MO640/MC668

Guilherme P. Telles

IC-Unicamp

Avisado está

- Estes slides são incompletos.
- Estes slides contêm erros.

Parte I

Rearranjo de genomas

Rearranjo de genomas

- Para comparar genomas inteiros pode ser mais adequado considerar mutações que afetam blocos maiores, e não simplesmente inserções, remoções e substituições.
- Blocos representam segmentos conservados dos genomas.
- Blocos podem ser orientados ou não orientados.
  - A orientação indica que o bloco indica que a orientação da molécula é conhecida.
  - A falta de orientação indica que a ordem relativa dos blocos é conhecida, mas não a direção.

Rearranjo de genomas

- Operações que atuam sobre blocos são por exemplo
  - reversão,
  - transposição,
  - fissão e fusão,
  - translocação,
  - inserção, remoção e duplicação.
- Essas operações são chamadas de operações de **rearranjo de genomas**

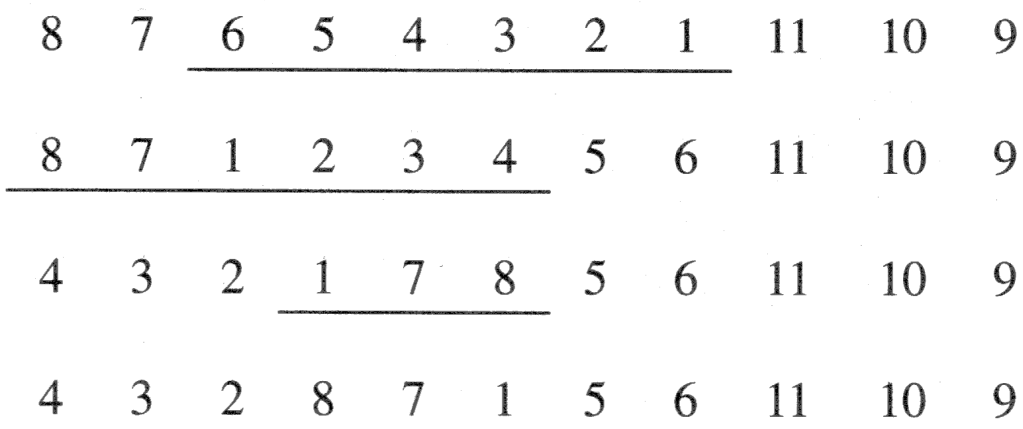
Rearranjo de genomas

- O problema biológico é comparar o genomas em termos do número mínimo de operações de um ou mais tipos que transformam um genoma no outro.
- As reversões são mutações que mais levam a diferenças significativas entre genomas.

Reversão


8 7 6 5 4 3 2 1 11 10 9
8 7 6 5 4 3 2 1 11 10 9
8 2 3 4 5 6 7 1 11 10 9
8 2 3 4 5 1 7 6 11 10 9
4 3 2 8 5 1 7 6 11 10 9
4 3 2 8 7 1 5 6 11 10 9
4 3 2 8 7 1 5 6 11 10 9
4 3 2 8 7 1 5 6 11 10 9

Reversão


8 7 6 5 4 3 2 1 11 10 9
8 7 1 2 3 4 5 6 11 10 9
4 3 2 1 7 8 5 6 11 10 9
4 3 2 8 7 1 5 6 11 10 9

Definições

- Vamos supor que os blocos nos dois genomas que serão comparados estão rotulados entre 1 e n.
- Seja $\mathcal{L}$ um conjunto finito de rótulos. Um conjunto $\mathcal{L}^o$ de rótulos orientados para $\mathcal{L}$ é

$$\mathcal{L}^o = \bigcup_{a \in \mathcal{L}} \{\overrightarrow{a}, \overleftarrow{a}\}$$

- Para um rótulo $x \in \mathcal{L}^o$, $|x|$ é um rótulo de $\mathcal{L}$ obtido pela remoção da orientação.
- Então para todo $x \in \mathcal{L}$ temos $|\overrightarrow{x}| = |\overleftarrow{x}| = a$.
- Para um rótulo x, $\overline{x}$ é o rótulo x com sua orientação invertida.

