Distância de reversão de cromossomos circulares

João Meidanis*
Instituto de Computação-UNICAMP

Maria Emilia M. T. Walter†
Departamento de Ciência da Computação-UnB
Instituto de Computação-UNICAMP

Zanoni Dias‡
Instituto de Computação-UNICAMP

Campinas, março de 1997

Sumário

Estudamos o problema de comparar dois cromossomos circulares, que evoluiram a partir de um ancestral comum por reversões, supondo que são conhecidas a ordem dos genes correspondentes e suas orientações. A distância de reversão é o menor número de reversões que transforma uma seqüência de números inteiros com sinais, que define um cromossomo, em outra, onde uma reversão toma uma subseqüência de elementos e reverte a sua ordem, trocando também os seus sinais. Apresentamos o primeiro algoritmo polinomial para determinar esta distância, baseado no algoritmo de Kaplan, Shamir e Tarjan para resolver o problema da distância de reversão de cromossomos lineares com sinais. Esclarecemos ainda alguns pontos sobre comparação de cromossomos lineares, e calculamos o diâmetro de reversão para cromossomos com sinais.

Sumário

We study the problem of comparing two circular chromosomes that evolved from a common ancestor by reversals, assuming that the order and orientations of corresponding genes are known. The reversal distance is the smallest number of reversals that transforms a sequence of signed integers, which defines a chromosome, into another, where a reversal acts on a subsequence of the elements, reverting their order and flipping their signs. We present the first polinomial algorithm to determine this distance, based on the algorithm by Kaplan, Shamir, and Tarjan that solves the corresponding problem for signed, linear chromosomes. We also clarify some issues on the comparison of linear chromosomes, and we compute the reversal diameter for signed chromosomes.

*e-mail:meidanis@dcc.unicamp.br
†e-mail:emilia@dcc.unicamp.br
‡e-mail:zanoni@dcc.unicamp.br

1 Introdução

O grande volume de dados, oriundos de seqüenciamento de genes em Biologia Molecular, vem suscitando um crescente interesse pelo desenvolvimento de algoritmos para comparação de genomas de espécies relacionadas, em termos de mutações ocorrendo em porções grandes dos seus cromossomos. Existem diversos tipos de eventos de mutação que podem ocorrer em genomas de organismos, dentre os quais, citamos *reversão*, que substitui uma seqüência de genes de uma região arbitrária do cromossomo pela sua seqüência complementar reversa. Isto tem o efeito de reverter a ordem dos genes nesta região, e de mudar a orientação de cada gene. Neste artigo estudamos a comparação de dois genomas constituídos por um único cromossomo circular, com base na ordem e orientação dos seus genes comuns, e em termos do evento de reversão.

Um cromossomo circular pode ser visto como um arranjo circular de blocos de genes, sendo que cada um dos blocos tem uma orientação. A Figura 1 mostra exemplos de cromossomos circulares de duas espécies de plantas, onde cada número representa um bloco contendo um ou mais genes, e as setas indicam as orientações relativas dos blocos de genes de uma espécie em relação a outra.

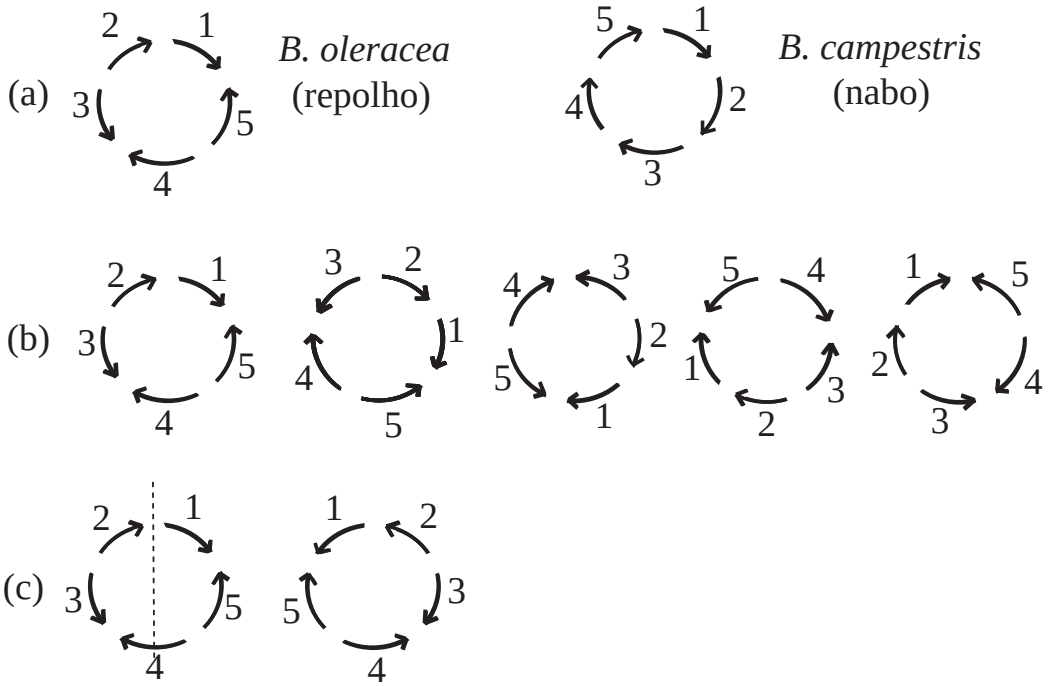

Figura 1: Exemplos de cromossomos circulares de duas espécies de plantas. (a) As setas indicam as orientações relativas dos blocos de genes de uma espécie em relação a outra. (b) Os exemplos mostram diferentes representações para o mesmo cromossomo. (c) Os exemplos mostram o mesmo cromossomo, considerando duas formas possíveis de observar os blocos de genes de um cromossomo circular. Estas duas formas são consideradas equivalentes, e estes cromossomos são obtidos um a partir do outro por reflexão em relação ao eixo da figura.

