Complexidade da Distância de Translocação para Genomas Sem Sinal

Lucas Angelo da Silveira & Mauricio Ayala-Rincón

Grupo de Teoria da Computação
Programa de Pós-Graduação em Informática
Universidade de Brasília

Sumário

Sumário

Introdução

Um genoma é constituído de toda informação hereditária de um ser, formado por um conjunto de diferentes genes que se encontra em cada núcleo de uma determinada espécie.

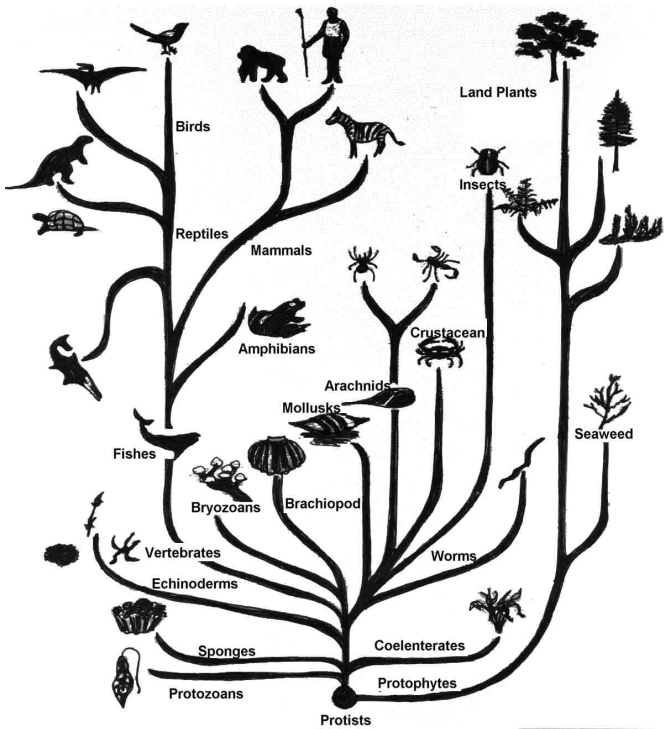


Figura : fonte:imgkid.com/ape-phylogenetic-tree.shtml

Definindo um Genoma

Genomas, cromossomos e genes

$$A=\{(a_{11}, a_{12}, \cdots, a_{1m_1}), (a_{21}, a_{22}, \cdots, a_{2m_2}) \cdots\cdots (a_{N1}, a_{N2}, \cdots, a_{Nm_N})\}$$

$$B=\{(b_{11}, b_{12}, \cdots, b_{1n_1}), (b_{21}, b_{22}, \cdots, b_{2n_2}) \cdots\cdots (b_{N1}, b_{N2}, \cdots, b_{Nn_N})\}$$

Um cromossomo não apresenta orientação, baseado nesta definição podemos visualizar um cromossomo X=$(x_1, x_2, ..., x_k)$ como $(x_k, x_{k-1}, \ldots, x_1)$.

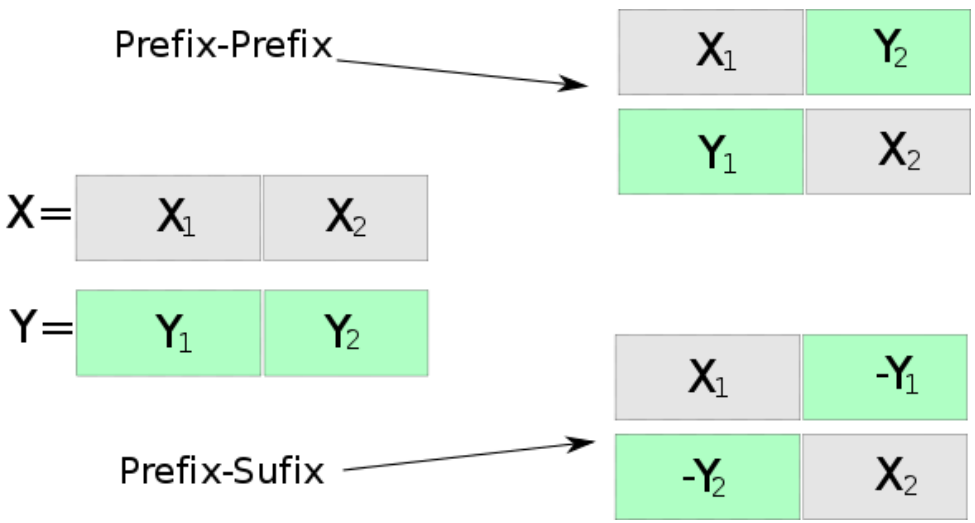
Sumário

Translocação e Suas Operações

Primeiros estudos

O problema de translocação para genomas foi primeiramente estudado por *Kececioglu e Ravi* em 1995, no artigo "Algorithms for Evolutionary Distances Between Genomes With Translocation".

Operações sobre os Segmentos X_1, X_2, Y_1 e Y_2



Igualdade entre Genomas

Dado um cromossomo $X = (x_1, x_2, \ldots, x_k)$, os genes x_1 e $-x_k$ são chamados de caudas. Assim temos que dois genomas são co-caudais, se seus conjuntos de caudas são iguais. Exemplo, seja A e B dois genomas.

$$A = \{(+1, -2, +4, +3, -5, +6), (7, -9, -8, +10)\},$$

$$B = \{(+1, +2, +3, +4, +5, +6), (+7, +8, +9, +10)\},$$

com $C_1 = \{1, -6, 7, -10\}$ e C_2=$\{1, -6, 7, -10\}$.

Definindo o Problema Formalmente

Ordenar Genomas por Translocação

Entrada: Dois genomas A e B, onde B é a identidade.
Questão: Existe uma sequência de translocações $\rho_1, \rho_2, \ldots, \rho_t$ que transforma A em B, e t é mínimo.