Permutação orientada

- Uma *permutação orientada* de $\mathcal{L}$ é um mapeamento $\alpha : [1..n] \rightarrow \mathcal{L}^o$ tal que para qualquer rótulo $a \in \mathcal{L}$ existe exatamente um $i \in [1..n]$ com $|\alpha(i)| = a$.
- Uma permutação α pode ser representada como uma sucessão dos seus elementos $\alpha(1), \alpha(2), \ldots, \alpha(n)$.
- A permutação identidade orientada é a permutação I tal que $I(i) = \overrightarrow{i}$, para $1 \leq i \leq n$.

Reversão

- Uma *reversão* transforma uma permutação orientada em outra invertendo a ordem dos elementos de uma porção contínua da permutação e complementando as orientações desses elementos.
- Uma reversão que envolve os elementos $\alpha(i)\,..\,\alpha(j)$ de α é denotada $[i\,..\,j]$.
- Uma reversão $\rho = [i\,..\,j]$ transforma uma permutação α em uma permutação $\alpha\rho$, definida da seguinte forma

$$\alpha[i\,..\,j](k) = \begin{cases} \overline{\alpha(i+j-k)} & \text{se } i \leq k \leq j \\ \alpha(k) & \text{caso contrário} \end{cases}$$

Reversão

- Há $n(n+1)/2$ reversões possíveis entre n rótulos, incluindo as reversões unitárias.

Distância de reversão

- Dadas duas permutações orientadas α e β sobre o mesmo conjunto $\mathcal{L}$ de rótulos, o problema da ordenação por reversões é determinar o número mínimo de reversões que transformam α em β.
- Ou seja, identificar uma série de reversões $\rho_1, \rho_2, \ldots, \rho_t$, com t mínimo, tal que

$$\alpha\rho_1\rho_2 \ldots \rho_t = \beta.$$

- O número t é a *distância de reversão* de α com relação a β e é denotada por $d_\beta(\alpha)$.
- t é único para um par α, β mas a série de reversões que ordena α não necessariamente é única.

Reversão ordenante

- Uma reversão ρ é chamada *reversão ordenante* de α com relação β se

$$d_\beta(\alpha\rho) < d_\beta(\alpha).$$

- (São essas reversões que queremos encontrar para resolver o problema.)
- Como as distâncias de reversão de α a β e de $\alpha\rho$ a β não podem diferir de mais do que uma unidade, temos:

$$d_\beta(\alpha\rho) = d_\beta(\alpha) - 1.$$

- Vamos considerar alguns limites inferiores para a distância de reversão.

Permutação estendida

- Dada uma permutação α, a *versão estendida* de α é obtida adicionando-se a ela um rótulo artificial L antes do primeiro rótulo $\alpha(1)$ e um rótulo artificial R depois do último rótulo $\alpha(n)$.

Ponto de quebra

- Um *ponto de quebra* de α com respeito a β é um par x, y de elementos de $\mathcal{L}^o$ tal que xy aparece na versão estendida de α mas nem xy e nem $\overline{yx}$ aparece na versão estendida de β.
- O número de pontos de quebra de uma permutação orientada α com respeito a β é denotado por $b_\beta(\alpha)$.

Limite inferior para a ordenação

- Uma reversão pode remover no máximo dois pontos de quebra de uma permutação.
- Logo para qualquer α e ρ

$$b(\alpha) - b(\alpha\rho) \leq 2,$$

Pontos de quebra

- Seja $\rho_1, \rho_2, \ldots, \rho_t$ uma seqüência de reversões que converte α em β. Ou seja $\alpha\rho_1\rho_2 \ldots \rho_t = \beta$.

  Então $b(\alpha\rho_1\rho_2 \ldots \rho_t) = b(\beta) = 0$.

