Uma abordagem computacional para a determinação de polimorfismo de base única

Miguel Galves

Resumo

A publicação da sequência do genoma humano gerou um grande entusiasmo em grupos de pesquisa e empresas farmacêuticas, que visam utilizar estas informações para o estudo e desenvolvimento de tratamentos individualizados. Estima-se que muitas doenças tenham uma base genética, gerada por mutações no genoma de certos indivíduos. Assim, o estudo dos polimorfismos, e em particular dos SNPs (polimorfimos de base única, que representam 90% dos polimorfismos presentes nos seres humanos) é de grande importância. Este trabalho propõe um estudo das etapas necessárias para a análise de SNPs, visando obter uma metodologia capaz de efetuar um estudo global destes polimorfismos.

1 Introdução

Este documento tem o objetivo de apresentar um plano a ser executado para o desenvolvimento do trabalho de mestrado. Nas seções 2 e 3, apresentaremos os conceitos básicos de genética e do processo de sequenciamento necessários para o entendimento do problema a ser resolvido. Na seção 5, definiremos o conceito de SNP e qual o seu interesse para a pesquisa atualmente, e nas seções 6, 7 e 8 descreveremos os problemas a serem estudados neste trabalho. Finalmente, as seções 9, 10 e 11 tratarão da proposta, do projeto PIPE no qual este trabalho está inserido e do cronograma das atividades a serem executadas.

2 Conceitos básicos de genética

As informações necessárias para o desenvolvimento dos seres vivos estão codificadas em cadeias de nucleotídeos. O conjunto completo de sequências é chamado de genoma.

2.1 DNA e RNA

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 para leitura das bases é $5' \rightarrow 3'$.

O código genético pode ser armazenado tanto na forma de **DNA** (ácido desoxiribonucléico) quanto na forma de **RNA** (ácido ribonucléico), exemplificados na Figura 1.

O DNA é formado por duas fitas de nucleotídeos, formadas por uma sequência de A, C, G e T e onde a pentose dos nucleotídeos é uma desoxiribose. As duas fitas são acopladas pelas bases

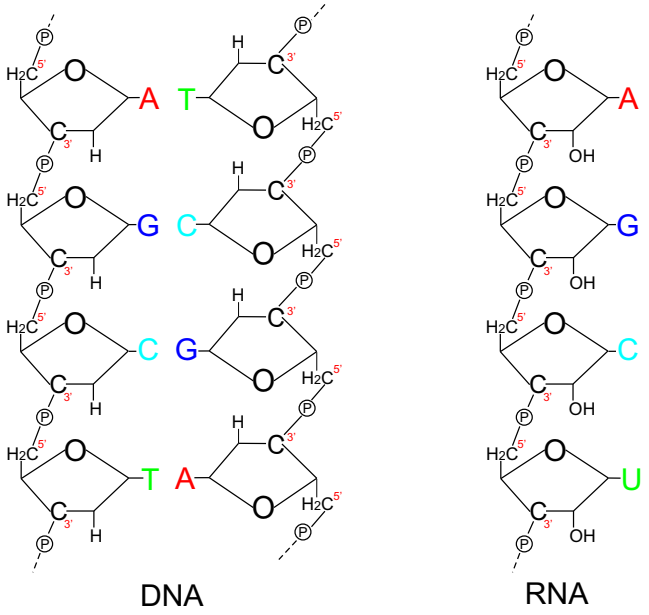

Figura 1: Fragmento de DNA e de RNA.

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 sequência 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 Expressão gênica

Um gene é um fragmento de DNA que contém informação genética codificada. Para que esta informação seja ativada, é necessário que ocorra 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.

2.3 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 contém 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).

Nesta seção, vamos nos interessar apenas pela transcrição de DNA em mRNA. Este processo 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 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 sequência 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 **exons**, e regiões não codificadoras, chamadas de **íntrons**. Os íntrons são removidos em um processo chamado de **splicing**, o qual cumpre uma função de grande importancia no processo de diversidade genética, uma vez que nem todos os exons são preservados no processo. Assim, um mesmo transcrito primário pode gerar vários mRNAs diferentes. Após este processo obtém-se um transcrito maduro.

2.4 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 **codons**. Existem $4^3 = 64$ codons: cada codon pode ser traduzido em um aminoácido (e um aminoácido pode ser originado por vários codons, uma vez que existem 20 aminoácidos) ou pode ter um significado especial, sendo utilizado como marcador de inicio e fim de uma região de tradução (respectivamente, **start-codon** e **stop-codon**). Existem várias tabelas de conversão de **codons** em aminoácidos, dependendo do organismo considerado. A Tabela 1 mostra o *Standard Genetic Code* [42], utilizado por muitos seres vivos, inclusive os seres humanos.

Aminoácido	**Cód. 3**	**Cód. 1**	**codon**
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 (Start-Codon)
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 1: Standard Genetic Code: Tabela de Conversão de Aminoácidos Standard utilizada por muitos seres vivos.

