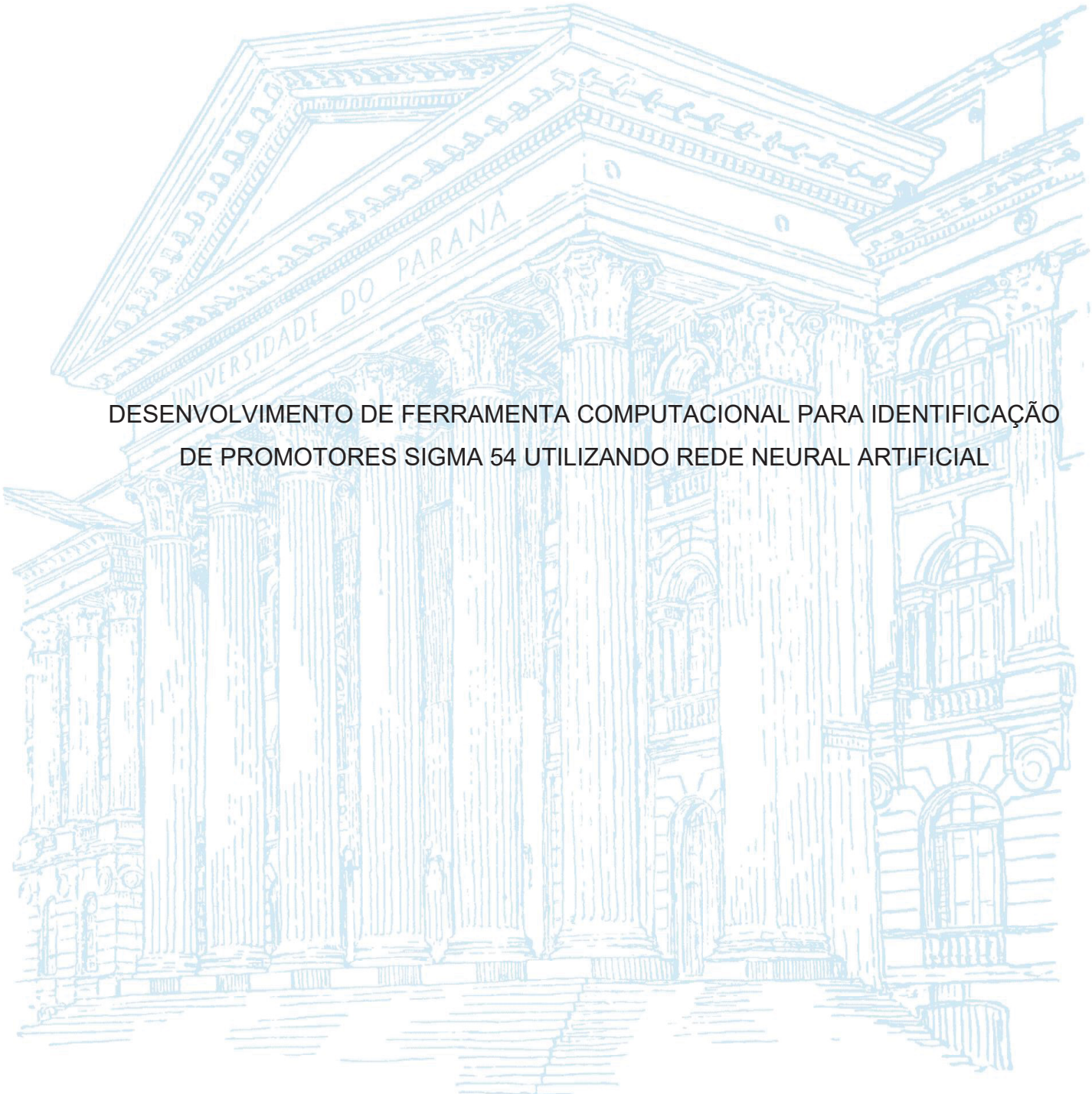
UNIVERSIDADE FEDERAL DO PARANÁ

LUCAS MARTINS FERREIRA



DESENVOLVIMENTO DE FERRAMENTA COMPUTACIONAL PARA IDENTIFICAÇÃO DE PROMOTORES SIGMA 54 UTILIZANDO REDE NEURAL ARTIFICIAL

CURITIBA

2012

LUCAS MARTINS FERREIRA

DESENVOLVIMENTO DE FERRAMENTA COMPUTACIONAL PARA IDENTIFICAÇÃO DE PROMOTORES SIGMA 54 UTILIZANDO REDE NEURAL ARTIFICIAL

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de Mestre em Bioinformática.

Orientadora: Profa. Dra. Liu Un Rigo
Co-orientador: Roberto Tadeu Raittz
Co-orientador: Paulo Afonso Bracarense Costa

CURITIBA

2012

FICHA CATALOGRÁFICA ELABORADA PELO SISTEMA DE BIBLIOTECAS/UFPR –
BIBLIOTECA DE EDUCAÇÃO PROFISSIONAL E TÉCNICA COM OS DADOS FORNECIDOS PELO AUTOR

Ferreira, Lucas Martins
Desenvolvimento de ferramenta computacional para identificação de promotores Sigma 54 utilizando rede neural artificial**.** / Lucas Martins Ferreira. – Curitiba, 2012**.**

Dissertação (Mestrado em Bioinformática) – Setor de Educação Profissional e Tecnológica da Universidade Federal do Paraná.

Orientadora: Profa. Dra. Liu Um Rigo
Coorientador: Roberto Tadeu Raitz
Coorientador: Paulo Afonso Bracarense Costa

1. RNA Polimerase Sigma 54. 2. Sequenciamento de nucleotídeo. 3. Bioinformática. 4. Software - Desenvolvimento. I. Universidade Federal do Paraná. II. Rigo, Liu Um. III. Raitz, Roberto Tadeu. IV. Costa, Paulo Afonso Bracarense. V. Título.

CDD – 006.3

Bibliotecária: Thays Luciana Barbosa de Farias – CRB 9/1995

TERMO DE APROVAÇÃO

LUCAS MARTINS FERREIRA

Desenvolvimento de ferramenta computacional para identificação de promotores sigma 54 utilizando rede neural artificial

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: Profª. Drª. Liu Un Rigo

Coorientador: Prof. Dr. Roberto Tadeu Raittz

Coorientador: Prof. Dr. Paulo Afonso Bracarense Costa

Profª. Drª. Gertrudes Aparecida Dandolini
Universidade Federal de Santa Catarina - UFSC

Profª. Drª. Rose Adele Monteiro
Universidade Federal do Paraná - UFPR

Curitiba, 28 de fevereiro de 2012

Dedico este trabalho aos meus pais: José Isaias e Josefina;

Às minhas irmãs: Maria Emilia e Lycianne;

Aos meus sobrinhos: Helena, João Davi e Pedro Henrique;

À minha namorada: Mariana;

Aos tios: Maria de Jesus e Miguel;

E aos meus cunhados: Paulo Afonso e Danilo.

AGRADECIMENTOS

A minha orientadora Professora Dra. Liu Un Rigo pelo apoio, grandes ensinamentos, confiança e por ter acreditado no projeto.

Ao meu co-orientador Professor Dr. Roberto Tadeu Raittz, pelos ensinamentos, paciência no laboratório, apoio, confiança e acreditar no projeto.

Ao co-orientador Professor Dr. Paulo Afonso Bracarense Costa, pelas aulas, apoio e acreditar no projeto.

Ao professor Dr. Fábio de Oliveira Pedrosa, pelas discussões sobre o assunto, ajudando a entender melhor o funcionamento, pelas correções no trabalho.

Ao professor Dr. Emanuel Maltempi de Souza, pelas discussões sobre o assunto, ajudando a entender melhor o funcionamento, pelas correções no trabalho.

Ao professor Dr. Adriano Barbosa da Silva, pela ajuda no desenvolvimento do trabalho, diversas conversas e correções.

Aos amigos e colegas de laboratório, Vanely, Juliana, Sérgio, Paula, Danhylo, Jesse, Alysson, Gustavo, Ricardo, Rodrigo, Rosa, Dieval.

Aos meus amigos Luciane, Juciellen, Leandro, Thiago, Odair, Joaquim, Rogério pela amizade e conversas animadas.

Ao Núcleo de Fixação de Nitrogênio do departamento de Bioquímica e Biologia Molecular, da Universidade Federal do Paraná.

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

Aos órgãos fomentadores: CAPES, CNPq e REUNI.

RESUMO

A transcrição de vários genes bacterianos é regulada por fatores sigma alternativos da RNA polimerase como o sigma 54 ou sigma N. A sequência consenso (mrNrYTGGCACG-N4-TTGCWNNw) do fator sigma 54 esta localizada 12 pares de bases acima do sitio de iniciação da transcrição e as bases mais conservadas estão localizadas nas posições -25/-24 (GG, 100% de conservação) e -13/-12 (GC, 96% de conservação). Algumas abordagens utilizando modelos ocultos de Markov (HMM) são reportadas em literatura para identificação de sequências promotoras em genomas completos. No entanto, estes métodos não apresentam resultados satisfatórios. Neste teste utilizamos um algoritmo que pré-seleciona candidatos a promotores sigma 54 baseando no padrão de conservação. Os candidatos são então classificados utilizando uma rede neuronal artificial treinada com um conjunto de sequências de promotores sigma 54 validados e um conjunto de sequências improváveis composto por metade apresentando o dois nucleotídeos “GG” e “GC” mais conservados e a outra metade com bases aleatórias. O método foi testado com o genoma da bactéria *Herbaspirillum seropedicae*, resultando em 3148 sequências candidatas com os padrões de conservação “GG” e “GC”. Dentre estes, 126 são considerados regiões verdadeiras de ligação do fator de transcrição sigma 54 pela rede neuronal. Todas as sequências validadas de fatores sigma 54 em *H. seropedicae* foram identificadas pelo programa . Portanto, nossa abordagem é capaz de detectar fatores de transcrição sigma 54.

Palavras-chave: Sequência de nucleotídeos. DNA. Fator sigma. Promotor.

ABSTRACT

The transcription of many bacterial genes is regulated by alternative sigma factors of the RNA polymerase such as the sigma 54 or sigma N. The consensus sequence (mrNrYTGGCACG-N4-TTGCWNNw) of the sigma 54 promoter is located 12 base pairs upstream of the transcription start site and the most conserved bases are located at the positions -25/-24 (GG, 100% conservation) and -13/-12 (GC, 96% conservation). Several approaches using Hidden Markov Models (HMM) have been reported in the literature to identify promoter sequences in whole genomes. However, these methods frequently do not show satisfactory results. In this test we used an algorithm that pre-sort candidates for sigma 54 promoter sequences based on the presence of the conserved motifs. The candidates are then screened using an artificial neural network trained with a set of validated sigma 54 promoter sequences and another set of improbable sequences composed by half showing the two dinucleotides “GG” and “GC” most highly conserved and the another half with random bases. We also tested the method in the whole genome sequence of the bacterium *Herbaspirillum seropedicae*, resulting in 3148 candidate regions with the conserved GG and GC motifs. Out of these, 126 were considered true sigma 54-dependent promoter by the trained neural network. All the validated sigma 54 promoters of H. seropedicae were identified by our software. Therefore, our approach is capable of reliable detection of sigma 54 promoters.

Keyword: Nucleotide sequence. DNA. Sigma factor. Promoter.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

DNA	-	Ácido Desoxirribonucléico
NCBI	-	National Center for Biotechnology Information
RNA	-	Ácido Ribonucléico
RNAm	-	Ácido Ribonucleico Mensageiro
RNAr	-	Ácido Ribonucleico Ribossomal
RNAt	-	Ácido Ribonucleico Transportador
EMBL	-	European Molecular Biology Laboratory Nucleotide Sequence Database
DDBJ	-	DNA Databank of Japan
HMM	-	Hidden Markov Models
FBN	-	Fixação Biológica de Nitrogênio
MATLAB	-	Matrix Laboratory
FAN	-	Free Associative Neurons
MLP	-	Multilayer Perceptron
PAM	-	Point Accepted Mutation
pH	-	Potencial Hidrogeniônico
µm	-	Micrometros
DDR	-	Double Data Rate
Gb	-	Giga Bytes
HDD	-	Hard Drive Disk
SSD	-	Solid State Disk
ORF	-	Open Reading Frame

SUMÁRIO

1. INTRODUÇÃO

1.1. BIOINFORMÁTICA

Ao longo dos anos a informática vem se tornando uma área do conhecimento muito importante para todos nós, fazendo associações com diversas outras áreas, ajudando-as a chegar a lugares nunca esperados. Com a Biologia não foi diferente, a associação entre estas duas áreas deram origem a uma nova área interdisciplinar, a Bioinformática.

Esta nova área do conhecimento auxilia no processamento da grandiosa quantidade de dados e informações que a Biologia Molecular, um ramo da Biologia que gera sequências genômicas.

Para o Centro Nacional de Informações para Biotecnologia (NCBI, sigla em inglês) o conceito de Bioinformática é o campo da ciência no qual a Biologia, a Ciência da Computação e a Tecnologia da Informação se unem para formar uma única disciplina, sendo o seu objetivo final do campo a descoberta de novos conhecimentos biológicos (NCBI, 2004).

Também para o NCBI, a Biologia do século XXI está sendo transformada de uma biologia baseada somente no laboratório para uma ciência da informação, e a Informática ajuda no entendimento de vários processos biológicos (NCBI, 2004).

Segundo Fox (2009), a Bioinformática deve envolver a integração de computadores, ferramentas de *software* e bancos de dados em um esforço para o direcionamento de questionamentos biológicos. Assim, a bioinformática é a conversão de informações biológicas em modelos computacionais processáveis (FOX, 2009).

A Bioinformática como uma nova área de conhecimento, trouxe empolgação para a classe ciêntifica, pela possibilidade de imersão em um mundo totalmente novo e desconhecido (FOX, 2009). Para Bayat (2002), a Bioinformática é uma matéria interdisciplinar que abrange várias áreas do conhecimento como Biologia, Medicina, Matemática, Física, Ciências da Computação e Estatística.

O profissional da área de Bioinformática deve ter conhecimentos específicos nas disciplinas Biologia e Ciências da Computação, detendo a capacidade de entender assuntos relacionados à Biologia Molecular além da aptidão de desenvolver *softwares* ou utilizar outros desenvolvidos por terceiros. Algumas das atividades realizadas por esta nova disciplina envolvem estudar e simular o metabolismo de células, construir árvores evolutivas, estudar estruturas tridimensionais de moléculas, analisar imagens e sinais biológicos (ARAGUAIA, 2011).

A Bioinformática remete à década de 1960, quando a pesquisadora Margaret Dayhoff (1925-1983) organizou e disponibilizou o primeiro atlas de sequências protéicas, publicado com o seguinte título *Atlas of Protein sequence and structure* (DAYHOFF, 1969 *apud* FOX, 2009). Outro grande feito para a Bioinformática da mesma pesquisadora foi o desenvolvimento da PAM *(Point Accepted Mutation)* em 1966, sendo uma matriz para a substituição de aminoácidos, esta que é largamente utilizada até os dias de hoje.

Os avanços no campo da computação trouxeram várias facilidades para a Bioinformática, podendo-se armazenar uma maior quantidade de dados com qualidade e velocidade no processamento das informações. Com os aprimoramentos na tecnologia houve um aumento do número de projetos de montagem de genomas. Um exemplo é o próprio genoma humano com o seu sequenciamento anunciado no dia 26 de junho de 2000, 5 anos antes da data que fora anunciada em 1987.

Para efeito de comparação um gene com uma média de 12 mil pares de bases, hoje leva menos de um minuto para ser sequenciado, há três anos a mesma sequência levaria cerca de uma hora e por fim há 20 anos, a mesma tarefa levaria um ano para ser concluída (VOGT, 2003).

As principais bases de dados que armazenam informações para a Bioinformática são o NCBI (Centro Nacional para Informações de Biotecnologia, sigla em inglês), EMBL (Instituto Europeu de Bioinformática, sigla em inglês) e por ultimo, mas não menos importante o DDBJ (Base de dados de DNA do Japão, sigla em inglês).

O objetivo destas bases de dados é o fomento e armazenamento de dados importantes para o desenvolvimento das mais variadas atividades recorrentes ao estudo como armazenamento de dados, análise e manipulação de dados genéticos e a análise da transcrição e seus reguladores.

1.2. TRANSCRIÇÃO

Durante o processo de transcrição um sistema enzimático converte a informação genética de um segmento de DNA em uma fita de RNA com uma sequência de bases complementares a uma das fitas do DNA (Nelson e Cox, 2000), ou seja, a transcrição faz a passagem de informações contidas na molécula de DNA para uma simples fita de RNA.

A transcrição assemelha-se a replicação em seu mecanismo químico fundamental, sua polaridade e seu uso de um molde, possuindo também semelhanças em suas fases, iniciação, elongação e terminação (Nelson e Cox, 2000).

