

# Identificação de sítios de ligação de fatores de transcrição com a integração de dados epigenéticos



Eduardo Gade Gusmão  
Centro de Informática  
Universidade Federal de Pernambuco

Dissertação de Mestrado  
*Ciência da Computação*

Outubro 2012

# **Identificação de sítios de ligação de fatores de transcrição com a integração de dados epigenéticos**

*Dissertação apresentada ao Centro de Informática da Universidade Federal de Pernambuco, como parte dos requisitos necessários para obtenção do título de Mestre em Ciência da Computação.*

Eduardo Gade Gusmão  
Centro de Informática  
Universidade Federal de Pernambuco

Orientador  
Ivan Gesteira Costa Filho

Co-orientador  
Marcilio Carlos Pereira de Souto

Dissertação de Mestrado  
*Ciência da Computação*

Outubro 2012

Dissertação submetida ao corpo docente do programa de pós-graduação do Centro de Informática da Universidade Federal de Pernambuco como parte dos requisitos necessários para obtenção do grau de mestre em Ciência da Computação.

Aprovado: \_\_\_\_\_

Katia Silva Guimarães – Centro de Informática - UFPE

---

Ana Maria Benko Iseppon – Departamento de Genética - UFPE

---

Paulo Gustavo Soares da Fonseca – Centro de Informática - UFPE

**IDENTIFICAÇÃO DE SÍTIOS DE LIGAÇÃO DE FATORES DE  
TRANSCRIÇÃO COM A INTEGRAÇÃO DE DADOS EPIGENÉTICOS.**

**Por**

*Eduardo Gade Gusmão*

UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

Cidade Universitária – Tels. (81) 2126-8414 – Fax: (81) 2126-8410.

RECIFE – BRASIL

Outubro – 2012

## Agradecimentos

Agradeço primeiramente à minha família, em especial à minha mãe Christiani Gade Gusmão, por fornecer todo o apoio necessário, permitindo dar continuidade a este projeto de pesquisa. Desde momentos necessários de lazer até pedidos de revisão gramatical não remunerados, eles sempre estiveram presentes e fizeram toda a diferença.

Agradeço ao meu orientador Dr. Ivan Gesteira Costa Filho pelos ensinamentos, sugestões, dicas e ajudas. Seu interesse em meu trabalho e disponibilidade para sanar dúvidas foram cruciais para a completude deste estudo. Agradeço também ao meu co-orientador Dr. Marcílio Carlos Pereira de Souto e à Dra. Thaís Gaudencio do Rêgo pelos ensinamentos paralelos e expansão da minha visão sobre a área da bioinformática. Também reservo um agradecimento especial ao Dr. Christoph Dieterich por contribuição no desenho experimental do trabalho; e aos membros da banca, Dra. Katia Silva Guimarães, Dra. Ana Maria Benko Iseppon e Dr. Paulo Gustavo Soares da Fonseca, que me deram a honra de poder mostrar o trabalho realizado. Agradeço também aos professores e funcionários do Centro de Informática, que me passaram valiosos ensinamentos e trabalharam para manter uma estrutura digna de um centro de referência em computação.

Agradeço às instituições FACEPE, CNPq e CAPES. Em especial à FACEPE, pelo auxílio financeiro na forma de bolsa de mestrado. Ao CNPq e à CAPES, pelos auxílios financeiros relativos à infra-estrutura.

Gostaria também de agradecer aos meus colegas Gilderlânio Santana de Araújo, Paulo Ricardo da Silva Soares, Felipe Kühner Câmara dos Santos, Nelson Gutemberg Rocha da Silva, João Rufino da Costa Neto, Yane Wanderley dos Santos, Diogo da Silva Severo, Everson Veríssimo da Silva, Arthur Felipe Melo Alvim, Flávia Roberta Barbosa de Araújo, Pablo Andretta Jaskowiak, André Kunio de Oliveira Tiba, Luciano Soares de Souza, Rebecca Cristina Linhares de Carvalho e Kalil Araújo Bispo. Os debates em ambiente de trabalho ou momentos de lazer permitiram meu crescimento em todos os sentidos. Por fim, agradeço ao meu amigo e companheiro Eduardo Henrique Farias de Carvalho por me fornecer o suporte necessário para que eu conseguisse completar todas as etapas deste trabalho e do curso de pós graduação.

Dedico este trabalho à minha família, que me forneceu todo o apoio necessário para o meu crescimento em todos os aspectos e ao orientador Ivan Gesteira Costa Filho, por estar presente em todos os momentos de dúvida e incentivar meu interesse na carreira acadêmica.

## Resumo

A identificação de elementos cis-regulatórios no DNA é crucial para o entendimento das redes regulatórias que governam diversos mecanismos celulares tais como diferenciação celular, desenvolvimento ou apoptose. Entretanto, essa tarefa é bastante complexa, dada a grande quantidade de diferentes fatores de transcrição no genoma humano. Atualmente, são estimados 1.500 fatores que podem se ligar, diretamente ou indiretamente, em múltiplos loci genômicos. O procedimento computacional padrão para a detecção de tais regiões consiste no uso de matrizes de pontuação, que são representações probabilísticas da afinidade de ligação desses fatores em determinadas sequências de DNA. Porém tal abordagem resulta em um grande número de falsos positivos pelo fato de não ser possível distinguir entre regiões ativas e inativas e pelos motivos estruturais serem pequenos e degenerados. Esses problemas têm sido superados através da consideração de características epigenéticas. A ideia básica é que algumas regiões da cromatina encontram-se densamente empacotadas em uma estrutura fechada, não permitindo ligação de proteínas reguladoras; enquanto outros sítios estão menos empacotados (cromatina descondensada), permitindo tais ligações. Pesquisas atuais mostram que fontes de dados capazes de sinalizar tais regiões descondensadas, tais como digestão de DNase I (obtida através de DNase-seq) e modificações de histonas (obtidas através de ChIP-seq), podem melhorar a detecção de sítios de ligação dos fatores de transcrição.

Neste trabalho, é proposta a construção de um modelo escondido de Markov contínuo bivariado com objetivo de integrar fontes de dados epigenéticas para avaliar se há melhora nos resultados, em relação à predições realizadas com o método computacional padrão ou através da utilização de fontes de dados epigenéticas de forma individual. Além disso, uma nova forma de estimativa de parâmetros para tal modelo foi desenvolvida, removendo a necessidade de se realizar procedimentos tradicionais custosos. Foi observado que o modelo proposto melhora significativamente a sensibilidade, com pouco ou nenhum efeito negativo na especificidade, em comparação com modelos existentes baseados em cromatina descondensada apenas.

**Palavras-chave:** Sítios de Ligação de Fatores de Transcrição; DNase-seq; ChIP-seq; Modificações de Histonas; Modelos Escondidos de Markov.

## Abstract

The identification of cis-regulatory elements on DNA is crucial for the understanding of the complex regulatory networks that orchestrate diverse cell mechanisms such as differentiation, development and apoptosis. However, this task is very complex, given the great number of different transcription factors in the human genome. Currently, it is believed that there are around 1,500 factors, each of which can bind directly or indirectly to multiple loci. The standard computational approach for the detection of such regions consists in using Position Weight Matrices, which are probabilistic representations of the factor's binding affinities, to search the genome for regions likely to be binding sites. However, such approach results in a very high number of false positive hits, since it cannot distinguish between active / inactive binding sites and also because motifs are usually small and degenerate. To overcome these problems, recent techniques are being based on epigenetic features. The main idea is that some regions of the chromatin are densely packed in a closed structure, preventing the binding of regulatory proteins, while other regions are less packed (open chromatin), allowing such binding. Current research shows that data sources that are capable of signaling open regions, such as DNase I digestion (obtained by DNase-seq) and histone modifications (obtained by ChIP-seq) can improve transcription factor binding sites prediction.

In this work, a continuous bivariate hidden Markov model is proposed which is capable of integrating epigenetic data sources, in order to evaluate if the results can be improved when compared to standard computational approaches or to single data source approaches. Besides that, a novel technique to estimate the parameters of the model was developed, making costly traditional procedures no longer necessary. It was observed that the proposed model significantly improves the sensitivity with low or no negative effect on the specificity when compared to open chromatin-only models.

**Keywords:** Transcription Factor Binding Sites; DNase-seq; ChIP-seq; Histone Modifications; Hidden Markov Models.

# Sumário

<b>Lista de Figuras</b>	<b>x</b>
<b>Lista de Tabelas</b>	<b>xi</b>
<b>Glossário</b>	<b>xii</b>
<b>1 Introdução</b>	<b>1</b>
1.1 Motivação . . . . .	1
1.2 Contribuições . . . . .	4
1.3 Estrutura do Documento . . . . .	5
<b>2 Contextualização Biológica</b>	<b>7</b>
2.1 Conceitos Básicos em Biologia Molecular . . . . .	8
2.1.1 DNA e RNA . . . . .	9
2.1.2 Proteínas . . . . .	10
2.1.3 Estrutura da Cromatina . . . . .	12
2.1.4 Dogma Central da Biologia Molecular . . . . .	13
2.2 Regulação Gênica em Eucariotos . . . . .	18
2.2.1 Maquinaria Regulatória Proximal . . . . .	19
2.2.2 Elementos Regulatórios Transcpcionais . . . . .	21
2.2.2.1 Núcleo do Promotor . . . . .	21
2.2.2.2 Elementos Promotores Proximais . . . . .	22
2.2.2.3 Amplificadores . . . . .	23
2.2.2.4 Silenciadores . . . . .	23
2.2.2.5 Insuladores . . . . .	24
2.2.2.6 Regiões de Controle de Locus . . . . .	24
2.3 Identificação de Sítios de Ligação de Fatores de Transcrição . . . . .	24
2.3.1 DNase I Footprinting . . . . .	26
2.3.2 Imunoprecipitação da Cromatina . . . . .	27
2.3.3 Motif Matching . . . . .	29
2.4 Solução Epigenética . . . . .	32

2.4.1	Conceitos e Elementos Epigenéticos	33
2.4.2	Métodos de Obtenção de Dados Epigenéticos	35
2.4.3	Geração de Sinais	37
2.5	Revisão da Literatura	39
2.6	Considerações Finais	40
<b>3</b>	<b>Modelos Escondidos de Markov</b>	<b>42</b>
3.1	Modelos Escondidos de Markov	43
3.2	Métodos de Predição Baseados em HMMs	45
3.2.1	Algoritmo de Viterbi	46
3.2.2	Probabilidade Posterior	48
3.3	Estimação de Parâmetros em HMMs	50
3.4	Considerações Finais	52
<b>4</b>	<b>Metodologia</b>	<b>53</b>
4.1	Bases de Dados	54
4.2	Motif Matching	57
4.3	Análises de Enriquecimento	57
4.4	Processamento dos Sinais Epigenéticos	58
4.5	Footprinting com HMMs	60
4.6	Estimação de Parâmetros e Aplicação dos HMMs	62
4.7	Gold Standard	64
4.8	Considerações Finais	65
<b>5</b>	<b>Resultados e Discussão</b>	<b>66</b>
5.1	Análise dos Sinais Epigenéticos	67
5.2	Acurácia do Modelo Proposto	76
5.3	Tempo de Execução e Armazenamento	85
5.4	Considerações Finais	87
<b>6</b>	<b>Conclusão</b>	<b>88</b>
6.1	Objetivos Atingidos	88
6.2	Dificuldades e Limitações de Escopo	89
6.3	Trabalhos Futuros	89
<b>Referências</b>		<b>91</b>

# Listas de Figuras

2.1	Célula eucariótica animal . . . . .	9
2.2	Estrutura do DNA . . . . .	10
2.3	Comparação entre as estruturas moleculares da proteína e do DNA . . . . .	11
2.4	Visão global da estrutura da cromatina . . . . .	13
2.5	Dogma central da Biologia Molecular . . . . .	14
2.6	Etapas do processo de transcrição . . . . .	15
2.7	Diferentes tipos de elementos cis-atuantes . . . . .	19
2.8	Maquinaria transcrecional eucariótica . . . . .	20
2.9	Funcionamento dos elementos regulatórios distais . . . . .	22
2.10	Esquema do método <i>DNase I Footprinting</i> . . . . .	28
2.11	Esquema do método ChIP . . . . .	29
2.12	Método para gerar PWMs . . . . .	30
2.13	Elementos epigenéticos . . . . .	34
2.14	Modificações de histonas . . . . .	35
2.15	Geração de Sinais Genômicos . . . . .	38
3.1	Esquema de um modelo escondido de Markov . . . . .	44
4.1	Fases do processo experimental . . . . .	54
4.2	Modificação dos sinais ao longo do processamento . . . . .	60
4.3	HMM que utiliza dados de DNase-seq apenas . . . . .	61
4.4	Modelagem do HMM e exemplo de aplicação . . . . .	62
5.1	Análise das melhores regiões de MPBS para o CTCF . . . . .	68
5.2	Regiões de TFBS com e sem evidência de ChIP-seq Pt.1 . . . . .	70
5.3	Regiões de TFBS com e sem evidência de ChIP-seq Pt.2 . . . . .	71
5.4	Regiões de TFBS com e sem evidência de ChIP-seq e <i>footprint</i> associado Pt. 1 .	73
5.5	Regiões de TFBS com e sem evidência de ChIP-seq e <i>footprint</i> associado Pt. 2 .	74
5.6	Exemplo de uma região com resultados melhorados pelo modelo proposto . . . . .	84
5.7	Exemplo do problema das previsões amplas . . . . .	85

# **Lista de Tabelas**

2.1	Impacto das modificações de histonas na estrutura da cromatina e expressão gênica	36
4.1	Fontes dos dados	55
4.2	Sinais epigenéticos e fatores estudados	56
5.1	Quantidade de <i>footprints</i> encontrados com cada modelo	76
5.2	Resultados para o fator ATF3	78
5.3	Resultados para o fator CTCF	78
5.4	Resultados para o fator CTCF	79
5.5	Resultados para o fator GABP	79
5.6	Resultados para o fator GABP	80
5.7	Resultados para o fator REST	80
5.8	Resultados para o fator REST	81
5.9	Comparação da sensibilidade e especificidade entre o modelo prévio e o proposto	82
5.10	Tempo de execução e memória	86
5.11	Espaço necessário para armazenamento	86

# Glossário

<b>Acetilação</b>	Reação que introduz um grupo funcional acetila em um composto orgânico.	massivo); técnica biológica para identificar regiões genômicas onde uma proteína de interesse está ligada através da realização de ChIP seguida de sequenciamento massivo dos fragmentos genômicos recuperados.
<b>ATP</b>	Adenosina Trifosfato; nucleotídeo responsável pelo armazenamento de energia em suas ligações químicas, utilizado em reações que exigem tal energia.	<b>DCE</b> Downstream Core Elements; elemento presente no núcleo do promotor de alguns genes relacionado com a formação do complexo pré-iniciação.
<b>Biopython</b>	Conjunto de bibliotecas para a linguagem Python contendo implementações de diversas ferramentas biológicas necessárias em várias áreas de bioinformática.	<b>DNase-chip</b> DNase I digestion followed by chip (digestão por DNase I seguida de chip); técnica biológica para identificação de regiões de cromatina descondensada através da clivagem do DNA com a endonuclease DNase I seguida de experimentos com <i>tiling arrays</i> .
<b>bp</b>	Base Pair (pares de bases); representa um par de bases (nucleotídeos) no DNA, isto é, uma coordenada genômica.	<b>DNase-seq</b> DNase I digestion followed by massive sequencing (digestão por DNase I seguida de sequenciamento massivo); técnica biológica para identificação de regiões de cromatina descondensada através da clivagem do DNA com a endonuclease DNase I seguida de sequenciamento massivo dos fragmentos genômicos recuperados.
<b>BRE</b>	TFIIB-Recognition Element; elemento presente no núcleo do promotor de alguns genes relacionado com a formação do complexo pré-iniciação.	<b>DNase I</b> Desoxirribonuclease I; endonuclease codificada pelo gene <i>DNASE1</i> capaz de clivar o DNA em várias diferentes condições.
<b>ChIP</b>	Chromatin Immunoprecipitation (imunoprecipitação da cromatina); técnica biológica para recuperar regiões genômicas onde uma proteína de interesse está ligada, através da imunoprecipitação da mesma utilizando um anticorpo (ou outros materiais).	<b>dNTP</b> desoxirribonucleotídeo trifosfato; monômero do DNA em seu formato com três grupos fosfato, necessários para produzir a energia suficiente para a interação com a macromolécula de DNA.
<b>ChIP-chip</b>	Chromatin Immunoprecipitation followed by chip (imunoprecipitação da cromatina seguida de chip); técnica biológica para identificar regiões genômicas onde uma proteína de interesse está ligada através da realização de ChIP seguida de experimentos com <i>tiling arrays</i> .	<b>EDTA</b> Ethylenediamine Tetraacetic Acid (ácido etilenodiamino tetra-acético); composto orgânico que age como agente quelante, formando complexos muito estáveis com diversos íons metálicos. Das várias utilizações destaca-se o controle em experimentos de ChIP.
<b>ChIP-seq</b>	Chromatin Immunoprecipitation followed by massive sequencing (imunoprecipitação da cromatina seguida de sequenciamento	<b>EM</b> Expectation Maximization (maximização da esperança); Algoritmo iterativo com objetivo de encontrar a estimativa de parâmetros de máxima verossimilhança

	utilizando dados sem rótulos (isto é, não se sabe a classe dos padrões).	
<b>ENCODE</b>	Encyclopedia of DNA Elements; Iniciativa dentro do programa <i>Genome Browser</i> da Universidade da Califórnia em Santa Cruz que disponibiliza diversas faixas de dados relativos à genômica funcional.	
<b>Endonuclease</b>	Classe de proteínas que clivam as ligações fosfodiéster dentro de uma cadeia de DNA.	
<b>FAIRE</b>	Formaldehyde-Assisted Identification of Regulatory Elements; técnica biológica para identificação de regiões de cromatina descondensada através de um protocolo menos denso do que o do DNase-seq.	
<b>FMR1</b>	Fragile X Mental Retardation 1; gene responsável pela codificação da proteína FMRP, comumente encontrada no cérebro e essencial para o desenvolvimento cognitivo e reprodução em fêmeas.	
<b>Fosforilação</b>	Reação que introduz um grupo funcional fosfato em um composto orgânico.	
<b>GHMM</b>	General Hidden Markov Model Library; Biblioteca disponível em C e em Python que implementa de forma eficiente HMMs com emissões discretas ou contínuas.	
<b>GTF</b>	General Transcription Factors (fatores de transcrição gerais); conjunto de proteínas que, junto com a RNA polimerase e o mediador, constituem o aparato básico para que a transcrição ocorra em nível basal em eucariotos.	
<b>HMM</b>	Hidden Markov Model (modelo escondido de Markov); técnica para modelagem estatística de séries temporais baseada em processos estocásticos de Markov.	
<b>HS</b>	DNase I Hypersensitive Sites (sítios hiper-sensíveis à DNase I); regiões no DNA que permitem a clivagem através da endonuclease DNase I.	
<b>Inr</b>	Elemento Iniciador; elemento presente no núcleo do promotor de alguns genes relacionado com a formação do complexo pré-iniciação.	
<b>LCR</b>	Locus Control Regions; região composta por vários elementos cis-atuantes distais cuja composição representa a sua funcionalidade regulatória.	
<b>MACS</b>	Model-based Analysis for ChIP-Seq; Ferramenta utilizada para analisar (processar e encontrar picos) dados de ChIP-seq.	
<b>Metilação</b>	Reação que introduz um grupo funcional metila em um composto orgânico.	
<b>Microarray</b>	Microarranjo; técnica experimental para medir níveis de expressão gênica (ou alguns outros atributos) que utiliza um chip que contém diversos fragmentos de DNA que representam regiões de interesse (genes ou exons, por exemplo).	
<b>MM</b>	Motif Matching; técnica computacional que utiliza representações probabilísticas de <i>motifs</i> (PFMs, PSSMs ou PWMs) para atribuir um grau de afinidade para regiões genômicas a respeito da probabilidade de um fator de transcrição se ligar àquela região	
<b>Motif</b>	padrão frequente ou assinatura; sequência genômica ou proteômica com padrão reconhecível e que tenha significado biológico.	
<b>MPBS</b>	Motif Predicted Binding Sites (sítios de ligação preditos através de <i>motifs</i> ); termo utilizado para referenciar sítios de ligação de fatores de transcrição preditos através de <i>motif matching</i> .	
<b>MPSS</b>	Massively Parallel Signature Sequencing; abordagem utilizada para identificar e quantificar transcritos de mRNA presentes em uma amostra.	
<b>MTE</b>	Motif Ten Element; elemento presente no núcleo do promotor de alguns genes relacionado com a formação do complexo pré-iniciação.	
<b>PCR</b>	Polymerase Chain Reaction (reação em cadeia da polimerase); método de amplificação (de criação de múltiplas cópias) de DNA.	

<b>PFM</b>	Position Frequency Matrix (matriz de frequência de posição); representação matricial de um <i>motif</i> onde as linhas representam os nucleotídeos e as colunas representam as posições do <i>motif</i> .	pontuais (em apenas um nucleotídeo) no genoma.
<b>PIC</b>	Transcription Preinitiation Complex (complexo pré-iniciação de transcrição); complexo de proteínas montados na região promotora necessárias para a transcrição.	
<b>PSSM</b>	Position Specific Scoring Matrix (matrizes de pontuação específica por posição); neste trabalho está sendo utilizado como sinônimo de PWM.	
<b>PWM</b>	Position Weight Matrix (matrizes de peso de posição); representação matricial logarítmica de um <i>motif</i> criada através de uma PFM.	
<b>Python</b>	Linguagem de programação de alto nível, interpretada, imperativa, orientada a objetos, de tipagem dinâmica e forte. Utilizada para analisar os dados, aplicar os métodos e gerar os gráficos em todo o projeto.	Transcription Factor (fator de transcrição); elementos regulatórios transatuantes. São proteínas que se ligam em regiões específicas no genoma para regular a transcrição de um ou mais genes.
<b>rNTP</b>	ribonucleotídeo trifosfato; monômero do RNA em seu formato com três grupos fosfato, necessários para produzir a energia suficiente para a interação com a macromolécula de RNA.	Transcription Factor Binding Site (sítio de ligação de fatores de transcrição); elementos regulatórios cis-atuantes. São as regiões onde os fatores de transcrição se ligam.
<b>SNP</b>	Single Nucleotide Polymorphism (polimorfismos de único nucleotídeo); variações	Tiling array técnica experimental semelhante ao microarranjo, porém neste caso os fragmentos de DNA no chip representam regiões contíguas no genoma dada uma janela e deslocamento específico.
		<b>TSS</b> Transcription Start Site (sítios de início de transcrição); sítio onde a transcrição se inicia.
		<b>Ubiquitinação</b> Marcação através de moléculas ubiquitina.

# 1

## Introdução

### 1.1 Motivação

Em outubro de 1990, iniciou-se o chamado *Projeto Genoma Humano* com o objetivo, na época extraordinário, de sequenciar o genoma humano completo. Dessa época até os dias de hoje, as tecnologias de sequenciamento avançaram de forma muito rápida. Para se ter uma ideia, em Setembro de 2001 o custo para sequenciar 1Mb de sequência de DNA era cerca de \$5.300,00 (totalizando aproximadamente \$95.300.000,00 por genoma humano); Enquanto em Julho de 2011 o custo para 1Mb era \$0,12 (fazendo um total aproximado de \$10.500,00 por genoma humano) [DNA Sequencing Consortiums, 2012]. O *Projeto Genoma Humano* levou 13 anos para ser completado, porém hoje em dia somos capazes de sequenciar o genoma humano completo com cerca de 3,194 bilhões de pares de bases (bp, do Inglês *Base Pairs*) em apenas três dias.

Há algum tempo atrás, achava-se que, de posse do genoma completo de um dado organismo, se poderia determinar com exatidão seu fenótipo, sua suscetibilidade a doenças, fornecer diagnósticos com alta precisão e que os tratamentos para doenças complexas como o câncer evoluiriam a ponto de curarem a maior parte das ocorrências. Porém percebeu-se que a simples definição da sequência de nucleotídeos que compõem o genoma não é suficiente para explicar os diversos processos regulatórios e metabólicos que ocorrem nos organismos dos seres vivos. Tais processos fazem parte de uma complexa cadeia de eventos que podem sim ocorrer no nível genômico e regulatório: transcripcional, pós-transcricional, traducional ou pós-traducional.

A execução correta dos processos biológicos tais como desenvolvimento, proliferação, envelhecimento, diferenciação e apoptose requer um conjunto de passos preciso e cuidadosamente orquestrado que depende da expressão espacial e temporal dos genes apropriada. Isso resulta no fato de que a desregulação da expressão gênica muitas vezes é relacionada a doenças [Rosenblom *et al.*, 2011]. Na era da pós-genômica, as atenções estão se voltando para o entendimento

## **1. INTRODUÇÃO**

---

de como os genes codificantes de proteínas (cerca de 20.000 – 25.000 em humanos) e seus produtos funcionam, principalmente sobre como seus padrões de expressão espacial e temporal são estabelecidos tanto no nível celular quanto considerando o organismo como um todo [Maston *et al.*, 2006].

Para entender esses mecanismos moleculares que governam os padrões de expressão gênica em uma escala global, é importante identificar os elementos regulatórios envolvidos nessas atividades. Exemplos desses componentes são elementos regulatórios trans-atuantes (ou fatores de transcrição (TFs, do Inglês *Transcription Factors*)), cis-atuantes (tais como silenciadores, amplificadores e insuladores) e fatores epigenéticos (tais como modificações de histonas, remodelamento da cromatina e metilação do DNA), cada um deles participando para que a expressão gênica ocorra de forma apropriada em processos biológicos específicos para cada célula, comuns entre alguns grupos de células ou ubíquos (presentes em todas as células do organismo) [Maston *et al.*, 2006; Rosenbloom *et al.*, 2011].

A identificação desses elementos, em especial os elementos regulatórios cis-atuantes nos quais os fatores de transcrição se ligam, pode ser uma tarefa bastante complexa, já que é estimado que existam mais que 1500 diferentes fatores de transcrição no genoma humano [Boyle *et al.*, 2011]. Além disso, sítios de ligação de fatores de transcrição (TFBSs, do Inglês *Transcription Factor Binding Sites*), com seus padrões frequentes ou assinaturas (em Inglês, *motifs*), são pequenos, com tamanhos geralmente variando entre 6 – 12 bp dos quais não mais que 4 – 6 bp ditam a especificidade da ligação [Maston *et al.*, 2006]. Além disso, apenas um subconjunto deles está ativo durante um determinado estado da célula, com os elementos deste subconjunto variando bastante entre diferentes tipos celulares [Cuellar-Partida *et al.*, 2012]. Também são fatores complicadores o fato de que vários fatores de transcrição têm múltiplos sítios de ligação possíveis (com diferentes *motifs*) e a existência de fatores que se ligam a DNA indiretamente, juntamente com outro fator ou complexo proteico [Alberts, 2007].

A abordagem computacional padrão para a identificação de TFBSs – *Motif Matching* (MM) – utiliza representações probabilísticas das afinidades dos sítios de ligação, seguido de um procedimento estatístico para detectar regiões genômicas com uma alta probabilidade de serem sítios de ligação para um fator em particular [Stormo, 2000]. Não obstante, *motif matching* é um método altamente sensível ao poder estatístico do algoritmo que está sendo utilizado para realizar tal procedimento e da qualidade da representação probabilística do *motif* utilizada. Várias desvantagens e impraticabilidades podem ser citadas como: (1) esse método é incapaz de distinguir entre regiões ativas e inativas; (2) os *motifs* geralmente são pequenos (sendo fácil encontrar por acaso regiões que não são sítios de ligação) ou degenerados (especificidade de ligação muito pequena) [Boyle *et al.*, 2011; Maston *et al.*, 2006]; (3) representações de *motifs* são difíceis de serem geradas e existe uma quantidade muito pequena de fatores com tais representações disponíveis em repositórios curados [Boyle *et al.*, 2011]; (4) a identificação de sítios

## **1.1. MOTIVAÇÃO**

---

de ligação de fatores que se ligam ao DNA de forma indireta é difícil, dado que eles não têm *motifs* bem definidos. A abordagem padrão para identificação de TFBSS são os experimentos de *DNase I Footprinting* utilizando DNase I como agente de clivagem, que é um método de alta acurácia e alta resolução [Gross & Garrard, 1988; Keene *et al.*, 1981]. Porém este método é altamente técnico e só consegue analisar < 1Kb por experimento o tornando impraticável em estudos pangenômicos (em Inglês, *genome-wide*), isto é, estudos cuja amplitude da análise é o genoma inteiro [Boyle *et al.*, 2011; Lodish *et al.*, 2007].

Novas tecnologias surgiram para suprir as dificuldades de aplicação dos métodos tradicionais. As principais técnicas para identificação de TFBSS atualmente são as abordagens baseadas em imunoprecipitação, seguidas de análises em *tiling arrays* (ChIP-chip) [Buck & Lieb, 2004] ou de sequenciamento em grande escala (ChIP-seq) [Park, 2009]. Porém tais técnicas são condicionais (específicas para as condições em que as células estão), falham para alguns fatores de transcrição em particular por motivos diversos e são experimentalmente e financeiramente custosas [Park, 2009]. O principal problema dessas técnicas está no fato de que elas fornecem um mapa geral (isto é, pangenômico) dos sítios de ligação apenas para um fator específico por experimento. Em estudos que analisam apenas um ou poucos destes fatores de transcrição, essas técnicas quase sempre são aplicadas por gerarem resultados com alta acurácia e boa resolução. Porém caso o objetivo seja criar um mapa de todos os possíveis sítios de ligação para uma célula num determinado momento, o número total de fatores de transcrição possíveis juntamente com o alto custo e dificuldades técnicas fazem com que ChIP-chip e ChIP-seq tenham pouco uso prático.

Tecnologias baseadas na junção de experimentos baseados em clivagem a partir da enzima de restrição DNase I com análises em *tiling arrays* (DNase-chip) [Crawford *et al.*, 2006a] ou sequenciamento em alta escala (DNase-seq) [Crawford *et al.*, 2004; Song & Crawford, 2010] estão se mostrando particularmente úteis para atingir o objetivo de caracterizar todos os sítios de ligação de uma determinada linha celular em escala genômica. Apesar da acurácia deste estudo ser extremamente dependente da técnica computacional e estatística associada à análise dos padrões de clivagem da DNase I, sua alta resolução está dando possibilidade a estudos bem sucedidos [Boyle *et al.*, 2008a, 2011; Crawford *et al.*, 2004, 2006b]. Está se tornando comum a utilização destas técnicas para gerar mapas, a nível genômico, de regiões de cromatina descondensada, em diversos tipos de células humanas expandindo nossos conhecimentos de diferenciação celular ou simplesmente aumentando a quantidade de elementos regulatórios com suporte de evidências [Song & Crawford, 2010]. Porém as técnicas baseadas em DNase I não fornecem a informação de quais são os fatores de transcrição que se ligam nos locais encontrados. Além disso, as técnicas estatísticas utilizadas estão atingindo um grau de complexidade bastante elevado e mostrando que ainda existem grandes quantidades de falsos positivos ou falsos negativos dependendo da situação [Boyle *et al.*, 2011; Cuellar-Partida *et al.*, 2012; Pique-Regi *et al.*, 2011].

## **1. INTRODUÇÃO**

---

Além do uso de métodos baseados em DNase I, pesquisas recentes têm focado na busca de padrões específicos de modificações pós-traducionais (tais como acetilação ou metilação) em proteínas chamadas histonas em diferentes tipos celulares e dados diversos padrões de expressão gênica. De fato, muitos desses estudos têm mostrado claros padrões (assinaturas) na cromatina e têm sugerido a aplicação destes resultados na identificação de elementos regulatórios [Barski *et al.*, 2007; Ernst & Kellis, 2010; Heintzman *et al.*, 2007; Hon *et al.*, 2009; Spivakov & Fisher, 2007]. Em particular, as modificações de histonas H3K4me2, H3K4me3, H3K9ac, H3K27ac e a histona variante H2A.Z são ótimos marcadores de regiões onde a cromatina se encontra em um estado menos enovelado (cromatina descondensada). Portanto, a presença destes marcadores epigenéticos é capaz de delimitar regiões ricas em sítios de ligação de elementos regulatórios [Barski *et al.*, 2007; Ramsey *et al.*, 2010; Schones & Zhao, 2008].

Alguns estudos atuais têm investigado a possibilidade de integração de diferentes metodologias biológicas como ChIP-seq ou ChIP-chip para padrões de histonas ou DNase-chip e DNase-seq com metodologias computacionais e probabilísticas, aplicadas diretamente ao contexto da identificação de elementos regulatórios [Cuellar-Partida *et al.*, 2012; Pique-Regi *et al.*, 2011; Won *et al.*, 2010]. Além disso, estudos que comparam diferentes padrões epigenéticos, fora de algum contexto específico, fornecem conceitos importantes que devem ser considerados durante a criação de uma metodologia aplicada a um problema específico [Shu *et al.*, 2011]

Cientistas estão entrando em consenso de que, no contexto de identificação de sítios de ligação para fatores de transcrição a nível genômico, abordagens que agregam diferentes tipos de informação atingem os objetivos de forma mais acurada e confiável do que a aplicação de técnicas individuais [Lassig, 2007]. Neste trabalho, várias fontes de dados epigenéticos provenientes de experimentos de identificação de cromatina descondensada com DNase-seq e modificações de histonas com ChIP-seq serão integradas utilizando uma abordagem probabilística baseada em modelos escondidos de Markov multivariados com emissões representando funções gaussianas. Para que os resultados sejam positivos, uma metodologia será claramente definida envolvendo o tratamento dos diferentes tipos de dados, implementação de técnicas especiais para que as cadeias de Markov não tenham problemas numéricos associados a dimensionalidade e quantidade de exemplos e verificação da acurácia do modelo sem nenhum tipo de viés resultante da aplicação das técnicas escolhidas.

### **1.2 Contribuições**

A contribuição deste projeto consiste na construção de um modelo escondido de Markov bivariado contínuo capaz de predizer sítios de ligação de fatores de transcrição em humanos. Este modelo será alimentado sempre com dados de cromatina descondensada e uma específica modificação de histona, de um conjunto maior de modificações. Para que este modelo seja construído,

### **1.3. ESTRUTURA DO DOCUMENTO**

---

análises dos padrões médios simples ao redor de regiões de TFBS experimentalmente determinadas serão realizadas. Determinados tais padrões, o modelo é construído, treinado (isto é, seus parâmetros são estimados) e testado. Com base em um conjunto de validação bem definido na literatura é possível avaliar tal modelo de forma eficaz.

Além da construção de um novo modelo capaz de integrar fontes de dados epigenéticas, um novo algoritmo de estimação de parâmetros será proposto. A motivação para a criação deste algoritmo está no fato de que os métodos presentes na literatura, na amplitude pesquisada, utilizavam dados provenientes da aplicação de técnicas biológicas custosas como base para o treinamento. De forma simples, estes conjuntos representam as informações biologicamente validadas a respeito de sítios de ligação. Constantemente tais dados eram obtidos em estudos mais antigos na literatura, resultando em conjuntos de treinamento pequenos e que, em vários casos, não correspondiam às regiões mais interessantes de se aplicar o treinamento. Portanto, o novo método de treinamento se baseia exclusivamente na aplicação de uma ferramenta computacional para avaliação de *motifs* chamada STAMP [Mahony & Benos, 2007].

Através da metodologia proposta, pretende-se verificar se modelos integrativos conseguem melhorar a acurácia em comparação a modelos que utilizam apenas cromatina descondensada como base preditiva, com base nos fatos: (1) Existem diversos locais observados com baixo sinal de digestão de DNase I porém com alta concentração de sítios de ligação ativos (falso negativos) e (2) algumas regiões hipersensíveis à DNase I não apresentam sítios de ligação (falso positivos). Nossa hipótese é que sinais de histonas, como uma fonte de dados adicionais, contribuirão para resolver algumas dessas ambiguidades.

## **1.3 Estrutura do Documento**

No capítulo seguinte serão realizadas as principais definições biológicas necessárias para o entendimento deste projeto de pesquisa. Após uma breve introdução revisando os conceitos básicos de biologia molecular (direcionado a leitores com embasamento puramente computacional), serão abordados temas como: regulação gênica, elementos regulatórios (*cis*-atuantes e *trans*-atuantes) e epigenética. Também serão revisados os principais métodos computacionais, estatísticos e biológicos que contêm alguma relação com a proposta deste trabalho. Finalmente, trabalhos relacionados serão brevemente descritos na última seção desse capítulo, traçando sempre um paralelo com a abordagem proposta.

O Capítulo 3 contém toda a formalização matemática do principal método utilizado neste trabalho: as cadeias escondidas de Markov. Após uma apresentação dos conceitos básicos de probabilidade e estatística, com objetivo principal de definir a nomenclatura utilizada, será

## **1. INTRODUÇÃO**

---

realizada uma introdução a este modelo probabilístico. Em sequência, são formalizados os métodos de predição e estimação de parâmetros utilizados neste projeto.

No Capítulo 4 serão definidos todos os procedimentos metodológicos realizados neste trabalho. Serão descritos os repositórios onde os dados foram obtidos, os métodos de busca genômica baseada em *motifs* (*motif matching*), os métodos de processamento dos sinais epigenéticos, as técnicas estatísticas de identificação de regiões enriquecidas de picos, a aplicação dos modelos probabilísticos e seu treinamento e a forma como a acurácia dos modelos foi aferida.

No Capítulo 5 todos os resultados serão exibidos. Tais resultados contêm descrições visuais do processamento dos sinais, resultados da aplicação dos modelos probabilísticos e tabelas contendo as acuráncias calculadas com base nos métodos estatísticos mais utilizados na literatura. Resultados serão exibidos tanto para o método proposto neste trabalho quanto para a replicação de métodos já existentes para efeito de comparação. Além disso, será realizada uma discussão a respeito dos resultados obtidos. Todos os pontos metodológicos e vieses são claramente exibidos para introduzir as asserções feitas com base nos resultados. Será mostrado que os modelos propostos conseguem superar modelos já existentes na literatura. Essa discussão tem o objetivo de motivar posteriores estudos com base na automatização de processos laboriosos, melhoramento das acuráncias observadas e construção de modelos mais complexos baseados na integração de múltiplos sinais epigenéticos.

Finalmente, no Capítulo 6, o trabalho é sumarizado. Os principais pontos serão destacados, incluindo as realizações e limitações dos modelos e técnicas propostos. Por fim, uma descrição detalhada da continuação deste trabalho é realizada, com destaque principal para o objetivo final: a construção de um modelo generalizado e capaz de integrar um número maior de sinais.

