Universidade de São Paulo
Instituto de Química

USP

Anotação de genomas

João Carlos Setubal

2020

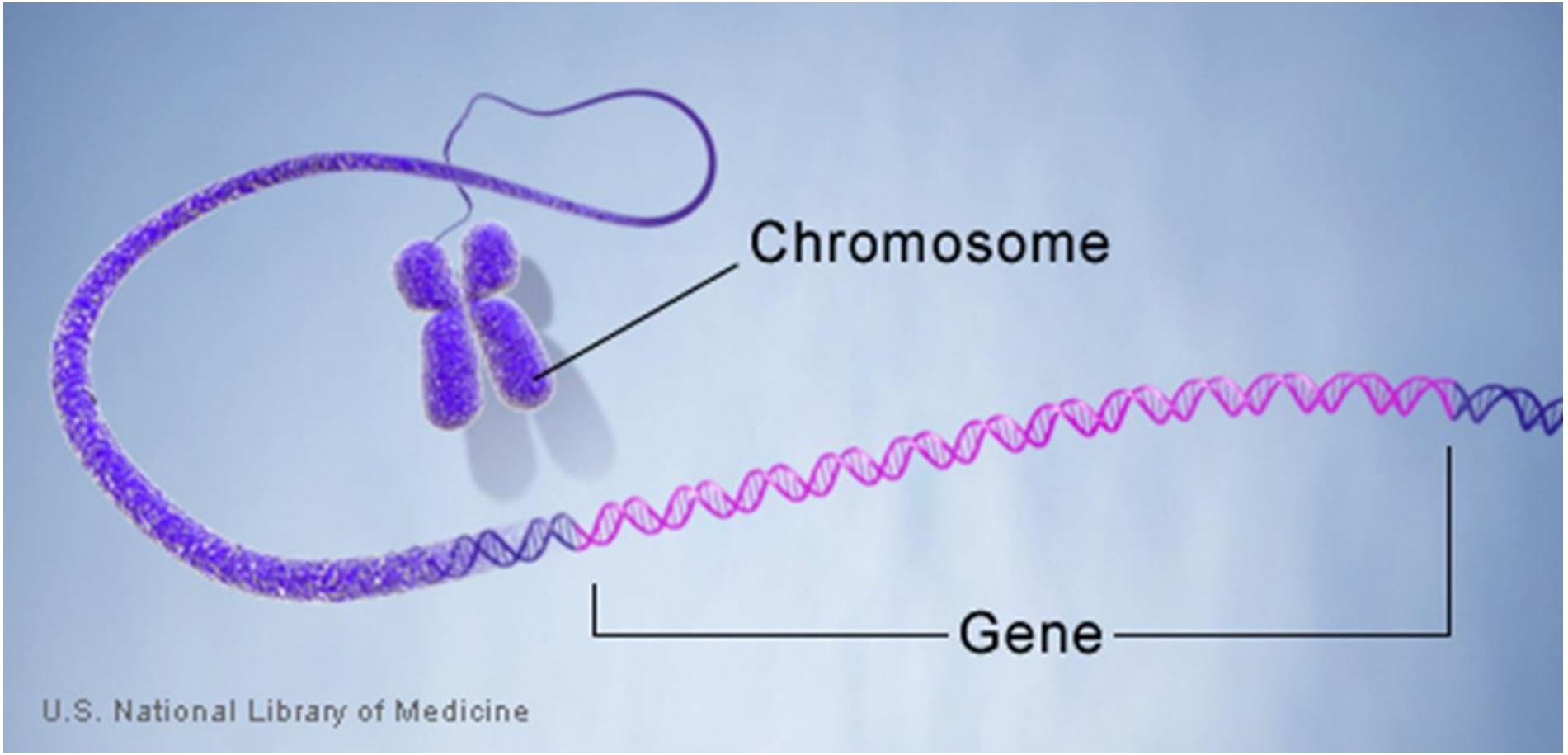
O que é anotação de um genoma?

- Dada a sequência em nucleotídeos de um genoma completo, sem buracos ou erros
- Achar os genes codificadores de proteína
  - Sequência codificadora (coding sequence, ou CDS) (às vezes usa-se (indevidamente) o termo ORF)
  - Promotores
  - Sítio de ligação ribossomal (RBS)
- Achar genes de RNA
  - RNA ribossomal
  - tRNA
  - Outros RNAs
- Atribuir função aos genes codificadores de proteína
- Esta aula: **genomas de procariotos**

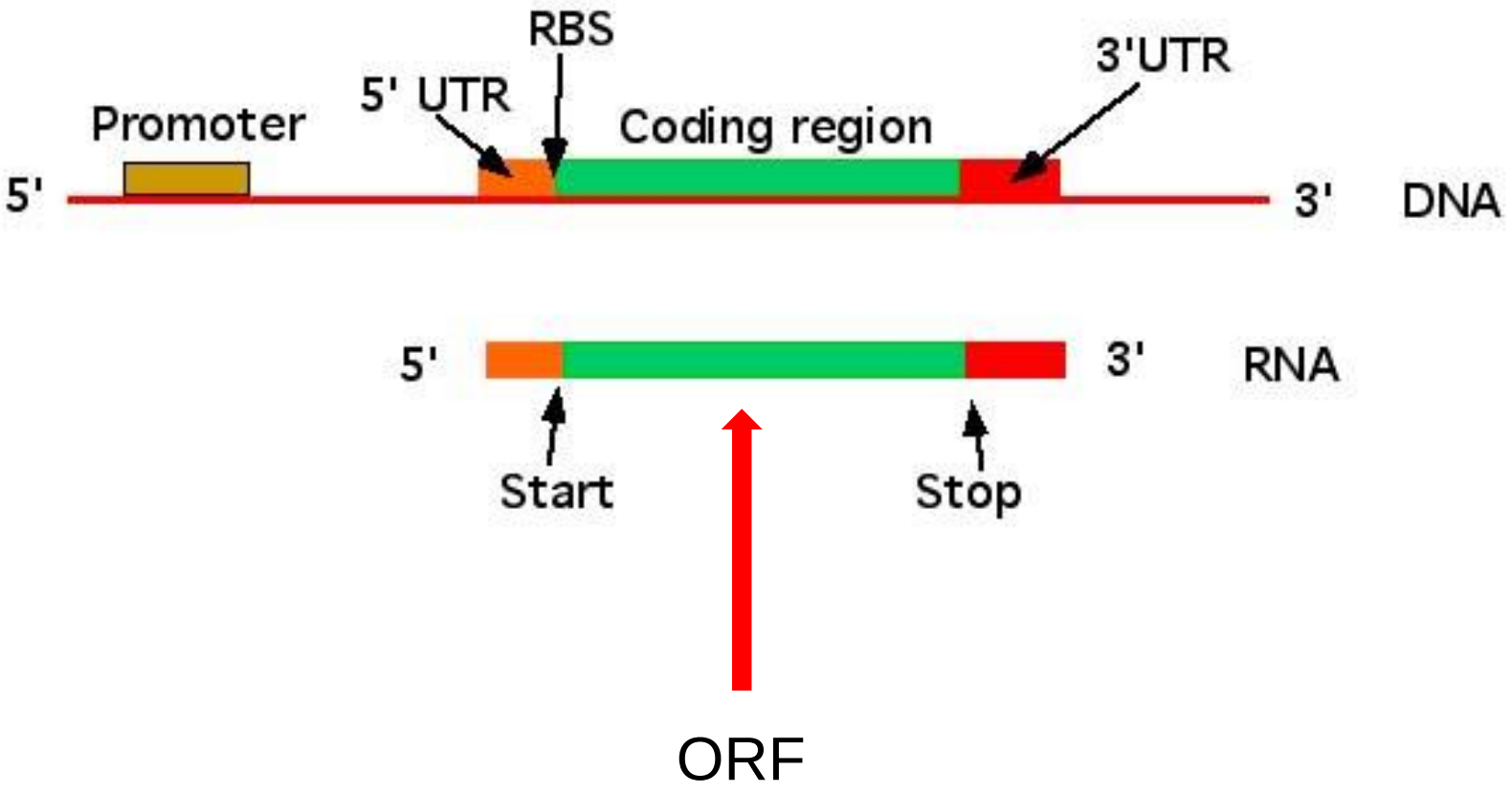
Achar genes codificadores de proteína

- *Gene finding*

Genes



Estrutura de um gene de procarioto



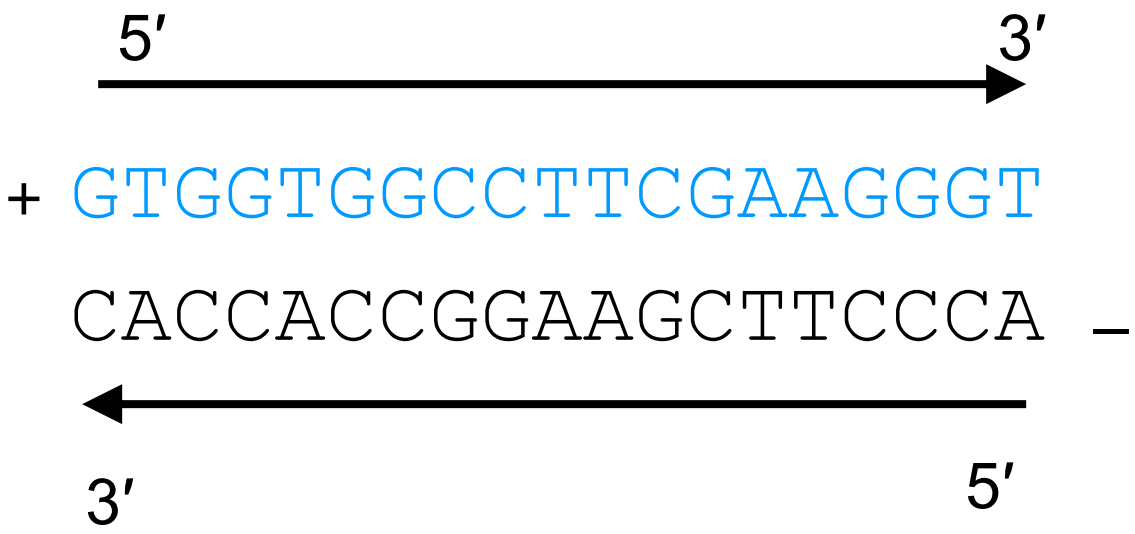
DNA tem *quadros de leitura*

+1: GTGGTGGCCTTCGAAGGGT

+2: TGGTGGCCTTCGAAGGGT

+3: GGTGGCCTTCGAAGGGT

DNA tem duas fitas (+ e –)



