotossíntese oxigenada, do metabolismo do nitrogênio, da biossíntese de toxinas e da aclimatação salina (Wilde e Hihara 2016).

As cianobactérias desempenham papéis essenciais em quase todos os ambientes bióticos, pois tem a capacidade de sobreviver em diferentes condições ambientais de nichos ecológicos, desde a água do mar até os desertos. Devido à grande capacidade de converter fotossinteticamente o dióxido de carbono atmosférico para biomassa, as cianobactérias estão sendo exploradas para a conversão de energia solar e dióxido de carbono em biocombustíveis, como etanol, butanol, entre outros (Tiruveedula e Wangikar 2017).

As cianobactérias são alvo de diversas pesquisas, pois possuem um grande potencial em atividades terapêuticas e farmacêuticas (Galica 2017; Ramaswany 2006; Sivonen et al. 2010; Wang et al. 2011). As cianobactérias também são fonte de metabólitos secundários conhecidos como produtos naturais, na qual são fontes de agentes terapêuticos utilizados para tratar o câncer, infecções, inflamações e muitos outros estados patológicos (Leao 2017). Além disso, as cianobactérias atraem atenção para estudos com biofertilizantes, bioativos e conservantes de alimentos (Tiruveedula e Wangikar 2017; Sivonen et al. 2010; Wang et al. 2011).

2.6 Fatores de Transcrição

Em cada organismo vivo, mecanismos de desenvolvimento, morfológicos e fisiológicos, como aqueles que permitem aclimatação a mudanças ambientais, são resultados da modulação da expressão do genoma. Um nível desta modulação está relacionado à expressão gênica, em que os fatores de transcrição estão entre os principais atores (Heydarizadeh et al. 2014).

Um dos principais componentes das redes reguladoras de transcrição (*transcriptional regulatory networks* - TRNs) são os TFs. A anotação dos genomas de diversos microrganismos é fundamental para compreender os mecanismos moleculares de regulação transcricional em procariotos, comparação de conteúdo de genes e topologia de TRNs em espécies relacionadas e construção de modelos realistas de evolução das TRNs (Novichkov et al. 2010).

Os fatores de transcrição são caracterizados por um domínio de ligação ao DNA (DNA *Binding Domain* - DBD), um domínio de oligomerização, que permite a interação com outros TFs e outros reguladores de transcrição, e por um domínio de regulação de transcrição, na qual permite o controle da expressão gênica (Thiriet-Rupert et al. 2016). Os TFs compartilham um grau significativo de similaridade estrutural do DBD, podendo ser classificados em várias famílias com base na diferença da estrutura do DBD (Wu et al. 2007).

Os fatores de transcrição (TFs) representam uma grande proporção de todas as proteínas codificadas, portanto, estudos sobre TFs são de extrema importância e podem coletar informações sobre o mecanismo das redes reguladoras de transcrição (*transcriptional regulatory networks* - TRNs) (Wu et al. 2007), e compreender aspectos evolutivos dos organismos através de características específicas da linhagem (Charoensawan et al. 2010; Thiriet-Rupert et al. 2016).

2.7 Predição Gênica

A predição de genes é um passo importante para a anotação de novas sequências e genomas montados, e é responsável por analisá-las e buscar sequências de nucleotídeos correspondentes a cada um de seus genes ou de outras regiões de interesse (Mathé et al. 2002).

Existem diversos programas de bioinformática para essa finalidade, com diferentes técnicas e metodologias. Entretanto, o princípio básico da predição de genes consiste em fazer com que o programa reconheça nucleotídeos que são característicos de um determinado tipo de elemento gênico. Desta forma é possível identificar: regiões promotoras, junção dos éxons com os íntrons, os códons de início e parada da tradução e onde começam as regiões 5' e 3' UTR (regiões não traduzidas "*Untranslated Regions*") (Mathé et al. 2002; Moreira 2015).

2.7.1 Modelo Oculto de Markov

Um modelo oculto de Markov (*Hidden Markov Model* - HMM) é um modelo probabilístico. Para o propósito de encontrar genes, consiste em estados correspondentes para um significado biológico, por exemplo, íntrons, éxons, códons de início) e permite uma relação entre esses estados de forma biológica, como a junção de éxons com os íntrons, ou códons de início e de parada (Stanke 2003).

AUGUSTUS é um programa de predição gênica baseado em modelos ocultos de Markov Generalizado (*Generalized Hidden Markov Model* - GHMM), que define distribuições de pro-

babilidade para as várias seções de seqüências genômicas. Íntrons, éxons, regiões intergênicas, entre outras, correspondem a estados no modelo e cada estado é pensado para criar seqüências de DNA com certas probabilidades de emissão pré-definidas (Stanke 2005).

AUGUSTUS encontra a melhor análise de uma determinada sequência genômica, ou seja, uma segmentação das seqüências em estados que é mais provável de acordo com o modelo estatístico subjacente (Stanke 2005; Stanke 2006). Segundo Stanke (2005), os modelos probabilísticos são criados a partir das sequências onde se localizam os pontos de emenda (*splicing*), a sequência da região do ponto de derivação, as bases que antecedem o início da tradução, as regiões codificantes e não codificantes, as primeiras bases de codificação de um gene, éxons simples, éxons iniciais, éxons internos, éxons finais, regiões intergênicas, número de éxons por gene e a distribuição do comprimento dos íntrons.

2.8 Trabalhos Relacionados

Com o objetivo de facilitar o entendimento e levantar os diferentes tipos de abordagem sobre identificação de fatores de transcrição, foram selecionados três trabalhos da literatura para estudos, que apresentam propostas semelhantes aos objetivos deste trabalho. O primeiro trabalho, de Thiriet-Rupert et al. (2016) trata da identificação de fatores de transcrição em microalgas. O segundo trabalho, de Wu et al. (2007) apresenta o banco de dados cTFbase para identificação e análise de fatores de transcrição em cianobactérias. O terceiro trabalho, de Novichkov et al. (2010) apresenta o banco de dados RegPrecise para identificação e análise de fatores de transcrição reguladores preditos em procariotos.

Thiriet-Rupert et al. (2016) apresenta um trabalho sobre identificação e comparação de fatores de transcrição entre linhagens de microalgas, destacando a grande importância do estudo dos fatores de transcrição para entender a complexa história evolutiva das algas e suas características. Para tal, foi criado um *pipeline* utilizando os programas de computador BLAST, HMMER e InterProScan para realizar a identificação e classificação dos fatores de transcrição, passando por um processo de filtragem utilizando *Scripts* em PERL. Após identificados, os fatores de transcrição foram classificados automaticamente em famílias específicas de acordo com o DBD (*DNA Binding Domain*) com auxílio de um *Script* em PERL, seguindo uma série de regras de atribuição.

O trabalho de Wu et al. (2007) apresenta o banco de dados cTFbase para classificar e analisar os fatores de transcrição em genomas de cianobactérias, seguidos de análise comparativa do genoma. Para tal, foram utilizadas 21 genomas de cianobactérias completamente sequenciados encontrados no banco de dados IMG (*Integrated Microbial Genomes*). Logo depois, uma série de ferramentas de bioinformática, como HMMER, BLAST e Helixturnhelix, foram utilizados para identificar os fatores de transcrição, incluindo a pesquisa de fatores de transcrição candidatos e sua verificação. Após isso, todos os fatores de transcrição candidatos foram verificados manualmente

e os falsos positivos foram removidos com base na atribuição de domínio e similaridade de sequência em bancos de dados principais, como Swiss-prot e Refseq. Por fim, foram identificados 1288 fatores de transcrição de 21 genomas de cianobactérias.

O terceiro trabalho correlato, de Novichkov et al. (2010), apresenta o banco de dados RegPrecise desenvolvido para captura, visualização e análise de fatores de transcrição reguladores preditos em procariotos. Para tal, a parte principal do RegPrecise contém anotações de alta confiança obtidas por meio de análise genômica comparativa e curadoria manual. Já a segunda parte é composta pelos resultados da propagação precisa dos reguladores curados manualmente para novos genomas relacionados. Os dados são organizados por grupos taxonômicos, por fatores de transcrição, por famílias de fatores de transcrição e por caminho ou subsistema.

3 MATERIAIS E MÉTODOS

Esta seção aborda o processo de desenvolvimento de um *pipeline* para identificação de TFs de cianobactérias, que foi baseado no modelo proposto por Thiriet-Rupert et al. (2016) para microalgas.

A análise de *pipeline* é essencial para a identificação de TF em genoma completo. Como não existe um *pipeline* universal, cada estudo usa o seu próprio. No entanto, cada *pipeline* baseia-se nas mesmas ferramentas ou na combinação de várias para atingir o resultado esperado.