Além do mRNA, existem dois elementos envolvidos no processo de tradução: o ribossomo e o tRNA. O **ribossomo** é uma organela responsável por percorrer o mRNA, iniciando no start-codon e passando por todos os codons da fita. Para cada codon, 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-codon**. O anti-codon (codon formado pelos nucleotídeos complementares) se acopla ao codon do mRNA, e o aminoácido é liberado, se juntando à cadeia de aminoácidos.

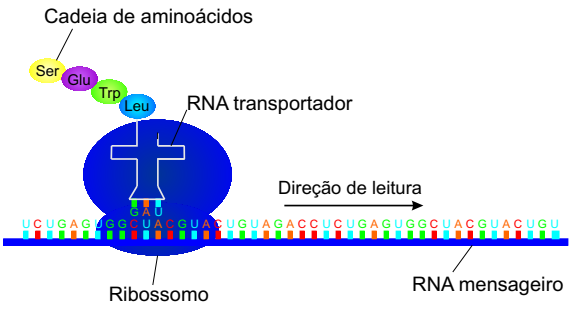

Figura 2: Mecanismo de tradução de mRNA em cadeia de aminoácido.

3 Sequenciamento genômico

O sequenciamento 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.

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 sequencimento é a **fragmentação**, que consiste em quebrar a cadeia de nucleotídeos a ser determinada em pequenos fragmentos sequenciáveis. 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, ou Polymerase Chain Reaction).

3.2 Método da terminação de cadeia

Uma vez obtidos os fragmentos de tamanho adequado, pode-se iniciar o processo de sequenciamento das bases de cada fragmento. Para isso, o método mais utilizado é o da terminação de cadeia [49]. O princípio básico deste método é gerar várias cópias de cadeias de nucleotídeos diferindo em tamanho por apenas uma base, que podem ser separadas e ordenadas por tamanho em um processo chamado de eletroforese. Utiliza-se marcadores para poder se reconhecer qual a última base de cada fragmento, possibilitando assim se obter a sequência de nucleotídeos do fragmento de interesse.

Os fragmentos de fita simples de DNA obtidos anteriormente são utilizados como moldes para uma reação de síntese de DNA complementar. Para que esta reação ocorra, acopla-se um oligonucleotídeo chamado 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 utilizado 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 possue 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 desoxiribonucleotídeos e dideoxinucleotídeos, fazendo com que algumas sequências fiquem mais longas do que outras. Cada tipo de dideoxinucleotídeo é acoplado a um marcador químico, utilizado na próxima fase do sequenciamento. A Figura 3 mostra o funcionamento deste processo.

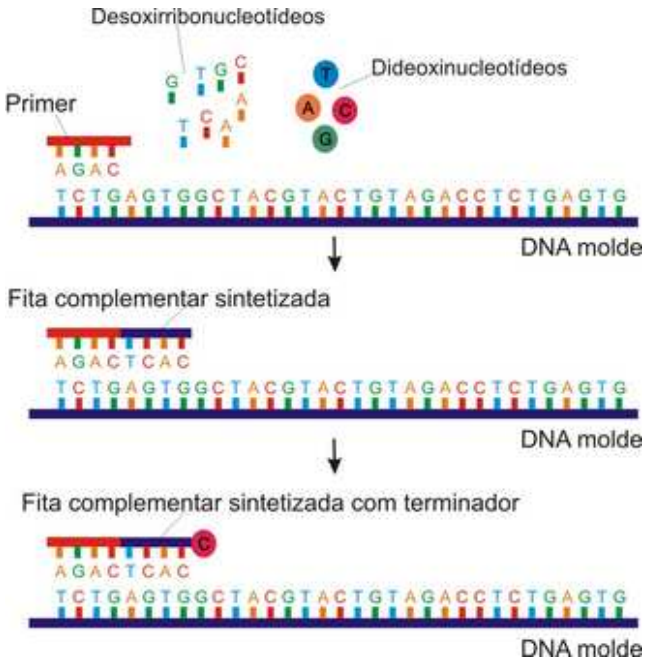

Figura 3: Método da terminação da cadeia.

As sequências sintetizadas com terminadores obtidas são separadas por tamanho em um processo chamado de eletroforese: neste processo, as cadeias são colocadas sobre uma placa contendo um gel, sobre o qual se aplica uma corrente elétrica fazendo com que os fragmentos migrem do polo negativo onde se encontram inicialmente para o polo positivo. Neste processo, os menores fragmentos tendem a migrar mais rapidamente, e ao final do procedimento, estarão mais perto do polo positivo do que os maiores fragmentos, sendo assim possível determinar seus tamanhos.

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 4a). Atualmente, cada marcador emite uma luz diferente, permitindo que se analise os quatro sinais em uma mesma canaleta (Figura 4b).

3.3 Leitura das bases

A partir da placa obtida, efetua-se o processo de **leitura das bases**.

A leitura se faz a partir de um cromatograma (Figura 4c), obtido a partir da placa, onde 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.

