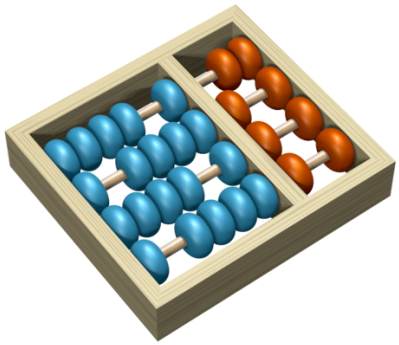
**Universidade Estadual de Campinas**

Instituto de Computação



**Exame de Qualificação Específico de Doutorado**

---

# Problemas de Rearranjos de Genomas Considerando Regiões Intergênicas ou Genes Repetidos

---

*Aluno:* Klairton de Lima Brito

*Orientador:* Prof. Dr. Zanoni Dias

*Coorientador:* Prof. Dr. Ulisses Martins Dias

**Resumo**

Rearranjos de genomas são eventos mutacionais que podem atuar em grandes trechos do genoma afetando o material genético. Dentre os eventos de rearranjos de genomas mais estudados podemos citar reversão, transposição, inserção e deleção. Determinar a sequência de rearranjos capaz de transformar o genoma de um organismo em outro é uma questão importante na genômica comparativa. Além disso, modelagens propostas para essa questão são linhas de pesquisas relevantes no contexto de Otimização Combinatória. Diversos estudos foram realizados e existem modelos que caracterizam o genoma desde como uma permutação até como uma cadeia de caracteres (string), representando a sequência de genes encontrada no genoma. Estudos mais recentes consideram, além da sequência de genes, informações adicionais presentes no genoma. Em particular, estruturas chamadas de regiões intergênicas, presentes entre cada par de genes consecutivos, têm atraído a atenção dos pesquisadores. Quando o genoma não apresenta genes repetidos, a representação é dada na forma de uma permutação e o problema é conhecido como Ordenação de Permutações. Caso contrário, a representação é dada na forma de uma cadeia de caracteres e o problema recebe o nome de Distância de Strings. O trabalho em questão tem como objetivo investigar as versões do problema de Ordenação de Permutações com Regiões Intergênicas considerando os eventos de reversão, transposição, inserção e deleção. Além disso, pretendemos investigar o problema de Distância de Strings considerando os eventos de reversão e transposição em instâncias em que cada gene apresenta no máximo $k$ cópias.

# 1 Introdução

A genômica comparativa consiste no estudo de similaridades entre os genomas de diferentes organismos. Uma forma bem aceita de realizar essa comparação é determinando o menor número de eventos de rearranjos de genomas capaz de transformar o genoma de um organismo em outro. Rearranjos de genomas são eventos mutacionais que podem atuar em grandes trechos do genoma modificando, inserindo ou removendo material genético [1].

Reversão e transposição são os eventos de rearranjo mais estudados na literatura [2, 3]. Uma reversão atua em um segmento do genoma invertendo a posição e a orientação dos

genes contidos no segmento, enquanto uma transposição troca dois segmentos consecutivos do genoma, mas sem afetar a posição e a orientação dos genes nos segmentos. Os eventos de reversão e transposição são chamados de conservativos, pois não alteram a quantidade de material genético do genoma. Existem também eventos não conservativos, como é o caso dos eventos de inserção e deleção [4], que inserem e removem material genético de uma região específica do genoma, respectivamente. Um modelo de rearranjo determina o conjunto de eventos de rearranjo permitidos para transformar um genoma em outro. O tamanho da menor sequência de eventos de rearranjos de genomas capaz de realizar essa tarefa é chamado de distância de rearranjo.

Um genoma pode ser representado computacionalmente de diferentes maneiras. Quando o genoma é tratado como uma sequência ordenada de genes, podemos encontrar casos em que determinados genes apresentam múltiplas cópias, sendo comum utilizarmos uma representação na forma de uma cadeia de caracteres, onde cada caractere é associado a um gene. Por outro lado, se existir apenas uma cópia de cada gene, podemos associar um número inteiro para cada gene e a representação é dada na forma de uma permutação. Em ambos os casos, quando a orientação dos genes é conhecida, um sinal de positivo ou negativo é atribuído para cada elemento e a representação é chamada com sinais (string com sinais e permutação com sinais). Caso contrário, o sinal é omitido e a representação é chamada sem sinais (string sem sinais e permutação sem sinais).

Ao utilizar a representação de um genoma como uma permutação, podemos simplificar o problema como sendo um problema de ordenação. Nesse caso, o objetivo consiste em transformar uma permutação $\pi$ qualquer em uma permutação específica na qual os elementos encontram-se ordenados de maneira crescente e com sinal positivo para o caso com sinais, essa permutação é chamada de identidade.

A representação do genoma como uma sequência ordenada de genes é uma abordagem simples e prática, mas acarreta na perda de informação referente às estruturas genéticas que não fazem parte da sequência de genes. Estudos recentes [5, 6] mostram que considerar informações adicionais contidas no genoma, além da sequência de genes, pode tornar a comparação entre genomas mais realista. Em particular, os pesquisadores apontaram a

importância de considerar o tamanho das regiões presentes entre cada par de genes consecutivos, chamadas de regiões intergênicas.

Esse projeto tem como objetivo a investigação de versões de dois problemas importantes no campo da Biologia Computacional que estão diretamente conectados com o ramo da Otimização Combinatória, sendo eles:

i. Problema de Ordenação de Permutações com Regiões Intergênicas: nesse caso, assumimos que existe apenas uma cópia de cada gene e a representação do genoma é dada por meio de permutações com e sem sinais, juntamente com as informações dos tamanhos das regiões intergênicas. Os eventos de rearranjo considerados serão reversão, transposição, inserção e deleção.

ii. Problema de Distância de Strings: nesse caso, podem existir cópias de genes e a representação do genoma é dada por meio de strings de caracteres com e sem sinais. Os eventos de rearranjo considerados serão reversão e transposição. Para as versões desse problema, estamos interessados em instâncias onde cada gene tem no máximo $k$ cópias.

O restante desse documento encontra-se dividido da seguinte forma. A Seção 2 apresenta uma revisão da literatura. A Seção 3 introduz definições e apresenta de maneira formal os problemas que serão investigados. A Seção 4 mostra os objetivos que pretendemos alcançar. A Seção 5 exibe um cronograma das atividades previstas até o fim do projeto. A Seção 6 resume os métodos que serão aplicados no desenvolvimento das atividades e a Seção 7 finaliza o documento apresentando as formas previstas para a análise dos resultados.

## 2 Revisão da Literatura

Essa seção apresenta uma síntese da bibliografia fundamental dos trabalhos existentes na literatura, bem como os melhores resultados obtidos até então.

