2
Seqüenciamento por Hibridação

A técnica de seqüenciamento de DNA por hibridação é apresentada neste capítulo. Também são discutidas a tarefa de seqüenciamento de DNA e as motivações para o desenvolvimento de novos métodos de seqüenciamento. Na Seção 2.1 apresenta-se o que é uma molécula de DNA e sua relevância. O seqüenciamento de DNA e as técnicas utilizadas atualmente para realizá-lo são discutidos na Seção 2.2. O processo bioquímico do SBH é apresentado na Seção 2.3. O problema computacional envolvido no SBH é discutido na Seção 2.4. Nas Seções 2.5 e 2.6 apresenta-se algumas dificuldades encontradas na aplicação da técnica do SBH. A formulação matemática do problema computacional tratado neste trabalho é apresentada na Seção 2.7. Outras aplicações do processo bioquímico envolvido no SBH são discutidas na Seção 2.8.

PUC-Rio - Certificação Digital Nº 0310813/CA

2.1
DNA, proteínas e os seres vivos

Diversos aspectos dos organismos dos seres vivos são controlados através de *proteínas*. Os anticorpos, por exemplo, são proteínas responsáveis pelo reconhecimento de organismos "invasores", como os vírus. A digestão de alimentos é realizada através da ação de proteínas chamadas enzimas, presentes no aparelho digestivo. Os diversos hormônios presentes no organismo também são proteínas. Enfim, existem diversas proteínas exercendo, permanentemente, diferentes tarefas em todo organismo. As proteínas são produzidas nas células de diversos tecidos dos seres vivos. Entretanto, os fatores que determinam sua produção ainda não são totalmente compreendidos.

O *ácido desoxirribonucleico* (DNA) é uma molécula presente em praticamente todas as células do organismo. Embora seja sabido que a produção de proteínas é influencida por fatores externos às células produtoras, sabe-se que, no mais baixo nível, é o DNA das células que determina o conteúdo

das proteínas produzidas. Em 1953, Watson e Crick [32] determinaram a estrutura segundo a qual as moléculas de DNA são organizadas. Foi quando se descobriu que o DNA é composto por outras moléculas denominadas *nucleotídeos*, organizados em duas fitas que se entrelaçam em forma de hélice. Uma ilustração de uma molécula de DNA é apresentada na Figura 2.1. Os nucleotídeos são compostos por duas moléculas: um açúcar e uma *base*. Existem quatro nucleotídeos distintos que são diferenciados pelas suas bases: adenina (A), timina (T), citosina (C) e guanina (G). Uma fita é uma seqüência de nucleotídeos ligados entre si pelos seus açúcares através de uma molécula de fosfato (PO_4, indicada na Figura 2.1 pela letra P circulada). As duas fitas de uma molécula de DNA estão conectadas uma na outra por toda sua extensão através de pontes de hidrogênio (H, na Figura 2.1) entre as bases de seus nucleotídeos. Estas fitas são complementares, as bases A são ligadas somente em bases T e as bases C são ligadas somente em bases G. Por isto, uma molécula de DNA pode ser definida por uma palavra sobre o alfabeto $\{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\}$, representando a seqüência de nucleotídeos de uma das fitas. Seu tamanho (comprimento) é medido em *pares de bases* (bp, "base pairs"), que indicam a quantidade de nucleotídeos (ou de bases) em cada uma das fitas.

Há algum tempo, sabe-se que cada proteína é determinada por alguma parte da seqüência de nucleotídeos do DNA da célula produtora. Estas partes são denominadas *regiões codificadoras* e cada três nucleotídeos na seqüência de DNA de uma região codificadora determinam um *aminoácido* na proteína produzida. Os aminoácidos são as moléculas básicas que compõem as proteínas e é a seqüência de aminoácidos de uma proteína que determina sua estrutura e, por fim, sua função no organismo. Daí a importância crucial no estudo do conteúdo do DNA para a compreensão dos mecanismos que regem o funcionamento dos organismos dos seres vivos.

