Montagem de regiões gênicas

Pedro Ivo Gomes de Faria

Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo

Orientador: Prof. Dr. Alan Durham

IME - Instituto de Matemática e Estatística

Sumário

IME - Instituto de Matemática e Estatística

Sumário

IME - Instituto de Matemática e Estatística

Sumário

IME - Instituto de Matemática e Estatística

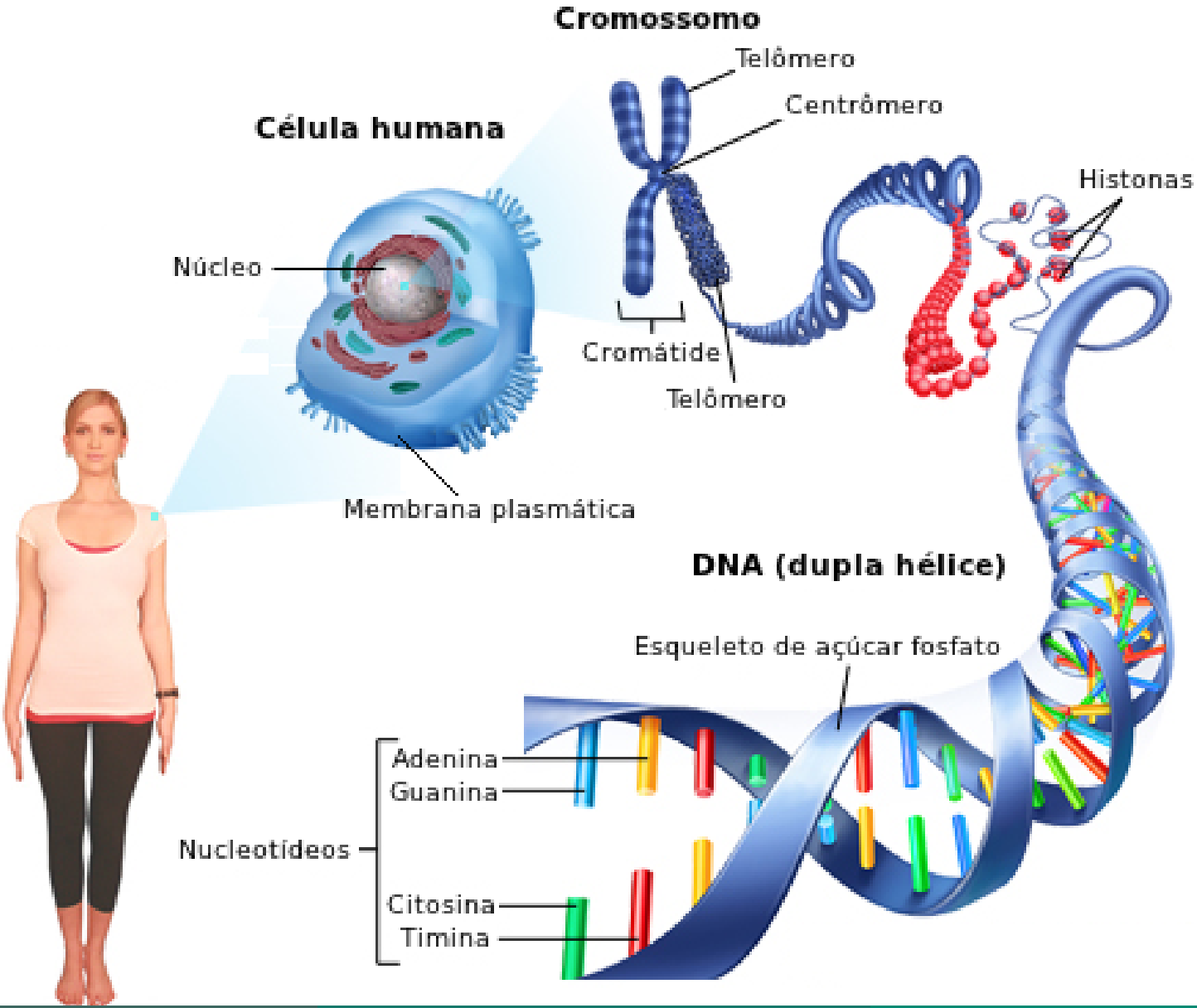
Classificação

Os seres vivos podem ser divididos em dois grupos (táxons):

- eucariotos: células com núcleo delimitado por uma membrana (animais, fungos, plantas);
- procariotos: células sem núcleo (bactérias, arqueias);

IME - Instituto de Matemática e Estatística

Localização e estrutura do DNA (em eucariotos)



IME - Instituto de Matemática e Estatística

Dogma central da biologia molecular

Descreve a transmissão e a expressão da hereditariedade, que ocorre utilizando três tipos de moléculas:

