

Computação para análise de genomas

Waldeyr Mendes Cordeiro da Silva

Faculdade Cambury 2018

Agenda

1. Conceitos Básicos de Biologia Molecular

Aspectos biológicos do curso

2. Comparação de sequências

Tipos de alinhamentos

3. Dados de sequenciamento de alto desempenho

Formatos de arquivos

❖ Filtragem e montagem de fragmentos

Controle de qualidade

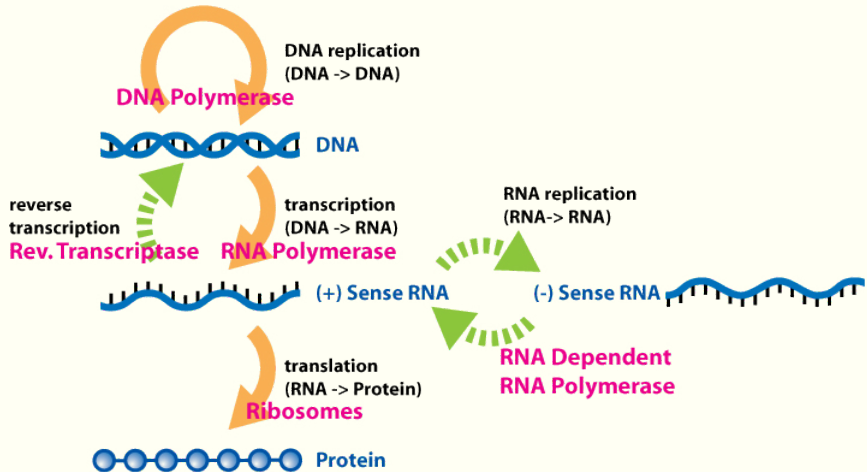
Montagem *de novo*

❖ Anotação

Significado biológico das sequências montadas

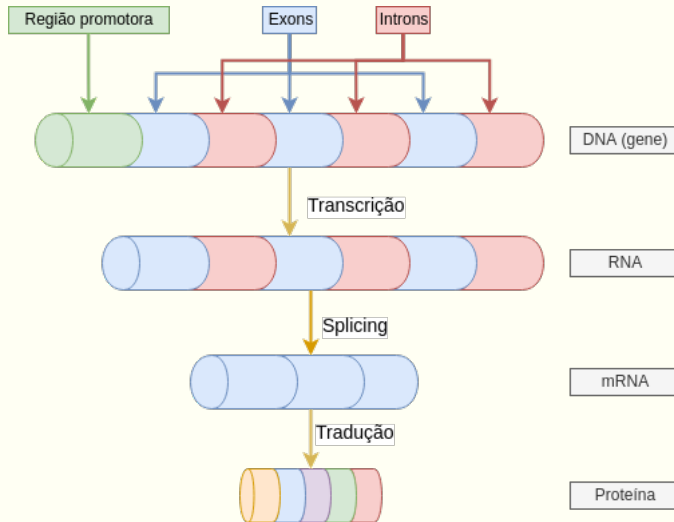
Conceitos Básicos de Biologia Molecular

Dogma

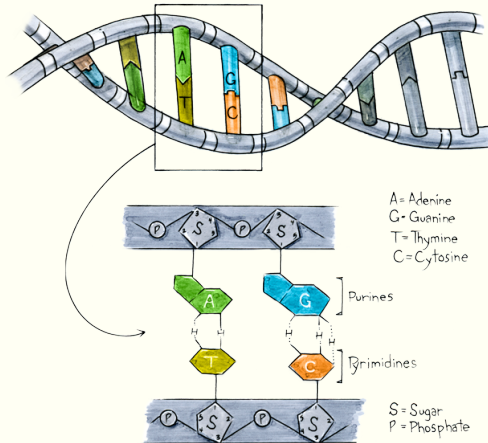


Fonte: https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

Dogma



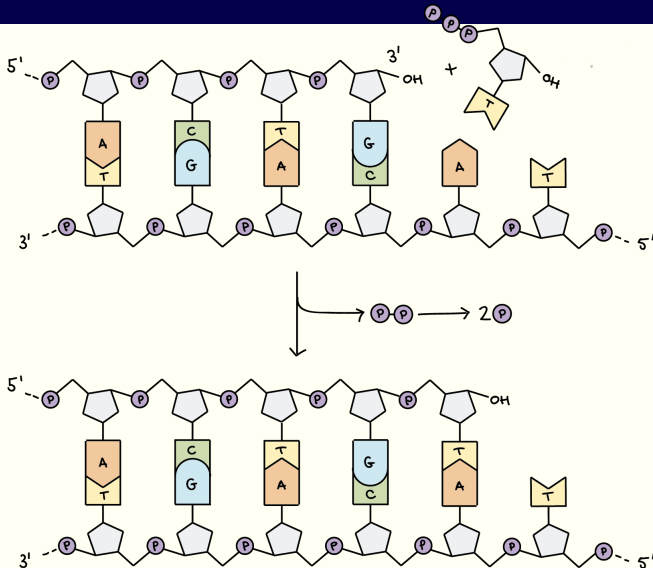
DNA



Copyright © 2012 University of Washington

Fonte: <https://www.my46.org/intro/what-is-dna>

DNA



Fonte: <https://www.khanacademy.org>

Sequenciamento de DNA

- ❖ Obter string(s) representando as moléculas que compõem o DNA
- ❖ Ainda não é possível sequenciar toda a molécula diretamente
- ❖ Sequenciar pedaços da molécula, começando em alguma posição na direção $5' \rightarrow 3'$
- ❖ Fragmento (*read*): substring de uma das fitas da molécula alvo de DNA
- ❖ Não sabemos:
 - ❖ A que fita pertence