6 quadros no total

GTGGTGGCCTTCGAAGGGT

TGGTGGCCTTCGAAGGGT

GGTGGCCTTCGAAGGGT

CACCACCGGAAGCTTCCCA

CACCACCGGAAGCTTCCC

CACCACCGGAAGCTTCC

Trecho de DNA de bactéria; sem interpretação, não passa de uma sopa de letrinhas

...
AGCTCGCGCTCCGCATCCATCCAGTAGGGTTCGGTGTCGACGAGCGTGCC
GTCCATATCCCAGAAGACGGCGGCCGGCATCGCGTGCGGAGTCAGTTCGG
TCACGGCTGACAAGTCTATCCCGGCGGCCCCGGGCCTATTCTTGAGGGAC
GGCGTCCTGACCGGTCGCCGGATGAAAGGACCAGAACGCCCCGTGACTGA
CGCGAACAGCATCCTCGGAGGGCGCATCCTCATGGTGGCCTTCGAAGGGT
GGAACGACGCTGGCGAGGCCGCCAGCGGGGCCGTCAAGACGCTCAAGGAC
CAGCTGGATGTCGTCCCGGTCGCCGAGGTCGATCCCGAGCTGTACTTCGA
CTTCCAGTTCAACCGGCCGGTCGTCGCGGACGACGACGGCCGCCGGCGCC
TCATCTGGCCGTCCGCGGAGATCCTGGGCCCAGCTCGCCCCGGCGACACC
GGCGATGCGCGCCTGGACGCCACCGGCGCCAACGCGGGCAATATCTTCCT
TCTCCTCGGCACCGAGCCGTCGCGCAGCTGGCGCAGCTTCACCGCGGAGA
TCATGGATGCGGCCCTGGCCTCCGACATCGGCGCCATCGTCTTCCTCGGT
GCGATGCTGGCGGACGTACCGCACACCCGCCCCATCTCCATCTTCGCTTC
GAGCGAGAACGCGGCCGTCCGTGCGGAGCTCGGCATCGAACGCTCTTCGT
ACGAGGGGCCGGTCGGTATCCTGAGCGCGCTCGCCGAAGGGGCGGAGGAC
GTGGGCATTCCGACCATCTCCATCTGGGCGTCGGTTCCGCACTATGTCCA
CAATGCGCCCAGCCCGAAGGCGGTGCTCGCACTGATCGACAAGCTCGAAG
AGCTGGTGAATGTCACCATCCCGCGTGGCTCGCTGGTGGAGGAGGCCACG
GCCTGGGAAGCCGGGATCGACGCGCTGGCTCTGGACGACGACGAGATGGC
TACGTACATCCAGCAGCTGGAGCAGGCACGCGACACCGTGGACTCCCCTG
AGGCCAGCGGCGAGGCGATCGCCCAGGAGTTCGAGCGCTACCTCCGCCGC
CGCGACGGCCGCGCCGGCGATGACCCCCGCCGTGGCTGACGTCACCCCCT
CTCTGCGTCCGCCGTCCTCTGTTCCCCCCGCTCGGCCTCCCCTGAGGCCG
AGGAGTCGCGCCCACATGCCGGAAACTCCTCCTTTCCTGACTTTCTGGAG ...

Esse trecho tem uma CDS; RBS em vermelho

...
AGCTCGCGCTCCGCATCCATCCAGTAGGGTTCGGTGTCGACGAGCGTGCC
GTCCATATCCCAGAAGACGGCGGCCGGCATCGCGTGCGGAGTCAGTTCGG
TCACGGCTGACAAGTCTATCCCGGCGGCCCCGGGCCTATTCTTGAGGGAC
GGCGTCCTGACCGGTCGCCGGATGAAAGGACCAGAACGCCCCGTGACTGA
CGCGAACAGCATCCTCGGAGGGCGCATCCTCATGGTGGCCTTCGAAGGGT
GGAACGACGCTGGCGAGGCCGCCAGCGGGGCCGTCAAGACGCTCAAGGAC
CAGCTGGATGTCGTCCCGGTCGCCGAGGTCGATCCCGAGCTGTACTTCGA
CTTCCAGTTCAACCGGCCGGTCGTCGCGGACGACGACGGCCGCCGGCGCC
TCATCTGGCCGTCCGCGGAGATCCTGGGCCCAGCTCGCCCCGGCGACACC
GGCGATGCGCGCCTGGACGCCACCGGCGCCAACGCGGGCAATATCTTCCT
TCTCCTCGGCACCGAGCCGTCGCGCAGCTGGCGCAGCTTCACCGCGGAGA
TCATGGATGCGGCCCTGGCCTCCGACATCGGCGCCATCGTCTTCCTCGGT
GCGATGCTGGCGGACGTACCGCACACCCGCCCCATCTCCATCTTCGCTTC
GAGCGAGAACGCGGCCGTCCGTGCGGAGCTCGGCATCGAACGCTCTTCGT
ACGAGGGGCCGGTCGGTATCCTGAGCGCGCTCGCCGAAGGGGCGGAGGAC
GTGGGCATTCCGACCATCTCCATCTGGGCGTCGGTTCCGCACTATGTCCA
CAATGCGCCCAGCCCGAAGGCGGTGCTCGCACTGATCGACAAGCTCGAAG
AGCTGGTGAATGTCACCATCCCGCGTGGCTCGCTGGTGGAGGAGGCCACG
GCCTGGGAAGCCGGGATCGACGCGCTGGCTCTGGACGACGACGAGATGGC
TACGTACATCCAGCAGCTGGAGCAGGCACGCGACACCGTGGACTCCCCTG
AGGCCAGCGGCGAGGCGATCGCCCAGGAGTTCGAGCGCTACCTCCGCCGC
CGCGACGGCCGCGCCGGCGATGACCCCCGCCGTGGCTGACGTCACCCCCT
CTCTGCGTCCGCCGTCCTCTGTTCCCCCCGCTCGGCCTCCCCTGAGGCCG
AGGAGTCGCGCCCACATGCCGGAAACTCCTCCTTTCCTGACTTTCTGGAG ...

Quadro aberto de leitura (ORF)

- Um trecho contíguo do genoma em que
  - O número de nucleotídeos é múltiplo de 3
  - O último codon é de parada
  - O primeiro codon é de início de tradução (ATG)
  - Não existe nenhum outro codon de parada entre o codon de início e o codon de parada do final

Método (rudimentar) para achar genes em procariotos

Ache todas as ORFs com pelo menos 900 bp

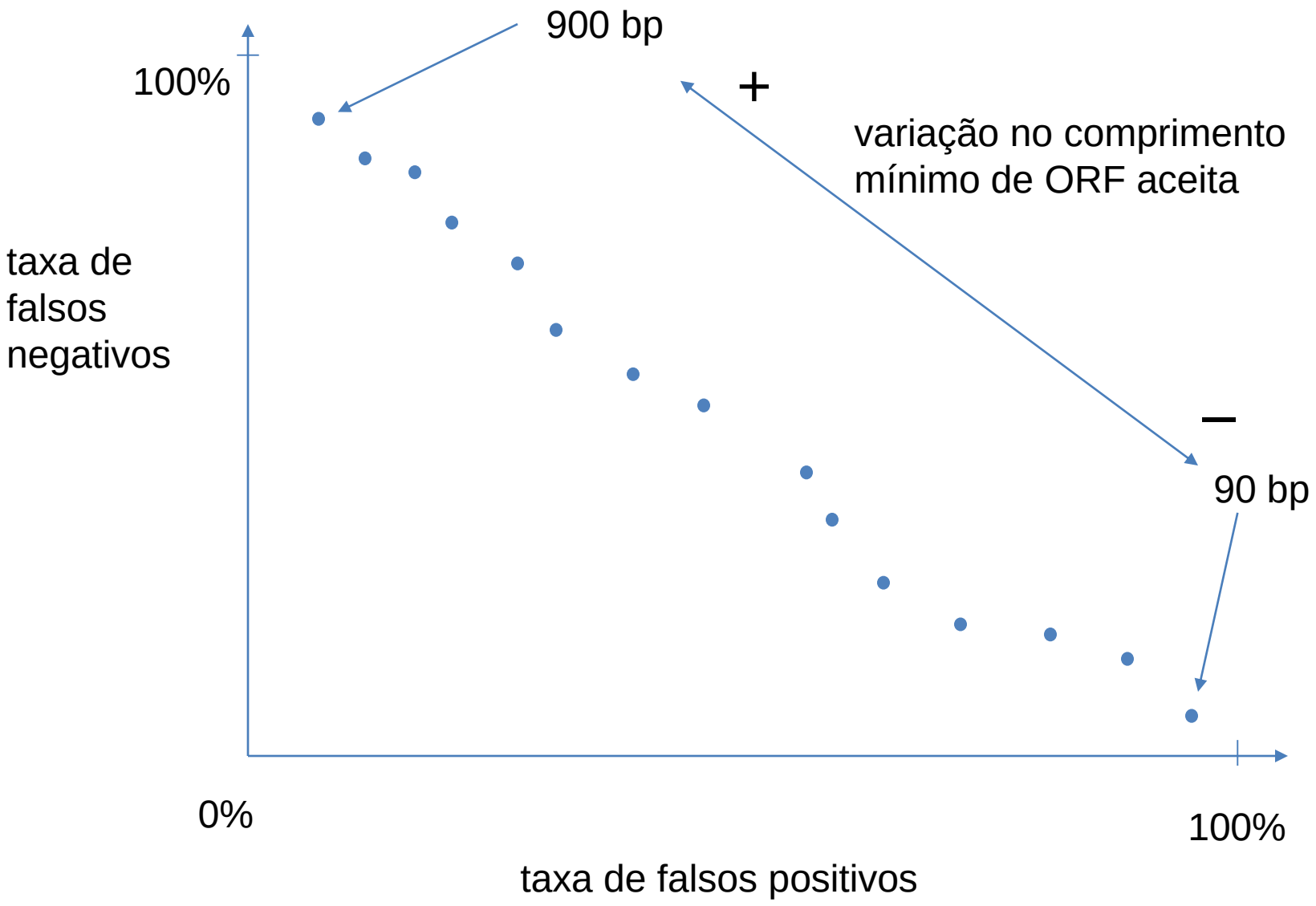
Bom quanto a falsos positivos

falsos positivos são ORFs com pelo menos 900 bp que não são CDSs

Ruim quanto a falsos negativos

falsos negativos são ORFs com menos de 900 bp que são CDSs

- 900 bp é um limiar muito conservador
- Gostaríamos de baixar esse limiar para diminuir a taxa de falsos negativos
- Mas sem aumentar muito a taxa de falsos positivos
- Uma ilustração (não é dado real) de como poderíamos abordar esta questão segue no próximo slide


