Uma abordagem computacional para a determinação de polimorfismo de base única

Miguel Galves

Dissertação de Mestrado

Instituto de Computação
Universidade Estadual de Campinas

Uma abordagem computacional para a determinação de polimorfismo de base única

Miguel Galves[1]

01 de Dezembro de 2006

Banca Examinadora:

- Prof. Dr. Zanoni Dias (Orientador)
- Prof. Dr. João Meidanis
  Instituto de Computação – IC – Unicamp
- Prof. Dr. Guilherme Pimentel Telles
  Instituto de Ciências Matemáticas e de Computação – ICMC – USP
- Prof. Dr. Cid Carvalho de Souza (Suplente)
  Instituto de Computação – IC – Unicamp

[1]**Apoio financeiro da Scylla Bioinformática.**

Uma abordagem computacional para a determinação de polimorfismo de base única

Este exemplar corresponde à redação final da Dissertação devidamente corrigida e defendida por Miguel Galves e aprovada pela Banca Examinadora.

Campinas, 01 de Dezembro de 2006.

Prof. Dr. Zanoni Dias (Orientador)

Dissertação apresentada ao Instituto de Computação, UNICAMP, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Agradecimentos

Gostaria de agradecer primeiramente a João Meidanis e Zanoni Dias, sócios da Scylla Bioinformática, por terem oferecido toda a estrutura necessária para que este trabalho fosse executado com êxito. Agradeço ao meu orientador Zanoni Dias, por ter ajudado a organizar a bagunça intrinseca que permeia meu trabalho.

Agradeço também à Flavia, Alexandre, Christian, Zé Augusto e André por ajudarem a tornar o dia-a-dia de trabalho mais produtivo e agradável, fator essencial para que o cotidiano se torne menos cansativo. Muita coisa não teria sido possível sem a colaboração efetiva destas pessoas.

Não posso deixar de citar meus amigos Dani (que inclusive colaborou com alguns desenhos), Ricardo, Vivi, Mário, João, Camila, Raphael, Thais, Talita, Laurent, Patricia, Leonardo, Júlia, Carol, Mari, Fernanda, Hugues e Tereza por não me deixarem esquecer que existe vida fora da pesquisa de mestrado.

Last, but not least, quero agradecer à minha família, e em particular aos meus pais, por todo apoio emocional, financeiro, intelectual, estrutural e outros tantos. Sem isto nada seria possível.

Não necessariamente nesta ordem...

Resumo

A pesquisa genômica é de grande interesse para a área médica. Por isso o entendimento de como os genes influenciam na aparição de doenças é de grande relevância para a criação de métodos de diagnóstico e criação de drogas apropriadas. A maioria dos genes apresenta uma grande freqüência de variações alélicas, conhecida como polimorfismo. Estas variações podem ser a chave para a predisposição de certos indivíduos a certas doenças. Dentre os polimorfismos, os SNPs têm uma grande importância, por representarem cerca de 90% dos polimorfismos encontrados no genoma humano [21]. Neste trabalho iremos estudar três etapas no processo de detecção e análise de SNPs.

A primeira etapa consiste no processo de alinhamento de seqüências de EST e cDNA com DNA genômico. A identificação de genes em seqüências de DNA não-caracterizadas é um dos grandes problemas na pesquisa genômica. Os algoritmos tradicionais [45, 55, 83, 84, 106] descrevem métodos para alinhar duas seqüências arbitrárias. Iremos descrever as estratégias de alinhamento de duas seqüências, discutir os métodos existentes para alinhamento de cDNA com DNA genômico e propor um conjunto de coeficientes apropriados a serem usados com os algoritmos clássicos para resolver este tipo de alinhamento.

A segunda etapa consiste na detecção de SNPs, seja através de alinhamentos múltiplos ou da análise de cromatogramas. Iremos descrever o funcionamento dos dois métodos citados acima e discutir uma nova metodologia para detectar SNPs em seqüências de vírus HIV.

A terceira etapa consiste em correlacionar SNPs. Sabe-se que a predisposição genética para muitas doenças não é devido a apenas uma mutação, ou à presença ou não de um certo alelo. Em muitos casos, vários SNPs agem em conjunto e aumentam ou não a chance de uma doença se manifestar em um indivíduo. Assim, é muito importante desenvolver métodos de correlação entre diversos SNPs para se entender como eles interagem entre si. Iremos descrever medidas de correlação e estudar a presença de LDs e LDs múltiplos em genes da cana-de-açúcar, mapeados pelo projeto SUCEST, e em genes humanos.

Abstract

Genomic research is of great interest in the medical field. Therefore, understanding how genes impact the ocurrence of diseases is of significant relevance, so that proper diagnosis can be made and appropriate drugs can be developed. Most genes present great variantion and allele frequency, known as polymorphism. These variantions may be key to understanding the predisposition in individuals to certain diseases. Among polymorhisms, SNPs are of great importance, representing circa 90% of all polymorphism found in the human genome. [21]. For this work, we will study three phases in the process of detecting and analysis of SNPs.

The first phase consists in the process of aligning EST sequences and cDNA to genomic DNA. Identiying genes in non-caracterized DNA sequences is one of the challenging problems in genomic research. Traditional algorithms [45, 55, 83, 84, 106] describe methods to align two arbitrary sequences. We shall describe alignment strategies of two sequences, discuss over existing existing methods for aligning cDNA with genomic DNA and propose a set of apropriate coefficients to be used in the classical algorithms to perform this kind of alignment.

The seconde phase consists in detecting SNPs, whether through multiple alignments or cromatogram analysis. We shall describe how the two above mentioned methods work and discuss a new methodology to detect SNPs and HIV sequences.

The third phase consists of correlating SNPs. It is known that the genetic predisposition for many diseases is not only due to a single mutation, or to the presence or absence of a single allele. In many cases, several SNPs act together and may increase or decrease the chance for a disease to manifest in a individual. Thus, it is very important to develop methods of correlation between SNPs to better understand how they interact. We shall describe correlation measures and study the presence of LD or multiple LDs in sugarcane genes,which were mapped by the SUCEST project, and in human genes.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Os trabalhos conjuntos da comunidade científica na área de biologia molecular têm produzido uma quantidade enorme de dados brutos. O objetivo da bioinformática é produzir ferramentas para aquisição, armazenamento, organização, análise e visualização desses dados, visando produzir resultados relevantes, com aplicações práticas.

1.1 Motivação e Objetivos

A pesquisa genômica é de grande interesse para a área médica, tendo em vista que muitas doenças graves em seres humanos possuem uma causa genética. Por isso o entendimento de como os genes influenciam na aparição de doenças é de grande relevância para a criação de métodos de diagnóstico e criação de drogas apropriadas.

A publicação do mapeamento do genoma humano gerou um grande entusiasmo na comunidade científica e médica [75]. Este entusiasmo baseia-se no fato que a maioria dos genes apresentam uma grande freqüência de variações alélicas, conhecidas como polimorfismos, e que estas variações podem ser a chave para a predisposição de alguns indivíduos a certas doenças. Dentre os polimorfismos, os SNPs tem tido um espaço de destaque, por representarem cerca de 90% dos polimorfismos encontrados no genoma humano [21].

O objetivo deste trabalho é estudar três etapas envolvidas no processo de detecção e análise de SNPs, e propor novas estratégias para aprimoramento de algoritmos relacionados a essas etapas. Cada fase será estudada de forma separada, com problemas específicos a serem resolvidos e conjuntos de dados distintos. Ao final do trabalho, teremos um conjunto de métodos para cada etapa, permitindo a criação de um fluxo de análise.

1.2 Organização do texto

A seguir, descreveremos brevemente a estrutura e os assuntos abordados em cada capítulo deste trabalho.

1.2.1 Conceitos

No Capítulo 2 iremos definir alguns conceitos básicos de bioquímica importantes para a compreensão do resto do trabalho. Além disso, mostraremos o contexto no qual este trabalho se insere e a importância do assunto na pesquisa genética atual.

1.2.2 Alinhamento de DNA com cDNA

A identificação de genes em seqüências de DNA não-caracterizadas (seqüências sobre as quais não se sabe quais são as regiões codificantes e quais são as regiões não codificantes) é um dos grandes problemas na pesquisa genômica.

Um dos métodos que tem sido mais utilizado para esta tarefa é alinhar pedaços de seqüência com seqüências genômicas. O grande número de ESTs seqüenciados tem sido um fator importante na adoção desta estratégia [80].

Os algoritmos tradicionais [45, 55, 83, 84, 106] descrevem métodos para alinhar duas sequências arbitrárias, utilizando coeficientes de penalização para inserção de buracos, e pesos para alinhamentos de bases. Porém, nenhuma destas estratégias leva em conta a natureza das sequências a serem alinhadas. Em particular, não levam em conta a existência de grandes blocos contíguos de regiões codificadoras e de regiões não codificadoras nos genes. Várias estratégias [19, 32, 60, 80, 122] foram apresentadas visando resolver este problema.

No Capítulo 3 iremos descrever as estratégias de alinhamento de duas sequências, discutir os métodos existentes para alinhamento de cDNA com DNA genômico e propor um conjunto de coeficientes apropriados a serem usados com os algoritmos clássicos para resolver este tipo de alinhamento.

Os resultados apresentados neste capítulo foram descritos no artigo "*Comparison of genomic DNA to cDNA alignment methods*" por Galves e Dias [35], publicado no *Lecture Notes on Bioinformatics* e apresentado no Brazilian Symposium on Bioinformatics – BSB 2005, realizado entre 25 e 29 de julho de 2005 em Porto Alegre – RS.

1.2.3 Detecção de SNPs

Existem basicamente duas técnicas computacionais para se detectar a existência de um SNP em um cadeia de DNA obtida por meio de seqüenciamento. A primeira técnica

baseia-se no cromatograma obtido ao final do processo de seqüenciamento. Analisando o sinal extraído da eletroforese, é possível se detectar a presença de um SNP na molécula seqüenciada [89].

A outra técnica utilizada baseia-se na análise de seções transversais de alinhamentos múltiplos de sequências de DNA ou ESTs. Utilizando métodos estatísticos [77] e dados de qualidade obtidos por programas de determinação de bases [30, 31], é possível se determinar a probabilidade de uma base ser um SNP, ou apenas uma mutação pontual do genoma.

No Capítulo 4 iremos descrever o funcionamento dos dois métodos citados acima e discutir uma nova metodologia para detectar SNPs em sequências de vírus HIV. Iremos também descrever alguns estudos para definição de um método que forneça parâmetros para a avaliação da confiabilidade estatística na identificação da SNPs.

Os resultados apresentados neste capítulo sobre detecção de SNPs foram descritos no artigo "*New strategy to detect single nucleotide polymorphisms*" por Galves, Quitzau e Dias [36], publicado na revista *Genetic Molecular Research* 2006, vol. 5, e apresentado no congresso X-Meetings 2006, realizado entre os dias 24 e 28 de julho de 2006 na cidade de Caxambu – MG. O relatório técnico "Comparação de métodos para determinação de SNPs com medidas de confiabilidade" (IC-06-15) por Baudet, Galves e Dias [9], sobre o estudo de confiabilidade estatística, foi depositado no Instituto de Computação da UNICAMP.

1.2.4 Correlação de SNPs

Sabe-se que a predisposição genética para muitas doenças não se deve apenas a uma mutação, ou à presença ou não de um certo alelo. Em muitos casos, vários SNPs agem em conjunto e aumentam ou não a chance de uma doença se manifestar em um indivíduo. Assim, é muito importante se desenvolver métodos de correlação entre diversos SNPs para se entender como eles interagem entre si.

Para se definir correlação entre SNPs estuda-se o princípio de Linkage Disequilibrium (LD) que pode ser definido como uma associação não-aleatória de SNPs [6, 101]. Várias medidas para quantificar LDs foram criadas, como D, D' e r^2 [27, 70].

No Capítulo 5 iremos descrever estas medidas e estudar a presença de LDs em genes da cana-de-açúcar, mapeados pelo projeto SUCEST, e genes humanos extraídos do cromossomo 6. Além disso, iremos definir heurísticas para detecção de LDs múltiplos nos genes mencionados.

O relatório técnico "Um algoritmo para identificação de correlações múltiplas de polimorfismos" (IC-06-14) por Almeida, Galves e Dias [3], descrevendo os resultados obtidos neste capítulo, foi depositado no Instituto de Computação da UNICAMP.

1.2.5 Conclusão

No Capítulo 6, serão resumidos os principais resultados obtidos ao longo deste trabalho, e possíveis extensões para futuros trabalhos sobre o mesmo tema.

Capítulo 2

Conceitos básicos

Neste Capítulo iremos definir alguns conceitos básicos para o entendimento do resto do trabalho. A Seção 2.1 resume o início das pesquisas genômicas. A Seção 2.2 define alguns conceitos básicos de biologia molecular e genética necessários para o entendimento do tema. A Seção 2.3 descreve os processos envolvidos em um projeto de seqüenciamento genético. A Seção 2.5 define o conceito de SNP, ou polimorfismos de base única, e seu interesse para a pesquisa genética e médica. A Seção 2.6 descreve brevemente duas categorias de métodos para análise e deteção de SNPs.

2.1 O início da pesquisa genômica

Em meados do século XIX, experimentos permitiram que se começasse a entender os mecanismos básicos da genética e da hereditariedade. A teoria vigente na época, e aceita por grande parte dos cientistas, era de que as características herdadas por um indivíduo eram o resultado da mistura das características dos pais. **Charles Darwin** acreditava que cada indivíduo possuía partículas no corpo afetadas por ações efetuadas durante a vida, e que estas partículas eram transmitidas através do sangue às células reprodutivas. Esta teoria, chamada de pangênese, era uma variante da teoria de **Jean Baptiste Lamarck**, publicada em 1809 no livro *Philosophie Zoologique*, da hereditariedade de características adquiridas.

A primeira hipótese correta foi formulada por **Gregor Mendel**, em seu trabalho *Experiments with Plant Hybrids* [79] publicado em 1866 na revista da Brunn Natural History Society, onde descrevia seu experimento com ervilhas. Esta planta foi escolhida porque possibilitava a Mendel observar duas gerações de plantas ao longo de um ano, permitindo obter resultados de forma relativamente rápida.

O experimento consistiu em fazer o cruzamento de várias plantas, e observar a evolução de sete características distintas de uma geração para a outra. Estas características foram

escolhidas por serem facilmente observáveis e por em geral terem apenas duas formas distintas: cor e posição da flor, tamanho do caule, forma e cor da semente, e forma e cor da vagem.

Mendel observou que em plantas de ervilhas algumas características não eram resultantes da mistura de características da geração anterior, invalidando a teoria aceita na época. Três conclusões importantes foram tiradas dos experimentos:

1. A hereditariedade de uma certa característica fenotípica é determinada por "unidades" ou "fatores" (hoje conhecidos como genes) que são passado aos descendentes sem nenhuma modificação.

2. Cada descendente herda uma "unidade" de cada característica da geração anterior.

3. Uma dada característica pode não aparecer em um indivíduo, mas pode ser passada para as próximas gerações.

Com estas conclusões, foram estabelecidos dois princípios do processo de hereditariedade:

Princípio da segregação Apenas um alelo de cada gene por indivíduo é transmitido às próximas gerações.

Princípio do arranjo independente os alelos são transmitidos aos filhos de forma aleatória, de forma que combinações não existentes nos pais podem aparecer nos filhos.

Estes resultados, que foram aceitos pela comunidade científica em 1900, estabeleciam a base para o entendimento do mecanismo genético em seres vivos. Porém não se sabia ainda como as características de cada ser vivo eram "armazenadas", decodificadas e transmitidas a descendentes.

A descoberta chave para o entendimento completo do mecanismo genético foi feita por James Watson e Francis Cricks, que publicaram seus resultados obtidos em 1953 no artigo *Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid* [119]. No modelo definido por eles, e que será descrito de forma mais detalhada na próxima seção, os genes estavam armazenados em uma estrutura helicoidal chamada de DNA.

Com esta descoberta, que lhes valeu o Prêmio Nobel em 1962, estava aberto o caminho para a pesquisa genômica, que teve seu auge no mapeamento do genoma humano. Na seção a seguir, serão descritos os princípios básicos do mecanismo genético dos seres vivos.

2.2 Princípios bioquímicos de genética

As informações necessárias para o desenvolvimento dos seres vivos estão codificadas em cadeias de **nucleotídeos**. O conjunto completo de seqüências é chamado de **genoma**. Esta seção irá descrever brevemente seu funcionamento.

2.2.1 DNA e RNA

O código genético pode ser armazenado tanto na forma de **DNA** (ácido desoxiribonucléico) quanto na forma de **RNA** (ácido ribonucléico), exemplificados na Figura 2.1. A estrutura básica dessas duas moléculas é o nucleotídeo.

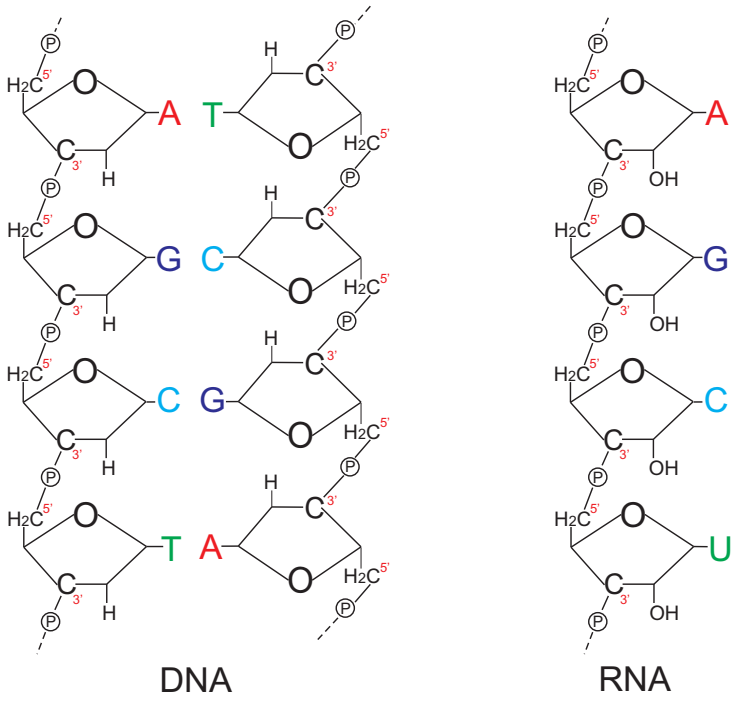

Figura 2.1: Fragmento de DNA e de RNA. Pode se observar que o DNA possui duas fitas com sentidos $5' \rightarrow 3'$ opostos, e que no RNA a base T é substituída pela base U.

Um nucleotídeo é um composto químico formado por uma base orgânica, uma pentose (molécula de açúcar com 5 carbonos) e um grupo fostato. Cada nucleotídeo é caracterizado pela sua base orgânica, que pode ser adenina (A), citosina (C), guanina (G), timina (T) ou uracila (U).

A ligação entre os nucleotídeos se faz por um grupo químico chamado hidroxil que liga o terceiro carbono da pentose de um nucleotídeo ao fosfato do nucleotídeo seguinte. Desta

ordenação, cria-se uma fita com uma extremidade chamada de 3', onde fica um grupo hidroxil livre, e uma extremidade 5', onde fica um fosfato livre. A direção convencionada para leitura das bases é $5' \rightarrow 3'$.

O DNA é formado por duas fitas de nucleotídeos, formadas por uma seqüência de A, C, G e T e onde a pentose dos nucleotídeos é uma desoxiribose. As duas fitas são acopladas pelas bases orgânicas seguindo a seguinte regra: a base A sempre se acopla à base T e a base C sempre se acopla à base G. As duas fitas são ditas complementares, e a orientação $5' \rightarrow 3'$ das duas fitas é oposta uma em relação à outra.

O RNA é formado apenas por uma fita composta por uma seqüência de bases A, C, G e U (com A se acoplando a U e C se acoplando a G) e onde a pentose dos nucleotídeos é uma ribose. Seu tamanho é bem menor do que o tamanho de uma molécula de DNA.

2.2.2 Expressão gênica

Um gene é um fragmento de DNA que contém informação genética codificada. Para que esta informação se torne uma realidade, é necessário que se faça um processo chamado de expressão gênica, na qual uma cópia do gene é transcrita sob a forma de RNA, que é utilizado para a síntese de uma proteína.

Uma proteína é uma macromolécula que possui uma função específica dentro de um organismo: ela é formada por uma cadeia de aminoácidos (existem 20 aminoácidos na natureza), e possui uma forma tridimensional que tem ligação direta com sua função. Os dois principais processos envolvidos na expressão gênica, transcrição e tradução, serão descritos a seguir.

A transcrição

O processo de síntese de um gene em RNA é chamado de transcrição, e pode gerar vários tipos de RNA: o **mRNA** (RNA mensageiro, que contêm as informações para produção de proteínas), o **tRNA** (RNA transportador, utilizado na fase de tradução que será explicada a seguir) e o **rRNA** (RNA ribossomal).

O processo de transcrição de DNA em mRNA, essencial para a síntese de proteínas, é ligeiramente diferente nos organismos **eucariotos** (como os seres humanos) e nos organismos **procariotos** (como as bactérias).

Em ambos os casos, o processo de transcrição é efetuado por uma enzima chamada de RNA polimerase, que inicia a síntese do mRNA em um marcador conhecido como promotor, e processa o DNA na direção $3' \rightarrow 5'$, gerando RNA na direção $5' \rightarrow 3'$. Nos procariotos, o mRNA é gerado diretamente a partir do DNA e não sofre modificações. Nos eucariotos o processo tem duas fases: na primeira fase, é gerado um RNA a partir do DNA, chamado de transcrito primário.

Na segunda fase, o transcrito primário sofre duas modificações. O RNA recebe uma seqüência de bases A chamada de poly-A (que dá estabilidade ao RNA) à extremidade 3' e uma molécula chamada **cap** à extremidade 5'. Além disso, a cadeia de bases é modificada devido ao fato que seus genes contêm dois tipos de regiões: regiões codificadoras (ou seja, que podem ser traduzidas em proteínas) chamadas de **éxons**, e regiões não codificadoras, chamadas de **íntrons**. Os íntrons são removidos em um processo chamado de **splicing**, o qual cumpre uma função de grande importância no processo de diversidade genética, uma vez que nem todos os éxons são preservados no processo. Assim, um mesmo transcrito primário pode gerar vários mRNAs diferentes. Após este processo obtém-se um transcrito maduro.

A tradução

O processo de síntese de uma proteína a partir de um mRNA é chamado de tradução, e se baseia em triplas de nucleotídeos chamados de **códons**. Existem $4^3 = 64$ combinações de códons: cada códon pode ser traduzido em um aminoácido (e um aminoácido pode ser originado por vários códons, uma vez que existem 20 aminoácidos) ou pode ter um significado especial, sendo utilizado como marcador de início e fim de uma região de tradução (respectivamente, **start-códon** e **stop-códon**).

Cada organismo pode interpretar os códons de forma diferente, gerando seqüências de aminoácidos distintas para uma mesma cadeia de nucleotídeos. Assim, existem várias tabelas de conversão de **códons** em aminoácidos, dependendo do organismo considerado. A Tabela 2.1 mostra o *Standard Genetic Code* [85], utilizado por muitos seres vivos, inclusive os seres humanos.

Além do mRNA, existem 2 elementos envolvidos no processo de tradução: o ribossomo e o tRNA. O **ribossomo** é uma organela responsável por percorrer o mRNA, iniciando no start-códon e passando por todos os códons da fita. Para cada códon, é ativado um tRNA, que funciona como um adaptador entre o códon e o aminoácido correspondente: de um lado, ele transporta um aminoácido, e do outro um **anti-códon**. O anti-códon (códon formado pelos nucleotídeos complementares) se acopla ao códon do mRNA, e o aminoácido é liberado e se junta à cadeia de aminoácidos. A Figura 2.2 exemplifica o mecanismo da tradução.

2.3 Seqüenciamento genômico

O seqüenciamento de um gene consiste em se determinar a cadeia de nucleotídeos que o compõe. Este processo envolve várias etapas e procedimentos experimentais, que serão brevemente descritos a seguir.

Standard Genetic Code

Aminoácido	**Cód. 3**	**Cód. 1**	**códon**
Alanina	ALA	A	GCA,GCC,GCG,GCU
Arginina	ARG	R	AGA,AGG,CGA,CGC,CGG,CGT
Asparagina	ASN	N	AAC,AAT
Ácido Aspartico	ASP	D	GAC,GAT
Cysteína	CYS	C	TGC,TGT
Ácido Glutâmico	GLU	E	GAA,GAC
Glutamina	GLN	Q	CAA,CAG
Glycina	GLY	G	GGA,GGC,GGG,GGT,CAG
Histidina	HIS	H	CAC,CAT
Isoleucina	ILE	I	ATA,ATC,ATT
Leucina	LEU	L	CTA,CTC,CTG,CTT,TTA,TTG
Lysina	LYS	K	AAA,AAG
Metyonina	MET	M	ATG
Phenilalanina	PHE	F	TTC,TTT
Prolina	PRO	P	CCT,CCC,CCA,CCG
Serina	SER	S	AGT,TCA,TCC,TCT,TCG
Threonina	THR	T	ACA,ACC,ACG,ACT
Tryptophan	TRP	W	TGG
Tyrosina	TYR	Y	TAC,TAT
Valina	VAL	V	GTA,GTC,GTG,GTT
Stop-Codon	.	*	TAA,TAG,TGA

Tabela 2.1: Standard Genetic Code: Tabela Padrão de Conversão de Aminoácidos utilizada por muitos seres vivos. Cada tripla de bases, ou códon, pode ser traduzida em um aminoácido ou interpretada como um códon especial. Nesta tabela, o códon ATG é interpretado como start-códon. O primeiro código representa um aminoácido com 3 letras, enquanto que o segundo o representa com apenas uma letra.

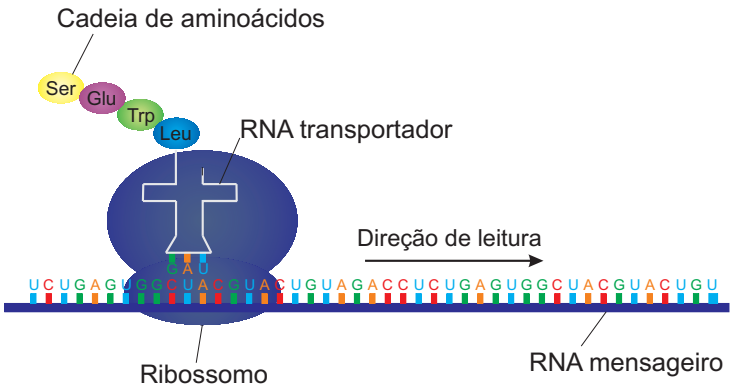

Figura 2.2: Mecanismo de tradução de mRNA em cadeia de aminoácido. O ribossomo "lê" o mRNA, acoplando a cada códon um tRNA contendo o anti-códon correspondente. Uma vez acoplado, o tRNA libera o aminoácido correspondente, que se junta à cadeia de aminoácidos existente.

2.3.1 Preparação do material genético

Apesar do desenvolvimento da tecnologia utilizada neste processo, ainda existe uma limitação física que impede que sejam sequenciadas cadeias de nucleotídeos maiores do que 1000 bases. Devido a esta limitação, o primeiro passo de um seqüenciamento é a **fragmentação**, que consiste em quebrar a cadeia de nucleotídeos a ser determinada em pequenos fragmentos seqüenciáveis. A fragmentação pode ser feita de duas formas: por **digestão** ou por **shotgun**.

No método de **digestão**, são utilizadas enzimas especiais chamadas de **enzimas de restrição**, que cortam o DNA em regiões conhecidas como **sítios de restrição**. No método **shotgun**, o DNA é submetido a altas taxas de vibrações, fazendo com que a cadeia de nucleotídeos se quebre em diversos pontos.

Os fragmentos são então replicados em um processo chamado de **amplificação**, que pode ser feita tanto inserindo os fragmentos em bactérias (processo chamado de clonagem), quanto utilizando enzimas que sintetizam novas fitas de DNA (processo chamado de PCR).

2.3.2 Método da terminação de cadeia

Uma vez obtidos os fragmentos de tamanho adequado, pode-se iniciar o processo de seqüenciamento das bases de cada fragmento. Para isso, o método mais utilizado é o da terminação de cadeia [103].

Os fragmentos de fita simples de DNA obtidos anteriormente são utilizados como moldes para uma reação de síntese de novas fitas símples de DNA complementares aos moldes (chamadas de DNA complementar). A Figura 2.3 mostra o funcionamento deste

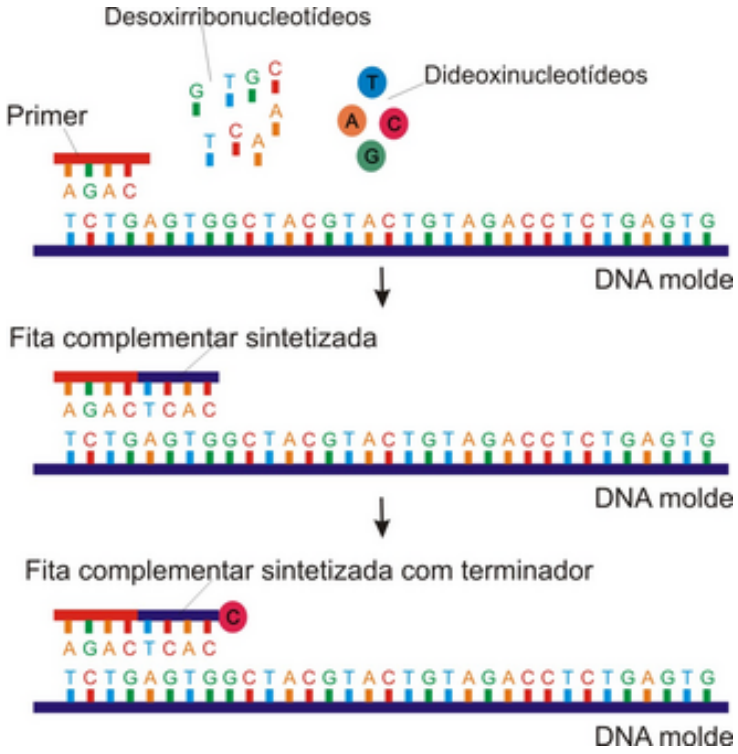

Figura 2.3: Método da terminação da cadeia. Acopla-se um primer a uma fita simples de DNA molde, e se efetua a síntese de DNA complementar em um meio contendo nucleotídeos normais e nucleotídeos terminadores. Assim que um terminador é acoplado à fita de DNA, o processo termina.

processo.

Para que esta reação ocorra, acopla-se uma cadeia de nucleotídeos chamada de **primer**, que servirá de iniciador para o processo de síntese da nova fita, e definirá qual a região do DNA será sintetizada.

O processo de síntese da fita complementar ao molde é catalizado por uma enzima de DNA polimerase, e ocorre em um meio contendo os quatro tipos de desoxiribonucleotídeos dATP, dCTP, dGTP e dTTP (contendo respectivamente as bases A, C, G e T) e alguns dideoxinucleotídeos ddATP, ddCTP, ddGTP e ddTTP (também contendo respectivamente as bases A, C, G e T).

Um dideoxinucleotídeo pode se acoplar a um desoxiribonucleotídeo, mas não possui o grupo 3'-hidroxil, necessário para que um novo nucleotídeo se conecte a ele, fazendo com que o processo de síntese seja interrompido. A enzima polimerase não distingue entre dideoxinucleotídeo e dideoxinucleotídeos, fazendo com que algumas seqüências fiquem mais longas do que outras. Cada tipo de dideoxinucleotídeo é acoplado a um marcador

químico, utilizado na próxima fase do seqüenciamento.

Os fragmentos com terminadores de seqüência assim obtidos são separados por tamanho em um processo chamado de eletroforese: neste processo, os fragmentos são colocados sobre uma placa contendo um gel, sobre o qual se aplica um campo elétrico fazendo com que os fragmentos migrem do pólo negativo onde se encontram inicialmente para o pólo positivo. Os menores fragmentos tendem a migrar mais rapidamente, e ao final do procedimento, estarão mais perto do pólo positivo do que os maiores fragmentos, sendo assim possível determinar seus tamanhos. A Figura 2.4 exemplifica o o arranjo final das seqüências após a eletroforese.

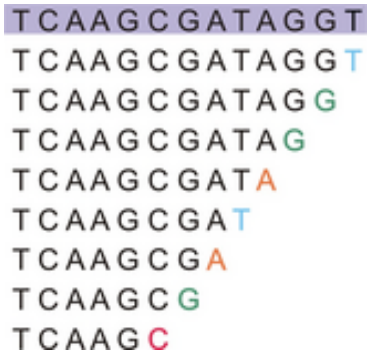

Figura 2.4: A seqüência original esta destacada com fundo colorido. Geram-se vários fragmentos diferindo apenas por uma base, que são ordenados por tamanho. Lendo-se a última base de cada fragmento é possível se reconstruir a seqüência original.

Aplica-se então uma radiação sobre a placa, de forma a excitar os marcadores químicos, que emitem luz, permitindo que sejam detectados visualmente. Em procedimentos mais antigos, era necessário dividir os fragmentos por terminador, e colocá-los em canaletas separadas para a eletroforese, pois não era possível analisar os 4 terminadores ao mesmo tempo. Assim, se obtinha uma placa com quatro sinais separados, cada um representando uma base (Figura 2.5 a). Atualmente, cada marcador emite uma luz diferente, permitindo que se analise os quatro sinais em uma mesma canaleta (Figura 2.5 b).

2.3.3 Leitura das bases

A partir da placa obtida, cria-se um **cromatograma**, exemplificado na Figura 2.5 c, a partir do qual é feita a **leitura das bases**.

A luz emitida pelos marcadores em cada posição é mostrada na forma de uma curva: cada curva representa uma base distinta, e a presença de um pico em uma dada posição do cromatograma indica a base que se encontra naquela posição.

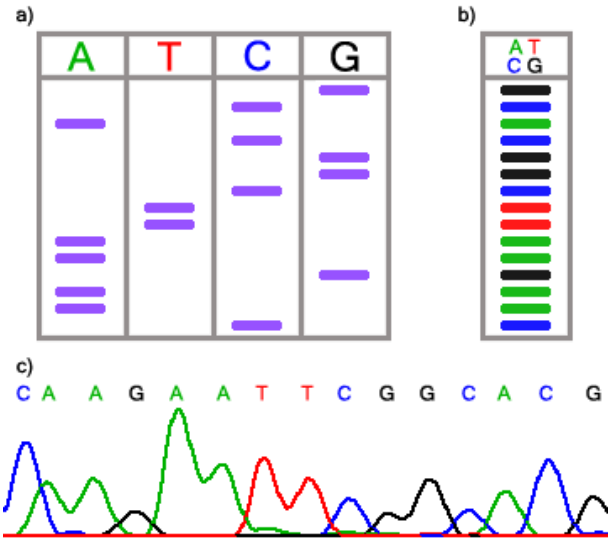

Figura 2.5: Leitura de bases. Em (a) pode se observar o resultado de eletroforese onde cada terminador foi analisado separadamente. Em (b), o resultado da eletroforese com todos os terminadores analisados juntos. Em (c), tem se o cromatograma final obtido a partir da leitura da placa.

Este procedimento é efetuado por programas como o `phred` [30, 31], que analisam o cromatograma e determinam a seqüência de bases de um DNA, em conjunto com um parâmetro indicando a qualidade (ou o erro) associado à cada base.

Nos projetos de seqüenciamento de genomas completos, deve-se reconstruir a seqüência completa original: este processo é chamado de **montagem**, e consiste em alinhar os diversos fragmentos, procurando regiões de sobreposição entre fragmentos, e utilizando-se algoritmos estatísticos para determinar fragmentos consecutivos, colocando-os em seqüência e obtendo o DNA original. O pacote `phrap` [46] executa este processamento.

2.4 Projetos de seqüenciamento

Existem atualmente dois tipos de projetos de seqüenciamento genético: projetos que visam determinar a seqüência genética completa de um organismo, e projetos de seqüenciamento de **ESTs**, que visam determinar apenas cadeias de nucleotídeos que são expressos pelo organismo (ou seja, traduzidos em aminoácidos, cumprindo alguma função).

2.4.1 Projetos genomas

Os projetos genomas têm como objetivo a obtenção da seqüência genética completa de um organismo vivo. Com certeza o mais importante foi o Projeto do Genoma Humano, iniciado em 1988 e finalizado em 2003 e contando com a participação de laboratórios em todo o mundo [2, 6, 23, 102].

Porém, outros projetos de organismos menores têm sido muito úteis para o aperfeiçoamento das técnicas e tecnologias utilizadas para seqüenciamento, além de apresentarem resultados práticos em várias áreas como medicina e agronomia. Dentre esses projetos, deve-se destacar projetos de seqüenciamento de organismos procariotos, que são menores e mais fáceis de serem obtidos. O primeiro grande projeto de seqüenciamento do genoma completo de um organismo foi o projeto da *Drosophila melanogaster* [10].

Na página Entrez Genome mantida pelo NCBI [85] é possível encontrar na versão de 15 de Agosto de 2006 os genomas completos de 1668 vírus, 28 arqueobactérias, 357 bactérias e 22 eucariotos com genomas completos, e seus dados estão em constante evolução. No início da pesquisa deste trabalho, a versão de 14 de Maio de 2004 apresentava 290 vírus, 18 arqueobactérias, 146 bactérias bem como de diversos eucariotos como o homem e o rato.

No Brasil, a FAPESP organizou em 1997 a rede **ONSA** (Organization for Nucleotide Sequencing and Analisys), um instituto virtual de genômica formado inicialmente por 30 laboratórios distantes geograficamente, e ligados a instituições de pesquisa do Estado de São Paulo, e por um centro de bioinformática. O primeiro projeto concluído por esta rede foi o da *Xyllela fastidiosa* [34], primeiro organismo causador de doenças de uma planta a ter seu genoma completamente seqüenciado. Outros projetos desenvolvidos pela rede ONSA foram os projetos *Xanthomonas citri* e *Xanthomonas campestri*.

2.4.2 Projetos EST

Os ESTs (Expressed Sequence Tags) [2] são seqüências genômicas expressas por um organismo, capturadas após o processo de transcrição. Assim, são de grande interesse pois representam regiões dos genes que são relevantes para a compreensão do funcionamento de um ser vivo, não contendo seqüências provenientes de regiões entre genes ou de íntrons, produzindo dados sobre os genes que estão sendo expressos no organismo de maneira imediata.

Os ESTs foram originalmente utilizados para identificação de genes. Posteriormente, descobriu-se que eram extremamente úteis como ferramenta de mapeamento de genoma [105]. Além disso ESTs foram utilizados para se estudar a frequência de SNPs no genoma humano e o fenômeno de *Alternative Splicing* [112].

Os ESTs são obtidos a partir do seqüenciamento de cDNA, que é uma fita de DNA

produzida a partir do complemento do mRNA com a utilização da enzima transcriptase reversa.

O processo de obtenção destes tem várias vantagens, permitindo que o seqüenciamento se torne muito rápido. Dentre as vantagens, pode-se citar o fato de que é possível executar o seqüenciamento de ESTs contendo em média 500 pares de bases em apenas um único procedimento, e o fato de que não existe a necessidade de se repetir os experimentos para verificação da validade do resultado [16].

Os ESTs seqüenciados podem ser obtidos em bases de dados públicas, como a db-EST [25], constantemente atualizada. Na versão 070204 de 02 de julho de 2004 estavam disponíveis 22.482.289 seqüências públicas de ESTs de 712 organismos diferentes. A versão 081106, de 11 de Agosto de 2006, apresentava 38.056.628 seqüências públicas de ESTs de 1179 organismos diferentes.

No Brasil, vários projetos de captura de ESTs foram desenvolvidos. Entre eles, pode-se citar:

- Genoma Cana-de-Açúcar (SUCEST) - Projeto EST que teve início em 1999 e a sua etapa de seqüenciamento foi concluída em 2000 [111, 114, 117].

- Projeto Genoma Humano do Câncer - Projeto EST realizado internacionalmente e que teve como objetivo a descoberta de genes que possuem ligação com os diversos tipos de câncer existentes. A parcela brasileira deste projeto foi formada por diversos grupos, incluindo o Instituto Ludwig para Pesquisa do Câncer [115].

- Projeto FORESTs - Projeto EST de seqüenciamento do eucalipto [33].

- Projeto *Schistossoma mansoni* - Projeto EST de seqüenciamento do organismo causador da esquistossomose [104].

- Projeto Genomas Agronômicos e Ambientais - Projeto multi-genômico de seqüenciamento completo ou EST de organismos ligados à agronomia ou ao ambiente [37].

2.5 SNPs e Farmacogenética

A pesquisa genômica é de grande interesse para a pesquisa médica, tendo em vista que muitas doenças graves em seres humanos são influenciadas por uma predisposição genética. O entendimento de como os genes influenciam na aparição destas doenças é portanto de grande relevância para a detecção prematura e cura destas [43, 75].

2.5.1 SNPs

Um polimorfismo ocorre em uma seqüência genética quando são encontradas duas ou mais formas genéticas (alelos) diferentes em indivíduos da mesma espécie. Para que um alelo seja considerado um polimorfismo, ele deve aparecer em pelo menos 1% da população analisada. Caso contrário, considera-se que o alelo é uma mutação pontual. SNPs, sigla para Single Nucleotide Polymorphisms, ou Polimorfismos de Base Única, são polimorfismos que ocorrem em apenas uma base em indivíduos normais [15].

A priori, os SNPs poderiam ser bi, tri ou tetra alélicos, ou seja, possuirem duas, três ou quatro formas diferentes. Porém os dois últimos tipos são extremamente raros:as variações mais freqüentes são substituições entre bases nitrogenadas de mesma característica estrutural (A/G ou G/A e C/T ou T/C), que são chamadas de transições. As outras substituições são conhecidas com transversões.

Um SNP pode ser sinônimo ou não: no primeiro caso, o aminoácido codificado pelo códon contendo SNP é o mesmo que aquele codificado pelo códon sem SNP. No segundo caso, o códon modificado gera um aminoácido diferente, podendo modificar a estrutura e função da proteína codificada. Estima-se que aproximadamente metade das causas genéticas de doenças são causadas por substituições de aminoácidos [24].

2.5.2 Farmacogenética

O interesse da medicina pela genética nasceu na década de 1950: em 1957, Arno Motulsky foi o primeiro pesquisador a articular pesquisas nesta área de interface entre as duas disciplinas [81].

Observando a resposta de indivíduos a certas drogas, ele argumentou que elementos genéticos poderiam ser subjacentes às variações de sensibilidade a substâncias. Nesta mesma época nasceu o termo **farmacogenética**, criado em 1959 por Friedrich Vogel para designar esta nova área de pesquisa

A publicação do mapeamento do genoma humano gerou um grande entusiasmo em grupos de pesquisa e empresas farmacogenéticas, que visam utilizar estas informações para o estudo e desenvolvimento de tratamentos individualizados. Este entusiasmo se baseia no fato de que a grande maioria dos genes apresentam uma grande freqüência de variações alélicas, conhecidos como polimorfismos, e que essas variações podem ser a chave para a predisposição de certos indivíduos a certas doenças.

Dentre os polimorfismos, os SNPs tem tido um espaço de destaque, por representarem cerca de 90% dos polimorfismos encontrados no genoma humano [21].

2.5.3 SNP e seu interesse para a pesquisa farmacogenética

Existe atualmente um grande esforço para se mapear estes polimorfismos, de forma a se obter dados sobre os quais possam ser feitas pesquisas. Os projetos SNP Consortium e o Human Genome Sequencing Consortium foram os principais responsáveis pelo mapeamento de 1.42 milhões de SNPs no genoma humano disponibilizados publicamente em novembro de 2000 [102].

Com estes dados obtidos, visa-se determinar as modificações do DNA humano que contribuem na sua variação fenotípica. No Brasil, laboratórios participam do Human Cancer Genome Project [51], visando detectar os SNPs ligados ao Câncer de seres humanos.

Estudos conseguiram mapear centenas de genes relacionados a doenças. Porém, raramente apenas uma mutação em um gene é responsável pela contração de uma doença. Em geral o fenótipo é causado por um conjunto de genes atuando de forma complexa. Além do mais a diversidade genética humana não se limita apenas a polimorfismos individuais dentro de genes, mas a uma combinação de alelos próximos um dos outros, atuando em conjunto [44, 64].

Muito do sucesso do mapeamento genético de doenças no passado foi devido ao fato de que as primeiras doenças humanas estudadas eram bastante simples, monogênicas, obedecendo às regras mendelianas de herança genotípica. Porém, em muitos casos, é necessário se obter dados de uma grande população para se conseguir definir de forma precisa regiões responsável pela aparição de uma doença.

Recentemente, geneticistas voltaram seus esforços para mapear doenças mais comuns, cuja base genética é mais complexa e que afetam grandes fatias da população, e onde os métodos tradicionais têm se mostrado menos eficientes [6]. Métodos novos foram idealizados para se determinar com mais precisão possíveis regiões onde os genes causadores de uma dada doença estariam localizados, utilizando para isso marcadores, como os SNPs.

2.6 Métodos de detecção de SNPs

Existem basicamente duas categorias de métodos para se detectar SNPs. A primeira utiliza procedimentos químicos, e a segunda é baseada em comparação de seqüências genômicas com o auxílio de ferramentas computacionais. É muito comum a utilização de métodos computacionais para análises preliminares, devido ao baixo custo e possibilidade de execução de testes em grandes quantidades de dados, e posteriormente o uso de métodos químicos para validação dos dados. A seguir serão descritos brevemente cada método.

2.6.1 Análise de PCR-RFLP

A maneira mais usada atualmente para estudar SNPs é determinar sua existência através da análise de **PCR-RFLP**. RFLP, ou Restricition Fragment Length Polymorphisms, foram os primeiros tipos de marcadores de DNA estudados [16]. Fragmentos de restrição são produzidos quando uma molécula de DNA é tratada por uma enzima de restrição, que corta a molécula em seqüências pré-definidas (por exemplo, um enzima corta a molécula sempre que encontrar a seqüência ACGTTG). Os locais onde uma enzima de restrição age sobre uma molécula de DNA são chamados de **sítios de restrição**.

O fato de que uma enzima de restrição age sempre sobre uma determinada seqüência faz com que o tratamento de uma fita de DNA com uma enzima deveria sempre produzir o mesmo conjunto de fragmentos. Porém isto não acontece no caso de existir um polimorfismo em um sítio de restrição.

Assim, quando um polimorfismo cria ou destrói um sítio de restrição, basta realizar uma reação PCR, digerir o produto e verificar em um gel se aquele polimorfismo existe ou não. Apesar de ser simples e barato, este método limita os SNPs que podem ser estudados, pois só permite trabalhar com SNPs já conhecidos, e não permite trabalhar com SNPs que não criem ou destruam sítios de restrição. Mesmo nestes casos, por vezes a enzima de restrição é cara ou indisponível, e o estudo não pode ser feito. Este método ainda é uma herança dos tempos em que seqüenciar era difícil e caro.

2.6.2 Comparação de seqüências

Uma nova maneira de estudar SNPs consiste em utilizar seqüenciamento de DNA. Escolhe-se uma região genômica de interesse e seqüencia-se esta região de vários indivíduos. As seqüências obtidas são alinhadas, utilizando algoritmos de alinhamento [19, 32, 45, 55, 60, 80, 83, 84, 106, 122]. O alinhamento obtido permite a comparação entre as seqüências, e a detecção de possíveis SNPs.

Este método está se popularizando cada vez mais, inclusive no Brasil, pois o custo do seqüenciamento vem caindo muito rapidamente, devido ao aumento do parque de seqüenciadores já instalados em nosso país. Além disto, uma análise por seqüenciamento, acoplada a um software adequado, tem grande poder para identificar novos SNPs, pois certamente ainda há muitos que não foram identificados, e não possui as limitações apontadas na técnica clássica descrita acima.

Os capítulos a seguir apresentam técnicas que se baseiam em métodos computacionais para detectar e analisar SNPs a partir de dados disponíveis em bases de dados públicas.

Capítulo 3

Alinhamento de DNA com cDNA

Neste capítulo, iremos discutir problemas relacionados ao alinhamento de cDNA com DNA genômico. Esta operação é de grande importância para a detecção de SNPs, pois o alinhamento permite que duas ou mais seqüências sejam comparadas base a base. Os resultados apresentados neste capítulo foram descritos no artigo "*Comparison of genomic DNA to cDNA alignment methods*" por Galves e Dias [35], publicado no *Lecture Notes on Bioinformatics* e apresentado no Brazilian Symposium on Bioinformatics – BSB 2005, realizado entre 25 e 29 de julho de 2005 em Porto Alegre – RS.

Vários pacotes computacionais foram especialmente desenvolvidos para resolver este problema, utilizando diversas heurísticas. O objetivo deste capítulo é identificar um conjunto de parâmetros que permita bons alinhamento entre DNA genômico e cDNA utilizando algoritmos clássicos de alinhamento.

Os resultados obtidos serão comparados com os resultados obtidos pelos pacotes computacionais `sim4` [32], `Spidey` [122] e `est_genome` [80], que foram especialmente desenvolvidos para alinhar cDNA e DNA genômico.

Na Seção 3.1 descreveremos os problemas relacionados ao alinhamento de cDNA com DNA. Na Seção 3.2 descreveremos os algoritmos clássicos de alinhamento de duas seqüências quaisquer. Na Seção 3.3 discutiremos estratégias utilizadas por alguns pacotes computacionais para obter alinhamentos corretos entre cDNA e DNA. Na Seção 3.4 serão descritos brevemente os algoritmos dos alinhadores utilizados neste capítulo. Na Seção 3.5 descreveremos o conjunto de dados utilizado neste capítulo. Na Seção 3.6 descreveremos a metodologia de testes utilizada para testar e avaliar os alinhadores. Os resultados obtidos serão analisados na Seção 3.7. Finalmente, na Seção 3.8 faremos uma breve conclusão dos resultados obtidos.

3.1 Alinhamento de cDNA com DNA genômico

Como visto no Capítulo anterior, os genes podem possuir regiões codificadoras (que serão usadas como base para a síntese de proteínas), e regiões não-codificadoras.

No processo de transcrição de genes em organismos eucariotos (como os seres humanos), as regiões não-codificadoras (também chamadas de íntron) são removidas, gerando um transcrito maduro. Além disso, nesse momento pode ocorrer `alternative splicing`, onde apenas um subconjunto dos éxons de um gene são transcritos.

Os projetos EST capturam o mRNA resultante deste processo, que sai do núcleo da célula para que seja efetuada a fase de tradução. Os ESTs são a chave para o entendimento do funcionamento interno de um organismo [122]. Porém, para que se entenda completamente o seu funcionamento, seqüências expressas tem que ser postas no seu contexto genômico, ou seja, é necessário que sejam identificadas as regiões genômicas correspondentes aos ESTs. Por isso, alinhadores de mRNA com genomas são de grande importância.

O grande desafio para os alinhadores de ESTs com DNA genômico é conseguir delimitar os diferentes éxons dentro do mRNA capturado pelos projetos ESTs e encontrar as regiões de origem na seqüencia do gene ou do cromossomo.

3.2 Alinhamento de duas seqüências

Define-se um alinhamento entre duas seqüências como sendo uma operação de inserção de espaços nas duas seqüências de forma que elas tenham o mesmo tamanho, e que se possa sobrepô-las permitindo a comparação das bases [106].

Por exemplo, dadas as seqüências ACGTTTG e ACGTTTTG, podemos facilmente observar que inserindo um espaço entre os dois últimos caracteres da primeira seqüência, teremos duas seqüências praticamente idênticas:

```
ACGTTT-G
ACGTTTTG
```

Dado um alinhamento como definido acima, podemos criar uma pontuação para avaliar a qualidade do resultado obtido. A pontuação mais simples que se pode dar é uma penalização para o alinhamento de uma base com um espaço (chamada de `gap`), um valor para alinhamento de bases distintas (chamada de `mismatch`) e um valor para alinhamento de bases iguais (chamada de `match`). Por exemplo se considerarmos $\texttt{gap} = -2$, $\texttt{match} = 1$ e $\texttt{mismatch} = -1$ teremos a seguinte pontuação para o alinhamento acima: $7 \times 1 - 2 = 5$.

O objetivo dos algoritmos de alinhamento existentes é obter um alinhamento ótimo, ou seja, que possua a maior pontuação possível. A seguir serão descritas três tipos de alinhamentos: global (3.2.1), semi-global (3.2.2) e local (3.2.3).

Além disso, serão descritos esquemas de pontuação baseados em funções lineares, mais apropriados para alinhamentos de seqüências reais (3.2.4). Nos exemplos utilizados nestas seções utilizaremos os coeficientes `gap`, `match` e `mismatch` definidos acima.

3.2.1 Alinhamento global

O objetivo do alinhamento global é gerar o melhor alinhamento possível entre duas seqüências. Os espaços podem ser inseridos em qualquer posição das seqüências, de forma a se obter a pontuação ótima.

Por exemplo, as seqüências ACCG e CG poderiam produzir os seguintes resultados

```
ACCG
-C-G
 ou
ACCG
--CG
```

produzindo a mesma pontuação -2.

O algoritmo, proposto por Gotoh [45], tem como princípio básico a determinação do alinhamento de duas seqüências a partir da construção de uma matriz de similaridade, calculando a pontuação do alinhamentos de prefixos arbitrários.

Sejam s e t duas seqüências a serem alinhadas, e m e n seus tamanhos respectivos. Cria-se uma matriz sim contendo $(m+1) \times (n+1)$ posições, onde a posição $sim[i,j]$ da matriz representa o melhor alinhamento obtido do prefixo $s[1..i]$ com o prefixo $t[1..j]$.

Inicialmente, os elementos da primeira coluna recebem o valor $-(i \times gap)$ e os elementos da primeira linha recebem o valor $-(j \times gap)$. O algoritmo percorre a matriz, da esquerda para a direita e de cima para baixo, com $1 \leq i \leq m$ e $1 \leq j \leq n$, e para cada posição $sim[i,j]$ da matriz, calcula a pontuação do alinhamento naquela posição da seguinte forma:

$$sim[i,j] = max \begin{cases} sim[i,j-1] - gap \\ sim[i-1,j-1] + p(i,j) \\ sim[i-1,j] - gap \end{cases}$$

onde $p(i,j)$ é a função que retorna a pontuação para o alinhamento do símbolo $s[i]$ com o símbolo $t[j]$.

No sistema definido acima, a primeira equação representa o alinhamento de $s[1..i]$ com $t[1..j-1]$ e o casamento de $t[j]$ com um espaço, a segunda equação representa o alinhamento de $s[1..i-1]$ com $t[1..j-1]$ e o casamento de $s[i]$ com $t[j]$, e a terceira equação representa o alinhamento de $s[1..i-1]$ com $t[1..j]$ e o casamento de $s[i]$ com um espaço.

Ao final do algoritmo, a posição $sim[m,n]$ contém o valor do alinhamento ótimo. Para se construir este alinhamento, deve-se partir da posição $sim[m,n]$ e percorrer a matriz passando pelas posições que geraram a pontuação ótima. Para isso, aplica-se o seguinte algoritmo: se $sim[i,j]$ for a posição corrente, verifica-se a partir de qual das três equações enunciadas acima foi definido seu valor. Se $sim[i,j] = sim[i-1,j-1] + p(i,j)$, então devemos ir para $sim[i-1,j-1]$, se $sim[i,j] = sim[i-1,j] - gap$, então devemos ir para $sim[i-1,j]$ e finalmente se $sim[i,j] = sim[i,j-1] - gap$ então devemos ir para $sim[i,j-1]$.

