Universidade de Brasília
Instituto de Ciências Biológicas
Departamento de Biologia Celular
Programa de Pós Graduação em Biologia Molecular

Fluxograma computacional para detecção e análise de sequências potencialmente formadoras de Z-DNA utilizando Bioconductor

Halian Gonçalves Vilela

Brasília, 27 de junho de 2012

Universidade de Brasília
Instituto de Ciências Biológicas
Departamento de Biologia Celular
Programa de Pós Graduação em Biologia Molecular

Fluxograma computacional para detecção e análise de sequências potencialmente formadoras de Z-DNA utilizando Bioconductor

Dissertação de Mestrado apresentada como requisito parcial à obtenção do título de Mestre em Biologia Molecular.

Halian Gonçalves Vilela

Orientador:
Marcelo Brígido

Co-orientadora:
Tainá Raiol

Brasília, 27 de junho de 2012

Dissertação de Mestrado sob o título **"Fluxograma computacional para detecção e análise de sequências potencialmente formadoras de Z-DNA utilizando Bioconductor"**, defendida por Halian Gonçalves Vilela e aprovada em 27 de junho de 2012, em Brasília, Distrito Federal, pela banca examinadora constituída pelos doutores:

Prof. Dr. Marcelo de Macedo Brígido
IB/Biomol-UnB
Orientador

Drª. Tainá Raiol de Alencar
IB/Biomol-UnB
Co-Orientadora

Prof. Dr. David John Bertioli
IB/GEM-UnB
Examinador Externo

Drª. Natália Florêncio Martins
EMBRAPA/CENARGEN
Examinadora Externa

Drª. Maria Emília M. T. Walter
CiC-UnB
Membro Suplente

Dedicatória

Dedico este trabalho à minha querida irmã Nathália Gonçalves Vilela. Porque uma dissertação de mestrado com certeza é muito mais importante do que um convite de graduação. Te amo irmãzinha!

Agradecimentos

A seção de agradecimentos é sempre a mais desorganizada em um trabalho como este. Talvez porque seja a hora em que o autor esquece as formalidades, deixa as emoções fluirem e se preocupa somente em não esquecer das pessoas importantes de sua vida, sem as quais, seria muito mais difícil completar qualquer tipo de objetivo.

No meu caso não poderia ser diferente, a completude deste trabalho deve-se muito à inúmeras pessoas importantes, espero não esquecer de nenhuma, ou pelo menos citá-las impessoalmente em expressões como "e todo o pessoal"e "a galera"(por favor, não se sintam excluídos!).

Começando dos mais próximos, não poderia deixar de agradecer primeiramente à minha perfeita namorada e parceira Pollyana por todo o tempo que me apoiou durante esses 8 anos lindos de namoro. Por todas as reclamações, alegrias, explicações e "nerdismos"que teve que ouvir, por todos os momentos de dúvidas em que soube me responder e pela paciência nos momentos mais tensos! Obrigado, amor!!

Em relação à família, é até difícil falar, meus pais, Denise e Pedro, por serem tão compreensivos com um filho tão ausente, que vai embora de casa muito cedo pra morar com os avós e ainda se dá ao luxo de ficar pegando o carro emprestado sempre que precisa!!! Hahahaha, é muita pretensão! Pai, muito obrigado por todos os dias em que se sacrificou para facilitar minha vida, não vou esquecer isso, com certeza sem sua compreensão teria sido muito mais difícil completar este trabalho. Mãe, obrigado por sempre ouvir com paciência minhas reclamações, por me dar colo (claro que não literalmente, hehehehe) nos momentos de procrastinação em que minha inspiração acabava e a vontade de estudar ia embora sem nem ter chegado ainda!! Desculpem-me, os dois, pelos momentos que eu perdia a paciência e os tratava de maneira inadequada, vocês sabem que isso é só um artefato da minha personalidade, mas que no fundo amo vocês demais!!!

Vó e vô, Lourdes e Márcio, meus segundos pais (não só pelo dito popular, mas nesse caso literalmente). Obrigado pelo abrigo, por me dar tudo que eu sempre precisei quase que instantaneamente. Vó, obrigado por me mimar tanto, hahahaha, com certeza a vida fica muito mais fácil e agradável com uma vozinha tão atenciosa como você! Não sei o

que seria das minhas sessões enormes de trabalho se você pra me oferecer uma comida ou um lanche sempre em boa hora! Vô, que eu nunca chamo assim e não será aqui a primeira vez, então corrigindo... Velhinho, obrigado pela preocupação com meu futuro, pelas inúmeras caronas até minha casa com papos efusivos sobre carreira e sobre o mundo em geral, apesar de nossas opiniões muitas vezes serem opostas, nossos debates são muito engrandecedores e com certeza sempre pesam (por mais que possa não parecer) nas minhas decisões. Me desculpem vocês também pelos momentos de falta de paciência, como disse para os meus pais, são só momentos, o amor por vocês é enorme e inabalável.

Meu tio Eduardo que me apresentou ao mundo da ciência, me fazendo acompanhar a sua saga desde o finalzinho da graduação, passando pelo mestrado, até o doutorado. Sempre com muitos conhecimentos acumulados sobre diversas áreas, respondendo com paciência e detalhismo todos os questionamentos daquele moleque curioso que eu era quando criança. Valeu véi!!

Vó Júlia e Tia Peta, a prontidão de vocês para ajudar não pode deixar de ser citada. Não me lembro do dia em que precisei de alguma coisa de vocês e que não recebido na hora com a maior atenção e preocupação do mundo, seja almoços de emergência, docinhos (ah, o manjar!) e comidinhas ou quaisquer outras coisas. Muito obrigado por tudo!

Minha irmãzinha Nathália, desculpa ter te esquecido no convite tá? Acho que a página anterior compensa isso né? Hehehee, mas antes que pareça só uma retratação, eu tenho que te agradecer por ser tão legal, apesar das brigas você é uma excelente pessoa, muito divertida, carinhosa (às vezes, heuaheuaheuae) e talentosa. Quero ver você brilhar nas pistas de dança daqui há uns anos!! Te amo! Obrigado pela paciência e compreensão em que você e o Vini viam para cá e eu pouco podia interagir por estar vidrado na construção dessa dissertação, obrigado aos dois!

Obrigado também aos meus queridos sogros, Clésio e Dora, por me acolherem em sua casa como um filho em todos os momentos, por confiarem em mim desde quando eu era um maltrapilho com cara de metaleiro maluco (hahahaha, tá, eu não era tão ruim assim né?) e por terem colocado no mundo a filha maravilhosa que eu tenho a honra de me relacionar.

Ao pessoal do laboratório, tenho que agradecer muito ao meu orientador Marcelo Brígido, por ter acreditado na minha capacidade, que mesmo não sendo nem cientista da computação e nem biólogo, poderia fazer um mestrado em Bioinformática! Obrigado por todos os esclarecimentos, paciência, compreensão, liberdade e piadinhas infames! Com certeza ter um orientador como você facilita muito o trabalho de qualquer estudante.

O mesmo vale para minha co-orientadora Tainá, a nossa querida e poderosa PÓS-DOC! (barulhos de raios e trovões) por todas as dicas, explicações e disponibilidade, mesmo que fora de hora. Pelas dicas e pela ajuda imensa nos complicados experimentos de bancada que tomaram bastante tempo e apesar de não terem constado nos resultados do trabalham me ajudaram a crescer como pesquisador. Obrigados à todos os meus outros co-orientadores informais, Prof. Maria Emília, Prof. Andrea Maranhão e Prof. Ildinete que me deram dicas importantíssimas para que fosse possível completar o trabalho a tempo. Prof. Maria Emília, muito obrigado pela confiança em dar atribuições tão importante como gerenciar o site do BSB2011 e compor o grupo dos seletos organizadores desse importante congresso!

Aos meus *brothers* da computação, Paulo, Saad, Lessa, Túlio e Ruben. Creio que o aprendizado que obtive com vocês foi uma das coisas mais importantes desse mestrado. Seria impossível eu ter aprendido tanto sem a ajuda de vocês, o nível de conhecimento é assustador, me dá muito orgulho de ter trabalhado com vocês, espero que possamos manter contato sempre e trabalharmos juntos novamente em seja lá qual empreitada resolvamos nos meter!! Valeu mesmo!

À minha querida amiga bióloga Bia! Que compartilhou comigo muitos momentos de incerteza na parte da multidisciplinariedade, computação + biologia AO MESMO tempo não é pra qualquer um né? Mas quem disso que somos "qualquer um"né Bia?? Muito obrigado pela sua tutoria na parte dos experimentos na bancada, com certeza sem a sua ajuda paciente eu não teria conseguido fazer sequer a mais simples das PCRs!

Ao meu grande amigo Robson, que é um dos responsáveis por eu ter feito esse mestrado, empolgando com o assunto e me contagiando com o espírito de cientista nato! Obrigado por nossas enormes conversas sobre a situação da pesquisa no Brasil, sobre nossas incertezas em relação à carreira, obrigado por me ouvir e por me pedir conselhos. Sempre me envaidecia muito todas as vezes que me pedia opinião por confiar muito em mim. És um cara que respeito e admiro muito, e quando alguém que você respeita e admira te elogia, o ego incha! Agradeço também o Prof. Dr. Márcio Poças, pois veio dele a notícia que o laboratório de Bioinformática estava precisando de gente pra trabalhar.

Obrigado a todo o pessoal do laboratório de imunologia molecular, especialmente à Galina, que também me ajudou imensamente na parte de bancada e ao Rafael Burtet por ter feito meu nome rodar o mundo junto com interessante trabalho dele! Obrigado a todos os funcionários do Biomol, especialmente ao Thompson por sua enorme disponibilidade para resolver todos os nossos problemas!

Obrigado também aos amigos que ficaram de fora, observando o processo, e que por causa do mestrado eu muitas vezes me fiz ausente, o pessoal do Dimensão, grandes amigos de infância, e ao pessoal do La-Salle (mesmo sem nunca ter estudado lá), grandes amigos de adolescência. Obrigado ao Rodolfo por ter me dado a oportunidade na hora certa de testar minhas habilidades e achar um possível caminho a seguir, o site deu certo, com muitos perrengues mas deu certo! E também ao Jorge por confiar tanto em mim e dar a oportunidade de aprender sempre em nossos trabalhos em conjunto.

E por fim, mas não menos importante agradeço a mim mesmo. Quem me conhece sabe que eu gosto de desafios e passar 2 anos estudando um assunto de ponta como Bioinformática é um desafio enorme. Estou feliz de ter conseguido vencê-lo.

Resumindo, para não entristecer os que não foram diretamente citados... OBRIGADO A TODOS!

「一つのことから一万のことを知れ」
- - 宮本武蔵
五輪書、地の巻

Resumo

O Z-DNA é uma conformação alternativa da molécula de DNA envolvida na regulação da expressão gênica. Porém, a função específica desta estrutura no metabolismo celular ainda não foi totalmente elucidada. Este trabalho apresenta um fluxograma de análise que utiliza o ambiente `R` para investigar regiões potencialmente formadoras de Z-DNA (ZDRs) ao longo de genomas. Tal método combina a análise termodinâmica empregada pelo conhecido *software* `Z-Catcher` com a capacidade de manipulação de dados biológicos dos pacotes do `Bioconductor`. A metodologia desenvolvida foi aplicada no cromossomo 14 do genoma humano como estudo de caso e com isso foi possível estabelecer uma correlação entre as ZDRs e os sítios de início da trancrição (TSSs), que se mostrou de acordo com resultados de estudos anteriores. Além disso, foi possível demonstrar que ZDRs posicionadas no interior de genes tendem a ocorrer preferencialmente em *introns* ao invés de *exons* e que ZDRs à montante dos TSSs podem ter correlação positiva com estimulação da atividade da RNA polimerase.

Palavras-chave: Z-DNA, ZDR, Z-Catcher, R, Bioconductor

Abstract

Z-DNA is an alternative conformation of the DNA molecule implied in regulation of gene expression. However, the exact role of this structure in cell metabolism is not yet fully understood. Presented in this work is a novel Z-DNA analysis workflow which employs the `R` software environment to investigate Z-DNA forming regions (ZDRs) throughout genomes. It combines thermodynamic analysis of the well-known software `Z-Catcher` with biological data manipulation capabilities of several `Bioconductor` packages. The methodology was applied in the human chromosome 14 as a case study. With that, a correlation was established between ZDRs and transcription start sites (TSSs) which is in agreement with previous reports. In addition, the workflow was able to show that ZDRs which are positioned inside genes tend to occur in intronic sequences rather than exonic and that ZDRs upstream to TSSs may have a positive correlation with the up-regulation of RNA polymerase activity.

Keywords: Z-DNA, ZDR, Z-Catcher, R, Bioconductor

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos, Siglas e Abreviaturas

ΔG Variação de energia livre de Gibbs

σ Densidade de *supercoiling*

ADAR1 *Double-stranded RNA-specific adenosine deaminase* (Desaminase de adenosina de RNA fita-dupla 1)

A-DNA *Deoxyribobucleic acid, conformation A* (Ácido desoxiribonucleico, conformação A)

BAM *Binary Sequence Alignment/Map*

B-DNA *Deoxyribobucleic acid, conformation B* (Ácido desoxiribonucleico, conformação B - canônica)

C Linguagem de Programação C

C-DNA *Deoxyribobucleic acid, conformation C* (Ácido desoxiribonucleico, conformação C)

ChIP-Seq *Chromatin Immunoprecipitation with massively parallel DNA sequencing* (Imunoprecipitação da cromatina com sequenciamento de alto desempenho)

c-MYC *Avian myelocytomatosis viral oncogene homolog* (Homólogo ao oncogene viral aviário de mielocitomatose)

CPU *Central processing unit* (Unidade central de processamento)

CSF-I *Colony stimulating factor-1* (Fator estimulador de colônia-1)

ddNTP *Dideoxy nucleoside triphosphate* (Dideoxi nucleosídeo trifosfato)

DLM1 Um dos nomes da proteína ZBP1 (Z-DNA binding protein 1)

DNA *Deoxyribonucleic acid* (Ácido desoxirribonucleico)

dNTP *Deoxy nucleoside triphosphate* (Deoxi nucleosídeo trifostato)

E3L Fator de virulência do Vaccinia virus

ENCODE *ENCyclopedia Of DNA Elements* (Enciclopédia de elementos do DNA)

Ensembl Projeto conjunto do EMBL-EBI (European Molecular Biology Laboratory - European Bioinformatics Institute) e Wellcome Trust Sanger Institute para anotação automática de genomas eucarióticos

FASTA Formato de arquivo de texto proveniente do antigo pacote de alinhamento FASTA (FAST-All)

FORTRAN *The IBM Mathematical FORmula TRANslating System*, antiga linguagem de programação

FTP *File transfer procotol* (Protocolo de transferência de arquivos)

GB Gigabyte, 10^9 bytes

GHz Gigahertz, 10^9 Hertz

GRCh37 *Genome Reference Consortium human genome 37*

HTS *High throughput sequencing* (Sequenciamento de alto desenpenho)

IDE *Integrated development environment* (Ambiente de desenvolvimento integrado)

MCF7 *Michigan Cancer Foundation-7* (linhagem de células de carcinoma mamário humano)

MHz Megahertz, 10^6 hertz

mRNA *Messenger ribonucleic acid* (Ácido ribonucleico mensageiro)

NCBI *National Center for Biotechnology Information*

PCR *Polymerase chain reaction* (Reação em cadeia da polimerade)

RAM *Random access memory* (Memória de acesso aleatório)

RNA *Ribonucleic acid* (Ácido ribonucleico)

RNA-Seq *Ribonucleic acid sequencing* (Sequenciamento de ácido ribonucleico)

RPM Revoluções por minuto

SAM *Sequence Alignment/Map*

SATA-II *Serial Advanced Technology Attachment-II*

SIBZ *Stress induced B-Z*

SQL *Structured Query Language* (Linguagem de consulta estruturada)

SRA *Short reads archive*

TSS *Transcription start site* (Sítio de início da transcrição)

UCSC *University of California Santa Cruz*

Z-DNA *Deoxyribobucleic acid, conformation Z* (Ácido desoxiribonucleico, conformação Z)

ZDR *Z-DNA forming region* (Região potencialmente formadora de Z-DNA)

1 Introdução

O DNA é uma molécula de estrutura dinâmica, coexistindo várias conformações diferentes em equilíbrio umas com as outras. A forma canônica, mais conhecida, dessa molécula é a chamada B-DNA, sua presença é dominante ao longo dos diferentes genomas e muito já se sabe sobre a sua estrutura e comportamento. Outras formas como o A-DNA e o Z-DNA podem surgir em condições específicas. A estrutura do Z-DNA, porém chama a atenção por ser muito distinta do B-DNA, sua hélice gira para a esquerda ao invés da direita, suas bases demonstram uma disposição alternada onde há uma rotação ao redor das ligações glicosídicas e por fim, o *backbone* da molécula exibe uma estrutura de *zig-zag*, característica que deu origem ao nome **Z**-DNA. Essa diferença estrutural faz com que o DNA na conformação Z difira suficientemente da B a ponto de haver ligantes seletivos para essa conformação, assim podemos observar uma alta antigenicidade e também uma especificidade de ligação por parte de algumas proteínas como a ADAR1 (Rich e Zhang, 2003).

Estas características peculiares do Z-DNA, associadas às descobertas de que ele está presente *in vivo* em regiões transcricionalmente ativas, levaram a crer que deveria haver alguma importância em termos de função biológica inerente à essa conformação. Há fortes evidências que sugerem a participação ativa do Z-DNA na transcrição. Estudos mostraram que a formação de Z-DNA após a abertura de um nucleossomo impede que esse nucleossomo volte a se formar, mantendo assim o gene transcricionalmente ativo por mais tempo (Garner e Felsenfeld, 1987). Também foi mostrado que regiões potencialmente formadoras de Z-DNA estão presentes em abundância próximos aos sítios de início da transcrição (TSS) por todo o genoma (Li *et al.*, 2009). Diante deste panorama, este trabalho sugere um fluxograma computacional que busca facilitar a análise de regiões potencialmente formadoras de Z-DNA, possibilitando a busca por padrões de distribuição e correlação com TSSs ou outros motivos gênicos importantes.

1.1 A alternância conformacional do DNA

Existem várias conformações descritas para a molécula do DNA que podem surgir em determinadas circustâncias e ambientes aos quais a molécula possa vir a ser submetida. Algumas dessas são raras ou transientes, como o C-DNA, que é uma estrutura que tende a ocorrer em um ambiente de umidade mais baixa e na presença de íons Li^+ em excesso. Essa estrutura foi descrita como simplesmente uma pequena variação estrutural da forma B, devido às condições específicas do ambiente (Dam e Levitt, 2000). A forma A-DNA foi uma das primeiras a serem descobertas, sua ocorrência se dá preferencialmente em condições de desidratação, e sua característica estrutural mais marcante é a hélice mais curta e larga em relação à conformação B; os pares de bases são mais inclinados e distantes do eixo de rotação da hélice e o período da hélice é ligeiramente maior que o da forma B (11bp por rotação comparados à 10-10,5bp da forma B) (Basham, Schroth e Ho, 1995). Exemplos de conformações do DNA podem ser vistos na figura 1.

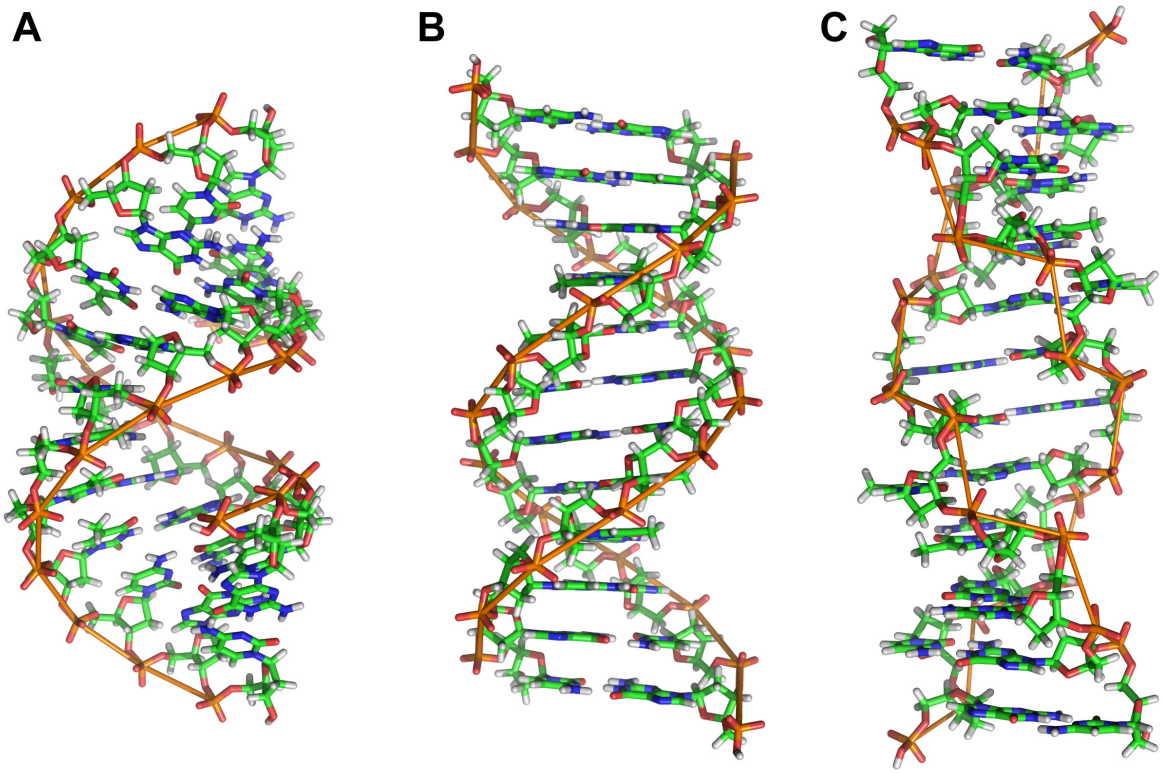

Figura 1: **Diferentes Estruturas do DNA.** A-DNA **(a)**, hélice mais curta (volta completa = 2,3Å) e larga (diâmetro = 23Å) em comparação à forma canônica B-DNA **(b)** que exibe uma hélice com altura de 3,32Å, diâmetro de 20Å e período menor. A forma Z-DNA **(c)**, possui uma hélice cujo o giro é para esquerda e exibe um padrão de *zig-zag* no *backbone* da molécula, sua altura é de 45,6Å e diâmetro 18Å (Wheeler, 2007a).

1.1.1 Z-DNA

Conforme pode ser visto na figura 1, a estrutura do Z-DNA difere bastante da estrutura do B-DNA, os detalhes dessas diferenças podem ser vistos na figura 2 abaixo.

