UNIVERSIDADE FEDERAL DO RIO GRANDE
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL
CURSO DE MESTRADO EM MODELAGEM COMPUTACIONAL

Dissertação de Mestrado

PROBLEMAS INVERSOS EM SISTEMAS BIOFÍSICOS: DESCOMPACTAÇÃO E SEQUENCIAMENTO DO DNA

Tiago da Silva Gautério

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal do Rio Grande- FURG, como requisito parcial para a obtenção do grau de Mestre em Modelagem Computacional

Orientador: Prof. Dr. Adriano De Cezaro
Co-orientadora: Prof[a]. Dr[a] Fabiana Travessini De Cezaro

Rio Grande, 2016

Ficha catalográfica

G275p Gautério, Tiago da Silva.
Problemas inversos em sistemas biofísicos: descompactação e sequenciamento do DNA / Tiago da Silva Gautério. – 2016.
71 f.

Dissertação (mestrado) – Universidade Federal do Rio Grande – FURG, Programa de Pós-graduação em Modelagem Computacional, Rio Grande/RS, 2016.
Orientador: Drª. Adriano De Cezaro.
Coorientadora: Drª. Fabiana Travessini De Cezaro

1. Descompactação de DNA 2. Regularização 3. Inferência Bayesiana 4. Sequenciamento I. De Cezaro, Adriano II. De Cezaro, Fabiana Travessini III. Título.

CDU 51:004

Catalogação na Fonte: Bibliotecário Me. João Paulo Borges da Silveira CRB 10/2130

Banca examinadora:

Prof. Dr. Adriano De Cezaro - FURG/BR (Orientador)

Profª. Drª. Diana Francisca Adamatti - FURG/BR

Prof. Dr. Juan Pablo Agnelli - UNC/AR

Dedico este aos meu pais, irmãos, meu cônjuge Félix e a toda minha família que sempre estiveram presentes em todos os desafios de minha vida. O carinho e o apoio de vocês me trouxe paz e a segunrança necessária para acreditar que meus sonhos podem se tornar realidade.

AGRADECIMENTOS

Gostaria de agradecer a Deus pelo dom da vida e por colocar em meu caminho pessoas tão especiais, como minha família, meu cônjuge e meus amigos.

À minha família, em especial meus pais e irmãos, que amo incondicionalmente, pelo carinho, paciência, apoio e por serem meu porto seguro.

Aos meus colegas e ex-colegas de profissão, aos meus alunos e aos meus amigos por estarem comigo nos momentos felizes e também nos de angústia, tornando minha caminhada suave.

A meu orientador e minha co-orientadora, por acreditarem em mim e fazerem parte da minha história nos momentos bons e ruins, por serem exemplos de pessoa e profissional os quais sempre farão parte da minha vida, são grandes amigos.

A CAPES pelo apoio financeiro.

Finalmente, à Universidade Federal do Rio Grande - FURG, em especial ao Programa de Pós-Graduação em Modelagem Computacional que abriram as portas para que eu pudesse realizar o sonho que é esta dissertação de mestrado.

Obrigado a todos!

Eu não disse que seria fácil. Apenas disse que valeria a pena.
— DOM BOSCO

RESUMO

Desde a década de 70 modelos matemáticos são estudados para melhor compreender a estrutura e o funcionamento do DNA. Neste trabalho faremos a análise de modelos de descompactação para a molécula de DNA e das informações que são possíveis de serem recuperadas sobre o sequenciamento a partir dos sinais obtidos pelos modelos de descompactação, conhecidos na literatura como o problema inverso da descompactação. Basicamente, consiste em determinar informações importantes sobre a estrutura original da molécula de DNA apenas com os dados gerados pelo processo de descompactação. A novidade deste trabalho reside no fato de que usaremos informações probabilísticas da distribuição dos sinais em conexão com o Teorema de Bayes para estabelecermos métodos de regularização para o problema inverso. Finalmente, descrevemos alguns algoritmos numéricos para o processo de sequenciamento.

Palavras-chave: Descompactação do DNA, Regularização, Inferência Bayesiana, Sequenciamento.

ABSTRACT

Since the 70's mathematical models are studied in order to better understand the structure and function of DNA. In this work we will analyze uncompressing models for the DNA molecule and the information that is possible to be recovered on the sequencing from the signals obtained by the models of uncompressing, known in the literature as the inverse problem of uncompressing. Basically, it consists in determining important information about the original structure of the DNA molecule only with the data generated by the uncompressing process. The innovation of this work lies in the fact that we use probabilistic information of the distribution of signals in connection with the Bayes theorem to establish regularization methods for the inverse problem. Finally, we describe some numerical algorithms for the sequencing process.

Keywords: DNA Unzipping, Regularization, Bayesian Inference, Sequencing.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

O ácido desoxirribonucleico (ADN), mais conhecido como DNA, sigla de ”deoxyribonucleic acid”, vem sendo estudado por muitos anos. Desde seu descobrimento, vários estudos tem sido desenvolvido a fim de caracterizar e entender o seu comportamento, haja visto que as moléculas de DNA são as responsáveis por toda a informação genética. Em outras palavras, sequências específicas, chamadas de genes, são os responsáveis por codificar as proteínas que regulam a maioria das funções vitais, além de formarem a maioria das estruturas celulares.

Como mostram as descobertas recentes, a maioria das doenças possuem fatores genéticos associados [17]. Deste modo, caso haja alterações na formação, composição e/ou mutação das moléculas que compõem a sequência DNA, as proteínas formadoras desta sequência podem ser incapazes de desenvolver suas funções normais, as quais resultam em uma disordem genética, acarretando, por sua vez, doenças associadas.

O conhecimento da sequência de DNA, assim, possui um papel de importância central tanto nos diagnósticos como em medidas terapêuticas da detecção e tratamento de doenças. A importância deste problema despertou o interesse de diversos grupos de pesquisa em todas as partes do mundo. Um resultado especial deste esforço é o grande Projeto Genoma Humano [10]. Atualmente, testes genéticos estão disponíveis para detectar mutações na sequência de DNA. Alguns destes testes são usados tanto para elucidar alguns diagósticos prévios quanto para direcionar as ações de tratamento, bem como ajudar a identificar riscos de doenças que possam ser prevenidas. Um exemplo classico destes disgnósticos são os testes nos recém nacidos. Por outro lado, genes específicos, codificados a partir de genes conhecidos, podem ser aplicados para tratar certas doenças, ou ainda, para induzir a secreção de proteínas específicas que possuem as funções terapêuticas desejadas [20]. A possibilidade de determinar, detectar e manipular a sequência de DNA passa, sem sombra de dúvidas, pelo conhecimento de sua dinâmica física, química e biológica. Para compreender estes processos, necessitamos entender o processo de descompactação e também do sequenciamento do DNA.

Neste trabalho nossos objetivos não são tão ambiciosos quanto os descritos acima, ou seja, o de descrever os processos físicos de descompactação e sequenciamento mais

avançados e sim, pretendemos expor uma breve introdução ao assunto. Para tal, faremos uma análise de alguns modelos de descompactação discutidos em [1, 2]. Devido ao avanço dos estudos nessa área e, consequentemente, o descobrimento de novas maneiras de explorar as propriedades do DNA desperta-nos o interesse de analisar a descompactação do mesmo. A descompactação do DNA de fita dupla com a aplicação de uma força, depende diretamente da sequência do DNA. No entanto, outros fatores, como flutuações devido as diferentes forças necessárias para romper as ligações devem ser consideradas no modelo. Apresentaremos no Capítulo 4 a modelagem deste problema para o caso de uma força constante. A dinâmica de descompactação, levando em consideração flutuações randômicas serão fruto de trabalhos futuros.

Uma vez conhecida a dinâmica de forças para descompactar uma sequência de DNA, surge a pergunta: É possível determinar a sequência de proteínas que compõem o DNA a partir do conhecimento das forças necessárias para descompactá-la? A resposta a esta pergunta pode ser formulada como um problema inverso [14, 21], o do sequenciamento, relacionado ao problema direto, que neste caso é a dinâmica da descompactação. Para resolver tal problema, é necessário construir uma força microscópica diferencial que usa ambos, o DNA nativo e o de teste, e nos fornece como resultado a força diferencial como uma função do comprimento da distância. Através do sinal da força diferencial, pretende-se descrever a natureza da mutação. Além disso, utilizaremos técnicas de calibração para determinar a posição da cadeia de DNA a partir das posições dos extremos da curva da força diferencial. O processo de sequenciamento (problema inverso) será descrito, para um caso simples, no Capítulo 5. Neste mesmo capítulo descrevemos o algoritmo que pode ser usado no sequenciamento, embora não apresentamos, neste trabalho, a implementação do mesmo. No entanto, apresentamos, para uma cadeia bem curta, o resultado do sequenciamento, feito “por força bruta”.

A originalidade de nossa proposta está em utilizar a fórmula de Bayes associada a distribuição de probabilidade dos dados observados (forças de rompimento) para determinar possíveis mutações na cadeia de formação de DNA.

A relação entre a fórmula de Bayes, e a teoria de problemas inversos é feita no Capítulo 3. Em especial, relacionamos a teoria bayesiana com a regularização de Tikhonov (Tikhonov iterado), teoria pela qual podemos garantir que as soluções obtidas não sofrem, muito, com os erros na modelagem, bem como com os erros nas observações, característicos de problemas mal postos (problemas inversos).

Finalizamos este trabalho no Capítulo 6 onde apresentamos as conclusões e os trabalhos futuros.

Como aplicação imediata do problema estudado neste trabalho podemos citar: a determinação de bactérias que infectaram um paciente e a determinação da variação genética responsável pela doença através da detecção das mutações da cadeia de DNA. No entanto, isso só é possível se conhecermos um pouco, ao menos, da estutura químico-

física do DNA. Logo, reservamos um espaço neste capítulo para apresentarmos um estudo mais detalhado acerca da molécula de DNA e suas características. Com este breve apanhado, conseguiremos avançar em direcão aos objetivos deste trabalho.

1.1 O DNA

Para a compreensão das características do DNA e até mesmo das ações como, por exemplo, o processo de descompactação, a flexibilidade e o sequenciamento, se faz necessário um estudo acerca da estrutura dessa molécula, envolvendo propriedades químicas, mecânica e a separação de cadeias. Neste sentido, faremos um resumo dos principais conceitos para o entendimento desta estrutura, também disponível em [1] [27].

1.1.1 Estrutura e propriedades químicas

O DNA é uma macromolécula feita de deoxiribonucleotideos, onde o açúcar e os grupos fosfatos tem um papel estrutural, enquanto que as bases carregam informações genéticas. A espinha dorsal do DNA é constante, hidroxilas de açúcar são unidas através de grupos de fosfatos por meio de reações. A parte variável do DNA é constituído pelas bases, ligadas a açúcares, que são de 4 tipos, Adeninda (A), Guanina(G), que são as duas purinas, e Timina (T) e (Citosina), que são as duas pirimidinas. Nestas sequências de purinas e piramidinas é que esta contida toda a informação do DNA.

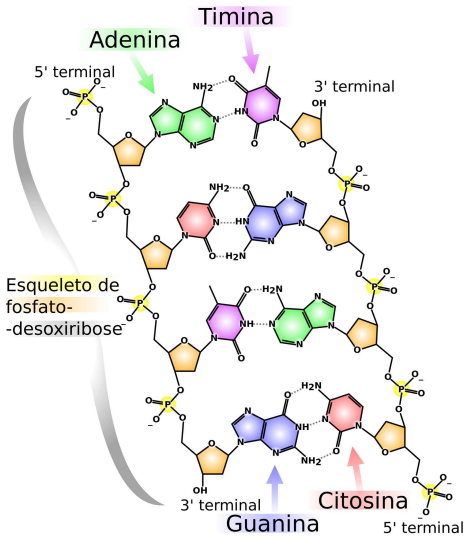

Figura 1: Estrura da molécula de DNA, destacando as ligações entre os pares de bases. Também disponível em [1]

A estrutura tridimensional do DNA foi primeiramente estudada por James Watson e Francis Crick através de uma análise de difrações de raio- X das fibras do DNA. De seus estudos eles concluíram que o DNA é uma dupla Hélice com duas cadeias de polinucleotideos, correndo em direções opostas enrolados em torno de um eixo comum, as bases

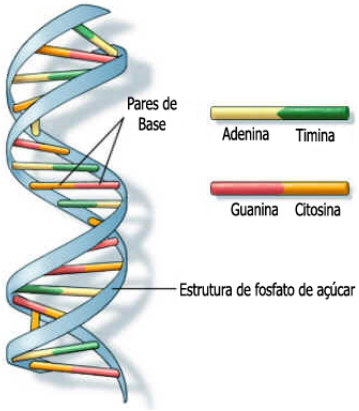

Figura 2: Estrura da molécula de DNA no formato de dupla hélice, destacando principalmente a formação dos pares de bases. Também disponível em [1]

ocupam o núcleo da hélice enquanto que a espinha dorsal se contorce em seu entorno formando ondulações que permitem os pares de bases interagirem. Watson e Crick também descobriram que as purinas e as pirimidinas não se ligam entre si devido ao espaço disponível para as ligações. Esta estrutura é denotada por Forma-B, em tradução livre, existem ainda outros tipos de formas como a A e a Z, que oferecem algumas vantagens em relação a B, por exemplo, a forma A é bem mais compacta porém em contra partida oferece menos informações, na figura 3, trazemos ilustrações das formas A, B e Z de uma molécula de DNA vista superiormente e frontal.

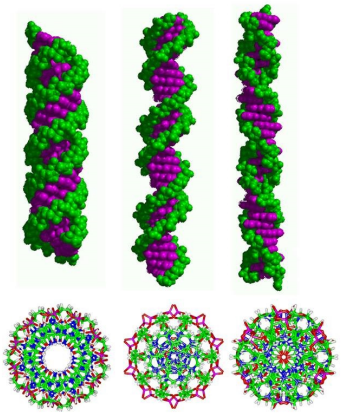

Figura 3: Ilustrações das formas A, B e Z de uma molécula de DNA vista superiormente e frontal. Também disponível em [1]

Uma contrinuição bastante importante desse estudo, é a interação no empilhamento entre as bases adjacentes na mesma cadeia. As investigações têm demonstrado que as bases têm uma tendência intrínseca para empilhar-se, que é reforçada por um solvente aquoso. Também cabe salientarmos, que as interações de empilhamento são dependentes

da sequência, ou seja, diferentes conjuntos de bases têm energias de empilhamento diferentes. O estudo acerca dos valores de empilhamento, Baldazzi [1] cita que foi feito por Santa Lucia, como resultado desse estudo destacamos a matriz de empilhamento completa na figura 4.

i \ i+1	A	T	C	G
A	-1,0	-0.88	-1.44	-1.28
T	-0.58	-1.0	-1.30	-1.45
C	-1.45	-1.28	-1.84	-2.17
G	-1.30	-1.44	-2.24	-1.84

Figura 4: Matriz de empilhamento completa. Também disponível em [1]

1.1.2 Mecânica do DNA

Devido a sua particular estrutura, as propriedades físicas do DNA são diferentes de qualquer outro polímero e se fazem essenciais em muitos processos celulares. A mecânica do DNA afeta, por exemplo, a maneira de como o DNA interage com as proteínas. Graças aos trabalhos de muitos grupos de estudos ao redor do mundo, as propriedades mecânicas do DNA são muito bem conhecidas.

1.1.2.1 Experimentos com moléculas únicas

Estes experimentos surgiram no inicio da década de 90. Pela primeira vez uma simples molécula de DNA pode ser manipulada, puxada e até mesmo rotacionada. Esses experimentos revolucionaram a bioquímica tradicional, que trabalhava com o comportamento médio de uma série de moléculas. Estes experimentos com moléculas únicas proporcionaram uma série de propriedades sobre o DNA tal como suas reações quando submetido a estresse de torções e tensões.

1.1.2.2 Alongamento do DNA

DNA de cadeia dupla

Em 1992 as propriedades elásticas do DNA de cadeia Dupla (dsDNA) foram primeiramente estudadas na Califórnia e em Milão. O trabalho destes pesquisadores é baseado em um método de *"cut and paste"* da biologia molecular que torna a molécula de DNA maleável. Neste tipo de experimento o DNA é fixado por uma espécie de alça e nele são aplicadas forças, primeiramente eram aplicadas forças através de campos magnéticos, atualmente são utilizados diversos tipos de instrumentos como micro agulhas e pinças

ópticas. O comportamento elástico do DNA esta mostrado na figura 5, mais informações desse processo podem ser encontrados em [1].

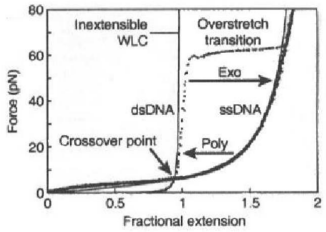

Figura 5: O comportamento elástico do DNA. Também disponível em [1]

DNA de cadeia simples

A elasticidade de um DNA de cadeia simples (ssDNA) é bastante diferente do de uma hélice. O ssDNA é mais flexível que o dsDNA e se mantem muito compactado para forças de 6pN. Para essa quantidade de força o comprimento do ssDNA é menor que o do dsDNA. Porém uma vez que o ssDNA começa a esticar, por não ser limitado por uma estrutura helicoidal, torna-se duas vezes mais longo que o de dupla hélice. Diferentemente do dsDNA, o ssDNA é mais vulnerável as condições de sal e portanto não exibe um comportamento elástico como o dsDNA.

