Universidade de São Paulo
Instituto de Química

USP

Anotação de genomas

João C. Setubal

2017

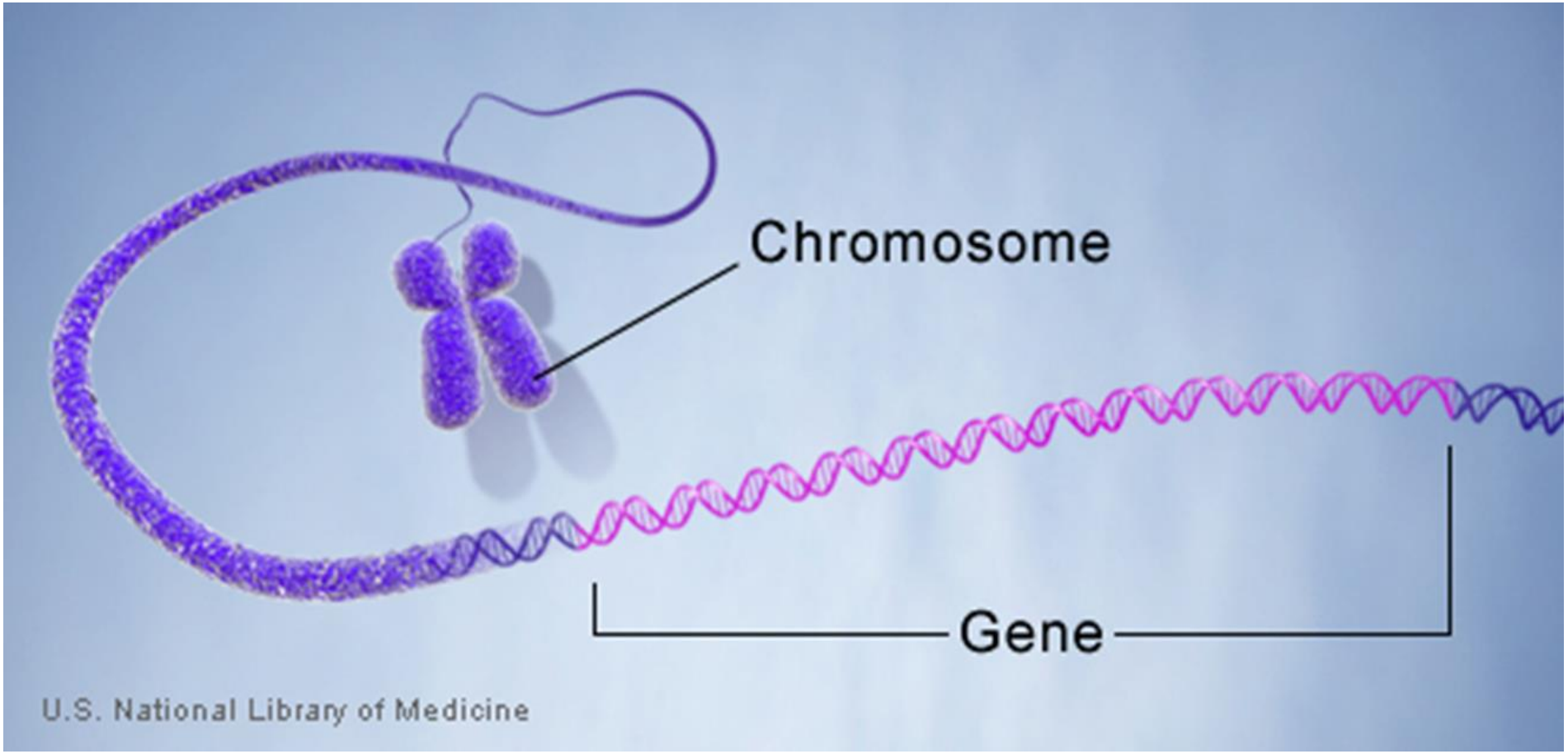
Sumário

- Dado um genoma completo, sem buracos ou erros
- Achar os genes codificadores de proteína
  - Sequência codificadora (CDS) (às vezes aparece ORF)
  - promotores
- Achar genes de RNA
  - RNA ribossomal
  - tRNA
  - Outros RNAs
- Atribuir função aos genes codificadores de proteína
- Esta aula: **genomas de procariotos**