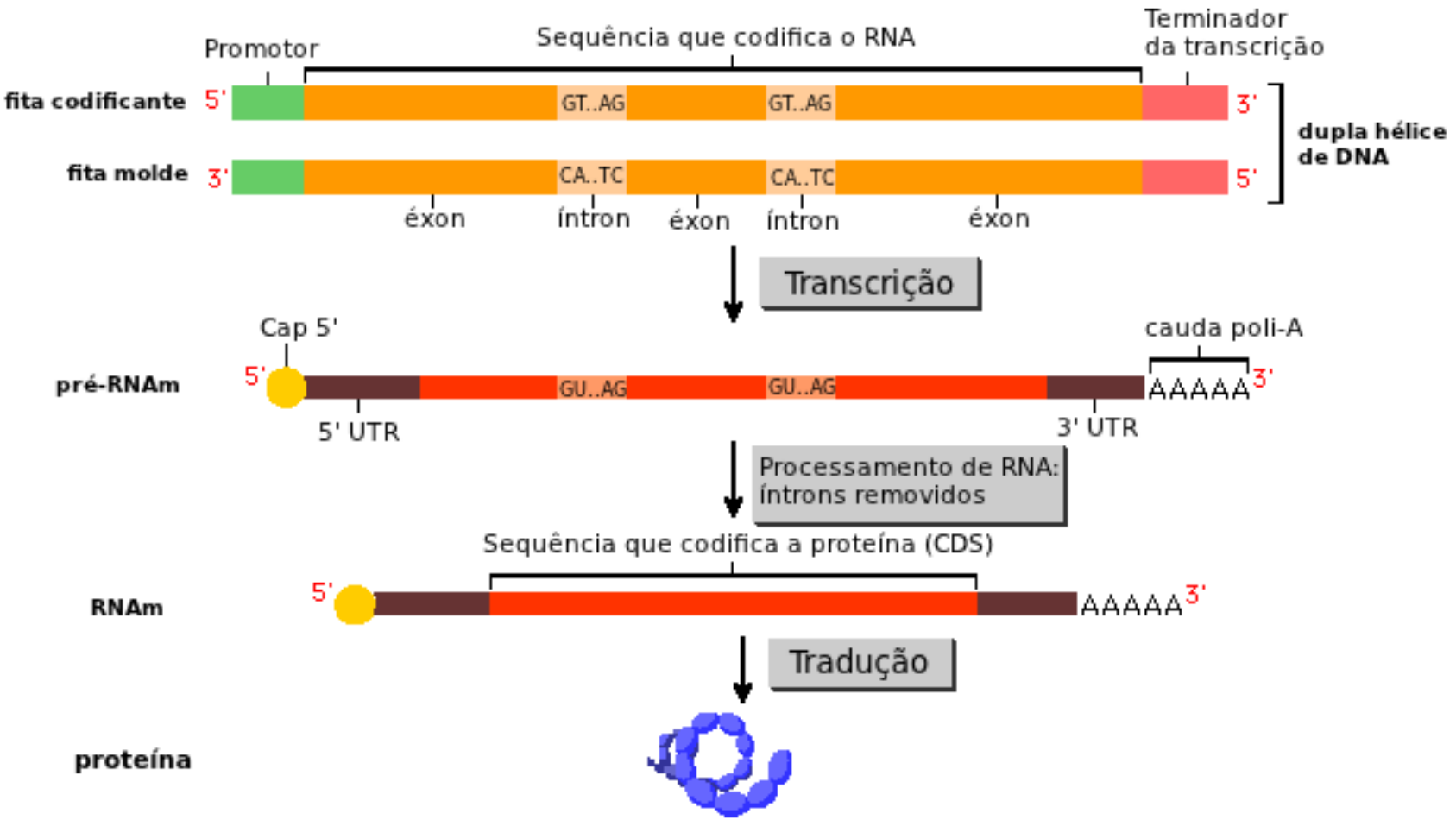
- DNA: armazena a informação;
- RNA: transfere a informação;
- proteína: executa uma função.

$$\text{duplicação} \circlearrowright \text{DNA} \xrightarrow{\text{transcrição}} \text{RNA} \xrightarrow{\text{tradução}} \text{proteína}$$

IME - Instituto de Matemática e Estatística

Estrutura dos genes (em eucariotos)

Gene: unidade molecular da hereditariedade; sequência de DNA que pode ser transcrita em uma versão de RNA.



Sumário

IME - Instituto de Matemática e Estatística

Motivação

- modo de organizar sequências de DNA, RNA, ou proteínas para identificar regiões de similaridade;
- regiões similares podem ser consequência de relações funcionais, estruturais ou evolutivas entre as sequências;
- processo que possui várias aplicações em bioinformática.

IME - Instituto de Matemática e Estatística

Alinhamento global

- compara sequências ao longo de toda a sua extensão;
- usado para identificar genes ou proteínas com funções similares, ou para estudar relações evolutivas entre sequências supostamente homólogas (i.e., que possuem um ancestral em comum).

