Adriana Neves dos Reis

Reconhecimento e Predição de Promotores Procarióticos: investigação de uma metodologia *in silico* baseada em HMMs

Dissertação apresentada à Universidade do Vale do Rio dos Sinos (UNISINOS) como requisito parcial para obtenção do título de Mestre em Computação Aplicada

Orientador: Prof. Dr. Ney Lemke

São Leopoldo

2005

UNISINOS

UNIVERSIDADE DO VALE DO RIO DOS SINOS – UNISINOS
CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA INTERDISCIPLINAR DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

Aluno: **Adriana Neves dos Reis**

Título da Dissertação: **Reconhecimento e Predição de Promotores Procarióticos: Investigação de uma Metodologia *in silico* baseada em HMMs.**

Banca:

Dr. Ney Lemke
Presidente/Orientador

Dr. Georgios Joannis Pappas Junior
Membro da banca externo

Dr. Adelmo Luis Cechin
Membro da banca interno

A banca examinadora da Dissertação, sob registro de Ata nº 33/2005 - PIPCA, em cumprimento ao Regimento do Programa Interdisciplinar de Pós-graduação em Computação Aplicada, julga esta Dissertação aprovada para o processo de obtenção do título de Mestre a Adriana Neves dos Reis.

São Leopoldo, 3 de março de 2005

Ao meu amado, Daniel Rodrigo Metz

Agradecimentos

Ao meu orientador Ney Lemke, por sua dedicação ao longo da elaboração da dissertação, não só de caráter técnico, mas também humano.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, pela bolsa integral que viabilizou a realização do mestrado com dedicação exclusiva.

À HP Brazil R&D, pela colaboração no desenvolvimento desta pesquisa.

Ao coordenador do Programa Interdisciplinar de Pós-Graduação em Computação Aplicada – PIPCA, Arthur Tórgo Gómez, e à secretária do programa, Rejane Weissheimer, pela disponibilidade bem humorada para tratar das questões burocráticas.

Ao professor Adelmo Luis Cechin, pelas relevantes contribuições para minha formação e para esta pesquisa.

Aos colegas do Laboratório de Bioinformática e Biologia Computacional - LBBC, pelo clima familiar de trabalho. Especialmente à Norma Machado da Silva, pela amizade, companheirismo e valiosas discussões a respeito dos resultados e da “magia” envolvida na interação da RNA-polimerase com os promotores.

Finalmente, ao meu pai, Nilton, que esteve ao meu lado para a concretização de mais este objetivo.

Resumo

A expressão dos genes em procariotos é desencadeada quando a enzima RNA-polimerase interage com uma região adjacente ao gene, chamada de promotor, onde se encontram os principais elementos regulatórios do processo de transcrição. Apesar do crescente avanço das técnicas experimentais em biologia molecular, caracterizar e identificar um número significante de promotores, presentes em um dado genoma, continua sendo uma tarefa demorada e cara. Abordagens *in silico* são bastante utilizadas para reconhecer essas regiões em procariotos. Entretanto, além do alto número de falsos positivos obtidos, elas enfrentam a inexistência de um número adequado de promotores conhecidos para identificar padrões conservados entre as espécies. Logo, um método criterioso e confiável para predizê-los em qualquer organismo procariótico ainda é um desafio. Esta dissertação propõe um protocolo de uso de *hidden* Markov *models* (HMMs) que emprega Estimação de Limiar de Decisão (ELD) e Análise de Discriminação (AD) neste problema. Quatro espécies procarióticas são analisadas (*Escherichia coli*, *Bacillus subtilis*, *Helicobacter pylori* e *Helicobacter hepaticus*), verificando-se a influência do tamanho das bases de dados de promotores, do conjunto de informações estruturais e funcionais, e do conteúdo A+T de cada genoma no desempenho do reconhecimento e da predição de promotores. Os resultados mostram que o protocolo proposto aumenta a capacidade do modelo HMM reconhecer essas regiões, alcançando uma redução de 44,96% na taxa de erro, comparada a trabalhos anteriores, para os promotores de *Escherichia coli*. Para o *Bacillus subtilis*, a exatidão é de 95% no reconhecimento e de 78% na predição. Para as espécies do gênero *Helicobacter*, o

protocolo tem uma baixa capacidade de predizer os promotores, gerando um alto número de falsos positivos.

Palavras-chave: promotores, procariotos, reconhecimento de padrões, HMMs

Abstract

Gene expression on prokaryotes initiates when the RNA-polymerase enzyme interacts with DNA regions called promoters. In these regions are located the main regulatory elements of the transcription process. Despite the improvement of *in vitro* techniques for molecular biology analysis, characterizing and identifying a great number of promoters on a genome is a complex task. *In silico* approaches are usually employed to recognize theses regions on prokaryotes. Nevertheless, the main drawback is the absence of a large set of promoters to identify conserved patterns among the species. Hence, a *in silico* method to predict them on any species is a challenge. This work proposes a protocol to use hidden Markov models (HMMs) methodology with Decision Threshold Estimation and Discrimination Analysis on this problem. Four prokaryotic species are investigated (*Escherichia coli*, *Bacillus subtilis*, *Helicobacter pylori* e *Helicobacter hepaticus*). The influence of different aspects in the recognition and prediction are examined: the size of promoter datasets, structural and functional available information, and A+T content of each genome. The results show that the proposed protocol increases the recognition ability of the HMM, obtaining a reduction in 44.96% of error rate compared with previous works on *Escherichia coli* promoters. For *Bacillus subtilis*, the accuracy is 95% on recognition and 78% on prediction. However, the protocol presents a high error rate on promoter prediction of *Helicobacter* species, since it generates a large number of false positives.

Key-words: promoters, prokaryotes, pattern recognition, HMMs

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

A	adenina
aa	amino ácido
AD	Análise de Discriminação
AM	Aprendizado de Máquina
C	citosina
DNA	ácido desoxirribonucléico
ELD	Estimação de Limiar de Decisão
ELDAD	Estimação de Limiar de Decisão e Análise de Discriminação
FN	falso negativo
FP	falso positivo
G	guanina
HMMs	*Hidden Markov Models*
KBANN	Knowledge Based Neural Network
mRNA	RNA mensageiro
NNPP	*Neural Networks Promoter Prediction*
nt	Nucleotídeo
pb	pares de base
PEC	Profile of *E. coli* Chromosome
RNA	ácido ribonucléico
RNAp	RNA-polimerase
RNs	Redes Neurais
RP_c	complexo fechado do promotor
RP_i	complexo intermediário RNAp-promotor
RP_o	complexo aberto RNAp-promotor
T	Timina
TIGR	The Institute for Genomic Research
TSS	transcription start site
VN	verdadeiro negative
VP	verdadeiro positive

Sumário

1 Introdução

Decifrar o código genético e compreender sua influência na organização e no funcionamento dos seres vivos é o desafio que tem motivado anos de pesquisa em diferentes áreas desde genética, biologia molecular e bioquímica até áreas exatas, como estatística e computação.

Após a postulação do modelo de dupla hélice da molécula de DNA por James Watson e Francis Crick em 1953, o grande marco na corrida para desvendar o código genético e os mecanismos bioquímicos e biomoleculares em que ele está envolvido foi o desenvolvimento da tecnologia *in vitro* para seqüenciar genomas em larga escala [Watson, 2003]. O resultado do seqüenciamento é uma longa cadeia de caracteres que representa a constituição do DNA, na qual estão contidos não apenas os genes, ou seja, as informações que precisam ser decodificadas, mas também outros elementos que governam a sua decodificação, chamados de regulatórios. O conjunto de mecanismos em que esses elementos participam é conhecido como Regulação da Expressão Gênica.

O estudo da Regulação da Expressão Gênica pode ajudar a responder importantes questões, como: "*Por que organismos com alta complexidade funcional não possuem, necessariamente, o número total de genes muito superior ao de espécies mais simples?*" Por exemplo, estima-se que o homem possua 35 mil genes, enquanto a *Arabdopis thaliana*, uma pequena planta parente da mostarda, possui 27 mil [Wang et al., 1999]. Além desta, se os diferentes tipos celulares de um mesmo organismo possuem uma cópia idêntica da molécula de DNA, "*O que proporciona a diferenciação celular se todas as células possuem o mesmo DNA?*".

Compreender o funcionamento da Regulação da Expressão dos Genes, mesmo com o surgimento de poderosas técnicas experimentais *in vitro* e *in silico*, é um dos maiores desafios da biologia molecular. Dentre os mecanismos regulatórios, a interação da enzima RNA-polimerase (RNAp) com a região de DNA antecedente ao gene (região promotora) desempenha o papel desencadeador de todo o processo de decodificação gênica.

Estudos mostram que a região promotora possui sítios específicos que auxiliam no reconhecimento dela pela RNAp [Oppon, 2000]. Iniciativas computacionais de descrever esses sítios e analisar sua influência no mecanismo de expressão utilizam abordagens como Reconhecimento de Padrões, Análise Probabilística e Aprendizado de Máquina.

Esta dissertação tem como tema investigar uma metodologia para reconhecimento de regiões promotoras com base em *hidden Markov models* (HMMs), a fim de estabelecer um protocolo de aplicação desses modelos no reconhecimento e na predição de promotores procarióticos.

1.1 Motivação

As iniciativas *in silico* para análise de regiões promotoras acompanham a biologia molecular desde a primeira geração de genomas seqüenciados. Elas são impulsionadas, principalmente, por dois fatores: a complexa rede molecular ainda indecifrável envolvida na regulação da expressão gênica e o desenvolvimento da tecnologia que permite o seqüenciamento de genomas completos de uma grande variedade de espécies [Qiu, 2003]; ainda mais para organismos procarióticos que possuem genomas menores, logo, mais fáceis de seqüenciar.

Entretanto, apesar da disponibilidade de muitos genomas procarióticos e das pesquisas em (*Escherichia coli*) *E. coli* e (*Thermus aquaticus*) *T. aquaticus* que revelam dados estruturais sobre a RNAp e seus sítios de interação com o promotor, as ferramentas

computacionais pouco utilizam esses dados e hipóteses levantadas para aprimorar suas predições. Isso pode ser a explicação para os baixos índices de acerto das mesmas, que variam de 13 a 54%. Além disso, o maior problema delas é o alto número de falsos positivos gerados [Oppon, 2000; Qiu, 2003; Stormo, 2000].

A inferência de outras características que possam ser agregadas aos modelos de reconhecimento de promotores, para reduzir o número de falsos positivos obtidos, é pouco discutida para procariotos. Recentemente, para promotores eucarióticos, a combinação de métodos até então aplicados individualmente se mostrou uma boa alternativa para acréscimo da sensitividade e especificidade da predição [Qiu, 2003].

Seguindo na mesma idéia de aproveitar diversas evidências para predição, Bockhorst et al. (2003) propõe um modelo de Rede Bayesiana para predição de operons procarióticos. Essa rede é um modo de representar a distribuição de probabilidade de um conjunto de variáveis aleatórias e explorar as relações de independência condicional entre elas.

Ainda para eucariotos, a comparação entre genomas de espécies relacionadas se apresenta como uma ferramenta poderosa na identificação de sítios de interação, baseada na hipótese de que regiões funcionais são mais conservadas que as demais, não exigindo o conhecimento prévio dos mesmos [Kellis et al., 2003; Qiu, 2003]. Estudos com *Saccharomyces cerevisiae* e espécies próximas mostram que a comparação de genomas produz definições mais exatas da estrutura do gene, auxiliando na definição de promotores e outras regiões intergênicas conservadas [Kellis et al., 2003]. Essa abordagem pode contribuir para solucionar o problema da ausência de dados para treinar modelos de outros procariotos, reduzindo o número de falsos positivos.

Assim, a investigação de metodologias que considerem informações estruturais e funcionais das regiões promotoras, o uso de diferentes métodos computacionais e estatísticos

em conjunto para aproveitar seus diferentes enfoques é relevante e de caráter promissor conforme o estado da arte.

1.2 Objetivo Geral

Estabelecer um protocolo consistente para compreender, reconhecer e predizer regiões promotoras em organismos procarióticos, integrando dados genômicos experimentais e a teoria de *hidden* Markov *models* (HMMs).

1.3 Objetivos Específicos

- ⇒ Treinar modelos HMMs para reconhecimento de regiões promotoras de *Escherichia coli*.
- ⇒ Investigar padrões nos modelos HMMs treinados, e relacioná-los com características estruturais da região promotora.
- ⇒ Confrontar padrões identificados pelos HMMs com dados *in vitro* e *in silico* disponíveis.
- ⇒ Estabelecer um protocolo para utilização de HMMs.
- ⇒ Avaliar o uso do protocolo para reconhecer e predizer promotores procarióticos.
- ⇒ Aplicar a metodologia HMMs com os diferentes protocolos propostos em um grupo diverso de espécies procarióticas.

1.4 Organização do Texto

Essa dissertação envolve conhecimento teórico das áreas de Biologia Molecular, Modelagem Computacional e Bioinformática. O texto está organizado como descrito a seguir.

No capítulo 2 são apresentados os principais conceitos de biologia molecular envolvidos na regulação gênica, com enfoque na caracterização estrutural dos promotores procarióticos e nos mecanismos em que ele está envolvido.

No capítulo 3 é realizada a revisão bibliográfica das principais abordagens *in silico* para análise de promotores. Neste se destaca a metodologia de HMM, que tem sua teoria e aplicação em Bioinformática apresentadas no capítulo 4.

O capítulo 5 descreve as propostas de protocolo para empregar HMMs nas tarefas de reconhecimento e predição de promotores. O desempenho de cada uma delas é testado analisando-se quatro espécies procarióticas, cujos resultados compõem o capítulo 6.

Finalmente, o capítulo 7 reúne as conclusões finais e propostas de trabalhos futuros.

2 O Papel dos Promotores na Expressão Gênica de Organismos Procarióticos

O DNA é um polímero e sua estrutura consiste de um conjunto de resíduos de fosfato e pentose (açúcar) alternados, com uma base nitrogenada purínica ou pirimidínica ligada ao açúcar. Ocorrem quatro tipos de base nesta molécula: duas purinas, adenina (A) e guanina (G), e duas pirimidinas, citosina (C) e timina (T). Estas estão dispostas ao longo de duas cadeias, denominadas fitas, antiparalelas que se entrelaçam, formando a estrutura dupla hélice ou dupla fita, padrão do DNA [Lewin, 2001].

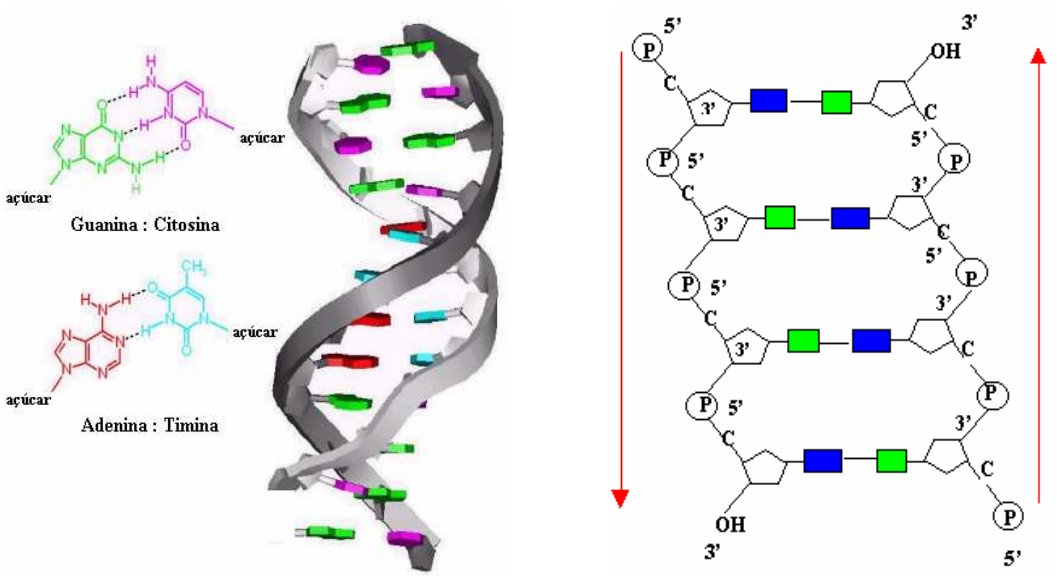
A ligação entre a base, o fosfato e o açúcar constitui um nucleotídeo (nt). Cada nucleotídeo de uma fita se liga ao complementar da outra fita, conforme a regra de pareamento A↔T e C↔G. A ligação A↔T é fraca por ocorrer através de duas pontes de hidrogênio, enquanto a C↔G é forte em função de suas três pontes de hidrogênio [Souto et al., 2003]. Mesmo que as pontes de hidrogênio sejam fracas, o grande número delas existente no DNA é suficiente para manter as duas cadeias unidas.

A partir de estudos de difração de raios-X, descobriu-se que a hélice possui um diâmetro médio característico ao longo de toda a molécula, sendo essa propriedade garantida porque as fitas são complementares e antiparalelas. A regra de complementaridade é justificada porque, de acordo com a estrutura química, se o pareamento ocorresse entre as duas purinas (A↔G), o diâmetro da hélice nessa ligação comparado ao diâmetro na ligação entre duas pirimidinas (C↔T) seria menor. Além desta ligação purina↔pirimidina, deve-se considerar que as cadeias possuem polaridade, tendo um fosfato 5’ (P) em uma extremidade e

um grupo hidroxila 3’ (OH) na outra [Lewin, 2001] (veja Figura 2.1). Toda nova fita de DNA é sintetizada na direção 5’→ 3’ [Lewin, 2001], sendo cada nucleotídeo ligado à região 3’(OH) livre do nucleotídeo anterior, em um processo conhecido como replicação.

O gene, por sua vez, é uma região nucleotídica em uma das cadeias do DNA, chamada de fita molde, que comporta a informação necessária para produção de uma unidade biomolecular específica, a qual realiza alguma função na célula. A maioria dessas unidades são proteínas, elementos essenciais para o ser vivo que podem ter função tanto estrutural, regulatória ou catalítica.

A regulação da expressão dos genes compreende um conjunto de mecanismos, incluindo reações químicas e interações físicas entre determinadas proteínas e a molécula de DNA. A expressão dos genes necessários para a execução de um processo celular é chamada de transcrição. Qual gene expressar, em que momento e em que quantidade são definidos em um fenômeno chamado de regulação gênica.



(a) (b)

Figura 2.1 – Estrutura da molécula de DNA. Em (a) são apresentadas as estruturas químicas da ligação complementar entre as bases. (b) Arquitetura esquemática da molécula mostrando as extremidades 3’ e 5’.

Mesmo para organismos procarióticos, ou seja, aqueles que não possuem membrana nuclear e, conseqüentemente, possuem o DNA livre na célula, a regulação em nível de transcrição é complexa. De modo geral, trata-se do controle subjacente à interação da enzima RNA-polimerase (RNAp) com a região de DNA *upstream* à localização do gene. Essa região, em que são encontrados os elementos regulatórios da expressão gênica, é chamada de promotor.

As próximas seções abordam os conceitos básicos envolvidos na relação dos promotores com a Regulação Gênica dos procariotos durante a transcrição. Primeiro, é apresentado o processo de transcrição dos genes como um todo, seguido da caracterização estrutural e funcional da região promotora e de seu papel regulatório.

2.1 Transcrição dos Genes

Expressar um gene significa transferir a informação contida em uma seqüência de DNA para uma molécula intermediária (mRNA), a qual é necessária para a construção de uma proteína. Segundo o dogma central da biologia molecular (Figura 2.2), essa transferência se dá em duas etapas que ocorrem simultaneamente em organismos procarióticos:

1. **Transcrição**: em que é realizada uma cópia da informação do gene numa molécula de RNA mensageiro (mRNA).
2. **Tradução**: quando estruturas celulares conhecidas como ribossomos confeccionam a seqüência de aminoácidos que constitui a proteína a partir do mRNA [Setubal e Meidanis, 1997].

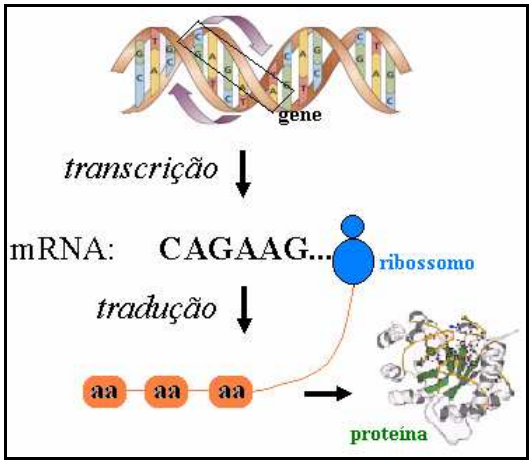


Figura 2.2 - Dogma Central da biologia molecular para a expressão gênica em organismos procarióticos. Um gene é transcrito em uma molécula de mRNA, na qual os ribossomos atuam para gerar uma cadeia de aminoácidos (aa) que constitui a proteína.

A RNA-polimerase, enzima conservada em todos os organismos vivos, é o alvo, direta ou indiretamente, da maioria dos processos regulatórios da transcrição [Naryshkin et al., 2000]. Em procariotos, ela possui seis subunidades que quando associadas compõem a RNAp holoenzima (ver Tab. 2.1). Dessas, a subunidade σ é o fator responsável pela fase de reconhecimento do promotor.

Tabela 2.1 – Descrição das subunidades da RNA-polimerase holoenzima de *Escherichia coli*.

Subunidade	Gene Codificante	Quantidade	Função na RNAp
α	rpoA	2	montagem
β	rpoB	1	Ligação dos nucleotídeos
β’	rpoC	1	Ligação ao molde
σ	rpoD	1	Ligação ao promotor
ω	rpoZ	1	acréscimo na força de associação entre as demais subunidades

A transcrição possui dois momentos chaves: a interação da RNAp com o promotor, e a cópia da seqüência do gene em um mRNA. O início de todo o processo ocorre quando a RNAp associada com o fator σ se liga a 40-60 pares de base (pb) do promotor para formar o complexo fechado do promotor (RP_c). Ela se prende, então, fortemente a ele para produzir um

complexo intermediário RNAp-promotor (RP_i), e abrir a dupla fita aproximadamente a 14 nucleotídeos antes do primeiro nucleotídeo do gene, dando acesso à informação genética contida na fita molde para gerar, finalmente, o complexo aberto RNAp-promotor (RP_o) [Mooney e Landick, 1999; Murakami et al., 2002; Naryshkin et al., 2000]. Na Figura 2.3, os dois complexos são mostrados. Quando uma cadeia de DNA não está pareada com a cadeia complementar, diz-se que ela está na forma fita simples.

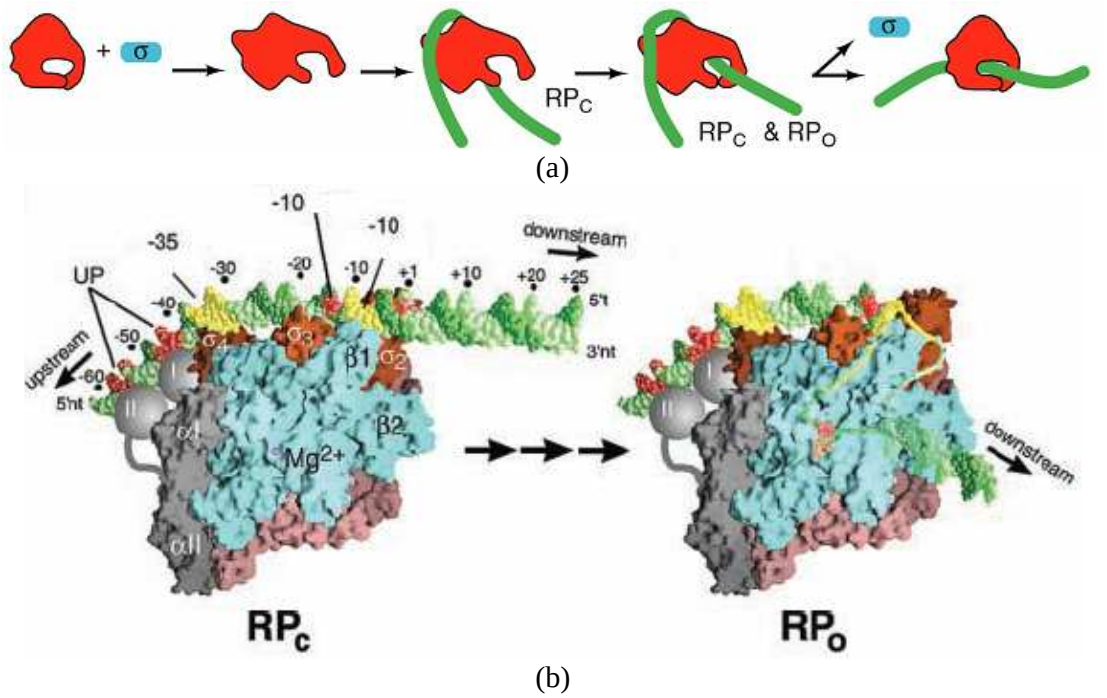


Figura 2.3 – Modelos dos complexos RP_c e RP_o formados entre a RNAp e o promotor. (a) Esquema das transições entre os complexos. (b) Visão detalhada da formação dos dois complexos [Murakami et al., 2002]. A fita molde está em verde escuro, e a complementar em verde claro. A RNAp tem suas subunidades mapeadas por cor: em azul, a subunidade β; em cinza, as duas subunidades α, em rosa, a subunidade β'; e em laranja σ. As setas representam diversas etapas intermediárias entre RP_c e RP_o

Essa etapa inicial termina com a inserção dos dois primeiros nucleotídeos na molécula de mRNA, seguida da liberação do fator σ e de proteínas regulatórias existentes [Lewin, 2001], descritas mais adiante.

No promotor é identificada a posição de início da transcrição, a partir da qual o DNA serve como molde para a síntese de RNA mensageiro. Esse sítio é conhecido como sítio de início da transcrição ou TSS – *transcription start site*, numerado como +1. A posição dos nucleotídeos anteriores a ele é localizada por inteiros negativos (região *upstream*), e os posteriores, por inteiros positivos (região *downstream*).

O término da transcrição é sinalizado por uma seqüência de nucleotídeos encontrada no seguimento posterior ao final do gene, denominada região terminadora.

Além da interação da RNAp com o genoma, existem proteínas regulatórias que se ligam em partes específicas do DNA, chamadas de operadores, inibindo ou ativando a transcrição em diferentes taxas.

Assim, podemos dizer que os elementos estruturais presentes no DNA atuantes na transcrição procariótica são: promotor, gene, operador e terminador. Veja a Figura 2.4.

