Fábio Manoel França Lobato

Abordagem Probabilística para Caracterização do Sistema de Marcação do Sequenciamento Multiplex na Plataforma ABI SOLID

Belém - PA, Brasil

2011

Fábio Manoel França Lobato

Abordagem Probabilística para Caracterização do Sistema de Marcação do Sequenciamento Multiplex na Plataforma ABI SOLID

Dissertação apresentada à Banca examinadora do Programa de Pós-Graduação em Engenharia Elétrica como um dos requisitos para a obtenção do Título de Mestre em Engenharia Elétrica com ênfase em Computação Aplicada.

Orientador:
Prof. Dr. Ádamo Lima de Santana

INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA ELÉTRICA
UNIVERSIDADE FEDERAL DO PARÁ

Belém - PA, Brasil

2011

Abordagem Probabilística para Caracterização do Sistema de Marcação do Sequenciamento Multiplex na Plataforma ABI SOLID

Dissertação de mestrado submetida à avaliação da banca examinadora aprovada pelo colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Pará e julgada adequada para obtenção do grau de mestre em engenharia elétrica na área de computação aplicada.

Prof. Dr. Ádamo Lima de Santana
ORIENTADOR - ITEC - UFPa

Prof. Dr. Nandamudi Lankalapalli Vijaykumar
MEMBRO EXTERNO - INPE

Prof. Dr. Sylvain Henri Darnet
MEMBRO EXTERNO - ICB - UFPa

Prof. Dr. Marcus Vinicius Alves Nunes
Coordenador do PPGEE

Agradecimentos

À toda minha família, não há palavras que descrevam minha gratidão! Ivan e Eluiza, meus pais, obrigado pela dedicação esforços incomensuráveis à minha educação, e ao meu irmão Ivan Jr., por sempre apoiar minhas decisões.

Ao meu orientador Prof. Dr. Ádamo Santana, por ter confiado e me aceito como aluno de mestrado, proporcionando discussões fundamentais para a evolução, não somente de minhas pesquisas, mas de minha pessoa.

Aos Professores Dr. Renato Francês e Dr. João Weyl, e todos os demais integrantes Laboratório de Planejamento de Redes de Alto Desempenho - UFPa, pela oportunidade que me concederam de participar do grupo de pesquisa e pelas inúmeras contribuições a este trabalho.

Aos Professores Dra. Ândrea Ribeiro-dos-Santos e Dr. Sylvain Darnet, integrantes do Laboratório de Genética Humana e Médica - UFPa, pela disponibilização dos dados e incansável auxílio.

Aos integrantes do Laboratório de Inteligência Computacional - USP, pelo carinho com que me acolheram e pela grande contribuição às minhas pesquisas, em especial à Profa. Dra. Solange Rezende, pelos livros e conselhos que muito me ajudaram.

Aos meus amigos, por serem isso com todas as letras. Principalmente à Daniela Leite, responsável pela efetivação da parceria LPRAD e LGHM, e pela constante motivação concedida; ao Péricles Machado e ao Diego Damasceno, pelas discussões e auxílio nas implementações em C++.

Ao Programa de Pós-Graduação em Engenharia Elétrica por ter me acolhido como aluno de mestrado e ao CNPQ, pelas bolsas concedidas.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Os sequenciadores de nova geração como as plataformas Illumina e SOLiD geram uma grande quantidade de dados, comumente, acima de 10 Gigabytes de arquivos-texto. Particularmente, a plataforma SOLiD permite o sequenciamento de múltiplas amostras em uma única corrida (denominada de corrida *multiplex*) por meio de um sistema de marcação chamado *Barcode*. Esta funcionalidade requer um processo computacional para separação dos dados por amostra, pois, o sequenciador fornece a mistura de todas amostras em uma única saída. Este processo deve ser seguro a fim de evitar eventuais embaralhamentos que possam prejudicar as análises posteriores. Neste contexto, o presente trabalho propõe desenvolvimento de um modelo probabilístico capaz de caracterizar sistema de marcação utilizado em sequenciamentos multiplex. Os resultados obtidos corroboraram a suficiência do modelo obtido, o qual permite, dentre outras coisas, identificar faltas em algum passo do processo de sequenciamento; adaptar e desenvolver de novos protocolos para preparação de amostras, além de atribuir um Grau de Confiança aos dados gerados e guiar um processo de filtragem que respeite as características de cada sequenciamento, não descartando sequências úteis de forma arbitrária.

Palavras-chave: *Bioinformática, Mineração de Dados, SOLiD, Barcode, Sequenciamento Multiplex, Modelagem Matemática.*

Abstract

The next generation sequencers such as Illumina and SOLiD platforms generate a large amount of data, commonly above 10 Gigabytes of text files. Particularly, the SOLiD platform allows the sequencing of multiple samples in a single run (called multiplex run) through a marking system called Barcode. This feature requires a computational process for separation of data per sample, therefore, the sequencer provides a mixture of all samples in a single output. This process must be secure to avoid any harm that may scramble further analysis. In this context, this dissertation proposes development of a probabilistic model capable of characterizing the marking system used in multiplex sequencing. The results corroborate the adequacy of the model obtained, which allows, among other things, identify faults in some step in the sequencing process, adapt and develop new protocols for sample preparation, and assign a grade to the reliability of data generated and guide a filtering process that respects the characteristics of each sequence, without discarding sequences useful in an arbitrary manner.

Keywords: *Bioinformatics, Data Mining, SOLiD, Barcode, Multiplex Sequencing, Mathematical Modeling.*

1 Introdução

"*Para obter conhecimento, adicione coisas todos os dias.*
Para ganhar sabedoria, elimine coisas todos os dias."
Lao Tsé

A comparação de indivíduos entre si é um processo natural e um dos motivadores do estudo da transferência de características dos genitores para a prole, a hereditariedade. Atualmente as pesquisas envolvendo fatores hereditários estão no nível molecular. Os constantes avanços tecnológicos fizeram com que a quantidade de dados destinados a estes estudos obtivessem um crescimento exponencial, como mostra a Figura 1.1, disponibilizada em 2009 pelo *National Center for Biotechnology Information* - NCBI. Este é apenas um dos vários Bancos de Dados Biológicos disponíveis, a exemplo do EMBL *Data Library* (*European Bioinformatics Institute*) e do DNA *Bank Of Japan* (*National Institute of Genetics*), administrados pelo Reino Unido e Japão, respectivamente.

Esta grande quantidade de dados produzida requer auxílio computacional para que sejam realizadas as análises necessárias, a fim de se identificar e validar informações biologicamente úteis. Neste contexto surge a bioinformática, área de pesquisa multidisciplinar que envolve profissionais das mais diversas áreas como biólogos, físicos, estatísticos, engenheiros, cientistas da computação, *etc*. Estes profissionais não atuam exclusivamente na análise dos dados, mas no desenvolvimento de mecanismos para sua obtenção.

Neste âmbito, diversas tecnologias de sequenciamento surgiram nos últimos anos, com destaque para as plataformas pertencentes a classe *High-Throughput Sequencing* (HTS) como o 454 da Roche, o Illumina da Solexa e o SOLiD (*Sequence by Oligonucleotide Ligation and Detection*) desenvolvido pela Applied Biosystems. Eles são conhecidos também como sequenciadores de nova geração e produzem uma grande quantidade da dados, comumente acima dos 10 Gigabytes de arquivos-texto.

Particularmente, o SOLiD permite o sequenciamento de até 256 amostras em uma única corrida, denominado de sequenciamento *multiplex*. Dentre os mecanismos que permitem esta

funcionalidade destaca-se o sistema de marcação de amostras proprietário, chamado de *barcode*. Em sua terceira versão, o SOLiD pode utilizar-se de 16 *barcodes* disponibilizados pela Applied Biosystems em uma biblioteca padrão, o *Small RNA Expression Kit*. Os *barcodes* são uma sequência de seis bases nucleotídicas, iniciadas sempre por uma Guanina e escolhidos, dentre outros pontos, por possuírem ortogonalidade quando representados no sistema de codificação intrínseco do SOLiD, o *colorspace*.

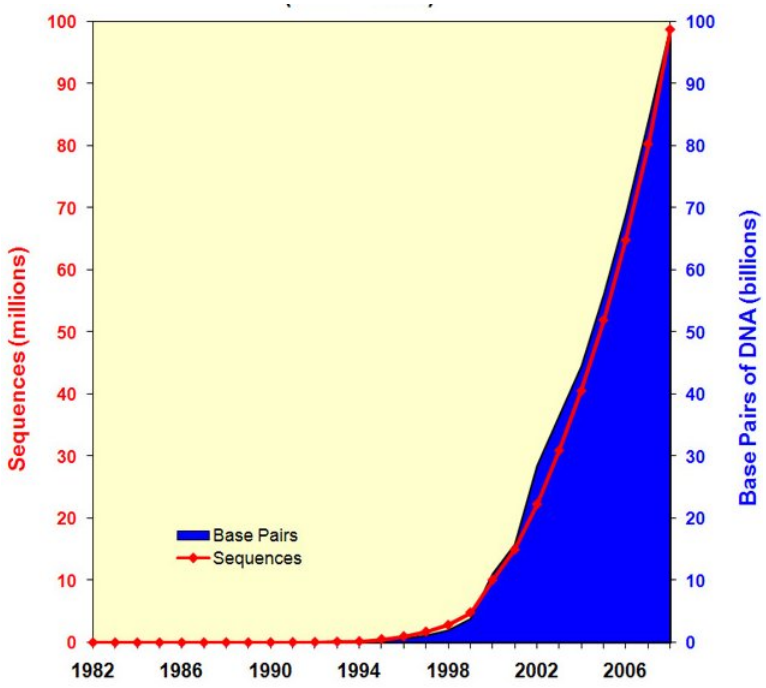


Figura 1.1: Crescimento do Banco de Dados de DNA administrado pelo NCBI.

O sistema de codificação clássico utiliza as iniciais das bases nucleotídicas e é conhecido por *basespace* ou *letterspace*. Uma sequência em *colorspace* inicia por uma base nucleotídica e depois seguem-se números compreendidos entre 0 e 3, cada número representa uma transição, conforme explicitado em (BREU, 2010). Inicialmente, esta nova codificação fez com que fosse adicionado um passo para conversão de *colorspace* para *basespace*, entretanto, isso representava um retrocesso tecnológico. Para tal, iniciou-se uma corrida para desenvolver ferramentas que trabalhassem com o *colorspace* nativamente.

Em sequenciamentos *multiplex*, a plataforma SOLiD disponibiliza em sua saída a mistura de todas as amostras sequenciadas. A recuperação das sequências, por amostra, depende da detecção dos *barcodes* nas sequências de marcação. Logo, uma discrepância nesta proporção e a presença de sequências de marcação que não correspondem aos *barcodes* da biblioteca padrão, prejudicam as análises biológicas subsequentes. Outro possível problema a ser investigado é a presença de sequências com baixa qualidade, tanto no sistema de marcação quanto nas

sequências de interesse, o que prejudicaria a confiabilidade do sequenciamento.

Visando amenizar as consequências das problemáticas supracitadas, definiram-se os seguintes objetivos-gerais para este trabalho:

- pesquisar na literatura sobre problemáticas similares e soluções adotadas;
- investigar os motivos das discrepâncias observadas na quantidade de *barcodes* detectados;
- investigar as possíveis causas e consequências da baixa-qualidade observadas no sistema de marcação utilizado nos dados disponíveis.

O cumprimento do primeiro objetivo-geral aperfeiçoou a compreensão das problemáticas apresentadas, o que culminou na delimitação do escopo da pesquisa e definição dos seguintes objetivos-específicos:

- desenvolver um modelo probabilístico capaz de descrever uma sequência conforme o Valor de Qualidade fornecido pelo SOLiD;
- desenvolver um modelo matemático capaz de caracterizar o sistema de marcação utilizado em corridas *multiplex* da plataforma SOLiD;
- abalizar todo o processo servindo-se de princípios utilizados no processo de extração de conhecimento de base de dados (KDD - *Knowledge Discovery in Data-bases*).

Atualmente, não há na literatura ferramentas que realizem a avaliação da qualidade do sequenciamento de forma direta. Esta tarefa ocorre normalmente por meio da análise da taxa de sequências que obtiveram *match* com o genoma de referência. Esta estratégia foi a adotada para a avaliação do trabalho de (SASSON; MICHAEL, 2010), que propõe um filtro de dados de baixa qualidade do SOLiD, cuja eficiência é medida justamente na taxa de sequencias que obtiveram *match* com o genoma de referência.

Este trabalho possui o diferencial de utilizar-se de uma abordagem probabilística para modelar o sistema de marcação utilizado na plataforma SOLiD. Os valores de qualidade fornecidos são mapeados em um valor de probabilidade, e estes utilizados para calcular a probabilidade de um determinada sequência não ter sido obtida “ao acaso”. Esta informação motivou a busca de estatísticas-suficientes capazes de descrever a qualidade do sequenciamento *multiplex* quanto ao sistema de marcação.

Isto permite a avaliação criteriosa dos protocolos de sequenciamentos utilizados, possibilitando a identificação de faltas em algum passo do processo; a adaptação e desenvolvimento de novos protocolos para preparação de amostras; a atribuição um Grau de Confiança aos dados gerados, além de guiar um processo de filtragem que respeite as características de cada sequenciamento, não descartando sequências úteis de forma arbitrária.

Buscando facilitar a compreensão, esta dissertação está organizada da seguinte forma: o Capítulo 1 apresenta uma introdução acerca do trabalho. O Capítulo 2, Fundamentos Biológicos, apresenta alguns conceitos pertinentes para o entendimento do trabalho, abrangendo o histórico da hereditariedade, alguns conceitos relevantes na área de Biologia Molecular e a descrição da plataforma de sequenciamento de nova geração ABI SOLiD.

O Capítulo 3 versa sobre tópicos básicos de estatística e probabilidade até abordar Processos Markovianos, enquanto a segunda parte deste capítulo aborda tópicos em computação, com ênfase na tarefa de Mineração de Dados, também conhecido na literatura por *Knowledge Discovery in Data-bases* (KDD), enquanto o Capítulo 4 apresenta os trabalhos correlatos.

No Capítulo 5, apresenta-se a abordagem desenvolvida, iniciando com a descrição dos materiais e métodos, para prosseguir com algumas considerações sobre as implementações, testes iniciais e estratégia adotada para cumprir com os objetivos traçados. Ao final, este capítulo apresenta e discute os resultados obtidos. O sexto e último Capítulo apresenta as conclusões, dificuldades encontradas e sugere alguns trabalhos que possam dar prosseguimento neste estudo.

2 *Fundamentos Biológicos*

"Só se dedicará a um assunto com toda a seriedade
alguém que esteja envolvido de modo imediato e que se ocupe dele com amor."
Arthur Schopenhauer

A comparação de indivíduos de uma espécie entre si é um processo natural e um dos motivadores do estudo da transferência de características dos genitores para a prole, denominado hereditariedade. Gravuras presentes na Babilônia, datadas de 6.000 anos mostram genealogias sobre a transmissão de certas características das crinas de cavalo. No entanto, a descoberta da entidade biológica que contém as informações que são repassadas para os descendentes realizou-se recentemente. Salvaguarda-se que a forma como estas informações estão codificadas e as vias metabólicas pelas quais se tornam características, ainda não estão completamente esclarecidas. As pesquisas recentes abrangem desde a forma com que as informações são coletadas até a associação de determinado(s) código(s) a característica(s) presente(s) no organismo em estudo. Este capítulo iniciará expondo brevemente sobre hereditariedade, apresentando conceitos iniciais da biologia molecular, atividades básicas das análises realizadas nesta área bem como suas aplicações e perspectivas futuras; ao final, apresenta também o sequenciador de nova geração SOLiD.

2.1 Histórico da Hereditariedade e Conceitos Iniciais

Desde os primórdios o ser humano busca comparar-se com outros de sua espécie, em características como cor dos olhos e tipo de nariz observava-se a transferência dos genitores para a prole. Isto ocorre por meio da transferência de um *material hereditário* que, segundo (SNUSTAD; SIMMONS, 2008) possui três características básicas:

1. Deve ser capaz de se replicar;
2. Deve conter informação;

3. Deve ser capaz de modificar.

A primeira característica significa que cópias da informação são transmitidas dos genitores para a prole; a segunda implica que o material hereditário deve orientar o desenvolvimento, funcionamento e o comportamento do organismo gerado; por fim, mesmo que uma vez em muito tempo, a informação do material hereditário deve ter a capacidade de mudar, para adaptação e evolução da espécie.

Este material hereditário é estudado pela genética e nela há três grandes marcos. O primeiro diz respeito aos trabalhos do monge austríaco Gregor Mendel. Ele estudou a herança de características diferentes em ervilhas, cultivadas no jardim do monastério. A análise de seus experimentos permitiu-lhe discernir padrões, que o levaram a postular a existência de fatores hereditários (MENDEL, 1866), atualmente denominados de *genes*.

Os resultados dos trabalhos de Mendel foram publicados em 1866 nos anais da *Natural History Society de Brünn*, sociedade científica da cidade onde Mendel vivia e trabalhou. No entanto, o impacto científico viria anos depois com a redescoberta da publicação de Mendel, resultando em uma série de estudos sobre herança em plantas, animais e micro-organismos. Tais estudos originaram outro questionamento: O que é um gene? (SNUSTAD; SIMMONS, 2008) .

Esta pergunta foi respondida em 1953 por James Watson e Francis Crick, que descobriram que os genes são constituídos de substâncias chamadas ácidos nucleicos e estes são feitos de blocos estruturais elementares denominados nucleotídeos. Cada nucleotídeo possui três componentes: uma molécula de açúcar, uma molécula de fosfato e uma molécula de nitrogênio. A molécula de açúcar que define o *ácido desoxirribonucleico* (DNA) é a desoxirribose, enquanto a do *ácido ribonucleico* (RNA) é a ribose. Porém, o grande impacto nos estudos dos genes ocorreu em 1953 em um artigo publicado na Nature, com o título *"Molecular Structure of Nucleic Acids"*, que corresponde ao segundo marco da genética. Nele, Watson e Crick descrevem a estrutura do DNA como uma sequência linear de nucleotídeos ligados uns aos outros, perfazendo uma cadeia e a sequência das bases permite a distinção de um gene do outro (WATSON; CRICK, 1953). Os nucleotídeos que constituem o DNA são a Adenina, Guanina, Timina e Citosina (representadas por "A", "G", "T", e "C", respectivamente); no RNA, a Timina é substituída pela Uracila (representada pela inicial, "U").

Após essas descobertas, várias pesquisas objetivaram a determinação da sequência de todo DNA de um organismo. À coleção de moléculas de DNA dá-se o nome de *genoma*, o primeiro estudado foi o Bacteriófago ϕ X174 (SANGER et al., 1977). Em 1990 iniciou-se o Projeto

do Genoma Humano, considerado o terceiro marco da biologia molecular, um esforço mundial para determinar a sequência de aproximadamente três bilhões de pares de nucleotídeos no DNA humano.

O término do projeto ocorreu em 2001 com a publicação de dois artigos (VENTER et al., 2001) e (LINTON et al., 2001). Neles, foram relatados que 2,7 bilhões de pares de nucleotídeos do DNA humano foram sequenciados. Por meio de análise computacional destas sequências estimou-se que o homem possuía entre 30.000 e 40.000 genes. Pesquisas recentes apontam que este número esteja compreendido entre 20.000 à 25.000 e que cerca de 98% à 99% do DNA humano é lixo, ou seja, não codifica proteínas (SOUTHAN, 2004) (GRIFFITHS et al., 2005).

A consequência inicial deste DNA redundante é que, se pegarmos dois seres humanos ao acaso e compararmos seu genoma, estima-se que 99,99% será idêntico ao outro. Em outras palavras, um ser humano difere em seu genótipo - informação genética de um organismo, em apenas 0,1%. Contudo, sabe-se que a variabilidade das características de cada indivíduo, fenótipo, são praticamente infinitas.

Um raciocínio adaptado de (KEEDWELL; NARAYANAN, 2005) que, embora simplista, auxilia o entendimento desta grande variabilidade do genótipo é atribuir a cada único gene, uma variável binária (escuro ou claro; ausência ou presença; *etc*.). Sabe-se que apenas 0,1% dos genes variam entre os indivíduos e que cada um possui cerca de 23.000 genes, logo o número de combinações possíveis é de 23, é igual à 8.388.608. Apesar de parecer pouco, este número mostra apenas o número de características caso os genes assumissem apenas características binárias e não interagissem entre si. De fato, os fenótipos são infinitos.

Os marcos supracitados permitem dividir a genética em três modos de estudo: Genética Clássica, Genética Molecular e Genética de Populações (SNUSTAD; SIMMONS, 2008). A clássica está compreendida no período anterior a descoberta da estrutura do DNA, os estudos eram realizados analisando os dados obtidos por meio do cruzamento entre linhagens diferentes de organismos.

O segundo modo de estudo é a Genética Molecular que se baseia na avaliação de sequências de DNA por meio de comparação entre as sequências conhecidas, permitindo o geneticista, realizar a tarefa de classificação dada sua função biológica, predição de função, interação entre as moléculas geradas, dentre outras atividades. Por fim, a Genética de Populações estuda populações inteiras de organismo. Neste caso os genes são avaliados estudando-se a variabilidade entre indivíduos em um mesmo grupo de organismos.

2.1.1 Dogma Central e Periférico da Biologia Molecular

Como visto na seção anterior, o DNA é o material hereditário pois possui as três características necessárias: é capaz de se replicar, contém informação (Genes) e é capaz de modificar. Estas características foram melhor entendidas no trabalho de Francis Crick, intitulado *"Central Dogma of Molecular Biology"* (CRICK, 1970). Nele o autor apresenta o fluxo de transferência da informação genética até a expressão das características em mecanismos reguladores, como apresentado na Figura 2.1.

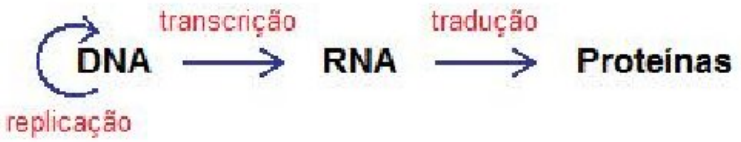


Figura 2.1: Dogma Central da Biologia Molecular.

O processo de implantação da informação genética apresentado pela Figura 2.1 inicia com a transcrição do DNA codificante em RNA, responsável por transferir e regular os mecanismos necessários para produção de proteínas. Os principais componentes envolvidos na codificação das proteínas são os códons. Cada *códon* é uma trinca de nucleotídeos que codifica um único *aminoácido*, unidades básicas de constituição de proteínas. Estas, são produtos da combinação de diversos aminoácidos, tipicamente 200 à 400 (MARZZOCO; TORRES, 2007), formando uma longa cadeia polipeptídica. Estas estruturas tridimensionais são responsáveis por expressar as características dos organismos, controlando funções hormonais, energéticas, enzimáticas, estruturais, *etc*.

Elas funcionam no modelo chave-fechadura, ou seja, há uma proteína específica para cada substrato pois há um encaixe perfeito entre eles. Deste modo uma determinada proteína não atuará em outro mecanismo regulador, por exemplo, a amilase, uma enzima digestiva que interage como catalisador no processo de digestão de carboidratos (BRANDEN; TOOZE, 1999).

Os avanços recentes na área da biologia molecular permitiram a quebra deste dogma, visto que inúmeras outras vias metabólicas foram descobertas. Primeiramente, descobriu-se que apenas uma pequena parcela dos transcritos torna-se proteína e que partes do DNA, antes considerada lixo (*"junk DNA"*), na verdade transcreve importantes entidades biológicas, como *long non-coding RNA*, smallRNA, microRNA (miRNA) e *small interfering RNA* (siRNA).

Tais entidades controlam importantes mecanismos reguladores, que atuam nos níveis de expressão de genes para que haja diferenciação celular, pois todas as células do organismo

possuem o mesmo DNA; em mecanismos reparadores, evitando a pertinência de mutações; na longevidade celular e em várias outras funções ainda não completamente esclarecidas (AUTEXIER; LUE, 2006) (CARTHEW; SONTHEIMER, 2009) (PAVANELLO et al., 2011).