1.1.3 A separação das cadeias

A duplicação celular é um processo natural de todas as células, neste processo uma célula original se duplica dando origem a outra. No caso do DNA dizemos que esse processo é semiconservativo, pois a duplicação gera combinações de pares de bases diferentes do original. Infelizmente o processo de replicação ainda é desconhecido, as únicas informações que temos é sobre o que temos antes e depois do processo de replicação.

1.1.3.1 Desnaturação do DNA

Uma maneira de quebrar ligações de hidrogênio é aquecer o DNA a uma temperatura muito perto do ponto de fervura ou utilizando agentes de desnaturação. Nessas condições a estrutura do DNA colapsa, isto é, as duas cadeias se separam e assumem uma configuração aleatória. Esse processo é acompanhado de uma perda qualitativa das propriedades físicas da molécula de DNA, que podem ser usadas para compreender a desnaturação do mesmo. Em analogia ao processo de fusão dos sólidos, a temperatura do ponto médio da curva da figura 6 é definida como temperatura de fusão T_m. Esta temperatura é a necessária para abrir uma mólecula e ela varia dependendo da composição dos nucleotídeos. Se o houverem mais ligações do tipo G-C a temperatura será mais elevada do que as ricas em A-T.

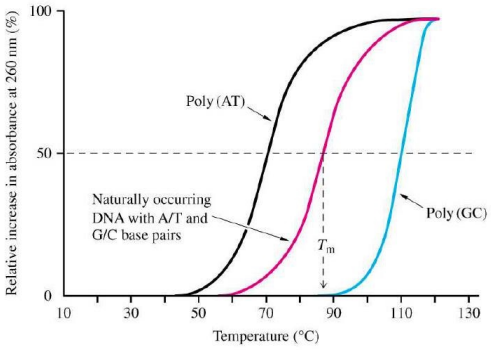

Figura 6: Representação da curva da absorvância de $260nm$ do DNA. Também disponível em [1]

Quando aquecidas as moléculas de DNA ricas em G-C, por exemplo, teram as regiões ricas em A-T quebradas primeiro, nessas condições, qualquer arrefecimento pode fazer com que a molécula torne a se ligar. Somente quando aproximadamente $\frac{3}{4}$ da separação ocorrer é que o processo de rompimento se torna irreversível.

A Cinética da descompatcação é um processo muito complexo, as taxas de desenrolamento são muito lentas porém estudos de 1972 conseguiram aproximar essa taxa de desenrolamento, que variava ao longo do tempo, por uma constante K_{inf} que é inversamente proporcional ao peso molécular do DNA.

1.1.3.2 Replicação do DNA

Nas últimas décadas os conhecimentos sobre a replicação do DNA tem aumentado, hoje já existem mecanismos suficientes para reprodução in vitro. A replicação inicia basicamente com uma separação das fitas que é realizada, dentro das células, por uma enzima denominada DNA-helicase, que ”anda” pela dupla hélice e mecanicamente descompacta as duas fitas. Durante esta replicação são formadas bolhas, chamadas de θ estruturas, que auxiliam o DNA-helicate a descompactar em ambas as direções. Este processo gera 2 problemas, o primeiro seria de a molécula começar a rotacionar, o que naturalmente não acontece devido a uma outra enzima que atua contra este sentido. O segundo seria uma descompactação muito rápida, para este problema uma enzima denominada DNA topoisomerase controla este problema. A replicação é realizada pela enzima DNA polymerase, que ”anda” sobre a fita simples e usa ela como um modelo para sintetizar seus complementares. Essa replicação é feita em forma de Y como mostra a figura 7.

1.1.3.3 Sequenciamento do DNA

Os conhecimentos a cerca do sequenciamento do DNA são de extrema importância tanto biologicamente quanto medicamente falando. Dada a importância do sequenci-

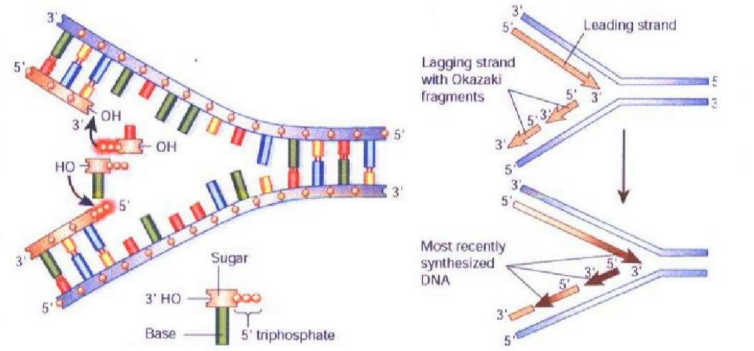

Figura 7: Replicação em forma de y. Também disponível em [1]

amento falaremos aqui de alguns métodos de sequenciamento enfatizando vantagens e limitações.

Método Eletroforetico

É um método tradicional baseado no método Sanger que utiliza um nucleotídeo modificado de dideoxynucleotides (ddNTP). Neste processo, esse nucleotídeo interrompe o processo de replicação e a partir de uma solução previamente preparada com DNA-polymerase começa completar as fitas já fechadas, através de uma análise por cromatografia é possível construir curvas que analisam a quantidade de nucleotídeos utilizados em determinado momento. Uma das vantagens desse método é que os resultados podem ser lidos em poucas horas, porém nestes métodos podem haver muitos ruídos fazendo com que a leitura não se torne tão clara e também devido a estrutura ser do tipo grampo podem haver problemas de deslocamento dos picos das curvas, fugindo do que seria um traço ideal.

Pirosequenciamento

Este método baseia-se na liberação de pirofosfato, liberado durante a síntese de DNA. No mesmo molde do anterior, uma mistura é preparada, porém aqui quando o complemento é construído de forma correta é liberada uma quantidade de pirofosfato que deve ser detectada. Se o caminho é construído de forma errada pouco ou nenhum pirofosfato deve ser liberado. O problema desta técnica é que ela é pouco eficaz para sequências muito longas, devido as limitações da leitura, porém é um método que evita a rotulagem fluorescente dos nucleotídeos o que o torna um método economicamente mais viável.

1.1.3.4 Estudos em moléculas simples

Devido aos recentes avanços na manipulação de moléculas simples de DNA, tem se direcionado amplos estudos para esse tipo de molécula em relação a mecanismos moleculares e iterações entre DNA e proteínas. Esses estudos tem revelado importantes

detalhes em relação a cinética da descompactação e informações a respeito da termodinâmica do processo. Em um experimento clássico, a fita é presa em dois suportes sólidos distintos e progressivamente separada. Existem basicamente duas maneiras de se fazer essa separação, na primeira possibilidade mantemos a velocidade constante e analisamos o sinal da força, na segunda mantemos a força constante e analisamos a dinâmica da descompactação.

Descompactação com velocidade constante

Neste experimento, um dos lados do conjunto molecular é ligado a superfície da lâmina do microscópio enquanto que o outro é ancorado a um talão microscópico. Em ambos os casos o produto químico fixador é obtido por meio de iterações corpo/anti-corpo. Uma micro agulha de vidro é introduzida através da meia-lua livre e fixada ao cordão. Esta micro agulha serve como um sensor de força. A amostra é deslocada lateralmente a uma velocidade constate. A principal fonte de ruídos nesse tipo de experimento advém de vibrações mecânicas decorrentes de perturbações sísmicas e acústicas, este ruído pode ser parcialmente reduzido pela baixa iteragem e mudança nos níveis de detecção. Na figura 8, temos um aparato para o experimento de descompactação utilizando pinças opticas.

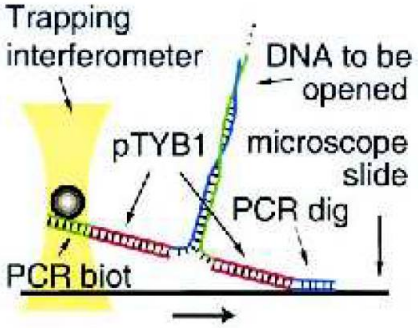

Figura 8: Aparato para o experimento de descompactação utilizando pinças ópticas. Também disponível em [1]

Descompactação com força constante

Neste tipo de experimento uma molécula de $\lambda - phage$ DNA é ligada a um capilar e a uma esfera magnética. A molécula a ser aberta é ligada por um gancho ao capilar através de três núcleos de timina da cadeia simples. Este experimento é realizado dentro de uma microcelula de vidro. Um gradiente de campo magnético permite exercer uma força controlada de forma gradual a separar as duas vertentes. Cinco imãs são colocados em um estágio de translação, realizada em um lado da microcelula, a consequência disso é uma força praticamente perpendicular ao vidro a qual o DNA esta ligado. A magnitude da força é determinada pela distância entre os imãs e a esfera. A força pode variar em até 20% dependendo do granulo do ferro, mas como muitas medidas podem ser feitas em paralelo utilizando a mesma molécula, o valor da força reportado é a media de diferentes

esferas. A figura 9, ilustra o experimento.

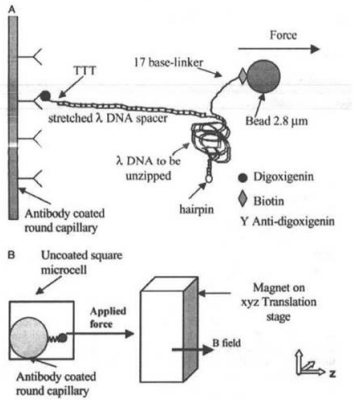

Figura 9: Ilustração do experimento de descompactação com força constante. Também disponível em [1]

2 PROBLEMAS INVERSOS

De forma grosseira, dizemos que um problema é um *problema inverso* quando o objetivo é convertermos medidas observadas em informações acerca de um objeto ou sistema, os quais são, neste sentido, "inversos" aos problemas conhecidos como *diretos*, que, naturalmente buscam determinar as consequências ligadas a uma dada causa. A solução destes problemas inversos, são de extrema importância, porque ela nos fornece informações que não podem ser observadas diretamente. Sendo assim, descreveremos neste capítulo um apanhado da teoria de problemas inversos que nos possibilite realizar um estudo, nos capítulos subsequentes, principalmente do sequenciamento (reconstrução) da molécula de DNA, um dos objetivos deste trabalho.

2.1 Síntese histórica de problemas inversos

A área de estudo denominada Problemas Inversos é uma área bastante antiga quando pensamos nos primeiros problemas que se tem conhecimento, [13]. Se criássemos uma linha do tempo para descrever tal história, partiríamos do problema de reconstrução da realidade a partir da imagem de objetos projetados como sombras no interior de uma caverna, proposto por Platão. Também o problema de determinar o diâmetro da terra dado medições feitas em duas cidades distintas, proposto por Eratósthenes, ambos relatados a.c.. Na continuidade, o problema de determinar a órbita de um cometa a partir de sua órbita anterior, proposto por Gauss em 1800. Mais recente Randon, com a "transformada de Randon"em 1917, método matemático para a resolução de problemas específicos. Estes formam um apanhado, bastante sintetizado, da área de estudo em questão. Atualmente, o estudo de problemas inversos vem atraindo muitos pesquisadores e tem se desenvolvido rapidamemte abrangindo diversas áreas do conhecimento. Além disso, a resolução desses problemas envolvem diversos setores da matemática e ciências aplicadas, caracterizando um estudo multidisciplinar.

2.2 Definindo problema inverso

A palavra problema em certas áreas do conhecimento pode ser definida como a dificuldade de obtenção de determinado objetivo. Já nas ciências exatas, podemos interpretar esta palavra como um desafio, que pode possui ou não soluções e, que por vezes possuem várias soluções. Neste sentido, os problemas que possuem uma causa, dado um efeito, são de grande importância e possuem várias aplicações.

Para a compreensão destes problemas e para definirmos de forma mais precisa os conceitos de problema direto e problema inverso, vamos considerar um processo dividido em três etapas, como mostra a figura 10.

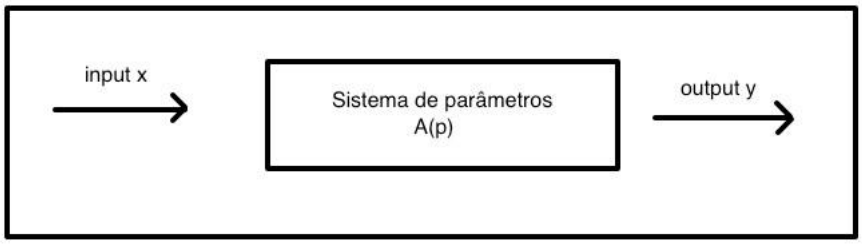

Figura 10: Modelagem do fenômeno em estudo, [13].

Do ponto de vista matemática (de modelagem), podemos representar o processo acima por uma equação do tipo

$$A(p).x = y\,. \tag{2.1}$$

Neste sentido, o *problema direto* é definido quando temos conhecimento acerca da causa (*input* "x") e do sistema de parâmetros (A(p)), procurando determinar o efeito (*output* "y"). Já o *problema inverso* pode aparecer na equação 2.1 de duas formas:

1) Temos informações acerca do sistema de parâmetros (A(p)) e os efeitos (*output* "y"), desejamos conhecer a causa (*input* "x"). Esse problema é conhecido como problema de reconstrução.

2) Temos informações acerca da causa (*input* "x") e os efeitos (*output* "y"), desejamos conhecer o sistema de parâmetros (A(p)). Esse problema é conhecido como problema de identificação.

Na resolução de problemas inversos, devemos levar em consideração a existência de ruídos "δ", pelo fato do efeito (*output* "y") ser uma medida e dificilmente obtido de forma precisa. Com isso, o efeito usado na resolução desse tipo de problema é denotado por y^{δ}

e assumimos que satisfaça:

$$\|y - y^{\delta}\| \leqslant \delta \tag{2.2}$$

Outro fato de extrema importância é que os problemas inversos, ao contrário dos problemas diretos, são mal-postos no sentido de Hadamard, [16].

Definição 2.1 (Problema bem/mal posto no sentido de Hadamard)**.** *Um problema é dito bem posto, no sentido de Hadamard, se satisfaz as condições:*

- *Existência da solução;*
- *Unicidade da solução;*
- *Dependência contínua da solução com relação aos dados.*

Caso uma dessas condições não seja satisfeita, o problema é dito mal posto.

Um dos efeitos mais presentes nos problemas inversos é a falta da dependência contínua com relação aos dados. Na próxima subseção apresentaremos um exemplo simples que mostra como este efeito aparece. Este exemplo também servirá para motivar a necessidade de consideramos as chamadas estratégias de regularização para problemas inversos.

2.2.1 Exemplo: o problema inverso da diferenciação

Sejam $y, y^{\delta} : [0,1] \to \mathbb{R}$ funções contínuas satisfazendo a desigualdade

$$||y(t) - y^{\delta}(t)||_{\infty} \leqslant \delta, \ \forall \ \mathrm{t} \in [0,1]. \tag{2.3}$$

Gostaríamos de reconstruir $x = y^{'}$ de y.

Do ponto de vista numérico, a estratégia é considerar aproximações por diferenças simétricas, ou seja, $\forall \tau \in (0,1)$,

$$x^{\delta,h}(\tau) := \frac{y^{\delta}(\tau + h) - y^{\delta}(\tau - h)}{2h}. \tag{2.4}$$

Gostaríamos de mostrar que quando $h \to 0$, $x^{\delta,h}(\tau)$ se aproxima de x. Sendo assim,

temos:

$$
\begin{aligned}
& ||x^{\delta,h}(\tau) - x(\tau)||_\infty = ||\frac{y^\delta(\tau+h) - y^\delta(\tau-h)}{2h} - x(\tau)||_\infty \\
= \; & ||\frac{y^\delta(\tau+h) - y^\delta(\tau-h) + y(\tau+h) - y(\tau+h) + y(\tau-h) - y(\tau-h)}{2h} - x(\tau)||_\infty \\
= \; & ||\frac{y^\delta(\tau+h) - y(\tau+h) - y^\delta(\tau-h) + y(\tau-h)}{2h} + \frac{y(\tau+h) - y(\tau-h)}{2h} - x(\tau)||_\infty \\
= \; & ||\frac{(y^\delta - y)(\tau+h) - (y^\delta - y)(\tau-h)}{2h} + \frac{y(\tau+h) - y(\tau-h)}{2h} - x(\tau)||_\infty \\
\leqslant \; & ||\frac{(y^\delta - y)(\tau+h) - (y^\delta - y)(\tau-h)}{2h}||_\infty + ||\frac{y(\tau+h) - y(\tau-h)}{2h} - x(\tau)||_\infty
\end{aligned}
$$

Supondo que

$$||x^{'}(t)||_\infty \leqslant E, \forall \text{ t} \in [0,1] \tag{2.5}$$

Por 2.3, temos que:

$$||\frac{(y^\delta - y)(\tau+h) - (y^\delta - y)(\tau-h)}{2h}||_\infty \leqslant \frac{\delta}{2h} \leqslant \frac{\delta}{h}$$

Por 2.5, temos que:

$$||\frac{y(\tau+h) - y(\tau-h)}{2h} - x(\tau)||_\infty \leqslant \frac{h}{h}||\frac{y(\tau+h) - y(\tau-h)}{2h} - x(\tau)||_\infty \leqslant ||x^{'}(\tau).h|| \leqslant E.h$$

Sendo assim,

$$||x^{\delta,h}(\tau) - x(\tau)||_\infty \leqslant E.h + \frac{\delta}{h} \tag{2.6}$$

Note que, quando $h \to 0$, $||x^{\delta,h}(\tau) - x(\tau)||_\infty$ não tende para zero. Sendo assim, não podemos simplesmente tomar um h pequeno, pois a desigualdade 2.6 mostra que $x^{\delta,h}$ pode estar longe da solução. No entanto, o h deve ser escolhido de tal forma que se aproximamos da solução exata.