Aplica-se este algoritmo iterativamente até chegarmos a $sim[0,0]$. O movimento efetuado define qual a operação a ser feita: alinhamento de $s[i]$ com um espaço (movimento vertical, i varia), alinhamento de $t[j]$ com um espaço (movimento horizontal, j varia) ou alinhamento de $s[i]$ com $t[j]$ (movimento diagonal, i e j variam). Este algoritmo utiliza tempo e espaço $O(mn)$.

3.2.2 Alinhamento semi-global

A estratégia semi-global tenta obter o melhor alinhamento entre um prefixo de uma seqüência com o sufixo da outra. Para isso, tenta agrupar o maior número de espaços no início e no final do alinhamento, não penalizando a criação destes.

O propósito desta estratégia é tentar obter o melhor alinhamento possível descartando as extremidades das seqüencias, que em geral possuem taxas de erros de seqüenciamento maiores do que as regiões internas.

Por exemplo, um alinhamento global das seqüências ACTGACCTCGGG e ACCGTCGGGCGG produziria o resultado

```
ACTGACCTCGGG
ACCGTCGGGCGG
```

com pontuação 0, enquanto que o alinhamento semi-global com os mesmos coeficientes produziria o resultado

```
ACTGACC-TCGGG---
----ACCGTCGGGCGG
```

com pontuação 6.

O algoritmo de construção deste tipo de alinhamento é muito semelhante ao global. A primeira linha ($i = 0$) e primeira coluna ($j = 0$) da matriz de similaridade são inicializadas com valor 0. Os valores da cada posição são calculados de forma idêntica ao da matriz de similaridade global. Para se reconstruir o alinhamento, procura-se o maior valor dentre os elementos da última linha ($i = m$) e da última coluna da matriz ($j = n$), e efetua-se o percurso reverso nas posições da matriz que geraram o alinhamento semi-global ótimo, até se chegar em $i = 0$ ou $j = 0$.

3.2.3 Alinhamento local

O alinhamento local tem como objetivo encontrar duas sub-seqüências (uma em cada seqüência original) que produzem o alinhamento com maior pontuação possível.

Por exemplo, um alinhamento local das seqüências ACCATCTTGC e TCCCGTGTA-AAA produziria o resultado

```
CC
CC
```

com pontuação 2.

O algoritmo de construção deste tipo de alinhamento inicializa a primeira linha e primeira coluna da matriz com valor 0, como o semi-global, e não admite valores negativos durante a execução do algoritmo. Assim, a regra de cálculo do valor de uma dada posição é

$$sim[i,j] = max \begin{cases} sim[i-1,j] - gap \\ sim[i-1,j-1] + p(i,j) \\ sim[i,j-1] - gap \\ 0 \end{cases}$$

Para reconstruir o alinhamento, o algoritmo encontra a posição $sim[i_1, j_1]$ contendo o maior valor da matriz de similaridade e a percorre em sentido inverso, indo sempre de uma posição para a posição que gerou o seu valor (de forma idêntica ao algoritmo global e semi-global). O algoritmo pára ao chegar na posição $sim[i_2, j_2]$ que contenha valor 0. Assim, o alinhamento com maior pontuação ocorrerá entre $s[i_2..i_1]$ e $t[j_2..j_1]$.

3.2.4 Função linear para criação de buracos

Nas seções anteriores, utilizou-se um sistema de pontuação que não discrimina a criação de buracos separados ou contíguos. Assim, as seqüências ACCG e CG poderiam produzir os alinhamentos

```
ACCG
-C-G
 ou
ACCG
--CG
```

com a mesma pontuação.

Porém observou-se que de forma geral é muito mais comum a existência de buracos contíguos de tamanho k do que a existência de k buracos isolados [106, Seção 3.3, pg. 64]. Assim foi necessário desenvolver uma estratégia que agrupasse o máximo possível os buracos inseridos em um alinhamento, penalizando mais a criação de buracos isolados do que a criação de buracos contíguos.

Para resolver este problema, substituiu-se o coeficiente *gap* por uma função linear $w(k) = g + hk$, onde k é o número de buraco contíguos, g o custo de se abrir um novo buraco e h o custo de se estender um buraco aberto. Os coeficientes g e h também são chamados de *open gap* e *extended gap*. Por exemplo, supondo $g = -2$ e $h = -1$, e calculando a pontuação para os dois alinhamentos obtidos acima, teríamos -4 para o primeiro alinhamento e -2 para o segundo.

A implementação desta função nos algoritmos de alinhamento global, semi-global e local requer a utilização de três matrizes A, B e C ao invés de uma. $A[i, j]$ representa a pontuação máxima do alinhamento de $s[1..i]$ com $t[1..j]$ que termina com $s[i]$ casado com $t[j]$, $B[i, j]$ representa a pontuação máxima do alinhamento de $s[1..i]$ com $t[1..j]$ que termina com um espaço casado com $t[j]$, e $C[i, j]$ representa a pontuação máxima do alinhamento de $s[1..i]$ com $t[1..j]$ que termina com $s[i]$ casado com um espaço.

Para o alinhamento global, inicializam-se as matrizes da seguinte forma:

$$\begin{cases} A[0,0] = 0, \\ A[i,0] = -\infty & \text{para } 1 \leq i \leq m \text{ e} \\ A[0,j] = -\infty & \text{para } 1 \leq j \leq n \end{cases}$$

$$\begin{cases} B[i,0] = -\infty & \text{para } 0 \leq i \leq \text{e} \\ B[0,j] = -(h + gj) & \text{para } 1 \leq j \leq n \end{cases}$$

$$\begin{cases} C[i,0] = -(h + gj) & \text{para } 1 \leq i \leq m \text{ e} \\ C[0,j] = -\infty & \text{para } 0 \leq j \leq n \end{cases}$$

A regra de cálculo dos coeficientes das matrizes é a seguinte:

$$A[i,j] = p(i,j) + max \begin{cases} A[i-1,j-1] \\ B[i-1,j-1] \\ C[i-1,j-1] \end{cases}$$

$$B[i,j] = max \begin{cases} -(h+g) + A[i,j-1] \\ -g + B[i,j-1] \\ -(h+g) + C[i,j-1] \end{cases}$$

$$C[i,j] = max \begin{cases} -(h+g) + A[i-1,j] \\ -(h+g) + B[i-1,j] \\ -g + C[i-1,j] \end{cases}$$

A execução do algoritmo para construção do alinhamento global é muito semelhante ao algoritmo descrito anteriormente. Procura-se o maior valor dentre $A[m,n]$, $B[m,n]$ e $C[m,n]$: a partir do elemento contendo o maior valor das três matrizes, percorrem-se os elementos que geraram o alinhamento ótimo, sempre indo da posição corrente (i, j) para a posição que gerou o valor atual. Se o valor veio da primeira equação de algum dos sistema, então move-se para $A(i-i, j-i)$ casando-se $s[i]$ com $t[j]$. Se o valor veio da segunda equação de algum dos sistemas, então move-se para $B(i-i, j)$ casando-se $s[i]$ um espaço. Se o valor veio da terceira equação de algum dos sistemas, então move-se para $C(i, j-1)$ casando-se um espaço com $t[j]$.

3.3 Pacotes para alinhamento de cDNA com DNA genômico

Várias estratégias foram desenvolvidas para resolver o problema de alinhamento de cDNA e ESTs com DNA genômico, utilizando algoritmos baseados em várias etapas e ferramentas distintas.

Nas Seção 3.3.1 descreveremos brevemente os pacotes computacionais utilizados, e na Seção 3.3.2 descrevemos algumas comparações de performance e qualidade de resultados obtidos por esses pacotes.

3.3.1 Descrição dos pacotes

O programa `est_genome` [80] foi desenvolvido para alinhar pedaços de seqüências com seqüências genômicas, permitindo a existência de grandes íntrons, reconhecendo sítios de splice e utilizando memória limitada. Seu algoritmo principal é uma modificação do algoritmo de Smith e Waterman [108].

O programa `sim4` [32] assume que as diferenças entre as seqüências a serem alinhadas se resumem a presença de íntrons na seqüência genômica e erros de seqüenciamento em ambas

as seqüências. A ferramenta foi utilizada para alinhar dados obtidos através do projeto de seqüenciamento da *Drosophila Melanogaster*, desenvolvido pelo Berkeley Drosophila Genome Project (BDGP) [10].

A ferramenta `Spidey` [122] produz alinhamentos de mRNAs com genomas, utilizando o toolkit do NCBI [85]. Seus principais objetivos são produzir bons alinhamentos a despeito do tamanho de íntrons (regiões não codificadoras) e não gerar erros devido a genes parálogos (genes que tem uma origem comum e aparecem no mesmo genoma) e pseudo-genes.

3.3.2 Comparação de desempenho dos programas analisados

Os autores dos projetos descritos acima fizeram testes comparativos entre as três ferramentas, e os resultados serão brevemente descritos nesta seção.

Florea [32] comparou o pacote `sim4` com o pacote `est_genome`, efetuando um teste de alinhamento de 184 CDSs (região do mRNA que será efetivamente traduzida em uma proteína) com a seqüência original, em uma máquina com processador Intel Pentium II dual 266MHz com 128Mb de memória rodando com o sistema operacional Red Hat Linux 5.0. O `est_genome` demorou 20 segundos por seqüência e acertou 143 alinhamentos, e o `sim4` demorou 0.06s por seqüência, acertou 166 alinhamentos no modo normal e acertou 172 alinhamentos no modo otimizado.

Wheelan [122] efetuou dois testes comparativos entre o pacote `Spidey` e as duas outras ferramentas analisadas. O primeiro teste consistiu em se fazer alinhamentos com seqüências de referência, de onde foram extraídos 646 mRNAs anotados, contendo um total de 3915 éxons. Estes mRNAs foram então alinhados com a seqüência original (tendo portanto uma semelhança de 100%): `Spidey` reconheceu 3873 éxons, dos quais 98.7% estavam corretos, `sim4` reconheceu 3909 éxons, dos quais 97.9% estavam corretos, e `est_genome` reconheceu 3716 éxons, dos quais 97.4% estavam corretos.

O segundo teste foi feito alinhando-se genes ortólogos (genes que tem uma origem comum e aparecem em genomas diferentes) de ratos com seqüências humanas de referência. `Spidey` foi configurado no modo inter-espécie, no qual são utilizados parâmetros diferentes de configuração do BLAST visando criar mais buracos maiores e não penalizar muito mismatch. Os programas `sim4` e `est_genome` foram utilizados com suas configurações normais. `Spidey` acertou 81.4% dos éxons, `sim4` acertou 53.9% e `est_genome` acertou apenas 37.2%.

Em relação ao tempo de processamento, `Spidey` e `sim4` se mostraram muito superiores a `est_genome`: para alinhar um mRNA com 5164bp com um contig de 1.03Mb em uma Sun Ultra 10 300MHz com 192Mb de memória, `Spidey` levou 14s, `sim4` levou 2s e `est_genome` levou 1h21m. Para processar 35 mRNAs com suas seqüências de referências, `Spidey` levou

1m11s, `sim4` levou 25s segundos e `est_genome` levou 2h56m.

3.4 Implementação de alinhadores global e semi-global com pontuação afim e espaço linear

Para desenvolvermos o trabalho proposto neste capítulo, foi necessário ter um conjunto de alinhadores com pontuação afim e espaço linear. Optamos por implementar uma versão própria destes alinhadores em vez de usar pacotes já existentes, como o pacote **fasta** [93].

Para implementar o alinhador global, utilizamos o algoritmo de espaço linear proposto por Miller e Myers [84]. O alinhador semi-global foi implementado baseado no algoritmo global. A seguir apresentaremos os resultados de testes de validação dos alinhadores obtidos. Os algoritmos foram implementados em Java.

3.4.1 Validação dos alinhadores

Para validar os alinhadores, fizemos testes de alinhamento com 1.000.000 de pares de seqüências geradas aleatoriamente com tamanhos variando entre 50bp e 400bp. Os pares de seqüências foram alinhados com nossos alinhadores e com os alinhadores do pacote `fasta`, e todos os resultados foram idênticos.

Efetuamos testes para comparar o tempo de execução dos alinhadores lineares, quadráticos e do pacote fasta. Foram efetuados alinhamentos com seqüências de tamanhos 50, 100, 200, 400, 1600, 3200 e 6400 bases. Para cada um dos tamanhos, geramos 50 pares de seqüências distintos.

Executamos os testes em uma máquina com processador Pentium 4 1.7MHz com 512MB de memória RAM, rodando o sistema Linux Fedora Core 3. Os resultados se encontram nas tabelas 3.1 (global) e 3.2 (semi-global).

Podemos observar que para o alinhador quadrático somente foram efetuados testes com seqüências de tamanho 50, 100, 200, 400 e 800 bases. Isto ocorreu porque a máquina virtual não conseguiu executar os testes com seqüências maiores por falta de memória.

É interessante notar que apesar dos resultados teóricos afirmarem que o algoritmo quadrático é duas vezes mais rápido que o algoritmo linear, os resultados obtidos com nossos alinhadores mostram que o algoritmo linear é significativamente mais eficiente. Isto se deve ao impacto do processo de alocação de memória, não desprezível para as quantidades exigidas para o alinhamento de seqüências genéticas.

Alinhador Global

Tamanho (bases)	**Linear (ms/alin)**	**Quadrático (ms/alin)**	**align (ms/alin)**
50	3.18	3.92	11.16
100	13.16	60.46	34.96
200	26.1	204.36	62.06
400	77.88	663.24	111.22
800	261.6	2165.70	207.36
1600	919.94	-	434.88
3200	3594.12	-	1197.88
6400	14400.84	-	4077.40

Tabela 3.1: Tempo médio em milisegundos por alinhamento de execução dos alinhadores globais com espaço quadrático, linear e do alinhador `align` do pacote **fasta**. Pode-se observar que para o alinhador quadrático somente foram efetuados testes com seqüências de tamanho 50, 100, 200, 400 e 800 bases. Isto ocorreu porque a máquina virtual não conseguiu executar os testes com seqüências maiores por falta de memória.

Alinhador Semi-Global

Tamanho (bases)	**Linear (ms/alin)**	**Quadrático (ms/alin)**	**align0 (ms/alin)**
50	1.90	2.96	35.38
100	4.28	17.34	75.30
200	11.90	121.58	106.20
400	38.54	597.46	148.68
800	142.28	2095.18	228.62
1600	523.48	-	452.30
3200	2055.14	-	1203.84
6400	9974.84	-	4091.98

Tabela 3.2: Tempo médio em milisegundos por alinhamento de execução dos alinhadores semi-globais com espaço quadrático, linear e do alinhador `align0` do pacote **fasta**. Pode-se observar que para o alinhador quadrático somente foram efetuados testes com seqüências de tamanho 50, 100, 200, 400 e 800 bases. Isto ocorreu porque a máquina virtual não conseguiu executar os testes com seqüências maiores por falta de memória.

3.5 Dados para testes

Para testar as estratégias desenvolvidas, construímos uma base de dados contendo informações sobre cromossomos, genes, mRNAs e CDSs do genoma humano.

3.5.1 Fontes de dados

Todos os dados foram obtidos através do site do NCBI [85]. As seqüências completas dos cromossomos utilizados foram obtidos no formato FASTA, disponibilizados no repositório no dia 09/10/2004. As informações sobre genes, mRNAs, CDSs e outros tipos de seqüência de cada cromossomo foram obtidas no formato GenBank (.gbk), disponibilizados no repositório no dia 26/08/2004.

3.5.2 Dados utilizados

Foi necessário filtrar os dados sobre o genoma humano, visando a construção de um conjunto consistente de dados para os testes. Construímos uma base de dados relacional para armazenamento, pesquisa e análise dos dados coletados. As informações de interesse armazenadas nesta base são genes, mRNAs, CDSs e UTRs (regiões de um mRNA não traduzidas em proteína).

O modelo utilizado para a seleção dos dados foi o seguinte: cada cromossomo deve possuir um ou mais genes, cada gene possui um ou mais mRNAs, e cada mRNA é formado por regiões de CDS e por nenhum, um ou dois UTRs. A região de um mRNA deve ser totalmente coberta pelas regiões de CDS e UTRs, e CDSs e UTRs não podem conter regiões não presentes no mRNA. Os padrões de mRNAs e CDSs aceitos para inserção na base de dados são exemplificados na Figura 3.1, e a Figura 3.2 exemplifica os padrões de mRNAs e CDSs eliminados.

Do total de genes obtidos dos arquivos, 6.72% foram removidos por serem incompletos (não conterem mRNAs) e 6.78% foram removidos por serem pseudo-genes. Apenas 0.06% dos mRNAs foram removidos por serem pseudo-mRNAs. Do total de CDSs, 0.17% foram removidos por serem pseudo-CDSs e 1.7% foram removidos por estarem fora do padrão.

A Tabela 3.3 mostra a quantidade de dados obtidos por cromossomo após a filtragem. Foram inseridos 23124 genes (86.5% do total de genes) e 27448 mRNAs, dos quais 9.48% não contêm UTR, 4.74% contém apenas 5'-UTR, 4.33% contém apenas 3'-UTR e 81.45% contém 3'-UTR e 5'-UTR.

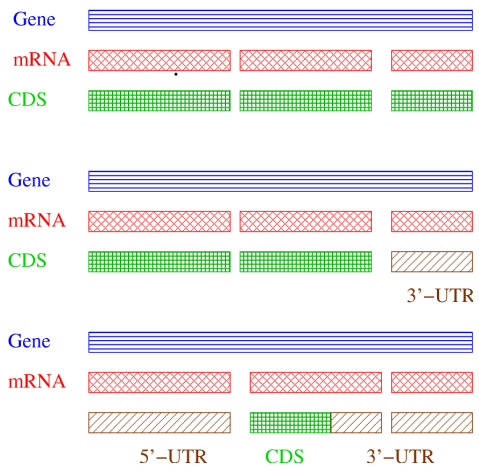

Figura 3.1: Exemplos de padrão de mRNA e CDS considerados corretos. Na Figura temos um mRNA formado por três CDS sem UTR; um mRNA formado por dois CDSs e uma ou mais regiões que formam um 3'-UTR; um mRNA formado por um CDS, uma região ou mais regiões de que formam um 5'-UTR e duas regiões que formam um 3'-UTR.

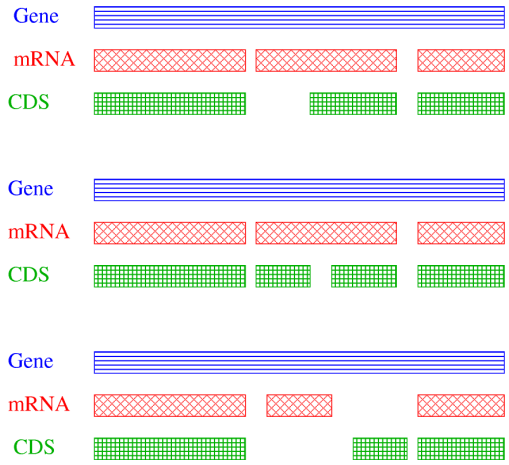

Figura 3.2: Exemplos de padrão de mRNA e CDS considerados errados. Na Figura temos uma das regiões internas do mRNA não completamente coberta pelo CDS; uma das regiões do mRNA parcialmente coberta por duas regiões de CDS; regiões do mRNA não estão presentes no CDS e/ou vice-versa.

	Genes	**mRNAs**	**Regiões de CDS**	**Regiões de UTR**
1	2391	2796	24783	6595
2	1523	1816	19033	4530
3	1242	1504	14278	4009
4	904	1028	9144	2512
5	1069	1264	10610	2764
6	1195	1404	12144	3331
7	1139	1349	12381	3213
8	839	987	8700	2378
9	947	1101	9693	2586
10	884	1097	11051	2649
11	1393	1633	13457	3654
12	1149	1380	13294	3531
13	423	474	4103	1123
14	676	827	7277	2093
15	783	921	8851	2372
16	970	1137	10500	2828
17	1289	1515	14569	3742
18	360	423	3822	953
19	1496	1724	13188	4197
20	596	833	6950	2175
21	271	360	3061	917
22	544	653	5400	1581
X	947	1142	9348	2893
Y	72	80	545	158
Total	**23102**	**27448**	**246182**	**66784**

Tabela 3.3: Numero de genes, mRNAs, regiões de CDS e de UTR por cromossomo inseridos na base de dados após processamento e filtragem dos dados do genoma humano.

3.6 Metodologia de testes

O objetivo desta fase do estudo é determinar o algoritmo (global ou semi-global) e o esquema de pontuação mais adequados para obter bons alinhamentos de ESTs com DNA genômico. Para testar as estratégias propostas, optou-se por trabalhar com subconjuntos de genes, de forma a permitir gerar uma grande quantidade de resultados em um tempo menor, mantendo a validade dos resultados. A seguir serão descritos os diferentes conjuntos utilizados para os testes.

Conjunto 1 Este conjunto é formado por genes com menos de 100000 bases do cromossomo Y. No total, são 66 mRNAs, com tamanho médio de 1305 bases, a serem alinhados com 64 genes, com tamanho médio de 14160 bases. Este conjunto possui genes com grandes blocos de símbolos N (bases desconhecidas), e portanto é mais complicado de ser alinhado corretamente.

Conjunto 2 Este conjunto é formado por genes com menos de 100000 bases e completos (portanto, não contendo bases N) do cromossomo Y. No total, são 50 mRNAs com tamanho médio de 1208 bases a serem alinhados com 40 genes, com tamanho médio de 11272 bases.

Conjunto 3 Este conjunto é formado por todos os genes do genoma inseridos na bases de dados com menos de 10000 bases e completos (portanto, não contendo bases N). No total, são 8056 mRNAs com tamanho médio de 1530 bases a serem alinhados com 7376 genes, com tamanho médio de 4233 bases.

Conjunto 4 Conjunto formado por sub-seqüências de genes completos do cromossomo 6 com menos de 10000 bases (493 genes selecionados), nas quais foram inseridos erros aleatórios. Para cada gene selecionado, foram extraídas 10 seqüências com tamanho variando entre 200 e 1000 bases. Em 5 delas, foram inseridos 1% de erros e nas 5 restantes foram inseridos 10% de erros (inserções, remoções e mutações). O conjunto final é composto por 4930 ESTs gerados artificialmente, a serem alinhados com 493 genes.

Para cada pontuação, geramos 8 tipos de alinhamentos distintos: Gene x mRNA (CDS+UTR), Gene x CDS, Gene+200 bases (100 bases em cada extremidade do gene) x mRNA e Gene+200 bases x CDS, utilizando alinhador global e semi-global. Inicialmente utilizamos apenas os conjuntos 1 e 2, de forma a ser possível testar uma grande quantidade de pontuações.

Uma vez determinados o método de alinhamento e o esquema de pontuação mais adequado, efetuamos os 8 tipos de alinhamentos definidos acima apenas utilizando os conjuntos de dados 3 e 4.

Além disso, para todos os conjuntos, efetuamos alinhamentos com os programas `sim4` [32], `Spidey` [122] e `est_genome` [80], utilizando as configurações padrão de cada um.

3.6.1 Métodos de avaliação

Foram definidas algumas avaliações para, em um primeiro momento, determinar o melhor método e esquema de pontuação, e a seguir comparar os resultados obtidos por estes com os resultados obtidos pelos programas externos.

Buracos inseridos no DNA genômico (Extra Gap) Conta o número de buracos inseridos no DNA genômico pelo alinhador. Utilizando mRNAs ou CDS sem erros, o resultado esperado é 0, e todos os valores devem ser superiores ou iguais a 0.

Número de éxons gerados (Delta Exon) Esta avaliação obtém o número de éxons que formam o mRNA e subtrai o número de éxons obtidos pelo alinhador.

Valores positivos indicam que o alinhador criou menos éxons do que o esperado, e valores negativos indicam que o alinhador criou mais éxons que o esperado. Idealmente o valor deve ser 0 para todos os alinhamentos produzidos.

É importante ressaltar que como o alinhador semi-global utilizado não tem como função determinar as regiões de éxons, consideramos como éxons as regiões contíguas de bases no EST ou mRNA alinhado. Em alinhamentos exatos (conjuntos 1, 2 e 3) esta aproximação é válida uma vez que os CDS representam regiões contíguas do gene e portanto não contém buracos.

Similaridade entre bases de éxons Este método de avaliação visa comparar o grau de similaridade entre os blocos de éxons do cDNA predito com os blocos de bases obtidos pelo alinhador. Para isso, faz a comparação das bases em uma dada posição da seqüência de cDNA predita com o cDNA obtido: o resultado da avaliação é dado pela divisão do número de bases iguais pela soma das bases do cDNA original.

Por exemplo, supondo que a primeira seqüência abaixo corresponda ao cDNA predito e a segunda seqüência ao cDNA obtido pelo alinhador

```
Esperado: ACTGACC------TCGGG---ACCGGCTTCGC
Obtido:   -ACTGACC------TCGGG--ACCGGCTTCGC
```

temos $14/23 = 60{,}87\%$ de bases iguais nas mesmas posições entre as duas seqüências.

Pontuação do alinhamento (Score) Esta avaliação subtrai a pontuação do alinhamento predito com a pontuação obtida pelo alinhamento gerado. Idealmente o valor deve ser 0. Esta avaliação foi apenas utilizada para os alinhamentos efetuados com o alinhador semi-global e global, uma vez que as outras ferramentas utilizam esquemas distintos de avaliação dos resultados obtidos.

Porcentagem de substituições (Mismatch) Esta avaliação conta o número de substituições geradas pelo alinhador e divide pelo número total de bases do EST ou cDNA original. Idealmente o valor deve ser 0.

3.7 Análise dos resultados obtidos

O estudo comparativo dos resultados produzidos por um alinhador convencional e alinhadores especiais externos foi dividido em várias etapas. Num primeiro momento os testes foram orientados para definir uma estratégia (global ou semi-global) e pontuação adequadas para alinhar ESTs com DNA. Definidos estes elementos, foram feitas comparações com os resultados obtidos com os alinhadores externos, primeiramente com dados exatos, e depois com seqüências com erros.

3.7.1 Definição da estratégia de alinhamento e do esquema de pontuação

Com base na metodologia descrita anteriormente, testamos algumas dezenas de esquemas de pontuação, utilizando os conjuntos de dados 1 e 2. Nas análises a seguir, iremos considerar apenas os 10 esquemas de pontuação que apresentaram melhores resultados. Denotaremos o esquema de pontuação pela tupla `(match,mistmatch,open gap,extended gap)`. Vale ressaltar que a pontuação (5, -4, -12, -4) é a pontuação padrão utilizada pelos programas do pacote fasta [93].

O primeiro critério observado para tentar definir uma pontuação e um algoritmo apropriado foi o número de buracos inseridos no DNA genômico: uma vez que mRNAs, CDSs e UTRs foram extraídos da seqüência original, um bom alinhador deveria inserir buracos apenas no cDNA. Neste quesito, o alinhador global se mostrou ideal, pois não inseriu nenhum buraco no DNA genômico e nenhum alinhamento efetuado com os dados dos conjuntos 1 e 2 (Tabela 3.4 e Tabela 3.6). Em relação ao alinhador semi-global (Tabela 3.5 e Tabela 3.7), os únicos esquemas de pontuação que apresentaram resultados adequados neste critério foram os esquemas $(1, -2, -1, 0)$ e $(1, -2, -10, 0)$.

Analisando os resultados obtidos com estes esquemas de pontuação, observa-se que no critério de número de éxons criados em relação ao número de éxons esperado, o alinhador

semi-global obtém resultados melhores. De fato, no conjunto de dados 1 (Tabela 3.8 e Tabela 3.9), o alinhador global obtém resultados semelhantes em todos os esquemas de pontuação utilizados, com média de éxons extras entre 0.98 e 1.2. O alinhador semi-global apresenta resultados diferenciados para este mesmo conjunto, utilizando os esquemas de pontuação $(1, -2, -1, 0)$ e $(1, -2, -10, 0)$: a média de éxons extras gerados é 1.2, enquanto que para os outros esquemas a média varia de -74 a 2.5. Além disso, a porcentagem de alinhamentos com pontuação 0 neste quesito utilizando os esquemas de pontuação citados acima é superior a 74%, enquanto que nos outros esquemas a porcentagem é inferior a 32%.

Com os dados do conjunto 2 (Tabela 3.10 e Tabela 3.11), o alinhador semi-global com pontuação $(1, -2, -1, 0)$ gera o número correto de éxons em 100% dos alinhamentos, enquanto que o alinhador global ainda gera éxons a mais em alguns casos. O esquema $(1, -2, -10, 0)$ obtém bons resultados, semelhantes aos do esquema anterior para alinhamento global e levemente inferior em dois casos com alinhamento semi-global. A melhora dos resultados do conjunto 2 em relação ao conjunto 1 se deve ao fato de que o primeiro não possui blocos de bases N, que podem levar a erros de alinhamento.

Os resultados com a avaliação de similaridade também mostram que os esquemas de pontuação $(1, -2, -1, 0)$ e $(1, -2, -10, 0)$ apresentam bons resultados. Em relação ao conjunto de dados 1 (Tabela 3.12 e Tabela 3.13), a porcentagem de similaridade é bastante baixa, em grande parte devido à grande quantidade de blocos de N. Porém, tanto o alinhador global quanto o alinhador semi-global apresentam uma similaridade altíssima com estes esquemas de pontuação no conjunto 2 (Tabela 3.14 e Tabela 3.15). Nos outros esquemas testados, em geral o alinhador global se mostra mais adequado.

Observando os resultado obtidos pela avaliação de Score (diferença entre a pontuação esperada do alinhamento ótimo com a pontuação gerada), mostrados nas Tabelas 3.16, 3.17, 3.18 e 3.19, podemos ver que os dois alinhadores obtém resultado semelhantes, e que os esquema de pontuação $(1, -2, -1, 0)$ e $(1, -2, -10, 0)$ produzem alinhamentos mais próximos dos alinhamentos esperados, sendo que nos outros esquemas a variação de pontuação é muito grande.

Levando em conta os resultados obtidos, observou-se que o alinhador semi-global apresenta resultados levemente melhores do que o alinhador global. De fato, espera-se que o alinhador semi-global seja mais adequado, por não penalizar a criação de buracos nas extremidades, favorecendo alinhamentos de grandes seqüências com pequenas seqüências. Como foi observado nos resultados, os esquemas $(1, -2, -1, 0)$ e $(1, -2, -10, 0)$ apresentam resultados muito semelhantes. Por isso definiu-se que seriam efetuados alinhamentos semi-globais com os dois esquemas para comparação com alinhadores externos e para testes com dados contendo erros.

Extra Gap (Conjunto 1 - Alinhador Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-10,-10,-1)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-10,-10,-10)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-10,-100,-1)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-2,-10,0)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-2,-1,0)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(10,-1,-10,-1)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(10,-40,-10,-1)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(5,-16,-12,-4)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
(5,-4,-12,-4)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%

Tabela 3.4: Mínimo, máximo, média e desvio padrão de buracos inseridos na seqüência de DNA genômico pelo alinhador global no conjunto 1, e porcentagem de alinhamentos com nenhum buraco inserido (% OK).

Extra Gap (Conjunto 1 - Alinhador Semi-Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	0	6	1.06	1.36	48.48%
	Gene x mRNA	0	17	1.95	3.26	42.42%
	Gene+200 x CDS	0	6	1.06	1.36	48.48%
	Gene+200 x mRNA	0	17	1.95	3.26	42.42%
(1,-10,-10,-1)	Gene x CDS	0	4	0.88	0.85	34.85%
	Gene x mRNA	0	11	1.05	1.61	34.85%
	Gene+200 x CDS	0	4	0.86	0.84	34.85%
	Gene+200 x mRNA	0	11	1.05	1.61	34.85%
(1,-10,-10,-10)	Gene x CDS	0	10	1.23	1.95	27.27%
	Gene x mRNA	0	11	1.17	1.89	22.73%
	Gene+200 x CDS	0	10	1.23	1.95	27.27%
	Gene+200 x mRNA	0	11	1.17	1.89	22.73%
(1,-10,-100,-1)	Gene x CDS	0	5	0.83	1.13	48.48%
	Gene x mRNA	0	6	0.89	1.29	50.00%
	Gene+200 x CDS	0	1	0.64	0.48	36.36%
	Gene+200 x mRNA	0	1	0.64	0.48	36.36%
(1,-2,-10,0)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-2,-1,0)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(10,-1,-10,-1)	Gene x CDS	0	27	2.44	5.00	51.52%
	Gene x mRNA	0	77	4.12	10.90	53.03%
	Gene+200 x CDS	0	27	2.44	5.00	51.52%
	Gene+200 x mRNA	0	77	4.12	10.90	53.03%
(10,-40,-10,-1)	Gene x CDS	0	196	14.05	28.40	43.94%
	Gene x mRNA	0	221	20.12	35.64	43.94%
	Gene+200 x CDS	0	196	14.05	28.40	43.94%
	Gene+200 x mRNA	0	221	20.12	35.64	43.94%
(5,-16,-12,-4)	Gene x CDS	0	92	11.18	16.84	34.85%
	Gene x CDS	0	92	11.21	16.83	34.85%
	Gene x mRNA	0	144	20.68	29.90	33.33%
	Gene x mRNA	0	144	20.73	29.87	33.33%
(5,-4,-12,-4)	Gene x mRNA	0	89	8.91	14.35	34.85%
	Gene x CDS	0	58	6.26	9.49	34.85%
	Gene+200 x CDS	0	58	6.26	9.49	34.85%
	Gene+200 x mRNA	0	89	8.91	14.35	34.85%

Tabela 3.5: Mínimo, máximo, média e desvio padrão de buracos inseridos na seqüência de DNA genômico pelo alinhador semi-global no conjunto 1, e porcentagem de alinhamentos com nenhum buraco inserido (% OK).

Extra Gap (Conjunto 2 - Alinhador Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-10,-10,-1)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-10,-10,-10)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-10,-100,-1)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-2,-10,0)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-2,-1,0)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(10,-1,-10,-1)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(10,-40,-10,-1)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(5,-16,-12,-4)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
(5,-4,-12,-4)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%

Tabela 3.6: Mínimo, máximo, média e desvio padrão de buracos inseridos na seqüência de DNA genômico pelo alinhador global no conjunto 2, e porcentagem de alinhamentos com nenhum buraco inserido (% OK).

Extra Gap (Conjunto 2 - Alinhador Semi-Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	0	6	1.30	1.45	38.00%
	Gene x mRNA	0	17	2.08	3.37	40.00%
	Gene+200 x CDS	0	6	1.30	1.45	38.00%
	Gene+200 x mRNA	0	17	2.08	3.37	40.00%
(1,-10,-10,-1)	Gene x CDS	0	4	0.98	0.89	30.00%
	Gene x mRNA	0	6	0.96	1.11	32.00%
	Gene+200 x CDS	0	4	0.96	0.88	30.00%
	Gene+200 x mRNA	0	6	0.96	1.11	32.00%
(1,-10,-10,-10)	Gene x CDS	0	10	1.18	1.70	15.00%
	Gene x mRNA	0	10	1.04	1.35	20.00%
	Gene+200 x CDS	0	10	1.18	1.70	15.00%
	Gene+200 x mRNA	0	10	1.04	1.35	20.00%
(1,-10,-100,-1)	Gene x CDS	0	5	0.80	1.05	46.00%
	Gene x mRNA	0	6	0.86	1.25	42.00%
	Gene+200 x CDS	0	1	0.68	0.47	34.00%
	Gene+200 x mRNA	0	1	0.66	0.48	32.00%
(1,-2,-10,0)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-2,-1,0)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(10,-1,-10,-1)	Gene x CDS	0	27	2.22	4.28	54.00%
	Gene x mRNA	0	77	4.00	11.79	50.00%
	Gene+200 x CDS	0	27	2.22	4.28	54.00%
	Gene+200 x mRNA	0	77	4.00	11.79	50.00%
(10,-40,-10,-1)	Gene x CDS	0	196	13.94	29.80	42.00%
	Gene x mRNA	0	221	19.56	36.78	40.00%
	Gene+200 x CDS	0	196	13.94	29.80	42.00%
	Gene+200 x mRNA	0	221	19.56	36.78	40.00%
(5,-16,-12,-4)	Gene x CDS	0	48	11.32	14.24	30.00%
	Gene x CDS	0	48	11.36	14.21	30.00%
	Gene x mRNA	0	144	19.30	27.24	30.00%
	Gene x mRNA	0	144	19.36	27.20	30.00%
(5,-4,-12,-4)	Gene x CDS	0	58	6.72	9.91	32.00%
	Gene x mRNA	0	89	9.26	15.30	30.00%
	Gene+200 x CDS	0	58	6.72	9.91	32.00%
	Gene+200 x mRNA	0	89	9.26	15.30	30.00%

Tabela 3.7: Mínimo, máximo, média e desvio padrão de buracos inseridos na seqüência de DNA genômico pelo alinhador semi-global no conjunto 2, e porcentagem de alinhamentos com nenhum buraco inserido (% OK).

Delta Exon (Conjunto 1 - Alinhador Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	-1	16	1.06	2.96	65.15%
	Gene x mRNA	-1	16	1.20	2.96	75.76%
	Gene+200 x CDS	-1	16	1.00	2.98	62.12%
	Gene+200 x mRNA	-1	16	1.03	3.03	57.58%
(1,-10,-10,-1)	Gene x CDS	-1	16	1.03	2.92	66.67%
	Gene x mRNA	-1	16	1.20	2.96	75.76%
	Gene+200 x CDS	-1	16	0.98	2.95	62.12%
	Gene+200 x mRNA	-1	16	1.02	3.04	59.09%
(1,-10,-10,-10)	Gene x CDS	-1	16	1.03	2.92	66.67%
	Gene x mRNA	-1	16	1.20	2.96	75.76%
	Gene+200 x CDS	-1	16	0.98	2.95	62.12%
	Gene+200 x mRNA	-1	16	1.02	3.04	59.09%
(1,-10,-100,-1)	Gene x CDS	-1	16	1.06	2.96	65.15%
	Gene x mRNA	-1	16	1.20	2.96	75.76%
	Gene+200 x CDS	-1	16	1.00	2.98	62.12%
	Gene+200 x mRNA	-1	16	1.03	3.03	57.58%
(1,-2,-10,0)	Gene x CDS	-1	16	1.02	2.97	60.61%
	Gene x mRNA	-1	16	1.20	2.96	74.24%
	Gene+200 x CDS	-1	16	1.02	2.99	59.09%
	Gene+200 x mRNA	-1	16	1.02	3.06	56.06%
(1,-2,-1,0)	Gene x CDS	-1	16	1.03	2.92	66.67%
	Gene x mRNA	-1	16	1.20	2.96	75.76%
	Gene+200 x CDS	-1	16	0.98	2.95	62.12%
	Gene+200 x mRNA	-1	16	1.02	3.04	59.09%
(10,-1,-10,-1)	Gene x CDS	-1	16	1.03	2.92	66.67%
	Gene x mRNA	-1	16	1.20	2.96	75.76%
	Gene+200 x CDS	-1	16	0.98	2.95	62.12%
	Gene+200 x mRNA	-1	16	1.02	3.04	59.09%
(10,-40,-10,-1)	Gene x CDS	-1	16	1.03	2.92	66.67%
	Gene x mRNA	-1	16	1.20	2.96	75.76%
	Gene+200 x CDS	-1	16	0.98	2.95	62.12%
	Gene+200 x mRNA	-1	16	1.02	3.04	59.09%
(5,-16,-12,-4)	Gene x CDS	-1	16	0.98	2.95	62.12%
	Gene x CDS	-1	16	1.03	2.92	66.67%
	Gene x mRNA	-1	16	1.20	2.96	75.76%
	Gene x mRNA	-1	16	1.02	3.04	59.09%
(5,-4,-12,-4)	Gene x CDS	-1	16	1.05	2.96	66.67%
	Gene x mRNA	-1	16	1.20	2.96	75.76%
	Gene+200 x CDS	-1	16	1.00	2.98	62.12%
	Gene+200 x mRNA	-1	16	1.02	3.04	59.09%

Tabela 3.8: Mínimo, máximo, média e desvio padrão de diferença de éxons criados pelo alinhador global com o número de éxons esperados no conjunto 1, e porcentagem de alinhamentos com delta éxons = 0 (% OK).

Delta Exon (Conjunto 1 - Alinhador Semi-Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	-1	25	3.03	4.54	27.27%
	Gene x mRNA	-4	16	2.74	4.00	27.27%
	Gene+200 x CDS	-1	25	3.03	4.54	27.27%
	Gene+200 x mRNA	-4	16	2.74	4.00	27.27%
(1,-10,-10,-1)	Gene x CDS	-1	25	3.71	4.69	21.21%
	Gene x mRNA	-1	26	4.11	4.82	22.73%
	Gene+200 x CDS	-1	25	3.70	4.69	21.21%
	Gene+200 x mRNA	-1	26	4.09	4.83	22.73%
(1,-10,-10,-10)	Gene x CDS	-3	25	3.85	4.73	18.18%
	Gene x mRNA	-2	26	4.30	4.91	9.09%
	Gene+200 x CDS	-3	25	3.85	4.73	18.18%
	Gene+200 x mRNA	-2	26	4.30	4.91	9.09%
(1,-10,-100,-1)	Gene x CDS	-1	16	2.56	3.25	34.85%
	Gene x mRNA	-2	16	2.85	3.69	36.36%
	Gene+200 x CDS	-1	25	3.88	4.67	22.73%
	Gene+200 x mRNA	-1	26	4.32	4.93	22.73%
(1,-2,-10,0)	Gene x CDS	0	16	1.20	2.90	74.24%
	Gene x mRNA	0	16	1.21	2.95	75.76%
	Gene+200 x CDS	0	16	1.20	2.90	74.24%
	Gene+200 x mRNA	0	16	1.21	2.95	75.76%
(1,-2,-1,0)	Gene x CDS	0	16	1.15	2.86	78.79%
	Gene x mRNA	0	16	1.21	2.95	77.27%
	Gene+200 x CDS	0	16	1.15	2.86	78.79%
	Gene+200 x mRNA	0	16	1.21	2.95	77.27%
(10,-1,-10,-1)	Gene x CDS	-365	16	-28.45	55.15	25.76%
	Gene x mRNA	-408	16	-39.89	64.23	30.30%
	Gene+200 x CDS	-365	16	-28.45	55.15	25.76%
	Gene+200 x mRNA	-408	16	-39.89	64.23	30.30%
(10,-40,-10,-1)	Gene x CDS	-620	16	-49.61	94.97	28.79%
	Gene x mRNA	-690	16	-74.89	126.06	31.82%
	Gene+200 x CDS	-620	16	-49.61	94.97	28.79%
	Gene+200 x mRNA	-690	16	-74.89	126.06	31.82%
(5,-16,-12,-4)	Gene x CDS	-22	25	-0.74	7.73	24.24%
	Gene x CDS	-22	25	-0.79	7.69	24.24%
	Gene x mRNA	-87	26	-5.29	15.70	22.73%
	Gene x mRNA	-87	26	-5.32	15.68	22.73%
(5,-4,-12,-4)	Gene x CDS	-65	16	-3.33	10.51	28.79%
	Gene x mRNA	-50	16	-4.45	9.50	21.21%
	Gene+200 x CDS	-65	16	-3.33	10.51	28.79%
	Gene+200 x mRNA	-50	16	-4.45	9.50	21.21%

Tabela 3.9: Mínimo, máximo, média e desvio padrão de diferença de éxons criados pelo alinhador semi-global com o número de éxons esperados no conjunto 1, e porcentagem de alinhamentos com delta éxons = 0 (% OK).

Delta Exon (Conjunto 2 - Alinhador Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	-1	1	-0.14	0.40	82.00%
	Gene x mRNA	-1	0	-0.02	0.14	98.00%
	Gene+200 x CDS	-1	0	-0.22	0.42	78.00%
	Gene+200 x mRNA	-1	1	-0.22	0.46	74.00%
(1,-10,-10,-1)	Gene x CDS	-1	0	-0.16	0.37	84.00%
	Gene x mRNA	-1	0	-0.02	0.14	98.00%
	Gene+200 x CDS	-1	0	-0.22	0.42	78.00%
	Gene+200 x mRNA	-1	0	-0.24	0.43	76.00%
(1,-10,-10,-10)	Gene x CDS	-1	0	-0.16	0.37	84.00%
	Gene x mRNA	-1	0	-0.02	0.14	98.00%
	Gene+200 x CDS	-1	0	-0.22	0.42	78.00%
	Gene+200 x mRNA	-1	0	-0.24	0.43	76.00%
(1,-10,-100,-1)	Gene x CDS	-1	1	-0.14	0.40	82.00%
	Gene x mRNA	-1	0	-0.02	0.14	98.00%
	Gene+200 x CDS	-1	0	-0.22	0.42	78.00%
	Gene+200 x mRNA	-1	1	-0.22	0.46	74.00%
(1,-2,-10,0)	Gene x CDS	-1	0	-0.20	0.40	80.00%
	Gene x mRNA	-1	0	-0.02	0.14	98.00%
	Gene+200 x CDS	-1	0	-0.22	0.42	78.00%
	Gene+200 x mRNA	-1	0	-0.26	0.44	74.00%
(1,-2,-1,0)	Gene x CDS	-1	0	-0.16	0.37	84.00%
	Gene x mRNA	-1	0	-0.02	0.14	98.00%
	Gene+200 x CDS	-1	0	-0.22	0.42	78.00%
	Gene+200 x mRNA	-1	0	-0.24	0.43	76.00%
(10,-1,-10,-1)	Gene x CDS	-1	0	-0.16	0.37	84.00%
	Gene x mRNA	-1	0	-0.02	0.14	98.00%
	Gene+200 x CDS	-1	0	-0.22	0.42	78.00%
	Gene+200 x mRNA	-1	0	-0.24	0.43	76.00%
(10,-40,-10,-1)	Gene x CDS	-1	0	-0.16	0.37	84.00%
	Gene x mRNA	-1	0	-0.02	0.14	98.00%
	Gene+200 x CDS	-1	0	-0.22	0.42	78.00%
	Gene+200 x mRNA	-1	0	-0.24	0.43	76.00%
(5,-16,-12,-4)	Gene x CDS	-1	0	-0.16	0.37	84.00%
	Gene x CDS	-1	0	-0.22	0.42	78.00%
	Gene x mRNA	-1	0	-0.24	0.43	76.00%
	Gene x mRNA	-1	0	-0.02	0.14	98.00%
(5,-4,-12,-4)	Gene x CDS	-1	0	-0.16	0.37	84.00%
	Gene x mRNA	-1	0	-0.02	0.14	98.00%
	Gene+200 x CDS	-1	0	-0.22	0.42	78.00%
	Gene+200 x mRNA	-1	0	-0.24	0.43	76.00%

Tabela 3.10: Mínimo, máximo, média e desvio padrão de diferença de éxons criados pelo alinhador global com o número de éxons esperados no conjunto 2, e porcentagem de alinhamentos com delta éxons = 0 (% OK).

Delta Exon (Conjunto 2 - Alinhador Semi-Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	-1	25	1.96	3.83	36.00%
	Gene x mRNA	-4	13	1.46	2.70	36.00%
	Gene+200 x CDS	-1	25	1.96	3.83	36.00%
	Gene+200 x mRNA	-4	13	1.46	2.70	36.00%
(1,-10,-10,-1)	Gene x CDS	-1	25	2.62	3.97	28.00%
	Gene x mRNA	-1	26	2.94	4.15	30.00%
	Gene+200 x CDS	-1	25	2.60	3.97	28.00%
	Gene+200 x mRNA	-1	26	2.92	4.15	30.00%
(1,-10,-10,-10)	Gene x CDS	-3	25	2.80	4.09	24.00%
	Gene x mRNA	-2	26	3.20	4.19	12.00%
	Gene+200 x CDS	-3	25	2.80	4.09	24.00%
	Gene+200 x mRNA	-2	26	3.20	4.19	12.00%
(1,-10,-100,-1)	Gene x CDS	-1	12	1.56	2.38	46.00%
	Gene x mRNA	-2	13	1.58	2.59	48.00%
	Gene+200 x CDS	-1	25	2.84	4.00	30.00%
	Gene+200 x mRNA	-1	26	3.14	4.18	30.00%
(1,-2,-10,0)	Gene x CDS	0	1	0.02	0.14	98.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	1	0.02	0.14	98.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-2,-1,0)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(10,-1,-10,-1)	Gene x CDS	-365	0	-31.44	58.37	34.00%
	Gene x mRNA	-408	1	-39.96	65.71	40.00%
	Gene+200 x CDS	-365	0	-31.44	58.37	34.00%
	Gene+200 x mRNA	-408	1	-39.96	65.71	40.00%
(10,-40,-10,-1)	Gene x CDS	-620	0	-51.62	97.94	38.00%
	Gene x mRNA	-690	1	-74.68	131.13	42.00%
	Gene+200 x CDS	-620	0	-51.62	97.94	38.00%
	Gene+200 x mRNA	-690	1	-74.68	131.13	42.00%
(5,-16,-12,-4)	Gene x CDS	-22	25	-2.04	7.48	32.00%
	Gene x CDS	-22	25	-2.10	7.41	32.00%
	Gene x mRNA	-87	26	-5.90	15.26	28.00%
	Gene x mRNA	-87	26	-5.94	15.23	28.00%
(5,-4,-12,-4)	Gene x CDS	-65	3	-4.92	10.97	34.00%
	Gene x mRNA	-50	1	-6.42	9.36	26.00%
	Gene+200 x CDS	-65	3	-4.92	10.97	34.00%
	Gene+200 x mRNA	-50	1	-6.42	9.36	26.00%

Tabela 3.11: Mínimo, máximo, média e desvio padrão de diferença de éxons criados pelo alinhador semi-global com o número de éxons esperados no conjunto 2, e porcentagem de alinhamentos com delta éxons = 0 (% OK).

Similaridade (Conjunto 1 - Alinhador Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	0.00%	100.00%	87.56%	25.91%	34.85%
	Gene x mRNA	22.78%	100.00%	90.56%	19.93%	36.36%
	Gene+200 x CDS	0.00%	100.00%	86.99%	26.82%	30.30%
	Gene+200 x mRNA	22.78%	100.00%	89.72%	21.43%	25.76%
(1,-10,-10,-1)	Gene x CDS	0.00%	100.00%	87.57%	25.92%	34.85%
	Gene x mRNA	22.78%	100.00%	90.56%	19.93%	36.36%
	Gene+200 x CDS	0.00%	100.00%	87.00%	26.83%	30.30%
	Gene+200 x mRNA	22.78%	100.00%	89.76%	21.44%	25.76%
(1,-10,-10,-10)	Gene x CDS	0.00%	100.00%	87.57%	25.92%	34.85%
	Gene x mRNA	22.78%	100.00%	90.56%	19.93%	36.36%
	Gene+200 x CDS	0.00%	100.00%	87.00%	26.83%	30.30%
	Gene+200 x mRNA	22.78%	100.00%	89.76%	21.44%	25.76%
(1,-10,-100,-1)	Gene x CDS	0.00%	100.00%	87.56%	25.91%	34.85%
	Gene x mRNA	22.78%	100.00%	90.56%	19.93%	36.36%
	Gene+200 x CDS	0.00%	100.00%	86.99%	26.82%	30.30%
	Gene+200 x mRNA	22.78%	100.00%	89.72%	21.43%	25.76%
(1,-2,-10,0)	Gene x CDS	0.00%	100.00%	86.16%	28.63%	28.79%
	Gene x mRNA	22.78%	100.00%	89.47%	22.48%	36.36%
	Gene+200 x CDS	0.00%	100.00%	85.93%	28.92%	30.30%
	Gene+200 x mRNA	22.78%	100.00%	89.30%	22.58%	25.76%
(1,-2,-1,0)	Gene x CDS	0.00%	100.00%	87.57%	25.92%	34.85%
	Gene x mRNA	22.78%	100.00%	90.56%	19.93%	36.36%
	Gene+200 x CDS	0.00%	100.00%	87.00%	26.83%	30.30%
	Gene+200 x mRNA	22.78%	100.00%	89.76%	21.44%	25.76%
(10,-1,-10,-1)	Gene x CDS	0.00%	100.00%	87.57%	25.92%	34.85%
	Gene x mRNA	22.78%	100.00%	90.56%	19.93%	36.36%
	Gene+200 x CDS	0.00%	100.00%	87.00%	26.83%	30.30%
	Gene+200 x mRNA	22.78%	100.00%	89.76%	21.44%	25.76%
(10,-40,-10,-1)	Gene x CDS	0.00%	100.00%	87.57%	25.92%	34.85%
	Gene x mRNA	22.78%	100.00%	90.56%	19.93%	36.36%
	Gene+200 x CDS	0.00%	100.00%	87.00%	26.83%	30.30%
	Gene+200 x mRNA	22.78%	100.00%	89.76%	21.44%	25.76%
(5,-16,-12,-4)	Gene x CDS	0.00%	100.00%	87.00%	26.83%	30.30%
	Gene x CDS	0.00%	100.00%	87.57%	25.92%	34.85%
	Gene x mRNA	22.78%	100.00%	89.76%	21.44%	25.76%
	Gene x mRNA	22.78%	100.00%	90.56%	19.93%	36.36%
(5,-4,-12,-4)	Gene x CDS	0.00%	100.00%	87.57%	25.92%	34.85%
	Gene x mRNA	22.78%	100.00%	90.56%	19.93%	36.36%
	Gene+200 x CDS	0.00%	100.00%	87.00%	26.83%	30.30%
	Gene+200 x mRNA	22.78%	100.00%	89.76%	21.44%	25.76%

Tabela 3.12: Mínimo, máximo, média e desvio padrão de porcentagem de similaridade das bases dos éxons criados pelo alinhador global no conjunto 1, e porcentagem de alinhamentos com 100% de similaridade (% OK).

Similaridade (Conjunto 1 - Alinhador Semi-Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	0.00%	100.00%	54.17%	34.96%	21.21%
	Gene x mRNA	0.00%	100.00%	54.81%	33.85%	21.21%
	Gene+200 x CDS	0.00%	100.00%	54.07%	35.01%	21.21%
	Gene+200 x mRNA	0.00%	100.00%	54.72%	33.89%	21.21%
(1,-10,-10,-1)	Gene x CDS	0.00%	100.00%	46.16%	41.37%	21.21%
	Gene x mRNA	0.00%	100.00%	44.24%	42.64%	22.73%
	Gene+200 x CDS	0.00%	100.00%	45.99%	41.45%	21.21%
	Gene+200 x mRNA	0.00%	100.00%	44.10%	42.69%	22.73%
(1,-10,-10,-10)	Gene x CDS	0.00%	100.00%	24.02%	39.78%	15.15%
	Gene x mRNA	0.00%	100.00%	19.18%	35.76%	9.09%
	Gene+200 x CDS	0.00%	100.00%	23.91%	39.75%	15.15%
	Gene+200 x mRNA	0.00%	100.00%	19.01%	35.72%	9.09%
(1,-10,-100,-1)	Gene x CDS	0.00%	100.00%	65.64%	32.87%	27.27%
	Gene x mRNA	0.00%	100.00%	65.25%	33.85%	25.76%
	Gene+200 x CDS	0.00%	100.00%	38.79%	43.76%	22.73%
	Gene+200 x mRNA	0.00%	100.00%	39.66%	43.73%	22.73%
(1,-2,-10,0)	Gene x CDS	0.00%	100.00%	87.15%	27.45%	39.39%
	Gene x mRNA	3.95%	100.00%	86.39%	27.34%	37.88%
	Gene+200 x CDS	0.00%	100.00%	86.73%	28.14%	39.39%
	Gene+200 x mRNA	3.95%	100.00%	85.68%	28.41%	37.88%
(1,-2,-1,0)	Gene x CDS	0.00%	100.00%	85.79%	30.03%	40.91%
	Gene x mRNA	3.95%	100.00%	85.31%	29.10%	37.88%
	Gene+200 x CDS	0.00%	100.00%	85.70%	30.15%	40.91%
	Gene+200 x mRNA	3.95%	100.00%	85.22%	29.22%	37.88%
(10,-1,-10,-1)	Gene x CDS	0.00%	100.00%	61.18%	33.78%	24.24%
	Gene x mRNA	0.00%	100.00%	62.11%	34.04%	25.76%
	Gene+200 x CDS	0.00%	100.00%	61.10%	33.83%	24.24%
	Gene+200 x mRNA	0.00%	100.00%	62.02%	34.09%	25.76%
(10,-40,-10,-1)	Gene x CDS	0.00%	100.00%	63.03%	33.88%	25.76%
	Gene x mRNA	0.00%	100.00%	63.33%	34.11%	25.76%
	Gene+200 x CDS	0.00%	100.00%	62.95%	33.93%	25.76%
	Gene+200 x mRNA	0.00%	100.00%	63.25%	34.16%	25.76%
(5,-16,-12,-4)	Gene x CDS	0.00%	100.00%	53.63%	37.00%	21.21%
	Gene x CDS	0.00%	100.00%	53.74%	36.95%	21.21%
	Gene x mRNA	0.00%	100.00%	52.75%	36.69%	21.21%
	Gene x mRNA	0.00%	100.00%	52.84%	36.66%	21.21%
(5,-4,-12,-4)	Gene x CDS	0.00%	100.00%	55.96%	33.36%	21.21%
	Gene x mRNA	0.00%	100.00%	55.35%	32.72%	19.70%
	Gene+200 x CDS	0.00%	100.00%	55.88%	33.39%	21.21%
	Gene+200 x mRNA	0.00%	100.00%	55.26%	32.74%	19.70%

Tabela 3.13: Mínimo, máximo, média e desvio padrão de porcentagem de similaridade das bases dos éxons criados pelo alinhador semi-global no conjunto 1, e porcentagem de alinhamentos com 100% de similaridade (% OK).