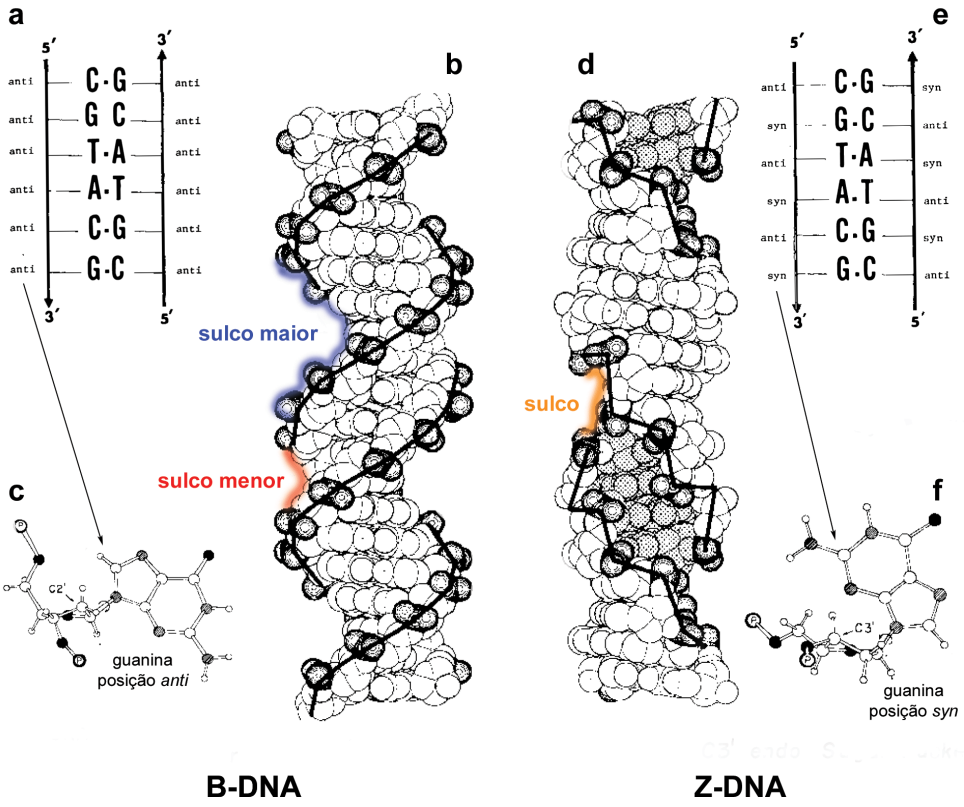

Figura 2: **Diferenças conformacionais entre Z e B-DNA.** Em **(a)** e **(e)** podemos ver como as bases se alternam nas conformações *anti* e *syn*, **(c)** e **(f)** mostram as diferença entre essas conformações com mais detalhes. As diferença nas hélices podem claramente ser vistas em **(b)** e **(d)**, onde estão evidenciadas as diferenças entre os sulcos, o lado da rotação e o *zig-zag* da forma Z (Rich, Nordheim e Wang, 1984, adaptado).

A diferença mais perceptível entre Z e B-DNA é o giro da hélice. Na molécula de Z-DNA a rotação da hélice é levógira, ou seja, o giro é para a esquerda enquanto que na forma B a rotação é dextrógira, para a direita. O *backbone* exibe um padrão de *zig-zag* ao longo da molécula formando somente um sulco por período, ao contrário dos dois sulcos, maior e menor, da forma B (Fig.2 **b** e **d**). A conformação de bases nitrogenadas também difere, essas conformações dizem respeito à orientação da base nitrogenada das purinas em relação à pentose correspondente. Como não há nenhum impedimento estérico, a base nitrogenada pode girar ao redor da ligação glicosídica, que liga a base à pentose. Nas conformações *anti* a base nitrogenada projeta-se de maneira a afastar-se da pentose, enquanto que na conformação *syn* ocorre o giro em torno da ligação de maneira que seus átomos mantém-se próximos à pentose (Fig.2 **c** e **f**). No B-DNA, todos os nucleotídeos estão na conformação

anti, enquanto que na forma Z há a alternância entre *anti* e *syn* ao longo de toda a hélice (Fig.2 **a** e **e**), esta alternância modifica a maneira como os nucleotídeos se empilham formando então o padrão de *zig-zag* característico (Rich, Nordheim e Wang, 1984).

1.1.2 *Supercoiling*

A formação do Z-DNA é um processo físico-químico complexo. A maior proximidade dos grupos PO_4 (fosfato) e a conformação *syn* das bases faz da conformação Z uma estrutura de maior energia livre em comparação com a B (Rich e Zhang, 2003). Isso indica que, para a transição de uma forma a outra, é necessário haver um ganho de energia. Um elemento importante, para esta formação, é o fenômeno mecânico conhecido como *supercoiling*, que armazena energia potencial capaz de estabilizar a transição da forma B para a Z.

Sabe-se que a molécula do DNA é extremamente longa e, para se acomodar no núcleo, faz-se necessário um processo de compactação que forma a cromatina. A dupla-hélice de DNA, associa-se quimicamente a proteínas chamadas histonas, cujo caráter alcalino, oposto ao caráter ácido do DNA, garante uma forte interação eletrostática entre as duas partes. A partir desta interação primária, observa-se vários níveis de compactação intermediários que culminam com a acomodação final na forma dos cromossomos, presente na divisão celular. A figura 3 mostra os diferentes níveis de compactação da mólecula de DNA. Definindo *supercoiling* por sua etimologia, é possível perceber o panorama em que ocorre no DNA. *Coil* pode ser traduzido como "bobina", ou seja, um segmento de corda ou fio enovelado em torno de um mesmo eixo de rotação, formando um segmento helicoidal (ᔓᔓᔓᔓᔓᔓᔓ). É possível que este segmento, por sua vez, seja enovelado novamente em torno de um segundo eixo, isto definiria um processo de superenovelamento, ou *supercoiling*. Neste contexto, ao fazer uma analogia do DNA com um fio já enovelado (formando a dupla-hélice), no processo de formação da cromatina temos a ocorrência de *supercoiling*. O *supercoiling* é importante não só para a compactação do DNA, mas também para que o processo de transcrição seja facilitado.

Em um segmento retilíneo de DNA, com as extremidades livres e em condições fisiológicas, a estrutura helicoidal é muito estável, a probabilidade de abertura da dupla hélice, mesmo que somente entre pares de base individuais, é extremamente baixa (da ordem de 10^{-5}) (Lukashin *et al.*, 1976). Para que a transição de um estado a outro do par de bases (pareado para aberto) seja possível, é necessária a variação da energia livre do segmento. Nestas condições, esta variação só ocorre com variação de temperatura.

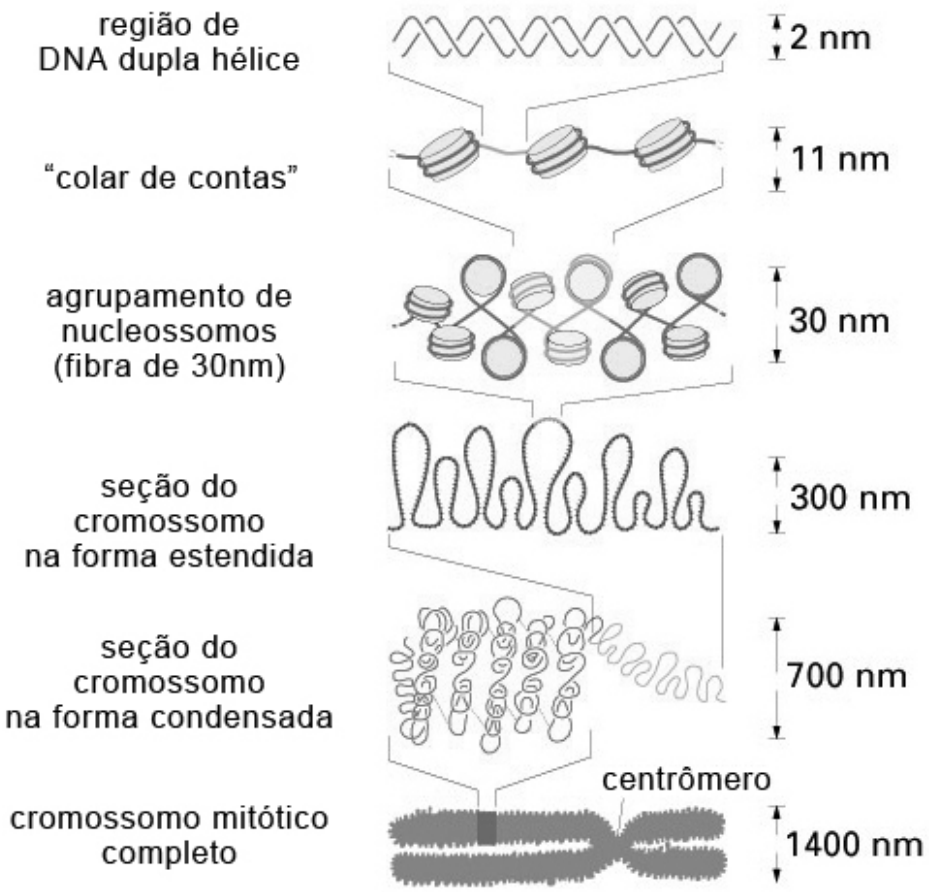

Figura 3: **Níveis de compactação do DNA.** De cima para baixo, observa-se o DNA em sua forma nativa de dupla hélice, a seguir, a interação da molécula com as histonas, formando nucleossomos que se organizam em uma estrutura conhecida como "colar de contas". Cada uma das três "contas" mostradas é um nucleossomo. Mediante a presença da histona H1, a estrutura anterior se compacta ainda mais formando uma fibra de 30nm. As fases subsequentes, cada vez mais compactadas, surgem mediante à necessidade da divisão celular. A compactação começa na fase da intérfase e culmina com o cromossomo totalmente compactado que pode ser observado durante a metáfase (Alberts *et al.*, 2008).

Porém, em um segmento circular de DNA, como um plasmídio, ou em uma situação onde ambas as extremidades da molécula estejam fixas, o panorama energético da estrutura não depende somente da temperatura (Vologodskii *et al.*, 1979). Como o sistema sempre procura manter-se em equilíbrio, caso este equilíbrio seja perturbado, por exemplo, ao adicionar mais uma volta ou tentar abrir a dupla-hélice, será necessário para o sistema introduzir uma mudança de conformação a fim de tentar restabelecê-lo. Para descontar essa perturbação, o sistema tende a introduzir voltas sobre si próprio modificando a macro-estrutura do segmento. Assim, o que antes era um segmento circular, pode passar a ser um segmento em forma de 8 (oito) ou cruciforme , dependendo do nível de perturbação. Cada volta extra da macro-estrutura pode ser definida como um *supercoil*, e o tipo de

perturbação determina qual o tipo de *supercoil*, se negativo ou positivo. Este processo e definições podem ser vistos na figura 4.

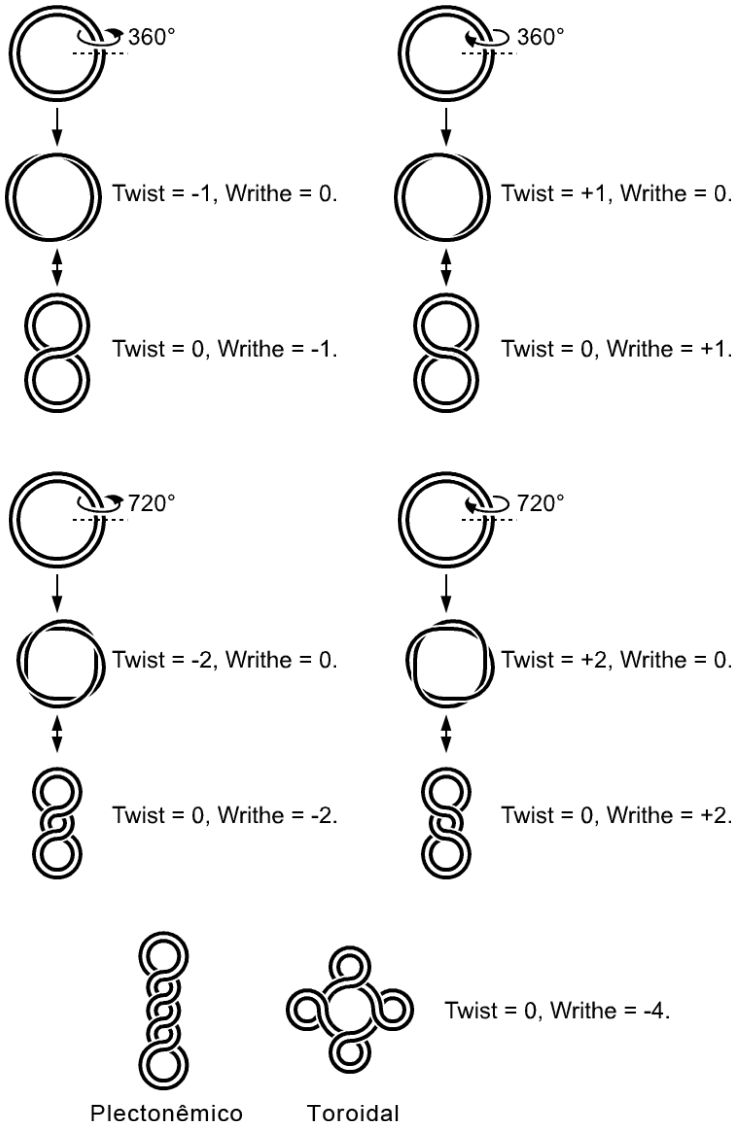

Figura 4: **Diferentes níveis de *supercoiling* em um segmento circular de DNA.** dependendo do sentido da perturbação, representada pelas setas circulares, é gerado *supercoiling* negativo ou positivo (respectivamente). Os parâmetros *twist* e *writhe* representam respectivamente a quantidade de voltas além da quantidade basal (determinada pela estrutura da dupla-hélice) e a quantidade de voltas da macro-estrutura sobre si mesma, ou seja, o *supercoiling* propriamente dito. Pode-se notar que quanto mais voltas além da quantidade basal forem introduzidas, mais retorcida ficará a estrutura (Wheeler, 2007b).

Apesar do DNA em eucariotos não estar na forma circular, a maneira como está compactado faz com que as regiões onde ocorre transcrição se comporte como segmentos onde as duas extremidades estão fixas. Sendo assim, o panorama energético da estrutura exibe um equilíbrio entre o *supercoiling* e a abertura da hélice, na qual o relaxamento do *supercoiling* é capaz de diminuir a energia livre necessária para a abertura da dupla-hélice por meio das topoisomerases (Wang, 1974). A via oposta também ocorre, assim, há um aumento de *supercoiling* decorrente da abertura da dupla-hélice para a passagem da maquinaria de transcrição. Um exemplo deste processo pode ser visto na figura 5.

A quantidade de energia livre presente na estrutura superenovelada é proporcional ao quadrado da quantidade de *supercoils* presentes. No entanto, se no segmento principal, uma porção da dupla-hélice mudar a rotação da direita para a esquerda (o que ocorre na transição de B para Z-DNA), esta energia livre pode também estabilizar este segmento (da mesma maneira como facilita a abertura da hélice) e consequentemente diminuir o número de *supercoils*. Por este motivo, o processo de *supercoiling* é tão importante para a formação do Z-DNA (Nordheim e Rich, 1983).

Um parâmetro chamado densidade de *supercoiling* (σ) fornece informação sobre quão superenovelado está o segmento de DNA. Este pode ser definido como a razão entre a variação da quantidade de voltas atualmente presentes no segmento e a quantidade "natural" de voltas presentes quando o segmento está em equilíbrio. A equação a seguir define esse parâmetro:

$$\sigma = \frac{\Delta Lk}{Lk_0} \tag{1.1}$$

Onde ΔLk é a variação ($Lk - Lk_0$) do número de ligação (*linking number*) da hélice em relação ao equilíbrio. Esse número é dado por $\frac{N}{h_0}$ que representa a razão entre o número de bases (N) e o número de bases por volta da hélice (h_0). A partir deste parâmetro σ é possível estimar por meio de cálculos termodinâmicos a quantidade de energia livre necessária para que ocorram as transições conformacionais da dupla hélice, seja para a abertura, seja para a transição de B para Z-DNA (Liu e Wang, 1987).

1.1.3 Importância biológica do Z-DNA

Durante anos especulou-se sobre qual seriam as prováveis funções do Z-DNA nos organismos, à medida que os estudos foram avançando, evidências apontavam cada vez mais para o fato de que a estrutura não era simplesmente fruto de equilíbrio termodinâmico, mas sim, que poderia ter algum papel ativo em eventos biológicos. Os principais fatos que contribuíram para tal hipótese foram a correlação da estrutura com a transcrição, a

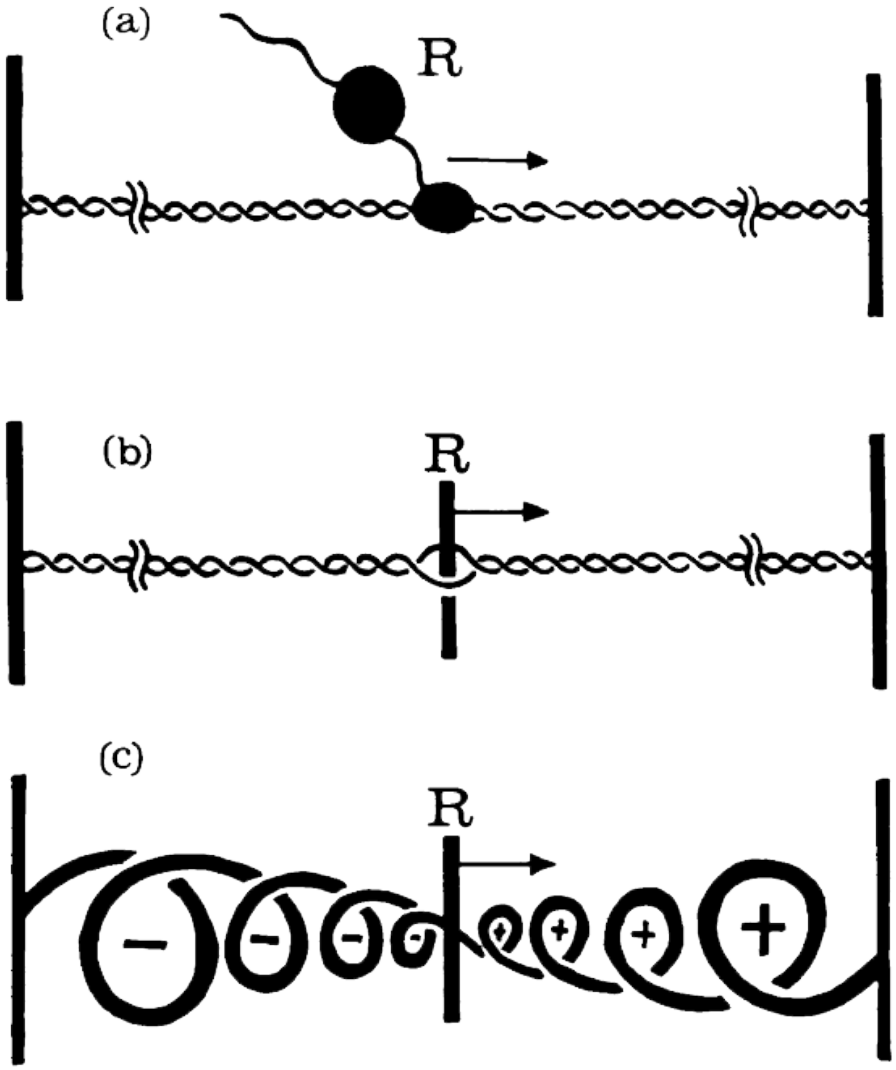

Figura 5: **Processos de *supercoiling* decorrente da passagem do aparato transcricional.** Em **(a) R** representa o aparato transcricional completo, composto pela RNA polimerase, o fragmento de mRNA nascente e as proteínas a este acopladas. O aparato move-se no sentido da transcrição, representado pela seta, e as barras negras nas extremidades representam as unidades maiores de compactação da qual o fragmento faz parte. Em **(b)** o aparato é representado como um divisor da dupla hélice em duas partes distintas, estas partes sofrem tensões torcionais de sinais opostos **(c)** à medida que a transcrição avança. *Supercoiling* negativo ocorre a montante do aparado, enquanto que a jusante, observa-se *supercoiling* positivo. (Liu e Wang, 1987)

antigenicidade (em oposição ao B-DNA que não é antigênico) em conjunto com a participação em doenças auto-imunes e por fim, a existência de proteínas com domínios de ligação específico (indicando importância em eventos evolutivos) (Rich e Zhang, 2003).

O primeiro estudo a correlacionar o Z-DNA com a transcrição mostrou que há formação de Z-DNA após a passagem do complexo da RNA polimerase como consequência do *supercoiling* negativo introduzido pela abertura da mesma (Liu e Wang, 1987), conforme

discutido na seção anterior. Baseados nestes fatos, e aliados aos conhecimentos adquiridos por diversos estudos conformacionais, pesquisadores puderam desenvolver ferramentas computacionais capazes de prever o potencial de formação de Z-DNA de sequências genômicas. Três *softwares* foram desenvolvidos: `Z-hunt`(Ho *et al.*, 1986), `Z-huntII` (Schroth, Chou e Ho, 1992) e `Z-Catcher`(Xiao, Dröge e Li, 2008). Apesar de algoritmos bem diferentes entre si, todos os programas usam uma abordagem similar, primeiro buscam por sequências repetitivas, ricas em alternância de purinas e pirimidinas, requisitos estruturais para a formação de Z, e depois executam cálculos termodinâmicos para inferir se a disposição dessas sequências favorece a formação da Z-DNA.

Com o avanço das possibilidades de detecção e utilizando tais programas foi possível mapear regiões genômicas inteiras e demonstrar que a disposição dessas sequências não se davam ao acaso, mas sim que havia uma certa preferência às proximidades dos TSSs, fortalecendo as evidências entre a correlação desta estrutura com o ambiente transcricional. A natureza antigênica do Z-DNA também foi um fator importante que chamou a atenção de pesquisadores. Inclusive, a utilização de anticorpos monoclonais *anti*-Z auxiliou nas pesquisas elucidando de forma experimental as evidências puramente computacionais até então (Rich e Zhang, 2003). Descobriu-se por meio destes estudos que nas regiões próximas a promotores do gene c-MYC[1], Z-DNA é formado durante a transcrição do gene e rapidamente revertido em B-DNA caso a transcrição cesse (Wölfl, Wittig e Rich, 1995). Também emergiram hipóteses sobre um possível papel regulador desta formação sobre a transcrição, na qual a formação de Z-DNA poderia ser responsável por impedir a formação de nucleossomos, mantendo assim a estrutura susceptível à ligação de fatores de transcrição e do aparato transcricional. Estes efeitos foram observados em estudos com os genes do fator estimulador de colônias I (CSF-I)[2] (Liu *et al.*, 2001). A descoberta de proteínas como a ADAR1 (Desaminase de adenosina de RNA fita-dupla 1), que exibem domínios de ligação específicos ao Z-DNA, também ajudaram a elucidar o papel biológico da conformação Z. Estas proteínas tem o papel de ligar-se a segmentos de pre-mRNA dupla-fita formados pelo pareamento de *exons* com *introns*. Uma vez ligada, a enzima cataliza o processo de deaminação da adenosina, transformando-a em inosina que, ao ser processada pelos ribossomos, é interpretada como guanina (Herbert *et al.*, 1995). Este processo demonstra um importante fator de variabilidade proteica, e o domínio de ligação ao Z-DNA desta proteína pode indicar um mecanismo de guia para genes transcricionalmente ativos que necessitam da edição em nível do pre-mRNA.

[1] gene supressor de tumor que codifica fatores de transcrição que controlam o ciclo celular. A mutação e consequentemente perda de função leva a um descontrole do ciclo celular e tumorigênese.

[2] uma das citocinas que induzem a diferenciação de células tronco hematopoiéticas

Após a descoberta e caracterização do motivo de ligação ao Z-DNA da proteína ADAR1 ($Z\alpha_{ADAR1}$), foi possível caracterizar diversos outros motivos semelhantes em proteínas tanto do genoma humano quanto de outros organismos, assim foram descobertas proteínas como a DLM1, encontrada em tecidos adjacentes a tumores e relacionada à resposta a interferons, e E3L, importante para garantir patogenicidade viral de certas variedades dos *vaccinia* vírus (Silva, 2010). Ambas exibiam motivos muito semelhantes ao $Z\alpha_{ADAR1}$, sugerindo que fossem capazes de se ligar ao Z-DNA. Isto foi demonstrado no estudo feito com E3L, em que alterações no motivo de ligação ao Z-DNA resulta no enfraquecimento da força de ligação, causando a perda da capacidade de ligação aos TSSs e consequentemente permitindo que o hospedeiro responda à infecção, o que reduz drasticamente a patogenicidade do vírus (Kim *et al.*, 2003), mostrando mais uma vez a importância biológica do Z-DNA.

