UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

Relatório de Pesquisa - SAAP

Título: Identificação de sítios de ligação para fatores de transcrição baseada em dados relativos à acessibilidade da cromatina e modificações de histonas.
Palavras-chave: DNase-seq, ChIP-seq, modificações de histonas, footprinting.
Área: Bioinformática / Inteligência Artificial
Orientador: Ivan Gesteira Costa Filho
Aluno: Eduardo Gade Gusmão

Recife, 2 de Março de 2012

Conteúdo

1. Introdução

Em outubro de 1990, iniciou-se o chamado "Projeto Genoma Humano" com o objetivo, na época extraordinário, de sequenciar o genoma humano completo. Desta época até os dias de hoje, as tecnologias de sequenciamento avançaram de forma muito rápida. Para se ter uma idéia, em Setembro de 2001 o custo para sequenciar 1Mb de sequência de DNA era $5.292,39 (totalizando $95.263.072,00 por genoma humano); Enquanto em Julho de 2011 o curso para 1Mb era $0,12 (totalizando $10.497,00 por genoma humano) [51]. O Projeto Genoma Humano levou 13 anos para ser completado, porém hoje em dia somos capazes de sequenciar um genoma humano completo em apenas três dias.

Porém percebeu-se que a simples definição da sequência de nucleotídeos que compõem o genoma não é suficiente para explicar os diversos processos metabólicos que ocorrem nos organismos dos seres vivos. Tais processos decorrem de uma complexa cadeia de eventos regulatórios que podem ocorrer tanto no nível genômico (por exemplo, SNPs), transcricional (elementos regulatórios), pós-transcricional (splicing) ou na tradução e pós-tradução.

No âmbito da regulação gênica, a identificação dos elementos no DNA que ativam, reprimem, bloqueiam, silenciam ou aumentam é necessária para que seja completamente compreendida a expressão gênica e as redes de interações metabólicas dentro de uma célula. Além disso, uma série de processos regulatórios (consistindo a chamada epigenética) está sendo associada a processos importantes e cruciais para o entendimento de um organismo, como o da diferenciação celular.

Existem diversos métodos para a identificação desses elementos regulatórios no DNA. Porém nenhum deles é auto-suficiente, no sentido de que sempre existe um *trade-off* entre a facilidade de realização, tempo de execução, porção do genoma a ser analisado em uma rodada do experimento, acurácia e resolução. Alguns desses métodos são exclusivamente biológicos como o *footprinting* tradicional, isto é, ele é realizado exclusivamente em laboratório. Porém outros exigem uma grande demanda computacional como a análise de dados de métodos que terminam em chip (ChIP-chip ou DNase-chip) ou em sequenciamento em massa (ChIP-seq ou DNase-seq).

Este trabalho propõe a integração de dados relativos à cromatina aberta (obtidos através de DNase-seq) e de modificação de histonas (obtidos através de ChIP-seq) para realizar um processo chamado *footprinting*, isto é, a procura (no genoma) de prováveis sítios de ligação de elementos regulatórios tais como *activators*, *repressors*, *silencers*, *enhancers* e *insulators*. Tal integração será realizada através de um modelo probabilístico chamado Cadeias de Markov Escondidas (HMM, do inglês *Hidden Markov Models*). Tal modelo permite realizar uma mistura de distribuições de probabilidade, que é a forma de representação de sinais contínuos genômicos que geram as maiores acurácias quando se trata de predição.

2. Revisão da literatura

Nesta seção, será realizada uma revisão geral da literatura disponível. Primeiramente, serão detalhados os conceitos biológicos e computacionais básicos utilizados na definição de conceitos mais avançados neste texto. Em seguida, será realizada uma descrição dos principais métodos disponíveis na literatura que possuam alguma interseção com a proposta deste trabalho. Finalmente, os objetivos deste projeto serão traçados com clareza, apresentando a justificativa necessária e estabelecendo as hipóteses propostas.

As siglas definidas nesta seção serão utilizadas durante todo o restante deste texto. Grande parte das palavras reservadas foram deixadas em inglês, por ser o único idioma onde existe um consenso sobre a nomenclatura utilizada. Tais palavras em inglês foram representadas em itálico. Para que este texto não se torne demasiadamente longo, é assumido que o leitor esteja familiarizado com definições e nomenclaturas básicas tanto biológicas quanto computacionais e probabilísticas.

2.1. Regulação genética e epigenética em eucariotos

2.1.1. Introdução

A regulação da expressão gênica eucariótica pode ocorrer em vários níveis, tanto na transcrição (nas fases de iniciação ou elongamento) quanto após a transcrição (processamento, transporte, estabilidade e tradução do mRNA). Porém acredita-se que a maioria dos processos regulatórios ocorra na fase de iniciação da transcrição. Em eucariotos, os genes transcritos pela RNA polimerase II (genes codificantes de proteínas) tipicamente possuem dois tipos de elementos regulatórios *cis-acting*: um promotor, que é constituído pelo core promoter ("núcleo" do promotor onde a RNA polimerase II (Pol2) se liga para iniciar a transcrição) e por elementos regulatórios proximais; E elementos regulatórios distais, que podem ser *enhancers*, *silencers*, *insulators* ou *locus control regions* (LCR). Esses elementos regulatórios *cis-acting* correspondem a sítios de ligação que contêm sequências específicas reconhecidas por proteínas chamadas de fatores de transcrição *trans-acting* [39]. A Figura 2.1 mostra um esquema de uma região regulatória típica.

Cada gene pode ser regulado por uma grande quantidade de fatores de transcrição e cada fator pode regular diversos genes, simultaneamente ou não [38]. Isto caracteriza um fenômeno chamado regulação combinatória, isto é, a possibilidade de um gene ser produzido temporalmente e espacialmente de formas muito distintas dependendo das necessidades da célula através de diferentes combinações dos fatores que se ligam em suas regiões regulatórias proximais ou distais. Torna-se clara, então, a necessidade da caracterização de todos esses

elementos regulatórios, dado que isso é fundamental para o entendimento dos diversos processos mencionados no início deste tópico.

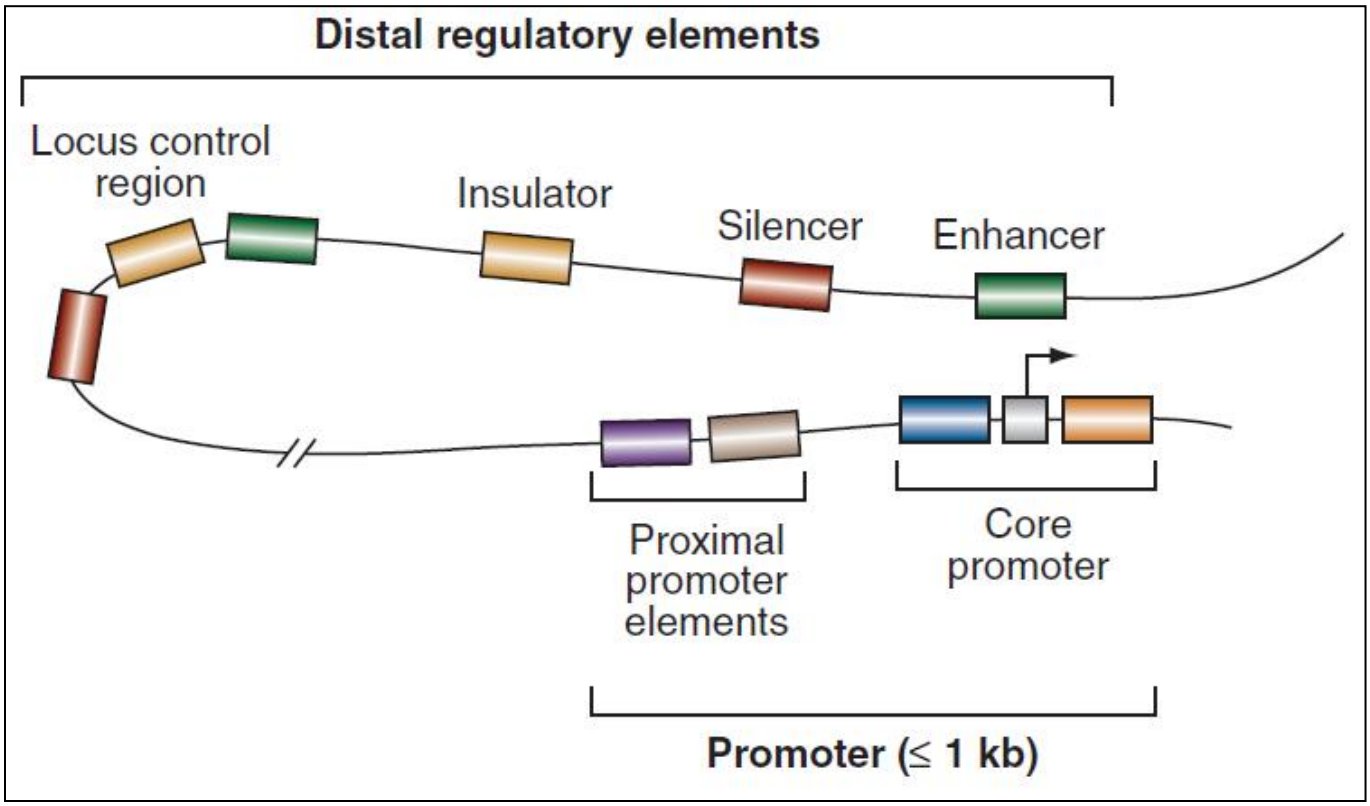

Figura 2.1. Esquema gráfico dos elementos regulatórios discutidos neste trabalho. Fonte: [39].

2.1.2. Maquinaria regulatória eucariótica proximal

Os fatores na região promotora envolvidos na transcrição acurada de genes codificantes de proteínas pela Pol2 podem ser divididos em três categorias: fatores de transcrição gerais (GTF), ativadores e co-ativadores. A Figura 2.2 exibe um esquema desse complexo, mostrando as possíveis interações entre alguns componentes. Nesta figura, foram colocadas interrogações nas interações que ainda estão sendo validadas experimentalmente.

Os fatores de transcrição gerais são necessários e suficientes para a transcrição *in vitro* (nível mais baixo de transcrição, intitulado nível basal). Essa classe inclui a própria Pol2 e os complexos auxiliares TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, além de uma larga estrutura chamada de mediador. Essas estruturas se montam de forma organizada para a formação do complexo pré-iniciação, momento onde a Pol2 é recrutada para o DNA.

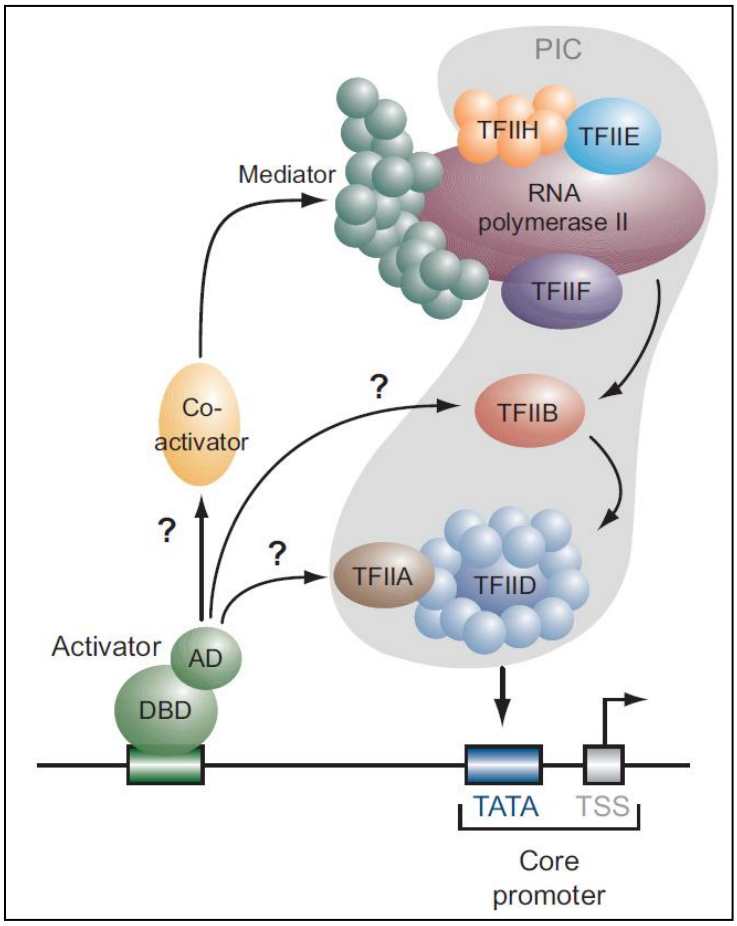

Figura 2.2. Esquema gráfico dos fatores de transcrição gerais que se ligam no core promoter para formar o complexo pré-iniciação e dos elementos regulatórios proximais (ativadores e co-ativadores). As setas representam interação proteína-proteína ou proteína DNA. As setas marcadas com uma interrogação representam interações possíveis porém ainda não definidas. Fonte: [39].

Os níveis de transcrição são fortemente estimulados pela presença do segundo tipo de fatores envolvidos neste processo: os ativadores. Tais elementos possuem pelo menos dois domínios. Um deles contém uma estrutura que reconhece trechos no DNA, podendo assim se ligar em locais específicos do DNA (TFBS, do inglês *Transcription Factor Binding Site*), com sequências específicas. O tamanho destas sequências reconhecidas, ou *motifs*, variam de 6 a 12bp porém as posições que realmente são fundamentais para a interação proteína-DNA variam de 4 a 6bp. Uma forma de representação muito comum desses motivos são as *Position Weight Matrix* (PWM), onde é definido um grau de afinidade de um certo fator de transcrição para cada base (A, C, G ou T) presente em cada posição do *motif*. Além do domínio de reconhecimento do DNA, os ativadores contêm um domínio de ativação, que é necessário para que esse fator ative ou estimule a transcrição. O ativadores podem funcionar de diversas maneiras: ajudando na formação do complexo pré-iniciação, atuando na própria iniciação, elongação ou re-iniciação, ou através do recrutamento de outros fatores que modificam a estrutura da cromatina (isso será mais detalhado no tópico 2.1.4 onde serão mostrados conceitos básicos de epigenética).

Finalmente, os co-ativadores não possuem domínios de reconhecimento de DNA. Eles são inseridos no sistema através de conexões proteína-proteína entre eles e os ativadores. Sua forma de atuação é bastante semelhante aos ativadores, através do estímulo da formação do complexo pré-iniciação ou através da modificação de cromatina.

2.1.3. Elementos regulatórios transcricionais

Neste tópico serão definidas brevemente as estruturas regulatórias usuais, descritas na literatura até o presente momento. A Figura 2.3, no final deste tópico, sumariza graficamente tais estruturas. Todas as descrições são baseadas em [39].

2.1.2.1. Core Promoter: É a região no início do gene que possui elementos onde a maquinaria geral de transcrição se liga. Alguns desses elementos foram bastante estudados tais como o elemento iniciador (Inr), o *TATA box* (sítio de ligação para a subunidade TBP da TFIID), o *Downstream Core Element* (DCE) e o TFIIB-Recognition Element (BRE). 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.

