UNIVERSIDADE FEDERAL DO PARANÁ

IDENTIFICAÇÃO E ANÁLISE DE PROMOTORES SIGMA 70 NO GENOMA DE *Herbaspirillum seropedicae* SmR1 UTILIZANDO MÉTODOS DE INTELIGÊNCIA ARTIFICIAL

CURITIBA

2014

RODNEI DAMACENO FREIRE

IDENTIFICAÇÃO E ANÁLISE DE PROMOTORES SIGMA 70 NO GENOMA DE *Herbaspirillum seropedicae* SmR1 UTILIZANDO MÉTODOS DE INTELIGÊNCIA 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
Coorientador: Prof Dr. Roberto Tadeu Raittz

CURITIBA

2014

TERMO DE APROVAÇÃO

RODNEI DAMACENO FREIRE

IDENTIFICAÇÃO E ANÁLISE DE PROMOTORES SIGMA 70 NO GENOMA DE *Herbaspirillum seropedicae* SmR1 UTILIZANDO MÉTODOS DE INTELIGÊNCIA 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:

Orientadora: Profª Drª Liu Un Rigo

Coorientador: Prof. Dr. Roberto Tadeu Raittz

Profª. Drª. Roseli Wassem

Universidade Federal do Paraná

Dr. Vinicius Almir Weiss

Universidade Federal do Paraná

Curitiba, 29 de setembro de 2014.

*“*Dedico este trabalho aos meus pais que me deram apoio nos momentos mais difíceis da minha vida, a minha esposa e filho que estiveram sempre ao meu lado ao longo deste estudo, aos meus professores que me ensinaram que por mais que achamos que o nosso conhecimento já está bem profundo, estamos enganados, pois o conhecimento é algo que estará sempre se renovando”

AGRADECIMENTOS

- À minha orientadora Profª. Drª. Liu Un Rigo pelo apoio, grandes ensinamentos, confiança e por ter sempre acreditado no projeto.

- Ao meu coorientador Prof. Dr. Roberto Tadeu Raittz, pelos ensinamentos, paciência, apoio, e confiança no projeto.

- À Profª. Drª Maria Berenice Reynaud Steffens, pelos aconselhamentos e pela dedicação ao curso de Pós- Graduação em Bioinformática.

- À Profª. Drª. Jeroniza Marchaukoski, pela dedicação e auxílio nos assuntos relacionados ao curso de Pós-Graduação em Bioinformática.

- Ao Prof. Dr. Emanuel Maltempi de Souza, pelas sugestões e explicações.

- À Profª. Drª. Rose Adele Monteiro pelo apoio e cooperação neste projeto.

- Ao Prof. Dr. Leonardo Magalhaes Cruz, pelas sugestões e aconselhamentos.

- Aos demais professores e funcionários do Programa de Pós-graduação em Bioinformática.

- À Drª. Heladia Salgado, Departamento de Microbiologia Molecular, Instituto de Biotecnologia, Programa de Genômica Computacional, Centro de Ciências Genômicas, Universidade Nacional Autônoma do México, pela gentileza e contribuições importantes neste trabalho.

- Ao companheiro de bancada Dr. Fernando Bachega Ruggiero, pela grande amizade, aconselhamentos e colaborações relevantes a este projeto.

- Ao Ms. Lucas M. Ferreira, pelos aconselhamentos e auxilio prestado.

- À Suzana, pela atenção e dedicação ao Programa de Pós-graduação em Bioinformática.

- Aos amigos Eslei Xavier, Fausto Koga e Helba Cirino Barboza, pelo apoio, incentivo e todos os momentos de descontração ao longo deste curso.

- Aos demais colegas de laboratório, pelo apoio, incentivo e companheirismo.

- Ao Núcleo de Fixação de Nitrogênio da Universidade Federal do Paraná.

- Aos órgãos fomentadores: CAPES, CNPq e REUNI.

- A Deus, acima de tudo.

RESUMO

Sigma 70 ou sigma N constituem fatores complementares da RNA-polimerase, cuja principal função é promover a transcrição de genes procarióticos. No caso de *Escherichia coli*, o consenso da região -35 (TTGACA) e -10 (TATAAT) do fator sigma 70, está localizado a partir do intervalo da décima à trigésima quinta base a montante do sítio de início de transcrição e as bases mais conservadas estão localizadas nas posições -10 (A^2 = 95% - T^6 = 96%) e -35 (T^1 = 82% - T^2 = 84%). Propusemos neste trabalho identificar sequências promotoras de transcrição dependentes do fator sigma 70, utilizando um algoritmo que pré-seleciona candidatos aos promotores sigma 70 com base no padrão de conservação. Os candidatos são então classificados através de treinamento de rede artificial, com conjunto de sequências de promotores sigma 70 validados e um conjunto de sequências improváveis, geradas aleatoriamente. O método foi testado *in silico* no genoma da betaproteobactéria *Herbaspirillum seropedicae* SmR1, resultando em 4.998 sequências candidatas a promotores fator sigma 70. Deste grupo foram selecionados 288 candidatos a partir das regiões intergênicas de genes com alto nível de expressão. Isto tornou possível validar os resultados obtidos para identificação de sequências promotoras sigma 70 e propor uma sequência consenso para o promotor de transcrição sigma 70 em *Herbaspirillum seropedicae* SmR1. A metodologia utilizada para identificar os sítios de ligação sigma 70 mostrou-se eficaz na identificação de candidatos aos promotores sigma 70 em *H. seropedicae* SmR1 e possivelmente em outras proteobactérias.

Palavras-chave: *Herbaspirillum*, Promotores, Fatores de Transcrição, sigma 70.

ABSTRACT

Sigma 70 or sigma N constitute complementary sigma factors of RNA-polymerase, whose main function is to promote the transcription of prokaryotic genes. In the case of *Escherichia coli*, the consensus of the -35 region (TTGACA) and -10 (TATAAT) sigma 70 factor sequence, located from the range of the tenth to the thirty-fifth base upstream of the transcription start site and the bases more conserved are located at positions -10 (A^2 = 95% - T^6 = 96%) and -35 (T^1 = 82% - T^2 = 84%). We proposed in this work to identify promoter sequences of the sigma 70 dependent transcription factor, using an algorithm that pre-selects candidates for sigma 70 promoters based on conservation pattern. The candidates sequences are ranked using artificial neural network training set of validated sigma 70 promoter sequences and a set of randomly generated sequences. The method was tested *in silico* using the Betaproteobacteria *Herbaspirillum seropedicae* SMR1 genome, resulting in 4.998 candidate sequences for promoters to sigma 70 factor with standard conservation. Among these candidates 288 were manually selected from the intergenic regions of genes with high expression level. This made it possible to validate the results obtained for indentification of sigma 70 sequences and propose a consensus sequence for transcriptional promoter sigma 70 in *Herbaspirillum seropedicae* SMR1. The methodology used to predict sigma 70 binding sites showed effectiveness to identify candidates for sigma 70 promoters in *H. seropedicae* SMR1 and possibly in other proteobacteria.

Keywords: *Herbaspirillum*, Promoters, Transcription Factors, sigma 70.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS, SÍMBOLOS E ABREVIATURAS

α – alpha

β – beta

β´ – beta linha

σ – sigma

µm – micrômetros

Conteúdo GC – quantidade de bases “G” em adição às bases “C” em relação ao total de bases de uma determinada sequência

DDBJ – DNA Databank of Japan

DDR – Double Data Rate

DNA – Ácido Desoxirribonucleico

EMBL – European Molecular Biology Laboratory Nucleotide Sequence Database

Gb – Giga Bytes

Gram negativa – bactéria que possui lipopolissacarídeos na membrana externa, o que resulta em colaração avermelhada quando coradas pela técnica de Gram

I.A – Inteligência Articial

MATLAB – Matrix Laboratory

MLP – Multilayer Perceptron

NCBI – National Center for Biotechnology Information

N – nitrogênio (elemento químico)

N2 – nitrogênio (gás atmosférico)

NH_4Cl – cloreto de amônio

Ns – bases indeterminadas

RBS – Ribosome Binding Site

RNA – Ácido Ribonucleico

RNAm – Ácido Ribonucleico Mensageiro

RNAr – Ácido Ribonucleico Ribossomal

RNAt – Ácido Ribonucleico Transportador

SM – Similaridae Máxima

***spp.* –** conjunto de espécies distintas pertencentes ao mesmo gênero

TSS – Transcription Start Site (sítio de início de transcrição)

SUMÁRIO

1 INTRODUÇÃO

1.1 Bioinformática

A informática ao longo dos anos vem se destacando como uma área de conhecimento específico e tecnológico de extrema importância para a humanidade, associando-se com diversas áreas, auxiliando e acelerando pesquisas científicas de diversas formas. A associação entre a área da Informática com a área das Ciências Biológicas deu origem a uma nova área interdisciplinar, a Bioinformática, sendo esta responsável pelo processamento da grande quantidade de dados e informações biológicas geradas a partir de estudos e análises em laboratórios de Biologia Molecular.

Para o Centro Nacional de Informações sobre Biotecnologia (NCBI) 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 objetivo final a descoberta de novos conhecimentos biológicos. De acordo com 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 químicos e biológicos. Já para 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 conhecimentos biológicos em modelos computacionais processáveis (FOX, 2009). Como uma nova área de conhecimento, a Bioinformática trouxe exaltação para a comunidade científica, justamente pela possibilidade de imersão em um mundo totalmente novo e desconhecido (FOX, 2009). De acordo com Bayat (2002), a Bioinformática é uma matéria interdisciplinar que abrange várias áreas do conhecimento como Biologia, Medicina, Matemática, Física, Ciências da Computação e Estatística. Um profissional adequado para a área de Bioinformática deve ter noções específicas nas disciplinas de Biologia e Ciências da Computação, detendo a capacidade de entender assuntos relacionados à Biologia Molecular além da aptidão de desenvolver ou aplicar softwares. 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 teve sua origem na década de 1960, quando a pesquisadora Margaret Dayhoff (1925-1983) organizou e disponibilizou o primeiro atlas de sequências proteicas, 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, uma matriz para a substituição de aminoácidos, amplamente utilizada nos dias atuais.

Os progressos na área da computação trouxeram várias facilidades para a Bioinformática, permitindo o armazenamento de uma maior quantidade de dados com qualidade e velocidade no processamento das informações. Com estes avanços na tecnologia surgiu um aumento no número de projetos de montagem de genomas. Um exemplo clássico é o próprio genoma humano, com o seu sequenciamento anunciado no dia 26 de junho de 2000, 60 meses antes da data estimada em 1987 (VOGT, 2003).

Para termos uma ideia da progressão que a Bioinformática vem apresentando, há vinte anos atrás uma sequência nucleotídica com uma média de 12 mil pares de bases levaria um ano para ser sequenciada, há três anos a mesma sequência levaria cerca de uma hora para ser concluída e atualmente leva menos de um minuto para que o sequenciamento seja concluído (VOGT, 2003).

O NCBI (Centro Nacional de Informações sobre Biotecnologia), EMBL-EBI (Instituto Europeu de Bioinformática) e o DDBJ (Base de Dados de DNA do Japão) são as principais bases de dados que armazenam informações para a Bioinformática, tendo como principal objetivo 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

A transcrição faz a passagem de informações contidas na molécula de DNA para uma fita simples de RNA e no decorrer do processo de transcrição, um sistema enzimático converte a informação genética de um segmento de DNA em uma fita de RNA

mensageiro com uma sequência de bases complementares a uma das fitas do DNA (NELSON E COX, 2000). A figura 1 mostra como ocorre o processo de trancrição em procariotos.

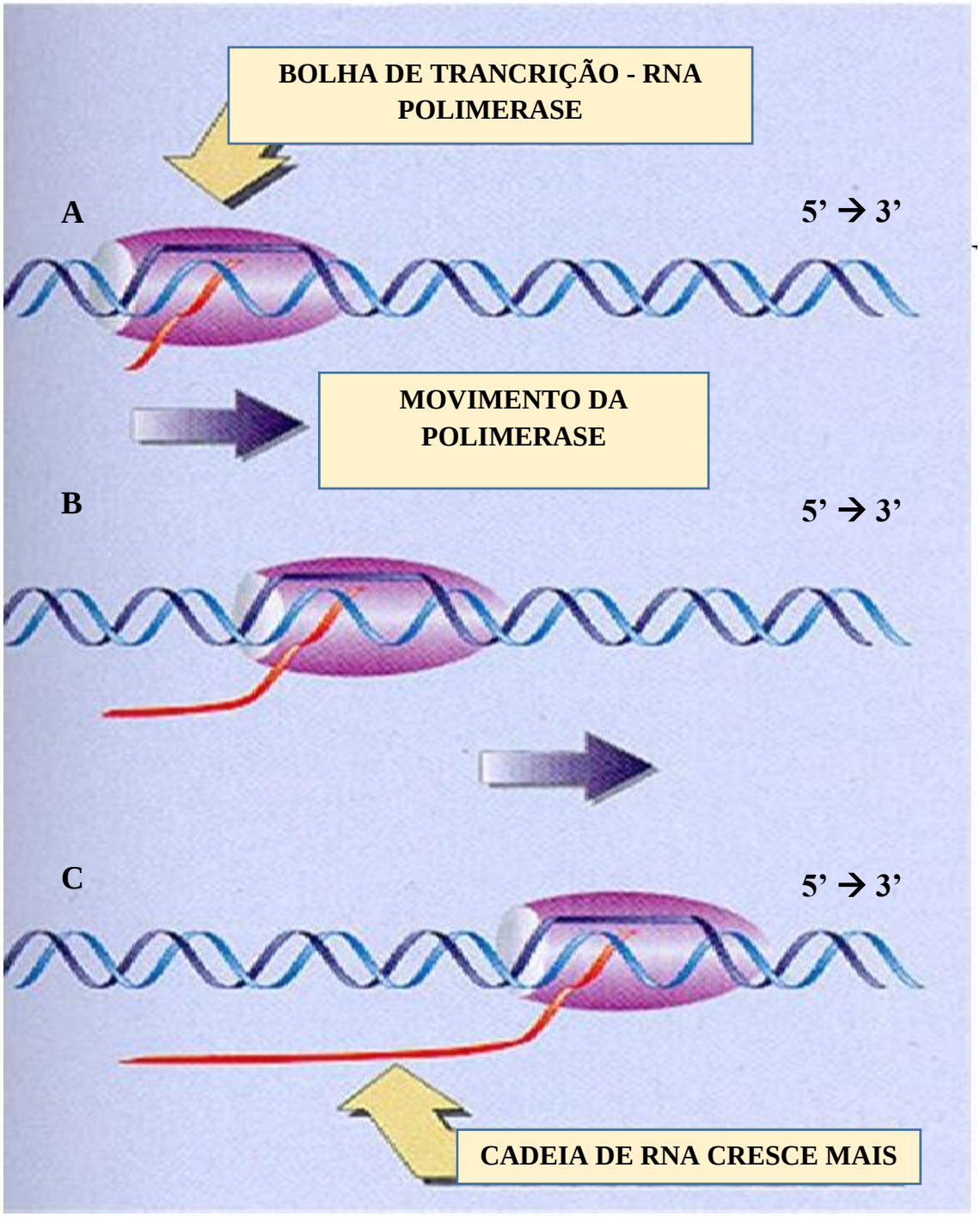


FIGURA 1 - ESQUEMA DO PROCESSO DE TRANSCRIÇÃO.

Em (**A**), o acoplamento da RNA-polimerase na região promotora de transcrição do DNA e o início de transcrição compondo os primeiros nucleotídeos do mRNA. Em (**B**), podemos observar que à medida que a RNA-polimerase desliza sobre o DNA, mais nucleotídeos são adicionados na cadeia ribonucleica, compondo uma fita simples de mRNA. Em (**C**), temos a fase de alongamento da cadeia ribonucleica.

FONTE: ADAPTADO DE LEWIN B., GENES VII (2000).

O processo de transcrição é muito similar ao processo de replicação em seu mecanismo químico fundamental: sua polaridade, e o uso de um molde, possuindo também semelhanças nas fases de iniciação, alongamento e terminação (NELSON E COX, 2000). Nas regiões a serem transcritas existem sinalizadores compostos por sequências reguladoras específicas que indicam o ponto onde deve ser iniciada a transcrição e onde deve ocorrer a terminação (NELSON E COX, 2000).

Vários tipos de RNAs são gerados no processo de transcrição, como por exemplo: um RNA mensageiro (mRNA), que codifica a sequência de aminoácidos de um ou mais polipeptídios especificados por um gene ou conjunto de genes, RNA transportador (ou de transferência, tRNA), que faz a leitura da informação codificada no RNA mensageiro e leva os aminoácidos correspondentes a ela até os ribossomos e RNA ribossomal (rRNA), que são constituintes do ribossomo, sendo estes maquinarias celulares responsáveis pela síntese das proteínas. Além destes três tipos existem outros RNAs sintetizados no processo de transcrição através da holoenzima de participação fundamental denominada RNA-polimerase, presente em procariotos e em eucariotos.

Em eucariotos temos presentes três tipos de RNA-polimerase (NELSON E COX, 2000). A nova fita de RNA é sintetizada na direção 5' → 3', antiparalelo à fita molde de DNA, e os nucleotídeos são adicionados respeitando interações de pareamento de bases Watson-Crick, havendo uma substituição na ligação das bases nitrogenadas Timina - Adenina pela Uracila – Adenina, configurando uma molécula de RNA (NELSON E COX, 2000).

De acordo com Kumar (1981), a RNA-polimerase procariótica pode ser isolada nas células das seguintes formas:

A) Completa com as cinco subunidades, formando a holoenzima completa;

B) Somente quatro delas, compondo a apoenzima.

As cinco subunidades do complexo enzimático da Holoenzima RNA-polimerase são representados por β (beta), β' (beta linha), α (alfa), σ (sigma). Tendo em vista que β e β´ compõem o centro catalítico da enzima, participando de todas as fases da transcrição e 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, como demonstrado na figura 2. Esta subunidade ou fator σ não é fixa na RNA-polimerase e reconhecendo o sítio de ligações ao DNA, regiões específicas denominadas promotoras (WÖSTEN, 1998).

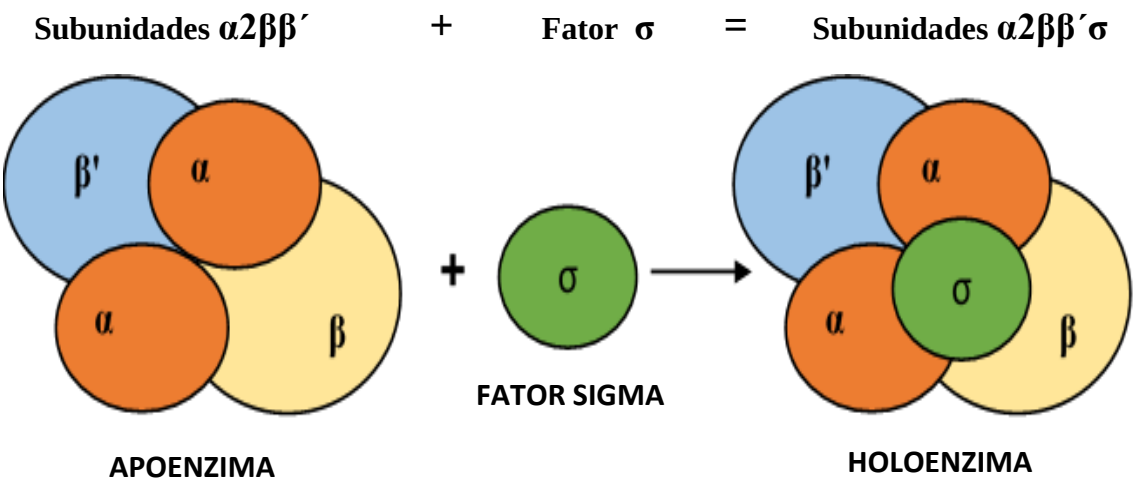


FIGURA 2 - REPRESENTAÇÃO DA INTERAÇÃO DO FATOR SIGMA À APOENZIMA COMPONDO A HOLOENZIMA.

Essa figura enfatiza as diferenças entre a apoenzima e a holoenzima. A adição do fator σ na apoenzima determina a composição da holoenzima.

FONTE: O AUTOR, (2014), BASEADO EM KUMAR, (1981).

1.3 Fatores sigma

Os fatores sigma são subunidades da RNA-polimerase capazes de reconhecer uma determinada região do DNA denominada de região promotora. De modo inerente, 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). A holoenzima abre a dupla hélice para a iniciação do processo de transcrição (DOUCLEFF *et al.*, 2005).