1.1.4 Métodos Computacionais para Detecção de Z-DNA

Conforme já discutido na seção 1.1.3, a criação de métodos computacionais para a detecção de sequências potencialmente formadoras de Z-DNA impulsionou várias descobertas no campo. Tais métodos se mostram bastante importantes para a triagem inicial de sequências a se estudar, direcionando os dispendiosos experimentos biológicos de bancada. `Z-Hunt` (Ho *et al.*, 1986) foi o primeiro método a ser criado. O processo de detecção é feito introduzindo partes da sequência de tamanhos fixos (16 a 24 nucleotídeos), em um plasmídio virtual de 4.263 pares de base sob condições padronizadas (em termos de energia livre). Caso o fragmento não apresente alternância de purinas e pirimidinas, ele já é descartado de início, caso contrário a análise continua. Neste plasmídio, é permitida a transição de B para Z-DNA somente para o fragmento introduzido. Então, sob estas condições controladas, é calculada a propensão deste fragmento para formar Z-DNA considerando as energias de transição de cada dinucleotídeo (estimados em diversos estudos anteriores) em função da densidade de *supercoiling* do plasmídio. A partir dos resultados dos cálculos (solução analítica de uma função), uma pontuação é dada ao fragmento, esta pontuação, chamada *Z-score*[3] é decorrente da comparação entre este fragmento e um conjunto de fragmentos gerados aleatoriamente, portanto representa um certo número médio de nucleotídeos aleatórios que devem ser buscados para se achar uma sequência com potencial de formação de Z-DNA igual ou maior que o fragmento sendo analisado. A primeira versão do `Z-Hunt` foi inovadora, porém pouco prática, visto que sua implementação em `FORTRAN` permitia somente análise de sequências de até 1Mb. Posteriormente o algoritmo

[3]não relacionado com o *z-score* da estatística tradicional

foi atualizado gerando o programa Z-HuntII (Schroth, Chou e Ho, 1992), implementado em `C`, que seguia basicamente o mesmo princípio de busca e pontuação.

O outro método, chamado `Z-Catcher` (Xiao, Dröge e Li, 2008), será utilizado neste trabalho, portanto o detalhamento do algoritmo encontra-se no capítulo 3. Em termos gerais, o `Z-Catcher` difere-se do *Z-Hunt* por considerar a variabilidade na densidade de *supercoiling* (σ) no contexto da análise, tanto que σ é um dos parâmetros de entrada do programa. A busca pelo pontencial formador de Z-DNA se dá por meio de um ciclo de cálculos que considera a energia de transição de cada dinucleotídeo individualmente, comparando o σ calculado ao σ introduzido pelo usuário, diferenciando-se do `Z-Hunt` no ponto em que as sequências resultantes não são expressas por meio de um modelo probabilístico.

O mais recente método para detecção de Z-DNA, chamado SIBZ (*Stress Induced B-Z*) (Zhabinskaya e Benham, 2011) difere dos anteriores por ser o único a considerar o equilíbrio termodinâmico de toda a sequência ao invés de testar somente os dinucleotídeos individualmente. Assim, este método e capaz de detectar a formação de Z-DNA levando em consideração o contexto competitivo das transições B-Z, onde cada base pode estar hora na conformação B, hora na conformação Z, sendo que cada transição modifica o perfil de equilíbrio, afetando assim as transições subsequentes. Este panorama é o mais próximo do que ocorre de fato *in vivo*, o que tornaria este método o mais próximo das predições experimentais. O método só esta disponível ao público através de uma interface *web* (`http://benham.genomecenter.ucdavis.edu`) e devido à impossibilidade de integração com as ferramentas aqui apresentadas, não foi considerado neste trabalho.

1.2 ChIP-Seq - Imunoprecipitação da cromatina associada à sequenciamento de alto desempenho (HTS)

Devido à natureza antigênica do Z-DNA, o uso de anticorpos específicos tornou-se uma ferramenta muito útil na investigação e localização dessas sequências em experimentos biológicos. Uma técnica muito promissora para esse tipo de investigação é o ChIP-Seq (*Chromatin Immunoprecipitation sequencing*), que alia a especificidade dos anticorpos com a resolução do sequenciamento de alto desempenho. Essa técnica consiste na utilização de um anticorpo com especificidade contra uma determinada macromolécula, geralmente proteínas associadas ao DNA como fatores de transcrição ou histonas. Para

estudos sobre Z-DNA, o alvo seria o próprio DNA na conformação Z, visto que, conforme já mencionado, esta apresenta antigenicidade. A princípio, o primeiro passo para se realizar um experimento de ChIP é fazer o *cross-linking*, ou seja, tratar a célula com algum agente químico, tal como o formaldeído, para que as ligações entre as proteínas de interesse e o DNA se tornem covalentes. Após essa ligação o DNA é fragmentado por sonicação ou digestão enzimática e os anticorpos são então adicionados ao sistema. Isso fará com que seja formado um complexo anticorpo-proteína-DNA. Os anticorpos necessitam dispor de algum mecanismo físico que facilite a separação entre os fragmentos que foram ligados a este, e o restante, não ligados. Para isto, eles podem ser acoplados a uma matriz fixa contendo anticorpos secundários[4], ou a *beads* magnéticos de maneira que a extração seja possível após a centrifugação, este processo de separação é chamado de imunoprecipitação.

Após a imunoprecipitação, os anticorpos são lavados para reduzir a precipitação inespecífica e o *cross-linking* é revertido por meio de calor. Enzimas (proteinases) são introduzidas na solução para digerir proteínas e o que resta é o DNA de interesse, ou seja, os fragmentos de DNA que estavam em interação com as proteínas ou que estavam na forma Z. Esse material então pode ser submetido a alguma das técnicas de sequenciamento de alto desempenho ou a algum outro método de detecção. A figura 6 demonstra de maneira geral o fluxograma de um experimento de ChIP.

Empregando o sequenciamento de alto desempenho após o experimento de ChIP é possível acessar de fato qual a sequência de bases de cada fragmento obtido, facilitando assim as análises subsequentes relativas à localização e descrição de tais fragmentos. Os métodos de sequenciamento de alto desempenho diferem do método de sequenciamento tradicional Sanger principalmente por minimizarem as etapas de preparação da amostra a ser sequenciada e por terem um resultado que gera milhões de fragmentos de sequência se comparados às centenas produzidas pelo método Sanger. A relação custo/benefício portanto é muito maior se levarmos em consideração os métodos de alto desempenho. A figura 7 demonstra as principais diferenças entre os dois métodos, ressaltando que apesar das diferenças operacionais entre as diversas tecnologias de sequenciamento de alto desempenho disponíveis, o fluxo de trabalho, de uma maneira geral, é muito semelhante.

[4]anticorpos que se ligam a outros anticorpos. Neste caso utilizam-se anticorpos específicos contra o alvo e anticorpos secundários, associados à algum método físico, que se ligam aos primeiros facilitando a extração.

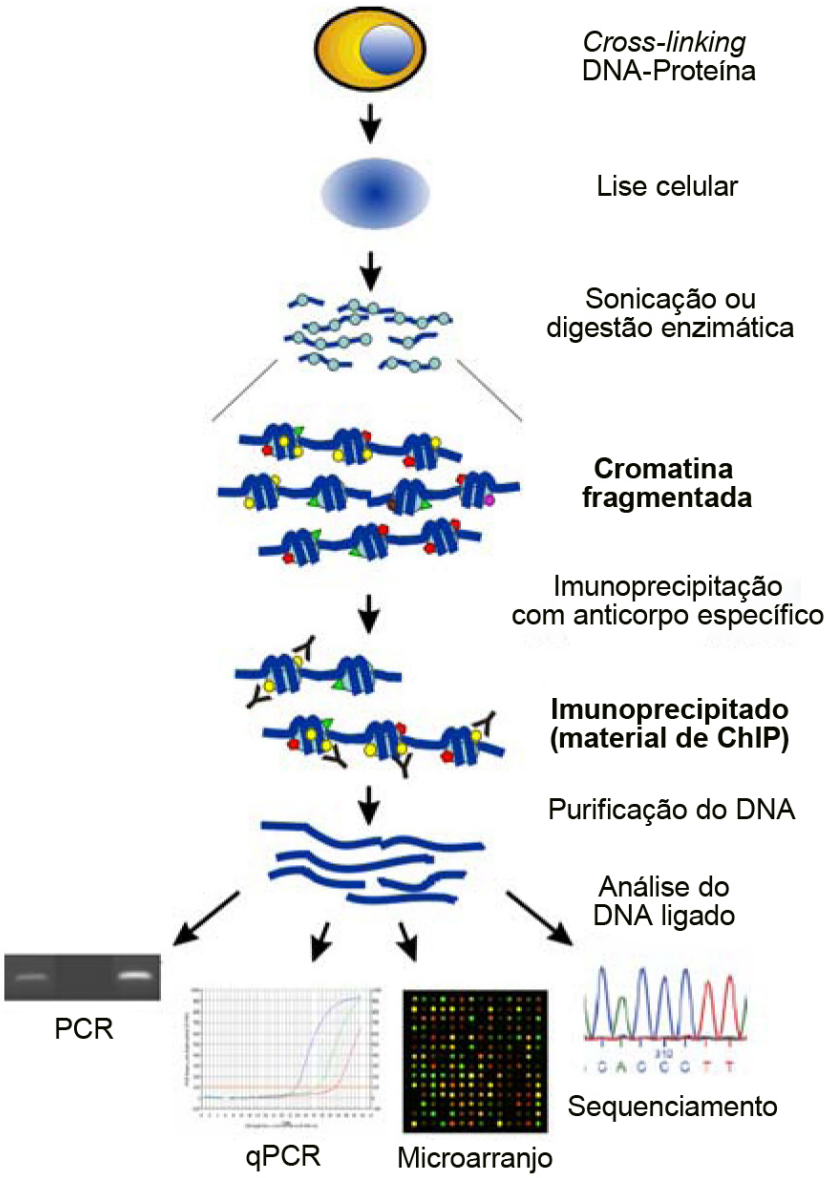

Figura 6: **Fluxograma genérico de um experimento de ChIP.** Quando o fluxo culmina com sequenciamento de alto desempenho, chamamos o experimento de ChIP-Seq (Collas, 2010, adaptado).

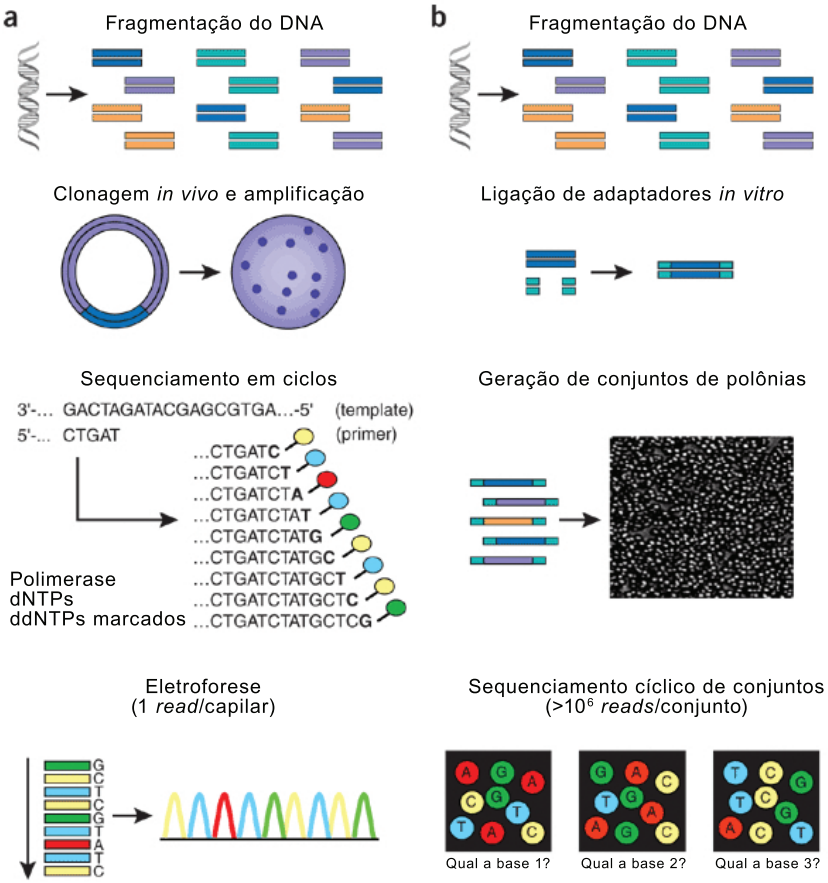

Figura 7: **Diferenças essenciais entre sequenciamento Sanger e sequenciamento de alto desempenho.** Em **(a)** podemos perceber que após a fragmentação do DNA a amostra deve passar por um laborioso processo de clonagem *in vivo* um vetor de clonagem, neste caso um vetor plasmidial. A partir de cada colônia, o DNA plasmidial é purificado e o processo de sequenciamento iniciado, ocorrendo em ciclos onde são adicionados nucleotídeos modificados marcados com sondas fluorescentes (ddNTPs) que interrompem a reação de polimerização. Estas interrupções geram diversos fragmentos de tamanhos progressivos, cuja separação e leitura são feitos através do processo de eletroforese capilar. Durante a eletroforese, o ddNTP de cada fragmento é excitado por um laser e a fluorescência é lida por um sensor capaz de interpretar as quatro cores diferentes (variação no comprimento de onda da fluorescência) das sondas, ao final são gerados os gráficos mostrados, chamados **eletroferogramas**. Já em **(b)** podemos perceber que as etapas de clonagem e amplificação não são mais necessárias, pois a ligação de adaptadores à amostra permite que essa amplificação seja feita já no próprio sequenciador. Essa amplificação gera *clusters* de amostras iguais chamadas de *PCR colonies* ou *polonies*. Em cada placa, milhões de *polonies* são formadas, o que permite que a cada ciclo de extensão seja possível detectar qual base foi anexada a várias sequências de uma vez. Para cada ciclo é obtida uma imagem fotográfica que registra a fluorescência da base adicionada, sendo estas processadas posteriormente para revelar as sequências finais (Shendure e Ji, 2008).

1.3 Pesquisas com Z-DNA no laboratório de Imunologia Molecular

O laboratório de Imunologia Molecular da Universidade de Brasília tem como grande área de interesse o estudo de anticorpos que se ligam a ácidos nucleicos. Tais pesquisas têm notável importância na elucidação de componentes que contribuem para os quadros de doenças auto-imunes. Desde de 1994, o labotório vem trabalhando com anticorpos anti-Z-DNA como modelo de interação DNA-proteína, sendo que grande parte desse trabalho focou na caracterização do anticorpo Z22 (Andrade, 1997; Andrade *et al.*, 2000; Maranhão e Brígido, 2000) que inclusive tornou-se modelo para caracterizar novas formas de anticorpos (Andrade *et al.*, 2005). Atualmente o grupo tem voltado a atenção para o papel do Z-DNA no controle da expressão gênica, o trabalho mais recente estabeleceu uma técnica de ChIP para isolamento de sequências em Z-DNA sem a necessidade de tratamento prévio (*cross-linking*). Os resultados mostraram que é possível isolar regiões em Z-DNA e corroborar previsões feitas por experimentos *in silico* (Silva, 2010).

O presente trabalho pretende colaborar com os resultados anteriores do grupo de maneira a fornecer uma ferramenta que aliada às técnicas de isolamento de Z-DNA e engenharia de anticorpos possa contribuir para comprovação e utilização do Z-DNA como possível regulador da expressão gênica. Por este motivo, para os testes do estudo de caso, foi escolhido o cromossomo 14 humano, pois neste está localizado o *locus* da cadeia pesada da imunoglobulina (IgH), que abriga os genes que codificam a maior subunidade peptídica da estrutura dos anticorpos (Tomlinson *et al.*, 1995).

2 Objetivos

2.1 Justificativa

Os métodos computacionais para análise de Z-DNA disponíveis atualmente não proveem uma grande capacidade analítica. Tais ferramentas geram resultados que necessitam de muito trabalho de pós-processamento para gerar dados interpretáveis, essa carga de trabalho pode ser facilmente diminuída com a automatização de parte dessas análises.

2.2 Objetivo Geral

- Criar um novo fluxo integrado de detecção e análise de regiões potencialmente formadoras de Z-DNA em genomas utilizando o ambiente estatístico `R` e pacotes do `Bioconductor`

2.3 Objetivos Específicos

- Possibilitar mapeamento de regiões potencialmente formadoras de Z-DNA no genoma de interesse.
- Caracterizar a distribuição de tais regiões em termos de localização e proximidade do TSS em *exons*, *introns* e junções de *splicing*.
- Analisar a correlação das potenciais localizações de Z-DNA com ocupação da RNA polimerase ou outros dados de ChIP-Seq.
- Fazer um estudo de caso no cromossomo 14 do genoma humano.

3 Materiais e Métodos

3.1 Descrição dos Equipamentos

As análises descritas neste trabalho foram realizadas em duas máquinas distintas. Para a maioria das análises, que não necessitavam de capacidade computacional elevada, foi utilizado um *desktop* simples com processador *Intel Core 2 Quad Q6600* de 2.4GHz, 4GB de memória *RAM* (DDR2-800MHz), disco rígido de 500GB (7200RPM, SATA-II) e sistema operacional *Windows 7 Ultimate 64bits*.

Para as análises que demandam maior capacidade computacional, ou para utilização de *softwares* disponíveis somente em ambiente UNIX, foi utilizado um servidor *Linux* com sistema operacional *Ubuntu Server 10.10*, 8 processadores *Intel(R) Xeon(R) CPU E5506* de 2.13GHz, 22GB de memória *RAM* e disco rígido de 300GB (7200RPM, SATA-II).

3.2 Fluxograma Analítico

Para possibilitar a análise de correlação entre as regiões potencialmente formadoras de Z-DNA (ZDRs) e elementos do genoma, um fluxograma de bioinformática foi desenvolvido utilizando o ambiente estatístico `R` e pacotes de análise do projeto `Bioconductor`. Os passos desse fluxograma, de um maneira geral, podem ser vistos na figura 8. Inicialmente, ocorre a previsão das ZDRs em toda a sequência de entrada utilizando uma versão ligeiramente modificada do programa `Z-Catcher` (Xiao, Dröge e Li, 2008). As modificações são simplesmente para possibilitar a integração do programa, escrito na linguagem `Perl`, ao ambiente `R`. Essas ZDRs tem então a sua localização confrontada com a localização dos sítios de início da transcrição dos genes, cujas anotações podem ser criadas pelo usuário ou retiradas de bancos de dados. No estudo de caso, as anotações foram retiradas do banco de dados ENCODE (*Encyclopedia of DNA Elements*) (Rosenbloom *et al.*, 2010), que faz parte do projeto *genome browser* da universidade da California Santa Cruz (Fujita *et al.*, 2010).

Para contextualizar as ZDRs em relação aos modelos gênicos, é possivel analisar a distribuição destas em termos de elementos gênicos tais como *exons*, *introns* e junções de *splicing*, assim como posicionar as ZDRs em relação aos TSSs analisados, assim elas podem ser classificadas como estando à montante (*upstream*), à jusante (*downstream*) ou no interior (*inside*) dos transcritos.

Também é possível fazer a análise de correlação das ZDRs com dados de ChIP-Seq. Estes dados geralmente consistem em milhões de *reads* geradas por sequenciadores de alto desempenho como Illumina® ou 454®. Tais *reads* são alinhadas contra sequências de referência por meio do software de alinhamento `Bowtie` (Langmead *et al.*, 2009). Após todos os devidos pré-processamentos, os dados são convertidos e condensados em estruturas específicas para que possam ter suas análises de correlação realizadas dentro do ambiente estatístico R, tais estruturas são disponibilizadas pelos pacotes de bioinformática do projeto Bioconductor (Bioconductor, 2011) e serão delhadas nas seções a seguir. Os gráficos referentes às análises foram gerados no `R` utilizando-se o pacote gráfico `ggplot2` (Wickham, 2011).

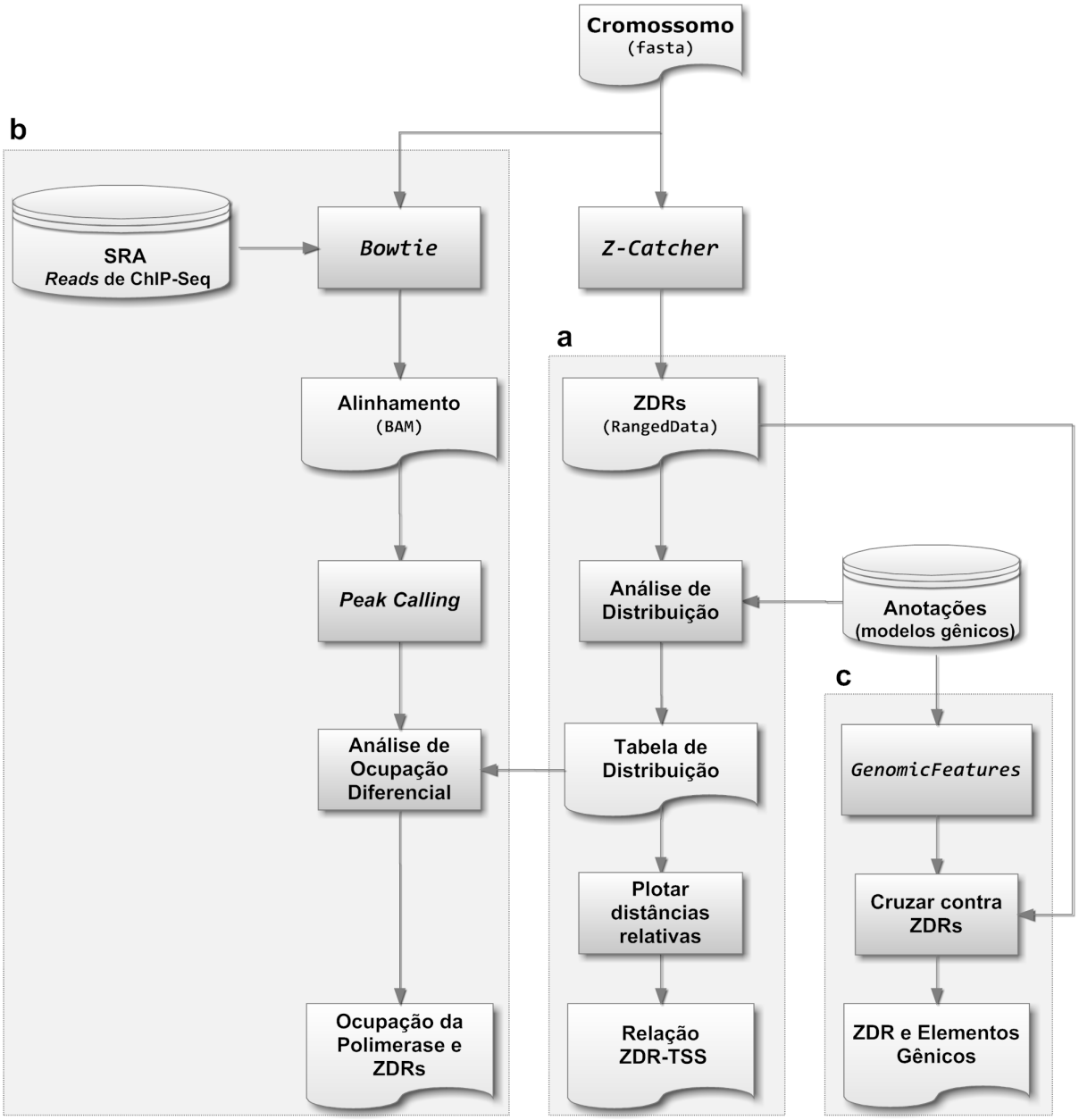

Figura 8: **Representação esquemática do fluxograma de análise.** A partir da sequência de entrada em formato `fasta` pode-se derivar uma série de análises cruzando informações entre as ZDRs preditas pelo `Z-Catcher` com informações inseridas pelo usuário como anotações gênicas ou *reads* de sequenciamento de alto de desempenho (*HTS - High Throughput Sequencing*). As caixas retangulares representam processos, as caixas com a parte inferior curvada representam dados (em formato de texto ou formatos específicos) e os cilindros representam informação retirada de banco de dados. Em **(a)**, análise das distâncias relativas aos TSS; **(b)**, análise de ocupação diferencial da RNA polimerase e **(c)** análise das distribuição de ZDRs em relação a elementos funcionais do genoma.

