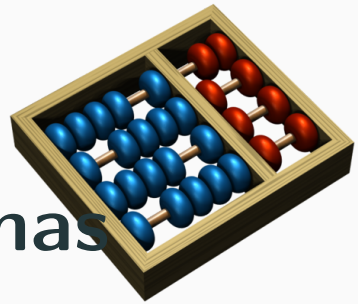


Problemas de Rearranjos de Genomas Considerando Regiões Intergênicas ou Genes Repetidos

Exame de Qualificação Específico de Doutorado

Aluno: Klairton de Lima Brito
Orientador: Prof. Dr. Zanoni Dias
Coorientador: Prof. Dr. Ulisses Martins Dias

9 de Dezembro de 2019

Instituto de Computação - Universidade Estadual de Campinas

Roteiro

1. Motivação
2. Conceitos
3. Problemas
4. Objetivos
5. Cronograma
6. Resultados Preliminares

Motivação

Motivação

Biologia Computacional

- Genômica comparativa
- Evolução genética
- Reconstrução filogenética
- Rearranjo de genomas
- Representações computacionais de um genoma:
  - Permutações
  - Strings

Motivação

Informações Incorporadas aos Modelos

- Ordem e orientação dos genes
- Demais estruturas genéticas tendem a ser descartadas

Investigações Recentes [1, 2]

- Importância de incorporar aos modelos informações além da ordem e orientação dos genes
- Regiões intergênicas
- Melhoria na estimativa da distância evolutiva entre organismos

Motivação

Incorporando Mais Informação aos Modelos

- Foco na investigação de versões dos problemas:
  - Ordenação de Permutações com Regiões Intergênicas
  - Distância de Strings

Conceitos

Ordenação de Permutações com Regiões Intergênicas

Ordenação de Permutações com Regiões Intergênicas

Representação do Genoma

- $(\pi, \breve{\pi}) = \{\breve{\pi}_1, \pi_1, \breve{\pi}_2, \pi_2, \ldots, \breve{\pi}_n, \pi_n, \breve{\pi}_{n+1}\} =$
  $\begin{cases} (\pi_1\ \pi_2\ \ldots\ \pi_n) \\ [\breve{\pi}_1, \breve{\pi}_2, \ldots, \breve{\pi}_n, \breve{\pi}_{n+1}] \end{cases}$
- $\breve{\pi}_i \geq 0$

Caso Com Sinais

- $\pi_i \in \{-n,\ -(n-1),\ \ldots,\ -1,\ +1,\ \ldots,\ +(n-1),\ +n\}$
- $|\pi_i| = |\pi_j| \leftrightarrow i = j$