As regiões a serem transcritas possuem sinalizadores, que são sequências reguladoras específicas que indicam o ponto onde deve ser iniciada a transcrição e onde deve ocorrer a terminação e também qual fita de DNA será utilizada como fita molde, sendo a fita antiparalela a fita molde recebe o nome de fita codificadora (Nelson e Cox, 2000).

O processo de transcrição gera vários tipos de RNAs, podendo ser RNA mensageiro (RNAm), o RNA transportador (ou de transferência, RNAt) e o RNA ribossomal (RNAr), além destes três tipos existem outros RNAs sintetizados no processo de transcrição, sendo estes três os mais importantes.

- RNA mensageiro - representam apenas 2% de todo o RNA presente nas células, codifica a sequência de aminoácidos de um ou mais polipeptídios especificados por um gene ou um conjunto de genes.
- RNA transportador - lê a informação codificada no RNA mensageiro e leva os aminoácidos correspondentes a ela, até os ribossomos, para a síntese da cadeia polipeptídica crescente durante a síntese das proteínas.
- RNA ribossomal - é o RNA constituinte do ribossomo, que são máquinas celulares que são responsáveis pela síntese das proteínas.

A transcrição é realizada por uma holoenzima denominada RNA polimerase, sendo esta DNA dependente. Sendo presente em procariotos e em eucariotos. Neste ultimo estando presente três tipos de RNA polimerase (Nelson e Cox, 2000).

A nova fita de RNA é sintetizada na direção 5´-> 3´, antiparalelo a fita molde de DNA, os nucleotídeos são adicionados respeitando interações de pareamento de bases Watson-Crick, havendo uma substituição do ligante timina com a adenina, no caso havendo o pareamento da uracila (Nelson e Cox, 2000).

A RNA polimerase pode ser isolada nas células de duas formas, completa com as cinco subunidades, formando a holoenzima completa, ou apenas com quatro delas, formando a apoenzima (KUMAR, 1981), sendo elas, β e β´ (beta e beta linha), formam o centro catalítico da enzima e estão presentes em todas as fases da transcrição, duas subunidades α (alfa), estas quatro formando a apoenzima e a última, a subunidade, ou fator, σ (sigma), que quando ligada às demais forma a holoenzima, assim demonstrado na Figura 1, esta subunidade não sendo fixa na RNA polimerase e reconhecendo o sitio de ligações ao DNA específicas, regiões denominadas promotoras (WÖSTEN, 1998).

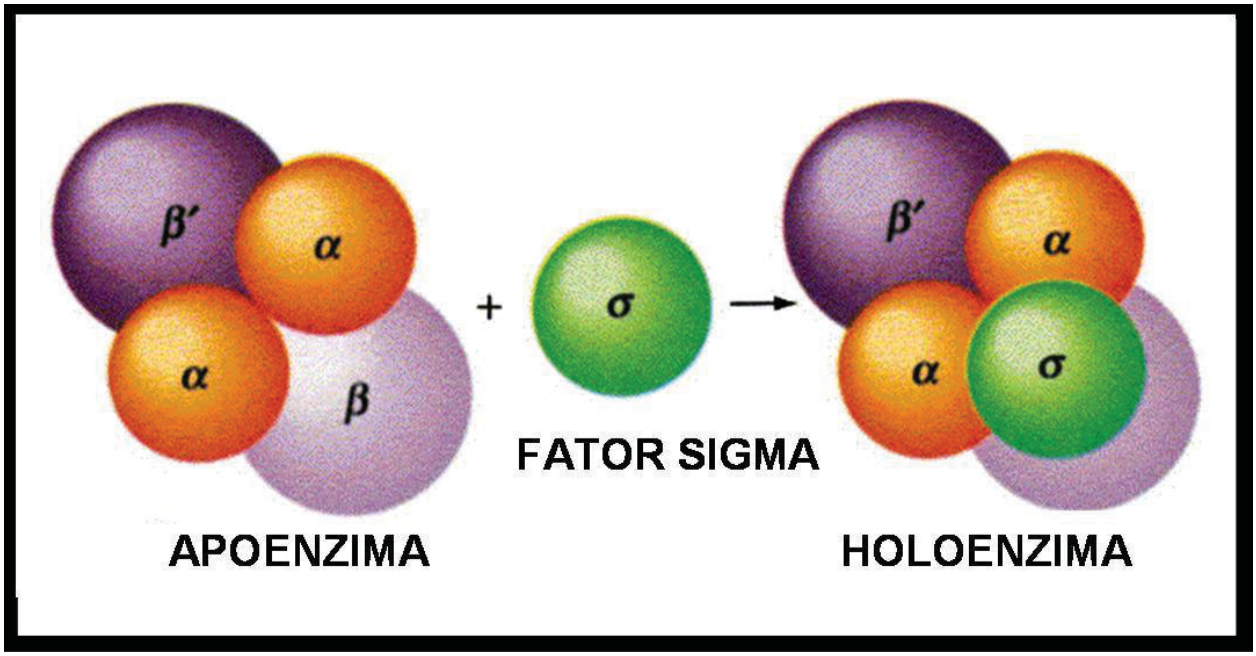


FIGURA 1 - FIGURA ILUSTRATIVA DA LIGAÇÃO DO FATOR SIGMA À APOENZIMA FORMANDO A HOLOENZIMA
FONTE: Genetica On Line (2012)

O processo de transcrição é dividido em quatro etapas, sendo elas o reconhecimento do promotor, a iniciação, a elongação e a terminação da transcrição (KUMAR, 1981, VON HIPPEL, 1984).

O reconhecimento do promotor é feito pelo fator sigma da RNA polimerase e a partir deste local será formado o complexo aberto e dada à iniciação da transcrição, a elongação ocorre após a liberação do fator sigma pela RNA polimerase e a apoenzima percorre a sequência realizando a sua cópia. A terminação se dá quando o transcrito força o seu desligamento da molécula de DNA, dependente de fator protéico ou não, como demonstrado na Figura 2.

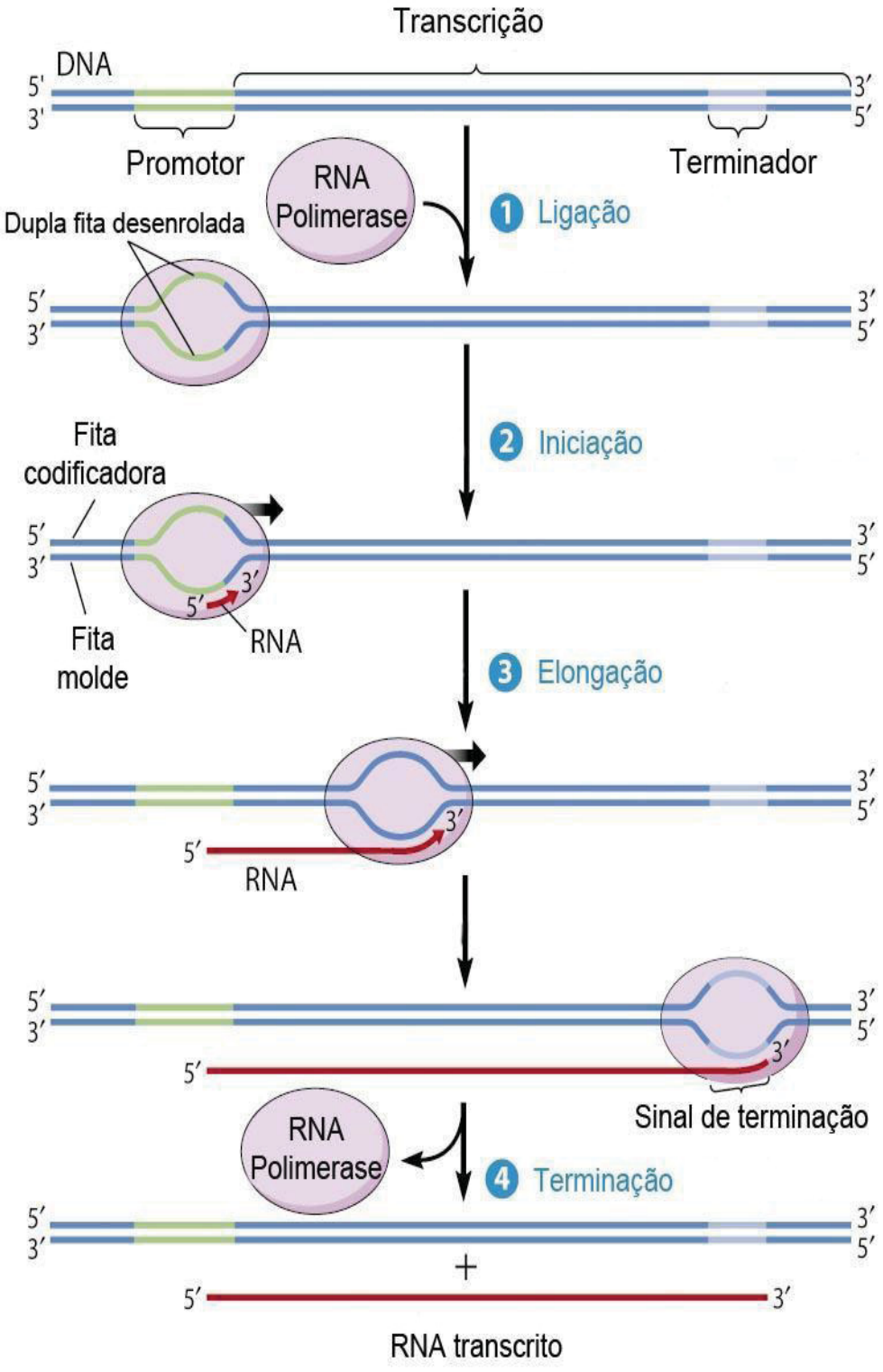


Figura 2- IMAGEM ILUSTRATIVA DO PROCESSO DE TRANSCRIÇÃO FONTE: O autor (2012)

1.3. REGULAÇÃO DA TRANSCRIÇÃO EM PROCARIOTOS

A célula não é uma ilha, não estando isolada do ambiente no qual está inserida respondendo a sinais intra e extracelulares. Estes sinais são captados e respondidos de acordo com a necessidade de adaptação que os organismos são submetidos como a variação de temperatura, de pH, de osmolaridade e disponibilidade de nutrientes (PARKINSON,1993; STOCK, NINFA E STOCK, 1989).

Essas redes de transdução de sinal são formadas por proteínas capazes de interagir entre si para produzir respostas condizentes ao ambiente (STOCK, ROBINSON & GOUDREAU, 2000).

As respostas adaptativas das bactérias modificam o comportamento da transcrição, modificando quais enzimas são transcritas para reorganizar o metabolismo celular. Para esta resposta as principais vias metabólicas têm as atividades de suas enzimas alteradas (WHITE, 2000) reorganizando o fluxo metabólico de acordo com o novo ambiente. Após ocorre mudanças mais profundas havendo modificação do padrão da expressão gênica em resposta aos sinais recebidos resultado em um novo conjunto de proteínas transcritas (STROCK, NINFA e STROCK, 1989).

Para atuação deste mecanismo de sinalização e regulação, a célula desenvolveu alguns sistemas de monitoramento dos sinais, um exemplo é o sistema regulador de dois componentes. Neste sistema, um dos componentes funciona como sensor e outro como regulador, fornecendo aos organismos a capacidade de perceber, interpretar e responder aos sinais (STOCK, ROBINSON, GOUDREAL, 2000; FOUSSARD et al, 2001; GALPERIN, 2004).

Neste contexto, a proteína sensora atua como quinase ou fosfatase onde através do processo de fosforilação ou desfosforilação sinalizando ao outro componente, este por sua vez atuará como regulador, promovendo ou reprimindo a transcrição (WEST E STOCK, 2001).

Como exemplos de regulação de transcrição que ocorre no processo da fixação biológica de nitrogênio, podemos citar os genes *ntrBC*(NIXON, RONSON e

AUSUBEL, 1986). Neste processo, a função de sensor é atribuído à proteína *ntrB* a função de regulador do processo de transcrição é a proteína *ntrC.*

As proteínas que possuem a função ativação da transcrição atuam juntamente com os fatores sigmas para a iniciação da transcrição regulando os genes a serem utilizados no processo.

1.4. FATORES SIGMA

Os fatores sigma são subunidades da RNA polimerase capazes de reconhecimento do promotor, àssim somente a holoenzima completamente formada é capaz de fazer a ligação com a molécula de DNA e fazer as mudanças conformacionais necessárias para a separação da dupla fita e iniciação da transcrição (ISHIHAMA, 1990).

Os fatores sigma facilitam a transcrição em pontos específicos da sequência do DNA fazendo a ligação com a RNA polimerase completando a holoenzima e então realiza o reconhecimento da região especifica de ligação à sequência de DNA. A holoenzima abre a dupla hélice para a iniciação do processo de transcrição (DOUCLEFF et al, 2005).

Esse ciclo de ligação com os fatores sigma ajuda no aprimoramento do metabolismo celular nas respostas às mudanças que ocorrem no ambiente onde as células estão inseridas, além das respostas aos sinais, orquestrando o desenvolvimento da célula utilizando conjuntos diferentes de genes transcritos (MOONEY, 2005).

Os fatores sigma são subdivididos em duas grandes famílias, a família sigma 70 (σ70), relacionada com a manutenção/sobrevivência da célula (MOONEY, 2005) e família sigma 54 (σ54) ou proteína *rpoN*, uma família de fatores sigmas alternativos (BARRIOS et al, 1999). Eles são denominados em relação ao peso molecular do primeiro membro da família identificado (DOUCLEFF et al, 2005 *apud* WÖSTEN, 1998).

A ligação da apoenzima ao fator sigma 70 constitui a Holoenzima que tem a capacidade de inicialização da transcrição por si mesma, pois consegue completar a formação do complexo aberto (McClure, 1985; GRALLA, 1990). Quando a apoenzima se liga ao sigma 54, não existe a mesma capacidade de formação do complexo aberto, e há a necessidade de ligação com outros fatores protéicos para a ativação da transcrição (SASSE-DWIGHT & GRALLA, 1988; MORETT & BUCK, 1989; POPHAM et al., 1989 e MORETT & SEGOVIA, 1993).

Estudos realizados em *Escherichia coli,* reportaram a presença de sete fatores sigma formadores de duas famílias principais de fatores sigma. A família do sigma 70, constituída por seis destes fatores, o próprio sigma 70 (σ70) mais os fatores sigma 38 (σ38), sigma 32 (σ32), sigma 28 (σ28), sigma 24 (σ24) e sigma 19 (σ19). O último fator identificado nos estudos é o fator sigma 54 (σ54) único que compõe a família do sigma 54 (WÖSTEN, 1998).

Cada uma das famílias de fatores sigma reconhece uma região de ligação ao DNA especifica, não havendo possibilidade de uma família se ligar à região de ligação da outra. Outras características que diferenciam as famílias são a isomerização e a regulação (BARRIOS et al, 1999).

O sítio de ligação da família sigma 70 compreende os hexâmetros posicionados nas bases -35 e -10 em relação a primeira base de transcrição, diferente do que ocorre para a família do sigma 54 onde a holoenzima reconhece as bases -24 e -12, também em relação a primeira base de transcrição (BARRIOS et al, 1999).

Os padrões de conservação para as regiões de ligação para os sigmas 70 e 54 estão demonstrados na Figura 3.