  Pela desigualdade anterior temos

$$\begin{aligned} b(\alpha) - b(\alpha\rho_1) &\leq 2, \\ b(\alpha\rho_1) - b(\alpha\rho_1\rho_2) &\leq 2, \\ &\vdots \\ b(\alpha\rho_1 \ldots \rho_{t-1}) - b(\alpha \ldots \rho_t) &\leq 2, \end{aligned}$$

Pontos de quebra

- Somando essas desigualdades e considerando que $b(\alpha\rho_1\rho_2\ldots\rho_t) = b(\beta) = 0$, temos $b(\alpha) \leq 2t$.
- Isso é verdade para qualquer ordenação de α com respeito a β, inclusive para a ótima.
- Nesse caso, $t = d(\alpha)$.
- Então

$$\frac{b(\alpha)}{2} \leq d(\alpha).$$

Diagrama realidade-desejo

- Um limite inferior melhor pode ser estabelecido a partir do grafo conhecido como *diagrama realidade-desejo*.
- Dadas duas permutações orientadas α e β, o diagrama realidade-desejo de α com relação a β, denotado $RD_\beta(\alpha)$, pode ser construído da seguinte forma:

Diagrama realidade-desejo

- Para cada elemento l de α, inclua dois vértices $l-$ e $l+$ em $RD_\beta(\alpha)$, um vértice para cada orientação distinta.
  - $l-$ corresponde à cauda da seta na orientação de l.
  - $l+$ corresponde à cabeça da seta na orientação de l.
- Inclua os vértices L e R em $RD_\beta(\alpha)$.
- Conecte os vértices dos elementos consecutivos de α de acordo com suas orientações (arestas realidade).
- Conecte L com o primeiro vértice e R com o último vértice.
- Conecte os vértices dos elementos de α de acordo com a permutação β (arestas desejo).
- Posicione todos os vértices em um círculo começando com L e posicione os outros vértices no sentido anti-horário.

Diagrama realidade-desejo

- Sobre um diagrama realidade-desejo temos:
  - As arestas realidade circulam o diagrama e as desejo cortam o diagrama.
  - Laços representam pontos onde o desejo e a realidade se encontram (pontos que não são de quebra).
  - O grau de cada vértice é dois, sendo incidentes uma aresta realidade e uma desejo.
  - Todo ciclo tem um número par de arestas, sendo metade delas realidade e metade delas desejo.

Diagrama realidade-desejo

- O número de ciclos de $RD_\beta(\alpha)$ é denotado por $c_\beta(\alpha)$.
- Observe que $c_\beta(\beta) = n + 1$ já que β não possui pontos de quebra. Disso, todas os seus ciclos correspondem a *loops*. Como existem $2n + 2$ nós, temos $n + 1$ ciclos.
- Então podemos pensar no processo de ordenação de uma permutação α por reversões como o processo de transformar $RD_\beta(\alpha)$ em um diagrama com número máximo de ciclos.

Como uma reversão afeta os ciclos em RD

- Sejam (s, t) e (u, v) duas arestas realidade que caracterizam uma reversão ρ, com (s, t) precedendo (u, v) na permutação α. $RD_\beta(\alpha)$ difere de $RD_\beta(\alpha\rho)$ nos seguintes aspectos:
  - As arestas realidade (s, t) e (u, v) são substituídas pelas arestas (s, u) e (t, v).
  - As arestas desejo continuam inalteradas.
  - A seção do círculo que vai do nó t ao nó u, incluindo suas extremidades, no sentido anti-horário, é revertida.

Convergência, divergência

- Sejam e e f duas arestas realidade pertencentes ao mesmo ciclo em $RD_\beta(\alpha)$. Assuma inicialmente que ambas estão orientadas no sentido anti-horário. Cada uma delas induz então uma orientação do ciclo em comum. Se essas orientações são as mesmas, dizemos que e e f *convergem*. senão, dizemos que e e f *divergem*.