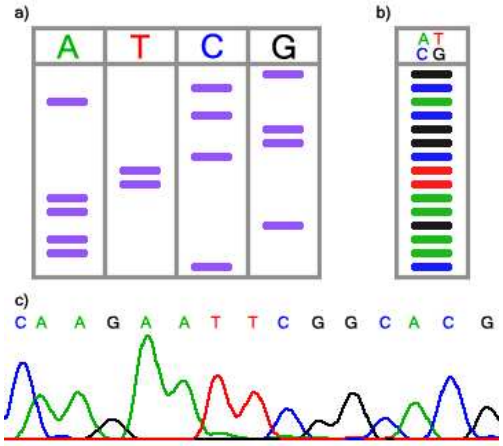

Figura 4: Leitura de bases.

Em um cromatograma ideal, os picos ficam espaçados de forma regular e não se sobrepõem, de forma 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, a problemas na eletroforese e análise da imagem obtida. Neste contexto, o objetivo do processo de *base-calling* é obter a melhor sequência possível.

Um dos pacotes computacionais mais antigos a executarem este processamento fazia parte do sistema da primeira máquina sequenciadora da Applied Biosystems (ABI) [11]. Este pacote obtém resultados de alta qualidade, sendo considerado como ponto de referência para análise de outros métodos. Porém, seu algoritmo não foi divulgado.

O pacote computacional *phred* [19, 18] determina uma sequência de nucleotídeos a partir de um cromatograma, e para cada base determinada calcula uma probabilidade de erro (ou qualidade). Após testes comparativos, observou-se que os resultados deste programa são tão precisos ou mais quanto os resultados do pacote ABI.

3.4 Montagem

Nos projetos de sequenciamento de genomas completos, deve se reconstruir a sequência completa original: este processo é chamado de **montagem**, e consiste em alinhar os diversos fragmentos, procurando regiões de sobreposição entre fragmentos, e utilizando algoritmos estatísticos para determinar fragmentos consecutivos, colocando-os em sequência e obtendo o DNA original.

4 Projetos de sequenciamento

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

4.1 Projetos genomas

Os projetos genomas tem como objetivo a obtenção da sequência genética completa de um organismo vivo. Dentre os projetos terminados ou em andamento, com certeza o mais importante tem sido o Projeto do Genoma Humano, iniciado em 1988 e contando com a participação de laboratórios em todo o mundo [12, 48, 2, 4].

Porém, outros projetos de organismos menores tem sido muito úteis para o aperfeiçoamento das técnicas e tecnologias utilizadas para sequenciamento, além de apresentarem resultados práticos em várias áreas como medicina e agronomia. O primeiro grande projeto de sequenciamento do genoma completo de um organismo foi o projeto da *Drosophila* [5]. Na página Entrez Genome mantida pelo NCBI [42] é possível encontrar genomas completos, além da lista de todos os projetos de seqüenciamento que estão em andamento.

No Brasil, a FAPESP organizou em 1997 a rede **ONSA** (Organization for Nucleotide Sequencing and Analysis), 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* [53], primeiro organismo causador de doenças de uma planta a ter seu genoma completamente sequenciado. Outros projetos desenvolvidos pela rede ONSA foram os projetos *Xanthomonas citri* e *Xanthomonas campestri*.

4.2 Projetos EST

Os ESTs (Expressed Sequence Tags) [2] são sequências genômicas expressas por um organismo, capturadas após o processo de transcrição e portanto não contendo sequências provenientes de regiões entre genes ou de íntrons. Originalmente utilizados para identificação de genes, descobriu-se posteriormente que eram extremamente úteis como ferramenta de mapeamento de genoma [51].

Os ESTs são obtidos a partir do sequenciamento de cDNA, que é uma fita de DNA produzida a partir do complemento do mRNA com a utilização da enzima transcriptase reversa. O seu processo de obtenção tem várias vantagens, permitindo que o processo de sequencimento possa ser executado de forma muito rápida.

Os ESTs sequenciados podem ser obtidos em bases de dados públicas, como a dbEST [14]. No Brasil, vários projetos de EST foram desenvolvidos ou estão em desenvolvimento. Entre eles, pode-se citar:

- Genoma Cana-de-Açúcar (SUCEST) [56, 58, 55].
- Projeto Genoma Humano do Câncer [57].
- Projeto FORESTs [21].
- Projeto *Schistossoma mansoni* [50].
- Projeto Genomas Agronômicos e Ambientais [22].

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.

5.1 Farmacogenética

Este interesse da medicina pela genética nasceu na década de 1950. Arno Motulsky foi o primeiro pesquisador a articular pesquisas nesta área de interface entre as duas disciplinas em 1957 [39]. Observando a resposta de indivíduos a certas drogas, ele argumentou que elementos genéticos poderiam ser subjacentes as 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 da sequência do genoma humano gerou um grande entusiasmo em grupos de pesquisa e empresas farmacêuticas, que visam utilizar estas informações para o estudo e desenvolvimento de tratamentos individualizados, baseados no fato que a grande maioria dos genes apresentam uma grande frequência de variações alélicas, conhecidos como polimorfismos, e que estas variações podem ser a chave para a predisposição de certos indivíduos a certas doenças [23, 36].

