UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
PÓS-GRADUAÇÃO EM ENGENHERIA ELÉTRICA

ENNIO DOS SANTOS BAPTISTA

UMA ABORDAGEM ALTERNATIVA PARA SEQÜENCIAMENTO POR HIBRIDIZAÇÃO

Recife

2003

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
PÓS-GRADUAÇÃO EM ENGENHERIA ELÉTRICA

ENNIO DOS SANTOS BAPTISTA

UMA ABORDAGEM ALTERNATIVA PARA SEQÜENCIAMENTO POR HIBRIDIZAÇÃO

DISSERTAÇÃO APRESENTADA À PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA DO CENTRO DE TECNOLOGIA E GEOCIÊNCIAS/ESCOLA DE ENGENHARIA DE PERNAMBUCO DA UNIVERSIDADE FEDERAL DE PERNAMBUCO COMO REQUISITO PARCIAL PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA.

ORIENTADOR: Prof. Dr. Rafael Dueire Lins

CO-ORIENTADORA: Prof.ª Dr.ª Katia Silva Guimarães

Recife

2003

ENNIO DOS SANTOS BAPTISTA

UMA ABORDAGEM ALTERNATIVA PARA SEQÜENCIAMENTO POR HIBRIDIZAÇÃO

DISSERTAÇÃO APRESENTADA À PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA DO CENTRO DE TECNOLOGIA E GEOCIÊNCIAS/ESCOLA DE ENGENHARIA DE PERNAMBUCO DA UNIVERSIDADE FEDERAL DE PERNAMBUCO COMO REQUISITO PARCIAL PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA.

Aprovada em 22 de setembro de 2003.

BANCA EXAMINADORA

_____

Prof. Fernando Menezes Campello de Souza, Pós-Doutor – UFPE

_____

Prof. Hélio Magalhães de Oliveira, Doutor – UFPE

_____

Prof.ª Katia Silva Guimarães, Pós-Doutora – UFPE

À Patrícia e
às nossas três amadas filhas.

AGRADECIMENTOS

À professora Katia Guimarães, meus sinceros agradecimentos
por ter me acompanhado nessa difícil jornada;

Ao professor Rafael Lins, por todo o seu esforço à frente do curso;

A todos os colegas do curso, por termos vividos juntos
as dificuldades e alegrias desse caminho;

À turma do Biolab–UFPE e da Telemática, por todo o apoio;

Às amigas Andréa e Lucynier, pelas conversas e dicas;

Aos amigos de apartamento, Manu, Mirlem e Nívia, pelo convívio,
apoio e amizade;

Aos amigos da SEFAZ, por toda a ajuda, carinho e, claro, por terem
me tolerado nesses últimos tempos;

Aos meus dedicados tradutores, Mili-Mili e Rômulo;

Às minhas secretárias, Leuda e Nalva, pelo inestimável apoio
no dia-a-dia de casa;

Às minhas queridas irmãs, Lara, Ló e France, pelo amor e incentivo
recebidos;

A todos os demais familiares e amigos que se dispuseram a ouvir
um pouco sobre *Seqüenciamento de DNA*, e até me incentivaram
a seguir em frente;

Em especial, agradeço à minha amada mãe, por toda
a sua dedicação e amor, e ao meu amado pai, pelo exemplo ímpar
de bravura;

Em especial, agradeço, também, à minha querida Patrícia
e às nossas queridas filhas, Aline, Maysa e Beatriz, por... por tudo;

E, acima de tudo, agradeço ao meu Senhor por esse momento.

Só quem se mortifica em vós floresce,
só é senhor de si quem se vos rende,
só sabe pretender quem vos pretende,
e só sobe por vós quem por vós desce.

(Jerónimo Baía)

RESUMO

Uma questão central no emergente campo da Biologia Molecular Computacional diz respeito ao problema de seqüenciamento de DNA. Seqüenciar uma molécula de DNA significa determinar a ordem das suas bases componentes – adenina (A), citosina (C), guanina (G) e timina (T). Em vista do comprimento de tais moléculas, muitas vezes da ordem de bilhões de bases, e das limitações existentes nos processos laboratoriais, os quais são capazes de manipular, no máximo, apenas 700 bases, esse se tornou um problema de natureza combinatorial e normalmente requer técnicas matemáticas e recursos computacionais para a sua solução.

Dentre os vários métodos de seqüenciamento de DNA desenvolvidos nas últimas décadas, um que se tem mostrado particularmente promissor é o método denominado Seqüenciamento por Hibridização (do inglês *Sequencing by Hybridization – SBH*), o qual se caracteriza por utilizar um *chip* de DNA para identificar o *espectro* da seqüência investigada, isto é, o conjunto de todas as subseqüências de um determinado tamanho que a compõem; e por tentar seqüenciá-la a partir das informações nele contidas. Recentemente, Halperin et al. (2002) apresentou duas variantes para o SBH. A primeira é baseada em um algoritmo, denominado algoritmo $\mathcal{A}$, projetado para lidar com os dados gerados pelo chip clássico de seqüenciamento; e a segunda, mais abrangente, inclui um novo modelo de chip que conta com bases universais distribuídas randomicamente, e, para lidar com os dados provenientes dele, inclui também um outro algoritmo, denominado algoritmo $\mathcal{B}$. Halperin et al. (2002) ainda sugeriu que a combinação adequada de alguns aspectos positivos dessas abordagens talvez pudesse gerar resultados práticos melhores do que os obtidos com a solução baseada apenas no algoritmo $\mathcal{B}$.

Este trabalho de pesquisa aponta os problemas de se implementar tal sugestão e, então, propõe uma abordagem alternativa que tende a superá-los, a qual mostrou-se ser mais geral, tendo, inclusive, a solução baseada no algoritmo $\mathcal{B}$ como um caso particular. Além disso, as simulações realizadas evidenciaram que os demais casos conseguem alcançar melhor rendimento em termos do tamanho da seqüência que pode ser corretamente determinada, empregando chips de menor custo.

Palavras-chave: seqüenciamento de DNA; seqüenciamento por hibridização; SBH; chip de DNA; chip de seqüenciamento; bases universais.

ABSTRACT

A central question in the emerging field of Computational Molecular Biology is the problem of DNA sequencing. Sequencing a DNA molecule requires determining the order of its component bases – adenine (A), cytosine (C), guanine (G) and thymine (T). Often, however, due to the length of these molecules that may contain billions of bases and because of limitations in existing laboratory methods, which are only capable of handling a maximum of 700 bases, the DNA sequencing process has become combinatorial in nature and usually requires mathematical techniques and computational means to be solved.

Among the various DNA sequencing techniques developed in the last decades, Sequencing by Hybridization (SBH) has shown itself to be particularly promising. SBH is characterized by the use of a DNA chip to identify the spectrum of the sequence being investigated, that is, the set of all the subsequences of a particular size which make up the primary sequence; and by attempting to sequence it based on the information contained within the spectrum. Halperin et al. (2002) recently presented two variations of this technique. The first variation includes an algorithm, called algorithm $\mathcal{A}$, designed to operate with data generated by the classic sequencing chip. The second, and more general variation includes a new type of chip, characterized by the presence of universal, randomly distributed bases and another algorithm called algorithm $\mathcal{B}$, designed to work on data produced by the chip itself. In addition, Halperin et al. suggests that perhaps a third technique that combines certain aspects of the previously described two, may be able to provide better practical results than those ones derived by the original algorithm $\mathcal{B}$.

This study identified the problems associated with implementing the aforementioned suggested technique while, simultaneously, presenting an alternative approach which tends to surpass them. This approach has shown itself to be more general, including, as a particular case, the solution based on the algorithm $\mathcal{B}$. In addition, simulations carried out show that other cases are also able to generate better results in terms of the size of the sequence that can be correctly determined, while using lower-cost chips.

Key Terms: DNA sequencing; sequencing by hybridization; SBH; DNA chip; sequencing chip; universal bases.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

Esta pesquisa se insere no emergente campo da Biologia Molecular Computacional, área multidisciplinar que emprega técnicas matemáticas e computacionais para solucionar problemas biológicos (SETUBAL; MEIDANIS, 1997), mais especificamente, as novas questões advindas da descoberta dos mecanismos celulares responsáveis pelos processos genéticos nos seres vivos.

De modo geral, podemos nos referir a tais questões como sendo aquelas cujas soluções envolvem dados representantes dos fatores bioquímicos que, mesmo ocorrendo em nível molecular, regem a vida de todos os seres vivos, desde o mais simples até o mais complexo dos organismos, determinando suas características estruturais e funcionais, e garantindo a transmissão dessas às gerações seguintes.

Exemplos de problemas que têm sido largamente investigados no âmbito dessa nova ciência são: construção de árvore filogenética, seqüenciamento de DNA, mapeamento físico de cromossomo, mapeamento genético, predição de gene e predição de estrutura molecular, para mencionar apenas alguns.

Em particular, neste trabalho, abordamos o problema do seqüenciamento de DNA. Esse é um procedimento básico da genética experimental, cujo objetivo é determinar a seqüência dos nucleotídeos constituintes de uma molécula de DNA. Mais especificamente, na nossa pesquisa, lidamos com um método de seqüenciamento relativamente novo denominado Seqüenciamento por Hibridização (do inglês *Sequencing by Hybridization* – SBH) (BAINS; SMITH, 1988; DRMANAC et al., 1989; LYSOV et al., 1988; SOUTHERN, 1988).

Recentemente, Halperin et al. (2002) apresentou duas abordagens para o SBH e, além disso, sugeriu que talvez uma nova estratégia, combinando aspectos particulares delas, pudesse gerar melhores resultados práticos. Esta dissertação aponta os problemas de se implementar tal sugestão e, ao mesmo tempo, apresenta uma estratégia alternativa que tende a superá-los, bem como os resultados decorrentes de sua avaliação experimental.

Um fator que impulsiona as pesquisas na área de seqüenciamento é a esperança de que, uma vez seqüenciado o DNA de um organismo, os estudos possam prosseguir até que se consiga interpretar as informações genéticas nele codificadas. Drmanac e Drmanac (2001) destacam a importância dessa tarefa quando afirmam que “pela determinação da seqüência do DNA de um organismo, pesquisadores podem obter informações críticas sobre o seu desenvolvimento e fisiologia, as suas relações taxonômicas e a sua susceptibilidade a doenças”. Nesse contexto, várias áreas podem se beneficiar dos resultados dessas pesquisas, como é o caso da Medicina e da Agricultura.

O seqüenciamento de DNA é um dos problemas mais estudados em Biologia Molecular. Normalmente, não é um procedimento com fim em si mesmo, mas é etapa inicial de muitas outras investigações nessa área. Assim, os resultados que fornece – cadeias

seqüenciadas – são os dados de entrada de outras etapas. Isso significa que o sucesso de tais investigações depende direta e decisivamente do sucesso do método de seqüenciamento empregado. Tal dependência justifica grande parte do esforço despendido por pesquisadores do mundo inteiro no desenvolvimento de métodos e ferramentas de seqüenciamento de DNA que sejam cada vez mais eficientes. É exatamente nesse contexto de crescente busca pela "melhor" técnica de seqüenciamento que a nossa pesquisa se justifica.

No Capítulo 1, esta dissertação introduz conceitos biológicos básicos necessários ao entendimento dos capítulos seguintes; no Capítulo 2, discute o problema de seqüenciamento e apresenta duas técnicas que são empregadas tradicionalmente em sua solução; no Capítulo 3, apresenta a relativamente nova técnica de Seqüenciamento por Hibridização – SBH; no Capítulo 4, apresenta as variantes propostas por Halperin et al. (2002) para o SBH; no Capítulo 5, apresenta a estratégia alternativa para o SBH que elaboramos a partir daquelas propostas por Halperin et al. (2002), bem como os resultados de sua avaliação experimental. Por fim, a dissertação conclui, resumindo os resultados da pesquisa e sugerindo trabalhos futuros.

1 UM POUCO DE BIOLOGIA MOLECULAR COMPUTACIONAL

1.1 BIOLOGIA COMPUTACIONAL

Ao entrarmos no Terceiro Milênio, é pouco provável que alguma área de atividade humana ainda não tenha se apoiado, direta ou indiretamente, em algum recurso da Ciência da Computação. A Biologia, de modo algum, é exceção, mas, pelo contrário, devido a essa cooperação vem experimentando um extraordinário avanço nas últimas décadas.

Nos laboratórios, vários procedimentos foram automatizados, permitindo que experimentos de melhor qualidade fossem realizados em uma fração do tempo e do custo normalmente necessários. Também, muitos problemas biológicos tidos como insolúveis ou que só podiam ser considerados em pequena escala foram superados com aplicação de técnicas e ferramentas computacionais adequadas. Talvez essa seja a contribuição mais interessante da Ciência da Computação para a Biologia, mas certamente não é a última. Atualmente, estudiosos da área contam com imensos bancos de dados distribuídos ao redor do mundo e acessíveis através de rede de computadores, tanto como fonte de informações biológicas para subsidiar as suas pesquisas quanto como meio de divulgação dos resultados obtidos.

A aplicação de técnicas computacionais, mais precisamente de técnicas algorítmicas, na solução de problemas biológicos tem sido tão intensa que os conhecimentos adquiridos ao longo do tempo fomentaram o surgimento de uma nova área de estudo denominada Biologia Molecular Computacional.

As questões principais tratadas no âmbito dessa emergente área referem-se geralmente a problemas da Biologia Molecular ou da Biologia Evolucionária (PEDERSEN, 1999). Em ambos os casos, o foco das investigações se concentra sobre os mecanismos genéticos responsáveis pelas características estruturais e funcionais de cada ser vivo e pela transmissão delas aos seus descendentes.

Estudos recentes comprovam que o material genético, que armazena informações a respeito dos organismos, como se fosse o seu projeto de construção, está intimamente relacionado com um conjunto de moléculas que se encontram no interior de cada uma de suas células. Essas moléculas são os ácidos nucléicos e, em última instância, as proteínas.

O próximo item tem duplo propósito. O primeiro, é o de apresentar os conceitos biológicos básicos relativos ao material genético, necessários à compreensão do restante do trabalho. O segundo, é o de deixar patente o tipo de informação normalmente manipulada pelos algoritmos projetados na Biologia Molecular Computacional.

1.2 SEQÜÊNCIAS BIOLÓGICAS

Os ácidos nucléicos e as proteínas são polímeros, isto é, moléculas construídas a partir de subunidades menores, denominadas monômeros. Ambas as moléculas são ditas moléculas

informacionais por guardarem informações genéticas em suas estruturas (LEHNINGER; NELSON; COX, 1995).

1.2.1 Ácidos Nucléicos

Ácido nucléico é um termo genérico que serve para designar tanto o ácido desoxirribonucléico (do inglês *desoxyribonucleic acid* – DNA) quanto o ácido ribonucléico (do inglês *ribonucleic acid* – RNA). Essas substâncias foram assim denominadas em virtude de serem ácidas e porque se acreditava que elas ocorriam exclusivamente no núcleo das células.

Os ácidos nucléicos são polinucleotídeos, isto é, polímeros cujas unidades monoméricas são nucleotídeos. Um nucleotídeo consiste de três moléculas menores: uma base nitrogenada ligada a uma pentose, por sua vez, ligada a um fosfato. As pentoses que compõem esses ácidos são de dois tipos. No RNA, a pentose é a ribose; e no DNA, é a desoxirribose. Ambas possuem 5 átomos de carbono, numerados convencionalmente desde 1’ até 5’. As bases nitrogenadas estão categorizadas em bases púricas e bases pirimídicas. Fazem parte do primeiro grupo a adenina e a guanina (comumente denotadas por A e G); e do segundo, a citosina, a timina e a uracila (normalmente denotadas por C, T e U), sendo que a timina ocorre exclusivamente no DNA e a uracila, exclusivamente no RNA. Essas duas bases mais o tipo de pentose usada fazem a distinção entre os dois tipos de ácidos nucléicos.

Nos ácidos nucléicos, os nucleotídeos se sucedem, sendo que o local 3’ da pentose de um está ligado, através do grupo fosfato, ao local 5’ da pentose do seu sucessor. Como ilustrado na Figura 1, o ácido nucléico pode ser visto como um arcabouço formado pela

sucessão alternada de fosfato e pentose, tendo as bases como grupos laterais. Essa estrutura permite que um ácido nucléico seja caracterizado simplesmente pela seqüência de suas bases, a qual, por convenção, é sempre considerada no sentido de 5' para 3', como ilustra a Figura 2.

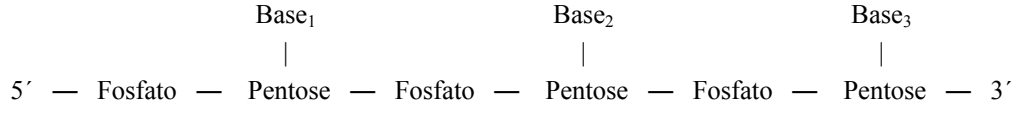


Figura 1: Esquema geral de um fragmento de ácido nucléico. Adaptado de (PEDERSEN, 1999).

Essa simplificação é conveniente para os propósitos computacionais, pois permite que essas duas seqüências biológicas sejam modeladas como seqüências de caracteres sobre um alfabeto de 4 letras {A, C, G, T}, no caso do DNA, e sobre {A, C, G, U}, no caso do RNA. Desse modo, problemas biológicos podem muitas vezes ser interpretados como problemas de manipulação de cadeias de caracteres (do inglês *strings*), que é um assunto bastante conhecido da Teoria da Computação, e para os quais há disponível uma imensa quantidade de algoritmos.

5′ — A C G T C A T A C G — 3′

Figura 2: Representação de um fragmento de DNA através da sua seqüência de bases.

Na realidade, esse modelo de seqüência de caracteres estaria mais próximo da estrutura do RNA do que da estrutura do DNA. A razão é que, diferentemente do RNA, que é formado por uma única seqüência de nucleotídeos, o DNA, segundo a estrutura de hélice dupla, postulada por James D. Watson e Francis Crick[1], em 1953, é formado por duas seqüências

[1] A estrutura postulada por Watson-Crick, conhecida como estrutura B, não é a única. Além dela, há ainda as estruturas A e Z, porém ela é o padrão de referência em qualquer estudo (LEHNINGER; NELSON; COX, 1995).

enroladas helicoidalmente em torno de um eixo imaginário comum, sendo que as bases de uma formam pares com as bases da outra, através de pontes de hidrogênio. O par de base, denotado por *pb*, é comumente usado como unidade de comprimento da molécula de DNA. As pontes de hidrogênio, que ocorrem especificamente entre bases complementares, ou seja, entre A e T, e entre C e G, fazem com que essas seqüências, também ditas complementares, se mantenham unidas. A essa interação entre ácidos nucléicos, por meio de pontes de hidrogênio, denomina-se *hibridização* (CHETVERIN; KRAMER, 1994). Essa regra de complementaridade das bases garante que, se uma das seqüências for conhecida, a outra poderá ser imediatamente inferida. Os algoritmos geralmente aproveitam essa informação e manipulam somente uma das seqüências. Nesse contexto, o modelo descrito volta a servir para representar igualmente RNA e DNA. Outra característica da hélice dupla de DNA é que as cadeias são antiparalelas, ou seja, enquanto uma está orientada no sentido convencional de 5′ para 3′, a outra está no sentido contrário, ou seja, de 3′ para 5′. A Figura 3 representa um trecho de DNA em hélice dupla.

```
5′ — A C G T C A T A C G — 3′
     | | | | | | | | | |
3′ — T G C A G T A T G C — 5′
```

Figura 3: Representação de um fragmento de DNA em hélice dupla. As linhas verticais representam pontes de hidrogênio.

1.2.2 Proteínas

Nas proteínas, as unidades monoméricas são os aminoácidos, sendo que um aminoácido é uma molécula composta por um átomo de hidrogênio, um grupo amina, um grupo carboxila e uma cadeia lateral, todos ligados a um carbono central, denominado α-carbono. Apesar de existir um número significativo de proteínas, apenas 20 tipos diferentes de aminoácidos são

encontrados em suas composições, sendo que a diferença entre quaisquer dois deles é somente a cadeia lateral.

Como ilustrado na Figura 4, numa proteína, os aminoácidos estão ligados seqüencialmente uns aos outros através de ligações denominadas ligações peptídicas, que se estabelecem entre o carbono do grupo carboxila de um e o nitrogênio do grupo amina do seguinte. Essa estrutura das proteínas permite que elas também, a exemplo dos ácidos nucléicos, sejam especificadas simplesmente pelas suas cadeias laterais. Mais uma vez essa simplificação interessa à computação. Nesse caso, as proteínas podem ser modeladas como seqüências de caracteres sobre um alfabeto de 20 caracteres (cada um correspondendo a um tipo diferente de aminoácido).

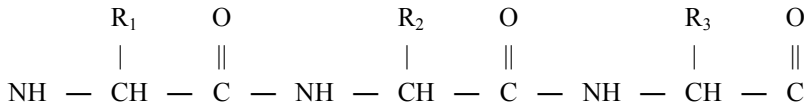


Figura 4: Esquema geral de um fragmento de proteína. Adaptado de (PEDERSEN, 1999).

Pereira (2001) cita que são as proteínas "que vão formar e dirigir a formação e o funcionamento do organismo". Realmente, as proteínas estão relacionadas às mais diversas funções nos organismos. Elas são principalmente enzimas, proteínas transportadoras, proteínas nutrientes e de armazenamento, proteínas contráteis ou de motilidade, proteínas estruturais, proteínas de defesa e proteínas reguladoras (LEHNINGER; NELSON; COX, 1995). Pereira (2001) descreve essas funções afirmando que as proteínas dirigem a construção de todas as estruturas que compõem as células; que algumas constituem elas mesmas outras partes das células; e, ainda, que outras são responsáveis por milhões de reações bioquímicas.

1.2.3 Síntese de Proteínas

Essas seqüências biológicas, tanto ácidos nucléicos quanto proteínas, apesar de suas estruturas seqüenciais simples, baseadas na repetição de um pequeno número de unidades básicas, são moléculas que desempenham papéis capitais na vida dos seres vivos, visto que constituem os seus mecanismos genéticos.

Atualmente sabe-se que, na base desses mecanismos, está o *genoma* do organismo. O genoma pode ser visto como o projeto do organismo com as informações para a sua construção e manutenção. São informações que dizem respeito à formação das suas diversas estruturas, bem como do seu funcionamento. É como se fosse uma receita com todas as instruções de que a natureza precisa para desenvolvê-lo.

Diversos estudos revelaram que o DNA é tipicamente o meio ou material usado pelos organismos para armazenar concretamente as suas instruções genômicas, sendo que cada uma delas está codificada em termos de A's, C's, G's e T's em algum trecho da sua extensa cadeia de nucleotídeos. Cada um desses trechos ou instruções denomina-se *gene*.

Sempre que necessário, a instrução genética é lida e executada no interior da própria célula por uma estrutura celular denominada ribossomo. A leitura não é feita diretamente no DNA, mas cabe ao RNA, mais especificamente, ao RNA mensageiro, o papel de fazer uma cópia da instrução e de transportá-la até os ribossomos. Tendo sido sintetizado a partir de um molde da seqüência complementar do gene, o RNA mensageiro é, na realidade, ele próprio, a cópia do gene, com a diferença de que nele as bases T's foram substituídas por bases U's. Esse processo de cópia chama-se *transcrição*, significando que a informação codificada no

alfabeto do DNA – {A, C, G, T} – passou a ser codificada no alfabeto do RNA – {A, C, G, U}. É importante ressaltar que o processo de transcrição descrito corresponde ao tipo que ocorre nos organismos *procariotos* (cujas células não possuem membrana nuclear). Nos demais organismos, nos *eucariotos* (cujas células possuem membrana nuclear), a transcrição é ligeiramente diferente, pois muitos dos seus genes possuem trechos na seqüência, denominados *introns*, que precisam ser descartados do RNA mensageiro após a transcrição.

Primeira Posição	Segunda Posição				Terceira Posição
	G	A	C	U	
G	Gly	Glu	Ala	Val	G
	Gly	Glu	Ala	Val	A
	Gly	Asp	Ala	Val	C
	Gly	Asp	Ala	Val	U
A	Arg	Lys	Thr	Met	G
	Arg	Lys	Thr	Ile	A
	Ser	Asn	Thr	Ile	C
	Ser	Asn	Thr	Ile	U
C	Arg	Gln	Pro	Leu	G
	Arg	Gln	Pro	Leu	A
	Arg	His	Pro	Leu	C
	Arg	His	Pro	Leu	U
U	Trp	-	Ser	Leu	G
	-	-	Ser	Leu	A
	Cys	Tyr	Ser	Phe	C
	Cys	Tyr	Ser	Phe	U

Figura 5: O Código Genético.

A execução de cada instrução genética levada ao ribossomo resulta na síntese de uma proteína específica, o que faz com que algumas vezes essa estrutura celular seja chamada de fábrica de proteínas. Nela, a cadeia de RNA mensageiro é lida seqüencialmente, sendo que cada conjunto de três nucleotídeos ou *códon*, como é conhecido, determina um aminoácido para compor a proteína. Esse processo denomina-se *tradução*, significando que a informação genética codificada no alfabeto do RNA foi traduzida para o alfabeto de aminoácidos. A

relação entre todos os aminoácidos e as trincas de nucleotídeos correspondentes, ilustrada na Figura 5, denomina-se código genético.

Uma vez que, via transcrição e tradução, é o DNA que determina a seqüência de aminoácidos de uma proteína e que muitos estudos evidenciam que essa seqüência, por sua vez, é fator determinante da estrutura espacial e, por conseguinte, da própria função biológica da proteína, é comum que as investigações que envolvem esta extraordinária molécula comecem pela determinação da seqüência de nucleotídeos correspondente, isto é, pelo seqüenciamento do DNA.

2 O PROBLEMA DE SEQÜENCIAMENTO DE DNA

2.1 VISÃO GERAL

O seqüenciamento de DNA é uma questão central na área da Biologia Molecular Computacional. Seqüenciar uma molécula de DNA significa determinar a ordem dos seus nucleotídeos componentes – adenina (A), citosina (C), guanina (G) e timina (T). Lehninger, Nelson e Cox (1995) afirmam que, "na sua capacidade de reservatório da informação, a mais importante propriedade de uma molécula de DNA é a sua seqüência de nucleotídeos".

Um fator que justifica a busca por técnicas de seqüenciamento cada vez mais confiáveis, velozes e de menor custo é que conhecer a seqüência da molécula de DNA de um organismo pode ser a chave para explicar o que o organismo é, tanto do ponto de vista estrutural quanto do ponto de vista funcional. Meidanis e Setubal (1994) afirmam que "muita informação pode ser obtida se tivermos a seqüência completa de um trecho importante de DNA".

Percebe-se, assim, que o seqüenciamento não é um procedimento com fim em si mesmo, mas, ao contrário, seus resultados interessam a muitas outras investigações. Burks (1994) aponta que as informações codificadas no DNA podem servir, entre outras aplicações,

de plataforma para aumentar nossa habilidade em: caracterizar e entender doenças e infecções humanas; projetar e desenvolver medicina terapêutica e preventiva; desenvolver fontes nutricionais melhoradas e empregar ferramentas microbiais em ambiente.

No mesmo sentido, Meidanis e Setubal (1994) asseguram que, se a seqüência de um gene for conhecida, "pode-se, através do código genético, obter a seqüência de aminoácidos da proteína correspondente, pode-se ainda fazer análises de preferência de códons, verificar se há pontos de corte para enzimas de restrição conhecidas, comparar o mesmo gene de diferentes indivíduos para localizar diferenças (que podem indicar doenças hereditárias) etc".

Drmanac e Drmanac (2001) destacam a importância desta tarefa quando afirmam que, "pela determinação da seqüência do DNA de um organismo, pesquisadores podem obter informações críticas sobre o seu desenvolvimento e fisiologia, as suas relações taxonômicas, e a sua susceptibilidade a doenças".

O grande desafio dos métodos de seqüenciamento empregados atualmente é de que consigam ser úteis em projetos que lidam com seqüenciamento em larga escala, isto é, com seqüenciamento de longas moléculas de DNA, com até mesmo bilhões de pares bases, como no caso do ousado Projeto Genoma Humano, iniciado em 1988 por um consórcio de pesquisadores americanos e concluído em 2001, depois de determinar os $3x10^9$ pares de bases que compõem o genoma do homem e de gerar uma imensa expectativa a respeito dos benefícios que advirão desse conhecimento.

Hoje, o seqüenciamento em larga escala é possível graças aos aprimoramentos que ocorreram tanto na Bioquímica quanto na Ciência da Computação durante as últimas décadas.

Pevzner (2000) mostra que no início foi diferente. Ele lembra que, na origem dos trabalhos de seqüenciamento de DNA, está o trabalho que resultou na determinação da seqüência dos aminoácidos da pequena molécula de proteína de insulina e que rendeu a Frederick Sanger, em 1953, seu primeiro prêmio Nobel de Química.

Frederick Sanger usou a estratégia do tipo dividir para conquistar. Naquela época, já se tinha conhecimento do processo denominado *Degradação de Edman*, que consiste na retirada e identificação de um aminoácido de cada vez do final da seqüência de proteína. A repetição desse processo permitia que os aminoácidos fossem lidos um a um. O problema, entretanto, é que os resultados se tornavam difíceis de interpretar a partir do quarto ou quinto aminoácido.

Diante desse obstáculo, Frederick Sanger adotou uma estratégia com três etapas bem definidas. Na primeira, as moléculas de insulina eram quebradas em vários fragmentos de tamanhos adequados. Na segunda, cada fragmento era seqüenciado diretamente pelo Método de Degradação de Edman. Finalmente, na terceira etapa, os fragmentos eram montados, com base nas sobreposições existentes entre eles, recriando a molécula original. Por exemplo, os três fragmentos da Figura 6 definem a seqüência *Gly Ile Val Glu Gln Cys Cys Ala*.

Gly Ile Val Glu
Ile Val Glu Gln
Gln Cys Cys Ala

Gly Ile Val Glu Gln Cys Cys Ala

Figura 6: Esquema de reconstrução de uma seqüência de aminoácidos com base na sobreposição dos seus fragmentos. Extraído de (PEVZNER, 2000).

É importante lembrar que, guardadas as proporções, o problema enfrentado por biólogos atuais, quando seqüenciam longas moléculas de DNA, é semelhante ao que foi enfrentado por

Sanger, de modo que esse princípio de “fragmentar-seqüenciar-montar” é largamente usado até hoje.

2.2 MÉTODOS TRADICIONAIS DE SEQÜENCIAMENTO

Nas últimas décadas, diversos métodos de seqüenciamento de DNA foram propostos, os quais podem ser categorizados em métodos diretos e métodos indiretos de seqüenciamento (DRMANAC et al., 2002). Fazem parte do primeiro grupo os métodos em que a determinação da seqüência total da molécula de DNA se dá pela identificação individual de cada uma das suas bases componentes. Outra característica marcante desses métodos é que limitações práticas impedem que eles sejam usados com seqüências com comprimentos maiores que algumas poucas centenas de bases. Nesse grupo, o principal exemplo, por ser o mais empregado, é o método denominado *Terminação de Cadeia* ou *Dideoxi*, o qual rendeu a Frederick Sanger, em 1980, o seu segundo prêmio Nobel (SANGER, 1977). Por outro lado, são considerados métodos indiretos aqueles que primeiramente determinam as partes ou fragmentos que constituem determinada seqüência, usando para isso normalmente um método direto, e depois as processam para tentar reconstruir a seqüência original, a qual é usualmente referida nesse contexto como *seqüência alvo*. Esses métodos são usualmente empregados no seqüenciamento em larga escala, isto é, com longas seqüências, sendo que o de maior importância é o método denominado *Shotgun*.

2.2.1 O Método de Sanger

Apesar de ter sido inventado há quase 30 anos, o Método de Terminação de Cadeia de Sanger, hoje, aprimorado e com alguns processos automatizados, continua sendo bastante

usado em virtude de ser tecnicamente simples. Basicamente, o método consiste de três etapas. Na primeira, são adequadamente gerados fragmentos de todos os tamanhos possíveis da molécula de DNA, de uma maneira que o nucleotídeo terminal de cada um deles pode facilmente ser identificado. Na segunda, todos os fragmentos são ordenados por tamanho. Finalmente, na terceira etapa, a leitura adequada dos nucleotídeos terminais dos fragmentos fornece a seqüência de nucleotídeos da seqüência alvo.

Na prática, o método envolve quatro reações bioquímicas, as quais têm por componentes cópias da molécula de DNA a ser seqüenciada: iniciadores (do inglês *primers*), deoxinucleotídeos (dATP, dCTP, dGTP, dTTP), dideoxinucleotídeos (ddATP, ddCTP, ddGTP, ddTTP) e enzima DNA polimerase. Os iniciadores são pequenas seqüências de nucleotídeos complementares ao trecho inicial da molécula de DNA. O deoxinucleotídeo e o dideoxinucleotídeo são componentes químicos que liberam os nucleotídeos que irão compor os fragmentos. Na realidade, o dideoxinucleotídeo libera um tipo de nucleotídeo ligeiramente modificado, com capacidade de se ligar apenas a um outro nucleotídeo e não a dois, como é o normal na formação da fita simples. A DNA polimerase é a enzima que catalisa as reações.

Inicialmente, para gerar os fragmentos com diversos tamanhos, o método aproveita o mecanismo da síntese de DNA pela reação da DNA polimerase. Nessa reação de síntese, o processo começa com cada iniciador se ligando ao trecho inicial da molécula de DNA. Em seguida, os conjuntos iniciador-DNA formados são colocados para reagir, sob a ação da DNA polimerase, com os deoxinucleotídeos. O papel da enzima é de estender o iniciador pela adição sucessiva dos nucleotídeos liberados pelos deoxinucleotídeos. Cada nucleotídeo, quando é adicionado, faz duas ligações, uma com o nucleotídeo terminal do iniciador e outra com a fita de DNA, conforme a regra de complementaridade (A–T e C–G). Posteriormente,

fará ainda uma terceira ligação com o próximo nucleotídeo a entrar na seqüência. O processo termina quando o nucleotídeo terminal da fita de DNA é alcançado. Nesse momento, tem-se uma nova fita de DNA complementar à fita original que serviu de molde.

No Método de Sanger, o processo de formação dessa nova fita é ligeiramente diferente devido à presença adicional de moléculas de dideoxinucleotídeos nas reações. Sob a ação da DNA polimerase, os dideoxinucleotídeos também irão liberar os seus nucleotídeos modificados para que estendam o iniciador. A diferença é que esse tipo de nucleotídeo, ao ser incluído na seqüência, encerra prematuramente o processo, antes mesmo do final da fita de DNA ser atingido, não permitindo que, depois dele, nenhum outro seja adicionado à seqüência. Assim, se em uma reação o único tipo de dideoxinucleotídeo presente for o ddATP, então todos os fragmentos obtidos pela interrupção prematura da reação terão como base terminal a adenina (A). Além disso, se a quantidade desse composto for adequada é possível que, no conjunto desses fragmentos terminados com a base A, existam representantes de todos os tamanhos possíveis. Da mesma forma, outras três reações podem ser feitas para os outros três tipos de dideoxinucleotídeos – ddCTP, ddGTP e ddTTP.

Dando prosseguimento ao processo de seqüenciamento, os fragmentos obtidos são submetidos a um processo de eletroforese, onde são colocados em gel dentro de um campo elétrico, para que, sob o efeito deste, fiquem ordenados por tamanho. Os fragmentos de mesmo tamanho tendem a ficar juntos formando bandas no gel. Então, a partir de uma leitura adequada dessas bandas, que leve em consideração tanto a posição delas no gel quanto o tipo de nucleotídeo terminal dos seus fragmentos constituintes, pode-se determinar a ordem dos nucleotídeos da molécula de DNA sendo seqüenciada. No início, para que as bandas fossem lidas, empregava-se radiação para marcar os fragmentos e um filme de raios X para registrar o

posicionamento delas. Atualmente, o processo está automatizado, a marcação é feita com corantes fluorescentes e a leitura, com laser. A Figura 7 ilustra as três etapas do método.

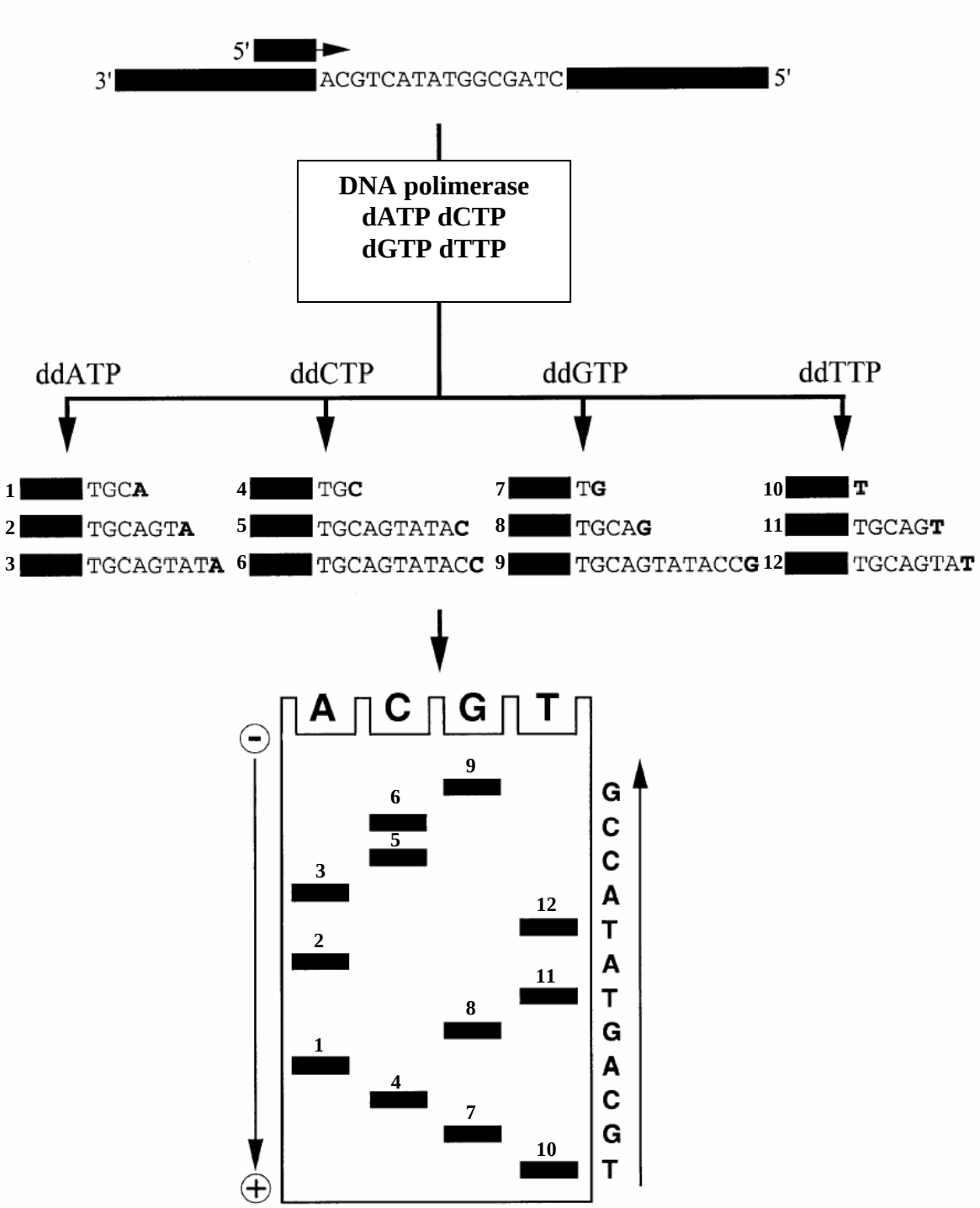


Figura 7: Esquema de funcionamento do método de Terminação de Cadeia de Sanger. Adaptado de (STERKY; LUNDEBERG, 2000).

A maior desvantagem dessa técnica é que, mesmo com os avanços tecnológicos atuais, ela não pode ser empregada diretamente no seqüenciamento de longas moléculas de DNA. A razão é que o tamanho máximo de seqüenciamento é limitado pela qualidade da resolução do gel empregado na eletroforese, de modo que, na prática, o tamanho da maior seqüência que se consegue determinar é de aproximadamente de 700 bases (SETUBAL; MEIDANIS, 1997).