2.1.2.2. Proximal Promoter Elements: É a região imediatamente *upstream* (até no máximo algumas centenas de pares de bases) do *core promoter*, contendo vários sítios de ligação para ativadores (conforme mencionado anteriormente).

2.1.2.3. Enhancers: Regulam a expressão temporalmente e espacialmente e sua atividade independe da distância do promotor (que pode chegar à ordem de Mb) e da sua orientação em relação a este. Essa região é tipicamente composta por vários sítios de ligação bastante próximos uns dos outros, onde os *enhancers* se ligam para aumentar a expressão do gene. A distinção entre os *enhancers* e os elementos proximais ainda é bastante nebulosa. De fato, grande parte dos fatores que se ligam em regiões proximais também se liga em *enhancers*. Existem fortes evidências de que esses elementos distais (como os *enhancers*) consigam atuar a partir de regiões tão distantes através do modelo de DNA *looping*. Neste modelo, o DNA se conforma de tal maneira que, apesar de estar vários bps longe do *core promoter*, 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 complexo pré-iniciação se forme em regiões de *enhancers* e que esse complexo se agregue ao restante dos fatores gerais através do processo de DNA *looping*.

2.1.2.4. Silencers: Elementos que reprimem a expressão de um gene. Assim como os *enhancers*, a atuação da maioria dos *silencers* não depende da distância para a região promotora nem da orientação, porém alguns *silencers* dependentes da posição foram encontrados. Os *silencers* podem estar em regiões proximais, em regiões distais de *enhancers* ou em regiões distais

independentes. O fator de transcrição que se liga em um *silencer* é chamado de repressor, nos quais os co-repressores podem se ligar (de forma semelhante aos ativadores e co-ativadores). Os *silencers* podem reprimir a expressão não permitindo a ligação de um ativador, inibindo a formação do complexo pré-iniciação ou recrutando modificadores de cromatina para "fechar" a região de forma que nenhum fator consiga se ligar naquele local.

2.1.2.5. Insulators: Bloqueiam a atuação de outros elementos regulatórios definindo uma espécie de partição do genoma em blocos com sistema interno de regulação. Os *insulators* têm duas funções: bloquear a influência de um *enhancer* sobre a expressão de um determinado gene ou bloquear a disseminação do *silenciamento* de uma região por estruturas que "fecham" (condensam) a cromatina (geralmente agindo numa reação em cadeia, parando apenas ao encontrar o *insulator*). Os *insulators* geralmente são dependentes de posição porém independentes de orientação. Apesar de vários fatores *trans-acting* que mediam a função do *insulator* serem conhecidos para a *drosophila*, em vertebrados se conhece apenas o CTCF (*CCCTC-binding factor*).

2.1.2.6. Locus Control Regions: São um grupo de elementos *cis-acting* (como *enhancers*, *silencers* ou *insulators*) reunidos em uma região com o objetivo de controlar a expressão de um determinado grupo de genes. Apesar da grande maioria dos LCR serem encontrados upstream, eles são independentes de posição, porém são dependentes de *copy-number*.

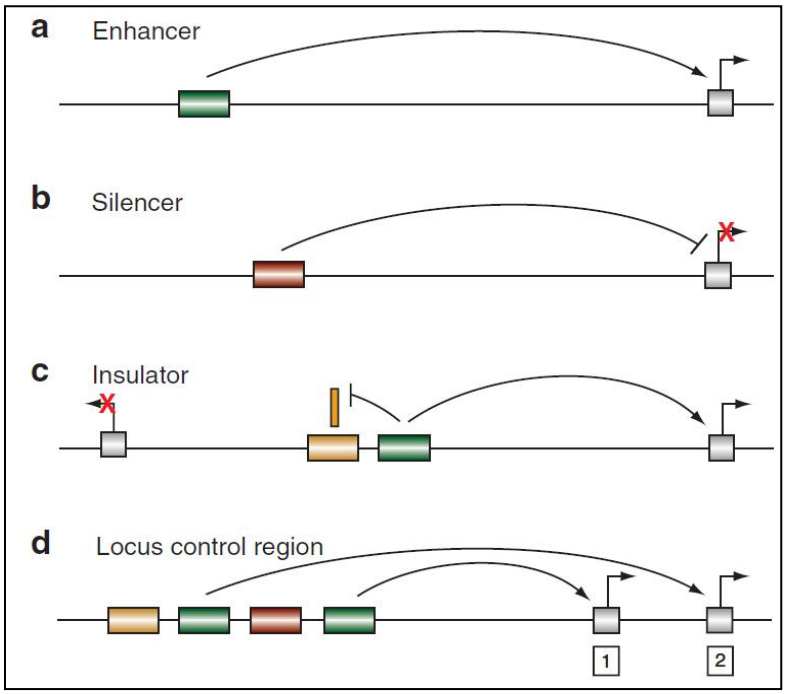

Figura 2.3. Esquema de atuação dos elementos regulatórios *cis-acting*. As linhas com uma terminação em seta representam estimulação e as linhas com uma terminação de linha perpendicular representam repressão. Fonte: [39].

2.1.4. Epigenética

Além dos elementos regulatórios no DNA, com sequências de nucleotídeos específicas onde os fatores de transcrição se ligam para aumentar ou diminuir o nível de expressão, existem diversos fatores relativos à organização da cromatina que modulam a transcrição de um gene. Esses fatores formam o que se chama de epigenética, do grego *epi* significa acima, isto é, é o conjunto de fatores que formam uma espécie de código em um nível mais alto do que o código genético (sequência de nucleotídeos) [52].

Cromatina é o nome que se dá ao complexo formado por DNA + proteínas. Nos organismos com núcleo real (eucariotos), o DNA se encontra enovelado em proteínas chamadas histonas. Essas proteínas formam um complexo de oito proteínas, formado por quatro pares das histonas H2A, H2B, H3 e H4. Cada uma dessas histonas possui uma cauda, que se propaga para fora do complexo histona + DNA (chamado de *nucleossomo*). As Figuras 2.4 e 2.5 mostram, respectivamente, um modelo gráfico para o enovelamento da cromatina e os diferentes níveis de enovelamento.

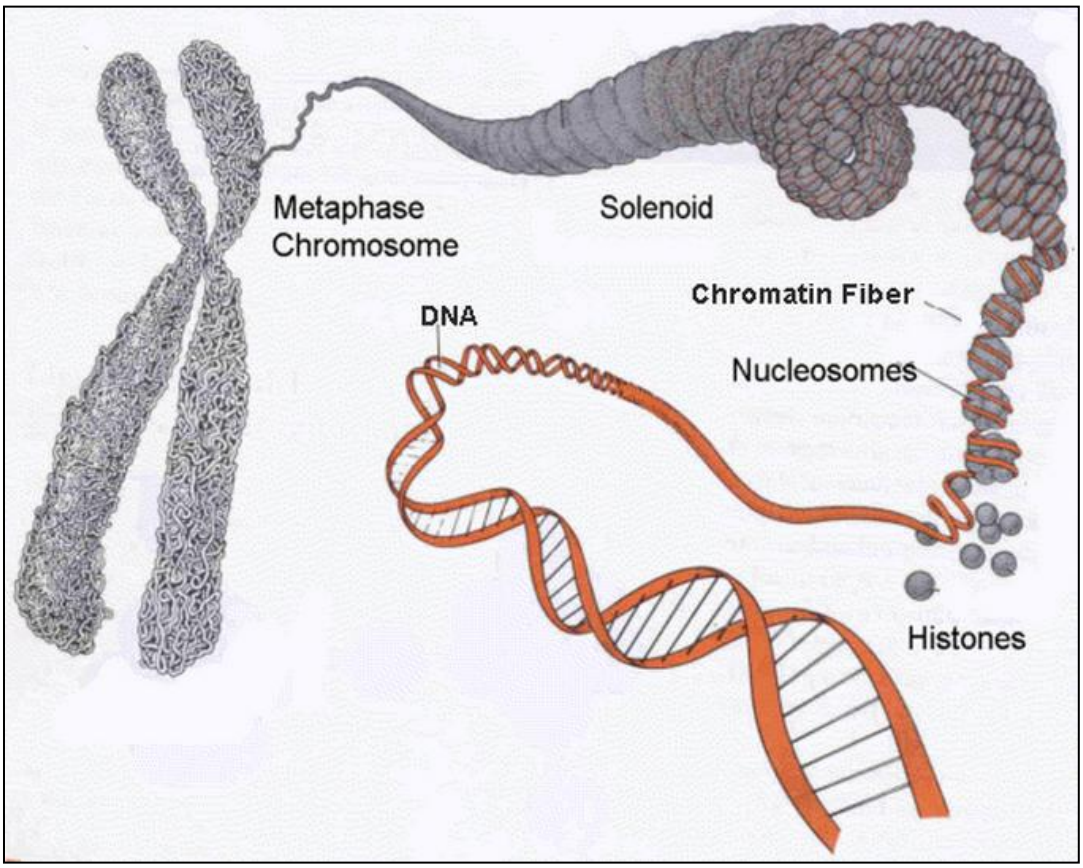

Figura 2.4. Esquema do enovelamento da cromatina.

Acreditava-se que o complexo de proteínas que forma o nucleossomo era apenas uma estrutura passiva cujo único objetivo era atuar na compactação do DNA para que o mesmo coubesse dentro do núcleo celular (o DNA é uma molécula grande que possui vários níveis de enovelamento (Figura 2.5). Completamente desenovelado chega a possuir 2 metros de comprimento). Porém observou-se que a estrutura da cromatina pode ser alterada, e essa alteração é regida por vários fatores (aos quais dá-se o nome de fatores epigenéticos).

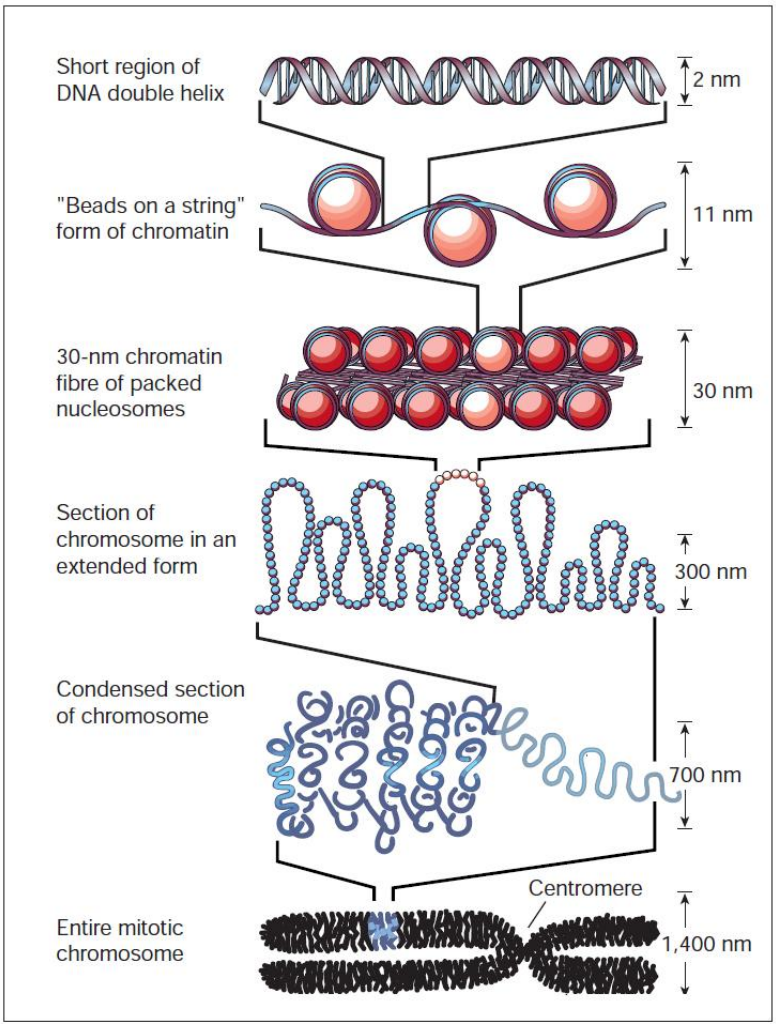

Figura 2.5. Os diferentes níveis de enovelamento da cromatina. Fonte: [10].

Tais fatores epigenéticos podem ser [52]:

- Modificações pós-tradução de aminoácidos na cauda das histonas.
- Remodelamento da cromatina através de processos dependentes de energia (ATP) que modificam o posicionamento dos nucleossomos.
- A inserção ou remoção de histonas variantes.
- Atuação de pequenos RNAs não codificantes.
- Metilação do próprio DNA geralmente em dinucleotídeos CpG.

Um dos fatores mais estudados é a modificação pós-tradução das histonas. A cauda das histonas podem sofrer modificações químicas em aminoácidos específicos. Entre essas modificações estão a fosforilação, acetilação, metilação e ubiquitilação. Essas modificações possuem uma nomenclatura específica, seguindo a ordem: tipo da histona, aminoácido que sofre a modificação e tipo de modificação [10]. Por exemplo, H3K4me2 se refere à bimetilação (me2) da lisina na posição 4 (K4) na cauda da histona H3. A Figura 2.6 mostra um mapa das possíveis modificações de histonas.

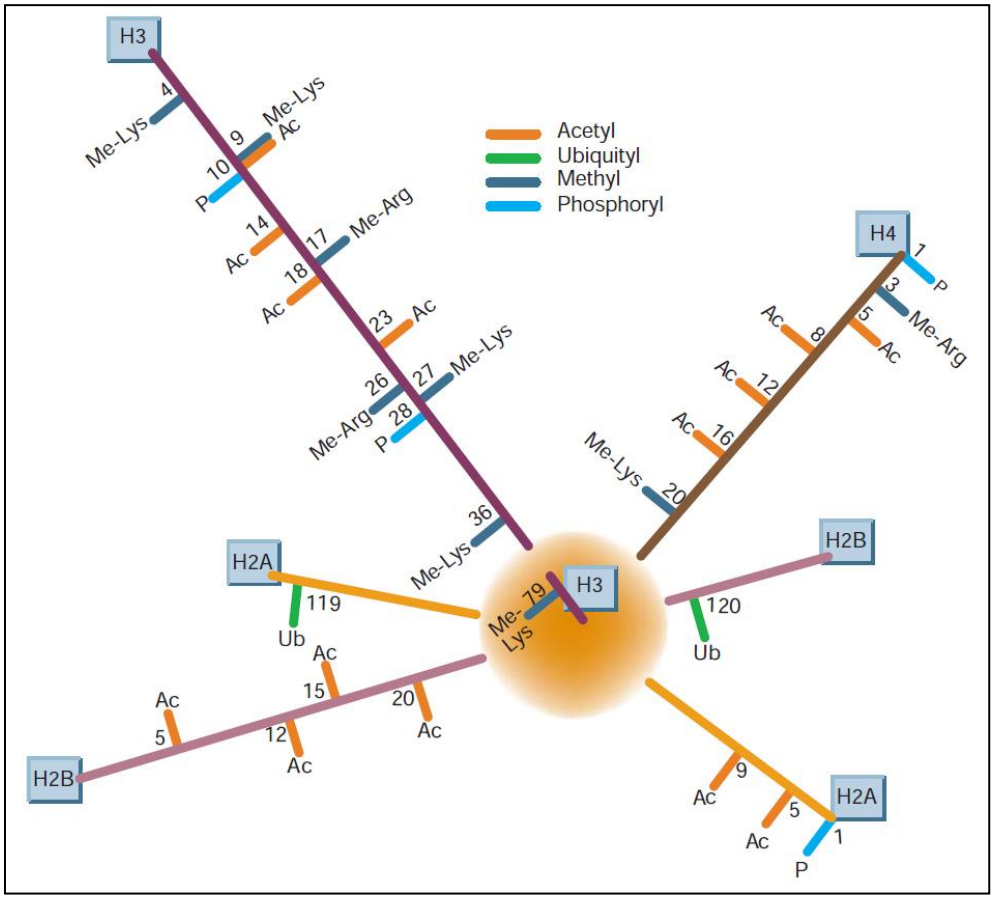

Figura 2.6. Mapa das possíveis modificações pós-tradução que as caudas das histonas podem sofrer. Fonte: [10].

