Trabalho Prático: Montagem de Genomas

Profª Janaína Rolan Loureiro

Estruturas de Dados e Programação

FACOM/UFMS

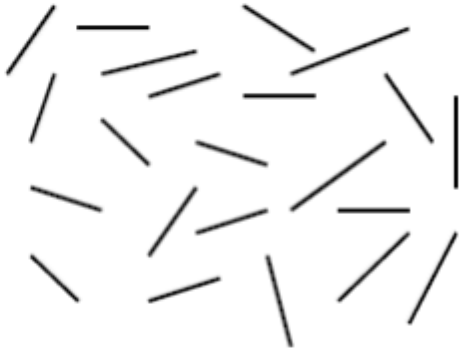
Preparação das Amostras

Extração

DNA

DNA original

Fragmentos

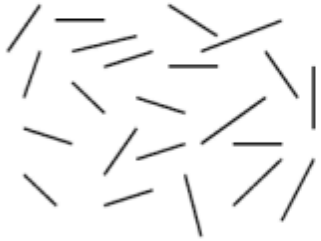


SEQUENCIADOR

Obs: Dependente da tecnologia utilizada

Sequenciamento

Fragmentos



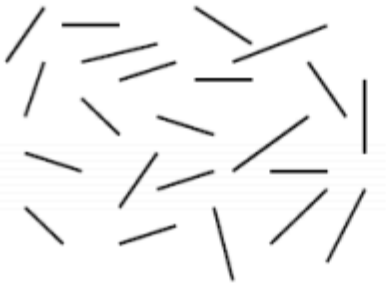
Sequenciar os fragmentos

Leituras (*reads*)