900 bp
100%
+
variação no comprimento
mínimo de ORF aceita
taxa de
falsos
negativos
–
90 bp
0%
100%
taxa de falsos positivos

A ferramenta ORFfinder já “implementa” esse método rudimentar

- https://www.ncbi.nlm.nih.gov/orffinder/

Exercício

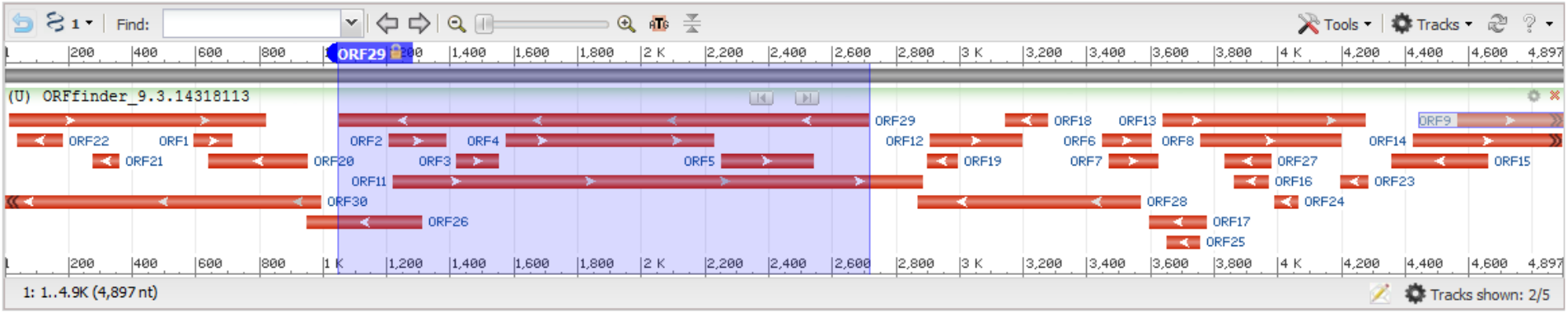
- Usando esta sequência
  - procure ORFs usando ORFfinder
  - você deve obter um resultado igual ao mostrado no próximo slide
  - brinque com os parâmetros
    - minimal ORF length
    - ORF start codon to use
    - Ignore nested ORFs

Open Reading Frame Viewer

Help

Sequence

ORFs found: 30 Genetic code: 1 Start codon: 'ATG' only



Six-frame translation...

ORF29 (555 aa) Display ORF as... Mark

```
>lcl|ORF29
MCIMGFSAAGKGKKMTSGRQSFVQGAAILGAAAFFTKLLGAVYRVPYQNI
TGNEGMFVYQQVYPLYSTLLILATAGFPLAISKLVSERLAEGDEAGARRV
FIVSSVTLTLTGFLFFLLLFAGAPWIAGWMGNREWLTLPIRAVSFALLVV
PLMSAIRGYFQGHQNMIPTALSQSVEQVVRVATILFAAWWFMSREGDVVS
AGAGAVFGAFTGAVGALLVLLMFLRGSGLLRQVPQAGPGAFAGRDSVRYV
AREIWRLSLPICLGSLVLPLFSLVDSFTVANLLKWSGWSVASSVEAKGIF
DRGQPLIQFASFFATAIALSIVPAVAEAKARGESGKMEERSSLAFRLTLL
LGLPASVGLAVVVRSANVMLFEDASGSDALAILALTTLFYTLGVTSAGIL
QGMGNVILPARNLLAGVAVKLVLNVLLIPWWDIRGAAAATLIAYAAATFL
NLAAFRGQLGRMFRLRETGWALATATLLMAGAAAGAQALLTAVTADWASH
RLAMTLTSLGSVAAGAAVYGIVLLGLGGVRREELRVVPKWGPRLVSLLES
RGLIR
```

Mark subset... Marked: 0 Download marked set as Protein FASTA

Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF29	-	**3**	**2717**	**1050**	**1668 | 555**
ORF11	+	3	1218	2885	1668 | 555
ORF30	-	3	992	>3	990 | 329
ORF10	+	3	15	821	807 | 268
ORF28	-	3	3569	2868	702 | 233
ORF4	+	1	1573	2229	657 | 218
ORF13	+	3	3639	4274	636 | 211
ORF14	+	3	4425	>4895	471 | 156
ORF8	+	2	3755	4198	444 | 147

ORF29

SmartBLAST

BLAST

Marked set (0)

SmartBLAST best hit titles...

BLAST

BLAST Database:

UniProtKB/Swiss-Prot (swissprot)

Método (um pouco melhor) para achar genes em procariotos

1. Ache todas ORFs
2. Traduza cada uma usando o código genético
3. Compare cada uma com seqüências de genes conhecidos
   - Se achar algum *hit* estatisticamente significativo, guarde; senão jogue fora
   - vamos estudar esta etapa em detalhes em aulas futuras
   - Você deve experimentar isto com o botão BLAST do ORFfinder
4. As sobreposições precisam ser resolvidas (ou seja, em geral, não podem existir CDSs que se sobrepõem, mesmo em fitas diferentes)

BLAST

- É um programa que permite comparar sequências, tanto em nt (blastn), quanto em aa (blastp), quanto uma mistura (blastx, tblastx)
- É a ferramenta de busca no site do NCBI e em outros sites
- Vamos estudá-lo em detalhe em aulas futuras

Na prática

- São usados métodos que usam técnicas bem mais sofisticadas
- Buscam padrões estatisticamente significativos no DNA
- Teoria: a composição em nucleotídeos das CDSs dos genes codificadores de proteína segue um padrão que é diferente das demais regiões
- Nome de uma dessas técnicas: modelos de Markov de estados ocultos (Hidden Markov Models, ou HMMs)

Programas mais usados

- Glimmer
  - http://ccb.jhu.edu/software/glimmer/index.shtml
- Prodigal
  - http://prodigal.ornl.gov/
- geneMark
  - http://exon.gatech.edu/

Limitações

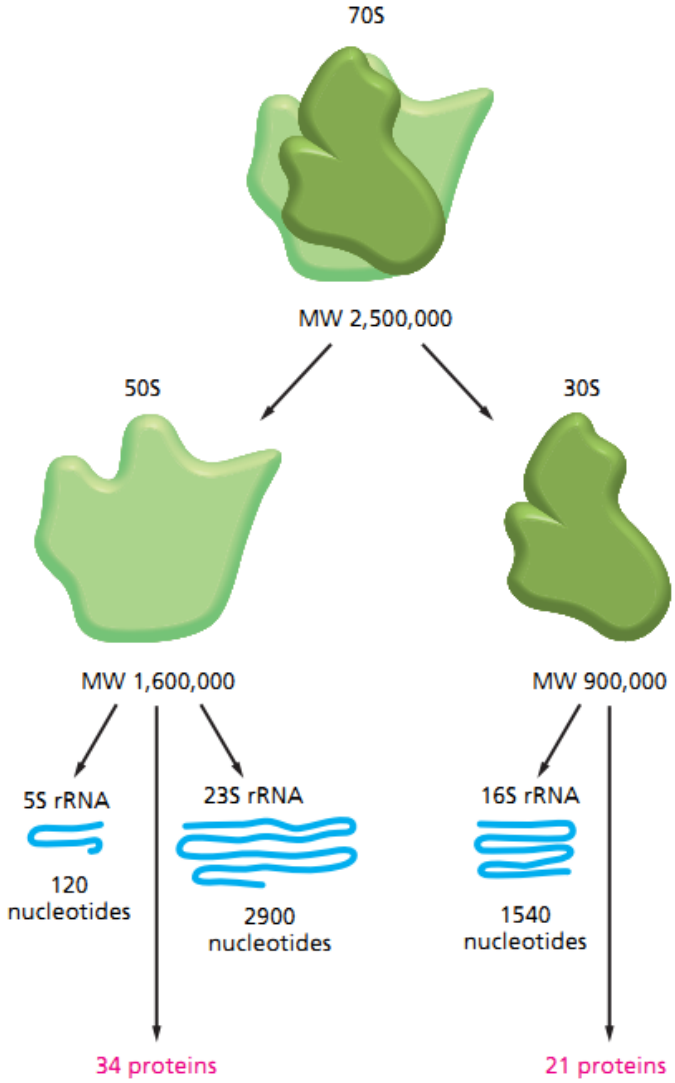
- Genes pequenos (menores do que 150 bp) geralmente são perdidos
  - Se aumentamos a sensibilidade (ou seja, abaixamos o tamanho mínimo dos genes que queremos encontrar), aparecem muitos falsos positivos
- Início de tradução nem sempre é correto