Quando consideramos um modelo de rearranjo composto apenas pelo evento de reversão e permutações com sinais, temos o problema de Ordenação de Permutações com Sinais por Reversões. Hannenhalli e Pevzner [7] apresentaram o primeiro algoritmo exato em tempo

polinomial para o problema, sendo posteriormente simplificado por Bergeron [2]. Atualmente, temos um algoritmo com complexidade subquadrática para determinar a sequência de reversões capaz de ordenar uma permutação com sinais [8]. Entretanto, se estivermos interessados somente na distância de reversão, existe um algoritmo que executa em tempo linear [9]. Adotando o mesmo modelo, mas agora com permutações sem sinais, temos o problema de Ordenação de Permutações sem Sinais por Reversões. Caprara [10] provou que o problema faz parte da classe de problemas NP-Difícil. Um dos primeiros algoritmos para o problema apresenta um fator de aproximação 1.75 [11]. Em seguida, Christie [12] apresentou um algoritmo com fator de aproximação 1.5. Atualmente, o melhor algoritmo para o problema apresenta um fator de aproximação 1.375 [13].

Quando consideramos um modelo de rearranjo composto apenas pelo evento de transposição, a orientação dos genes não precisa ser considerada, tendo em vista que o evento de transposição não altera a orientação dos genes. Dessa forma, ao adotar permutações sem sinais, temos o problema de Ordenação de Permutações sem Sinais por Transposições. O problema também pertence à classe de problemas NP-Difícil, sendo a prova apresentada por Bulteau *et al.* [14]. O primeiro algoritmo para o problema foi proposto por Bafna e Pevzner [3] com um fator de aproximação 1.5. Posteriormente, Christie [15] apresentou uma implementação mais simples para esse algoritmo. Atualmente, o melhor algoritmo para o problema apresenta um fator de aproximação 1.375 [16] e heurísticas foram apresentadas por Dias e Dias [17] visando a obtenção de resultados práticos melhores.

Ao considerar um modelo de rearranjo composto pelos eventos de reversão e transposição em permutações com e sem sinais, obtemos os problemas de Ordenação de Permutações com Sinais por Reversões e Transposições, e Ordenação de Permutações sem Sinais por Reversões e Transposições, respectivamente. Recentemente, foi provado que ambos os problemas pertencem à classe de problemas NP-Difícil [18]. Os melhores algoritmos para os problemas apresentam fatores de aproximação 2 [19] e $2k$ [20] (onde $k$ é o fator de aproximação para a decomposição de ciclos [21]) para os casos com e sem sinais, respectivamente. Diversas heurísticas considerando esses problemas foram apresentadas na literatura [22, 23].

Quando passamos a considerar que o genoma pode apresentar genes repetidos, em 2001,

Christie e Irving [24] mostraram que o problema de Distância de Strings sem Sinais por Reversões pertence à classe de problemas NP-Difícil, mesmo se considerarmos um alfabeto binário (os caracteres das strings comparadas pertencem ao conjunto $\{0, 1\}$). Para isso, os autores apresentaram uma redução do problema 3-partition [25]. Em 2005, Radcliffe *et al.* [26] mostraram que a Distância de Strings com Sinais por Reversões e Distância de Strings sem Sinais por Transposições também pertencem à classe de problemas NP-Difícil, mesmo se considerarmos um alfabeto binário. Outra contribuição importante do trabalho foi que os autores caracterizaram um conjunto de instâncias em que é possível obter uma solução ótima em tempo polinomial.

Uma relação entre o problema de Distância de Strings por Reversões e o problema de Partição Mínima em Strings foi apresentada por Chen *et al.* [27]. Com essa relação entre os problemas, foi apresentado por Kolman e Waleń [28] um algoritmo de aproximação com fator $\Theta(k)$ para o problema de Distância de Strings com e sem Sinais por Reversões, onde $k$ representa o número máximo de cópias de um caractere nas strings consideradas.

Trabalhos que levam em conta a sequência de genes e também consideram os tamanhos das regiões intergênicas começaram a ser apresentados recentemente. Fertin *et al.* [29] apresentaram um modelo que permite o uso do evento de rearranjo Double-Cut and Join (DCJ), mostraram que o problema pertence à classe de problemas NP-Difícil e desenvolveram um algoritmo de aproximação com fator 4/3. O evento de rearranjo DJC [30] atua fragmentando o genoma em dois pontos e, em seguida, as extremidades dos segmentos resultantes são unidas obedecendo certas restrições. Bulteau *et al.* [31] apresentaram um modelo que permite o uso do evento DCJ juntamente com os eventos não conservativos de inserção e deleção restritos a atuarem apenas sobre as regiões intergênicas. Para esse problema, os autores apresentaram um algoritmo exato em tempo polinomial. Oliveira *et al.* [32] apresentaram um modelo que permite o uso apenas de reversões super curtas (esse evento de rearranjo possui uma restrição adicional que faz com que todo evento de reversão afete um segmento com no máximo dois genes). Juntamente com o modelo, os autores desenvolveram algoritmos de aproximação para o problema de forma geral e para instâncias do problema com características específicas.

Trabalhos considerando a ordem dos genes e o tamanho das regiões intergênicas são recentes, sendo uma promissora linha de pesquisa, tendo em vista as melhorias que podem ser obtidas nas estimativas para a distância evolutiva entre os organismos.

# 3 Problemas

Nessa seção, apresentamos formalmente algumas definições e os problemas que serão investigados.

## 3.1 Ordenação de Permutações com Regiões Intergênicas

Para esse problema, assumimos que os genomas não possuem genes repetidos e compartilham o mesmo conjunto de genes. Um genoma $\mathcal{G}$ é dado como uma sequência de $n$ genes ($g_i$) intercalados por uma sequência de $n+1$ regiões intergênicas ($r_i$), isto é, $\mathcal{G} = (r_1, g_1, r_2, g_2, \ldots, r_n, g_n,\ r_{n+1})$. Cada região intergênica possui uma quantidade bem definida de nucleotídeos que, a partir de agora, chamaremos de tamanho da região intergênica. Os eventos de rearranjo podem atuar sobre essas regiões dividindo-as em porções com diferentes tamanhos. Dessa forma, visando simular esse comportamento e manter a representatividade de um genoma qualquer, podemos utilizar dois elementos: (i) uma permutação $\pi$ e (ii) uma lista ordenada $\breve{\pi}$ que representam a sequência de genes e os tamanhos das regiões intergênicas, respectivamente.