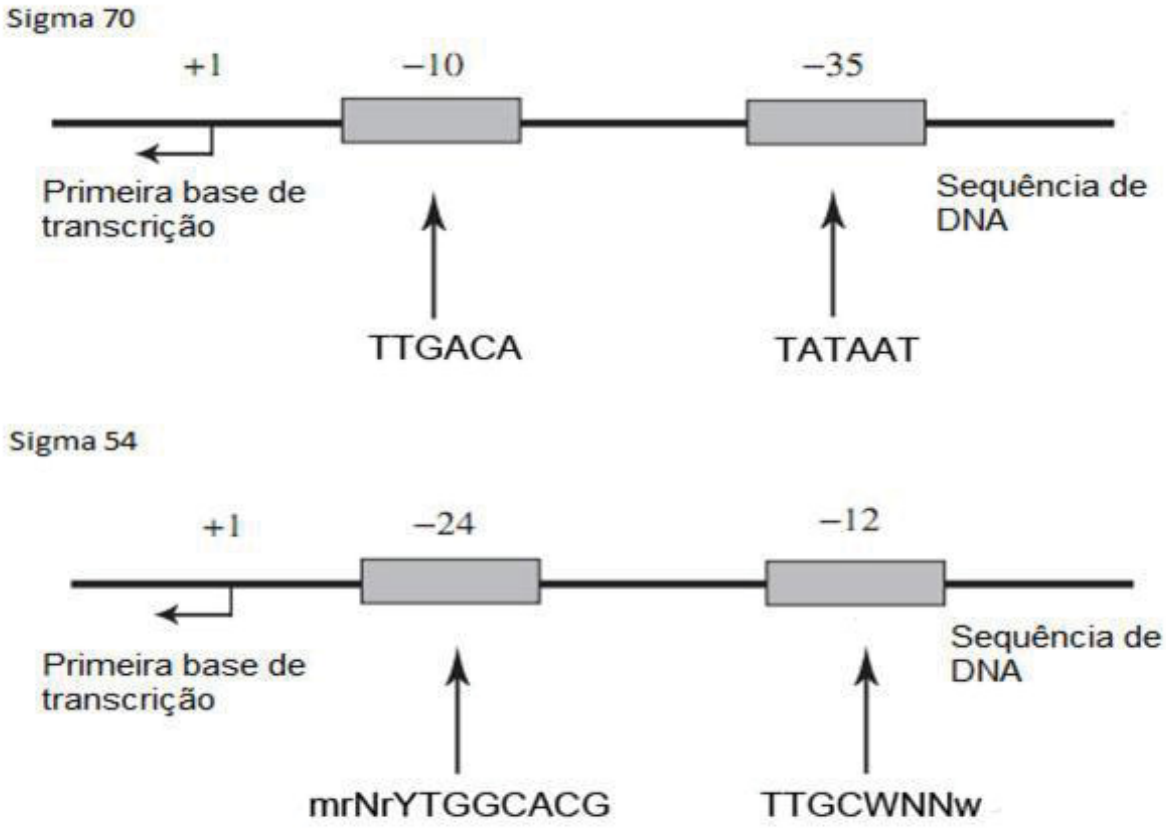


FIGURA 3- REPRESENTAÇÃO GRÁFICA DAS SEQUENCIAS CONSENSO PARA OS FATORES SIGMA 70 E 54
FONTE: Adaptado de Potvin (2005) e referências de Harley e Reynolds (1987), Helmann (1995), Graves & Rabinowitz (1986) e Barrios et al (1999).

Os fatores sigma pertencentes à família sigma 70 são subdivididos em quatro grupos estruturalmente relacionados, mas funcionalmente diferentes (WÖSTEN, 1998; LONETTO, GRIBSKOV & GROSS, 1992). Os fatores sigma pertencentes ao grupo 1, realizam a transcrição de genes responsáveis pela manutenção da sobrevivência da célula, já os pertencentes aos grupos 2 e 3 transcrevem genes responsáveis pela resposta a sinais apropriados, que fazem parte do último grupo possuem funções extracelulares.

Aqueles fatores que pertencem à família do sigma 54 apresentam pouca ou nenhuma semelhança com outras famílias. Até 1986 acreditava-se que o σ54 estivesse ligado somente à regulação da transcrição de genes responsáveis pelo metabolismo de fontes alternativas de nitrogênio, mas posteriormente outras funções

foram atribuídas a ele. Entre estas funções está à degradação de tolueno e xileno, transporte de ácidos de carboxílicos, captação de hidrogênio, construção de flagelos, são algumas das novas funções atribuídas.

A região de ligação à molécula não sofre muitas mutações apresentando uma grande conservação nas bases mais presentes. No estudo realizado por Barrios et al; (1999), a base disposta na posição -24, chega a conservação de 100%, e a menor conservação presente nas bases posicionadas nos locais -13/-12 com 96% de conservação. Já a base posicionada na base -25 apresenta conservação de 99% assim formando os dois dinucleotídeos com maior conservação presente no promotor do sigma 54.

1.5. FIXAÇÃO BIOLÓGICA DE NITROGÊNIO

O nitrogênio é um elemento químico encontrado em abundância na atmosfera terrestre, compreende aproximadamente 80% da mesma, estando em uma forma estável, inerte, onde apenas alguns seres vivos são capazes da realização de assimilação do mesmo em forma gasosa (HOWARD e REES, 1996).

Estando presente em aminoácidos, proteínas, DNA, RNA e outras estruturas celulares, o nitrogênio é deveras importante para todos os seres vivos, levando-o assim a um alto patamar de relevância no processo de fixação biológica de nitrogênio.

A fixação biológica de nitrogênio é responsável por 65% de todo o nitrogênio fixado e 10% é resultado de processos naturais, como raios, erupções vulcânicas e radiação ultravioleta (HUNGRIA e CAMPO, 2005 e MOREIRA E SIQUEIRA, 2006). Os outros 25% restantes são resultantes de processos industriais em forma de adubo químico, constituindo um processo de alto custo.

Economicamente, a fixação biológica de nitrogênio também tem sua importância, já que é utilizado como fertilizante natural para o desenvolvimento de culturas como, sorgo, arroz, cana de açúcar, milho, entre outros.

O processo de fixação consiste na quebra da tripla ligação entre as duas moléculas do dinitrogênio realizado pelo complexo enzimático denominado nitrogenase onde bactérias reduzem o dinitrogenio (N_2) atmosférico em amônia (NH_3). (POSTGATE et al).

A relação estequiométrica da reação catalisada pela nitrogenase é representada na seguinte equação (BURRIS, 1991).

$$N_2 + 8H^+ + 8e^- + 16MgATP \rightarrow 2NH_4^+ + H_2 + 16MgADP + 16Pi$$

O processo de fixação biológica de nitrogênio requer grande quantidade de energia, e por esta razão requer um elevado grau de regulação, principalmente a nível transcricional.

Esta regulação da transcrição dos genes que realizam a fixação biológica do nitrogênio, genes *nif*, ocorre através do gene *nifA* e do fator de transcrição sigma 54 (DIXON e KHAN, 2004). A expressão do *nifA* é regulada pelos produtos de genes que atuam em cascata. Este grupo de proteínas recebe o nome de sistema regulatório de nitrogênio, ou simplesmente *ntr*.

O sistema *ntr* é constituido por sete proteínas, que incluem o sistema de dois componentes, *ntrBC,* as proteínas PII, *GlnB* e *GlnK*, as enzimas uridilidtransferase *GlnD*, adenililtransferase *GlnE* e a glutamina sintetase, *GlnA* (MERRICK e EDWARDS, 1995).

Os organismos capazes de realizar a fixação biológica de nitrogênio são todos procariotos, Bacteria e Arquea, e receberam a denominação de diazotrófos (POSTGATE, 1982).

1.6. DIAZOTRÓFOS

Os organismos diazotróficos são encontrados entre diferentes reinos. No reino bactéria pode ser dividido entre bactérias de vida livre, bactérias associativas e bactérias simbióticas (EVANS & BURRIS, 1992). Estas últimas por sua vez ainda podem ser divididas entre as facultativas e as obrigatórias (BALDANI et al; 1997).

Os primeiros organismos descritos como fixadores biológicos de nitrogênio eram encontrados na rizosfera e no rizoplano (DÖBEREINER, 1992). Outros organismos que foram identificados como diazotróficos habitam o interior da planta e por esta ação são chamados de endofíticos (OLIVARES et al, 1996; REINHOLD-HUREK et al, 2000; URETA et al, 1995).

A utilização de bactérias diazotróficas para o fornecimento de compostos nitrogenados para as culturas pode substituir a utilização de fertilizantes nitrogenados, diminuindo assim a poluição de águas, solo e atmosfera e contribuir para a redução nos custos de produção (PEDROSA, 2005).

Em um estudo realizado por Young (1992) foi descrita uma série de bactérias que realizam o processo de fixação biológica de nitrogênio. Este estudo tem sido utilizado como base para vários outros trabalhos relacionados como o de GEHLEN (2011), por exemplo, sendo desenvolvido um protocolo para a procura de genes *nif* presente em diversas bactérias depositadas no banco de dados do NCBI utilizando de redes neurais artificiais.

Algumas das bactérias listadas como fixadoras biológicas de nitrogênio são: *Azospirillum brasiliense, Azospirillum amazonense e Herbaspirillum seropedicae.* Este último sendo utilizado como referência para o desenvolvimento deste trabalho.

1.7. Herbaspirillum

Herbaspirillum é gênero bacteriano que foi isolado durante buscas realizadas em rizosfera e raízes de cereais (BALDANI et al; 1984). Inicialmente estas bactérias foram classificadas no gênero *Azospirillum* devido a semelhanças entre as mesmas, mesmo tendo menor tamanho e baixa mobilidade (BALDANI et al; 1984). Entretanto, hibridizações RNA-RNA demostraram que não havia relação filogenética entre elas, o que permitiu a criação de um novo gênero bacteriano, *Herbaspirillum* (FALK et al; 1986).

O gênero *Herbaspirillum* apresenta características como ser diazotrófico, endofítico, podendo ser encontrados tanto em tecidos vegetais quanto em órgãos. As espécies são geralmente vibrióide, mas pode apresentar formato helicoidal; tem uma maior tolerância a variação de pH do que o gênero *Azospirillum*, podendo crescer entre pH 3 a 8,0; possui de um a três flagelos em um ou em ambos os polos; seu diâmetro pode variar de 0,6 a 0,7 µm e o comprimento de 1,5 a 5 µm (BALDANI et al; 1986).

O nome *Herbaspirillum* surgiu em referência ao habitat do micro organismo vindo do latim significado de para *herbas* de planta herbácea e *spirillum* pequena espiral (BALDANI et al; 1986).

Atualmente o gênero *Herbaspirillum* é constituído por diversas espécies entre elas *Herbaspirillum seropedicae* e *Herbaspirillum rubrisubalbicans* (BALDANI et al, 1996). *Herbaspirillum frisingense* (KIRCHHOF et al; 2001), *Herbaspirillum putei, Herbaspirillum huttiense, Herbaspirillum autotrophicum* (DING e YOKOTA et al; 2004), *Herbaspirillum lusitanum* (VALVERDE et al; 2003) e *Herbaspirillum chlorophenolicum* (IM et al; 2004).

1.8. Herbaspirillum seropedicae

H. seropedicae, além das características do gênero *Herbaspirillum*, é um micro-organismo muito móvel perto de uma fonte de O_2 e o crescimento na presença de N_2 é mais lento que no gênero *Azospirillum* (BALDANI et al; 1986).

Como todas as espécies do gênero, *H. seropedicae* é uma proteobactéria da classe β (YOUNG et al, 1992). Esta espécie recebeu o *seropedicae* em homenagem à cidade de Seropédica localizada no Estado do Rio de Janeiro, onde foi originalmente isolada (BALDANI et al; 1986). A Figura 4 mostra a bactéria *Herbaspirillum seropedicae*.

H. Seropédica, é uma bactéria endofítica, apresentando uma baixa sobrevivência em solos naturais (OLIVARES et al; 1996). Esta associação podendo ser estabelecida nas sementes, por propagação vegetativa ou de resíduos vegetais mortos (BALDANI et al, 1992; OLIVARES et al, 1996). Esta espécie de diazotrófo é capaz de se associar com gramíneas, de grande importância agrícola como trigo, milho, sorgo, cana de açúcar e outras plantas como banana, abacaxi e outras (BALDANI et al, 1986; YOUNG et al, 1992; CRUZ et al, 2001).

Por essa característica a bactéria pode ser utilizada como biofertilizante, tornando-se de grande importância econômica. Esta finalidade foi demostrada em trabalhos realizados onde regiões ficaram por vários anos sem fertilização nitrogenada e foram reportadas sinais de fixação biológica de nitrogênio, levando a crer na ação de bactérias endolíticas como o *Herbaspirillum seropedicae* (DOBERËINER et al; 1992).

FIGURA 4- ELETROMICROGRAFIA DA BACTÉRIA DE *Herbaspirillum seropedicae*
FONTE: (BALDANI, 1986).

O genoma completo da estirpe SmR1 de *Herbaspirillum seropedicae* foi sequenciado e anotado pelo Programa Genoma no Paraná (Consorcio Genopar, www.genopar.org). Ele é constituído por um único cromossomo que contêm 5.513.887 pares de bases, G+C de 63,4% e um total de 4,804 genes (PEDROSA et al, 2011).

Como em outros diazotrófos, em *H. seropedicae* a transcrição dos genes *nif* (*nifB*, *nifHDK*) requer a presença do fator de transcrição σ54 para a ativação do processo de transcrição (MERRICK, 1992; MACHADO et al, 1996; PEDROSA et al, 1997).

1.9. REDES NEURAIS ARTIFICIAIS

Redes neurais artificiais são técnicas computacionais que implementam uma abordagem de aprendizado semelhante a de seres vivos, simulando o funcionamento do sistema nervoso humano se referindo a capacidade de aprendizagem com as mais diferentes ações, adquirindo conhecimento através da experiência e observações.

Para Rezende (2003, p. 142) as redes neurais artificiais são modelos matemáticos que se assemelham as estruturas neurais biológicas e que tem capacidade computacional de adquiridas por meio de aprendizagem e generalização. Para Nievola (1998, p. 12) as redes neurais consistem em uma abordagem de inteligência artificial para as soluções de problemas que tem por base o modelo de inteligência conhecido: o cérebro humano.

Com a grande complexidade das redes neurais biológicas, possuindo bilhões de neurônios, enquanto que as redes artificiais apenas apresentam de dezenas a milhares de unidades de processamento, a limitação deste método é a velocidade de aprendizagem.

Os neurônios das redes artificiais, que podem ser chamados de nós ou nodos, são interconectados imitando o funcionamento do cérebro humano, sendo dispostos em camadas e conectados a um ou mais neurônios. Essas conexões possuem pesos para o nivelamento da resposta com uma facilidade de adaptação, imitando o funcionamento das sinapses humanas (HAYKIN, 1999). Assim os neurônios são fundamentais para o processamento nas redes neurais (HAYKIN, 2001).

Nas Figuras 5 e 6 é mostra um neurônio biológico e um modelo de rede neuronal artificial, respectivamente.

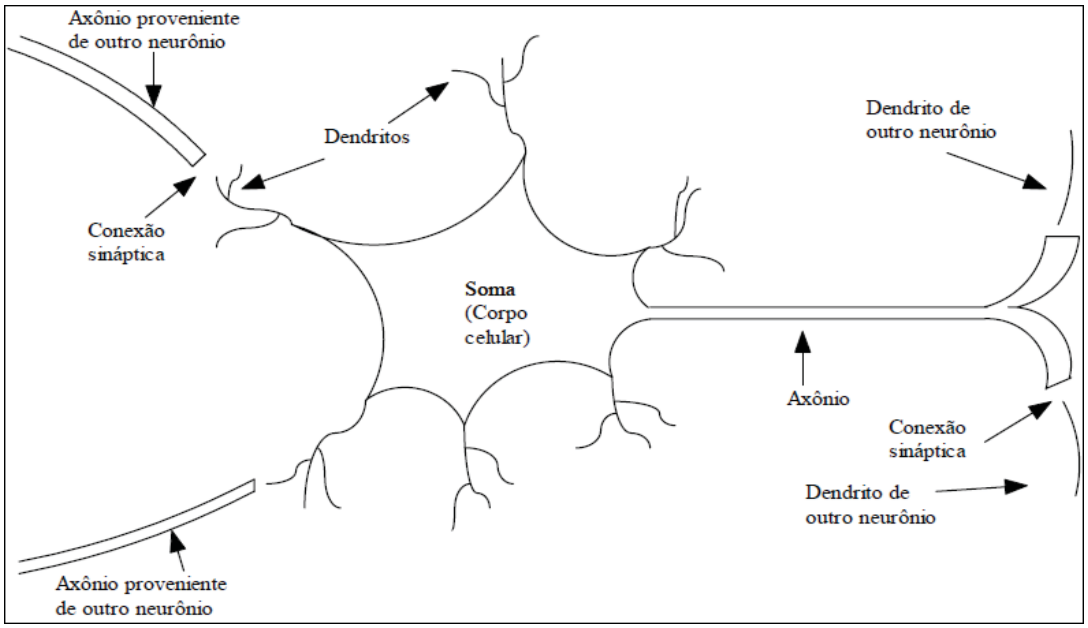


FIGURA 5– NEURÔNIO BIOLÔGICO
FONTE: Adaptado de Fausett (1994)

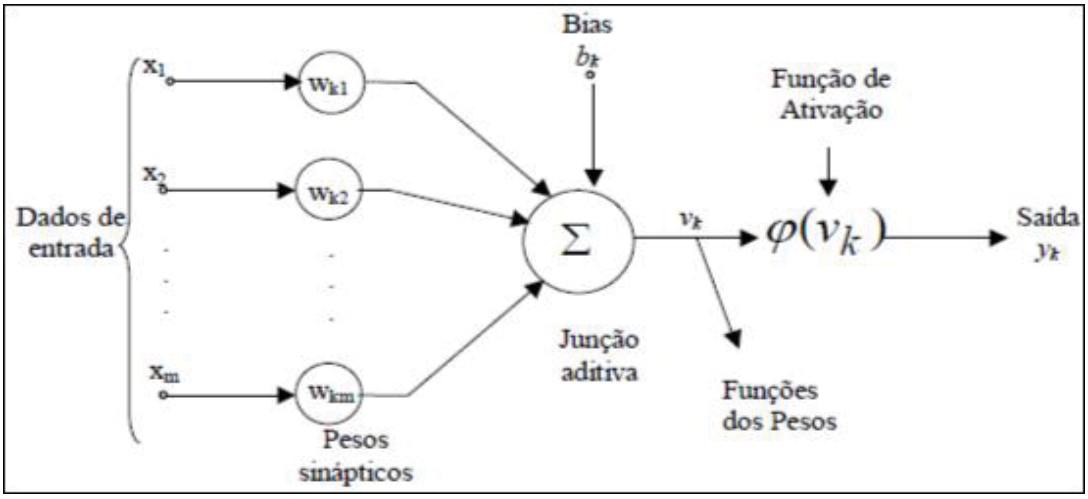


FIGURA 6- DESENHO REPRESENTATIVO DE UMA REDE NEURONAL ARTIFICIAL
FONTE: Adaptado de Haykin (2001)

Estruturalmente as redes apresentam uma disposição em paralelo, para que quando houver uma falha de um ou mais neurônios os demais possam assumir o processamento utilizando uma rota diferente e minimizando os efeitos da falha para o resultado do processamento da rede neural, com isso fazendo o procedimento de tolerância a falhas.