Dentre os polimorfismos, os SNPs tem tido um espaço de destaque, por representarem cerca de 90% dos polimorfismos encontrados no genoma humano [10].

5.2 SNPs

Um polimorfismo em uma seqûencia genética é a existencia de uma 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 índivíduos normais [6].

A priori, os SNPs poderiam ser polimorfismos 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 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.

Um SNP pode ser sinônimo ou não: no primeiro caso, o aminoácido codificado pelo codon contendo SNP é o mesmo que aquele codificado pelo codon sem SNP (mas o polimorfismo pode afetar a estabilidade do mRNA codificador); no segundo caso (chamado de nsSNP), o codon modificado gera um aminoácido diferente, podendo modificar a estrutura e função da proteína codificada.

5.3 SNP e seu interesse para a pesquisa farmacogenética

Um dos maiores interesses da pesquisa sobre o genoma humano é determinar se um SNP não sinônimo, 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 [13].

Existe atualmente um grande esforço para se mapear 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 [48]. 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 [27], 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 [32, 24].

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 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 responsável pela aparição de uma doença.

Recentemente, geneticistas voltaram seus esforços em mapear doenças mais comuns, cuja base genética é mais complexa e que afetam grandes fatias da população, onde os métodos tradicionais tem se mostrado menos eficientes [4]. 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.

6 Detecção e estudo de SNPs

Existem basicamente dois métodos para se detectar SNPs. O primeiro utiliza procedimentos químicos, e o segundo é baseado em comparação de sequências genômicas com o auxílio de ferramentas computacionais.

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 [7, pg. 18]. Fragmentos de restrição são produzidos quando uma molécula de DNA é tratada por uma enzima de restrição, que corta a molécula em sequências pré-definidas (por exemplo, a enzima BamHI corta a molécula sempre que encontrar a sequência GGATCC). 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 sequência 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 sitio 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 aqueles 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, impossibilitando o estudo. Este método ainda é uma herança dos tempos em que seqüenciar era difícil e caro.

6.2 Utilização de ferramentas computacionais

Uma nova maneira de estudar SNPs consiste em utilizar seqüenciamento de DNA aliado a ferramentas computacionais. Escolhe-se uma região genômica de interesse e seqüencia-se esta região de vários indivíduos. As sequências obtidas são alinhadas, utilizando algoritmos de alinhamento [29, 25, 40, 41, 38, 30, 9, 52, 20, 60]. 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.

6.3 Pacotes computacionais existentes para detecção de SNPs

O método de determinação de bases e cálculo de suas respectivas qualidades é a base de dois métodos para detecção de SNPs utilizados por dois pacotes computacionais: polyphred e polybayes.

6.3.1 polyphred

O pacote *polyphred* [44] utiliza os resultados obtidos pelos programas *phred* [19] e *phrap* [26], que monta sequências de consenso, para detectar SNPs.

O seu algoritmo se baseia em duas características observadas em cromatogramas contendo sequências com SNPs: uma significante redução ($\sim 50\%$) no tamanho do pico normalizado observado no cromatograma e a presença de um segundo pico menor que o principal na posição em questão [34, 45].

Assim, para cada posição de uma sequência 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.

6.3.2 polybayes: detecção de SNPs por análise bayesiana

O programa *polybayes* [37] utiliza um algoritmo de inferência Bayesiana para calcular a probabilidade de um dado alelo ser polimórfico.

O programa utiliza o valor 0.003 (um locus polimórfico a cada 333 bp) como a probabilidade total a priori de que um locus é polimórfico [8, 17]. Este valor é distribuído entre as bases para criar uma probabilidade a priori para cada permutação. Um valor a priori de $(1 - 0.003)/4$ é atribuído a cada uma das quatro permutações não polimórficas.

A probabilidade Bayesiana a posteriori de uma permutação em um nucleotídeo em particular é calculada considerando 4^N permutações diferentes como conjunto de modelos conflitantes.

A probabilidade a posteriori Bayesiana de um 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 um valor de limiar.

7 Alinhamento de DNA com cDNA

Define-se um alinhamento entre duas sequências como sendo uma operação de inserção de espaços nas duas sequências de forma a que elas tenham o mesmo tamanho, e que se possa sobrepô-las permitindo a comparação das bases [52, pg. 49].

Por exemplo, dadas as sequências ACGTTTG e ACGTTTTG, podemos facilmente observar que inserindo um espaço entre o T e o G da primeira sequência, teremos duas sequências praticamente idênticas:

```
ACGTTTTG
ACGTTT-G
```

Dado um alinhamento como definido acima, podemos criar uma pontuação para avaliar a qualidade do resultado obtido. O objetivo dos algoritmos de alinhamento existentes é obter um alinhamento ótimo, ou seja, que possua a maior pontuação possível

7.1 Esquemas de pontuação de alinhamento

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 $gap = -2$, $match = 1$ e $mismatch = -1$ teremos a seguinte pontuação para o alinhamento acima: $7 \times 1 - 2 = 5$.