Uma reversão no cromossomo circular é obtida determinando-se dois pontos de

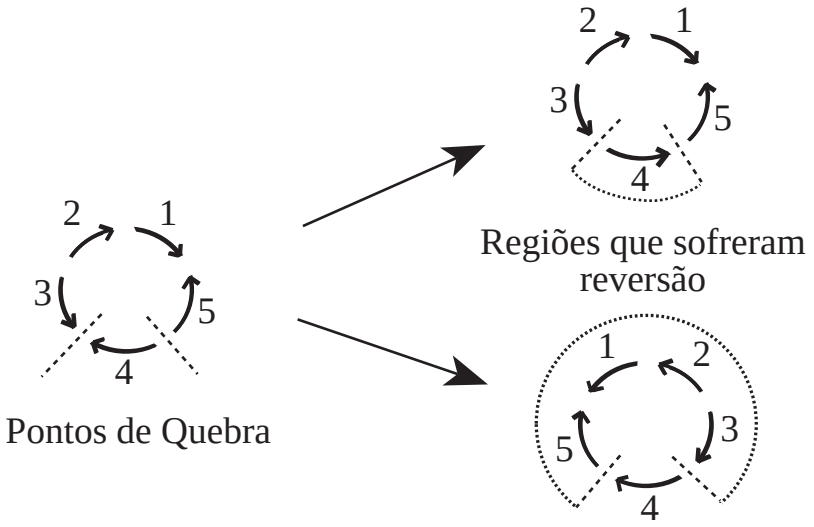

Figura 2: Este exemplo mostra as duas formas possíveis de ocorrer uma reversão em um cromossomo circular, dados os pontos onde ocorreram as quebras.

quebra no cromossomo, e invertendo-se uma das regiões que estes pontos delimitam (ver Figura 2).

De forma genérica, o *problema da distância de reversão para cromossomos circulares com sinais* é modelado da seguinte forma. Dados dois cromossomos circulares A e B, queremos encontrar a menor série de reversões necessárias para transformar A em B. O número mínimo de reversões é chamado de *distância de reversão*. A Figura 3 mostra um exemplo de transformação de um cromossomo circular em outro.

Uma variante deste problema surge quando não são conhecidas as orientações dos genes nos cromossomos. Neste caso, surge a versão *sem sinal* do problema, onde as reversões apenas invertem a ordem dos genes. Existem ainda outras versões do mesmo problema, considerando cromossomos lineares, e também outros eventos de mutação, além de reversão. A seguir, citamos apenas os trabalhos mais diretamente relacionados a este, pois a literatura é extremamente vasta.

Kececioglu e Sankoff [5] estudaram o problema da distância de reversão de permutações lineares sem sinais, e desenvolveram o primeiro algoritmo de aproximação para o problema. Bafna e Pevzner [1] posteriormente introduziram uma estrutura chamada de *grafo de pontos-de-quebra* de uma permutação inicial em relação a uma permutação alvo, e consideraram um novo parâmetro, baseado numa decomposição de ciclos alternantes de máxima cardinalidade. Hannenhalli e Pevzner [3] apresentam outros dois parâmetros no grafo de pontos-de-quebra: o número de obstáculos e o indicador de fortalezas, que, juntamente com o número de ciclos alternantes, permitiram provar um *teorema da dualidade*. Com base neste teorema, eles apresentaram o primeiro algoritmo polinomial para o problema da distância de reversão de permutações lineares com sinais, com complexidade de tempo $O(n^4)$. Berman e Hannenhalli [2] introduziram novas estruturas de dados neste algoritmo, e baixaram a sua complexidade para $O(n^2\alpha(n))$. Finalmente, Kaplan, Shamir e Tarjan [4], baseados na teoria de Hannenhalli e Pevzner, e utilizando parte do algoritmo de

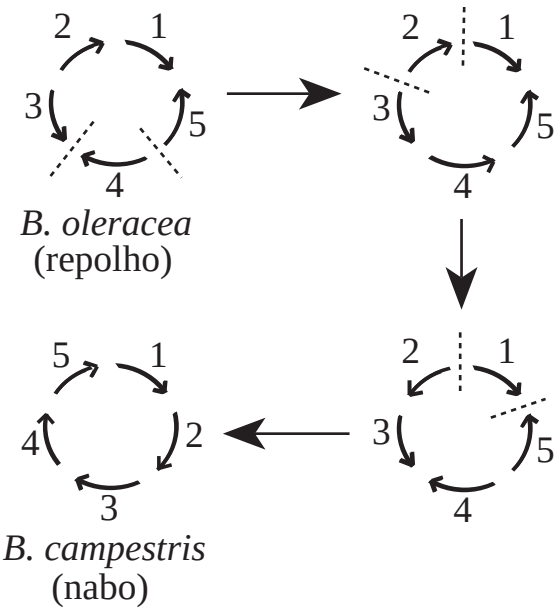

Figura 3: Este exemplo mostra uma série de reversões que transforma *B. oleracea* (repolho) em *B. campestris* (nabo).

