Generalizações de Problemas Envolvendo Partição de Strings e Eventos de Rearranjo

Candidato: Gabriel Henriques Siqueira
Orientador: Prof. Dr. Zanoni Dias

28 de março de 2022

Instituto de Computação, Universidade Estadual de Campinas

Roteiro

Introdução

Motivação

- Rearranjos de Genomas são mutações que afetam uma grande porção do genoma.
- O número de rearranjos capazes de transformar um genoma em outro pode ser usado como uma medida da distância evolutiva entre os genomas.
- Essa distância pode ser usada em outros problemas no campo da Biologia Computacional como a construção de árvores filogenéticas e a detecção de genes ortólogos.

Eventos de Rearranjo

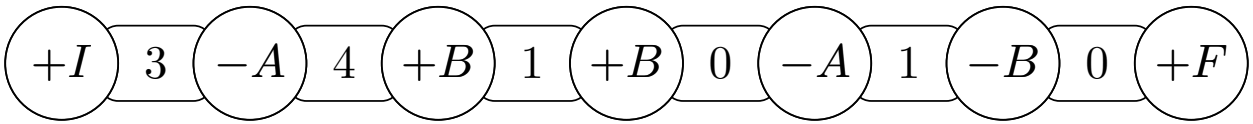
- Chamamos de modelo de rearranjo o conjunto de eventos de rearranjo sendo considerado.
- Alguns eventos de rearranjo: transposição, reversão, inserção, deleção, translocação e *double cut and join* (DCJ).

Representação do Genoma

- Depende das características dos genomas e do modelo de rearranjo.
- Vamos considerar apenas genomas lineares que possuem um único cromossomo.
- Para tornar a notação uniforme, adicionamos um gene artificial no começo e no final do genoma.
- A quantidade de nucleotídeos entre genes (tamanho das regiões intergênicas) é considerada na representação.

Representação do Genoma

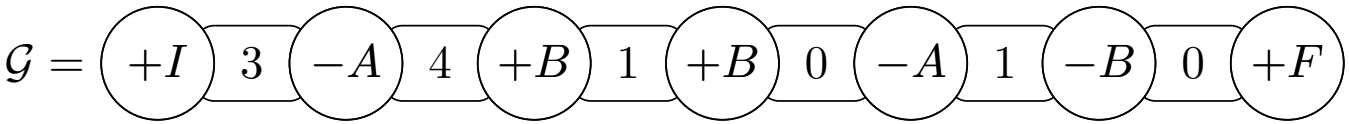
- Um genoma $\mathcal{G}$ é representado por uma string S e uma lista de inteiros $\breve{S}$.
- Cada caractere da string S representa um gene.
- A orientação dos genes é representada por um sinal $+$ ou $-$ associado a cada caractere.
- Quando omitimos os sinais, temos um genoma sem sinais. Caso contrário, temos um genoma com sinais.
- Cada inteiro de $\breve{S}$ representa o tamanho de uma região intergênica.



$$S = [+I\ -A\ +B\ +B\ -A\ -B\ +F],\ \breve{S} = [3\ 4\ 1\ 0\ 1\ 0]$$

Representação do Genoma

- $|S|$ e $|\check{S}|$ são os tamanhos de S e de $\check{S}$ ($|S| = |\check{S}| + 1$).
- S_i é o caractere na i-ésima posição de S.
- $\check{S}_i$ é a região intergênica entre S_i e S_{i+1}.
- Σ_S é o conjunto de caracteres distintos, sem considerar os sinais.
- Os elementos de Σ_S são chamados de rótulos.
- $occ(\alpha, S)$ é a ocorrência de α em S.
- $occ(S) = \max_{\alpha \in \Sigma_S}(occ(\alpha, S))$.



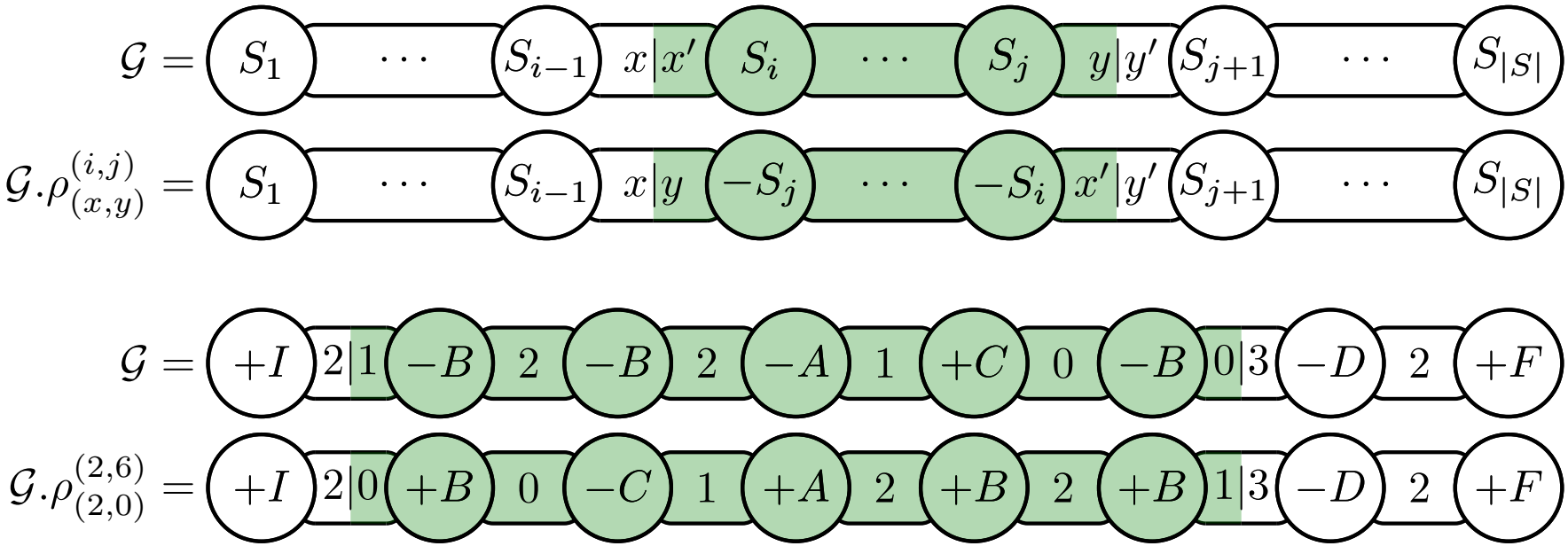
$$|S| = 7,\ |\check{S}| = 6,\ \Sigma_S = \{I, A, B, F\},\ occ(S) = occ(B, S) = 3$$

Representação do Genoma

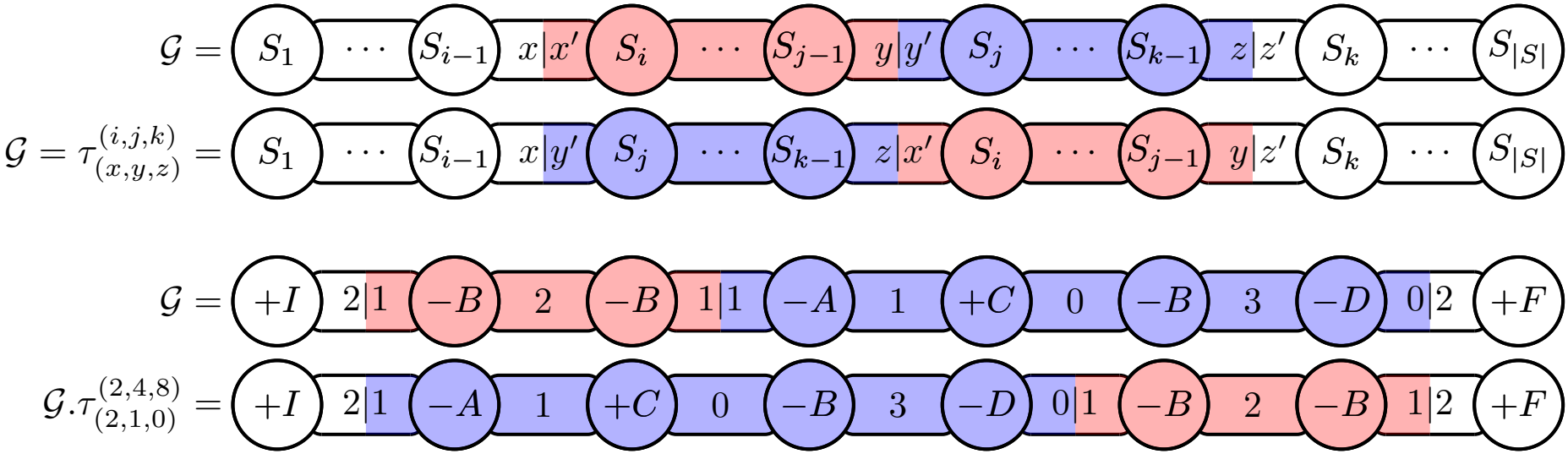
- Quando comparamos um par de genomas por eventos conservativos, esses genomas devem ser balanceados.
- Dois genomas $\mathcal{G} = (S, \breve{S})$ e $\mathcal{H} = (P, \breve{P})$ são balanceados se:
  - $\Sigma_S = \Sigma_P$;
  - $occ(\alpha, S) = occ(\alpha, P)$, para todo $\alpha \in \Sigma_S$;
  - $\sum_{i=1}^{|\breve{S}|} \breve{S}_i = \sum_{i=1}^{|\breve{P}|} \breve{P}_i$.
- Dois genomas são desbalanceados caso contrário.