Complexidade para Genomas Via Translocação

Autor	Complexidade
Hannenhalli	$O(n^3)$
Bergeron	$O(n^3)$
Lusheng Wang	$O(n^2)$

Tabela : Problema na versão com genomas com sinais.

Autor	Complexidade	Raio
Yun Cui	$O(n^2)$	1.75
Yun Cui	$O(n^2 + (\frac{4}{\epsilon})^{1.5}\sqrt{\log(\frac{4}{\epsilon})2^{\frac{4}{\epsilon}}})$	$1.5 + \epsilon$
Haitao Jiang	$O(n^2 + n\log^2 n \log\log n)$	$1.408 + \epsilon$

Tabela : Problema na versão com genomas sem sinais.

Sumário

Definindo uma Permutação π e o Grafo $G(\pi)$

Permutação π

Uma permutação π é uma função bijetiva sobre o conjunto $\{1, \ldots, n\}$.

O grafo de pontos-de-quebra para $\pi = (2, 4, 1, 3)$

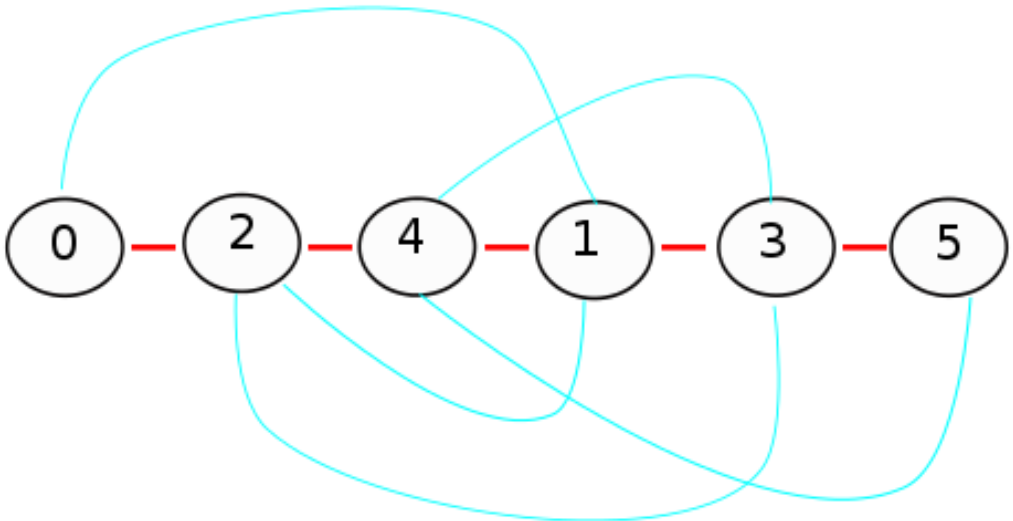


Figura : Grafo de pontos-de-quebra $G(\pi)$ para uma permutação estendida $\pi = (0, 2, 4, 1, 3, 5)$ e $\iota = (0, 1, 2, 3, 4, 5)$.

R é o conjunto de arestas vermelhas e **B** é o conjunto de arestas azuis.

Relações de um G(π)

G(π) satisfaz:

- cada subgrafo a partir de G(R) e G(B) é um caminho simples.
- cada nodo $i \in V$ tem o mesmo grau (0, 1 ou 2) em G(R) e G(B).
- Uma aresta em G(R) não está contida em G(B), e vice-versa.

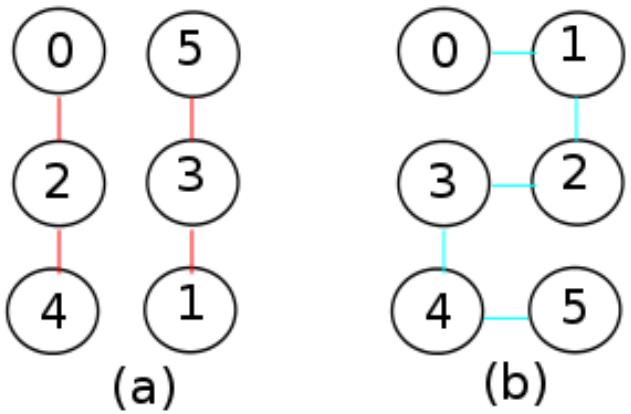


Figura : Subgrafos a partir de $\pi = (0, 2, 4, 1, 3, 5)$ e $\iota = (0, 1, 2, 3, 4, 5)$. (a) Subgrafo G(R), (b) Subgrafo G(B)

Ciclos alternados de G(π)

É uma sequência de arestas $r_1, b_1, r_2, b_2, \ldots, r_m, b_m$, onde $r_i \in R$ e $b_i \in B$.
Exemplo:$\{(0,2),(2,1),(1,4),(4,3),(3,1),(1,0)\}$

