

Universidade Federal do Espírito Santo
Centro de Ciências da Saúde
Programa de Pós-Graduação em Biotecnologia



Projetos Genomas

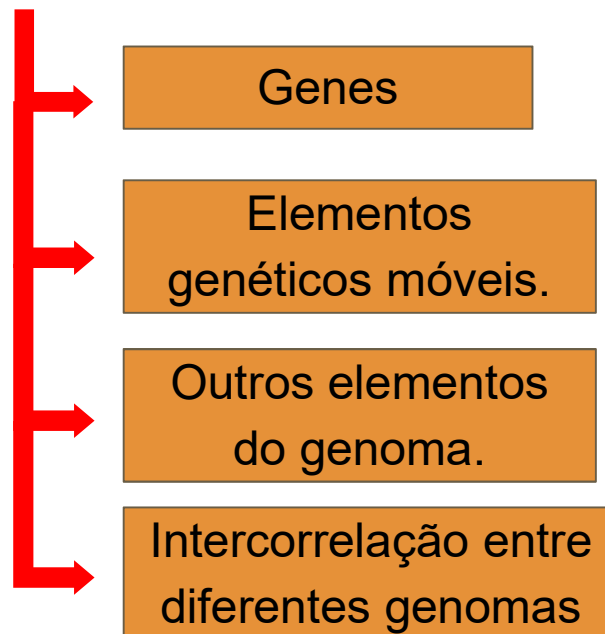
BIOINFORMÁTICA

Mestrandos: Jeisdens Fernandes
Marcos Hell
Matheus Lage



Introdução

A análise in silico das sequências nucleotídicas de cromossomos de um dado organismo, ou simplesmente genoma, constitui uma das mais importantes aplicações da bioinformática.

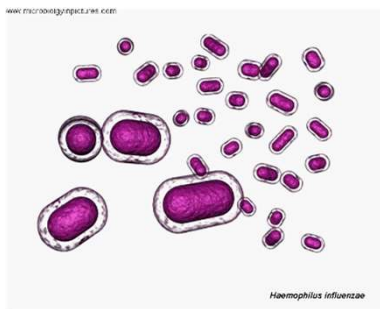


Introdução

- Bactéria *Haemophilus influenza*

Primeiro organismo a ter a sequência de nucleotídeos determinada (1995).

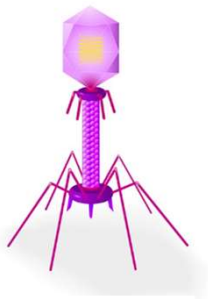
1,8 milhão de bases e 1 cromossomo circular.



Introdução

Geração de algoritmos para analisar as sequências genômicas.

Previamente às análises do genoma de *H. influenzae*, programas para montagem de genomas já existiam.



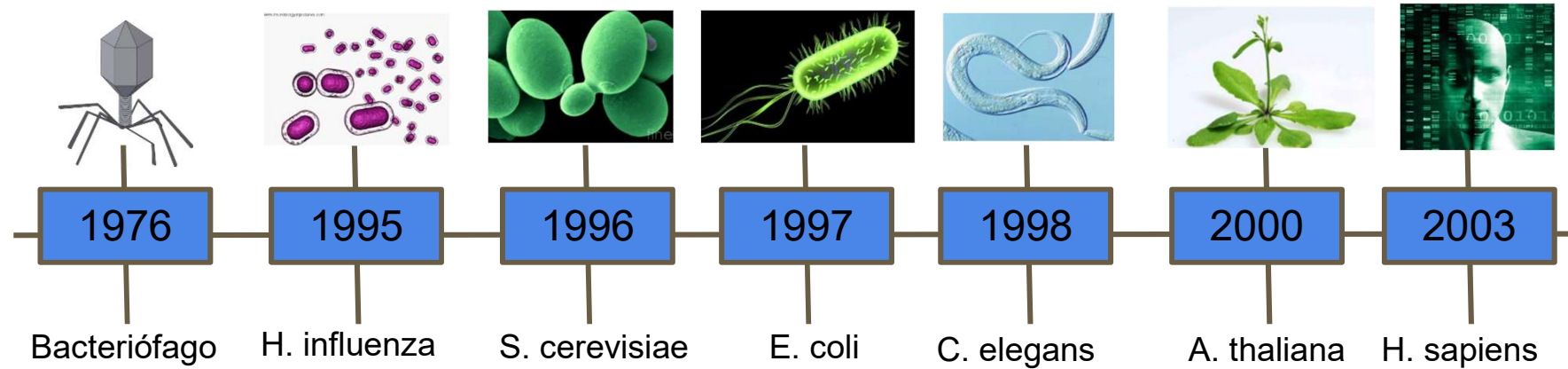
Bacteriófago
48.000 pb



Citomegalovírus (CMV) - 229.000pb

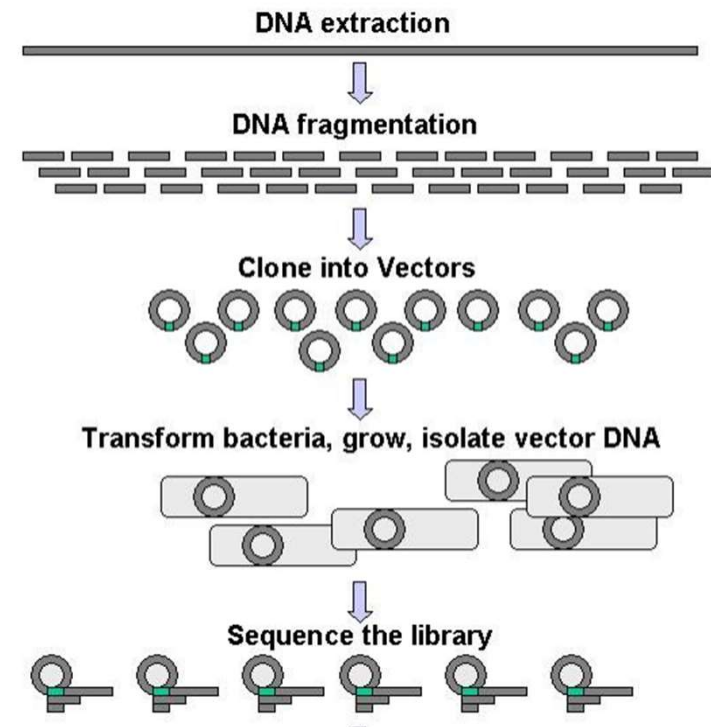
Introdução

Linha temporal do sequenciamento



Montagem de genomas

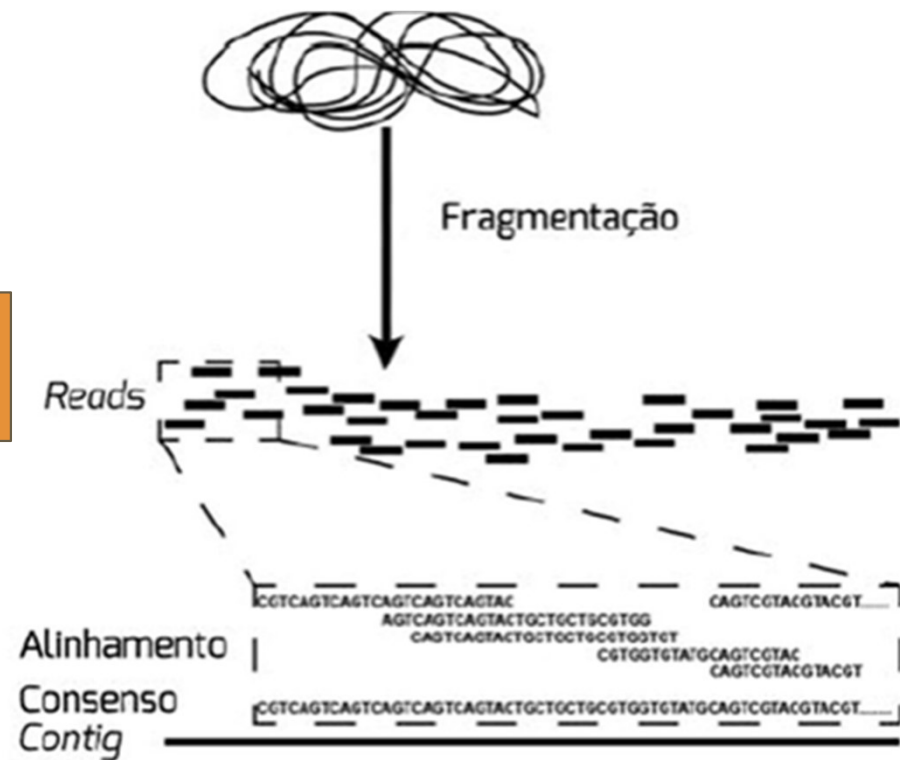
Sequenciamento na metodologia de Sanger nos primeiros anos da era genômica.