Etapas de sequenciamento de genomas

Montagem

Leituras



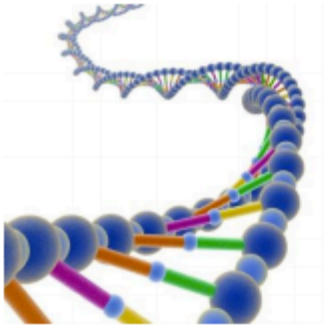
DNA

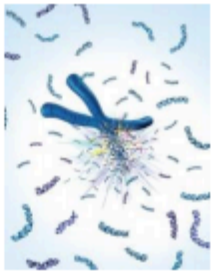
Bioinformática

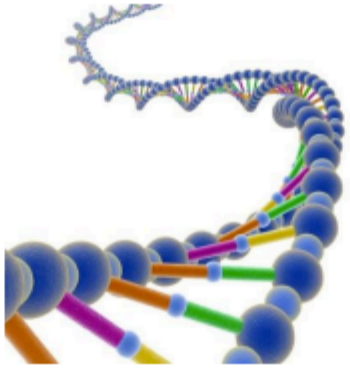
Programas Montadores

Etapas de sequenciamento de genomas

Montagem







Etapas de sequenciamento de genomas

Montagem

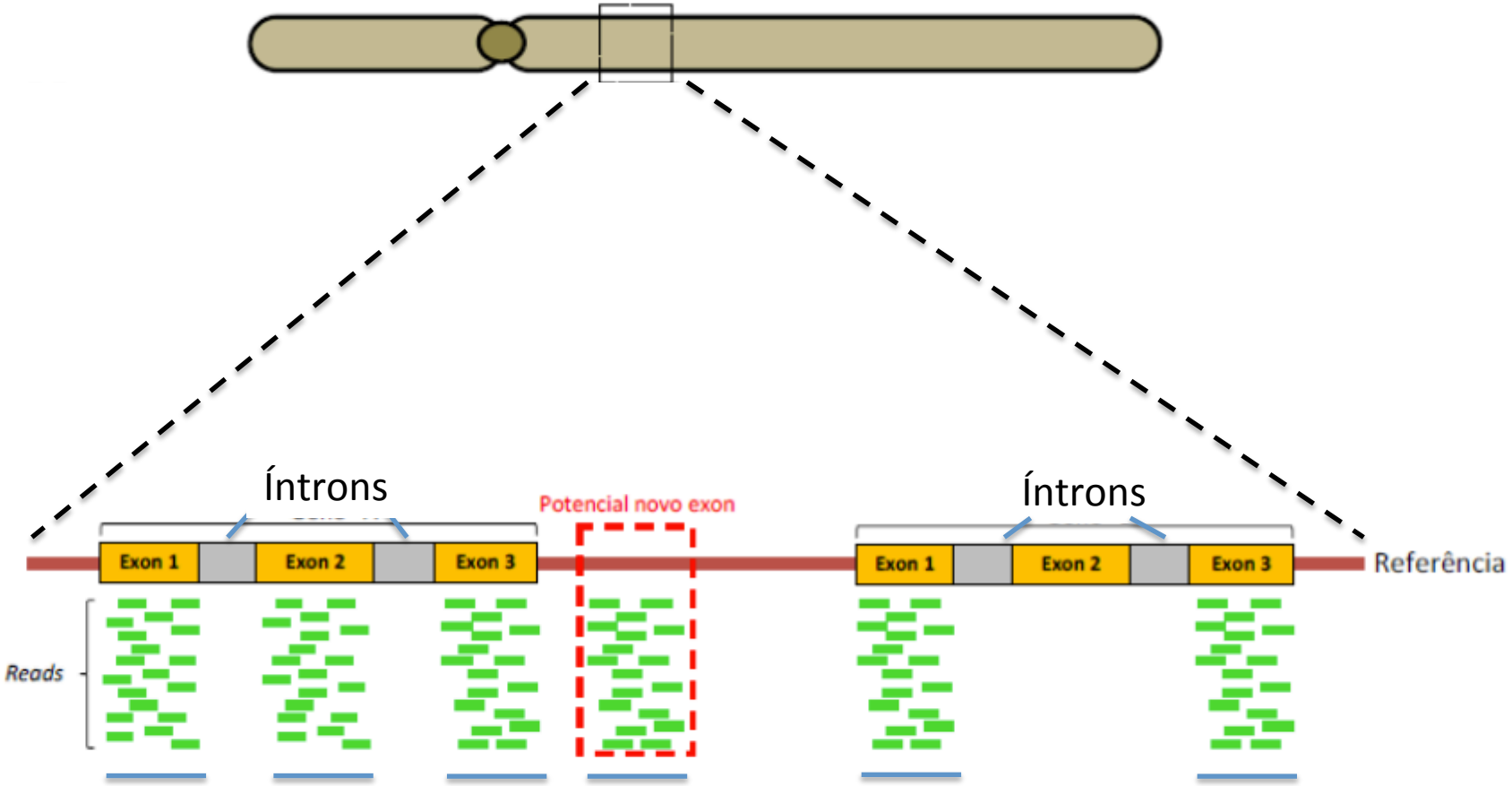
Aligned reads

```
ACGCGATTCAGGTTACCACG
   GCGATTCAGGTTACCACGCG
      GATTCAGGTTACCACGCGTA
         TTCAGGTTACCACGCGTAGC
            CAGGTTACCACGCGTAGCGC
               GGTTACCACGCGTAGCGCAT
                  TTACCACGCGTAGCGCATTA
                     ACCACGCGTAGCGCATTACA
                        CACGCGTAGCGCATTACACA
                           CGCGTAGCGCATTACACAGA
                              CGTAGCGCATTACACAGATT
                                 TAGCGCATTACACAGATTAG

ACGCGATTCAGGTTACCACGCGTAGCGCATTACACAGATTAG
```

Reads

Etapas de sequenciamento de genomas



Trabalho Prático

- Você foi contratado por uma empresa de sequenciamento genético para desenvolver uma ferramenta de montagem de genomas. Sua tarefa será, dados um conjunto *R* contendo os *reads* gerados pelo sequenciador e um gene de referência *g*, identificar quais são os éxons e íntrons deste gene.

Trabalho Prático

- É muito importante distinguir regiões de éxons (codificantes) de regiões de íntros (não-codificantes) em um sequenciamento de gene, porque a função biológica deles é também distinta. Enquanto os éxons codificam aminoácidos para a síntese protéica, principal processo relacionado ao DNA, os íntrons não participam deste procedimento, estando a sua real função ainda sob investigação.

