Predição e anotação gênica: princípios gerais e comparação entre abordagens

Drª Desirrê Petters-Vandresen

Por que anotar um genoma?

- Sequência do genoma sem anotação: baixa utilidade e aplicabilidade em abordagens funcionais

- Busca por padrões e características que identifiquem genes e sequências relevantes dentro de uma montagem

Quais aspectos biológicos precisamos levar em consideração ao predizer e anotar genes?

Como os softwares utilizam informações sobre estes aspectos para anotação e análise das sequências?

Dogma central da Biologia Molecular

- Fluxo de informação na célula

DNA
Nucleotídeos (4 bases)

Replicação

Transcrição

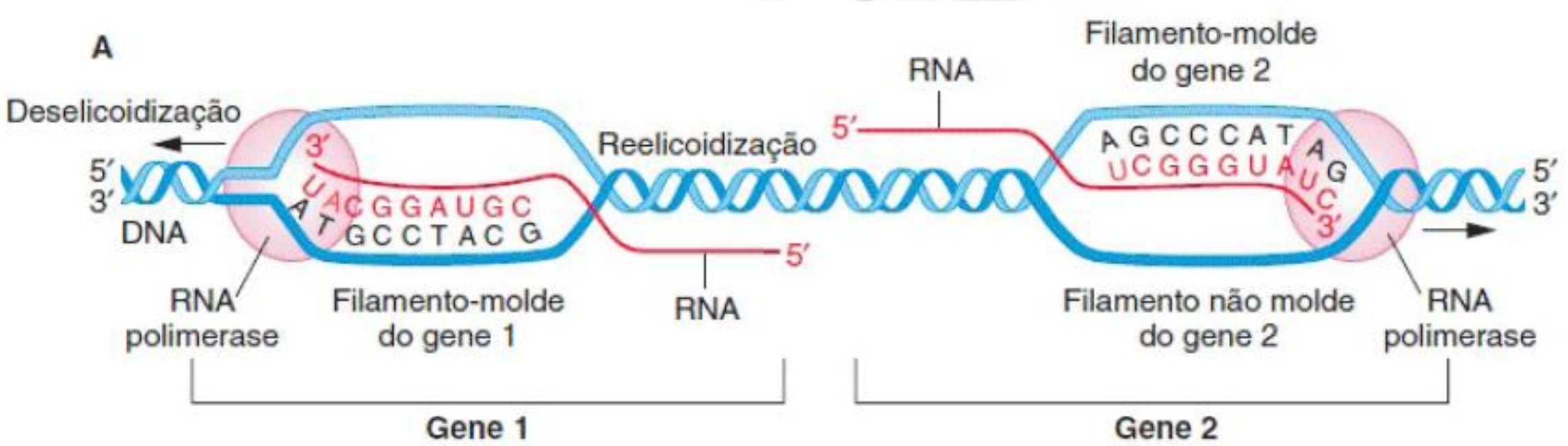
Transcrição reversa

RNA
Nucleotídeos (4 bases)

Tradução
Código genético degenerado (64 trincas)

Proteínas
Aminoácidos (20 resíduos)