A subseção a seguir apresentará as análises básicas realizadas na biologia molecular, desde os estudos referentes ao DNA até o estado da arte das pesquisas envolvendo RNA-seq e biologia de sistemas.

2.1.2 Atividades Básicas das Análises Biológicas

As análises realizadas na biologia molecular envolvem tanto os estudos das entidades biológicas , RNA, DNA, Proteína, e seus subtipos como será abordado adiante; o estudo das transformações do material genético, tradução e transcrição; e a interação entre todos os agentes e processos relacionados. Existem outros estudos envolvendo mecanismos de auto-regulação, genômica funcional, redes de interação, contudo este subseção abordará três análises básicas: a genômica, transcriptômica e a proteômica.

A primeira análise abordada nesta seção será a *genômica*, subárea da biologia molecular que estuda o genoma completo de um organismo: identificando, quantificando, analisando correlação e predizendo a função de genes ou grupo de genes. Esta área busca relacionar uma determinada característica de uma célula de determinado organismo com a expressão gênica da mesma, confrontando genótipo e fenótipo (BIRD, 2007) (REIK, 2007).

Já o estudo de todas as moléculas de RNA fica a cargo da *transcriptômica*. Para (WANG; GERSTEIN; SNYDER, 2009) *"o entendimento de todo o transcriptoma é essencial para o entendimento dos elementos funcionais do genoma, revelando as moléculas constituintes das células e dos tecidos, e também estudar o desenvolvimento e doenças celulares."* Estes mesmos autores ressaltam que, para o transcriptoma, é importante catalogar todo o tipo de transcritos, incluindo RNA mensageiro (mRNA), RNA não-codificante (non-coding RNA), microRNA, *etc*.

Como dito anteriormente, as proteínas são estruturas compostas por aminoácidos, *"para cada sequência de aminoácido natural, há um estado nativo estável, o qual, sob condições adequadas, é adotado espontaneamente."* (LESK, 2008). Este processo é conhecido por enovelamento da proteína, onde ela assume sua forma tridimensional que atuará segundo o modelo chave-substrato. A subárea da biologia molecular que estuda a composição, formação estrutural e predição de funcionalidade de proteínas é a *proteômica*. Atualmente, o maior desafio desta área é a predição da estruturas, pois é um processo complexo, envolvendo estudos nas áreas de física quântica, computação e bioquímica (BRANDEN; TOOZE, 1999).

Estas são as atividades básicas da biologia molecular. Juntas elas auxiliam no entendimento dos diversos mecanismos reguladores dos organismos. Entretanto, o funcionamento destes mecanismos tem um caráter extremamente não-linear, pois existem redes de interações entre proteínas, RNAs, como também outras substâncias presentes no organismo que alteram o funcionamento padrão.

Neste contexto, surge um novo campo de pesquisa, a *biologia de sistemas*, que concentra na interação e no controle da atividade de genes e proteínas. Juntos, estes campos de estudos da biologia molecular produzem conhecimentos aplicáveis as mais diversas áreas, como será abordado na subseção a seguir.

2.1.3 Aplicabilidade e Perspectivas Futuras

Os estudos envolvendo biologia molecular possuem uma vasta aplicabilidade, uma vez que todos os seres vivos possuem um material genético,responsável pelas características do organismo em questão, seja ele planta, bactéria, anfíbio, manífero, *etc*.

Na agricultura, as pesquisas buscam o melhoramento genético dos alimentos. Culturas como soja, milho e tomate são manipuladas geneticamente visando o aumento da produtividade, resistência a pragas e a duração de seu estado maduro (BARRY et al., 2000) (FISCHHOFF, 1987) (ARTES; ESCRICHE, 1994). Outros trabalhos buscam o estudo do genoma e de sua evolução, como no trabalho de (ARGOUT et al., 2011) que publicou recentemente o genoma do cacau, *Theobroma cacao*, e sua relação evolutiva com outras espécies.

Na pecuária, as pesquisas auxiliam na análise genética da domesticação do gado contemporâneo, que inclui espécies familiares no oeste da Europa e norte da América, *Bos Taurus*, e o zebu encontrado na África e Índia, *Bos indicus*, estudando as características físicas e bioquímicas intrínsecas das variações destas espécies (LOFTUS et al., 1994) (BRADLEY et al., 1996).

Na área médica, os estudos não envolvem somente análises do genoma humano, envolvem também a análise de organismos patogênicos como bactérias, ou estudo da similaridade evolutiva entre os genomas humano e de outras espécies, possibilitando o comparação metabólica em testes de fármacos interespécies (BLUMBERG, 1996) (THOMPSON; MCINNES; WILLARD, 2008).

As recentes pesquisas envolvendo tecnologias de sequenciamento possibilitaram o vislumbramento da popularização das análises genômicas. Novos sequenciadores como o *Ion Torrent* ^{TM} *Semiconductor Sequencing* prometem o sequenciamento de fragmentos de DNA em poucas

horas a um custo reduzido. Em 17 de abril de 2011 o (ESTADÃO, 2011) noticiou o lançamento do *Ion Torrent*^{TM} ao custo de US$ 70 mil, um décimo do que os concorrentes.

Isto permitirá, em um futuro próximo, a investigação de doenças endêmicas, como o câncer gástrico na região norte do Estado do Pará, e ainda exames que proporcionarão a identificação de provável desenvolvimento de doenças como câncer ou hipertensão arterial em determinado indivíduo. A seção a seguir descreve a plataforma de sequenciamento ABI SOLiD^{TM} e suas peculiaridades.

2.2 O Sequenciador de Nova Geração SOLiD

As tecnologias de sequenciamento como a 454 da Roche, o Illumina da Solexa e a SOLiD^{TM} da Applied Biosystems são pertencentes à classe *High-Throughput Sequencing* e revolucionaram as pesquisas biológicas devido a geração milhões de leituras de pequenas sequências, entre 35 à 400 bases, em um espaço curto de tempo (PETTERSON et al., 2009). Adicionalmente, estas plataformas permitem o sequenciamento de diversas amostras em uma única corrida, denominado de corrida *multiplex*.

Os dados gerados por estas plataformas são apresentados no formato de arquivos-texto e comumente ultrapassam os 10 *Gigabytes* (Gb). No entanto, cada sequenciador tem características intrínsecas na sintaxe e codificação dos arquivos. Especificamente a plataforma desenvolvida pela *Applied Biosystems*, o *Sequence by Oligonucleotide Ligation and Detection* SOLiD, possui algumas nuances que serão descritas nas subseções a seguir.

2.2.1 A Plataforma de Sequenciamento SOLiD

A plataforma SOLiD inclui o analisador, formado por alguns instrumentos envolvidos no processo de sequenciamento de amostras, por um sistema computacional acoplado e um conjunto de softwares proprietários. A figura 2.2 mostra a plataforma vista de frente.

Anterior à etapa de sequenciamento é necessário que as amostras sejam preparadas segundo um protocolo estabelecido. Na plataforma SOLiD, seguem-se obrigatoriamente quatro etapas (PANDEY; NUTTER; PREDIGER, 2008):

1. **Preparação das bibliotecas**: determina de qual forma o DNA será manipulado, pode ser de dois tipos: *"fragment library"* ou *"mate-paired library"*, Figura 2.3 A;

2. **Emulsão em *Polymerase Chain Reaction* (PCR)**: amplifica uma ou algumas cópias da

amostra obtida no passo anterior e se encontram anexadas a um componente metálico denominado *bead*, Figura 2.3 B ;

3. **Purificação**: a emulsão é quebrada para liberação das *beads* de onde se remove as amostras contaminadas;

4. **Depósito das *beads***: as *beads* amplificadas e purificadas são ligadas covalentemente à lâminas de vidro por meio dos adaptadores anexados nas etapas anteriores, Figura 2.3 C;

Em seguida, as lâminas obtidas são submetidas ao sequenciador que faz a leitura da transição entre bases nucleotídicas da sequência por meio da detecção dos fluorocromos FAM, Cy3, TXR, ou Cy5. Estas leituras estão codificadas em *"two base color codes"* também chamado de *"colorspace"*.

No SOLiD, o sequenciamento *multiplex* pode ser realizado de três formas (BIOSYSTEM, 2008):

- Utilizando-se lâminas especiais dotadas de segmentos onde as amostras podem ser depositadas em separado, Figura 2.4;
- Utilizando-se do sistema de marcação próprio denominado de *SOLiD System Barcodes*;

Figura 2.2: *SOLiD* ^{TM} *System*.

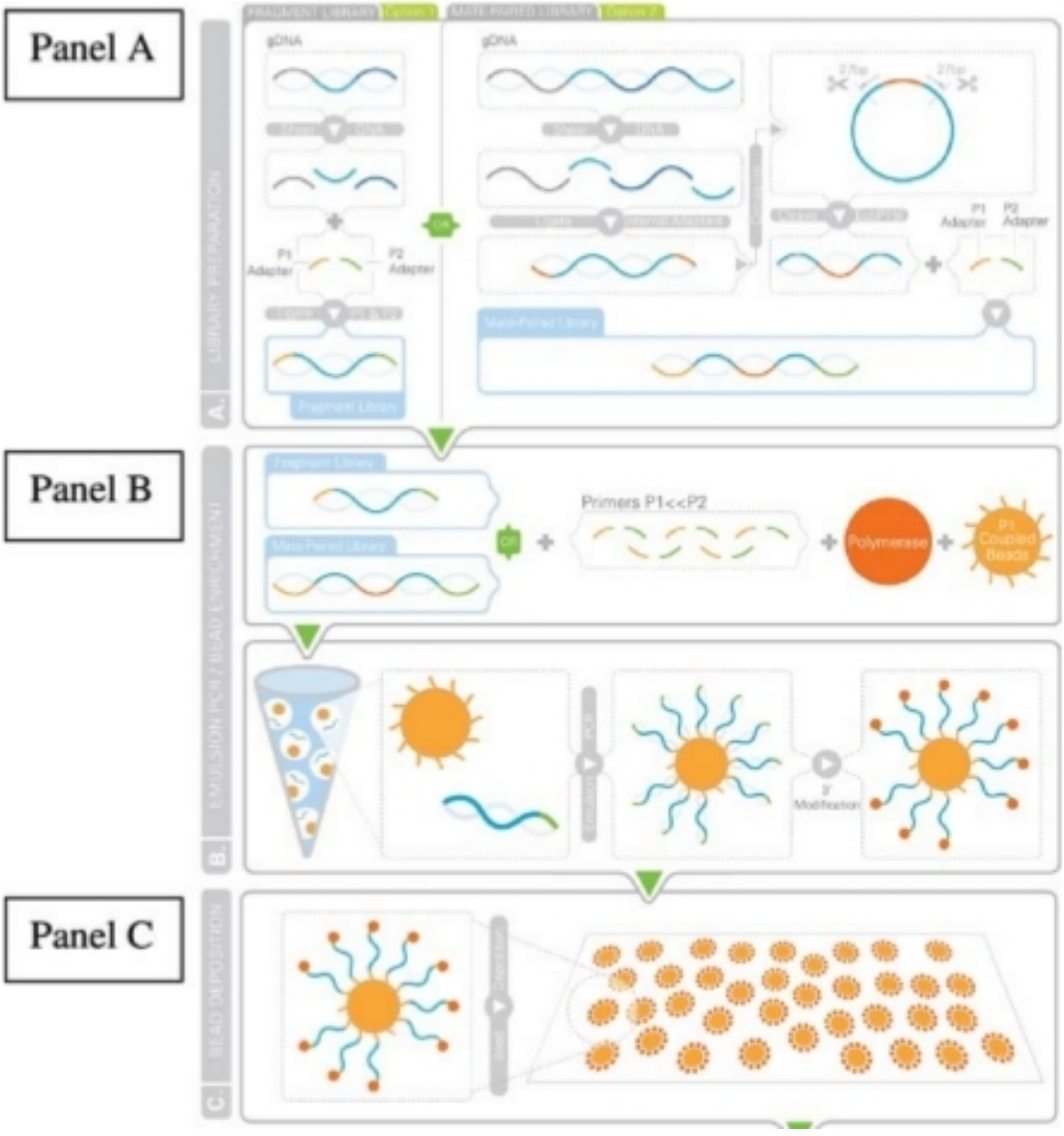


Figura 2.3: Representação esquemática da etapa de preparação das amostras na plataforma SOLiD^{TM} .

- Combinando-se os dois mecanismos supracitados.

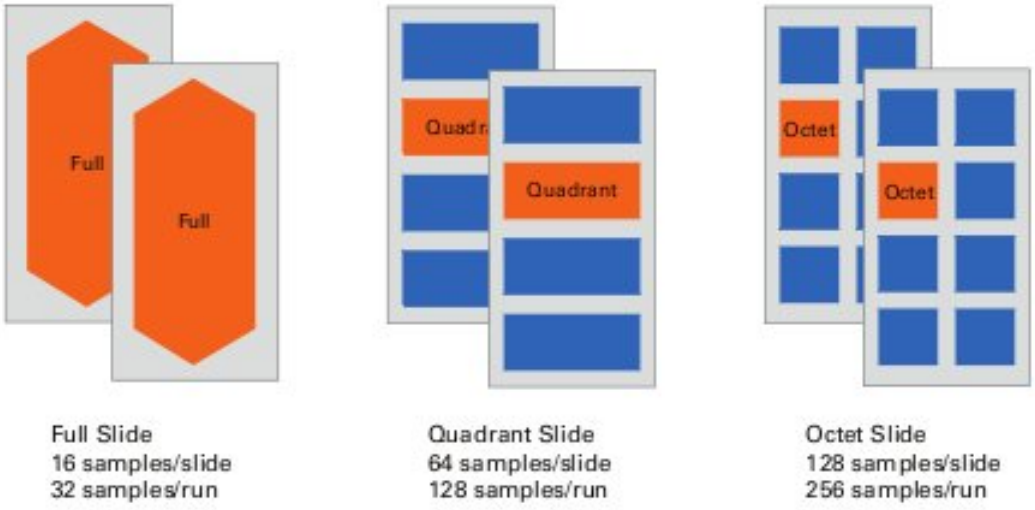


Figura 2.4: Lâminas disponíveis para o sequenciamento *multiplex*, extraído de (BIOSYSTEM, 2008).

Os *barcodes* da biblioteca padrão fornecida pela *Applied Biosystem*, por meio do SOLiD *Small RNA Expression Kit*, contém 16 sequências de DNA que são adaptadas aos fragmentos das amostra antes da etapa de PCR, Figura 2.3 B. Estes *barcodes* são formados por 6 bases nucleotídicas, sempre iniciada pela Guanina e foram escolhidas por possuir a mesma temperatura de fusão, baixa taxa de erro e serem únicas e ortogonais em *colorspace* (BIOSYSTEM, 2008). A Tabela 2.1 apresenta os 16 *Barcodes* da biblioteca padrão, tanto em *colorspace* quanto em *basespace*.

A combinação dos dezesseis marcadores com as lâminas dotadas de segmentos, permite o sequenciamento de até 256 amostras em uma única corrida, o que implica na flexibilização dos estudos, dado a possibilidade de comparação das regiões de interesse de diversas amostras a tempo e custo reduzidos.

2.2.2 A Codificação *"two base color enconding"*

O sistema de codificação *"two base color enconding"*, adotado pela plataforma SOLiD baseia-se na teoria dos grupos e é resultado da arguição di-base, ou seja, cada cor obtida é resultado da combinação de duas bases adjacentes. A Figura 2.5 e a Tabela 2.2 apresentam o esquema visual desta codificação e a associação entre códigos, fluorocromos e pares de bases, respectivamente (BREU, 2010).

Tabela 2.1: Os dezesseis *Barcodes* da biblioteca padrão

Barcode	*colorspace*	*basespace*
1	“0032”	GGGCCT
2	“0111”	GGTGTG
3	“0200”	AAGGGG
4	“0323”	CCGATG
5	“1013”	CAACGA
6	“1130”	GTGCCC
7	“1221”	GTCTGG
8	“1302”	ACGGAG
9	“2020”	GAAGGG
10	“2103”	GACCGC
11	“2212”	CTCAGG
12	“2331”	AGCGTT
13	“3001”	CGGGTC
14	“3122”	CGTCTG
15	“3233”	TAGCGT
16	“3310”	GCGTTA

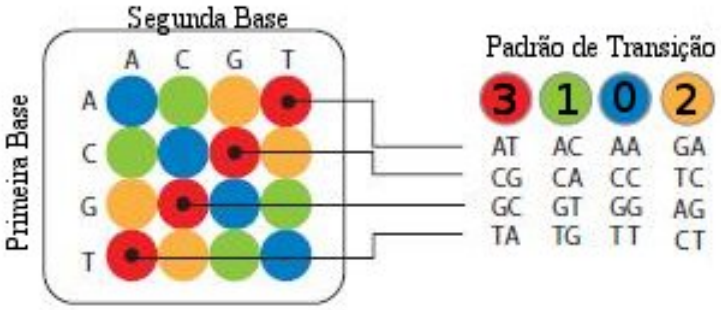


Figura 2.5: Esquema visual da codificação utilizada no SOLiD.

Tabela 2.2: Regras de associação entre códigos, fluorocromos e pares de base do sistema de codificação *"two base color enconding"*

Código	0	1	2	3
Fluorocromos	FAM	Cy3	TXR	Cy5
	AA	AC	AG	AT
	CC	CA	GA	TA
	GG	GT	CT	CG
	TT	TG	TC	GC

Este esquema de codificação implica na necessidade de se adicionar um passo para conversão dos dados, visando obter a sequência do DNA nucleotídeos propriamente dita, também chamada de conversão de *basespace* para *colorspace*. Vale ressaltar que o processo inverso, de *colorspace* para *basespace*, segue o mesmo racionício.

No processo de conversão de *basespace* para *colorspace*, mantém-se a primeira base nucleotídica da sequência e inicia-se o passo de conversão seguindo-se a lógica apresentada na Tabela 2.2. Por exemplo, na conversão da sequência ATCAAGCCTC para *colorspace*, mantém-se o primeiro nucleotídeo A e verifica o código referente à transição AT, que é 3, então adiciona-se à sequência convertida, A3, a próxima avaliação será realizada na dupla TC, consultando a Tabela 2.2, sabe-se que o código referente é 2, portanto, A32. Esse processo é realizado até a última base, como mostra a Tabela 2.3.

Tabela 2.3: Exemplo de conversão entre *basespace* e *colorspace*.

Basespace	A	T	C	A	A	G	C	C	T	C
Colorspace	A	3	2	1	0	2	3	0	2	2

Em corridas *multiplex* o SOLiD fornece em sua saída as sequências de marcação (*barcodes*) e as sequências da amostra propriamente dita, tratadas aqui por sequências de interesse, todas em *colorspace*. A plataforma também atribui um valor de qualidade às transições di-base, com intervalo compreendido de -1 à 35. O valor negativo indica uma ausência de leitura pelo sequenciador enquanto o valor 35 indica o máximo valor de qualidade associada à uma transição. Estas informações estão condensadas em arquivos e representam a saída do sequenciador, tema da próxima subseção.

2.2.3 A Saída do Sequenciador

As informações contidas na saída de corridas *multiplex* da plataforma SOLiD encontram-se condensadas em quatro arquivos, dois referem-se às sequência de marcação e sua qualidade e outros dois referem-se as sequências de interesse, também com a qualidade associada.

Os arquivos que representam as sequências obtidas estão no formato “csfasta”. Os que contêm o valor da qualidade associada a cada transição de base das sequências lidas estão no formato “qual”. A sintaxe é a mesma para os quatro arquivos, como apresenta a Tabela 2.4;

Os arquivos iniciam com três linhas de cabeçalho sinalizadas pelo caractere “#”, seguido pelo par identificador e informação. Os identificadores (ID) também sinalizam a natureza da informação logo abaixo, se ela é pertencente ao *barcode* (R3) ou à sequência de interesse (F3), apresentados na primeira coluna da Tabela 2.4 por “>identificador_*3. A informação pode ser

Tabela 2.4: Sintaxe básica da saída do SOLiD.

Sintaxe	Exemplo
#Cabeçalho	Title: miRNA_20090819_1
>Identificador_*3	>1_44_108_R3
Sequência de marcação ou de Interesse ou Qualidade	G01130

do sistema de marcação como a sequência "G01130" presente na segunda coluna da referida Tabela, ou a sequência de interesse ou ainda a qualidade associada, como mostram as Tabelas 2.5 e 2.6.

Tabela 2.5: Trecho de um arquivo csfasta referente a sequências de interesse.

```
# Fri Sep 4 21:37:24 2009 (...)
# Cwd: /home/pipeline
# Title: miRNA_CANCERGASTRICO_20090819_1_CANCERGASTRICO_
>9_42_20_F3
T233020103231212231201300022000023000
>9_42_360_F3
T230020333032010221101200022000020000
>9_42_455_F3
T233020133233211201202303310000020000
```

Tabela 2.6: Trecho de um arquivo qual referente a qualidade associada à sequências de interesse

```
# Fri Sep 4 21:37:24 2009 (...)
# Cwd: /home/pipeline
# Title: miRNA_CANCERGASTRICO_20090819_1_CANCERGASTRICO_
>9_42_20_F3
31 27 28 18 13 18 12 5 6 17 2 3 2 6 4 20 2 8 3 28 19 3 13 2 18 11 3 19 18 16 15 5 13 11 11
>9_42_360_F3
10 2 10 5 2 10 13 20 7 11 2 4 9 11 4 22 18 3 9 17 6 4 19 15 25 5 6 23 21 13 13 8 12 21 13
>9_42_455_F3
31 26 29 22 21 27 15 5 17 4 21 4 4 21 5 20 8 5 14 24 20 4 19 17 21 14 10 23 24 11 7 17 10 22
20
```

As Tabelas 2.5 e 2.6 mostram um trecho inicial dos arquivos que contém as sequências de interesse e valores de qualidade associadas. As três primeiras linhas correspondem ao cabeçalho com as informações do sequenciamento, no entanto alguns pontos foram suprimidos para facilitar a legibilidade. Após o cabeçalho há os Identificadores seguidos da informação.

A recuperação das sequências de interesse por amostra ocorre da seguinte forma, algoritmos buscam o Identificador, por exemplo o ">9_42_20_*3" nos arquivos R3 e F3 e concatena as sequências de marcação e interesse juntamente com as qualidades associada, depois verifica a qual *barcode* a sequência pertence, separando as informações para tornar possíveis as análises posteriores.

2.3 Considerações Finais

Esta capítulo abordou alguns conceitos biológicos relacionados à transmissão de características dos genitores para a prole, hereditariedade. O material responsável por esta transmissão deve possuir três características: capacidade de replicação, contenção de informação e capacidade de mudança - mutação. Diversos estudos se dedicaram ao material genético, os mais relevantes foram os trabalhos de Gregor Mendel, 1866; Watson e Crick, 1953; e Craig Venter, 2001; este último significou o término do projeto genoma humano, um grande marco para os estudos genômicos.

Este capítulo também apresentou o Dogma Central da Biologia Molecular destacando os recentes avanços na área que permitiu a quebra deste dogma devido a descoberta de inúmeras entidades biológicas como os *long non-coding RNA*, smallRNA, miRNA, *etc*. responsáveis por controlar importantes mecanismos reguladores. Neste contexto, existem diversas atividades nas análises biológicas a fim de estudar estas vias metabólicas, dentre as quais, destacam-se a genômica, a transcriptômica, a proteômica e a mais recente, biologia de sistemas.

O conhecimento destas atividades permite ao leitor, compreender o processo necessário para a descoberta de conhecimento biologicamente útil, os quais são aplicáveis nas mais diversas áreas, desde melhoramento genético de espécies destinadas à alimentação como o tomate, cacau e gado, até na medicina, com testes de fármacos e diagnóstico de doenças genéticas.

Por fim, o capítulo mostrou a plataforma SOLiD - *Sequence by Oligonucleotide Ligation and Detection* da Applied Biosystems, a qual permite sequenciar até 256 amostras em uma única corrida, denominado de sequenciamento *multiplex*. Ressaltou-se também o sistema de codificação próprio, chamado de *colorspace*, que provocou uma corrida para o desenvolvimento de métodos computacionais que trabalhassem nativamente com essa codificação.