2.2 Seqüenciamento de DNA

O *seqüenciamento de DNA* consiste em determinar a seqüência de nucleotídeos de uma dada molécula de DNA. Esta informação é denominada estrutura primária do DNA porque representa a estrutura da molécula no seu mais baixo nível. Entretanto, obter esta informação não é uma tarefa simples porque não se conhece uma maneira de "ler" diretamente os nucleotídeos de uma molécula de DNA. Em 1977, Sanger *et al.* [29] e Maxam *et al.* [21] propuseram, independentemente, duas metodologias para

PUC-Rio - Certificação Digital Nº 0310813/CA

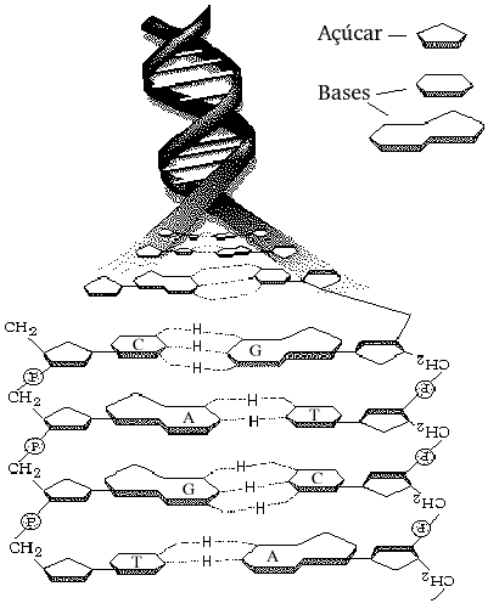

Figura 2.1: Duas visões de uma molécula de DNA.

PUC-Rio - Certificação Digital Nº 0310813/CA

realizar o seqüenciamento de uma molécula de DNA. A idéia de Sanger *et al.* é vastamente utilizada nos projetos de seqüenciamento, pois supera a idéia de Maxam *et al.* na questão do tempo necessário para realizar o seqüenciamento (embora em alguns casos particulares a utilização da técnica de Maxam *et al.* seja mais vantajosa).

A técnica de Sanger *et al.* realiza o seqüenciamento gerando várias moléculas de diferentes tamanhos com conteúdo igual ao início da molécula original, ou seja, moléculas que são "prefixos" da molécula original. Em seguida, estas moléculas são separadas por tamanho e determina-se o último nucleotídeo de cada molécula, podendo inferir-se a posicão relativa entre os nucleotídeos. Esta técnica também é chamada de *corrida em gel*, porque a fase de separação das moléculas é realizada colocando-as em um gel. As moléculas atravessam todo o gel, sendo que as menores o fazem em uma velocidade maior do que as maiores, permitindo determinar-se o tamanho relativo entre elas. Os maiores problemas com a corrida em gel é o tempo necessário para separar as moléculas e a qualidade da informação nas pontas da seqüência. Na ponta maior, o problema é que a diferença relativa do tamanho das moléculas grandes é pequena, dificultando a separação. Na ponta menor, o problema vem da fase anterior (quando são criados os prefixos) onde prefixos pequenos possuem probabilidade menor de serem gerados. Outra implicação do problema de separação é a limitação no tamanho das moléculas a serem seqüenciadas (hoje em dia consegue-se seqüenciar moléculas com aproximadamente 800 bp). A técnica de *Maxam et al.* também sofre destes males.

Um genoma completo (toda a informação genética de um organimo) pode chegar a milhões ou até bilhões de pares de base. Então como foi possível realizar o seqüenciamento de vários genomas completos (inclusive o humano, que possui aproximadamente três bilhões de pares de base)? É possível realizar o seqüenciamento de moléculas grandes utilizando uma técnica denominada *shotgun*. A idéia básica é criar vários clones da molécula original, quebrá-los aleatoriamente em partes menores que possibilitem o seqüenciamento por corrida em gel, seqüenciar cada parte independentemente e realizar a montagem destas partes baseando-se na sobreposição entre elas. Por isto, realizar rapidamente o seqüenciamento de moléculas de tamanho médio é muito importante. Esta tarefa é realizada um número muito grande de vezes em alguns projetos de seqüenciamento.

