MO640/MC668

Guilherme P. Telles

IC-Unicamp

Avisado está

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

Parte I

Montagem de fragmentos de DNA

Seqüenciamento de DNA

- Conhecer as cadeias do DNA permite muitas análises biológicas.

Seqüenciamento de DNA

- Conhecer as cadeias do DNA permite muitas análises biológicas.
- Seqüenciamento de DNA são técnicas que permitem obter a seqüenciar de moléculas de DNA.

Seqüenciamento de DNA

- Conhecer as cadeias do DNA permite muitas análises biológicas.
- Seqüenciamento de DNA são técnicas que permitem obter a seqüenciar de moléculas de DNA.
- A técnica mais antiga é de Sanger, 1977.

Seqüenciamento de DNA

- A limitação mais importante é que só é possível obter a cadeia de trechos de DNA de até 1000 bases.

Seqüenciamento de DNA

- A limitação mais importante é que só é possível obter a cadeia de trechos de DNA de até 1000 bases.
- Para seqüenciar uma molécula longa de DNA ela é replicada e as cópias são quebradas em fragmentos, que são seqüenciados.

Seqüenciamento de DNA

- A limitação mais importante é que só é possível obter a cadeia de trechos de DNA de até 1000 bases.
- Para seqüenciar uma molécula longa de DNA ela é replicada e as cópias são quebradas em fragmentos, que são seqüenciados.
- O problema computacional é reconstruir a molécula original com base nas sobreposições entre os fragmentos.

Seqüenciamento de DNA

- A limitação mais importante é que só é possível obter a cadeia de trechos de DNA de até 1000 bases.
- Para seqüenciar uma molécula longa de DNA ela é replicada e as cópias são quebradas em fragmentos, que são seqüenciados.
- O problema computacional é reconstruir a molécula original com base nas sobreposições entre os fragmentos.
- Fragmentos seqüenciados também são chamados de *reads*.

Seqüenciamento de DNA

- A limitação mais importante é que só é possível obter a cadeia de trechos de DNA de até 1000 bases.
- Para seqüenciar uma molécula longa de DNA ela é replicada e as cópias são quebradas em fragmentos, que são seqüenciados.
- O problema computacional é reconstruir a molécula original com base nas sobreposições entre os fragmentos.
- Fragmentos seqüenciados também são chamados de *reads*.
- Podemos ter ou não uma estimativa do tamanho da molécula de DNA original.

Seqüenciamento de DNA

- A limitação mais importante é que só é possível obter a cadeia de trechos de DNA de até 1000 bases.
- Para seqüenciar uma molécula longa de DNA ela é replicada e as cópias são quebradas em fragmentos, que são seqüenciados.
- O problema computacional é reconstruir a molécula original com base nas sobreposições entre os fragmentos.
- Fragmentos seqüenciados também são chamados de *reads*.
- Podemos ter ou não uma estimativa do tamanho da molécula de DNA original.
- Pode haver um read para cada ponta de um fragmento de DNA, do qual se tem uma estimativa de tamanho, formando um par-de-reads.

Tecnologias de seqüenciamento (mais usadas)

- Tecnologias mais recentes (pirosseqüenciamento) permitem produzir dezenas de milhares de cadeias por dia, a um custo mais baixo que na tecnologia anterior (Sanger).

Tecnologias de seqüenciamento (mais usadas)

- Tecnologias mais recentes (pirosseqüenciamento) permitem produzir dezenas de milhares de cadeias por dia, a um custo mais baixo que na tecnologia anterior (Sanger).
- Exigem muito menos intervenção humana.

Tecnologias de seqüenciamento (mais usadas)

- Tecnologias mais recentes (pirosseqüenciamento) permitem produzir dezenas de milhares de cadeias por dia, a um custo mais baixo que na tecnologia anterior (Sanger).
- Exigem muito menos intervenção humana.
- Números (2011):

Tecnologias de seqüenciamento (mais usadas)

- Tecnologias mais recentes (pirosseqüenciamento) permitem produzir dezenas de milhares de cadeias por dia, a um custo mais baixo que na tecnologia anterior (Sanger).
- Exigem muito menos intervenção humana.
- Números (2011):
  - Sanger: 60 kbp a cada 7h. Reads de até 1000 bp.

Tecnologias de seqüenciamento (mais usadas)

- Tecnologias mais recentes (pirosseqüenciamento) permitem produzir dezenas de milhares de cadeias por dia, a um custo mais baixo que na tecnologia anterior (Sanger).
- Exigem muito menos intervenção humana.
- Números (2011):
  - Sanger: 60 kbp a cada 7h. Reads de até 1000 bp.
  - 454: 1 Gbp a cada 24h. Reads de até 1000 bp.

Tecnologias de seqüenciamento (mais usadas)

- Tecnologias mais recentes (pirosseqüenciamento) permitem produzir dezenas de milhares de cadeias por dia, a um custo mais baixo que na tecnologia anterior (Sanger).
- Exigem muito menos intervenção humana.
- Números (2011):
  - Sanger: 60 kbp a cada 7h. Reads de até 1000 bp.
  - 454: 1 Gbp a cada 24h. Reads de até 1000 bp.
  - Illumina: 1.8 Gbp a cada 4 dias. Reads de até 300 bp.

Tecnologias de seqüenciamento (mais usadas)

- Tecnologias mais recentes (pirosseqüenciamento) permitem produzir dezenas de milhares de cadeias por dia, a um custo mais baixo que na tecnologia anterior (Sanger).
- Exigem muito menos intervenção humana.
- Números (2011):
  - Sanger: 60 kbp a cada 7h. Reads de até 1000 bp.
  - 454: 1 Gbp a cada 24h. Reads de até 1000 bp.
  - Illumina: 1.8 Gbp a cada 4 dias. Reads de até 300 bp.
  - SOLiD: 120 Gbp a cada 10 dias. Reads de 50 bp.

Tecnologias de seqüenciamento (mais usadas)

- Tecnologias mais recentes (pirosseqüenciamento) permitem produzir dezenas de milhares de cadeias por dia, a um custo mais baixo que na tecnologia anterior (Sanger).
- Exigem muito menos intervenção humana.
- Números (2011):
  - Sanger: 60 kbp a cada 7h. Reads de até 1000 bp.
  - 454: 1 Gbp a cada 24h. Reads de até 1000 bp.
  - Illumina: 1.8 Gbp a cada 4 dias. Reads de até 300 bp.
  - SOLiD: 120 Gbp a cada 10 dias. Reads de 50 bp.
- Além da cadeia de DNA, produzem uma indicação da precisão da leitura de cada letra, chamada de qualidade.

Tecnologias de seqüenciamento (mais usadas)

- Tecnologias mais recentes (pirosseqüenciamento) permitem produzir dezenas de milhares de cadeias por dia, a um custo mais baixo que na tecnologia anterior (Sanger).
- Exigem muito menos intervenção humana.
- Números (2011):
  - Sanger: 60 kbp a cada 7h. Reads de até 1000 bp.
  - 454: 1 Gbp a cada 24h. Reads de até 1000 bp.
  - Illumina: 1.8 Gbp a cada 4 dias. Reads de até 300 bp.
  - SOLiD: 120 Gbp a cada 10 dias. Reads de 50 bp.
- Além da cadeia de DNA, produzem uma indicação da precisão da leitura de cada letra, chamada de qualidade.
- Estão evoluindo rapidamente.

Estratégias principais

- Shotgun total: quebrar o DNA inteiro em fragmentos que podem ser seqüenciados diretamente.

Estratégias principais

- Shotgun total: quebrar o DNA inteiro em fragmentos que podem ser seqüenciados diretamente.
- Shotgun hierárquico: quebrar o DNA em pedaços ainda grandes (150kbp, BACs), selecionar um conjunto deles que cobre a molécula original e seqüenciar cada pedaço por shotgun total.

Estratégias principais

- Shotgun total: quebrar o DNA inteiro em fragmentos que podem ser seqüenciados diretamente.
- Shotgun hierárquico: quebrar o DNA em pedaços ainda grandes (150kbp, BACs), selecionar um conjunto deles que cobre a molécula original e seqüenciar cada pedaço por shotgun total.
- Deve haver redundância no seqüenciamento para que haja alta probabilidade de que toda a molécula seja coberta.

Problema computacional

- É conhecido como problema da montagem de fragmentos de DNA.

Problema computacional

- É conhecido como problema da montagem de fragmentos de DNA.
- Consiste em reconstruir a molécula original com base nas sobreposições entre os fragmentos.

Problema computacional

- É conhecido como problema da montagem de fragmentos de DNA.
- Consiste em reconstruir a molécula original com base nas sobreposições entre os fragmentos.
- A montagem produz uma organização das cadeias dos fragmentos, similar a um alinhamento múltiplo.

Problema computacional

- É conhecido como problema da montagem de fragmentos de DNA.
- Consiste em reconstruir a molécula original com base nas sobreposições entre os fragmentos.
- A montagem produz uma organização das cadeias dos fragmentos, similar a um alinhamento múltiplo.
- A resposta para o problema é uma *cadeia-consenso* ou apenas *consenso*, obtida por uma votação da maioria das letras em uma coluna.

Porque é difícil

- Os genomas são grandes.

Porque é difícil

- Os genomas são grandes.
- Os genomas têm repetições.

Porque é difícil

- Os genomas são grandes.
- Os genomas têm repetições.
- O seqüenciamento não é perfeito.

Porque é difícil

- Os genomas são grandes.
- Os genomas têm repetições.
- O seqüenciamento não é perfeito.
- A quantidade de reads é muito grande.

Experimento de shotgun ideal

- Todas as regiões da molécula de interesse estão cobertas por fragmentos.

Experimento de shotgun ideal

- Todas as regiões da molécula de interesse estão cobertas por fragmentos.
- Todos os fragmentos são seqüenciados com exatidão.