3.3 Dados de Referência (estudo de caso)

3.3.1 `hg19` - Genoma Humano

O genoma de referência utilizado neste trabalho foi obtido diretamente do servidor FTP do NCBI (Genome Reference Consortium, 2011), sendo que somente o cromossomo 14 foi utilizado no estudo de caso para testar a metodologia. A versão utilizada foi a última versão base lançada até o momento, chamada de `hg19/GRCh37`.

3.3.2 Anotação de Elementos Funcionais do Genoma

Para obter as anotações sobre posicionamento dos elementos gênicos, foi utilizada a base de dados **ENCODE**. Os dados foram obtidos diretamente do *site* da UCSC, especificamente na seção *Table Browser* (`http://genome.ucsc.edu/cgi-bin/hgTables?command=start`), onde é possível obter todos os dados do *genome browser* disponíveis no *site*. A figura 9 mostra uma captura de tela com os parâmetros utilizados para obter o banco, enquanto a figura 10 mostra a captura de tela de uma tabela exemplificando quais os dados presentes no ENCODE.

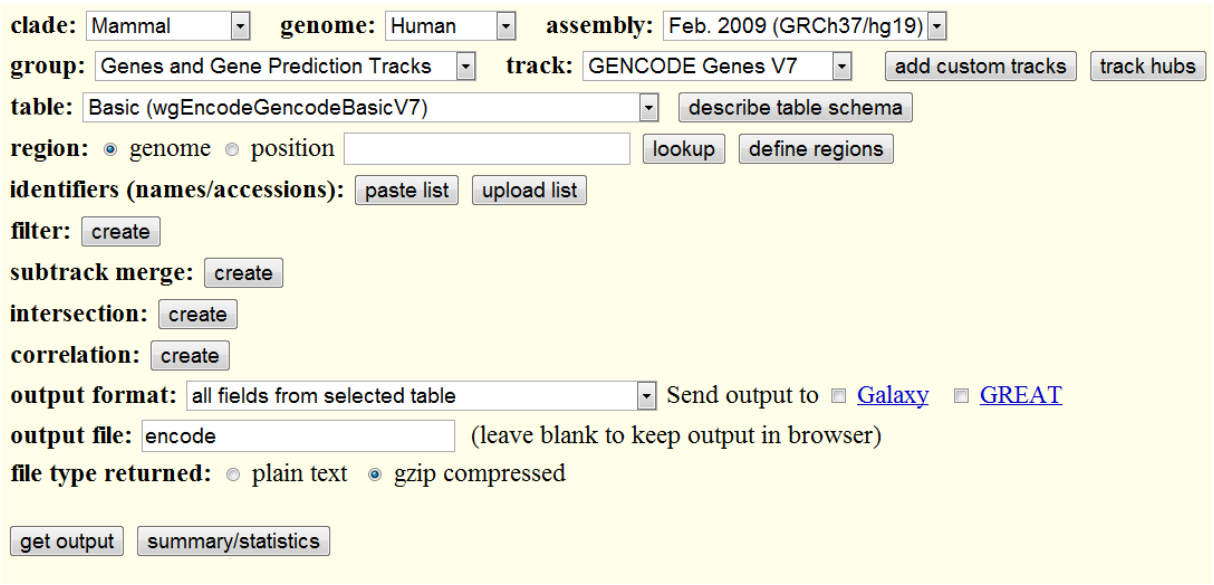

Figura 9: **Captura de tela mostrando os parâmetros para obter o banco de dados de elementos funcionais.** A versão `V7` é a mais recente. Configurando outros parâmetros, é possivel filtrar a tabela para que a saída mostre somente dados de interesse.

Schema for GENCODE Genes V7 - Gene Annotations from ENCODE/GENCODE Version 7

Database: hg19 **Primary Table:** wgEncodeGencodeBasicV7 **Row Count:** 86,046
Format description: A gene prediction with some additional info.

field	example	SQL type	info	description
bin	585	smallint(5) unsigned	range	Indexing field to speed chromosome range queries.
name	ENST00000456328.2	varchar(255)	values	Name of gene (usually transcript_id from GTF)
chrom	chr1	varchar(255)	values	Reference sequence chromosome or scaffold
strand	+	char(1)	values	+ or - for strand
txStart	11868	int(10) unsigned	range	Transcription start position
txEnd	14409	int(10) unsigned	range	Transcription end position
cdsStart	14409	int(10) unsigned	range	Coding region start
cdsEnd	14409	int(10) unsigned	range	Coding region end
exonCount	3	int(10) unsigned	range	Number of exons
exonStarts	11868,12612,13220,	longblob		Exon start positions
exonEnds	12227,12721,14409,	longblob		Exon end positions
score	0	int(11)	range	score
name2	DDX11L1	varchar(255)	values	Alternate name (e.g. gene_id from GTF)
cdsStartStat	none	enum('none', 'unk', 'incmpl', 'cmpl')	values	enum('none','unk','incmpl','cmpl')
cdsEndStat	none	enum('none', 'unk', 'incmpl', 'cmpl')	values	enum('none','unk','incmpl','cmpl')
exonFrames	-1,-1,-1,	longblob		Exon frame {0,1,2}, or -1 if no frame for exon

Figura 10: **Esquema detalhado da saída do banco de dados do ENCODE no UCSC.** O arquivo de saída é um arquivo texto simples (*plain text*) cujas colunas estão listadas no campo ***field***. O arquivo possui 86.046 linhas, cada uma correpondendo a um transcrito diferente. Os dados que compõem cada linha são mostrados no campo ***example***.O campo ***SQL type*** mostra como os dados são armazenados no banco de dados do UCSC, ***info*** mostra alguns detalhes do arquivo diretamente no *site* e ***description*** mostra uma breve descrição de cada coluna.

3.3.3 Ocupação da RNA polimerase a partir de *reads* do SRA

As *reads* de ChIP-Seq utilizadas foram escolhidas após uma extensa busca nos arquivos do SRA (*Sequence Read Archive*) do NCBI (*National Center for Biotechnology Information*) (Leinonen, Sugawara e Shumway, 2011). O objetivo era selecionar um conjunto de *reads* referente às regiões de ocupação da RNA Polimerase que tivesse sido isolado de células MCF7. Essa característica era importante pois tal linhagem celular já havia sido utilizada em estudos anteriores sobre Z-DNA conduzidos no Laboratório de Imunologia Molecular (Silva, 2010) e a descoberta de novas informações contribuiria para o desenvolvimento de trabalhos futuros.

No banco SRA, havia somente um estudo (*accession number:* GSE23701) que apresentava as condições especificadas acima. Nesse estudo foi realizada uma investigação acerca de quais parâmetros podem influenciar a seleção de sítios de ligação dos fatores de transcrição ao DNA. Para tal, os autores utilizaram o receptor de hormônio nuclear, ER-α (receptor de estrogênio), como modelo. Utilizando as técnicas de ChIP-Seq, com as sequências de fragmentos de DNA identificados pelo sequenciador de alto desempenho Illumina®, todos os sítios de ligação ao DNA deste fator foram mapeados, bem como as

marcas de cromatina e ocupação da polimerase[1]. Sucedeu-se então uma análise de correlações entre esses sítios e as regiões selecionadas tanto em situações de indução como de não-indução do fator pelo seu ligante, o estradiol (Joseph *et al.*, 2010). Para o estudo de caso do presente trabalho, foram utilizadas as *reads* referentes à ocupação da RNA polimerase nas duas situações testadas: (i) induzida, com a estimulação por estradiol e (ii) não induzida, sem estimulação. [2]

3.4 Softwares

3.4.1 `Z-Catcher`

Para se fazer a predição de sequências potencialmente formadoras de Z-DNA (ZDRs) foi utilizado o programa `Z-Catcher` (Xiao, Dröge e Li, 2008). O programa é implementado na linguagem `Perl` e utilizado via linha de comando. Sua organização consiste de *scripts* cujas implementações exibem duas maneiras distintas de funcionamento, uma específica para sequências de cromossomos ou sequências muito longas e outra genérica para outros tipos de sequências menores. Por se tratar de uma série de *scripts*, o `Z-Catcher` pode ser utilizado em qualquer sistema operacional, desde que os interpretadores `Perl` estejam instalados. O fluxograma de funcionamento deste programa pode ser visto na figura 11.

Basicamente, o programa procura, na sequência fornecida pelo usuário, por regiões cuja estrutura denota que a energia livre liberada (ΔG) em um processo de relaxamento da dupla hélice seria o suficiente para estabilizar a transição de B para Z-DNA. Primeiro, analisa-se a sequência de entrada para verificar se esta possui um perfil de alternância entre purinas e pirimidinas, visto que este é um dos requisitos para formação de Z-DNA. Se confirmado, então a sequência é percorrida em janelas de 12 nucleotídeos que são analisados de dois em dois (dinucleotídeos). Essa análise é feita assimilando um perfil *Anti-Syn* ou *Syn-Anti* para cada dinucleotídeo e então calculando o ΔG para sua estabilização, a soma dos ΔG de todos os dinucleotídeos é a energia necessária para estabilizar o processo de transição do fragmento. A partir desta energia, o valor de σ é estimado e confrontado com um valor fornecido pelo usuário (σ_0), se o valor calculado for inferior ao fornecido ($\sigma < \sigma_0$) isso indica que esta sequência não exibe potencial formador de Z-DNA neste contexto de densidade de *supercoiling*, a sequência então é descartada e a análise recomeça com os

[1] sítios do DNA onde há interação entre o ácido nucleico e histonas ou onde a RNA polimerase se acopla para iniciar o processo de transcrição

[2] A escolha do sistema estradiol-ER-α foi meramente devido à disponibilidade do estudo no SRA. Para a utilização neste trabalho, quaisquer outros sistemas de indução-repressão teriam igual valia.

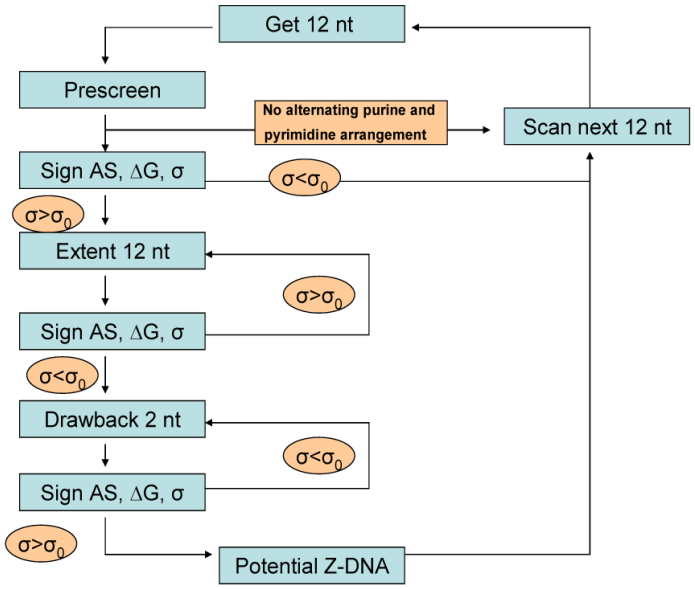

Figura 11: **Fluxograma do software `Z-Catcher`.** As janelas de 12 nucleotídeos são avaliadas em busca de perfis de alternância purina-pirimidina para assimilação de características *anti-syn* ou *syn-anti* (AS) e posteriores testes termodinâmicos que calculam a energia livre (ΔG) gerando valores de densidade de *supercoiling* (σ) a serem comparados com os limites atribuídos pelo usuário (σ_0) (Xiao, Dröge e Li, 2008).

próximos 12 nucleotídeos, caso $\sigma \geq \sigma_0$ a sequência é extendida por mais 12 nucleotídeos e o cálculo é repetido para estes. Esse processo se repete enquanto se sustentar a condição $\sigma \geq \sigma_0$. No momento em quem esta tornar-se falsa, dinucleotídeos do final deste fragmento vão sendo retirados e o cálculo refeito até que a condição volte a ser verdadeira, neste momento tenta-se adicionar um único nucleotídeo no final e caso a condição ainda se sustente este fragmento em sua totalidade é anotado como potencialmente formador de Z-DNA. Caso o último nucleotídeo invalide a condição $\sigma \geq \sigma_0$, ele é retirado e o fragmento anotado.

As implementações possuem variantes que dependem do valor de *supercoiling* a ser utilizado, a versão `lowerenergy` é designada especificamente para σ = -0.065, -0.060, -0.055, -0.050, -0.045, -0.040 ou -0.035. Para outros valores ($\sigma \leq -0.07$) deverá ser utilizada a versão padrão. A inserção dos parâmetros no programa é interativa e estes vão sendo solicitados ao usuário um por vez, devendo ser informados: o limite de *supercoiling* negativo, o cromossomo (somente no *script* específico para cromossomos), o caminho para o arquivo `fasta` a ser analisado e finalmente o nome do arquivo de saída que conterá os resultados.

Este fluxo interativo, apesar de didático, se mostra inconveniente caso haja a necessidade de integrar o `Z-Catcher` a uma sequência de programas, pois desta maneira não é

possível executá-lo com somente um comando já especificando todos os parâmetros de uma só vez, diretamente na linha de comando. Para facilitar a dinâmica e a integração com o `R`, essa característica foi ligeiramente modificada. Utilizando o módulo `Getopt::Std` da biblioteca padrão da linguagem `Perl`, os *scripts* foram alterados para possibilitar a execução a partir de apenas um comando.

3.4.2 `R` e Bioconductor

`R` (R Development Core Team, 2011) é um sistema de código livre voltado para análises estatísticas e confecção de gráficos amplamente utilizado no ambiente científico. Este provê um rico ambiente interativo, manipulável por meio da inserção de comandos em uma linha de comando, e uma linguagem de programação, que permite a criação de *scripts* e programas mais complexos. O sistema tem suas funcionalidades extensíveis por meio de um sistema de pacotes, no qual usuários podem criar métodos de análises ou procedimentos computacionais para fins específicos, que ao serem consolidados em pacotes podem ser disponibilizados para outros usuários. Sendo assim, uma gama desses pacotes, para as mais variadas aplicações, já estão inclusos na versão base do sistema, pacotes mais específicos podem ser instalados a partir de repositórios oficiais distribuídos em servidores ao redor do mundo ou em repositórios de propriedade dos desenvolvedores. Suas amplas aplicações incluem análises matemáticas e financeiras, estudo de topografias, dentre outros. Neste trabalho foi utilizada a versão de número 2.14.2. deste *software*.

Para análises biológicas, além dos pacotes já inclusos na versão base, existe um projeto chamado *Bioconductor* que agrega inúmeros pacotes adicionais ao `R`. Tais pacotes, que estendem as funcionalidades dos pacotes originais, são específicos para análises de dados biológicos que, em sua maioria, abrangem o campo da Bioinformática (Bioconductor, 2011). Os softwares, listados a seguir juntamente com suas versões, foram amplamente utilizados nas análises conduzidas neste trabalho.

3.4.2.1 `IRanges (1.12.6 - Bioconductor)`

Pacote para manipulação de intervalos, dispõe de uma série de métodos e estruturas de dados voltadas para organização e armazenamento de sequências que possam ser representadas via intervalos numéricos (Pages, Aboyoun e Lawrence, 2011).

3.4.2.2 `GenomicRanges (1.6.7 - Bioconductor)`

Especialização do pacote `IRanges` voltado especificamente para manipulação e representação de intervalos genômicos. Estruturas mais especializadas permitem o eficiente armazenamento de sequências provenientes de sequenciamento de alto desempenho ou alinhamentos contra genomas de referência (Aboyoun, Pages e Lawrence, 2011).

3.4.2.3 `ChIPpeakAnno (2.2.0 - Bioconductor)`

Pacote voltado para anotação[3] de picos[4] indentificados por experimentos de ChIP-seq ou experimentos que gerem grande quantidade de dados no formato de intervalos genômicos, como é o caso das ZDRs, anotadas por este pacote em relação aos transcritos de modelos gênicos (Zhu *et al.*, 2011).

3.4.2.4 `GenomicFeatures (1.6.8 - Bioconductor)`

Este pacote provê uma estrutura de dados chamada `TranscriptDB` que trata-se de um container para um banco de dados local do tipo `SQLite` e permite o armazenamento de anotações genômicas de maneira eficiente. Estas anotaçãoes podem, opcionalmente, ser obtidas diretamente das bases de dados da UCSC ou do BioMart (Haider *et al.*, 2009) ou criadas pelo usuário. Métodos de manipulação possibilitam a extração de elementos gênicos tais como *exons* e *introns* em formatos convenientes para integração com outros pacotes do Bioconductor (Carlson *et al.*, 2011).

3.4.2.5 `RSQLite (0.11.1 - Bioconductor)`

Interface que integra o gerenciador do sistema de banco de dados `SQLite` ao `R` permitindo que comunicação entre os pacotes do ambiente estátistico e o sistema de banco de dados. Através desta interface, o pacote `GenomicFeatures` é capaz de interagir com a estrutura interna de seu formato `TranscriptDB`, adicionando, removendo ou lendo dados (James, 2011).

[3] **anotação** no contexto de bioinformática é o processo de correlacionar uma sequência desconhecida, oriunda de algum experimento de sequenciamento ou similiar, com sua função/posicionamento no DNA de origem.

[4] regiões que, em um experimento de ChIP-seq, concentram grande quantidade de sequências identificadas, sugerindo uma região onde há interação entre a molécula investigada e o DNA.

3.4.2.6 `Rsamtools (1.6.3 - Bioconductor)`

Este pacote adiciona ao `R` a possibilidade de importar arquivos no formato `BAM`, geralmente oriundos de *softwares* externos de alinhamento, com opções de filtragem de dados e conversão de formatos. Por intermédio dele as *reads* advindas do `Bowtie` são importadas para o `R` (Morgan e Pagès, 2010).

3.4.2.7 `BayesPeak (1.6.0 - Bioconductor)`

Provê métodos estatísticos que realizam uma filtragem conhecida como *peak-calling* em dados de experimentos de ChIP-seq. Esta filtragem, comum a qualquer análise de dados oriundos desse tipo de experimento, consiste basicamente em contar o número de *hits*[5] por intervalo genômico para que seja possível identificar as áreas de enriquecimento[6] e consequentemente filtrar o ruído de fundo do experimento, ou seja, aquelas *reads* sequenciadas referentes a artefatos e impurezas, tais como fragmentos que, por limitações da técnica, continuaram presentes na amostra mesmo não tendo sido ligados aos anticorpos (Spyrou *et al.*, 2009; Cairns *et al.*, 2011).

3.4.2.8 `DESeq (1.6.1 - Bioconductor)`

Detecta expressão diferencial de genes. Utilizado em experimentos que geram grande quantidade de dados de sequenciamento, tais como RNA-Seq e ChIP-Seq, cujas amostras estão divididas por meio de tratamentos ou condições diferenciadas. A análise de expressão diferencial quantifica as *reads* reportadas para cada grupo de amostra e utiliza-as em testes estatísticos, baseados na distribuição binomial negativa, que apontam se um gene está mais ou menos expresso em uma condição que em outra (Anders e Huber, 2010).

3.4.2.9 `multicore (0.1-8 - R)`

Pacote que provê ao R maneiras de executar cálculos em paralelo em máquinas com múltiplos processadores (ou *cores*) melhorando a velocidade de análises mais pesadas computacionalmente (Urbanek, 2011).

[5] sequências oriundas de experimentos de sequenciamento que foram alinhadas com sucesso a alguma região do genoma de referência.

[6] regiões com clara concentração de *hits* em comparação às regiões adjacentes, podem indicar atividade transcricional ou locais de interação com o DNA, dependendo do tipo de experimento.

3.4.2.10 `ggplot2 (0.9.0 - R)`

Pacote para geração de gráficos baseado em um sistema chamado *grammar of graphics*. Esse sistema gera gráficos altamente personalizáveis baseados em camadas que se sobrepõem umas sobre a outras criando a figura final. Cada camada é personalizável por si só, o que permite a criação de gráficos complexos, de boa aparência e com alta capacidade informativa (Wickham, 2011).

3.4.3 *Softwares* Auxiliares

3.4.3.1 `RStudio (0.95.263 - Windows)`

IDE [7] que disponibiliza uma interface gráfica que organiza e aumenta a eficiência da utilização do `R`. Provê facilidades para edição de dados, instalação de pacotes e gerenciamento de arquivos. Disponível em `http://rstudio.org`

3.4.3.2 `bowtie (0.12.7 - Linux)`

Software de alinhamento múltiplo de sequências oriundas de plataformas de sequenciamento de alto-desempenho. Alinha sequências curtas, de até 35bp, contra genomas de referência. Emprega um algoritmo de compressão de dados conhecido como transformada Burrows-Wheeler que comprime os genomas de referência em um índice buscável, diminuindo a carga de memória utilizada no processo de alinhamento, mantendo a busca rápida e eficiente (Langmead *et al.*, 2009).

3.4.3.3 `samtools (0.1.16 - Linux)`

Conjunto de ferramentas para manipulação de arquivos de alinhamento no formato SAM/BAM. Possibilita operações básicas tais como combinar arquivos, indexá-los para otimizar a busca e ordená-los para facilitar a extração de informação (Handsaker *et al.*, 2009). Sua utilização no arquivo de saída do `bowtie` permite a conversão de SAM para BAM, sua versão binária, para que possa ser importado para o `R` pelo pacote `rsamtools`.

[7] *Integrated Development Environment* - ambiente de desenvolvimento integrado, ferramentas que facilitam o desenvolvimento de *softwares*, fornecendo ferramentas importantes para aumentar a eficiência de utilização das linguagens de programação.

3.4.3.4 SRA toolkit (2.1.2 0 - Linux)

Grupo de ferramentas para manipulação de arquivos do SRA (Leinonen, Sugawara e Shumway, 2011) (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software). Dados disponibilizados na base de dados do SRA estão comprimidos em um formato de arquivos específico chamado sra. Estes podem ser convertidos para diversos outros formatos tais como FASTA ou SAM por meio deste grupo de ferramentas.

4 Resultados

4.1 Fluxograma do Estudo de Caso

Para implementar o fluxograma mostrado na figura 8, foi feito um estudo de caso utilizando o cromossomo 14 como ambiente genômico de referência para a busca e análise das ZDRs. As modificações ao fluxograma, especificamente para aplicá-lo ao cromossomo 14, podem ser vistas na figura 12. A previsão das ZDRs foi feita utilizando a sequência genômica do cromossomo 14 como entrada e a partir deste ponto as análises subsequentes foram realizadas. Conforme já especificado, as anotações de modelos gênicos foram retiradas do ENCODE e as *reads* de ChIP-seq do repositório SRA. Nas seções seguintes serão mostrados os resultados de cada análise do fluxograma juntamente com o detalhamento dos procedimentos analíticos.