Eventos de Rearranjo

Reversão


$\mathcal{G} = S_1 \cdots S_{i-1}\ x|x'\ S_i \cdots S_j\ y|y'\ S_{j+1} \cdots S_{|S|}$
$\mathcal{G}.\rho^{(i,j)}_{(x,y)} = S_1 \cdots S_{i-1}\ x|y\ -S_j \cdots -S_i\ x'|y'\ S_{j+1} \cdots S_{|S|}$
$\mathcal{G} = +I\ 2|1\ -B\ 2\ -B\ 2\ -A\ 1\ +C\ 0\ -B\ 0|3\ -D\ 2\ +F$
$\mathcal{G}.\rho^{(2,6)}_{(2,0)} = +I\ 2|0\ +B\ 0\ -C\ 1\ +A\ 2\ +B\ 2\ +B\ 1|3\ -D\ 2\ +F$

Transposição



Inserção

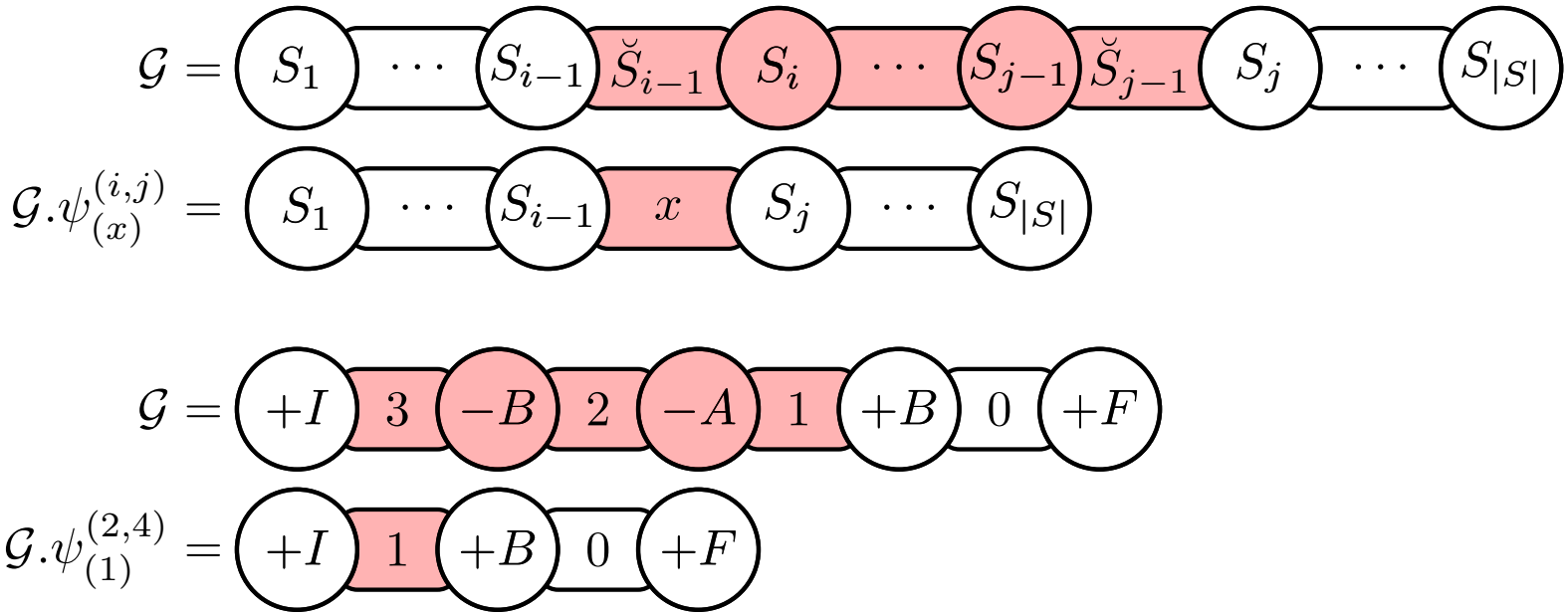
$$\mathcal{G} = \boxed{S_1} \cdots \boxed{S_i} \; x|x' \; \boxed{S_{i+1}} \cdots \boxed{S_{|S|}}$$

$$\mathcal{G}.\phi^{(i,(A,\breve{A}))}_{(x,y,z)} = \boxed{S_1} \cdots \boxed{S_i} \; x|y \; \boxed{A_1} \cdots \boxed{A_{|A|}} \; z|x' \; \boxed{S_{i+1}} \cdots \boxed{S_{|S|}}$$

$$\mathcal{G} = \boxed{+I} \; 3 \; \boxed{-B} \; 1|1 \; \boxed{-B} \; 2 \; \boxed{-A} \; 1 \; \boxed{+F}$$

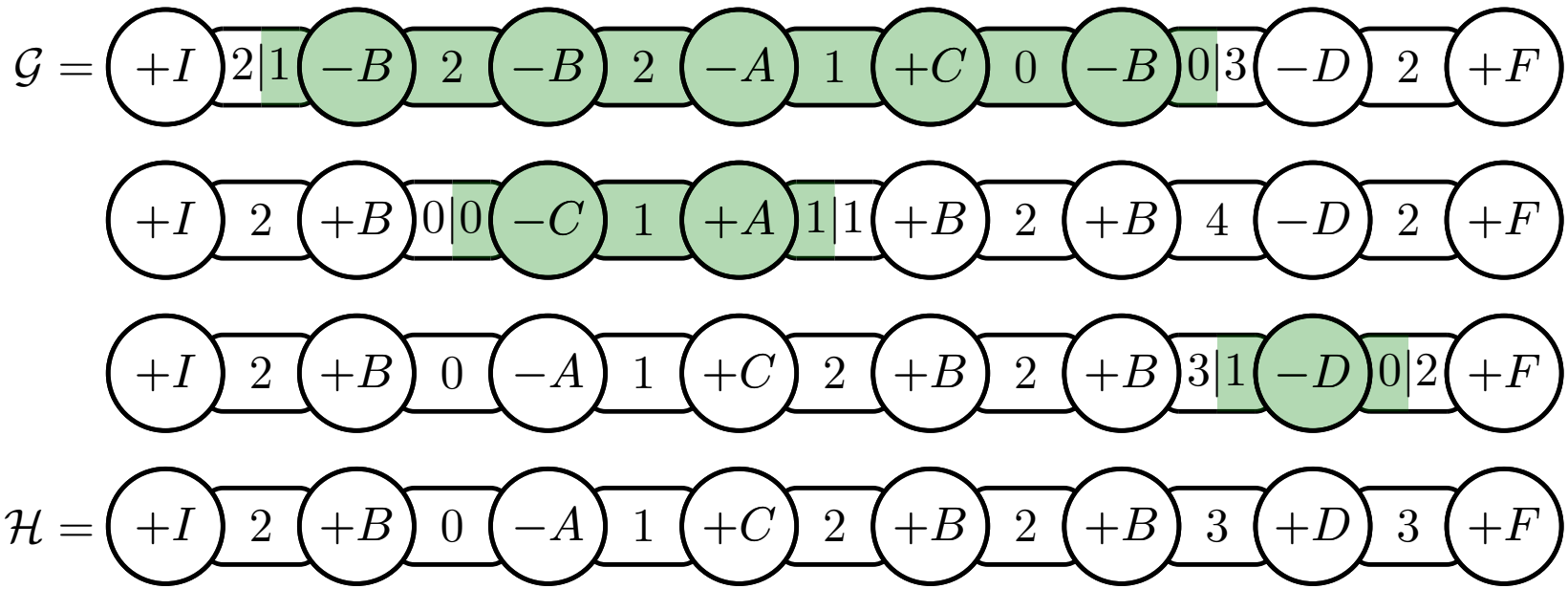
$$\mathcal{G}.\phi^{(2,((+A-E),[5]))}_{(1,0,3)} = \boxed{+I} \; 3 \; \boxed{-B} \; 1|0 \; \boxed{+A} \; 5 \; \boxed{-E} \; 3|1 \; \boxed{-B} \; 2 \; \boxed{-A} \; 1 \; \boxed{+F}$$

Deleção


$\mathcal{G} =$ S_1 $\cdots$ S_{i-1} $\breve{S}_{i-1}$ S_i $\cdots$ S_{j-1} $\breve{S}_{j-1}$ S_j $\cdots$ $S_{|S|}$
$\mathcal{G}.\psi_{(x)}^{(i,j)} =$ S_1 $\cdots$ S_{i-1} x S_j $\cdots$ $S_{|S|}$
$\mathcal{G} =$ $+I$ 3 $-B$ 2 $-A$ 1 $+B$ 0 $+F$
$\mathcal{G}.\psi_{(1)}^{(2,4)} =$ $+I$ 1 $+B$ 0 $+F$

Distância de Rearranjo

- Distância de Reversão:



$$dist(\mathcal{G}, \mathcal{H}) = 3$$

Distância de Rearranjo

- Os primeiros estudos de distância de rearranjo [Bafna e Pevzner 1996; Bafna e Pevzner 1998; Walter et al. 1998; Watterson et al. 1982] possuem as seguintes limitações:
  - Os genomas devem ser balanceados;
  - Regiões intergênicas não eram consideradas;
  - Cada gene aparecia apenas uma vez no genoma.
- Nesse cenário, a distância de reversão em genomas com sinais pode ser encontrada com um algoritmo polinomial exato [Hannenhalli e Pevzner 1999].
- Entretanto, encontrar a distância de reversão em genomas sem sinais ou as distâncias envolvendo transposição pertence à classe de problemas NP-difícil [Bulteau et al. 2012; Caprara 1999; Oliveira et al. 2019].