O ciclo de ligação com os fatores sigma auxilia 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 função do peso molecular do primeiro membro da família identificado (DOUCLEFF *et al.*, 2005; 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 (MC CLURE, 1985; GRALLA, 1996). Quando a apoenzima se liga ao sigma 54, não existe a mesma capacidade de formação do complexo aberto, e há necessidade de ligação com outros fatores proteicos para a ativação da transcrição (SASSE-DWIGHT E GRALLA, 1988; MORETT E BUCK, 1989; e MORETT E SEGOVIA, 1993).

A figura 3 mostra interação da RNA-polimerase com a dupla fita de dna sobre a região promotora dependente do fator sigma 70.

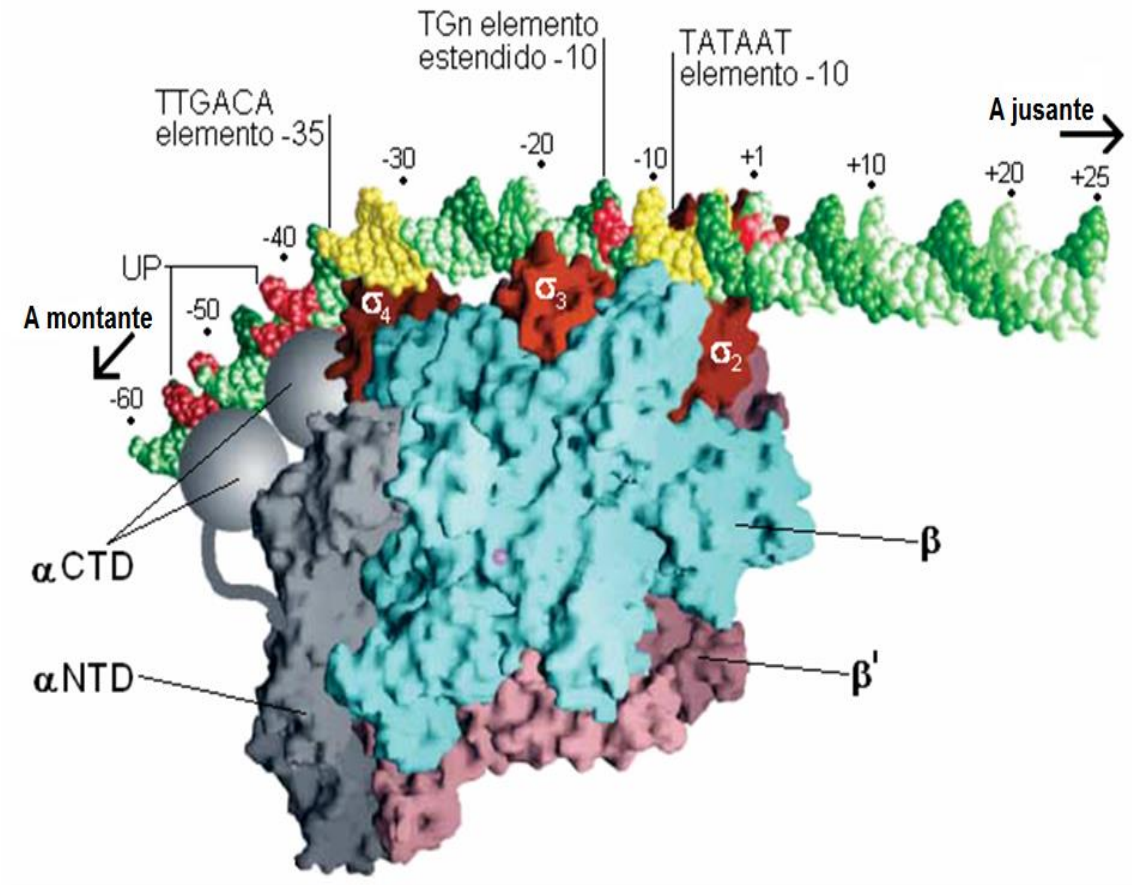


FIGURA 3 - INTERAÇÃO DA RNA-POLIMERASE COM A DUPLA FITA DE DNA SOBRE A REGIÃO PROMOTORA DEPENDENTE DO FATOR SIGMA 70.

As subunidades σ são responsáveis pelo reconhecimento do trecho promotor do DNA. A região -35 (TTGACA) é reconhecida pelo σ4. A região -10 (TATAAT) é reconhecida pelo σ2, podendo haver a interação da região -10 estendida (TGn) reconhecida pelo σ3. A subunidade β' é responsável pela ligação ao DNA. A subunidade β está envolvida na elongação e no início da cadeia ribonucleotídica. As duas subunidades α estão relacionadas com o início da transcrição e na interação com proteínas regulatórias que controlam os processos transcritivos.

FONTE: ADAPTADO DE MURAKAMI, 2002.

Estudos realizados em *Escherichia coli*, reportaram a presença de sete fatores sigma constituintes das 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) demonstrados no quadro abaixo.

QUADRO 1 - TIPOS DE FATORES SIGMA.

SIGMA		SIMBOLOGIA	FAMÍLIA	FUNÇÃO/RELAÇÃO
1	sigma 70	(σ70)	sigma 70	Manutenção Celular
2	sigma 38	(σ38)	sigma 70	Manutenção Celular
3	sigma 32	(σ32)	sigma 70	Manutenção Celular
4	sigma 28	(σ28)	sigma 70	Manutenção Celular
5	sigma 24	(σ24)	sigma 70	Manutenção Celular
6	sigma 19	(σ19)	sigma 70	Manutenção Celular
7	sigma 54	(σ54)	sigma 54	Fixação de Nitrogênio

Quadro apresentando os tipos de fatores sigma relacionados com a família e função na célula.

FONTE: ADAPTADO DE 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 hexâmeros posicionados nas bases -35 e -10 em relação à 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 à primeira base de transcrição (BARRIOS *et al.*, 1999).

Outra diferença é que os promotores da família sigma 70 regulam genes relacionados à manutenção e organização celular, já os promotores da família sigma 54 estão envolvidos com a regulação de genes bacterianos que promovem a fixação de

nitrogênio (BARRIOS *et al.*, 1999). A figura 4 representa as sequências consenso para os fatores sigma 70 e sigma 54.

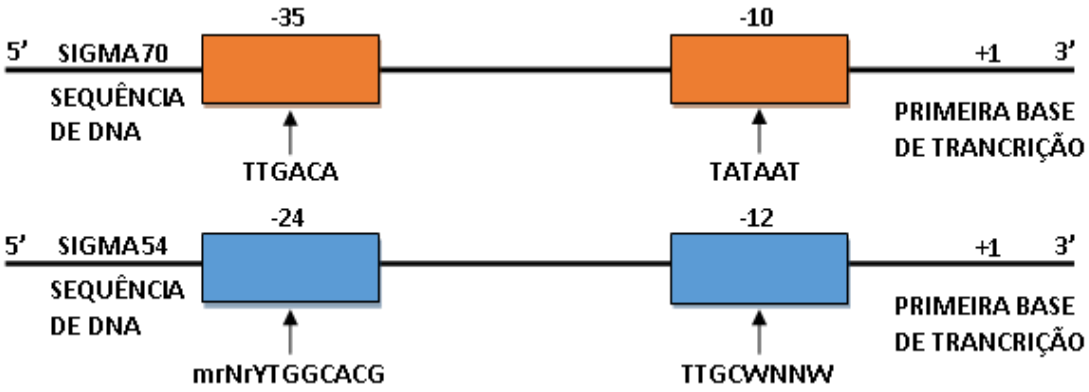


FIGURA 4 - REPRESENTAÇÃO GRÁFICA DAS SEQUÊNCIAS CONSENSO PARA OS FATORES SIGMA 70 E SIGMA 54.

Esquema inicial de representação de promotores σ70 e σ54 em procariotos.

FONTE: ADAPTADO DE POTVIN (2007) E BARRIOS (1999).

A região -10 pode ser chamada de Pribnow box, sendo esta região a mais conservada. Enquanto para a região -35 não há outra denominação. As diferenças principais entre os dois promotores estão relacionadas com a sequência de bases que as compõem e a distância onde estas estão dispostas em relação ao +1 (sítio de início de transcrição). A figura 5 representa os promotores sigma 70 em *Escherichia coli* com as proporções de bases conservadas nos hexâmeros que compõem a regiâo -35 e -10.

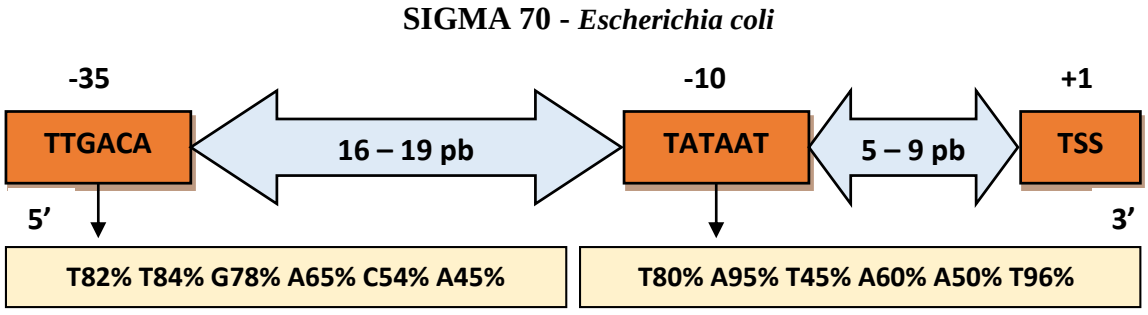


FIGURA 5 - REPRESENTAÇÃO DE PROMOTORES SIGMA 70 EM *Escherichia coli* COM AS PROPORÇÕES DE BASES CONSERVADAS NOS HEXÂMEROS QUE COMPÕEM A REGIÂO -35 E -10.

Entre os hexâmeros poderá haver 16 aos 19 pares de bases, sem consenso descrito. Entre a região -10 e o sitio de início de transcrição (TSS) pode haver de 5 - 9 pares de bases, sem consenso descrito.

FONTE: ADAPTADO DE LEWIN B., 2000.

1.4 Regulação da expressão gênica em procariotos

Naturalmente a célula responde a sinais intra e extracelulares. Estes sinais são captados e respondidos de acordo com a necessidade de adequação, tendo em vista que comumente os organismos são expostos de forma repentina as mais diversas formas de variação do meio em que estas se encontram, como por exemplo a variação de temperatura, níveis de pH, osmolaridade e a disponibilidade de diversos tipos de nutrientes (PARKINSON,1993; STOCK, NINFA E STOCK, 1989). Os sinais são interpretados através de eficientes redes de transdução de sinal. Essas redes são formadas por proteínas que interagem entre si, produzindo desta forma, respostas adaptativas e condizentes ao meio que a célula se encontra (STOCK, ROBINSON E GOUDREAU, 2000).

Em resposta aos sinais recebidos, ocorrem mudanças mais intensas, estabelecendo uma necessária alteração do padrão da expressão gênica, onde será transcrito um conjunto alternativo de proteínas (STOCK, NINFA e STOCK, 1989). A transcrição é devidamente reprogramada de acordo com as respostas adaptativas das células, modificando desta forma as enzimas que serão transcritas e quais enzimas serão dispensadas para uma eventual reorganização do metabolismo celular, cuja alteração das atividades de suas enzimas, reorganizam o fluxo metabólico de acordo com o novo meio em que esta esteja inserida, definindo desta forma, vias metabólicas alternativas (WHITE, 2000).

Evolutivamente, as células desenvolveram alguns sistemas de monitoramento dos sinais, como por exemplo o sistema regulador de dois componentes. Neste sistema, um dos componentes funcionará como um sensor, sendo a atuação desta proteína sensora uma quinase ou fosfatase sinalizando ao outro componente através do processo de fosforilação ou desfosforilação, enquanto o outro componente atuará como regulador, promovendo ou reprimindo a transcrição (WEST E STOCK, 2001). Desta forma o sistema regulador de dois componentes é responsável pela percepção, interpretação e resposta aos sinais recebidos, proporcionando uma devida adaptação dos organismos às sucessivas e inesperadas mudanças de ambiente (STOCK, ROBINSON, GOUDREAL, 2000; FOUSSARD *et al*, 2001; GALPERIN, 2004).

1.5 *Herbaspirillum seropedicae*

Herbaspirillum, palavra derivada do latim *herba* (herbáceo) e *spirillum* (pequena espiral) é um gênero de bactérias pertencentes à subclasse β das proteobactérias, aeróbias e não fermentadoras de açúcares, Gram-negativas, móveis, vibrióides, algumas vezes helicoidais, possuindo de um a três flagelos em um ou ambos os pólos e possuem entre 0,6-0,7 µm de diâmetro e seu comprimento pode variar de 1,5 a 5,0 µm (YOUNG, 1992).

Essas bactérias são encontradas em associação com raízes, caules e folhas de diversas plantas monocotiledôneas, geralmente gramíneas de grande importância econômica como arroz, sorgo, milho, trigo, cevada e cana-de- açúcar, entre outras plantas forrageiras e até mesmo em plantas dicotiledôneas como banana e abacaxi (JAMES *et al*. 1998; RONCCATO-MACARI *et al.*, 2003).

O quadro 2 mostra como é a classificação taxonômica de *Herbaspirillum seropedicae* atualmente.

QUADRO 2 - CLASSIFICAÇÃO TAXONÔMICA DE *Herbaspirillum seropedicae*.

DOMÍNIO	Bactéria
FILO	Proteobactérias
CLASSE	Proteobactérias beta
ORDEM	Burkholderiales
FAMÍLIA	Oxalobacteraceae
GÉNERO	*Herbaspirillum*
ESPÉCIE	*Herbaspirillum seropedicae*

O quadro acima mostra a classificação taxonômica atualizada para o gênero *Herbaspirillum*.

FONTE: www.ncbi.nlm.nih.gov/ , ACESSADO EM 20/05/2013.

No quadro 3 a seguir, estão destacados alguns exemplares de espécies do gênero *Herbaspirillum*.

QUADRO 3 - EXEMPLOS DE ESPÉCIES DO GÊNERO *HERBASPIRILLUM*.

Herbaspirillum hiltneri	*Herbaspirillum huttiense*
Herbaspirillum frisingense	*Herbaspirillum lusitanum*
Herbaspirillum putei	*Herbaspirillum chlorophenolicum*
Herbaspirillum rubrisubalbicans	*Herbaspirillum autotrophicum*

O quadro acima destaca oito espécies do gênero *Herbaspirillum*.

FONTE: ADAPTADO DE BALDANI *et al*, 1986.

A espécie Herbaspirillum seropedicae também foi isolada de outras plantas não-leguminosas, como por exemplo a palmácea conhecida como dendê, onde foi encontrada no interior de raízes e caules (DÖBEREINER *et al.*, 1995). A disseminação natural de *H. seropedicae* ainda não está bem evidente, podendo esta ocorrer por sementes ou através da propagação vegetativa, como no caso de *A. diazotrophicus* em plantas de cana-de-açúcar. A disseminação por propagação vegetativa foi confirmada pela presença da *Herbaspirillum* em plantas de cana-de-açúcar originadas por processos de micropropagação, nos quais o meristema apical não foi cuidadosamente extraído (OLIVARES *et al.*, 1996). A evidência de que sua disseminação comumente ocorre através de sementes, foi comprovada a partir de isolados de sementes de cereais, como por exemplo o arroz (BALDANI *et al.*, 1997). Podemos afirmar que *H. seropedicae* não sobrevive bem no solo natural, como outros endófitos, sendo sua sobrevivência menos afetada em solo estéril, o que indica que fatores biológicos interferem na sobrevivência desta bactéria. Todavia, a taxa de sobrevivência de *H. seropedicae* em ambos os solos foi mais alta do que a observada para *A. diazotrophicus* (BALDANI *et al.*, 1997).

A outra espécie do gênero *Herbaspirillum*, *H. rubrisubalbicans*, foi derivada da reclassificação de *Pseudomonas rubrisubalbicans*, considerada um agente fitopatogênico causador da doença conhecida como estria mosqueada em algumas variedades susceptíveis de cana-de-açúcar cultivadas no Brasil (PIMENTEL *et al.*,1991). Esta espécie foi incluída no gênero *Herbaspirillum* com base na homologia DNA:rRNA e em algumas características fisiológicas, como a incorporação do gás $15N_2$ (BALDANI *et al.*, 1997).

Estudos de caracterização ecológica dessa nova espécie demonstram claramente a ocorrência de *H. rubrisubalbicans* em raízes, caules e folhas de cana-de-açúcar de todas as partes do mundo, sendo também encontrada em arroz e palmeira (FERREIRA *et al.*, 1995; BALDANI *et al.*, 1997a; OLIVARES *et al.*, 1996). A capacidade de *Herbaspirillum* formar uma associação com gramíneas de interesse econômico como milho, sorgo e cana-de-açúcar (DÖBEREINER, 1992), sem provocar doença, tem despertado interesse para o seu estudo.

O mecanismo de colonização dos tecidos vegetais por essas bactérias ainda não está completamente esclarecido. Análises microscópicas de plantas colonizadas sugerem que ocorra a ligação à superfície da planta e proliferação das bactérias, seguida de penetração na planta e ocupação de espaços intercelulares e feixes vasculares com posterior colonização e estabelecimento nas partes aéreas e vasos do xilema (OLIVARES *et al.*,1995, RONCATO-MACCARI *et al.*, 2003). O potencial como biofertilizante de *H. seropedicae* torna importante o estudo dos mecanismos celulares deste microrganismo.

1.6 Inteligência Artificial

De acordo com Luger (2004), a Inteligência Artificial (I.A) é um ramo da ciência da computação que se ocupa com o comportamento inteligente e aprendizagem de máquinas. Já para Rich (1994), a I.A é o estudo de como fazer os computadores realizarem coisas que, atualmente, os humanos fazem melhor.

O objetivo principal da I.A se baseia em entender entidades inteligentes e reproduzir o comportamento inteligente desenvolvendo sistemas para realizar tarefas que não possuem solução algorítmica satisfatória pela computação convencional (LUGER, 2004).

Pensando em algumas características básicas desses sistemas, como a capacidade de raciocínio, os métodos de I.A visam aplicar regras lógicas a um conjunto de dados disponíveis para chegar a uma conclusão lógica, bem como a aprendizagem de máquina visando o aperfeiçoamento baseado em erros e acertos, para que futuramente, tenham uma ação de maneira mais eficaz (LUGER, 2004). Uma das atividades que envolvem as técnicas de I.A aborda o reconhecimento de padrões, tanto padrões visuais e sensoriais, como também padrões de comportamento, tendo como objetivo a inferência computacional, sendo esta, a capacidade de conseguir aplicar o raciocínio das máquinas em situações reais do nosso cotidiano através de treinamento e aprendizagem de redes neurais artificiais (LUGER, 2004).

1.7 Redes neurais artificiais

Para Nievola (1998), as redes neurais consistem em uma abordagem de inteligência artificial para as soluções de problemas que tem por base o cérebro humano. Já para Rezende (2003), as redes neurais artificiais são modelos matemáticos que se parecem com as estruturas neurais biológicas, tendo a capacidade computacional adquirida através da aprendizagem e generalização.

Com a grande complexidade das redes neurais biológicas, possuindo bilhões de neurônios, enquanto que as redes neurais artificiais apresentam apenas dezenas a milhares de unidades de processamento, sendo a velocidade de aprendizagem das redes artificiais torna-se o grande diferencial.

Os neurônios das redes artificiais, que podem ser chamados de nós ou nodos, são interconectados imitando o funcionamento do cérebro humano, e dispostos em camadas estão 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).

A figura 6 é uma representação de um neurônio biológico e uma rede neural artificial.