 - ❖ A posição relativa ao início da fita

Comparação de sequências

Alinhamento

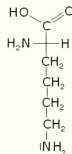
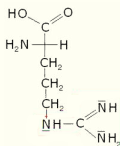
Posicionamento das sequências, preservando a ordem dos nucleotídeos ou aminoácidos e indicando as posições em que as sequências são iguais ou diferentes

- ❖ Ferramenta básica da Bioinformática
- ❖ Alfabeto
 - ❖ DNA/RNA - 4 nucleotídeos (ACGT/ACGU)
 - ❖ Proteínas - 20 aminoácidos (A, R, N, D, E, C, G, Q, H, I, L, K, M, F, P, S, Y, T, W, V)
- ❖ Interesse no alinhamento ótimo: o máximo de similaridade e o mínimo de diferenças

Alinhamento

- ❖ Identidade → Porcentagem de aminoácidos (ou nucleotídeos) com um *match* direto no alinhamento
- ❖ Similaridade → Porcentagem de *matches* idênticos e similares (substituição conservativa)

Exemplo: arginina ↔ lisina



- ❖ Homologia → Similaridade entre sequências que dividem ancestral comum

Alinhamento

Exemplo

ROSAVERMELHA
AMOROSOVERME

8% de identidade (1 em 12).

Alinhamento

Exemplo

⊖ ⊖ ⊖ ROSA VERMELHA
AMO ROSOVERME ⊖ ⊖ ⊖

53% de identidade (8 em 15).

Alinhamento

Tipos de alinhamento

Quanto à quantidade de entradas

- a) Pairwise - pareamento de 2 sequências
- b) Alinhamento múltiplo - múltiplas sequências