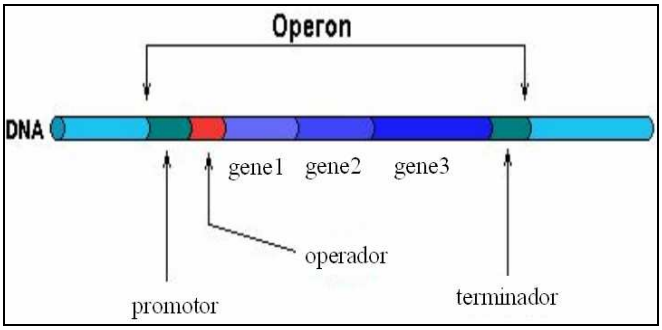


Figura 2.4 – Elementos estruturais do DNA que participam da transcrição.

A maior parte dos genes de procariotos está organizada na forma de operons. Um operon é definido como o conjunto de genes adjacentes, cuja transcrição é regulada pela mesma região promotora. O produto fundamental da transcrição é a seqüência de um ou mais genes que, dadas condições específicas, são transcritos como uma molécula de mRNA. Esse produto é denominado de unidade de transcrição [Bockhorst et al., 2003]. As proteínas

geradas a partir desta unidade geralmente estão envolvidas em uma mesma tarefa celular. Enquanto o mRNA é criado, os ribossomos executam a tradução do mesmo em proteína (Figura 2.2). As proteínas são constituídas de aminoácidos, os quais são codificados a partir de cada grupo de três nucleotídeos, chamado códon, do mRNA. Na natureza existem 20 aminoácidos, mas há 64 códons possíveis. Logo, alguns deles são mapeados por mais de um códon [Lewin, 2001].

Neste trabalho, o interesse está concentrado nas características que configuram uma região de DNA como promotora. Desta forma, a seguir, são apresentados os conhecimentos moleculares já consolidados sobre a estrutura e função dos promotores.

2.2 Promotores Procarióticos

Os estudos experimentais sobre promotores utilizam, em sua maioria, a bactéria *Escherichia coli* (*E. coli*) como modelo biológico procariótico, enquanto os relativos à estrutura da RNAp usam o núcleo desta mesma enzima do *Thermus aquaticus* (*T. aquaticus* ou Taq) [Mooney e Landick, 1999], bactéria resistente a altas temperaturas. Assim, as características descritas aqui se referem a esses dois organismos.

Como já foi dito anteriormente, os promotores são regiões do DNA que antecedem genes, sendo reconhecidas por proteínas específicas. Os procarióticos aparentam ser menos complexos que os eucarióticos (organismos com núcleo celular), apresentando um tamanho menor e um número reduzido de elementos reconhecidos pelo fator σ [Oppon, 2000].

Pesquisas em laboratório identificaram diferentes tipos de fatores σ, estando cada um deles associado à transcrição de uma classe de promotores que regula um conjunto de genes que precisam ser expressos em um mesmo momento celular. Para a *E. coli*, as informações sobre os fatores σ são apresentadas na Tab. 2.2. O número correspondente ao fator σ é dado pelo seu respectivo peso molecular [Lewin, 2001].

Tabela 2.2 – Propriedades dos fatores σ de *E. coli*.

Fator σ	Gene codificante do respectivo fator σ	Momento celular de ativação
28	rpoF	resposta a estresse por falta de alimento
32	rpoH	resposta a choque térmico
38	rpoS	genes flagelares
54	rpoN	assimilação de nitrogênio
70	rpoD	condição padrão; sigma primário ou constitutivo

Murakami et al.. (2002), em seus estudos a respeito da estrutura cristalizada da RNAp de *T. aquaticus* complexada com fragmentos de promotores, revelaram que todas as ligações específicas com elementos desta região são mediadas por regiões conservadas da subunidade σ . A Figura 2.5 mostra esquematicamente esses resultados.

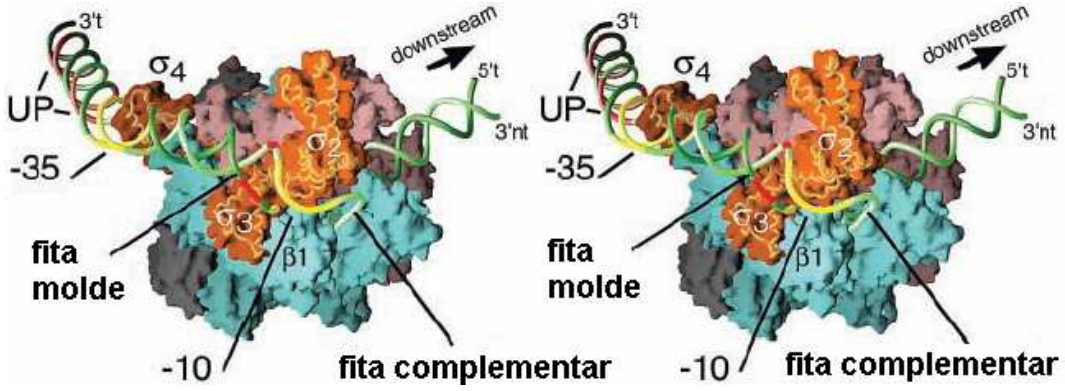


Figura 2.5 – Modelo RP_o com destaque para a subunidade σ (em laranja) [Murakami et al., 2002]. Observe que partes de σ (2,3,4) interagem com o DNA nas regiões próximas aos nucleotídeos -35 e -10, em amarelo.

Entre as funções dos promotores estão: a identificação da base de início da transcrição (+1), geralmente uma adenina (A) ou citosina (C); e a determinação da taxa de transcrição ou força do promotor, a qual depende da afinidade de ligação da RNAp e da taxa de

isomerização do complexo fechado do promotor para aberto, ou seja, da mudança de DNA dupla fita para fita simples [Lewin, 2001].

Os promotores do tipo σ^{70} e seus equivalentes em outros procariotos, devido à sua condição basal, isto é, por transcreverem a maioria dos genes, são os mais estudados. Segundo os resultados dessas pesquisas, um promotor é constituído de três regiões características (Figura 2.6):

- ⇒ uma seqüência de 6 nucleotídeos (hexâmero) centrada em -35 do sítio inicial de transcrição +1;
- ⇒ um hexâmero centrado a -10 ;
- ⇒ e a região que separa os 2 hexâmeros.

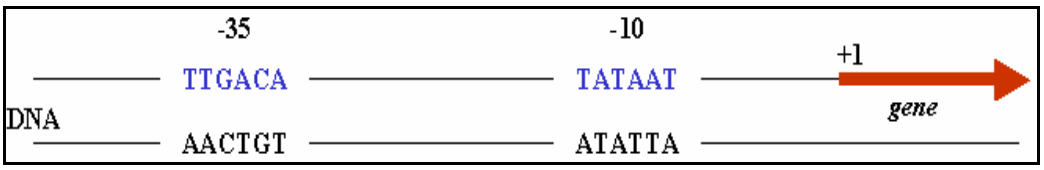


Figura 2.6 – Região promotora procariótica e seus elementos principais.

A constituição nucleotídica dos hexâmeros apresenta um alto grau de conservação quando são comparados promotores de um organismo, e até mesmo de organismos diferentes. O -35 funciona como sinal para reconhecimento pela RNAp, enquanto o -10 permite converter o complexo de fechado para aberto. Além disso, a distância entre eles, em média de 17 nucleotídeos, parece ser relevante, apesar do tamanho variável e da baixa conservação nas bases que a compõe, pois pode ser crítica na interação dos hexâmeros com a geometria da RNAp [Oppon, 2000].

A formação dos complexos entre a RNAp e o promotor influencia essa caracterização, porém não há registro de informação estrutural em alta resolução sobre esses [Naryshin et al.,

2000]. É sabido que, na transição do RP_c para o RP_o (ver seção 2.1), existe pelo menos um estado intermediário que depende de fatores como a concentração de Mg^{2+}, temperatura e região promotora [Murakami et al., 2002].

Análises do RP_o para identificar sítios específicos de interação no promotor *lac(ICAP)UV5* do *T. aquaticus* foram realizados com técnicas de *Photocrosslinking*. Como resultado, foram mapeadas as interações das diferentes subunidades da RNAp (Figura 2.7) nas duas fitas do DNA. É importante destacar que os sítios de interação das partes da RNAp, para promotores em geral, podem ser alterados em função da ação de proteínas regulatórias [Naryshkin et al., 2000].

Apesar dos livros de biologia molecular descreverem os promotores de uma maneira bastante simplificada, os trabalhos experimentais recentes, como os de Murakami et al. (2002) e Naryshkin et al. (2000), revelam propriedades muito mais complexas, mas sem impacto considerável no entendimento minucioso dos mecanismos regulatórios.

Identificar e caracterizar a estrutura dos promotores através de técnicas *in vitro* ainda é uma tarefa cara e demorada, de modo que somente para a *E. coli* é conhecido um número relevante dessas regiões. Logo, ferramentas computacionais para reconhecimento e predição de promotores podem impulsionar o conhecimento sobre a relação dessas regiões com o processo regulatório da expressão dos genes.

A primeira tentativa de caracterizar promotores procarióticos baseou-se na freqüência de cada tipo de nucleotídeo nos sítios dessas regiões da *E. coli*, as quais foram alinhadas uma abaixo da outra.

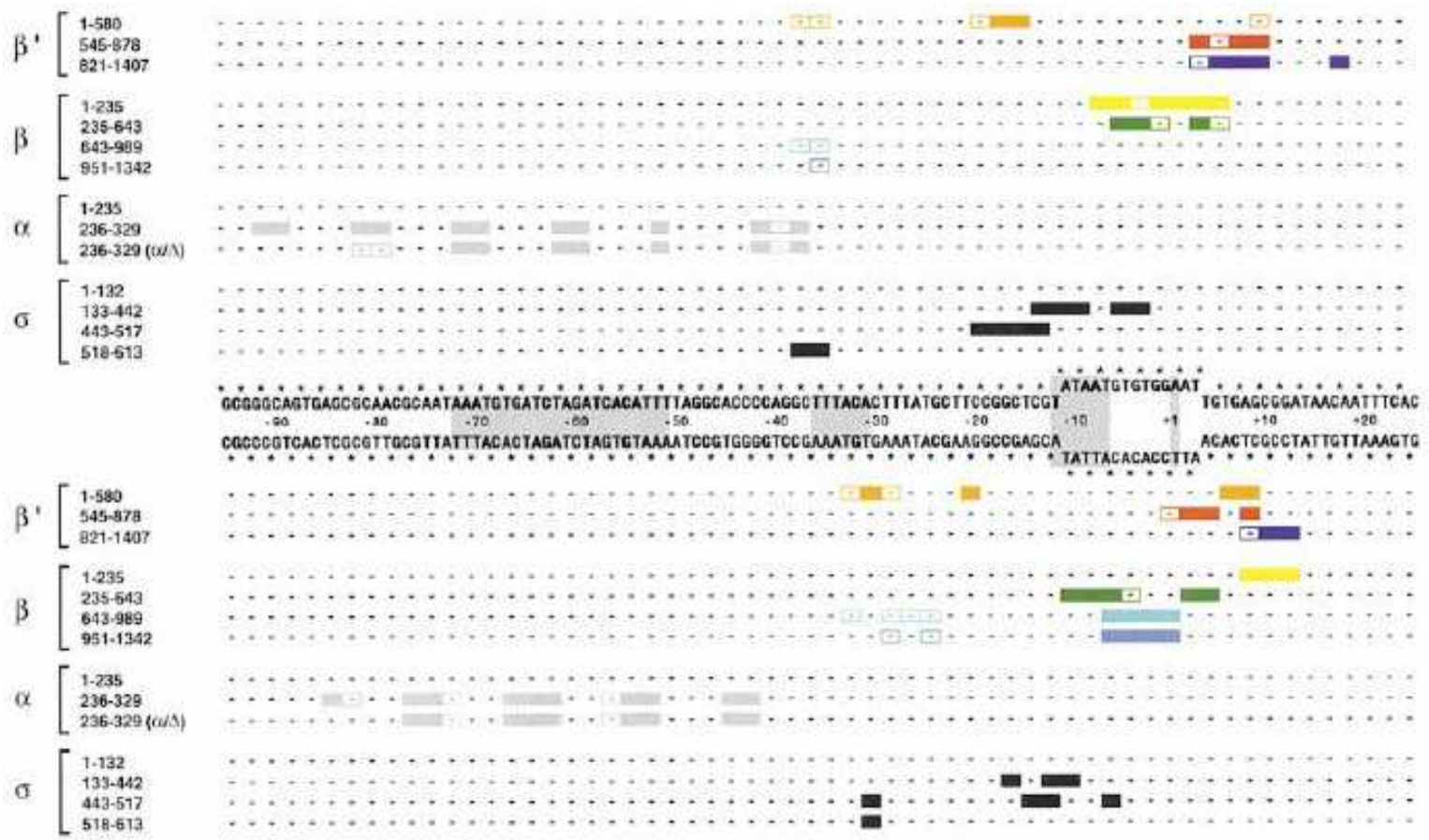


Figura 2.7 – Sítios específicos de interação RNAp-DNA identificados por *Photocrosslinking* para o promotor *lac(ICAP)UV5* do *T. aquaticus* [Naryshkin et al., 2000]. Acima está a fita molde, e abaixo, sua complementar. O * sinaliza os fosfatos considerados no experimento *in vitro*. Barras preenchidas indicam ligações fortes, enquanto as abertas, ligações fracas. Cada cor sinaliza uma parte específica da RNAp que interage com o promotor.

Assim, resultou o promotor ideal ou consenso, definido por TTGACAN_{17}TATAAT, onde *N* corresponde a qualquer uma das bases. Porém, esse padrão não é encontrado com identidade total em nenhuma região promotora já identificada experimentalmente [Pevzner, 2000]. Além disso, os promotores considerados fortes tendem a não possuir mais do que três nucleotídeos diferentes do consenso [Lewin, 2001].

Desta forma, muitas pesquisas já foram desenvolvidas para análise de promotores. O próximo capítulo discute os principais trabalhos de abordagem *in silico* com enfoque em procariotos.

3 Análise de Promotores *in silico*: Revisão Bibliográfica

Os conceitos biológicos sobre a expressão gênica em procariotos vistos no capítulo anterior fortalecem o aspecto chave da região promotora nos mecanismos de regulação.

Ainda que os promotores sejam estruturas de importância indiscutível, a habilidade em identificá-los é menos desenvolvida comparada à de encontrar regiões codificantes em DNA. Isso acontece porque os promotores são muito divergentes, e até os seus padrões mais característicos, como os hexâmeros centrados nos sítios -35 e -10, não são conservados [Qiu, 2003].

Nos últimos anos, o uso de ferramentas computacionais tem se mostrado importante para inferir funções e estruturas de seqüências e proteínas regulatórias [Oppon, 2000]. Para promotores, os algoritmos se enquadram em três abordagens [Qiu, 2003]:

- ⇒ Baseada em sinal: que opera no reconhecimento de sinais relativamente conservados, assim como de distâncias entre esses elementos;
- ⇒ Baseada em conteúdo: que utiliza as diferenças do conteúdo das seqüências para classificá-las em promotor ou não-promotor, por exemplo, preferência de códons para codificação de aminoácidos na região próxima ao sítio +1;
- ⇒ Aprendizado de Máquina (AM): que usam um conjunto de informações estruturais e funcionais disponíveis sobre os promotores para “aprender” automaticamente a reconhecê-los, e produzir hipóteses relevantes sobre os mesmos [Baldi e Brunak, 2001].

Diferentes aplicações dessas abordagens para o reconhecimento e a predição de promotores são encontradas na literatura. Com elas, almeja-se não apenas criar recursos

eficientes para classificação dessas regiões, mas também descobrir outras propriedades físico-químicas com influência na ação dos promotores na expressão em procariotos.

Os principais métodos computacionais das abordagens baseadas em sinal e em aprendizado de máquina encontrados na literatura são apresentados a seguir, acompanhados da discussão das limitações que tornam esse campo de pesquisa latente. Especificamente para o reconhecimento de promotores procarióticos, a abordagem baseada em conteúdo não apresenta uso relevante, logo, não será considerada.

3.1 Reconhecimento Baseado em Sinal

3.1.1 Seqüência Consenso

O método clássico para analisar promotores é a determinação da região consenso. Ele consiste em alinhar um conjunto de seqüências identificadas previamente como promotoras pelo seu sítio de início da transcrição, para, posteriormente, pesquisar por regiões conservadas no interior delas. Para cada coluna no alinhamento é fornecida a variação encontrada nesta posição do promotor.

Apesar dos hexâmeros -35 e -10 terem sido encontrados manualmente considerando a idéia desta técnica, ela é muito simples e imprecisa [Souto et al., 2003; Stormo, 2000]. Lisser e Margalit (1993) apresentam a distribuição de bases para os consensos dos hexâmeros, de acordo com 298 promotores identificados experimentalmente em sua pesquisa. Os resultados constam na Tab. 3.1.

O uso da seqüência consenso para pesquisar novos promotores é restringida pela variação nucleotídica encontrada na maioria das posições do alinhamento. Logo, uma abordagem alternativa para aproveitar essa informação de variação entre as seqüências pode ser utilizada para criar um método mais eficaz.

Tabela 3.1 – Freqüências em % dos nucleotídeos para os padrões -35 e -10, de acordo com um conjunto de 298 da compilação de Lisser e Margalit (1993).

	-35						**-10**					
A	10	6	9	56	21	54	5	76	15	61	56	6
C	10	7	12	17	54	13	10	6	11	13	20	7
G	10	8	61	11	9	16	8	6	14	14	8	5
T	69	79	18	16	16	17	77	12	60	12	15	82
consenso:	**T**	**T**	**G**	**A**	**C**	**A**	**T**	**A**	**T**	**A**	**A**	**T**

3.1.2 Matriz de Posições Ponderadas

As Matrizes de Posições Ponderadas (*weighted position matrix*) fundamentam-se na modelagem dos sinais [Souto et al., 2003]. Cada linha da matriz corresponde a um dos quatro nucleotídeos, e cada coluna a uma posição do alinhamento. Os elementos da matriz são os pesos utilizados para pontuar uma seqüência teste, conforme uma medida que quantifica o quanto esta se adere ao modelo [Hertz e Stormo, 1999].

A pontuação é dada pela soma dos pesos de cada letra alinhada em cada posição.

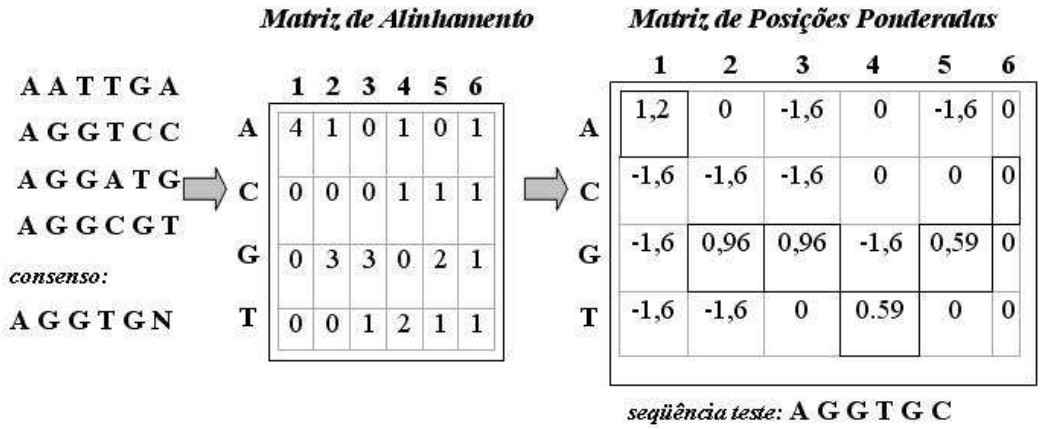


Figura 3.1 – Exemplo da transformação da Matriz de Alinhamento para a Matriz de Posições Ponderadas, em que é apresentado como identificar os pesos para a seqüência teste AGGTGC [Hertz e Stormo, 1999].

A Figura 3.1 exemplifica como criar uma matriz de peso a partir de uma matriz de alinhamento. Essa transformação é feita pela fórmula:

$$\ln \frac{(n_{i,j} + p_i)/(N+1)}{p_i},$$

onde $n_{i,j}$ é o número de vezes que a letra *i* é encontrada na posição *j* do alinhamento; p_i é a probabilidade *a priori* da letra *i*, no caso, 0,25 para todas as bases; e *N* é o número total de seqüências analisadas.

3.1.3 Entropia Relativa

Muitas vezes para encontrar padrões em regiões de DNA, os métodos computacionais perdem desempenho devido a características intrínsecas aos genomas dos organismos analisados.

A característica mais comum é a tendência na freqüência de algum dos nucleotídeos. Por exemplo, se um dado genoma possui a freqüência de adeninas em torno de 70%, detectar padrões formados por muitos *A*s seria freqüente, os quais podem não conter relevância biológica [Pevzner, 2000].

O cálculo de entropia relativa é apontado como uma solução para identificar sinais em regiões com tendência nucleotídica. Ela é definida por [Pevzner, 2000]:

$$-\sum_{j=1}^{k} \sum_{n=A,T,C,G} p_{nj} \log_2 \frac{p_{nj}}{t_n},$$

onde *k* é o tamanho da seqüência; p_{nj} é a freqüência do nucleotídeo *n* na posição *j* entre as ocorrências do sinal; e t_n é a freqüência do nucleotídeo *n* no conjunto de seqüências.

3.1.4 Logos de Seqüências

Os logos de seqüências podem ser vistos como uma técnica que resume graficamente informações quantitativas e qualitativas, representando a conservação de seqüências em um conjunto de sítios de interação de proteínas com o DNA.

A representação é feita através de letras empilhadas ordenadas por tamanho, com a maior localizada no topo da pilha. A altura de cada letra é proporcional à freqüência do nucleotídeo que ela representa, sendo a visualização da quantidade de informação contida em um dado sítio. Em suma, este é mais um modo de investigar o alinhamento de seqüências [Gibas e Jambeck, 2001; Schneider, 2001; Schneider, 2003].

Esse método foi desenvolvido por Thomas D. Schneider com base em Teoria da Informação. Sua metodologia consiste de seis etapas [Schneider, 2003]:

1. Calcular o número de ocorrências de seqüências com nucleotídeo em um dado sítio, visto que as regiões promotoras não precisam ter tamanho idêntico, dado por:

$$n(l) = \sum_{b=A,T,C,G} n(b,l) .$$

O l representa cada posição das seqüências alinhadas.

2. Calcular a freqüência de bases em cada posição, definida por:

$$f(b,l) = \frac{n(b,l)}{n(l)} .$$

3. Estimar a incerteza de *Shannon* a partir de:

$$H = -\sum_{i=1}^{M} f_i \log_2 f_i + e(n) \text{ bits/símbolo.}$$

M é o número de símbolos; e $e(n)$ é um fator de correção para substituir a probabilidade do *i*-ésimo símbolo com freqüência f_i, que permite reduzir o caráter de tendência quando a amostra de seqüências alinhadas é pequena.

Os padrões de interação proteína-DNA são modelados conforme dois estados termodinâmicos: *antes* e *depois* da ligação.

4. Calcular *H* para o estado "antes da interação", dado por:

$$H_{antes} \cong 2 \text{ bits/base.}$$

Obtém-se esse resultado porque se considera todas as quatro bases igualmente possíveis antes da interação, logo f_i está em torno de 0,25.

E calcular *H* para o estado "depois da interação", dado por:

$$H_{depois} = H(l) = -\sum_{b=A,T,C,G} f(b,l)\log_2 f(b,l) + e(n(l)) \text{ (bits/base).}$$

5. A informação em cada posição é dada pelo decréscimo na incerteza gerado pela transição entre os dois estados termodinâmicos:

$$R_{seqüência}(l) = H_{depois} - H_{antes} \text{ (bits/base).}$$

R é o ganho de informação em bits por base.

6. Calcular o total de informação, definido por:

$$R_{seqüência} = \sum R_{seqüência}(l) \text{ (bits/base).}$$

Schneider (2001), em sua análise de seqüências biológicas por logos, encontrou uma alta conservação do hexâmero -10. Além disso, o logo para regiões promotoras procarióticas apresenta uma extremidade altamente conservada próxima ao final 3' das seqüências consideradas. Segundo a discussão do mesmo trabalho, o *T* altamente conservado no sítio -7 é

justificado por estudos experimentais que apontam ele como pertencente ao primeiro pareamento rompido na dupla fita de DNA durante o início da transcrição [Schneider, 2001]. Veja a Figura 3.2.

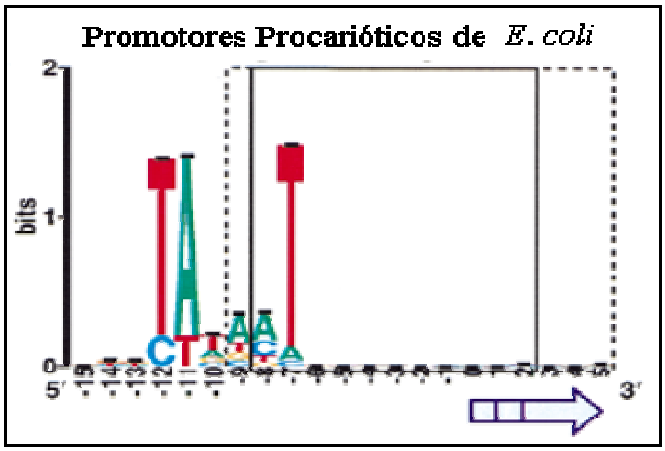


Figura 3.2 – Logo para análise do hexâmero -10 de promotores de *E. coli*. A seta representa a direção da transcrição. Observe o sítio -7, possível base em que a dupla fita do DNA começa a ser rompida [Schneider, 2001].

3.2 Análise por Aprendizado de Máquina

3.2.1 Redes Neurais

As Redes Neurais Artificiais (RNAs) tratam-se de um sistema de aprendizado de máquina inspirado no funcionamento de redes neurais biológicas. Assim, é um modelo computacional paralelo constituído de elementos interconectados, chamados de neurônios. As conexões entre eles possuem pesos que são ajustados na etapa de treinamento. Pode-se afirmar que as RNs "aprendem" dos exemplos e apresentam alguma capacidade de generalização do conjunto de treinamento [Wu e McLarty, 2000].

Em 1990, Towell et al. (1990) estabeleceram um sistema híbrido de Redes Neurais e Regras Simbólicas, o KBANN (*Knowledge Based Neural Network*), que faz uso de métodos de aprendizado empíricos para refinar a correção do conhecimento adquirido pelo modelo.