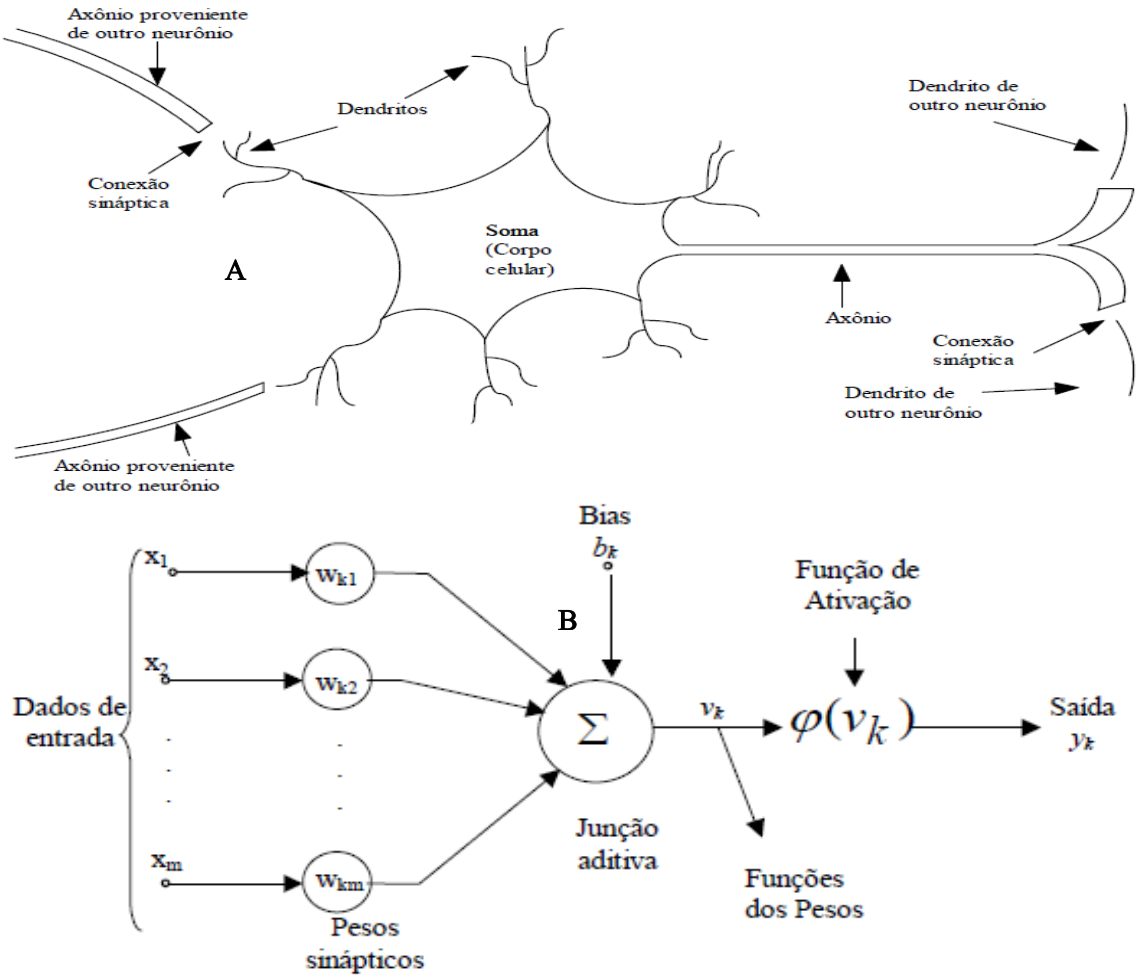


FIGURA 6 - DESENHO REPRESENTATIVO DE UM NEURÔNIO BIOLÓGICO E UMA REDE NEURAL ARTIFICIAL.

Em (**A**), representação das conexões sinápticas entre o axônio de um neurônio com o dendrito de um neurônio vizinho. O núcleo desta célula está contido no corpo celular do neurônio (Soma). Em (**B**), representação de um modelo de rede neural artificial, com os dados de entrada sendo convertidos para pesos sinápticos, associados à junção aditiva (Soma). Funções dos pesos e ativação determinam a saída dos dados que irão interagir com o próximo neurônio artificial, como se fosse uma ligação sináptica.

FONTE: ADAPTADO DE FAUSETT (1994) E 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, minimizando os efeitos da falha para o resultado do processamento da rede neural. Isso torna o procedimento de tolerante às falhas. No entanto, o meio de aprendizado utilizado para o reconhecimento do padrão de conservação dos fatores sigma 70 a partir dos candidatos selecionados pelos algoritmos foi a rede neural artificial pela sua facilidade de aprendizagem e precisão na resposta obtida.

1.7.1 Estudo do reconhecimento de padrões

O reconhecimento de padrões é um algoritmo incumbido na identificação de certas estruturas nos dados de entrada em comparação a estruturas conhecidas e sua posterior classificação dentro de categorias, sendo o grau de associação maior entre estruturas de mesma categoria e menor entre as categorias de estruturas diferentes.

Desta forma os dados de entrada são medidos por sensores e selecionados DE acordo com o conteúdo de informações relevantes para a decisão, e passam por um processo de redução de sua dimensionalidade para que possam ser usados por um classificador, que o designará à classe que melhor o represente.

Segundo TOU e GONZALES (1981), o estudo do reconhecimento de padrões pode ser dividido em duas categorias básicas:

1) Estudo de todos organismos vivos, visando estabelecer os modos pelos quais os mesmos desenvolvem e aprimoram suas capacidades de reconhecimento de padrões;

2) Desenvolvimento e/ou aplicações de teorias e técnicas, visando a construção de máquinas capazes de apresentar características semelhantes a dos seres humanos em reconhecimento de padrões.

1.7.2 Conceito de Padrão e classe

De acordo com Tou e Gonzáles (1981), podemos definir conceitos básicos de padrão e classe da seguinte forma:

A) Padrão: são propriedades que possibilitam o agrupamento de objetos semelhantes dentro de uma determinada classe ou categoria, de acordo com a interpretação dos dados de entrada, consentindo desta forma, a extração das características condescendentes destes objetos;

B) Classe: a classe de um padrão pode ser definida como um conjunto de atributos comuns aos objetos de estudo.

1.7.3 Fases de um sistema para reconhecimento de padrões

Segundo MARQUES, (1999), podemos dividir um sistema para reconhecimento de padrões em 3 grandes fases. Estas fases são descritas da seguinte forma:

1) Representação dos dados de entrada e sua mensuração: essa etapa refere-se à representação dos dados de entrada que podem ser mensurados a partir do objeto a ser estudado. Essa mensuração descreve os padrões característicos do objeto, possibilitando a sua posterior classificação em uma determinada classe. O vetor que caracteriza perfeitamente um objeto seria de dimensionalidade infinita, descrito por um vetor: $Z = (z1, z2, z3, z4, zN)$, onde $(z1, z2, z3,...,zN)$ são suas características.

2) Extração das características: essa etapa consiste na extração de características intrínsecas e atributos do objeto e consequente redução da dimensionalidade do vetor padrão. A escolha das características é de fundamental importância para um bom desempenho do classificador. Esta escolha é feita objetivando os fenômenos que se pretendem classificar. Exige-se, portanto, um conhecimento específico sobre o problema em estudo. Nesta etapa, os objetivos básicos são: a redução da dimensionalidade do vetor característico, sem que isso cause perda de informação inerente a classificação, visando a redução do esforço computacional e a seleção das características significativas para a tarefa de classificação.

3) Classificação do objeto em estudo: essa etapa de reconhecimento de padrões envolve a determinação de procedimentos que permitam a identificação e classificação do objeto em uma determinada classe de objetos. De modo distinto da segunda etapa, a concepção do classificador pode ser abrangida abstratamente e independente da natureza do problema, tendo em vista que os métodos usados em reconhecimento de voz, análise de imagens, identificação de caracteres, entre outros, são muitas vezes os mesmos, possibilitando a aplicação dessas técnicas em contextos variados, sem perda de sua eficiência (MARQUES, 1999).

O Extrator de Características tem como função determinar e extrair as características mais significativas que contribuam para a descrição do objeto, dentre as infinitas características que possam descrevê-lo. Outra informação importante é que o extrator de características sofre uma determinada variação, de acordo com o sistema a ser analisado.

O quadro 4, a seguir, exemplifica várias tarefas de classificação, propostas por um sistema de reconhecimento de padrões, com seus dados de entrada e respectivos dados de saída (MARQUES, 1999).

QUADRO 4 - EXEMPLOS DE TAREFAS DE CLASSIFICAÇÃO.

TAREFAS DE CLASSIFICAÇÃO	**DADOS DE ENTRADA**	**DADOS DE SAÍDA**
Reconhecimento de espécies	Conjunto de espécies	Identificação da espécie
Diagnósticos Médicos	Sintomas	Identificação da Patologia
Previsão do tempo	Mapas atmosféricos	Chuva, sol, vento, etc...
Reconhecimento de sequências de DNA	Sequências de DNA padronizadas	Identificação de sequências de DNA

Quadro apresentando exemplos de tarefas de classificação e suas respectivas aplicações.

FONTE: ADAPTADO DE MONTEIRO A. A, 2002.

Uma vez extraídas as características é necessário a classificação do objeto. Esta classificação pressupõe a designação do objeto a uma determinada classe, dentre as várias que se apresentam. Nesta etapa o classificador "aprende" a distinguir dentre as classes, aquela à qual o objeto pertence. Os padrões de uma mesma classe tendem a se aglomerar, compondo os agrupamentos (MARQUES, 1999).

Quando o treinamento do classificador exigir amplo conhecimento da estrutura estatística dos padrões a serem estudados e o padrão de entrada for apontado como membro de uma classe pré-definida pelos padrões de treinamento, o classificador será chamado de Classificador Paramétrico e a classificação se processa de forma supervisionada. Entretanto, quando o classificador utilizar determinado modelo estatístico, sofrendo ajustes mediante processos adaptativos e a associação entre padrões se fizer com base em similaridades dos padrões de treinamento, o classificador será chamado de Classificador Não Paramétrico e a classificação se processará de forma não supervisionada (MARQUES, 1999).

A maior dificuldade no desenvolvimento de um projeto de reconhecimento de padrões está exatamente na escolha da técnica adequada para que as fases do

reconhecimento de padrões ocorram de modo a representar satisfatoriamente os fenômenos do mundo real. Este presente trabalho utiliza o Classificador Paramétrico. A figura 7 mostra as distintas fases do reconhecimento de padrões.

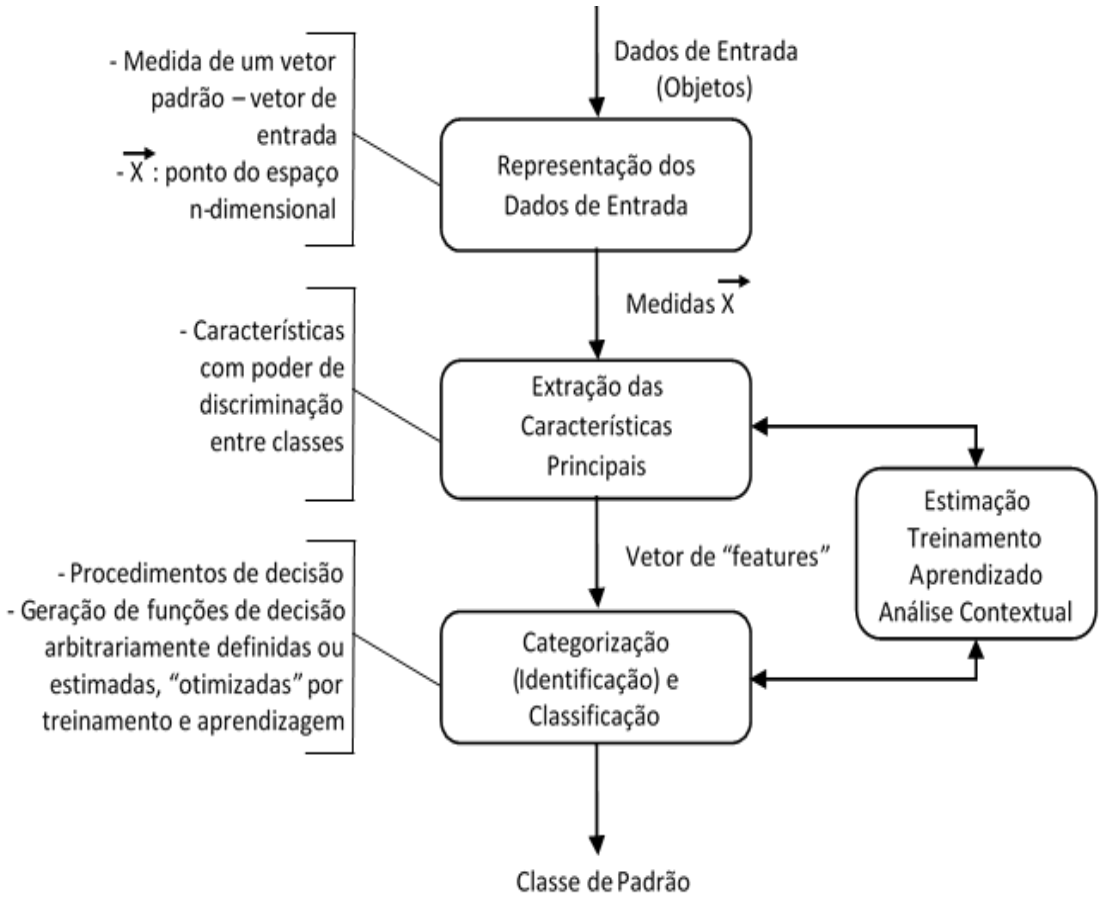


FIGURA 7 - FASES DO RECONHECIMENTO DE PADRÕES.

Ilustração, de forma mais detalhada, das diversas fases do reconhecimento de padrões desde a representação dos dados de entrada até categorização e padronização dos dados atribuídos para classe de padrão.

FONTE: MONTEIRO A. A, 2002.

1.7.4 As propriedades comuns *("feature matching")*

A principal característica das propriedades comuns *("feature matching")* está relacionada com o extrator das características do sistema, sendo esta abordagem responsável pelo bom desempenho do sistema de reconhecimento de padrões. Quando

são determinadas a partir de uma amostra, todas as características de um padrão de classe, o processo de reconhecimento remetese simplesmente em estabelecer comparações com os novos objetos submetidos para análise. Porém é extremamente difícil, identificar todas as características determinantes de uma classe de padrão. A utilização deste conceito implica frequentemente no desenvolvimento de técnicas que permitam aperfeiçoar e otimizar a extração das características dos objetos em estudo (TOU E GONZÁLES, 1981).

Tendo em vista propriedades comuns, a caracterização de padrões é efetivada de acordo com algumas "características principais" pertinentes aos elementos desta classe (TOU e GONZÁLES, 1981). Os padrões pertencentes a uma mesma classe possuirão propriedades comuns de discriminação dessa classe. Desta forma, suas características são extraídas e comparadas com aquelas armazenadas como discriminantes das classes, ocorrendo sempre que um desconhecido padrão é identificado pelo sistema. O classificador então, "classificará" este novo padrão em uma das classes existentes ou então designará o objeto a uma nova classe.

1.7.5 Sistema de reconhecimento de padrões supervisionado

O sistema aprende a reconhecer padrões por meio de esquemas de adaptação quando estes padrões representativos em cada classe estão disponíveis. Assim um sistema de reconhecimento de padrões supervisionado consiste na disponibilidade de "padrões de treinamento" e de um "procedimento de aprendizado" (TOU e GONZÁLES, 1981). A poderosa ferramenta perceptron, gradiente, erro quadrático mínimo e funções potenciais são alguns exemplos de algoritmos utilizados no reconhecimento de padrões supervisionados.

1.7.6 Classificação de padrões baseada em verossimilhança

A classificação por distância de funções é um dos primeiros conceitos em reconhecimento automático de padrões. Esta técnica uma boa ferramenta para a solução

de problemas em que cada padrão de classe apresenta um grau de variabilidade limitado, como por exemplo a identificação de impressões digitais através da leitura biométrica.

A verossimilhança compõe uma abordagem apropriada para o problema de classificação. A classificação por verossimilhança estabelece a distância entre um padrão "y" de classificação desconhecida em relação ao protótipo de cada classe e nomeia o padrão à classe que se encontra mais próxima (TOU E GONZÁLES, 1981).

1.7.7 Técnica de extração de características

A extração de características se tornou uma aliada importante no desenvolvimento de aplicações pertinentes a área de Bioinformática, deixando de ser um simples exemplo ilustrativo (SAEYS, 2007).

Sendo uma técnica utilizada para o aprendizado de máquina, a extração de características pode ser compreendida como qualquer medição útil extraída no processo de identificação de um determinado padrão. A extração de características pode ser de modo simbólico, numérico ou ambos ao mesmo tempo, sendo as variáveis apresentadas de forma contínua ou discreta (SOUZA, 1999).

Nesta técnica é escolhido um subconjunto das funcionalidades disponíveis para a aplicação destes a um algoritmo de aprendizagem, sendo que, o melhor conjunto é sempre aquele que revela uma melhor precisão com uma menor quantidade de dimensões (SEWELL, 2007).

De acordo com Sewell, (2007), podem ser utilizadas as seguintes abordagens para extração de características:

A) Forward selection: este processo é iniciado sem nenhuma variável e elas são adicionadas uma a uma, com o erro minimizado a cada etapa de diminuição do erro. Quando o erro não é minimizado de forma significativa o processo é interrompido.

B) Backward selection: inicialmente neste processo estão presentes todas as variáveis, sendo removidas uma a uma para que o erro seja minimizado até ser estabilizado, ainda que uma remoção não minimize o erro de forma significativa.

3 JUSTIFICATIVAS

- A identificação e análise dos genes regulados por proteínas ativadoras de transcrição dependentes do fator sigma 70 permitirá a compreensão da função deste fator na regulação e na expressão de genes em *Herbaspirillum seropedicae* SmR1 e possivelmente em outras proteobactérias.

- A abordagem de candidatos a promotores de transcrição utilizando as redes neurais, busca um padrão mais confiável na predição de promotores de transcrição dependentes do fator sigma 70 em *Herbaspirillum seropedicae* SmR1.

4 OBJETIVOS

4.1 Objetivos gerais

- Identificar e propor um padrão de conservação nas bases reconhecidas pela holoenzima RNA-polimerase associado ao fator sigma no genoma da proteobactéria *Herbaspirillum seropedicae* SmR1.

- Desenvolver uma metodologia automatizada baseada em reconhecimento de padrões através de treinamento de redes neurais artificiais para a identificação de regiões promotoras de transcrição dependentes do fator sigma 70 em genomas de bactérias.

4.2 Objetivos específicos

- Identificar o padrão de conservação de bases reconhecido pela holoenzima formada em ligação com o fator sigma 70;

- Desenvolver um programa computacional para a busca de candidatos às sequências de regiões promotoras com padrão de conservação do fator sigma 70;

- Introduzir uma metodologia para tomada de decisão na identificação de regiões promotoras reguladas pelo fator sigma 70, através de técnicas de reconhecimento de padrões e classificação por redes neurais artificiais;

- Localizar no genoma de *Herbaspirillum seropedicae* SmR1 as sequências de DNA válidas dependentes do fator sigma 70, identificando possíveis genes regulados por este fator;

- Propor um consenso relacionados às proporções de bases nitrogenadas que compõem as regiões promotoras de transcrição dependentes do fator sigma 70 em *Herbaspirillum seropedicae* SmR1.

5 MATERIAL E MÉTODOS

Com exceção do software MATLAB® e do MICROSOFT EXCEL, os materiais utilizados neste trabalho são disponibilizados gratuitamente na Rede Internet. A longo da descrição dos materiais, será fornecida todas as fontes e endereços eletrônicos onde estes materiais estão disponibilizados e como estes foram empregados neste presente trabalho.

5.1 Matlab

Para este trabalho foi utilizada uma ferramenta para o desenvolvimento de cálculos científicos denominada Matlab, cuja palavra, é proveniente da língua inglesa, significando 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).

O MATLAB® conta com diversos tipos de bibliotecas que contém várias funções desenvolvidas pela Mathworks, como por exemplo, a Toolbox de Bioinformática já inclusa na versão R2012a. Em complemento a todas estas bibliotecas disponibilizadas pela desenvolvedora, programadores espalhados pelo mundo fazem o desenvolvimento de novas funções complementares, como o caso da Toolbox de Bioinformática-UFPR, para o acréscimo das atividades utilizando o MATLAB®.

5.2 Microsoft Excel

O Microsoft Excel® foi 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 Microsoft Excel® possibilita a análise, o gerenciamento e o compartilhamento de informações ajudando a tomada de decisão (MICROSOFT, 2011).

5.3 Rede Neural MLP (Multilayer Perceptron)

A rede neural artificial utilizada neste trabalho é uma rede MLP (Multilayer Perceptron). O treinamento da rede foi realizado em MATLAB® utilizando comandos previamente gravados e disponibilizados pelo grupo colaborativo do Laboratório de Bioinformática - UFPR. A qualidade da rede neural está diretamente relacionada ao número de possíveis candidatos ao fator de transcrição dependente do sigma 70 adquiridos no banco de dados biológicos RegulonDB.

As redes neurais MLP são redes neurais, cuja a função é mapear os conjuntos de entradas de dados em conjuntos de saída apropriados. Desta forma as redes MLPs vem

apresentando êxito em aplicações nas mais diversas áreas, como no reconhecimento de padrões e processamento de sinais.