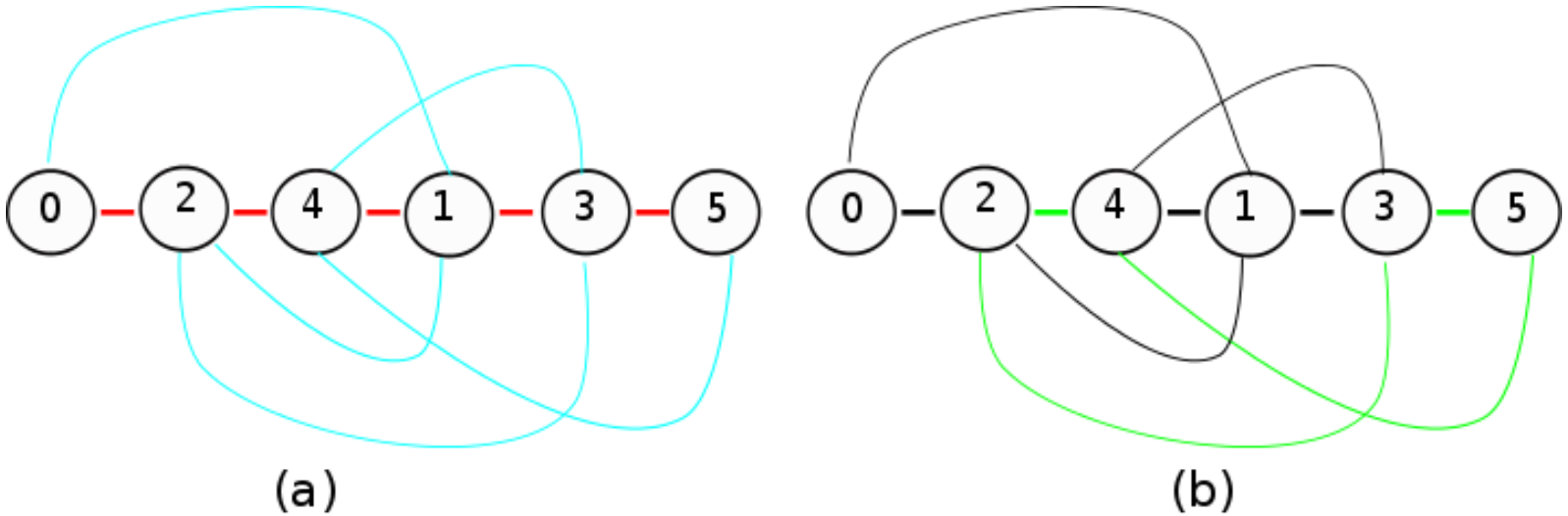


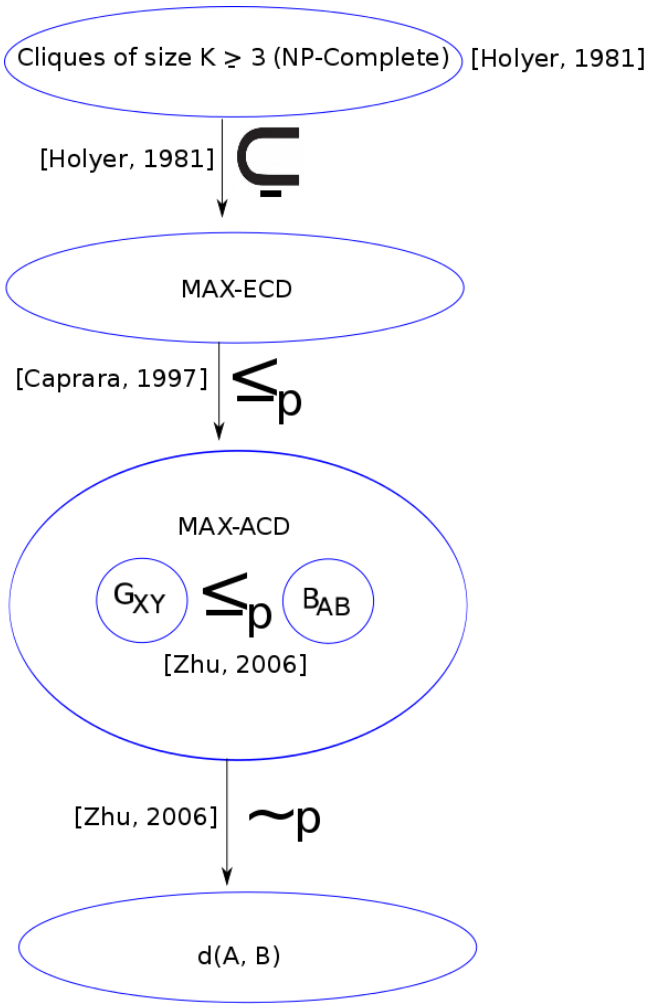
Figura : (a) Grafo de pontos-de-quebra G(π), (b) máxima decomposição em ciclos alternados de G(π).

Maximizar ciclos em G(π)

Conhecido como o problema MAX-ACD, que consiste a partir de G(π) encontrar uma máxima decomposição de ciclos alternados.

Sumário

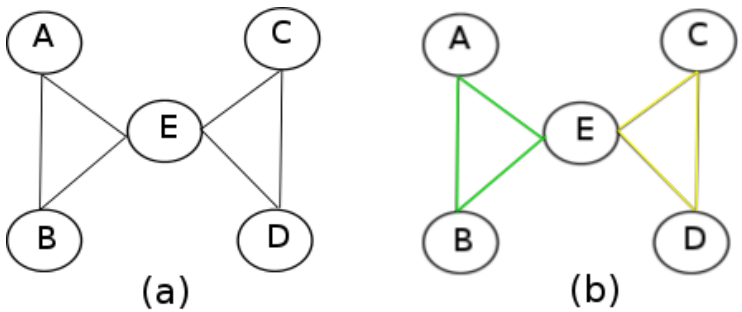
Sequência de Reduções Polinomiais



Relação entre MAX-ACD e Grafos Eulerianos

MAX-ECD

Consiste em particionar o conjunto de arestas E de um grafo euleriano H no máximo número de ciclos.



MAX-ECD é NP-Hard

Holyer provou em [Holyer, 1981] que decompor o conjunto de arestas de um grafo H em cliques de tamanho k para $k \geq 3$ é *NP-Completo*. Para k=3, consiste em particionar H em triângulos, assim podemos assumir que H é euleriano.