Como estamos representando a sequência de genes como uma permutação, podemos tratar o problema como um problema de ordenação em que queremos transformar a permutação $\pi$ na permutação $\iota$ [33]. Além disso, queremos alterar os tamanhos das regiões intergênicas $\breve{\pi}$ de maneira a obtermos tamanhos específicos definidos por $\breve{\iota}$. Dessa forma, uma instância para o problema consiste de três elementos $(\pi, \breve{\pi}, \breve{\iota})$, tendo em vista que a permutação identidade $\iota$ pode ser facilmente computada a partir da quantidade de elementos da permutação $\pi$. A seguir, definimos formalmente os eventos de interesse para essa linha de pesquisa levando em conta as regiões intergênicas.

**Definição 3.1** *Uma inserção intergênica $\phi_x^i$, tal que $1 \leq i \leq (n+1)$, $x > 0$ e $x \in \mathbb{N}$ atua na região intergênica $\breve{\pi}_i$ inserindo uma quantidade $x$ de nucleotídeos.*

**Definição 3.2** *Uma deleção intergênica $\psi_x^i$, tal que $1 \leq i \leq (n+1)$, $0 < x \leq \breve{\pi}_i$ e $x \in \mathbb{N}$ atua na região intergênica $\breve{\pi}_i$ removendo uma quantidade $x$ de nucleotídeos.*

A Figura 1 mostra de maneira genérica uma inserção e deleção intergênica.

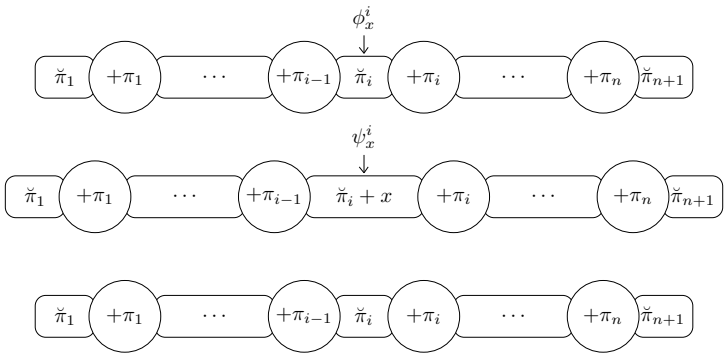



Figura 1: Ilustração de uma inserção intergênica seguida por uma deleção intergênica.

**Definição 3.3** *Uma reversão intergênica $\rho_{(x,y)}^{(i,j)}$, com $1 \leq i \leq j \leq n$, $0 \leq x \leq \breve{\pi}_i$, $0 \leq y \leq \breve{\pi}_{j+1}$ e $\{x, y\} \subset \mathbb{N}$, divide as regiões intergênicas $\breve{\pi}_i$ e $\breve{\pi}_{j+1}$ em partes com tamanhos $(x, x' = \breve{\pi}_i - x)$ e $(y, y' = \breve{\pi}_{j+1} - y)$, respectivamente. O segmento $(x', \pi_i, \ldots, \pi_j, y)$ é invertido e as partes de tamnho $(x, y)$ e $(x', y')$ se unem formando as novas regiões intergênicas $\breve{\pi}_i$ e $\breve{\pi}_{j+1}$, respectivamente. Caso a permutação $\pi$ possua sinais, a orientação dos elementos do segmento também é invertida.*

A Figura 2 mostra de maneira genérica uma reversão intergênica.

**Definição 3.4** *Uma transposição intergênica $\tau_{(x,y,z)}^{(i,j,k)}$, com $1 \leq i < j < k \leq n+1$, $0 \leq x \leq \breve{\pi}_i$, $0 \leq y \leq \breve{\pi}_j$, $0 \leq z \leq \breve{\pi}_k$ e $\{x, y, z\} \subset \mathbb{N}$, divide as regiões intergênicas $\breve{\pi}_i$, $\breve{\pi}_j$ e $\breve{\pi}_k$ em partes com tamanhos $(x, x' = \breve{\pi}_i - x)$, $(y, y' = \breve{\pi}_j - y)$ e $(z, z' = \breve{\pi}_k - z)$, respectivamente. Os segmentos $(x', \pi_i, \ldots, y)$ e $(y', \pi_j, \ldots, z)$ trocam de posição sem alterar as orientações dos genes e as partes de tamnho $(x, y')$, $(z, x')$ e $(y, z')$ se unem formando as novas regiões intergênicas $\breve{\pi}_i$, $\breve{\pi}_{k+i-j}$ e $\breve{\pi}_k$, respectivamente.*

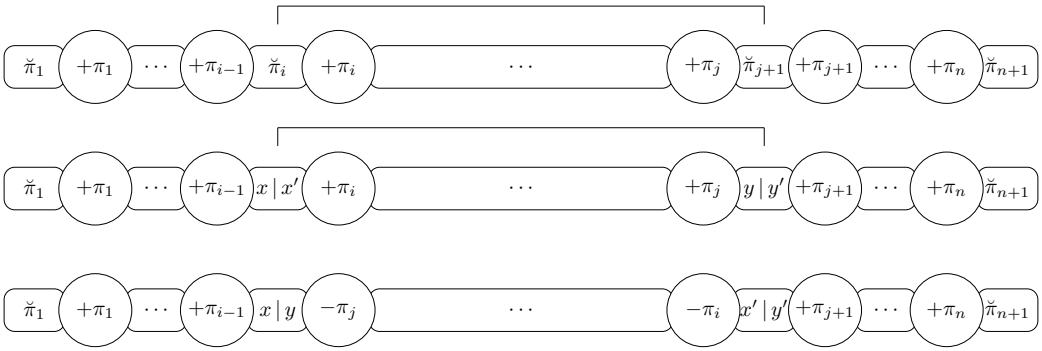



Figura 2: Ilustração de uma reversão intergênica.

A Figura 3 mostra de maneira genérica uma transposição intergênica.

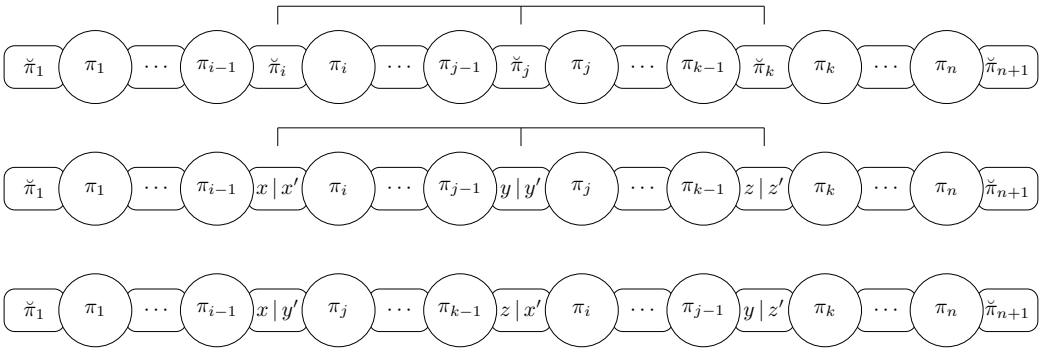



Figura 3: Ilustração de uma transposição intergênica.