Em seus experimentos, o KBANN é utilizado também para reconhecer promotores, sendo a topologia da rede e seus pesos inicias estabelecidos por um biólogo. As regras consideram dois aspectos: a região de contato e a região de conformação. As de contato consideram os hexâmeros -35 e -10, enquanto as de conformação estabelecem maneiras alternativas de como e onde esses podem ocorrer [Craven e Shavlik, 1994]. Como resultado, as RNs apresentaram desempenho superior a outros métodos referenciados na literatura, o que confirma a eficácia do uso de AM para estudo de promotores. Sobre o uso de regras, os experimentos mostraram que as conformacionais não agregam ganho significativo à predição. Esse ponto se revela contraditório aos estudos *in vitro* abordados no capítulo anterior, pois é esperado que as mudanças conformacionais que ocorrem na transição de RP_c para RP_o estejam relacionadas com os sinais da região promotora.

Pedersen e Engelbrecht (1995) utilizaram RNs para predizer se um dado nucleotídeo era ou não um sítio inicial de transcrição (+1). Além disso, a rede neural não buscava predizer regiões promotoras, mas sim medir o conteúdo de informação em diferentes segmentos das regiões. Mais que a descoberta dos hexâmeros nos sítios -35 e -10, os resultados apontam outras regiões conservadas correlacionadas com o sítio +1, as quais estão distanciadas por uma volta da hélice do DNA (≈ 10,5 pb), observação consistente com a hipótese de que a RNAp interage com o promotor principalmente ao longo de uma das fitas [Pedersen e Engelbrecht, 1995].

Outra ferramenta baseada em RN é o *Neural Networks Promoter Prediction* (NNPP). Oppon (2000) executou um teste neste sistema a partir de um conjunto composto de 31 seqüências de 75 bases, sendo 5 de regiões promotoras e 26 de regiões codificantes de *E. coli*. Com um limiar de 6,0, o NNPP acerta ao dizer que é promotor 60% das vezes, e em afirmar que não é promotor 50% das vezes. Uma provável explicação para essa performance baixa é a questão do sistema ter sido projetado para um organismo procariótico específico.

3.2.2 Modelos Ocultos de Markov – HMMs

Como os promotores podem ser tratados como seqüências de nucleotídeos que apresentam bases com diferentes graus de conservação em cada posição, o HMM, o qual pode ser definido com um autômato estocástico de estados finitos, se mostra adequado frente à sua capacidade de capturar regularidades em seqüências de caracteres, considerando a variação nos símbolos observados em cada estado [Clote e Backofen, 2000; Mount, 2000].

Os HMMs tornaram-se a técnica de Aprendizado de Máquina bastante utilizada no estudo de promotores. Essa preferência está firmada na hipótese de que regiões características de promotores, relevantes para que a RNAp se direcione corretamente ao sítio +1, devem se apresentar conservadas entre os promotores de um genoma ou até mesmo entre os promotores de genomas de organismos próximos evolutivamente [Oppon, 2000]. Além disso, sua aplicação não precisa de um alinhamento prévio das seqüências, e se mostra melhor que os métodos estatísticos mais simples já abordados na seção 3.1 [Pedersen et al., 1996], pois esses falham em considerar as posições ao longo da cadeia como estatisticamente independentes [Souto et al., 2003]; o que sabemos ser incorreto no caso de análise de segmentos de DNA.

Pedersen et al. (1996) treinou um modelo HMM com um conjunto de 166 seqüências promotoras de *E. coli* da base de dados de Lisser e Margalit (1993). A ênfase do trabalho foi a análise do fato dos promotores serem divididos em classes de acordo com o fator σ que o reconhece. Os resultados mostram que o HMM apresenta um excelente índice de classificação para promotores desconhecidos com respeito à classe sigma, além de conseguir “aprender” a estrutura seqüencial presente em promotores procarióticos [Pedersen et al., 1996].

A Figura 3.3 apresenta o gráfico com os padrões encontrados pelo modelo HMM de Pedersen et al. (1996). Note que os hexâmeros conservados em -35 e -10 são detectados.

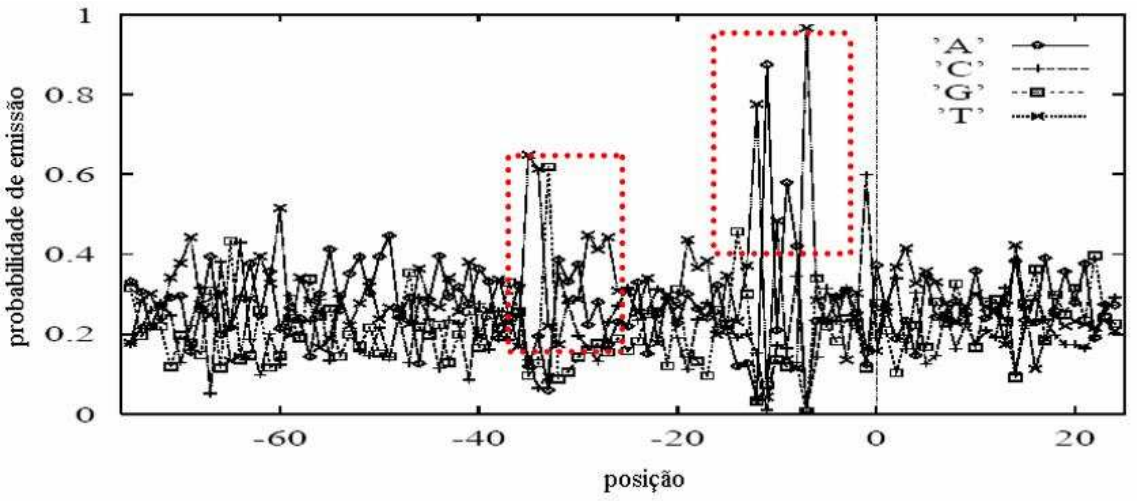


Figura 3.3 – Gráfico com as probabilidades de emissão para os estados de pareamento do modelo HMM de Pedersen et al. (1996). Observe a identificação dos padrões -35 e -10 (linhas pontilhadas), sendo este último mais evidente.

Uma das restrições do HMM é o tamanho do conjunto de treinamento, visto à grande quantidade de parâmetros que precisam ser estimados no modelo. Oppon (2000) é o primeiro a investigar especificamente esta restrição. Segundo sua pesquisa, um modo de aumentar a eficiência no treinamento de modelos para identificar promotores, quando se tem um conjunto mínimo de treinamento, é integrar diferentes sistemas de predição, no caso, RNs, HMMs e análise estatística [Oppon, 2000].

3.3 Considerações

A revisão bibliográfica sobre os métodos computacionais de análise de promotores procarióticos mostra uma alta diversidade de técnicas, assim como de hipóteses estruturais e funcionais para essas regiões. Apesar de algumas conclusões contraditórias, percebe-se que a identificação dos hexâmeros conservados em -35 e -10 sempre está entre os resultados.

Além disso, entre as técnicas discutidas, vê-se o HMM como a mais relevante, principalmente por sua capacidade de representar de certa forma a estrutura do DNA, considerando alguma dependência entre os sítios do promotor.

Diante dessas observações, o próximo capítulo descreve os conceitos básicos da teoria de HMMs, base deste trabalho, e seu emprego na análise de seqüências biológicas.

4 HMMs Aplicados em Bioinformática

Os modelos de Markov foram desenvolvidos por Andrei A. Markov em 1913. Com propósito de estudos em lingüística, ele realizava a modelagem da seqüência de letras em textos da literatura russa. Desde então, esses modelos foram desenvolvidos como uma ferramenta estatística genérica [Manning, 2001].

Baum e Petrie (1966) publicaram o artigo "*Statistical inference for probabilistic functions of finite state Markov chains*", o qual teve grande contribuição para a teoria dos modelos ocultos de Markov – HMMs. Amplamente utilizados em estudos de reconhecimento de voz, apenas na década de 80, seus fundamentos teóricos foram explicitados em diversos trabalhos, o que tornou a técnica bastante conhecida [Meneses, 2002].

Na década de 90, o HMM tornou-se o modelo probabilístico com ampla aceitação nas tarefas de Bioinformática, por permitir a modelagem, o alinhamento e a análise de seqüências em combinação com técnicas de Aprendizado de Máquina [Baldi e Brunak, 2001; Guimarães e Melo, 2003].

Neste capítulo é realizada uma revisão geral sobre os conceitos de Modelos de Markov, Modelos Ocultos de Markov, e as aplicações destes últimos nos diferentes contextos de análises de seqüências em Bioinformática.

4.1 Conceitos de Cadeias de Markov

Um processo de Markov é um processo estocástico que possui um comportamento dinâmico, tal que as distribuições de probabilidade para o seu desenvolvimento futuro depende de *n* estados passados. Sua representação formal se dá por um sistema de transições

de estados, sendo estes representados por vetores probabilísticos que podem variar no espaço temporal (discreto ou contínuo), e as transições entre eles são probabilísticas e dependentes do estado atual [Dimuro et al., 2002].

Cadeia de Markov é o modelo de Markov em que o espaço de estados é discreto e enumerável. Quando o processo de Markov é uma aproximação, ou seja, em que nem todos os estados são perfeitamente conhecidos, tem-se um modelo oculto ou HMM. O interesse central neste tipo de modelo é sua capacidade de capturar a essência do processo escondido sobre eles [Dimuro et al., 2002].

4.2 Modelos Ocultos de Markov

HMM é um modelo de Markov de tempo discreto com algumas características extras que o tornam mais genérico e flexível, permitindo a modelagem de fenômenos para os quais o modelo de Markov convencional se mostra insuficiente. Entre essas características, destaca-se o fato de que toda vez que um estado é visitado pela cadeia, este "emite" um símbolo de um alfabeto de forma tempo-independente e estado-dependente da distribuição de probabilidades sobre o alfabeto [Durbin et al., 1998; Ewens e Grant, 2001].

Assim, tem-se um processo dividido em 2 etapas: transição e emissão. Isso faz com que, ao final do processo, seja gerada uma seqüência de estados visitados e uma seqüência de símbolos observada [Ewens e Grant, 2001].

Um modelo oculto de Markov também pode ser definido como uma máquina de estados estocástica que gera um símbolo cada vez que uma transição ocorre de um estado para outro, sendo esses estados conectados em uma estrutura de grafo direcionado.

O modelo HMM é composto de cinco elementos:

- $\Rightarrow$ um conjunto de N estados $S_1, S_2, \ldots S_N$.
- $\Rightarrow$ um alfabeto de M símbolos observados distintos $A = \{a_1, a_2, \ldots, a_M\}$.

- $\Rightarrow$ uma matriz de probabilidades de transição $T = (p_{ij})$, onde $p_{ij} = P(q_{t+1} = S_j \mid q_t = S_i)$.
- $\Rightarrow$ a matriz de probabilidade de emissão de cada símbolo a em A para cada estado S_i, isto é, $B = b_i(a) = P(S_i$ emitir o símbolo $a)$, sendo B de ordem $N \times M$.
- $\Rightarrow$ um vetor de distribuições iniciais $\pi = (\pi_i)$, onde $i = P(q_1 = S_i)$.

Observe que os dois primeiros elementos fornecem a estrutura do modelo, enquanto os demais descrevem os parâmetros do mesmo, em suma $\lambda = (T, B, \pi)$ representa o conjunto completo de parâmetros.

HMM possui dois estados especiais, o *início* e o *fim*, que determinam o começo e o final da cadeia respectivamente. O sistema sempre se encontra inicialmente no estado *início*, e passa de estado em estado até o estado *fim*, enquanto emite símbolos a cada transição ocorrida. Estando em um dado estado i, ele possui uma probabilidade p_{ji} de realizar a transição para o estado j e uma probabilidade $b_{i(a)}$ de emitir o símbolo a pertencente ao alfabeto A [Baldi e Brunak, 2001; Guimarães e Melo, 2003].

Ser de primeira ordem significa que no modelo as emissões e transições dependem exclusivamente do estado ocupado no instante imediatamente anterior [Durbin et al., 1998]. Apenas os símbolos emitidos nos estados são observados, e não os caminhos entre os eles, assim esses podem ser considerados como variáveis escondidas, o que explica a denominação oculto (*hidden*) do modelo [Baldi e Brunak, 2001; Durbin et al., 1998; Guimarães e Melo, 2003].

Uma importante característica dos HMMs é que se pode responder eficientemente algumas questões sobre a seqüência de símbolos observada O e a seqüência de estados visitados Q. Os três problemas básicos são [Dimuro et al., 2002]:

1. Qual a probabilidade de um HMM com parâmetros λ gerar uma dada seqüência, $P(O \mid \lambda)$?

2. Qual a seqüência de estados que possui mais alta probabilidade de ocorrer dada uma seqüência observada, $\text{argmax}_Q P(Q \mid O)$?
3. Qual o conjunto de parâmetros λ que maximiza a probabilidade de uma seqüência observada dada uma topologia fixa do modelo, $P(O \mid \lambda)$?

Para responder essas perguntas, existe um conjunto de algoritmos descritos na próxima seção.

4.2.1 Algoritmos

A teoria de HMMs requer freqüentemente o cálculo das questões que encerram a seção anterior. Para isso, existe um conjunto de algoritmos baseados em programação dinâmica a fim de encontrar a melhor resposta [Eddy, 1998; Guimarães e Melo, 2003]. A seguir é apresentado cada algoritmo que resolve o respectivo problema mencionado anteriormente.

4.2.1.1 Algoritmo *forward-backward*

O algoritmo *forward-backward* resolve o primeiro problema, em que queremos determinar a probabilidade de uma seqüência observada ter sido gerada por um modelo HMM. Esse cálculo é utilizado para a tarefa de reconhecimento de seqüências [Dimuro et al., 2002].

Assumindo uma seqüência de estados $Q = \{q_1, q_2, \ldots q_T\}$, a probabilidade de uma seqüência observada ter sido gerada por um HMM é dada por:

$$P(O \mid Q, \lambda) = \prod_{t=1}^{T} P(O \mid q_t, \lambda) = b_{q1}(o_1) b_{q2}(o_2) \ldots b_{qT}(o_T),$$

enquanto a probabilidade da seqüência de estados dado um HMM é:

$$P(Q \mid \lambda) = \pi_{q1} a_{q1q2} a_{q2q3} \ldots a_{qT-1qT}.$$

Dessa forma, a probabilidade conjunta da seqüência observada e da seqüência de estados dado um HMM é obtido pelo produto do resultado das duas equações anteriores, ou seja,

$$P(O,Q \mid \lambda) = P(O \mid Q,\lambda)P(Q \mid \lambda).$$

Assim, o cálculo da probabilidade de uma seqüência observada dado um HMM é resultante da soma da probabilidade conjunta considerando todas as seqüências de estados possíveis:

$$P(O \mid \lambda) = \sum_{todosQ} P(O,Q \mid \lambda)$$

Devido ao grande número de multiplicações, o cálculo dessa probabilidade é lento mesmo computacionalmente. Para contornar isso, Baum propôs um algoritmo progressivo-recursivo, chamado de *forward-backward* [Dimuro et al., 2002].

A idéia do algoritmo baseia-se em computar:

- ⇒ a variável progressiva $\alpha_t(i)$: a probabilidade de observar uma seqüência até um instante *t*, considerando a ocorrência do estado S_i no instante *t*.
- ⇒ a variável regressiva $\beta_t(i)$: a probabilidade de observar uma seqüência entre *t*+1 e *T* dado que ocorreu o estado S_i no instante *t*.

Logo, a resposta para o primeiro problema é respondida pela equação:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \beta_t(i)\alpha_t(i).$$

4.2.1.2 Algoritmo de *Viterbi*

Determinar a seqüência de estados correspondente a uma seqüência observada requer que seja adotado um critério, uma vez que uma mesma seqüência de observações pode ser gerada por diferentes seqüências de estados.

O critério mais aceito é escolher a seqüência de estados que gera a seqüência observada com mais alta probabilidade, ou seja, maximizar $P(Q, O|\lambda)$. O algoritmo de *Viterbi* realiza essa maximização de modo eficiente. Ele está dividido em duas partes: primeiro ele encontra a probabilidade máxima, depois ele volta para descobrir a Q que gera tal probabilidade.

Esse algoritmo é definido pelos quatro passos a seguir [Baldi e Brunak, 2001; Meneses, 2002]:

1. Inicialização:

   Para $1 \leq i \leq N$:

$$\delta_1(i) = \pi_i b_i(O_1)$$

$$\psi_1(i) = 0$$

2. Recursão:

   Para $2 \leq t \leq T$, $1 \leq j \leq N$ e $1 \leq i \leq N$:

$$\delta_t(j) = \max(b_j(o_t)\delta_{t-1}(i)a_{ij})$$

$$\psi_t(i) = \arg\max(\delta_{t-1}(i)a_{ij})$$

3. Finalização:

   Para $1 \leq i \leq N$:

$$P^* = \max(\delta_t(i))$$

$$q_T^* = \arg\max(\delta_t(i))$$

4. Definição da melhor seqüência:

   Para $T \geq t \geq 1$:

$$q_t^* = \psi_{t+1} q_{t+1}^*$$

4.2.1.3 Algoritmo de *Baum-Welch*

Não existe um solução ótima para o problema de determinar os parâmetros de um HMM que maximize $P(O|\ \lambda)$. A solução mais adotada baseia-se na criação de um modelo inicial aleatório, sobre o qual é aplicado um método de reestimação iterativo, em que cada novo modelo gera a seqüência de observações com mais alta probabilidade que o anterior. Utilizando o conceito de freqüências de ocorrência, o novo modelo $\lambda' = (T', B', \pi')$ é calculado com o algoritmo de reestimação de *Baum-Welch*, onde, de acordo com Meneses (2002):

$$\pi'_i = \frac{número\_de\_vezes\_em\_S_i\_no\_tempo\_t=1}{número\_total\_de\_ocupações\_no\_tempo\_t=1}$$

$$\alpha'_{ij} = \frac{número\_de\_transições\_de\_S_i\_para\_S_j}{número\_total\_de\_transições\_de\_S_i}$$

$$\beta'_i(k) = \frac{número\_de\_vezes\_que\_se\_observou\_k\_em\_S_i}{número\_total\_de\_vezes\_em\_S_i}.$$

Os parâmetros são calculados sempre a partir do modelo atual na iteração em questão. Baum provou que esse algoritmo melhora a probabilidade de uma seqüência observada, isto é, $P(O \mid \lambda') \geq P(O \mid \lambda)$.

4.2.2 HMM para seqüências biológicas

Aplicar um HMM resume-se a computar as matrizes de probabilidades de transição e emissão para que o modelo reproduza o conjunto de seqüências observadas.

Em Bioinformática, o estudo de regularidades é realizado pelo método do alinhamento de seqüências, ou seja, a comparação do tipo de nucleotídeo que aparece em cada posição num conjunto de seqüências para analisar a similaridade entre elas de forma global ou localmente [Setúbal e Meidanis, 1997]. O modelo de Markov executa essa tarefa com a vantagem de considerar a possibilidade de deleção de uma ou mais bases da seqüência, e não

apenas a hipótese de uma base estar ou não pareada com as demais que ocupam uma mesma posição [Mount, 2000]. A seguir é descrito como esses elementos são modelados no HMM.

4.2.2.1 Estrutura

A estrutura de um HMM pode ser definida como um grafo direcionado. A arquitetura padrão é do tipo esquerda para direita, que não permite o retorno para um estado do qual já partiu uma transição.

Baseado na função de alinhamento, o modelo de Markov para seqüências biológicas é composto de 3 tipos de estados:

- ⇒ principal (**P**), que modela a situação de pareamento de bases, logo, seu conjunto define a seqüência esperada de estados de emissão [Guimarães e Melo, 2003];
- ⇒ inserção (**I**), que modela o caso de não pareamento entre as bases;
- ⇒ e deleção (**D**), que modela a remoção de um símbolo na seqüência, não emitindo símbolos ao serem alcançados.

Além disso, os dois estados especiais *início* e *fim* também não emitem símbolos ao serem atingidos. Quando não emitem símbolos, os estados são chamados de silenciosos.

A Figura 4.1 mostra a arquitetura genérica de HMM usada para modelar um processo de alinhamento de seqüências.

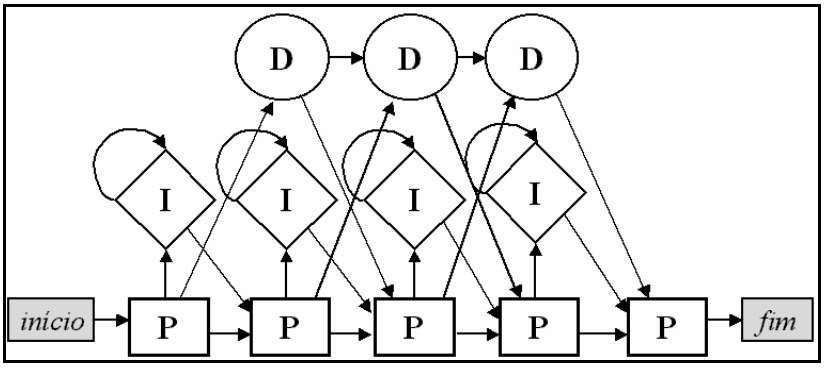


Figura 4.1 – Estrutura do HMM para seqüências biológicas. Além da arquitetura, são apresentados os três tipos de estados: pareamento (P), inserção (I) e deleção (D).

Para seqüências de tamanho n e alfabeto de k elementos, a arquitetura tem aproximadamente $2 \cdot n \cdot k$ parâmetros de emissão e $9 \cdot n$ parâmetros de transição [Guimarães e Melo, 2003].

4.2.3 Reconhecedor de Padrões

Utilizando HMMs, o reconhecimento de uma seqüência pertencer a uma dada classe, baseia-se na existência de modelos probabilísticos para os elementos que compõem a linguagem gerado dessa seqüência. O protocolo padrão de aplicação de um HMM (HMMP), para reconhecer os sinais desta linguagem, segue as seguintes etapas [Meneses, 2002]:

1. Definição das classes de seqüências a serem reconhecidas.
2. Definição de uma topologia para o HMM através da especificação do tipo de modelo, do número de estados e de observações em cada estado.
3. Criação de um modelo para cada classe de acordo com as especificações da etapa anterior.
4. Obtenção de um conjunto razoável de dados para treinar um HMM para cada classe.
5. Treinamento dos modelos.

6. Aplicação dos modelos para reconhecer seqüências.

Utilizar HMM como reconhecedor de uma classe de seqüências requer uma métrica para avaliar o quanto uma dada seqüência se adere ao modelo treinado. Para isso, utilizamos um escore dado pelo logaritmo negativo da verossimilhança do modelo através do algoritmo de *Viterbi*, o qual seleciona o melhor caminho de transição ao longo dos estados possíveis [Clote e Backofen, 2000].

A verossimilhança é a probabilidade de um evento já ocorrido ter gerado uma saída específica, ao contrário da probabilidade que se refere a eventos futuros [Mount, 2000]. Para HMM, o evento passado é o modelo ter atingido um dado estado. É importante ressaltar que, estando o valor dessa probabilidade entre 0 e 1, o seu logaritmo negativo é um escore de energia positiva, em que altos valores de verossimilhança correspondem à baixa energia [Clote e Backofen, 2000]. Assim, quanto mais alto o escore, maior a chance da seqüência não pertencer à classe modelada pelo HMM.

Para definir até que valor de verossimilhança uma seqüência pode ser dita de uma dada classe, é comum considerar outra classe de seqüência e calcular qual o valor de verossimilhança em que todos os exemplos da outra classe ficam acima, enquanto o da classe do HMM fica abaixo.

4.3 Usos em Bioinformática

Um HMM é um modelo estatístico apropriado para muitas tarefas em biologia molecular. Seu uso mais comum é como um *perfil probabilístico* de uma família de proteínas, também chamado de *perfil HMM*, com o qual se podem realizar buscas em banco de dados de proteínas à procura de outros membros da mesma família [Krogh, 1998]. A estrutura de modelo aqui utilizada é a mesma descrita na seção 4.2.2.1.

HMMs também são úteis para tratar de problemas que envolvem uma “estrutura gramatical”, como o caso de encontrar genes em genomas (*gene finding*). Neste caso, vários sinais devem ser reconhecidos e combinados na predição de exons e íntrons, sendo que a predição deve considerar várias regras para tornar a predição gênica confiável [Krogh, 1998]. Para essa aplicação estão disponíveis diferentes ferramentas, tais como o GLIMMER e o GeneMark [Salzberg et al., 1998; Lukashin e Borodovsky, 1998].

4.4 Considerações

Diante da grande aceitação do uso de HMMs para análises de Bioinformática, essa técnica de reconhecimento de padrões foi tomada como alicerce metodológico para reconhecer e predizer promotores procarióticos neste trabalho. A construção e a definição do protocolo de aplicação desta metodologia são discutidas no próximo capítulo.

5 Metodologia

Compreender o funcionamento da RNA-polimerase e sua regulação aparece como objetivo chave para biólogos moleculares, desde de sua descoberta no final da década de 50 por Samuel Weiss [Mooney e Landick, 1999]. Dos elementos estruturais com os quais ela interage, o promotor ganha atenção por ser o maior responsável pelo início de todo o processo de transcrição e da expressão correta dos genes.

O avanço de técnicas de seqüenciamento de genomas gerou um acúmulo de dados nos últimos anos, mostrando que os meios convencionais de análise *in vitro* são restritos tanto pelo custo quanto pela dificuldade em capturar informações subjacentes à regulação gênica.

Neste sentido, ferramentas de Aprendizado de Máquina ganham aplicabilidade ao problema por serem capazes de aprender de forma automatizada a partir de dados disponíveis e levantar hipóteses relevantes sobre mecanismos biológicos ainda ocultos [Baldi e Brunak, 2001; Souto et al., 2003].

Para análise de promotores procarióticos, as abordagens utilizadas, conforme discutido no Capítulo 3, não consideram em seus modelos informações resultantes de estudos experimentais já disponíveis. Na função de reconhecimento dessas regiões, as técnicas se mostram eficientes apenas para genomas cujos promotores são conhecidos, ou seja, somente para a *E. coli*. A maior barreira no uso de AM é a falta de promotores identificados para formar os conjuntos de treinamento de cada organismo, pois os modelos treinados com promotores de *E. coli* não capturam propriedades genéricas o bastante para reconhecer qualquer promotor procariótico.