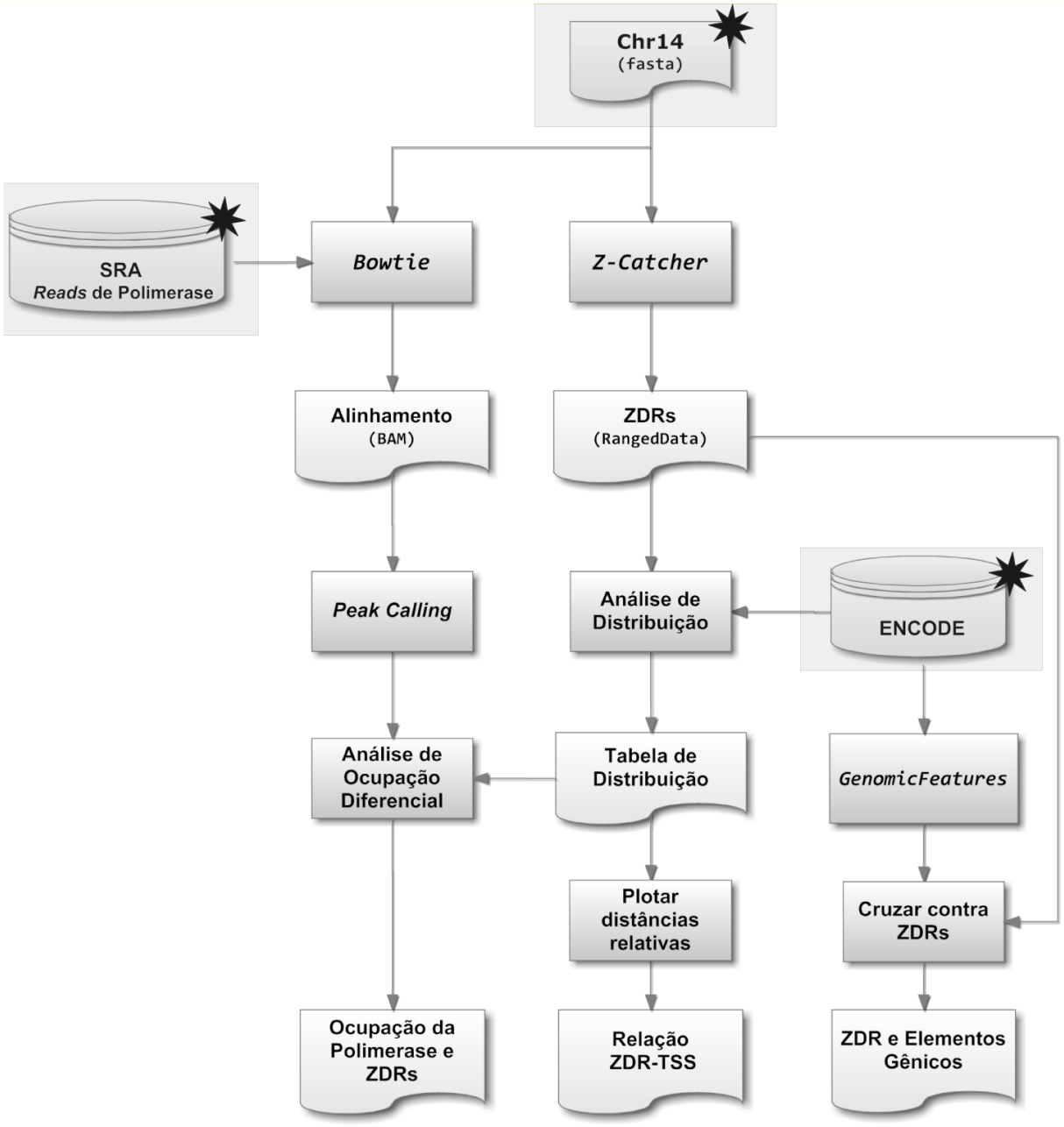

Figura 12: **Representação esquemática do fluxograma de análise aplicado ao cromossomo 14.** Caixas destacadas e com estrelas representam as entradas específicas para o estudo de caso. No centro, o arquivo `fasta` com a sequência genômica do cromossomo 14: a entrada principal para o `Z-Catcher` e o contexto onde ocorrerão as análises. À esquerda, as *reads* do SRA referentes às amostras de ChIP-seq de ocupação da RNA polimerase, e à direita, o ENCODE como referência de modelos gênicos para o genoma humano.

4.2 Etapas Preliminares

4.2.1 ZDRs

4.2.1.1 Integração com `Z-Catcher` e obtenção de ZDRs

Uma vez que os *scripts* do `Z-Catcher` foram adaptados para possibilitarem a chamada com todos os parâmetros de uma só vez (ao invés de um por vez como nos *scripts* originais), foi construída uma função do `R` capaz de fazer esta chamada e em seguida inserir os resultados diretamente em seu ambiente sob a forma de um objeto interno da linguagem.

Para o estudo de caso em particular, foi utilizado o *script* do `Z-Catcher` que possibilita busca genômica, tendo o cromossomo 14 como entrada. Em um primeiro momento, três diferentes densidades de *supercoiling* foram testadas visando avaliar a diferença exercida pelo perfil energético nas sequências. O arquivo de saída obtido diretamente do `Z-Catcher` é um arquivo de texto puro com 4 colunas separadas entre si por tabulações, como visto na figura 13.

```
#cutoff supercoiling density is -0.07
Chromosome ZDRstart ZDRlength ZDRsequence
Chr14      19022419 12        ATGTGCACGTGC
Chr14      19050701 14        GTGCGCATGTACCC
Chr14      19066182 32        GTGCGCACACACACTGGCCTGCGCCCACTGTC
Chr14      19077502 58        GTGTGTGTGTGTGTGTGAGTGTGTGTGTGTGA[...]
```

Figura 13: **Exemplo do arquivo de saída do `Z-Catcher`.** Os quatro campos correspondem ao cromossomo analisado (*Chromosome*), à localização do início da ZDR (*ZDRStart*), ao tamanho da ZDR encontrada (*ZDRlength*) e à sequência de nucleotídeos da ZDR (*ZDRsequence*)

Com os dados dentro do ambiente `R`, foram analisadas a quantidade e o tamanho médio das ZDRs obtidas para cada uma das três densidades utilizadas. Na tabela 1 podemos ver que as diferenças são marcantes. Fica claro que num cenário de menor energia (σ mais próximo de zero), as sequências preditas são mais longas, porém, bem menos frequentes. Enquanto que em cenários de maior energia (valores de σ distantes de zero) as sequências preditas são muito abundantes porém mais curtas. Um ponto principal dessa análise é a diferença nas quantidades de sequências preditas, os números sugerem uma consideração acerca da estringência do algoritmo em apontar prováveis ZDRs e consequentemente sobre a presença de falsos positivos e negativos. Para evitar os extremos (nem muito estringente e nem muito permissivo) foram utilizadas nas análises subsequentes as sequências preditas com densidade de até -0.070, uma vez que estudos

anteriores apontaram que nesta densidade observa-se ampla distribuição de ZDRs ao longo dos cromossomos em condições fisiológicas (Li *et al.*, 2009).

Tabela 1: **Quantidade e tamanho médio (em pares de base) das ZDRs preditas pelo `Z-catcher` no cromossomo 14**

	Densidade de *Supercoiling*	Número de Sequências	Tamanho Médio (pb)
1	-0.050	1786	41.48
2	-0.070	7523	34.94
3	-0.090	367636	22.90

4.2.1.2 Conversão de formatos

Para a preparação das ZDRs, assim como dos outros dados, foi utilizado o pacote `GenomicRanges`. Esse pacote disponibiliza uma estrutura de dados que facilita muito análises que necessitam manipular sequências genômicas das mais diversas maneiras. Essa estrutura, chamada `GRange`, é mostrada na figura 14.

O diferencial e grande vantagem de se utilizar as `GRanges` em relação a outras estruturas de dados é a possibilidade de manipulação dos intervalos. Na figura, podemos ver uma coluna chamada `IRanges` onde aparecem os intervalos do cromossomo em que cada ZDR ocorre. Obtendo tais intervalos de diferentes grupos de dados é possível cruzá-los utilizando a função `findOverlaps()`, disponibilizada pelo mesmo pacote, e assim obter diversas informações a respeito da co-localização das ZDRs com outros dados a serem analisados. A estrutura também tem como propriedade a fácil interconversão entre vários formatos diferentes utilizados em outros pacotes do Bioconductor, permintindo exportação de sequências individuais, ou de grupos de sequências definidas por filtros inseridos pelo usuário. Em suma, a partir dos `GRanges` das ZDRs, foi possível fazer as análises em relação ao TSS, elementos gênicos e ocupação da polimerase.

4.2.2 ENCODE

4.2.2.1 Filtragem e inserção no R

Para inserir os dados do ENCODE no R e permitir as análises de localização foi necessário filtrar a tabela original do banco para obter somente os dados relevantes. Como mostrado na figura 10, a tabela do ENCODE possui muitos dados referentes aos modelos gênicos que representa, porém, muitas dessas informações podem ser descartadas para

```
GRanges with 7523 ranges and 2 elementMetadata values:
               seqnames                ranges strand   |      size sequence
                  <Rle>             <IRanges>  <Rle>   | <integer> <factor>
     >CHR14_z1    chr14   [19022419, 19022430]      *   |        12 ATGTGCACGTGC
     >CHR14_z2    chr14   [19050701, 19050714]      *   |        14 GTGCGCATGTAC[...]
     >CHR14_z3    chr14   [19066182, 19066213]      *   |        32 GTGCGCACACAC[...]
     >CHR14_z4    chr14   [19077502, 19077559]      *   |        58 GTGTGTGTGTGT[...]
     >CHR14_z5    chr14   [19090397, 19090418]      *   |        22 GTGTGTGTGTGT[...]
     >CHR14_z6    chr14   [19152420, 19152441]      *   |        22 GCACACACACAC[...]
     >CHR14_z7    chr14   [19162840, 19162867]      *   |        28 GTGTGTGTGTGT[...]
     >CHR14_z8    chr14   [19188744, 19188759]      *   |        16 ACACACACACAC[...]
     >CHR14_z9    chr14   [19196434, 19196449]      *   |        16 GTGTGTGTGTGT[...]
           ...      ...                     ...    ... ...       ... ...
  >CHR14_z7515    chr14 [107180565, 107180592]      *   |        28 ACACACACACGC[...]
  >CHR14_z7516    chr14 [107188298, 107188393]      *   |        96 ACACACACACAC[...]
  >CHR14_z7517    chr14 [107188414, 107188433]      *   |        20 ACACACACACAC[...]
  >CHR14_z7518    chr14 [107196348, 107196379]      *   |        32 ACACACACACAC[...]
  >CHR14_z7519    chr14 [107234453, 107234476]      *   |        24 GTGCACGGGCAC[...]
  >CHR14_z7520    chr14 [107243641, 107243678]      *   |        38 GTGTGTGTGTGT[...]
  >CHR14_z7521    chr14 [107247824, 107247837]      *   |        14 GTGCGGGTGCAC[...]
  >CHR14_z7522    chr14 [107253662, 107253679]      *   |        18 ACGCGCACGTAC[...]
  >CHR14_z7523    chr14 [107284330, 107284381]      *   |        52 ACACACACACAC[...]
  ---
  seqlengths:
       chr14
   107349540
```

Figura 14: **Exemplo da estrutura de uma `GRange`.** Pode-se perceber que a saída do `Z-Catcher` está completamente contida nesse formato, apenas algumas colunas extras, próprias da estrutura, foram adicionadas.

os propósitos desse trabalho. Assim, foram selecionados para cada entrada somente o código do transcrito, o cromossomo, a fita onde se encontra e a posição de início e fim da transcrição, além de algumas informações complementares como o tamanho, o nome do gene do qual faz parte e o número de *exons*. Essas informações foram consolidadas em um `GRange` cuja estrutura é mostrada na figura 15.

As quatro primeiras colunas desse `GRange` são obrigatórias para compor a estrutura, as colunas restantes são metadados que adicionam informações extras às sequências e não são utilizadas nos processos das análises. Para o estudo de caso foram selecionados somente os transcritos referentes ao cromossomo 14.

```
GRanges with 2317 ranges and 3 elementMetadata values:
                    seqnames                ranges strand   |      size    niceName exonNumbers
                       <Rle>             <IRanges>  <Rle>   | <integer> <character>   <integer>
  ENST00000315266.5    Chr14   [66974124, 67648515]      +   |    674391        GPHN          22
  ENST00000478722.1    Chr14   [66974124, 67648520]      +   |    674396        GPHN          23
  ENST00000459628.1    Chr14   [66974855, 67525746]      +   |    550891        GPHN          11
  ENST00000543237.1    Chr14   [66975221, 67647740]      +   |    672519        GPHN          25
  ENST00000305960.9    Chr14   [66975230, 67647914]      +   |    672684        GPHN          21
  ENST00000346562.2    Chr14   [33408448, 34273382]      +   |    864934       NPAS3          11
  ENST00000341321.4    Chr14   [33408458, 34149849]      +   |    741391       NPAS3           7
  ENST00000356141.4    Chr14   [33408522, 34270315]      +   |    861793       NPAS3          12
  ENST00000357798.5    Chr14   [33408522, 34270315]      +   |    861793       NPAS3          12
                ...      ...                   ...    ... ...       ...         ...         ...
  ENST00000390630.2    Chr14 [107095125, 107095662]      -   |       537    IGHV4-61           2
  ENST00000454421.2    Chr14 [107113740, 107114274]      -   |       534    IGHV3-64           2
  ENST00000390632.2    Chr14 [107131032, 107131560]      -   |       528    IGHV3-66           2
  ENST00000390633.2    Chr14 [107169930, 107170428]      -   |       498    IGHV1-69           2
  ENST00000390634.2    Chr14 [107178819, 107179338]      -   |       519    IGHV2-70           2
  ENST00000433072.2    Chr14 [107198931, 107199471]      -   |       540    IGHV3-72           2
  ENST00000390636.2    Chr14 [107210931, 107211471]      -   |       540    IGHV3-73           2
  ENST00000424969.2    Chr14 [107218675, 107219365]      -   |       690    IGHV3-74           2
  ENST00000390639.2    Chr14 [107282791, 107283280]      -   |       489    IGHV7-81           2
  ---
  seqlengths:
       Chr14
   107349540
```

Figura 15: `GRange` **obtido do ENCODE.** Da esquerda para a direita, as colunas denotam: ID do transcrito no Ensembl, cromossomo, localização (intervalo), fita, tamanho do transcrito, nome do gene, quantidade de exons do transcrito.

4.2.3 *Reads* de ChIP-Seq da RNA polimerase

4.2.3.1 Obtenção

Para as análises de ocupação da RNA polimerase, os dois conjuntos de *reads* foram obtidos diretamente do SRA, conforme descrito na seção 3.3.3, no formato `sra` e convertidos para `fastq` através do `SRA toolkit`. As *reads* referentes aos experimentos de ChIP-Seq de células MCF7 induzidas e não-induzidas por estradiol continham respectivamente 916,3 milhões de bases e 957,3 milhões de bases.

4.2.3.2 Pré-processamento

Para que as *reads* pudessem ser utilizadas nas análises subsequentes, foi necessário determinar suas localizações no genoma. Elas foram alinhadas contra o cromossomo 14 utilizando o *software* de alinhamento `Bowtie` calibrado para retornar somente os melhores alinhamentos em um arquivo de formato `SAM`. Posteriormente o arquivo foi inserido no R por intermédio do pacote `Rsamtools` (Morgan e Pagès, 2010) e em seguida convertido em `GRanges` mantendo somente as reads com alinhamento exato.

4.3 Análises

4.3.1 Distâncias relativas aos TSSs

Detectar correlações entre ZDRs e genes é importante para auxiliar na elucidação das funções biológicas do Z-DNA. Conforme dito anteriormente, há vários indícios que ligam o Z-DNA a eventos transcricionais, assim como mostram a localização aparentemente predominante de ZDRs nas proximidades dos TSSs.

O fluxograma desenvolvido neste trabalho possui como uma das principais funcionalidades um método que facilita a localização dessas regiões nos cromossomos ou sequências de interesse. A estratégia é confrontar a localização de cada ZDR, predita pelo `Z-Catcher`, com o TSS mais próximo. Isso pode ser feito facilmente utilizando uma função do pacote `ChIPpeakAnno` (Zhu *et al.*, 2011). A função, chamada `annotatePeakInBatch`, faz os cálculos de distância entre as ZDRs e o início do elemento mais próximo (nesse caso o transcrito) resultando em uma tabela de correlações onde é possível observar exatamente a posição relativa de cada ZDR.

Para automatizar todo processo, foi criada no `R` uma função chamada `zDistr`, responsável pela análise de distribuição (Fig.8 (a), segunda caixa). Essa função aceita como parâmetros de entrada um arquivo de ZDRs oriundo da etapa de detecção no `Z-Catcher` (em `DataFrame`[1] ou `GRanges`) e um arquivo de modelos gênicos (no estudo de caso foi utilizado o ENCODE no formato `GRanges`). Então, no corpo da função, é feita uma chamada à `annotatePeakInBatch`. O resultado pode ser reportado diretamente em um gráfico ou simplesmente retornado sob forma da tabela original, caso haja a necessidade de armazenar os resultados para utilização posterior. A tabela possui muitos campos de resultados, mas os principais estão mostrados na figura 16.

peak	feature	insideFeature	distancetoFeature
>CHR14_z1	ENST00000384179.1	upstream	-97095
>CHR14_z10	ENST00000359695.2	upstream	-56051
>CHR14_z100	ENST00000315957.4	downstream	13715
>CHR14_z1000	ENST00000346562.2	upstream	-266732

Figura 16: **Principais campos da saída da função `annotatePeakInBatch` aplicada às ZDRs contra o ENCODE.** A coluna `peak` representa as ZDRs e `features` os transcritos aos quais as distâncias foram comparadas. As outras duas colunas mostram respectivamente qual a posição relativa entre a ZDR e o transcrito e qual a distância entre eles.

[1] `DataFrame` é uma estrutura de dados do ambiente `R` que consiste basicamente em uma tabela cujas linhas e colunas podem ser nomeadas e utilizadas individualmente em diversos processos e cálculos.

Os cálculos da função `annotatePeakInBatch` são feitos, por padrão, utilizando o início da ZDR contra o início do elemento (ou final caso o elemento esteja na fita negativa), mas a função aceita parâmetros que modifiquem essas características, sendo possível calcular as distância utilizando o meio ou o final de ambas. Os resultados são consolidados de forma a mostrar qual a posição relativa entre as entidades comparadas, bem como a distância de uma à outra. Por exemplo, na primeira linha da figura 16 observa-se que a ZDR entitulada `CHR14_z1` está a montante (*upstream*) do transcrito `ENST00000384179.1` a uma distância de 97.095 pares de base.

O gráfico, resultante da função `zDistr`, é gerado ao plotar essas distâncias em uma curva de frequência, assim é possível obter uma estimativa da distribuição das ZDRs em relação aos TSS. Aplicando-a ao cromossomo 14, resulta no gráfico mostrado na figura 17.

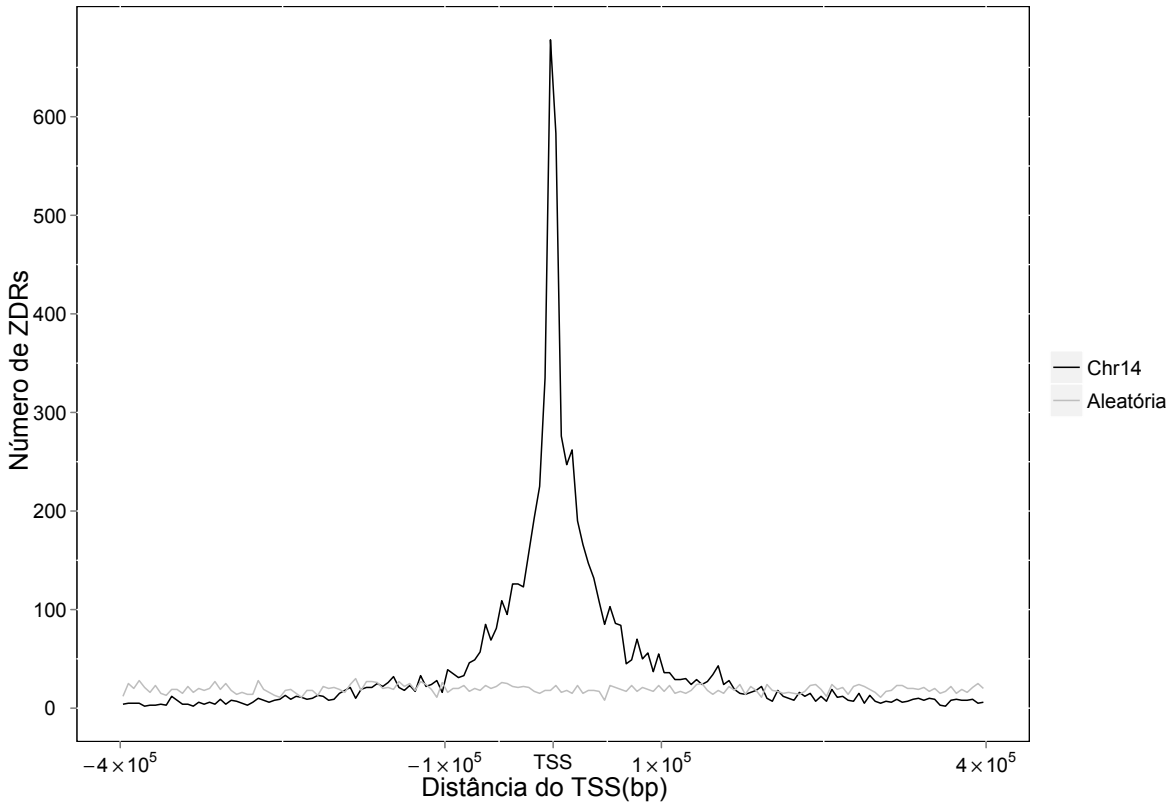

Figura 17: **Gráfico de distribuição de ZDRs ao redor de TSSs:** A linha preta representa a distribuição das ZDRs identificadas pelo `Z-Catcher` cujas localizações foram confrontadas com a localização dos TSSs de cada transcrito do ENCODE. A linha em cinza mostra a distribuição aleatória dessas distâncias em toda a extensão do cromossomo.

4.3.2 Distribuição das ZDRs em relação a elementos funcionais

Para fazer o mapeamento e correlação das ZDRs em relação aos outros elementos gênicos (*exons*, *introns* e *splice junctions*) foi utilizado o pacote `GenomicFeatures` (Carlson *et al.*, 2011). Utilizando este pacote, uma vez montadas as estruturas gênicas, várias análises podem ser feitas por meio da filtragem de elementos funcionais específicos.

Como já citado, dados de referência podem ser obtidos diretamente no `R` através das funções que se conectam aos bancos de dados *online*. Essas funções, respectivamente `makeTranscriptDbFromUCSC` e `makeTranscriptDbFromBiomart`, aceitam parâmetros que definem quais dados serão extraídos. No caso da função `makeTranscriptDbFromUCSC` existe uma função auxiliar, `supportedUCSCtables`, que lista quais as tabelas disponíveis para cada genoma cadastrado no banco. Obtendo o nome da tabela, a requisição pode ser feita e o objeto resultante é salvo no `R` em formato `TranscriptDB`.

4.3.2.1 Construção do banco de dados

Apesar da funcionalidade de obtenção automática de dados aumentar a praticidade das análises, versões mais recentes dos bancos não podem ser obtidas por meio desse método devido ao fato de não haver, no pacote, um mecanismo de sincronia com a fonte original. Por esse motivo, a versão mais recente do ENCODE, utilizada no estudo de caso, foi obtida de maneira manual, conforme já descrito na seção 9. O pacote disponibiliza uma função chamada `makeTranscriptDB` que permite construir manualmente um banco de dados no formato `TranscriptDB` a partir de dados inseridos pelo usuário. Essa função exige como parâmetros de entrada: informações sobre identificação e localização genômica dos transcritos, juntamente com cada um de seus *exons*; nome dos genes a qual esses transcritos estão associados e informações (nome e tamanho) dos cromossomos dos quais esses transcritos fazem parte.

Foi necessário escrever algumas funções no `R` para reaver esses dados através de filtragem e processamento das colunas contidas na tabela do ENCODE. Um fluxograma do processo é mostrado na figura 18.