Caso Sem Sinais

- $\pi_i \in \{1,\ \ldots,\ n-1,\ n\}$
- $\pi_i = \pi_j \leftrightarrow i = j$

Ordenação de Permutações com Regiões Intergênicas

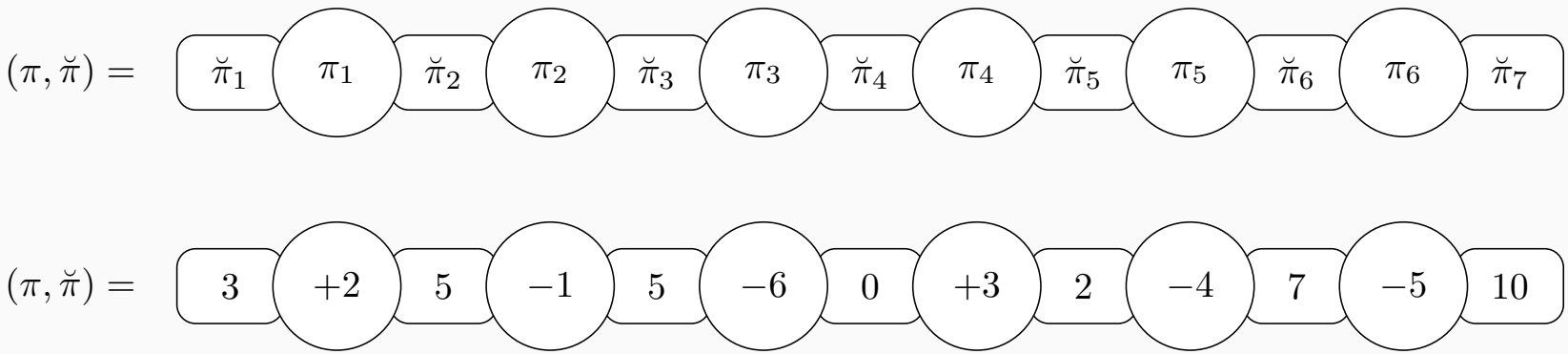


Figura 1: Exemplo da representação de um genoma com $\pi = (+2\ -1\ -6\ +3\ -4\ -5)$ e $\breve{\pi} = (3, 5, 5, 0, 2, 7, 10)$.

Ordenação de Permutações com Regiões Intergênicas

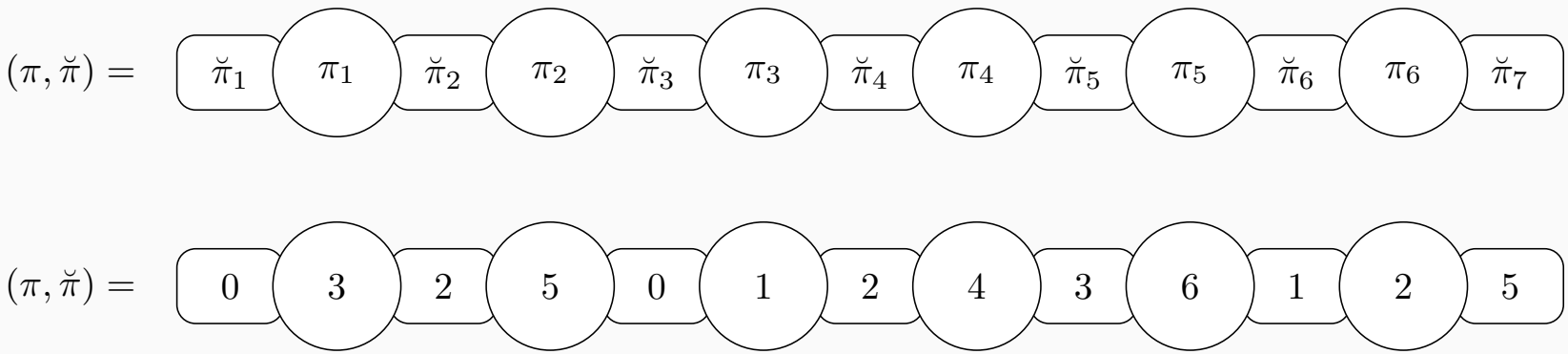


Figura 2: Exemplo da representação de um genoma com $\pi = (3\ 5\ 1\ 4\ 6\ 2)$ e $\breve{\pi} = (0, 2, 0, 2, 3, 1, 5)$.

Ordenação de Permutações com Regiões Intergênicas

Reversão Intergênica

- Denotado por $\rho_{(x,y)}^{(i,j)}$
- Inverte um segmento do genoma
- Regiões intergênicas $\breve{\pi}_i$ e $\breve{\pi}_{j+1}$ são afetadas
- Evento conservativo

