

Ordenação por Block-Interchanges e Reversões com Sinais

Cleber Mira João Meidanis

{cleber,meidanis}@ic.unicamp.br
Instituto de Computação
Universidade Estadual de Campinas

01/12/2005

Conteúdo

- 1 Introdução
 - Rearranjos em Genomas
 - Formalismo Algébrico

- 2 Ordenação

Conteúdo da Apresentação

- 1 **Introdução**
 - Rearranjos em Genomas
 - Formalismo Algébrico
- 2 Ordenação

EVOLUÇÃO EM GENOMAS.

Mutações Locais vs. Mutações Globais.

- Mutações Locais → alterações em bases (edit distance).
- Mutações Globais → Rearranjos e Recombinações.

Foco desse trabalho: Rearranjos em Genomas.

MOTIVAÇÃO E APLICAÇÕES.

- **Análise comparativa da evolução das espécies**
- Análise de rearranjos pode ser mais efetiva na construção de hipóteses evolutivas.
- Reversões com sinais e Block-interchanges parece ser eventos de rearranjo freqüentes em cenários evolutivos parcimoniosos.
- Problemas combinatórios interessantes (novas estruturas de dados, algoritmos e etc.).

MOTIVAÇÃO E APLICAÇÕES.

- Análise comparativa da evolução das espécies
- Análise de rearranjos pode ser mais efetiva na construção de hipóteses evolutivas.
- Reversões com sinais e Block-interchanges parece ser eventos de rearranjo freqüentes em cenários evolutivos parcimoniosos.
- Problemas combinatórios interessantes (novas estruturas de dados, algoritmos e etc.).

MOTIVAÇÃO E APLICAÇÕES.

- Análise comparativa da evolução das espécies
- Análise de rearranjos pode ser mais efetiva na construção de hipóteses evolutivas.
- Reversões com sinais e Block-interchanges parece ser eventos de rearranjo freqüentes em cenários evolutivos parcimoniosos.
- Problemas combinatórios interessantes (novas estruturas de dados, algoritmos e etc.).

MOTIVAÇÃO E APLICAÇÕES.

- Análise comparativa da evolução das espécies
- Análise de rearranjos pode ser mais efetiva na construção de hipóteses evolutivas.
- Reversões com sinais e Block-interchanges parece ser eventos de rearranjo freqüentes em cenários evolutivos parcimoniosos.
- Problemas combinatórios interessantes (novas estruturas de dados, algoritmos e etc.).

GENOMA.

- *Genoma*: lista ordenada de blocos de genes.
- Genoma Linear: Lista linear \leftrightarrow Genoma Circular: Lista circular
- Genomas são geralmente modelados por meio de funções.

Exemplo: $\pi = x_1, \dots, x_n$ é representado como

$$\pi = [x_1, \dots, x_n]$$

onde π passou a ser uma função tal que $\pi(i) = x_i$ para $1 \leq i \leq n$.

REVERSÕES COM SINAIS.

$$\begin{array}{cccccc} +1 & \underline{-5} & +4 & \underline{-3} & \underline{+2} & -6 \\ +1 & -2 & +3 & -4 & +5 & -6 \end{array}$$

Figure: Reversão com sinais aplicada sobre o genoma $\pi = [1, -5, 4, -3, 2, -6]$. Os blocos sublinhados (-5 a 2) são afetados pela reversão: a seqüência de blocos é invertida e a orientação de cada bloco é trocada.

BLOCK-INTERCHANGES.

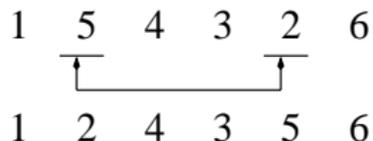


Figure: Transposição aplicada sobre o genoma $\pi = [1, 5, 4, 3, 2, 6]$. As seqüências de blocos sublinhados (5 e 2) são movidas de sua posição original em π : suas posições são trocadas.

PROBLEMAS DE REARRANJO EM GENOMAS.

O *Problema de Rearranjo em Genomas* consiste em encontrar uma seqüência mínima de determinados eventos de mutação (reversões, transposições, reversões com sinais e etc) que transforme um genoma em outro.