Portanto, o meio de aprendizado utilizado para o reconhecimento do padrão de conservação dos fatores sigma 54 a partir dos candidatos selecionados pelos algoritmos foi à rede neuronal artificial pela sua facilidade de aprendizagem e precisão na resposta obtida.

1.10. EXTRAÇÃO DE CARACTERÍSTICAS

A extração de características é uma técnica utilizada para o aprendizado de máquina, esta pode ser entendida como qualquer medição útil extraída no processo de identificação de um padrão. Podendo ser simbólicas, numéricas ou dos dois tipos ao mesmo tempo. As variáveis sendo contínuas ou discretas (SOUZA, 1999).

Na técnica de extração de características é selecionado um subconjunto das funcionalidades disponíveis para a aplicação destes a um algoritmo de aprendizagem. O melhor conjunto é aquele que apresenta uma melhor precisão com uma baixa quantidade de dimensões (SEWELL, 2007).

As seguintes abordagens para extração de características podem ser utilizadas segundo Sewell (2007)

- *Forward selection*; O processo é iniciado sem nenhuma variável e elas são adicionadas uma a uma, com o erro diminuído a cada etapa de diminuição do erro. O processo é parado quando o erro não é diminuído de forma significativa.
- *Backward selection*: No inicio do processo estão todas as variáveis presentes. Estas sendo removidas uma a uma sendo o erro diminuído até ser estabilizado. Até que em uma remoção não diminua o erro de forma significativa.

Para a bioinformática a extração de características se tornou uma importante aliada no desenvolvimento de suas aplicações, deixando de ser um mero exemplo ilustrativo (SAEYS, 2007).

1.11. FAN

A rede neural Free Associative Neurons, ou FAN, com tradução livre para neurônios de associação livre foi desenvolvida por pesquisadores brasileiros e primeiramente publicada no ano de 1997.

As redes FAN são híbridas contendo suas bases nas redes conexionistas, modelagem difusa e representação de padrões, Além de ter bases nas noções de aprendizado neural. O grande poder das redes FAN se dá pela sua granularidade de informação facilitando a aprendizagem da rede através desta característica (RAITTZ, 1998).

Nas redes FAN cada padrão que entra é expandido de forma a cobrir de forma difusa a sua vizinhança, assim sendo, cada conjunto de entrada acaba por se tornar uma grade com os valores originais sendo o ponto de imprecisão entre as classes próximas. O aprendizado da rede é realizado pela projeção de toda a rede difusa no espaço FAN. Para cada classe do domínio do problema existe uma classe FAN, sendo representada por uma grade com todas as combinações possíveis presentes entre os conjuntos difusos. No treinamento cada célula é representada por uma célula difusa que contém o peso correspondente à sua frequência de ocorrência e grau de pertinência. O treinamento é baseado no reforço da célula, se a classificação for correta, ou no esquecimento se for errada (RAITTZ, 1997).

Graças a essa facilidade com o trabalho com reconhecimento de padrões a rede neural FAN foi escolhida para este trabalho podendo reconhecer as classes para a classificação com maior facilidade.

Um exemplo de utilização da FAN é o FControl, sendo um sistema inteligente para a detecção de fraudes em comercio eletrônico (COELHO, 2006)

1.12. MLP

As redes neuronais *Multilayer perceptrons* (perceptrons de multicamada, sigla em inglês) são redes neuronais que mapeiam conjuntos de entradas de dados em conjuntos de saída apropriados. Estas redes vêm sendo aplicadas em diversas áreas com sucesso, como no reconhecimento de padrões e processamento de sinais.

Uma rede do tipo MLP é constituída de um conjunto de nós fonte, os quais formam a camada de entrada de dados na rede, uma ou mais de uma camada oculta e uma camada de saída. A camada de entrada é a única não constituída por neurônios e assim não possuindo capacidade computacional (HAYKIN, 1994).

As redes *multilayer perceptrons* são progressivas, ou seja, a saída de uma camada alimenta unicamente a entrada da próxima camada sem a presença de realimentação, assim o sinal se propaga através da rede de forma progressiva (HAYKIN, 1994).

Outra característica das MLPs é o algoritmo de treinamento de *backpropagation* que se baseia na heurística de aprendizado por correção do erro, sendo o erro retro propagado das camadas de saída para as camadas ocultas (HAYKIN, 1994).

1.13. HMM

Os modelos ocultos de Markov (Hidden Markov Models, sigla em inglês) é muito utilizado nas ciências aplicadas e na engenharia. E são dois processos estocásticos, o primeiro é uma cadeia de Markov sendo este oculto. Já o segundo produz sequências de símbolos observáveis a cada momento (RABINER & JUANG, 1986).

Esta técnica é aplicada nas mais diversas áreas como, por exemplo: identificar regiões de interesse em códigos genéticos, classificação de proteínas, classificar textos, identificação de sinais linguísticos, dentre outras.

2. OBJETIVOS

2.1. OBJETIVO GERAL

Desenvolver ferramenta computacional capaz de identificar regiões promotoras reguladas pelo fator sigma 54 (σ54) da RNA polimerase, utilizando redes neurais artificiais para a classificação dos candidatos em genomas de bactérias.

2.2. OBJETIVOS ESPECÍFICOS

Identificar o padrão de conservação de bases reconhecido pela holoenzima formada em ligação com o fator sigma 54 (σ54).

Desenvolver um programa computacional para a busca de candidatos à sequências de regiões promotoras com padrão de conservação do sigma 54.

Gerar dois conjuntos de dados aleatórios. Um com as bases mais conservadas presentes nas sequências confirmadas em literatura. E um segundo conjunto com bases completamente aleatórias.

Treinar de redes neurais, MLP, FAN para comparação de resultados e escolha do modelo a ser aplicado.

Integração da rede neural artificial com o algoritmo de seleção para testes de buscas de sequencias reguladas pelo fator sigma 54 (σ54).

Buscas de regiões de falsos positivos classificados erroneamente pela rede neural.

3. MATERIAIS E MÉTODOS

Para o desenvolvimento deste trabalho foi utilizado um computador fornecido pela Universidade Federal do Paraná – UFPR fabricado pela Lenovo de modelo: M90-P e um notebook de próprio fabricado pela LG de modelo: R590, a descrição dos microcomputadores esta contida na tabela 1.

TABELA 1- LISTAGEM DAS CONFIGURAÇÕES DOS COMPUTADORES UTILIZADOS DURANTE O DESENVOLVIMENTO DA APLICAÇÃO

	Lenovo ThinkCentre M90P	**Lg R590-5700**
Plataforma	Sff	Notebook
Processador	Core i5 650 (3,2Ghz)	Core i7 740QM (1,73Ghz)
Memória	4Gb DDR3	6Gb DDR3
Disco Rígido	320Gb	640Gb (HDD) + 120Gb (SSD)

FONTE: O autor (2012)

3.1. MATLAB

O software mais utilizado para o desenvolvimento deste trabalho foi o MATLAB, nome vindo do inglês com o significado de *Matrix Laboratory*. Segundo a desenvolvedora do programa, MathWorks, o MATLAB é uma linguagem computacional técnica de alto nível e um ambiente interativo para o desenvolvimento de algoritmos, visualização e análise de dados e computação numérica (MATHWORKS, 2011).

O Matlab pode ser utilizado para um grande leque de aplicações como processamento de sinais e imagens, comunicação, controle de *desing*, medição e teste, modelamento e análise financeira e computação biológica (MATHWORKS, 2011).

Para o desenvolvimento destas atividades o MATLAB conta com diversos tipos de bibliotecas que contém várias funções desenvolvidas pela Mathworks, como por exemplo, a *Toolbox* de Bioinformática. Em complemento a todas estas bibliotecas disponibilizadas pela desenvolvedora, programadores espalhados pelo

mundo fazem o desenvolvimento de novas funções complementares para o acréscimo das atividades utilizando o MATLAB.

3.2. MICROSOFT EXCEL

Em paralelo ao MATLAB o Microsoft Excel também foi bem utilizado para a implementação deste trabalho. O mesmo fora empregado na tabulação em planilhas eletrônicas dos resultados obtidos e criação de gráficos que representassem as respostas obtidas pela execução do algoritmo.

O Microsoft Excel faz parte da suíte de aplicativos para escritório chamada *Office* desenvolvidos pela Microsoft e segundo a mesma o Excel possibilita a análise, o gerenciamento e o compartilhamento de informações ajudando a tomada de decisão (MICROSOFT, 2011).

3.3. EASYFAN

Com a escolha da rede neural FAN para a realização deste trabalho o Software utilizado para a classificação dos grupos foi o Easyfan. Este software tem a capacidade de suportar aprendizado supervisionado e permite a visualização gráfica das características e dos neurônios das redes treinadas (SOURCEFORGE, 2012).

A rede neural foi treinada de acordo com as características requeridas pelo programa, os conjuntos de dados foram divididos entre três grupos, cada uma deles contendo padrões verdadeiros e falsos, sendo eles os grupos de treinamento e teste.

O programa foi utilizado para a classificação em verdadeiro ou falso para os candidatos coletados no genoma avaliado, a partir da rede treinada.

3.4. ARTEMIS

Para a visualização dos resultados obtidos a partir da execução da aplicação desenvolvida neste trabalho foi utilizado o programa Artemis que foi desenvolvido pelo instituto Sanger (http://www.sanger.ac.uk/resources/software/artemis/).

Este programa é utilizado para a visualização de sequências de DNA e como uma ferramenta para a anotação de genomas, podendo apresentar o resultado de uma única análise ou de um conjunto delas (RUTHERFORD, 2000).

O programa pode ser executado nas mais diversas plataformas computacionais, Microsoft Windows, Apple Mac OS ou Linux já que é um programa desenvolvido em linguagem Java possuindo assim um alto grau de portabilidade (RUTHERFORD, 2000).

No programa podem ser utilizados tipos de dados, sendo eles do Genbank ou EMBL. Além de ser utilizado localmente pode estar contido em um site na internet através de uma applet do java (RUTHERFORD, 2000).

3.5. AMOSTRAS VERDADEIRAS E FALSAS DE LIGAÇÃO DO FATOR DE TRANSCRIÇÃO

Para o treinamento da rede neural os dados de promotores verdadeiros foram coletados em literatura, sendo eles confirmados biologicamente e com as pesquisas em bibliografia foi obtido um total de 445 sequências verdadeiras em diversos organismos, como demonstrado na Tabela 2.

TABELA 2- QUANTIDADE DE SEQUENCIAS DE LIGAÇÃO AO FATOR DE TRANSCRIÇÃO SIGMA 54 ADQUIRIDAS COM SUAS RESPECTIVAS REFERÊNCIAS

Quantidade	**Referência**
189 sequências	BARRIOS et al, 1999
36 Sequências	STUDHOLME et al, 2000
220 Sequências	LEANG et al, 2009

FONTE: O autor (2012)

Após a aquisição dos dados contidos em literatura e com a continuação da busca por sequências chegou-se a um banco de dados com os resultados obtidos em estudos realizados em diversas bactérias para a localização de sítios de ligação de fatores sigma 54.

Este banco de dados foi levantado pelo pesquisador Kyle Conway (2009) no laboratório de pesquisas da universidade de Ottawa Canadá, o banco de dados pode ser acessado a partir do link www.sigma54.ca.

Dentro destas planilhas continham as sequencias resultantes dos experimentos com um total de 5662 sequências, estes dos quais alguns exemplos apresentados na tabela 3.

TABELA 3- EXEMPLOS DE REGIÕES DE LIGAÇÃO DO FATOR DE TRANSCRIÇÃO SIGMA 54 RETIRADAS DO BANCO DE DADOS UTILIZADO PARA O TREINAMENTO DA REDE NEURONAL

Bactéria	**Gene**	**Sequência**
***E. coli* str. K-12 substr. MG1655**		
	rpoH	CTGGCACAGTTGTTGCT
	prpB	GTGGCACACCCCTTGCT
	glnK	CTGGCACACCGCTTGCA
	glnA	TTGGCACAGATTTCGCT
***B. subtilis* subsp. subtilis str. 168**		
	levD	TTGGCACGATCCTTGCA
	acoA	CTGGCACACTTCTTGCA
	rocD	TTGGCACAGAACTTGCA
	rocA	ATGGCATGATTCTTGCA

FONTE: O autor (2012)

As sequências falsas foram geradas aleatoriamente a partir de funções desenvolvidas no MATLAB, estas com o tamanho apropriado e dividido entre dois grupos para cobrir a maior quantidade possível de casos falsos. Os grupos foram divididos da seguinte forma.

- Grupo 1 – Contém a mesma quantidade de bases encontrada para o padrão verdadeiro e também as bases com maior conservação segundo BARRIOS et al (1999). Grupo gerado a partir de função desenvolvida no MATLAB, exemplo na tabela 4.
- Grupo 2 – Também contém a mesma quantidade de bases, mas se diferencia por não conter as bases conservadas assim sendo composto por bases completamente aleatórias. Grupo gerado a partir da função *DNARAND* do MATLAB, exemplo na tabela 4.

TABELA 4– SEQUÊNCIAS IMPROVÁVEIS OU FALSAS RETIRADAS COMO EXEMPLO DOS GRUPOS 1 E 2 UTILIZADOS PARA O TREINAMENTO DA REDE NEURONAL

Grupo	**Sequência Sigma 54**
1	
	TGGTCGGGCGCAGGCT
	TGGCTATAACCCTGCC
	AGGTTTGGATTCCGCG
	TGGATTCATATGGGCG
2	
	CATTCATCTAGTGGGT
	TCGGAATTCAAAGCAA
	CAGCACCCATAAGGCT
	TAAGAACACTCACAGT

FONTE: O autor (2012)

Estes conjuntos demonstrados na tabela 4 foram utilizados para a construção dos conjuntos de treinamento e teste no treinamento da rede neural.

3.6. TREINAMENTO DA REDE NEURAL

A rede neural foi treinada utilizando o programa EasyFan empregando os conjuntos criados a partir da junção dos conjuntos de sequências verdadeiras e falsas, foi utilizado as especificações originais do programa apenas fazendo pequenas alterações no peso para o padrão verdadeiro afim de se conseguir um melhor resultado do treinamento.

As características dos dados foram extraídas por uma função desenvolvida no MATLAB para que com estes atributos a rede neuronal possa reconhecer e classificar as regiões candidatas.

O resultado deste algoritmo é composto de um vetor de 19 colunas, as 16 primeiras são respectivamente as bases nitrogenadas traduzidas para números, seguindo a tabela 5.

TABELA 5- TRADUÇÃO DE BASES NITROGENADAS PARA NÚMEROS

Bases	Números
A	0.0
C	1.0
G	2.0
T	3.0

FONTE: O autor (2012)

As três últimas colunas são preenchidas respectivamente com o valor do alinhamento contra a sequência consenso retirado do banco de dados utilizado para o treinamento da rede, alinhamento contra o anti-consenso retirado do mesmo banco e o último valor é sorteado aleatoriamente para a classificação entre verdadeiro ou falso para a rede neuronal.

Tomando uma sequência como exemplo, a tabela 6 apresenta o resultado da extração de características para a mesma.

TABELA 6 – EXEMPLO DE EXTRAÇÃO DE CARACTERISTICA

Tipo	
Sequência	TGGTCGGGCGCAGGCT
Tradução da sequência para números	3.0 2.0 2. 3.0 1.0 2.0 2.0 2.0 1.0 2.0 1.0 0 2.0 2.0 1.0 3.0
Alinhamento contra o consenso das sequências	111.5
Alinhamento contra o anti-consenso das sequências	0.02
Valor sorteado para a classificação	1.0

FONTE: O autor (2012)

Na tabela 7 são apresentados alguns exemplos de sequências submetidas à extração de características.