Neste trabalho, foram utilizadas as ferramentas: Prinseq, ferramenta para verificar, filtrar e gerar estatísticas resumidas das sequências; MAFFT (http://mafft.cbrc.jp/alignment/server/) um software de alinhamento múltiplo de sequências; HMMER versão 3.1 (http://hmmer.org/), um software que utiliza Modelo Oculto de Markov para realizar análise de sequências biológicas; AUGUSTUS (http://bioinf.uni-greifswald.de/webaugustus/index.gsp) um programa de predição gênica; e o banco de dados Pfam 31.0 (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam31.0/) na qual é uma grande coleção de famílias de proteínas, cada uma representada por alinhamentos de sequências múltiplas e modelos ocultos de Markov (HMMs).

Este *pipeline* utilizou três estratégias: (1) criação e predição de um banco de dados proteico com auxílio da ferramenta Augustus; (2) criação de modelos HMMER com as 29 TFs de cianobactérias disponíveis no cTFbase. Em seguida, é realizada a busca de cada modelo de TF no banco de dados proteico, também utilizando a ferramenta HMMER; (3) análise de domínios utilizando a ferramenta HMMER e o banco de dados proteico Pfam, para confirmar um fator de transcrição putativo.

O ambiente de teste utilizado foi um computador desktop, onde as características deste podem ser visualizadas no Quadro.

Quadro 1 – Ambiente de Execução.

Sistema Operacional	**Ubuntu 14.04 x64**
Processador	**Intel i5**
Memória	**8 GB**

Fonte: Elaborado pelo autor (2018)

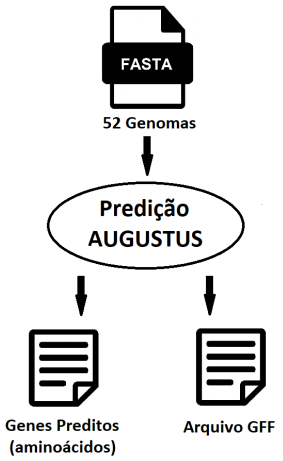
3.1 Construção de um Banco de Dados de Cianobactérias

Foram baixadas 154 dados genômicos de Cianobactérias disponíveis no NCBI, porém utilizamos somente as Cianobactérias que possuem genoma completo. Para isso, foi utilizado o Prinseq para realizar um filtro, obtendo 52 Cianobactérias com genoma completo (Apêndice A).

3.2 Pipeline

Todos os 52 genomas completos de cianobactérias foram concatenados em um arquivo .fasta (Apêndice B), com tamanho aproximado de 360 Mb. Em seguida, esse arquivo foi submetido ao Augustus, que por sua vez retorna alguns arquivos, são eles: "augustus.gff"e "augustus.aa", que são predição de genes em formato gff e predição de genes como sequências de proteína em formato fasta, respectivamente, como podemos visualizar na Figura 3.

Figura 3 – Predição de Genes.



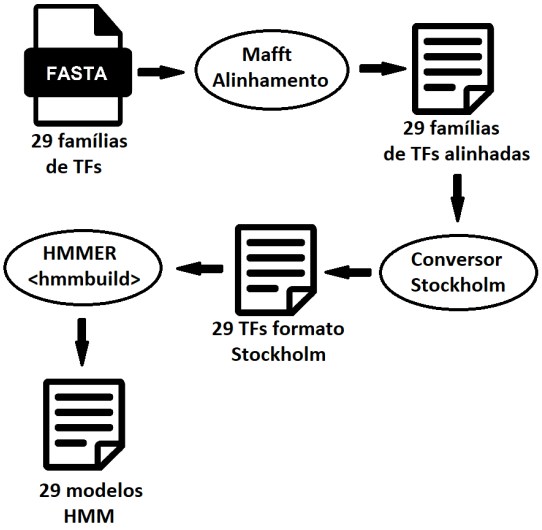
Fonte: Elaborada pelo autor (2018)

A Figura 4 representa a criação dos modelos HMM para cada uma das 29 famílias de TFs obtidas no cTFbase. Para isso, cada família passou por um tratamento, alinhamento com o Mafft e em seguida a conversão para o formato Stockholm. Por fim, utilizou-se o HMMER para criação de um modelo HMM para cada família de TF.

Após isso, utilizou-se novamente o HMMER para a identificação de sequências putativas. Para isso, usamos os 29 modelos HMM obtidos no passo anterior e os genes preditos que foram gerados pelo AUGUSTUS. Neste processo o HMMER a partir dos modelos HMM identifica padrões nas sequências preditas e nos retorna as sequências putativas, ou seja, as sequências que podem ser fatores de transcrição. Em seguida, um *Script* em Python foi desenvolvido para obter informações de cada fator de transcrição obtido, como reconhecer a qual cianobactéria cada sequência pertence utilizando o arquivo GFF gerado pelo AUGUSTUS. Por fim, o *Script* nos retorna o mapeamento das cianobactérias e fatores de transcrição com a quantidade obtida, como mostra na Figura 5.

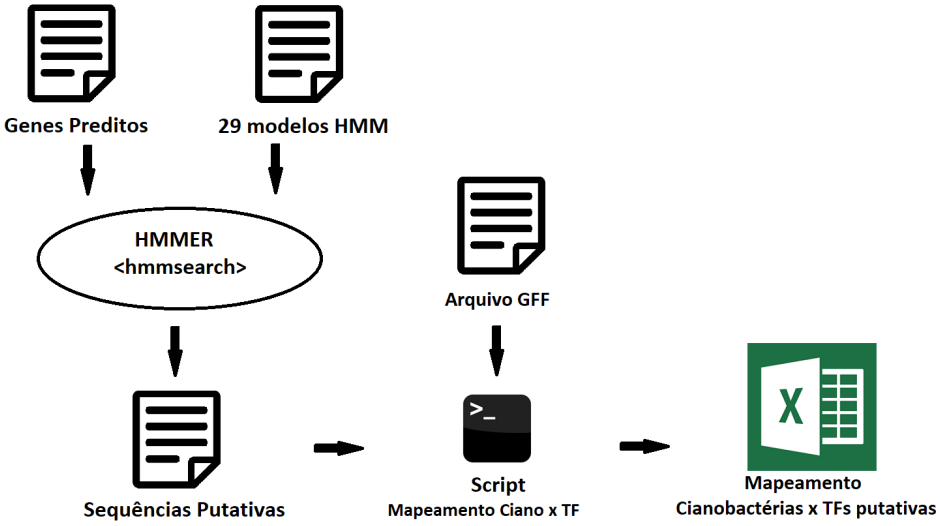
Em seguida, foi necessário realizar uma análise de domínios para validar as TFs putativas, ou seja, saber se uma TF putativa é de fato um TF ou não. Para isso, utilizou-se o bando de dados Pfam junto com os genes preditos pelo AUGUSTUS, na Figura 3. Os dois arquivos foram usados como entrada para rodar o HMMER e obter os perfis de domínios do cTFbase, como podemos ver na Figura 6. Os perfis de domínios contém informações dos domínios mais relevantes para

Figura 4 – Criação dos Modelos HMM.



Fonte: Elaborada pelo autor (2018)

Figura 5 – Identificação de Sequências Putativas.

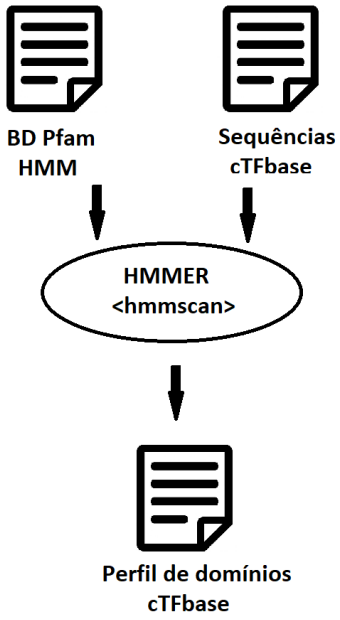


Fonte: Elaborada pelo autor (2018)

cada sequência. Esse passo foi realizado novamente, dessa vez utilizando o bando de dados Pfam com as sequências putativas obtidas no passo da Figura 5. Dessa forma, obteve-se os perfis de domínios de cada família de TF do cTFbase e um perfil de domínios das sequências putativas.

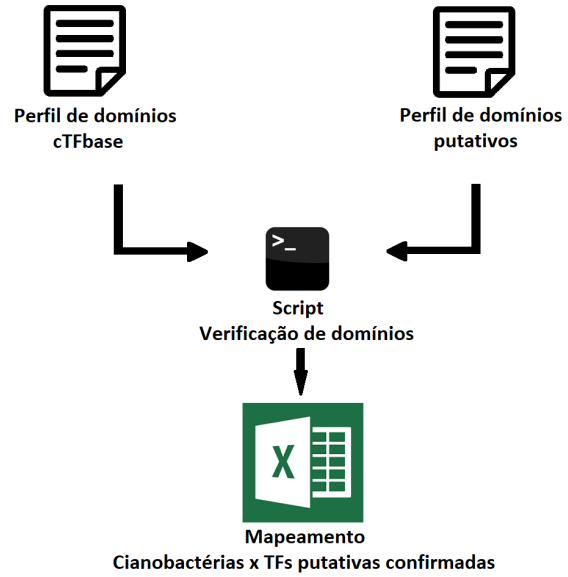
Por fim, criou-se um *Script* em Python que realiza uma análise dos perfis de domínios gerados no passo anterior. Ou seja, o *Script* verificou cada sequência putativa e seus respectivos domínios, caso esses domínios correspondessem aos domínios do cTFbase, a sequência em análise era validada como um fator de transcrição. Retornando o mapeamento de cianobactérias, fatores de transcrição e quantidade (Apêndice C), como mostra a Figura 7.