Similaridade (Conjunto 2 - Alinhador Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	98.39%	100.00%	99.76%	0.40%	44.00%
	Gene x mRNA	99.18%	100.00%	99.86%	0.19%	48.00%
	Gene+200 x CDS	98.27%	100.00%	99.75%	0.40%	40.00%
	Gene+200 x mRNA	97.07%	100.00%	99.77%	0.44%	34.00%
(1,-10,-10,-1)	Gene x CDS	98.45%	100.00%	99.78%	0.35%	44.00%
	Gene x mRNA	99.18%	100.00%	99.86%	0.19%	48.00%
	Gene+200 x CDS	98.27%	100.00%	99.77%	0.35%	40.00%
	Gene+200 x mRNA	99.02%	100.00%	99.82%	0.20%	34.00%
(1,-10,-10,-10)	Gene x CDS	98.45%	100.00%	99.78%	0.35%	44.00%
	Gene x mRNA	99.18%	100.00%	99.86%	0.19%	48.00%
	Gene+200 x CDS	98.27%	100.00%	99.77%	0.35%	40.00%
	Gene+200 x mRNA	99.02%	100.00%	99.82%	0.20%	34.00%
(1,-10,-100,-1)	Gene x CDS	98.39%	100.00%	99.76%	0.40%	44.00%
	Gene x mRNA	99.18%	100.00%	99.86%	0.19%	48.00%
	Gene+200 x CDS	98.27%	100.00%	99.75%	0.40%	40.00%
	Gene+200 x mRNA	97.07%	100.00%	99.77%	0.44%	34.00%
(1,-2,-10,0)	Gene x CDS	98.39%	100.00%	99.74%	0.37%	36.00%
	Gene x mRNA	98.89%	100.00%	99.86%	0.22%	48.00%
	Gene+200 x CDS	98.39%	100.00%	99.74%	0.39%	40.00%
	Gene+200 x mRNA	98.77%	100.00%	99.82%	0.24%	34.00%
(1,-2,-1,0)	Gene x CDS	98.45%	100.00%	99.78%	0.35%	44.00%
	Gene x mRNA	99.18%	100.00%	99.86%	0.19%	48.00%
	Gene+200 x CDS	98.27%	100.00%	99.77%	0.35%	40.00%
	Gene+200 x mRNA	99.02%	100.00%	99.82%	0.20%	34.00%
(10,-1,-10,-1)	Gene x CDS	98.45%	100.00%	99.78%	0.35%	44.00%
	Gene x mRNA	99.18%	100.00%	99.86%	0.19%	48.00%
	Gene+200 x CDS	98.27%	100.00%	99.77%	0.35%	40.00%
	Gene+200 x mRNA	99.02%	100.00%	99.82%	0.20%	34.00%
(10,-40,-10,-1)	Gene x CDS	98.45%	100.00%	99.78%	0.35%	44.00%
	Gene x mRNA	99.18%	100.00%	99.86%	0.19%	48.00%
	Gene+200 x CDS	98.27%	100.00%	99.77%	0.35%	40.00%
	Gene+200 x mRNA	99.02%	100.00%	99.82%	0.20%	34.00%
(5,-16,-12,-4)	Gene x CDS	98.27%	100.00%	99.77%	0.35%	40.00%
	Gene x CDS	98.45%	100.00%	99.78%	0.35%	44.00%
	Gene x mRNA	99.02%	100.00%	99.82%	0.20%	34.00%
	Gene x mRNA	99.18%	100.00%	99.86%	0.19%	48.00%
(5,-4,-12,-4)	Gene x CDS	98.45%	100.00%	99.78%	0.35%	44.00%
	Gene x mRNA	99.18%	100.00%	99.86%	0.19%	48.00%
	Gene+200 x CDS	98.27%	100.00%	99.77%	0.35%	40.00%
	Gene+200 x mRNA	99.02%	100.00%	99.82%	0.20%	34.00%

Tabela 3.14: Mínimo, máximo, média e desvio padrão de porcentagem de similaridade das bases dos éxons criados pelo alinhador global no conjunto 2, e porcentagem de alinhamentos com 100% de similaridade (% OK).

Similaridade (Conjunto 2 - Alinhador Semi-Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	0.00%	100.00%	65.11%	31.89%	28.00%
	Gene x mRNA	0.00%	100.00%	66.05%	29.83%	28.00%
	Gene+200 x CDS	0.00%	100.00%	65.09%	31.93%	28.00%
	Gene+200 x mRNA	0.00%	100.00%	66.03%	29.86%	28.00%
(1,-10,-10,-1)	Gene x CDS	0.00%	100.00%	55.37%	42.39%	28.00%
	Gene x mRNA	0.00%	100.00%	53.03%	44.50%	30.00%
	Gene+200 x CDS	0.00%	100.00%	55.36%	42.40%	28.00%
	Gene+200 x mRNA	0.00%	100.00%	53.02%	44.52%	30.00%
(1,-10,-10,-10)	Gene x CDS	0.00%	100.00%	28.50%	44.19%	20.00%
	Gene x mRNA	0.00%	100.00%	21.31%	39.93%	12.00%
	Gene+200 x CDS	0.00%	100.00%	28.47%	44.21%	20.00%
	Gene+200 x mRNA	0.00%	100.00%	21.21%	39.96%	12.00%
(1,-10,-100,-1)	Gene x CDS	2.05%	100.00%	77.73%	25.96%	36.00%
	Gene x mRNA	2.05%	100.00%	79.01%	25.06%	34.00%
	Gene+200 x CDS	0.00%	100.00%	48.01%	46.16%	30.00%
	Gene+200 x mRNA	0.00%	100.00%	48.27%	46.41%	30.00%
(1,-2,-10,0)	Gene x CDS	98.39%	100.00%	99.79%	0.36%	52.00%
	Gene x mRNA	98.89%	100.00%	99.86%	0.22%	50.00%
	Gene+200 x CDS	98.39%	100.00%	99.79%	0.36%	52.00%
	Gene+200 x mRNA	98.89%	100.00%	99.86%	0.22%	50.00%
(1,-2,-1,0)	Gene x CDS	98.39%	100.00%	99.81%	0.35%	54.00%
	Gene x mRNA	99.18%	100.00%	99.86%	0.19%	50.00%
	Gene+200 x CDS	98.39%	100.00%	99.81%	0.35%	54.00%
	Gene+200 x mRNA	99.18%	100.00%	99.86%	0.19%	50.00%
(10,-1,-10,-1)	Gene x CDS	2.05%	100.00%	73.62%	27.10%	32.00%
	Gene x mRNA	2.05%	100.00%	74.95%	27.28%	34.00%
	Gene+200 x CDS	2.05%	100.00%	73.62%	27.10%	32.00%
	Gene+200 x mRNA	2.05%	100.00%	74.95%	27.28%	34.00%
(10,-40,-10,-1)	Gene x CDS	2.19%	100.00%	76.04%	26.16%	34.00%
	Gene x mRNA	2.19%	100.00%	76.55%	26.63%	34.00%
	Gene+200 x CDS	2.19%	100.00%	76.04%	26.16%	34.00%
	Gene+200 x mRNA	2.19%	100.00%	76.55%	26.63%	34.00%
(5,-16,-12,-4)	Gene x CDS	0.00%	100.00%	64.67%	34.36%	28.00%
	Gene x CDS	0.00%	100.00%	64.70%	34.31%	28.00%
	Gene x mRNA	0.00%	100.00%	63.97%	34.02%	28.00%
	Gene x mRNA	0.00%	100.00%	63.98%	34.00%	28.00%
(5,-4,-12,-4)	Gene x CDS	0.00%	100.00%	67.50%	28.52%	28.00%
	Gene x mRNA	0.00%	100.00%	66.75%	27.81%	26.00%
	Gene+200 x CDS	0.00%	100.00%	67.50%	28.52%	28.00%
	Gene+200 x mRNA	0.00%	100.00%	66.75%	27.81%	26.00%

Tabela 3.15: Mínimo, máximo, média e desvio padrão de porcentagem de similaridade das bases dos éxons criados pelo alinhador semi-global no conjunto 2, e porcentagem de alinhamentos com 100% de similaridade (% OK).

Score (Conjunto 1 - Alinhador Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	-170	0	-12.91	31.03	75.76%
	Gene x mRNA	-150	0	-11.06	27.52	77.27%
	Gene+200 x CDS	-170	0	-14.12	32.12	74.24%
	Gene+200 x mRNA	-170	0	-14.55	32.78	75.76%
(1,-10,-10,-1)	Gene x CDS	-170	0	-12.58	30.70	78.79%
	Gene x mRNA	-150	0	-11.06	27.52	77.27%
	Gene+200 x CDS	-170	0	-13.79	31.85	77.27%
	Gene+200 x mRNA	-170	0	-14.55	32.78	75.76%
(1,-10,-10,-10)	Gene x CDS	-170	0	-12.58	30.70	78.79%
	Gene x mRNA	-150	0	-11.06	27.52	77.27%
	Gene+200 x CDS	-170	0	-13.79	31.85	77.27%
	Gene+200 x mRNA	-170	0	-14.55	32.78	75.76%
(1,-10,-100,-1)	Gene x CDS	-1700	0	-130.32	310.75	74.24%
	Gene x mRNA	-1500	0	-110.61	275.19	74.24%
	Gene+200 x CDS	-1700	0	-142.44	321.77	77.27%
	Gene+200 x mRNA	-1700	0	-146.14	327.52	75.76%
(1,-2,-10,0)	Gene x CDS	-170	0	-12.77	30.99	75.76%
	Gene x mRNA	-150	0	-11.06	27.52	77.27%
	Gene+200 x CDS	-170	0	-13.98	32.07	77.24%
	Gene+200 x mRNA	-170	0	-14.55	32.78	75.76%
(1,-2,-1,0)	Gene x CDS	-17	0	-1.26	3.07	78.79%
	Gene x mRNA	-15	0	-1.11	2.75	77.27%
	Gene+200 x CDS	-17	0	-1.38	3.19	77.27%
	Gene+200 x mRNA	-17	0	-1.45	3.28	75.76%
(10,-1,-10,-1)	Gene x CDS	-170	0	-12.58	30.7	78.79%
	Gene x mRNA	-150	0	-11.06	27.52	77.27%
	Gene+200 x CDS	-170	0	-13.79	31.85	77.27%
	Gene+200 x mRNA	-170	0	-14.55	32.78	75.76%
(10,-40,-10,-1)	Gene x CDS	-170	0	-12.58	30.7	78.79%
	Gene x mRNA	-150	0	-11.06	27.52	77.27%
	Gene+200 x CDS	-170	0	-13.79	31.85	77.27%
	Gene+200 x mRNA	-170	0	-14.55	32.78	75.76%
(5,-16,-12,-4)	Gene x CDS	-204	0	-16.55	38.22	77.27%
	Gene x CDS	-204	0	-15.09	36.84	78.79%
	Gene x mRNA	-204	0	-17.45	39.33	75.76%
	Gene x mRNA	-180	0	-13.27	33.02	77.27%
(5,-4,-12,-4)	Gene x CDS	-204	0	-15.14	36.94	78.79%
	Gene x mRNA	-180	0	-13.27	33.02	77.27%
	Gene+200 x CDS	-204	0	-16.59	38.32	77.27%
	Gene+200 x mRNA	-204	0	-17.45	39.33	75.76%

Tabela 3.16: Mínimo, máximo, média e desvio padrão de diferença de pontuação do alinhamento gerado pelo alinhador global com o pontuação esperado no conjunto 1, e porcentagem de alinhamentos com delta score = 0 (% OK).

Score (Conjunto 1 - Alinhador Semi-Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	-63208	0	-9959.76	14458.22	21.21%
	Gene x mRNA	-66948	0	-11937.02	16196.47	21.21%
	Gene+200 x CDS	-63208	0	-9959.67	14458.18	21.21%
	Gene+200 x mRNA	-66948	0	-11937.05	16196.49	21.21%
(1,-10,-10,-1)	Gene x CDS	-63029	0	-9890.32	14447.12	21.21%
	Gene x mRNA	-66569	0	-11805.95	16096.44	22.73%
	Gene+200 x CDS	-63029	0	-9890.21	14446.94	21.21%
	Gene+200 x mRNA	-66569	0	-11806.08	16096.36	22.73%
(1,-10,-10,-10)	Gene x CDS	-641729	0	-102774.45	147308.67	15.15%
	Gene x mRNA	-680936	0	-123806.09	165248.31	9.09%
	Gene+200 x CDS	-641729	0	-102774.17	147308.74	15.15%
	Gene+200 x mRNA	-680936	0	-123805.94	165248.49	9.09%
(1,-10,-100,-1)	Gene x CDS	-62745	0	-9910.17	14411.75	22.73%
	Gene x mRNA	-65609	0	-11869.23	16085.76	22.73%
	Gene+200 x CDS	-64739	0	-10213.48	14726.59	21.21%
	Gene+200 x mRNA	-68369	0	-12168.65	16383.15	22.73%
(1,-2,-10,0)	Gene x CDS	-160	0	-11.65	28.80	77.27%
	Gene x mRNA	-160	0	-12.12	29.48	77.27%
	Gene+200 x CDS	-160	0	-11.65	28.80	77.27%
	Gene+200 x mRNA	-160	0	-12.12	29.48	77.27%
(1,-2,-1,0)	Gene x CDS	-16	0	-1.15	2.86	78.79%
	Gene x mRNA	-16	0	-1.21	2.95	77.27%
	Gene+200 x CDS	-16	0	-1.15	2.86	78.79%
	Gene+200 x mRNA	-16	0	-1.21	2.95	77.27%
(10,-1,-10,-1)	Gene x CDS	-59366	0	-9032.02	13523.29	25.76%
	Gene x mRNA	-62265	0	-10636.59	14926.1	30.30%
	Gene+200 x CDS	-59366	0	-9032.02	13523.29	25.76%
	Gene+200 x mRNA	-62265	0	-10636.59	14926.1	30.30%
(10,-40,-10,-1)	Gene x CDS	-57954	0	-8761.11	13256.91	28.79%
	Gene x mRNA	-60516	0	-10241.06	14542.82	31.82%
	Gene+200 x CDS	-57954	0	-8761.11	13256.91	28.79%
	Gene+200 x mRNA	-60516	0	-10241.06	14542.82	31.82%
(5,-16,-12,-4)	Gene x CDS	-250264	0	-39274.12	57333.95	21.21%
	Gene x CDS	-250264	0	-39274.48	57334.23	21.21%
	Gene x mRNA	-264755	0	-46936.95	64069.89	21.21%
	Gene x mRNA	-264755	0	-46937.02	64069.86	21.21%
(5,-4,-12,-4)	Gene x CDS	-252313	0	-39705.91	57702.81	21.21%
	Gene x mRNA	-267282	0	-47640.45	64694.66	19.70%
	Gene+200 x CDS	-252313	0	-39705.91	57702.81	21.21%
	Gene+200 x mRNA	-267282	0	-47640.45	64694.66	19.70%

Tabela 3.17: Mínimo, máximo, média e desvio padrão de diferença de score do alinhamento gerado pelo alinhador semi-global com o score esperado no conjunto 1, e porcentagem de alinhamentos com delta score = 0 (% OK).

Score (Conjunto 2 - Alinhador Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	-4	0	-0.16	0.79	96.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	-6	0	-0.20	1.01	96.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-10,-10,-1)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-10,-10,-10)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-10,-100,-1)	Gene x CDS	-67	0	-2.68	13.26	96.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	-78	0	-2.90	14.39	96.00%
	Gene+200 x mRNA	-45	0	-0.90	6.36	98.00%
(1,-2,-10,0)	Gene x CDS	-1	0	-0.04	0.20	96.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	-4	0	-0.10	0.58	96.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-2,-1,0)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(10,-1,-10,-1)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(10,-40,-10,-1)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(5,-16,-12,-4)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
(5,-4,-12,-4)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%

Tabela 3.18: Mínimo, máximo, média e desvio padrão de diferença de score do alinhamento gerado pelo alinhador global com o score esperado no conjunto 2, e porcentagem de alinhamentos com delta score = 0 (% OK).

Score (Conjunto 2 - Alinhador Semi-Global)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-1,-10,-1)	Gene x CDS	-41615	0	-7729.20	11860.84	28.00%
	Gene x mRNA	-42099	0	-9107.08	12976.88	28.00%
	Gene+200 x CDS	-41615	0	-7729.08	11860.75	28.00%
	Gene+200 x mRNA	-42099	0	-9107.12	12976.91	28.00%
(1,-10,-10,-1)	Gene x CDS	-41384	0	-7653.40	11853.09	28.00%
	Gene x mRNA	-41361	0	-8975.12	12893.41	30.00%
	Gene+200 x CDS	-41383	0	-7653.26	11852.77	28.00%
	Gene+200 x mRNA	-41361	0	-8975.28	12893.30	30.00%
(1,-10,-10,-10)	Gene x CDS	-421821	0	-80817.72	121789.06	20.00%
	Gene x mRNA	-439239	0	-95882.02	133556.55	12.00%
	Gene+200 x CDS	-421819	0	-80817.34	121789.11	20.00%
	Gene+200 x mRNA	-439239	0	-95881.82	133556.8	12.00%
(1,-10,-100,-1)	Gene x CDS	-41526	0	-7615.26	11972.26	40.00%
	Gene x mRNA	-41601	0	-8980.48	13055.00	42.00%
	Gene+200 x CDS	-41743	0	-7880.28	12049.00	30.00%
	Gene+200 x mRNA	-41721	0	-9233.20	13097.07	30.00%
(1,-2,-10,0)	Gene x CDS	-1	0	-0.02	0.14	98.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	-1	0	-0.02	0.14	98.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(1,-2,-1,0)	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
(10,-1,-10,-1)	Gene x CDS	-40666	0	-6788.04	10904.90	34.00%
	Gene x mRNA	-39949	0	-7838.72	11700.37	40.00%
	Gene+200 x CDS	-40666	0	-6788.04	10904.90	34.00%
	Gene+200 x mRNA	-39949	0	-7838.72	11700.37	40.00%
(10,-40,-10,-1)	Gene x CDS	-40551	0	-6519.40	10677.07	38.00%
	Gene x mRNA	-39355	0	-7457.12	11343.72	42.00%
	Gene+200 x CDS	-40551	0	-6519.40	10677.07	38.00%
	Gene+200 x mRNA	-39355	0	-7457.12	11343.72	42.00%
(5,-16,-12,-4)	Gene x CDS	-165422	0	-30363.06	46950.39	28.00%
	Gene x CDS	-165422	0	-30363.54	46950.92	28.00%
	Gene x mRNA	-165254	0	-35648.48	51202.83	26.00%
	Gene x mRNA	-165254	0	-35648.56	51202.78	26.00%
(5,-4,-12,-4)	Gene x CDS	-166371	0	-30815.64	47325.92	28.00%
	Gene x mRNA	-168365	0	-36366.72	51857.04	28.00%
	Gene+200 x CDS	-166371	0	-30815.64	47325.92	28.00%
	Gene+200 x mRNA	-168365	0	-36366.72	51857.04	28.00%

Tabela 3.19: Mínimo, máximo, média e desvio padrão de diferença de score do alinhamento gerado pelo alinhador semi-global com o score esperado no conjunto 2, e porcentagem de alinhamentos com delta score = 0 (% OK).

3.7.2 Análise dos alinhamentos exatos

O alinhador semi-global utilizado tem como única tarefa encontrar o melhor alinhamento possível entre duas seqüências, não levando em conta a existência de regiões de éxons separadas por grandes regiões de íntrons. Neste sentido pode-se observar que os resultados obtidos com as pontuações escolhidas são muito satisfatórios para alinhamentos de ESTs com DNA genômico sem erros.

Pela Tabela 3.20, pode-se observar que em 100% dos alinhamentos exatos gerados, o alinhador semi-global não inseriu nenhum espaço na seqüência genômica. O pacote `Spidey` não insere buracos em mais de 97% dos alinhamentos gerados, tendo uma média inferior a 0.2 buracos por alinhamento. O pacote `est_genome` obtém os piores resultados, com média de buracos inseridos superior a 14 por alinhamento.

A diferença entre o número de éxons esperados e o número de éxons gerados, mostrada na Tabela 3.21, é na grande maioria igual a 0, a menos para o pacote `Spidey`, que não obtém nenhum alinhamento com exatamente o número esperado de éxons. O pacote `sim4` obteve os melhores resultados, com mais de 94% dos alinhamentos com o mesmo número de éxons determinados. Os resultados obtidos com o alinhador semi-global são muito bons, sobretudo levando-se em conta que neste caso consideramos que um éxon é uma região de bases contíguas, considerando um simples buraco como sendo um íntron.

Além disso, pode-se ver na Tabela 3.22 que o grau de similaridade dos éxons das seqüências é bastante alto para o alinhador semi-global (acima de 99.8% com desvio padrão inferior a 0.65%). Esses resultados são compatíveis com os resultados obtidos pelos alinhadores externos, e em alguns casos, melhores. O pacote `sim4` obteve similaridade superior a 99%, porém o pacote `Spidey` obteve em média 81% de similaridade, e `est_genome` não passou dos 62%.

Finalmente, pode-se observar na Tabela 3.23 que o alinhador semi-global com esquema $(1, -2, -1, 0)$ não gera nenhum mismatch em 100% dos alinhamentos, enquanto que os alinhadores externos geram alinhamentos com substituições (e neste quesito os piores resultados são obtidos pelo pacote `est_genome`). O alinhador semi-global com esquema $(1, -2, -10, 0)$ gera alguns mismatches, devido ao fato do custo de abertura de um buraco ser bastante maior que o custo de gerar um mismatch.

Extra Gap (Conjunto 3)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-2,-1,0)	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
(1,-2,-10,0)	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene x CDS	0	0	0.00	0.00	100.00%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	0	0.00	0.00	100.00%
sim4	Gene x mRNA	0	14	1.11	1.69	54.56%
	Gene x CDS	0	16	0.96	1.63	60.85%
	Gene+200 x mRNA	0	14	1.11	1.69	54.51%
	Gene+200 x CDS	0	16	0.96	1.64	60.82%
est_genome	Gene x mRNA	0	201	16.99	21.49	27.84%
	Gene x CDS	0	201	14.13	21.12	41.66%
	Gene+200 x mRNA	0	201	17.00	21.49	27.84%
	Gene+200 x CDS	0	201	14.13	21.12	41.66%
Spidey	Gene x mRNA	0	36	0.15	1.39	97.43%
	Gene x CDS	0	23	0.10	1.03	98.01%
	Gene+200 x mRNA	0	36	0.15	1.34	97.18%
	Gene+200 x CDS	0	23	0.10	1.02	98.03%

Tabela 3.20: Mínimo, máximo, média e desvio padrão de buracos inseridos no DNA genômico pelos alinhadores `sim4`, `est_genome`, `Spidey` e semi-global com pontuações (1, -2, -1, 0) e (1, -2, -10, 0) no conjunto 3, e porcentagem de alinhamentos com nenhum buraco inserido (% OK).

Delta Exon (Conjunto 3)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-2,-1,0)	Gene x mRNA	0	0	0.00	0.00	100.00%
	Gene x CDS	0	1	0.00	0.00	99.91%
	Gene+200 x mRNA	0	0	0.00	0.00	100.00%
	Gene+200 x CDS	0	1	0.00	0.00	99.91%
(1,-2,-10,0)	Gene x mRNA	0	1	0.01	0.07	99.47%
	Gene x CDS	0	2	0.02	0.16	97.54%
	Gene+200 x mRNA	0	1	0.01	0.07	99.47%
	Gene+200 x CDS	0	2	0.02	0.16	97.54%
sim4	Gene x mRNA	-3	8	-0.01	0.20	97.46%
	Gene x CDS	-3	13	-0.05	0.30	94.02%
	Gene+200 x mRNA	-3	6	-0.01	0.20	97.44%
	Gene+200 x CDS	-3	13	-0.05	0.30	94.00%
est_genome	Gene x mRNA	-4	0	-0.14	0.30	76.79%
	Gene x CDS	-4	0	-0.21	0.40	80.24%
	Gene+200 x mRNA	-4	0	-0.14	0.30	76.85%
	Gene+200 x CDS	-4	0	-0.21	0.40	80.24%
Spidey	Gene x mRNA	-27	-1	-4.04	3.10	0.00%
	Gene x CDS	-27	-1	-3.60	3.00	0.00%
	Gene+200 x mRNA	-27	-1	-4.04	3.10	0.00%
	Gene+200 x CDS	-27	-1	-3.60	3.00	0.00%

Tabela 3.21: Mínimo, máximo, média e desvio padrão de diferença de éxons criados pelos alinhadores `sim4`, `est_genome`, `Spidey` e semi-global com pontuações (1, -2, -1, 0) e (1, -2, -10, 0) com o número de éxons esperados no conjunto 3, e porcentagem de alinhamentos com delta éxons = 0 (% OK).

Similaridade (Conjunto 3)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-2,-1,0)	Gene x mRNA	80.85%	100.00%	99.89%	0.49%	53.56%
	Gene x CDS	80.31%	100.00%	99.83%	0.63%	59.35%
	Gene+200 x mRNA	80.31%	100.00%	99.83%	0.63%	53.56%
	Gene+200 x CDS	80.85%	100.00%	99.89%	0.49%	59.35%
(1,-2,-10,0)	Gene x mRNA	80.85%	100.00%	99.89%	0.49%	53.45%
	Gene x CDS	80.31%	100.00%	99.83%	0.64%	59.15%
	Gene+200 x mRNA	80.85%	100.00%	99.89%	0.49%	53.45%
	Gene+200 x CDS	80.31%	100.00%	99.83%	0.64%	59.15%
sim4	Gene x mRNA	36.00%	100.00%	99.39%	1.34%	22.72%
	Gene x CDS	10.29%	100.00%	99.08%	2.29%	33.19%
	Gene+200 x mRNA	36.00%	100.00%	99.39%	1.34%	22.72%
	Gene+200 x CDS	10.29%	100.00%	99.08%	2.29%	33.19%
est2genome	Gene x mRNA	1.80%	100.00%	53.83%	35.09%	18.11%
	Gene x CDS	2.68%	100.00%	62.45%	35.09%	31.05%
	Gene+200 x mRNA	1.80%	100.00%	53.80%	35.10%	18.11%
	Gene+200 x CDS	2.68%	100.00%	62.44%	35.09%	31.05%
Spidey	Gene x mRNA	0.00%	100.00%	80.34%	36.49%	44.25%
	Gene x CDS	0.00%	100.00%	81.47%	37.06%	50.92%
	Gene+200 x mRNA	0.00%	100.00%	80.19%	36.75%	44.19%
	Gene+200 x CDS	0.00%	100.00%	81.53%	37.02%	50.93%

Tabela 3.22: Mínimo, máximo, média e desvio padrão de porcentagem de similaridade das bases dos éxons criados pelos alinhadores `sim4`, `est_genome`, `Spidey` e semi-global com pontuações (1, -2, -1, 0) e (1, -2, -10, 0) no conjunto 3, e porcentagem de alinhamentos com 100% de similaridade (% OK).

Mismatch (Conjunto 3)

Score	Alinhamento	Min	Max	Med	σ	% OK
(1,-2,-1,0)	Gene x mRNA	0.00%	0.00%	0.00%	0.00%	100.00%
	Gene x CDS	0.00%	0.00%	0.00%	0.00%	100.00%
	Gene+200 x mRNA	0.00%	0.00%	0.00%	0.00%	100.00%
	Gene+200 x CDS	0.00%	0.00%	0.00%	0.00%	100.00%
(1,-2,-10,0)	Gene X Mrna	0.00%	0.96%	0.00%	0.03%	99.47%
	Gene x CDS	0.00%	2.08%	0.01%	0.09%	97.63%
	Gene+200 x mRNA	0.00%	0.96%	0.00%	0.03%	99.47%
	Gene+200 x CDS	0.00%	2.08%	0.01%	0.09%	97.63%
sim4	Gene x mRNA	0.00%	2.15%	0.17%	0.21%	36.68%
	Gene x CDS	0.00%	3.23%	0.24%	0.33%	45.47%
	Gene+200 x mRNA	0.00%	2.15%	0.17%	0.21%	36.67%
	Gene+200 x CDS	0.00%	3.23%	0.24%	0.33%	45.47%
est_genome	Gene x mRNA	0.00%	7.25%	1.19%	1.26%	21.55%
	Gene x CDS	0.00%	13.36%	1.48%	1.70%	35.18%
	Gene+200 x mRNA	0.00%	7.25%	1.19%	1.26%	21.56%
	Gene+200 x CDS	0.00%	13.36%	1.48%	1.70%	35.18%
Spidey	Gene x mRNA	0.00%	23.69%	0.15%	0.98%	90.65%
	Gene x CDS	0.00%	20.56%	0.20%	1.28%	89.67%
	Gene+200 x mRNA	0.00%	28.68%	0.20%	1.38%	90.75%
	Gene+200 x CDS	0.00%	20.06%	0.21%	1.30%	89.59%

Tabela 3.23: Mínimo, máximo, média e desvio padrão de porcentagem de substituições geradas pelos alinhadores `sim4`, `est_genome`, `Spidey` e semi-global com pontuações (1, -2, -1, 0) e (1, -2, -10, 0) no conjunto 3, e porcentagem de alinhamentos com 0% de mismatch (% OK).

Delta Exon (Conjunto 4)						
Alinhador	**Erros**	**Min**	**Max**	**Med**	σ	**% OK**
(1,-2,-1,0)	1%	-15	16	-2.46	3.27	49.90%
	10%	-94	2	-33.89	15.56	46.00%
(1,-2,-10,0)	1%	-10	17	-0.25	2.58	20.28%
	10%	-24	12	-7.21	4.83	2.31%
sim4	1%	-17	2	-1.37	2.13	45.80%
	10%	-16	2	-1.34	2.08	49.13%
est_genome	1%	-17	0	-1.48	2.17	0.00%
	10%	-16	0	-1.48	2.15	0.00%
Spidey	1%	-18	-1	-3.58	2.75	0.00%
	10%	-18	-1	-3.58	2.75	0.00%

Tabela 3.24: Mínimo, máximo, média e desvio padrão de diferença de éxons criados pelos alinhadores `sim4`, `est_genome`, `Spidey` e semi-global com pontuações (1, -2, -1, 0) e (1, -2, -10, 0) com o número de éxons esperados no conjunto 4, e porcentagem de alinhamentos com delta éxons = 0 (% OK).

3.7.3 Análise dos alinhamentos com erros

Avaliando todos os resultados obtidos com alinhamentos de ESTs com erros artificiais, pode-se ver que a qualidade dos alinhamentos obtidos cai de forma igual para todos alinhadores. Ainda assim, o alinhador semi-global com a pontuação proposta ainda obtém os melhores resultados, comparativamente com os alinhadores externos específicos.

Em relação à avaliação de éxons gerados, mostrada na Tabela 3.24, tanto `sim4` quanto `est_genome` apresentam os melhores resultados, com médias inferior de 1.5 éxons a mais gerados por alinhamento. O alinhador semi-global com pontuação $(1, -2, -1, 0)$ gera em média 2.4 éxons a mais com ESTs contendo 1% de erros, e 33.8 com ESTs contendo 10% de erros. Estes valores melhoram muito com o alinhador semi-global com pontuação $(1, -2, -10, 0)$: 0.25 éxons a mais com 1% de erro e 7.21 a mais com 10% de erros.

Estes resultados se devem ao fato de que o segundo esquema tenta agrupar ao máximo as bases em regiões contíguas, devido ao custo de abertura de um buraco, o que é extremamente desejável em alinhamentos de cDNA com DNA genômico. Estes valores podem ser considerados bons, tendo em vista que buracos são considerados como íntrons, e que a média se aproxima muito dos valores obtidos pelos alinhadores externos. O pacote `Spidey` obteve os piores resultados.

A porcentagem de similaridade das bases dos éxons (Tabela 3.25) cai de 99.8% em média para 54% (taxa de 1% de erros) e 48% (taxa de 10% de erros) para o alinhador semi-global. O pacote `sim4` obtém resultados parecidos, e o pacote `Spidey` obtém uma média levemente inferior aos dois anteriores (47% para 1% de taxa de erros e 41% para

Similaridade (Conjunto 4)						
Alinhador	**Erros**	**Min**	**Max**	**Med**	σ	**% OK**
(1,-2,-1,0)	1%	4.04%	100.00%	54.14%	30.42%	0.45%
	10%	2.99%	95.88%	48.20%	27.64%	0.00%
(1,-2,-10,0)	1%	4.04%	100.00%	54.20%	30.45%	0.45%
	10%	4.34%	95.28%	47.71%	26.87%	0.00%
sim4	1%	4.04%	100.00%	53.98%	30.33%	0.41%
	10%	3.55%	95.88%	50.41%	28.54%	0.00%
est2genome	1%	1.36%	100.00%	27.78%	22.11%	1.54%
	10%	1.24%	67.23%	17.77%	12.18%	0.00%
Spidey	1%	0.00%	100.00%	47.43%	32.61%	0.49%
	10%	0.00%	95.28%	41.47%	29.67%	0.00%

Tabela 3.25: Mínimo, máximo, média e desvio padrão de porcentagem de similaridade das bases dos éxons criados pelos alinhadores `sim4`, `est_genome`, `Spidey` e semi-global com pontuações (1, -2, -1, 0) e (1, -2, -10, 0) no conjunto 4, e porcentagem de alinhamentos com 100% de similaridade (% OK).

10% de taxa de erros). Já o pacote `est_genome` obtém resultados bastante ruins neste quesito, com uma média de similaridade inferior à 28%.

3.7.4 Análise de sensibilidade dos alinhadores à taxa de erros

Para avaliar a sensibilidade dos alinhadores utilizados neste estudo à taxa de erro, efetuamos alinhamentos com ESTs com taxas de erros variando de 0.1% a 3%. Esta faixa de valores é mais realista que os 10% utilizados anteriormente, uma vez que os erros produzidos por máquina seqüenciadoras são inferiores a 3% [109], e que em grande parte dos casos as regiões de baixa qualidade são removidas por procedimentos de limpeza.

Para cada alinhamento, foram avaliados o número de éxons gerados e o grau de similaridade dos éxons. Os resultados estão sintetizados nas Figuras 3.3 e 3.4.

Analisando a variação de número de éxons gerados em relação ao número de éxons esperados, observa-se que os três alinhadores externos geram resultados praticamente constantes, sendo que `sim4` e `est_genome` tem resultados superiores a `Spidey`. O alinhador semi-global tem uma degradação quase-linear neste quesito, o que não chega a ser um resultado ruim, uma vez que o limite inferior, com taxa de 3%, é próximo do resultado obtido por `Spidey`. Além disso, deve-se levar em conta que no caso do alinhador semi-global, que não define limites de éxons/íntrons, cada buraco inserido é considerado como sendo um íntron.

Em relação à porcentagem de similaridade de bases de éxons, observa-se que todos os alinhadores obtém resultados praticamente constantes, que `sim4` e o alinhador semi-global

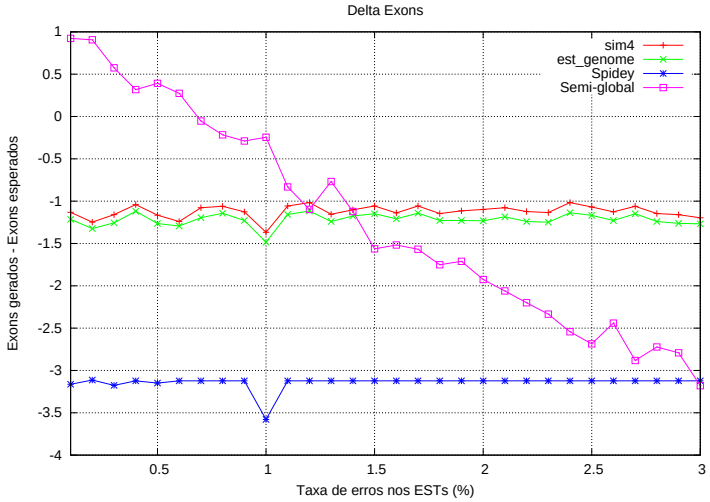

Figura 3.3: Variação do número de éxons gerados em relação ao número de éxons esperados para alinhamentos de DNA com ESTs com taxa de erro variando de 0.1% a 3%, com os alinhadores `sim4`, `Spidey`, `est_genome` e semi-global.

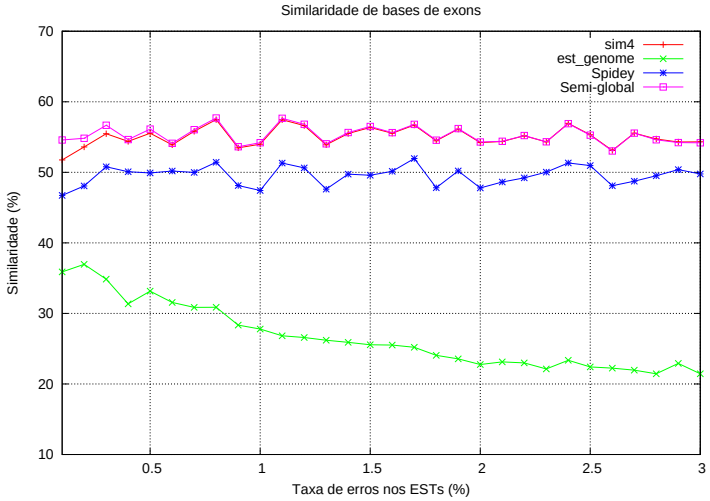

Figura 3.4: Variação da porcentagem de similaridade de éxons para alinhamentos de DNA com ESTs com taxa de erro variando de 0.1% a 3%, com os alinhadores `sim4`, `Spidey`, `est_genome` e semi-global.

Tempos de execução		
	EST-to-DNA (seg/alinhamento)	**mRNA-to-DNA (seg/alinhamento)**
sim4	0.013	0.017
Spidey	0.066	0.140
est_genome	0.640	3.400
Semi-global	0.670	5.170

Tabela 3.26: Comparação dos tempos médio de execução de cada alinhador, por tipo de alinhamento, em segundos por alinhamento.

obtém resultados bem semelhantes e superiores às duas outras ferramentas.

3.7.5 Análise de performance

Todos os testes foram efetuados em uma máquina com processador Intel Pentium 4 1.7GHz com 512Mb de memória RAM, rodando sistema operacional Linux Fedora Core 3. A Tabela 3.26 mostra os tempos médios de execução de cada alinhador utilizado neste trabalho. Os tempos foram obtidos após a execução dos alinhadores nos dados do conjunto 2 (50 alinhamentos).

Os pacotes `sim4` e `Spidey` são bem mais rápidos do que o alinhador semi-global, enquanto que este último obteve performance semelhante ao pacote `est_genome`. Esses resultados eram esperados, uma vez que o alinhador semi-global utilizado foi implementado em Java, e os outros sistemas foram implementados em C.

Os testes comparativos entre o alinhador semi-global em Java e os alinhadores do pacote fasta (Seção 3.4) mostram que as implementações em C, tem um desempenho bastante superior, com o aumento do tamanho das instâncias. Assim pode se esperar que apesar de mais lento que os pacotes `sim4` e `Spidey`, que utilizam heurísticas, a utilização de alinhadores convencionais com consumo de espaço linear são viáveis para este tipo de alinhamento.

3.8 Conclusão e trabalhos futuros

De forma geral, pode-se dizer que o alinhador semi-global com esquemas de pontuação $(1, -2, -1, 0)$ e $(1, -2, -10, 0)$ produzem alinhamentos entre DNA genômico e cDNA extremamente satisfatórios. Os resultados obtidos com alinhamentos sem erros são próximos do ideal e os resultados produzidos com dados contendo poucos erros estão muito próximos dos resultados obtidos pelos alinhadores externos, desenvolvidos especialmente para resolver este tipo de problema. Em relação aos alinhamentos de seqüências com erros, pode

se dizer que o esquema $(1, -2, -10, 0)$ é mais adequado, por ter uma tendência a preferir gerar blocos de bases contíguas e grandes regiões de buracos. Como foi dito anteriormente esta característica é muito desejável em alinhamentos de cDNA com DNA genômico.

Comparando os alinhadores externos, pode se dizer que o melhor dentre eles foi o pacote `sim4`, tanto nos testes com cDNA sem erros quanto nos testes com ESTs com erros. É importante ressaltar que o alinhador semi-global obteve resultados tão bons quanto os resultados obtidos por `sim4`. Uma desvantagem do uso de um alinhador simples em relação aos alinhadores `sim4`, `Spidey` e `est_genome` é que o primeiro não indica claramente os limites dos éxons e íntrons.

Uma possível extensão deste trabalho seria utilizar alguns critérios para adequar a pontuação do alinhador a certas regiões previamente definidas. Um critério possível seria utilizar a porcentagem de GC (porcentagem de bases G e C em uma seqüencia genética, estudos mostram que éxons e íntrons possuem porcentagens diferenciadas de GC [82]) para tentar determinar regiões com maior probabilidade de serem éxons, e para estas regiões utilizar pontuações distintas das outras. Além disso, é interessante fazer testes com ESTs com erros gerados de forma mais realistas. Uma possibilidade seria utilizar a ferramenta Seq-Gen [99], que utiliza modelos matemáticos para simular a evolução de uma seqüência genética.

Capítulo 4

Detecção de SNPs

Neste capítulo iremos discutir problemas relacionados à análise de cromatogramas e detecção de SNPs e à determinação da confiabilidade estatística dos SNPs detectados. O objetivo do trabalho a seguir é determinar algoritmos de análise de cromatograma apropriados para detecção de SNPs em lotes de seqüencias genéticas virais. Iremos também analisar alguns métodos de determinação de confiabilidade estatística dos resultados obtidos na fase de detecção de SNPs.

Os resultados apresentados neste capítulo foram descritos no artigo "*New strategy to detect single nucleotide polymorphisms*" por Galves, Quitzau e Dias [36], publicado na revista *Genetic Molecular Research* 2006, vol. 5, e apresentado no congresso X-Meetings 2006, realizado entre os dias 24 e 28 de julho de 2006 na cidade de Caxambu – MG. O relatório técnico "Comparação de métodos para determinação de SNPs com medidas de confiabilidade" (IC-06-15) por Baudet, Galves e Dias [9], sobre o estudo de confiabilidade estatística, foi depositado no Instituto de Computação da UNICAMP.

Na Seção 4.1, vamos descrever o método de determinação de bases de uma seqüência de DNA utilizado pelo pacote computacional `phred`. Na Seção 4.2 descreveremos o funcionamento de dois pacotes computacionais que tentam detectar novos SNPs, utilizando os resultados obtidos após o processo de seqüenciamento. Na Seção 4.3 descreveremos os lotes de seqüências genéticas do vírus HIV utilizados neste Capítulo. Na Seção 4.4 descreveremos as estratégias criadas para detectar SNPs no conjunto de dados utilizado. Na Seção 4.5 descreveremos os resultados obtidos pelos pacotes `polybayes` e `polyphred`. Na Seção 4.6 apresentaremos os resultados obtidos com as estratégias definidas neste Capítulo. Na Seção 4.7 descreveremos um estudo sobre métodos para estabelecer um grau de confiabilidade de SNPs determinados a partir de alinhamentos múltiplos de sequências de cana-de-açúcar. Finalmente na Seção 4.8 apresentaremos as conclusões.

4.1 Base-calling

Como foi dito no Capítulo 2, o método mais utilizado para seqüenciamento de DNA é o método de Sanger [103], cujo resultado final é um cromatograma contendo curvas representando o sinal lido a partir da placa de eletroforese. Essas curvas devem ser posteriormente interpretadas e transformadas na seqüências de bases A, C, G, T. Em um cromatograma ideal, os picos ficam espaçados de forma regular, e não se sobrepõe, de forma que cada pico corresponda a apenas uma base. Porém, cromatogramas reais não apresentam sinais ideais, por uma série de fatores ligados ao processo químico utilizado, a problemas na eletroforese e a problemas na análise da imagem obtida. O objetivo do processo de *base-calling* é obter a melhor seqüência possível.

Um dos pacotes computacionais mais antigos a executarem este processamento fazia parte do sistema da primeira máquina sequenciadora da ABI [22]. Este pacote obtém resultados de alta qualidade, e é considerado como ponto de referência para análise de outros métodos. Porém, seu algoritmo não foi divulgado.

Nas Seções 4.1.1 e 4.1.2 vamos brevemente descrever o funcionamento do pacote computacional `phred` [31, 30], que determina uma seqüência de nucleotídeos a partir de um cromatograma e cujos resultados são tão precisos quanto os resultados do pacote ABI.

4.1.1 Análise do cromatograma

O algoritmo de base-calling utilizado pelo `phred` é baseado em 4 fases, e foi refinado de forma empírica. A primeira fase do algoritmo tenta definir a localização de picos ideais, utilizando métodos de Fourier. Inicialmente, o sinal é analisado à procura de picos, onde um pico é identificado como sendo um ponto de valor máximo ou, caso este não exista, o ponto médio entre dois pontos de inflexão consecutivos. Um pico é levado em conta apenas se for maior que 10% do pico anterior e se for o maior valor entre os quatro sinais obtidos do gel. Os pontos são então varridos à procura de regiões de espaçamento uniforme de picos, onde considera-se como região janelas de 200 pontos. Para cada região centrada num pico detectado anteriormente, determina-se a distância entre picos adjacentes, assim como a média e o desvio padrão das distâncias. A região que tiver menor desvio padrão é considerada como região inicial na busca de picos ideais.

A segunda fase do algoritmo consiste em se localizar picos observados nos pontos obtidos. Os 4 sinais são varridos em busca de regiões côncavas, satisfazendo a equação $2v(i) \geq v(i+1) + v(i-1)$ onde $v(i)$ é o valor do sinal no ponto i. Para cada região encontrada, os valores dos sinais são somados para se determinar a área do pico: se esta for maior que 10% da média das áreas dos 10 últimos picos observados, e maior que 5% da área do pico imediatamente anterior, então o pico é definido como um pico observado. A relação entre a área do pico encontrado e a média dos 10 picos anteriores é armazenada

como a área relativa do pico.

A terceira fase do algoritmo consiste em se designar um pico observado na segunda fase para cada pico ideal predito na primeira fase. Esta fase consiste em 3 estágios distintos:

1. procurar por relações triviais
2. utilizar programação dinâmica para alinhar picos preditos e observados que não foram casados na primeira fase
3. alinhar picos preditos e observados que não foram casados nas duas primeiras fases mas que possam representar bases reais.

Para cada pico predito, procura-se por todos os picos observados que são mais próximos deste pico predito do que de qualquer outro, e destes picos observados, aquele com maior área relativa é designado como sendo o `best_obs_peak`. As áreas relativas dos picos observados são então recalculadas, utilizando o valor corrente da área média dos 10 últimos picos designados como `best_obs_peak`.

A quarta fase do algoritmo consiste em se analisar os picos não considerados como bases e que claramente correspondem a bases da seqüência. Isto pode ocorrer quando uma compressão, ou ruído extenso ou um erro no processamento do gel interfere com a predição de picos, resultando num número de bases inferior à realidade. Para se recuperar estes picos, observa-se se são respeitados os seguintes critérios:

1. o pico tem o maior valor dentre os 4 sinais na mesma região.
2. o pico tem um tamanho mínimo.
3. o pico não está particionado.
4. o pico está cercado por picos resolvidos.
5. o pico se encontra num local tal que adicioná-lo melhora o espaçamento entre picos.

4.1.2 Cálculo da qualidade da base

Define-se a qualidade q de uma base como sendo

$$q = -10 \times \log_{10}(p)$$

onde p é a probabilidade de erro estimada para uma dada base. Esta informação é de grande importância para análises posteriores da seqüência.

Os erros produzidos por programas de base-calling em geral são causados por erros de análise em picos em uma região do sinal, e indicações do erro podem aparecer na

vizinhança do pico errado, e não forçosamente no pico em questão. Por isso, uma análise de uma janela tende a ser mais efetiva para detecção de erros.

O algoritmo utilizado pelo pacote `phred` [30] define quatro parâmetros para cálculo da qualidade: espaçamento entre picos, razão entre bases definidas e não definidas para intervalos de sete picos, razão entre bases definidas e não definidas para intervalos de três picos e resolução de picos.

O algoritmo utilizado produz uma tabela de referência formada por um conjunto de linhas onde cada linha contém um conjunto de valores de limiar (um para cada parâmetro), junto com uma probabilidade de erro e um valor de qualidade.

O processo de atribuição de qualidades em bases consiste em se calcular o valor dos quatro parâmetros e procurar uma linha na tabela tal que todos os parâmetros calculados sejam inferiores aos valores da tabela. O valor de qualidade associado à linha é atribuído à base. Se nenhum corte satisfazendo as condições for encontrado, então atribui-se qualidade 0 à base.

4.2 Pacotes computacionais para detecção de SNPs

Nesta seção, discutiremos métodos existentes para determinação de SNPs em lotes de seqüências genéticas. Na Seção 4.2.1, descreveremos a notação IUPAC para polimorfismos em seqüencias genéticas. Nas Seções 4.2.2 e 4.2.3, descreveremos os pacotes computacionais `polyphred` e `polybayes`, que utilizam métodos distintos para detecção de SNPs baseados nos algoritmos de determinação de bases e suas qualidade descritos anteriormente.

4.2.1 Notação IUPAC para polimorfismos

A IUPAC [65] (International Union of Pure and Applied Chemistry) é a associação responsável por determinar, entre outros, padrões de notação, nomenclatura e métodos de medida de dados em química. Dentre as notações definidas pela IUPAC, estão os símbolos que definem polimorfismos.

Um polimorfismo em uma dada posição de uma seqüência genética significa que nesta posição, são conhecidas diferentes formas alélicas (dentre as quatro bases A, C, G ou T). Podemos definir 3 tipos de polimorfimos: bi-alélicos (duas formas alélicas conhecidas), tri-alélicos (três formas alélicas conhecidas) e tetra-alélicos (que pode significar ou que não se conhece ao certo quais são as formas alélicas, ou que quatro formas alélicas são conhecidas).

Para se permitir que em uma seqüência essas variações sejam representadas, foram criados símbolos para cada uma das 11 combinações possíveis dos quatro nucleotídeos A,

	M	R	S	V	W	Y	H	K	D	B	N
A	X	X		X	X		X		X		X
C	X		X	X		X	X			X	X
G		X	X	X				X	X	X	X
T					X	X	X	X	X	X	X

Tabela 4.1: Tabela de correlação entre símbolos representando polimorfismos (colunas) e as bases simples que os compõem (linhas). O X indica que um dado polimorfismo engloba uma base. Por exemplo, M representa o polimorfos A/C.

C, G e T. Os símbolos estão listados na Tabela 4.1. Os nucleotídeos estão representados nas linhas e os símbolos de polimorfismos estão representados nas colunas. Um X em uma célula da tabela indica que o símbolo da coluna representa um polimorfismo que inclui a forma alélica da base representada pela linha.

4.2.2 `polyphred`: detecção de SNPs por análise de picos do cromatograma

O pacote `polyphred` [89] utiliza os resultados obtidos pelos programas `phred` [31], que executa um algoritmo de base-calling e `phrap`, que monta seqüências de consenso, para detectar SNPs.

O seu algoritmo se baseia em duas características observadas em cromatogramas contendo seqüências com SNPs. Estas características são as seguintes:

- Uma significante redução ($\sim 50\%$) no tamanho do pico normalizado observado no cromatograma.
- Um segundo pico menor que o principal na posição em questão [71, 95].

Assim, para cada posição de uma seqüência alinhada montada pelo programa `phrap`, o programa `polyphred` analisa as áreas normalizadas e as qualidades de cada base obtidas através do programa `phred`: se for detectado um pico menor que um certo valor e a saída produzida por `phred` indicar um segundo pico, então o programa grava a posição como sendo um candidato a SNP.

4.2.3 `polybayes`: detecção de SNPs por análise bayesiana

O programa `polybayes` [77] utiliza um algoritmo de inferência Bayesiana para calcular a probabilidade de um dado alelo ser polimórfico. O algoritmo considera uma seção transversal de um alinhamento com N seqüências $R_1, \ldots, R_N$ como sendo uma permutação

de N elementos que podem assumir os valores A, C, G ou T, num total de 4^N permutações de nucleotídeos.

A detecção de SNPs em um alinhamento múltiplo é efetuada avaliando-se a probabilidade de heterogeneidade de nucleotídeos em uma seção transversal do alinhamento múltiplo, ou seja, avaliando a probabilidade de que uma dada posição de uma seqüencia possa ter várias formas alélicas.

4.3 Descrição dos lotes de seqüências genéticas do vírus HIV

No trabalho descrito neste capítulo, foram utilizadas seqüências genéticas obtidas de indivíduos que contraíram o vírus HIV. A região sequenciada em cada indivíduo é um trecho de 1302bp bem conservado do DNA do vírus HIV, ou seja, um trecho que aparece quase sempre idêntico em diversas cepas, eventualmente em posições diferentes do genoma.

Os dados utilizados foram obtidos do Instituto Fleury [63]. As seqüências foram divididas em 35 lotes, cada um correpondendo a um indivíduo. Cada lote contém seis seqüências com tamanho médio de 670bp, e uma seqüência modelo com 1302 bases com os diversos polimorfismos detectados manualmente pelos pesquisadores do Instituto Fleury. Além disso, utilizamos uma seqüência de referência do vírus HIV, comum a todos os lotes, que serve como âncora para alinhamentos múltiplos entre as seqüências dos lotes. As seqüências utilizadas cobrem uma região maior do que a referência utilizada.

Ao fazer uma pesquisa usando a ferramenta de BLAST [4] contra genomas completos de HIV disponíveis publicamente [56], observamos que os 50 melhores hits possuem pontuação 0.0 (estatisticamente idêntico) quando comparados com a seqüência de referência utilizada neste trabalho, com similaridade entre bases sempre superior 97%.

Com as 50 seqüências altamente similares obtidas pelo BLAST, foi possível construir um mapa da região de referência contendo os polimorfismos conhecidos e catalogados. Na região de referência foi possível identificar 250 posições polimórficas, das quais 19 tri-alélicas, e uma tetra-alélica. A seqüência de consenso gerada se encontra na Tabela 4.2, com os polimorfismos destacados em caixa alta.