TABELA 7– EXEMPLOS DE SEQUÊNCIAS SUBAMETIDAS A EXTRAÇÃOD E CARACTERISTICAS

Sequência	**Resultado**
TGGTCGGGCGCAGGCT	3.0 2.0 2. 3.0 1.0 2.0 2.0 2.0 1.0 2.0 1.0 0 2.0 2.0 1.0 3.0 111.5 0.02 1.0
TGGCTATAACCCTGCC	3.0 2.0 2.0 1.0 3.0 0 3.0 0 0 1.0 1.0 1.0 3.0 2.0 1.0 1.0 48.5 0.0004 1.0
AGGTTTGGATTCCGCG	0 2.0 2.0 3.0 3.0 3.0 2.0 2.0 0 3.0 3.0 1.0 1.0 2.0 1.0 2.0 0.7578 0.0118 1.0
TGGATTCATATGGGCG	3.0 2.0 2.0 0 3.0 3.0 1.0 0 3.0 0 3.0 2.0 2.0 2.0 1.0 2.0 0.3298 0.001 1.0
CATTCATCTAGTGGGT	1.0 0 3.0 3.0 1.0 0 3.0 1.0 3.0 0 2.0 3.0 2.0 2.0 2.0 3.0 0.0002 0.0022 2.0
TCGGAATTCAAAGCAA	3.0 1.0 2.0 2.0 0 0 3.0 3.0 1.0 0 0 0 2.0 1.0 0 0 0.0272 1.7413 2.0
CAGCACCCATAAGGCT	1.0 0 2.0 1.0 0 1.0 1.0 1.0 0 3.0 0 0 2.0 2.0 1.0 3.0 0.7578 0.1435 2.0
TAAGAACACTCACAGT	3.0 0 0 2.0 0 0 1.0 0 1.0 3.0 1.0 0 1.0 0 2.0 3.0 0.0272 0.0051 2.0

FONTE: O autor (2012)

Após a extração de características os dados foram submetidos ao treinamento da rede em si, o processo que foi realizado diversas vezes até chegar a um nível de confiabilidade da rede neural desejado.

Estes testes foram realizados utilizando dois tipos de redes neurais para a identificação de qual tipo seria utilizado no programa, verificando-se aquela que apresentava uma melhor qualidade nos resultados. Os tipos utilizados para teste foram a MLP (*Multilayer percptron*) e a FAN (*Free Associative Neurons*).

A comparação entre os tipos de redes neurais foi realizada com o objetivo de escolha da melhor rede a ser utilizada para a resolução do problema considerando a quantidade de acertos a que cada uma alcançava para uma melhor identificação dos candidatos a fator sigma 54.

Durante os testes realizados a FAN apresentou um melhor resultado sendo assim escolhida entre as duas testadas para ser utilizada no processamento dos candidatos garantindo uma melhor confiabilidade ao resultado.

Na tabela 8 se observa uma comparação entre as duas redes testadas, exibindo a superioridade do resultado apresentado pela rede neural FAN.

TABELA 8- TABELA COMPARATIVA ENTRE AS REDES NEURONAIS PARA A DECISÃO DE QUAL REDE SERIA UTILIZADA NO DESENVOLVIMENTO DA APLICAÇÃO

	MLP	FAN
Percentual de acerto	95.19%	97.95%

FONTE: O autor (2012)

Nas tabelas 9 e 10, que são um complemento da tabela 8, é demonstrada as matrizes de confusão paras as duas redes neurais testadas.

TABELA 9- MATRIZ DE CONFUSÃO DA REDE NEURONAL MLP

	Classificação verdadeira	**Classificação falsa**	**Total**
Dado verdadeiro	**92,15%**	7,85%	2831
Dado falso	3,29%	**96,71%**	5662
		Total geral	8493

FONTE: O autor (2012)

TABELA 10– MATRIZ DE CONFUSÃO DA REDE NEURONAL FAN

	Classificação verdadeira	**Classificação falsa**
Dado verdadeiro	**95,51%**	4,49%
Dado falso	2,09%	**97,91%**

FONTE: O autor (2012)

Mesmo após a incorporação da rede neural no programa, alguns testes ainda se faziam necessários para garantir a qualidade dos resultados, assim várias redes foram treinadas com diferentes configurações nos arquivos de entrada.

Foram sendo alteradas as quantidades de sequências de promotores utilizados tanto do grupo das verdadeiras, como das sequências falsas.

O treinamento final da rede levou 12h seguidas de processamento, tempo estipulado para o EasyFan identificar a melhor relação de acertos/erros para o padrão inserido na rede.

3.7. GENOMAS

Todas as sequências de genomas utilizadas neste trabalho foram adquiridas no FTP do NCBI. O genoma da bactéria *Herbaspirillum seropedicae* está identificado sob o número de acesso NC_014323.

No início do desenvolvimento do programa eram utilizados dois arquivos por genoma. Um terceiro passou a ser utilizado na última versão do programa em substituição aos dois primeiros.

O primeiro arquivo é formado pelo cabeçalho e pela sequência completa do genoma estudado, está em formato *fasta* e é utilizado para a cópia das regiões intergênicas.

O segundo arquivo, em formato *multi-fasta*, contém as *ORFs* que estão dentro do genoma e um cabeçalho para cada uma das *ORFs*. identificando as regiões codificadoras.

O terceiro arquivo, em formato *Genbank* substitui os outros dois, pois contém todas as informações necessárias para a execução da aplicação em um único arquivo.

3.8. DESENVOLVIMENTO DA FERRAMENTA S54FINDER

Muitos algoritmos foram desenvolvidos para a execução deste trabalho sendo eles desde leitura dos arquivos que contém o genoma completo do organismo estudado, passando por algoritmos para a obtenção dos candidatos e chegando até o final com a seleção dos candidatos para treinamento e teste das redes neurais.

Para a obtenção da versão definitiva do algoritmo, muitas versões intermediárias foram desenvolvidas, modificando de uma versão para outra a forma de obtenção dos candidatos, quais arquivos seriam os de entrada do software e os

arquivos de manipulação de dados para a entrada da rede neural e sua forma de chamada.

Na primeira versão desenvolvida não foi apresentado um grande aprimoramento lógico para a realização das buscas pelos fatores sigma 54, apenas contando com a seleção de candidatos que estivessem no tamanho padronizado e com as bases mais conservadas presentes nos locais corretos.

Este padrão foi retirado da publicação sobre este fator de transcrição desenvolvido por BARRIOS et al (1999) com o seguinte padrão de conservação mrNrYTGGCACGnnnnTTGCWNNw. Deste, as bases mais conservadas e o tamanho total de quatorze pares de bases formaram a *string* que é procurada nas moléculas de DNA bacteriano.

A representação da Figura 7 demonstra como seria o promotor a ser buscado.

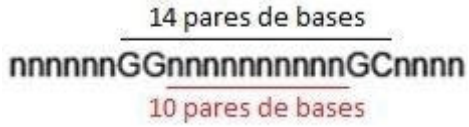


FIGURA 7– DISPOSIÇÃO DAS BASES CONSERVADAS NO PROMOTOR DO FATOR DE TRANSCRIÇÃO SIGMA 54
FONTE: O autor (2012)

A partir da segunda versão do algoritmo foram sendo realizadas várias alterações na lógica de programação para o melhoramento das respostas trazendo uma maior fidedignidade a resposta do algoritmo.

As alterações mais importantes foram sendo realizadas nas versões mais recentes com a adaptação de várias informações adquiridas em bibliografia, como mutações que possam ocorrer na região de ligação do fator sigma.

Dentre estas atualizações das versões foram desenvolvidas vinte e duas versões sem a utilização de redes neuronais artificiais e mais dezoito versões contendo esta última funcionalidade.

Para a validação dos resultados foram utilizadas algumas regiões promotoras

já publicadas em literatura (SOUZA et al, 1991; BARRIOS et al, 1999; SCHWAB et al, 2006), estas regiões servem para a confirmação do método de forma a validar as buscas. A tabela 11 lista as regiões com os respectivos genes regulados.

TABELA 11– REGIÕES DE LIGAÇÃO CONFIRMADAS BIOLOGICAMENTE

Gene	**Sitio de ligação do sigma 54**	**Referência**
nifA	GGGCATGAAGTTTGCT	SOUZA et al, 1991
nifB	TGGCACGGTTTTGGCT	SOUZA et al, 1991
nifH	TGGCACGGCATTTGCA	BARRIOS et al, 1999
glnA	TGGCATGGAATCTGCT	SCHWAB et al, 2006

FONTE: O autor (2012)

Para as mais variadas versões desenvolvidas algumas apresentam alterações importantes em relação às demais como na apresentação dos resultados, melhoria na seleção de candidatos e adição de redes neurais. Estas alterações são mostradas na tabela 12.

TABELA 12- RELACIONA AS VERSÕES DA APLICAÇÃO COM AS ALTERAÇÕES E MARCOS IMPORTANTES

Versão	**Informação**
2ªà5ª	Não apresenta resultados.
6ª a 12ª	Otimização do algoritmo de pré-seleção.
15ª	Divisão na análise em ambas as orientações do DNA.
16ª	Implementação da área de avaliação variável.
17ª	Adição do limite mínimo para a área de avaliação.
19ª	Alteração no algoritmo de pré-seleção tornando-o mais flexível.
22ª	Adição da rede neural artificial.
39ª	Adição da rede com melhor treinamento.
40ª	Implementação da seleção de número mínimo de bases conservadas e arquivos de entrada em formato GENBANK.

FONTE: O autor (2012)

Ao final da execução da aplicação o resultado é armazenado em quatro diferentes arquivos, estes listados na tabela 13.

TABELA 13- LISTA O FORMATO DOS ARQUIVOS E SEUS CONTEÚDOS

Formato	**Conteúdo**
1 em formato genbank (.gbk)	Posição dos candidatos classificados como verdadeiros pela rede neural.
2 em formato texto (.txt)	Nome das ORF´s respectivas as sequências classificadas como verdadeiras, um para cada orientação.
1 em formato matlab (.mat)	Armazena os dados da execução para posterior avaliação.

FONTE: O autor (2012)

4. RESULTADOS E DISCUSSÃO

A aplicação desenvolvida neste trabalho será apresentada e terá os seus resultados discutidos e comparados entre os resultados das versões construídas durante o desenvolvimento, além da comparação com outras aplicações que desempenham os mesmos tipos de buscas.

4.1. FLUXO DE EXECUÇÃO DO PROGRAMA DESENVOLVIDO

Os algoritmos desenvolvidos seguem uma ordem determinada de execução, com dados de entrada e saída para cada algoritmo, essa ordem é respeitada para um ótimo funcionamento do programa. Essa ordem de execução é demonstrada na Figura 8.

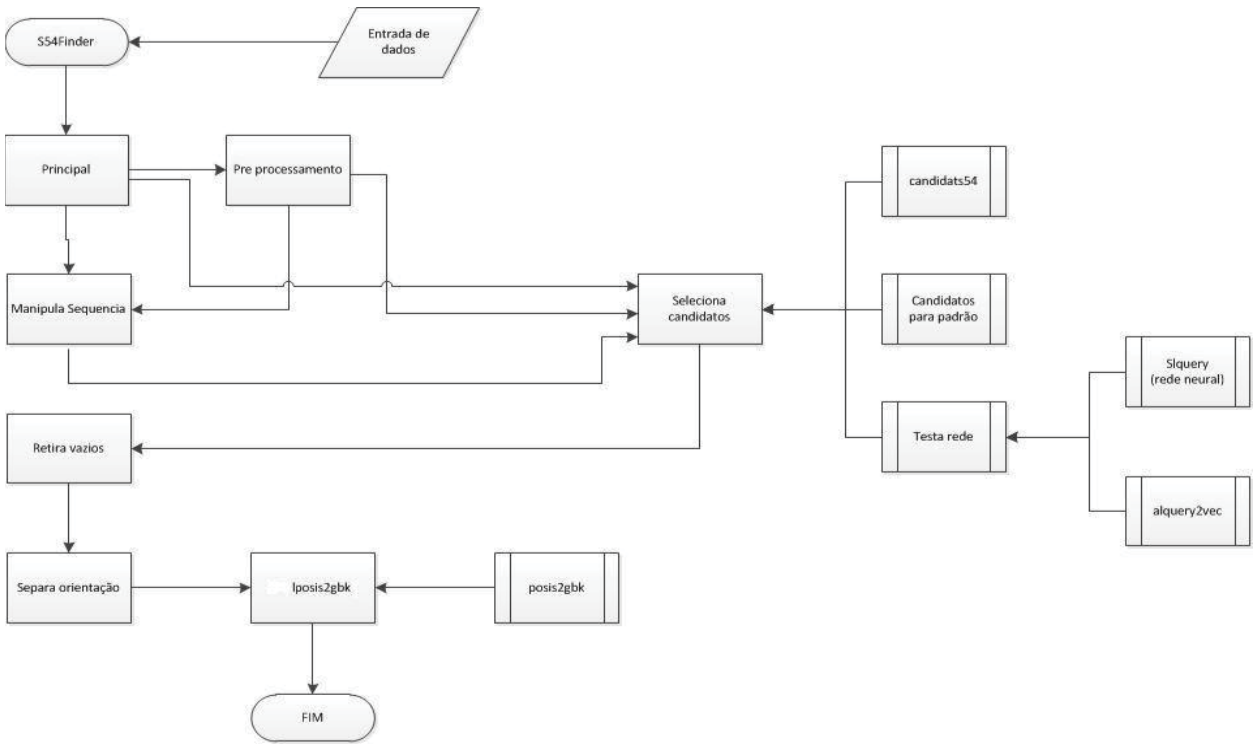


FIGURA 8- ILUSTRAÇÃO DO FLUXO DE EXECUÇÃO DO PROGRAMA DESENVOLVIDO
FONTE: O autor (2012)

Na próxima seção deste trabalho será explicado o funcionamento dos algoritmos e a função que cada um desempenha durante a execução da aplicação.

4.2. EXPLICAÇÃO DAS FUNÇÕES

Cada algoritmo desenvolvido para o programa desempenha uma função única dentro da execução do mesmo, a tabela 14 apresenta o nome e a função executada de cada um dos algoritmos desenvolvidos.

TABELA 14- DESCRIÇÃO DOS NOMES DAS FUNÇÕES E COM SUAS RESPECTIVAS FUNÇÃO NA APLICAÇÃO

Nome	Função
S54Finder	Interface do programa.
Principal	Função principal da execução do programa chama todas as outras funções.
Pré-processamento	Aquisição de posicionamento das ORF´s
Manipula sequencia	Substitui bases das ORF´s por letras “z”.
Seleciona candidatos	Seleciona a intergênica de cada ORF e seus candidatos.
Candidats54	Seleciona os candidatos dentro da intergênica já copiada pela função acima.
Candidatos para padrão	Padroniza os candidatos para o teste com a rede neural.
Testa rede	Envia os candidatos selecionados para a rede neural.
Slquery	Chama a rede neural e testa os candidatos.
Slquery2vec	Alinha os resultados da função anterior em vetor
Retira vazios	Retira ORF´s que não possuam candidatos em suas regiões intergênicas.
Separa orientação	Separa os candidatos de acordo com a orientação da sequência.
Lposis2gbk	Organiza os resultados para a criação do arquivo .GBK.
Posis2gbk	Gera o arquivo .GBK resultante a execução da aplicação.

FONTE: O autor (2012)

4.3. INTERFACE

Depois do desenvolvimento de todos os algoritmos que fazem parte do programa, a interface para o mesmo foi desenvolvida contendo as informações necessárias para a execução do programa, como demonstrada na Figura 9.

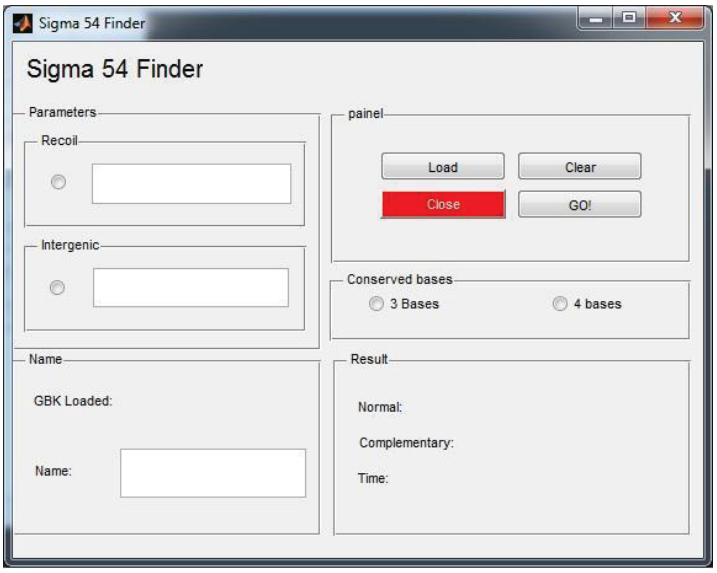


FIGURA 9 - FIGURA ILUSTRATIVA DA INTERFACE CRIADA PARA A UTILIZAÇÃO DO PROGRAMA
FONTE: O autor (2012)

A figura demonstra todas as informações que devem ser inseridas como a quantidade de bases à serem utilizadas para as buscas, a quantidade de bases conservadas que devem ser inseridas na busca entre outros parâmetros.