Achar genes de RNA

- RNA ribossomal
  - Operon
    - 16S, 5S, 23S
- tRNA
- Outros RNAs

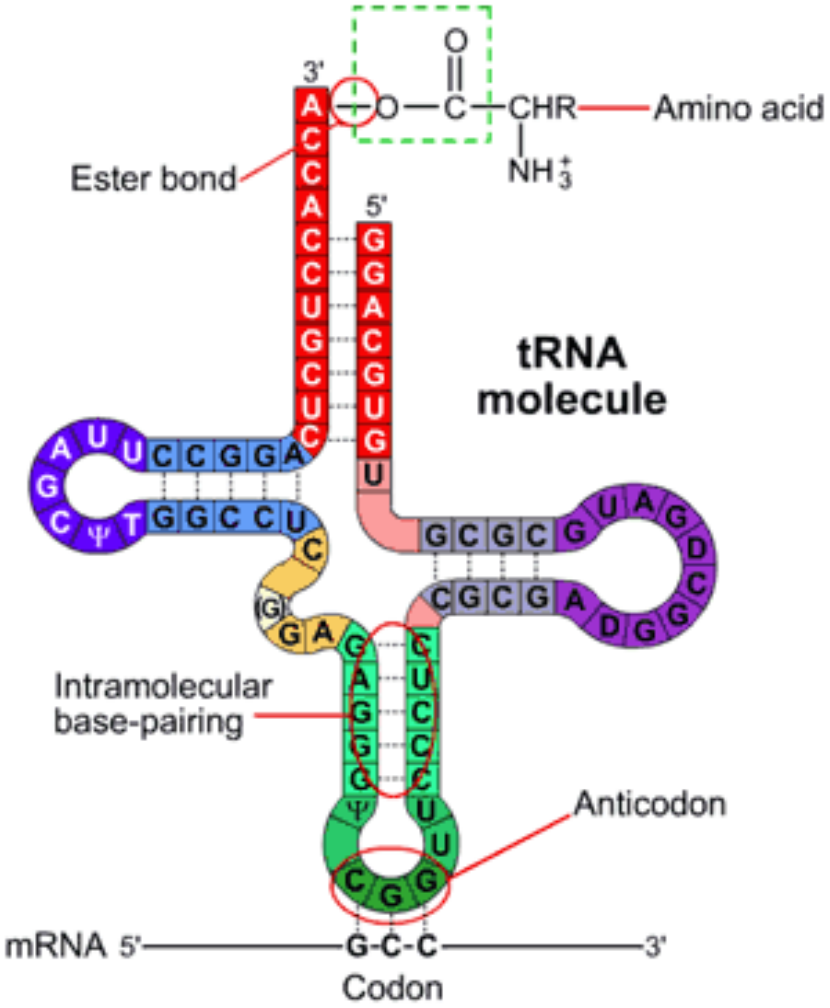
RNA ribosomal de procariotos

unidades do rRNA



Alberts et al. 2008

tRNA



Em procariotos tipicamente existem cerca de 50 genes de tRNA

Outros RNAs

- tmRNA
  - Resgata ribossomos emperrados
- Ribonuclease P RNA
- 6S RNA
  - Regulação gênica por ligação com RNA polimerase
- SRP RNA
- etc

Como achá-los?

- rRNA
  - similaridade: BLASTN, RNAmmer
  - características intrínsecas: programa Infernal
  - É difícil determinar as fronteiras exatas dos genes
- tRNA
  - tRNAscan-SE
  - Aragorn
- Outros RNAs
  - RFAM (por similaridade ou por Infernal)

RFAM

Famílias de RNA são descritas por esse grupo na Wikipedia

Infernal: inference of RNA alignments

Overview:

Infernal ("INFERence of RNA ALignment") is for searching DNA sequence databases for RNA structure and sequence similarities. It is an implementation of a special case of profile stochastic context-free grammars called *covariance models* (CMs). A CM is like a sequence profile, but it scores a combination of sequence consensus and RNA secondary structure consensus, so in many cases, it is more capable of identifying RNA homologs that conserve their secondary structure more than their primary sequence.

The latest release of Infernal is **1.1.1 [23 July 2014].**

Documentation:

- **User's Guide [PDF, 125 pages]** .
- **README** from the current release.
- **Release notes** for the current release.

Reference:

- The recommended citation for using Infernal 1.1 is E. P. Nawrocki and S. R. Eddy, **Infernal 1.1: 100-fold faster RNA homology searches** , *Bioinformatics* 29:2933-2935 (2013).

Download:

- The current source code distribution: **infernal 1.1.1, source only [tarball, 19.5 MB]**
  Source with binaries: **infernal 1.1.1 with Linux/Intel binaries [tarball, 32.1 MB]** , **infernal 1.1.1 with MacOSX/Intel binaries [tarball, 32.0 MB]** , **infernal 1.1.1 with Windows/Cygwin binaries [tarball, 35.1 MB]** , **README for using Cygwin binaries in Windows** ,
  Infernal is **freely available** under the GNU General Public License version 3 **[GPLv3]**.

Contact us:

- We welcome bug reports, feature requests, and code contributions. Email us at: **infernal@janelia.hhmi.org** .

Rfam CMs:

- You can download a single file with all 2450 Rfam release 12.0 CMs in Infernal 1.1 format **here**. Infernal 1.1's cmfetch program can be used to fetch individual CMs from this file.

Internal benchmark:

- The Infernal 1.1 Bioinformatics publication contains results from our internal RMARK3 benchmark. Files necessary for reproducing that benchmark are available **here**.

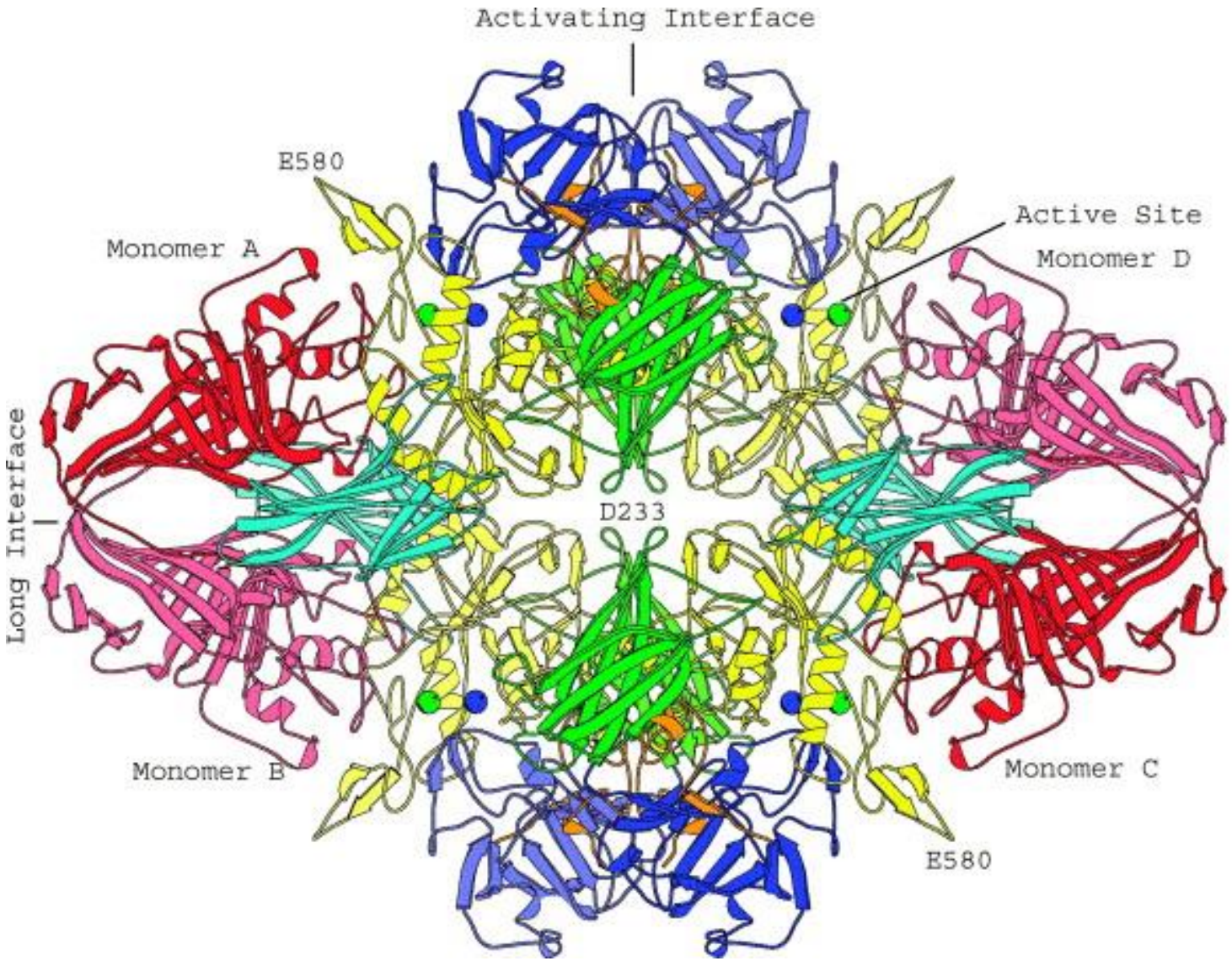
Further reference:

- **This book chapter** explains how to use Infernal and Rfam to annotate RNAs in genomes.
- The **Rfam database** of RNA families is based on the Infernal software. The most recent release 12.0 is described in **(Nawrocki, 2015)**

Anotação funcional

atributo	exemplo
Nome da proteína	Beta-galactosidase
Nome do gene	**lacZ**
organismo	***Escherichia coli* (strain K12)**
comprimento	1024 AA
função	Hydrolysis of terminal non-reducing beta-D-galactose residues in beta-D-galactosides
sequencia	MTMITDSLAVVLQRRDWENPG VTQLNRLAA(…)
estrutura	Próximo slide
Evidência de existência	Referências da literatura

Número EC, sítios ativos, interações, massa, etc



R.H. Jacobson, X.-J. Zhang, R.F. DuBose, B.W. MatthewsThree-dimensional structure of β-galactosidase from *E. coli* Nature, 369 (1994), pp. 761–766

B.W. Matthews, C. R. Biologies 328 (2005)

Como anotar?