3 *Fundamentos Matemáticos e Computacionais*

"É notável uma ciência que começou com jogos de azar tenha se tornado o mais importante objeto do conhecimento humano."

Pierre Simon Laplace

Théorie Analytique des Probabilités, 1812

A bioinformática é uma ciência altamente multidisciplinar, envolvendo profissionais das mais diversas áreas como biólogos, biomédicos, bioquímicos, físicos, estatísticos, engenheiros, cientistas da computação, *etc*. É necessário que estes profissionais tenham noções básicas das diversas áreas que compõem a bioinformática, a fim de facilitar o desenvolvimento das atividades requeridas. Desta forma, este capítulo abordará alguns conceitos matemáticos e computacionais, pertinentes para o entendimento do presente estudo, enfatizando tópicos básicos em estatística, probabilidade e mineração de dados.

3.1 Tópicos Básicos de Estatística e de Probabilidade

A *estatística* é a ciência que fornece os princípios e métodos para coleta, organização, resumo análise e interpretação de dados. Ela baseia-se em *experimentos*, que consistem em procedimentos que coletam medidas ou observações. O espaço com todos os possíveis valores observáveis é chamado de *espaço-amostral* (*S*), e o subconjunto de unidades retiradas de uma população a fim de se obter a informação desejada chama-se *amostra* (KENNEDY; NEVILLE, 1986) (BUSSAB; MORETTIN, 2002) (VIEIRA, 2008).

Em algumas situações, é interessante selecionar subconjuntos do espaço-amostral com características pré-definidas. Por exemplo, o lançamento de uma moeda com duas faces, cara e coroa, e deseja-se saber apenas a quantidade de coroas, delimitou-se um subconjunto, o qual é denominado *evento* (LEON-GARCIA, 2008).

Para cada evento do espaço-amostral S, deve-se atribuir um número não-negativo chamado *probabilidade*, portanto, probabilidade é uma função dos eventos definidos. A literatura adota a notação de P(A) para a "probabilidade de ocorrência do evento A". No entanto, dado que A é um evento do espaço-amostral S, para que P(A) seja considerada uma probabilidade, é necessário obedeça três axiomas (PAPOULIS, 1965) (CLARKE; DISNEY, 1970):

Axioma 3.1 *P(A)*≥ 0

Axioma 3.2 *P(S) = 1*

O Axioma (3.1) denota que a probabilidade é um número não negativo. O Axioma (3.2) implica que o conjunto de todos os valores possíveis, o espaço-amostral é um evento certo, portanto, possui a máxima probabilidade, 1. Em outro extremo, o conjunto vazio $\emptyset$ é um evento com nenhum valor e é conhecido por *evento impossível* cuja probabilidade é 0.

$$\text{Se } A \cap B = \emptyset, \quad \text{então} \quad P(A \cup B) = P(A) + P(B)$$

Se $A_1, A_2, \ldots$ é uma sequência de eventos tais que $A_i \cap A_j = \emptyset$ para todo $i \neq j$, então:

Axioma 3.3 $P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k)$

Este terceiro axioma denota que a probabilidade do evento igual à união de n eventos mutuamente exclusivos é igual à soma das probabilidades individuais de cada evento. Os conceitos acima expostos, de espaço-amostra, eventos e probabilidade perfazem o modelo matemático de um experimento. Para (CHARTRAND, 1985) um modelo matemático representa ou identifica situações da vida-real ou problema, em um sistema matemático. Este mesmo autor apresenta como exemplo a representação de um plano pela geometria Euclidiana, o qual fornece bons resultados na medição de pequenas distâncias.

No mundo real, as etapas de atribuição de um espaço-amostral, definição dos eventos de interesse e atribuição de probabilidade para cada evento, satisfazendo os axiomas; representam o passo mais difícil na resolução de problemas probabilísticos (ROSS, 1972). Este terceiro passo pode ser realizado por meio da frequência relativa de um determinado evento, a qual é definida por:

$$\lim_{n\to\infty}(n_H/n) = P(H) \tag{3.1}$$

A equação 3.1 representa o limite de n experimentos quando estes tendem ao infinito, enquanto o numerador n_H denota o número de ocorrências do evento H, logo n_H/n apresenta a probabilidade de ocorrência deste evento $P(H)$ segundo sua frequência relativa.

Esta abordagem baseada na contagem é intuitiva e satisfaz vários problemas práticos (PEEBES, 2001). Dela pode-se extrair diversas informações que indicam qual evento possui mais incidência, denominado *moda*; ou qual o valor mais provável de ser encontrado, denominado de *valor esperado* ou simplesmente *média*. Estas duas informações estão contidas no grupo das medidas de posição. Enquanto informações como a *variância* e o *desvio padrão*, indicam quão distantes os valores de determinada amostra estão do valor esperado, fazem parte do grupo de medidas de dispersão (KENNEDY; NEVILLE, 1986).

Para se calcular a média e a variância, de uma sequência de eventos $x_1, x_2, x_3, ...x_n$, a média μ é calculada por meio da equação 3.2, enquanto a variância é obtida pela equação 3.3.

$$\mu = \frac{1}{n}\sum_{n=1}^{n} x_i \tag{3.2}$$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{3.3}$$

Outras medidas também são consideradas *estatísticas*, já que resumem as informações contidas na amostra, são elas: valores máximos e mínimos encontrados, amplitude amostral, *etc*. Quando estes valores, ou conjunto de valores representam toda uma série, diz-se que eles resumem a série, por isso são chamados de *medidas-resumo*.

Para facilitar a manipulação de eventos criou-se uma *função* que associa a cada ponto do espaço amostral, um número, geralmente pertencente ao conjunto dos números Reais, denominada de *variável aleatória*. A Figura 3.1 apresenta o lançamento de uma moeda, com o espaço amostral S representado pelos dois eventos possíveis, "cara" e "coroa". Ao evento "coroa", atribui-se o número 0 enquanto ao evento "cara" atribuiu-se o número 1.

Este exemplo, embora simplista, permite visualizar a definição de eventos, no caso, a ocorrência de "cara" e "coroa" no lançamento de uma moeda, por meio de variáveis aleatórias. Por convenção, representa-se a VA por uma letra maiúscula. Então sendo X uma VA e x um número, define-se o evento:

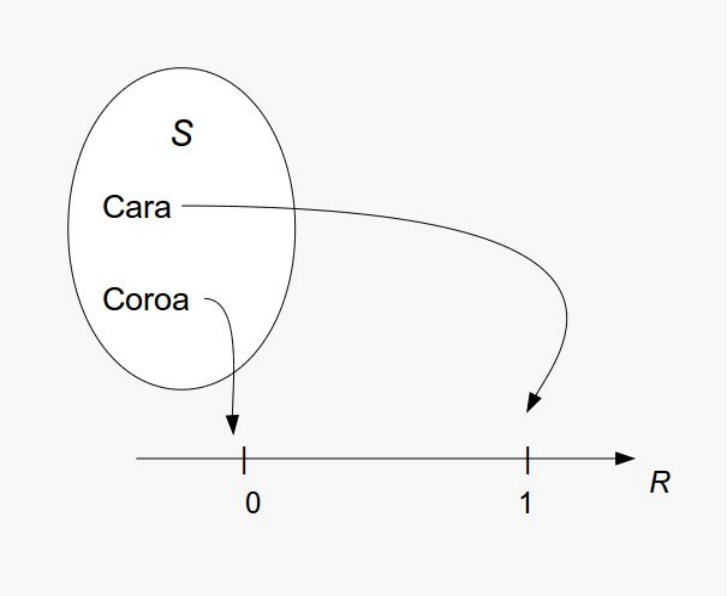


Figura 3.1: Variável aleatória associada com o arremesso de uma moeda com duas faces.

$$[X \leq x] = s : X(s) \leq x \tag{3.4}$$

A probabilidade $P(X \leq x)$ depende de x, logo, é uma função de x. Este tipo de função apresentado na Equação 3.4 denomina-se *função de distribuição de probabilidade cumulativa* ou simplesmente *função de distribuição* $F_X(x)$. Existe também uma classe de função probabilística denominada *função densidade de probabilidade* ou simplesmente *função densidade*, a qual é definida como a derivada da função de distribuição e denotada por $f_X(x)$.

$$f_X(x) = \frac{dF_X(x)}{d_x} \tag{3.5}$$

Existem diversas funções de densidade já definidas como a *gaussiana*, *binomial*, *uniforme*, *etc* (PAPOULIS, 1965). Na equação 3.6 apresenta-se a função de densidade de probabilidade Gaussiana, também chamada de distribuição Normal. Esta distribuição depende de dois parâmetros, média (μ) e variância σ^2.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.6}$$

A Figura 3.2 mostra uma distribuição Normal com parâmetros $\mu = 0$ e $\sigma^2 = 1$, vale ressaltar que neste gráfico, o valor esperado é o mesmo da moda, 0. Por meio desta análise gráfica é possível inferir a aproximação de uma função não conhecida à uma distribuição conhecida.

Contudo, (VIEIRA, 2008) destaca que as análises gráficas são falhas quando comparado com as análises numéricas.

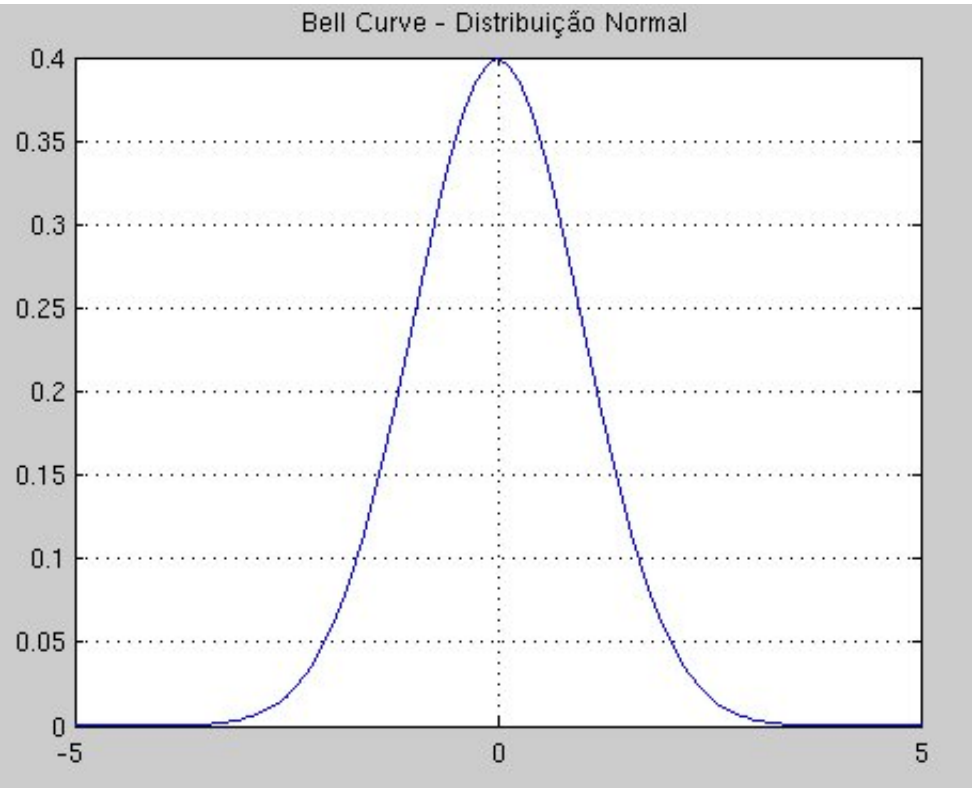


Figura 3.2: *"Bell Curve"* da distribuição Normal com parâmetros $\mu = 0$ e $\sigma^2 = 1$.

A confiabilidade dos métodos numéricos permite a realização de afirmações a partir de uma amostra representativa da população, tarefa conhecida por *inferência*. Isto é feito baseado no conhecimento prévio da distribuição das VA que definem o comportamento de um experimento, as informações inferidas auxiliam na *predição* do comportamento futuro (valores das variáveis) do sistema em estudo.

Existem dois paradigmas principais para realização da inferência estatística, a *inferência clássica*, também chamada de frequencista, e a *inferência bayesiana*, baseado no Teorema de Bayes (PAULINO; TURKMAN; MURTEIRA, 2003). A vantagem do desenvolvimento e utilização de técnicas inferenciais reside na utilização de uma parcela representativa dos dados e não o conjunto em sua totalidade, diminuindo o tempo de coleta e análise. Contudo, em alguns experimentos é necessário a análise de toda população, como será abordado na seção 3.3. Os conceitos de probabilidade e estatística apresentados nesta seção servem de base para o entendimento da seção posterior que aborda Processos Markovianos.

3.2 Processos Markovianos

Considere-se uma *sequência aleatória* ou um processo aleatório de parâmetro discreto X_n, como na Eq. 3.7.

$$\{X_n\} = \{X_0, X_1, X_2, ...\} \tag{3.7}$$

Considere-se também que o espaço-amostra, ou seja, o conjunto de valores possíveis das variáveis aleatórias X_n, é discreto. Neste caso em particular, chamaremos o espaço-amostra de *espaço de estados*, portanto a afirmação "$X_n = i$", pode ser lida da seguinte forma: "o sistema está no estado i após n passos" (PUTERMAN, 1994) .

Para calcular a estrutura de probabilidade de um processo aleatório de parâmetro discreto, determina-se as probabilidades conjuntas:

$$p(j_0, j_1, \ldots, j_k) = Pr(X_0 = j_0, X_1 = j_1, \ldots, X_k = j_k) \tag{3.8}$$

para todo k finito e para toda sequência $j_0, j_1, \ldots, j_k$ de estados. Tal processo é dominado um *processo de Markov* ou *cadeia de Markov* se, para cada k, a probabilidade condicional de que o sistema esteja em um dado estado após k passos, conhecendo-se os estados do sistema em todos os passos anteriores é a mesma que a probabilidade condicional, conhecendo-se apenas o estado em um passo imediatamente anterior (ROSS, 1972) (TIJMS, 1994) (KOVACS, 1996). Portanto, para a Eq. 3.8 seja considerada uma cadeia de Markov, ela deve satisfazer a seguinte identidade:

$$p(j_k|j_0, j_1, \ldots, j_k) = Pr(X_k = j_k|X_{k-1} = j_{k-1}) = p(j_k|jk-1) \tag{3.9}$$

O conjunto de todas as probabilidades conjuntas necessárias à descrição de um processo markoviano chama-se *probabilidades de transição*, a disposição destas probabilidades ($p_{ij}, i, j = 1, \ldots N$) em uma matriz **P** dá origem à chamada *matriz de transição* (BOLCH et al., 2006):

$$\mathbf{P} = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,N} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,N} \\ \vdots & \vdots & \vdots & \vdots \\ p_{N,1} & p_{N,2} & \cdots & p_{N,N} \end{pmatrix}$$

Convém observar que qualquer matriz que tenham entradas não negativas, com somas de linhas iguais a um é chamada de *matriz estocástica*. Este tipo de matriz pode ser representado por uma máquina de estados (CARROL; LONG, 1989) (KAM, 1997) ou por um grafo orientado (CHARTRAND, 1985). Um exemplo de aplicação destes conceitos é a representação de uma sequência finita de nucleotídeos, por um modelo probabilístico. Nesta caso, a Figura 3.3 mostra os nucleotídeos como os estados possíveis $S = \text{A, C, G, T}$ e as conectividades *a priori* como todas as transições possíveis.

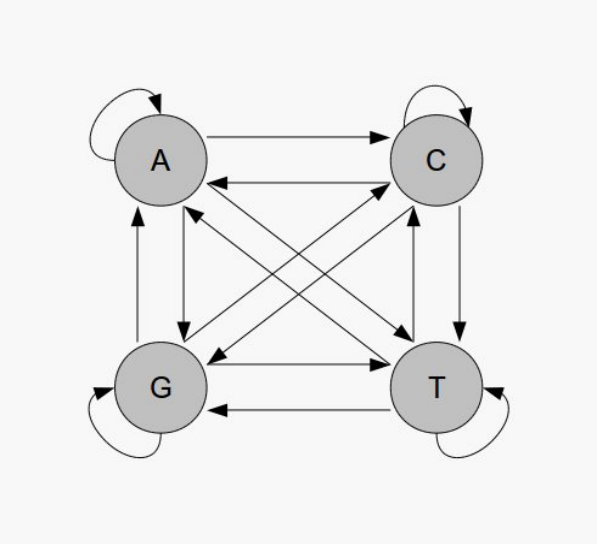


Figura 3.3: Cadeia de Markov que modela a geração de uma sequência de DNA.

É possível obter informações de um DNA real para verificar as probabilidades de transições entre os estados e assim, alimentar a cadeia da Figura 3.3 com as probabilidades obtidas. Como visto, as Cadeias de Markov são modelos matemáticos versáteis, sendo utilizado nas mais diversas áreas, dentre as que se destacam têm-se: pesquisa operacional, avaliação de desempenho de sistemas, bioinformática e mineração de dados (CHING; MICHAEL, 2010). Esta última aplicação envolve a extração de conhecimento de bases de dados para realizar tarefas como classificação, *clusterização* e predição, tópicos discutidos na seção posterior.

3.3 Tópicos em Computação

O desenvolvimento tecnológico na área de componentes eletrônicos possibilitou o grande aumento da geração, armazenamento e processamento de informação. Dois exemplos ilustram este aumento, o primeiro refere-se à evolução do *Hard Drive* (HD), em 1956 a IBM lançou o primeiro HD, o IBM 350, com capacidade de 5 *megabytes* (IBM, 2011), atualmente, os computadores pessoais vêm equipados com HD com capacidade de armazenamento superiores a 500 *Gybabytes*. O segundo exemplo vem de uma previsão realizada por (GANTZ; REINSEL, 2010), nela, os autores afirmam que em 2020 a quantidade de informação criada e replicada no mundo deve chegar a 35 milhões de *petabytes*, a Tabela 3.1 mostra a relação de grandeza entre *Byte, Megabyte, Gigabyte e Petabyte*.

Tabela 3.1: Múltiplos do Byte.

Denominação	Símbolo	Múltiplo
Byte	B	10^0
Megabyte	MB	10^6
Gigabyte	GB	10^9
Petabyte	PB	10^{15}

As informações em formato digital comumente estão em formato tabular, cada linha de uma tabela é uma instância ou registro e cada coluna representa um atributo que contém informação (DATE, 2004). O conjunto de atributos e registros armazenam conhecimento, o qual deve ser identificado, extraído, validado e utilizado. Estes processos serão abordados na próxima subseção.

Mineração de Dados

O processo de identificação, extração, validação e utilização do conhecimento refere-se ao campo de pesquisa chamado de Extração de Conhecimento de Base de Dados e geralmente denominado na literatura tanto como *Knowledge Discovery in Data-Base* (KDD) quanto como Mineração de Dados (MD). No entanto, alguns autores discordam desta denominação, argumentando que a MD é uma subetapa do processo de KDD (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b). Neste trabalho adotaremos estes termos tratando do mesmo processo com o esquema de etapas apresentado na Figura 3.4.

Para (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a) KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis, a partir de bases de dados. Para (HAN; KAMBER, 2006) o

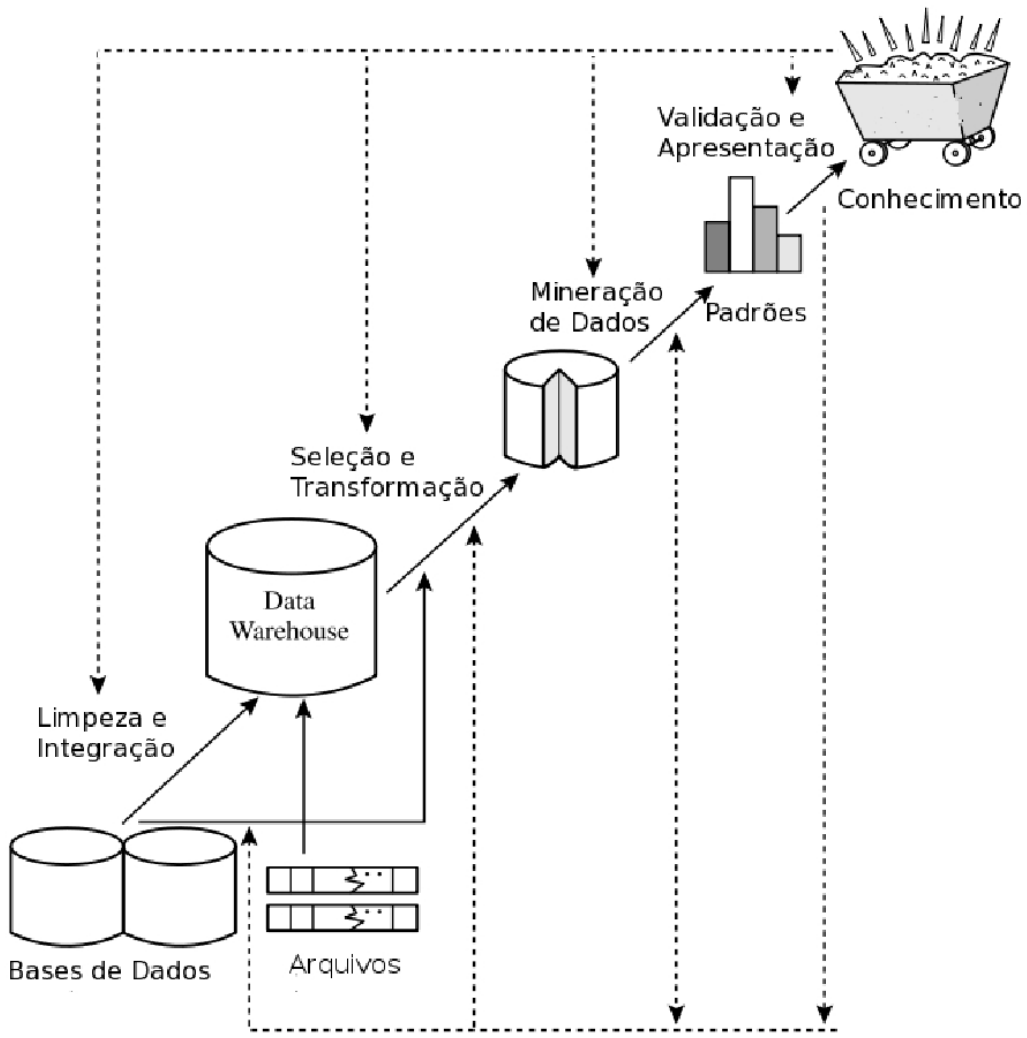


Figura 3.4: Fluxo do processo de KDD, extraída de (HAN; KAMBER, 2006).

processo de descoberta de padrões nas seguintes etapas:

1. *Data Cleaning*;
2. Integração dos dados;
3. Seleção dos dados;
4. Transformação dos dados;
5. Mineração dos dados;
6. Avaliação de Padrões;
7. Visualização do conhecimento;

O processo inicia com a aquisição de dados, em (GARDNER, 1998) o autor define o processo de *Data Wareshousing* como o de criação de um repositório de informação, coletada de diferentes fontes, armazenadas em um esquema unificado e, usualmente residente em um único lugar, o produto deste processo é denominado *Data Warehouse*. Estes repositórios são construídos pelos passos de 1 à 4 propostos por (HAN; KAMBER, 2006). A Figura 3.5 ilustra o arcabouço típico para construção e uso de um *data warehouse*.