O objetivo do problema de Ordenação de Permutações com Regiões Intergênicas consiste em encontrar a menor sequência de rearranjos de genomas permitidos por um modelo $\mathcal{M}$ capaz de transformar a permutação $\pi$ na permutação $\iota$ e as regiões intergênicas $\breve{\pi}$ nas regiões intergênicas $\breve{\iota}$. O tamanho da menor sequência de rearranjos de genomas capaz de realizar essa tarefa é chamado de distância de rearranjo, denotada por $d_{\mathcal{M}}(\pi, \breve{\pi}, \breve{\iota})$.

Dado um modelo de rearranjo que permite apenas reversões intergências, a Figura 4 mostra uma sequência de reversões intergênicas $S_\rho$ capaz de transformar $(\pi, \breve{\pi})$ em $(\iota, \breve{\iota})$.

Dado um modelo de rearranjo que permite apenas transposições intergências, a Figura 5 mostra uma sequência de transposições intergênicas $S_\tau$ capaz de transformar $(\pi, \breve{\pi})$ em $(\iota, \breve{\iota})$.

Dado um modelo de rearranjo que permite reversões e transposições intergências, a Fi-

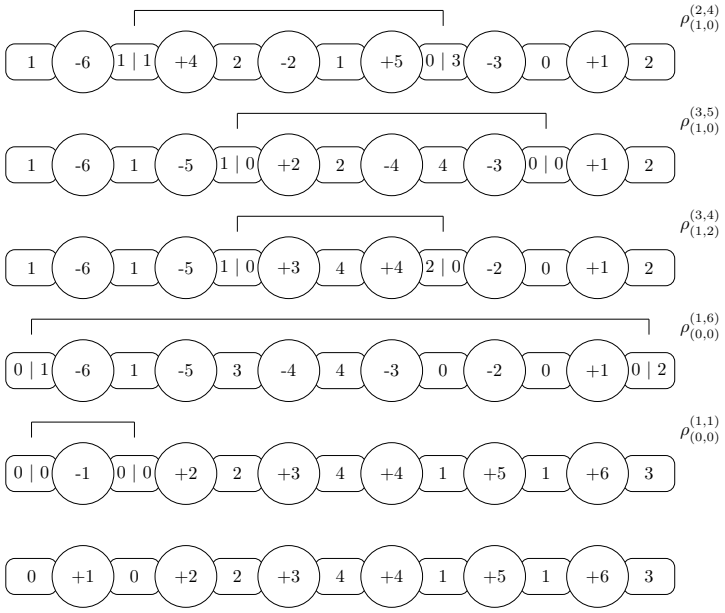



Figura 4: Sequência de reversões intergênicas $S_\rho = \langle \rho^{(2,4)}_{(1,0)}, \rho^{(3,5)}_{(1,0)}, \rho^{(3,4)}_{(1,2)}, \rho^{(1,6)}_{(0,0)}, \rho^{(1,1)}_{(0,0)} \rangle$ sendo aplicada na instância $(\pi, \breve{\pi}, \breve{\iota})$, tal que $\pi = (-6\ +4\ -2\ +5\ -3\ +1)$, $\breve{\pi} = (1, 2, 2, 1, 3, 0, 2)$ e $\breve{\iota} = (0, 0, 2, 4, 1, 1, 3)$).

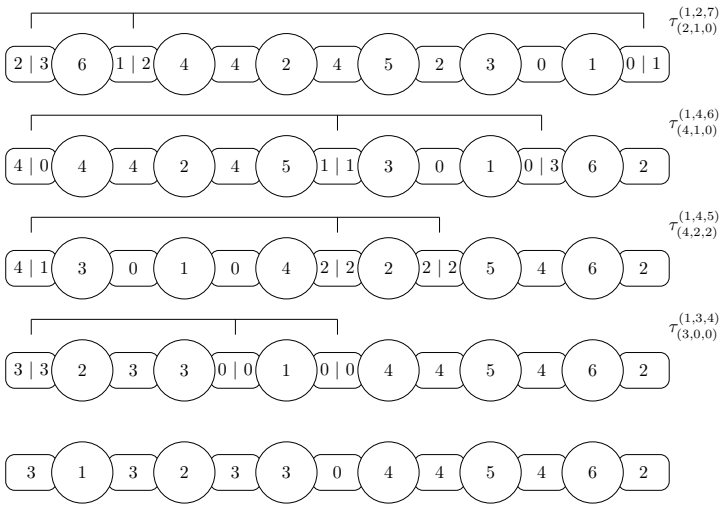



Figura 5: Sequência de transposições intergênicas $S_\tau = \langle \tau^{(1,2,7)}_{(2,1,0)}, \tau^{(1,4,6)}_{(4,1,0)}, \tau^{(1,4,5)}_{(4,2,2)}, \tau^{(1,3,4)}_{(3,0,0)} \rangle$ sendo aplicada na instância $(\pi, \breve{\pi}, \breve{\iota})$, tal que $\pi = (6\ 4\ 2\ 5\ 3\ 1)$, $\breve{\pi} = (5, 3, 4, 4, 2, 0, 1)$ e $\breve{\iota} = (3, 3, 3, 0, 4, 4, 2)$).

gura 6 mostra uma sequência de reversões e transposições intergênicas $S_{\rho\tau}$ capaz de transformar $(\pi, \breve{\pi})$ em $(\iota, \breve{\iota})$.

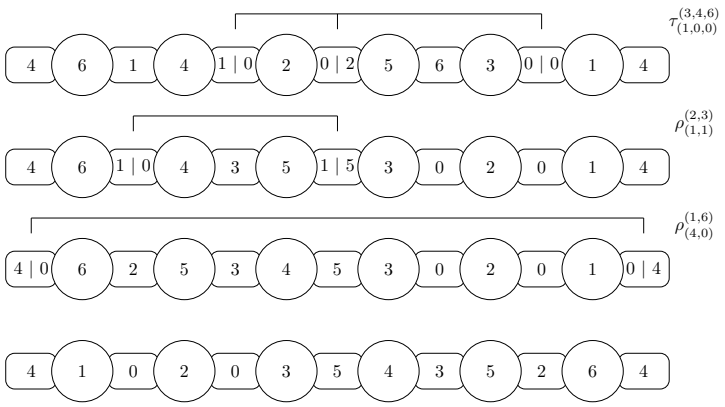



Figura 6: Sequência de reversões e transposições intergênicas $S_{\rho\tau} = \langle \tau^{(3,4,6)}_{(1,0,0)}, \rho^{(2,3)}_{(1,1)}, \rho^{(1,6)}_{(4,0)} \rangle$ sendo aplicada na instância $(\pi, \breve{\pi}, \breve{\iota})$, tal que $\pi = (6\ 4\ 2\ 5\ 3\ 1)$, $\breve{\pi} = (4, 1, 1, 2, 6, 0, 4)$ e $\breve{\iota} = (4, 0, 0, 5, 3, 2, 4)$).