Berman e Hannenhalli, mostraram um novo algoritmo com complexidade $O(n^2)$. Chamaremos este último de algoritmo KST.

Watterson e co-autores [9], num trabalho pioneiro, mostraram um algoritmo bastante simples para encontrar a distância de reversão de permutações circulares sem sinais. Kececioglu e Sankoff [6] apresentam um algoritmo exato branch-and-bound para o problema da distância de reversão para permutações circulares com sinais. Este algoritmo determinou limites extremamente precisos para a distância de reversão, em vários experimentos. Os autores relatam porém que não encontraram motivos para que estes limites fossem tão próximos.

Neste artigo mostramos um algoritmo polinomial para o problema da distância de reversão de cromossomos circulares com sinais. Para tanto, apresentamos uma formalização para cromossomos circulares e para reversões atuando neles. O algoritmo é baseado na teoria para o problema linear dada por Hannenhalli e Pevzner [3]. Além disso, elucidamos algumas questões envolvendo cromossomos lineares e calculamos o diâmetro de reversão para ambos os casos, linear e circular.

Na Seção 2, formalizamos um cromossomo circular por meio de uma classe de equivalência, e definimos reversões atuando em cromossomos circulares por meio de reversões atuando em cromossomos lineares. A Seção 3 traz um algoritmo polinomial para resolver o problema da distância de reversão de cromossomos circulares, que basicamente utiliza o algoritmo KST, tendo como entrada duas seqüências especiais que representam os cromossomos circulares. Na Seção 4 apresentamos resultados relativos a distâncias de reversão de cromossomos lineares e circulares. Na Seção 5 apresentamos o cálculo do diâmetro de reversão. Finalmente, a última seção traz as conclusões do trabalho e direções futuras.

2 Formalização do problema

Iniciamos esta seção com um breve resumo dos principais resultados conhecidos sobre cromossomos lineares com sinais, devidos principalmente a Bafna e Pevzner [1] e Hannenhalli e Pevzner [3]. Um cromossomo linear com sinais é representado por uma permutação com sinais. Uma *permutação com sinais* é uma permutação comum, mas na qual cada elemento tem sinal positivo (+) ou negativo (−) indicando a orientação relativa do bloco. Neste caso, uma *reversão* ϱ do intervalo $[i, j]$ é denotada por

$$\varrho(i,j) \cdot \pi = (\pi_1 \dots \pi_{i-1} \overline{\pi}_j \overline{\pi}_{j-1} \dots \overline{\pi}_{i+1} \overline{\pi}_i \pi_{j+1} \dots \pi_n)$$

onde $\overline{\pi}_k$ indica inversão do sinal de π_k.

O problema da distância de reversão para cromossomos lineares é modelado como se segue. Dadas duas permutações π e σ representando dois cromossomos lineares com sinais, o **problema da distância de reversão** entre π e σ é encontrar uma série de reversões $\varrho_1, \varrho_2, \dots, \varrho_t$ tal que $\varrho_t \cdot \varrho_{t-1} \cdot \dots \cdot \varrho_2 \cdot \varrho_1 \cdot \pi = \sigma$ e t é mínimo. Chamamos t de **distância de reversão entre** π **e** σ, denotado também por $d(\pi, \sigma)$.

Os trabalhos de Bafna e Pevzner [1] e de Hannenhalli e Pevzner [3] baseiam-se numa estrutura chamada de *grafo de pontos-de-quebra*. Este grafo é construído a partir de π e σ como segue. Cada inteiro com sinal é representado por uma seta, sendo esta da esquerda para a direita quando o sinal é +, e da direita para a esquerda caso contrário. Os pontos iniciais e finais destas setas são os vértices do grafo. Adicionam-se ainda dois pontos de referência, um à esquerda da seqüência e um à sua direita. A seguir, arestas ditas *de realidade* são colocadas entre extremos de setas adjacentes em π, e arestas ditas *de desejo* são incluidas entre extremos de setas adjacentes em σ. Propriedades importantes deste grafo são:

1. O grafo resultante é composto de uma coleção de ciclos pares. Quando $\pi = \sigma$, o número destes ciclos atinge seu valor máximo, que é $n + 1$. Para qualquer outro par de permutações, há menos de $n + 1$ ciclos.

2. Qualquer aresta realidade pertencente a um ciclo de tamanho maior do que 2 representa um *ponto de quebra* na permutação, isto é, um ponto no qual alguma reversão terá que passar para transformar π em σ. Quando dois vértices pertencem a um ciclo de tamanho 2, ou seja, são ligados por duas arestas paralelas, sendo uma de realidade e a outra de desejo, dizemos que *não há* uma quebra nesta posição.