Devido ao fato dos problemas inversos serem mal postos, torna-se indispensável os métodos de regularização para garantirmos uma melhor aproximação da real solução.

2.3 Metodos de regularização

Regularização se faz necessário, quando precisamos resolver um problema mal posto. A maioria dos problemas inversos, pelo fato de apresentarem, principalmente, falta da depêndencia contínua da solução em relação aos dados, são mal postos. Logo, para determinar uma solução que se aproxime da solução exata é necessário uma estratégia ou método de regularização, determinando assim uma solução aproximada (x_α^δ) de maneira

estável e que convirja, quando o nível de ruídos converge para zero, para a solução $x^\dagger$ do problema inverso, [13].

Definição 2.2. *Sejam $A : \mathcal{H}_1 \to \mathcal{H}_2$ um operador linear limitado e $\alpha_0 \in (0, +\infty)$. Para todo $\alpha \in (0, \alpha_0)$, seja*

$$R_\alpha : \mathcal{H}_2 \to \mathcal{H}_1$$

um operador contínuo (não necessariamente linear). A família $\{R_\alpha\}$ é chamada de uma regularização ou uma família de operadores de regularização (para $A^\dagger$, a inversa generalizada de A, veja [13] para definição e propriedades) se, para todo $y \in \mathcal{D}(A^\dagger)$, existir uma regra para escolha do parâmetro $\alpha := \alpha(\delta, y^\delta)$ tal que

$$\limsup_{\delta \to 0} \{\|R_\alpha y^\delta - A^\dagger y\| : y^\delta \in \mathcal{H}_2, \|y - y^\delta\| \leq \delta\} = 0$$

é satisfeita para $\alpha := \alpha(\delta, y^\delta) \xrightarrow{\delta \to 0} 0$.

Observação 2.3. *Note que não estamos requerendo que a família de aperadores de regularização $\{R_\alpha\}$ seja de operadores lineares. No caso em que $\{R_\alpha\}$ é linear, então dizemos que o método de regularização é linear.*

Definição 2.4. *Uma estratégia de regularização (R_α, α) é dita convergente se $x_\alpha := R_\alpha y$ converge para $x^\dagger$, onde $x^\dagger$ representa uma solução generalizada do problema* (2.1).

Existem diversos métodos de regularização [13]. Destacaremos nas subseções seguintes dois destes que se fazem necessários para o entendimento da discrição dos problemas em questão.

2.3.1 Regularização de Tikhonov para problemas lineares

As estratégias ou Métodos de Regularização para problemas inversos, se fazem necessários para contornarmos fatores como instabilidade e mal-condicionamento, [13]. A Regularização de Tikhonov tem este compromisso de precisão e estabilidade, que firmaremos ao longo deste capítulo.

2.3.1.1 Convergência

Estamos interessados em determinar, de forma estável, uma aproximação para a solução do problema inverso

$$Ax = y^\delta. \tag{2.7}$$

para medidas conhecidas do erro $||y - y^\delta|| \leqslant \delta$ e um operador A, linear.

No entanto, determinar uma solução regularizada requer estratégias diferentes do que tentar resolver equações normais

$$A^*Ax = A^*y^\delta. \tag{2.8}$$

Ou, até mesmo, encontrar um mínimo para o problema variacional de quadrados mínimos

$$J(x) = \frac{1}{2}||Ax - y^\delta||^2. \tag{2.9}$$

Como ponto de partida, queremos inverter o operador A de forma a manter o resíduo $||Ax - y^\delta||$ controlado. Pelo Teorema da Aplicação Espectral [22], é equivalente a calcular $g(A^*A)$, onde $g(\lambda) = \frac{1}{\lambda}$, $\lambda \neq 0$. Logo, uma solução aproximada para equação 2.8 é dada por

$$x^\delta = g(A^*A)A^*y^\delta. \tag{2.10}$$

Agora, caso o operador A seja mal condicionado (compacto, por exemplo), sabemos que o espectro de A^*A vai estar muito próximo ou ser igual a zero [22]. Devido a isso, a estratégia 2.10 não é possível ou é muito instável. Logo, para determinarmos uma solução aproximada para o problema inverso, de maneira estável, devemos afastar o espectro de A^*A de zero.

Sendo assim, seja $0 < \alpha \in [0, \alpha_0]$, defina

$$f_\alpha(\lambda) := g(\lambda + \alpha) = \frac{1}{\lambda^2 + \alpha} \tag{2.11}$$

A função $f_\alpha(.)$ é dita ser uma *função filtro* para o método de regularização de Tikhonov. Do Teorema da Aplicação Espectral [22], temos que

$$f_\alpha(A^*A) = (A^*A + \alpha I)^{-1}. \tag{2.12}$$

Segue que a escolha de x_α^δ, da forma

$$x_\alpha^\delta = (A^*A + \alpha I)^{-1}A^*y^\delta \tag{2.13}$$

é uma solução regularizada, definida via equação linear

$$(A^*A + \alpha I)x_\alpha^\delta = A^*y^\delta \tag{2.14}$$

Esta equação 2.14, chamada de Regularização de Tikhonov, pode ser pensada como uma regularização para as equações do tipo 2.8

Vejamos no exemplo abaixo como esta regularização se pronuncia.

Suponha que A seja um operador linear e compacto entre espaços de Hilbert com um sistema singular dado por (σ_j, e_j, f_j). A solução regularizada x_α^δ na equação 2.13 tem a forma

$$x_\alpha^\delta = \sum_{j=1}^{\infty} \frac{\sigma_j}{\sigma_j^2 + \alpha} < y^\delta, f_j > e_j. \tag{2.15}$$

Temos ainda, que $x^\delta = g(A^*A)A^*y^\delta$ satisfaz

$$x^\delta = \sum_{j=1}^{\infty} \frac{1}{\sigma_j} < y^\delta, f_j > e_j. \tag{2.16}$$

Note que, comparando as equações 2.15 e 2.16, podemos observar o resultado de estabilidade da equação 2.15, onde o erro $< y^\delta, f_j >$ é propagado com um fator de $\frac{\sigma_j}{\sigma_j^2+\alpha}$ que é sempre limitado quando $j \to 0$. De forma semelhante podemos observar a equação 2.16. Porém, determinar um sistema de um operador é uma tarefa muito custosa e uma alternativa é determinar uma solução pelo teorema abaixo que trata de uma versão variacional da regularização de Tikhonov.

Teorema 2.5. *Seja x_α^δ como na equação 2.13. Então x_α^δ é o único minimizador do funcional de Tikhonov*

$$J_\alpha(x) := ||Ax - y^\delta||^2 + \alpha||x||^2 \tag{2.17}$$

Demonstração: Disponível em [13], página 78.

Minimização em 2.17 é um compromisso entre minimizar a norma do resíduo $||Ax - y^\delta||$ e tomar o tamanho do termo de penalização $||x||$ pequeno e, assim, forçar a estabilidade. O parâmetro α no funcional é o parâmetro de regularização.

2.3.1.2 Semi-Convergência

A definição da solução regularizada, pela minimização do funcional de Tikhonov 2.17, nos fornece diretamente resultados de convergência e estabilidade, como:

Teorema 2.6. *Seja x_α^δ como na equação 2.13, $y \in Im(A)$ com $||y - y^\delta|| \leqslant \delta$. Se $\alpha := \alpha(\delta)$ é tal que*

$$\lim_{\delta \to 0} \alpha(\delta) = 0 \; e \; \lim_{\delta \to 0} \frac{\delta^2}{\alpha(\delta)} = 0, \tag{2.18}$$

então

$$\lim_{\delta \to 0} x_{\alpha(\delta)}^\delta = A^\dagger y \tag{2.19}$$

Demonstração: Disponível em [13], páginas 79 e 80.

2.3.1.3 Taxas de convergência

Segue da definição de solução regularizada pelo método de Tikhonov que

$$||x_\alpha^\delta - x^\dagger|| \leqslant \sup_{\lambda \in \Sigma(A)} |\lambda f_\alpha(\lambda)|||y - y^\delta|| \leqslant \frac{\delta}{\sqrt{\alpha}}. \tag{2.20}$$

Assim, se $\alpha \sim \delta$, obtemos a seguinte ordem de convergência

$$||x^\dagger - x_\alpha^\delta|| = \mathcal{O}(\sqrt{\delta}) \tag{2.21}$$

Na continuidade, também temos como uma alternativa para a regularização de operadores os métodos iterativos de regularização, como segue na subseção seguinte.

2.3.2 Regularização por Métodos Iterativos para problemas lineares

Os métodos iterativos de regularização têm a vantagem de possuírem propriedades auto-regularizantes. Gauss demonstrou que a melhor maneira de determinar um parâmetro desconhecido de uma equação do tipo 2.1, é minimizar a soma dos quadrados dos resíduos, isto é,

$$\min_{x \in \mathcal{H}_1} \frac{1}{2}||A(x) - y||^2. \tag{2.22}$$

Assumindo algumas propriedades do operador A, podemos provar que o minimizador de 2.22, caso exista, deve satisfazer a condição necessária de primeira ordem

$$A^{'}(x)^* A(x) = A^{'}(x)^* y \tag{2.23}$$

Uma possibilidade para encontrar uma solução de 2.23 é interpretá-la como uma iteração de ponto fixo

$$x_{k+1} = \Phi(x_k) \tag{2.24}$$

para o operador

$$\Phi(x) = x + A^{'}(x)^*(y - A(x)). \tag{2.25}$$

Muitos métodos iterativos para resolver 2.1 são baseados na solução da equação normal 2.23, via sucessivas iterações, partindo de um chute inicial x_0. Destacaremos aqui, o método de Landweber, um dos métodos do tipo gradiente, trazendo alguns teoremas e lemas que estão demonstrados de forma clara no capítulo 6 da referência [13].

2.3.2.1 *Método de Landweber*

Uma maneira de resolvermos a equação 2.23 consiste em considerarmos que o operador A é linear e limitado e a iteração

$$x_{k+1} = x_k + \gamma A^*(y - Ax_k), \quad k = 0, 1, 2, ... \tag{2.26}$$

em que $||A||^{-2} \geqslant \gamma > 0$ é um parâmetro de relaxação, de forma que a iteração tenha a propriedade de descida.

No caso de dados com ruídos y^δ, denotando as iterações por x_k^δ, chegamos a iteração de Landweber

$$x_{k+1}^\delta = x_k^\delta + A^*(y^\delta - Ax_k^\delta). \tag{2.27}$$

2.3.2.2 *Convergência*

Nesta subseção, provaremos que a sequência de iterados pelo método de Landeweber $\{x_k\}$ converge para a solução aproximada do problema inverso 2.1.

Começaremos denotando a solução de quadrados mínimos com norma mínima para o problema 2.1 como $x^\dagger := A^\dagger y$ e também dando condições necessárias e suficientes para a iteração 2.27 convergir.

Teorema 2.7. *Se $y \in \mathcal{D}(A^\dagger)$, então a sequência x_k gerada pela iteração de Landweber 2.26 converge para $x^\dagger = A^\dagger y$ quando $k \to \infty$. Se $y \notin \mathcal{D}(A^\dagger)$, então $||x_k|| \to \infty$ quando $k \to \infty$.*

Do teorema 2.7, temos que a sequência $\{x_k\}$, gerada pela iteração de Landweber, converge para a solução de quadrados mínimos da equação 2.1 quando $y \in \mathcal{D}(A^\dagger)$. Temos também, que x_k^δ diverge, pois em geral os dados perturbados y^δ são tais que $y^\delta \notin \mathcal{D}(A^\dagger)$. Sendo assim, destacaremos também a propagação destes erros.

Lema 2.8. *Sejam y, y^δ com $||y - y^\delta|| \leqslant \delta$ e x_k e x_k^δ obtidos pelas respectivas iterações de Landweber 2.26 e 2.27. Então,*

$$||x_k - x_k^\delta|| \leqslant \sqrt{k}\delta, \quad k \geqslant 0 \tag{2.28}$$

Logo, se há a presença de erro nos dados, temos uma estimativa que é fundamental para iteração de Landweber, que segue

$$||A^\dagger y - x_k^\delta|| \leqslant ||A^\dagger y - x_k|| + ||x_k - x_k^\delta||. \tag{2.29}$$

Tal estimativa, mostra que o erro total consiste em um erro de aproximação que diminui lentamente e um erro dos dados que cresce na ordem de no máximo $\sqrt{k}\delta$. Sendo

assim, para valores pequenos de k o erro nos dados é despresível e a iteração parece convergir para a solução exata $A^{\dagger}y$. Porém, quando $\sqrt{k}\delta$ atinge a magnitude da ordem do erro de aproximação, o erro propagado nos dados torna-se grande. Assim, para a resolução de problemas mal postos, a propriedade de regularização depende de um critério de parada que detecte a transição entre convergência e divergência do método.

Uma alternativa para a escolha do critério de parada é o princípio da discrepância. Este princípio diz que a iteração é parada no índice $k_* = k(\delta, y^{\delta})$ quando, pela primeira vez,

$$||y^{\delta} - Ax_{k(\delta,y^{\delta})}|| \leqslant \tau\delta, \quad \tau > 2 \text{ fixo.} \tag{2.30}$$

E para garantirmos que enquanto a discrepância não é atingida, a aproximação para a solução não piore, segue o teorema.

Teorema 2.9 (Monotonia)**.** *Sejam $y \in \mathcal{D}(A)$, $x^{\dagger}$ a solução de norma mínima de 2.1 e y^{δ} satisfazendo $||y - y^{\delta}|| \leqslant \delta$. Se 2.30 é satisfeita, então*

$$||x_{k+1}^{\delta} - x^{\dagger}|| \leqslant ||x_k^{\delta} - x^{\dagger}||. \tag{2.31}$$

Sendo assim, no caso de dados corrompidos por ruídos, a iteração de Landweber deve ser parada após uma quantidade finita de passos.

Estes métodos apresentados, nos possibilitam determinar uma boa solução para um problema inverso, uma solução aproximada com uma taxa de erro aceitável.

3 PROBLEMAS INVERSOS E MODELOS PROBABILÍSTICOS

O objetivo deste capítulo é relacionar a teoria de problemas inversos e métodos de regularização que apresentamos brevemente no capítulo anterior com a teoria de probabilidade. Embora a relação que será feita aqui é bastante superficial, cremos que seja o suficinte para motivar a utilização de métodos probabilísticos, mais especificamente da fórmula de Bayes para tratar o problema inverso de nosso interesse no restante deste trabalho.

3.1 Modelos Probabilísticos

As variáveis aleatórias, relacionadas a grande parte dos experimentos, podem ser representadas por famílias de distribuições estatísticas, conhecidas por modelos probabilísticos. Estes modelos possuem propriedades que são, basicamente, funções de poucos parâmetros. Na literatura, existem diversas distribuições estatísticas, contínuas e discretas. Dentre as distribuições contínuas que apresentam um grande grau de importância, destacamos: a Uniforme, a Normal, a Qui-quadrado, a T de Studant, a Gama e a Exponencial. Já as distribuições discretas, temos: a Bernoulli, a Binomial, a Poisson e a Geométrica. Estas são as distribuições que trataremos resumidamente neste capítulo e maiores detalhes acerca das mesmas podem ser encontradas em [19][24].

Definição 3.1 (Função de Distribuição de Probabilidades)**.** *Chama-se função de distribuição de probabilidades de uma variável aleatória X, a uma função $F : \mathbb{R} \to [0, 1]$, tal que:* $\forall x \in \mathbb{R}, F(x) = P(X \leq x) = P(\{w \in \Omega : X(w) \leq x\})$.

Propiedades da Função Distribuição:

1. $\lim_{x\to\infty} F(x) = 1, \lim_{x\to-\infty} F(x) = 0$;
2. $F(x)$ é contínua à direita;

3. $F(x)$ é não decrescente $\forall a,b \in \mathbb{R}, a < b$, $F(a) \leqslant F(b)$;

4. $\forall a, b \in \mathbb{R}, a < b, P(a < X \leq b) = F(b) - F(a)$;

5. $\forall a \in \mathbb{R}, P(X = a) = F(a) - \lim_{h \to 0} F(a - h)(h > 0)$

Definição 3.2 (Função Densidade de Probabilidade)**.** *Chama-se função densidade de probabilidade da variável aleatória discreta X, a função $f(x)$ que atenda às seguintes condições:*

1. $f(x) \geqslant 0$*, para* $a < x < b$*;*

2. $\int_a^b f(x)dx = 1$*, onde* a *e* b *podem ser, respectivamente,* $-\infty$ *e* ∞*.*

3.1.1 Distribuições Contínuas

3.1.1.1 Distribuição Uniforme

Uma variável aleatória X tem distribuição uniforme contínua no intervalo $[a, b], a < b$, se sua função densidade de probabilidade é dada por:

$$f(x) = \begin{cases} \frac{1}{b-a}, \ a \leqslant x \leqslant b; \\ 0, \ \text{caso contrário.} \end{cases} \tag{3.1}$$

Usaremos a notação $X \sim U[a, b]$ para indicar que X segue o modelo uniforme contínuo no intervalo considerado.

3.1.1.2 Distribuição Normal

Uma variável aleatória contínua X tem distribuição Normal com parâmetros μ e σ^2, se sua função densidade é dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ para \ -\infty < x < \infty. \tag{3.2}$$

Usaremos a notação $X \sim N(\mu, \sigma^2)$ para indicar que X tem distribuição Normal com parâmetros μ e σ^2.