Exemplo:

```
QUERIDA---ROSAVERMELHA
||||       |||  |||||
QUEROUMAMOROSOVERME---
```

Alinhamento local

- encontra *substrings* das sequências com alta similaridade entre si;
- utilizado para detectar padrões de nucleotídeos ou aminoácidos (domínios proteicos) conservados.

Exemplos (de dois alinhamentos locais entre as sequências `QUERIDAROSAVERMELHA` e `QUEROUMAMOROSOVERME`):

```
QUER          ROSAVERME
||||    e     ||| |||||
QUER          ROSOVERME
```

Alinhamento semiglobal

- encontra sufixos de uma sequência que tenham alta similaridade com prefixos de outra sequência;
- utilizado na montagem de sequências.

Exemplo:

```
---ROSAVERMELHA
   |||  |||||
AMOROSOVERME---
```

Medidas
Identidade e cobertura

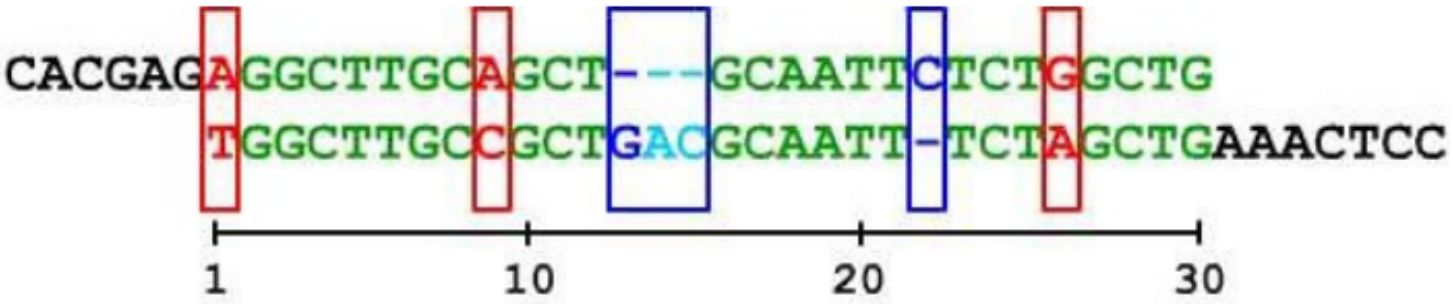
- identidade (do alinhamento): porcentagem de colunas idênticas do alinhamento;
- cobertura (de uma sequência): porcentagem de caracteres da sequência presentes na região alinhada.
- exemplo: para as sequências $s = \texttt{QUERIDAROSAVERMELHA}$ ($|s| = 19$), $t = \texttt{QUEROUMAMOROSOVERME}$ ($|t| = 19$) e para o alinhamento local $a = \begin{Bmatrix}\text{ROSAVERME}\\ \text{ROSOVERME}\end{Bmatrix}$, temos:
  - cobertura$(s, a) = |\texttt{ROSAVERME}| \,/\, |s| = 9/19 \approx 0,47$;
  - cobertura$(t, a) = |\texttt{ROSOVERME}| \,/\, |t| = 9/19 \approx 0,47$;
  - identidade$(a) = 8/9 \approx 0,89$.

IME - Instituto de Matemática e Estatística

Medidas
Pontuação (*score*)

- é uma medida da similaridade entre as sequências na região do alinhamento;
- colunas similares são “premiadas” (pontuação positiva) e colunas não similares são “penalizadas” (pontuação negativa);
- objetivo é encontrar um alinhamento de pontuação máxima.

Exemplo (para um alinhamento semiglobal):



3 x (-2) = -6
2 x (-4) = -8
2 x (-3) = -6
23 x (+1) = 23

3

Gap opening	-4 pontos
Gap extension	-3 pontos
Mismatch	-2 pontos
Match	+1 ponto

IME - Instituto de Matemática e Estatística

Sumário

IME - Instituto de Matemática e Estatística

Sumário

IME - Instituto de Matemática e Estatística

Motivação e definição

- genomas possuem tamanhos variando de milhões (bactérias) a bilhões (animais, plantas) de pares de bases (pb);
- tecnologia atual apenas consegue sequenciar fragmentos (*reads*) de cerca de 1000 pb (no máximo);
- o problema consiste em determinar a sequência original a partir dos fragmentos (processo chamado de montagem do genoma).

IME - Instituto de Matemática e Estatística

Modelagem

A modelagem pode ser feita pelo problema da *superstring* comum mais curta (versão de otimização):

- instância: um alfabeto finito Σ (= $\{$A,T,C,G$\}$) e um conjunto finito de *strings* $R \subset \Sigma^*$;
- solução viável: uma *string* $w \in \Sigma^*$ tal que cada *string* $x \in R$ seja uma *substring* de w (i.e, $\forall x \in R, \exists w_0, w_1 \in \Sigma^* : w = w_0 x w_1$) ;
- objetivo: minimizar o tamanho de w ($|w|$).