Quanto à estratégia de alinhamento

- a) Global
- b) Local

Quanto ao tipo de entrada

DNA x RNA x Proteína

Alinhamento

Erros

-	-	A	C	C	G	T	-	-
-	-	-	-	C	G	T	G	C
T	T	A	C	-	-	-	-	-
-	T	A	G	C	G	T	-	-

erro de substituição C/G

T T A C C G T G C

consenso: votação da maioria

Alinhamento de sequências

Erros

-	-	A	C	C	-	G	T	-	-
-	-	-	-	C	A	G	T	G	C
T	T	A	C	-	-	-	-	-	-
-	T	A	G	C	-	G	T	-	-
T	T	A	C	C	-	G	T	G	C

erro de inserção de A

consenso: votação da maioria
impressão: - não aparece

Alinhamento de sequências

Erros

-	-	A	C	C	G	T	-	-
-	-	-	-	C	G	T	G	C
T	T	A	C	-	-	-	-	-
-	T	A	C	-	G	T	-	-

erro de remoção de C
no último fragmento

T T A C C G T G C

consenso: votação da maioria

Alinhamento de sequências

Modelos de pontuação

- ❖ Substituições
- ❖ Gaps (inserções/deleções)
- ❖ Matriz de substituição

Alinhamento

Modelos de pontuação

- ❖ Tomando as sequências: GACGGATTAG e GATCGGAATAG
- ❖ *Match* = +1
- ❖ *Mismatch* = -1
- ❖ *Gap* = -2

G	A	-	C	G	G	A	T	T	A	G
G	A	T	C	G	G	A	A	T	A	G
<hr/>										
+1	+1	-2	+1	+1	+1	+1	-1	+1	+1	+1

- ❖ Obs: Valores das penalidades podem ser escolhidos

Alinhamento

Pairwise

- ❖ Alinhamento global (algoritmo de Needleman-Wunsch)
- ❖ Alinhamento local - *substrings* - (algoritmo de Smith-Waterman)
- ❖ Alinhamento semi-global - *alinhar prefixos e sufixos*

Alinhamento

Alinhamento global

J. Mol. Biol. (1970) **48**, 443–453

A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins

SAUL B. NEEDLEMAN AND CHRISTIAN D. WUNSCH

*Department of Biochemistry, Northwestern University, and
Nuclear Medicine Service, V. A. Research Hospital
Chicago, Ill. 60611, U.S.A.*

(Received 21 July 1969)

Alinhamento

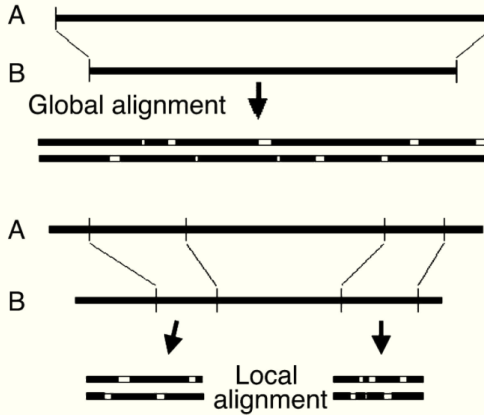
Alinhamento local

J. Mol. Biol. (1981), **147**, 195–197

Identification of Common Molecular Subsequences

The identification of maximally homologous subsequences among sets of long sequences is an important problem in molecular sequence analysis. The problem is straightforward only if one restricts consideration to contiguous subsequences (segments) containing no internal deletions or insertions. The more general problem has its solution in an extension of sequence metrics (Sellers 1974; Waterman *et al.*..

Alinhamento



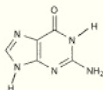
Penacchio et al. 2003

Alinhamento

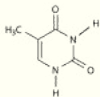
Matrizes de substituição



Adenina (A)



Guanina (G)



Timina (T)



Citosina (C)



Uracila (U)

	A	C	G	T
A	+20	+5	+10	+5
C	+5	+20	+5	+10
G	+10	+5	+20	+5
T	+5	+20	+5	+10

Alinhamento

PAM (Percent Accepted Mutation) (Dayhoff et al., 1972)

- ❖ Baseada em alinhamento global de proteínas relacionadas alinhadas sem gaps ()mutações são muito significantes)
- ❖ Probabilidade de um aminoácido ser substituído por outro (eventos independentes HMM)
- ❖ PAM1: 1% de mutações aceitas ou seja as sequências tem 99% de similaridade (aminoácidos idênticos)
- ❖ PAM250: PAM1 x PAM1, 250 vezes - representando milhões de anos de evolução: melhor para sequências menos relacionadas