Reversões sobre ciclos

- Seja ρ uma reversão atuando em duas arestas realidade e e f de $RD_\beta(\alpha)$. Então:
  - se e e f pertencem a ciclos distintos, $c(\alpha\rho) = c(\alpha) - 1$.
  - se e e f pertencem ao mesmo ciclo e convergem então $c(\alpha\rho) = c(\alpha)$.
  - se e e f pertencem ao mesmo ciclo e divergem então $c(\alpha\rho) = c(\alpha) + 1$.

Outro limite inferior

- Do resultado anterior podemos concluir que o número de ciclos se altera em no máximo uma unidade a cada reversão.
- Isso fornece outro limite inferior para a distância de reversão de duas permutações orientadas α e β.

Outro limite inferior

- Seja $\rho_1, \rho_2, \ldots, \rho_t$ uma série de reversões (não necessariamente ótima) que converte α em β. Ou seja

$$\alpha\rho_1\rho_2\ldots\rho_t = \beta.$$

Então

$$c(\alpha\rho_1\rho_2\ldots\rho_t) = c(\beta) = n + 1.$$

Outro limite inferior

- Sabemos que

$$\begin{aligned} c(\alpha\rho_1) - c(\alpha) &\leq 1, \\ c(\alpha\rho_1\rho_2) - c(\alpha\rho_1) &\leq 1, \\ &\vdots \\ c(\alpha\rho_1 \ldots \rho_t) - c(\alpha\rho_1 \ldots \rho_{t-1}) &\leq 1, \end{aligned}$$

Somando temos

$$n + 1 - c(\alpha) \leq t.$$

Isso é verdade para qualquer ordenação de α com respeito a β, inclusive para a ótima. Nesse caso, $t = d(\alpha)$.

Então

$$n + 1 - c(\alpha) \leq d(\alpha).$$

Ciclos bons e ruins

- Um ciclo pode ser classificado com base nos efeitos de uma reversão no número de ciclos de um diagrama:
  - Ciclo bom: se tem duas arestas realidade divergentes.
  - Ciclo ruim: se não tem duas arestas realidade divergentes.
- A classificação se aplica a ciclos com pelo menos quatro arestas chamados *ciclos próprios*.
- Se tivermos apenas ciclos bons no diagrama realidade-desejo de α, então o limite inferior é exato: $n + 1 - c(\alpha) = d(\alpha)$.