Uma rede do tipo MLP é composta por um conjunto de nós fonte, compondo 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, 1999).

A rede MLP apresenta característica progressiva, tendo em vista que a saída de uma camada alimenta exclusivamente a entrada da próxima camada sem a presença de realimentação, desta forma o sinal se propaga através da rede de forma progressiva (HAYKIN, 1999). O algoritmo de treinamento de *backpropagation*, é outra característica da rede MLP. Este algoritmo é baseado na heurística de aprendizado por correção do erro, sendo este erro retro propagado através das camadas de saída para as camadas ocultas.

5.4 Artemis

Para a visualização dos resultados obtidos a partir da execução da aplicação desenvolvida neste trabalho foi utilizado o programa ARTEMIS®, desenvolvido pelo instituto Sanger. Este software 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 diversos 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). O programa ARTEMIS® pode ser gratuitamente adquirido para diversas plataformas no portal do NCBI ou através deste endereço eletrônico: http://www.sanger.ac.uk/resources/software/artemis/

5.5 NCBI - Centro Nacional de Informações sobre Biotecnologia

O Centro Nacional de Informações sobre Biotecnologia (NCBI) disponibilizado no seguinte endereço eletrônico http://www.ncbi.nlm.nih.gov/ , é parte da Biblioteca Nacional dos Estados Unidos de Medicina (NLM), uma filial do National Institutes of Health. O NCBI está localizado em Bethesda, Maryland, e foi fundada em 1988 por meio de legislação patrocinada pelo senador Claude Pepper. O NCBI abriga genomas sequenciados e disponíveis no formato fasta ou GenBank, além de um vasto índice de artigos de pesquisa biomédica na PubMed Central e PubMed, bem como outras informações relevantes para a biotecnologia. Todos esses bancos de dados estão disponíveis on-line através do motor de busca Entrez.

5.6 RegulonDB

RegulonDB - Centro de Ciências Genômicas – UNAM, disponibilizado no seguinte endereço eletrônico http://regulondb.ccg.unam.mx/ é o banco de dados biológicos da rede de regulação especializado em um organismo de referência primária *Escherichia coli* K-12. A empresa tem expandido seu contexto biológico para que a regulação da transcrição seja parte de uma unidade que inicia com o sinal e continua com a transdução de sinal para o núcleo de regulação, modificando a expressão dos genes-alvo afetados e responsáveis pela resposta (COLLADO-VIDES, MAGASANIK E GRALLA, 1991).

Esse banco de dados biológicos fornece informações com curadoria de organização e regulação do gene em *E. coli.* A informação corrente é fornecida sobre o gene ou operon. Futura expansão irá incluir informações sobre a regulação para além da iniciação da transcrição. O RegulonDB tem informações sobre as unidades de transcrição e os detalhes mecânicos de regulação das referidas unidades incluindo: promotores e fatores sigma, terminadores e regulons. Além disso, os sítios de ligação do ribossomo e reguladores da transcrição estão incluídos (COLLADO-VIDES, MAGASANIK E GRALLA, 1991).

Há também informações sobre o produto do gene. Inicialmente é recolhido e analisado um grande grupo de promotores e sua regulação em *E. coli*, sendo publicado

no início dos anos 90 (COLLADO-VIDES, MAGASANIK E GRALLA, 1991). Alguns anos mais tarde, uma versão expandida (GRALLA E COLLADO-VIDES, 1996), converte o conteúdo biológico para a versão electrónica, tal como uma base de dados relacional no portal RegulonDB, onde foi publicado inicialmente em 1998 (Huerta *et al.*, 1998). Alguns anos depois, houve o ingresso na equipe EcoCyc com Monica Riley, Milton Saier e Pedro Karp e a primeira versão de EcoCyc com informações de regulamentação foi publicada em 2002 (KARP *et al.*, 2002). Tal equipe é a verdadeira fonte de curadoria especializada em *Escherichia coli* K-12, alimentando tanto a base de dados biológicos do RegulonDB quanto o EcoCyc.

5.7 Hardware

Foi utilizado um computador fornecido pela Universidade Federal do Paraná - Laboratório de Bioinformática.

QUADRO 5 - CONFIGURAÇÕES DE HARDWARE.

MODELO	Lenovo ThinkCentre M90P
PROCESSADOR	Core i5 650 (3,2Ghz)
DISCO RÍGIDO	320Gb
SISTEMA OPERACIONAL	Windows®

Quadro apresentando as configurações de hardware utilizado neste trabalho.

FONTE: O AUTOR, (2014).

A metodologia adotada consiste em um treinamento de rede neural artificial utilizando dados biológicos de sequências de promotores descritas e disponibilizadas em banco de dados biológicos especializados e a incorporação dos dados da rede treinada em um algoritmo especifico, adequado para tarefa de predição computacional de possíveis candidatos a promotores de transcrição dependentes do fator sigma 70 em *Herbaspirillum* e outras proteobactérias.

5.8 Algoritmo de busca utilizado

O algoritmo utilizado para a predição de regiões promotoras dependentes do fator de transcrição sigma 70 recebeu o nome de S70FINDER e o *script* utilizado, bem como as funções, são encontradas nos apêndices.

A funções utilizadas na construção deste algoritmo não foram criadas exclusivamente para os testes realizados, apenas adaptadas para o caso em questão, tendo em vista que estas funções já existiam anteriormente, armazenadas na biblioteca de funções em MATLAB® mantidas e disponibilizadas pelo grupo de Bioinformática-UFPR. O resultado da modelagem deste algoritmo, aliado ao conjunto de dados de treinamento disponibilizados pela rede neural, deu origem a uma poderosa ferramenta computacional, utilizada na identificação de regiões promotoras de transcrição dependentes do fator sigma 70, introduzindo desta forma uma nova metodologia capacitada na predição computacional de promotores sigma 70 em *Herbaspirillum seropedicae* SmR1 e possivelmente em trechos disponíveis de genomas pertencentes a outras proteobactérias.

A vantagem deste algoritmo em relação a outros é a rapidez no processamento e a praticidade no retorno das informações, tendo em vista que este algoritmo realiza a predição de promotores sigma 70 com rapidez em pequenos trechos do DNA, levando menos de 7 minutos para processar genomas completos, proporcionando uma maior comodidade ao usuário desta ferramenta.

Para execução deste algoritmo, é necessário fornecer uma sequência genômica no formato GenBank® (*.gb), disponibilizado gratuitamente através do portal eletrônico do Centro Nacional de Informações sobre Biotecnologia (NCBI), referente à proteobactéria a ser analisada pelo programa.

Importantes informações como posição no mapa genômico, número de bases, sentido de leitura de sequência e pontuação dos possíveis candidatos a promotores sigma 70, ficam armazenadas em um arquivo (*.gb), gerado ao final do processo, sendo considerado como dado de saída de toda a leitura realizada pelo algoritmo. Para visualização e análise dos dados obtidos, é necessário carregar os arquivos referentes ao dado de saída e a sequência genômica escolhida no software ARTEMIS®.

5.9 Conjunto de dados

Para realização deste trabalho foi de extrema importância à obtenção e seleção de dados pertinentes a sequências de regiões promotoras de transcrição dependentes do sigma 70. Os dados coletados referentes a promotores sigma 70, foram salvos como arquivo de texto, com a extensão (*.txt), pois possui fácil leitura em funções desenvolvidas especificamente para mineração de dados. Essas funções foram utilizadas na estruturação do algoritmo incumbido no treinamento e aprendizagem de rede neural artificial. Para obter um bom treinamento da rede neural foi necessário ter uma quantidade elevada de dados referente ao problema em questão. Estes dados devem ser confiáveis para que não sejam induzidos erros ou vícios no treinamento da rede neural artificial.

Atualmente, a quantidade de dados disponibilizados referente a sequências de regiões promotoras dependentes do fator sigma 70 em *Herbaspirillum seropedicae* SmR1 é relativamente inferior à quantidade de dados disponibilizados, referente a sequências de regiões promotoras dependentes do fator sigma 70 em *Escherichia coli*. Portanto, para se obter um bom treinamento de rede neural artificial, foram utilizados somente dados de regiões promotoras dependentes do fator sigma 70 encontrados e mapeados em genoma da proteobactérias *Escherichia coli* K12. O banco de dados com maior número de regiões promotoras catalogados para esta proteobactérias foi encontrado no site RegulonDB.

5.10 Primeiro treinamento e aprendizagem de rede neural artificial

Inicialmente foram utilizados um total de 4.341 sequências de promotores sigma 70, da família sigma 70, obtidas no portal eletrônico do RegulonDB. Essa quantidade de dados de promotores, com inferência computacional, sem supervisão humana, gerou um número elevado de candidatos a promotores sigma 70 considerados falsos positivos.

No primeiro treinamento de rede neural artificial foram utilizadas seis características, cada característica representa uma base nitrogenada exclusiva de DNA. O conjunto dessas seis bases representam o hexâmero da região promotora -35 (TTGACA). Para cada hexâmero da lista de possíveis candidatos a promotores sigma 70, foi gerado aleatoriamente, através de função especifica, hexâmeros artificiais, representando os falsos candidatos a promotores sigma 70. Para o conjunto dessas características foram

atribuídas duas classes. A primeira classe representa os verdadeiros hexâmeros para as região -35 e a segunda classe representa os falsos hexâmeros para as região -35.

5.11 Segundo treinamento e aprendizagem de rede neural artificial

Para este segundo treinamento de rede neural artificial, foram utilizadas seis características, cada característica representa uma base nitrogenada exclusiva de DNA. O conjunto dessas seis bases representam o hexâmero da região promotora -10 (TATAAT). Para cada hexâmeros da lista de possíveis candidatos a promotores sigma 70, foi gerado aleatoriamente, através de função especifica, hexâmeros artificiais, representando os falsos candidatos a promotores sigma 70. Para o conjunto dessas características foram atribuídas duas classes. A primeira classe representa os verdadeiros hexâmeros para as região -10 e a segunda classe representa os falsos hexâmeros para as região -10.

5.12 Terceiro treinamento e aprendizagem de rede neural artificial

Para realizar o terceiro treinamento foram utilizadas 820 sequências de candidatos a promotores sigma 70 depositadas e submetidas à curadoria do RegulonDB, baseada em supervisão humana, ou seja, candidatos a promotores sigma 70 analisados e validados. Para cada um dos 820 possíveis candidatos selecionados foi designado um falso candidato, derivado do resultado do primeiro e do segundo treinamento da rede neural artificial, totalizando 820 falsos positivos.

O critério principal adotado para a seleção manual destes falsos candidatos está relacionado com a distância entre o promotor e o gene, e o número de pares de bases entre as regiões promotoras -35 e -10. Promotores que apareceram em regiões gênicas foram considerados como falso positivo, pois não possuem sentido biológico, já que devem estar dispostos em regiões intergênicas, a montante dos genes que codificam para um produto final (proteínas ou RNA).

Com esses novos dados inseridos no terceiro treinamento, houve uma diminuição do número de falsos positivos, aumentando a capacidade de a rede neural

artificial de reconhecer um determinado padrão, com forte potencial de identificação de sequências promotoras dependentes do fator sigma 70. No 3° treinamento foram atribuídas doze características que representam a combinação dos dois hexâmeros que compõem a região -35 e -10. Desta forma foram reduzido os casos onde apareceram elevadas repetições de hexâmeros individualizados ao longo do genoma bacteriano, tendo em vista que a principal característica de um promotor dependente do fator sigma 70 é ser representado pelos dois hexâmeros que compõem a região -35 e região -10. Para estas doze características foram atribuídas novamente duas classes. A primeira classe representa os verdadeiros hexâmeros das regiões -35 e -10 unidas e a segunda classe representa os falsos hexâmeros das regiões -35 e -10 unidas.

Foi estabelecido uma linha de corte para as regiões promotoras que apresentaram pares de bases que não obedeceram ao limite mínimo de 12 pares de bases e o limite máximo de 20 pares de bases entre as regiões -35 e -10. Desta forma foi obtido o melhor treinamento da rede neural artificial baseando-se em regiões promotoras completas, incluindo os pares de bases localizados entre as regiões BOX -35 e BOX -10 dos promotores de transcrição dependentes do fator σ^{70}. Para os pares de bases entre as sequências que compõe o BOX -35 e BOX -10 dos promotores sigma 70 não temos um padrão de conservação descrito ou preestabelecido através dos testes realizados neste trabalho, podendo ter estas sequências padrão variável.

Para melhor compreensão do trabalho, foi elaborado um fluxograma das etapas mais relevantes deste projeto desde a obtenção dos dados referentes aos promotores de transcrição dependentes do fator sigma 70, obtidos no portal eletrônico do RegulonDB, até os dados referentes à sequência genômica da proteobactéria *Herbaspirillum seropedicae* SmR1, obtida no portal eletrônico do NCBI.

A figura 8 mostrada a seguir representa as etapas da construção do algoritimo.

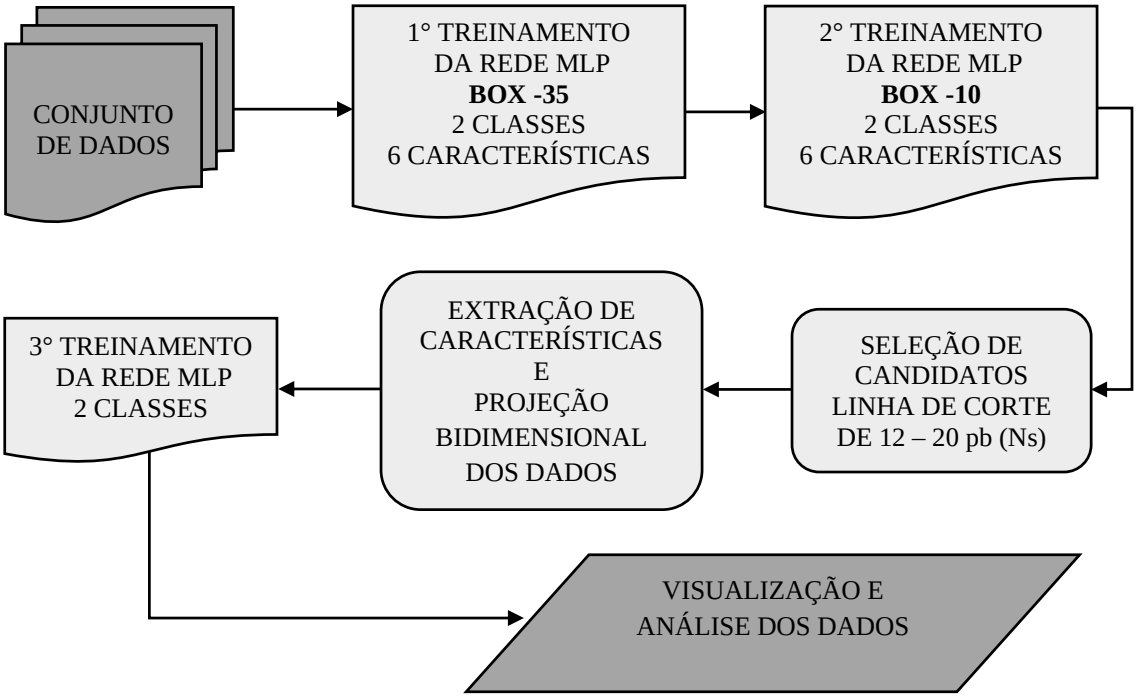


FIGURA 8 - ETAPAS DA CONSTRUÇÃO DO ALGORITIMO.

No fluxograma acima podemos observar as etapas da construção da ferramenta computacional utilizada na predição de regiões promotoras de transcrição em *H. seropedicae* SmR1 desde a obtenção dos dados de entrada até a análise dos dados de saída obtidos através do processamento do programa.

FONTE: O AUTOR, (2014).

6 RESULTADOS

Os resultados obtidos através da rede neural artificial treinada, apontam 4.498 candidatos a promotores de transcrição dependentes do fator sigma 70 em *H. seropedicae* SmR1. Destes 4.998 candidatos foram selecionados 288 candidatos entre as regiões intergênicas dos 100 genes mais expressos em *H. seropedicae* SmR1, segundo BARBOZA, (2014). Destes 288 candidatos, separamos uma listagem com 100 promotores com maio valor de pontuação, de acordo com os testes de verossimilhança. Estes 100 candidatos estão localizados a montante dos 100 genes mais expressos em *H. seropedicae* SmR1 e podem ser facilmente visualizados no programa Artemis.

6.1 Listagem de candidatos a promotores sigma 70

A tabela 1 a seguir representa um subgrupo retirado da listagem final obtida (em anexo), contendo dez sequências de candidatos a promotores de transcrição dependentes do fator sigma 70 identificados pela rede neural e os respectivos genes, mapeados ao longo do genoma da betaproteobactéria *Herbaspirillum seropedicae* SmR1.

TABELA 1 - CANDIDATOS A PROMOTORES SIGMA 70 ENCONTRADOS NO GENOMA DE *Herbaspirillum seropedicae* SmR1.

Cand.	Box -35	Ns	Nº (Ns)	Box -10	GC%	Pont.	Gene Dow.	Gene Up.
1	Ttggaa	tggcttcttataccgc	16	aagaat	39.29	20,3	Hsero_0068	cdsA
2	Ttgaca	gtctagagatgttcctc	17	tatagt	37.93	21,2	tufB	ampG
3	Ttgaca	gtctagagatgttcctc	17	tatagt	37.93	21,2	secE	ampG
4	Atcaat	aaggagccgtcatggca	17	aagaag	44.83	17,7	rplJ	nusG
5	Atcaat	aaggagccgtcatggca	17	aagaag	44.83	17,7	rplL	nusG
6	Ttgcca	acgaagagcgaatctcc	17	tatcat	44.83	20,9	fusA	rpoC
7	Ttgact	tccgcgccgcccacctttt	18	tactat	53.33	20,0	Hsero_0326	Hsero_0344
8	Ttgacc	gagcagctaagtgcctg	17	tataat	44.83	21,5	rplU	Hsero_0361
9	Ttgcat	cctccctaacccgccac	17	tagaat	51.72	21,0	nrdB	Hsero_0378
10	Tttcgt	catcttgtttgtggtggtg	19	tacaat	38.71	19,6	cspC	Hsero_0503

A primeira coluna está relacionada ao número do possível candidato a promotor sigma 70 encontrado. A segunda coluna se refere a região -35 do promotor sigma 70 também chamado de BOX -35. A terceira coluna, refere-se as bases indeterminadas (Ns) entre as regiões -35 e -10 do promotor. A quarta coluna se refere a quantidade de pares de bases encontradas em Ns. A quinta coluna se refere a região -10 do promotor sigma 70 também chamado de BOX -10. A conteúdo GC de cada região promotora completa foi colocada na sexta coluna. A sétima coluna refere-se a uma pontuação baseado na máxima similaridade. Com essa pontuação foi possível selecionar apenas um candidato a promotor para um determinado conjunto de candidatos a promotores preditos para um único gene, tendo em vista que os demais candidatos a promotores não foram descartados, sendo relacionados em outra listagem de resultados mais detalhada. A oitava coluna se refere ao gene dowstream localizado logo abaixo do possível candidato a promotor sigma 70, ou seja, o nome do gene no qual o candidato a promotor sigma 70 está relacionado no momento da transcrição. A nona coluna faz referência ao gene upstream, ou seja, o nome do gene que está acima do candidato a promotor sigma 70, fazendo parte da vizinhança do gene dowstream.

FONTE: O AUTOR, (2014).

Os dados bioestatísticos foram obtidos através uma amostragem de 100 genes mais expressos, segundo BARBOSA (2014), e seus possíveis candidatos listados podem ser observados na tabela 2 abaixo:

TABELA 2 - DADOS ESTATISTICOS DOS CANDIDATOS À PROMOTORES SIGMA 70 EM GENOMA DE *Herbaspirillum seropedicae* SmR1.

	MÉDIA	**MEDIANA**	**MODA**	***D. P**	**MÁXIMO**	**MÍNIMO**
% GC	40,6	40,7	44,8	10,1	65,6	21,2
N° (Ns)	17,6	17,0	17,0	1,2	20,0	16,0
PONTUAÇÃO	20,0	20,1	20,1	0,9	21,6	17,7

% GC - Conteúdo GC; N° (Ns) - Número de pares de bases entre o box -35 e box -10; PONTUAÇÃO - Valor atribuído para cada candidato baseando-se em verossimilhança; (*) - Desvio padrão.

FONTE: O AUTOR, (2014) BASEADO EM ANÁLISE ESTATÍSTICA DE 288 CANDIDATOS.