Um protocolo eficaz e robusto para reconhecer e predizer promotores em procariotos é uma questão latente entre as pesquisas de Bioinformática, devido, principalmente, à alta variação nucleotídica dessas regiões e dos diferentes fatores σ que se associam à RNAp para o início da transcrição.

Este trabalho se propõe a enfrentar o desafio de criar um procedimento *in silico* para análise de promotores procarióticos, que considere características funcionais e estruturais dessas regiões, auxiliando tanto para a tarefa de reconhecimento como para a de predição.

A revisão bibliográfica sobre as abordagens *in silico* para análise de regiões promotoras mostra o HMM como uma ferramenta adequada para a tarefa de reconhecimento de seqüências. Assim, a partir do protocolo padrão de aplicação de HMM para reconhecer cadeias de caracteres (veja seção 4.2.3) foram estabelecidas modificações e extensões, no intuito de aumentar o desempenho do modelo no reconhecimento de promotores, e, posteriormente, tornar possível a utilização desse mesmo protocolo para predição de promotores em outros organismos, mesmo que adaptações sejam necessárias.

Além disso, procurou-se analisar o HMM como uma "caixa aberta", investigando as propriedades capturadas pelo modelo treinado. Essas características confrontadas com a medição da performance do reconhecedor de promotores possibilitaram a proposta de um novo protocolo, resultado final desta dissertação.

É importante ressaltar que o protocolo aqui proposto ganhou forma à medida que resultados preliminares eram obtidos, sempre considerando o acréscimo no desempenho na tarefa de reconhecimento assim como a viabilidade do mesmo ser usado na de predição.

Os protocolos para o uso de HMMs na análise de regiões promotoras procarióticas são tema deste capítulo.

5.1 Recursos Disponíveis

5.1.1 Dados

Antes de aplicar uma metodologia de Aprendizado de Máquina, é essencial realizar o levantamento dos dados disponíveis para que a etapa de treinamento tenha um número adequado de exemplos para estimar os parâmetros dos modelos.

As regiões promotoras e os dados relacionados ao seu funcionamento foram extraídos de uma coleta de informações em artigos científicos de biologia molecular e de bancos de dados públicos, contemplando a estrutura da RNAp e do promotor, e os mecanismos envolvidos em sua interação.

De posse das informações levantadas na revisão bibliográfica, ficou clara a necessidade de uma base de dados com mais instâncias para estimar os parâmetros do HMM com maior qualidade e, conseqüentemente, obter maiores taxas de acerto no reconhecimento. Além disso, o número baixo de exemplos para o treinamento é a causa apontada, quase que unanimente, para o grande número de seqüências não-promotoras com alto ajuste ao modelo promotor, tanto utilizando HMMs como outros métodos computacionais e estatísticos.

Diante desta necessidade, os bancos de dados utilizados foram:

⇒ **EcoCyc:** banco de dados com diferentes categorias de informações sobre a *E. coli*, tais como mapas metabólicos, organização gênica, regulação transcricional, entre outras [Karp et al., 2004] (http://www.ecocyc.org).

⇒ **GenBank:** a maior base de seqüências genéticas e genes anotados, de onde serão extraídos os genomas completos dos organismos estudados (http://www.ncbi.nlm.nih.gov/Genbank/index.html).

⇒ **PEC – Profiling of *E. coli* Chromosome:** mais completa lista de genes essenciais e genes dispensáveis em *E. coli*, baseada não apenas em evidências experimentais,

mas também em uma revisão sistemática da literatura experimental [Gerdes et al., 2003] (http://www.shigen.nig.ac.jp).

⇒ **RegulonDB:** base de informações confiáveis sobre a rede regulatória de *Escherichia coli* com conhecimento experimental, que inclui dados da organização dos operons e de sua decomposição em unidades transcricionais, dos promotores e sua respectiva classe de fator σ associado, entre outros. (http://www.cifn.unam.mx/Computational_Genomics/regulondb/) [Salgado et al., 2004]. As regiões promotoras propriamente ditas foram extraídas deste banco, sendo consideradas as versões 3.1 [Huerta et al., 2003] e 4.0 [Salgado et al., 2004], disponível, respectivamente, em dezembro de 2003 e em julho de 2004.

⇒ **TIGR:** base de dados do *The Institute for Genomic Research* reunindo análises estruturais, funcionais e comparativas de genomas e de genes para diferentes organismos (http://www.tigr.org).

A *Escherichia coli* foi o organismo escolhido para verificação da correção dos resultados da metodologia, pois ele tem o maior número de dados experimentais com alta qualidade disponíveis, além de ser considerado o modelo biológico padrão entre os procariotos.

Entre as poucas bases públicas de informações de promotores de *E. coli*, como a compilação de Lisser e Margalit (1993) com 300 exemplos e o Promec [Hershberg et al., 2001] com 427, o RegulonDB [Huerta et al., 1998] foi selecionado como fonte principal de dados em virtude de possuir o maior e mais detalhado conjunto de regiões promotoras de *Escherichia coli.*

Além dessas regiões, o banco reúne dados referentes aos mecanismos que regulam o processo de transcrição e aos aspectos estruturais relevantes para a expressão gênica neste

organismo, o que permite também a análise de correlação entre a taxa de reconhecimento e outras propriedades regulatórias pertinentes.

Diante da utilização do protocolo baseado em HMMs para predição, outros organismos também foram investigados, sendo eles: *Bacillus subtilis* 168, o gênero *Helicobacter*, sendo analisadas as espécies *Helicobacter hepaticus* ATCC51449 (*H. hepaticus*) e *Helicobacter pylori* (*H. pylori*) nas linhagens J99 e 26695. O *Bacillus subtilis* é uma bactéria encontrada no solo, na água e em associação com plantas, sendo uma fonte importante de proteínas sintéticas . O *H. pylori* é uma bactéria patogênica que habita o epitélio gástrico do homem, que junto com outros fatores pode causar gastrite, a qual pode evoluir para outras doenças, como o câncer [Vanet et al., 2000]. O *H. hepaticus*, por sua vez, causa hepatite crônica e câncer de fígado em camundongo [Suerbaum et al., 2003]. É importante ressaltar que as duas linhagens são mais próximas evolutivamente do que as duas espécies.

Tabela 5.1 – Informações genômicas a respeito dos organismos investigados. A sigla é um código estabelecido para facilitar a referência ao organismo.

Organismo	Sigla	Tamanho do Genoma (pb)	Conteúdo A+T (entre 0 e 1)
Escherichia coli K12-MG1655	eco	4.639.221	0.492
Bacillus subtilis 168	bsu	4.214.810	0.565
Helicobacter pylori 26695	hpy26695	1.667.867	0.61
Helicobacter pylori J99	hpyJ99	1.643.831	0.61
Helicobacter hepaticus ATCC 51449	hhe	1.799.146	0.641

Apesar da definição de um conjunto limitado de procariotos para análise, o protocolo de aplicação dos HMMs para análise de promotores almeja ser robusto para investigação em qualquer organismo procariótico.

5.1.2 Ferramentas Computacionais

Para construir o reconhecedor de promotores e estudar suas características foi utilizada a ferramenta HMMpro, um simulador de HMMs específico para análise de seqüências biológicas [Net-ID, 2003], que permite entre outras funções: treinar o modelo, computar as composições das seqüências, visualizar as probabilidades de emissão e transição e escolher um método de avaliação.

A linguagem Perl foi utilizada para programar *scripts* de automatização de tarefas, tais como: a preparação dos dados para as etapas de treinamento e teste, a integração com informações de outras bases que não o RegulonDB, e a computação de resultados estatísticos.

5.2 Criação dos HMMs

5.2.1 Preparação dos Dados

A medição do desempenho de um HMM para reconhecimento necessita que sejam consideradas seqüências de pelo menos duas classes, uma que será usada para treinar o modelo reconhecedor e outra para avaliar sua especificidade. As regiões pertencentes à classe que o HMM reconhece são chamadas de verdadeiros positivos (*VP*), enquanto as não-pertencentes são chamadas de verdadeiros negativos (*VN*). Os *VN* são utilizados para se estimar valor máximo de verossimilhança para uma seqüência ser dita promotora.

Neste trabalho, os verdadeiros positivos são as regiões promotoras extraídas da base *promoter_table.dat* do RegulonDB, constituídas de 81 nucleotídeos, sendo 60 deles correspondente à região *upstream* do gene (região promotora). Os verdadeiros negativos, por sua vez, são fragmentos de regiões gênicas também de 81 nucleotídeos mapeados dos arquivos *.fna* do organismo em questão disponível no GenBank (veja Tab. 5.2), que contém todas as seqüências nucleotídicas dos genes no formato FASTA.

O uso de apenas duas classes de informações, região gênica e promotora, é importante por eximir a incorporação de outras informações como dados de expressão, unidades de transcrição caracterizadas e bem definidas, classe funcional dos genes, entre outras usadas em outros trabalhos.

Tabela 5.2 – Arquivos *.fna* disponível no GenBank para cada organismo investigado.

Organismo	**Arquivo *.fna***
Escherichia coli K12-MG1655	U00096.fna
Bacillus subtilis 168	AL009126.fna
Helicobacter pylori 26695	AE000511.fna
Helicobacter pylori J99	AE001439.fna
Helicobacter hepaticus ATCC 51449	AE017125.fna

Para a extração das seqüências promotoras e gênicas, filtragem dos dados (devido à inconsistências das bases de dados: existência de registros sem a região propriamente dita identificada e/ou a existência de fator σ associado), assim como para a geração de seqüências aleatórias necessárias para algumas análises, foram implementados programas em Perl. Esses têm como saída os dados já formatados para a leitura da ferramenta Hmmpro.

5.2.2 Topologia e treinamento do HMM

A arquitetura de todos os HMMs criados, independente da classe, é do tipo linear com transições possíveis para todos os tipos de estados, ou seja, de pareamento (*match*), de não-pareamento (*mismatch*) e de deleção (*gap*).

O HMM possui 81 estados principais, mesmo número de bases das regiões promotoras do RegulonDB. O alfabeto para a emissão é constituído dos 4 nucleotídeos, sendo a probabilidade inicial de emissão de cada um igual a 0,25.

Definidos a estrutura e os parâmetros iniciais, o HMM é treinado por 30 épocas no modo *on-line*, período considerado adequado para este tipo de problema [Baldi e Brunak, 2001]. Após o treinamento, tem-se o melhor modelo que descreve as seqüências do conjunto de treinamento submetidas ao HMM nesta etapa.

5.3 Protocolos para avaliar o desempenho dos HMMs

A revisão bibliográfica do Capítulo 3 e a descrição da teoria de HMMs no Capítulo 4 mostram a ausência da utilização de uma técnica de validação estatística do desempenho do HMM quando aplicados ao reconhecimento de seqüências.

Na intenção de utilizar o HMM de forma criteriosa, a metodologia de *10-fold cross validation* foi adicionada ao protocolo padrão para criação dos modelos, principalmente em razão do grande número de parâmetros a serem computados em relação ao número de instâncias disponíveis. Ela baseia-se na divisão do conjunto de treinamento em 10 partes, sendo 9 deles utilizadas para treino e 1 delas para teste do modelo [Wu e McLarty, 2000]. O conjunto de teste é constituído desses 10% do de treinamento mais o mesmo número de seqüências da classe gênica. A realização de 10 simulações de modelos com essa metodologia tem como vantagem a possibilidade de analisar o erro médio do HMM como reconhecedor de seqüências.

A seguir, são apresentadas as demais extensões ao protocolo de uso da metodologia de HMMs para análise de promotores procarióticos.

5.3.1 Com Estimação de Limiar de Decisão - ELD

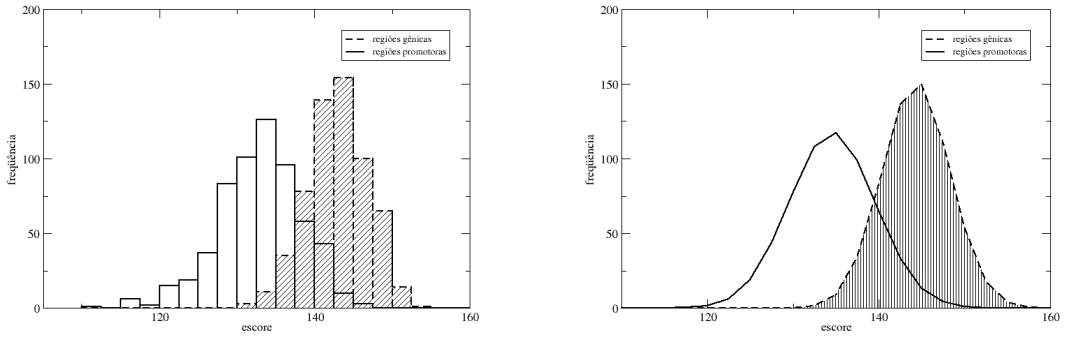
Definir o valor máximo para uma seqüência ser considerada promotora, baseado na comparação da verossimilhança de outra classe de seqüência, torna esse limiar fortemente

dependente da classe considerada. Ademais, isso gera a expectativa de uma alta taxa de erro quando o modelo for empregado para predizer essas regiões em diferentes organismos.

Assim, considerou-se que a aplicação de HMMs como uma ferramenta de classificação exige um método mais rigoroso para determinar o valor de escore limite ou crítico (S_c), tal que as seqüências com pontuação menor que ele sejam classificadas como promotoras.

Para especificar o S_c, utilizamos dois conjuntos de seqüências: o de promotores submetido para treinar o HMM e o de fragmentos gênicos de mesmo tamanho. Esses dois conjuntos são avaliados pelo HMM através do algoritmo de *Viterbi* (ver seção 4.2.1.2).

Como resultado, temos os escores para o conjunto de *VP* ($S_{promotores}$) e para o conjunto de *VN* (S_{genes}). Observa-se que essas duas variáveis podem ser modeladas como variáveis aleatórias normalmente distribuídas. Assim, para cada conjunto são computados os histogramas. Ajustando esses dados a Gaussianas, conforme mostra a Figura 5.2, têm-se duas distribuições $D_{promotores}$ e D_{genes}, respectivamente para verdadeiros positivos e verdadeiros negativos.



(a) (b)

Figura 5.1 – Distribuições de probabilidade dos escores do algoritmo de *Viterbi* para regiões promotoras e gênicas. Em (a) são apresentados os histogramas. Em (b), as curvas de ajuste a Gaussianas.

Para definir S_c, utilizou-se a Estimação do Limiar de Decisão com base na Regra de Decisão de Bayes, na qual, usando a melhor estimativa das probabilidades de duas classes, calcula-se o valor esperado de decisão para cada uma delas, escolhendo a alternativa com máximo valor.

Considerando que as distribuições *D* representam fielmente os dados, podem-se determinar as funções $D_{promotores}(S)$ e $D_{genes}(S)$, as quais denotam a fração de verdadeiros negativos e de verdadeiros positivos em função do escore limiar *S*. S_c é o valor esperado que maximiza a soma dessas duas funções.

Com base nesta idéia, escolheu-se o critério de exatidão (*A*) para escolher o S_c. Assim, o S_c é determinado pelo valor que maximiza a exatidão do modelo, a partir do número de verdadeiros positivos, verdadeiros negativos, falsos positivos (genes reconhecidos como promotores) e falsos negativos (promotores, como genes), *VP*, *VN*, *FP*, *FN*, respectivamente. O cálculo de exatidão é dado por:

$$A = \frac{VN + VP}{VN + VP + FN + FP}.$$

O cálculo de exatidão também é utilizado como medida de desempenho sobre os conjuntos de teste. Assim, o protocolo computa dois valores de exatidão: a esperada (A_e), resultante da soma das distribuições, e a observada (A_o), resultante do cálculo sobre o conjunto de teste.

5.3.2 Com Estimação de Limiar de Decisão e Análise de Discriminação - ELDAD

Quando um HMM é treinado com as regiões promotoras, o modelo captura não apenas padrões específicos dessas seqüências, mas também outras características intrínsecas à constituição e organização do genoma do organismo. Um caso conhecido é a influência do conteúdo A+T do genoma.

Assim, além do uso de ELD, estendeu-se o protocolo para o uso de um par de HMMs, uma para regiões promotoras e o outro para gênicas, com o objetivo de eliminar o ruído causado por outros padrões genômicos no modelo promotor. Essa alteração está fundamentada na Teoria de Estimação de Limite de Decisão de Verossimilhança para Análise de Discriminação [Arslan e Hansen, 1998].

Para isso, criam-se dois modelos:

$\Rightarrow$ um para representar seqüências promotoras de *E. coli*;

$\Rightarrow$ um *modelo nulo* para representar demais regiões.

Visto que os genomas em procariotos são bastante compactos, com baixa porcentagem de regiões não-codificantes (*junk-DNA*; DNA-lixo), pode-se reduzir sua constituição a três classes de seqüências: gênicas, promotoras e terminadoras. Como as regiões regulatórias, promotores e terminadores, representam um baixo percentual do número total de bases do genoma, optou-se por usar as seqüências gênicas como verdadeiros negativos, assim como quando usado apenas o ELD. Sendo o genoma procariótico caracteristicamente compacto, ou seja, quase que totalmente composto por regiões codificantes, supõe-se que a RNAp requerer um ajuste fino para reconhecer no máximo 100 pb onde se encontra o promotor.

Para computar o escore (S_p) de uma seqüência *x* ser promotora, consideram-se 2 modelos: o de promotor (*hmm_p*) e o de genes (*hmm_g*), sendo dado por:

$$S_p = \log \frac{P(x \mid hmm\_p)}{P(x \mid hmm\_g)}$$

ou

$$S_p = P(x \mid hmm\_p) - P(x \mid hmm\_g)\,.$$

Esse cálculo tem a vantagem de não requerer que sejam conhecidas previamente as probabilidades das duas classes de seqüências mencionadas, pois, de acordo com a regra de *Bayes*:

$$P(hmm\_p \mid x) = \frac{P(x \mid hmm\_p)P(hmm\_p)}{P(x)}$$

e

$$P(hmm\_g \mid x) = \frac{P(x \mid hmm\_g)P(hmm\_g)}{P(x)}.$$

Além disso, a razão da adoção deste critério em relação à modelagem é reduzir o ruído das características intrínsecas do genoma no modelo promotor, principalmente o conteúdo A+T. Com o S_p, verifica-se o quanto mais provável uma seqüência é promotora do que gênica através do cálculo da subtração entre as verossimilhanças. O uso deste parâmetro na utilização de HMMs para reconhecer promotores reduz o número de *FP*, quando analisados genomas com conteúdo A+T diferente de 0.5, como será discutido no Capítulo 6.

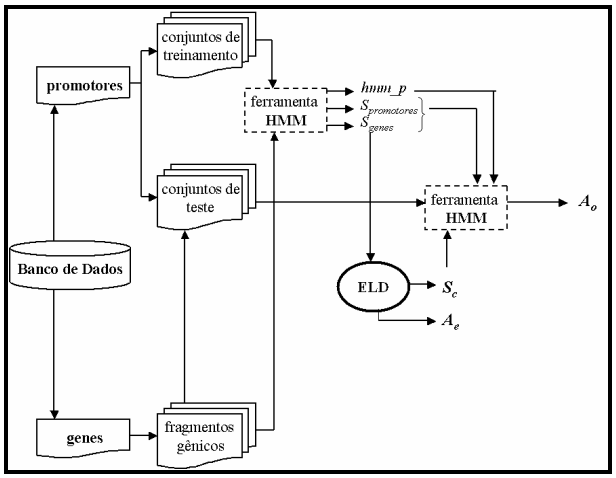
5.4 Aplicações dos Protocolos: Reconhecimento e Predição

Apesar da descrição anterior dos protocolos, é necessário especificar suas adaptações para sua utilização na tarefa de reconhecimento, para a qual são conhecidos um número relevante de regiões promotoras do organismo em questão, e na de predição, quando a única informação de maior confiança são as localizações dos genes.

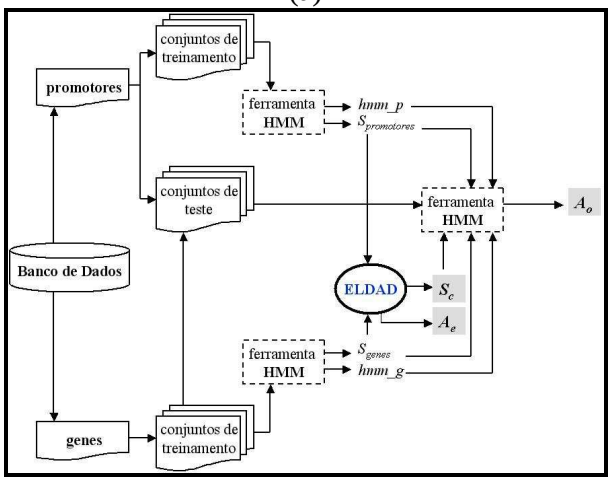
Para o reconhecimento é possível utilizar os três métodos: HMM com protocolo padrão, HMM com ELD, e HMM com ELDAD. A Figura 5.2 representa esquematiamente as diferenças entre dois protocolos.

Para a predição, como não existe um número adequado de promotores procarióticos para criar HMMs que os reconheçam em cada organismo, utiliza-se como modelo promotor o resultante da melhor simulação do *10-fold cross validation* de *E. coli*. Assim, o uso do protocolo ELD não pode ser aplicado, uma vez que não existem verdadeiros positivos para

criar a $D_{promotores}$, sendo este substituído pelo cálculo da média e desvio-padrão de D_{genes}, que servirá de critérios para avaliar a predição.



(a)



(b)

Figura 5.2 – Esquemas da aplicação adaptada do uso de HMM com os protocolos propostos para cada análise de promotores. Em (a) para reconhecimento usando ELD e em (b) para reconhecimento com ELDAD.

Esses dois critérios podem, então, ser associados à análise de discriminação, sendo consideradas duas abordagens:

⇒ Larga escala: baseada apenas na distribuição de *VN* para o genoma do organismo para o qual será executada a predição, sendo os escores das seqüências computados seguindo a equação de S_p.

⇒ Curta escala: em que são analisados segmentos específicos do genoma, a fim de estudar como diferentes sub-partes de um segmento são ajustáveis ao HMM promotor, obtendo o comportamento de S_p. O foco é a investigação de operons.

A Figura 5.3 representa esquematicamente a adaptação do protocolo para a tarefa de predição.

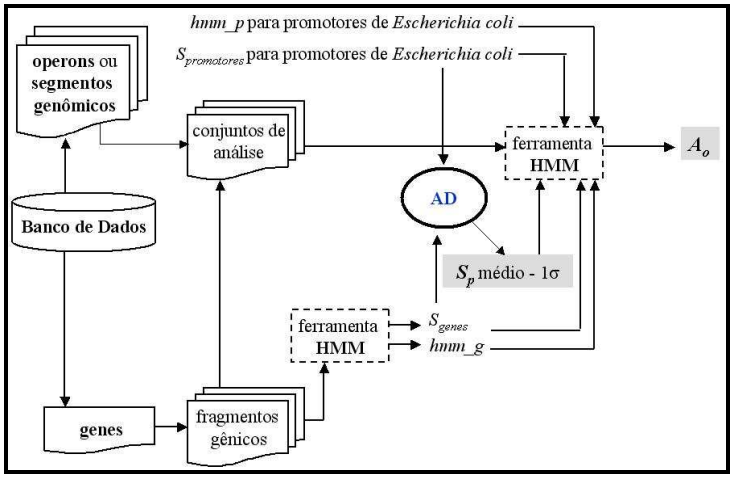


Figura 5.3 – Esquema da aplicação adaptada do uso de HMM com as extensões propostas para a tarefa de predição.

5.5 Considerações

Não há dúvida de que HMM é um modelo teoricamente adequado para o problema de reconhecimento e predição de promotores em procariotos. A metodologia descrita neste capítulo mostra o uso da abordagem de modelos de Markov, com quatro extensões:

1. uso de base de dados com mais de 600 instâncias;
2. cálculo de limiar de decisão;
3. análise de discriminação;
4. especificação de ajustes para a aplicação dos protocolos de reconhecimento na predição de promotores.

Os resultados da aplicação dos protocolos em diferentes procariotos são discutidos no próximo capítulo.

6 Resultados

A metodologia de HMMs utilizando os protocolos com Estimação de Limiar de Decisão (ELD) e com ELD e Análise de Discriminação (ELDAD) foram aplicados aos organismos especificados na seção 5.1.1.

A aplicação de cada protocolo depende do conjunto de informações disponíveis para cada um desses procariotos. Assim, em função das limitações e ajustes necessários no momento de obter os resultados para cada organismo, os resultados foram organizados de acordo com as espécies analisadas.

6.1 *Escherichia coli*

O grande número de dados disponíveis a respeito da regulação gênica em *Escherichia coli* permitiu uma análise mais abrangente, sendo estruturada em diferentes abordagens, dentre as quais:

- ⇒ Avaliação do desempenho de reconhecimento de promotores utilizando os dois protocolos propostos.
- ⇒ Comparação dos padrões capturados pelo modelo HMM de promotor com informações de biologia molecular relacionadas à regulação e outras propriedades moleculares.
- ⇒ Investigação do uso dos protocolos para predição aplicado em operons (predição em curta escala).

Esta seção descreve e discute esses resultados.

6.1.1 Reconhecimento

O protocolo ELD foi aplicado nas duas versões do RegulonDB, a fim de investigar a possível influência do número de seqüências disponíveis para a etapa de treinamento na capacidade do modelo reconhecer os promotores.

Na versão 3.1, a base para o *10-fold cross validation* continha 600 promotores. Enquanto, para a versão 4.0 do RegulonDB, a base possuía 929 promotores. O desempenho dos modelos, confrontando o tamanho dos conjuntos de treinamento do modelo com sua respectiva taxa de acerto, é explicitado na tabela abaixo.

Tabela 6.1 – Medições de escore limite e exatidão estimada e observada para o reconhecimento de promotores de *E. coli* nas 2 versões do RegulonDB.

Base	Medida	Média	Desvio-padrão
RegulonDB versão 3.1	S_c	139,177	0,957
	A_e	0,843	0,03
	A_o	0,805	0,047
RegulonDB versão 4.0	S_c	136,587	0,615
	A_e	0,851	0,011
	A_o	0,817	0,07

Observe que o S_c médio, sobre as 10 simulações, para a versão 3.1 foi de 136,587, muito próximo de 137,6, valor obtido por Pedersen et al. (1996), com um conjunto de dados menor para treinamento e sem considerar uma validação estatística como o *10-fold cross validation.*

Considerando os valores de exatidão observados, conclui-se que o aumento de 600 para 929 seqüências promotoras no conjunto de dados diminui o erro de reconhecimento em

1,05%, acréscimo de performance irrelevante visto o custo para a identificação de mais 329 promotores.