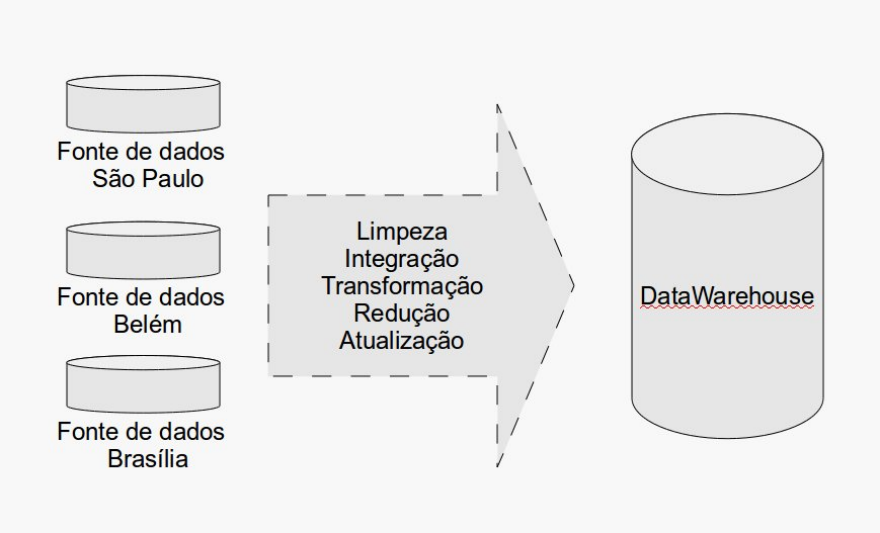


Figura 3.5: Arcabouço para a criação de um *Data Warehouse*, adaptada de (HAN; KAMBER, 2006).

A seta tracejada indica as principais tarefas necessárias para criação de um *data warehouse*. A primeira, *limpeza dos dados*, refere-se a remoção de falhas nos dados como valores ausentes, ruídos e inconsistências, como uma data de nascimento onde o ano é posterior ao ano atual. Estes dados falhos são bastantes frequentes em sistemas de informação que possuem a entrada de dados realizada de forma manual, suscitando erros de digitação ou confusão no preenchimento dos campos; ou ainda, em sistemas de informação gerados pela integração de diferentes sistemas, nestes, as falhas mais frequentes são a presença de valores ausentes e dados inconsistentes, como no caso da discrepância de entre códigos de cadastro entre diversos departamentos que tiveram seus sistemas integrados.

A segunda tarefa consiste na reconfiguração de múltiplas bases de dados para assegurar formatos consistentes e é chamada de *integração*. Após a coleta, limpeza e integração dos dados, comumente é necessário transformá-los para que possam ser processados pelos algoritmos de extração de padrões. Um exemplo trivial de transformação de dados está na conversão de idade para faixa-etária ou na agregação de vendas em totalizadores como vendas mensais e trimestrais.

Em grande parte dos casos, o número de atributos de uma base de dados é muito grande, o que aumenta o tempo aprendizagem e dificulta a interpretação do conhecimento gerado, degradando processo de extração de conhecimento. Para contornar esta problemática realiza-se a redução da base de dados, que pode ser uma redução dimensional, removendo-se atributos ou suprimindo-se instâncias. Por fim, é necessário que a *data warehouse* seja atualizada com novas instâncias, evitando a obsolescência da informação contida.

Todo o processo criação de um *data warehouse*, independentemente da abordagem para a estruturação das etapas do processo de Extração de Conhecimento de Base de Dados, está inserida na etapa de pré-processamento, o qual demanda cerca de 80% do tempo de todo processo, sendo portanto, primordial para o êxito do intento.

Na quinta etapa, mineração de dados, aplica-se os métodos de inteligência computacional para o reconhecimento de padrões. (HAN; KAMBER, 2006) destaca os seguintes objetivos desta etapa:

- Classificação: prediz à qual classe um item pertence;
- Associação: identifica grupos de dados que apresentam co-ocorrência entre si;
- Agrupamento: mais conhecido por *clustering*, identifica grupos de dados associando-os à rótulos;
- Regressão ou predição: mapeia valores dos dados em uma função preditiva, resultando em um ou mais valores reais.

Existem diversos métodos de inteligencia computacional desenvolvidos para a satisfazer os objetivos acima listados. Algumas destas técnicas consistem aplicação de um determinado algoritmo de extração de padrão, outras combinam diversas técnicas visando prover uma melhor adaptabilidade e maior confiabilidade ao resultado final. Portanto, a etapa de processamento engloba a escolha do objetivo e a escolha do algoritmo (WU et al., 2008), culminando na extração de padrões, os quais serão avaliados e utilizados (WITTEN; FRANK, 2005).

As sete etapas propostas por (HAN; KAMBER, 2006) são agrupadas por (REZENDE et al., 2003) em cinco etapas: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento, como mostra a Figura 3.6. Convém observar que as três etapas que, de fato, representam o processo de extração de conhecimento são o pré-processamento, que envolve o *data warehousing*; a extração de padrões, que envolve a escolha do objetivo e a escolha e aplicação do algoritmo; e por fim o pós-processamento que visa validar os padrões extraídos e proporcionar uma interface para visualização do conhecimento.

Figura 3.6: Etapas do processo de Mineração de Dados, extraído de (REZENDE et al., 2003).

3.4 Considerações Finais

Este capítulo abordou alguns conceitos matemáticos e computacionais necessários para o entendimento deste trabalho. A parte matemática teve um enfoque estatístico/probabilístico, apresentando os conceitos de experimento, espaço-amostra, eventos e os axiomas de probabilidade propostos por Kolmogorov. Neste âmbito, mostrou-se sobre algumas medidas estatísticas que analisam a concentração dos dados, como a média e a moda; e a dispersão como a variância e o desvio padrão. Ao conjunto de medidas que representam toda uma série chamou-se de medidas-resumo. Apresentou-se também conceitos de variável aleatória, funções de distribuição e de densidade de probabilidade e alguns tópicos em Processos Markovianos, como Cadeias de Markov.

A seção computacional tratou de um campo da pesquisa destinado à Extração de Conhecimento de Base de Dados, referenciadas na literatura por *Knowledge Discovery In Data-Base* ou por Mineração de dados. O Processo de KDD possui várias abordagens para encadeamento das etapas, no presente trabalho adotou-se a visão de (REZENDE et al., 2003) que possui cinco fases: Identificação do Problema, Processamento, Extração de Padrões, Pós-Processamento e Utilização do conhecimento.

4 *Trabalhos Correlatos*

"Dificuldades são como as montanhas:
só se aplainam quando avançamos sobre elas."
Émile Zola

Diversos estudos vêm sendo direcionados à plataforma SOLiD, em especial ao seu sistema de codificação. Algumas ferramentas desenvolvidas são provenientes da própria Applied Biosystem, que disponibiliza em seu site uma lista de softwares destinado às mais diversas atividades, desde conversores para *basespace*, como o BaseQV Tool, a *pilelines*, que representam um conjunto de softwares hierarquizados responsáveis por automatizar as tarefas de análise de dados, como o "Corona Lite" (BIOSYSTEM, 2011).

Contudo, grande parte dos esforços para desenvolvimento de ferramentas para o SOLiD concentram-se no alinhamento de genoma, também chamado na literatura de mapeamento. (LI; HOMER, 2010) realiza um levantamento dos algoritmos destinados ao alinhamento de sequencias produzidas pelas plataformas de nova geração.

Apesar de desempenhar outra função no processo de análise de dados do SOLiD, alguns itens destes trabalhos auxiliaram no desenvolvimento do presente estudo, portanto serão discutidos na seção posterior, a qual é seguida por um tralho de (SASSON; MICHAEL, 2010) que propõe um *framework* para filtragem de erros presentes nos dados produzidos pelo SOLiD.

4.1 Ferramentas de Alinhamento

As novas tecnologias de sequenciamento proporcionaram um aumento substancial nas aplicações biológicas, desde a análise quantitativa dos transcritos (RNA-seq) à montagem de genomas e transcriptomas em um curto espaço de tempo. Todas estas aplicações biológicas requerem um passo fundamental, o alinhamento das sequências obtidas contra um genoma de referência. Até 2010 cerca de 20 algoritmos para alinhamento de pequenas sequências haviam sido publicadas, (LI; HOMER, 2010) descreve de forma sucinta a fundamentação teórica que embasam

estas ferramentas bem como alguns testes.

Neste mesmo trabalho, o autor divide os alinhadores quanto os algoritmos utilizados em três categorias: baseados em tabelas *hash*, baseados na estrutura de dados *suffix tree*, e os baseados em *merge sorting*. Da primeira categoria que foram analisadas pelo autor, destacam-se as ferramentas SOAP (LI et al., 2008) e ZOOM (LIN et al., 2008). A SOAP é a primeira ferramenta a utilizar *spaced seeds*, uma evolução é a ferramenta PerM (CHEN; SOUAIAIA; CHEN, 2009) de onde adaptou-se o esquema do funcionamento das *spaced seeds* contido na Tabela 4.1.

Tabela 4.1: *Spaced seeds* realizado de forma periódica.

Posição	8	9	10	11	12	13	14
Slide 0	1	1	1	*	1	*	*
Slide 1	*	1	1	1	*	1	*
Slide 2	*	*	1	1	1	*	1
Slide 3	1	*	*	1	1	1	*
Slide 4	*	1	*	*	1	1	1
Slide 5	1	*	1	*	*	1	1
Slide 6	1	1	*	1	*	*	1

O símbolo *, presente na Tabela 4.1 denota um *mismatch* na sequência e 1 a base nucleotídica pertencente à referida posição da sequência analisada. Esta abordagem, utilizando *spaced seeds* aliados a um deslocamento dos *mismatchs* auxilia no aumento da cobertura como também na diminuição do tempo direcionado ao alinhamento.

Isto é atestado por (LI; HOMER, 2010), onde conclui que o alinhamento de sequências curtas não representa um gargalo para a análise dos dados. A relevância deste trabalho também reside na listagem e explanação de algoritmos apropriados à bioinformática, abrangendo estrutura de dados e técnicas otimizadas para realizar o alinhamento de genoma, as quais podem guiar o processo de desenvolvimento de ferramentas destinadas a outras análises.

Por fim, trabalhos envolvendo alinhamento permitem a avaliação dos sequenciamentos em relação à porcentagem de sequências mapeadas contra um genoma de referências. No entanto, leituras com baixa qualidade podem não corresponder a sequência da amostra, gerando resultados faltosos. Este fato, portanto, não permite a avaliação do sequenciamento quanto a aspectos de confiabilidade dos dados.

4.2 Filtragem de Erros do SOLiD

Os trabalhos envolvendo alinhamento e mapeamento de sequências negligenciam as características intrínsecas de cada sequenciamento, por este motivo, (SASSON; MICHAEL, 2010)

realizaram um estudo sobre os dois erros mais comuns na plataforma SOLiD, os erros policlonais e as chamadas errôneas às cores. Os erros policlonais ocorrem quando duas amostras diferentes são amplificadas em uma mesma *bead*, resultando em uma sequência híbrida, que é parcialmente filtrada pela plataforma durante o sequenciamento por meio da intensidade da imagem obtida, relação sinal ruído e o angulo de captura da *bead*. As chamadas errôneas às cores são independentes e podem ocorrer diversas vezes em uma leitura, o que conduz também a uma sequência incorreta.

(SASSON; MICHAEL, 2010) analisaram dados de sequenciamento *mate-paired* no SOLiD dos organismos, sob os seguintes mnemônicos e descrições: DH10B_R3, *Escherichia coli, reverse mate*; DH10B_F3, *forward read*; HuX_F3, Humano; e AtCol_F3, *Arabiodopsis thaliana*. A Figura 4.1 apresenta a relação entre erro e a localização em leituras do SOLiD.

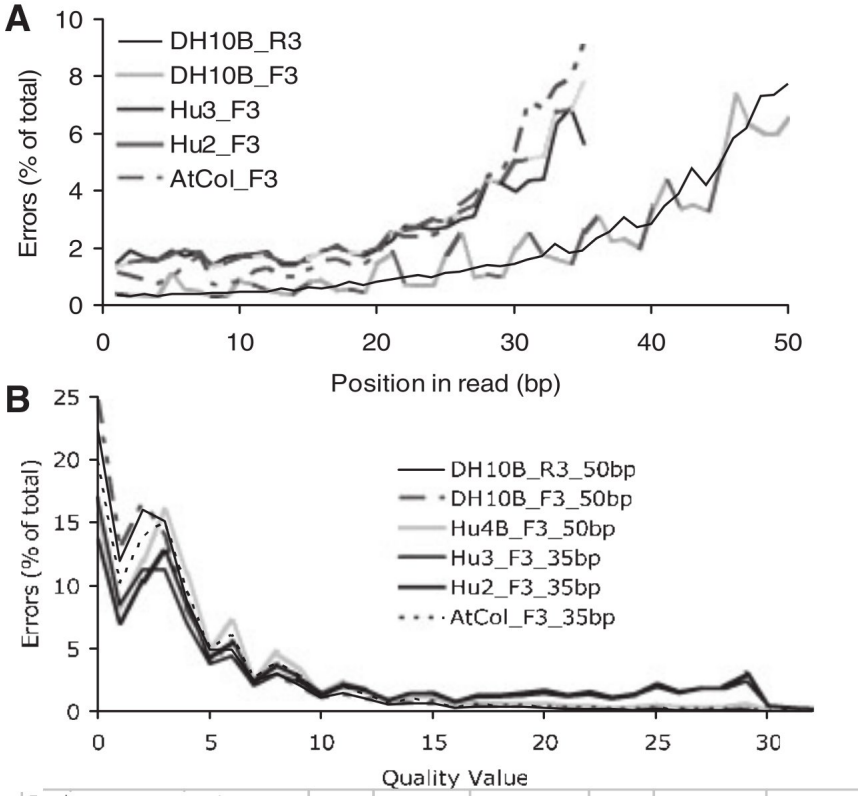


Figura 4.1: Relação entre erro e localização em leituras do SOLiD.

Na Figura 4.1 A, percebe-se que a taxa de erro aumenta consideravelmente a partir de 20 bases lidas, enquanto o valor de qualidade associada a uma taxa de erro considerável está em 10. Desta forma, os autores definiram valores-padrão: para erros policlonais o número mínimo de erros(p) é 1, enquanto o valor de qualidade mínimo é p_QV = 25 e máximo de erros identificados $\varepsilon = 3$; para erros independentes o valor mínimo de qualidade é e_QV = 10. Caso uma sequência apresente um Valor de Qualidade inferior ao patamar estabelecido, a sequência

é descartada.

Os resultados apresentados permitem afirmar que, apenas a taxa de sequências mapeadas com um genoma de referência não são suficientes para caracterizar um sequenciamento pois, a aplicação do filtro a sequências do organismo *Caenorhabditis elegans* utilizando as heurísticas: p=3, P_QV=22, ε = 3 e e_QV = 10, os dados brutos foram reduzidos de 40 M leituras para 5 M. O mapeamento destas leituras aumentou de 56% para 95% sem *mismatches*, dos dados que não passaram pelas heurísticas, 38% alinharam com até dois *mismatches*.

No entanto, apenas a taxa de mapeamento de leituras sem *missmathces* não permite a avaliação da qualidade de um sequenciamento e, a filtragem extrema dos dados ainda afeta análises posteriores como a detecção de variação, já que diminui consideravelmente o número de sequências que apresentam *mismatch*.

5 *Abordagem Probabilística para Caracterização do Sequenciamento Multiplex na Plataforma SOLiD*

"Questionamento constante e frequente
é a primeira chave para a sabedoria...
Através do duvidar que somos levados a inquirir,
e pelo inquérito nós percebemos a verdade."
Pedro Abelardo

Este capítulo abordará o desenvolvimento da abordagem probabilística para caracterização do sistema de marcação do sequenciamento *multiplex* realizado na plataforma ABI SOLiD. Inicialmente serão apresentadas características iniciais dos dados disponibilizados e a estrutura computacional envolvida. Em seguida inicia-se a descrição da abordagem proposta, a qual baseia-se nas etapas de mineração de dados proposta por (REZENDE et al., 2003), abrangendo desde a identificação do problema até a utilização do conhecimento extraído. Ao final do capítulo apresentam-se os resultados obtidos.

5.1 Materiais e Métodos

Os materiais biológicos utilizados neste trabalho foram produzidos pela segunda versão da plataforma ABI SOLiD e correspondem ao sequenciamento de cinco pacientes com câncer, de onde se extraíram duas amostras de cada. Todos os pacientes forneceram consentimento por escrito e o estudo foi aprovado pelo Comitê de Ética em Pesquisa (CEP) do Hospital Universitário João Barros Barreto (HUJBB) – Universidade Federal do Pará (UFPa), protocolo número 14052004 / HUJBB.

As dez amostras de RNA dos pacientes foram fragmentadas e preparadas utilizando-se o SOLiD *Small RNA Expression* Kit (Ambion Inc., US). Os fragmentos obtidos foram acrescido

de marcadores, no caso, os 10 primeiros *barcodes* da biblioteca padrão fornecida pela Applied para então serem amplificados e depositados em lâminas com apenas um segmento.

As sequências de interesse obtidas pelo SOLiD em sua versão dois possuem 35 bases de comprimento e iniciam sempre por uma Timina, enquanto os *barcodes* possuem 6 bases de comprimento e iniciam por uma Guanina.

As amostras foram sequenciadas duas vezes, a primeira em 19 de agosto de 2009 e o segundo em 15 de março de 2010. Entretanto, devido a uma queda no abastecimento da energia elétrica, a segunda corrida foi interrompida e recomeçou no dia 17 de março de 2010, implicando em três saídas do sequênciador que juntas, somavam 300 *Gigabytes* de dados, dos quais cerca de 50 Gb correspondiam aos arquivos "csfasta" e "qual", com as informações das sequências de interesse e marcadores.

A Tabela 5.1 condensa algumas informações sobre os dados analisados. A discrepância observada entre alguns atributos será discutida na seção a seguir.

Tabela 5.1: Dados disponíveis.

	Corrida 1	Corrida 2	Corrida 3
Data	07/09/2009	15/03/2010	17/03/2010
Número de sequências de marcação	142.453.565	29.602.637	51.832.867
Número de sequências de interesse	158.673.424	19.523.621	27.634.981
Tamanho dos arquivos	32,8Gb	4,4Gb	6,9Gb

Para a análise dos testes utilizou-se de um servidor equipado com processador Intel Xeon© modelo X3430 com frequência de 2.40GHz, 8Gb de memória secundária e 1 Tb de armazenamento. O sistema operacional é o Ubuntu Server 10.04 com arquitetura 64 bits padrão, com *Java Runtime Enviroment* em sua versão 1.6 e g++ na versão 4.3.

5.2 Abordagem Proposta

A abordagem proposta baseia-se nas cinco etapas de mineração de dados estabelecidas por (REZENDE et al., 2003) e encontra-se resumida na Figura 5.1. No presente estudo, a hierarquia de atividades segue o seguinte raciocínio: na Identificação do Problema pesquisou-se sobre as análises preliminares dos dados produzidos pela plataforma SOLiD, em especial, em corridas *multiplex*. Neste ponto encontrou-se alguns problemas nos dados disponíveis, como

descrito na Subseção 5.2.1. Posterior às diversas abordagens testadas identificou-se os atributos que deveriam ser analisados por meio de uma abordagem probabilística e desenvolveu-se uma abordagem para padronização dos dados, itens vistos na Subseção 5.2.2.

Na etapa de Identificação de Padrões, também chamada de Processamento, aplicou-se o modelo probabilístico gerado e identificou-se medidas-resumo capazes de caracterizar o sequenciamento *multiplex* da plataforma SOLiD quanto o sistema de marcação, bem como propõe uma filtragem-adaptativa dos dados, como visto na Subseção 5.2.3. Os dados processados e relatórios das atividades anteriores são encaminhados para a etapa de Pós-Processamento, responsável por condensar e apresentar o conhecimento gerado e separar os dados das sequências de interesse por amostra. Por fim, o conhecimento gerado permitirá a avaliação do sequenciamento quanto ao sistema de marcação e a separação dos dados por amostra para as análises subsequentes.

5.2.1 Identificação do Problema

O primeiro passo para as análises biológicas de dados obtidos por sequenciamento *multiplex* é separar as sequências de interesse por amostra. Nos dados analisados, essa separação é realizada detectando os *barcodes* nas sequências de marcação. Espera-se que estas sequências possuam *match* perfeito com os *barcodes* utilizados no experimento, e que sua proporção quantitativa siga uma distribuição uniforme como apresenta a Figura 5.2.

No gráfico apresentado na Figura 5.2, as barras azuis denotam a quantidade de sequências de marcação que obtiveram um *match* perfeito com os *barcodes* da biblioteca padrão, enquanto as barras laranjas representam a quantidade de sequências de marcação mapeadas com um *mismatch*, neste caso a proporção entre os marcadores tende à uniformidade como esperado.

No entanto, na análise dos dados disponíveis observou-se uma discrepância acentuada entre os *barcodes* recuperados, como mostram as Figuras 5.3, 5.4 e 5.5. A taxa de sequências de marcação que obtiveram *match* perfeito com os *barcodes*, no sequenciamento realizado em 2009 é satisfatória, 81,45%, portanto, a discrepância observada no gráfico da Figura 5.3 representa de fato uma problemática a ser investigada. As corridas realizadas em 2010 obtiveram uma taxa de *match* perfeito entre sequências de marcação baixa, respectivamente 1,63% e 48,18%, contudo a proporção faz-se representativa ao estudo.

O trabalho de (GONCALVES et al., 2010) apresenta esta problemática e propõem uma abordagem para melhorar a detecção de sequências de marcação com presença de *mismatch*. Porém, a recuperação dos 18,55% de sequências de marcação restantes não é suficiente para

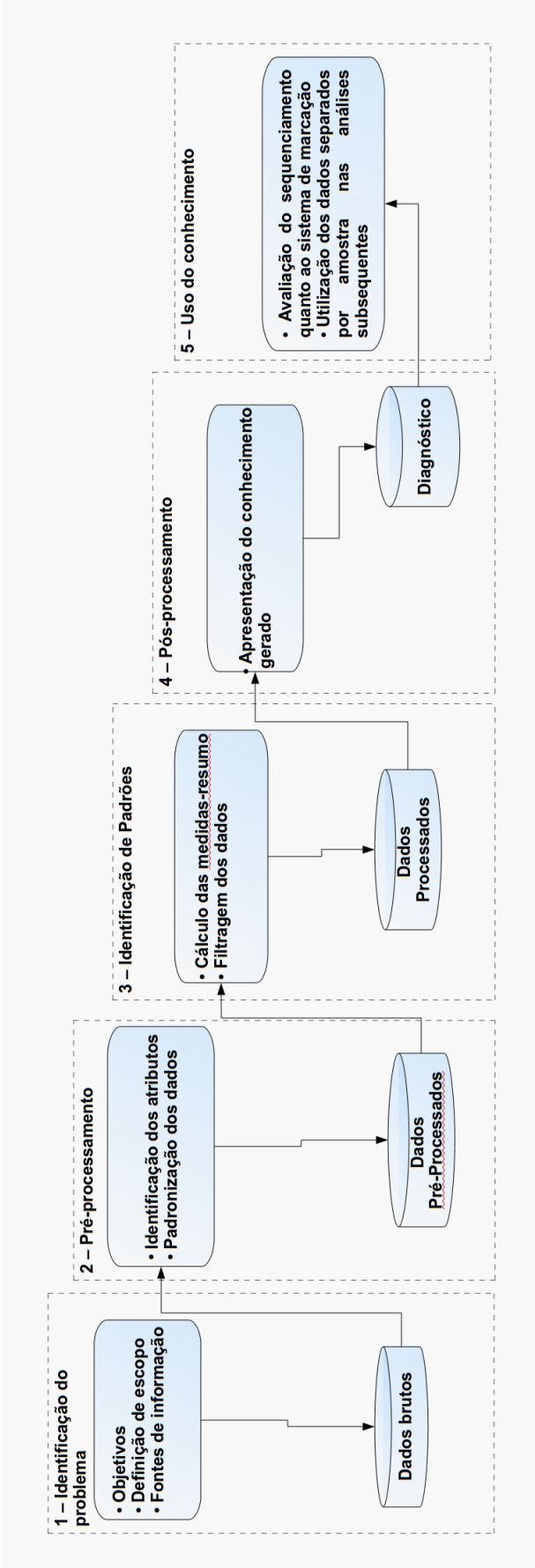


Figura 5.1: Visão geral da abordagem proposta.