IME - Instituto de Matemática e Estatística

Sumário

IME - Instituto de Matemática e Estatística

Complicações teóricas

Considerando as classes de complexidade computacional, o problema:

- é NP-completo (versão de decisão), ou seja, está em NP e é NP-difícil;
- é APX-completo (versão de otimização), ou seja, está em APX (melhor aproximação conhecida - devido a Sweedyk - possui fator 2,5) e é APX-difícil (não possui PTAS - aproximação de fator $(1 + \epsilon), \forall \epsilon > 0$ - a menos que P = NP).

IME - Instituto de Matemática e Estatística

Sumário

IME - Instituto de Matemática e Estatística

Erros de sequenciamento

- Os métodos de sequenciamento geralmente identificam cada uma das bases através de sequências de reações químicas que geram algum sinal detectável;
- Interferências ou interpretações incorretas do sinal geram fragmentos com erros.

Sequência original:

```
TTACCGTGC
```

Entrada:

```
ACCGT
CGTGC
TTAC
TGCCGT
```

Saída:

```
--ACCGT--
----CGTGC
TTAC-----
-TGCCGT--
---------
TTACCGTGC
```

IME - Instituto de Matemática e Estatística

Orientação desconhecida

- Comumente apenas tentamos montar uma das fitas de cada molécula do DNA (pois a outra é obtida por complementaridade), mas os fragmentos podem ter vindo de qualquer uma das fitas;
- Logo, é necessário decidir (para cada fragmento) se sua versão original (→) ou a complementar (←) será utilizada na montagem.

Sequência original (ambas as fitas):

```
5′ CACGTAGTAC 3′
3′ GTGCATCATG 5′
```

Entrada:

```
5′ CACGT   3′
5′ ACGT    3′
5′ ACTACG  3′
5′ GTACT   3′
```

Saída:

```
→  5′  CACGT-----   3′
→  5′  -ACGT-----   3′
←  5′  --CGTAGT--   3′
←  5′  -----AGTAC   3′
   5′  CACGTAGTAC   3′
```

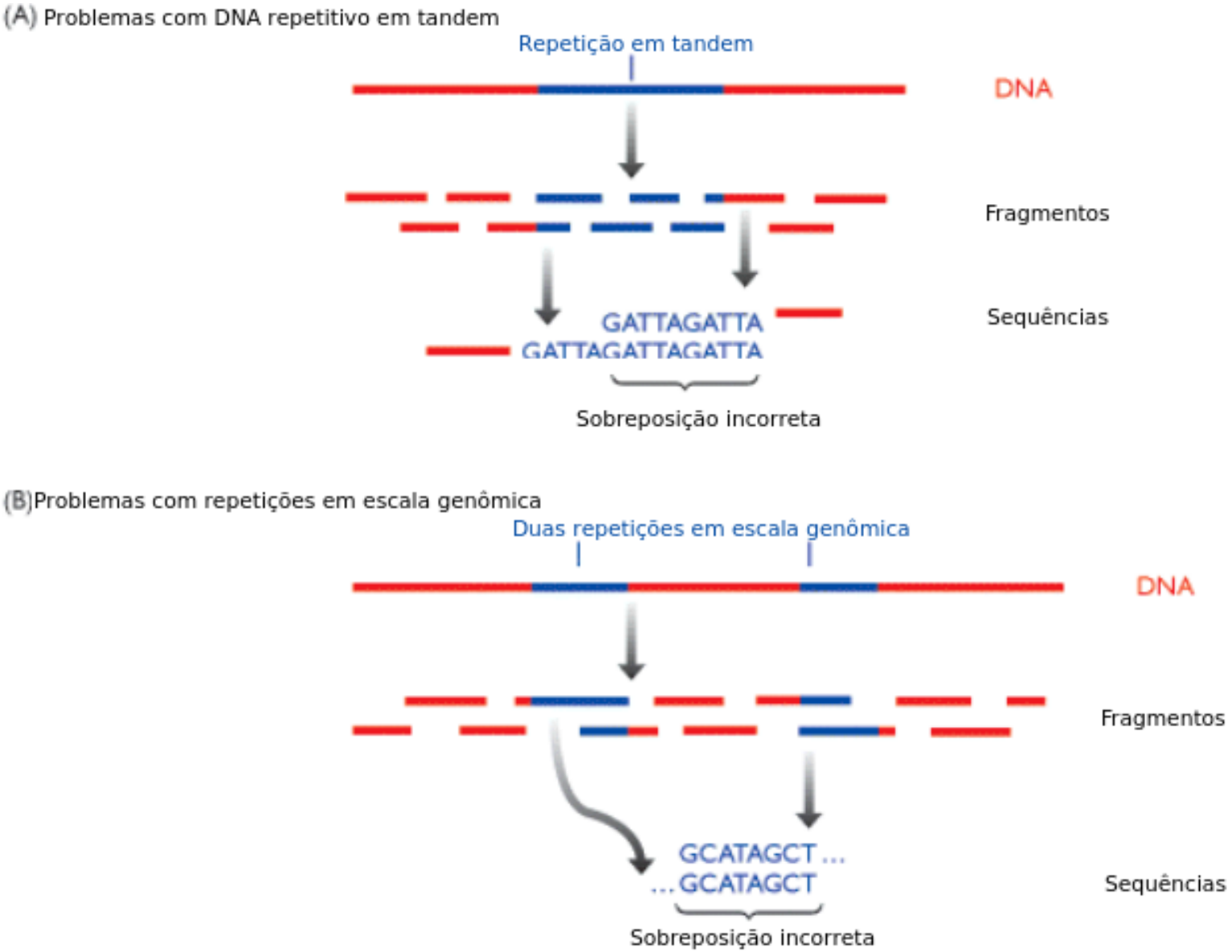
IME - Instituto de Matemática e Estatística