Explicando mais detalhadamente os parâmetros e opções a serem inseridos e o seu papel durante a execução da aplicação são demonstrado na tabela 15.

TABELA 15- RELAÇÃO ENTRE AS OPÇÕES DA INTERFACE E SUAS RESPECTIVAS FUNÇÕES PARA A EXECUÇÃO DA APLICAÇÃO

Opção	Função
Recuo (*Recoil*)	Recuo na primeira base a ser avaliada durante a busca, levando ao programa a buscar apenas a partir dela, utilizando uma distancia mínima entre a primeira base da ORF e a primeira base onde será efetuada a busca pelo candidato a fator sigma 54, possui o valor padrão de zero, não havendo assim recuo.
Intergênica (*Intergenic*)	Tamanho da região intergênica utilizada na busca, pode variar de acordo coma necessidade, aumentando a quantidade de bases gera uma maior quantidade de candidatos e assim elevando o tempo de execução.
GBK Carrergado (*GBK Loaded*)	Mostra qual arquivo GBK esta carregado para a execução do programa, somente mostra o arquivo quando o mesmo esta carregado.
Nome (*Name*)	Nome dado aos arquivos de resposta da execução da aplicação, sendo atribuídos aos arquivos .GBK e .MAT que contém as informações de saída da aplicação.
3 Bases e 4 Bases	Seleção entre a quantidade de bases conservadas que são consideradas na seleção de candidatos.
Normal (*Normal*)	Mostra a quantidade de candidatos a fator sigma 54 presentes na fita normal.
Complementar (*Complementary*)	Mostra a quantidade de candidatos a fator sigma 54 presentes na fita complementar.
Tempo (*Time*)	Tempo de execução da aplicação, medido em minutos.
Carregar (*Load*)	Cerraga o arquivo GBK para a execução da aplicação.
Limpar (*Clear*)	Limpa as informações contidas nos campos, preparando pra uma nova execução.
Vai (*GO*)	Inicia a execução do programa, só inicia após a inserção de todos os parâmetros, evitando assim erros durante a execução.
Fechar (*Close*)	Fecha a aplicação.

FONTE: O autor (2012)

4.4. PRÉ-SELEÇÃO DE CANDIDATOS

Para a primeira versão desenvolvida apenas para as regiões intergênicas de alguns genes selecionados da bactéria *Herbaspirillum seropedicae*, 32 no total. Foram encontradas um total de 108 sequências candidatas das quais poucas eram de fato regiões promotoras.

Neste primeiro teste apenas uma pequena parte do genoma bacteriano foi coberta, justificando um teste completo envolvendo o genoma completo da bactéria resultando no número de candidatos apresentados na tabela 16.

TABELA 16- NÚMERO DE CANDIDATOS A REGIÕES DE LIGAÇÃO DO FATOR SIGMA 54 RESULTANTE DA EXECUÇÃO A PRIMEIRA VERSÃO DA APLICAÇÃO UTILIZANDO O GENOMA COMPLETO DA BACTÉRIA *Herbaspirillum seropedicae*

Orientação	**Nº de Candidatos**
Normal	61191
Complementar	60552

FONTE: O autor (2012)

Com a evolução do trabalho várias versões para a aplicação foram sendo desenvolvidas com alguns aprimoramentos em várias partes gerando uma grande quantidade de diferentes resultados com uma variação no número de candidatos a regiões de ligação do fator de transcrição. A comparação entre os resultados é apresentada na tabela 17.

TABELA 17- COMPARAÇÃO ENTRE O NÚMERO DE CANDIDATOS A REGIÕES DE LIGAÇÃO DO FATOR SIGMA 54 ENTRE ALGUMAS VERSÕES DO PROGRAMA, UTILIZANDO O GENOMA COMPLETO DA BACTÉRIA *Herbaspirillum seropedicae*

Versão	**Nº de candidatos**
11ª	3045
12ª	1418
13ª	2709
19ª	1959

FONTE: O autor (2012)

Após algumas outras alterações no algoritmo passando a ter uma variação no tamanho da região intergênica que é avaliada pelo algoritmo de pré-seleção de candidatos, mostrando o resultado na Tabela 18. Ainda com a confirmação da presença de algumas regiões intergênicas confirmadas biologicamente.

TABELA 18- COMPARAÇÃO DA QUANTIDADE DE CANDIDATOS PARA CADA TESTE COM A VARIAÇÃO DO TAMANHO DA REGIÃO INTERGÊNICA AVALIADA COM A CONFIRMAÇÃO DA PRESENÇA DE REGIÕES PROMOTORAS QUE POSSUEM CONFIRMAÇÃO BIOLÓGICA

Tam.	Normal	Complement.	*nifA*	*nifB*	*nifH*	*glnA*
50	212	176	-	-	-	-
100	811	746	-	x	-	-
125	1093	993	-	x	-	-
150	1324	1238	-	x	-	-
175	1536	1445	x	x	-	x
200	1790	1682	x	x	x	x
250	2225	2091	x	x	x	x
300	2680	2498	x	x	x	x
400	3545	3376	x	x	x	x

FONTE: O autor (2012)

Além da variação da região intergênica avaliada um limite mínimo foi implementado para evitar a pré-seleção de falsos positivos pelo algoritmo de pré-seleção, gerando uma nova tabela 19 ainda apresentando as regiões confirmadas biologicamente.

TABELA 19- COMPARAÇÃO DA QUANTIDADE DE CANDIDATOS COM A VARIAÇÃO DO TAMANHO DA REGIÃO INTERGÊNICA AVALIADA E LIMITE MÍNIMO DE DISTÂNCIA PARA AS ORFS E CONFIRMAÇÃO DA PRESENÇA DE REGIÕES PROMOTORAS QUE POSSUEM CONFIRMAÇÃO BIOLÓGICA

Tam.	L.C.	Normal	Comple.	nifA	nifB	nifH	glnA
200	-	1790	1682	X	x	X	x
200	10	1685	1646	X	X	X	X
200	25	1536	1519	X	X	X	X
200	50	1324	1233	X	-	X	X
250	-	2225	2091	X	X	X	X
250	10	2147	2055	X	X	X	X
250	25	2005	1928	X	X	X	X
250	50	1790	1642	x	-	x	X

FONTE: O autor (2012)

4.5. ADIÇÃO DE REDES NEURONAIS ARTIFICIAIS

O grande número de candidatos à região de ligação do fator de transcrição sigma 54 comprova que somente o algoritmo de pré-seleção de candidatos não foi capaz de classificar como verdadeiros ou falsos os candidatos pré-selecionados.

Para cumprir essa função de classificação foram adicionadas ao algoritmo as redes neurais artificiais para o contorno desta limitação encontrada.

A tabela 20 lista alguns treinamentos da rede neuronal FAN.

TABELA 20- COMPARAÇÃO ENTRE OS TREINAMENTOS PARA A ESCOLHA DA MAIS APROPRIADA PARA INCORPORAÇÃO À EXECUÇÃO DA APLICAÇÃO

Teste	Verdadeiros Positivos	Verdadeiros Negativos	Falsos Positivos	Falsos Negativos	Acertos de Verdadeiros	Acertos de Falsos
1	7795	16021	169	300	97,88%	98,16%
2	7768	16037	153	327	98,07%	98,00%
3	7809	16022	168	286	97,89%	98,25%
4	7786	16024	166	307	97,91%	98,12%

FONTE: O autor (2012)

Após a escolha da rede neuronal que apresenta um melhor comportamento frente aos candidatos pré-selecionados, novos testes foram realizados para a comparação com as versões da aplicação desenvolvidas anteriormente. A tabela 21 apresenta uma comparação do número de candidatos encontrados por algumas versões da aplicação com e sem o uso de redes neuronais.

Algumas regiões promotoras sofreram algumas mutações assim não possuindo uma das quatro bases mais conservadas. Para que o algoritmo de pré-seleção considerasse estas regiões como sendo possíveis candidatos a regiões de ligação para o fator sigma 54. Passou a ser adotada uma variação entre três ou quatro bases gerando um novo conjunto de resultados cujos são apresentados na tabela 21.

TABELA 21- COMPARAÇÃO DO NÚMERO DE CANDIDATOS ENCONTRADOS ANTES E APÓS ADOÇÃO DE REDES NEURAIS ARTIFICIAIS. COM A ADIÇÃO DE SELEÇÃO DO NÚMERO DE BASES CONSERVADAS

	Fita normal	**Fita complementar**
1ª versão	61191	60552
19ª versão	976	983
38ª sem rede neural	1628	1520
38ª com rede neural	219	213
39ª sem rede neural	1628	1520
40ª com 3 bases conservadas e rede neural	166	115
40ª com 4 bases conservadas e rede neural	74	52

NOTA: Sequências e nomes das sequências candidatas a fator de transcrição sigma 54 da 40ª versão com 4 bases conservadas e rede neural estão listadas nos apêndices 1 e 2.
FONTE: O autor (2012)

4.6. COMPARAÇÃO COM OUTRAS APLICAÇÕES

Outras aplicações disponíveis e já publicadas realizam buscas em genomas do mesmo padrão de conservação das regiões de ligação ao fator de transcrição sigma 54, então a comparação com estes métodos serve para a validação do método desenvolvido neste trabalho.

A primeira ferramenta a ser comparada é o *GenomeMatScan*, desenvolvida por Cases (2003) e utiliza de modelos ocultos de Markov como motor de suas buscas.

Testes empregando o genoma bacteriano do *Herbaspirillum seropedicae* no *GenomeMatScan* são apresentados nas tabelas 22 e 23 exibindo as sequências candidatas identificadas para a região compreendida da região intergênica do *nifS* até a região intergênica do *nifA*.

TABELA 22- SEQUÊNCIAS IDENTIFICADAS PELO GENOMEMATSCAN COMO REGIÕES PROMOTORAS DO TRECHO DA FITA NORMAL DE HERBASPIRILLUM SEROPEDICAE DA REGIÃO INTERGÊNICA DO NIFS ATE A REGIÃO INTERGÊNICA DO NIFA

Sequências	**Gene**
TGGCATGCTCTTTGCT	*nifS*
TGGCACATGAATTGCT	*nifV*
TGGTCCAACCCTTGCA	Não identificada.*
TGGGATAGACCTTGCA	Não identificada.**

* - Região identificada dentro da região codificadora do gene *fdxB*.
** - Não existe região codificadora próxima.
FONTE: O autor (2012)

TABELA 23- SEQUÊNCIAS IDENTIFICADAS PELO GENOMEMATSCAN COMO REGIÕES PROMOTORAS DO TRECHO DA FITA COMPLEMENTAR DE HERBASPIRILLUM SEROPEDICAE DA REGIÃO INTERGÊNICA DO NIFS ATE A REGIÃO INTERGÊNICA DO NIFA

Sequências	**Gene**
TGGCATGGTTTTGGCT	*Hsero_2872*
GGGCATGAAGTTTGCT	*nifA*
TGGCACGGTTTTGGCT	*nifB*
TGGTGCGCATGTTGCA	Não identificado.*
TGGCCAGTGAATTGCA	Não identificado.**
TGGTACGCGCCATGCA	Não identificado.***
TGGCACGGCATTTGCA	*nifH*
TGGCCCTCGTATTGCT	Não identificado.****
TGGGATAGACCTTGCA	Não identificado.*****

* - Região identificada dentro da região codificadora do gene *Hsreo_2867*.
** - Região identificada dentro da região codificadora do gene *Hsreo_2860*.
*** - Região identificada dentro da região codificadora do gene *Hsreo_2855*.
**** - Região identificada dentro da região codificadora do gene *nifD*.
***** - Região identificada dentro da região codificadora do gene *nifK*.
FONTE: O autor (2012)

Para a mesma região o teste utilizando o S54Finder foi realizado apresentando o resultado na tabela 24.

TABELA 24- SEQUÊNCIAS IDENTIFICADAS PELO S54FINDER COMO REGIÕES PROMOTORAS DO TRECHO DE AMBAS AS FITAS DE HERBASPIRILLUM SEROPEDICAE DA REGIÃO INTERGÊNICA DO NIFS ATE A REGIÃO INTERGÊNICA DO NIFA. COM SUAS RESPECTIVAS ORIENTAÇÕES

Sequências	**Gene**	**Orientação**
TGGCATGCTCTTTGCTG	*nifS*	Normal
TGGCACATGAATTGCTA	*nifV*	Normal
GTGGCACGGCATTTGCA	*nifH*	Complementar
GGGAACGATGTTTGCTA	*Hsero_2854*	Normal
CTGGCACGGTTTTGGCT	*nifB*	Complementar
TGGGCATGAAGTTTGCT	*nifA*	Complementar
TGGCATGGTTTTGGCTA	*Hsero_2872*	Normal

FONTE: O autor (2012)

O resumo comparativo entre os resultados do *GenomeMatScan* e *S54Finder* é apresentado na tabela 25.

TABELA 25- RESUMO COMPARATIVO DAS EXECUÇÕES DAS APLICAÇÕES PARA A REGIÃO COMPREENDIDA ENTRE A REGIÃO INTERGÊNICA DO NIFS E A REGIÃO INTERGÊNICA DO NIFA, CONSIDERANDO O TIPO DO RESULTADO ENCONTRADO

	GenomeMatScan	**S54Finder**
Verdadeiros	5	5
Falsos	7	0
Hipotéticos	1*	2*
Total	13	7

* - Carece de confirmação biológica. FONTE: O autor (2012)

O segundo programa a ser utilizado para comparações foi o PromScan desenvolvido por Studholme (2000), que também usa HMM.

Os testes com este programa não foram realizados, pois a execução do mesmo em ambiente web retorna um erro de arquivo não encontrado e quando executado localmente o mesmo erro foi apresentado.

A tabela 26 apresenta um comparativo entre os programas com suas principais características.

TABELA 26 - COMPARATIVO DAS PRINCIPAIS CARACTERÍSTICAS DOS APLICATIVOS DE BUSCA PARA FATORES SIGMA 54 AVALIADOS

	S54Finder	**GenomeMatScan**	**PromScan**
Versão Web	Não*	Sim	Sim**
Processamento Local	Sim	Não	Sim**
Método	FAN	HMM	HMM
Forma de entrada	Arquivo em formato GenBank (.gbk)	Bases copiadas diretamente no navegador	Arquivo em formato Fasta e ptt
Resultado	Saidas em formatos Genbank (.gbk), MatLab (.mat)e Texto	Mostra no navegador	Mostra no navegador
Processamento em grandes regiões de DNA.	Sim	Não	?***

* - Esta prevista a Implementação da aplicação na Web como um trabalho futuro.
** - É necessário que haja um compilador Perl instalado.
*** - Teste não pode ser executado.
FONTE: O autor (2012)

Em complemento à tabela que faz a comparação entre os programas às próximas tabelas 27, 28 e 29 apresentam os prós e contras da execução de cada aplicação.

TABELA 27- PRÓS E CONTRAS PARA O PROGRAMA S54FINDER

S54Finder	
Prós	
	Altas taxas de acerto de regiões promotoras.
	Baixa taxa de falsos positivos.
	De fácil utilização.
	Interface amigável.
	Versão disponível para uso local.
	Não limitado a pequenas regiões de DNA.
	Uso para genomas completos.
	Flexibilidade das opções de busca.
Contras	
	Ainda não publicado.
	Não dispõe de versão web.

FONTE: O autor (2012)

TABELA 28- PRÓS E CONTRAS PARA O PROGRAMA GENOMEMATSCAN

GenomeMatScan	
Prós	Versão disponível via Web.
	Interface simples.
Contras	
	Resultado de difícil interpretação.
	Não roda para genomas completos.
	Não realiza buscas em grandes regiões de DNA.
	Poucas opções de busca.

FONTE: O autor (2012)

TABELA 29- PRÓS E CONTRAS PARA O PROGRAMA PROMSCAN

PromScan	
Prós	
	Não pode ser avaliado.
Contra	
	Uso desaconselhado pelo autor.

5. CONCLUSÕES

A ferramenta de busca de sequências de ligação ao DNA de fatores de transcrição sigma 54 desenvolvido neste trabalho, *S54Finder*, apresentou resultados convincentes quanto às pesquisas realizadas em *Herbaspirillum seropedicae* e em outros genomas bacterianos.

A pré-seleção de candidatos sem a utilização de redes neuronais mostrou que com o padrão de conservação, baseados em sequências confirmadas biologicamente e publicadas, é capaz de localizar os sítios de ligação em genomas bacterianos, mas gerando uma grande quantidade de falsos positivos.

O número de candidatos a sequências de ligação ao DNA de fatores sigma 54 com a utilização de redes neuronais artificiais foi reduzido de 3148 sequências candidatas para 126 sequências candidatas prováveis em *Herbaspirillum seropedicae*.