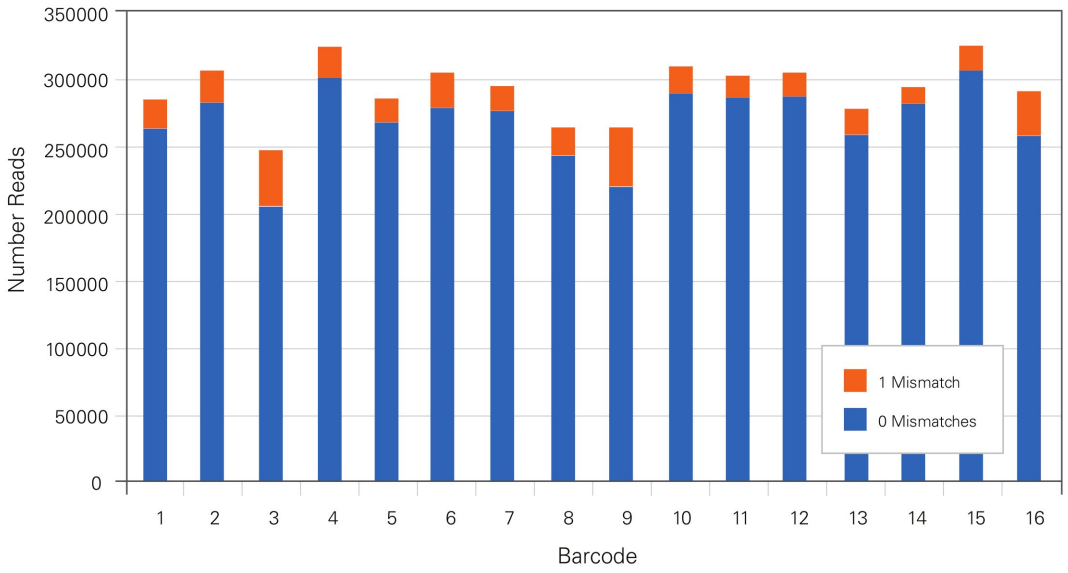


Figura 5.2: Proporção quantitativa de um experimento utilizando os 16 *barcodes*, extraído de (BIOSYSTEM, 2008).

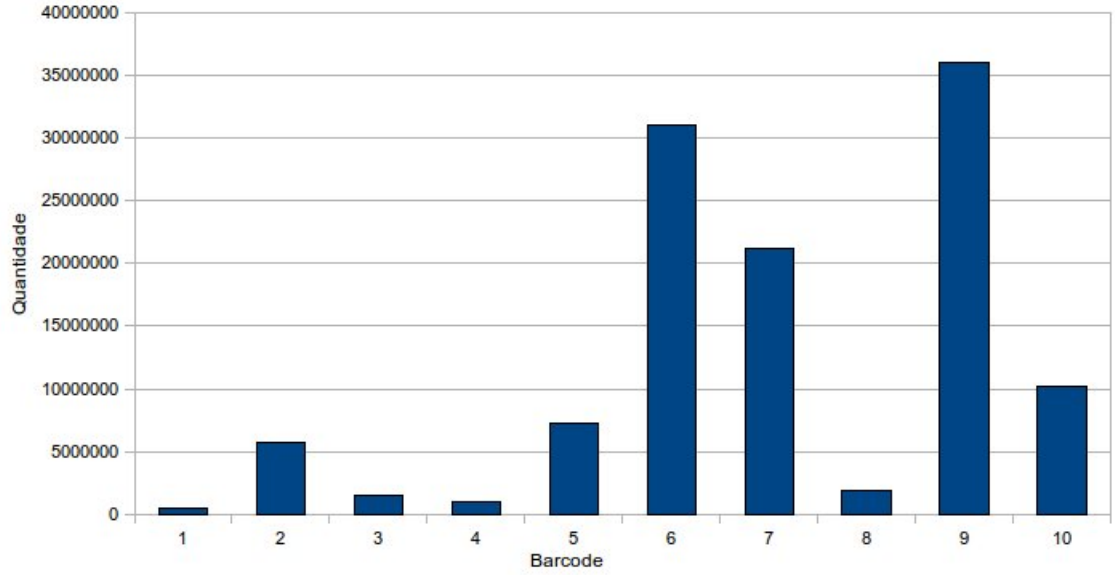


Figura 5.3: Proporção quantitativa entre os *barcodes* recuperados com *match* perfeito no sequenciamento de 2009.

suavizar a discrepância encontrada.

Neste contexto, sucederam-se diversas discussões entre a equipe de Bioinformática e Biólogos responsáveis pela preparação das amostras a fim de identificar as causas desta discrepância, algo até então não solucionado. Ademais, os dados provenientes do segundo sequenciamento contêm poucas leituras quando comparado à primeira corrida com o agravante destas leituras possuírem baixa qualidade.

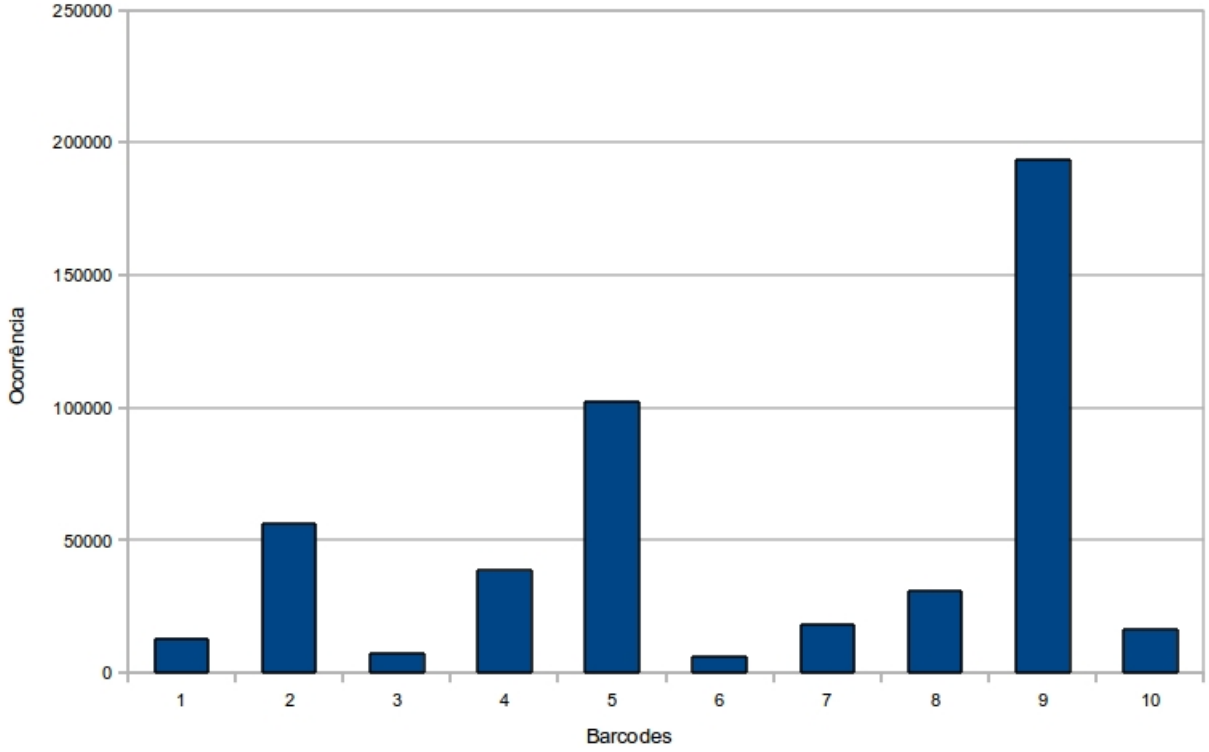


Figura 5.4: Proporção quantitativa entre os *barcodes* recuperados com *match* perfeito no primeiro sequenciamento 2010.

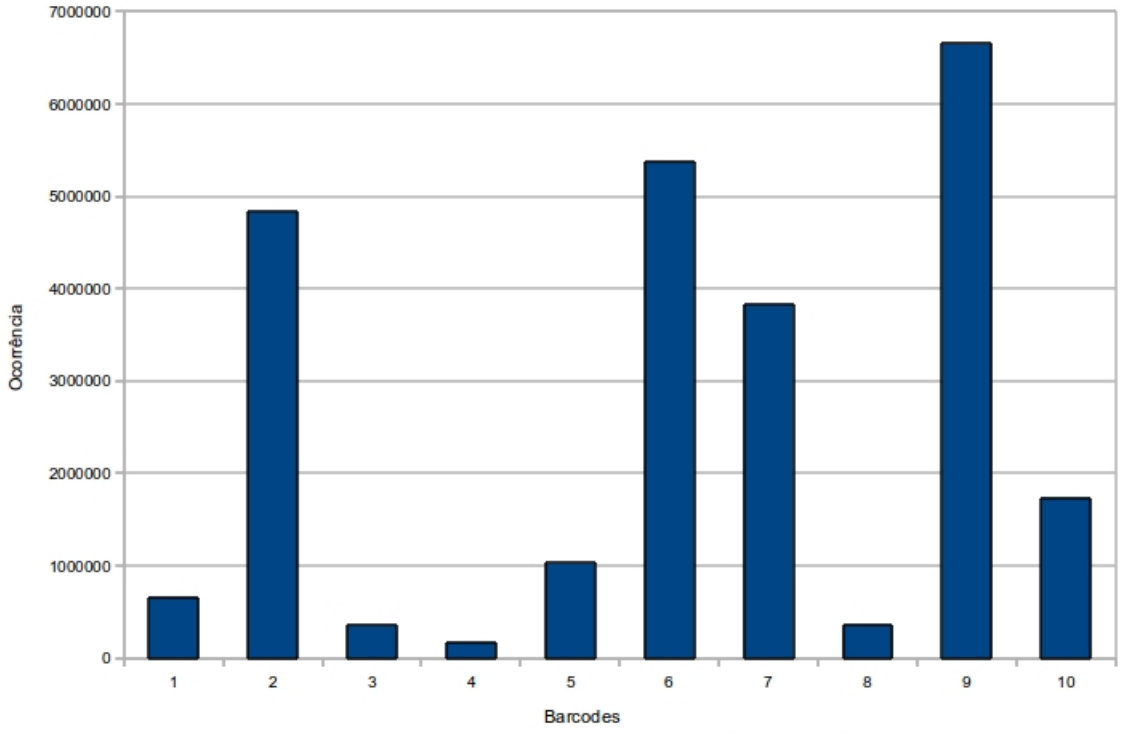


Figura 5.5: Proporção quantitativa entre os *barcodes* recuperados com *match* perfeito no segundo sequenciamento de 2010.

Outro item identificado nesta etapa foi o não pareamento entre os Identificadores dos arquivos com informações do sistema de marcação (R3) e com informações das sequências de interesse (F3), ou seja, haviam IDs em um par de arquivos que estavam ausentes no outro par. Isto ocorre, pois o sistema computacional acoplado ao sequênciador, responsável por transformar as imagens em *colorspace*, associar o valor de qualidade respectivo e pré-filtrar dados de baixa qualidade, não suporta a grande vazão de dados produzidos, permitindo esta falha.

O encadeamento das descobertas destes problemas suscitaram mais discussões entre as equipes, onde geravam-se muitas conjecturas e poucas conclusões. Diante disto, fez-se necessário o desenvolvimento de um estudo com enfoque matemático, capaz de caracterizar o sistema de marcação utilizado em corridas *multiplex* na plataforma SOLiD. Permitindo, dentre outras coisas, a avaliação dos protocolos utilizados na preparação das amostras e a análise de confiabilidade dos dados obtidos.

5.2.2 Pré-Processamento

Nesta etapa, os dados obtidos na saída do sequênciador devem ser preparados para a separação de dados por amostras, visando o início das análises biológicas. No âmbito da mineração de dados, esta etapa consome cerca de 80% do tempo total destinado à extração de padrões. O presente trabalho não foge desta proporção, tanto pelo tamanho dos arquivos quanto pela complexidade do problema.

Implementação

O tamanho acentuado dos arquivos torna necessário o desenvolvimento e utilização de técnicas de programação otimizadas, evitando ao máximo leitura/escrita, atividades ainda que representam um gargalo (responsável pela queda de desempenho) em qualquer sistema computacional (OUSTERHOUT; DOUGLIS, 1989) (STALLINGS, 2002) (TANENBAUM, 2006), por isso, realizou-se uma extensa pesquisa de algoritmos aplicáveis à esta problemática assim como diversos testes.

Neles, os primeiros algoritmos desenvolvidos para as análise dos dados disponíveis foram implementados na Linguagem Java em sua Versão 1.6. Devido extensas discussões acerca de custo de processamento e consumo de memória desta linguagem. O código-fonte Java é compilado em *bytecodes* os quais são interpretados pela *Java Virtual Machine* (JVM) (GOSLING; MCGILTON, 1996), pensou-se que a ineficiência das ferramentas desenvolvidas era causada pela linguagem adotada. Outro fato que corroborou esta impressão foi a leitura do texto inti-

tulado: *How Perl Saved the Human Genome Project* (STEIN, 1996) e a grande veiculação de códigos escritos em Perl, em livros-texto de Bioinformática (LESK, 2008).

Alguns aspectos confrontavam com esta tendência, como o fato de ferramentas implementadas nesta linguagem a exemplo de (SALMELA, 2010). Para dirimir esta dúvida, traduziu-se os algoritmos para outras linguagens como Perl e C++, os resultados obtidos divergiram do esperado, com Java mostrando-se mais eficiente que Perl. Apenas o desempenho de C++ foi como o esperado, reduzindo o custo de processamento e de memória.

Em (FOURMENT; GILLINGS, 2008) encontraram-se as respostas para este questionamento. Nele, os autores implementaram diversos algoritmos como BLAST e Neighbor-Joining em C, C++, C, Java, Perl e Phyton. Os resultados reiteraram os testes realizados, C, C++ e Java apresentaram desempenho similar na maior parte dos algoritmos comparados em (FOURMENT; GILLINGS, 2008), enquanto Perl e Python, linguagens puramente interpretadas apresentaram desempenho inferior ao outro grupo.

Portanto, escolheu-se utilizar uma abordagem híbrida. Tarefas que necessitam de alto consumo de memória e processamento foram implementadas em C. Tarefas mais triviais foram codificadas em Java, com o objetivo de se obter um equilíbrio entre desempenho e tempo de codificação por que a linguagem Java, apesar de consumir bastante processamento, possui uma grande quantidade de rotinas já implementadas diminuindo o tempo destinado a codificação, como apontado em (FOURMENT; GILLINGS, 2008).

Pareamento de Arquivos

Como apresentado na Subseção 5.2.1, os arquivos gerados pela plataforma SOLiD apresentam uma falha quanto ao não pareamento dos Identificadores. É necessário a realização de um procedimento computacional para encontrar a interseção dos Identificadores entre os arquivos contendo as informações sobre o sistema de marcação e os arquivos que possuem as informações das sequências de interesse, como mostra a Figura 5.6.

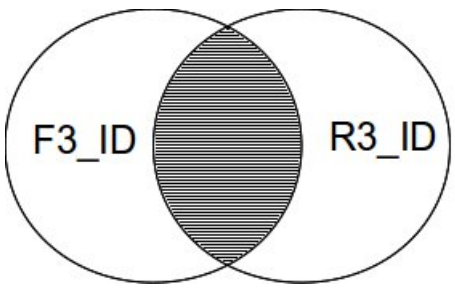


Figura 5.6: Diagrama da interseção entre os IDs dos arquivos R3 e F3.

A estratégia utilizada para realizar este pareamento foi utilizar uma estrutura de dados chamado de Mapa (CORMEN et al., 2002), composta por dois campos: chave e valor. Neste caso, a chave será o ID e o valor receberá um contador o qual é incrementado a cada vez que o ID seja verificado em um arquivo. Ao final, apenas os IDs que tiverem um valor igual a quatro (número de arquivos) serão recuperados. Aliado à esta estrutura de dados utilizou-se também rotinas de leitura e escrita populares em competições de programação, conhecidas como *Fast-I/O*, por serem mais eficientes do que os métodos de entrada e saída padrão de C/C++ uma vez que suprimem alguns métodos de verificação, desnecessários neste tipo de aplicação. Os códigos-fonte com estes métodos encontram-se dispostos no Anexo 1.

A taxa de Identificadores pareados permite verificar a necessidade de se realizar um *upgrade* no *cluster* acoplado ao sequênciador, pois uma grande quantidade de IDs não pareados denota uma baixa eficiência frente ao volume de dados gerado, desperdiçando espaço em disco e implicando na necessidade de preparação dos dados antes das análises biológicas.

Grau de Confiança da Transição Di-base

A plataforma SOLiD associa à transição di-base detectada pelo sequênciador um valor de qualidade, compreendido de -1 à 35. O valor negativo indica ausência de leitura, enquanto 35 indica o valor máximo de confiança associado a uma transição. Visando utilizar uma abordagem probabilística é necessário converter a qualidade fornecida pela plataforma em um valor de probabilidade que satisfaça os Axiomas da Probabilidade (vide Capítulo 3).

Para tanto, solicitou-se esta informação à *Applied Biosystems*, contudo, por motivos mercadológicos a fabricante da plataforma não disponibilizou a função que realizasse o mapeamento entre Valor de Qualidade e Probabilidade. Em (EWING; GREEN, 1998) encontrou-se uma função que, adaptada ao intervalo dos valores de qualidade informados pela plataforma SOLiD, consegue aproximar o mapeamento Qualidade-Probabilidade, como segue:

$$P(Q) = 1 - 10^{-\frac{(Q+1)}{10}} \tag{5.1}$$

Na Equação 5.1 Q é o Valor de Qualidade informado pela plataforma, o qual sofre uma normalização a fim de $P(Q)$ assuma apenas valores positivos. O gráfico gerado é apresentado na Figura 5.7, reitera-se que todos os valores de $P(Q)$ obedecem aos Axiomas da Probabilidade.

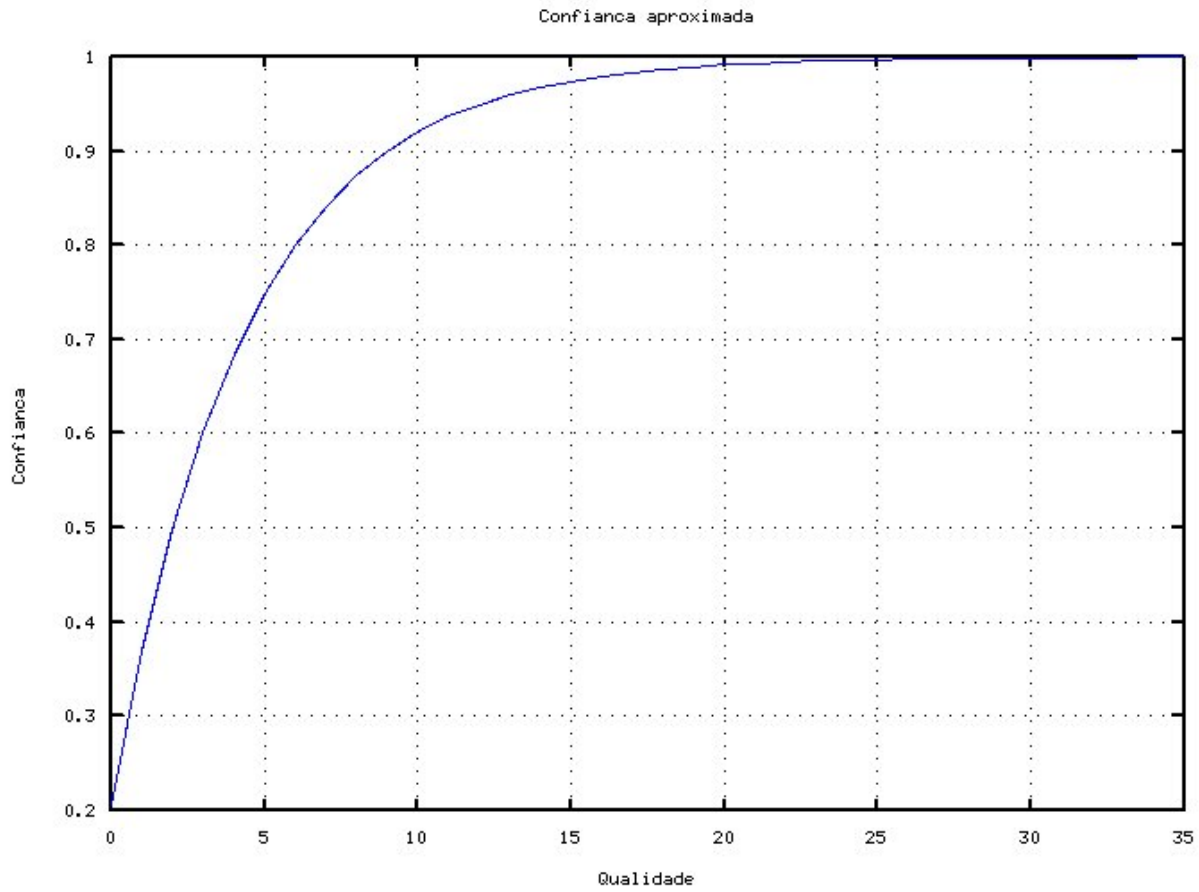


Figura 5.7: Mapeamento Qualidade-Probabilidade obtido por meio da Equação 5.1.

Grau de Confiança de Uma Sequência

A Equação 5.1 apresenta o grau de confiança de uma transição di-base, no entanto, pretende-se estudar o grau de confiança de uma sequência, objetivando o estudo da qualidade do sistema de marcação utilizado em corridas *multiplex*. Para tal, modelou-se uma sequência como uma Cadeia de Markov, com o código em *colorspace* sendo os estados possíveis $S = \{A, T, C, G, 0, 1, 2, 3\}$ e $P(Q)$ obtido sendo a probabilidade de transição. O espaço-amostral S contém as bases nucleotídicas, pois são elas que iniciam uma sequência em *colorspace*, a partir da primeira base é que segue o código numérico, como mostra o exemplo da Figura 5.8.

A Figura 5.8 apresenta a base nucleotídica Guanina iniciando a sequência, como acontece com os *barcodes* e seguindo-se duas transições "00", em *basespace* a sequência seria "GGG". Os valores de qualidade associados são "10 24", que, quando aplicado a Equação 5.1, obtêm-se os valores de confiança 0,92057 e 0,99684. Para se calcular o grau de confiança da sequência θ multiplicam-se as probabilidades de todas as transições existentes, como mostra a Equação 5.2.

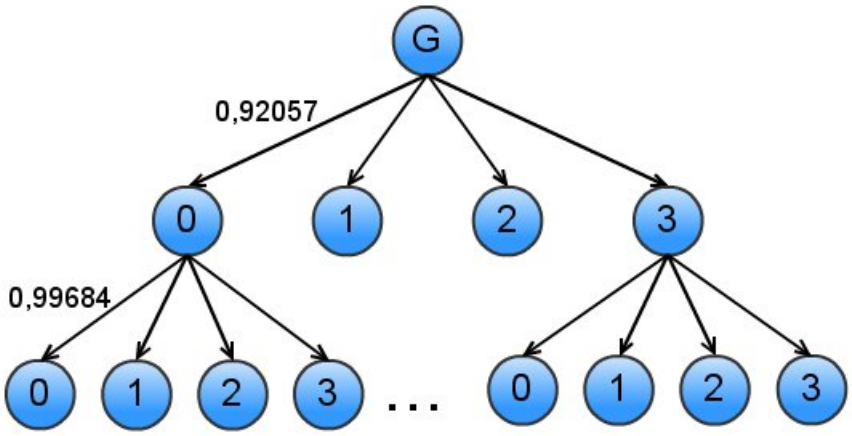


Figura 5.8: Cadeia de Markov que modela a sequência “00” com qualidade associada “10 24”.

$$P(\theta) = \prod P(Q) \tag{5.2}$$

No exemplo apresentado, a probabilidade da sequência “G00” não ser produto do acaso é 0,91766, resultado da multiplicação dos valores de confiança 0,92057 e 0,99684. Este modelo probabilístico permite, dentre outros pontos, a descoberta de medidas-resumo capazes de caracterizar o sequenciamento quanto ao sistema de marcação; guiar um processo de filtragem, respeitando as características intrínsecas de cada corrida multiplex; e desenvolver um processo de recuperação de sequências que não correspondem exatamente com algum *barcode* da biblioteca padrão.