Experimento de shotgun ideal

- Todas as regiões da molécula de interesse estão cobertas por fragmentos.
- Todos os fragmentos são seqüenciados com exatidão.
- Cada fragmento tem sobreposição com outros.

Experimento de shotgun ideal

- Todas as regiões da molécula de interesse estão cobertas por fragmentos.
- Todos os fragmentos são seqüenciados com exatidão.
- Cada fragmento tem sobreposição com outros.
- Temos uma

Erros de seqüenciamento

- O seqüenciamento de de cada fragmento pode não ser exato.

Erros de seqüenciamento

- O seqüenciamento de de cada fragmento pode não ser exato.
- Cada tecnologia é susceptível a certos tipos de erros, que podem ser inserções, remoções e substituições de bases.

Erros de seqüenciamento

- O seqüenciamento de de cada fragmento pode não ser exato.
- Cada tecnologia é susceptível a certos tipos de erros, que podem ser inserções, remoções e substituições de bases.
- Para Sanger, de 1 a 5% concentrados nas extremidades, principalmente $3'$.

Erros de seqüenciamento

- O seqüenciamento de de cada fragmento pode não ser exato.
- Cada tecnologia é susceptível a certos tipos de erros, que podem ser inserções, remoções e substituições de bases.
- Para Sanger, de 1 a 5% concentrados nas extremidades, principalmente $3'$.
- Para piroseqüenciamento, menor que 1% (pesquisa preliminar).

Quimeras

- Partes não consecutivas do DNA se juntam e formam um fragmento falso.

Falta de cobertura

- Uma ou mais regiões do DNA não são cobertas por nenhum fragmento.

Falta de cobertura

- Uma ou mais regiões do DNA não são cobertas por nenhum fragmento.
- Acontece por limitações da técnica de seqüenciamento para um trecho da molécula com uma composição específica de bases.

Repetições

- É comum que existam regiões que se repetem exatamente ou com grande similaridade.

Repetições

- É comum que existam regiões que se repetem exatamente ou com grande similaridade.
- Se a similaridade é grande as diferenças são tomadas por erros de seqüenciamento.

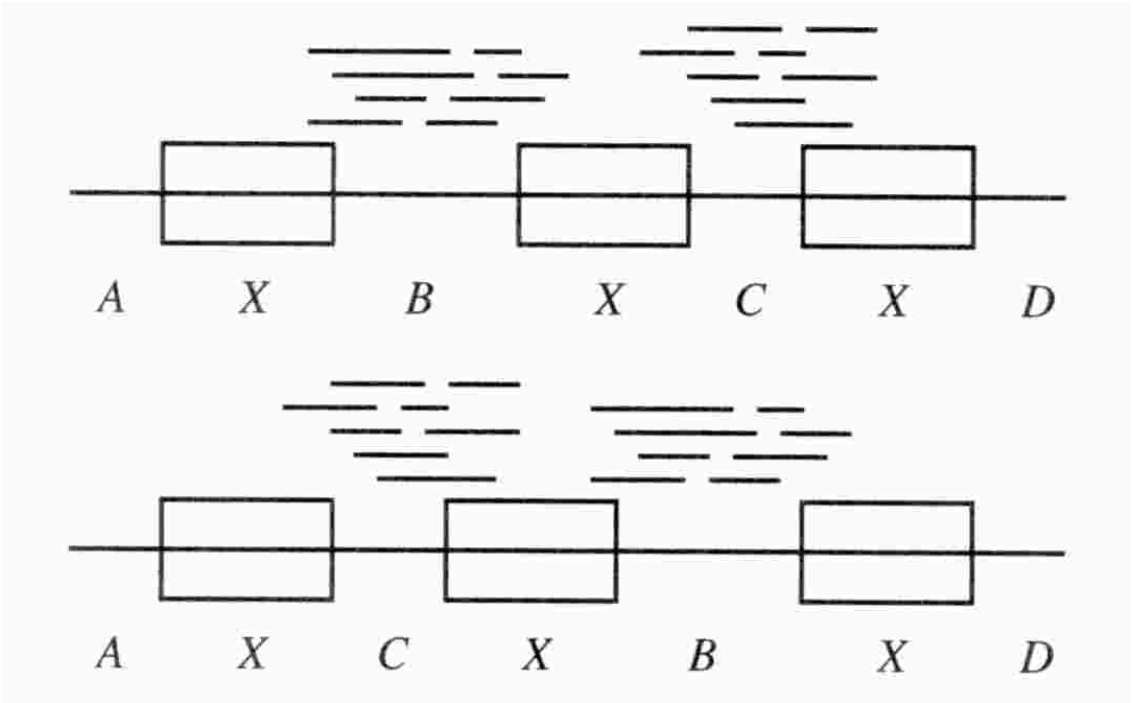
Repetições

- É comum que existam regiões que se repetem exatamente ou com grande similaridade.
- Se a similaridade é grande as diferenças são tomadas por erros de seqüenciamento.
- Repetições cobertas totalmente por um read não são um problema.

Repetições

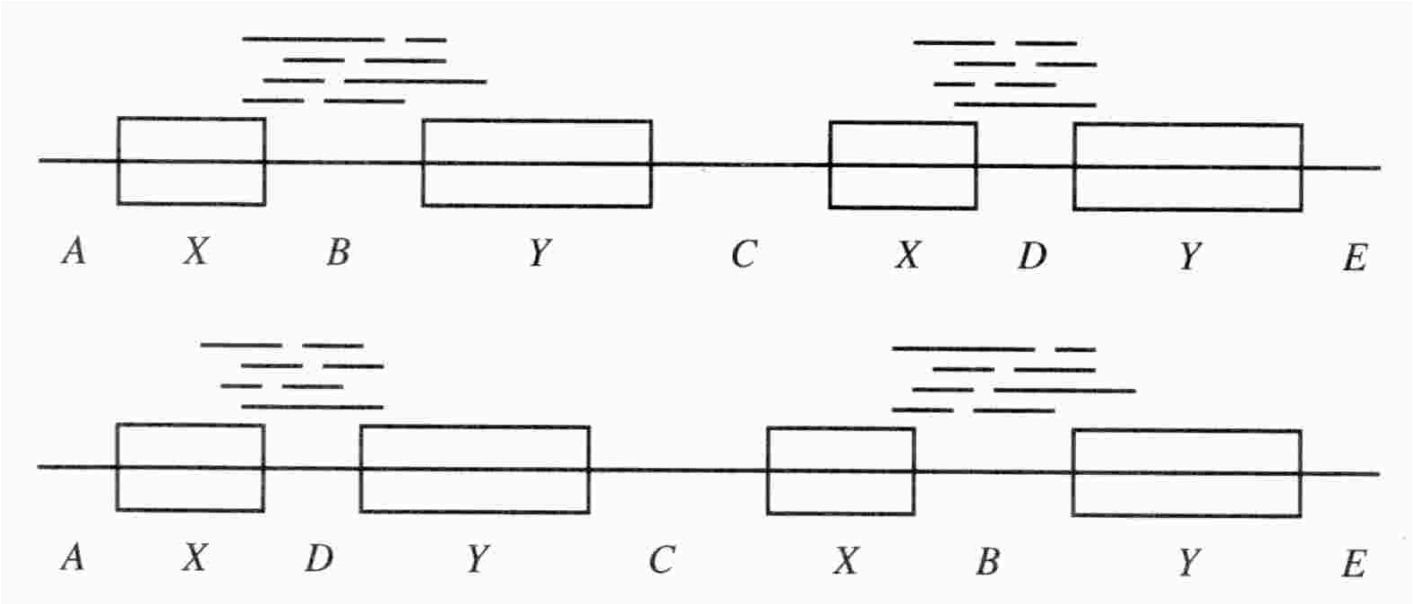
- É comum que existam regiões que se repetem exatamente ou com grande similaridade.
- Se a similaridade é grande as diferenças são tomadas por erros de seqüenciamento.
- Repetições cobertas totalmente por um read não são um problema.
- Repetições podem ser grandes ou pequenas com um grande número de cópias.

Repetições



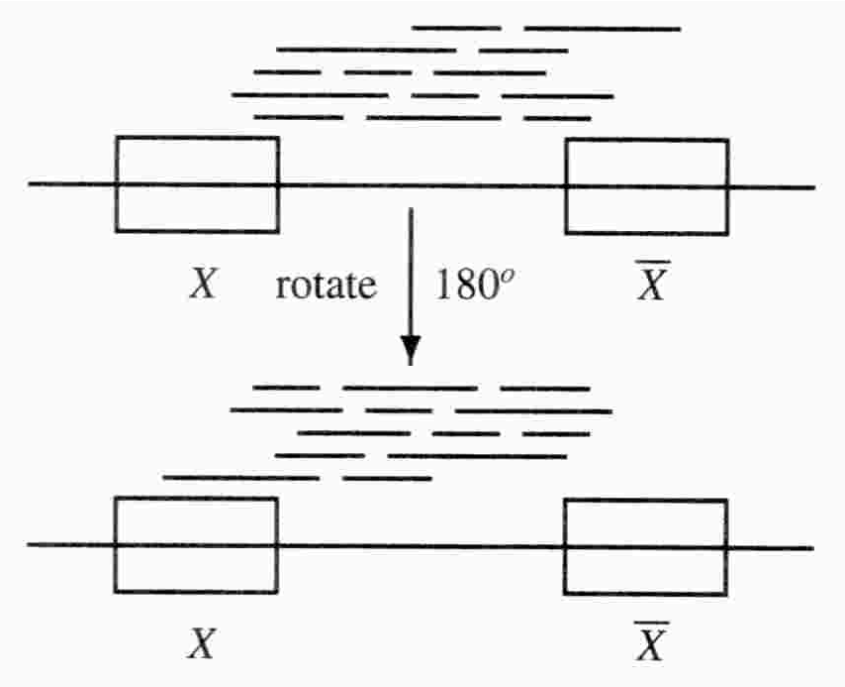
[Setubal e Meidanis, 1996]