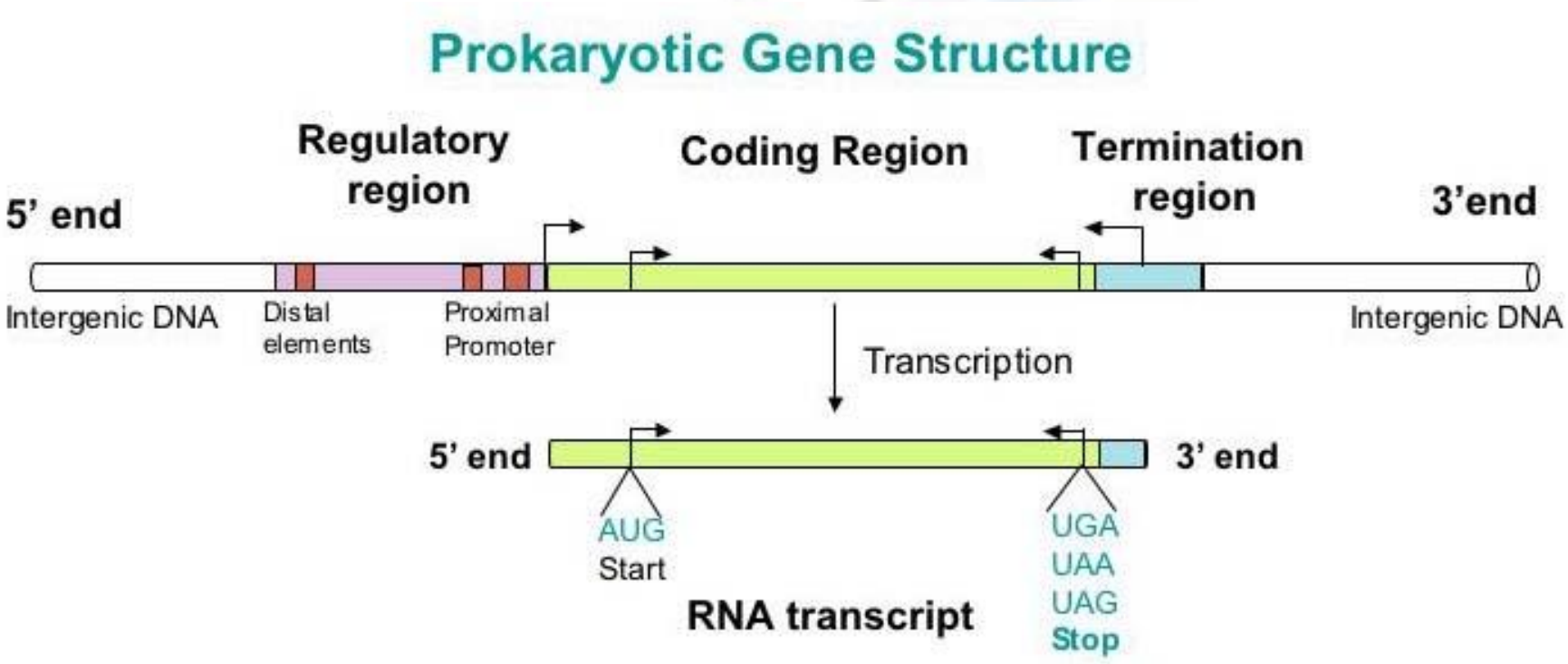
Organização da informação codificante



Adaptado de:
GRIFFITHS, A. J. F.; WESSLER, S. R.; CARROLL, S. B.; DOEBLEY, J. RNA – Transcrição e Processamento. In: _____. (org.) **Introdução à Genética**. 11ª Ed. Rio de Janeiro: Guanabara Koogan, 2016.

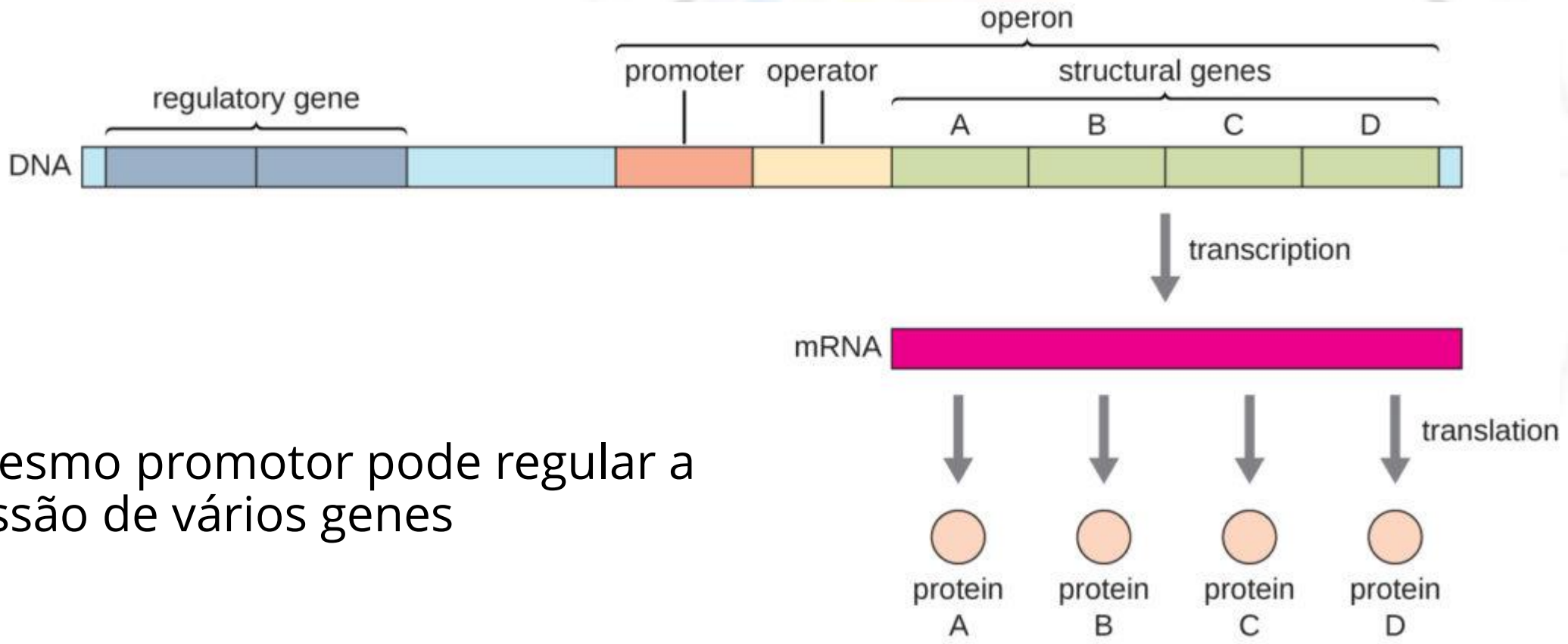
- Ambas as fitas podem codificar proteínas
- O sentido de codificação é o mesmo em cada uma das fitas
- Os transcritos sempre estão no sentido 5’ – 3’
- As sequências codificantes estão presentes na fita codificante (e não na fita que é usada como molde)