Figura 6 – Perfil de Domínios do cTFbase.



Fonte: Elaborada pelo autor (2018)

Figura 7 – Validação de TFs pela Análise de domínios.



Fonte: Elaborada pelo autor (2018)

4 RESULTADOS E DISCUSSÃO

O *pipeline* apresentou resultados significantes na identificação de famílias de fatores de transcrição em novas cianobactérias, e na identificação de novos fatores de transcrição em cianobactérias trabalhadas pelo cTFbase, contribuindo para uma análise mais detalhada dos fatores de transcrição de tais genomas.

No Quadro 2, podemos analisar a relação entre as 7 melhores cianobactérias, ou seja, as cianobactérias que obtiveram os melhores resultados, e quantidade de fatores de transcrição identificados ao lado da quantidade de fatores de transcrição confirmados pela análise de domínio. Também destacamos as famílias de TF *"GerE"* e *"OmpR"*, que apresentaram grandes quantidades de fatores de transcrição identificadas e confirmadas.

Quadro 2 – Mapeamento Cianobactérias e Fatores de Transcrição com a quantidade putativa e a quantidade de putativas confirmadas entre parênteses. (NC_009925.1 - *Acaryochloris marina MBIC11017*; NC_019738.1 - *Microcoleus PCC 7113*; NC_019695.1 - *Chroococcidiopsis thermalis PCC 7203*; NC_019771.1 - *Anabaena cylindrica PCC 7122*; NZ_CP019636.1 - *Nostocales cyanobacterium HT-58-2*; NC_019729.1 - *Oscillatoria nigro-viridis PCC 7112*; NC_019678.1 - *Rivularia PCC 7116*)

	Acaryochloris marina		*Microcoleus PCC 7113*		*Chroococcidiopsis thermalis*		*Anabaena cylindrica*		*Nostocales cyanobacterium*		*Oscillatoria nigro-viridis*		*Rivularia PCC 7116*	
	Put	Conf	Put	Conf	Put	Conf	Put	Conf	Put	Conf	Put	Conf	Put	Conf
copG	56	48	114	106	81	73	77	62	83	74	96	89	85	62
GerE	138	138	203	201	150	149	128	125	148	147	170	167	153	152
GntR	3	3	8	7	12	12	5	5	23	23	5	5	7	5
HTH3	28	22	34	30	21	17	28	23	26	23	38	33	26	24
LysR	13	13	8	8	6	6	5	5	10	10	7	7	17	17
OmpR	124	121	190	189	143	142	115	114	135	135	157	155	136	135

Fonte: Elaborado pelo autor (2018)

No trabalho de análise desenvolvido pelo cTFbase, dentre as cianobactérias utilizadas, 8 cianos apresentaram resultados importantes na identificação de novos fatores de transcrição. Esses resultados além de confirmar os fatores de transcrição descobertos pelo cTFbase, contribuímos para a identificação de novos fatores de transcrição, resultando também em uma análise mais detalhada dos fatores de transcrição dos 8 genomas anteriormente analisados. O Quadro 3 mostra a relação entre as 8 cianobactérias e a quantidade de fatores de transcrição anotados pelo cTFbase ao lado da quantidade de fatores de transcrição obtidos pelo nosso *pipeline*. Destacamos as cianobactérias *Anabaena Variabilis*, que é um organismo modelo para estudar os primórdios da vida multicelular, e *Nostoc Punctiforme* que possui capacidade para diferenciar estruturas celulares e acomodar caminhos metabólicos em conformidade, razão pela qual é uma das

bactérias mais versáteis e adaptativas estudadas atualmente.

Quadro 3 – Contribuição na Identificação de novos Fatores de Transcrição. (NC_007413.1 - *Anabaena variabilis ATCC 29413*; NC_005125.1 - *Gloeobacter violaceus PCC 7421*; NC_010628.1 - *Nostoc punctiforme PCC 73102*; NC_005042.1 - *Prochlorococcus marinus CCMP1375*; NC_006576.1 - *Synechococcus elongatus PCC 6301*; NC_009481.1 - *Synechococcus WH7803*; NC_004113.1 - *Thermosynechococcus elongatus*; NC_008312.1 - *Trichodesmium erythraeum*)

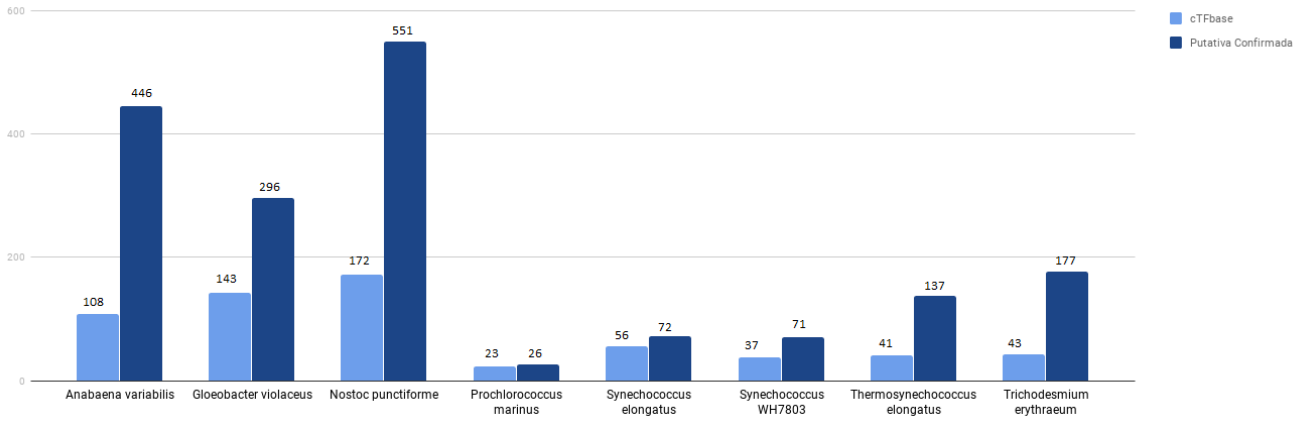
	Anabaena variabilis		*Gloeobacter violaceus*		*Nostoc punctiforme*		*Prochloroco-ccus marinus*		*Synechococ-cus elongatus*		*Synechococ-cus WH7803*		*Thermosynecho-coccus Elongatus*		*Trichodesmium Erythraeum*	
	Put	Conf	Put	Conf	Put	Conf	Put	Conf	Put	Conf	Put	Conf	Put	Conf	Put	Conf
arsR	8	16	6	17	7	13	1	0	3	10	0	2	2	4	1	0
copG	5	57	10	42	11	75	2	0	1	7	1	4	0	11	1	17
Crp	8	14	6	11	6	18	3	2	4	8	2	4	3	4	5	10
GerE	13	127	20	55	21	166	3	6	2	28	8	17	5	31	4	53
HTH3	14	21	8	15	26	24	0	1	9	12	2	2	1	16	4	13
MarR	2	18	7	25	3	19	0	2	1	9	1	6	0	9	0	10
OmpR	16	119	11	45	23	150	4	6	9	23	9	13	7	30	6	44

Fonte: Elaborado pelo autor (2018)

A fim de analisar a contribuição na identificação de novos fatores de transcrição para as 8 cianobactérias, fez-se a comparação da quantidade identificada pelo cTFbase contra a quantidade obtida pelo *pipeline* desenvolvido nestre trabalho, na Figura 8. Levando em consideração as cianobactérias *Anabaena Variabilis*, *Gloebacter Violaceus* e *Nostoc Punctiforme*, o cTFbase identificou 423 TFs, enquanto nosso *pipeline* identificou 1293 TFs, ou seja, 870 TFs a mais.