Montagem de genomas

Reads: sequência obtida de cada clone.

Contigs: união de sequências reads formadas pela sobreposição de elementos comuns.



Montagem de genomas

Next-generation sequencing (NGS):

- ❖ Fragmentação aleatória;
- ❖ Sem clonagem;
- ❖ Obtenção mais rápida de reads.



Illumina - HiSeq 2000: Produz acima de 600Gb por corrida em 13 dias.

Montagem de genomas

Os primeiros algoritmos se baseavam no alinhamento dos reads e na concatenação de sequências obtidas dos reads com os maiores alinhamentos.

A montagem do genoma a partir de reads seguia os passos abaixo:

- i)* cálculo de alinhamentos aos pares de todos os fragmentos;
 - ii)* escolha de dois fragmentos com a maior sobreposição;
 - iii)* fusão dos dois fragmentos;
 - iv)* repetição dos passos anteriores até obtenção de uma única sequência.
-

Montagem de genomas

Para as novas metodologias, devido ao tamanho pequeno dos fragmentos, diferentes algoritmos tiveram que ser desenvolvidos.

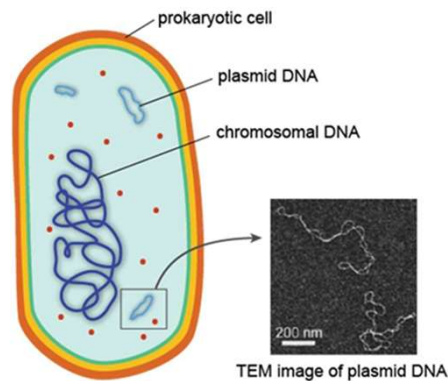
Diferentes programas surgiram para montagem de genomas utilizando diferentes algoritmos.

Nome	Análise
ABYSS	grandes genomas
ALLPATHS-LG	grandes genomas
Celera WGS Assembler	grandes genomas
CLC Genomics Workbench	genomas e transcriptomas
Geneious	genomas
Newbler	genomas e transcriptomas
Phrap	genomas e transcriptomas
SOAPdenovo	genomas e transcriptomas
Staden gap4 package	genomas pequenos e transcriptomas
Trans-ABYSS	transcriptomas
Velvet	genomas pequenos e transcriptomas

Montagem de genomas

O sequenciamento do genoma procarioto é com obtenção de uma sequência única a qual representa toda a sequência nucleotídica do cromossomo.

Nº. de contigs dependerá do nº. de plasmídios.



Montagem de genomas

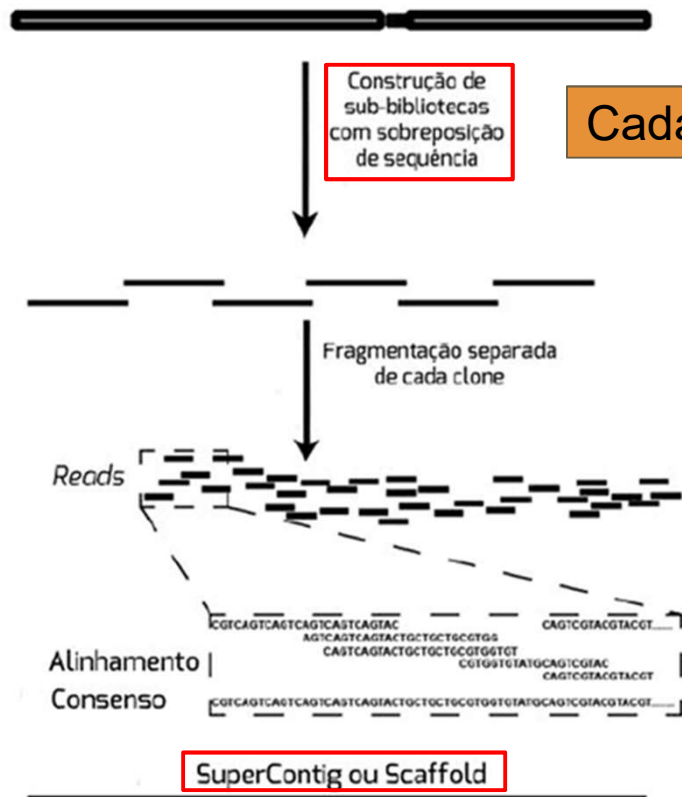
O genoma eucarioto possui grande variação no número de cromossomos, sendo esperado maior nº. de contigs e maior complexidade do mesmo.

Sequências repetitivas torna mais difícil a montagem.

espécie	nome comum	tamanho estimado do genoma (pb)
<i>Oryza sativa</i>	arroz	5.000.000.000
<i>Mus musculus</i>	camundongo	3.454.200.000
<i>Homo sapiens</i>	homem	3.400.000.000
<i>Rattus norvegicus</i>	rato	2.900.000.000
<i>Drosophila melanogaster</i>	mosca-da-fruta	180.000.000

Internet: www.cbs.dtu.dk

Montagem de genomas



Cada sub-biblioteca será sequenciada para gerar contigs.

Conjunto de diferentes contigs oriundos de diferentes sub- bibliotecas será utilizado para a formação de scaffolds.

Montagem de genomas

- ❖ Alguns programas permitem montar genomas complexos com repetições baseadas em reads maiores.
- ❖ Realizam montagens em duas ou mais fases distintas, nas quais as sequências repetitivas são processadas separadamente.



Montagem de genomas

Qualidade da montagem do genoma: pode ser acompanhada por alguns índices.

- ❖ A cobertura reflete na quantidade de reads associados a um determinado fragmento de DNA.
Ex.: cobertura de 10X = cada nucleotídeo foi encontrado em pelo menos 10 reads.
- ❖ -N50: medida estatística muito utilizada para avaliar a quantidade de montagem, pois revela o quanto de um genoma é coberto por contigs grandes.

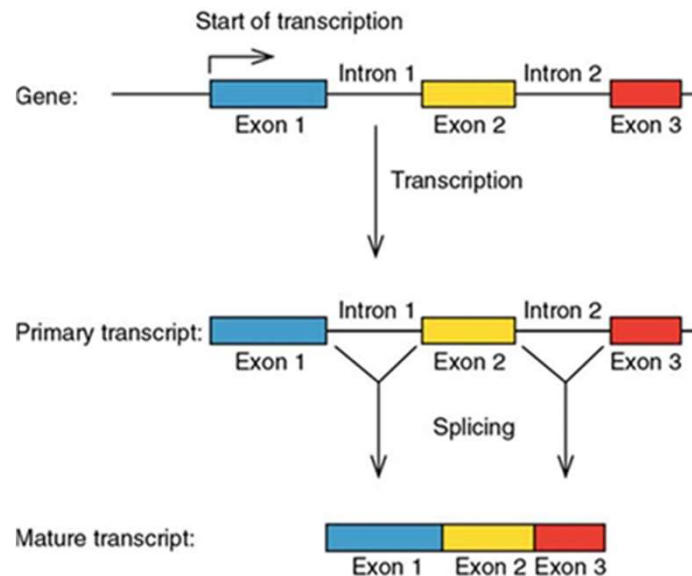
$$N50 = n$$

n= significa que 50% dos reads estão montados em um contig de tamanho n ou maior.