A partir deste grafo são calculados três parâmetros que determinam a distância entre π e σ: o número de ciclos $c(\pi, \sigma)$, o número de obstáculos $h(\pi, \sigma)$ e o indicador de fortalezas $f(\pi, \sigma)$, este último podendo ser igual a zero ou um apenas. A distância é dada então pela fórmula:

$$d(\pi, \sigma) = n + 1 - c(\pi, \sigma) + h(\pi, \sigma) + f(\pi, \sigma).$$

Remetemos o leitor aos artigos relevantes [3, 2, 4] ou ao texto introdutório de Meidanis e Setubal [7] para uma explicação mais detalhada sobre estes parâmetros. Procuraremos esclarecer outras propriedades fundamentais desta construção à medida que forem necessárias neste texto.

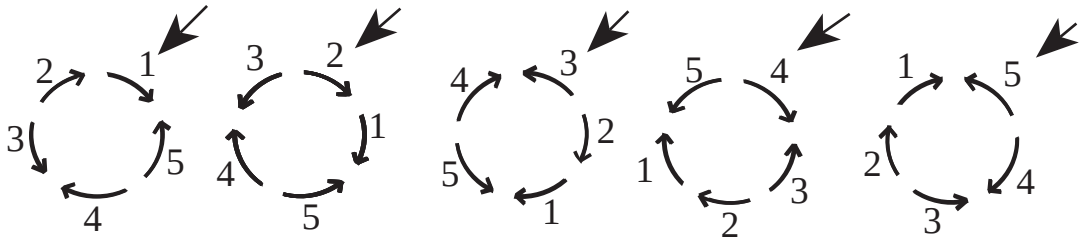

Figura 4: No cromossomo circular podemos considerar qualquer um dos blocos de genes como sendo o primeiro. Portanto, todas estas seqüências são consideradas equivalentes, e representam o cromossomo circular de *B. oleracea* mostrado na Figura 1 (a).

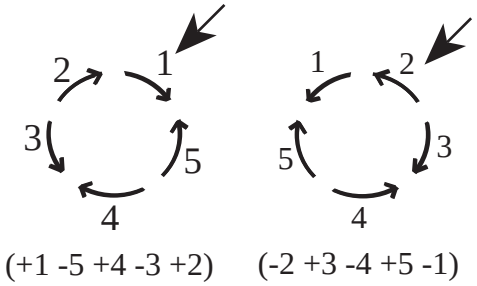

Figura 5: No cromossomo circular duas seqüências onde uma é obtida a partir da outra por reflexão são consideradas equivalentes. O cromossomo circular representado é *B. oleracea* mostrada na Figura 1 (a).

Mostramos agora uma representação de um cromossomo circular por uma classe de equivalência, e definimos como reversão atuará no cromossomo circular.

Dado um *ponto inicial* e uma *direção* (horária ou anti-horária), podemos representar o cromossomo circular por uma seqüência da forma seguinte. Um *bloco* de genes do cromossomo será modelado por um inteiro com sinal. O sinal "+" indica a seta com sentido horário da Figura 1, e o sinal "−" indica a seta com sentido anti-horário. Assim, $\pi = (\pi_1 \pi_2 \ldots \pi_n)$ denotará um cromossomo circular, com n blocos de genes. Por exemplo, o cromossomo de *B. oleracea* da Figura 1 pode ser representado pela seqüência $(+1\ -5\ +4\ -3\ +2)$.

Como podemos escolher qualquer um dos blocos como sendo o primeiro, temos diversas seqüências representando o mesmo cromossomo (ver Figura 4), e todas estas seqüências são consideradas equivalentes. Além disso, duas seqüências onde uma é obtida a partir da outra por reflexão são consideradas equivalentes (ver Figura 5). Por convenção, os blocos são lidos no sentido horário.

Definiremos a seguir as operações de *rotação* e *reflexão*, que formalizarão as duas características descritas acima, e a classe de equivalência que representará um cromossomo circular.

Seja S_n o conjunto de todas as possíveis seqüências de inteiros distintos com sinais, onde cada seqüência tem tamanho n. Os inteiros devem estar no intervalo $[1..n]$. Seja $\pi = (\pi_1 \pi_2 \ldots \pi_n)$ uma seqüência de S_n. Definimos dois tipos de operações atuando em π da seguinte forma:

- *Rotações.* Denotaremos pela letra r a rotação básica que desloca os elementos de uma permutação de uma posição para a esquerda, como segue:

$$r \cdot \pi = (\pi_2 \pi_3 \ldots \pi_n \pi_1).$$

As operações da forma r^i são chamadas de *rotações.*

- *Reflexões.* Denotaremos por s a reflexão básica que inverte a ordem dos elementos de uma permutação e também o sinal de todos eles. Assim,

$$s \cdot \pi = (\overline{\pi}_n \overline{\pi}_{n-1} \ldots \overline{\pi}_2 \overline{\pi}_1).$$

De modo geral, as operações da forma sr^i são chamadas de *reflexões.* Cada reflexão é igual à sua própria inversa.

A seguinte relação vale:

$$rs = sr^{-1}. \tag{1}$$

Definiremos agora uma relação de equivalência entre duas seqüências π e γ: $\pi \sim \gamma$ se e somente se existem $i, j \in Z$ tais que $\gamma = r^i s^j \cdot \pi$.

A relação acima é de equivalência. A verificação deste fato é simples. A Equação (1) é útil nesta verificação.