Nos projetos de seqüenciamento em larga escala, esse problema normalmente é contornado pelo emprego de uma estratégia que fragmenta a longa molécula de DNA, seqüencia os fragmentos obtidos e, finalmente, a partir deles, reconstrói a seqüência original. Essa estratégia é conhecida como Shotgun e será descrita na seção seguinte.

2.2.2 O Método Shotgun

Um problema bastante atual e que cada vez mais vem se tornando rotina na área da Biologia Molecular diz respeito à tarefa de seqüenciar longas moléculas de DNA. Tal problema pode ser visto como uma versão ampliada do problema enfrentado por Frederick Sanger, no passado, no seqüenciamento da molécula de insulina, de modo que a sua solução segue o mesmo princípio da solução adotada naquela ocasião, ou seja, permanece a idéia de se reduzir a seqüência original a vários fragmentos que podem ser seqüenciados diretamente para, em seguida, serem montados com base em informações de sobreposição existentes entre o trecho final de um e o trecho inicial do seguinte.

Notadamente, o Método Shotgun é o método indireto mais utilizado no seqüenciamento de longas moléculas de DNA. Ele tem as características positivas de ser *paralelizável*, *econômico*, *automatizável* (SANGER et al., 1982) e de possuir *escalabilidade* (TAMMI,

2003). Ele já não é mais um método puramente experimental, mas, em virtude da complexidade e do volume dos dados normalmente envolvidos, alguns de seus procedimentos requerem obrigatoriamente o apoio de ferramentas computacionais.

2.2.2.1 Fases do Método Shotgun

De modo geral, a estratégia Shotgun consiste nas seguintes tarefas (MEIDANIS; SETUBAL, 1994):

Fragmentação aleatória de diversas cópias do DNA – em um momento inicial, a molécula de DNA passa por um processo de replicação ou clonagem. Em seguida, as várias cópias resultantes são quebradas em posições aleatórias de suas seqüências. A replicação é necessária para que haja pontos de sobreposição entre fragmentos oriundos de cópias diferentes e, por conseqüência, informação que ajude na reconstrução da seqüência alvo. Freqüentemente, a fragmentação das moléculas é feita por um processo que emprega ultra-som, fazendo com que as moléculas de DNA vibrem até que se rompam em pontos aleatórios. Uma outra abordagem possível, mas menos empregada, utiliza enzimas de restrição para cortar as longas moléculas de DNA em posições específicas, determinadas pela própria composição da enzima empregada.

Seleção dos fragmentos – nem todos os fragmentos obtidos são aproveitados. Normalmente, os fragmentos grandes ou pequenos demais são descartados, não sendo usados nas etapas seguintes do seqüenciamento.

Clonagem dos fragmentos selecionados – os fragmentos selecionados são em seguida clonados para que cada fragmento não seja único no experimento.

Seqüenciamento dos fragmentos – todos os fragmentos clonados são seqüenciados por algum processo direto de seqüenciamento, como por exemplo, pelo Método de Terminação de Cadeia de Sanger.

Montagem de fragmentos – no Método Shotgun, a última tarefa consiste exatamente em reunir todos os fragmentos seqüenciados de modo que se consiga deduzir a seqüência original. Esse problema é comumente conhecido como o problema de *Montagem de Fragmentos*. Na realidade, essa é a etapa do seqüenciamento que realmente desperta o interesse do cientista da computação, e para a qual diversos algoritmos têm sido propostos. A próxima seção apresenta os seus principais aspectos.

2.2.2.2 O Problema de Montagem de Fragmentos

Supondo que as primeiras fases do Método Shotgun tenham rendido um conjunto de fragmentos seqüenciados de uma dada molécula de DNA, o problema de Montagem de Fragmentos ou, como também é conhecido, o problema de Reconstrução da Seqüência Alvo, consiste em descobrir a ordem em que tais fragmentos devem ser considerados para que a seqüência original seja reconstruída e, por conseguinte, a seqüência de suas bases componentes determinada.

De modo geral, o esquema de montagem conta apenas com a informação de sobreposição que se espera existir entre o conjunto de fragmentos. Ocorre sobreposição entre

dois fragmentos quando a seqüência de um é subseqüência do outro ou quando o trecho inicial de um coincide com o trecho final do outro. Para ilustrar, tem-se que os fragmentos TCGACTAG e GACT se sobrepõem de acordo com o primeiro caso, e que os fragmentos GCCATGAC e TGACCGTA se sobrepõem de acordo com o segundo. A Figura 8 ilustra os vários casos de sobreposição que podem ocorrer entre dois fragmentos (TAMMI, 2003).

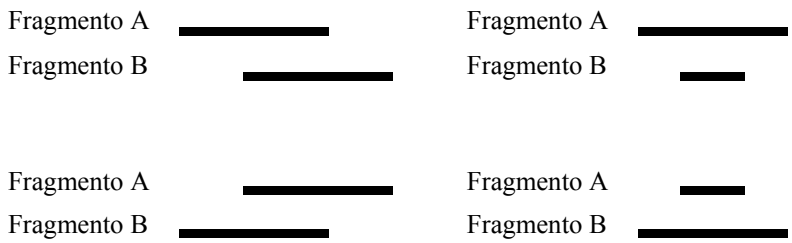


Figura 8: A sobreposição entre dois fragmentos pode ser de quatro tipos. Adaptado de (TAMMI, 2003).

Apesar de várias soluções já terem sido propostas para o problema de Montagem de Fragmentos, freqüentemente elas seguem o mesmo princípio clássico de dividir o problema nas três subtarefas elementares de *overlap*, *layout* e *consensus* (PEVZNER, 2000). Para a discussão nos próximos itens de cada um desses passos, deve-se considerar, como exemplo, que a etapa de montagem recebeu como entrada os fragmentos ilustrados na Figura 9.

1. T A C G T
2. A G A
3. C G T A
4. G A C G
5. T A G

Figura 9: Exemplo de fragmentos de entrada para a fase de montagem.

Fase de *overlap* – essa é a fase inicial do processo de montagem dos fragmentos, na qual todos os pares de fragmentos que apresentam sobreposições significativas entre si são

identificados. Nessa fase, a definição de um número mínimo de bases que se sobrepõem é importante e corresponde a uma tentativa de excluir do processo aqueles pares cujos fragmentos, tendo se originado em locais não contíguos da seqüência de DNA, ocasionalmente se sobrepõem. Para o conjunto de fragmentos do exemplo, se a sobreposição mínima for fixada em duas bases, somente os seguintes pares serão identificados nessa etapa de *overlap*: (1,3), (2,4), (2,5) e (3,5).

Fase de *layout* – nessa fase, são identificados, com base no grau de sobreposição existente entre os fragmentos, todos os posicionamentos possíveis que os mesmos podem assumir, uns em relação aos outros, para reconstrução da seqüência alvo. Para ilustrar, a Figura 10 apresenta dois *layouts* válidos para o conjunto de fragmentos apresentados na Figura 9. Nesse caso, se não houver nenhuma informação adicional, não é possível identificar qual das opções é a correta. Conhecer o tamanho da seqüência alvo previamente é um fator que ajuda a esclarecer a questão. Atualmente é possível que esse tamanho seja determinado com margem de erro de 10% (SETUBAL; MEIDANIS, 1997).

```
T A C G T - - - - -        - - - - - - - T A C G T
- - - - - A G A - -          A G A - - - - - - -
- - C G T A - - - -        - - - - C G T A - - -
- - - - - - - G A C G      - - G A C G - - - - -
- - - - T A G - - -        T A G - - - - - - - -
```

Figura 10: Dois *layouts* possíveis para o conjunto de fragmentos de entrada da Figura 9.

Fase de *consensus* – a cada *layout* encontrado, corresponde uma seqüência que é candidata à seqüência alvo. Normalmente, esta seqüência é conhecida como *seqüência de consenso* devido à forma como é obtida. Em um *layout*, os diversos fragmentos estão alinhados de modo que cada coluna corresponde a um tipo de base. A seqüência de consenso

nada mais é do que a seqüência dessas bases (ou colunas). A Figura 11 ilustra os *layouts* apresentados na Figura 10 com suas respectivas seqüências de consenso.

```
T A C G T - - - - -        - - - - - - - T A C G T
- - - - - - A G A - -          A G A - - - - - - - -
- - C G T A - - - - -        - - - - C G T A - - - -
- - - - - - - G A C G        - - G A C G - - - - - -
- - - - - T A G - - -        T A G - - - - - - - - -
T A C G T A G A C G         T A G A C G T A C G T
```

Figura 11: Seqüências de consenso resultantes dos *layouts* da Figura 10.

A tarefa de reunir os diversos fragmentos como num quebra-cabeça unidirecional, por si só, é muito complexa. Não bastasse isso, alguns problemas ainda podem aumentar o grau de dificuldade. Por exemplo, é comum a ocorrência de fragmentos contendo erros. O desconhecimento da orientação desses fragmentos, se de 5' para 3' ou de 3' para 5', também é outro complicador, juntamente com a ocorrência de regiões repetidas ou com falta de cobertura.

Erros – é possível que os procedimentos para a geração dos fragmentos introduzam erros nas suas seqüências de bases, fazendo com que os mesmos não sejam subseqüências exatas da seqüência alvo, mas apenas aproximadas. Três tipos de erros podem ocorrer. O primeiro, conhecido como *base-calling error*, resulta da inserção, da deleção ou da substituição de algumas bases do fragmento. Esse tipo de erro é o mais simples e pode nem afetar o cômputo da seqüência de consenso. O segundo tipo de problema diz respeito à formação de fragmentos quiméricos, os quais resultam da união indevida de fragmentos advindos de posições não contíguas da molécula sendo seqüenciada. Por fim, o terceiro tipo de problema ocorre quando fragmentos estranhos à molécula de DNA do experimento conseguem se misturar ao conjunto legítimo dos seus fragmentos.

Orientação desconhecida – cada fragmento obtido pelo Método Shotgun pode ter sua origem em uma das duas fitas da molécula de DNA, de modo que as duas possibilidades devem ser exploradas durante a montagem.

Regiões repetidas – seqüências com repetições longas são mais difíceis de determinar e, normalmente, conduzem a situações de ambigüidade, em que não se tem como identificar seguramente qual das seqüências reconstruídas corresponde à seqüência alvo.

Falta de cobertura – a falta de cobertura é um problema relacionado à qualidade do conjunto de entrada de fragmentos. Devido à forma randômica como os mesmos são obtidos, é comum que em determinado trecho faltem fragmentos, ou seja, informações para determinar a seqüência de consenso. A Figura 12 ilustra essa situação. Uma solução para esse tipo de problema é fazer seqüenciamento adicional de fragmentos, mas mesmo isso pode falhar por não se poder especificar exatamente a região descoberta (MEIDANIS; SETUBAL; 1994).

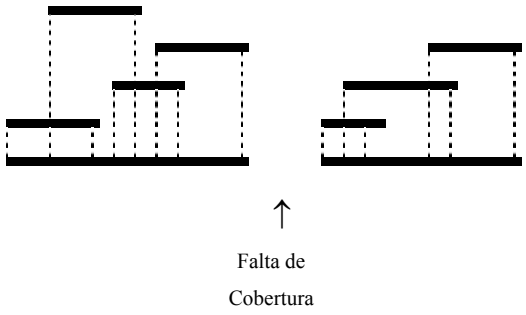


Figura 12: Esquema para ilustrar o problema de falta de cobertura.

2.2.2.3 Formalização do Problema de Montagem de Fragmentos

O problema de Montagem de Fragmentos tem sido tradicionalmente modelado como

um problema de manipulação de cadeias de caracteres, mais especificamente como o problema da *Mínima Supercadeia Comum* (MSC) (PETOLA et al, 1984; TUNNER, 1986; KECECIOGLU, 1991), o qual pode ser formalmente definido como:

Definição 1 (MEIDANIS; SETUBAL, 1994) – *Dadas as seqüências* $u_1, u_2, ..., u_n$ *sobre um alfabeto* Σ*, o problema da Mínima Supercadeia Comum (MSC) consiste em encontrar uma seqüência S mais curta possível, tal que cada* u_i *seja um fator de S...*

Na realidade, esse é um modelo com pouca ou nenhuma aplicabilidade prática. Ele é mais importante como recurso teórico para compreensão das questões envolvidas no problema de montagem. Meidanis e Setubal (1994) apontam que essa formalização falha em modelar aspectos biológicos importantes do problema, como:

a) o modelo não admite que os fragmentos a serem montados contenham erros;

b) o modelo pressupõe que a orientação dos fragmentos é conhecida;

c) a opção pela seqüência mais curta não, necessariamente, conduz à solução correta do problema;

d) o problema de falta de cobertura não é considerado.

Adicionalmente, segundo Maier (1978), é provavelmente que não exista nenhum algoritmo eficiente computacionalmente para todas as instâncias do problema. Por outro lado, a vantagem desse modelo é que o problema de determinar a Mínima Supercadeia Comum

para um conjunto Γ de seqüências é uma questão da área da Teoria da Computação bastante conhecida, sendo que existe uma abordagem matematicamente elegante para ela, a qual consiste em mudar o contexto do problema para o domínio dos grafos. Essa mudança faz com que o problema se reduza à tarefa de encontrar um Caminho Hamiltoniano em um grafo, denominado grafo de sobreposição, construído adequadamente com as seqüências do conjunto Γ. Formalmente, Caminho Hamiltoniano pode ser definido como:

Definição 2 – *Um Caminho Hamiltoniano em um grafo $\mathcal{G}$ é um caminho que atravessa cada vértice de $\mathcal{G}$ exatamente uma vez.*

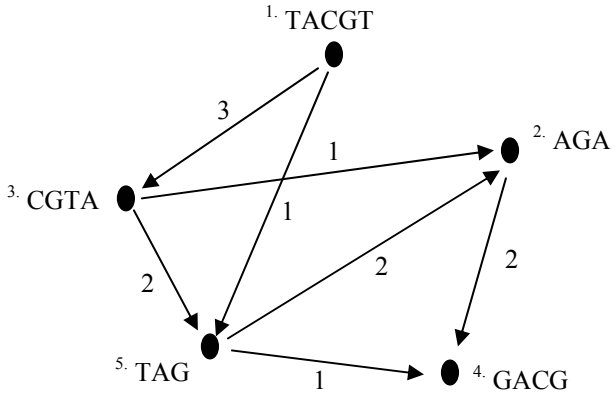


Figura 13: Exemplo de um grafo de sobreposição para os fragmentos da Figura 9.

No grafo de sobreposição, cada vértice representa uma seqüência de Γ, e cada aresta, a sobreposição existente entre as duas seqüências que liga. As arestas são orientadas e ponderadas. A orientação será no sentido da seqüência *u* para a seqüência *v*, se a sobreposição ocorrer entre as bases terminais de *u* e as bases iniciais de *v,* e será no sentido de *v* para *u*, caso contrário. O peso das arestas corresponde justamente ao tamanho das sobreposições a que se referem. A Figura 13 representa o grafo de sobreposição para os fragmentos apresentados na Figura 9.

Num grafo de sobreposição, cada caminho corresponde a uma supercadeia dos vértices pertencentes a ele. Na realidade, tem-se que cada caminho gera um *layout* com esses vértices, a partir do qual a supercadeia pode ser deduzida. Por exemplo, para o caminho com a seqüência de vértices 1, 3, 5 do grafo da Figura 13, obtém-se o *layout* e a supercadeia ilustrados na Figura 14.

```
T A C G T - -
- - C G T A -
- - - - T A G
---------------
T A C G T A G
```

Figura 14: Supercadeia correspondente ao caminho 1, 3, 5 do grafo da Figura 13.

Obviamente, para a supercadeia que envolve todos os fragmentos dados, o caminho a ser procurado é algum que percorra todos o vértices do grafo sem repeti-los, ou seja, um Caminho Hamiltoniano. Ademais, se o interesse for pela supercadeia de menor extensão, isto é, pela supercadeia que maximiza as sobreposições entre os fragmentos, o Caminho Hamiltoniano a ser procurado no grafo de sobreposição será o de maior peso nas arestas.

2.3 OUTROS MÉTODOS DE SEQÜENCIAMENTO

Além dos métodos tradicionalmente empregados – o Método de Terminação de Cadeia de Sanger e o Método Shotgun –, os pesquisadores ainda contam com várias outras abordagens para o seqüenciamento de DNA. Usualmente, também se aplicam os Métodos de *Degradação Química*, *Primer Walking*, *Nested Deletion*, *Seqüenciamento por Hibridização* dentre outros.

Degradação Química – A invenção desse método direto de seqüenciamento rendeu a Alan Maxam e Walter Gilbert um prêmio Nobel em 1980 (MAXAM; GILBERT, 1977). Sua idéia básica é a mesma do Método de Terminação de Cadeia de Sanger (LEHNINGER; NELSON; COX, 1995), ou seja, busca-se decompor a seqüência de DNA em quatro conjuntos de fragmentos de acordo com as bases terminais que possuem. A diferença está apenas no tipo das reações empregadas. As fases seguintes, de separação dos fragmentos por eletroforese e de leitura do auto-radiograma, são idênticas. De todo modo, o Método de Sanger, por ser tecnicamente mais simples, é o mais amplamente usado.

Primer Walking – Dentro da categoria dos métodos indiretos de seqüenciamento de DNA, esse método classifica-se como um método dirigido, significando que, diferentemente do Método Shotgun, no qual os fragmentos a serem seqüenciados são obtidos de forma aleatória, nele, os fragmentos são obtidos e seqüenciados de acordo com a ordem que aparecem na seqüência de DNA investigada. O método consiste no uso de um iniciador adequado para o seqüenciamento das j primeiras bases da seqüência, onde j é o tamanho usualmente permitido pelos métodos diretos de seqüenciamento. A primeira subseqüência obtida, além de ser parte da resposta procurada, o seu trecho final serve para a confecção de um novo iniciador, o qual começará o seqüenciamento das próximas j bases. Esse processo é repetido até que a seqüência de DNA seja completamente conhecida. As abordagens dirigidas possuem duas grandes desvantagens. A primeira é que elas são essencialmente seqüenciais, isto é, são não-*paralelizáveis*, e por conseguinte, lentas. A segunda está relacionada ao fato de que, se o processo for interrompido em algum lugar da seqüência, o seqüenciamento do trecho além daquele ponto ficará prejudicado.

Nested Deletion – a exemplo do Método *Primer Walking*, esse é um método indireto e dirigido de seqüenciamento. A idéia básica dessa abordagem se resume em duas etapas que se repetem sucessivamente até que a seqüência do DNA tenha sido determinada por completo. Na primeira, o trecho inicial da seqüência é determinado pela aplicação de um método direto de seqüenciamento. Na segunda etapa, uma enzima é aplicada para remover o que foi seqüenciado e, assim, expor um novo trecho inicial da seqüência restante para que o processo recomece. Sendo um método dirigido, o *Nested Deletion* apresenta as mesmas desvantagens descritas para o Método *Primer Walking*.

Seqüenciamento por Hibridização – essa é uma abordagem relativamente nova para o problema de seqüenciamento de DNA que tem despertado grande interesse nos pesquisadores, por ser um método alternativo para as tradicionais abordagens baseadas em eletroforese com gel. A idéia central do novo método é a de que uma seqüência de DNA possa ser reconstruída a partir do conjunto de todas as suas subseqüências de um determinado tamanho. Uma vez que o Seqüenciamento por Hibridização é a base que fundamenta este trabalho, o capítulo seguinte é reservado a sua apresentação.

3 SEQÜENCIAMENTO POR HIBRIDIZAÇÃO

3.1 INTRODUÇÃO

Por volta do final dos anos 80, alguns grupos de pesquisadores propuseram um método alternativo de seqüenciamento de DNA para as abordagens clássicas baseadas em eletroforese com gel, denominado Seqüenciamento por Hibridização (do inglês *Sequencing by Hybridization* – SBH) (BAINS; SMITH, 1988; DRMANAC et al., 1989; LYSOV et al., 1988; SOUTHERN, 1988).

A idéia central desse novo método é a de que uma seqüência de DNA desconhecida pode ser determinada se o conjunto de todas as suas subseqüências de um tamanho especificado l (l-tuplas) for conhecido. Esse conjunto denomina-se *espectro* da seqüência. Por exemplo: para $l = 4$, a seqüência ACTGTTAG tem como espectro o conjunto {ACTG, CTGT, TGTT, GTTA, TTAG}.

O método SBH consiste em um procedimento com duas etapas complementares bem definidas (BLAZEWICZ et al., 2000). A primeira é uma etapa experimental que se passa nos laboratórios, na qual milhares de reações bioquímicas de hibridização são processadas

simultaneamente para gerar o espectro de uma seqüência de DNA dada. A segunda é uma etapa combinatorial que, devido ao volume de dados e à complexidade das operações envolvidas, normalmente, exige o emprego de recursos computacionais. Essa etapa recebe como entrada o espectro gerado na fase anterior e, a partir das informações nele contidas, busca determinar a ordem em que as bases aparecem ao longo da seqüência original.

3.2 ETAPA DE HIBRIDIZAÇÃO

Dada uma seqüência de DNA desconhecida, o trabalho realizado na fase experimental do SBH consiste em determinar todas as suas l-tuplas constituintes. Para isso, emprega-se um dispositivo denominado *chip de seqüenciamento* ou *chip de DNA*, o qual possui uma superfície estruturada na forma de matriz, sobre a qual são fixados, em posições individuais e bem definidas, todos os 4^l oligonucleotídeos (pequenas cadeias de nucleotídeos) de tamanho k ou *sondas* (do inglês *probes*). Na realidade, esse modelo de chip corresponde ao tipo clássico, comumente representado por $C(l)$.

No início do seqüenciamento, o chip é colocado em contato com uma solução contendo cópias da seqüência de DNA a ser determinada, todas previamente marcadas por fluorescência ou radiação, e, em seguida, é lavado, sendo que somente aquelas cópias que tiverem hibridizado com alguma de suas sondas permanecerão ligadas à sua superfície. A hibridização ocorrerá sempre que a sonda for complemento, segundo Watson-Crick, de alguma l-tupla da seqüência. Em seguida, um procedimento adequado de leitura do chip revela o conjunto de todas as posições marcadas do chip e, conseqüentemente, de todas as sondas que hibridizaram. E, uma vez que estas são seqüências complementares às l-tuplas da seqüência de DNA, o espectro pode ser facilmente deduzido.

É importante ressaltar que algumas vezes o conjunto de sondas é referido como sendo o próprio espectro. Nesse caso, deve-se observar que a seqüência reconstruída ao final do processo não será exatamente a seqüência alvo, mas sim o seu complemento. Este trabalho adota essa segunda definição para espectro.

A Figura 15 ilustra o resultado do processo de hibridização da pequena seqüência TAGACTTGAC (complementar da seqüência ATCTGAACTG) com o chip clássico $C(4)$ que possui todas as 256 sondas de tamanho 4. Para esse caso, tem-se que o espectro é $\Gamma = \{ATCT, TCTG, CTGA, TGAA, GAAC, AACT, ACTG\}$.

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA								AACT								
AC															ACTG	
AG																
AT								ATCT								
CA																
CC																
CG																
CT									CTGA							
GA		GAAC														
GC																
GG																
GT																
TA																
TC															TCTG	
TG	TGAA															
TT																

Figura 15: Hibridização da seqüência TAGACTTGAC com o chip clássico de seqüenciamento $C(4)$. Adaptado de (PEVZNER, 2000).

Na realidade, raramente o espectro resultante de um experimento prático corresponde perfeitamente ao conjunto das l-tuplas encontradas na seqüência de DNA. A razão é que, a despeito da sofisticação e da precisão dos procedimentos laboratoriais empregados, é impossível eliminar completamente os erros experimentais. Sendo assim, é fundamental que os procedimentos de determinação da seqüência a partir do seu espectro saibam lidar tanto

com os *erros falsos positivos* (sondas presentes indevidamente no espectro) quanto com *erros falsos negativos* (sondas ausentes indevidamente do espectro).

Deve-se observar que o espectro, na realidade, é um conjunto e não um multiconjunto, o que significa que cada sonda aparece nele apenas uma única vez, independente de poder, potencialmente, hibridizar em mais de um local da seqüência. Essa é uma situação que ocorre quando há repetição de um trecho de tamanho l no DNA.

3.3 ETAPA COMBINATORIAL

No SBH, conhecer o espectro da seqüência é apenas a parte inicial de todo o trabalho de seqüenciamento, pois ainda é preciso determinar a ordem em que as sondas devem ser consideradas na reconstrução[2] da seqüência alvo. Essa é uma tarefa de caráter combinatorial que normalmente exige o emprego de técnicas matemáticas e algorítmicas associado ao uso de computadores poderosos. Pode-se dizer que é nessa etapa que o seqüenciamento efetivamente ocorre. Formalmente, o problema de determinação da seqüência alvo pelo método SBH pode ser definido como:

Definição 3 (GUSFIELD, 1997) – *O problema SBH consiste em se determinar o máximo possível da cadeia de DNA alvo S a partir da lista Γ de todas as subseqüências de tamanho l que aparecem em S. Em particular, se possível, determinar exclusivamente a seqüência S original a partir da lista Γ.*

[2] O termo "reconstrução" certamente é mais apropriado ao contexto do método Shotgun, onde a seqüência de DNA alvo é inicialmente reduzida a fragmentos para posteriormente ser montada novamente.

Embora a fase experimental do SBH e do tradicional Método Shotgun sejam bem diferentes, tanto com relação aos procedimentos propriamente ditos quanto com relação aos resultados por eles gerados, ambos os métodos possuem pontos em comum importantes. Por exemplo, eles seguem o mesmo princípio pelo qual a seqüência investigada é primeiramente determinada em termos das suas partes constituintes para, só então, num momento posterior, ser montada e seqüenciada por inteiro no computador. Além disso, o procedimento combinatorial, no SBH, aproveita a mesma estratégia de montagem aplicada pelo Shotgun, a qual se baseia no uso de informações sobre as sobreposições existentes entre as extremidades dos seus fragmentos.

Na realidade, o tipo de sobreposição existente entre as sondas do SBH é bem diferente. A característica peculiar de formação do espectro, decorrente da estrutura de construção do chip clássico de seqüenciamento, que garante que todas as l - 1 bases finais de um sonda se sobrepõem às l - 1 bases iniciais da sonda que aparecerá na posição imediatamente à frente na seqüência, é extremamente útil, pois sugere a um algoritmo de seqüenciamento como ele deve proceder para ir montando um sonda sobre a outra até reconstruir a seqüência completa desejada.

```
T A G A C T T G A C
| | | | | | | | | |
A T C T G A A C T G
---------------------
A T C T
  T C T G
    C T G A
      T G A A
        G A A C
          A A C T
            A C T G
---------------------
```

Figura 16: Esquema de determinação da seqüência ATCTGAACTG a partir do espectro gerado com o chip $C(4)$.

Como exemplo, a Figura 16 esquematiza a reconstrução da seqüência TAGACTTGAC (complementar de ATCTGAACTG) a partir das informações de sobreposição das sondas obtidas com um chip $C(4)$.

3.3.1 O Caminho Hamiltoniano

Da mesma forma que o problema de montagem dos fragmentos no Método Shotgun, o problema combinatorial do SBH também pode ser formulado como o problema da Mínima Supercadeia Comum – MSC (definido no item 2.2.2.3), de modo que o objetivo passe a ser a identificação da menor seqüência possível que contenha todas as sondas do espectro como subseqüências. Na realidade, o SBH deve ser visto como um caso especial da MSC, pois todas as subseqüências (ou sondas) a serem consideradas possuem o mesmo tamanho e ocupam uma posição exclusiva na seqüência. Além do mais, deve ser observado que, enquanto no Shotgun o tamanho mínimo de sobreposição entre os fragmentos tem de ser especificado, no SBH, ele é fixo em $l - 1$.

Uma outra forma de considerar o problema foi proposta por Khrapko et al. (1989). Ele aplicou a teoria dos grafos e sugeriu que o problema do SBH é equivalente à tarefa de encontrar um Caminho Hamiltoniano em um grafo orientado, construído adequadamente a partir do espectro da seqüência. No grafo proposto, os vértices representam as sondas do espectro e as arestas, as sobreposições de tamanho $l - 1$ existentes entre elas. Cada aresta é orientada no sentido do vértice u para o vértice v, se a sobreposição ocorre entre as bases finais da sonda relacionada ao vértice u e as bases iniciais da sonda relacionada ao vértice v, caso contrário, a orientação será de v para u.

Em um grafo construído dessa forma, cada Caminho Hamiltoniano corresponde a uma Mínima Supercadeia Comum compatível com o espectro da seqüência alvo. Se no grafo for encontrado apenas um Caminho Hamiltoniano, o problema estará solucionado, e a seqüência alvo determinada. Caso exista mais de um caminho, é necessário que informações adicionais sejam obtidas para que se possa definir com precisão qual das seqüências resultantes corresponde efetivamente àquela que está sendo investigada. Além dessa dificuldade, essa abordagem tem dois grandes obstáculos. O primeiro é que a tarefa de encontrar um Caminho Hamiltoniano é NP-completo. O segundo é que ela não funciona com espectro incompleto, isto é, com espectro contendo erros falsos negativos (PEVZNER, 2000; BLAZEWICZ et al., 1996).

$$\Gamma = \{ATG,\ TGG,\ TGC,\ GTG,\ GGC,\ GCA,\ GCG,\ CGT\}$$

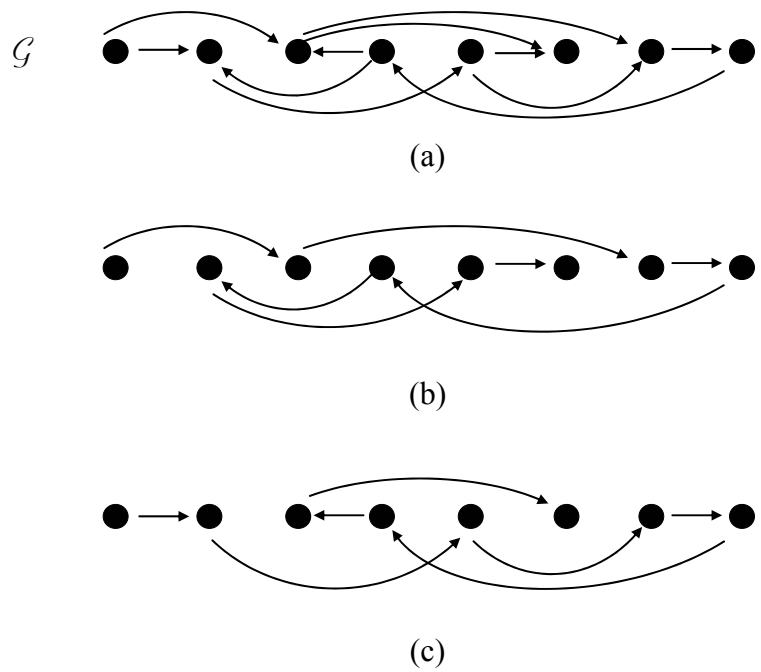


Figura 17: (a) Grafo $\mathcal{G}$ para o espectro Γ com dois Caminhos Hamiltonianos correspondendo às seqüências (b) ATGCGTGGCA e (c) ATGGCGTGCA. Adaptado de (PEVZNER, 2000).

Para ilustrar a solução proposta por Khrapko et al. (1989), a Figura 17(a) representa o grafo $\mathcal{G}$ construído a partir do espectro Γ = {ATG, TGG, TGC, GTG, GGC, GCA, GCG, CGT}. Verifica-se que nele há dois Caminhos Hamiltonianos. A Figura 17(b) representa o caminho que gera a seqüência ATGCGTGGCA, e a Figura 17(c) representa o caminho que gera a seqüência ATGGCGTGCA.

3.3.2 O Caminho Euleriano

Pevzner (1989) propôs uma nova abordagem para o problema SBH. Uma abordagem que, além de ser matematicamente elegante, tem a vantagem de usar um algoritmo de tempo linear para determinar a seqüência alvo. O objetivo da proposta é reduzir o problema à tarefa de encontrar um Caminho Euleriano, e não mais um Caminho Hamiltoniano, em um grafo dirigido construído adequadamente a partir das informações disponíveis no espectro da seqüência. Formalmente, Caminho Euleriano pode ser definido como:

Definição 4 – *Um Caminho Euleriano em um grafo $\mathcal{G}$ é um caminho que atravessa cada aresta de $\mathcal{G}$ exatamente uma vez.*

A característica principal do grafo sugerido por Pevzner (1989) é que nele cada aresta representa uma sonda do espectro. O grafo pode ser construído da seguinte maneira: para cada sonda do espectro obtido a partir de um chip $C(l)$, deve ser gerada uma aresta orientada, partindo de um vértice u em direção a um vértice v, sendo que o vértice u representa as l - 1 bases finais e o vértice v, as l - 1 bases iniciais da sonda associada a essa aresta. O vértice que ainda não fizer parte do grafo em construção deve ser prévia e adequadamente criado.

A Figura 18 ilustra um grafo que foi construído segundo a forma mencionada acima, a partir do seguinte espectro Γ = {ATG, TGG, TGC, GTG, GGC, GCA, GCG, CGT}.

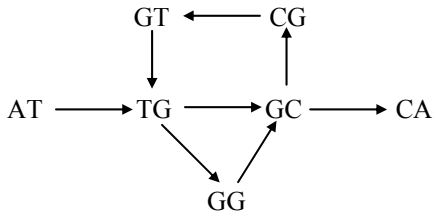


Figura 18: Construção do grafo proposto por Pevzner para o espectro Γ = {ATG, TGG, TGC, GTG, GGC, GCA, GCG, CGT}. Extraído de (PEVZNER, 2000).

3.3.2.1 Limitações do Caminho Euleriano

Apesar do problema de encontrar um Caminho Euleriano em um grafo ser bastante conhecido e de existir para ele solução em tempo linear, a abordagem de Pevzner (1989) sofre de grandes limitações que a impedem de ter aplicabilidade prática imediata, a saber:

a) *Erro no espectro* – para que a abordagem do Caminho Euleriano funcione adequadamente, o espectro resultante dos experimentos laboratoriais deve corresponder exatamente ao conjunto de sondas que efetivamente hibridizam com a amostra de DNA analisada. Normalmente, essa suposição é irreal, pois, na prática, não se consegue eliminar completamente os erros falsos positivos nem os falsos negativos, de modo que o problema torna-se NP-completo (GALLANT; MAIER; STORER, 1980; GALLANT, 1983). No intuito de contornar essa dificuldade, diversos algoritmos têm sido propostos. Por exemplo, para lidar com falsos negativos, tem-se os algoritmos que foram apresentados em (BAINS, 1991; BAINS; SMITH, 1988; BLAZEWICZ et al. 1997; GUENOCHE, 1992; KHRAPKO et al.,

1989; PEVZNER, 1989); para falsos positivos, os que foram apresentados em (DRMANAC; LABAT; CRKVENJAKOV, 1991; LIPSHUTZ, 1993). Entretanto, o mais geral, capaz de lidar com ambos os tipos de erros, foi proposto em (BLAZEWICZ et al. 1999);

b) *Multiplicidade desconhecida* – a reconstrução de uma seqüência de DNA com o chip clássico C(*l*) fica mais complicada se, ao longo dela, existirem *l*-tuplas repetidas. Desde que cada sonda é única no chip de seqüenciamento, não há como capturar informações a respeito da ocorrência de repetições. Isso significa que não importa o número de vezes que uma *l*-tupla aparece ao longo da seqüência alvo, pois apenas uma sonda, complementar a ela, fica registrada no espectro. Para contornar essa questão, diversos trabalhos partem do pressuposto de que a multiplicidade das sondas é conhecida. Nesse caso é assumido que o espectro é um multiconjunto (e não mais um conjunto) em que as sondas podem aparecer várias vezes, refletindo as repetições existentes na seqüência (SHAMIR, 2001);

c) *Soluções ambíguas* – mesmo no caso ideal, com experimentos livres de erros e seqüências livres de repetições de tamanho *l*, pode ser difícil deduzir uma seqüência alvo a partir do seu espectro. Infelizmente, o grafo sugerido por Pevzner (1989) pode ter diversos Caminhos Eulerianos, cada um correspondendo à reconstrução de uma seqüência distinta (GUSFIELD, 1997). Nesse caso, cada seqüência reconstruída é claramente uma Mínima Supercadeia Comum para as sondas do espectro e, portanto, uma candidata à seqüência alvo. Assim, a menos que alguma informação adicional sobre a seqüência alvo seja obtida, a ambigüidade não pode ser resolvida, isto é, a solução correta não pode ser inequivocamente indicada

(DRMANAC et al., 2002). A próxima seção aborda o problema de soluções ambíguas de forma mais detalhada.

3.3.2.2 Soluções Ambíguas

Naturalmente, saber identificar a situação em que as soluções ambíguas ocorrem é a chave para tentar resolver a questão. Nos grafos, as soluções ambíguas se manifestam como ramificações, ou melhor, como vértices (do inglês *branching vertex*) de onde partem mais de um aresta (BEN-DOR et al., 1999). Por outro lado, nas seqüências, elas se manifestam como repetições de tamanho l - 1. Na realidade, essas são duas formas equivalentes de caracterizar a ocorrência de ambigüidade. Mas é importante notar que nem toda bifurcação no grafo conduz, necessariamente, a múltiplas seqüências.

Essa situação de ambigüidade pode ser observada no grafo ilustrado na Figura 18. Percebe-se que o vértice TG, com l - 1 bases, é um ponto de bifurcação que contribui para a formação de dois Caminhos Eulerianos. Um é o caminho que percorre as arestas {AT, TG}, {TG, GC}, {GC, CG}, {CG, GT}, {GT, TG}, {TG, GG}, {GG, GC}, {GC, CA} e o outro, o que percorre as arestas {AT, TG}, {GG, GC}, {GC, CG}, {CG, GT}, {GT, TG}, {TG, GC}, {GC, CA}. O primeiro corresponde à seqüência ATGCGTGGCA e o segundo, à seqüência ATGGCGTGCA.

Sobre soluções ambíguas, Pe'er e Shamir (2000) citam que, "a menos que o número de tais bifurcações seja muito pequeno, não existe uma maneira boa de determinar a seqüência verdadeira". É importante ressaltar que a quantidade de bifurcações no grafo de Pevzner

(1989) depende do tamanho da sonda e da própria composição da seqüência alvo (DRMANAC et al., 2002).

O problema de ambigüidade afeta de tal forma o desempenho da abordagem pelo Caminho Euleriano que, por exemplo, mesmo em experimento livre de erros, o tamanho máximo do fragmento de DNA que pode ser reconstruído sem ambigüidade a partir do chip $C(8)$, contendo todos as 65.536 sondas de tamanho 8, é de apenas 200 bases em 94% dos casos (PEVZNER et al, 1991).

Para caracterizar melhor o problema de ambigüidade, a Tabela 1 relaciona o desempenho do SBH, em termos do tamanho reconstruído em 90% dos casos, com o tamanho da sonda usada (DRMANAC et al., 2002 apud DRMANAC et al., 1989). Os valores apresentados referem-se a seqüências artificiais geradas randomicamente, ou seja, seqüências em que as bases são mutuamente independentes e são distribuídas igualmente com probabilidade de 1/4 cada uma.

Tabela 1
Tamanho da sonda *versus* tamanho da seqüência reconstruída
sem ambigüidade em 90% dos casos

Tamanho da sonda (*pb*)	Tamanho da seqüência (*pb*)
8	200
9	400
10	800
11	1600
12	3200

A análise das variáveis envolvidas no SBH revela o poder do método de Pevzner (1989). Partindo do princípio de que a condição suficiente para que não ocorra ambigüidade é

a inexistência de repetições de tamanho[3] l - 1, ficou provado que, para uma seqüência randômica, o tamanho máximo m que pode ser reconstruído sem ambigüidade com o chip clássico $C(l)$, contendo n sondas é (SHAMIR, 2002):

$$m = \sqrt{\frac{1}{2}n}$$

Ou, assintoticamente,

$$m = O(\sqrt{n})$$

Ou, ainda, dado que n é igual a 4^l para o chip $C(l)$,

$$m = O(2^l)$$

Esse resultado tem sido reafirmado tanto por estudos teóricos quanto por experimentos práticos. É fácil constatar que se trata de uma solução muito insatisfatória. Basta observar que, para sondas de tamanho l = 8, o valor máximo para m é de 256 bases, o que está bem abaixo das 700 bases normalmente obtidas com o seqüenciamento direto pelo método de Sanger.