2.3 Arranjo de DNA

PUC-Rio - Certificação Digital Nº 0310813/CA

Uma alternativa interessante para realizar o seqüenciamento de moléculas de DNA de tamanho médio é o *seqüenciamento por hibridação* ("sequencing by hybridization" ou, simplesmente, SBH). Esta técnica ainda está em desenvolvimento e pode se tornar mais rápida, mais barata e obter resultados mais confiáveis do que a corrida em gel. O SBH foi proposto independentemente e simultaneamente por quatro grupos em 1988 [1, 10, 20, 30]. O SBH implica a criação de um *arranjo de DNA*, que consiste de uma grade bidimensional em miniatura contendo, em cada célula, um fragmento de DNA sintetizado. Cada fragmento é denominado *prova* e o conjunto de todas as provas do arranjo é chamado de *biblioteca de provas*. Um experimento bioquímico realizado com o arranjo e uma solução contendo várias cópias de uma molécula de DNA desconhecida (*DNA alvo* ou *seqüência alvo*) revela informações sobre esta molécula. Durante o experimento, as cópias do DNA alvo se ligam quimicamente às provas que são complemento de alguma subseqüência da sua seqüência de nucleotídeos. Esta reação química é denominada *hibridação*. Como as cópias do DNA alvo são previamente marcadas com algum material radioativo ou fluorescente, é possível "ler" o arranjo após o experimento de hibridação e determinar quais provas reagiram com o DNA alvo. Em resumo, é possível determinar quais provas são subseqüências do DNA alvo. Este subconjunto da biblioteca de provas é denominado *espectro*.

No SBH, o DNA alvo é a molécula cuja seqüência de nucleotídeos se deseja obter e a biblioteca de provas é constituída por todas as 4^l

provas possíveis de tamanho l. O arranjo de DNA construído desta forma é denominado *arranjo padrão* $C(l)$. O esquema de um $C(4)$, após um experimento com a seqüência `ATAGGCAGGA`, é ilustrado na Figura 2.2.

AAAA	AAAC	AAAG	AAAT	AACA	AACC	AACG	AACT	AAGA	AAGC	AAGG	AAGT	AATA	AATC	AATG	AATT
ACAA	ACAC	ACAG	ACAT	ACCA	ACCC	ACCG	ACCT	ACGA	ACGC	ACGG	ACGT	ACTA	ACTC	ACTG	ACTT
AGAA	AGAC	AGAG	AGAT	AGCA	AGCC	AGCG	AGCT	AGGA	AGGC	AGGG	AGGT	AGTA	AGTC	AGTG	AGTT
ATAA	ATAC	ATAG	ATAT	ATCA	ATGC	ATCG	ATCT	ATGA	ATGC	ATGG	ATGT	ATTA	ATTC	ATTG	ATTT
CAAA	CAAC	CAAG	CAAT	CACA	CACC	CACG	CACT	CAGA	CAGC	CAGG	CAGT	CATA	CATC	CATG	CATT
CCAA	CCAC	CCAG	CCAT	CCCA	CCCC	CCCG	CCCT	CCGA	CCGC	CCGG	CCGT	CCTA	CCTC	CCTG	CCTT
CGAA	CGAC	CGAG	CGAT	CGCA	CGCC	CGCG	CGCT	CGGA	CGGC	CGGG	CGGT	CGTA	CGTC	CGTG	CGTT
CTAA	CTAC	CTAG	CTAT	CTCA	CTCC	CTCG	CTCT	CTGA	CTGC	CTGG	CTGT	CTTA	CTTC	CTTG	CTTT
GAAA	GAAC	GAAG	GAAT	GACA	GACC	GACG	GACT	GAGA	GAGC	GAGG	GAGT	GATA	GATC	GATG	GATT
GCAA	GCAC	GCAG	GCAT	GCCA	GCCC	GCCG	GCCT	GCGA	GCGC	GCGG	GCGT	GCTA	GCTC	GCTG	GCTT
GGAA	GGAC	GGAG	GGAT	GGCA	GGCC	GGCG	GGCT	GGGA	GGGC	GGGG	GGGT	GGTA	GGTC	GGTG	GGTT
GTAA	GTAC	GTAG	GTAT	GTCA	GTCC	GTCG	GTCT	GTGA	GTGC	GTGG	GTGT	GTTA	GTTC	GTTG	GTTT
TAAA	TAAC	TAAG	TAAT	TACA	TACC	TACG	TACT	TAGA	TAGC	TAGG	TAGT	TATA	TATC	TATG	TATT

Figura 2.2: Esquema de um $C(4)$ após um experimento de hibridação com a seqüência `ATAGGCAGGA`.