A porcentagem do conteúdo GC obteve a média de 40,69% apresentando uma frequência modal de 44,83%. Não foram identificados candidatos à promotor com conteúdo GC abaixo de 19,35% ou acima de 65,63%. O número médio de pares de bases esperadas entre os hexâmeros dos box -35 e -10 descritos em *Escherichia coli* coincidiu com a frequência do número de pares de bases entre os hexâmeros dos box -35 e -10 preditos computacionalmente em *Herbaspirillum seropedicae* SmR1. A pontuação estipulada para os candidatos a promotores sigma 70 indicados pela rede neural artificial nos testes realizados, foi fundamentada com métodos de verossimilhança. As sequências com maior similaridade em relação ao consenso proposto para sequências promotoras sigma 70 em *Herbaspirillum seropedicae* SmR1 e as proporções das bases nitrogenadas visualizadas em cada posição dos hexâmeros, receberam uma pontuação máxima de 21,6. As sequências que apresentaram baixa similaridade em relação ao consenso proposto para as sequências promotoras sigma 70 em *Herbaspirillum seropedicae* SmR1, receberam uma pontuação com valor mínimo de 17,7.

6.2 Análises dos dados através de modelos baseados em regressão logística

Os candidatos a promotores sigma 70 são apresentados com vetor de características bidimensional, proporcionando maior facilidade de visualização dos resultados obtidos.

Ao final do processamento dos dados são plotados gráficos bidimensionais representando o momento da separação entre bons candidatos considerados pela rede neural como verdadeiros e candidatos com baixo potencial considerados pela rede neural artificial como falsos positivos.

Os gráficos gerados para as regiões -35 e -10 serão respectivamente apresentados nas figuras 9 e 10:

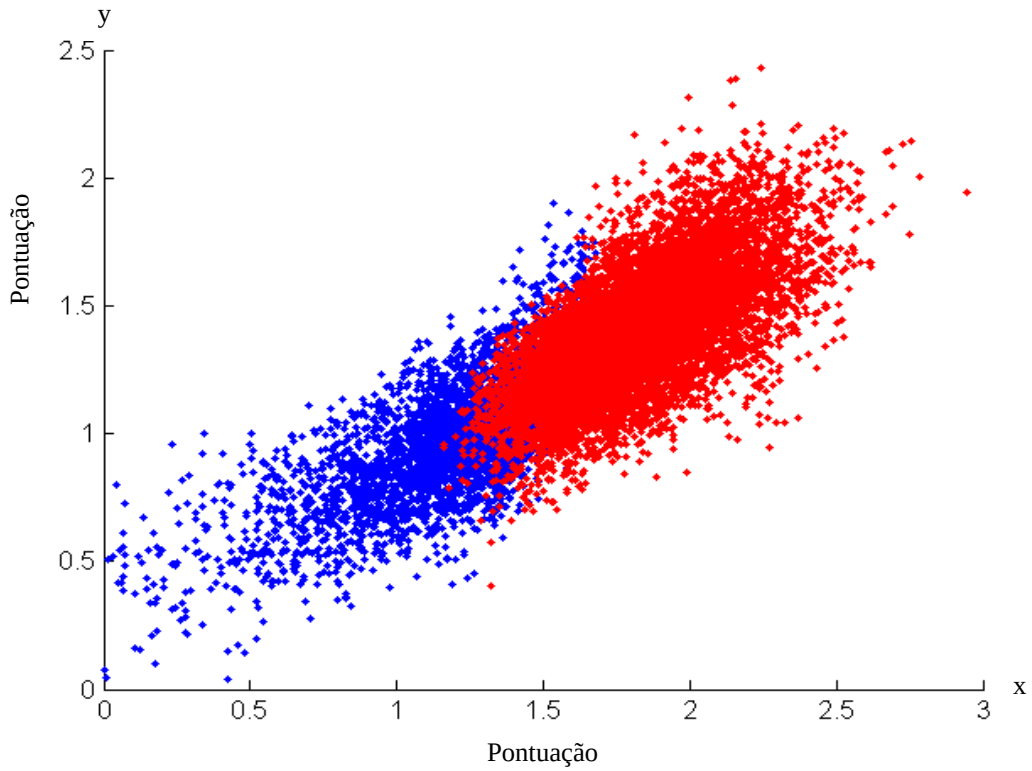


FIGURA 9 - GRÁFICO BIDIMENSIONAL DE CLASSIFICACÃO PARA AS REGIÕES -35 σ^{70} .

O gráfico gerado ao final do processamento possibilita visualizar em 2 dimensões o momento da separação entre os bons candidatos (em azul), e os candidatos com baixo potencial (em vermelho) a promotores de transcrição dependentes do fator sigma 70.

FONTE: O AUTOR (2014), ATRAVÉS DO PROCESSAMENTO DO ALGORITIMO.

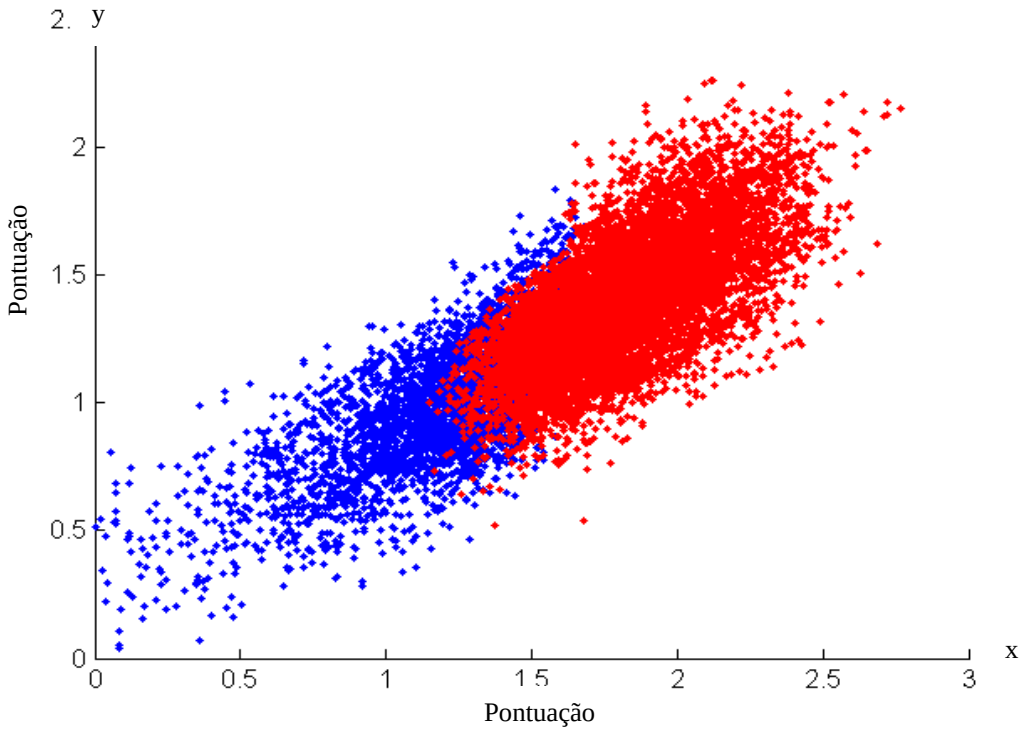


FIGURA 10 - GRÁFICO BIDIMENSIONAL DE CLASSIFICACÃO PARA AS REGIÕES -10 σ^{70}.

O gráfico gerado ao final do processamento possibilita visualizar em 2 dimensões o momento da separação entre os bons candidatos (em azul), e os candidatos com baixo potencial (em vermelho) a promotores de transcrição dependentes do fator sigma 70.

FONTE: O AUTOR (2014), ATRAVÉS DO PROCESSAMENTO DO ALGORITIMO.

6.3 Análises de dados através de modelos baseados em verossimilhança

Os candidatos a promotores sigma 70 foram submetidos a análise baseada em modelos de verossimilhança, com o propósito de obter e atribuir valores de pontuação para os candidatos a promotores sigma 70 mais próximos do consenso proposto em *Herbaspirillum seropedicae* SmR1. Foram selecionados 288 candidatos a promotores sigma 70, situados a montante dos 100 genes mais expressos segundo BARBOSA (2014), "Re-anotação do genoma de *Herbaspirillum seropedicae* SmR1 com dados de transcriptoma por RNA-Seq".

A figura 11 mostra o cálculo utilizado para realização de testes de verossimilhança nos possíveis candidatos a promotores sigma 70 em *H. Seropedicae*.

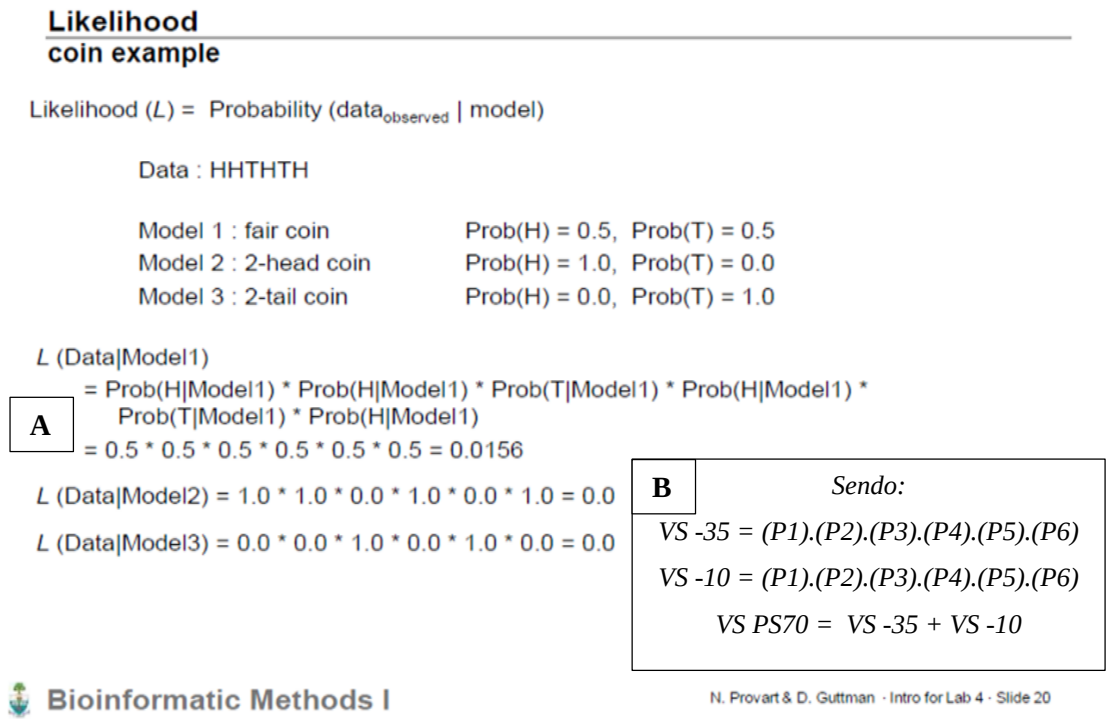


FIGURA 11 - MÉTODO DE VEROSSIMILHANÇA UTILIZADO.

Em **A** o método mais adequado para calcular a verossimilhança neste caso. Em **B** a adaptação deste cálculo para aplicação nos candidatos a promotores sigma 70 neste projeto.

FONTE: APTADO DE N. PROVART E D. GUTTMAN, BIOINFORMATIC METHODS I - INTRO FOR LAB 4 · SLIDE 20, ACESSADO GRATUITAMENTE EM 22/05/2014, NO SEGUINTE ENDEREÇO ELETRÔNICO https://www.coursera.org/ .

Nesta adaptação, as variáveis *P1*, *P2*, *P3*, *P4*, *P5* e *P6* correspondem as frequências de bases que compõem os hexâmeros das regiões box -35 e box -10 em *Herbaspirillum seropedicae* SmR1. V*S -35* igual ao produto das frequências de bases observadas nas seis posições do hexâmero que compões o box -35. V*S -10* é igual ao produto das frequências de bases observadas nas seis posições do hexâmero que compõem o box -10. V*S PS70* é igual a soma da verossimilhança do box -35 e da verossimilhança do box -10, resultando em um valor utilizado como pontuação para sequências completas de candidatos a promotores sigma 70.

A comparação entre os valores de pontuação para a região -35 em *E. coli* e para a região -35 em *Herbaspirillum seropedicae* SmR1, revelou em sua maioria, um agrupamento destes valores, permitindo uma sincronização destes dados como demonstrado na figura 12 a seguir.

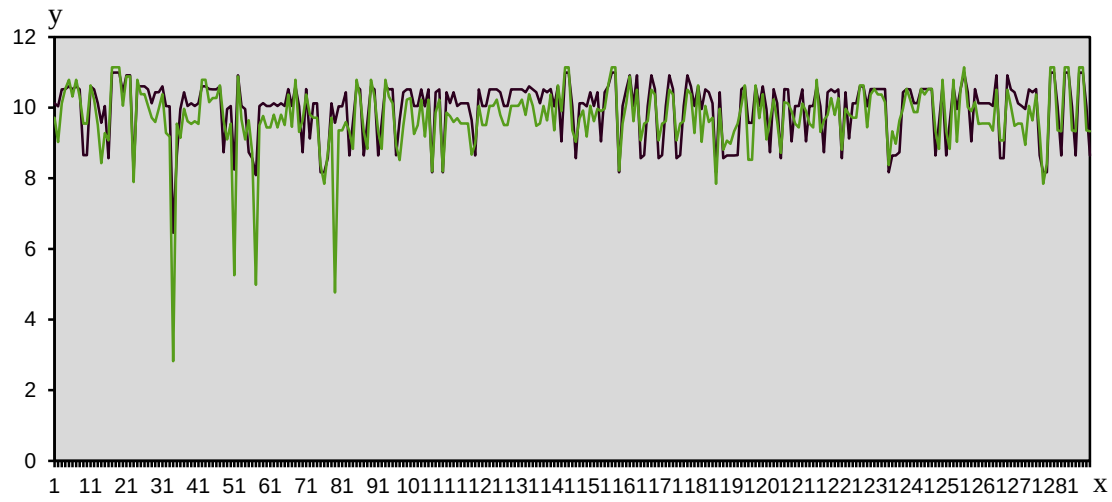


FIGURA 12 - GRÁFICO COMPARATIVO ENTRE AS PONTUAÇÕES PARA A REGIÃO -35 DE PROMOTORES σ^{70} EM *E. coli* E *H. seropedicae SmR1*

O eixo y representa os valores de pontuação para cada região -35 (0 – 12). O eixo x representa a quantidade de candidatos a promotores sigma 70 analisados (1 – 288). Em preto temos as pontuações atribuídas para regiões do box -10 através de uma matriz baseada em proporções de bases conservadas, descritas em *E. coli*. Em verde temos as pontuações atribuídas para regiões do box -10 através de uma matriz baseada em proporções de bases conservadas observadas em *H. seropedicae* SmR1.

FONTE: O AUTOR (2014).

Na figura 13 podemos observar a sobreposição dos valores de pontuação para a região -35.

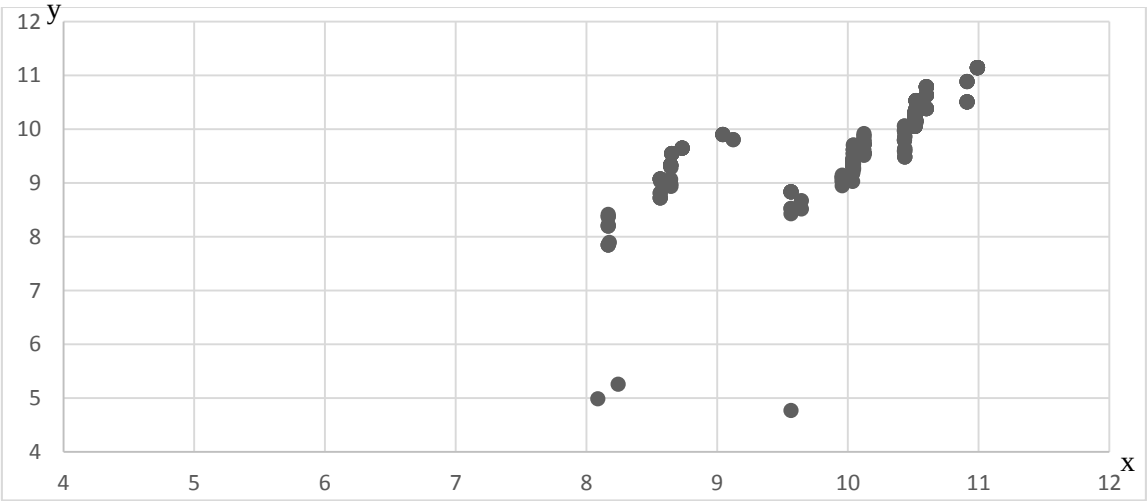


FIGURA 13 - GRÁFICO REPRESENTANDO A SOBREPOSIÇÃO DA PONTUAÇÃO PARA A REGIÃO -35 DE PROMOTORES σ^{70} EM *E. coli* E *H. seropedicae* SmR1.

O eixo x e y representam os valores de pontuação para a região -35 nas duas espécies. A reta em diagonal, representa a aproximação destes candidatos sobrepostos. Nos pontos, temos a sobreposição dos valores de pontuação atribuídos para região -35 em *E. coli* e *H. seropedicae* SmR1.

FONTE: O AUTOR, (2014).

A figura 14 mostra uma comparação entre os valores de pontuação para a região -10 em *E. coli* e a região -10 em *Herbaspirillum seropedicae* SmR1, revelando em muitos picos uma proximidade destes valores.

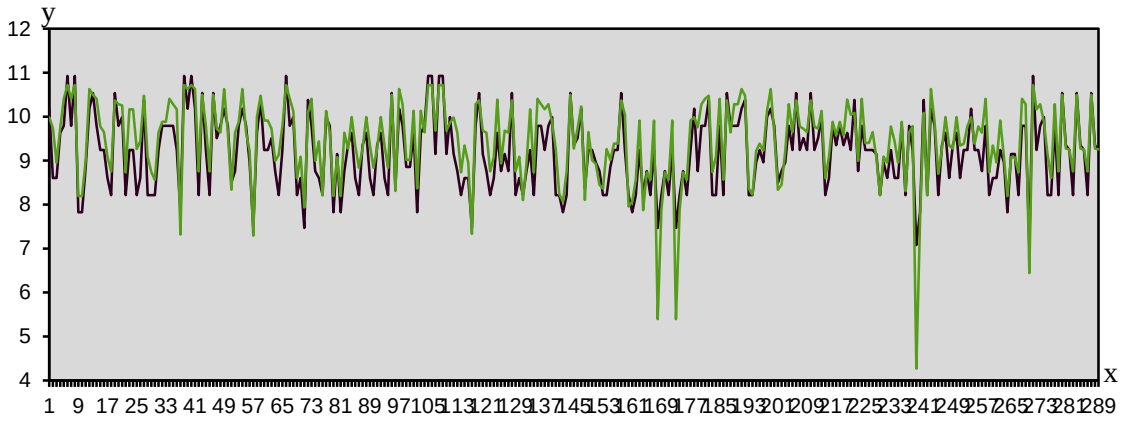


FIGURA 14 - GRÁFICO COMPARATIVO ENTRE AS PONTUAÇÕES PARA A REGIÃO -10 DE PROMOTORES σ^{70} EM *E. coli* E *H. seropedicae* SmR1.

O eixo y representa os valores de pontuação para cada região -10 (0 – 12). O eixo x representa a quantidade de candidatos a promotores sigma 70 analisados (1 – 288). Em preto temos as pontuações atribuídas para regiões do box -10 através de uma matriz baseada em proporções de bases conservadas, descritas em *E. coli*. Em verde temos as pontuações atribuídas para regiões do box -10 através de uma matriz baseada em proporções de bases conservadas observadas em *H. seropedicae* SmR1.

FONTE: O AUTOR, (2014).

Em outro gráfico podemos observar a sobreposição dos valores de pontuação para a região -10, sugerindo um agrupamento destes valores em determinadas regiões do gráfico.