Na tentativa de reduzir a ocorrência de soluções ambíguas, diversas estratégias alternativas têm sido propostas. Shamir e Tsur (2001) mencionam o uso de projetos de chips alternativos, de protocolos interativos, de informação de localização e de seqüências homólogas conhecidas. Especialmente, com relação a projetos de chips alternativos baseados em diferentes estruturas de sonda, um limite superior importante foi estabelecido.

[3] Pevzner (2000) considera as repetições de tamanho l, e não de tamanho l - 1, mas chega ao mesmo resultado: $m = O(2^l)$.

Considerando o espectro como um vetor binário, que registra para cada sonda do chip apenas a informação que diz se ela hibridizou ou não com o DNA, Preparata, Frieze e Upfal (1999) chegaram à conclusão de que o tamanho *m* de uma seqüência randômica que pode ser reconstruída sem ambigüidade, a partir de qualquer variação do chip clássico *C*(*l*) com *n* sondas, é:

$$m = O(n)$$

Ou, no caso especial de $n = 4^l$,

$$m = O(4^l)$$

Percebe-se que esse é um resultado bastante significativo. Por exemplo, para uma sonda de tamanho $l = 8$, uma seqüência randômica pode teoricamente atingir o tamanho de 65.536 bases. É um resultado que tem estimulado a busca por esquemas de chips mais eficientes. Em (PREPARATA; FRIEZE; UPFAL, 1999) foi proposto um novo esquema de chip de seqüenciamento que usa *bases universais* na construção das suas sondas, além das bases naturais – A, C, G e T –, sendo que esse tipo especial de base tem a característica importante de hibridizar com qualquer uma das outras quatro. Abandonando o uso de grafos, os autores apresentaram um algoritmo para esse novo modelo de chip que, com alta probabilidade, aproxima-se desse limite superior dentro de um fator constante.

Em publicação recente, um outro projeto de chip foi proposto por Halperin et al. (2002). Foi usado um esquema de sonda que, a exemplo do modelo anterior de Preparata, Frieze e Upfal (1999), também usa bases universais. Para esse novo modelo foi apresentado um algoritmo que, com alta probabilidade, reconstrói seqüências randômicas aproximando-se

do limite teórico dentro de um fator $\log^2$. Com relação à proposta de Preparata, Frieze e Upfal (1999), essa abordagem tem mais aplicabilidade prática, pois lida tanto com erros falsos positivos quanto com erros falsos negativos.

Com base nesse último modelo de chip e em outras idéias apresentadas na mesma publicação, esta dissertação também oferece uma abordagem alternativa para o problema SBH, baseada em um novo esquema de chip e em um algoritmo de reconstrução projetado para lidar com os dados gerados por ele.

4 AS PROPOSTAS DE HALPERIN ET AL.

Recentemente, Halperin et al. (2002) apresento duas novas abordagens para o SBH. A primeira consiste de um algoritmo de reconstrução, denominado algoritmo $\mathcal{A}$, que funciona em tempo polinomial sobre os dados gerados pelo chip clássico $C(l)$. A segunda inclui um esquema alternativo de chip de seqüenciamento que tem na sua estrutura, além das quatro bases naturais – A, C, G e T –, um tipo diferente de base, denominado *base universal*; e para lidar com os dados provenientes dele, inclui, ainda, um novo algoritmo polinomial, denominado algoritmo $\mathcal{B}$. Além dessas duas propostas, Halperin et al. (2002) ainda levantou a hipótese de que talvez o algoritmo $\mathcal{B}$ pudesse produzir melhores resultados práticos se lhe fosse atribuído certo aspecto do algoritmo $\mathcal{A}$.

As próximas seções descrevem, inicialmente, uma abordagem pioneira no uso de bases universais no SBH, apresentada por Preparata, Frieze e Upfal (1999), a qual, de certa forma, influenciou as proposta de Halperin et al. (2002). Em seguida, apresentam as duas soluções de Halperin et al. (2002), bem como um algoritmo que foi elaborado no curso desta pesquisa para funcionar de acordo com a hipótese por eles levantada, o qual, por extensão, foi denominado algoritmo $\mathcal{AB}$. Portanto, o objetivo deste capítulo é apresentar os aspectos relevantes das técnicas que serviram de ponto de partida para a elaboração do esquema

alternativo de chip e do algoritmo de reconstrução propostos como contribuição desta dissertação.

4.1 CONSIDERAÇÕES PRELIMINARES

Preliminarmente, algumas observações devem ser feitas a respeito do funcionamento das abordagens que serão apresentadas em seguida.

De modo geral, essas abordagens iniciam com uma subseqüência da seqüência alvo e tentam estendê-la, uma base por vez, até alcançar o seu final. Destaca-se que essa estratégia é diferente da estratégia de reconstrução baseada na construção de grafos e na posterior busca por Caminhos Hamiltonianos ou Eulerianos, como nos métodos anteriores. Mas, ao contrário disso, os algoritmos usados nessas abordagens manipulam os dados do problema na sua forma original, ou seja, como cadeias de caracteres. Para cada posição da seqüência que está sendo reconstruída, essas abordagens trabalham consultando o seu espectro e apurando o que um conjunto específico de sondas "diz" a respeito daquela posição. A forma de seleção desse conjunto e o modo de apuração são característicos de cada abordagem e acabam por determinar o seu poder de lidar com os problemas de multiplicidade de sondas, de ambigüidade e de erros no espectro.

Todas essas abordagens lidam com o problema de multiplicidade. Considerando que o espectro é consultado a cada nova posição a ser determinada e que ele nunca sofre alteração durante todo o processo de seqüenciamento, é suficiente que cada sonda apareça nele uma única vez, mesmo que possa hibridizar em mais de um local da seqüência.

Com exceção da abordagem de Preparata, Frieze e Upfal (1999), todas as demais podem lidar com certos níveis de erros falsos positivos e de erros falsos negativos no espectro. Na realidade, a sensibilidade das abordagens a erros está relacionada ao modo como as sondas são consideradas na determinação da base a estender.

No início do processo de seqüenciamento, todas as propostas assumem que o prefixo da seqüência alvo, cujo tamanho é igual ao tamanho da sonda usada menos um, é conhecido. É importante destacar que essa suposição foi feita originalmente no trabalho de Preparata, Frieze e Upfal (1999), que a justificaram apresentando duas soluções bioquímicas e uma algorítmica capazes de atender a esse requisito.

Para os algoritmos $\mathcal{A}$, $\mathcal{B}$ e $\mathcal{AB}$, deve-se assumir que o tamanho da seqüência a ser determinada é conhecido. A razão para isso está no fato de que, embora não esteja registrado nos algoritmos, esse é o parâmetro-chave que determina o instante em que os mesmos devem encerrar a execução. Essa também é uma suposição plausível, pois, de acordo com Setubal e Meidanis (1997), o tamanho da seqüência alvo pode ser determinado com uma margem de erro de 10%.

Antes de se prosseguir com a descrição das abordagens, é importante que as seguintes definições sejam apresentadas:

Definição 5 – *Diz-se que uma seqüência S' apóia uma seqüência S se ela é subseqüência de S.*

Definição 6 – *Admitindo-se que uma seqüência* $S = s_1,...,s_i$, *em reconstrução, seja estendida por uma seqüência* $a = a_1,...,a_j$, *denomina-se caminho de S a subseqüência* $s_{i-l+2},...,s_i,a_1,...,a_j$, *onde l é o tamanho das sondas usadas.*

4.2 PREPARATA ET AL. E AS BASES UNIVERSAIS

Em (PREPARATA; FRIEZE; UPFAL, 1999), os autores apresentaram uma nova solução para o SBH com a perspectiva de superar os resultados alcançados pelo chip clássico. Tal solução é baseada na construção de um modelo alternativo de chip de seqüenciamento e na proposta de um algoritmo que opera sobre o espectro gerado por ele. Os detalhes sobre essa abordagem serão descritos em seguida.

4.2.1 O Chip de Preparata et al.

O novo chip de Preparata, Frieze e Upfal (1999) caracteriza-se principalmente pela presença adicional de bases universais na estrutura das sondas, as quais, em face disso, são normalmente conhecidas como *gapped probes*. As bases universais podem ser imaginadas como bases curingas (normalmente denotadas por "*") que podem assumir o papel de qualquer uma das quatro bases naturais (A, C, G, e T). Vistas dessa forma, pode-se dizer que sempre colaboram com o processo de hibridização, agindo como se fossem o complemento da base a que se opõem. Mas, além disso, as sondas seguem um padrão periódico de formação bem definido: elas são formadas por um grupo de *x* bases naturais contíguas – *body* –, seguido por *y* bases naturais não contíguas, separadas umas das outras por *x* - 1 bases universais – *tail*. E dessa forma, cada uma possui um comprimento total de $l = x\ (y + 1)$ bases. Considerando $k = x + y$ o número de bases naturais ou especificadas, isto é, não-universais, o chip completo

de Preparata, Frieze e Upfal (1999) corresponde ao conjunto de todas as $n = 4^k$ sondas que podem ser geradas pela combinação das bases A, C, G e T de acordo com o padrão de formação escolhido. Formalmente, o chip de Preparata, Frieze e Upfal (1999) é definido como:

Definição 7 (PREPARATA; FRIEZE; UPFAL, 1999) – *Para os parâmetros dados x e y, o conjunto C(x,y) de sondas (x,y) consiste de todos as sondas da forma* $N^x(*^{x-1}N)^y$*, onde N é uma das quatro bases padrão de DNA (A, C, G e T),* $*$ *é a base universal, e cada expoente expressa o número de vezes que a base ou a seqüência de bases a ele associada se repete.*

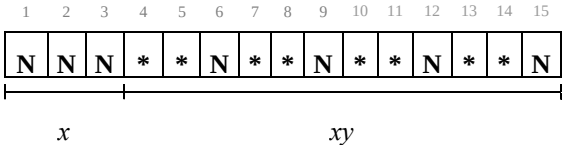


Figura 19: Esquema de uma sonda de Preparata, Frieze e Upfal (1999) com $x = 3$ e $y = 4$.

A Figura 19 ilustra uma sonda construída, segundo essa definição, para os parâmetros $x = 3$ e $y = 4$. Como exemplo, tem-se que a sonda ATG**T**G**A**T, que segue esse padrão, pode hibridizar tanto com a seqüência TACGGACTCAGTAGA como com a seqüência TACTCACTCTGTCAA, ou ainda com qualquer outra da forma TAC**A**C**T**A.

4.2.2 O Algoritmo de Preparata et al.

O algoritmo proposto para reconstruir a seqüência alvo, a partir do espectro gerado por esse modelo de chip, funciona como descrito a seguir. Considerando que uma seqüência S é conhecida desde a sua base inicial s_1 até a base s_i, para cada nova posição que se pretende

determinar, o algoritmo identifica o conjunto de sondas cujas as *l* - 1 bases iniciais apóiam as *l* - 1 bases finais da seqüência sendo reconstruída. O conjunto resultante poderá:

a) *ser vazio*: nesse caso, o algoritmo considera que o processo de seqüenciamento chegou ao fim;
b) *ter apenas uma sonda*: nesse caso, o algoritmo considera que a extensão procurada foi encontrada;
c) *ter mais de uma sonda*:
   c.1) *se todas as sondas sugerem a mesma base*: nesse caso a base procurada estará determinada;
   c.2) *se as sondas sugerem bases distintas*: se não há consenso, o algoritmo não pode identificar a base correta que estende a seqüência. Para tentar resolver a questão, ele considera um deslocamento de *s* bases à direita na seqüência alvo e reinicia o processo, identificando um novo conjunto formado por aquelas sondas cujas *l* - *d* - 1 primeiras bases apóiam as *l* - *d* - 1 últimas bases da seqüência alvo, onde *d* é a quantidade de bases deslocadas desde o início da execução. Caso a ambigüidade ainda persista, o algoritmo apela para o próximo deslocamento à direita do atual, e assim sucessivamente, até atingir o número máximo *y* de deslocamentos.

É importante destacar o modo pelo qual o conjunto de sondas que confirma a posição a estender é avaliado: na verdade procura-se sempre chegar ao consenso, isto é, a uma situação em que, após um número necessário de deslocamentos, todos as sondas indiquem uma mesma base. Para essa abordagem, ficou provado que o tamanho *m* de uma seqüência escolhida

randomicamente que pode ser reconstruída com alta probabilidade a partir de um chip com n sondas é dado por $m = \Theta(n)$ (PREPARATA; FRIEZE; UPFAL, 1999).

A Figura 20 esquematiza o funcionamento desse algoritmo sobre dados obtidos a partir de um chip $C(3,3)$ para a reconstrução da uma seqüência S que já está com as suas 14 primeiras bases determinadas. Para identificar a posição seguinte, foi identificado, inicialmente, um conjunto com duas sondas cujas 11 primeiras bases apóiam as 11 últimas bases da seqüência – uma indicando a base T e a outra, a base G para essa posição. Nesse caso, como não houve consenso, o algoritmo considera um deslocamento de $s = 3$ bases à direita e com relação a ele, identifica um outro conjunto com três sondas cujas 8 primeiras bases apóiam as 8 últimas bases da seqüência. Como, novamente, não houve consenso, o algoritmo considera um novo deslocamento à direita. Dessa vez, como apenas uma sonda é encontrada, a extensão da seqüência fica determinada pela base dessa sonda que está relacionada à posição 15, ou seja, pela base T.

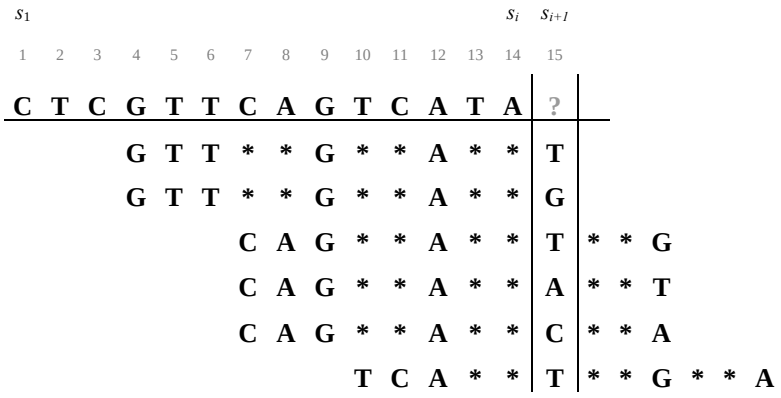


Figura 20: Exemplo esquemático de execução do algoritmo de Preparata, Frieze e Upfal (1999) sobre dados obtidos com o chip $C(3,3)$.

4.2.3 Aspectos Relevantes

Da análise dessa proposta, pode-se destacar as seguintes características relevantes:

a) o aspecto construtivo do chip faz com que, em um experimento livre de erros, existam no espectro *x* sondas indicando corretamente a posição a estender, obviamente, cada uma associada a um deslocamento diferente. Na realidade, ocasionalmente, podem existir outras sondas, mas elas terão hibridizado com outras localizações da seqüência alvo. Essa é uma forma de lidar com a ambigüidade, pois, enquanto for possível, recorre-se a um novo deslocamento para decidir a base correta (SHAMIR, 2002);

b) apesar de existirem essas *x* sondas, a forma de funcionamento do algoritmo, baseada na avaliação de um deslocamento por vez, acaba comprometendo a resistência da abordagem a erros, pois faz com que, em última instância, a base a estender seja determinada por uma única sonda. Isso aumenta a probabilidade de que, no caso de falsos negativos, o algoritmo reconstrua a seqüência incorretamente, e, no caso de falsos positivos, ele tenha de recorrer ao deslocamento seguinte;

c) o fato de haver apenas uma sonda correta em cada deslocamento está relacionado ao padrão rígido de posicionamento das bases universais no chip;

d) o uso de bases universais praticamente desvincula as sondas de um determinado deslocamento das sondas do deslocamento anterior, reduzindo a probabilidade de que

uma situação de ambigüidade se propague (SHAMIR, 2002). Na realidade, essa característica faz com que o mecanismo de deslocamento funcione adequadamente.

4.3 HALPERIN ET AL. E O CHIP CLÁSSICO

4.3.1 O Algoritmo $\mathcal{A}$ para o Chip Clássico

A primeira solução apresentada por Halperin et al. (2002) para o SBH consiste de um algoritmo bastante trivial, denominado algoritmo $\mathcal{A}$[4], o qual opera sobre os dados produzidos pelo chip clássico $C(l)$ ou, considerando que nele a quantidade k de bases naturais presentes em cada sonda determina o próprio tamanho l da sonda, $C(k)$.

Considerando que a seqüência alvo S é conhecida desde a base inicial s_1 até a base s_i, o algoritmo $\mathcal{A}$ tem de executar os seguintes passos para determinar a base s_{i+1}:

1. gerar todas as 4^k seqüências $a = a_1,...,a_k$ como extensão de $s_1,...,s_i$;

2. para cada seqüência a', contar o número de sondas do espectro que apóiam a subseqüência $s_{i-k+2},...,s_i, a_1',...,a_k'$ ou *caminho*;

3. para a seqüência vencedora, fazer $s_{i+1} = a_1'$ (o empate é resolvido arbitrariamente).

---

[4] Destaca-se que esse algoritmo foi apresentado como solução para o problema de número 35 da lista de problemas em aberto na biologia elaborada por Pevzner e Waterman (1995).

Para uma seqüência com m bases, o tempo de execução desse algoritmo é dado por $O(4^k km)$. Esse resultado reflete o fato de que, em cada posição dessa seqüência, devem ser testadas k sondas para cada uma da 4^k extensões geradas.

A Figura 21 esquematiza o funcionamento do algoritmo $\mathcal{A}$ sobre dados obtidos a partir do chip $C(8)$ para a reconstrução da posição 9 da seqüência S. De acordo com o algoritmo, deve-se gerar todas as extensões possíveis no trecho entre as posições 9 e 16, e contar, para cada uma delas, o número de sondas que apóiam o caminho que vai da posição 2 até a posição 16. Tem-se que nesse exemplo existem 8 sondas para a extensão GAGTTAGT, e, caso nenhuma outra extensão supere esse escore, o algoritmo irá estender a seqüência com a base G.

s_1							s_i	a'_1							a'_k
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
C	**T**	**C**	**A**	**C**	**G**	**T**	**T**	**G**	**A**	**G**	**T**	**T**	**A**	**G**	**T**
	T	**C**	**A**	**C**	**G**	**T**	**T**	**G**							
		C	**A**	**C**	**G**	**T**	**T**	**G**	**A**						
			A	**C**	**G**	**T**	**T**	**G**	**A**	**G**					
				C	**G**	**T**	**T**	**G**	**A**	**G**	**T**				
					G	**T**	**T**	**G**	**A**	**G**	**T**	**T**			
						T	**T**	**G**	**A**	**G**	**T**	**T**	**A**		
							T	**G**	**A**	**G**	**T**	**T**	**A**	**G**	
								G	**A**	**G**	**T**	**T**	**A**	**G**	**T**

Figura 21: Exemplo esquemático de execução do algoritmo $\mathcal{A}$ sobre dados obtidos com o chip clássico $C(8)$.

4.3.2 Aspectos Relevantes

Da análise dessa proposta, pode-se destacar as seguintes características relevantes:

a) à semelhança do que foi visto na abordagem de Preparata, Frieze e Upfal (1999), os aspectos construtivos do chip clássico fazem com que, em caso de experimento livre de erros, existam no espectro k sondas determinando corretamente a posição a estender, obviamente, cada uma associada a um deslocamento diferente. Na realidade, ocasionalmente, podem existir outras sondas, mas elas terão hibridizado com outras localizações da seqüência alvo. Vale ressaltar que, no caso do chip clássico, o deslocamento é de apenas uma base;

b) enquanto a abordagem de Preparata, Frieze e Upfal (1999) avalia as sondas, deslocamento por deslocamento, a solução de Halperin et al. (2002) apura de uma única vez todos as sondas que hibridizam em um caminho com a seqüência alvo, fazendo, assim, com que essa abordagem seja menos sensível a erros. O benefício se verifica tanto com relação aos erros falsos positivos como com relação aos falsos negativos, de modo que, mesmo que um erro de laboratório deixe de fora do espectro determinada sonda que tenha hibridizado com a seqüência, ainda existirão outras sondas que serão usadas pelo algoritmo para determinar a extensão procurada. Por outro lado, se existirem sondas no espectro indicando incorretamente uma base, a tendência é que o efeito delas seja minimizado em face da quantidade de sondas que apóiam a base correta. Efetivamente, o algoritmo irá falhar somente quando o número de sondas que apóiam o caminho correto for menor que o número de sondas que apóiam algum caminho errado;

c) da mesma forma que o chip de Preparata, Frieze e Upfal (1999), o chip clássico gera apenas uma sonda para cada posição da seqüência alvo, entretanto, em decorrência

dessa forma de apuração das sondas que informam algo sobre a posição a estender, esse aspecto não compromete tanto o poder do algoritmo de lidar com erros;

d) a idéia de caminhos permite que o algoritmo “olhe” para a frente da seqüência para determinar o valor da posição atual, ou melhor, que ele se apóie também na informação de sondas que correspondem a posições que ainda serão, no futuro, determinadas por ele. Na realidade, isso equivale a simular o estado da seqüência algumas posições à frente, para tentar descobrir se determinada escolha feita será posteriormente confirmada;

e) o primeiro passo do algoritmo consiste em gerar todas as extensões possíveis de tamanho *k*. Considerando que *k* coincide com o tamanho da sondas usadas, o custo dessa tarefa certamente é o maior obstáculo para essa abordagem.

4.4 HALPERIN ET AL. E O NOVO CHIP

Além do algoritmo $\mathcal{A}$, elaborado para lidar com os dados obtidos com o chip clássico, Halperin et al. (2002) propôs uma nova abordagem para o problema SBH, baseada em um projeto de chip alternativo e na elaboração de um novo algoritmo, denominado algoritmo $\mathcal{B}$, capaz de lidar com o espectro gerado por ele.

4.4.1 O Novo Chip de Halperin et al.

A exemplo do projeto de Preparata, Frieze e Upfal (1999), o projeto de chip desenvolvido por Halperin et al. (2002) também emprega bases universais na estrutura das sondas além das bases naturais. Por outro lado, uma diferença importante entre esses dois

projetos é que, enquanto as sondas de Preparata, Frieze e Upfal (1999) possuem uma estrutura periódica determinística, as sondas projetadas por Halperin et al. (2002) possuem as bases naturais distribuídas randomicamente entre as bases universais.

Para o novo esquema de sonda proposto, ficou demonstrado (HALPERIN et al., 2002) que, com um chip contendo $n = \Theta(k4^k)$ sondas, uma seqüência de tamanho $m \approx (\frac{4^k}{k})$ pode ser, com alta probabilidade, determinada sem ambigüidade, na presença de erros. No caso de $m = \alpha \frac{4^k}{\beta k}$, para alguma constante α e alguma constante β, que depende das taxas de erro, o chip pode ser descrito da seguinte forma:

a) cada uma das sondas constituintes é composta de duas partes: uma parte inicial com *k* bases naturais posicionadas aleatoriamente e separadas por bases universais com um fator *c* de distanciamento, onde *c* é um inteiro convenientemente grande; e uma parte final correspondendo a uma única base natural. Por conseqüência, o tamanho total de cada sonda é dado por *ck* + 1. A Figura 22 ilustra uma sonda construída segundo a forma descrita acima para *c* = 2 e *k* = 4;

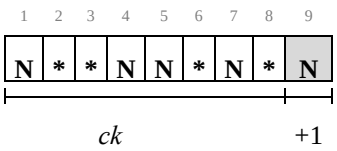


Figura 22: Exemplo de uma sonda do chip de Halperin et al. (2002) com *c* = 2 e *k* = 4.

b) no total, ele possui βk famílias de sondas. Para cada família, é selecionado randomicamente um conjunto de *k* posições no intervalo de 1 a *ck* e são formadas

todas as 4^{k+1} sondas possíveis, colocando as bases A, C, G e T nas k posições especificadas e na última. Dessa forma, o número total de sondas no chip é dado por $\beta k 4^{k+1}$. A Figura 23 ilustra um chip construído segundo a forma descrita acima. Observa-se que nele existem 3 famílias de sondas com $c = 2$ e $k = 4$.

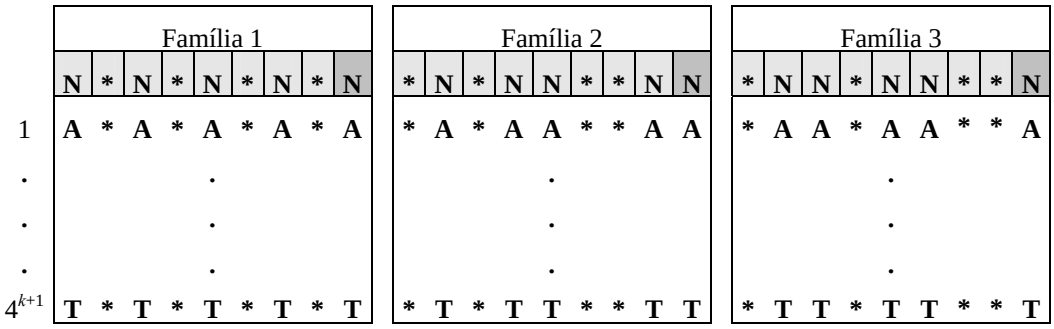


Figura 23: Exemplo de um chip de Halperin et al. (2002) com 3 famílias de sondas com $c = 2$, $k = 4$. Cada família é formada por todas as combinações possíveis das bases A, C, G e T nas posições especificadas.

4.4.2 O Algoritmo $\mathcal{B}$ para o Novo Chip

O algoritmo $\mathcal{B}$ foi proposto para ser aplicado sobre os dados do espectro obtido com o chip especificado acima. Considerando que uma seqüência alvo S é conhecida desde a sua primeira base s_1 até a base s_i, esse algoritmo tem de executar os seguinte passos para determinar a base s_{i+1}:

1. para cada base natural X, contar quantas sondas apóiam a subseqüência $s_{i-ck+1},...,s_i, X$;

2. para a base vencedora, fazer $s_{i+1} = X$ (o empate é resolvido arbitrariamente).

Para uma seqüência com *m* bases, o tempo de execução desse algoritmo é dado por $O(km)$. Esse resultado reflete o fato de que, em cada posição dessa seqüência, devem ser testadas *k* sondas para cada uma das 4 extensões geradas.

A Figura 24 esquematiza o funcionamento do algoritmo $\mathcal{B}$ sobre dados obtidos a partir de um chip com $\beta k = 3$, $c = 2$ e $k = 3$ para a reconstrução da posição 9 da seqüência *S*. De acordo com o algoritmo, deve-se atribuir as quatro bases naturais a essa posição e contar, para cada caso, quantas sondas apóiam a subseqüência que vai da posição 3 até a posição 9. Tem-se, nesse exemplo, que existem 3 sondas indicando a base G, sendo que o algoritmo irá usá-la para estender a seqüência, caso nenhuma outra base supere esse escore.

s_1							s_i	X
1	2	3	4	5	6	7	8	9
C	**T**	**C**	**A**	**C**	**G**	**T**	**T**	**G**
		C	*****	*****	**G**	*****	**T**	**G**
		*****	**A**	*****	**G**	**T**	*****	**G**
		C	*****	**C**	*****	*****	**T**	**G**

Figura 24: Exemplo esquemático de execução do algoritmo $\mathcal{B}$ sobre dados obtidos com o chip de Halperin et al. (2002) com $c = 2$ e $k = 3$.

4.4.3 Aspectos Relevantes

Da análise dessa proposta, pode-se destacar as seguintes características relevantes:

a) a principal característica dessa técnica consiste em se usar um chip dividido em famílias de sondas, conseqüência do uso de bases universais posicionadas aleatoriamente. A vantagem disso é que agora não se tem no espectro somente uma sonda hibridizando com cada subseqüência da cadeia alvo, mas, sim, um número de

sondas que é igual à quantidade de famílias existentes. Como discutido anteriormente, ter um número maior de sondas apoiando a escolha da base a estender é um fator importante que ajuda a minimizar os efeitos dos erros experimentais;

b) enquanto os algoritmos anteriores usavam sondas deslocadas à direita para confirmar a posição atual, o algoritmo $\mathcal{B}$ considera apenas as sondas cujas *ck* bases iniciais apóiam a subseqüência final da cadeia em construção. Com relação a essa questão, Halperin et al. (2002) fez uma afirmação importante que serviu de ponto de partida para este trabalho de pesquisa:

> Uma extensão natural do algoritmo $\mathcal{B}$ é um algoritmo similar ao algoritmo $\mathcal{A}$, isto é, ao contrário de se contar o número de *supporting probes*[5] para somente quatro extensões possíveis, nós poderíamos contar o número de *supporting probes* de um conjunto de caminhos possíveis que estendem a seqüência atualmente sendo reconstruída. É possível que esse algoritmo produza melhores resultados na prática que o algoritmo $\mathcal{B}$ (HALPERIN et al., 2002).

O próximo item apresenta um algoritmo que funciona segundo a forma sugerida.

4.5 A NOVA PROPOSTA DE HALPERIN ET AL.

4.5.1 O Algoritmo $\mathcal{AB}$

O início deste trabalho de pesquisa foi dedicado justamente à elaboração de um algoritmo que funcionasse de acordo com a sugestão de Halperin et al. (2002) de introduzir no seu algoritmo $\mathcal{B}$ o mecanismo que garante que o algoritmo $\mathcal{A}$ use todas as sondas do espectro

[5] Um *probe* (sonda, em português) é um *supporting probe* de uma seqüência se ele apóia a seqüência e se aparece no seu espectro (HALPERIN et al., 2002).

que dispõem de informação sobre a posição a estender. De outra forma, o que está por trás da idéia de Halperin et al. (2002) é o interesse de que o novo algoritmo contemple o conceito de caminho, ou melhor, de que apure o "valor" dessa posição com base nas sondas que o apóiam. É importante frisar que a sugestão proposta prevê alteração apenas no algoritmo, e não na estrutura do chip, o qual deve permanecer como descrito no item 4.4.1.

O algoritmo resultante está descrito abaixo, sendo que para ele também valem as considerações preliminares a respeito do conhecimento prévio do prefixo da seqüência alvo, bem como do seu comprimento total.

Considerando que uma seqüência *S* é conhecida desde a base inicial s_1 até a base s_i, o algoritmo $\mathcal{AB}$ tem de executar os seguintes passos para determinar a base s_{i+1}:

1. gerar todas as 4^{ck+1} seqüências $a = a_1,...,a_{ck+1}$ como extensão de $s_1,...,s_i$;

2. para cada seqüência a', contar o número de sondas do espectro que apóiam a subseqüência $s_{i-ck+1},...,s_i, a_1',...,a_{ck+1}'$ ou *caminho*;

3. para a seqüência vencedora, fazer $s_{i+1} = a_1'$ (o empate é resolvido arbitrariamente).

O tempo de execução desse algoritmo é dado por $O(4^{ck}k^2cm)$. Esse tempo expressa o fato de que, para cada uma das *m* posições da seqüência alvo, são consideradas todas as 4^{ck+1} extensões de tamanho $ck + 1$, sendo que, para cada uma de suas bases, é testada uma sonda de cada família.

A Figura 25 esquematiza o funcionamento do algoritmo $\mathcal{AB}$ sobre dados obtidos a partir de um chip com $\beta k = 2$, $c = 2$ e $k = 2$ para a reconstrução da posição 7 de uma seqüência S. De acordo com o primeiro passo do algoritmo, deve-se gerar todas as extensões possíveis no trecho que vai da posição 7 até a posição 11, e contar, para cada uma delas, o número de sondas que apóiam o caminho, que vai da posição 3 até a posição 11. Tem-se que existem 5 grupos de duas sondas apoiando a extensão TTGAC, sendo que 6 delas indicam a base T para a posição 7.

s_1					s_i	a_1				a_{ck+1}
1	2	3	4	5	6	7	8	9	10	11
C	**T**	**C**	**A**	**C**	**G**	**T**	**T**	**G**	**A**	**C**
		*	**A**	*	**G**	**T**				
		C	*	**C**	*	**T**				
			*	**C**	*	**T**	**T**			
			A	*	**G**	*	**T**			
				*	**G**	*	**T**	**G**		
				C	*	**T**	*	**G**		
					*	**T**	*	**G**	**A**	
					G	*	**T**	*	**A**	
						*	**T**	*	**A**	**C**
						T	*	**G**	*	**C**

Figura 25: Exemplo esquemático de execução do algoritmo $\mathcal{AB}$ sobre dados obtidos com o chip de Halperin et al. (2002), com $\beta k = 2$, $c = 2$ e $k = 2$.

4.5.2 Aspectos Relevantes

Da análise dessa proposta, pode-se destacar as seguintes características relevantes:

a) como fora planejado, o algoritmo $\mathcal{AB}$ herdou do algoritmo $\mathcal{A}$ a característica de usar o conceito de caminho na apuração da base a estender;

b) da mesma forma, herdou do algoritmo $\mathcal{B}$ a capacidade de trabalhar com chips estruturados em famílias de sondas, característica alcançada pelo uso de sondas contendo bases universais distribuídas randomicamente;

c) o grande obstáculo para o funcionamento desse algoritmo está no fato de que, para cada posição da seqüência alvo, devem ser geradas todas as extensões possíveis de tamanho $ck + 1$. Para se ter uma idéia mais concreta do problema, deve-se saber que Halperin et al. (2002) sugere os valores 4 e 6 como razoáveis para c e k, respectivamente, o que implica em um tamanho de 25 bases para a sonda, e, por conseguinte, para a extensão. Na realidade, esse problema já existia no algoritmo $\mathcal{A}$. A diferença é que no algoritmo $\mathcal{AB}$ ele se agrava, visto que, normalmente, esse algoritmo trabalha com sondas muito maiores;

d) outra questão que se observa na implementação do algoritmo $\mathcal{AB}$ é que nem todas as sondas que apóiam um determinado caminho indicam explicitamente uma base para a posição a estender. É o caso das sondas que têm uma base universal associada à tal posição. Esse problema realmente não afeta a resposta do algoritmo. Tentar evitá-lo é apenas uma questão de eliminar um esforço computacional desnecessário.

5 UMA NOVA ABORDAGEM PARA O SBH

Este capítulo tem por objetivo apresentar uma nova abordagem para o problema SBH, a qual é baseada na construção de um novo chip de seqüenciamento e na elaboração de um algoritmo de reconstrução correspondente que, atuando sobre os dados gerados por ele, tenta reconstruir a seqüência alvo. Adicionalmente, apresentam-se e analisam-se, inclusive comparativamente com o desempenho do algoritmo $\mathcal{B}$, os resultados obtidos da sua avaliação experimental. Destaca-se que, como parte deste trabalho de pesquisa, essa nova proposta e alguns dos seus resultados foram, preliminarmente, apresentados em artigo publicado sob o título *Uma Abordagem Alternativa para Seqüenciamento por Hibridização* nos anais do XXIII Congresso da Sociedade Brasileira de Computação – XXX SEMISH (BAPTISTA; GUIMARÃES, 2003). Esse artigo compõe o Apêndice B.

5.1 CONSIDERAÇÕES PRELIMINARES

É importante frisar que, assim como no caso do algoritmo $\mathcal{AB}$, o desenvolvimento dessa nova proposta teve origem na sugestão de Halperin et al. (2002) de aprimorar o algoritmo $\mathcal{B}$ pela adição de uma característica particular do algoritmo $\mathcal{A}$. Na realidade, procurou-se elaborar uma estratégia de seqüenciamento que contemplasse os aspectos positivos do algoritmo $\mathcal{AB}$ e ao mesmo tempo contornasse as suas restrições.

Especificamente, o objetivo foi chegar a uma estratégia que, assegurando o uso de um conjunto de sondas para apoiar cada subseqüência da seqüência alvo e o uso de sondas relacionadas a posições imediatamente à frente da que está sendo reconstruída, evitasse o problema não somente de gerar grandes extensões como também de considerar sondas que não acrescentam nenhuma informação que auxilie o processo de reconstrução.

A elaboração dessa nova abordagem foi precedida do estudo de todas as soluções apontadas por Halperin et al. (2002) para o SBH. Nesse estudo, procurou-se principalmente identificar os fatores associados às características que se pretendia incluir na nova proposta, bem como os associados às que deveriam ser evitadas. Efetivamente, esse estudo preliminar evidenciou que:

a) o mecanismo que permite o uso de sondas deslocadas, umas em relação às outras, na determinação da base que deve estender a seqüência, é reflexo da característica peculiar de construção do chip de seqüenciamento, a qual garante que sondas relacionadas às posições consecutivas na seqüência alvo se sobreponham por l - 1 bases, onde l corresponde ao tamanho da sonda usada. Entretanto, observa-se que nem todas as sondas deslocadas, que guardam relação com a posição a estender, realmente indicam a base correta para ela. No caso do chip clássico, isso realmente ocorre. Entretanto, para os chips propostos por Preparata, Frieze e Upfal (1999) e por Halperin et al. (2002), não. No caso do chip de Preparata, Frieze e Upfal (1999), somente as sondas deslocadas por uma quantidade múltipla de s são consideradas, onde s é a quantidade de bases universais iniciais da sonda. Para o chip de Halperin et al. (2002), as sondas a serem consideradas dependem da distribuição aleatória das bases naturais;

b) no caso dos algoritmos $\mathcal{B}$ e $\mathcal{AB}$, o uso de bases universais distribuídas aleatoriamente possibilita a existência de múltiplas famílias de sondas no chip, o que, em última instância, significa a existência de múltiplas sondas no espectro para cada posição da seqüência alvo;

c) com respeito ao tamanho da extensão a ser gerada pelos algoritmos $\mathcal{A}$ e $\mathcal{AB}$, observa-se que o mesmo corresponde ao tamanho da própria sonda que está sendo usada, fazendo com que todas as sondas relacionadas com a posição a estender sejam consideradas. Entretanto, observa-se que nada impede que extensões menores sejam usadas. Seria necessário apenas modificar o algoritmo para isso. Mas, nesse caso, é intuitivo que os resultados obtidos seriam inferiores, devido ao fato de se estar usando uma menor quantidade de informação;

d) por usarem extensões do tamanho da própria sonda, os algoritmos $\mathcal{A}$ e $\mathcal{AB}$ acabam por considerar, na determinação da base a estender, todas as sondas do espectro que guardam vínculo com a posição a estender. No caso do algoritmo $\mathcal{AB}$, são consideradas inclusive aquelas que têm uma base universal associada à essa posição, e, portanto, não podem ajudar a defini-la. Entretanto, percebeu-se que no caso da extensão ser reduzida, como mencionado na alínea acima, isso não ocorreria se, ocasionalmente, os chips fossem construídos com bases naturais ocupando as *e* posições finais de cada sonda, onde *e* corresponde ao tamanho adotado para essa extensão.

Diante dessas observações, percebeu-se que, para atingir os objetivos traçados, a nova estratégia para o problema SBH deveria incluir, obrigatoriamente, um novo projeto de chip e

um novo algoritmo de reconstrução. Os dois itens seguintes descrevem o chip e o algoritmo propostos como contribuição principal desse trabalho de pesquisa.

5.2 UM NOVO ESQUEMA DE CHIP

O chip proposto deve ser construído da seguinte forma:

a) cada uma de suas sondas é composta de duas partes: uma parte inicial com *k* bases naturais (denotadas por N) posicionadas aleatoriamente e separadas por bases universais com um fator *c* de distanciamento, onde *c* é um inteiro convenientemente grande; e uma parte final composta de *e* bases naturais contíguas. Por conseqüência, o tamanho total de cada sonda é dado por *ck* + *e*. A Figura 26 ilustra um sonda construída segundo a descrição acima para $c = 2$, $k = 3$ e $e = 3$;

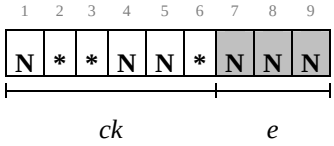


Figura 26. Exemplo de uma sonda do novo chip com $c = 2$, $k = 3$ e $e = 3$.

b) no total, o chip possui *f* famílias de sondas. Para cada família, é selecionado randomicamente um conjunto de *k* posições no intervalo de 1 a *ck* e são formadas todas as 4^{k+e} sondas possíveis, colocando as bases A, C, G e T nas *k* posições selecionadas e nas *e* últimas. Dessa forma, o número total *n* de sondas no chip é dado por $f4^{k+e}$. A Figura 27 ilustra um chip construído segundo a descrição acima para $f = 3$, $c = 2$, $k = 3$ e $e = 3$.