- Manualmente
  - Seguir protocolos
  - Impraticável para a avalanche de genomas que existe hoje
- Automaticamente
  - Pipelines de anotação

The Standard Operating Procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4)

Marcel Huntemann[1], Natalia N. Ivanova[1], Konstantinos Mavromatis[1,3], H. James Tripp[1], David Paez-Espino[1], Krishnaveni Palaniappan[2], Ernest Szeto[2], Manoj Pillay[2], I-Min A. Chen[2], Amrita Pati[1],Victor M. Markowitz[2], Nikos C. Kyrpides[1]

[1]Genome Biology Program, Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, USA

[2]Biosciences Computing, Computational Research Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, USA
[3]Current address: Computational Biology Group, Celgene Corporation

Abstract

The DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4) performs structural and functional annotation for microbial genomes datasets that are then included into the Integrated Microbial Genome (IMG) comparative analysis system. MGAP is applied on assembled nucleotide sequence datasets that are provided via the IMG submission site (http://img.jgi.doe.gov/submit). Dataset submission for annotation first requires project and associated metadata description in GOLD (http://www.genomesonline.org/). The MGAP sequence data processing consists of feature prediction including identification of protein-coding genes, non-coding RNAs and regulatory RNA features, as well as CRISPR elements. Structural annotation is followed by assignment of protein product names and functions.

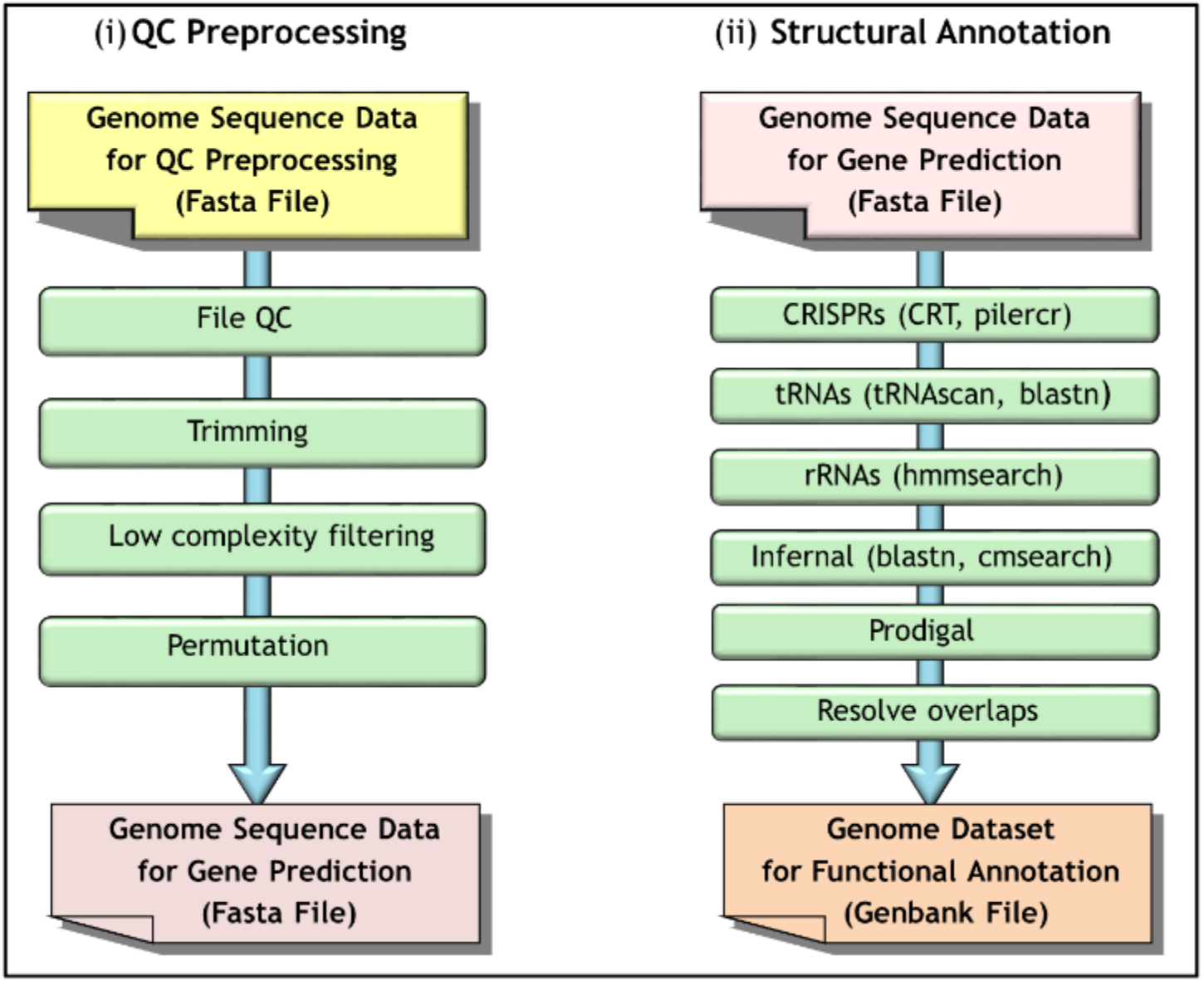


Figure 1. Genome sequence data preprocessing and structural annotation steps.

Protein Families

1. **COG & KOG assignment**: protein sequences are compared to COG PSSMs obtained from the CDD database [11] using the program RPS-BLAST at an e-value cutoff of 1e-2, with the top hit retained. The alignment length needs to be at least 70% of the consensus sequence length.

2. **KEGG Orthology (KO) term assignment**: Genes are associated with KO terms [12] as follows. First, the genes that can be unambiguously mapped to the entries in KEGG Genes database are assigned the KO terms associated with the corresponding KEGG gene. The gene to KEGG gene mapping is based on NCBI's GI numbers and GeneIDs. For genes that are not mapped to KEGG genes, USEARCH is run against the database of KEGG genes by applying UBLAST [18]. The results of this search are organized in a list of candidate KO assignments. KO terms are assigned to genes using a subset of this list, whereby the threshold is defined by an E-value cutoff of 1e-5, KO assignments are selected from the top 5 hits, with 30% or better alignment sequence identity, and alignment percentage of at least 70% over the length of the query gene and KEGG subject gene.

3. **MetaCyc assignment**: genes are associated with MetaCyc [13] reactions as follows. First, genes are mapped to KO terms as described above, whereby KO terms are associated with Enzyme Commission numbers (EC numbers) using the KEGG KO term to Enzyme relationship provided by KEGG. Next, genes are associated with MetaCyc reactions via EC numbers.

4. **Pfam & TIGRfam assignments**: protein sequences are searched against Pfam [14] and TIGRfam [15] databases using HMMER 3.0. For TIGRfam, the noise cutoff (--cug_nc) is used, with hits below the trusted cutoff and at/above the noise cutoff flagged as "marginal". For Pfam, the gathering threshold (--cut_ga) is used inside the pfam_scan.pl script (see: ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/OldPfamScan/HMMER2/pfam_scan.pl). The script also helps resolving overlaps between hits to Pfam models from the same clan in order to generate final Pfam assignments.

5. **InterPro Scan**: Additional protein family annotations for SMART, PrositeProfiles, PrositePatterns, and SuperFamily are provided by InterPro Scan (run with default parameters) [16].

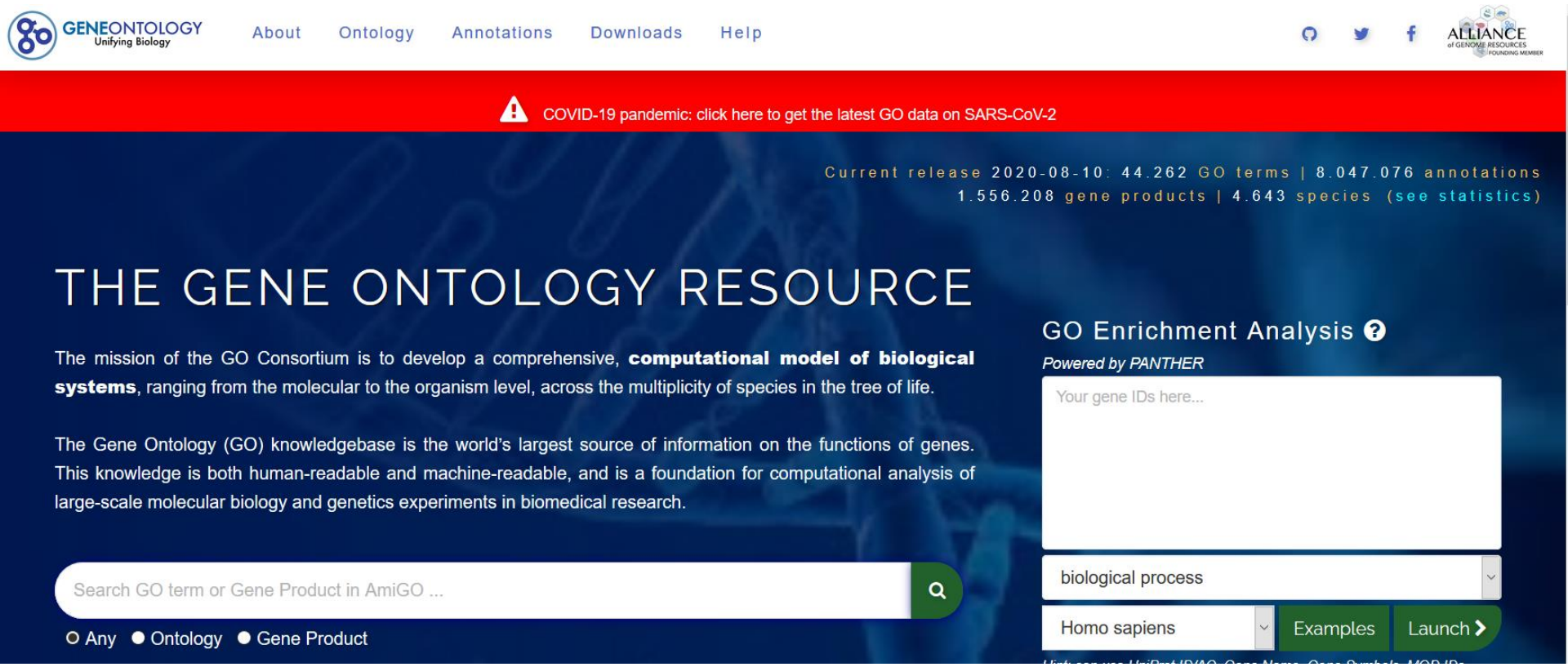
Plataformas de anotação

- NCBI
  - É possível submeter um genoma para o NCBI (apenas a sequência) e pedir que ele seja anotado
  - Eles rodam o pipeline PGAP no genoma
  - https://www.ncbi.nlm.nih.gov/genome/annotation_prok/
- RAST
  - Rapid Annotation using Subsystems Technology
  - https://rast.nmpdr.org/

