Montagem de DNA

Dados 500 a 2000 fragmentos de DNA, com tamanho de 200 a 700 bases, encontrar a sequência alvo, ou seja, a sequência que contenha todos os fragmentos. Em casos reais, uma sequência alvo tem tamanho de 30.000 a 100.000 pares de bases.

- Sequenciar uma molécula de DNA é determinar o string de bases que ele contém. A sequência de DNA a ser sequenciada é chamada sequência alvo.

- Sabe-se aproximadamente (com 10% de erro) o tamanho da sequência alvo.

- É impossível sequenciar toda a sequência alvo de uma vez. Ao invés disso, pega-se pedaços dessa molécula, iniciando em posições escolhidas ao acaso, e sequencia-se estes pedaços. Casa uma dessas sequências é chamada fragmento. Note que, como o DNA é uma fita dupla, essa sequência pode ter a orientação $3\prime \longrightarrow 5\prime$ ou $5\prime \longrightarrow 3\prime$. Em qualquer dos casos, o fragmento é sequenciado como se fosse $5\prime \longrightarrow 3\prime$.

A montagem de fragmentos de DNA para obter a sequência alvo original é como um quebra-cabeças, quie consiste em tentar alinhar todos os fragementos de maneira a obter uma sequência, sendo que inicialmente não se sabe nem a posição e nem a orientação de cada fragmento.

O Caso Ideal

entrada:

ACCGT
CGTGC
TTAC
TACCGT

montagem:

--ACCGT--
----CGTGC
TTAC-----
-TACCGT--

consenso → TTACCGTGC

A partir de um conjunto de fragmentos, tenta-se alinhá-los para encontrar a sequência alvo. O alinhamento é baseado nas sobreposições (overlaps) entre os fragmentos. A sequência gerada é chamada de sequência de consenso, ou consenso, pois é obtida tomando-se os *votos* de cada fragmento para cada posição da sequência.

Complicações...

Erros:

- *substituições, inserções e deleções* de bases nos fragmentos (probabilidade de ocorrência: 1 a 5%);
- *chimeras*: fragmentos de partes distintas da molécula se unem para formar um fragmento único;
- *contaminação:* presença de fragmentos de outras moléculas de DNA, que não a molécula a ser sequenciada.

Chimeras e contaminação devem ser removidas antes da montagem...

Orientação desconhecida: como todos os fragmentos são sequenciados na orientação 5′ $\rightarrow$ 3′, pode-se ter fragmentos que estão na orientação *errada*. Neste caso, deve-se considerar o seu complementar para a montagem. Com n fragmentos, teríamos 2^n combinações para montar, o que inviável na prática...

Regiões repetidas: algumas moléculas tem sequências que aparecem muitas vezes na molécula alvo. Esse fato gera dois tipos de problemas:

- se um fragmento está totalmente contido em uma região repetida, existem *muitos lugares onde podemos alinhar este fragmento* (às vezes, as cópias não são exatamente iguais, e isso pode enfraquecer o consenso);

- o posicionamento de regiões repetidas pode deixar o *alinhamento ambíguo*...

Falta de cobertura: *Cobertura* de uma posição i é o número de fragmentos que contém esta posição. Se a cobertura for zero em uma ou mais posições, não existe informação suficiente nos fragmentos para montar a sequência alvo. Um *contig* é uma região coberta continuamente por fragmentos. O caso ideal da montagem é obter apenas um contig para um conjunto de fragmentos.

Modelos

Shortest Common Superstring:

Problema: SCS

Entrada: Uma coleção $\mathcal{F}$ de strings

Saída: O menor string possível S tal que para todo string $f \in \mathcal{F}$, S é superstring de f.

Este modelo *não considera erros* e nem orientação desconhecida. Além disso, repetições podem causar respostas equivocadas.

Este problema é *NP-difícil*, mas existem algoritmos aproximados. Este modelo é apenas de importância teórica, pois tem muitas limitações com relação ao caso biológico real.

Reconstrução:

Problema: Reconstrução

Entrada: Uma coleção $\mathcal{F}$ de strings e uma *tolerância* ε entre 0 e 1 (nível de erro aceitável).

Saída: O menor string possível S tal que para todo string $f \in \mathcal{F}$ tem-se

$$min(d_s(f, S), d_s(\overline{f}, S)) \leq \varepsilon |f|$$

onde $\overline{f}$ é o complemento de f, d_s é a *função de distância de substrings* $d_s(a, b) = min_{s \in S(b)} d(a, s)$, $S(b)$ é o conjunto de todos os substrings de b e d é a função de distância clássica (uma unidade é adicionada para cada inserção, deleção ou erro, exceto deleção na extremidade da segunda sequência).

A idéia é encontrar um string S onde cada f ou seu complemento seja um substring aproximado de S com nível de erro ε.

Este modelo leva erros e orientações desconhecidas em consideração, mas não modela repetições, falta de cobertura e tamanho do alvo. O problema é NP-difícil (contém SCS).

Multicontig:

Problema: Multicontig

Entrada: Uma coleção $\mathcal{F}$ de strings, um inteiro $t \geq 0$, e uma tolerância de erros ε entre 0 e 1.

Saída: Uma partição de $\mathcal{F}$ no menor número de coleções C_i, $i \leq k$, tal que cada C_i admite um t-contig com ε-consenso.

Um *t-contig* é um conjunto de fragmentos alinhados onde a menor ligação (sobreposição que não está contida em outros fragmentos) tem t posições.

Um string S é um *ε-consenso* quando a distância de edição de substrings entre cada fragmento alinhado e sua imagem no consenso é no máximo $\varepsilon|f|$.

Este problema é NP-difícel (contém o problema do caminho Hamiltoniano).