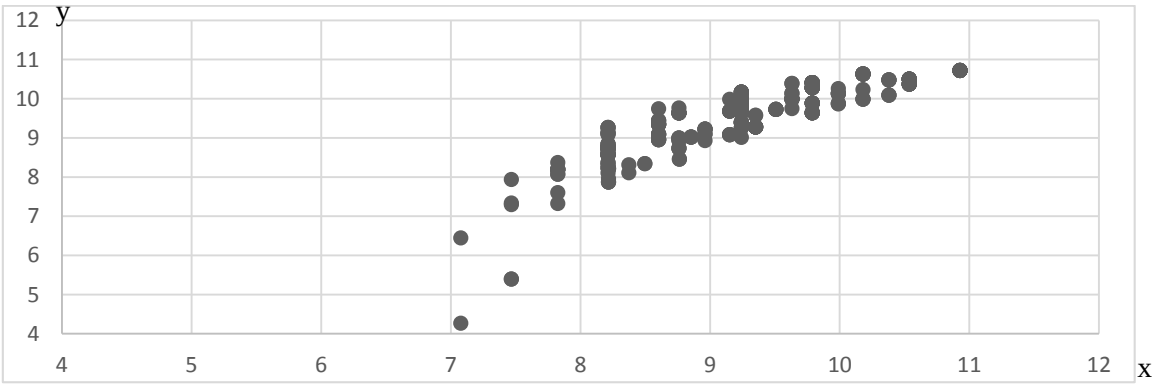


FIGURA 15 - GRÁFICO REPRESENTANDO A SOBREPOSIÇÃO DA PONTUAÇÃO PARA AS REGIÕES -10 DE PROMOTORES σ^{70} EM *E. coli* E *H. seropedicae* SmR1

O eixo x e y representam os valores de pontuação para a região -10 nas duas espécies. A reta em diagonal, representa a aproximação destes candidatos sobrepostos. Nos pontos, temos a sobreposição dos valores de pontuação atribuídos para região -10 em *E. coli* e *H. seropedicae* SmR1.

FONTE: O AUTOR, (2014).

A figura 16 mostra uma comparação entre os valores de pontuação para a região -35 e -10 em *E. coli* e para região -35 e -10 em *Herbaspirillum seropedicae* SmR1, revelando um agrupamento destes valores.

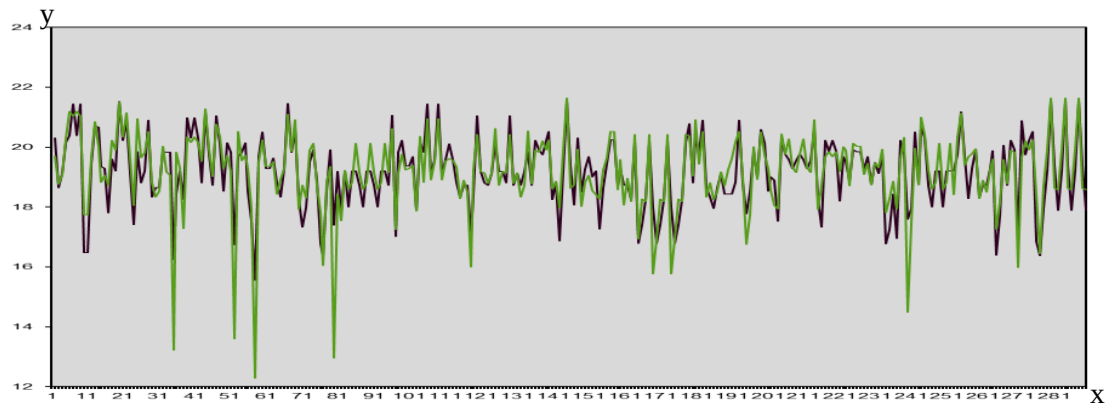


FIGURA 16 - GRÁFICO COMPARATIVO ENTRE AS PONTUAÇÕES PARA AS REGIÕES -35 e -10 DE PROMOTORES σ^{70} EM *E. coli* E *H. seropedicae* SmR1.

O eixo y representa os valores de pontuação para cada candidato (0 – 24). O eixo x representa a quantidade de candidatos a promotores sigma 70 analisados (1 – 288). Em preto temos as pontuações atribuídas para regiões do box -35 e -10 através de uma matriz baseada em proporções de bases conservadas, descritas em *E. coli*. Em verde temos as pontuações atribuídas para regiões do box -35 e -10 através de uma matriz baseada em proporções de bases conservadas observadas em *H. seropedicae* SmR1.

FONTE: O AUTOR (2014).

Em outro gráfico gerado, podemos observar a sobreposição dos valores de pontuação para as regiões -35 e -10, revelando um agrupamento destes valores em determinadas regiões do gráfico, significando a proximidade dos valores de pontuação para as duas espécies.

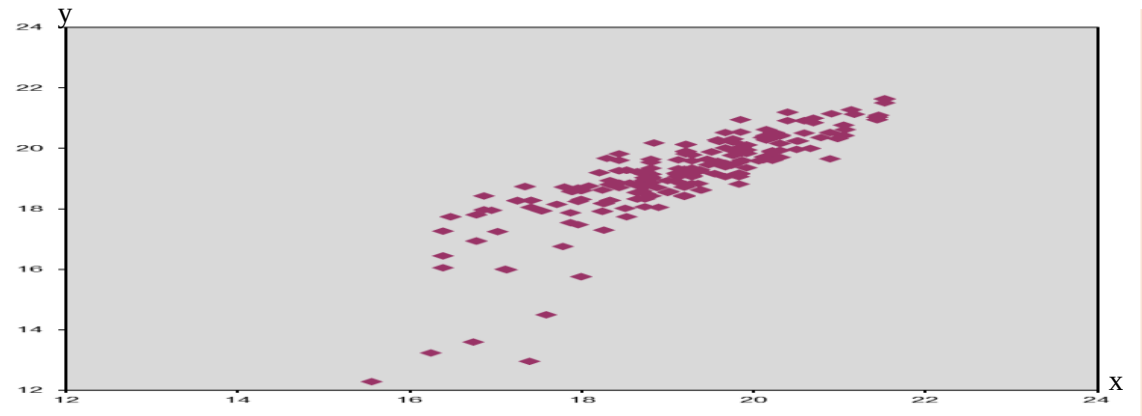


FIGURA 17 - GRÁFICO REPRESENTANDO A SOBREPOSIÇÃO PARA A PONTUAÇÃO DA REGIÃO -35 e -10 DE PROMOTORES σ^{70} EM *E. coli* E *H. seropedicae* SmR1

O eixo x e y representam os valores de pontuação da regiões -35 e -10 em *E. coli* e *H. seropedicae* SmR1. A reta em diagonal, representa a aproximação destes candidatos sobrepostos. Nos pontos, temos a sobreposição dos valores de pontuação atribuídos para as regiões -35 e -10 nas duas espécies.

FONTE: O AUTOR, (2014).

6.4 Consenso Proposto para regiões promotoras sigma 70 em *H. seropedicae* SmR1

Após obtenção da lista composta por candidatos a promotores de transcrição dependentes do fator sigma 70 em *H. seropedicae* SmR1, validados através da análise de regiões intergênicas de genes com alto nível de expressão, podemos propor uma sequência consenso para estes candidatos. Para isso, foi utilizada uma função (ambiente Matlab), desenvolvida pelo grupo de Bioinformática-UFPR, específica para esta tarefa. Essa função tem como objetivo alinhar de forma consensual as letras que representam as bases nitrogenadas de várias sequências de DNA, retornando, como resultado final, um consenso. Neste caso, são sequências que compõem os hexâmeros dos BOX-35 e BOX-10 retirados de uma amostragem de 100 genes mais expressos em *H. seropedicae* SmR1 e seus respectivos candidatos a regiões promotoras de transcrição dependentes do fator sigma 70, obtidos através de treinamento da rede neural artificial, validados e testados no genoma desta proteobactéria.

A figura 18 mostra uma comparação entre as sequências consenso para a região do BOX-35 obtidas para *H. seropedicae* e *E. coli*.

BOX -35 SIGMA 70

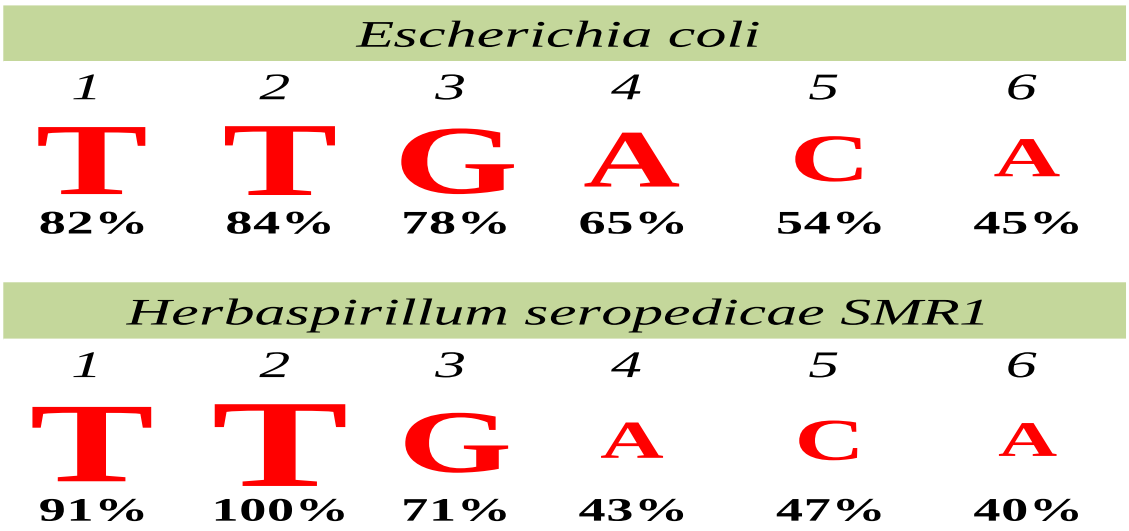


FIGURA 18 - COMPARAÇÃO DAS PROPORÇÕES DE BASES ENTRE O CONSENSO DA REGIÃO -35 EM *E. coli* E *H. seropedicae* SmR1

Nessa figura observamos as diferenças nas proporções de bases dispostas para as seis posições dispostas no box -35 em candidatos a promotores de transcrição dependentes do fator σ^{70} em *H. seropedicae* SmR1 comparados com o consenso e proporções de bases descritas na região -35 de promotores de transcrição dependentes do fator σ^{70} em *E. coli.*

FONTE: O AUTOR (2014).

A sequência consenso proposta para região -35 em *H. seropedicae* SmR1 apresentou alta similaridade em relação a sequência consenso para região -35 em *E. coli*. Em uma amostragem de 100 candidatos a promotores validados, todos apresentavam 100% da base nitrogenada timina (T) na segunda posição. Isso sugere uma forte conservação desta base, obtendo pouca variação.

A figura 19 mostra uma comparação entre as sequências consenso para a região do -10 obtidas para *H. seropedicae* e *E. coli*.

BOX -10 SIGMA 70

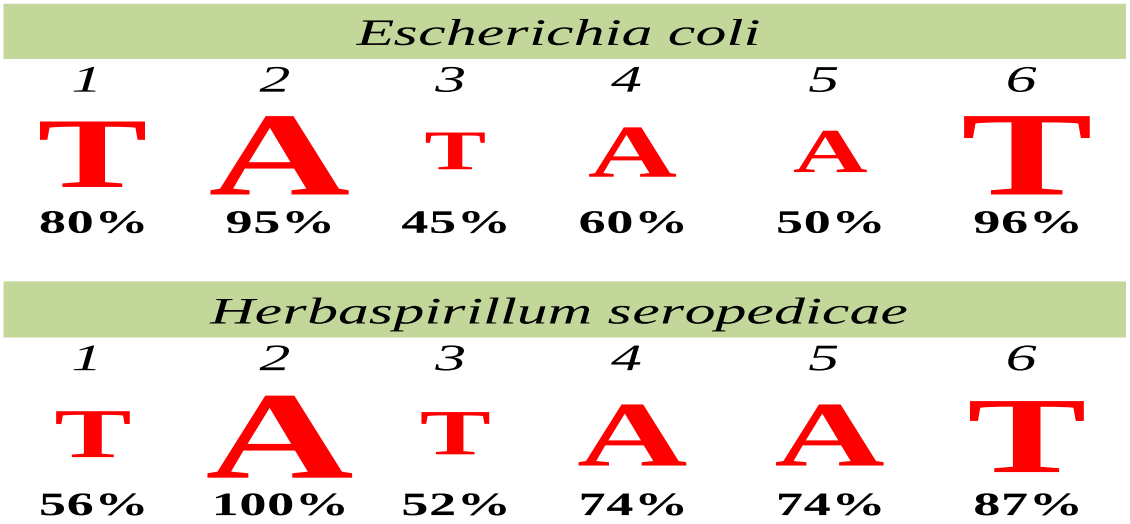


FIGURA 19 - COMPARAÇÃO DAS PROPORÇOES DE BASES ENTRE O CONSENSO DA REGIÃO -10 EM *E. coli* E *H.seropedicae* SmR1

Nessa figura observamos as diferenças nas proporções de bases dispostas nas seis posições do box -10 em candidatos a promotores de transcrição σ^{70} em *H. seropedicae* SmR1 comparados com o consenso e proporções de bases descritas na região -10 de promotores de transcrição Sigma 70 em *E. coli.*

FONTE: O AUTOR (2014).

A sequência consenso proposta para região -10 em *H. seropedicae* SmR1 apresentou alta similaridade em relação a sequência consenso para região -10 em *E. coli* conforme demonstrada na figura 19. Em uma amostragem de 100 candidatos a promotores validados todos apresentaram a base nitrogenada adenina (A) na segunda posição. A similaridade revelada entre o consenso proposto e o consenso descrito em *E. coli*, bem como nas proporções de bases encontradas, reforçam a comprovação destes candidatos, bem como a validação dos dados obtidos através do algoritimo.

O consenso proposto permitiu a análise das proporções das bases dispostas nos hexâmeros do box -35 e do hexâmero do box -10. A porcentagem de ocorrência das bases que compõem os dois hexâmeros proporcionou a construção de uma matriz utilizada nos testes envolvendo métodos de verossimilhança.

Podemos observar na tabela 3 as proporções das bases dispostas no hexâmero da região -35 retiradas de uma amostragem de 100 candidatos validados.

TABELA 3 - PROPORÇÕES DAS BASES DO BOX -35 EM *H. seropedicae* SmR1

BOX -35		**BASES NITROGENADAS**			
		T	**A**	**C**	**G**
POSIÇÕES	**1**	* 91%	7%	2%	0%
	2	*100%	0%	0%	0%
	3	16%	5%	8%	*71%
	4	14%	*43%	35%	8%
	5	11%	27%	*47%	15%
	6	24%	*40%	18%	18%

(*) Em destaque, as bases nitrogenadas com maior frequência nas seis posições do box -10. Bases nitrogenadas T = Timina A = Adenina C = Citosina G = Guanina

FONTE: O AUTOR (2014), TENDO COMO BASE UMA AMOSTRAGEM DE 100 CANDIDATOS A PROMOTORES σ^{70} EM *H. seropedicae* SmR1 VALIDADOS.

Através de uma função denominada ‘*seqlogo’* encontrada na biblioteca de funções do software Matlab®, plotamos um gráfico representando as proporções das bases nitrogenadas dispostas na região -35. As bases são empilhadas de acordo com a ordem crescente das proporções em cada posição. As bases nitrogenadas deste gráfico são diferenciadas por cores variáveis. Quanto maior for a proporção de uma determinada base, maior será seu tamanho em relação as demais bases.

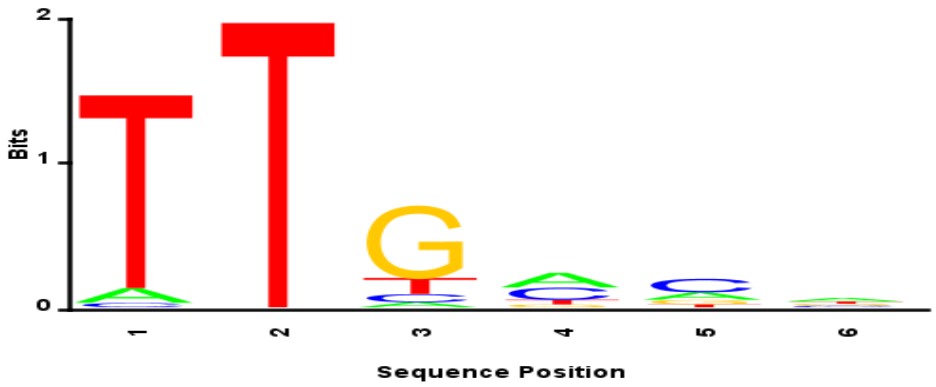


FIGURA 20 - GRÁFICO DE LOGO GERADO APARTIR DAS PROPORÇÕES DE BASES DO BOX -35 EM *H. seropedicae* SmR1.

A figura mostra as maiores proporções medidas em Bits ao longo das seis posições que compõem o hexâmero da região -35 em candidatos a promotores Sigma 70 em *H. seropedicae* SmR1. Bases nitrogenadas T = Timina A = Adenina C = Citosina G = Guanina

FONTE: O AUTOR (2014), TENDO COMO BASE UMA AMOSTRAGEM DE 100 CANDIDATOS A PROMOTORES DE TRANSCRIÇÃO DEPENDENTES DO FATOR σ^{70} EM *H. seropedicae* SmR1.

Podemos observar na tabela 4 as proporções das bases dispostas no hexâmero da região -10 retiradas de uma amostragem de 100 candidatos validados.

TABELA 4 - PROPORÇÕES DAS BASES DO BOX -10 EM *H. seropedicae* SmR1.

BOX -10		**BASES NITROGENADAS**			
		T	**A**	**C**	**G**
POSIÇÕES	**1**	*56%	31%	13%	0%
	2	0%	*100%	0%	0%
	3	*52%	16%	9%	23%
	4	10%	*74%	9%	7%
	5	9%	*74%	6%	11%
	6	*87%	0%	4%	9%

(*) Em destaque, as bases nitrogenadas com maior frequência nas seis posições do box -10. Bases nitrogenadas T = Timina A = Adenina C = Citosina G = Guanina.

FONTE: O AUTOR (2014), TENDO COMO BASE UMA AMOSTRAGEM DE 100 CANDIDATOS A PROMOTORES σ^{70} EM *H. seropedicae* SmR1 VALIDADOS.

Novamente, através de uma função denominada ‘*seqlogo’* encontrada na biblioteca de funções do software Matlab®, plotamos um gráfico representando as proporções das bases nitrogenadas dispostas na região -10. As bases são empilhadas de acordo com a ordem crescente das proporções em cada posição. As bases nitrogenadas deste gráfico são diferenciadas por cores variáveis. Quanto maior for a proporção de uma determinada base, maior será seu tamanho em relação as demais bases. A figura 21 mostra o gráfico gerado com a função ‘*seqlogo’*.

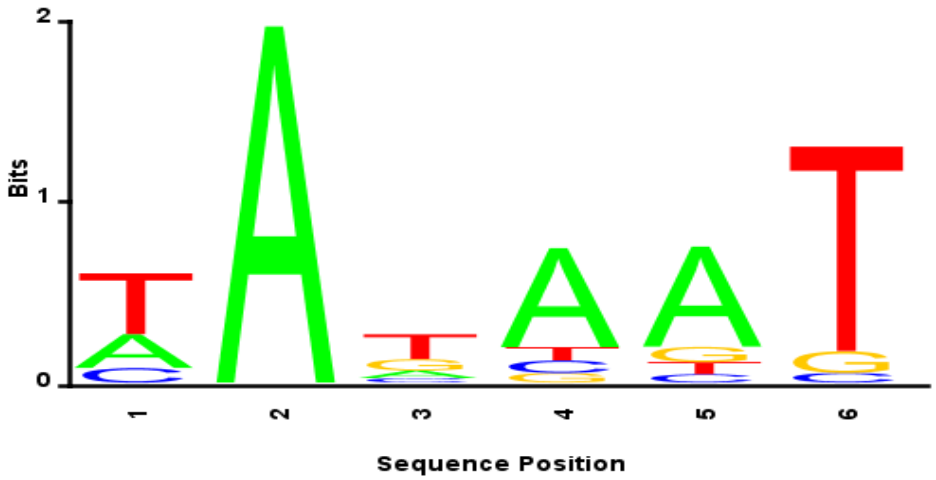


FIGURA 21 - GRÁFICO GERADO APARTIR DAS PROPORÇÕES DE BASES DO BOX -10 EM *H. seropedicae* SmR1.