A partir do resultado do experimento com o arranjo, é possível determinar a seqüência de nucleotídeos no DNA alvo solucionando o *problema de montagem* das provas do espectro. A *sobreposição máxima* (ou, simplesmente, *sobreposição*) entre duas provas é o tamanho da maior seqüência que é um sufixo próprio da primeira prova e, ao mesmo tempo, um prefixo da segunda. Então, o problema de montagem consiste em encontrar uma seqüência de todas as provas do espectro de forma que provas consecutivas possuam sobreposição de $l-1$ bases (*sobreposição ótima*). Considerando que o comprimento da seqüência alvo é n, o espectro conterá $n-l+1$ provas.

2.4 O problema de montagem

O problema de montagem do SBH pode ser reduzido ao conhecido *problema do caminho Hamiltoniano* (PCH), que consiste em encontrar um caminho que visita exatamente uma vez *todos os vértices* de um grafo. Um caminho com estas características é chamado de Hamiltoniano. A redução consiste em criar um vértice para cada prova do espectro e uma aresta entre cada par de provas que possuam sobreposição de $l-1$ nucleotídeos. Um caminho Hamiltoniano neste grafo corresponde a uma seqüência de todas as provas com sobreposição ótima entre provas consecutivas, ou seja, uma solução para o problema de montagem. Na Figura 2.3 é apresentado o grafo resultante da redução do problema de montagem do exemplo da Figura 2.2 a um PCH (o caminho Hamiltoniano está em destaque). Entretanto,

PUC-Rio - Certificação Digital Nº 0310813/CA

sabe-se que o PCH é um problema NP-difícil e, portanto, não se conhecem algoritmos eficientes (de complexidade polinomial) para resolvê-lo.

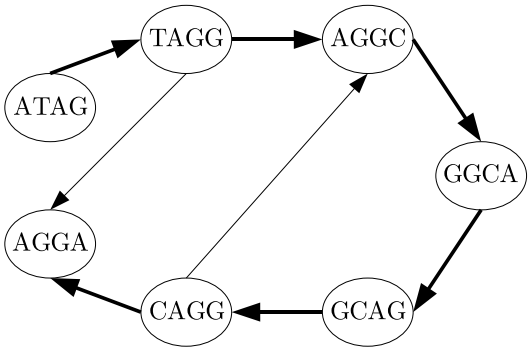

Figura 2.3: Grafo resultante da redução do problema de montagem do exemplo da Figura 2.2 a um PCH.

Existe uma redução do problema de montagem para o não menos conhecido *problema do caminho Euleriano* (PCE), para o qual se conhece algoritmos eficientes. O PCE consiste em encontrar um caminho Euleriano em um grafo. Um caminho Euleriano passa por *todas as arestas* de um grafo exatamente uma vez. Nesta redução, os vértices do grafo são todas as seqüências de tamanho $l-1$ e uma aresta (u, v) é criada no grafo se existe uma prova no espectro que possua prefixo igual a u e sufixo igual a v. Como para cada prova do espectro é criada uma aresta no grafo, o problema de montagem corresponde a encontrar um caminho euleriano neste grafo. O grafo resultante da redução do problema de montagem do exemplo da Figura 2.2 a um PCE é apresentado na Figura 2.4. Cada aresta está rotulada com a prova correspondente. Por esta redução foi provado que o problema de montagem é polinomial neste caso (na Seção 2.6 são apresentadas algumas características que complicam o problema de montagem).

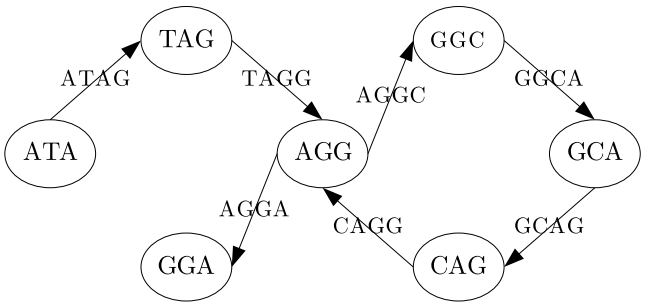

Figura 2.4: Grafo resultante da redução do problema de montagem do exemplo da Figura 2.2 a um PCE.