3.1.1.3 Distribuição Qui-quadrada

Uma variável aleatória contínua X segue uma distribuição de um Qui-quadrado, com k graus de libertadade, se a função densidade de probabilidade é:

$$f(x) = \begin{cases} \frac{e^{\frac{-x}{2}} x^{\frac{k}{2}-1}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}, \ x > 0; \\ 0, outros \ casos. \end{cases} \left(\alpha > 0, k \in \mathbb{N} \right) \tag{3.3}$$

Usaremos a notação $X \sim \chi^2_{(k)}$ para indicar que X segue uma distribuição Qui-quadrada de k graus de libertadade.

3.1.1.4 Distribuição T de Studant

Uma variável aleatória contínua X segue uma distribuição T de Studant, com k graus de libertadade, se a função densidade de probabilidade é:

$$f(x) = \begin{cases} \frac{\Gamma(\frac{(k+1)}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}(1+\frac{x^2}{k})^{-\frac{k+1}{2}},\ x \in \mathbb{R}; \\ 0,\ outros\ casos. \end{cases} \quad (k \in \mathbb{N}) \tag{3.4}$$

Usaremos a notação $X \sim t_{(k)}$ para indicar que X segue uma distribuição T de Studant de k graus de libertadade.

3.1.1.5 Distribuição Gama

Uma variável aleatória contínua X segue uma distribuição Gama, de parâmetros $ne\alpha$, se tem função densidade de probabilidade da forma:

$$f(x) = \begin{cases} \frac{\alpha^n e^{-\alpha x} x^{n-1}}{\Gamma(n)},\ x > 0; \\ 0,\ outros\ casos. \end{cases} \tag{3.5}$$

onde:

- $(\alpha > 0, n \in \mathbb{N})$;
- $\Gamma(n) = \int_0^\infty e^{-x}x^{n-1}dx$;
- $(\Gamma(n) = (n-1)!)$.

Usaremos a notação $X \sim G(n, \alpha)$ para indicar que X segue uma distribuição Gama de parâmetros n e α.

3.1.1.6 Distribuição Exponencial

Uma variável aleatória contínua X, assumindo valores não negativos, segue o modelo Exponencial com parâmetro $\alpha > 0$ se sua densidade é:

$$f(x) = \begin{cases} \alpha e^{-\alpha x},\ x \geq 0; \\ 0,\ outros\ casos. \end{cases} \tag{3.6}$$

Usaremos a notação $X \sim Exp(\alpha)$ para indicar que X tem distribuição Exponencial de parâmetro α.

3.1.2 Distribuições Discretas

3.1.2.1 Distribuição de Bernoulli

Uma variável aleatória discreta X, definida por:

$$X = \begin{cases} 1, \text{se A ocorre;} \\ 0, \text{se A não ocorre.} \end{cases}$$

sendo A um evento qualquer, segue uma distribuição de Bernoulli, se tem função de probabilidade tal que:

$$f(x) = P[X = x] = \begin{cases} p^x q^{1-x}, \ se\ x = 0, 1; \\ 0, \ outros\ casos. \end{cases} , (0 < p < 1, p + q = 1). \tag{3.7}$$

Usaremos a notação $X \sim B(1, p)$ para indicar que X segue uma distribuição de Bernoulli.

3.1.2.2 Distribuição Binomial

Uma variável aleatória discreta X, segue uma distribuição Binomial de parâmetros n e p se sua função de probabilidade é da forma:

$$f(x) = P[X = x] = \begin{cases} \binom{n}{x} p^x q^{n-x}, \ se\ x = 0, 1, 2, ..., n; \\ 0, \ outros\ casos. \end{cases} , (0 < p < 1, p + q = 1). \tag{3.8}$$

Usaremos a notação $X \sim B(n, p)$ para indicar que X segue uma distribuição Binomial de parâmetros n e p.

3.1.2.3 Distribuição Poisson

Uma variável aleatória discreta X, segue uma distribuição de Poesson de parâmetro λ se sua função de probabilidade é da forma:

$$f(x) = \begin{cases} e^{-\lambda} \frac{\lambda^x}{x!}, \ x = 1, 2, ...; \\ 0, \ outros\ casos. \end{cases} , (\lambda > 0). \tag{3.9}$$

Usaremos a notação $X \sim Pois(\lambda)$ para indicar que X segue uma distribuição de Poesson de parâmetro λ.

3.1.2.4 Distribuição Geométrica

Uma variável aleatória discreta X, segue uma distribuição Geométrica (ou de Pascal) de parâmetro p se sua função de probabilidade é da forma:

$$f(x) = P[X = x] = \begin{cases} pq^{x-1}, x = 1, 2, 3, ...; \\ 0, outroscasos. \end{cases} , (0 < p < 1, p + q = 1). \quad (3.10)$$

Usaremos a notação $X \sim BN(1, p)$ para indicar que X segue uma distribuição Geométrica de parâmetro p.

Na biologia quantitativa, muitos modelos encontrados são melhores descritos por modelos probabilísticos. Os sistemas dinâmicos ativados termicamente, por exemplo, são muitos difundidos e por causa disso muitos sistemas biológicos exibem comportamento estocástico, comportamentos que não conseguimos determinar um padrão, mas os mesmos tem origem em eventos aleatórios, ao realizarmos experimentos de descompactação do DNA observaremos tal comportamento [1] [2] [3] [5] [14].

Existem também, muitos sistemas biológicos que apresentam um número de variáveis que podem ser tratadas como aleatórias, porque elas são modeladas para um conjunto possíveis de realizações e não mudam durante o experimento. Um modelo probabilístico assegurará uma probabilidade para a resposta de um experimento. E com isso, o problema inverso é de interesse, pois podemos garantir uma probabilidade para um modelo ou um conjunto de parâmetros dado a resposta de um experimento.

3.2 Teorema de Bayes

Nesta seção vamos olhar para o Teorema de Bayes, o qual foi desenvolvido por Thomas Bayes e publicado em 1763 [23] como uma ferramenta para determinar uma solução regularizada para problemas inversos. Para tal, considere o par de vetores x e y e represente por $P(x)$, $P(y)$ e $P(x, y)$ as funções distribuição de probabilidade do evento $x \in [x, x + dx]$ ser $P(x)dx$, do evento $y \in [y, y + dy]$ ser $P(y)dy$ e do evento $(x, y) \in [x, x + dx, y, y + dy]$ ser $P(x, y)dxdy$, respectivamente. Finalmente, assuma que $P(y|x)$ representa a função de distribuição de probilidade para determinar y dado que x é conhecido.

Assim, $P(x, y)dxdy$ pode ser escrita de forma equivalente á:

$$P(x, y)dxdy = P(x)dxP(y|x)dy \qquad ou \qquad P(x, y)dxdy = P(y)dyP(x|y)dx \quad (3.11)$$

Igualando os dois lados na equação (3.11), obtemos o Teorema de Bayes

$$P(x \mid y) = \frac{P(y \mid x).P(x)}{P(y)} \quad (3.12)$$

3.3 Inferência Bayesiana

Nesta seção, destacaremos o que é a Inferência Bayesiana, também como ela pode ser aplicada e suas características. A Inferência Bayesiana consiste na aplicação iterativa do teorema de Bayes, com a finalidade de atualizar o que conhecemos sobre as variáveis aleatórias envolvidas no problema, que podem ser um parâmetro do modelo em questão, por exemplo. Essa não é a única forma de uma inferência estatística, mas ela apresenta algumas características que fazem com que ela seja mais util que outras técnicas, um exemplo temos a inferência frequentista, onde a frequência é vista como uma probabilidade.

Uma vantagem da inferência Bayesiana é que esta retornará uma distribuição de probabilidade, que em geral, contém muito mais informação que um valor inferido e um intervalo de confiança. Em contra partida, esta pode depender fortemente da escolha de uma distribuição *a priori* que nem sempre pode ser uma escolha natural, [2].

Sendo um pouco mais formal. Suponha que o evento x seja uma sucessão de eventos mutuamente excludentes (x_1,x_2,x_3,...,x_n), que formam um espaço amostral Ω. Considerando y um evento qualquer, podemos reescrever 3.12 como:

$$P(x_i \mid y) = \frac{P(y \mid x_i).P(x_i)}{\sum_{i=1}^{n} iP(y \mid x_i).P(x_i)} \tag{3.13}$$

Este teorema, nos permite determinar as probabilidades dos diversos eventos x_i que podem ser a causa da ocorrencia do evento y.

Dentre as aplicações existentes, destacamos aqui um exemplo bastante clássico que é o cálculo do número de falsos positivos em testes médicos, maiores detalhes em [2]. Para tal, vamos supor que há uma doença muito rara que ocorre em apenas uma pequena fração ϵ da população. Um teste para a doença, retorna um resultado falso com probabilidade p.

$$\begin{aligned} P(negativo \mid doente) = P(positivo \mid \textbf{\textit{saudável}}) &= p \\ P(positivo \mid doente) = P(negativo \mid \textbf{\textit{saudável}}) &= 1-p \\ P(doente) &= \epsilon \end{aligned}$$

Pelo teorema de Bayes, temos que:

$$P(falso\ negativo) = P(doente \mid negativo) = \frac{p\epsilon}{p\epsilon + (1-p)(1-\epsilon)} \tag{3.14}$$

$$P(falso\ positivo) = P(\textbf{\textit{saudável}} \mid positivo) = \frac{p(1-\epsilon)}{p(1-\epsilon) + (1-p)\epsilon} \tag{3.15}$$

Note que as probabilidades de falsos negativos 3.14 e falsos positivos 3.15 são bastante

diferentes. Acontece que o número de falsos positivos é muito elevado devido a raridade da doença. Sendo possível observarmos que mais da metade de falsos positivos são a menos de uma probabilidade p de ter um resultado impreciso é menor que a prevalência da doença ϵ.

3.4 Fórmula de Bayes e Problemas Inversos

Nesta seção veremos como a fórmula de Bayes (3.12) se relaciona com os problemas inversos descritos no Capítulo 2.

Inicialmente, note que $P(x)$ representa a função densidade de probabilidade do vetor de estados x estimada anteriormente as medidas. Desta forma, $P(x)$ representa a informação *a priori* da função densidade de probabilidade. Esta desempenha um papel muito importante na determinação de soluções aproximadas para problemas inversos. Não nos deteremos nos detalhes aqui, mas sugerimos [6] e [7]. Já, $P(y|x)$ é a função densidade de probabilidade de observar y, dado o valor exato x. Assim, $P(y|x)$ está intimamente ligado com o modelagem do problema.

Por outro lado, $P(x|y)$ é a função densidade de probabilidade *a posteriori* para o vetor de estados x dados as medidas y. Note que este é exatamente o objetivo dos problemas inversos. Neste contexto, para obtermos as soluções x para um dado problema inverso de maneira que a resposta seja a " melhor possível" buscamos por um vetor x que maximize $P(x|y)$. De forma mais precisa, desejamos resolver o problema de maximização *a posteriori* (chamada MAP)

$$\max_x P(x|y)\,. \tag{3.16}$$

Observe que como $P(y)$ no denominador de (3.12) independe de x, este é somente um fator de escala. Por simplicidade, podemos reescalonar o problema de forma que $P(y) = 1$.

3.5 Relação entre MAP e regularização de Tikhonov

Nesta seção, formalmente, apresentaremos a relação entre a fórmula de Bayes e a regularização de Tikhonov apresentada na Seção 2.3. Para tal, assuma que sejam conhecidas informações *a priori* x_0 para o problema (2.1). Assuma ainda que saibamos uma variância do erro cometido no propor x_0 como *a priori*, dada por δ_0^2, tal que a estimativa *a priori* da $||x - x_0|| \geq \delta_0$. Por fim, suponha que as medidas sejam tomadas com um erro de variância δ^2, que incorpora os erros de modelagem e do aparelho de medidas, de forma que o problema (2.1) tenha a forma

$$y^\delta = A(p)x \pm \delta\,. \tag{3.17}$$

Desta forma, temos que

$$||y - y^\delta|| = ||A(p)x - y^\delta|| \leq \delta\,.$$

Por simplicidade, assumiremos que o a relação entre y e x é linear, ou seja que $A(p)$ é um operador linear.

Suponha que os erros de distribuição sejam Gaussianos (ver seção 3.1). Assim, temos que

$$P(x) = \frac{1}{\delta_0\sqrt{2\pi}} \exp -\frac{||x - x_0||^2}{2\delta_0^2} \tag{3.18}$$

e

$$P(y|x) = \frac{1}{\delta\sqrt{2\pi}} \exp -\frac{||y - A(p)x||^2}{2\delta^2}\,. \tag{3.19}$$

Aplicando a fórmula de Bayes (3.12) (lembrando que estamos assumindo que $P(y) = 1$ haja visto que este independe de x), temos que

$$\begin{aligned} P(x|y) = &\quad \frac{1}{\delta_0\sqrt{2\pi}} \exp -\frac{||x - x_0||^2}{2\delta_0^2} \frac{1}{\delta\sqrt{2\pi}} \exp -\frac{||y - A(p)x||^2}{2\delta^2} \\ = &\quad \frac{1}{2\pi\delta\delta_0} \exp -\frac{||y - A(p)x||^2}{2\delta^2} - \frac{||x - x_0||^2}{2\delta_0^2}. \end{aligned} \tag{3.20}$$

Desta forma, o problema de maximização (3.16) (ou mais precisamente logMAP) é dado por

$$\begin{aligned} \max_x \log P(x|y) &= \max_x \log \left(\frac{1}{2\pi\delta\delta_0} \exp -\frac{||y - A(p)x||^2}{2\delta^2} - \frac{||x - x_0||^2}{2\delta_0^2} \right) \\ &= \max_x \left(C - \frac{||y - A(p)x||^2}{2\delta^2} - \frac{||x - x_0||^2}{2\delta_0^2} \right)\,. \end{aligned}$$

Tomando $J_\alpha(x) = ||y - A(p)x||^2 + \alpha||x - x_0||^2$ o funcional de Tikhonov com $\sqrt{\alpha} = \alpha\left(\frac{\delta}{\delta_0}\right)$ e observando que $max - J_\alpha(x)$ é equivalente a $\min J_\alpha(x)$ ([25] e [18]) temos que, encontrar x que maximize $P(x|y)$ é equivalente a encontrar x que minimize o funcional de Tikhonov $J_\alpha(x)$.

Desta forma, toda a teoria de regularização descrita no Capítulo 2 na seção 2.3 pode ser aplicada, utilizando a fórmula de Bayes.

Observação 3.3. *A equivalência feita acima entre a hipótese de que a distribuição dos erros nos dados seja Gaussiana e da minimização de um funcional de Tikhonov com erros quadráticos pode ser generalizada para outras distribuições de probabilidade descritas anteriormente. Nós nos ateremos a estes casos. Interessados podem consutar [9] e [6].*

Com o exposto acima, podemos anunciar os seguintes teoremas:

Teorema 3.4 (Existência)**.** *Dada qualquer escolha da distribuição de probabilidade a priori* $P(x)$ *tal que a esta esteja associada a uma escolha do parâmetro de regularização* $\alpha = \alpha(\frac{\delta}{\delta_0}) > 0$*. Então existe uma solução do problema* (3.16)*.*

Demonstração: Segue diretamente da equivalência entre a minimização de J_α (que exite pela teoria clássica de otimização [15] [25]) e uma solução de (3.16) derivada acima.

Teorema 3.5 (Convergência e estabilidade)**.** *Suponha que escolha da distribuição de probabilidade a priori* $P(x)$ *seja feita de forma que gere uma escolha do parâmetro de regularização* $\alpha = \alpha(\frac{\delta}{\delta_0})$ *satisfazendo* $\alpha \to 0$ *e* $\alpha/\delta^2 \to 0$ *para o funciional de Tikhonov, então a solução aproximada, aqui denotada por* x_α^δ*, de* (3.16) *é uma solução regularizada para o problema 2.1.*

Demonstração: Segue diretamente das propriedades de regularização dos minimizadores do funcional de Tikhonov com a escolha de α de forma apropriada e da equivalência entre o problema (3.16) e a minimização do funcional J_α.

3.5.1 Inferência Bayesiana e Tikhonov iterado

Faremos aqui um breve comentário a respeito das propriedades de regularização ao utilizarmos a inferência Bayesiana como um método de identificação. Em outras palavras, suponha dado x_k e considere

$$x_{k+1} = argmax P(x|y) = argmax P(x_k)P(y|x) \tag{3.21}$$

onde utilizamos a fórmula de Bayes.

As vantages de considerar (3.21) é que a informação *a priori* $P(x_k)$ é atualizada a cada iteração.

Fazendo a mesma análise como na seção 3.5 obtemos que x_{k+1} em (3.21) pode ser visto como

$$x_{k+1} = arminJ_\alpha(x : x_k) \tag{3.22}$$

onde

$$J_\alpha(x : x_k) = ||A(p)x - y||^2 + \alpha||x - x_k||^2$$

é o funcional de Tikhonov-iterado considerado em [8] e [4]. Como o método de Tikhonov iterado é de fato um método iterativo, os resutlados de convergência e estabilidade seguem, de maneira muito próximas, aos resultados apresentados para o método de Landweber na Subseção 2.3.2. Para detalhes sobre as propriedades de regularização do método de funcional de Tikhonov iterado, veja [8], [4] e referências.

4 MODELAGEM DO PROBLEMA DIRETO: DESCOMPACTAÇÃO DO DNA

Como já mencionado anteriormente, é importante e fundamental na biologia o estudo da descompactação das cadeias e o sequenciamento da molécula de DNA, 1. Neste capítulo, nos deteremos em descrever o problema direto de descompactação do DNA aplicando uma força fixa. A apresentação que faremos aqui está baseada em estudos anteriores, que podem ser consultados nas referências [1, 2, 3].