## 2

# Contextualização Biológica

Neste capítulo, serão descritos os conceitos biológicos necessários para o entendimento deste projeto de pesquisa. Em primeiro lugar, os conceitos básicos em Biologia Molecular serão apresentados. Tal apresentação será conduzida superficialmente, com objetivo único de suprir as necessidades do leitor não familiarizado com a área da Biologia Molecular. Explicações mais detalhadas a respeito de assuntos como Genética ou Biologia Molecular, podem ser encontradas em livros didáticos tais como [Alberts, 2007; Allis *et al.*, 2007; Lewin, 2003; Lodish *et al.*, 2007; Watson *et al.*, 2003].

A seguir, será realizada uma introdução ao conceito de regulação gênica em eucariotos. Posteriormente, o mecanismo regulatório será descrito em mais detalhes através da apresentação esquemática dos elementos que participam na transcrição de forma proximal e distal. Nesse momento, serão definidos os conceitos de elementos regulatórios *cis*- e *trans*- atuantes. Em seguida, o conceito de epigenética será detalhado e mais informações serão dadas a respeito de características epigenéticas exploradas neste trabalho como as modificações das histonas. Finalmente, serão exibidos os métodos biológicos mais importantes neste tema e serão mencionados alguns estudos que fazem parte do estado da arte da identificação de sítios de ligação de fatores de transcrição.

Ao longo de todo o documento, foi optado por deixar alguns termos nas suas versões originais em Inglês. Alguns destes não possuem tradução direta, enquanto outros não possuem tradução consensual, fazendo com que suas respectivas traduções tornem a leitura um pouco mais difícil.

## **2. CONTEXTUALIZAÇÃO BIOLÓGICA**

---

### **2.1 Conceitos Básicos em Biologia Molecular**

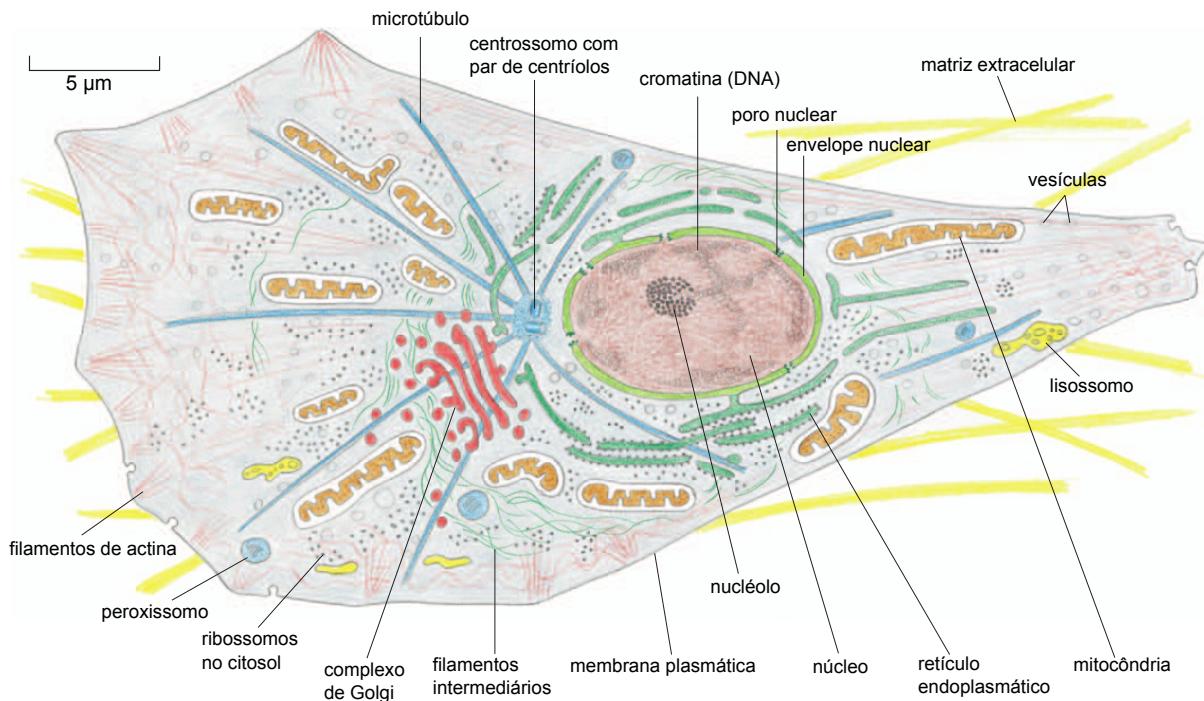
A Biologia Molecular consiste, de forma bastante sucinta, no estudo da célula no nível molecular. O principal foco desta área de conhecimento, que agrega conhecimentos, ferramentas e objetivos em comum com áreas como Bioquímica e Genética, é o estudo do material genético contido dentro da célula dos organismos e os seus produtos, as proteínas. Esta seção será baseada nos livros e artigos [Alberts, 2007; Allis *et al.*, 2007; Lewin, 2003; Lodish *et al.*, 2007; Maston *et al.*, 2006; Setubal & Meidanis, 1997; Watson *et al.*, 2003], nos quais mais detalhes podem ser encontrados sobre os processos aqui exibidos.

Estima-se que existam mais de 10 milhões, provavelmente 100 milhões, de organismos vivos no nosso planeta atualmente [Alberts, 2007]. Cada espécie possui características próprias e é capaz de se reproduzir gerando descendentes da mesma espécie, isto é, com atributos específicos na definição dessas espécies. Esse fenômeno, chamado hereditariedade, é central para a definição de vida, distinguindo-a de outros processos químicos naturais. A maioria dos organismos vivos são compostos por uma única célula (organismos unicelulares); outros, como nós humanos, são compostos por mais de uma célula (organismos multicelulares). As células são o meio pelo qual a informação hereditária se propaga através das gerações, possuindo toda a maquinaria necessária para agregar materiais naturais do ambiente e construir novas células a partir deles, contendo uma cópia completa da informação hereditária. A esse tipo de informação é dado o nome de carga genética, por motivos que ficarão mais claros no decorrer do texto. A Figura 2.1 mostra um exemplo de uma célula animal e seus principais componentes

Dos diversos componentes presentes dentro das células, existem quatro tipos de macromoléculas. Essas macromoléculas são polímeros, isto é, são longas sequências de unidades menores agregadas umas às outras, chamadas de monômeros. São elas: carboidratos (formados por açúcares), lipídeos (formados por componentes como ácidos graxos ou glicerol), proteínas (formadas por aminoácidos) e ácidos nucleicos (formados por nucleotídeos). As duas últimas serão focadas, já que possuem relação com as características hereditárias de interesse para este trabalho.

As proteínas possuem diversas funções no organismo, entre elas: catálise de reações químicas (enzimas), processamento de metabólitos, sinalização celular, regulação da produção das próprias proteínas e função estrutural. Pela grande frequência nas atividades metabólicas, número de diferentes tipos proteicos e variedade de processos em que as proteínas atuam, pode-se dizer que elas possuem um papel central para a manutenção dos organismos vivos. Os ácidos nucleicos, por sua vez, encontram-se nos formatos de ácido desoxirribonucleico (DNA) e ácido ribonucleico (RNA). A função do DNA é guardar a informação hereditária mencionada no início deste texto. O RNA, por sua vez, desempenha um papel fundamental nos processos necessários para a manifestação destas informações. O restante desta seção será focada na definição das estruturas

## 2.1. CONCEITOS BÁSICOS EM BIOLOGIA MOLECULAR



**Figura 2.1: Célula eucariótica animal** - Os principais componentes da célula eucariótica animal.  
Fonte: [Alberts, 2007]

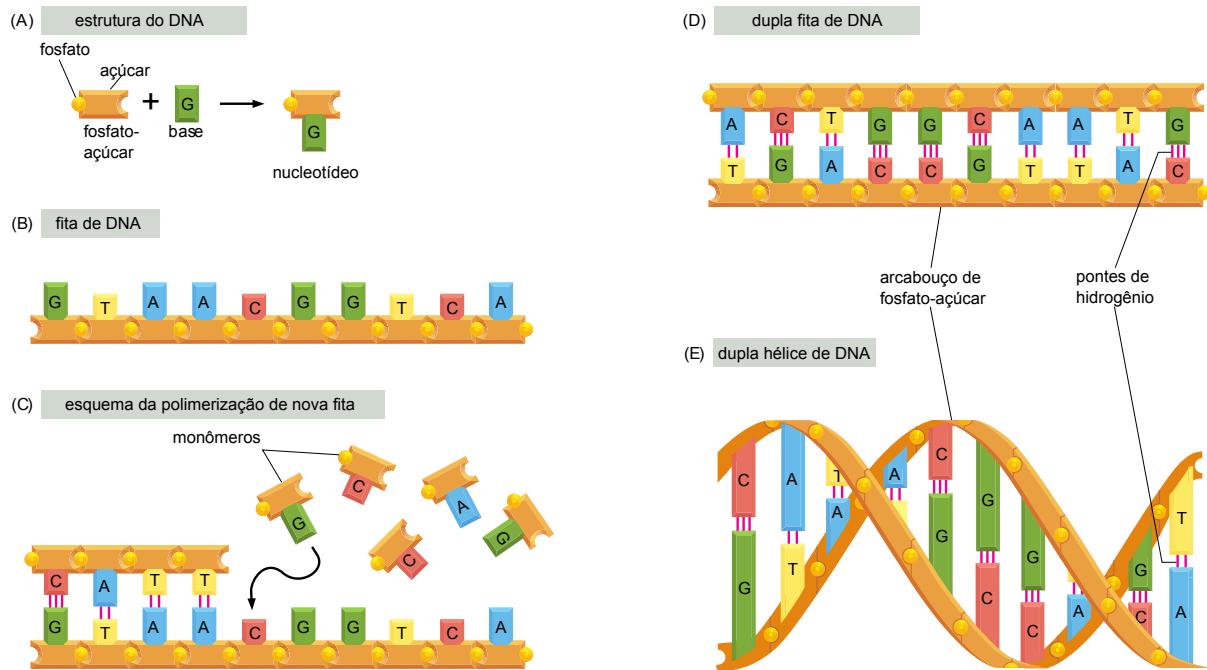
do DNA, RNA e proteínas e no detalhamento do processo chamado dogma central da Biologia Molecular, onde as proteínas são criadas a partir da informação contida no DNA.

### 2.1.1 DNA e RNA

A molécula de DNA é formada por uma dupla hélice de cadeias poliméricas emparelhadas dos mesmos quatro tipos de monômeros, os nucleotídeos adenina (A), citosina (C), guanina (G) e timina (T) (Figura 2.2). Cada nucleotídeo é composto por um açúcar (desoxirribose), um grupo fosfato e uma base nitrogenada (que define o tipo do nucleotídeo). Cada nucleotídeo é ligado a outro pertencente à mesma fita através de ligações fosfodiéster formando um arcabouço (em Inglês, *backbone*) de açúcar fosfato. As duas fitas são conectadas através de pontes de hidrogênio formadas entre as bases nitrogenadas, que se projetam para o interior das fitas. Duas pontes de hidrogênio são formadas entre adenina e timina e três pontes de hidrogênio são formadas entre citosina e guanina. Por esta razão, é comum citar nucleotídeos como pares de bases (bp) ou apenas bases.

A molécula de RNA difere da molécula de DNA por possuir o açúcar ribose ao invés da desoxirribose, por geralmente existir no formato de fita simples, e não dupla (a ribose confere

## 2. CONTEXTUALIZAÇÃO BIOLÓGICA



**Figura 2.2: Estrutura do DNA** - (A) Esquema dos componentes que formam o nucleotídeo, unidade básica do DNA. (B) Vários nucleotídeos, dos diferentes tipos possíveis (A, C, G ou T), ligados através de ligações fosfodiéster formando uma fita simples de DNA. (C) O DNA é abundante em fita dupla. Processos biológicos permitem a adição de nucleotídeos a uma fita simples, formando uma fita dupla de DNA, em um processo nomeado polimerização. Os nucleotídeos do tipo A sempre formam duas pontes de hidrogênio com o tipo T e os nucleotídeos do tipo C sempre formam três pontes de hidrogênio com o tipo G. Também é comum o uso do termo hibridização para quando duas fitas pré-existentes se ligam devido à complementaridade de seus nucleotídeos; e o termo desnaturação, para quando algum evento, como o aumento da temperatura, separa as duas fitas preservando as ligações fosfodiéster de ambas. (D) Fita dupla exibida em um esquema linear, com objetivo meramente ilustrativo, já que o DNA geralmente ocorre em formato de dupla hélice. (E) O DNA em seu formato comum na natureza – dupla hélice. Fonte: [Alberts, 2007]

uma maior estabilidade à esta estrutura, que inclusive possui capacidade de se hibridizar consigo própria) e pelo fato de que o nucleotídeo timina é substituído pela uracila (U). As moléculas de RNA possuem várias funções, das quais algumas serão descritas adiante. Por este motivo, existe uma extensa nomenclatura para os RNAs, de acordo com sua função. Os mais comuns são o mRNA (RNA mensageiro), tRNA (RNA transportador) e rRNA (RNA ribossômico), cujas funções ficarão claras durante a explicação do dogma central da Biologia Molecular.

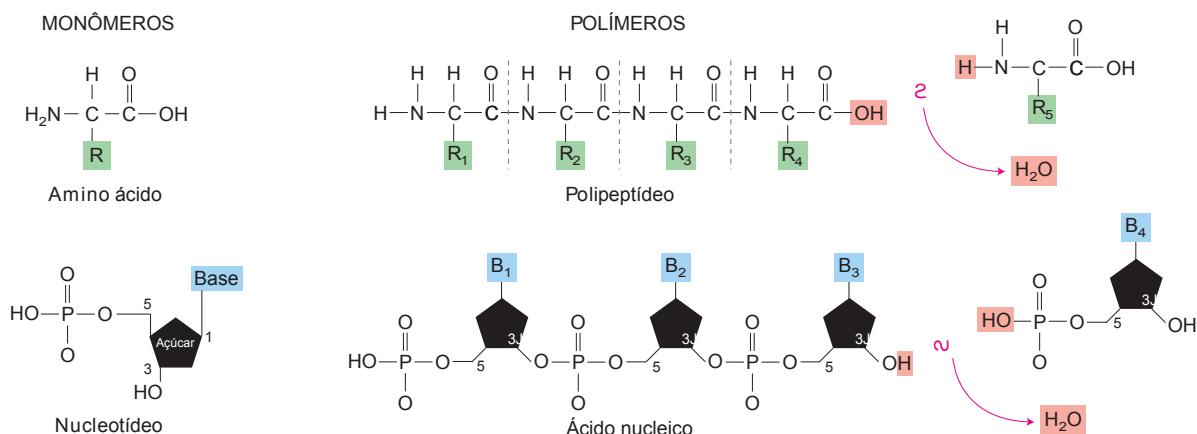
### 2.1.2 Proteínas

As proteínas são compostos químicos de alto peso molecular formados por uma longa cadeia de aminoácidos. Elas consistem em, aproximadamente, 80% do peso seco de uma célula. Essas

## 2.1. CONCEITOS BÁSICOS EM BIOLOGIA MOLECULAR

macromoléculas são formadas por blocos de aminoácidos que, por sua vez, são moléculas que possuem um carbono central ligado a um grupo carboxila, um grupo amina, um hidrogênio e uma cadeia lateral. Essa cadeia lateral pode assumir um entre vinte valores diferentes, definindo o tipo do aminoácido. A ordem específica dos aminoácidos que formam a cadeia polipeptídica determina a estrutura tridimensional da proteína, pelo fato de que cada tipo de aminoácido possui certas características físico-químicas e a estrutura dos aminoácidos permite certas rotações em torno do carbono central.

Sabe-se que a forma da proteína está diretamente relacionada com a sua função. A simples substituição de um aminoácido da cadeia é suficiente para que a proteína modifique sua conformação levando a um mal funcionamento ou a um funcionamento incompleto. Finalmente, as proteínas possuem sítios específicos onde elas interagem com outras proteínas, moléculas ou metabólitos chamados sítios ativos. A Figura 2.3 mostra a comparação das estruturas químicas da proteína e do DNA.



**Figura 2.3: Comparação entre as estruturas moleculares da proteína e do DNA** - Na primeira linha estão definidos o monômero e o polímero que correspondem à proteína. Na segunda linha temos o mesmo esquema para a estrutura do DNA. Em ambos os polímeros, novos monômeros são adicionados através de uma reação de condensação. Fonte: [Lodish *et al.*, 2007]

Utiliza-se o termo domínio para se referir a uma parte da proteína que parece uma estrutura estável em solução por si só. A maioria das proteínas varia, em tamanho, entre 100 e 2.000 resíduos de aminoácidos. Proteínas que possuem peso molecular maior do que 20.000 daltons geralmente são formadas por dois ou mais domínios; entretanto, proteínas de alto peso molecular (entre 500.000 a 2.500.000 daltons) são compostas por diversas cadeias polipeptídicas. Cada proteína possui uma certa quantidade de sítios ativos, que realizam alguma atividade metabólica através da capacidade de se ligar com outras moléculas, como DNA, RNA, metabólitos ou até outras proteínas.

## **2. CONTEXTUALIZAÇÃO BIOLÓGICA**

---

Finalmente, diferentemente do que se acreditava na época em que as primeiras estruturas de proteínas foram determinadas, as proteínas possuem um número relativamente pequeno de *motifs* estruturais dada a grande quantidade de proteínas diferentes que se conhece. Alguns tipos específicos de *motifs* de domínios são associados a atividades específicas, como o domínio intitulado *dinucleotide fold*, frequentemente encontrado em enzimas que se ligam à ATP.

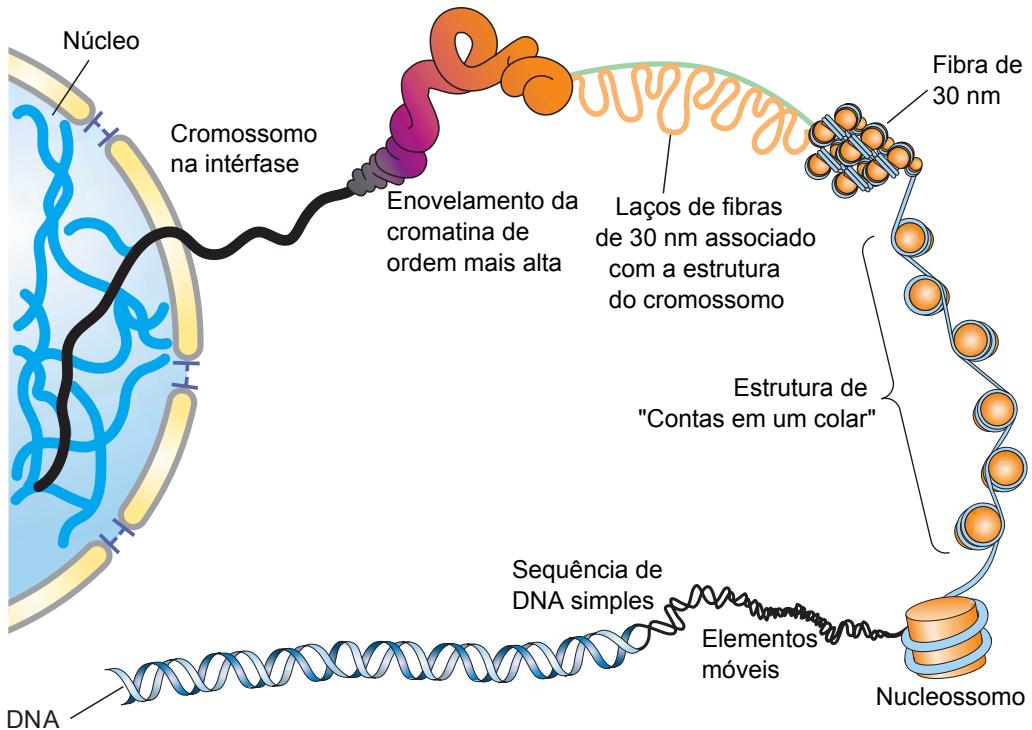
### **2.1.3 Estrutura da Cromatina**

Os organismos podem ser divididos em dois grandes grupos: procariotos e eucariotos. Os procariotos são organismos nos quais a carga genética, isto é, o DNA, está disposto no citoplasma da célula. Já os eucariotos, possuem um núcleo celular que contém, entre outras coisas, o DNA. Este projeto focará apenas nos organismos eucariotos. A grande maioria dos organismos eucariotos possui mais de uma molécula de DNA, que são chamadas de cromossomos, e o conjunto de todos os cromossomos de um organismo é chamado de genoma. Além disso, cada cromossomo pode conter uma certa quantidade de cópias, definindo a sua haploidia. No caso dos seres humanos, foco deste trabalho, existem duas cópias de um total de 22 cromossomos (nomeados de 1 a 22), mais dois cromossomos sexuais (chamados X e Y), formando um total de 46 cromossomos e definindo os humanos como seres diploides.

O DNA não se apresenta isolado no núcleo celular. Ao invés disso, ele se conforma em diversos níveis organizacionais (Figura 2.4), envolvendo elementos como as proteínas histonas, o que permite sua compacidade e confere outras funções regulatórias que ainda estão sendo estudadas e serão discutidas mais adiante. De forma simples, a cromatina pode estar condensada em uma estrutura não propensa para a iniciação da transcrição (nesse caso, recebe o nome de heterocromatina) ou pode estar descondensada, permitindo que a transcrição ocorra (eucromatina).

O DNA encontra-se envolto em um conjunto de oito histonas, formado por quatro pares dos diferentes tipos de histonas chamadas H2A, H2B, H3 e H4. Essa unidade formada pelo DNA dando, em um estado padrão, aproximadamente 1.65 voltas ( 147bp) [Allis *et al.*, 2007] em torno do complexo de histonas é chamada de nucleossomo. A partir desse nível mais baixo, a estrutura da cromatina se condensa em diversos graus. De fato, caso estiquemos o genoma humano diploide de uma célula apenas, teremos uma molécula linear com aproximadamente dois metros de comprimento. Portanto a compactação do DNA deve ser realizada de forma bastante eficaz para que a cromatina caiba dentro do núcleo celular.

## 2.1. CONCEITOS BÁSICOS EM BIOLOGIA MOLECULAR



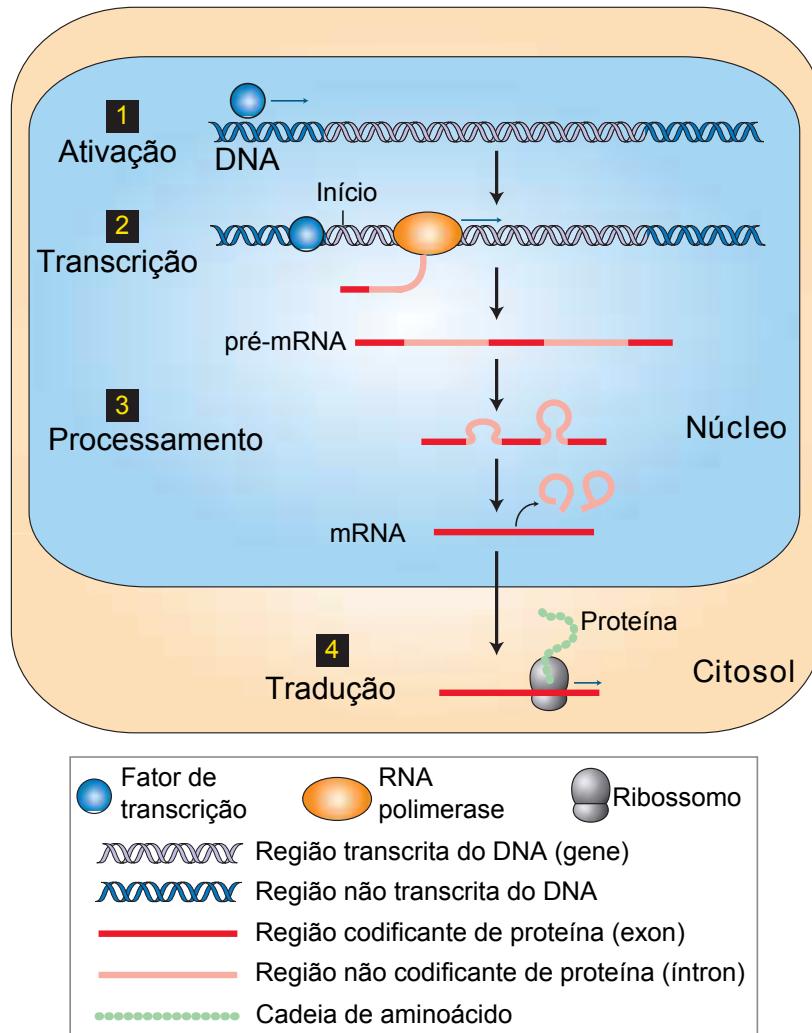
**Figura 2.4: Visão global da estrutura da cromatina** - A cromatina possui vários níveis de enovelamento, o que confere ao DNA seu caráter compacto e é de extrema importância para mecanismos regulatórios mais complexos. Fonte: [Lodish *et al.*, 2007]

### 2.1.4 Dogma Central da Biologia Molecular

Conforme mencionado previamente, as proteínas são sintetizadas a partir da informação genética contida no DNA, constituindo o processo conhecido como dogma central da Biologia Molecular (Figura 2.5). Além da produção de proteínas, o dogma central também engloba a replicação do DNA, processo pelo qual a informação genética é transmitida durante a divisão celular. Neste trabalho, entretanto, será focada apenas a síntese de proteínas, mais especificamente a transcrição.

A transcrição é a etapa responsável pela geração de uma molécula de RNA a partir de um trecho específico da molécula de DNA, chamado gene. De forma simplificada, podemos dizer que genes são trechos da molécula de DNA que possuem informação codificante, isto é, serão transformados em RNA. Tais genes podem apresentar algumas variações entre indivíduos de uma mesma espécie. De forma simplificada, cada uma destas versões é chamada de alelo. Uma parte do RNA produzido, chamado de mRNA, será posteriormente traduzido em uma proteína, e outra parte desse RNA realizará outras funções que fogem do escopo deste trabalho. Várias proteínas participam da transcrição e algumas delas serão descritas em detalhes nas seções

## 2. CONTEXTUALIZAÇÃO BIOLÓGICA

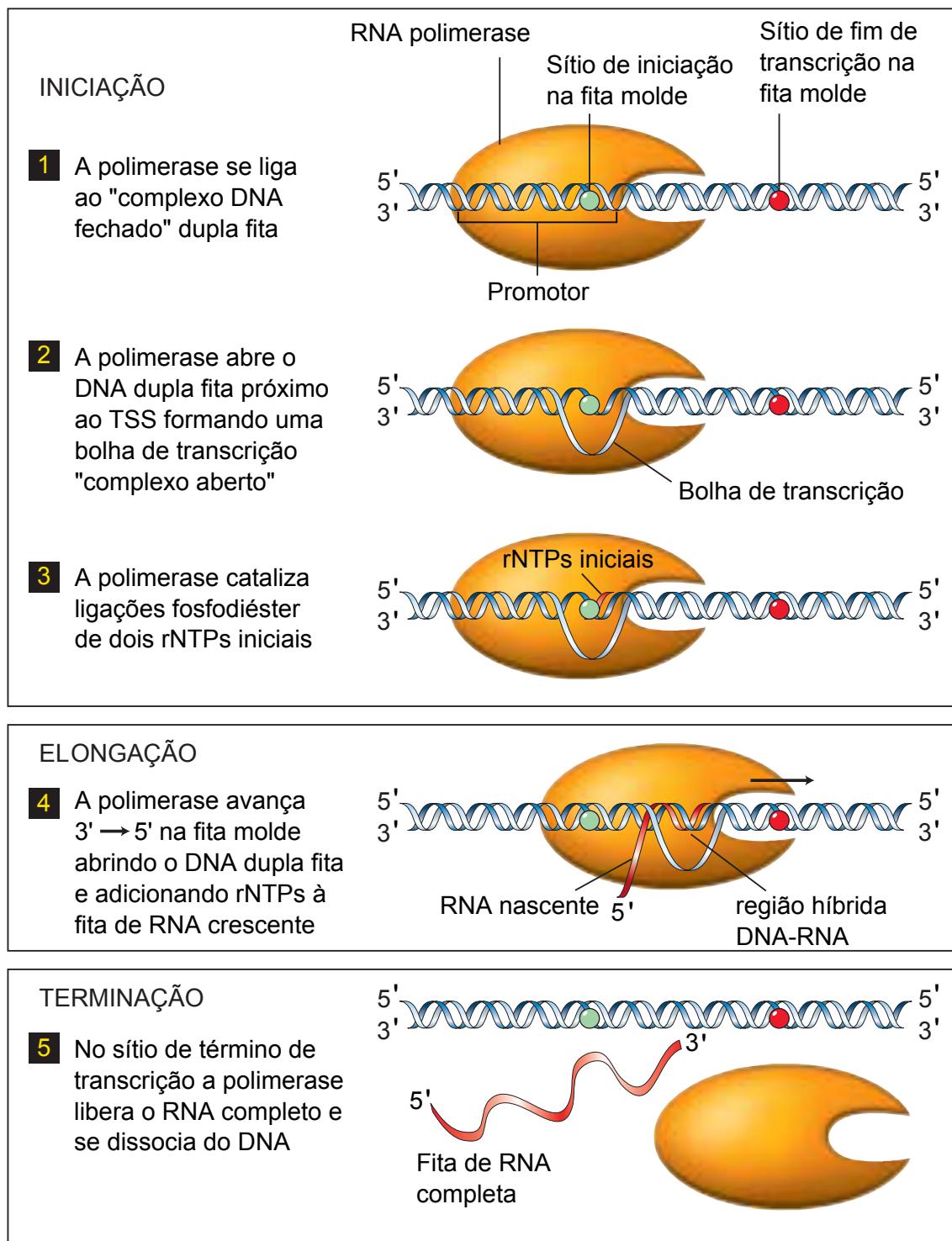


**Figura 2.5: Dogma central da Biologia Molecular** - Os quatro principais processos biológicos, dentro do contexto celular eucarioto, para a síntese de proteínas. (1) Ativação – Proteínas reguladoras da transcrição (fatores de transcrição) se acoplam no início do gene, preparando-o para a fase seguinte. (2) Transcrição – O DNA é lido pela proteína RNA polimerase e uma molécula de mRNA é criada contendo a informação complementar à fita de DNA lida. (3) Processamento do mRNA – O mRNA é processado e transportado para fora do núcleo celular. (4) Tradução – A molécula de mRNA processada é convertida em uma proteína, em estruturas chamadas ribossomos. Fonte: [Lodish *et al.*, 2007]

posteriores, porém por simplicidade, apenas a principal proteína, chamada RNA polimerase, estará em foco. Para transcrever um gene, a RNA polimerase procede por uma série de passos bem definidos que podem ser agrupados em três fases: iniciação (ou ativação), elongação e terminação (Figura 2.6).

Durante a fase de iniciação, a RNA polimerase se liga em uma região específica no DNA chamada de região promotora. Após a ligação, a fita de DNA em volta do ponto onde a

## 2.1. CONCEITOS BÁSICOS EM BIOLOGIA MOLECULAR



**Figura 2.6: Etapas do processo de transcrição** - Os três estágios que compõem o processo de transcrição. Siglas introduzidas nesta figura são definidas no glossário. Fonte: [Lodish *et al.*, 2007]

## **2. CONTEXTUALIZAÇÃO BIOLÓGICA**

---

transcrição está se iniciando se desenovela permitindo que a RNA polimerase, que está ligada a uma das fitas, continue o processo. Começa então a fase de elongação, onde o RNA, após sintetizar um pequeno trecho de RNA de aproximadamente 10 bases, começa a percorrer o gene. A cada base “lida” pela RNA polimerase, uma base é introduzida na cadeia de RNA correspondente à base com a qual a base “lida” possui afinidade. Além disso, conforme a RNA polimerase se desloca, ela abre a dupla fita de DNA à sua frente e re-hibridiza as fitas previamente abertas cujo conteúdo já foi lido. Finalmente, na fase de terminação, a RNA polimerase se desestabiliza, para e libera a cadeia de RNA produzida. Em algumas células existem sequências bem definidas que correspondem a essa terminação; porém em outras ainda não está claro o que faz com que a enzima cesse o processo de transcrição. Apenas a RNA polimerase foi citada, porém, como será visto mais adiante, diversas outras proteínas participam do processo de transcrição.

É importante mencionar que apenas uma das fitas é lida durante o processo (chamada de fita senso), porém as duas fitas contém informação necessária para produzir mRNA. Outro ponto importante é a orientação das fitas. Cada fita tem duas extremidades: uma corresponde a um grupo hidroxila ligado ao carbono 3' do açúcar, e outra corresponde ao grupo fosfato, ligado ao carbono 5' do açúcar. Por esta razão, processos que envolvem deslocamento no DNA podem possuir orientação 3' → 5' (antisenso) ou 5' → 3' (senso). Além disso, as duas fitas que compõem a dupla hélice de DNA estão ligadas em sentidos opostos. A transcrição sempre ocorre no sentido 5' → 3'.

Após a transcrição, as moléculas de mRNA que servem como molde para produção de proteínas (pré-mRNA) passam por uma série de procedimentos para torná-las aptas para o processo de tradução. Esse processo, também intitulado de *splicing* do mRNA, inicia com a exclusão de certos trechos do pré-mRNA. Os genes possuem dois tipos básicos de regiões chamadas introns e exons. Nessa fase inicial do processamento do mRNA, as regiões de introns são totalmente removidas. Em adição, algumas regiões de exons podem ser removidas da mesma forma. Além do *splicing*, a sequência do pré-mRNA pode ser alterada através de outros processos tais como o rearranjo de mRNA, o qual modifica o mRNA não processado através de uma desaminação sítio-específica e guiando a inserção ou deleção de uridinas. Após essa etapa, a molécula de mRNA sofre algumas alterações químicas em sua extremidade 5' conhecida como revestimento do terminal 5' e um fragmento adicional contendo apenas moléculas de adenina é introduzido em sua extremidade 3' em um processo intitulado poliadeniilação. O pré-mRNA passa então a ser chamado de mRNA processado e deve ser transportado para fora do núcleo celular. Porém antes do transporte, o mRNA deve ter uma coleção de características que o distinguem de outros tipos de RNA (que devem permanecer no núcleo) tais como certas proteínas que reconhecem sequências de exons.

## **2.1. CONCEITOS BÁSICOS EM BIOLOGIA MOLECULAR**

---

No processo de *splicing*, exons também podem ser removidos. Isso permite que um só gene seja capaz de gerar vários mRNAs diferentes pelo fato de que diferentes exons podem ser mantidos em resposta a diferentes estímulos celulares. Essa característica, conhecida como *splicing* alternativo, explica em grande parte (juntamente com outros processos como as modificações pós-traducionais) o fato de que existe uma quantidade muito maior de diferentes proteínas do que de genes codificantes de proteínas.

O processo de tradução consiste na leitura do mRNA processado e na criação de uma cadeia polipeptídica através da junção de aminoácidos. A principal estrutura associada à tradução é o ribossomo, que se situa no citoplasma da célula e é composto por proteínas e por rRNA. Por este motivo, o mRNA deve sair do núcleo celular para que o processo de tradução ocorra. Assim como a transcrição, a tradução pode ser dividida em várias etapas, porém como este processo não é fundamental para o entendimento deste trabalho, uma explanação mais breve será fornecida. A tradução inicia quando o mRNA é acoplado ao ribossomo. Cada trinca de bases do mRNA (chamada códon) é “lida” pelo ribossomo, que irá acoplar um aminoácido correspondente à trinca na sequência de aminoácidos que está sendo gerada. Cada códon possui um aminoácido correspondente e, pelo fato de existirem 64 possíveis combinações de códons e apenas 20 aminoácidos, alguns aminoácidos correspondem a mais de um códon. Existem também códons específicos para indicar a posição onde esse processo de tradução irá começar e terminar. Os tRNAs são as estruturas responsáveis por armazenar cada aminoácido que será posteriormente acoplado à cadeia. Eles são formados por um códon específico de um lado e um aminoácido ligado ao outro e estão presentes em número muito grande no citoplasma. Quando determinado códon do mRNA é lido, um rRNA que estiver próximo do ribossomo é alinhado com este códon, acarretando na junção do aminoácido que está em uma de suas extremidades à sequência de aminoácidos corrente.

A proteína formada irá se conformar de acordo com as propriedades físico-químicas dos aminoácidos influenciadas pelo meio aquoso do citoplasma. Após essa conformação, a proteína está pronta para realizar suas atividades. Entretanto, algumas proteínas sofrem modificações pós-traducionais, podendo acarretar em uma modificação em sua estrutura. Essas modificações geralmente envolvem a adição de grupos metil, acetil e vários outros em determinados aminoácidos. As histonas, que fazem parte da estrutura da cromatina, são exemplos de proteínas que sofrem modificações pós-traducionais e serão abordadas em detalhes mais adiante.

O dogma central da Biologia Molecular é o procedimento chave para manutenção da vida como conhecemos. Algumas fases desse complexo processo foram descritas de forma bastante simplificada. As próximas seções correspondem ao detalhamento da fase de transcrição, principalmente a fase de iniciação, explicando os principais mecanismos conhecidos atualmente que contribuem para a regulação espacial e temporal das regiões gênicas que serão transcritas.

## **2. CONTEXTUALIZAÇÃO BIOLÓGICA**

---

### **2.2 Regulação Gênica em Eucariotos**

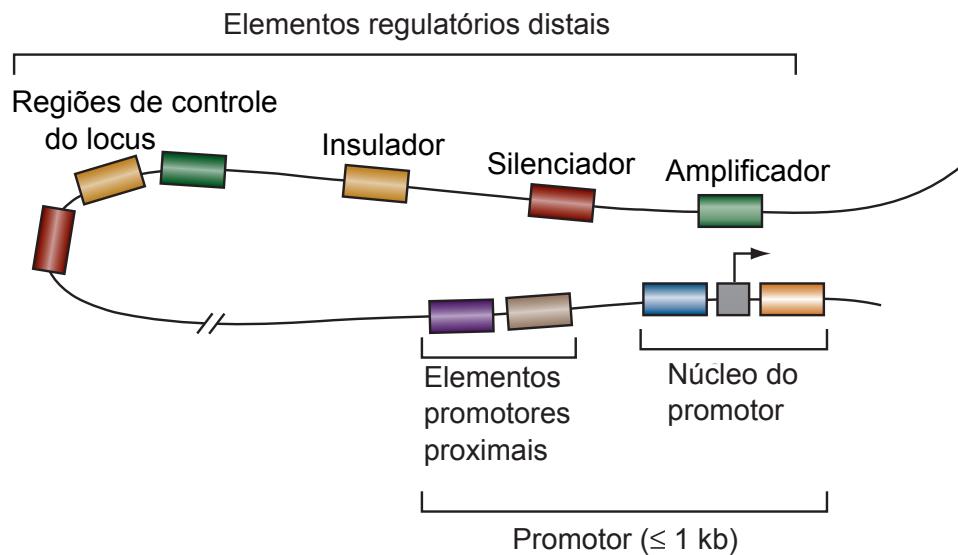
Na seção anterior foram discutidos alguns conceitos básicos a respeito do processo de criação de proteínas a partir do DNA. É importante mencionar que a transcrição não tem como objetivo exclusivo a produção de RNA que será transformado em proteínas. Existem vários outros tipos de RNA que atuam em diversos tipos de processos moleculares. A etapa de iniciação da transcrição foi descrita anteriormente como sendo a etapa onde a RNA polimerase deve se ligar à região promotora para que o procedimento possa começar, porém vários fatores contribuem para que os genes sejam transcritos. Dá-se o nome de regulação gênica a todos os processos que as células utilizam para regular a forma como os genes são convertidos em moléculas de RNA.