Estrutura de genes procarióticos



- Ausência de íntrons
- Genes de estrutura simples
- Possibilidade de genes sobrepostos
- Ausência de processamento complexo do mRNA

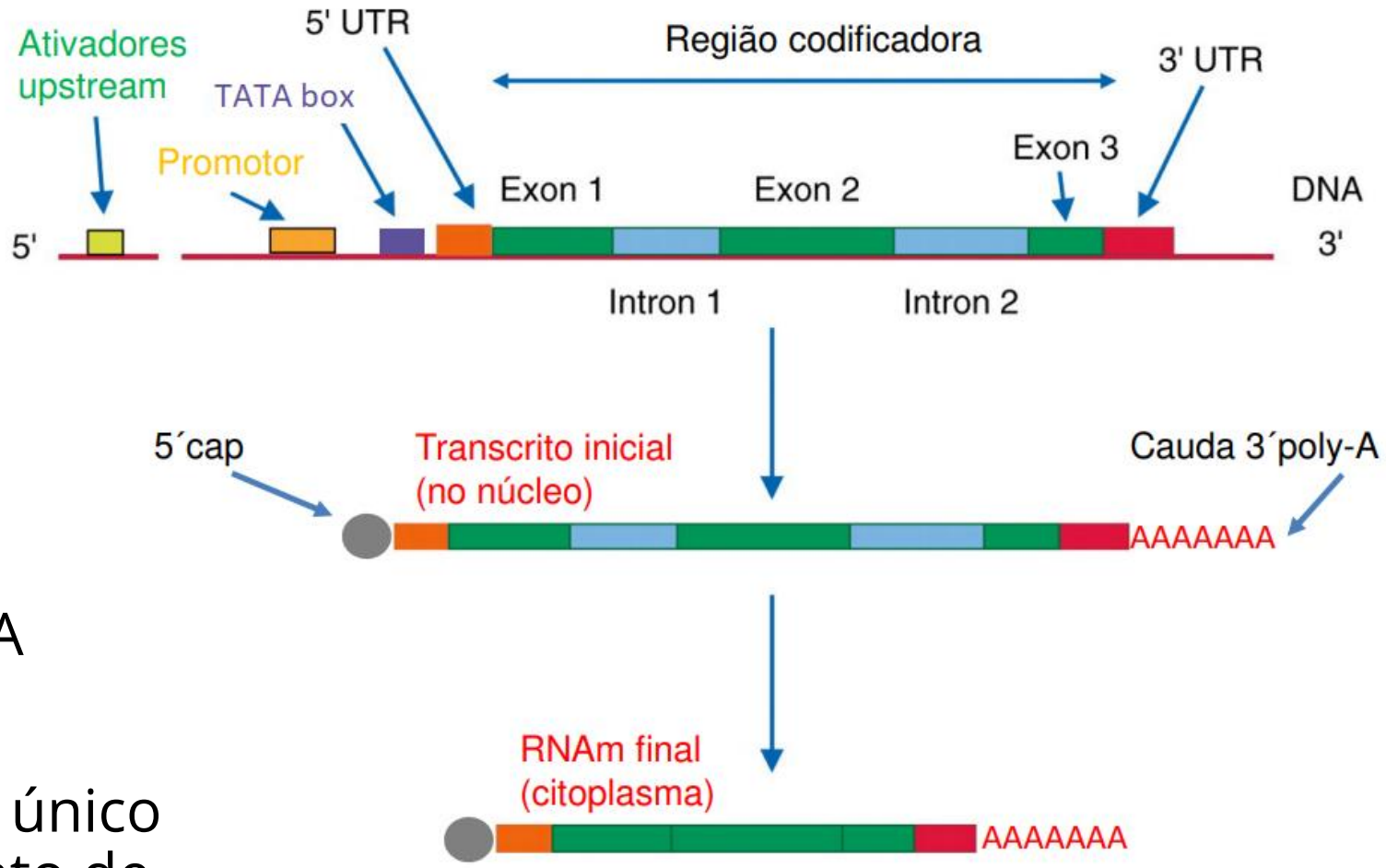
Estrutura de operons



- Um mesmo promotor pode regular a expressão de vários genes

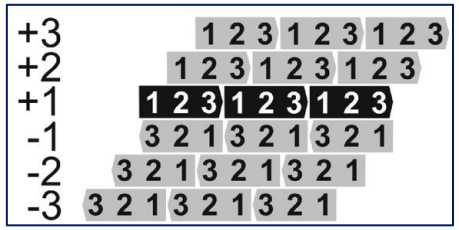
Estrutura de genes eucarióticos

- Genes de estrutura complexa
- Presença de íntrons, que podem ter tamanhos grandes (<300Kb)
- Splicing alternativo e processamento do mRNA
- Um promotor regula um único gene, não há agrupamento de genes em operons



Identificação de fases de leitura

- Cada fita do DNA possui 3 fases de leitura
- Total de 6 fases de leitura (fita direta e fita reversa)