Alinhamento

Blast

- ❖ Basic Local Alignment Search Tool (Altschul *et al.*, 1990)
 - ❖ Abordagem heurística
 - ❖ Acha palavras curtas e monta tabela hash
 - ❖ Procura palavras curtas em um banco de dados e estende-as quando há um HSP (*High Scoring Segment Pair*)
 - ❖ Calcula estatística do alinhamento e interrompe-o quando o *e-value* torna-se menor do que um determinado limite

Dados de sequenciamento de alto desempenho

Formatos (FASTQ)

```
@SEQ_ID
TTCAACTCGTTAGTAAATATCAAACGATCAGTACCATTTTGGGGTTCAAAGTGACAGTTT
+
!'>>>CCC'*(((((***(***-+*'')+))%%%+)))*55CCF>>%%%%).1CCCC65
```

Exemplo Illumina: **@HWUSI-EAS100R:6:73:941:1973#0/1**

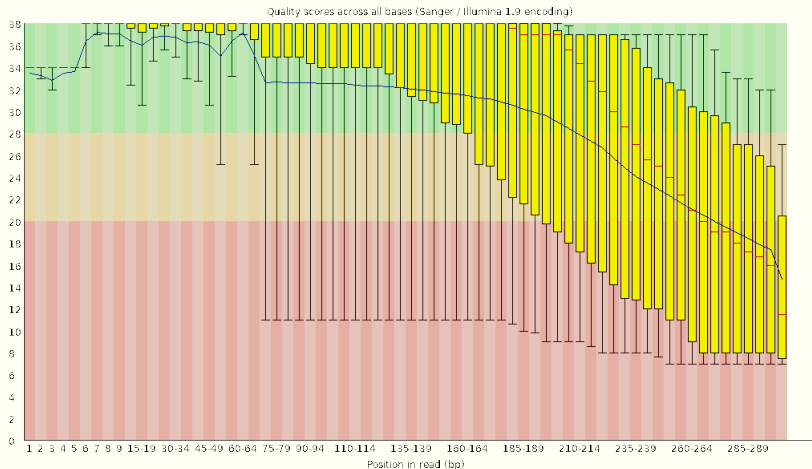
- ❖ HSWUSI-EAS100R → Unique instrument name
- ❖ 6 → Flowcell lane
- ❖ 73 → Tile number within the flow cell lane
- ❖ 941 → x-coordinate of the cluster within the tile
- ❖ 1973 → y-coordinate of cluster within the tile
- ❖ #0 → Index number for multiplexed sample
- ❖ /1 → Member of a pair

Formatos (FASTQ)

A qualidade (varia de 33 a 126) de cada nucleotídeo sequenciado é representado pelo caractere correspondente da tabela ASCII. Os valores *shifted down* para 0 a 93 por compatibilidade com a escala PHRED de qualidade (0 a 60)

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Formatos (FASTQ)



Formatos (FASTA)

```
>gi|13959657|sp|Q9PTU8|VSP3_B0TJA Venom serine proteinase A precursor  
MVLIRVIANLLILQLSNAQKSSELVIGGDECNITEHRFLVEIFNSSGLFCGGTLIDQEWVLSAAHCDMRN  
MRIYLGVHNEGVQHADQQRRFAREKFFCLSSRNYTKWDKDIMLIRLNRPVNNSEHIAPLSLPSNPPSVGS  
VCRIMGWTITSPNATFPDVPHCANINLFNYTVCRGAHAGLPATSRTLCAQVLQGGIDTCGGDSGGPLIC  
NGTFQGIVSWGHHPCAQPGEPALYTKVFDYLPWIIQSIIAGNTTATCPP
```

1. Cabeçalho

- ❖ GenBank/EMBL → gi|gi_number|*|accession.version|locus
- ❖ NCBI refseq → ref|accession|locus
- ❖ PRF Protein Research Foundation → pir|entry
- ❖ SWISS-PROT → sp|accession|locus
- ❖ PDB Protein Data Bank → pdb|entry|chain

2. Sequência

- ❖ nucleotídeos ou aminoácidos

Formatos (FASTA)