A regulação gênica pode ocorrer em vários níveis diferentes do dogma central: iniciação da transcrição, elongação da transcrição, processamento de mRNA, transporte do mRNA do núcleo até o citoplasma, tradução e estabilidade do mRNA. Entretanto, acredita-se que a maior parte dos eventos regulatórios ocorram no nível de iniciação da transcrição. Parte da regulação nesta etapa é baseada em proteínas chamadas elementos regulatórios, que utilizam propriedades físicas e químicas para fazer com que os genes sejam transcritos em diversos níveis de intensidade, desde nenhuma transcrição (gene silenciado ou inativo) até o nível máximo de transcrição comportado por aquele gene, dado o seu locus na cromatina. Os genes transcritos pela RNA polimerase II (eucariotos) tipicamente contêm dois tipos de elementos regulatórios: os elementos cis-atuantes e os elementos trans-atuantes.

Os elementos cis-atuantes constituem as regiões no DNA onde os elementos trans-atuantes se ligam. Neste trabalho, essa nomenclatura será extrapolada, sendo os elementos trans-atuantes também chamados de fatores de transcrição (TFs) e os elementos cis-atuantes, também chamados de sítios de ligação de fatores de transcrição (TFBSs). Os elementos cis-atuantes podem ser divididos em duas famílias distintas (Figura 2.7): (1) um promotor, composto por um núcleo e por elementos regulatórios proximais; (2) elementos regulatórios distais, divididos atualmente em amplificadores, silenciadores, insuladores e regiões de controle do locus (LCRs, do Inglês *Locus Control Regions*).

A estrutura (ou disposição) dos elementos cis- e trans-atuantes pode chegar a ser bastante complexa. Essa complexidade se faz necessária dado que existem 20.000 – 25.000 genes no genoma humano, cada um requerente de um padrão específico de expressão espacial/temporal, existindo apenas pouco mais do que 1.500 fatores de transcrição. A presença de múltiplos elementos regulatórios em regiões proximais ou distais conferem a possibilidade de uma regulação combinatória, que aumenta de forma exponencial o número total de padrões de expressão possíveis.

## 2.2. REGULAÇÃO GÊNICA EM EUCAΡIOTOS



**Figura 2.7: Diferentes tipos de elementos cis-atuantes** - Região regulatória típica de um gene, contendo um promotor (núcleo do promotor e elementos proximais) e elementos regulatórios distais (amplificador, silenciador, insulador e região de controle do locus) Fonte: [Maston *et al.*, 2006]

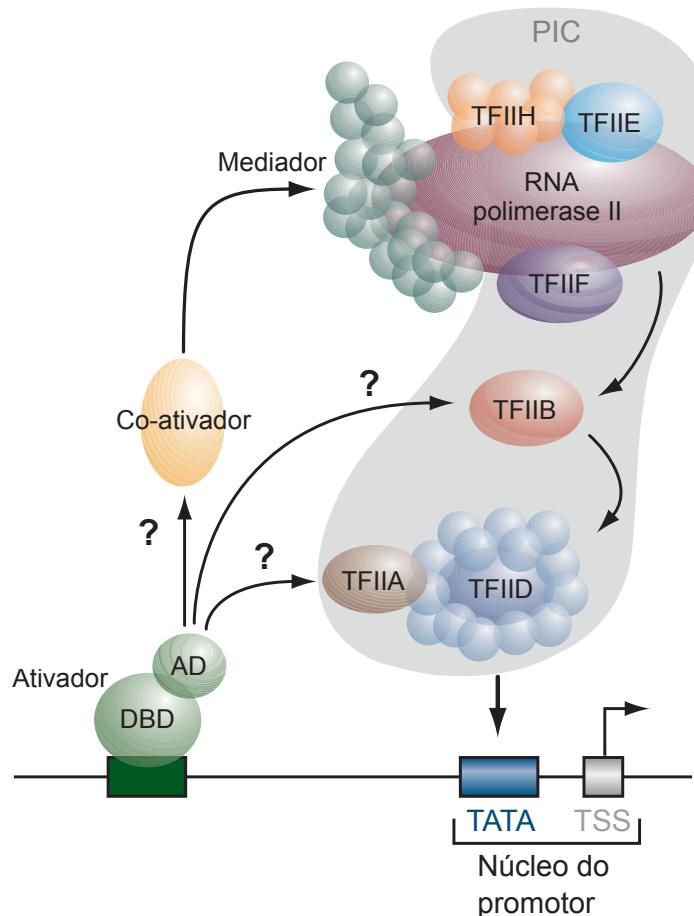
### 2.2.1 Maquinaria Regulatória Proximal

Os fatores proximais envolvidos na transcrição eucariótica podem ser divididos em três grupos (Figura 2.8): (1) fatores de transcrição gerais (ou básicos), que incluem a RNA polimerase II e vários componentes auxiliares (TFIIB, TFIID, TFIIE, TFIIF, TFIIG e TFIIH); (2) ativadores; (3) co-ativadores. Em adição a esses componentes, o Mediador – uma estrutura grande e altamente conservada – também é importante para a transcrição acurada.

Os fatores de transcrição gerais (GTF, do Inglês *General Transcription Factors*) se montam na região promotora em uma ordem específica, para formar o complexo pré-iniciação (PIC, do Inglês *Preinitiation Complex*), que direciona a RNA polimerase II para o sítio de iniciação da transcrição (TSS, do Inglês *Transcription Start Site*). Primeiramente, o TFIID se liga numa região chamada caixa TATA (em Inglês, *TATA box*). Após isso, alguns eventos ocorrem antes da fase de elongação, incluindo a fusão do promotor, liberação e escape. Quando a RNA polimerase II procede para a etapa de elongação, uma armação composta pelos fatores TFIID, TFIIE, TFIIF e mediador, permanece no núcleo do promotor, fazendo com que a re-iniciação da transcrição necessite apenas do recrutamento da RNA polimerase II e dos fatores TFIIF e TFIIB.

A montagem do PIC no núcleo do promotor é suficiente para permitir níveis baixos e acurados de transcrição (nível basal). Os ativadores possuem a capacidade de estimular bastante o nível da transcrição. Em geral, esses fatores são proteínas que se ligam ao DNA, reconhecendo

## 2. CONTEXTUALIZAÇÃO BIOLÓGICA



**Figura 2.8: Maquinaria transcrecional eucariótica** - Fatores de transcrição gerais, ativadores e co-ativadores se montam na região promotora de uma forma ordenada, formando o complexo pré-iniciação. As interrogações representam as conexões que ainda estão sendo estudadas, cuja ordem de ligação, até o presente momento, ainda não foi conclusivamente identificada. Fonte: [Maston *et al.*, 2006]

sequências que geralmente ocorrem à montante do núcleo do promotor. Eles contêm domínios de ligação no DNA e de ativação, necessários para a estimulação da transcrição. A estimulação da transcrição pode se dar de várias formas: (1) ajudando na formação rápida e apropriada do PIC através de interações diretas com um ou mais componentes da maquinaria transcrecional (alvos); (2) promovendo outras etapas transpcionais como elongação ou re-iniciação; (3) re-crutando complexos modificadores de estrutura da cromatina (que, através de modificações pós-tradicionais nas caudas das histonas, fazem com que a cromatina fique em um estado mais aberto e propício para a transcrição).

O funcionamento dos ativadores pode ser modulado pelos co-ativadores. Tipicamente, os co-ativadores não contêm domínios para reconhecimento de sequências específicas no DNA. Ao invés disso, eles contêm domínios necessários para realizar interações proteína-proteína com um

## 2.2. REGULAÇÃO GÊNICA EM EUCA RIOTOS

---

ou mais ativadores no DNA. O modo como este tipo de fator aumenta o nível transcrecional é basicamente o mesmo dos ativadores, porém eles possuem uma propriedade adicional na qual são capazes de regular o funcionamento de um ativador para que estes realizem uma regulação positiva ou negativa.

Uma das características mais interessantes observadas nos ativadores é que eles são capazes de estimular a transcrição sinergicamente. Neste fenômeno, o efeito de múltiplos fatores trabalhando juntos é maior do que a soma dos efeitos que eles teriam se estivesse trabalhando individualmente. Esse efeito pode ocorrer de forma promiscua, na qual diversos fatores de diferentes tipos encontram-se nesse estado cooperativo, ou de forma não-promiscua, na qual várias cópias de um mesmo fator estão presentes. Apesar de ter sido observado, esse fenômeno ainda não é completamente conhecido.

### 2.2.2 Elementos Regulatórios Transcricionais

A seguir são descritos brevemente os elementos regulatórios transcricionais apresentados. A Figura 2.9 sumariza os elementos regulatórios que atuam de forma distal. Cada elemento regulatório apresentado funciona de forma diferente, contribuindo para o aumento do nível transcrecional ou diminuição deste nível (e possível silenciamento total do gene) ou para ambos dependendo do contexto em que é inserido.

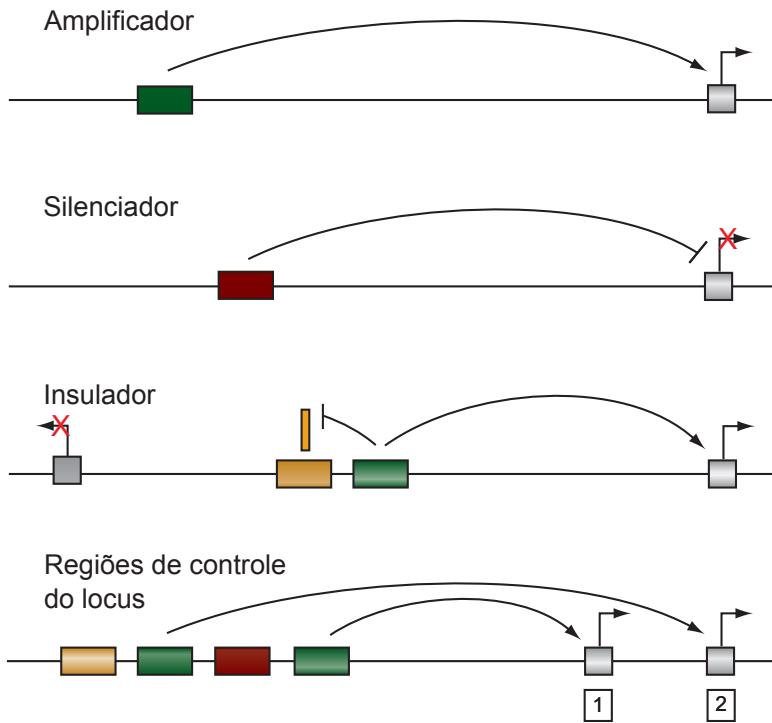
#### 2.2.2.1 Núcleo do Promotor

É a região no início do gene que possui elementos onde a maquinaria geral de transcrição se liga e o PIC se forma, definindo a posição do TSS e a direção da transcrição. Alguns desses elementos foram bastante estudados tais como o elemento iniciador (Inr), a caixa TATA, o elemento central à jusante (DCE, do Inglês *Downstream Core Element*), o elemento de reconhecimento do TFIIB (BRE, do Inglês *TFIIB-Recognition Element*) e o *motif* na posição 10 (MTE, do Inglês *Motif Ten Element*). Com exceção do BRE, todos os outros elementos descritos até então interagem com o fator TFIID.

Análises estatísticas em 10.000 diferentes promotores mostraram que tais elementos não são tão universais quanto se pensava. De fato, aproximadamente um quarto dos promotores analisados não possuía nenhum desses quatro elementos mencionados, sugerindo que talvez existam arquiteturas mais complexas a serem descobertas. De fato, pesquisas recentes apontam para arquiteturas menos usuais tais como os desertos de ATG. Além disso, foi descoberto recentemente que as propriedades estruturais de ordens mais altas do promotor, que são determinadas em parte pela sequência de nucleotídeos e sua curvatura, dobrabilidade e estabilidade, podem ser usadas para identificar e classificar esses núcleos dos promotores.

## 2. CONTEXTUALIZAÇÃO BIOLÓGICA

---



**Figura 2.9: Funcionamento dos elementos regulatórios distais** - A função dos amplificadores e silenciadores é de, respectivamente, ativar e reprimir a transcrição. Insuladores evitam que genes sejam afetados por elementos regulatórios na vizinhança. Regiões de controle do locus são trechos compostos por vários elementos regulatórios cujo funcionamento em conjunto confere um padrão de expressão singular que afeta agrupamentos de genes nas proximidades. Fonte: [Maston *et al.*, 2006]

### 2.2.2.2 Elementos Promotores Proximais

Os elementos promotores proximais estão localizados imediatamente à montante (até no máximo algumas centenas de pares de bases) do núcleo do promotor, contendo vários sítios de ligação para ativadores. Uma característica interessante está no fato de que aproximadamente 60% dos promotores situam-se próximos à ilhas de CpG – trechos que variam de 500 bp a 2 kb que contêm uma alta quantidade de nucleotídeos C+G e uma frequência de CpG mais alta do que outras regiões do DNA. A maioria dos dinucleotídeos CpG no genoma são metilados no quinto carbono da citosina, entretanto os nucleotídeos em ilhas CpG geralmente não são metilados. Existem várias correlações interessantes a esse respeito, como o fato de que promotores que contêm caixa TATA geralmente não estão próximos a ilhas CpG, porém promotores baseados em BREs são bastante associados a essas ilhas. O fato de que a metilação do DNA está associada ao silenciamento da transcrição sugere que a função das ilhas CpG seja de impedir a metilação dessa região, consequentemente, silenciando-a.

## **2.2. REGULAÇÃO GÊNICA EM EUCA RIOTOS**

---

### **2.2.2.3 Amplificadores**

Elementos amplificadores regulam a expressão temporalmente e espacialmente e sua atividade independe da distância ao promotor (que pode chegar à ordem de Mb) ou da sua orientação em relação a este. Essa região é tipicamente composta por vários sítios de ligação bastante próximos uns dos outros, onde os amplificadores se ligam para aumentar a expressão do gene. Amplificadores também são modulares, isto é, a atividade de um único promotor pode ser modificada por diferentes amplificadores em tempos diferentes ou tecidos diferentes, em resposta a diferentes estímulos. Além disso, a organização espacial e orientação dos sítios de ligação que formam o amplificador podem ser vitais para sua atividade regulatória.

Amplificadores são funcionalmente similares aos elementos proximais e a distinção entre eles ainda é bastante nebulosa. De fato, grande parte dos fatores que se liga em regiões proximais também se liga em amplificadores. Existem fortes evidências de que esses elementos distais (como os amplificadores) consigam atuar a partir de regiões tão distantes através do modelo de laço do DNA (em Inglês, *DNA looping*). Neste modelo, o DNA se conforma de tal maneira que, apesar de estar vários bps longe do núcleo do promotor, fisicamente estas estruturas podem estar próximas umas das outras (como na junção das duas extremidades de um cadarço de tênis). Alguns modelos propõem até que parte do PIC se forme em regiões amplificadoras e que esse complexo se agregue ao restante dos fatores gerais através do processo de laço do DNA.

### **2.2.2.4 Silenciadores**

Silenciadores são elementos que reprimem a expressão de um gene (efeito transcripcional negativo). Assim como os amplificadores, a atuação da maioria dos silenciadores não depende da distância à região promotora nem da orientação, porém alguns silenciadores dependentes da posição foram encontrados. Os silenciadores podem estar em regiões proximais, em regiões distais de amplificadores ou em regiões distais independentes. Além disso, silenciadores podem se ligar ao DNA cooperativamente e também possuem características sinérgicas.

O fator de transcrição que se liga em um elemento silenciador é chamado de repressor, nos quais os co-repressores podem se ligar (de forma semelhante aos ativadores e co-ativadores). Como mencionado anteriormente, ativadores podem se tornar repressores através do recrutamento de alguns co-fatores específicos. Os silenciadores podem reprimir a expressão de diversas formas: (1) não permitindo a ligação de um ativador ou componente da maquinaria transcripcional, bloqueando fisicamente suas ligações ou competindo diretamente por um mesmo sítio; (2) inibindo a formação do complexo pré-iniciação; (3) recrutando modificadores de cromatina para condensar a região de forma a dificultar a ligação de ativadores ou da própria maquinaria transcripcional.

## **2. CONTEXTUALIZAÇÃO BIOLÓGICA**

---

### **2.2.2.5 Insuladores**

Insuladores, também conhecidos como elementos de fronteira, bloqueiam a atuação de outros elementos regulatórios definindo uma espécie de participação do genoma em blocos com sistema interno de regulação. Os insuladores têm duas propriedades específicas: (1) bloquear a influência de um amplificador sobre a expressão de um determinado gene, bloqueando a comunicação amplificador-promotor; (2) bloquear a disseminação do silenciamento de uma região por estruturas que condensam a cromatina (que geralmente agem como uma reação em cadeia, parando apenas ao encontrar o insulador). Esses elementos geralmente são dependentes de posição porém independentes de orientação.

Apesar de vários fatores trans-atuantes que mediam a função do insulador serem conhecidos para a *Drosophila*, em vertebrados se conhece apenas o CTCF (do Inglês *CCCTC-binding factor*). A atividade deste fator pode ser regulada de várias formas, incluindo metilação do DNA, modificação pós-traducionais e interação com co-fatores.

A forma como os insuladores realizam suas funções de bloqueio de comunicação amplificador-promotor ou barreira para heterocromatina ainda não é conhecida. Os modelos propostos podem ser agrupados em duas categorias. A primeira associa os insuladores com a maquinaria regulatória transcricional, e a segunda os associa com a organização estrutural da cromatina.

### **2.2.2.6 Regiões de Controle de Locus**

Regiões de controle de locus são grupos de elementos regulatórios, tais como amplificadores, silenciadores e insuladores, envolvidos na regulação de um locus inteiro ou de um agrupamento de genes. Tais regiões são definidas operacionalmente como elementos que direcionam a expressão fisiológica específica por tecido de uma forma independente de posição e dependente de variação do número de cópias gênicas (CNV, do Inglês *Copy-Number Variation*). Os elementos que se ligam nestas regiões (ativadores, co-ativadores, repressores, co-repressores ou modificadores de cromatina) podem afetar a expressão de forma distinta e sua atividade coletiva que confere a função específica de cada LCR.

## **2.3 Identificação de Sítios de Ligação de Fatores de Transcrição**

Na Seção 2.2 foi apresentada uma introdução superficial à área de regulação gênica. Várias propriedades dos elementos regulatórios foram definidas, pretendendo com isso motivar estudos que propõem métodos para identificar a localização de tais estruturas no DNA. De fato, redes regulatórias complexas governam diversos mecanismos celulares críticos para a célula, tais

## **2.3. IDENTIFICAÇÃO DE SÍTIOS DE LIGAÇÃO DE FATORES DE TRANSCRIÇÃO**

---

como a proliferação, desenvolvimento, diferenciação, envelhecimento e apoptose. Para que esses mecanismos funcionem de forma correta e consistente, um número muito grande de diferentes componentes regulatórios devem desempenhar seus papéis, que podem variar de acordo com as circunstâncias, em diversas vias metabólicas. Todos os elementos mencionados na seção anterior, colaboram para a orquestração espacial/temporal apropriada da expressão gênica de processos celulares ubíquos, comuns entre certos tipos de células ou totalmente específicos por célula. Consequentemente, a identificação desses elementos regulatórios é crucial para a compreensão da função (ou funções) que cada um deles desempenha nas numerosas redes regulatórias das quais participam. Isso permite, por exemplo, a melhor compreensão de doenças causadas pela desregulação (regulação imprópria por um grande número de diferentes razões).

Conforme mencionado anteriormente, estima-se que o número de diferentes fatores de transcrição em humanos seja maior do que 1.500. Cada um desses fatores pode se ligar no DNA diretamente ou através do recrutamento de outros fatores (por exemplo, em um esquema ativador – co-ativador, como revisado na Seção 2.2.2.2). Além disso, alguns elementos distais compostos por várias estruturas regulatórias menores (como os LCRs) possuem função diretamente equivalente às suas configurações, isto é, aos tipos de elementos que compõem estas regiões e à disposição dos mesmos dentro destes loci. Ademais, as sequências onde tais fatores trans-atuantes têm maior afinidade de ligação geralmente são pequenas, variando entre 6 – 12 bp, dos quais apenas um número ainda menor de nucleotídeos está presente de forma quase consensual. Somando todas essas características, a identificação destas regiões se torna bastante complexa, sendo necessários esforços (e avanços) nas áreas biológica e computacional para que esta tarefa tenha bons resultados.

Finalmente, uma das maiores dificuldades está no fato de que tais elementos regulatórios são específicos por tipo (ou linha) celular. O genoma humano consiste, em teoria, na mesma sequência de nucleotídeos para todas as células do organismo. Sabe-se atualmente que existem diferenças significativas até entre células de um mesmo tipo, como variações no número de cromossomos observadas recentemente em neurônios, porém tais diferenças não excluem a hipótese atualmente aceita de que as diferenças entre as células do organismo se dão majoritariamente devido ao controle regulatório, que ativa ou desativa, em diferentes graus, diferentes genes, modificando o padrão da expressão e consequentemente gerando diferenças estruturais significativas. A partir disso, define-se a maior limitação dos métodos computacionais automáticos, baseados em busca por sequência, como o fato de tais métodos não conseguirem distinguir quais os sítios de afinidade de ligação de proteínas no DNA estão ativos ou inativos.

Nas Seções 2.3.1 e 2.3.2 a seguir, serão explorados os dois métodos biológicos tradicionais mais comuns para a identificação de sítios de ligação de fatores de transcrição. Adicionalmente, será definida na Seção 2.3.3 a abordagem computacional padrão para o problema, que são as buscas baseadas em sequência. Tais métodos possuem limitações bem evidentes, seja terem

## **2. CONTEXTUALIZAÇÃO BIOLÓGICA**

---

baixo rendimento (não sendo possível a aplicação em escala genômica) ou pelas dificuldades mencionadas nos parágrafos anteriores. Entretanto, na Seção 2.4.2 serão realizadas extensões desses métodos, cuja aplicação se enquadra no estado da arte das soluções deste problema, sendo esta a motivação para a apresentação de tais tecnologias.

### **2.3.1 DNase I Footprinting**

Este método tradicional consiste em observar padrões de digestão no DNA de algum agente de clivagem capaz de quebrar as ligações fosfodiéster desta molécula. Estes agentes podem ser, por exemplo, radicais hidroxila ou radiação ultravioleta. Porém neste trabalho será dado foco à endonuclease Desoxirribonuclease I (DNase I). Esta enzima é capaz de se ligar no sulco menor (ou secundário) da dupla hélice de DNA e produzir uma quebra na ligação fosfodiéster. A DNase I é perfeita em experimentos desse gênero pois o seu grande tamanho faz com que ela seja realmente sensível a proteínas que estão ligadas no DNA e também porque sua ação é facilmente controlada com EDTA (ver glossário).

O método se inicia com a obtenção do DNA genômico. De posse do DNA de várias células do tipo específico sob estudo, a porção onde se deseja verificar se existem indícios de elementos funcionais (isto é, se possuem sítios de ligação de fatores de transcrição) é amplificada via reação em cadeia da polimerase (PCR, do Inglês *Polymerase Chain Reaction*). Amplificação é o processo de geração de várias moléculas de DNA idênticas à original. O tamanho ideal para tal região deve ser entre 50 e 200 pares de bases. Neste momento se torna claro que a principal desvantagem deste método é o baixo rendimento, isto é, uma rodada deste método demora um tempo razoavelmente alto e é capaz de analisar somente um trecho bastante pequeno, tornando impraticável a aplicação deste método em estudos pangenômicos.

Após a amplificação, os fragmentos resultantes são rotulados com uma molécula fluorescente e são separadas duas porções deste material. Em uma delas é adicionada a proteína de interesse enquanto a outra é reservada para posterior comparação (controle). O agente de clivagem é então adicionado em ambas as porções, permitindo que ele corte o DNA em várias posições aleatórias. Além destes cortes aleatórios com a DNase I, são realizados cortes em regiões especificadas anteriormente com enzimas de restrição, para permitir a análise posterior. Em seguida, o DNA contendo a proteína e o DNA controle são colocados numa cuba para realização de uma eletroforese com gel de poliacrilamida. Nesse experimento, DNA é colocado sobre um gel sobre o qual é aplicada uma diferença de potencial. Pelo fato de o DNA ser eletronegativo ele irá migrar para o outro lado da cuba, porém os fragmentos menores irão migrar mais rapidamente por passarem mais facilmente entre os poros do gel. Após a eletroforese, é aplicado algum agente que possibilite visualizar o marcador fluorescente (como luz ultravioleta).

## **2.3. IDENTIFICAÇÃO DE SÍTIOS DE LIGAÇÃO DE FATORES DE TRANSCRIÇÃO**

---

A distribuição dos fragmentos assemelha-se a uma escada, com os fragmentos menores mais próximos da extremidade negativa da cuba e os fragmentos maiores, mais próximos da origem, na extremidade positiva. As amostras com a proteína de interesse e de controle são então comparadas. Pelo fato de a enzima DNase I não ser capaz de cortar o DNA em regiões onde se encontram outras proteínas ligadas, fragmentos com o tamanho exato produzido, caso a DNase tivesse cortado aquela região, não estarão presentes na amostra que a enzima de interesse foi aplicada, porém estarão presentes na outra amostra. Portanto a falta de bandas na amostra de interesse em uma região onde houve presença de bandas fluorescentes na amostra de controle sinaliza que a proteína de interesse estava ligada naquela região. A esta região é dado o nome de *footprint*. A Figura 2.10 detalha este processo de forma visual.

Obviamente, o processo é muito mais complexo do que o descrito neste texto. Etapas adicionais incluem o tratamento apropriado dos fragmentos obtidos, como a inserção de ligantes. Sua vantagem está no fato de que ele é realmente preciso e é capaz de encontrar as posições exatas onde a proteína estava ligada, com um grau de confiabilidade bastante alto. Sua desvantagem, como mencionado anteriormente, é que, por ser complexo e longo, ele definitivamente possui um baixo rendimento.

### **2.3.2 Imunoprecipitação da Cromatina**

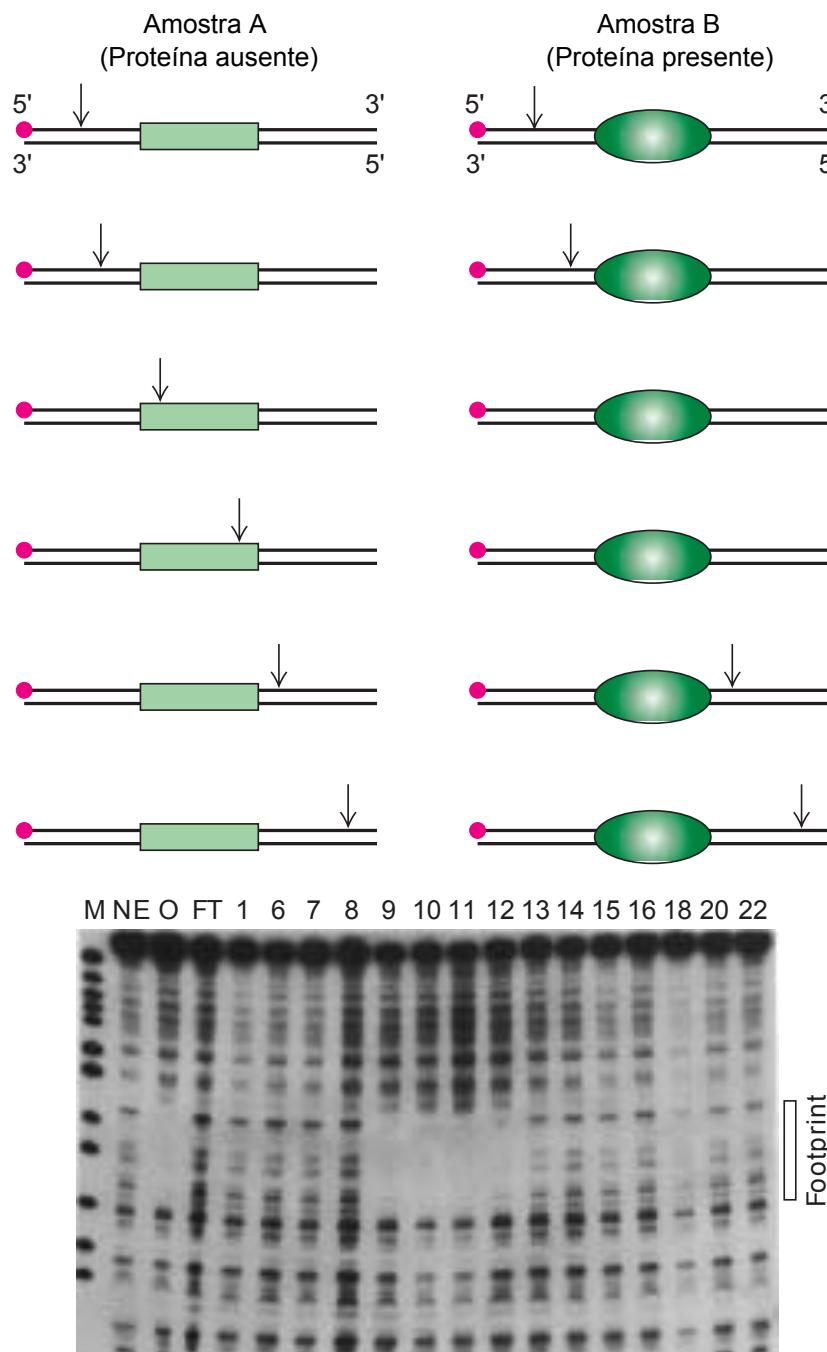
A imunoprecipitação da cromatina (ChIP, do Inglês *Chromatin Immunoprecipitation*) é uma técnica experimental utilizada para investigar as interações entre proteína-DNA na célula. O objetivo é identificar os locais exatos onde proteínas específicas, tais como fatores de transcrição, estão ligadas. Essa técnica também pode ser utilizada para se identificar proteínas com algum tipo de modificação pós traducional, como as modificações nas caudas das histonas.

De forma resumida o método funciona da seguinte forma: primeiramente a célula é quebrada para que se possa acessar o complexo DNA-proteína (cromatina). Esse complexo é clivado através de algum método (como sonicação, raios ultravioleta ou proteínas endonucleases) e os fragmentos contendo a proteína de interesse são extraídos através de imunoprecipitação. Neste método, é utilizado um anticorpo específico para a proteína de interesse para recuperar os complexos DNA-proteína fragmentados (Figura 2.11). Tais fragmentos possuem tamanho médio de 200 bp, porém isso varia bastante de acordo com a abordagem utilizada.

A partir disso, o DNA é purificado e os fragmentos resultantes podem ser determinados através de métodos semelhantes aos descritos para o método de *DNase I Footprinting* (basicamente, PCR com eletroforese em seguida, com algumas diferenças no tratamento dos complexos). As coordenadas genômicas recuperadas estarão associadas à proteína de interesse. É importante observar que, enquanto no método de *DNase I Footprinting* as regiões de depleção de digestão de DNase I são as regiões de interesse, no método de ChIP as regiões enriquecidas

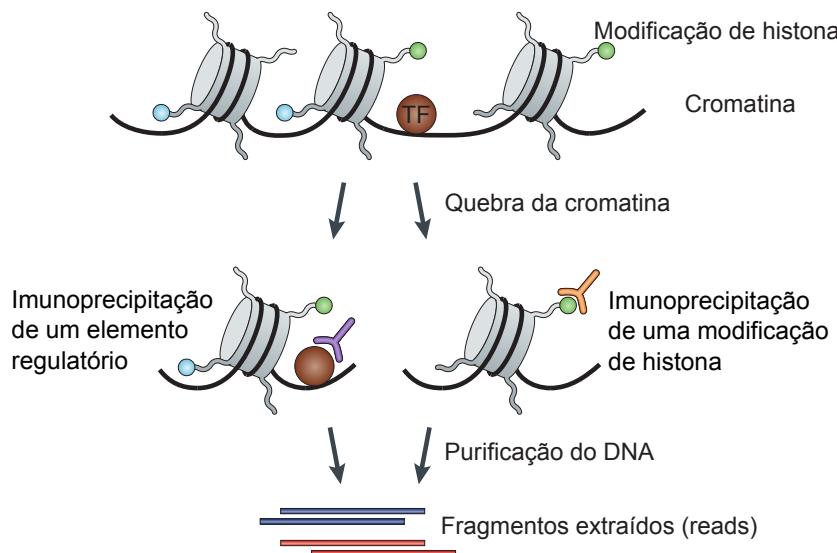
## 2. CONTEXTUALIZAÇÃO BIOLÓGICA

---



**Figura 2.10:** Esquema do método *DNase I Footprinting* - A amostra (A) não contém a proteína de interesse, enquanto a amostra (B) contém tal proteína (parte de cima da figura). Ao aplicar a enzima DNase I, todo o comprimento da amostra (A) será digerido enquanto que a região que contém a proteína na amostra (B) não será digerida. Essa depleção na atividade digestiva se mostra como um intervalo sem sinal fluorescente, nos resultados da eletroforese (parte de baixo da figura). Fonte: [Lodish *et al.*, 2007]

## 2.3. IDENTIFICAÇÃO DE SÍTIOS DE LIGAÇÃO DE FATORES DE TRANSCRIÇÃO



**Figura 2.11: Esquema do método ChIP** - Este simples esquema exibe as duas possibilidades de aplicação do método de ChIP: proteínas (como elementos regulatórios) ou proteínas modificadas (como histonas). Fonte: [Park, 2009]

são as buscadas. Além disso, vale a pena enfatizar que no método descrito na subseção anterior, os resultados representam, dentro da região onde o método é aplicado, todos os possíveis sítios de ligação DNA-proteína (sem especificar quais são as proteínas que se ligam nestas regiões), enquanto que no método de ChIP, apenas os sítios onde uma proteína de interesse estava ligada são identificados.

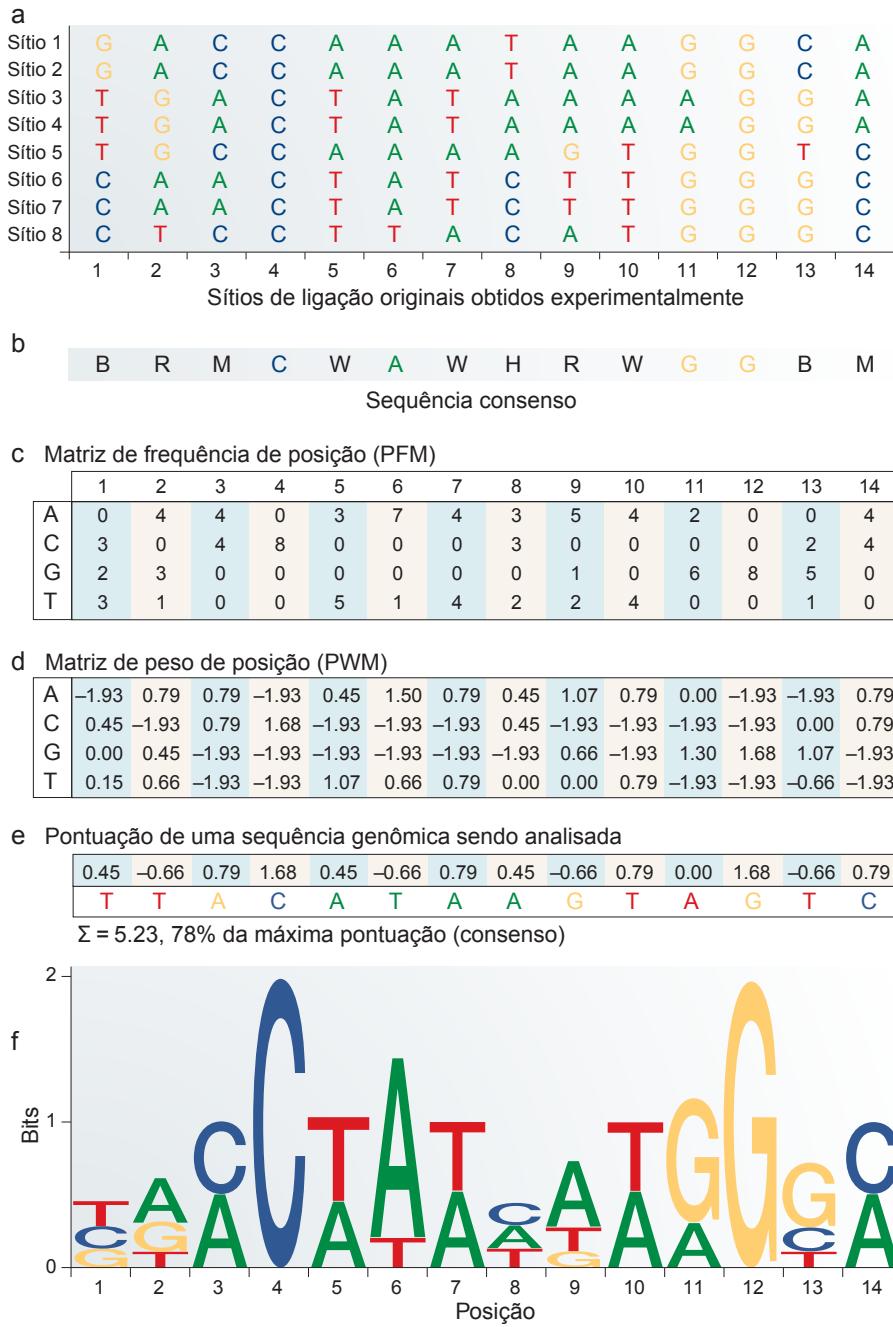
### 2.3.3 Motif Matching

Conforme mencionado anteriormente, ambos *DNase I Footprinting* e ChIP são métodos com baixo rendimento, isto é, são capazes de analisar apenas um pequeno trecho do genoma a cada rodada. Isso faz com que a aplicação de tais métodos seja financeiramente e tecnicamente custosa. Com a crescente demanda por métodos que consigam analisar o genoma inteiro, algumas abordagens computacionais baseadas em busca por sequência se tornaram bastante comuns. Será descrito o *Motif Matching* (MM), método que se baseia em análises biológicas em primeira mão para a geração de estruturas capazes de serem aplicadas através de meios puramente computacionais, ao longo de todo o genoma e com complexidade que permite sua aplicação em diversos genomas em um curto período de tempo.