Essas modificações de histonas estão sendo associadas a processos de mudança na estrutura da cromatina. Sabe-se, por exemplo, que o domínio H3K79 possui características de anti-silenciamento. Já H3K9, H3K27 e H4K20 podem sofrer modificações que reprimam a expressão de um gene. Ainda não são conhecidas todas as funções de todas as possíveis modificações das histonas. Estudos como este que está sendo documentado neste texto também têm como objetivo a caracterização dos padrões de modificação de histonas em regiões de silenciamento ou ativação, dadas várias linhas celulares com características e objetivos diferentes dentro do organismo [52].

2.2. Identificação de sítios de ligação para fatores de transcrição

Tornou-se claro a partir da revisão realizada na seção 2.1 que a identificação dos elementos regulatórios que governam a expressão espacial e temporal dos genes é de fundamental importância para o entendimento de como os genes interagem de forma a orquestrar processos metabólicos complexos como a diferenciação celular, apoptose ou desenvolvimento.

Nesta seção serão revisados alguns métodos biológicos para identificação destes elementos funcionais no DNA. Cada um dos métodos a seguir possui vantagens e desvantagens. Nos métodos mais recentes, baseados em microarranjos (com terminação -chip) ou em sequenciamento massivo de pequenos fragmentos (com terminação -seq), a aplicação de métodos robustos de biologia computacional se torna fundamental. Nestes métodos, a quantidade de dados gerada é bastante volumosa e geralmente estes dados possuem ruídos e vieses que precisam ser tratados com métodos estatísticos igualmente robustos.

2.2.1. DNA Footprinting

Este método tradicional e puramente biológico, consiste em observar padrões de digestão no DNA de algum agente de clivagem. Estes agentes podem ser, por exemplo, radicais hidroxila ou radiação ultravioleta. Porém neste texto será dado foco à endonuclease Desoxirribonuclease I (DNase I). Esta enzima é capaz de se ligar no *minor groove* 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 [27].

Segundo [26], esse método se inicia com a obtenção do material genético. De posse do DNA de várias células do tipo específico que se deseja estudar, 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. 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 seu *low-throughput*, isto é, uma rodada deste método demora um tempo razoavelmente alto e é capaz de analisar somente um trecho bastante pequeno, tornando impossível a aplicação deste método em estudos *genome-wide*.

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. O agente de clivagem é então adicionado a ambas as porções, permitindo que ele corte o DNA em várias posições aleatórias. 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). A distribuição dos fragmentos estará parecendo 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 2.7 detalha este processo de forma visualmente agradável.

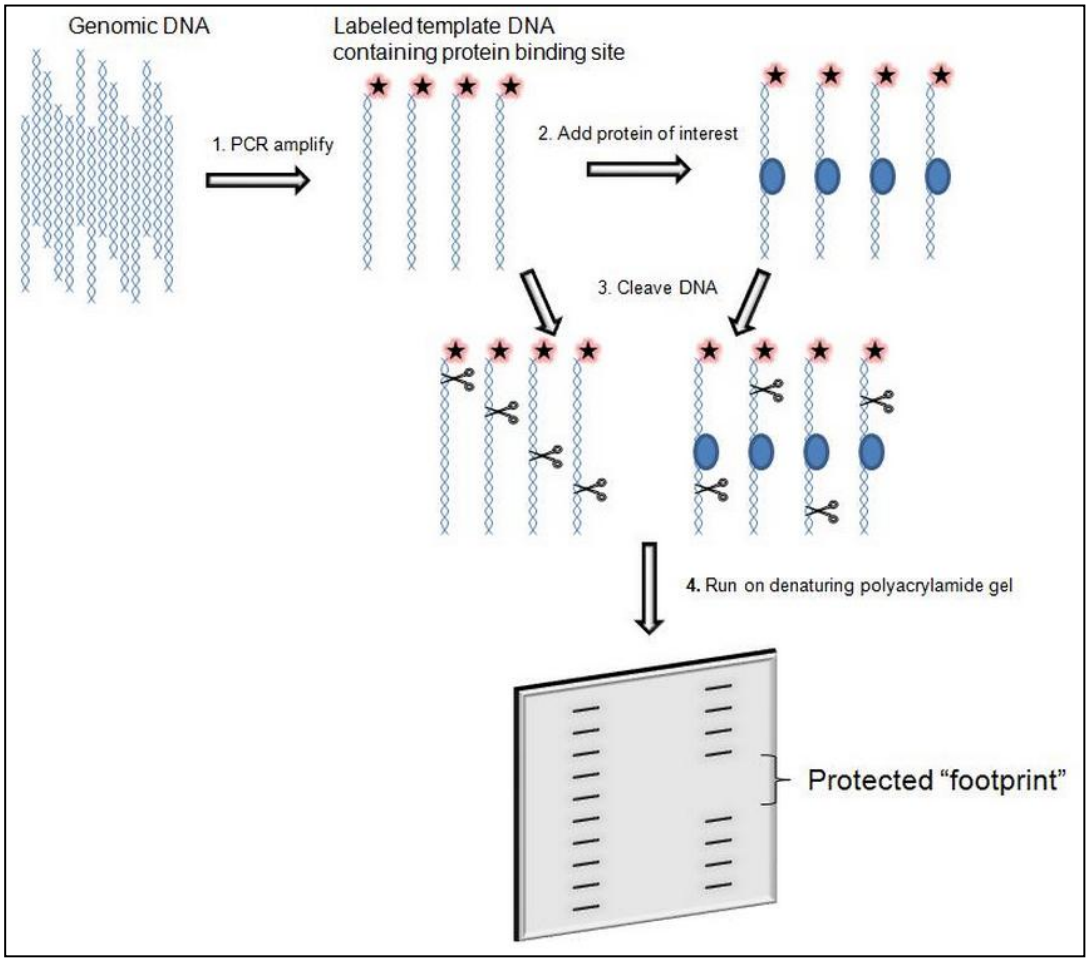

Figura 2.7. DNA Footprinting.

Obviamente, o processo é muito mais complexo do que o descrito neste texto. 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 é definitivamente *low-throughput*. Este método foi exibido aqui apenas como um exemplo de um método mais tradicional para realizar *footprinting* e para introduzir este conceito, porém tal método não necessita de tecnologias computacionais tais como algoritmos de aprendizado de máquina ou mineração de dados. As técnicas descritas a seguir são mais recentes e foram propostas justamente para permitir estudos *genome-wide*, cujo objetivo é pesquisar as diferenças entre a configuração e distribuição dos elementos regulatórios em diversas linhas celulares ou em diferentes momentos do desenvolvimento ou diferenciação de uma linha celular.

2.2.2. ChIP-chip

ChIP-chip (ou ChIP-on-chip) é uma técnica que combina imunoprecipitação da cromatina (ChIP) com tecnologias de microarranjo (chip) para investigar as interações entre proteínas e DNA *in vivo*. Como ela utiliza, em um de seus estágios, a imunoprecipitação, é permitida a identificação não só de elementos funcionais no genoma (*activators*, *enhancers*, *silencers* ou *insulators*) como também de modificação de histonas [1].

Essa tecnologia foi aplicada com sucesso pela primeira vez em três estudos publicados em 2000 e 2001. Os autores identificaram sítios de ligação para fatores de transcrição específicos na levedura (S. cerevisiae). Em 2002, o grupo de Richard Young determinou a posição, a nível *genome-wide*, de 106 fatores de transcrição usando um *c-Myc tagging system* na levedura. Outras aplicações para esta técnica incluíram replicação do DNA, recombinação e estrutura da cromatina. Desde então ChIP-chip se tornou uma poderosa ferramenta no mapeamento de fatores de transcrição e modificações de histonas ao longo do genoma. A aplicação desta tecnologia em células de mamíferos encontrou maior dificuldade devido à quantidade de regiões repetitivas no DNA. Por esse motivo, vários estudos foram restritos a regiões promotoras conhecidas. Porém, mais atualmente, empresas como a *NibleGen* já comercializam chips *genome-wide* para mamíferos.

Antes de explicar como funciona o experimento de ChIP-chip, será descrito brevemente o procedimento de imunoprecipitação de cromatina (ChIP). De forma resumida o método funciona da seguinte forma: primeiramente a célula é quebrada para que possamos acessar o complexo DNA-proteína (cromatina). Esse complexo é partido através de algum método (como sonicação) e os fragmentos contendo a proteína de interesse são extraídos através de imunoprecipitação. A partir disso, o DNA é purificado e os fragmentos resultantes podem ser determinados através de algum método (PCR, *tiling array*, sequenciamento). No caso de ChIP-chip são utilizados *tiling arrays*, que são microarranjos especiais onde cada spot corresponde a um trecho genômico e

spots consecutivos correspondem a trechos genômicos consecutivos (obviamente, com algum grau de espaçamento e tamanho de janela). A seguir será revisado o método de *ChIP-seq* que utiliza o sequenciamento *high-throughput* para determinar os fragmentos capturados na ChIP.

Começando com uma questão biológica, um experimento de ChIP-chip pode ser dividido em três partes: a primeira é o design experimental, a segunda é o experimento no *wet-lab* e finalmente a análise dos dados no *dry-lab*, respondendo às questões levantadas ou resultando em novas questões que reinicializarão o ciclo. A Figura 2.8 sumariza a descrição deste método, realizada a seguir.

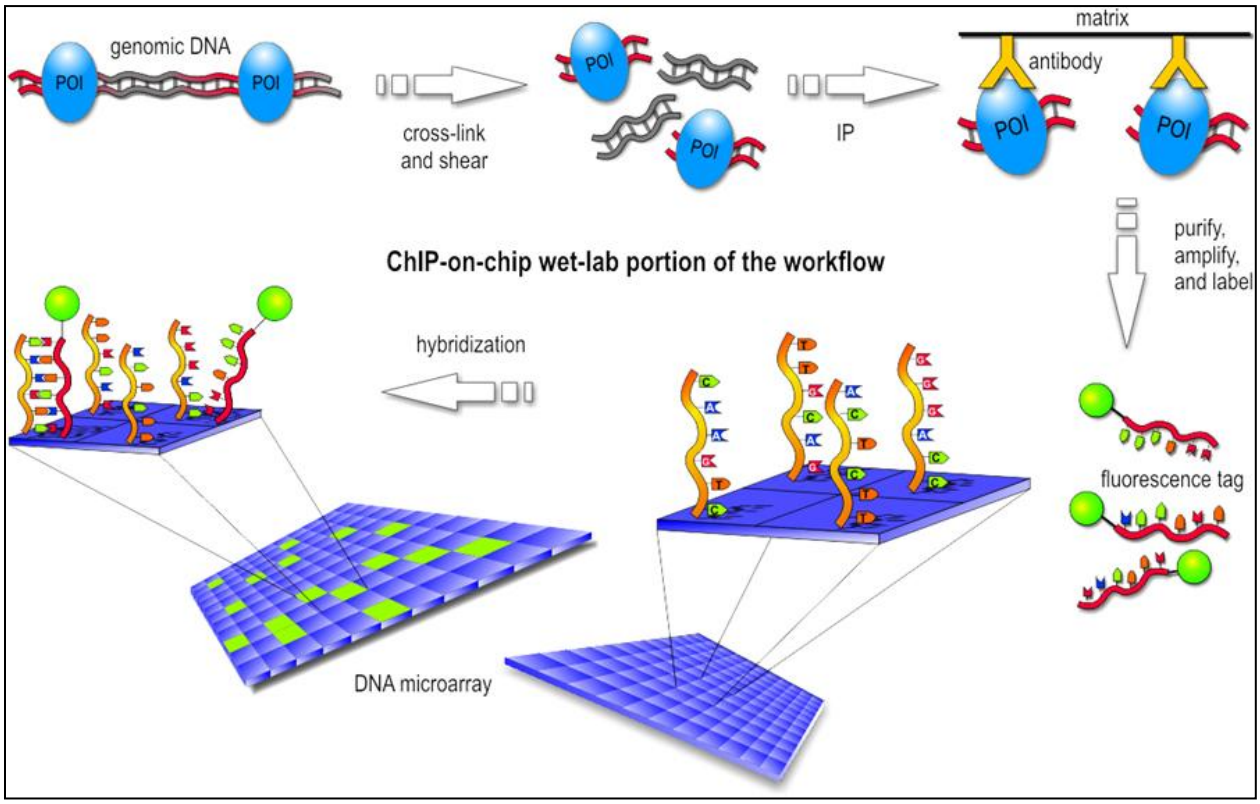

Figura 2.8. ChIP-chip.

A primeira etapa do *wet-lab* corresponde à imunoprecipitação da cromatina como foi descrito no início deste texto. Após amplificação e desnaturação do DNA, os fragmentos em fita simples são rotulados com um composto fluorescente tal como Cy5 ou Alexa 647. Por fim, os fragmentos rotulados são aplicados no microarranjo que contém as sequências curtas e que cobrem a região de interesse. Os fragmentos irão se hibridizar com o DNA contido nas sondas formando novamente fitas duplas.

Após um tempo suficientemente grande para que a hibridização ocorra, o array é iluminado com uma luz de fluorescência. As sondas do chip que foram hibridizadas com os fragmentos rotulados emitem sinais de luz que são capturados pela câmera. Essa imagem contém os dados brutos que serão utilizados no restante do workflow. A imagem dos dados brutos em formato *false-color image* deve ser convertida em valores numéricos para que a análise possa ser realizada. Esta é a parte mais desafiadora de todo o processo. Problemas podem surgir nas próximas partes do processo tais como: encontrar os métodos apropriados para remover ruídos, normalização dos dados, etc. No final da preparação dos dados, as análises estatísticas podem ser realizadas com intuito de explicar as questões levantadas no início do projeto. Por causa da diferença nos tipos de plataforma e na não padronização entre elas, as análises podem variar bastante dependendo das escolhas iniciais. Existem três passos básicos para a análise dos dados:

- 1. Durante a primeira etapa, os sinais de fluorescência capturados são normalizados usando sinais de controle derivados do mesmo chip ou de um segundo chip. Tal controle tem o objetivo de informar quais sondas foram hibridizadas corretamente e quais não foram.
- 2. No segundo passo, análises numéricas e testes estatísticos são aplicados aos dados (da amostra e de controle) para a identificação das regiões enriquecidas para a proteína de interesse.
- 3. No terceiro passo, as regiões enriquecidas são analisadas mais profundamente. Tais análises dependem do tipo de experimento. Se a proteína de interesse for um fator de transcrição, por exemplo, pode-se desejar inferir os motifs ou outros padrões funcionais que permitam a anotação funcional do genoma.