Regiões de repetição

- consistem de *substrings* que aparecem duas ou mais vezes na molécula de DNA original;
- são o maior desafio da montagem de genomas, principalmente em eucariotos;
- podem fazer com que fragmentos não contíguos possuam sobreposição em suas extremidades, o que faz com que sejam incorretamente sobrepostos durante a montagem (considerando a modelagem como o problema da *superstring* comum mais curta).

IME - Instituto de Matemática e Estatística

Problemas causados por regiões de repetição



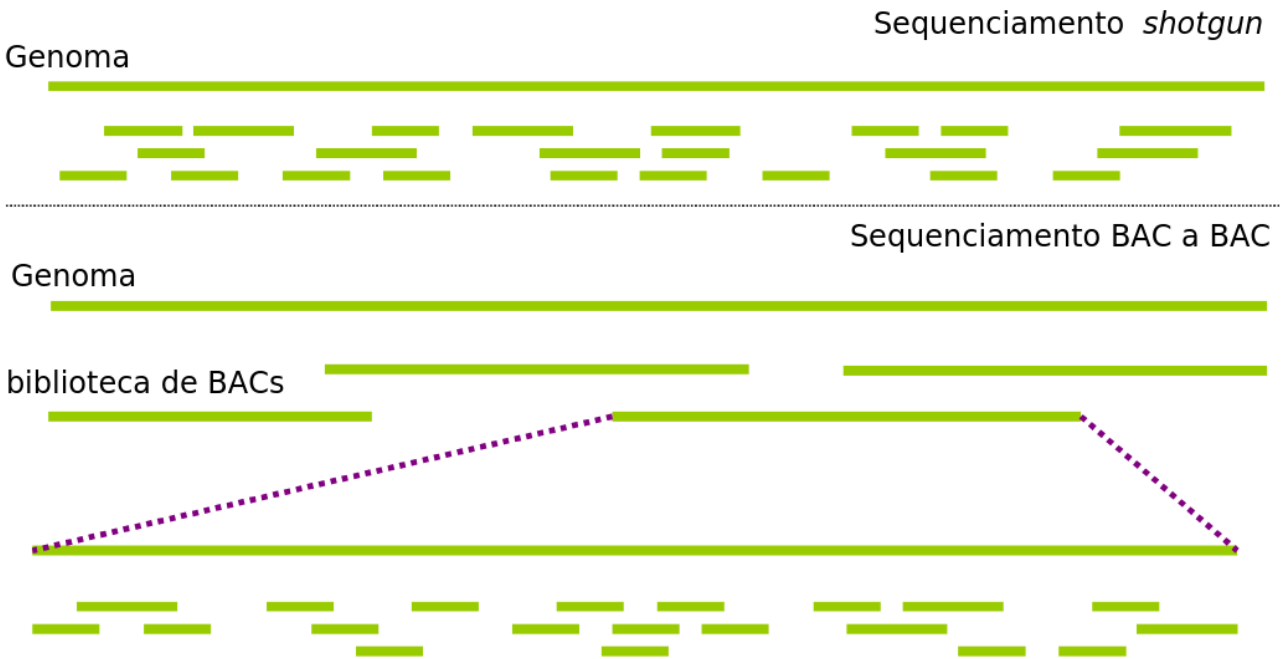
IME - Instituto de Matemática e Estatística