Deve ser observado que a diferença básica desse esquema de chip para aquele proposto por Halperin et al. (2002) é que, enquanto o primeiro garante *e* bases naturais no final da sonda, o segundo garante apenas uma. Deve-se ressaltar, entretanto, que, para o caso especial de $e = 1$, os chips são idênticos.

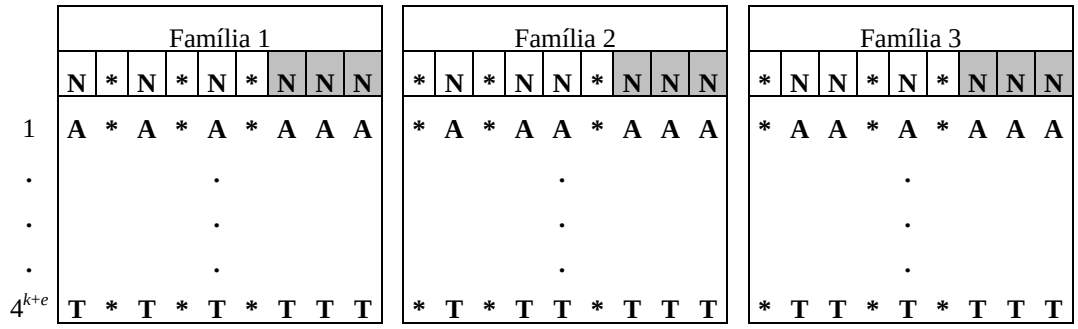


Figura 27: Exemplo do novo chip com 3 famílias de sondas com $c = 2$, $k = 3$ e $e = 3$. Cada família é formada por todas as combinações possíveis das bases A, C, G e T nas posições especificadas.

5.3 O ALGORITMO $\mathcal{AB}$-ESTENDIDO

De modo resumido, o que se deve esperar do procedimento de seqüenciamento pelo método SBH é que o chip forneça informação adequada sobre a seqüência alvo, e que o algoritmo de reconstrução seja eficiente em manipulá-la. A segunda etapa dessa nova abordagem consiste exatamente na elaboração de um algoritmo que opere sobre os dados gerados pelo novo esquema de chip proposto, denominado algoritmo $\mathcal{AB}$-estendido.

Devido à própria origem da nova proposta e aos objetivos que se buscou alcançar com ela, esse algoritmo pode ser visto como extensão dos algoritmos $\mathcal{A}$ e $\mathcal{B}$. Inclusive continuam valendo para ele as considerações feitas sobre o conhecimento prévio de um prefixo da seqüência alvo e do número de suas bases constituintes.

Considerando que uma seqüência S é conhecida desde a base inicial s_1 até a base s_i, o algoritmo $\mathcal{AB}$-estendido executa os seguintes passos sobre o espectro obtido com o novo chip para determinar a base s_{i+1}:

1. gerar todas as 4^e seqüências $a = a_1,...,a_e$ como extensão de $s_1,...,s_i$;

2. para cada seqüência a', contar o número de sondas do espectro que casam com a subseqüência $s_{i-ck-e+2},...,s_i, a_1',...,a_e'$ ou *caminho*;

3. para a seqüência vencedora, fazer $s_{i+1} = a_1'$ (o empate é resolvido arbitrariamente).

Deve-se notar que esse algoritmo é bastante similar ao algoritmo $\mathcal{AB}$ original, sendo que a principal diferença está no tamanho da extensão que é gerada por cada um. Enquanto no algoritmo $\mathcal{AB}$ o tamanho da extensão é o próprio tamanho da sonda, dado por $l = ck + 1$, no $\mathcal{AB}$-estendido, o tamanho da extensão é dado por e.

Tabela 2
Tempo de execução dos algoritmos de reconstrução

Algoritmo	Tempo de execução
$\mathcal{A}$	$O(4^k km)$
$\mathcal{B}$	$O(km)$
$\mathcal{AB}$	$O(4^{ck} k^2 cm)$
$\mathcal{AB}$-estendido	$O(4^e efm)$

O tempo de execução desse algoritmo é dado por $O(4^e efm)$. Esse tempo expressa o fato de que, para cada uma das m posições da seqüência alvo, são consideradas todas as 4^e

extensões de tamanho *e*, sendo que para cada uma de suas bases é testada uma sonda de cada família. Para fins comparativos, esse resultado é apresentado na Tabela 2, juntamente com as expressões relativas aos algoritmos sugeridos em Halperin et al. (2002).

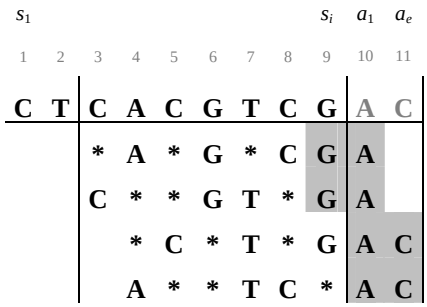


Figura 28. Exemplo esquemático de execução do algoritmo $\mathcal{AB}$-estendido sobre dados obtidos com o novo chip com $f = 2$, $c = 2$, $k = 3$ e $e = 2$.

A Figura 28 esquematiza o funcionamento do algoritmo $\mathcal{AB}$-estendido para dados obtidos por um chip com $f = 2$, $k = 3$, $c = 2$ e $e = 2$, no momento da reconstrução da posição 10 de uma seqüência *S*. De acordo com o algoritmo, devem-se gerar todas as extensões possíveis de tamanho *e* para o trecho que vai da posição 10 até a posição 11 e, então, avaliar, para cada uma delas, a quantidade de sondas que casam com o caminho que vai da posição 3 até a posição 11. Tem-se que, para a extensão AC, existem 4 sondas confirmando a base A para a posição 10.

5.4 ASPECTOS RELEVANTES

A nova proposta conseguiu reunir as seguintes características positivas dos algoritmos $\mathcal{A}$ e $\mathcal{B}$:

a) o aspecto mais interessante dessa nova abordagem talvez seja o fato da mesma

poder ser vista como uma solução de caráter mais geral, na qual a solução proposta por Halperin et al. (2002) – correspondente ao algoritmo $\mathcal{B}$ – é apenas uma de suas instâncias, mais especificamente, corresponde ao caso em que $e = 1$;

b) ao usar bases universais distribuídas randomicamente ao longo da sonda, o esquema do novo chip possibilita o uso de famílias de sondas;

c) o algoritmo usa o conceito de caminho e dessa forma apóia-se nas informações de sondas relacionadas a posições que ainda serão, no futuro, determinadas por ele.

É verdade que o algoritmo $\mathcal{AB}$ já guardava essas características positivas, porém, a nova abordagem, em relação a ele, avançou nas seguintes direções:

a) como o tamanho da extensão não está vinculado ao tamanho da sonda, mas, sim, ao valor definido para e durante a construção do chip, existe a oportunidade de minimização do problema prático de gerar grandes extensões: basta apenas que seja escolhido um valor “pequeno” para e;

b) eliminou-se o problema de haver várias sondas apoiando um caminho sem, no entanto, esclarecer nada sobre a posição a estender. Isso pode ser visto na Figura 28, onde todas as sondas indicam a base A para a posição 10;

c) a dificuldade desse algoritmo está no fato de se ter de gerar repetidas vezes todas as extensões possíveis de tamanho e. Como visto, esse problema já estava presente tanto no algoritmo $\mathcal{A}$ quanto no algoritmo $\mathcal{AB}$. O fato de esse novo algoritmo usar extensões de tamanho e, ao invés de k e de $ck + 1$, como naqueles dois algoritmos,

assintoticamente, não altera a situação, mas ainda assim é possível que, com relação ao algoritmo $\mathcal{B}$, ele produza resultados práticos superiores. É exatamente isso que se esperava comprovar com os testes realizados.

5.5 TESTES E RESULTADOS

A fim de avaliar o desempenho dessa nova estratégia para o SBH, diversos testes foram realizados com o chip e o algoritmo alternativos propostos. Através desses testes, buscou-se revelar os principais efeitos das variáveis envolvidas no problema e ainda fazer uma análise comparativa com a proposta apresentada por Halperin et al. (2002).

É importante destacar que normalmente a comparação de estratégias que envolvem tipos e quantidades de parâmetros diferentes não é uma tarefa trivial, fazendo com que os resultados obtidos venham cercados de considerações. Nesse caso, a comparação foi um pouco mais simples, visto que, na realidade, não se trata de abordagens diferentes, pois, como foi citado anteriormente, a solução de Halperin et al. (2002) pode ser vista como um caso específico da solução proposta por esta pesquisa.

Os teste foram baseados na simulação computacional de todo o processo de SBH, incluindo tanto a fase experimental de hibridização, com vistas à geração do espectro, quanto a fase combinatorial, com vistas à reconstrução propriamente dita da seqüência alvo. Entretanto, vale ressaltar que os resultados foram apreciados como produto do processo SBH como um todo. Os próximos itens descrevem os testes realizados e discutem os resultados obtidos.

5.5.1 Simulação

De modo geral, o programa de simulação desenvolvido trata da reconstrução de uma única seqüência alvo a partir dos dados gerados por um chip de seqüenciamento especificado. O programa completo inclui, além das rotinas que implementam a duas etapas básicas do SBH – a bioquímica e a combinatorial –, rotinas responsáveis pela preparação da seqüência de DNA a serem usadas nos testes, pela simulação de erros no espectro e pela verificação do sucesso do processo.

O programa foi codificado na linguagem C e os testes foram executados em uma máquina SparcStation20 com sistema operacional Linux.

Antes de detalhar o programa de simulação é importante apresentar as principais variáveis envolvidas na solução alternativa proposta:

[k] – número de bases naturais distribuídas aleatoriamente ao longo da sonda;

[e] – número de bases naturais fixas localizadas no final da sonda;

[c] – fator de distanciamento;

[p] – probabilidade de uma sonda ser falsa positiva;

[q] – probabilidade de uma sonda ser falsa negativa;

[m] – tamanho da seqüência;

[n] – número de sondas no chip.

Os itens a seguir descrevem cada uma das cinco principais rotinas que compõem o programa de simulação.

a) *rotina de geração da seqüência de DNA* – todos os testes foram executados com seqüências artificiais de DNA. Para produzir tais seqüências, o programa de simulação possui uma rotina que recebe, como entrada, o tamanho *m* da seqüência e gera randomicamente cada uma de suas bases componentes. Naturalmente, como se trata de DNA, a saída da rotina é uma seqüência escrita sobre o alfabeto {A, C, G, T};

b) *rotina de geração do espectro* – como apresentado no Capítulo 3, a obtenção, na prática, do espectro de uma seqüência de DNA resulta de uma série de reações de hibridização que ocorrem entre um conjunto de cópias dessa seqüência e o conjunto de sondas fixadas sobre a superfície do chip de seqüenciamento usado no experimento. Entretanto, para fins de teste da nova abordagem, nenhum experimento precisou ser efetivamente realizado, mas, ao invés disso, essas reações bioquímicas foram simuladas *in virtuo*. O programa de simulação inclui uma rotina que recebe como entrada a seqüência de DNA a ser determinada e os parâmetros do chip de seqüenciamento a ser "construído" e "usado", e gera, como saída, o conjunto de sondas que apóiam essa seqüência em pelo menos uma localização ao longo da sua extensão, ou seja, gera o seu espectro. Vale ressaltar que a seqüência de entrada dessa rotina é proveniente da *rotina de geração da seqüência*, descrita acima; e os parâmetros que definem o chip são *c*, *k*, *e* e *n*;

c) *rotina de simulação de erros no espectro* – deve-se observar que o espectro gerado através da rotina descrita acima é um espectro livre dos erros que comumente afetam os experimentos práticos. Assim, para tentar aproximar a simulação da realidade, foi criada uma rotina que introduz erros no espectro original, com uma taxa *p* de erros

falsos positivos e *q* de erros falsos negativos. O espectro e essas taxas constituem os parâmetros de entrada da rotina;

d) *rotina de reconstrução* – essa rotina está associada à segunda etapa do SBH, o que significa dizer que é responsável por tentar reconstruir a seqüência original a partir do conteúdo do espectro. Enquanto o chip fornece informações importantes sobre a constituição da seqüência, essa rotina se preocupa em fazer uso adequado delas. Praticamente, essa rotina é a implementação direta do algoritmo alternativo proposto. Foi preciso apenas incluir um critério de parada para garantir que a rotina interrompa a execução após reconstruir as *m* bases da seqüência alvo. Da mesma forma que no caso dos algoritmos $\mathcal{A}$ e $\mathcal{B}$, é assumido que o tamanho *m* é conhecido. Como parâmetros de entrada, a rotina recebe *m* e o espectro da seqüência a ser reconstruída, e, ainda, para que possa manipular adequadamente as sondas desse espectro, informações sobre a estrutura do chip usado, isto é, sobre f, c, k e e. Por outro lado, a saída da rotina é uma seqüência que, em caso de sucesso do procedimento, corresponde à seqüência que deu origem ao espectro usado;

e) *rotina de verificação* – além das rotinas relacionadas ao processo de SBH propriamente dito, o programa de simulação inclui uma rotina que verifica se o mesmo foi bem sucedido ou não, comparando, base a base, a seqüência resultante da rotina de reconstrução com a seqüência gerada, no início do processo, pela rotina de geração de DNA. Essa rotina recebe como entrada essas duas seqüências e fornece como saída uma informação binária para representar o sucesso ou o fracasso.

5.5.2 Testes

Os testes realizados consistem na chamada sistemática das rotinas de simulação com diferentes valores para os parâmetros de entrada. Todos os resultados foram organizados na forma de gráficos e estão apresentados no Apêndice A. Deve-se observar que cada gráfico individualmente enfatiza alguns aspectos particulares da nova proposta, mas que a visão completa do seu desempenho passa pela análise mais detida do conjunto desses gráficos.

Nessa etapa do estudo procurou-se investigar principalmente a influência de certos fatores no sucesso do processo de reconstrução, quais sejam:

a) distanciamento *c*;

b) taxas de erros;

c) número de famílias;

d) número e configuração das bases naturais e universais;

e) tamanho da seqüência.

Os teste realizados podem ser categorizados nos cinco tipos descritos a seguir:

a) tipo 1: com esses testes buscou-se determinar o tamanho da maior seqüência que pode ser corretamente reconstruída em aproximadamente 90% dos casos como

função do distanciamento *c*. Para tanto, adotou-se para *c* valores inteiros no intervalo de 2 a 40. Também foram considerados diferentes valores para os parâmetros *k*, *e* e *n*, de modo que cada gráfico corresponde a um valor particular de *n* e um valor particular de *k* + *e* (número de bases naturais); e cada curva, dentro de um gráfico, a uma das possíveis combinações de *k* e *e* que resulte no valor *k* + *e* definido para aquele gráfico. Observa-se que a curva para *e* = 1 corresponde à solução proposta por Halperin et al. (2002). Os valores adotados para *n* foram escolhidos para que os testes fossem feitos com 4 e 16 famílias de sondas. Para *k* + *e* adotaram-se os valores 5, 6 e 7. Em todos os testes, considerou-se o espectro livre de erros, ou seja, adotou-se $p = q = 0$. Os resultados desses testes estão sintetizados nas Figuras 1A, 2A e 3A;

b) tipo 2: com os testes desse tipo, visou-se caracterizar o comportamento da nova abordagem na presença de erros no espectro. Mais especificamente, buscou-se determinar o tamanho da maior seqüência que pode ser corretamente reconstruída em aproximadamente 90% dos casos em função dos erros falsos positivos, dos erros falsos negativos, ou ainda, de ambos atuando ao mesmo tempo. Nesse último caso, adotou-se *p* = *q*. Para todos os casos, *p* e *q* variaram de 0% a 50% com incremento de 5%, e adotou-se *c* = 10. Também foram considerados diferentes valores para os parâmetros *k*, *e* e *n*, de modo que cada gráfico corresponde a um valor particular de *n* e a um valor particular de *k* + *e*, e cada curva, dentro de um gráfico, a uma das possíveis combinações de *k* e *e* que resulte no valor *k* + *e* definido para aquele gráfico, com exceção da curva em que *k* = 1. Essa restrição foi considerada, porque ficou evidente dos primeiros experimentos que, para esse valor, não se conseguiria construir o chip com o número de famílias desejado. Observa-se que a curva para

e = 1 corresponde à solução proposta por Halperin et al. (2002). Os valores adotados para n foram escolhidos para que os testes fossem feitos com 4 e 16 famílias de sondas. Para $k + e$ adotaram-se os valores 5, 6 e 7. Os resultados desses testes estão sintetizados nas Figuras de 4A a 12A;

c) tipo 3: certamente a métrica mais importante para definir o rendimento de um método de seqüenciamento de DNA é o tamanho da seqüência que ele consegue reconstruir. O terceiro tipo de teste realizado procura justamente evidenciar a taxa de sucesso de reconstrução da seqüência alvo como função do seu tamanho m. Para cada valor de m foram realizadas 200 simulações e para elas foi computada a porcentagem das tentativas bem sucedidas. Além de variar m, foram também considerados diferentes valores para os parâmetros k, e, n, p, q, de modo que cada gráfico corresponde a um valor particular de n , de $k + e$, e de p e q, e cada curva dentro de um gráfico, a uma das possíveis combinações de k e e que resulte no valor $k + e$ definido para aquele gráfico, com exceção da combinação onde k = 1. Para n, adotaram-se valores que garantissem chips com 4 e 16 famílias de sondas, e para $k + e$, os valores 5, 6 e 7. Adotaram-se também $p = q$ = 1% e $p = q$ = 5%. Em todos os casos, considerou-se c = 10. Os resultados desses testes estão sintetizados nas Figuras de 13A a 18A;

d) tipo 4: à semelhança dos testes do tipo 2, esses testes visaram a caracterizar o desempenho da nova abordagem em lidar com espectro contendo erros. Mais especificamente, buscou-se determinar o tamanho da maior seqüência que pode ser corretamente reconstruída em aproximadamente 90% dos casos como função dos erros falsos positivos, dos erros falsos negativos e, ainda, de ambos atuando ao

mesmo tempo. Nesse último caso, adotou-se $p = q$. Para todos os casos, p e q variaram de 0% a 50% com incremento de 5%. Os testes realizados buscaram melhor caracterizar essa relação – tamanho reconstruído *versus* taxa de erro – pela comparação de pares de chips com o mesmo "custo" de construção, isto é, formados pela mesma quantidade de sondas, as quais devem ter o mesmo tamanho, o mesmo número de bases naturais e, por conseqüência, o mesmo número de bases universais. No total, foram considerados 3 pares de valores para k, c e e: {(4,4,3), (6,3,1)}; {(4,7,3), (6,5,1)}; {(4,10,3), (6,7,1)}. Para esses valores construíram-se chips contendo 4 e 16 famílias de sondas. Os resultados dos testes então organizados de modo que cada gráfico gerado corresponde a cada um desses pares, a um número de famílias e a um tipo de erro; e cada curva, dentro dos gráficos, a uma das combinações de k, c e e. Deve-se notar que, para efeito comparativo, adotou-se em cada gráfico a curva com $e = 1$, a qual corresponde à solução proposta por Halperin et al. (2002). Os resultados desses testes estão sintetizados nas Figuras de 18A a 27A;

e) tipo 5: à semelhança dos testes do tipo 3, esses testes evidenciam a taxa de sucesso de reconstrução da seqüência alvo como função do seu tamanho m. Nesse caso, o objetivo foi comparar o desempenho de pares de chips com o mesmo "custo" de construção, conforme fora realizado nos testes do tipo 4. Inclusive foram considerados os mesmos 3 pares de valores adotados para k, c e e. Usando esses valores, foram construídos chips com 4 e 16 famílias, de modo que cada gráfico gerado corresponde a um par de chips; e cada curva dentro de um gráfico, a um deles em particular. Em todos os casos adotou-se $p = q = 0{,}5\%$. Os resultados desses testes estão sintetizados nas Figuras de 28A a 30A.

5.5.3 Resultados

Os resultados dos testes realizados com essa nova abordagem para o SBH estão sumarizados na forma de gráficos no Apêndice A. Esta seção apresenta e discute os principais aspectos evidenciados a partir da análise deles:

a) das figuras 1A, 2A, e 3A, fica evidente que, a partir de certo ponto, o distanciamento *c* tende a não influenciar mais no rendimento da abordagem, no que diz respeito ao tamanho da seqüência que pode ser reconstruída. Na realidade, esse aspecto já havia sido mencionado em Halperin et al. (2002), então o que merece ser destacado nesse ponto é que ele permanece válido para a nova proposta, ou melhor, para o caso geral. Deve-se notar que esse é um resultado importante, visto que o tamanho da sonda é um aspecto limitante da tecnologia SBH. Com base nisso, em alguns testes, foi fixado um valor para *c* (no geral, $c = 10$), a fim de que se pudesse avaliar o comportamento de outras variáveis envolvidas no problema;

b) escolhida uma quantidade de bases naturais ($k + e$) para as sondas, verifica-se que os piores valores alcançados pela nova abordagem, em termos do tamanho da seqüência que pode ser corretamente reconstruída, correspondem exatamente às curvas em que as variáveis *k* e *e* assumem os valores extremos possíveis. Por exemplo, assumindo que a quantidade de bases naturais é igual a 7, os menores desempenhos se verificam quando $k = 6$ e $e = 1$ ou quando $k = 1$ e $e = 6$. Na realidade, isso poderia ser intuitivamente esperado da análise teórica do método, pois, para $e = 1$, a solução leva em conta apenas as informações das sondas que apóiam a subseqüência final da seqüência alvo, isto é, as sondas associados às posições à direita da que se deseja

estender não são considerados; e para $k = 1$, a menos que o valor de c seja bastante elevado, não se consegue construir o chip de seqüenciamento com o número de famílias de sondas planejado. Observa-se que essa limitação fez com que em certos gráficos não se traçasse mais a curva correspondente ao valor $k = 1$. De qualquer forma, um ponto que deve ser destacado é o seguinte: tanto em um caso como no outro, o problema de baixo desempenho está relacionado ao uso de uma quantidade menor de sondas, ou seja, de informação;

c) da alínea anterior, tem-se que, na prática, o pior caso da nova proposta ocorre para $e = 1$, ou seja, para o caso particular que corresponde ao algoritmo $\mathcal{B}$ de Halperin et al. (2002). Indiretamente, isso significa dizer que, com a nova proposta, é possível obter melhores resultados. Contudo, devem-se fazer algumas ponderações sobre essa afirmação. Deve-se observar que o indicador de desempenho que está sendo considerado, nesse caso, é exclusivamente o tamanho reconstruído com sucesso. Com relação ao tempo de processamento, talvez, outros testes apontem para uma situação inversa. De qualquer modo, um fato que deve ser considerado, quando da avaliação desse compromisso entre o tempo de execução e o tamanho da seqüência reconstruída, é que, geralmente, as soluções com maior valor de e, ou de outra forma, com menor valor de k, tendem a usar sondas menores. Realmente, esse é um aspecto muito importante da nova proposta, principalmente diante da afirmação feita por Halperin et al. (2002) de que, apesar de as bases universais terem sido geradas com sucesso em laboratório, ainda não é certo se sondas longas, com muitas bases universais, podem hibridizar confiavelmente;

d) como visto anteriormente, uma questão importante a ser considerada na avaliação de um método de seqüenciamento por hibridização é o seu poder em lidar com espectros ruidosos. Com relação a essa métrica, teoricamente, deveria se esperar que a nova abordagem, por usar mecanismos análogos aos usados nas abordagens relacionadas aos algoritmos $\mathcal{A}$ e $\mathcal{B}$, baseados no uso de múltiplas sondas para confirma a posição a estender, também fosse resistente tanto a erros falsos positivos quanto a erros falsos negativos. Por isso, pode-se afirmar que os testes serviram muito mais para quantificar essa métrica do que para comprovar a sua efetividade. No geral, esses testes evidenciaram que os resultados apresentados por Halperin et al. (2002) – algoritmo $\mathcal{B}$ – são extensíveis à nova abordagem, ou seja, continua valendo a constatação de que os erros falsos positivos afetam menos o desempenho do método, em termos do tamanho da seqüência que pode se reconstruído, do que os falsos negativos; e que existe uma relação quase linear entre o tamanho reconstruído *m* e *p*. Além disso os testes evidenciaram que existe semelhança no comportamento das curvas para chips com o mesmo número de bases naturais ($k + e$), independentemente dos valores assumidos para *k* e *e*, frente às mesmas variações de erros;

e) com relação à composição do chip em famílias de sondas, os resultados evidenciaram o que era intuitivamente esperado da análise teórica do método. O uso de um maior número de famílias melhora o desempenho da abordagem, independentemente da combinação de *k* e *e*, e dos níveis de erro *p* e *q* adotados. Nos chips simulados nos testes, adotaram-se 4 e 16 famílias de sondas. Para esses dois valores, nota-se que o aumento correspondente no tamanho da seqüência corretamente reconstruída não chega a ser proporcional a essa variação.

CONCLUSÕES

As principais investigações no âmbito da Biologia Molecular Computacional lidam com moléculas de DNA e partem do princípio de que as seqüências de seus nucleotídeos componentes são conhecidas. Por isso, é de se esperar que o sucesso de tais investigações seja dependente do sucesso do método de seqüenciamento empregado, o qual, normalmente, é medido em termos do tamanho da molécula que consegue seqüenciar.

Entretanto, é importante ressaltar que nenhum método pode ser apresentado como o melhor em absoluto, mas o desempenho de cada um depende de fatores específicos, como tamanho da seqüência alvo, taxas de erros no experimento e, mesmo, da própria seqüência de bases em si. Foi exatamente nesse contexto que a proposta apresentada por este trabalho de pesquisa se desenvolveu.

Subsidiados pelas idéias apresentadas em Halperin et al. (2002), elaboramos uma solução alternativa para o problema de seqüenciamento de DNA, mais especificamente para o SBH, a qual inclui a proposta de um novo projeto de chip de seqüenciamento, que contempla o uso de bases universais distribuídas aleatoriamente, bem como o algoritmo de reconstrução correspondente.

Neste trabalho, ficou patente que essa solução atingiu o objetivo proposto, conquanto alcançou melhores resultados práticos do que o algoritmo $\mathcal{B}$ de Halperin et al. (2002), ao mesmo tempo que superou os problemas que afetavam o algoritmo $\mathcal{AB}$ (e.g., a necessidade de gerar grandes extensões e de considerar sondas que não ajudam na determinação da posição a estender).

Efetivamente, avançou-se mais do que isso, posto que a nossa proposta abrange a solução baseada no algoritmo $\mathcal{B}$ como uma de suas instâncias, mais especificamente, esta corresponde ao caso em que $e = 1$. Na realidade, as simulações evidenciaram que esse é o pior caso. Isso, se a análise for feita em termos do tamanho de reconstrução da cadeia alvo. Se, no entanto, a questão envolver tempo de processamento, é possível que os testes sigam em outra direção. Mas mesmo assim, um fator ainda pesa em favor das demais instâncias que a nossa solução pode assumir – é que existe a tendência de que elas usem sondas menores. Esse resultado também é bastante promissor, visto que não há garantia de que sondas contendo grandes seqüências de bases universais podem hibridizar confiavelmente.

Ficou patente, também, que a nova abordagem pode lidar tanto com o problema de multiplicidade de sondas quanto com o problema de erros no espectro. Entretanto, com relação à capacidade de lidar com reconstruções ambíguas, no momento, pode-se apenas, conjecturar sobre isso. Por usar os mesmos princípios, dos algoritmos $\mathcal{A}$ e $\mathcal{B}$, é possível que a nossa abordagem também lide em algum grau com esse problema.

A avaliação dessa nova abordagem considerou testes realizados sobre um conjunto restrito de valores. No futuro, pretendemos usar um conjunto mais amplo e ainda avaliar o

comportamento da nossa estratégia com seqüências reais de DNA, ao invés de seqüências geradas artificialmente. É possível que os resultados de todos esses experimentos nos remetam à análise matemática da nossa proposta.

Ensaios laboratoriais ajudariam a revelar o alcance real da nossa abordagem e, ainda, a sua viabilidade técnica e econômica. Talvez, através deles, pudéssemos nos deparar com questões ainda não abordadas e com dificuldades não previstas e, até mesmo, impeditivas de sua efetividade. De toda maneira, acreditamos que este trabalho tenha lançado alguma luz sobre o importante campo que é o seqüenciamento de DNA, mormente, sobre a técnica de SBH, de modo que esperamos que sirva de subsídio e ajude em pesquisa futuras voltadas para a exploração e compreensão dos mecanismos biológicos que garantem a “seqüência” da vida das espécies.

REFERÊNCIAS BIBLIOGRÁFICAS

BAINS, W. Hybridization Methods for DNA Sequencing. *Genomics*, v. 10, p. 294-301, 1991.

_____; SMITH, G. A Novel Method for Nucleic Acid Sequence Determination. *Journal of Theoretical Biology*, v. 135, p. 303-307, 1988.

BAPTISTA, E. S.; GUIMARÃES, K. S. Uma Abordagem Alternativa para Seqüenciamento por Hibridização. In: Congresso da Sociedade Brasileira de Computação, 23 / Seminário Integrado de Software e Hardware – SEMISH, 30, 2-8 ago. 2003, Campinas. *Anais*. Campinas: SBC, 2003. p. 697-708.

BEN-DOR, A. et al. On the Complexity of Positional Sequencing by Hybridization. In: International Conference On Combinatorial Pattern Matching – CPM, 10, 1999, New York. *Proceedings*. New York: ACM Press, 1999. p. 88-100.

BLAZEWICZ, J. et al. Sequential algorithms for DNA sequencing. *Computational Methods in Science and Technology*, v. 1, p. 31-42, 1996.

_____. Sequential and parallel algorithms for DNA sequencing. *CABIOS*, v. 13, p. 151-158, 1997.

_____. DNA sequencing with positive and negative errors. *Journal of Computational Biology*, v. 6, p. 113-123, 1999.

_____. Tabu search for DNA sequencing with false and false positives. *European Journal of Operational Research,* v. 125, p. 257-265, 2000.

BURKS, C. DNA Sequence Assembly. *IEEE Engineering in Medicine and Biology Magazine,* v. 13, n. 5, p. 771-773, 1994.

CHETVERIN, A. B.; KRAMER, F. R. Oligonucleotide Arrays: New Concepts and Possibilities. *Biotechnology*, v. 12, p. 1093-1099, 1994.

DRMANAC, R.; DRMANAC, S. Sequencing by Hybridization Arrays. In: RAMPAL, J. B. (ed.). *DNA Arrays*: Methods and Protocols. Totowa, NJ: Humana Press, 2001. p. 39-51. (Methods in Molecular Biology, v. 170).

DRMANAC, R. et al. Sequencing of megabase plus DNA by hybridization: theory and the method. *Genomics*, v. 4, p. 114-128, 1989.

_____. Sequencing by Hybridization (SBH): Advantages, Achievements, and Opportunities. In: HOHEISEL, J. (ed.). *Chip Technology*. Berlin: Springer-Verlag Heidelberg, 2002. p. 75-101. (Advances in Biochemical Engineering/Biotechnology, v. 77).

DRMANAC, R.; LABAT, I.; CRKVENJAKOV, R. An algorithm for DNA sequence generation from k-tuple word contents of the minimal number of random fragments. *Journal of Biomolecular Structure and Dynamics*, v. 8, p. 1085-1102, 1991.

GALLANT, J. The complexity of the overlap method for sequencing biopolymers. *Journal of Theoretical Biology*, v. 101, p. 1-17, 1983.

_____; MAIER, D.; STORER, J. On finding minimal length superstrings. *Journal of Computer and System Sciences*, v. 20, p. 50-58, 1980.

GUENOCHE, A. Can we recover a sequence, just knowing all its subsequence of given length? *CABIOS*, v. 8, p. 569-574, 1992.

GUSFIELD, D. *Algorithms on Strings, Trees and Sequences:* Computer Science and Computational Biology. New York: Cambrige University Press, 1997.

HALPERIN, E. et al. Handling long targets and errors in sequencing by hybridization. In: Annual International Conference on Research in Computational Molecular Biology – RECOMB, 6, 2002, Washington, DC. *Proceedings*. New York: ACM Press, 2002. p. 176-185.

KECECIOGLU, J. D. *Exact and approximate algorithms for DNA sequence reconstruction*. 1991. Ph.D. Thesis – Dept. of Computer Science, University of Arizona, Tucson.

KHRAPKO. K. R. et al. An oligonucleotide hybridization approach to DNA sequencing. *FEBS Letters*, v. 256, p. 118-122, 1989.

LEHNINGER, A. L.; NELSON, D. L.; COX, M. M. *Princípios de Bioquímica*. Tradução de W.R. Loodi, e A. A. Simões. São Paulo: Sarvier, 1995. 839 p.

LIPSHUTZ, R. J. Likelihood DNA sequencing by hybridization. *Journal of Biomolecular Structure and Dynamics*, v. 11, p. 637-653, 1993.

LYSOV, Y. et al. DNA sequencing by hybridization with oligonucleotides. *Doklady Academy of Science of USSR*, v. 303, p. 1508-1511, 1988.

MAIER, D. The complexity of some problems on subsequences and supersequences. *Journal of the ACM*, v. 25, p. 322-336, 1978.

MAXAM, A.; GILBERT, W. A new method for sequencing DNA. *Proc. Natl. Acad Sci*, v. 74, p. 560-564, 1977.

MEIDANIS, J.; SETUBAL, J. C. Uma Introdução à Biologia Computacional. In: Escola de Computação, 9, 24-31 jul. 1994, Recife. *Livro*. Recife: UFPE-DI, 1994. p. 131.

PEDERSEN, C. N. S. *Algorithms in computational Biology*. 1999. Ph. D. Thesis – Faculty of Science, University of Aarhus, Denmark.

PE'ER, I.; SHAMIR, R. Spectrum Alignment: Efficient Resequencing by Hybridization. In: International Conference on Intelligent Systems for Molecular Biology – ISMB, 8, 2000, San Diego. *Proceedings*. Menlo Park: AAAI Press, 2000. p. 260-268.

PEREIRA, L. V. *Seqüenciaram o Genoma Humano... E agora?* São Paulo: Moderna, 2001. 132 p.

PETOLA, H., SODERLUND, H. UKKONEN, E. SEQAIDS: A DNA sequence assembling program based on a mathematical model. *Nucleic Acids Research*, v. 12, p. 307-321, 1984.

PEVZNER, P. A. *Computational Molecular Biology:* An Algorithm Approach. Cambridge, MA: The MIT Press, 2000. 332 p.

_____. L-tuple DNA sequencing: Computer analysis. *Journal of Biomolecular Structure and Dynamics*, v. 7, p. 63-73, 1989.

PEVZNER, P. A. et al. Improved Chips for sequencing by Hybridization. *Journal of Biomolecular Structure and Dynamics*, v. 9, p. 399-410, 1991.

PEVZNER, P. A.; WATERMAN, M. S. Open combinatorial problems in computational molecular biology. In: Israel Symposium on Theory of Computing and Systems – ISTCS, 3, 1995, Tel Aviv. *Proceedings*. Los Alamitos: IEEE Computers Society Press, 1995. p.158-173.

PREPARATA, F. P.; FRIEZE, A. M.; UPFAL, E. On the power of universal bases in sequencing by hybridization. In: Annual International Conference on Research in Computational Molecular Biology – RECOMB, 3, 1999, Lyon. *Proceedings*. New York: ACM Press, 1999. p. 295-301.

SANGER, F. et al. Nucleotide sequence of bacteriophage λ DNA. *Journal of Molecular Biology*, v. 162, p. 729-773, 1982.

SANGER, F.; NICKLEN, S.; COULSON, A. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Science USA*, v. 74, p. 5463-5467, 1977.

SETUBAL, J. C.; MEIDANIS, J. *Introduction to Computational Molecular Biology*. Boston: PWS Publishing Company, 1997. 296 p.

SHAMIR, R. *Analysis of Gene Expression Data*. Lecture 3, 2002. Disponível em: <http://www.math.tau.ac.il/~rshamir/ge/02/scribes/lec03.pdf>. Acesso em: 26 mai. 2003.

SHAMIR, R.; TSUR, D. Large Scale Sequencing By Hybridization. In: Annual International Conference on Research in Computational Molecular Biology – RECOMB, 5, 2001, Montreal. *Proceedings*. New York: ACM Press, 2001. p. 269-277.

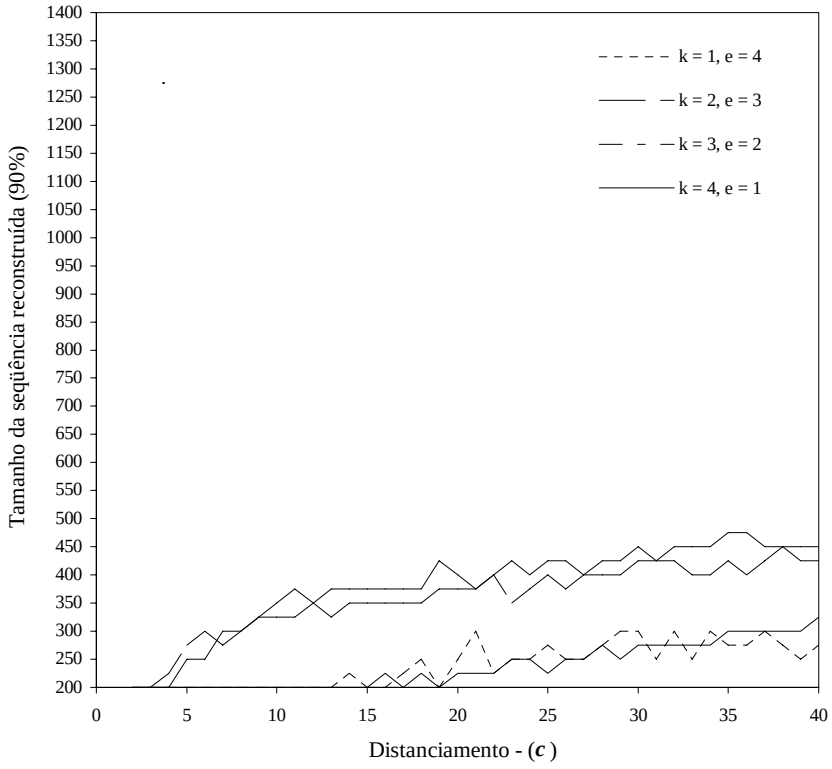
SOUTHERN, E. *United Kingdom Patent Application GB8810400*. 1988.

STERKY, F.; LUNDEBERG, J. Sequence analysis of genes and genomes. *Journal of Biotechnology,* v. 76, p. 1-31, 2000.

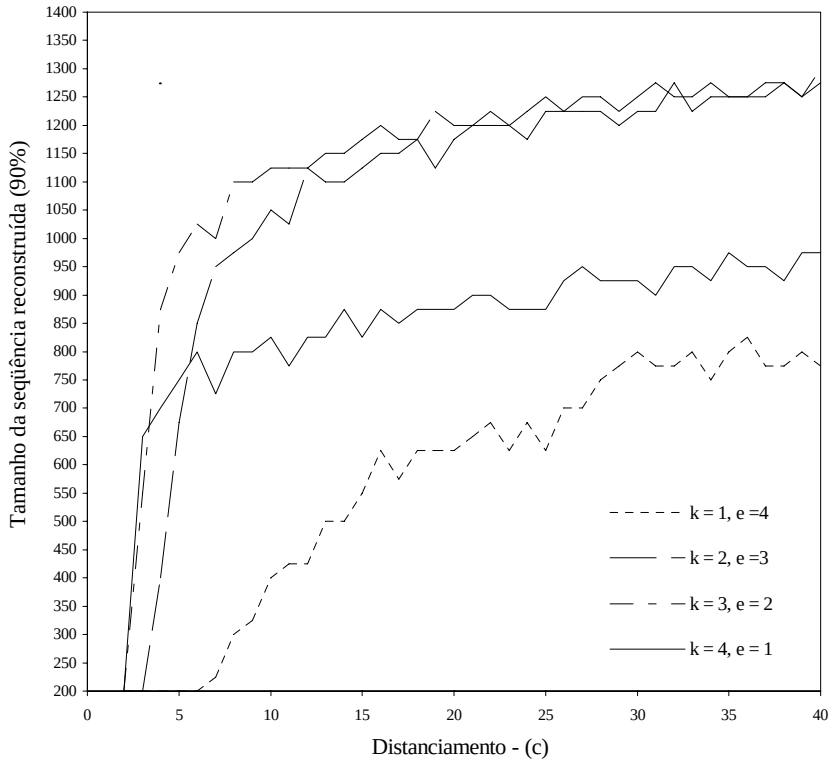
TAMMI, M. T. *The Principles of Shotgun Sequencing and Automated fragment Assembly*: Special excerpt for lecture. Stocholm: Center for Genomics and Bioinformatics, Karolinska Institutet, 2003. Disponível em: <http://web.cgb.ki.se/student/sfa.pdf>. Acesso em: 26 mai. 2003.