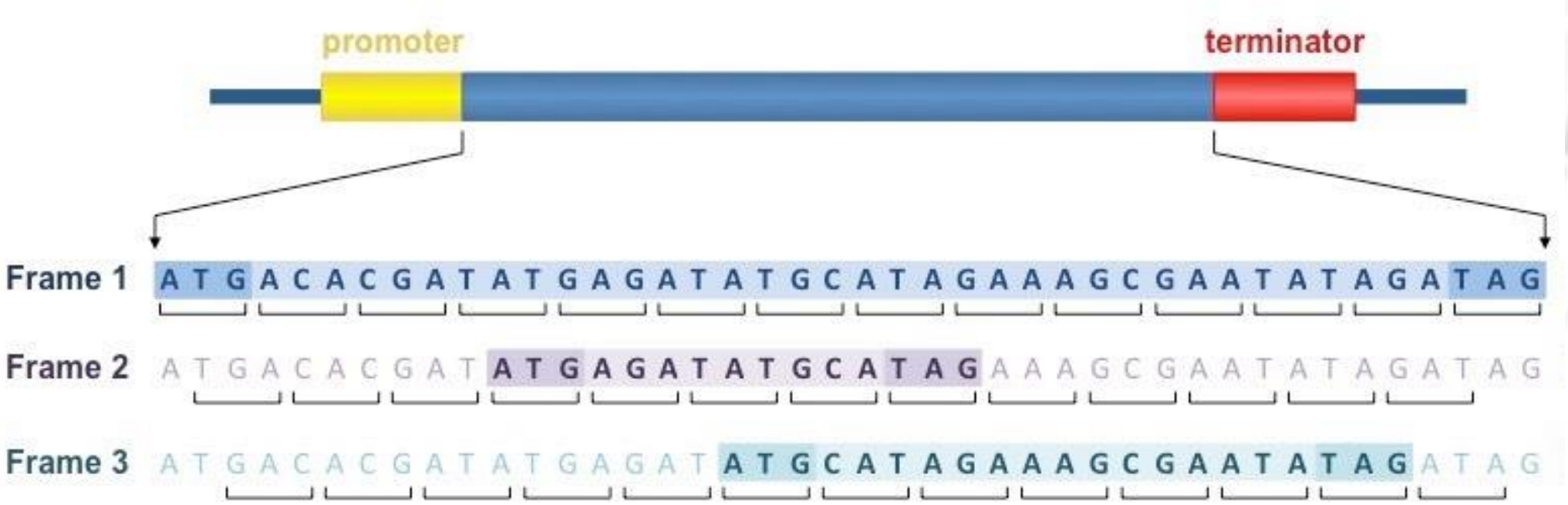

```
1   Atg ccc aag ctg aat agc gta gag ggg ttt tca tca ttt gag gac gat gta taa
     M   P   K   L   N   S   V   E   G   F   S   S   F   E   D   D   V   *

2   a Tgc cca agc tga ata gcg tag agg ggt ttt cat cat ttg agg acg atg tat aa
      C   P   S   *   I   A   *   R   G   F   H   H   L   R   T   H   Y

3   at Gcc caa gct gaa tag cgt aga ggg gtt ttc atc att tga gga cga tgt ata a
       A   Q   A   E   *   R   R   G   V   F   I   I   *   G   R   C   I
```

```
4   tac ggg ttc gac tta tcg cat ctc ccc aaa agt agt aaa ctc ctg cta cat atT
     H   G   L   Q   I   A   Y   L   P   K   *   *   K   L   V   I   Y   L

5   ta cgg gtt cga ctt atc gca tct ccc caa aag tag taa act cct gct aca taT t
       G   L   S   F   L   T   S   P   N   E   D   N   S   S   S   T   Y

6   t acg ggt tcg act tat cgc atc tcc cca aaa gta gta aac tcc tgc tac atA tt
      A   W   A   S   Y   R   L   P   T   K   N   N   Q   P   R   H   I
```

Open Reading Frame (ORF)