É importante observar que o trabalho de descompactar as cadeias polímeras que compõem o DNA dependem, primeiramente, da sequência, porque os pares de bases GC são mais estáveis do que os de AT, por exemplo [1, 2, 3]. Segundo, a estrutura de hélice dupla apresenta empilhamentos de pares de bases e isso acarreta interações entre os pares vizinhos. Ambos os fenômenos devem ser considerado na descompactação do DNA.

Para descompactar uma molécula de DNA, separar as ligações de hidrogênio, existem diversos experimentos e modelos, mas, neste trabalho vamos nos deter a separação por aplicação de uma força mecânica. Para entendermos este processo, iniciaremos definindo o que podemos chamar de problema direto (descompactação) e nas seções seguintes caracterizar a descompactação utilizando a hipótese de que ela será realizada por um processo de força mecânica.

4.1 O problema direto

Nesta seção, descreveremos de forma mais clara o problema direto que consiste na descompactação das cadeias polímeras da molécula de DNA. Por sua vez, na modelagem deste problema deve ser levado em consideração características como extensão, energia livre e a elasticidade.

Em uma molécula de DNA, ao se aplicar uma determinada força, de qualquer natureza, nas cadeias polímeras, cadeias de açúcar-fosfato, em condições e intensidade propícias é possível romper as ligações que são formadas pelos pares de bases. A aplicação desta tal força, resulta em duas cadeias polímeras, além de poder ser observado (medido) características como: a energia livre resultante do rompimento desses pares de

bases e a elasticidade de cada cadeia, sinais estes que são fundamentais e auxiliarão para resolução do problema inverso, o sequenciamento.

4.2 Modelando o problema direto

Nesta seção, vamos descrever matematicamente modelos teóricos para a elasticidade e para o processo de descompactação do DNA. Apesar de se conhecer e compreender a estrutura de uma cadeia polímera e sua elasticidade, existem estudos atuais acerca de modelos que propiciam uma entendimento maior com relação as características bastante específicas, como controle e medições quantitativas, que não se conhecia com as técnicas mais antigas. Descrevemos aqui, modelos padrão da física de polímeros, direcionado a compreensão biopolímera e elasticidade do DNA.

4.2.1 Modelando a elasticidade

Para obter as propriedades físicas do polímero, foram construídos diversos modelos hipotéticos que quando aplicados, por exemplo, em cadeias curtas apresentam certas limitações. Para este problema, vamos considerar que uma cadeia hipotética consiste em n ligações, que podem ser representadas por vetores ligando $i-1$ e i. Um conjunto de vetores t_i (vetores que conectam a ligação $i-1$ a i) acerca do mesmo pode ser constituído. Veja Figura 11. Note que, uma possível interpretação dos vetores t_i podem ser obtidos com os vetores tangentes (sob uma certa normalização) de uma curva suave no espaço, passando pelos pontos que representam as ligações. Assim, a representação que vamos construir aqui pode ser vista como uma aproximação discreta de um modelo contínuo para esta dinâmica.

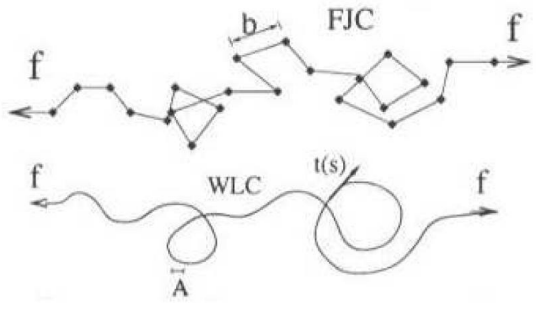

Figura 11: Representação esquemática de modelos de polímero. Também disponível em [1]

Para esta modelagem, uma das quantidades mais importante é a distância entre os extremos dos polímeros, que pode ser dado pela equação abaixo:

$$R = \sum_i t_i \tag{4.1}$$

Geralmente, a magnitude da escala do R que é interessante. Esta pode ser calculada pelo quadrado da distância entre extremos, ou seja, o produto escalar de R com R, dado por:

$$R^2 = \sum_{ij} t_i t_j \tag{4.2}$$

Cadeia livremente articulada (FJC)

É o modelo mais simples para descrever uma cadeia polímera simples. Consiste em um modelo polímero simples descrito como uma cadeia de N segmentos $t_i = b\hat{t}_i$ de comprimento fixo b. É assumido ainda que as juntas podem se mover livremente em qualquer direção independentemente das demais. Portanto os vetores são não-correlacionados. Por isso, este modelo pode ser visto como um modelo de um caminho totalmente randômico. Assim, matematicamente, a distribuição de probabilidade é dada pela equação abaixo:

$$P(t_i) = \frac{1}{4\pi b^2}\delta(|t_i| - b) \tag{4.3}$$

A constante $\frac{1}{4\pi b^2}$ fornece a correção de normalização em um espaço tridimensional e a função δ impõem que o comprimento da ligação seja exatamente b.

Sob estas hipóteses, a média da distância ponto-a-ponto do polímero deve ser igual a zero, pois, não há uma direção preferencial. Ou seja,

$$< t_i >= \int t_i P(t_i) dt_i = 0 \tag{4.4}$$

Que por sua vez, o quadrado da média da distância é dada por:

$$\begin{aligned} < R^2 > &= < \sum_i \sum_j t_i t_j > & (4.5)\\ &= \sum_{ij} < t_i, t_j > & (4.6)\\ &= \sum_i t_j^2 + \sum_{i \neq j} < t_i t_j > & (4.7)\\ &= Nb^2 \,. & (4.8) \end{aligned}$$

Portanto, a distribuição de probabilidade dos extremos da cadeia, R, que é dada por

$$P(R) = \int \prod_i [P(t_i)dt_i]\delta(R - \sum_i t_i)\,. \tag{4.9}$$

Note que, a existência da distribuição δ na expressão acima nos leva a concluir que temos que avaliar a integral sobre todas as possibilidades de junções e, dentre estas, escolher somente as que possuem distância entre extremos R.

Calculando a distribuição de probabilidade em (4.9), obtemos

$$P(R) = \Big(\frac{3}{2\pi b^2 N}\Big)^{\frac{3}{2}} e^{\frac{-3R^2}{2b^2 N}}\,. \tag{4.10}$$

Ou seja, a distribuição de probabilidade (considerando este modelo) é uma gaussiana (ver seção 3.1), cuja dependência é apenas do módulo da distância entre extremos R.

Observação 4.1. *Da equação* (4.10) *podemos deduzir o seguinte: Suponha que a primeira ligação esteja localizada na origem do sistema de coordenadas que estamos usando. Assim,* (4.10) *determina qual a probabilidade de encontrar o outro extremo da cadeia que está a uma distância R.*

Além disso, o máximo de $P(R)$ acontece quando R é o menor possível, ou seja, quando o final da sequência está exatamente na próxima ligação. Essa informação é importante, pois nos fornece uma primeira motivação para usar um método iterativo do tipo proposto na Subseção 3.5.1 como um método para usar a maximização da probabilidade.

Esticando uma FJC

Agora estudaremos o que acontece se aplicarmos uma força externa (constante) ao modelo FJC descrito acima. Neste caso, ao aplicamos uma força externa a cadeia é esticada, fazendo com que as ligações se alinhem ao longo da direção da força, como acontece com um par de partículas carregadas eletricamente, com cargas distintas em um campo elétrico. A energia liberada desse processo (a uma força constante) deve ser diretamente proporcional a força aplicada. Em outras palavras, é dado pela equação:

$$H_{FJC} = E = -fR = -f\sum_i t_i \tag{4.11}$$

Multiplicando cada fator por e^{-E/k_BT}, que é fator de Boltzmann, responsável por expressar a probabilidade de um estado de energia E com relação a um estado com energia zero, e calculando a média, sob uma sequência com N juntas, temos

$$< z >_{FJC} = Nb\Big[coth\Big(\frac{fb}{k_BT}\Big) - \frac{k_BT}{fb}\Big] \tag{4.12}$$

Para forças pequenas, a cadeia comporta-se como uma mola de rigidez $\frac{3k_BT}{b^2N}$ e comprimento restante zero, dado por:

$$< z >_{FJC}= \frac{Nb^2}{3k_BT}f \tag{4.13}$$

Já para forças grandes, temos

$$< z > \to bN\Big(1 - \frac{k_BT}{fb}\Big) \tag{4.14}$$

e assim, a força necessária para destender a molécula diverge quando o tamanh Nb é alcançado.

ssDNA

Depois de descrever como algumas cadeias polímeras se comportam, passaremos a modelar o comportamento de uma fita simples de DNA (single stand DNA, denotado por ssDNA).

Como o ssDNA é uma cadeia muito fina e flexível, esta é frequentemente descrito por um modelo de cadeia livremente articulada FJC. No entanto, para forças de grande intesidade, resultados experimentais mostram que um modelo como em (4.14) não é o suficiente para modelar a cadeia de ssDNA. De fato, numa cadeia como o ssDNA tem que levar em consideração auto-atrações de algumas de suas ligações, bem como repulsões causadas pelos elementos consistentes das mesmas ligações. Outros fatores que devem ser levados em conta são as concentrações dos solutos onde a cadeia está imersa. Veja figura 2.2 na página 32 da Tese de Baldazzi [1]. De qualquer maneira, tipo de paridade das bases, a iteração entre as bases no processo de esticamento e a elasticidade intíncica dos segmentos devem ser levados em consideração.

Excluindo os efeitos do volume, Zhang et al. [29] propuseram que a energia para esticar uma cadeia de ssDNA deve ser regida pela equação:

$$H_{ssDNA} = H_{mFJC} + \sum_{i=0}^{N_p} V_p + \frac{v^2}{E_W}\int ds_i ds_j \frac{exp\Big(-\frac{|r_i - r_j|}{l_n}\Big)}{|r_i - r_j|} \tag{4.15}$$

onde:

- $H_{mFJC} = \frac{Y_{ss}}{2}\sum_{i=1}^{N}\Big(|r_i - r_j| - b\Big)^2 - fz_N$ é uma modificação de FJC descrito anteriormente, incorporando segmentos que podem esticar (com módulo de Young Y_ss, ou contrário da rigides assumida no modelo FJC. O termo
- $\sum_{i=0}^{N_p} V_p$ representa o emparelhamento de base, já N_p é o número de nós emparelhados e V_p é o potencial médio de emparelhamento;

- O último termo, representa a repulsão eletrostática entre o segmento de DNA, onde v é a densidade de carga eficaz e ϵ_W é a constante dielétrica da água.

O sistema que descrevemos, pode ser simulado com um procedimento de Monte Carlo. Devido as características já citadas e as variações que as mesmas sofrem, o mFJC representa um bom modelo para a elasticidade de um DNA de cadeia simples, quando restrita de 10 a 70 pN. Assim, descrevemos a distância entre extremidades da molécula pela expressão

$$< z >_{mFJC} = L_{ss}\Big[coth\Big(\frac{fb}{k_BT}\Big) - \frac{K_BT}{fb}\Big]\Big(1 + \frac{f}{y_{ss}}\Big) \tag{4.16}$$

onde, L_{ss} representa o comprimento do contorno da molécula e S é o módulo de alongamento de DNA_{ss}.

Modelo Kraty-Porod (KP)

No modelo FJC não existe restrição quanto aos ângulos entre duas ligações adjacentes e assim, são permitidas curvas acentuadas. Por outro lado, polímeros possuem rigidez específicas que fazem com que flexão acentuadas sejam custosas energeticamente. Tal característica pode ser assumida no modelo, simplesmente incorporando a quantidade de energia

$$\frac{H_{KP}}{K_BT} = -a\sum_i \hat{t}_i\hat{t}_{i+1} \tag{4.17}$$

para o ângulo entre dois segmentos adjacentes na cadeia. Na equação anterior, $\hat{t}_i$ é o versor do vetor t_i e a é uma constante que descreve a rigidez da molécula, onde $a >> 1$ (muito rigido) e $a < 1$ e (muito flexível).

Devido a existência dessa inflexibilidade de direção entre diferentes ligações implica que a direção de uma dada ligação influência a ligação seguinte, fazendo com que elas sejam, agora, relacionadas. Para determinar os efeitos nos pares que são influenciados, faz-se interessante calcular o alcance efetivo $|i - j|$ no qual as mudanças são sentidas. Para tal, podemos determinar a correlação térmica entre os extremos, dada por:

$$< \hat{t}_N, \hat{t}_0 >= \frac{\int d^2t_0...d^2t_N\hat{t}_N.\hat{t}_0P(\hat{t}_0...\hat{t}_N)}{\int d^2t_0...d^2t_NP(\hat{t}_0...\hat{t}_N)} \tag{4.18}$$

onde $P(\hat{t}_0...\hat{t}_N)\alpha\prod_{i=0}^{N-1} e^{a\hat{t}_i\hat{t}_{i+1}}$.

Usando $P(\hat{t}_0...\hat{t}_N)$ como acima, obtemos que

$$< \hat{t}_N, \hat{t}_0 >= e^{Nln\Big[coth(a)-\frac{1}{a}\Big]} \tag{4.19}$$

Como resultado da equação (4.19) concluímos que, para $a > 0$ temos que $coth(a) - \frac{1}{a} < 1$, ou seja, a correlação cai exponencialmente a uma distância

$$l_c = \frac{1}{ln\Big[coth(a) - \frac{1}{a}\Big]} \tag{4.20}$$

Logo, dados dois vetores a uma distância maior ou igual a l_c, estes podem ser considerados como independentes, uma vez que sua correlação mútua tende a zero.

Modelo como cadeia de Worm (WLC)

Este modelo consiste no limite contínuo do módulo de (KP) quando $b \to 0$. Assim podemos escrever a flexão de um polímero em função do seu vetor tangente $\hat{t}(s)$, com s sendo a parametrização, como:

$$E_{WLC} = -K_B T a \sum_{i=1}^{N} \hat{t}_i.\hat{t}_{i+1} = -\frac{K_B T a b}{2} \sum_{i=1}^{N} b\Big(\frac{\hat{t}_i - \hat{t}_{i+1}}{b}\Big)^2 + constante \tag{4.21}$$

$$\to \frac{B}{2} \int_0^L ds\Big(\frac{d\hat{t}}{ds}\Big)^2$$

cujo limite é tomado de forma que ab seja constante quando $b \to 0$. Assim, a deve ir para infinito. As constantes elasticas a e B são relacionadas por $B = K_B T a b$ e, L é simplesmente Nb.

A correlação 4.19 pode ser reescrita substituindo $ln\Big[coth(a) - \frac{1}{a}\Big]$ por simplesmente $-\frac{1}{a}$, quando $a \to \infty$, resultando em

$$< \hat{t}(s), \hat{t}(s') >= e^{-\frac{K_B T|s-s'|}{B}} = e^{-A|s-s'|} \tag{4.22}$$

onde a constante $A = \frac{B}{k_B T}$, chamada de comprimento de persistência. Esta fornece uma medida da flexibilidade do polímero. Para sermos um pouco mais claro e entendermos melhor o significado da constante A, vamos calcular a distância de dois pontos que estão a separados um do outro por um comprimento L ao longo da curva (agora contínua) que descreve a cadeia polímera. Em termos do vetor tangente, sabemos que este está relacionado com a função comprimento de arco, dado por $\hat{t}(s) = r(L) - r(0) = \int_0^L ds\hat{t}(s)$. Portanto, a média quadrática é simplesmente

$$< |r(L) - r(0)|^2 >= 2AL + 2A^2(e^{-L/A} - 1) \tag{4.23}$$

Deste modo, dados dois pontos que estão mais próximos que o comprimento de

persistência A, temos que $< |r(L) - r(0)|^2 > \approx L^2$. Neste limite, a cadeia polímera não dobra muito. Como consequência, a distância média dos extremos é aproximadamente L. Assim, a cadeia se comporta quase como uma haste rígida e a correlação entre os diferentes segmentos é máxima. Note que, neste sentido, a constante de persistência de comprimento A age como uma informação sobre a distância na qual a cadeia persiste na mesma direção. Isto nos leva a concluir que a rigidez na cadeia possui um efeito favorável para a compreensão da mesma. No caso em que $A >> L$, a cadeia polímera age como se tivesse juntas livres e os pontos podem ser considerados não-correlacionados. Para cadeias longas, i.e. $L >> A$ a distância média quadrática é $2AL$. Portanto, se parece com um caminho aleatorio de $\frac{L}{2A}$ passos, cada um com comprimento $2A$.

Esticando uma cadeia longa modelada por WLC

Veremos aqui que ao esticar um polímero semi-flexível longo, i.e., com $L >> A$, novos termos aparecem na formulação do modelo WLC. Este é um dos indicativos de que, para cadeias longas, as dificuldades de solução do problema de sequenciamento que será estudado no próximo capítulo é mais difícil. Os termos adicionais são os responsáveis por acoplar a força f a distância entre extremos no modelo WLC, dado por:

$$\beta E = \frac{K_B T A}{2} \int_0^L d\Big(\frac{d\hat{t}}{ds}\Big)^2 - f\hat{z}[r(L) - r(0)] \tag{4.24}$$

Podemos ainda reescrever esta última equação como uma integral sobre $\hat{t}$

$$\beta E = \int_0^L \Big[\frac{A}{2}\Big(\frac{d\hat{t}}{ds}\Big)^2 - \beta f \hat{z}.\hat{t}\Big] \tag{4.25}$$

Podemos notar que em (4.25) o hamiltoniano é controlado pelo parâmetro $\beta A f$. Assim, temos dois regimes de forças a serem considerados como casos extremos, por estarem acima e abaixo da força $K_B T/A$.