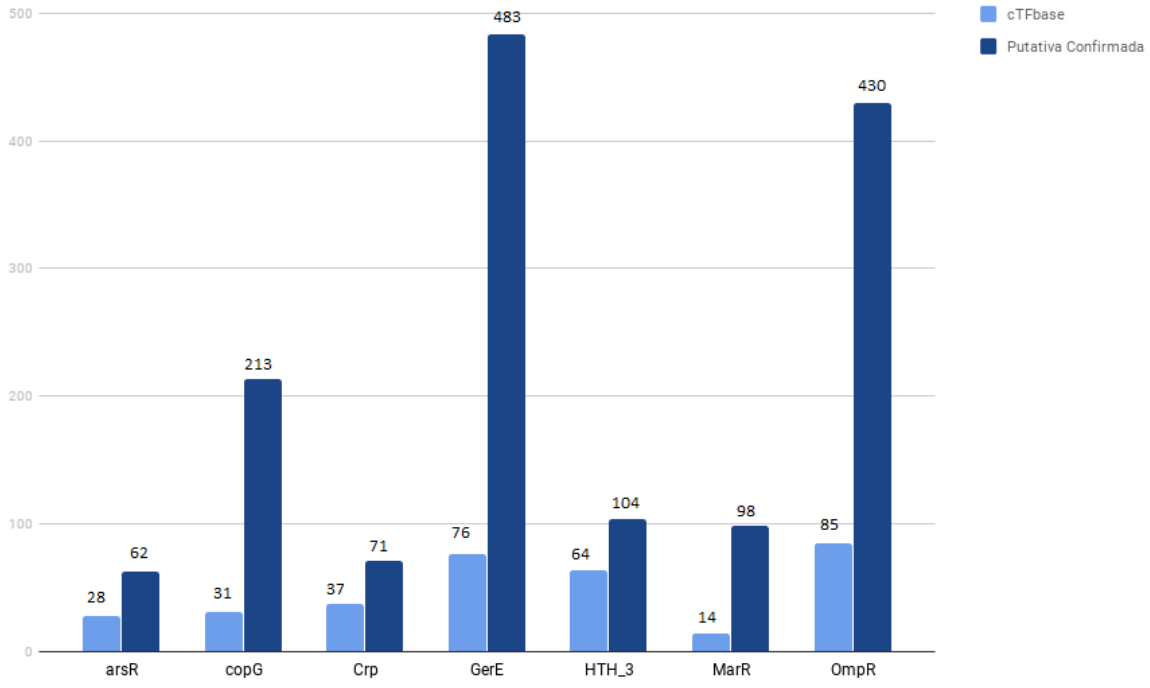
O Figura 9 representa as melhores contribuições por família de fator de transcrição, ou seja, as famílias que obtiveram os resultados com maior quantidade de fatores de transcrição nas 8 cianobactérias. Considerando as famílias de TF analisadas, o cTFbase identificou 335 TFs, e nosso *pipeline* obteve 1461 TFs, identificando 1126 TFs a mais que o cTFbase.

Figura 8 – Comparação da quantidade de TFs descritas no cTFbase e identificadas neste trabalho por Cianobactéria.



Fonte: Elaborado pelo autor (2018)

Figura 9 – Comparação da quantidade de TFs descritas no cTFbase e identificadas neste trabalho por família de TF.



Fonte: Elaborado pelo autor (2018)

5 CONCLUSÃO

As cianobactérias são um antigo grupo de bactérias gram negativas com forte variação do tamanho do genoma variando de 1,6 a 9,1 Mb e ainda pouco se sabe sobre seus fatores de transcrição. Os fatores de transcrição são um grupo de genes e é de grande importância estudá-los para a investigação da história evolutiva dos organismos e suas características.

Neste trabalho, várias técnicas foram aplicadas para implementar um pipeline automático e abrangente usando uma combinação de softwares como AUGUSTUS, HMMER e alguns *Scripts* a fim de descobrir novos fatores de transcrição, por meio da predição por genômica comparativa, em cianobactérias que estão publicadas em bancos de dados públicos, ajudando na análise evolutiva dos organismos e suas características.

Dessa forma, o pipeline desenvolvido apresentou resultados significantes na identificação de famílias de fatores de transcrição em novas cianobactérias, e na identificação de novos fatores de transcrição em cianobactérias trabalhadas pelo cTFbase, contribuindo para uma análise mais detalhada dos fatores de transcrição de tais genomas.

O que se pôde concluir quanto a utilização do *pipeline* proposto neste trabalho foi que:

- Identificou fatores de transcrição putativos e realiza uma análise de domínios para confirmar se o fator de transcrição putativo é de fato um fator de transcrição ou não.
- O esquema obteve resultados significativos na identificação de famílias de fatores de transcrição em novos genomas de cianobactérias, como mostrado na Seção 4.
- O esquema contribuiu confirmando e identificando novos fatores de transcrição em 8 cianobactérias também estudadas pelo cTFbase, colaborando para uma análise mais detalhadas dos fatores de transcrição de tais genomas.

5.1 Trabalhos Futuros

Como trabalhos futuros, pretende-se continuar o desenvolvimento e otimização do *pipeline* para que apresentem resultados mais detalhados e mais precisos. Também pretende-se aplicar o cálculo de acurária de Thiriet-Rupert et al. (2016) para avaliar a qualidade de predição do *pipeline* obtendo-se informações de verdadeiro-positivos e falso-positivos, ou seja, verificar a precisão de fatores de transcrição confirmados e fatores de transcrição putativos.

Para o desenvolvimento deste trabalho, foi considerado 29 famílias de fatores de transcrição obtidas no banco de dados cTFbase e um conjunto com 52 genomas completos de cianobactérias obtidos no NCBI. Acredita-se que é possível utilizar outras famílias de fatores de transcrição no *pipeline* para obter novas informações de fatores de transcrição em genomas de cianobactérias novos ou já estudados.

Além disso, é pretendido fazer uma análise *in vitro* dos fatores de transcrição putativos confirmados para validar e avaliar a precisão das técnicas utilizadas neste trabalho.

REFERÊNCIAS

ADAMS, J. Complex genomes: Shotgun sequencing. Nature Education, v. 1, n. 1, p. 186, 2008.

ARAVIND L, KOONIN EV. DNA-binding proteins and evolution of transcription regulation in the archaea. Nucleic Acids Res. 1999;27:4658–70.

BENSON, Dennis A. et al. GenBank. Nucleic acids research, v. 41, n. D1, p. D36-D42, 2013. Charoensawan V, Wilson D, Teichmann SA. Genomic repertoires of DNAbinding transcription factors across the tree of life. Nucleic Acids Res. 2010; 38:7364–77.

BURGE, C. B. and KARLIN, S. (1998) Finding the genes in genomic DNA. Curr. Opin. Struct. Biol. 8, 346-354.

CRICK, Francis. Central dogma of molecular biology. Nature, v. 227, n. 5258, p. 561-563, 1970.

DAHM, R. Friedrich Miescher and the Discovery of DNA. 2004.

DELCHER, A. L. et al. Identifying bacterial genes and endosymbiont DNA with Glimmer, Bioinformatics 23:6 (2007), 673-679.

ESPINDOLA, Foued Salmen et al. Bioinformatic resources applied on the omic sciences as genomic, transcriptomic, proteomic, interatomic and metabolomic. Bioscience Journal, v. 26, n. 3, 2010.

GALICA, Tomáš; HROUZEK, Pavel; MAREŠ, Jan. Genome mining reveals high incidence of putative lipopeptide biosynthesis NRPS/PKS clusters containing fatty acyl-AMP ligase genes in biofilm-forming cyanobacteria. Journal of Phycology, 2017.

GIBAS, C. JAMBECK, P. Desenvolvendo a Bioinformática. Campus. 2001.

GENOMES ONLINE DATABASE (GOLD), 2017. Disponível em: <https://gold.jgi.doe.gov/organisms>. Acesso em 2 de setembro de 2017.

GÖTZ, Stefan et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research, v. 36, n. 10, p. 3420-3435, 2008.

GRIFFITHS, A.; WESSLER, S.; LEWONTIN, R.; CARROLL, S. Introduction to Genetics, 9th ed. 2008.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. 3th ed. 2011.

HEYDARIZADEH, P et al. Functional investigations in diatoms need more than a transcriptomic approach. Diatom Res. 2014;29:75–89.

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.

KLUG, W.; CUMMINGS, M.; SPENCER, C.; PALLADINO, M. Concepts of Genetics, 2012.

KLUG, William S.; CUMMINGS, Michael R.; SPENCER, Charlotte A.; PALLADINO, Michael A.;. Conceitos de Genética. 9. ed. Brasil: Artmed, 2010. 896 p.

KOPF, Matthias; HESS, Wolfgang R. Regulatory RNAs in photosynthetic cyanobacteria. FEMS microbiology reviews, v. 39, n. 3, p. 301-315, 2015.

LANG D, et al. Genome-Wide Phylogenetic Comparative Analysis of Plant Transcriptional Regulation: A Timeline of Loss, Gain, Expansion, and Correlation with Complexity. Genome Biol Evol. 2010;2:488–503.

LEAO, Tiago et al. Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus Moorea. Proceedings of the National Academy of Sciences, v. 114, n. 12, p. 3198-3203, 2017.

LOWE, T.M. e CHAN, P.P. (2016) tRNAscan-SE On-line: Search and Contextual Analysis of Transfer RNA Genes. Nucl. Acids Res. 44: W54-57.

LUSCOMBE, Nicholas M. et al. What is bioinformatics? A proposed definition and overview of the field. Methods of information in medicine, v. 40, n. 4, p. 346-358, 2001.

MATHÉ, Catherine et al. Current methods of gene prediction, their strengths and weaknesses. Nucleic acids research, v. 30, n. 19, p. 4103-4117, 2002.