Montagem de transcriptomas

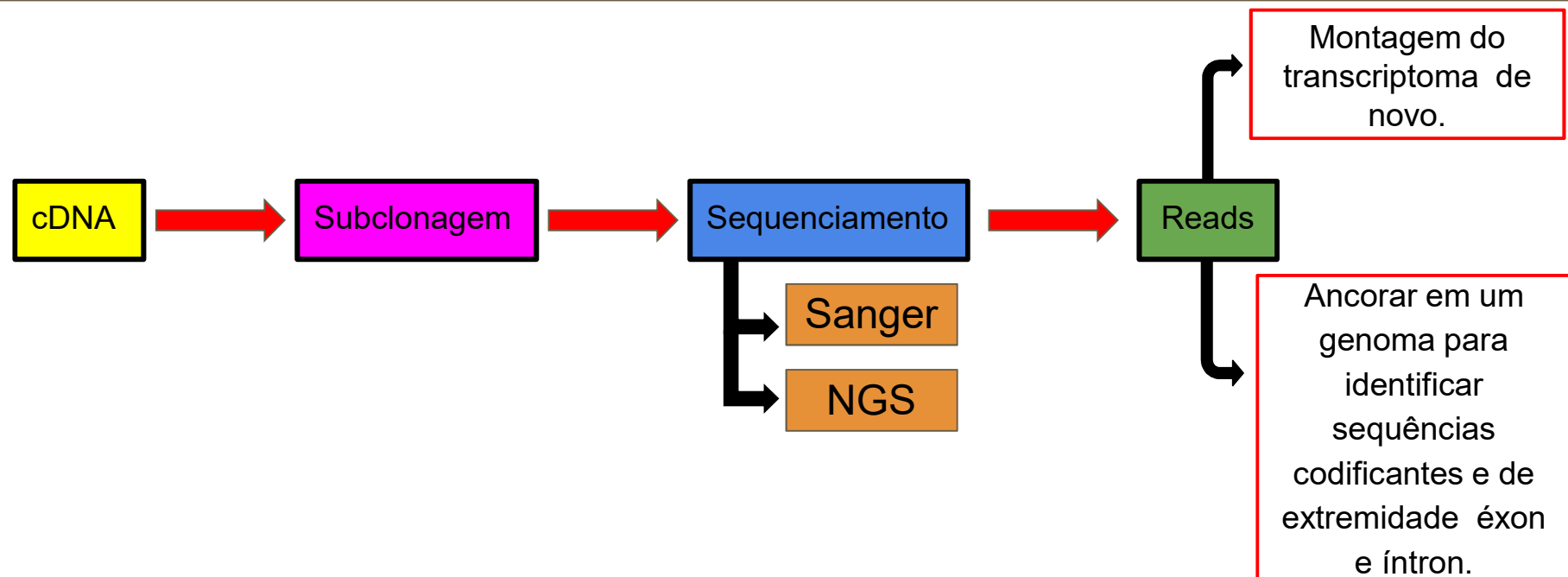
Identificação de transcritos:

- ❖ Fornece indícios sobre quais genes estão sendo expressos em uma situação fisiológica específicas;
- ❖ Sequenciamento: aplicação importante na procura de sequências codificantes.



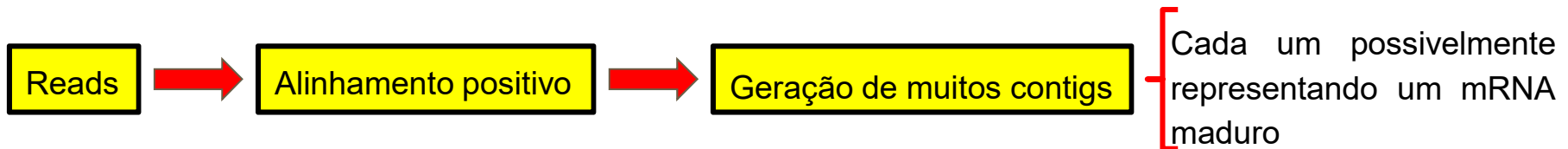
Montagem de transcriptomas

Ao contrário de genomas, em transcriptomas o material de partida é o cDNA, obtido através de transcrição reversa de mRNA.



Montagem de transcriptomas

Alinhamento de novo



Programas:

- ❖ Realizam a montagem;
- ❖ Alinhamento de genomas;
- ❖ Análise representativa de cada transcrito dentro do conjunto total de RNA analisado.

Cálculo da frequência relativa de cada transcrito identificado.

Montagem de transcriptomas

Cálculo da frequência relativa:

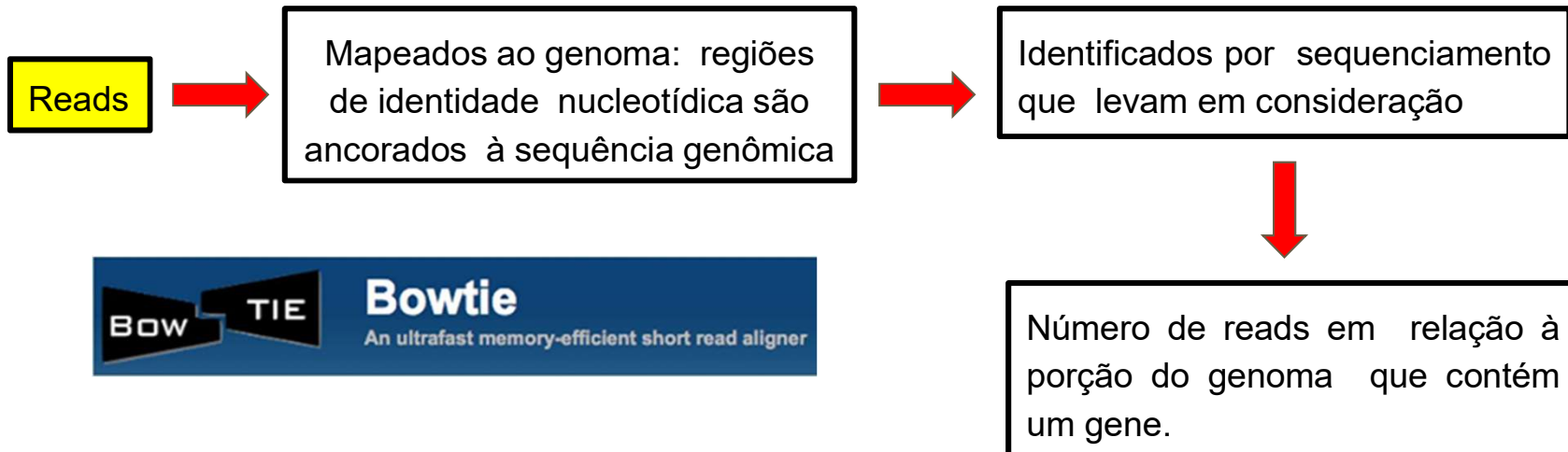
- ❖ Realizar análises de expressão diferencial de genes.
- ❖ Programas:
 - Cufflink-Cuffdiff
 - DegSeq
 - DESeq
 - EdgeR

Duas estratégias:

1. Mapeamento a uma sequência genômica previamente conhecida;
2. Análise de novo, independente da sequência genômica e baseada na montagem dos transcritos diretamente a partir dos reads.

Montagem de transcriptomas

Mapeamento a uma sequência genômica previamente conhecida



Resultado: determinada sequência do genoma é representada por um grande número de reads.

Montagem de transcriptomas

Mapeamento de sequência cujo genoma não foi ainda determinado

RNAseq:

- ❖ Uso de NGS para sequenciar cDNA à capturar a informação do transcriptoma;
- ❖ Não necessita uma lista pré-definida dos genes que se deseja detectar;
- ❖ A princípio, qualquer transcrito que estiver sendo expresso pode ser detectado;
- ❖ Detecta também eventos de splicing alternativo.