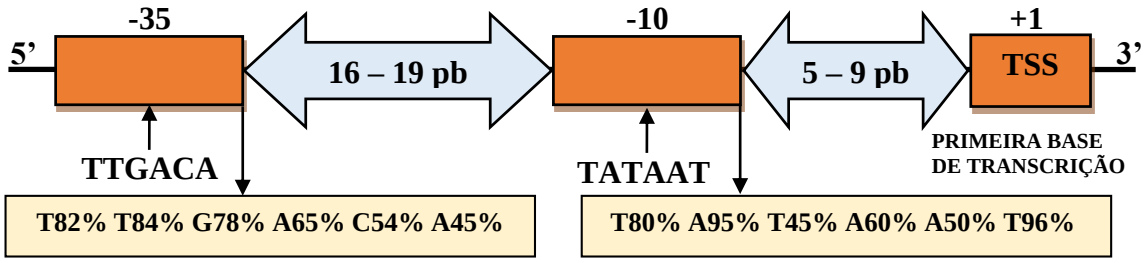
A figura mostra as maiores proporções medidas em Bits ao longo das seis posições que compõem o hexâmero da região -10 em candidatos a promotores Sigma 70 em *H. seropedicae* SmR1. Bases nitrogenadas T = Timina A = Adenina C = Citosina G = Guanina

FONTE: O AUTOR (2014), TENDO COMO BASE UMA AMOSTRAGEM DE 100 CANDIDATOS PROMOTORES DE TRANSCRIÇÃO DEPENDENTES DO FATOR σ^{70} EM *H. seropedicae* SmR1.

Verificamos que o consenso das regiões promotoras de transcrição dependentes do fator σ^{70} descritas em *E. coli*, mostraram-se conservadas em *H. seropedicae* SmR1.

A figura 22 apresenta uma comparação entre as regiões promotoras σ^{70} de *E. coli* e regiões promotoras σ^{70} propostas em *H. seropedicae* SmR1.

Sequência consenso promotor σ^{70} em *Escherichia coli*



Proposição para sequência consenso σ^{70} em *Herbaspirillum seropedicae* SmR1

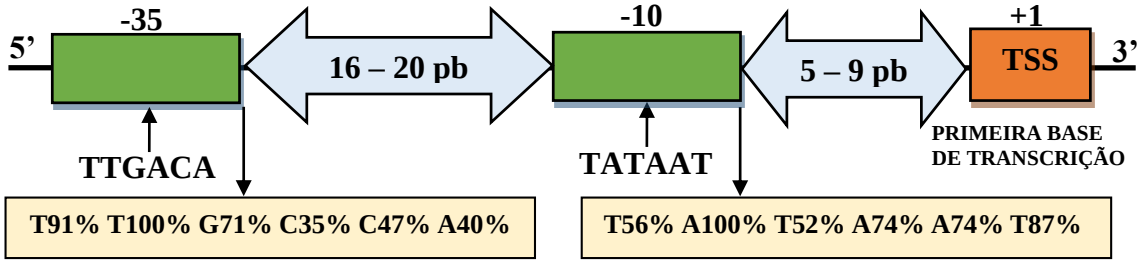


FIGURA 22 - COMPARAÇÃO ENTRE AS REGIÕES PROMOTORAS SIGMA 70 DE *Escherichia coli* E REGIÕES PROMOTORAS SIGMA 70 PROPOSTAS EM *Herbaspirillum seropedicae* SmR1.

A figura revela, tendo como base uma amostragem de 100 candidatos a promotores sigma 70 em *H. seropedicae* SmR1 validados, as proporções de bases nitrogenadas apresentaram pequenas diferenças. O número de pares de bases entre os box -35 e -10 tiveram um aumento de apenas 1 par de bases em relação ao número de pares de bases entre os box -35 e -10 descrito em *E. coli.* O número de pares de bases entre o box -10 e o sítio de início de transcrição (TSS), não tiveram alterações. O consenso proposto em *H. seropedicae* SmR1 possui alta similaridade em relação ao consenso descrito em *E. coli.* Bases nitrogenadas: T = Timina A = Adenina C = Citosina G = Guanina.

FONTE: O AUTOR (2014), BASEADO EM UMA AMOSTRAGEM DE 100 CANDIDATOS A PROMOTORES σ^{70} EM *H. seropedicae* SmR1 VALIDADOS.

7 CONCLUSÕES

- O algoritmo utilizado neste trabalho identificou 4.998 candidatos a promotores de transcrição dependentes do fator sigma 70 em *Herbaspirillum seropedicae* SmR1.

- O consenso proposto de 288 candidatos a promotores sigma 70 validados através dos 100 genes mais expressos em *Herbaspirillum seropedicae* SmR1 possui alta similaridade em relação ao consenso descrito em *Escherichia coli*, diferenciando apenas nas proporções de frequência de bases nitrogenadas nas posições dos hexâmeros.

- O padrão de conservação das 3 primeiras bases da região -35 permaneceu presente em grande parte dos candidatos, sendo o trímero TTG considerado o mais conservado da região -35 do promotor sigma 70 em ambas as espécies analisadas.

- O box -10 manteve-se conservado no consenso proposto em relação ao consenso descrito em *E. coli*.

- A segunda base do hexâmero do box -35 (T^2) assim como a segunda base do hexâmero do box -10 (A^2) revelaram ser as mais conservadas nas sequências de promotores sigma 70, em uma amostragem de 100 candidatos.

- O número de pares de bases, localizadas entre os box -35 e box -10 em *Herbaspirillum seropedicae* SmR1 manteve-se com média igual ao número de pares de bases localizadas entre os box -35 e box -10 descritos em *E. coli*.

- O método de predição computacional melhora proporcionalmente a adição de dados de promotores sigma 70 em *Herbaspirillum seropedicae* SmR1 confirmados em laboratório

- Os resultados obtidos fornecem dados importantes para trabalhos futuros, visando a confirmação destes candidatos a promotores de transcrição dependentes do fator sigma 70 em *Herbaspirillum seropedicae* SmR1, através de técnicas de Biologia Molecular.

REFERÊNCIAS BIBLIOGRÁFICAS

ARAGUAIA, M. Bioinformática, 2011. Extraído de: http://www.brasilescola.com/biologia/bioinformatica.htm. Último acesso em: 22/11/2012.

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.

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.

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.

BARBOSA, H.C. 100 f. Re-anotação do genoma de *Herbaspirillum seropedicae* SmR1 com dados de transcriptoma por RNA-Seq. Tese (Mestrado em Bioinformática) - Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, Curitiba, 2014.

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

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

COLLADO-VIDES J, MAGASANIK B, GRALLA JD. **Control site location and transcriptional regulation in *Escherichia coli***. Microbiol Rev. Sep; v.55, p. 371-94. 1991

DAYHOFF, M.O. Computer analysis of protein evolution. **Sci Am**., Jul, v.221, p. 86-95, 1969.

DÖBEREINER, J. History and new perspectives of diazotrophs in association with non leguminous plants. **Symbiosis, Rehovot,** v. 13, p. 1-13, 1992.

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.

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

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.

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: 21/11/2012.

GRALLA J.D; COLLADO-VIDES J. **Organization and Function of Transcription Regulatory Elements**, Chap. 79 In: Neidhardt F.C., Curtiss III R., Ingraham J. 1996.

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

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

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

HUERTA A.M.; SALGADO H.; THIEFFRY D., and COLLADO-VIDES J. **RegulonDB: A Database on Transcription Regulation in *Escherichia coli***, Nucleic Acids Res. v.26, p. 55-60, 1998.

ISHIHAMA, A. Molecular Assembly and functional modulation of *Escherichia coli* RNA polymerase. **Adv. Biophys.** V.26, p. 19-31, 1990.

JAMES, E. K.; OLIVARES F. L. Infection and colonization of sugar cane and other graminaceous plants by endophytic diazotrophs. **Crit. Rev. Plant Sci.** v.17, p.77-119, 1998.

KARP P.D., RILEY M., SAIER M., PAULSEN I.T., COLLADO-VIDES J., PALEY S.M., PELLEGRINI-TOOLE A., BONAVIDES C., GAMA-CASTRO S, **The EcoCyc Database**, Nucleic Acids Res. v.30, p. 56-58, 2002.

KUMAR, S. A. The Structure and Mechanism of Action of Bacterial DNA Dependent RNA polymerase. **Prog. Biophys. Molec. Biol.** v.38, p.163-210, 1981.

LEWIN, B. **Genes VII**, Cambridge, Oxford University Press, 2000.

LUGER, GEORGE F. Inteligência Artificial: **Estruturas e Estratégias para a Solução de Problemas Complexos**. 4ª ed. Porto Alegre: Bookman. p. 774, 2004.

MARQUES, J,S. Reconhecimento de Padrões Métodos Estatísticos e Neurais. IST Press, Portugal, 1999.

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

MONTEIRO A. A., CASTRO P. P. L. P., Algoritmos para Reconhecimento de Padrões, Departamento de Engenharia Elétrica, Universidade de Taubaté, **Rev. Ciênc. Exatas, Taubaté,** v. 5-8, p. 129-145, 1999.

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

MORETT, E. E BUCK, M. In vivo studies on the interaction of RNA polymerase-σ54 with the Klebsiella pneumoniae and Rhizobium meliloti niH promoters. **J. Mol. Biol.** V.210, p. 65-77, 1989.

MORETT, E. E SEGOVIA, L. J**. Bacteriol**, V.175, p. 6067-6074, 1993.

MURAKAMI, K. S. *et al.* Structural Basis of Transcription Initiation: an RNA-polimerase Holoenzime-DNA Complex. Science, New York, n.296, p. 1285-1290, May 2002.

NELSON DL, COX MM. Lehninger, Principles of Biochemistry. 3º ed. New York: Worth Publishers, 2000.

NIEVOLA, J. C . Artificial Neural Networks Training with an Inconsistent Data Subset. In: International Symposium on Engineering of Intelligent Systems'98, 1998, Tenerife, Espanha. International Symposium on Engineering of Intelligent Systems'98. Canadá : ICSC Academic Press. p. 529-531, 1998.

OLIVARES, F.L; DOS REIS JR, F. B.; REIS, V. M.; BALDANI, J. I.; DÖBEREINER, J. Infection of sugarcane roots by the endophytic diazotrophs *Herbaspirillum seropedicae* and *H. rubrisubalbicans*. In: International Symposium on Sustainable Agriculture for the Tropics – **The Role of Biological Nitrogen Fixation**, Angra dos Reis, p. 65-66,1995.

PIMENTEL, J.P.; OLIVARES, F.L.; PITARD, R.M.; URQUIAGA, S.C.; AKIBA, F.; DÖBEREINER, J. Dinitrogen fixation and infection of grasses leaves by *Pseudomonas rubrisubalbicans* and *Herbaspirillum seropedicae*. **Plant and Soil, Dordrecht,** v.137, n.1, p.61-65, 1991.

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

REZENDE, S.O. Sistemas inteligentes: fundamentos e aplicações. Ed. Manole, São Paulo, 2003.

RICH, ELAINE; KNIGHT, KEVIN. Inteligência Artificial. 2ª ed. Rio de Janeiro: McGraw-Hill, 1994. 722 p.

RONCATO-MACCARI, L.D.; RAMOS, R.J.O.; PEDROSA, F.O.; ALQUINI, Y.; CHUBATSU, L.S.; YATES, M.G.; RIGO, L.U.; STEFFENS, M.B.R. e SOUZA, E. M. Endophytic *Herbaspirillum seropedicae* Expresses nif Genes in Gramineous Plants. **FEMS Microbiol. Eco.,** V. 45(1), p. 39-47, 2003.

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.

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

SASSE-QWIGHT, S. E GRALLA, J. D. Probing the *Eschirichia coli gln*ALG upstream activation mechanism in vivo. **Proc. Natl. Acad. Sci.** USA V.85, p. 8934-8938, 1988.

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

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

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.

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

TOU, J. T.; GONZALES, R. C. Pattern recognition principles. Reading: Addison-Wesley. p. 377, 1981.

VOGT, C. Bioinformática, genes e inovação, 2003. Extraído de: http://www.comciencia.br/reportagens/bioinformatica/bio01.shtml. Último acesso em: 23/11/2012.

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

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

WÖSTEN, M. M. Eubacterial sigma-factors, **FEMS Microbiol. Rev**. V. 22, p.127-150, 1998.

YOUNG, J.P.W. Phylogenetic classification of nitrogen-fixing organisms. In: STACEY, G.; BURRIS, R. H.; EVANS, H. J. ed. **Biological Nitrogen Fixation London**: Chapman E Hall, New York, p.43-86. 1992.

APÊNDICES

Listagem contendo 288 sequências candidatas a promotores de transcrição dependentes do fator sigma 70 em *H. seropedicae* SmR1 segundo dados de transcriptoma por RNA-Seq em Re-anotação do genoma de *Herbaspirillum seropedicae* SmR1 (BARBOSA, 2014).

Gene Dow		GC%	BOX -35	BOX -10	Ns	Nº (Ns)	Pon	Gene Up	Sent.
1	Hsero_0068	50.0	ttggca	taaatt	gatgctgcgacccgatgg	18	19,7	Hsero_0073	R
		36.67	tttcaa	tatgct	tcaaagacccagaggaaa	18	18,8		
		38.71	ttttta	tatgat	tgccttgctgcggctttgt	19	19,1	cdsA	
		39.29	ttggaa	aagaat	tggcttcttataccgc	16	20,3		
2	tufB	44.83	ttgcct	tagaat	ggaatgctgatgttgcc	17	21,2	ampG	F
		37.93	ttgaca	tatagt	gtctagagatgttcctc	17	21,0		
3	secE	44.83	ttgcct	tagaat	ggaatgctgatgttgcc	17	21,2	ampG	F
		37.93	ttgaca	tatagt	gtctagagatgttcctc	17	21,0		
4	rplJ	44.83	atcaat	aagaag	aaggagccgtcatggca	17	17,7	nusG	F
5	rplL	44.83	atcaat	aagaag	aaggagccgtcatggca	17	17,7	nusG	F
6	fusA	41.38	atgaag	taaaat	agcctgggtctgcatct	17	19,6	rpsG	F
		44.83	ttgcca	tatcat	acgaagagcgaatctcc	17	20,9	rpoC	
7	Hsero_0326	53.33	ttgact	tactat	tccgcgccgcccaccttt	18	20,0	Hsero_0344	R
		43.33	ttgcct	aactgt	ggttactgccaatcgaaa	18	18,8		
		43.75	ttgctt	aaaact	gcctggttactgccaatcga	20	19,1		
		53.57	ttgctg	aatagg	gtgccgacgatgacga	16	18,7		
8	rplU	41.94	ttttga	tataat	ccgagcagctaagtgcctg	19	20,2	Hsero_0361	F
		43.33	tttgac	tataat	cgagcagctaagtgcctg	18	19,9		
		44.83	ttgacc	tataat	gagcagctaagtgcctg	17	21,5		
		58.06	ttgccg	aatagt	atctcggcgctgttgccgg	19	20,3		
		34.48	ttgaat	aataat	tggctatccggaagtga	17	21,1		
		26.67	tttttt	aataat	gtccttgatgtacaagga	18	19,6		
9	nrdB	57.69	ttgcat	cactag	cctccctaacccgc	14	18,1	Hsero_0378	F
		51.72	ttgcat	tagaat	cctccctaacccgccac	17	21,0		
10	cspC	38.71	tttcgt	tacaat	catcttgtttgtggtggtg	19	19,6	Hsero_0503	F
11	rpoD	51.72	ttccaa	tacaat	tcgggtgccagtgctgc	17	19,8	Hsero_0865	F
12	groES1	41.38	ttgaaa	tatact	attccccaacccatccc	17	20,5	nodD	R
		21.88	ttccac	taaatt	atattccatatattcagaaa	20	18,8		
		35.71	ttttgc	catatt	cttacctcacattcca	16	18,3		

		28.13	ttctgt	aatatt	catttgagttggagttaatc	20	18,5		
13	Hsero_1181	39.29	ttgtac	tacaat	gtttgtccacacagct	16	20,0	Hsero_1179	**F**
14	hflB	35.71	ttgtcc	cagtat	ctcaatgagataaaac	16	19,2	greA	**F**
		35.48	ttgtcc	tattct	ctcaatgagataaaaccag	19	19,1		
15	rpsT	41.38	ttgcct	ttataa	gctcacgcaaatctgg	16	13,2	uup	**F**
		53.33	ttgccg	aagaag	gcaaaggccatttcctgg	18	19,8		
16	cspD	30.0	ttgcat	aatcgt	tctcgatttaatgagtat	18	19,3	ispF	**F**
		30.0	ctaaag	aatatt	cacccgaaaaagtcatta	18	17,3		
17	clpS	39.29	ttgaca	aagcat	gtacctgttgccataa	16	20,3	icd	**F**
		25.81	ttgcca	taaact	taaaagcatctttcaaata	19	20,2		
18	clpA	39.29	ttgaca	aagcat	gtacctgttgccataa	16	20,3	icd	**F**
		25.81	ttgcca	taaact	taaaagcatctttcaaata	19	20,2		
19	Hsero_1639	27.59	tttgac	tagaat	tttatgcacgattaaac	17	19,5	Hsero_1642	**R**
		28.57	ttgact	tagaat	ttatgcacgattaaac	16	21,3		
		50.0	ttgcgc	cagaat	gcacgggattgataac	16	19,9		
20	rpmF	25.81	tttgac	tattat	tggtaaggaaattttctgt	19	19,0	Hsero_1908	**F**
		26.67	ttgact	tattat	ggtaaggaaattttctgt	18	20,8		
		40.63	tttacg	taaaat	gaaacctggatccggtgtgt	20	20,4		
21	acpP	41.94	ttggcg	taaaag	caatcttgccaggaatagt	19	19,3	fabG	**F**
		25.0	ttgcca	aatttt	ggaatagttaaaagta	16	19,7		
		40.63	ttggcg	aaaagt	caatcttgccaggaatagtt	20	19,2		
		41.18	ctgcta	acctat	aaatgcgcgcacttttgta	19	13,6		
22	rpoE	46.88	tttccg	aataat	cgccgcagatcccattgatc	20	20,5	fabH	**F**
		43.33	ttgtcg	aagtat	accaggaatccctgaaca	18	19,5		
		25.0	ttgcca	aatttt	ggaatagttaaaagta	16	19,7	fabG	
		41.94	ttggcg	taaaag	caatcttgccaggaatagt	19	19,3		
23	infC	28.13	tttatt	aatagc	ggattttaaaaggaaactgc	20	17,9	thrS	**F**
24	Hsero_2062	20.69	atcgta	aattca	ttgaaaatatgcaatta	17	12,3	Hsero_2064	**R**
		34.38	ttgcta	aaatat	gcctggctatatcgtattga	20	19,5		
25	flaG	37.93	ttgaaa	taggat	taagcctgatagctagg	17	20,2	Hsero_2071	**R**
		35.48	ttgcgt	aagact	aaaaccctaaacttttcc	19	19,3	Hsero_2072	
26	fliC	35.48	ttgcgt	aagact	aaaaccctaaacttttcc	19	19,3	Hsero_2072	**R**
		28.57	tttacc	taaagt	acaacagatattttcc	16	19,5		
		34.38	tttgcg	aagact	taaaaccctaaacttttcc	20	18,4		
		28.13	tttctt	taaagt	taccacaacagatattttcc	20	18,9		
27	Hsero_2071	56.67	ttgcgc	aaggat	gtcatcgaaatcgccccc	18	19,3	lexA	**R**
		33.33	ttgaca	aaaaat	caaggattgcacctgtat	18	21,1	phoH	
		50.0	ttgcct	aaacat	tccgtcggcatcaatcgtcc	20	19,9		
28	rpsB	44.83	ttgaac	tagaat	aaagtgacggggtcgag	17	20,9		