TUNER, J. Approximation algorithms for the shortest common superstring problem. *Information and Computation*, v. 83, p. 1-20, 1989.

APÊNDICE A – Resultados dos Testes (Gráficos)

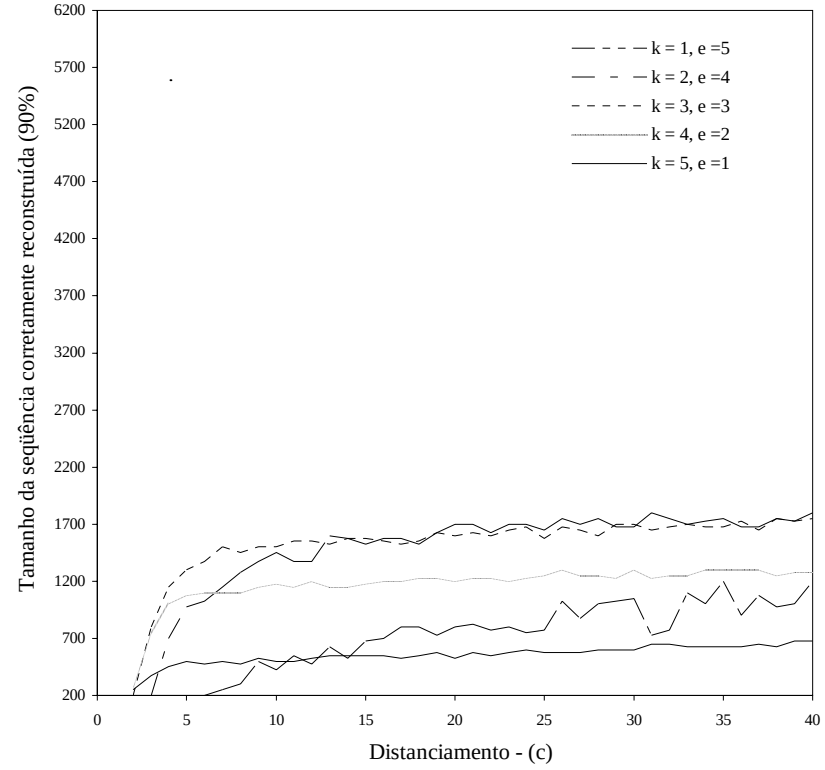


(a)

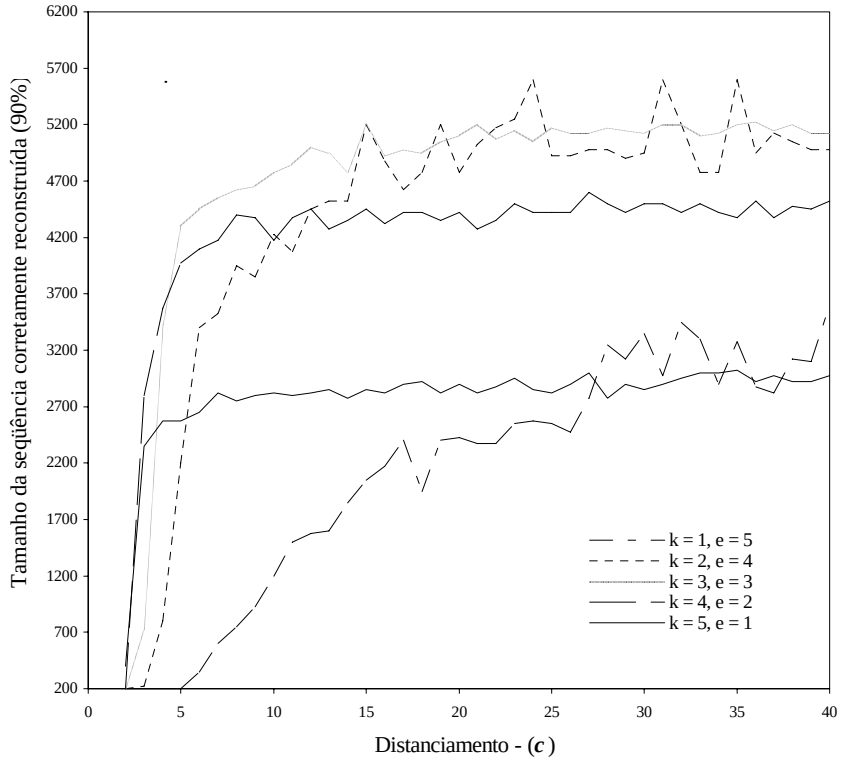


(b)

Figura 1A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* o distanciamento *c* das bases naturais, quando $k + e = 5$, $p = q = 0$, (a) $n = 4^6$ e (b) $n = 4^7$.

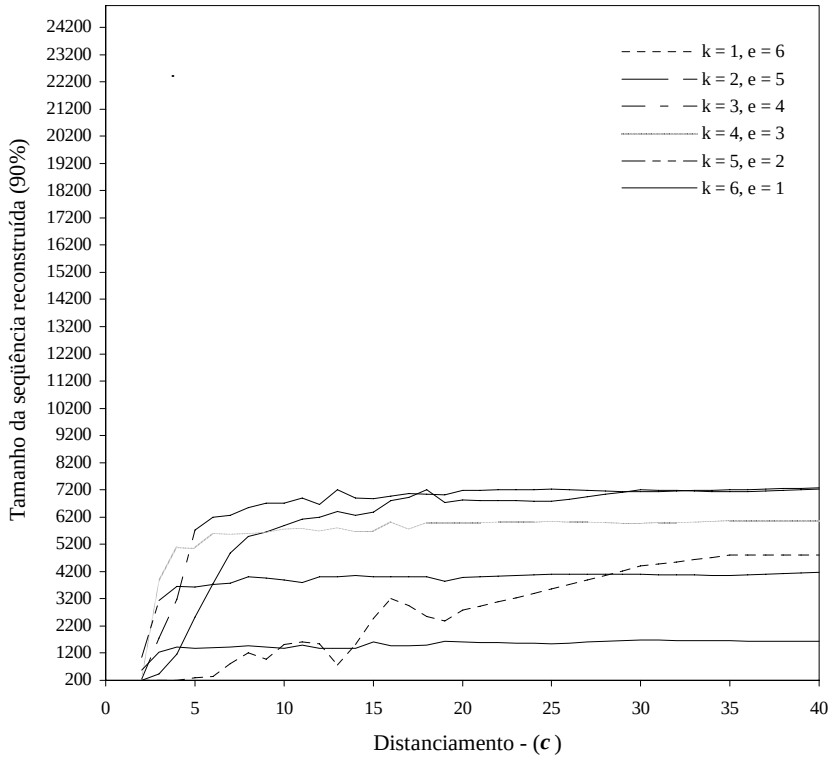


(a)

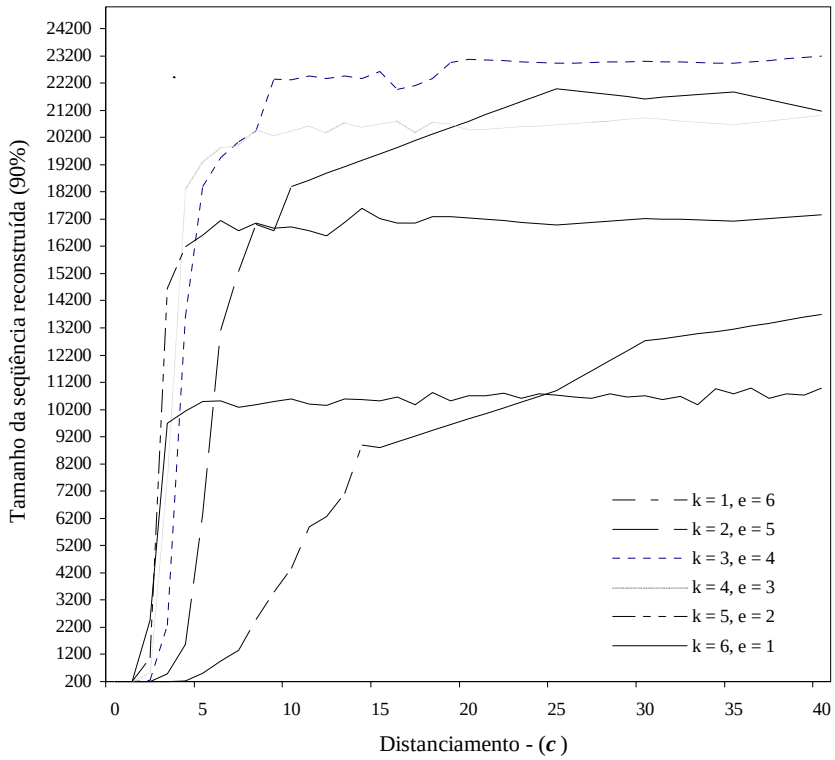


(b)

Figura 2A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* o distanciamento *c* das bases naturais, quando $k + e = 6$, $p = q = 0$, (a) $n = 4^7$ e (b) $n = 4^8$.

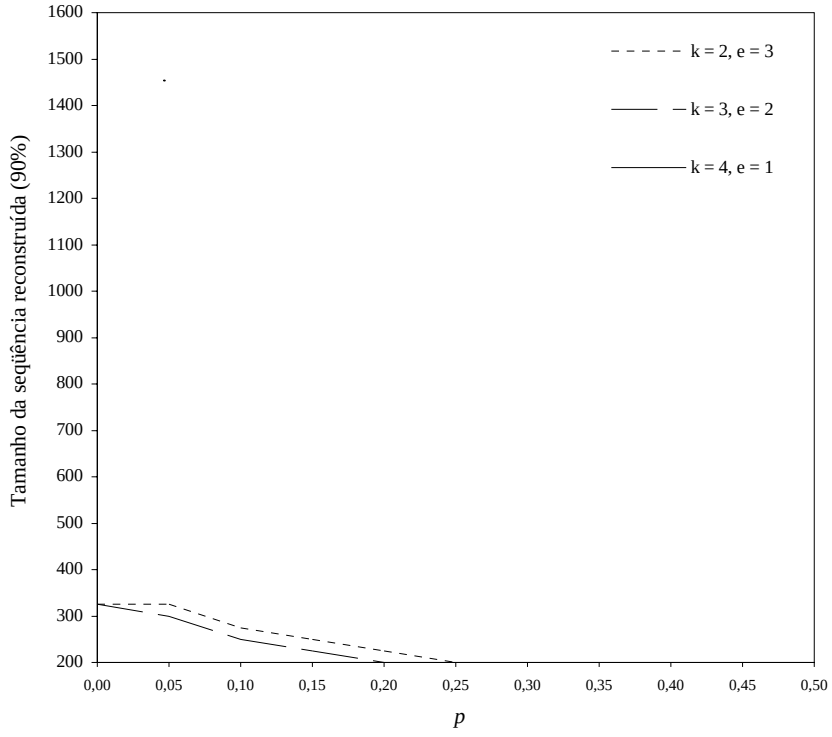


(a)

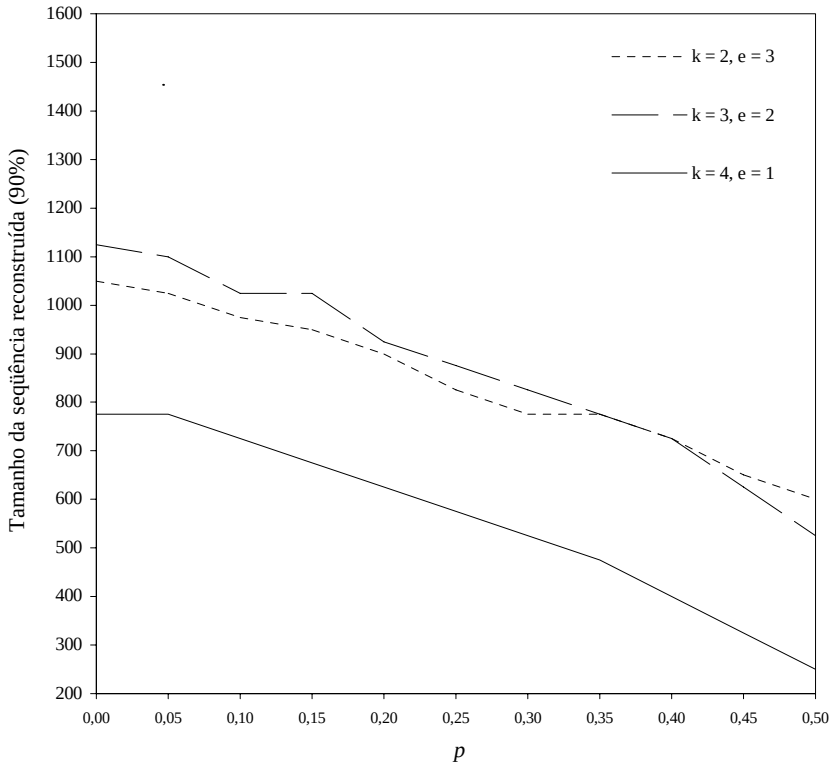


(b)

Figura 3A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* o distanciamento *c* das bases naturais, quando $k + e = 7$, $p = q = 0$, (a) $n = 4^8$ e (b) $n = 4^7$.

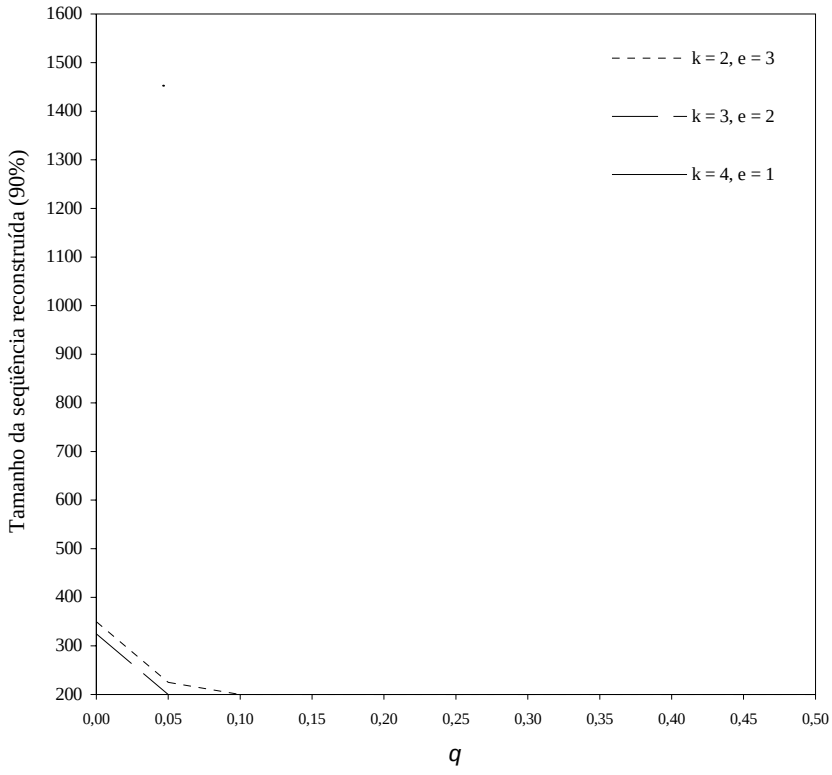


(a)

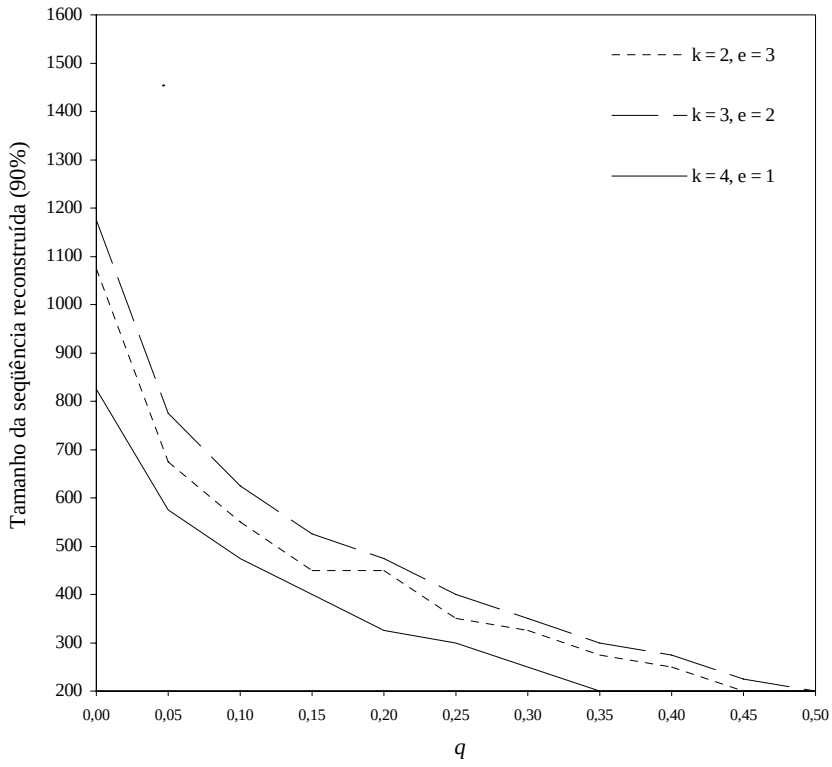


(b)

Figura 4A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* a taxa p de erro falso positivo, quando $k + e = 5$, $q = 0$, $c = 10$, (a) $n = 4^6$ e (b) $n = 4^7$. Em (a), a linha ($k = 4$, $e = 1$) é menor que 200.

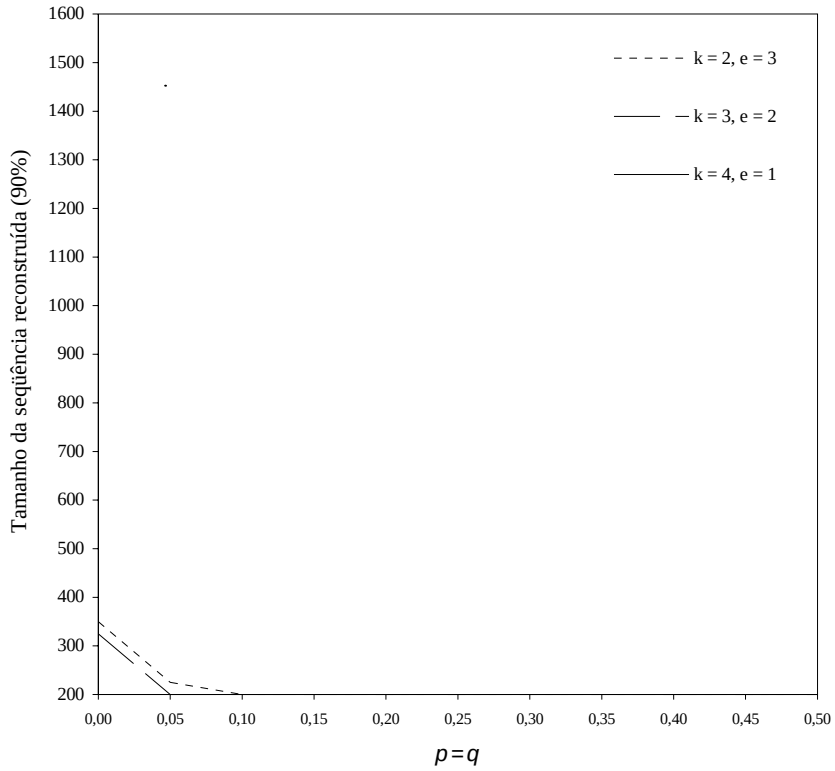


(a)

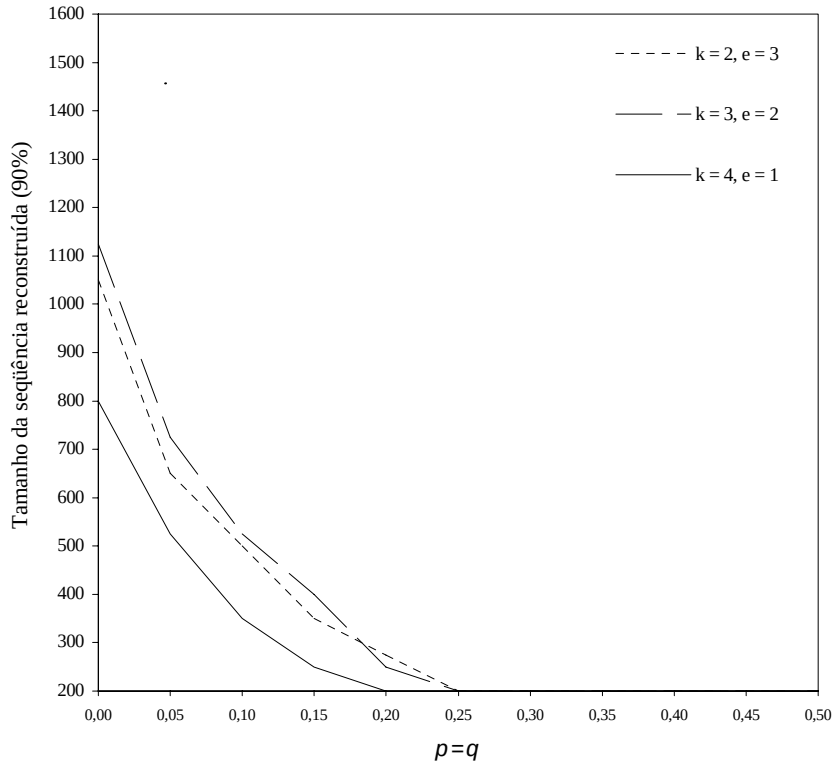


(b)

Figura 5A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* a taxa q de erro falso negativo, quando $k + e = 5$, $p = 0$, $c = 10$, (a) $n = 4^6$ e (b) $n = 4^7$. Em (a), a linha ($k = 4$, $e = 1$) é menor que 200.

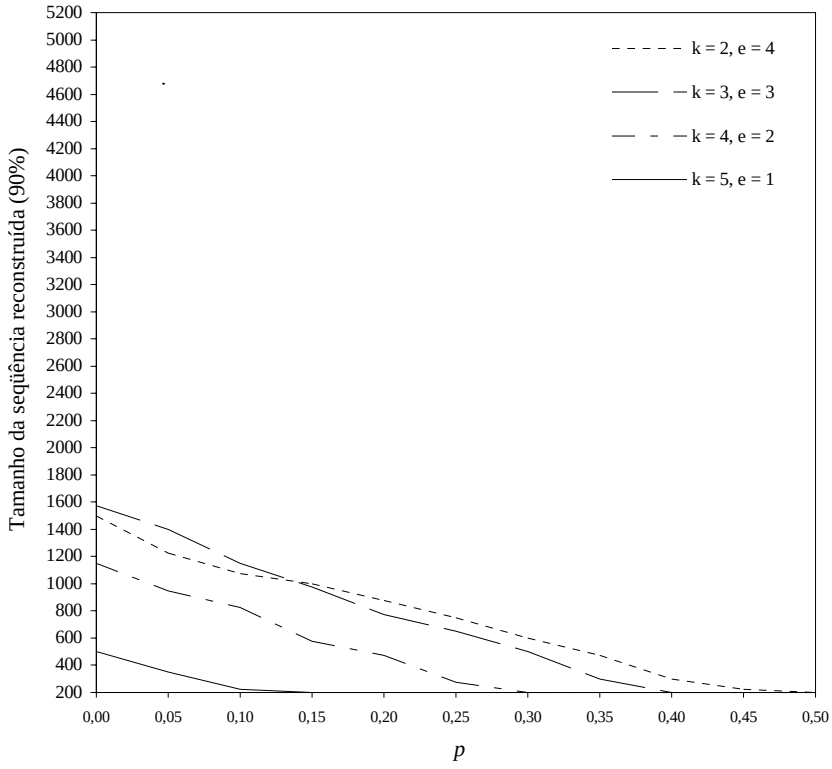


(a)

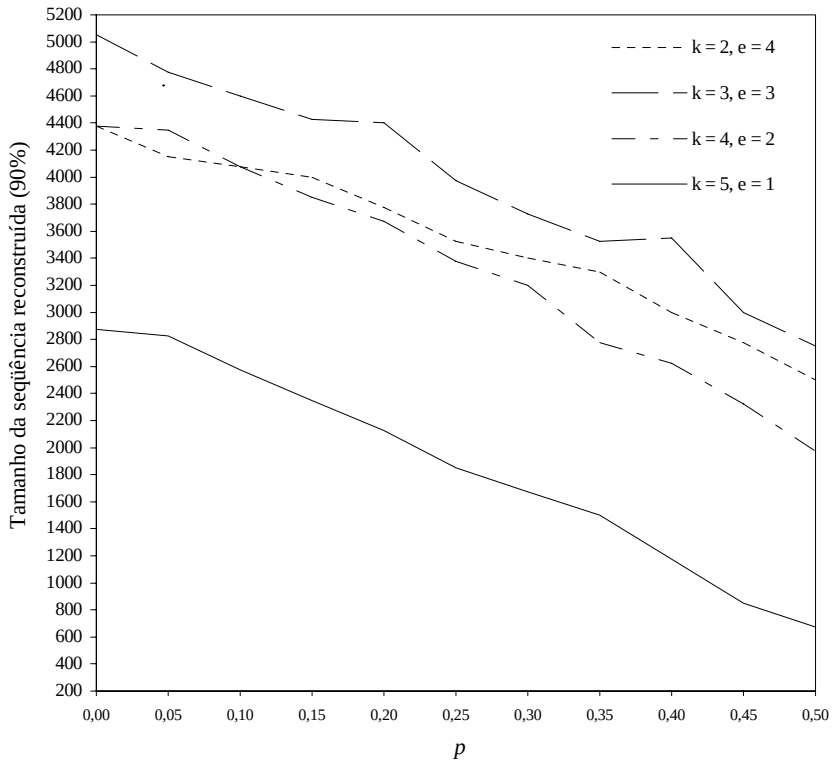


(b)

Figura 6A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* as taxas *p e q* ($p = q$) de erros, quando $k + e = 5$, $c = 10$, (a) $n = 4^6$ e (b) $n = 4^7$. Em (a), a linha ($k = 4$, $e = 1$) é menor que 200.

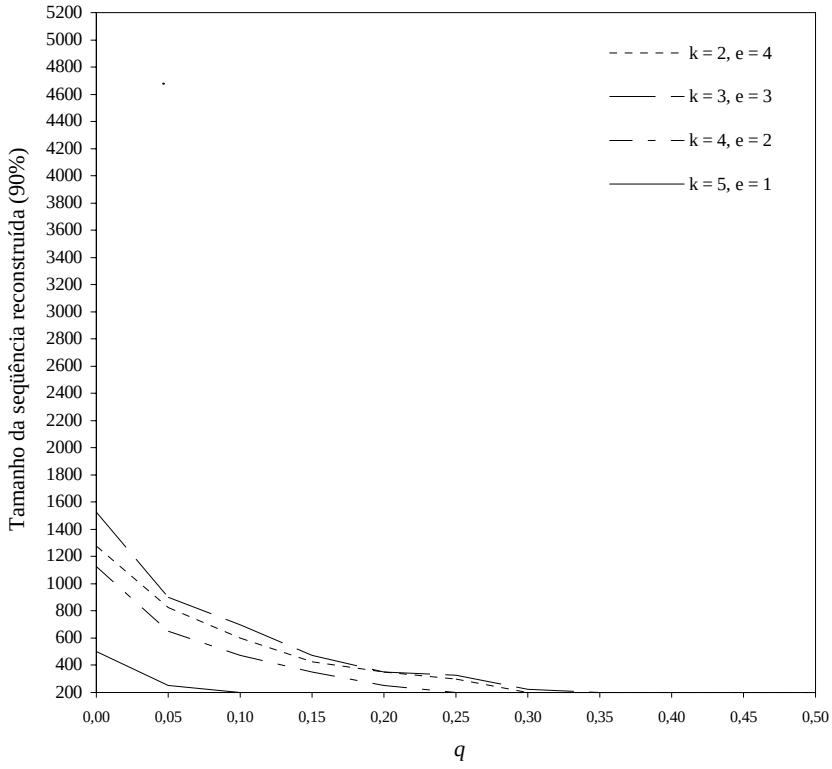


(a)

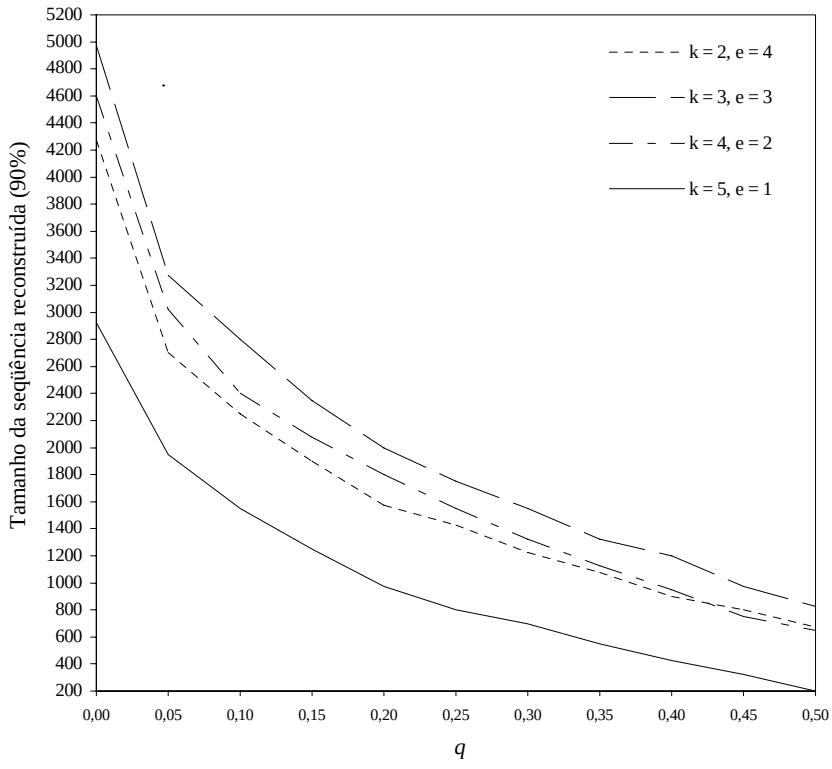


(b)

Figura 7A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* a taxa p de erro falso positivo, quando $k + e = 6$, $q = 0$, $c = 10$, (a) $n = 4^7$ e (b) $n = 4^8$.

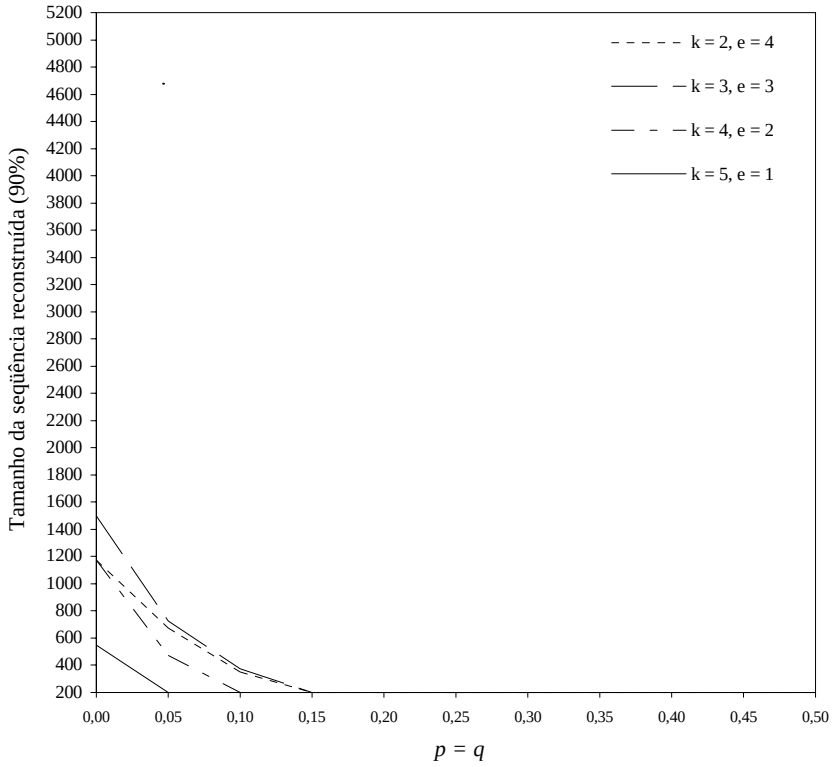


(a)

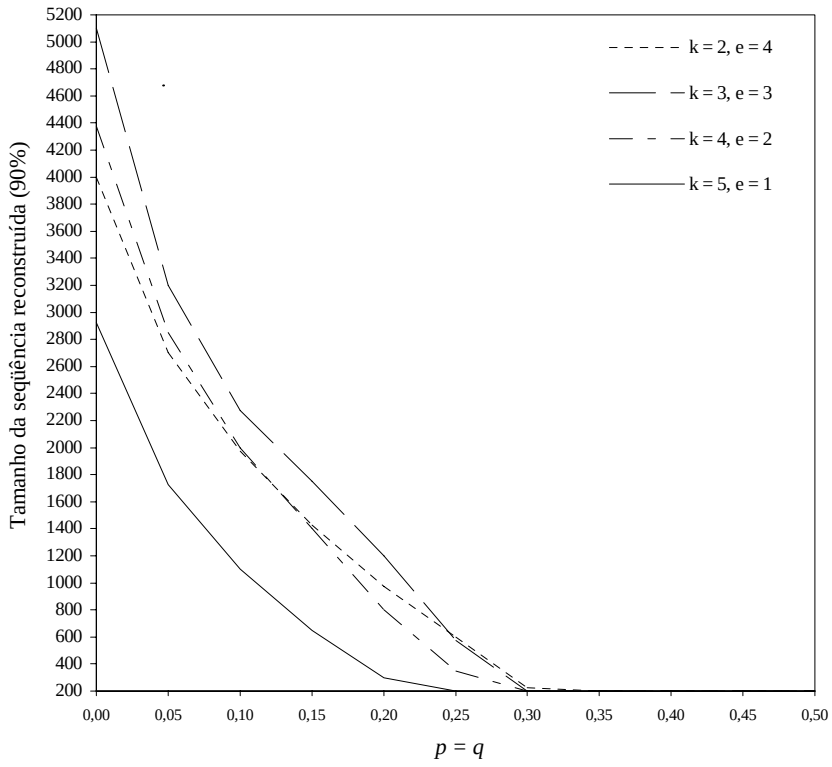


(b)

Figura 8A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* a taxa q de erro falso negativo, quando $k + e = 6$, $p = 0$, $c = 10$, (a) $n = 4^7$ e (b) $n = 4^8$.

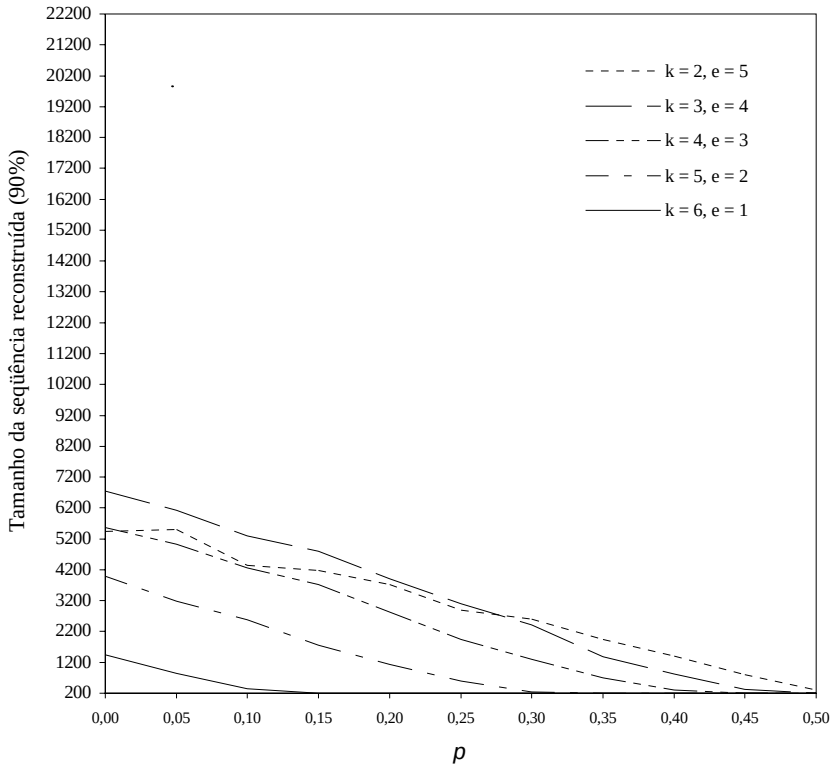


(a)

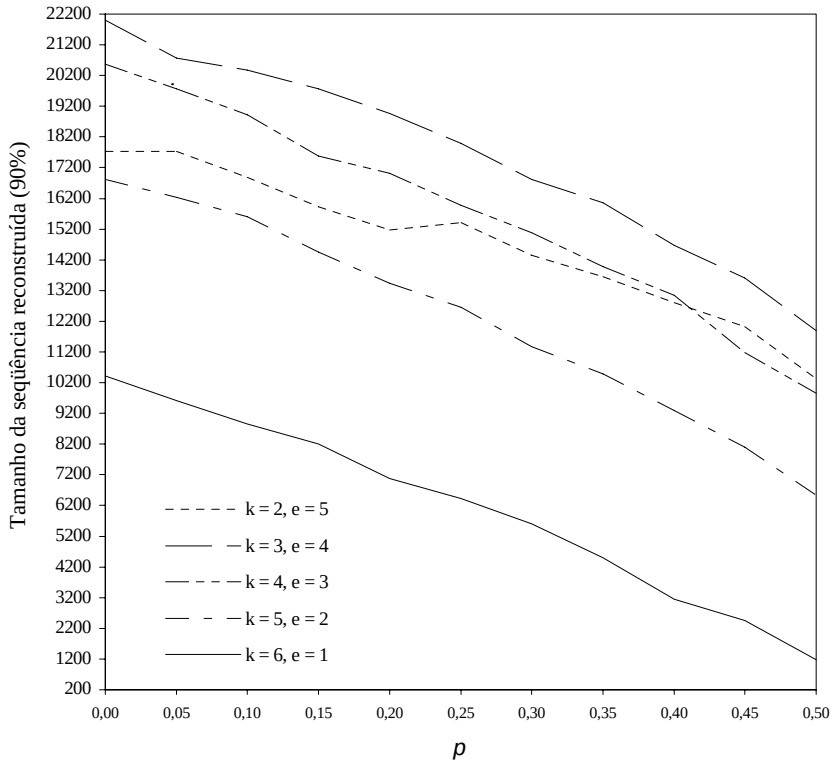


(b)

Figura 9A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* as taxas p e q $(p = q)$ de erros, quando $k + e = 6$, $c = 10$, (a) $n = 4^7$ e (b) $n = 4^8$.