Este sistema simples de pontuação não discrimina a criação de buracos separados ou contíguos. Assim, as sequências ACCG e CG poderiam produzir os seguintes resultados

```
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 [52]. 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 parâmetro *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 (*open gap*) e h o custo de se estender um buraco aberto (*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.

7.2 Tipos de alinhamento

Existem basicamente três tipos de alinhamento: global, semi-global e local, que serão brevemente descritos a seguir. Nos exemplos utilizados, consideraremos um esquema de pontuação simples com $gap = -2$, $match = 1$ e $mismatch = -1$.

7.2.1 Alinhamento global

O objetivo do alinhamento global é gerar o melhor alinhamento possível entre duas sequências. Os espaços podem ser inseridos em qualquer posição das sequências, de forma a se obter a pontuação ótima. Por exemplo, as sequências ACCG e CG poderiam produzir o alinhamento

```
ACCG
-C-G
```

com pontuação -2.

7.2.2 Alinhamento semi-global

A estratégia semi-global tem como objetivo alinhar sequências incompletas. Para isso, tenta obter o melhor alinhamento entre um prefixo de uma sequência com o sufixo da outra agrupando o maior número de espaços no início e no final do alinhamento, não penalizando a criação destes.

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

```
ACTGACCTCGGG
ACCGTCGGGCGG
```

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

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

com pontuação 6.

7.2.3 Alinhamento local

O alinhamento local tem como objetivo encontrar duas subsequências (uma em cada sequência original) que produzem o alinhamento com maior pontuação possível.

Por exemplo, um alinhamento local das sequências ACCATCTTGC e TCCCGTGTAAAA produziria o resultado