Para as informações sobre identificação e localização dos transcritos, foram selecionados da tabela e armazenados em um `DataFrame` (*transcripts*): o cromossomo do qual o transcrito faz parte, a fita onde se localiza, começo e fim de sua sequência e o nome (ID do Ensembl). Um segundo `DataFrame` (*splicing*) foi criado para conter ordem e posicionamento de cada *exon* para cada um dos transcritos da tabela. Primeiro, a função

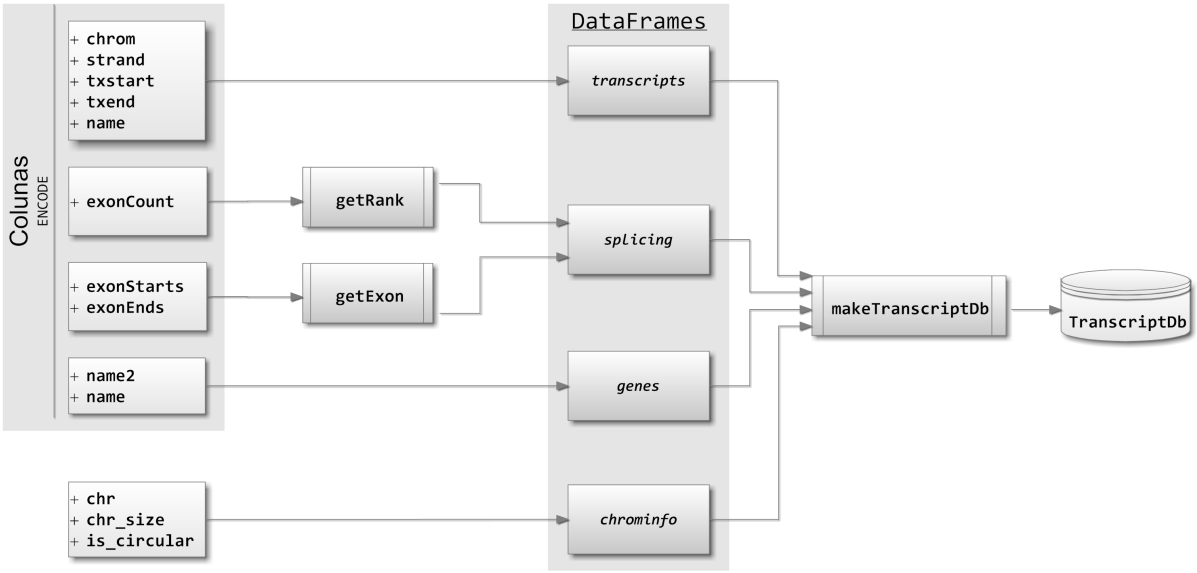

Figura 18: **Fluxograma para criação do banco de dados em formato `TranscriptDb`.** os dados iniciais são retirados da tabela original do ENCODE (superior esquerdo, fundo escurecido) ou, no caso das informações sobre os cromossomos, de dados da literatura. As funções `getRank` e `getExon` processam parte dos dados. Estes são consolidados em quatro `DataFrames` que servem de entrada para a função `makeTranscriptDb` que, por fim, gera um banco de dados no formato `TranscriptDb`

`getRank` utiliza a coluna `exonCount`, que informa a quantidade de *exons* presentes em cada transcrito, para criar uma lista ordenando e numerando cada um deles (*e.g.* se o primeiro transcrito possuísse três *exons* e o segundo quatro, a lista seria: `1,2,3,1,2,3,4`). Em seguida, a função `getExon` faz a varredura das colunas `exonStarts` e `exonEnds`, que possuem respectivamente posições de início e fim de cada *exon*, associando cada início ao fim correspondente, consolidando desta maneira, a localização individual dos *exons* dentro do transcrito. Por fim, os *exons* já separados foram associados às suas posições de acordo com a ordem gerada pela função `getRank`.

Outros dois `DataFrames` foram criados para conter informações sobre os genes dos quais cada transcrito faz parte e descrever os cromossomos. O primeiro (*genes*) é formado pela associação da coluna `name2`, que contém o nome dos genes, à coluna `name`, que contém o nome do transcrito (ID no Ensembl). O segundo (*chrominfo*) é formado pelo nome dos cromossomos do genoma humano e seus tamanhos, juntamente com uma variável *booleana*[2] `is_circular` indicando se o cromossomo é circular ou não.

Ao fim do processo de consolidação, cada um dos `DataFrames` foi utilizado como argumentos para a função `makeTranscriptDB`, gerando então o banco de dados em formato

[2]variável formada somente por valores binários: verdadeiro ou falso. Indica simplesmente se alguma condição está presente ou não.

`TranscriptDB` para ser manipulado por meio das outras funções disponibilizadas pelo pacote.

4.3.2.2 Separação dos elementos gênicos

A separação dos transcritos em *exons* e *introns* foi efetuada utilizando as funções `exonsBy` e `intronsByTranscript`, também disponibilizadas pelo pacote `GenomicFeatures`. Ambas recebem como argumento um banco em `TranscriptDB` e geram uma saída em um formato chamado `GRangesList`, que consiste em uma lista onde cada elemento é um `GRange`, representando, neste caso todos os *exons* ou *introns* de cada transcrito.

4.3.2.3 Intersecção com ZDRs

Para o estudo de caso, a separação dos elementos gênicos foi efetuada para o cromossomo 14. As ZDRs no formato `GRanges`, obtidas pelo processo descrito anteriormente, foram filtradas de modo a selecionar somente aquelas localizadas exclusivamente no interior dos transcritos. Então, a função `findOverlaps`, do pacote `GenomicRanges`, foi aplicada para calcular as possíveis intersecções entre elas e os elementos gênicos. O resultado desse cálculo consiste em uma tabela de correlação com duas colunas, ambas são preenchidas pelos índices dos elementos intersectados entre si, por exemplo, se alguma porção do transcrito **1** se intersecta com a ZDR **4**, na tabela irá constar | **1** | **4** |. Devido a este resultado ser estritamente numérico, o pacote disponibiliza uma função cuja finalidade é recuperar exatamente os transcritos onde foram encontradas intersecções, sendo assim, passando um objeto contendo o resultado da intersecção para a função `queryHits` obtém-se uma nova `GRangesList` listando um subconjunto dos transcritos cujos elementos intersectam com ZDRs. Para clarificar os resultados, esse processo foi feito separadamente para *exons* e *introns*, e depois foi contada a quantidade de intersecções únicas, indicando qual a fração das ZDRs contidas em cada elemento gênico. A distribuição das ZDRs em todo o cromossomo 14, levando em consideração a posição relativa aos TSSs pode ser vista na figura 19. As ZDRs que foram classificadas como *inside* foram subdivididas em *exons*, *introns* e *splicing junctions*.

4.3.3 Ocupação diferencial da RNA polimerase

As relações entre Z-DNA e processos de transcrição (Liu e Wang, 1987) e a proximidade em relação aos TSS (Xiao, Dröge e Li, 2008) levantou a hipótese sobre a possibilidade

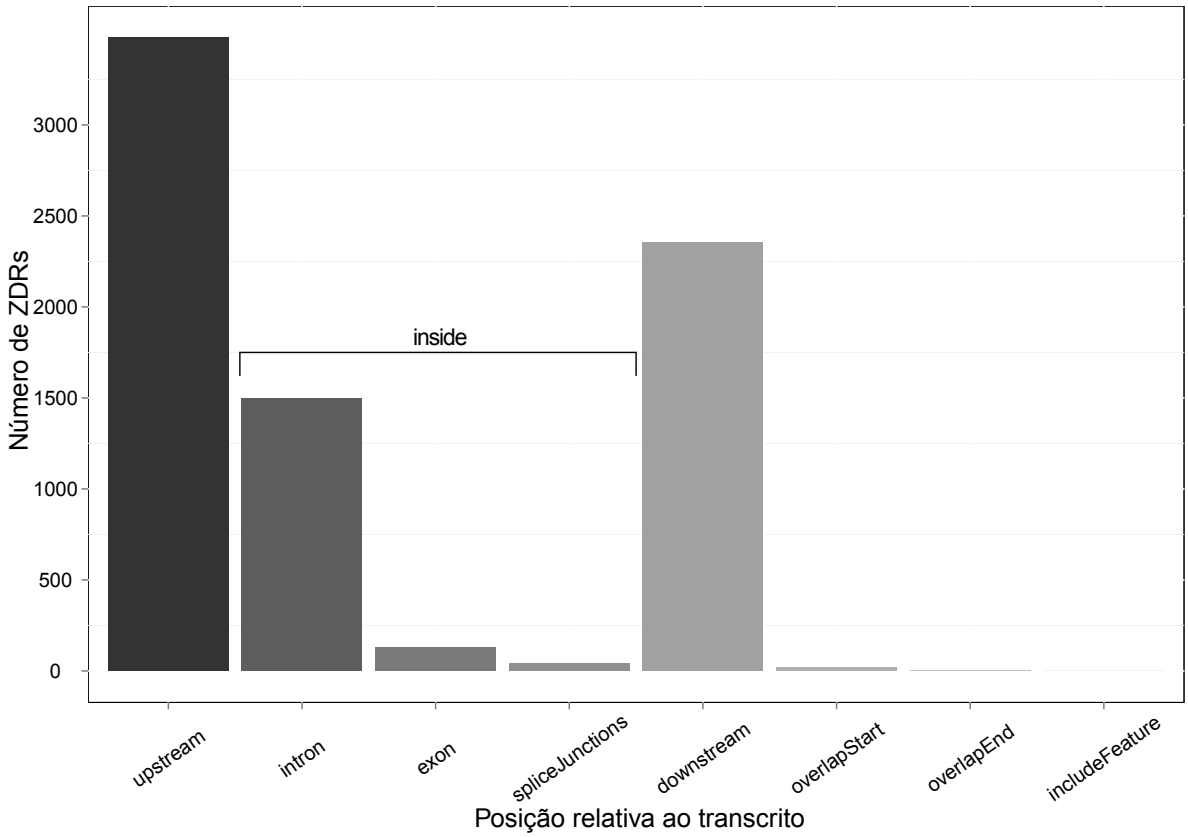

Figura 19: **Localização relativa das ZDRs em função dos transcritos:** De um total de 7.523 ZDRs, grande parte concentra-se a montante (*upstream*) e a jusante (*downstream*) dos transcritos, constituindo respectivamente ∼46% (3.476 ZDRs) e ∼31% (2.357 ZDRs) do total. A outra grande parte, aproximadamente 22% (1.667 ZDRs), é formada por ZDRs localizadas no interior dos transcritos (*inside*). Estas são mostradas subdivididas em termos de elementos gênicos, e pode-se perceber que a grande maioria concentra-se nas regiões intrônicas (∼90% das localizadas *inside*) enquanto que somente ∼8% se encontra nos *exons* e aproximadamente 2% nas junções de *splicing*. As outras localizações que representam respectivamente, a sobreposição com o início e fim do transcrito (*overlapStart* e *overlapEnd*) e transcritos contidos no interior de ZDRs (*includeFeature*), somam menos de 1% do total.

dessas ZDRs influenciarem de alguma maneira a ocupação da RNA polimerase nas regiões próximas ao TSS. Para investigar a validade dessa hipótese, foi feita para o estudo de caso uma análise de correlação entre as ZDRs e as *reads* de ChIP-Seq que demonstraram enriquecimento diferenciado, entre os casos induzido e não-induzido com estradiol, visando encontrar algum tipo de relação causa-efeito que pudesse corroborar ou não a hipótese. Nesses termos, enriquecimento diferenciado significa que ao alinhar as *reads* ao genoma, na mesma região do cromossomo, encontram-se presentes para cada caso (induzido ou não), uma maior ou menor quantidade de *reads* alinhadas. Isto indica que durante o experimento, nessa região, a atividade da RNA polimerase foi modulada pela diferente condição de indução. A hipótese tem por fim investigar se há algum padrão de distribuição

dessas *reads* que indique a participação de regiões formadoras de Z-DNA modulando a ocupação da RNA polimerase.

Por se tratar de *reads* de ChIP-Seq os dados originais devem primeiro passar pelo processo de *peak-calling*, realizado pelo pacote `BayesPeak`. Os detalhes deste processo são explicados a seguir.

4.3.3.1 *Peak Calling*

As *reads* pré-processadas pelo procedimento mencionado na seção 4.2.3.2 foram convertidas de `GRange` para `RangedData` para que pudessem ser utilizadas pela função `bayespeak` do pacote homônimo. Esse formato faz parte do pacote `IRanges` (Pages, Aboyoun e Lawrence, 2011) e é muito semelhante ao `GRange`, porém mais genérico, podendo tratar outros tipos de dados com intervalos que não sejam necessariamente genômicos. Por exemplo, no caso de um `RangedData`, informações sobre a fita de DNA localizam-se na coluna de metadados, por se tratarem de informação não essencial para caracterizar o conjunto. A função `bayespeak` possibilita a utilização de múltiplos processadores, devido ao fato dos cálculos estatísticos de *peak calling* serem muito exigentes computacionalmente. Para utilizar essa opção foi necessário carregar o pacote `multicore` (Urbanek, 2011). A função então foi aplicada às *reads* utilizando 8 processadores para realizar a tarefa.

4.3.3.2 Expressão diferencial

A análise de expressão diferencial baseia-se primeiramente na contagem de *reads* que se sobrepõem às ZDRs (quantidade de *hits*), o primeiro passo é fazer essa contagem por meio da função `countOverlaps` (pacote `IRanges`), e consolidar esses dados em uma matriz, cujas primeiras linhas são mostradas na tabela 2.

Tabela 2: **Primeiras linhas da matriz de contagem de sobreposições.** Os números representam a quantidade de *reads* de ChIP-Seq da RNA polimerase que se sobrepõem à ZDR indicada nos dois conjuntos de dados.

ZDR	Número de Reads	
	controle	estradiol
CHR14_z1	28	15
CHR14_z2	12	12
CHR14_z3	38	43
CHR14_z4	17	23
CHR14_z5	6	1
CHR14_z10	11	12

Esta matriz então é dada como argumento para a função do pacote `DESeq` chamada `newCountDataSet` que converte a tabela de contagem para um formato próprio, utilizado pelo pacote para fazer suas análises internas, chamado `CountDataSet`. Os dados então passam pelo processo de estimação de parâmetros através das funções `estimateSizeFactors` e `estimateDispersions`. Este processo seria dispensável para o conjunto de dados deste trabalho devido a ausência de replicatas biológicas, pois neste caso não há como estimar a dispersão da expressão pelos dados e isso é feito empiricamente pelo algoritmo, porém o processo é exigido para que se possa usar as outras funções do pacote. Após estimados, os dados são finalmente usados como entrada para a função `nbinomTest` que aplica um teste que usa a distribuição binomial negativa[3] para definir a diferença de enriquecimento entre as *reads* dos dois grupos de dados. Uma amostra do resultado dessa função, antes de ser ordenada e processada, é mostrada na figura 20.

id	baseMean	baseMeanA	baseMeanB	foldChange	log2FoldChange	pval	padj
CHR14_z1	21.5	28	15	0.5357143	-0.9004643	0.3665566	1
CHR14_z2	12.0	12	12	1.0000000	0.0000000	1.0000000	1
CHR14_z3	40.5	38	43	1.1315789	0.1783372	0.8364666	1
CHR14_z4	20.0	17	23	1.3529412	0.4360991	0.6991440	1
CHR14_z5	3.5	6	1	0.1666667	-2.5849625	0.4438057	1
CHR14_z10	11.5	11	12	1.0909091	0.1255309	1.0000000	1

Figura 20: **Saída da função `nbinomTest`.** as colunas denotam respectivamente a identificação da ZDR, a média entre a contagem de *reads*, o número de *reads* no grupo controle, número de *reads* no grupo tratado com estradiol, o enriquecimento de um grupo em relação a outro, log2 desse enriquecimento, o p-value da distribuição e o p-value ajustado para taxa de falsos positivos.

Para recuperar as ZDRs que apresentaram maior diferença entre a quantidade de *reads* em cada grupo, a tabela foi filtrada de maneira a separar aquelas cujo `foldChange` era maior ou igual a **2**, representando as regiões com enriquecimento, ou *upregulated*, e aquelas com `foldChange` menor que **0,5**, representando as que não tiveram enriquecimento, ou *downregulated*. Em seguida, os IDs de cada ZDR foram cruzados com os IDs da tabela de distribuição das ZDRs contra o ENCODE, afim de correlacionar o enriquecimento das regiões com o posicionamento das mesmas em relação aos transcritos. Os resultados da aplicação deste processo ao cromossomo 14 e suas posições podem ser vistos na figura 21.

[3]distribuição utilizada quando se observa um conjunto de dados composto por contagem de valores que demonstrem grande dispersão (Cameron e Trivedi, 1998, p. 71). Neste contexto utiliza a comparação entre a média e a dispersão biológica, representada pela variância.

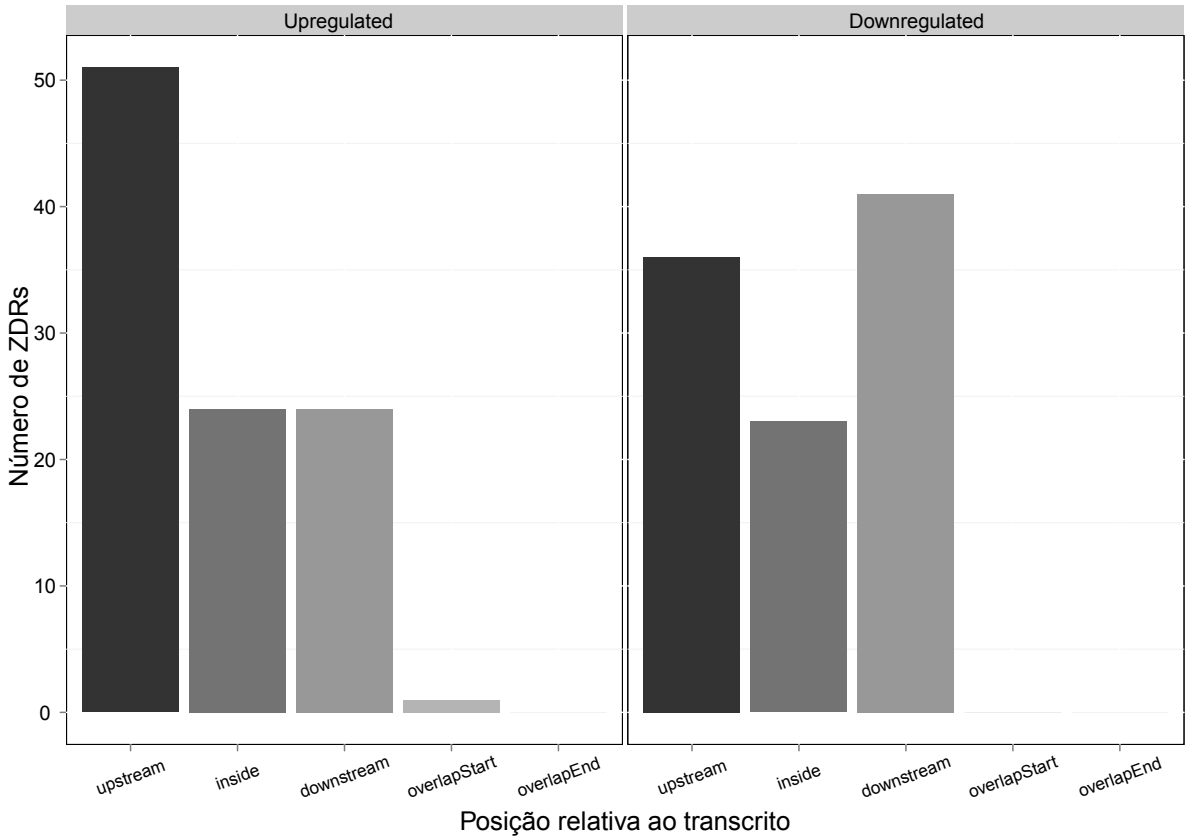

Figura 21: **Localização relativa de ZDRs correlacionadas com reads da RNA polimerase.** Cada painel representa as *reads* associadas a ZDRs que apresentaram maior ou menor enriquecimento (respectivamente *up* e *downregulated*). As barras mostram a quantidade de ZDRs relacionadas às *reads* posicionadas em relação a cada transcrito. É possível observar uma tendência para as ZDRs relacionadas com enriquecimento de *reads* (*upregulated*) estarem posicionadas a montante (*upstream*) dos transcritos, enquanto que uma tendência oposta, embora fraca, é observada no caso do enriquecimento negativo (*downregulated*).

5 Discussão e Conclusões

O fluxograma apresentado neste trabalho teve como foco principal a utilização de ferramentas científicas gratuitas associadas de uma maneira que facilitasse a análise de sequências potencialmente formadoras de Z-DNA. Em primeira instância, a metodologia foi testada com o cromossomo 14 humano afim de validá-la assim como contribuir com informações inovadoras.

A figura 17 mostra que a primeira fase do fluxograma, na qual buscava-se a distribuição das ZDRs em relação aos TSSs, mostrou resultados satisfatórios, sendo que o gráfico apresentado em que as ZDRs concentram-se ao redor dos TSSs de fato replica os resultados da literatura, principalmente aqueles de Xiao *et al*, 2008. Apesar destes resultados já serem bastante conhecidos neste campo de estudo, não se conhecia métodos automatizados para se chegar a ele. O fluxograma apresenta uma metodologia que pode facilitar a investigação de novos organismos, nos quais a distribuição de ZDRs não é conhecida.

Já na figura 19 é possível observar resultados inovadores. Apesar dos resultados de Li *et al*, 2009 terem mostrado *hotspots* de Z-DNA ao longo do genoma, que haviam sido detectados experimentalmente utilizando o motivo $Z\alpha_{ADAR1}$ como sonda, nenhum outro trabalho mostrou a distribuição destas em relação a todas as outras regiões adjacentes aos transcritos. Mais uma vez, os resultados corroboram a afirmação da literatura de que as ZDRs estão correlacionadas à transcrição gênica, visto que aproximadamente 50% das regiões analisadas no cromossomo 14 estão localizadas a montante dos TSSs. Este mesmo resultado mostra também que dos 22% das ZDRs localizadas no interior de transcritos, praticamente 90% destas estão em *introns*, porém esta taxa está dentro do esperado para a distribuição aleatória dado que os *introns* deste cromossomo são em média 24 vezes maiores [1] que os *exons* e portanto a probabilidade de uma sequência no interior de um transcrito fazer parte de um *intron* é significativamente maior. Ainda assim, o fluxograma apresenta uma maneira eficaz de localizar as ZDRs distribuídas nas adjacências dos trans-

[1] tamanhos médios: *introns* = 6319,5 pb e *exons* = 257,4 pb

critos, assim como localizá-las em relação aos elementos gênicos, caso estas estejam no interior de transcritos.

Os resultados da investigação entre a relação das taxas de ocupação da RNA polimerase com a presença de ZDRs são mostrados na figura 21. Pode-se perceber que em se tratando do grupo de genes com expressão estimulada (*upregulated*), mediante à presença de estradiol, a quantidade de ZDRs relacionadas às *reads* da RNA polimerase está concentrada à montante (*upstream*) dos TSSs, enquanto que no grupo com expressão diminuída (*downregulated*) há uma fraca tendência para o agrupamento à jusante (*downstream*). Isto poderia sugerir uma possível correlação positiva entre essas regiões e o estado de ativação de genes durante a transcrição, uma vez que a formação de ZDRs à montante pode favorecer este processo. No caso do acúmulo à jusante, não é possível afirmar categoricamente nenhum tipo de correlação pois as diferenças são pouco significativas, portanto, um tratamento estatístico adequado somado à experimentos biológicos para validação seriam necessários para comprovar ambas as afirmações.

Uma questão crucial no estudo de Z-DNA é o paradoxo entre a sua função e formação. Diversos estudos apontaram a correlação entre o processo transcricional e a formação do Z-DNA. Sabe-se que as forças de *supercoiling* negativo e energias associadas, provenientes da passagem do aparato transcricional, são essenciais para a estabilização das transições de B para Z-DNA, porém, alguns estudos de função também apontam que a sua formação pode ser essencial para o controle da transcrição, como mostrado nos experimentos com os genes c-MYC e CSF-I (Wittig *et al.*, 1992; Wölfl, Wittig e Rich, 1995; Liu *et al.*, 2001). Assim, a possível correlação aqui apresentada das ZDRs com a ocupação da polimerase pode trazer um novo grupo de informações para somar ao que já é conhecido, mostrando que possivelmente, o Z-DNA esteja envolvido na manutenção do equilíbrio termodinâmico transcricional, onde sua presença auxilie o processo de transcrição contrabalanceando as forças de torção. Alguns dados porém, já sugeriram que em *E.coli* a presença de sequências formadoras de Z-DNA em plasmídeos pode impedir a passagem do aparato transcricional (Peck, Wang *et al.*, 1985), embora o fato tenha sido observado somente nos experimentos *in vitro*. Outros experimentos com a RNA polimerase do bacteriófago T7 também exibiram as propriedades inibitórias do Z-DNA, neste caso não impedindo totalmente a transcrição porém diminuindo sua atividade (Dröge e Pohl, 1991). As prováveis explicações para este efeito podem estar novamente associadas ao panorama energético, visto que quanto maior a densidade de *supercoiling* do fragmento, maior é o efeito sobre o bloqueio da transcrição, de maneira que o excesso de *supercoiling* negativo tende a favorecer a formação de Z ao invés de B-DNA. Este cenário tornaria o

processo de movimentação do aparato transcricional sobre a região em Z energeticamente desfavorável, e por isso o bloqueio (Ditlevson *et al.*, 2008). Levando em consideração esses cenários divergentes, é possível perceber que o DNA na forma Z é de natureza extremamente dinâmica, apresentando uma infinidade de possibilidades funcionais que ainda devem ser estudadas a fundo considerando diversos panoramas energéticos, que por si só podem alterar a função desta estrutura.