Conteúdo da Apresentação

- 1 **Introdução**
 - Rearranjos em Genomas
 - **Formalismo Algébrico**
- 2 Ordenação

FORMALISMO ALGÉBRICO.

- *Permutação*: função bijetora π sobre um conjunto E .
- *Identidade*: permutação ι tal que $\iota(x) = x$ para qualquer $x \in E$.
- *Órbita*: $orb(\pi, x) = \{y \in E \mid y = \pi^k x \text{ para um inteiro } k\}$.
- $Orb(\pi, E)$: conjunto das órbitas da permutação π sobre E .
- *Ciclo*: permutação com no máximo uma órbita não trivial.
- *k-ciclo*: ciclo cuja órbita não trivial tem k elementos. Obs.: A identidade é um 1-ciclo.

FORMALISMO ALGÉBRICO.

- *Permutação*: função bijetora π sobre um conjunto E .
- *Identidade*: permutação ι tal que $\iota(x) = x$ para qualquer $x \in E$.
- *Órbita*: $orb(\pi, x) = \{y \in E \mid y = \pi^k x \text{ para um inteiro } k\}$.
- $Orb(\pi, E)$: conjunto das órbitas da permutação π sobre E .
- *Ciclo*: permutação com no máximo uma órbita não trivial.
- *k-ciclo*: ciclo cuja órbita não trivial tem k elementos. Obs.: A identidade é um 1-ciclo.

FORMALISMO ALGÉBRICO.

- *Permutação*: função bijetora π sobre um conjunto E .
- *Identidade*: permutação ι tal que $\iota(x) = x$ para qualquer $x \in E$.
- *Órbita*: $orb(\pi, x) = \{y \in E \mid y = \pi^k x \text{ para um inteiro } k\}$.
- $Orb(\pi, E)$: conjunto das órbitas da permutação π sobre E .
- *Ciclo*: permutação com no máximo uma órbita não trivial.
- *k-ciclo*: ciclo cuja órbita não trivial tem k elementos. Obs.: A identidade é um 1-ciclo.

FORMALISMO ALGÉBRICO.

- *Decomposição em k -ciclos de uma permutação:* representação de uma permutação π como um produto de k -ciclos não necessariamente disjuntos.

Theorem

Qualquer permutação π sobre um conjunto E pode ser representada como um produto de ciclos disjuntos.

- *Norma de π :* É o número mínimo de 2-ciclos cujo produto é π . Notação: $\|\pi\|$.
- *Divisibilidade:* Uma permutação α *divide* uma permutação β (ambas sobre E), denotado por $\alpha|\beta$, quando $\|\beta\alpha^{-1}\| = \|\beta\| - \|\alpha\|$.

PRODUTO DE PERMUTAÇÕES.

Dados: $E = \{-3, -2, -1, 1, 2, 3\}$,
 $\alpha = (-3\ 2\ 1)(3\ -2)$ e $\beta = (-1\ 3\ -3\ 2)$
temos:

$\rightarrow \alpha\beta x = \alpha(\beta(x))$ para todo $x \in E$.

Ex.: $\alpha\beta -3 = \alpha 2 = 1$.

$$\alpha\beta = (1\ -3)(2\ -1\ -2\ 3)$$

MODELANDO O GENOMA.

Γ : Permutação que associa um bloco de genes em uma fita do genoma ao bloco reverso na fita complementar.

$$E_+ = \{x_1, x_2, \dots, x_n\}.$$

$$E_- = \{\Gamma x_1, \Gamma x_2, \dots, \Gamma x_n\}.$$

$$E = E_+ \cup E_-.$$

Genoma Unicromossomal (Cromossomo):

$$\pi = (x_1 \ x_2 \ \dots \ x_n)(\Gamma x_n \ \dots \ \Gamma x_2 \ \Gamma x_1)$$

Genoma: Produto de cromossomos.

Propriedade que define o genoma: $\Gamma \pi \Gamma = \pi^{-1}$.

EXEMPLO DE GENOMAS.