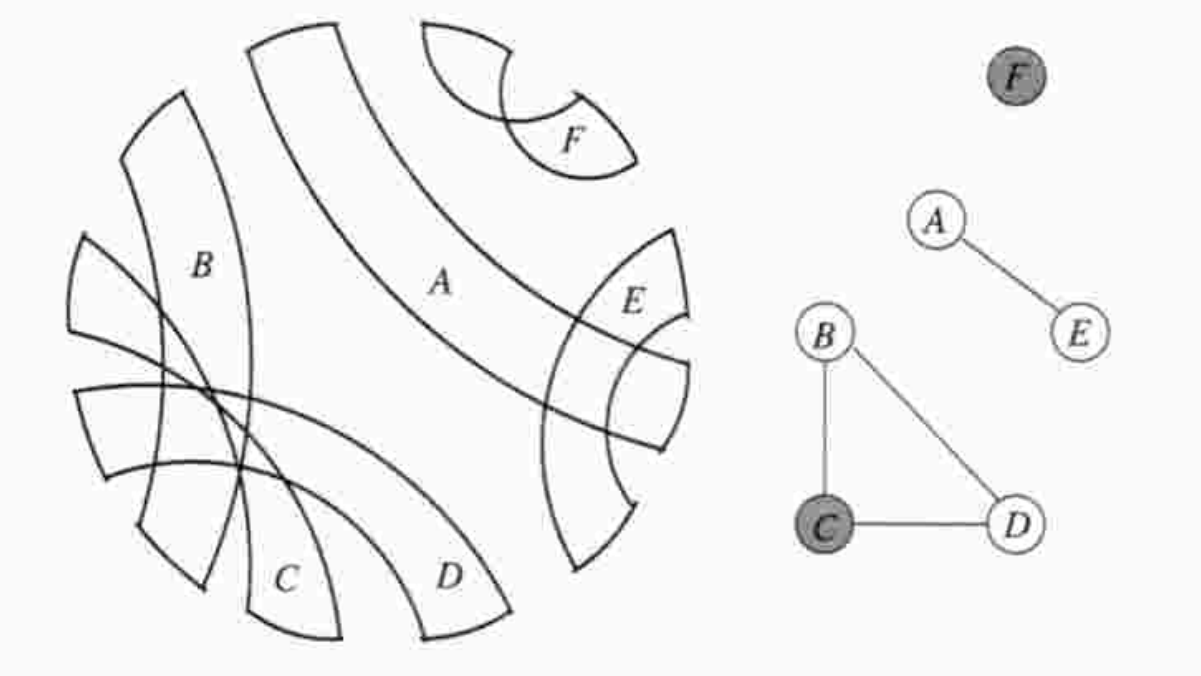
Ciclos bons e ruins

- Quando há ciclos ruins às vezes é possível chegar a β adicionando um ciclo por reversão: uma reversão que quebra um ciclo bom também pode dobrar um ciclo ruim transformando-o em um ciclo bom.
- Esse relacionamento é caracterizado no grafo de intercalação.

Grafo de intercalação

- Ciclos intercalados são ciclos em que alguma aresta desejo de um cruza uma aresta desejo do outro.
- O *grafo de intercalação* de α com relação a β, denotado por $I_\beta(\alpha)$ é tal que
  - os vértices são os ciclos próprios de $RD_\beta(\alpha)$ e
  - dois vértices são adjacentes se e somente se os ciclos correspondentes se intercalam.

Grafo de intercalação



Componentes bons e ruins

- Se um componente de um grafo de intercalação é composto inteiramente de ciclos ruins, ele é um *componente ruim*. Caso contrário, ele é um *componente bom*.
- Uma reversão definida por duas arestas divergentes de um mesmo ciclo é uma reversão ordenante se e somente se sua aplicação não cria componentes ruins no grafo.
- É fato que se houver um ciclo bom em um diagrama realidade-desejo então vai sempre existir uma reversão ordenante.

Idéia do algoritmo

- A idéia para um algoritmo que ordena permutações com sinal por reversões é:
  - Processe todos os componentes bons, escolhendo arestas realidade que induzem uma reversão ordenante.
  - Processe todos os componentes ruins em uma ordem específica, de acordo com certas características desses componentes.

Separação

- Um componente B *separa* dois componentes A e C se todas as arestas desejo entre A e C teriam que cruzar uma aresta de B.
- Uma reversão envolvendo arestas de A e C dobra B. Ao ser dobrado, um componente ruim se torna bom. Mas um componente bom pode se tornar ruim.

Obstáculos

- Um *obstáculo* é um componente ruim que não separa nenhum outro par de componentes ruins.
- Um *não-obstáculo* é um componente ruim que separa pelo menos um par de componentes ruins.
- O número de obstáculos em uma permutação α é $h(\alpha)$.

Obstáculos

- Dizemos que um obstáculo A *protege* um não-obstáculo B quando a remoção de A transforma B em um obstáculo.
- Um obstáculo A é um *super-obstáculo* se ele protege algum não-obstáculo B. Caso contrário, ele é chamado *obstáculo simples*.
- Uma permutação α é uma *fortaleza* quando o seu diagrama realidade-desejo contém um número ímpar de obstáculos e todos eles são super-obstáculos.

Fórmula para distância de reversão

- A fórmula exata para a distância de reversão de duas permutações orientadas é

$$d(\alpha) = n + 1 - c(\alpha) + h(\alpha) + f(\alpha),$$

onde $c(\alpha)$ corresponde ao número de ciclos no diagrama, $h(\alpha)$ ao número de obstáculos no diagrama e $f(\alpha)$ é 0 caso a permutação não seja uma fortaleza e 1 caso contrário.

Algoritmo

- O algoritmo de ordenação produz uma seqüência de reversões ordenantes.
- Quando não há ciclos bons, ele usa uma reversão em arestas convergentes ou uma reversão em arestas de ciclos diferentes.