Distância de Rearranjo

- Novos estudos começaram a introduzir as operações de inserção e deleção (indels) para lidar com genomas desbalanceados [Braga et al. 2011; El-Mabrouk 2000].
- Nesse cenário, a distância de reversão e indels em genomas com sinais pode ser encontrada com um algoritmo polinomial exato [Willing et al. 2021].
- Entretanto, encontrar a distância de reversão e indels em genomas sem sinais ou as distâncias envolvendo transposição pertence à classe de problemas NP-difícil [Alexandrino et al. 2021b].

Distância de Rearranjo

- Outros estudos incluíram regiões intergênicas ao modelar os genomas [Brito et al. 2020; Fertin et al. 2017].
- Nesse cenário, encontrar as distâncias envolvendo reversão ou transposição em genomas com ou sem sinais pertence à classe de problemas NP-difícil [Brito et al. 2020; Oliveira et al. 2021a; Oliveira et al. 2021b].
- Estudos recentes combinam regiões intergênicas e indels [Alexandrino et al. 2021a; Alexandrino et al. 2021c].

Distância de Rearranjo

- Existem ainda estudos que consideram que genes podem ter múltiplas cópias e lidam com genomas balanceados sem considerar as regiões intergênicas [Chen et al. 2005; Rubert et al. 2017].
- Nesse cenário, encontrar as distâncias envolvendo reversão ou transposição em genomas com ou sem sinais pertence à classe de problemas NP-difícil [Christie e Irving 2001; Radcliffe et al. 2005].

Distância de Rearranjo

Tabela 1: Aproximações conhecidas para distâncias de rearranjo em genomas com sinais levando em conta a presença de repetição de genes ou regiões intergênicas (RI).

Distância de...	**Sem Repetição de Genes**		**Com Repetição de Genes**	
	Sem RI	**Com RI**	**Sem RI**	**Com RI**
Reversão	Exato	2	$16occ(S)$, **4occ(S)**	**8occ(S)**
Reversão e Transposição	2	3	$24occ(S)$, **6occ(S)**	**9occ(S)**
Reversão e Indels	Exato	3	–	–
Reversão, Transposição e Indels	3	–	–	–

Distância de Rearranjo

Tabela 2: Aproximações conhecidas para distâncias de rearranjo em genomas sem sinais levando em conta a presença de repetição de genes ou regiões intergênicas (RI).

Distância de...	**Sem Repetição de Genes**		**Com Repetição de Genes**	
	Sem RI	**Com RI**	**Sem RI**	**Com RI**
Reversão	1.375	4	$16occ(S)$, **4occ(S)**	**8occ(S)**
Transposição	1.375	3.5	$12occ(S)$, **6occ(S)**	**6occ(S)**[1]
Reversão e Transposição	$2.8334 + \epsilon$	4	$24occ(S)$, **6occ(S)**	**8occ(S)**
Reversão e Indels	2	4	–	–
Transposição e Indels	3	4.5	–	–
Reversão, Transposição e Indels	3	6	–	–

[1] aproximação assintótica.

Partição Intergênica

Reverso de um Genoma

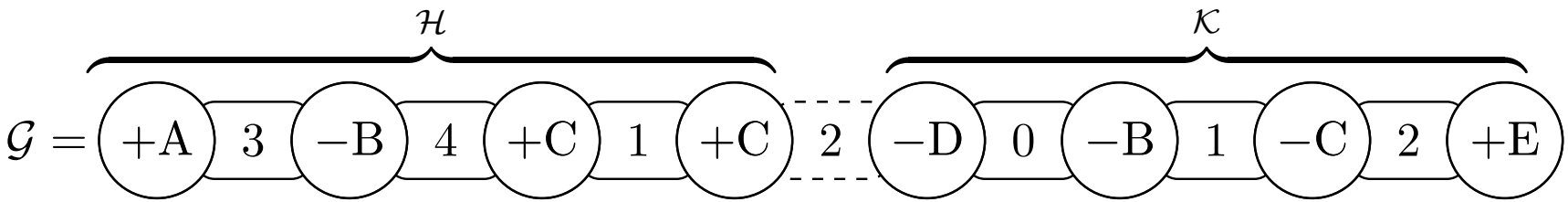
- $rev(\mathcal{G}) = (R, \breve{R})$ é o reverso de um genoma $\mathcal{G} = (S, \breve{S})$:
  - $|R| = |S|$;
  - $R_i = -S_{|S|-i+1}, \forall 1 \leq i \leq |S|$, se $\mathcal{G}$ é um genoma com sinais;
  - $R_i = S_{|S|-i+1}, \forall 1 \leq i \leq |S|$, se $\mathcal{G}$ é um genoma sem sinais;
  - $\breve{R}_i = \breve{S}_{|\breve{s}|-i+1}$.

$$\mathcal{G} = \left(+A\right)\ 1\ \left(-B\right)\ 0\ \left(-B\right)$$

$$rev(\mathcal{G}) = \left(+B\right)\ 0\ \left(+B\right)\ 1\ \left(-A\right)$$

Quebra

- Uma *quebra* de um genoma $\mathcal{G} = (S, \breve{S})$ é uma operação que separa $\mathcal{G}$ em dois genomas $\mathcal{H} = (P, \breve{P})$ e $\mathcal{K} = (Q, \breve{Q})$ a partir de uma região intergênica $\breve{S}_i$.



Partição com Sinais

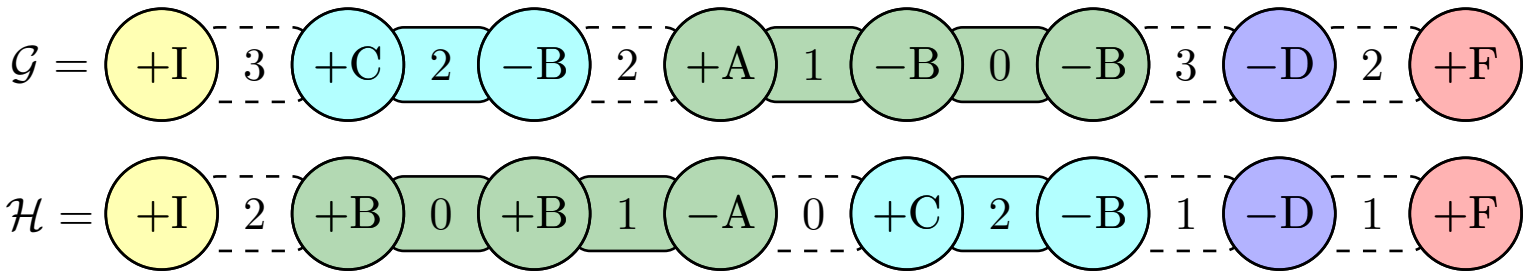
- Uma *partição com sinais comum* entre duas strings com sinais balanceadas S e P é um par de sequências de strings $(\mathbb{S}, \mathbb{P})$ tal que:
  1. As strings de $\mathbb{S}$ quando concatenadas correspondem a S.
  2. As strings de $\mathbb{P}$ quando concatenadas correspondem a P.
  3. As strings de $\mathbb{S}$ podem ser reordenadas e revertidas para obtermos as strings de $\mathbb{P}$.

$$S = \boxed{+I}\boxed{+C}\boxed{-B}\boxed{+A}\boxed{-B}\boxed{-B}\boxed{-D}\boxed{+F}$$

$$P = \boxed{+I}\boxed{+B}\boxed{+B}\boxed{-A}\boxed{+C}\boxed{-B}\boxed{-D}\boxed{+F}$$

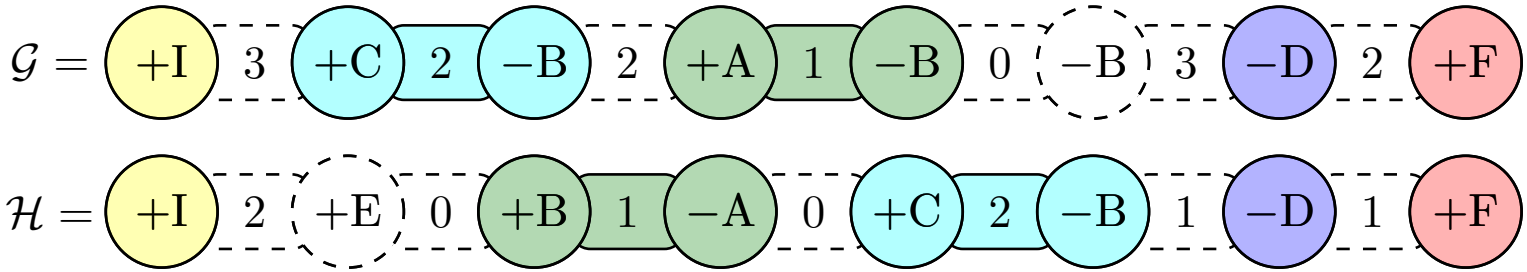
Partição Intergênica com Sinais

- Uma *partição intergênica com sinais comum* entre dois genomas com sinais balanceados $\mathcal{G} = (S, \check{S})$ e $\mathcal{H} = (P, \check{P})$ é um par de sequências de genomas com sinais $(\mathbb{S}, \mathbb{P})$ tal que:
  1. O genoma $\mathcal{G}$ pode ser quebrado nos genomas de $\mathbb{S}$.
  2. O genoma $\mathcal{H}$ pode ser quebrado nos genomas de $\mathbb{P}$.
  3. Os genomas de $\mathbb{S}$ podem ser reordenados e revertidos para obtermos os genomas de $\mathbb{P}$.