- Genoma do vírus da imunodeficiência humana 2:
(*gag-pol HIV2gp2 HIV2gp3 HIV2gp4*
HIV2gp5 HIV2gp6 HIV2gp7 HIV2gp8)
- Hepatitis B virus:
(*HBVgp1 HBVgp2 HBVgp3 HBVgp4*)
(Γ *HBVgp4* Γ *HBVgp3* Γ *HBVgp2* Γ *HBVgp1*)

Fonte: GeneBank (www.ncbi.nlm.nih.gov)

MODELANDO EVENTOS DE REARRANJO.

Reversão com sinal

- 1 Produto de dois 2-ciclos: $(u \pi \Gamma v)(v \pi \Gamma u)$;
- 2 $(u v)|\pi$.

A reversão com sinal é chamada de *aplicável* ao genoma π .
Efeito da reversão com sinal:

$$\begin{aligned}(u \pi \Gamma v)(v \pi \Gamma u)(\dots u \dots v \dots)(\dots \Gamma v \dots \Gamma u \dots) = \\ (\dots \pi^{-1} u \pi \Gamma v \dots \Gamma \pi u \Gamma u v \pi v \dots) \\ (\dots \Gamma \pi v \Gamma v u \pi u \dots \pi^{-1} v \pi \Gamma u \dots)\end{aligned}$$

MODELANDO EVENTOS DE REARRANJO.

Block-Interchange

Formato: $\rho = (u\ x)(\pi\Gamma\ x\ \pi\Gamma\ u)(v\ y)(\pi\Gamma\ y\ \pi\Gamma\ v)$; tal que

- 1 $u \neq x, u \neq y$ e $v \neq y$,
- 2 $v, x, y \in \text{orb}(\pi, u)$,
- 3 $(u\ v\ x\ y)|\pi$.

A transposição é chamada de *aplicável* ao genoma π .

Efeito do block-interchange:

$$\begin{aligned} &\rho(\dots u \dots v \dots x \dots y \dots)(\dots \Gamma y \dots \Gamma x \dots \Gamma v \dots \Gamma u \dots) = \\ &(\dots \pi^{-1}u\ x \dots \pi^{-1}y\ v \dots \pi^{-1}x\ u \dots \pi^{-1}v\ y \dots) \\ &(\dots \Gamma y\ \pi\Gamma v \dots \Gamma u\ \pi\Gamma x \dots \Gamma v\ \pi\Gamma y \dots \Gamma x\ \pi\Gamma u \dots) \end{aligned}$$

REARRANJO POR REVERSÕES COM SINAIS E BLOCK-INTERCHANGES.

- O peso de um evento de rearranjo ρ é $w(\rho) = \|\rho\|/2$.
- O problema de rearranjo por reversões com sinais e block-interchanges consiste em encontrar, dados os genomas π e σ e a função Γ , todos sobre E , uma seqüência com peso mínimo de reversões com sinal e block-interchanges ρ_1, \dots, ρ_k tal que $\rho_k \dots \rho_1 \pi = \sigma$, o evento de rearranjo ρ_i é aplicável a $\rho_{i-1} \dots \rho_1 \pi$ para $1 \leq i \leq k$ e $\sum_{i=1}^k w(\rho_i)$ é mínimo. O valor mínimo de $\sum_{i=1}^k w(\rho_i)$ será denotado por $W(\pi, \sigma)$.

REARRANJO POR REVERSÕES COM SINAIS E BLOCK-INTERCHANGES.

- O peso de um evento de rearranjo ρ é $w(\rho) = \|\rho\|/2$.
- O problema de rearranjo por reversões com sinais e block-interchanges consiste em encontrar, dados os genomas π e σ e a função Γ , todos sobre E , uma seqüência com peso mínimo de reversões com sinal e block-interchanges ρ_1, \dots, ρ_k tal que $\rho_k \dots \rho_1 \pi = \sigma$, o evento de rearranjo ρ_i é aplicável a $\rho_{i-1} \dots \rho_1 \pi$ para $1 \leq i \leq k$ e $\sum_{i=1}^k w(\rho_i)$ é mínimo. O valor mínimo de $\sum_{i=1}^k w(\rho_i)$ será denotado por $W(\pi, \sigma)$.

EVENTOS BONS E GENOMAS EQUIORBITAIS.