Ion Proton™ System

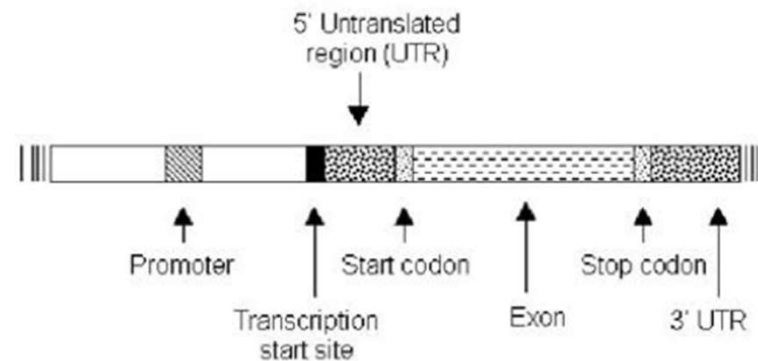
Identificação/anotação gênica

- Anotação de genomas:
 - Utilizada para delimitar possíveis genes e predizer a sua função com base nas similaridades com sequências conservadas.
 - Dois grandes grupos de genes:
 - Genes cujo produto é reconhecido pelos ribossomos e dará origem a proteína (mRNA);
 - Genes cujo produto terá funções estruturais e funcionais dependente da própria moléculas de RNA (tRNA e rRNA).

Identificação/anotação gênica

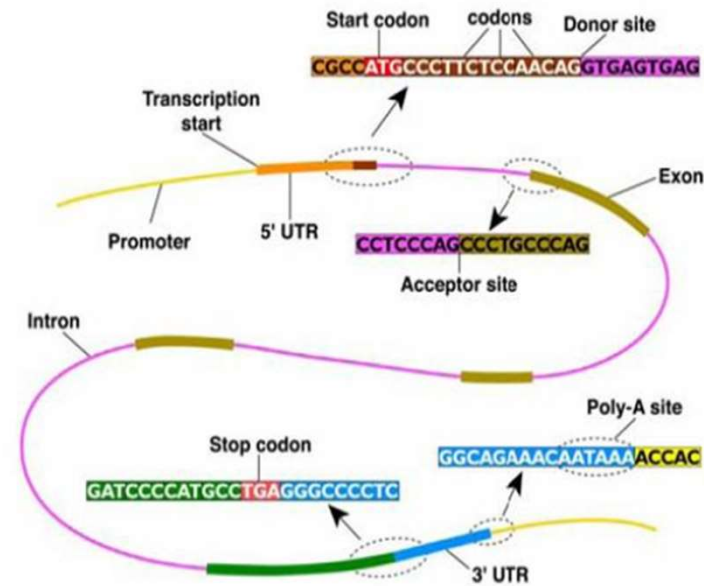
- Identificação de regiões codificantes
- ORF (open reading frame): toda região delimitada por um códon de início e término.

- Procariotos



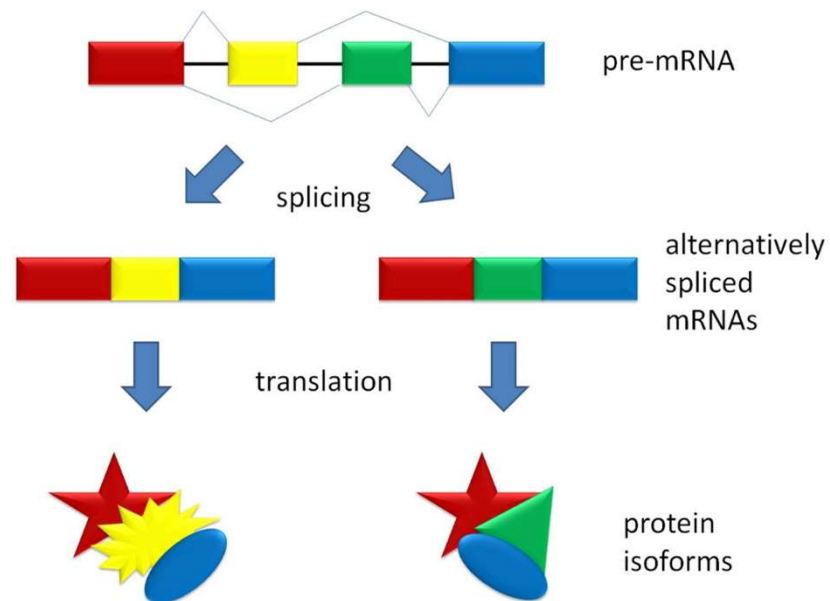
Identificação/anotação gênica

- Eucariotos:
- Com íntrons;
papel regulação;
- Íntrons sofrem
splicing.



Identificação/anotação gênica

- Splicing alternativo: produção de diferentes proteínas do mesmo gene



Identificação/anotação gênica

- Procariotos:
- Procura de genes: baseada apenas na identificação de ORFs , resultando em falsos positivos e falsos negativos .

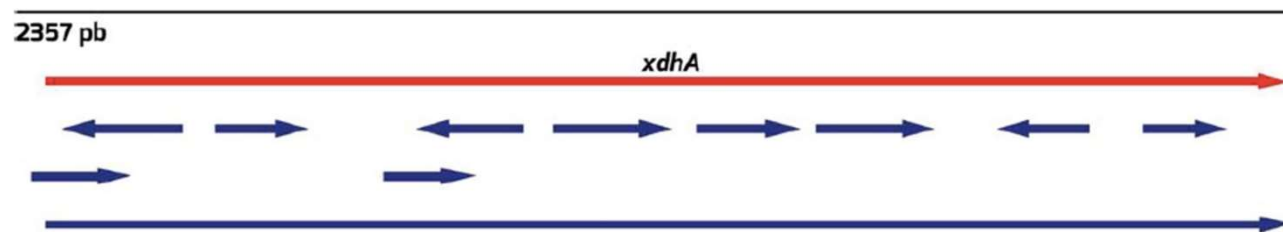
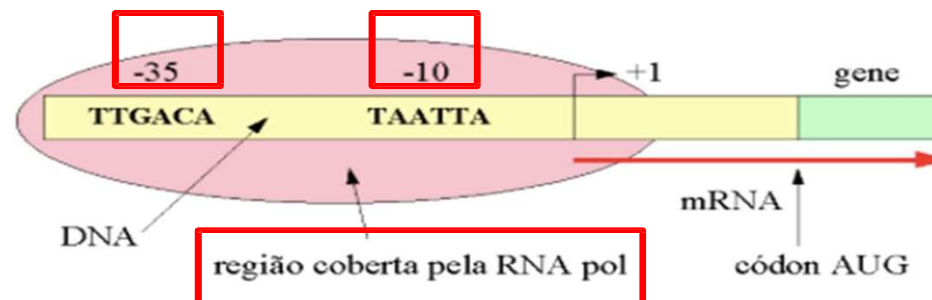


Figura 3-4: A simples procura de ORFs pode gerar resultados falso positivos na procura de genes em organismos procarióticos. Como exemplo, uma sequência de DNA de 2357 pb da bactéria *E. coli* HS (nucleotídeos 3027764 ao 3030120 – [Código de Acesso junto ao NCBI NC_009880.1](#)), a qual contém o gene *xdhA*, foi avaliada quanto à presença de ORFs com mais de 150 pb com o programa ORF Finder. A sequência anotada do gene encontra-se em vermelho, ao passo que as possíveis ORFs estão demarcadas em azul.

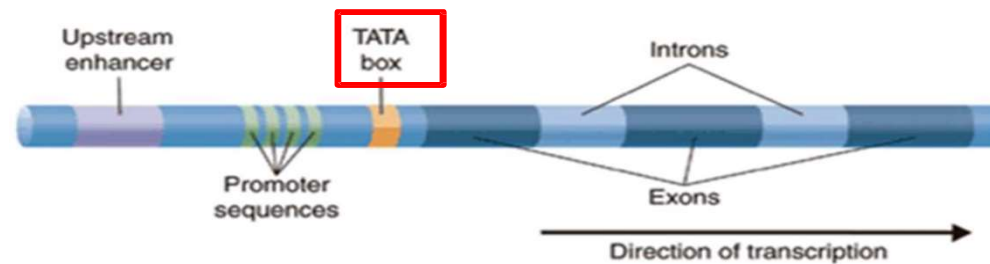
Identificação/anotação gênica

- Sinais transcrpcionais: sequências conservadas utilizadas para o processo de transcrição.