Métodos baseados em microarranjos são completamente *high-throughput* e, apesar de alguns anos atrás o custo desse experimento ter sido bastante elevado, atualmente é possível realizar os mesmos experimentos com custos bem mais baixos. Porém os dados gerados por este método são bastante ruidosos. Análises estatísticas estado-da-arte são fundamentais para que um estudo seja aceito pela comunidade. Além disso, este método possui resolução muito baixa. Isto quer dizer que não somos capazes de dizer com exatidão onde a proteína está ligada, e sim somente a região provável onde ela se encontra. Esta técnica é útil em estudos onde se pretende demonstrar a presença de fatores ou modificações de histonas específicas em algumas regiões, porém essa técnica não deve ser utilizada em casos onde se queira criar um mapa contendo as localizações exatas de elementos regulatórios, em especial em regiões de alta densidade de tais elementos (tais como *enhancers*).

2.2.3. ChIP-seq

Este método é bastante semelhante ao método de ChIP-chip descrito no tópico anterior. O procedimento de imunoprecipitação é realizado da mesma forma porém os fragmentos capturados são sequenciados através de algum método de sequenciamento de alto desempenho. Tais métodos podem ser o sequenciamento com terminadores reversíveis de molécula única ou múltipla, pirosequenciamento, sequenciamento por ligação [2,4]. A Figura 2.9 exibe a sequência de eventos básica de um experimento de ChIP-seq.

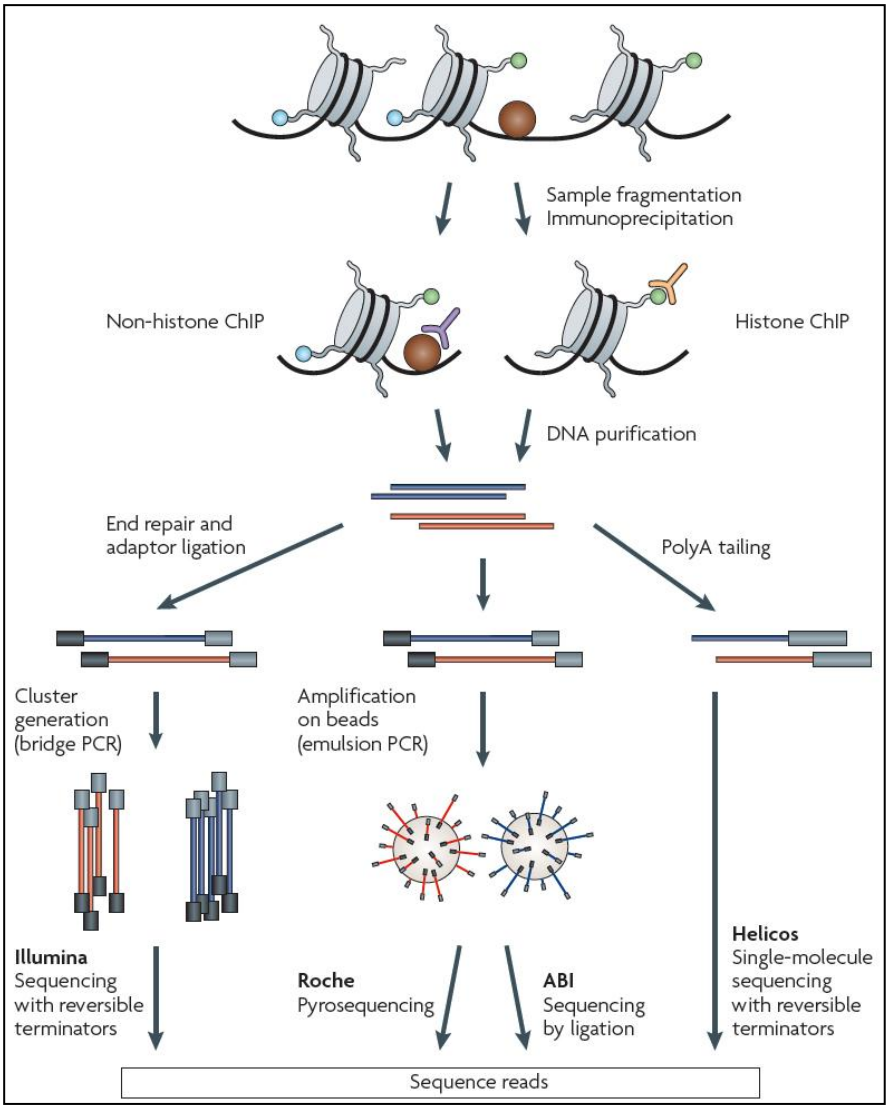

Figura 2.9. Esquema para a geração de *reads* com ChIP-seq. Fonte: [2]

De posse dos trechos genômicos sequenciados a partir dos fragmentos de interesse capturados (a partir deste momento, chamados de *reads*), os mesmos são alinhados no genoma através de

algum método de alinhamento. Nesta etapa alguns cuidados devem ser tomados. Devido à presença de diversas regiões repetitivas no genoma, estes alinhamentos devem ser cuidadosamente estudados para que apenas os trechos relevantes sejam levados em consideração. Além disso, testes estatísticos devem ser realizados para que se removam regiões possivelmente ruidosas devido à não exatidão dos métodos de manipulação biológica (ChIP, amplificação, reparo das extremidades e ligação dos adaptadores, etc.). Este sinal, formado pelos *reads* alinhados no genoma, formam uma espécie de série temporal, conforme exibido na Figura 2.10.

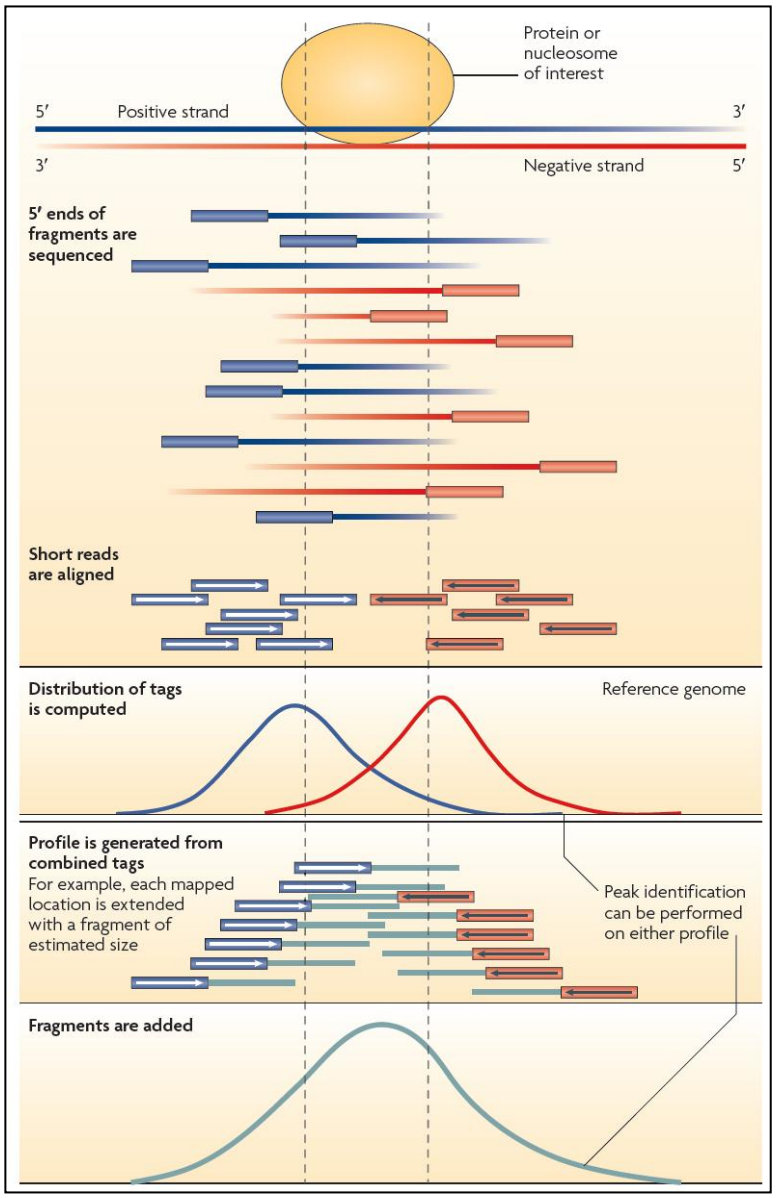

Figura 2.10. Exemplo do sinal genômico produzido pelo ChIP-seq. Fonte: [2]

A análise de regiões enriquecidas de sinal, isto é, regiões de picos, pode ser realizada de várias formas. A forma mais simples consiste na adequação do sinal a uma distribuição estatística (usualmente a distribuição gama) e estabelecimento do ponto de corte a partir de um p-value (1% ou 5%, dependendo do nível de confiança desejado). Atualmente existem vários softwares que realizam o processamento desses sinais de forma a encontrar essas regiões enriquecidas. Dois softwares bastante usados são o MACS (*Model-based Analysis of ChIP-seq*) [5] e o F-seq [21]. O primeiro utiliza uma distribuição de Poisson de forma dinâmica (variando o parâmetro λ baseado em uma janela definida pelo usuário) para capturar vieses locais no genoma. O segundo realiza uma estimação da densidade suavizada do sinal através do método da janela de Parzen, que utiliza como parâmetros o tamanho estimado dos *reads* (ou um valor estipulado pelo usuário) e um parâmetro de suavização. Além disso, o F-seq corrige vieses gerados por *copy-number variations* permitindo a utilização de conjuntos de controle de sinais de *background* tanto específico para o método em questão quanto para a linha celular em questão.

Este método vem sendo bastante utilizado por permitir uma predição de elementos regulatórios ou epigenéticos acurada, específica por fator desejado, com boa resolução e de *high-throughput*. Porém este método requer um grande número de células inicialmente e um anticorpo (para a ChIP) de alta qualidade para se obter um sinal aceitável. Além disso, apesar da resolução ser maior do que nos métodos baseados em chip, este método ainda é incapaz de resolver as interações DNA-proteína com a resolução em pares de bases.

2.2.4. DNase-seq

O método DNase-seq tem a mesma finalidade e idéia do método de *footprinting* apresentado na seção 2.2.1. Apesar de algumas diferenças específicas relativas ao protocolo biológico a ser seguido, a primeira fase segue exatamente a mesma idéia. Estamos interessados na versão *in vivo* desse protocolo, que será descrita superficialmente a seguir [20, 23, 25, 33].

A partir de uma linha celular específica, adicionam-se os componentes necessários para que as proteínas que estavam ligadas no DNA, permaneçam ligadas nas fases subsequentes (geralmente formaldeído). Dessa forma, estamos conservando o estado natural da célula naquele momento, por isso o método é considerado *in vivo*. Após a extração do material genético, a enzima DNase I é colocada na solução para digerir o DNA em vários locais aleatórios. As regiões dos fragmentos pré-digestão que continham proteínas ligadas não são digeridas, pela incapacidade da enzima DNase I de cortar o DNA em locais onde já exista uma proteína (como foi explicado na seção 2.2.1).

Após a digestão, os fragmentos de DNA que não possuem proteínas ligadas são extraídos (perceba a diferença para o método de ChIP-seq, onde os fragmentos *com* a proteína de interesse que são extraídos). Esses fragmentos são sequenciados através de qualquer método de sequenciamento *high-throughput* (como os descritos na seção 2.2.3). Gerando um sinal

genômico com o mesmo formato do exibido na Figura 2.10, porém com uma forma de interpretação completamente diferente.

A Figura 2.11 exibe um exemplo deste sinal para a região promotora do gene FMR1 (Fragile-*X Mental Retardation 1*). O primeiro sinal exibido (*DNase-seq smoothed*) é o sinal de DNase-seq suavizado através do método F-seq (que utiliza o método da janela de Parzen para realizar tal suavização, como descrito na seção anterior). O sinal abaixo deste (*DNase-seq raw*) é a simples contagem da quantidade de *reads* alinhados em cada bp. Este sinal é ampliado e comparado com *footprints* gerados através do método descrito na seção 2.2.1 por [22] (Druin et al *footprints*) e com os *footprints* preditos utilizando os dados de DNase-seq (DNase-seq *footprints*). Observe que os *footprints* se caracterizam por regiões de depleção de sinal de digestão de DNase I.