MENDEL, G. Experiments in Plant Hybridization, 1865.

MOREIRA, L (Org.). Ciências Genômicas: Fundamentos e Aplicações, 2015.

PEVSNER, J. Bioinformatics and Functional Genomics, 2nd ed. 2009.

PROSDOCIMI, Francisco et al. Bioinformática: manual do usuário. Biotecnologia Ciência E Desenvolvimento, v. 29, p. 12-25, 2002.

RAMASWAMY, Aishwarya V. et al. The secondary metabolites and biosynthetic gene clusters of marine cyanobacteria. Applications in biotechnology. Frontiers in marine biotechnology. Horizon Bioscience, p. 175-224, 2006.

RAYKO E, et al. Transcription fator families inferred from genome sequences of photosynthetic stramenopiles. New Phytol. 2010;188:52–66.

RODRÍGUEZEZPELETA, Naiara; HACKENBERG, Michael; ARANSAY, Ana M. (Ed.). Bioinformatics for high throughput sequencing. Springer Science and Business Media, 2011.

SIVONEN, Kaarina et al. Cyanobactins—ribosomal cyclic peptides produced by cyanobacteria. Applied microbiology and biotechnology, v. 86, n. 5, p. 1213-1225, 2010.

STANKE, Mario; WAACK, Stephan. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics, v. 19, p. ii215-ii225, 2003.

STANKE, Mario; MORGENSTERN, Burkhard. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic acids research, v. 33, p. W465-W467,

2005.

STANKE, Mario et al. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC bioinformatics, v. 7, n. 1, p. 62, 2006.

TER-HOVHANNISYAN, Vardges et al. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome research, v. 18, n. 12, p. 1979-1990, 2008.

TIRUVEEDULA, Gopi Siva Sai; WANGIKAR, Pramod P. Gene essentiality, conservation index and co-evolution of genes in cyanobacteria. PloS one, v. 12, n. 6, p. e0178565, 2017.

VERLI, H (Org). Bioinformática: da Biologia à Flexibilidade. 2014.

WALTER, Juline M. et al. Proposal of a new genome-based taxonomy for Cyanobacteria. PeerJ Preprints, 2017.

WANG, Hao; FEWER, David P.; SIVONEN, Kaarina. Genome Mining Demonstrates the Widespread Occurrence of Gene Clusters Encoding Bacteriocins in. 2011.

WATSON, J. The Double Helix: A Personal Account of the Discovery of the Structure of DNA. 1968.

WATSON, J.; BAKER, T.; BELL, S.; GANN, A.; LEVINE, M.; LOSICK, R.; Biologia Molecular do Gene, 5th ed. 2006.

WATSON, James D. et al. DNA Recombinante: Genes e genomas. 3. ed. [s. i.]: Artmed, 2009. 474 p. Tradução: Elio Hideo Babá et al.

WILDE, Annegret; HIHARA, Yukako. Transcriptional and posttranscriptional regulation of cyanobacterial photosynthesis. Biochimica et Biophysica Acta (BBA)-Bioenergetics, v. 1857, n. 3, p. 296-308, 2016.

Wu J, Zhao F, Wang S, Deng G, Wang J, Bai J, et al. cTFbase: a database for comparative genomics of transcription factors in cyanobacteria. BMC Genomics. 2007;8:104.

APÊNDICE

A Lista com 52 genomas completos de cianobactérias obtidos no NCBI

Número de Acesso	Cianobactéria
NC_009925.1	Acaryochloris marina MBIC11017
NC_019427.1	Anabaena sp. 90 chromosome chANA01
NC_019439.1	Anabaena sp. 90 chromosome chANA02
NC_019771.1	Anabaena cylindrica PCC 7122
NC_007413.1	Anabaena variabilis ATCC 29413
NZ_FO818640.1	Arthrospira sp. str. PCC 8005
NC_016640.1	Arthrospira platensis NIES-39
NC_019682.1	Calothrix sp. PCC 7507
NC_013771.1	Candidatus Atelocyanobacterium thalassa isolate ALOHA
NC_019697.1	Chamaesiphon minutus PCC 6605
NC_010175.1	Chloroflexus aurantiacus J-10-fl
NC_019695.1	Chroococcidiopsis thermalis PCC 7203
NC_019753.1	Crinalium epipsammum PCC 9333
NC_019776.1	Cyanobacterium aponinum PCC 10605
NZ_AP012549.1	Cyanobacterium endosymbiont of Epithemia turgida isolate EtSB Lake Yunoko
NC_019675.1	Cyanobium gracile PCC 6307
NC_014501.1	Cyanothece sp. PCC 7822
NC_019757.1	Cylindrospermum stagnale PCC 7417
NC_019780.1	Dactylococcopsis salina PCC 8305
NC_019703.1	Geitlerinema sp. PCC 7407
NZ_AP014815.1	Geminocystis sp. NIES-3708
NZ_CM001775.1	Geminocystis herdmanii PCC 6308
NC_022600.1	Gloeobacter kilaueensis JS1
NC_005125.1	Gloeobacter violaceus PCC 7421
NZ_CP017675.1	Candidatus Gloeomargarita lithophora strain D10
NC_019745.1	Gloeocapsa sp. PCC 7428

NC_019779.1	Halothece sp. PCC 7418
NC_019683.1	Leptolyngbya sp. PCC 7376
NC_019738.1	Microcoleus sp. PCC 7113
NC_010296.1	Microcystis aeruginosa NIES-843
NZ_CP017708.1	Moorea producens JHB sequence
NZ_CP007203.1	Nodularia spumigena CCY9414
NC_014248.1	Nostoc azollae 0708
NC_019676.1	Nostoc sp. PCC 7107
NZ_CP019636.1	Nostocales cyanobacterium HT-58-2
NZ_CP012036.1	Nostoc piscinale CENA21
NC_010628.1	Nostoc punctiforme PCC 73102
NC_019693.1	Oscillatoria acuminata PCC 6304
NZ_CM001633.1	Oscillatoriales cyanobacterium JSC-12
NC_019729.1	Oscillatoria nigro-viridis PCC 7112
NZ_CM002803.1	Planktothrix agardhii NIVA-CYA 126/8
NC_019689.1	Pleurocapsa sp. PCC 7327
NZ_CP007753.1	Prochlorococcus sp. MIT 0604
NC_005042.1	Prochlorococcus marinus subsp. marinus str. CCMP1375
NC_019701.1	Pseudanabaena sp. PCC 7367
NC_019678.1	Rivularia sp. PCC 7116
NC_019748.1	Stanieria cyanosphaera PCC 7437
NC_009481.1	Synechococcus WH7803
NC_006576.1	Synechococcus elongatus PCC 6301
NC_023033.1	Thermosynechococcus sp. NK55
NC_004113.1	Thermosynechococcus elongatus BP-1
NC_008312.1	Trichodesmium erythraeum IMS101