## 3.2 Distância de Strings

Para esse problema, assumimos que os genomas podem apresentar genes repetidos, mas que compartilham o mesmo conjunto de genes. Dado um genoma $\mathcal{G}$ como uma sequência de $n$ genes $(g_1, \ldots, g_n)$, a sua representação é dada no formato de uma string $\beta = (\beta_1, \ldots, \beta_n)$ em que cada gene é mapeado como um caractere. Um caractere na posição $i$ de uma string $\beta$ é representado por $\beta_i$. Além disso, o tamanho de uma string $\beta$ é dado pela quantidade de caracteres que compõem a string e é denotado por $|\beta|$. Um alfabeto $\Sigma$ é um conjunto de caracteres permitidos para a representação de strings. Dizemos que duas strings $\beta'$ e $\beta''$ fazem parte de um mesmo alfabeto $\Sigma$ se todos os caracteres das strings $\beta'$ e $\beta''$ pertencem à $\Sigma$ e se todos os caracteres de $\Sigma$ fazem parte das strings $\beta'$ e $\beta''$ simultaneamente.

**Definição 3.5** *A ocorrência de um caractere $\alpha$ em uma string $\beta$, denotada por $occ(\alpha, \beta)$, representa o número de cópias do caractere $\alpha$ na string $\beta$. A maior ocorrência de um caractere em uma string $\beta$ é denotada por $occ(\beta) = \max_{\alpha \in \beta}(occ(\alpha, \beta))$.*

**Definição 3.6** *Duas strings $\beta'$ e $\beta''$ são chamadas de balanceadas se fazem parte de um*

*mesmo alfabeto* $\Sigma$ *e a ocorrência dos caracteres em ambas as strings é a mesma, ou seja,* $occ(\alpha, \beta') = occ(\alpha, \beta'')$, $\forall\ \alpha \in \Sigma$.

**Definição 3.7** *Uma reversão* $\rho(i, j)$, *onde* $1 \leq i \leq j \leq |\beta|$, *é um evento que inverte a ordem dos elementos do segmento da posição* $i$ *até* $j$ *do genoma. Quando a orientação dos genes do genoma é conhecida, os elementos do segmento afetado também tem a orientação invertida.*

$$\begin{aligned} \beta &= (+\beta_1 \ldots +\beta_{i-1}\underline{+\beta_i \ldots +\beta_j}+\beta_{j+1} \ldots +\beta_n) \\ \beta \circ \rho(i,j) &= (+\beta_1 \ldots +\beta_{i-1}\underline{-\beta_j \ldots -\beta_i}+\beta_{j+1} \ldots +\beta_n) \end{aligned}$$

O exemplo a seguir mostra uma reversão $\rho(2, 4)$ sendo aplicada em uma string com e sem sinais.

$$\begin{aligned} \beta' &= (+A\ \underline{+B\ +C\ -C}\ -B\ +A\ -D) \\ \beta' \circ \rho(2,4) &= (+A\ \underline{+C\ -C\ -B}\ -B\ +A\ -D) \end{aligned}$$

$$\begin{aligned} \beta'' &= (A\ \underline{B\ C\ C}\ B\ A\ D) \\ \beta'' \circ \rho(2,4) &= (A\ \underline{C\ C\ B}\ B\ A\ D) \end{aligned}$$

**Definição 3.8** *Uma transposição* $\tau(i, j, k)$, *onde* $1 \leq i < j < k \leq |\beta| + 1$, *é um evento que troca o segmento da posição* $i$ *até* $j - 1$ *com o segmento da posição* $j$ *até* $k - 1$ *do genoma, mas sem alterar a ordem e a orientação dos genes nos segmentos.*

$$\begin{aligned} \beta &= (+\beta_1 \ldots +\beta_{i-1}\underline{+\beta_i \ldots +\beta_{j-1}}\ \underline{+\beta_j \ldots +\beta_{k-1}}+\beta_k \ldots +\beta_n) \\ \beta \circ \tau(i,j,k) &= (+\beta_1 \ldots +\beta_{i-1}\underline{+\beta_j \ldots +\beta_{k-1}}\ \underline{+\beta_i \ldots +\beta_{j-1}}+\beta_k \ldots +\beta_n) \end{aligned}$$

O exemplo a seguir mostra uma transposição $\tau(2, 5, 7)$ sendo aplicada em uma string com e sem sinais.

$$\begin{aligned} \beta' &= (+A\ \underline{+B\ +C\ -C}\ \underline{-B\ +A}\ -D) \\ \beta' \circ \tau(2,5,7) &= (+A\ \underline{-B\ +A}\ \underline{+B\ +C\ -C}\ -D) \end{aligned}$$

$$\begin{aligned} \beta'' &= (A\ \underline{B\ C\ C}\ \underline{B\ A}\ D) \\ \beta'' \circ \tau(2,5,7) &= (A\ \underline{B\ A}\ \underline{B\ C\ C}\ D) \end{aligned}$$

Dadas duas strings balanceadas $\beta'$ e $\beta''$, o objetivo consiste em encontrar a menor sequência de eventos de rearranjo permitidos por um modelo $\mathcal{M}$ capaz de transformar $\beta'$

em $\beta''$. O tamanho da menor sequência de rearranjos de genomas capaz de realizar essa tarefa é chamado de distância de strings, denotada como $d_{\mathcal{M}}(\beta', \beta'')$.

Dado um modelo de rearranjo que permite apenas o uso de reversões, o exemplo a seguir mostra uma sequência de reversões $S_\rho$ capaz de transformar a string com sinais $\beta'$ na string com sinais $\beta''$.

$$\begin{aligned}
\beta' &= (+B\ -B\ -A\ +C\ -C\ +A) \\
\beta'' &= (-B\ -C\ -A\ -A\ +C\ -B) \\
S_\rho &= \langle \rho(2,4), \rho(1,6), \rho(1,1), \rho(1,3) \rangle
\end{aligned}$$

$$\begin{aligned}
\beta' \circ S_\rho &= (+B\ \underline{-B\ -A\ +C}\ -C\ +A) \\
&= (\underline{+B\ -C\ +A\ +B\ -C\ +A}) \\
&= (\underline{-A}\ +C\ -B\ -A\ +C\ -B) \\
&= (\underline{+A\ +C\ -B}\ -A\ +C\ -B) \\
&= (+B\ -C\ -A\ -A\ +C\ -B) = \beta''
\end{aligned}$$

Dado um modelo de rearranjo que permite apenas o uso de transposições, o exemplo a seguir mostra uma sequência de transposições $S_\tau$ capaz de transformar a string sem sinais $\beta'$ na string sem sinais $\beta''$.