O algoritmo toma como entrada um genoma (sequência de nucleotídeos) e uma matriz de pontuação, específica por fator a ser estudado, que será definida a seguir (ver esquema completo na Figura 2.12). O primeiro procedimento para gerar tal matriz consiste na obtenção de diversos fragmentos onde o elemento regulatório alvo se liga. Isso pode ser feito através de vários métodos

## 2. CONTEXTUALIZAÇÃO BIOLÓGICA

---



**Figura 2.12: Método para gerar PWMs** - (a) Sítios de ligação são obtidos experimentalmente e alinhados. (b) Os sítios obtidos contêm boas estimativas sobre a preferência de ligação da proteína em questão, o que pode ser visto através da seqüência consenso. (c) A PFM é criada através da contagem de nucleotídeos em cada posição. (d) Uma PWM é criada a partir da PFM através do modelo descrito pelas Equações 2.1 e 2.2. (e) Dada uma nova seqüência, uma pontuação pode ser avaliada a partir da PWM. (f) Gráficos baseados em entropia (ou *logos*) são representações visuais comuns dessas matrizes de posição. Fonte: [Wasserman & Sandelin, 2004]

### **2.3. IDENTIFICAÇÃO DE SÍTIOS DE LIGAÇÃO DE FATORES DE TRANSCRIÇÃO**

---

biológicos que fogem ao escopo deste trabalho (*DNase I Footprinting* e ChIP são alguns deles). De posse desses fragmentos, eles são alinhados e as posições que são importantes para a ligação DNA-proteína são aproximadas. Uma primeira matriz, chamada de matriz de frequência de posição (PFM, do Inglês *Position Frequency Matrix*) [Wasserman & Sandelin, 2004] é criada da seguinte forma: as linhas  $i = \{A, C, G, T\}$  correspondem a cada um dos 4 nucleotídeos do DNA e as colunas  $j = 1, 2, \dots, N$ , onde  $N =$  comprimento total do *motif*, correspondem a cada posição deste *motif* alinhado. Cada entrada  $X_{ij}$  da matriz corresponde à quantidade de nucleotídeos do tipo  $i$  na posição  $j$  do conjunto de fragmentos alinhados. Quanto mais sequência tivermos obtido inicialmente, mais confiável será essa estimativa da afinidade no DNA para esta proteína específica.

A partir de uma PFM, é comum serem criadas representações logarítmicas chamadas matrizes de peso de posição (PWMs, do Inglês *Position Weight Matrices*) ou matrizes de pontuação específica por posição (PSSM, do Inglês *Position-Specific Scoring Matrices*, pronunciada *possums*) [Wasserman & Sandelin, 2004]. PWMs e PSSMs são termos usados como sinônimos neste trabalho, sendo o termo PWM usado com maior frequência. Vários métodos podem ser utilizados para criar PWMs a partir de PFMs. O mais comum consiste no cálculo da probabilidade corrigida  $p(i, j)$  de se encontrar a base  $i$  na posição  $j$ , isto é:

$$p(i, j) = \frac{f_{ij} + s(i)}{N + \sum_{i' \in \{A, C, G, T\}} s(i')} , \quad (2.1)$$

onde  $f_{ij}$  é a frequência da base  $i$  na posição  $j$  e  $s(i)$  é uma função simples de *pseudocounts*. Esta função normalmente gera pequenos valores para evitar probabilidade nula de eventos de ligação raros mas factíveis. Tal função é crucial quando a amostra de sequência de sítios de ligação usada para estimar a PWM é pequena, algo comum. A partir da probabilidade corrigida, as entradas  $W_{ij}$  da PWM podem ser calculadas por:

$$W_{ij} = \log_2 \frac{p(i, j)}{p(i)} , \quad (2.2)$$

onde  $p(i)$  é a probabilidade geral de fundo do carácter  $i$  (para o *motif*, região ou genoma inteiro).

A partir de uma PWM é possível calcular a probabilidade de ligação, em um genoma, do fator para o qual a PWM foi calculada. Para cada sequência contígua de nucleotídeos do genoma de tamanho  $N$  (comprimento do *motif*), pode ser calculado um *bit score*  $B$ . Existem várias formas de se calcular tal pontuação, sendo a mais simples delas a soma de todas as entradas  $W_{ij}$  para todos os nucleotídeos  $i$  da sequência, dadas as coordenadas genômicas  $j$ . Isso criará um *ranking* a respeito da probabilidade de ligação do fator em todas as sequências contíguas no genoma. Técnicas estatísticas podem ser aplicadas para determinar qual a pontuação de corte que determinará quais sequências representam sítios de ligação.

## **2. CONTEXTUALIZAÇÃO BIOLÓGICA**

---

Versões dessa técnica possuem taxas de acerto bastante razoáveis e suas complexidades computacionais superam bastante a tecnicidade dos métodos puramente biológicos, isto é, o MM é aplicável de forma pangenómica. Entretanto, esta técnica possui desvantagens bem críticas: (1) MM é incapaz de diferenciar sítios de ligação ativos ou inativos, produzindo sempre os mesmos resultados para todas as linhas celulares onde aplicada. (2) Apesar de serem boas representações, PWMs geralmente são pequenas e degeneradas. Isto se dá pelo fato de que a maioria dos *motifs* possuem comprimentos entre 6 – 12 bp, com especificidade de ligação (posições onde apenas uma base possui frequência alta) variando, em geral, entre 4 – 6 bp. Como consequência dos pontos (1) e (2), o número de falsos positivos é extremamente alto. (3) A análise biológica das sequências nos quais os fatores estão ligados faz com que seja difícil a criação de PWMs para todos os fatores possíveis, ainda mais pelo fato de que alguns ainda estão sendo estudados. (4) Alguns fatores se ligam no genoma por intermédio de outros (por exemplo, co-ativadores e co-repressores), de forma que a criação de PWMs para estes fatores é complexa. (5) A acurácia desta técnica depende bastante da forma como a PWM foi criada, do algoritmo utilizado para realizar o MM e método estatístico utilizado para determinar os verdadeiros TFBSSs. Tais variáveis podem mudar bastante entre fatores diferentes, tornando o desenho experimental bastante complexo.

### **2.4 Solução Epigenética**

Os problemas encontrados pelas técnicas computacionais baseadas em busca de sequências de afinidade estão sendo amenizados por novas técnicas que estão atualmente no estado da arte no que concerne a identificação de TFBSSs. Tais técnicas utilizam dados epigenéticos para encontrar regiões que contêm sítios de ligação atuantes no momento em que tais dados foram mensurados. Utilizando esta abordagem, é possível criar um mapa consistente dos sítios de ligação presentes em uma determinada linhagem celular ou dadas determinadas condições. De fato, vários estudos estão mostrando que tais mapas geram uma assinatura da cromatina bastante consistente e com múltiplas aplicações em diversos tipos de estudos [Barski *et al.*, 2007; Heintzman *et al.*, 2007; Hon *et al.*, 2009; Ramsey *et al.*, 2010; Shu *et al.*, 2011].

O sucesso da utilização de características epigenéticas é explicado através da hipótese da cromatina descondensada/condensada. Em algumas regiões, a cromatina se encontra em um estado altamente condensado (enovelado), formando uma estrutura compacta que impede o acesso da maquinaria regulatória (e de fatores trans-atuantes) às regiões cis-regulatórias. Entretanto, em outras regiões, a cromatina é encontrada em um estado menos enovelado, formando estruturas mais permissivas à ligação de proteínas. Fatores epigenéticos, como as modificações pós-traducionais nas caudas das histonas, estão sendo diretamente relacionadas a mecanismos

## **2.4. SOLUÇÃO EPIGENÉTICA**

---

de abertura ou fechamento da cromatina. Sabendo que os fatores de transcrição se ligam preferencialmente em regiões mais permissivas, a utilização de características epigenéticas, como as modificações de histonas, faz com que o espaço de busca por sítios de ligação ativos possa ser reduzido. Tal delineamento epigenético das regiões mais prováveis de conter um sítio de ligação ativo consiste não só em uma abordagem com fundamentos biológicos concretos, como facilita a aplicação de metodologias computacionais (tais como o *motif matching*).

O termo *epigenética* tem origem na observação de padrões de hereditariedade não-Mendelianos em vários organismos. Mutações Mendelianas clássicas resultam de diferenças nos alelos causadas por variações de diversos tipos na estrutura de DNA, que coletivamente definem os tratos fenotípicos e contribuem para a determinação das fronteiras entre as espécies. É bastante evidente que tais fronteiras sofrem pressão da seleção natural. Em contraste, estão fenômenos tais como a variação do crescimento embrionário, alterações de coloração por mosaico genético, inativação aleatória do cromossomo X, paramutação em plantas e vários outros, que podem se manifestar, por exemplo, da expressão de apenas um (dos dois) alelo [Allis *et al.*, 2007].

A partir da discussão realizada, epigenética pode ser definida como o estudo das variações hereditárias na expressão gênica ou fenótipo celular causadas por outros motivos que não as variações na sequência de nucleotídeos do DNA. A partícula *epi-* do grego, significa sobre, acima, exterior. Em resumo, esse termo se refere às modificações funcionais relevantes para o genoma que não envolvem uma mudança na sequência de DNA. Evidências conclusivas que suportam as hipóteses epigenéticas mostram que esses mecanismos habilitam a transferência de *experiências* entre gerações. De forma relacionada, esses eventos ainda seriam capazes de explicar as variações que ocorrem entre, por exemplo, gêmeos univitelinos.

Vários elementos podem compor as variações englobadas pela epigenética, entre eles estão as modificações pós-traducionais nas caudas das histonas e as histonas variantes, utilizadas neste trabalho. Além disso, neste projeto de pesquisa assumiu-se como verdadeira a hipótese do DNA aberto/fechado. Nas subseções a seguir serão definidos brevemente tais conceitos e também serão detalhados os métodos que possibilitam a obtenção de dados epigenéticos.

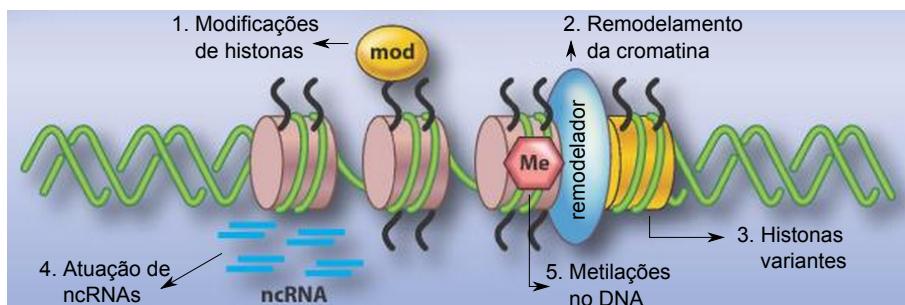
### **2.4.1 Conceitos e Elementos Epigenéticos**

Anteriormente foram definidos dois estados em que regiões da cromatina podem se apresentar: heterocromatina – estado de cromatina condensada, e eucromatina – estado de cromatina descondensada. Entretanto, estudos recentes sugerem que existe um espectro de estados da cromatina, sendo esta uma macromolécula com estrutura bastante dinâmica, propensa a remodelações e reestruturações à medida que recebe entradas relevantes das vias de sinalização. Esses diversos estados em que a cromatina se apresenta fornecem dicas importantes sobre as interações proteína-DNA que ocorrem em vizinhanças distintas.

## 2. CONTEXTUALIZAÇÃO BIOLÓGICA

---

A estrutura macromolecular da cromatina, bem como efeitos de ordens mais baixas como a disposição dos nucleossomos, pode ser alterada por fatores cis, fatores trans ou substituições de elementos do nucleossomo. A Figura 2.13 sumariza os principais elementos epigenéticos. Nela estão representados: (1) Modificações pós-traducionais de aminoácidos na cauda das histonas; (2) Remodelamento da cromatina através de processos dependentes de energia (ATP) que modificam o posicionamento dos nucleossomos; (3) A inserção ou remoção de histonas variantes; (4) Atuação de pequenos ncRNAs; (5) Metilação do DNA, geralmente em dinucleotídeos CpG fora de ilhas (definidas na Seção 2.2.2.1). Neste trabalho será dado foco apenas às modificações das histonas.

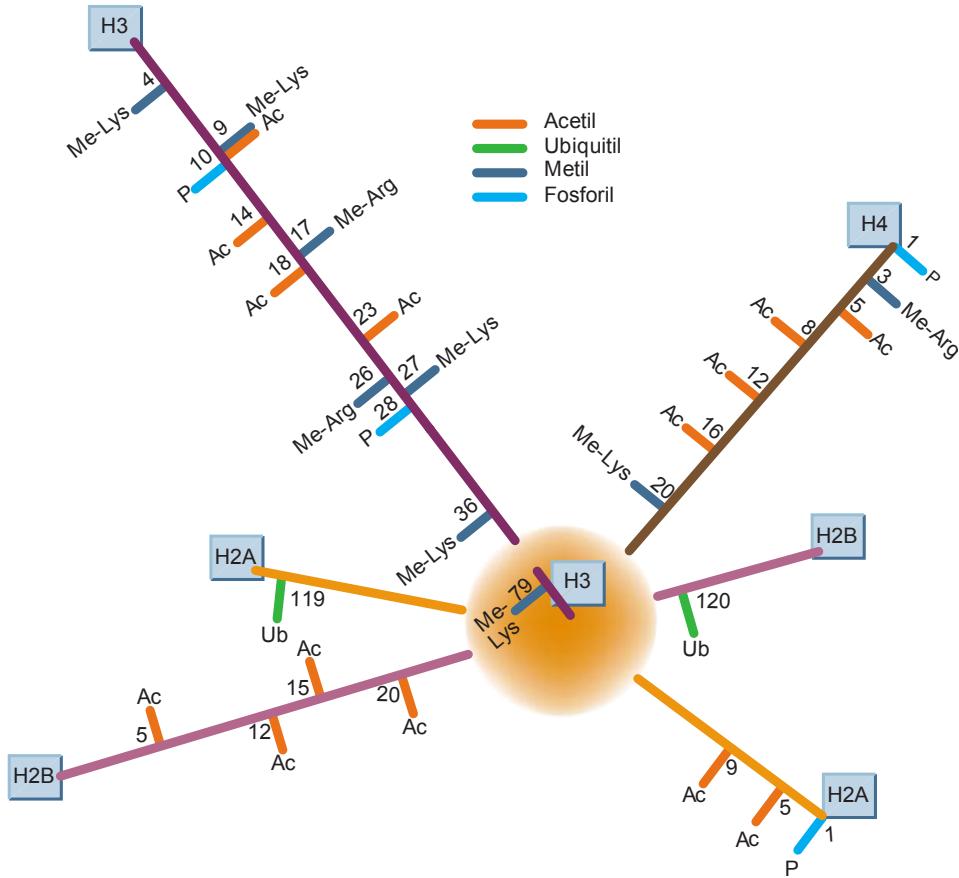


**Figura 2.13: Elementos epigenéticos** - Um esquema sumarizando os principais elementos epigenéticos. O objetivo desta figura é meramente ilustrativo e não representa toda a extensão da epigenética nem um esquema funcional de como tais elementos ocorrem. Fonte: [Allis *et al.*, 2007]

Um dos fatores mais estudados é a modificação pós-traducional na cauda das histonas. As caudas das histonas podem sofrer modificações químicas em aminoácidos específicos. Entre essas modificações estão a fosforilação, acetilação, metilação e ubiquitinação. Essas modificações possuem uma nomenclatura específica, seguindo a ordem: tipo da histona, aminoácido que sofre a modificação e tipo de modificação [Allis *et al.*, 2007]. Por exemplo, H3K4me2 se refere à dimetilação (me2) da lisina na posição 4 (K4) na cauda da histona H3. A Fig 2.14 mostra um mapa das principais modificações de histonas observadas até o momento.

O estudo mais aprofundado das modificações nas histonas e histonas variantes (neste texto, a nomenclatura será ocasionalmente extrapolada, sendo ambas chamadas de *modificações de histonas*) têm permitido maior entendimento sobre o impacto das mesmas na estrutura da cromatina e na expressão gênica [Grant, 2001; Spivakov & Fisher, 2007]. Alguns exemplos mais conhecidos são descritos na Tabela 2.1. Os padrões gerais de metilação e de acetilação são analisados com mais detalhes em [Barski *et al.*, 2007] e [Ramsey *et al.*, 2010], respectivamente. Por fim, algumas funções para modificações específicas ainda estão sendo estudadas, como por exemplo a modificação H3K27ac, que parece ser capaz de separar regiões amplificadoras ativas de regiões estacionárias [Creyghton *et al.*, 2010].

## 2.4. SOLUÇÃO EPIGENÉTICA



**Figura 2.14: Modificações de histonas** - Esquema gráfico representando as principais modificações de histonas detectadas até o presente momento. Fonte: [Felsenfeld & Groudine, 2003]

Entre as modificações mostradas, a H2A.Z, H3K4me2, H3K4me3 e H3K9ac parecem exibir forte capacidade de separar regiões de cromatina descondensada e condensada, como evidenciado em [Hon *et al.*, 2009; Won *et al.*, 2010] e nos estudos realizados internamente (mais detalhes na Seção 5.1). Por esta razão, neste estudo tais modificações nas histonas serão chamadas de modificações ativadoras, sendo as análises posteriores focadas neste grupo de modificações.

### 2.4.2 Métodos de Obtenção de Dados Epigenéticos

Sequenciamento de próxima geração (*Next-Generation Sequencing*) tem proporcionado meios para se realizar métodos biológicos tradicionais, baseados em eletroforese ou outra técnica de baixo rendimento, de forma pangenômica (isto é, com alto rendimento). A ideia básica consiste em substituir os procedimentos de baixo rendimento para obtenção das sequências de interesse (como a eletroforese para os métodos descritos nas Seções 2.3.1 e 2.3.2) por técnicas de sequenciamento de alto desempenho.

## 2. CONTEXTUALIZAÇÃO BIOLÓGICA

---

**Tabela 2.1: Impacto das modificações de histonas na estrutura da cromatina e expressão gênica.** Fonte: [Allis *et al.*, 2007]

Modificação	Impacto
H2A.Z	Suspensão dos genes para a iniciação da transcrição e prevenção de silenciamento da eucromatina.
H3K4me1	Ativação de transcrição. Relações com amplificadores foram identificadas.
H3K4me2	Eucromatina permissiva e ativação de transcrição.
H3K4me3	Eucromatina permissiva. Regiões de ponto de início da transcrição de genes que são transcrecionais iniciados, mas não necessariamente completamente transcritos.
H3K9ac	Ativação da transcrição e deposição de histonas.
H3K9me1	Silenciamento e repressão da transcrição.
H3K9me3	Altamente enriquecida em gene inativos. Relações com metilação no DNA foram identificadas.
H3K27ac	Ativação da transcrição. Relações com amplificadores foram identificadas.
H3K27me3	Inibição da transcrição.
H3K36me3	Associada a regiões transcritas. No corpo gênico, evita o início da transcrição em locais aberrantes.
H3K79me2	Alongamento da transcrição e ponto de verificação crítico no controle transcrecional.
H4K20me1	Heterocromatina e silenciamento da transcrição.

Existem várias técnicas de sequenciamento de alto desempenho, propostas por diferentes plataformas que comercializam seus sequenciadores. Entre elas estão: (1) sequenciamento massivo paralelo de assinaturas (MPSS, do Inglês *Massively Parallel Signature Sequencing*), que baseia-se em *esferas* e utiliza uma complexa abordagem de ligação e decodificação de adaptadores; (2) pirosequenciamento, que utiliza PCR de emulsão para amplificação e reação de DNA nascente com luciferase para identificar picos luminosos em rodadas revezadas de adição de nucleotídeos; (3) sequenciamento Illumina (Solexa), com amplificação via ponte e identificação de sequências via fotografias de nucleotídeos com rótulos fluorescentes. Esses são apenas alguns exemplos de uma quantidade imensa de técnicas. Cada método tradicional é adaptado mais facilmente com um subconjunto dessas técnicas, porém tais detalhes não serão abordados.

O método de DNase-seq [Crawford *et al.*, 2004; Song & Crawford, 2010] consiste na digestão de sequências de DNA com a enzima DNase I (conforme detalhado na Seção 2.3.1) e poste-

## **2.4. SOLUÇÃO EPIGENÉTICA**

---

rior identificação dos trechos através de sequenciamento de alto desempenho. Algumas etapas adicionais de tratamento de sequência são necessários, porém eles não adicionam graus muito mais elevados de tecnicidade ao método. A maior vantagem de tal abordagem se dá pelo fato de que agora é possível realizar o método de *DNase I Footprinting* ao longo de todo o genoma, obtendo resultados com alta resolução (na ordem de pares de bases) e acurados [Boyle *et al.*, 2008a, 2011]. Através deste método é possível medir locais onde a cromatina estava acessível, ou regiões hipersensíveis à DNase I. Além disso, é possível identificar regiões específicas onde proteínas estão ligadas ao DNA, porém sem especificar quais proteínas são estas.

O método de ChIP-seq [Park, 2009] consiste na realização do procedimento de imunoprecipitação da cromatina (ChIP – conforme detalhado na Seção 2.3.2) e posterior identificação das regiões enriquecidas para o tipo específico de proteína através de sequenciamento de alto desempenho. Assim como no método de DNase-seq, algumas etapas adicionais são necessárias entre as etapas mencionadas. Tal método é capaz de identificar, com boa resolução, regiões onde proteínas específicas se ligam no DNA. É importante observar que o método de ChIP-seq, por si só, já é capaz de identificar TFBSS com uma acurácia bastante alta, mas apenas para o caso de fatores de transcrição onde anticorpos que tenham alta afinidade de ligação com a proteína estejam disponíveis, o que se aplica apenas a um fração dos fatores de transcrição conhecidos. Em estudos onde é necessária a identificação dos sítios de ligação de uma pequena quantidade de fatores, tal método, quando disponível, representa a melhor opção atualmente. Porém, estudos atuais estão focando na identificação de assinaturas celulares, isto é, eles pretendem identificar o maior número de TFBSSs possível, para todos os fatores existentes. Em tais estudos, a aplicação de ChIP-seq é bastante complexa pois um experimento completo teria que ser realizado para todos os fatores que se tem conhecimento (ou para um grande número destes), processo que é altamente custoso e técnico. Porém, tal método também é capaz de identificar fatores epigenéticos como as modificações de histonas, o que fornece dados interessantes para direcionar a identificação total de TFBSSs sem que um grande número de experimentos seja conduzido.

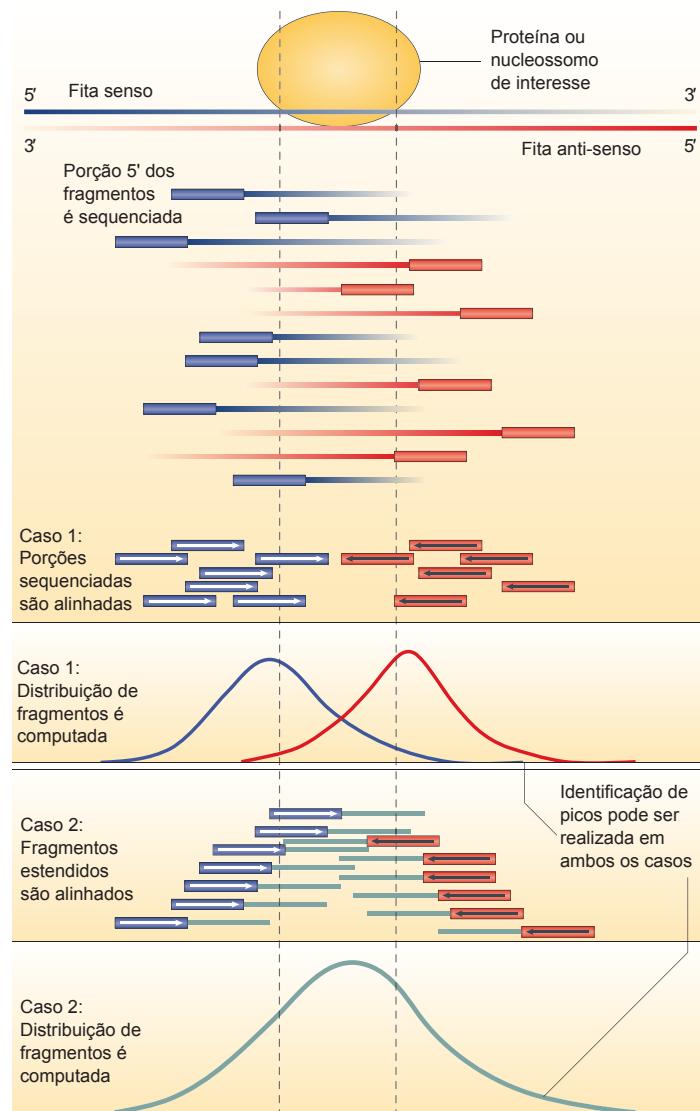
### **2.4.3 Geração de Sinais**

Os métodos descritos na Seção 2.4.2 resultam em diversas sequências de nucleotídeos dos locais recuperados. As tecnologias de sequenciamento de alto desempenho geralmente sequenciam apenas um pequeno número de bases a partir da posição 5'. O próximo passo então consiste em alinhar os fragmentos obtidos no genoma. Nesta etapa, alguns filtros podem ser impostos. É comum, por exemplo, descartar regiões que alinharam de forma significativa em 4 ou mais locais, devido a problemas gerados por regiões repetitivas. Alguns estudos também removem regiões onde vários fragmentos foram perfeitamente alinhados sem que qualquer fragmento tenha alinhado com regiões vizinhas, para excluir problemas devido à amplificação indevida ou outras fases específicas da técnica de sequenciamento utilizada [Boyle *et al.*, 2008b; Zhang *et al.*, 2008].

## 2. CONTEXTUALIZAÇÃO BIOLÓGICA

---

A partir das sequências (em Inglês, *reads*) alinhadas, podemos calcular um sinal genômico. Tal sinal consiste na simples contagem de quantos fragmentos se sobrepujaram em cada bp do genoma. A Figura 2.15 mostra duas abordagens comumente utilizadas. Na primeira, são considerados os *reads* inteiros provenientes tanto da fita senso quanto da anti-senso. Isso gera picos bimodais, que podem ser utilizados de forma diferenciada ou igualitária. Na segunda abordagem, os *reads* são estendidos englobando todo o trecho onde o fator de interesse esteve presente, gerando apenas um pico onde as regiões mais altas representariam os trechos enriquecidos [Park, 2009].



**Figura 2.15: Geração de Sinais Genômicos** - O esquema mostra duas abordagens possíveis (entre diversas abordagens existentes): No primeiro caso o sinal é gerado a partir dos fragmentos originais que foram sequenciados (o tamanho varia de acordo com o método de sequenciamento utilizado). No segundo caso, o fragmento é modificado (neste caso estendido) para atender a algumas características técnicas, como o fato de que os fragmentos obtidos através de ChIP têm, em média, 200 bp (muito maior do que os fragmentos sequenciados). Fonte: [Park, 2009]

## **2.5. REVISÃO DA LITERATURA**

---

A Figura 2.15 representa a geração de sinais genômicos para o método ChIP-seq. Algumas particularidades a respeito dos métodos DNase-seq e ChIP-seq serão explorados na seção onde os métodos deste estudo são detalhados. Por ora, apresenta-se apenas o fato de que é comum gerar sinais epigenéticos com esses dois métodos de forma que a representação das regiões enriquecidas seja bem diferente. No caso do DNase-seq, é comum considerar apenas o bp da extremidade 5' para gerar os sinais. Foi demonstrado que com tal abordagem, as regiões de TFBSSs são representadas como trechos de depleção de sinal, após trechos de picos [Boyle *et al.*, 2011]. Esse sinal é dito possuir alta resolução pois como apenas um bp foi utilizado para calcular as sobreposições, os sinais tendem a mostrar regiões bastante específicas, delineando picos bem claros das regiões exatas onde a DNase I digeriu o DNA. No caso do ChIP-seq o sinal possui uma resolução um pouco mais baixa, já que é comum que os fragmentos sequenciados e alinhados sejam estendidos até o tamanho médio dos fragmentos obtidos através do método de ChIP. Consequentemente, a proteína de interesse poderia estar ligada em quase toda a região estendida. Nesse caso, que se assemelha à segunda abordagem descrita anteriormente, as regiões enriquecidas seriam representadas como picos, estando a proteína de interesse, ligada em alguma região dentro destes picos.

## **2.5 Revisão da Literatura**

Este projeto é parcialmente baseado em um estudo recente por Boyle et al., onde várias propriedades relativas aos resultados do DNase-seq foram discutidas e TFBSSs foram preditos utilizando dados de DNase-seq e um modelo probabilístico [Boyle *et al.*, 2011]. Dois resultados em especial são interessantes. Primeiro, foi demonstrado que as regiões de depleção em sinais gerados a partir de DNase-seq são ótimos preditores de regiões de TFBSSs. Análises estatísticas mostram que a significância de tais regiões está bastante relacionada com o nível de enriquecimento de técnicas para fatores específicos como ChIP-seq ou com pontuações de técnicas como MM. Outra parte do estudo consistiu na criação de um modelo escondido de Markov simples univariado para a identificação automática de regiões de TFBSSs a partir dos sinais de digestão de DNase I. Este estudo foi replicado e várias características, como o conjunto de validação, foi seguido de forma idêntica, para permitir uma comparação com máxima precisão.

Outros estudos que se baseiam em DNase foram publicados [Boyle *et al.*, 2008a; Crawford *et al.*, 2004, 2006a,b; He *et al.*, 2012; Song & Crawford, 2010; Song *et al.*, 2011]. Crawford et al. [Crawford *et al.*, 2004] utilizaram padrões de digestão de DNase I para recuperar regiões hipersensíveis e mostrou que essas regiões são bons preditores de sítios de ligação ativos no estado corrente da célula. Tal técnica serve como hipótese central para diversos outros estudos baseados na identificação específica de tais regiões. A partir do sucesso de tal protocolo, ele foi devidamente formalizado em Song e Crawford [Song & Crawford, 2010]. Mais recentemente,

## **2. CONTEXTUALIZAÇÃO BIOLÓGICA**

---

estudos como He et al. [He *et al.*, 2012] estão mostrando, através de padrões em regiões de hipersensibilidade à DNase I, que as estruturas da cromatina realmente são bastante variáveis, por uma grande quantidade de características, e além de específicos por célula, parecem ser específicos por elementos regulatórios ou módulos regulatórios.

Em relação à abordagens mais integrativas, isto é, que utilizaram várias fontes de dados epigenéticas em um só modelo (assumindo dependência ou não), alguns algoritmos provaram ser mais eficazes do que aqueles baseados apenas em DNase-seq [Cuellar-Partida *et al.*, 2012; Ernst & Kellis, 2010; Pique-Regi *et al.*, 2011; Whitington *et al.*, 2009; Won *et al.*, 2010]. Talvez o método mais simples entre as abordagens integrativas seja a busca por ocorrências de um *motif* específico utilizando filtros determinísticos baseados em modificações de histonas [Whitington *et al.*, 2009]. Vários outros métodos integrativos foram propostos, de forma a combinar *motifs* no DNA com informações a respeito da estrutura da cromatina [Ernst & Kellis, 2010; Won *et al.*, 2010]. Pique-Regi et al. [Pique-Regi *et al.*, 2011], criaram um modelo bem utilizado chamado CENTIPEDE, que utiliza um modelo de mistura Bayesiana hierárquico que incorpora informações sobre a sequência de DNA, a conservação evolucionária, a distância do sítio de início de transcrição (TSS), hipersensibilidade à DNase I e marcas de histona ativadoras e repressoras.

Ainda a respeito dos modelos integrativos, Cuellar-Partida et al. [Cuellar-Partida *et al.*, 2012] combinaram dados relativos às modificações de histonas H3K4me1, H3K4me3, H3K9ac, H3K27ac e digestão de DNase I para criar um modelo Bayesiano simples, baseado em razões logarítmicas de probabilidade posterior. Foi mostrado que este modelo simples consegue melhorar o desempenho em relação a modelos mais complexos como o CENTIPEDE ou os modelos propostos em [Ernst & Kellis, 2010; Whitington *et al.*, 2009; Won *et al.*, 2010]. Consideramos a validação realizada por estes estudos levemente divergentes da metodologia do Boyle et al. [Boyle *et al.*, 2011], não possibilitando a comparação direta.

Finalmente, pesquisas recentes têm focado na busca por padrões epigenéticos (tais como as modificações de histonas) em diferentes linhas celulares, condições e padrões de expressão. De fato, diversos estudos mostram claras assinaturas da cromatina e sugeriram a aplicação de tais padrões em diversos problemas, incluindo a predição de sítios de ligação [Barski *et al.*, 2007; Heintzman *et al.*, 2007; Hon *et al.*, 2009; Ramsey *et al.*, 2010]. Estudos que compararam as diferentes fontes de dados epigenéticas também são interessantes e elucidam várias questões sobre a dependência de uma sobre outra [Shu *et al.*, 2011].

### **2.6 Considerações Finais**

Neste capítulo, foi realizada uma revisão sobre os principais conceitos de Biologia Molecular, Genética, epigenética e regulação gênica. A partir desse conhecimento, o problema de iden-

## **2.6. CONSIDERAÇÕES FINAIS**

---

tificação de sítios de ligação de fatores de transcrição foi delineado, deixando bem claras as fronteiras e níveis de dificuldade diferentes em diversas abordagens do problema. Foi mostrado que sinais epigenéticos estão sendo utilizados para melhorar a predição de TFBSS e que sequenciamento de próxima geração permite a mensuração de tais dados de forma pangenômica. Por fim, uma discussão sobre os estudos situados no estado da arte foi realizada, apontando as semelhanças e diferenças com a forma como o problema será abordado neste projeto de pesquisa.

# 3

## Modelos Escondidos de Markov

Neste capítulo, será descrito o método de aprendizagem de máquina que será aplicado posteriormente ao problema de identificação de TFBSS: o Modelo Escondido de Markov. Outros métodos matemáticos serão utilizados durante o processamento dos sinais epigenéticos e em outras etapas, porém apenas este método será exibido por fazer parte do núcleo deste estudo. Os modelos escondidos de Markov (HMMs, do Inglês *Hidden Markov Models*), é uma técnica probabilística baseada na teoria de Bayes e em processos estocásticos de Markov. Serão abordados algoritmos de predição e estimativa de parâmetros baseados em HMMs. Não necessariamente todos os algoritmos mostrados serão utilizados, sendo estes exibidos por motivos didáticos. Toda teoria exibida será baseada nos livros e artigos [Bilmes, 1997; Bishop, 2006; Duda *et al.*, 2000; Durbin *et al.*, 1998; Dymarski, 2011; Hair *et al.*, 1998; Hastie *et al.*, 2009; Lesk, 2005; Levin *et al.*, 2008; Mitchell, 1997; Rabiner, 1989; Russell & Norvig, 2002], onde mais informações podem ser obtidas.

A área de aprendizagem de máquina é uma ramificação da grande área de inteligência artificial, dentro da ciência da computação. Essa disciplina tem como objetivo a análise de dados provenientes das mais diversas fontes de modo a realizar inferências sobre tais dados. A tarefa de inferência mais comum é a classificação, onde um método é treinado de forma a capturar características de interesse a partir de padrões existentes nos dados utilizados e, após esse treinamento, é capaz de classificar novos padrões com base no que *aprendeu*. Esse treinamento pode seguir diversos paradigmas, entre eles estão a aprendizagem supervisionada, não-supervisionada e por reforço.

Na aprendizagem supervisionada, os exemplos (ou instâncias) são mostrados ao algoritmo, juntamente com as *respostas* ou classe de cada instância. O treinamento é dito supervisionado pois o classificador tem completo conhecimento das classes da amostra de dados de treino e deve aprender baseado nesta característica. Na abordagem não-supervisionada, o algoritmo

### 3.1. MODELOS ESCONDIDOS DE MARKOV

---

recebe as instâncias dos dados sem suas respectivas classes. O objetivo é encontrar padrões em comum entre múltiplas instâncias, criando sua própria categorização (isto é, separação dos dados) interna com base nessas características intrínsecas. O método HMM descrito irá conter algumas instâncias teóricas supervisionadas e não supervisionadas, porém apenas as técnicas supervisionadas serão utilizadas no projeto.

## 3.1 Modelos Escondidos de Markov

Cadeias de Markov são modelos probabilísticos compostos por uma coleção de estados e uma coleção de transições entre esses estados, que correspondem à probabilidade da mudança de um estado para o outro. Os modelos escondidos de Markov seguem esta mesma ideia, porém neles, além da sequência de estados conhecida, existe uma sequência de estados, chamada de caminho (em inglês, *path*), que não é conhecida e cada estado emite símbolos conhecidos (que fazem parte de um alfabeto  $\Sigma$ ) a partir de uma determinada probabilidade. O objetivo deste modelo é, considerando a sequência de estados conhecida como sendo uma sequência de “emissões” de símbolos dentro de um alfabeto específico, determinar qual é a sequência de estados mais provável de ter gerado esta sequência de símbolos.

Os HMMs são formalizados a seguir. Um modelo escondido de Markov consiste em: (1) um conjunto de estados  $S = \{S_1, S_2, \dots, S_n\}$ ; (2) uma matriz  $A$  de dimensões  $n \times n$  onde cada célula  $a_{ij}$  dessa matriz representa a probabilidade de se transitar do estado  $i$  para o estado  $j$ ; (3) uma matriz  $E$  de tamanho  $|\Sigma| \times n$  onde cada entrada  $e_i(b)$  representa a probabilidade de se emitir, no estado  $i$ , a entrada observada  $b \in \Sigma$ . Esse modelo recebe como entrada uma sequência  $x = x_1x_2\dots x_L$  de observações e possui uma instância especial  $\pi = \pi_1\pi_2\dots\pi_L$ , onde  $\pi_i \in S$ , chamada caminho (ou sequência de estados escondidos), que pode assumir o papel de entrada ou saída do algoritmo dependendo dos objetivos da prova ou modelagem que se deseja obter. A Figura 3.1 sumariza essas definições de forma gráfica. Modelos gráficos deste gênero serão utilizados mais adiante quando soluções para o problema de predição de TFBs forem propostas.