O uso da metodologia de HMM com ELD funcionou satisfatoriamente para o reconhecimento de regiões promotoras de *E. coli* (veja Tab. 6.1), apresentando as seguintes vantagens no uso dessa técnica:

- ⇒ Desnecessidade do alinhamento prévio das seqüências de treinamento para gerar o modelo.
- ⇒ Possibilidade de utilizar o modelo como reconhecedor em outros genomas procarióticos, pois a *E. coli* ainda é o único organismo com um número relevante de promotores identificados.

Na tentativa de utilizar o modelo HMM da *E. coli* para predizer regiões promotoras em outros procariotos, aplicou-se a mesma metodologia e protocolo ao organismo *Bacillus subtilis* (*B. subtilis*). Nesta situação, não se obteve sucesso, pois os fragmentos de genes desse organismo se ajustam muito bem ao modelo, aumentando a taxa de falsos positivos. A possível causa desta falha é a distribuição do conteúdo dos nucleotídeos nos genomas, conteúdo A+T e G+C, a qual já foi discutida em estudos de identificação de sinais em DNA [Pevzner, 2000].

Assim, testaram-se conjuntos de seqüências aleatórias com diferentes percentuais de conteúdo A+T no modelo criado com a versão 4.0 do RegulonDB (10 conjuntos para cada valor de percentual), para verificar sua influência no reconhecimento. Comparando esses resultados com os de genomas com percentuais semelhantes aos testados (ver Figura 6.1), fica claro que a tendência nucleotídica é uma barreira para o uso do HMM de *E. coli* apenas com o protocolo ELD para predição genérica de promotores procarióticos. Isso porque, quando mais alto for o valor do conteúdo A+T de organismo comparado ao de *E. coli* (em torno de 0,5), maior a probabilidade deste HMM reconher a seqüência como promotora.

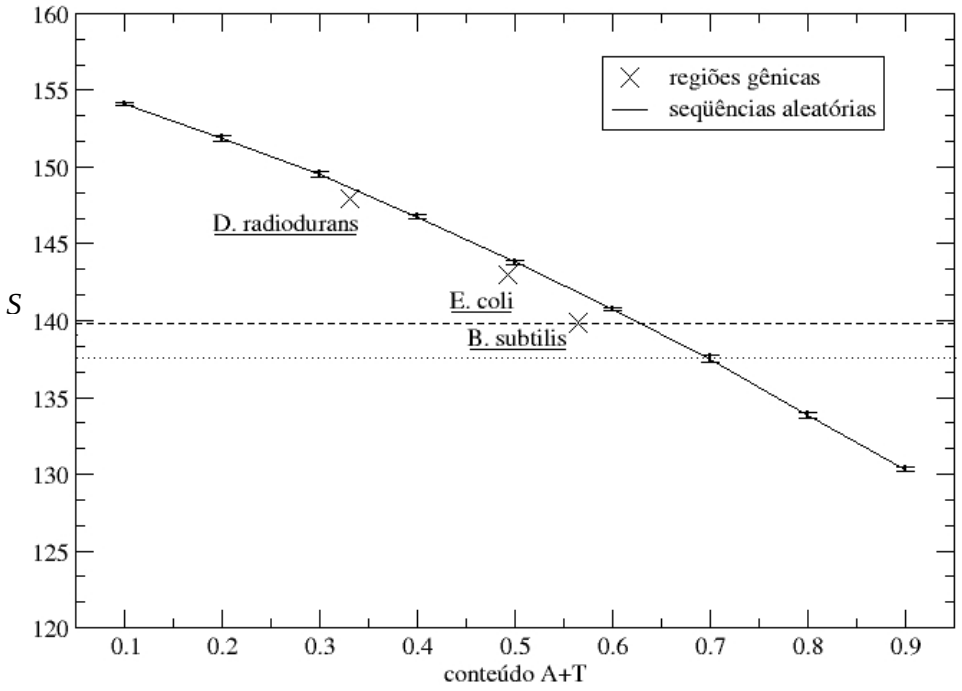


Figura 6.1 - Gráfico do escore (*S*) médio para seqüências geradas aleatoriamente com diferentes percentuais de conteúdo A+T. Observe que organismos com percentuais de A+T próximos aos analisados possuem resultados semelhantes. A linha tracejada representa o S_c para o modelo criado com a versão 4.0 do RegulonDB, e a linha pontilhada é o S_c do trabalho de Pedersen et al. (1996). O organismo *D. radiodurans* foi adicionado à análise por conter um baixo conteúdo A+T.

Para o RegulonDB versão 4.0, além do protocolo ELD, também foi investigado o desempenho do ELDAD para reconhecer os promotores. A comparação dos resultados dos dois protocolos é apresentada na Tab. 6.2.

Tabela 6.2 – Medições de escore limite e exatidão para o reconhecimento de promotores de *E. coli*, utilizando os protocolos ELD e ELDAD no RegulonDB versão 4.0.

Protocolo	Medida	Média	Desvio-padrão
ELD	S_c	136,587	0,615
	A_e	0,851	0,011
	A_o	0,817	0,07
ELDAD	S_c	1,343	0,604
	A_e	0,921	0,005
	A_o	0,919	0,03

Com o ELDAD, a exatidão, tanto observada quanto a esperada, ultrapassa 0,9. Além disso, comparando-se os mesmos resultados com os obtidos apenas com ELD, adquiriu-se um aumento de performance em 12,48% e uma redução na taxa de erro em 44,26%.

6.1.2 HMMs e os padrões conservados

Sabendo que as seqüências promotoras possuem regiões conservadas em torno das posições -35 e -10, analisou-se a probabilidade de emissão dos nucleotídeos em cada posição do HMM de promotor em busca desses padrões. Selecionando o nucleotídeo que apresentou maior probabilidade de emissão em cada estado principal, obteve-se o gráfico da Figura 6.2. Nesta análise, são encontrados nucleotídeos altamente conservados além dos hexâmetos -10 e -35, sendo que alguns estão bem distantes desses dois padrões. É interessante observar, por exemplo, os dois sítios (-34 e -33) *T*s altamente conservados e que não possuem interação física com a RNAp quando estudado o promotor de *T. aquaticcus* [Naryshkin et al., 2000].

Assumindo que a estrutura da RNAp de *T. aquaticus* é semelhante à de *E. coli*, as bases mais prováveis do HMM foram comparadas com os sítios de interação do promotor

lac(ICAP)UV5, verificando-se que as timinas conservadas não interagem com a RNAp. Isso levanta a hipótese da existência de bases conservadas com outras atribuições para o processo de transcrição não envolvendo ligação entre a RNAp e o DNA.

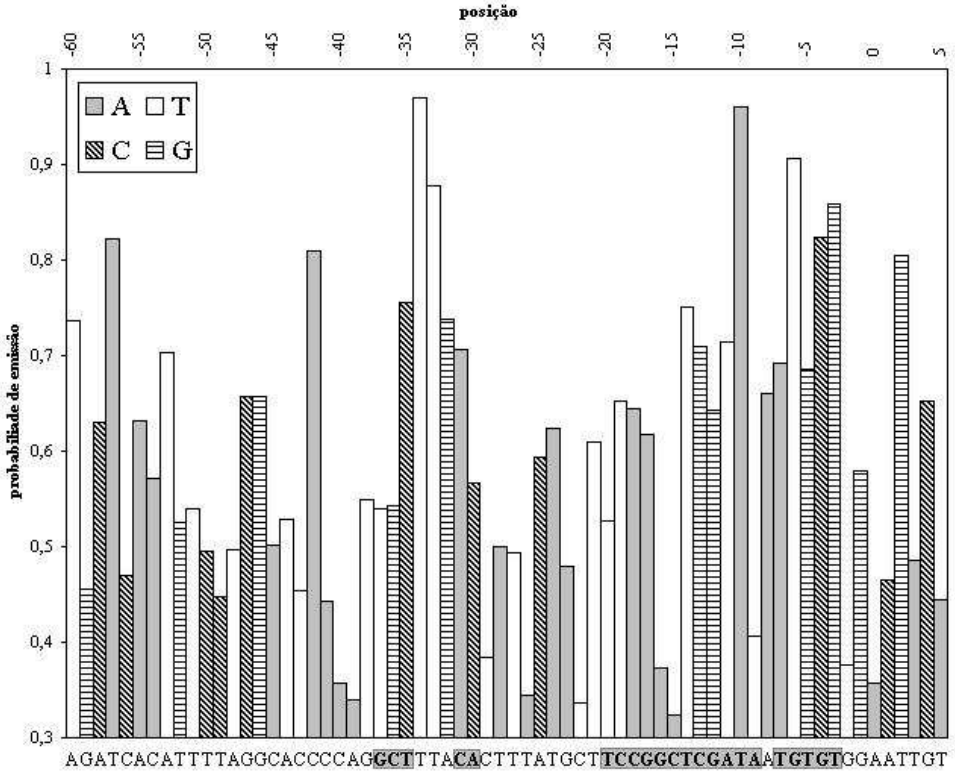


Figura 6.2 – Gráfico descrevendo a base de maior probabilidade para cada estado de pareamento do modelo HMM. Abaixo está a seqüência do promotor *lac(ICAP)UV5* do *T. aquaticus* analisado experimentalmente por Naryshkin et al. (2000). As caixas em cinza representam sítios de interação da RNAp ao promotor. A flutuação média foi calculada, obtendo-se 0,27. Esse valor elimina a hipótese das probabilidades mais baixas resultarem meramente de ruído.

Focando nas regiões de interesse, verifica-se que de -35 a -31 há grande chance de ocorrência do padrão *CTTGA*, e de -14 a -10, do padrão *TGGTA*; os quais comparados com o promotor ideal $TTGACAN_{17}TATAAT$ mostram similaridade relevante.

Ainda para investigar os padrões capturados pelo HMM, foram criados logos a partir de 30 seqüências com alta propabilidade de serem geradas pelo modelo. Essas foram obtidas

por uma função disponível no HMMpro, a partir do modelo treinado de *E. coli*. Essa análise foi realizada com a ferramenta WebLogo [Crooks et al., 2004], resultando na Figura 6.3.

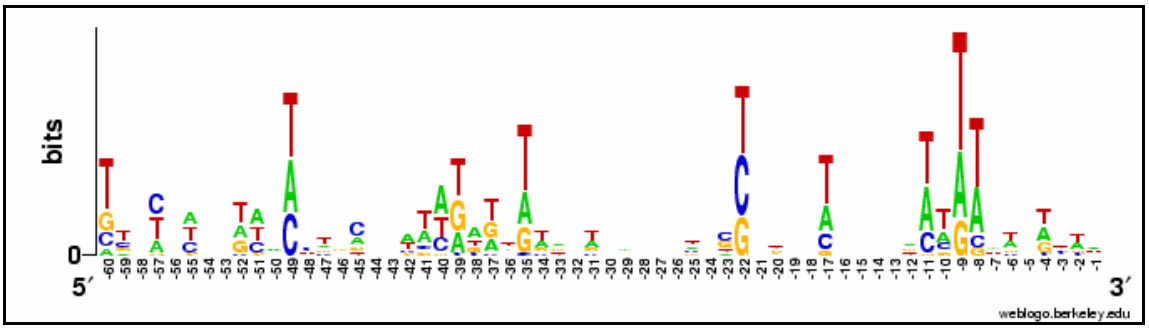


Figura 6.3 – Logo de seqüência para promotores gerados com o modelo promotor de *E. coli*. Nas 60 posições *upstream* são encontrados os padrões nas regiões dos hexâmeros -35 e -10, sendo esse último mais conservado. Observa-se, novamente, a conservação do *T* na posição -35, como no gráfico da Figura 6.2.

Assim, analisando as características de natureza estatística capturadas pelo modelo, conclui-se que o HMM permite a:

- ⇒ representação das regularidades do promotor de forma probabilística, considerando os diferentes graus de conservação das bases.
- ⇒ identificação de outras regiões conservadas, além da -35 e -10;
- ⇒ inferência de outros mecanismos regulatórios que envolvam interação de moléculas com o DNA.

É importante notar que o HMM não exige que os elementos de uma seqüência sejam idênticos ao do seu padrão para reconhecê-la como promotora; mas sim que a probabilidade de ocorrência de uma dada seqüência seja alta com base no limiar S_c.

6.1.3 Correlação com propriedades funcionais

Para verificar outras características que pudessem estar envolvidas na capacidade de reconhecimento de um promotor a partir dos padrões capturados pelo HMM, mediu-se a

relação de *S* do modelo HMM para promotor com algumas propriedades funcionais. A primeira foi a essencialidade gênica, característica que indica a não sobrevivência do organismo caso um dado gene seja nocauteado. Se um dos genes do operon regulado pelo promotor é essencial conforme o banco de dados PEC [Gerdes et al., 2003], o promotor foi considerado ser de gene essencial. A Tab. 6.3 apresenta os resultados.

Tabela 6.3 – Análise de escore versus Essencialidade.

Classe	Número de Promotores	*S* médio	Desvio-padrão
Essencial	95	131,886	4,168
Não-Essencial	714	131,109	4,334
Desconhecida	48	131,901	5,206
Não encontrada	72	128,945	16,285

Os resultados indicam não existir uma correlação entre o escore do algoritmo de *Viterbi* no HMM para promotor com o fato deste regular um gene essencial ou não. O único valor com maior desvio-padrão é o da classe "*Não encontrada"*, que agrupa promotores existentes no RegulonDB para os quais não se encontrou registro de gene no PEC.

Utilizando outras informações disponíveis no RegulonDB e EcoCyc, duas outras características foram analisadas: o número de configurações de interações regulatórias (IR) e e a classe de fator σ associado ao promotor. O número de IRs se refere à existência de diferentes sítios possíveis de interação da RNAp com o promotor, além da possibilidade de um mesmo promotor interagir com outras proteínas regulatórias em momentos distintos. Os resultados para essas propriedades são descritos, respectivamente, nas Tab. 6.4 e 6.5.

Tabela 6.4 – Análise de escore versus Número de Interações Regulatórias.

Número de configurações de IR	Número de Promotores	*S* médio	Desvio-padrão
0	436	132,36	4,626
1	205	130,624	4,359
2	104	129,963	3,243
3	62	129,95	4,079
4	50	129,392	4,032

Também não foi encontrada correlação entre o número de configurações de interação regulatória com *S*. O que chama a atenção, pois era esperado que promotores com baixo número de IR possuíssem maior aderência ao modelo, visto que, quanto menos configurações, menor a complexidade na regulação e, conseqüentemente, mais focada nos hexâmeros -10 e -35 está a interação da RNAp.

Tabela 6.5 – Análise de escore versus Classe do fator σ. Fatores associados a poucos promotores foram desconsiderados.

Classe fator σ	Número de Promotores	*S* médio	Desvio-padrão
70	647	130,718	4,244
70 e 38	27	130,852	3,504
38	31	131,723	4,322
24	32	137,071	3,868
Não identificado	123	130,565	4,483

Os resultados para o fator σ também surpreenderam, pois sendo o σ^{70} um fator constitutivo, ou seja, que reconhece promotores de genes que precisam ser transcritos continuamente, era previsto que seus promotores teriam alta aderência ao modelo HMM. Isso em função da possível necessidade do reconhecimento ocorrer de forma otimizada, para que a transcrição fosse mais rápida.

6.1.4 Operons

A organização dos genes em operons é considerada comum, mesmo havendo pouca evidência experimental para sua estrutura, conservação e evolução em diferentes espécies. Apenas para *Escherichia coli* existe um número relevante de dados disponíveis sobre seus operons e unidades de transcrição [Daruvar et al., 2002].

De posse destes dados retirados da base do RegulonDB e do EcoCyc, aplicou-se o protocolo ELDAD para análise de predição em curta escala para alguns operons. A região de cada operon foi mapeada a partir do primeiro nucleotídeo do primeiro gene até o último nucleotídeo que constitui o último gene. Para inferir a região promotora da unidade de transcrição formada por todos os genes do operon, foram adicionados ao mapeamento os 100 pb *upstream* à posição do primeiro nucleotídeo do primeiro gene localizado no operon.

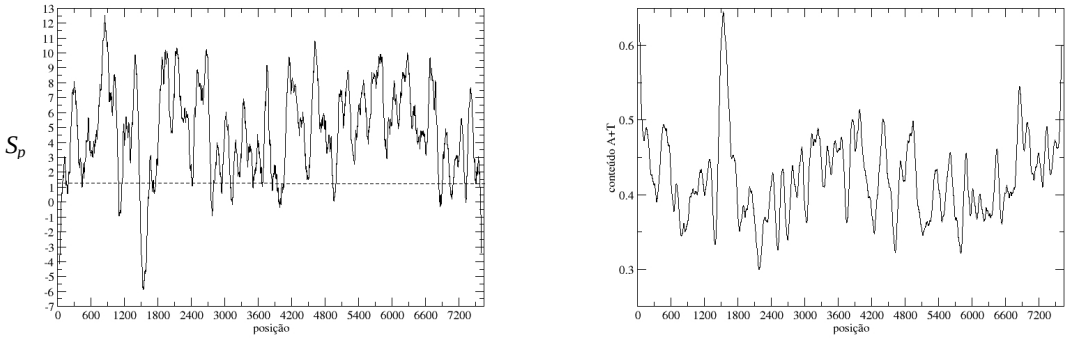
Como os HMMs treinados possuem 81 estados principais, por essa região gênica foi passada um janelamento de mesmo tamanho, de forma que todas as suas bases fossem a primeira de uma seqüência de tamanho 81 para ser submetida ao modelo.

Para analisar a influência do conteúdo A+T nestes resultados, foi computado o gráfico do comportamento desta variável para este mesmo conjunto de dados.

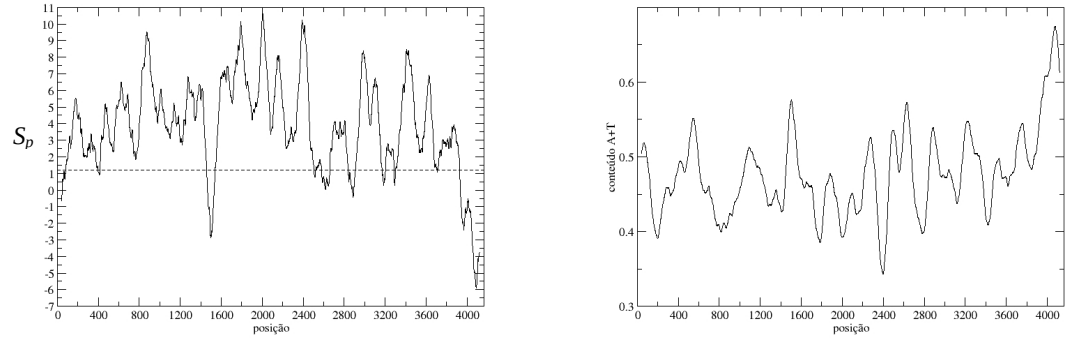
Os operons estudados possuem um grande número de genes, sendo esses pertencentes a uma mesma classe funcional de acordo com Daruvar et al. (2002). Na Figura 6.4, são apresentados os resultados para três operons de *E. coli*: *fecIRABCDE*, *fliLMNOPQR* e *trpLEDCBA*. A Tab. 6.6 especifica a localização de cada operon no genoma.

Tabela 6.6 – Mapeamento da localização dos operons de *E. coli* analisados.

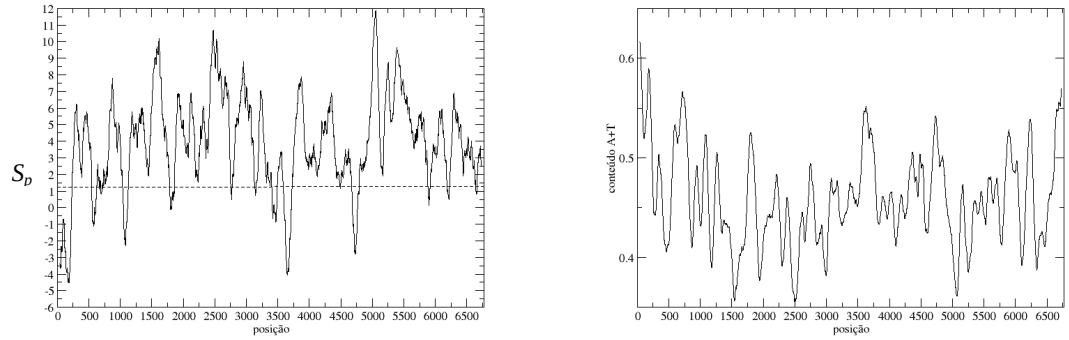
Operon	Base Inicial	Base Final	Fita	Número de Genes
fecIRABCDE	4.508.258	4.515.803	-	7
fliLMNOPQR	2.017.640	2.021.700	+	7
trpLEDCBA	1.314.440	1.321.106	-	6



(a) (a’)



(b) (b’)



(c) (c’)

Figura 6.4 – Análise de predição com ELDAD em curta escala os operons de *E. coli*: (a) *fecIRABCDE*, (b) *fliLMNOPQR* e (c) *trpLEDCBA*. (a), (b) e (c) são os gráficos dos S_p para as seqüências resultantes do uso de uma *sliding window* de 81 posições. A linha tracejada marca o S_c, e o promotor está possivelmente localizado nas primeiras 100 posições. Em (a’), (b’) e (c’) são apresentados os gráficos do conteúdo A+T do respectivo operon.

Os gráficos na figura 6.4 com o cálculo de S_p mostram que o ELDAD aponta várias regiões gênicas e intergênicas que seriam preditas como promotoras (as abaixo de S_c). Entretanto, deve-se considerar o número alto de genes no operon pode acarretar a existência de diferentes configurações de unidades de transcrição. Muitos operons possuem uma arquitetura simples, com um promotor regulado por uma única proteína. Entretanto, é comum encontrar estruturas mais complexas, com mais de uma proteína atuando na regulação ou com mais de um promotor. Sem contar, com estruturas com padrões de organização ainda mais complexos com diversos promotores localizados a *upstream* ou internamente no operon [Daruvar et al., 2002].

Os gráficos do conteúdo A+T confirmam que o modelo consegue capturar a hipótese de que promotores procarióticos também podem ser caracterizados por um alto conteúdo A+T internamente ou em suas proximidades [Kozobay-Avraham et al., 2004], de modo que mesmo as regiões em que não se espera a existência de promotor, mas que são classificadas como tal, são identificados picos de A+T.

6.2 *Bacillus subtilis*

O *Bacillus subtilis* foi utilizado para comparação dos resultados com *Escherichia coli* por duas razões: ele é a bactéria com o genoma maior e mais bem conhecido, distante suficientemente de *E. coli* [Daruvar et al., 2002], e possui um conjunto de 220 promotores conhecidos [Helmann, 1995], permitindo que a metodologia também seja testada para a tarefa de reconhecimento dessas regiões.

6.2.1 Reconhecimento

O reconhecimento de promotores de *B. subtilis* também considerou os protocolos ELD e ELDAD, utilizando os 220 promotores analisados por Helmann (1995). A comparação do desempenho entre os dois consta na Tab. 6.7.

Tabela 6.7 – Medições de escore limite e exatidão para o reconhecimento de promotores de *B. subtilis* com ELD e ELDAD para os dados de Helmann (1995).

Protocolo	Medida	Média	Desvio-padrão
ELD	S_c	132,966	1,374
	A_e	0,983	0,004
	A_o	0,95	0,031
ELDAD	S_c	-22,635	1,454
	A_e	0,986	0,003
	A_o	0,95	0,031

Os resultados do cálculo de exatisão são bastante próximos e indicam um excelente desempenho na tarefa de reconhecimento. Uma explicação para esta alta taxa de acerto pode ser o fato de todos os promotores dessa base serem regulados pelo mesmo fator σ, no caso σ^A, o qual é equivalente ao σ^{70} de *E. coli*. Mas deve-se ressaltar, que tal característica não pode ser generalizada, uma vez que a análise de *S* com fator σ não mostrou correlação.

Para verificar possíveis alterações nos padrões das regiões promotoras em *B. subtilis*, computou-se um logo de seqüências para um conjunto de dados gerados com o HMM promotor. Na Figura 6.5, observa-se uma alta conservação de *T*s, o que deve estar relacionado com o percentual de conteúdo A+T deste organismo (0.565) ser maior que o de *E. coli* (0.492).

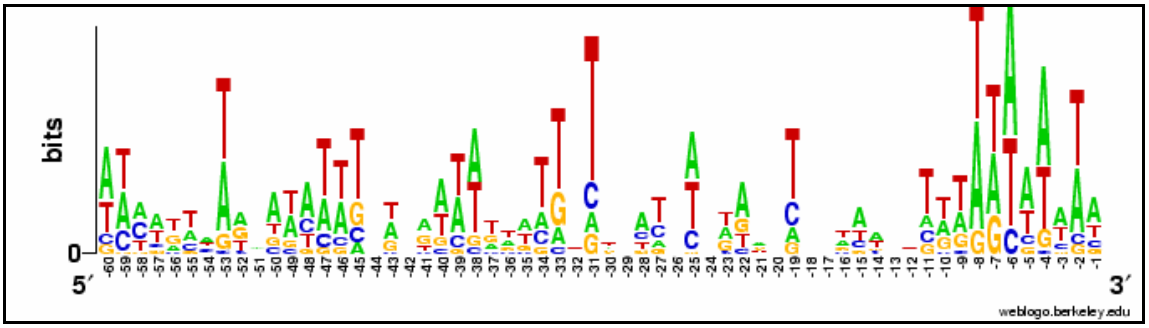


Figura 6.5 – Logo de seqüência para promotores gerados com o modelo promotor de *B. subtilis*. Nas 60 posições *upstream* são encontrados os padrões nas regiões dos hexâmeros -35 e -10, assim como uma alta conservação de *T*s ao longo de todas as posições, como pode ser comparado com o logo para o caso de *E. coli* (Figura 6.3).

6.2.2 Predição

A disponibilidade de um conjunto, mesmo que pequeno comparado ao RegulonDB, de promotores de *B. subtilis* possibilitou que fosse testada a utilização de predição com Análise de Discriminação, conforme especificado para o caso de larga escala, utilizando modelo HMM para promotor de *E. coli* com maior valor de exatidão.

Desta forma, foram criados 10 modelos para regiões gênicas de *B. subtilis*, com os quais foram computadas as distribuições D_{genes} baseadas na aderência dessas regiões ao HMM de promotor de *E. coli* e cada HMM para gene simulado com os dados de *B. subtilis*.