		24.14	tttttg	tatctt	gcaatgtttgtagaaaa	17	17,9	Hsero_2174	**F**
		41.38	ttttga	taaaag	aacgcgaatccgttcgc	17	18,7		
29	lon	33.33	atccta	aaaaat	gtagctctatgtctgaga	18	18,3	clpP	**R**
		38.71	ttgata	tataac	aatggcactggcagtgaca	19	19,9		
		32.26	ttgcct	taaatt	ttctgcgcaaatttttgac	19	20,1		
		31.25	tttgcc	taaatt	tttctgcgcaaatttttgac	20	18,7		
30	Hsero_2905	25.81	ttccaa	tatttg	atttcactgtcagtaaatt	19	17,8	antB	**R**
31	hfq	28.57	ataaaa	tattcc	ttggtaaagcacttga	16	16,1	Hsero_2949	**R**
		26.67	ttgaag	aatagc	ctaaaatcgtttaaattg	18	18,8		
		31.03	ttggcg	taaatt	ttgaagctaaaatcgtt	17	19,4		
		31.03	atcaat	gaagct	gtttttggcgtt	12	13,0		
32	ndk	25.81	ttatca	tatttt	attcttggctttaagcctt	19	18,4	rumT	**R**
		27.59	atcaat	tatttt	tcttggctttaagcctt	17	17,5		
		39.29	tttccc	catcat	tcatgcaacaagacat	16	19,2	rpoS	
33	cheW	21.43	ttcaga	tatagg	agaaataagaaaaata	16	18,6	metC	**R**
		26.67	ttgtaa	aagtgt	aaaaacgcaagttttctt	18	18,8		
		44.83	ttttaa	tagaat	gcccggctgtgtggtca	17	20,1		
		72.41	ttccgg	tatagt	cgccggcgcgctgcgcg	17	19,2	Hsero_2991	
34	flhC	21.43	ttcaga	tatagg	agaaataagaaaaata	16	18,6	metC	**R**
		26.67	ttgtaa	aagtgt	aaaaacgcaagttttctt	18	18,8		
		44.83	ttttaa	tagaat	gcccggctgtgtggtca	17	20,1		
		72.41	ttccgg	tatagt	cgccggcgcgctgcgcg	17	19,2	Hsero_2991	
35	flhD	21.43	ttcaga	tatagg	agaaataagaaaaata	16	18,6	metC	**R**
		26.67	ttgtaa	aagtgt	aaaaacgcaagttttctt	18	18,8		
		44.83	ttttaa	tagaat	gcccggctgtgtggtca	17	20,1		
		72.41	ttccgg	tatagt	cgccggcgcgctgcgcg	17	19,2	Hsero_2991	
36	cspD	37.93	ttgact	aacaat	tttgcatttcgaccgag	17	20,6	Hsero_1400	**R**
		31.25	atcacg	tattac	taagatctagatgagataga	20	17,2		
		45.16	ttgcca	taagct	tgcgtttgttcgaccatca	19	19,1		
		56.25	ttgccg	catgat	ttggaagccgactgcacgtg	20	19,8		
37	efp	14.29	atgaac	tatatt	gatataaataattaac	16	19,2	Hsero_3062	**F**
		13.33	atgaac	tattat	gatataaataattaacta	18	19,3		
		23.33	ttgcga	aactat	tgaacgatataaataatt	18	19,4		
		21.88	attaac	aaatat	tatattattgcgcacaatct	20	17,9		
		21.21	ttgcga	tatatt	tgaacgatataaataattaac	21	20,4		
38	Hsero_3196	45.16	ttggcg	tattct	tgggttaccgtttctaagg	19	18,8	acd	**F**
39	qor	41.38	ttgaca	tatatt	aacggaacacagctccg	17	20,9	fabG	**F**
		45.16	ttgaca	tattcg	aacggaacacagctccgta	19	18,9		

		46.88	ttccca	aatgat	gtattgaggaccgcaccaga	20	19,5		
40	Hsero_3500	41.38	ttgaca	tatatt	aacggaacacagctccg	17	20,9	fabG	**F**
		45.16	ttgaca	tattcg	aacggaacacagctccgta	19	18,9		
		46.88	ttccca	aatgat	gtattgaggaccgcaccaga	20	19,5		
41	ihfB	31.03	ttgatt	taatat	gacaagggtttttctgc	17	19,6	Hsero_3700	**R**
		57.14	tttcca	catcat	atccgcgcaaggccgg	16	19,6		
		48.28	tttccc	aagagt	tcgaagaagagatgcgc	17	19,3		
		18.75	ttttt	tacagt	aagaatatccttatttttca	20	18,3		
		20.0	ttttaa	tacagt	gaatatccttatttttca	18	18,9		
		19.35	ttttta	tacagt	agaatatccttatttttca	19	18,5		
42	Hsero_3696	20.69	atatta	taatct	ttcatttttgatgtcg	17	16,0	Hsero_3672	**F**
		57.14	ttgccc	tatagc	tggccatcgactaggg	16	19,3		
43	Hsero_3845	40.0	ttgacg	tatact	ctaagtccttgaaaaggc	18	20,4	ubiF	**R**
		48.28	ttccca	aaggat	tgcgctcgggaagataa	17	19,2		
		46.67	tttccc	aaggat	atgcgctcgggaagataa	18	19,1		
		38.71	tttgac	tatact	gctaagtccttgaaaaggc	19	18,8		
		37.5	ttttga	tatact	cgctaagtccttgaaaaggc	20	19,1		
		41.38	ttgcaa	tatatt	cgcacaaagcgctttgc	17	20,6		
		51.61	ttctcc	aatact	agcatgatcgaggccgagc	19	18,7		
44	rplM	48.28	ttccca	aaggat	tgcgctcgggaagataa	17	19,2	ubiF	**R**
		46.67	tttccc	aaggat	atgcgctcgggaagataa	18	19,1		
		40.0	ttgacg	tatact	ctaagtccttgaaaaggc	18	20,4		
		38.71	tttgac	tatact	gctaagtccttgaaaaggc	19	18,8		
		37.5	ttttga	tatact	cgctaagtccttgaaaaggc	20	19,1		
		37.5	ttattg	tatatt	caacgcacaaagcgctttgc	20	18,3		
		51.61	ttctcc	aatact	agcatgatcgaggccgagc	19	18,7		
45	Hsero_3857	58.06	ttgcat	tacaat	gacccgggggcagatggcg	19	20,5	gdhA	**R**
46	yhbH	25.0	tttttt	tatact	gctttgctaattcgtt	16	18,8	Hsero_3965	**F**
		65.63	ttgcct	catgat	gcggcccagcccagcgtgac	20	19,9		
		61.29	ttgccg	tacagt	caccacggccacgatctgg	19	19,8		
47	Hsero_4198	25.0	ttgcat	tatact	aatttttaaagcgtc	16	20,2	pth	**F**
		46.43	ttgccg	cattat	atcgagatcagcgaga	16	19,9	Hsero_4186	
		36.67	ttgagc	aaaaat	ttgtgaacgggagcaaaa	18	20,2	dgt	
		16.67	tttcgt	tatttt	tacatttttctgaaaat	18	18,6		
48	Hsero_4295	29.03	ttatgc	taaaat	cccacaatcgacatttttt	19	18,9	dapF	**F**
		40.63	atcaag	aataag	tttctgcagttgcagccgat	20	18,0		
49	atpE	29.03	tttgac	tataat	ttaccagaaaacgaacctt	19	19,9	Hsero_4370	**R**
		30.0	ttgact	tataat	taccagaaaacgaacctt	18	21,6		
		31.25	ttcaga	tatttt	aaacgtttactccaactgag	20	18,6		

50	Hsero_4678	41.38	tttacc	aattag	gcaggttggagtgaaga	17	18,8	Hsero_4685	F
		65.52	ttgtca	tagact	gcgccggccaggctccc	17	19,9		
		51.61	ataacg	tagtat	atcgcctggcgtgagcagg	19	18,0		
		45.16	ttgctg	tattct	atgccggaatagccatcga	19	18,8		
		48.39	ttggtg	aatagt	gaccagttgaacttgcgcc	19	19,1		
51	dnaN	28.13	ttcagc	aagatt	tttatagttaggctgtttat	20	18,5	fdx	F
		37.93	ataaca	aatcat	cacaacgtcaaggatcg	17	18,4		
		53.57	tttgtc	aataag	gatggggccgtggtgg	16	18,3	Hsero_4803	
52	rpsJ	28.13	tttcgt	aatcat	gcattaaagacttaggaata	20	19,2	rpsG	F
		41.38	atgaag	taaaat	agcctgggtctgcatct	17	19,6		
53	rplD	37.93	ttgtgc	tataat	gcttttgttttacgggc	17	20,5	tufB	F
54	rplW	37.93	ttgtgc	tataat	gcttttgttttacgggc	17	20,5	tufB	F
55	rpsC	40.63	ttgacg	tatgtg	agctgaaagtcaagaccatt	20	18,6	rplB	F
		46.43	ttgcag	aaaagt	acctgatccgcggtaa	16	19,6		
		32.26	ataaga	tatgat	aggctaagaaataaggtcc	19	18,1		
		39.29	atcaac	aataat	ttcaccgaggaggcta	16	19,0	rplD	
56	rplE	41.38	ttctgt	aagatt	aaggagtggtcatggca	17	18,2	rpsQ	F
		41.38	ttgcgt	cataat	gctcgatattatgctgc	17	20,4		
57	rplR	35.71	ctcaaa	aatagg	gaaaccaagaagaagt	16	16,9	rpsH	F
		46.67	ttatcg	aagaag	acgaccaatcgaagtcgg	18	18,3	rplE	
		41.38	ttctgt	aagatt	aaggagtggtcatggca	17	18,2	rpsQ	
		41.38	ttgcgt	cataat	gctcgatattatgctgc	17	20,4		
58	infA	33.33	gttcta	aaaaat	aggattatcgatcatggc	18	15,8	rpsF	F
		35.71	ctcaaa	aatagg	gaaaccaagaagaagt	16	16,9	rpsH	
		46.67	ttatcg	aagaag	acgaccaatcgaagtcgg	18	18,3	rplE	
		41.38	ttctgt	aagatt	aaggagtggtcatggca	17	18,2	rpsQ	
		41.38	ttgcgt	cataat	gctcgatattatgctgc	17	20,4		
59	rpsM	33.33	gttcta	aaaaat	aggattatcgatcatggc	18	15,8	rpsF	F
		35.71	ctcaaa	aatagg	gaaaccaagaagaagt	16	16,9	rpsH	
		46.67	ttatcg	aagaag	acgaccaatcgaagtcgg	18	18,3	rplE	
		41.38	ttctgt	aagatt	aaggagtggtcatggca	17	18,2	rpsQ	
		41.38	ttgcgt	cataat	gctcgatattatgctgc	17	20,4		
60	dksA	46.43	ttggca	tacaat	tccggaatcgttggtg	16	20,4	thiF	F
61	dnaK	29.03	tttcct	aaaact	ccacatattaaaaccattc	19	19,0	hemH	F
		59.38	ttgccg	taaaat	gcccgctcctactggcctgc	20	20,9		
62	sucC	56.67	ttgcct	aatggt	caccgttgcgatgccctg	18	19,4	recA	F
63	groEL1	41.38	ttgaaa	tatact	attccccaacccatccc	17	20,5	nodD	R
		35.71	ttttgc	catatt	cttacctcacattcca	16	18,3		
		21.88	ttccac	taaatt	atattccatatattcagaaa	20	18,8		

		42.86	ttgcct	tattcc	tacctcacattccaca	16	18,2		
		28.13	ttctgt	aatatt	catttgagttggagttaatc	20	18,5		
		56.67	ttgacc	aatacg	tgagaaccggcgccgaac	18	19,2		
		60.71	ttggcg	tatacg	tccgtaccggtgaggg	16	18,7		
		61.29	ttgccg	tatagc	gcatcggtaccggtgagcg	19	19,3		
		61.29	ttgccg	tatagg	gcatcggtaccgctgaggg	19	19,6		
64	Hsero_1104	58.62	ttgcca	aagaag	aggcggccagtgcgttg	17	20,2	edd	**F**
65	rpmG	27.59	ttgaaa	tatact	agtctaagttttttccc	17	20,5	radC	**R**
66	fdxA	26.67	tttgtg	taaaat	ggaaaaatcgaattacct	18	19,0	acb	**R**
		25.81	ttattt	aagcct	acttatacgtaatctgtat	19	16,7		
		25.0	tttatt	aagcct	tacttatacgtaatctgtat	20	17,7		
		27.59	ttgtgg	taaaat	gaaaaatcgaattacct	17	20,0		
67	nuoA	35.71	tttagc	aagagt	tgattgtgcattgaac	16	18,9	tpiA	**F**
		30.0	ttgaac	tacaat	aagagtaacaaatgaagg	18	20,5		
68	fabF1	25.0	ttgcca	aatttt	ggaatagttaaaagta	16	19,7	fabG	**F**
		41.94	ttggcg	taaaag	caatcttgccaggaatagt	19	19,3		
		41.18	ctgcta	tatcat	aaatgcgcgcacttttgtaacc	22	18,6		
69	rseA	35.71	ataaca	taagat	aagcgggatatggaac	16	18,1	rpoE	**F**
70	rplT	28.13	tttatt	aatagc	ggattttaaaaggaaactgc	20	17,9	thrS	**F**
71	flgG	35.48	ttgccc	cagaat	tgtatcaatagttgatttc	19	20,4	fliA	**F**
72	rpsF	34.48	ttgctg	tatgat	ttctatccaaaagcttg	17	19,8	Hsero_2068	**F**
		58.62	ttgacg	aataag	ccggggtgggcgtcttc	17	20,3		
		41.38	ttgctt	aaaagt	tatcgggcagtcttgga	17	19,3		
		25.0	tttacg	aagact	caaaaaaatctttttc	16	19,2		
73	rpsR	34.48	ttgctg	tatgat	ttctatccaaaagcttg	17	19,8	Hsero_2068	**F**
		58.62	ttgacg	aataag	ccggggtgggcgtcttc	17	20,3		
		41.38	ttgctt	aaaagt	tatcgggcagtcttgga	17	19,3		
		25.0	tttacg	aagact	caaaaaaatctttttc	16	19,2		
74	tsf	44.83	ttgaac	tagaat	aaagtgacggggtcgag	17	20,9	Hsero_2174	**F**
		24.14	tttttg	tatctt	gcaatgtttgtagaaaa	17	17,9		
		41.38	ttttga	taaaag	aacgcgaatccgttcgc	17	18,7		
75	Hsero_2186	55.17	ttgtcg	aatact	ggcgcagccacggcaaa	17	19,7	Hsero_2185	**F**
		35.71	tttaga	tattat	tggcggtcatttgctg	16	19,8	cdsA	
76	Hsero_2187	55.17	ttgtcg	aatact	ggcgcagccacggcaaa	17	19,7	Hsero_2185	**F**
		35.71	tttaga	tattat	tggcggtcatttgctg	16	19,8	dxr	

77	Hsero_2243	50.0	ttgcaa	aatacg	tgcatgcggtcaacccca	18	19,2	Hsero_2252	**R**
		60.0	ttgcag	aatcat	cgacagggcgactgccgc	18	20,0		
78	clpX	38.71	ttgata	tataac	aatggcactggcagtgaca	19	19,9	clpP	**R**
		31.25	tttgcc	taaatt	tttctgcgcaaatttttgac	20	18,7		
		32.26	ttgcct	taaatt	ttctgcgcaaatttttgac	19	20,1		
79	clpP	38.71	ttggac	taaaat	tatcggcgctttttccgta	19	20,0	creA	**R**
80	tig	38.71	ttggac	taaaat	tatcggcgctttttccgta	19	20,0	creA	**R**
81	hns	44.83	ttgctc	tattgt	tagtttgccagctggct	17	19,1	Hsero_2877	**R**
		19.35	ttatca	aaaaat	tcctgaaagaataaattga	19	19,5		
		23.33	ataaaa	aagaat	gtctttttatcatcctga	18	18,7		
		18.75	tttatc	aaaaat	atcctgaaagaataaattga	20	19,5		
		38.71	ttgctt	cagaat	ttgttattttctgcgcagg	19	19,9		
82	Hsero_2904	25.81	ttccaa	tatttg	atttcactgtcagtaaatt	19	17,8	Hsero_2906	**R**
83	Hsero_2928	32.14	ttttta	tatagg	gcaaattggtctgtca	16	18,3	pbpC	**R**
		50.0	ttgtcg	tatagc	aagaacaccgatgcgg	16	18,9		
		37.93	ttcttg	taaaag	ctgatgtcgttcagttg	17	17,9		
		50.0	ttggcg	aatcat	ccttgtcgagatcggt	16	19,6		
		46.67	ttgcgc	aagaat	aaagatttgccgccgatg	18	20,3		
84	Hsero_2957	26.67	gtattt	tatcat	tagcatgaaaccattaga	18	14,5	Hsero_2951	**F**
		27.59	ataatt	tagcat	cgatgaaccttgtattt	17	17,5		
		28.13	ttgata	tagcat	attcgatgaaccttgtattt	20	20,0		
		25.0	ataaaa	aagaat	ataaaggcttaaagcc	16	18,7		
		28.57	ttgcca	aaaaat	tgagaaaaactccaat	16	21,0		
		58.06	ttgtcg	aagaat	gcgcgacggatctcgaccc	19	20,4		
85	cheY	42.86	ttcctg	aagaat	aagatagcctgccatg	16	19,2	dnaJ	**F**
86	Hsero_2984	21.43	ttcaga	tatagg	agaaataagaaaaata	16	18,6	metC	**R**
		26.67	ttgtaa	aagtgt	aaaaacgcaagttttctt	18	18,8		
		44.83	ttttaa	tagaat	gcccggctgtgtggtca	17	20,1		
87	motB	21.43	ttcaga	tatagg	agaaataagaaaaata	16	18,6	metC	**R**
		26.67	ttgtaa	aagtgt	aaaaacgcaagttttctt	18	18,8		
		44.83	ttttaa	tagaat	gcccggctgtgtggtca	17	20,1		
88	glnA	30.0	ttgttt	aatggt	cttccattagtgcatttt	18	18,4	Hsero_3130	**R**
89	iscU	56.67	ttgcgg	aagaat	atggccaggtagggcagg	18	20,3	Hsero_3148	**R**
90	slyD	43.75	ttggca	tataat	gccaccaagtttgatgcagg	20	21,1	Hsero_3274	**R**
		31.03	ttgtgc	aattat	gatgaataattaccggt	17	19,4		
91	spoVK	41.94	ttgctt	aataag	gagttcatccaacgacgtc	19	19,7	Hsero_3475	**R**

		51.61	tttccg	cagaat	atggacgtccgtaaacggc	19	19,8	Hsero_3479	
		30.0	ttgcct	taaact	cacaggaatttaatcaat	18	19,9	Hsero_3481	
92	rpsA	18.75	tttttt	tacagt	aagaatatccttatttttca	20	18,3	Hsero_3700	**R**
		20.0	ttttaa	tacagt	gaatatccttatttttca	18	18,9		
		19.35	ttttta	tacagt	agaatatccttatttttca	19	18,5		
		41.38	ttgcgt	aaagat	attacgcttgcatctcg	17	19,3		
93	Hsero_3880	44.83	tttatc	cataat	aatcaacggcgcagcgg	17	19,6	fabG	**F**
		50.67	atcaat	aatagg	caacggcgcagcggcat	17	17,3		
		48.39	ttatca	aatagg	atcaacggcgcagcggcat	19	18,1		
		46.43	ttatca	cataat	atcaacggcgcagcgg	16	19,6		
94	rpoN	25.0	tttttt	tatact	gctttgctaattcgtt	16	18,8	Hsero_3965	**F**
		65.63	ttgcct	catgat	gcggcccagcccagcgtgac	20	19,9		
		61.29	ttgccg	tacagt	caccacggccacgatctgg	19	19,8		
95	petC	21.43	ctattc	aaaagt	agcttatatataatca	16	16,0	Hsero_4051	**R**
		46.67	ttgaca	aatgtt	atgggttgcgctgcgtca	18	19,7		
96	Hsero_4196	25.0	ttgcat	tatact	aatttttttaaagcgtc	16	20,2	pth	**F**
		46.43	ttgccg	cattat	atcgagatcagcgaga	16	19,9	Hsero_4186	
		36.67	ttgagc	aaaaat	ttgtgaacgggagcaaaa	18	20,2		
97	wecB	34.48	tttcgg	tattag	cattgtcaaaaatgcac	17	18,4	Hsero_2738	**F**
		28.57	tttttg	tattcc	gtcaaaattgcaacat	16	16,4		
		54.84	ttgccg	tatcgg	gtgagcatgatgaccttgg	19	18,7	cheD	
98	atpA	29.03	tttgac	tataat	ttaccagaaaacgaacctt	19	19,9	Hsero_4370	**R**
		30.0	ttgact	tataat	taccagaaaacgaacctt	18	21,6		
		31.25	ttcaga	tatttt	aaacgtttactccaactgag	20	18,6		
		50.0	ttgttg	tatagg	atgtcagccaggccag	16	18,6	Hsero_4372	
99	atpH	29.03	tttgac	tataat	ttaccagaaaacgaacctt	19	19,9	Hsero_4370	**R**
		30.0	ttgact	tataat	taccagaaaacgaacctt	18	21,6		
		31.25	ttcaga	tatttt	aaacgtttactccaactgag	20	18,6		
		50.0	ttgttg	tatagg	atgtcagccaggccag	16	18,6	Hsero_4372	
100	atpB	29.03	tttgac	tataat	ttaccagaaaacgaacctt	19	19,9	Hsero_4370	**R**
		30.0	ttgact	tataat	taccagaaaacgaacctt	18	21,6		
		31.25	ttcaga	tatttt	aaacgtttactccaactgag	20	18,6		
		50.0	ttgttg	tatagg	atgtcagccaggccag	16	18,6	Hsero_4372	