Repetições



[Setubal e Meidanis, 1996]

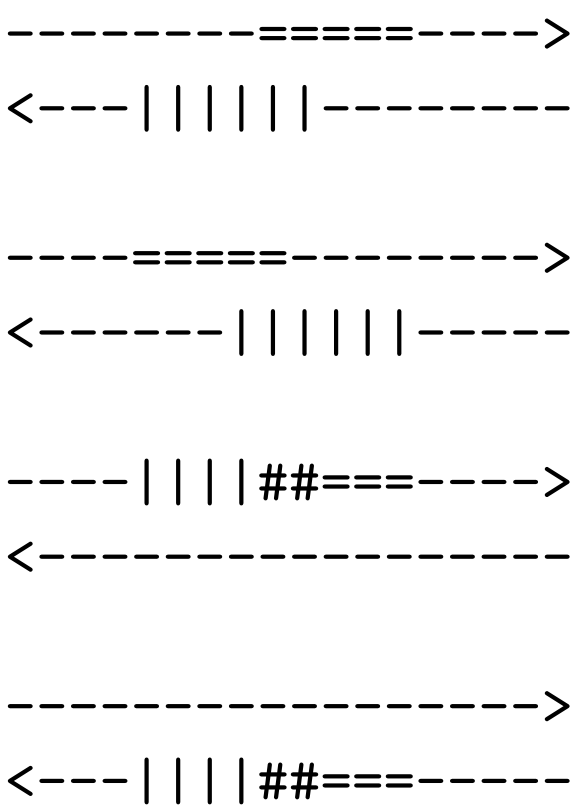
Repetições



[Setubal e Meidanis, 1996]

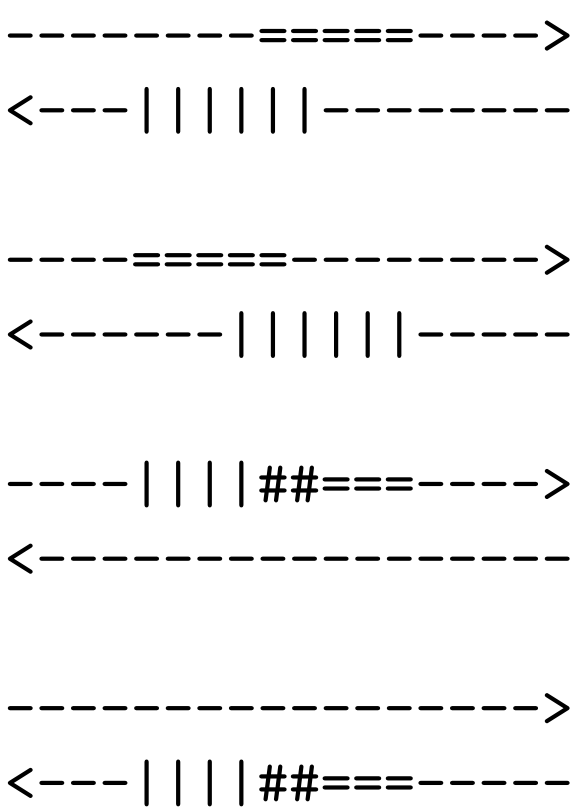
Orientação desconhecida

- Os fragmentos são seqüenciados de $5'$ para $3'$ mas não sabemos de que fita eles vieram.



Orientação desconhecida

- Os fragmentos são seqüenciados de $5'$ para $3'$ mas não sabemos de que fita eles vieram.



- Temos uma indicação da qualidade de cada base.

Modelos

- Vamos olhar para algumas formas em que o problema já foi formulado.

Modelos

- Vamos olhar para algumas formas em que o problema já foi formulado.
- Na maioria dos modelos vamos considerar montagem mais parcimoniosa, isto é, que a molécula mais curta é a melhor resposta.

Cobertura e contigs

- Para n reads com tamanho médio t originados de uma cadeia de tamanho T e montados com sobreposição mínima o, a cobertura média é

$$c = \frac{nt}{T}.$$

Cobertura e contigs

- Para n reads com tamanho médio t originados de uma cadeia de tamanho T e montados com sobreposição mínima o, a cobertura média é

$$c = \frac{nt}{T}.$$

- A montagem pode resultar em mais de um subconjunto de cadeias. Cada subconjunto é chamado de contig. O número esperado de contigs é

$$p = ne^{\frac{-n(t-o)}{T}}.$$

Cobertura e contigs

- Para n reads com tamanho médio t originados de uma cadeia de tamanho T e montados com sobreposição mínima o, a cobertura média é

$$c = \frac{nt}{T}.$$

- A montagem pode resultar em mais de um subconjunto de cadeias. Cada subconjunto é chamado de contig. O número esperado de contigs é

$$p = ne^{\frac{-n(t-o)}{T}}.$$

- O número esperado de contigs com pelo menos dois reads é

$$p' = ne^{\frac{-n(t-o)}{T}} - ne^{\frac{-2n(t-o)}{T}}.$$

Consenso

- A seqüência consenso ou consenso para um alinhamento de cadeias em uma montagem é obtida tomando a maioria em cada coluna.

Notação

- O complemento-reverso de uma cadeia de DNA s é $\overline{s}$.

Notação

- O complemento-reverso de uma cadeia de DNA s é $\overline{s}$.
- Dizemos que duas cadeias s_i e s_j têm sobreposição (sufixo-prefixo) se $s_i[q, |s_i|] = s_j[1, p]$. O tamanho de uma sobreposição é $ovl(s_i, s_j)$.

Super-cadeia comum mais curta - SCS

- A idéia é encontrar uma cadeia de tamanho mínimo que contenha os reads como subcadeia supondo que

Super-cadeia comum mais curta - SCS

- A idéia é encontrar uma cadeia de tamanho mínimo que contenha os reads como subcadeia supondo que
  - Os reads têm orientação conhecida.

Super-cadeia comum mais curta - SCS

- A idéia é encontrar uma cadeia de tamanho mínimo que contenha os reads como subcadeia supondo que
  - Os reads têm orientação conhecida.
  - Os reads não têm erros de seqüenciamento.

Super-cadeia comum mais curta - SCS

- A idéia é encontrar uma cadeia de tamanho mínimo que contenha os reads como subcadeia supondo que
  - Os reads têm orientação conhecida.
  - Os reads não têm erros de seqüenciamento.
- Dado um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$, o *problema da super-cadeia comum mais curta* é encontrar a menor cadeia C que seja super-cadeia de toda cadeia em $\mathcal{S}$.

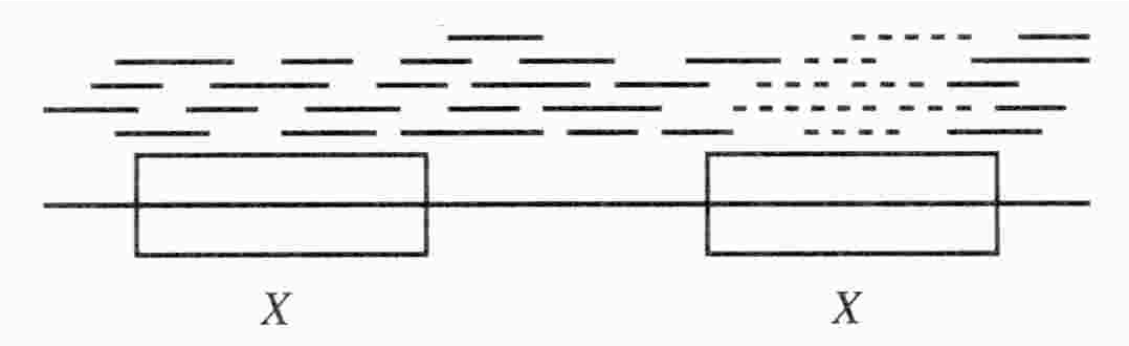
Super-cadeia comum mais curta - SCS

- A idéia é encontrar uma cadeia de tamanho mínimo que contenha os reads como subcadeia supondo que
  - Os reads têm orientação conhecida.
  - Os reads não têm erros de seqüenciamento.
- Dado um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$, o *problema da super-cadeia comum mais curta* é encontrar a menor cadeia C que seja super-cadeia de toda cadeia em $\mathcal{S}$.
- É NP-difícil. Existem aproximações.