Ordenação de Permutações com Regiões Intergênicas

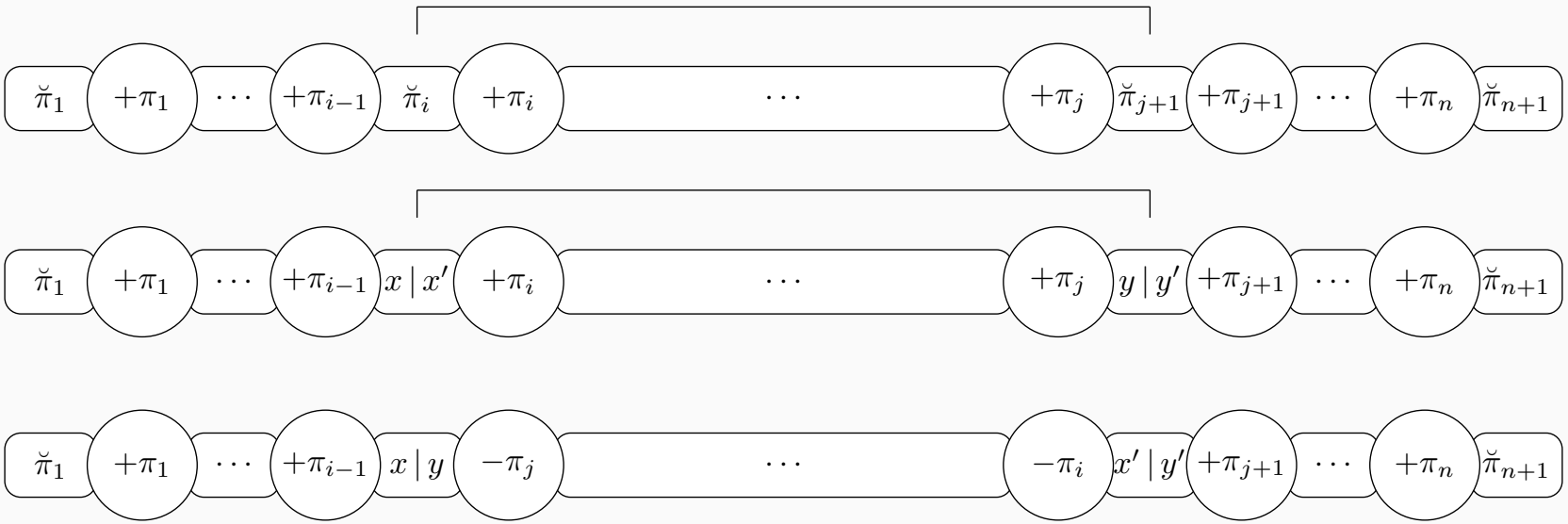


Figura 3: Exemplo de uma reversão intergênica $\rho^{(i,j)}_{(x,y)}$.

Ordenação de Permutações com Regiões Intergênicas

Transposição Intergênica

- Denotado por $\tau^{(i,j,k)}_{(x,y,z)}$
- Troca dois segmentos consecutivos do genoma
- Regiões intergênicas $\breve{\pi}_i$, $\breve{\pi}_j$ e $\breve{\pi}_k$ são afetadas
- Evento conservativo