- **Forças Pequenas -** Neste regime podemos usar uma aproximação de primeira ordem. Uma vez que conhecemos a flutuação para as forças de ordem zero e o média quadrática das distâncias extremas (veja 4.23 para $L >> A$), temos, por simetria,

$$z =< \hat{z}, [r(L) - r(0)] >= \frac{2AL}{3} \tag{4.26}$$

Usando a aproximação de primeira ordem para a força, isto é, somente o termo de

primeria ordem da série de potências em 4.23, que possui três termos, temos:

$$f = \frac{3k_BT}{2AL}z + ... \tag{4.27}$$

- **Forças Grandes -** Neste regime podemos usar a expansão $\frac{1}{\sqrt{f}}$ para calcular o comportamento não linear de um polímero. De fato, suponha que a cadeia polimera já estaja bem esticado. Neste caso, podemos usar a seguinte aproximação o vetor tangente $\hat{t}(s) = \hat{z}t_{||} + u$, onde $u << 1$. Como $\hat{t}$ deve ser unitário (versor), temos que $\hat{t}_{||} = \sqrt{1 - |u|^2} = 1 - \frac{1}{2}|u|^2 + ...$, cuja expansão em Taylor converge. Neste contexto, podemos escrever as flutuações de $|u|^2$ como:

$$< |u(s)|^2 >= \frac{1}{\sqrt{\beta A f}} \tag{4.28}$$

e a extensão, na direção da força, é dada por

$$z = L < t_{||} >= L\Big(1 - \frac{1}{2}|u|^2 + ...\Big) = L\Big(1 - \frac{1}{\sqrt{4\beta A f}}\Big) \tag{4.29}$$

mostrando o comportamento assintótico da forma $\frac{1}{\sqrt{f}}$.

O DNA em hélice dupla (dsDNA)

Usaremos este espaço para comentar alguns pontos de interesse sobre a elasticidade da hélice dupla do DNA, que denotaremos aqui por dsDNA. Se considerarmos o dsDNA em um regime de forças pequenas, então este é bem descrito como um polímero semi-flexível, veja referências em [1]. Comparado com o ssDNA, o dsDNA pode ser considerado bastante rígido e assim, pode ser modelado como um polimero rígido. Um dos modelos mais aceitos para o dsDNA é uma aproximação do modelo WLC visto acima, onde

$$f = \frac{1}{\beta A}\left[\frac{1}{4(1 - z/L)^2} - \frac{1}{4} + \frac{z}{L} - \frac{f}{Y_{ds}}\right] \tag{4.30}$$

onde $Y_{ds} \approx 1000pN$ é uma constante de elasticidade da cadeia de dsDNA.

4.2.2 Modelando o processo de descompactação: Caso Estático

Nesta subseção vamos nos deter ao processo de separação da cadeia de DNA. Nesta modelagem alguns ingredientes devem ser levados em conta. Extensão, a energia livre (que é uma quantidade uni-dimensional) e a elasticidade da molécula, cujos modelos foram descritos brevemente acima.

Só relembrando, teremos que considerar a energia necessária para separar uma molécula de DNA em hélice dupla (dsDNA).

Assumindo que esta molécula possui N pares e denotanto por $s_i = A, C, T, G$ as bases ao longo da cadeia de DNA, temos que a energia gasta para abrir os pares de bases, do primeiro ao n-ésimo (claro que $n < N$), pode ser expresso pela soma das energias necessárias para abrir cada um dos pares de bases, denotado por $g_0(\cdot)$, isto é,

$$G(n) = G_{ds}(n) = \sum_{i \leq n} g_0(s_i, s_{i+1}). \tag{4.31}$$

onde $g_0(s_i, s_{i+1})$ leva em consideração a paridade, bem como os efeitos de alongamento entre as bases vizinhas na cadeia. A figura 4, no capítulo 1, espressa 10 valores para g_0, dependendo da combinação entre as bases i e $i + 1$. Os valores expressos também dependem das concentrações do soluto onde a fita de DNA estiver imersa, entre outros fatores. Veja [1] e referências para maiores detalhes.

Outras quantidades que devem ser levadas em consideração para a modelagem são as energias potenciais elásticas que agem em moléculas sob uma dada tensão. Estas quantidades levam em considerações as ligações entre o ssDNA e o dsDNA, que conectam a molécula a ser aberta ao mecanismo de abertura usado (ver figura 12).

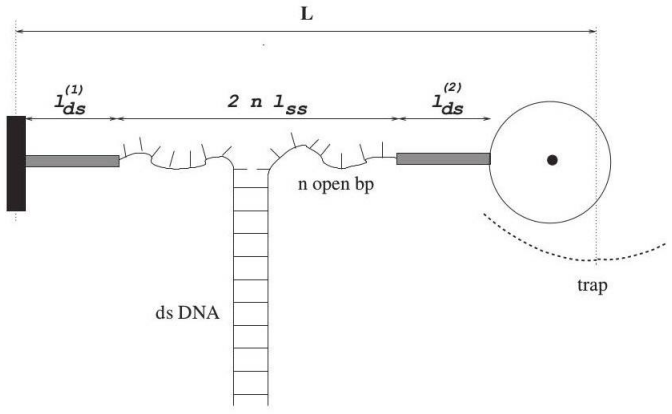

Figura 12: Ilustração do processo de descompactação do DNA. Também disponível em [2]

Na figura 12 temos uma constante de rígidez K_{ss}, comprimento nl_{ss} e energia livre ng_{ss}. Segundo [3],[11], [26], $k_{ss}(n)$, l_{ss} e g_{ss} são parâmetros obtidos a partir de uma aproximação harmônica local do modelo de cadeia livremente articulada, conhecido por ser preciso para as propriedades elásticas do ssDNA em torno da força média de descompactação (f_{av}). Além disso, temos dois ligantes rígidos de comprimento l_{ds}, uma armadilha óptica, com rígidez K_{trap} constante e L indica a posição desta armadilha. Tais parâmetros e componentes deste processo de descompactação pode ser visualizado na figura 12, também disponível em [3].

4.2.2.1 Descompactação do ssDNA

Para o espectro de forças de interesse, a elasticidade do ssDNA é bem descrito pelo modelo FJC (ver subseção 4.2.1) modificado, dado por (veja a figura 12)

$$G_{ss}(n,f) = ng_{ss}(f) = nb_0 \log\left[k_BT\frac{\sinh(df/k_BT)}{df}\right] \tag{4.32}$$

para os parâmetros $d = 0.59A^o$ e $b_0 = 1.15A^o$.

A energia livre da cadeia de ssDNA a uma distância fixa x_{ss} entre suas duas extremidades é

$$G_{ss}(n,x_{ss}) = f(x_{ss})x_{ss} - ng_{ss}(f(x_{ss}))\,. \tag{4.33}$$

Nesta equação, temos:

- $f(x_{ss})$ é a força necessária para uma cadeia simples com n bases abertas ter uma extenção x_{ss}.

- $f(x_{ss})x_{ss}$ que representa a quantidade de "energia" que foi aplicada, suficientemente para romper n pares de bases.

- $ng_{ss}(f(x_{ss}))$ representando a quantidade exata necessária para o rompimento de n pares de bases.

No ponto de equilíbrio, (lembre que a cadeia é elástica e se não for esticada ela se comprime e assim o caso em que a energia livre $G_{ss}(n,x_{ss})$ estiver no estado de equilíbrio), podemos a partir das equações 4.32 e 4.33 determinar uma expressão para x_{ss} dependente da energia necessária para romper cada par de base (g_{ss}), de forma implícita. De fato, derivando a equação 4.32, em relação a f, temos:

$$\frac{\partial G_{ss}}{\partial f}(n,f) = n\frac{dg_{ss}(f(x_{ss}))}{df} = x_{ss} - n\frac{dg_{ss}(f(x_{ss}))}{df} \tag{4.34}$$

$$x_{ss} = n\frac{dg_{ss}(f(x_{ss}))}{df} = 0\,. \tag{4.35}$$

Assim, de 4.34, temos x_{ss} dada de forma expícita por:

$$x_{ss} = \frac{\partial G_{ss}}{\partial f}(n,f) = n\frac{dg_{ss}(f(x_{ss}))}{df} \tag{4.36}$$

Agora, vamos reescrever a expressão 4.36 através de uma expansão em torno da força média de descompactação (f_{av}). Para isso, temos que escolher um valor de referência

para f_{av}, e assim, definir a extensão do ssDNA por pares de bases, dada por:

$$l_{ss} = \frac{dg_{ss}(f_{av})}{df} \tag{4.37}$$

Estamos considerando que pequenas variações em x_{ss} do valor de equilibrio nl_{ss} correpondam a força $f_{av} = f(nl_{ss})$, resultarão em pequenas variações na força f aplicada nas estremidades da cadeia de ssDNA. Linearizando a equação 4.36 em torno dos valores $x_{ss} = nl_{ss}$ e $f = f_{av}$, temos

$$f - f_a v \cong K_{ss}(n)(x_{ss} - nl_{ss}) \tag{4.38}$$

onde a constante de rigidez K_{ss} é dada por

$$K_{ss}(n) = \left[n\frac{d^2 g_{ss}}{df^2}(f_a v) \right]^{-1} . \tag{4.39}$$

Note que a expressão para K_{ss} dada pela equação (4.39) implica que a rigidez diminui com o número de pares de base n abertos.

Com esta aproximação dada pela equação (4.38), a energia livre para a cadeia de ssDNA a uma extensão fixa x_{ss} em (4.33) fica

$$G_{ss}(n, x_{ss}) = f_{av}x_{ss} + \frac{1}{2}K_{ss}(n)(x_{ss} - nl_{ss}) - ng_{ss} \tag{4.40}$$

onde $g_{ss} = g_{ss}(f_{av})$.

4.2.2.2 Descompactação do dsDNA

No processo de descompactação experimental que estamos analizando, além do modelo de descompactação do ssDNA, temos que levar em consideração a influência da armadilha óptica que assumiremos possuir uma constante de rigidez K_{trap} e as conexões entre os elementos que estão em cada uma das cadeias de ssDNA da élice dupla (veja figura 12). Consideraremos estas conexões rígidas para a energia f_{av} que será usada. Por fim, temos que modelar a cadeia de dsDNA.

Consideraremos aqui que a energia gasta para separar os primeiros n pares $\{s_1, \cdots, s_n\}$ seja dado por (4.31), para g_0 dado na figura 4, disponível no 1.

Nosso próximo passo é descrever a energia livre total, que leva em conta as duas fitas simples de ssDNA que estão entrelaçadas na élice do dsDNA, cujas distâncias fixas serão denotadas, respectivamente, por x_{ss}^1 e x_{ss}^2, a uma posição L da armadilha. Ainda, consideramos que as bases das duas fitas estão a uma distância fixa l_{ds}. Supondo que o centro da armadilha está localizado em $L - x_{ss}^1 - x_{ss}^2 - l_{ds}$, temos que a energia livre total

pode ser escrita por,

$$G(x_{ss}^1, x_{ss}^2, n|L) = G(n) + G_{ss}(n, x_{ss}^1) + G_{ss}(n, x_{ss}^2) + \frac{1}{2}K_{trap}(L - x_{ss}^1 - x_{ss}^2 - l_{ds})^2 \tag{4.41}$$

Usando o modelo para G_{ss} dada na equação (4.40), para cada uma das fitas simples do DNA, respectivamente, a equação (4.41) pode ser reescrita por

$$\begin{aligned} G(x_{ss}^1, x_{ss}^2, n|L) = G(n) - 2ng_{ss} + f_{av}(x_{ss}^1 + x_{ss}^2) + \frac{1}{2}(x_{ss}^1 - nl_{ss})^2 + \\ + \frac{1}{2}(x_{ss}^2 - nl_{ss})^2 + \frac{1}{2}K_{trap}(L - x_{ss}^1 - x_{ss}^2 - l_{ds})^2 . \end{aligned} \tag{4.42}$$

Para obtermos a expressão para a energia livre dependendo somente do número de pares de bases e distância L, $G(n|L)$, estimamos a distância entre x_{ss}^1 e x_{ss}^2 e escrevemos $l_{av} = f_{av}/K_{trap}$ e $K(n) = K_{ss}K_{trap}/(K_{ss} + 2K_{trap})$. Assim, obtemos de (4.42) que

$$G(n|L) = G(n) - 2ng_{ss} + \frac{1}{2}K(n)(L - l_{av} - l_{ds} - 2nl_{ss})^2 . \tag{4.43}$$

A rigidez efetiva $K(n)$ não varia significantemente quando, para L fixo, o número de pares de bases descompactadas variam em torno do número médio de bases descompactadas para o dado L fixo. Assim, passaremos a considerar $K(n)$ como uma função que depende somente de L, denotado a partir de agora por $K(L)$.

Nosso objetivo aqui é escrever a energia livre total $G(n, L)$, dependendo somente de L. Para tal, escrevemos

$$Z(L) = \sum_{n=0}^{N} e^{-G(n|L)} . \tag{4.44}$$

Assim, podemos escrever a energia livre como a contribuição

$$G(L) = -\log Z(L) \tag{4.45}$$

4.2.3 Algumas evidências da má-colocação do problema

Nesta subseção daremos uma breve ideia (sem de fato provar rigorosamente), que o problema que estamos estudando é mal posto no sentido de Hadamard, como descrito na seção 2.2.

Dado a energia livre, temos que a o valor da força, para L fixo é dada por

$$f = f_{av} - \frac{dG}{dL}(L) . \tag{4.46}$$

Dado que a energia livre $G(L)$ é medida nos experimentos, a qual denotamos por G^{δ},

temos que o problema de determinar a força correspondente, digamos f^{δ}.

Como na Subseção (2.2.1), temos que este problema é mal posto no sentido de Hadamard. Logo, a solução de qualquer problema inverso que envolva dados experimentais relacionados a força f deve utilizar algum tipo de regularização.

4.3 Modelo dinâmico para a descompactação

A modelagem apresentada anteriormente é independente do tempo. Nesta seção apresentaremos uma dinâmica alternativa para a modelagem da descompactação da cadeia de DNA. Seguiremos o modelo proposto em [12].

Diferentemente da modelagem apresentada anteriormente, nesta modelagem, dado a sequência de bases $\{b_1, \cdots, b_N\}$, queremos determinar o tempo de descompactação em função do número de pares.

Neste trabalho, vamos considerar que a energia livre para descompactar os n primeiros pares de base é dada por

$$g(n, f) = \sum_{i \geq n} g_0(i, i+1) + 2n g_{ss}(f)\,, \tag{4.47}$$

onde g_0 representa a força necessária para separar os pares de bases, enquanto g_{ss} representa a força necessária para separar dois pares na mesma base. Como anteriormente, g_{ss} é dado pela equação (4.32).

Cocco e Monasson em [12] propuseram o seguinte modelo dinâmico para o movimento da interface entre as partes abertas e fechadas da molécula, de acordo com as seguintes taxas básicas

$$r_o(n) = re^{g_0(n)/k_BT}\,, \qquad r_c(f, n) = re^{2g_{ss}(f)/k_BT} \tag{4.48}$$

de abertura r_o e fechamento r_c, respectivamente. O parâmetro r representa a taxa (microscópica) constante para um par de bases abrir ou fechar na ausência da aplicação de forças externas na iteração entre os pares de base. Neste sentido, r pode ser considerado como o inverso de tempo de difusão entre objetos a alguns nanomilímetros de distância. Em particular, $r = k_BT/2\pi\eta l^3$.

5 MODELAGEM DO PROBLEMA INVERSO: SEQUENCIAMENTO

Realizar a modelagem matemática do sequenciamento do DNA, que consiste na resolução de um problema inverso, é uma tarefa bastante difícil devido a complexibilidade do mesmo. Existem diversos pesquisadores que dedicam-se ao conhecimento biológico e ao estudo do processo de predição de uma sequência de DNA, uns apresentam uma modelagem mais realística, abordando um conhecimento avançado no que diz repeito a estrutura molecular e funcionalidades, outros um pouco mais superficiais.

Nos capítulos que antecedem a este, além de caracterizar a molécula de DNA e realizar uma abordagem teorica a cerca de problemas inversos, modelos probabilístico, teorema de Bayes, metodos de regularização e relações entre esses temas, destacamos também que em uma experiência de descompactação da sequência de DNA, observando as energias livres consequentes deste processo ocasionando a abertura/fechamento da cadeia de DNA. Com isso, existem modelos capazes de reproduzir a força (para experiências de velocidade constante) ou posições (para experiências de forças constantes) desse processo, tornando viável obter inversamente informações sobre a sequência.

Nesta seção, realizaremos um estudo de como a molécula de DNA pode ser reconstruída a partir de modelos descritos anteriormente. Neste ponto estamos usando algumas das técnicas desevolvidas em [1], com o diferencial de que, uma vez que construímos a relação entre a distribuição de probabilidades e a regularização de Tikhonov, temos a garantia de que alguns dos processos apresentados neste capítulo são, de fato, métodos de regularização.

Como um dos objetivos deste capítulo é de fato caracterizar a sequência da molécula de DNA, cabe ressaltar que a estrutura dessa molécula é composta por duas cadeias polímeras enroladas, uma em torno da outra, formando uma hélice. Além disso, esta molécula só é estável se as duas cadeias transportarem sequências complementares, ou seja, pares de bases AT e GC justapostos. Devido a isso, dado um único fio (fita) podemos reconstruir a molécula de DNA por síntese de uma cadeia complementar, já que as mesmas obedecem tal restrição. Em outras palavras, o método de sequenciamento depende, primordialmente, do sequenciamento de uma cadeia simples.