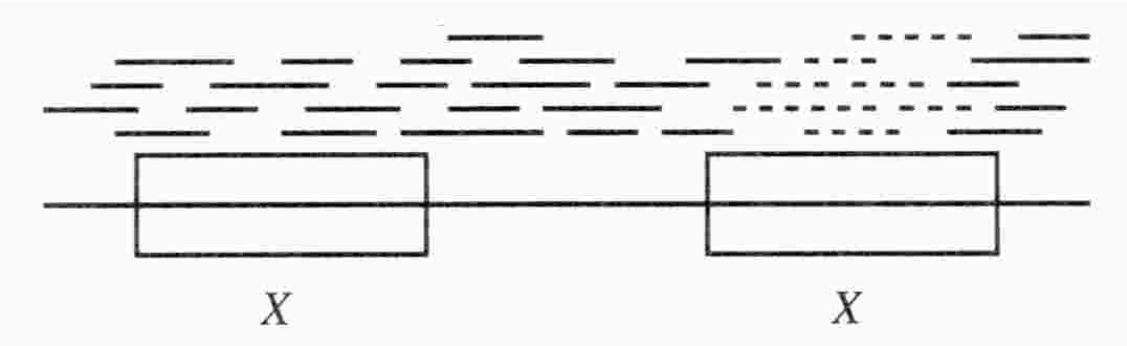
Super-cadeia comum mais curta

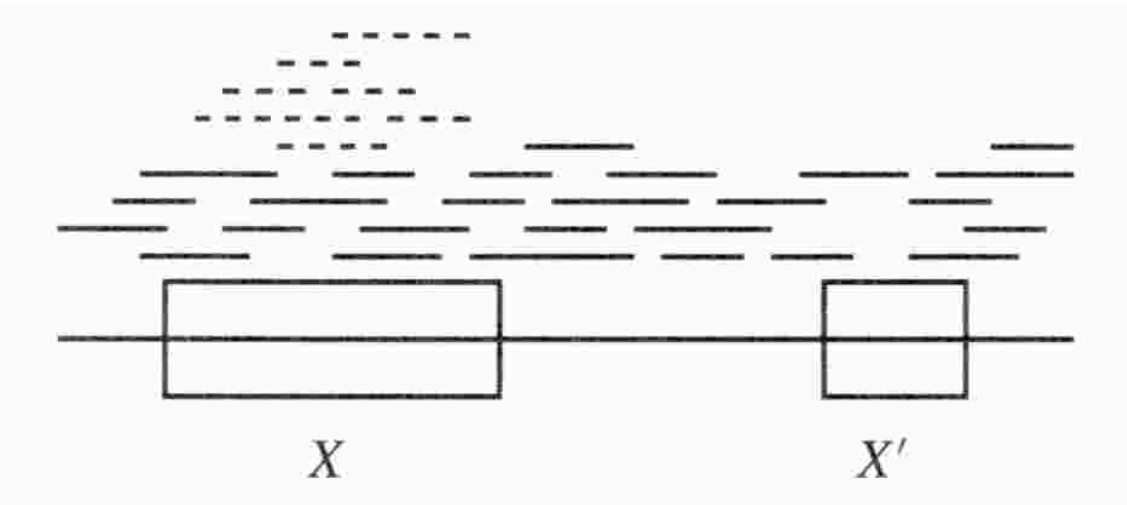
- Mesmo com a suposição de seqüenciamento perfeito e orientação conhecida, repetições levam a uma super-cadeia mais curta, que vai ser preferida.

Super-cadeia comum mais curta



Super-cadeia comum mais curta





[Setubal e Meidanis, 1996]

Reconstrução

- A idéia é encontrar uma cadeia de tamanho mínimo C tal que todo read ou o complemento reverso dele seja subcadeia de C com erro limitado por ε.

Reconstrução

- Seja $d_s(s, t)$ a distância de edição entre duas cadeias s e t que não penaliza espaços nas extremidades de t.

Reconstrução

- Seja $d_s(s,t)$ a distância de edição entre duas cadeias s e t que não penaliza espaços nas extremidades de t.
- Formalmente

$$d_s(s,t) = \min_{x \in SC(t)} \{ed(s,x)\},$$

onde $SC(t)$ é o conjunto de todas as subcadeias de t e $ed(\cdot,\cdot)$ é a distância de edição entre duas cadeias.

Reconstrução

- Seja $d_s(s,t)$ a distância de edição entre duas cadeias s e t que não penaliza espaços nas extremidades de t.
- Formalmente

$$d_s(s,t) = \min_{x \in SC(t)} \{ed(s,x)\},$$

onde $SC(t)$ é o conjunto de todas as subcadeias de t e $ed(\cdot,\cdot)$ é a distância de edição entre duas cadeias.

- d_s pode ser calculada usando PD semi-global com pontuação 1 para espaço ou mismatch e 0 para match.

Reconstrução

- Dado um conjunto de cadeias $\mathcal{S}$ e um real $\varepsilon \in [0, 1]$, o *problema da reconstrução* é encontrar uma menor cadeia C tal que para todo $s \in \mathcal{S}$ tenhamos

$$\min\{d_s(s, C), d_s(\overline{s}, C)\} \leq \varepsilon|s|.$$

Reconstrução

- Dado um conjunto de cadeias $\mathcal{S}$ e um real $\varepsilon \in [0, 1]$, o *problema da reconstrução* é encontrar uma menor cadeia C tal que para todo $s \in \mathcal{S}$ tenhamos

$$\min\{d_s(s, C), d_s(\overline{s}, C)\} \leq \varepsilon|s|.$$

- ε é a tolerância de erro.

Reconstrução

- Dado um conjunto de cadeias $\mathcal{S}$ e um real $\varepsilon \in [0, 1]$, o *problema da reconstrução* é encontrar uma menor cadeia C tal que para todo $s \in \mathcal{S}$ tenhamos

$$\min\{d_s(s, C), d_s(\overline{s}, C)\} \leq \varepsilon |s|.$$

- ε é a tolerância de erro.
- A super-cadeia comum mais curta é um caso particular da reconstrução, quando $\varepsilon = 0$.

Reconstrução

- É NP-difícil.

Reconstrução

- É NP-difícil.
- Inclui erros e orientação mas não modela repeats, falta de cobertura ou o tamanho da molécula original.

Multicontig

- A idéia é permitir a formação de vários contigs, exigindo que haja uma boa ligação entre os fragmentos.

Multicontig

- Dados um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$, um inteiro t e um real $\varepsilon \in [0, 1]$, o *problema multicontig* é particionar $\mathcal{S}$ em $\{B_1, \ldots, B_l\}$ e encontrar a menor cadeia C_i para cada B_i tal que

Multicontig

- Dados um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$, um inteiro t e um real $\varepsilon \in [0, 1]$, o *problema multicontig* é particionar $\mathcal{S}$ em $\{B_1, \ldots, B_l\}$ e encontrar a menor cadeia C_i para cada B_i tal que
  - para todo $f \in B_i$, ou f ou $\overline{f}$ é subcadeia de C_i,

Multicontig

- Dados um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$, um inteiro t e um real $\varepsilon \in [0, 1]$, o *problema multicontig* é particionar $\mathcal{S}$ em $\{B_1, \ldots, B_l\}$ e encontrar a menor cadeia C_i para cada B_i tal que
  - para todo $f \in B_i$, ou f ou $\overline{f}$ é subcadeia de C_i,
  - em C_i toda sobreposição entre duas cadeias x e y de B_i não contida em outra cadeia w de B_i tem tamanho pelo menos t.

Multicontig

- Dados um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$, um inteiro t e um real $\varepsilon \in [0, 1]$, o *problema multicontig* é particionar $\mathcal{S}$ em $\{B_1, \ldots, B_l\}$ e encontrar a menor cadeia C_i para cada B_i tal que
  - para todo $f \in B_i$, ou f ou $\overline{f}$ é subcadeia de C_i,
  - em C_i toda sobreposição entre duas cadeias x e y de B_i não contida em outra cadeia w de B_i tem tamanho pelo menos t.
- t define o tamanho da sobreposição mínima não contida entre dois reads.

Multicontig

- Dados um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$, um inteiro t e um real $\varepsilon \in [0, 1]$, o *problema multicontig* é particionar $\mathcal{S}$ em $\{B_1, \ldots, B_l\}$ e encontrar a menor cadeia C_i para cada B_i tal que
  - para todo $f \in B_i$, ou f ou $\overline{f}$ é subcadeia de C_i,
  - em C_i toda sobreposição entre duas cadeias x e y de B_i não contida em outra cadeia w de B_i tem tamanho pelo menos t.
- t define o tamanho da sobreposição mínima não contida entre dois reads.
- NP-difícil.

Multicontig com erros

- Seja f um fragmento e sejam $\ell(f)$ e $r(f)$ as colunas mais à esquerda e mais à direita de f no alinhamento.

Multicontig com erros

- Seja f um fragmento e sejam $\ell(f)$ e $r(f)$ as colunas mais à esquerda e mais à direita de f no alinhamento.
- A imagem de f sobre o consenso C do alinhamento é $C[\ell(f), r(f)]$.

Multicontig com erros

- Dados um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$, um inteiro t e um real $\varepsilon \in [0, 1]$, o *problema multicontig* é particionar $\mathcal{S}$ em $\{B_1, \ldots, B_l\}$ e encontrar a menor cadeia C_i para cada B_i tal que

Multicontig com erros

- Dados um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$, um inteiro t e um real $\varepsilon \in [0, 1]$, o *problema multicontig* é particionar $\mathcal{S}$ em $\{B_1, \ldots, B_l\}$ e encontrar a menor cadeia C_i para cada B_i tal que
  - para todo $f \in B_i$, ou f ou $\overline{f}$ é subcadeia de C_i,

Multicontig com erros

- Dados um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$, um inteiro t e um real $\varepsilon \in [0, 1]$, o *problema multicontig* é particionar $\mathcal{S}$ em $\{B_1, \ldots, B_l\}$ e encontrar a menor cadeia C_i para cada B_i tal que
  - para todo $f \in B_i$, ou f ou $\overline{f}$ é subcadeia de C_i,
  - em C_i toda sobreposição entre duas cadeias x e y de B_i não contida em outra cadeia w de B_i tem tamanho pelo menos t e

Multicontig com erros

- Dados um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$, um inteiro t e um real $\varepsilon \in [0, 1]$, o *problema multicontig* é particionar $\mathcal{S}$ em $\{B_1, \ldots, B_l\}$ e encontrar a menor cadeia C_i para cada B_i tal que
  - para todo $f \in B_i$, ou f ou $\overline{f}$ é subcadeia de C_i,
  - em C_i toda sobreposição entre duas cadeias x e y de B_i não contida em outra cadeia w de B_i tem tamanho pelo menos t e
  - para todo fragmento f $ed(f, C_i[\ell(f), r(f)]) \leq \varepsilon|s|$.

Multicontig com erros

- Dados um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$, um inteiro t e um real $\varepsilon \in [0, 1]$, o *problema multicontig* é particionar $\mathcal{S}$ em $\{B_1, \ldots, B_l\}$ e encontrar a menor cadeia C_i para cada B_i tal que
  - para todo $f \in B_i$, ou f ou $\overline{f}$ é subcadeia de C_i,
  - em C_i toda sobreposição entre duas cadeias x e y de B_i não contida em outra cadeia w de B_i tem tamanho pelo menos t e
  - para todo fragmento f $ed(f, C_i[\ell(f), r(f)]) \leq \varepsilon|s|$.
- t define o tamanho da sobreposição mínima não contida e ε é a tolerância de erro no alinhamento entre um fragmento e o consenso.