O problema dos termos

- Diferentes pessoas usam diferentes palavras para descrever a mesma função
  - mixirica, tangerina, bergamota
  - sinônimos
- Diferentes pessoas usam as mesmas palavras para descrever funções diferentes
  - manga (a fruta, de camisa)
  - homógrafos
- É necessário uma padronização
  - *Gene Ontology*

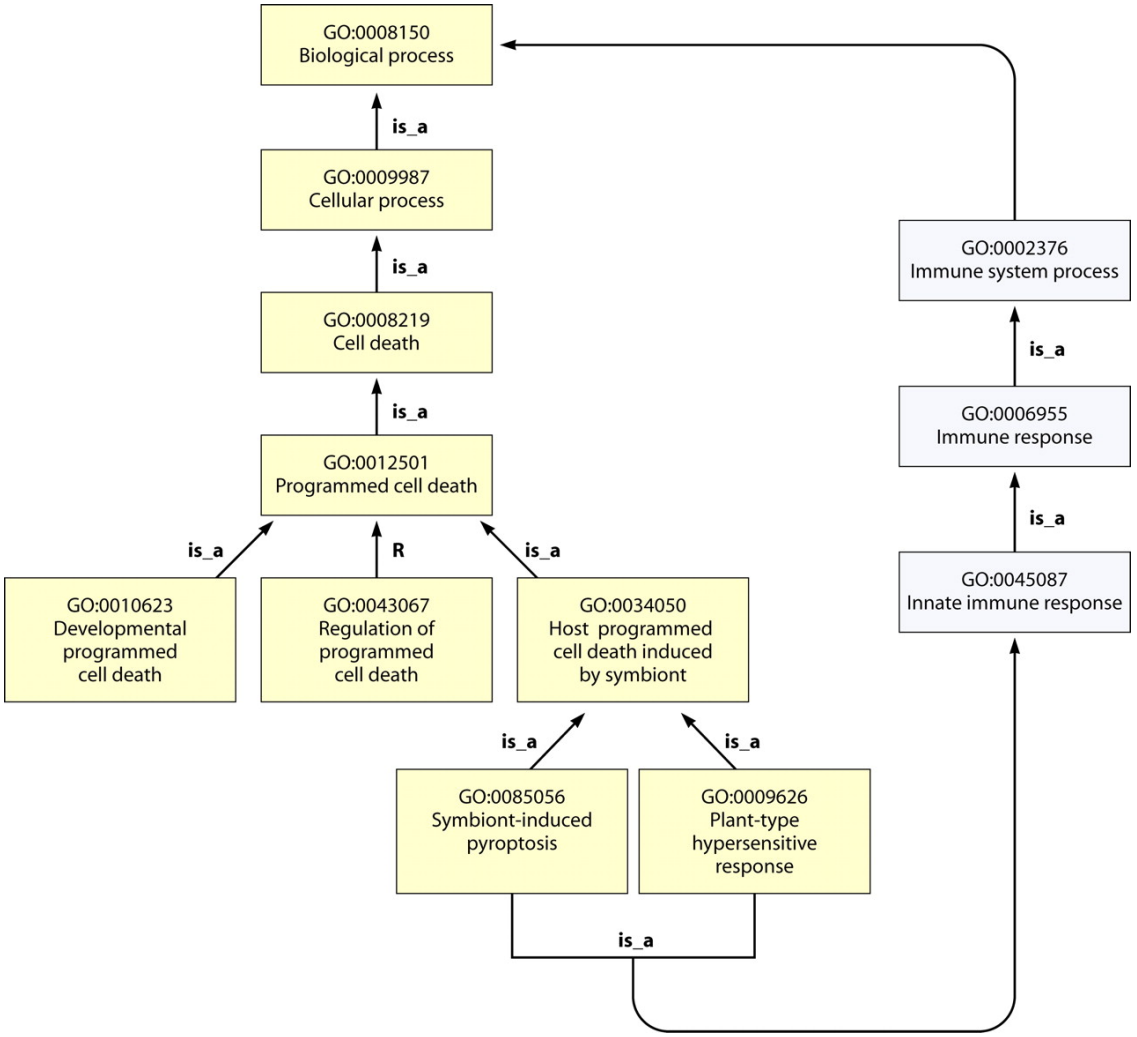
http://geneontology.org



Gene Ontology

- Sistema que faz 2 coisas básicas
  - Padroniza os termos
  - Padroniza a relação entre eles
- 3 grandes áreas
  - Função molecular
  - Processo biológico
  - Componente celular

Simplified directed acyclic graph (DAG) illustrating several terms describing different types of programmed cell death (PCD).



Trudy Torto-Alalibo et al. Microbiol. Mol. Biol. Rev. 2010;74:479-503

Microbiology and Molecular Biology Reviews

A padronização de GO permite processamento computacional

- esse é o principal uso dos números de GO
- exemplo
  - enriquecimento de funções

Análise de enriquecimento

- Muito comum em expressão gênica
  - genes diferencialmente expressos em condição A em relação a um controle (para + ou para -)
- Há um enriquecimento de categorias GO (ou COG, etc) dos genes diferencialmente expressos?
  - Super-representação
  - Sub-representação
- Análise de enriquecimento depende de um teste estatístico para afirmar se existe super- ou sub-representação

Códigos de evidência

- Usados no processo de anotação para indicar como a anotação foi feita

supported the association of a GO term. The *Experimental Evidence codes* are:

- Inferred from Experiment (EXP)
- Inferred from Direct Assay (IDA)
- Inferred from Physical Interaction (IPI)
- Inferred from Mutant Phenotype (IMP)
- Inferred from Genetic Interaction (IGI)
- Inferred from Expression Pattern (IEP)

estes são os códigos para inferência experimental

Use of the computational analysis evidence codes indicates that the annotation is based on an in silico analysis of the gene sequence and/or other data as described in the cited reference. The evidence codes in this category also indicate a varying degree of curatorial input. The *Computational Analysis evidence codes* are:

- Inferred from Sequence or structural Similarity (ISS)
- Inferred from Sequence Orthology (ISO)
- Inferred from Sequence Alignment (ISA)
- Inferred from Sequence Model (ISM)
- Inferred from Genomic Context (IGC)
- Inferred from Biological aspect of Ancestor (IBA)
- Inferred from Biological aspect of Descendant (IBD)
- Inferred from Key Residues (IKR)
- Inferred from Rapid Divergence(IRD)
- Inferred from Reviewed Computational Analysis (RCA)

em anotação automática, este é o código mais comum

Author statement codes indicate that the annotation was made on the basis of a statement made by the author(s) in the reference cited. The *Author Statement evidence codes* used by GO are:

- Traceable Author Statement (TAS)
- Non-traceable Author Statement (NAS)

Use of the curatorial statement evidence codes indicates an annotation made on the basis of a curatorial judgement that does not fit into one of the other evidence code classifications. The *Curatorial Statement codes* are:

- Inferred by Curator (IC)
- No biological Data available (ND) evidence code

All of the above evidence codes are assigned by curators. However, GO also used one evidence code that is assigned by automated methods, without curatorial judgement. The *Automatically-Assigned evidence code* is:

- Inferred from Electronic Annotation (IEA)

Evidence codes are **not** statements of the quality of the annotation. Within each evidence code classification, some methods produce annotations of higher confidence or greater specificity than other methods, in addition the way in which a technique has been applied or interpreted in a paper will also affect the quality of the resulting annotation. Thus evidence codes **cannot** be used as a measure of the quality of the annotation.

Gene Ontology não padroniza nomes de proteínas

- ‘gene symbol’ (exemplo: lacZ)
- Ou mesmo…
  - A frase curta que supostamente descreve a função dos genes (a proteína)
    - exemplo: gene *metK*
      - methionine adenosyltransferase
      - S-adenosylmethionine synthase
      - Estas duas são equivalentes
- Então alguns problemas de comunicação persistem

Propagação de erros

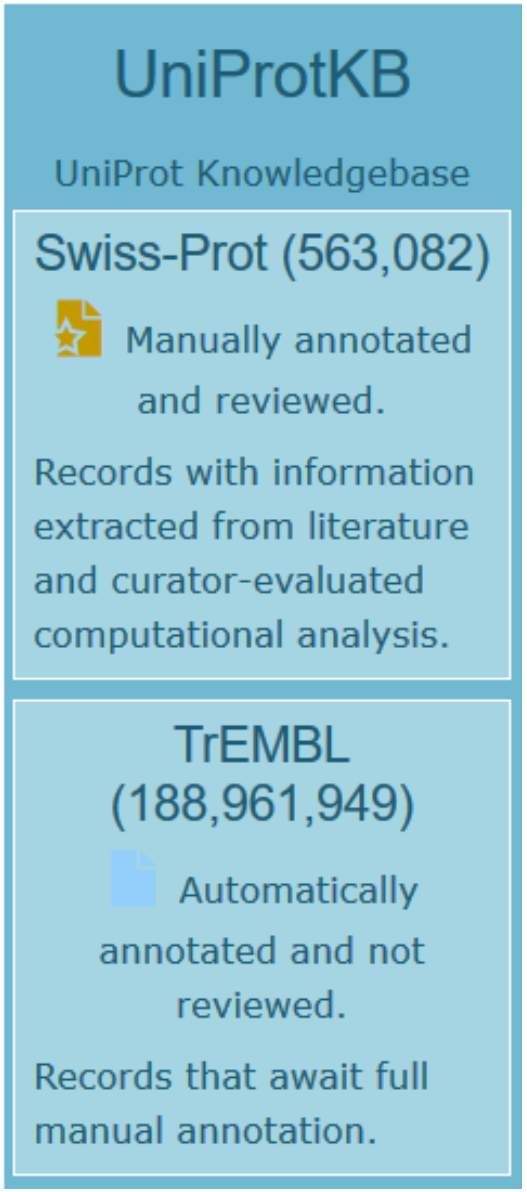
- Já faz anos que estamos convivendo com um tsunami de sequências por causa da revolução genômica
- Anotação automática de genomas é essencial
- uma técnica utilizada é propagação automática
- f(A) ⇨ f(B) ⇨ f(C) ⇨ ... ⇨ f(i)
- é frequente a situação em que f(i) é errada, pois uma ou mais propagações indevidas ocorreram no meio do caminho

Como lidar com isto?