Para cada simulação, no lugar de usar ELD, calculou-se o S_p médio e seu desvio-padrão, e a partir desses valores computou-se a exatidão utilizando três possíveis valores de limiar: o próprio S_p médio, o S_p médio menos um desvio-padrão e o S_p médio menos 2 desvios-padrão. A exatidão foi calculada para um conjunto de teste formado pelos 220 promotores já citados mais 220 fragmentos gênicos do mesmo organismo. Essas medições constam na Tab. 6.8.

Tabela 6.8 – Medições de escore médio e exatidão para a predição de promotores de *B. subtilis* com Análise de Discriminação.

Simulação	S_p médio	Desvio-padrão (σ)	Exatidão		
			S_p médio	S_p médio - 1σ	S_p médio - 2σ
1	-19,623	4,656	0,725	0,775	0,693
2	-19,378	4,987	0,714	0,784	0,673
3	-19,320	4,96	0,714	0,775	0,68
4	-19,221	5,002	0,71	0,773	0,68
5	-19,046	4,943	0,702	0,775	0,693
6	-18,873	4,945	0,704	0,777	0,693
7	-18,480	4,889	0,693	0,786	0,709
8	-18,604	4,792	0,695	0,784	0,711
9	-18,613	4,874	0,695	0,777	0,702
10	-18,797	4,994	0,698	0,773	0,693

Os dados indicam o uso de S_p médio menos um desvio-padrão como o valor de limiar com maior exatidão sobre o conjunto de teste. Mas é importante notar que a exatidão que no reconhecimento está em torno de 0,95, na predição não alcança 0,8. Isso, possivelmente, pela falta de informação dos promotores de *B. subtilis* no caso de predição.

Para comparar o reconhecimento com a predição, foram exibidas em um mesmo gráfico: as D_{genes}, sem considerar a existência de promotores, e a $D_{promotores}$, obtida com base nos promotores de *E. coli*. A Figura 6.6 apresenta este gráfico. Observe que o ponto de divisão da área da região formada com o cruzamento dos dois tipos de distribuição está próximo do valor de S_p médio menos um desvio-padrão.

Além disso, é importante destacar que a D_{genes} do *B. subtilis* cruza a $D_{promotores}$ para promotores de *E. coli* pela direita. Isso mostra como o uso do protocolo ELDAD contribui para um melhor desempenho do HMM, pois, caso fosse utilizado o padrão, todas as seqüências gênicas do *B. subtilis* seriam reconhecidas como promotores (veja a Figura 6.1).

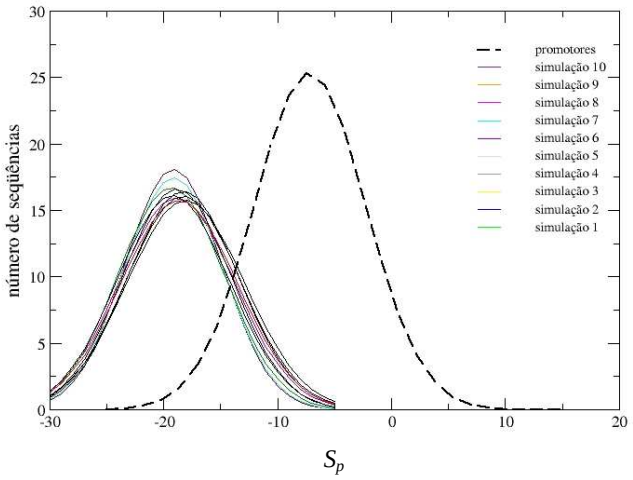


Figura 6.6 – Gráficos das distribuições de S_p consideradas para reconhecimento e predição. As curvas coloridas representam as D_{genes} para *B. subtilis*, enquanto a curva tracejada representa a $D_{promotores}$ que obteve maior exatidão para promotores de *E. coli*.

Esse deslocamento da ordem das distribuições de S_p no gráfico mostra que a análise prévia da distribuição dos promotores de *E. coli* em confronto com a de genes do organismo, para o qual se deseja realizar a predição, pode trazer compreensões mais abrangentes de como o HMM promotor executará a predição frente ao conteúdo A+T deste procarioto. Além disso, permitir a identificação de novas características próprias do novo genoma que influência no reconhecimento do promotor.

6.3 *Helicobacter pylori* (26695 e J99) e *Helicobacter hepaticus*

O uso de predição em curta escala foi selecionado para um estudo com todos os procariotos tratados neste trabalho, para o qual operons de *E. coli* foram mapeados nos demais organismos. Esta escolha ocorre porque, na ausência de promotores para validar os resultados, a análise somente de operons parece mais atrativa. Outra possibilidade seria utilizar apenas

regiões intergênicas, mas a complexidade envolvida na organização dos operons, como já discutido, descarta essa alternativa.

Aqui detalhamos os resultados de dois operons: *trpLEBCDA*, ausente em *Helicobacter hepaticus*, e *fliDST*, presente em todos os cinco genomas considerados. De acordo com Itoh et al. (1999), esses operons possuem uma alta conservação de sua estrutura tanto em *B. subtilis* quanto em *Helicobacter pylori*. O mapeamento da localização e da presença dos genes destes dois operons de *E. coli* nas outras espécies é indicado na Tab. 6.9.

Para identificar um possível padrão de comportamento da S_p na região promotora, ao invés de 100 pb *upstream* ao primeiro gene, foram mapeados 500 pb. As Figuras 6.7 e 6.8 apresentam os gráficos de S_p para o *trpLEBCA* e *fliDST*, respectivamente.

Os gráficos da Figura 6.7 mostram que o protocolo, apesar de errar, funciona bem para *E. coli*, pois uma grande parte de possíveis *FP* encontra-se na região entre S_p médio e S_p menos 1 desvio-padrão. Para o *B. subtilis* o resultado é um pouco inferior. No entanto, a taxa de erro da predição para as duas cepas de *Helicobacter pylori* é altíssima, chegando a 100% para a 26695. Contudo, é importante observar que a possível região promotora entre todos os organismos não é predita como seqüência gênica.

Esses resultados são confirmados com os gráficos para o operon *fliDST* (Figura 6.8), em que, novamente identifica-se a alta taxa de erro para o gênero *Helicobacter*.

Tabela 6.9 – Mapeamento dos operons *trpLEDCBA* e *fliDST* nas espécies investigadas.

Operon	**Gene**	***E. coli***			***B. subtilis***			***H. pylori* 26695**			***H. pylori* J99**			***H. hepaticus***		
		Base inicial	Base final	fita	Base inicial	Base final	fita	Base inicial	Base final	fita	Base inicial	Base final	fita	Base inicial	Base final	fita
trp	L	1.321.062	1.321.106	-	X			X			X			X		
	E	1.319.408	1.320.970	-	2.375.071	2.376.615	-	1.357.267	1.358.769	-	1.336.310	1.337.812	-	310.451	311.947	+
	D	1.317.813	1.319.408	-	2.374.083	2.375.096	-	1.356.686	1.357.270	-	1.334.725	1.335.732	-	308.843	310.447	+
	C	1.316.451	1.317.812	-	2.373.338	2.374.087	-	1.354.331	1.355.689	-	1.333.374	1.334.732	-	2.484	3.989	+
	B	1.315.246	1.316.439	-	2.371.503	2.372.702	-	1.353.148	1.354.329	-	1.332.191	1.333.372	-	1.354.518	1.355.780	+
	A	1.314.440	1.315.246	-	2.370.707	2.375.507	-	1.352.363	1.353.151	-	1.331.406	1.332.194	-	1.353.570	1.354.373	+
fli	D	2.001.896	2.003.302	+	3.631.944	3.633.437	-	806.840	808.864	+	769.101	771.158	+	766.121	768.169	+
	S	2.003.327	2.003.373	+	3.631.521	3.631.919	-	808.936	809.316	+	771.200	771.580	+	768.180	768.563	+
	T	2.003.737	2.004.102	+	3.631.183	3.631.521	-	X			X			X		

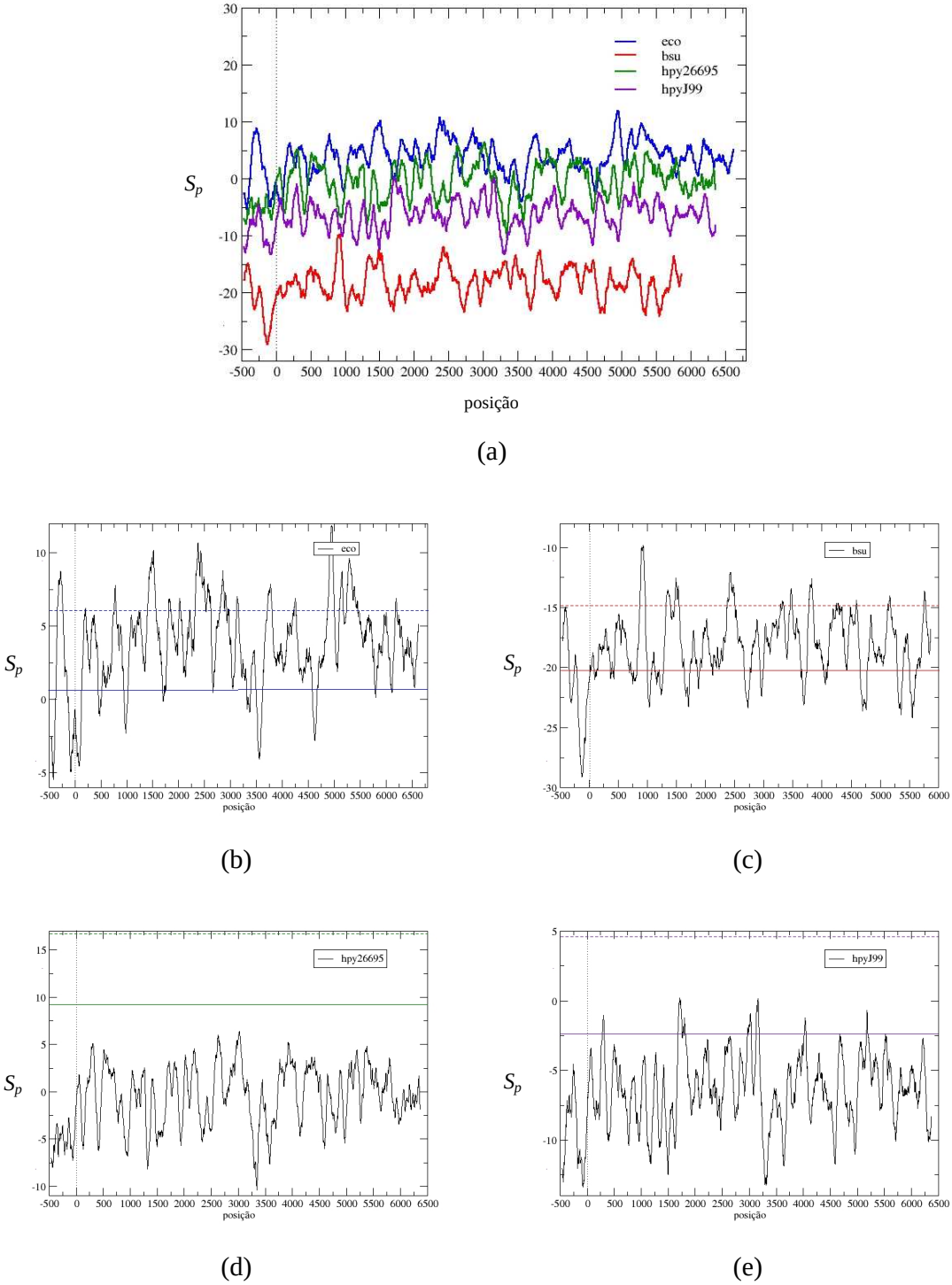


Figura 6.7 – Gráficos de S_p para o operon *trp*. Em (a) são apresentadas as curvas para todos os organismos. Cada uma delas é separada por organismo nos gráficos de (b) a (e). As linhas tracejadas representam o valor de S_c médio e as linhas representam o valor de S_c menos 1 desvio-padrão, este último considerado, conforme os estudos com *B. subtilis*, a melhor métrica para S_c na tarefa de predição em larga escala.

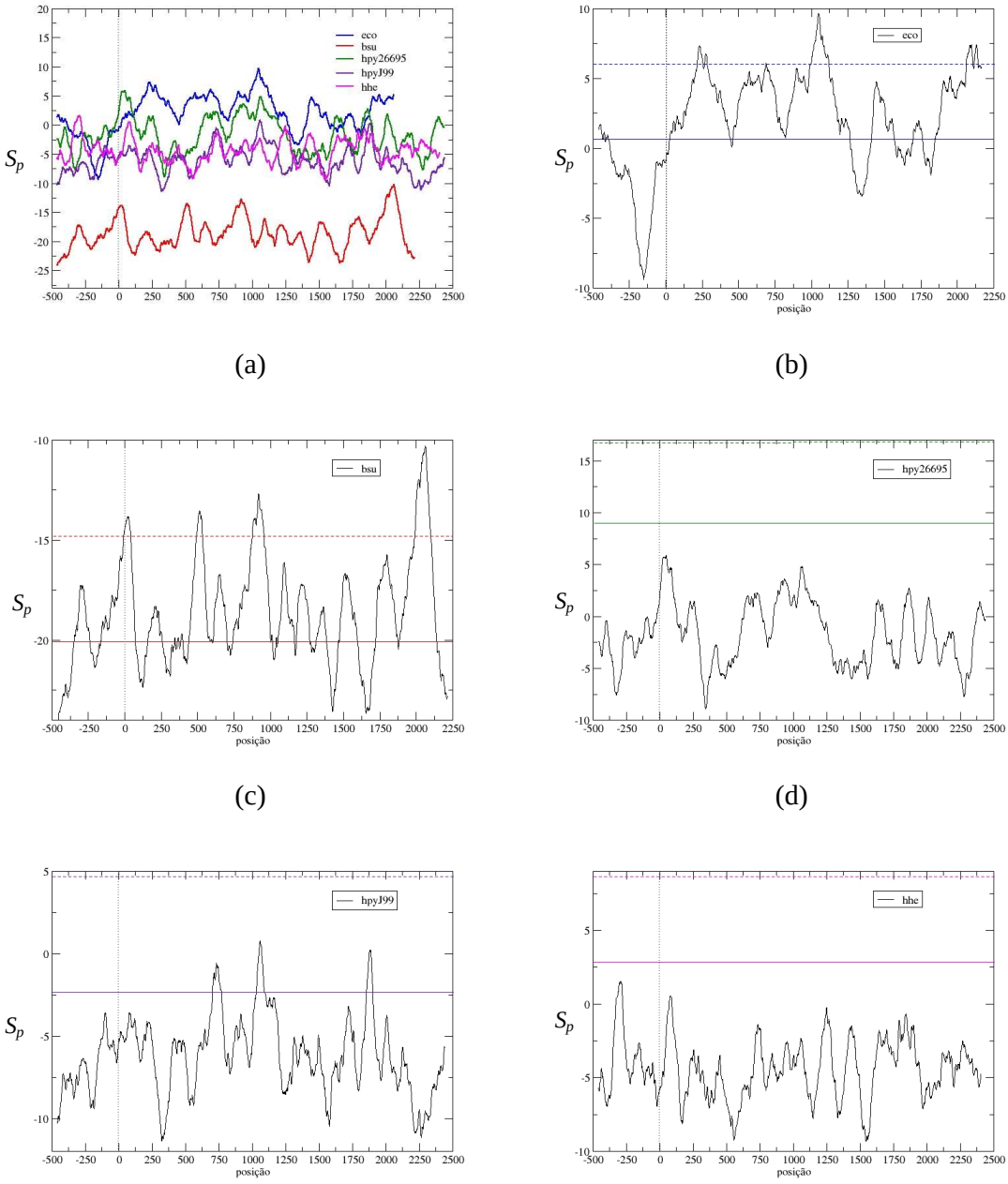
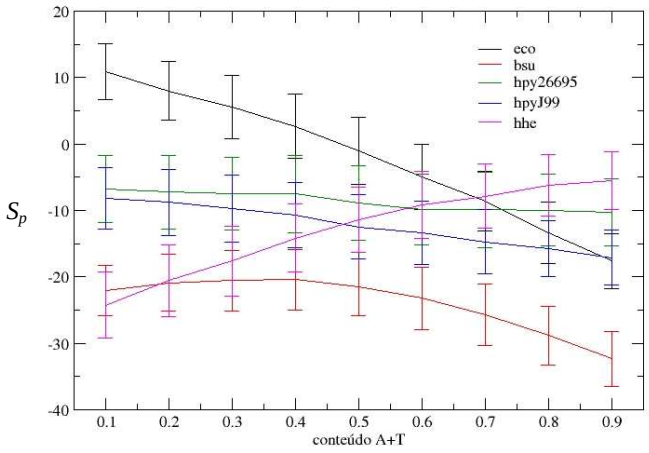
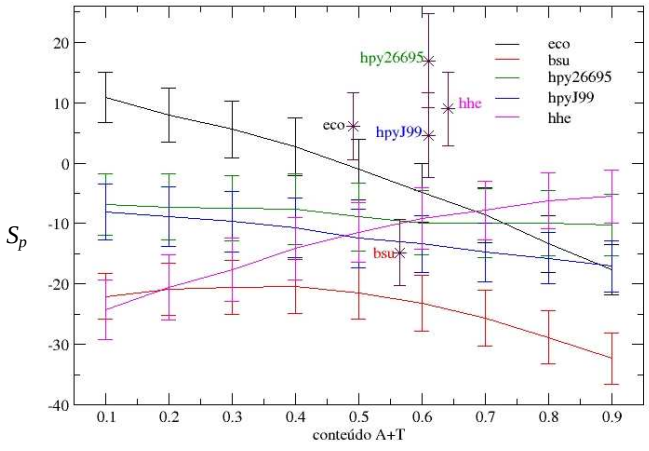


Figura 6.8 – Gráficos de S_p para o operon *fliDST*. Em (a) são apresentadas as curvas para todos os organismos. Cada uma delas é separada por organismo nos gráficos de (b) a (f). As linhas tracejadas representam o valor de S_c médio e as linhas representam o valor de S_c menos um desvio-padrão, este último considerado, conforme os estudos com *B. subtilis*, a melhor métrica para S_c na tarefa de predição em larga escala.

Frente ao baixo desempenho do uso do protocolo na predição em curta escala para *Helicobacter*, optou-se por repetir a análise dos modelos HMMs para seqüências aleatórias com diferentes percentuais de conteúdo A+T. O comportamento de S_p computado com o HMM para fragmentos gênicos de cada organismo e o HMM de *E. coli* constituem a Figura 6.9.



(a)



(b)

Figura 6.9 – Gráficos de S_p para seqüências aleatórias com diferentes valores de conteúdo A+T. Em (a) são indicados exclusivamente os resultados dos HMMs com as seqüências aleatórias, aos quais, em (b), são adicionados os resultados de S_p para seqüências gênicas das espécies estudadas.

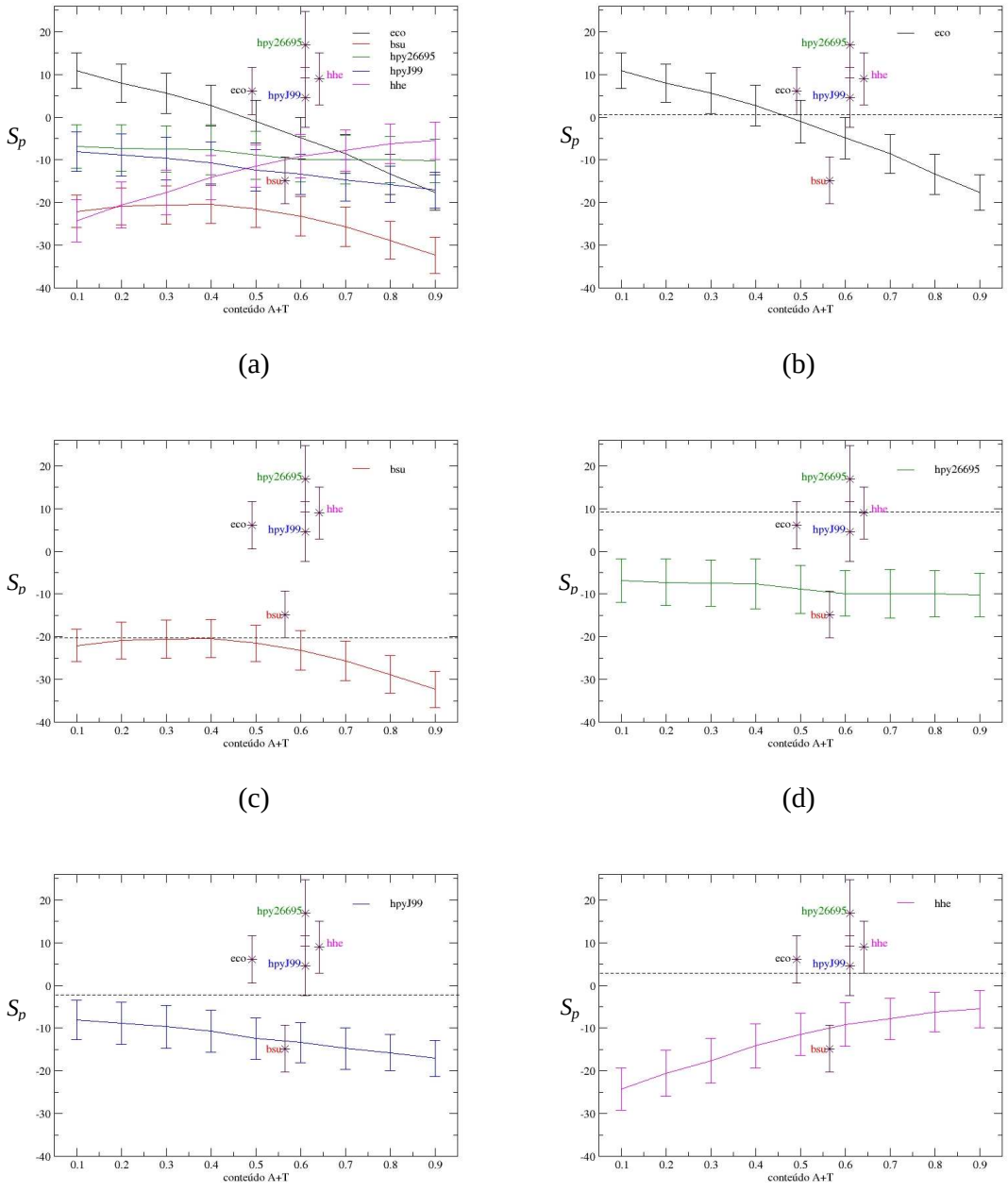


Figura 6.10 – Gráficos comparando o valor de S_p para seqüências gênicas das espécies estudadas com a curva de S_p para seqüências aleatórias com diferentes percentuais de conteúdo A+T. Em (a) constam todos os resultados, separados por espécie nos gráficos de (b) a (f). A linha tracejada indica o limiar para reconhecimento de promotor em cada organismo.

Um comportamento decrescente do valor de S_p conforme cresce o conteúdo A+T é obtido em todas as espécies com exceção do *Helicobacter hepaticus*. Além disso, apesar das duas cepas de *Helicobacter pylori* possuírem desempenhos diferentes na análise em curta escala (veja as Figuras 6.7 e 6.8), seus modelos HMMs resultam em comportamentos de S_p bastante semelhantes.

Comparando os resultados para seqüências gênicas de cada organismo com a curva de S_p obtida com o seu respectivo modelo HMM, verifica-se que para *E. coli*, o valor de S_p menos 1 desvio-padrão divide a curva praticamente na metade, indicando a forte influência do conteúdo A+T na aderência das seqüências ao modelo. Para *B. subtilis* esse comportamento é mais suave, sendo que o valor de S_p sofre uma queda relevante para valores de conteúdo A+T a partir de 0,5. Para o gênero *Helicobacter*, não é possível identificar correlação entre o valor de S_p para as suas seqüências gênicas e o resultado das seqüências aleatórias com variação do conteúdo A+T, além de todas as seqüências aleatórias serem reconhecidas como promotoras utilizando os S_c obtidos.

O baixo desempenho do HMM mesmo com a análise de discriminação para o *Helicobacter pylori* motivou o estudo das distribuições de escore de suas seqüências gênicas (D_{genes}), pois ocorrem muitos casos em que seqüências com apenas um nucleotídeo de diferença em sua constituição resultam valores de S_p distantes. Veja o exemplo na Tab. 6.10.

Tabela 6.10 – Exemplo de duas seqüências semelhantes do operon *fliDST* de *H. pylori* com escores distantes.

Seqüência	***S*$_p$**
>**AGTGCATCGCTAAAGGGTATGGGGATAAGCCCGGCCGCTCCAGCCGCTCCTGATGC AACATGGATAATGGTTTTACTTTCA**	–3,885
> **GTGCATCGCTAAAGGGTATGGGGATAAGCCCGGCCGCTCCAGCCGCTCCTGATGC AACATGGATAATGGTTTTACTTTCAT**	–13,862

Confrontando D_{genes} de *H. pylori* com a $D_{promotores}$ de *E. coli*, como apresentado na Figura 6.11, verificou-se que as duas distribuições praticamente se sobrepõem, não existindo discriminação evidente entre as duas.

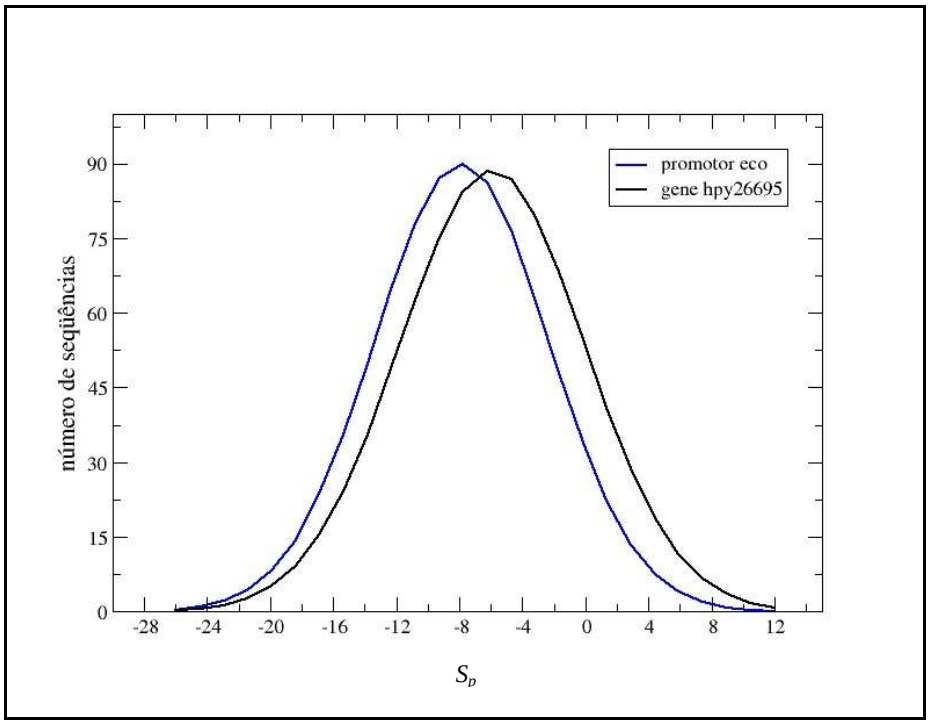


Figura 6.11 – Distribuição $D_{promotores}$ de *E.coli* e distribuição D_{genes} de *H. pylori*.