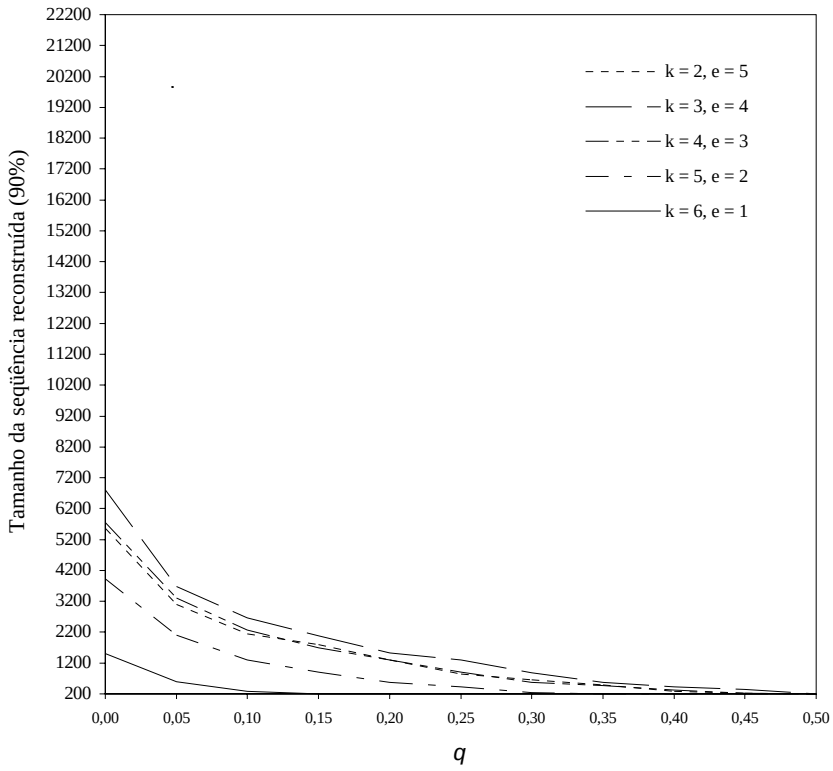


(a)

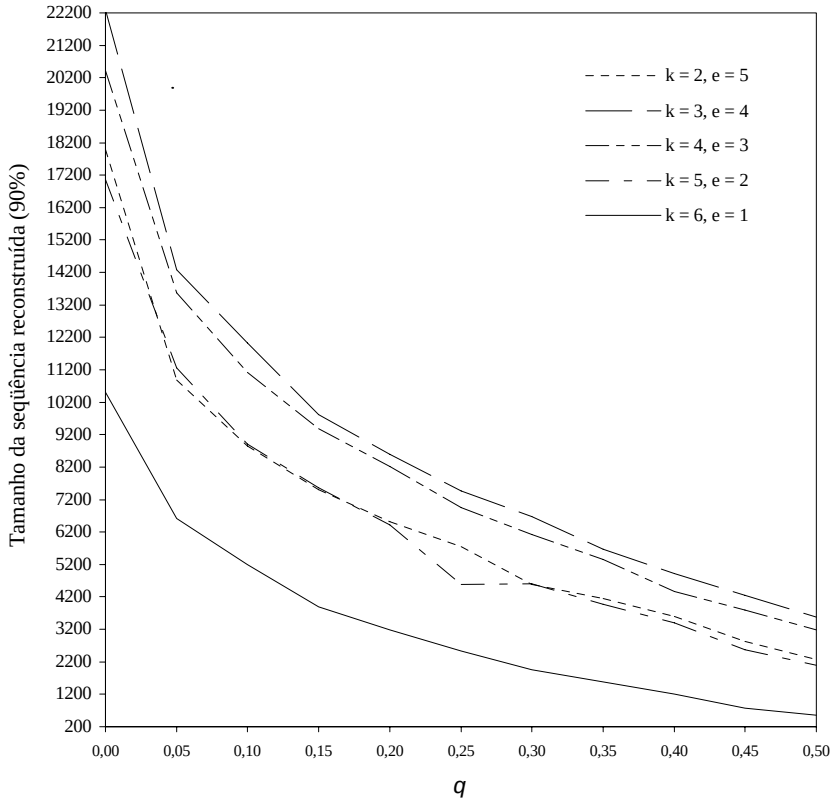


(b)

Figura 10A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* a taxa *p* de erro falso positivo, quando $k + e = 7$, $q = 0$, $c = 10$, (a) $n = 4^8$ e (b) $n = 4^9$.

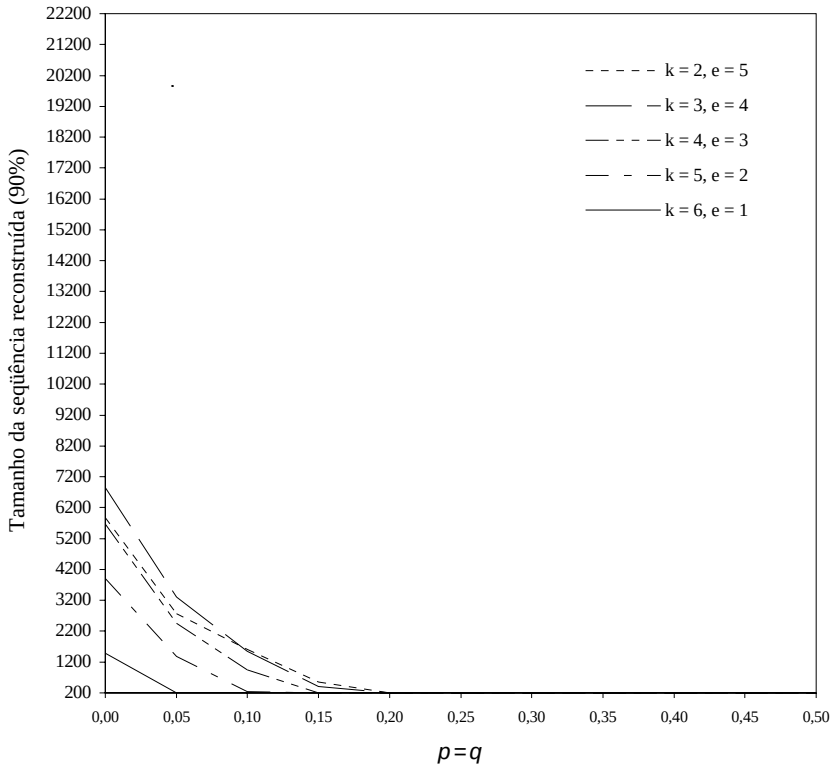


(a)

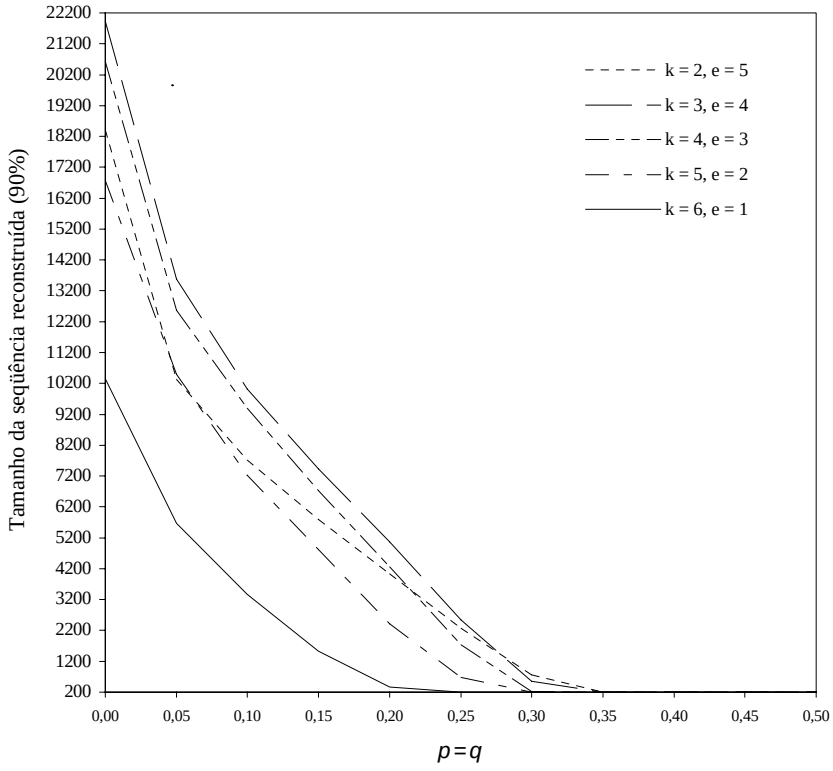


(b)

Figura 11A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* a taxa q de erro falso negativo, quando $k + e = 7$, $p = 0$, $c = 10$, (a) $n = 4^8$ e (b) $n = 4^9$.

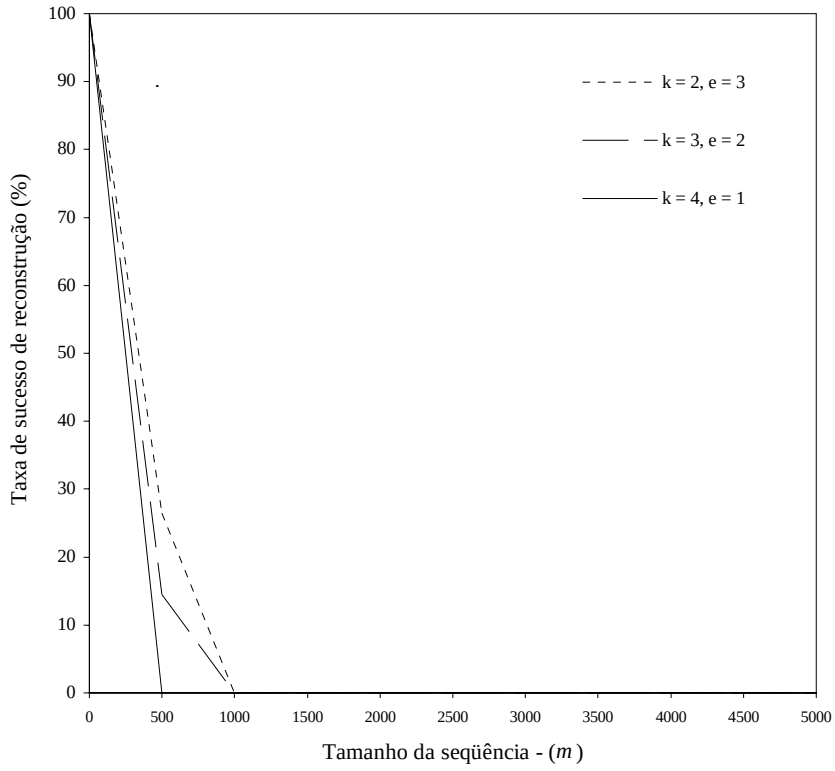


(a)

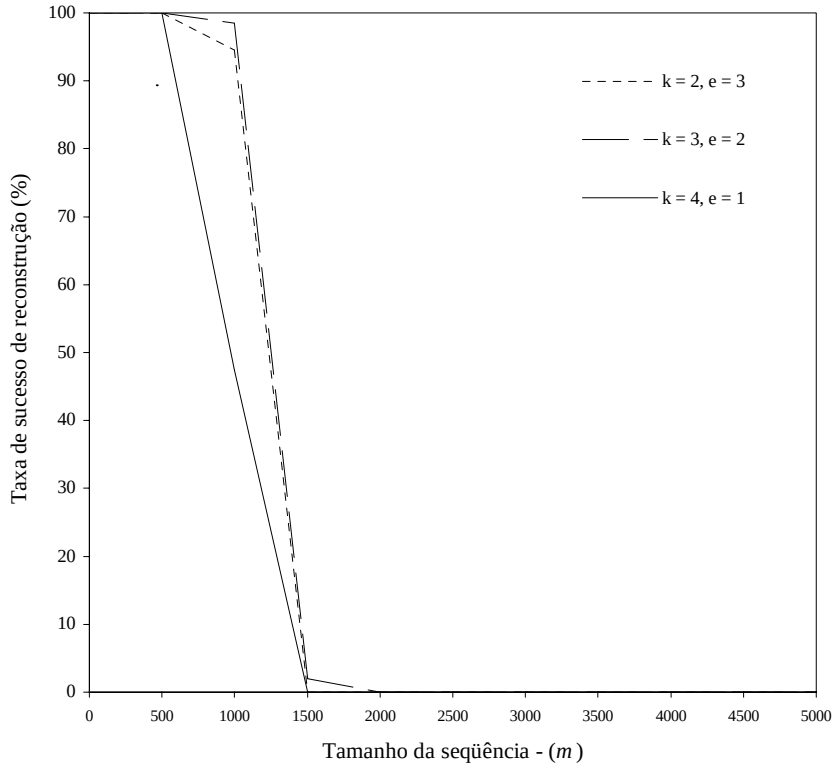


(b)

Figura 12A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* as taxas p e q ($p = q$) de erros, quando $k + e = 7$, (a) $n = 4^8$ e (b) $n = 4^9$.

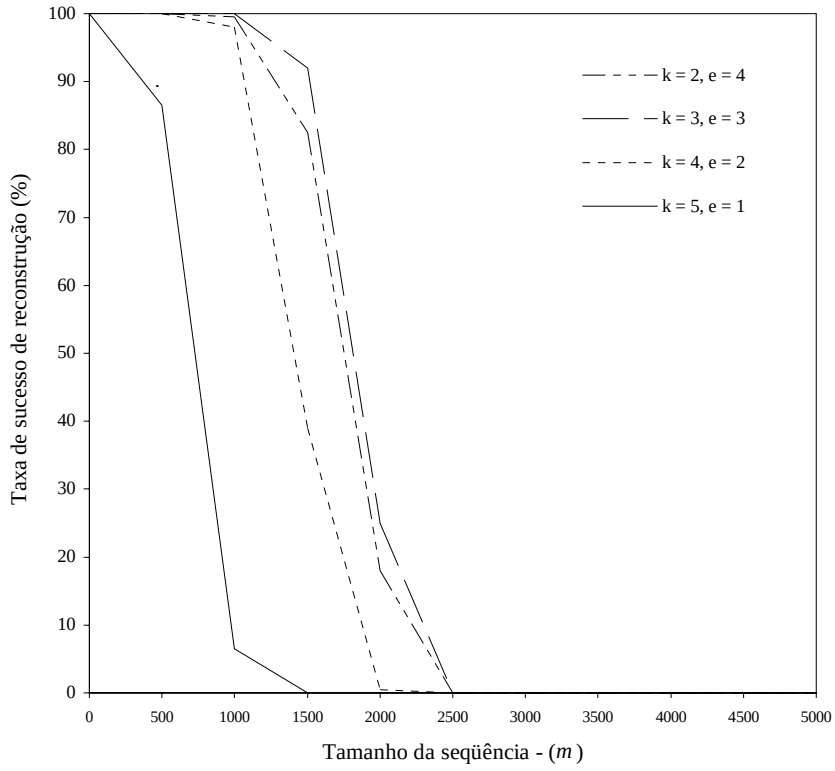


(a)

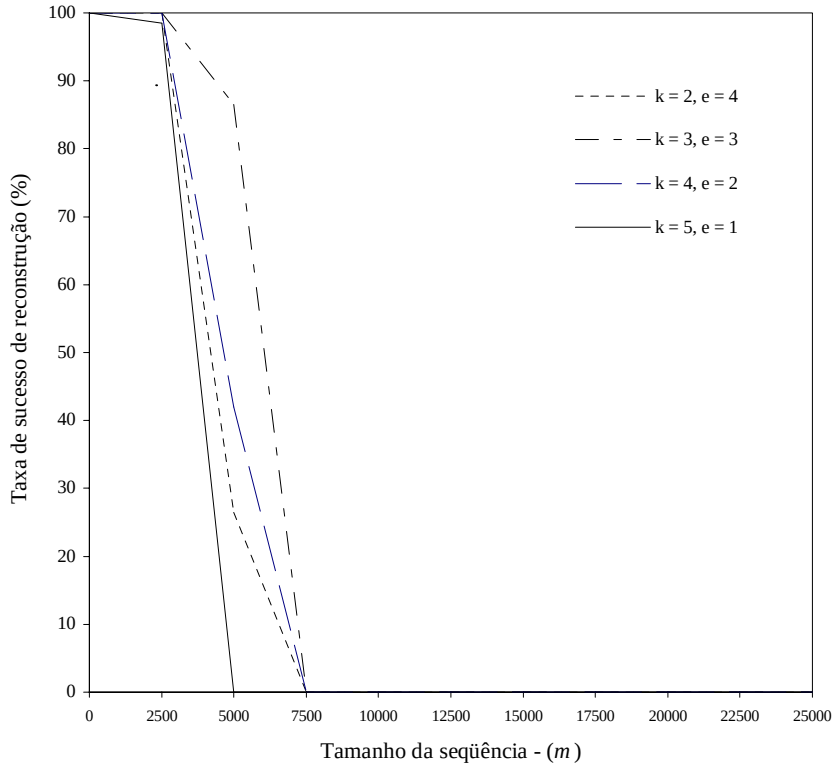


(b)

Figura 13A: Taxa de sucesso de reconstrução da seqüência *versus* o tamanho da seqüência, quando $k + e = 5$, $c = 10$, $p = q = 0{,}001$, (a) $n = 4^6$ e (b) $n = 4^7$.

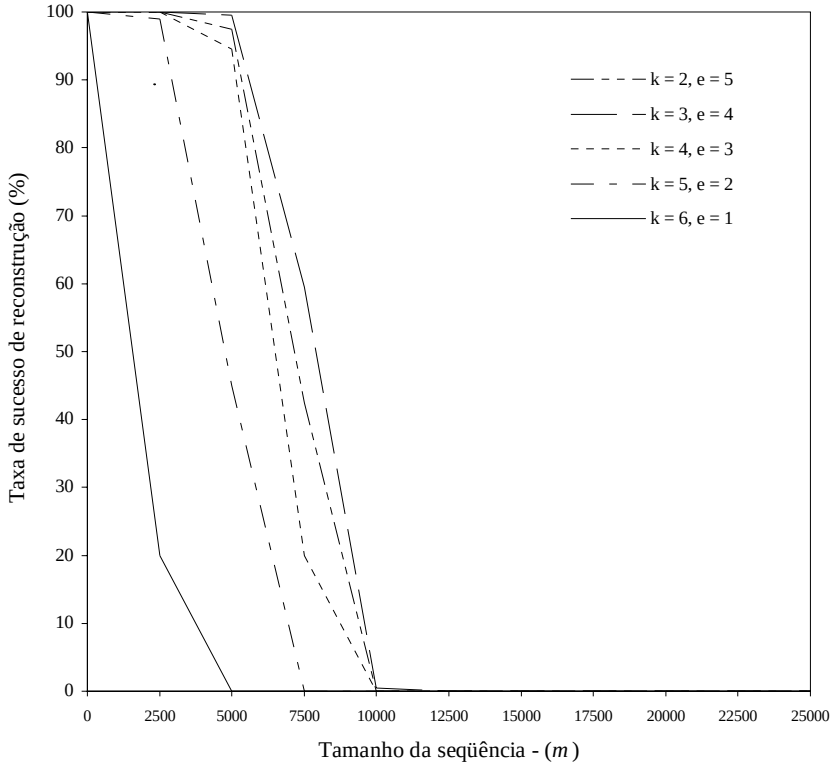


(a)

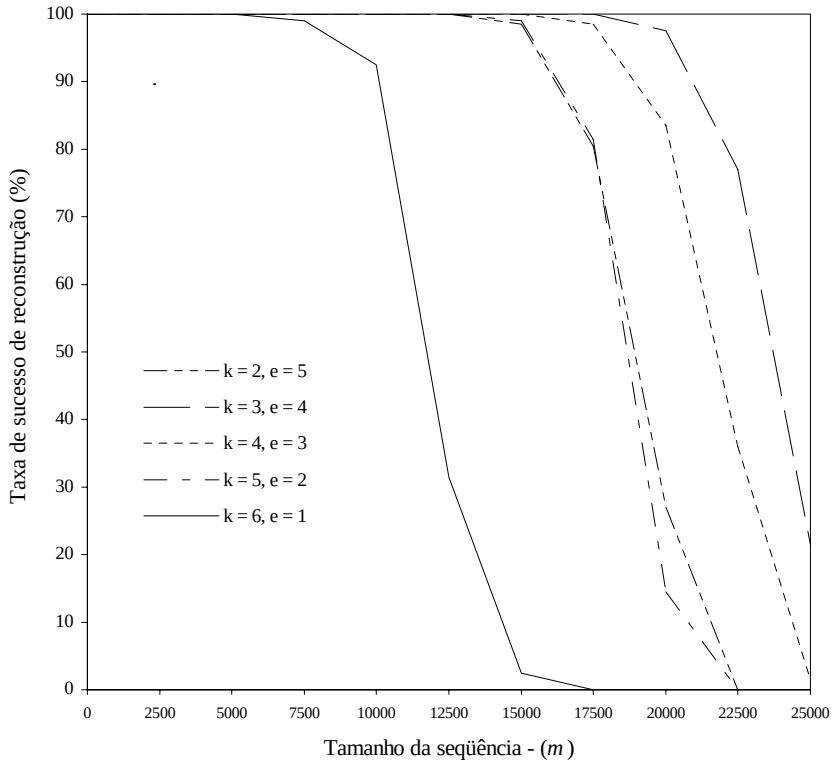


(b)

Figura 14A: Taxa de sucesso de reconstrução da seqüência *versus* o tamanho da seqüência, quando $k + e = 6$, $c = 10$, $p = q = 0{,}001$, (a) $n = 4^7$ e (b) $n = 4^8$.

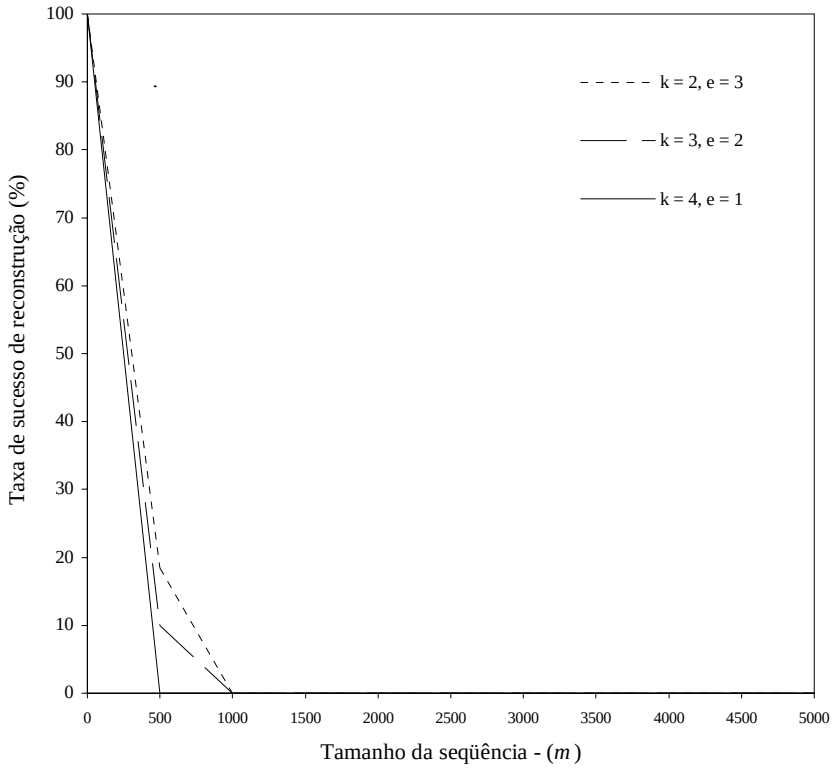


(a)

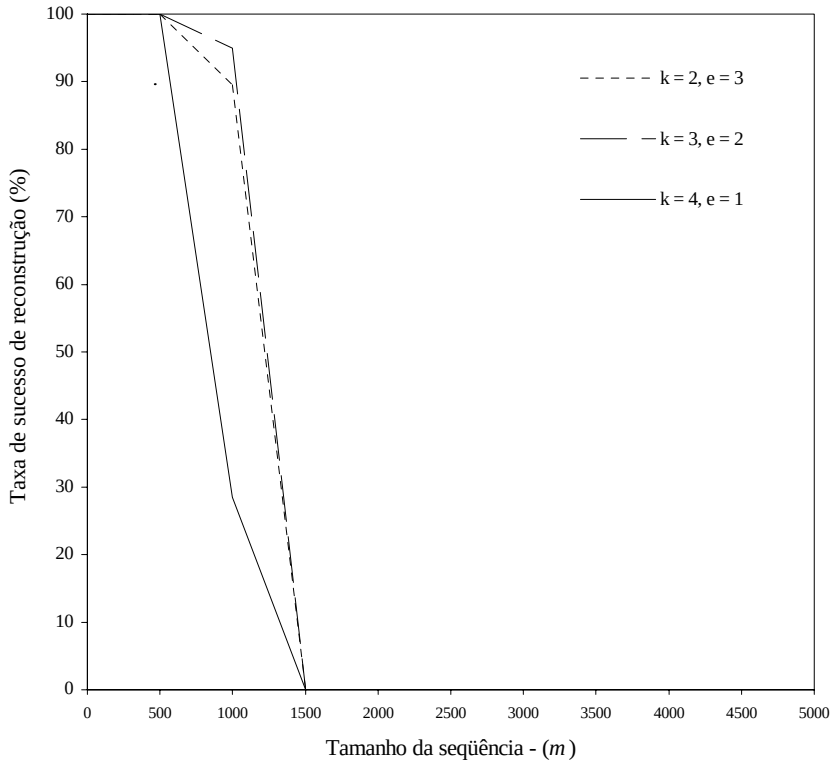


(b)

Figura 15A: Taxa de sucesso de reconstrução da seqüência *versus* o tamanho da seqüência, quando $k + e = 7$, $c = 10$, $p = q = 0{,}001$, (a) $n = 4^8$ e (b) $n = 4^9$.

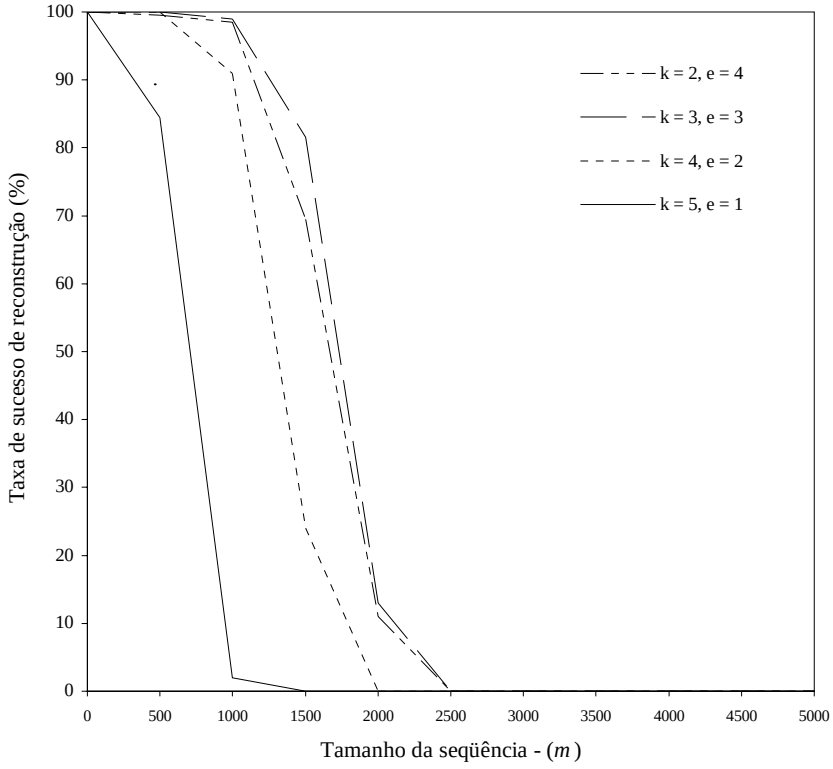


(a)

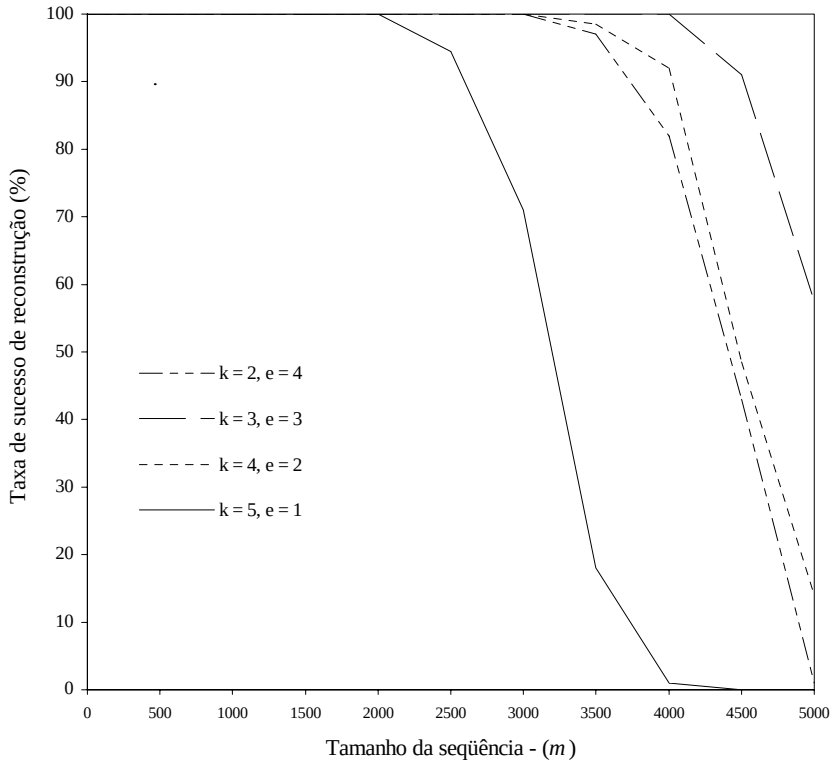


(b)

Figura 16A: Taxa de sucesso de reconstrução da seqüência *versus* o tamanho da seqüência, quando $k + e = 5$, $c = 10$, $p = q = 0{,}005$, (a) $n = 4^6$ e (b) $n = 4^7$.

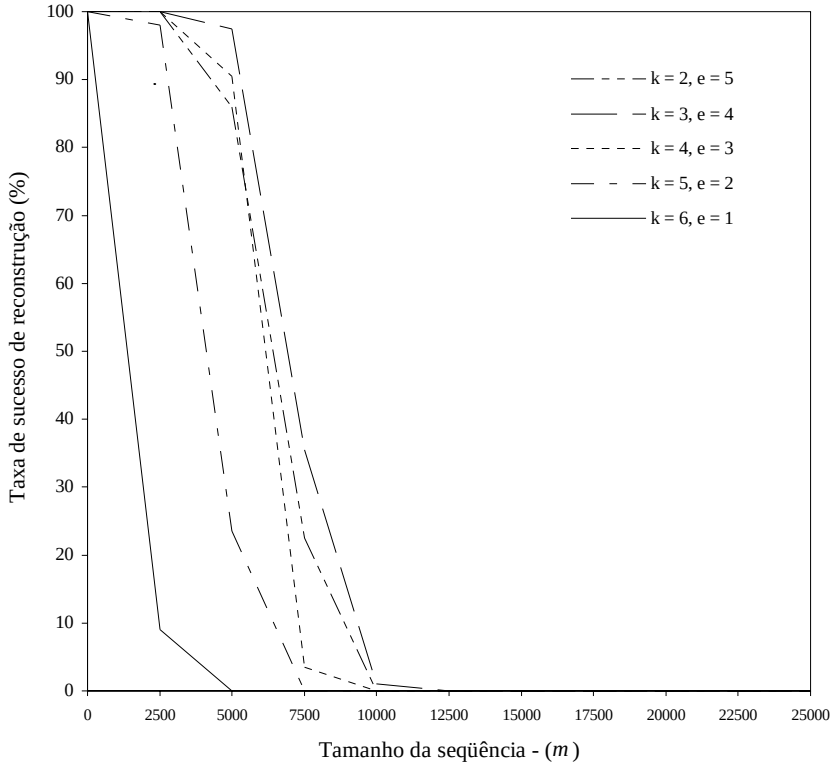


(a)

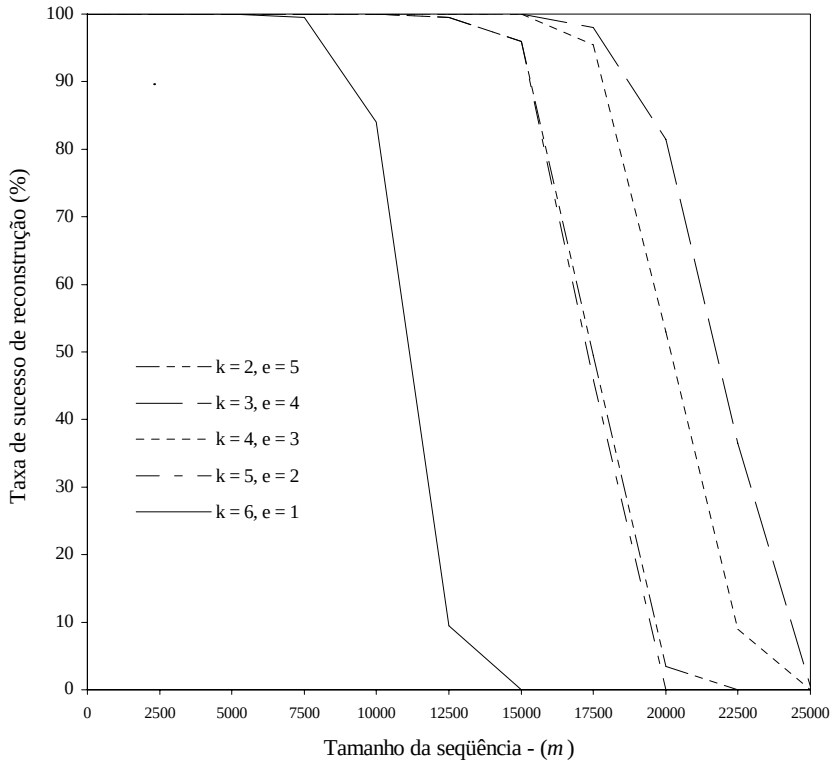


(b)

Figura 17A: Taxa de sucesso de reconstrução da seqüência *versus* o tamanho da seqüência, quando $k + e = 6$, $c = 10$, $p = q = 0{,}005$, (a) $n = 4^7$ e (b) $n = 4^8$.

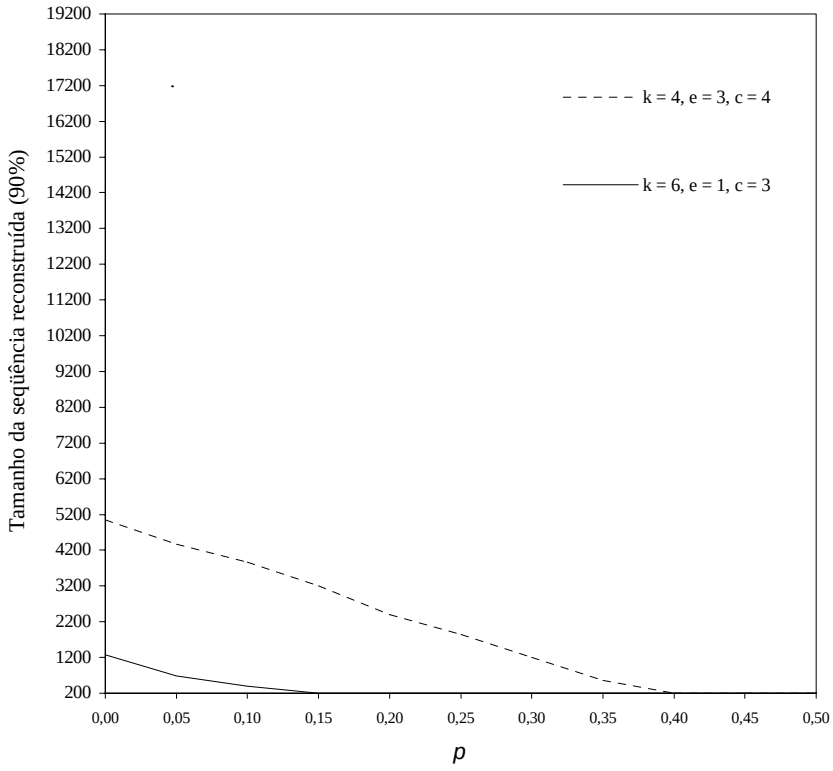


(a)

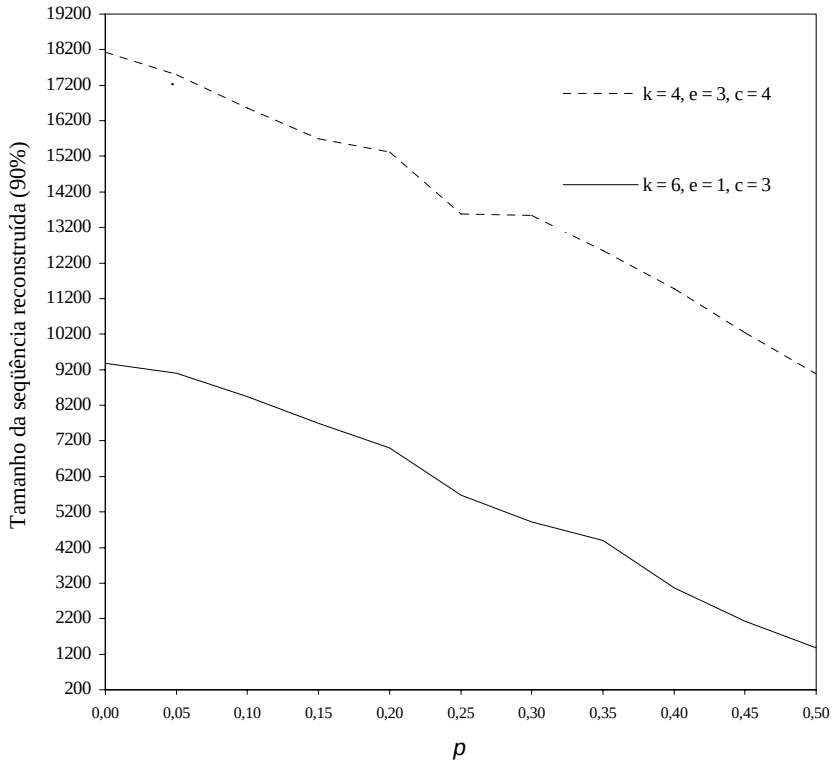


(b)

Figura 18A: Taxa de sucesso de reconstrução da seqüência *versus* o tamanho da seqüência, quando $k + e = 7$, $c = 10$, $p = q = 0{,}005$, (a) $n = 4^8$ e (b) $n = 4^9$.

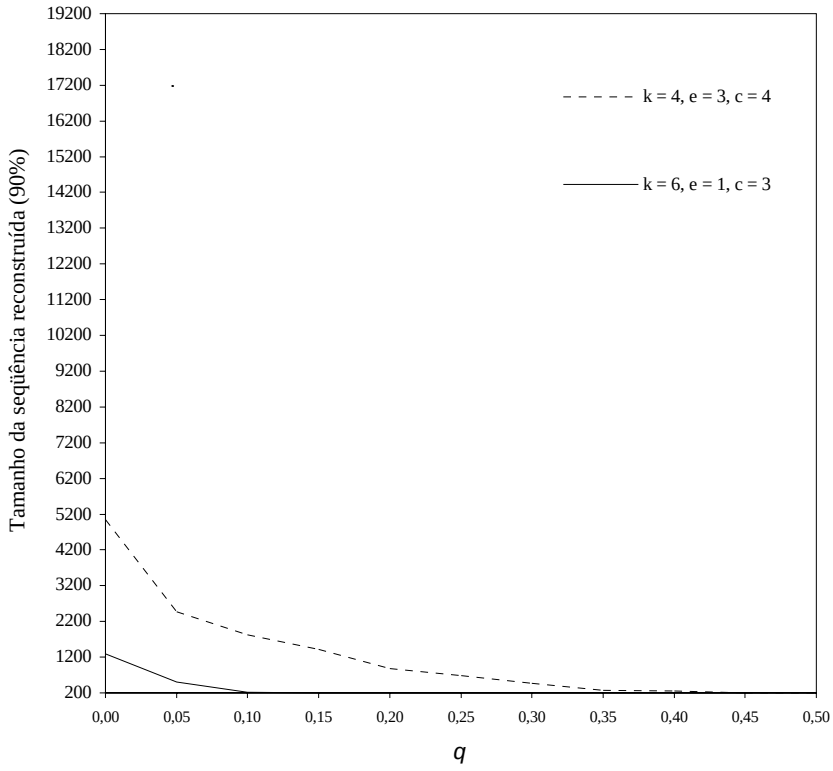


(a)

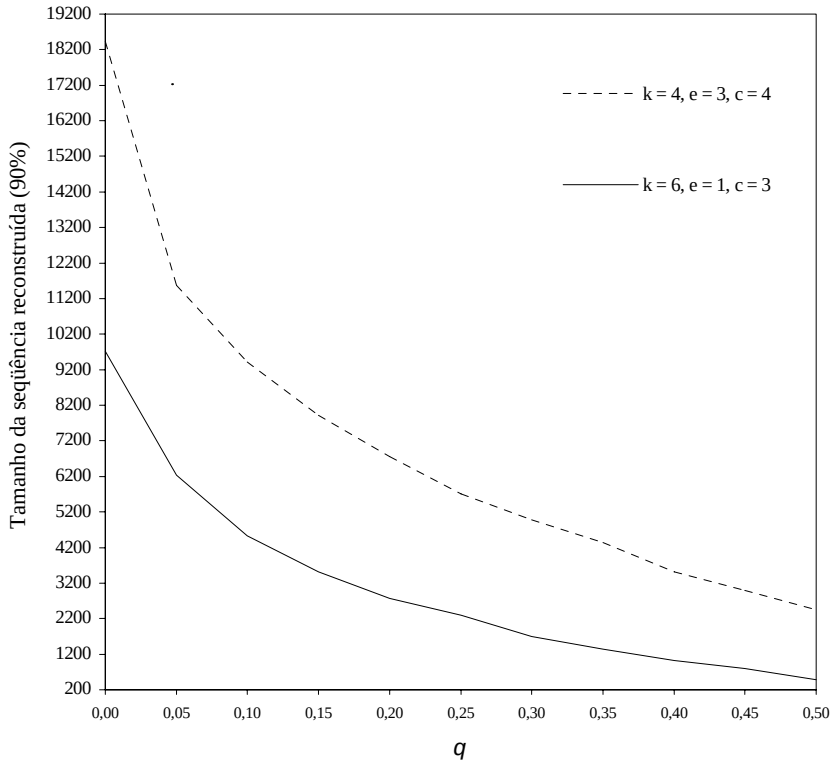


(b)

Figura 19A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* a taxa *p* de erro falso positivo, quando $k + e = 7$, $q = 0$, $ck + e = 19$, (a) $n = 4^8$ e (b) $n = 4^9$.