Por fim, este trabalho mostrou com sucesso que a automatização de métodos de análise para regiões potencialmente formadoras de Z-DNA pode facilitar a consolidação das informações obtidas por outros métodos já disponíveis, permitindo assim uma rápida interpretação de resultados *in silico* para otimizar futuros experimentos biológicos de bancada.

6 Perspectivas

- Gerar um pacote de `R` e disponibilizá-lo como parte dos pacotes distribuídos pelo `Bioconductor`, permitindo que cada análise mostrada seja feita de forma individual com a possibilidade da inserção de testes estatísticos pertinentes a cada tipo de projeto.
- Reconsiderar parte das análises apresentadas para o estudo de caso, tanto em outros cromossomos como em genomas de outros organismos, aplicando-se testes estatísticos adequados.
- Refatorar o algoritmo do `Z-Catcher`, reimplementando-o em uma linguagem mais eficiente no intuito de melhorar o tempo das análises e interface com o usuário.
- Entrar em contato com os desenvolvedores do método SIBZ, na tentativa de obter o código fonte da ferramenta afim de integrá-la à metodologia como uma alternativa à busca de ZDRs apresentada pelo `Z-Catcher`.

Um artigo científico, fruto deste trabalho, entitulado *A Bioconductor based workflow for Z-DNA region detection and biological inference* será apresentado oralmente no *Brazilian Symposium of Bioinformatics* (BSB2012, Campo Grande-MS, `http://bsb2012.facom.ufms.br/`) e será publicado pela editora *Springer Verlag* como parte da série *Lecture Notes in Bioinformatics*. Este artigo encontra-se no Anexo A deste trabalho.

Os códigos-fonte de `R`, juntamente com instruções básicas de utilização serão disponibilizados no repositório online em `https://github.com/Lianzinho/z-analyst`.

APÊNDICE A – Cálculos Termodinâmicos utilizados pelo Z-Catcher

Para a realização das análises de Z-DNA, o `Z-Catcher` faz uso de cálculos termodinâmicos aplicados aos dinucleotídeos que permitem a estimativa da densidade de *supercoiling* (σ) de cada sequência analisada. Tal densidade é obtida a partir da energia livre necessária para a realização da transição de B para Z-DNA (ΔG_m). Esta energia é proveniente do relaxamento do *supercoiling* negativo da dupla hélice no contexto de um plasmídio virtual (da mesma maneira como feito para o `Z-Hunt`), e pode ser obtida pela equação

$$\Delta G_m = 10RNT[(\sigma + \frac{1.8m}{N})^2 - \sigma^2] \tag{A.1}$$

onde σ é a densidade de *supercoiling*, T é temperatura, N é o tamanho do plasmídio e m é o numero de pares de base envolvidos na transição. Essa equação A.1 deriva da equação geral de relaxamento de *supercoiling* (Frank-Kamenetskii e Vologodskii, 1984):

$$\Delta E = AN[(\sigma + \frac{km}{N})^2 - \sigma^2] \tag{A.2}$$

Assim, como a sequência é colocada em um plasmídio virtual (pBR322, 4263pb) para ser analisada, os parâmetros são definidos como $A = 10RT$, $k = 1.8$ e $N = 4263$ (Frank-Kamenetskii e Vologodskii, 1984). Para que a energia do relaxamento seja suficiente para compensar o consumo da transição de B para Z-DNA e a estabilização desta estrutura, a seguinte condição deve ser verdadeira:

$$2F_j^* + m\Delta F_{BZ} + \Delta G_m = 0 \tag{A.3}$$

onde F_j^* é a energia livre necessária para a formação da junção B-Z e ΔF_{BZ} é a energia necessária para estabilização dos nucleotídeos já na forma de Z-DNA. Esses valores fixos

foram estimados a partir de vários estudos termodinâmicos e foram consolidados por Ho *et al.*, 1986. A tabela 3 mostra os valores.

Tabela 3: **Energias de transição B para Z-DNA.** As conformações AS e SA indicam se os dinucleotídeos estão respectivamente *anti-syn* ou *syn-anti* na junção B-Z. As colunas correspondentes às siglas em parênteses $(AS)^1$ e $(SA)^1$ indicam os valores necessários para a estabilização de nucleotídeos já na forma Z-DNA, em relação aos dinucleotídeos vizinhos (Ho *et al.*, 1986 apud Xiao; Dröge; Li, 2008, adaptado).

Dinucleotídeo	Conformação			
	AS	SA	$(AS)^1$	$(SA)^1$
CG	0.7	4	4	4
GC	4	0.7	4	4
CA	1.3	4.6	4.5	4.5
AC	4.6	1.3	4.5	4.5
TG	1.3	4.6	4.5	4.5
GT	4.6	1.3	4.5	4.5
TA	2.5	5.9	5.6	5.6
AT	5.9	2.5	5.6	5.6
CC	2.4	2.4	4	4
GG	2.4	2.4	4	4
CT	3.4	3.4	6.3	6.3
TC	3.4	3.4	6.3	6.3
GA	3.4	3.4	6.3	6.3
AG	3.4	3.4	6.3	6.3
AA	3.9	3.9	7.4	7.4
TT	3.9	3.9	7.4	7.4

Levando em consideração estes cálculos, é possível chegar ao valor de σ (equivalente à energia liberada pelo relaxamento do *supercoiling*) utilizando as equações A.1 e A.3, e assim comparar com o valor inserido pelo usuário a cada iteração do algoritmo, determinando assim o potencial de formação de Z-DNA da sequência sendo analisada.

ANEXO A – Artigo Científico - Brazilian Symposium of Bioinformatics, Agosto de 2012 - Campo Grande-MS

A Bioconductor based workflow for Z-DNA region detection and biological inference

Halian Vilela[1], Tainá Raiol[1], Andrea Queiroz Maranhão[1], Maria Emília Walter[2], and Marcelo M. Brígido[1]

[1] Department of Cellular Biology, Institute of Biology, University of Brasilia, 70910-900, Brasília, DF, Brazil (brigido@unb.br)
[2] Department of Computer Science, Institute of Exact Sciences, University of Brasilia, 70910-900, Brasília, DF, Brazil

Abstract. Z-DNA is an alternative conformation of the DNA molecule implied in regulation of gene expression. However, the exact role of this structure in cell metabolism is not yet fully understood. Here we present a novel Z-DNA analysis workflow using the R software environment which aims to investigate Z-DNA forming regions (ZDRs) throughout the genome. It combines thermodynamic analysis of the well-known software `Z-Catcher` with biological data manipulation capabilities of several `Bioconductor` packages. We employed our methodology in the human chromosome 14 as a case study. With that, we established a correlation of ZDRs with transcription start sites (TSSs) which is in agreement with previous reports. In addition, our workflow was able to show that ZDRs which are positioned inside genes tend to occur in intronic sequences rather than exonic and that ZDRs upstream to TSSs may have a positive correlation with the up-regulation of RNA polymerase activity.

Keywords: Z-DNA, ZDR, Z-Catcher, R, Bioconductor

1 Introduction

The knowledge of the distribution of Z-DNA forming regions (ZDRs) throughout a genome is a valuable resource that helps to elucidate the role of this alternative DNA form in biological systems. Several evidences indicate that these regions may account for some level of gene expression regulation [6, 17]. Although, it remains a challenge to determine Z-DNA genomic distribution and the regulatory networks involved in Z-DNA formation at ZDRs. Xiao *et al.* [22] developed a Z-DNA map for the human genome by searching whole chromosomes for ZDRs and locating them in relation to the transcription start sites (TSSs) of all the annotated gene models available at that moment. Although it was shown that Z-DNA has an uneven distribution along the genome, their data lack consistent biological evidence implicating Z-DNA in gene expression. Many questions remain still opened, which makes room for further investigation.

In this study, we propose a workflow for a deep investigation of the Z-DNA distribution and the possible interaction with RNA polymerase activity. The

search for Z-DNA using our proposed workflow was performed along the human chromosome 14 as a case study, due to our interest in immunology research and background in antibody engineering. The immunoglobulin heavy chain locus (IgH) is located at this chromosome, which harbours genes that code for the larger polypeptide subunit of the antibody molecule [21]. In addition to a potential way to modulate the IgH genes, the understanding of how Z-DNA modulates gene expression may provide researchers with useful strategies for genetic engineering and biopharmaceutic production, therefore contributing to broaden the toolset of this research field.

1.1 The structure of Z-DNA

Z-DNA is an alternative conformation of the canonical DNA molecule (B-DNA). It is generally formed by varying length stretches of alternating purine-pyrimidine nucleotides exposed to conditions of high supercoiling. Which are usually formed right after the passage of the RNA polymerase during transcription [16, 12]. Its structure differs dramatically from the canonical form, displaying a helix which turns to the opposite direction (left) and a backbone which forms a zig-zag, hence the name "Z". This zig-zag backbone is formed due to the alternating *syn-anti* conformation of the nucleotides. The *syn* conformation, which occurs only in Z-DNA induces changes towards the phosphate bond causing the DNA backbone, formed by such bonds, to look fairly sinuous. Figure 1 summarizes the main differences among these two conformations. Due to the unique form of Z-DNA, it exhibits some notorious properties like antigenicity and binding specificity to some proteins like ADAR1[1] and the yeast protein zuotin [19].

1.2 Z-DNA prediction softwares: Z-Catcher and Z-Hunt

Detecting the occurrence of Z-DNA throughout the genome was first addressed by pure biophysical conformation studies [17]. After years of research some knowledge was acquired and it was possible to create prediction methods by which Z-DNA structure could be evaluated. `Z-Catcher` [22] and `Z-Hunt` [8] are computer softwares which were developed from two of such prediction methods. Both of them focus on thermodynamic analysis of an input sequence. Basically, they search for patterns of alternating purine-pyrimidine and apply free energy calculations to assess whether a given sequence is likely to change from canonical to Z conformation (B-to-Z).

Both methods try to classify potential Z forming sequences, `Z-Hunt` uses a comparative measurement to do so. It outputs sequences associated with a `z-score`, which represents the number of random base pairs that must be scanned, on average, to find a sequence with equal or better Z-forming likelihood relative to the sequence at issue [8]. On the other hand, `Z-Catcher` calculates the free energy based only on sequence analysis, taking into account a user defined threshold for a parameter called superhelical density, expressed by the greek letter σ (sigma). This parameter represents the free energy needed to

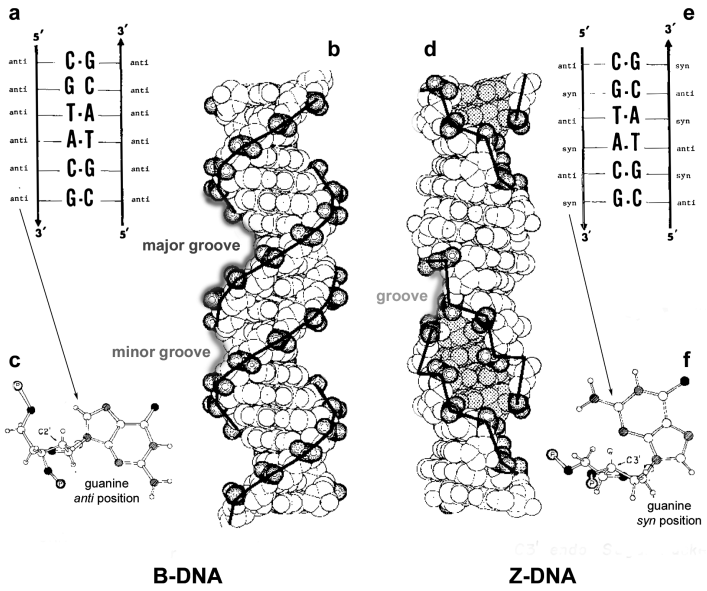

Fig. 1. Main differences between B and Z-DNA: The helix and its structural changes can be seen in **b** and **d**. *Anti* and *syn* conformations are shown respectively in **a** and **e** and their structures are detailed in **c** and **f**. [16, adapted]

meet the energetic requirement of a sequence to perform the B-to-Z transition. Higher values of σ means that the free energy released from the relaxation of the DNA helix would easily stabilize several sequences with different Z-DNA forming potentials, while lower values tend to be more strict, meaning that the released free energy would stabilize fewer sequences. The user defines a σ value which will then be compared in each iteration with the transition required value. This process will be repeated while it stays within the threshold limits. When the limit is finally reached, the algorithm stops and the sequence is annotated as a potential ZDR.

1.3 R and Bioconductor

`R` is an open source software environment and programming language for statistical computing which is widely used in the scientific community. It is suited to many different tasks, such as financial, mathematical, geoprocessing and weather studies [15]. For instance, a project called `Bioconductor` [3], which is a repository of `R` add-ons packages, was created specifically for biological purposes. Most of these packages provide bioinformatic capabilities and facilitate manipulation and conversion of different formats and data.

To store and represent almost all biological sequences in this work, we used a data structure called `RangedData`, which is part of the `Bioconductor` package

`IRanges` [13]. This structure represents biological sequences mainly by storing start-end ranges as well as information such as names, spaces (e.g. chromosomes) and other miscellaneous user-defined descriptions. The advantage of using this structure is that many other `Bioconductor` packages provide methods for comparison, trimming, flanking and sequence analysing using this format. `ChIPpeakAnno` is one example and was extensively used in this work. Its essential functionality is to find the relation between user supplied sequences and annotation data. Using its capabilities, our workflow is able to investigate the distribution of the overall distances of Z-DNA forming regions (ZDRs) relative to the nearest gene transcription start site (TSS).

All plots presented in this work were made with `ggplot2` graphics package, which is a highly customizable plotting system based on "the grammar of graphics". This system builds graphics based on separate layers overlaping each other to create a final picture, being each layer totally customizable, it is possible to create complex, nice-looking and very informative graphics.

2 Workflow overview

Our workflow aims to provide a simple manner to analyse data allowing comparison to previous studies and bringing new information about the relationship between ZDRs and the genome. The main advantage of our approach is the direct visualization of results, eliminating the need to condense and interpret data using other softwares. Figure 2 depicts a general view of our workflow design.

The search for ZDRs is made by a slightly modified version of `Z-Catcher`, which was adapted to receive parameter input directly from the command line instead of the interactive input method of the original version. This version is integrated into R where a function may be called receiving as arguments a `fasta` file containing the sequences to be investigated and the density threshold. The function then manages all details and format converting, returning the result as an R internal object.

2.1 ZDR distribution analysis

Once ZDR data are available, one may build a profile of ZDRs relative to TSSs. ZDR data and an annotation file from ENCODE [18] are used as input for an `R` function that uses the `ChiPpeakAnno` package internally to match ZDRs and TSSs. The matches result in a distribution table of relative distances and positions which is then used to investigate where most ZDRs lie within the genome. These results will be discussed in more detail in the case study section.

2.2 ZDRs and differential RNA polymerase occupation data

In order to expand our analysis of ZDR positioning, we attempted to correlate it with ChIP-seq data regarding the RNA Polymerase II (PolII) occupation sites near the TSS. Our intent was to investigate whether ZDR sites can somehow

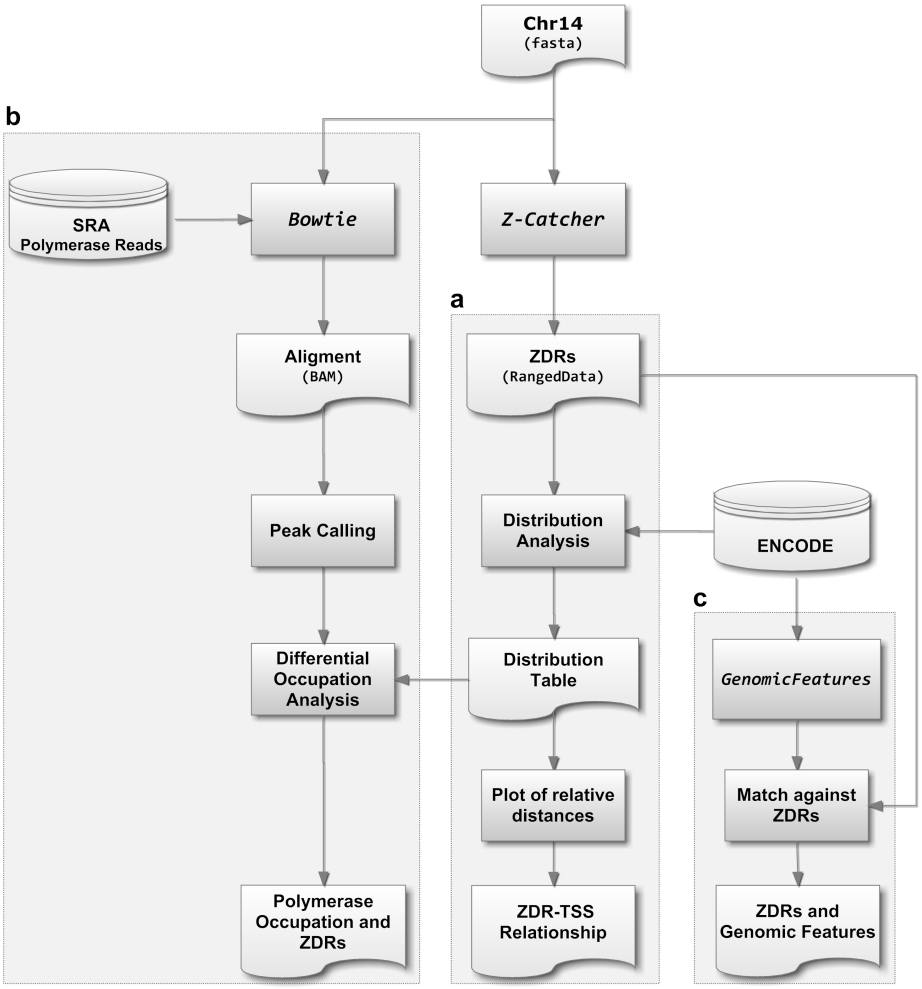

Fig. 2. ZDR analysis workflow: rectangles represent functions or processes, curved-bottom rectangles represent general data and cylinders represent database entries. Italic text represents software names or `R` packages and parenthesis represents file formats. Chromosome 14 `fasta` sequences are submitted to `Z-Catcher` to search for ZDRs. They are then used in the **distribution analysis (a)** by cross-matching with transcript TSS' positions from ENCODE, which results in a distribution table regarding relative distances between each ZDR and its nearest TSS. Plotting these distances gives the ZDR-TSS relationship for Chromosome 14. The same distribution table is used to search for correlations with **RNA Polymerase II differential occupation (b)** tags that were obtained by peak calling aligned reads from the SRA to Chromosome 14. At last, ENCODE transcripts are divided into **genomic features (c)** (exons, introns and splice junctions) and matched against ZDRs to investigate how they are distributed inside genes.

interact with PolII and therefore change the occupation rates of the enzyme. This analysis was carried out using public data retrieved from the NCBI Short Reads Archive (SRA) [11], particularly from Joseph *et al.* [9]. Reads derived from Illumina platform were divided in two groups: (i) cells treated with estradiol and (ii) untreated cells (control). Both groups were then aligned to the human chromosome 14 by the `Bowtie` aligner [10] with default parameters, except for `--best` which reports only the best alignments (those with fewer mismatches). The aligned reads were filtered out by a process called "peak calling" which identifies genome enriched areas where reads cluster together [14]. This process was performed in R by the `BayesPeak` package [4] and the differential occupation fold-change analysis was calculated by the `DESeq` package. [2]

2.3 ZDRs and genomic features

To further investigate the distribution of ZDRs in relation to genomic structure features, we searched for ZDRs within exons, introns, UTRs and intergenic regions. Although this analysis used the same `ENCODE` annotation file, it differed by being processed through the `GenomicFeatures` [5] suite so that the genes' annotation would be subdivided into their features. Then, those features were matched against the ZDRs to return an overall percentage of ZDRs relative to each genomic feature.

3 Case study: human chromosome 14

3.1 Genomic feature analysis

As stated above, the first step of our analysis was to correlate ZDRs found by `Z-Catcher` with the gene annotation from `ENCODE`. Figure 3 shows that our approach was able to reproduce literature findings by showing an overall clustering of ZDRs around TSSs [22], as opposed to randomized distances to each TSS. Once the ZDRs are not equally spread over the genome, this distribution suggests that the ZDRs may play a role in [17] or be dependent upon transcription events [12]. To further address this issue, we looked deeper into the distribution and plotted the exact location of ZDRs relative to their nearest TSS and its respective transcript. With this analysis, we wanted to investigate if there was any bias towards specific ZDR hotspots around or within transcripts. Indeed it is possible to observe in Figure 4 that ZDRs seem to be more concentrated upstream of TSSs, which would corroborate the hypothesis on the Z-DNA relationship with transcription events.

To date, no Z-DNA mapping approach has focused on the ZDR distribution throughout the genome in relation to its genomic features. Taking that into account, it is important to further investigate this correlation, since it may reveal some unknown distribution pattern and may also help to elucidate Z-DNA function. Some works had suggested that the presence of ZDRs within introns would enable and guide the coupling of proteins from the ADAR1 family, which

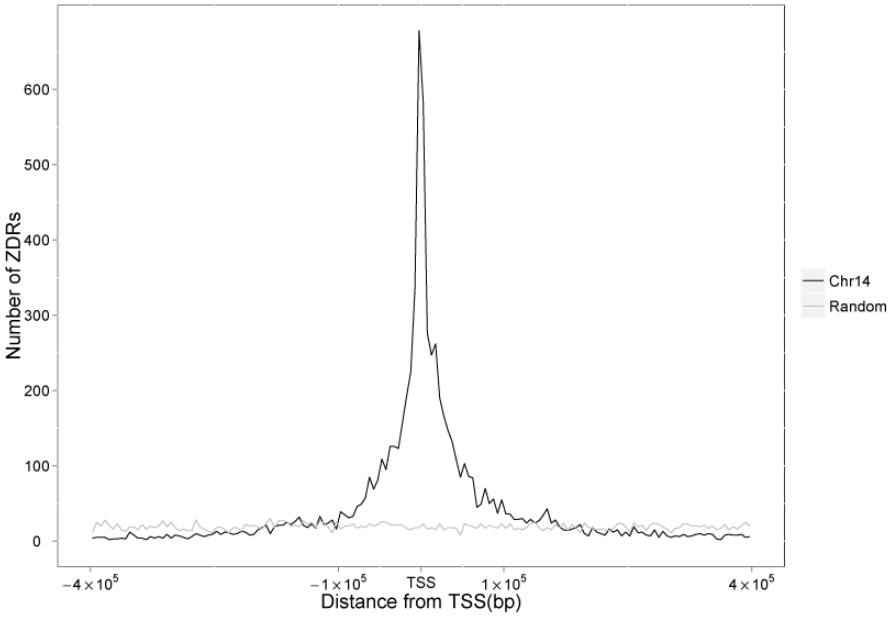

Fig. 3. Distribution of ZDRs throughout human chromosome 14: black lines show ZDR clusters around the TSS in contrast to random quasi-uniform distibution depicted in light grey.

are responsible for mRNA editing [1]. These proteins are not only known to be present in Z-DNA binding sites with high affinity but also to be responsible for the deamination of adenosines to inosines (which are translated as guanines). These editing events act as a source of phenotypic variation [7] and could play an important role at modulation of the nervous system [20]. If it is found that this interaction is dependent on Z-DNA formation, an important function for ZDRs would be revealed.

ZDRs found inside transcripts lie almost exclusively within introns, which account for roughly 18% of the total number of detected ZDRs (Figure 4). Considering that genes are composed mostly by intronic sequences, this percentage may not represent a strict preference of the ZDRs' distribution. Anyhow, in Table 1 it is possible to see five of the transcripts and their associated genes, which exhibit the largest number of ZDRs within introns. Such genes may be good candidates for further investigation of ADAR1 family mechanism of action, which would contribute to understand the potential role of Z-DNA guiding RNA editing enzymes.