```
CC
CC
```

com pontuação 2.

7.3 Alinhamento de cDNA com DNA genômico

A identificação de genes em sequências de DNA 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 sequenciados tem sido um fator importante na adoção desta estratégia.

Os ESTs são uma das chaves 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 [12], 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.

Ferramentas padrão para alinhamento de sequências genéricas não são ideais para este 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.

A seguir descreveremos as estratégias utilizadas por três pacotes computacionais que alinham cDNA com DNA genômico.

7.3.1 est_genome

O programa *est_genome* [38] foi desenvolvido para alinhar pedaços de sequências (mRNA, EST ou cDNA) com sequências genômicas, permitindo a existência de grandes íntrons, reconhecendo sítios de splicing e utilizando memória limitada.

O algoritmo utilizado tem os seguintes passos:

- Faz uma primeira passada utilizando o algoritmo Smith-Waterman [54], que produz alinhamentos locais, para encontrar o início e fim dos segmentos de pontuação máxima. Subsequências correspondentes a estes segmentos são extraídas.
- 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 [43], que produz alinhamentos globais.
- 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 seja menor do que o limiar. As sequências divididas são alinhadas separadamente e intercaladas.
- Efetua-se uma busca dos segmentos na sequência genômica, de forma direta e reversa, levando-se em conta uma direção de splicing direta (ou seja, com consenso GT/AG). Depois, o alinhamento é feito levando-se em conta splicing com direção reversa (consenso CT/AC).

7.3.2 sim4

O programa *sim4* [20] tem como objetivo obter o alinhamento de ESTs com DNA 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, High Similarity Pairs).
- Seleção de um conjunto de HSPs que poderiam representar um gene.
- Encontrar os limites dos exons.
- Determinar o alinhamento de cada exon, utilizando o método de Chao [9].

7.3.3 Spidey

A ferramenta *Spidey* [60] produz alinhamentos de mRNAs com genomas, utilizando o toolkit do NCBI [42]. 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.

O algoritmo utilizado tem os seguintes passos:

- Cria janelas genômicas no DNA. Para isso, compara cada mRNA com a sequência genômica utilizando BLAST com alto grau de semelhança (*e-value* $= 10^{-6}$), e define regiões do DNA que são candidatas a exons.
- Tenta cobrir todo o mRNA com as janelas genômicas criadas. Para isso, compara cada mRNA com as janelas criadas no passo anterior, utilizando um BLAST com grau de semelhança menor que o inicial (*e-value* $= 10^{-3}$). Utiliza um algoritmo guloso para gerar um subconjunto de alinhamentos com alta similaridade e sem sobreposições.

- Caso existam regiões do mRNA não cobertas após o segundo passo, procura regiões no DNA compatíveis utilizando o BLAST com baixíssimo grau de similaridade (*e-value* = 1), e caso ainda sobre algum buraco, utiliza-se o DotView [3].

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.

7.3.4 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 [20] comparou o pacote *sim4* com o pacote *est_genome*, efetuando um teste de alinhamento de 184 ESTs com a sequência original: *est_genome* demorou 20 segundos por sequência e acertou 143 alinhamentos, e *sim4* demorou 0.06s por sequência, acertou 166 alinhamentos no modo normal e acertou 172 alinhamentose no modo otimizado.

Wheelan [60] efetuou dois testes comparativos entre o pacote *Spidey* e as duas outras ferramentas analisadas.

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, e *est_genome* reconheceu 3716 exons, dos quais 97.4% estavam corretos.

O segundo 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* 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.

8 Correlação de SNPs

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

O processo tradicional de análise 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 desequilíbrio de ligação (LD) trarão resultados mais completos e precisos sobre estas doenças [47]. LD tem sido apontado como uma ferramenta de grande utilidade para facilitar o mapeamento de genótipos complexos [4].

8.1 Definição de LD

Desequilíbrio de ligação, 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.

À medida que haplótipos (combinação de alelos encontrados em loci próximos dentro de um cromossomo) de ancestrais se propagam em uma população, a distância física tende a ser reduzida por eventos de recombinação, e eventos de recombinação entre marcadores próximos 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 desequilíbrio de ligação.

8.2 Medidas utilizadas para quantificar um LD

Várias medidas foram criadas para se quantificar LDs. A mais antigas das medidas propostas para desequilíbrio é chamada D. Esta medida quantifica um LD como sendo a diferença entre a frequência observada entre um haplótipo 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 [15]). As mais comuns são os valores absolutos D' [35] e r^2 [28].

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 nunca são 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 pelo produto das frequências dos SNPs nos dois loci. O caso $r^2 = 1$, conhecido como LD perfeito, acontece se, e somente se, os marcadores nunca são separados por recombinação e têm a mesma frequência alélica. Portanto, quando se tem um LD completo, informações sobre um marcador fornecem informações completas sobre o outro marcador, fazendo as duas informações redundantes.

A medida r^2 tem sido muito utilizada para se definir o que são LDs úteis [33]. 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 $r^2 > 1/3$ são considerados como úteis em processos de mapeamento.

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 efetivo da população (Effective Population Size [46]), 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).

8.3 Pacotes computacionais que analisam LD

Nesta seção iremos descrever duas ferramentas existentes, ambas escritas em Java, para analisar LDs em sequências de nucleotídeos.