4.3.1 Remoção de regiões de baixa qualidade

O algoritmo de remoção de regiões de baixa qualidade utilizado identifica a subcadeia com máxima pontuação: a seqüência de qualidades é convertida em uma seqüência de probabilidade de erros segundo a fórmula de erro utilizada pelo `phred` (ver Seção 4.1.2), $q = -10 \times \log_{10}(p)$. Uma vez convertido, o algoritmo subtrai de cada probabilidade de

Mapa dos polimorfismos gerados a partir do BLAST

```
cctcaRRtcactctttggcaRcgacccMtMgtcWcaRtaaRRMtagRggggcaRBtaaaggaRgctctat
tagaYacaggagcagatgatacagtaKtRRaagaMatRRRYttgccWggRaRatggaaRccaaaaatgat
agggggMattggaggttttatcaaagtaagacagtatgaKcaRRtWVYcRtagaDatHtgYggaMataRa
gctRYaggtacagtVttaRtaggacctacaccYgtcaacataattggaagaaatYtgttgactcWRMttg
gYtgYacYttaaatttYccYattagYccYattgRRactgtaccagtRaRattaaaRccaggRatggatgg
YccaaaagtYaRacaatggccattgacagaagaaaaaataaaagcattRVtagaMatYtgtRcagaRatg
gaRaaRRaaggRaaRatttcaaRaattgggcctgaRaaYccataYaatactccaRtatttgcYataaRga
aaaaRgacagtactaRatggagRaaattagtagaYttYagagaacttRataRRaRaactcaRgaYttYtg
ggaagttcaattRggaataccMcatccHKcRggDYtaaRRMagaaaaaatcagtaMcagtactRgatgtg
ggHgatgcataYttYtcagttccHYtaSaYRaRRaHttYagRaagtaYMctRcattYaccataccWagtD
YaaaYaatSagacaccMggRRBtRgRtaYcaRtaYaatgtRctKccacaRggatggaaRggRtcaccaKc
DatWttccaRWgYagcatgacaaNaatcttRgagccKtttaRaaaacaRaatccagaMMtaRtYatYtat
caRtaYRtRgatgaYttgtaYgtRggatctgaYttRgaaatagRRcMgcatagRRcaaaaatagaggaRc
tgagacaVcatctgttRaRgtggggRYtKacYacaccagacaaRaaacatcagaMRgaRcctccattYct
YtggatgggKtatgaactccatcctgaYaaatggacRRtacaRcctataRHRctRccagaRaaRgaHagc
tggactgtcaaYgacatacaRaaRttagtgggRaaattRaattgggcaagtcaRatYtaYScaggRatta
RagtaaRgcaRttRtgYaRactYcttagRggRRccaaRgcactWacagaWgtRRtaScaYtRacaVaagR
agcagaRctagaRctRgcagaaaaYagRgaRattYtaaaagaRccRgtacatggRgtgtattatgaccca
tcaaaagacYtRRtagYagaaVtacaRaaRcaRggRMaMSgc
```

Tabela 4.2: Mapa de polimorfismos gerados a partir dos resultados obtidos pela ferramenta BLAST, utilizando o trecho de referência do HIV. O mapa foi construído anotando as posições polimórficas detectadas nas colunas do alinhamento, utilizando o modo de visualização Master-Slave.

erro o valor 0.05 e procura a subcadeia da seqüência genética com maior pontuação. O algoritmo está descrito com mais detalhes em Baudet e Dias [8].

A Tabela 4.3 mostra o tamanho médio das seqüências de HIV utilizadas, divididas por lote, antes e depois das regiões de baixa qualidade serem removidas. Em alguns casos, pode-se observar que metade das seqüências é removida pelo algoritmo de filtragem. A Tabela 4.4 mostra a cobertura média das referências antes e depois do processo de remoção de regiões de baixa qualidade. O cálculo da cobertura média é feito somando o total de bases cobrindo uma dada posição na seqüência de referência, e dividindo o resultado pelo número de bases cobertas na referência. Pode-se observar que as seqüências originais cobrem toda a referência, com uma cobertura média variando entre 2.15 e 2.89.

4.4 Estratégias para a identificação de polimorfismos

A seguir serão descritas as estratégias de análise de seqüências desenvolvidas para a detecção de SNPs nas seqüências virais, levando em conta a baixa cobertura dos lotes e a baixa qualidade de uma parte das bases. A estratégia desenvolvida utiliza três passos: análise e correção do base-calling gerado pelo programa `phred`, descrito na Seção 4.4.1, filtragem dos polimorfismos gerados no passo anterior, descrito na Seção 4.4.2 e geração de consenso, descrito na Seção 4.4.3.

4.4.1 Correção de Base-calling

Esta seção descreve as estratégias implementadas para a identificação de polimorfismos com base na combinação das leituras da intensidades dos marcadores fluorecentes obtidas dos arquivos de uma máquina seqüenciadora com os picos inferidos pelo programa `phred`.

Ao longo do texto, nos referiremos ao pico correspondente à base indicada pelo `phred` como *pico de referência* e aos picos das outras bases no mesmo ponto como *picos polimórficos*. A identificação de um polimorfismo consiste portanto em comparar o pico de referência com os picos polimórficos em uma determinada posição e decidir se a posição é um polimorfismo ou se os picos polimórficos são na realidade erros de seqüenciamento.

No que diz respeito a picos num cromatograma, nos referiremos ao ponto mais alto do pico como *ponto de referência* ou *posição de referência* e aos pontos onde estão as extremidades do pico como *início* (extremidade esquerda) e *fim* (extremidade direita) do pico.

A seguir são descritas as seguintes estratégias: "Relação de Áreas", "Relação das Médias das Alturas", "Limite Variável", "Pico Único por Janela", "Eliminação de Picos Ruins" e "Pico Mais Baixo".

Lote	**Tamanho médio**	**Tamanho médio limpo**	**% removida**
1000090130	677	441	34%
1200003161	722	304	56%
1600033676	712	357	49%
1600035083	663	391	41%
3000460567	712	357	49%
3000462186R	722	304	57%
3000464645	712	357	49%
3000556396	664	489	26%
3000559731	710	326	54%
3800144129	663	391	41%
4200129683	712	357	49%
4600600055	664	489	26%
4600601614	663	391	41%
4600675963	722	304	57%
4600678332	710	326	54%
5000986228	664	489	26%
5001004479	685	324	52%
5001006959	664	489	26%
5200205498	685	324	52%
5200250019	709	367	47%
5200259064	664	489	26%
5400042344	722	304	57%
6300067797	712	357	49%
6300105258	709	367	48%
6400115894	729	398	45%
6400116358	664	489	26%
6400125551	710	326	54%
6500000467	722	304	57%
6600196125	663	391	41%
7000152272	639	281	56%
7000174441	657	412	37%
7200151026	722	304	57%
7400271960	722	304	57%
7800412280	685	324	52%
7800470275	664	489	26%

Tabela 4.3: Tamanho médio das seqüências de HIV utilizadas neste trabalho, separadas por lote, antes e depois de remover as regiões de baixa qualidade.

Lote	**Cobertura média não trimado**	**% Bases não cobertas**	**Cobertura média trimado**	**% Bases não cobertas**
1000090130	2.71	0.00%	2.15	0.00%
1200003161	2.62	0.00%	1.77	2.83%
1600033676	2.15	0.00%	1.55	7.17%
1600035083	2.62	0.00%	1.29	6.43%
3000460567	2.72	0.00%	1.78	3.55%
3000462186R	2.85	0.00%	1.24	13.58%
3000464645	2.61	0.00%	2.08	0.00%
3000556396	2.79	0.00%	1.38	3.77%
3000559731	2.80	0.00%	1.56	3.41%
3800144129	2.68	0.00%	1.86	5.48%
4200129683	2.71	0.00%	2.51	0.00%
4600600055	2.62	0.00%	2.14	0.00%
4600601614	2.88	0.00%	1.78	1.52%
4600675963	2.74	0.00%	2.51	0.00%
4600678332	2.65	0.00%	1.81	0.00%
5000986228	2.59	0.00%	1.77	2.56%
5001004479	2.61	0.00%	1.69	0.00%
5001006959	2.82	0.00%	1.88	0.00%
5200205498	2.74	0.00%	1.70	11.20%
5200250019	2.67	0.00%	1.58	22.52%
5200259064	2.57	0.00%	1.27	32.54%
5400042344	2.69	0.00%	1.99	2.36%
6300067797	2.69	0.00%	2.11	0.10%
6300105258	2.79	0.00%	1.71	14.13%
6400115894	2.82	0.00%	1.79	0.99%
6400116358	2.62	0.00%	1.65	8.20%
6400125551	2.87	0.00%	1.77	3.15%
6500000467	2.63	0.00%	1.92	0.00%
6600196125	2.73	0.00%	1.70	10.60%
7000152272	2.61	0.00%	1.45	7.40%
7000174441	2.70	0.00%	1.96	5.16%
7200151026	2.79	0.00%	1.53	4.68%
7400271960	2.89	0.00%	1.43	3.53%
7800412280	2.44	0.00%	1.56	3.87%
7800470275	2.69	0.00%	2.12	0.00%

Tabela 4.4: Cobertura média e número de bases não cobertas nos consensos de HIV por lote, antes e depois da remoção de regiões de baixa qualidade. O cálculo da cobertura média é feito dividindo-se o número total de bases cobrindo o consenso pelo número de bases cobertas.

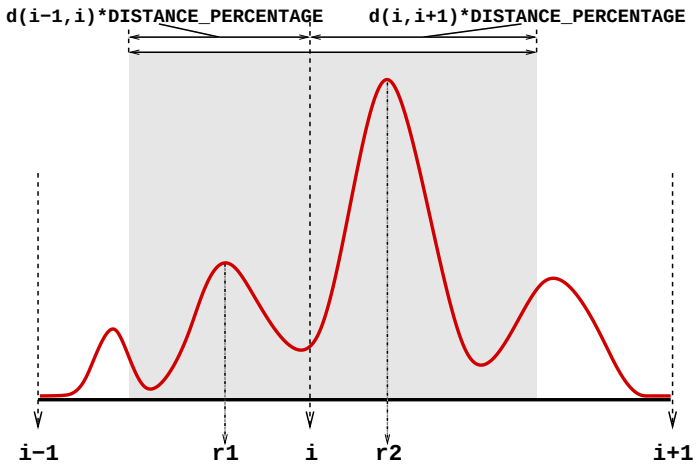

Figura 4.1: Região onde são procurados os picos para a identificação de polimorfismos. A curva representa um sinal secundário do cromatograma, não selecionado pelo phred como base na posição indicada, e as posições $i-1$, i e $i+1$ correspondem a posições dos picos de referência consecutivos. A função $d(a,b)$ determina a distância entre dois picos. No exemplo, apenas os picos cujas posições de referência estão na área escura ($r1$ e $r2$) serão considerados na procura por polimorfismos.

Relação de Áreas

Esta estratégia tenta identificar polimorfismos usando a relação entre a área do pico de referência e as áreas dos picos polimórficos. Para isso, a operação de base-calling é dividida em duas etapas: a identificação de picos para as quatro bases e a comparação entre os picos.

O primeiro passo para a identificação de picos polimórficos é definir a região no cromatograma onde eles serão procurados. A Figura 4.1 mostra como é determinada esta região. A posição i corresponde ao ponto do cromatograma onde o software `phred` identificou um pico de referência, e $i-1$ e $i+1$ correspondem aos picos de referência anterior e posterior, respectivamente. A curva representa um sinal secundário do cromatograma, não selecionado pelo phred como base na posição indicada

Para cada pico de referência i, calcula-se as distâncias $d_1 = d(i, i-1)$ e $d_2 = d(i, i+1)$, onde $d(a,b)$ determina a distância entre dois picos. A região onde os picos secundários serão procurados é dada por

$$p(i) - \texttt{DISTANCE\_PERCENTAGE} \times d_2 \leq p(x) \leq p(i) + \texttt{DISTANCE\_PERCENTAGE} \times d_1$$

onde $p(x)$ corresponde à uma posição do cromatograma e `DISTANCE_PERCENTAGE` é um parâmetro variável tal que $0 \leq \texttt{DISTANCE\_PERCENTAGE} \leq 1$.

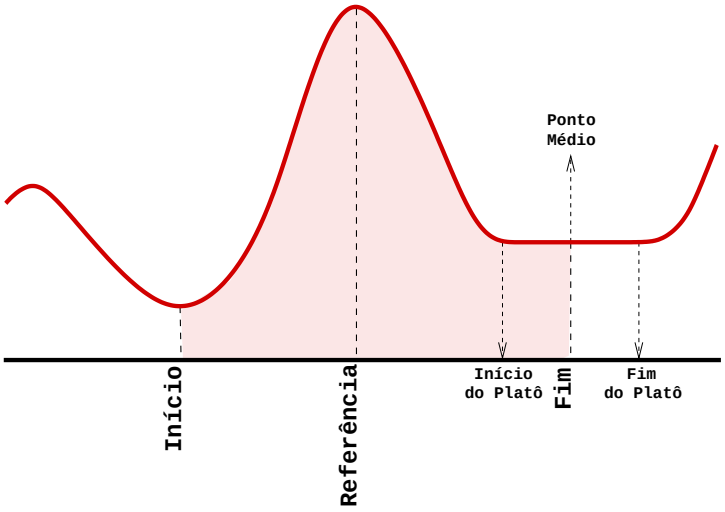

Figura 4.2: Limites considerados na análise de um pico. A região marcada corresponde à área usada como base para a determinação de polimorfismos. O ponto indicado como `Referência` indica o ponto no cromatograma usado como referência para a identificação do pico. Os pontos `Início` e `Fim` correspondem às posições limite do pico no cromatograma. Note que a posição `Fim` corresponde à mediana de um platô.

O algoritmo percorre a região posição por posição, identificando pontos que satisfaçam a condição:

$$t(p-1) \leq t(p) > t(p+1)$$
$$\text{ou}$$
$$t(p-1) < t(p) \geq t(p+1),$$

onde $t(p)$ corresponde à intensidade do marcadores fluorecentes observada na p-ésima leitura, ou seja, na posição p do cromatograma. Todos os pontos encontrados são considerados picos, mas somente o ponto mais próximo da posição indicada pelo software `phred` como o ponto correspondente a uma base é considerado para análise.

A análise começa com a determinação da região correspondente ao pico. A princípio, tanto o início quanto o fim do pico coincidem com o seu ponto de referência. Num primeiro momento, estas extremidades são afastadas do ponto de referência enquanto os valores logo à esquerda do início e à direita do fim forem menores que os valores nas posições inicial e final, respectivamente. No final deste processo, os pontos além destas extremidades correspondem a platôs ou mínimos locais. Mínimos locais são considerados extremidades de picos. No caso de platôs, todo o trecho de platô é identificado e o ponto considerado como limite do pico é o ponto médio do platô.

Tomando a Figura 4.2 como exemplo, o primeiro ponto identificado é o ponto de

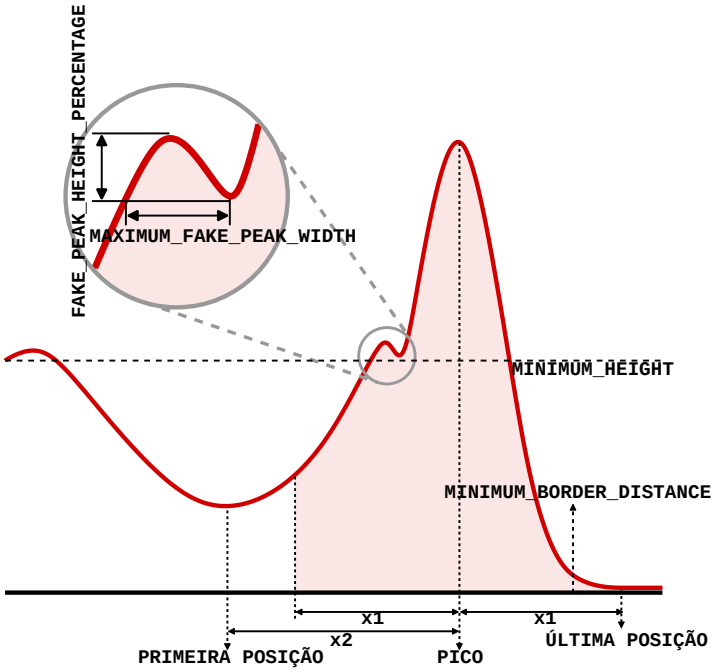

Figura 4.3: Pico identificado no método de relação de médias das alturas. Um pico falso é apresentado em destaque, bem como os parâmetros usados para identificá-lo.

`Referência` do pico, que corresponde ao seu máximo. Num segundo momento, os pontos de `Início` e `Fim` do pico são definidos. No caso deste exemplo, o `Fim` do pico determinado nesta etapa se encontra no ponto indicado como `Início do Platô`. Na terceira fase, este ponto será estendido até o `Fim do Platô` e o ponto considerado como `Fim` do pico coincide com o `Ponto Médio` do platô.

O procedimento de identificação de picos é repetido para as quatro bases nitrogenadas. Depois que os picos para as quatro bases são identificados, a área do pico da base indicada pelo software `phred` é comparada com a maior área dentre os picos das demais bases. Para essa comparação, dividimos a área do pico polimórfico pela área do pico de referência da posição. Se esta razão superar um limite determinado pelo parâmetro `MIN_RELATION`, então a posição é considerada um polimorfismo e a base original é trocada por um dos símbolos para pares de base determinado pelo código IUPAC [65] para bases nitrogenadas.

Relação das Médias das Alturas

Assim como a estratégia baseada em relação de áreas de picos, esta estratégia também é dividida em identificação de picos e comparação do sinal da base identificada pelo `phred` com os sinais das demais bases.

A identificação de picos nesta estratégia segue a mesma idéia básica da identificação

de picos usada na estratégia anterior, com a diferença de que esta é capaz de identificar e ignorar pequenos picos que correspondem a variações pontuais no sinal e não a bases. Isso é feito com base em dois parâmetros, `FAKE_PEAK_HEIGHT_PERCENTAGE` e `MAXIMUM_FAKE_PEAK_WIDTH`, que determinam a altura e largura máxima de um pico falso.

A Figura 4.3 ilustra o efeito da alteração destes parâmetros. Nesta estratégia, quando uma das extremidades do pico pára de decrescer, antes de considerar o ponto como fim do pico, é verificado se as próximas leituras não correspondem a um pico falso. Um pico falso é definido como uma região que não contenha mais do que `MAXIMUM_FAKE_PEAK_WIDTH` de leituras de intensidade do marcador fluorecente e na qual nenhuma das leituras ultrapasse em `FAKE_PEAK_HEIGHT_PERCENTAGE` da leitura da posição de referência do pico a leitura que antecede a região.

Assim como na estratégia anterior, aqui a região na qual os picos são procurados é definida pelo parâmetro `DISTANCE_PERCENTAGE`, mas neste caso, para ser considerado na análise, não basta que o pico seja o mais próximo da posição de referência fornecida pelo `phred`. Outros critérios devem ser satisfeitos antes de o pico ser considerado para análise. São eles:

Distância Mínima do Centro: Ambas as extremidades do pico devem estar a uma distância maior que a distância mínima permitida da posição de referência. Esta distância mínima é determinada por `MINIMUM_BORDER_DISTANCE`.

Altura Mínima: A maior leitura do marcador fluorecente encontrada no pico deve ultrapassar um limite estabelecido por `MINIMUM_HEIGHT`.

Há basicamente duas diferenças entre o modo como esta estratégia trata os picos e o modo como eles eram tratados pelo estratégia anterior. Neste caso, não é a área do pico que é considerada, mas sim a sua altura média, assim, o valor comparado é a intensidade média do marcador fluorecente medida no trecho correspondente ao pico. A segunda diferença é que, neste caso, as extremidades do pico analisado são eqüidistantes da posição de referência. Isso é feito alterando a primeira ou a última posição do pico de modo que ambas distem de um mesmo número de leituras da posição de referência. Esta distância é determinada pela menor distância entre os picos e a posição de referência. A Figura 4.3 ilustra esta alteração das posições. Nela é possível ver que a área considerada para o cálculo da média das alturas não começa exatamente na primeira posição do pico, mas a uma posição que dista $x1$ leituras da posição de referência do pico.

As médias das alturas são comparadas da mesma forma que a área dos picos na estratégia anterior.

Limite Variável

Esta estratégia tem o objetivo de adaptar a estratégia de identificação de polimorfismos automaticamente para seqüências de maior ou menor qualidade. A idéia é identificar o menor limite para a relação entre a altura média dos picos polimórficos e a do pico de referência para a qual a estratégia da altura média produz uma seqüência sem pares consecutivos de polimorfismos.

Para isso, é feita uma busca binária por um valor inteiro entre 10 e 100 que, usado como limite, produz uma seqüência sem pares consecutivos de polimorfismos. Além disso, o valor do limite é tal que não haja nenhum outro valor menor capaz de produzir uma seqüência com tal característica.

Pico Único por Janela

Esta estratégia tem a intenção de evitar falsos positivos criados pelo atraso no sinal de algumas bases, o que acaba criando picos deslocados no cromatograma, e falsos positivos criados pela diferença de intensidade de leitura dos diferentes marcadores usados durante o seqüenciamento. Para isso, os picos de todas as bases são identificados de uma só vez, de acordo com os passos descritos a seguir.

1. **Separação das leituras:** Para cada base nitrogenada, cria-se um vetor que armazena as intensidades dos sinais em cada ponto do cromatograma lido pela máquina sequenciadora.

2. **Normalização das Leituras:** Os valores das leituras contidas nos quatro vetores são normalizados, sendo divididos pela média dos valores não nulos contidos no vetor.

3. **Identificação dos Picos:** São identificados picos nas leituras das quatro bases em cada posição sugerida pelo `phred`. A altura média de cada pico em cada um dos vetores de leitura é armazenada em um vetor de alturas de pico. Um quinto vetor é criado para armazenar as alturas médias dos picos de referência e as alturas médias desses picos são reduzidas para zero em seus respectivos vetores.

4. **Remoção de Picos Menores:** Os vetores contendo as alturas médias dos picos para cada base são processados e, para cada conjunto de tamanho `WINDOW` de picos consecutivas, todas as alturas médias, com exceção da maior, são zeradas.

5. **Identificação de Polimorfismos:** A identificação de polimorfismos se dá da mesma maneira que na estratégia de relação de altura média, mas usando os valores de altura média restantes nos vetores de alturas.

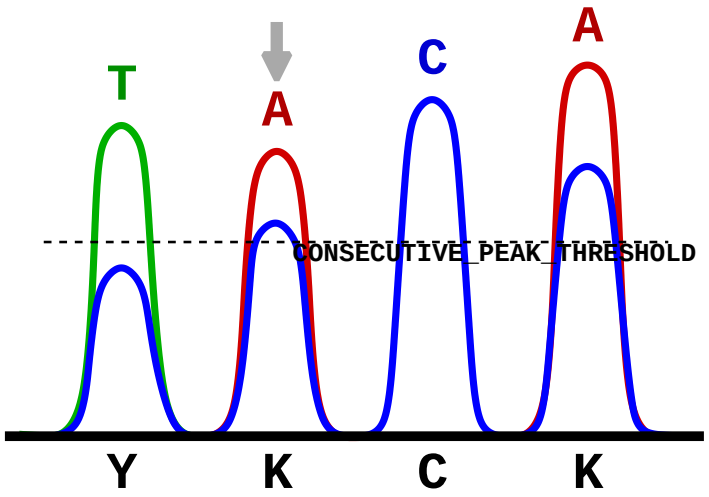

Figura 4.4: Exemplo de interferência de picos ruins num cromatograma. O suposto polimorfismo na posição indicada pela seta não é confirmado devido à presença de outros picos polimórficos vizinhos, que são interpretados como erros no seqüenciamento.

Eliminação de Picos Ruins

Assim como a estratégia anterior, esta tenta minimizar o efeito de atrasos de trechos da seqüência que podem causar picos deslocados no cromatograma. Para isso, quando um pico polimórfico é encontrado, como na posição indicada pela seta na Figura 4.4, os picos na sua vizinhança são analisados.

A análise dos picos vizinhos é feita da seguinte forma: são considerados os picos consecutivos antes e depois do pico analisado que tem altura maior ou igual a uma porcentagem da altura do pico polimórfico, determinada pelo parâmetro `CONSECUTIVE_PEAK_THRESHOLD` do basecaller. Se pelo menos um destes picos não for um pico de referência, então o polimorfismo é descartado.

A Figura 4.4 mostra um exemplo em que o polimorfismo não é confirmado. Isso porque, apesar de à esquerda da base analisada não haver nenhum pico com altura que ultrapasse o limite especificado, à direita temos dois picos com altura suficiente. Se dependesse do primeiro, o polimorfismo seria confirmado, pois o primeiro se trata de uma base indicada pelo `phred`, mas o segundo também é um suposto pico polimórfico, de modo que o basecaller passa a considerar o pico em análise como sendo um pico originado por problemas de seqüenciamento.

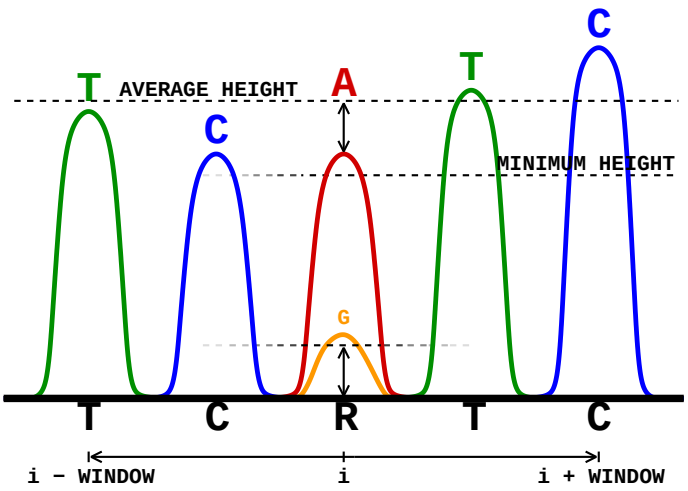

Figura 4.5: Exemplo de um polimorfismo identificado pelo critério de pico abaixo da média. No exemplo, o valor do parâmetro `WINDOW` é 2 e a média das alturas dos picos vizinhos está indicada por uma linha tracejada. A linha que indica a altura máxima de um pico para ser considerado um pico mais baixo também está indicada por uma linha tracejada, assim como a altura mínima que um pico deve ter para ser considerado um pico polimórfico.

Pico Mais Baixo

Esta estratégia tenta identificar um polimorfismo com base na diminuição do sinal da base mais freqüente causada pela base menos freqüente. Para isso, polimorfismos encontrados em bases cuja altura do pico está abaixo de um certo limite determinado pela média das alturas dos picos mais próximos, denominados de *picos mais baixos*, são tratados de maneira diferente dos demais picos. Os detalhes de como os picos são tratados são descritos a seguir.

Picos Mais Baixos: Picos mais baixos são picos cuja posição mais alta está abaixo de um limite especificado. O limite é especificado por dois parâmetros: `WINDOW` e `HEIGHT_PERCENTAGE`. O parâmetro `WINDOW` determina o número de picos à esquerda e à direita do pico analisado que serão usados para o cálculo da média das alturas. A média das alturas é então determinada com base nos $2 * \texttt{WINDOW}$ picos vizinhos do pico analisado. Para ser considerado um pico mais baixo, a altura do pico analisado tem que ser menor que `HEIGHT_PERCENTAGE` da altura média dos vizinhos.

Para um possível pico polimórfico ser considerado um polimorfismo, ele tem que ter no mínimo uma altura que, somada à altura do pico original, ultrapasse a altura média dos picos na janela considerada, tal como mostra a Figura 4.5.

Picos Normais: Todos os demais picos são tratados de maneira semelhante à das primeiras estratégias, ou seja, se a relação entre suas alturas e a altura média dos picos vizinhos for maior que um limite especificado, a posição é considerada um polimorfismo. No caso, no entanto, o limite que a razão entre as alturas deve ultrapassar é maior que nas demais estratégias.

4.4.2 Filtro de polimorfismos

Para reduzir o número de possíveis falsos positivos gerados pela fase de base-calling, foi implementado um filtro que analisa a distribuição dos SNPs ao longo de uma seqüência. O objetivo do filtro é eliminar regiões com muitos polimorfismos em série, gerados provavelmente devido a erros de seqüenciamento.

O algoritmo implementado para o filtro utiliza um janela de tamanho 9, centrada no polimorfismo que se deseja analisar. A vizinhança do polimorfismo (4 bases para cada lado) é percorrida, e para cada polimorfismo encontrado, adiciona-se uma pontuação ao polimorfismo central, que varia com a distância ao centro da janela: polimorfismos vizinhos do polimorfismo central geram uma pontuação 70; polimorfismos a distância 2 geram uma pontuação 40; polimorfismos a distância 3 geram uma pontuação 20 e polimorfismos a distância 4 geram uma pontuação 10. Ao final do processamento, se a pontuação do polimorfismo central for superior a um limiar (definido com valor 40), o polimorfismo é removido e substituído pela base original.

Por exemplo, levando em conta a seqüência abaixo, com o polimorfismo K e sua vizinhança (K, W e S foram gerados pela fase de base calling):

A	W	C	T	K	C	S	T	C

O polimorfismo W gera uma pontuação 20 para K, e o polimorfismo S gera uma pontuação 40. Portanto, a pontuação total de K ao final do algoritmo é $20+40 = 60 > 40$, o que faz com que o filtro substitua o polimorfismo pela base original gerada pelo `phred`.

4.4.3 Geração de consenso

O algoritmo de consenso utilizado analisa as bases de uma seção transversal como um todo. As regras utilizadas foram definidas empiricamente.

Inicialmente o algoritmo conta quantas vezes cada base aparece, e calcula as qualidades médias. Caso apareçam polimorfismos, estes são contabilizados individualmente e as bases que formam o polimorfismos também são incrementadas. Por exemplo, se uma seção transversal é composta pelas bases A, A e M, com qualidades 30, 20 e 10, então ao final

teremos três aparições de A com qualidade média 20, uma aparição de M com qualidade média 10 e uma aparição de C com qualidade média 10.

Uma vez contabilizadas as freqüências das bases e polimorfimos, as seguintes regras são aplicadas:

- Se apenas uma base tiver sido contabilizada, define esta base como consenso;
- Se nenhum polimorfismo tiver sido contabilizado, mas existirem mais de um tipo de base na seção, verifica se a segunda base com maior freqüência tem qualidade média superior a 50% da qualidade média da base mais freqüente, ou se esta tem qualidade média inferior a 20. Caso uma das condições se verifique, retorna o símbolo IUPAC correspondente ao polimorfismo. Caso contrário, retorna a base mais freqüente.
- Se existir um polimorfismo definido na fase de base-calling, representado pelo símbolo IUPAC correspondente, na seção transversal, verifica primeiramente se ele é compatível com as duas bases mais freqüentes. Se não for, retorna a base mais freqüente como sendo o consenso. Caso contrário, verifica se o polimorfismo aparece em pelo menos 2/3 das bases da seção do transversal ou se a qualidade média do polimorfismo é superior à 50% da qualidade da base mais freqüente: caso alguma das condições seja satisfeita, retorna o polimorfismo, caso contrário retorna a base mais freqüente.

4.5 Resultados obtidos com `polybayes` e `polyphred`

Os resultados obtidos pelos dois pacotes descritos anteriormente estão muito aquém do desejado, quando aplicados aos dados utilizados neste trabalho. Para efetuar os testes, os seguintes passos foram executados: a seqüência de referência foi utilizada como âncora para gerar um **contig** e uma seqüência de consenso com o programa `phrap` [46]. O arquivo **ACE** gerado foi processado pelos pacotes `polybayes` e `polyphred`, que geraram um arquivo de saída com os eventuais polimorfismos identificados. Para comparar o consenso obtido com a seqüência modelo de cada lote, utilizou-se o programa `cross_match` [46] para gerar um alinhamento.

Dos 35 lotes, todos com polimorfismos marcados, `polybayes` detectou polimorfismos em apenas 2 lotes, e `polyphred` detectou em 4 lotes. Os resultados estão resumidos na Tabela 4.5 e na Tabela 4.6 respectivamente.

O desempenho ruim do pacote `polybayes` pode ser explicado por seu algoritmo estatístico. Para cada seção transversal do alinhamento, uma probabilidade a priori de polimorfismo é atribuída a cada base, e se calcula a probabilidade a posteriori analisando todas as bases da seção: ao final, se a probabilidade a posteriori for superior a um certo

Detecção de SNPs em seqüências de HIV com `polybayes`

Lote	SNPs existentes	SNPs detectados	SNPs corretos	Falsos positivos	Falsos negativos
4600601614	12	1	1	0	11
4600678332	5	1	0	1	5

Tabela 4.5: Lotes de seqüências genéticas de HIV nos quais o pacote `polybayes` encontrou SNPs. Cada lote representa um indivíduo distinto e é composto por 6 leituras de uma mesma região do vírus HIV.

Detecção de SNPs em seqüências de HIV com `polyphred`

Lote	SNPs existentes	SNPs detectados	SNPs corretos	Falsos positivos	Falsos negativos
3000464645	10	1	0	1	10
4600675963	4	3	0	3	4
7200151026	26	1	0	1	26
7800470275	15	8	1	7	14

Tabela 4.6: Lotes de seqüências genéticas de HIV nos quais o pacote `polyphred` encontrou SNPs. Cada lote representa um indivíduo distinto e é composto por 6 leituras de uma mesma região do vírus HIV.

valor de limiar, o programa considera a seção como contendo um SNP. Para se obter resultados confiáveis, é portanto necessário que cada seção contenha um número relativamente grande de bases, o que não ocorre com os dados utilizados neste trabalho. Nos dois casos onde o programa reportou polimorfismos (dos quais apenas um estava correto), a configuração é a mesma: 4 bases na seção transversal do alinhamento múltiplo, com três bases A de baixa qualidade e um G de alta qualidade. Na maioria dos outros casos, as seções transversais continham no máximo três bases.

Em relação ao pacote `polyphred`, o desempenho ruim se deve provavelmente à baixa qualidade das leituras dos cromatogramas. De fato, em Nickerson et al. [89], explicita-se claramente o fato de que a qualidade dos resultados de `polyphred` depende da regularidade dos sinais do cromatograma, o que não ocorre nos lotes utilizados. Além disso, observa-se que em alguns casos o programa não leva em conta que a probabilidade de se ter muitos polimorfismos próximos um do outro é muito baixa, reportando dois ou mais polimorfismos em série.

4.6 Resultados obtidos por nossos algoritmos

A análise do melhor algoritmo para detecção de SNPs foi feita em duas etapas distintas. Na primeira etapa, foram escolhidos os dois melhores algoritmos, executados com seus valores padrão, utilizando seqüências com e sem regiões de baixa qualidade. Na segunda etapa, os dois algoritmos escolhidos foram executados com diferentes conjuntos de parâmetros.

A Seção 4.6.1 descreve os parâmetros considerados para análise dos algoritmos, a Seção 4.6.2 descreve a escolha dos melhores algoritmos de base-calling, e a Seção 4.6.3 descreve o processo de variação dos parâmetros dos algoritmos selecionados.

4.6.1 Parâmetros analisados

Para avaliar os algoritmos gerados, os seguintes parâmetros foram avaliados: verdadeiros positivos, falsos positivos, falsos negativos e taxa de erros. A seguir definimos cada parâmetro.

Verdadeiro positivo: polimorfismo predito pelo algoritmo e confirmado pelo modelo.

Falso positivos: polimorfismo predito pelo algoritmo e não confirmado pelo modelo.

Falsos negativo: polimorfismo anotado no modelo e não encontrado pelo algoritmo.

Taxa de erros: Número total de erros gerados dividido pelo número total de bases na seqüência, em porcentagem.

Detecção de SNPs em seqüências de HIV sem filtro de baixa qualidade

	Verdadeiros positivos (%)	Falsos positivos (%)	Falsos negativos (%)	Taxa de erros (%)
1	62.0339	783.3842	35.3722	2.5939
2	61.7974	684.1075	35.8253	2.3773
3	42.8509	439.8737	55.0539	2.0952
4	65.7616	638.3753	32.9799	1.2585
5	53.6839	749.5413	42.9885	3.3276
6	64.3437	1891.1288	32.8445	2.8118

Tabela 4.7: Resultado das diferentes estratégias de base-calling e consenso ao analisar seqüências de HIV. Cada linha um algoritmo de base-calling (respectivamente, 1 - Relação de Áreas, 2 - Relação das Médias das Alturas, 3 - Limite Variável, 4 - Pico Único por Janela, 5 - Eliminação de Picos Ruins e 6 - Pico Mais Baixo).

Por exemplo, suponha que o modelo é uma seqüência de 100 bases e que indica a presença de 5 polimorfismos. Se o algoritmo detectar 10 polimorfismos dos quais 3 são idênticos (posição e tipo corretos), 1 for detectado na posição correta, mas for de tipo diferente, e 6 forem encontrados em posições incorretas, então teríamos 60% de verdadeiros positivos, 40% de falsos negativos, 140% de falsos positivos e uma taxa de erro de 7%.

4.6.2 Escolha dos melhores algoritmos

Para se definir, dentre os algoritmos desenvolvidos, qual o melhor para resolver o problema proposto, efetuamos testes com todos os algoritmos de base-calling, utilizando as seqüências originais e seqüências sem as regiões de baixa qualidade.

Os resultados estão sumarizados nas Tabelas 4.7 e 4.8. Cada linha representa um algoritmo de base-calling (respectivamente, 1 - Relação de Áreas, 2 - Relação das Médias das Alturas, 3 - Limite Variável, 4 - Pico Único por Janela, 5 - Eliminação de Picos Ruins e 6 - Pico Mais Baixo). Os dados estão em porcentagem, relativa ao número de polimorfismos previamente marcados no modelo, e representam a média dos resultados obtidos nos 35 lotes.

Analisando as tabelas, podemos perceber primeiramente que o número de falsos positivos cai bastante quando se remove as regiões de baixa qualidade das seqüências de HIV: utilizando as seqüências originais, a porcentagem de falsos positivos varia entre 439% e 1891%, sendo que a média é de 864%. Nas seqüências previamente limpas, a média de porcentagem de falsos positivos em relação ao número esperado de polimorfismos é de 422%. É interessante analisar a variação de erros, mostrados nas tabelas: pode se obser-

Detecção de SNPs em seqüências de HIV com filtro de baixa qualidade

	Verdadeiros positivos (%)	**Falsos positivos (%)**	**Falsos negativos (%)**	**Taxa de erros (%)**
1	48.3452	213.9017	50.3623	1.2925
2	48.5329	286.5049	50.6905	0.7766
3	24.9277	154.0852	74.4601	0.6122
4	51.2401	223.6307	48.3857	0.3741
5	38.3020	319.6037	60.1596	1.5385
6	51.6472	1340.1315	46.4944	1.8584

Tabela 4.8: Resultado das diferentes estratégias de base-calling e consenso ao analisar seqüências de HIV sem regiões de baixa qualidade. Cada linha representa um algoritmo de base-calling (respectivamente, 1 - Relação de Áreas, 2 - Relação das Médias das Alturas, 3 - Limite Variável, 4 - Pico Único por Janela, 5 - Eliminação de Picos Ruins e 6 - Pico Mais Baixo).

var que o uso de um filtro de remoção de regiões de baixa qualidade e o uso de um filtro para remover polimorfismos espúrios diminui consideravelmente o número de erros por seqüência. Isto aumenta a confiabilidade nos resultados, e é muito desejável no caso do uso de seqüências com baixa qualidade.

Todos os algoritmos apresentaram uma leve queda na porcentagem de acertos após remoção de regiões de baixa qualidade. Isto era esperado devido ao fato que a remoção de regiões de baixa qualidade diminui a cobertura média de uma seção transversal do alinhamento, fazendo com o algoritmo de consenso não disponha de elementos necessários para detecção do polimorfismo. Além disso, uma porcentagem das bases da referência ficou sem cobertura: em alguns casos, a porcentagem de bases não cobertas chega a mais de 10%.

O algoritmo ‘‘`Pico Mais Baixo`’’ gera um número muito grande de falsos positivos, tanto em seqüências com e sem regiões de baixa qualidade, e não obtém resultados superiores aos outros algoritmos em termos de porcentagem de acertos. O algoritmo ‘‘`Limite Variável`’’ obtém um maior número de falsos negativos do que de acertos com as seqüências originais, e dentre os algoritmo é o que apresenta número de acertos mais baixo em seqüências limpas, sendo portanto descartado. Da mesma forma, ‘‘`Eliminação de Picos Ruins`’’ obtém um maior número de falsos negativos do que de acertos, e a distância entre estes valores é maior do que a distância entre os resultados obtidos pelos outros algoritmos.

Os algoritmos ‘‘`Relação de Áreas`’’, ‘‘`Relação das Médias das Alturas`’’‘ e ‘‘`Pico Único por Janela`’’ obtém os melhores resultados, e são muito semelhantes.

Porém, os dois primeiros são mais simples e possuem menos parâmetros, sendo portanto mais simples de serem regulados. Assim sendo, foram escolhidos os dois primeiros algoritmos de base-calling para se variar parâmetros, buscando a melhor combinação para detecção de polimorfimos.

4.6.3 Variação de parâmetros

Todas as estratégias de identificação de polimorfismos durante a operação de base calling possuem parâmetros que podem ser variados com o intuito de melhorar a qualidade dos resultados. Estudamos a variação de alguns destes parâmetros nas duas estratégias mais promissoras identificadas nos primeiros testes: ‘‘`Relação das Áreas`’’ e ‘‘`Relação das Médias das Alturas`’’. Os resultados são descritos nas seções a seguir.

Relação de Áreas

Os parâmetros variados foram `MIN_RELATION` e `DISTANCE_PERCENTAGE`, respectivamente com valores entre 0.2 e 0.4 e 0.2 e 0.6, realizando uma operação de base calling para cada par do produto cartesiano:

$$\begin{gathered}\texttt{MIN\_RELATION} = \{0.2, 0.25, 0.3, 0.35, 0.4\} \\ \times \\ \texttt{DISTANCE\_PERCENTAGE} = \{0.2, 0.3, 0.4, 0.5, 0.6\}\end{gathered}$$

Para cada lote de seqüências, criamos cinco gráficos de superfície:

Polimorfismos Corretos: Apresenta a porcentagem dos polimorfismos assinalados nos modelos que foram encontrados usando a estratégia de base calling em questão.

Falsos Positivos: Apresenta a quantidade de polimorfismos encontrados que não correspondem a polimorfismos previstos nos modelos em relação à quantidade de polimorfismos assinalados nos modelos.

Falsos Negativos: Apresenta a quantidade de polimorfismos previstos no modelo e não encontrados pela estratégia em questão em relação ao total de polimorfismos assinalados pelo modelo.

Razão: Apresenta a razão entre o número de polimorfismos encontrados e o número de polimorfismos assinalados no modelo.

Erro: Apresenta a razão entre o número total de erros na classificação entre posições polimórficas e não-polimórficas (soma de falsos positivos e falsos negativos) e o comprimento da seqüência consenso.

Todos os gráficos foram feitos para todos os lotes, mas para simplificarmos a análise, nesta seção serão apresentados os padrões mais comuns nos gráficos de polimorfismos corretos e de falsos positivos. Isto se deve por um lado pelo fato dos gráficos de falsos negativos e verdadeiros positivos serem altamente correlatos, uma vez que a soma dos falsos negativos com os polimorfismos corretos é sempre igual ao número de polimorfismos assinalados no modelo. Por outro lado, o número de falsos positivos domina o gráfico da razão entre número de polimorfismos encontrados e o número de polimorfismos esperados, uma vez que o número de falsos positivos é normalmente muito maior que o de polimorfismos esperados.

Pudemos identificar quatro padrões de gráficos de polimorfismos corretos, mostrados nas Figuras 4.6 a 4.9, que abrangem a grande maioria dos gráficos. Vemos pelos padrões que as regiões:

$$0.5 \leq \texttt{DISTANCE\_PERCENTAGE} \leq 0.6$$
$$\text{e}$$
$$0.2 \leq \texttt{MIN\_RELATION} \leq 0.25$$

na maioria das vezes compreende o maior número de acertos, o que parece bastante razoável, pois estes valores maximizam o tamanho da região em torno do ponto no qual o `phred` identificou um pico no qual picos polimórficos são procurados, ao mesmo tempo que minimizam a relação entre as áreas que caracteriza um polimorfismo.

A primeira vista, um bom par de parâmetros seria:

$$\begin{cases} \texttt{MIN\_RELATION} = 0.2 \\ \texttt{DISTANCE\_PERCENTAGE} = 0.6 \end{cases}$$

Por outro lado, vemos no padrão 5 mostrado na Figura 4.10 que este é justamente o ponto em que o número de falsos positivos é maior. Além disso, para valores de `DISTANCE_PERCENTAGE` superiores a 0.5, começa a haver intersecção na área de picos vizinhos nas quais picos polimórficos são procurados, o que cria a chance de um mesmo pico polimórfico ser considerado na análise de duas bases consecutivas, podendo dar origem a falsos positivos. Por estes motivos, o par de parâmetros:

$$\begin{cases} \texttt{MIN\_RELATION} = 0.25 \\ \texttt{DISTANCE\_PERCENTAGE} = 0.5 \end{cases}$$

parece ser mais indicado para esta estratégia.

Relação das Médias das Alturas

Esta estratégia possui seis parâmetros que podem ser explorados. Os três parâmetros `MIN_RELATION`, `DISTANCE_PERCENTAGE` e `MINIMUM_HEIGHT` parecem influenciar mais os re-

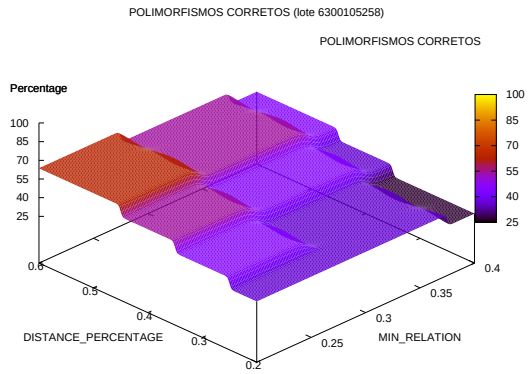

Figura 4.6: Padrão 1, encontrado na grande maioria dos gráficos de polimorfismos corretos. É caracterizado pelo ponto máximo na região em que o parâmetro `DISTANCE_PERCENTAGE` é máximo e o parâmetro `MIN_RELATION` é mínimo e pelo ponto de mínimo no extremo oposto, com uma queda suave, apresentando portanto vários níveis intermediários.

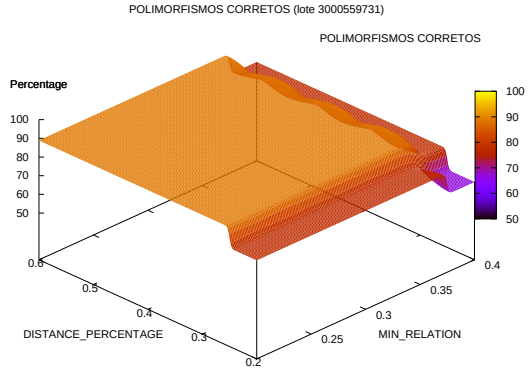

Figura 4.7: Padrão 2, encontrado em vários gráficos de polimorfismos corretos. Semelhante ao padrão 1, este padrão é caracterizado pelos pontos de máximo e mínimo nas mesmas regiões em que se encontram os pontos equivalentes do padrão 1, mas com predominância de apenas quatro planos.

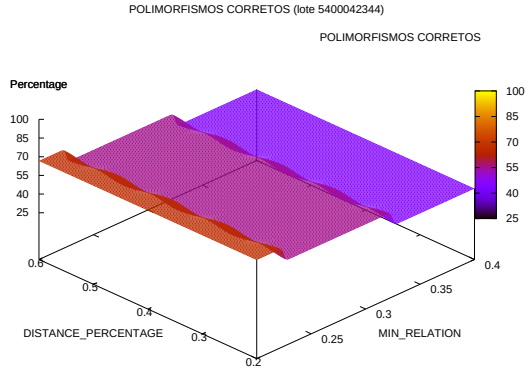

Figura 4.8: Padrão 3, encontrado em vários gráficos de polimorfismos corretos. Caracteriza-se pela variação do número de polimorfismos corretos apenas na direção do eixo `MIN_RELATION`, tendo o máximo em `MIN_RELATION` $= 0.2$.

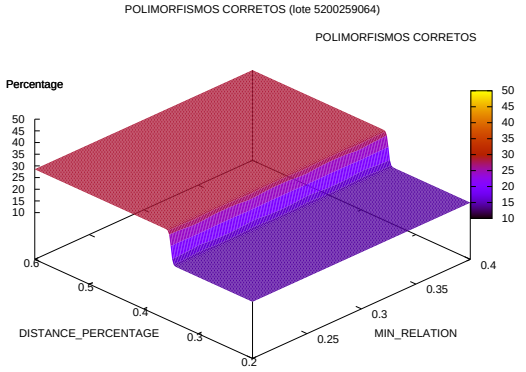

Figura 4.9: Padrão 4, encontrado em vários gráficos de polimorfismos corretos. Ao contrário do padrão 3, caracteriza-se pela variações no número de polimorfismos corretos apenas no eixo `DISTANCE_PERCENTAGE` e pelo máximo em `DISTANCE_PERCENTAGE` $= 0.6$.

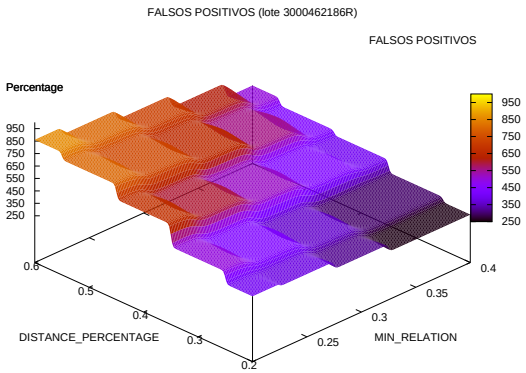

Figura 4.10: Padrão 5, observado quase na totalidade dos gráficos de falsos positivos.

sultados e foram os parâmetros que usamos nos testes cujos resultados apresentamos nesta seção.

Como ponto de partida, concentramos nossa análise na relação entre os parâmetros `MIN_RELATION` e `DISTANCE_PERCENTAGE`, tal como na análise da estratégia anterior. Para isso realizamos o base calling para cada par de parâmetros determinado pelo produto cartesiano dos seguintes conjuntos de parâmetros:

$$\texttt{MIN\_RELATION} = \{0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$$
$$\times$$
$$\texttt{DISTANCE\_PERCENTAGE} = \{0.2, 0.3, 0.4, 0.5, 0.6\}$$

Os gráficos construídos para esta análise foram os mesmos descritos na seção anterior, mas não procuramos neste momento um par de parâmetros que maximize a qualidade do resultado produzido pela estratégia, mas sim um valor para o parâmetro `DISTANCE_PERCENTAGE` que possa ser usado na análise da relação entre os parâmetros `MIN_RELATION` e `MINIMUM_HEIGHT`. As Figuras 4.11 a 4.14 mostram os padrões encontrados nos gráficos de polimorfismos corretos, enquanto as Figuras 4.15 e 4.16 apresentam os padrões encontrados nos gráficos de falsos positivos.

Da mesma forma como no teste da estratégia anterior, o valor mais indicado para o parâmetro `DISTANCE_PERCENTAGE` foi 0.5. Com este valor, realizamos novamente o base calling para os pares de parâmetro determinados pelo produto cartesiano

$$\texttt{MINIMUM\_HEIGHT} = \{20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120\}$$
$$\times$$
$$\texttt{MIN\_RELATION} = \{0.1, 0.2, 0.3, 0.4, 0.5\}$$

Aproximadamente 70% dos gráficos de polimorfismos corretos não apresentam variação alguma. Todos os demais gráficos de polimorfismos corretos se enquadram em um dos padrões mostrados nas Figuras 4.17 e 4.18. Estes padrões de gráficos mostram que a altura mínima de um pico para ser considerado um pico polimórfico, quando influencia o resultado do base calling, não o faz de forma regular. Isso se deve principalmente ao fato de as alturas variarem de um cromatograma para o outro. Uma alteração possível para a estratégia no futuro, que possivelmente ampliaria o efeito deste parâmetro no número de polimorfismos corretos seria o uso de uma altura relativa e não dos valores obtidos diretamente do arquivo de saída do `phred`.

Os falsos positivos, ao contrário dos polimorfismos corretos, normalmente sofrem a influência da altura mínima do pico. Por outro lado, assim como nos poucos lotes em que o número de polimorfismos corretos sofreu influência da altura mínima do pico, no caso dos falsos positivos há lotes em que o número diminui com o aumento da altura mínima (Figuras 4.19 a 4.21) e casos em que ocorre exatamente o contrário (Figuras 4.22 a 4.24),

mas o padrão observado em aproximadamente 40% dos casos possuia vários pontos de mínimos e máximos (Figuras 4.25 a 4.27), o que dificulta a escolha de um valor para este parâmetro que reduza a quantidade de falsos positivos na maioria das operações de base calling.

Ao final da análise, chegamos à conclusão de que o mesmo par de valores para os parâmetros `DISTANCE_PERCENTAGE` e `MIN_RELATION` são bons para esta estratégia. Quanto ao parâmetro `MINIMUM_HEIGHT`, embora não seja possível estabelecer um valor que aumente o número de acertos, é possível alterar a estratégia para utilizar alturas normalizadas com base na média das alturas dos cromatogramas, por exemplo. Desta forma a diferença de sinal entre cromatogramas diferentes é atenuada, o que, como dito anteriormente, pode ajudar na determinação de uma valor bom para a maioria dos cromatogramas. Para fins de análise, na próxima seção utilizaremos o valor `MINIMUM_HEIGHT` = 20, por ser o menos restritivo.

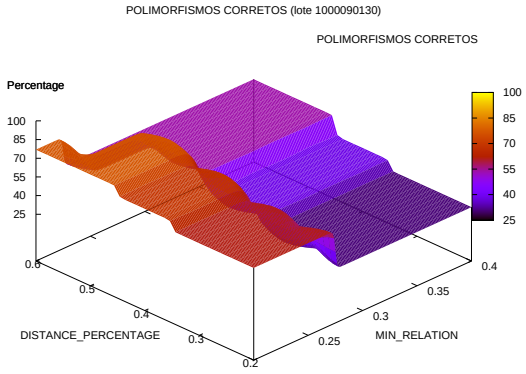

Figura 4.11: Padrão 6, encontrado na grande maioria dos gráficos de polimorfismos corretos da estratégia de média das alturas na análise conjunta dos parâmetros `MIN_RELATION` e `DISTANCE_PERCENTAGE`.

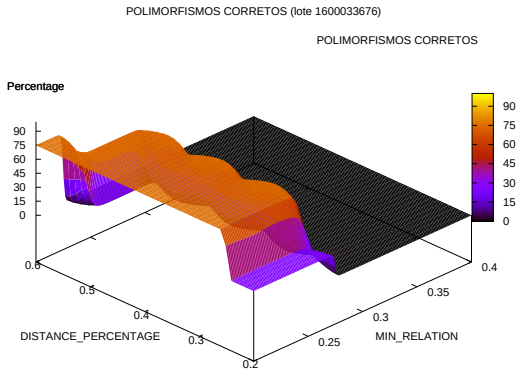

Figura 4.12: Padrão 7, encontrado em gráficos de polimorfismos corretos da estratégia de média das alturas na análise conjunta dos parâmetros `MIN_RELATION` e `DISTANCE_PERCENTAGE`.