$$\begin{aligned}
\beta' &= (C\ A\ B\ A\ C\ B) \\
\beta'' &= (A\ A\ B\ B\ C\ C) \\
S_\tau &= \langle \tau(1,4,6), \tau(2,3,7) \rangle
\end{aligned}$$

$$\begin{aligned}
\beta' \circ S_\tau &= (\underline{C\ A\ B}\ \underline{A\ C}\ B) \\
&= (A\ \underline{C\ C}\ \underline{A\ B\ B}) \\
&= (A\ A\ B\ B\ C\ C) = \beta''
\end{aligned}$$

Dado um modelo de rearranjo que permite o uso de reversões e transposições, o exemplo a seguir mostra uma sequência de reversões e transposições $S_{\rho\tau}$ capaz de transformar a string com sinais $\beta'$ na string com sinais $\beta''$.

$$
\begin{aligned}
\beta' &= (+B\ -B\ -A\ +C\ -C\ +A) \\
\beta'' &= (-C\ -C\ -B\ -A\ -A\ -B) \\
S_{\rho\tau} &= \langle \tau(1,5,7), \rho(2,4), \rho(2,6), \rho(3,5) \rangle
\end{aligned}
$$

$$
\begin{aligned}
\beta' \circ S_{\rho\tau} &= (\underline{+B\ -B\ -A\ +C}\ \underline{-C\ +A}) \\
&= (-C\ \underline{+A\ +B\ -B}\ -A\ +C) \\
&= (-C\ \underline{+B\ -B\ -A\ -A\ +C}) \\
&= (-C\ -C\ \underline{+A\ +A\ +B}\ -B) \\
&= (-C\ -C\ -B\ -A\ -A\ -B) = \beta''
\end{aligned}
$$

# 4 Objetivos

Nessa seção, apresentamos os objetivos que pretendemos alcançar durante a execução do projeto. Para isso, realizamos uma divisão em duas etapas, sendo que cada etapa consiste na investigação de um determinado problema.

## 4.1 Ordenação de Permutações com Regiões Intergênicas

Pretendemos realizar uma investigação de maneira gradual das versões do problema considerando os diferentes modelos de rearranjo. As versões que serão investigadas são:

- Ordenação de Permutações com Sinais por Reversões Intergênicas;
- Ordenação de Permutações sem Sinais por Reversões Intergênicas;
- Ordenação de Permutações sem Sinais por Transposições Intergênicas;
- Ordenação de Permutações com Sinais por Reversões e Transposições Intergênicas;
- Ordenação de Permutações sem Sinais por Reversões e Transposições Intergênicas.

Além disso, para cada uma das versões mencionadas, pretendemos estender a investigação considerando também os eventos de inserção e deleção intergênica.

Problemas que consideram a representação de um genoma como uma permutação tendem a ser mais apropriados para a aplicação em genomas de organismos em que é possível realizar

um pré-processamento para remover cópias dos genes repetidos sem gerar grande impacto na solução final. Além da sequência de genes, todas as versões estudadas consideram o tamanho das regiões intergênicas, o que é totalmente descartado nos modelos clássicos. Resultados anteriores argumentam que incorporar essas informações referente aos tamanhos das regiões intergênicas ao modelos pode melhorar as estimativas para a distância evolutiva entre os organismos [5]. Os eventos de deleção e inserção nas versões estudadas atuam somente sobre as regiões intergênicas. Investigações futuras podem considerar a atuação desses eventos tanto nas regiões intergênicas quanto sobre a sequência de genes.

O foco da investigação será estudar a complexidade de cada uma das versões do problema, propor algoritmos (exatos, de aproximação ou probabilísticos), realizar experimentos e comparar os resultados obtidos com os resultados conhecidos da literatura.

Como resultado preliminar, tivemos um artigo aceito no "15th International Symposium on Bioinformatics Research and Applications" (`ISBRA'2019`) [34]. Nesse artigo, mostramos que as versões de Ordenação de Permutações sem Sinais por Reversões Intergênicas e Ordenação de Permutações sem Sinais por Reversões, Inserções e Deleções Intergênicas pertencem à classe de problemas NP-Difícil e apresentamos um algoritmo com fator de aproximação 4 para ambas as versões. Além disso, apresentamos um algoritmo com fator de aproximação 6 para as versões do problema de Ordenação de Permutações sem Sinais por Reversões e Transposições Intergênicas e Ordenação de Permutações sem Sinais por Reversões, Transposições, Inserções e Deleções Intergênicas. Uma versão estendida desse trabalho apresentando novos algoritmos de aproximação e resultados experimentais foi aceita para publicação na edição especial do `Journal of Computational Biology` dedicada ao `ISBRA'2019`.

## 4.2 Distância de Strings

Para esse problema, pretendemos estudar as versões considerando strings balanceadas do problema de Distância de Strings e considerando caracteres com no máximo $k$ cópias. As versões do problema que serão investigadas são:

- Distância de Strings com Sinais por Reversões;

- Distância de Strings sem Sinais por Reversões;
- Distância de Strings sem Sinais por Transposições;
- Distância de Strings com Sinais por Reversões e Transposições;
- Distância de Strings sem Sinais por Reversões e Transposições.

As versões desse problema refletem a realidade de genomas de organismos em que a repetição de genes ocorre com mais frequência e o pré-processamento dos dados visando remover as repetições pode provocar um grande impacto negativo no resultado final. Dessa forma, uma representação utilizando strings e permitindo a repetições de genes torna-se mais adequada. As versões propostas abrangem todos os modelos considerando os eventos de reversão e transposição. Entretanto, modelos considerando os eventos de deleção e remoção podem ser estudados futuramente.

O foco dessa investigação é o desenvolvimento de melhores algoritmos para o problema, tanto de aproximação, quanto heurísticas (por exemplo, baseadas em Algoritmos Genéticos).

# 5 Plano de Trabalho

Nessa seção, apresentamos um cronograma das atividades previstas para o programa de doutorado. No âmbito do programa CAPES-COFECUB (Projeto 831/15, coordenado pelo Prof. Dr. Zanoni Dias), em julho de 2018, o aluno iniciou o doutorado sanduíche com duração de um ano no *Laboratoire des Sciences du Numérique de Nantes* (LS2N), *Université de Nantes* (França), sob a supervisão do Prof. Dr. Guillaume Fertin. Esse programa é proveniente de uma parceria entre o Brasil e a França tendo como objetivos o incetivo à produção científica, o intercâmbio acadêmico e a formação de novos pesquisadores. A seguir, listamos as atividades propostas para o doutorado e na Tabela 1 posicionamos essas etapas no cronograma.