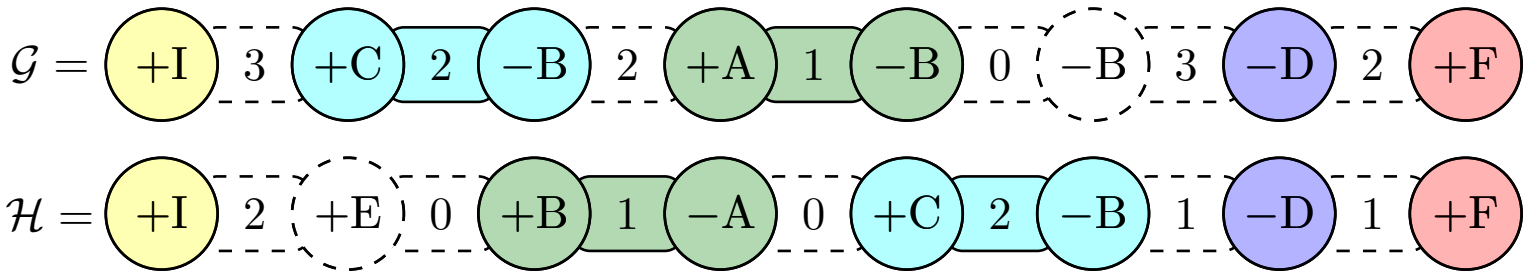
Partição Intergênica com Sinais

- Uma *partição intergênica com sinais comum* entre dois genomas com sinais $\mathcal{G} = (S, \check{S})$ e $\mathcal{H} = (P, \check{P})$ é um par de sequências de genomas com sinais $(\mathbb{S}, \mathbb{P})$ tal que:
  1. O genoma $\mathcal{G}$ pode ser quebrado nos genomas de $\mathbb{S}$.
  2. O genoma $\mathcal{H}$ pode ser quebrado nos genomas de $\mathbb{P}$.
  3. Existe uma bijeção ϕ entre uma subsequência $\mathbb{S}'$ de genomas de $\mathbb{S}$ e uma subsequência $\mathbb{P}'$ de genomas de $\mathbb{P}$, tal que: (i) Se $\phi(i) = j$, onde $\mathbb{S}_i \in \mathbb{S}'$ e $\mathbb{P}_j \in \mathbb{P}'$, então $\mathbb{S}_i = \mathbb{P}_j$ ou $rev(\mathbb{S}_i) = \mathbb{P}_j$; e (ii) Seja X o conjunto de rótulos dos caracteres de $\mathbb{S}$ que não pertencem a $\mathbb{S}'$ e Y o conjunto de rótulos dos caracteres de $\mathbb{P}$ que não pertencem a $\mathbb{P}'$, então $X \cap Y = \emptyset$.



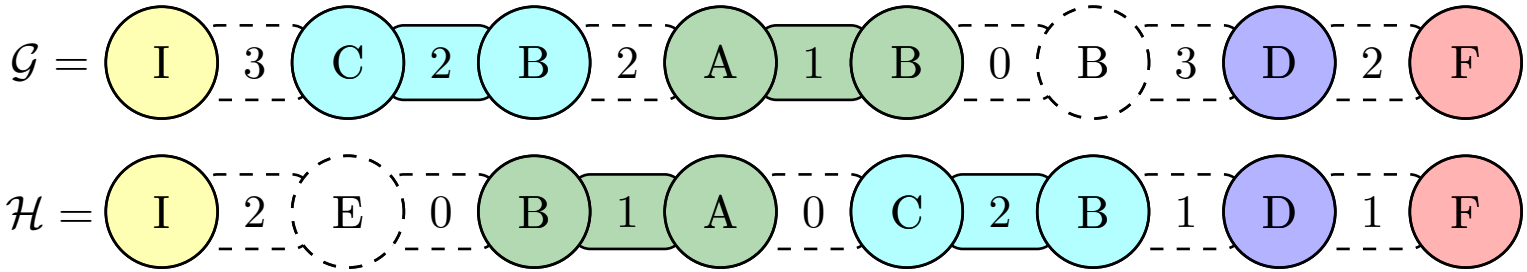
Partição Intergênica com Sinais

- Uma *partição intergênica com sinais comum* entre dois genomas com sinais $\mathcal{G} = (S, \check{S})$ e $\mathcal{H} = (P, \check{P})$ é um par de sequências de genomas com sinais $(\mathbb{S}, \mathbb{P})$ tal que:
  1. O genoma $\mathcal{G}$ pode ser quebrado nos genomas de $\mathbb{S}$.
  2. O genoma $\mathcal{H}$ pode ser quebrado nos genomas de $\mathbb{P}$.
  3. Existe uma bijeção ϕ entre uma subsequência $\mathbb{S}'$ de genomas de $\mathbb{S}$ e uma subsequência $\mathbb{P}'$ de genomas de $\mathbb{P}$, tal que: (i) Se $\phi(i) = j$, onde $\mathbb{S}_i \in \mathbb{S}'$ e $\mathbb{P}_j \in \mathbb{P}'$, então $\mathbb{S}_i = \mathbb{P}_j$ ou $rev(\mathbb{S}_i) = \mathbb{P}_j$; e (ii) Seja X o conjunto de rótulos dos caracteres de $\mathbb{S}$ que não pertencem a $\mathbb{S}'$ e Y o conjunto de rótulos dos caracteres de $\mathbb{P}$ que não pertencem a $\mathbb{P}'$, então $X \cap Y = \emptyset$.



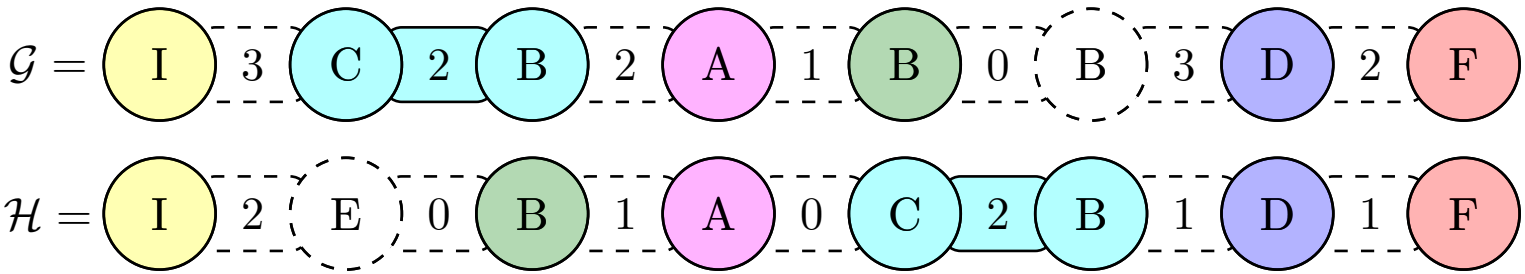
Partição Intergênica Reversa

- Uma *partição intergênica reversa comum* entre dois genomas sem sinais $\mathcal{G} = (S, \check{S})$ e $\mathcal{H} = (P, \check{P})$ é um par de sequências de genomas sem sinais $(\mathbb{S}, \mathbb{P})$ tal que:
  1. O genoma $\mathcal{G}$ pode ser quebrado nos genomas de $\mathbb{S}$.
  2. O genoma $\mathcal{H}$ pode ser quebrado nos genomas de $\mathbb{P}$.
  3. Existe uma bijeção ϕ entre uma subsequência $\mathbb{S}'$ de genomas de $\mathbb{S}$ e uma subsequência $\mathbb{P}'$ de genomas de $\mathbb{P}$, tal que: (i) Se $\phi(i) = j$, onde $\mathbb{S}_i \in \mathbb{S}'$ e $\mathbb{P}_j \in \mathbb{P}'$, então $\mathbb{S}_i = \mathbb{P}_j$ ou $rev(\mathbb{S}_i) = \mathbb{P}_j$; e (ii) Seja X o conjunto de rótulos dos caracteres de $\mathbb{S}$ que não pertencem a $\mathbb{S}'$ e Y o conjunto de rótulos dos caracteres de $\mathbb{P}$ que não pertencem a $\mathbb{P}'$, então $X \cap Y = \emptyset$.