Ordenação de Permutações com Regiões Intergênicas

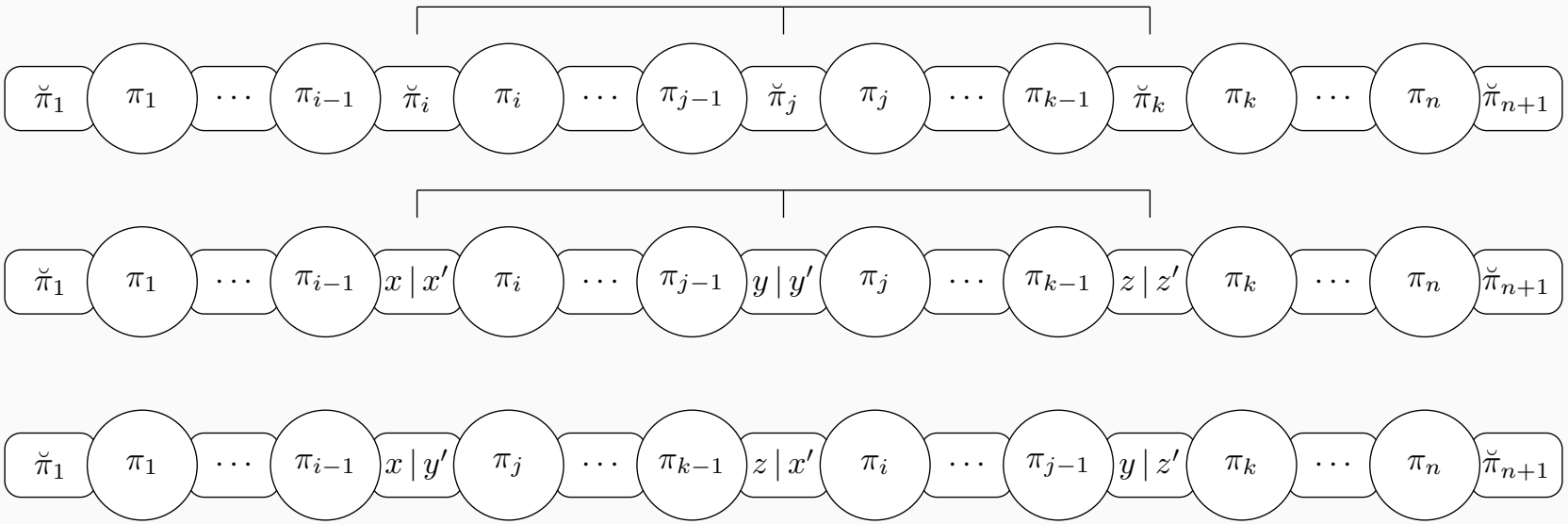


Figura 4: Exemplo de uma transposição intergênica $\tau^{(i,j,k)}_{(x,y,z)}$.

Ordenação de Permutações com Regiões Intergênicas

Inserção Intergênica

- Denotado por ϕ_x^i
- Insere nucleotídeos em uma região intergênica
- Região intergênica $\breve{\pi}_i$ é afetada
- Evento não conservativo

Deleção Intergênica

- Denotado por ψ_x^i
- Remove nucleotídeos de uma região intergênica
- Região intergênica $\breve{\pi}_i$ é afetada
- Evento não conservativo