Sumário

IME - Instituto de Matemática e Estatística

Estratégias de sequenciamento

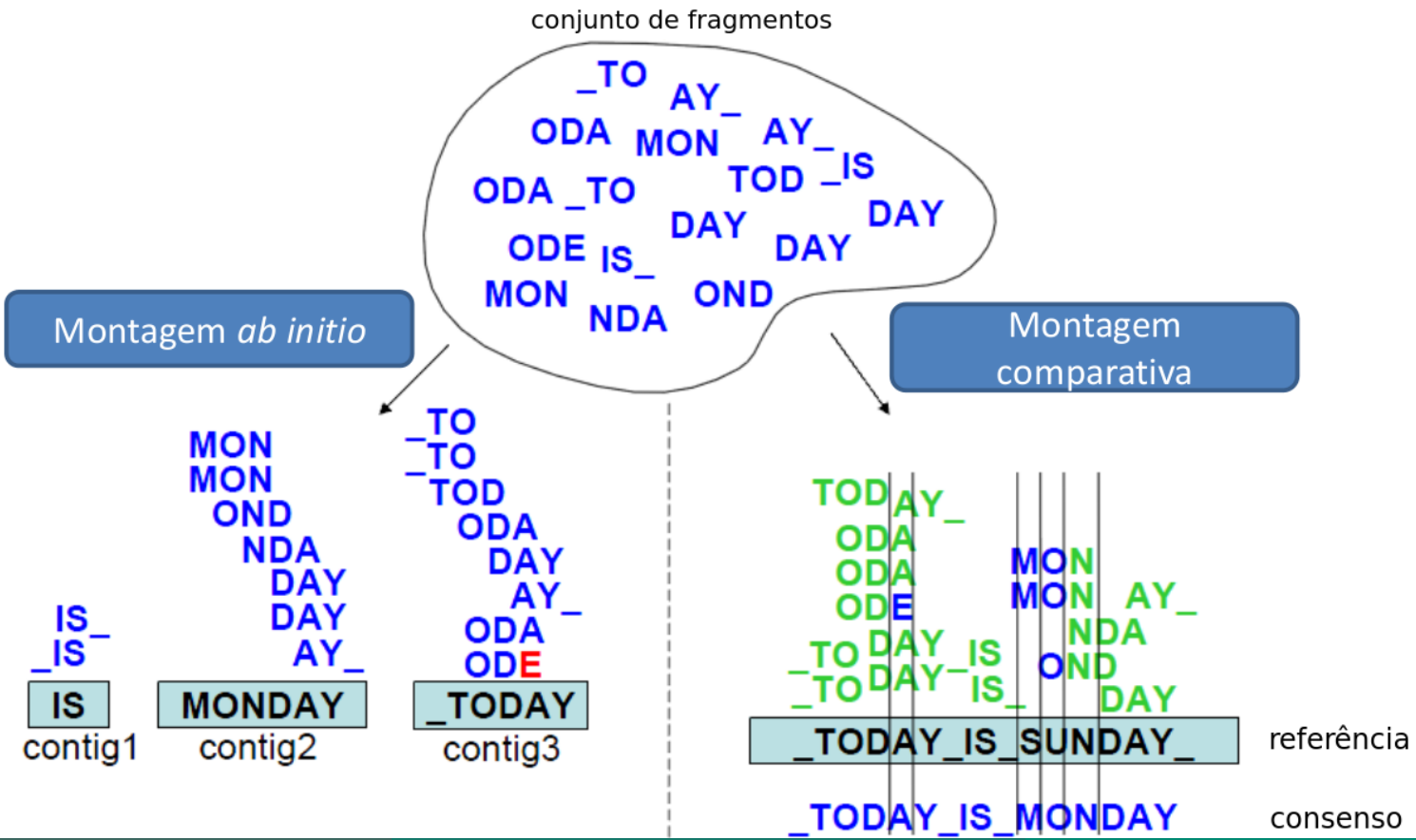
- em vez de gerar fragmentos para todo o genoma (sequenciamento *shotgun*), dividi-lo em partes menores, determinar qual a ordem entre essas partes e sequenciar cada uma individualmente (sequenciamento BAC a BAC);
- sequenciamento *shotgun* é rápido, barato e impreciso, enquanto sequenciamento BAC a BAC é lento, custoso e preciso.



IME - Instituto de Matemática e Estatística

Estratégias de montagem

- em vez de utilizar apenas o conjunto de fragmentos (ab initio), utilizar uma sequência de referência (montagem comparativa);
- sequência de referência deve possuir similaridade suficiente com a sequência que originou os fragmentos.