Partição Intergênica Direta

- Uma *partição intergênica direta comum* entre dois genomas sem sinais $\mathcal{G} = (S, \check{S})$ e $\mathcal{H} = (P, \check{P})$ é um par de sequências de genomas sem sinais $(\mathbb{S}, \mathbb{P})$ tal que:
  1. O genoma $\mathcal{G}$ pode ser quebrado nos genomas de $\mathbb{S}$.
  2. O genoma $\mathcal{H}$ pode ser quebrado nos genomas de $\mathbb{P}$.
  3. Existe uma bijeção ϕ entre uma subsequência $\mathbb{S}'$ de genomas de $\mathbb{S}$ e uma subsequência $\mathbb{P}'$ de genomas de $\mathbb{P}$, tal que: (i) Se $\phi(i) = j$, onde $\mathbb{S}_i \in \mathbb{S}'$ e $\mathbb{P}_j \in \mathbb{P}'$, então $\mathbb{S}_i = \mathbb{P}_j$; e (ii) Seja X o conjunto de rótulos dos caracteres de $\mathbb{S}$ que não pertencem a $\mathbb{S}'$ e Y o conjunto de rótulos dos caracteres de $\mathbb{P}$ que não pertencem a $\mathbb{P}'$, então $X \cap Y = \emptyset$.

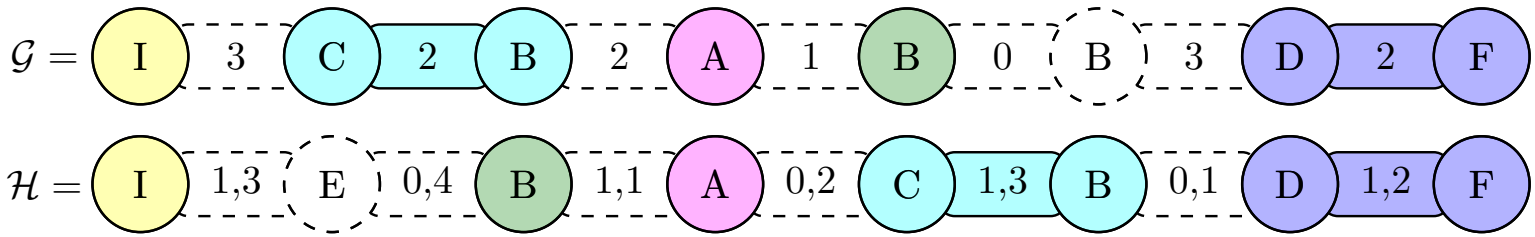


Partições Flexíveis

- Nas versões do problema de partição apresentadas até aqui, temos uma condição muito restrita ao exigirmos que as regiões intergênicas sejam iguais.
- Podemos aceitar pequenas diferenças nas regiões intergênicas.
- Um *genoma flexível* $\mathcal{H} = (P, \check{P})$ é representado por uma string P e uma lista de pares de inteiros $\check{P}$. Cada par $(\check{P}_i^i, \check{P}_i^s)$ de $\check{P}$ indica que a região intergênica correspondente a esse par tem seu tamanho no intervalo $[\check{P}_i^i, \check{P}_i^s]$.
- Um genoma $\mathcal{G} = (S, \check{S})$ é dito *compatível* com um genoma flexível $\mathcal{H} = (P, \check{P})$, denotado por $\mathcal{G} \approx \mathcal{H}$ se $S = P$ e $\check{P}_i^i \leq \check{S}_i \leq \check{P}_i^s, \forall 1 \leq i \leq |\check{S}|$.

Partição Intergênica Flexível Direta

- Uma *partição intergênica flexível direta comum* entre um genoma sem sinais $\mathcal{G} = (S, \check{S})$ e um genoma flexível sem sinais $\mathcal{H} = (P, \check{P})$ é um par $(\mathbb{S}, \mathbb{P})$, composto por uma sequência de genomas sem sinais $\mathbb{S}$ e uma sequência de genomas flexíveis sem sinais $\mathbb{P}$, tal que:
  1. O genoma $\mathcal{G}$ pode ser quebrado nos genomas de $\mathbb{S}$.
  2. O genoma $\mathcal{H}$ pode ser quebrado nos genomas de $\mathbb{P}$.
  3. Existe uma bijeção ϕ entre uma subsequência $\mathbb{S}'$ de genomas de $\mathbb{S}$ e uma subsequência $\mathbb{P}'$ de genomas de $\mathbb{P}$, tal que: (i) Se $\phi(i) = j$, onde $\mathbb{S}_i \in \mathbb{S}'$ e $\mathbb{P}_j \in \mathbb{P}'$, então $\mathbb{S}_i \approx \mathbb{P}_j$; e (ii) Seja X o conjunto de rótulos dos caracteres de $\mathbb{S}$ que não pertencem a $\mathbb{S}'$ e Y o conjunto de rótulos dos caracteres de $\mathbb{P}$ que não pertencem a $\mathbb{P}'$, então $X \cap Y = \emptyset$.



Problemas de Partição Mínima

Tabela 3: Problemas de partição mínima de strings de acordo com o tipo de partição e com a representação das regiões intergênicas (RI). Esses problemas buscam minimizar o número de quebras necessárias para formar a partição $(\mathbb{S}, \mathbb{P})$.

	Partição Direta	**Partição Reversa**	**Partição com Sinais**
Sem RI	PCMS	PRCMS	PSCMS
RI Fixas	PICMS	PIRCMS	PISCMS
RI Flexíveis	PFCMS	PFRCMS	PFSCMS

1. PCMS: Partição Comum Mínima de Strings;
2. PRCMS: Partição Reversa Comum Mínima de Strings;
3. PSCMS: Partição com Sinais Comum Mínima de Strings;
4. PICMS: Partição Intergênica Comum Mínima de Strings;
5. PIRCMS: Partição Intergênica Reversa Comum Mínima de Strings;
6. PISCMS: Partição Intergênica com Sinais Comum Mínima de Strings;
7. PFCMS: Partição Intergênica Flexível Comum Mínima de Strings;
8. PFRCMS: Partição Intergênica Flexível Reversa Comum Mínima de Strings;
9. PFSCMS: Partição Intergênica Flexível com Sinais Comum Mínima de Strings.

Relações entre Partições e Rearranjos

- Para genomas balanceados com sinais e sem regiões intergênicas, uma ℓ-aproximação para PSCMS garante uma 2ℓ-aproximação para a distância de reversão [Chen et al. 2005] e uma 3ℓ-aproximação para a distância de reversão e transposição [Siqueira 2021].
- Para genomas balanceados sem sinais e sem regiões intergênicas, uma ℓ-aproximação para PRCMS garante uma 2ℓ-aproximação para a distância de reversão e uma 3ℓ-aproximação para a distância de reversão e transposição [Siqueira 2021].
- Para genomas balanceados sem sinais e sem regiões intergênicas, uma ℓ-aproximação para PCMS garante uma 3ℓ-aproximação para a distância de transposição [Shapira e Storer 2007].

Relações entre Partições e Rearranjos

- Para genomas balanceados com sinais e com regiões intergênicas, uma ℓ-aproximação para PISCMS garante uma 4ℓ-aproximação para a distância de reversão e uma 4.5ℓ-aproximação para a distância de reversão e transposição [Siqueira et al. 2022].
- Para genomas balanceados sem sinais e com regiões intergênicas, uma ℓ-aproximação para PIRCMS garante uma 4ℓ-aproximação para a distância de reversão e uma 4ℓ-aproximação para a distância de reversão e transposição [Siqueira et al. 2021a].
- Para genomas balanceados sem sinais e com regiões intergênicas, uma ℓ-aproximação para PICMS garante uma 3ℓ-aproximação assintótica para a distância de transposição [Siqueira et al. 2021a].

Algoritmos para Partição

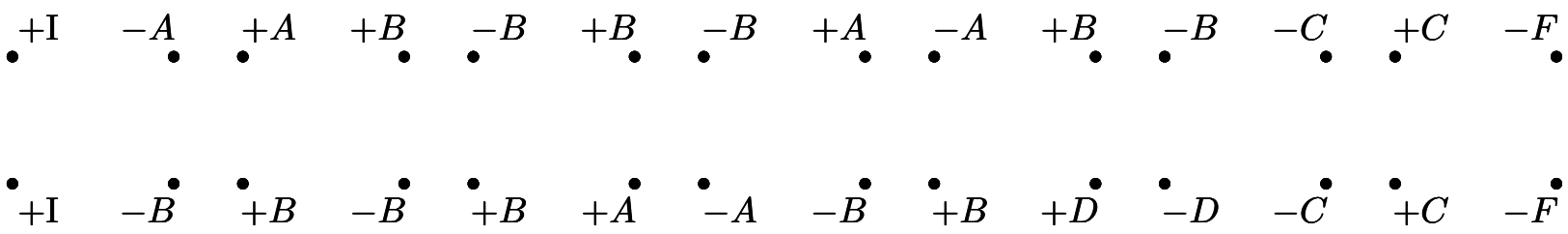
- Algoritmos da literatura:
  - Não consideram regiões intergênicas e assumem que os genomas são balanceados.
  - O melhor fator de aproximação conhecido era $4occ(S)$ para PCMS e $8occ(S)$ para PRCMS e PSCMS [Kolman e Waleń 2007].
- Em genomas balanceados, o algoritmo que propomos [Siqueira et al. 2021a] garante uma aproximação com fator $2occ(S)$ para os problemas sem regiões intergênicas e com regiões intergênicas fixas.
- Adaptamos esse algoritmo, como uma heurística, para lidar com genomas não balanceados.