Realizadas as definições iniciais sobre os parâmetros e entradas do modelo, podemos formalizar de maneira probabilística o conceito de *transição* e *emissão*, respectivamente, segundo as Equações 3.1 e 3.2. Tais definições correspondem à base de todos os resultados subsequentes e devem ser entendidos como cláusulas básicas para a teoria dos HMMs.

$$a_{kl} = P(\pi_i = l | \pi_{i-1} = k) \tag{3.1}$$

### 3. MODELOS ESCONDIDOS DE MARKOV

---

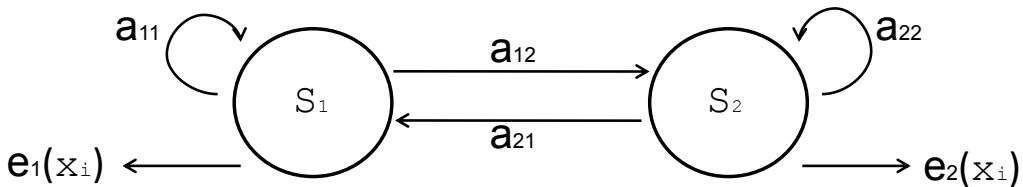
$$\begin{array}{lll} \text{Conjunto de estados} & \text{Conjunto de transições} & \text{Conjunto de emissões} \\ \mathbf{s} = \{s_1, s_2\} & \mathbf{a} = \{a_{11}, a_{12}, a_{21}, a_{22}\} & \mathbf{e} = \{e_1, e_2\} \end{array}$$

Observação

$$\mathbf{x} = x_1 x_2 x_3 \dots x_L$$

Estados escondidos

$$\boldsymbol{\pi} = \pi_1 \pi_2 \pi_3 \dots \pi_L$$



**Figura 3.1: Esquema de um modelo escondido de Markov** - Neste esquema exemplo, existem 2 estados  $S_1$  e  $S_2$ . Cada um dos dois estados possui transição para si e para o outro estado. A emissão de cada estado, isto é  $e_1(x_i)$  e  $e_2(x_i)$ , correspondem a probabilidades pontuais atribuídas a cada possível valor  $x_i$ . Observe que a matriz de transição está representada em sua forma vetorial para facilitar a visualização.

$$e_k(b) = P(x_i = b | \pi_i = k) \quad (3.2)$$

Além das ações básicas de transição e emissão, a teoria dos HMMs possui uma propriedade chave: a probabilidade de prosseguir do estado  $i$  para o estado  $i + 1$  depende apenas da probabilidade no estado  $i$ . Dessa forma, o processo estocástico faz com que as probabilidades sejam summarizadas em cada estado, de forma indutiva. Podemos generalizar a propriedade chave como: a probabilidade de prosseguir do estado  $i$  para o estado  $i + 1$  depende apenas da probabilidade dos  $T$  estados anteriores, definindo um HMM de ordem  $T$ . Utilizando um estado auxiliar inicial 0, no qual o modelo se encontra no início do processo, e um estado auxiliar final  $L + 1$  (também denotado posteriormente como  $\epsilon$ ), no qual o modelo se encontra no fim do processo, podemos representar esse conceito chave, para o caso de ordem 1, segundo a Equação 3.3.

$$P(x, \pi) = a_{0\pi_1} \prod_{i=1}^L e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}} \quad (3.3)$$

Podemos definir as emissões como discretas ou contínuas. A diferença não irá afetar a modelagem teórica a seguir, pelo fato de que: no caso discreto, basta que as probabilidades (denotadas por  $P(\cdot)$ ) sejam funções (massa) de probabilidade; em contrapartida, no caso contínuo, as probabilidades  $P(\cdot)$  seriam funções densidade de probabilidade. Os sinais utilizados neste projeto são de natureza contínua, portanto as emissões irão corresponder a distribuições gaussi-

### **3.2. MÉTODOS DE PREDIÇÃO BASEADOS EM HMMS**

---

anas (Equação 3.4). Isto significa que cada emissão, em cada estado, será representada através dos parâmetros de uma função densidade de probabilidade do tipo normal: a média  $\mu$  e o desvio padrão  $\sigma$ .

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty, \sigma > 0 \quad (3.4)$$

Além de contínuos, os modelos escondidos de Markov podem ser multivariados. Novamente, o formalismo a seguir se modificará apenas no que concerne à adição de dimensões. No modelo, a única diferença seria que a matriz de emissões  $E$  teria uma dimensão adicional de tamanho  $d$ , onde  $d$  é a dimensionalidade do modelo (isto é, a quantidade de sinais que serão simultaneamente inseridos). A entrada  $e_{ij}(b)$  desta matriz com três dimensões representaria a emissão para o  $i$ -ésimo estado, para o  $j$ -ésimo sinal, para um valor observado  $b$ .

Os algoritmos apresentados na Seção 3.2 consistem em métodos para se descobrir o caminho  $\pi$  a partir de uma sequência de caracteres  $x$  utilizando um modelo com os parâmetros  $A$  e  $E$  definidos. Nesses métodos a Equação 3.3 será explorada e serão criadas novas variáveis para ajudar no entendimento. Os algoritmos apresentados na Seção 3.3 mostram formas de se estimar os parâmetros  $A$  e  $E$  para um modelo de Markov escondido, de forma supervisionada ou não supervisionada.

## **3.2 Métodos de Predição Baseados em HMMs**

Dado o formalismo definido na Seção 3.1, existem, basicamente, três problemas que devem ser resolvidos para que o modelo tenha aplicações práticas:

- Problema 1** Dada a sequência observada  $x = x_1 x_2 \dots x_L$  e um modelo composto por  $\theta = \{A, E\}$ , como é escolhida a sequência  $\pi = \pi_1 \pi_2 \dots \pi_L$  que é ótima dado algum critério significativo (isto é, que melhor explica as observações)?
- Problema 2** Dada a sequência observada  $x = x_1 x_2 \dots x_L$  e um modelo composto por  $\theta = \{A, E\}$ , como é computada  $P(x|\theta)$ , isto é, a probabilidade da sequência observada, dado o modelo?
- Problema 3** Como os parâmetros  $\theta = \{A, E\}$  podem ser ajustados de forma a maximizar  $P(x|\theta)$ ?

O primeiro problema proposto, que aborda a parte *escondida* do HMM, será abordado na Seção 3.2.1, ao definir o método de Viterbi. O segundo problema será utilizado para avaliar a probabilidade posterior na Seção 3.2.2 (correspondente, mais especificamente, aos métodos

### 3. MODELOS ESCONDIDOS DE MARKOV

---

*forward* ou *backward*). E finalmente, o terceiro problema, diretamente solucionado através do simples método da verossimilhança na Seção 3.3, faz com que sejamos capazes de treinar o modelo.

Nesta seção serão definidos os dois principais métodos para se predizer sequências de estados escondidos  $\pi$  a partir de um HMM e de entradas (sequências de símbolos  $x$ ). O primeiro método segue diretamente das definições anteriores, a partir da utilização do paradigma de programação dinâmica para resolver o problema da exaustão inicial. O segundo método resulta em um vetor de probabilidades posterior de tamanho igual ao número de estados, para cada elemento do vetor de entrada. Neste método, que geralmente produz previsões mais acuradas que o primeiro, o caminho  $\pi$  pode ser avaliado de várias formas, incluindo a aceitação do estado que possui a maior probabilidade posterior para cada posição da sequência de entrada.

#### 3.2.1 Algoritmo de Viterbi

Ao introduzir uma sequência de estados escondidos  $\pi$  no modelo, se torna impossível descrever deterministicamente em qual estado do modelo estamos apenas através da observação do símbolo correspondente da sequência de entrada  $x$ . Encontrar o significado da sequência de entrada em termos da sequência de estados escondidos se chama decodificação, no jargão original de reconhecimento de padrões sonoros.

O Algoritmo de Viterbi foi proposto por Andrew Viterbi, em 1976, como um algoritmo de decodificação para códigos convolucionais sobre conexões digitais de comunicação que continham alto nível de ruído. Após sua proposição, esse algoritmo foi aplicado em áreas como celulares digitais CDMA e GSM, modems discados, satélites, comunicações espaciais, redes sem fio 802.11 e atualmente, é bastante utilizado em reconhecimento de fala, linguística computacional e bioinformática.

O Algoritmo de Viterbi pertence ao paradigma da programação dinâmica e consiste em descobrir qual é o caminho mais provável  $\pi^*$  dada a sequência de emissão  $x$ . A Equação 3.5 descreve em termos formais essa proposição.

$$\pi^* = \operatorname{argmax}_{\pi} P(x, \pi) \quad (3.5)$$

A forma exaustiva de resolução de tal algoritmo seria calcular as probabilidades  $P(x, \pi)$  para todas as sequências  $\pi$  existentes. Entretanto, conforme aumentamos o tamanho  $L$  da sequência de entrada, o número total de combinações de estados que constituem as sequências  $\pi$  cresce exponencialmente, e quanto maior o número de estados, mais agressivo é tal crescimento.

### **3.2. MÉTODOS DE PREDIÇÃO BASEADOS EM HMMS**

---

Felizmente, Viterbi apontou uma solução baseada em programação dinâmica, onde o caminho mais provável  $\pi^*$  pode ser encontrado recursivamente.

Suponha que criemos variáveis de Viterbi  $v_k(i)$ , que correspondem à probabilidade do caminho mais provável do prefixo  $x_1 \dots x_i$  que termina no estado  $S_k$ . Supondo que tais probabilidades são conhecidas para todos os estados  $k$  podemos calcular essas probabilidades para o prefixo  $x_1 \dots x_{i+1}$  como descrito na Equação 3.6.

$$v_l(i+1) = e_l(x_{i+1}) \max_k(v_k(i)a_{kl}) \quad (3.6)$$

Dado que todas as sequências se iniciam em um estado inicial 0, podemos definir as variáveis de Viterbi para este estado inicial como  $v_0(0) = 1$  e  $v_k(0) = 0$  para todos os outros estados que não o inicial. A partir destas variáveis iniciais, podemos continuar calculando as variáveis dos próximos estados segundo a Equação 3.6 e manter um ponteiro  $ptr$  para os estados que possuíram a maior probabilidade em cada iteração. Tal algoritmo, que é possível dada a propriedade chave das cadeias de Markov, é definido a seguir:

---

#### **Algoritmo de Viterbi**

---

- 1. Inicialização:**
  - 1.1.  $v_0(0) = 1$
  - 1.2.  $v_k(0) = 0$  para  $k > 0$
- 2. Recursão ( $i = 1, \dots, L$ ):**
  - 2.1.  $v_l(i) = e_l(x_i) \max_k(v_k(i-1)a_{kl})$
  - 2.2.  $ptr_i(l) = \operatorname{argmax}_k(v_k(i-1)a_{kl})$
- 3. Terminação:**
  - 3.1.  $P(x, \pi^*) = \max_k(v_k(L)a_{k\epsilon})$
  - 3.2.  $\pi_L^* = \operatorname{argmax}_k(v_k(L)a_{k\epsilon})$
- 4. Remontagem ( $i = L, \dots, 1$ ):**
  - 4.1.  $\pi_{i-1}^* = ptr_i(\pi_i^*)$

Existem alguns problemas práticos de implementação em relação ao Algoritmo de Viterbi. O mais severo decorre do fato de que multiplicar diversas probabilidades baixas irá gerar números de ordens extremamente baixas, o que ocasiona em erros de estouro negativo (*underflow*) quando não tratado de forma correta. A solução mais utilizada consiste em realizar o algoritmo no espaço logarítmico, o que faria com que todas as multiplicações virassem somatórios. Esse tipo de detalhe foge ao escopo deste trabalho e não será abordado.

### 3. MODELOS ESCONDIDOS DE MARKOV

---

#### 3.2.2 Probabilidade Posterior

Além do Algoritmo de Viterbi, podemos realizar a decodificação através do cálculo da probabilidade posterior de estar em cada estado escondido, em cada posição da sequência de entrada. Extrair o conjunto mais provável de estados escondidos desta abordagem pode ser realizado de forma simples como observar qual estado possui a maior probabilidade posterior para cada posição da sequência, ou de formas mais complexas como fixar um ponto de corte para aceitação de estados escondidos baseado nestas probabilidades. Além de permitir a extração do conjunto mais provável de estados de uma forma mais elaborada, o cálculo das probabilidades posteriores permite que seja visualizada a forma como as transições estão ocorrendo. Por essas razões, geralmente esta abordagem é preferível em relação ao Algoritmo de Viterbi.

A probabilidade posterior pode ser definida mais formalmente como sendo a probabilidade de, em uma certa posição da cadeia de caracteres, observarmos o estado escondido  $k$ , dada a sequência observada. Pelo teorema de Bayes, é possível colocar essa proposição em termos matemáticos (Equação 3.7).

$$P(\pi_i = k|x) = \frac{P(x, \pi_i = k)}{P(x)} \quad (3.7)$$

Primeiramente, será focado o cálculo da cláusula  $P(x)$ , isto é, a evidência de uma certa cadeia de caracteres  $x$  dentro de todas as possibilidades de cadeias de tamanho  $L$ . Formalmente, isso pode ser definido em relação ao caminho segundo a Equação 3.8.

$$P(x) = \sum_{\pi} P(x, \pi) \quad (3.8)$$

O cálculo exaustivo da Equação 3.8 é impossível pois o número de caminhos cresce exponencialmente com o tamanho da sequência (conforme já foi visto no contexto de Viterbi). Porém podemos avaliar esta expressão com a mesma ideia de Viterbi mostrada, apenas modificando os passos de maximização por somatórios. Neste novo algoritmo a variável  $f_k(i)$ , chamada variável *forward*, é utilizada assim como a variável de Viterbi (Equação 3.9). A variável *forward* corresponde à probabilidade de observar a sequência  $x$  até (e incluindo)  $x_i$  de tal forma que  $\pi_i = k$ . A recursão utilizada pelo algoritmo é definida na Equação 3.10.

$$f_k(i) = P(x_1 \dots x_i, \pi_i = k) \quad (3.9)$$

### 3.2. MÉTODOS DE PREDIÇÃO BASEADOS EM HMMS

---

$$f_l(i+1) = e_l(x_{i+1}) \sum_k f_k(i) a_{kl} \quad (3.10)$$

O algoritmo é mostrado a seguir. Assim como o Algoritmo de Viterbi, este método está sujeito a estouros negativos. Tal problema não pode ser resolvido da mesma forma como a Equação 3.5 foi por conter somatórios em sua própria natureza. A solução está novamente em se trabalhar em um espaço logarítmico, porém utilizando abordagens mais complexas.

---

#### *Algoritmo Forward*

---

- 1. Inicialização:**
    - 1.1.  $f_0(0) = 1$
    - 1.2.  $f_k(0) = 0$  para  $k > 0$
  - 2. Recursão ( $i = 1, \dots, L$ ):**
    - 2.1.  $f_l(i) = e_l(x_i) \sum_k f_k(i-1) a_{kl}$
  - 3. Terminação:**
    - 3.1.  $P(x) = \sum_k f_k(L) a_{k\epsilon}$
- 

Continuando a busca pela probabilidade posterior, podemos explorar o termo  $P(x, \pi_i = k)$ . Ao aplicar a propriedade chave dos modelos de Markov, podemos realizar a decomposição demonstrada na Equação 3.11. A segunda linha desta equação ocorre porque tudo que ocorre depois do estado  $k$  depende apenas do que ocorre no estado  $k$ .

$$\begin{aligned} P(x, \pi_i = k) &= P(x_1 \dots x_i, \pi_i = k) P(x_{i+1} \dots x_L | x_1 \dots x_i, \pi_i = k) \\ &= P(x_1 \dots x_i, \pi_i = k) P(x_{i+1} \dots x_L | \pi_i = k) \end{aligned} \quad (3.11)$$

É bastante claro que o primeiro termo da segunda linha da Equação 3.11 corresponde à variável *forward*  $f_k(i)$  cujo cálculo foi apresentado anteriormente. Para calcular a probabilidade posterior precisamos apenas abordar o segundo termo da segunda linha da Equação 3.11. É possível, então, criar outra variável, chamada *backward*, para calcular o termo restante. Obviamente, essa variável é definida como na Equação 3.12.

$$b_k(i) = P(x_{i+1} \dots x_L | \pi_i = k) \quad (3.12)$$

Para calcular tais variáveis é mostrado o Algoritmo *backward* a seguir. Tal algoritmo é análogo ao *forward* porém ao invés de proceder do início da sequência até o ponto desejado, ele procede do fim da sequência até o ponto desejado.

### 3. MODELOS ESCONDIDOS DE MARKOV

---

---

#### *Algoritmo Backward*

---

- 1. Inicialização:**  
1.1.  $b_k(L) = a_{k\epsilon}, \forall k$
  - 2. Recursão ( $i = L - 1, \dots, 1$ ):**  
2.1.  $b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i+1)$
  - 3. Terminação:**  
3.1.  $P(x) = \sum_l a_{0l} e_l(x_1) b_l(1)$
- 

A partir dos Algoritmos *forward* e *backward* podemos calcular a probabilidade posterior conforme definida na Equação 3.7 através de uma simples substituição dos termos nesta equação pelas respectivas variáveis criadas (Equação 3.13). O termo  $P(x)$  nesta equação pode ser calculado através da aplicação de um dos algoritmos, *forward* ou *backward* na sequência inteira.

$$P(\pi_i = k|x) = \frac{f_k(i)b_k(i)}{P(x)} \quad (3.13)$$

### 3.3 Estimação de Parâmetros em HMMs

Na Seção 3.2, algoritmos para determinar a sequência de estados escondidos foram definidos. Nesta seção, será demonstrado um método para a criação de tais HMMs, isto é, a estimação dos parâmetros que compõem o HMM (a matriz de transições  $A$  e o vetor de emissões  $E$ ). A técnica escolhida, máxima verossimilhança, consiste na estimação mais simples possível. A ideia é que os parâmetros sejam o mais próximo possível dos observados nos dados de treinamento. Esta abordagem é, portanto, supervisionada. Caso fosse necessária a estimação de parâmetros um HMM sem informações de classe a priori, um método não-supervisionado como o Baum-Welch teria que ser utilizado. Neste método, estimativas são feitas através de aproximações baseadas no algoritmo de Maximização da Esperança (EM, em Inglês *Expectation Maximization*).

Como mencionado, podemos estimar os parâmetros de forma supervisionada ou não supervisionada. Entretanto, o modelo geral, isto é, a sequência de estados  $S$ , já deverá estar corretamente modelada. A criação de um modelo oscila bastante entre os que acreditam nesta tarefa como uma arte e naqueles que desenvolvem métodos específicos, geralmente baseados em duração probabilística dos estados. De qualquer forma, tal tarefa não será mencionada. Os modelos originais desenvolvidos neste trabalho foram idealizados com base nos padrões dos dados e sua robustez foi aferida de forma puramente empírica.

O método da máxima verossimilhança é a forma mais simples de se estimar os parâmetros  $A$  e  $E$  dos modelos escondidos de Markov. Neste tipo de estimação, é utilizada uma sequência de

### 3.3. ESTIMAÇÃO DE PARÂMETROS EM HMMS

---

símbolos  $x$  com sequência de estados conhecida  $\pi$  para calcular os parâmetros mais verossímeis.

Para o caso discreto, de forma intuitiva, será realizada a simples contagem do número de vezes em que acontece cada evento relacionado aos parâmetros. Denotando por  $A_{kl}$  o número de ocorrências de transições entre os estados  $k$  e  $l$  (não confundir com  $a_{kl}$ , que é a probabilidade desta transição), e  $E_k(b)$  o número de emissões do símbolo  $b$  no estado  $k$ , o estimador de máxima verossimilhança consiste na simples aplicação das Equações 3.14 e 3.15.

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \quad (3.14)$$

$$e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')} \quad (3.15)$$

Generalizando para o caso contínuo, tem-se uma função de densidade  $p(x|\Theta)$  governada pelo conjunto de parâmetros  $\Theta$ . No caso de uma gaussiana, por exemplo,  $\Theta$  corresponde às médias e desvios padrões das entradas utilizadas. Suponha que tenhamos também um conjunto de dados de tamanho  $T$  obtido a partir desta distribuição, isto é,  $X = \{X_1, \dots, X_T\}$ . A densidade resultante das amostras é dada pela Equação 3.16.

$$p(X|\Theta) = \prod_{i=1}^T p(X_i|\Theta) = L(\Theta|X) \quad (3.16)$$

Essa função  $L(\Theta|X)$  é chamada de verossimilhança dos parâmetros dado o conjunto de entradas  $X$ . De forma intuitiva, ela pode ser pensada como uma função dos parâmetros  $\Theta$  onde o conjunto de dados  $X$  se encontra fixo. No problema da máxima verossimilhança, o objetivo é encontrar o conjunto de parâmetros  $\Theta$  que maximize a função  $L$  (Equação 3.17).

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} L(\Theta|X) \quad (3.17)$$

Esse problema pode ser facilmente resolvido para o caso da gaussiana (onde  $\Theta = \{\mu, \sigma\}$ ), bastando igualar a derivada de  $\log(L(\Theta|X))$  a zero e resolver diretamente para  $\mu$  e  $\sigma$ . O motivo para o uso da função  $\log$  é que ela torna o problema analiticamente mais fácil. Para outras distribuições, entretanto, técnicas mais elaboradas são necessárias, dado que a solução para as expressões analíticas não podem ser encontradas diretamente. Tais detalhes não serão expostos, visto que neste projeto serão utilizadas apenas gaussianas para representar os sinais de entrada.

### **3. MODELOS ESCONDIDOS DE MARKOV**

---

#### **3.4 Considerações Finais**

Neste capítulo, foi descrita a técnica do modelo escondido de Markov sob a ótica do aprendizado de máquina. Primeiramente, foi mostrada a teoria dos modelos escondidos de Markov. Após uma introdução, foram abordadas as principais técnicas de decodificação (predição de estados escondidos a partir de observações) e estimativa de parâmetros. Esta técnica é a principal ferramenta deste estudo, aplicada diretamente aos sinais epigenéticos (observações) gerados pelos métodos descritos no capítulo anterior para predizer sítios de ligação de fatores de transcrição (estados escondidos).

# 4

## Metodologia

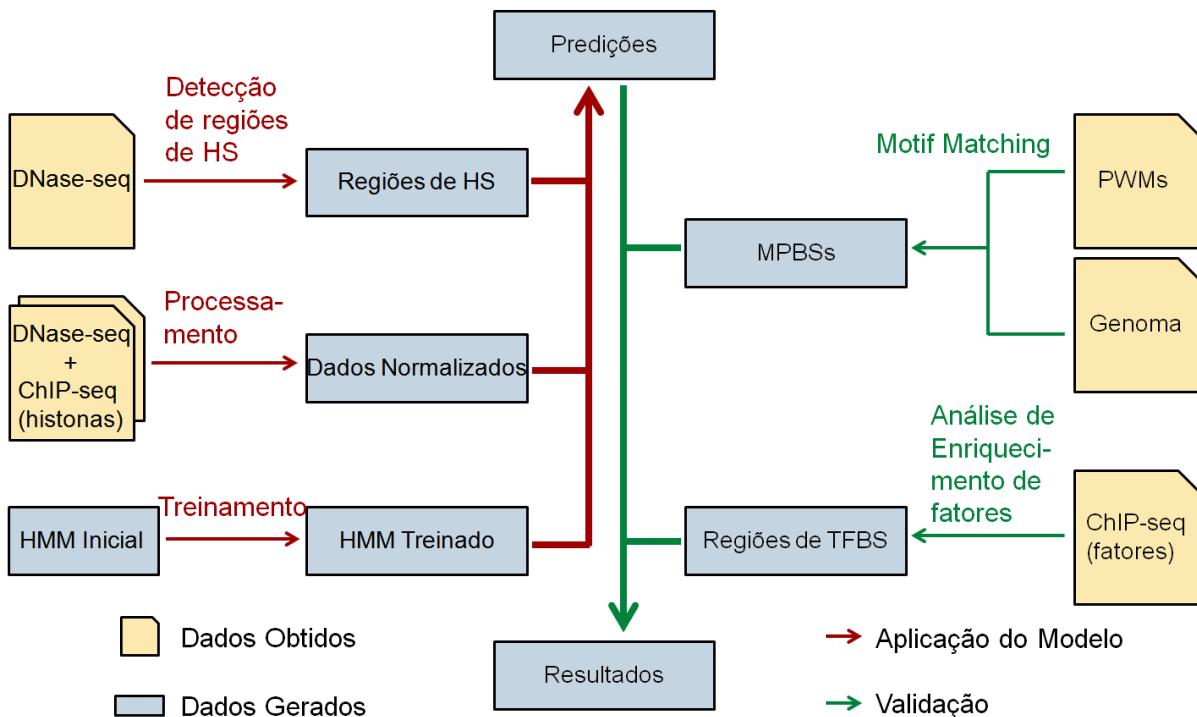
Neste capítulo, será descrita a forma como os experimentos foram realizados. Serão dados detalhes a respeito das bases de dados utilizadas e os repositórios onde elas foram obtidas. Então, todos os procedimentos realizados serão descritos, envolvendo: *motif matching*, análise de enriquecimento dos dados de digestão de DNase I (regiões hipersensíveis à DNase I) e de dados obtidos através de ChIP-seq (para os fatores de transcrição), processamento dos sinais genômicos obtidos com DNase-seq e ChIP-seq e aplicação dos HMMs para realizar *footprinting* automático. Será descrita também a forma como os resultados foram validados utilizando *gold standards* bem estabelecidos na literatura.

Deve-se destacar que a principal finalidade dos experimentos descritos a seguir é o melhoramento da identificação de sítios de ligação de fatores de transcrição. A partir da discussão realizada anteriormente sobre os fatores epigenéticos, propomos que a adição de sinais genômicos relativos às modificações nas caudas das histonas acrescente informações ao modelo capazes de suprir algumas deficiências a partir do uso de dados relativos à digestão da DNase apenas. Deste estudo, duas contribuições maiores são apontadas: a construção de um modelo capaz de melhorar o desempenho e a criação de um novo método para treinar o modelo sem precisar se basear em dados validados através de técnicas experimentais custosas.

Após a obtenção dos dados nos repositórios específicos (Seção 4.1), o processo experimental começa com a aplicação da técnica *motif matching* para gerar os resultados necessários para formação do *gold standard* (Seção 4.2). Após, é realizada a identificação das regiões hipersensíveis à DNase I (HS, do Inglês *DNase I Hypersensitivity Site*) e regiões de picos nos dados de ChIP-seq para os TFs (Seção 4.3). As regiões enriquecidas nos dados de ChIP-seq também são necessárias para a criação do *gold standard*. Depois, os sinais epigenéticos (cromatina descondensada e modificações de histonas) são processados, gerando a entrada para os HMMs (Seção 4.4). De posse de tais sinais processados, os HMMs são construídos (Seção 4.4),

## 4. METODOLOGIA

treinados e aplicados nas regiões de HS (Seção 4.6). Os resultados da aplicação de tal modelo, isto é, as previsões dos sítios de ligação de fatores de transcrição são avaliados a partir de um *gold standard* bastante utilizado na literatura (Seção 4.7). A Figura 4.1 mostra, de forma esquemática, todo o processo experimental. A seguir, todos os procedimentos exibidos nesta figura serão descritos.



**Figura 4.1: Fases do processo experimental** - Esquema que demonstra todas as fases do processo experimental. Neste diagrama, o experimento foi dividido em *Aplicação do Modelo* (linhas vermelhas) e *Validação* (linhas verdes). Retângulos exibem dados obtidos (amarelos) ou gerados (azuis) e as setas conectando os retângulos representam as fases do experimento.

### 4.1 Bases de Dados

O ENCODE (do Inglês, *Encyclopedia of DNA Elements*) [Rosenbloom *et al.*, 2011; The ENCODE Project Consortium, 2004, 2007, 2011] é um projeto que pretende estudar o genoma funcional nos humanos. Este projeto está atualmente hospedado no *Genome Browser*[Kent *et al.*, 2002]. Esse consórcio, com pouco mais do que 5 anos, consiste em um esforço por parte de vários laboratórios para criar anotações funcionais de forma pangenômica. Tais anotações incluem interações na cromatina, metilação no DNA, modificações de histonas, cromatina descondensada (digestão de DNase I e FAIRE), perfis de RNA, sítios de ligação de fatores de transcrição e outros. Atualmente, tais dados estão disponíveis para cerca de 200 linhagens celulares humanas diferentes. Diversos dados, como será descrito em seguida, foram obtidos

## 4.1. BASES DE DADOS

---

através do projeto ENCODE. Todos os dados utilizados neste projeto se referem à linha celular de leucemia mielóide aguda, K562.

A Tabela 4.1 sumariza todas as faixas de dados do *Genome Browser* utilizadas e exibe os endereços virtuais para o acesso das mesmas. Os endereços virtuais exibidos contêm informações detalhadas sobre os protocolos sob o quais os dados foram gerados, incluindo a forma como foram realizadas a digestão com DNase I, a imunoprecipitação, o sequenciamento e o alinhamento. Além disso, esta tabela também contém os repositórios onde as PWMs foram obtidas. Informações sobre os sinais epigenéticos, fatores de transcrição e PWMs utilizadas são exibidas na Tabela 4.2. Acredita-se que os fatores analisados neste estudo sejam bem representativos, sendo alguns deles bastante utilizado em estudos do gênero [Boyle *et al.*, 2011; Cuellar-Partida *et al.*, 2012; Pique-Regi *et al.*, 2011]. As modificações de histonas nas quais o experimento foi focado possuem forte presença em regiões de cromatina descondensada. Por este motivo elas foram escolhidas e serão chamadas de *histonas ativadoras*.

**Tabela 4.1: Fontes dos dados.**

Fonte	Tipo	URL
Human Genome hg19	genoma completo	<a href="http://bit.ly/oHXPgq">http://bit.ly/oHXPgq</a>
Duke DNase	cromatina descondensada (DNase-seq)	<a href="http://bit.ly/wOwc8R">http://bit.ly/wOwc8R</a>
Broad Histone	modificação de histona (ChIP-seq)	<a href="http://bit.ly/xKQLS7">http://bit.ly/xKQLS7</a>
SYDH TFBS	TFBS (ChIP-seq)	<a href="http://bit.ly/A0VxYz">http://bit.ly/A0VxYz</a>
HAIB TFBS	TFBS (ChIP-seq)	<a href="http://bit.ly/zqnhn8">http://bit.ly/zqnhn8</a>
UTA TFBS	TFBS (ChIP-seq)	<a href="http://bit.ly/z9b0o1">http://bit.ly/z9b0o1</a>
Jaspar	PWM	<a href="http://bit.ly/92ebHi">http://bit.ly/92ebHi</a>
Transfac	PWM	<a href="http://bit.ly/PfTeA1">http://bit.ly/PfTeA1</a>
Uniprobe	PWM	<a href="http://bit.ly/Qn0kT3">http://bit.ly/Qn0kT3</a>
Renlab	PWM	<a href="http://bit.ly/RV5c4R">http://bit.ly/RV5c4R</a>

Os sinais epigenéticos de cromatina descondensada relativos à digestão com DNase I através de DNase-seq foram obtidos no ENCODE na faixa *Duke DNase*. Nesta faixa estão disponíveis os fragmentos brutos recuperados pelo método de DNase-seq, os fragmentos alinhados, o sinal genômico relativo à aplicação do método F-seq [Boyle *et al.*, 2008b], o sinal genômico relativo à simples contagem da sobreposição dos fragmentos obtidos e as regiões enriquecidas. Para este projeto, os fragmentos alinhados foram utilizados para gerar os sinais que posteriormente servirão como entrada para o modelo preditivo e o sinal genômico relativo à aplicação do método F-seq foi utilizado para identificar as regiões enriquecidas, isto é, as regiões hipersensíveis à

## 4. METODOLOGIA

---

DNase I. Não foram utilizados, diretamente, o sinal genômico relativo à contagem da sobreposição dos fragmentos e as regiões enriquecidas calculadas pelo próprio ENCODE pelo fato de a abordagem utilizada nesta faixa ter algumas divergências em relação estudo com o qual se pretende comparar o método proposto.

**Tabela 4.2: Sinais epigenéticos e fatores estudados** – Cada fator estudado possui uma trinca no formato (J,T,R) associado (abaixo do mesmo). Os três números de cada trinca representam, respectivamente, o número de PWMs obtidas nos repositórios Jaspar, Transfac e Renlab.

Sinais Epigenéticos	DNase	H2A.Z	H3K4me2	H3K4me3	H3K9ac
Fatores	ATF3 (0,1,0)	CEPB (0,2,0)	CTCF (1,0,1)	E2F4 (0,2,0)	GABP (1,1,0)
(J,T,R)	MEF2A (1,0,0)	P300 (0,1,0)	REST (1,1,0)		

Os sinais epigenéticos relativos às modificações de histonas gerados com ChIP-seq foram obtidos no ENCODE na faixa *Broad Histone*, proposta pelo *Broad Institute* e pelo laboratório *Bernstein lab*. Nesta faixa estão disponíveis os fragmentos brutos recuperados pelo método de ChIP-seq, os fragmentos alinhados e o sinal genômico gerado com o programa *Scripture* [Guttman *et al.*, 2010]. Novamente, apenas os dados relativos aos fragmentos alinhados foram utilizados. Tais dados também servirão de entrada para o modelo preditivo e também para calcular as regiões onde o modelo será aplicado (regiões hipersensíveis à DNase I).

Os dados relativos aos TFBSs dos fatores utilizados foram obtidos, no ENCODE, a partir das faixas *SYDH TFBS*, *HAIB TFBS* e *UTA TFBS*. A primeira faixa representa o consórcio formado pelas universidades de Stanford, Yale, sul da Califórnia e Harvard; a segunda é provida pelo *Myers Lab* do instituto HudsonAlpha de biotecnologia; e a terceira é provida pela universidade do Texas em Austin. Foram obtidos os sinais genômicos relacionados à sobreposição de fragmentos alinhados, criado de maneira diferente em cada faixa. A faixa *SYDH TFBS* utilizou métodos próprios para criação do sinal genômico (descritos no endereço eletrônico providenciado). A faixa *HAIB TFBS* utilizou o método MACS [Zhang *et al.*, 2008] para criar tais sinais. Finalmente, a faixa *UTA TFBS* gerou sinais genômicos através do programa F-seq [Boyle *et al.*, 2008b].

Além dos sinais e regiões enriquecidas obtidos no ENCODE, foram obtidas PWMs para realizar o procedimento de *motif matching*, em repositórios específicos. Foram obtidos dados nos repositórios Jaspar [Bryne *et al.*, 2008], Transfac [Matys *et al.*, 2006; Wingender *et al.*, 1996],

Uniprobe [Newburger & Bulyk, 2009] e um *motif* de alta qualidade para o insulador CTCF foi obtido no laboratório Renlab [Essien *et al.*, 2009]. O critério mínimo para que um *motif* fosse considerado é que ele tivesse sido criado a partir de um vertebrado. Como pode ser visto na Tabela 4.2 podem existir mais de uma PWM para cada fator, até para cada repositório. Como cada uma dessas PWMs redundantes foram geradas com um processo específico que possui sua própria qualidade, foi optado por utilizar todos os *motifs* encontrados para todos os fatores. O processo de *motif matching* é utilizado, assim como os dados de TFBS com ChIP-seq, apenas para criar o *gold standard*.

## 4.2 Motif Matching

Todas as PWMs obtidas foram utilizadas para realizar *motif matching* no genoma completo. Essa técnica produz *bit scores* que podem ser utilizados para avaliar a qualidade de cada emparelhamento. Para permitir uma comparação direta com o modelo prévio, foi seguida a sua metodologia para aceitação de TFBSS obtidos através desta técnica. Essa metodologia consiste em descartar todos os emparelhamentos que obtiveram *bit scores* menores do que o mínimo entre: 70% do maior *bit score* possível (sequência consenso do PWM) e 90% da diferença entre o maior e o menor possível *bit score* [Boyle *et al.*, 2011]. Neste trabalho será utilizada a nomenclatura MPBS (do Inglês, *Motif Predicted Binding Site*), para denotar os TFBSS preditos através deste método.

Para realizar o procedimento de MM, foi utilizado o módulo *Bio.Motif* para análise de *motifs* da ferramenta *Biopython* [Cock *et al.*, 2009]. Essa ferramenta utiliza um modelo baseado em probabilidade de fundo (*background*) assim como visto na Seção 2.3.3. Esse parâmetro, que assume um valor real  $v$  em escala logarítmica, permite selecionar os emparelhamentos que ocorreram com probabilidade  $2^v$  vezes maior do que o esperado por chance, dadas as frequências dos nucleotídeos naturais do genoma. Ao utilizar o valor  $v = 0$ , foram selecionados todos os resultados que ocorriam com maior probabilidade do que o esperado por acaso (pois  $2^0 = 1$ , portanto todos os valores mais prováveis que o fundo são selecionados). Apenas após essa filtragem inicial o método descrito em [Boyle *et al.*, 2011] foi aplicado (em concordância com o proposto por estes).

## 4.3 Análises de Enriquecimento

Para ter acesso às regiões hipersensíveis à DNase I e regiões de picos nos dados de ChIP-seq foi realizada uma análise estatística de enriquecimento simples. Tal análise foi realizada nos dados relativos aos sinais genômicos de DNase-seq ou ChIP-seq (para os TFs), obtidos a

## 4. METODOLOGIA

---

partir da aplicação dos métodos específicos para contagem de sinais descrita em cada repositório avaliado. A análise de enriquecimento consistiu no ajuste destes sinais contínuos à distribuições  $\Gamma$ , definida nas Equações 4.1 (distribuição  $\Gamma$  com parâmetros  $k$  e  $\theta$ ) e 4.2 (função  $\Gamma$  utilizada na definição da distribuição).

$$f(x; k, \theta) = \frac{1}{\theta^k} \frac{1}{\Gamma(k)} x^{k-1} e^{-\frac{x}{\theta}} \quad (4.1)$$

$$\Gamma(n) = (n - 1)! \quad (4.2)$$

O procedimento de ajuste é simples. É calculada a média  $\mu$  e a variância  $\sigma^2$  da amostra, isto é, dos sinais epigenéticos, em todo o genoma. A média e a variância são então utilizadas para estimar os parâmetros  $k$  e  $\theta$  através da resolução de um sistema de equações com os seguintes resultados probabilísticos:  $\mu = k\theta$  e  $\sigma^2 = k\theta^2$ . Esses parâmetros são então utilizados para inferir o *p-value* de corte, para o qual os valores inferiores serão as regiões não-enriquecidas e os valores superiores serão as regiões enriquecidas. Esta função é comumente utilizada para este propósito por possuir características semelhantes às distribuições reais dos sinais obtidos através de métodos como DNase-seq e ChIP-seq. A distribuição exponencial também é bastante utilizada, porém como a distribuição  $\Gamma$  foi utilizada no estudo com o qual se pretende realizar as comparações, a última foi escolhida para reproduzir mais fielmente os resultados.