1. Obtenção dos créditos obrigatórios em disciplinas do programa de doutorado;
2. Doutorado sanduíche no exterior;

Tabela 1: Cronograma das atividades.

| Atividades | Semestres | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 2018/1 | 2018/2 | 2019/1 | 2019/2 | 2020/1 | 2020/2 | 2021/1 | 2021/2 |
| 1 | x | | | x | | | | |
| 2 | | x | x | | | | | |
| 3 | | | | x | | | | |
| 4 | | | | x | | | | |
| 5 | | x | x | | x | x | | |
| 6 | | x | x | x | x | | | |
| 7 | | | | x | x | x | x | |
| 8 | | | | | | | x | x |
| 9 | | | | | | | | x |
| 10 | | | | | | | | x |

3. Exame de Qualificação Específico (EQE);
4. Participação no Programa de Estágio Docente (PED);
5. Revisão da literatura;
6. Investigação das variações do problema de Ordenação de Permutações com Regiões Intergênicas;
7. Investigação das variações do problema de Distância de Strings;
8. Escrita da tese;
9. Revisão da tese;
10. Defesa da tese.

Vale ressaltar que os tempos alocados em algumas atividades podem sofrer alterações no decorrer do desenvolvimento da pesquisa, uma vez que alguns resultados obtidos podem ser mais promissores que outros, fazendo com que mais tempo seja despendido em uma atividade em detrimento de outra.

# 6 Materiais e Métodos

Para as duas etapas principais desse projeto, temos o desenvolvimento de uma parte teórica seguida por uma parte prática. Para o problema de Ordenação de Permutações com Regiões

Intergênicas, pretendemos inicialmente manter uma abordagem com o foco no estudo e desenvolvimento de provas, teoremas e algoritmos para as versões consideradas do problema. Em seguida, implementaremos os algoritmos propostos e realizaremos experimentos computacionais.

Para o problema de Distância de Strings, pretendemos investigar o quanto os resultados práticos (considerando heurísticas) se aproximam dos resultados teóricos (considerando algoritmos de aproximação) conhecidos para as versões do problema. Dessa forma, pretendemos obter uma visão geral dos pontos que podem ser melhorados e, se possível, propor melhorias para os resultados teóricos.

Inicialmente, para realizarmos os experimentos computacionais, será necessária a criação de bases de dados com as representação de genomas para cada cenário das versões dos problemas investigados no âmbito desse projeto [35]. Dessa maneira, essas bases de dados serão utilizadas como entrada para os programas que serão desenvolvidos durante a parte experimental de cada etapa. As bases de dados sintéticas serão criadas para cada problema visando obter as características esperadas nos cenários estudados.

A adoção das bases de dados sintéticas nos proporcionam um processo de análise mais rápido, tendo em vista que os dados já estarão em conformidade com as restrições impostas pelos modelos e nenhum tratamento será necessário. De maneira resumida, cada base de dados será composta por uma quantidade fixa de triplas, que caracterizam um genoma de origem, um genoma alvo e a sequência de rearranjo de genomas utilizada para transformar um genoma em outro. Para a criação de cada tripla, uma sequência aleatória de eventos de rearranjo de genomas será aplicada em um genoma origem para a obtenção do genoma alvo. Dessa forma, será possível comparar os resultados fornecidos pelos algoritmos com limitantes teóricos, bem como com a sequência aleatória utilizada para a criação de cada instância. Nessa parte, os resultados serão analisados com base no fator de aproximação obtido. Para cada conjunto de dados, analisaremos o fatores de aproximação mínimo, médio e máximo obtidos, o que nos possibilita constatar o desempenho de cada algoritmo em cada conjunto de dados.

Posteriormente, pretendemos aplicar os algoritmos em uma base de dados de genomas re-

ais, por exemplo a base Cyanorak [1] [36,37], com intuito de construir uma árvore filogenética e comparar os resultados obtidos com os resultados já existentes. Nesse contexto, o grupo de genomas escolhido deve possuir obrigatoriamente um histórico evolutivo consolidado para que uma árvore filogenética possa ser usada como "ground-truth". Assim, podemos verificar quais algoritmos fornecem resultados mais próximos desse "ground-truth" e poderemos indicar quais modelos refletem melhor a evolução observada nessa base de dados. Vale ressaltar que, ao seguir essa linha de experimentos utilizando genomas reais, um pré-processamento nos dados de genomas reais deve ser realizado, tendo em vista que os mesmos devem estar em conformidade com as restrições impostas pelos modelos.

# 7 Análise dos Resultados

Em ambas as etapas, os resultados práticos obtidos serão comparados com os limitantes inferiores conhecidos na literatura ou com os limitantes inferiores que podem ser apresentados nas partes teóricas de cada etapa. Essa comparação nos permite medir a qualidade das soluções obtidas e podem apontar possíveis pontos de melhorias. Sempre que possível, serão realizadas comparações com resultados fornecidos por outros trabalhos. Além disso, esperamos que a composição dos resultados teóricos e práticos resultem em publicações de artigos em congressos e revistas internacionais.

# Referências

[1] G. Fertin, A. Labarre, I. Rusu, É. Tannier, and S. Vialette, *Combinatorics of Genome Rearrangements*. Computational Molecular Biology, London, England: The MIT Press, 2009.

[2] A. Bergeron, "A Very Elementary Presentation of the Hannenhalli-Pevzner Theory," *Discrete Applied Mathematics*, vol. 146, no. 2, pp. 134–145, 2005.

[1] http://application.sb-roscoff.fr/cyanorak

[3] V. Bafna and P. A. Pevzner, "Sorting by Transpositions," *SIAM Journal on Discrete Mathematics*, vol. 11, no. 2, pp. 224–240, 1998.

[4] M. Aigner and D. B. West, "Sorting by Insertion of Leading Elements," *Journal of Combinatorial Theory Series A*, vol. 45, no. 2, pp. 306–309, 1987.

[5] P. Biller, L. Guéguen, C. Knibbe, and E. Tannier, "Breaking Good: Accounting for Fragility of Genomic Regions in Rearrangement Distance Estimation," *Genome Biology and Evolution*, vol. 8, no. 5, pp. 1427–1439, 2016.

[6] P. Biller, C. Knibbe, G. Beslon, and E. Tannier, "Comparative Genomics on Artificial Life," in *Pursuit of the Universal*, pp. 35–44, 2016.

[7] S. Hannenhalli and P. A. Pevzner, "Transforming Cabbage into Turnip: Polynomial Algorithm for Sorting Signed Permutations by Reversals," *Journal of the ACM*, vol. 46, no. 1, pp. 1–27, 1999.

[8] E. Tannier, A. Bergeron, and M.-F. Sagot, "Advances on Sorting by Reversals," *Discrete Applied Mathematics*, vol. 155, no. 6-7, pp. 881–888, 2007.