Multicontig com erros

- Dados um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$, um inteiro t e um real $\varepsilon \in [0, 1]$, o *problema multicontig* é particionar $\mathcal{S}$ em $\{B_1, \ldots, B_l\}$ e encontrar a menor cadeia C_i para cada B_i tal que
  - para todo $f \in B_i$, ou f ou $\overline{f}$ é subcadeia de C_i,
  - em C_i toda sobreposição entre duas cadeias x e y de B_i não contida em outra cadeia w de B_i tem tamanho pelo menos t e
  - para todo fragmento f $ed(f, C_i[\ell(f), r(f)]) \leq \varepsilon|s|$.
- t define o tamanho da sobreposição mínima não contida e ε é a tolerância de erro no alinhamento entre um fragmento e o consenso.
- NP-difícil.

Caminhos

- Um caminho Hamiltoniano em um grafo é um caminho que passa em cada vértice exatamente uma vez.

Caminhos

- Um caminho Hamiltoniano em um grafo é um caminho que passa em cada vértice exatamente uma vez.
- Um caminho Euleriano em um grafo é um caminho que passa em cada aresta exatamente uma vez.

Caminhos

- Um caminho Hamiltoniano em um grafo é um caminho que passa em cada vértice exatamente uma vez.
- Um caminho Euleriano em um grafo é um caminho que passa em cada aresta exatamente uma vez.
- Um caminho chinês em um grafo é um caminho que passa em cada aresta pelo menos uma vez.

Grafo de sobreposições

- Dois fragmentos x e y se sobrepõem se existir uma cadeia não-vazia z de tamanho maximal que é sufixo de x e prefixo de y.

Grafo de sobreposições

- Dois fragmentos x e y se sobrepõem se existir uma cadeia não-vazia z de tamanho maximal que é sufixo de x e prefixo de y.
- Um grafo de sobreposições H para um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$ é um grafo orientado completo e com pesos

Grafo de sobreposições

- Dois fragmentos x e y se sobrepõem se existir uma cadeia não-vazia z de tamanho maximal que é sufixo de x e prefixo de y.
- Um grafo de sobreposições H para um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$ é um grafo orientado completo e com pesos
  - que tem um vértice para cada cadeia em $\mathcal{S}$ e

Grafo de sobreposições

- Dois fragmentos x e y se sobrepõem se existir uma cadeia não-vazia z de tamanho maximal que é sufixo de x e prefixo de y.
- Um grafo de sobreposições H para um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$ é um grafo orientado completo e com pesos
  - que tem um vértice para cada cadeia em $\mathcal{S}$ e
  - em que $w(s_i, s_j) = |s_j| - ovl(s_i, s_j)$ para $i \neq j$.

Grafo de sobreposições

- Dois fragmentos x e y se sobrepõem se existir uma cadeia não-vazia z de tamanho maximal que é sufixo de x e prefixo de y.
- Um grafo de sobreposições H para um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$ é um grafo orientado completo e com pesos
  - que tem um vértice para cada cadeia em $\mathcal{S}$ e
  - em que $w(s_i, s_j) = |s_j| - ovl(s_i, s_j)$ para $i \neq j$.
- Um caminho que passa por todos os vértices exatamente uma vez e tem custo mínimo corresponde a uma montagem mais curta dos fragmentos.

Grafo de sobreposições

- Dois fragmentos x e y se sobrepõem se existir uma cadeia não-vazia z de tamanho maximal que é sufixo de x e prefixo de y.
- Um grafo de sobreposições H para um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_k\}$ é um grafo orientado completo e com pesos
  - que tem um vértice para cada cadeia em $\mathcal{S}$ e
  - em que $w(s_i, s_j) = |s_j| - ovl(s_i, s_j)$ para $i \neq j$.
- Um caminho que passa por todos os vértices exatamente uma vez e tem custo mínimo corresponde a uma montagem mais curta dos fragmentos.
- Encontrar esse caminho é NP-difícil.

Grafo de Bruijn

.

- P.A. Pevzner, H. Tang e M.S. Waterman, 2001.

An Eulerian path approach to DNA fragment assembly. PNAS, 98, pág. 9748, 2001.

Grafo de Bruijn

.

- P.A. Pevzner, H. Tang e M.S. Waterman, 2001.
- O espectro de um read é o conjunto de suas subcadeias de tamanho k (k-mers).

An Eulerian path approach to DNA fragment assembly. PNAS, 98, pág. 9748, 2001.

Grafo de Bruijn

.

- P.A. Pevzner, H. Tang e M.S. Waterman, 2001.
- O espectro de um read é o conjunto de suas subcadeias de tamanho k (k-mers).
- Para um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_m\}$, um grafo de Bruijn $G_k(\mathcal{S})$

An Eulerian path approach to DNA fragment assembly. PNAS, 98, pág. 9748, 2001.

Grafo de Bruijn

.

- P.A. Pevzner, H. Tang e M.S. Waterman, 2001.
- O espectro de um read é o conjunto de suas subcadeias de tamanho k (k-mers).
- Para um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_m\}$, um grafo de Bruijn $G_k(\mathcal{S})$
  - tem como vértices cada um dos k-mers na união dos espectros de $\{s_1, \ldots, s_m\}$ e

An Eulerian path approach to DNA fragment assembly. PNAS, 98, pág. 9748, 2001.

Grafo de Bruijn

.

- P.A. Pevzner, H. Tang e M.S. Waterman, 2001.
- O espectro de um read é o conjunto de suas subcadeias de tamanho k (k-mers).
- Para um conjunto de cadeias $\mathcal{S} = \{s_1, \ldots, s_m\}$, um grafo de Bruijn $G_k(\mathcal{S})$
  - tem como vértices cada um dos k-mers na união dos espectros de $\{s_1, \ldots, s_m\}$ e
  - uma aresta $(s[1, k], s[2, k+1])$ para cada subcadeia s de tamanho $k+1$ em cada cadeia de $\mathcal{S}$.

An Eulerian path approach to DNA fragment assembly. PNAS, 98, pág. 9748, 2001.

Grafo de Bruijn

- Cada read s_i corresponde a um caminho

$$p(s_i) = s[1, k] \to s[2, k+1] \to \ldots \to s[|s_i| - k + 1, |s_i|].$$

Grafo de Bruijn

- Cada read s_i corresponde a um caminho

$$p(s_i) = s[1, k] \to s[2, k+1] \to \ldots \to s[|s_i| - k + 1, |s_i|].$$

- Um caminho é um super-caminho de $G_k(\mathcal{S})$ se ele contém o caminho $p(s_i)$ para todo s_i em $\mathcal{S}$.

Grafo de Bruijn

- Cada read s_i corresponde a um caminho

$$p(s_i) = s[1, k] \to s[2, k+1] \to \ldots \to s[|s_i| - k + 1, |s_i|].$$

- Um caminho é um super-caminho de $G_k(\mathcal{S})$ se ele contém o caminho $p(s_i)$ para todo s_i em $\mathcal{S}$.
- Um super-caminho corresponde a uma montagem válida dos reads.

Grafo de Bruijn

- Cada read s_i corresponde a um caminho

$$p(s_i) = s[1, k] \to s[2, k+1] \to \ldots \to s[|s_i| - k + 1, |s_i|].$$

- Um caminho é um super-caminho de $G_k(\mathcal{S})$ se ele contém o caminho $p(s_i)$ para todo s_i em $\mathcal{S}$.
- Um super-caminho corresponde a uma montagem válida dos reads.
- Uma montagem parcimoniosa é um super-caminho de tamanho mínimo em $G_k(\mathcal{S})$.

Grafo de Bruijn

- Cada read s_i corresponde a um caminho

$$p(s_i) = s[1, k] \to s[2, k+1] \to \ldots \to s[|s_i| - k + 1, |s_i|].$$

- Um caminho é um super-caminho de $G_k(\mathcal{S})$ se ele contém o caminho $p(s_i)$ para todo s_i em $\mathcal{S}$.
- Um super-caminho corresponde a uma montagem válida dos reads.
- Uma montagem parcimoniosa é um super-caminho de tamanho mínimo em $G_k(\mathcal{S})$.
- Problema NP-difícil (super-caminho de Bruijn).

Grafo de Bruijn bi-orientado

- P. Medvedev *et al.*.

Computability of models for sequence assembly, WABI, pág. 289, 2007.

Grafo de Bruijn bi-orientado

- P. Medvedev *et al.*.
- Uma k-molécula é um par de k-mers que são complementos reversos um do outro.

Computability of models for sequence assembly, WABI, pág. 289, 2007.

Grafo de Bruijn bi-orientado

- P. Medvedev *et al.*.
- Uma k-molécula é um par de k-mers que são complementos reversos um do outro.
- O k-espectro-molécula de uma molécula de DNA é o conjunto de todas as k-moléculas que correspondem aos k-mers do k-espectro de ambas as fitas do DNA.

Computability of models for sequence assembly, WABI, pág. 289, 2007.

Grafo de Bruijn bi-orientado

- P. Medvedev *et al.*.
- Uma k-molécula é um par de k-mers que são complementos reversos um do outro.
- O k-espectro-molécula de uma molécula de DNA é o conjunto de todas as k-moléculas que correspondem aos k-mers do k-espectro de ambas as fitas do DNA.
- Em toda k-molécula, um dos k-mers é rotulado $+$ e o outro é rotulado $-$. Essa orientação é arbitrária mas uma k-molécula é sempre orientada da mesma forma.

Computability of models for sequence assembly, WABI, pág. 289, 2007.