8.3.1 LDA - Linkage Disequilibrium Analyser

O software LDA [16] (*Linkage Disequilibrium Analyser*) fornece uma interface gráfica 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.

O algoritmo executado pela ferramenta LDA é o seguinte: aplica-se um teste para verificar se os alelos em cada locus seguem o equilíbrio de Hardy-Weinberg (HWE) [59]. Para os loci que seguem HWE, aplica-se um algoritmo para estimar as frequências dos quatro haplótipos nos pares de loci [31]. São aplicadas várias medidas nos LD, baseadas nos coeficientes padrões D, D' e o quadrado do coeficiente de correlação r^2.

8.3.2 GOLD - Graphical Overview of Linkage Disequilibrium

A ferramenta GOLD [1] (*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.

9 Proposta

O estudo de SNPs envolve várias etapas, formando um fluxo de processamento e análise. Neste trabalho, pretendemos estudar as três etapas descritas anteriormente (alinhamento de cDNA com DNA genômico, detecção e correlação de SNPs), tentando propor uma metodologia que englobe todas estas etapas.

Em relação ao problema de alinhamento, vimos que os programas analisados tentam alinhar cDNA e DNA genômico utilizando várias etapas, e em particular, todos trabalham com a idéia de encontrar alinhamentos locais com alto grau de semelhança. A ferramenta *est_genome* utiliza mais parâmetros do que os algoritmos tradicionais para obter seus alinhamentos, e *sim4* e *Spidey* utilizam o algoritmo do BLAST, para encontrar possíveis zonas de exons. Após análise dos resultados, observa-se que *est_genome* tem desempenho bem inferior às duas outras ferramentas.

Neste trabalho, tentaremos obter bons alinhamentos de cDNA ou mRNA com DNA genômico utilizando os algoritmos tradicionais, definindo um conjunto de pontuações e penalizações que se adequem a este fim, mantendo uma boa performance de execução. Dentre as possíveis variações no cálculo da pontuação, podemos citar:

- Utilização de mais parâmetros, além dos quatro conhecidos (open gap, extended gap, match e mismatch).
- Estimativa do número de íntrons esperados no DNA genômico.
- Utilização de parâmetros sensíveis ao tamanho das sequências.

Os pacotes analisados que fornecem suporte para análise dos SNPs permitem que sejam calculadas medidas de interesse e permitem a visualização gráfica dos dados. Porém requerem que sejam fornecidos dados sobre os SNPs de forma manual.

Entende-se que em um sistema de análise de SNPs, os mecanismos de detecção e análise devem ser integrados, de forma a que seja possível obter informações sobre os possíveis candidatos a SNP.

Assim, neste trabalho pretendemos propor um mecanismo de análise semi-automática dos dados obtidos nas fases anteriores do processo de estudo de SNPs (alinhamento e detecção), que dêem suporte à análise imediata dos resultados obtidos.

Finalmente, iremos analisar as metodologias de detecção de SNPs existentes, tentando identificar pontos a serem melhorados, e formas de integração dos diferentes métodos.

10 PIPE

O programa PIPE (Programa de Inovação Tecnológica em Pequenas Empresas) é um programa da FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) destinado a financiar o desenvolvimento de pesquisas inovadoras a serem executadas por pequenas empresas de base tecnológica, e que tenham retorno comercial ou social.

O trabalho proposto se enquadra no projeto PIPE "Sistema de Identificação de Polimorfismos" (Processo Fapesp 03/07748-9) concedido à empresa Scylla Bioinformática S.A, sob coordenação do Prof. João Meidanis, e com vigência no período de 01/07/2004 a 30/06/2006.

O objetivo principal deste projeto é realizar pesquisas que permitam ao software SIP (Sistema de Identificação de Polimorfismos, desenvolvido pela Scylla Bioinformática) se fixar como um software inovador, competitivo no mercado nacional e internacional, que confere aos seus usuários vantagens decisivas na análise de SNPs.

O trabalho aqui proposto compreenderá a documentação das metodologias desenvolvidas no contexto deste projeto PIPE, e será desenvolvido nas instalações da Scylla Bioinformática com recursos da própria empresa e da FAPESP, e com o auxílio de mais duas pessoas dedicadas a este projeto.

11 Cronograma

A Tabela 2 descreve a distribuição das atividades a serem realizadas durante a execução deste trabalho.

	2004				2005												2006	
	Set	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Jan	Fev
1	I			II		III												
2						IV			V		VI							
3											VII			VIII		IX		
4																	X	XI

Tabela 2: Cronograma de atividades.

1. Alinhamento de cDNA com DNA genômico:
   - `I` - Estudo e identificação de parâmetros ideais para alinhamento.
   - `II` - Testes com os novos métodos de alinhamento desenvolvidos.
   - `III` - Escrita dos resultados obtidos nos testes de alinhamento.
2. Correlação de SNPs:
   - `IV` - Análise dos métodos existentes e formulação de uma nova metodologia de correlação de SNPs.
   - `V` - Testes computacionais com os novos métodos de correlação de SNPs.

- `VI` - Escrita dos resultados obtidos nos testes de correlação de SNPs.

3. Detecção de SNPs:
   - `VII` - Análise das metodologias aplicadas por programas existentes e formulação de uma nova metodologia para detecção de SNPs.
   - `IX` - Testes computacionais com os novos métodos de detecção de SNPs.
   - `X` - Escrita dos resultados obtidos nos testes de detecção.
4. Dissertação:
   - `XI` - Revisão final do texto da dissertação.
   - `XII` - Defesa da dissertação.

Referências

[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] F. Aklilu. Dotview. Another local alignment tool.

[4] K. G. Ardlie, L. Kruglyak, and M. Seielsta. Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 3:299–309, 2002.

[5] Berkeley Drosophila Genome Project. http://www.fruitfly.org/.

[6] A. J. Brookes. The essence of SNPs. *Gene*, 234:177–186, 1999.

[7] T. A. Brown. *Genomes*. John Wiley and Sons, Inc, 1999.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] The Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[13] D. N. Cooper, E. V. Ball, and M. Krawczak. The human gene mutation database. *Nucleic Acid Research*, 26:285–287, 1998.