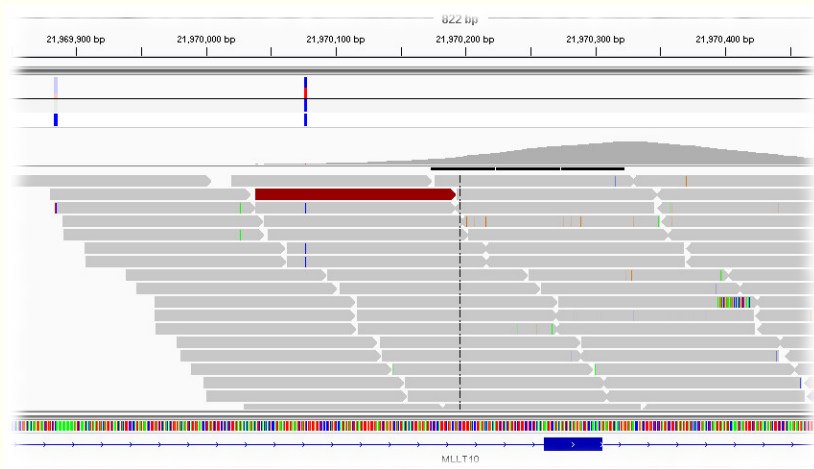
- ❖ .fasta, .fa → arquivo fasta genérico
- ❖ .fna → FASTA nucleotídeos
- ❖ .ffn → FASTA regiões codificadoras (nucleotídeos)
- ❖ .faa → FASTA aminoácidos
- ❖ .frn → FASTA RNA não codificador
- ❖ Multi-fasta → múltiplas sequências em um único arquivo

Formatos (SAM, BAM)

SAMTools fazem pós-processamento de alinhamentos de *reads*, as quais são sequências de DNA em formato FASTQ.

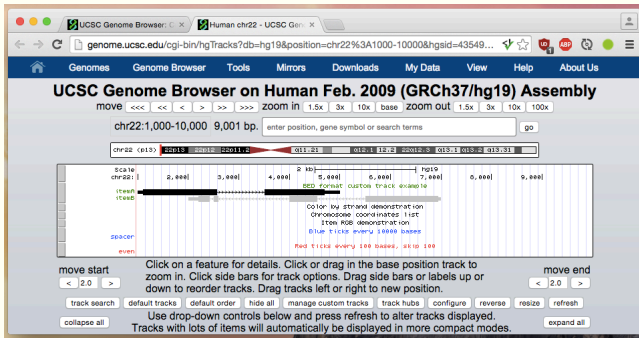
- ❖ SAM (Sequence Alignment/MAP) guarda o alinhamento das *reads* e pode ser lido por diversos softwares como o IGV (Integrated Genome Viewer).
- ❖ BAM (Binary Alignment/MAP) é uma versão comprimida de um alinhamento das *reads*. Pode ser obtido diretamente do alinhamento ou convertido a partir de um arquivo SAM.

Formatos (SAM, BAM)



Formatos (BED)

- ❖ BED é um arquivo organizado em colunas separadas por tabulação (tab) com anotações da sequência
- ❖ Pode ser aberto em um genome browser



Formatos (BED)

- ❖ Arquivos BED têm 12 colunas, 1-3 obrigatórias, 4-12 opcionais

1. **chrom** → nome do cromossomo no qual a *feature* existe
2. **start** → posição inicial na sequência
3. **end** → posição final na sequência
4. **name** → nome da *feature*
5. **score** → 0 and 1000 (nível de cinza¹)
6. **strand** → direção da fita “+” ou “-”
7. **thickStart** → posição inicial onde a *feature* é desenhada
8. **thickEnd** → posição final onde a *feature* é desenhada
9. **itemRgb** → determina a cor dos dados
10. **blockCount** → número de bloco (exons)
11. **blockSizes** → lista de blocos separados por vírgula
12. **blockStarts** → lista de posições iniciais dos blocos