Procarioto:

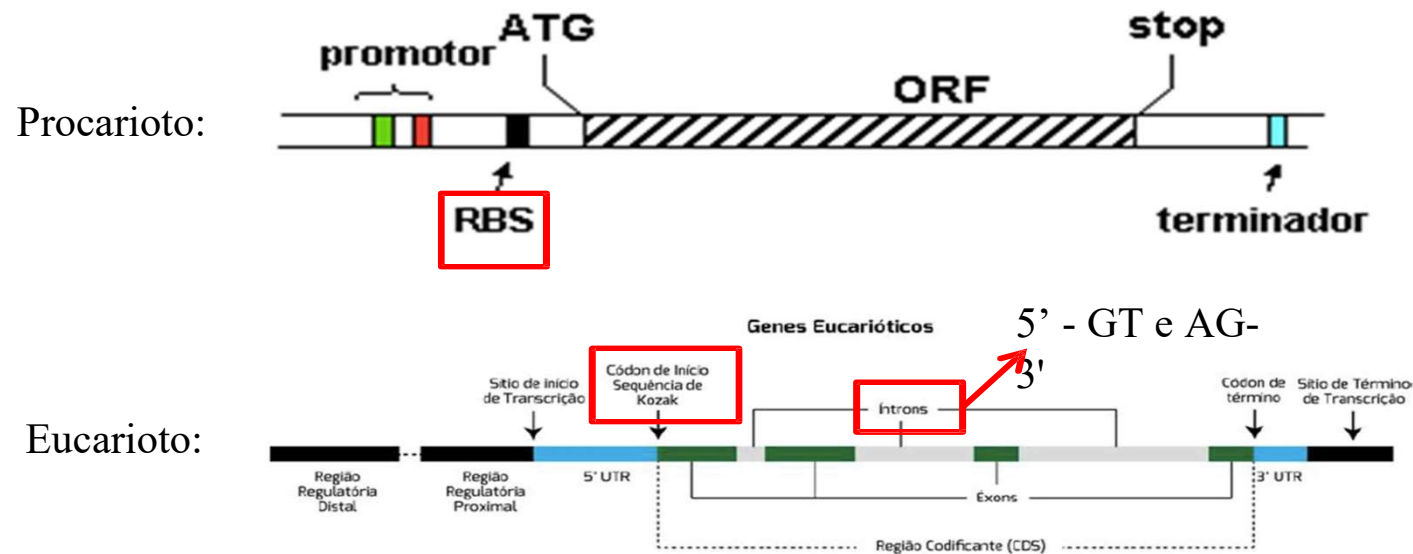


Eucarioto:



Identificação/anotação gênica

- Sinais traducionais: regiões importantes para o recrutamento de ribossomos.



Identificação/anotação gênica

- Detectores de conteúdos:
 - Classificam a sequência de DNA em codificantes e não-codificantes (íntrons, regiões intergênicas, regiões não traduzidas).
 - Detectores extrínsecos: baseiam-se nas regiões codificantes por serem mais conservadas do que as não codificantes □ permite identificação de exons conservados para procurar homologias.
 - Detectores intrínsecos: tem como foco características inatas ao DNA as quais permitem a predição do potencial de uma sequência codificar ou não uma proteína.

Identificação/anotação gênica

- Detectores intrínsecos:

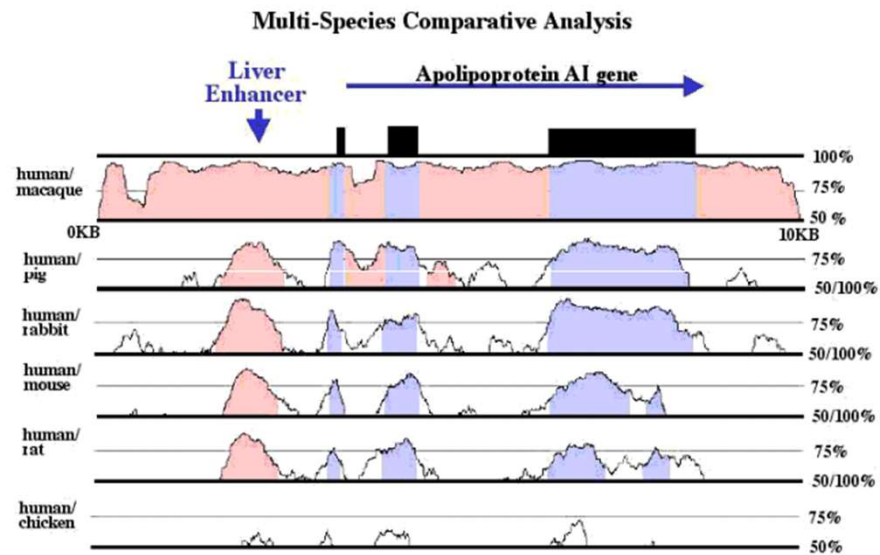
Exemplos:

- i) preferência por bases G ou C na terceira posição do códon;
- ii) uso diferencial de códons sinônimos (# códons codificam mesmo aas);
- iii) frequência de sequências distintas hexâmeras;
- iv) periodicidade de ocorrência de bases.

Identificação/anotação gênica

Procura empírica de genes

Busca por similaridades com outros bancos de dados



Identificação/anotação gênica

- Detectores de sinais

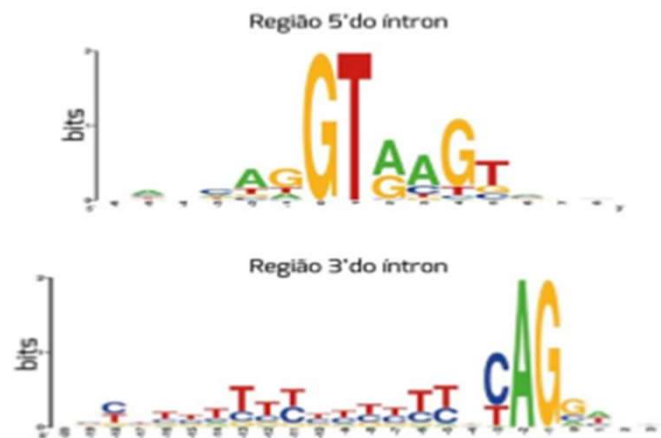
- Conteúdo para delimitar a sequência gênica.
- Os algoritmos são treinados com sequências conhecidas do genoma em questão.

O grau de conservação dos nucleotídeos presentes na sequência de Kozak de *Drosophila melanogaster*, perfil este que pode ser utilizado na predição de novas sequências codificantes neste organismo.



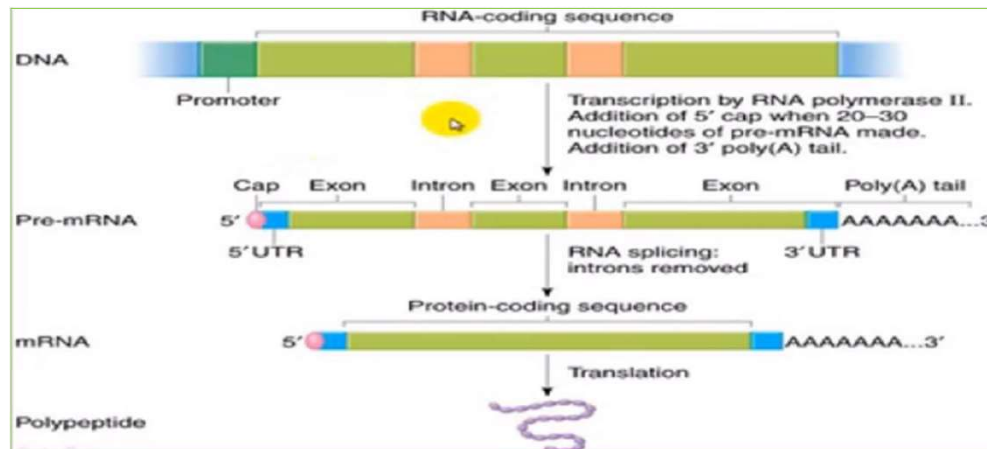
Identificação/anotação gênica

Padrão de conservação de nucleotídeos nas regiões 5' (painel superior) e 3' (painel inferior) de íntrons humanos.