[14] dbEST – The International Expressed Sequence Tags Database, September 2002. www.ncbi.nlm.nih.gov/dbEST.

[15] B. Devlin and N. A. Risch. A comparison of linkage disequilibrium measures for fine scaling mapping. *Genomics*, 29:311–322, 1995.

[16] 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.

[17] M. K. Halushka et al. Patterns of single-nucleotide polymorphisms in candidate genes regulating blood-pressure homeostasis. *Nature Genetics*, 22:239–247, 1999.

[18] B. Ewing and P. Green. Base-Calling of Automated Sequencer Traces Using *Phred* II. Error Probabilities. *Genome Research*, 8:186–194, 1998.

[19] 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.

[20] 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.

[21] FORESTs: Eucalyptus Genome Sequencing Consortium, July 2004. http://forests.esalq.usp.br/.

[22] Agronomical & environmental genomes, July 2004. http://watson.fapesp.br/AEG/agro.htm.

[23] D. B. Goldstein, S. K. Tate, and S. M. Sisodiya. Pharmacogenetics goes genomics. *Nature Reviews*, 4:937–947, December 2003.

[24] 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.

[25] O. Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162:705–708, 1982.

[26] P. Green. Phrap documentation, September 2002. www.phrap.org.

[27] Human Cancer Genome Project - SNP. http://bit.fmrp.usp.br/.

[28] W. G. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38:226–231, 1968.

[29] D. S. Hirschberg. A linear space algorithm for computing longest common subsequences. *Communications of the ACM*, 175:341–343, 1975.

[30] X. Huang. On global sequence alignment. *Computer Applications in the Bioscience*, 10:227–235, 1994.

[31] R. R. Hudson. Linkage disequilibrium and recombination. In *Handbook of Statistical Genetics*, pages 309–324. Wiley, 2001.

[32] 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.

[33] L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, 22:139–144, 1999.

[34] 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.

[35] R. C. Lewontin. The interaction of selection and linkage. *Genetics*, 49:49–67, 1964.

[36] 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.

[37] 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.

[38] R. Mott. EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *Computer Applications in the Biosciences*, 13:477–478, 1997.

[39] A. Motulsky. Drug reactions, enzymes, and biochemical genetics. *Journal of the American Medical Association*, 165:835–837, 1957.

[40] E. W. Myers and W. Miller. A file comparison program. *Software - Practice and Experience*, 15:1025–1040, 1985.

[41] E. W. Myers and W. Miller. Optimal alignment in linear space. *Computer Applications in the Biosciences*, 4(1):11–17, 1988.

[42] National Center for Biotecnology Information. http://www.ncbi.nlm.nih.gov/.

[43] 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.

[44] 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.

[45] R. S. Phelps, R. B. Chadwick, M. P. Conrad, M. N Kronick, and A. Kamb. *BioTechniques*, 19:984–989, 1995.

[46] J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *American Journal of the Human Genetics*, 69:1–14, 2001.

[47] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Nature*, 273:1516–1517, 1996.

[48] 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.

[49] 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.

[50] *Schistosoma mansoni* EST Genome Project, September 2002. http://verjo18.iq.usp.br/schisto.

[51] G. D. Schuler. Pieces of the puzzle: Expressed sequence tags and the catalogue of human genes. *Journal of Molecular Medicine*, 75:694–698, 1997.

[52] J. C. Setubal and J. Meidanis. *Introduction to Computional Molecular Biology*. PWS Publishing Company, 1997.

[53] A. J. G. Simpson, F.C. Reinach, P. Arruda, F. A. Abreu, M. Acencio, R. Alvarenga, L. M. C. Alves, J. E. Araya, G. S. Baia, C. S. Baptista, M. H. Barros, E. D. Bonaccorsi, S. Bordin, J. M. Bove, M. R. S. Briones, M. R. P. Bueno, A. A. Camargo, L. E. A. Camargo, D. M. Carraro, H. Carrer, N. B. Colauto, C. Colombo, F. F. Costa, M. C. R. Costa, C. M. Costa-Neto, L. L. Coutinho, M. Cristofani, E. Dias-Neto, C. Docena, H. El-Dorry, A. P. Facincani, A. J. S. Ferreira, V. C. A. Ferreira, J. A. Ferro, J. S. Fraga, S. C. França, M. C. Franco, M. Frohme, L. R. Furlan, M. Garnier, G. H. Goldman, M. H. S. Goldman, S. L. Gomes, A. Gruber, P. L. Ho, J. D. Hoheisel, M. L. Junqueira, E. L. Kemper, J. P. Kitajima, J. E. Krieger, E. E. Kuramae, F. Laigret, M. R. Lambais, L. C. C. Leite, E. G. M. Lemos, M. V. F. Lemos, S. A. Lopes, C. R. Lopes, J. A. Machado, M. A. Machado, A. M. B. N. Madeira, H. M. F. Madeira, C. L. Marino, M. V. Marques, E. A. L. Martins, E. M. F. Martins, A. Y. Matsukuma, C. F. M. Menck, E. C. Miracca, C. Y. Miyaki, C. B. Monteiro-Vitorello, D. H. Moon, M. A. Nagai, A. L. T. O. Nascimento, L. E. S. Netto, A. Nhani, F. G. Nobrega, L. R. Nunes, M. A. Oliveira, M. C. De Oliveira, R. C. De Oliveira, D. A. Palmieri, A. Paris, B. R. Peixoto, G. A. G. Pereira, H. A. Pereira, J. B. Pesquero, R. B. Quaggio, P. G. Roberto, V. Rodrigues, A. J. De M. Rosa, V. E. De Rosa, R. G. De Sá, R. V. Santelli, H. E. Sawasaki, A. C. R. Da Silva, A. M. Da Silva, F. R. Da Silva, W. A. Silva, J. F. Da Silveira, M. L. Z. Silvestri, W. J. Siqueira, A. A. De Souza,

A. P. De Souza, M. F. Terenzi, D. Truffi, S. M. Tsai, M. H. Tsuhako, H. Vallada, M. A. Van Sluys, S. Verjovski-Almeida, A. L. Vettore, M. A. Zago, M. Zatz, J. Meidanis, and J. C. Setubal. The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature*, 406(6792):151–159, July 2000.

[54] T. E. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[55] The Sugar Cane EST Genome Project, September 2002. http://www.sucest.lad.ic.unicamp.br.

[56] 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.

[57] The Human Cancer Genome Project, September 2002. http://www.ludwig.org.br/ORESTES.

[58] 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.

[59] B. S. Weir. *Genetic Data Analysis II*. Sinauer Associates, 2nd edition, April 1996. ISBN 0878939024.

[60] 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.