Baseando-se nos ajustes realizados, foram consideradas como regiões enriquecidas aquelas que possuíram valores maiores ou iguais ao valor correspondente ao *p-value* de 0.05. É importante constatar que o próprio ENCODE disponibiliza tal análise estatística, porém tais dados não foram utilizados pois a metodologia, de uma forma geral, era diferente. O *p-value* utilizado foi escolhido em conformidade com objetivo de comparar este modelo com o previamente proposto em [Boyle *et al.*, 2011].

### 4.4 Processamento dos Sinais Epigenéticos

A primeira fase do processamento dos sinais dos dados de DNase-seq consiste na contagem das sobreposições dos fragmentos alinhados. Neste caso, foi considerado apenas o bp na extremidade 5' dos fragmentos, correspondendo à posição exata no qual a enzima DNase I digeriu o DNA. Tal abordagem gera um sinal de alta resolução bastante específico, capaz de delinear claramente as proteínas ligadas ao DNA.

#### **4.4. PROCESSAMENTO DOS SINAIS EPIGENÉTICOS**

---

Para a geração dos sinais de contagem bruta para os dados de modificação de histonas obtidos através de ChIP-seq, o mesmo procedimento foi aplicado. Entretanto, como o nucleossomo alvo pode se encontrar em qualquer posição do fragmento recuperado através de ChIP, os fragmentos foram estendidos até o tamanho de 200 bp, que representa a média de tais fragmentos reais (os fragmentos são sequenciados apenas nas primeiras 36 bases). A diferença na resolução entre os dois sinais gera padrões específicos, analisados em mais detalhes na Seção 4.5.

Os dados de cromatina descondensada (DNase-seq) foram normalizados de forma a minimizar a variação entre o tamanho dos picos ao longo do genoma. Tal normalização seguiu o método local proposto em [Boyle *et al.*, 2011]. Neste método, cada sinal (em cada coordenada genômica) é dividido pela média de todas as entradas maiores que 0 em uma janela de tamanho igual a 1 kb ao redor desta coordenada genômica. A principal característica desta normalização é a preservação das nuances dadas pela alta resolução do método DNase-seq. Os dados de ChIP-seq foram submetidos a uma simples função logarítmica, com objetivo de suavizar as curvas ao longo do genoma. O método utilizado para os dados de DNase-seq não foi utilizado para os sinais de modificações de histonas, pelo fato de que a intensidade deste sinal é importante para o modelo, enquanto a intensidade do sinal de DNase-seq não tem grande importância (como será visto na Seção 4.5).

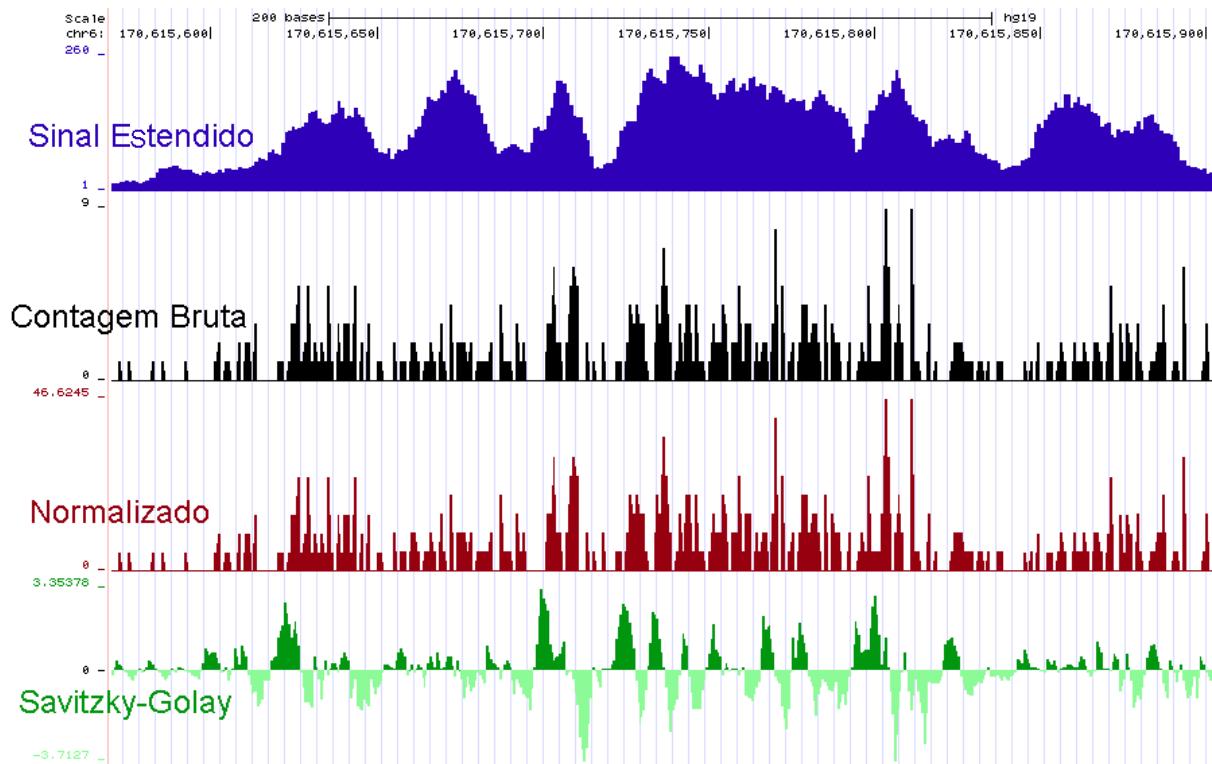
Os dados de DNase-seq passam por mais uma etapa, com objetivo de extrair as características necessárias para o modelo que será descrito em detalhes na Seção 4.5. Essa etapa consiste em duas fases. Na primeira, os dados são suavizados através do filtro estatístico de Savitzky-Golay [Gorry, 1990; Leach *et al.*, 1984; Luo *et al.*, 2005; Madden, 1978; Press *et al.*, 1992]. Tal suavização remove ruídos naturais deste sinal epigenético. A suavização é baseada no ajuste dos sinais normalizados a um polinômio de grau 2 através de uma convolução com uma janela de tamanho 8 bp (excluindo o bp central). A segunda etapa consiste na diferenciação deste sinal epigenético, através da computação da 1<sup>a</sup> derivada [Boyle *et al.*, 2011].

Os sinais gerados após a suavização e derivação representam a inclinação (em Inglês, *slope*) do sinal normalizado. Isto quer dizer que, nos locais onde o sinal normalizado tinha um movimento crescente, o sinal relativo à inclinação assumia valores positivos; e nos locais onde o sinal normalizado tinha um movimento decrescente, o sinal relativo à inclinação assumia valores negativos. Além disso, quanto mais íngreme a elevação ou queda do sinal normalizado, maiores são os valores da inclinação correspondente (em termos absolutos).

A Figura 4.2 exibe um exemplo do sinal obtido através de DNase-seq, em todas as fases do processamento, para um trecho do cromossomo 6. Esta figura foi gerada utilizando o *Genome Browser* e contém um formato adicional para os dados processados, não utilizado no processo experimental: o sinal estendido. Este sinal é gerado a partir da extensão dos fragmentos alinhados em 5 bp para a esquerda e para a direita da coordenada onde a enzima DNase I digeriu o DNA. O objetivo de tal sinal é facilitar a visualização e a interpretação dos outros sinais.

## 4. METODOLOGIA

---

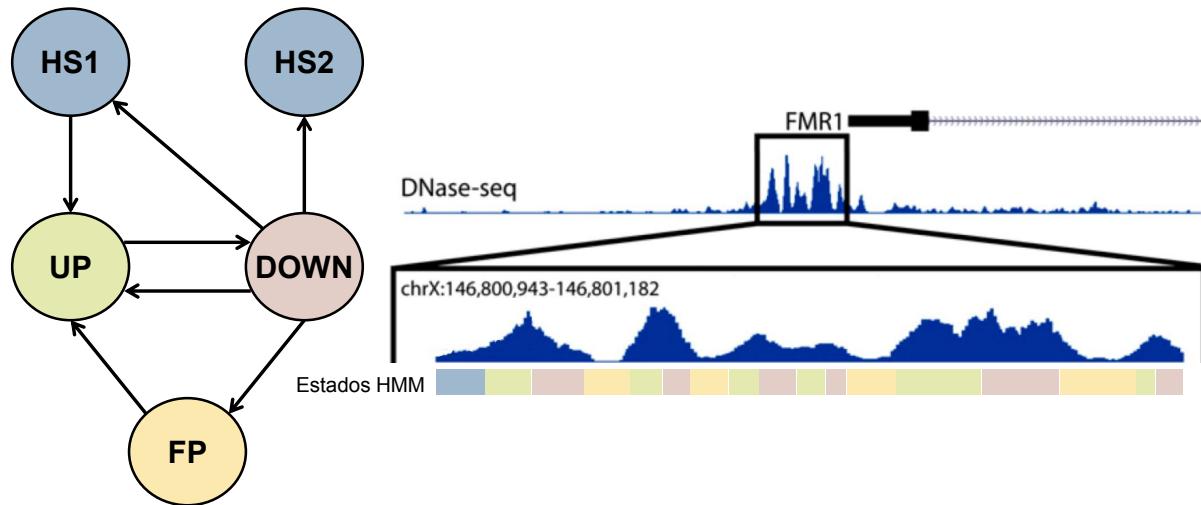


**Figura 4.2: Modificação dos sinais ao longo do processamento** - Esquema que exibe os sinais de DNase-seq em todas as fases do processamento para um trecho do cromossomo 6. Em azul está o sinal estendido (ver descrição no texto), em preto o sinal correspondente à contagem bruta dos dados, em vermelho o sinal normalizado e em verde o sinal obtido após a aplicação da suavização e diferenciação através do método de Savitzky-Golay.

## 4.5 Footprinting com HMMs

Foi constatado em [Boyle *et al.*, 2011] que um TFBS poderia ser caracterizado através dos sinais de cromatina descondensada gerados com DNase-seq como uma depleção de sinal entre dois picos. Isto se explica pelo fato de que naquela região onde a proteína estava ligada não havia digestão da DNase, porém nas regiões imediatamente anterior e posterior a clivagem ocorre. Tal padrão será intitulado pico-vale-pico. O padrão que se deseja reconhecer é formado, nos sinais normalizados, por uma subida e descida (primeiro pico), então uma região relativamente plana (vale) e outra subida e descida (segundo pico). Tal padrão é facilmente representado através dos sinais de inclinação, dado que subidas são representadas por valores positivos e descidas são representadas por valores negativos. Boyle et al. utilizaram essa ideia para construir seu HMM capaz de predizer TFBSSs (Figura 4.3).

Ao serem adicionados os sinais de histonas, um padrão levemente diferente ocorre. Dado que os sinais gerados através de ChIP-seq possuem resolução um pouco menor, uma região inteira de HS (que correspondem à blocos com vários picos agrupados no sinal obtido com DNase-seq)



**Figura 4.3: HMM que utiliza dados de DNase-seq apenas** - Esquema gráfico do HMM proposto por Boyle et al. (esquerda) para predizer TFBSS com base apenas em sinais obtidos através de DNase-seq. Exemplo dos estados obtidos, em cada coordenada genômica, a partir da aplicação deste modelo em um trecho da região promotora do gene FMR1 no cromossomo X (direita). Fonte: [Boyle et al., 2011]

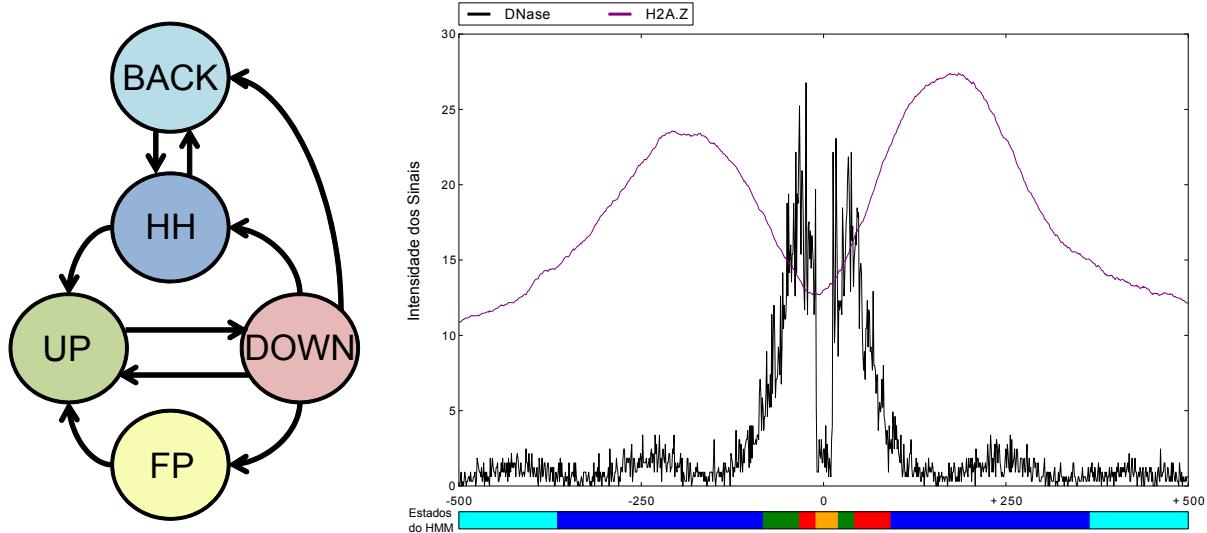
corresponde a uma região de depleção nas histonas ativadoras. Foi observado (ver Seção 5.1) que o sinal das modificações de histona ativadores constituem outro padrão de pico-vale-pico, porém em um nível mais alto do que o padrão gerado por cromatina descondensada. Dois picos de histonas ativadoras (sinais intensos) delimitam regiões de HS, que por sua vez possuem vários padrões pico-vale-pico correspondentes aos sítios de ligação.

Realizada esta discussão sobre as características dos sinais epigenéticos, pode-se definir a estrutura do novo HMM proposto. Tal modelo deve ser capaz de reconhecer tal padrão formado por, simultaneamente, um sinal de cromatina descondensada e um sinal de histona (isto é, um modelo bivariado). O modelo possui um estado para sinais de fundo (*background* – *BACK*), que correspondem aos sinais baixos para ambas cromatina descondensada e modificação de histona, geralmente no início ou fim das regiões onde os modelos foram aplicados. Ao encontrar valores significativamente altos de modificação de histonas (primeiro pico), o modelo procede para o estado *High Histone* (*HH*). Então esse valor irá reduzir um pouco entrando na região de HS. Nesta região o modelo irá variar entre os estados de crescimento de sinal de cromatina descondensada (*UP*), decrescimento deste sinal (*DOWN*) e regiões de vale (*Footprint* – *FP*). Esta última região corresponde aos sítios de ligação de fatores de transcrição. Após a região de HS, o método poderá: (1) retornar para o estado *HH*, caso o segundo pico exista (ou, mais comumente, existirem várias regiões de HS na região sendo analisada, delimitadas por vários picos de modificações de histonas); (2) retornar para o estado *BACK*, quando os sinais de histona forem demasiadamente baixos ou não existirem mais regiões de HS. A Figura 4.4 demonstra o HMM criado (lado esquerdo) e o explica através de um gráfico com os sinais de digestão de

#### 4. METODOLOGIA

---

DNase e a histona variante H2A.Z (lado direito).



**Figura 4.4: Modelagem do HMM e exemplo de aplicação** - HMM utilizado neste estudo (esquerda) contendo 5 estados. O estado *BACK* (azul claro) representa as regiões de pequena intensidade de sinais. O estado *HH* (azul escuro) representa as regiões de alta intensidade de histonas modificadas. O estado *UP* (verde) e *DOWN* (vermelho) representam, respectivamente, as regiões onde os sinais de digestão de DNase I crescem e decrescem. E o estado *FP* (amarelo) representa as regiões de vale que correspondem aos TFBSSs (ou *footprints*). O gráfico (direita) corresponde à média dos sinais (para a digestão de DNase I e a histona variante H2A.Z) obtidos em 100 regiões de tamanho 1000 centralizadas nos 100 MPBSs com maior *bit score*. O mapa de cores abaixo do gráfico mostra os estados do HMM correspondentes à cada posição, baseado nas cores dos estados da figura do HMM.

## 4.6 Estimação de Parâmetros e Aplicação dos HMMs

Este HMM foi treinado, isto é, seus parâmetros foram estimados, utilizando duas abordagens. A primeira, intitulada *FMR1* consiste na forma proposta em [Boyle *et al.*, 2011]. Esta estratégia é baseada, inicialmente, em regiões biologicamente validadas através de métodos de baixo rendimento como o *DNase I Footprinting*. A segunda estratégia, chamada *STAMP* foi elaborada com o intuito de anotar mais regiões inicialmente sem ter que se basear em metodologias biológicas iniciais.

Na estratégia *FMR1*, primeiramente deve ser obtida uma região onde TFBSSs foram experimentalmente validados através de algum método de alta acurácia. No caso, foi utilizado o resultado de um experimento de *DNase I Footprinting* tradicional da região promotora do gene

## **4.6. ESTIMAÇÃO DE PARÂMETROS E APLICAÇÃO DOS HMMS**

---

FMR1 (*Fragile X Mental Retardation 1*) no cromossomo X [Drouin *et al.*, 1997]. Essa região é anotada manualmente, isto é, para cada coordenada genômica, é atribuído um estado do HMM com base nos TFBSSs comprovados. Posteriormente, a anotação é utilizada para estimar os parâmetros de um primeiro modelo, através da técnica de máxima verossimilhança (Seção 3.3). Este primeiro modelo é então utilizado para anotar automaticamente uma região maior. No caso, as 1000 regiões de HS do cromossomo 6 que possuem maior evidência de enriquecimento foram utilizadas [Boyle *et al.*, 2011]. Com base nesta anotação mais abrangente, o modelo final é criado novamente através de máxima verossimilhança.

A segunda abordagem de treinamento, intitulada STAMP, consiste em um novo método proposto. A motivação da criação deste novo método é que, dessa forma, mais regiões poderão ser inicialmente anotadas sem a necessidade de realizar métodos biológicos a priori ou procurar na literatura por regiões que coincidam com trechos onde o treinamento é interessante. Este método utiliza a ferramenta STAMP [Mahony & Benos, 2007], que consiste em uma técnica para se realizar *motif matching* em cadeias de nucleotídeos com base em um repositório contendo diversas PWMs (e não apenas uma PWM).

Em detalhes, o método STAMP é aplicado nas regiões iniciais que serão utilizadas para se realizar a anotação (geralmente, locais onde existem sinais para todas as características epigenéticas em questão, e possuem bom nível de enriquecimento de DNase I). O algoritmo realiza um *motif matching* mais elaborado na região em questão, utilizando cada uma das PWMs em cada repositório utilizado. Neste caso, foram utilizados os repositórios completos do Jaspar, Transfac (público), Uniprobe e o *motif* CTCF do Renlab. Os resultados são listas contendo probabilidades de afinidade de ligação de cada fator nesta região. Conforme proposto [Boyle *et al.*, 2011; Mahony & Benos, 2007], foram consideradas como significativas os emparelhamentos que obtiveram afinidade de ligação menor ou igual à  $10^{-6}$ . Tais resultados correspondem a um conjunto de TFBSSs de alta qualidade, suficiente para realizar as anotações iniciais.

A metodologia completa consiste em: a partir dos resultados desta técnica aplicada nas 10 melhores regiões de HS do cromossomo 6 (utilizado apenas em conformidade com a metodologia prévia) que possuem maior evidência de enriquecimento, tais regiões são manualmente anotadas de forma idêntica à realizada no treinamento FMR1. Tais anotações são então utilizadas para gerar o modelo final através de máxima verossimilhança. Uma segunda rodada de anotação e treinamento, como no treinamento FMR1, não foi realizada por verificar que os parâmetros já eram bastante robustos.

Os modelos treinados são aplicados nas regiões de HS identificadas da forma descrita na Seção 4.3. Tais regiões não são regiões de HS no sentido biológico literal, mas regiões onde observou-se um enriquecimento na atividade de digestão de DNase I. Se torna claro, então, a razão de ter escolhido um *p-value* de enriquecimento relativamente alto (0.05). Dessa forma as

## 4. METODOLOGIA

---

regiões são um pouco mais largas do que regiões de HS literais, permitindo que os padrões (principalmente os das histonas, que são mais largos) sejam completamente incluídos nas mesmas.

A implementação dos HMMs foi realizada utilizando o pacote em *Python* da *General Hidden Markov Model Library (GHMM)* [Schliep *et al.*, 2004]. A probabilidade posterior foi utilizada em todos os casos para aferir a sequência de estados escondidos. Os TFBSSs preditos correspondem às coordenadas genômicas onde a probabilidade posterior do estado *FP* foi maior do que a dos demais estados (ver Seção 3.2.2). Tal região foi estendida em 3 bp para a esquerda e para a direita para tornar as previsões mais robustas e facilitar a visualização pelos métodos de validação.

### 4.7 Gold Standard

O *gold standard* utilizado neste projeto foi baseado em uma abordagem bastante utilizada na literatura [Boyle *et al.*, 2011; Cuellar-Partida *et al.*, 2012; Pique-Regi *et al.*, 2011]. Ele consiste em um conjunto contendo TFBSSs considerados verdadeiros e falsos, criado a partir das MPBSs em conjunto com os dados de ChIP-seq para os fatores de transcrição. TFBSSs verdadeiros são todos os MPBSs que se possuem evidência de ChIP-seq e TFBSSs falsos são aqueles que não possuem tal evidência. A evidência se apresenta quando pelo menos 1 bp do MPBS apresenta sobreposição com as regiões enriquecidas nos dados de ChIP-seq. Essas regiões enriquecidas foram avaliadas como descrito na Seção 4.3.

Após a identificação dos TFBSSs verdadeiros e falsos para cada fator, uma tabela de contingência pode ser criada através da consideração das previsões (ou *footprints*) realizadas. Os verdadeiros positivos (*TP*) são os verdadeiros TFBSSs que possuem sobreposição com algum *footprint*; os falsos negativos (*FN*) são os verdadeiros TFBSSs que não possuem *footprint* associado; os verdadeiros negativos (*TN*) são falsos TFBSSs que não possuem sobreposição com algum *footprint*; e falsos positivos (*FP*) são falsos TFBSSs que possuem *footprint* associado. Novamente, a mínima sobreposição de 1 bp já é válida. A partir desta tabela de contingência, é possível calcular as estatísticas utilizadas para avaliar o modelo, apresentadas na Tabela.

O modelo proposto foi comparado apenas com a abordagem prévia em [Boyle *et al.*, 2011]. O modelo não foi comparado à abordagem CENTIPEDE descrita em [Pique-Regi *et al.*, 2011] e a abordagem Bayesiana detalhada em [Cuellar-Partida *et al.*, 2012] pelo fato de que o conjunto de validação utilizado por eles diferia da proposta do Boyle et al. Primeiramente, em Pique-Regi et al. e Cuellar et al. os TFBSSs verdadeiros são aqueles MPBSs que contêm evidência de ChIP-seq (assim como em Boyle et al.), porém os TFBSSs falsos consistem nos MPBSs que se sobrepõem em regiões com uma quantidade de fragmentos de ChIP-seq sobrepostos menor ou igual ao experimento controle para esta linhagem celular (também disponível no ENCODE). Apontamos

## **4.8. CONSIDERAÇÕES FINAIS**

---

então o fato de que essa abordagem faz com que apenas um subconjunto das instâncias negativas estejam sendo consideradas. Além disso, as instâncias negativas consideradas são apenas aquelas que possuem níveis muito baixos de evidência de ChIP-seq, isto é, são as instâncias negativas mais fáceis de serem classificadas corretamente. Além disso, ao comparar os resultados de sensibilidade vs. taxa de falsos positivos (curva ROC), esses MPBSs que não foram considerados verdadeiros TFBSS nem falsos TFBSS, foram descartados sobre a premissa de que o *gold standard* estaria contaminado com instâncias na fronteira de classificação. A mesma observação anterior, a respeito de isto representar um problema mais fácil do ponto de vista de aprendizado de máquina, se aplica a este argumento.

### **4.8 Considerações Finais**

Neste capítulo foram descritos em detalhes os procedimentos realizados neste trabalho. Em resumo, foram descritos os procedimentos de obtenção dos dados, o *motif matching*, a análise de enriquecimento dos sinais obtidos através de DNase-seq e ChIP-seq para os fatores de transcrição, o processamento dos sinais de DNase-seq e de ChIP-seq para as modificações de histonas, a modelagem, treinamento e aplicação dos HMMs multivariados, e a forma de validação utilizada para comparar o novo modelo proposto ao modelo prévio. Finalmente, tal modelo prévio também foi replicado, para que os resultados entre os dois fosse comparado dadas as ferramentas utilizadas neste projeto.

# 5

## Resultados e Discussão

Neste capítulo serão mostrados os resultados referentes à aplicação do método proposto. Tais resultados serão exibidos no formato de: (1) gráficos de sinais epigenéticos, que mostram padrões médios ao redor de regiões de interesse; (2) tabelas, com as estatísticas representando a acurácia de modelos, isto é, a eficiência preditiva dos sinais. Além disso serão exibidos dados a respeito do tempo computacional e capacidade de armazenamento necessário para a realização dos experimentos.

Tais resultados apresentados também serão discutidos, sob a ótica das considerações feitas durante a apresentação da fundamentação teórica biológica e computacional. Serão realizadas observações a respeito de: (1) análises realizadas envolvendo os sinais epigenéticos presentes nas regiões de interesse; (2) análises envolvendo os estados do HMM nas regiões de interesse; (3) acurácia dos modelos. Serão discutidos também alguns exemplos da aplicação dos modelos propostos, mostrando as ocasiões em que o modelo funcionou conforme previsto e as melhorias que ainda precisam ser realizadas.

A apresentação dos resultados foi dividida em duas partes. Na primeira, é realizada uma análise mais profunda das características que os sinais epigenéticos possuem em certas regiões de interesse (Seção 5.1). Essa foi a primeira análise realizada neste trabalho e teve como objetivo o entendimento do comportamento dos sinais epigenéticos que seriam utilizados posteriormente no modelo probabilístico. Na segunda parte, os resultados da aplicação do HMM descrito na Seção 4.5 serão exibidos (Seção 5.2). São exibidos resultados tanto para o HMM proposto quanto para o HMM segundo a abordagem anterior.

Conforme mencionado anteriormente, foram investigados os padrões epigenéticos relativos à digestão de DNase I (nomeado *DNase*), à histona variante H2A.Z e às histonas ativadoras H3K4me2, H3K4me3 e H3K9ac. Esse conjunto de características epigenéticas foi utilizado pelo fato de que ele marca, de forma eficaz, regiões de cromatina descondensada (ver Seção 2.4.1).

## **5.1. ANÁLISE DOS SINAIS EPIGENÉTICOS**

---

Em relação aos fatores de transcrição (e aos seus respectivos *motifs*), a análise dos sinais epigenéticos médios foi realizada em todos os fatores presentes na Tabela 4.2. A acurácia do modelo, entretanto, foi acessada apenas para um conjunto representativo destes, a saber, ATF3 (com PWM obtida no Transfac), CTCF (com *motifs* do Jaspar e Renlab), GABP (PWMs do Jaspar e Transfac) e REST (com *motifs* do Jaspar e Transfac). Os fatores CTCF, GABP e REST foram escolhidos por terem sido também utilizados em [Boyle *et al.*, 2011]. O fator REST, em especial, foi utilizado como forma de avaliar fatores que possuem baixos níveis de marcas epigenéticas. O fator ATF3 foi escolhido pois observou-se que este fator possui a maior razão entre instâncias negativas e instâncias positivas (ver Tabela ??), tendo sido este caso o mais desafiador para a nova abordagem proposta (ver discussão realizada na Seção 5.2).

### **5.1 Análise dos Sinais Epigenéticos**

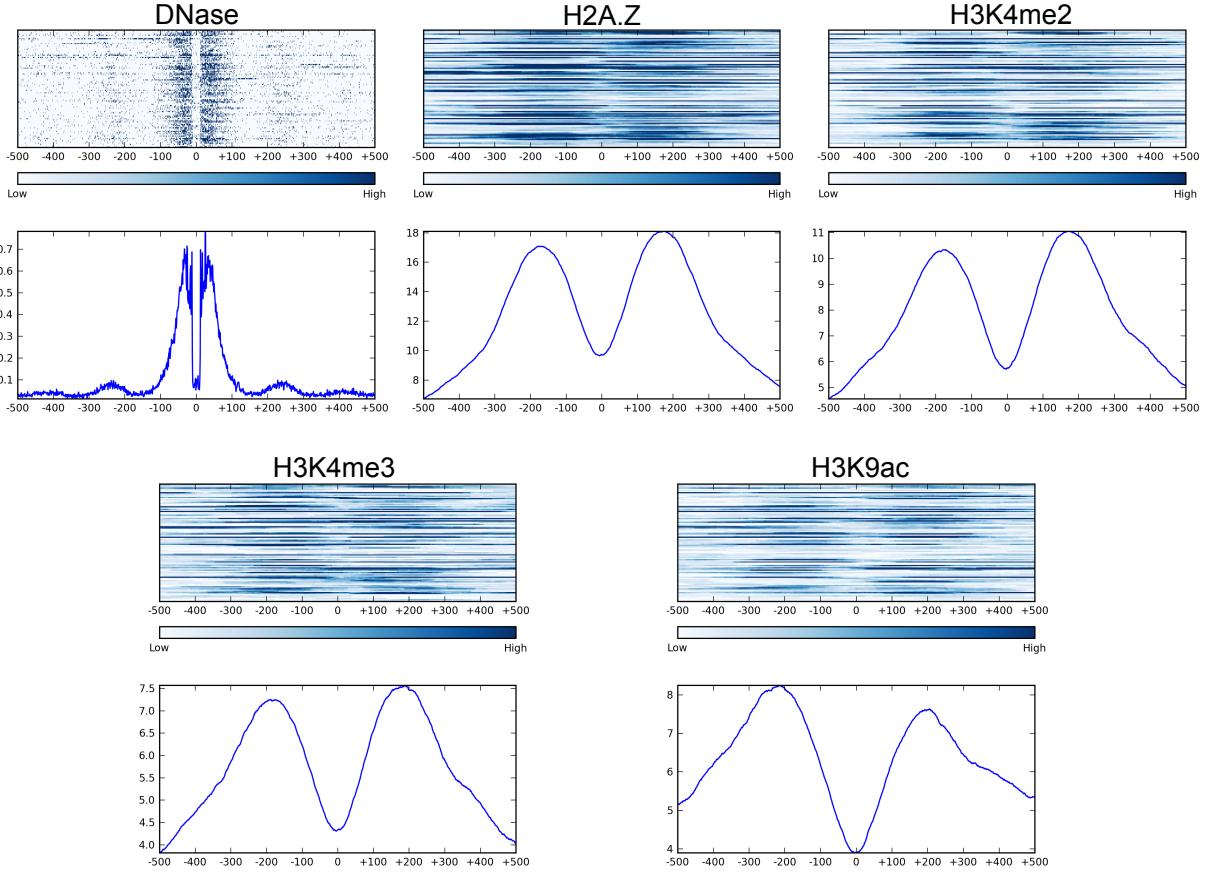
Serão realizados três tipos diferentes de análises nesta seção. A primeira análise consiste na investigação do comportamento dos sinais ao redor de regiões de MPBSs. O objetivo disto é a apresentação dos sinais epigenéticos para que o leitor se familiarize com os padrões observados. A segunda análise corresponde à investigação destes sinais em regiões de MPBSs com e sem evidência de ChIP-seq, com objetivo de mostrar a capacidade de separação de cada sinal epigenético, em diferentes fatores de transcrição, com base nas definições do *gold standard*. Finalmente, a terceira análise engloba MPBSs, evidência de ChIP-seq, e predições realizadas com o modelo previamente proposto, com objetivo de entender os pontos positivos e negativos deste modelo baseado em DNase apenas.

A primeira análise consiste na visualização dos sinais epigenéticos ao redor dos 100 MPBSs com maior *bit score* (Figura 5.1). Cada região analisada consiste na extensão de 500 bp para a esquerda e direita do local onde a PWM foi identificada no genoma. Para cada sinal, é mostrado um gráfico de cores (parte superior) onde as linhas correspondem às regiões analisadas e as colunas correspondem às coordenadas genômicas. A intensidade de cada ponto neste gráfico corresponde à intensidade do respectivo sinal epigenético (*low* = baixa intensidade e *high* = alta intensidade, nas escalas de cores). Além do gráfico de cores, existe um gráfico de linha (parte inferior) correspondente à média do sinal ao longo de toda a extensão analisada, para cada região. Nesta análise, cujo objetivo é apenas apresentar as características epigenéticas usuais, são apresentados os resultados apenas o fator CTCF com *motif* obtido no repositório Jaspar.

Através da análise da Figura 5.1 é possível constatar claramente os padrões de depleção de DNase e de modificações de histonas nas regiões com alta afinidade de ligação para o *motif* CTCF. A alta resolução do sinal de digestão de DNase I faz com que a depleção seja bastante específica, em média, delineando os fatores de transcrição de forma precisa. A adição de tais

## 5. RESULTADOS E DISCUSSÃO

---



**Figura 5.1:** Análise das melhores regiões de MPBS para o CTCF - Análise dos sinais epigenéticos nas 100 regiões com maior *bit score*.

dados dão ao modelo uma capacidade maior para realizar a principal tarefa proposta: a de identificar de forma precisa os TFBSs. Com resolução um pouco mais baixa, os sinais das histonas (obtidos através de ChIP-seq) possuem, em média, depleções mais abrangentes, que geralmente englobam áreas inteiras de HS (isto é, diversos picos e depleções de DNase).

A segunda análise consiste na visualização, para cada fator de transcrição, dos sinais epigenéticos ao redor das 100 regiões de MPBSs com maiores *bit score* que possuem evidência de ChIP-seq e das 100 regiões de MPBSs com maiores *bit score* que não possuem evidência de ChIP-seq (Figuras 5.2 e 5.3). Cada região analisada consiste na extensão de 500 bp para a esquerda e para a direita do local onde o *motif* foi identificado no genoma. Na figura, são exibidos gráficos de linha contendo a média dos sinais para todas estas regiões sobre toda a extensão analisada. A linha verde corresponde aos MPBSs sem evidência de ChIP-seq e a linha vermelha corresponde aos MPBSs com evidência de ChIP-seq.

Todos os fatores analisados são mostrados neste caso, para todos os sinais epigenéticos. Os rótulos dos fatores estão no formato *NOME\_XN*, onde *NOME* corresponde ao nome do fator,

## **5.1. ANÁLISE DOS SINAIS EPIGENÉTICOS**

---

$X$  corresponde à inicial do repositório onde tal fator foi obtido e  $N$  corresponde ao número do *motif*, em ordem de entrada no repositório, deste fator (como mencionado anteriormente, alguns fatores possuem mais de um PWM por repositório). Caso existam menos de 10 sinais (do máximo de 100) para qualquer categoria descrita (com evidência ChIP e sem evidência ChIP), a curva correspondente a esta categoria não é exibida, para que problemas relativos à computação da média de poucas regiões não enviesasse a visualização. Esse caso ocorreu apenas para o fator REST com *motif* obtido no repositório Transfac.

O objetivo deste gráfico, que junta informação de MPBSs com enriquecimento de ChIP-seq, é visualizar os padrões epigenéticos com base no que foi considerado o *gold standard* deste projeto. A partir de tal visualização, é possível observar o comportamento dos sinais epigenéticos em regiões onde se deseja que o modelo reconheça como TFBS e em regiões onde se deseja que o modelo não considere um TFBS.

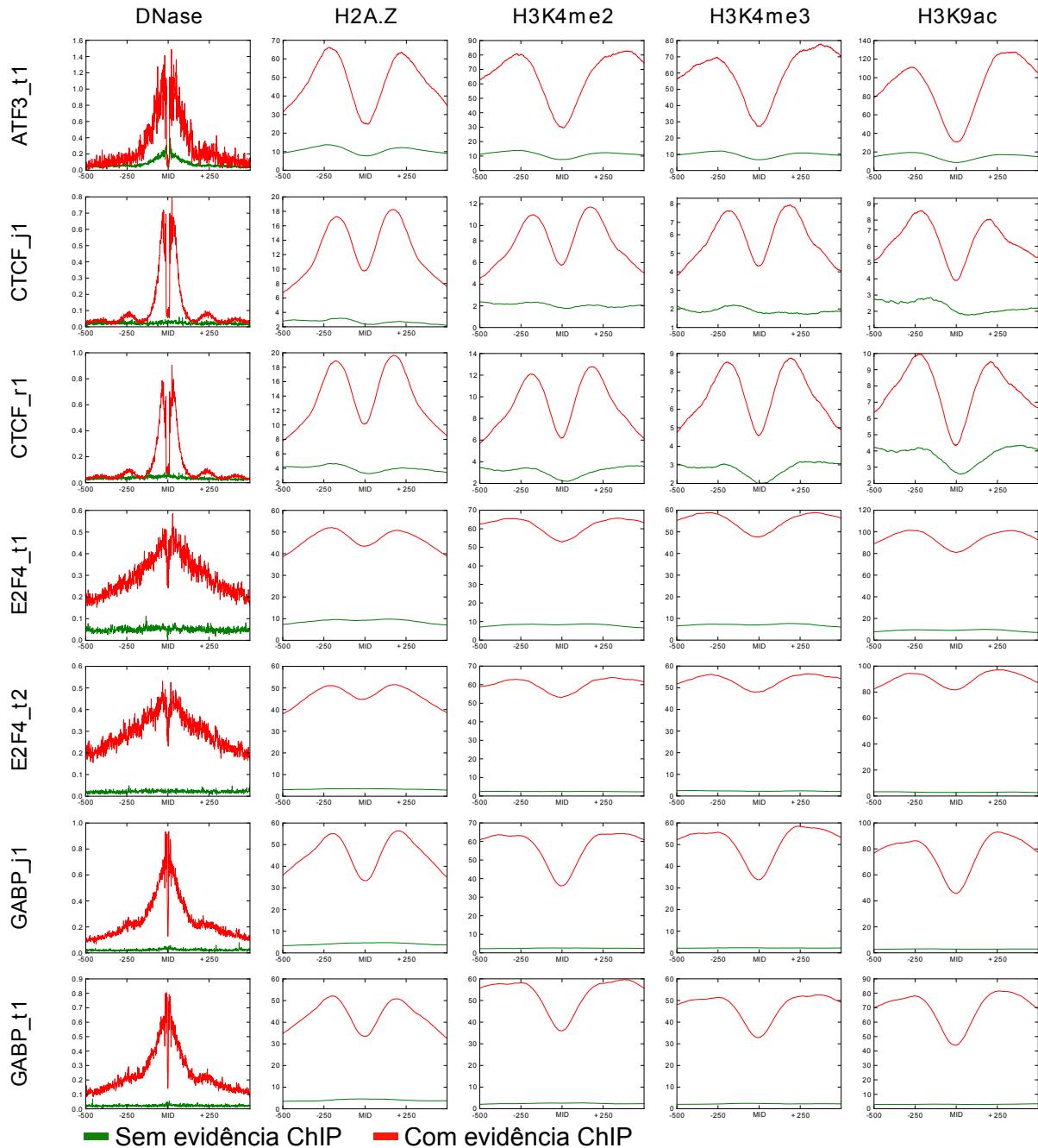
Os gráficos presentes na Figura 5.2 mostram os principais argumentos em favor da utilização de dados epigenéticos. Observa-se com clareza, neste gráfico, a diferença de intensidade e formato dos picos/vales entre regiões de MPBSs com evidência de ChIP-seq e sem tal evidência. É importante ressaltar que tal diferença ocorre mesmo tendo sido consideradas as melhores regiões sem evidência de ChIP-seq, isto é, as regiões que possuem maiores *bit score*. Os fatores ATF3, CTCF e GABP possuem padrões de depleção (pico-vale-pico) bem delineados, enquanto o fator E2F4 possui depleções mais suaves tanto para os dados de DNase-seq quanto para obtidos com ChIP-seq.