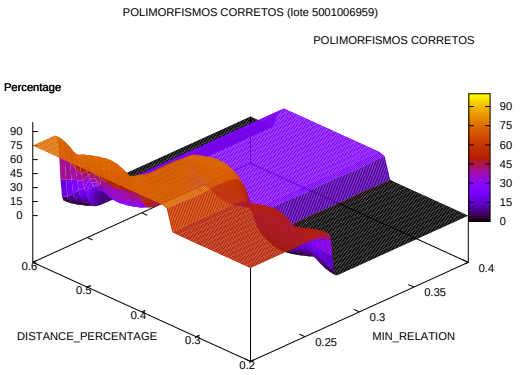

Figura 4.13: Padrão 8, encontrado em gráficos de polimorfismos corretos da estratégia de média das alturas na análise conjunta dos parâmetros `MIN_RELATION` e `DISTANCE_PERCENTAGE`.

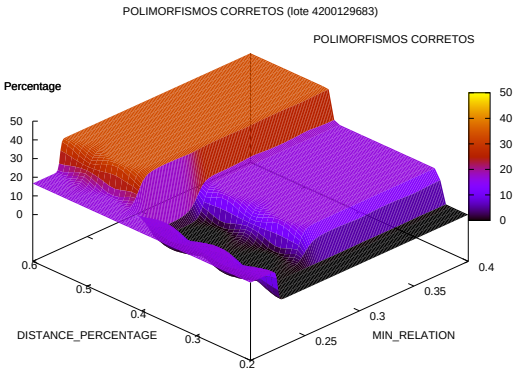

Figura 4.14: Padrão 9, encontrado em alguns gráficos de polimorfismos corretos de média das alturas na análise conjunta dos parâmetros `MIN_RELATION` e `DISTANCE_PERCENTAGE`.

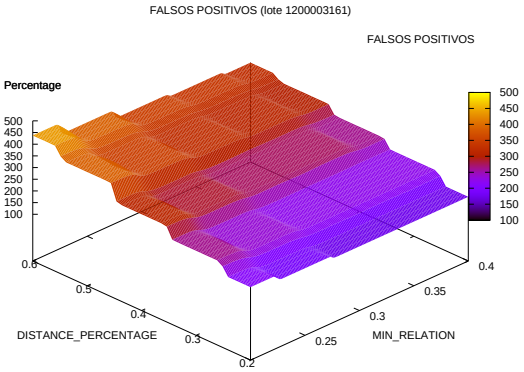

Figura 4.15: Padrão 10, observado em alguns gráficos de falsos positivos na análise conjunta dos parâmetros `MIN_RELATION` e `DISTANCE_PERCENTAGE`.

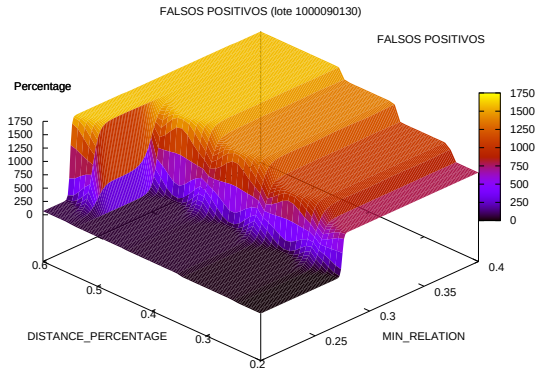

Figura 4.16: Padrão 11, observado em quase todos os gráficos de falsos positivos na análise conjunta dos parâmetros `MIN_RELATION` e `DISTANCE_PERCENTAGE`.

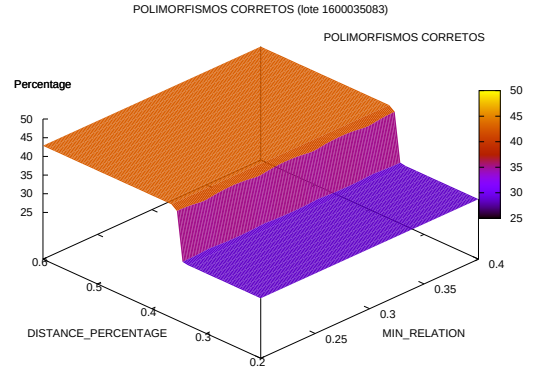

Figura 4.17: Padrão 12, encontrado em gráficos de polimorfismos corretos na análise conjunta dos parâmetros `MIN_RELATION` e `MINIMUM_HEIGHT`.

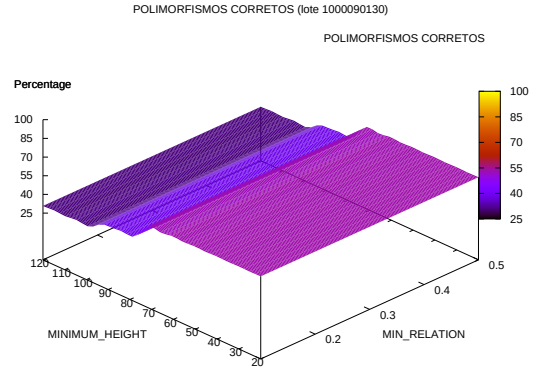

Figura 4.18: Padrão 13, observado em gráficos de polimorfismos corretos na análise conjunta dos parâmetros `MIN_RELATION` e `MINIMUM_HEIGHT`.

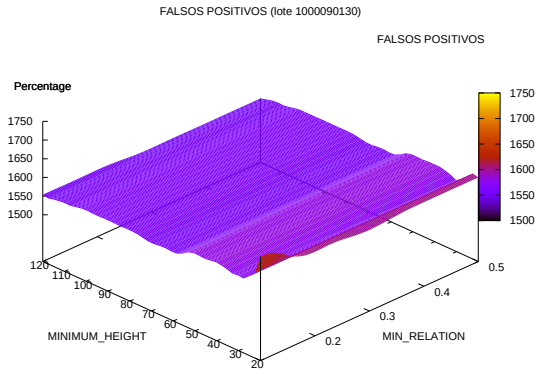

Figura 4.19: Padrão 14, observado em gráficos de falsos positivos da análise conjunta dos parâmetros `MINIMUM_HEIGHT` e `MIN_RELATION`.

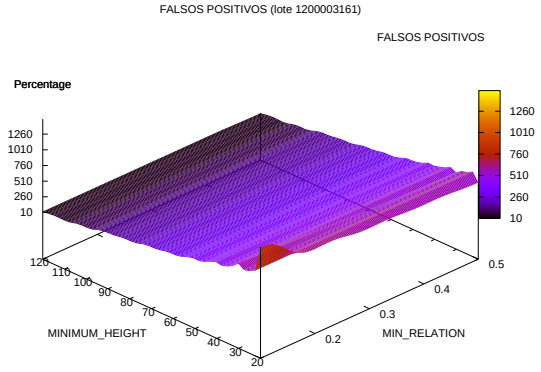

Figura 4.20: Padrão 15, observado em gráficos de falsos positivos da análise conjunta dos parâmetros `MINIMUM_HEIGHT` e `MIN_RELATION`.

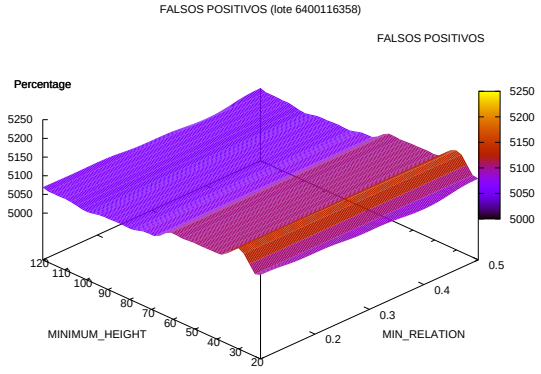

Figura 4.21: Padrão 16, observado em gráficos de falsos positivos da análise conjunta dos parâmetros `MINIMUM_HEIGHT` e `MIN_RELATION`.

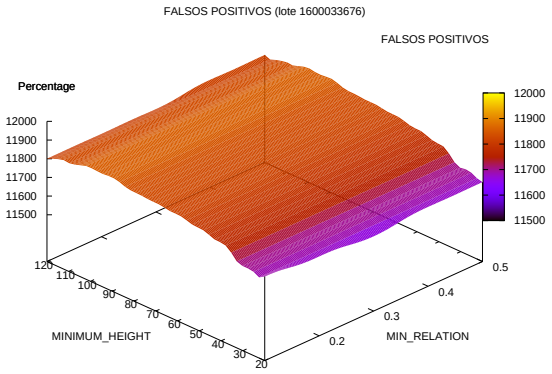

Figura 4.22: Padrão 17, observado em gráficos de falsos positivos da análise conjunta dos parâmetros `MINIMUM_HEIGHT` e `MIN_RELATION`.

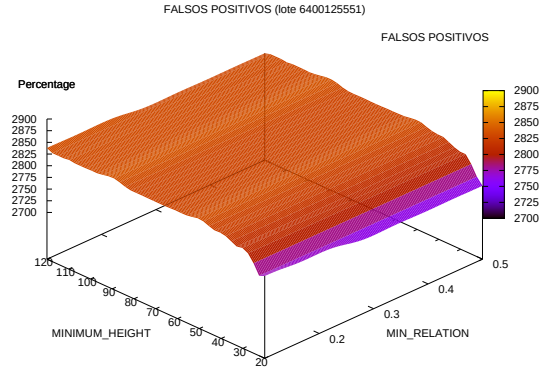

Figura 4.23: Padrão 18, observado em gráficos de falsos positivos da análise conjunta dos parâmetros `MINIMUM_HEIGHT` e `MIN_RELATION`.

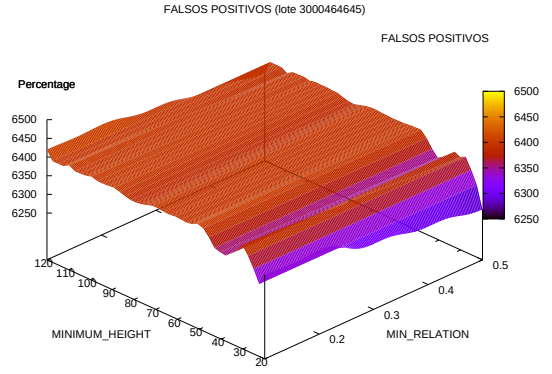

Figura 4.24: Padrão 19, observado em gráficos de falsos positivos da análise conjunta dos parâmetros `MINIMUM_HEIGHT` e `MIN_RELATION`.

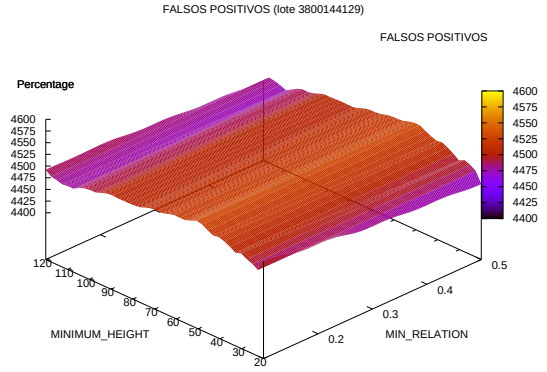

Figura 4.25: Padrão 20, observado em gráficos de falsos positivos da análise conjunta dos parâmetros `MINIMUM_HEIGHT` e `MIN_RELATION`.

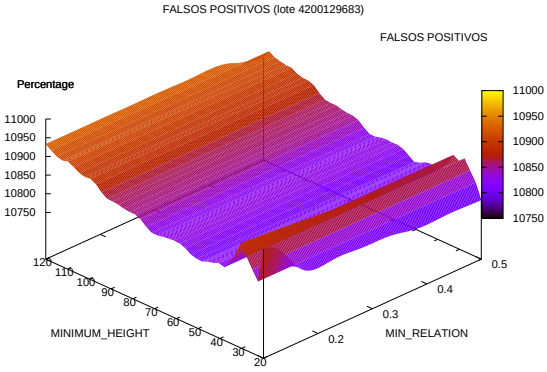

Figura 4.26: Padrão 21, observado em gráficos de falsos positivos da análise conjunta dos parâmetros `MINIMUM_HEIGHT` e `MIN_RELATION`.

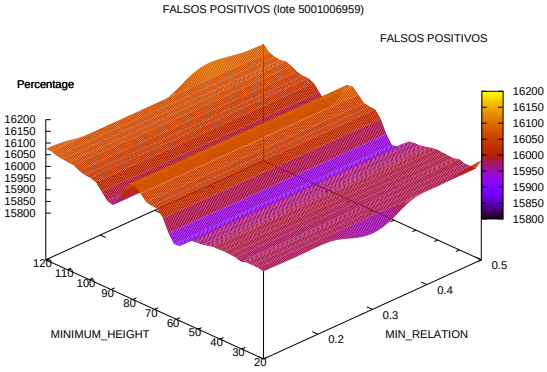

Figura 4.27: Padrão 22, observado em gráficos de falsos positivos da análise conjunta dos parâmetros `MINIMUM_HEIGHT` e `MIN_RELATION`.

Comparação de Desempenho Entre as Estratégias

Os gráficos mostrados nas Figuras 4.28, 4.29 e 4.30 apresentam uma comparação entre os resultados obtidos usando as estratégias mencionadas e os parâmetros determinados nas seções anteriores.

Claramente a estratégia de relação entre as áreas tem um desempenho melhor que a de relação entre médias das alturas. No gráfico mostrado na Figura 4.28 vemos que ela foi tão eficiente ou melhor que a outra estratégia em 100% dos casos, quando comparamos a porcentagem de polimorfismos corretos. Como conseqüência, o gráfico mostrado na Figura 4.29 comprova a eficiência da primeira estratégia quando comparamos o número de falsos negativos.

Apenas quando comparamos as porcentagens de falsos positivos encontrados a maior eficiência da relação entre as área não é tão evidente. Mesmo assim, vemos pelo gráfico que em apenas 12 lotes, aproximadamente 34% do total, esta estratégia foi superada pela relação entre as médias das alturas, o que comprova sua superioridade.

4.7 Confiabilidade estatística de SNPs

Foram feitos alguns estudos na busca de métodos estatísticos que forneçam parâmetros de confiabilidade quanto a definição de um SNP. Avaliamos dois métodos. O primeiro é a ferramenta `polybayes`. O segundo é um sistema simples que leva em conta as qualidades de cada base, determinadas pelo pacote `phred`.

Para tentar avaliar a probabilidade de uma seção transversal de um alinhamento múltiplo ou *cluster* conter um SNP, definimos um sistema simples que leva em conta as qualidades de cada base, determinadas pelo pacote `phred`. Para efeito de comparação com o método aqui descrito executamos o `polybayes` na mesma base. A seguir serão descritos o método e os resultados obtidos na aplicação do pacote computacional nos dados da cana-de-açúcar. No texto que segue chamaremos este sistema simples, por nós definido, de Método Simples de Avaliação de SNP ou, simplesmente, MSASNP.

4.7.1 Descrição do método MSASNP

Dada uma seção transversal, calcula-se para cada base a probabilidade de erro P_{erro} (ou seja, a probabilidade que a base não seja aquela determinada pelo `phred`), que pode ser obtida a partir da qualidade Q da base, através da fórmula:

$$P_{erro} = 10^{\frac{-Q}{10}}.$$

A probabilidade de acerto P_{acerto} (probabilidade da base ser de fato aquela determinada por `phred`) é portanto definida por $P_{acerto} = 1 - P_{erro}$.

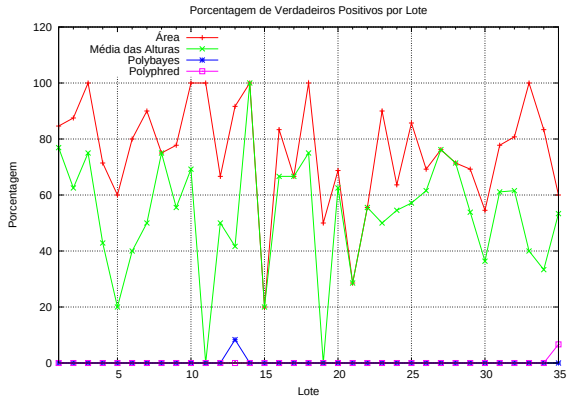

Figura 4.28: Porcentagem de polimorfismos encontrados usando as estratégias de relação de áreas (Area) e relação de média das alturas (Media) quando os parâmetros usados são os sugeridos na Seção 4.6.3. Cada ponto do eixo horizontal representa um lote de seqüências virais.

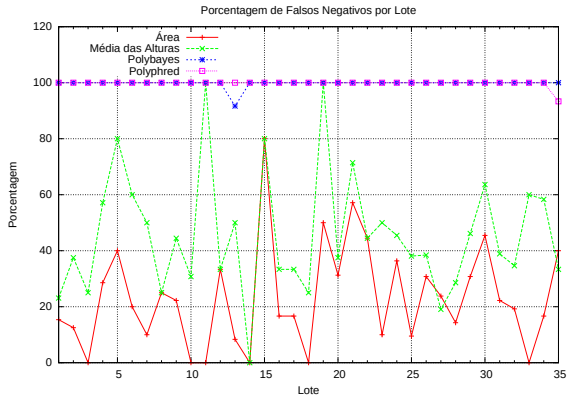

Figura 4.29: Porcentagem de falsos negativos encontrados usando as estratégias de relação de áreas (Area) e relação de média das alturas (Media) quando os parâmetros usados são os sugeridos na Seção 4.6.3. Cada ponto do eixo horizontal representa um lote de seqüências virais.

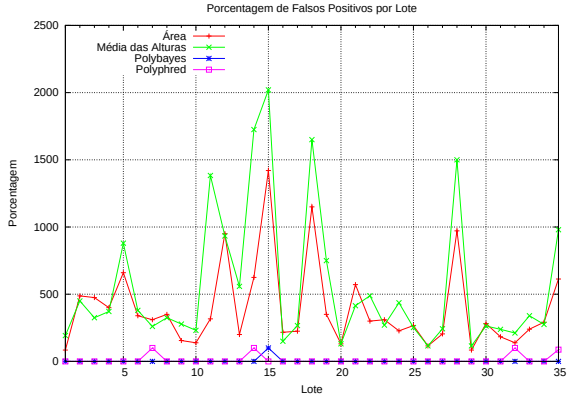

Figura 4.30: Porcentagem de falsos positivos encontrados usando as estratégias de relação de áreas (Area) e relação de média das alturas (Media) quando os parâmetros usados são os sugeridos na Seção 4.6.3. Cada ponto do eixo horizontal representa um lote de seqüências virais.

Exemplo de análise de seção transversal

	Qualidade	P_{erro}	A	C	G	T
A	10	0.100	0.900	0.033	0.033	0.033
A	20	0.010	0.990	0.003	0.003	0.003
A	15	0.032	0.968	0.011	0.011	0.011
A	8	0.158	0.842	0.053	0.053	0.053
C	11	0.079	0.026	0.921	0.026	0.026
C	11	0.079	0.026	0.921	0.026	0.026
T	3	0.501	0.167	0.167	0.167	0.499
C	9	0.126	0.042	0.874	0.042	0.042
		P_B	3.57×10^{-6}	7.66×10^{-9}	3.04×10^{-13}	9.08×10^{-13}
		$P_{B'}$	1.000	0.999	0.317	0.589

Tabela 4.9: Exemplo da fase inicial do algoritmo para cálculo de probabilidade de SNP. Cada linha mostra a base determinada pelo `phred` na seção transversal, sua qualidade associada, a probabilidade de erro associada à qualidade e as probabilidades de acerto atribuídas a cada possível base para aquela posição da sequência. As duas últimas linhas indicam a probabilidade de existir somente a base B na seção transversal (P_B) e a probabilidade de a base B aparecer pelo menos uma vez ($P_{B'}$).

Para cada posição da seção transversal, assume-se que a base definida por `phred` tem probabilidade P_{acerto} de ser a correta, e que as três outras bases têm a mesma probabilidade $P_{erro}/3$ de ser a real base existente na sequência na posição do alinhamento.

Ao final do cálculo de probabilidade para cada base em cada posição da seção transversal, temos uma tabela $4 \times N$, onde N representa tamanho da seção transversal e cada coluna representa as probabidades de uma dada base (A, C, G ou T). Para cada coluna representando uma base calculamos as probabilidades P_A, P_C, P_T, P_G da seção do alinhamento conter apenas aquela base. Assim temos que:

$$P_A = \prod_{i=1}^{N} P_{Ai},\ P_C = \prod_{i=1}^{N} P_{Ci},\ P_G = \prod_{i=1}^{N} P_{Gi},\ P_T = \prod_{i=1}^{N} P_{Ti}$$

onde P_{Ai}, P_{Ci}, P_{Ti} e P_{Gi} representam a probabilidade da seqüência i de uma seção transversal conter respectivamente as bases A, C, T, ou G. Um exemplo desse produtório está mostrado na Tabela 4.9.

A probabilidade da seção de um alinhamento conter ao menos uma base B é dada por:

$$P_{B'} = 1 - \prod_{i=1}^{N} (1 - P_{Bi})$$

onde B pode ser A, C, G ou T. A Tabela 4.9 também ilutra o cálculo destes valores.

Exemplo de cálculo de probabilidades de variações de SNP

	A	C	G	T
A	-	0,999	0,317	0,589
C	0,999	-	0,317	0,589
G	0,317	0,317	-	0,187
T	0,589	0,589	0,187	-

Tabela 4.10: Exemplo de cálculo de valores de probabilidade de variações de SNP ($P_{XY} = P_{X'} \times P_{Y'}$) utilizando os dados de $P_{B'}$ da Tabela 4.9.

A probabilidade de uma coluna do alinhamento conter um SNP é dada por:

$$P_{SNP} = 1 - (P_A + P_C + P_T + P_G)$$

e a probabilidade do SNP ser formado por um dado par de bases é calculado pela multiplicação das probabilidades de cada uma das bases aparecer ao menos uma vez, ou seja:

$$P_{XY} = P_{X'} \times P_{Y'}$$

(por exemplo $P_{AC} = P_{A'} \times P_{C'}$). Aqui discutiremos apenas SNPs bialélicos, contudo este conceito pode ser estendido para SNPs tri ou tetra alélicos.

Considerando a seção transversal do exemplo exibido na Tabela 4.9, a probabilidade de ocorrência de um SNP na posição seria de $P_{SNP} = 1 - 3.57 \times 10^{-6} - 7.66 \times 10^{-9} - 3.04 \times 10^{-13} - 9.08 \times 10^{-13} = 0.999996$. Utilizando o mesmo exemplo, a Tabela 4.10 exibe as probabilidades de ocorrência de cada uma das possíveis variações bialélicas de SNPs.

Após o cálculo das probabilidades de cada variação de SNP, o método realiza uma filtragem dos resultados com o objetivo de eliminar candidatos com menores chances de serem SNP. O filtro considera apenas variações em que ambas as bases apareçam na seção tranversal pelo menos 2 vezes e que tenham freqüência maior do que 1%. Além disso, cada base da variação deve mostrar qualidade máxima maior que 10.

Analisando o nosso exemplo, podemos ver que as bases A, C, G e T aparecem, respectivamente, 4, 3, 0 e 1 vezes na seção. Além disso, as suas freqüências são de 50.0%, 37.5%, 0.0% e 12.5%. As variações que incluem as bases G e T são desconsideradas pois as duas aparecem menos que 2 vezes na seção. Note que as variações com base T são descartadas apesar de ela apresentar freqüência maior do que 1%.

Após eliminar as variações com bases G e T, resta apenas a variação "AC". Como as qualidades máximas das bases A e C são 20 e 11, esta variação é considerada válida e apontada pelo método como o SNP encontrado para a posição.

No exemplo, existia apenas um SNP considerado válido pelo método. Contudo, podem ocorrer casos em que mais de uma variação é considerada viável. Nestes casos, o desempate é feito considerando a que possuir maior probabilidade P_{XY}. Se ocorrer um novo empate, o método escolherá a variação que tiver maior média de qualidade. Um último critério de desempate considerará a variação que apresentar maior média de valores máximos de qualidade.

4.7.2 Comparação dos métodos `polybayes` e MSASNP

Para avaliar os métodos utilizamos um conjunto de dados formado por 8198 *clusters* de seqüências de cana-de-açúcar, que descrevemos na Seção 5.4.1. Nestes *clusters*, os pesquisadores do projeto SUCEST apontaram a presença de 42853 posições de SNPs (5.23 SNP/*cluster*).

Como estamos trabalhando apenas com posições bialélicas, nós retiramos todas as posições que apresentaram variações tri ou tetra alélicas ou que apresentaram eventos de INDEL. Após a filtragem, obtivemos uma lista de 41558 posições, contendo SNPs bialélicos, distribuídas em 8115 *clusters* (5.07 SNP/*cluster*).

Executamos o software `polybayes` com e sem filtro de seqüências parálogas. Este filtro do `polybayes` analisa as seqüências e, a partir das discrepâncias que elas apresentarem em relação à âncora (consenso do *cluster*, neste caso), separa-as em dois grupos: nativas e parálogas. Para o cálculo de SNPs, o programa considera apenas as seqüências nativas. Quando o filtro é aplicado o número de seqüências a ser analisada em busca de SNPs é menor e, portanto, a execução é mais rápida. Se o filtro é desligado, além da execução ser mais lenta, geralmente o programa retorna um número maior de posições de SNPs.

A execução do `polybayes` sem o filtro de parálogos produziu um total de 172842 posições de polimorfismo (21.08 SNP/*cluster*). Deste total, 131622 posições eram SNPs bialélicos distribuídos em 8195 *clusters* (16.06 SNP/*cluster*). O tempo de execução foi de 661 minutos e 45 segundos em uma máquina com 2 processadores INTEL Xeon 3.2 GHz, 4 GB DDR ECC e 4 discos 320 ULTRA SCSI 133 GB rodando Fedora Core 4.

Utilizando o filtro, o número de posições polimórficas obtido foi de 138695, distribuídas em 8042 *clusters* (16.05 SNP/cluster). Destas, 103325 eram posições bialélicas distribuídas em 8029 *clusters* (12.60 SNP/*cluster*). Para produzir estes dados, o programa gastou 578 minutos e 59 segundos na mesma máquina.

Aplicamos o método MSASNP no mesmo conjunto de *clusters*. O método gastou 302 minutos e 27 segundos na mesma máquina utilizada para executar o `polybayes`.

Os dados brutos, produzido pelo método, indicavam os valores para todas as posições que tinham pelo menos duas bases diferentes em uma seção transversal do alinhamento. Um total de 4144426 posições apresentaram esta característica mínima, resultando em

505, 54 SNP/*cluster*.

Este número de posições é muito maior do que o número de posições indicadas pelo SUCEST, por exemplo. Obviamente, a maior parte não se trata de SNPs e, portanto, um critério deve ser criado para separar as posições que realmente são polimorfismos.

Decidimos utilizar como critério o valor de probabilidade P_{SNP} calculado. Contudo, este número, pela própria natureza do cálculo, tende a ser, na maioria dos casos, muito próximo de 1. Por exemplo, se utilizamos o valor mínimo de 0.9 para considerar a posição como um SNP, temos um total de 4115998 posições (502.07 SNP/*cluster*), ou seja, apenas 0.68% do conjunto total é descartado.

Para avaliar o efeito da escolha de diferentes valores para a probabilidade mínima requerida, utilizamos a fórmula $f(x) = 1 - 10^{-x}$, com x variando no intervalo [1.20], para definir o conjunto de probabilidades mínimas a ser testado. Os números de posições definidas como sendo SNPs, segundo cada valor utilizado, é exibido no gráfico da Figura 4.31 (curva vermelha). Neste gráfico podemos ver também o número de posições em que ocorreu correspondência com o conjunto de SNPs definido pelo SUCEST (curva verde). A curva azul representa o número de polimorfismos bialélicos encontrados pelo `polybayes` sem o filtro de parálogos. Já a curva magenta indica a intersecção entre `polybayes` e SUCEST.

Podemos observar no gráfico que o método MSASNP aponta muitas posições, apresentando número maior de SNPs que o `polybayes` em grande parte dos casos. Podemos notar também que apresenta sempre um número maior de SNPs que o apresentado pela intersecção entre SUCEST e `polybayes`.

O gráfico da Figura 4.32 exibe a porcentagem de posições de SNPs apontadas pelo `polybayes` que conferem com os dados do SUCEST (curva verde). E na curva vermelha apresentamos os valores obtidos com o método MSASNP que conferem com os apresentados pelo SUCEST. Podemos observar que o `polybayes` acerta bastante e que o método MSASNP acerta cada vez menos quando impomos mais restrições.

O gráfico nos mostra que mesmo com a utilização de um valor para probabilidade mínima, o método MSASNP continua a apontar muitas posições. Isso ocorre porque os alinhamentos dos *clusters* possuem muitas regiões com baixa qualidade, produzindo uma grande quantidade de SNPs em posições consecutivas.

Assim, decidimos aplicar um filtro de janela deslizante que percorre as posições do alinhamento e elimina SNPs consecutivos. A janela inicia a procura pelo primeiro candidato a SNP existente no alinhamento. Ao encontrar esta posição, a janela a indica como SNP e pula 5 posições, ignorando qualquer candidato a SNP existente neste intervalo. Este procedimento, portanto, não permite que exista um SNP distante do outro a menos de 5 posições.

Os gráficos das Figuras 4.33 e 4.34 são equivalentes aos das Figuras 4.31 e 4.32 só que

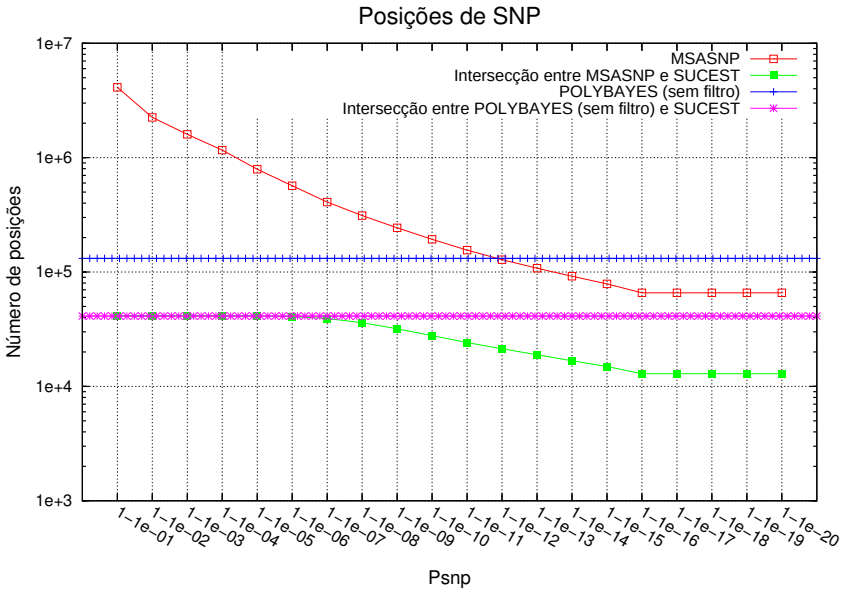

Figura 4.31: Gráfico comparativo no número de posições marcadas como SNP. No eixo X temos um limite inferior, para a probabilidade de ser SNP, onde consideramos que seja um SNP. No eixo Y temos o número de posições marcadas como sendo SNP. A curva vermelha refere-se ao método MSASNP. A curva azul apresenta o número de SNP apontados pelo `polybayes` (sem filtro). A curva magenta apresenta o número de SNPs que aparecem no SUCEST (dados de referência) e `polybayes` ao mesmo tempo. A curva verde aponta o número de SNPs que aparecem no SUCEST e método MSASNP ao mesmo tempo.

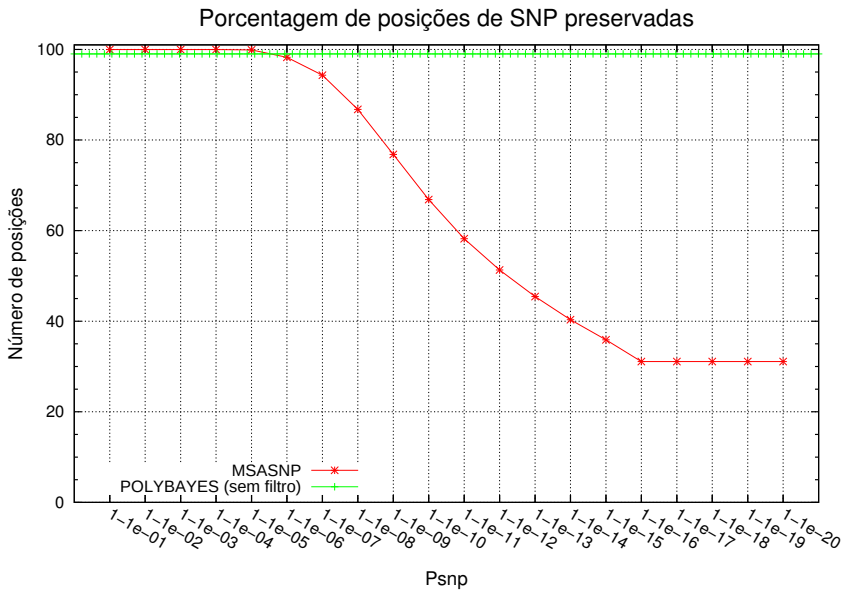

Figura 4.32: Gráfico comparativo no número de SNPs preservados, tomando como referência os dados do SUCEST. No eixo X temos um limite inferior, para a probabilidade de ser SNP, onde consideramos que seja um SNP. No eixo Y temos o número de posições preservadas. A curva verde refere-se ao `polybayes` (sem filtro) e a curva vermelha ao método MSASNP.

agora utilizando a janela deslizante. Como podemos ver, o número de posições indicadas como SNP pelo método MSASNP caiu bastante. Contudo, a porcentagem de posições apontadas pelo SUCEST e pelo `polybayes` também caíram. Isso indica que este filtro não é capaz de eliminar falsos positivos sem afetar os verdadeiros positivos.

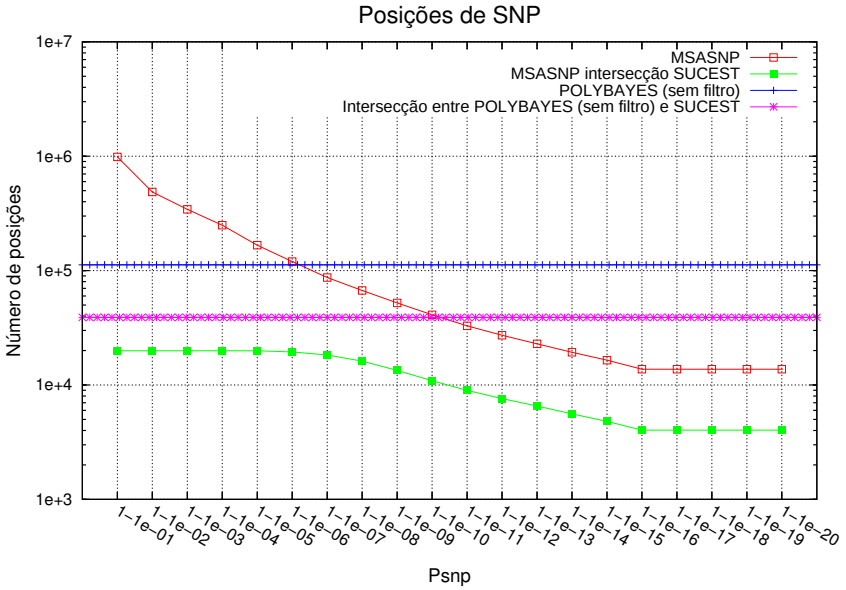

Figura 4.33: Gráfico comparativo no número de posições marcadas como SNP quando utilizamos uma janela deslizante de cinco posições entre dois SNPs. No eixo X temos um limite inferior, para a probabilidade de ser SNP, onde consideramos que seja um SNP. No eixo Y temos o número de posições marcadas como sendo SNP. A curva vermelha refere-se ao método MSASNP. A curva azul apresenta o número de SNP apontados pelo `polybayes` (sem filtro). A curva magenta apresenta o número de SNPs que aparecem no SUCEST (dados de referência) e `polybayes` ao mesmo tempo. A curva verde aponta o número de SNPs que aparecem no SUCEST e método MSASNP ao mesmo tempo.

Além disso, analisando o `polybayes`, verificamos que a intersecção de suas posições de SNP com as do SUCEST é de 41138, ou seja, 98.99% das posições bialélicas.

Para isso, o `polybayes` produziu 131622 posições, ou seja 3.17 vezes mais do que o apontado pelo SUCEST.

Por outro lado, o método MSASNP usando probabilidade mínima de $1-10^{-6}$ produziu uma intersecção de posições de SNP com as do SUCEST é de 40828, ou seja, 98.27% das posições bialélicas. Porém produziu 567618 posições, ou seja 13.66 vezes mais do que o apontado pelo SUCEST.

Se utilizarmos o filtro de janelas nos dados do `polybayes`, o número de posições de SNP cai para 112247 (2.70 vezes mais que o conjunto SUCEST). A intersecção entre estes

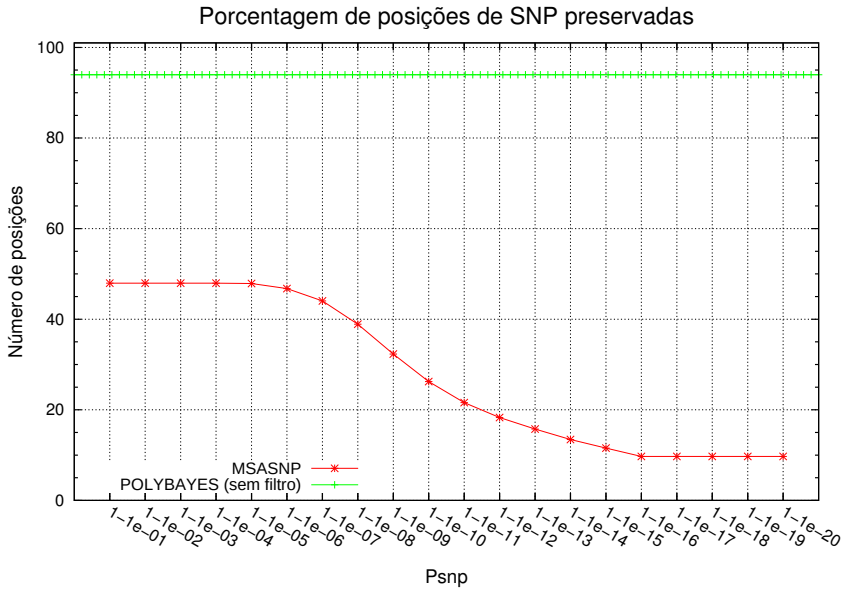

Figura 4.34: Gráfico comparativo no número de SNPs preservados, tomando como referência os dados do SUCEST, quando utilizamos uma janela deslizante de cinco posições entre dois SNPs. No eixo X temos um limite inferior, para a probabilidade de ser SNP, onde consideramos que seja um SNP. No eixo Y temos o número de posições preservadas. A curva verde refere-se ao `polybayes` (sem filtro) e a curva vermelha ao método MSASNP.

dois conjuntos foi de 39034 posições, ou seja, 93.96% do total.

Usando o filtro de janelas no método MSASNP com probabilidade mínima de $1 - 10^{-6}$ temos 120089 posições de SNP (2.89 vezes mais que o conjunto SUCEST) e uma intersecção de 19437 posições. Apesar de atingirmos um número de posições próximo do obtido pelo `polybayes`, o número de verdadeiros positivos caiu para 46.79% dos SNPs apontados pelo SUCEST.

De acordo com os resultados apresentados nos gráficos concluimos que o método MSASNP não foi capaz de obter resultados satisfatórios. Como pudemos ver, o método gera uma quantidade enorme de falsos positivos e apresenta um menor índice de acerto que o `polybayes`. Pudemos verificar também que a ferramenta `polybayes` apresenta bons resultados.

4.8 Conclusão e trabalhos futuros

Os resultados obtidos no trabalho de detecção de SNPs por análise de cromatograma foram bastante satisfatórios, se comparados com os resultados gerados pelos programas `polybayes` e `polyphred`. Devemos também levar em conta que os lotes utilizados possuem baixa cobertura média de cada base da seqüência de referência, com alta porcentagem de bases com baixa qualidade e grande quantidade de polimorfismos, dificultando o trabalho de detecção.

Após executarmos os diversos algoritmos de detecção de SNPs propostos com diversas parametrizações, estabelecemos que o melhor algoritmo é o de "Relação de Áreas", com parâmetros:

$$\begin{cases} \texttt{MIN\_RELATION} = 0.25 \\ \texttt{DISTANCE\_PERCENTAGE} = 0.5 \end{cases}$$

Seqüências virais, como o HIV utilizado neste trabalho, possuem um alto número de mutações. Seria portanto interessante repetirmos o experimento utilizando seqüencias genéticas de seres vivos mais conservados, como, por exemplo, mamíferos, de forma a validar os algoritmos desenvolvidos.

Quanto aos estudos sobre métodos estatísticos de confiabilidade, definimos um método simples de determinação da confiabilidade de SNPs, que chamamos de MSASNP. Analisamos também a ferramenta `polybayes` e comparamos os resultados. Pudemos observar que a ferramenta `polybayes` exigiu um pouco mais de tempo para realizar o processamento dos dados, porém apresentou resultados melhores.

Capítulo 5

Correlação de polimorfismos

Neste capítulo discutiremos o uso de SNPs em desequilíbrio de ligação, também chamado de Linkage Disequilibrium ou LD, para mapeamento fino de genes. O relatório técnico "Um algoritmo para identificação de correlações múltiplas de polimorfismos" (IC-06-14) por Almeida, Galves e Dias [3], descrevendo os resultados obtidos neste capítulo, foi depositado no Instituto de Computação da UNICAMP.

Estima-se que o mapeamento de genes baseado em Linkage Disequilibrium trará resultados mais completos e precisos, permitindo obter informações de grande utilidade para compreensão e tratamento de doenças com causas genéticas [101]. LD tem sido apontado como uma ferramenta de grande utilidade para facilitar o mapeamento de genótipos complexos [6], e para refinar a busca por locus responsáveis por doenças. Neste Capítulo apresentamos uma análise sobre LDs e LDs múltiplos em genes do cromossomo 6 humano e da cana-de-açúcar, obtidos através do projeto SUCEST.

Na Seção 5.1 descreveremos brevemente os problemas relacionados à metodologias de mapeamento de genes, e os mecanismos genéticos que levam à criação de LDs. Na Seção 5.2, vamos definir algumas medidas utilizadas para quantificar o grau de associação entre loci de um cromossomo. Na Seção 5.3 definiremos o conceito de LD múltiplo, e o algoritmo para cálculo de LDs múltiplos a partir de uma estrutura de grafo. Na Seção 5.4 apresentaremos os dados que foram utilizados nos testes. Na Seção 5.5 apresentaremos a análise de LDs múltiplos em dados do genoma da cana-de-açúcar. Na Seção 5.6 apresentaremos a análise de LDs nos dados do genoma humano. E, finalmente, na Seção 5.7 apresentaremos a conclusão do trabalho.

5.1 Correlação de polimorfismos

Na Seção 5.1.1 descreveremos a metodologia tradicional de mapeamento de genes. Na Seção 5.1.2, vamos definir o conceito de *Linkage Disequilibrium*, e como ele é usado para

mapeamento fino de genes e haplótipos. Na Seção 5.2 vamos definir algumas medidas usadas para descrever associação entre alelos.

5.1.1 Mapeamento de genes relacionados a doenças

Estabelecer a correlação entre um fenótipo observado com um genótipo é um dos objetivos fundamentais da genética. Obter a seqüência genética relacionada a uma doença é essencial para se produzir terapias e tratamentos adequados. Métodos gerais para descobrir os genes responsáveis por doenças, que possuem características mendelianas simples, só foram determinadas no início da década de 80, quando a análise de ligação de genes foi proposta pela primeira vez.

Genes que controlam características mendelianas podem ser identificados e isolados com base em informações sobre as características hereditárias, utilizando a técnica de clonagem posicional [13]. Esta técnica permite a delimitação de regiões cromossomais de 1 a 2 cM (um a dois *centimorgans*). Para um mapeamento mais fino de genes, é recomendável utilizar técnicas baseadas em desequilíbrio de ligação, que permitem marcar regiões menores e mais próximas dos genes de interesse.

5.1.2 Mapeamento por Desequilíbrio de Ligação

Linkage Disequilibrium, ou LD, é uma associação não aleatória de alelos. Um conjunto de SNPs estatisticamente associados é chamado de haplótipo. Quando dois ou mais alelos específicos, em *loci* distintos, em um mesmo cromossomo são sempre encontrados em conjunto, então os *loci* estão em desequilíbrio de ligação [91, pg. 177]. Quando isto acontece, a identificação de um um SNP em um locus fornece informações sobre SNPs em outros loci.

Muito do conhecimento e entendimento de como LDs são formados na natureza veio do estudo feito em espécies de *Drosophila* [10], e em particular, da *Drosophila Melanogaster*, de onde foram tirados estudos mais detalhados de LDs.

Análises de LD são mais efetivas em populações isoladas, que em geral possuem menor heterogeneidade alélica, ou em análises de doenças causadas por mutações mais antigas e bastante comuns na população humana. Como exemplo de doença que se aplica ao primeiro caso podemos citar a displasia distrófica em finlandeses e como exemplo do segundo caso podemos citar a fibrose cística e a doença de Huntington em populações européias [13].

Estudos foram feitos para se analisar os padrões de LD na população humana. Pode-se observar que estes são diferentes conforme as regiões, com europeus apresentando menor diversidade nucleotídica e maior número de LDs que africanos. Comparações são dificultadas pelo fato de que diversos estudos foram feitos utilizando medidas diferentes.

Um ponto que pode ser extraído destes estudos é que a variação de LD em qualquer distância é grande, e não é previsível de uma região geográfica para a outra. Com isso, deve-se tomar muito cuidado ao se fazer qualquer predição sobre LDs em regiões do globo terrestre onde não foram feitos estudos empíricos.

LD é um fenômeno muito útil para mapeamento de genes, por permitir produzir mapas com definição de 0.1 cM, o que é equivalente a aproximadamente 100 kbp (cem mil pares de base) quando falamos de genoma humano. Assim, a distância entre um marcador e um gene de interesse é bem menor, facilitando o seu alcance e posterior clonagem e análise.

5.2 Medidas utilizadas para quantificar um LD

Várias medidas foram criadas para se quantificar LD [27]. A mais antiga das medidas propostas para desequilíbrio é chamada D. Esta medida quantifica um LD como sendo a diferença entre a freqüência observada entre um haplótipo de dois loci e a freqüência que seria esperada se os alelos fossem aleatórios. Considerando os alelos A, a, B, b, temos:

$$D = P_{AB} - P_A P_B$$

onde P_B e P_A são as probabilidades de aparição dos alelos separadamente e P_{AB} é a probabilidade dos dois alelos aparecerem juntos. Podemos afirmar que: $P_A = P_{AB} + P_{Ab}$, $P_a = P_{aB} + P_{ab}$, $P_B = P_{AB} + P_{aB}$, $P_b = P_{Ab} + P_{ab}$ e $P_{AB} + P_{Ab} + P_{aB} + P_{ab} = 1$. Observe que:

$$P_A P_B = P_{AB}(P_{AB} + P_{Ab} + P_{aB}) + P_{Ab}P_{aB}$$

então:

$$\begin{aligned} P_{AB} - P_A P_B &= P_{AB} - P_{AB}(P_{AB} + P_{Ab} + P_{aB}) - P_{Ab}P_{aB} \\ &= P_{AB}(1 - P_{AB} - P_{Ab} - P_{aB}) - P_{Ab}P_{aB} \\ &= P_{AB}P_{ab} - P_{Ab}P_{aB} \end{aligned}$$

Seu valor numérico tem pouco uso na comparação de LDs. Isto se deve ao fato de que D depende da freqüência de alelos, que dificulta a comparação e a avaliação dos resultados. Assim, várias outras medidas, com escala variando de 0 a 1, baseadas em D foram propostas, e serão descritas a seguir.

Medida D'

A medida D', proposta por Lewontin [74], é dada pela seguinte fórmula:

$$D' = \frac{D}{D_{max}}$$

com

$$D_{max} = \begin{cases} min[P_A P_b, P_a P_B] & \text{se } D > 0 \\ min[P_A P_B, P_a P_b] & \text{se } D < 0 \end{cases}$$

O denominador da fórmula corresponde ao maior valor que D pode assumir dadas as probabilidades de aparição de cada alelo. O caso $D' = 1$ é conhecido como LD completo, ou seja, quando dois SNPs não foram separados por recombinação. Valores intermediários de D' são difíceis de serem interpretados.

Medida r^2

A medida r^2, também denotada por Δ^2, é apontada por Hill e Weir [54] como sendo a mais utilizada para análise de LDs. É obtida pela fórmula:

$$r^2 = \frac{D^2}{P_A P_a P_B P_b}$$

O caso $r^2 = 1$, conhecido como LD perfeito, acontece se e somente se os marcadores não foram separados por recombinação e tem a mesma freqüência alélica.

A medida r^2 tem sido muito utilizada para se definir o que são LDs úteis [70]. De fato, o aumento do número de amostras em estudos de associação tem custo alto, e aumentar o número de amostras para compensar LDs fracos é praticamente inviável. LDs com $r^2 \geq 1/3$ são considerados como úteis em processos de mapeamentos.

Medida δ

A medida δ é derivada da medida $\delta*$, usada em epidemiologia. Esta medida é muito robusta para análise de amostragens de conjuntos de indivíduos onde haplótipos doentes são mais freqüentes do que o observado na população (amostragens de casos de controle).

A medida é dada por:

$$\delta = \frac{D}{P_a P_{AB}}$$

Neste caso, considera-se que dos dois loci analisados, um é o locus de interesse, com A sendo o alelo normal e a sendo o alelo causador da doença, e o segundo locus é utilizado como marcador com alelos B e b. Para doenças raras com haplótipos amostrados aleatoriamente, temos que $\delta = \delta* = D'$.

Medida d

A medida d também é muito utilizada em epidemiologia, e recomendada em Kaplan e Weir [67] para estudos de LD quando são usadas amostragens de casos de controle.

A medida é dada pela fórmula:

$$d = \frac{D}{P_a P_A}$$

Assim como na medida δ, considera-se que dos dois loci analisados, um é o locus que se pretende estudar (com A sendo o alelo normal e a sendo o alelo causador da doença), e o segundo locus é utilizado como marcador (com alelos B e b).

Como D' e r^2 são as mais utilizadas e conhecidas, decidimos por limitar nosso estudo a estas.

5.3 LDs Múltiplos

Dada uma região genômica contígua contendo SNPs marcados (que doravante chamaremos de *contig*), podemos construir um grafo [121] onde cada vértice representa um SNP. Se dois SNPs definem um LD então os vértices correspondentes são ligados por uma aresta. A Figura 5.1 nos apresenta um exemplo representando 16 SNPs.

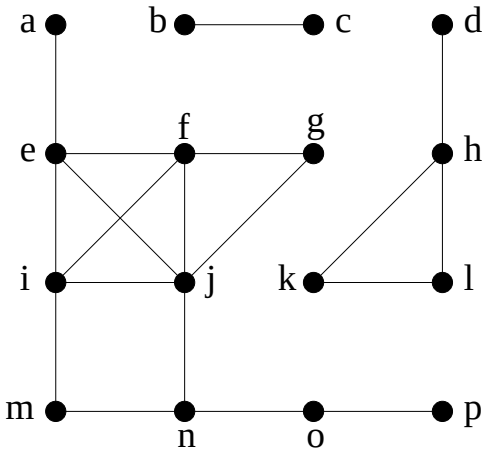

Figura 5.1: Um exemplo de grafo representando 16 SNPs. Cada vértice representa um SNP, e se dois SNPs definem um LD então os vértices correspondentes são ligados por uma aresta.

Definimos tamanho de um grafo como o número de vértices no grafo. Dizemos que o vértice a é vizinho do vértice b caso exista uma aresta ligando a e b. Definimos vizinhança

de um vértice como o conjunto de vértices que são vizinhos ao vértice. Definimos grau de um vértice como a cardinalidade de sua vizinhança, ou em outras palavras, como o número de vértices vizinhos. Definimos caminho como uma seqüência de vértices tal que para cada vértice na seqüência, exceto o último, exista uma aresta que o ligue com seu sucessor. Seja C um conjunto. Dizemos que o conjunto $S1 \subset C$ é maximal, em C, para a propriedade P se $S1$ atende P e não existe um conjunto $S2 \neq S1$ tal que $S1 \subset S2 \subset C$ e $S2$ atende P. Dizemos que $S1 \subset C$ é máximo, em C, para a propriedade P se $S1$ atende P e não existe um conjunto $S2 \subset C$ tal que $S2$ atende P e a cardinalidade de $S2$ é maior que a cardinalidade de $S1$.

Uma componente conexa de um grafo é um conjunto maximal de vértices tal que existe pelo menos um caminho que liga cada par de vértices. Por exemplo, no grafo da Figura 5.1 temos três componentes conexas, a saber: $\{a, e, f, g, i, j, m, n, o, p\}$, $\{b, c\}$ e $\{d, h, k, l\}$. Uma clique em um grafo é um conjunto de vértices tal que cada vértice é vizinho de todos os outros do conjunto. São exemplos de clique no grafo da Figura 5.1: $\{b, c\}$, $\{h, k, l\}$, $\{f, g, j\}$ e $\{e, f, i, j\}$. Observe que todas estas são cliques maximais e que $\{e, f, i, j\}$ é a clique máxima no grafo. Definimos o LD múltiplo de um SNP como a clique máxima, de tamanho maior ou igual a três, que contém o SNP, ou em outras palavras, o LD múltiplo do SNP x é o maior conjunto S de SNPs tal que $x \in S$ e existe um LD entre todo par de elementos de S.

Os LDs (simples) podem ser empregados como marcadores de regiões cromossômicas. Os LDs múltiplos permitem que qualquer par de seus SNPs sejam empregados como marcadores de regiões cromossômicas.

Na Seção 5.3.1 apresentamos como calcular os LDs múltiplos de um dado *contig*.

5.3.1 Heurística para definição de LDs múltiplos

Dado um *contig*, construímos seu respectivo grafo. Calculamos as componentes conexas e, então, passamos a buscar pelos LDs múltiplos para cada SNP.

Objetivando encontrar o LD múltiplo para cada SNP, realizamos o seguinte procedimento. Para cada componente conexa, buscamos pela maior clique. Definimos esta como a maior clique para cada um de seus vértices. Em seguida, passamos a buscar pela maior clique dos demais vértices na componente conexa.

A busca por cliques máximas em grafos é um problema NP-Difícil [68], ou seja, é um problema para o qual não é conhecido um algoritmo eficiente em termos computacionais. Sendo assim, decidimos por adotar uma heurística gulosa baseada no grau dos vértices no procedimento de busca por clique máxima em componentes conexas. A busca é realizada da seguinte forma:

- Eliminamos os vértices de grau um. Esta etapa é repetida diversas vezes até que

todos os vértices de grau um sejam eliminados, uma vez que a remoção de um vértice de grau um pode gerar um novo. Este procedimento irá contribuir com a redução no número de possibilidades a verificar. Observe que esta operação não afeta a busca por cliques de tamanho maior ou igual a três (LDs múltiplos).

- Criamos uma lista de vértices ordenada pelos seus respectivos graus. Esta lista é utilizada, mais uma vez, na redução do número de possibilidades. Observe que, se buscamos por uma clique de tamanho n, um vértice de grau menor que $n-1$ não pode estar presente.