PUC-Rio - Certificação Digital Nº 0310813/CA

2.5 Reconstrução única

O problema de montagem também é chamado de *problema de reconstrução* e nem sempre possui uma solução única. Para um dado arranjo, experimentos de hibridação com diferentes seqüências alvo podem resultar em um mesmo espectro. Nestes casos, o grafo resultante da redução do problema de reconstrução a um PCE possui mais de um caminho euleriano. Um exemplo de um problema de reconstrução com duas soluções possíveis é ilustrado na Figura 2.5. Neste exemplo, é apresentado um grafo correspondente à redução do problema de reconstrução resultante do experimento com um $C(3)$. O grafo possui dois caminhos eulerianos que correspondem às seqüências ATGGCGTGCA e ATGCGTGGCA. Em casos como este, é impossível determinar qual caminho corresponde à seqüência alvo.

A possibilidade de um experimento resultar em um problema de reconstrução ambíguo está diretamente ligada ao arranjo de DNA utilizado. Por exemplo, para um arranjo padrão, quanto maior o tamanho das provas, menor é a chance de uma seqüência possuir um espectro compartilhado por outras seqüências. Uma forma de analisar o *poder de seqüenciamento* de um arranjo C é calcular a probabilidade de uma seqüência aleatória de tamanho n não possuir um espectro único para este arranjo. Quanto maior esta probabilidade, menor é o poder de seqüenciamento do arranjo. Denominando-se esta probabilidade como $p(C, n)$, pode-se pensar no valor máximo para n tal que $p(C, n) \leq q$, para uma dada probabilidade limite q, ou seja, qual o tamanho máximo das seqüências que podem ser seqüenciadas por C com certa garantia de sucesso na reconstrução (probabilidade de reconstrução única maior ou igual a $1 - q$). Para um arranjo C e uma dada probabilidade limite q, este tamanho é denominado $n_{max}(C, q)$.

Alguns trabalhos foram desenvolvidos para analisar e melhorar o poder de seqüenciamento dos arranjos de DNA. Em [24] foi provado que, para um arranjo padrão, $n_{max}(C, q) \approx \frac{1}{3} \cdot |C| \cdot q$, onde $|C|$ é a *dimensão* do arranjo. A dimensão de um arranjo é a quantidade de provas sintetizadas e fixadas na sua superfície (no caso de um arranjo padrão, $|C(l)| = 4^l$). Em [25] foram propostos diferentes projetos de arranjos que possuem um poder de seqüenciamento maior do que os arranjos padrão. Estes arranjos utilizam provas compostas por *bases universais* (bases que se ligam a qualquer uma das quatro bases padrão). Um dos arranjos propostos é denominado *arranjo binário*, $C_{bin}(l)$, e possui $n_{max}(C_{bin}(l), q) \approx \frac{1}{\sqrt{12}} \cdot |C_{bin}(l)| \cdot \sqrt{q}$, o que implica em um fator de melhoria de $\sqrt{\frac{3}{4q}}$ em relação ao arranjo padrão. Como