Reduzir de MAX-ECD para MAX-ACD

Fase 1

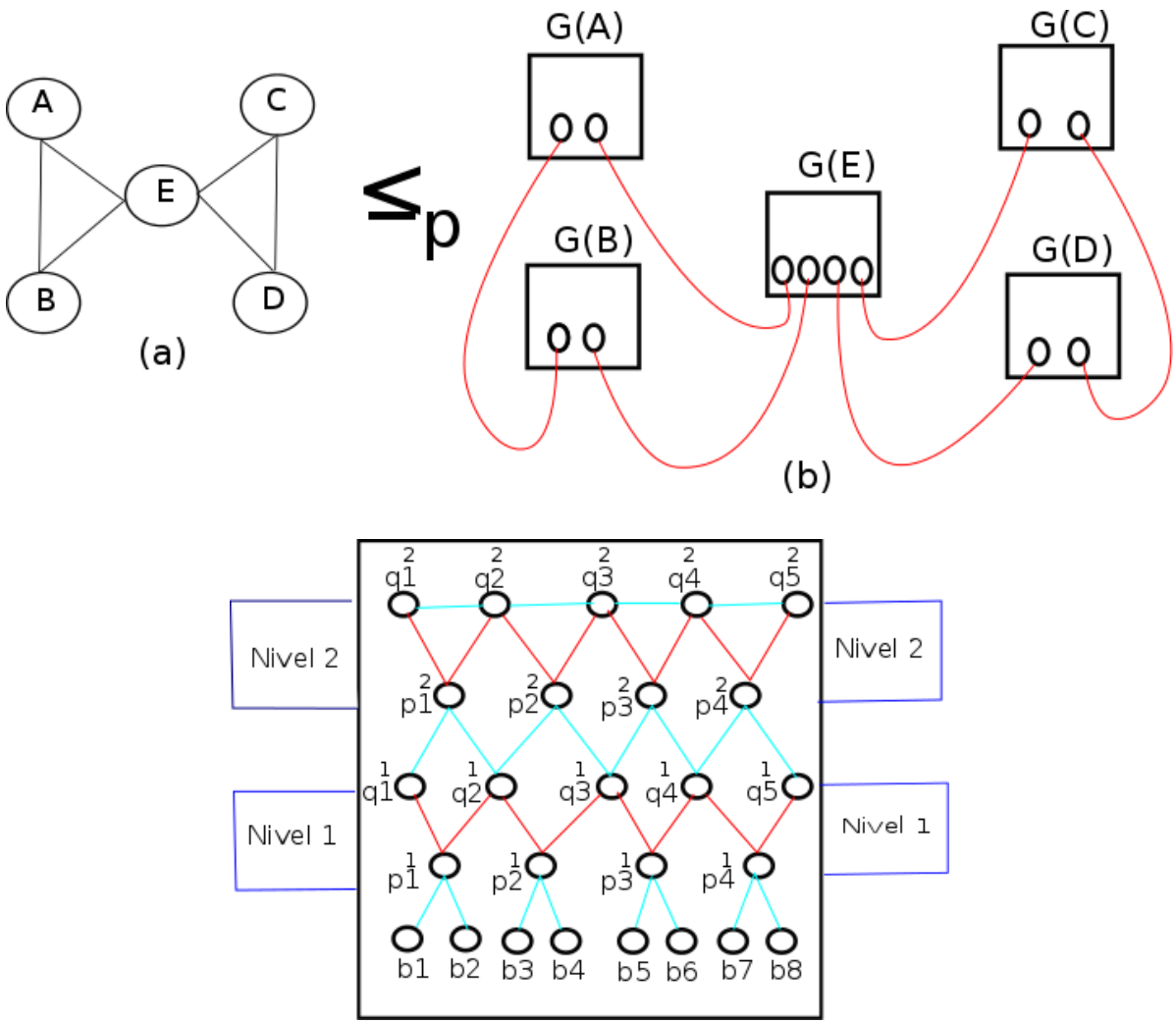


Figura : G(d, m), pra d=8 e m=2, s= $\frac{d}{2}$, r=$\frac{d}{4}$

Afirmação 1: $m \geq l + r$

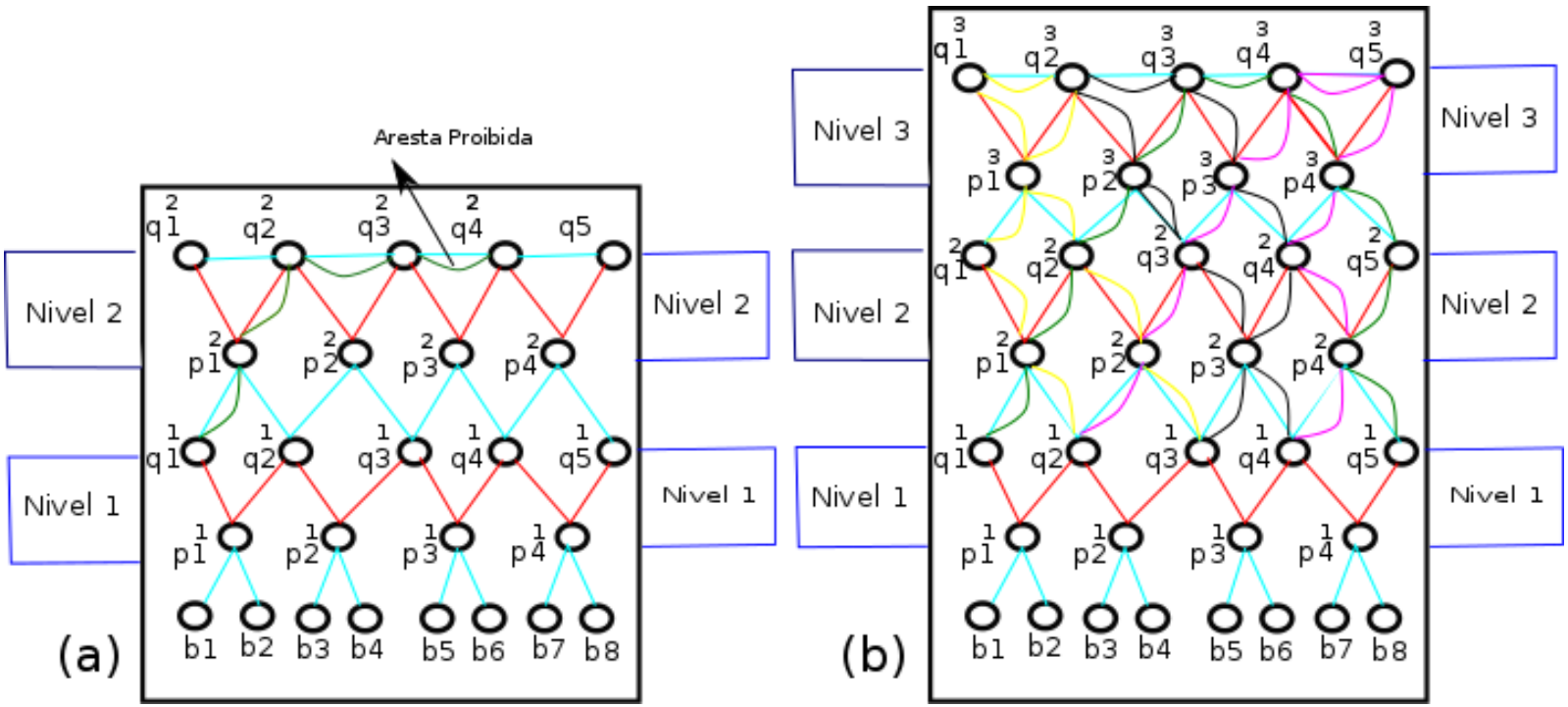