Considerações

Nesta etapa do desenvolvimento do trabalho pesquisou-se diversas técnicas e tecnologias aplicadas à analise de dados biológicos, optando-se por codificar as tarefas que requerem maior carga de processamento em C/C++ e tarefas triviais em Java. A tarefa de pareamento dos Identificadores utilizou-se de métodos de leitura e escrita mais eficientes pois, suprimem métodos de verificação desnecessários neste tipo de aplicação. Para se calcular o grau de confiança de uma sequência em termos probabilísticos, adaptou-se uma função capaz de mapear o Valor de Qualidade fornecido pela plataforma SOLiD em um Valor de Probabilidade. Modelou-se a sequência, dada em *colorspace*, como uma Cadeia de Markov com o mapeamento obtido sendo a probabilidade de transição. A confiança de determinada sequência é obtida pelo produtório de todas as transições existentes.

5.2.3 Identificação de Padrões

A etapa de Pré-processamento buscou a preparação dos dados para as análises e o desenvolvimento de um modelo probabilístico capaz de mapear os valores de qualidade fornecidos pelo sequênciador em valores probabilísticos. Na etapa de processamento buscou-se utilizar as probabilidades obtidas para se caracterizar o sistema de marcação do sequenciamento *multiplex* da plataforma SOLiD.

Neste ponto, estudou-se métodos estatísticos capazes de realizar a tarefa desejada culminando na seleção das seguintes medidas-resumo: média, moda, variância, Taxa de Identificadores Pareados, Taxa de Sequências de marcação que apresentam *match* perfeito com os *Barcodes* além da Proporção dos *Barcodes* encontrados.

As três primeiras medidas auxiliam, dentre outros pontos, na análise do protocolo de preparação de amostras para corrida multiplex, dado que se tem a visão geral do grau de confiança do sistema de marcação; e na elaboração e adoção de uma estratégia de filtragem dos dados, que pode ser tanto como no trabalho de (SASSON; MICHAEL, 2010), utilizando-se heurísticas baseadas no Valor de Qualidade puro, quanto no Grau de Confiabilidade desenvolvido no presente estudo.

A Taxa de Identificadores Pareados é obtida por meio da contagem do número de sequências dos Dados Brutos e dos Dados Pré-Processados, permitindo a avaliação do desempenho do *cluster* acoplado ao sequênciador. Uma alta Taxa de Identificadores Pareados indica que o *cluster* processou os arquivos corretamente e, infere-se, que a carga de processamento destinada à pré-filtragem foi baixa, denotando uma boa qualidade no sequenciamento como um todo.

A Taxa de Sequências de Marcação que apresentam *match* perfeito com os *Barcodes* da biblioteca padrão é uma das medidas mais importantes para a abordagem proposta, juntamente com a Proporção dos *Barcodes* encontrados. Pois uma taxa superior a 75% e uma proporção tendendo a uniformidade denota que não houveram falhas na etapa de preparação das amostras, sendo assim, tem-se um sequenciamento confiável quanto ao sistema de marcação.

Nesta etapa também acontece a filtragem dos dados, que é realizada de duas formas: baseada no Valor de Qualidade puro e baseada na abordagem probabilística. A primeira utilizou a heurística proposta por (SASSON; MICHAEL, 2010), sequências de marcação que possuíam pelo menos uma transição com qualidade inferior a um *threshold* eram descartadas. A segunda abordagem aplicou o modelo probabilístico proposto, utilizando o Grau de Confiança da sequência. Os valores de corte utilizados e os resultados obtidos são apresentados na subseção 5.3.

5.2.4 Pós-Processamento

A etapa de Pós-Processamento tem por objetivo prover a visualização das informações coletadas nas etapas anteriores, bem como encaminhar os Dados Processados para as análises biológicas subsequentes. Neste âmbito, optou-se por apresentar, além de uma tabela com as medidas-resumo obtidas, a Função Densidade de Probabilidade tendo em vista que a representação pictórica dos dados é mais intuitiva (VIEIRA, 2008), o que facilita na interpretação do conhecimento extraído. As Figuras 5.9, 5.10 e 5.11 apresentam as Funções Densidade de Probabilidade das Corridas 1, 2 e 3, respectivamente.

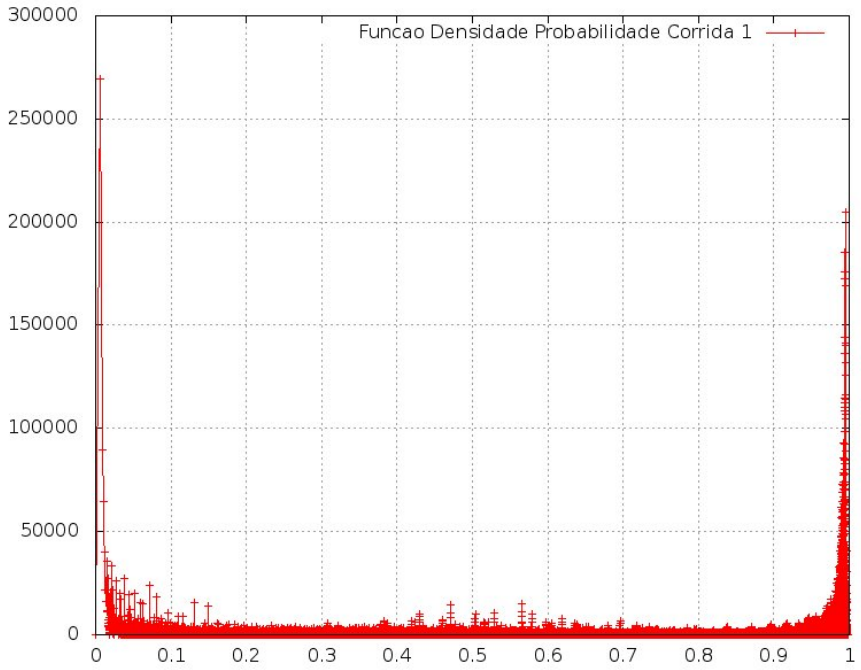


Figura 5.9: Função Densidade de Probabilidade da Corrida 1.

Espera-se que ocorra uma alta concentração na extremidade direita do gráfico e ausência de picos em demais localidades, denotando uma boa qualidade de sequenciamento e baixa variabilidade no Grau de confiança. Primeiro gráfico apresenta alta concentração nas pontas, possuindo também sobressaltos na faixa de 0,4 à 0,5, enquanto o segundo possui apenas a concentração próxima a zero. Já o gráfico da Figura 5.11 possui concentração tanto em probabilidades baixas quanto altas, no entanto, a alta frequência da Moda fez com que a escala do gráfico fosse aumentada para 800.000, dificultando a análise gráfica. Deste modo faz-se necessário a comparação das medidas-resumo e disposição desta informação em formato tabular como apresentado na seção a seguir, que também disserta sobre os possíveis causadores das problemáticas identificadas.

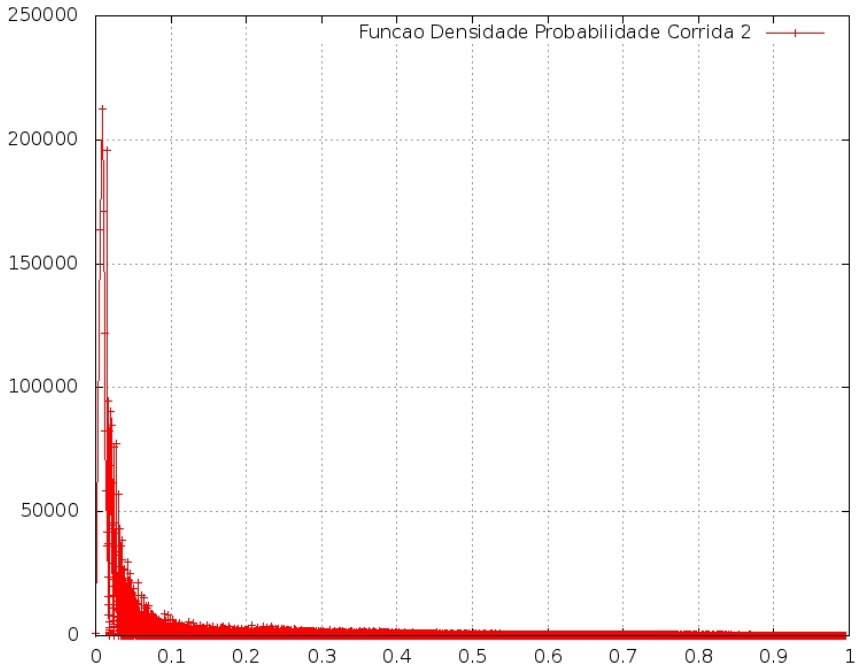


Figura 5.10: Função Densidade de Probabilidade da Corrida 2.

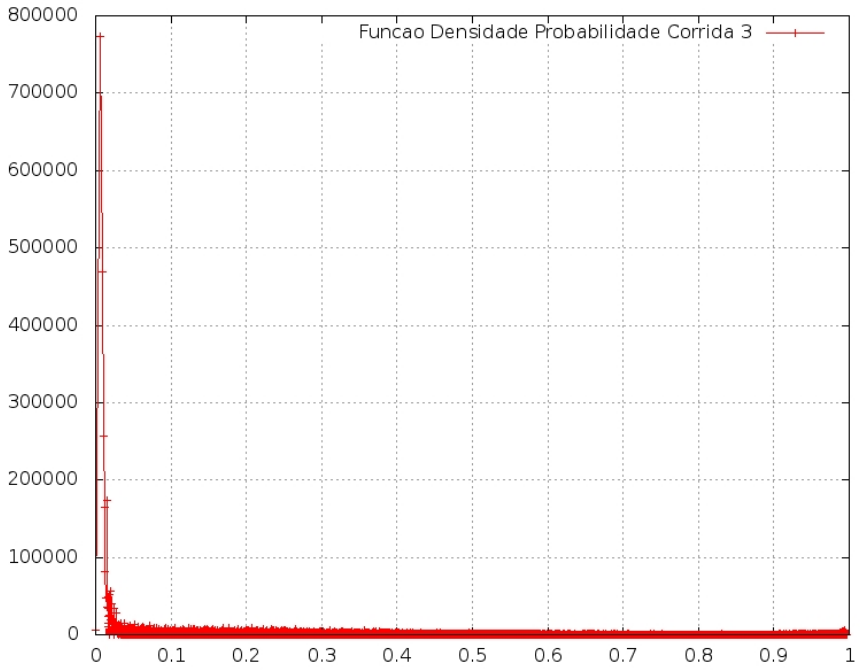


Figura 5.11: Função Densidade de Probabilidade da Corrida 3.

5.3 Apresentação e Discussão dos Resultados

A análise inicial dos dados permitiu a identificação de uma discrepância na proporção de ocorrência dos *barcodes* detectados sem *mismatch*. Este fato foi o principal motivador do presente estudo. A análise dos gráficos apresentados nas Figuras 5.3, 5.4 e 5.5 permitiu a iden-

tificação da alta ocorrência dos *barcodes* 5, 6, 7, e 10; destes, destaca-se o *barcode* 9, com frequência elevada nos três sequenciamentos, como mostra a Figura 5.12.

Dentre as possíveis causas deste problema estão: falha na preparação das bibliotecas, falhas no depósito das *beads* nas lâminas além de problemas com os reagentes, em particular com o *Small RNA Expression Kit* utilizado no processo.

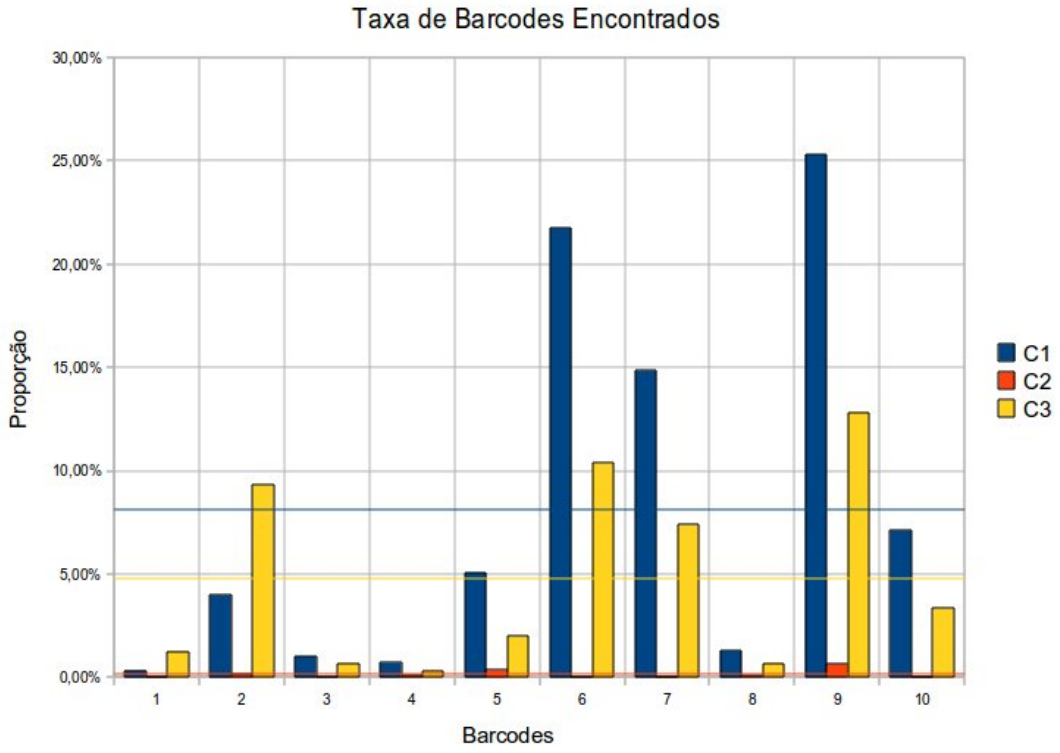


Figura 5.12: Proporção de Barcodes em Níveis Percentuais.

Nas etapas de Pré-Processamento e Processamento os arquivos são pareados para se calcular as medidas-resumo e realizar a filtragem dos dados. As primeiras informações encontram-se dispostas na Tabela 5.2 enquanto os dados sobre as filtragens realizadas encontram-se nas Tabelas 5.3 e 5.4.

Como esperado, a Taxa de *Barcodes* Pareados foi superior na Corrida 1 (C1), aproximadamente 85%, enquanto a Corrida 3 apresentou o pior resultado, 67%. Isto ocorreu pois o número de sequências presentes nos arquivos R3 é quase duas vezes superior ao número de sequências encontradas nos arquivos F3. O motivo plausível desta diferença reside no fato das amostras terem permanecido do dia 15 ao dia 17 de março no sequênciador, nestes dois dias as sequências mais longas sofreram uma degeneração, sendo descartadas pelo pré-filtro da plataforma.

A Taxa de *Match* perfeito em C2 foi muito abaixo do esperado, aproximadamente 1,5%. Este fato, *per se*, indica que os dados deste sequenciamento não devem ser utilizados em análises

Tabela 5.2: Medidas-resumo extraídas dos dados disponíveis.

	Corrida 1 (C1)	Corrida 2 (C2)	Corrida 3 (C3)
Taxa de Barcodes Pareados	84%	69%	67%
Taxa de *Match* Perfeito	81,44%	1,63%	48,18%
Média	81,23%	47,15%	59,17%
Moda	30,88%	37,26%	30,88%
Variância	5,53%	6,87%	8,06%

posteriores. Aliado à este fato, esta corrida apresenta uma média do Grau de Confiança abaixo dos 50% com baixa variabilidade de 6,87%, denotando alta concentração de leituras com baixa qualidade, mesmo que a Moda aponte um valor superior ao das outras corridas, 37,26%. A Corrida 3, apesar de ter a menor Taxa de Pareamento e uma alta variabilidade, 8,06%, mostrou-se razoável nas outras medidas-resumo.

As análises dos resultados obtidos das últimas corridas permite inferir que o causador da baixa confiança de C2 foi a queda na qualidade no abastecimento, e posterior interrupção do fornecimento de energia elétrica, dado que o mau desempenho não foi mantido na retomada do sequenciamento. Já as falhas ocorridas na Corrida C3 foi advinda do tempo de permanência excessivo no sequênciador, ocasionando degeneração das sequências e possível desprendimento das *beads* devido ao decaimento magnético natural.

Tabela 5.3: Resultados das filtragens baseadas no Valor de Qualidade.

Qualidade	Corrida 1 (C1)	Corrida 2 (C2)	Corrida 3 (C3)
6	65,68%	20,15%	32,27%
7	60,78%	14,16%	26,38%
8	58,51%	10,85%	23,62%
9	55,09%	7,85%	20,12%
10	53,11%	6,30%	18,40%
12	47,23%	3,18%	13,78%
13	45,03%	2,45%	12,40%
15	39,02%	1,13%	9,06%

A Tabela 5.3 apresenta a taxa de sequências que não apresentam transição com Valor de Qualidade inferior ao *threshold* estipulado. Por exemplo, a segunda linha da referida Tabela apresenta a taxa de sequências que possuem transições com qualidade superior à 6. O mesmo acontece na Tabela 5.4, contudo esta abordagem utiliza o Grau de Confiança como *threshold*. Comparando-se os resultados das duas estratégias de filtragem, percebe-se que a baseada no valor de qualidade realiza cortes mais abruptos, dado sua natureza exponencial, enquanto a abordagem probabilística permite a flexibilização da filtragem, já que suaviza o nível de corte.

As figuras 5.13 e 5.14 apresentam o percentual de sequências de marcação que passaram

Tabela 5.4: Resultados das filtragens baseadas no Grau de Confiança.

Grau de Confiança	Corrida 1 (C1)	Corrida 2 (C2)	Corrida 3 (C3)
70%	62,7%	15,44%	27,94%
75%	59,79%	11,68%	24,51%
78%	58,23%	9,76%	22,74%
80%	56,84%	8,50%	21,36%
82%	55,74%	7,40%	20,24%
85%	53,57%	5,76%	18,23%
88%	50,91%	4,19%	16,05%
90%	48,70%	3,22%	14,43%
95%	40,04%	3,22%	9,35%

pelo processo de filtragem, com constantes: 75%, 80% e 90% para o Grau de Confiança; e correspondentes 8, 10 e 12 para o valor de qualidade. A comparação entre as duas figuras ratificam o mapeamento entre valor de qualidade e Grau de Confiança. A única taxa das figuras que não seguem o padrão refere-se à corrida dois, pois foi acometida por anomalias.

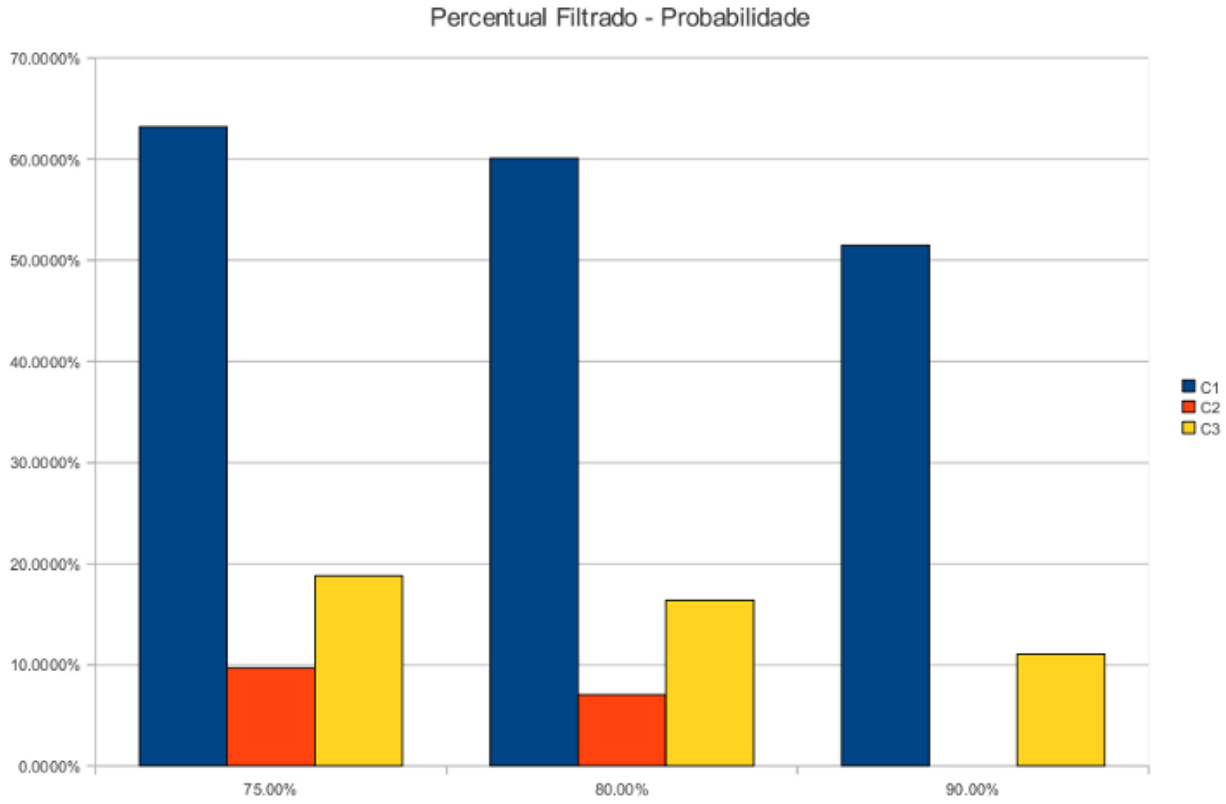


Figura 5.13: Taxa de sequências filtradas baseado no Grau de Confiança.

As figuras 5.15 e 5.16 apresentam a ocorrência dos *barcodes* após a filtragem com Grau de Confiança 70% e 90% respectivamente. É possível notar que a discrepância na ocorrência dos *barcodes* persiste mesmo aumentando o valor de corte da filtragem, isso reafirma a existência de falhas na etapa de preparação das amostras.

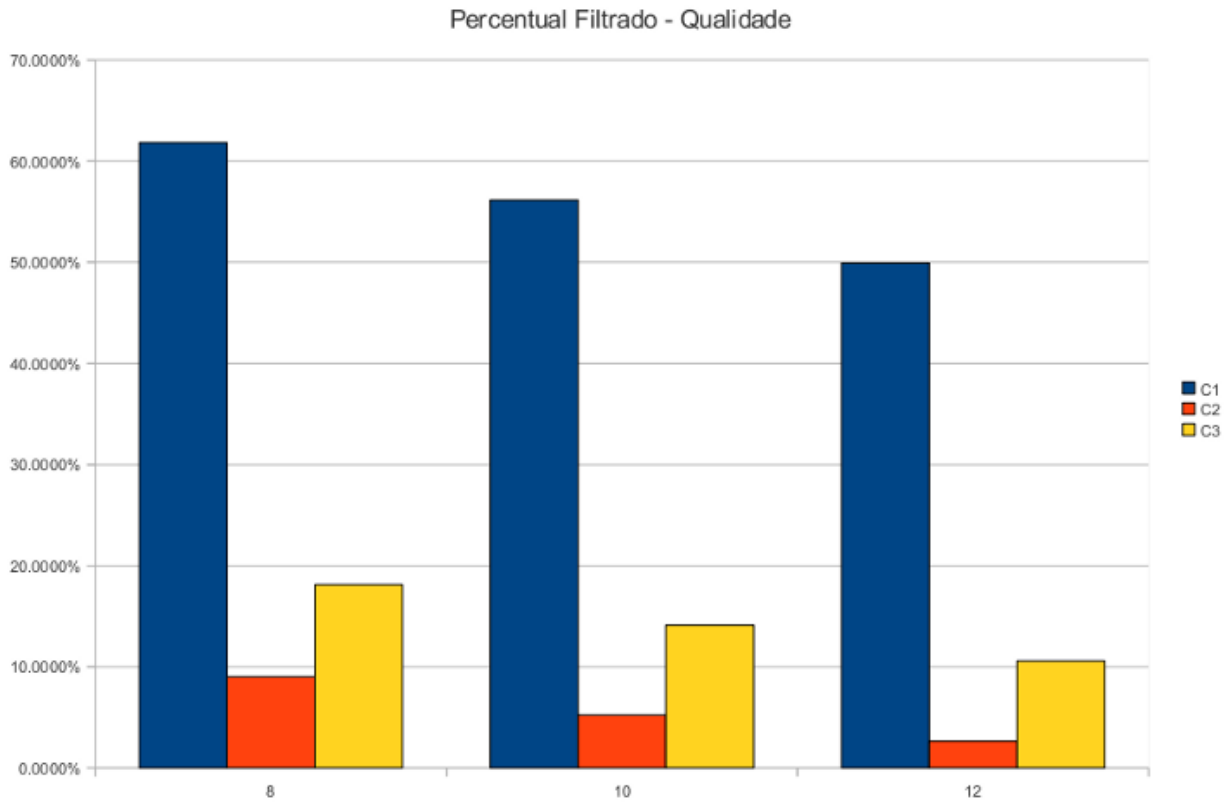


Figura 5.14: Taxa de sequências filtradas baseado no valor de qualidade.