Trabalho Prático

- Dados *R={r1, r2,…, rn}* e *g,* seu programa deve:

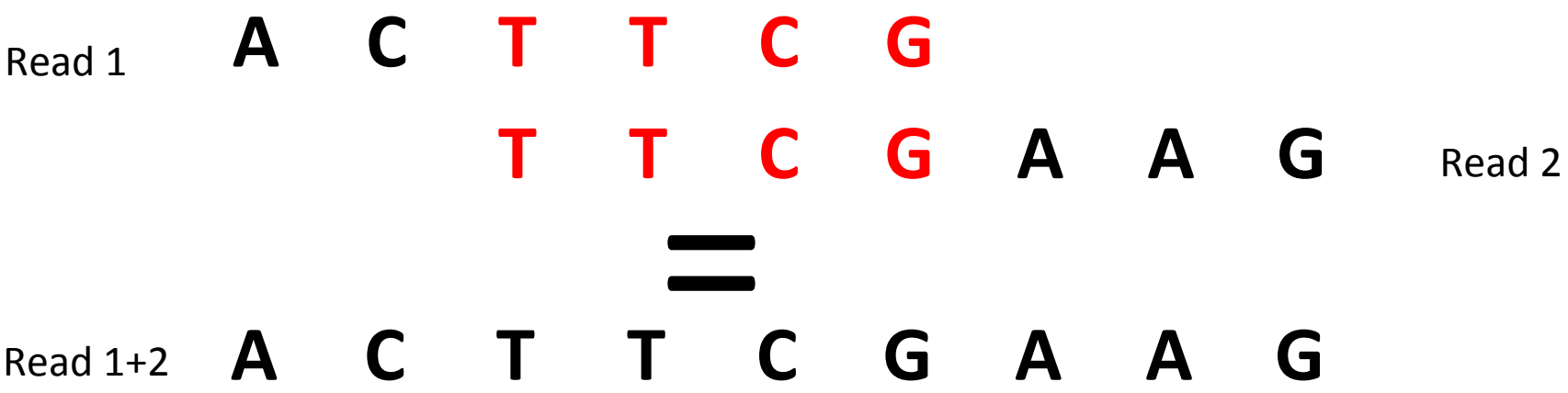
1. Encontrar o maior casamento possível entre **cada par de reads**, considerando o sufixo de um e o prefixo do outro:

Read 1 **A C T T C G**

T T C G A A G Read 2

Trabalho Prático

2. Selecionar o par que tenha o maior padrão em comum para unir as sequências, gerando um "read" maior. O intuito é encontrar *reads* que eram vizinhos na sequência original.



Trabalho Prático

2. Caso mais do que um par contenha padrões em comum do mesmo tamanho máximo, dê preferência para aqueles que gerem a **menor** sequência após unidos. Isso porque a chance de que eles fossem vizinhos inicialmente é maior, uma vez que a proporção da coincidência em relação ao seu tamanho é maior. Se os empates ainda persistirem, respeite a ordem de entrada dos *reads*, ou seja, escolha *reads* que foram informados primeiro.

Trabalho Prático

3. Retorne ao passo 1, agora com 1 *read* a menos, até que não seja mais possível unir nenhum dos pares.

4. Com os *reads* finais, faça o casamento deles com a sequência de referência. Por simplificação do problema biológico, você pode assumir que TODOS os reads finais irão casar exatamente uma vez com a sequência *g*.

Trabalho Prático

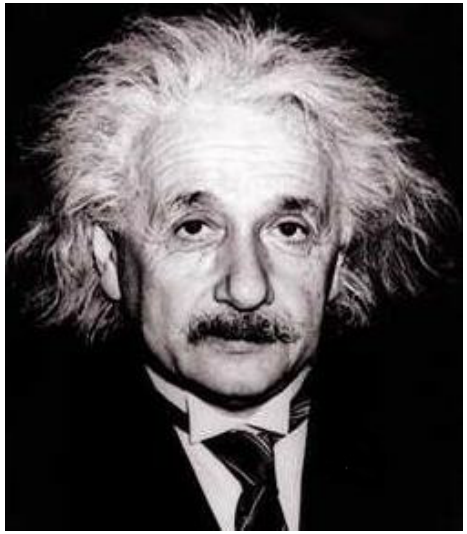
5. Por fim, retorne como éxons esses casamentos dos *reads* com *g* e como íntros os segmentos de *g* sem casamento. Indique as posições de início e final de cada éxon/íntron em *g,* assim como a sequência deles em si.