Ordenação de Permutações com Regiões Intergênicas

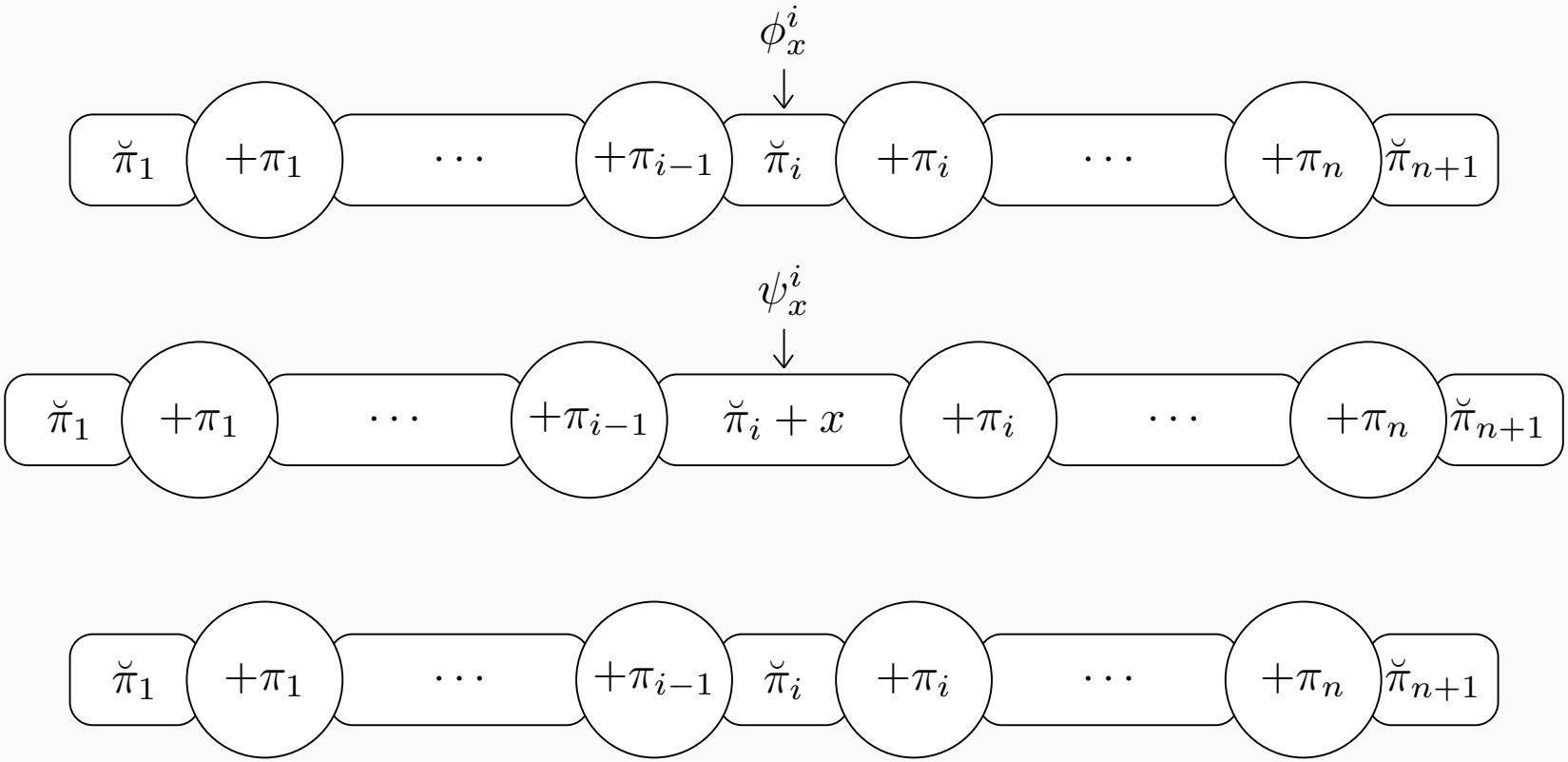


Figura 5: Exemplo de uma inserção intergênica ϕ_x^i e uma deleção intergênica ψ_x^i.

Conceitos

Distância de Strings

Distância de Strings

Representação do Genoma

- Uma string $\beta = (\beta_1, \ldots, \beta_n)$
- $|\beta|$: tamanho da string β
- Σ: alfabeto
- $occ(\alpha, \beta)$: ocorrência do caractere α na string β
- Strings balanceadas

Caso Com Sinais

- $\beta = (+A\ +B\ +C\ -C\ -B\ +A\ -D)$

Caso Sem Sinais

- $\beta = (A\ B\ C\ C\ B\ A\ D)$

Distância de Strings

Reversão

- Denotado por $\rho(i, j)$
- Inverte um segmento do genoma
- Exemplo:

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

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

Distância de Strings

Transposição

- Denotado por $\tau(i, j, k)$
- Troca dois segmentos consecutivos do genoma
- Exemplo:

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

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

Problemas

Ordenação de Permutações com Regiões Intergênicas

Ordenação de Permutações com Regiões Intergênicas

- Considerando apenas os eventos conservativos:
  1. Ordenação de Permutações com Sinais por Reversões Intergênicas (SSbIR)
  2. Ordenação de Permutações sem Sinais por Reversões Intergênicas (SbIR)
  3. Ordenação de Permutações sem Sinais por Transposições Intergênicas (SbIT)
  4. Ordenação de Permutações com Sinais por Reversões e Transposições Intergênicas (SSbIRT)
  5. Ordenação de Permutações sem Sinais por Reversões e Transposições Intergênicas (SbIRT)