Sorgo: genoma modelo para cana-de-açúcar

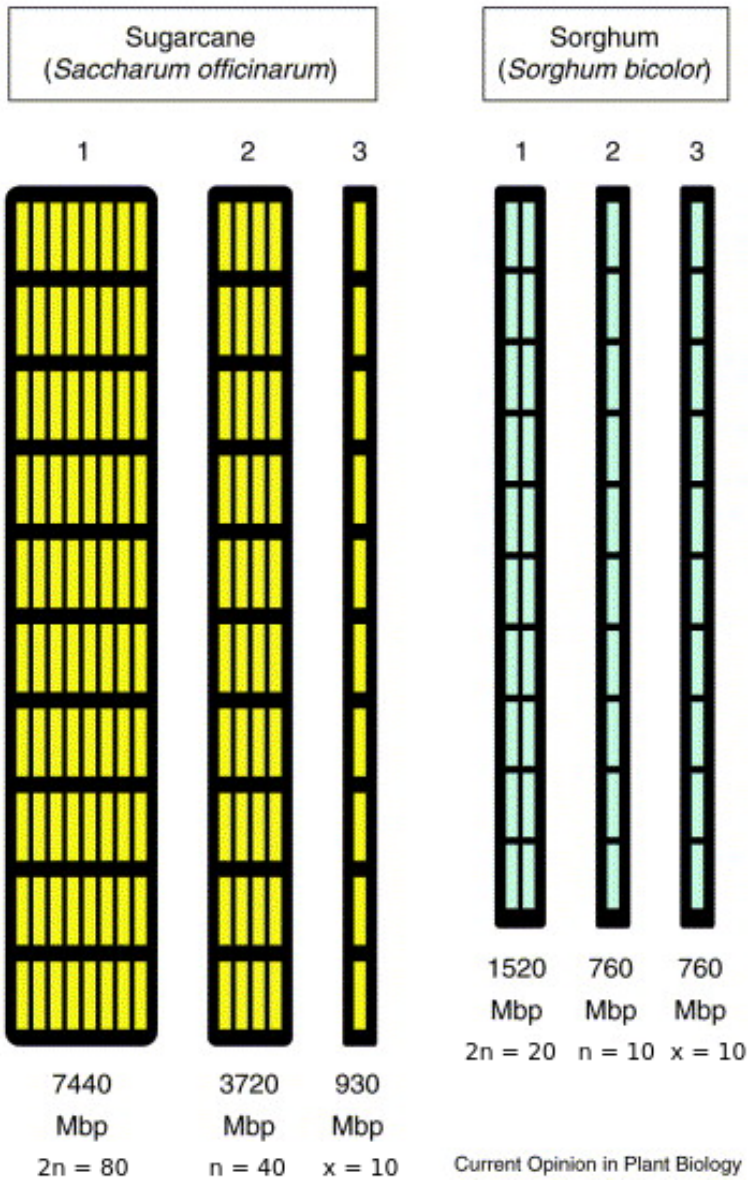
- sorgo (*S. bicolor*) é a planta cultivada mais próxima evolutivamente da cana-de-açúcar (*S. officinarum*);
- divergência entre ambas ocorreu há cerca de 5 milhões de anos;
- ambas pertencem à família das gramíneas (Poaceae), subfamília Panicoideae, tribo Andropogoneae.

Figura: Sorgo (*Sorghum bicolor*).

Figura: Cana-de-açúcar (*Saccharum officinarum*).

IME - Instituto de Matemática e Estatística

Comparação dos genomas de cana-de-açúcar e sorgo



- 1 - número diploide (2n): número de cromossomos numa célula somática.
- 2 - número haploide (n): número de cromossomos num gameta.
- 3 - número monoploide (x): número de cromossomos num conjunto não redundante.

IME - Instituto de Matemática e Estatística

Passos dos pipelines de montagem e validação desenvolvidos

Para cada conjunto de sequências de DNA de um BAC de cana-de-açúcar:

1. mascarar as sequências contaminantes (EGene);
2. alinhar as sequências (usando Blat) de DNA filtradas no proteoma de sorgo ($\approx$ 30000 proteínas);
3. montar separadamente (usando Newbler) cada conjunto de sequências que se alinhou na mesma proteína;
4. tentar estender o máximo possível (usando Genseed) cada conjunto montado anteriormente, utilizando as sequências que não foram mapeadas nas proteínas de sorgo;
5. alinhar os *contigs* gerados na sequência já disponível do BAC (supostamente correta);
6. verificar a ocorrência de quimeras (i.e., *contigs* montados incorretamente) via cobertura dos *contigs* e identidade dos alinhamentos.

IME - Instituto de Matemática e Estatística

Sumário

IME - Instituto de Matemática e Estatística

Resultados

Para o BAC `SHCRBa_218_D04`, os resultados foram:

nome do contig	nome da prot.	tam. reg. 5' (pb)	tam. reg. 3' (pb)	ident. na prot.	cobert. no BAC	ident. no BAC
C_1	P_1	708	X	0,99	0,95	1
C_2	P_2	515	X	0,90	1	0,99
C_3	P_2	X	X	0,97	0,99	1
C_4	P_2	X	266	0,92	0,98	0,99
C_5	P_3	20	X	0,96	1	1
C_6	P_4	282	X	0,93	0,94	1
C_7	P_5	740	X	0,96	0,97	0,99
C_8	P_5	X	855	0,96	0,93	0,99
C_9	P_6	940	X	0,94	0,93	0,99
C_{10}	P_6	X	605	0,96	1	0,99
C_{11}	P_7	618	450	0,96	0,91	0,99

IME - Instituto de Matemática e Estatística

Conclusão

Para os 6 BACs que puderam ser montados e validados:

- aproximadamente 70% das regiões gênicas puderam ser estendidas em algum sentido;
- de modo geral, não houve ocorrência de quimeras.