A partir desta relação de equivalência, definiremos a classe de equivalência da seqüência π, denotada por $[\pi]$, que representa um cromossomo circular com sinais, como

$$[\pi] = \{\gamma \in S_n | \pi \sim \gamma\}$$

Esta formalização é interessante sob o ponto de vista biológico, pois não fixa o primeiro elemento da seqüência, e portanto qualquer bloco de genes pode ser o primeiro, bastando aplicar uma rotação. Além disso, duas seqüências onde uma é obtida a partir da outra por reflexão podem ser produzidas por aplicação do operador s.

Formalizaremos agora como uma reversão atuará numa classe A que representa um cromossomo circular. Não será permitida uma escolha aleatória da seqüência em A na qual a reversão atuará. Definiremos uma *representante canônica* de A, denotada por $can(A)$, com as características de ter o bloco 1 fixado como sendo o primeiro, e ter a orientação $+$. Por exemplo:

$$\begin{aligned} A \;=\; \{\; & (+1-5+4-3+2)\,(-5+4-3+2+1)\,(+4-3+2+1-5) \\ & (-3+2+1-5+4)\,(+2+1-5+4-3)\,(-2+3-4+5-1) \\ & (+3-4+5-1-2)\,(-4+5-1-2+3)\,(+5-1-2+3-4) \\ & (-1-2+3-4+5)\;\} \\ can(A) \;=\; & (+1-5+4-3+2) \end{aligned}$$

Note que toda classe de equivalência possui uma única representante canônica. Em termos do formalismo, uma reversão será aplicada apenas na representante canônica.

Definiremos agora reversão atuando num cromossomo circular utilizando o formalismo já disponível de reversão atuando num cromossomo linear. Inicialmente apresentaremos um exemplo intuitivo e em seguida a formalização.

A Figura 6 mostra um exemplo para as duas possíveis formas de ocorrer uma reversão em um cromossomo circular, dados os dois pontos onde ocorreram as quebras. Podemos observar que as seqüências resultantes pertencem à mesma classe de equivalência.

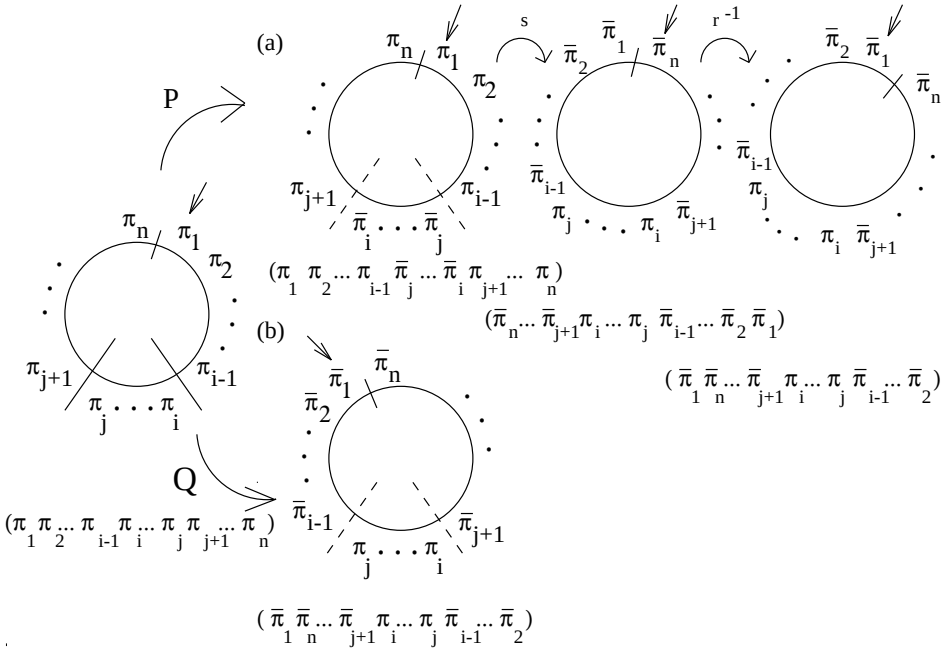

Figura 6: Este exemplo mostra que os dois cromossomos circulares resultantes de reversão são representados por duas seqüências que pertencem à mesma classe de equivalência. Observe que a seta, antes de ocorrer a reversão, aponta para o primeiro bloco da representante canônica da classe de equivalência que representa este cromossomo circular. O trecho do cromossomo que sofre a reversão pode envolver ou não a seta. (a) Neste caso, a reversão leva uma representante canônica em outra. (b) Neste caso, a reversão leva uma representante canônica em uma seqüência que não é representante canônica. Podemos aplicar reflexão e rotação numa das seqüências para obter a outra.

Teorema 2.1 *Para qualquer reversão circular P, existem inteiros i e j com $2 \leq i \leq j \leq n$ tais que*

$$P \cdot A = [\varrho(i,j) \cdot can(A)].$$

Demonstração: A prova baseia-se na Figura 6. Há dois casos, conforme a região a ser revertida contenha ou não o símbolo $+1$. Em ambos os casos, é fácil ver pela figura que obtem-se uma reversão $\varrho(i,j)$ com $2 \leq i \leq j \leq n$.

□

Nas condições do Teorema 2.1 denotamos P por $\varrho^c(i,j)$.

Podemos agora enunciar o problema da encontrar o menor número de reversões atuando em cromossomos circulares com orientações relativas conhecidas.