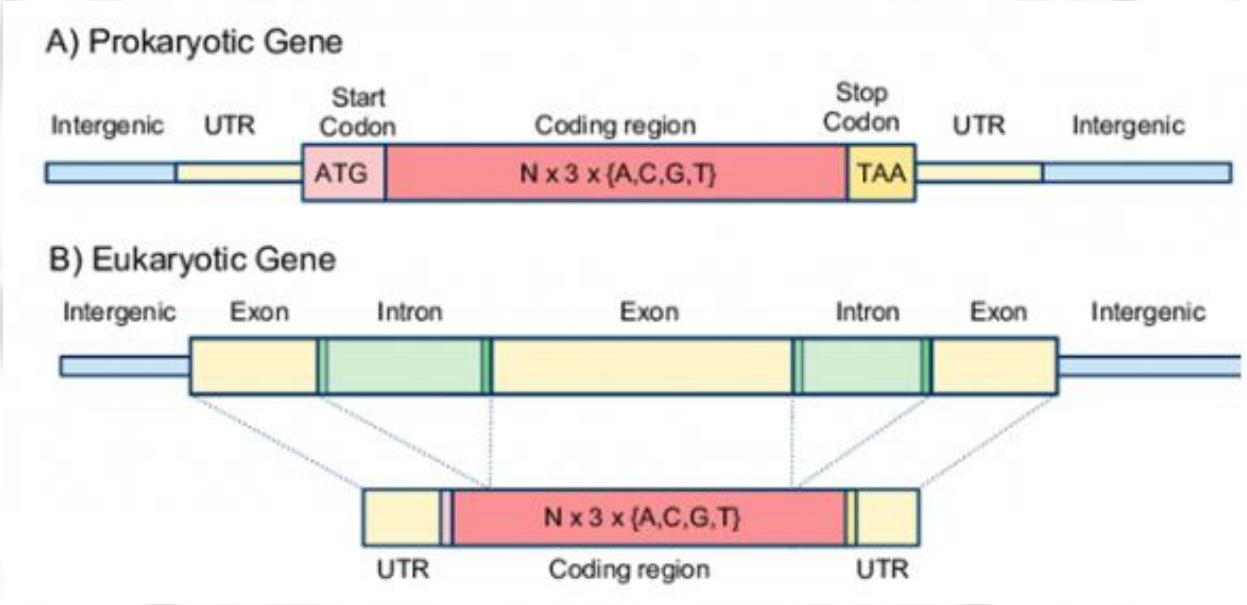
- Uma região de sequências de nucleotídeos presentes em uma mesma fase de leitura, a partir de um códon de ínicio (ATG) até um códon de parada (TAA, TAG, TGA)

- Múltiplas ORFs são possíveis... Qual realmente representa uma CDS?



Coding sequence (CDS)

- É a região do DNA que codifica uma proteína
- Em procariotos as CDS são iguais às ORFs, pela ausência de íntrons
- Em eucariotos as CDS não correspondem às ORFs devido à presença de íntrons



Principais abordagens de anotação

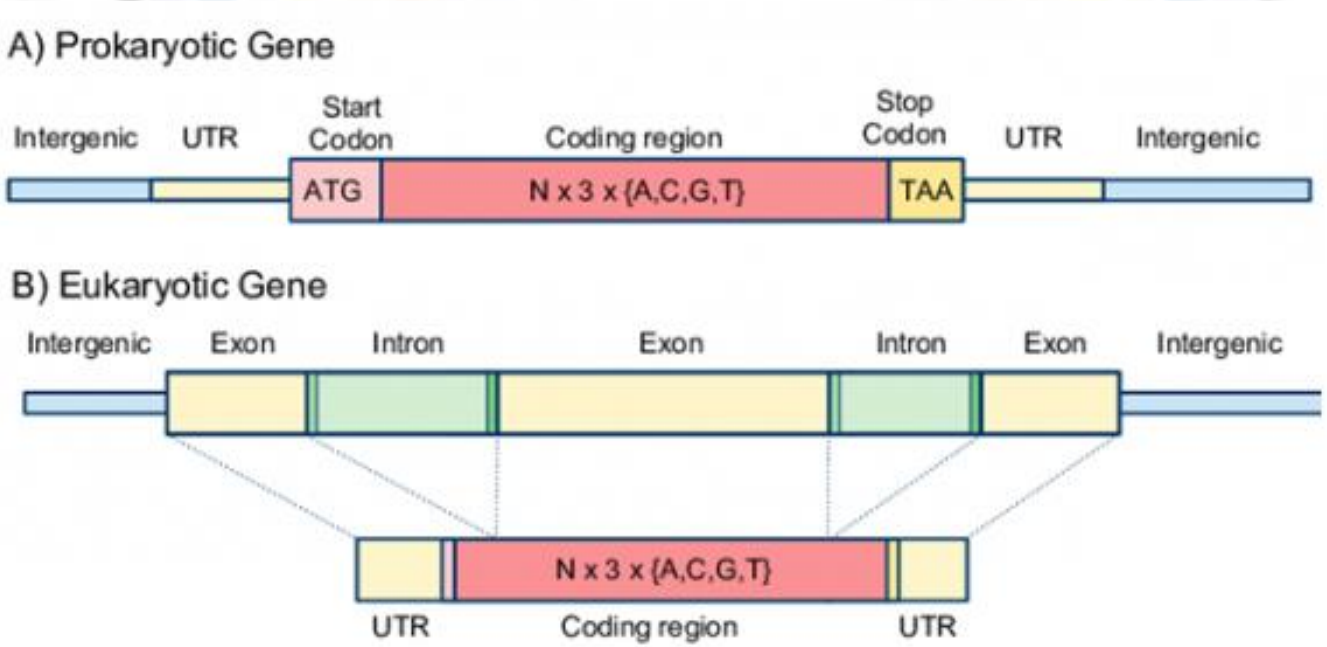
- *Ab initio*
- Baseada em homologia

Ab initio – Visão geral

- Diretamente baseada na sequência genômica analisada

- Modelos estatísticos treinados para encontrar características presentes em genes:

  - Códons de início e parada
  - Éxons e íntrons
  - Sítios de splicing
  - Sequências intergênicas
  - Sequências transcritas mas não traduzidas



Referência da Imagem

Ab initio – Vantagens

- Pode ser utilizada mesmo quando só há disponibilidade da sequência do genoma a ser analisado, sem informações adicionais como transcriptoma ou proteoma

- Alguns métodos mais avançados realizam um processo iterativo de anotação e auto-treinamento

Ab initio – Desvantagens

- Exigem um bom “treinamento” a partir de conjuntos de genes previamente anotados com alta qualidade para melhorar a precisão
- Melhores resultados com conjuntos de treinamento que apresentem tamanhos de éxons e íntrons similares e apresentem o mesmo tipo de sítios de splicing
- Os melhores conjuntos de treinamento normalmente são de espécies evolutivamente próximas, e construídos a partir de dados de homologia de sequências

Baseada em homologia

- Métodos que usam homologia para encontrar os genes
  - Evidências experimentais de genes e proteínas estudados e validados em abordagens funcionais
  - Alinhamentos com sequências de espécies evolutivamente próximas

- Abordagens principais
  - Baseada em RNA
  - Baseada em proteínas

Baseada em homologia - RNA

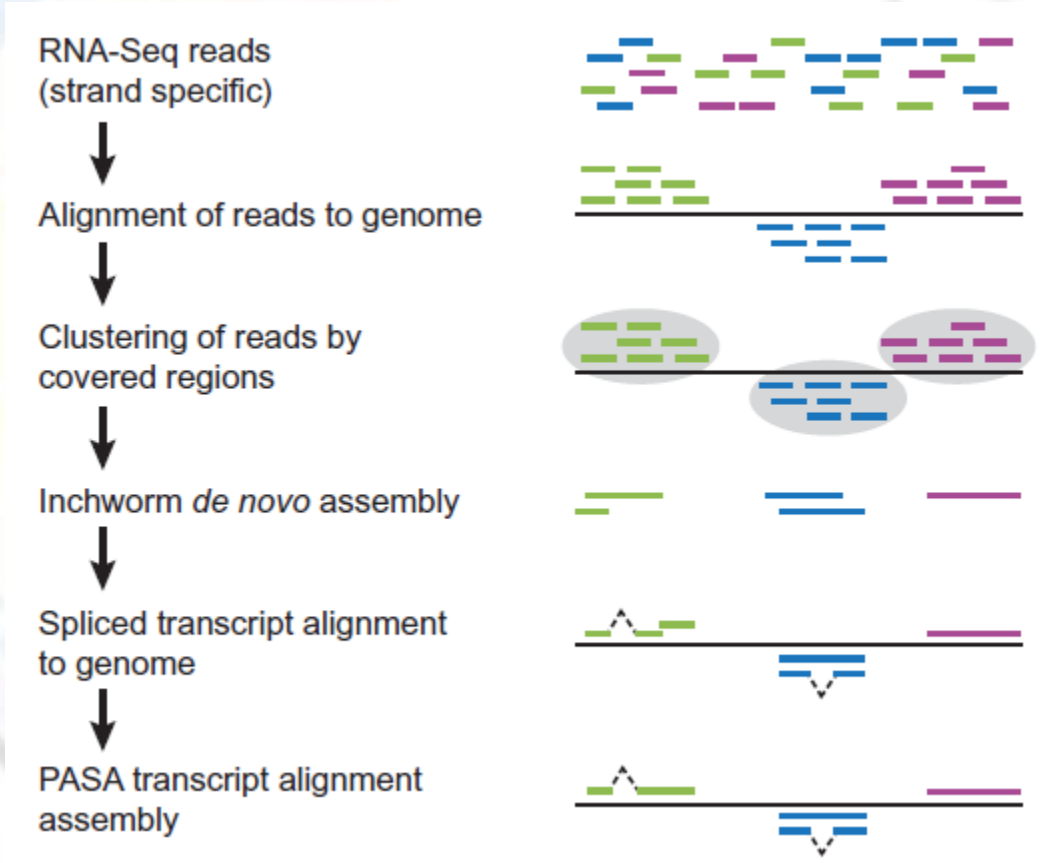
- Transcritos provenientes do mesmo organismo do qual a sequência do genoma foi obtido são um tipo de evidência muito preciso:
  - Altamente idênticos ao genoma
  - Determinação precisa dos limites entre éxons-íntrons
  - Determinação de sítio de poli-adenilação
  - Determinação das regiões UTRs
  - Determinação de transcritos alternativos

Baseada em homologia - RNA

- Sequências de transcritos podem ser provenientes de:
  - Expressed sequence tags (ESTs)
  - Sequências de cDNA de transcritos completos
  - Sequências de RNA-Seq, no formato de reads ou transcritos montados (transcriptoma)

Baseada em homologia – RNA-Seq

- Abordagens gerais
  - **Baseada em alinhamento:** alinhar os reads contra o genoma e montar os transcritos localmente com base nos alinhamentos
  - **Baseada em transcritos:** montar os reads em transcritos, e posteriormente alinhar contra o genoma para determinar as estruturas dos genes
  - Híbrida entre as abordagens anteriores