Hurdle cutting

- Uma reversão em arestas convergentes não muda o número de ciclos.
- A melhor escolha é um obstáculo: isso transforma a componente ruim em boa sem aumentar o número de ciclos.
- (Escolher um não-obstáculo não muda o número de obstáculos ou fortaleza. Um super obstáculo também não ajuda porque o não-obstáculo que ele protege se torna obstáculo).

Hurdle cutting

- A reversão sobre arestas realidade de um ciclo de um obstáculo é chamada de hurdle cutting.
- Ela não muda $c(\alpha)$ e diminui $h(\alpha)$ quando o obstáculo é simples, mas para não aumentar $f(\alpha)$ ela só é empregada quando $h(\alpha)$ é par.

Hurdle merging

- Uma reversão em arestas de ciclos diferentes diminui o número de ciclos, o que é ruim, mas o número de obstáculos pode ser diminuído de 2.
- Para isso as arestas devem ser de ciclos em obstáculos diferentes. Essa reversão é chamada de hurdle merging.
- Os dois obstáculos e quaisquer não-obstáculo que os separam se tornam componentes bons.

Hurdle merging

- Existe a possibilidade dessa operação transformar um não obstáculo em obstáculo.
- Para evitar isso precisamos escolher obstáculos opostos.
- Dois obstáculos A e B são opostos se o número de obstáculos entre A e B no sentido horário é igual no sentido anti-horário. Só existem se $h(\alpha)$ é par.

Fortaleza

- Se $h(\alpha)$ é par, hurdle merging vai deixar $h(\alpha)$ par.
- Se $h(\alpha)$ é ímpar e existe obstáculo simples não pode haver obstáculo oposto. Então aplicamos hurddle cutting e $h(\alpha)$ será par. Se $h(\alpha)$ é ímpar e não existe obstáculo simples já temos uma fortaleza e $f(\alpha)$ não pode aumentar.
- Então essas operações não transforma a permutação em fortaleza.

Algoritmo

SORTING-REVERSAL(α, β)

```
1  if there is a good component in RD_β(α)
2        Pick two divergent edges e and f in this component
         making sure the corresponding reversal does not
         create any bad components
3        return the reversal characterized by e and f
4  else
5        if h(α) is even
6              return merging of two opposite hurdles
7        else
8              if h(α) is odd and there is a simple hurdle
9                    return a reversal cutting this hurdle
10             else // fortress
11                   return merging of any two hurdles
```

Componentes ruins no algoritmo

- A primeira operação diminui o número de ciclos em uma unidade, mas diminui o número de obstáculos em duas, tornando-os componentes bons e não transforma o diagrama em uma fortaleza.
- A segunda operação transforma o componente ruim em um componente bom, sem modificar o número de ciclos e não transforma o diagrama em uma fortaleza.
- A terceira operação diminui o número de ciclos em uma unidade, mas diminui o número de obstáculos em duas, tornando-os componentes bons.

Complexidade

- O grafo para representar o diagrama, os componentes e fortaleza podem ser identificados em tempo $O(n^2)$.
- São $O(n^2)$ reversões possíveis e cada uma pode ser verificada para a formação de componentes ruins em tempo $O(n^2)$. Então os componentes bons podem ser tratados em tempo $O(n^4)$.
- A função vai ser executada dba vezes e então o algoritmo é $O(n^5)$.

Permutação não-orientada

- Uma *permutação não-orientada* sobre um conjunto de n rótulos $\mathcal{L}$ é um mapeamento $\alpha : [1..n] \rightarrow \mathcal{L}$.
- A permutação identidade não-orientada é a permutação I tal que $I(i) = i$, para $1 \leq i \leq n$.