Dadas duas classes de equivalência A e B, que representam dois cromossomos circulares com orientações relativas conhecidas, o **problema da distância de reversão de cromossomos circulares com sinais** é encontrar uma série de reversões $P_1, P_2, \ldots, P_u$ de tal forma que $P_u \cdot P_{u-1} \cdot \ldots \cdot P_2 \cdot P_1 \cdot A = B$ e u seja mínimo. Chamamos u de distância de reversão entre A e B, denotado por $d^c(A,B)$.

3 Algoritmo MWD

A distância de reversão do cromossomo circular com sinais será calculada utilizando o algoritmo KST [4].

Temos imediatamente, pelas definições de r e s e de P e Q da Figura 6, um algoritmo exaustivo para calcular um limite superior para $d^c(A,B)$, que consiste em obter o mínimo entre $d(\pi,\sigma)$, para todos $\pi \in A, \sigma \in B$, utilizando o algoritmo KST. Temos $4n^2$ possíveis entradas para o algoritmo KST, pois temos $2n$ seqüências em A e $2n$ em B. Assim, a complexidade do algoritmo exaustivo é de $O(n^4)$.

Mas existe um meio de chamar o algoritmo KST uma única vez, fornecendo como entrada duas seqüências que levam a uma distância de reversão mínima, conforme descrito em seguida. O resultado principal nesta direção é o Teorema 3.1. Antes, porém, necessitamos de um resultado preliminar.

Lema 3.1 *Se $\pi_1 = +1$ e $2 \leq i \leq j \leq n$ então $[\varrho(i,j) \cdot \pi] = \varrho^c(i,j) \cdot [\pi]$.*

Esta prova é imediata das definições de reversão circular e de representante canônica.

O seguinte teorema resolve o problema da distância de reversão para cromossomos circulares com sinal.

Teorema 3.1 *Dados dois cromossomos circulares representados pelas classes A e B temos que $d^c(A,B) = d(can(A), can(B))$.*

Demonstração: Primeiro vamos mostrar que $d^c(A,B) \leq d(can(A), can(B))$. Sejam $\pi = can(A)$ e $\sigma = can(B)$. Um resultado que decorre da teoria de Hannenhalli e Pevzner afirma que é sempre possível obter uma seqüência mínima de reversões tal que nenhuma destas reversões acrescenta pontos de quebra onde não existiam. As reversões usadas por esta seqüência mínima especial para π, σ fornecem uma seqüência de reversões para o caso circular da seguinte forma.

De acordo com a definição, a primeira reversão será aplicada numa representante canônica da classe de equivalência que modela a seqüência inicial π, resultando numa representante canônica de uma outra classe de equivalência. E assim sucessivamente, as reversões sempre produzirão representantes canônicas, pois nunca vão atuar em $\pi_1 = +1$, pois isto criaria um ponto de quebra onde não havia. Estas reversões portanto sempre atuam no intervalo $[2,n]$ das permutações que representam as classes de equivalência.

Isto nos leva a uma série de reversões $\varrho_1, \ldots, \varrho_t$ tal que $\varrho_t \cdot \varrho_{t-1} \cdot \ldots \cdot \varrho_1 \cdot \pi = \sigma$, onde cada ϱ_k é igual a uma reversão da forma $\varrho_k(i,j)$, com $2 \leq i \leq j \leq n$. Assim,

$$[\varrho_t \cdot \varrho_{t-1} \cdot \ldots \cdot \varrho_2 \cdot \varrho_1 \cdot \pi] = [\sigma] = B$$

Pelo Lema 3.1,

$$\varrho_t^c \cdot \varrho_{t-1}^c \cdot \ldots \cdot \varrho_2^c \cdot \varrho_1^c \cdot [\pi] = B$$

$$\varrho_t^c \cdot \varrho_{t-1}^c \cdot \ldots \cdot \varrho_2^c \cdot \varrho_1^c \cdot A = B$$

Assim, $d^c(A,B) \leq t$ onde $t = d(can(A), can(B))$.

Em seguida, mostraremos que $d^c(A,B) \geq d(can(A), can(B))$. Para resolver o problema da distância de reversão do cromossomo circular com sinais, utilizamos

reversões no intervalo $[2, n]$, que atuam sempre na seqüência representante canônica. Em termos do cromossomo linear, no início, $\pi_1 = +1$ está na posição correta, e portanto não precisa ser modificado. Portanto elas fornecem uma série de reversões para o caso linear também. □

O algoritmo MWD consiste em chamar o algoritmo KST fornecendo como entrada as representantes canônicas de A e B. O algoritmo MWD está correto pelo Teorema 3.1, e sua complexidade é $O(n^2)$ (da complexidade do algoritmo KST), onde n é o número de blocos de genes dos cromossomos circulares.

4 Outros resultados

Nesta seção reunimos alguns resultados importantes a respeito de distâncias de reversões de permutações com sinais. Em primeiro lugar, damos uma fórmula para a distância de reversão de cromossomos circulares.