Grafo de Bruijn bi-orientado

- Um grafo de Bruijn bi-orientado $B_k(\mathcal{S})$ para cadeias $\mathcal{S} = \{s_1, \ldots, s_m\}$ tem como vérices as k-moléculas na união de todos os k-espectro-moléculas das cadeias em $\mathcal{S}$.

Grafo de Bruijn bi-orientado

- Um grafo de Bruijn bi-orientado $B_k(\mathcal{S})$ para cadeias $\mathcal{S} = \{s_1, \ldots, s_m\}$ tem como vérices as k-moléculas na união de todos os k-espectro-moléculas das cadeias em $\mathcal{S}$.
- Uma aresta representa uma $(k+1)$-molécula e conecta duas k-moléculas que aparecem consecutivamente em alguma k-espectro-molécula para as cadeias em $\mathcal{S}$.

Grafo de Bruijn bi-orientado

- Um grafo de Bruijn bi-orientado $B_k(\mathcal{S})$ para cadeias $\mathcal{S} = \{s_1, \ldots, s_m\}$ tem como vérices as k-moléculas na união de todos os k-espectro-moléculas das cadeias em $\mathcal{S}$.
- Uma aresta representa uma $(k+1)$-molécula e conecta duas k-moléculas que aparecem consecutivamente em alguma k-espectro-molécula para as cadeias em $\mathcal{S}$.
- Cada aresta tem duas orientações, que indicam qual orientação do k-mer foi considerada em cada vértice.

Grafo de Bruijn bi-orientado

- Se (x, y) é uma $(k+1)$-molécula que corresponde a $w[i, i+k-1]$ e a $w[i+1, i+k]$ então

Grafo de Bruijn bi-orientado

- Se (x, y) é uma $(k+1)$-molécula que corresponde a $w[i, i+k-1]$ e a $w[i+1, i+k]$ então
  - ela é positiva na incidência com x se $w[i, i+k-1]$ é positiva e é negativa caso contrário.

Grafo de Bruijn bi-orientado

- Se (x, y) é uma $(k+1)$-molécula que corresponde a $w[i, i+k-1]$ e a $w[i+1, i+k]$ então
  - ela é positiva na incidência com x se $w[i, i+k-1]$ é positiva e é negativa caso contrário.
  - ela é positiva na incidência com y se $w[i+1, i+k]$ é positiva e é negativa caso contrário.

Grafo de Bruijn bi-orientado

- Se (x, y) é uma $(k+1)$-molécula que corresponde a $w[i, i+k-1]$ e a $w[i+1, i+k]$ então
  - ela é positiva na incidência com x se $w[i, i+k-1]$ é positiva e é negativa caso contrário.
  - ela é positiva na incidência com y se $w[i+1, i+k]$ é positiva e é negativa caso contrário.
- Cada $(k+1)$-mer e seu complemento são representados por exatamente uma aresta do grafo.

Grafo de Bruijn bi-orientado

- Cada montagem válida dos fragmentos corresponde a um caminho no grafo $B_k(\mathcal{S})$.

Grafo de Bruijn bi-orientado

- Cada montagem válida dos fragmentos corresponde a um caminho no grafo $B_k(\mathcal{S})$.
- Um caminho de tamanho mínimo que contém cada aresta pelo menos uma vez é uma solução parcimoniosa para a montagem.

Grafo de Bruijn bi-orientado

- Cada montagem válida dos fragmentos corresponde a um caminho no grafo $B_k(\mathcal{S})$.
- Um caminho de tamanho mínimo que contém cada aresta pelo menos uma vez é uma solução parcimoniosa para a montagem.
- Esse problema é o do caminho chinês de custo mínimo e pode ser resolvido em tempo polinomial (lembre que esse é o caso sem erros.)

Grafo de cadeias

- Seja $\mathcal{S}$ um conjunto de cadeias em que nenhuma está contida em outra.

E. Myers. The fragment assembly string graph, ECCB/JBI, pág. 85, 2005.

Grafo de cadeias

- Seja $\mathcal{S}$ um conjunto de cadeias em que nenhuma está contida em outra.
- Sejam x e y cadeias que se sobrepõem. Denotamos a sobreposição por $x[b_{xy}, e_{xy}] = y[b_{yx}, e_{yx}]$.

E. Myers. The fragment assembly string graph, ECCB/JBI, pág. 85, 2005.

Grafo de cadeias

- Seja $\mathcal{S}$ um conjunto de cadeias em que nenhuma está contida em outra.
- Sejam x e y cadeias que se sobrepõem. Denotamos a sobreposição por $x[b_{xy}, e_{xy}] = y[b_{yx}, e_{yx}]$.
- O grafo de cadeias preliminar tem um vértice para cada cadeia de $\mathcal{S}$.

E. Myers. The fragment assembly string graph, ECCB/JBI, pág. 85, 2005.

Grafo de cadeias

- Seja $\mathcal{S}$ um conjunto de cadeias em que nenhuma está contida em outra.
- Sejam x e y cadeias que se sobrepõem. Denotamos a sobreposição por $x[b_{xy}, e_{xy}] = y[b_{yx}, e_{yx}]$.
- O grafo de cadeias preliminar tem um vértice para cada cadeia de $\mathcal{S}$.
- Para cada sobreposição entre x e y adicionamos uma aresta bi-direcionada (x, y).

E. Myers. The fragment assembly string graph, ECCB/JBI, pág. 85, 2005.

Grafo de cadeias

- Cada aresta tem dois rótulos, um para cada parte não sobreposta dos reads.

Grafo de cadeias

- Cada aresta tem dois rótulos, um para cada parte não sobreposta dos reads.
- Cada aresta tem dois tipos, um para cada ponta.

Grafo de cadeias

- Cada aresta tem dois rótulos, um para cada parte não sobreposta dos reads.
- Cada aresta tem dois tipos, um para cada ponta.
- Cada aresta tem os dados $(type_{xy}, type_{yx}, label_{xy}, label_{yx})$.

Grafo de cadeias

- $type_{xy}$ é igual a B se um prefixo de x se sobrepõe ou E se um sufixo de x se sobrepõe:

$$type_{xy} = \begin{cases} B & \text{se } b_{xy} = 1 \\ E & \text{se } e_{xy} = |x| \end{cases}$$

Grafo de cadeias

- $type_{xy}$ é igual a B se um prefixo de x se sobrepõe ou E se um sufixo de x se sobrepõe:

$$type_{xy} = \begin{cases} B & \text{se } b_{xy} = 1 \\ E & \text{se } e_{xy} = |x| \end{cases}$$

- $type_{yx}$ é igual a B se um prefixo de y se sobrepõe ou E se um sufixo de y se sobrepõe:

$$type_{yx} = \begin{cases} B & \text{se } b_{yx} = 1 \\ E & \text{se } e_{yx} = |y| \end{cases}$$

Grafo de cadeias

- $label_{xy}$ é igual à parte não sobreposta de y:

$$label_{xy} = \begin{cases} y[e_{yx}+1, |y|] & \text{se } b_{yx} = 1 \\ y[1, b_{yx}-1] & \text{se } e_{yx} = |y| \end{cases}$$

Grafo de cadeias

- $label_{xy}$ é igual à parte não sobreposta de y:

$$label_{xy} = \begin{cases} y[e_{yx}+1, |y|] & \text{se } b_{yx} = 1 \\ y[1, b_{yx}-1] & \text{se } e_{yx} = |y| \end{cases}$$

- $label_{yx}$ é igual à parte não sobreposta de x:

$$label_{yx} = \begin{cases} x[e_{xy}+1, |x|] & \text{se } b_{xy} = 1 \\ y[1, b_{xy}-1] & \text{se } e_{xy} = |x| \end{cases}$$

Grafo de cadeias

- $label_{xy}$ é igual à parte não sobreposta de y:

$$label_{xy} = \begin{cases} y[e_{yx}+1, |y|] & \text{se } b_{yx} = 1 \\ y[1, b_{yx}-1] & \text{se } e_{yx} = |y| \end{cases}$$

- $label_{yx}$ é igual à parte não sobreposta de x:

$$label_{yx} = \begin{cases} x[e_{xy}+1, |x|] & \text{se } b_{xy} = 1 \\ y[1, b_{xy}-1] & \text{se } e_{xy} = |x| \end{cases}$$

- Se a sobreposição entre x e y é complementar reversa, digamos $\overline{x[b_{xy}, e_{xy}]} = y[b_{yx}, e_{yx}]$, então $label_{xy}$ e $label_{yx}$ são complementares reversos também.

Grafo de cadeias

- $label_{xy}$ é igual à parte não sobreposta de y:

$$label_{xy} = \begin{cases} y[e_{yx}+1, |y|] & \text{se } b_{yx} = 1 \\ y[1, b_{yx}-1] & \text{se } e_{yx} = |y| \end{cases}$$

- $label_{yx}$ é igual à parte não sobreposta de x:

$$label_{yx} = \begin{cases} x[e_{xy}+1, |x|] & \text{se } b_{xy} = 1 \\ y[1, b_{xy}-1] & \text{se } e_{xy} = |x| \end{cases}$$

- Se a sobreposição entre x e y é complementar reversa, digamos $\overline{x[b_{xy}, e_{xy}]} = y[b_{yx}, e_{yx}]$, então $label_{xy}$ e $label_{yx}$ são complementares reversos também.
- Uma aresta que representa uma sobreposição complementar reversa tem suas duas orientações do mesmo tipo.

Grafo de cadeias

- Sejam cadeias x, y e z que se sobrepõem duas-a-duas.

Grafo de cadeias

- Sejam cadeias x, y e z que se sobrepõem duas-a-duas.
- Se y e z se sobrepõem com a mesma ponta de z ($type_{xy} = type_{xz}$) então existe um caminho que contém as três cadeias em sucessão.

Grafo de cadeias

- Sejam cadeias x, y e z que se sobrepõem duas-a-duas.
- Se y e z se sobrepõem com a mesma ponta de z ($type_{xy} = type_{xz}$) então existe um caminho que contém as três cadeias em sucessão.
- Digamos que esse caminho seja $x \rightarrow y \rightarrow z$.