O *S54Finder* se mostrou uma ferramenta eficiente para a identificação de sequências de ligação ao DNA de fatores de transcrição sigma 54, mas deve passar por outras fases de desenvolvimento para implementação de novas funcionalidades como o acesso via web e de documentação para facilitar o seu entendimento e uso, dentre outras futuras necessidades.

6. REFERÊNCIAS

1. ARAGUAIA, M. Bioinformática, 2011. Extraído http://www.brasilescola.com/biologia/bioinformatica.htm. Último acesso em: 22/11/2011.

2. BALDANI, J.L.; BALDANI, V.L.D.; OLIVARES, F.; DÖBEREINER, J.; Identification and ecology of Herbaspirillum seropedicae and closely related Pseudomonas rubrisubalbicans. Symbiosis, v. 13, p. 65-73, 1992.

3. BALDANI, J.L.; BALDANI, V.L.D.; SELDIN, L.; DÖBEREINER, J. Characterization of Herbaspirillum seropedicae gen. Nov., sp. Nov., a new rootassociated nitrogen-fixing bacterium. International Journal of systematic Bacteriology, v. 36, p. 86-93, 1986.

4. BALDANI, J. I.; CARUSO, L.; BALDANI, V. L. D.; GOI, S.; DÖBEREINER, J. Recent advances in BNF with non-legume plants. Soil Biology and Biochemistry, v.29, p. 911-922, 1997.

5. BARRIOS, H.; VALDERRAMA, B. e MORETT, E. Compilation and analysis of s54-dependent promoter sequences. Nucleic Acids Res. v.27, p. 4305-4313, 1999.

6. BAYAT, A. Science, medicine, and the future Bioinformatics. BMJ, v. 324 p. 1018–22, 2002.

7. BURRIS, R.H. Nitrogenases. The Journal of Biological Chemistry, v. 266, n. 15, p. 9339-9342, 1991.

8. CASES, I.; USSERY, D. W.; LORENZO, V. The s^{54} regulon (sigmulon) of *Pseudomonas putida* Envoirimental Microbiology v.5, p.1281-1293, 2003.

9. CONWAY,K. s^{54} Promoter database. Disponível em: http://www.sigma54.ca/PromoterApp/Web/data.aspx. Último acesso em: 16/02/2012.

10. COELHO, L. Dos S. RAITTZ, R. T. TREZUB, M. FControl: sistema inteligente inovador para detecção de fraudes em operações de comércio eletrônico. Gest. Prod. São Carlos, v. 13, n. 1, 2006.

11. DAYHOFF, M.O. Computer analysis of protein evolution. Sci Am., v. 221(1), p.86-95, 1969.

12. DING, L. YOKOTA, A. Proposals of Curvibacter gracilis gen. nov., sp. nov. and Herbaspirillum putei sp. nov. for bacterial strains isolated from well water and reclassification of [Pseudomonas] huttiensis, [Pseudomonas] lanceolata, [Aquaspirillum] delicatum and [Aquaspirillum] autotrophicum as Herbaspirillum huttiense comb. nov., Curvibacter lanceolatus comb. nov., Curvibacter delicatus comb. nov. and Herbaspirillum autotrophicum comb. nov. International Journal of Systematic and Evolutionary Microbiology, v. 54, p. 2223–2230, 2004.

13. DIXON, R.; KAHN, D. Geneticregulationof biological nitrogen fixation. Nat. Rev. Microbiol., v. 2, p. 621-631, 2004.

14. DÖBEREINER, J. Recent changes in concepts of plant bacteria interactions: Endophytic N2 fixing bactéria. Ciência e Cultura, São Paulo, v. 44, n. 5, p. 310-313, 1992.

15. DOUCLEFF, M.; MALAK, L. T.; PELTON, J. G. e WEMMER, D. E. The C-terminal RpoN domainof s54 forms an unpredicted helix-turn-helix motif similar to domains s70. J. Biol. Chem. v. 208, p. 41530-41536, 2005.

16. EVANS, H. J.; BURRIS R. H.; Highlights in biological nitrogen fixation during the last 50 years. Biological Nitrogen Fixation, p. 1-92, 1992.

17. FAUSETT, L. Fundamentals of Neural Networks Prentice Hal , Englewood, New Jersey. 1994, 461p.

18. FOUSSARD, M.; CABANTOUS, S.; PÉDELACQ, J-D.; GUILLERT, V.; TRAINER, S.; MOUREY, L.; BRICK, C; SAMAMA, J-P. The molecular puzzle of two-component signaling cascades. Microbes and Infection, v. 3, p. 417-424, 2001.

19. FOX, J. What is Bioinformatics? The Science Creative Quarterly, Issue Four, 2009. Extraído de: http://www.scq.ubc.ca/what-is-bioinformatics/. Último acesso em: 21/11/2011.

20. GALPERIN, M. Bacterist signal transduction network in a genomic pernspective. Environ Microbiol, v. 6, p. 552-567, 2004.

21. GEHLEN, M. A. C. ESTUDO E LEVANTAMENTO DOS GENES *NIF* PUBLICADOS NO NCBI USANDO CONCEITOS DE MINERAÇÃO DE DADOS E INTELIGÊNCIA ARTIFICIAL. Dissertação de mestrado, Universidade federal do Paraná, 2011.

22. GENETICA ON LINE. GENÉTICA MOLECULAR – AULA 3. Dinsponível em: http://aprendendogenetica.blogspot.com/2011/03/genetica-molecular-aula-3-transcricao.html. Último acesso em: 16/02/2012.

23. GRALLA, J. D. Methods Enzymol. V. 185, p. 37-54, 1990.

24. GRAVES, M. C. & RABINOWITZ, J. C. *In vivo* and *in vitro* transcription of the *Clostridium pasteurianum* ferrodoxin gene. Evidence for “extended” promoter elements in gram-positive organisms. J. Biol. Chem. v. 261, p. 11409-11415, 1986.

25. HARLEY, C. B. & REYNOLDS, R. P. Analysis of *E. coli* promoter sequences. Nucleic Acids Res. v.15, p. 2343-2361, 1987.

26. HAYKIN, S. Neural Networks – A Comprehensive Foundation, Macmillan College Publishing Inc., 1994.

27. HELMANN, J. D. Compilation and analysis of *Bascillus subtilis* A-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA. Nucleic Acids Res. v.23 p. 2351-2360, 1995.

28. IM, W. T. BAE, H. S. YOKOTA, A. LEE, S. T. Herbaspirillum chlorophenolicum sp. nov., a 4-chlorophenol-degrading bacterium. International Journal of Systematic and Evolutionary Microbiology. v. 54, p. 851-855, 2004

29. ISHIHAMA, A. Molecular Assembly and Functional Modulation of *Escherichia coli* RNA Polymerase. Adv. Biophys. v.26, p.19-31, 1990.

30. HAYKIN, S. Neural Networks – A Compreensive Foundation. Prentice-Hall, New Jersey 2nd Edition, 1999.

31. HAYKIN, S. Redes Neurais Princípios e Prática. Tradução de: Paulo Martins Engel. Porto
Alegre: Bookman, 2001.

32. HOWARD, J.B.; REES, D.C. Structural basis of biological nitrogen fixation. Chem. Rev., v. 96, p. 2965-2982, 1996.

33. HUNGRIA, M.; CAMPO, R.J. Fixação biológica do nitrogênio em sistemas agrícolas. In: Congresso Brasileiro de Ciência do Solo, 30, 2005, Recife. Anais Recife: SBS, 2005.

34. KIRCHHOF, G. ECKERT, B. STOFFELS, M. BALDANI, J. I. REIS, V. M. E HARTMANN, A. *Herbaspirillum frisingense* sp. nov., a new nitrogen-fixing bacterial species that occurs in C4-fibre plants. Int J Syst Evol Microbiol v. 1, p. 157-168, 2001.

35. KUMAR, S. A. The Structure and Mechanism of Action of Bacterial DNA-Dependent RNA Polymerase. Prog. Byophis. Molec. Biol. v.38, p-163-201, 1981.

36. LEANG, C. KRUSHKAL, J. UEKI, T. PULJIC, M. SUN, J. JUÁREZ, K. NÚÑEZ, C. REGUERA, G. DIDONATO, R. POSTIER, B. ADKINS R. M. LOVLEY D. R. Genome-wide analysis of the RpoN regulon in *Geobacter sulfurreducens*. BMC Genomics, 2009.

37. LONETTO, M.; GRIBSKOV, M. & GROSS C. A. The s70 family: sequence conservation and evolutionary relationships. J. Bacteriol, v. 174, p. 3843-3849, 1992.

38. MATHWORKS. Matlab – The language of technical computing. DisponÍvel em: http://www.mathworks.com/products/matlab/. Último acesso em: 20/12/2011.

39. MC CLURE, W. R. Mechanism and Control of Transcription initiation in Prokaryotes. Ann. Rev. Biochem. v. 54 p. 171-204, 1985.

40. MERRICK, M.J.; EDWARDS, R.A. Nitrogen control in bacteria. Microbiol. Rev., v. 59, p. 604-622, 1995.

41. MICROSOFT, Recursos e benefícios do Excel. Disponível em: http://office.microsoft.com/pt-br/excel/recursos-e-beneficios-do-excel-2010-HA101806958.aspx. Último acesso em: 20/12/2011.

42. MOONEY, R. A; DARST, S. A e LANDICK R. Sigma and RNA Polymerase: An On-Again, Off-Again Relationship? Molec. Cell v. 20, p. 335-345, 2005.

43. MOREIRA, F.M.S.; SIQUEIRA, J.O. Microbiologia e bioquímica do solo. Lavras: Universidade Federal de Lavras, 2006. 729p.

44. MORETT, E. & BUCK, M. *In vivo* studies on the interaction of RNA polymerase-σ^{54} with the *Klebsiella pneumoniae* and *Rhizobium meliloti niH* promoters. J. Mol. Biol. V.210, p. 65-77, 1989.

45. MORETT, E. & SEGOVIA, L. J. Bacteriol, v.175, p. 6067-6074, 1993.

46. NELSON, David L.; COX, Michael M. Lehninger Principles of Biochemistry. 5. ed. W. H. Freeman, 2008, 1100 p.

47. NCBI. A science primer - just the facts: a basic introduction to the science underlying NCBI Resources. Disponível em: <http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html>, Último acesso: 21/11/2011.

48. NIXON, B.T.; RONSON, C.W.; AUSUBEL, F.M. Two-component regulatory systems responsive to environmental stimuli share strongly conserved domains with the nitrogen assimilation regulatory genes ntrB and ntrC . Proc. Natl. Acad. Sci. USA, v. 83, p. 7850-7854, 1986.

49. OLIVARES, F. L.; BALDANI, V. L. D.; REIS, V. M.; BALDANI, J. I.; DÖBEREINER, J. Occurrence of endophytic diazotrophic Herbaspirillum spp in roots, stems and leaves predominantly of gramineae. Biology and Fertility of Soils, Berlin, v. 21, n. 3, p. 197-200,1996.

50. PARKINSON, J.S. Signal transduction schemes of bacteria. Cell, v. 73, p. 857-871, 1993.

51. PEDROSA, F.O. Sequenciada bacteria que produz "adubo natural" Genoma mapeado por cientistas paranaenses trara economia de milhoes para a agricultura brasileira. Ciencia Hoje On-line. Disponivel em:<http://cienciahoje.uol.com.br/controlPanel/materia/view/3234>, 2005

52. POPHAM, D. L.; SZETO, D.; KEENER, J. e KUSTO, S. Function of a bacterial activator protein that binds to transcriptional enhancers. Science v. 243, p. 629-635, 1989.

53. POSTGATE, J.R. Biological nitrogen fixation: Fundamental. Phil. Trans. R. Soc. Lond., v. 296, p. 375-85, 1982.

54. POTVIN, E.; SANSCHAGRIN, F. & LEVESQUE, R. C. Sigma factors in *Pseudomonas aeruginosa.* FEMS Microbiol Rev. v. 32, p. 38-55, 2007.

55. RABINER, L. R.; JUANG B. H. An Introduction to hidden Markov Models, IEEE ASSAP Magazine, p. 4-15, 1986.

56. RAITTZ, R. T. SOUZA, J. A. DANDOLINI, G. A. PACHECO, R. C. S. MARTINS, A. GAUTHIER, F. e BARCIA, R. Learning by means of free associative neurons. NAFIPS 97: Annual Meeting of the North American Fuzzy Information Processing Society, September 21-24, Syracuse New York, 1997.

57. RAITTZ, R. T. SOUZA, J. A. DANDOLINI, G. A. et al. FAN: Learning by means of free associative neurons. WCCI98 – IEEE World Congress on Computational Intelligence. p. 425-430 Anchorage, Alaska, 1998.

58. RAMOS, J. R. L. S. Análises Moleculares Comparativas de Estirpes de Herbaspirillum por PFGE, RAPD, RFLP e Sequenciamento do gene que codifica o 16SrRNA. Universidade Federal do Paraná, Tese de Doutorado, 2003.

59. REINHOLD-HUREK, B.; HUREK, T. Reassessment of the taxonomic structure of the diazotrophic genus Azoarcus sensu lato and description of three new genera and new
species, Azovibrio restrictus gen. nov., sp. nov., Azospira oryzae gen. nov., sp. nov. and Azonexus fungiphilus gen. nov., sp. nov. Int. J. Syst. Evol. Microbiol., v. 50, p. 649-659, 2000.

60. RUTHERFORD, K. PARKHILL, J. CROOK, J. HORSNELL, T. RICE, P. RAJANDREAM, M. BARRELL, B. Artemis: Sequence visualization and annotation. Bioinformatics Aplications Note. v. 16, n. 10, p. 944-945, 2000.

61. SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. Bioinformatics, v. 23, n. 19, p 2507–2517, 2007.

62. SASSE-QWIGHT, S. & GRALLA, J. D. Probing the *Eschirichia coli* glnALG upstreamactivation mechanism *in vivo.* Proc. Natl. Acad. Sci. USA V.85, p. 8934-8938, 1988.

63. SCHWAB, S. SOUZA, E. M. YATES, M. G. PERSUHN, D. C. STEFFENS, M. B. R. CHUBATSU, L. S. PEDROSA, F. O. RIGO, L. U. The glnAntrBC operon of *Herbaspirillum seropedicae* is transcribed by two oppositely regulated promoters upstream of *glnA*. Can. J. Microbiol. v. 53, p. 100-105. 2007.

64. SOURCEFORGE, EasyFan, Disponívem em: http://easyfan.sourceforge.net/. Último acesso
em: 16/02/2012.

65. SOUZA, E. M. FUNAYAMA, S. RIGO, L. U. YATES, M. G. PEDROSA, F O. Sequence and structural organization of a nifA-like gene and part of a nifB-like gene of *Herbaspirillurn seropedicae* strain 278. Journal of General microbiology. v. 137, p. 1511-1522. 1991.

66. SOUZA, J.A. Reconhecimento de padrões usando indexação recursiva. Tese de Doutorado, Universidade Federal de Santa Catarina, 1999.

67. SEWELL, M. Feature Selection. 2007. Disponível em <http://machine-learning.martinsewell.com/feature-selection/>. Último acesso: 05/01/2011.

68. STUDHOLM, D. J. BUCK, M. NIXON, B. T. Identification of potential s^n-dependent promoters in bacterial genomes. Microbiology comments, 2000.

69. STOCK, A. M.; ROBINSON, V. L. & GOUDREAU, P. N. Two-component signal transduction Annu. Rev. Biochem. v. 69, p. 183-215, 2000.

70. STOCK, A. NINFA, A. J. STOCK A. M. Protein phosposrylation and regulation of adaptive responses in bacteria . Microbiol. Rev. v.53, p. 450-490, 1989.

71. URETA, A; ALVAREZ, B.RÁMON, A.; VERA, M. A.; MARTINÉZ-DRETS, G. Identification of Acetobacter diazotrophicus, Herbaspirillum seropedicae and Herbaspirillum rubrisubalbicans using biochemichal and genetic criteria. Plant and Soil., v. 127, p. 271-277, 1995.

72. VALVERDE, S. R. et al. Efeitos multiplicadores da economia florestal brasileira. Revista Árvore, v. 27, n. 3, p. 285-293, 2003.

73. VOGT, C. Bioinformática, genes e inovação, 2003. Extraído de: http://www.comciencia.br/reportagens/bioinformatica/bio01.shtml. Último acesso em: 23/11/2011.

74. VON HIPPEL, P. H.; BEAR, D. G.; MORGAN, W. D. e MC SWIGGEN Protein-Nucleic Acid Interactions in transcription: A molecular Analysis. Ann. Rev. Biochem.. v. 53, p. 389-446, 1984.

75. WEST, A; STOCK, A., Histidine kinases and response regulator proteins in two-component signalling systems. Trends Biochem. Sci., v. 26, p. 369-376, 2001.