B Exemplo de Arquivo FASTA

```
>NC_009925.1 Acaryochloris marina MBIC11017, complete genomeAATAAATACTTACAGGTATTCCACCTGAAACTCTTTCTATGAATGA
CTTTCAAGTCTATATCCTATATTTATCCTCAATAAAATATGCACAATAGATCTCTACTGAGAAAACTTTATATTTTAGAAGCAATTCATCTCCCTTTTAAAATACA
AACTTGCATAGGATTGCCATAAATTAATGCACCTAAAATTAAGTCGTTTCTATGAAATTTATTTTAGGTGTGAAATTATATATTAGGTGTACAATTAGCCATACCA
TTCCGACAATATCGACATAGGAATATTGAGTAGAACAAGATGTACCATTATTAGCTTCTGGGTTGGCATAAGGGAATACAGGCAAAGACTTAAACCTATTCGGACT
CTGTAGAGCTGGCTCAATATTCAAAAGAATAATGTCCCTCGATACAAGCATCAGAATAGAGATCGCGCTCTTTACACCTATTAAATACATCATCTAAAACAACTGC
ATATTGTTCTATGCTCAAGCTGTAAGCATCAATATCTAGGCAAGCTCATTGACTAAAACAGTCTCTATCGAAATGTTCCAGCTCAATTCAAGGTCGAGATAGTACA
CCTATTAAATTAATTTCATACCTATAAATTCGAGATTCCCAGTAGTCTATTGGTGACTACATCAATCGGGCAACACCACCATGGCAAATATGCCGAGGGCAAGAAG
AACAAAGGGCACGGGAGTCATCGGGACGCGAGCACAATACGAAAACCGATGTTGTTGTTCGTGTTGTCGGGATTGTTGTGGTTGCGCGTGGCAGACCGACAGTTCC
TCGGATTGTTGTTCCAAGACCCGCCGTGTTTTGGATGTCTGTCCCGCTTTGACCATAAGGAGTTTTAGGGGTGGTTCGTGAATGTCAATGGCCTAAATTAACGATC
GCAAGCTGGATACGCAATCTCTGGATGTCGCTTAAATCAGGTGTCATCAAAGATTCGTCGTCTCAGTTGGTAGGTATCACCATGTTTAAGGTGGGCTGCCCAACTC
TGTAAGCACTGAATAAGATCTTGAGGGTCACATTGATAGGTTCCATAGGCTTGTTGCAGGTTTTTGATTCGTTGACGACCTCGGCGGAGGTTGTCGTTTCTGACTC
GGATGCGGTCATGCAGCACTCAGAAGCCGACGAAGTTGGCTCCTTGTTGAGTTTCAAATTGCTGACTTTTTACGGGATGGATCTTGAGGCGTAAGGTCGCTAAGTA
GTTTTCAATGGAGGTGCGGCAGTGTTGGAGATATTGCCATTCATCTGAGAATAAGGCGAGGTCATCTACGTAGCGTAAATACTTTGAACAACGCAGGTTCTCTTTA
ACGTAATGATCAAAGCTATTGAGATAGACGTTGGCAAAGAATTGGCTGGTTAGGTTACCGATGAGCAACCCTCGTTTGCGAGTCAGTGGGGTTAGCAGGGTGCCTC
CAGGAAAATATTCAATGGAAGGGTCTTGTTCATTGCTGTTGTCGATAATTTTATCGATCAACCAAAGGGTTGAGGGACAACGAATCTTCTTGTGCAGGATATTTTT
GAGGATGTCTAGATCGATGGAGGGGAAGTATTTGCGGATATCGCACTGCAAGATATATCGGCTGCTACACAGCCCTACTTTTCAAAAGCTTTAATCTCAAACCTCG
AAGTTAAATCCCTGTTGTTTCAATAGCTCGTATTTGCTTTGGTCGAACTGATAGAGTCTGGCAGCACGATGAGCAACTCCAGTTTGTTGCTCATCAAGGGGAATTA
AGAGATCCATTTTTAGCAATTTCTTACGAAAGTTCCGTTTATCTAGAGGATGGCCAAGAACTGTTTCGTAAAGCTGCTGAATTTGGGTGAGTGTAAATTTTTGCGG
CAGTAGTTCGAAGCCAATCGGTTCGTAGCGAACTTTGCCCTGCAATCGCCGGATGGCCTGTTTCAAGATTTGAGCATGATCAAAAGCTAGAGATGGCAAATTCTCA
ATGGAGTACCAGGCTGCGTCTTCAGCATCAGTAGAGGCTTGTAAAGGATATTCGATGAGGTTGATTAGAGCGTAGTAGGCAACGGAGATGATGCGATCGCGCGGAT
CTCGACCCAAATCGCCAAAGGTATAGAGCTGCTCTAGGAAAATACCCTGAACCCCCGTTTCTTCTCGCAGCTCCCGGCGAGCCGCGTCTTCTAAAGATTCATCCAT
CTGCACAAAGCCTCCGGGTAATGCCCACTGATGCTGGAAAGGGGGAATCTGGCGCTGAATCAGTAGGACTTTGAGATCGATCTGTTCGTCTAGGCCAAACACCACG
CAGTCAACGGTAAGACCAGGGCGCGGATAATCATAGGTATAGGGCATAGGATAGGGGCAATCTTGACATCGTGTAAAAACGACACTATGATGGTTTTATTCAAGTT
AGTGTCGTTTTTACACTATAGAGCAATGCAACAGCATTTTGAAGTCGCGGCCGGGTCTATTTTGGGCCAGAACCATCGACGAGTGAGTCAGAATAATCAAGATGCT
TGGGCAAGTCTGACGCTAGGTCAAGCAACCATTGCGGTGGTATGCGATGGGTGCGGAAGTCGACCTAAAAGTGAAGTTGGCGCGCAGCTTGGGGCTCAGTTAACGG
TGCATGCTCTTGCCCAACTCGTAGCAGAGGAGCAACCGGAAGACGAGCTGTTTTGGCAGATGTTGAGGAAGCACCTATTAGGAGGGATGCAGACTTTACAGGAGCG
GCTAGGAGCAGATTCGCAGAGGCTGGTGGATTACTGGTTATTCACTCTGATGGGGTGTTTAGTGACCCCAGAAACAACCTGGATTTTTGGCCTAGGGGATGGGGTA
TTTGCCATCAATGAGGAGTTCCATCAGATTGGCCCATTTACTAACAATGCCCCGCCCTATCTGGCCTACAGCTTGCTATGGAACACCGAGAAGGGCCTCGATCCGT
TTCAACTCCAGATTTATGCTCAACACCCCACAGAACAGGTGCAGTCTGTGCTGATTGGGAGCGATGGAGTGGGGGATTTAGTTAATGTAGCAGAAGATTATCTCCC
TGGGAAGTCAGAGCTGGTAGGACCACTGAATCAGTTTTGGTGTTGCGATCGCTATTTCCAGAACCCTGATCAGGTGCGACGACGGCTCGCTCAAATTAATCGAGAG
```

C Mapeamento completo de cianobactérias e fatores de transcrição com a quantidade putativa e a quantidade de putativas confirmadas

	>NC_009925.1	>NC_019427.1	>NC_019439.1	>NC_019771.1	>NC_007413.1	>NZ_FO818640.1	>NC_016640.1	>NC_019682.1	>NC_013771.1	>NC_019697.1	>NC_010175.1	>NC_019695.1	>NC_019753.1
AraC	22	4	1	13	10	6	6	13	0	14	10	33	10
AraC_CONFIRMADO	18	1	0	7	4	1	2	11	0	11	4	27	1
arsR	19	3	0	11	18	9	15	13	2	19	23	18	10
arsR_CONFIRMADO	18	3	0	10	16	5	5	11	2	15	22	16	9
Bac_DNA	1	1	1	2	2	3	3	1	1	1	1	1	1
Bac_DNA_CONFIRMADO	1	1	1	2	1	3	3	1	1	1	1	1	1
bolA	1	0	1	1	1	2	1	1	1	1	0	1	1
bolA_CONFIRMADO	1	0	1	1	1	1	1	1	1	1	0	1	1
copG	56	36	6	77	64	52	50	74	6	36	33	81	82
copG_CONFIRMADO	48	18	4	62	57	41	42	63	5	26	28	73	74
Crp	12	5	???	11	14	9	8	17	2	6	10	18	8
Crp_CONFIRMADO	12	5		10	14	9	8	17	2	6	10	17	8
DnaA	4	1	???	1	1	2	???	1	???	2	2	1	1
DnaA_CONFIRMADO	1	1		1	1	1		1			1	1	1
DUF24	5	1	???	4	4	2	2	1	1	5	5	6	3
DUF24_CONFIRMADO	4			4	4	2	2	1		3	1	4	2
DUF387	1	???	1	1	1	1	1	1	2	1	1	1	1
DUF387_CONFIRMADO	1		1	1	1	1	1	1	1	1	1		1
FUR	10	3	1	4	4	3	3	5	???	8	3	5	3
FUR_CONFIRMADO	9	3	1	4	3	3	3	4		5	1	4	3
GerE	138	41	9	128	127	104	104	117	7	122	98	150	127
GerE_CONFIRMADO	138	40	9	125	127	101	102	116	6	119	97	149	126
GntR	3	3	???	5	4	5	5	9	1	7	9	12	3
GntR_CONFIRMADO	3	3		5	4	5	5	9	1	7	9	12	3
HetR	1	1	???	1	1	1	1	2	???	???	???	???	1
HetR_CONFIRMADO		1		1	1	1	1	1					1
HrcA	1	1	???	2	1	1	1	1	1	2	2	1	1
HrcA_CONFIRMADO	1	1		1	1	1	1	1	1	1	2	1	1
HTH_3	28	17	???	28	26	14	9	29	1	22	17	21	18
HTH_3_CONFIRMADO	22	15		23	21	14	9	20	1	19	15	17	17