Esse resultado aponta para a necessidade de adicionar outras propriedades genômicas na metodologia para predição de promotores em procariotos. Segundo Lathe III et al. (2000), a ordem e o conteúdo dos genes, assim como os mecanismos regulatórios do operon podem ser muito diferentes, mesmo para espécies filogeneticamente próximas.

O fato das espécies do gênero *Helicobacter* serem patogênicas, de conterem mais de 60% de conteúdo A+T, valor observado nas regiões promotoras de *E. coli* (veja a Figura 6.4, (a') e (c')), de sua RNAp conter partes substancialmente diferentes da de *E. coli*, também são

outros aspectos que podem indicar um processo regulatório diferenciado ou até mesmo mais complexo [Marais, 1999; Vanet et al., 2000].

6.4 Considerações

Os resultados discutidos neste capítulo firmam a necessidade da existência de regiões promotoras para que o modelo HMM aplicado, com um protocolo adequado, resulte em baixas taxas de erro na tarefa de reconhecimento. Além disso, a tarefa de predição mostra que o uso de HMMs, independente do protocolo de aplicação utilizado, menos ou mais rigoroso, ainda requer ajustes e uma modelagem mais ampla do ponto de vista genômico e regulatório. Percebe-se, claramente, que características intrínsecas à organização do genoma ou aos mecanismos de regulação gênica em cada espécie possuem forte influência para a RNAp reconhecer o promotor.

As conclusões do uso dos protocolos ELD e ELDAD, para reconhecer e predizer promotores procarióticos, são apresentadas no próximo capítulo. Além disso, possíveis alternativas para aprimorar a capacidade de predição da metodologia baseada em HMMs com os protocolos propostos são discutidas.

7 Conclusões

O problema do reconhecimento e predição de promotores em organismos procarióticos ainda é um tema aberto na área de Bioinformática, apesar do crescente número de estudos experimentais *in vitro* e *in silico*. Essa dissertação propôs um protocolo para compreender, reconhecer e predizer regiões promotoras em procariotos, considerando dados genômicos experimentais e a teoria de *hidden* Markov *models* (HMMs).

Para isso foi formalizado um protocolo de aplicação de HMMs com Estimação do Limiar de Decisão (ELD) e Análise de Discriminação (ELDAD) para caracterizar promotores. O protocolo diferencia-se dos estudos anteriores por: considerar um conjunto maior de exemplos para treinamento (coletados do RegulonDB), adotar uma metodologia de validação estatística dos resultados através do *10-fold cross validation*, definir um método mais rigoroso para cálculo do limiar de decisão baseado na Regra de Bayes, utilizar um par de HMMs com análise de discriminação para deter-se nas propriedades exclusivas dos promotores, eliminando, por exemplo, a influência do conteúdo A+T na performance de reconhecimento, e aplicar o protocolo para predizer promotores em outros procariotos em larga e curta escala.

Os HMMs foram empregados tanto com o protocolo padrão, assim como os 2 propostos neste trabalho. Os testes foram realizados para 4 espécies procarióticas de acordo com suas informações disponíveis, sendo elas: *Escherichia coli*, *Bacillus subtilis*, *Helicobacter pylori* (cepas 26695 e J99) e *Helicobacter hepaticus*.

Em *Escherichia coli* e *Bacillus subtilis* foi possível analisar o reconhecimento e predição de promotores. Com a *E. coli* verificou-se que o acréscimo do número de seqüências para compor o conjunto de treinamento não foi suficiente para gerar aumento de exatidão

relevante no reconhecimento, o que acabou sendo obtido com a adoção do ELDAD, com o qual houve uma redução no erro em 44,26%. Em *B. subtilis* os resultados são excelentes, ultrapassando 90% de taxa de acerto. A partir da análise de predição com 220 promotores desta mesma espécie, foi possível ajustar o uso de HMM com ELDA para predição em outras espécies, sendo aqui tratadas as do gênero *Helicobacter*.

Para o *H. pylori* e *H. hepaticus*, o protocolo falha, pois as seqüências gênicas destes dois organismos possuem aderência semelhante ao modelo promotor de *E. coli*, de modo que o protocolo não consegue discriminar as duas classes de seqüências, gene e promotor. Isso indica que a comparação do comportamento de aderência das seqüências promotoras de *E. coli* com as gênicas do organismo, para o qual se pretende executar a predição, pode definir a viabilidade do protocolo ser aplicado em outras espécies não consideradas neste trabalho.

Contudo, foi possível detectar pontos em que a metodologia de HMM pode ser complementada com outras propriedades biomoleculares, e em que o protocolo pode adotar outras técnicas de Aprendizado de Máquina. Como trabalho futuro, um modo de empregar a tendência na distribuição nucleotídica dos genomas e outras informações estruturais ao modelo aparece como uma direção interessante no estabelecimento de uma metodologia robusta para predição de promotores procarióticos.

Como a função que rege a relação entre a maioria dos elementos hoje conhecidos para dizer se uma seqüência é ou não promotora é desconhecida, Redes Bayesianas aparecem como uma alternativa interessante capaz de utilizar diferentes tipos de informação de modo integrado para predição de promotores, por se tratar de uma abordagem que explora diferentes evidências. Uma rede Bayesiana pode ser definida como um grafo acíclico e direcionado que representa dependência entre variáveis em um modelo probabilístico [Bockhorst et al., 2003].

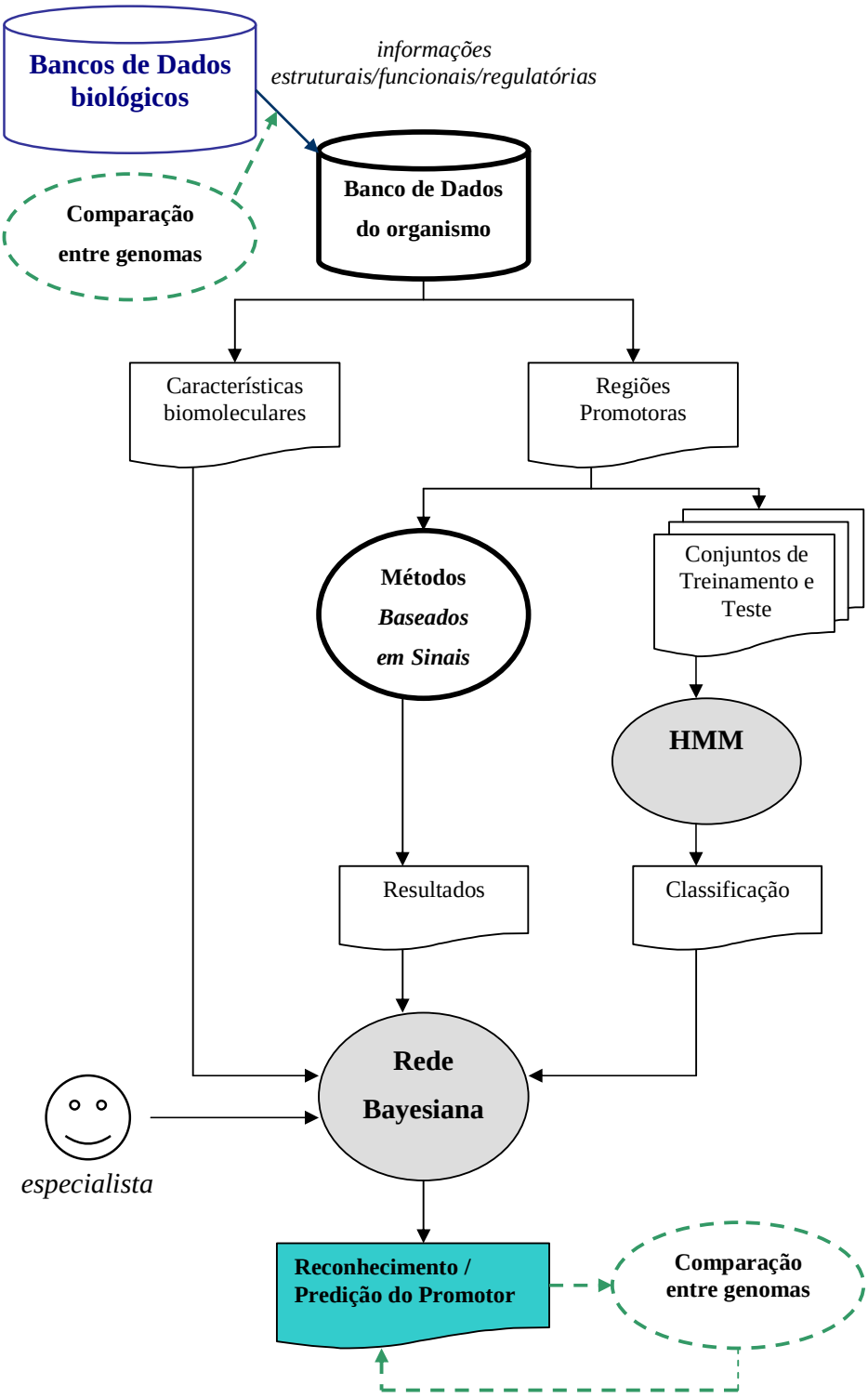


Figura 7.1 – Proposta de um novo protocolo para reconhecimento e predição de promotores procarióticos. Em cinza, são indicadas as ferramentas de Aprendizado de Máquina. A Comparação entre genomas está em tracejado por ser um módulo de análise opcional.

Para tanto, visto a dificuldade de obtenção de dados para treinamento, a premissa seria coletar o maior número de dados disponíveis em diferentes bases públicas para constituir o banco de dados de cada organismo. E, a partir desse, contemplar 3 cenários de integração:

- ⇒ *Integração de Métodos Computacionais:* integrar dados gerados por métodos baseados em sinal, HMMs e métricas estatísticas.
- ⇒ *Integração de Características Biomoleculares:* integrar diferentes classes de informações estruturais, conformacionais e funcionais sobre os promotores, tais como essencialidade do gene, o conteúdo A+T do genoma, entre outras.
- ⇒ *Integração de Métodos Computacionais e Características Biomoleculares:* reunir as duas abordagens acima definidas.

A proposta desse possível novo protocolo é apresentado esquematicamente na Figura 7.1. Observe que o resultado de cada método pode ser analisado individualmente, ou ainda em combinação com outros também computados. A Rede Bayesiana, além de possibilitar a integração de diferentes subconjuntos de características existentes para um dado organismo, permite integrar esses resultados utilizando o conhecimento prévio do especialista, que definirá as relações de causa e conseqüência.

Esta dissertação, assim como a proposta de trabalho futuro, reafirma a importância de considerar e agregar informações experimentais de caráter estrutural e funcional dos organismos nos HMMs para promotores. De modo a não apenas obter menores taxas de erro ao reconhecer e predizer essas regiões, mas também inferir mecanismos pertinentes de interação destas com a RNA-polimerase, a fim de contribuir de forma relevante para a compreensão da expressão gênica em procariotos.

Referências Bibliográficas

[1] Arslan, L., Hansen, J. H. L.: **Likelihood Decision Boundary Estimation between HMM pairs in Speech Recognition**. IEEE Transactions on Speech & Audio Processing, vol. 6, n. 4, 410-414 (1998).

[2] Baldi, P., Brunak, S.: **Bioinformatics: the machine learning approach**. MIT Press, 2ª ed. (2001).

[3] Baum, L., Petrie, T.: **Statistical inference for probabilistic functions of finite state Markov chains**. Annals of Mathematical Statistics, n. 37, 1554-1563 (1966).

[4] Bockhorst, J., Craven, M., Page, D., Shavlik, J., Glasner, J.: **A Bayesian network approach to operon prediction**. Bioinformatics, vol. 19, n. 10, 1227-1235 (2003).

[5] Clote, P., Backofen, R.: **Computational Molecular Biology: an introduction**. Wiley (2000).

[6] Craven M. W., Shavlik, J. W.: **Machine Learning Approaches to Gene Recognition**. IEEE Expert: Intelligent Systems and Their Applications, vol. 9, n. 2, 2-10 (1994).

[7] Crooks, G. E., Hon, G., Chandonia, J-M., Brenner, S.: **WebLogo: A Sequence Logo Generator.** Genome Research, n. 14, 1188-1190 (2004).

[8] Daruvar, A., Collado-Vides, J., Valencia, A.: **Analysis of the Cellular Functions of *Escherichia coli* Operons and Their Conservation in *Bacillus subtilis***. Journal of Molecular Evolution, 52, 211-221 (2002).

[9] Dimuro, G. P., Reiser, R. H. S., Costa, A. C. R., Souza, P. L. R.: **Modelos de Markov e Aplicações**. Em: Anais da VI Oficina de Inteligência Artificial, 37-59 (2002).

[10] Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: **Biological Sequence Analysis: probabilistic models of proteins and nucleic acids**. Cambridge (1998).

[11] Eddy, S. R.: **Profile hidden Markov models.** Bioinformatics, vol. 14, n. 9, 755-763 (1998).

[12] Ewens, W. J., Grant, G. R.: **Statistical Methods in Bioinformatics: an introduction**. Springer (2001).

[13] Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balázsi, G., Ravasz, E., Daugherty, M. D., Somera, A. L., Kyrpides, N. C., Anderson, I., Gelfand, M. S., Bhattacharya, A., Kapatral, V., D'Souza, M., Baev, M. V., Grechkin, Y., Mseeh, F., Fonstein, M. Y., Overbeek, R., Barabási, A.-L., Oltvai, Z. N., Osterman, A. L.: **Experimental Determination and System Level Analysis of Essential Genes in *Escherichia coli* MG1655**. Journal of Bacteriology, vol. 185, n.19, 5673-5684 (2003).

[14] Gibas, C., Jambeck, P.: **Desenvolvendo Bioinformática**. Campus (2001).

[15] Guimarães, K. S., Melo, J. C. B.: **Uma Introdução à Análise de Seqüências e Estruturas Biológicas**. Em: Anais da XXII Jornada de Atualização em Informática (JAI) (2003).

[16] Helmann, J. D.: **Compilation and analysis of *Bacillus subtilis* σ^A-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA**. Nucleic Acids Research, vol. 23, n. 13, 2351-2360 (1995).

[17] Hershberg, R., Bejerano, G., Santos-Zavaleta, A., Margalit, H.: **PromEC: An updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites**. Nucleic Acids Research, vol. 29, n. 1, 277 (2001).

[18] Hertz, G. Z., Stormo, G. D.: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences**. Bioinformatics, vol. 15, n. 7/8, 563-577 (1999).

[19] Huerta, A. M., Salgado, H., Thieffry, D., Collado-Vides, J.: **RegulonDB: a database on transcriptional regulation in *Escherichia coli***. Nucleic Acids Research, vol. 26, n. 1, 55-59 (1998).

[20] Huerta, A. M., Collado-Vides, J.: **Sigma70 Promoters in *Escherichia coli*: Specific Transcription in Dense Regions of Overlapping Promoter-like Signals**. Journal of Molecular Biology, vol. 333, 261-278 (2003).

[21] Itoh, T., Takemoto, K., Mori, H., Gojoboru, T.: **Evolutionary Instability of Operon Structures Disclosed by Sequence Comparisons of Complete Microbial Genomes**. Molecular Biology and Evolution, vol. 16, n. 3, 332-346 (1999).

[22] Karp, P. D., Arnaud, M., Collado-Vides, J., Ingraham, J., Paulsen, T., Saier, M. H. Jr.: **The *E. coli* EcoCyc Database: No Longer Just a Metabolic Pathway Database**. ASM News, vol. 70, n. 1, 25-30 (2004).

[23] Kellis, M., Patterson, N., Endrizzi, M., Birren, B., Lander, E. S.: **Sequencing and comparison of yeast species to identify genes and regulatory elements**. Nature, vol. 423, 241-254 (2003).

[24] Kozobay-Avraham, L., Hosid, S., Bolshoy, A.: **Curvature distribution in prokaryotic genomes**. In Silico Biology, vol. 4, n. 29 (2004).

[25] Krogh, A.: **An Introduction to Hidden Markov Models for Biological Sequences**. Em: Computational Methods in Molecular Biology, Elsevier, 45-63 (1998).

[26] Lathe III, W. C., Snel, B., Bork, P.: **Gene context conservation of a higher order than operons**. Trends in Biochemical Sciences, vol. 25, n.10 (2000).

[27] Lewin, B.: **Genes VII**. Artmed Editora (2001).

[28] Lisser, S., Margalit, H.: **Compilation of *E. coli* mRNA promoter sequences**. Nucleic Acids Research, vol. 21, n. 7, 1507-1516 (1993).

[29] Lukashin, A. V., Borodovsky, M.: **GeneMark.hmm: new solutions for gene finding**. Nucleic Acids Research, vol. 26, n. 4, 1107-1115 (1998).

[30] Manning, C. D.: **Foundations of Statistical Language Processing**. Cambridge-MIT (2001).

[31] Marais, A., Mendz, G. L., Hazell, S. L., Mégraud, F.: **Metabolism and Genetics of *Helicobacter pylori*: the Genoma Era**. Microbiology and Molecular Biology Reviews, vol. 63, n. 3, 642-674 (1999).

[32] Meneses, C.: **Modelos de Markov não observáveis**. Processamento Digital da Fala (http://www.deetc.isel.ipl.pt/comunicacoesep/disciplinas/pdf/index.html) Último acesso: 28/12/2004 (2002)

[33] Mooney, R. A., Landick, R.: **RNA Polymerase Unveiled**. Cell, vol. 98, 687-690 (1999).

[34] Mount, D. W.: **Bioinformatics: sequence and genome analysis**. CSHL Press (2000).

[35] Murakami, K. S., Masuda, S., Campbell, E. A., Muzzin, O., Darst, S. A.: **Structural Basis of Transcription Initiation: An RNA Polymerase Holoenzyme-DNA complex**. Science, vol. 296, 1285-1290 (2002).

[36] Naryshkin, N., Revyakin, A., Kim, Y., Mekler, V., Ebright, R. H.: **Structural Organization of the RNA Polymerase-Promoter Open Complex**. Cell, vol. 101, 601-611 (2000).

[37] Net-ID, I.: HMMpro: **A Hidden Markov Model (HMM) Simulator**. http://www.netid.com/html/hmmpro.html (último acesso: 23 de julho de 2003).

[38] Nilsson, N. J.: **Artificial Intelligence: a new synthesis**. Morgan Kaufmann (1998).

[39] Oppon, E. C.: **Synergistic Use of Promoter Prediction Algorithms: a choice for a small training dataset?**. Tese de Doutorado, South African National Bioinformatics Institute (SANBI) (2000).

[40] Pedersen, A. G., Engelbrecht, J.: **Investigations of *Escherichia coli* Promoter Sequences With Artificial Neural Networks: New Signals Discovered Upstream of the Transcriptional Startpoint**. Em: Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology (ISMB-95), vol. 3, 292-299 (1995).

[41] Pedersen, A. G., Baldi, P., Brunak, S., Chauvin, Y.: **Characterization of prokaryotic and eukaryotic promoters using hidden Markov models**. Em: Proceedings for Fourth International Conference on Intelligent Systems for Molecular Biology, 182-191 (1996).

[42] Pevzner, A. P.: Computational Molecular Biology: an algorithmic approach. MIT Press (2000).

[43] Qiu, P.: **Recent advances in computational promoter analysis in understanding the transcriptional regulatory network**. Biochemical and Biophysical Research Communications, vol. 309, 495-501 (2003).

[44] Reis, A. N., Lemke, N.: **Aplicando HMMs no Reconhecimento de Regiões Promotoras em Genomas Procarióticos**. Em: Anais do I WorkComp Sul, mídia digital (2004).

[45] Reis, A. N., Lemke, N.: **Análise de um Modelo HMM para Predição de Regiões Promotoras Procarióticas com Base em Dados de *Escherichia coli***. Em: XXVII Congresso Nacional de Matemática Aplicada e Computacional - CNMAC (2004).

[46] Salgado, H., Gama-Castro, S., Martínez-Antonio, A., Díaz-Peredo, E., Sánchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jiménez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martínez, C., Collado-Vides, J.: **RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12**. Nucleic Acids Research, vol. 32, D303-D306 (2004).

[47] Salzberg, S. L., Delcher, A. L., Kasif, S., White, O.: **Microbial gene identification using interpolated Markov models**. Nucleic Acids Research, vol. 26, n. 2, 544-548 (1998).

[48] Schneider, T. D.: **Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation**. Nucleic Acids Research, vol. 29, n. 23, 4881-4891 (2001).

[49] Schneider, T. D.: **Some lessons for molecular biology from information theory**. Special Series on Studies in Fuzziness and Soft Computing, vol. 119, 229-237 (2003).

[50] Setubal, J. C., Meidanis, J.: **Introduction to Computational Molecular Biology**. ITP (1997).

[51] Souto, M. C. P., Delbem, A. C. B., Carvalho, A. C. P. L. F.: **Técnicas de Aprendizado de Máquina para Problemas de Biologia Molecular**. Em: Anais da III Jornada de Mini-Cursos de Inteligência Artificial (MCIA), 103-152 (2003).

[52] Stormo, G. D.: **DNA binding sites: representation and discovery**. Bioinformatics, vol. 16, 16-23 (2000).

[53] Suerbaum, S., Josenhans, C., Sterzenbach, T., Drescher, B., Brandt, P., Bell, M., Dröge, M., Fartmann, B., Fischer, H., Ge, Z., Hörster, A., Holland, R., Klein, K., König, J., Macko, L., Mendz, G. L., Nyakatura, G., Schauer, D. B., Shen, Z., Weber, J., Frosch, M., Fox, J. G.: **The complete genome sequence of the carcinogenic bacterium *Helicobacter hepaticus***. PNAS, vol. 100, n. 13, 7901-7906 (2003).

[54] Towell, G., Shavlik, J., Noordewier, M.: **Refinement of Approximate Domain Theories by Knowledge-based Neural Networks**. Em: Proceedings of the National Conference on Artificial Intelligence, 861-866 (1990).

[55] Vanet, A., Marsan, L., Labigne, A., Sagot, M.: **Inferring Regulatory Elements from a Whole Genome. An Analysis of *Helicobacter pylori* σ^{80} Family of Promoter Signals**. Journal of Molecular Biology, 257, 335-353 (2000).

[56] Wang, J. T. L., Shapiro, B. A., Shasha, D.: **Pattern Discovery in Biomolecular Data: tools, techniques, and applications**. Oxford (1999).

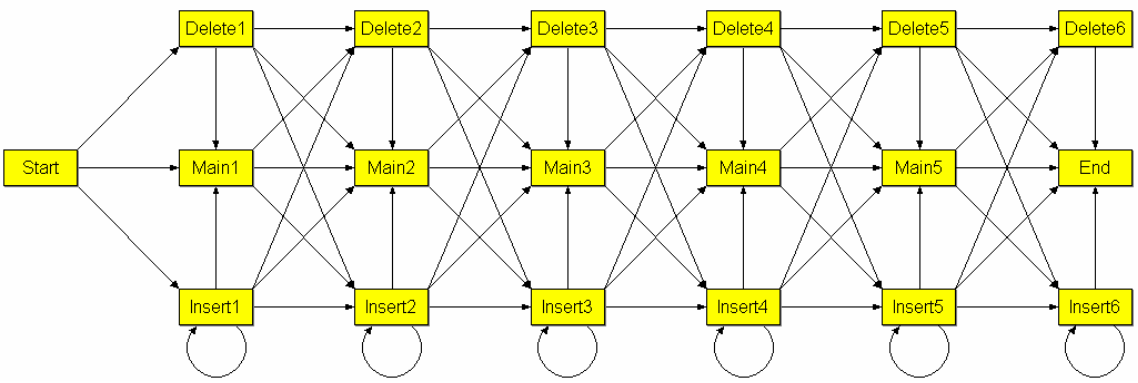
[57] Watson, J. D.: **DNA: the secret of life**. Alfred A. Knopf (2003).

[58] Wu, C. H., McLarty, J. W**.: Neural Networks and Genome Informatics**. Elsevier Science (2000).

Apêndice A – Topologia Padrão dos HMMs Treinados

Neste trabalho, é seguido um padrão de topologia para todos os modelos HMMs treinados. Ela é do tipo linear, mas, diferente da estrutura apresentada na Figura 4.1, possui os estados completamente conectados.

Por questões de espaço, a topologia é apresentada abaixo considerando apenas 5 estados principais, ao invés dos 81. A imagem foi extraída da ferramenta HMMpro.