- *Par*: Em sua representação como um produto de ciclos disjuntos, é possível distinguir fatores que obedecem ao formato $\alpha\pi\Gamma\alpha^{-1}(\pi\Gamma)^{-1}$. Esses fatores são chamados *pares*.
- Número de pares de $\sigma\pi^{-1}$ é denotado por $c(\pi, \sigma)$.
- $c(\rho\pi, \sigma) - c(\pi, \sigma) = \Delta c(\rho, \pi, \sigma)$ onde ρ é aplicável a π .
- *Bom Evento*: Um evento de rearranjo ρ é chamado um *bom evento* para (π, σ) quando $\Delta c(\rho, \pi, \sigma) = w(\rho)$.
- *Genomas Equiorbitais*: Dados os genomas π e σ sobre E , eles são *genomas equiorbitais* quando $\text{Orb}(\pi, E) = \text{Orb}(\sigma, E)$.

EXEMPLO.

- $E = \{-3, -2, -1, 1, 2, 3\}$,
- $\Gamma = (1 \ -1)(2 \ -2)(3 \ -3)$,
- $\pi = (-2 \ 1 \ 3)(-3 \ -1 \ 2)$,
- $\sigma = (1 \ 2 \ 3)(-3 \ -2 \ -1)$,
- $\sigma\pi^{-1} = (1 \ -1 \ -2)(-3 \ 3 \ 2)$.
- $\rho = (-2 \ -1)(3 \ -3)$

VARIAÇÃO NO NÚMERO DE PARES.

Dados π , σ e Γ sobre E :

- Se ρ é uma reversão com sinais aplicável a π , então $\Delta c(\rho, \pi, \sigma) \in \{-1, 0, 1\}$.
- Se ρ é um block-interchange aplicável a π então $\Delta c(\rho, \pi, \sigma) \in \{-2, 0, 2\}$.
- Como $w(\rho) = 1$ se ρ é uma reversão com sinais e $w(\rho) = 2$ se ρ é um block-interchange, então *bons eventos* maximizam a variação do número de pares e portanto são bons candidatos a participar de uma seqüência de eventos de rearranjo que transforme π em σ .

GENOMAS EQUIORBITAIS.

Theorem

Dados os genomas equiorbitais π e σ , a função Γ , todos sobre E , então existe um block-interchange ρ aplicável a π tal que ρ é um bom evento para (π, σ) e os genomas $\rho\pi$ e σ são equiorbitais.

GENOMAS NÃO EQUIORBITAIS.

Theorem

Dados os genomas não equiorbitais π e σ , a função Γ , todos sobre E , então existe uma reversão com sinais ρ aplicável a π tal que ρ é um bom evento para (π, σ) .

ALGORITMO.

- 1 Faça $r = 0$, $\theta = \pi$ e $W = 0$.
- 2 Enquanto $\theta \neq \sigma$ faça
 - Incremente r .
 - se θ e σ são genomas equiorbitais então encontre um block-interchange ρ_r aplicável a θ e que seja um bom evento para (θ, σ) .
 - senão encontre uma reversão com sinais ρ_r aplicável a θ e que seja um bom evento para (θ, σ) .
 - Faça $\theta = \rho_r \theta$ e $W = W + w(\rho_r)$.
- 3 Devolva ρ_1, \dots, ρ_r e W .

COMPLEXIDADE.

Seja $n = |E|/2$.

- Determinar se $\theta = \sigma$: $O(n)$;
- Determinar se θ e σ são equiorbitais: $O(n)$;
- Encontrar block-interchange que seja bom evento: $O(n)$;
- Encontrar reversão com sinais que seja bom evento: $O(n)$;
- Aplicar evento: $O(n)$;
- O loop principal é executado no máximo n vezes.

A complexidade geral do algoritmo é $O(n^2)$.

Conclusão

- O **Formalismo Algébrico** representa adequadamente rearranjos em genomas envolvendo reversões com sinais e block-interchanges.
- A **Medida de Rerranjos** baseada em reversões com sinais e block-interchanges pode ser obtida através de um algoritmo polinomial.
- A **Norma** de um evento de rearranjo mostrou-se um parâmetro adequado adequado na atribuição de pesos a eventos de rearranjo.
- Problemas a resolver:
 - Verificar se é possível obter $W(\pi, \sigma)$ em tempo subquadrático.
 - Incluir o evento de translocação e tratar genomas multicromossomais.