Empacotamento Máximo de Ciclos

Grafo de Adjacências

$$S = (+I \quad +A \quad -B \quad -B \quad -A \quad -B \quad +C \quad +F)$$

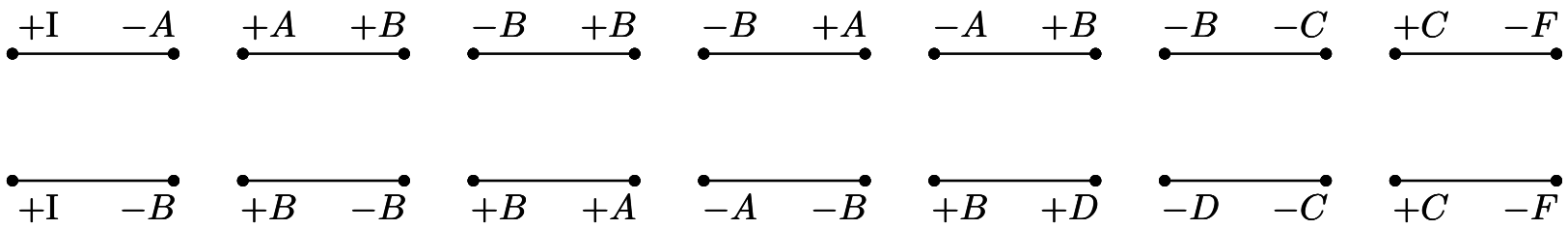
$$P = (+I \quad +B \quad +B \quad -A \quad +B \quad -D \quad +C \quad +F)$$



Grafo de Adjacências

$$S = (+I \quad +A \quad -B \quad -B \quad -A \quad -B \quad +C \quad +F)$$

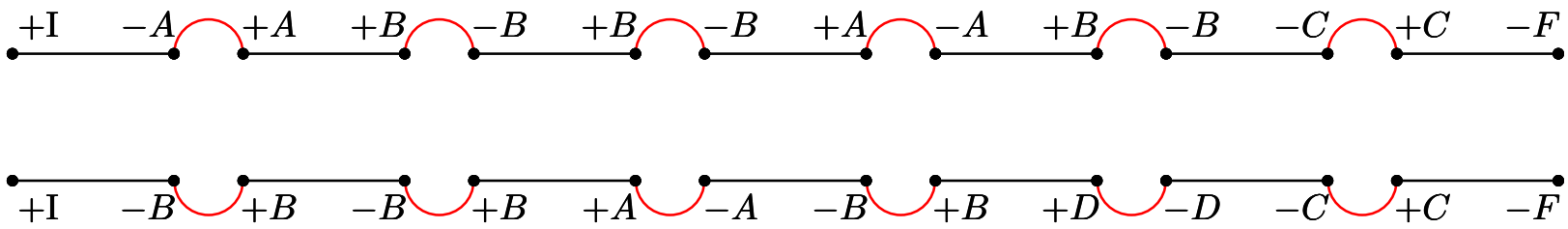
$$P = (+I \quad +B \quad +B \quad -A \quad +B \quad -D \quad +C \quad +F)$$



Grafo de Adjacências

$$S = (+I \quad +A \quad -B \quad -B \quad -A \quad -B \quad +C \quad +F)$$

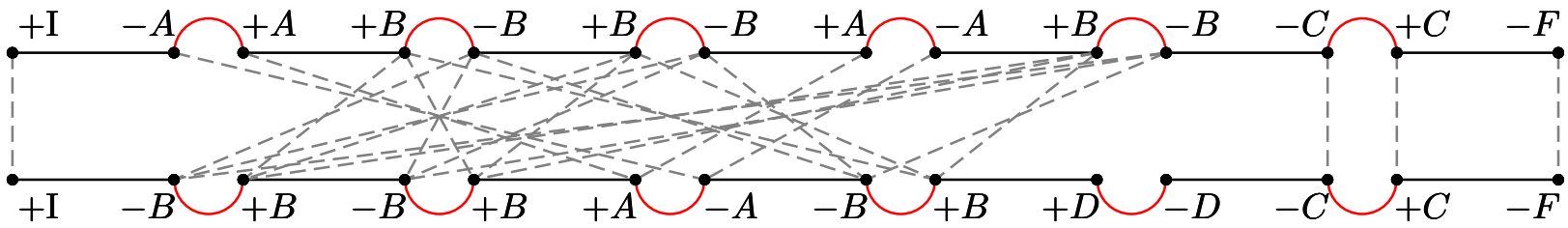
$$P = (+I \quad +B \quad +B \quad -A \quad +B \quad -D \quad +C \quad +F)$$



Grafo de Adjacências

$$S = (+I \quad +A \quad -B \quad -B \quad -A \quad -B \quad +C \quad +F)$$

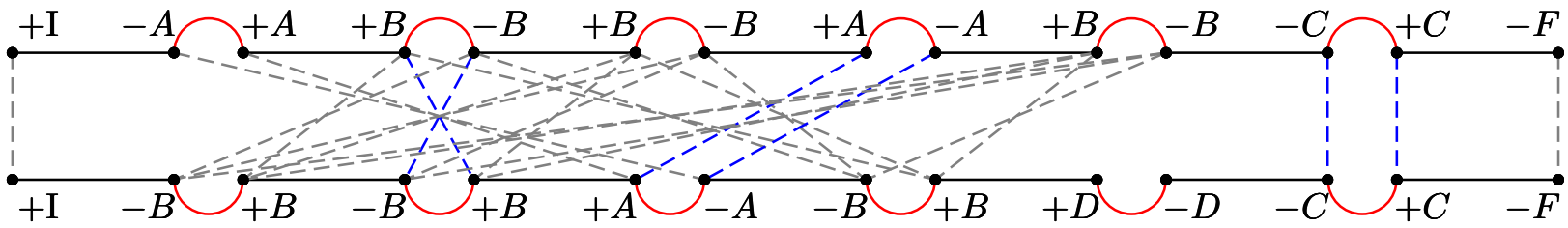
$$P = (+I \quad +B \quad +B \quad -A \quad +B \quad -D \quad +C \quad +F)$$



Grafo de Adjacências

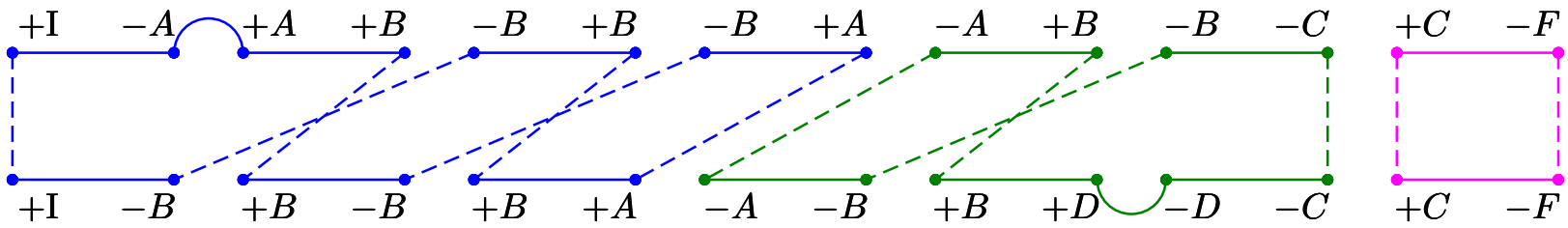
$$S = (+I \quad +A \quad -B \quad -B \quad -A \quad -B \quad +C \quad +F)$$

$$P = (+I \quad +B \quad +B \quad -A \quad +B \quad -D \quad +C \quad +F)$$



Grafo de Adjacências

- Um *empacotamento de ciclos alternados* é um conjunto de ciclos alternados disjuntos, tais que:
  - Cada vértice pertence a exatamente um ciclo.
  - Uma aresta cinza pertence a um ciclo see ela não tem uma aresta gêmea ou sua aresta gêmea também pertence a algum ciclo do empacotamento.
  - A quantidade de arestas vermelhas que conectam dois vértices $+\alpha$ e $-\alpha$ em algum ciclo é igual a $max(0, occ(\alpha, S) - occ(\alpha, P))$, se a aresta conecta vértices advindos de S, ou a $max(0, occ(\alpha, P) - occ(\alpha, S))$, se a aresta conecta vértices advindos de P.