Reversão

- Uma *reversão* transforma uma permutação orientada em outra revertendo a ordem dos elementos de uma porção contínua da permutação.
- Uma reversão que envolve os elementos $\alpha(i)\,..\,\alpha(j)$ de α é denotada $[i\,..\,j]$.
- Uma reversão $\rho = [i\,..\,j]$ transforma uma permutação α em uma permutação $\alpha\rho$, definida da seguinte forma

$$\alpha[i\,..\,j](k) = \begin{cases} \alpha(i+j-k) & \text{se } i \leq k \leq j \\ \alpha(k) & \text{caso contrário} \end{cases}$$

Distância de reversão

- Dadas duas permutações não-orientadas α e β sobre o mesmo conjunto $\mathcal{L}$ de rótulos, o problema da ordenação por reversões é determinar o mínimo de reversões que transformam α em β.
- Ou seja, identificar uma série de reversões $\rho_1, \rho_2, \ldots, \rho_t$, com t mínimo, tal que

$$\alpha\rho_1\rho_2 \ldots \rho_t = \beta.$$

- O número t é a *distância de reversão* de α com relação a β e é denotada por $d_\beta(\alpha)$.

Permutação estendida

- Dada uma permutação α, a *versão estendida* dessa permutação é obtida adicionando-se a ela um rótulo artificial L antes do primeiro rótulo $\alpha(1)$ e um rótulo artificial R depois do último rótulo $\alpha(n)$.

Ponto de quebra

- Um *ponto de quebra* de α com respeito a β corresponde a um par x, y de elementos de $\mathcal{L}$ tal que x e y são consecutivos em α mas não são consecutivos em β.
- O número de pontos de quebra de uma permutação não-orientada α com respeito a β é denotado por $b_\beta(\alpha)$.
- Para toda permutação $\alpha \neq \beta$ temos $b_\beta(\alpha) \geq 2$.
- Uma reversão pode remover no máximo 2 pontos de quebra, então

$$d_\beta(\alpha) \geq \frac{b_\beta(\alpha)}{2}.$$

Tiras

- Uma seqüência consecutiva de rótulos delimitada por pontos-de-quebra é uma *tira*.
- Uma tira pode ser crescente ou decrescente.
- Tiras unitárias são crescentes e decrescentes.
- L e R formam uma única tira crescente LR.

Pontos-de-quebra e tiras

- Teorema: Se o rótulo k pertence a uma tira decrescente e $k-1$ pertence a uma tira crescente então existe uma reversão que remove pelo menos um breakpoint.
- Teorema: Se o rótulo k pertence a uma tira decrescente e $k+1$ pertence a uma tira crescente então existe uma reversão que remove pelo menos um breakpoint.
- Teorema: se o rótulo k pertence a uma tira decrescente e ou $k+1$ ou $k-1$ pertence a uma tira crescente então há uma reversão que remove pelo menos um ponto-de-quebra.
- Teorema: Seja α uma permutação com tira decrescente. Se todas as reversões que removem um ponto-de-quebra de α não deixam restar tiras decrescentes então existe uma reversão que remove 2 pontos de quebra de α.

Algoritmo

- O algoritmo de aproximação a seguir remove um ponto-de-quebra em média a cada iteração. Quando a primeira iteração não tem uma tira decrescente então a última iteração produz a identidade removendo 2 pontos-de-quebra, compensando a primeira.

Algoritmo

SORTING-REVERSAL(α, β)

```
1   list = ∅
2   while α ≠ 1
3       if a has a decreasing strip
4           k = the smallest label in a decreasing strip
5           ρ = the reversal that cuts after k and k − 1
6           if αρ has no decreasing strip
7               l = the largest label in a decreasing strip
8               ρ = the reversal that cuts before l and l + 1
9       else
10          ρ = the reversal that cuts the first two breakpoints
11          α = αρ
12      list = list + ρ
13  return list
```

Outras operações

Operação	Complex.	Melhor aproximação
reversões com sinal	P	1
reversões	NPD	1.375
transposições	NPD	1.375
reversões e transposições	aberto	2.8334
reversões com sinal e transposições	aberto	2
reversões de prefixo	NPD	2
reversões de prefixo com sinal	aberto	2
transposições de prefixo	aberto	2
transposições de prefixo e transposiçõs	aberto	$2+\varepsilon$