5.1 Predição: um caso ideal

A resolução do problema inverso, que consiste na reconstrução da molécula de DNA, está apoiada na teoria Bayesiana. Sendo assim, podemos estudar este problema da seguinte forma: dado um sinal x do experimento de descompactação, podemos olhar a sequência da molécula S para compreendermos como ela foi gerada.

Um caso ideal é aquele em que o experimento não é afetado por qualquer ruído instrumental e os dados são adquiridos com uma resolução de espaço e tempo perfeita. Mesmo na ausência de ruídos instrumentais, o sinal é estocástico devido aos ruídos térmicos e a repetição da experiência não gerará o mesmo sinal.

A probabilidade da sequência de DNA S dado o sinal observado x, no âmbito da inferência Bayesiana é dado por

$$P(S \mid x) = \frac{P(x \mid S).P(S)}{P(x)} \tag{5.1}$$

Nesta equação 5.1, o valor de S maximizando sua probabilidade consiste na predição para sequência. Aqui, se considerarmos *a priori* uma distribuição uniforme de sequências $P(S)$, a maximização de $P(S \mid x)$ reduz-se a de $P(x \mid S)$.

5.1.1 Construindo a $P(x \mid S)$

Inicialmente, vamos considerar a abertura de uma molécula de DNA no intervalo de tempo, 0 a T. A probabilidade de ter um sinal de descompactação x dado um sequência S, $P(x \mid S)$, pode ser construída a partir do modelo dinâmico disponível no capítulo 4. Assim, podemos escrever $P(x \mid S)$ como um produto de T operadores, cada um correspondendo a um intervalo de tempo, da discretização entre 0 e T, como segue:

$$P(x \mid S) = \prod_{t=0}^{T} U_S(x_{t+1}, x_t) = \prod_{t=0}^{T} [1 + dt.H_S(x_{t+1}, x_t)] \tag{5.2}$$

onde

$$H_S(x_{t+1}, x_t) = r_c \delta_{x_{t+1}, x_{t-1}} + r_o \delta_{x_{t+1}, x_{t+1}} - (r_o + r_c)\delta_{x_{t+1}, x_t}, \tag{5.3}$$

e $\delta_{\cdot,\cdot}$ determina se houve uma abertura ou um fechamento da base, dada a força e o efeito das elasticidades.

O sinal $x = \{t_i, u_i, d_i\}$ é representado através do número de intervalos elementares t_i/dt gasto em cada base i, o número u_i ($i \rightarrow i + 1$) e d_i ($i \rightarrow i - 1$) de transições, respectivamente, para as bases seguintes e anteriores. O produto dependente do tempo

em 5.2 pode ser reescrito em função dos pares de bases i

$$P(x \mid S) = \prod_i [1 - dt(r_o + r_c)]^{t_i \backslash dt} . dtr_o{}^{u_i} . dtr_c{}^{d_i} \tag{5.4}$$

onde a probabilidade de descanso/abertura/fechamento de bases i é introduzido com o número de vezes que a transição correspondente ocorreu no caminho dinâmico $x = \{t_i, u_i, d_i\}$.

No processo de Monte Carlo, o tempo gasto em uma base é contínuo de modo que a $P(x \mid S)$ pode ser reescrita como um produto de matrizes de transferência

$$P(x \mid S) = C(f, d) \prod_i M(b_i, b_{i+1}; u_i, t_i, d_i) \tag{5.5}$$

onde

$$M(b_i, b_{i+1}; u_i, t_i, d_i) = exp[-t_i(r_o(b_i, b_{i+1}) + r_c)] \times (dtr_o(b_i, b_{i+1}))^{u_i}. \tag{5.6}$$

$C(f, d) = (dtr_c)^d$ não dependem da sequência e $d = \sum_i d_i$ é o número total de transições para trás.

5.1.2 Verificação de Normalização

Nesta subseção, iremos verificar a normalização correta para a fórmula acima, ou seja, de forma que

$$\sum_x P(x \mid S) = 1\,, \tag{5.7}$$

seja de fato uma distribuição de probabilidades.

Observando a equação 5.4, podemos reescrever a expressão acima como

$$\sum_x P(x \mid S) = \sum_{\{t_i, u_i, d_i\}} N(x \mid \{t_i, u_i, d_i\}) P(x \mid S) \tag{5.8}$$

Note que em 5.5, $P(x \mid S)$ é uma função de $\{t_i, u_i, d_i\}$. Logo, $N(x \mid \{t_i, u_i, d_i\})$ representa o número de caminhos x que correspondem ao mesmo conjunto $\{t_i, u_i, d_i\}$. O $\sum_{\{t_i, u_i, d_i\}}$, deve ser concebido como uma soma sobre todos os caminhos dinâmicos possíveis que são constituído por $N_{MC} = \sum_i u_i + d_i + t_i/dt$ etapas. A partir da equação dinâmica 4.48 cada caminho é fornecido por um produto de N_{MC} operadores de evolução elementar $\hat{H}$. Portanto,

$$\sum_{\{t_i, u_i, d_i\}} \leftrightarrow \sum_j (\hat{H}^{N_{MC}})_{1j} = \sum_m C_m \prod_i (H_{i,i})^{\alpha i} (H_{i,i+1})^{\beta i} (H_{i,i-1})^{\gamma i} \tag{5.9}$$

A soma sobre o índice j significa que estamos considerando todas as possibilidades iniciando em um, independentemente de sua posição final j. O multiplicador $N(x \mid \{t_i, U_i, d_i\})$ correponde ao coeficiente C_m dos termos com $\alpha_i = t_i/dt$, $\beta_i = u_i$ e $\gamma_i = d_i$. Se considerarmos a matriz $\hat{M} = 1 + dt \cdot \hat{H}$, podemos preceber que a probabilidade $P(x|S)$ em 5.4 é equivalente ao m-ésimo termo de 5.9 com

$$H_{i,i} \to 1 - H_{i,i} = M_{i,i} \tag{5.10}$$

$$H_{i,i+1} \to M_{i,i+1} \tag{5.11}$$

$$H_{i,i-1} \to M_{i,i-1} \tag{5.12}$$

e com a mesma $C_m = N(x \mid \{t_i, U_i, d_i\})$. Com isso, podemos reescrever a condição de normalização 5.8 como

$$\sum_x P(x \mid S) = \sum_j (M^{N_{MC}})_{1j} \tag{5.13}$$

A matriz M tem a propriedade de que cada linha é normalizada para 1, ou seja,

$$\sum_j M_{ij} = 1 \tag{5.14}$$

Portanto,

$$\sum_j M_{ij}^2 = \sum_j \sum_k M_{ik} M_{kl} \,. \tag{5.15}$$

Da mesma maneira para qualquer potência de M, até N_{MC}. Consequentemente,

$$\sum_x P(x \mid S) = 1 \tag{5.16}$$

como queriamos demonstrar.

5.1.3 Otimização

Uma vez conhecida a probabilidade $P(x \mid S)$, nesta subseção vamos explorar o problema de maximização de (3.16) sobre todas as possíveis sequências S. Como vimos no Subseção 3.5.1, se utilizarmos um método iterativo, este será uma estratégia de regularização, escolhidas as probabilidades adequadas.

Nesta seção apresentaremos um método iterativo, recursivo para maximizar $P(x|S)$. De fato, utilizaremos um algoritmo conhecido como algoritmo de Viterbi [28] para o processo de maximização da equação 5.5. O algoritmo de Viterbi pode ser descrito como um algoritmo de programação dinâmica para encontrar a mais provável sequência de um estado inobservável.

De forma suscinta, o algoritmo de Viterbi traduzido para o problemas que estamos considerando consiste em:

- Começando pela primeira base, seleciona-se o valor ótimo para esta base, com relação a todos os possíveis valores da segunda base. Neste sentido, atribuímos uma probabilidade a cada valor da segunda base como

$$P_2(b_2) = \max_{b_1} M(b_1, b_2) \tag{5.17}$$

- Otimiza-se para a segunda base, obtendo-se

$$P_3(b_3) = \max_{b_2}(P_2(b_2)M(b_2, b_3)) \tag{5.18}$$

  e assim por diante.

- Podendo assim, estender para i-ésima base como

$$P_{i+1}(b_{i+1}) = \max_{b_i}(P_i(b_i)M(b_i, b_{i+1})) \tag{5.19}$$

  até chegarmos na última base N da sequência.

  Note que, em cada passo o máximo é atingido para alguma base $b_i^{\max}(b_{i+1})$ que depende da escolha da base seguinte b_{i+1}.

- Ao escolhermos o valor b_N^* que otimiza $P_N(b_N)$, a sequência ótima é obtida utilizando a relação recursiva $b_{i-1} = b_{i-1}^{\max}(b_i^*)$ até a primeira base da cadeia.

5.1.4 Resultado Numérico - Programa de Reconstrução

Queremos deixar claro desde o início desta seção que não implementamos os algorítmos que descreveremos. Estes são objetivos de nossos trabalhos futuros. Nesta subseção, só daremos mais detalhes sobre os algoritmos que descrevemos anteriormente.

A predição da molécula de DNA que consiste na resolução de um problema inverso, foco deste capítulo, está fundamentada na teoria Bayesiana e apoiada em métodos de regularização. Para tal predição, destacamos, nesta subseção, a descrição de dois algoritmos que propiciam a reconstrução da sequência da molécula de DNA, o algoritmo de Viterbi e o algoritmo de comprimento de banda infinita, disponíveis, respectivamente, em [1] e [2].

Os programas de predição de uma sequência de DNA que apresentaremos abaixo, baseia-se em algumas condições experimentais, que são: a força f aplicada, a temperatura $T = 300K$, a matriz de empilhamento (figura 4) e a concentração do soluto $[NaCl] = 150mM$.

5.1.4.1 Algoritmo de Viterbi: Reescrito

Para o problema que estamos considerando, o algoritmo de Viterbi, recebe como entrada o sinal de abertura $x = \{t_i, u_i, d_i\}$ e constrói as matrizes $M(b_i, b_{i+1}; u_i, t_i, d_i)$ para cada par de base i. Sendo assim, existem N matrizes distintas, onde N consiste no número de moléculas abertas. E com a finalidade de evitar números grandes, o valor do elemento máximo é fatorado de tal modo que cada matriz tenha um elemento igual a 1 e todos os outros menores que 1 (como no procedimento de normalização). É importante destacarmos, que este procedimento de normalização não afeta o resultado final, pois a constante global não depende das bases (b_i, b_{i+1}). Lembre que, o que queremos encontrar é a sequência S para o qual o máximo $P(x \mid S)$ é alcançado e não o próprio valor de máximo. Este máximo é calculado usando o algoritmo Viterbi, que foi previamente introduzido.

De maneira a evitar instabilidades numéricas advindas de cálculos oriundos de números muito pequenos, aplicamos, de forma equivalente, o procedimento resursivo descrito em 5.19 para o logaritmo da probabilidade $\pi_i = ln(P_i(b_i))$. Assim, o algorítmo de Viterbi se traduz em

$$\pi_{i+1}(b_{i+1}) = \max_{b_i}(\pi_i(b_i) + lnM(b_i, b_{i+1})) \tag{5.20}$$

Uma representação gráfica deste algoritmo é mostrada na figura 13, retirada de [1]. Neste procedimento, as quatro possibilidades para o valor da base b_i é testada e o valor máximo $b_i^{\max(b_{i+1})}$ é calculado em função da base b_{i+1}.

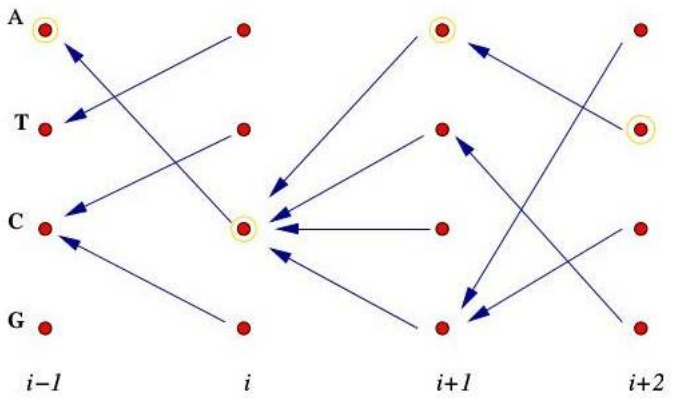

Figura 13: Representação gráfica do algoritmo Viterbi: para cada escolha de uma base, uma ligação com o tipo de base mais provável anterior é desenhado.

Note que, na figura 13, em cada etapa uma ligação é estabelecida entre cada possível b_{i+1} e a escolha correspondente $b_i^{\max}(b_{i+1})$ para a base anterior. Assim, quatro cadeias são construídas. Quando a última base é determinada pela condição $\max P_N(b_N)$ o caminho que descreve a sequência correta é selecionada e toda a sequência é reconstruída de trás

para frente.

Podemos também observar na figura 13, que o máximo de 5.19 é atingido para o mesmo tipo de base, independente da escolha de um lado. Com isso, a previsão para a base i é exata, ou seja, $P(b_i) = 1$

5.1.5 Aproximação SP: uma aproximação para o número de pares abertos

Nesta seção, vamos nos deter em encontrar uma aproximação para o número de pares abertos em um processo de descompactação do DNA. Para determinarmos o número de pares abertos ($n^{SP}(L)$ neste processo, minimizamos $G(n, L)$ encontrada em 4.2.2. Essa aproximação SP, consiste em aproximar a soma dos valores de n em 4.44 $Z(L)$, através de sua contribuição dominante, 4.45, vindo de $n = n^{SP}(L)$. Sendo assim, a força de equilíbrio é dado pela equação

$$< f > (L) - f_{av} \simeq -\frac{\partial G}{\partial L}(n^{SP}(L), L) \tag{5.21}$$

$$\simeq K(L)(L - l_{av} - l_{ds} - 2n^{SP}(L)l_{ss}), \tag{5.22}$$

onde a energia livre corresponde a $G(L) = G(n^{SP}(L)|L)$.

Como os resultamos da descompactação, a energia livre do emparelhamento e o empilhamento (dsDNA) é convertida em energia livre elástica (ssDNA) por pares de bases descompactadas seguindo o modelo harmônico, descrito na seção anterior, definida por

$$g(L) \equiv 2g_{ss} + 2(< f > (L) - f_{av})l_{ss} \tag{5.23}$$

Em 5.23, $g(L)$ é a média dos valores da energia livre dos pares de base, $g_0(s_n, s_{n+1})$, através da distribuição do número de pares de bases descompactadas quando L é fixado.

Agora, de 5.21 e 5.23, temos que o número de pares de bases descompactadas é dado por:

$$n^{SP} = \frac{1}{2l_{ss}}(L - l_{av} - l_{ds} - \frac{f_{exp}(L) - f_{av}}{K(L)}), \tag{5.24}$$

que quando em equilíbrio, a energia livre dos pares de bases, também na posição L corresponde

$$g^{SP}(L) = 2s_{ss} + 2(f_{exp}(L) - f_{av})l_{ss}. \tag{5.25}$$

Sendo assim, o comportamento da energia livre dos pares de bases da molécula de DNA, pode ser obervada representando $(n^{SP}(L), g^{SP}(L))$ para vários valores de L.

A aproximação SP, como mostrado nesta seção, é fácil e rápido de ser implementada, porém ela não leva em consideração as flutuações do número de pares de bases descompactadas, ela deixa em torno de um valor mais provável. Pensando neste fator,

apresentaremos na seção a seguir um outro esquema de aproximação.

5.1.6 Aproximação Box: uma outra aproximação para o número de pares abertos

A aproximação Box também é calculada através da soma de todos os valores possíveis de n em 4.45, porém considera a energia livre dos pares de bases acumulativas, G_{ds} em 4.31, dependendo de um número limitado de parâmetros a qual pode ser otimizada para obter um sinal de força experimental. Para tal consideração, a energia livre acumulada é reescrita como uma soma de funções de caixas de largura b, como

$$G_{ds}^{Box}(n) = b \sum_{k=0}^{\textit{parte inteira de } n/b} g_k \tag{5.26}$$

onde g_k, representa a média do box de energias livres dos pares de bases no intervalo $i = kb + 1, ..., (k + 1)b$.

O valor b, da equação 5.26, pode ser escolhido de acordo com a conveniência, e a ordem de grandeza coincide com as flutuações típicas sobre a posição do par aberto na armadilha ótica na posição L fixa, em unidades de l_{ss}. O mesmo é dado por

$$b_B(L) = \sqrt{\frac{k_B T}{4K(L) l_{ss}^2}} \tag{5.27}$$

Agora, b deve ser esolhido de forma a se adaptar as flutuações existentes e as características do aparelho, ou seja, escolhido de forma que a precisão esteja de acordo com a rigidez da instalação. Neste trabalho, utilizaremos $b = b_B(L)/2$, mostrado no material complementar da tese de Baldazzi [1], que se trata de uma valor ótimo para este tipo de problema.

A resolução do problema inverso, tem por objetivo inferir os parâmetros $g_0, g_1, ..., g_{N/b-1}$ a partir da curva experimental de descompactação ($f_{exp}(L)$). Como resultado de flutuações térmicas do número de pares de bases descompactadas em L fixo, esperamos que as medidas de forças $f_{exp}(L)$ e $f_{exp}(L')$ para serem correlacionadas, enquanto $|L' - L| < \sqrt{k_B T/K} \sim b l_{ss}$.