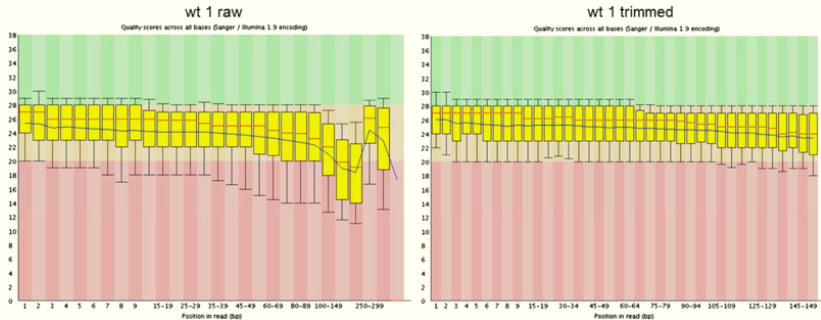
1) Pode ser usado para outras medidas como p-value, up/down, ...

Formatos (GFF)

- ❖ GFF são similares aos BED e têm 9 colunas obrigatórias
 1. seqname → nome da sequência
 2. source → origem da *feature*
 3. feature → tipo de *feature*, equivalente ao campo *name* do BED
 4. start → posição inicial
 5. end → posição final
 6. score → assim como o arquivo BED permite níveis de valores representando a expressividade da anotação
 7. strand → direção da fita “+” ou “-”
 8. frame → frame da sequência codificadora: “0”, “1”, “2” ou “.”
 9. attribute → muda conforme a versão do GFF (GFF1, GFF2, GFF3) e denota texto livre com algum significado biológico

Filtragem e montagem de fragmentos

Filtragem

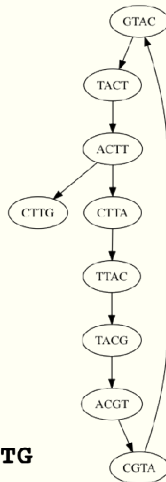


RNA-Seq analysis of isolated satellite cells in Prmt5 deficient mice

Montagem (Grafo De Brujin)

ACTT
CTTA
TTAC
TACG
ACGT
CGTA
GTAC
TACT
ACTT
CTTG

ACTTACGTACTTG



Montagem (Grafo De Brujin)

- ❖ Um caminho *Euleriano* visita cada aresta exatamente uma vez
- ❖ Existe um caminho Euleriano se, e somente se:
 1. máximo de 2 nós semi-balanceados
 2. todos os demais nós são balanceados
- ❖ Um nó é balanceado se *indegree* e *outdegree* são iguais
- ❖ Um nó é semi-balanceado se a diferença máxima de *indegree* e *outdegree* é 1
- ❖ um grafo é conectado se cada nó pode ser alcançado por algum outro nó

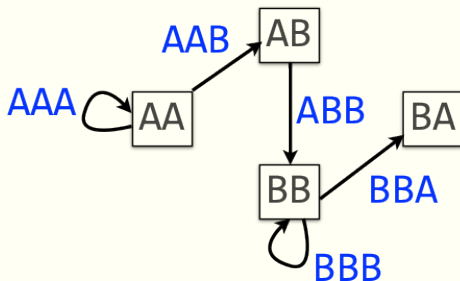
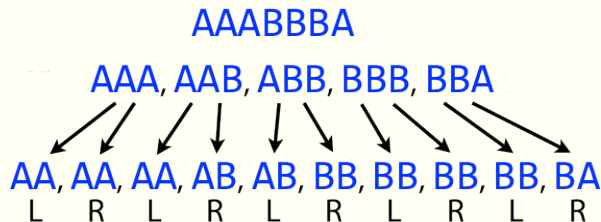
Montagem (Grafo De Brujin)

k-mer é uma substring de tamanho **k**

Exemplo: Seja a String **GGCGATTCATCG**, então todos os 3-mer:

GGC
GCG
CGA
GAT
ATT
TTC
TCA
CAT
ATC
TCG

Montagem (Grafo De Brujin)



Prática em Laboratório

GitHub



Prática