- É bom ter uma dose de ceticismo em relação a qualquer anotação
- Procurar saber a metodologia que foi usada
- Uma anotação confiável precisa estar ancorada em dados experimentais
  - Estes são escassos

Registros curados no SwissProt

https://www.uniprot.org/



UniProtKB - P0A817 (METK_ECOLI)

Basket 1

Display

BLAST | Align | Format | Added to basket | History

Help video | Add a publication | Feedback

- Entry
- Publications
- Feature viewer
- Feature table

None

- Function
- Names & Taxonomy
- Subcellular location
- Pathology & Biotech
- PTM / Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar proteins
- Cross-references
- Entry information

Protein	**S-adenosylmethionine synthase**
Gene	**metK**
Organism	*Escherichia coli (strain K12)*
Status	Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level[i]

Function[i]

Catalyzes the formation of S-adenosylmethionine (AdoMet) from methionine and ATP. The overall synthetic reaction is composed of two sequential steps, AdoMet formation and the subsequent tripolyphosphate hydrolysis which occurs prior to release of AdoMet from the enzyme (PubMed:6251075, PubMed:7629147, PubMed:7629176, PubMed:9753435, PubMed:10551856, PubMed:10660564). Is essential for growth (PubMed:11952912). 7 Publications

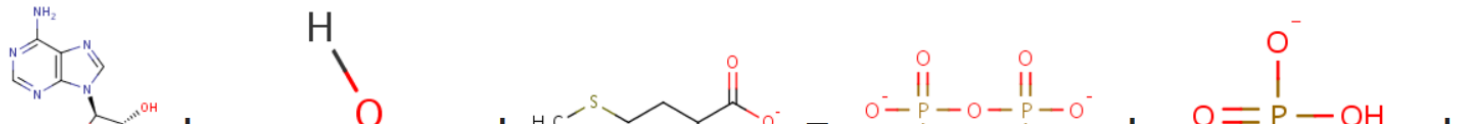
Caution

Was originally thought to differ from MetX, which was assigned as a second AdoMet synthase before being shown to be identical to MetK. 1 Publication

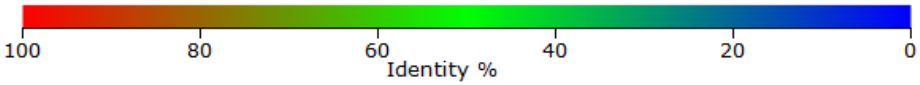
Catalytic activity[i]

- ATP + H_2O + L-methionine = diphosphate + phosphate + *S*-adenosyl-L-methionine UniRule annotation 6 Publications
  EC:2.5.1.6 UniRule annotation 6 Publications
  Source: Rhea. « Hide

ATP | H_2O | L-methionine | diphosphate | phosphate



Procurar estrelinha entre os alinhamentos



Edit and resubmit Order by: Score

Overview

Show all 50

Entry	Protein names	Match hit (50 100 150 200 250 300)	Identity
A0A0U1YJD2	DODA1 (Mirabilis jalapa)		100.0%
B6F0W8	4,5-DOPA dioxygenase extradiol (Mirabilis jalapa)		99.3%
A0A0G2UWA2	4,5-dioxygenase-like protein (Mirabilis multiflora)		94.5%
A0A0G2UYR6	4,5-dioxygenase-like protein (Mirabilis jalapa)		100.0%

Alignments

BLAST Align Download Add to basket Columns

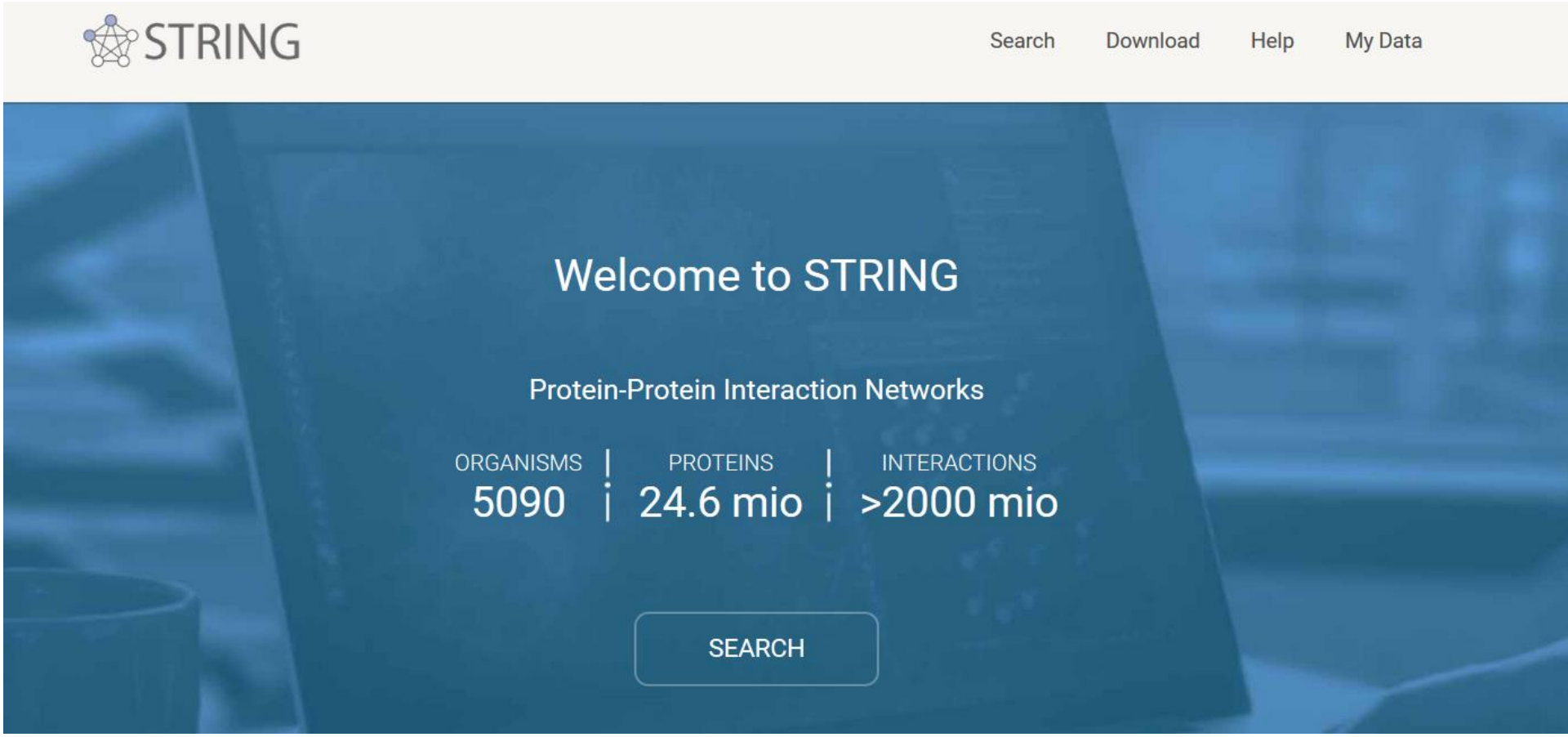
1 to 25 of 50 Show 25

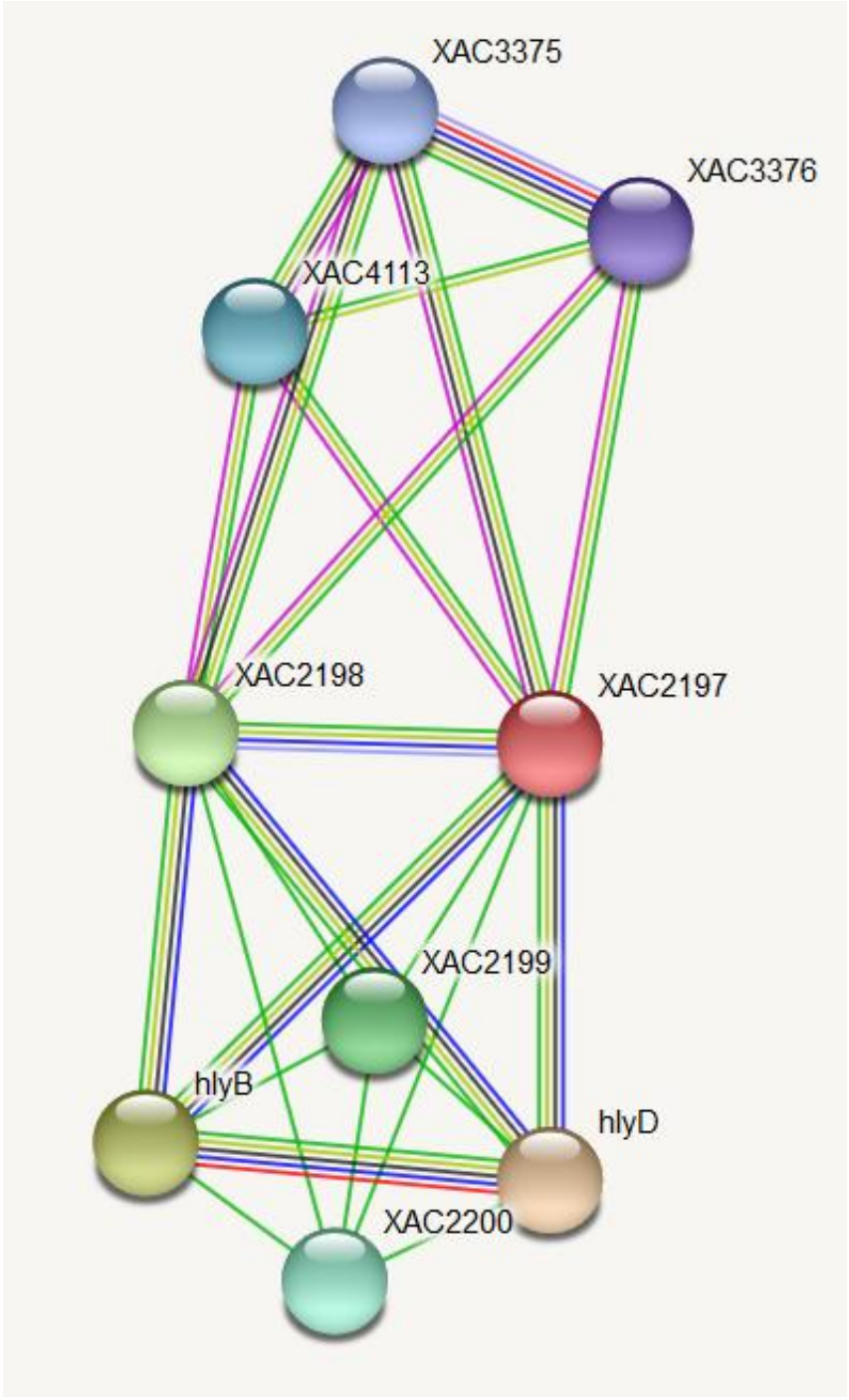
Entry	Alignment overview	Info	Status
Query: test B2016081714483A1C7ED25EE8374758DF3FD545FD25B74A9			
A0A0U1YJD2	A0A0U1YJD2_MIRJA - DODA1 - Mirabilis jalapa... - View alignment	E-value: 0.0 Score: 1,479 Ident.: 100.0%	
B6F0W8	DODA_MIRJA - 4,5-DOPA dioxygenase extradiol - Mirabilis jalapa... - View alignment	E-value: 0.0 Score: 1,468 Ident.: 99.3%	
A0A0G2UWA2	A0A0G2UWA2_9CARY - 4,5-dioxygenase-like protein Mirabilis multiflora - View alignment	E-value: 0.0 Score: 1,411 Ident.: 94.5%	