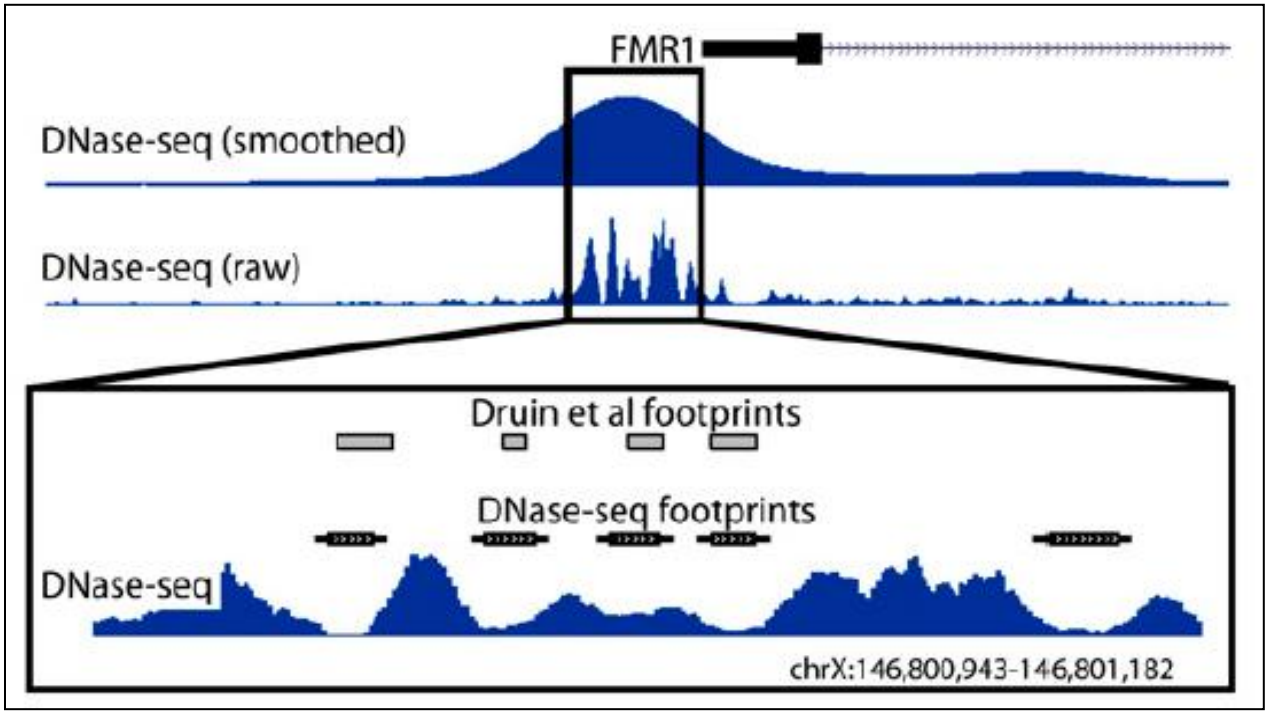

Figura 2.11. Exemplo do sinal genômico produzido pelo DNase-seq. A Figura exibe sinais gerados a partir do método DNase-seq no cromossomo X das posições 146.800.943 até 146.801.182. Neste trecho genômico se encontra a região promotora para o gene FMR1. O sinal suavizado (*smoothed*) foi gerado a partir do software F-seq, enquanto o sinal bruto (*raw*) foi gerado pela simples contagem de quantos *reads* se sobrepunham em cada bp do genoma. Os *footprints* de Druin et al foram gerados através do método de *footprint* tradicional, enquanto os *footprints* a partir do sinal de DNase-seq foram gerados utilizando uma cadeia de Markov com cinco estados (descrita mais adiante). Fonte: [22]

Uma análise de enriquecimento (isto é, encontrar as regiões com picos significantes) no sinal gerado pelo método DNase-seq nos fornece as regiões de hipersensibilidade da DNase I. Tais regiões foram detalhadamente descritas em [26,27]. Essas regiões representam o que é chamado de cromatina aberta, isto é, são as regiões onde a cromatina encontra-se no estado ideal (desenovelado) para permitir a ligação de fatores de transcrição. Além disso, conforme explicado anteriormente, a maior vantagem deste método é o fato de que podemos acessar os sítios de ligação desses fatores *trans-acting* com a resolução em pares de bases. Isto é, somos capazes de indicar os sítios exatos de ligação com a ajuda de algum método que analise os motifs encontrados nas regiões de *footprint*.

Uma desvantagem significativa deste método está no fato de que ele não é específico por fator de transcrição. A partir da definição de *footprints* através de algum método de predição computacional, somos capazes de dizer que fatores se ligam em determinadas regiões, porém não somos capazes de especificar quais fatores. Essa é a principal diferença e *trade-off* entre os métodos de DNase-seq e ChIP-seq.

2.3. Modelos Escondidos de Markov

Neste tópico, será descrita a teoria básica para modelos de Markov escondidos. A integração dos sinais genômicos das tecnologias DNase-seq e ChIP-seq será realizada através de um modelo de Markov com emissões representadas por misturas de gaussianas. No texto a seguir serão descritas duas formas de treinamento: o algoritmo de Baum-Welch e por máxima verossimilhança. Foi observado que a forma de treinamento é geralmente a parte mais fraca dos principais trabalhos disponíveis na literatura [16,22], portanto pretende-se estudar cuidadosamente a melhor forma de treinamento para aproveitar ao máximo a informação destes sinais.

Por simplicidade, foi optado por realizar tais definições levando em consideração uma distribuição discreta de probabilidade. A extensão para o caso contínuo (caso deste trabalho), porém, é simples. É necessário apenas que todas as funções discretas de probabilidade (onde cada valor num conjunto discreto tem uma probabilidade de ocorrer) sejam transformadas em funções contínuas (onde cada valor num fluxo contínuo de valores tem uma probabilidade definida por alguma distribuição de probabilidade).

Também foi optado por não definir os modelos de mistura específicos utilizados neste trabalho para que o texto não fique muito longo e enfadonho, porém a idéia, novamente, é bastante simples: basta substituir as emissões contínuas por modelos de misturas de probabilidade. Lembre-se de que um modelo de misturas é necessário para que os diferentes sinais genômicos sejam integrados.

2.3.1. Introdução

O modelo escondido de Markov é uma extensão das cadeias de Markov clássicas. A cadeia de Markov é um modelo probabilístico composto 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 [48]. O modelo escondido de Markov segue esta mesma idéia, porém nele, além da sequência de estados conhecida, temos uma sequência de estados, chamada de caminho, que não é conhecida. 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 o estado escondido mais provável de ter gerado esta sequência de símbolos [49].

Formalizando esta idéia, são feitas as seguintes definições [50]: seja **x** a cadeia de símbolos conhecida, **π** a sequência de estados desconhecidos (caminho), **A** a matriz de transições onde uma célula $\mathbf{a_{ij}}$ desta matriz representa a probabilidade de se transitar do estado i para o estado j, **E** a matriz de emissão no qual será denotada a probabilidade de emissão do símbolo b no estado k por $\mathbf{e_k(b)}$, e **Σ** o alfabeto de símbolos.

Dessa forma podemos definir as seguintes probabilidades:

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

$$e_k(b) = P(x_i = b | \pi_i = k). \quad (2)$$

$$P(x, \pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}. \quad (3)$$

O resultado (3) é obtido a partir da propriedade chave das cadeias de Markov na qual a probabilidade de ir para o estado i+1 depende apenas da probabilidade atual no estado i.

Os algoritmos dos tópicos 2.3.2 e 2.3.3 (a seguir) consistem em métodos para se descobrir o estado escondido **π** a partir de uma sequência de caracteres **x** utilizando um modelo com os parâmetros **A** e **E** definidos. Nesses métodos a equação (3) será explorada e serão criadas novas variáveis para ajudar no entendimento. Os algoritmos dos tópicos 2.3.4 e 2.3.5 mostram formas de se estimar os parâmetros **A** e **E** para um modelo de Markov escondido.

2.3.2. Algoritmo de Viterbi

O algoritmo de Viterbi pertence ao paradigma da programação dinâmica e consiste em descobrir qual é o caminho mais provável **π** dada a sequência de emissão **x**. Em termos formais, queremos encontrar:

$$\pi^* = argmax_\pi P(x, \pi). \quad (4)$$

Seja **L** o tamanho da sequência **x**. O algoritmo é definido da seguinte forma:

Algoritmo Viterbi:
1. Inicialização: $v_0(0) = 1$ e $v_k(0) = 0$ para k > 0.
2. Recursão (i = 1, ..., L):
2.1. $v_l(i) = e_l(x_i)\, max_k(v_k(i-1)a_{kl})$;
2.2. $ptr_i(l) = argmax_k(v_k(i-1)a_{kl})$.
3. Terminação:
3.1. $P(x, \pi^*) = max_k(v_k(L)a_{k0})$;
3.2. $\pi_L^* = argmax_k(v_k(L)a_{k0})$.
4. Traceback (i = L, ..., 1): $\pi_{i-1}^* = ptr_i(\pi_i^*)$.

2.3.3. Probabilidade Posterior

Para calcular a probabilidade posterior, definida em detalhes mais adiante, é necessária a definição de dois algoritmos: *forward* e *backward*.

Neste momento, o objetivo é calcular P(x), isto é, a evidência de uma certa cadeia de caracteres x dentro de todas as possibilidades de cadeias de tamanho L. Formalmente, podemos definir, em relação ao caminho:

$$P(x) = \sum_\pi P(x, \pi). \tag{5}$$

O cálculo exaustivo da equação (5) é impossível pois o número de caminhos cresce exponencialmente com o tamanho da sequência. Porém podemos avaliar esta expressão com o mesmo algoritmo de Viterbi mostrado, apenas modificando os passos de maximização por somatórios. A variável de Viterbi $v_k(i)$, corresponderá então à variável:

$$f_k(i) = P(x_1 \ldots x_i, \pi_i = k). \tag{6}$$

O algoritmo é definido a seguir:

Algoritmo Forward:
1. Inicialização: $f_0(0) = 1$ e $f_k(0) = 0$ para k > 0.
2. Recursão (i = 1, ..., L):
$f_l(i) = e_l(x_i) \sum_k f_k(i-1)a_{kl}$.
3. Terminação:
$P(x) = \sum_k f_k(L)a_{k0}$.

A probabilidade posterior que estamos procurando pode ser definida 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 temos:

$$P(\pi_i = k \mid x) = P(x, \pi_i = k\ ) / P(x). \tag{7}$$

Explorando o termo P(x, π_i = k) e aplicando a propriedade chave das cadeias de Markov temos:

$$\begin{aligned} P(x, \pi_i = k) &= P(x_1 \dots x_i, \pi_i = k) P(x_{i+1} \dots x_L | x_1 \dots x_i, \pi_i = k) \\ &= P(x_1 \dots x_i, \pi_i = k) P(x_{i+1} \dots x_L | \ \pi_i = k). \end{aligned} \tag{8}$$

É possível, então, criar outra variável, chamada *backward*, para calcular o segundo termo do lado direito da equação (8). Obviamente, essa variável é definida da seguinte forma:

$$b_k(i) = P(x_{i+1} \dots x_L | \ \pi_i = k). \tag{9}$$

O procedimento utilizado para computar a expressão (9) é semelhante ao *forward*, porém com recursão inversa. O seguinte algoritmo pode ser utilizado para computar a variável *backward*:

Algoritmo Backward:
1. Inicialização: $b_k(L) = a_{k0}$ para todo k.
2. Recursão (i = L-1, ..., 1):
$b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i+1)$.
3. Terminação:
$P(x) = \sum_l a_{0l} e_l(x_1) b_l(1)$.

A partir dos algoritmos de forward e backward podemos calcular a probabilidade posterior conforme definida na equação (7). Através de uma simples substituição nesta equação é dada origem à seguinte expressão:

$$P(\pi_i = k \mid x) = f_k(i) b_k(i) \ / \ P(x) \tag{10}$$

O termo P(x) da equação (10) é calculado a partir de um dos algoritmos *forward* ou *backward* na sequência inteira.

2.3.4. Método da Máxima Verossimilhança

O método da máxima verossimilhança (MMV) 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 símbolos **x** com rótulo conhecido $\boldsymbol{\pi}$ para calcular os parâmetros mais verossímeis.

Neste caso, 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, e $E_k(b)$ o número de emissões do símbolo b no estado k, o estimador de máxima verossimilhança realiza os seguintes cálculos:

$$a_{kl} = A_{kl} \,/\, \sum_{l'} A_{kl'}. \tag{11}$$

$$e_k(b) = E_k(b) / \sum_{b'} E_k(b'). \tag{12}$$

2.3.5. Algoritmo de Baum-Welch

Quando a sequência de estados escondidos é desconhecida, isto é, não temos uma sequência rotulada, o problema da estimação de parâmetros se torna um pouco mais complexo. Como não existe uma forma de se calcular os parâmetros diretamente como no método da máxima verossimilhança, deve-se usar um procedimento iterativo onde, em cada etapa, é verificado um certo critério de ajuste. Qualquer algoritmo padrão para otimização de funções contínuas pode ser utilizado, porém o método mais comum é o algoritmo de Baum-Welch.

Este algoritmo, derivado do Expectation-Maximization (EM), calcula A_{kl} e $E_k(b)$ como o número esperado de vezes com que essas transições e emissões ocorreriam nos dados. O objetivo, então, é maximizar a seguinte probabilidade (onde θ denota o conjunto de parâmetros do modelo):

$$P(\pi_i = k, \pi_{i+1} = l | x, \theta) = \frac{f_k(i) a_{kl} e_l(x_{i+1}) b_l(i+1)}{P(x)}. \tag{13}$$

O termo do lado direito da equação (13) é obtido a partir da diferenciação do lado esquerdo a respeito dos estados escondidos. Tal manipulação não será exibida pois foge do escopo deste projeto. A partir deste resultado, podemos derivar o número esperado de vezes que a_{kl} é utilizado, através do somatório sobre todas as posições da sequência de caracteres, sobre todas as sequências contidas no conjunto de treinamento:

$$A_{kl} = \sum_j \frac{1}{P(x^j)} \sum_i f_k^j(i) a_{kl} e_l\left(x_{i+1}^j\right) b_l^j(i+1). \tag{14}$$

Onde $f_k^j(i)$ é a i-ésima variável *forward* sobre o estado k, calculada para a sequência j do conjunto de treinamento não rotulado. E $b_l^j(i)$ segue a mesma idéia, porém para a variável *backward*. De forma similar, podemos derivar também o número esperado de vezes nos quais o símbolo b ocorrerá no estado k:

$$E_k(b) = \sum_j \frac{1}{P(x^j)} \sum_{\{i|x_i^j=b\}} f_k^j(i) b_k^j(i). \quad (15)$$

A partir das equações derivadas acima, a idéia geral do algoritmo de Baum-Welch é definida da seguinte forma: Inicializamos os parâmetros de forma aleatória. Procedemos então num processo iterativo composto por duas etapas: na primeira, é calculado o número esperado de vezes nas quais as transições (equação 14) e emissões (equação 15) devem ocorrer no modelo; na segunda etapa, são utilizadas as equações (11) e (12) para calcular os novos parâmetros desta iteração.

Esse procedimento iterativo faz parte do paradigma de busca heurística do tipo subida da encosta (*hill-climbing*). Neste tipo de procedimento garantimos apenas convergência para um máximo local, sendo este dependente exclusivamente de como os parâmetros foram inicializados. Vários critérios de parada podem ser utilizados, entre eles, o término quando o aumento na log-verossimilhança entre o modelo calculado na iteração atual e o modelo calculado na iteração anterior não for maior do que um limite pré-estabelecido, ou a estipulação de um número máximo de iterações, para que o algoritmo não se torne muito custoso.

O algoritmo completo, consistindo apenas na organização das idéias expostas, é exibido abaixo:

Algoritmo Baum-Welch:

1. Inicialização: Inicialize $\theta=\{A, E\}$ aleatoriamente.

2. Iteração (Expectation-Maximization):

2.1. Inicialize A_{kl} e $E_k(b)$ com *pseudocounts*;

2.2. Para cada sequência de treinamento j=1,...,n:

2.2.1. Calcule $f_k^j(i)$ e $b_l^j(i)$;

2.2.2. Use (14) e (15) para incrementar A_{kl} e $E_k(b)$ com as novas variáveis calculadas.

2.3. Calcule os novos parâmetros a_{kl} e $e_k(b)$ formando A_{new} e E_{new} com (11) e (12);

2.4. Calcule a log-verossimilhança de θ_{new} e θ. Após isso faça $\theta \leftarrow \theta_{new}$.

3. Terminação: Pare a iteração da etapa 2 se a log- verossimilhança de θ_{new} não for significativamente maior do que a de θ; ou se um número máximo de iterações for atingido.

2.4. Estudos relacionados

A execução bem sucedida de processos biológicos tais como desenvolvimento, proliferação, apoptose, envelhecimento e diferenciação requer uma série de passos que dependem da expressão espacialmente e temporalmente apropriada. Por este motivo, a desregulação da expressão de um pequeno conjunto de genes pode acarretar no aparecimento de diversas doenças.