Identificação/anotação gênica

- Dentre as limitações da predição *ab initio* está o fato de que, usualmente, o resultado obtido se refere às regiões codificantes.
- Para tanto, alguns destes algoritmos são treinados com modelos gênicos já conhecidos, de organismos filogeneticamente próximos e, assim, provavelmente possuem uma estrutura gênica muito parecida com a do organismo que está em análise.



Identificação/anotação gênica

A forma mais simples de anotação automática se dá pela análise de uma série de diferentes mecanismos de predição e delimitação de sequências gênicas e, então, utilização de um algoritmo de seleção, também denominado de *combiner*.

Função de selecionar a predição que melhor represente os modelos gênicos

Os combiners estimam os tipos e as freqüências de erros oriundos de cada programa de predição

Escolhendo as combinações de evidências que minimizam tais erros.

Identificação/anotação gênica

Tabela 2-4: Principais algoritmos utilizados na predição de genes e a sua funcionalidade.

Algoritmo	Descrição	Aplicação
Predições <i>ab initio</i> e baseados em evidência		
Augustus	Aceita evidências baseadas em transcriptomas e banco de dados de proteínas	Eucariotos
FGNESH	Arquivos para treino derivados de análise do fabricante	Eucariotos
fgenesB	Predição de genes e operons em bactérias baseadas em padrões e cadeias de Markov	Procariotos
Genemark	Arquitetura de busca baseada em <i>self-training</i>	Procariotos e eucariotos
Twinscan	Extensão do algoritmo Genscan que utiliza homologia entre dois genomas para guiar a predição de genes	Eucariotos
GenomeScan	Extensão do algoritmo Genscan que utiliza BLASTx para guiar a predição de genes	Eucariotos
Glimmer	Utiliza modelos de Markov interpolados	Procariotos
Combiners		
Evidence Modeler	Tem como resultado um modelo gênico pela combinação de evidências obtidas a partir de alinhamento de dados transcriptômicos e proteômicos com predições <i>ab initio</i>	Eucariotos
Evigan	Algoritmo de evidências probabilísticas que usa redes Bayesianas para pontuar e integrar predições <i>ab initio</i> e baseadas em evidência para produzir modelos gênicos.	Eucariotos

Identificação/anotação gênica

- Diversidade de métodos estatísticos, geralmente treinados com sequências proteicas conhecidamente pertencentes a algum sub-compartimento celular

Tabela 3-4: Principais algoritmos utilizados na predição da localização celular de proteínas.

Algoritmo	Descrição	Aplicação
BaCellLo	Com base na composição de aminoácidos e sequências de treino, prediz em 5 localizações (secretada, citoplasmática, nuclear, mitocondrial e cloroplástica)	Plantas, animais e fungos
LOCtree	Com base na sequência N-terminal, prediz a localização em secretada, citoplasmática, nuclear, mitocondrial, cloroplástica e organelar.	Eucariotos e procarlotos
TARGETp	Com base na sequência N-terminal, prediz a localização como secretada, mitocondrial e cloroplástica, dentre outras.	Eucariotos e procarlotos
Wolf PSORT	Com base na sequência N-terminal e regras empíricas, classifica o endereçamento em cloroplástico, citosólico, citosqueleto, retículo endoplasmático, extracelular, golgi, lisossômico, mitocondrial, nuclear, peroxissomal, membrana plasmática e membrana vacuolar. Permite localização múltipla.	Animais, fungos e plantas
Cell-PLoc	Permite realizar a localização de proteínas em mais de 25 diferentes locais, baseados em treino com sequências cuja proteína tem localização conhecida.	Eucariotos, procarlotos e vírus

Identificação/anotação gênica

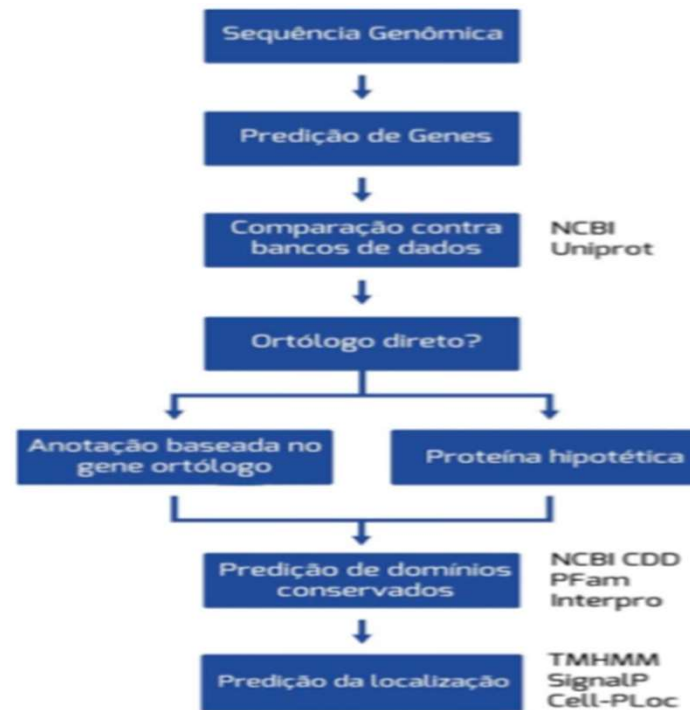
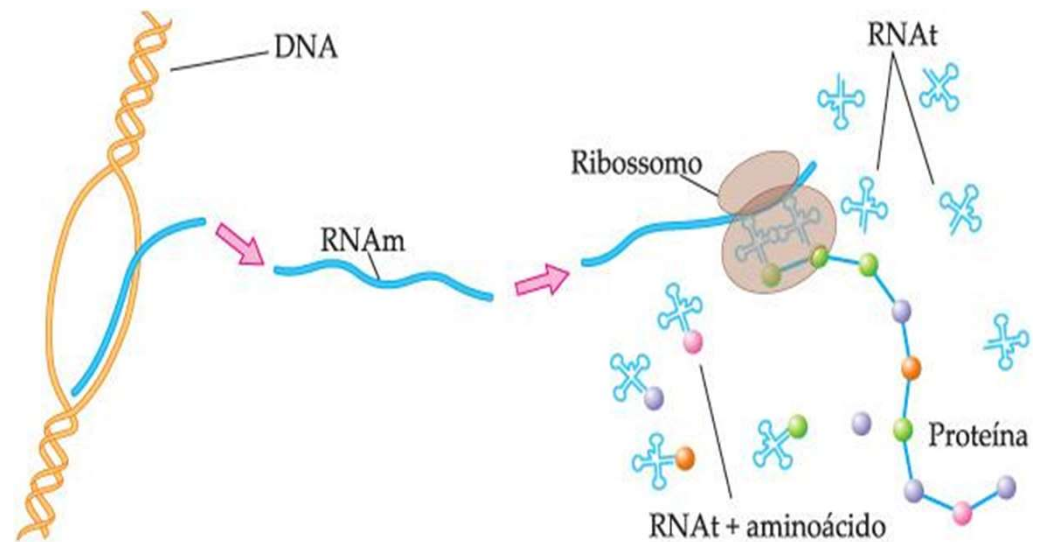


Figura 7-4: Um fluxo de trabalho genérico para anotação de genes.