	>NC_019776.1	>NZ_AP012549.1	>NC_019675.1	>NC_014501.1	>NC_019757.1	>NC_019780.1	>NC_019703.1	>NZ_AP014815.1	>NZ_CM001775.1	>NC_022600.1	>NC_005125.1	>NZ_CP017675.1	>NC_019745.1
AraC	10	1	3	6	4	0	1	5	6	8	10	2	48
AraC_CONFIRMADO	9	0	3	3	2	0	0	2	4	8	9	1	42
arsR	10	3	8	8	14	9	10	9	7	17	19	7	22
arsR_CONFIRMADO	9	3	7	7	11	9	10	7	7	14	17	5	20
Bac_DNA	1	1	1	1	1	1	1	2	1	3	4	2	0
Bac_DNA_CONFIRMADO	1	1	1	1	1	1	1	1	1	3	4	2	0
bolA	1	1	1	2	1	1	1	1	1	1	1	1	1
bolA_CONFIRMADO	1	1	1	1	1	1	1	1	1	1	1	1	1
copG	32	7	10	77	64	22	54	37	38	42	50	15	64
copG_CONFIRMADO	25	5	7	61	49	18	49	27	30	37	42	13	61
Crp	4	3	12	14	14	6	10	8	10	6	12	5	11
Crp_CONFIRMADO	4	3	12	14	14	6	10	8	10	6	11	5	11
DnaA	???	???	1	1	2	1	1	1	???	1	1	1	1
DnaA_CONFIRMADO			1	1	1	1	1			1	1	1	1
DUF24	1	1	4	2	3	1	4	2	1	4	8	2	6
DUF24_CONFIRMADO			2	2	2	1	4	1		1	4	2	4
DUF387	1	???	1	2	1	1	1	1	1	1	3	???	1
DUF387_CONFIRMADO	1		1	1	1		1	1	1	1	1		1
FUR	5	1	3	4	5	2	2	6	8	5	8	2	4
FUR_CONFIRMADO	5	1	3	3	3	2	1	5	6	4	4	2	4
GerE	76	15	21	129	113	27	97	80	75	67	55	47	105
GerE_CONFIRMADO	74	14	21	122	110	26	95	80	73	66	55	47	105
GntR	2	2	2	4	7	5	6	1	3	3	4	4	11
GntR_CONFIRMADO	2	2	2	4	6	5	5	1	2	3	4	4	11
HetR	???	???	???	???	1	???	1	???	???	???	???	???	???
HetR_CONFIRMADO					1		1						
HrcA	1	1	1	1	1	1	1	1	1	1	1	1	1
HrcA_CONFIRMADO	1	1	1	1	1	1	1	1	1	1	1	1	1
HTH_3	7	5	5	32	22	18	7	12	12	20	17	7	16
HTH_3_CONFIRMADO	7	5	1	28	22	15	6	8	11	15	15	5	15

	>NC_019779.1	>NC_019683.1	>NC_019738.1	>NC_010296.1	>NZ_CP017708.1	>NZ_CP007203.1	>NC_019676.1	>NZ_CP019636.1	>NZ_CP012036.1	>NC_010628.1	>NC_019693.1	>NZ_CM001633.1	>NC_019729.1
AraC	3	11	12	5	10	4	13	19	14	13	8	13	9
AraC_CONFIRMADO	1	5	4	3	4	3	8	14	8	10	1	7	4
arsR	8	7	13	10	18	11	18	42	17	14	16	19	9
arsR_CONFIRMADO	7	6	13	9	15	10	18	39	17	13	13	17	9
Bac_DNA	1	1	1	1	1	1	1	1	1	2	3	1	1
Bac_DNA_CONFIRMADO	1	1	1	1	1	1	1	1	1	2	2	1	1
bolA	1	1	1	1	2	1	1	2	0	1	1	1	1
bolA_CONFIRMADO	1	1	1	1	1	1	1	1	0	1	1	1	1
copG	38	44	114	40	74	42	79	83	94	88	114	53	96
copG_CONFIRMADO	27	36	106	15	54	26	64	74	79	75	97	50	89
Crp	8	8	17	8	16	8	9	18	14	18	10	10	13
Crp_CONFIRMADO	8	8	17	8	16	8	9	17	14	18	10	10	12
DnaA	1	1	1	1	1	1	1	1	1	1	1	1	3
DnaA_CONFIRMADO	1	1	1	1	1	1	1	1	1	1	1	1	1
DUF24	2	5	7	1	3	???	7	6	3	4	3	5	7
DUF24_CONFIRMADO		4	1	1	2		3	3	1	2	2	2	4
DUF387	1	1	1	1	1	1	10	2	1	1	1	1	1
DUF387_CONFIRMADO	1	1	1	1	1	1	1	1	1	1	1	1	1
FUR	3	2	5	3	5	5	4	4	4	6	4	3	6
FUR_CONFIRMADO	2	2	4	2	4	4	4	4	4	4	4	3	3
GerE	72	96	203	37	133	67	136	148	165	167	156	116	170
GerE_CONFIRMADO	68	95	201	36	131	67	133	147	161	166	154	116	167
GntR	4	2	8	4	5	7	4	23	5	8	6	5	5
GntR_CONFIRMADO	3	2	7	4	5	5	4	23	5	8	5	5	5
HetR	???	1	1	???	1	1	2	1	2	3	2	1	1
HetR_CONFIRMADO		1	1		1	1	1	1	1	2	1	1	1
HrcA	1	1	1	3	2	1	1	1	1	1	2	1	1
HrcA_CONFIRMADO	1	1	1	1	1	1	1	1	1	1	1	1	1
HTH_3	15	11	34	27	24	20	39	26	25	28	29	8	38
HTH_3_CONFIRMADO	12	9	30	25	19	18	35	23	23	24	29	6	33

	>NZ_CM002803.1	>NC_019689.1	>NZ_CP007753.1	>NC_005042.1	>NC_019701.1	>NC_019678.1	>NC_019748.1	>NC_009481.1	>NC_006576.1	>NC_023033.1	>NC_004113.1	>NC_008312.1	>NC_014248.1
AraC	6	9	1	0	7	25	9	0	0	2	1	1	2
AraC_CONFIRMADO	0	3	0	0	2	18	4	0	0	2	1	0	0
arsR	7	14	0	0	12	18	11	3	11	6	4	3	5
arsR_CONFIRMADO	6	13	0	0	11	17	8	2	10	6	4	0	2
Bac_DNA	2	0	1	0	2	1	1	1	1	1	1	1	1
Bac_DNA_CONFIRMADO	2	0	1	0	2	1	1	1	1	1	1	1	0
bolA	1	1	1	0	2	1	1	1	1	1	1	1	1
bolA_CONFIRMADO	1	1	1	0	2	0	1	1	1	1	1	1	0
copG	39	37	4	1	33	85	50	4	10	15	12	44	35
copG_CONFIRMADO	32	30	3		28	62	41	4	7	13	11	17	26
Crp	10	12	2	2	10	14	11	4	8	5	4	31	9
Crp_CONFIRMADO	10	12	2	2	10	13	10	4	8	5	4	10	7
DnaA	???	1	1	1	2	1	1	1	1	1	2	1	2
DnaA_CONFIRMADO		1	1	1	1	1	1	1	1	1	1	1	1
DUF24	2	4	???	???	2	4	5	???	4	1	1	2	???
DUF24_CONFIRMADO	1	2			1	3	5		1	1	1	1	
DUF387	1	1	???	1	1	1	1	1	1	???	???	1	1
DUF387_CONFIRMADO	1	1		1	1	1	1	1	1			1	
FUR	4	7	???	1	4	5	4	3	6	3	5	1	3
FUR_CONFIRMADO	4	7		1	3	4	3	3	5	3	4	1	1
GerE	68	116	6	7	85	153	124	19	28	40	33	54	74
GerE_CONFIRMADO	67	114	6	6	85	152	122	17	28	39	31	53	70
GntR	4	6	1	2	4	7	5	3	3	3	4	3	6
GntR_CONFIRMADO	4	6		2	4	5	5	3	3	3	3	3	2
HetR	1	???	???	???	???	1	???	???	???	???	???	2	1
HetR_CONFIRMADO	1					1						1	
HrcA	1	1	???	???	1	2	1	1	1	1	1	1	1
HrcA_CONFIRMADO	1	1			1	1	1	1	1	1	1	1	
HTH_3	11	13	1	1	12	26	18	2	12	5	16	13	8
HTH_3_CONFIRMADO	10	11		1	10	24	16	2	12	5	16	13	6

	>NC_009925.1	>NC_019427.1	>NC_019439.1	>NC_019771.1	>NC_007413.1	>NZ_FO818640.1	>NC_016640.1	>NC_019682.1	>NC_013771.1	>NC_019697.1	>NC_010175.1	>NC_019695.1	>NC_019753.1
HTH_11	4	2	1	3	3	21	6	6	???	7	10	6	3
HTH_11_CONFIRMADO	3	2	1	3	3	20	5	5		7	8	6	2
LexA	6	1	???	1	1	???	???	1	1	1	1	1	1
LexA_CONFIRMADO	1	1		1	1			1	1	1	1	1	1
LysR	13	6	???	5	10	6	7	8	1	12	6	6	4
LysR_CONFIRMADO	13	5		5	7	6	6	7	1	12	5	6	4
MarR	18	6	1	21	32	14	17	20	2	21	31	27	14
MarR_CONFIRMADO	13	4	1	15	18	9	9	14	2	15	29	25	13
MerR	10	3	1	5	10	14	9	13	2	17	5	3	13
MerR_CONFIRMADO	6			2	3	7	2	6	1	14	3	2	8
NifT	???	1	???	1	1	???	1	1	???	???	???	???	???
NifT_CONFIRMADO		1		1	1			1					
OmpR	124	34	7	115	119	92	94	102	5	103	93	143	115
OmpR_CONFIRMADO	121	34	7	114	119	92	93	102	5	101	88	142	114
PadR	8	2	1	4	5	3	3	5	???	5	4	6	4
PadR_CONFIRMADO	8	1	1	3	5	3	3	5		5	2	5	4
PatA	119	32	6	108	115	89	92	101	4	102	86	138	112
PatA_CONFIRMADO	17	3		14	8	10	12	7		3	11	14	6
ROK	1	3	???	3	4	3	3	3	1	1	3	4	3
ROK_CONFIRMADO	1	2		3	3	2	2	3	1	1	3	3	2
Rrf2	4	2	???	3	4	1	1	4	???	2	5	5	2
Rrf2_CONFIRMADO	4	2		3	4	1	1	4		2	5	5	2
SfsA	1	1	1	1	1	1	1	1	???	1	1	1	1
SfsA_CONFIRMADO	1		1	1	1	1	1	1		1	1	1	1
tetR	22	2	???	11	11	3	4	14	???	23	11	16	3
tetR_CONFIRMADO	22	2		8	10	2	3	12		20	7	16	2
unclassified	6	4	???	4	9	3	4	6	2	6	15	5	6
unclassified_CONFIRMAD	5	4		3	6	2	3	6	2	5	13	4	4