Para entender os mecanismos moleculares que governam padrões específicos de expressão gênica é necessário entender as razões pelas quais os genes são expressos nos níveis observados. Dessa forma, é necessário identificar os elementos que regulam a expressão dos genes. A identificação e anotação funcional dos elementos regulatórios no DNA são, portanto, um novo desafio para a genômica moderna.

A literatura recente mostrou que dados de DNase-seq são bastante poderosos para se predizer tais elementos regulatórios. Boyle et al e Crawford et al conduziram uma série de estudos em que as regiões de hipersensibilidade foram vastamente caracterizadas em mapeamentos de alta resolução e ao longo de todo o genoma. Além disso, footprints foram preditos através de abordagens computacionais e estatísticas [20,22,23,25]. Sabo et al [30] desenvolveu uma metodologia quantitativa baseada em algoritmos de agrupamento para realizar uma avaliação digital das regiões de hipersensibilidade. Song et al [34] usaram dados de DNase-seq para gerar mapas genome-wide de visualização de regiões enriquecidas de digestão de DNase I em diversas linhas celulares distintas, aumentando significantemente o número de elementos regulatórios com evidência experimental.

Além de dados de cromatina aberta (DNase-seq), outros tipos de dados epigenéticos têm mostrado grande poder para identificar regiões de interesse regulatório. Entre eles estão os dados relativos às modificações de histonas [8]. Isso provavelmente se dá pelo fato de que as modificações pós-traducionais covalentes dos terminais N das histonas regulam eventos importantes relativos à estrutura da cromatina [7]. Pesquisadores já descobriram mais de 100 modificações de histonas possíveis de acontecerem em um nucleossomo. Como estas estruturas podem conter múltiplas modificações, o número total de possíveis combinações de modificações de histonas excede bastante o número total de nucleossomos no corpo humano [19]. Consequentemente, pesquisas atuais têm focado na identificação de padrões específicos para essas modificações pós-traducionais em diferentes condições ambientais, linhas celulares, tecidos ou em comparação a diferentes padrões de expressão. De fato, vários estudos têm mostrado padrões claros (descritos pelo termo recentemente criado: *chromatin signature*) e têm sugerido a aplicação dos seus resultados em diversos problemas, inclusive na predição de elementos regulatórios [7,9,12,19,15].

2.5. Objetivos

O principal objetivo deste trabalho é, portanto, contribuir para a identificação de tais elementos regulatórios através de um novo método que integra dados de cromatina aberta (obtidos através de experimentos de DNase-seq) e de modificações de histonas (obtidos através de ChIP-seq). Alguns estudos mostraram que tal integração poderia resultar numa espécie de redundância, observando nenhuma melhoria na predição de alguns sítios de ligação de fatores de transcrição conhecidos através de modelos Bayesianos [28]. Porém, vários outros estudos discordaram desta

visão e mostraram análises exploratórias que sugerem que a integração desses dados possa ser a chave para uma identificação probabilística bastante acurada [8,9,22,32].

Será feito uso tanto dos dados disponíveis em grandes repositórios [42], quanto em repositórios com informações sobre motivos conhecidos de fatores de transcrição [40,45,46,47] para criar um modelo probabilístico integrado. Pretende-se que tal modelo seja capaz de capturar padrões específicos nos dados. Sabe-se através de análises iniciais que o modelo tem especificidade, em relação aos dados de cromatina aberta, relativa à linha celular, isto é, os parâmetros dos modelos probabilísticos gerados variam conforme a linha celular, porém dentro de uma linha celular um modelo já consegue definir bem vários elementos anotados. Pretende-se avaliar a especificidade deste modelo integrado proposto, que pode ser específico para linha celular, tipo de modificação de histona (por exemplo, abertura ou fechamento da cromatina) ou tipo de elemento regulatório (por exemplo, *insulator* ou *enhancer*).

De posse dos resultados, pretende-se estabelecer uma série de proposições sobre como tais dados podem ser integrados e sobre como cada tipo de fator de transcrição se comporta em relação aos diferentes sinais epigenéticos presentes nos trechos genômicos onde ele governa o nível de expressão de um ou mais genes. Além de uma análise global baseada nas características computacionais e estatísticas do tema, pretende-se realizar uma análise detalhada baseada nos padrões biológicos resultantes da aplicação deste novo modelo ao longo de todo o genoma.

3. Metodologia

Nesta seção, será descrita a metodologia aplicada no desenvolvimento deste trabalho. Será apresentada uma visão geral do *pipeline* experimental seguida de uma descrição dos conjuntos de dados utilizados. Após isso, serão descritas as principais fases experimentais.

3.1. Visão geral

Este trabalho pretende gerar dois tipos de resultados. O primeiro é uma análise exploratória que pretende identificar padrões a partir dos sinais de DNase-seq e de modificações de histonas. Nesta fase serão criados gráficos com o objetivo de comparar os seguintes itens: as regiões de hipersensibilidade de DNase I (cromatina aberta), regiões enriquecidas a partir de sinais de ChIP-seq de diferentes fatores de transcrição pertencentes a diferentes classes (*enhancers*, *silencers*, *repressors*, *insulators* e *activators*), regiões enriquecidas de sinais de ChIP-seq de diferentes modificações de histonas tanto repressivas quanto ativadoras e diferentes regiões onde foram encontrados *motifs* de preferência de ligação de fatores de transcrição a partir da realização de um *motif matching* em todo o genoma com PWMs obtidas em diferentes repositórios.

A segunda análise pretende agregar a esses elementos os *footprints* gerados através de um modelo probabilístico construído a partir da observação dos resultados da análise exploratória. Será seguida, basicamente, a metodologia descrita em [22] para a geração dos diferentes sinais, normalização e identificação de *footprints*. A principal hipótese estabelecida consiste em melhorar a acurácia do modelo proposto em [22] através da adição dos sinais de modificação de histonas.

3.2. Conjunto de Dados

Os dados de DNase-seq e ChIP-seq são provenientes de um grande estudo chamado ENCODE (*ENCyclopedia Of Dna Elements*) [42]. Este estudo, hospedado no UCSC Genome Browser [43], tem como objetivo a catalogação de todos os elementos regulatórios presentes no DNA através de diversas categorias de dados gerados através de protocolos rígidos por universidades e grupos de pesquisa associados.

No ENCODE, estão disponíveis dados, para diversas linhas celulares, de DNase-seq e ChIP-seq tanto para diversos fatores de transcrição quanto para diversas modificações de histonas. Pretende-se utilizar apenas um subconjunto pequeno de linhas celulares, fatores de transcrição e modificações de histonas inicialmente, porém pretende-se ampliar esse conjunto futuramente.

Além desses dados, também estão sendo utilizadas PWMs obtidas em três repositórios. O primeiro é o Jaspar [40] que possui uma grande quantidade de motifs não-redundantes para diversos tipos de organismo. O segundo é o Transfac [45,47] versão 7.0 pública, que possui diversos motivos obtidos em diversos estudos diferentes. O terceiro é o Uniprobe [46] que disponibiliza motivos gerados através do método de PBM (*Universal Protein Binding Microarray*). Redundâncias entre as PWMs obtidas, isto é, motivos para o mesmo fator obtidos em bases de dados distintas, serão tratados separadamente. Esta medida será tomada pelo fato de cada motivo ter sido criado a partir de métodos experimentais e computacionais distintos, possuindo assim uma qualidade única. Aplicar algoritmos de alinhamento e agrupamento para permitir a junção dos motivos não exclui o fato de que um motivo de boa qualidade pode ser corrompido por um motivo de má qualidade [17].

3.3. Motif Matching

Em geral, elementos regulatórios possuem preferência de ligação em certas regiões do DNA com motivos (motifs) específicos. Um dos mecanismos que serão utilizados para acessar a confiabilidade e acurácia do modelo dependerá de um conjunto de teste criado a partir do *motif matching* dos fatores analisados no genoma e sua sobreposição com dados provenientes da tecnologia de ChIP-seq [22]. Além disso, para a análise exploratória, pretende-se utilizar as coordenadas genômicas mais bem pontuadas pelo algoritmo de *motif match* descrito abaixo como referência central para analisar os sinais de digestão de DNase e modificações de histonas ao redor desta região.

Para realizar o *motif matching*, será utilizada a ferramenta Biopython [41], que consiste em uma grande coleção de métodos especialmente desenvolvidos para bioinformática baseados na linguagem Python. Este procedimento consiste em calcular, para todas as sequências genômicas contíguas de tamanho igual ao motivo (PWM) fornecido, a verossimilhança daquela sequência ser mais provável de ocorrer por chance do que uma sequência aleatória dadas as frequências dos nucleotídeos observados para o genoma em questão. De forma mais simples, este método tem como objetivo retornar as coordenadas genômicas mais semelhantes ao motivo (PWM) fornecido. Caso a sequência seja mais provável de ocorrer do que o background genômico, a pontuação retornada, chamada de bit score, é positiva, caso contrário, é negativa [18,19].

De forma a filtrar os motivos obtidos, serão mantidos apenas aqueles que possuírem um *bit score* maior do que o mínimo entre 70% do máximo bit score possível e 90% da diferença entre o máximo e o mínimo bit score possível [22].

3.4. Identificação das regiões de hipersensibilidade e de enriquecimento de fatores de transcrição

As regiões de hipersensibilidade para a digestão da DNase e de enriquecimento de fatores de transcrição (regiões de picos) serão identificadas através do uso da ferramenta F-seq [21] e de uma adequação simples à uma distribuição gama [22].

O F-seq é uma ferramenta que consiste na implementação do método da Janela de Parzen com parâmetros de suavização adaptados de forma dinâmica. Além disso, o F-seq é bastante útil pois corrige o sinal para vieses causados por *Copy Number Variations* (CNV) e por ruídos nos próprios métodos experimentais. A partir do sinal genômico gerado pelo software F-seq, será realizada uma adequação à uma distribuição gama. Calculados os parâmetros da distribuição, um ponto de corte será estabelecido a partir do *P-value* de 5%.

Existem diversas outras ferramentas para realizar tal identificação de regiões enriquecidas em um conjunto de sinais genômicos. Uma destas é o MACS [5] que foi criada especialmente para dados de ChIP-seq, mas recentemente modificada para permitir a utilização em dados de DNase-seq. Serão estudadas as possibilidades de utilização de mais ferramentas neste estudo. O grande empecilho para o estudo com mais de uma ferramenta está no fato de que a quantidade de dados gerados para cada linha celular ou fator de transcrição é enorme, sendo difícil o armazenamento e manutenção do mesmo.

3.5. Análise dos sinais de digestão da DNase I e modificação de histonas

Estão disponíveis, nos repositórios mencionados, as coordenadas genômicas correspondentes ao alinhamento global dos fragmentos gerados pelo método de DNase-seq ou ChIP-seq. Para fornecer estes sinais como entrada para o modelo probabilístico, serão calculados, além das contagens da quantidade de fragmentos que se sobrepunham em cada bp, sinais correspondentes à normalização e curvatura. A normalização será realizada para evitar que regiões com grandes quantidades de fragmentos se sobressaiam em relação à regiões com pequena quantidade de fragmentos, pois divergências pequenas na quantidade de fragmentos entre duas regiões de hipersensibilidade decorrem apenas do fato de a digestão da enzima DNase ser um evento aleatório, não sendo necessariamente uma medida acurada de confiabilidade. Os sinais de curvatura são de tal forma que assumem valores positivos quando o sinal está aumentando e um valor negativo quando o sinal está diminuindo. Tal sinal tem como objetivo capturar padrões específicos destes métodos (que serão detalhados na seção resultados).

A contagem será realizada de forma trivial. Para cada bp, serão avaliados quantos fragmentos possuem sobreposição com o mesmo. Nesse caso há uma diferença fundamental entre a contagem dos dados de DNase-seq e ChIP-seq para modificações de histonas. No caso do

DNase-seq, o interesse está apenas na região onde houve digestão da enzima DNase I. Esta região corresponde ao início dos fragmentos alinhados, pois, como foi visto na seção 2.2.4, apenas o início (região 5') dos fragmentos é sequenciado pelos métodos de sequenciamento paralelo massivo. Portanto são contabilizados apenas o primeiro bp das sequências (observando, obviamente, se elas foram alinhadas na fita molde ou anti-molde). Para as histonas, estamos interessados em toda a região do nucleossomo capturado pela imunoprecipitação. Novamente, como sequenciamos apenas o trecho inicial dos fragmentos, precisamos estende-lo até que ele possua um tamanho razoavelmente aceito como média para os tamanhos capturados em imunoprecipitação.

O sinal relativo à contagem bruta servirá como entrada para a geração do sinal normalizado, onde, em cada posição genômica, será calculado um score normalizado correspondente a divisão do sinal original pela média de todos os sinais diferentes de 0 dentro de uma janela de tamanho pré-definido. Finalmente, a curvatura será calculada a partir do sinal normalizado como a primeira derivada da curva suavizada através do método de Savitzky-Golay, onde se é interpolado um polinômio de segunda ordem ao sinal baseando-se numa janela de tamanho pré-estabelecido [22].

Um dos objetivos da análise primária é indicar quais destes sinais descreverão os dados de forma mais apropriada. Os dados correspondentes à curvatura do sinal foram aplicados de forma bem sucedida para a predição de elementos regulatórios previamente [22], porém existem motivos para acreditar que a integração dos dados relativos às modificações das histonas seja melhor realizada com os dados da contagem bruta [13,16,32].

3.6. Construção do modelo probabilístico

Após a observação dos resultados da primeira etapa exploratória, pretende-se construir um modelo escondido de Markov para capturar os padrões existentes. Como existem sinais de várias fontes distintas (DNase e as várias modificações de histonas) o modelo deverá possuir emissões correspondentes à misturas de gaussianas. A disposição dos estados deverá seguir a metodologia amplamente aceita de modelagem de duração [49], onde cada estado será analisado em separado e replicado o quanto for necessário para tornar as transições que o envolvem o mais próximas do desejado possível.

O treinamento deste tipo de modelo, para o problema geral de predição de elementos funcionais no DNA, é geralmente a parte mais fraca dos trabalhos analisados na literatura [22,16]. A dificuldade está no fato de existirem poucas regiões com comprovações irrefutáveis de fatores de transcrição quando se trata de diversas linhas celulares distintas. Além disso, estas regiões são pequenas (geralmente do tamanho de uma região de hipersensibilidade comum, em torno de 400 a 600 bp), não fornecendo ao modelo de Markov os parâmetros necessários para uma predição

acurada. Pretende-se focar, portanto, no treinamento do modelo. Existem duas soluções iniciais a serem exploradas. A primeira é o algoritmo de Baum-Welch, que é capaz de encontrar padrões nos dados de forma autossuficiente (sem conjuntos de treinamento), com um algoritmo derivado do *Expectation-Maximization* [49]. A segunda se trata no treinamento com o próprio Viterbi. Nesta forma de treinamento, parâmetros são inicializados aleatoriamente e convergem à medida que são feitas sucessivas iterações entre a predição pelo Viterbi e o treinamento por máxima verossimilhança.