	Éxon 1				**Íntron 1**			**Éxon 2**			**Íntron 2**			**Éxon 3**	
g	A	T	G	C	A	T	A	C	G	C	G	C	T	A	T
reads	A	T	G	C				C	G	C			T	A	T

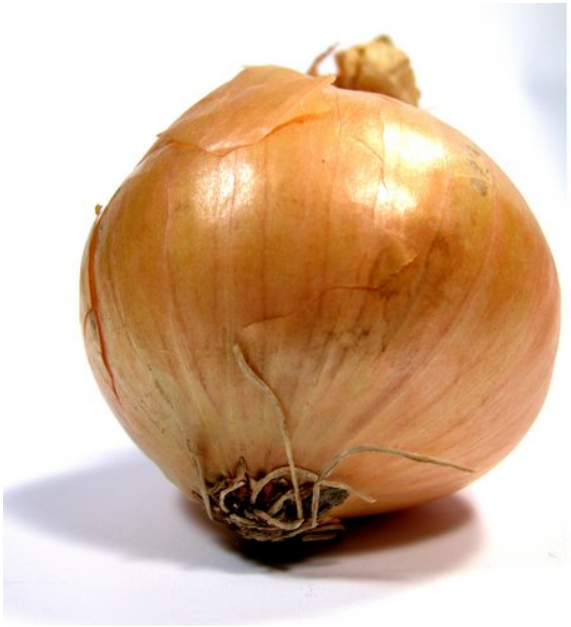
Observação

- Lembre-se de que quando lidamos com informações genéticas, as sequências a serem tratadas são extremamente grandes.
- Portanto, evitar o desperdício de memória e processamento é crucial em problemas da Bioinformática.
- Embora não seja exigido que seu programa suporte instâncias reais, evite que ele consuma tempo/memória além do necessário para resolver o problema, pois a eficiência também será avaliada.

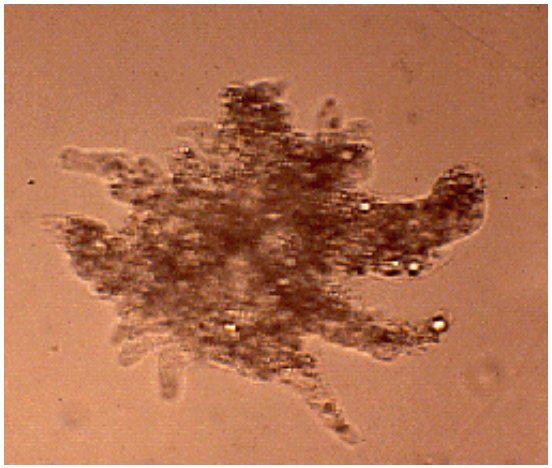
Tamanho do genoma



3,4 Gbp
Homo sapiens



15 Gbp
Allium cepa



680 Gbp
Amoeba dubia

Bp = par de bases, unidade de medida das sequências genéticas
Ex: ACTGACA = 7 bp

Recomendações

- O trabalho deve ser desenvolvido individualmente;
- Entrega do trabalho:
- Será realizada pelo sistema www.ead.facom.ufms.br no campo destinado;
- Apenas o arquivo .c do código deve ser postado;
- Data limite: 27/06/2014 às 23 horas e 55 minutos;
- Observação: Trabalhos entregues em local/hora diferentes dos pré-estabelecidos receberão nota ZERO.
- Plágios ou cópias acarretarão em nota ZERO para todas as partes envolvidas;
- Códigos que não compilem não serão corrigidos e receberão nota ZERO;
- Caso a professora julgue necessário, você poderá ser submetido a uma entrevista referente ao código entregue com data previamente combinada entre as partes;
- Este trabalho tem valor de 0 a 8.