Figura : G(d, m), pra d=8 e m=2, s= $\frac{d}{2}$, r=$\frac{d}{4}$ encontrar um caminho de q_1^1 para q_5^1

Afirmação 2: $m \geq r(s-1)+1$

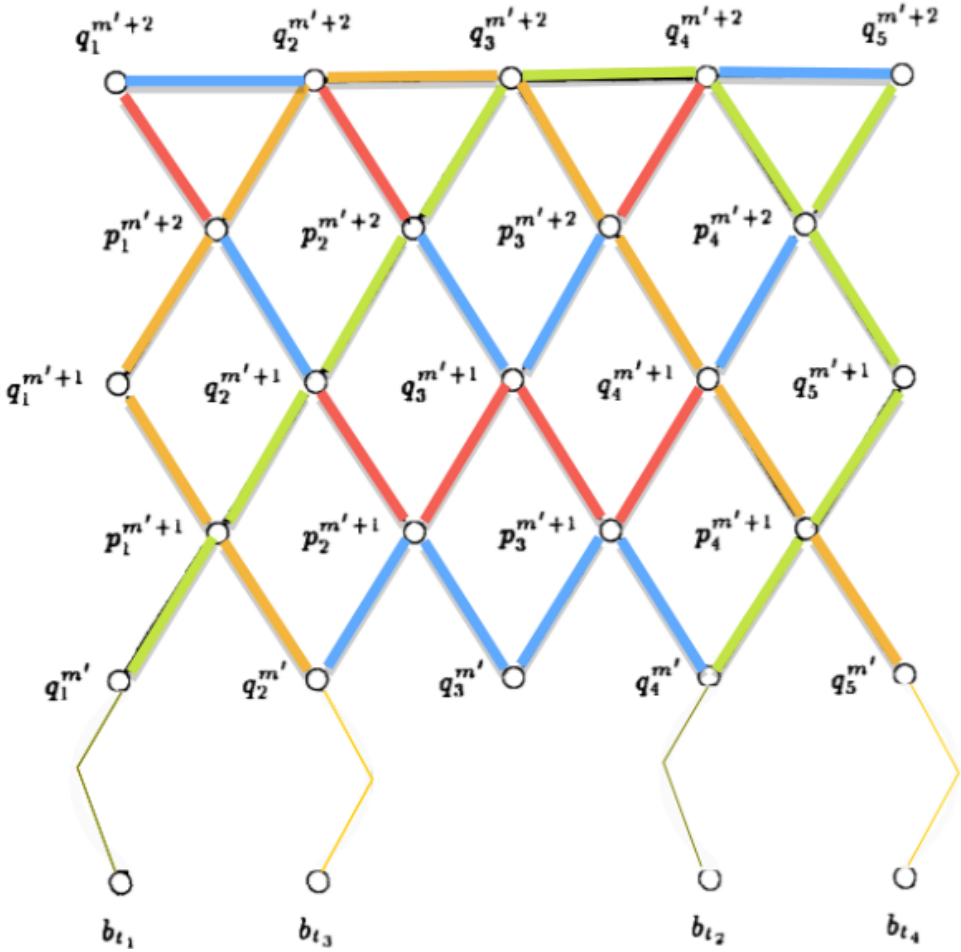
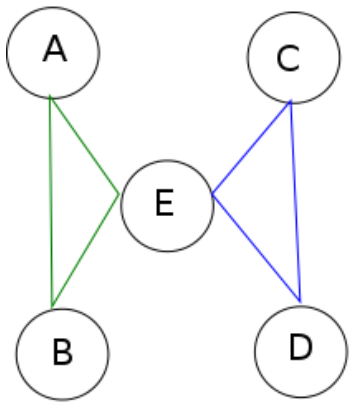
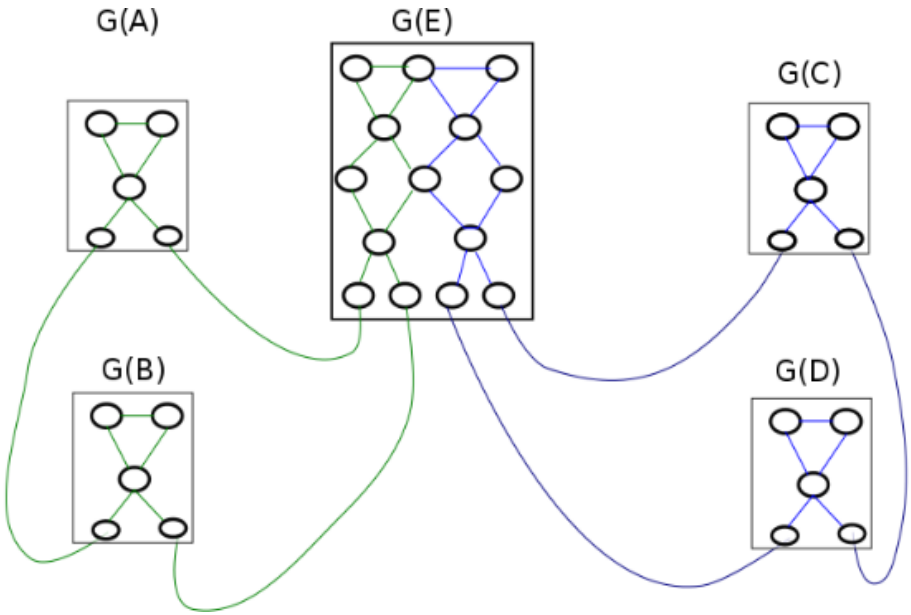