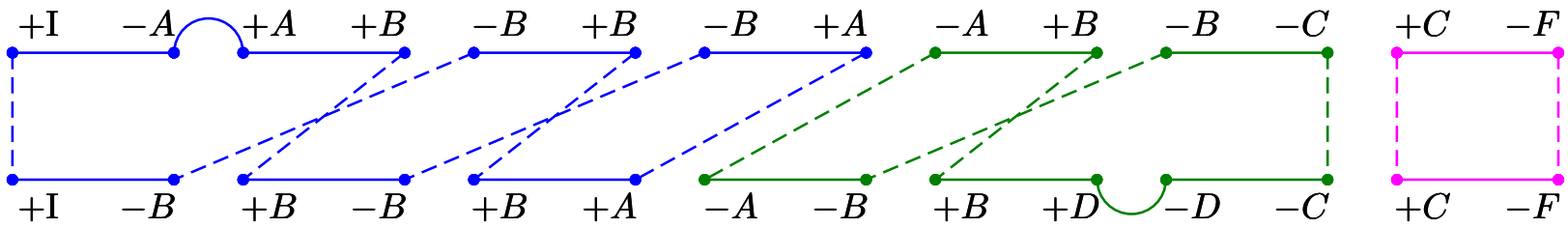


Grafo de Adjacências

- Um *empacotamento de ciclos alternados* codifica uma correspondência entre os caracteres, o que permite transformar as strings em permutações (strings sem repetição de caracteres).

$$S = (+I \ +A \ -B \ -B \ -A \ -B \ +C \ +F) \qquad \pi = (+I \ +A' \ -B' \ -B'' \ -A'' \ -B''' \ +C \ +F)$$

$$P = (+I \ +B \ +B \ -A \ +B \ -D \ +C \ +F) \qquad \sigma = (+I \ +B' \ +B'' \ -A'' \ +B''' \ -D \ +C \ +F)$$



Experimentos Práticos

Bases de Dados

- Criamos bases para cada modelo de rearranjo considerando genomas com e sem sinais.
- Apresentaremos aqui as bases para os genomas com sinais e para o modelo considerando apenas o evento de reversão.
- As bases foram geradas a partir dos parâmetros $O \in \{25, 50, 75, 100\}$, e $L \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$. Para cada par desses parâmetros, construímos duas base contendo 1000 genomas cada.

Bases de Dados

- R_O^L: Para a construção dos genomas dessas bases, geramos uma string S escolhendo 100 caracteres de um alfabeto de tamanho L. Em seguida, transformamos S em uma nova string P aplicando O operações de reversão, seguidas de $\lfloor \frac{O}{2} \rfloor$ operações de deleção de um único caractere, seguidas de $\lfloor \frac{O}{2} \rfloor$ operações de inserção de um único caractere.
- $\check{R}_O^L$: Essas bases são construídas por um processo similar ao usado nas bases R_O^L, mas além da string S foi gerada também um lista de inteiros $\check{S}$. Para a lista $\check{S}$ foram escolhidos aleatoriamente, com uma distribuição uniforme, 101 inteiros no intervalo $[0, 100]$. Além disso, as operações aplicadas em P envolvem também a lista $\check{S}$, o que produz uma lista $\check{P}$.

Experimentos Práticos

- Para cada par de genomas, geramos 100 correspondências entre genes de ambos os genomas.
- Aplicamos o algoritmo de partição para cada par de genomas.
- Para cada novo par, geramos novamente 100 correspondências entre genes de ambos os genomas.
- Com as correspondências, computamos as distâncias de reversão usando um algoritmo para o caso em que não há mais de uma cópia de cada gene [Alexandrino et al. 2021b].
- Para as bases R_O^L, computamos também as distâncias usando duas heurísticas para o problema de Empacotamento Máximo de Ciclos (a heurística de Algoritmo Genético (AG) e a heurística de Empacotamentos Aleatórios (EA)).

Resultados

Tabela 4: Cada linha corresponde a um conjunto de bases (unindo todos os valores de L) e apresenta a média das distâncias de reversão (considerando o mínimo e a média de todas as correspondências) e a média dos tempos de execução (para calcular as distâncias ou as partições).

Base	Sem Partição			Com Partição			
	Distância		Tempo (s)	Distância		Tempo (s)	
	Min.	Med.	Distância	Min.	Med.	Partição	Distância
R^L_{25}	88.19	95.48	19.04	65.11	70.36	0.90	11.92
R^L_{50}	89.00	94.39	18.25	81.18	85.80	0.96	16.40
R^L_{75}	84.41	88.77	16.94	80.92	84.91	1.04	16.05
R^L_{100}	74.81	78.44	14.50	73.05	76.48	1.13	13.91
$\bar{R}^L_{25}$	102.35	115.13	21.89	75.47	84.94	0.89	14.01
$\bar{R}^L_{50}$	101.32	111.26	20.62	93.84	102.78	0.99	17.40
$\bar{R}^L_{75}$	93.50	101.53	18.71	91.28	98.92	1.06	16.91
$\bar{R}^L_{100}$	80.68	86.99	15.93	80.06	86.22	1.15	15.20

Resultados

Tabela 5: Médias das distâncias de reversão e dos tempos de execução considerando o mínimo das correspondências (MAP) ou as heurísticas de empacotamento (AG e EA). A primeira parte da tabela apresenta os resultados sem o algoritmo de partição e a segunda com o algoritmo de partição.

Base	Distância			Tempo (s)			
	MAP	AG	EA	Partição	MAP	AG	EA
R^L_{25}	88.19	80.62	83.59	–	19.04	0.18	0.12
R^L_{50}	89.00	83.33	84.42	–	18.25	0.16	0.11
R^L_{75}	84.41	71.89	72.57	–	16.94	0.12	0.09
R^L_{100}	74.81	52.68	53.34	–	14.50	0.07	0.07
R^L_{25}	65.11	66.02	66.34	0.90	11.92	0.11	0.08
R^L_{50}	81.18	76.76	77.38	0.96	16.40	0.13	0.08
R^L_{75}	80.92	68.11	68.70	1.04	16.05	0.11	0.09
R^L_{100}	73.05	50.51	51.11	1.13	13.91	0.06	0.06

Planejamento

Objetivos

- Encontrar novas soluções para as diferentes variações dos problemas de partição.
- Estabelecer novas relações entre esses problemas e os problemas de rearranjo de genomas.
- Buscar por novos algoritmos para os problemas de rearranjo utilizando mapeamentos das strings em permutações e empacotamento máximo de ciclos.

Metodologia

- Para cada um dos problemas, iniciaremos com um estudo teórico, desenvolvendo heurísticas, algoritmos exatos ou algoritmos que garantam um fator de aproximação.
- Implementaremos os algoritmos descritos e realizaremos experimentos práticos para verificar o desempenho obtido.
- Realizaremos testes primariamente em bases de dados construídas artificialmente (quando possível, utilizaremos bases já presentes na literatura).
- Pretendemos realizar testes com genomas reais para validar o uso dos algoritmos desenvolvidos.
- Realizaremos, quando possível, uma comparação com outros algoritmos existentes na literatura.

Cronograma das Atividades

	Semestres							
	2021/1	2021/2	2022/1	2022/2	2023/1	2023/2	2024/1	2024/2
1	X		X		X		X	
2		X	X					
3		X						
4			X					
5					X	X		
6	X	X						
7		X	X	X				
8			X	X	X			
9					X	X	X	
10							X	X
11								X

1. Revisão da literatura;
2. Participação no Programa de Estágio Docente (PED);
3. Escrita da proposta de doutorado;
4. Exame de Qualificação Específico (EQE);

Cronograma das Atividades

	Semestres							
	2021/1	2021/2	2022/1	2022/2	2023/1	2023/2	2024/1	2024/2
1	X		X		X		X	
2		X	X					
3		X						
4			X					
5					X	X		
6	X	X						
7		X	X	X				
8			X	X	X			
9					X	X	X	
10							X	X
11								X

5. Estágio de Pesquisa no Exterior;
6. Investigação dos problemas de partição e distância de rearranjo com e sem regiões intergênicas em genomas balanceados;

Cronograma das Atividades

	Semestres							
	2021/1	2021/2	2022/1	2022/2	2023/1	2023/2	2024/1	2024/2
1	X		X		X		X	
2		X	X					
3		X						
4			X					
5					X	X		
6	X	X						
7		X	X	X				
8			X	X	X			
9					X	X	X	
10							X	X
11								X

7. Investigação dos problemas de partição e distância de rearranjo com e sem regiões intergênicas em genomas não balanceados;
8. Investigação dos problemas de partição e distância de rearranjo com regiões intergênicas flexíveis em genomas balanceados;

Cronograma das Atividades

	Semestres							
	2021/1	2021/2	2022/1	2022/2	2023/1	2023/2	2024/1	2024/2
1	X		X		X		X	
2		X	X					
3		X						
4			X					
5					X	X		
6	X	X						
7		X	X	X				
8			X	X	X			
9					X	X	X	
10							X	X
11								X

9. Investigação dos problemas de partição e distância de rearranjo com regiões intergênicas flexíveis em genomas não balanceados;
10. Escrita e revisão da tese;
11. Defesa da tese.