Apêndice B – Modelo HMM para promotores de *E. coli*

Abaixo são explicitados os parâmetros do modelo HMM para promotores de *E. coli* treinado com o RegulonDB versão 4.0, e que obteve maior exatidão dentre os do *10-fold cross validation.*

A formatação segue a seguinte estrutura:

Insert x	T	Main x	Insert x	Main x+1	Insert x+1	Delete x+1
	E	P(A)	P(T)	P(C)	P(G)	
Delete x	T	Main x	Main x+1	Insert x+1	Delete x+1	
Main x	T	Main x+1	Insert x+1	Delete x+1		
	E	P(A)	P(T)	P(C)	P(G)	

O “T” indica que na linha consta as probabilidades de transição para cada estado na ordem apresentada, enquanto o “E” representa as probalidades de emissão em cada estado. Abaixo, os parâmetros do HMM são listados.

```
------------------------------------------------------------------------
Start       T     0.313252 0.631498 0.0552499
Insert 1    T     0.0284052 0.210259 0.0621462 0.315198 0.383991
            E     0.0787851 0.201863 0.655269 0.064083
Delete 1    T     0.601505 0.0415358 0.266937 0.0900217
Main 1      T     0.660906 0.317479 0.0216147
            E     0.569301 0.165659 0.056022 0.209018
Insert 2    T     0.0330935 0.37273 0.141533 0.285142 0.167502
            E     0.616294 0.301752 0.0330609 0.0488928
Delete 2    T     0.040616 0.00668886 0.908219 0.044476
Main 2      T     0.626193 0.363817 0.00998983
            E     0.00806868 0.0982503 0.0380526 0.855628
Insert 3    T     0.303244 0.234416 0.0979181 0.0342999 0.330122
            E     0.0387387 0.412919 0.199891 0.348451
Delete 3    T     0.172514 0.270995 0.213115 0.343376
Main 3      T     0.22202 0.671218 0.106762
            E     0.0238491 0.1748 0.688815 0.112536
Insert 4    T     0.0957057 0.166644 0.0150568 0.711613 0.0109811
            E     0.822762 0.041927 0.125857 0.00945448
Delete 4    T     0.296026 0.20137 0.174337 0.328267
Main 4      T     0.68322 0.131913 0.184867
            E     0.0386748 0.501462 0.348673 0.11119
Insert 5    T     0.314329 0.0501395 0.123301 0.366954 0.145276
            E     0.485961 0.0955944 0.0359569 0.382488
Delete 5    T     0.322296 0.183217 0.4073 0.0871874
Main 5      T     0.632844 0.269961 0.0971943
```

```
            E     0.0192001 0.515882 0.0175486 0.44737
Insert 6    T     0.0674376 0.394058 0.444672 0.041274 0.0525577
            E     0.170301 0.158152 0.189419 0.482128
Delete 6    T     0.330496 0.176941 0.3363 0.156264
Main 6      T     0.0860326 0.841865 0.0721027
            E     0.130455 0.0144899 0.829799 0.0252563
Insert 7    T     0.0437104 0.395015 0.328257 0.194233 0.0387858
            E     0.149878 0.214615 0.499209 0.136299
Delete 7    T     0.372618 0.128103 0.428737 0.0705421
Main 7      T     0.449807 0.323596 0.226598
            E     0.281828 0.551209 0.0883194 0.078643
Insert 8    T     0.0201457 0.450021 0.0817862 0.176297 0.27175
            E     0.832003 0.1424 0.011658 0.0139392
Delete 8    T     0.646337 0.280539 0.0320372 0.0410867
Main 8      T     0.187331 0.742902 0.0697669
            E     0.010505 0.457738 0.00284364 0.528913
Insert 9    T     0.248941 0.0906195 0.0664981 0.0208549 0.573087
            E     0.0279019 0.012155 0.906168 0.0537753
Delete 9    T     0.211147 0.70348 0.053268 0.0321052
Main 9      T     0.0757736 0.909827 0.0143994
            E     0.630086 0.0107291 0.0106992 0.348486
Insert 10   T     0.351881 0.204761 0.184604 0.0415592 0.217195
            E     0.0741618 0.247727 0.0396919 0.63842
Delete 10   T     0.253366 0.195608 0.199486 0.351539
Main 10     T     0.23409 0.724698 0.0412122
            E     0.166618 0.452623 0.350744 0.0300149
Insert 11   T     0.0325255 0.092342 0.816637 0.0213153 0.0371807
            E     0.0787166 0.346402 0.0670127 0.507869
Delete 11   T     0.259965 0.0458089 0.466539 0.227687
Main 11     T     0.0556157 0.0457595 0.898625
            E     0.0420087 0.857843 0.00738945 0.0927591
Insert 12   T     0.0615913 0.100226 0.0561322 0.228268 0.553782
            E     0.718702 0.11427 0.0879306 0.0790979
Delete 12   T     0.0557328 0.28268 0.594849 0.0667377
Main 12     T     0.795265 0.0325715 0.172163
            E     0.0213813 0.159604 0.472481 0.346534
Insert 13   T     0.144932 0.0434615 0.632559 0.0641595 0.114888
            E     0.655371 0.0186493 0.272696 0.053283
Delete 13   T     0.0941786 0.0847502 0.773695 0.0473763
Main 13     T     0.251472 0.231658 0.51687
            E     0.130278 0.0245331 0.789989 0.0551998
Insert 14   T     0.0132954 0.0422826 0.0296508 0.900353 0.0144186
            E     0.860712 0.0848026 0.0187514 0.0357341
Delete 14   T     0.149173 0.773061 0.0124644 0.0653023
Main 14     T     0.520413 0.0311731 0.448414
            E     0.0207114 0.513288 0.0973798 0.368621
Insert 15   T     0.0306836 0.0608415 0.0419142 0.381777 0.484784
            E     0.811354 0.0427104 0.0103839 0.135552
Delete 15   T     0.282137 0.573223 0.0795387 0.0651009
Main 15     T     0.296511 0.567728 0.135762
            E     0.00422992 0.415962 0.012136 0.567672
Insert 16   T     0.137508 0.0589504 0.0275962 0.743293 0.032653
            E     0.430321 0.0523315 0.233172 0.284176
Delete 16   T     0.0446135 0.912913 0.0221431 0.0203302
Main 16     T     0.933387 0.0223945 0.044218
            E     0.0110839 0.00871527 0.968154 0.0120471
Insert 17   T     0.0975589 0.0259608 0.0411888 0.770054 0.0652371
            E     0.0217928 0.490765 0.132917 0.354525
Delete 17   T     0.5241 0.0963202 0.268112 0.111467
Main 17     T     0.822481 0.112508 0.0650115
            E     0.228518 0.38028 0.368924 0.0222781
```

```
Insert 18   T     0.0133388 0.0108669 0.8899 0.0211163 0.0647781
            E     0.0194674 0.640647 0.118431 0.221454
Delete 18   T     0.248116 0.135132 0.547617 0.0691347
Main 18     T     0.222703 0.619906 0.157391
            E     0.287236 0.10617 0.0172814 0.589313
Insert 19   T     0.810056 0.0159136 0.00597077 0.152836 0.0152233
            E     0.0505723 0.0152753 0.830293 0.103859
Delete 19   T     0.28662 0.598527 0.0912015 0.0236518
Main 19     T     0.494294 0.192694 0.313012
            E     0.693557 0.0515567 0.193551 0.0613354
Insert 20   T     0.0139793 0.0498603 0.00823502 0.919748 0.0081777
            E     0.836764 0.00916855 0.0108601 0.143207
Delete 20   T     0.109499 0.818201 0.0358011 0.0364989
Main 20     T     0.975755 0.00721485 0.0170304
            E     0.0374159 0.286556 0.261101 0.414928
Insert 21   T     0.228462 0.0373898 0.14825 0.23897 0.346927
            E     0.609375 0.0562521 0.00778716 0.326586
Delete 21   T     0.57521 0.169373 0.179702 0.0757147
Main 21     T     0.21241 0.469537 0.318053
            E     0.00395473 0.580161 0.0142544 0.401629
Insert 22   T     0.0750639 0.0453384 0.798976 0.0218459 0.058776
            E     0.0539248 0.111942 0.799193 0.0349397
Delete 22   T     0.235586 0.420019 0.198733 0.145661
Main 22     T     0.0270143 0.95751 0.0154753
            E     0.586262 0.0362016 0.155157 0.222379
Insert 23   T     0.0121823 0.0549738 0.0391426 0.87281 0.020891
            E     0.336912 0.2143 0.0541571 0.39463
Delete 23   T     0.645052 0.192122 0.0963964 0.0664299
Main 23     T     0.543627 0.0545388 0.401835
            E     0.115905 0.334826 0.543082 0.00618749
Insert 24   T     0.0167282 0.0776711 0.0437001 0.840428 0.0214722
            E     0.0872881 0.612121 0.00888342 0.291707
Delete 24   T     0.428335 0.280165 0.219843 0.071657
Main 24     T     0.933547 0.0332894 0.0331634
            E     0.182576 0.354916 0.0137699 0.448739
Insert 25   T     0.0135868 0.150855 0.800383 0.00857452 0.0266004
            E     0.183107 0.0235511 0.49323 0.300112
Delete 25   T     0.794638 0.0974269 0.0706873 0.0372476
Main 25     T     0.0353293 0.836654 0.128016
            E     0.239324 0.0053753 0.732504 0.022796
Insert 26   T     0.152573 0.0553278 0.0229924 0.754037 0.0150699
            E     0.40464 0.247127 0.228867 0.119366
Delete 26   T     0.228331 0.634222 0.0392783 0.0981689
Main 26     T     0.974723 0.00658483 0.0186918
            E     0.0142659 0.365261 0.305116 0.315357
Insert 27   T     0.0119001 0.0424751 0.375024 0.471114 0.0994874
            E     0.657851 0.0159922 0.00680252 0.319354
Delete 27   T     0.686608 0.097876 0.153317 0.0621987
Main 27     T     0.251981 0.486323 0.261696
            E     0.0065008 0.842123 0.0650116 0.0863645
Insert 28   T     0.205543 0.0101757 0.741441 0.0153698 0.0274706
            E     0.331166 0.148477 0.147139 0.373218
Delete 28   T     0.46375 0.150433 0.336775 0.0490424
Main 28     T     0.0769299 0.279082 0.643988
            E     0.165844 0.0163748 0.811854 0.00592662
Insert 29   T     0.0472835 0.325936 0.0538791 0.511651 0.0612506
            E     0.842565 0.0289722 0.0931869 0.035276
Delete 29   T     0.110067 0.121716 0.490869 0.277349
Main 29     T     0.669531 0.152347 0.178122
            E     0.0333115 0.278302 0.307452 0.380934
Insert 30   T     0.0617074 0.0359062 0.745631 0.0378422 0.118913
```

```
            E     0.15968 0.313013 0.279794 0.247512
Delete 30   T     0.217206 0.26562 0.471489 0.0456852
Main 30     T     0.0281496 0.947269 0.0245811
            E     0.00722253 0.586256 0.0189316 0.38759
Insert 31   T     0.915483 0.0191666 0.0264518 0.0236396 0.0152588
            E     0.317854 0.0088543 0.534277 0.139014
Delete 31   T     0.505337 0.261887 0.157675 0.0751009
Main 31     T     0.655737 0.316397 0.0278667
            E     0.240963 0.264242 0.378904 0.115891
Insert 32   T     0.756082 0.0598628 0.0437866 0.111846 0.0284223
            E     0.22196 0.456362 0.284599 0.0370792
Delete 32   T     0.8352 0.0572412 0.0718608 0.0356977
Main 32     T     0.425657 0.535599 0.038744
            E     0.127784 0.59352 0.00747816 0.271217
Insert 33   T     0.0275554 0.178198 0.0279944 0.741379 0.0248726
            E     0.135985 0.0195686 0.548055 0.296392
Delete 33   T     0.753638 0.0618172 0.148993 0.0355515
Main 33     T     0.605876 0.286467 0.107657
            E     0.506424 0.0180641 0.455926 0.0195853
Insert 34   T     0.135904 0.188223 0.362434 0.0207112 0.292729
            E     0.0114326 0.379995 0.41272 0.195852
Delete 34   T     0.552599 0.283405 0.0608322 0.103163
Main 34     T     0.0381586 0.955427 0.00641485
            E     0.912752 0.0159037 0.0439588 0.0273856
Insert 35   T     0.104328 0.0950128 0.256447 0.0371933 0.507018
            E     0.554846 0.0832152 0.00592633 0.356012
Delete 35   T     0.418124 0.456181 0.0795368 0.0461587
Main 35     T     0.0936662 0.589212 0.317121
            E     0.121511 0.359711 0.112313 0.406466
Insert 36   T     0.276924 0.13688 0.025247 0.538025 0.022924
            E     0.0445488 0.00858843 0.403896 0.542967
Delete 36   T     0.176774 0.288838 0.507603 0.0267852
Main 36     T     0.756457 0.0524338 0.191109
            E     0.284011 0.136809 0.485843 0.0933367
Insert 37   T     0.0162231 0.0111798 0.0197425 0.940865 0.0119895
            E     0.295247 0.205537 0.259613 0.239604
Delete 37   T     0.61278 0.121664 0.202975 0.0625809
Main 37     T     0.936044 0.0218279 0.0421277
            E     0.299193 0.372472 0.323908 0.00442734
Insert 38   T     0.00739513 0.00811821 0.0840426 0.834049 0.0663946
            E     0.00957294 0.0521754 0.00985634 0.928395
Delete 38   T     0.589604 0.126994 0.210392 0.0730096
Main 38     T     0.801607 0.0544515 0.143941
            E     0.329179 0.558205 0.00985066 0.102764
Insert 39   T     0.400346 0.0409373 0.111556 0.360599 0.0865618
            E     0.0150271 0.285254 0.434935 0.264784
Delete 39   T     0.331846 0.46045 0.110211 0.097493
Main 39     T     0.342899 0.597355 0.0597458
            E     0.219522 0.046807 0.665224 0.0684478
Insert 40   T     0.244141 0.0543063 0.0138276 0.680477 0.00724767
            E     0.457265 0.340696 0.0266929 0.175346
Delete 40   T     0.71866 0.0503568 0.113889 0.117095
Main 40     T     0.114879 0.14437 0.740752
            E     0.00692666 0.977619 0.00957808 0.00587635
Insert 41   T     0.0326327 0.0176201 0.033748 0.822152 0.0938478
            E     0.432433 0.174546 0.0990394 0.293982
Delete 41   T     0.170417 0.162564 0.471075 0.195945
Main 41     T     0.953385 0.0284823 0.0181325
            E     0.108039 0.146877 0.0435751 0.701508
Insert 42   T     0.106687 0.0154155 0.0227541 0.820295 0.034848
            E     0.222849 0.132752 0.472527 0.171872
```

```
Delete 42   T     0.145185 0.7146 0.106168 0.0340461
Main 42     T     0.858092 0.111388 0.03052
            E     0.00474981 0.00417269 0.979009 0.0120686
Insert 43   T     0.0692526 0.0625661 0.129248 0.697728 0.041205
            E     0.230264 0.343887 0.0837145 0.342134
Delete 43   T     0.599048 0.201013 0.129325 0.0706139
Main 43     T     0.710221 0.244748 0.0450314
            E     0.181393 0.0250281 0.771124 0.0224551
Insert 44   T     0.0446156 0.0925308 0.165763 0.623271 0.0738195
            E     0.0295306 0.241682 0.0768174 0.65197
Delete 44   T     0.842274 0.0546769 0.0694681 0.0335809
Main 44     T     0.714327 0.252009 0.0336642
            E     0.154578 0.592339 0.0414033 0.211681
Insert 45   T     0.0324632 0.265807 0.518604 0.0792141 0.103911
            E     0.0167578 0.238138 0.391434 0.353669
Delete 45   T     0.702138 0.0874846 0.167107 0.04327
Main  45    T     0.417877 0.538792 0.0433309
            E     0.410457 0.0625806 0.485064 0.0418986
Insert 46   T     0.130119 0.297775 0.0797601 0.41047 0.0818759
            E     0.682985 0.196369 0.032749 0.0878966
Delete 46   T     0.788522 0.103423 0.069749 0.0383063
Main 46     T     0.777633 0.147544 0.0748224
            E     0.205505 0.126244 0.283653 0.384598
Insert 47   T     0.0283527 0.0276021 0.804174 0.0480839 0.0917876
            E     0.330549 0.0289172 0.0846797 0.555854
Delete 47   T     0.446751 0.139792 0.373745 0.0397123
Main 47     T     0.163218 0.334829 0.501952
            E     0.0477351 0.603161 0.0749704 0.274133
Insert 48   T     0.021461 0.284493 0.617397 0.0219144 0.0547347
            E     0.0775709 0.386039 0.191662 0.344728
Delete 48   T     0.117622 0.251791 0.604896 0.0256898
Main 48     T     0.151525 0.374867 0.473608
            E     0.0165902 0.127673 0.607682 0.248054
Insert 49   T     0.0312855 0.336288 0.0755482 0.530802 0.0260759
            E     0.956783 0.011012 0.0145341 0.0176707
Delete 49   T     0.243168 0.652416 0.0489677 0.0554485
Main 49     T     0.208006 0.264275 0.527719
            E     0.0478418 0.00962673 0.811512 0.131019
Insert 50   T     0.0561317 0.0294624 0.287353 0.552503 0.0745496
            E     0.224793 0.420592 0.122715 0.2319
Delete 50   T     0.177273 0.0270355 0.778266 0.0174256
Main 50     T     0.845342 0.08009 0.074568
            E     0.122256 0.602654 0.232209 0.0428796
Insert 51   T     0.0620942 0.182631 0.105778 0.581483 0.0680137
            E     0.0389145 0.937166 0.0117076 0.0122116
Delete 51   T     0.66783 0.105198 0.179403 0.0475684
Main 51     T     0.320393 0.622357 0.0572499
            E     0.164756 0.0832648 0.0117098 0.740269
Insert 52   T     0.374033 0.151728 0.0661652 0.300772 0.107302
            E     0.164712 0.0366047 0.762635 0.0360479
Delete 52   T     0.768257 0.125493 0.0615356 0.0447144
Main 52     T     0.853423 0.0900299 0.056547
            E     0.392491 0.13457 0.303078 0.169861
Insert 53   T     0.0320174 0.0270157 0.00547351 0.929941 0.00555256
            E     0.751431 0.192671 0.025209 0.0306884
Delete 53   T     0.498688 0.30308 0.0575877 0.140644
Main 53     T     0.274576 0.0343908 0.691033
            E     0.00365128 0.418647 0.0268248 0.550877
Insert 54   T     0.14538 0.164787 0.169402 0.254606 0.265824
            E     0.789549 0.0630517 0.0114598 0.135939
Delete 54   T     0.313728 0.175816 0.422489 0.0879673
```

```
Main 54     T     0.227067 0.156005 0.616927
            E     0.00688872 0.152124 0.776842 0.0641448
Insert 55   T     0.0192337 0.242141 0.0574522 0.65169 0.0294828
            E     0.5559 0.044951 0.0314262 0.367723
Delete 55   T     0.193515 0.689395 0.0687257 0.0483643
Main 55     T     0.22893 0.492345 0.278724
            E     0.0279704 0.0269926 0.893093 0.0519436
Insert 56   T     0.351865 0.0252386 0.00624749 0.607461 0.00918738
            E     0.0390107 0.18843 0.0906607 0.681898
Delete 56   T     0.400571 0.434558 0.12723 0.0376407
Main 56     T     0.362392 0.0361541 0.601454
            E     0.176912 0.75728 0.0600852 0.00572331
Insert 57   T     0.0097703 0.020558 0.626229 0.0179544 0.325488
            E     0.0102217 0.377262 0.0248861 0.58763
Delete 57   T     0.428837 0.269079 0.21944 0.0826436
Main 57     T     0.120019 0.845522 0.0344585
            E     0.273195 0.180919 0.00809832 0.537788
Insert 58   T     0.404325 0.0887705 0.296574 0.0455527 0.164778
            E     0.00314588 0.00541648 0.98335 0.0080877
Delete 58   T     0.426572 0.09812 0.443251 0.0320568
Main 58     T     0.0817514 0.887414 0.0308343
            E     0.422089 0.0111029 0.389977 0.176831
Insert 59   T     0.126382 0.0416057 0.683179 0.0983516 0.0504815
            E     0.265891 0.310125 0.0832387 0.340745
Delete 59   T     0.725132 0.117784 0.104857 0.0522271
Main 59     T     0.310531 0.635991 0.0534777
            E     0.387961 0.337832 0.238089 0.0361184
Insert 60   T     0.108756 0.275321 0.0495088 0.51953 0.0468851
            E     0.680046 0.255434 0.0546569 0.0098633
Delete 60   T     0.887398 0.0414867 0.0439153 0.0272
Main 60     T     0.480498 0.0212313 0.498271
            E     0.030435 0.49723 0.00636178 0.465973
Insert 61   T     0.0105638 0.300127 0.0449509 0.0384061 0.605952
            E     0.704042 0.101705 0.164585 0.0296676
Delete 61   T     0.42636 0.242392 0.155562 0.175686
Main 61     T     0.101174 0.780747 0.118078
            E     0.0579556 0.0477581 0.768846 0.125441
Insert 62   T     0.353679 0.293299 0.0693154 0.199812 0.0838942
            E     0.0774887 0.267153 0.52618 0.129178
Delete 62   T     0.124048 0.393065 0.383825 0.0990621
Main 62     T     0.873434 0.106553 0.0200128
            E     0.0540258 0.832896 0.027232 0.0858459
Insert 63   T     0.188005 0.354593 0.120604 0.299454 0.037344
            E     0.159263 0.023198 0.063708 0.753831
Delete 63   T     0.621076 0.0606075 0.248349 0.0699676
Main 63     T     0.656699 0.20663 0.136671
            E     0.265997 0.0117285 0.614411 0.107864
Insert 64   T     0.158269 0.0844609 0.0880674 0.625717 0.0434858
            E     0.437869 0.173087 0.360311 0.0287332
Delete 64   T     0.300325 0.11053 0.542882 0.0462626
Main 64     T     0.588089 0.030103 0.381808
            E     0.0643644 0.336006 0.419625 0.180005
Insert 65   T     0.0412989 0.022169 0.0391983 0.87949 0.017844
            E     0.434667 0.206505 0.0108411 0.347987
Delete 65   T     0.453773 0.21533 0.297044 0.0338541
Main 65     T     0.211862 0.0600992 0.728038
            E     0.00688171 0.891917 0.00557486 0.095626
Insert 66   T     0.383106 0.0222219 0.00849841 0.578889 0.00728477
            E     0.265115 0.011577 0.0142243 0.709084
Delete 66   T     0.285611 0.438108 0.0532985 0.222983
Main 66     T     0.676809 0.230584 0.0926064
```

```
            E     0.0729196 0.00291023 0.913909 0.0102609
Insert 67   T     0.043728 0.062549 0.553568 0.0185371 0.321618
            E     0.00839785 0.856992 0.0231696 0.11144
Delete 67   T     0.539263 0.260848 0.125147 0.0747419
Main 67     T     0.204847 0.171012 0.624141
            E     0.600064 0.0536432 0.289327 0.0569653
Insert 68   T     0.0258954 0.514574 0.0159453 0.0371129 0.406473
            E     0.93717 0.0484737 0.00625689 0.00809895
Delete 68   T     0.241111 0.354576 0.361361 0.0429516
Main 68     T     0.844776 0.112725 0.0424993
            E     0.107218 0.0997789 0.0128598 0.780143
Insert 69   T     0.0675912 0.287509 0.0250962 0.092653 0.527151
            E     0.184744 0.143963 0.0353047 0.635989
Delete 69   T     0.0435997 0.403607 0.506418 0.0463758
Main 69     T     0.415981 0.417729 0.16629
            E     0.0106471 0.0417059 0.892997 0.0546501
Insert 70   T     0.339352 0.17885 0.0678808 0.0530858 0.360831
            E     0.0259577 0.815105 0.0953577 0.0635793
Delete 70   T     0.0292229 0.0801144 0.865779 0.0248833
Main 70     T     0.736442 0.12388 0.139678
            E     0.261289 0.0253571 0.700022 0.0133315
Insert 71   T     0.0564513 0.118077 0.69652 0.0801757 0.0487766
            E     0.408209 0.264801 0.0366281 0.290362
Delete 71   T     0.116207 0.771297 0.0841078 0.0283883
Main 71     T     0.013104 0.151217 0.835679
            E     0.376559 0.596494 0.00519077 0.021756
Insert 72   T     0.156809 0.146551 0.148477 0.209095 0.339068
            E     0.708654 0.21619 0.0549354 0.0202204
Delete 72   T     0.0199731 0.387098 0.553044 0.0398854
Main 72     T     0.901141 0.0169903 0.0818691
            E     0.175525 0.113606 0.0187393 0.69213
Insert 73   T     0.0611818 0.0678764 0.481409 0.167573 0.22196
            E     0.584042 0.294295 0.102704 0.0189586
Delete 73   T     0.355866 0.0932691 0.220583 0.330282
Main 73     T     0.809168 0.137434 0.0533988
            E     0.0068553 0.216034 0.617602 0.159508
Insert 74   T     0.0547462 0.405417 0.105404 0.151947 0.282486
            E     0.910234 0.0516406 0.0103282 0.0277976
Delete 74   T     0.0919392 0.0212328 0.865668 0.0211601
Main 74     T     0.265084 0.269618 0.465298
            E     0.221347 0.157592 0.571253 0.049808
Insert 75   T     0.0135697 0.0337813 0.0182981 0.926929 0.00742231
            E     0.0309716 0.201807 0.198352 0.568869
Delete 75   T     0.0476942 0.0102375 0.013304 0.928764
Main 75     T     0.850946 0.136167 0.0128872
            E     0.203124 0.557341 0.00902116 0.230514
Insert 76   T     0.442466 0.0882017 0.0116167 0.232737 0.224978
            E     0.0101453 0.359143 0.126076 0.504636
Delete 76   T     0.0150938 0.674843 0.0195126 0.290551
Main 76     T     0.283796 0.573203 0.143002
            E     0.228561 0.0155482 0.743249 0.0126413
Insert 77   T     0.0292474 0.0174168 0.050301 0.882767 0.0202675
            E     0.246293 0.337321 0.188904 0.227482
Delete 77   T     0.016478 0.00745905 0.967265 0.00879793
Main 77     T     0.970115 0.0192498 0.0106354
            E     0.430459 0.139765 0.30252 0.127256
Insert 78   T     0.00851613 0.0408776 0.0521616 0.884162 0.0142828
            E     0.0780065 0.167341 0.115642 0.639011
Delete 78   T     0.717627 0.0588279 0.145736 0.0778096
Main 78     T     0.967016 0.0160194 0.0169649
            E     0.232681 0.615218 0.0514099 0.100691
```

```
Insert 79   T     0.909621 0.00688882 0.0174875 0.0479736 0.0180291
            E     0.00868577 0.211381 0.496232 0.283701
Delete 79   T     0.62503 0.0912686 0.151082 0.13262
Main 79     T     0.641255 0.325616 0.033129
            E     0.310074 0.193055 0.375862 0.121008
Insert 80   T     0.00676821 0.0108651 0.614398 0.347483 0.0204854
            E     0.760695 0.100863 0.0694198 0.0690218
Delete 80   T     0.494259 0.148603 0.0634713 0.293666
Main 80     T     0.516524 0.025547 0.457929
            E     0.0330718 0.244652 0.20562 0.516655
Insert 81   T     0.0349682 0.229849 0.356456 0.0222697 0.356456
            E     0.953607 0.0117845 0.00664611 0.027962
Delete 81   T     0.244769 0.0948092 0.565612 0.0948092
Main 81     T     0.469862 0.0602755 0.469862
            E     0.0229882 0.453568 0.263012 0.260432
Insert 82   T     0.598272 0.401728
            E     0.121496 0.283404 0.398068 0.197032
Delete 82   T     1.0

End
--------------------------------------------------------------------------------
```