Figura : G(d, m), para d=8, s= $\frac{d}{2}$, r=$\frac{d}{4}$, dois caminhos (b_{t1}, b_{t3}), (b_{t2}, b_{t4})

Correspondência um-para-um





Complexidade

$m = r(s-1)+1$
Quantidade de vértices é dado por $V = ((m.(d+1)+d).n$

Sumário

Grafo de Pontos-de-Quebra para Genomas sem Sinais

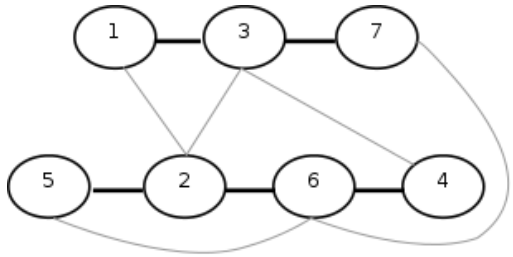
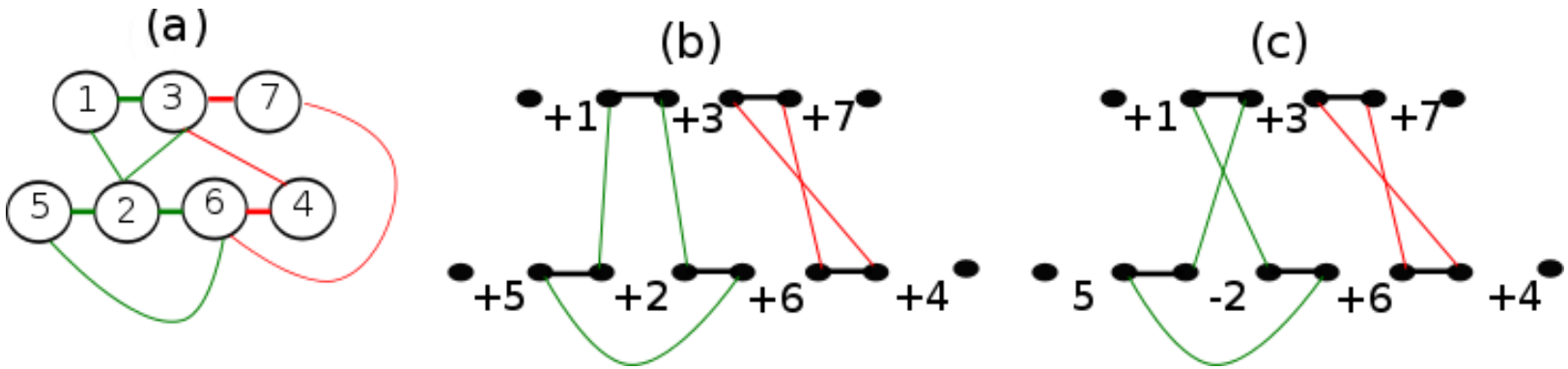


Figura : Genomas A=$\{(1,3,7),(5,2,6,4)\}$ e B=$\{(1,2,3,4),(5,6,7)\}$



Noção de SP e MinSP

Segmento de pelos menos 3 elementos em um cromossomo. Exemplo:
$A = \{(1,3,2,4,5,8,6),(7,9)\}$ e $B = \{(1,2,3,4,5,6),(7,8,9)\}$
$\{1,3,2,4,5\}$ é uma SP e $\{1,3,2,4\}$ é uma MinSP.

Processo de Redução de G_{XY} para B_{AB}

dado dois cromossomo X e Y sem sinal. Construímos dois genomas A e B. Onde $A = \{X_1, X_2\}$ e $B = \{Y_1, Y_2\}$. Exemplo: X=$(1, 3, 4, 2, 5)$, Y=$(1, 2, 3, 4, 5)$.

X_1

$t_{1,k} = 3n - 2 + k$ para $1 \leq k \leq n - 1$.
$X_1 = (1, 14, 3, 15, 4, 16, 2, 17, 5)$.

X_2

X_2 contém dois novos tipos de genes.
$T_{2,l} = n + l$, para $1 \leq l \leq 2(n - 1)$ e $s_i = 4n - 3 + i$, para $1 \leq i \leq (n - 2)d$.
$X_2 = (6, 7, 18, 19, 20, 21, 8, 9, 22, 23, 24, 25, 10, 11, 26, 27, 28, 29, 12, 13)$.