Ordenação de Permutações com Regiões Intergênicas

- Considerando eventos conservativos e não conservativos:
  6. Ordenação de Permutações com Sinais por Reversões, Inserções e Deleções Intergênicas (SSbIRID)
  7. Ordenação de Permutações sem Sinais por Reversões, Inserções e Deleções Intergênicas (SbIRID)
  8. Ordenação de Permutações sem Sinais por Transposições, Inserções e Deleções Intergênicas (SbITID)
  9. Ordenação de Permutações com Sinais por Reversões, Transposições, Inserções e Deleções Intergênicas (SSbIRTID)
  10. Ordenação de Permutações sem Sinais por Reversões, Transposições, Inserções e Deleções Intergênicas (SbIRTID)

Ordenação de Permutações com Regiões Intergênicas

Modelo	Sinal	Clássico	Com Região Intergênicas
Reversão	Com	P [8]	NP-Difícil [9]
	Sem	NP-Difícil [6]	NP-Difícil [4]
Transposição	Sem	NP-Difícil [5]	NP-Difícil [9]
Reversão e Transposição	Com	NP-Difícil [10]	NP-Difícil [9]
	Sem	NP-Difícil [10]	NP-Difícil [3]

Tabela 1: Complexidade dos modelos clássicos e dos modelos que incorporam a informação das regiões intergênicas.

Problemas

Distância de Strings com Número Máximo de Cópias

Distância de Strings

- Versões considerando no máximo k cópias por caractere:
  1. Distância de Strings com Sinais por Reversões (SKSDbR) [11]
  2. Distância de Strings sem Sinais por Reversões (KSDbR) [7]
  3. Distância de Strings sem Sinais por Transposições (KSDbT) [11]
  4. Distância de Strings com Sinais por Reversões e Transposições (SKSDbRT)
  5. Distância de Strings sem Sinais por Reversões e Transposições (KSDbRT)

Objetivos

Objetivos

Estudo das Versões dos Problemas

- Foco teórico e prático
- Ordenação de Permutações com Regiões Intergênicas:
  - Investigação sobre a complexidade das versões do problema
  - Desenvolvimento de algoritmos de aproximação
- Distância de Strings com Número Máximo de Cópias:
  - Estudo de teoremas e provas de trabalhos existentes na literatura
  - Desenvolvimento e aplicação de meta-heurísticas
- Experimentos utilizando bases de dados sintéticos ou reais

Cronograma

Cronograma

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

1. Obtenção dos créditos obrigatórios em disciplinas do programa de doutorado
2. Doutorado sanduíche no exterior (Université de Nantes)
3. Exame de Qualificação Específico (EQE)
4. Participação no Programa de Estágio Docente (PED)

Cronograma

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

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

Cronograma

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

8. Escrita da tese
9. Revisão da tese
10. Defesa da tese

Resultados Preliminares

Resultados Preliminares

Sorting by Reversals, Transpositions, and Indels on both Gene Order and Intergenic Sizes. *K. L. Brito, G. Jean, G. Fertin, A. R. Oliveira, U. Dias, and Z. Dias (ISBRA'2019)*

- 15th International Symposium on Bioinformatics Research and Applications
- Barcelona, Espanha
- 03 a 06 de junho de 2019

Resultados Preliminares

Sorting by Reversals, Transpositions, and Indels on both Gene Order and Intergenic Sizes. *K. L. Brito, G. Jean, G. Fertin, A. R. Oliveira, U. Dias, and Z. Dias (ISBRA'2019)*

Lema

O problema Ordenação de Permutações sem Sinais por Reversões Intergênicas (SbIR) é NP-Difícil.

Lema

O problema Ordenação de Permutações sem Sinais por Reversões, Inserções e Deleções Intergênicas (SbIRID) é NP-Difícil.