Adaptado de:
HAAS et al. 2011. **Mycology**. DOI: 10.1080\21501203.2011.606851

Baseada em homologia - Proteínas

- Proteínas são uma fonte de homologia importante para uma anotação, principalmente em termos funcionais
- Particularmente úteis quando não há informação de RNA disponível
- Proteínas conservadas podem auxiliar na anotação de espécies distantes entre si
- Limitação: restrita à proteínas que já tenham sido estudadas e caracterizadas

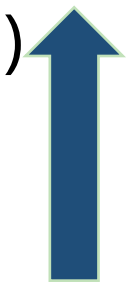
Abordagem híbrida: consenso entre *ab initio* e homologia

- Pipelines para execução de softwares de anotação ab initio e homologia

- Softwares desenvolvidos para realizar consenso entre todos os resultados, atribuindo pesos distintos para cada um dos métodos

- Em geral:

  - RNA (transcrito)
  - RNA (parcial)
  - Proteínas
  - *Ab initio*



Maior peso

Menor peso

Como avaliar a qualidade de uma anotação?

- Avaliação de conteúdo (BUSCO)
- Comparação com anotações prévias confiáveis
  - A quantidade de genes é similar?
  - Anotações funcionais resultam em informações similares?
- Avaliação manual

Formato GFF3 (General Feature Format)

- Uma feature por linha, 9 colunas delimitadas por tabulações

- **1 (seqid):** nome do cromossomo, contig ou scaffold em que a feature está localizada
- **2 (source):** nome do programa que gerou a anotação, ou da base de dados em que a anotação foi obtida
- **3 (type):** categoria da feature (ex: gene, CDS, mRNA, exon)
- **4 (start):** posição do início da feature no cromossomo, contig ou scaffold
- **5 (end):** posição do final da feature no cromossomo, contig ou scaffold
- **6 (score):** score de confiabilidade, mas muitos softwares não atribuem nenhum valor (.)
- **7 (strand):** indica se a feature está na fita direta/forward (+) ou reversa/reverse (-)
- **8 (phase):** fase de leitura
- **9 (attributes):** informações adicionais sobre a feature, como nome, ou vínculo com outra feature anterior

```
scaffold_10 EVM gene    2697    4438    .   +   .   ID=scaffold_10.1;Name=scaffold_10.1
scaffold_10 EVM mRNA    2697    4438    .   +   .   ID=scaffold_10.1;Parent=scaffold_10.1;Name=scaffold_10.1
scaffold_10 EVM exon    2697    3078    .   +   .   ID=scaffold_10.1.exon1;Parent=scaffold_10.1
scaffold_10 EVM CDS 2697    3078    .   +   0   ID=cds.scaffold_10.1;Parent=scaffold_10.1
scaffold_10 EVM exon    3145    4023    .   +   .   ID=scaffold_10.1.exon2;Parent=scaffold_10.1
scaffold_10 EVM CDS 3145    4023    .   +   2   ID=cds.scaffold_10.1;Parent=scaffold_10.1
scaffold_10 EVM exon    4068    4438    .   +   .   ID=scaffold_10.1.exon3;Parent=scaffold_10.1
scaffold_10 EVM CDS 4068    4438    .   +   2   ID=cds.scaffold_10.1;Parent=scaffold_10.1
```