- Seja d o maior grau entre os vértices. Começamos por buscar uma clique com $d+1$ vértices durante t segundos. Procuramos por todos os vértices com grau maior ou igual a d (a lista ordenada por graus facilita a busca). Digamos que foram encontrados n vértices. Se $n \geq d+1$ então há possibilidade de encontrarmos a clique e assim fazemos as combinações destes n vértices em grupos de $d+1$ elementos. Para cada grupo de $d+1$ SNPs, verificamos se é clique. Na primeira resposta positiva o procedimento é encerrado retornando a clique corrente. Caso esta busca por uma clique, com $d+1$ vértices, exceda t segundos o procedimento é encerrado sem nada retornar. Caso não seja encontrada a clique com $d+1$ vértices, passamos a buscar por uma clique com d vértices novamente durante t segundos e assim sucessivamente até encontrar uma clique. Em último caso, uma clique com dois vértices é encontrada. O limite de tempo nas buscas serve para limitar o processo. Observe que cada vez que o procedimento atinge o limite de tempo a busca torna-se mais flexível, cliques menores são buscadas.

Na Figura 5.2 apresentamos o resultado da aplicação do primeiro passo do algoritmo na maior componente conexa do exemplo que exibimos na Figura 5.1. Pelo segundo passo temos a seguinte lista de vértice:

$$j \quad f \quad i \quad e \quad g \quad m \quad n$$

com seus respectivos graus:

$$5 \quad 4 \quad 4 \quad 3 \quad 2 \quad 2 \quad 2\,.$$

No terceiro passo começamos a buscar por uma hipotética clique com 6 vértices, uma vez que o maior grau é 5. Como apenas um dos vértices tem grau maior ou igual a 5 então podemos afirmar que esta clique não existe. Passamos a buscar por uma clique de tamanho 5. Também não é possível encontrar. Há apenas três vértices com grau maior ou igual a 4. Passamos então a buscar uma clique de tamanho 4. Desta vez há possibilidade,

pois temos quatro vértices (e, f, i e j) com grau maior ou igual a 3. Esta é realmente uma clique e assim encerramos nossa busca. Observe que neste caso tivemos apenas uma possibilidade a verificar, mas isto nem sempre ocorre. Por exemplo, caso estivéssemos procurando por uma clique de tamanho 3 teríamos $\frac{7!}{3!4!} = 35$ possibilidades a analisar, pois há sete vértice com grau maior ou igual a 2.

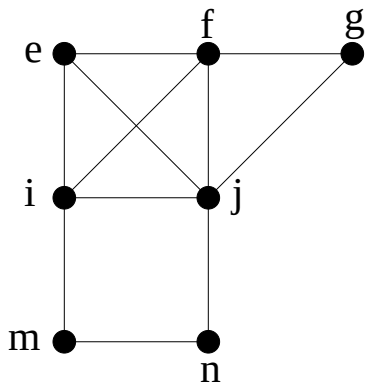

Figura 5.2: Exemplo de grafo obtido após aplicação do primeiro passo do algoritmo de busca pela clique máxima na maior componente conexa do de um grafo. O resultado acima foi obtido a partir do exemplo da Figura 5.1.

Uma vez encontrada a maior clique na componente conexa, um procedimento semelhante é usado para buscar pela maior clique de cada um dos vértices restantes. A diferença está no fato de termos um elemento fixo na clique e os testes, na busca por clique, limitam-se ao elemento e seus vizinhos. Como adotamos uma heurística, nosso algoritmo não garante que a maior clique de cada vértice foi encontrada.

Interpretamos os vértices cuja maior clique possui tamanho dois como LDs simples e os vértices cuja maior clique possui tamanho um como SNPs isolados.

5.4 Fontes de dados

Nesta seção iremos apresentar as duas fontes de dados que utilizamos para identificação de SNPs e de LDs múltiplos. A primeira, proveniente do projeto SUCEST [111], consiste num conjunto de genes de cana-de-açúcar. A segunda, proveniente da base de dados mantida pelo NCBI [85], consiste num conjunto de genes extraídos dos cromossomo 6 humano.

5.4.1 Dados do genoma da cana-de-açúcar

ESTs de cana-de-açúcar com SNPs anotados foram extraídos da base de dados do projeto SUCEST [111]. Inicialmente um conjunto de 291689 ESTs foi produzido. Este conjunto é composto por seqüências com um tamanho médio de $829.44 \pm 182,60$ bp com qualidade média de 23.15 ± 15.71. Posteriormente as seqüências genéticas foram agrupadas em *clusters* utilizando o pacote `cap3` [61]. Foram gerados 43141 *clusters* onde 16338 são *singlets* (*clusters* formados por um único EST).

Descrição do conjunto inicial

O método de identificação dos polimorfismos no SUCEST está descrito em Grivet *et al.* 2001 [47] e Grivet *et al.* 2003 [48]. A detecção de polimorfismos em cada *cluster* foi feita em dois passos: inicialmente define-se como SNP uma posição onde o alelo menos frequente aparece no mínimo duas vezes na seção transversal do alinhamento, com qualidade superior ou igual a 20. O segundo passo consiste em filtrar os SNPs, mantendo apenas as posições cuja vizinhança de 10 bases (5 para cada lado) esteja perfeitamente alinhada com todos os outros ESTs do *cluster*.

Para cada *cluster*, o projeto anotou as posições de SNPs observados, as bases observadas nos ESTs alinhados e suas respectivas freqüências. As freqüências dos alelos nas posições de SNPs foram utilizadas para o cálculo das métricas de LD descritas na Seção 5.2.

Validação dos SNPs

Ao todo foram obtidos 8198 arquivos representando *clusters* (um *cluster* por arquivo), com 43029 posições de SNPs anotados. Para validar os dados obtidos, todos os *clusters* foram processados pelo `polybayes` [77], utilizando sua configuração padrão. Os arquivos phd necessários para execução do programa foram gerados a partir das seqüências em formato `fasta` e `qual` dos ESTs. Para cada *cluster*, comparamos o conjunto de SNPs obtidos pelo `polybayes`, que chamaremos de PB, e o conjunto de SNPs mapeados pelo projeto SUCEST, que chamaremos de SC. Os resultados foram agrupados da seguinte forma:

- *clusters* onde $PB = SC$
- *clusters* onde $PB \supset SC$ (onde $PB \neq \emptyset$)
- *clusters* onde $PB \subset SC$
- *clusters* onde $PB \cap SC \neq \emptyset$ (onde $PB \not\supset SC$ e $PB \not\subset SC$)

Análise dos SNPs da cana-de-açúcar					
		Média/Cluster			
	Clusters	$PB \cap SC$	$SC \setminus PB$	$PB \setminus SC$	Total
$PB = SC$	181	1.8	0.0	0.0	1.8
$PB \supset SC$	6310	4.6	0.0	12.1	16.7
$PB \subset SC$	35	1.9	1.8	0.0	3.7
$PB \cap SC \neq \emptyset$	1261	7.8	2.2	17.4	27.4
$PB \cap SC = \emptyset$	254	0.0	2.0	5.2	7.2
$PB = \emptyset$	157	0.0	2.6	0.0	2.6
Total	8198	4.8	0.5	12.1	17.4

Tabela 5.1: Comparação entre resultados obtidos pelo projeto SUCEST e `polybayes`. As colunas $PB \cap SC$, $SC \setminus PB$ e $PB \setminus SC$ representam respectivamente SNPs que pertencem tanto ao conjunto SC quanto ao PB, apenas a SC e apenas a PB.

- *clusters* onde $PB \cap SC = \emptyset$ (onde $PB \neq \emptyset$)
- *clusters* onde $PB = \emptyset$

Os resultados estão sumarizados na Tabela 5.1.

A grande maioria de *clusters* com $PB \supset SC$ deve-se ao fato do projeto SUCEST ter removido SNPs cuja vizinhança de tamanho 10 não tivesse qualidade mínima de 20 e cujo alinhamento no *cluster* não fosse perfeito. Observando as seções transversais de posições onde `polybayes` não detectou SNPs marcados pelo projeto SUCEST, observamos que a grande maioria tem baixa cobertura, e possui apenas duas bases polimórficas.

Para efetuar as análises posteriores, foi montado um conjunto de *clusters* contendo apenas SNPs marcados tanto por `polybayes` quanto pelo projeto SUCEST. O conjunto contém 7787 *clusters*, contendo 39049 posições de SNPs (90.75% dos SNPs marcados inicialmente).

5.4.2 Dados do genoma humano

Foram escolhidos genes do cromossomo 6, de uma região conhecida como MHC, ou *Major Histocompatibility Complex*. Essa região de aproximadamente 4 Mbp é muito densa em genes, tendo mais de 120 genes, dos quais 40% codificam proteínas relacionadas a funções imunológicas. Mais de 20000 artigos foram escritos nos últimos 30 anos, estudando a correlação dos genes dessa região com doenças [57, 110].

Resumo dos dados coletados por gene				
	Antes da filtragem		**Após filtragem**	
	ESTs	**SNPs**	**ESTs**	**SNPs**
HLA-A	2535	2109	1475	280
HLA-B	2503	336	1740	144
HLA-DOB	42	77	42	41

Tabela 5.2: Número de ESTs e SNPs obtidos para cada gene selecionado da região MHC do cromossomo 6 humano, antes e depois do processo de filtragem. Os ESTs foram obtidos em formato `fasta` e os SNPs em `flat file`.

Descrição do conjunto inicial

Através do site do NCBI [85] foram obtidos três genes, selecionados por se encontrarem em regiões com alta densidade de polimorfismos dentro do MHC: HLA-A (3324 bp, da base 30.018.310 à base 30.021.633 na montagem de referência), HLA-B (3287 bp, da base 31.429.628 à base 31.432.914 na montagem de referência) e HLA-DOB (4236 bp, da base 32.888.527 à base 32.892.762 na montagem de referência).

Para cada um dos genes, foi obtida a lista de ESTs e cDNAs utilizados para efetuar a montagem da seqüência do cromossomo, assim como a lista de SNPs de referência disponíveis na base de dados dbSNP [26] marcados como pertencentes aos genes escolhidos. Os ESTs e cDNAs foram filtrados de forma a se obter um conjunto de seqüências sem bases indefinidas (símbolos N). A lista de SNPs também foi filtrada, eliminando os seguintes elementos:

- INDELs.
- SNPs cujas posições não são compatíveis com os limites dos genes respectivos.
- SNPs onde nenhum alelo corresponde ao alelo da seqüência de referência do gene.
- Posições redundantes: Em alguns casos, foram encontradas variações alélicas diferentes anotadas na mesma posição do genoma. Neste caso as variações foram agrupadas, de forma a que a lista final tivesse posições únicas. Por exemplo: as variações A/C e A/G anotadas na posição 10 são agrupadas, sendo considerada a variação A/C/G.

A Tabela 5.2 sumariza os dados obtidos antes e após a filtragem, para cada gene.

Validação dos SNPs

O primeiro passo para a validação dos SNPs foi a montagem de *clusters* para análise das seções transversais. Não foi possível agrupar todas as seqüências utilizando os pacotes `cap3` e `phrap` (ambos criam mais de um *cluster* ou *contig*). Por isso, optou-se por criar um alinhamento múltiplo usando o algoritmo de alinhamento estrela. O alinhamento estrela alinha cada EST com uma seqüência de referência (nesse caso, a seqüência do gene completo) e propaga os buracos inseridos nos outros alinhamentos gerados.

Os ESTs obtidos após a filtragem foram alinhados um a um com o gene, usando o algoritmo clássico semi-global e pontuação $match = 1$, $mismatch = -2$, $opengap = -10$ e $extendgap = 0$. Esta pontuação foi escolhida por ter sido definida como a mais apropriada para efetuar alinhamentos de ESTs com DNA no estudo realizado no Capítulo 3. Para cada EST, foi efetuado o alinhamento da seqüência original com o gene e do complemento com o gene, escolhendo aquele que gerasse a melhor pontuação.

Os *clusters* gerados foram processados com `polybayes`, para validação da lista de polimorfismos obtidos através do site do dbSNP do NCBI. Para ser considerado como um SNP válido, considerou-se que uma posição deveria ser marcada pelo `polybayes` e a variação alélica menos freqüente deveria aparecer no mínimo duas vezes e representar 1% ou mais de todas as bases na seção transversal do alinhamento.

Analisando os alinhamentos dos *clusters* utilizando o critério acima, detectamos que 462 colunas são preservadas no gene HLA-A, 377 colunas são preservadas no gene HLA-B, e 24 colunas são preservadas no gene HLA-DOB.

Como os arquivos de qualidade dos ESTs não estão disponíveis, foi necessário atribuir um valor de qualidade à cada base. Um estudo foi realizado para determinar o valor de qualidade que produz o melhor resultado com `polybayes`: valores muito baixo podem gerar muitos falsos negativos, enquanto que valores muito altos tendem a gerar muitos falsos positivos.

O estudo foi feito com o *cluster* do gene HLA-DOB, por ser o menor e portanto mais rápido de ser processado. Os valores de qualidade atribuidos foram 15, 20, 25, 30, 35 e 40. Para analisar os resultados, os seguintes aspectos foram observados:

- Número de polimorfismos anotados pelo `polybayes`.
- Número de polimorfismos anotados pelo `polybayes` após remoção de INDELs e de seções transversais contendo buracos na âncora.
- Número de posições que foram anotadas tanto pelo NCBI quanto pelo `polybayes` cujas variações alélicas são idênticas.
- Número de posições que foram anotadas tanto pelo NCBI quanto pelo `polybayes` cujas variações alélicas são diferentes (por exemplo AG e GT).

Sensibilidade de `polybayes` em função da qualidade

Qualidade	Total	Filtrado	Idênticos	Diferentes	Não encontrados
15	24	11	6	0	35
20	37	23	6	0	35
25	38	24	6	0	35
30	242	84	8	1	32
35	335	161	9	1	31
40	355	181	9	1	31

Tabela 5.3: Número de SNPs obtidos pelo `polybayes` com diferentes qualidades de base atribuidas aos ESTs do gene HLA-DOB: número de polimorfimos anotados pelo `polybayes`, número de polimorfismos anotados pelo `polybayes` após remoção de INDELs, posições idênticas às marcadas pela base de referência (NCBI) com alelos iguais, posições idênticas às marcadas pelo NCBI com alelos diferentes, posições marcadas pelo NCBI e não marcadas pelo `polybayes`.

Tempo de execução de `polybayes`

	HLA-A	HLA-B	HLA-DOB
25	1h16m33s	2h30m39s	0h00m07s
30	2h36m27s	4h06m17s	0h00m32s

Tabela 5.4: Tempo de execução do `polybayes` em função do gene e da qualidade atribuída às bases dos ESTs.

- Número de posições que foram anotadas pelo NCBI e que o `polybayes` não identificou.

Os resultados obtidos pelo `polybayes` estão sumarizados na Tabela 5.3. Pode-se observar que com qualidade maior ou igual a 30, `polybayes` gera um número muito maior de SNPs, mas que o número de posições que são anotados também no NCBI não aumenta praticamente nada, e o número de SNPs anotados no NCBI não detectados pelo `polybayes` sofre uma variação muito pequena.

A Tabela 5.4 mostra a diferença de tempo de execução do `polybayes` por gene e por qualidade atribuída às bases dos ESTs. É interessante observar que a variação de qualidade influi bastante no tempo de execução, provavelmente forçando o software a analisar um maior número de possibilidades.

Optou-se por trabalhar com qualidade 25. Executamos `polybayes` nos outros genes. Os resultados estão sumarizados na Tabela 5.5. Os resultados mostram que `polybayes` encontra muito mais SNPs do que os anotados no dbSNP.

Análise dos resultados de `polybayes`

	Total	Filtrado	Idênticos	Diferentes	Não encontrados
HLA-A	640	340	79/79	28/28	1/173
HLA-B	757	387	42/42	7/9	4/93
HLA-DOB	38	24	6/6	0/0	0/35

Tabela 5.5: Resumo da validação de SNPs nos genes selecionados do cromossomo 6 humano. Na primeira coluna temos o gene, na segunda o número de polimorfismos anotados pelo `polybayes` sem filtros, na terceira o número de polimorfismos anotados pelo `polybayes` filtrando INDELs. As três últimas são comparações com a base de referência (NCBI): posições idênticas com alelos iguais, posições idênticas com alelos diferentes, posições marcadas pela base de referência e não marcadas pelo `polybayes`. As três últimas colunas mostram dois valores X/Y. O valor Y representa os resultados obtidos sem filtros e o valor X representa os resultados obtidos depois da aplicação de filtros. Os filtros consideram apenas posições cuja variação alélica menos freqüente apareça pelo menos duas vezes e represente no mínimo 1% de todas as bases da seção transversal do alinhamento.

Isso ocorre pelo fato do `polybayes` fazer uma análise computacional, baseada puramente na qualidade das bases e a profundidade de cada seção transversal. Além disso, os SNPs anotados no dbSNP são validados experimentalmente, removendo eventuais falsos positivos. Observa-se também que, para os genes analisados, todos os polimorfimos validados tanto pelo NCBI quanto pelo `polybayes` (com variações alélicas iguais ou não) correspondem aos critérios de SNP definidos anteriormente. Das posições anotadas pelo NCBI mas não encontradas pelo `polybayes`, apenas uma no gene HLA-A e quatro no gene HLA-B correspondem aos critérios, nos lotes gerados.

Os resultados cujas posições foram marcadas tanto por `polybayes` quanto por NCBI mas cujas variações alélicas são diferentes foram analisados de forma mais detalhada. Os seguintes casos foram observados:

- $NCBI \cap PB = \emptyset$: as bases anotadas pelo NCBI não correspondem às bases observadas pelo `polybayes`. Exemplo: AC e GT.
- $NCBI \supset PB$: as bases anotadas pelo NCBI contém as bases anotadas pelo `polybayes` e mais outras. Exemplo: ACG e CG.
- $NCBI \subset PB$: as bases anotadas pelo `polybayes` contém as bases anotadas pelo NCBI e mais outras. Exemplo: ACT e AT.
- $NCBI \cap PB \neq \emptyset$: as bases anotadas pelo `polybayes` e as bases anotadas pelo NCBI

Comparação das posições marcadas por NCBI e polybayes que apresentam alelos distintos

	$NCBI \cap PB = \emptyset$	$NCBI \supset PB$	$NCBI \subset PB$	$NCBI \cap PB \neq \emptyset$
HLA-A	0/0	20/20	4/4	4/4
HLA-B	0/0	2/2	4/4	1/3
HLA-DOB	0/0	0/0	0/0	0/0

Tabela 5.6: Análise das posições anotadas tanto por NCBI quanto por `polybayes` com variações alélicas diferentes. As colunas mostram dois valores X/Y. O valor Y representa os resultados obtidos sem filtros e o valor X representa os resultados obtidos depois da aplicação de filtros. Os filtros consideram apenas posições cuja variação alélica menos freqüente apareça, pelo menos, duas vezes e represente, pelo menos, 1% de todas as bases presentes na seção transversal do alinhamento.

tem uma intersecção comum, mas nenhuma variação contém a outra. Exemplo: CT e TG.

Os resultados dessa análise estão sumarizados na Tabela 5.6. É interessante notar que em nenhum dos casos os conjuntos anotados pelo NCBI e pelo `polybayes` são totalmente disjuntos.

5.5 LDs múltiplos no Projeto SUCEST

Os ESTs de cana-de-açúcar com SNPs anotados foram extraídos da base de dados do projeto SUCEST, de acordo com a explicação da Seção 5.4.1. Um estudo sobre SNPs foi realizado nestes *clusters* e foram catalogados SNPs em 8198 *contigs*. Sobre estes SNPs, estudamos LDs múltiplos.

Inicialmente selecionamos apenas os SNPs bialélicos. Depois selecionamos os *contigs* onde havia pelo menos dois SNPs validados pelo `polybayes` de acordo com o procedimento descrito na Seção 5.4.1. Com isto, nos restaram 6178 *contigs*. Nestes dados aplicamos o procedimento descrito na Seção 5.3.1.

Na Seção 5.5.1 definimos as métricas e os limiares utilizados para definir LD. Resultados de testes com diferentes configurações do parâmetro t do algoritmo são mostrados na Seção 5.5.2. Finalmente, na Seção 5.5.3 apresentamos os resultados obtidos com a aplicação de nosso algoritmo em dados oriundos do Projeto `SUCEST`.

5.5.1 Definição de LD

Durante o procedimento de seqüenciamento de material genético é atribuído um valor de qualidade a cada base seqüenciada. Este número relaciona-se com a probabilidade daquela leitura estar correta. Podemos definir um limiar e classificar as bases como sendo de baixa qualidade, caso tenha valor inferior ao limiar, ou de alta qualidade, caso contrário.

Anteriormente dissemos que uma aresta é criada sempre que dois SNPs formam um LD. Mas quando considerar que dois SNPs formam um LD? Das diversas métricas existentes, consideramos apenas duas: D' e r^2. Em nosso estudo, fizemos uma análise de quatro critérios distintos:

- Consideramos LD apenas quando $D' = 1$ e as bases de baixa qualidade foram desprezadas nos cálculos.
- Consideramos LD apenas quando $D' = 1$ e as bases de baixa qualidade foram utilizadas nos cáculos tal como as de alta qualidade.
- Consideramos LD apenas quando $r^2 \geq 1/3$ e as bases de baixa qualidade foram desprezadas nos cálculos.
- Consideramos LD apenas quando $r^2 \geq 1/3$ e as bases de baixa qualidade foram utilizadas nos cálculos tal como as de alta qualidade.

Observe que em cada um dos casos temos, possivelmente, grafos distintos.

5.5.2 Verificando limite para o tempo de busca por clique

A Tabela 5.7 nos mostra o tempo total e o tempo médio por *contig* gasto na busca por LDs múltiplos utilizando limite de tempo para busca por cliques máximas de 5 e 60 segundos. Mostra-nos também o quão mais lenta foi a busca com limite de 60 segundos. Em geral, usar a métrica r^2 é cerca de 10 vezes mais rápido do que usar a métrica D. Observe que, na comparação entre $t = 5$ e $t = 60$, o tempo, quando utilizamos r^2 como métrica, cresce aproximadamente 100% enquanto que quando utilizamos D' o tempo é 5 a 6 vezes o tempo inicial. Apresentamos os valores para cada situação: $r^2 \geq 1/3$ eliminando bases de baixa qualidade, $r^2 \geq 1/3$ considerando bases de baixa qualidade, $D' = 1$ eliminando bases de baixa qualidade e $D' = 1$ considerando bases de baixa qualidade. A seguir uma descrição detalhada de cada coluna da tabela:

- A coluna **Métrica** indica a situação a que se referem os valores. O símbolo * indica que consideramos as bases de baixa qualidade.

- A coluna **T5** indica o tempo total gasto em segundos para calcular os LDs múltiplos de todos os *contigs* com limite de tempo de 5 segundos para a busca por clique de determinado tamanho.
- A coluna **T5/c** apresenta o tempo médio gasto por *contig* para cálculo de LDs múltiplos quando limitamos o tempo de busca em 5 segundo.
- As colunas **T60** e **T60/c** são semelhantes a **T5** e **T5/c** só que a limitação do tempo de busca agora é 60 segundos.
- Na coluna **T60/T5** é exibido o valor da divisão de **T60** por **T5**.

Métrica	**T5**	**T5/c**	**T60**	**T60/c**	**T60/T5**
$r^2 \geq 1/3$	325.31	0.05	740.99	0.12	2.27
$r^2 \geq 1/3$ *	345.61	0.06	750.85	0.12	2.17
$D' = 1$	2110.83	0.34	10386.84	1.68	4.92
$D' = 1$ *	3285.97	0.53	19824.78	3.21	6.03

Tabela 5.7: Comparação no tempo total de execução, em segundos, utilizando configurações distintas para o parâmetro t, do programa para busca por LDs múltiplos nos dados do `SUCEST`. Foram realizados testes com $t = 5$ e $t = 60$. O parâmetro t define o tempo máximo, em segundos, de busca por uma clique de determinado tamanho. Maiores informações podem ser encontradas no texto.

Realizamos estes dois testes, com limite de $t = 5$ e $t = 60$ segundos, para avaliar se $t = 5$ segundos é suficiente para respostas satisfatórias. Para realizar a comparação dos resultados observamos parâmetros relativos às componentes conexas e outros relativos a vértices (SNPs).

Quanto a componentes conexas, foram três os parâmetros observados: tamanho da maior clique, número de cliques de tamanho 1 e número de cliques de tamanho 2. Não foram observadas diferenças em qualquer dos quatro casos abordados.

Quanto a SNPs, observamos dois parâmetros: maior clique possível (de acordo com seu grau e os graus dos vizinhos) e tamanho da clique encontrada. Obviamente, o primeiro parâmetro não varia. Já no segundo, percebemos um pequeno número de variações. O maior número concentrou-se nos casos relativos a métrica D'.

Ressaltamos aqui os números. Observe que foram comparados os resultados de 6178 *contigs*, com um total de 39608 SNPs, em quatro situações distintas. Não foram encontradas diferenças em relação às componentes conexas. Em apenas 507 casos, de um total de 158432, tivemos um incremento no tamanho da clique. Note que 507 refere-se a apenas 0.32% dos casos.

Chegamos a conclusão de que $t = 5$ segundos é suficiente para apresentar resultados satisfatórios.

5.5.3 Resultados

Doravante chamaremos as componentes conexas do grafo de "grupos de SNPs relacionados indiretamente", e as cliques de "grupos de SNPs relacionados diretamente".

As Figuras 5.3, 5.4, 5.5 e 5.6 nos apresentam gráficos comparativos das quatro situações estudadas.

No gráfico da Figura 5.3, para cada *contig* montamos o grafo e contamos as componentes conexas. Depois agrupamos os *contigs* pelo número de componentes conexas. O número de *contigs* é cumulativo. As médias obtidas foram 1.38 ± 0.66, 1.32 ± 0.60, 2.54 ± 2.18 e 2.38 ± 1.95 respectivamente para os casos $D' = 1$, $D' = 1*$, $r^2 \geq 1/3$ e $r^2 \geq 1/3*$. Neste grafo, podemos observar que, utilizando $D' = 1$ como parâmetro para definir LD, há uma maior tendência a criação de componentes conexas maiores, o que leva a um menor número de componentes conexas.

No gráfico da Figura 5.4, para cada *contig* montamos o grafo e calculamos a maior clique. Depois agrupamos os *contigs* pela tamanho da maior clique. O número de *contigs* é cumulativo. As médias obtidas foram 4.34 ± 1.96, 4.38 ± 1.91, 3.25 ± 1.79 e 3.34 ± 1.88 respectivamente para os casos $D' = 1$, $D' = 1*$, $r^2 \geq 1/3$ e $r^2 \geq 1/3*$. Neste gráfico, podemos observar que, utilizando $D' = 1$ como parâmetro para definir LD, há uma maior tendência a formação de cliques maiores.

No gráfico da Figura 5.5, para cada *contig* montamos o grafo e contamos as componentes conexas com apenas um vértice. Depois agrupamos os *contigs* pelo número de componentes unitárias. O número de *contigs* é cumulativo. As médias obtidas foram 0.26 ± 0.58, 0.23 ± 0.55, 1.18 ± 1.67 e 1.08 ± 1.54 respectivamente para os casos $D' = 1$, $D' = 1*$, $r^2 \geq 1/3$ e $r^2 \geq 1/3*$. Neste gráfico, podemos observar que, utilizando $D' = 1$ como parâmetro para definir LD, há uma tendência menor ao isolamento de vértices.

No gráfico da Figura 5.6, para cada *contig* montamos o grafo e contamos os vértices que possuem clique máxima de tamanho dois. Depois agrupamos os *contigs* pelo número vértices com clique máxima de tamanho dois. O número de *contigs* é cumulativo. As médias obtidas foram 0.80 ± 1.13, 0.76 ± 1.09, 1.71 ± 2.08 e 1.64 ± 2.02 respectivamente para os casos $D' = 1$, $D' = 1*$, $r^2 \geq 1/3$ e $r^2 \geq 1/3*$. Neste gráfico, podemos observar que, utilizando $D' = 1$ como parâmetro para definir LD, há uma tendência menor a formação de cliques de tamanho dois.

Os gráficos apresentados nas Figuras 5.3, 5.4, 5.5 e 5.6 nos mostram que as métricas D' e r^2 , com limiares 1 (SNP completo) e 1/3 (SNP útil) respectivamente, possuem variações mas a tendência é a mesma.

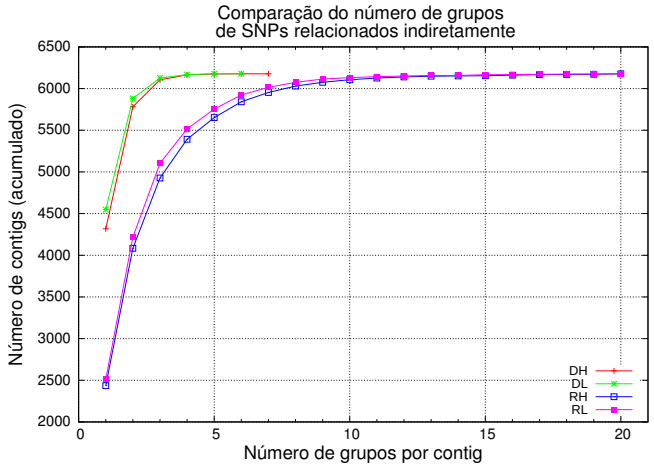

Figura 5.3: Gráfico comparando o número de grupos de SNPs relacionados indiretamente nos diversos casos estudados. No eixo X temos o número de componentes conexas (grupos de SNPs relacionados indiretamente) por *contig* e no eixo Y o número de *contigs*. A curva DH representa o caso onde ocorre LD quando $D' = 1$ e bases de baixa qualidade são excluídas. A curva DL, quando $D' = 1$ e bases de baixa qualidade são consideradas. A curva RH representa o caso onde ocorre LD quando $r^2 \geq 1/3$ e bases de baixa qualidade são excluídas. A curva RL, quando $r^2 \geq 1/3$ e bases de baixa qualidade são consideradas.

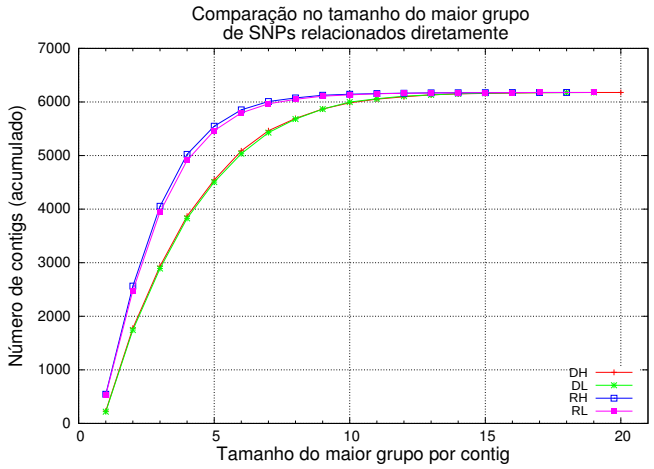

Figura 5.4: Gráfico comparando o tamanho do maior grupo de SNPs relacionados diretamente nos diversos casos estudados. No eixo X temos o tamanho da maior clique (grupo de SNPs relacionados diretamente) por *contig* e no eixo Y o número de *contigs*. A curva DH representa o caso onde ocorre LD quando $D' = 1$ e bases de baixa qualidade são excluídas. A curva DL, quando $D' = 1$ e bases de baixa qualidade são consideradas. A curva RH representa o caso onde ocorre LD quando $r^2 \geq 1/3$ e bases de baixa qualidade são excluídas. A curva RL, quando $r^2 \geq 1/3$ e bases de baixa qualidade são consideradas.

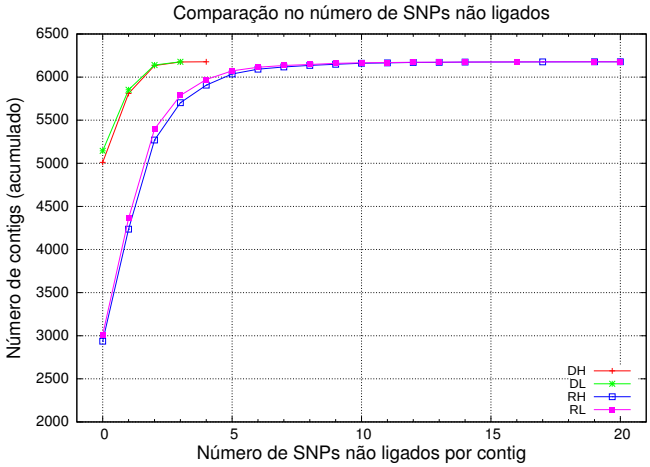

Figura 5.5: Gráfico comparando o número de SNPs não ligados nos diversos casos estudados. No eixo X temos o número de componentes conexas de tamanho um (SNPs não ligados) por *contig* e no eixo Y temos o número de *contigs*. A curva DH representa o caso onde ocorre LD quando $D' = 1$ e bases de baixa qualidade são excluídas. A curva DL, quando $D' = 1$ e bases de baixa qualidade são consideradas. A curva RH representa o caso onde ocorre LD quando $r^2 \geq 1/3$ e bases de baixa qualidade são excluídas. A curva RL, quando $r^2 \geq 1/3$ e bases de baixa qualidade são consideradas.

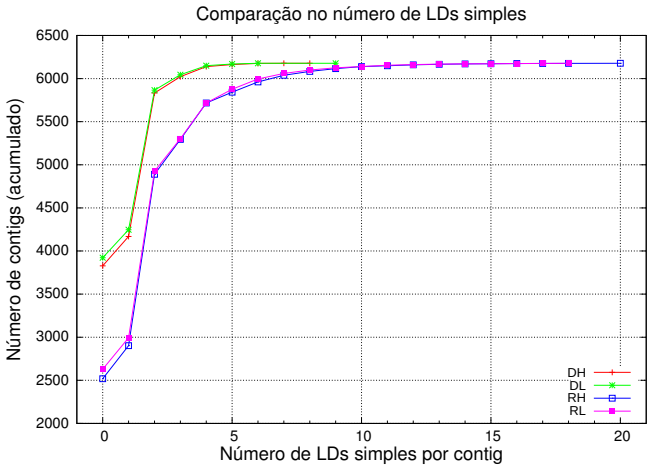

Figura 5.6: Gráfico comparando o número de LDs simples nos diversos casos estudados. No eixo X temos o número de vértices que possuem clique máxima de tamanho dois (LDs simples) por *contig* e no eixo Y temos o número de *contigs*. A curva DH representa o caso onde ocorre LD quando $D' = 1$ e bases de baixa qualidade são excluídas. A curva DL, quando $D' = 1$ e bases de baixa qualidade são consideradas. A curva RH representa o caso onde ocorre LD quando $r^2 \geq 1/3$ e bases de baixa qualidade são excluídas. A curva RL, quando $r^2 \geq 1/3$ e bases de baixa qualidade são consideradas.

Na Tabela 5.8 apresentamos um resumo de valores obtidos, para os quatro casos estudados, nos cálculos de LDs múltiplos nos dados da cana-de-açúcar. Valores para cinco parâmetros são listados, são eles: número de componentes conexas, tamanho da maior clique, número de SNPs isolados, número de LDs simples e média do tamanho das cliques associadas a cada SNP. Em primeiro lugar podemos observar que os valores considerando ou não bases de baixa qualidade são extremamente semelhantes. Em segundo lugar podemos validar tudo o que já havíamos observado nos gráficos. Utilizando a definição de LD completo há uma tendência de formação de menor número de componentes conexas, formação de componentes maiores, menor número de SNPs isolados e LDs simples.

Resultados obtidos nos casos estudados (SUCEST)

Parâmetro	$D' = 1$	$D' = 1*$	$r^2 \geq 1/3$	$r^2 \geq 1/3*$
Componentes Conexas (CCs)	1.4	1.3	2.5	2.4
Maior Clique (MC)	4.3	4.4	3.2	3.3
SNPs Isolados (C1)	0.3	0.2	1.2	1.1
LDs Simples (C2)	0.8	0.8	1.7	1.6
Média do Tamanho das Cliques (MTC)	4.0	4.1	2.8	2.9

Tabela 5.8: Comparação dos resultados dos cálculos de LDs múltiplos, nos dados da cana-de-açúcar no projeto SUCEST, utilizando a definição de LD completo ($D' = 1$) e de LD útil ($r^2 \geq 1/3$). O símbolo "$*$" indica que as bases de baixa qualidade foram consideradas nos cálculos. Na primeira coluna indicamos o parâmetro, na segunda apresentamos os resultados utilizando LD completo sem bases de baixa qualidade, na terceira utilizando LD completo com bases de baixa qualidade. As colunas 4 e 5 apresenta resultados semelhantes às colunas 2 e 3 só que agora utilizando a definição de LD útil. Os parâmetros listados são: número de componentes conexas, tamanho da maior clique, número de SNPs isolados, número de LDs simples e média do tamanho das cliques associadas a cada SNP.

5.6 LDs múltiplos no genoma humano

Aplicamos também o algoritmo que descrevemos na Seção 5.3 aos dados do genoma humano, descritos na Seção 5.4.2. Para testar a heurística que busca por LDs múltiplos, geramos três conjuntos de dados, descritos a seguir:

1. **NCBI Filtrado**: este primeiro conjunto de dados refere-se a um subconjunto do conjunto de SNPs marcados na base de referência (dbSNP). Conforme apresentamos na Tabela 5.9, inicialmente tínhamos 280, 144 e 41 SNPs respectivamente para HLA-A, HLA-B e HLA-DOB. Ao aplicarmos filtros restaram 98, 28 e 10 SNPs. Os filtros removeram SNPs tri ou tetra aléticos, além de remover posições que não pudemos mais considerar SNPs, depois dos filtros que aplicamos inicialmente, pois

não havia mais bases discordantes. Apresentamos os resultados dos cálculos para LDs múltiplos na Tabela 5.10.

2. **Simples**: o segundo conjunto foi obtido observando as seções transversais dos *contigs* que continham variações alélicas. Foram descartados SNPs tri e tetra alélicos e só foram considerados SNPs quando a base de menor freqüência ocorreu, pelo menos, duas vezes e possui, pelo menos, 1% de freqüências dentre as bases da coluna. Apresentamos os resultados dos cálculos para LDs múltiplos na Tabela 5.11.

3. **Intersecção**: o último dos conjuntos de dados é composto pela intersecção dos dois primeiros. Apresentamos os resultados dos cálculos para LDs múltiplos na Tabela 5.12.

Número de SNPs por conjunto de dados

Gene	NCBI	NCBI Filtrado	Simples	Intersecção
HLA-A	280	98	137	37
HLA-B	144	28	95	11
HLA-DOB	41	10	22	6

Tabela 5.9: Lista o número de SNPs em cada um dos conjuntos de dados do genoma humano onde foram calculados os LDs múltiplos. NCBI Filtrado, Simples e Intersecção são os conjuntos de dados. NCBI foi um conjunto inicial descrito na Seção 5.4.2 de onde foi extraído o conjunto NCBI Filtrado.

Comparação dos resultados (NCBI Filtrado)

	$D' = 1$					$r^2 \geq 1/3$				
	CCs	MC	C1	C2	MTC	CCs	MC	C1	C2	MTC
HLA-A	5	28	3	5	15.7	53	11	44	11	3.4
HLA-B	6	6	3	9	3.5	21	2	16	12	1.4
HLA-DOB	3	3	0	3	2.7	7	2	5	5	1.5

Tabela 5.10: Comparação dos resultados dos cálculos de LDs múltiplos, nos genes selecionados do cromossomo 6 humano, utilizando a definição de LD completo ($D' = 1$) e LD útil ($r^2 \geq 1/3$) com o conjunto "NCBI Filtrado". Na primeira coluna indicamos o gene e nos dois blocos que seguem apresentamos os números. O primeiro bloco refere-se ao dados com a definição de LD completo e o segundo a LD útil. Em cada um dos blocos temos os seguintes valores: número de componentes conexas (CC), tamanho da maior clique (MC), número de SNPs isolados (C1), número de LDs simples (C2) e média do tamanho das cliques associadas a cada SNP (MTC).

Comparação dos resultados (Simples)

	$D' = 1$					$r^2 \geq 1/3$				
	CCs	MC	C1	C2	MTC	CCs	MC	C1	C2	MTC
HLA-A	2	18	0	5	9.7	60	10	48	38	3.0
HLA-B	3	15	1	14	9.0	45	5	36	23	2.3
HLA-DOB	1	10	0	1	8.5	3	8	2	5	4.8

Tabela 5.11: Comparação dos resultados dos cálculos de LDs múltiplos, nos genes selecionados do cromossomo 6 humano, utilizando a definição de LD completo ($D' = 1$) e LD útil ($r^2 \geq 1/3$) para o conjunto "Simples". Na primeira coluna indicamos o gene e nos dois blocos que seguem apresentamos os números. O primeiro bloco refere-se ao dados com a definição de LD completo e o segundo a LD útil. Em cada um dos blocos temos os seguintes valores: número de componentes conexas (CC), tamanho da maior clique (MC), número de SNPs isolados (C1), número de LDs simples (C2) e média do tamanho das cliques associadas a cada SNP (MTC).

Assim como os resultados com a cana-de-açúcar pudemos observar uma tendência a formação de um menor número de componentes conexas, formação de componentes maiores e criação de um menor número de SNPs isolados e LDs simples quando utilizamos a definição de LD completo.

5.7 Conclusão

Neste capítulo, apresentamos uma definição de LDs múltiplos fundamentada em teoria dos grafos, onde chamamos de LD múltiplo de um SNP a clique máxima que o contém. Apresentamos um algoritmo que, utilizando uma heurística gulosa baseada no grau dos vértices, busca pelos LDs múltiplos. Implementamos uma heurística pois a busca por cliques máximas é um problema NP-Difícil. Este algoritmo não garante que a clique máxima será encontrada, mas ótimos resultados foram obtidos nos conjuntos de dados estudados.

O projeto SUCEST, financiado pela FAPESP, obteve um conjunto de seqüências de ESTs de cana-de-açúcar e realizou uma série de trabalhos com estas seqüências. Dentre estes trabalhos, foram catalogados uma série de SNPs nos *clusters* obtidos. Aplicamos este conjunto de dados ao algoritmo descrito neste trabalho utilizando quatro variações.

Na primeira variação removemos as bases de baixa qualidade e definimos LD como todo par de SNP cujo valor para métrica D' seja igual a 1. Na segunda, fizemos a mesma definição para LD só que consideramos as bases de baixa. A terceira e quarta variações são semelhantes a primeira e segunda, diferenciando apenas na definição de LD. Nestas é considerado LD o par de SNPs cujo valor para a métrica r^2 seja maior ou igual a 1/3.

Comparação dos resultados (Intersecção)

	$D' = 1$					$r^2 \geq 1/3$				
	CCs	MC	C1	C2	MTC	CCs	MC	C1	C2	MTC
HLA-A	2	12	1	2	8,6	17	7	11	2	3.2
HLA-B	3	4	0	4	3,1	8	2	5	6	1.5
HLA-DOB	2	3	1	1	2,5	4	2	3	3	1.5

Tabela 5.12: Comparação dos resultados dos cálculos de LDs múltiplos, nos genes selecionados do cromossomo 6 humano, utilizando a definição de LD completo ($D' = 1$) e LD útil ($r^2 \geq 1/3$) para o conjunto "Intersecção". Na primeira coluna indicamos o gene e nos dois blocos que seguem apresentamos os números. O primeiro bloco refere-se ao dados com a definição de LD completo e o segundo a LD útil. Em cada um dos blocos temos os seguintes valores: número de componentes conexas (CC), tamanho da maior clique (MC), número de SNPs isolados (C1), número de LDs simples (C2) e média do tamanho das cliques associadas a cada SNP (MTC) .

Na literatura, a primeira definição é chamada de LD completo e a segunda de LD útil.

Realizamos testes também com dados da região MHC do cromossomo 6 do genoma humano, extraídos do NCBI. Essa região é caraterizada por uma grande concentração de genes e SNPs e pelo fato de uma quantidade considerável destes genes estarem relacionados ao sistema imunológico. Obtivemos resultados semelhantes aos obtidos com os dados da cana-de-açúcar.

Comparando os resultados que apresentamos, chegamos a conclusão de que a definição de LD completo é mais adequada para o cálculo de LDs múltiplos. Tal conclusão deveu-se a dois fatores, são eles: a métrica apresentou uma maior capacidade de agrupamento e uma menor tendência ao isolamento de SNPs ou formação de LDs simples. Em contrapartida, a definição de LD útil permite um melhor desempenho da heurística, já que o grafo resultante possui menos arestas.

Capítulo 6

Conclusão

Ao longo deste trabalho, estudamos três etapas envolvidas no processo de detecção e análise de SNPs: alinhamento de ESTs e cDNA com DNA genômico, detecção de SNPs por análise de cromatograma e alinhamento múltiplo, e finalmente análise de correlação de polimorfismos e LDs múltiplos. O principal objetivo foi propor novas estratégias para o aprimoramento de algoritmos relacionados a essas três etapas.

6.1 Alinhamento de cDNA e ESTs com DNA genômico

A primeira etapa estudada foi o alinhamento de ESTs e cDNA com DNA genômico. Esta etapa é de grande importância, pois durante o processo de transcrição de genes em organismos eucariotos as regiões não-codificadoras são removidas, gerando um transcrito maduro. Além disso, nesse momento pode ocorrer `alternative splicing`, onde apenas um subconjunto dos exons de um gene são transcritos. Por isso, para que os ESTs possam ser comparados entre si e analisados de forma adequada, é necessário que sejam colocados dentro do contexto genômico através do alinhamento com a seqüência do gene de origem.

O grande desafio para os alinhadores de ESTs com DNA genômico é conseguir delimitar os diferentes exons dentro do mRNA capturado pelos projetos ESTs e encontrar as regiões de origem na seqüencia do gene ou do cromossomo. Várias estratégias foram desenvolvidas para resolver o problema de alinhamento de cDNA e ESTs com DNA genômico, utilizando algoritmos baseados em várias etapas e heurísticas distintas. Neste trabalho analisamos três ferramentas desenvolvidas para este fim: `est_genome` [80], `sim4` [32] e `Spidey` [122] .

O objetivo principal do trabalho foi obter um conjunto de parâmetros adequados para resolver este tipo de alinhamento utilizando algoritmos tradicionais de alinhamento (global

e semi-global com pontuação linear para buracos), utilizando como conjuntos de dados genes dos cromossomos humanos.

Os resultados obtidos mostram que o algoritmo semi-global com pontuação (1, −2, −10, 0) produz alinhamentos entre DNA genômico e cDNA extremamente satisfatórios. Os resultados obtidos com alinhamentos sem erros são próximos do ideal e os resultados produzidos com dados contendo poucos erros estão muito próximos dos resultados obtidos pelos alinhadores externos, desenvolvidos especialmente para resolver este tipo de problema. Dentre os alinhadores externos, nossos testes mostraram que o melhor foi o pacote `sim4`, tanto nos testes com cDNA sem erros quanto nos testes com ESTs com erros.

6.2 Detecção de SNPs

A segunda etapa estudada foi o processo de detecção de SNPs aplicadas a lotes de seqüências genéticas de HIV extraídas de pacientes soropositivos. Estes lotes apresentam a característica de possuírem baixa cobertura média de cada base da seqüência de referência, com alta porcentagem de bases com baixa qualidade e grande quantidade de polimorfismos. A motivação deste estudo foi o fato de que os pacotes computacionais `Polybayes` e `Polyphred` não obtiveram bons resultados no processo de detecção.

Foram usadas duas técnicas para detectar polimorfismos nos lotes: análise do cromatograma de cada seqüência, a procura de sinais secundários não levados em conta pelo pacote `phred`, e geração de consenso a partir da análise das seções transversais do alinhamento múltiplo.

Os resultados obtidos foram bastante satisfatórios, se comparados com os resultados obtidos pelos pacotes citados anteriormente. Ao final do processo de comparação entre os diversos algoritmos de análise ("Relação de Áreas", "Relação das Médias das Alturas", "Limite Variável", "Pico Único por Janela", "Eliminação de Picos Ruins"e "Pico Mais Baixo") de cromatograma e correção de base-calling, estabeleceu-se que o melhor algoritmo é o de "Relação de Áreas", com parâmetros:

$$\begin{cases} \texttt{MIN\_RELATION} = 0.25 \\ \texttt{DISTANCE\_PERCENTAGE} = 0.5 \end{cases}$$

Como foi dito anteriormente, seqüências genéticas de vírus como o HIV utilizado neste trabalho possuem um alto número de mutações. Seria portanto interessante repetir o experimento utilizando seqüencias genéticas de seres vivos mais conservados, como, por exemplo, mamíferos, de forma a validar os algoritmos desenvolvidos.

6.3 Correlação de SNPs

A terceira etapa deste trabalho foi a análise de correlação de SNPs, utilizando desequilíbrio de ligação, também chamado de LD - Linkage Disequilibrium. LD tem sido apontado como uma ferramenta de grande utilidade para facilitar o mapeamento de genótipos complexos [6], e para refinar a busca por locus responsáveis por doenças.

Apresentamos uma definição de LDs múltiplos fundamentada em teoria dos grafos. Foi definido um algoritmo que utiliza uma heurística gulosa baseada no grau dos vértices para buscar LDs múltiplos. Este algoritmo não garante que o LD múltiplo máximo será encontrado, mas ótimos resultados foram obtidos nos conjuntos de dados estudados.

O algoritmo descrito neste trabalho foi executado utilizando quatro variações. Na primeira variação removemos as bases de baixa qualidade e definimos LD como todo par de SNP cujo valor para métrica D' seja igual a 1. Na segunda, fizemos a mesma definição para LD só que consideramos as bases de baixa. A terceira e quarta variações são semelhantes a primeira e segunda, diferenciando apenas na definição de LD. Nestas é considerado LD o par de SNPs cujo valor para a métrica r^2 seja maior ou igual a 1/3. Na literatura, a primeira definição é chamada de LD completo e a segunda de LD útil.

Os algoritmos foram aplicados a SNPs mapeados em genes da cana-de-açúcar pelo projeto SUCEST, e em dados da região MHC do cromossomo 6 do genoma humano. Os resultados obtidos com os dois conjuntos de dados foram bastante semelhantes. Comparando os resultados que apresentamos, chegamos a conclusão de que a definição de LD completo é mais adequada para o cálculo de LDs múltiplos. Tal conclusão deveu-se a dois fatores: a métrica apresentou uma maior capacidade de agrupamento e uma menor tendência ao isolamento de SNPs ou formação de LDs simples.

Pudemos observar também que o algoritmo apresentou comportamento semelhante quando utilizou a definição de LD útil ao invés de LD completo. Sendo assim, LD útil pode oferecer bons resultados, caso o desempenho do algoritmo seja fator limitante.

6.4 Considerações finais

Este trabalho apresentou um estudo de vários métodos e algoritmos para análise de SNPs. Cada etapa foi estudada de forma separada, resolvendo problemas específicos distintos com conjuntos de dados diferentes. O resultado final pode ser considerado como um fluxo de processamento para SNPs, cobrindo várias etapas necessárias.

Uma extensão interessante deste trabalho seria a construção de um pacote computacional que implementasse um fluxo de trabalho completo com todas as etapas relacionadas ao estudo de SNPs, conforme os algoritmos de alinhamento, de identificação de SNPs e de LDs (simples e múltiplos) propostos nesta dissertação.

A validação desta ferramenta poderia ser feita através do processamento de conjuntos de dados genéticos previamente anotados de seres vivos bem conservados, como, por exemplo, o do genoma humano.

Apêndice A

Revisão Bibliográfica

Nesta Seção apresentaremos uma breve revisão bibliográfica, com os resumos de artigos estudados ao longo do desenvolvimento deste trabalho. A escolha dos textos foi feita baseada no interesse para entendimento e fundamentação das várias etapas do estudo de SNPs. O objetivo destes resumos é extrair as idéias centrais de cada artigo e citar os principais tópicos de interesse para a nossa pesquisa.

A.1 The essence of SNPs [15]

O projeto de sequenciamento do genoma humano criou um grande interesse por estudo sobre polimorfismos, visando entender a conexão destes com propenção a doenças, maior ou menor sensibilidade a certas drogas, entre outros.

Dentre os polimorfismos, os SNPs tem tido um espaço de destaque, por representarem cerca de 90% dos polimorfismos encontrados no genoma humano [21]. SNPs (Single Nucleotide Polymorphism, ou Polimorfismos de Nucleotídeo Único) são pares de bases nas quais podem existir variações alélicas em indivíduos normais com uma freqüência de 1% ou mais. Não são considerados SNPs inserções ou remoções simples de bases em uma seqüência genômica. A priori, SNPs poderiam ser polimorfismo bi, tri ou tetra alélicos. Porém os dois últimos tipos são extremamente raros e praticamente inexistentes.

A definição acima tem alguns problemas: alguns pesquisadores preferem não considerar polimorfismos de base única que predispõe o indivíduo a doenças como SNP. Seriam apenas considerados como SNPs variações que levem ação recessiva, aumento de risco de doença e outros desde que ocorram em indivíduos não doentes.

Existem quatro tipos de polimorfismos possíveis em SNPs: uma transição, C⇔T (G⇔A), que ocorre em dois terços do casos, e tres transversões C⇔A (G⇔T), C⇔G (G⇔C) T⇔A (A⇔T). A freqüência básica com que estes SNPs são observados é da ordem de 1 a cada 1000 pares de bases (1/1000bp). A maioria dos SNPs existentes no

genoma humano foram originados após a especiação, mas antes da aparição das diferentes populações. Com isso, a maioria dos SNPs humanos não existem nas populaçoes de primatas, mas 85% deles são encontrados em todas as populações humanas, e apenas 15% existem apenas dentro de uma população.

Um grande esforço tem sido feito com o objetivo de se mapear a maior quantidade possível de SNPs. Uma das grandes barreiras para se obter estes dados é a dificuldade de obter séries de seqüências genéticas de diversas populações permitindo a determinação das frequências alélicas dos SNPs.

A efetiva utilização dos dados obtidos sobre SNPs tem como pré requisito a construção de bases de dados públicas, sendo que as duas maiores são dbSNP [26] e HGBASE [52].

A enorme quantidade de fenótipos humanos podem ser causados tanto por fatores genéticos como por fatores ambientais. Claramente, muitos fenótipos associados a doenças tem uma forte componente genética. Com isso, espera-se que riscos de se contrair doenças como câncer, diabetes, doenças mentais e cardiovasculares sejam fortemente ligadas por padrões de SNPs. O mesmo raciocínio pode ser utilizado para respostas a certas drogas. Faz-se uma separação entre variações que predispõem fortemente a doenças e variações que simplesmente modificam o risco de se contrair a doença: SNPs se enquadram no segundo caso. Ou seja, SNPs não são necessários nem suficientes para se causar doenças, mas uma combinação de certas variações com fatores ambientais podem determinar ou não o aparecimento de doenças.

Se um fator contribui para se aumentar o risco de uma doença, então este deve ser encontrado em uma maior freqüência em individuos doentes do que em individuos sadios. O processo de se associar variações alélicas com fenótipos envolve simplesmente se determinar a freqüência de um SNP em vários indivíduos que apresentem o fenótipo. A validade dos resultados dependerá da seleção apropriada dos indivíduos. Esta abordagem traz dois problemas: primeiramente, muitas vezes a complexidade dos fatores que levam a um certo fenótipo é tao grande que não se pode definir fatores precisos a serem estudados. O segundo problema é que cada SNP estudado levaria a um resultado, e o estudo da combinação destes resultados poderia levar a milhares de associações falsas. Um método alternativo de se aprimorar a estratégia de estudo de associação é utilizar informações fornecidas por correlação entre alelos, conhecido como *Linkage Disequilibrium* [], doravante denominado de LD. Um LD consiste em um conjunto de mutações no código genético que gera uma configuração onde é possível se predizer a existência de um alelo a partir do mapeamento de um segundo alelo bem preciso.