Resultados Preliminares

Sorting by Reversals, Transpositions, and Indels on both Gene Order and Intergenic Sizes. *K. L. Brito, G. Jean, G. Fertin, A. R. Oliveira, U. Dias, and Z. Dias (ISBRA'2019)*

- Conceito de *Breakpoint Intergênico*
- Limitantes Inferiores
- 4-aproximação para os problemas SbIR e SbIRID
- 6-aproximação para os problemas SbIRT e SbIRTID

Resultados Preliminares

Sorting by Genome Rearrangements on both Gene Order and Intergenic Sizes. *K. L. Brito, G. Jean, G. Fertin, A. R. Oliveira, U. Dias, and Z. Dias (Journal of Computational Biology)*

Lema

O problema Ordenação de Permutações sem Sinais por Reversões e Transposições Intergênicas (SbIRT) é NP-Difícil.

Lema

O problema Ordenação de Permutações sem Sinais por Reversões, Transposições, Inserções e Deleções Intergênicas (SbIRTID) é NP-Difícil.

Resultados Preliminares

Sorting by Genome Rearrangements on both Gene Order and Intergenic Sizes. *K. L. Brito, G. Jean, G. Fertin, A. R. Oliveira, U. Dias, and Z. Dias (Journal of Computational Biology)*

- Propriedades do problema
- Melhoria no fator de aproximação dos algoritmos
- 4.5-aproximação para os problemas SbIRT e SbIRTID
- Resultados experimentais
- Comparativo entre resultados teóricos e práticos

Resultados Preliminares

Sorting by Genome Rearrangements on both Gene Order and Intergenic Sizes. *K. L. Brito, G. Jean, G. Fertin, A. R. Oliveira, U. Dias, and Z. Dias (Journal of Computational Biology)*

- Investigação de variações ponderadas
- Um cenário considerando os eventos de reversão, inserção e deleção intergênica
- Três cenários considerando os eventos de reversão, transposição, inserção e deleção intergênica
- Dois cenários considerando os eventos de reversão e transposição intergênica
- Um algoritmo de aproximação apresentado para cada cenário estudado

Resultados Preliminares

Distância de Strings

$$
\begin{aligned}
S &= (+A\ -B\ +A\ -B\ +A\ -B) \\
S^{m'} = \pi &= (+1\ \ -4\ +2\ \ -5\ +3\ -6) \\
T^{m''} = \sigma &= (+6\ \ -3\ +5\ \ -2\ +4\ -1) \\
T &= (+B\ -A\ +B\ -A\ +B\ -A)
\end{aligned}
$$

Lema

Existe um par de mapeamentos m' e m'' tal que $d(S, T) = d(\pi, \sigma)$.

Resultados Preliminares

Distância de Strings

- Aplicação de meta-heurísticas para geração dos mapeamentos
- Utilização dos algoritmos conhecidos para os problemas com permutações
- Trabalho inicial com o desenvolvimento de três heurísticas
  - Mapeamento Aleatório (MA)
  - Busca Local (BL)
  - Algoritmo Genético (AG)
- Distância de Strings com Sinais por Reversões
- Resultados promissores

Resultados Preliminares

Tabela 2: Distâncias médias fornecidas pelas heurísticas para strings de tamanho 100 com no máximo k ocorrências por caractere.

20 Reversões								
k	03	04	05	06	07	08	09	10
MA	35.98	37.09	38.51	39.78	40.48	41.39	42.25	43.71
BL	21.09	20.96	21.56	22.50	22.02	22.61	22.72	24.02
AG	19.89	19.90	19.90	20.04	20.17	20.85	21.63	22.54
30 Reversões								
k	03	04	05	06	07	08	09	10
MA	44.63	47.01	47.48	48.07	49.87	50.43	51.34	52.42
BL	32.42	33.67	33.92	33.50	36.00	36.57	36.94	37.90
AG	29.69	29.85	29.79	29.76	30.08	30.73	31.44	32.73

Resultados Preliminares

Tabela 3: Distâncias médias fornecidas pelas heurísticas para strings de tamanho 100 com no máximo k ocorrências por caractere.