Identificação e anotação de RNAnc

Processo de síntese proteica:

RNA mensageiro;
RNA ribossômico;
RNA transportador;

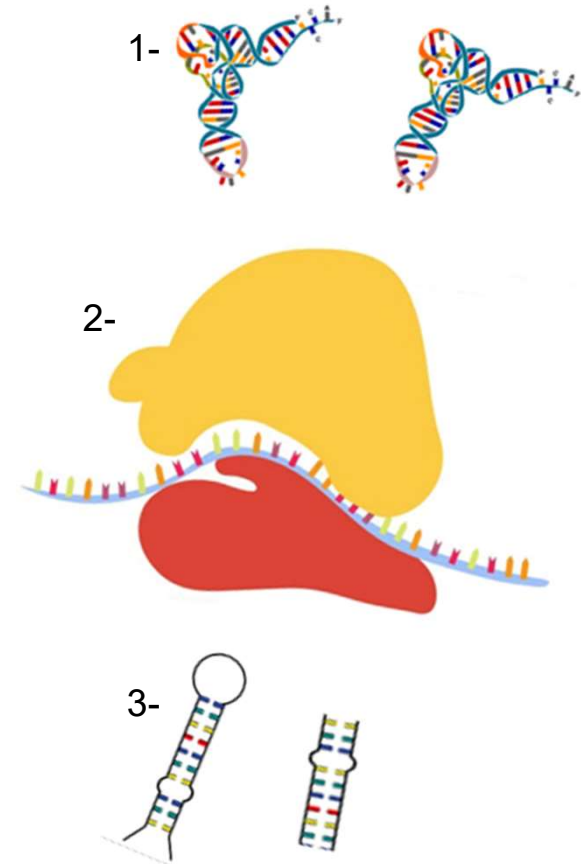


Identificação e anotação de RNAnc

RNAs não codificantes: RNAt (1), RNAr (2), pequenos RNAs (3);

A anotação dos genes de RNAnc ainda não apresenta um grande número de programas:

- **Heterogeneidade e pequena conservação** desses genes;
- Geralmente não apresentam conservação de **sequência primária**;



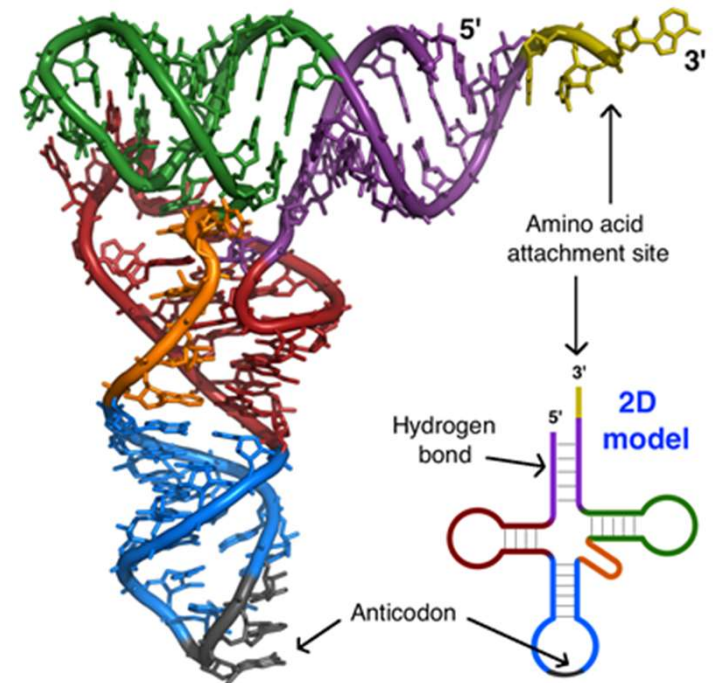
Identificação e anotação de RNAnc

Busca de RNAt nos genomas:

tRNAscan-SE: estrutura secundária, bases conservadas;

ARAGORN: predição da estrutura do tRNA baseada na homologia de sequências conservadas, estrutura secundária;

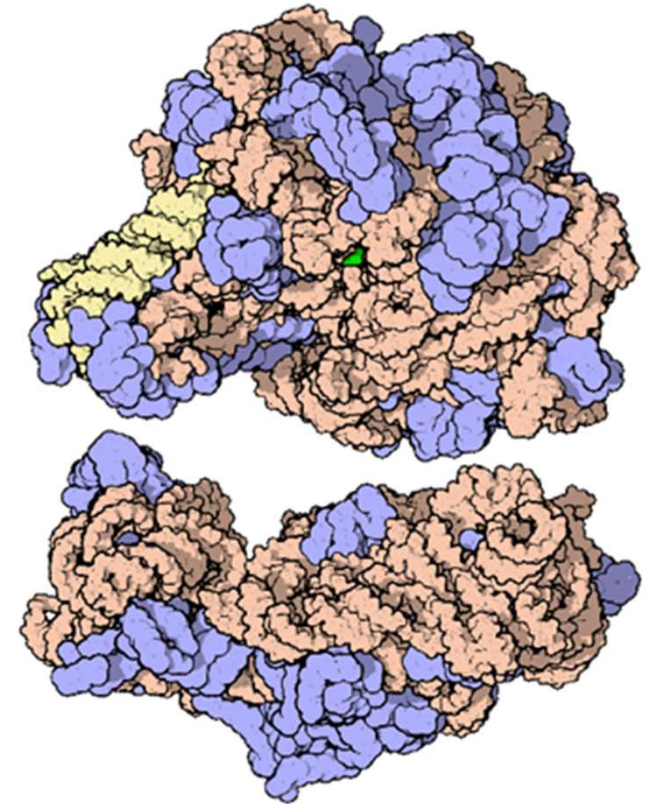
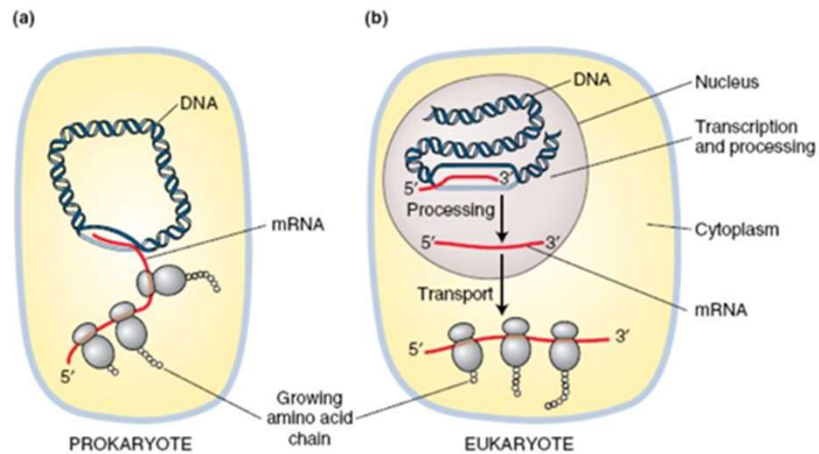
tRNAfinder: detecção da estrutura secundária do RNA predito para identificar genes de tRNA;



Identificação e anotação de RNAnc

A busca de **RNAs** é baseada na conservação de sequências;

- **Procariotos** (23S, 16S e 5S)
- **Eucariotos** (28S, 18S, 5.8S e 5S)



Identificação e anotação de RNAnc



Desenvolvido para delineamento dos **genes de RNAr**;



Banco de dados que contém informações sobre as famílias de RNA não codificantes e de outros elementos de RNA estruturados;

Divididos em três grandes grupos:

RNAs não codificantes (RNAnc);

Elementos estruturais regulatórios em *cis*;

RNAs que podem sofrer o processo de *auto-splicing*.