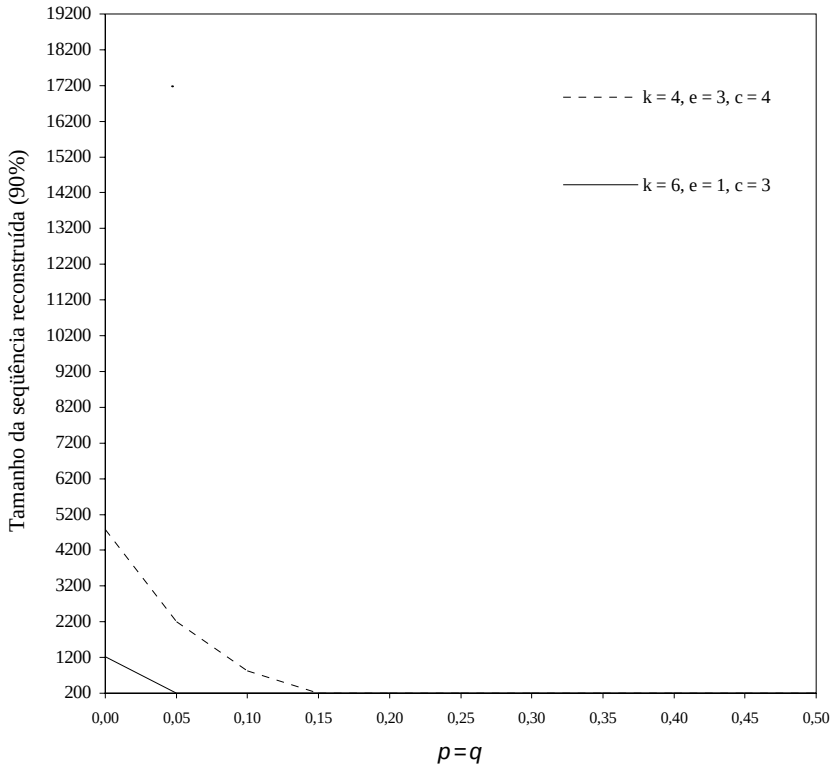


(a)

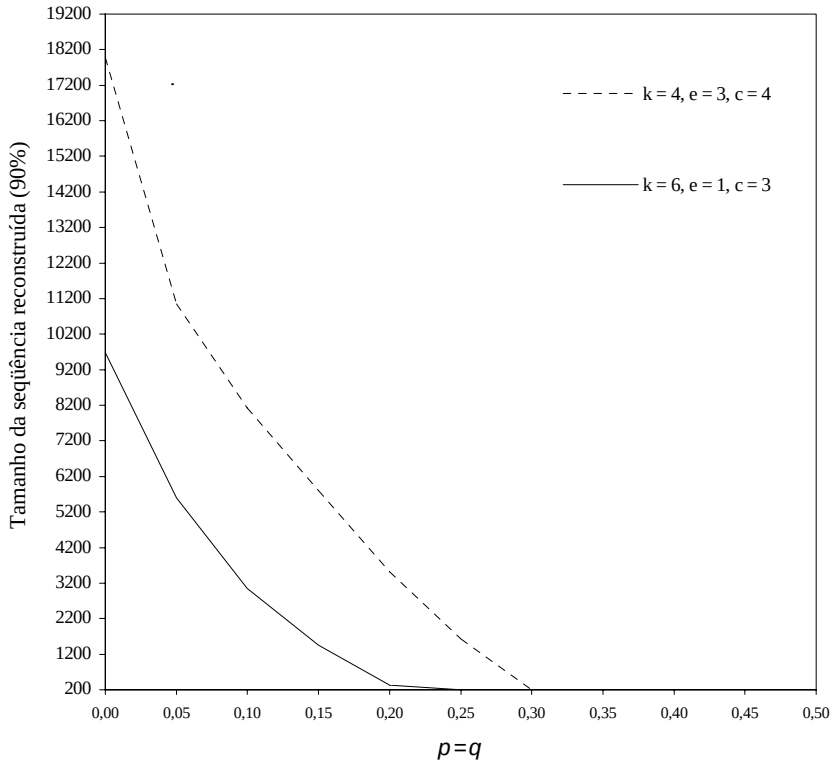


(b)

Figura 20A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* a taxa *q* de erro falso negativo, quando $k + e = 7$, $p = 0$, $ck + e = 19$, (a) $n = 4^8$ e (b) $n = 4^9$.

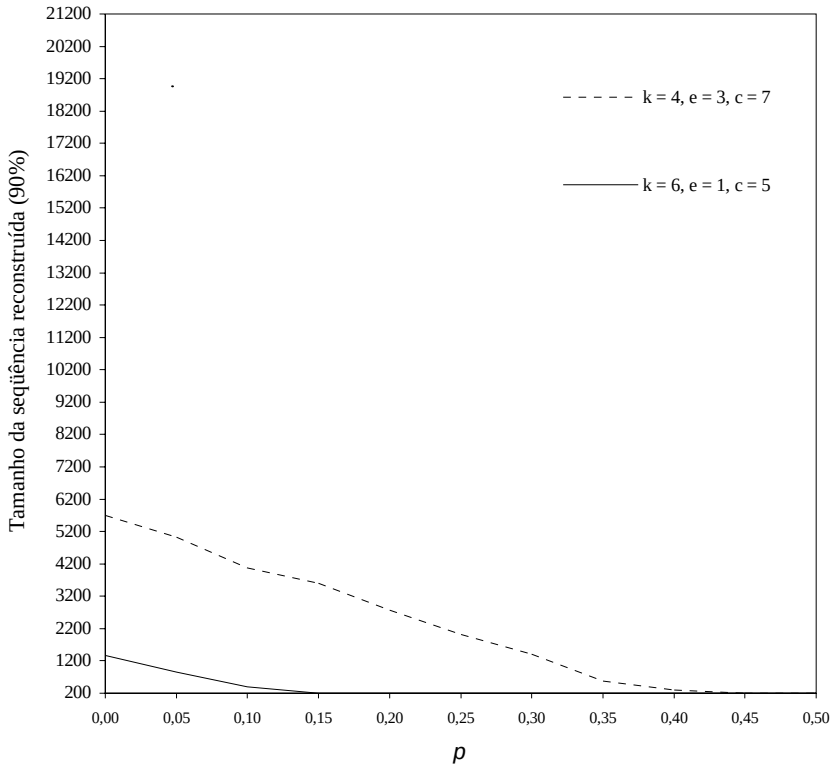


(a)

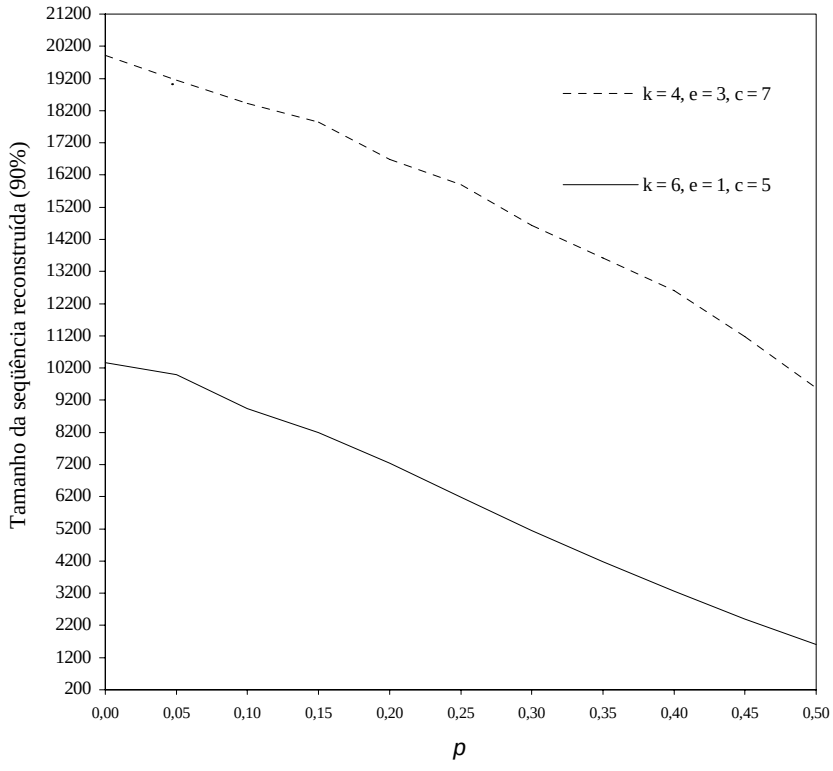


(b)

Figura 21A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* as taxas p e q ($p = q$) de erros, quando $k + e = 7$, $ck + e = 19$, (a) $n = 4^8$ e (b) $n = 4^9$.

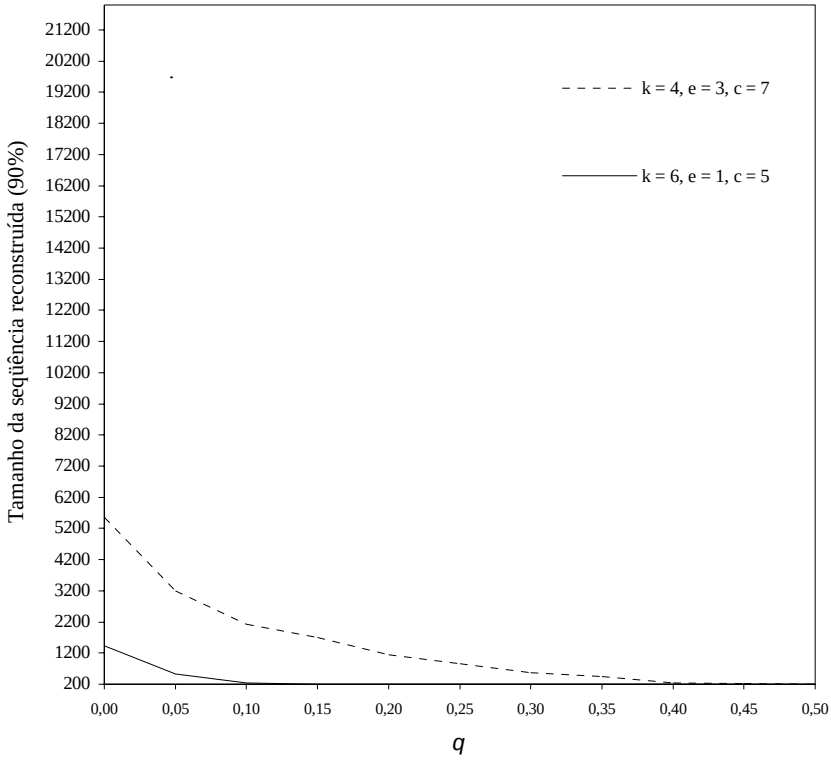


(a)

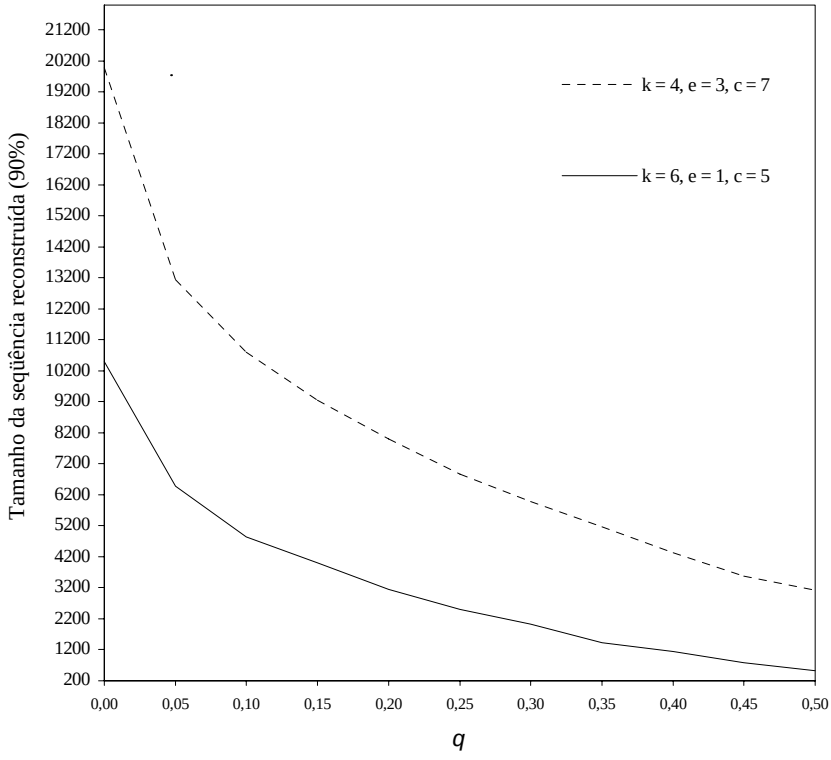


(b)

Figura 22A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* a taxa *p* de erro falso positivo, quando $k + e = 7$, $q = 0$, $ck + e = 31$, (a) $n = 4^8$ e (b) $n = 4^9$.

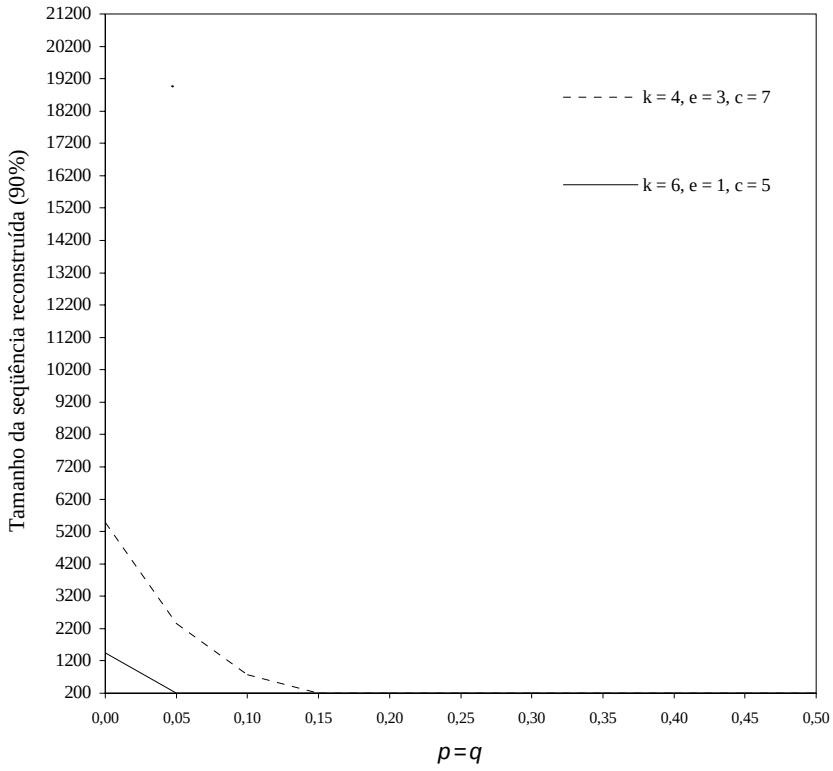


(a)

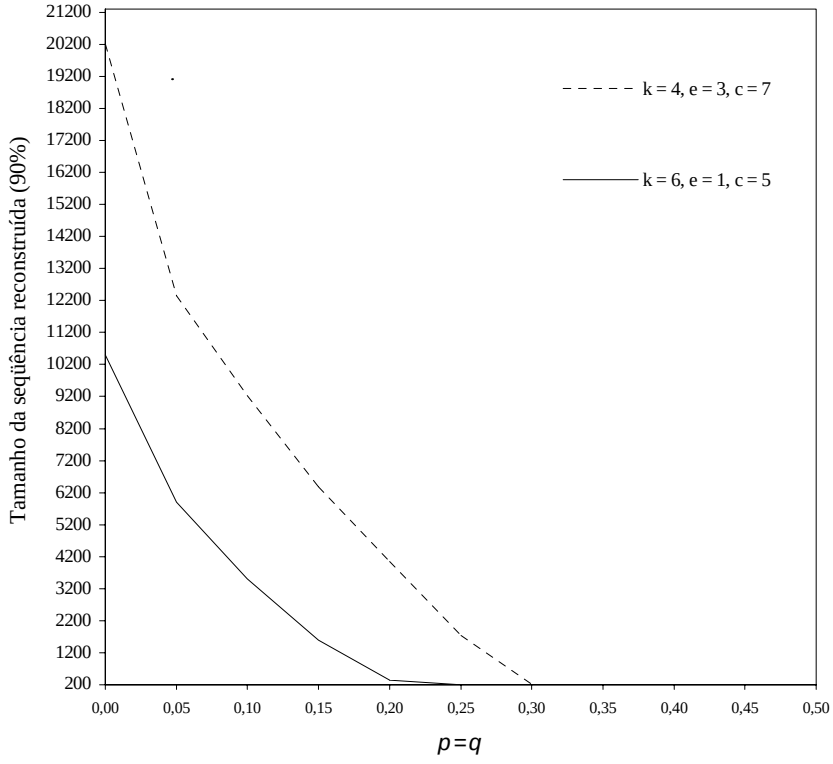


(b)

Figura 23A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* a taxa *q* de erro falso negativo, quando $k + e = 7$, $p = 0$, $ck + e = 31$, (a) $n = 4^8$ e (b) $n = 4^9$.

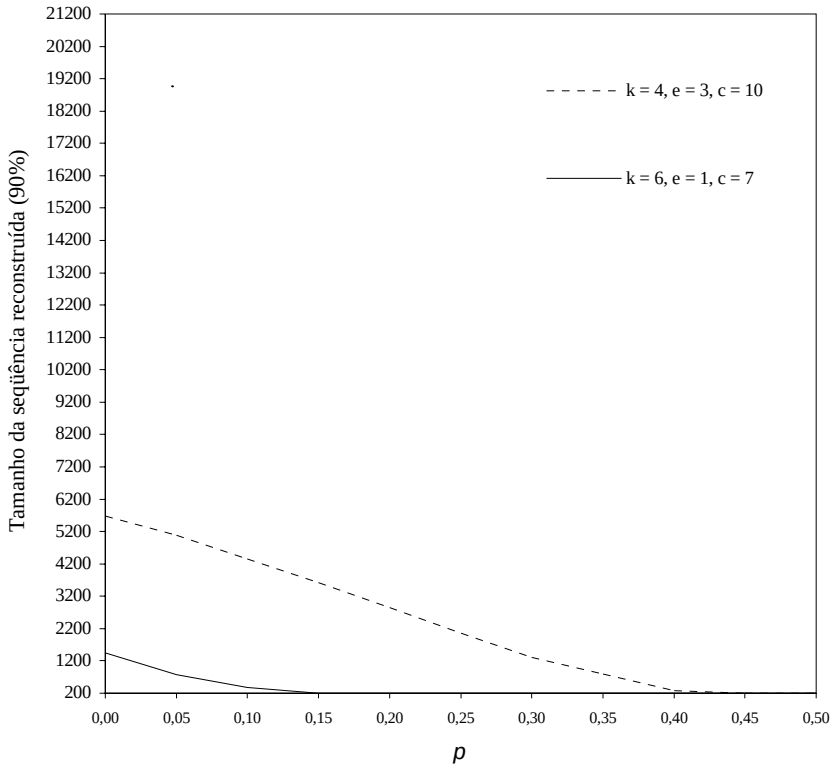


(a)

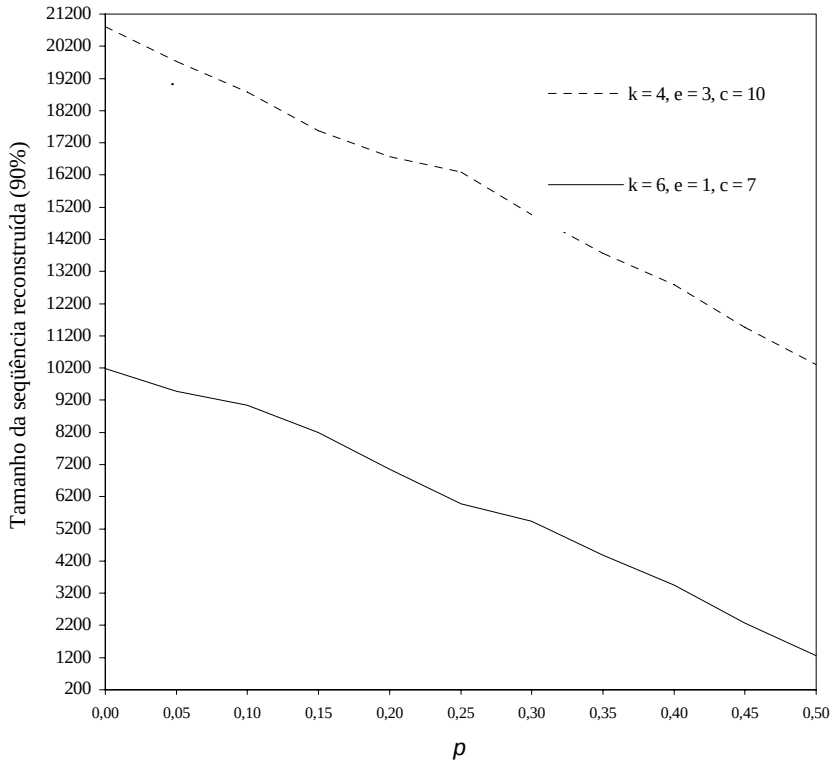


(b)

Figura 24A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* as taxas p e q ($p = q$) de erros, quando $k + e = 7$, $p = 0$, $ck + e = 31$, (a) $n = 4^8$ e (b) $n = 4^9$.

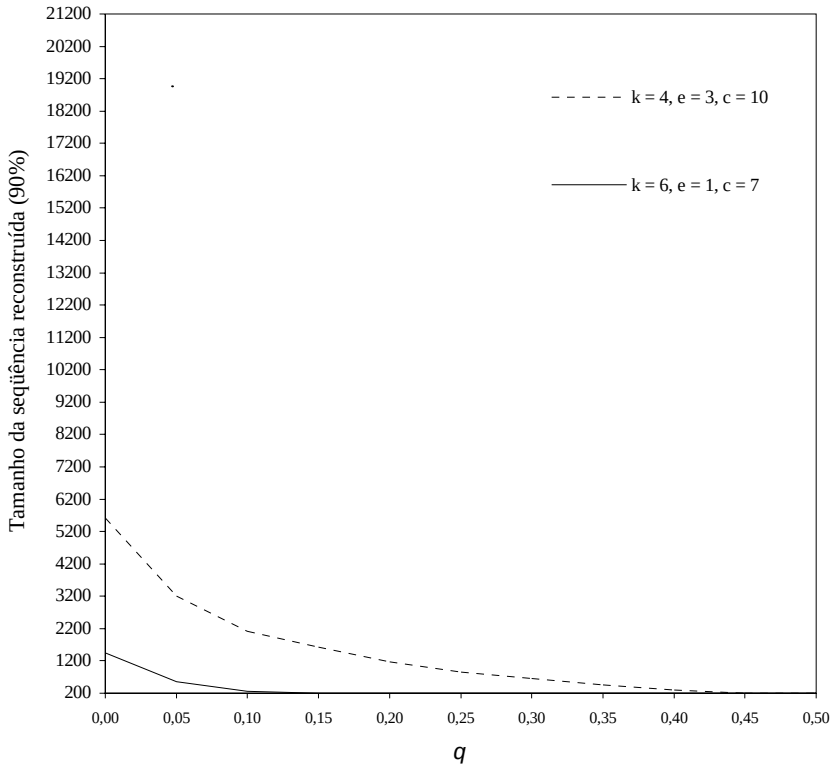


(a)

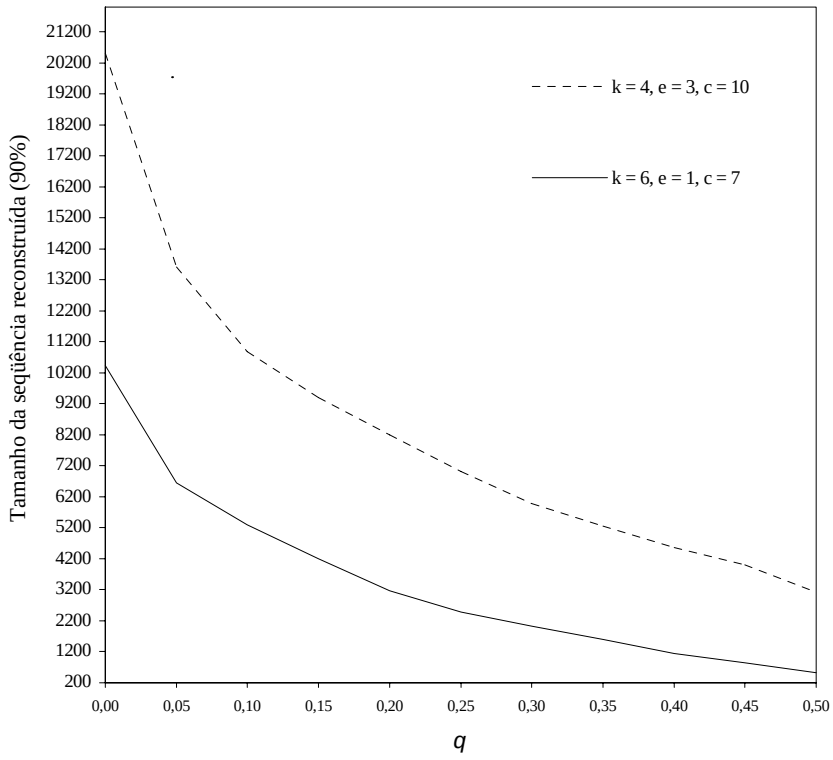


(b)

Figura 25A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* a taxa *p* de erro falso positivo, quando $k + e = 7$, $q = 0$, $ck + e = 43$, (a) $n = 4^8$ e (b) $n = 4^9$.

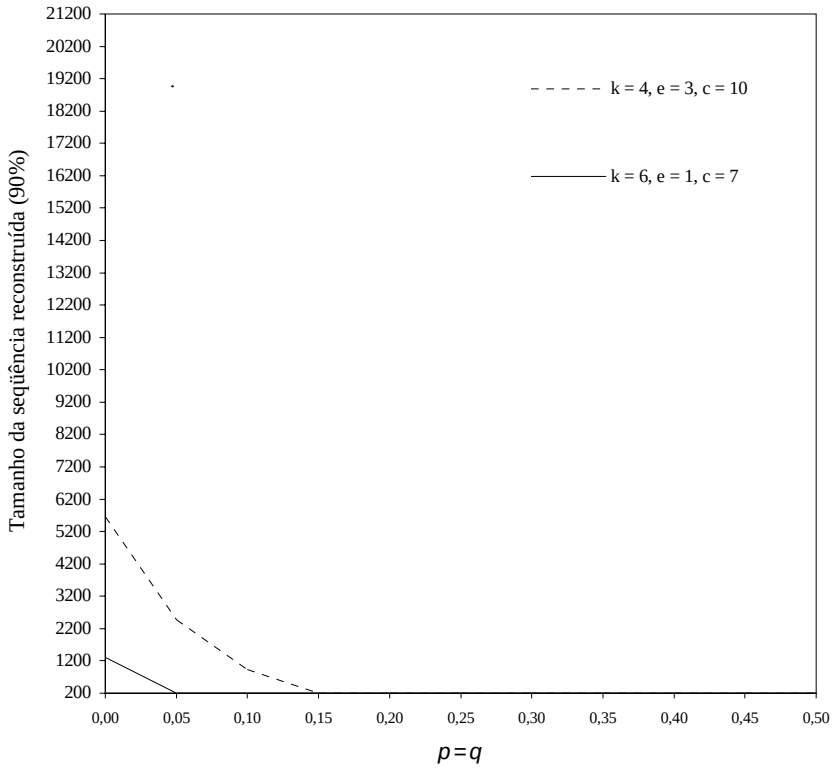


(a)

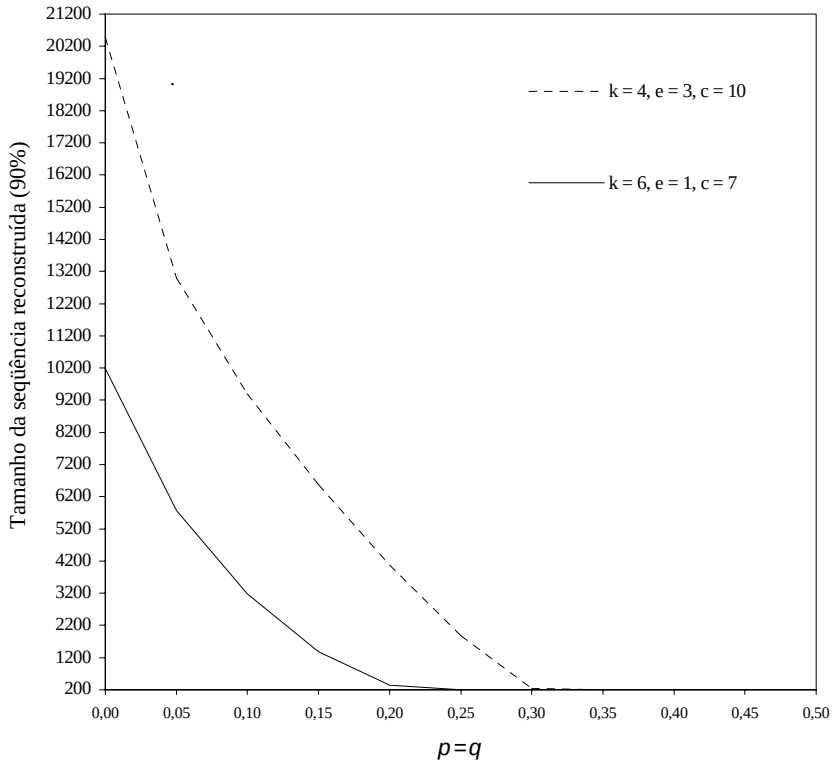


(b)

Figura 26A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* a taxa *q* de erro falso negativo, quando $k + e = 7$, $p = 0$, $ck + e = 43$, (a) $n = 4^8$ e (b) $n = 4^9$.

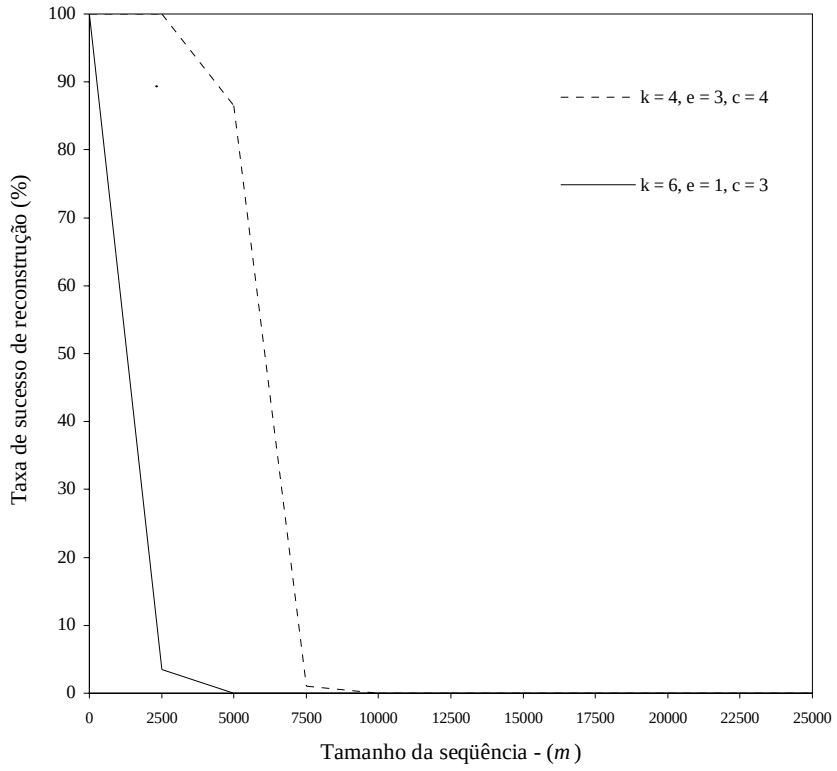


(a)

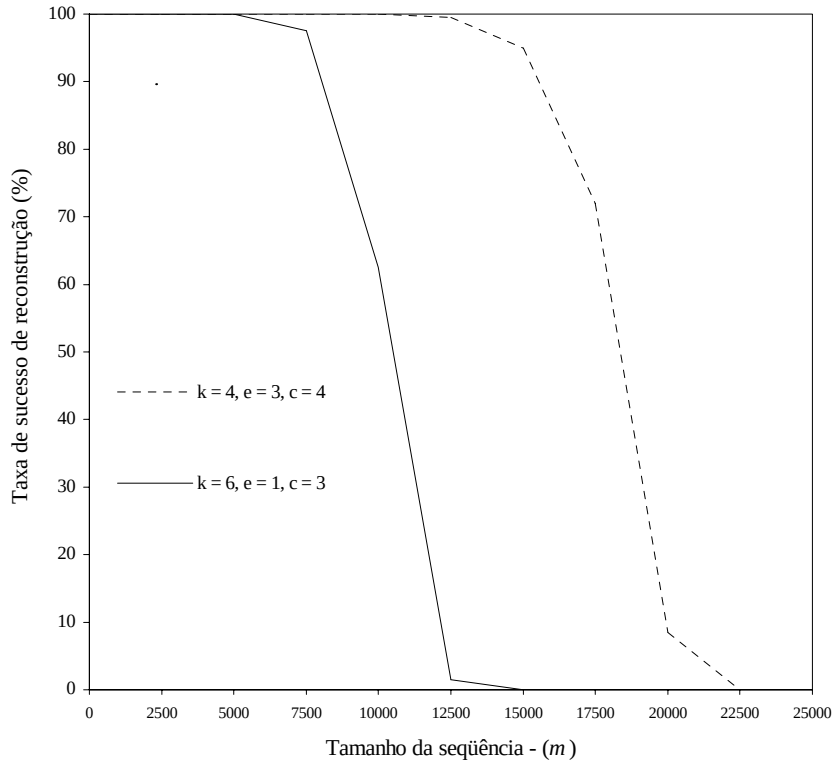


(b)

Figura 27A: Tamanho da seqüência corretamente reconstruída em aproximadamente 90% dos casos *versus* as taxas p e q ($p = q$) de erros, quando $k + e = 7$, $ck + e = 43$, (a) $n = 4^8$ e (b) $n = 4^9$.

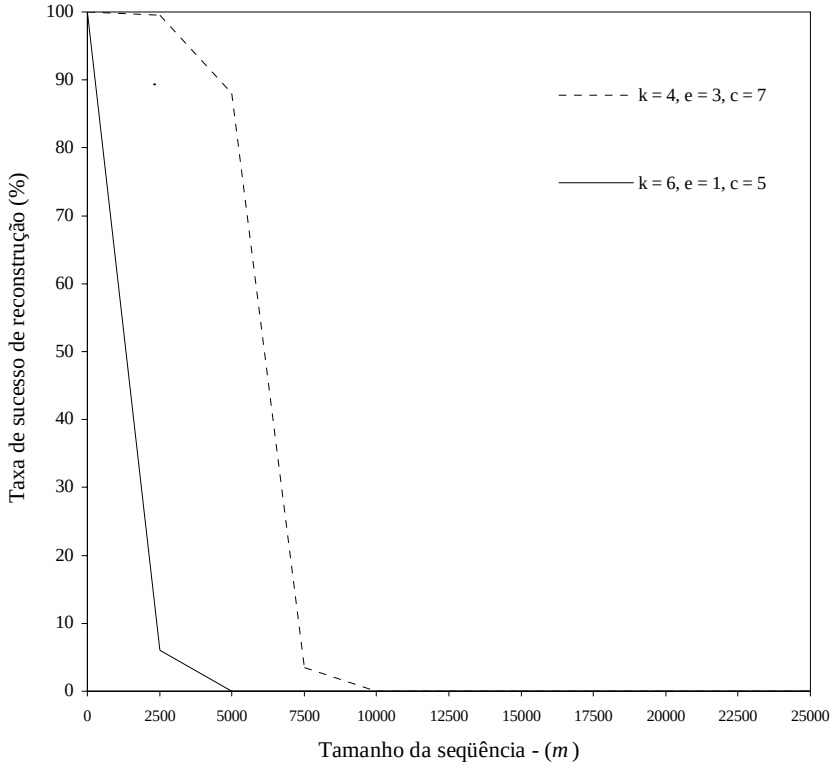


(a)

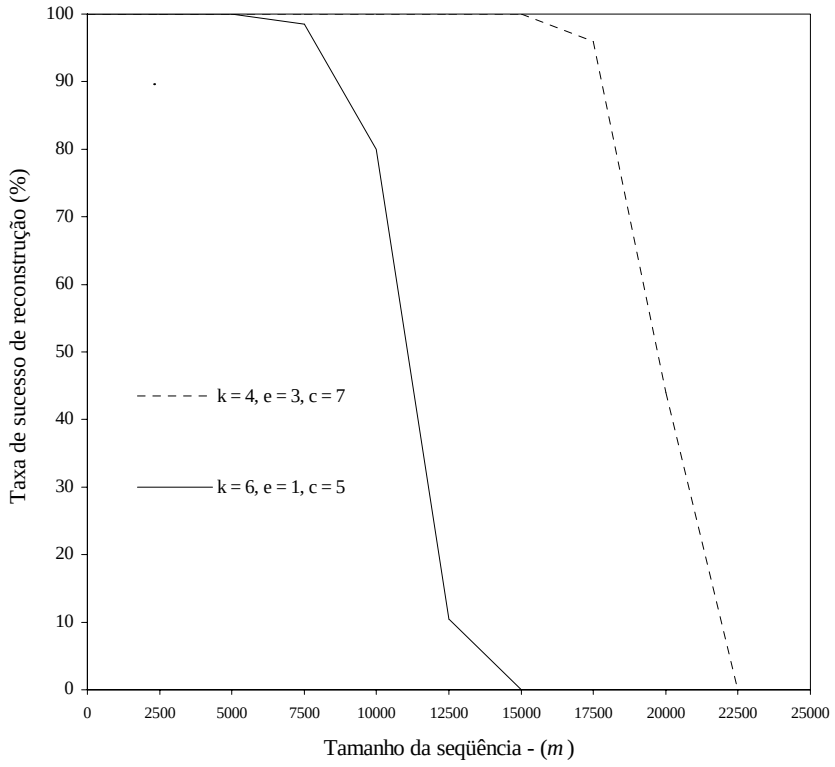


(b)

Figura 28A: Taxa de sucesso de reconstrução da seqüência *versus* o tamanho da seqüência, quando $k + e = 7$, $ck + e = 19$, $p = q = 0{,}005$, (a) $n = 4^8$ e (b) $n = 4^9$.