Convém destacar que a aplicabilidade desta abordagem não se limita apenas ao sistema de marcação, podendo ser expandida para as sequências de interesse, caracterizando o sequenciamento como um todo. Quanto a filtragem, dado a suavização da curva de corte proporcionada pela abordagem probabilística,

5.4 Considerações Finais

Este capítulo apresentou a abordagem probabilística para caracterização do sistema de marcação do sequenciamento *multiplex* na plataforma SOLiD, tema desta dissertação. O desenvolvimento da abordagem proposta utilizou-se do processo de Extração de Conhecimento de Base de dados apresentado por (REZENDE et al., 2003), tendo como etapas a identificação do problema, o Pré-Processamento, a Identificação de Padrões, o Pós-Processamento e a utilização do conhecimento extraído.

Na etapa de identificação do problema, pesquisou-se sobre as análises preliminares dos dados de sequenciamento *multiplex* produzidos pela plataforma SOLiD e analisou-se a integridade e consistências dos dados. Neste âmbito, identificou-se uma discrepância acentuada na

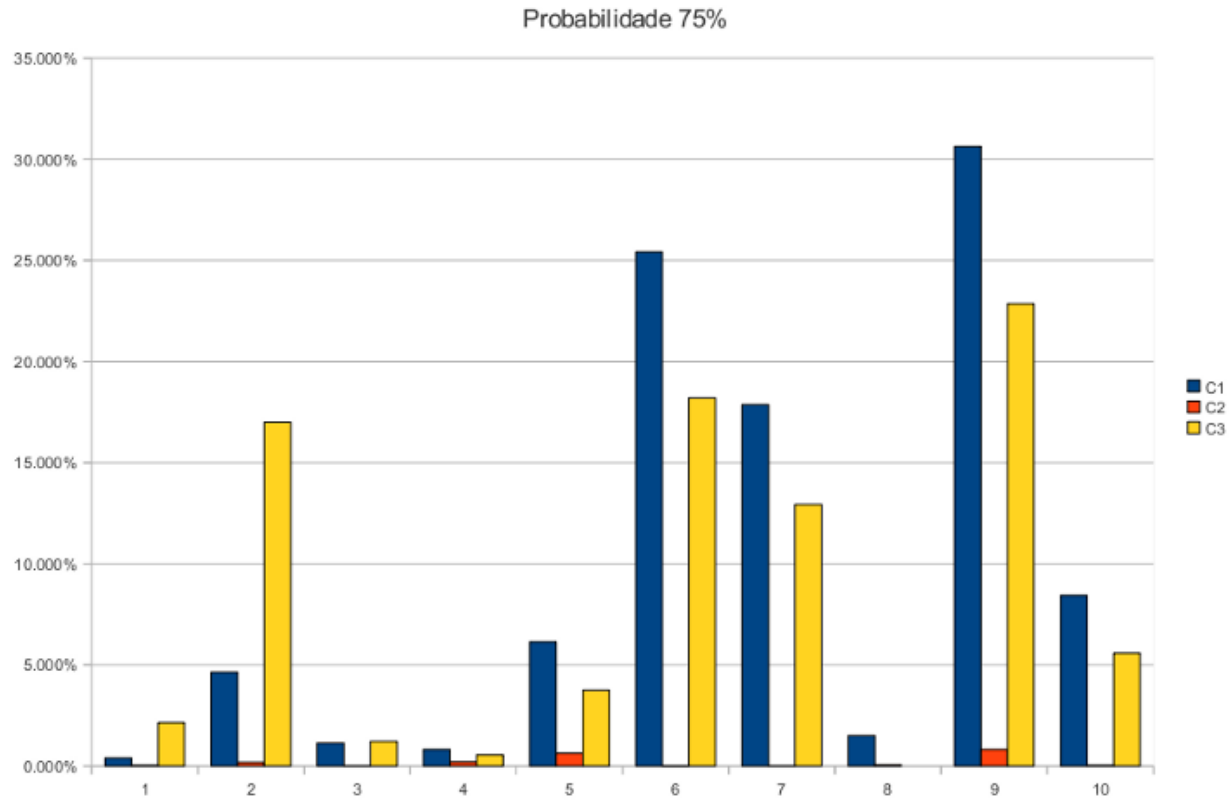


Figura 5.15: Ocorrência de *barcodes* filtrados para o Grau de Confiança de 75%.

ocorrência de determinadas sequências de marcação, os *barcodes* mais frequentes foram 5, 6, 7, e 10, em detrimento aos outros seis utilizados nos experimentos. Outro item identificado foi não pareamento entre os Identificadores dos arquivos com informações do sistema de marcação (R3) e com informações das sequências de interesse (F3), o que implica na necessidade de um passo para que os arquivos fossem pareados.

Esta tarefa foi realizada na fase de Pré-Processamento, que também incluiu pesquisas acerca de algoritmos e linguagens de programação mais adequados às problemáticas observadas. Outro ponto estudado nesta etapa foi a busca por uma função que mapeasse o Valor de Qualidade fornecido pelo SOLiD em um valor de probabilidade, por ser uma tecnologia proprietária, a *Applied Biosystems* não forneceu esta função. Para contornar este empecilho, adaptou-se uma função proposta por (EWING; GREEN, 1998) a qual realizou a contento o mapeamento requerido.

Com a função obtida, foi possível calcular a probabilidade de uma sequência não ter sido obtida ao acaso. Dado que uma sequência em *colorspace* pode ser modelada segundo uma Cadeia de Markov, com os seguintes estados possíveis: $\{A,T,C,G,0,1,2,3\}$ e a probabilidade calculada como a probabilidade de transição.

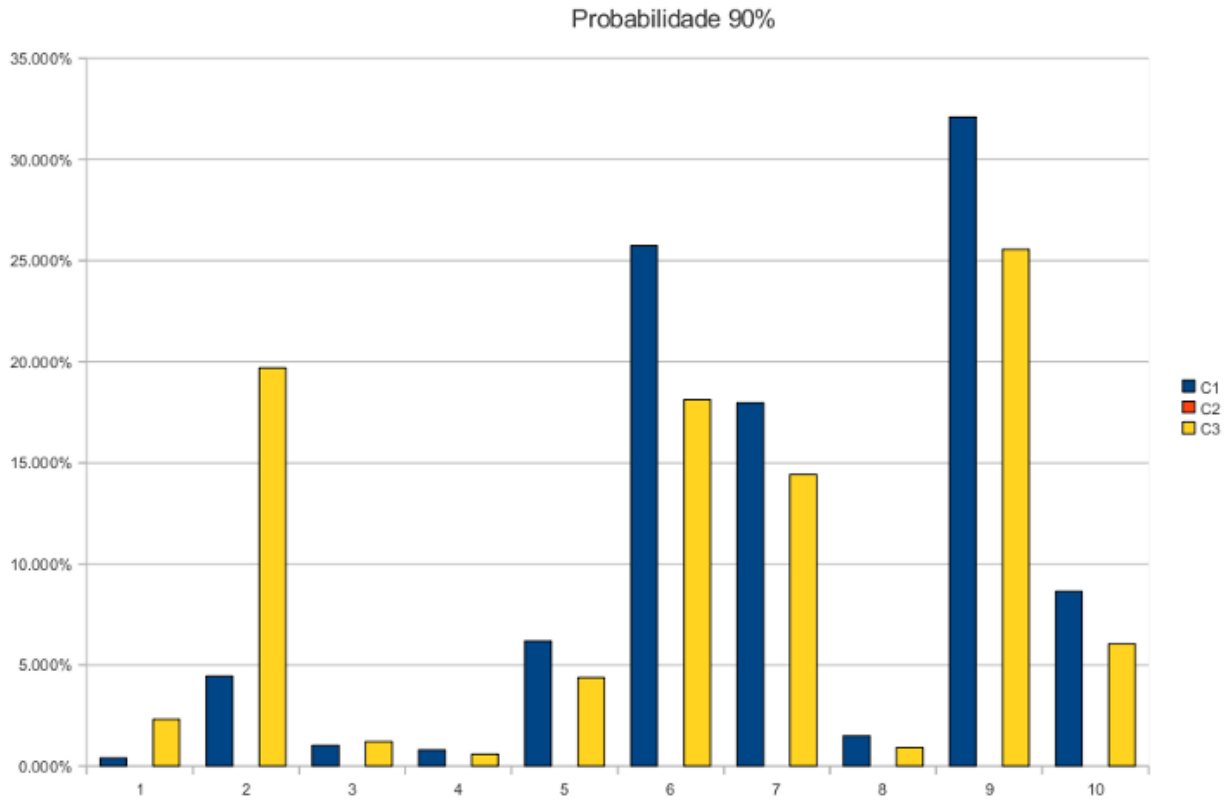


Figura 5.16: Ocorrência de *barcodes* filtrados para o Grau de Confiança de 90%.

A probabilidade final, denominada de Grau de Confiança, foi utiliza na etapa de Identificação de Padrões. Nela definiu-se as seguintes medidas-resumo capazes de caracterizar o sequenciamento *multiplex* quanto ao sistema de marcação: média, moda, variância, Taxa de Identificadores Pareados, Taxa de Sequências de marcação que apresentam *match* perfeito com os *Barcodes* além da Proporção dos *Barcodes* encontrados. Na etapa de Pós-Processamento, foram plotados os gráficos com a densidade de probabilidade dos dados e disponibilizou-se as medidas-resumo em formato tabular, como apresentado no Anexo 2.

Dentre as aplicações da utilização do conhecimento gerado, ressalta-se a avaliação dos passos realizados no sequenciamento *multiplex*, permitindo a identificação de faltas em algum passo do processo, e a adaptação e desenvolvimento de novos protocolos para preparação de amostras.

6 *Conclusões e Trabalhos futuros*

"A análise de longo prazo, necessariamente, inclui o exame das consequências de curto prazo.""

Ludwig von Mises

As novas tecnologias de sequenciamento pertencentes à classe *High-Throughput Sequencing*(HTS) como as plataformas 454 da Roche, o Illumina da Solexa e o SOLiD desenvolvido pela *Applied Biosystems*, modificaram a forma como as pesquisas biológicas eram feitas devido ao grande volume de dados produzidos, comumente acima dos 10 Gigabytes de arquivos-texto. Esta grande quantidade de informação requer auxílio computacional para que sejam realizadas as análises necessárias, visando se identificar e validar informações biologicamente úteis.

Cada plataforma de sequenciamento possui características intrínsecas no que dis respeito a disposição dos dados. A plataforma SOLiD, por exemplo, utiliza-se de um sistema de codificação próprio e permite também, o sequenciamento de múltiplas amostras em uma única corrida, denominada de corrida *multiplex*, por meio de um sistema de marcação chamado *barcode*. Esta funcionalidade requer um processo computacional para separação dos dados, pois o sequenciador fornece a mistura de todas amostras em uma única saída.

Motivação

A recuperação das sequências, por amostra, depende da detecção dos *barcodes* nas sequências de marcação. Logo, uma discrepância nesta proporção e a presença de sequências de marcação que não correspondem aos *barcodes* da biblioteca padrão, prejudicam as análises biológicas subsequentes. Outro possível problema a ser investigado é a presença de sequências com baixa qualidade tanto no sistema de marcação quanto nas sequências de interesse, o que prejudicaria a confiabilidade do sequenciamento. Estes problemas foram observados no estudo de caso, esperava-se uma distribuição uniforme entre os 10 *barcodes* utilizados no experimento, contudo, observou-se uma alta ocorrência dos *barcodes* 6, 7 e 9, em detrimento aos demais.

Estes problemas advêm de falhas em alguma etapa do sequenciamento, o que prejudica o processo de extração de informações biologicamente úteis. A busca pela atapa faltosa é um processo difícil, os responsáveis pela preparação das amostras submetidas ao sequenciamento (chamada no jargão da biologia de "bancada molhada") indicam falhas na análise dos dados pela equipe de bioinformática, e vice-versa. Para dirimir este problema pesquisou-se na literatura sobre problemáticas similares e suas respectivas soluções e também por uma metodologia para se investigar os motivos das discrepâncias observadas na quantidade de *barcodes* detectados.

No entanto, não se encontrou trabalhos que permitissem a avaliação de corridas *multiplex*, o que motivou o desenvolvimento de um modelo probabilístico capaz de atribuir um grau de confiança aos dados provenientes da plataforma SOLiD, além de se realizar buscas por medidas-resumo capazes de caracterizar sistema de marcação utilizado em sequenciamentos *multiplex*.

Contribuições

Os resultados obtidos corroboraram a suficiência do modelo desenvolvido, o qual permite, dentre outros pontos:

- avaliar de forma criteriosa os protocolos de sequenciamentos utilizados, possibilitando a identificação de faltas em algum passo do processo;
- adaptar e desenvolver novos protocolos para preparação de amostras;
- atribuir um Grau de Confiança aos dados gerados e guiar um processo de filtragem que respeite as características de cada sequenciamento, não descartando sequências úteis de forma arbitrária.

Vale salientar que as funcionalidades ganhas com o presente estudo não se aplicam somente ao sistema de marcação, o modelo proposto é extensível para análise de sequências de interesse. No estudo de caso, o modelo possibilitou a identificação de falhas na preparação das amostras - seja no momento da marcação ou na preparação do *pool* que foi submetido ao sequênciador, dado a alta ocorrência dos *barcodes* 6, 7 e 9, em detrimento aos demais.

Quanto a baixa qualidade da segunda corrida, corroborou-se que o causador foi a queda na qualidade no abastecimento, e posterior interrupção do fornecimento de energia elétrica, dado que o mau desempenho não foi mantido na retomada do sequenciamento. Já as falhas ocorridas na Corrida C3 foram advinda do tempo de permanência excessivo no sequênciador, ocasionando degeneração das sequências e possível desprendimento das *beads* devido ao decaimento

magnético natural. O modelo desenvolvido e os resultados encontrados foram publicados respectivamente em (LOBATO et al., 2011b) e (LOBATO et al., 2011a).

Dificuldades Encontradas

Para desenvolver soluções computacionais para a biologia é necessário entendimento no domínio da aplicação, principalmente no tocante aos conceitos biológicos. Sendo assim, a maior parte das dificuldades encontradas residem na falta de base nesta área, como segue:

- compreensão do processo de sequenciamento genético no SOLiD, fundamental para o processo de avaliação e interpretação das informações extraídas dos dados;
- compreensão do esquema de codificação *colorspace*, desde a sintaxe dos arquivos até estratégias para diminuir o consumo de memória ao armazenar os dados para processamento;
- dificuldade em se obter informações da Applied Biosystems, fabricante da plataforma;
- grande volume de dados a serem analisados, o que requer rotinas computacionais otimizadas para se diminuir o tempo de espera até a obtenção dos resultados.

Trabalhos futuros

Esta dissertação apresentou uma abordagem para a caracterização de dados de corrida *multiplex*. Para tal, produziu-se diversas ferramentas com um escopo bem definido, uma classe para detectar a ocorrência de *barcodes* utilizados no experimento, outra para calcular a média, *etc*. Para melhorar o estudo, indica-se os seguintes pontos:

- integrar as diversas ferramentas seguindo o fluxo de atividades proposto;
- realizar testes utilizando outros conjuntos de dados, se possível englobando os 16 *barcodes* da biblioteca padrão;
- preparar um arcabouço de *software* com arquitetura escalável a fim de facilitar a integração de atualizações da plataforma;
- desenvolver um método estatístico, baseando-se na frequência de determinadas transições da Cadeia de Markov para se recuperar sequências de marcação que não obtiveram *match* perfeito com os *barcodes* da biblioteca padrão;

- desenvolver um modelo capaz de descrever o decaimento magnético que ocorre nas *beads* metálicas na escala temporal.

Referências Bibliográficas

ARGOUT, X. et al. The genome of theobroma cacao. *Nature Genetics*, n. 43, p. 101–108, 2011.

ARTES, F.; ESCRICHE, A. . J. Intermitent warming reduces chilling injury and decay of tomato fruit. *Journal of Food Science*, v. 59, n. 5, 1994.

AUTEXIER, C.; LUE, N. F. The structure and function of telomerase reverse transcriptase. *Annual Review of Biochemistry*, n. 75, p. 493, 2006.

BARRY, B. D. et al. Performance of transgenic corn hybrids in missouri for insect control and yield. *Journal of Economic Entomology*, v. 93, n. 3, p. 991–999, 2000.

BIOSYSTEM, A. Solid system barcoding. *Application Note*, 2008.

BIOSYSTEM, A. Solid software development community. *http://solidsoftwaretools.com/gf/*, 2011. Capturado em 17 de maio de 2011.

BIRD, A. Perceptions of epigenetics. *Nature*, n. 447, p. 396–398, 2007.

BLUMBERG, B. Medical research for the next millenium. *The Cambridge Review*, n. 117, p. 3–8, 1996.

BOLCH, G. et al. *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications,*. 2. ed. [S.l.]: John Wiley & Sons, Ltd, 2006.

BRADLEY, D. et al. Mitochondrial diversity and the origins of african and european cattle. *National Academy of Science of USA*, v. 93, p. 5131–5135, 1996.

BRANDEN, C. I.; TOOZE, J. *Introduction to Protein Structure*. 2. ed. [S.l.]: Garland, 1999.

BREU, H. *A theoretical Understanding of 2 Base Color Codes and Its Application to Annotation, Error Detection, and Error Correction*. [S.l.], 2010.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística básica*. [S.l.]: Saraiva, 2002.

CARROL, J.; LONG, D. *Theory of Finite Automata with an Introduction to Formal Languages.* [S.l.]: Prentice Hall, 1989.

CARTHEW, R. W.; SONTHEIMER, E. J. Origins and mechanisms of mirnas and sirnas. *Cell*, 2009.

CHARTRAND, G. *Introductory graph theory*. [S.l.]: Dover Publications, Inc., 1985. Originally published: Graph as mathematical models., c1977.

CHEN, Y.; SOUAIAIA, T.; CHEN, T. Perm: efficient mapping of short sequencing reads with periodoc full sensitive spaced seeds. *Bioinformatics*, v. 25, n. 29, p. 2514–2521, 2009.

CHING, W.; MICHAEL, K. *Markov Chains: Models, Algorithms and Applications*. [S.l.]: Springer, 2010.

CLARKE, A. B.; DISNEY, R. L. *Porbability and Random Processes, An Introduction for Applied Scientists and Engineers*. [S.l.]: McGraw-Hill, 1970.

CORMEN, T. H. et al. *Algoritmos*. [S.l.]: Editora Campus, 2002.

CRICK, F. H. C. Central dogma of molecular biology. *Nature*, p. 561–563, 1970.

DATE, C. J. *Introdução a Sistemas de Bancos de Dados*. 8. ed. [S.l.]: Campus, 2004.

ESTADÃO. Portal estadão. *http://blogs.estadao.com.br/link/?s=rede%20paraense*, 2011. Capturado em 20 de abril de 2011.

EWING, B.; GREEN, P. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Res*, v. 8, n. 3, p. 186–94, mar. 1998. ISSN 1088-9051. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/9521922>.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, ACM, New York, NY, USA, v. 39, n. 11, p. 27–34, 1996. ISSN 0001-0782.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Knowledge discovery and data mining: Towards a unifying framework. *Proceedings of the Second International Conference on KDD and Data Mining*, 1996.

FISCHHOFF, D. A. Insect tolerant transgenic tomato plants. *Bio Technology*, v. 5, p. 807–813, 1987.

FOURMENT, M.; GILLINGS, M. R. A comparison of common programming languages used in bioinformatics. *BMC Bioinformatics BioMed Central Open Access Methodology article A comparison of common programming languages*, 2008.

GANTZ, J. F.; REINSEL, D. The digital universe decade – are you ready? *External Publication of IDC (Analyse the Future) Information and Data*, p. 1–16, 2010.

GARDNER, S. R. Building the data warehouse. *Communications of ACM*, v. 9, n. 41, p. 52–60, 1998.

GONCALVES, A. N. et al. Otimização da detecção dos barcodes em corrida multiplex da plataforma solid. *57º Congresso Brasileiro de Genética*, 2010.

GOSLING, J.; MCGILTON, H. *The Java Language Environment*. [S.l.], 1996.

GRIFFITHS, A. J. F. et al. *Introduction to Genetic Analysis*. 8. ed. [S.l.]: Freeman, 2005.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. 2. ed. [S.l.]: Elsevier, 2006.

IBM. Ibm archives: Ibm 350 disk storage unit. *http://www-03.ibm.com/ibm/history/exhibits/storage/storage350.html*, 2011. Capturado em 12 de maio de 2011.

KAM, T. *Synthesis of Finite State Machines: Functional Optimization*. [S.l.]: Kluwer Academic Publishers, 1997.

KEEDWELL, E.; NARAYANAN, A. *Intelligent Bioinformatics: the application of artificial intelligence techniques to bioinformatics problems*. [S.l.]: John Wiley & Sons, Ltd, 2005.

KENNEDY, J. B.; NEVILLE, A. M. *Basic statistical Methods for Engineers and Scientists*. [S.l.]: Harper & Row, 1986.

KOVACS, Z. L. *Teoria da Probabilidade e Processos Estocásticos*. [S.l.]: Acadêmica, 1996.

LEON-GARCIA, A. *Probability, statistics, and random processes for electrical engineering*. 3. ed. [S.l.]: Pearson Prentice Hall, 2008.

LESK, A. M. *Introdução à Bioinformática*. Segunda edição. [S.l.]: Artmed, 2008.

LI, H.; HOMER, N. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, v. 11, n. 5, p. 473–483, 2010.

LI, R. et al. Soap: short oligonucleotide aligment program. *Bioinformatics*, v. 24, 2008.

LIN, H. et al. Zoom! zillions of oligos mapped. *Bioinformatics*, 2008.

LINTON, M. et al. Initial sequencing and analysis of the human genome. *Nature*, p. 860–921, 2001.

LOBATO, F. M. F. et al. Abordagem probabilística para análise de confiabilidade de dados gerados em sequenciamentos multiplex na plataforma abi solid. *Anais do XLIII Simpósio Brasileiro de Pesquisa Operacional.*, 2011.

LOBATO, F. M. F. et al. A probabilistic approach for characterizing the marking system of multiplex sequencing in abi solid platform. *International Conference on Bioinformatics and Computational Biology*, 2011.

LOFTUS, R. et al. Evidence for two independent domestications of cattle. *National Academy of Science*, v. 91, p. 2757–2761, 1994.

MARZZOCO, A.; TORRES, B. B. *Bioquímica Básica*. 3. ed. [S.l.]: Guanabara Koogan, 2007.

MENDEL, J. Versuche über pflanzenhybridenverhandlungen des naturforschenden vereines in brünn. *Journal of the Royal Horticultural Society*, 1866.

OUSTERHOUT, J.; DOUGLIS, F. Beating the i/o bottleneck: a case for log-structured file systems. *ACM SIGOPS Operating Systems Review*, v. 23, 1989.