PUC-Rio - Certificação Digital Nº 0310813/CA

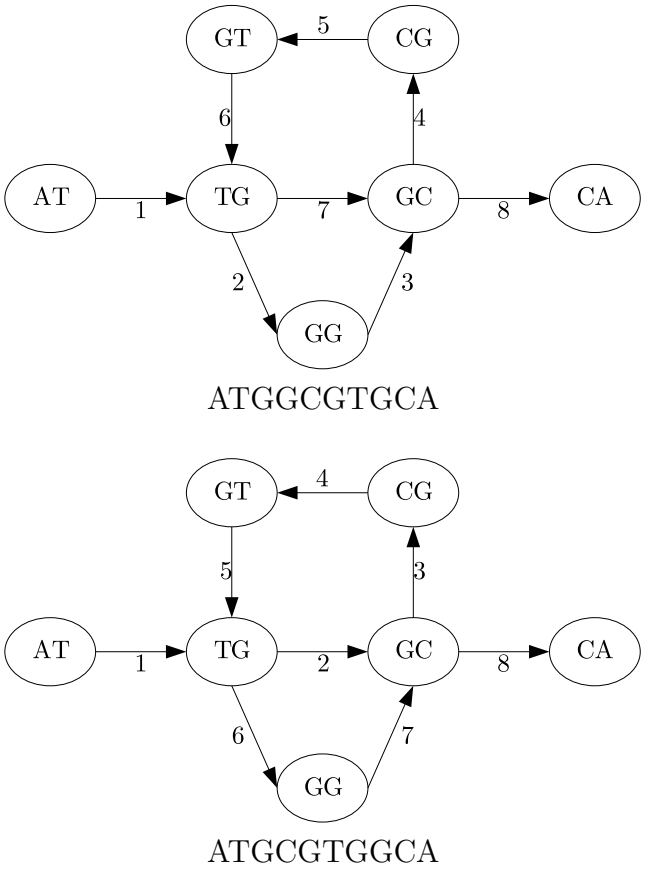

Figura 2.5: Duas reconstruções (dois caminhos eulerianos) no mesmo grafo. O grafo é resultado da redução do problema de reconstrução do espectro $S = \{\text{ATG}, \text{TGG}, \text{TGC}, \text{GTG}, \text{GGC}, \text{GCA}, \text{GCG}, \text{CGT}\}$ a um PCE. Os números nas arestas indicam a ordem delas no caminho que reconstroi cada seqüência. As seqüências correspondentes estão indicadas abaixo de cada grafo.

base de comparação, tomando $q = 0.01$ e arranjos com dimensão de 64-Kb, $n_{max} \approx 1800$ para os arranjos binários e, para os arranjos padrão, $n_{max} \approx 210$.

Existem outras propostas para melhorar o desempenho do SBH. Por exemplo, em [19] foi apresentada uma idéia que consiste em realizar experimentos adicionais para obter informações sobre a posição aproximada de cada prova na seqüência alvo. Desta forma, é menos provável obter ambigüidades na fase de reconstrução. Há ainda outras propostas [26, 18] de arranjos utilizando bases universais com poder de seqüenciamento maior do que os discutidos acima. É claro que o problema de reconstrução que surge para cada tipo de arranjo é diferente, apesar de haver algumas similaridades. Apesar das vantagens destes arranjos, o problema de reconstrução considerado neste trabalho é aquele referente ao arranjo padrão.

PUC-Rio - Certificação Digital Nº 0310813/CA

2.6 Erros de hibridação

Durante o processo de hibridação entre o arranjo de DNA e a molécula alvo podem ocorrer alguns problemas que resultam em informações erradas no espectro. Existem dois tipos de erros: *falsos positivos* e *falsos negativos*. Falsos positivos consistem de provas que aparecem no espectro, mas não fazem parte do DNA alvo. Falsos negativos são provas que fazem parte do DNA alvo, mas não aparecem no espectro.

Os erros de hibridação são devidos, principalmente, às condições peculiares nas quais o experimento deve ser realizado para que as reações de hibridação ocorram corretamente. Estas condições, principalmente a temperatura de fusão das provas, são dependentes do conteúdo de cada prova e podem variar mais do que o desejado de uma prova para outra. Nestes casos, uma hibridação que deveria ocorrer não ocorre porque a temperatura está acima da temperatura de fusão da prova em questão (falso negativo) ou uma hibridação entre seqüências que não são complementares pode ocorrer porque a temperatura está muito abaixo da temperatura de fusão da prova (falso positivo). Outra fonte de falsos negativos são as repetições de provas na seqüência de DNA. Tipicamente, o resultado do experimento apenas informa a presença ou não de cada prova. É difícil determinar se uma prova ocorre mais de uma vez e mais difícil ainda é determinar a quantidade de ocorrências. Por isto, se uma prova aparece em m posições distintas da seqüência alvo, serão gerados $m-1$ falsos negativos.

A presença de erros no espectro implica em um diferente problema computacional que é muito mais difícil do que o problema livre de erros. Na realidade, este novo problema é NP-difícil. A formulação é de um problema de otimização combinatória e utiliza algumas informações adicionais, além do espectro.