Publicações

Publicações

- "*Heuristics for Genome Rearrangement Distance with Replicated Genes*" (Gabriel Siqueira, Klairton Lima Brito, Ulisses Dias e Zanoni Dias), publicado na revista *IEEE/ACM Transactions on Computational Biology and Bioinformatics (2021)*.
- "*Heuristics for Cycle Packing of Adjacency Graphs for Genomes with Repeated Genes*" (Gabriel Siqueira, André Rodrigues Oliveira, Alexsandro Oliveira Alexandrino e Zanoni Dias), apresentado na conferência *Brazilian Symposium on Bioinformatics (BSB'2021)*. Esse trabalho recebeu o prêmio de melhor artigo da conferência.

Publicações

- "*Approximation Algorithm for Rearrangement Distances Considering Repeated Genes and Intergenic Regions*" (Gabriel Siqueira, Alexsandro Oliveira Alexandrino, André Rodrigues Oliveira e Zanoni Dias), publicado na revista *Algorithms for Molecular Biology (2021)*.
- "*Signed Rearrangement Distances Considering Repeated Genes and Intergenic Regions*" (Gabriel Siqueira, Alexsandro Oliveira Alexandrino e Zanoni Dias), apresentado na conferência *14th International Conference on Bioinformatics and Computational Biology (BICOB'2022)*.

Generalizações de Problemas Envolvendo Partição de Strings e Eventos de Rearranjo

Candidato: Gabriel Henriques Siqueira
Orientador: Prof. Dr. Zanoni Dias

28 de março de 2022

Instituto de Computação, Universidade Estadual de Campinas

Referências

Alexandrino, Alexsandro Oliveira, Klairton Lima Brito, Andre Rodrigues Oliveira, Ulisses Dias e Zanoni Dias (2021a). "Reversal Distance on Genomes with Different Gene Content and Intergenic Regions Information". Em: *Proceedings of the 8th International Conference on Algorithms for Computational Biology (AlCoB'2021)*. Cham: Springer International Publishing, pp. 121–133.

Alexandrino, Alexsandro Oliveira, Andre Rodrigues Oliveira, Ulisses Dias e Zanoni Dias (2021b). "Genome Rearrangement Distance with Reversals, Transpositions, and Indels". Em: *Journal of Computational Biology* 28.3, pp. 235–247.

Referências

- Alexandrino, Alexsandro Oliveira, Andre Rodrigues Oliveira, Ulisses Dias e Zanoni Dias (2021c). "Incorporating intergenic regions into reversal and transposition distances with indels". Em: *Journal of Bioinformatics and Computational Biology* 19.06, p. 2140011.
- Bafna, Vineet e Pavel A. Pevzner (1996). "Genome Rearrangements and Sorting by Reversals". Em: *SIAM Journal on Computing* 25.2, pp. 272–289.
- Bafna, Vineet e Pavel A. Pevzner (1998). "Sorting by Transpositions". Em: *SIAM Journal on Discrete Mathematics* 11.2, pp. 224–240.
- Braga, Marília D.V., Eyla Willing e Jens Stoye (2011). "Double Cut and Join with Insertions and Deletions". Em: *Journal of Computational Biology* 18.9, pp. 1167–1184.

Referências


Brito, Klairton Lima, Géraldine Jean, Guillaume Fertin, Andre Rodrigues Oliveira, Ulisses Dias e Zanoni Dias (2020). "Sorting by Genome Rearrangements on Both Gene Order and Intergenic Sizes". Em: *Journal of Computational Biology* 27.2, pp. 156–174.


Bulteau, Laurent, Guillaume Fertin e Irena Rusu (2012). "Sorting by Transpositions Is Difficult". Em: *SIAM Journal on Discrete Mathematics* 26.3, pp. 1148–1180.


Caprara, Alberto (1999). "Sorting Permutations by Reversals and Eulerian Cycle Decompositions". Em: *SIAM Journal on Discrete Mathematics* 12.1, pp. 91–110.

Referências

- Chen, Xin, Jie Zheng, Zheng Fu, Peng Nan, Yang Zhong, Stefano Lonardi e Tao Jiang (2005). "Assignment of Orthologous Genes via Genome Rearrangement". Em: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2.4, pp. 302–315.
- Christie, David A. e Robert W. Irving (2001). "Sorting Strings by Reversals and by Transpositions". Em: *SIAM Journal on Discrete Mathematics* 14.2, pp. 193–206.
- Fertin, Guillaume, Géraldine Jean e Eric Tannier (2017). "Algorithms for computing the double cut and join distance on both gene order and intergenic sizes". Em: *Algorithms for Molecular Biology* 12.1, p. 16.
- Hannenhalli, Sridhar e Pavel A. Pevzner (1999). "Transforming cabbage into turnip. polynomial algorithm for sorting signed permutations by reversals". Em: *Journal of ACM* 46.1, pp. 1–27.

Referências

- Kolman, Petr e Tomasz Waleń (2007). "Reversal Distance for Strings with Duplicates: Linear Time Approximation Using Hitting Set". Em: *Proceedings of the 4th International Workshop on Approximation and Online Algorithms (WAOA'2006)*. Vol. 4368. Lecture Notes in Computer Science. Berlin, Heidelberg, pp. 279–289.
- El-Mabrouk, Nadia (2000). "Genome Rearrangement by Reversals and Insertions/Deletions of Contiguous Segments". Em: *Combinatorial Pattern Matching*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 222–234.
- Oliveira, Andre Rodrigues, Klairton Lima Brito, Ulisses Dias e Zanoni Dias (2019). "On the Complexity of Sorting by Reversals and Transpositions Problems". Em: *Journal of Computational Biology* 26.11, pp. 1223–1229.

Referências

- Oliveira, Andre Rodrigues, Geraldine Jean, Guillaume Fertin, Klairton Lima Brito, Laurent Bulteau, Ulisses Dias e Zanoni Dias (2021a). "Sorting Signed Permutations by Intergenic Reversals". Em: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 18.6, pp. 2870–2876.
- Oliveira, Andre Rodrigues, Geraldine Jean, Guillaume Fertin, Klairton Lima Brito, Ulisses Dias e Zanoni Dias (2021b). "Sorting Permutations by Intergenic Operations". Em: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 18.6, pp. 2080–2093.
- Radcliffe, Andrew J., Alex D. Scott e Elizabeth Wilmer (2005). "Reversals and Transpositions Over Finite Alphabets". Em: *SIAM Journal on Discrete Mathematics* 19.1, pp. 224–244.

Referências

- Rubert, Diego P., Pedro Feijão, Marília Dias Vieira Braga, Jens Stoye e Fábio Henrique Viduani Martinez (2017). "Approximating the DCJ distance of balanced genomes in linear time". Em: *Algorithms for Molecular Biology* 12.1, p. 3.
- Shapira, Dana e James A. Storer (2007). "Edit distance with move operations". Em: *Journal of Discrete Algorithms* 5.2, pp. 380–392.
- Siqueira, Gabriel (2021). "Heurísticas para Problemas de Rearranjo de Genomas com Genes Multiplicados". Dissertação de mestrado. Instituto de Computação, Universidade Estadual de Campinas.

Referências


Siqueira, Gabriel, Alexsandro Oliveira Alexandrino e Zanoni Dias (2022). "Signed Rearrangement Distances Considering Repeated Genes and Intergenic Regions". Em: *Proceedings of 14th International Conference on Bioinformatics and Computational Biology (BICoB'2022)*. Vol. 83. EPiC Series in Computing. EasyChair, pp. 31–42.


Siqueira, Gabriel, Alexsandro Oliveira Alexandrino, Andre Rodrigues Oliveira e Zanoni Dias (2021a). "Approximation algorithm for rearrangement distances considering repeated genes and intergenic regions". Em: *Algorithms for Molecular Biology* 16.1, p. 21.


Siqueira, Gabriel, Klairton Lima Brito, Ulisses Dias e Zanoni Dias (2021b). "Heuristics for Genome Rearrangement Distance with Replicated Genes". Em: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 18.6, pp. 2094–2108.

Referências


Siqueira, Gabriel, Andre Rodrigues Oliveira, Alexsandro Oliveira Alexandrino e Zanoni Dias (2021c). "Heuristics for Cycle Packing of Adjacency Graphs for Genomes with Repeated Genes". Em: *Proceedings of the 14th Brazilian Symposium on Bioinformatics (BSB'2021)*. Cham: Springer International Publishing, pp. 93–105.


Walter, Maria Emília M.T., Zanoni Dias e João Meidanis (1998). "Reversal and transposition distance of linear chromosomes". Em: *Proceedings of the String Processing and Information Retrieval: A South American Symposium (SPIRE'1998)*. Los Alamitos, CA, USA, pp. 96–102.


Watterson, Geoffrey A., Warren J. Ewens, Thomas E. Hall e Alexander Morgan (1982). "The chromosome inversion problem". Em: *Journal of Theoretical Biology* 99.1, pp. 1–7.

Referências

 Willing, Eyla, Jens Stoye e Marilia Braga (2021). "Computing the Inversion-Indel Distance". Em: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 18.06, pp. 2314–2326.