Grafo de cadeias

- Sejam cadeias x, y e z que se sobrepõem duas-a-duas.
- Se y e z se sobrepõem com a mesma ponta de z ($type_{xy} = type_{xz}$) então existe um caminho que contém as três cadeias em sucessão.
- Digamos que esse caminho seja $x \to y \to z$.
- Então a aresta (x, z) é transitiva e pode ser removida do grafo sem perda de informação.

Grafo de cadeias

- Sejam cadeias x, y e z que se sobrepõem duas-a-duas.
- Se y e z se sobrepõem com a mesma ponta de z ($type_{xy} = type_{xz}$) então existe um caminho que contém as três cadeias em sucessão.
- Digamos que esse caminho seja $x \to y \to z$.
- Então a aresta (x, z) é transitiva e pode ser removida do grafo sem perda de informação.
- O grafo de cadeias é um subgrafo sem arestas transitivas maximal do grafo de cadeias preliminar.

Grafo de cadeias

- Além disso cada aresta do grafo é classificada (estatisticamente) como

Grafo de cadeias

- Além disso cada aresta do grafo é classificada (estatisticamente) como
- obrigatória se deve aparecer pelo menos uma vez em um caminho.

Grafo de cadeias

- Além disso cada aresta do grafo é classificada (estatisticamente) como
- obrigatória se deve aparecer pelo menos uma vez em um caminho.
- exata se deve aparecer exatamente uma vez em um caminho.

Grafo de cadeias

- Além disso cada aresta do grafo é classificada (estatisticamente) como
- obrigatória se deve aparecer pelo menos uma vez em um caminho.
- exata se deve aparecer exatamente uma vez em um caminho.
- opcional se pode aparecer um número qualquer de vezes em um caminho.

Grafo de cadeias

- A concatenação de x e $label_{xy}$ é uma montagem de x e y.

Grafo de cadeias

- A concatenação de x e $label_{xy}$ é uma montagem de x e y.
- Um caminho nesse grafo é válido se entra em um vértice por uma orientção do tipo B e sai por uma do tipo E ou vice-versa.

Grafo de cadeias

- A concatenação de x e $label_{xy}$ é uma montagem de x e y.
- Um caminho nesse grafo é válido se entra em um vértice por uma orientção do tipo B e sai por uma do tipo E ou vice-versa.
- Um caminho cíclico no grafo de cadeias que respeita as restrições de seleção da classificação representa uma montagem válida do genoma.

Grafo de cadeias

- A concatenação de x e $label_{xy}$ é uma montagem de x e y.
- Um caminho nesse grafo é válido se entra em um vértice por uma orientção do tipo B e sai por uma do tipo E ou vice-versa.
- Um caminho cíclico no grafo de cadeias que respeita as restrições de seleção da classificação representa uma montagem válida do genoma.
- Seja s uma função de seleção das arestas de I em $\{$obrigatória, exata, opcional$\}$.

Grafo de cadeias

- A concatenação de x e $label_{xy}$ é uma montagem de x e y.
- Um caminho nesse grafo é válido se entra em um vértice por uma orientção do tipo B e sai por uma do tipo E ou vice-versa.
- Um caminho cíclico no grafo de cadeias que respeita as restrições de seleção da classificação representa uma montagem válida do genoma.
- Seja s uma função de seleção das arestas de I em $\{$obrigatória, exata, opcional$\}$.
- Um s-path é um caminho em I que contém todas as arestas obrigatórias pelo menos uma vez, todas as arestas exatas uma vez e todas as arestas opcionais um número qualquer de vezes.

Grafo de cadeias

- A concatenação de x e $label_{xy}$ é uma montagem de x e y.
- Um caminho nesse grafo é válido se entra em um vértice por uma orientção do tipo B e sai por uma do tipo E ou vice-versa.
- Um caminho cíclico no grafo de cadeias que respeita as restrições de seleção da classificação representa uma montagem válida do genoma.
- Seja s uma função de seleção das arestas de I em $\{$obrigatória, exata, opcional$\}$.
- Um s-path é um caminho em I que contém todas as arestas obrigatórias pelo menos uma vez, todas as arestas exatas uma vez e todas as arestas opcionais um número qualquer de vezes.
- O problema de encontrar um s-path cíclico de peso mínimo em um grafo orientado com pesos e função de seleção s é NP-difícil.

Sobre a classificação prática

- Depois da eliminação de transitividades o autor propõe a redução de seqüências com vértices com grau de entrada e saída igual a 1 por uma única aresta.

Sobre a classificação prática

- Depois da eliminação de transitividades o autor propõe a redução de seqüências com vértices com grau de entrada e saída igual a 1 por uma única aresta.
- Arestas exatas são arestas com uma grande probabilidade de representarem uma região coberta apenas por aquela aresta. As contas são feitas a partir de uma estimativa do tamanho de genoma e da aresta.

Sobre a classificação prática

- Depois da eliminação de transitividades o autor propõe a redução de seqüências com vértices com grau de entrada e saída igual a 1 por uma única aresta.
- Arestas exatas são arestas com uma grande probabilidade de representarem uma região coberta apenas por aquela aresta. As contas são feitas a partir de uma estimativa do tamanho de genoma e da aresta.
- Arestas obrigatórias são as que restaram e têm um vértice interior.

Sobre a classificação prática

- Depois da eliminação de transitividades o autor propõe a redução de seqüências com vértices com grau de entrada e saída igual a 1 por uma única aresta.
- Arestas exatas são arestas com uma grande probabilidade de representarem uma região coberta apenas por aquela aresta. As contas são feitas a partir de uma estimativa do tamanho de genoma e da aresta.
- Arestas obrigatórias são as que restaram e têm um vértice interior.
- As que não têm um vértice interior são opcionais.

Heurísticas

- Grandes classes de heurísticas:

Heurísticas

- Grandes classes de heurísticas:
  - Gulosas.

Heurísticas

- Grandes classes de heurísticas:
  - Gulosas.
  - Baseadas em grafo de sobreposições.

Heurísticas

- Grandes classes de heurísticas:
  - Gulosas.
  - Baseadas em grafo de sobreposições.
  - Baseadas em grafo de cadeias.

Heurísticas

- Grandes classes de heurísticas:
  - Gulosas.
  - Baseadas em grafo de sobreposições.
  - Baseadas em grafo de cadeias.
  - Baseadas em grafo de Bruijn.

Heurísticas

- Grandes classes de heurísticas:
  - Gulosas.
  - Baseadas em grafo de sobreposições.
  - Baseadas em grafo de cadeias.
  - Baseadas em grafo de Bruijn.
- As três primeiras incluem uma primeira fase de comparação dos reads para encontrar sobreposições sufixo-prefixo.

Heurísticas

- Grandes classes de heurísticas:
  - Gulosas.
  - Baseadas em grafo de sobreposições.
  - Baseadas em grafo de cadeias.
  - Baseadas em grafo de Bruijn.
- As três primeiras incluem uma primeira fase de comparação dos reads para encontrar sobreposições sufixo-prefixo.
  - Normalmente acelerada usando assinaturas.

Heurísticas

- Grandes classes de heurísticas:
  - Gulosas.
  - Baseadas em grafo de sobreposições.
  - Baseadas em grafo de cadeias.
  - Baseadas em grafo de Bruijn.
- As três primeiras incluem uma primeira fase de comparação dos reads para encontrar sobreposições sufixo-prefixo.
  - Normalmente acelerada usando assinaturas.
  - Trivialmente paralelizável.

Heurísticas

- Grandes classes de heurísticas:
  - Gulosas.
  - Baseadas em grafo de sobreposições.
  - Baseadas em grafo de cadeias.
  - Baseadas em grafo de Bruijn.
- As três primeiras incluem uma primeira fase de comparação dos reads para encontrar sobreposições sufixo-prefixo.
  - Normalmente acelerada usando assinaturas.
  - Trivialmente paralelizável.
- Normalmente incluem uma estratégia de 'limpeza' dos dados para remoção de erros que podem ser identificados previamente.

Heurísticas

- Grandes classes de heurísticas:
  - Gulosas.
  - Baseadas em grafo de sobreposições.
  - Baseadas em grafo de cadeias.
  - Baseadas em grafo de Bruijn.
- As três primeiras incluem uma primeira fase de comparação dos reads para encontrar sobreposições sufixo-prefixo.
  - Normalmente acelerada usando assinaturas.
  - Trivialmente paralelizável.
- Normalmente incluem uma estratégia de 'limpeza' dos dados para remoção de erros que podem ser identificados previamente.
- Normalmente incluem alguma estratégia para lidar com pares-de-reads.

Gulosa

- Estratégia:

Gulosa

- Estratégia:
  - Definir alguma medida M para avaliar sobreposições entre reads e entre contigs.

Gulosa

- Estratégia:
  - Definir alguma medida M para avaliar sobreposições entre reads e entre contigs.
  - Agrupar o par de reads e/ou contigs com o maior M, enquanto houver algum par com M maior que um limiar mínimo.

Gulosa

- Estratégia:
  - Definir alguma medida M para avaliar sobreposições entre reads e entre contigs.
  - Agrupar o par de reads e/ou contigs com o maior M, enquanto houver algum par com M maior que um limiar mínimo.
  - Construir um consenso para cada contig por uma votação entre os reads no contig.

Gulosa

- Estratégia:
  - Definir alguma medida M para avaliar sobreposições entre reads e entre contigs.
  - Agrupar o par de reads e/ou contigs com o maior M, enquanto houver algum par com M maior que um limiar mínimo.
  - Construir um consenso para cada contig por uma votação entre os reads no contig.
- Phrap, TIGR assembler, CAP3.

Divergência

- Ocorre divergência quando um read A tem sobreposição com B e C, mas B e C não têm sobreposição.