A.2 SNPs: Sutis diferenças de um código [50]

O genoma humano possui um grande número de variações nas sequencias nucleotídica em segmentos correspondentes do genoma. Dentre esta variações, as mais comuns são chamadas de SNPs, que corresponde a posições onde os nucleotídeos se alternam com uma freqüência mínima de 1% em uma dada popolução [15].

As variações mais frequentes são substituições entre bases nitrogenadas de mesma característica estrutural (A/G ou G/A e C/T ou T/C), que são chamadas de transições. As outras substituições são conhecidas com transversões.

As mudanças podem ocorrer em regiões codificadoras ou não codificadoras do genoma. No primeiro caso, podem ser ou não sinônimas (caso a troca gere aminoácidos diferentes ou não), e mesmo sendo sinônimas, podem afetar o comportamento do gene.

O estudo dos SNPs tem grande importância na descoberta de alterações que possam levar à predisposição de indivíduos contraírem doenças como câncer, maior resistência a certas drogas, entre outros. O estudo detalhado destes efeitos pode levar ao desenvolvimento de produtos individualizados, permitindo maior eficiência nos tratamentos médicos.

Devido a isto, existem vários projetos que visam mapear SNPs no genoma humano. Entre eles, o SNP Consortium, que em novembro de 2000 havia mapeado 1.4 milhões de SNPs no genoma humano [102], e o HCGP-SNP [51] no Brasil.

Estes projetos de SNP necessitam de uma enorme quantidade de dados, e por isso a utilização de ferramentas bioinformáticas são de grande importância, assim como a construção de sistemas de banco de dados capazes de catalogar e disponibilizar estes dados via internet, para pesquisa. Paralelamente é necessário que existam métodos de validação experimental para que se possa avaliar de forma precisa cada SNP encontrado.

Todas estas características dos SNPs trazem a tona uma discussão ética: a legalidade de se patentear SNP para exploração comercial por uma empresa farmacêutica. A tendência atual é se patentear os produtos industrializados criados a partir das pesquisas de SNPs, mas não os polimorfismo em si.

A.3 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphism [102]

Uma das principais metas dos atuais estudos genéticos é descobrir variantes de DNA que contribuem na variação fenotípica humana. Estudos conseguiram mapear centenas de genes relacionados a doenças. Porém, raramente apenas uma mutação em um gene é

responsável pela contração de uma doença: ao contrário, em geral o fenótipo é causado por um conjunto de genes atuando de forma complexa. Alem do mais a diversidade genética humana não se limita apenas a polimorfismos individuais dentre de genes, mas a uma combinação de alelos próximos um dos outros.

Dois projetos, o SNP Consortium e o Human Genome Sequencing Consortium, foram os principais responsáveis pelo mapeamento de 1.42 milhões de SNPs no genoma humano ($\sim$ 95% dos SNPs mapeados pelos dois), disponibilizados publicamente em novembro de 2000.

Observou-se que os SNPs aparecem em média a cada 1900 bases, e que grande parte do genoma contêm uma alta densidade de polimorfismos: 90% de segmentos contíguos de 20k bases contem um ou mais SNPs, assim como 63% de segmentos de 5k bases e 28% de segmentos de 1k bases. Apenas 4% do genoma possui separação entre SNPs maiores do que 80000 bases.

Analisando a distribuição dos SNPs em regiões codificadoras do genoma, observou-se que 93% dos loci de genes contém ao menos um SNP, 59% contém 5 ou mais SNPs e 39% contém 10 ou mais SNPs. Observou-se também que 98% dos loci estão a uma distância menor que 5000 bases do SNP mais próximo.

A.4 Single-nucleotide polymorphisms in the public domain: how useful are they? [76]

Muitos grupos de pesquisa públicos e privados tem trabalhado com o objetivo de identificar um grande conjunto de SNPs. Os resultados destas pesquisas tem sido depositados em bases de dados públicas, como o dbSNP, no National Center for Biotecnology Information [26].

Uma grande parte dos SNPs armazenados nestas bases são na realidade candidatos a SNPs, encontrados por softwares de data-mining, que não foram caracterizados.

Dois estudos pilotos utilizando metodologias de análise diferentes foram efetuados para analisar os SNPs contidos na base dbSNP, afim de descobrir o grau de interesse dos dados disponibilizados. Em particular, foram feitas análises em 3 populações diferentes (Caucasianos, Africanos e Asiáticos) visando descobrir quantos SNPs eram comuns às 3, podendo então funcionar como marcadores genéticos.

Os resultados obtidos foram bem parecidos: em ambos os casos, uma pequena porcentagem dos dados eram incompletos (15%), sendo descartados; 8% dos SNPs não foram detectados em nenhuma população; 77% dos SNPs existem em apenas uma população, 53% em duas populações e 27% nas 3 populações.

Estes dados são bastante confiáveis do ponto de vista experimental. O problema é

saber se os SNPs não são resultados falso-positivos, gerados por regiões duplicadas ligeiramente diferentes tratadas como regiões diferentes. Por experiência, estima-se que apenas 5% dos SNPs que não são capturados por filtros computacionais de regiões repetidas são falsos-positivos.

Com estes dados, conclui-se que os repositórios públicos, apesar de conterem dados não totalmente caracterizados, podem ser muito úteis para a pesquisa, desde que de faça uma escolha judiciosa dos SNPs.

A.5 SNP Databases and Pharmacogenetics: A Great Start, but a Long Way to Go [75]

A publicação do mapeamento preliminar do genoma humano gerou um grande entusiasmo para o estudo de tratamentos individualizados, baseados no fato que a grande maioria dos genes apresentam uma grande frequência de variações alélicas.

Grande parte destas variações são SNPs (aproximadamente 1.42 milhão mapeados em novembro de 2000 [102]), e este grande número leva à necessidade da criação de estratégias eficientes para análise e compilação de bibliotecas mapeando todos os SNPs do genoma humano.

Muitas bases de dados disponíveis na internet para pesquisa foram criadas, e são de grande interesse para estudos de farmacogenética. No entanto, não haviam sido feitas comparações entre as bases para verificar de que forma os dados contidos nelas podem ajudar a se obter resultados em pesquisa genômica.

Este estudo analisou os recursos disponíveis em diversas bases. Foram procurados SNPs contidos em 74 genes humanos ligados ao processamento molecular (transporte, degradação, ativação, etc) de 3 classes de drogas usadas em quimioterapia. As buscas foram efetuadas em 7 diferentes bases públicas (CGAP-GAI [18], LEELAB [73], dbSNP [26], JSNP [66], HOWDY [58], HGVBase [52], GeneSNP [38], PharmGKB [94]), visando se determinar o grau de redundância dos SNPs armazenados.

Dos 74 genes pesquisados, 893 SNPs haviam sido descobertos até o final de 2001. Os resultados mostraram que havia pouquíssima redundância de SNPs armazenados nas bases. Por exemplo, considerando as bases CGP-GAI, LEELAB e HOWDY, apenas 8 SNPs (1%) foram encontrados nas três. 6 SNPs não foram encontrados em nenhuma delas.

Os dados contidos nestas bases devem ser analisados com muito cuidado. Muitos deles não são validados experimentalmente, sendo encontrados por ferramentas de busca computacional, que podem gerar resultados falsos. Um erro possível é o de se definir como SNPs através da comparação de ESTs obtidos de diferentes famílias de genes com alto

grau de semelhança.

Portanto, pesquisas por SNPs devem ser feitas em diferentes bases, para se obter resultados mais completos e fiáveis. Políticas de controle de inserção e validação de novos dados nas bases de dados são de extrema importância. Da mesma forma, o desenvolvimento de estratégias de sequenciamento de DNA apropriadas para procura e validação experimental de SNPs são essenciais para se obter resultados relevantes na pesquisa por polimorfismos relevantes ao tratamento de doenças.

A.6 Pharmacogenetics goes genomics [43]

A farmacogenética busca a redução das variação de resposta de diferentes indivíduos a certas drogas, criando terapias individuais, baseadas nas características genéticas de cada indivíduo.

Arno Motulsky foi o primeiro pesquisador a articular pesquisas em farmacogenética em 1957 [81]. Ele argumentou que elementos genéticos poderiam ser subjacentes as variações de respostas em cada indivíduo para certas drogas. O desafio é mapear estas variações genéticas e entender com elas interagem entre si, permitindo criar terapias de acordo com estes fatores.

Definir genes candidatos, ou seja, genes que podem conter polimorfismos que influenciem o comportamento do organismo a certas drogas, é um processo bastante complexo. Vários grupos de genes podem estar associados ao fenótipo, como por exemplo genes que codificam transportadores de substâncias ou enzimas de metabolização, e genes que codificam os sítios de recepção das drogas.

Atualmente, existe um conjunto de polimorfismos que foram fortemente associados a resposta a drogas, citados em pelo menos dois estudos diferentes cada um. Deste conjunto, aproximadamente 50% dos polimorfismos agem sobe o sítio de recepção da droga ou sobre proteínas que agem sobre a droga até o momento de chegarem ao sítio, enquanto que 33% agem sobre enzimas de metabolização, confirmando a importância destas categorias de polimorfismos.

O grande problema com estudo farmacogenéticos atualmente é a falta de planejamento. É necessario se fazer uma amostragem adequada da população, para que os resultados possam ser conclusivos: muitas das amostras apresentadas em estudos são muito menores do que as requeridas, tornando os estudos pouco efetivos. Por outro lado, poucos estudos levam em conta o problema da estratificação, que consiste em analisar uma população de indivíduos geneticamente estruturada, podendo causar resultados errados.

Grande parte dos estudos farmacogenéticos são retrospectivas, ou seja, são executados depois que respostas ja tenham sido observadas. Raramente são feitos estudo prévios para se definir como o conhecimento do genótipo pode melhorar o tratamento clínico de um

paciente. Em geral, transformar pesquisa farmacogenética em terapias avançadas requer uma drástica expansão dos estudos prospectivos sobre como variações do código genético influem na resposta a drogas.

A.7 Accounting for Human Polymorphisms Predicted to Affect Protein Function [88]

Um dos maiores interesses da pesquisa sobre o genoma humano é determinar se um SNP não sinonimo, chamado de nsSNP, afeta a produção da proteína e consequentemente tem impacto sobre a saúde do indivíduo. Atualmente, aproximadamente metade das causas genéticas de doenças são causadas por substituições de amino-ácidos [24].

Define-se como SNP não sinônimo polimorfismos que levem à tradução de um amino-ácido diferente, podendo modificar a estrutura e função de uma proteína. Um nsSNP é chamado de neutro quando sua modificação não altera a proteína produzida.

Dois grupos utilizaram ferramentas computacionais para predizer o efeito de um nsSNP na produção de proteínas: o primeiro [113] estimou que aproximadamente 20% dos nsSNPs ($\sim$ 2000) são prejudiciais, e o segundo [20] estimou que aproximadamente 30% ($\sim$ 9500) são prejudiciais.

As ferramentas descritas em Sunyaev et al.(2001)[113] e Chasman e Adams(2001)[20] utilizam como base de funcionamento o fato que aproximadamente 30% das proteínas codificadas pelo genoma humano são parecidas com proteínas cujas estruturas são conhecidas [49]: assim, as ferramentas utilizam as informaçoes estruturais para analisar os nsSNPs. O problema é que estas ferramentas são restritas às proteínas com estrutura conhecidas.

A ferramenta SIFT [87] (*Sorting Intolerant From Tolerant*) utilizada neste estudo não requer informações estruturais, fazendo análise apenas a partir de sequências homólogas. Num primeiro passo, a ferramenta escolhe sequências próximas à da proteína a ser analisada, e gera um alinhamento, a partir do qual será efetuada a análise. Em um segundo passo, calcula-se a probabilidade da substituição de interesse ser prejudicial baseando-se nos amino-ácidos observados na posição da substituição no alinhamento gerado no passo inicial: o nsSNP será considerado prejudicial se a probabilidade estiver abaixo de um certo limite pré-definido. As sequências são obtidas do projeto SWISS-PROT/TrEMBL [7] com a ajuda da ferramenta de blast PSI-BLAST [5]: as sequências encontradas com mais de 90% de similaridade são utilizadas.

O estudo utilizou as ferramenta com dados obtidos de 3 bases de dados distintas. No primeiro conjunto de dados, nsSNPs relacionados a doenças obtidos da base SWISS-PROT/TrEMBL, o software previu que 69% das substituições são prejudiciais. No se-

gundo conjunto de dados, nsSNPs obtidos de indivíduos sadios obtidos do Whitehead Institute [17], o software previu que 19% dos nsSNPs são prejudiciais. Porém estes dados podem ser errados devido ao fato do software ter previsto 19% de dados confirmadamente neutros como sendo prejudiciais. Na terceira conjunto de dados, nsSNPs não confirmados obtidos de dbSNP, 25% dos nsSNPs foram previstos como prejudiciais, e 19% de nsSNPs neutros foram erroneamente previstos como prejudiciais.

A vantagem de uma ferramenta de predição está no fato que atualmente existem mais de um milhão de SNPs mapeados [102], dos quais muitos podem não afetar a função da proteína codificada (serem neutros). Definir um conjunto de nsSNPs com probabilidade de serem prejudiciais permite que os estudos sejam mais concentrados, evitando perda de tempo, uma vez que estudos de SNP são custosos e demorados.

A.8 EST analysis online: WWW tools for detection of SNPs and alternative splice forms [14]

ESTs (Expressed Seguence Tags) foram originalmente utilizados para identificação de genes [2]. Posteriormente, descobriu-se que eram extremamente úteis como ferramenta de mapeamento de genoma [105]. Além disso ESTs foram utilizados para se estudar a frequência de SNPs no genoma humano e o fenômeno de Alternative Splicing (AS) [112]. Este trabalho apresenta uma ferramenta de análise de ASs e de SNPs através de ESTs, disponível na internet [78].

A ferramenta para identificação de SNPs primeiramente alinha os ESTs a serem analizados com uma sequência genética de busca (que pode ser um cDNA ou um outro EST), utilizando a ferramenta BLASTN [5]. Os ESTs alinhados são filtrados, de forma que apenas aqueles com semelhança de 95% são guardados. Para eliminar ruídos devido a sequências de baixa qualidade, a ferramenta utiliza o programa phred [31]para avaliar qualidade de cada base: apenas as bases com qualidade superior a 30 (numa escala logaritmica de 0 a 60) são analizadas, o que significa uma probabilidade de erro de 1/1000. Além destes filtros, utiliza-se uma regra para definir se uma base modificada corresponde a um SNP: o polimorfismo deve aparecer em pelo menos dois ESTs alinhados. Estudos mostraram que 74% dos SNPs detectados desta forma foram confirmados posteriormente por re-sequenciamento. A ferramenta exibe o alinhamento obtido, e os SNPs detectados destacados em vermelho. O cromatograma correpondente a cada EST e a qualidade obtida pelo phred também são exibidos.

A ferramenta também permite fazer buscas por ASs. Alternative Splicing é um importante mecanismo que permite a um gene expressar diferentes formas fenotípicas, algumas delas causadoras de doenças [69]. A ferramenta permite ao usuário selecionar uma

sequência de mRNA ou de proteína, colocá-la no formato FASTA e compará-la com ESTs humanos, utilizando BLAST. O resultado é analisado de forma a eliminar sequências que tenham semelhança inferior a 95% em 100 pares de bases ou 30 aminoácidos, visando eliminar pseudo-genes e sequências homólogas. O sistema elimina também candidatos a AS com repetição interna de proteínas, que podem ser confundidas com possíveis formas de AS. Feitas as filtragens, a ferramenta procura por sequências que estejam presentes no EST e não no resultado da busca e vice-versa, destacando os resultados.

A.9 Patterns of Linkage Disequilibrium in the Human Genome [6]

Muito do sucesso do mapeamento genético de doenças no passado se deveu ao fato de que as primeiras doenças humanas estudadas eram bastante simples, monogênicas, obedecendo as regras mendelianas de herança genotípica. Grande parte destas doenças foram identificadas por meio de estudos de ligação, que consiste em se coletar dados genéticos de famílias de indivíduos afetados por uma doença, e comparar os dados em busca de regiões do genoma compartilhados por indivíduos doentes e não presentes em indivíduos sadios. Porém, em muitos casos, é necessário se obter dados de uma grande população para se conseguir definir de forma precisa regiões causadoras de uma doença.

Métodos novos foram idealizados para se determinar com mais precisão possíveis regiões onde os genes causadores de uma dada doença estariam localizados, utilizando para isso marcadores.

Recentemente, geneticistas voltaram seus esforços em mapear doenças mais comuns, cuja base genética é mais complexa e que afetam grandes fatias da população, e onde os métodos tradicionais tem se mostrado menos eficientes. Neste contexto, o conceito de Linkage Disequilibrium tem se mostrado extremamente útil, devido ao fato de se acreditar que estudos de associação são mais eficientes do que estudos de ligação. Estudos de associação consistem em se associar alelos particulares a um determinado fenótipo.

De fato, uma análise feita por Risch e Merikangas [101] sugere que o número de linhagens genotípicas distintas necessárias para se fazer um estudo clássico de ligação visando o mapeamento de genes com efeito menor mas implicados na suscetibilidade de um indivíduo de contrair uma dada doença seria proibitivo.

Linkage Disequilibrium, ou LD, é uma associação não aleatória de alelos em loci adjacentes: quando dois alelos específicos em loci diferentes num mesmo cromossomo são encontrados em conjunto (quando um está presente, o outro também estará), então os loci estão em desequilíbrio . Este conceito pode ser formalizado por uma das mais antigas medidas propostas para desequilíbrio, simbolizada por D. Esta medida quantifica um LD

como sendo a diferença entre a frequência observada entre um haplótipo (combinação de alelos encontrados em loci próximos dentro de um cromossomo) de dois loci e a frequência que seria esperada se os alelos fossem aleatórios. A equação seria $D = P_{AB} - P_A \times P_B$, onde A e B são alelos, P_B e P_A são as probabilidades de aparição dos alelos separadamente e P_{AB} é a probabilidade dos dois alelos aparecerem juntos.

Apesar de D expressar o conceito intuitivo de um LD, seu valor numérico tem pouco uso na comparação de LDs. Isto se deve ao fato de que D depende da frequência de alelos. Assim, várias outras medidas tem sido propostas (apresentadas por Devlin e Risch [27]). As mais comuns são os valores absolutos D' e r^2.

O valor absoluto de D' é obtido divindo-se o valor de D por seu valor máximo possível, dada uma frequência alélica nos dois loci. O caso $D' = 1$ é conhecido como LD completo, ou seja, quando dois SNPs não foram separados por recombinação.

A medida r^2, também denotada por Δ^2, é complementar a D', e foi apontada como sendo a medida ideal para comparação de LDs em processos de mapeamento. É obtido dividindo-se D^2 pela produto das frequências dos 4 alelos nos dois loci. O caso $r^2 = 1$, conhecido como LD perfeito, acontece se, e somente se, os marcadores não foram separados por recombinação e tem a mesma frequência alélica. Portanto, quando se tem um LD completo, informações sobre um marcador provem informações completas sobre o outro marcador, fazendo as duas informações redundantes.

r^2 tem sido muito utilizado para se definir o que são LDs úteis [70]. De fato, aumento do número de amostras em estudos de associação tem custo alto, e aumentar o número de amostras para compensar LDs fracos é praticamente inviável. LDs com $r^2 > 1/3$ são considerados como úteis em processos de mapeamentos.

Uma outra medida utilizada para se quantificar um LD é o parâmetro de recombinação da população $4N_e r$ (também conhecido como ρ, $4N_e c$ ou C), onde r é a taxa de recombinação na região de interesse e N_e é o tamanho eficaz da população (Effective Population Size [97]), que corresponde ao tamanho de uma população gerada pelo cruzamento ideal e aleatório de uma outra população, mantendo o mesmo nível de variação observada na população real (N_e é bem menor do que o tamanho da população real)

Vários fatores influem na criação de LDs. Os mais evidentes são mutação e recombinação, mas outros podem ser citados, como a mudança aleatória da frequência de certos genes sobretudo em pequenas populações; o aumento de uma população; mistura de populações; seleção natural; taxa variável de recombinações; taxa variável de mutações, e outros.

Muito do conhecimento e entendimento de como LDs são formados na natureza veio do estudo feito em espécies de *Drosophilas*, e em particular, da *Drosophilas Melanogaster*, de onde foram tirados estudos mais detalhados de LDs.

É muito importante se analisar a evolução demográfica humana para se determinar a

densidade de marcadores necessárias para se efetuar estudos de associação úteis.

Uma outra questão de interesse é saber se existem blocos discretos de LDs, ou se deve se estudar diversas regiões para o mapeamento de uma dada doença.

Estudos foram feitos para se analisar os padrões de LD na população humana. Observa-se que este são diferentes conforme as regiões, com europeus apresentando menor diversidade nucleotídica e maior número de LDs que africanos. Comparações são dificultadas pelo fato de que diversos estudos foram feitos utilizando medidas diferentes. Um ponto que pode ser extraído destes estudos é que a variação de LD em qualquer distância é grande, e não é previsível de uma região para a outra. Com isso, deve-se tomar muito cuidado ao se fazer qualquer predição sobre LDs em regiões onde não foram feitos estudos empíricos.

A.10 Optimal alignment in linear space [84]

Em biologia computacional, um problema muito comum é o seguinte: dadas duas sequências $A = a_1a_2 \ldots a_M$ e $B = b_1b_2 \ldots b_N$, deseja-se encontrar um conjunto de operações que converta A em B com custo mínimo, considerando-se que as operações permitidas são substituição de um símbolo por outro, remoção de k símbolos consecutivos e inserção de k símbolos consecutivos. O custo de uma substituição de a por b é dado pela função $w(a, b)$. O custo de se inserir buracos devido a k inserções ou remoções é dado pela função $gap(k) = g + hk$ onde g representa o custo de se abrir um buraco e h representa o custo de se extender um buraco.

Gotoh [45] propôs um algoritmo que permite reconstruir o conjunto de operações utilizadas para transformar uma sequência na outra em tempo e espaço quadrático ($O(MN)$). Porém, o espaço necessário é, em muitas situações, proibitivo, inviabilizando seu uso em máquinas convencionais. Hirschberg [55] propôs um algoritmo que resolve o problema de se encontrar a maior subsequência comum entre duas sequências, utilizando espaço $O(N)$ onde N é a sequência de menor tamanho.

Este trabalho aplica a técnica usada por Hirschberg no algoritmo de Gotoh, para encontrar a conversão ótima entre duas sequências. As seguintes simplificaçôes são usadas na descrição do algoritmo: $w(a, b) = 1$ se $a \neq b$, $w(a, b) = 0$ se $a = b$ e $gap(k) = k$.

Gotoh define $A_i = a_1, a_2, \ldots, a_i$ e $B_j = b_1, b_2, \ldots, b_j$, e as seguintes funções:

- $C(i, j)$ = custo mínimo de conversão de a_i para b_j onde:

$$C(i,j) = \begin{cases} min\{D(i-1,j), C(i-1,j) + g\} & \text{se } i > 0 \text{ e } j > 0 \\ gap(j) & \text{se } i = 0 \text{ e } j > 0 \\ gap(i) & \text{se } i > 0 \text{ e } j = 0 \\ 0 & \text{se } i = 0 \text{ e } j = 0 \end{cases}$$

- $D(i,j)$ = custo mínimo de conversão de a_i para b_j que remova a_i onde:

$$D(i,j) = \begin{cases} min\{D(i-1,j), C(i-1,j)+g\} & \text{se } i>0 \text{ e } j>0 \\ C(0,j)+g & \text{se } i=0 \text{ e } j>0 \end{cases}$$

- $I(i,j)$ = custo mínimo de conversão de a_i para b_j que insira b_j onde:

$$I(i,j) = \begin{cases} min\{I(i,j-1), C(i,j-1)+g\}+h & \text{se } i>0 \text{ e } j>0 \\ C(i,0)+g & \text{se } i>0 \text{ e } j=0 \end{cases}$$

Os valores na i-ésima linha de C e D depende apenas dos valores nas linhas i e $i-1$, enquanto que valores na i-ésima linha de I depende apenas dos valores na linha i. Assim, é possível se definir dois vetores CC e DD e três escalares e, c e s tais que:

$$CC(k) = \begin{cases} C(i,k) & \text{se } k<j \\ C(i-1,k) & \text{se } k \geq j \end{cases}$$

$$DD(k) = \begin{cases} D(i,k) & \text{se } k<j \\ D(i-1,k) & \text{se } k \geq j \end{cases}$$

$$e = I(i,j-1)$$

$$c = C(i,j-1)$$

$$s = C(i-1,j-1)$$

O algoritmo original apresentado por Hirschberg é recursivo e utiliza a técnica de divisão e conquista. A idéia central é encontrar um ponto médio de uma conversão ótima aplicando o algoritmo original em duas fases: direta e reversa, recursivamente. Supondo $M > 1$ e $N > 0$, define-se o ponto médio $i^* = \lfloor M/2 \rfloor$. Na fase direta, aplica-se o algoritmo original nas sequências A_{i^*} e B, de forma a que $CC(j)$ represente o custo mínimo de conversão de A_{i^*} para B e $DD(j)$ represente o custo mínimo de conversão de A_{i^*} para B que termina com uma remoção.

Define-se também $rev(A)$ como sendo a_M, a_{M-1}, ..., a_1 e A_i^T como sendo a_{i+1}, a_{i+2}, ..., a_M. A fase reversa aplica o algoritmo em $rev(A)_{M-i^*}$ e $rev(B)$, obtendo-se dois vetores RR e SS homólogos a CC e DD, onde $RR(j)$ contém o custo mínimo de conversão de $rev(A)_{M-i^*}$ para $rev(B)$ e $SS(j)$ contém o custo mínimo de conversão de $rev(A)_{M-i^*}$ para $rev(B)$ que termina com uma remoção. Portanto pode se afirmar que $RR(N-j)$ contém o custo mínimo de conversão de $A_{i^*}^T$ para B_j^T e $SS(N-j)$ contém o custo mínimo de conversão de $A_{i^*}^T$ para B_j^T que se inicia com a remoção.

Com os vetores acima definidos, o ponto médio de uma conversão ótima (i^*, j^*) pode ser obtido fixando-se $i^* = \lfloor M/2 \rfloor$ e j^* tal que:

$$min_{j \in [0.N]}\{min(CC(j) + RR(N-j), DD(j) + SS(N-j) - j\}$$

Encontrado o ponto ótimo, a conversão ótima pode ser obtida com os seguintes passos:

1. Encontrar recursivamente uma conversão ótima de A_{i^*} para B_{j^*}
2. Encontrar recursivamente uma conversão ótima de $A^T_{i^*}$ para $B^T_{j^*}$
3. Concatenar essas conversões parciais

O algoritmo proposto usa espaço $O(N \log M)$, onde $O(N)$ é utilizado pelos vetores globais e $O(\log M)$ pela pilha gerada pelas chamadas recursivas. O tempo de execução é aproximadamente o dobro do algoritmo original de Hirschberg.

A.11 est_genome: A program to align spliced DNA sequences to unspliced genomic DNA [80]

Este trabalho apresenta o programa *est_genome*, desenvolvido para alinhar pedaços de sequências (mRNA, EST ou cDNA) com sequências genômicas. O pacote foi escrito em ANSI C.

A identificação de genes em sequências de DNA não-caracterizadas é um dos grandes problemas na pesquisa genômica. Um dos métodos que tem sido mais utilizado para esta tarefa é alinhar pedaços de sequência com sequências genômicas. O grande número de ESTs sequênciados tem sido um fator importante na adoção desta estratégia.

Ferramentas padrão para alinhamento de sequências genéricas não são ideais para esta tipo de tarefa, devido ao grande número de íntrons que podem ocorrer na sequência genômica. Além disso, alinhamento de longas sequências pode ser impossível utilizando-se algoritmos que consomem espaço quadrático com o tamanho da sequência, por consumirem muita memória.

O programa *est_genome* resolve estes problemas, permitindo a existência de grandes íntrons, reconhecendo sítios de splice e utilizando memória limitada. Por ser um programa lento se comparado ao BLAST [4], primeiro se compara a sequência genômica com o dbEST, utilizando BLASTN: todos os ESTs encontrados são realinhados com *est_genome*.

O algoritmo utiliza uma modificação do algoritmo de Smith e Waterman [108]. A estrutura de custos para pontuação do alinhamento é a seguinte (o valor numérico da

pontuação de encontra entre parênteses): bases alinhadas tem pontuação *+match(1)*, e bases desalinhadas tem custo -mismatch(1). Um indel em qualquer uma das sequências fora de um íntron tem custo -gap(2) (não existe custo de se abrir um buraco), e dentro de um íntron tem custo -intron(40), a menos que comece com GT e termine com AG (ou CT e AC se a direção for reversa), no qual o custo é -splice(20).

A diferença entre *splice* e *intron* pode causar alguns erros no reconhecimento do ponto final de exons. Se introns do alinhamento não tiverem GT/AC (CT/AC) como limites, então pode existir erro de sequênciamento. Em alguns casos, exons menores do que *splice* podem ser perdidos.

O algoritmo utilizado é o seguinte: Seja $X(i,j)$ a pontuação do melhor alinhamento local terminando na base i no pedaço de sequência e na base j na sequência genômica. Seja $B(i)$ a pontuação do melhor alinhamento local encontrado até o momento que se termina na base i no pedaço de sequência. Seja $C(i)$ a coordenada da sequência genômica à qual se refere $B(i)$. Sejam $S(i)$ e $G(j)$ os nucleotídeos respectivamente nas posições i no pedaço de sequência e j na sequência genômica.

Tem-se a seguinte equação:

$$X(i,j) = max \begin{cases} X(i-1,j) - gap \\ X(i-1,j-1) + D \\ X(i,j-1) - gap \\ B \\ 0 \end{cases}$$

com

$$D = \begin{cases} match & \text{se S(i)=G(j)} \\ -mismatch & \text{caso contrário,} \end{cases}$$

$$B = \begin{cases} B(i) - splice & \text{se C(i),j é um par doador-receptor} \\ B(i) - intron & \text{caso contrário} \end{cases}$$

e

$$(B(i), C(i)) = \begin{cases} (X(i,j), j) & \text{se X(i,j)>B(j)} \\ (B(i), C(i)) & \text{caso contrário.} \end{cases}$$

O termo B representa o custo do melhor alinhamento local terminado com um intron em i,j, de forma que $X(i,j)$ é o custo do melhor alinhamento geral terminado em i,j.

O programa utiliza um algoritmo com espaço linear usando uma estratégia de divisão e conquista [84, 60] para limitar o consumo de memória:

1. Faz uma primeira passada utilizando o algoritmo Smith-Waterman, que produz alinhamentos locais, para encontrar o início e fim dos segmentos de pontuação máxima. Subsequências correspondentes a estes segmentos são extraídas.

2. Se o produto do tamanho das subsequências é menor do que um limiar definido pelo usuário, os segmentos são realinhados utilizando o algoritmo de Needleman-Wunsch [86], que produz alinhamentos globais.

3. Se o produto ultrapassa o limiar, o alinhamento é feito recursivamente dividindo a sequência ao meio e encontrando a posição no genoma que alinha com o ponto médio. Este procedimento é repetido até que o produto dos comprimentos for menor do que o limiar. As sequências divididas são alinhadas separadamente e intercaladas.

4. Efetua-se uma busca dos segmentos na sequência genômica, de forma direta e reversa, levando em conta uma direção de splicing direta (ou seja, com consenso GT/AG). Depois, o alinhamento é feito levando se em conta splicing com diresão reversa (consenso CT/AT).

O programa levou 11s de processamento para alinhar um EST de 519bp com um cosmídeo de 33670bp retirado do cromossomo humano 16 em uma estação Digital Alpha 255/233.

A.12 A computer program for aligning cDNA sequence with genomic DNA sequence [32]

Com a grande quantidade de ESTs e sequências genômicas, é cada vez mais comun que se queira alinhar os dois. Este trabalho apresenta um programa chamado *sim4*, que tem como objetivo obter este alinhamento de forma eficiente e precisa, assumindo que as diferenças entre as sequências a serem alinhadas se resumem a presença de íntrons na sequência genômica e erros de sequênciamento em ambas as sequências.

O algoritmo utilizado tem os seguintes passos:

- Determinação de segmentos com pares de alta semelhança (HSP): Aplicando o mesmo algoritmo utilizado pelo software blast [4], consiste em encontrar sequências de tamanho 12 que casam perfeitamente, e extendendo-as em ambos os sentidos, utilizando apontuação de 1 para um casamento e -5 para um não casamento, até que a inclusão de novas bases não melhore a pontuação obtida.

- Seleção de um conjunto de HSPs que poderiam representar um gene: utiliza-se um algoritmo de programação dinâmica que seleciona a melhor sequência de HSPs

levando em conta as seguintes restrições: suas posições iniciais no EST estão ordenadas de forma crescente e as diagonais na matriz de alinhamento de dois HSPs consecutivos são ou muito semelhantes ou diferentes o suficiente para ser um possível íntron. A pontuação de cada HSP é multiplicada por 100 e subtraída da diferença entre as diagonais de HSPs consecutivos para determinar a pontuação da sequência.

- Encontrar os limites dos exons: se dois núcleos de exons consecutivos se sobrepõe, os finais são removidos, tentando se obter um íntron cujo padrão seja ou GT ... AG ou CT ... AC. Caso os núcleos não se sobreponham, utiliza-se um algoritmo guloso(Miller e Myers [83]), visando juntá-los e ajustar a junção de forma a se obter o padrão de íntron definido acima. Caso a extensão falhe, a região de separação é comparada com outros HSPs, utilizando critérios menos restritos de semelhança (iniciando a busca com uma sequência de 8 pares de bases idênticos). Os exons do início e do fim são extendidos utilizando os mesmos passos descritos acima.

- Determinar o alinhamento de cada exon, utilizando o método de Chao [19].

A ferramenta foi utilizada para alinhar dados obtidos através do projeto de sequenciamento da *Drosophila Melanogaster*, desenvolvido pelo Berkeley Drosophila Genome Project (BDGP) [10].

Para se avaliar o desempenho da ferramenta, foram utilizadas sequências bem conhecidas obtidas de genes de Drosophila, utilizados para treinar o software de detecção de genes *Genie* [100]. Destas sequências, foram obtidos 184 CDS, que foram passados para o programa para que este alinha-se cada CDS com sua sequência original: o erro foi medido pela semelhança entre o limite íntron-exon obtido pelo software e o limite anotado. Sem utilizar nenhum tipo de otimização, o programa acertou exatamente 166 alinhamentos. Dos 18 alinhamentos errados, 11 tinham um erro entre 1 e 10 bases, 6 tinham um erro entre 11 e 20 bases e 1 tinha um erro de 25 bases. Utilizando otimização, o programa acertou 172 sequências. Todos estes erros eram provocados pelo mesmo problema: o programa tem dificude de alinhar corretamente pequenso exons iniciais e finais.

Foram feitos testes comparativos com duas outras ferramentas para alinhamento de mRNA, cDNA ou ESTs com as sequências genômicas de origem: *est_genome* [80] e *est2gen* [12]. Utilizou-se o mesmo conjunto de dados descrito acima: *est2gen* demorou 156 segundos por sequência e não cometeu nenhum erro. *est_genome* demorou 20 segundos por sequência e acertou 143 alinhamentos, e *sim4* demorou 0.06s por sequência, e no modo normal acertou 166 alinhamentos e no modo otimizado acertou 172 alinhamentos.

Em outro teste, foram inseridos erros em sequências, com taxas de 1, 3 e 5% em séries de 500bp. Os resultados obtidos mostram que o programa produz o alinhamento correto apesar dos erros.

O software foi desenvolvido visando produzir alinhamentos corretos levando em conta introns e erros de sequênciamento, e não foi previsto para analisar corretamente mutações ligadas ao processo de evolução. Para se obter dados empíricos, foram feitos alinhamentos de 16 mRNAs humanos com sequência genômica ortóloga de ratos. Em 9 genes o programa acertou 100% das sequências codificadoras, e só acertou 100% do gene em 3 mRNAS.

A estratégia implementada no pacote pode ser integrada de forma interessante em uma grande variedade de pacotes computacionais de análise de sequências. Um deles é utilizar o programa para comparar uma sequência genômica com bases de dados de ESTs.

O pacote pode ser obtido no sítio do Globin Gene Server [40].

A.13 Spidey: A Tool for mRNA-to-Genomic Alignments [122]

Seqüências expressas, ou ESTs, são a chave para o entendimento do funcionamento interno de um organismo. Porém, para que se entenda completamente o seu funcionamento, sequências expressas tem que ser postas no seu contexto genômico. Estima-se que o ser humano possui entre 30000 e 35000 genes [23], fazendo com que o processo de Alternative Splicing seja um fator importante na geração da diversidade fenotípica humana. Por isso, alinhadores de mRNA com genomas são de grande importância.

Este trabalho tem como objetivo apresentar a ferramenta *Spidey*, que produz alinhamentos de mRNAs com genomas. O sistema foi desenvolvido em C e foi incorporado ao NCBI Toolkit [90], está disponível na forma de programa para download ou web service no site do NCBI [122]. Seus principais objetivos são produzir bons alinhamentos a despeito do tamanho de íntrons (regiões não codificadoras) e não gerar erros devido a genes parálogos (genes que tem uma origem comum e aparecem no mesmo genoma) e pseudo-genes. Para efetuar os alinhamentos, são utilizadas as ferramentas BLAST e DotView.

Primeiramente, o programa cria janelas genômicas, visando evitar erros com pseudo-genes e parálogos. Para isso, obtém os mRNAs e compara com sequências genômicas utilizando BLAST com alto grau de semelhança ($e = 10^{-6}$). Os alinhamentos obtidos são ordenados pela pontuação de forma crescente. Aplica-se um algoritmo recursivo para gerar as janelas, que pega o primeiro alinhamento da lista (potanto aquele com maior pontuação), cria uma janela para ele na sequência genômica, e o compara com todos os alinhamentos da lista: para cada alinhamento que for consistente com o primeiro (ou seja, da mesma fita de mRNA, com coordenadas não sobrepostas), cria uma nova janela. Nas iterações subsequentes, o algoritmo trata os alinhamentos restantes, colocando-os em suas próprias janelas, até que não sobre nenhum alinhamento.

Uma vez que as janelas foram definidas, obtém se regiões genômicas com grande probabilidade de serem exons. Efetua-se então um BLAST com grau de semelhança menor que o inicial ($e = 10^{-3}$), de todas a sequências de mRNA com as regiões genômicas definidas por cada janela, visando criar um alinhamento que cubra todo o mRNA. Utiliza-se um algoritmo guloso para se obter o melhor alinhamento possível. Faz-se uma análise cuidadosa para verificar se não existem buracos no alinhamento obtido. Caso existam, aplica-se o BLAST novamente com baixíssimo grau de similaridade ($e = 1$), e caso ainda sobre algum buraco, utiliza-se o DotView.

Ao final do procedimento, o programa analisa o alinhamento obtido para calcular a porcentagem de semelhança por exon, o número de buracoes por exon, a porcentagem de cobertura do mRNA, presença de poly(A) e outros. Se a porcentagem de semelhança e a porcentagem de cobertura estiverem acima de um dado valor de corte, o programa gera um relatório final.

Para analisar a performance do programa, foram feitos testes comparativos com duas outras ferramentas semelhantes: o *sim4* [32] e o *est_genome* [80].

O primeiro teste consistiu em se fazer alinhamentos com sequências de referência, de onde foram extraídos 646 mRNAs anotados, contendo um total de 3915 exons. Estes mRNAs foram então alinhados com a sequência original (tendo portanto uma semelhança de 100%):

- *Spidey* reconheceu 3873 exons, dos quais 98.7% estavam corretos.
- *sim4* reconheceu 3909 exons, dos quais 97.9% estavam corretos.
- *est_genome* reconheceu 3716 exons, dos quais 97.4% estavam corretos.

O segundo teste consistiu em se alinhar mais de 11000 seqüências de referência com contigs do genoma humano obtidos no site do NCBI [85], na versão de 01/04/2001. Aplicou-se o programa MEGABLAST para cada sequência com todos os contigs: foram aceitos os hits de no mínimo 75 bases com 97% de semelhança, gerando uma lista de mRNAs potenciais para cada contig. Esta lista foi dada como parâmetro de entrada para o Spidey, que alinhou os mRNAs potenciais com cada contig: os alinhamentos foram aceitos se cobrissem pelo menos 90% dos mRNAs, tivessem ao menos um exon com semelhança superior ou igual a 99% e não tivessem nenhum exon com semelhança inferior a 95%. Assim, foram encontrados 7848 mRNAs chamados de modelos, dos quais 7664 representavam mRNAs únicos. (ou seja, que possuiam apenas um alinhamento possível). Foram então feitos testes de alinhamento visando analisar a capacidade do software de alinhar os mRNAs com os contigs corretos.

Um terceiro teste foi efetuado de forma a verificar se o programa Spidey não gerava erros ao alinhar mRNAs em locais próximos a genes parálogos. Para isso, foram obtidos

mRNAs de clusters genéticos (genes relacionados próximos no cromossomo), e foram feitos alinhamentos entre eles. Os genes utilizados tinham uma semelhança entre 65 e 98%, em regiões entre 25 e 100% do tamanho de cada um. O programa *Spidey* alinhou corretamente os 16 mRNAs obtidos.

O último teste foi feito alinhando-se genes ortólogos (genes que tem uma origem comum e aparecem em genomas diferentes) de ratos com sequências humanas de referência. *Spidey* foi configurado no modo inter-espécie, no qual são utilizados parâmetros de configuração do BLAST diferentes (open gap=5, extended gap=1, mismatch=-1) visando criar mais buracos maiores e não penalizar muito mismatch. Os programas *sim4* e *est_genome* também foram utilizados com suas configurações normais. *Spidey* acertou 81.4% dos exons, *sim4* acertou 53.9% e *est_genome* acertou apenas 37.2%.

Em relação ao tempo de processamento, *Spidey* e *sim4* se mostraram muito superiores a *est_genome*: para alinhar um mRNA com 5164bp com um contig de 1.03Mb em uma Sun Ultra 10 300MHz com 192Mb de memória, *Spidey* levou 14s, *sim4* levou 2s e *est_genome* levou 1h21m. Para processar 35 mRNAs com suas sequências de referências, *Spidey* levou 1m11s, *sim4* levou 25s segundos e *est_genome* levou 2h56m.

A.14 A polymorphism in Endostatin, an Angiogenesis Inhibitor, Predisposes for the Development of Prostatic Adenocarcinoma [64]

Câncer de próstata é o segundo tipo de câncer que mais mata homens nos Estados Unidos. Por isso, o desenvolvimento de metodos que permitam indentificar pacientes com alta predisposição para a doença é essencial para se aumentar os índices de cura, através de diagnósticos precoces.

Apesar da maioria dos casos de câncer acontecer de forma esporádica, observa-se que o aspecto genético pode aumentar a predisposição em certas familias. Atualmente já foram mapeados pelo menos 5 genes que podem causar susceptibilidade para se contrair câncer.

O processo de angiogenese, onde são formados novas veias sanguíneas, é fundamental no processo de progressão do câncer e metastase. Algumas substâncias funcionam como inibidor deste processo. Entre elas, a endostatina, cuja aplicação em ratos levou a regressão de tumores.

Foi feito um estudo de associação que concluiu que um SNP no gene que gera a endostatina pode aumentar a probabilidade de se contrair o câncer de próstata: individuos heterozigotos N104 tem 2.5x mais chances de desenvolver um câncer de próstata do que indivíduos homozigotos D104. Este resultado foi idêntico tanto para caucasianos quanto para negros.

A.15 Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification [44]

Estudos clínicos com o inibidor da kinase tirosina ABL, STI571, em indivíduos com leucemia mielóide crônica demonstraram que muitos pacientes em estado avançado da doença repondem a droga apenas inicialmente. Através da análise bioquímica e molecular de material clínico, descobriu-se que a resistência a droga esta associada com a reativação da tradução do sinal BCR-ABL em todos os casos examinados.

Em 6 dos 9 pacientes analisados, a resistência estava associada a um SNP: a substituição de treonina por isoleucina foi suficiente para criar resistência ao STI-571. Em 3 pacientes, a resistência estava associada à amplificação progressiva do gene BCR-ABL. Estes estudos fornecem evidências que cânceres geneticamente complexos dependem de um evento inicial, e sugerem a criação de uma estratégia para identificação de possíveis resistências a inibidores de STI-571.

A.16 Dynamic allele-specific hybridization [59]

Este trabalho apresenta o método DASH (Dynamic Allele Specific Hybridization): um protocolo experimental para determinar a presença ou não de um dado SNP em uma sequência genética, sem a necessidade de se fazer sequenciamento desta última.

O principio do método é o seguite. O gene de interesse é amplificado através de PCR. A seguir, adiciona-se milhares de cópias da sequência de teste: pequeno trecho da região complementar que contém o SNP a ser estudado. Em baixas temperaturas, os dois trechos complementares de DNA se juntam. Acrescenta-se então um marcador que quando excitado emite fluorescência proporcinal à quantidade de duplas fitas de DNA presentes na solução. Aquecidas, as duplas fitas que contém o SNP tendem a se desnaturar mais rapidamente que as duplas fitas que não contém SNP (mais estáveis), fazendo com que a luz emitida pelo marcador diminua rapidamente, permitindo detectar-se facilmente a presença do SNP analisado.

Observou-se que o tamanho da sequência de prova influe no resultado do teste: sequências com tamanho variando de 15 a 21 bases, com o SNP a ser detectado posicionado no centro fornecem melhores resultados.

Observou-se também que este método detecta bem os quatro principais tipos de polimorfismos existentes (C⇔A, T⇔A, C⇔G e T⇔G).

A.17 Base-Calling of Automated Sequencer Traces Using *Phred* I. Acuracy Assesment [31]

Grande parte dos procedimentos de sequenciamento de DNA são feitos utilizando o método de Sanger [103]. Este processo consiste em se obter a sequência nucleotídica de interesse em fita simples, e se gerar uma sequência (chamada de primer) complementar a uma região de interesse. Faz se uma série de reações de forma a se gerar uma população de fragmentos de tamanho variável, todos complementares à sequência de interesse e iniciando com o primer. Adiciona-se terminadores distintos, que se acoplam às últimas bases de cada fragmento, permitindo que se identifique qual base termina um dado fragmento. Estes fragmentos são depositados em um gel e são separados por tamanho, através do processo de eletroforese, permitindo que se faça a leitura da sequência nucleotídica de interesse.

No processo de sequenciamento automático [107], os terminadores são marcados com substâncias que emitem fluorescência, uma para cada base. Estes marcadores podem se acoplar ao primer ou então à cadeia de terminadores [98]. Quando os fragmentos são colocados no gel, um laser é utilizado para excitar os marcadores, que emitem luz coletada por sensores. Assim, é possível se medir o sinal luminoso gerado em cada posição do gel após a eletroforese, determinando quais bases se encontram em determinadas posições: este processo é chamado de base-calling.

O sinal obtido do gel é processado e em geral apresentado sob forma de cromatograma (Figura A.1) que consiste em quatro curvas de cores diferentes, uma para cada base diferente (A,T,C ou G), que devem ser lidas da esquerda pra direita. De forma simplificada, podemos dizer que cada pico no cromatograma corresponde a uma base distinta.

Em um cromatograma ideal, os picos ficam espaçados de forma regular, e não se sobrepõe, de forma a que cada pico corresponda a apenas uma base. Porém, cromatogramas reais não apresentam sinais ideais, por um série de fatores ligados ao processo químico utilizado, à problemas na eletroforese e problemas na análise da imagem obtida.

Devido à migração anômala de fragmentos muito pequenos, os primeiros pontos obtidos ($\sim$ 50) contém muitos ruídos e ficam muito próximo uns dos outros. O mesmo acontece com os últimos pontos (correspondentes aos maiores fragmentos). Nas regiões centrais, o problemas mais comum é a compressão de picos, que acontece quando dois fragmentos se juntam, e migram de forma mais rápida do que o esperado, fazendo com que alguns picos do sinal obtido fiquem mais a esquerda do que o esperado. Este fenômeno ocorre mais em sequências ricas em pares GC do que sequências ricas em pares AT.

O objetivo do prcesso de *base-calling* é de se obter a melhor sequência possível, levando-se em conta os problemas descritos acima. Um dos pacotes computacionais mais antigos a executarem este processamento fazia parte do sistema da primeira máquina sequenciadora

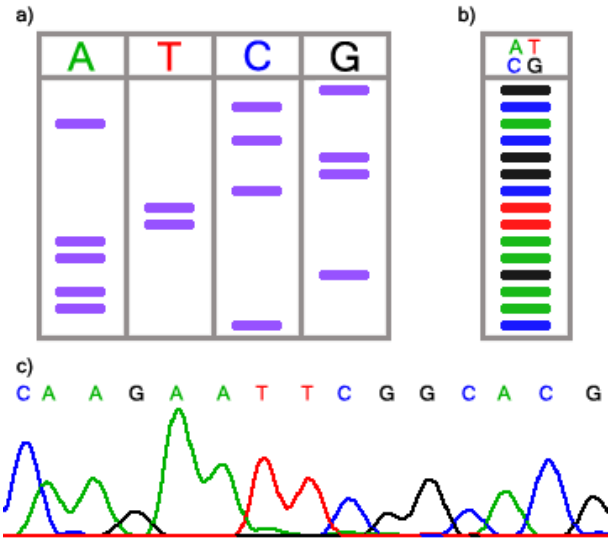

Figura A.1: Leitura de bases. Em (a) pode se observar o resultado de eletroforese onde cada terminador foi analisado separadamente. Em (b), o resultado da eletroforese com todos os terminadores analisados juntos. Em (c), tem se o cromatograma final obtido a partir da leitura da placa.

da ABI [22]. Este pacote, cujo algoritmo não foi detalhado foi publicado num documento de descrição do software [92], obtém resultados de alta qualidade, e é considerado como ponto de referência para análise de outros métodos. Com o aumento de número de sequenciamentos, vários programas de *base-calling* tem sido desenvolvidos [39, 42, 41, 11], porém nenhum deles obtém resultados de qualidade comparável ao ABI.

Este trabalho apresenta o pacote computacional *phred*, que efetua um processamento de base-calling e obtém precisão superior à precisão do ABI. O algoritmo utilizado é baseado em 4 fases, e foi refinado de forma empírica, a partir da análise de dados processados.

A primeira fase do algoritmo tenta definir a localização de picos ideais, utilizando métodos de Fourier. Inicialmente, o sinal é analisado à procura de picos. Um pico é identificado como sendo um ponto de valor máximo ou, caso este não exista, o ponto médio entre dois pontos de inflexão consecutivos. Um pico é levado em conta apenas se exceder 10% do pico anterior e se for o maior valor entre os 4 sinais obtidos do gel. Os pontos são então varridos à procura de regiões de espaçamento uniforme de picos, onde considera-se como região janelas de 200 pontos. Para cada região centrada num pico detectado anteriormente, determina-se a distância entre picos adjacentes, assim como a média e o desvio padrão das distâncias. A região que tiver menor desvio padrão é considerada como região inicial na busca de picos ideais.

A segunda fase do algoritmo consiste em se localizar picos observados nos pontos obtidos. Os 4 sinais são varridos em busca de regiões concavas, satisfazendo a equação $2v(i) \geq v(i+1) + v(i-1)$ onde $v(i)$ é o valor do sinal no ponto i. Para cada região encontrada, os valores dos sinal são somados para se determinar a área do pico: se esta exceder 10% da média das áreas dos 10 últimos picos observados, e 5% da área do pico imediatamente anterior, então o pico é definido como um pico observado. A relação entre a área do pico encontrado e a média dos 10 picos anteriores é armazenada como a área relativa do pico.

A terceira fase do algoritmo consite em se designar um pico observado na segunda fase para cada pico ideal predito na primeira fase. Esta fase consiste em 3 estágios distintos:

1. procurar por relações triviais
2. utilizar programação dinâmica para alinhar picos preditos e observados que não foram casados na primeira fase
3. alinhar picos preditos e observados que não foram casados nas duas primeiras fases mas que possam representar bases reais.

Para cada pico predito, procura-se por todos os picos observados que são mais próximos deste pico predito do que de qualquer outro, e destes picos observados, aquele com maior área relativa é designado como sendo o *best_obs_peak*. As áreas relativas dos picos observados são então recalculadas, utilizando o valor corrente da área média dos 10 últimos picos designados como *best_obs_peak*.

A quarta fase do algoritmo consiste em se analisar os picos não considerados como bases e que claramente correspondem a bases da sequência. Isto pode ocorrer quando uma compressão, ou ruído extenso ou um erro no processamento do gel interfere com a predição de picos, resultando num número de bases inferior à realidade. Para se recuperar estes picos, observa-se se são respeitados os seguintes critérios:

1. o pico tem o maior valor dentre os 4 sinais na mesma região.
2. o pico tem um tamanho mínimo.
3. o pico não está particionado.
4. o pico está cercado por picos resolvidos.
5. o pico se encontra num local tal que adicioná-lo melhora o espaçamento entre picos.

Para se comparar a qualidade dos resultados obtidos com *phred* e com o software ABI, foram utilizados 3 conjuntos de sequências bem conhecidas e anotadas: os conjuntos 1 e 2

são de sequências curtas com qualidade boa, e tem padrões de erros similares. O conjunto 3 é uma sequência longa com alta qualidade As análises foram feitas de forma separada para reads com marcadores no primer e reads com marcadores no terminador.

De forma geral, o total de taxas de erro do *phred* é inferior à taxa de erro do ABI. Para marcadores nos primers, *phred* produziu $\sim 41\%$ menos erros que o ABI nos conjuntos 1 e 2, e $\sim 52\%$ menos erros no conjunto 3. Para os reads com marcador no terminador, *phred* produziu 39% menos erros nos 3 conjuntos. As mesmas comparações foram feitas nos conjuntos após terem sido retiradas as 5% de sequências com maior taxa de erro no *phred* e as 5% de sequências com maior taxa de erro no ABI: foram observados entre 40% e 50% menos erros produzidos pelo *phred*. Foram analisados também a taxa erros de inserção e remoção de bases, na qual o *phred* obteve melhores resultados também. Em relação às substituições, *phred* produz aproximadamente o mesmo número de erros que o ABI.

Acredita-se que o algoritmo de detectação de picos observados e o método de alinhamento de picos observados com picos preditos irão funcionar de forma correta independentemente da fonte do sinal analisado. Porém isto não é válido para o algoritmo de predição de picos, que foi ajustado para analisar sinais processados por máquinas ABI.

Alguns aprimoramentos ainda devem ser feitos no pacote. Em particular, a capacidade de resolver compressões causadas por sequências de CC e GG, e o aumento da qualidade nas regiões iniciais e finais do sinal que são muito úteis no processamento de aplicações que utilizam sequências únicas, como ESTs.

O pacote foi escrito em C, e está disponível gratuitamente para fins acadêmicos.

A.18 Base-Calling of Automated Sequencer Traces Using *Phred* II. Error Probabilities [30]

A qualidade das bases sequenciadas automaticamente pode variar devido a um grande número de fatores [31], e para que se faça um uso efetivo dos resultados obtido é necessário que se tenha alguma medida de confiabilidade adequada. Funções de probabilidade de erro específicas por posição [72] são muito úteis para esta finalidade, e em conjunto com um montador de sequências, podem melhorar a precisão da montagem final, e obter melhores elementos para definir regiões do sequenciamento a serem melhorados.

Alguns desenvolvedores de algoritmos para base calling [39, 42, 11] criaram medidas de qualidade para as bases obtidas, mas os resultados não foram divulgados. O trabalho mais completo foi desenvolvido por Lawrence e Solovyev [72], que definiram os parâmetros importantes para análise de qualidade e determinação de probabilidades de erro.

Este trabalho apresenta um procedimento diferente para a determinação de qualidade utilizado pelo programa *phred* que não requer a utilização de hipóteses de distribuição

normal multivariada, utiliza parâmetros computados a partir de parâmetros de janelas do sinal de leitura e enfatisa o cálculo das probabilidades de regiões de alta qualidade do sinal (taxa de erro ≤ 0.01).