Agora, para reduzir a redudância nos dados, consideramos o conjunto de forças medidas ($f_{exp}(L_k)$) em posições discretas $L_k = L_0 + k \times 2b l_{ss}$, com k inteiro. Aqui, estamos também considerando que a posição de deslocamento L_0, está englobando o comprimento dos ligantes l_{ds} e o deslocamento médio da conta l_{av}. Assumimos ainda, que o erro experimental para medir a força é uma variável normal com média igual a zero e variância ϵ^2, onde $\epsilon = 0,1pN$. Assim, o logaritmo da probabilidade de um conjunto de medidas de

forças em posições L_k é dado por

$$\log P(\{f_{exp}(L)\}|\{g_k\}) = -\frac{1}{2\epsilon^2}\sum_{k=0}^{N/b-1}(f_{exp}(L_k) - < f >^{Box}(L_k))^2 - \frac{1}{2\Delta^2}\sum_{k=0}^{N/b-1}(g_k - \overline{g})^2$$

onde, no primeiro termo desta equação $< f >^{Box}(L_k)$ é a força de equilíbrio na posição L_k da armadilha, com a verdadeira energia livre por pares de base acumulada G_{ds} substituida por G_{ds}^{Box} e no segundo termo, representa a contribuição, a priori, para o log da probabilidade. Este segundo termo, regulariza o problema de inferência através da imposição dos parâmetros de energias livres que são inseridos.

Podemos ainda, maximizar o $\log P$ em torno dos coeficientes g_k, utilizando um procedimento de gradiente de subida.

5.1.7 Algoritmo de Comprimento de Banda Infinita

O algoritmo comprimento de banda infinita, ou ainda, infinite bandwidth algorithm, leva em consideração que há o conhecimento de variáveis resultantes ou envolvidas no processo de descompactação, disponível para nós na descrição matemática do problema direto no capítulo 4. Neste algoritmo, é suposto que conhecemos a posição da forquilha em todos os momentos, ou seja, é conhecida a informação acerca do número de pares de bases abertas. Uma aproximação para esta informação é dada na Seção 5.1.5. Isto, caracteriza uma situação idealizada, devida as mensurações que deverão ser feitas e a escala a qual se trabalha.

Inicialmente, descreveremos as probabilidades de transição do estado de abertura ($r_0(n)$) e taxa de fechamento ($r_c(f)$), para intervalos de tempo pequenos (Δt), como:

$$P(n(t+\Delta t)|n(t)) = \begin{cases} \Delta t r_o(n(t)) & \text{para} \quad n(t+\Delta t) = n(t)+1 \\ \Delta t r_c(f) & \text{para} \quad n(t+\Delta t) = n(t)-1 \\ 1 - \Delta t r_o(n(t)) - \Delta t r_c(f) & \text{para} \quad n(t+\Delta t) = n(t) \\ 0(\Delta t) & \text{para} \quad outroscasos. \end{cases} \tag{5.28}$$

Está equação pode ser utilizada para definir a probabilidade de um resultado de uma experiência. De fato, este pode ser visto como uma versão discreta da modelagem dinâmica apresentada na seção 4.3, onde

- t_n: tempo gasto com n bases abertas;
- u_n: número de transições de n para $n+1$;
- n_d: número de transições de n para $n-1$.

E escrevemos que a probabilidade de um traço experimental (T), condicionado à

sequência B, sobre a força externa F por:

$$P(T|B) = \prod_n (\Delta tr_o(n(t))^{u_n} (\Delta tr_c(f))^{d_n} (1 - \Delta tr_o(n(t)) - \Delta tr_c(f))^{t_n/\Delta t} \quad (5.29)$$

$$= C(T) \prod_n M(b_n, b_{n+1}; u_n, t_n) \quad (5.30)$$

Nesta equação, podemos separar a parte que depende da sequência da que não depende, definindo então:

$$C(T) = (\Delta t)^{u+d} exp(-t_{tot} r_c(f)); \quad (5.31)$$

$$M(b_n, b_{n+1}; u_n, t_n) = exp(g_o(b_n, b_{n+1}) u_n - r e^{g_o(b_n, b_{n+1} t_n)}; \quad (5.32)$$

onde:

- $u = \sum_n u_n$;
- $d = \sum_n d_n$;
- $t_{tot} = \sum_n t_n$.

Agora, definido o traço, podemos utilizar o teorema de Bayes para calcular a probabilidade de uma sequência dado um traço, como na equação 5.33.

$$P(B|T) = \frac{P(T|B)P(B)}{P(T)} \quad (5.33)$$

Para obtermos uma primeira estimativa aproximada da sequência dada, podemos supor que todas as sequências são equiprováveis e que $P(B)$ é uniforme, embora geralmente não seja verdade.

Na continuidade, podemos maximizar a expressão que temos para $P(T|B)$ sobre $g_o(b_n, b_{n+1})$, sem impor que a energia livre pode ter dez valores distintos (veja a tabela de energia livre na figura 4), para obter uma estimativa máxima da probabilidade, dada por:

$$g_o(b_n, b_{n+1}) = log(\frac{u_n}{r t_n}) \quad (5.34)$$

Este é um cálculo bom para uma primeira estimativa, pois equivale a busca em um espaço contínuo quando efetivamente têm apenas quatro valores possíveis para base.

Agora, a fim de determinarmos uma sequência mais provável, podemos utilizar o algoritmo de Viterbi, disponível em [2] e [1]. Este algoritmo, consiste no seguinte procedimento:

- Vamos considerar as duas primeiras bases e definir $P_2(b_2) =$

$max_{b_1} M(b_1, b_2; u_1, t_1)$, após, $b_1^{max}(b_2) = argmax_{b_1} M(b_1, b_2; u_1, t_1)$, para $n \neq 1$, temos:

$$P_{n+1}(b_{n+1}) = max_{b_n} M(b_n, b_{n+1}; u_n, t_n) P_n(b_n);$$
$$b_n^{max}(b_{n+1}) = arg_{b_n} maxM(b_n, b_{n+1}; u_n, t_n) P_n(b_n).$$

Isso significa que o valor ótimo para uma base depende da escolha para base seguinte.

- Resolvendo estas equações até o último $P_N(b_N)$, que é maximizada para obtermos $b_N^* = argmax_{b_N} P_N(b_N)$, sendo possível retornar ao primeiro valor de ajuste $b_n^* = b_n^{max}(b_{n+1}^*)$. Como mostra o diagrama da figura 14.

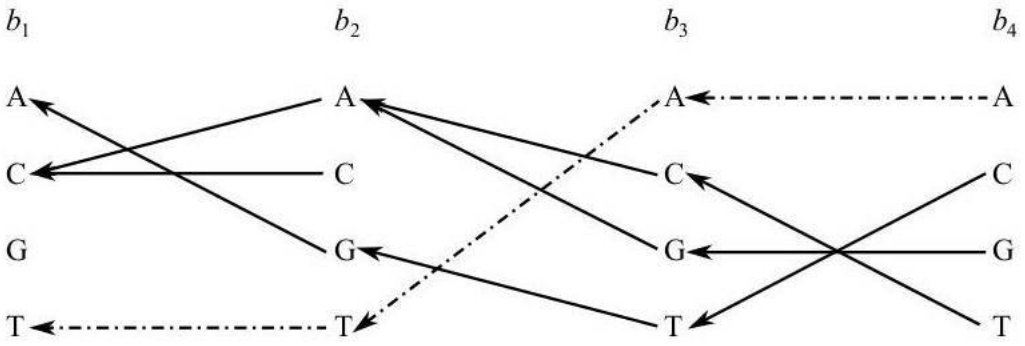

Figura 14: Ilustração do processo de predição da molécula de DNA.

Podemos observar, na figura 14, que ao escolher um $b_1^{max}(b_2)$ equivale a escolher o melhor b_1 para cada uma escolha de b_2, que está sendo representado com uma seta de b_2 para b_1. Após, iterando o procedimento até chegarmos ao b_N (no nosso exemplo figura 14, $N = 4$), podemos calcular a b_N ótima, neste caso A, e propagar para trás obtendo a sequência ótima $TTAA$.

Este algoritmo é relevante, pois a sua complexidade cresce linearmente em N e é preciso explorar somente um subconjunto pequeno de 4^N sequências possíveis. Outro ponto relevante, é que as experiências de descompactação podem ser repetidas várias vezes, obtendo diferentes traços que podem ser combinados pelo produto das probabilidades, como segue:

$$P(T_1, T_2, ..., T_M|B) = \prod_{i=1}^{N} P(T_i|B) \tag{5.35}$$

onde:

- T_i é o traço do i-ésimo experimento de uma série de M.

Portanto, podemos realizar diversos experimentos e combiná-los para a construção da sequência.

Estes, são apenas dois dos algoritmos existentes na bibliografia, que resolvem casos idealizados. A existência de casos mais realísticos e informações acerca da implementação dos mesmos será comentada no próximo capítulo.

6 CONSIDERAÇÕS FINAIS E TRABALHOS FUTUROS

Pesquisas recentes demonstram que o DNA regula diversos mecanismos que são os responsáveis por codificar as proteínas que regulam a maioria das funções vitais. Além de formarem a maioria das estruturas celulares. Assim, a existência de alterações na formação, composição e/ou mutação das moléculas que compõem a sequência de DNA, as proteínas formadoras desta sequência podem ser incapazes de desenvolverem suas funções normais, as quais resultam em uma desordem genética, acarretando, por sua vez, doenças associadas. Portanto, o conhecimento da sequência de DNA, possui um papel de importância central tanto nos diagnósticos como em medidas terapêuticas da detecção e tratamento de doenças. Hoje em dia há diversas técnicas, inclusive comerciais, para o sequenciamento do DNA, que são, caras, demoradas além de não apresentar garantia de precisão nos resultados.

Desda maneira, surge a necessidade de estudos mais aprofundados sobre a modelagem da descompactação da molécula de DNA, a qual é responsável pelo fornecimento e comparação de dados sobre o processo físico da cadeia, dos quais procura-se estrair informações para determinar o sequênciamento. Este último se traduz na determinação da sequência de DNA, que carrega a informação genética vital para a vida. Há uma enormidade de modelos que prevêem a modelagem dos processos físicos de descompactação do DNA. Neste trabalho nós apresentamos alguns dos modelos, assumindo que a cadeia de DNA seja um polímero elástico. Esta modelagem possui suas próprias limitações. Dentre elas estão a resolução espaciais sub-atômicas e a precisão da ordem de **piconewton** para a medida das forças que atuam no processo de descompactação. Embora estes pontos possam ser superados por aparelhos modernos, a existência de instabilidade térmicas da armadilha óptica que opera na descompactação, faz com que os dados para o sequenciamento sejam corrumpidos por ruídos. Como o sequenciamento é, de fato, um problema inverso e assim, mal posto no sentido de Hadamard, faz-se necessário a introdução de técnicas de regularização para resolver o problema, de maneira estável, com relação ao nível de ruído nos dados.

A principal contribuição e originalidade de nossa proposta está em utilizar a fórmula de Bayes associada a distribuição de probabilidade dos dados observados (forças de rom-

pimento) para determinar possíveis mutações na cadeia de formação de DNA. Provamos que, existe uma relação intrinseca entre a teoria Bayesiana e a regularização de Tikhonov (Tikhonov iterado). Com isso podemos garantir que as soluções obtidas não sofrem, muito, com os erros na modelagem, bem como com os erros nas observações, características de problemas mal postos (problemas inversos). Com base na fórmula de Bayes, propusemos um algorítmo iterativo, que pode ser considerado como um método de Tikhonov-iterado, para o processo de sequenciamento.

No entanto, gostaríamos de deixar claro que o estudo apresentado aqui é apenas o princípio de um grande estudo necessário para entender todo o processo complexo envolvendo a descompactação e sequenciamento de uma cadeia de DNA. Assim, deixamos diversos pontos a serem estudados nos trabalhos futuros, os quais listamos logo abaixo.

6.1 Trabalhos Futuros

Dentre os trabalhos futuros que temos em mente, apontamos

- Modelar o processo de descompactação como um sistema dinâmico (parecido com o apresentado na Seção 4.3) que incorpore os erros térmicos associados a armadilha óptica.

- Como a cadeia de DNA pode ser considerada como um polimero elástico, e assim o processo de abertura e fechamento da cadeia é influenciada por elementos de memória própria dos efeitos elásticos, pretendemos modelar os processos associados com a descompactação com derivadas de ordem fracionária [13].

- Implementar o algoritmo iterativo para o processo de descompactação apresentado neste trabalho, para cadeias longas.

- Estudar os processos apresentados neste trabalho sob distribuições de probabilidades mais complexas que a distribuição Gaussiana, de forma a ser mais preciso na modelagem. Uma vez identificada esta distribuição de probabilidade, derivar, via fórmula de Bayes, um método de regularização associado a um funcional de Tikhonov generalizado, de forma a garantir propriedades de regularização da solução associada. Bem como implementar o algoritmo de sequenciamento inerente ao processo.

REFERÊNCIAS

[1] V. Baldazzi. *Statistical Mechanics of Unzipping: Bayesian Inference of DNA Sequence*. PhD thesis, Universita' Degli Studi di Roma 'Tor Vergata', Roma, 2005.

[2] C. Barbieri. *Des problémes inverses en Biophysique*. PhD thesis, Universite Pierre et Marie Curie, Paris-FR, 2011.

[3] C. Barbieri, S. Cocco, T. Jorg, and R. Monasson. Reconstruction and identification of dna sequence landscapes from unzipping experiments at equilibrium. *Biophysical Journal*, (106):430–439, 2014.

[4] J. Baumeister and A. Leitão. Topics in inverse problem. In *25º Colóquio Brasileiro de Matemática*, Rio de Janeiro, 2005. IMPA.

[5] S. M. Bhattacharjee and D. Marenduzzo. Dna sequence from the unzipping force? : one mutation problem. 2001.

[6] D. Calvetti and E. Somersalo. *Introduction to Bayesian Scientific Computing - Ten Lectures on Subjective Computing*. Springer-Verlag New York, 2007.

[7] D. Calvetti and E. Somersalo. *Computational Mathematical Modeling. An Integrated Approach through Scales*. SIAM, Philadelphia, 2012.

[8] A. Cezaro and A. Leitão. Problemas inversos: Uma introdução. In *1º Colóquio de Matemática da Região Sul*, Santa Maria - RS, 2010. UFSM, UFSM.

[9] A. Cezaro, O. Scherzer, and J. Zubelli. A convex-regularization framework for local-volatility calibration in derivative markets: the connection with convex-risk measures and exponential families. *Universidade Federal do Rio Grande - FURG*, 2008.

[10] H. Chial. Dna sequencing technologies key to the human genome project. *Nature Education*, 1(1):219, 2008.

[11] S. Cocco, J. Marko, and R. Monasson. Theoretical models for single-molecule dna and rna experiments: from elasticity to unzipping. *C.R. Phisique*, (3):569–584, 2002.

[12] S. Cocco, R. Monasson, and J. Marko. *Slow nulcleic acid unzipping kinetics from sequence-defined barriers*. Eur. Phys. J. E 10, 2002.

[13] A. De Cezaro, A.and Oliveira and F. De Cezaro. *Identificação de parâmetros em equações diferenciais: teoria e aplicações*. EDUFPI, Teresina - PI, 2012.

[14] H. Engl, C. Flamm, P. Kugler, J. Lu, S. Muller, and P. Schuster. Inverse problems in systems biology. *Inverse Problems*, (25):123014, 51 pp, 2009.

[15] P. E. Gill, W. Murray, and M. H. Wright. *Pratical optimization*. Academic Press, London, 1981.

[16] J. Hadamard. *Lectures on Cauchy's problem in linear partial differential equations*. Dover Publications, New York, 1953.

[17] N. Holtzman, P. Murphy, M. Watson, and P. Barr. Predictive genetic testing: from basic research to clinical practice. *Science*, 278(5338):602–605, 1997.

[18] A. Izmailov and M. Solodov. *Otimização - Condições de Otimalidade, Elementos de Análise Convexa e de Dualidade*, volume 1. IMPA, Rio de Janeiro - RJ, 2005.

[19] B. James. *Probabilidade: um curso em nível intermediário*. IMPA, Rio de Janeiro - RJ, 2010.

[20] e. Justine Burley, John Harris. *A Companion to Genethics*. John Wiley & Sons, 2008.

[21] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Springer-Verlag New York, 2005.

[22] E. Kreyszig. *Introductory functional analysis with applications*, volume 1. Wiley Classics Library, Nwe York, 1989.

[23] P. Lee. *Bayesian Statistics*. Oxford University Press, UK, 2004.

[24] A. Magalhães, M.and LIMA. *Noções de Probabilidade e Estatística*. Edusp, São Paulo - SP, 2004.

[25] J. Nocedal and S. J. Wright. *Numeriacal Optimization*. Springer, New york, 2 edition, 2006.

[26] S. Smith, Y. Cui, and C. Bustamante. Overstretching b-dna: the elastic response of individual double-stranded and single-stranded dna molecules. *Science*, (271):795–799, 1996.

[27] P. Turner, A. McLennan, A. Bates, and M. white. *Biologia Molecular*. Guanabara Koogan S.A., Rio de Janeiro - RJ, 2004.

[28] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, (13):260–269, 1967.

[29] T. Zhang, C. Zhang, Z. Dong, and Y. Guan. *Determination of Base Binding Strength and Base Stacking Interaction of DNA Duplex Using Atomic Force Microscope*. Scientific Reports, 2015.