	>NC_019776.1	>NZ_AP012549.1	>NC_019675.1	>NC_014501.1	>NC_019757.1	>NC_019780.1	>NC_019703.1	>NZ_AP014815.1	>NZ_CM001775.1	>NC_022600.1	>NC_005125.1	>NZ_CP017675.1	>NC_019745.1
HTH_11	3	3	2	5	3	3	3	3	2	4	4	4	6
HTH_11_CONFIRMADO	3	3	1	5	3	3	3	3	2	3	4	3	6
LexA	1	1	2	1	1	???	1	1	2	3	???	???	1
LexA_CONFIRMADO	1	1	1	1	1		1	1	1				1
LysR	5	4	3	7	4	4	5	7	3	9	11	5	8
LysR_CONFIRMADO	5	3	2	6	4	4	4	5	3	8	10	5	8
MarR	11	3	6	20	23	11	18	10	7	22	29	7	25
MarR_CONFIRMADO	11	3	6	18	17	9	14	9	7	19	25	5	20
MerR	4	???	6	8	12	9	7	8	13	6	6	6	11
MerR_CONFIRMADO	2		4	2	5	5	3	4	9	4	4	1	7
NifT	???	1	???	1	1	???	???	1	1	???	???	???	1
NifT_CONFIRMADO		1		1	1								1
OmpR	72	12	19	116	99	23	87	75	69	57	47	42	97
OmpR_CONFIRMADO	70	11	18	114	99	23	86	74	68	54	45	42	97
PadR	2	1	2	4	3	6	1	3	4	7	12	4	7
PadR_CONFIRMADO		1	1	4	3	4	1	3	2	5	10	4	7
PatA	69	11	13	108	97	22	84	75	68	49	41	40	95
PatA_CONFIRMADO	17			17	7	1	8	11	11	3	1	5	6
ROK	2	1	1	1	3	2	1	1	2	3	1	2	4
ROK_CONFIRMADO	1	1	1	1	3	1	1	1	1	3	1	2	1
Rrf2	2	2	2	3	4	3	4	1	1	3	3	2	3
Rrf2_CONFIRMADO	2	1	2	3	4	3	4	1	1	3	2	2	2
SfsA	1	1	1	1	1	???	1	1	1	1	1	1	2
SfsA_CONFIRMADO	1	1	1	1	1		1	1	1	1	1	1	1
tetR	2	1	4	11	5	2	4	2	3	21	15	5	18
tetR_CONFIRMADO	2		4	8	5	2	4	2	2	21	15	5	17
unclassified	4	2	3	4	6	4	6	5	3	12	11	3	7
unclassified_CONFIRMAD	3	2	1	4	4	4	4	5	3	8	10	3	6

	>NC_019779.1	>NC_019683.1	>NC_019738.1	>NC_010296.1	>NZ_CP017708.1	>NZ_CP007203.1	>NC_019676.1	>NZ_CP019636.1	>NZ_CP012036.1	>NC_010628.1	>NC_019693.1	>NZ_CM001633.1	>NC_019729.1
HTH_11	3	4	6	3	8	5	6	7	7	4	3	6	3
HTH_11_CONFIRMADO	3	4	6	3	6	4	6	7	7	4	3	6	3
LexA	???	2	1	1	3	1	1	1	1	2	???	1	2
LexA_CONFIRMADO		1	1	1	1	1	1	1	1	2		1	
LysR	4	9	8	7	5	5	5	10	5	7	4	7	7
LysR_CONFIRMADO	4	9	8	7	5	5	5	10	5	7	4	7	7
MarR	11	17	26	15	14	9	20	50	28	25	16	21	18
MarR_CONFIRMADO	10	13	21	13	11	7	16	45	21	19	14	19	15
MerR	14	4	14	11	14	16	8	5	17	7	13	7	12
MerR_CONFIRMADO	8	1	6	2	10	12	3	3	4	2	3	2	6
NifT	1	???	1	???	???	1	1	1	1	1	???	1	???
NifT_CONFIRMADO	1		1			1	1	1	1	1		1	
OmpR	68	86	190	30	114	59	122	135	149	151	150	103	157
OmpR_CONFIRMADO	65	86	189	29	113	59	122	135	149	150	150	102	155
PadR	4	3	5	3	5	3	4	8	7	5	4	5	4
PadR_CONFIRMADO	3	3	5	2	3	3	4	6	6	4	4	5	3
PatA	66	85	187	27	108	56	121	134	146	146	150	97	148
PatA_CONFIRMADO	6	12	9	1	8	6	8	10	6	9	10	5	11
ROK	2	1	2	1	3	3	3	3	3	3	4	2	2
ROK_CONFIRMADO	1	1	2	1	2	3	3	3	3	3	2	2	2
Rrf2	3	1	2	2	1	2	4	5	3	4	2	1	6
Rrf2_CONFIRMADO	2	1	1	1	1	2	3	5	3	4	2	1	3
SfsA	1	1	1	1	1	1	1	1	1	1	1	1	1
SfsA_CONFIRMADO	1	1	1	1	1	1	1	1	1	1	1	1	1
tetR	6	10	9	1	8	5	13	17	15	11	5	5	3
tetR_CONFIRMADO	4	10	6	1	7	5	12	16	14	11	4	4	3
unclassified	4	3	7	3	4	6	5	8	8	9	7	6	7
unclassified_CONFIRMAD	4	3	5	3	4	3	4	7	5	6	4	6	4

	>NZ_CM002803.1	>NC_019689.1	>NZ_CP007753.1	>NC_005042.1	>NC_019701.1	>NC_019678.1	>NC_019748.1	>NC_009481.1	>NC_006576.1	>NC_023033.1	>NC_004113.1	>NC_008312.1	>NC_014248.1
HTH_11	3	6	???	???	2	5	7	2	3	2	2	1	3
HTH_11_CONFIRMADO	2	6			2	5	7	2	3	2	2	1	1
LexA	1	1	1	1	1	3	1	2	1	???	???	???	1
LexA_CONFIRMADO		1	1	1		1	1	1					
LysR	6	6	1	1	4	17	6	1	5	3	3	3	3
LysR_CONFIRMADO	6	6	1	1	4	17	6	1	4	3	3	3	1
MarR	10	23	???	2	14	30	14	7	12	11	9	14	9
MarR_CONFIRMADO	8	20		2	13	26	12	6	9	9	9	10	6
MerR	2	9	???	???	2	15	11	???	2	1	2	5	2
MerR_CONFIRMADO		4				9	2		1	1	2	3	1
NifT	???	1	???	???	???	1	???	???	???	???	???	1	1
NifT_CONFIRMADO		1				1						1	
OmpR	63	107	5	6	82	136	119	14	23	37	31	44	63
OmpR_CONFIRMADO	62	107	5	6	82	135	117	13	23	37	30	44	62
PadR	3	2	???	1	2	8	2	2	3	1	2	2	2
PadR_CONFIRMADO	3	2		1	1	5	2	2	1	1	2	1	1
PatA	58	105	4	4	80	125	115	10	20	37	30	40	59
PatA_CONFIRMADO	5	10			10	9	13		2	4	4	2	4
ROK	1	2	???	???	1	2	2	1	1	2	1	1	3
ROK_CONFIRMADO	1	1			1	1	2	1	1	2	1	1	1
Rrf2	2	5	???	???	2	4	1	???	3	2	1	2	2
Rrf2_CONFIRMADO	2	4			2	2	1		3	2	1	2	1
SfsA	1	2	???	1	1	2	1	1	1	1	1	1	1
SfsA_CONFIRMADO	1	1		1	1	1	1	1	1	1	1	1	
tetR	3	9	???	???	4	25	6	???	2	1	1	1	3
tetR_CONFIRMADO	2	8			4	24	6		1			1	1
unclassified	5	9	???	???	5	7	6	4	3	4	3	3	4
unclassified_CONFIRMAD	4	8			5	5	5	4	3	3	3	3	2