76. WHITE, D. The Physiology and Biochemistry of Prokaryotes Oxford University Press, Segunda edição, 2000.

77. WÖSTEN, M.M. Eubacterial Sigma-factors, FEMS Microbiol. Rev. v. 22, p.127-150, 1998.

78. YOUNG, J.P.W. Phylogenetic classification of nitrogen-fixing organisms, Em: STACEY, G.; BURRIS, H.; EVANS, H. J. (ed), Biological nitrogen fixation. Chapman & Hall, New York, N.Y., 1992. 943p.

APÊNDICES

APÊNDICE 1 – LISTA COM SEQUÊNCIAS E NOMES DOS CANDIDATOS A FATOR DE TRANSCRIÇÃO SIGMA 54 DA 40ª VERSÃO COM 4 BASES DE CONSERVAÇÃO E REDE NEURONAL. NA ORIENTAÇÃO NORMAL

Sequências	Nomes dos candidatos a fator de transcrição sigma 54
TGGCACGGCTCCTGCT	membrane protein [Herbaspirillumseropedicae SmR1]
TGGTCTGAAAATTGCT	AraC family transcription regulator protein [Herbaspirillumseropedicae SmR1]
TGGCGCAAATCCTGCT	ABC-type branched-chain amino acid transport system, periplasmic component protein [Herbaspirillumseropedicae SmR1]
CGGCGCGGGTTTTGCC	ATP-dependent protease HslVU (ClpYQ), ATPase subunit heat shock protein [Herbaspirillumseropedicae SmR1]
TGGTGCAGCCCTTGCT	hypothetical protein Hsero_0218 [Herbaspirillumseropedicae SmR1]
CGGCGCGGGTCTTGCC	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
TGGCGCAGATATTGCA	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
CGGCATGTTGTTTGCT	heat-shock Hsp70 protein [Herbaspirillumseropedicae SmR1]
TGGCGCGATGATCGCA	ABC-type transport system, permease component protein [Herbaspirillumseropedicae SmR1]
AGGCATGGCTCTGGCA	transmembrane protein [Herbaspirillumseropedicae SmR1]
TGGCACGCCGGCCGCG	phenylpropionatedioxygenase (Rieske 2Fe-2S family) protein [Herbaspirillum seropedicae SmR1]
TGGCCCGCCCAATGCA	phosphomannomutaseprotein [Herbaspirillum seropedicae SmR1]
TGGCAAAGTGGTTGCA	permease of the major facilitator superfamily protein [Herbaspirillumseropedicae SmR1]
AGGCATCATTTTTGCT	molecular chaperone protein [Herbaspirillumseropedicae SmR1]
TGGCAGGGCGATCGCA	phosphoribosylformylglycinamidinecyclo-ligase (AirS) protein [Herbaspirillum seropedicae SmR1]
TGGCACGACAATTGCA	poly-beta-hydroxyalkanoatedepolymerase protein [Herbaspirillum seropedicae SmR1]
TGGCAAGCCGGTGGCA	SAM-dependent methyltransferaseprotein [Herbaspirillum seropedicae SmR1]
TGGAACTGCCTTCGCA	type III secretion HrpQ protein [Herbaspirillumseropedicae SmR1]
GGGGACGGCGTTTGCC	hypothetical protein Hsero_0806 [Herbaspirillumseropedicae SmR1]
TGGCACGTTCTTCGCT	NADH-dependent FMN reductase oxidorreductaseprotein [Herbaspirillumseropedicae SmR1]
TGGCACTCGATTTGCT	hypothetical protein Hsero_1015 [Herbaspirillumseropedicae SmR1]
TGGTATCAAGCTTGCA	ABC-type dipeptide transporter, periplasmic peptide-bindingprotein [Herbaspirillumseropedicae SmR1]

Sequências	Nomes dos candidatos a fator de transcrição sigma 54
TGGCGCTGGGGCTGCT	small-conductancemechanosensitive channel protein [Herbaspirillum seropedicae SmR1]
GGGAACGGATTTTGCT	ABC-type branched-chain amino acid transport system, permease component protein [Herbaspirillumseropedicae SmR1]
TGGAATGGATTTTGCT	methyl-acceptingchemotaxis I protein [Herbaspirillum seropedicae SmR1]
GGGCATGAATCATGCT	methyl-acceptingchemotaxis I (serine chemoreceptor) transmembrane protein [Herbaspirillum seropedicae SmR1]
CGGCATAGGTCTTGCC	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
AGGCCCGGCTCTTGCA	permease of the major facilitator superfamily protein [Herbaspirillumseropedicae SmR1]
AGGCACCAGTCTTGCT	methyl-acceptingchemotaxis transducer transmembrane protein [Herbaspirillum seropedicae SmR1]
TGGCGCGCTTCTGGCC	LysR family transcription regulator protein [Herbaspirillumseropedicae SmR1]
TGGCATAGTCCTTGCA	ABC-type amino acid transport system, periplasmic component protein [Herbaspirillumseropedicae SmR1]
TGGGGCGGGGCTTGCC	permease of the major facilitator superfamily protein [Herbaspirillumseropedicae SmR1]
TGGCGCGGCGGCTGCA	short-chain dehydrogenase protein [Herbaspirillumseropedicae SmR1]
TGGCCGGCATCGTGCT	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
TGGCATGCGATTTGCC	NAD+ dependent acetaldehyde dehydrogenase protein [Herbaspirillumseropedicae SmR1]
CGGCATCGCTGCTGCA	LysR family transcription regulator protein [Herbaspirillumseropedicae SmR1]
CGGCACGATCATTGCT	ATP-dependent Clp protease subunit (heat-shock) protein [Herbaspirillumseropedicae SmR1]
TGGTACGGCGCTTGCG	chromosome segregation SMC ATPase protein [Herbaspirillumseropedicae SmR1]
TGGCACCAAGCAGGCT	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
CGGCAAAGCGGTTGCC	ABC-type branched-chain amino acid transport system, permease component protein [Herbaspirillumseropedicae SmR1]
TGGCACGGCGCGCGCG	TonB-dependent siderophore receptor protein [Herbaspirillumseropedicae SmR1]
TGGCAAAACTGATGCA	DnaK suppressor protein [Herbaspirillumseropedicae SmR1]
GGGCATGGCGCTTGCT	AraC family ethanolamine operon transcription regulator protein [Herbaspirillumseropedicae SmR1]
TGGCACGACAGTTGCC	glycine cleavage T (aminomethyltransferase)protein [Herbaspirillum seropedicae SmR1]
TGGTGCGCCAGTTGCA	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
TGGCATGCTCTTTGCT	cysteine desulfurase (NifS) protein [Herbaspirillumseropedicae SmR1]
TGGCACATGAATTGCT	homocitrate synthase protein [Herbaspirillumseropedicae SmR1]

Sequências	Nomes dos candidatos a fator de transcrição sigma 54
GGGAACGATGTTTGCT	two component response regulator protein [Herbaspirillumseropedicae SmR1]
TGGCATGGTTTTGGCT	oxygen-binding (globin) protein [Herbaspirillumseropedicae SmR1]
TGGCACAGCACCGGCA	ABC-type transport system involved in resistance to organic solvents, permease component protein [Herbaspirillumseropedicae SmR1]
TGGAATACGACTTGCT	phospholipid-bindingprotein [Herbaspirillum seropedicae SmR1]
TGGCACGACCTTTGCT	high affinity nitrate/nitritetransporter transmembrane protein [Herbaspirillumseropedicae SmR1]
TGGCAAGCAAGTGGCG	exported protein [Herbaspirillumseropedicae SmR1]
TGGCACTGCCTGCGCA	nucleoside-diphosphate-sugarepimerase protein [Herbaspirillum seropedicae SmR1]
TGGCACAATATTGGCG	RNA polymerase sigma-70 factor, ECF subfamily (sigma-24) transcription regulator protein [Herbaspirillumseropedicae SmR1]
TGGCAGTTTTCTTGCA	methionyl-tRNA synthetase protein [Herbaspirillumseropedicae SmR1]
TGGAACGGATTTTGCT	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
TGGCGCATCGCCTGCA	dihydroorotate dehydrogenase protein [Herbaspirillumseropedicae SmR1]
TGGCATGGAAATCGCT	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
TGGCGCGCCTCTGGCT	Na+(K+)/H+ antiporter protein [Herbaspirillumseropedicae SmR1]
TGGAACGAAATTTGCA	AraC family transcription regulator protein [Herbaspirillumseropedicae SmR1]
GGGCGCTGCCCTTGCT	exporter of the RND superfamily protein [Herbaspirillumseropedicae SmR1]
AGGCATGTTTCTTGCG	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
TGGCCCCCGCCTCGCT	extracellular polysaccharide synthase protein [Herbaspirillumseropedicae SmR1]
TGGCACGTCTGCGGCA	LysR family transcription regulator protein [Herbaspirillumseropedicae SmR1]
TGGCACCGATTTTGCA	acetyl-CoA acetyltransferaseprotein [Herbaspirillum seropedicae SmR1]
TGGTGCAGACATTGCA	TPR repeat containing protein [Herbaspirillumseropedicae SmR1]
TGGCAAGTTTTTTGCT	sugar phosphate isomerase (involved in capsule formation) protein [Herbaspirillumseropedicae SmR1]
GGGAACGTTTCTTGCA	GGDEF domain containing protein [Herbaspirillumseropedicae SmR1]
TGGCGCCGTGATGGCT	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
TGGCACACAACTTGCT	GntR family transcription regulator protein [Herbaspirillumseropedicae SmR1]
TGGCATGACGGTAGCA	ABC-type amino acid transport system, permease component protein [Herbaspirillumseropedicae SmR1]
TGGCACGATATCTGCT	porin precursor transmembrane protein [Herbaspirillumseropedicae SmR1]
TGGCATGCAACTTGCT	methyl-acceptingchemotaxis protein [Herbaspirillum seropedicae SmR1]
CGGCGCGGCTTTTGCT	arogenate dehydratase protein [Herbaspirillumseropedicae SmR1]
CGGCACGTCCATGGCA	cytochrome c family protein [Herbaspirillumseropedicae SmR1]

Sequências	Nomes dos candidatos a fator de transcrição sigma 54
TGGCAAACTCCTTGCG	NADPH-quinone reductase (modulator of drug activity B) protein [Herbaspirillumseropedicae SmR1]
TGGCCTGCTGTTTGCC	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
AGGCATGGAACTTGCG	ABC-type branched-chain amino acid transport system, periplasmic component protein [Herbaspirillumseropedicae SmR1]
TGGCACGCAATCTGCT	ABC-type urea transport system, periplasmic component protein [Herbaspirillumseropedicae SmR1]
GGGGGCGGCTTTTGCT	two component response regulator protein [Herbaspirillumseropedicae SmR1]
TGGCACAATAGGCGCA	hypothetical protein Hsero_4749 [Herbaspirillumseropedicae SmR1]

APÊNDICE 2 - LISTA COM SEQUÊNCIAS E NOMES DOS CANDIDATOS A FATOR DE TRANSCRIÇÃO SIGMA 54 DA 40ª VERSÃO COM 4 BASES DE CONSERVAÇÃO E REDE NEURONAL. NA ORIENTAÇÃO COMPLEMENTA

Sequências	Nomes dos candidatos a fator de transcrição sigma 54
TGGCACTCGACCTGCA	diguanylate cyclase protein [Herbaspirillumseropedicae SmR1]
TGGCCGCGCCCTTGCA	D-isomer specific 2-hydroxyacid dehydrogenase NAD-binding protein [Herbaspirillumseropedicae SmR1]
TGGAACGGAAATTGCT	ammonium transporter transmembrane protein [Herbaspirillumseropedicae SmR1]
TGGCACAAGCTTTGCA	glutathione S-transferase protein [Herbaspirillumseropedicae SmR1]
TGGAAAACTCCTTGCT	LysR family transcription regulator protein [Herbaspirillumseropedicae SmR1]
TGGCGGATACTTTGCT	nitrilase regulator protein [Herbaspirillumseropedicae SmR1]
TGGCCGGGCGCTTGCG	superoxide dismutase [Fe] protein [Herbaspirillumseropedicae SmR1]
TGGCGCAACGCCTGCG	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
TGGCGCTGCCCCTGCG	3-oxoacyl-(acyl-carrierprotein) reductase protein [Herbaspirillum seropedicae SmR1]
AGGCACGCTAGATGCT	AraC family transcription regulator protein [Herbaspirillumseropedicae SmR1]
TGGCATGAAGGATGCT	isoquinoline 1-oxidoreductase(alpha subunit) oxidoreductase protein [Herbaspirillum seropedicae SmR1]
GGGCGCGCACTTTGCG	signal peptide protein [Herbaspirillumseropedicae SmR1]
TGGCATGGCTCCTGCT	amidase family protein [Herbaspirillumseropedicae SmR1]
TGGCGTGATTCTTGCA	GGDEF family protein [Herbaspirillumseropedicae SmR1]
TGGCGCTGATGATGCG	LysR family transcription regulator protein [Herbaspirillumseropedicae SmR1]
TGGCAGCGTCCTTGCC	hypothetical protein Hsero_1625 [Herbaspirillumseropedicae SmR1]
TGGCACTGGCATTGCT	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
TGGCAAAGAGCCTGCA	GGDEF family protein [Herbaspirillumseropedicae SmR1]
TGGCACGGACCTTGCA	transmembrane protein [Herbaspirillumseropedicae SmR1]
TGGCATGCTTGCTGCT	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
TGGCTAGCGCTTTGCA	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
CGGCGCGGCATTTGCC	ABC-type glutamate/aspartatetransport system, permease component protein [Herbaspirillum seropedicae SmR1]
TGGTGCGAACCTTGCC	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]

Sequências	Nomes dos candidatos a fator de transcrição sigma 54
TGGCAGACCGGGTGCT	multidrug resistance efflux pump protein [Herbaspirillumseropedicae SmR1]
CGGCATGAAACCTGCT	PII uridylyltransferaseprotein [Herbaspirillum seropedicae SmR1]
TGGCAAGGGCCTGGCT	two component response regulator protein [Herbaspirillumseropedicae SmR1]
TGGCTGTTCTGCGGCC	two component response regulator protein [Herbaspirillumseropedicae SmR1]
TGGACCTGGGGTTGCA	transcription regulator, LysR family protein [Herbaspirillumseropedicae SmR1]
TGGCACTGGGGGTGCG	DNA polymerase II protein [Herbaspirillumseropedicae SmR1]
TGGCGCAGGGCGTGCT	dienelactone hydrolase protein [Herbaspirillumseropedicae SmR1]
TGGCATTAATTTGGCA	transmembrane sensor histidine kinase transcription regulator protein [Herbaspirillumseropedicae SmR1]
TGGCACGGAACTGGCA	D-3-phosphoglyceratedehydrogenase protein [Herbaspirillum seropedicae SmR1]
TGGCATGGGTATGGCA	exported protein [Herbaspirillumseropedicae SmR1]
TGGCGCACCAATAGCT	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
TGGCCCGGTCCTTGCT	hypothetical protein Hsero_2758 [Herbaspirillumseropedicae SmR1]
TGGCATGCCGATTGCT	hypothetical protein Hsero_2759 [Herbaspirillumseropedicae SmR1]
TGGCACGGCATTTGCA	dinitrogenase reductase protein [Herbaspirillumseropedicae SmR1]
TGGCACGGTTTTGGCT	FeMo cofactor biosynthesis protein [Herbaspirillumseropedicae SmR1]
GGGCATGAAGTTTGCT	nif-specific regulatory protein [Herbaspirillumseropedicae SmR1]
GGGCGCAGTAGTTGCT	ATPase involved in chromosome partitioning protein [Herbaspirillumseropedicae SmR1]
TGGCACGGCCGTTGCA	carboxyphosphonoenolpyruvatephosphonomutase protein [Herbaspirillum seropedicae SmR1]
TGGCATGGAATCTGCT	glutamine synthetase protein [Herbaspirillumseropedicae SmR1]
AGGCGCCGTTCTTGCA	nucleoside-diphosphate-sugarepimerase protein [Herbaspirillum seropedicae SmR1]
TGGCACGAAGCCTGCA	alpha-ketoglutaratepermease of the major facilitator superfamily protein [Herbaspirillum seropedicae SmR1]
CGGCGCAGCGCTTGCG	conserved hypothetical protein [Herbaspirillumseropedicae SmR1]
TGGCAGAGGCGATGCT	transcription regulator protein [Herbaspirillumseropedicae SmR1]
TGGCGCCAGCAGCGCC	transcription regulator protein [Herbaspirillumseropedicae SmR1]
TGGCGTGATTTTTGCT	methyl-acceptingchemotaxis protein I, serine sensor receptor transmembrane protein [Herbaspirillum seropedicae SmR1]
TGGCATGTTTTATGCG	hypothetical protein Hsero_3544 [Herbaspirillumseropedicae SmR1]
TGGCATCGTTATTGCT	diguanylate cyclase/phosphodiesterasewith PAS/PAC sensor domains protein [Herbaspirillum seropedicae SmR1]
TGGCATCCAAATTGCT	TRAP-type C4-dicarboxylatetransport system, periplasmic component protein [Herbaspirillum seropedicae SmR1]