Regiões codificantes (CDS) em formato FASTA

- **Linha 1:** identificador da sequência após o sinal de maior (>)
- **Linha 2:** sequência
- Cada sequência corresponde a um gene

```
>scaffold_10.1
ATGGACGGACCGCTCTCGACGCTGCTCTCTTACTCCAACGCCATTTTGCTGGTGCTGGGCCTCGCCGGGC
TGGCATACATCAGCTTCCGCGCCGCGTACGGCACCGACGTGGGGCGCATCACGGGCATTCCTGAGCCGGG
GCACGCGGTGGCGTTCTACGGACACCTCAACTCCAAGGCGCTCGGCAGCGACCACCCCACTGCGCTGCAG
GAGTATTCGGTGAAGAATGGGTGGCCGTTGGTGCAGGTGCGGTTTGGGCAGCGGCGGGTCGTGGTGCTGA
ATACGTTTGCGGCGGCGCAGCATTTCATCATTCGCAATGGGGGGGCGACAATTGACCGCCCGCTGTTTTG
GACATTTCATAAGTTTGTGAGCAATACGCAGGGCGCAACCATTGGCACGTCGCCGTGGGACGCATCGTGC
AAGCGCAAGCGCACGGCAATCGGGGCGTACATGACGCGGCCGGCCATCCAGCGCAATGCGCCACTCATCG
ACATCGAGGCGCTGGGGCTGGTCGAAGGCATCTTTAACGCCTCGCTGGACGACAACAACAACCCCAGTGT
CGAAGTGGACCCCCGCCTCTTTTTCCAGCGGGCTTCGCTCAACTTTGTGCTCATGCTCTGCTACGCGTCG
CGGTTCCCGGACATTGACGACCCGCTGCTGCACGAGATTCTGGCCACGGGCAAGACGGTCAGCACGTTTC
GCAGCACCAACAACAACATGGCCGACTACGTGCCGCTGCTGCGGTACCTGCCCAACGCGCGGACGGCGAT
GGCCAAGCAGGTGACCAAGAAGCGCGACGTGTGGCTCGAGGCGCTGCTGGAGCGCGTGCGCAAAGCCGTG
GCGGCCGGCAAGCCCGTGTCGTGCATTGCATCGTCGCTGCTCAAGGAAAAGGGGTCCGAGAAGCTGACAG
AGGCCGAGATTCGCTCCATCAACGTCGGGCTCGTCTCGGGCGGCAGCGACACGATTGCGACGACGGGGCT
CGGCGGGCTTGGGTTCCTCGCGTCCAAGGAGGGCCAGGCGATTCAGCAAAAGGCGTACGACGAGATTATG
AAGGTCTACGCGACGGCCGAGGAGGCGTGGGAGAATTGCGTGCTCGAGGAGAATGTCGAGTACGTCGTCG
CGCTCGTGCGCGAGATGCTGCGGTACTACTGCGCGATACAGCTGCTGCCACCGCGCAAGACGTGCAAGCC
GTTTGAGTGGCATGGCGCACAAATCCCTGCTGGTGTCACGGTATACATGAACGCGCAGGCTATCAATCAC
GACAAAACCGCATACGGACCAGACGCGCACATTTTCCGACCAGAGCGTTGGCTCGATCCCAGCAGTCCGT
ACCAGGTCGGGCTTCCCTACCACTACTCGTATGGCGCGGGCTCGCGAGCATGCACGGCCGTGGCGCTGTC
GAACCGGATTCTCTACTGCTACTTTGTGAGGCTGATTGTTTCGTTCCGCTTCACGGCCAGCGCAGACGCG
CCGCCGACGCTGGATTACATTGGATTCAACGAGAACCCGCAGGCGGCGACGGTCGTCCCAAAGACGTTTC
GGGTTAACATTGAGGAGAGGCGGCCGAGGGAGGAGCTGGCCAAGAATTTCGAGGCGAGTCGAAAGGCCAC
TTCTCACCTCGTCTTTACTTAG
```

Sequências de aminoácidos em formato FASTA

- **Linha 1:** identificador da sequência após o sinal de maior (>)
- **Linha 2:** sequência
- Cada sequência corresponde a um gene

```
>scaffold_10.1
MDGPLSTLLSYSNAILLVLGLAGLAYISFRAAYGTDVGRITGIPEPGHAVAFYGHLNSKALGSDHPTALQ
EYSVKNGWPLVQVRFGQRRVVVLNTFAAAQHFIIRNGGATIDRPLFWTFHKFVSNTQGATIGTSPWDASC
KRKRTAIGAYMTRPAIQRNAPLIDIEALGLVEGIFNASLDDNNNPSVEVDPRLFFQRASLNFVLMLCYAS
RFPDIDDPLLHEILATGKTVSTFRSTNNNMADYVPLLRYLPNARTAMAKQVTKKRDVWLEALLERVRKAV
AAGKPVSCIASSLLKEKGSEKLTEAEIRSINVGLVSGGSDTIATTGLGGLGFLASKEGQAIQQKAYDEIM
KVYATAEEAWENCVLEENVEYVVALVREMLRYYCAIQLLPPRKTCKPFEWHGAQIPAGVTVYMNAQAINH
DKTAYGPDAHIFRPERWLDPSSPYQVGLPYHYSYGAGSRACTAVALSNRILYCYFVRLIVSFRFTASADA
PPTLDYIGFNENPQAATVVPKTFRVNIEERRPREELAKNFEASRKATSHLVFT
>scaffold_10.2
MALQTCRRCRKRRIKCDLQLPACTSCQLVDLECLYFDDSLGHDVPRSYLHALSKKVENLESTINAIKSPA
AAAPSPTPFSQSDCPTPLQASLDPRGSSASSLGLGTSAGLLENLLKTLVQRSSTQDQSALSRFASRTRDV
EDDSALAFPPLKVNFSKLDTQSLQQPHLQRALIEYYAKTVQSSFPLLSKAQIDSLLRYEHPLRQCTAAER
LPIYGIFALASNLVSRDLDKDQSITASMWTERFHSYIAGFDSSNAHGAVRMKQNILALCFLALLDLVSPL
SPKGGVWEVVGAASRSYVKVLDDLSVSSPEIDDEFERLGHCIYLLESTLSIHFRIPSLYCNSAPTVIPSG
LSEPLVYHTLYTLTQLLNFPKDVSVDMESSIPACLRINLESGPSDVSLGQAQVYLTLHPLFTSPGAGIHC
CSPDLLSKIALAAAAFITHTHKLNKERRVVSIWVTAENVLQAGAAWAAYLMLHSQRDSPLHDYHVPKPID
KLPPMEPIVRCSSLLASFAERWKGGRRFCQAWEAFTELLLADDSLSKMATAPQA
```