Interações entre proteínas

- métodos experimentais
  - mais confiáveis
- métodos computacionais
  - mais sujeitos a erros

Programa STRING https://string-db.org/




XAC3375
XAC3376
XAC4113
XAC2198
XAC2197
XAC2199
hlyB
hlyD
XAC2200

Nodes:

Network nodes represent proteins

splice isoforms or post-translational modifications are collapsed, i.e. each node represents all the proteins produced by a single, protein-coding gene locus.

Node Color

colored nodes:
query proteins and first shell of interactors

white nodes:
second shell of interactors

Node Content

empty nodes:
proteins of unknown 3D structure

filled nodes:
some 3D structure is known or predicted

Edges:

Edges represent protein-protein associations

associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding each other.

Known Interactions

from curated databases

experimentally determined

Predicted Interactions

gene neighborhood

gene fusions

gene co-occurrence

Others

textmining

co-expression

protein homology

Your Input:

		Neighborhood	Gene Fusion	Cooccurence	Coexpression	Experiments	Databases	Textmining	[Homology]	Score
XAC2197	*Hemolysin-type calcium binding protein; Identified by sequence similarity; putative; ORF located using Blastx/Glimmer /Genemark (2183 aa)*									

Predicted Functional Partners:

		Neighborhood	Gene Fusion	Cooccurence	Coexpression	Experiments	Databases	Textmining	[Homology]	Score
hlyD	*Hemolysin secretion protein D; Identified by sequence similarity; putative; ORF located using Blastx/Glimmer/Genemark (491 ...*	●		●	●			●		0.958
hlyB	*Hemolysin secretion protein B; Identified by sequence similarity; putative; ORF located using Blastx/Glimmer/Genemark (760 ...*	●		●	●			●		0.928
XAC2198	*Hemolysin-type calcium binding protein; Identified by sequence similarity; putative; ORF located using Blastx/Glimmer/Gene...*	●		●				●	●	0.921
XAC2199	*Uncharacterized protein; Putative; ORF located using Glimmer/Genemark (213 aa)*	●								0.845
XAC2200	*Uncharacterized protein; Putative; ORF located using Glimmer/Genemark (209 aa)*	●								0.627
XAC4113	*YapH protein; Identified by sequence similarity; putative; ORF located using Blastx/Glimmer/Genemark (2411 aa)*	●				●		●		0.510
XAC3375	*Uncharacterized protein; Identified by sequence similarity; putative; ORF located using Blastx/Glimmer/Genemark (601 aa)*	●			●	●		●		0.442
XAC3376	*Uncharacterized protein; Identified by sequence similarity; putative; ORF located using Blastx/Glimmer/Genemark (323 aa)*	●				●		●		0.415

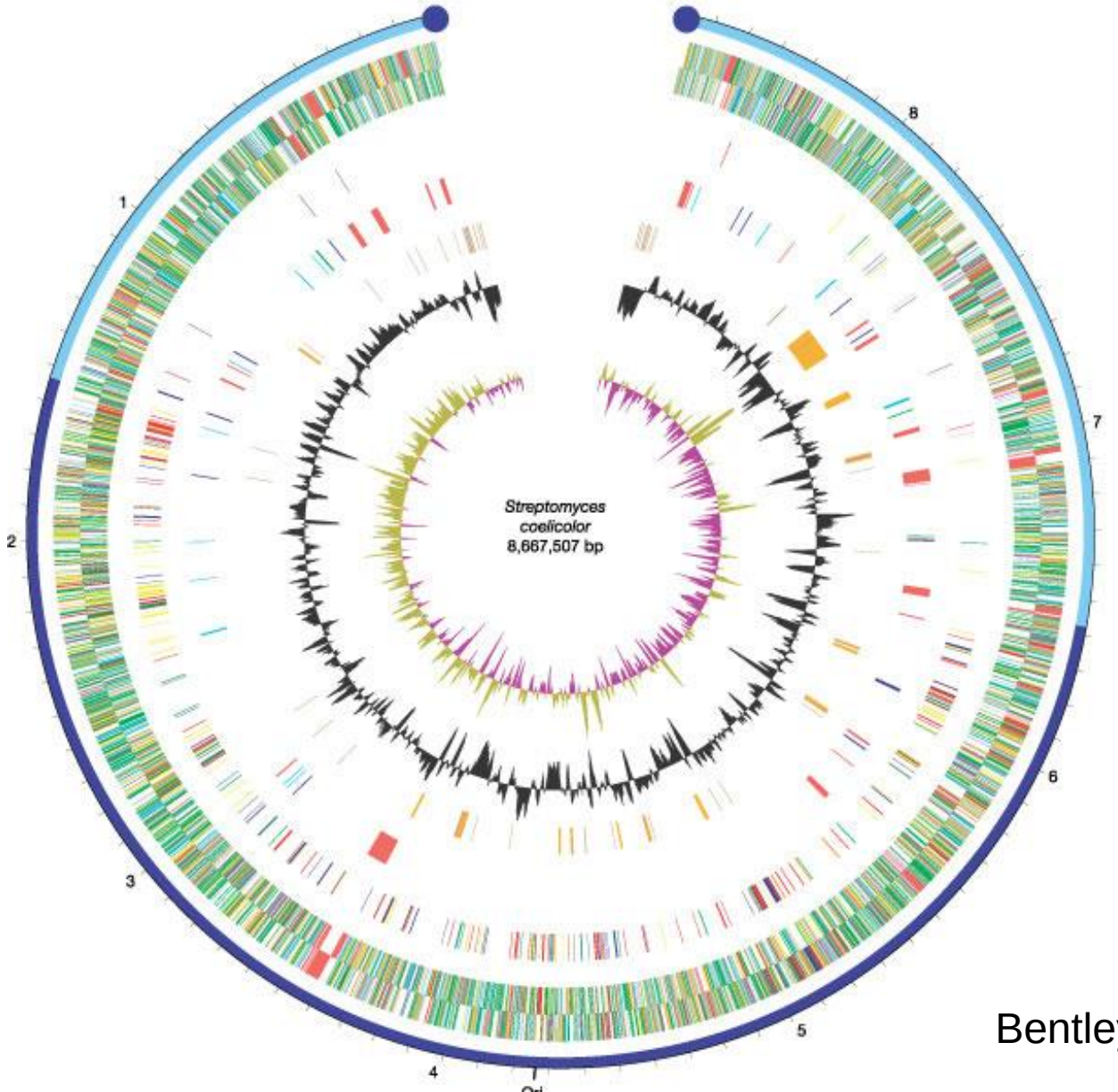
Your Current Organism:

Xanthomonas axonopodis citri

NCBI taxonomy Id: 190486

Other names: X. axonopodis pv. citri str. 306, Xanthomonas axonopodis citri, Xanthomonas axonopodis pv. citri str. 306, Xanthomonas citri pv. citri str. 306

Um produto final da anotação é um diagrama mostrando o genoma anotado



uma representação de um genoma (linear) com sua anotação

Bentley et al., *Nature*, 2002

Legenda das faixas, de fora para dentro (de Bentley et al., *Nature*, 2002)

- Circles 1 and 2 all genes (reverse and forward strand,respectively) colour-coded by function (black, energy metabolism; red, information transfer and secondary metabolism; dark green, surface associated; cyan, degradation of large molecules; magenta, degradation of small molecules; yellow, central or intermediary metabolism; pale blue, regulators; orange, conserved hypothetical; brown, pseudogenes; pale green, unknown; grey, miscellaneous)
- circle 3, selected 'essential' genes (for celldivision, DNA replication, transcription, translation and amino-acid biosynthesis, colour coding as for circles 1 and 2)
- circle 4, selected 'contingency' genes (red, secondary metabolism; pale blue, exoenzymes; dark blue, conservon; green, gas vesicle proteins)
- circle 5, mobile elements (brown, transposases; orange, putative laterally acquired genes)
- circle 6, G+C content
- circle 7, GC bias (G-C/G+C), khaki indicates values > 1, purple, < 1).
- The origin of replication (Ori) and terminal protein (blue circles) are also indicated.