O contraste das curvas entre as regiões enriquecidas e não enriquecidas, observado nestes gráficos, variou bastante. O GABP (ativador) possui contrastes enormes, tendo as regiões não enriquecidas de ChIP-seq praticamente nenhuma depleção visível em média. O ativador E2F4 também apresentou contrastes semelhantes ao do fator GABP, porém nesse caso o sinal médio relativo às regiões com evidência de ChIP-seq possuiu depleções menos acentuadas em relação ao GABP. Os ativadores, em geral, possuíam níveis mais altos de histonas consideradas ativadoras, enquanto o nível de DNase geralmente não variou de forma tão abrupta. Os fatores ATF3 e CTCF estão em leve discordância com esse fato, apresentando depleções suaves (porém visíveis) até em regiões sem evidência de ChIP-seq.

Por outro lado, os gráficos presentes na Figura 5.3 mostram sinais epigenéticos com padrões mais fracos e destoantes dos exibidos na Figura 5.2. O fator CEBPB possuiu os sinais mais fracos entre todos os fatores analisados, porém ainda assim é possível verificar diferenças na intensidade dos sinais epigenéticos, em especial nas modificações de histonas. Apesar de possuírem altos níveis de presença dos sinais epigenéticos, os fatores MEF2A e P300 diferem dos fatores da figura anterior pelo fato de que a depleção é bem menos caracterizada, em especial para o sinal de DNase. Finalmente, para o silenciador REST, foram observados padrões claros em relação à DNase, porém pouca evidência das histonas ativadoras.

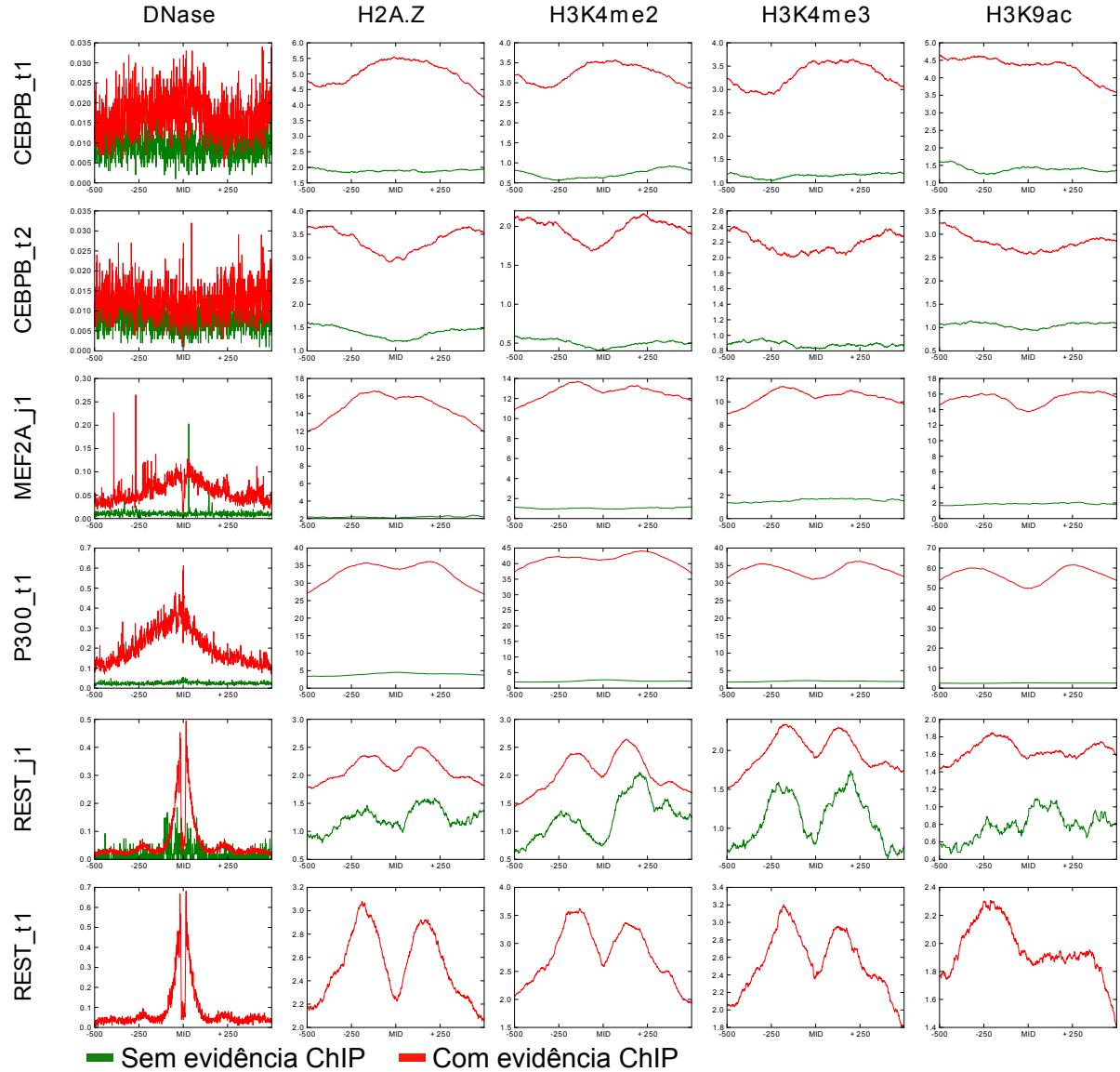
## 5. RESULTADOS E DISCUSSÃO

---



**Figura 5.2:** Regiões de TFBS com e sem evidência de ChIP-seq Pt.1 - Análise dos sinais epigenéticos ao redor dos 100 MPBSs com maior *bit score* que possuem ou não possuem evidência de ChIP-seq. São analisadas regiões de 1000 bp, sendo necessárias pelo menos 10 regiões para cada categoria, para que o sinal seja exibido (evitando vieses estatísticos). Nesta figura, são exibidos os fatores de transcrição que apresentaram os sinais epigenéticos mais delineados dentre os fatores estudados.

## 5.1. ANÁLISE DOS SINAIS EPIGENÉTICOS



**Figura 5.3:** Regiões de TFBS com e sem evidência de ChIP-seq Pt.2 - Análise dos sinais epigenéticos ao redor dos 100 MPBSs com maior *bit score* que possuem ou não possuem evidência de ChIP-seq. São analisadas regiões de 1000 bp, sendo necessárias pelo menos 10 regiões para cada categoria, para que o sinal seja exibido (evitando vieses estatísticos). Nesta figura, são exibidos os fatores de transcrição que apresentaram os sinais epigenéticos menos claros dentre os fatores estudados.

## 5. RESULTADOS E DISCUSSÃO

---

Os contrastes nesta segunda figura apresentaram variações ainda maiores. O fator CEBPB, apesar dos sinais fracos, apresentou um leve contraste para os sinais relativos às modificações de histonas. Os fatores MEF2A e P300, apesar da depleção suave, possuíram altos contrastes entre as regiões enriquecidas de ChIP-seq e não enriquecidas. O REST, como notado anteriormente, possuiu contrastes relevantes em relação à DNase, porém contrastes não tão precisos em relação às modificações de histonas.

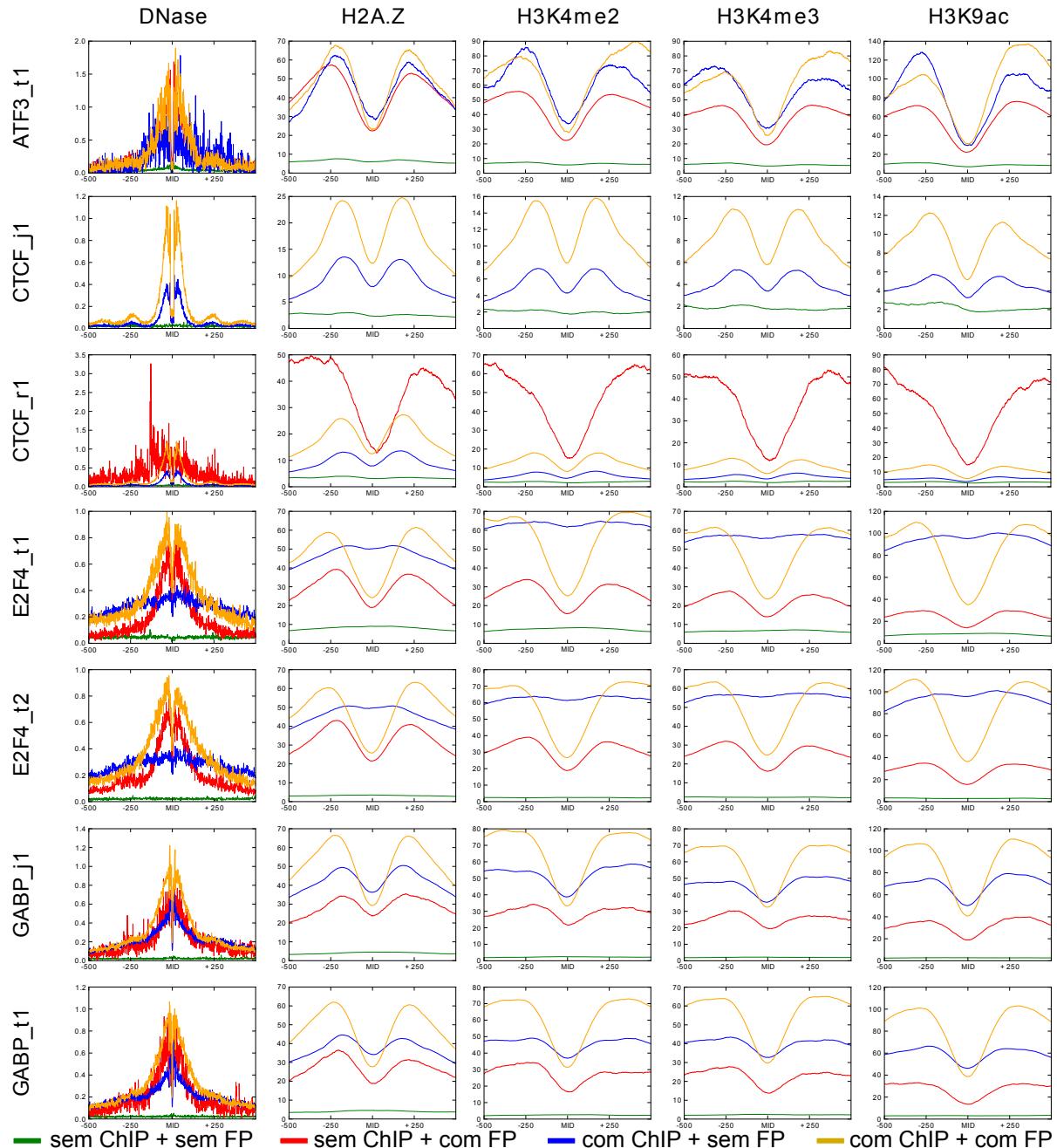
Além da intensidade entre MPBSs com e sem evidência de ChIP-seq e entre fatores de diferentes tipos, é interessante observar o formato em que o gráfico da média dessas regiões toma. As modificações das histonas, para o fator CTCF por exemplo, apresentam picos com comprimentos (frequência) menores do que os presentes no fator GABP ou E2F4. Para estes, a primeira subida e última descida (relativas ao aspecto bimodal dos padrões) não estão nem visíveis nesta janela de tamanho 1.000 bp para, por exemplo, as modificações H3K4me2 e H3K4me3. Tais padrões não foram especificamente analisados neste trabalho, porém podem representar um interessante estudo futuro, com hipótese de que o formato dos sinais epigenéticos ao redor de regiões enriquecidas de proteínas reflete o formato estrutural daquela proteína (os elementos regulatórios possuem *motifs* estruturais bem definidos).

A terceira análise consiste na visualização, para cada TF, dos sinais epigenéticos ao redor das 100 regiões de MPBSs com maiores *bit score* que: (1) não possuem evidência de ChIP-seq nem um *footprint* associado, isto é, verdadeiros negativos (linhas de cor verde); (2) não possuem evidência de ChIP-seq porém possuem um *footprint* associado, isto é, falsos positivos (linhas de cor vermelha); (3) possuem evidência de ChIP-seq porém não possuem *footprint* associado, isto é, falsos negativos (linhas de cor azul); (4) possuem evidência de ChIP-seq e *footprint*, isto é, verdadeiros positivos (linhas de cor amarela) (Figuras 5.4 e 5.5). Nestas figuras, são exibidos gráficos de linha contendo a média dos sinais para todas estas regiões sobre toda a extensão analisada.

Os rótulos dos fatores seguiram a mesma descrição dada para as Figuras 5.2 e 5.3. Caso existam menos de 10 sinais (do máximo de 100) para qualquer categoria descrita, a curva correspondente a esta categoria não é exibida, para que problemas relativos à computação da média de poucas regiões não enviesasse a visualização. Esse caso ocorreu para os fatores: CTCF com *motif* obtido no repositório Jaspar, REST com *motifs* obtidos nos repositórios Jaspar e Transfac.

O objetivo destes gráficos é analisar as previsões realizadas pelo modelo anterior, em relação aos sinais epigenéticos que se pretende inserir no modelo proposto. A partir destas análises o conjunto de histonas que seria utilizado no novo modelo foi determinado. Os fins de tal determinação foram apenas o teste empírico da hipótese proposta, e não a asserção determinística de quais histonas são melhores preditoras para cada caso. Um passo na direção deste tipo de informação será dado em estudos futuros.

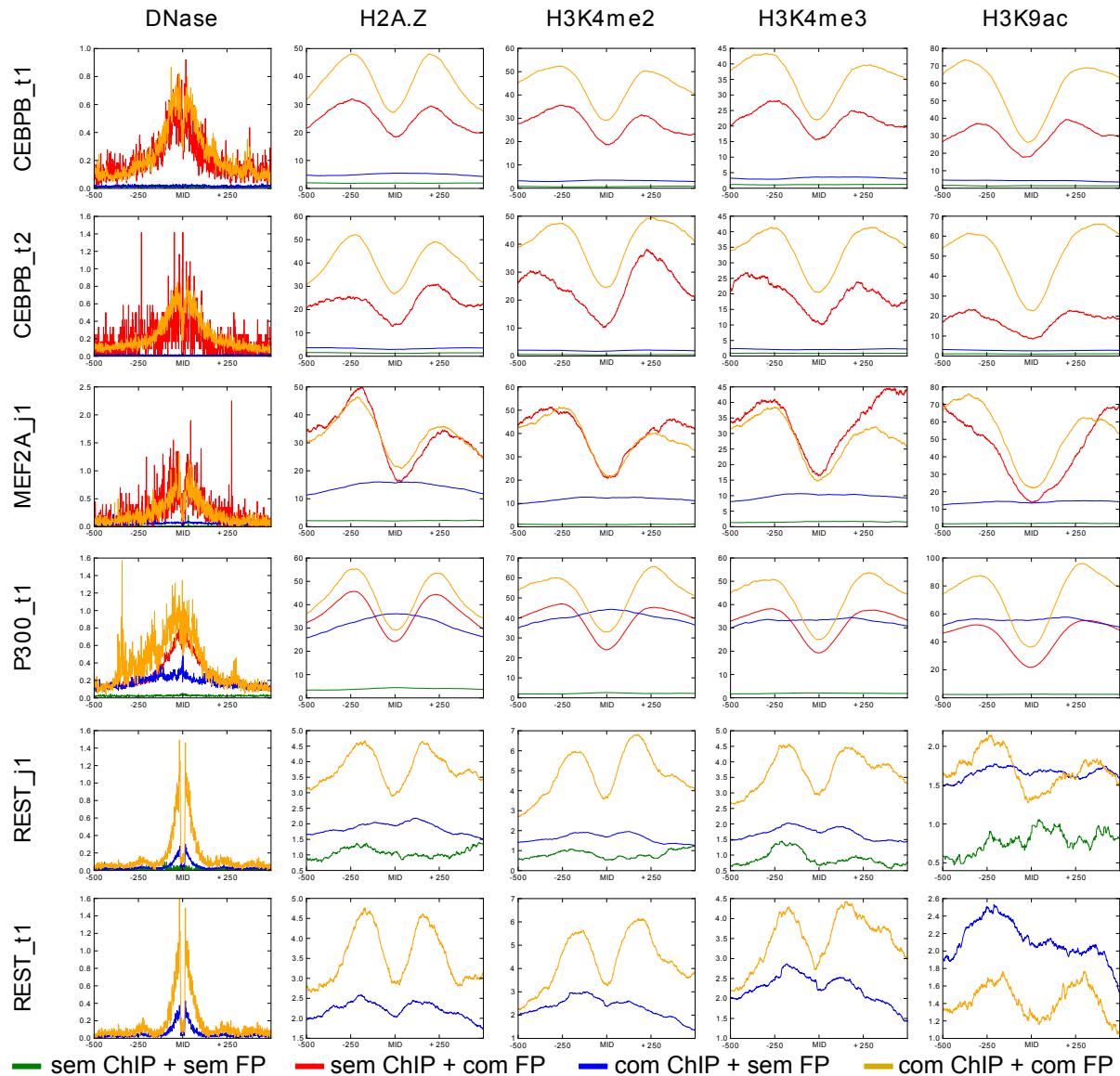
## 5.1. ANÁLISE DOS SINAIS EPIGENÉTICOS



**Figura 5.4:** Regiões de TFBS com e sem evidência de ChIP-seq e *footprint* associado  
**Pt. 1** - Análise dos sinais epigenéticos ao redor dos 100 MPBSs com maior *bit score* que possuem ou não possuem evidência de ChIP-seq e *footprint* associado. São analisadas regiões de 1000 bp, sendo necessárias pelo menos 10 regiões para cada categoria, para que o sinal seja exibido (evitando vieses estatísticos). Nesta figura, são exibidos os fatores de transcrição que apresentaram os sinais epigenéticos mais delineados dentre os fatores estudados.

## 5. RESULTADOS E DISCUSSÃO

---



**Figura 5.5: Regiões de TFBS com e sem evidência de ChIP-seq e footprint associado**  
**Pt. 2** - Análise dos sinais epigenéticos ao redor dos 100 MPBSs com maior *bit score* que possuem ou não possuem evidência de ChIP-seq e footprint associado. São analisadas regiões de 1000 bp, sendo necessárias pelo menos 10 regiões para cada categoria, para que o sinal seja exibido (evitando vieses estatísticos). Nesta figura, são exibidos os fatores de transcrição que apresentaram os sinais epigenéticos menos claros dentre os fatores estudados.

## **5.1. ANÁLISE DOS SINAIS EPIGENÉTICOS**

---

Através da observação dos gráficos para o fator ATF3, percebe-se que os sinais de falsos positivos (vermelho) se confundem com o sinal de verdadeiros positivos (amarelo) para a DNase e para a histona variante H2A.Z. Porém, para os outros sinais epigenéticos parece existir uma proximidade maior entre os sinais de falsos negativos (azul) e de verdadeiros positivos do que entre os verdadeiros positivos e falsos positivos, o que faria com que um modelo que utilizasse tais sinais ganhasse essa informação adicional. Alguns fatores, como o CEBP e MEF2A, entretanto, não possuem evidências interessantes para os sinais com evidência de ChIP e sem *footprint* (azul).

Para o fator CTCF, as curvas pareceram bastante consistentes, porém não é possível realizar inferências a respeito da adição de histonas no modelo, dado que os sinais de falsos positivos estão pouco representados. Para o caso do *motif* obtido no Jaspar, a quantidade de falsos positivos foi muito pequena, fazendo com que tal sinal fosse excluído da análise. Para o caso do *motif* obtido no Renlab, este sinal parece ter sido sobre-representado. Porém é possível observar a lógica recorrente de que os falsos negativos geralmente têm sinal mais baixo do que verdadeiros positivos, porém apresentam o mesmo formato de vale.

Os padrões médios observados para os fatores E2F4 e P300 possuem características semelhantes. No caso do E2F4, os verdadeiros positivos e falsos positivos se confundem, tornando a predição menos eficaz. Entretanto, os sinais relativos aos falsos negativos, apesar de não apresentarem tendência pico-vale-pico evidentes, possuem intensidades mais altas do que os falsos positivos (na mesma faixa dos verdadeiros positivos), o que poderia sinalizar um ponto positivo. Por outro lado, no caso do P300, os padrões são semelhantes porém a linha relativa aos falsos negativos está aproximadamente na mesma faixa dos falsos positivos, o que provavelmente acarretaria em piores inferências.

A análise dos padrões para o fator GABP fornece, assim como para o fator ATF3, um ponto positivo para a inserção das modificações de histonas. É possível visualizar que a linha representando os falsos positivos está bastante próxima da linha representando os falsos negativos para a DNase, porém para as histonas ela se apresenta consistentemente abaixo em todos os casos. É importante mencionar que uma análise relativa aos desvios padrões foi realizada, porém não exibidas nos gráficos pela dificuldade de leitura que ela apresentou. A análise de desvios padrões não demonstrou variância significativa entre tais sinais, porém espera-se que pelo menos os padrões tidos como falsos negativos sejam identificados pelo novo modelo, já que sua curvatura possui interseção consistente com as curvaturas dos verdadeiros positivos em todos os casos.

Assim como o fator CTCF, o fator REST possui *motif* grande e com grande quantidade de bases conservadas. Isso faz com que o número de falsos positivos não seja grande o suficiente para ser exibido nesses gráficos. Nos gráficos relativos ao *motif* obtido no Transfac, nem os

## 5. RESULTADOS E DISCUSSÃO

---

verdadeiros negativos (que são bem numerosos em outros casos) tiveram representatividade significativa.

Em geral, espera-se que as histonas acrescentem informações úteis ao novo modelo proposto. A análise destes gráficos para diversos fatores de transcrição diferentes mostra que esta é a tendência sobre uma quantidade razoável de fatores de transcrição. Na seção seguinte, tal hipótese será testada através do modelo descrito na Seção 4.5.

### 5.2 Acurácia do Modelo Proposto

Nesta seção, primeiramente serão mostradas estatísticas gerais a respeito da quantidade de regiões encontradas pelos métodos de enriquecimento, pelo *motif matching* e pela aplicação dos modelos. Então, serão apresentadas as tabelas correspondentes ao cálculo das estatísticas em relação à aplicação do modelo no genoma inteiro (Seção 4.5) e do *gold standard* definido na Seção 4.7. O objetivo da apresentação de tais resultados é a comparação do modelo anterior com o modelo proposto neste trabalho.

Na Tabela 5.1 são exibidas as quantidade de regiões preditas (isto é, *footprints*) utilizando ambos os modelos (prévio e proposto) e ambas as formas de treinamento (FMR1 e STAMP). O número total de regiões hypersensíveis à DNase I nas quais todos os métodos foram aplicados foi igual a 133.372. Todos os modelos foram aplicados somente nestas regiões, obtidas de forma idêntica às regiões enriquecidas de ChIP-seq (ver Seção 4.3).

**Tabela 5.1: Quantidade de *footprints* encontrados com cada modelo** – Nesta tabela são exibidas as quantidade de regiões preditas (#footprints) utilizando todos os modelos e formas de treinamento. Os modelos bivariados propostos são referenciados apenas pela histona correspondente.

Modelo	Treino	#footprints
DNase apenas	FMR1	109648
	STAMP	67758
H2A.Z	FMR1	422537
	STAMP	192274
H3K4me2	FMR1	436509
	STAMP	200293
H3K4me3	FMR1	475023
	STAMP	202496
H3K9ac	FMR1	460468
	STAMP	183744

## **5.2. ACURÁCIA DO MODELO PROPOSTO**

---

Devem ser consideradas duas informações contidas na Tabela 5.1. Primeiramente, pode-se perceber que os modelos propostos geram uma quantidade muito maior (até quase cinco vezes maior) de previsões do que o modelo baseado em DNase apenas. Esse fato possui algumas vantagens e desvantagens, que serão discutidas mais adiante. Também é possível observar que os modelos treinados com a abordagem STAMP produzem quantidades bem menores de previsões do que os modelos treinados com a abordagem FMR1. Novamente, as implicações serão descritas posteriormente.

São apresentadas, então, as tabelas contendo a comparação entre o método prévio e o novo modelo proposto. Para cada fator, são calculadas a sensibilidade ( $S_s$ ), especificidade ( $S_p$ ), *positive predictive value* ( $P_p$ ), *negative predictive value* ( $N_p$ ) e taxa de acerto ( $C_r$ ) (ver Tabela ??), relativas aos *footprints* gerados pela aplicação do modelo anterior e do modelo proposto. O modelo anterior foi replicado e aplicado com as ferramentas utilizadas neste projeto, para remover vieses gerados pelas mesmas. Em cada modelo foram aplicadas as duas formas de treinamento (FMR1 e STAMP).

As estatísticas apresentadas nas Tabelas 5.2 a 5.8 mostram que o modelo proposto, em geral, aumenta bastante a sensibilidade (em até 49.37% a mais) enquanto que apresenta uma pequena queda na especificidade (em, no máximo, 10.35%) (ver Tabela 5.9 a seguir). A taxa de acerto ( $C_r$ ) apresentou um aumento para os fatores CTCF e REST, enquanto para outros fatores os valores da precisão foram equivalentes. Isso ocorre, possivelmente, pela quantidade de exemplos negativos, para esses outros *motifs*, ser maior (dado que as PWMs têm qualidade inferior), fazendo com que a parcela de especificidade tenha uma maior contribuição na taxa de acerto geral do que a sensibilidade (ver Tabela ??). Em adição, para o fator REST, é interessante observar que houve grandes diferenças nos resultados entre PWMs provenientes de repositórios diferentes, mostrando que existe impacto relacionado com a qualidade dos *motifs*.

## 5. RESULTADOS E DISCUSSÃO

---

**Tabela 5.2: Resultados (em %) para o fator ATF3 (PWM obtida no Transfac)** – São exibidos resultados para o modelo prévio (DNase apenas) e para os modelos bivariados com DNase + modificação de histona (apenas o nome desta é exibido). Para cada modelo, ambas formas de treinamento são consideradas (FMR1 e STAMP). O melhor resultado para cada estatística é destacado em negrito.

Modelo	Treino	Sn	Sp	Pp	Np	Cr
DNase apenas	FMR1	58.75	96.8	10.28	99.73	96.57
	STAMP	71.25	<b>96.99</b>	<b>12.87</b>	99.82	<b>96.83</b>
H2A.Z	FMR1	31.25	94.34	3.33	99.55	93.95
	STAMP	70.0	90.32	4.31	99.79	90.19
H3K4me2	FMR1	32.5	92.66	2.69	99.55	92.29
	STAMP	76.25	89.68	4.41	<b>99.84</b>	89.6
H3K4me3	FMR1	35.0	92.08	2.68	99.56	91.72
	STAMP	<b>77.5</b>	89.32	4.33	<b>99.84</b>	89.25
H3K9ac	FMR1	25.0	93.11	2.21	99.5	92.69
	STAMP	67.5	89.88	3.99	99.77	89.74

**Tabela 5.3: Resultados (em %) para o fator CTCF (PWM obtida no Jaspar)** – São exibidos resultados para o modelo prévio (DNase apenas) e para os modelos bivariados com DNase + modificação de histona (apenas o nome desta é exibido). Para cada modelo, ambas formas de treinamento são consideradas (FMR1 e STAMP). O melhor resultado para cada estatística é destacado em negrito.

Modelo	Treino	Sn	Sp	Pp	Np	Cr
DNase apenas	FMR1	29.45	99.59	<b>99.87</b>	11.35	35.28
	STAMP	26.08	<b>99.86</b>	99.95	10.91	32.21
H2A.Z	FMR1	50.33	97.93	99.63	15.16	54.29
	STAMP	71.80	94.74	99.34	23.35	73.71
H3K4me2	FMR1	63.71	95.85	99.41	19.32	66.38
	STAMP	74.76	94.74	99.37	25.39	76.42
H3K4me3	FMR1	65.13	96.13	99.46	19.99	67.71
	STAMP	<b>75.45</b>	94.33	99.32	<b>25.83</b>	<b>77.02</b>
H3K9ac	FMR1	60.95	96.68	99.51	18.33	63.92
	STAMP	74.96	94.33	99.32	25.46	76.57

## 5.2. ACURÁCIA DO MODELO PROPOSTO

---

**Tabela 5.4: Resultados (em %) para o fator CTCF (PWM obtida no Renlab)** – São exibidos resultados para o modelo prévio (DNase apenas) e para os modelos bivariados com DNase + modificação de histona (apenas o nome desta é exibido). Para cada modelo, ambas formas de treinamento são consideradas (FMR1 e STAMP). O melhor resultado para cada estatística é destacado em negrito.

Modelo	Treino	Sn	Sp	Pp	Np	Cr
DNase apenas	FMR1	29.68	98.4	<b>98.82</b>	23.64	42.13
	STAMP	25.61	<b>98.61</b>	<b>98.82</b>	22.68	38.84
H2A.Z	FMR1	50.19	92.85	96.95	29.2	57.92
	STAMP	69.16	88.26	96.38	38.77	72.62
H3K4me2	FMR1	61.51	89.64	96.41	34.01	66.6
	STAMP	72.51	87.82	96.42	41.42	75.29
H3K4me3	FMR1	63.07	89.5	96.45	34.91	67.86
	STAMP	<b>72.98</b>	87.45	96.34	<b>41.73</b>	<b>75.6</b>
H3K9ac	FMR1	59.57	91.61	96.98	33.4	65.38
	STAMP	72.27	88.11	96.49	41.29	75.14

**Tabela 5.5: Resultados (em %) para o fator GABP (PWM obtida no Jaspar)** – São exibidos resultados para o modelo prévio (DNase apenas) e para os modelos bivariados com DNase + modificação de histona (apenas o nome desta é exibido). Para cada modelo, ambas formas de treinamento são consideradas (FMR1 e STAMP). O melhor resultado para cada estatística é destacado em negrito.

Modelo	Treino	Sn	Sp	Pp	Np	Cr
DNase apenas	FMR1	27.9	99.77	91.84	93.66	93.62
	STAMP	27.8	<b>99.86</b>	<b>94.96</b>	93.66	<b>93.69</b>
H2A.Z	FMR1	39.09	97.9	63.52	94.5	92.86
	STAMP	46.32	94.96	46.27	94.97	90.8
H3K4me2	FMR1	37.27	96.28	48.38	94.25	91.23
	STAMP	50.87	94.55	46.61	95.36	90.81
H3K4me3	FMR1	40.29	96.14	49.42	94.51	91.36
	STAMP	<b>53.11</b>	94.37	46.91	<b>95.56</b>	90.84
H3K9ac	FMR1	36.34	96.96	52.83	94.21	91.77
	STAMP	42.19	94.49	41.74	94.58	90.01

## 5. RESULTADOS E DISCUSSÃO

---

**Tabela 5.6:** Resultados (em %) para o fator GABP (PWM obtida no Transfac) – São exibidos resultados para o modelo prévio (DNase apenas) e para os modelos bivariados com DNase + modificação de histona (apenas o nome desta é exibido). Para cada modelo, ambas formas de treinamento são consideradas (FMR1 e STAMP). O melhor resultado para cada estatística é destacado em negrito.

Modelo	Treino	Sn	Sp	Pp	Np	Cr
DNase apenas	FMR1	26.26	99.75	86.62	95.61	95.46
	STAMP	25.19	<b>99.84</b>	<b>90.97</b>	95.56	<b>95.48</b>
H2A.Z	FMR1	38.48	97.84	52.45	96.25	94.37
	STAMP	44.81	95.36	37.48	96.53	92.41
H3K4me2	FMR1	36.73	96.44	38.99	96.09	92.95
	STAMP	49.38	95.04	38.16	96.8	92.37
H3K4me3	FMR1	40.61	96.36	40.9	96.32	93.1
	STAMP	<b>52.63</b>	94.87	38.89	<b>97.0</b>	92.4
H3K9ac	FMR1	36.01	97.04	43.04	96.07	93.48
	STAMP	41.62	94.97	33.9	96.33	91.85

**Tabela 5.7:** Resultados (em %) para o fator REST (PWM obtida no Jaspar) – São exibidos resultados para o modelo prévio (DNase apenas) e para os modelos bivariados com DNase + modificação de histona (apenas o nome desta é exibido). Para cada modelo, ambas formas de treinamento são consideradas (FMR1 e STAMP). O melhor resultado para cada estatística é destacado em negrito.

Modelo	Treino	Sn	Sp	Pp	Np	Cr
DNase apenas	FMR1	20.49	96.67	99.18	5.82	24.17
	STAMP	14.39	<b>98.33</b>	99.42	5.51	18.45
H2A.Z	FMR1	35.39	95.0	99.29	6.95	38.28
	STAMP	<b>55.21</b>	95.0	99.54	<b>9.73</b>	<b>57.13</b>
H3K4me2	FMR1	49.96	96.67	<b>99.66</b>	8.94	52.22
	STAMP	55.04	95.0	99.54	9.69	56.97
H3K4me3	FMR1	48.52	95.0	99.48	8.57	50.77
	STAMP	55.04	95.0	99.54	9.69	56.97
H3K9ac	FMR1	43.27	95.0	99.42	7.84	45.77
	STAMP	<b>55.21</b>	95.0	99.54	<b>9.73</b>	<b>57.13</b>

## 5.2. ACURÁCIA DO MODELO PROPOSTO

---

**Tabela 5.8: Resultados (em %) para o fator REST (PWM obtida no Transfac)** – São exibidos resultados para o modelo prévio (DNase apenas) e para os modelos bivariados com DNase + modificação de histona (apenas o nome desta é exibido). Para cada modelo, ambas formas de treinamento são consideradas (FMR1 e STAMP). O melhor resultado para cada estatística é destacado em negrito.

Modelo	Treino	Sn	Sp	Pp	Np	Cr
DNase apenas	FMR1	31.78	<b>100.0</b>	<b>100.0</b>	3.13	33.25
	STAMP	23.96	<b>100.0</b>	<b>100.0</b>	2.81	25.6
H2A.Z	FMR1	46.21	<b>100.0</b>	<b>100.0</b>	3.93	47.37
	STAMP	<b>69.44</b>	<b>100.0</b>	<b>100.0</b>	<b>6.72</b>	<b>70.1</b>
H3K4me2	FMR1	63.57	<b>100.0</b>	<b>100.0</b>	5.7	64.35
	STAMP	68.95	<b>100.0</b>	<b>100.0</b>	6.62	69.62
H3K4me3	FMR1	61.12	<b>100.0</b>	<b>100.0</b>	5.36	61.96
	STAMP	69.19	<b>100.0</b>	<b>100.0</b>	6.67	69.86
H3K9ac	FMR1	55.5	<b>100.0</b>	<b>100.0</b>	4.71	56.46
	STAMP	<b>69.44</b>	<b>100.0</b>	<b>100.0</b>	<b>6.72</b>	<b>70.1</b>

## 5. RESULTADOS E DISCUSSÃO

---

A Tabela 5.9 compara os resultados, em relação à sensibilidade e especificidade, de forma mais direta e analítica. Esta tabela compara a diferença entre os melhores resultados para o método proposto e os resultados para o método prévio, levando em consideração ambas as formas de treinamento. Esta tabela evidencia a proporção de quanto o método proposto aumentou a sensibilidade em razão da sensibilidade. Pode-se observar também que os maiores aumentos da sensibilidade ocorrem ao utilizar o método STAMP para treinar os modelos propostos. Também é interessante o fato de que as diferenças entre os modelos foram próximas para *motifs* diferentes de um mesmo fator de transcrição, evidenciando a robustez dos resultados em relação às análises considerando um fator específico.

**Tabela 5.9: Comparação da sensibilidade e especificidade entre o modelo prévio e o proposto** – Cada célula exibe (em %) a diferença, na sensibilidade ou especificidade, entre o melhor resultado obtido entre um dos métodos propostos e o resultado para o método prévio. Diferenças positivas representam melhoria dos resultados.