[9] D. A. Bader, B. M. E. Moret, and M. Yan, "A Linear-Time Algorithm for Computing Inversion Distance Between Signed Permutations with an Experimental Study," *Journal of Computational Biology*, vol. 8, pp. 483–491, 2001.

[10] A. Caprara, "Sorting Permutations by Reversals and Eulerian Cycle Decompositions," *SIAM Journal on Discrete Mathematics*, vol. 12, no. 1, pp. 91–110, 1999.

[11] V. Bafna and P. A. Pevzner, "Genome Rearrangements and Sorting by Reversals," *SIAM Journal on Computing*, vol. 25, no. 2, pp. 272–289, 1996.

[12] D. A. Christie, "A 3/2-Approximation Algorithm for Sorting by Reversals," in *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'1998)*, (Philadelphia, PA, USA), pp. 244–252, 1998.

[13] P. Berman, S. Hannenhalli, and M. Karpinski, "1.375-Approximation Algorithm for Sorting by Reversals," in *Proceedings of the 10th Annual European Symposium on Algorithms (ESA'2002)*, vol. 2461 of *Lecture Notes in Computer Science*, pp. 200–210, 2002.

[14] L. Bulteau, G. Fertin, and I. Rusu, "Sorting by Transpositions is Difficult," *SIAM Journal on Computing*, vol. 26, no. 3, pp. 1148–1180, 2012.

[15] D. A. Christie, *Genome Rearrangement Problems*. PhD thesis, Department of Computing Science, University of Glasgow, 1998.

[16] I. Elias and T. Hartman, "A 1.375-Approximation Algorithm for Sorting by Transpositions," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 3, no. 4, pp. 369–379, 2006.

[17] U. Dias and Z. Dias, "Extending Bafna-Pevzner Algorithm," in *Proceedings of the 1st International Symposium on Biocomputing (ISB'2010)*, (New York, NY, USA), pp. 1–8, 2010.

[18] A. R. Oliveira, K. L. Brito, U. Dias, and Z. Dias, "On the Complexity of Sorting by Reversals and Transpositions Problems," *Journal of Computational Biology*, vol. 26, pp. 1223–1229, 2019.

[19] M. E. M. T. Walter, Z. Dias, and J. Meidanis, "Reversal and Transposition Distance of Linear Chromosomes," in *Proceedings of the 5th International Symposium on String Processing and Information Retrieval (SPIRE'1998)*, (Los Alamitos, CA, USA), pp. 96–102, 1998.

[20] A. Rahman, S. Shatabda, and M. Hasan, "An Approximation Algorithm for Sorting by Reversals and Transpositions," *Journal of Discrete Algorithms*, vol. 6, no. 3, pp. 449–457, 2008.

[21] X. Chen, "On Sorting Unsigned Permutations by Double-Cut-and-Joins," *Journal of Combinatorial Optimization*, vol. 25, no. 3, pp. 339–351, 2013.

[22] U. Dias, G. R. Galvão, C. N. Lintzmayer, and Z. Dias, "A General Heuristic for Genome Rearrangement Problems," *Journal of Bioinformatics and Computational Biology*, vol. 12, no. 3, p. 26, 2014.

[23] K. L. Brito, A. R. Oliveira, U. Dias, and Z. Dias, "Heuristics for the Sorting Signed Permutations by Reversals and Transpositions Problem," in *Algorithms for Computational Biology*, vol. 10849, pp. 65–75, 2018.

[24] D. A. Christie and R. W. Irving, "Sorting Strings by Reversals and by Transpositions," *SIAM Journal on Discrete Mathematics*, vol. 14, no. 2, pp. 193–206, 2001.

[25] M. R. Garey and D. S. Johnson, *Computers and Intractability; A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co., 1990.

[26] A. J. Radcliffe, A. D. Scott, and E. L. Wilmer, "Reversals and Transpositions Over Finite Alphabets," *SIAM Journal on Discrete Mathematics*, vol. 19, no. 1, pp. 224–244, 2005.

[27] X. Chen, J. Zheng, Z. Fu, P. Nan, Y. Zhong, S. Lonardi, and T. Jiang, "Assignment of Orthologous Genes via Genome Rearrangement," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 4, pp. 302–315, 2005.

[28] P. Kolman and T. Waleń, "Reversal Distance for Strings with Duplicates: Linear Time Approximation Using Hitting Set," in *International Workshop on Approximation and Online Algorithms (WAOA'2006)*, pp. 279–289, 2006.

[29] G. Fertin, G. Jean, and E. Tannier, "Algorithms for Computing the Double Cut and Join Distance on both Gene Order and Intergenic Sizes," *Algorithms for Molecular Biology*, vol. 12, no. 1, p. 16, 2017.

[30] S. Yancopoulos, O. Attie, and R. Friedberg, "Efficient Sorting of Genomic Permutations by Translocation, Inversion and Block Interchange," *Bioinformatics*, vol. 21, no. 16, pp. 3340–3346, 2005.

[31] L. Bulteau, G. Fertin, and E. Tannier, "Genome Rearrangements with Indels in Intergenes Restrict the Scenario Space," *BMC Bioinformatics*, vol. 17, no. 14, p. 426, 2016.

[32] A. R. Oliveira, G. Jean, G. Fertin, U. Dias, and Z. Dias, "Super Short Reversals on Both Gene Order and Intergenic Sizes," in *Advances in Bioinformatics and Computational Biology*, pp. 14–25, 2018.

[33] G. R. Galvão and Z. Dias, "An Audit Tool for Genome Rearrangement Algorithms," *Journal of Experimental Algorithmics*, vol. 19, pp. 1–34, 2014.

[34] K. L. Brito, G. Jean, G. Fertin, A. R. Oliveira, U. Dias, and Z. Dias, "Sorting by Reversals, Transpositions, and Indels on both Gene Order and Intergenic Sizes," in *Proceedings of the 15th International Symposium on Bioinformatics Research and Applications (ISBRA'2019)*, Lecture Notes in Computer Science, pp. 28–39, 2019.

[35] K. de Lima Brito, "Sorting Signed Permutations by Reversals and Transpositions," Master's thesis, Institute of Computing, University of Campinas, Brazil, 2018. In Portuguese.

[36] F. Humily, F. Partensky, C. Six, G. K. Farrant, M. Ratin, D. Marie, and L. Garczarek, "A gene island with two possible configurations is involved in chromatic acclimation in marine synechococcus," *PloS one*, vol. 8, no. 12, p. e84459, 2013.

[37] C. Six, J.-C. Thomas, L. Garczarek, M. Ostrowski, A. Dufresne, N. Blot, D. J. Scanlan, and F. Partensky, "Diversity and evolution of phycobilisomes in marine synechococcus spp.: a comparative genomics study," *Genome biology*, vol. 8, no. 12, p. R259, 2007.