PANDEY, V.; NUTTER, R. C.; PREDIGER, E. Next-generation genome sequencing: Twoards personalized medicine. In: ____. [S.l.]: Wiley-VCH verlag GmbH & Co. KGaA, 2008. cap. Applied Biosystems SOLiD System: Ligation-Based Sequencing.

PAPOULIS, A. *Probability, Random Variables, and Stochastics Processes*. [S.l.]: McGraw-Hill, 1965.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. *Estatística Bayesiana*. [S.l.]: Fundação Caloust Gulbenkian, 2003.

PAVANELLO, S. et al. Shortened telomeres in individuals with abuse in alcohol consumption. *International Journal of Cancer*, 2011.

PEEBES, P. Z. *Probability, random variables, and random signal principles*. 4. ed. [S.l.]: McGraw-Hill, 2001.

PETTERSON, E. et al. Generations of sequencing technologies. *Genomics*, n. 93, p. 105–111, 2009.

PUTERMAN, L. M. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. [S.l.]: John Wiley & Sons, Ltd, 1994.

REIK, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, n. 447, p. 425–432, 2007.

REZENDE, S. O. et al. Mineração de dados. In: REZENDE, S. O. (Ed.). *Sistemas Inteligentes - Fundamentos e Aplicações*. 1. ed. [S.l.]: Manole, 2003. cap. 12, p. 307–335.

ROSS, S. M. *Introduction to probability models*. [S.l.]: Academic Press, 1972.

SALMELA, L. Correction of sequencing erros in a mixed set of reads. *Bioinformatics*, v. 26, n. 10, p. 1284–1290, 2010.

SANGER, F. et al. Nucleotide sequence of bacteriophage ϕ x174 dna. *Nature*, 1977.

SASSON, A.; MICHAEL, T. P. Filtering error from solid. *Bioinformatics*, v. 26, n. 6, p. 849–850, 2010.

SNUSTAD, D. P.; SIMMONS, M. J. *Fundamentos de genética*. Quarta edição. [S.l.]: Guanabara Koogan, 2008.

SOUTHAN, C. Has the yo-yo stopped? an assessment of human protein-coding gene number. *Proteomics*, v. 4, p. 1712–1726, 2004.

STALLINGS, W. *Arquitetura e Organização de Computadores*. 5. ed. [S.l.]: Prentice Hall, 2002.

STEIN, L. How perl saved the human genome project. Cambridge, England, in the conference room of the largest DNA sequencing center in Europe. 1996.

TANENBAUM, A. S. *Organização Estruturada de Computadores*. 5. ed. [S.l.]: Prentice Hall, 2006.

THOMPSON, M. W.; MCINNES, R.; WILLARD, H. F. *Genética Médica*. 7. ed. [S.l.]: Guanabara Koogan, 2008.

TIJMS, H. C. *Stochastic models: an algorithmic approach*. [S.l.]: John Wiley & Sons, Ltd, 1994.

VENTER, J. C. et al. The sequence of the human genome. *Science*, p. 1304–1351, 2001.

VIEIRA, S. *Introdução à bioestatística*. 4. ed. [S.l.]: Elsevier, 2008.

WANG, Z.; GERSTEIN, M.; SNYDER, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature*, Rev. Genetics, p. 57–63, 2009.

WATSON, J.; CRICK, F. H. C. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 1953.

WITTEN, I. H.; FRANK, E. *Data Mining: pratical machine learning tools and techniques*. 2. ed. [S.l.]: Elsevier, 2005.

WU, X. et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, v. 14, p. 1–37, 2008.

Anexo 1 - Códigos-Fonte

Código-Fonte do Header bioFastIo.h

```
1  #ifndef ____XXX___BIOFASTIO___H___XXX
2  #define ____XXX___BIOFASTIO___H___XXX
3
4  #include <cstdio>
5  #include <cstdlib>
6  #include <cstring>
7  #include <string>
8  namespace bioAppUFPA{
9  namespace bioFastIo{
10 class FastInput{
11
12 public:
13         FastInput(){F=NULL; buffer=0; bufferSize=1000000; }
14         FastInput(const char* fn, std::size_t bufferSize=1000000);
15         bool open(const char* fn, std::size_t bufferSize=1000000);
16         void close();
17         bool status();
18         bool eof();
19         void nextStr(char* s,int& l);
20         void nextLine(char* s,int& l);
21         int nextInt();
22         char nextChar();
23         ~FastInput();
24
25 private:
26         void update();
27         std::size_t bufferSize;
28         std::size_t ip;
29         char* buffer;
30         std::FILE* F;
31         bool isOpen;
32 };
33
34 class FastOutput{
```

```
35 public:
36         FastOutput(){F=NULL; buffer=0; size_num=1000; }
37         FastOutput(const char* fn, std::size_t size_num=1000);
38         bool open(const char* fn, std::size_t size_num=1000);
39         void close();
40         bool status();
41         void printInt(int i);
42         void printChar(char c);
43         void printStr(const char* s, int l);
44         ~FastOutput();
45 private:
46         std::size_t size_num;
47         char* buffer;
48         std::FILE* F;
49         bool isOpen;
50 };
51
52 };
53 void string_to_ArrayOfInt(std::string str,int arr[],int& arr_len);
54 bool checkArray_valueLessThan(char arr[],int arr_len,int value);
55 };
56 #endif
```

Código-Fonte da implementação das rotinas de *fast Input/Output*.

```
1
2 /*
3 ================================================================================
4 Fun  es para Leitura R pida
5 ================================================================================
6  */
7 bioAppUFPA::bioFastIo::FastInput::FastInput(const char* fn,std::size_t
      bufferSize)
8 {
9         this->bufferSize=bufferSize;
10         this->ip=0;
11         this->buffer=new char[this->bufferSize];
12         this->F=std::fopen(fn,"rb");
13         if(this->F!=NULL) update();
14 }
```

```
15 bioAppUFPA::bioFastIo::FastInput::~FastInput()
16 {
17         if(buffer!=NULL) delete buffer;
18         if(F!=NULL&&this->isOpen) std::fclose(F);
19         this->isOpen=false;
20 }
21 bool bioAppUFPA::bioFastIo::FastInput::status()
22 {
23         return F!=NULL&&buffer!=NULL;
24 }
25 bool bioAppUFPA::bioFastIo::FastInput::open(const char* fn, std::size_t
       bufferSize)
26 {
27         this->isOpen=true;
28         this->bufferSize=bufferSize;
29         this->ip=0;
30         if(this->buffer==0) this->buffer=new char[this->bufferSize];
31         this->F=std::fopen(fn,"rb");
32         if(this->F!=NULL) update();
33         return F!=NULL;
34 }
35 void bioAppUFPA::bioFastIo::FastInput::close()
36 {
37         if(F!=NULL&&this->isOpen) std::fclose(F);
38         this->isOpen=false;
39 }
40
41 bool bioAppUFPA::bioFastIo::FastInput::eof()
42 {
43         return bufferSize==0;
44 }
45 /**
46 Atualiza buffer de entrada
47  */
48 void bioAppUFPA::bioFastIo::FastInput::update()
49 {
50         std::size_t n=0;
51         ip=0;
52         if(!feof(F)){
53                 n=std::fread(buffer,sizeof(char),bufferSize,F);
54         }
55         bufferSize=n;
56 }
```

```
57
58 char bioAppUFPA::bioFastIo::FastInput::nextChar()
59 {
60         if(ip==bufferSize) update();
61         if(bufferSize==0) return 0;
62         char c=buffer[ip];
63         ip++;
64         return c;
65 }
66 void bioAppUFPA::bioFastIo::FastInput::nextStr(char* in,int& l)
67 {
68         char c=0;
69         l=0;
70         while(true){
71                 if(ip==bufferSize) update();
72                 if(bufferSize==0) break;
73                 c=buffer[ip];
74                 if(c==EOF) break;
75                 if(c!=' '&&c!='\t'&&c!='\n'&&c!='\r'&&c!='\a'&&c!=0) break;
76                 ip++;
77         }
78         while(true){
79                 if(c==EOF) break;
80                 if(c==' '||c=='\t'||c=='\n'||c=='\r'||c=='\a'||c==0) break;
81                 in[l]=c;
82                 l++;
83                 ip++;
84                 if(ip==bufferSize) update();
85                 if(bufferSize==0) break;
86                 c=buffer[ip];
87         }
88 }
89 void bioAppUFPA::bioFastIo::FastInput::nextLine(char* in,int& l)
90 {
91         char c=0;
92         l=0;
93         while(true){
94                 if(ip==bufferSize) update();
95                 if(bufferSize==0) break;
96                 c=buffer[ip];
97                 if(c==EOF||c==0) break;
98                 if(c!=' '&&c!='\t'&&c!='\n'&&c!='\r'&&c!='\a') break;
99                 ip++;
```

```
100             }
101             while(true){
102                     if(c==EOF) break;
103                     if(c==0||c=='\n') break;
104                     in[l]=c;
105                     l++;
106                     ip++;
107                     if(ip==bufferSize) update();
108                     if(bufferSize==0) break;
109             }
110     }
111     int bioAppUFPA::bioFastIo::FastInput::nextInt()
112     {
113             char c;
114             while(true){
115                     if(ip==bufferSize) update();
116                     if(bufferSize==0) break;
117                     c=buffer[ip];
118                     if( (c>='0'&&c<='9') || c==EOF) break;
119                     ip++;
120             }
121             int num=0;
122             while(true){
123                     if(c<'0' || c>'9' ) break;
124                     num=(int)(c-'0') +10*num;
125                     ip++;
126                     if(ip==bufferSize) update();
127                     if(bufferSize==0) break;
128                     c=buffer[ip];
129             }
130             return num;
131     }
```

Código-Fonte da ferramenta responsável pelo cálculo da média aritmética do Grau de Confiança

```
1   package Estatistica;
2
3
4   import java.io.BufferedReader;
5   import java.io.BufferedWriter;
6   import java.io.File;
```

```
7  import java.io.FileReader;
8  import java.util.Iterator;
9  import java.util.Map;
10 import java.util.Scanner;
11 import java.util.Set;
12 import java.util.StringTokenizer;
13 import java.util.TreeMap;
14 import java.util.TreeSet;
15
16 import arquivos.Arquivo2009;
17 import arquivos.Arquivo2010_15032010;
18 import arquivos.Arquivo2010_17032010;
19
20 /**
21  *
22  * @author fabio
23  */
24 public class CalculaMedia {
25
26 public static void main(String args[]) {
27
28 //declara  o das vari veis
29 Scanner input = new Scanner(System.in);
30 System.out.println("Insira o caminho absoluto do arquivo de qualidade a ser
       analisado: ");
31 String nomeArquivo = input.nextLine();
32 File arquivo = new File(nomeArquivo);
33 String s = null;
34 long contadorGrandeWhile = 01;
35 Float qualidadebarcode = 0.0f;
36 StringTokenizer st;
37 Integer auxiliar = 0;
38 int numeroOcorrencias =0 ;
39 double [] vetorProbabilidades = new double[37];
40 for(int i = -1 ; i<36; i++){
41 vetorProbabilidades[i+1] = 1 - Math.pow(10, (-(double)(i+1) / 10));
42 }
43 try {
44 //abre os arquivos e o leitor
45 BufferedReader buffreader = new BufferedReader(new FileReader(arquivo));
46 System.out.println("Inicio do Processamento");
47 TreeMap<Float, Integer> mapaBloco = new TreeMap<Float, Integer>();
48 bigwhile:
```

```
49 while ((s = buffreader.readLine()) != null) {
50 contadorGrandeWhile++;
51 if (contadorGrandeWhile <= 3) {
52 continue bigwhile;//pula linhas cabe alho
53 }
54 if ((contadorGrandeWhile % 2) != 0) {
55 //System.out.println(s);
56 st = new StringTokenizer(s);
57 qualidadebarcode = 1.0f;
58 while (st.hasMoreTokens()) {
59 qualidadebarcode =(float) ((float) qualidadebarcode*vetorProbabilidades[(
       Integer.parseInt(st.nextToken()))+1]);
60 }//fim do calculo da probabilidade do Barcode
61 if(mapaBloco.get(qualidadebarcode)==null)
62 {
63 numeroOcorrencias++;
64 mapaBloco.put(qualidadebarcode, 1);
65 }
66 else
67 {
68 numeroOcorrencias++;
69 auxiliar = (Integer) mapaBloco.get(qualidadebarcode)+1;
70 mapaBloco.put(qualidadebarcode, auxiliar);
71 }
72 }//fim do if linhaQUal
73 }//fim do while
74 System.out.println("In cio do calculo da M dia.");
75 Set set = mapaBloco.entrySet();
76 Iterator itr = set.iterator();
77 int sum = 0 ;
78 Double somatorio = 0.0;
79 while(itr.hasNext()){
80 Map.Entry me = (Map.Entry) itr.next();
81 somatorio += (Float)me.getKey()*(Integer)me.getValue();
82 sum+=(Integer)me.getValue();
83 }
84 Double media = somatorio/sum;
85 System.out.println(("M dia: "+media));
86 System.out.println("Somat rio: "+sum);
87 System.out.println("Numero de acessos ao if/else: "+numeroOcorrencias);
88 System.out.println("T rmino");
89 }//fim do m todo try
90 catch (Exception e) {
```

```
91 e.getStackTrace();
92 e.getLocalizedMessage();
93 e.printStackTrace();
94 }
95 }
96 }
```

Anexo 2 - Modelo Relatório

Informações que devem aparecer no relatório:

Corridas analisadas: 3;

Barcodes Utilizados no Processo: 10;

Barcode	*colorspace*	*basespace*
1	“0032”	GGGCCT
2	“0111”	GGTGTG
3	“0200”	AAGGGG
4	“0323”	CCGATG
5	“1013”	CAACGA
6	“1130”	GTGCCC
7	“1221”	GTCTGG
8	“1302”	ACGGAG
9	“2020”	GAAGGG
10	“2103”	GACCGC

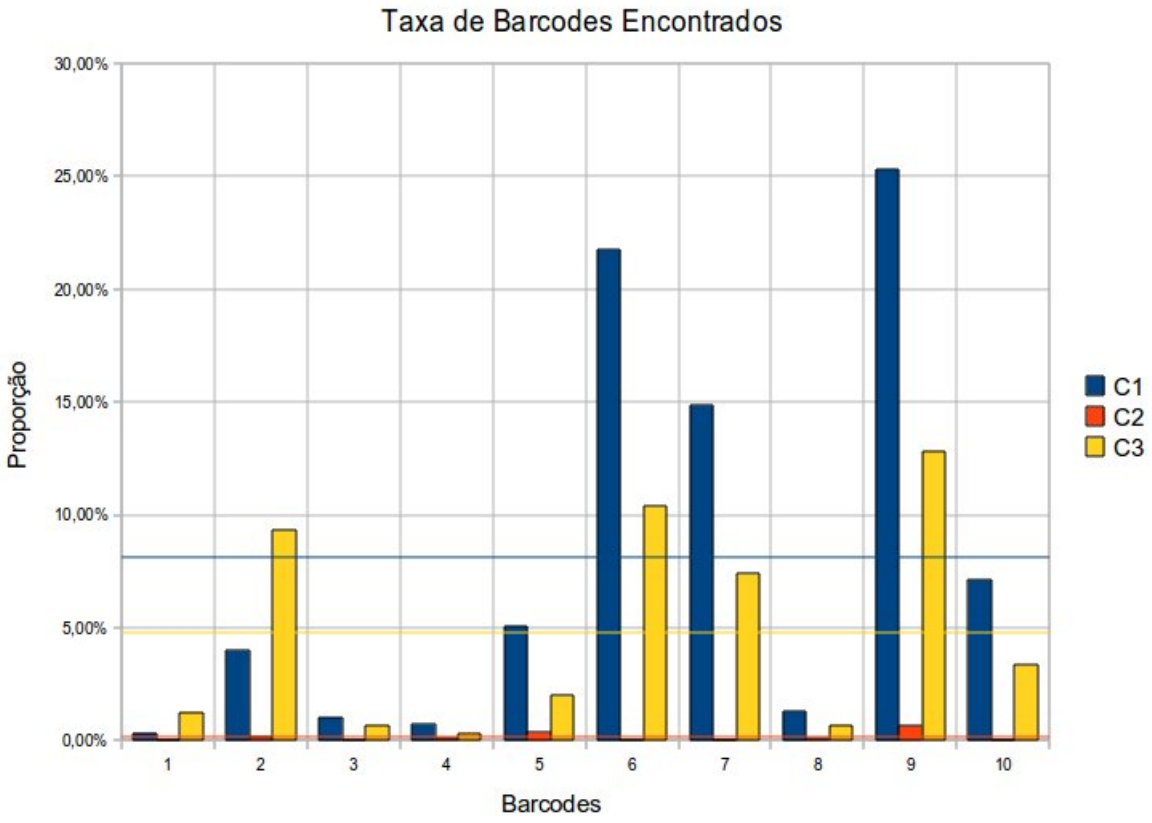
Informações sobre os dados brutos:

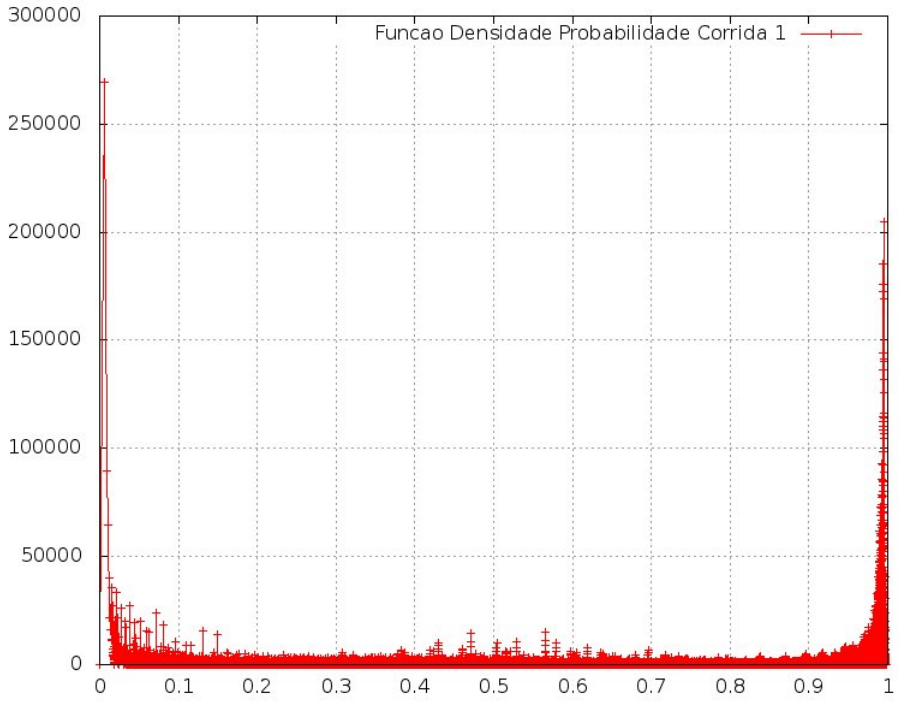
	Corrida 1	Corrida 2	Corrida 3
Data	07/09/2009	15/03/2010	17/03/2010
Número de sequências de marcação	142.453.565	29.602.637	51.832.867
Número de sequências de interesse	158.673.424	19.523.621	27.634.981
Tamanho dos arquivos	32,8Gb	4,4Gb	6,9Gb

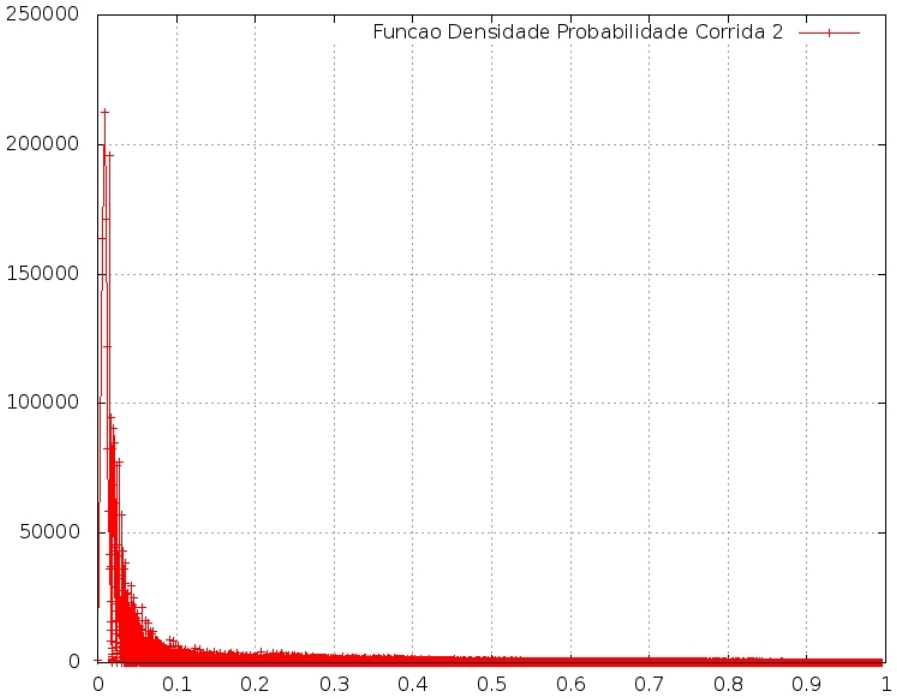
Medidas-Resumo Obtidas:

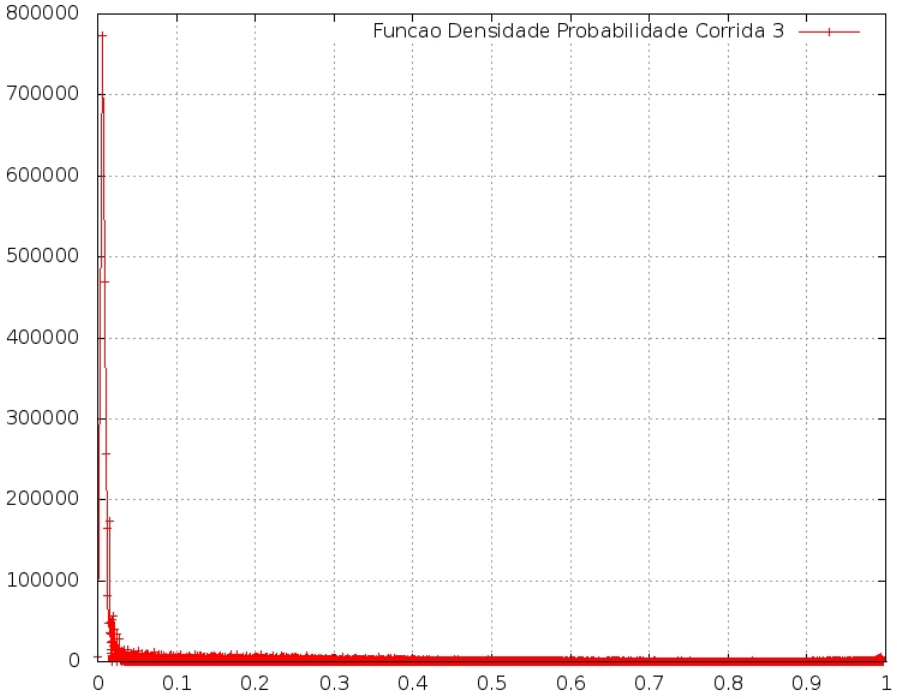
	Corrida 1 (C1)	Corrida 2 (C2)	Corrida 3 (C3)
Taxa de Barcodes Pareados	84%	69%	67%
Taxa de *Match* Perfeito	81,44%	1,63%	48,18%
Média	81,23%	47,15%	59,17%
Moda	30,88%	37,26%	30,88%
Variância	5,53%	6,87%	8,06%

Proporção dos Barcodes Encontrados e Funções Densidade de Probabilidade das Corridas: C1, C2 e C3, respectivamente:






Funcao Densidade Probabilidade Corrida 2
250000
200000
150000
100000
50000
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Funcao Densidade Probabilidade Corrida 3
800000
700000
600000
500000
400000
300000
200000
100000
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1