Define-se a qualidade q de uma base como sendo

$$q = -10 \times \log_{10}(p)$$

onde p é a probabilidade de erro estimada para uma dada base. Assim uma base com probabilidade de erro de 1/1000 fica com qualidade 30 (quanto maior a qualidade, menor a probabilidade de erro).

Probabilidades de erro devem ser atribuídas por métodos que não requerem nenhum conhecimento prévio da sequência real, e devem ser válidas no sentido de corresponder a taxas de erros observadas. Uma das formas de se atribuir probabilidades de erro é de atribuir a cada base o mesmo valor, dada pela razão entre o número de erros observados e o número de bases existentes. Uma outra forma é de se identificar subconjuntos, nos quais se observam quantidades de erro diferentes. Ambos os métodos são válidos, porém o segundo é mais apropriado por discrimar conjuntos mais precisos de outros menos precisos. Pode se definir uma função que mede a habilidade de discriminação de um algoritmo, dada por

$$P_r = \left(\frac{|B_r|}{|B|} \right)$$

onde B_r é o maior conjunto de bases com taxa de erro inferior a r e B é o conjunto total de bases. O objetivo deste trabalho é desenvolver um método de probabilidade de erro válido e com alto poder de discriminação para pequenos valores de r ($r \leq 0.01$).

Os erros produzidos por programas de base-calling em geral são causados por erros de análise em picos em uma região do sinal, e indicações do erro podem aparecer na vizinhança do pico errado, e não forçosamente no pico em questão. Por isso, uma análise de uma janela tende a ser mais efetiva para detecção de erros. No algoritmo apresentado, 4 parâmetros são levados em conta:

1. Espaçamento entre picos: dada uma janela de 7 picos centrada no pico corrente, é a razão entre o maior e o menor espaço entre picos. O menor valor encontrado corresponde a picos espaçados regularmente.

2. Razão entre bases definidas e não definidas: dada uma janela de 7 picos centrada no pico corrente, é a razão da amplitude do maior pico não determinado com a amplitude do menor pico não determinado. Um pico não determinado é um pico que não foi definido como um pico predito [31]. Se o pico tiver sido determinado como N, então o programa define o valor 100.

3. Igual ao ítem anterior, em uma janela de 3 picos centrada no atual.

4. Resolução de picos: O número de bases entre a base corrente e a mais próxima base não determinada vezes -1. Uma base é não determinada se tiver sido chamada de N.

Sejam r, s, t, e u os parâmetros acima e $r(b)$, $s(b)$, $t(b)$ e $u(b)$ os valores calculados de cada parâmetro em uma dada base b. Define-se os valores de limiar r_i, s_j, t_k, e u_l. Um conjunto de valores de limiar (r_i, s_j, t_k, u_l) é dito optimal para uma dada taxa de erro R se o conjunto de bases B, tal que $\forall b \in B, r(b) \leq r_i, s(b) \leq s_j, t(b) \leq t_k, u(b) \leq u_l$, tem uma taxa de erro inferior ou igual a R, e nenhum outro conjunto de limiares possui um conjunto de bases maior que $|B|$ com taxa de erro inferior ou igual a R.

O algoritmo utilizado produz uma tabela de referência formada por um conjunto de linhas onde cada linha contém um conjunto de valores de limiar (1 para cada parâmetro), junto com uma probabilidade de erro e um valor de qualidade.

Para um dado parâmetro p, são definidos valores $p_1 < p_2 < \ldots < p_{50}$ tal que $\forall b$, $p_1 < p(b) < p_{50}$ e o número de bases que satisfaz $p_{i-1} < p(b) \leq p_i$ é aproximadamente o mesmo para $\forall i$.

Define-se um corte como sendo um 4-upla (i, j, k, l) com i, j, k e l entre 0 e 50 (existem portanto $50^4 = 6.25$ milhões de cortes): cada corte possui um conjunto de limiares r_i, s_j, t_k, e u_l. Para cada corte, define-se $err_{(i,j,k,l)}$ como sendo o número de bases erradas b abaixo do corte, ou seja, satisfazendo $r(b) \leq r_i, s(b) \leq s_j, t(b) \leq t_k, u(b) \leq u_l$. De forma similar,define-se $corr_{(i,j,k,l)}$ o número total de bases corretas abaixo do corte. A taxa de erro abaixo do corte $e_{(i,j,k,l)}$ é dada por

$$e_{(i,j,k,l)} = \frac{1 + err_{(i,j,k,l)}}{1 + err_{(i,j,k,l)} + corr_{(i,j,k,l)}}$$

e a qualidade correspondente é dada por

$$q_{(i,j,k,l)} = -10 \times \log_{10}(e_{(i,j,k,l)})$$

A construção da tabela é feita com os seguintes passos:

1. Achar o corte (i, j, k, l) para o qual $q_{(i,j,k,l)}$ é o maior. Em caso de empate, selecionar aquele cujo valor $err_{(i,j,k,l)} + corr_{(i,j,k,l)}$ é maior, e em caso de novo empate, selecionar aquele cuja soma dos índices é maior. Colocar $e_{(i,j,k,l)}$, $q_{(i,j,k,l)}$ e os parâmetros r_i, s_j, t_k, e u_l na tabela e retirar o corte da lista.

2. Para cada corte restante (i', j', k', l'), ajustar $err_{(i,j,k,l)}$ e $corr_{(i,j,k,l)}$ removendo as bases abaixo de (i, j, k, l) e recalcular $e_{(i',j',k',l')}$ e $q_{(i',j',k',l')}$ usando os novos valores. Se *err* e *corr* forem 0 para todos os cortes restante, parar. Senão voltar para passo 1.

O processo de atribuição de qualidades em bases consiste em se calcular o valor dos quatro parâmetros e procurar uma linha na tabela contendo um corte tal que os parâmetros calculados estejam abaixo: o valor de qualidade associado à linha é atribuído à base. Se nenhum corte satisfazendo as condições for encontrado, então atribue-se qualidade 0 à base.

O pacote foi escrito em C, e está disponível gratuitamente para fins acadêmicos.

A.19 PolyPhred: automating the detection and genotyping of single nucleotide substitution using fluorescence-based resequencing [89]

SNPs são a forma mais frequente de variação no genoma humano, e a identificação destas variações representa um papel importante no estudo da evolução da população humana, e na exploração do relacionamento entre a estrutura genômica humana e a correlaçao genótipo-fenótipo.

Este trabalho apresenta o programa PolyPhred, capaz de detectar SNPs, utilizando saídas dos programas p*phred* [31], que executa um algoritmo de base calling e *phrap* que monta sequências de consenso. A saída produzida por *polyphred* pode ser analisada utilizando o editor *consed*.

A idéia utilizada pelo programa para encontrar SNPs é a seguinte: ao se comparar sequências homozigotas de sequências heterozigotas, observam-se duas grandes diferenças no segundo caso:

1. Uma significante redução ($\sim 50\%$) no tamanho do pico normalizado observado no cromatograma.

2. Um segundo pico menor que o principal na posição em questão [71, 95].

Assim, o programa *polyphred* analisa a áreas normalizadas e as qualidades de cada base obtidas através do programa *phred*, para cada posição de uma sequência alinhada montada pelo programa *phrap*: se o programa detecta um pico menor que um certo valor e a saída produzida por *phred* indica um segundo pico, então o programa grava a posição como sendo um candidato à SNP.

Ao se analisar os resultados obtidos experimentalmente, observa-se que a precisão do resultado fornecido pelo programa depende por um lado do tipo de protocolo experimental utilizado para se obter o sequenciamento (marcador associado ao primer ou marcador associado ao terminador de sequência), e por outro lado da qualidade das bases obtidas pelo programa *phred*. A razão entre verdadeiros-positivos e falsos-positivos (confirmada

por resequenciamento posteriormente) variava de 1:11 em qualidade 20 a 2:1 em qualidade 40 para sequências obtidas através de marcadores em primer, e de 1:13 em qualidade 20 a 1:1 em qualidade 40 para sequências obtidas através de marcadores em terminador.

A.20 A general approach to single-nucleotide polymorphism discovery [77]

Este trabalho apresenta um método para detecção de SNPs, utilizando o programa *polybayes* desenvolvido para esta finalidade, que utiliza um algoritmo de inferência Bayesiana para calcular a probabilidade de um dado alelo ser polimórfico. Para a validação da metodologia, fez-se uma análise dos resultados obtidos alinhando-se ESTs humanos com sequências genômicas terminadas ou em processo de sequenciamento.

Foram obtidos 1954 ESTs de dbEST disponíveis com cromatograma, que foram processados com *phred* para se determinar as qualidades das bases (foram utilizadas todas as bases para as análises subsequentes).

Foram retirados EST parálogos, deixando apenas aqueles que se originaram da mesma região genômica. A determinação se um EST era parálogo ou não foi feita observando se o número de diferenças entre o EST e a sequência genômica de referência era consistente com a taxa estimada de variação polimórfica em oposição à diferença entre duas sequências provenientes de duplicação de cromossomos.

Estimou-se que grande parte das sequências parálogas apresentava uma taxa de diferenças maior que $P_{PAR} = 0.02(2\%)$ em comparação com a taxa média de diferenças entre duas sequências polimórficas $P_{POLY,2} = 0.001(0.1\%)$. Em um alinhamento de tamanho L entre duas sequências espera-se portanto $L \times P_{POLY,2}$ diferenças devido a polimorfismos e $L \times P_{PAR}$ diferenças devido ao fato de serem parálogos. Em ambos os casos espera-se um número E de erros adicionais. Portanto considera-se dois modelos: um EST pode ser polimórfico($Model_{NAT}$), e espera-se $D_{NAT} = L \times P_{POLY,2} + E$ erros, ou pode pode ser parálogo($Model_{PAR}$), e espera-se $D_{PAR} = L \times P_{PAR} + E$ erros. Estima-se que a probabilidade de discrepância d entre duas sequências segue a distribuição de Poisson com parâmetro $\lambda = D_{NAT}$ para $Model_{NAT}$ e $\lambda = D_{PAR}$ para $Model_{PAR}$.

A probabilidade a posteriori $P(Model_{NAT}|d)$ de um EST representar uma variante polimórfica é dada por:

$$P(Model_{NAT}|d) = \frac{1}{1 + e^{(D_{NAT} - D_{PAR})} \times (D_{PAR}/D_{NAT})}$$

ESTs que obtiveram $P_{NAT} \geq P_{NAT,MIN} = 0.75$ foram considerados adequados. Assim, foram eliminados 23% dos ESTs.

Os restantes foram então agrupados em 147 clusters, e em cada cluster, produziu-se primeiramente um alinhamento entre cada EST e a sequência genômica com cross_match, e depois produziu-se um alinhamento múltiplo, propagando-se os buracos e inserções em cada EST.

A detecção de SNPs no alinhamento múltiplo foi efetuado avaliando-se a probabilidade de heterogeneidade de nucleotídeos em uma secção do alinhamento múltiplo. Cada um dos nucleotídeos $S_1, \ldots, S_N$ em uma secção transversal do alinhamento de N sequências $R_1, \ldots, R_N$ pode ser qualquer uma das 4 bases de DNA A, C, G ou T. A probabilidade $P(S_i|R_i)$ que um nucleotídeo S_i seja A, C, G, ou T é estimada à partir da probabilidade de erro $P_{ERROR,i}$ obtida da qualidade da base. Atribui-se $(1 - P_{ERROR,i})$ para a base determinada e $P_{ERROR,i}/3$ para cada uma das 3 outras bases. Inserções e remoções não são consideradas, devido à falta de valor de qualidade.

Cada permutação heterogênea é classificada de acordo com sua multiplicidade nucleotídica, a variação específica e a distribuição de alelos. Foi utilizado o valor $P_{POLY} = 0.003$ (um locus polimórfico a cada 333 bp) como a probabilidade total a priori de que um locus é polimórfico [17, 29] (taxa de 1/1000 polimorfismos entre quaisquer duas sequências). Este valor foi distribuído entre as bases para criar uma probabilidade a priori $P_{Prior}(S_1, \ldots, S_N)$ para cada permutação. Um valor a priori de $(1 - P_{POLY})/4$ foi atribuído a cada uma das 4 permutações não polimórficas, correspondendo a uma composição de base uniforme $P_{Prior}(S_i)$.

A probabilidade Bayesiana a posteriori de uma permutação em um nucleotídeo em particular foi calculada considerando 4^N permutações diferentes como conjunto de modelos conflitantes:

$$P(S_1, \ldots, S_N|R_1, \ldots, R_N) = \frac{F(S_1, \ldots, S_N)}{G(S_{i1}, \ldots, S_{iN})}$$

onde

$$F(S_1, \ldots, S_N) = \frac{P(S_{i1}|R_1)}{P_{Prior}(S_{i1})} \times \ldots \times \frac{P(S_{iN}|R_{iN})}{P_{Prior}(S_{iN})} \times P_{Prior}(S_1, \ldots, S_N)$$

e

$$G(S_{i1}, \ldots, S_{iN}) = \sum_{\forall s \in (S_{i1}, \ldots, S_{iN})} \left(\frac{P(S_{i1}|R_1)}{P_{Prior}(S_{i1})} \times \ldots \times \frac{P(S_{iN}|R_{iN})}{P_{Prior}(S_{iN})} \times P_{Prior}(S_{i1}, \ldots, S_{iN}) \right)$$

A probabilidade a posteriori Bayesiana de um SNP, P_{SNP}, é a soma das probabilidades a posteriori de todas as permutações heterogêneas. O cálculo é efetuado por um algoritmo recursivo. Um locus em um alinhamento múltiplo é considerado como SNP candidato se

a probabilidade a posteriori correspondente for maior que o um valor de limiar $P_{SNP,MIN}$. Neste trabalho utilizou-se o valor 0.4 como limiar.

Das bases analisadas, 99.3% apresentaram $P_{SNP} \leq 0.1$, indicando que ou as bases eram iguais, ou existiam erros devido à baixa qualidade das bases. Foram obtidas 97 candidatos com $P_{SNP} \geq 0.4$, dos quais 56% foram reconfirmados posteriormente.

A.21 Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease [13]

Fazer a ligação entre fenótipo e genótipo é um dos objetivos fundamentais da genética. Métodos gerais para relacionar doenças com características mendelianas simples com os genes responsáveis só foram determinados no início da década de 80, quando a análise de ligação de genomas completos utilizando marcadores em DNA anônimo foi proposto pela primeira vez.

Na década de 90, aproximadamente 1200 genes responsáveis por doenças ou características humanas foram mapeados, utilizando o processo de clonagem posicional. Utilizando este processo, genes que controlam características mendelianas podem ser identificados e isolados baseados em informação apenas de características hereditárias. A identificação do gene permite obter conhecimento dos aspectos fisiológicos e moleculares de um fenótipo.

O primeiro passo para a clonagem posicional é a análise de ligação. Famílias onde o fenótipo de uma doença segrega os indivíduos são analisadas usando um conjunto de marcadores genéticos. Inicialmente os marcadores utilizados eram RFLPs. Hoje em dia SNPs estão sendo cada vez mais utilizados para este fim. O principal pré-requisito para que o mapeamento de ligação seja bem sucedido é obter um conjunto de familias onde o fenótipo é segregador e pode ser determinado com um mínimo possível de ambiguidade. Erros de diagnóstico que levam à inclusão de um indivíduo com uma doença diferente ou sem doença alguma podem causar o fracasso de um estudo de ligação, mesmo utilizando uma grande população para estudo.

Uma busca nas bases de dados Online Mendelian Inheritance in Man (OMIM), Human Gene Mutation Database (HGMD) e LocusLink retorna respectivamente 1222, 1163 e 1338 genes humanos causadores de doenças identificados. Estes números representam apenas 3% do número de genes estimados para o ser humano, e possivelmente foram os mais fáceis de serem determinados. Existem razões para se acreditar que existam muito mais genes causadores de doenças mendelianas não detectados, entre elas muitas doenças raras.

Um modo de acelerar processo de fenótipos recessivos raros é o processo de mapea-

mento de homozigocidade, no qual indivíduos da mesma familia (com grau e parentesco definido) possuindo o fenótipo a ser estudado são examinados, em busca de regiões herdadas de um ancestral comum. Este método é muito útil em populações onde casamentos entre parentes próximos são comuns, e muitas doenças raras e recessivas cujo gene ainda não foi caracterizado são encontradas apenas nesta população.

Em muitos casos, mesmo com evidências estatísticas para produzir mapeamento, nenhum resultado concreto é obtido. Para estes fenótipos, LD (Linkage Disequilibrium) tem sido usado para delimitar melhor as regiões genéticas de interesse (onde o gene ligado à doença deve ser procurado) com sucesso. Assim como para mapeamento de homozigocidade, métodos baseados em LD dependem de indivíduos descendentes de um mesmo ancestral com uma mutação

Uma análise típica de LD envolve a comparação da freqüência de alelos marcadores entre indivíduos afetados por uma doença e indivíduos sadios. Mas em alguns casos informações podem ser extraídas a partir da análise da distribuição genotípica de indivíduos doentes, sem os indivíduos de controle.

É possível também se obter bons mapeamentos através da análise de haplótipos, na qual todos os marcadores são levados em conta simultaneamente como haplótipos, e não individualmente.

A justificativa inicial para o financiamento do Projeto do Genoma Humano (Human Genome Project, HPG) foi a criação de uma infraestrutura para estudo de genes humanos ligados a doenças. A quantidade de informações geradas por este projeto e disponíveis publicamente fez o custo de se clonar um gene cair aproximadamente 100 vezes. Procedimentos de busca por dados, comparação de dados de diversos laboratórios, além de dados de espécies diferentes, antes obtidas apenas experimentalmente podem ser atualmente efetuados em bases de dados públicas de forma fácil e barata.

Algumas lições podem ser tiradas dos dados acumulados pelos projetos de mapeamento de genomas. Analisando a distribuição dos tipos de polimorfismos existentes, pode se perceber que fenótipo de doenças mendelianas são associados principalmente a alterações na seqüência normal de codificação das proteínas. Existe uma clara relação entre a importância do amino ácido substituído com a possibilidade de observação do fenótipo. Além disso, pode se dizer que determinar herança mendeliana simples pode não ser um processo simples.

A.22 Linkage Disequilibrium in Humans: Models and Data [97]

Linkage Disequilibrium (LD) desempenha um papel importante tanto como ferramenta para mapeamento de genes ligados à doenças quanto para estudos de ligação de genomas. LD pode também ser utilizado para estudar a origem e história dos seres humanos.

LD representa o grau de interdependência entre alelos em loci diferentes. Dados o alelo A no locus 1 e o alelo B no locus 2, com freqüências respectivas π_A e π_B, se os dois loci forem independentes, a freqüência esperada do haplótipo AB seria $\pi_A\pi_B$. Caso a freqüência do haplótipo for maior ou menor (implicando que os alelos tendem a ser observados em conjunto), então os dois loci são considerados como estando em LD.

Diferentes medidas foram propostas para medir o grau de LD entre dois loci. este artigo utiliza a medida r^2, também denotada por Δ^2. Dados dois loci bi-alélicos em um mesmo cromossomo, com alelos A e a no primeiro locus e B e b no segundo locus, com freqüências respectiva π_A, π_a, π_B, π_b. As freqüências dos haplótipos são denotadas como π_{AB}, π_{aB}, π_{aB}, π_{ab}. Então

$$r^2 = \frac{(\pi_{AB} \times \pi_{ab} - \pi_{Ab} \times \pi_{aB})^2}{\pi_A\pi_a\pi_B\pi_b}$$

O valor do coeficiente de desequilíbrio r^2 entre dois loci é deduzido a partir de uma probabilidade de distribuição resultante do processo evolucionário. Cromossomos de uma população estão ligados por uma árvore genealógica. Marcadores genéticos muito próximos uns dos outros tendem a ter uma genealogia muito semelhante ou igual, induzindo a uma dependência entre os alelos.

O valor esperado de r^2 é uma função do parâmetro $\rho = 4N_ec$, onde c é a taxa de recombinação entre dois marcadores e N_e é o **Tamanho Efetivo da População**. Para valores grandes de ρ, $E(r^2) \sim 1/\rho$.

Um dos grandes interesses relacionados a LD é o mapeamento de genes ligados a doenças complexas. Risch e Merikangas [101] mostram que mapeamento de associações tem maior poder de detecção de mutações ligadas à doenças do que analise de ligação. Porém, a construção de mapas requer um grande número de marcadores, sendo importante se definir a densidade necessária de marcadores. A densidade por sua vez depende do tipo de testes de associação entre marcadores que serão executados para construir o mapa.

O entendimento completo do fenômeno de LD nos seres humanos passa por um grande conhecimento da historia da demografia humana, incluindo o histórico da variação de tamanho e estrutura das diversas populações. Atualmente, os resultados obtidos com LD entre SNPs são limitados pela quantidade de dados disponíveis. Uma grande quantidade de SNPs mapeados em diversas regiões permitiria os estudo tanto de LD entre marcadores

próximos (< 5kb) e marcadores distantes (≥1Mb). Neste contexto, a obtenção de dados de muitos indivíduos para um mesmo SNP é menos útil do que a detecção de diferentes SNPs ainda não conhecidos.

A.23 LDA - a java based linkage disequilibrium analyser [28]

Linkage Disequilibrium (LD) tem sido apontado como uma ferramenta de grande utilidade para facilitar o mapeamento de genótipos complexos [6]. Porém, este processo depende de forma crucial tanto do padrão de LD quanto da extensão dos LDs no genoma humano. Este trabalho apresenta um pacote computacional escrito em Java (LDA, *Linkage Disequilibrium Analyser*) que analisa LDs.

LD representa uma associação não aleatória de alelos em loci adjacentes. Neste trabalho são levados em conta SNPs autossomais (localizados em cromossomos diferentes de X e Y) com fase desconhecida.

O procedimento executado pela ferramenta LDA é o seguinte: aplica-se um teste para verificar se os alelos em cada locus seguem o equilibrio de Hardy-Weinberg (HWE) [120]. Para os loci que seguem HWE, aplica-se um algoritmos para estimar as frequências dos 4 haplótipos nos pares de loci [62]. São aplicadas várias medidas nao LD, baseados nos coeficientes padrões D, D' [74] e o quadrado do coeficiente de correlação r^2 [53].

O software LDA fornece uma interface gráfica e utiliza o algoritmo descrito acima para analisar LDs. A entrada do programa é um arquivo-texto contendo dados genotípicos organizados por locus. O usuário pode definir parâmetros para a análise.

A.24 GOLD - Graphical Overview of Linkage Disequilibrium [1]

O processo tradicional de analise de correlação entre SNPs identifica regiões cromossomais prováveis de conter genes responsáveis por doenças. Porém, estes métodos são limitados pelo número de eventos de recombinações em indivíduos de mesma linhagem genética, e impraticáveis para análise de doenças complexas. Estima-se que o mapeamento de genes baseado em Linkage Disequilibrium (LD) trarão resultados mais completos e precisos sobre estas doenças [101].

A medida que haplotipos de ancestrais se propagam em uma população, a distância física tende a ser reduzida por eventos de recombinação, e eventos de recombinação entre marcadores prõximos um do outro são bastante raros. Por isso, espera-se que indivíduos

que herdam uma mutação relacionada a uma doença de um mesmo ancestral compartilhem a região do haplótipo onde a mutação ocorreu. Os marcadores destes haplótipos comum não são associados de forma aleatória, e são considerados como estando em desequilibrio de ligação (Linkage Disequilibrium).

Enquanto marcadores de um LD estão fortemente relacionados, o padrão de haplótipos muda bastante e depende de vários fatores como migração, seleção, mutação e outros. Obter informações sobre estes padrões é de grande importância para os mapeamento genético humano.

Tradicionalmente, LDs são descritos por medidas genéticas comuns, como D', D e Δ^2(também chamada de r^2), definidas por Weir [120]. O número de medidas cresce exponencialmente com o número de marcadores, dificultando a análise a partir de resultados tabelados. A ferramenta GOLD (Graphical Overview of Linkage Disequilibrium) oferece um modo gráfico de visualizar os dados. Estes são apresentados em um gráfico onde para cada par de marcadores m_i e m_j adiciona-se um ponto de coordenadas $(x, y) = (m_i, m_j)$, cuja cor representa a estatística associada ao LD.

O software é livre e foi escrito em Java, podendo ser executado em qualquer plataforma que possua JVM superior à versão 1.4

A.25 SNPHunter: a bioinformatic software for single nucleotide polymorphism data acquisition and management [118]

SNPs são uma ferramenta importante no processo de localização de genes suscetíveis de agirem em doenças genéticas complexas. A seleção e obtenção de um conjunto ótimo de SNPs de bases de dados públicas tem se tornado um dos maiores problemas no planejamento de estudos em larga escala de LDs.

Apesar de muitos pacotes computacionais disponíveis atualmente trabalharem com SNPs (busca, visualização, design de primers), nenhum deles oferece boas soluções para seleção de um conjunto apropriado de SNPs. A seleção deste conjunto é importante pois o objetivo dos geneticistas é maximizar a probabilidade de se detectar um locus relacionado a doenças através do melhor conjunto possível de SNPs correlacionados, dado um orçamento limitado.

Apesar da base de dados dbSNP [26] prover um conjunto abrangente de ferramentas para busca de SNPs, ainda são necessárias ferramentas para busca eficiente e fácil de SNPs desejados, para avaliação de suas anotações e para a exportação dos dados para formatos apropriados para análises posteriores. A ferramenta SNPHunter, descrita neste trabalho, oferece uma interface gráfica e funciona como um intermediário entre o usuário e a base

dbSNP. O pacote permite a extração e exportação de SNPs da base dbSNP, importação de dados gravados na máquina local e permite seleção de SNPs baseado em sua posição, função e heterozigocidade. O pacote computacional se baseia em um parser HTTP, que delega as buscas a ferramentas públicas disponíveis na internet como dbSNP, LocusLink, MapViewer e ACEView, todos do NCBI [85]. O algoritmo utilizado se baseia no método descrito em Thompson et al. [116].

O projeto foi desenvolvido em Visual Basic .NET e funciona na plataforma Windows.

A.26 Sequence polymorphism from EST data in sugarcane: a fine analysis of 6-phosphogluconate dehydrogenase genes [47]

O objetivo deste trabalho é demonstrar o uso de ESTs de cana-de-açucar obtidos da base de dados SUCEST para detectar SNPs. É o primeiro trabalho sobre análise de SNPs em plantas superiores. O projeto do Genoma da cana-de-açucar (SUCEST [111]) produziu aproximadamente 260.000 ESTs gerados à partir de 230.000 cDNAs obtidos de 37 bibliotecas distintas. Os genes selecionados para análise foram os relacionados ao 6-fosfogluconato dehidrogenase.

A estratégia utilizada para detecção de SNPs foi baseada na estratégia de Picoult-Newbert *et al.* [96]. Os ESTs possivelmente gerados à partir dos genes de interesse (*Pgd1* e *Pgd2*) foram obtidos através da ferramenta de alinhamento local `BlastN` [4] e agrupados com a ferramenta `Phrap` [46] com alta estringência de forma a montar clusters contendo apenas ESTs altamente idênticos. Os clusters gerados pelo pacote `Phrap` geraram sequências de consenso, mais fáceis de serem utilizadas para análises posteriores por serem mais longas e possuirem maior número de bases com boa qualidade. Os consensos foram comparados dois a dois com a ferramenta `Blast` de forma a grupa-los por similaridade: apenas era considerados válidos para análise de similaridade resultados obtidos em sequências de no mínimo 100bps, e eram considerados do mesmo grupo dois consensos com similaridade mínima de 98%. Por fim, consensos relacionados a um mesmo gene foram alinhados, montando um super-cluster, a partir do qual seriam procurados os SNPs.

Um polimorfismo foi considerado como sendo SNP se a variação menos freqüente aparecesse no mínimo duas vezes na seção do alinhamento com qualidade definida pelo pacote Phred [31, 30] superior ou igual a 20, e as bases adjacentes estivessem alinhadas numa janela de 10bp.

Com o procedimento descrito acima, sessenta e quatro ESTs foram identificados e divididos em dois conjuntos de 14 e 50 ESTs respectivamente. O alinhamento das seqüências do primeiro grupo permitiu a detecção de um único SNP e o alinhamento das seqüências

do segundo grupo permitiu a detecção de 39 SNP, incluindo 27 na região codificante do gene. Trinta e oito SNP foram bi-nucleotídicos e um único tri-nucleotídico.

Apêndice B

Glossário

Alelo Uma das possíveis formas de informação de um locus genético.

cDNA Fita simples de DNA sintetizada em laboratório a partir de uma fita de mRNA.

CDS Sigla para Coding Sequence, ou sequência codificadora. É a região do mRNA que se encontra entre o códon de início de síntese proteica (start-códon) e o códon de terminação de síntese proteica. É a região que será efetivamente traduzida em uma proteína.

centimorgan (cM) Unidade de distância genética. Corresponde à porcentagem de freqüência de recombinação genética entre dois loci adjacentes: 1 cM = 1% de probabilidade de um alelo ser separado de outro em um evento de recombinação (crossing-over). Em seres humanos, 1 cM corresponde a aproximadamente 1Mbp.

Clonagem posicional Técnica que consiste em isolar e sequenciar um trecho de um gene à partir de mapas físicos e de marcadores na região de interesse.

Códon Tripla de nucleotídeo usada como base para processo de tradução. Cada códon é traduzido em um aminoácido.

Contig Conjunto de clones contíguos sobrepostos que cobrem uma região cromossômica.

DNA Formado por duas fitas de ácido desoxiribonucléico (A, C, G ou T). As duas fitas são acopladas pelas bases orgânicas seguindo a seguinte regra: a base A sempre se acopla à base T e a base C sempre se acopla à base G. As duas fitas são ditas complementares, e a orientação $5' \rightarrow 3'$ das duas fitas é oposta uma em relação à outra.

Eucarioto Seres vivos com células eucarióticas, ou seja, com um núcleo celular rodeado por uma membrana (ou seja, núcleo individualizado, separado do citoplasma) e com várias organelas.

EST Expressed Sequence Tag: Sequência genética expressa gerada a partir de um mRNA.

Falso negativo Resultado de um teste indicando resposta negativa, quando na verdade a resposta deveria ser positiva.

Falso positivo Resultado de um teste indicando resposta positiva, quando na verdade a resposta deveria ser negativa.

Fenótipo Aspecto visível de uma determinada característica de um indivíduo.

Gene Cadeia de nucleotídeos de tamanho variável cuja função é servir de base para a construção de proteínas. Um gene pode conter regiões que efetivamente codificam uma proteína (chamadas de exons) e regiões que não codificam nada (chamadas de íntrons). Os genes são armazenados em fitas de RNA.

Genes Parálogos Genes que tem uma origem comum e aparecem no mesmo genoma.

Genes Ortólogos Genes que tem uma origem comum e aparecem em genomas diferentes.

Genoma O genoma é o conjunto total de informações codificadas no DNA de um organismo, incluindo regiões codificadoras e não codificadoras.

Genótipo Constituição genética de um organismo.

Haplótipo Um haplótipo é a constituição genética do cromossomo de um indivíduo. No caso de organismos diplóides, como os seres humanos, o haplótipo contém um membro do par de alelos. Um haplótipo pode se referir a apenas um locus ou ao genoma inteiro. Um haplótipo relacionado a um genoma contém um alelo de cada par de genes.

Um outro sentido pode ser dado ao termo, relacionado a SNPs: um haplótipo pode se referir a um conjunto de polimorfismos de base única estatisticamente associados. Neste contexto, quando se tem um haplótipo, a presença de um polimorfismo fornece informações sobre a presença dos outros polimorfimos do haplótipo.

Heredograma Representação gráfica da manifestação de um dado caráter (característica) em uma famíla de múltiplas gerações. Exibe os indivíduos da família, destacando as relações de parentesco e a manifestação ou não do carater estudado, como mostrado na Figura B.1.

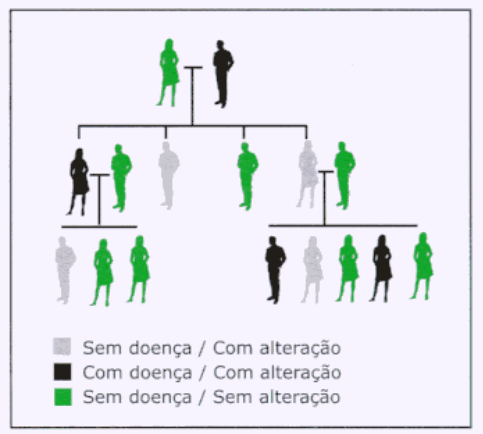

Figura B.1: Exemplo de heredograma: cada linha representa uma geração da família. Quadrados reperesentam indivíduos do sexo masculino, e círculos representam indivíduos do sexo feminino. As cores representam a manifestação de um dado caráter em cada indivíduo da família.

INDEL Insertion/Deletion. Alinhamento de uma base com um espaço, devido à inserção ou remoção de uma base em uma das seqüencias.

IUPAC International Union of Pure and Applied Chemistry [65].

Locus Posição física de um elemento dentro de um cromossomo.

Mapa genéticos Mapa que descreve a ordem de marcadores em um cromossomo e quantifica a distância entre eles em centimorgans (cM). Pode ser chamado mapas de ligação ou mapa meiótico [91, pg.159].

NCBI National Center for Biotechnology Information [85].

Nucleotídeo Unidade básica que forma uma fita de ácido nucléico, formada por uma base orgânica, um açúcar e fósforo.

mRNA RNA mensageiro gerado a partir dos genes em um processo chamado de transcrição, que basicamente gera uma fita simples de RNA a partir do DNA que compõe o gene, removendo as regiões de íntrons caso necessário (apenas organismos eucariotos, como os seres humanos, possuem íntrons). O mRNA é utilizado para a codificação de proteínas em um processo chamado de tradução, que utiliza o mRNA como fita de leitura.

Procarioto Organismos unicelulares sem a membrana que envolve o núcleo, a carioteca ou membrana nuclear, e sem presença de proteínas histônicas associadas ao

DNA, que por sua vez encontra-se disperso no citoplasma ou em forma de anéis (plasmídeos).

RNA Formado por uma fita de ácido ribonucléico (A, C, G e U, com A se acoplando a U e C se acoplando a G). Seu tamanho é bem menor do que o tamanho de uma molécula de DNA, e o RNA é mais instável que o DNA.

SNP Single Nucleotide Polymorphism, ou polimorfismo de base única. Polimorfismo em uma sequencia genética que afeta apenas uma base.

UTR Sigla para untranslated region, ou região não traduzida. É uma região pertencente ao mRNA que está fora do CDS, e que foi adicionadas ao mRNA durante o processo de transcrição. Existem dois tipos de UTRs:

5'-UTR Sequência de bases que vai do sítio de início de transcrição à base imediatamente anterior o códon de início de tradução. Um 5'-UTR pode conter regiões que regulam a eficiência do processo de tradução ou a estabilidade do mRNA.

3'-UTR Sequência de bases localizada na extremidade 3' do mRNA, após o códon de término de síntese proteica, não traduzida em proteína. Um 3'-UTR pode conter regiões que regulam a eficiência do processo de tradução, a estabilidade do mRNA ou sinais de poliadenilação.

Bibliografia

[1] G. R. Abecasis and W. O. C. Cookson. GOLD - Graphical Overview of Linkage Disequilibrium. *BioInformatics*, 16(2):182–183, February 2000.

[2] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, C. R. Merril H. Xiao, A. Wu, B. Olde, and R. F. Moreno. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 21:1651–1656, 1991.

[3] André A. M. Almeida, Miguel Galves, and Zanoni Dias. Um algoritmo para identificação de correlações múltiplas de polimorfismos. Technical Report IC-06-14, September 2006.

[4] S. F. Altschul, W. GIsh, E. W. Myers, W. Miller, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[5] S. F. Altschul, T. L Madden, A. A. Schaffer, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein databases search programs. *Nucleic Acid Research*, 25:3389–3402, 1997.

[6] K. G. Ardlie, L. Kruglyak, and M. Seielsta. Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 3:299–309, 2002.

[7] A. Bairoch and R. Appweiler. The SWISS-PROT protein sequence database and its supplement trembl in 2000. *Nucleic Acid Research*, 28:45–48, 2000.

[8] C. Baudet and Z. Dias. New EST TrimmingStrategy. In J.C. Setubal and S. Verjovski-Almeida, editors, *Lecture Notes on Bioninformatics*, volume 3594, pages 206–209. Springer-Verlag Berlin Heildelberg, July 2005. Brazilian Symposium on Bioinformatics (BSB 2005).

[9] Christian Baudet, Miguel Galves, and Zanoni Dias. Comparação de métodos para determinação de snps com medidas de confiabilidade. Technical Report IC-06-15, September 2006.

[10] Berkeley Drosophila Genome Project, July 2005. http://www.fruitfly.org/.

[11] A. J. Berno. A graph theoretic approach to the analysis of DNA sequencing data. *Genome Research*, 6:80–91, 1986.

[12] E. Birney and R. Durbin. Dynamite: a flexible code generating language for dynamic methods used in sequence comparison. *Porc. Fifth Int. Conf. Intelligent. Systems Mol. Biol.*, 5:56–64, 1997.

[13] D. Botstein and N. Risch. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics Supplement*, 33:228–237, March 2003.

[14] D. Brett, G. Lehmann, J. Hanke, S. Gross, J. Reich, and P. Bork. EST analysis online: WWW tools for detection of SNPs and alternative splice forms. *TIG*, 16(9):416–418, September 2000.

[15] A. J. Brookes. The essence of SNPs. *Gene*, 234:177–186, 1999.

[16] T. A. Brown. *Genomes*. John Wiley and Sons, Inc, 1999.

[17] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, C. R. Lane, E. P. Lim, N. Kalyanaraman, and J. Nemesh. Characterization of single-nucleotide polymorphisms in coding regions od human genes. *Nature Genetics*, 22:231–238, 1999.

[18] Cancer Genome Anatomy Project - Genetic Annotation Initiative. http://lpgws.nci.nih.gov.

[19] K-M. Chao, J. Zhang, J. Ostell, and W. Miller. A tool for aligning very similar DNA sequences. *Computer Applications in the Biosciences*, 13:75–80, 1997.

[20] D. Chasman and R. M. Adams. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: Structure based assessment of amino acid variation. *Journal of Molecular Biology*, 307:683–706, 2001.

[21] F. S. Collins, L. D. Brooks, and A. Chakravarti. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research*, 8:1229–1231, 1998.

[22] C. Connell, S. Fung, C. Heiner, J. Bridgham, V. Chakerian, E. Heron, B. Jones, S. Menchen, W. Mordan, and M. Raff. Automated DNA sequence analysis. *BioTechniques*, 5:342–348, 1987.

[23] The Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[24] D. N. Cooper, E. V. Ball, and M. Krawczak. The human gene mutation database. *Nucleic Acid Research*, 26:285–287, 1998.

[25] dbEST – The International Expressed Sequence Tags Database, September 2002. www.ncbi.nlm.nih.gov/dbEST.

[26] National Center for Biotecnology Information SNP Database, March 2006. http://www.ncbi.nlm.nih.gov/SNP.

[27] B. Devlin and N. A. Risch. A comparison of linkage disequilibrium measures for fine scaling mapping. *Genomics*, 29:311–322, 1995.

[28] K. Ding, K. Zhou, F. He, and Y. Shen. LDA - a Java-based linkage disequilibrium analyser. *BioInformatics*, 19(16):2147–2148, November 2003. Software available at http://www.chgb.org.cn/lda/lda.htm.

[29] M. K. Halushka et al. Patterns of single-nucleotide polymorphisms in candidate genes regulating blood-pressure homeostasis. *Nature Genetics*, 22:239–247, 1999.

[30] B. Ewing and P. Green. Base-Calling of Automated Sequencer Traces Using *Phred* II. Error Probabilities. *Genome Research*, 8:186–194, 1998.

[31] B. Ewing, L. Hillier, M. C. Wendl, and P. Green. Base-Calling of Automated Sequencer Traces Using *Phred* I. Acuracy Assesment. *Genome Research*, 8:175–185, 1998.

[32] L. Florea, G. Hartzell, Z. Zhang, G. Rubin, and W. Miller. A computer program for aligning cDNA sequence with genomic DNA sequence. *Genome Research*, 8:967–974, 1998.

[33] FORESTs: Eucalyptus Genome Sequencing Consortium, July 2004. http://forests.esalq.usp.br/.

[34] *Xylella* – Genoma Funcional, September 2002. www.lbm.fcav.unesp.br/fun.

[35] M. Galves and Z. Dias. Comparison of genomic DNA to cDNA alignment methods. In J.C. Setubal and S. Verjovski-Almeida, editors, *Lecture Notes on Bioninformatics*, volume 3594, pages 170–180. Springer-Verlag Berlin Heildelberg, July 2005. Brazilian Symposium on Bioinformatics (BSB 2005).

[36] M. Galves, J. A. A. Quitzau, and Z. Dias. New strategy to detect single nucleotide polymorphisms. *Genetics and Molecular Research*, 5(1):143–153, 2006.

[37] Agronomical & environmental genomes, July 2004. http://watson.fapesp.br/AEG/agro.htm.

[38] GeneSNPs, July 2004. http://www.genome.utah.edu/genesnps.

[39] M. C. Giddings, R. L Brumley Jr., M. Haker, and L. M Smith. An adaptative, object oriented strategy for base-calling in DNA sequence analysis. *Nucleic Acids Research*, 21:4530–4540, 1993.

[40] Globin Gene Server, July 2004. http://globin.cse.psu.edu/.

[41] J. Golden, E. Garcia, and C. Tibbets. Evolutionary optimization of a neural network-based signal processor for photometric data from an automated DNA sequencer. In *Evolutionary programming IV. Proceedings of the Fourth Annual Conference on Evolutionary Programming*, pages 579–601, 1995.

[42] J. B. Golden, D. Torgersen, and C. Tibbets. Pattern recognition for automated DNA sequencing: I. online signal conditioning and feature extraction for basecalling. In D. Searls L. Hunter and J. Shavlik, editors, *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, pages 136–144, Menlo Park, CA, 1993. AAAI Press.

[43] D. B. Goldstein, S. K. Tate, and S. M. Sisodiya. Pharmacogenetics goes genomics. *Nature Reviews*, 4:937–947, December 2003.

[44] M. E. Gorre, M. Mohammed, K. Ellwood, N. Hsu, R. Paquette, P. N. Rao, and C. L. Sawyers. Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science*, 293:876–880, August 2001.

[45] O. Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162:705–708, 1982.

[46] P. Green. Phrap documentation, September 2002. www.phrap.org.

[47] L. Grivet, J.C. Glaszmann, and P. Arruda. Sequence polymorphism from EST data in sugarcane: a fine analysis of 6-phosphogluconate dehydrogenase genes. *Genetics and Molecular Biology*, 24(1–4):161–167, 2001.

[48] L. Grivet, J.C. Glaszmann, M. Vincentz, F. da Silva, and P.Arruda. ESTs as a source for sequence polymorphism discovery in sugarcane: example of the Adh genes. *Theoretical Applied Genetics*, 106:190–197, 2003.

[49] N. Guex, A. Diemand, and M. C. Peitsch. Protein modelling for all. *Trends Biochem.*, 24:364–367, 1999.

[50] P. E. M. Guimarães and M. C. R. Costa. SNPs: Sutis difereças de um código. *BioTecnologia Ciência e Desenvolvimento*, 26:24–27, maio/junho 2002. In Portuguese.

[51] Human Cancer Genome Project - SNP, July 2004. http://bit.fmrp.usp.br/.

[52] Human Genome Variation Base, July 2004. http://hgvbase.cgb.ki.se.

[53] W. G. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38:226–231, 1968.

[54] W. G. Hill and B. S. Weir. Maximum-likelihood estimation of gene location by linkage disequilibrium. *American Journal of the Human Genetics*, 54:705–714, 1994.

[55] D. S. Hirschberg. A linear space algorithm for computing longest common subsequences. *Communications of the ACM*, 175:341–343, 1975.

[56] HIV Databases, March 2006. http://http://www.hiv.lanl.gov/.

[57] R. Horton, L. Wilming, V. Rand1, R. C. Lovering, E. A. Bruford, V. K. Khodiyar, M. J. Lush, S. Povey, C. Conover Talbot, M. W. Wright, H. M. Wain, J. Trowsdale, A. Ziegler, and S. Beck. Gene map of the extended human MHC. *Nature Reviews Genetics*, 5:889–899, December 2004.

[58] Human Organized Whole genome Database, July 2004. http://gdb.jst.go.jp/HOWDY.

[59] W. M. Howell, M. Jobs, U. Gyllensten, and A. J. Brookes. Dynamic allele-specific hybridization. *Nature Biotechnology*, 17:87–88, January 1999.

[60] X. Huang. On global sequence alignment. *Computer Applications in the Bioscience*, 10:227–235, 1994.

[61] X. Huang and A. Madan. CAP3: a DNA sequence assembly program. *Genome Research*, 9:868–877, 1999.

[62] R. R. Hudson. Linkage disequilibrium and recombination. In *Handbook of Statistical Genetics*, pages 309–324. Wiley, 2001.

[63] Instituto Fleury, September 2004. http://www.institutofleury.org.br/.

[64] P. Iughetti, O. Suzuki, P. H. C. Godoi, V. A. F. Alves, A. L. Sertie, T. Zorick, F. Soares, A. Camargo, E. S. Moreira, C. Loreto, C. A. Moreira-Filho, A. Simpson, G. Oliva, and M. R. Passos-Bueno. A polymorphism in endostatin, an angiogenesis inhibitor, predisposes for the development of prostatic adenocarcinoma. *Cancer Research*, 61:7375–7378, October 2001.

[65] International Union of Pure and Applied Chemistry, September 2005. http://www.chem.qmw.ac.uk/iupac/.

[66] Japanese Single Nucleotide Polymorphism, July 2004. http://snp.ims.u-tokyo.ac.jp.

[67] N. Kaplan and B. S. Weir. Expected behavior of conditional linkage disequilibrium. *American Journal of the Human Genetics*, 51:333–343, 1992.

[68] R. M. Karp. Reducibility among combinatorial problems. *Complexity of Computer Computations*, pages 85–103, 1972.

[69] B. Klamt. Frasier syndrome is caused by defective alternative splicing of WT1 leading to an altered ration of WT1+/-KTS splic isoforms. *Human Molecular Genetics*, 4:709–714, 1998.

[70] L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, 22:139–144, 1999.

[71] P. Y. Kwok, C. Carlson ant T. D. Yager, W. Ankener, and D. A. Nickerson. Comparative analysis of human dna variations by fluorescence-based sequencing of pcr products. *Genomics*, 23:138–144, 1994.

[72] C. B. Lawrence and V. V. Solovyev. Assigment of position-specific error probability to primary DNA sequence. *Nucleic Acids Research*, 20:2471–2483, 1994.

[73] Leelab Databases, July 2004. http://www.bioinformatics.ucla.edu/snp.

[74] R. C. Lewontin. The interaction of selection and linkage. I. general considerations; heterotic models. *Genetics*, 49:49–67, 1964.

[75] S. Marsh, P. Kwok, and H. L. McLeod. SNP databases and pharmacogenetics: A great start, but a long way to go. *Human Mutation*, 20:174–179, 2002.

[76] G. Marth, R. Yeh, M. Minton, R. Donaldson, Qun Li, R. Davenport S. Duan, R. D. Miller, and R. Kwok. Single-nucleotide polymorphisms in the public domain: how useful are they? *Nature Genetics*, 27:371–372, April 2001. brief communication.

[77] G. T. Marth, I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu, H. Zakeri, N. O Stitziel, L. Hillier, P-Y. Kwok, and W. R. Gish. A general approach to single-nucleotide polymorphism discovery. *Nature Genetics*, 23:452–456, December 1999.

[78] Max Delbrueck Center for Molecular Medicine. http://mahe.bioinf.mdc-berlin.de/home.html.

[79] Mendel's Paper in English, July 2003. http://www.mendelweb.org/Mendel.html.

[80] R. Mott. EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *Computer Applications in the Biosciences*, 13:477–478, 1997.

[81] A. Motulsky. Drug reactions, enzymes, and biochemical genetics. *Journal of the American Medical Association*, 165:835–837, 1957.

[82] A Muto and S Osawa. The guanine and cytosine content of genomic dna and bacterial evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 87:166–169, 1987.

[83] E. W. Myers and W. Miller. A file comparison program. *Software - Practice and Experience*, 15:1025–1040, 1985.

[84] E. W. Myers and W. Miller. Optimal alignment in linear space. *Computer Applications in the Biosciences*, 4(1):11–17, 1988.

[85] National Center for Biotecnology Information, July 2006. http://www.ncbi.nlm.nih.gov/.

[86] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[87] P. C. Ng and S. Henikoff. Predicting deleterious amino acid substitutions. *Genome Research*, 11:863–874, 2001.

[88] P. C. Ng and S. Henikoff. Accounting for human polimorphism predicted to affect protein function. *Genome Research*, 12:436–446, 2002.

[89] D. A. Nickerson, V. O. Tobe, and S. L. Taylor. Polyphred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Research*, 25(14):2745–2751, 1997. Disponível em http://droog.mbt.washington.edu/PolyPhred.html.

[90] J. M. Ostell. *The NCBI Software Tools. In Nucleic Acid and Protein Analysis: A Practical Approach*. IRL Press, Oxford, 1996. pp 31–43.

[91] J. J. Pasternak. *Genética Molecular Humana - Mecanismo das Doenças Hereditárias*. Editora Manole, 1. ed edition, 2002. Título original em inglês: An introduction to human molecular genetics: mechanisms of inherited diseases.

[92] PE Applied Biosystems, Foster City, CA. *ABI PRISM, DNA sequencing analysis software, user's manual*, 1996.

[93] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8):2444–2448, 1988. ftp://ftp.virginia.edu/pub/fasta/.

[94] The Pharmacogenetics and Pharmacogenomics Knowledge Base, July 2004. http://pharmgkb.org.

[95] R. S. Phelps, R. B. Chadwick, M. P. Conrad, M. N Kronick, and A. Kamb. *BioTechniques*, 19:984–989, 1995.

[96] L. Picoult-Newberg, T.E. Ideker, M.G.Pohl, S.L. Taylor, M.A. Donaldson, D.A. Nickerson, and M. Boyce-Jacino. Mining SNPs from EST databases. *Genome Research*, 9:167–174, 1999.

[97] J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *American Journal of the Human Genetics*, 69:1–14, 2001.

[98] J. M. Prober, G. L Trainor, R. J. Dam, F. W. Hobbs, C. W. Robertson, R. J. Zagursky, A. J. Cocuzza, M. A. Jensen, and K. Baumeister. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Sience*, 238:326–341, 1987.

[99] A. Rambaut and N. C. Grassly. Seq-gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, 13:235–238, 1997.

[100] M. G. Reese, F. H. Eeckman, D. Kulp, and D. Haussler. Improved splice site detection in genie. *Journal of Computational Biology*, 4:311–323, 1997.

[101] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Nature*, 273:1516–1517, 1996.

[102] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggil, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P.Y. Kwok, E. R. Mardin, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgings, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, and D. Altshuler. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphism. *Nature*, 409:928–933, February 2001.

[103] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74:336–341, 1977.

[104] *Schistosoma mansoni* EST Genome Project, September 2002. http://verjo18.iq.usp.br/schisto.

[105] G. D. Schuler. Pieces of the puzzle: Expressed sequence tags and the catalogue of human genes. *Journal of Molecular Medicine*, 75:694–698, 1997.

[106] J. C. Setubal and J. Meidanis. *Introduction to Computional Molecular Biology*. PWS Publishing Company, 1997.

[107] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. H Kent, and L. E. Hood. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321:674–679, 1986.

[108] T. E. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[109] R. Sorek and H. M. Safer. A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Research*, 31(3):1067–1074, 2003.

[110] C. A. Stewart, R. Horton, R. J. N. Allcock, J. L. Ashurst, A. M. Atrazhev, P. Coggill, I. Dunham, S. Forbes, K. Halls, J. M.M. Howson, S. J. Humphray, S. Hunt, A. J. Mungall, K. Osoegawa, S. Palmer, A. N. Roberts, J. Rogers, S. Sims, Y. Wang, L. Wilming, J. F. Elliott, P. J. de Jong, S. Sawcer, J. A. Todd, J. Trowsdale, and Stephan Beck. Complete MHC Haplotype Sequencing for Common Disease Gene Mapping. *Genome Research*, 14:1176–1187, 2004.

[111] The Sugar Cane EST Genome Project, September 2002. http://sucest.lbi.ic.unicamp.br.

[112] S. Sunyaev. Prediction of nonsynonymous single nucleotide polymorphims in human disease-associated. *Journal of Molecular Medicine*, 77:754–760, 1999.

[113] S. Sunyaev, V. Ramensky, I. Koch, W. Lathe III, A. S. Kondrashov, and P. Bork. Prediction of deleterious human alleles. *Human Molecular Genetics*, 10:591–597, 2001.

[114] G. P. Telles, M. D.V. Braga, Z. Dias, L. T. Li, J. A. A. Quitzau, F. R. da Silva, and J. Meidanis. Bioinformatics of the sugarcane EST project. *Genetics and Molecular Biology*, 24(1-4):9–15, December 2001.

[115] The Human Cancer Genome Project, September 2002. http://www.ludwig.org.br/ORESTES.

[116] D. Thompson, D. Stram, D. Goldgar, and J. S. Witte. Haplotype tagging single nucleotide polymorphisms and association studies, 2003.

[117] A. L. Vettore, F. R. da Silva, E. L. Kemper, G. M. Souza, A. M. da Silva, M. I. T. Ferro, F. Henrique-Silva, A. Giglioti, M. V. F. Lemos, L. L. Coutinho, M. P. Nobrega, H. Carrer, S. C. Fran, M. Bacci Jr., M. H. S. Goldman, S. L. Gomes, L. R. Nunes, L. E. A. Camargo, W. J. Siqueira, M. A. V. Sluys, O. H. Thiemann, E. E. Kuramae, R. V. Santelli, C. L. Marino, M. L. P. N. Targon, J. A. Ferro, H. C. S. Silveira, D. C. Marini, E. G. M. Lemos, C. B. Monteiro-Vitorello, J. H. M. Tambor, D. M. Carraro, P. G. Roberto, V. G. Martins, G. H. Goldman, R. C. de Oliveira, D. Truffi, C. A. Colombo, M. Rossi, P. G. de Araujo, S. A. Sculaccio, A. Angella, M. M. A. Lima, V. E. de Rosa Jr., F. Siviero, V. E. Coscrato, M. A. Machado, L. Grivet, S. M. Z. Di Mauro, F. G. Nobrega, C. F.M.Menck, M. D. V. Braga, G. P. Telles, F. A. A. Cara, G. Pedrosa, J. Meidanis, and P. Arruda. Analysis and functional annotation of an expressed sequence tag collection for the tropical crop sugarcane. *Genome Research*, 13:2725–2735, 2003. Submitted: 12/May/2003. Acepted: 08/September/2003.

[118] L. Wang, S. Liu, T. Liu, and X. Xun. SNPHunter: a bioinformatic software for single nucleotide polymorphism data acquisition and management. *BMC Bionformatics*, 6:60–66, 2005.

[119] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[120] B. S. Weir. *Genetic Data Analysis II.* Sinauer Associates, 2nd edition, April 1996. ISBN 0878939024.

[121] D. B. West. *Introduction to Graph Theory*. Prentice Hall, 1996.

[122] S. J. Wheelan, D. M. Church, and J. M. Ostell. Spidey: A Tool for mRNA-to-Genomic Alignments. *Genome Research*, 11:1952–1957, 2001. Disponível em http://www.ncbi.nlm.nih.gov/spidey.