Processo de Redução de G_{XY} para B_{AB}(Cont.)

$Y_1 = Y$.

Y_2

$Y_2 = (6, 14, 7, 18, 19, 20, 21, 8, 15, 9, 22, 23, 24, 25, 10, 16, 11, 26, 27, 28, 29, 12, 17, 13)$.

Grafo de Pontos-de-Quebra B_{AB}

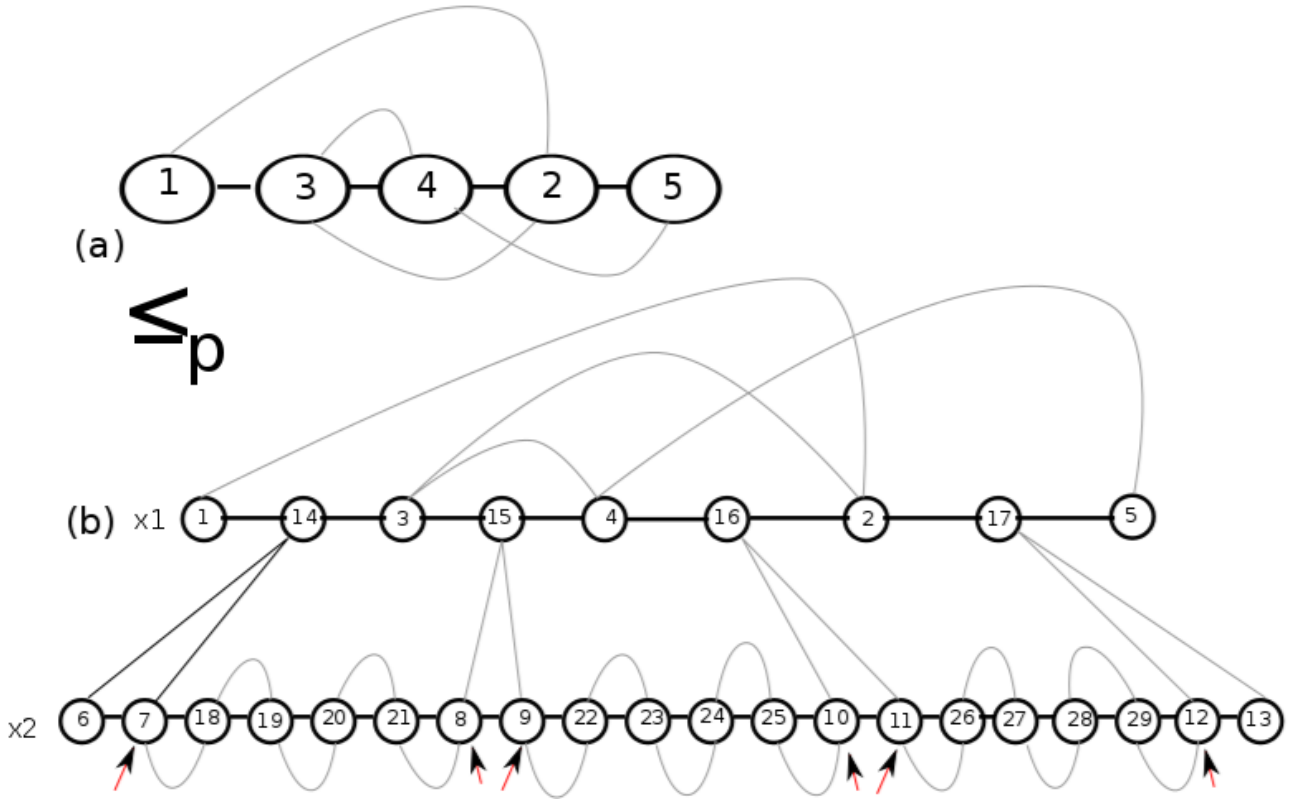
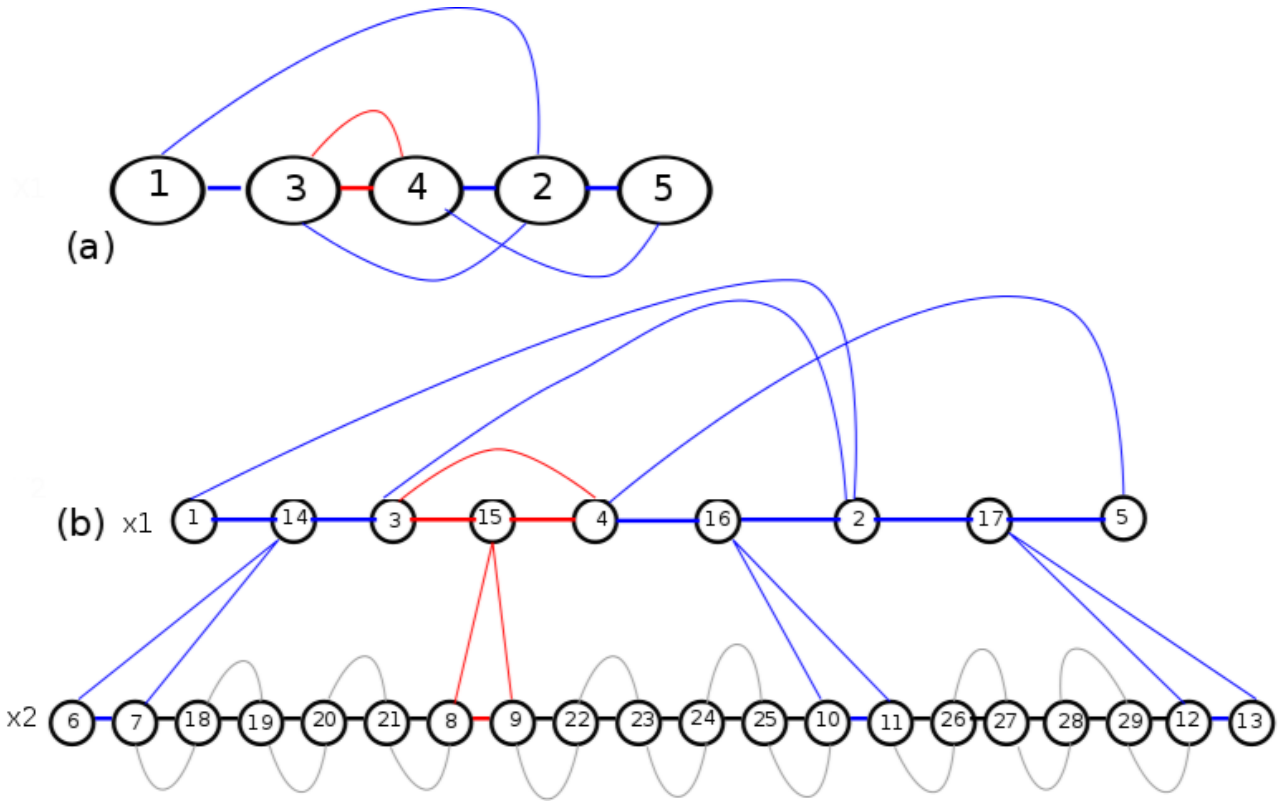