2.7 Formulação do problema de montagem com erros

O problema de SBH considerando falsos positivos e falsos negativos no espectro foi provado ser NP-difícil em [7]. O problema é formulado como um problema de otimização combinatória que tem como entrada o espectro e o tamanho da seqüência de DNA alvo. O objetivo é encontrar uma seqüência de provas que seja um sub-conjunto do espectro, reconstrua uma seqüência de DNA menor ou igual ao tamanho fornecido e possua o

PUC-Rio - Certificação Digital Nº 0310813/CA

maior número possível de provas. Neste trabalho considera-se mais uma informação como disponível: a primeira prova da seqüência alvo. Estimar o tamanho de uma seqüência de DNA é possível utilizando *gel electrophoresis*. Os seus primeiros nucleotídeos normalmente estão disponíveis pela ampla utilização do método PCR [28] na geração e ampliação de fragmentos de DNA para seqüenciamento (outros autores consideram mais informações, tais como a freqüência de erros, entretanto estas informações podem não estar disponíveis na prática). A formulação do problema de SBH tratada neste trabalho é a seguinte:

Problema de Seqüenciamento por Hibridação (PSBH)

Dados um espectro $S = \{s_1, s_2, ..., s_m\}$, o tamanho das provas l, o tamanho da seqüência original n e a primeira prova da seqüência $s_0 \in S$, encontrar uma seqüência que possua tamanho menor ou igual a n, contendo o maior número possível de provas do espectro.

Este problema é um caso particular do *problema do caixeiro viajante com coleta de prêmios* ("prize collecting traveling salesman problem"). Neste problema, o objetivo é encontrar um caminho em um grafo que obedeça o limite para a soma dos pesos das arestas utilizadas e colete a maior quantidade de prêmios possíveis (cada vértice oferece um prêmio). No PSBH, os vértices são as provas do espectro, os pesos nas arestas são iguais ao tamanho das provas menos o tamanho da sobreposição máxima entre as duas provas (incremento no tamanho da seqüência resultante causado quando uma prova é acrescentada ao final da seqüência) e o prêmio de cada vértice é igual a um (o objetivo é maximizar a quantidade de provas utilizadas).

Uma *instância* do PSBH é representada por um grafo dirigido $G = (V, E)$ com pesos nas arestas, onde $V = S$ é o conjunto de vértices do grafo e $E = \{(u, v) \mid \forall\ u, v \in S\}$ é o conjunto de arestas de G. O peso da aresta (u, v) é dado por

$$w(u,v) = \begin{cases} l - o(u,v), & \text{se } o(u,v) > 0 \\ \infty, & \text{caso contrario} \end{cases},$$

onde $o(u, v)$ indica a sobreposição máxima entre as provas u e v. O grafo que representa a instância do exemplo da Figura 2.2 é ilustrado na Figura 2.6. Neste caso, a solução para o problema de reconstrução consiste em um

PUC-Rio - Certificação Digital Nº 0310813/CA

caminho que visita todos os vértices do grafo e utiliza apenas arestas de peso unitário (a solução que reconstrói a seqüência original está em destaque no grafo). Estas características são devidas à ausência de erros no espectro. Inserindo-se um falso positivo e um falso negativo no espectro do exemplo (trocando a prova `GGCA` por `GGCG`), a solução ótima não passará por todos os vértices (devido ao falso positivo) e utilizará uma aresta com peso dois (devido ao falso negativo), como ilustrado na Figura 2.7.

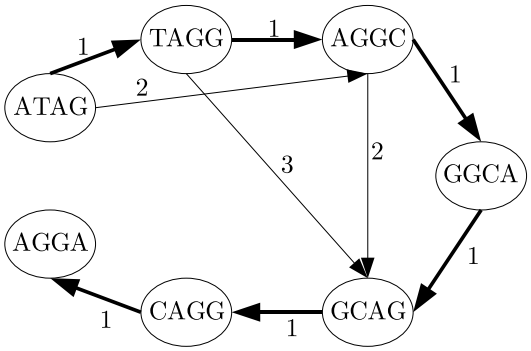

Figura 2.6: Grafo representando a instância do exemplo da Figura 2.2 (algumas arestas foram suprimidas para simplificar a leitura do grafo).

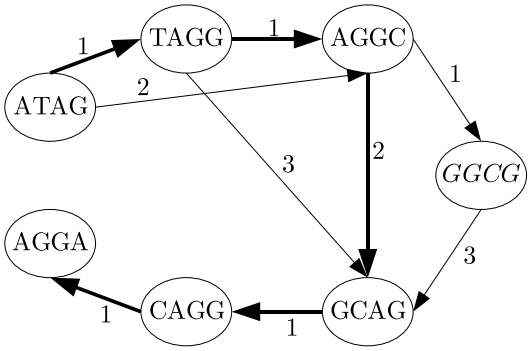

Figura 2.7: Grafo representando a instância do exemplo da Figura 2.2, com um falso positivo (`GGCG`) e um falso negativo (`GGCA`) (algumas arestas foram suprimidas para simplificar a leitura do grafo).