Fatores	Treino	Sn	Sp
ATF3 (Transfac)	FMR1	-23.75	-2.46
	STAMP	+6.25	-6.67
CTCF (Jaspar)	FMR1	+35.68	-1.66
	STAMP	+49.37	-5.12
CTCF (Renlab)	FMR1	+33.39	-5.55
	STAMP	+47.37	-10.35
GABP (Jaspar)	FMR1	+12.39	-1.87
	STAMP	+25.31	-4.9
GABP (Transfac)	FMR1	+14.35	-1.91
	STAMP	+27.44	-4.48
REST (Jaspar)	FMR1	+29.47	0.0
	STAMP	+40.82	-3.33
REST (Transfac)	FMR1	+31.79	0.0
	STAMP	+45.48	0.0

Considera-se que o modelo proposto foi bem sucedido pelo fato de que o reconhecimento de um número maior de regiões corretas (maior sensibilidade) é preferível, nestes casos, sobre a rejeição de tais TFBSS verdadeiros em razão de um aumento na especificidade. Tais resultados são utilizados por exemplo, como nos estudos [Barski *et al.*, 2007; Heintzman *et al.*, 2007; Hon *et al.*, 2009; Ramsey *et al.*, 2010], para criar mapas regulatórios consistentes, que possuem em sua natureza a preferência por uma quantidade maior de marcadores positivos.

## 5.2. ACURÁCIA DO MODELO PROPOSTO

---

Neste momento, é necessário traçar um paralelo dos resultados com o número de regiões preditas pelos modelos (Tabela 5.1). O número de *footprints* identificados pelos modelos baseados em FMR1 é grande devido ao fato de que os parâmetros estimados correspondem apenas à uma região anotada (a região promotora do gene FMR1 – ver Seção 4.6). A quantidade de *footprints* relacionados aos modelos propostos é bastante alta pelo fato de que, nesta região, o sinal das histonas não era tão intenso. Isto mostra a dificuldade da aplicação de métodos, como o FMR1, baseados na realização de experimentos biológicos custosos adicionais ou na busca por tais resultados na literatura. Estima-se que o sucesso da aplicação do método STAMP, bem como a identificação de uma quantidade de previsões mais real, têm origem no fato de que, por permitir uma quantidade maior de regiões anotadas, os parâmetros dos modelos são estimados de forma mais precisa.

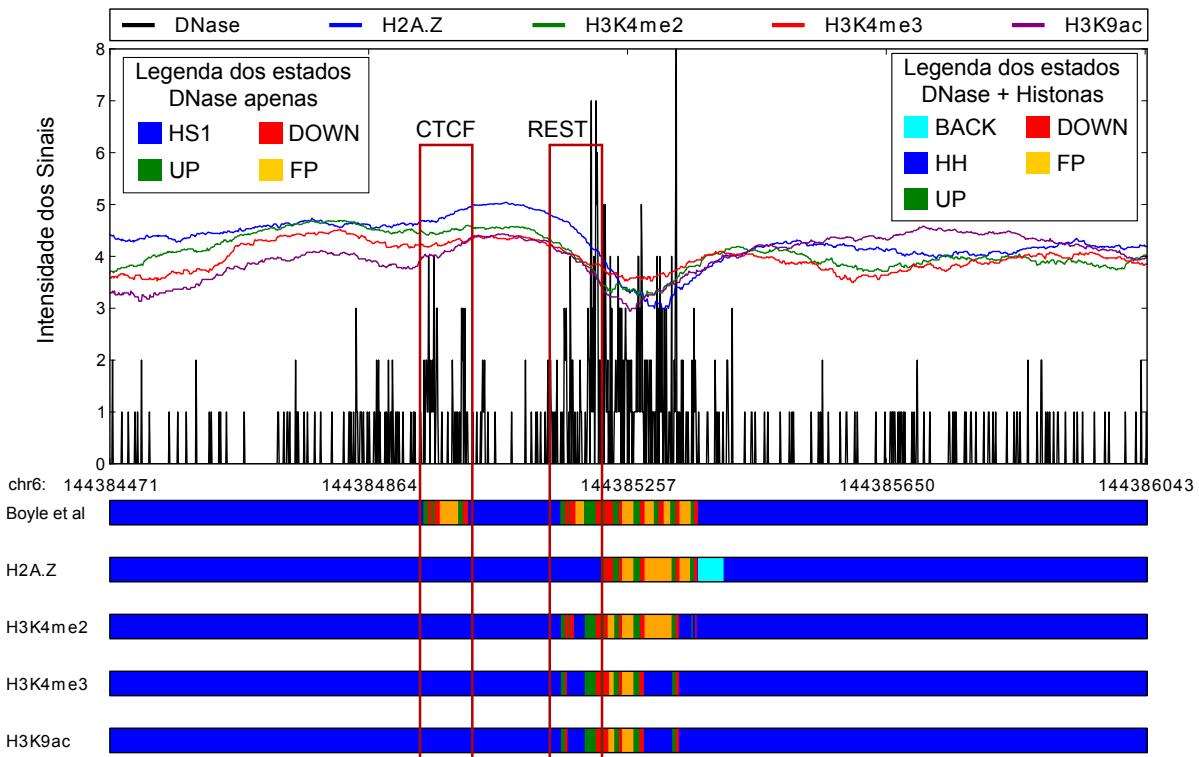
Foram observadas duas formas pelas quais o modelo proposto é capaz de produzir melhores resultados. A primeira, que aconteceu numa grande escala, corresponde ao aumento no número de verdadeiros positivos. Observou-se que as regiões de vale das histonas proveram uma permissividade maior de entrada no estado de *footprint* em regiões com baixos sinais de digestão de DNase. A segunda forma, que ocorreu em escala menor, corresponde à desconsideração de alguns falsos positivos críticos em regiões onde as histonas tinham sinais mais elevados. Esta segunda forma foi capaz de manter a especificidade em níveis altos, ainda que não melhores do que no modelo prévio. A Figura 5.6 mostra um exemplo relativo à este segundo ponto discutido.

A partir da Figura 5.6, podemos visualizar o quanto preciso é o modelo. Uma das principais vantagens da abordagem utilizada é que ela tira proveito do aspecto espacial dos dados, isto é, das características que os sinais epigenéticos tomam, ao longo do genoma. Além de prover uma base probabilística robusta, o aproveitamento espacial faz com que sejam possíveis previsões com alta precisão, dado que os sinais possuem boa resolução. Métodos que ignoram dados espaciais, apenas levando em consideração características obtidas ao se observar as regiões analisadas como um todo, não possuem tal precisão. Essa é uma das principais críticas ao método descrito em [Pique-Regi *et al.*, 2011].

A nova forma de treinamento (STAMP) foi aplicada ao modelo prévio e a forma de treinamento prévia (FMR1) foi aplicada aos novos modelos com objetivo de verificar o impacto das técnicas de treinamento nos resultados. Melhorias nos novos modelos poderiam ser devidas simplesmente ao uso de uma forma de treinamento mais consistente do que pela inserção de sinais epigenéticos. Observou-se que a nova forma de treinamento contribuiu para algumas estatísticas maiores, porém que ela não parece ter sido o motivo dos melhoramentos observados. Um exemplo de evidência neste sentido é o fato de que o novo método de treinamento de fato aumentou a especificidade do modelo prévio, reduzindo a sensibilidade do mesmo na maioria dos casos, o que corresponde ao caminho inverso da melhoria observada com a adição das modificações de histonas. É interessante observar também que as melhores estatísticas variam de acordo com a adição das diferentes modificações de histonas, para diferentes modelos. Para os ativadores ATF3 e

## 5. RESULTADOS E DISCUSSÃO

---

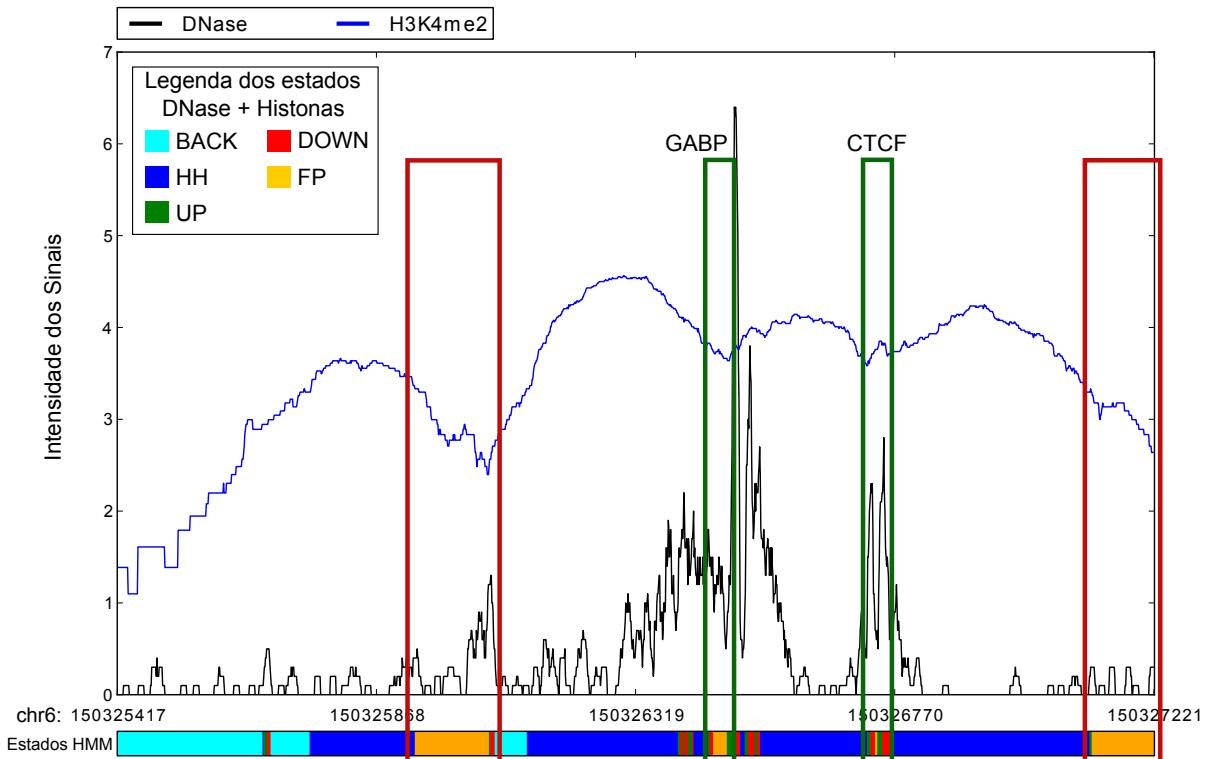


**Figura 5.6: Exemplo de uma região com resultados melhorados pelo modelo proposto**  
- São exibidos os sinais epigenéticos em uma região do cromossomo 6. Os mapas de cores abaixo do gráfico que mostra a intensidade dos sinais, demonstram os estados do HMM para cada coordenada, com cores correspondentes ao modelo exibido na Figura 4.4. Os retângulos vermelhos demonstram as duas regiões de falsos positivos pelo método prévio, que foram mascaradas pela adição das histonas na nova abordagem.

GABP e o insulador CTCF, as melhores sensibilidades com o novo método de treinamento foram observadas com a adição das histonas H3K4me2 e H3K4me3. Enquanto que o repressor REST obteve as melhores sensibilidades para as histonas H2A.Z e H3K9ac.

Apesar dos bons resultados observados, o modelo possui um problema que ocorre com mais frequência do que no modelo prévio. Esse problema consiste em previsões demasiadamente extensas. Em detalhes, o propósito desta abordagem ao problema de identificação de TFBSs consiste em utilizar tais dados de alta resolução para prever posições bastante específicas onde os TFs se ligam. Esses trechos preditos variam entre 5 e 30 bp em média, não devendo ser maior do que 50 bp. Porém a baixa resolução correspondente à inserção das histonas fez com que alguns fragmentos preditos tivessem mais do que 50 bp, às vezes chegando a 200 bp. Dessa forma, a ideologia do problema é ferida por tais previsões muito extensas. Estudos futuros pretendem focar nas diferenças de resolução entre os sinais para que se chegue a um consenso ideal. A Figura 5.7 mostra um exemplo dessas previsões demasiadamente longas.

### 5.3. TEMPO DE EXECUÇÃO E ARMAZENAMENTO



**Figura 5.7: Exemplo do problema das previsões amplas** - São exibidos os sinais de DNase e da modificação de histona H3K4me2 para uma região do cromossomo 6. O mapa de cores demonstra os estados do HMM para cada coordenada a respeito da aplicação do modelo bivariado baseado em DNase + H3K4me2, com cores correspondentes ao modelo exibido na Figura 4.4. Os retângulos verdes mostram regiões corretamente preditas, porém os retângulos vermelhos mostram regiões inapropriadamente extensas para a proposta de resolução deste problema.

## 5.3 Tempo de Execução e Armazenamento

Estima-se que o projeto necessitou de um total de 1.874 horas computacionais para ser executado completamente, sem levar em consideração os testes realizados ao longo do processo experimental. A Tabela 5.10 exibe o tempo computacional mínimo, médio e máximo para a realização de todas as etapas do processo. O tempo mínimo e máximo correspondem, respectivamente, aos menores e maiores tempos relativos à aplicação de uma tarefa que envolve diversas instâncias. Por exemplo, a aplicação do método *motif matching* era realizada para cada fator de transcrição, sendo estes considerados as instâncias neste caso. Esta tabela também exibe a quantidade de memória necessária para executar cada fase. No fim, a tabela exibe o tempo total, considerando a soma dos tempos relativos à multiplicação de todos os tempos individuais pelo número de instâncias. Pode-se dizer que o projeto só pôde ser realizado devido ao uso de um *grid engine* com 60 cores, que permitiu a execução em paralelo de várias fases do estudo.

A Tabela 5.11 exibe o tamanho médio para cada tipo de dado de entrada e saída dos

## 5. RESULTADOS E DISCUSSÃO

---

**Tabela 5.10: Tempo de execução e memória** – São exibidos os tempos de execução e quantidade de memória, mínimo (min), médio (med) e máximo (max), para cada etapa do processo experimental. Todos os valores desta tabela correspondem ao tempo de execução de uma instância da respectiva etapa, com exceção da linha *Total*, onde são exibidos o tempo total considerando todas as instâncias de todas as etapas. Quando maiores que 1h, os tempos mínimo e máximo foram truncados para a hora mais próxima. O total em relação à memória consumida corresponde ao máximo de memória necessária considerando as fases do experimento.

Etapa	Tempo			Memória		
	min	max	med	min	max	med
Motif Matching	8:00	14:00	11:47	413MB	413MB	413MB
Enriquecimento	15:00	35:00	21:32	1821MB	1849MB	1839MB
Contagem Bruta	5:00	12:00	7:45	1500MB	1600MB	1530MB
Normalização	9:00	9:00	9:00	1700MB	1700MB	1700MB
Savitzky-Golay	5:00	5:00	5:00	1500MB	1500MB	1500MB
Treino FMR1	1:00	3:00	2:21	812MB	835MB	814MB
Treino STAMP	0:25	0:37	0:28	812MB	841MB	815MB
Aplicação HMM	16:00	19:00	17:03	512MB	540MB	535MB
Validação	3:00	6:00	5:12	500MB	514MB	511MB
Gráficos	28:00	57:00	31:41	1900MB	1900MB	1900MB
Total	1874:00			1900MB		

**Tabela 5.11: Espaço necessário para armazenamento** – É exibido o espaço necessário para armazenar os arquivos que representam os dados e resultados utilizados neste projeto. É descrito o tipo (entrada ou saída), o nome do dado, os formatos nos quais o mesmo poderia se encontrar, o espaço médio (aproximado) necessário para armazenar um instância (Ind.), o número de instâncias (Inst.) e o espaço total médio necessário para armazenar os dados (Total).

Tipo	Dados	Formatos	Ind.	Inst.	Total
Entrada	DNase-seq	bed&wig	10GB	1	10GB
	ChIP-seq Histonas	bed	14GB	4	56GB
	ChIP-seq TFBS	bed&wig	14GB	8	112GB
	PWM	pwm	<1MB	13	10MB
Saída	MPBSs	bed	1GB	13	13GB
	Regiões enriquecidas	bed	0.5GB	9	4.5GB
	Sinais processados	bw	20GB	5	100GB
	Resultados	bed&txt	0.2GB	10	2GB
	Gráficos	eps	1GB	39	39GB

## **5.4. CONSIDERAÇÕES FINAIS**

---

métodos. Nesta tabela, são definidos os formatos desses dados, o tamanho médio para cada instância individual (Ind), o número de instâncias para cada dado (Inst) e o tamanho total considerando a todas as instâncias (Grp). Pode-se dizer que, em média, foram necessários 340 GB de armazenamento para a execução apropriada deste estudo, desconsiderando todos os arquivos gerados durante as fases de teste.

Três tipos de dados principais foram utilizados no decorrer do projeto. O primeiro tipo, chamado *bed*, consiste em um arquivo de texto simples contendo, em cada linha, informações de coordenadas genômicas. O tamanho de tais arquivos variou entre pequeno (por exemplo, TFBSS para um fator com *motif* de alta qualidade, isto é, poucos TFBSS) e grande (por exemplo, os fragmentos alinhados advindos das técnicas de DNase-seq ou ChIP-seq). O segundo tipo, chamado *wig* ou *wiggle*, consiste em um arquivo de texto simples contendo um valor de ponto flutuante para cada coordenada genômica de interesse. O tamanho de tais arquivos foi, em geral, grande, correspondendo principalmente aos sinais genômicos durante a etapa de contagem, normalização e aplicação do método de Savitzky-Golay. Tal tipo de arquivo pode ser comprimido em um formato nomeado *bw* ou *bigwig*. Finalmente, temos os arquivos *pwm* que representavam as PWMs para cada fator de transcrição analisado. Tais arquivos são geralmente pequenos, contendo apenas as informações de afinidade (ponto flutuante) para cada um dos quatro nucleotídeos e para cada posição do *motif* (não maior do que 20 bases). Os outros formatos mencionados são de uso comum.

## **5.4 Considerações Finais**

Neste capítulo foram exibidos os gráficos e tabelas referentes aos resultados obtidos neste estudo. Foram exibidos os gráficos necessários para a análise de regiões de interesse envolvendo MPBSs, regiões enriquecidas em ChIP-seq para os fatores de transcrição e resultados do modelo anterior. Além disso, após mostrar estatísticas gerais relacionadas com a quantidade de regiões produzidas durante o processo experimental, foram descritas as estatísticas avaliadas a partir da aplicação do modelo anterior e do modelo proposto. Finalmente, foi realizada uma discussão referente ao tempo computacional, processamento e armazenamento necessários durante a execução do projeto.

Após a apresentação dos resultados, em cada seção foram realizadas discussões a respeito dos mesmos. Primeiramente, foram discutidos os gráficos que visualizam tendências médias nos sinais epigenéticos em diversas regiões de interesse (e combinações dessas regiões). Após isso, foram discutidos os resultados da aplicação do método anterior e do método proposto. Foram apontadas as formas como o método proposto melhorou as previsões e também as limitações deste novo modelo. Finalmente, discutiu-se a infraestrutura necessária para realização de um projeto deste gênero.

# 6

## Conclusão

### 6.1 Objetivos Atingidos

Neste projeto de pesquisa foi proposto um método para melhorar a identificação de sítios de ligação para fatores de transcrição utilizando dados relativos à digestão da DNase e modificações de histonas. Tal abordagem é baseada no fato de que tais fatores epigenéticos são capazes de descrever regiões de cromatina descondensada, local com alta densidade de sítios de ligação. Além do método probabilístico, isto é, o modelo escondido de Markov, foi proposto um novo método de treinamento baseado na ferramenta STAMP, aumentando a viabilidade de regiões nas quais o HMM pode ser treinado.

Previamente à aplicação do modelo, foram criados três tipos de gráficos para melhor entender o comportamento dos sinais epigenéticos: (1) considerando regiões de MBPSs; (2) considerando a junção entre MPBSs e evidência de ChIP-seq; (3) considerando MPBSs, ChIP-seq e as predições realizadas pelo método prévio. Tais gráficos proveram as ideias necessárias para a construção do modelo probabilístico, integrando diferentes sinais epigenéticos. É importante observar que outros tipos de análises foram realizadas. Por exemplo, em relação às predições do modelo anterior em regiões específicas (e não médias de várias regiões). Porém tais resultados são bastante numerosos e são perfeitamente sumarizados pelos gráficos exibidos.

A criação do modelo foi realizada em várias etapas de tentativa e erro. O modelo preditivo que apresentou resultados mais próximos do que se esperava, durante as etapas experimentais, foi comparado ao método prévio e obteve algumas vantagens. Em especial, o método proposto aumentou a sensibilidade em níveis consideráveis enquanto sofreu uma pequena redução na especificidade. Através dos pontos discutidos no capítulo anterior, o novo método foi considerado bem sucedido. Além disso, é possível visualizar, graficamente, como os sinais de modificações

## **6.2. DIFICULDADES E LIMITAÇÕES DE ESCOPO**

---

de histonas ajudam na predição de alta resolução da DNase, fornecendo evidências a favor de abordagens integrativas de dados.

### **6.2 Dificuldades e Limitações de Escopo**

Os principais dados utilizados neste projeto foram obtidos no repositório ENCODE. Tais dados possuem uma restrição de uso que consiste em uma janela de tempo a partir do momento que são disponibilizados. Isso fez com que alguns dados não fossem reportados, e continuamos esperando tal liberação. Além disso, a criação do conjunto de validação possui a restrição de que os PWMs obtidos nos repositórios de *motifs* deveriam ter também dados de ChIP-seq para os fatores correspondentes. Entretanto, tal dificuldade não foi crítica, isto é, um número razoável de fatores pôde ser testado, expressando as tendências gerais de ambos os modelos de forma acurada.

Outra limitação está relacionada ao tamanho dos dados epigenéticos em larga escala, o que limita o numero de células e sinais considerados no estudo. Por exemplo, os dados do tipo wig (*wiggle*) com sinais de modificação de histonas são bem grandes (ver Tabela 5.11), fazendo com que a análise em mais de uma linha celular tenha uma alto custo computacional e de armazenamento. Para os dados discutidos aqui, foram necessários 340 GB de armazenamento e 1.874 horas de computação (ver Seção 5.3). Em especial, o tempo computacional só foi possível devido ao uso de um *grid engine* com 60 cores.

Apesar do modelo proposto ter contribuído para resultados mais interessantes do ponto de vista metodológico, alguns pontos negativos podem ser observados. A introdução de outra dimensão faz com que o procedimento, de uma forma geral, tome mais tempo para executar todas as etapas. Entretanto, como apontado na Seção 5.2, houve alguns casos onde as previsões feitas pelo modelo proposto foram mais extensas do que o esperado. Isso corresponde a um desvio na ideia de identificação absoluta de TFBSS defendida por Boyle et al. Estudos futuros deverão levar essa característica em consideração.

### **6.3 Trabalhos Futuros**

A primeira característica dos trabalhos futuros consiste no aumento do número de linhas celulares, modificações de histonas e fatores de transcrição, sobre os quais os métodos serão aplicados. Com o crescimento do repositório ENCODE, e de outras iniciativas do gênero, mais dados estarão disponíveis para serem utilizados, aumentando o leque de possibilidades. A análise de um número maior de modificações de histonas e de fatores de transcrição já é diretamente possível,

## **6. CONCLUSÃO**

---

assim que tais dados estiverem disponíveis nos repositórios mencionados (o que deverá acontecer num futuro próximo [Rosenbloom *et al.*, 2011]). A análise em um número maior de linhas celulares, entretanto, está completamente condicionada à capacidade computacional à disposição. A linha celular K562 foi escolhida por possuir os dados para a maior variedade de histonas e fatores entre todas as outras. Com o futuro aumento na capacidade computacional e nos experimentos realizados em outras linhas celulares, os métodos poderão ser aplicados e testados de forma mais extensa.

Além dos dados epigenéticos, métodos atuais estão utilizando outras informações como conservação e afinidade de ligação do fator baseado na sequência genômica [Pique-Regi *et al.*, 2011] ou regiões de aplicação [Won *et al.*, 2010]. Tal integração adicional pretende ser levada em consideração na modelagem futura de sistemas probabilísticos. Extensões diretas do modelo proposto, por exemplo, já poderiam utilizar informações de afinidade de ligação (isto é, o *bit score* do *motif matching*) a priori, ou a análise estatística mais robusta da ferramenta STAMP.

Em termos experimentais, pretende-se realizar uma análise consistindo na verificação do impacto de cada característica epigenética na predição de TFBSSs. Tais estudos procurariam padrões epigenéticos ao redor de MPBSs com e sem evidência de ChIP-seq e tentaria separá-los, utilizando alguma abordagem de aprendizado de máquina, através de combinações de diferentes sinais epigenéticos. Tal abordagem também poderia ser cuidadosamente estudada para que pudesse ser um possível classificador, aplicado ao reconhecimento de TFBSSs, utilizando as características epigenéticas e as informações de afinidade de ligação. Além disso, outra ideia que se pretende explorar consiste na relação entre padrões epigenéticos e diferentes atributos dos fatores de transcrição (tais como suas funções ou família proteica). Estudos deste gênero podem contribuir para a melhoria futura de sistemas de identificação de sítios de ligação de fatores de transcrição.

# Referências

- ALBERTS, B. (2007). *Molecular Biology of the Cell*. Other, 5th edn. 2, 7, 8, 9, 10
- ALLIS, C., JENUWEIN, T. & REINBERG, D. (2007). *Epigenetics*. Cold Spring Harbor Laboratory Press. 7, 8, 12, 33, 34, 36
- BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T.Y., SCHONES, D.E., WANG, Z., WEI, G., CHEPELEV, I. & ZHAO, K. (2007). High-Resolution Profiling of Histone Methylation in the Human Genome. *Cell*, **129**, 823–837. 4, 32, 34, 40, 82
- BILMES, J. (1997). A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. 42
- BISHOP, C.M. (2006). *Pattern recognition and machine learning*. Springer, 1st edn. 42
- BOYLE, A.P., DAVIS, S., SHULHA, H.P., MELTZER, P., MARGULIES, E.H., WENG, Z., FUREY, T.S. & CRAWFORD, G.E. (2008a). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322. 3, 37, 39
- BOYLE, A.P., GUINNEY, J., CRAWFORD, G.E. & FUREY, T.S. (2008b). F-seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**, 2537–2538. 37, 55, 56
- BOYLE, A.P., SONG, L., LEE, B.K., LONDON, D., KEEFE, D., BIRNEY, E., IYER, V.R., CRAWFORD, G.E. & FUREY, T.S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research*, **21**, 456–464. 2, 3, 37, 39, 40, 55, 57, 58, 59, 60, 61, 62, 63, 64, 67
- BRYNE, J.C.C., VALEN, E., TANG, M.H.E.H., MARSTRAND, T., WINTHER, O., DA PIEDADE, I., KROGH, A., LENHARD, B. & SANDELIN, A. (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic acids research*, **36**, D102–D106. 56
- BUCK, M.J. & LIEB, J.D. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349–360. 3
- COCK, P.J.A., ANTАО, T., CHANG, J.T., CHAPMAN, B.A., COX, C.J., DALKE, A., FRIEDBERG, I., HAMELRYCK, T., KAUFF, F., WILCZYNKI, B. & DE HOON, M.J.L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423. 57
- CRAWFORD, G.E., HOLT, I.E., MULLIKIN, J.C., TAI, D., BLAKESLEY, R., BOUFFARD, G., YOUNG, A., MASILO, C., GREEN, E.D., WOLFSBERG, T.G., COLLINS, F.S. & NATIONAL INSTITUTES OF HEALTH INTRAMURAL SEQUENCING CENTER (2004). Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 992–997. 3, 36, 39
- CRAWFORD, G.E., DAVIS, S., SCACHERI, P.C., REINAUD, G., HALAWI, M.J., ERDOS, M.R., GREEN, R., MELTZER, P.S., WOLFSBERG, T.G. & COLLINS, F.S. (2006a). DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature methods*, **3**, 503–509. 3, 39
- CRAWFORD, G.E., HOLT, I.E., WHITTLE, J., WEBB, B.D., TAI, D., DAVIS, S., MARGULIES, E.H., CHEN, Y., BERNAT, J.A., GINSBURG, D., ZHOU, D., LUO, S., VASICEK, T.J., DALY, M.J., WOLFSBERG, T.G. & COLLINS, F.S. (2006b). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research*, **16**, 123–131. 3, 39
- CREYGHTON, M.P., CHENG, A.W., WELSTEAD, G.G., KOOISTRA, T., CAREY, B.W., STEINE,

## REFERÊNCIAS

---

- E.J., HANNA, J., LODATO, M.A., FRAMPTON, G.M., SHARP, P.A. & ET AL. (2010). Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 21931–21936. 34
- CUELLAR-PARTIDA, G., BUSKE, F.A., MCLEAY, R.C., WHITINGTON, T., NOBLE, W.S. & BAILEY, T.L. (2012). Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, **28**, 56–62. 2, 3, 4, 40, 55, 64
- DNA SEQUENCING CONSORTIUMS (2012). Dna sequencing website. <http://www.dnasequencing.org>. 1
- DROUIN, R., ANGERS, M., DALLAIRE, N., ROSE, T.M., KHANDJIAN, E.W. & ROUSSEAU, F. (1997). Structural and functional characterization of the human fmr1 promoter reveals similarities with the hnrrnp-a2 promoter region. *Human Molecular Genetics*, **1**, 91–96. 63
- DUDA, R.O., STORK, D.G. & HART, P.E. (2000). *Pattern classification*. Wiley, 2nd edn. 42
- DURBIN, R., EDDY, S.R., KROGH, A. & MITCHISON, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. 42
- DYMARSKI, P., ed. (2011). *Hidden Markov Models, Theory and Applications*. InTech. 42
- ERNST, J. & KELLIS, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, **28**, 817–825. 4, 40
- ESSIEN, K., VIGNEAU, S., APRELEVA, S., SINGH, L., BARTOLOMEI, M. & HANNENHALLI, S. (2009). CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. *Genome Biology*, **10**, R131+. 57
- FELSENFELD, G. & GROUDINE, M. (2003). Controlling the double helix. *Nature*, **421**, 448–453. 35
- GORRY, P.A. (1990). General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method. *Analytical Chemistry*, **62**, 570–573. 59
- GRANT, P.A. (2001). A tale of histone modifications. *Genome biology*, **2**, reviews0003.1–reviews0003.6. 34
- GROSS, D.S. & GARRARD, W.T. (1988). Nuclease hypersensitive sites in chromatin. *Annual Review of Biochemistry*, **57**, 159–197. 3
- GUTTMAN, M., GARBER, M., LEVIN, J.Z., DONAGHEY, J., ROBINSON, J., ADICONIS, X., FAN, L., KOZIOL, M.J., GNIRKE, A., NUSBAUM, C., RINN, J.L., LANDER, E.S. & REGEV, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature biotechnology*, **28**, 503–510. 56
- HAIR, J.F., TATHAM, R.L., ANDERSON, R.E. & BLACK, W. (1998). *Multivariate Data Analysis*. Prentice Hall, 5th edn. 42
- HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J.H. (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics, Springer. 42
- HE, H.H., MEYER, C.A., CHEN, M.W., JORDAN, V.C., BROWN, M. & LIU, X.S. (2012). Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Research*, **22**, 1015–1025. 39, 40
- HEINTZMAN, N.D., STUART, R.K., HON, G., FU, Y., CHING, C.W., HAWKINS, R.D., BARRERA, L.O., VAN CALCAR, S., QU, C., CHING, K.A., WANG, W., WENG, Z., GREEN, R.D., CRAWFORD, G.E. & REN, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, **39**, 311–318. 4, 32, 40, 82
- HON, G., WANG, W. & REN, B. (2009). Discovery and Annotation of Functional Chromatin Signatures in the Human Genome. *PLoS Comput Biol*, **5**, e1000566+. 4, 32, 35, 40, 82
- KEENE, M.A., CORCES, V., LOWENHAUPT, K. & ELGIN, S.C. (1981). Dnase i hypersensitive sites in drosophila chromatin occur at the 5' ends of regions of transcription. *Proceedings of the National*

## REFERÊNCIAS

- Academy of Sciences of the United States of America, **78**, 143–146. 3
- KENT, W.J., SUGNET, C.W., FUREY, T.S., ROSKIN, K.M., PRINGLE, T.H., ZAHLER, A.M. & HAUSLER, D. (2002). The Human Genome Browser at UCSC. *Genome Research*, **12**, 996–1006. 54
- LASSIG, M. (2007). From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinformatics*, **8**, S7+. 4
- LEACH, R.A., CARTER, C.A. & HARRIS, J.M. (1984). Least-squares polynomial filters for initial point and slope estimation. *Analytical Chemistry*, **56**, 2304–2307. 59
- LESK, A.M. (2005). *Introduction to bioinformatics*. Oxford University Press. 42
- LEVIN, D.A., PERES, Y. & WILMER, E.L. (2008). *Markov Chains and Mixing Times*. American Mathematical Society, 1st edn. 42
- LEWIN, B. (2003). *Genes VIII*. Benjamin Cummings, united states ed edn. 7, 8
- LODISH, H., BERK, A., KAISER, C.A., KRIEGER, M., SCOTT, M.P., BRETSCHER, A., PLOEGH, H. & MATSUDAIRA, P. (2007). *Molecular Cell Biology*. W. H. Freeman, 6th edn. 3, 7, 8, 11, 13, 14, 15, 28
- LUO, J., YING, K., HE, P. & BAI, J. (2005). Properties of savitzky golay digital differentiators. *Digital Signal Processing*, **15**, 122–136. 59
- MADDEN, H.H. (1978). Comments on the Savitzky-Golay convolution method for least-squares fit smoothing and differentiation of digital data. *Anal.Chem.*, **50**, 1383–1386. 59
- MAHONY, S. & BENOS, P.V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic acids research*, **35**, gkm272–258. 5, 63
- MASTON, G.A., EVANS, S.K. & GREEN, M.R. (2006). Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics*, **7**, 29–59. 2, 8, 19, 20, 22
- MATYS, V., KEL-MARGOLIS, O.V., FRICKE, E., LIEBICH, I., LAND, S., BARRE-DIRRIE, A., REUTER, I., CHEKMENEV, D., KRULL, M., HORNISCHER, K., VOSS, N., STEGMAIER, P., LEWICKI-POTAPOV, B., SAXEL, H., KEL, A.E. & WINGENDER, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, **34**, D108–D110. 56
- MITCHELL, T.M. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1st edn. 42
- NEWBURGER, D.E. & BULYK, M.L. (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, **37**, D77–D82. 57
- PARK, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, **10**, 669–680. 3, 29, 37, 38
- PIQUE-REGI, R., DEGNER, J.F., PAI, A.A., GAFFNEY, D.J., GILAD, Y. & PRITCHARD, J.K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, **21**, 447–455. 3, 4, 40, 55, 64, 83, 90
- PRESS, W.H., TEUKOLSKY, S.A., VETTERLING, W.T. & FLANNERY, B.P. (1992). Numerical recipes in c: The art of scientific computing. second edition. 59
- RABINER, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**, 257–286. 42
- RAMSEY, S.A., KNIJNENBURG, T.A., KENNEDY, K.A., ZAK, D.E., GILCHRIST, M., GOLD, E.S., JOHNSON, C.D., LAMPANO, A.E., LITVAK, V., NAVARRO, G. & ET AL. (2010). Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics*, **26**, 2071–2075. 4, 32, 34, 40, 82
- ROSENBLUM, K.R., DRESZER, T.R., LONG, J.C., MALLADI, V.S., SLOAN, C.A., RANEY, B.J., CLINE, M.S., KAROLCHIK, D., BARBER, G.P., CLAWSON, H., DIEKHANS, M., FUJITA, P.A., GOLDMAN, M., GRAVELL, R.C., HARTE, R.A.,

## REFERÊNCIAS

---

- HINRICHES, A.S., KIRKUP, V.M., KUHN, R.M., LEARNED, K., MADDREN, M., MEYER, L.R., POHL, A., RHEAD, B., WONG, M.C., ZWEIG, A.S., HAUSSLER, D. & KENT, W.J. (2011). ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Research*, **1**, 2, 54, 90
- RUSSELL, S. & NORVIG, P. (2002). *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall series in artificial intelligence, Prentice Hall, 2nd edn. 42
- SCHLIEP, A., GEORGI, B., RUNGSARITYOTIN, W. & SCHÖNHUTH, A. (2004). The general hidden markov model library: Analyzing systems with unobservable states. *Proceedings of the ISMB 2004*. 64
- SCHONES, D.E. & ZHAO, K. (2008). Genome-wide approaches to studying chromatin modifications. *Nature Reviews Genetics*, **9**, 179–191. 4
- SETUBAL, C. & MEIDANIS, J. (1997). *Introduction to Computational Molecular Biology*. PWS Publishing. 8
- SHU, W., CHEN, H., BO, X. & WANG, S. (2011). Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains. *Nucleic Acids Research*, **39**, 7428–7443. 4, 32, 40
- SONG, L. & CRAWFORD, G.E. (2010). DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. *Cold Spring Harbor Protocols*, **2010**, pdb.prot5384+. 3, 36, 39
- SONG, L., ZHANG, Z., GRASFEDER, L.L., BOYLE, A.P., GIRESI, P.G., LEE, B.K., SHEFFIELD, N.C., GRÄF, S., HUSS, M., KEEFE, D., LIU, Z., LONDON, D., McDANIELL, R.M., SHIBATA, Y., SHOWERS, K.A., SIMON, J.M., VALES, T., WANG, T., WINTER, D., ZHANG, Z., CLARKE, N.D., BIRNEY, E., IYER, V.R., CRAWFORD, G.E., LIEB, J.D. & FUREY, T.S. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Research*, **21**, 1757–1767. 39
- SPIVAKOV, M. & FISHER, A.G. (2007). Epigenetic signatures of stem-cell identity. *Nat Rev Genet*, **8**, 263–271. 4, 34
- STORMO, G.D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23. 2
- THE ENCODE PROJECT CONSORTIUM (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640. 54
- THE ENCODE PROJECT CONSORTIUM (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816. 54
- THE ENCODE PROJECT CONSORTIUM (2011). A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol*, **9**, e1001046+. 54
- WASSERMAN, W.W. & SANDELIN, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature reviews. Genetics*, **5**, 276–287. 30, 31
- WATSON, J.D., BAKER, T.A., BELL, S.P., GANN, A., LEVINE, M. & LOSICK, R. (2003). *Molecular Biology of the Gene*. Benjamin Cummings, 5th edn. 7, 8
- WHITINGTON, T., PERKINS, A.C. & BAILEY, T.L. (2009). High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Research*, **37**, 14–25. 40
- WINGENDER, E., DIETZE, P., KARAS, H. & KNÜPPEL, R. (1996). TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites. *Nucleic Acids Research*, **24**, 238–241. 56
- WON, K.J., REN, B. & WANG, W. (2010). Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biology*, **11**, R7+. 4, 35, 40, 90
- ZHANG, Y., LIU, T., MEYER, C.A., EECKHOUTE, J., JOHNSON, D.S., BERNSTEIN, B.E., NUSBAUM, C., MYERS, R.M., BROWN, M., LI, W. & LIU, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology*, **9**, R137+. 37, 56