40 Reversões								
k	03	04	05	06	07	08	09	10
MA	54.13	55.31	56.27	56.84	58.45	58.31	59.69	60.63
BL	46.25	46.28	47.25	48.20	49.30	49.31	50.35	51.26
AG	39.52	39.25	39.53	39.57	39.86	40.00	41.74	42.28
50 Reversões								
k	03	04	05	06	07	08	09	10
MA	61.99	62.67	63.32	63.91	65.07	65.32	66.03	66.67
BL	57.47	58.68	58.62	59.07	59.81	59.90	60.18	61.31
AG	48.94	48.80	48.49	48.84	49.47	49.71	50.54	51.52

Resultados Preliminares

Próximos Passos:

- Refinamento dos parâmetros
- Comparação com os resultados existentes na literatura
- Aplicação nas demais versões do problema

Referências

[1] Biller, P., Guéguen, L., Knibbe, C., and Tannier, E.
Breaking Good: Accounting for Fragility of Genomic Regions in Rearrangement Distance Estimation.
Genome Biology and Evolution 8, 5 (2016), 1427–1439.

[2] Biller, P., Knibbe, C., Beslon, G., and Tannier, E.
Comparative Genomics on Artificial Life.
In *Pursuit of the Universal* (2016), Springer International Publishing, pp. 35–44.

[3] Brito, K. L., Jean, G., Fertin, G., Oliveira, A. R., Dias, U., and Dias, Z.
Sorting by Genome Rearrangements on both Gene Order and Intergenic Sizes.
Journal of Computational Biology (2019), 1–18.

Referências

[4] Brito, K. L., Jean, G., Fertin, G., Oliveira, A. R., Dias, U., and Dias, Z.
Sorting by Reversals, Transpositions, and Indels on both Gene Order and Intergenic Sizes.
In *International Symposium on Bioinformatics Research and Applications* (2019), Springer International Publishing, pp. 28–39.

[5] Bulteau, L., Fertin, G., and Rusu, I.
Sorting by Transpositions is Difficult.
SIAM Journal on Computing 26, 3 (2012), 1148–1180.

[6] Caprara, A.
Sorting Permutations by Reversals and Eulerian Cycle Decompositions.
SIAM Journal on Discrete Mathematics 12, 1 (1999), 91–110.

Referências

[7] Christie, D. A., and Irving, R. W.
Sorting Strings by Reversals and by Transpositions.
SIAM Journal on Discrete Mathematics 14, 2 (2001), 193–206.

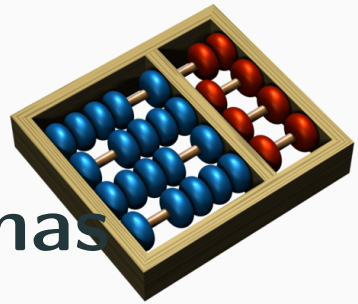
[8] Hannenhalli, S., and Pevzner, P. A.
Transforming Cabbage into Turnip: Polynomial Algorithm for Sorting Signed Permutations by Reversals.
Journal of the ACM 46, 1 (1999), 1–27.

[9] Oliveira, A. R.
***Modelos Restritos e Intergênicos para a Ordenação por Reversões e Transposições*.**
PhD thesis, University of Campinas, 2019.

Referências

[10] Oliveira, A. R., Brito, K. L., Dias, U., and Dias, Z.
On the Complexity of Sorting by Reversals and Transpositions Problems.
Journal of Computational Biology 26 (2019), 1223–1229.

[11] Radcliffe, A. J., Scott, A. D., and Wilmer, E. L.
Reversals and Transpositions Over Finite Alphabets.
SIAM Journal on Discrete Mathematics 19, 1 (2005), 224–244.



Problemas de Rearranjos de Genomas Considerando Regiões Intergênicas ou Genes Repetidos

Exame de Qualificação Específico de Doutorado

Aluno: Klairton de Lima Brito
Orientador: Prof. Dr. Zanoni Dias
Coorientador: Prof. Dr. Ulisses Martins Dias

9 de Dezembro de 2019

Instituto de Computação - Universidade Estadual de Campinas