Após a construção do modelo, pretende-se avaliá-lo de forma extensiva através de métodos similares aos descritos em [8,22], utilizando como conjunto de teste regiões verificadas experimentalmente por ChIP-seq somadas aos resultados do *motif matching*. Então, o modelo será refinado até que seja gerado um classificador robusto e que consiga capturar os padrões observados de forma não enviesada, livre de ruídos e com um *tradeoff* aceitável entre sensibilidade e especificidade.

4. Resultados preliminares

A metodologia descrita por Boyle et al [22] foi replicada como forma de verificar se os conceitos foram aprendidos de forma correta. Nesse estudo, foi verificado que footprints podem ser preditos a partir de dados de DNase I, sendo caracterizados como regiões de depleção de sinal entre duas regiões com picos significantes. Para realizar tal identificação automática, foi criado um HMM de cinco estados, mostrado em detalhes na Figura 4.1.

Os parâmetros estimados para o modelo da Figura 4.1 estão descritos na Tabela 4.1. Os parâmetros foram estimados da seguinte forma: Foi gerado um modelo inicial com parâmetros estabelecidos através do método da máxima verossimilhança aplicado em uma anotação manual da região promotora do gene FMR1 no cromossomo X. A partir disso, esse modelo inicial foi utilizado para anotar todo o cromossomo 6 partindo do pressuposto que nele a maioria dos elementos regulatórios são ubíquos, isto é, estão presentes na maioria das linhas celulares pesquisadas. A partir dessa nova anotação, o modelo final foi estimado através do método da máxima verossimilhança novamente.

As Figuras 4.2 e 4.3 mostram os *footprints* identificados para duas regiões de hipersensibilidade distintas no cromossomo 6 da linha celular GM12878 utilizando o genoma versão 18 como referência (para o estudo real, pretende-se utilizar a versão 19, mais recente). As figuras foram geradas a partir do *genome browser* disponibilizado pela UCSC [43].

Na parte superior, de ambas as figuras, estão presentes réguas genômicas com as coordenadas correspondentes às regiões analisadas. Abaixo estão as regiões onde foram encontrados footprints (pequenos quadrados pretos). Abaixo estão quatro sinais genômicos. O primeiro corresponde à contagem dos sítios de digestão da DNase I, porém nesse caso, estendia-se 5bp para o lado esquerdo e direito do bp mais à esquerda (5') dos fragmentos alinhados para permitir uma melhor visualização (sinal mais denso). O segundo, terceiro e quarto correspondem, respectivamente, à contagem, normalização e curvatura, da maneira que foram descritas na seção 3.5.

Além disso, os *footprints* encontrados foram anotados com a ferramenta STAMP [44]. Tal anotação foi realizada apenas para ilustrar a dificuldade no processo de anotação automática, dado a escassez de *motifs* (PWMs) disponíveis. *Footprints* onde não foram encontrados fatores de transcrição com um nível de significância menor do que 1.10^{-6} possuem rótulo "Nomatch".

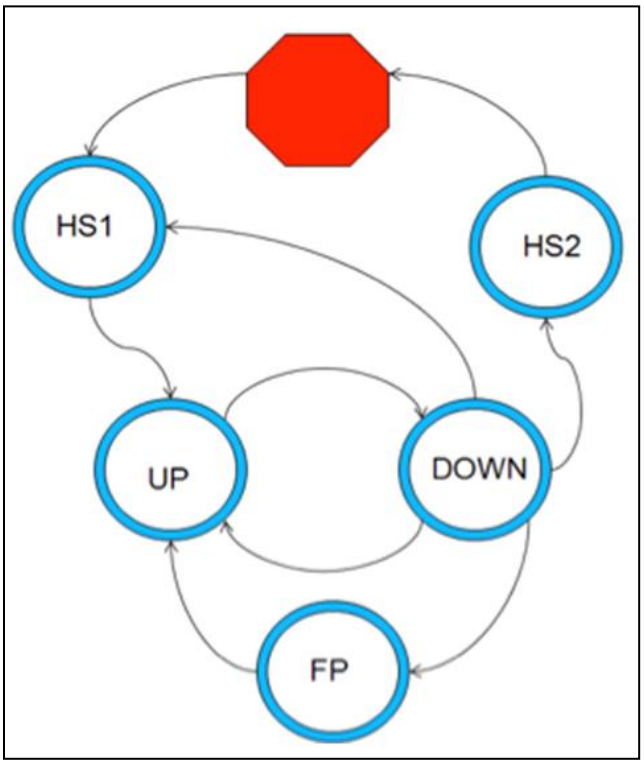

Figura 4.1. HMM de cinco estados. O estado HS1 corresponde ao *background* (região de hipersensibilidade anterior à região enriquecida de picos). Os estados UP e DOWN têm como objetivo capturar as regiões de subida e descida dos picos que ficam entre regiões de interesse. O estado FP corresponde à região de interesse (*footprint*) e o estado HS2 ao *background* após o término das regiões enriquecidas de picos. Fonte: [22]

Tabela 4.1. Parâmetros para o HMM treinado para a linha celular GM12878. As primeiras seis linhas mostram as probabilidades de transição entre os estados do modelo. A última linha exibe os parâmetros de emissão (média e desvio padrão da gaussiana, respectivamente) para cada estado. Fonte [22]

	HS1		UP		DOWN		FP		HS2	
HS1	0.9999		0.0001		0.0000		0.0000		0.0000	
UP	0.0000		0.9000		0.1000		0.0000		0.0000	
DOWN	0.0150		0.0450		0.9000		0.0350		0.0050	
FP	0.0000		0.0300		0.0000		0.9700		0.0000	
HS2	0.0000		0.0000		0.0000		0.0000		1.0000	
Emissão	0	0.85	1.6	1.960	-1.6	1.908	0	0.90	0	0.850

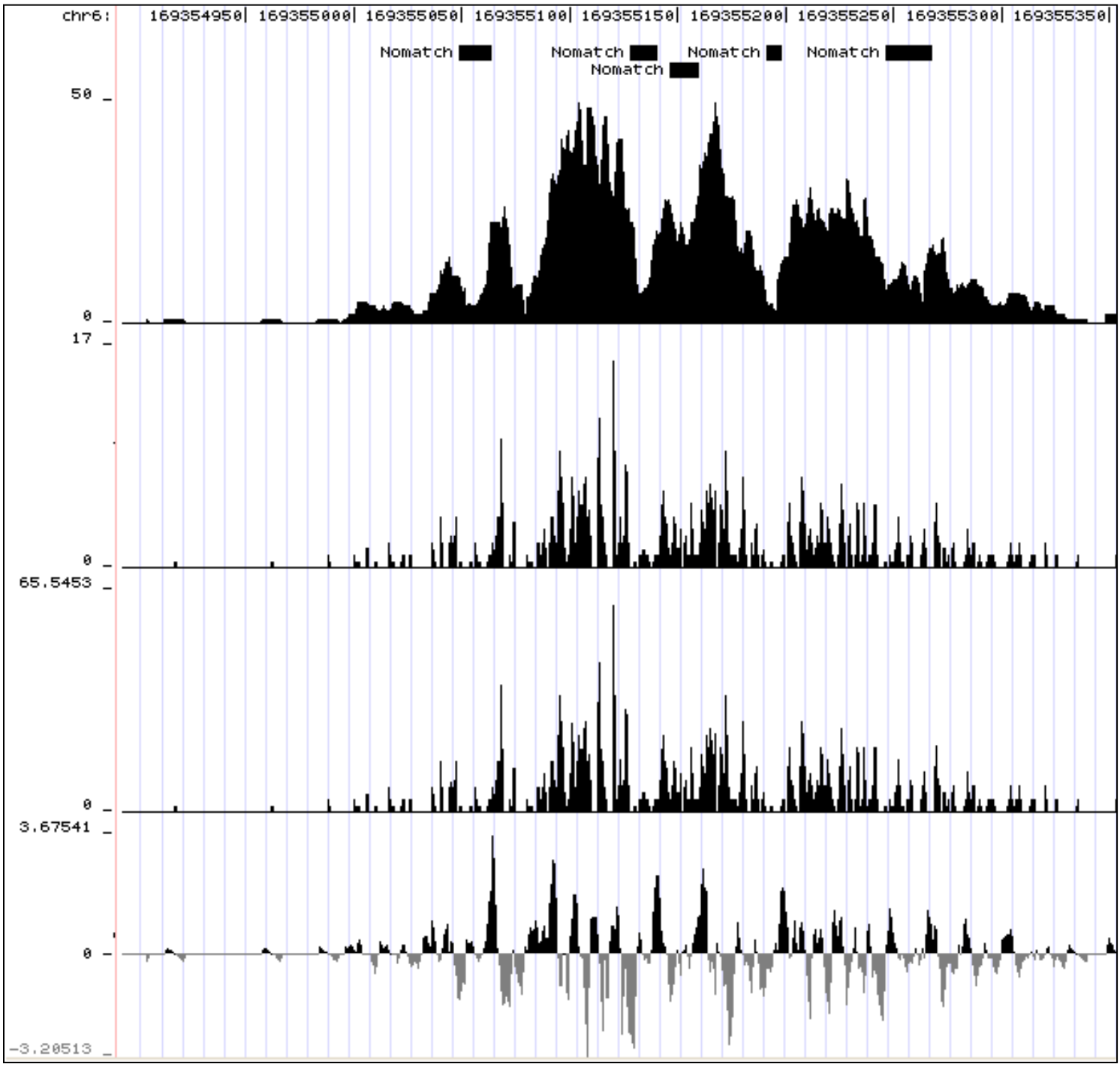

Figura 4.2. Exemplos de *footprints* e sinais de DNase-seq.

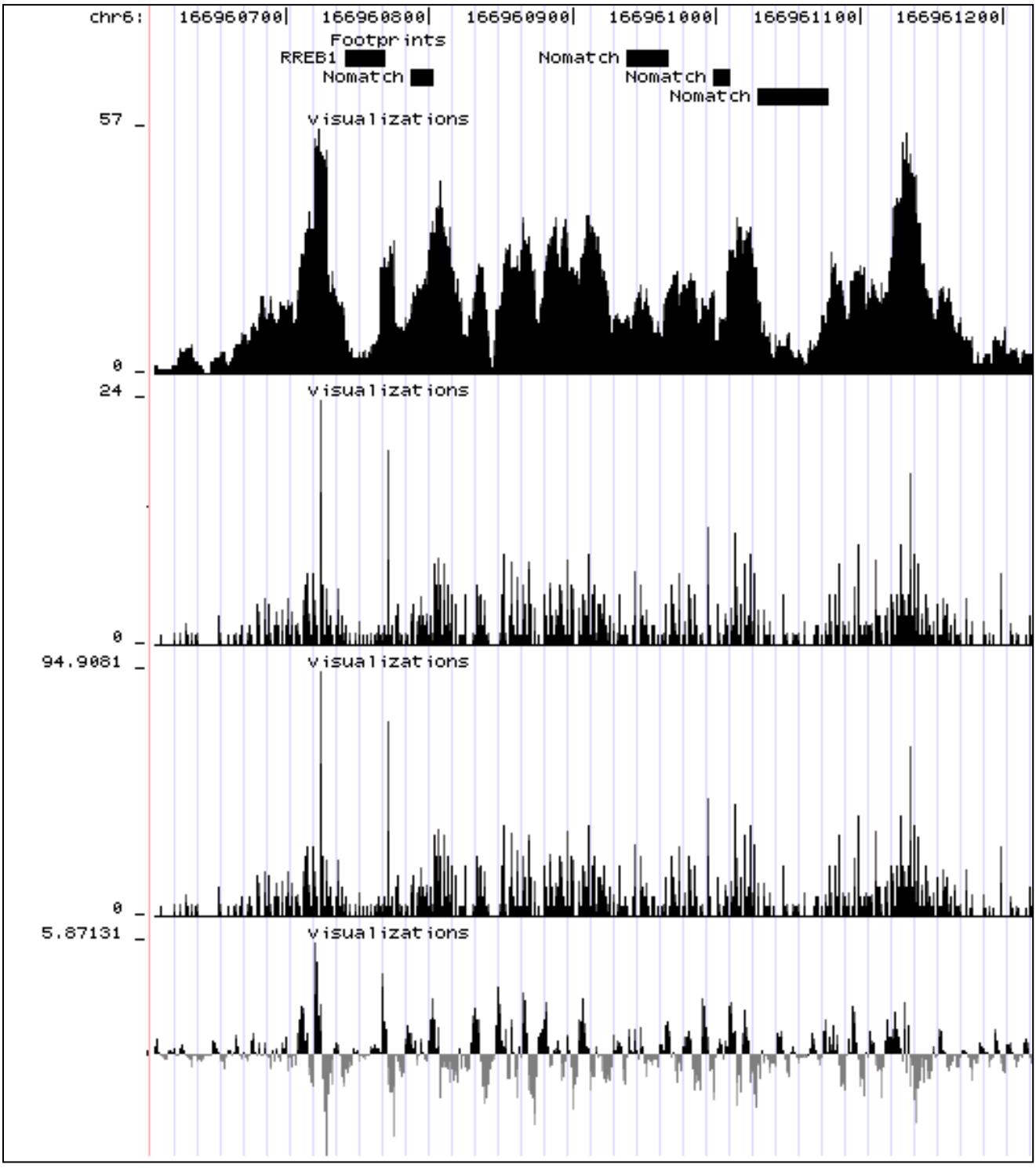

Figura 4.3. Exemplos de *footprints* e sinais de DNase-seq.

5. Cronograma

Foi estabelecido o cronograma de atividades para Janeiro a Julho de 2012 (Tabela 5.1). Tal esquema possui as principais etapas deste projeto de pesquisa, divididas por semana, cujas descrições se encontram abaixo:

- *Base de Dados:* Obtenção e padronização das bases de dados necessárias. Um grande período foi alocado para esta tarefa dado que as bases de dados são bastante volumosas, e pretende-se realizar este estudo para diferentes linhas celulares, fatores de transcrição e modificações de histonas.
- *Experimentos I:* Primeira rodada de experimentos. Esta etapa tem como objetivo a geração dos sinais de digestão de DNase I e de modificações de histonas para a procura de padrões dentro das regiões de hipersensibilidade, levando em consideração, entre outros fatores, os MPBS e os *footprints*.
- *Criação do Modelo:* A partir dos resultados da primeira rodada de experimentos, pretende-se criar um modelo probabilístico que consiga capturar os padrões observados. A principal característica deste modelo será a sua capacidade de integrar os dados de digestão da DNase I (DNase-seq) e de modificação de histonas (através de ChIP-seq).
- *Experimentos II:* Segunda rodada de experimentos. O objetivo é definir novos *footprints* com o novo modelo proposto.
- *Avaliação - Resultados:* Os resultados gerados pela segunda rodada de experimentos serão analisados. Os padrões encontrados serão reexaminados para permitir a comparação entre os resultados esperados e os observados.
- *Reavaliação - Modelo:* Após a avaliação dos novos padrões capturados, o modelo poderá ser reavaliado para permitir pequenas correções probabilísticas. O grande desafio é criar um modelo que seja específico o suficiente para permitir alta precisão, porém que seja generalista o suficiente para permitir sua aplicação em escala genômica, sem que ocorra *overfitting*.
- *Relatório preliminar:* Será escrito um relatório preliminar, que servirá como base e ponto de partida para a elaboração da dissertação. A data de entrega deste relatório servirá como uma espécie de *milestone* deste ciclo experimental inicial.
- *Experimentos III:* De posse de um modelo reavaliado, esta fase tem como objetivo: padronização e limpeza do código, re-execução de etapas do processo onde houver necessidade, aplicação do novo modelo, geração dos resultados quantitativos e qualitativos.
- *Análises Finais:* À medida que os resultados serão gerados, eles serão interpretados sob a ótica tanto da biologia quanto da ciência da computação, para que pequenos erros possam ser corrigidos e para permitir que os resultados gerais mostrem claramente uma aceitação total, parcial ou não-aceitação da hipótese proposta.