Uma *solução viável* para o PSBH é um caminho em G que parte do vértice inicial s_0, visita no máximo uma vez cada vértice e a soma dos pesos de suas arestas não ultrapassa $n - l$. Um caminho em G pode ser representado por uma seqüência de provas $a = (a_1, a_2, \ldots, a_k)$, tal que $a_i \in S$, $\forall\ i \in \{1, 2, \ldots, k\}$. Em relação ao caminho a, define-se que $S(a) = \{a_1, a_2, \ldots, a_k\}$ é o conjunto de provas do espectro utilizadas pelo caminho, $E(a) = \{(a_1, a_2), (a_2, a_3), \ldots, (a_{k-1}, a_k)\}$ é o conjunto de arestas utilizadas e $|a| = |S(a)|$ é a quantidade de provas utilizadas. Um caminho a qualquer representa uma solução viável para o PSBH se $a_1 = s_0$,

PUC-Rio - Certificação Digital Nº 0310813/CA

$a_i \neq a_j,\ \forall a_i, a_j \in S(a)$, e $w(a) \leq n - l$, onde $w(a) = \sum_{(u,v)\in E(a)} w(u,v)$ é a soma dos pesos das arestas utilizadas por a. Daqui em diante, uma solução será tratada como um caminho no grafo G. O objetivo no PSBH é encontrar uma solução a^* tal que $|a^*| = \max_{a\in A^v} |a|$, onde A^v é o conjunto de soluções viáveis para o problema. No exemplo da Figura 2.7, o espectro é $S = \{\text{ATAG}, \text{AGGA}, \text{AGGC}, \text{CAGG}, \text{GCAG}, \text{GGCG}, \text{TAGG}\}$ e os valores da função w estão especificados na Tabela 2.1. A solução ótima para esta instância é $a^* = (\text{ATAG}, \text{TAGG}, \text{AGGC}, \text{GCAG}, \text{CAGG}, \text{AGGA})$, portanto, $|a^*| = 6$ e $w(a^*) = 1 + 1 + 2 + 1 + 1 = 6 = n - l$.

	ATAG	AGGA	AGGC	CAGG	GCAG	GGCG	TAGG
ATAG	∞	2	2	∞	3	3	1
AGGA	3	3	3	∞	∞	∞	∞
AGGC	∞	∞	∞	3	2	1	∞
CAGG	∞	1	1	∞	3	2	∞
GCAG	∞	2	2	1	3	3	∞
GGCG	∞	∞	∞	∞	3	3	∞
TAGG	∞	1	1	∞	3	2	∞

Tabela 2.1: Valores da função w para o exemplo da Figura 2.7. As linhas correspondem ao primeiro parâmetro (origem da aresta) e as colunas ao segundo (destino da aresta).

PUC-Rio - Certificação Digital Nº 0310813/CA

2.8 Outras aplicações para o arranjo de DNA

Como o arranjo de DNA foi proposto inicialmente para realizar seqüenciamento de DNA, ele foi denominado seqüenciamento de DNA por hibridação. Entretanto, hoje em dia, eles são amplamente utilizados em outras tarefas e são denominados *DNA arrays* ou *DNA chips*. Uma aplicação muito bem sucedida é o reseqüenciamento de DNA, que consiste em detectar as diferenças entre uma molécula de DNA cuja seqüência já é conhecida e outra molécula mutante. Neste caso, é possível projetar um arranjo que utiliza provas maiores e consegue evitar os erros no espectro. Chee *et al.* em 1996 [8] projetaram um arranjo de DNA capaz de detectar mutações em moléculas completas do DNA mitocondrial humano (16569 bp) e conseguiram encontrar três mutações causadoras de doenças. Outra aplicação para o arranjo de DNA é a análise de expressão gênica [17], que consiste em detectar o nível de expressão de vários genes conhecidos através de um único experimento de hibridação.