Logo, o *pipeline* poderia ser utilizado como uma forma razoavelmente confiável (embora limitada) de montar regiões gênicas, com alguma chance de conseguir estender a montagem até a região promotora do gene.

Referências I

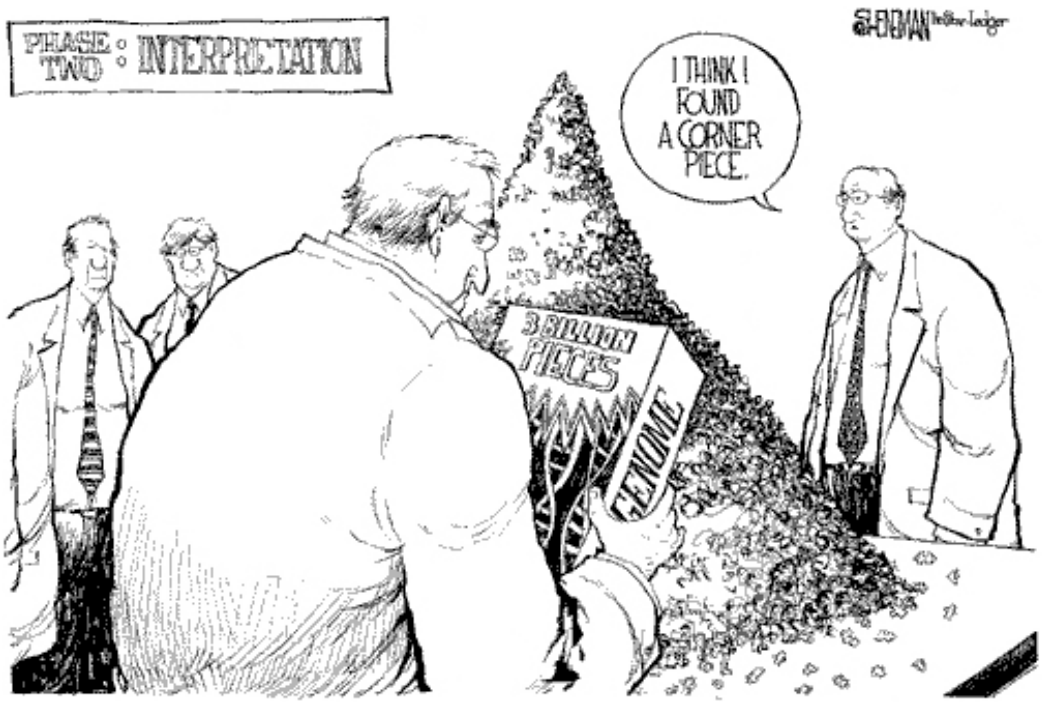
SETUBAL, J.; MEIDANIS, J.
Introduction to computational molecular biology.
1ª ed.
Boston: PWS, 1997.

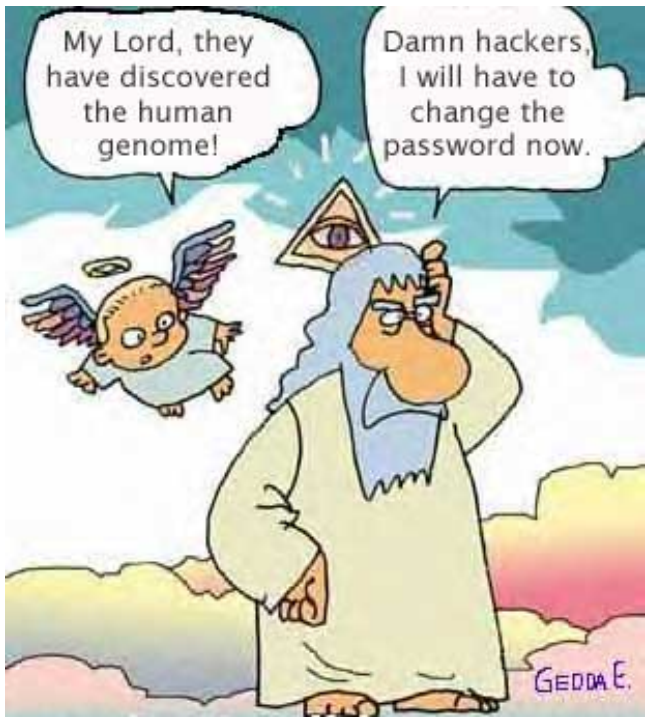
ALBERTS, B. et al.
Biologia molecular da célula.
5ª ed.
Porto Alegre: Artmed, 2010.

GRIVET, L.; ARRUDA, P.
Sugarcane genomics: depicting the complex genome of an important tropical crop.
Current Opinion in Plant Biology 5:122–127, 2001.

IME - Instituto de Matemática e Estatística

Obrigado!





IME - Instituto de Matemática e Estatística