Identificação de pequenos RNAs

Pequenos RNAs: miRNAs, siRNAs, TasiRNAs, tRFs, entre outros;

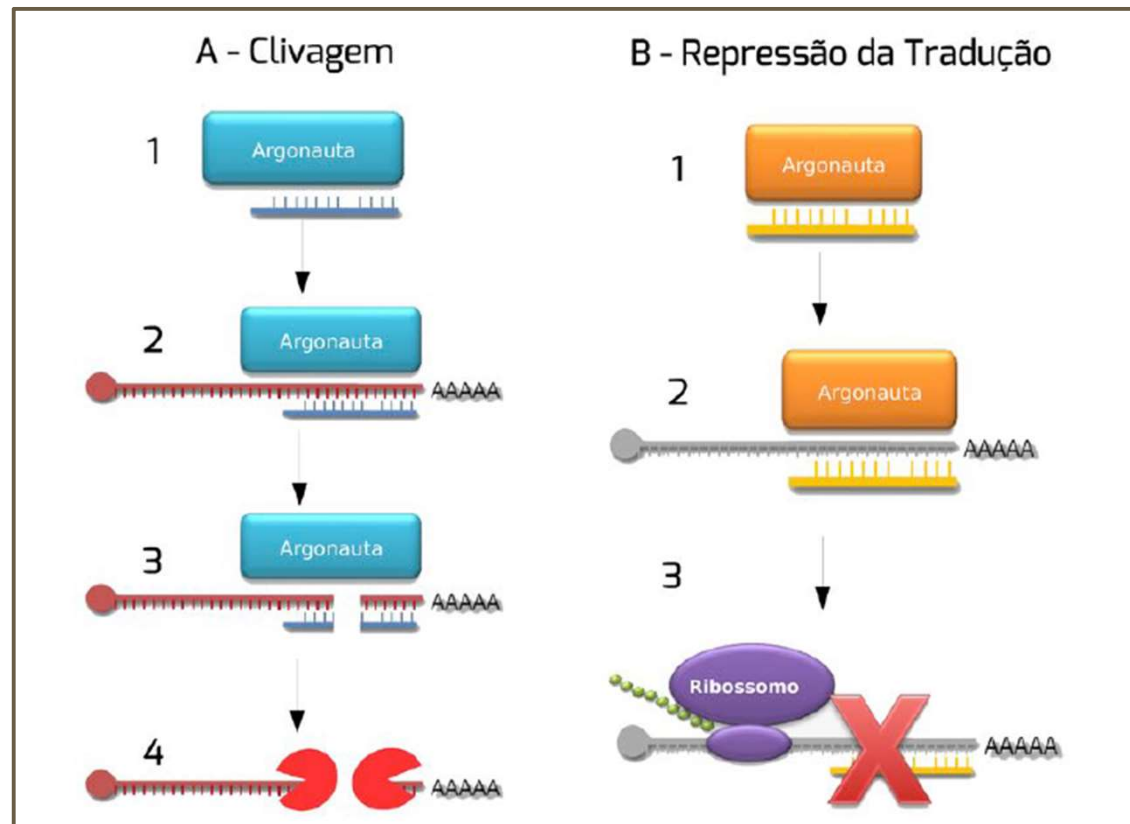
Características: não codificam proteínas, tamanho das moléculas, rotas de biogênese, papéis funcionais;

Inicialmente conhecidos como **produtos de degradação celular** > Identificação do fenômeno de **silenciamento gênico** > descoberta do **papel funcional** dessas moléculas;

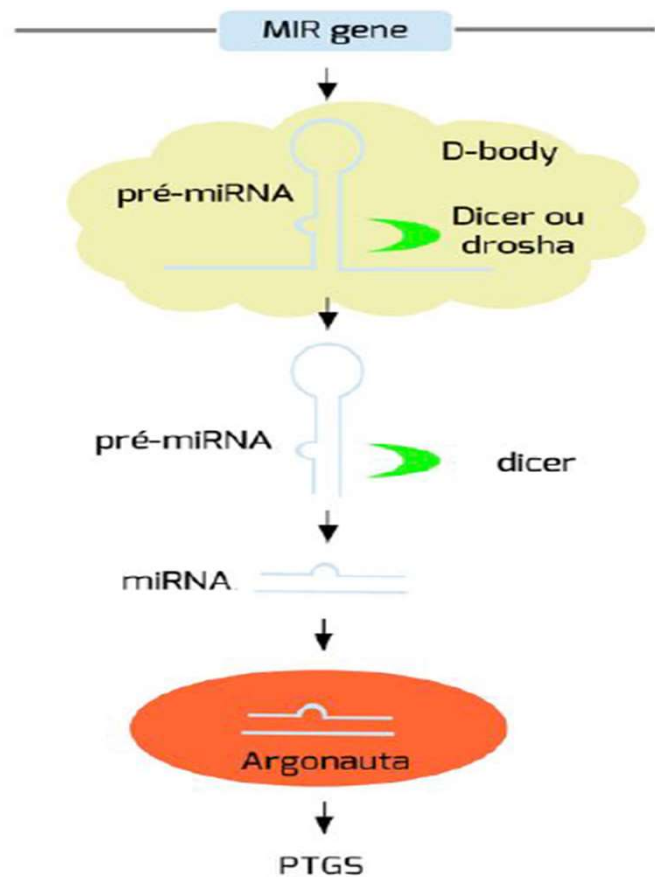
Tabela 1: principais classes de pequenos RNAs com função regulatória

Classe	Tamanho (nt)	Função biológica	Mecanismo de ação	Origem	Organismos
microRNA ou miRNA	21-24	PTGS	Clivagem e repressão da maquinaria de tradução	Intergênica e íntrons	Plantas, animais, fungos e vírus
siRNA	21-24	PTGS, TGS	Clivagem, repressão da maquinaria de tradução e metilação de DNA	Intergênica, éxons e íntrons	Plantas, animais, fungos e vírus
tasiRNA	21-22	PTGS	Clivagem	Transcritos alvo de microRNAs	Plantas, animais e fungos
natsiRNA	21-22	PTGS	Clivagem	Transcritos convergentes parcialmente sobrepostos	Plantas

Silenciamento gênico pós transcrição



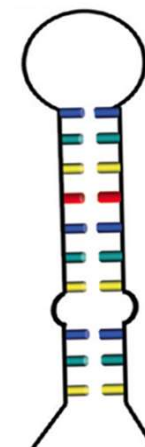
Biogênese de miRNA



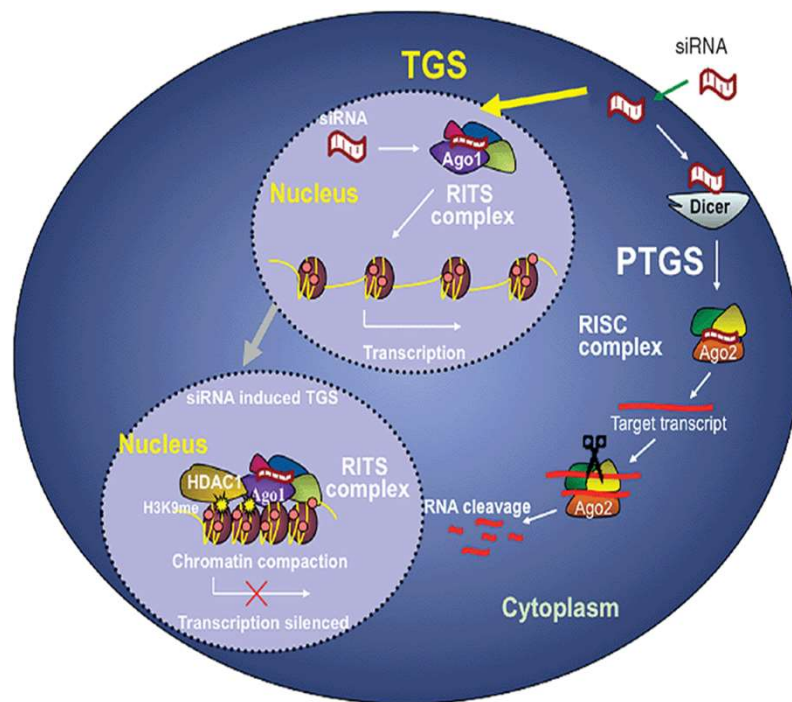
Ferramentas para identificação de miRNA:

miRTools, miRDeep, miRExpress,
miRAnalyser, miRanda miRCat.

Análise em bibliotecas de pequenos RNAs e
cálculos para formação de grampo



Silenciamento gênico transcricional



Ferramenta para identificar **siRNAs**:

SiLoCo

Laboratórios que pesquisam pequenos RNAs desenvolverem suas próprias ferramentas.

Demais classes de **pequenos RNAs**:

Variações da ferramenta **BLAST** são geralmente utilizadas

Identificação de pequenos RNAs

Principais desafios no emprego da bioinformática:

- Identificação da região, ou precursor, que da origem ao pequeno RNA;
- Identificação dos genes alvos regulados por estes;

Existem bancos de dados que disponibilizam os possíveis alvos para um determinado miRNA.

- microRNA.org
-

OBRIGADO!