Figura : (a)Grafo de pontos-de-quebra G_{XY} de X e Y, (b) O grafo de ponto de quebra B_AB dos genomas A e B.

Existe $4n - 3 + (n - 2).d$ genes em ambos os genomas A e B.

Relação de decomposição de G_{XY} com B_{AB}

Afirmativa 1: Assumindo que $d \geq n-1$.Existe J ciclos em G_{XY} e B_{AB} pode ser decomposto em $(n-2).(d+1)+J$ ciclos.



Considerando Arestas Restantes

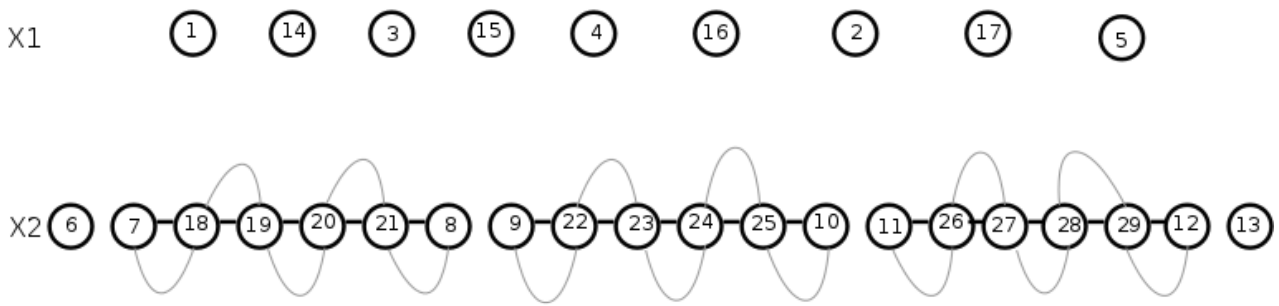


Figura : Cenário ao descartar arestas usadas no ciclos longos.

Condição para Redução em Tempo Polinomial de G_{XY} para B_{AB}

Temos $V = 4n - 3 + (n - 2).d$ genes em ambos os genomas A e B.
Fazendo $d = n - 1$. Temos que a redução é polinomial.

Distância de Translocação em função da Decomposição de B_{AB}

Afirmativa 2: $d(A, B) = 3n - 3 - J$.
Teorema 1 em [Zhu, 2006] mostra que a distância de translocação entre genomas A e B é dada por:$d(A, B) = n - N - C_{AB}$.

$$
\begin{aligned}
d(A,B) \quad \leq \quad & \\
& d(A', B') = n - N - C_{AB} \\
& \qquad = 4n - 3 + (n-2)d - 2 - ((n-2)(d+1) + J) \\
& \qquad = 4n - 3 + dn - 2d - 2 - (dn + n - 2d - 2 + J) \\
& \qquad = 4n - 3 + dn - 2d - 2 - dn - n + 2d + 2 - J \\
& \qquad = 4n - n - 3 + 2 - 2 - 2d + 2d - J \\
& \qquad = 3n - 3 - J.
\end{aligned}
$$

Teorema 12 em [Hannenhalli, 1996] mostra que se não há *minSPs* em A e B, $d(A', B') = n - N - C_{AB}$.
Logo $d(A, B) = 3n - 3 - J$.

Conclusão

Translocação é usada para estimar a distância evolutiva entre duas espécies. Essencialmente foi mostrado que o problema de distância de translocação para genomas sem sinais é *NP-Hard*. Para o qual, utilizou-se o fato que MAX-ECD é *NP-hard*, mostrando sequências de reduções polinomiais a partir deste problema.

Trabalhos Futuros

1. Implementar o algoritmo de raio de aproximação $1.408 + \epsilon$.
2. Implementar técnicas de algoritmos genéticos, afim de superar raio de aproximação mencionado acima.

A. Bergeron et. al. (J. of Computational Biology 13(2) : 567–578, 2006). On Sorting by Translocations.

A. Caprara et. al. (Conf. on Computational Molecular Biology, 1997, *pp.*75–83). Sorting by reversal is difficult.

D. Zhu et. al. (J. of Theor. Comput. Sci, 352(1 − 3) : 322–328, 2006). On the complexity of unsigned translocation distance.

S. Hannenhalli et. al. (J. Disc. Appl. Math., 71(1–3), 137–151, 1996). Polynomial-time algorithm for computing translocation distance between genomes.

Ian-Holyer (SIAM J. on Computing, 713–717, 1981). The NP-Completeness of Some Edge-Partition Problems