Divergência

- Ocorre divergência quando um read A tem sobreposição com B e C, mas B e C não têm sobreposição.
- Acontece nas fronteiras de regiões repetidas ou por causa de erros de seqüenciamento.

Outra gulosa

- Estratégia:

Outra gulosa

- Estratégia:
  - Definir algum critério para avaliar sobreposições entre reads. Pode considerar tamanho da sobreposição, qualidade das bases, número de seqüências que confirmam a sobreposição etc.

Outra gulosa

- Estratégia:
  - Definir algum critério para avaliar sobreposições entre reads. Pode considerar tamanho da sobreposição, qualidade das bases, número de seqüências que confirmam a sobreposição etc.
  - Selecionar um read para começar um contig.

Outra gulosa

- Estratégia:
  - Definir algum critério para avaliar sobreposições entre reads. Pode considerar tamanho da sobreposição, qualidade das bases, número de seqüências que confirmam a sobreposição etc.
  - Selecionar um read para começar um contig.
  - Estender o contig na extremidade 3' até que ocorra divergência.

Outra gulosa

- Estratégia:
  - Definir algum critério para avaliar sobreposições entre reads. Pode considerar tamanho da sobreposição, qualidade das bases, número de seqüências que confirmam a sobreposição etc.
  - Selecionar um read para começar um contig.
  - Estender o contig na extremidade 3' até que ocorra divergência.
  - Estender o contig na extremidade 5' até que ocorra divergência.

Outra gulosa

- Estratégia:
  - Definir algum critério para avaliar sobreposições entre reads. Pode considerar tamanho da sobreposição, qualidade das bases, número de seqüências que confirmam a sobreposição etc.
  - Selecionar um read para começar um contig.
  - Estender o contig na extremidade 3' até que ocorra divergência.
  - Estender o contig na extremidade 5' até que ocorra divergência.
  - Construir um consenso para cada contig.

Outra gulosa

- Estratégia:
  - Definir algum critério para avaliar sobreposições entre reads. Pode considerar tamanho da sobreposição, qualidade das bases, número de seqüências que confirmam a sobreposição etc.
  - Selecionar um read para começar um contig.
  - Estender o contig na extremidade 3' até que ocorra divergência.
  - Estender o contig na extremidade 5' até que ocorra divergência.
  - Construir um consenso para cada contig.
- SSAKE,VCAKE,SHARCGS.

Grafo de sobreposições, de cadeias e de Bruijn

- Estratégia:

Grafo de sobreposições, de cadeias e de Bruijn

- Estratégia:
  - Construir um grafo de sobreposições.

Grafo de sobreposições, de cadeias e de Bruijn

- Estratégia:
  - Construir um grafo de sobreposições.
  - Identificar caminhos que correspondem a regiões do genoma.

Grafo de sobreposições, de cadeias e de Bruijn

- Estratégia:
  - Construir um grafo de sobreposições.
  - Identificar caminhos que correspondem a regiões do genoma.
  - Construir consenso.

Grafo de sobreposições, de cadeias e de Bruijn

- Estratégia:
  - Construir um grafo de sobreposições.
  - Identificar caminhos que correspondem a regiões do genoma.
  - Construir consenso.
- Sobreposições: Celera assembler, Arachne, Edena.

Grafo de sobreposições, de cadeias e de Bruijn

- Estratégia:
  - Construir um grafo de sobreposições.
  - Identificar caminhos que correspondem a regiões do genoma.
  - Construir consenso.
- Sobreposições: Celera assembler, Arachne, Edena.
- Cadeias: Euler.

Grafo de sobreposições, de cadeias e de Bruijn

- Estratégia:
  - Construir um grafo de sobreposições.
  - Identificar caminhos que correspondem a regiões do genoma.
  - Construir consenso.
- Sobreposições: Celera assembler, Arachne, Edena.
- Cadeias: Euler.
- de Bruijn: Velvet, Allpaths.

Scaffolding

- As estratégias de montagem normalmente produzem vários contigs.

Scaffolding

- As estratégias de montagem normalmente produzem vários contigs.
  - Um contig é uma cadeia na qual a ordem das letras foi determinada com alta confiabilidade.

Scaffolding

- As estratégias de montagem normalmente produzem vários contigs.
  - Um contig é uma cadeia na qual a ordem das letras foi determinada com alta confiabilidade.
- Scaffolding é uma forma de reconstruir um DNA usando informação dos pares-de-reads para determinar a ordem relativa de contigs.

Scaffolding

- As estratégias de montagem normalmente produzem vários contigs.
  - Um contig é uma cadeia na qual a ordem das letras foi determinada com alta confiabilidade.
- Scaffolding é uma forma de reconstruir um DNA usando informação dos pares-de-reads para determinar a ordem relativa de contigs.
- Um scaffold (andaime) é composto por contigs e buracos.

Scaffolding

- As estratégias de montagem normalmente produzem vários contigs.
  - Um contig é uma cadeia na qual a ordem das letras foi determinada com alta confiabilidade.
- Scaffolding é uma forma de reconstruir um DNA usando informação dos pares-de-reads para determinar a ordem relativa de contigs.
- Um scaffold (andaime) é composto por contigs e buracos.
- Um buraco é definido por pelo menos um par-de-reads com um read em um contig e o outro em um contig diferente. O tamanho do buraco pode ser estimado.

Scaffolding

- Nem sempre é possível obter um único scaffold para os contigs.

Scaffolding

- Nem sempre é possível obter um único scaffold para os contigs.
- Scaffolds podem se sobrepor, por exemplo quando representam alelos.

Scaffolding

- Nem sempre é possível obter um único scaffold para os contigs.
- Scaffolds podem se sobrepor, por exemplo quando representam alelos.
- A ordem relativa dos scaffolds ao longo da molécula não é conhecida.

Scaffolding

- Nem sempre é possível obter um único scaffold para os contigs.
- Scaffolds podem se sobrepor, por exemplo quando representam alelos.
- A ordem relativa dos scaffolds ao longo da molécula não é conhecida.
- Quase todos os montadores fazem scaffolding.

Scaffolding

- Nem sempre é possível obter um único scaffold para os contigs.
- Scaffolds podem se sobrepor, por exemplo quando representam alelos.
- A ordem relativa dos scaffolds ao longo da molécula não é conhecida.
- Quase todos os montadores fazem scaffolding.
- A estratégia típica para determinar os scaffolds é gulosa, incorporando informação por ordem de confiabilidade até que haja um conflito.

Scaffolding

- Nem sempre é possível obter um único scaffold para os contigs.
- Scaffolds podem se sobrepor, por exemplo quando representam alelos.
- A ordem relativa dos scaffolds ao longo da molécula não é conhecida.
- Quase todos os montadores fazem scaffolding.
- A estratégia típica para determinar os scaffolds é gulosa, incorporando informação por ordem de confiabilidade até que haja um conflito.
- Outras estratégias incluem informação de mapas físicos de DNA. Um mapa físico indica a posição relativa de marcadores moleculares bem definidos ao longo do DNA.

Validação

- Vários montadores incluem estratégias de validação das montagens.

Validação

- Vários montadores incluem estratégias de validação das montagens.
- Estratégias de validação incluem:

Validação

- Vários montadores incluem estratégias de validação das montagens.
- Estratégias de validação incluem:
  - Análises de consistência com base nas colunas do consenso.

Validação

- Vários montadores incluem estratégias de validação das montagens.
- Estratégias de validação incluem:
  - Análises de consistência com base nas colunas do consenso.
  - Uso de pares de reads para analisar a consistência dos contigs.

Validação

- Vários montadores incluem estratégias de validação das montagens.
- Estratégias de validação incluem:
  - Análises de consistência com base nas colunas do consenso.
  - Uso de pares de reads para analisar a consistência dos contigs.
  - Avaliação da distribuição dos reads ao longo da molécula para identificar repetições.

Validação

- Vários montadores incluem estratégias de validação das montagens.
- Estratégias de validação incluem:
  - Análises de consistência com base nas colunas do consenso.
  - Uso de pares de reads para analisar a consistência dos contigs.
  - Avaliação da distribuição dos reads ao longo da molécula para identificar repetições.
  - Uso de mapas físicos para analisar a consistência dos contigs.

Seqüência consenso

- Freqüentemente há interesse em produzir uma seqüência que represente o alinhamento múltiplo.

Seqüência consenso

- Freqüentemente há interesse em produzir uma seqüência que represente o alinhamento múltiplo.
- Para cada coluna i do alinhamento o *caractere-consenso* é o caractere do alfabeto que maximiza a soma da pontuação para todos os caracteres da coluna.

Seqüência consenso

- Freqüentemente há interesse em produzir uma seqüência que represente o alinhamento múltiplo.
- Para cada coluna i do alinhamento o *caractere-consenso* é o caractere do alfabeto que maximiza a soma da pontuação para todos os caracteres da coluna.
- A cadeia consenso é a concatenação dos caracteres-consenso para todas as colunas do alinhamento.

Seqüência consenso

- Freqüentemente há interesse em produzir uma seqüência que represente o alinhamento múltiplo.
- Para cada coluna i do alinhamento o *caractere-consenso* é o caractere do alfabeto que maximiza a soma da pontuação para todos os caracteres da coluna.
- A cadeia consenso é a concatenação dos caracteres-consenso para todas as colunas do alinhamento.
- A maioria é possível, mas não é considerada muito interessante.

Seqüência consenso

- O problema do alinhamento múltiplo pode ser redefinido em termos da seqüência consenso. Seja $s(i)$ a pontuação do caracter-consenso.

Seqüência consenso

- O problema do alinhamento múltiplo pode ser redefinido em termos da seqüência consenso. Seja $s(i)$ a pontuação do caracter-consenso.
- O alinhamento múltiplo ótimo de consenso ótimo é o alinhamento múltiplo para o qual o somatório $\sum_i s(i)$ é máximo.