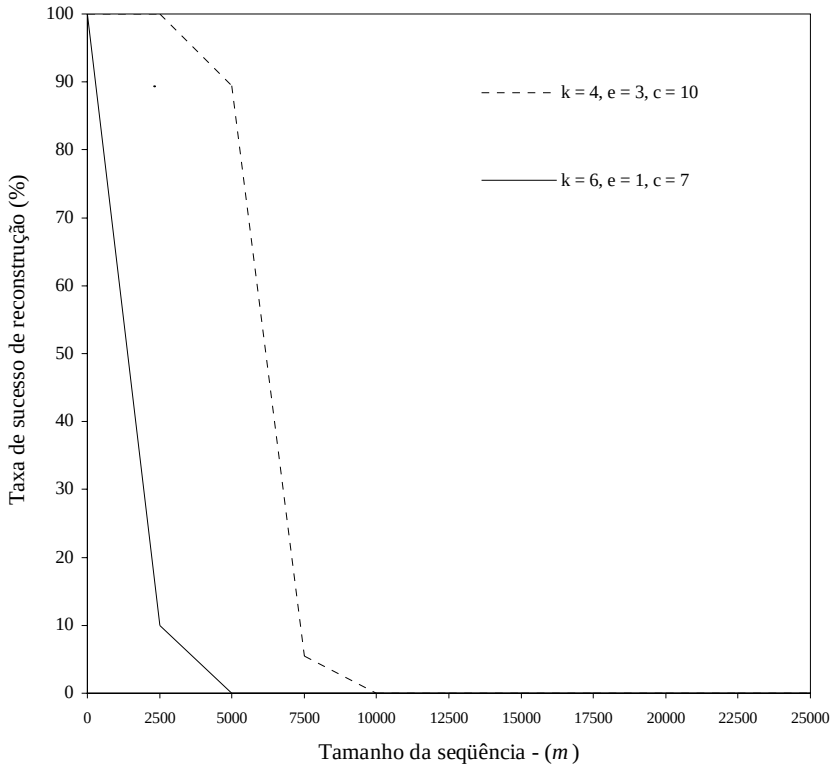


(a)

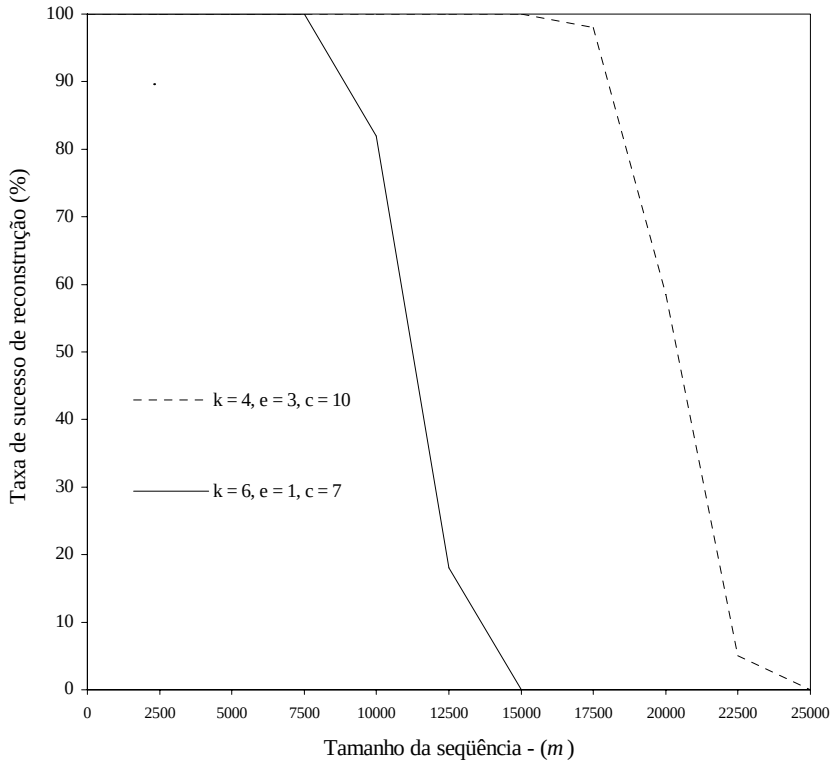


(b)

Figura 29A: Taxa de sucesso de reconstrução da seqüência *versus* o tamanho da seqüência, quando $k + e = 7$, $ck + e = 31$, $p = q = 0{,}005$, (a) $n = 4^8$ e (b) $n = 4^9$.



(a)



(b)

Figura 30A: Taxa de sucesso de reconstrução da seqüência *versus* o tamanho da seqüência, quando $k + e = 7$, $ck + e = 43$, $p = q = 0{,}005$, (a) $n = 4^8$ e (b) $n = 4^9$.

APÊNDICE B – Trabalho Preliminar Publicado

Uma Abordagem Alternativa para Seqüenciamento por Hibridização

Ennio dos Santos Baptista[1,2], Katia Silva Guimarães[3]

[1]Programa de Pós-Graduação em Engenharia Elétrica
Instituto de Tecnologia da Amazônia – UTAM
CEP 69.050.020 – Manaus – AM

[2]Programa de Pós-Graduação em Engenharia Elétrica
Departamento de Eletrônica e Sistemas
Universidade Federal de Pernambuco – UFPE
Caixa Postal 7800, Cidade Universitária – 50711-970 – Recife – PE

[3]Centro de Informática
Universidade Federal de Pernambuco – UFPE
Caixa Postal 7851, Cidade Universitária – 50732-970 – Recife – PE

{esb, katia}@cin.ufpe.br

Abstract. *Recently, Halperin et al. presented two new approaches to sequencing by hybridization and suggested that a new strategy which combines various aspects of the two approaches might generate better practical results. In this article, we address the problem by implementing the above-stated approach and present an alternative one, which tends to generate more positive results, including examples stemming from its experimental application.*

Resumo. *Recentemente, Halperin et al. apresentou duas novas abordagens para o seqüenciamento por hibridização e sugeriu que uma nova estratégia, combinando aspectos particulares delas, talvez pudesse poderia gerar resultados práticos melhores. Neste artigo, apontamos o problema de se implementar tal sugestão e apresentamos uma proposta alternativa que tende a superá-lo, bem como alguns resultados decorrentes de sua avaliação experimental.*

1. Introdução

1.1. Seqüenciamento por Hibridização

Uma questão central na emergente área da Biologia Computacional diz respeito ao seqüenciamento de DNA. Seqüenciar uma cadeia de DNA significa determinar a ordem dos seus nucleotídeos componentes – adenina (*A*), citosina (*C*), guanina (*G*) e timina (*T*). A importância dessa tarefa é que ela é o primeiro passo em direção à compreensão dos mecanismos bioquímicos que determinam as características estruturais e funcionais de cada ser vivo. Drmanac e Drmanac (2001) afirmam que, "pela determinação da seqüência do DNA de um organismo, pesquisadores podem obter informações críticas sobre o seu desenvolvimento, a sua fisiologia, as suas relações taxonômicas, e a sua susceptibilidade a doenças".

No fim dos anos 80, quatro grupos de pesquisadores propuseram um método alternativo de seqüenciamento para o então consagrado paradigma *overlap-layout-consensus*. A idéia por trás desse novo método, denominado Seqüenciamento por Hibridização (*Sequencing by Hybridization – SBH*), é de que uma cadeia de DNA pode ser seqüenciada a partir do conjunto de todas as suas subseqüências de tamanho *k*.

O SBH é um procedimento com duas etapas complementares bem definidas. A primeira etapa envolve experimentos bioquímicos. Emprega-se um *chip* de seqüenciamento – DNA *chip* – com uma superfície estruturada na forma de matriz, sobre a qual são fixados, em posições individuais e conhecidas, todos os oligonucleotídeos de tamanho *k*, ou *probes*. Esse *chip* é colocado em contato com uma solução contendo cópias da cadeia de DNA a ser seqüenciada – seqüência alvo. Em seguida, é lavado, e somente aquelas cópias que tiverem hibridizado com algum de seus *probes* permanecerão ligadas à sua superfície. A hibridização ocorrerá sempre que o *probe* for complemento, segundo Watson-Crick, de alguma subseqüência de tamanho *k* da cadeia. Finalmente, um procedimento adequado de leitura do *chip* revela o conjunto de todos os *probes* que hibridizaram. Esse conjunto denomina-se *spectrum* da seqüência. A segunda etapa do SBH é uma etapa combinatorial que emprega recursos computacionais para determinar a ordem em que os *probes* de um dado *spectrum* devem ser considerados na reconstrução da seqüência alvo. Essa é a etapa onde o seqüenciamento efetivamente ocorre.

A grande limitação do método é que podem existir diversas seqüências com o mesmo *spectrum*. O desafio, pois, do procedimento algorítmico é superar essa ambigüidade e descobrir, em uma situação específica, qual das seqüências candidatas é a seqüência alvo.

1.2. Propostas Anteriores

Diversas abordagens têm sido apresentadas para a reconstrução da seqüência alvo a partir do seu *spectrum*. O *chip* de seqüenciamento clássico, $C(k)$, que contém todos os 4^k *probes* de tamanho *k*, é o tradicionalmente usado.

Pevzner (1989) propôs uma abordagem matematicamente elegante, equiparando o problema de reconstrução da seqüência alvo ao problema de se encontrar um Caminho Euleriano em um grafo construído adequadamente com os *probes* gerados com esse *chip*. Contudo, para que essa abordagem funcione, deve-se assumir que a multiplicidade de cada *probe* é conhecida e que o experimento bioquímico é perfeito, gerando um *spectrum* livre de erros, tanto dos *probes* falsos positivos (*probes* presentes indevidamente no *spectrum*), quanto dos *probes* falsos negativos (*probes* ausentes indevidamente do *spectrum*). Na verdade, essas considerações são irreais e fazem com que essa abordagem não encontre aplicabilidade prática imediata.

Nessa estratégia, a ambigüidade continua sendo um fator crítico. Pevzner (1991) provou isso, demonstrando que, para o *chip* clássico $C(8)$, com 65536 *probes*, a taxa de sucesso na reconstrução de seqüências com apenas 200 bases é de no máximo 94%.

Várias outras estratégias foram propostas para superar o problema de reconstruções ambíguas. Dentre elas estão as que sugerem projetos alternativos para os *chips* de seqüenciamento, como é o caso da proposta de Preparata et al. (1999) que, de certa forma, serviu de base para os *chips* discutidos neste trabalho. Nela, os autores

apresentaram um novo esquema de *chip* que usa bases universais na construção dos seus *probes*, além das bases naturais – *A*, *C*, *G* e *T*. Esse tipo diferente de base tem a característica importante de hibridizar com qualquer uma das 4 bases naturais. Para esse novo modelo de *chip*, os autores apresentaram um algoritmo que, com alta probabilidade, aproxima-se do limite teórico dentro de um fator constante. Doi e Imai (2000) aperfeiçoaram esse algoritmo fazendo com que ele passasse a funcionar também com *spectrum* contendo erros.

Em artigo publicado recentemente, Halperin et al. (2002) apresentam um algoritmo, denominado algoritmo **A**, que roda em tempo polinomial sobre dados gerados com o *chip* clássico. Esse algoritmo foi apresentado como uma solução para o problema de número 35 da lista de problemas em aberto na biologia elaborada por Pevzner e Waterman (1995). Nesse mesmo artigo, os autores apresentaram um outro esquema para o *chip* de seqüenciamento, que, a exemplo do de Preparata et al. (1995), também usa bases universais. A diferença é que, nessa nova proposta, essas bases são dispostas ao longo dos *probes* de forma randômica, e não em posições fixas, como anteriormente. Para esse *chip* foi proposto um outro algoritmo, denominado algoritmo **B**, muito mais resistente a erros do que o de Doi e Imai (2000). Finalmente, levantaram a hipótese de que talvez uma combinação adequada do algoritmo **A** com o algoritmo **B**, à qual chamamos de algoritmo **AB,** venha a produzir ainda melhores resultados práticos do que o algoritmo **B** original.

Neste trabalho, apontamos a dificuldade encontrada na implementação dessa proposta e realizamos um estudo experimental com um esquema alternativo de *chip* e um algoritmo que visam superá-la.

A exemplo de outros trabalhos, usamos o símbolo "*" para representar as bases universais; e p e q para representar, respectivamente, a probabilidade de um *probe* ser falso positivo e de ser falso negativo. Da mesma forma, consideramos que no início da montagem da seqüência alvo, o seu prefixo de tamanho i - 1, onde i é o tamanho do *probe*, é dado.

2. As Propostas de Halperin et al.

2.1. O Algoritmo A para o *Chip* Clássico

O algoritmo **A** proposto para o *chip* clássico $C(k)$, apesar de ser bastante trivial, pode lidar com erros e com o problema de ambigüidade. Além disso não pressupõe o conhecimento da multiplicidade dos *probes*. Considerando que já conhecemos a seqüência S, desde s_1 até s_i, o algoritmo determina a base s_{i+1} executando os seguintes passos:

1. gerar todas as 4^k seqüências $a_1 = a_1,...,a_k$ como extensão de $s_1,...,s_i$;
2. para cada seqüência a', contar o número de *probes* do *spectrum* que casam com a subseqüência $s_{i-k+2},...,s_i,a'_1,...,a'_k$ ou *caminho*;
3. para a seqüência vencedora, fazer $s_{i+1} = a'_1$ (o empate é resolvido arbitrariamente).

Para esse algoritmo, o tempo de execução é dado por $O(4^k km)$, onde m é o tamanho da seqüência alvo.

A Figura 1 esquematiza a reconstrução da posição 9 da seqüência S para $k = 8$. De acordo com o algoritmo, temos de gerar todas as extensões no trecho entre a posição 9 e a posição 16 e contar, para cada uma delas, o número de *probes* que casam com o *caminho* de 2 a 16. Percebemos que nesse exemplo existem 8 *probes* para a extensão GAGTTAGT.

Uma característica importante do algoritmo **A** para o nosso trabalho é que, ao contar o número de *probes* pertencentes ao *caminho*, ele se apóia na informação de *probes* que estão relacionados a posições que ainda serão, no futuro, determinadas por ele. De outra forma, ele "olha" adiante na cadeia para confirmar a posição atual. Isso pode ser feito porque a natureza de construção do *chip* clássico $C(k)$ garante que se dois *probes* estão em posições adjacentes na seqüência. Então, as k - 1 bases finais de um se sobrepõem às k-1 bases finais do outro. É exatamente essa idéia de *caminho* que pode ser aproveitada para melhorar o algoritmo **B**. Por outro lado, o primeiro passo do algoritmo determina que sejam geradas todas as 4^k extensões de tamanho k. Certamente esse é o maior problema dessa abordagem.

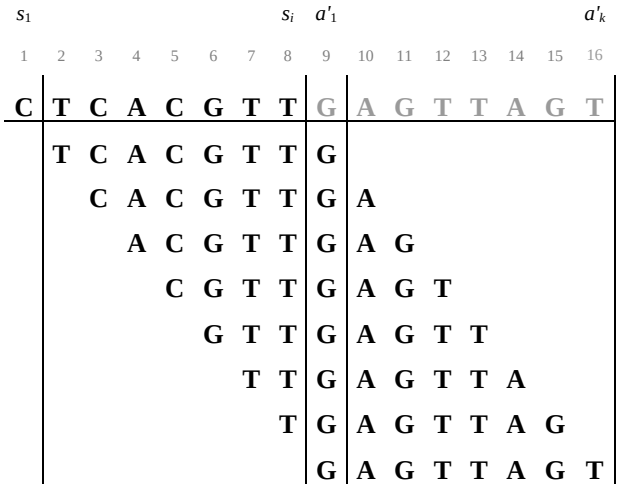


Figura 1. Esquema de execução do algoritmo A para o *chip* $C(8)$.

2.2. Halperin et al. e o Novo *Chip*

Halperin et al. apresentou um projeto de *chip* de seqüenciamento que, a exemplo do projeto de Preparata et al. (1999), também emprega bases universais. Esse novo *chip* tem as seguintes características:

a. cada um de seus *probes* é composto de duas partes: uma parte inicial com k bases naturais posicionadas aleatoriamente e separadas por bases universais com um fator c de distanciamento, onde c é um inteiro convenientemente grande; e uma parte final com uma base natural. Logo, o tamanho total de cada *probe* é dado por $ck+1$;

b. formamos βk famílias de probes, onde β, uma constante inteira maior que 1, depende do nível de erro. Para cada família selecionamos aleatoriamente um conjunto de k posições no intervalo de 1 a ck e formamos todos os 4^{k+1}

pobres possíveis colocando as bases *A*, *C*, *G* e *T* nas *k* posições selecionadas e na última;

c. o número *n* total de probes no chip é dado por $\beta k 4^{k+1}$.

2.3. O Algoritmo B para o Novo *Chip*

O algoritmo proposto para esse novo *chip* também é bastante trivial. A exemplo do algoritmo anterior, ele lida com erros no *spectrum* e com o problema de ambigüidade, e não pressupõe o conhecimento da multiplicidade dos *probes*. Considerando que já se conhece a seqüência *S* desde s_1 até s_i , ele funciona da seguinte maneira para determinar a base s_{i+1}:

1. para cada base natural *X*, contar quantos *probes* casam com a subseqüência $s_{i-ck+1},...,s_i, X$;

2. para a base vencedora, fazer $s_{i+1} = X$ (o empate é resolvido arbitrariamente);

O tempo de execução desse algoritmo é dado por $O(km)$, onde *m* é o tamanho da seqüência alvo.

A Figura 2 esquematiza o funcionamento do algoritmo **B** para $c = 2$ e $k = 3$. Ela apresenta uma seqüência com as 8 primeiras bases conhecidas. Deseja-se determinar a base da posição 9. De acordo com o algoritmo, devemos atribuir as 4 bases naturais a essa posição e contar, para cada uma, quantos *probes* casam com a subseqüência que vai da posição 3 até a posição 9. Percebemos que, para a base G, existem 3 *probes*.

s_1							s_i	X
1	2	3	4	5	6	7	8	9
C	**T**	**C**	**A**	**C**	**G**	**T**	**T**	**G**
		C	*****	*****	**G**	*****	**T**	**G**
		*****	**A**	*****	**G**	**T**	*****	**G**
		C	*****	**C**	*****	*****	**T**	**G**

Figura 2. Esquema de execução do algoritmo B para um chip com $c = 2$ e $k = 3$.

A característica principal dessa técnica consiste em termos um *chip* dividido em várias famílias de *probes*, conseqüência do uso de bases universais posicionadas aleatoriamente. A vantagem é que agora não temos no *spectrum* somente um *probe* hibridizando com cada subseqüência da cadeia alvo, mas, sim, uma quantidade de *probes* igual ao número de famílias do *chip*. Essa característica torna o método mais resistente a erros.

2.4. O algoritmo AB para o Novo *Chip*

Halperin et al. levantou a hipótese de que provavelmente chegaremos a resultados práticos melhores se fizermos o algoritmo **B** semelhante ao algoritmo **A**, isto é, se ao invés de contarmos o número de *probes* que casam com cada uma das 4 extensões possíveis, passássemos a contar o número de *probes* que casam com cada *caminho* que estende a seqüência alvo. Esse novo algoritmo também é bastante simples e, da mesma forma que os anteriores, lida com erros, ambigüidade e não pressupõe conhecimento sobre a multiplicidade dos probes. O seu tempo de execução é dado por $O(4^{ck}k^2cm)$.

Considerando que já se conhece a seqüência *S* desde s_1 até s_i, ele executa os seguintes passos para determinar a base s_{i+1}:

1. gerar todas as 4^{ck+1} seqüências $a = a_1,...,a_{ck+1}$ como extensão de $s_1,...,s_i$;
2. para cada seqüência a', contar o número de *probes* do *spectrum* que casam com a subseqüência $s_{i-ck+1},...,s_i, a'_1,...,a'_{ck+1}$ ou *caminho*;
3. para a seqüência vencedora, fazer $s_{i+1} = a'_1$ (o empate é resolvido arbitrariamente).

A Figura 3 esquematiza o funcionamento do algoritmo **AB**. Ela apresenta uma seqüência com as 6 primeiras bases conhecidas. Deseja-se determinar a base da posição 7. De acordo com o primeiro passo do algoritmo, devemos gerar todas as extensões possíveis no trecho que vai de 7 a 11. Para a extensão TTGAC, percebemos que existem 5 grupos de 2 *probes* casando com o caminho que vai desde a posição 3 até a posição 11, sendo que 6 deles indicam a base T para a posição 7.

Construído dessa forma, o algoritmo **AB** herda do algoritmo **A** a idéia de usar *caminhos* na contagem de *probes*; e do algoritmo **B**, a característica de usar várias famílias de *probes*.

s_1					s_i	a_1				a_{ck+1}
1	2	3	4	5	6	7	8	9	10	11
C	**T**	**C**	**A**	**C**	**G**	**T**	**T**	**G**	**A**	**C**
		*	**A**	*	**G**	**T**				
		C	*	**C**	*	**T**				
			*	**C**	*	**T**	**T**			
			A	*	**G**	*	**T**			
				*	**G**	*	**T**	**G**		
				C	*	**T**	*	**G**		
					*	**T**	*	**G**	**A**	
					G	*	**T**	*	**A**	
						*	**T**	*	**A**	**C**
						T	*	**G**	*	**C**

Figura 3. Esquema de execução do algoritmo AB para um *chip* com *k* = 2 e *c* = 2.

Um grande obstáculo para a implementação desse algoritmo é o fato de termos de gerar todas as extensões possíveis de tamanho *ck*+1. Para se ter uma idéia do problema, Halperin et al. sugerem os valores 4 e 6 como razoáveis para *c* e *k*, respectivamente, o que implica em um tamanho de 25 bases para o *probe*, e, por conseguinte, para a extensão *a*. O pior é pensar que esse problema se repete para todas as posições da seqüência alvo. Na realidade, esse problema já existia no algoritmo **A**, mas só sentimos os seus efeitos práticos no algoritmo **AB** porque, para este, o tamanho do *probe* é, geralmente, muito maior.

Outro problema que se observa na implementação do algoritmo **AB** é que nem todos os *probes* que casam com um *caminho* realmente indicam uma base para a posição a estender. É o caso dos *probes* que têm uma base universal associada à tal posição. Esse problema realmente não afeta a resposta do algoritmo. Tentar evitá-lo é apenas uma questão de se eliminar esforço computacional desnecessário.

3. A Nossa Proposta

Nesta seção apresentamos um esquema alternativo de *chip* e um novo algoritmo para montar a seqüência alvo a partir do seu *spectrum*. A construção desse *chip* levou em conta a sugestão de Halperin et al. de se incrementar o algoritmo **B** com aspectos do algoritmo **A**, mais especificamente com a idéia de usar *caminhos*. O objetivo dessa nova estratégia é chegar a melhores resultados práticos do que os obtidos com o algoritmo **B** original**.** Na realidade, isso já era esperado do algoritmo **AB**. A diferença é que a nossa proposta pretende superar as dificuldades enfrentadas por este.

3.1. O Nosso Chip

O *chip* que propomos é construído da seguinte forma:

a. cada um de seus *probes* é composto de duas partes: uma parte inicial com *k* bases naturais (*N*) posicionadas aleatoriamente e separadas por bases universais com um fator *c* de distanciamento, onde *c* é um inteiro convenientemente grande; e uma parte final composta de *e* bases naturais contíguas. Logo, o tamanho total de cada *probe* é dado por *ck*+*e*. Figura 4;

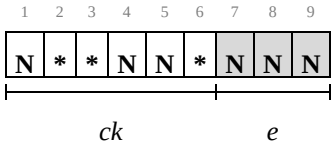


Figura 4. Esquema de um *probe* para *c* = 2, *k* = 3 e *e* = 3.

b. formamos *f* famílias de probes. Para cada família, selecionamos aleatoriamente um conjunto de *k* posições no intervalo de 1 a *ck* e formamos todos os 4^{k+e} *pobres* possíveis, colocando as bases *A*, *C*, *G* e *T* nas *k* posições selecionadas e nas *e* últimas. Figura 5;

c. o número *n* total de probes no chip é dado por $f4^{k+e}$.

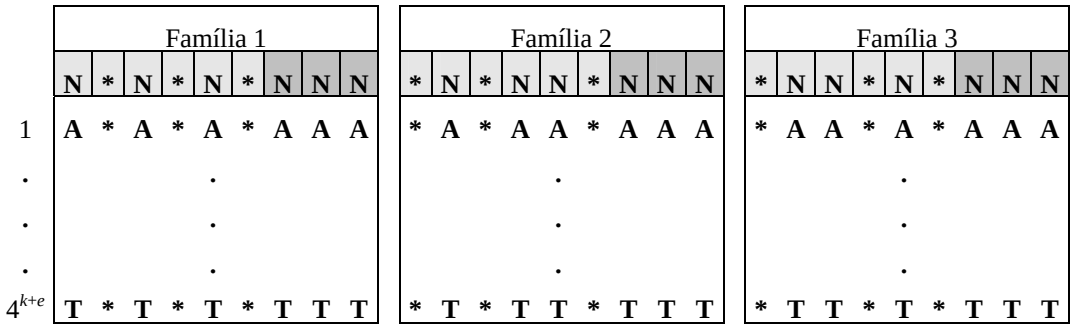


Figura 5. Esquema de um *Chip* com 3 famílias de *probes* para *c* = 2, *k* = 3 e *e* = 3. Cada família é formada por todas as combinações possíveis das base *A*, *C*, *G* e *T* nas posições especificadas.

Basicamente, a diferença do nosso projeto para aquele proposto por Halperin et al. é que o nosso esquema garante *e* bases naturais no final do *probe*, enquanto o deles garante apenas uma. Nota-se que para o caso especial de $e = 1$, os *chips* são iguais.

3.2. O Nosso Algoritmo

Considerando que a seqüência S é conhecida desde s_1 até s_i, o nosso algoritmo executa os seguintes passos sobre o *spectrum* obtido com o nosso *chip* para determinar a base s_{i+1}:

1. gerar todas as 4^e seqüências $a = a_1,...,a_e$ como extensão de $s_1,...,s_i$;
2. para cada seqüência a', contar o número de *probes* do *spectrum* que casam com a subseqüência $s_{i-ck-e+2},...,s_i, a'_1,...,a'_e$ ou *caminho*;
3. para a seqüência vencedora, fazer $s_{i+1} = a'_1$ (o empate é resolvido arbitrariamente).

Observamos que o nosso algoritmo é similar ao algoritmo **AB** proposto por Halperin et al. A diferença está no tamanho da extensão a ser gerada. Enquanto naquele algoritmo o tamanho da extensão é o próprio tamanho do *probe*, dado por $ck+1$, no nosso, o tamanho da extensão é e. E assim, o tempo de execução do nosso algoritmo é dado por $O(4^e efm)$.

A Figura 6 mostra uma seqüência sendo construída pelo nosso algoritmo. As suas 9 primeiras bases já foram determinadas. Desejamos determinar o valor da base posição 10. De acordo com o algoritmo, temos de gerar todas as extensões possíveis de tamanho *e* para o trecho que vai da posição 10 até a posição 11 e, então, avaliar, para cada uma delas, a quantidade de *probes* que casam com o caminho que vai da posição 3 até a posição 11. Essa figura mostra que para a extensão AC existem 4 *probes*.

								s_i	a_1	a_e
1	2	3	4	5	6	7	8	9	10	11
C	**T**	**C**	**A**	**C**	**G**	**T**	**C**	**G**	**A**	**C**
		*	**A**	*	**G**	*	**C**	**G**	**A**	
		C	*	*	**G**	**T**	*	**G**	**A**	
			*	**C**	*	**T**	*	**G**	**A**	**C**
			A	*	*	**T**	**C**	*	**A**	**C**

Figura 6. Esquema de funcionamento do nosso algoritmo para um *chip* com $c = 2$, $k = 3$ e $e = 2$.

Em nossa proposta conseguimos reunir as seguintes características positivas dos algoritmos **A** e **B**:

a. o nosso *chip* é formado por famílias de probes, o que aumenta a resistência do algoritmo a erros falsos negativos e falsos positivos;

b. o algoritmo usa o conceito de *caminho* e, assim, apóia-se nas informações de um número maior de *probes* para determina a base a estender;

c. o algoritmo usa bases universais que tendem a minimizar o problema de ambigüidade.

É verdade que o algoritmo **AB** já guardava essas características positivas, porém, nossa técnica, em relação a ele, avançou nas seguintes direções:

a. como o tamanho da extensão não está vinculado ao tamanho do *probe*, mas, sim, ao valor definido para *e* durante a construção do *chip*, temos a oportunidade de minimizar o problema prático de gerar grandes extensões. Basta apenas que escolhamos um valor "pequeno" para *e*.

b. eliminamos o problema de termos vários *probes* apoiando um caminho sem, no entanto, esclarecer nada sobre a posição a estender. Por exemplo, na Figura 6, todos os *probes* indicam a base **A** para a posição 10.

O problema do nosso algoritmo é o fato de termos de gerar repetidas vezes todas as extensões possíveis de tamanho *e*. Como vimos, esse problema já estava presente tanto no Algoritmo **A** quanto no Algoritmo **AB**. O fato de usarmos na nossa técnica extensões de tamanho *e*, ao invés de *k* e de *ck*+1, como naqueles dois algoritmos, assintoticamente, não altera nada, mas ainda assim é possível que, com relação ao algoritmo **B**, alcancemos resultados práticos superiores. É exatamente isso que esperamos comprovar com os nossos experimentos.

4. Resultados Experimentais

Nesta seção, apresentamos alguns resultados preliminares de simulações realizadas com o algoritmo e com o *chip* alternativos que desenvolvemos. O algoritmo foi implementado na linguagem C e executado sobre a plataforma operacional SunOS em uma máquina SPARCstation-20.

Para efeito comparativo, sempre que possível, os valores para os parâmetros do algoritmo foram escolhidos com base nos valores usados nos experimentos publicados por Halperin et al. Em todas as simulações, as seqüências de DNA foram geradas randomicamente.

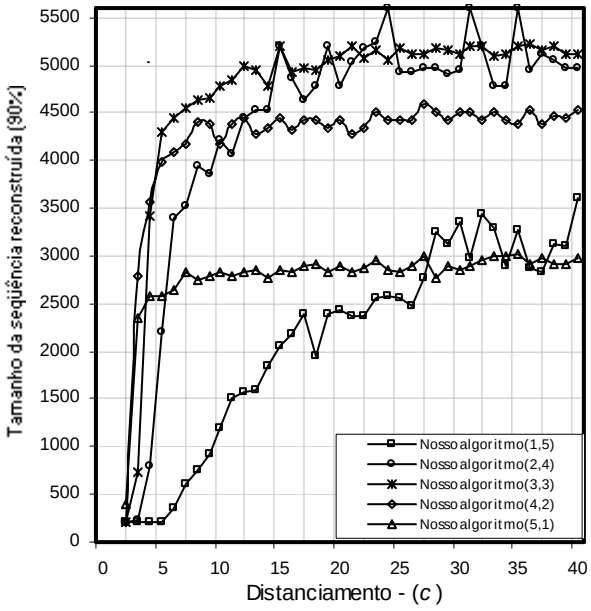


Figura 7. Desempenho do nosso algoritmo para *k* + *e* = 6.

Os gráficos da Figura 7 apresentam o *maior comprimento que se atinge com taxa de sucesso próxima de 90%* como função do *distanciamento c*. Para obter as estatísticas de cada ponto, usamos uma amostra de 200 seqüências. Em todos os casos, fizemos $p = q = 0$ e usamos 4^8 *probes*, quantidade que corresponde ao total de probes do *chip* clássico. Há um gráfico para cada combinação possível de *k* e *e* que resulte em $k + e = 6$. Para um valor adequado de *c*, garante-se a existência de 16 famílias de *probes* no *chip*.

Destacamos que para o caso $e = 1$, o nosso algoritmo (Nosso algoritmo (5,1)) funciona como o algoritmo **B** e, obviamente, gera a mesma saída. O pior resultado ocorre para $e = 5$ e $k = 1$, mesmo assim, para *c* próximo de 30, a curva resultante tende a acompanhar a do algoritmo **B**. Para todas as outras combinações de *k* e *e*, percebemos que os nossos resultados são superiores. Os resultados se mostram ainda mais promissores se considerarmos que, para um mesmo valor de *c*, o tamanho do nosso probe é geralmente bem menor do que o de Halperin et al.

Os gráficos da Figura 8 apresentam o *percentual de sucesso de reconstrução* como função do *tamanho da seqüência alvo* para o nosso algoritmo e para o algoritmo **B**. Em todos os casos, usamos 4^8 *probes* com 6 bases naturais. Fizemos $p = q$ e variamos os seus valores. Para obter as estatísticas de cada ponto, usamos uma amostra de 200 seqüências. Para o algoritmo **B**, usamos $c_B = 4$ e $k_B = 5$. Para o nosso, selecionamos $c_n = 6$, $k_n = 3$ e $e_n = 3$. Esses valores foram escolhidos de modo que a comparação pudesse ser feita entre probes com o mesmo tamanho, nesse caso, cada um com 21 bases.

Na presença de diferentes níveis de erro, percebe-se que os nossos resultados ainda permanecem acima dos resultados do algoritmo **B**. Percebe-se, ainda, que os desempenhos dos algoritmos caem a taxas mais ou menos constantes e aproximadas.

5. Conclusões Finais e Trabalhos Futuros

Observa-se que o algoritmo e o *chip* propostos minimizam o problema que ocorre com o algoritmo **AB** de se ter de gerar grandes extensões. Em relação ao algoritmo **B**, as simulações comprovam preliminarmente que com a nossa proposta é possível se reconstruir seqüências maiores. A contrapartida por esse avanço é um maior tempo de processamento. Na verdade, isso já era intuitivamente esperado. Mas um fato deve ser considerado quando da avaliação desse compromisso entre o tempo de execução e o tamanho da cadeia reconstruída, é que geralmente o tamanho dos probes que usamos é muito menor do que o dos probes usados por Halperin et al.

Realmente esse é um aspecto muito importante da nossa proposta, principalmente se considerarmos a afirmação dos próprios Halperin et al. de que, apesar de as bases universais terem sido geradas com sucesso em laboratório, ainda não é certo se probes longos, com muitas bases universais, podem hibridizar confiavelmente.

Neste artigo, investigamos o comportamento do algoritmo proposto para $k + e = 6$, que corresponde ao valor usado por Halperin et al. em seu experimento. Até aqui, o nosso objetivo foi fazer um estudo comparativo para determinar a viabilidade da nossa proposta. No futuro, pretendemos realizar simulações com outros valores. Também estudaremos o comportamento do algoritmo com seqüências naturais, ao invés de seqüências geradas randomicamente. É possível que, ao final, todos esses resultados experimentais nos remetam à análise matemática da nossa proposta.

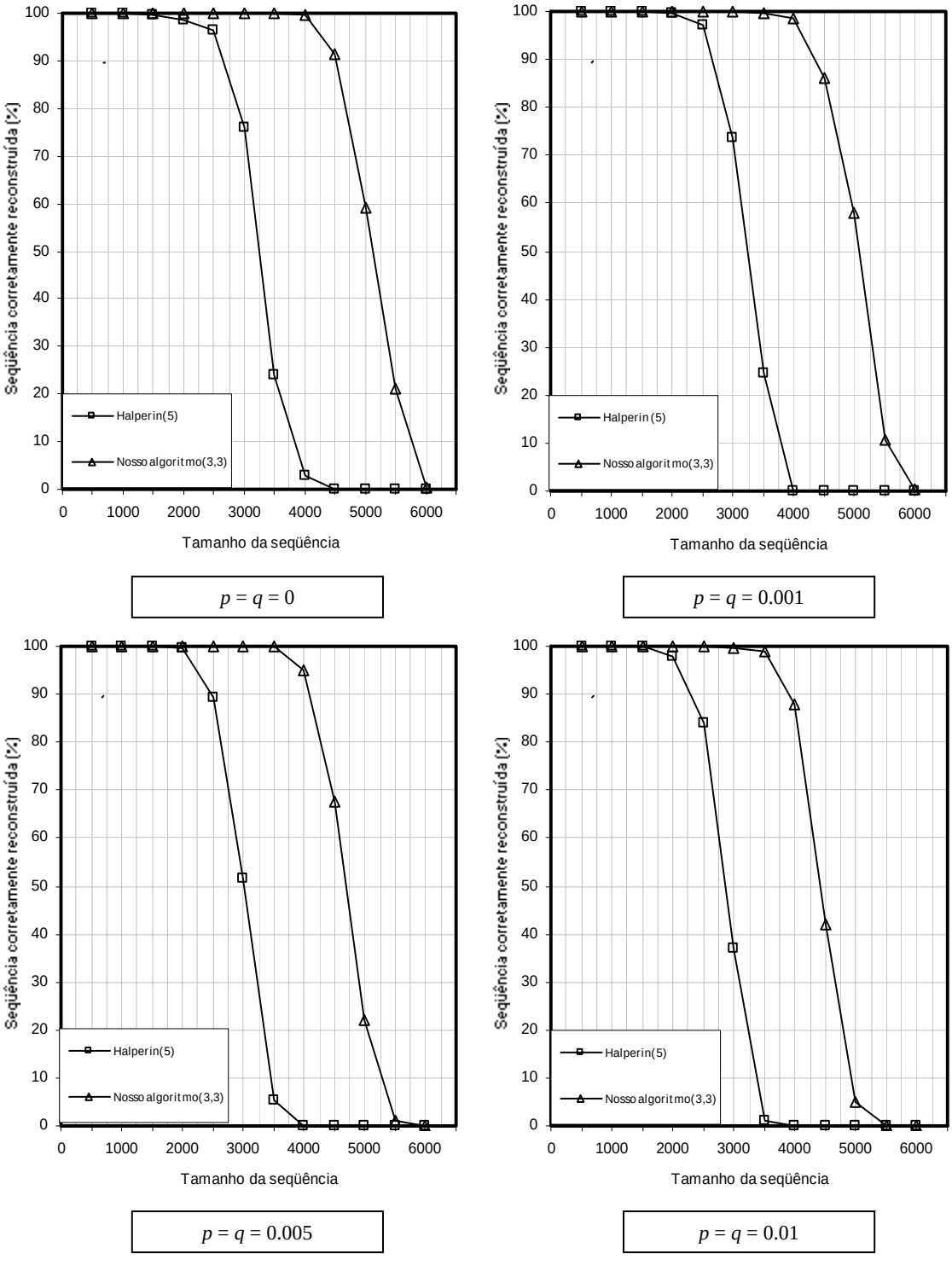


Figura 8. Desempenho do nosso algoritmo e do algoritmo B.

6. Referências Bibliográficas

Doi, K., Imai, H. (2000) "Sequencing by hybridization in the presence of hybridization errors", In Proceedings of the Workshop on Genome Informatics (*GIW* '00), volume 11, pages 53-62.

Drmanac, R., Drmanac, S. (2001) “Sequencing by hybridization arrays”, In:RAMPAL, J.B. DNA Arrays: Methods and Protocols. Totowa: Humana, 2001. 170v. (Methods in Molecular Biology).

Halperin, E., Halperin, S., Hartman, T., Shamir, R. (2002) “Handling long targets and errors in sequencing by hybridization”, Proceedings of 6th RECOMB (2002), 176-185. To appear in the Journal of Computational Biology.

Pevzner, P.A. (1989) “l-tuple DNA sequencing: computer analysis”, Journal of Biomolecular Structure & Dynamics, 7(1):63-73.

Pevzner, P.A., Lysov, Yu. P., Khrapko K.R., Belyavsky, A.V., Florentiev, V.L., Mirzabekov, A.D. (1991) “Improved chips for sequencing by hybridization”, Journal of Biomolecular Structure and Dynamics, 9(2):399-410.

Pevzner, P.A., Waterman, M.S. (1995) “Open combinatorial problems in computational molecular biology”, In 3rd Israel Symposium on Theory of Computing and Systems, pages 158-163. IEEE Computers Society Press.

Preparata, F.P., Frieze, A.M., and Upfal, E. (1999) “On the power of universal bases in sequencing by hybridization”, Proc. The Third Annual International Conference on Computational Molecular Biology, 295-301.