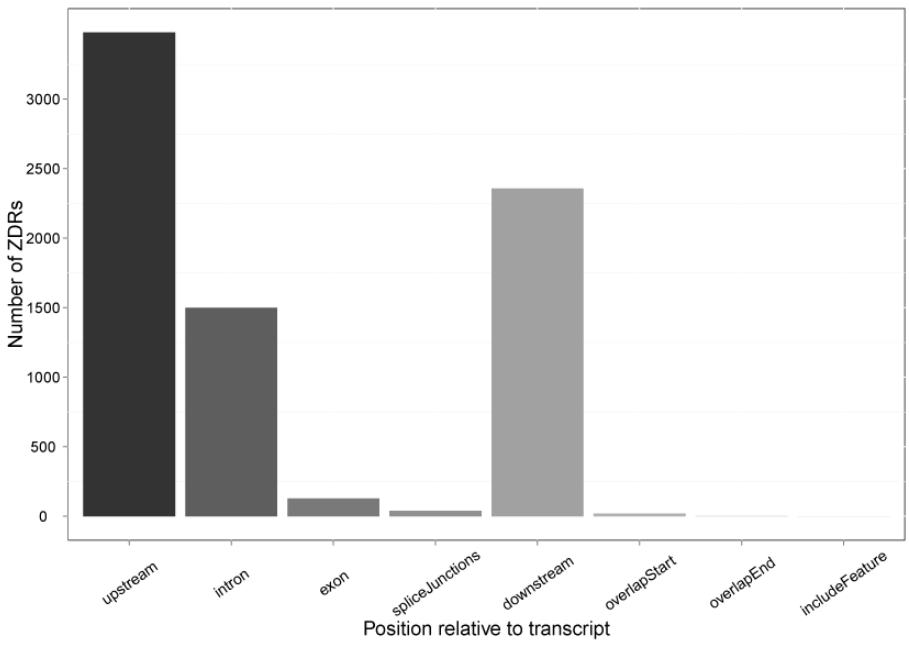

Fig. 4. Exact location of ZDRs in relation to transcripts: `includeFeature` means that a ZDR is larger enough do embrace the whole transcript, a rare situation. `intron`, `exon` and `spliceJunctions` represents ZDRs which fall within transcripts, where `spliceJunctions` represents those ZDRs shared both by introns and exons.

Table 1. Chromosome 14 transcripts with the largest number of inside-intron ZDRs.

transcript	gene	# of ZDRs
ENST00000330071.6	NRXN3	151
ENST00000332068.8	NRXN3	149
ENST00000335750.5	NRXN3	125
ENST00000488612.1	RAD51L1	82
ENST00000346562.2	NPAS3	71

3.2 Differential occupation analysis

Even though we were able to show the correlation between ZDRs and TSSs, it still remained unclear if it represents only a by-product of gene transcription events or if ZDRs indeed act as gene expression regulators. To address this point, we analysed whether ZDR presence could modify RNA Polymerase II occupation of transcription start sites.

Our dataset of PolII was taken from a ChIP-Seq experiment which analysed the occupation of the promoter region of the ER-α estrogen receptor of MCF-7 cells in two conditions: activated (with presence of its ligand, estradiol) and inac-

tivated (controlled set). With this analysis, we were able to investigate whether the differential enrichment of the PolII tags in specific locations may correlate to the presence of ZDRs.

We divided our fold-change results in two groups, one with transcripts with up-regulated occupation in the activated condition (fold-change ≥ 2) and the other with transcripts with down-regulated occupation (fold-change < 0.5). Next, we ranked the 100 topmost differentially occupied transcripts of each group and analysed their ZDRs related position. Figure 5 shows that the number of ZDRs overlapped by PolII tags at the 100 topmost differentially occupied transcripts exhibits an interesting pattern. Those ZDRs present in up-regulated genes tend to gather upstream of TSSs while those ZDRs present in down-regulated genes exhibit a weak tendency to gather downstream. This may suggest a positive correlation between ZDRs position and the activation state of genes, considering that the formation of ZDRs upstream could favour gene transcription, while ZDRs downstream could inhibit it. Nevertheless, our data need additional experimental evidence and further statistical analysis to confirm these statements.

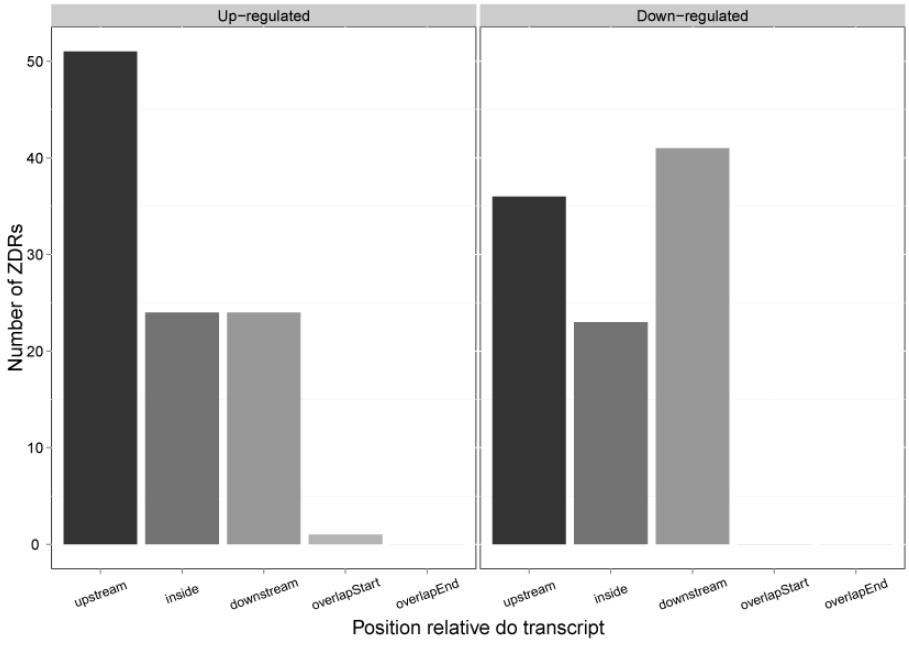

Fig. 5. Number of ZDRs overlapped by PolII tags at the 100 topmost differentially occupied transcripts: note that ZDRs tend to gather upstream of TSSs in up-regulated regions and downstream in down-regulated ones.

4 Conclusions

In this work we developed a workflow for analysis of ZDR regions in animal cells that merges *in silico* data with experimental ones. The workflow uses several bioconductor packages and retrieves biological data from high throughput sequencing. Hence, one could easily correlate data from ChIP-seq and RNA-seq to ZDR regions in whole chromosomes. Our case study focused on the human chromosome 14, and the results showed that our workflow approach was able to conduct ZDR distribution analysis that corroborates previous studies. It brought as well new information on how those ZDRs spread over the chromosomal sequences.

The role of Z-DNA in gene regulation has been debated for a long time. In our case study we showed that the majority of ZDRs appear upstream of the transcripts. We also showed that when accounted for internal genomic features, ZDRs tend to concentrate in introns rather than exons. Although this was expected, it showed that our approach is able to succesfully detect ZDRs' distribution within transcripts. Hence, one could investigate in other human chromosomes or another species genome the hypothesis of ZDRs serving as anchor sites for Z-DNA binding factors such as ADAR1, which is responsible for RNA edition.

The comparison of ZDRs prediction to PolII occupancy in steroid regulated genes suggests differences in ZDRs positioning in relation to TSSs. Up regulated genes seem to concentrate ZDRs upstream of TSSs as opposed to down regulated genes that tends to concentrate ZDRs downstream. However, further experimental studies and statistical investigation are still necessary to convincingly correlate ZDRs to gene expression.

We are presently working to assemble the `R` scripts developed for this workflow in a user friendly `R` package, where the user will be able to perform similar analysis as those previously shown. Our goal is to deliver an easy and fast way to perform basic distribution analysis associated to biological information in different kinds of genomes, allowing for an efficient computing platform for the Z-DNA biology researcher.

References

1. A. Herbert, K. Lowenhaupt, J.S., Rich, A.: Chicken double-stranded rna adenosine deaminase has apparent specificity for z-dna. Proc Natl Acad Sci USA 92(16), 7550 – 7554 (1995)
2. Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome Biology 11, R106 (2010), `http://www.bioconductor.org/packages/release/bioc/html/DESeq.html`
3. Bioconductor: Open Source Software for Bioinformatics (2011), `http://www.bioconductor.org/`
4. Cairns, J., Spyrou, C., Stark, R., Smith, M.L., Lynch, A.G., Tavaré, S.: BayesPeakan R package for analysing ChIP-seq data. Bioinformatics 27(5), 713–714 (2011)

5. Carlson, M., Pages, H., Aboyoun, P., Falcon, S., Morgan, M., Sarkar, D., Lawrence, M.: GenomicFeatures: Tools for making and manipulating transcript centric annotations (2011), http://www.bioconductor.org/packages/release/bioc/html/GenomicFeatures.html
6. Herbert, a., Rich, A.: The biology of left-handed Z-DNA. The Journal of biological chemistry 271(20), 11595–8 (1996)
7. Herbert, A.: Rna editing, introns and evolution. Trends in Genetics 12(1), 6–9 (1996)
8. Ho, P.S., Ellison, M.J., Quigley, G.J., Rich, A.: A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. The EMBO journal 5(10), 2737–44 (1986)
9. Joseph, R., Orlov, Y.L., Huss, M., Sun, W., Kong, S.L., Ukil, L., Pan, Y.F., Li, G., Lim, M., Thomsen, J.S., Ruan, Y., Clarke, N.D., Prabhakar, S., Cheung, E., Liu, E.T.: Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. Molecular systems biology 6(456), 456 (2010)
10. Langmead, B., Trapnell, C., Pop, M., Salzberg, S.: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10(3), R25 (2009), http://bowtie-bio.sourceforge.net/index.shtml
11. Leinonen, R., Sugawara, H., Shumway, M.: The sequence read archive. Nucleic acids research 39(Database issue), D19–21 (2011), http://www.ncbi.nlm.nih.gov/sra
12. Liu, L.F.: Supercoiling of the DNA Template during Transcription. Proceedings of the National Academy of Sciences 84(20), 7024–7027 (1987)
13. Pages, H., Aboyoun, P., Lawrence, M.: IRanges: Infrastructure for manipulating intervals on sequences, http://www.bioconductor.org/packages/release/bioc/html/IRanges.html, r package version 1.12.6
14. Pepke, S., Wold, B., Mortazavi, A.: Computation for chip-seq and rna-seq studies. Nature Methods pp. S22–S32
15. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2011), http://www.R-project.org, ISBN 3-900051-07-0
16. Rich, A., Nordheim, A., Wang, A.H.: The chemistry and biology of left-handed Z-DNA. Ann. Rev. Biochem. 53, 791–846 (1984)
17. Rich, A., Zhang, S.: Z-DNA : the long road to biological function. Nature reviews. Genetics 4(July), 566–573 (2003)
18. Rosenbloom, K.R., Dreszer, T.R., Pheasant, M., Barber, G.P., Meyer, L.R., Pohl, A., Raney, B.J., Wang, T., Hinrichs, A.S., Zweig, A.S., Fujita, P.a., Learned, K., Rhead, B., Smith, K.E., Kuhn, R.M., Karolchik, D., Haussler, D., Kent, W.J.: ENCODE whole-genome data in the UCSC Genome Browser. Nucleic acids research 38(Database issue), D620–5 (2010)
19. S. Zhang, C. Lockshin, A.H.E.W., Rich, A.: Zuotin, a putative z-dna binding protein in saccharomyces cerevisiae. EMBO J. 11(10), 3787 3796 (1992)
20. Sommer, B., Khler, M., Sprengel, R., Seeburg, P.H.: Rna editing in brain controls a determinant of ion flow in glutamate-gated channels. Cell 67(1), 11 – 19 (1991)
21. Tomlinson, I.M., Cook, G.P., Walter, G., Carter, N.P., Riethman, H., Buluwela, L., Rabbitts, T.H., Winter, G.: A complete map of the human immunoglobulin vh locus. Annals of the New York Academy of Sciences 764(1), 43–46 (1995)
22. Xiao, J., Dröge, P., Li, J.: Detecting Z-DNA Forming Regions in the Human Genome. International Conference on Genome Informatics 2008 (2008)

Referências

ABOYOUN, P.; PAGES, H.; LAWRENCE, M. **GenomicRanges: Representation and manipulation of genomic intervals**. [S.l.], 2011. Disponível em: <http://www-.bioconductor.org/packages/release/bioc/html/GenomicRanges.html>.

ALBERTS, B. *et al.* **Molecular biology of the cell**. 5. ed. New York: Garland science, Taylor & Francis Group, LLC, 2008.

ANDERS, S.; HUBER, W. Differential expression analysis for sequence count data. **Genome Biology**, v. 11, p. R106, 2010. Disponível em: <http://www.bioconductor-.org/packages/release/bioc/html/DESeq.html>.

ANDRADE, E. V. *et al.* Single-chain Fv with Fc fragment of the human IgG1 tag: construction, Pichia pastoris expression and antigen binding characterization. **Journal of biochemistry**, v. 128, n. 6, p. 891–5, Dec 2000.

ANDRADE, E. V. de. **Mapeamento dos Sítios de Interação Antigênica Em Construções Recombinantes de Dois Anticorpos Anti-Z-DNA.** Tese (Mestrado) — Universidade de Brasilia, Departamento de Biologia Molecular, 1997.

ANDRADE, E. V. de *et al.* Thermodynamic basis for antibody binding to Z-DNA: comparison of a monoclonal antibody and its recombinant derivatives. **Biochimica et biophysica acta**, v. 1726, n. 3, p. 293–301, Nov 2005.

BASHAM, B.; SCHROTH, G. P.; HO, P. S. An A-DNA triplet code: thermodynamic rules for predicting A- and B-DNA. **Proceedings of the National Academy of Sciences of the United States of America**, v. 92, n. 14, p. 6464–8, Jul 1995.

BIOCONDUCTOR. **Open Source Software for Bioinformatics**. 2011. Disponível em: <http://www.bioconductor.org/>.

CAIRNS, J. *et al.* BayesPeak—an R package for analysing ChIP-seq data. **Bioinformatics**, v. 27, n. 5, p. 713–714, 2011. Disponível em: <http://www-.bioconductor.org/packages/release/bioc/html/BayesPeak.html>.

CAMERON, A.; TRIVEDI, P. **Regression analysis of count data**. [S.l.]: Cambridge Univ Pr, 1998.

CARLSON, M. *et al.* **GenomicFeatures: Tools for making and manipulating transcript centric annotations**. [S.l.], 2011. Disponível em: <http://www-.bioconductor.org/packages/release/bioc/html/GenomicFeatures.html>.

COLLAS, P. The current state of chromatin immunoprecipitation. **Molecular biotechnology**, v. 45, n. 1, p. 87–100, May 2010.

DAM, L. van; LEVITT, M. H. BII nucleotides in the B and C forms of natural-sequence polymeric DNA: A new model for the C form of DNA. **Journal of molecular biology**, v. 304, n. 4, p. 541–61, Dec 2000.

DITLEVSON, J. *et al.* Inhibitory effect of a short Z-DNA forming sequence on transcription elongation by T7 RNA polymerase. **Nucleic acids research**, Oxford Univ Press, v. 36, n. 10, p. 3163–3170, 2008.

DRÖGE, P.; POHL, F. The influence of an alternate template conformation on elongating phage T7 RNA polymerase. **Nucleic acids research**, Oxford Univ Press, v. 19, n. 19, p. 5301–5306, 1991.

FRANK-KAMENETSKII, M.; VOLOGODSKII, A. Thermodynamics of the B–Z transition in superhelical DNA. **Nature**, Nature Publishing Group, v. 307, p. 481–482, Feb 1984.

FUJITA, P. A. *et al.* The UCSC Genome Browser database: update 2011. **Nucleic Acids Research**, 2010. Disponível em: <http://genome.ucsc.edu>.

GARNER, M. M.; FELSENFELD, G. Effect of Z-DNA on nucleosome placement. **Journal of Molecular Biology**, v. 196, n. 3, p. 581–590, 1987.

Genome Reference Consortium. **hg19/GRCh37**. 2011. Disponível em: <ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/Primary_Assembly/assembled_chromosomes/FASTA/>.

HAIDER, S. *et al.* BioMart Central Portal—unified access to biological data. **Nucleic Acids Research**, v. 37, n. suppl 2, p. W23–W27, Jul 2009. ISSN 1362-4962. Disponível em: <http://www.biomart.org>.

HANDSAKER, B. *et al.* The Sequence Alignment&Map format and SAMtools. **Bioinformatics**, Oxford University Press, Oxford, UK, v. 25, n. 16, p. 2078–2079, Aug 2009. ISSN 1367-4803. Disponível em: <http://samtools.sourceforge.net/>.

HERBERT, A. *et al.* Chicken double-stranded RNA adenosine deaminase has apparent specificity for Z-DNA. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 92, n. 16, p. 7550, 1995.

HO, P. S. *et al.* A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. **The EMBO journal**, v. 5, n. 10, p. 2737–44, Oct 1986.

JAMES, D. A. **RSQLite: SQLite interface for R**. [S.l.], 2011. Disponível em: <http://cran.r-project.org/package=RSQLite>.

JOSEPH, R. *et al.* Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. **Molecular systems biology**, Nature Publishing Group, v. 6, n. 456, p. 456, Dec 2010.

KIM, Y. *et al.* A role for Z-DNA binding in vaccinia virus pathogenesis. **Proceedings of the National Academy of Sciences of the United States of America**, National Acad Sciences, v. 100, n. 12, p. 6974, 2003.

LANGMEAD, B. *et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome Biology**, v. 10, n. 3, p. R25, 2009. Disponível em: <http://bowtie-bio.sourceforge.net/index.shtml>.

LEINONEN, R.; SUGAWARA, H.; SHUMWAY, M. The sequence read archive. **Nucleic acids research**, v. 39, n. Database issue, p. D19–21, Jan 2011. Disponível em: <http://www.ncbi.nlm.nih.gov/Traces/sra>.

LI, H. *et al.* Human genomic Z-DNA segments probed by the Z alpha domain of ADAR1. **Nucleic acids research**, v. 37, n. 8, p. 2737–46, May 2009.

LIU, L.; WANG, J. Supercoiling of the DNA Template during Transcription. **Proceedings of the National Academy of Sciences**, v. 84, n. 20, p. 7024–7027, Oct 1987.

LIU, R. *et al.* Regulation of CSF1 promoter by the SWI/SNF-like BAF complex. **Cell**, Elsevier, v. 106, n. 3, p. 309–318, 2001.

LUKASHIN, A. *et al.* Fluctuational opening of the double helix as revealed by theoretical and experimental study of DNA interaction with formaldehyde. **Journal of molecular biology**, Elsevier, v. 108, n. 4, p. 665–682, 1976.

MARANHÃO, A. Q.; BRÍGIDO, M. M. Expression of anti-Z-DNA single chain antibody variable fragment on the filamentous phage surface. **Brazilian Journal of Medical and Biological Research**, Scielo, v. 33, p. 569–579, 2000.

MORGAN, M.; PAGÈS, H. **Rsamtools: Binary alignment (BAM), variant call (BCF), or tabix file import**. [S.l.], 2010. R package version 1.6.3. Disponível em: <http://bioconductor.org/packages/release/bioc/html/Rsamtools.html>.

NORDHEIM, A.; RICH, A. The sequence $(dC\text{-}dA)_n$ X $(dG\text{-}dT)_n$ forms left-handed Z-DNA in negatively supercoiled plasmids. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 80, n. 7, p. 1821–1825, 1983.

PAGES, H.; ABOYOUN, P.; LAWRENCE, M. **IRanges: Infrastructure for manipulating intervals on sequences**. [S.l.], 2011. R package version 1.12.6. Disponível em: <http://www.bioconductor.org/packages/release/bioc/html/IRanges-.html>.

PECK, L.; WANG, J. *et al.* Transcriptional block caused by a negative supercoiling induced structural change in an alternating CG sequence. **Cell**, v. 40, n. 1, p. 129, 1985.

R Development Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2011. ISBN 3-900051-07-0. Disponível em: <http://www-.R-project.org>.

RICH, A.; NORDHEIM, A.; WANG, A. H. The chemistry and Z-DNA. **Ann. Rev. Biochem.**, v. 53, p. 791–846, 1984.

RICH, A.; ZHANG, S. Z-DNA : the long road to biological function. **Nature reviews. Genetics**, v. 4, n. Jul, p. 566–573, 2003.

ROSENBLOOM, K. R. *et al.* ENCODE whole-genome data in the UCSC Genome Browser. **Nucleic acids research**, v. 38, n. Database issue, p. D620–5, Jan 2010. Disponível em: <http://genome.ucsc.edu/ENCODE>.

SCHROTH, G. P.; CHOU, P. J.; HO, P. S. Mapping Z-DNA in the human genome. Computer-aided mapping reveals a nonrandom distribution of potential Z-DNA-forming sequences in human genes. **The Journal of biological chemistry**, v. 267, n. 17, p. 11846–55, Jun 1992.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. **Nature biotechnology**, v. 26, n. 10, p. 1135–45, Oct 2008.

SILVA, I. C. R. da. **Isolamento e análise de regiões de cromatina nativa contendo segmentos de DNA na conformação Z em diferentes tipos celulares.** Tese (Doutorado) — Universidade de Brasilia, Departamento de Patologia Molecular, 2010.

SPYROU, C. *et al.* BayesPeak: Bayesian analysis of ChIP-seq data. **BMC Bioinformatics**, v. 10, n. 1, p. 299, 2009.

TOMLINSON, I. *et al.* A complete map of the human immunoglobulin VH locus. **Annals of the New York Academy of Sciences**, Wiley Online Library, v. 764, n. 1, p. 43–46, 1995.

URBANEK, S. **multicore: Parallel processing of R code on machines with multiple cores or CPUs**. [S.l.], 2011. Disponível em: <http://www.rforge.net-/multicore/index.html>.

VOLOGODSKII, A. *et al.* Fluctuations in superhelical DNA. **Nucleic acids research**, Oxford Univ Press, v. 6, n. 3, p. 967–982, 1979.

WANG, J. Interactions between twisted DNAs and enzymes: the effects of superhelical turns. **Journal of molecular biology**, v. 87, n. 4, p. 797, 1974.

WHEELER, R. **A-DNA, B-DNA and Z-DNA**. Feb 2007. Disponível em: <http://en.wikipedia.org/wiki/File:A-DNA,_B-DNA_and_Z-DNA.png>.

_____. **Circular DNA supercoiling**. Apr 2007. Disponível em: <http://en.wikipedia-.org/wiki/File:Circular_DNA_Supercoiling.png>.

WICKHAM, H. **ggplot2**. 2011. Disponível em: <http://had.co.nz/ggplot2/>.

WITTIG, B. *et al.* Transcription of human c-myc in permeabilized nuclei is associated with formation of Z-DNA in three discrete regions of the gene. **The EMBO journal**, Nature Publishing Group, v. 11, n. 12, p. 4653, 1992.

WÖLFL, S.; WITTIG, B.; RICH, A. Identification of transcriptionally induced Z-DNA segments in the human c-myc gene. **Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression**, Elsevier, v. 1264, n. 3, p. 294–302, 1995.

XIAO, J.; DRÖGE, P.; LI, J. Detecting Z-DNA Forming Regions in the Human Genome. **International Conference on Genome Informatics 2008**, 2008.

ZHABINSKAYA, D.; BENHAM, C. Theoretical Analysis of the Stress Induced BZ Transition in Superhelical DNA. **PLoS computational biology**, Public Library of Science, v. 7, n. 1, p. e1001051, 2011.

ZHU, L. J. *et al.* **ChIPpeakAnno: Batch annotation of the peaks identified from either ChIP-seq, ChIP-chip experiments or any experiments resulted in large number of chromosome ranges.** [S.l.], 2011. Disponível em: <http://www.bioconductor.org/packages/release/bioc/html/ChIPpeakAnno.html>.