Em seguida, observamos que há menos reversões no caso circular do que no caso linear de mesmo tamanho. Os resultados subseqüentes ajudam a explicar uma prática muito freqüente em artigos onde se estuda como calcular distância de reversão de cromossomos lineares, que passamos a descrever. Dado um cromossomo linear, há essencialmente duas maneiras de escrevê-lo como uma permutação, sendo uma a reflexão da outra. Isto é devido ao fato de que moléculas de DNA livres não possuem uma extremidade distuingüível em geral, de modo que a leitura pode ser feita começando de qualquer uma das pontas. Assim sendo, ao comparar dois cromossomos, devemos na realidade considerar ambas as possibilidades para cada um deles e tomar o mínimo como sendo a distância. Contudo, em artigos, é comum a prática de tomar permutações com um extremo comum, se possível. Nossos resultados justificam isso mostrando que, neste caso, os representantes escolhidos conduzem ao mínimo.

Por fim, provamos um resultado ligando distâncias lineares e circulares.

Teorema 4.1 *Dadas duas classes A e B que representam dois cromossomos circulares com sinais, temos que*

$$d^c(A, B) = d(can(A), can(B)) =$$

$$(n+1) - c(can(A), can(B)) + h(can(A), can(B)) + f(can(A), can(B))$$

Este teorema decorre imediatamente do Teorema 3.1 e da distância de reversão de cromossomos lineares proposta por Hannenhalli e Pevzner [3].

Temos como motivação para os próximos dois teoremas, o fato de que a distância real de reversão para cromossomos lineares π e σ é dada por

$$\min\{d(\pi, \sigma), d(s \cdot \pi, \sigma), d(\pi, s \cdot \sigma), d(s \cdot \pi, s \cdot \sigma)\} \tag{2}$$

Então, considerando o Teorema 4.2, e se tivermos representantes com as características do Teorema 4.3, temos que os resultados encontrados nos artigos calculando apenas $d(\pi, \sigma)$ é realmente o mínimo entre os quatro.

Enunciamos agora o teorema que reduz para duas as possibilidades para calcular a menor distância de reversão, formulada na Equação (2). Sua demonstração é simples e será omitida.

Teorema 4.2 *Dadas duas permutações π e σ quaisquer, então*

$$d(\pi, \sigma) = d(s \cdot \pi, s \cdot \sigma) \qquad e \qquad d(s \cdot \pi, \sigma) = d(\pi, s \cdot \sigma)$$

O teorema seguinte justifica a prática de escolher, para o cálculo da distância, permutações que começam ou terminam com a mesma seqüência de genes e com a mesma orientação.

Teorema 4.3 *Se dadas duas seqüências π e σ tal que $\pi_1 = \sigma_1$ ou $\pi_n = \sigma_n$ então*

$$d(\pi, \sigma) \leq d(\pi, s \cdot \sigma)$$

Gostaríamos agora de saber qual é a relação entre $d(\pi, \sigma)$ e $d^c([\pi], [\sigma])$ para π e σ quaisquer. Temos o seguinte resultado.

Teorema 4.4 *Dadas duas permutações π e σ quaisquer,*

$$d(\pi, \sigma) \geq d^c([\pi], [\sigma])$$

A prova deste teorema encontra-se na versão estendida deste trabalho.

O teorema seguinte nos permite verificar que as representantes canônicas das classes que representam os cromossomos circulares, dadas como entrada para o algoritmo MWD, fornecem $u = d^c(A, B)$ mínimo, dentre todas as permutações que pertencem às duas classes, dadas como entrada para o algoritmo exaustivo.

Teorema 4.5 *Dadas duas permutações π e σ quaisquer, e as duas classes correspondentes $A = [\pi]$ e $B = [\sigma]$, $\pi \in A$ e $\sigma \in B$, temos que*

$$d(can(A), can(B)) = \min_{\substack{\pi \in A \\ \sigma \in B}} \{d(\pi, \sigma)\}$$

Este teorema decorre dos Teoremas 3.1 e 4.4.

5 O diâmetro de reversão de cromossomos com sinais

O **diâmetro de reversão circular**, denotado por $D^c(n)$, do conjunto das classes de equivalência em S_n, com respeito à distância de reversão circular, é o máximo número de reversões necessárias para transformar uma classe de equivalência em outra. Analogamente, o **diâmetro de reversão linear**, denotado por $D(n)$, do conjunto S_n de permutações de n elementos, com respeito à distância de reversão linear, é o máximo número de reversões necessárias para transformar uma permutação de n elementos em outra. Mostramos agora que o diâmetro de reversão de cromossomos circulares e lineares com sinais é respectivamente n e $n+1$ (exceto em alguns poucos casos). Isto vem a corrigir uma afirmação feita por Kececioglu e Sankoff de que $n - 2 \leq D(n) \leq n - 1$ [6].

Teorema 5.1 *O diâmetro de reversão para cromossomos circulares e lineares com sinais é respectivamente*

$$D^c(n) = \max_{\substack{A \in S_n^c \\ B \in S_n^c}} \{d^c(A,B)\} = \begin{cases} n-1 & \text{se } n = 1,\ n = 2 \text{ ou } n = 4 \\ n & \text{caso contrário} \end{cases}$$

$$D(n) = \max_{\substack{\pi \in S_n \\ \sigma \in S_n}} \{d(\pi,\sigma)\} = \begin{cases} n & \text{se } n = 1 \text{ ou } n = 3 \\ n+1 & \text{caso contrário} \end{cases}$$

Demonstração: Para o caso do cromossomo circular, é possível exibir duas classes de equivalência, $[\pi_n]$ e $[\sigma_n]$, para as quais $d^c([\pi_n],[\sigma_n])$ é n. Isto é feito construindo permutações que geram, no grafo de pontos-de-quebra, obstáculos com número ímpar de arestas realidade, ou combinações de obstáculos com 3 e 5 arestas realidade. A construção só não funciona nos casos excepcionais destacados no enunciado. Com isto temos que $D^c(n) \geq n$. Usando o Teorema de Kececioglu e Sankoff [6] ($n - 1 \leq D^c(n) \leq n$) temos o resultado do teorema.