- *Relatórios / Artigos:* Redação dos relatórios necessários e organização do estudo em formato de jornal / periódico para que as idéias sejam devidamente expostas e para que a contribuição deste trabalho para a pesquisa científica seja completa. Pequenos relatórios serão gerados ao final de todos os meses para permitir um melhor acompanhamento por parte do orientador.
- *Dissertação (início):* A dissertação têm início previsto para Maio. Pequenos trechos, principalmente relativos à revisão de literatura e base biológica e computacional para o estudo, estão sendo escritos à medida que são estudados. A previsão para finalização da primeira versão da dissertação é o início de Outubro.

Tabela 5.1. Cronograma de atividades para Janeiro a Julho de 2012

Período (2012)	Jan	Fev	Mar	Abr	Mai	Jun	Jul
Base de Dados	■	■					
Experimentos I		■	■				
Criação do Modelo			■	■			
Experimentos II			■	■	■		
Avaliação - Resultados				■	■		
Reavaliação - Modelo				■	■		
Relatório preliminar				■	■		
Experimentos III					■	■	
Análises Finais					■	■	■
Relatórios / Artigos			■	■	■	■	■
Dissertação (início)					■	■	■

6. Conclusão

Este documento teve como objetivo a descrição da proposta de projeto de mestrado do aluno Eduardo Gade Gusmão, na área de biologia computacional. Foram realizadas a revisão da literatura e definição dos experimentos a serem realizados e foi proposto um cronograma de atividades para o primeiro semestre de 2012.

Com este trabalho, pretende-se dar uma contribuição na área de transcriptômica (regulação gênica) no que concerne a identificação de elementos regulatórios no DNA através de sinais advindos de métodos de DNase-seq e ChIP-seq.

Duas possíveis extensões deste trabalho foram estipuladas e serão realizadas ainda este ano caso as análises descritas neste documento sejam cumpridas no tempo estabelecido:

- A primeira consiste na realização deste mesmo pipeline experimental para um grande número de linhas celulares (a saber, K562, HeLa-S3, H1-hESC, HepG2, NHEK, HUVEC, GM12878, GM12891, GM12892, GM18507, GM19238, GM19239, GM19240), um grande número de fatores de transcrição (por volta de 30) e um grande número de modificações de histonas (por volta de 12).
- A segunda consiste na integração de outros tipos de dados no modelo probabilístico. Tais dados podem ser originários, por exemplo, em experimentos FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements), experimentos chip (DNase-chip, ChIP-chip ou expressão gênica com microarranjos) ou dados relativos à conservação gênica.

O aluno está, nesta presente data, recebendo bolsa de mestrado pela instituição de fomento FACEPE. O aluno declara dedicação exclusiva ao projeto de pesquisa. Não existem conflitos de interesses declarados.

Referências Bibliográficas

[1] Buck, M.J., Lieb, J.D. "ChIP-chip Considerations: For the Design, Analysis, and Application of Genome-Wide Chromatin Immunoprecipitation Experiments". *Genomics*. 83(3), 349-360 (2004)

[2] Park, P.J. "ChIP–seq: Advantages and Challenges of a Maturing Technology". *Nature Reviews Genetics*. 10(10), 669-680 (2009).

[3] Song, J.S., Johnson, W.E., Zhu, X., Zhang, X., Li, W., Manrai, A.K., Liu, J.S., Chen, R., Liu, X.S. "Model-Based Analysis of Two-Color Arrays (MA2C)". *Genome Biology*. 8(8), R178.1-R178.13 (2007)

[4] Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M., Sidow, A. "Genome-Wide Analysis of Transcription Factor Binding Sites Based on ChIP-seq Data". *Nature Methods*. 5(9), 829-834 (2008).

[5] 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. "Model-Based Analysis of ChIP-seq (MACS)". *Genome Biology*. 9(9), R137 (2008).

[6] Ball, P. "Portrait of a Molecule". *Nature*. 421(6921), 421-422 (2003).

[7] Barski, A., Cuddapah, S., Cui, K., Roh, T., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., Zhao, K. "High-Resolution Profiling of Histone Methylations in the Human Genome". *Cell*. 129(4), 823-837 (2007).

[8] Cuellar-Partida, G., Buske, F.A., McLeay, R.C., Whitington, T., Noble, W.S., Bailey, T.L. "Epigenetic Priors for Identifying Active Transcription Factor Binding Sites". *Bioinformatics*. 28(1), 56-62 (2012).

[9] Ernst, J., Kellis, M. "Discovery and Characterization of Chromatin States for Systematic Annotation of the Human Genome". *Nature Biotechnology*. 28(8), 817-825 (2010)

[10] Felsenfeld, G., Groudine, M. "Controlling the Double Helix". *Nature*. 421(6921), 448-453 (2003).

[11] Grant, P.A. "A Tale of Histone Modifications". *Genome Biology*. 2(4), reviews0003.1-0003.6 (2001).

[12] Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, D., Barrera, L.O., Calcar, S.V., Qu, C., Ching, K.A., Wang, W., Weng, Z., Green, R.D., Crawford, G.E., Ren, B. "Distinct and Predictive Chromatin Signatures of Transcriptional Promoters and Enhancers in the Human Genome". *Nature Genetics*. 39(3), 311-318 (2007).

[13] Hon, G., Wang, W., Ren, B. "Discovery and Annotation of Functional Chromatin Signatures in the Human Genome". *PLoS Computational Biology*. 5(11), e1000566 (2009)

[14] Schones, D.E., Zhao, K. "Genome-Wide Approaches to Studying Chromatin Modifications". *Nature Reviews Genetics*. 9(3), 179-191 (2008).

[15] Spivakov, M.S., Fisher, A.G. "Epigenetic Signatures of Stem-Cell Identity". *Nature Reviews Genetics*. 8(4), 263-271 (2007).

[16] Won, K., Ren, B., Wang, W. "Genome-Wide Prediction of Transcription Factor Binding Sites Using an Integrated Model". *Genome Biology*. 11(1), R7 (2010)

[17] Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E. "WebLogo: A Sequence Logo Generator". *Genome Res*. 14(6), 1188-1190 (2004).

[18] D'haeseleer, P. "What are DNA sequence motifs?". *Nature Biotechnology*. 24(4), 423-425 (2006)

[19] Stormo, G.D. "DNA Binding Sites: Representation and Discovery". *Bioinformatics*. 16(1), 16-23 (2000)

[20] Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., Crawford, G.E. "High-Resolution Mapping and Characterization of Open Chromatin across the Genome". *Cell*. 132(2), 311-322 (2008).

[21] Boyle, A.P., Guinney, J., Crawford, G.E., Furey, T.S. "F-Seq: A Feature Density Estimator for High-Throughput Sequence Tags". *Bioinformatics*. 24(21), 2537-2538 (2008).

[22] Boyle, A.P., Song, L., Lee, B.K., London, D., Keefe, D., Birney, E., Iyer, V.R., Crawford, G.E., Furey, T.S. "High-Resolution Genome-Wide In Vivo Footprinting of Diverse Transcription Factors in Human Cells". *Genome Res. Biol.* 21(3), 456-464 (2011).

[23] Crawford, G.E., Holt, I.E., Mullikin, J.C., Tai, D., National Institutes of Health Intramural Sequencing Center, Green, E.D., Wolfsberg, T.G., Collins, F.S. "Identifying Gene Regulatory

Elements by Genome-Wide Recovery of DNase Hypersensitive Sites". *PNAS*. 101(4), 992-997 (2004).

[24] Crawford, G.E., Davis, S., Scacheri, P.C., Renaud, G., Halawi, M.J., Erdos, M.R., Green, R., Meltzer, P.S., Wolfsberg, T.G., Collins, F.S. "DNase-chip: A High Resolution Method to Identify DNase I Hypersensitive Sites Using Tiled Microarrays". Nature Methods. 3(7), 503-509 (2006).

[25] 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. "Genome-Wide Mapping of DNase Hypersensitive Sites Using Massively Parallel Signature Sequencing (MPSS)". Genome Res. 16(1), 123-131 (2006).

[26] Gross, D.S., Garrard, W.T. "Nuclease Hypersensitive Sites in Chromatin". *Ann. Rev. Biochem*. 57, 159-197 (1988).

[27] Keene, M.A., Corces, V., Lowenhaupt, K., Elgin, S.C.R. "DNase I Hypersensitive Sites in Drosophila Chromatin Occur at the 5' Ends of Regions of Transcription". *Proc. Natl. Acad. Sci.* 78(1), 143-146 (1981).

[28] Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y., Pritchard, J.K. "Accurate Inference of Transcription Factor Binding from DNA Sequence and Chromatin Accessibility Data". *Genome Res.* 21(3), 447-455 (2011).

[29] Sabo, P.J., Humbert, R., Hawrylycz, M., Wallace, J.C., Dorschner, M.O., McArthur, M., Stamatoyannopoulos, J.A. "Genome-Wide Identification of DNaseI Hypersensitive Sites Using Active Chromatin Sequence Libraries". *PNAS*. 101(13), 4537-4542 (2004).

[30] Sabo, P.J., Hawrylycz, M., Wallace, J.C., Humbert, R., Yu, M., Shafer, A., Kawamoto, J., Hall, R., Mack, J., Dorschner, M.O., McArthur, M., Stamatoyannopoulos, J.A. "Discovery of Functional Noncoding Elements by Digital Analysis of Chromatin Structure". *PNAS*. 101(48), 16837-16842 (2004).

[31] Sabo, P.J., Kuehn, M.S., Thurman, R., Johnson, B.E., Johnson, E.M., Cao, H., Yu, M., Rosenzweig, E., Goldy, J., Haydock, A., Weaver, M., Shafer, A., Lee, K., Neri, F., Humbert, R., Singer, M.A., Richmond, T.A., Dorschner, M.O., McArthur, M., Hawrylycz, M., Green, R.D., Navas, P.A., Noble, W.S., Stamatoyannopoulos, J.A. "Genome-Scale Mapping of DNase I Sensitivity In Vivo Using Tiling DNA Microarrays". *Nature Methods*. 3(7), 511-518 (2006).

[32] Shu, W., Chen, H., Bo, X., Wang, S. "Genome-Wide Analysis of the Relationships between DNaseI HS, Histone Modifications and Gene Expression Reveals Distinct Modes of Chromatin Domains". *Nucleic Acids Research*. 39(17), 7428-7443 (2011).

[33] Song, L., Crawford, G.E.: DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. Cold Spring Harb. Protoc. 2010(2), pdb.prot5384 (2010)

[34] Song, L., Zhang, Z., Grasfeder, L.L., Boyle, A.P., Giresi, P.G., Lee, B., Sheffield, N.C., Graf, 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. "Open Chromatin Defined by DNaseI and FAIRE Identifies Regulatory Elements that Shape Cell-Type Identity". *Genome Res*. 21(10), 1757-1767 (2011).

[35] Essien, K., Vigneau, S., Apreleva, S., Singh, L.N., Bartolomei, M.S., Hannenhalli, S. "CTCF Binding Site Classes Exhibit Distinct Evolutionary, Genomic, Epigenomic and Transcriptomic Features". *Genome Biology*. 10(1), R131 (2009).

[36] Giresi, P.G., Lieb, J.D. "Isolation of Active Regulatory Elements from Eukaryotic Chromatin Using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements)". *Methods*. 48(3), 233-239 (2009).

[37] Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., Lieb, J.D. "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) Isolates Active Regulatory Elements from Human Chromatin". *Genome Res*. 17(6), 877-885 (2007).

[38] Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., Ren, B. "Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome". *Cell*. 128(6), 1231-1245 (2007).

[39] Maston, G.A., Evans, S.K., Green, M.R. "Transcriptional Regulatory Elements in the Human Genome". *Annu Rev Genomics Hum Genet*. 7, 29-59 (2006)

[40] Byrne, J.C., Valen, E., Tang, M.E., Marstrand, T., Winther, O., Piedade, I., Krogh, A., Lenhard, B., Sandelin, A. "JASPAR, the Open Access Database of Transcription Factor-Binding Profiles: New Content and Tools in the 2008 Update". *Nucleic Acids Research*. 36, D102-D106 (2008).

[41] Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J.L. "Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics". *Bioinformatics*. 25(11), 1422-1423 (2009)

[42] Rosenbloom, 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., Hinrichs, 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. "ENCODE Whole-Genome Data in the UCSC Genome Browser: Update 2012". *Nucleic Acids Res*. 40, D912-D917 (2012).

[43] Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D. "The Human Genome Browser at UCSC". *Genome Res*. 12(6), 996-1006 (2002).

[44] Mahony, S., Benos, P.V. "STAMP: A Web Tool for Exploring DNA-Binding Motif Similarities". *Nucleic Acids Research*. 35, W253-W258 (2007).

[45] Matys, V., Kel-Margoulis, 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. "TRANSFAC and its Module TRANSCompel: Transcriptional Gene Regulation in Eukaryotes". *Nucleic Acids Research*. 34, D108-D110 (2006).

[46] Newburger, D.E., Bulyk, M.L. "UniPROBE: An Online Database of Protein Binding Microarray Data on Protein–DNA Interactions". *Nucleic Acids Research*. 37, D77-D82 (2009).

[47] Wingender, E., Dietze, P., Karas, H., Knuppel, R. "TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites". *Nucleic Acids Research*. 24(1), 238-241 (1996).

[48] Bishop, C.M. “Pattern recognition and machine learning”. *Springer*. 803p. (2006)

[49] Durbin, R., Eddy, S., Krogh, A., Mitchison, G. “Biological sequence analysis: probabilistic models of proteins and nucleic acids”. *Cambridge University Press*. 366p. (2002)

[50] Rabiner, L.R. “A tutorial on hidden markov models and selected applications in speech recognition”. *Proceedings of the IEEE*. 77(2), 257-286. (1989)

[51] DNA Sequencing Website. <http://www.dnasequencing.org>

[52] Allis, C.D., Jenuwein, T., Reinberg, D., Caparros, M. "Epigenetics". *Cold Spring Harbor Laboratory Press*. 502p. (2007)