Neste mesmo trabalho, Kececioglu e Sankoff dão limites para o diâmetro linear que são incorretos, e que corrigimos a seguir. Para o caso do cromossomo linear, de forma análoga ao circular, temos duas seqüências, π_n e σ_n, para as quais $d(\pi_n,\sigma_n)$ é $n+1$, para cada n, exceto os destacados no enunciado. Temos que provar ainda que $D(n) < n+2$, para obter o resultado desejado. Pela fórmula de Hannenhalli e Pevzner [3] temos que: $d(\pi_n,\iota_n) = (n+1) - c(\pi_n,\iota_n) + h(\pi_n,\iota_n) + f(\pi_n,\iota_n)$. Primeiro, temos que $h(\pi_n,\iota_n) \leq c(\pi_n,\iota_n)$, pela definição de $h(\pi_n,\iota_n)$. Assim, se $h(\pi_n,\iota_n) = c(\pi_n,\iota_n)$ (máximo possível), $d(\pi_n,\iota_n) \leq (n+1)+1$, ou seja, $d(\pi_n,\iota_n) \leq n+2$. Mas para $f(\pi_n,\iota_n) = 1$, $h(\pi_n,\iota_n) < c(\pi_n,\iota_n)$, e portanto, $d(\pi_n,\iota_n) < n+2$.

□

6 Conclusões

Neste texto procuramos sistematizar um pouco mais a teoria sobre problemas de distância de reversão para cromossomos com sinal. Para tanto, contribuímos nos aspectos descritos em seguida. Para o problema da distância de reversão de cromossomos circulares com sinais, propusemos um algoritmo polinomial, baseado no algoritmo polinomial de Kaplan, Shamir e Tarjan [4]. Esta solução inclui as diferentes possibilidades de visualizar um cromossomo circular com sinais, obtidas uma a partir da outra por rotações e reflexões. Determinamos o diâmetro de reversão para permutações lineares ($D(n) = n+1$) e circulares ($D(n) = n$), com sinais, corrigindo uma afirmação do artigo de Kececioglu e Sankoff [6] sobre o diâmetro linear $D(n)$. Finalmente, justificamos a prática, comum nos artigos, de calcular só $d(\pi,\sigma)$, fixando a priori uma das pontas das moléculas de DNA. Provamos que quando $\pi_1 = \sigma_1$ ou $\pi_n = \sigma_n$ é suficiente calcular $d(\pi,\sigma)$.

A partir destes estudos, surgiram algumas questões que passamos a comentar. Primeiro, em que condições temos $d(\pi,\sigma) \leq d(\pi, s \cdot \sigma)$? Neste trabalho, demos condições suficientes para que isto ocorra (Teorema 4.3). Contudo, um exemplo apresentado por Palmer e co-autores [8] não seguia este padrão, mas também levava a uma distância mínima. Por fim, chamamos de *representantes ótimas* de duas classes, que representam dois cromossomos circulares, duas permutações, uma de

cada classe, que levam a uma distância de reversão mínima. A questão neste caso é como caracterizar este conjunto das representantes ótimas.

Referências

[1] V. Bafna and P. Pevzner. Genome rearrangements and sorting by reversals. In *34th Annual IEEE Symposium on Foundations of Computer Science*, pages 148–157, 1993. Versao estendida foi publicada com o mesmo titulo em SIAM Journal of Computing, 25(2):272-289,1996.

[2] P. Berman and S. Hannenhalli. Fast sorting by reversals. In *Proceedings of Combinatorial Pattern Matching - CPM'96*, 1996.

[3] S. Hannenhalli and P. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, pages 178–189, 1995.

[4] H. Kaplan, R. Shamir, and R. E. Tarjan. Faster and simpler algorithm for sorting signed permutations by reversals. *SODA'97*, 1997.

[5] J. Kececioglu and D. Sankoff. Exact and approximation algorithms for the inversion distance between two permutations. In *Combinatorial Pattern Matching, Proc. 4th Annual Symposium (CPM'93) de Lecture Notes in Computer Science*, volume 684, pages 87–105, Berlin, 1993. Springer-Verlag. Versao estendida com titulo: "Exact and approximation algorithms for the inversion distance between two permutations" foi publicada em Algoritmica, 13: 180-210, 1995.

[6] J. Kececioglu and D. Sankoff. Efficient bounds for oriented chromosome inversion distance. *Lecture Notes in Computer Science*, 807:307–325, 1994.

[7] J. Meidanis and J. C. Setubal. *Introduction to Computational Molecular Biology.* PWS Publishing Co, 1997.

[8] J.D. Palmer, B. Osorio, and W.F. Thompson. Evolutionary significance of inversions in legume chloroplast dnas. *Current Genetics*, 14:65–74, 1988.

[9] G. A. Watterson, W. J. Ewens, and T. E. Hall. The chromosome inversion problem. *J. Theor. Biol.*, 99:1–7, 1982.