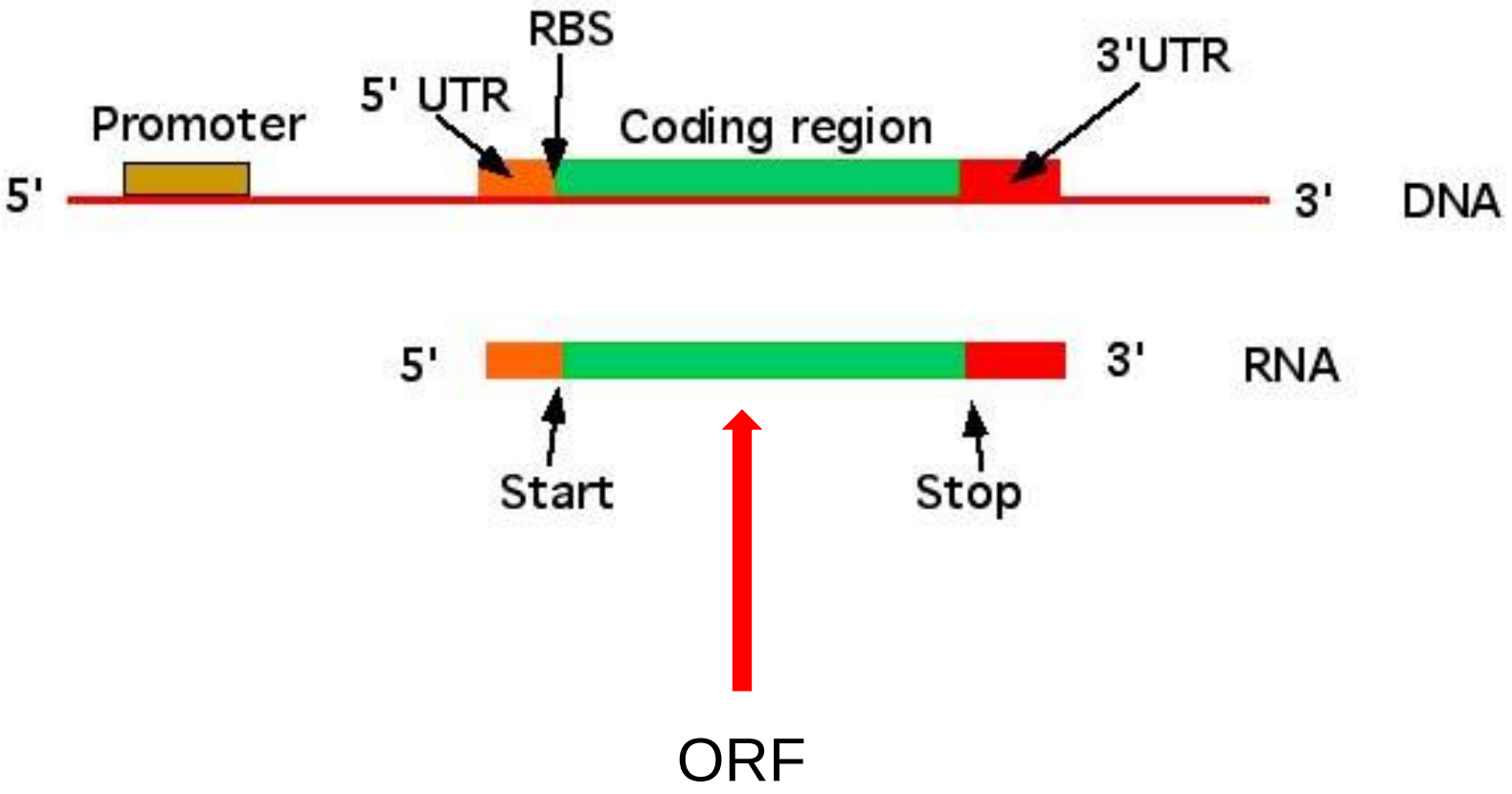
Achar genes codificadores de proteína

- *Gene finding*

Genes



Estrutura de um gene de procarioto



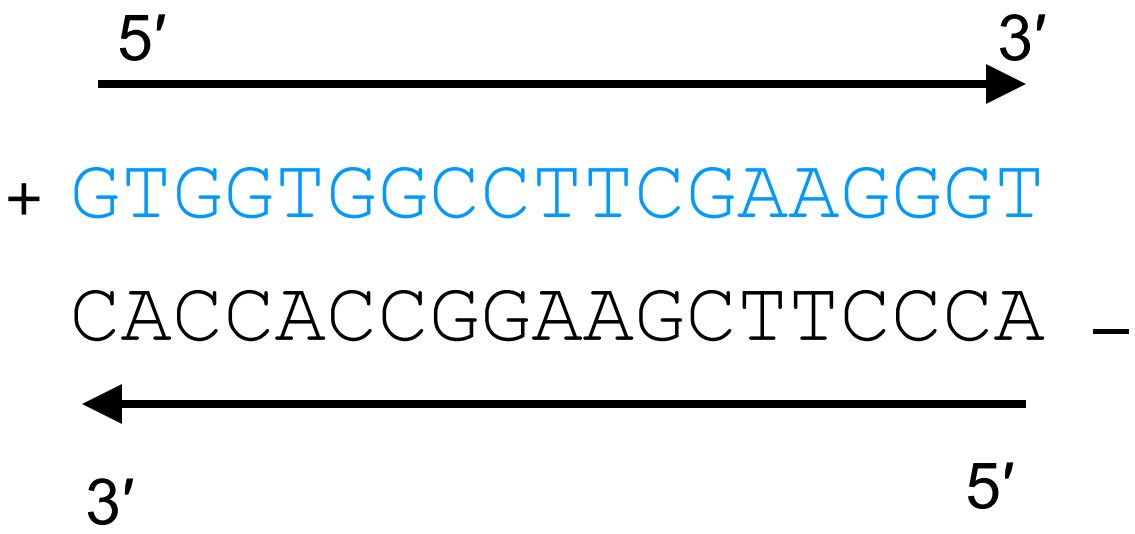
DNA tem *quadros de leitura*

+1: GTGGTGGCCTTCGAAGGGT

+2: TGGTGGCCTTCGAAGGGT

+3: GGTGGCCTTCGAAGGGT

DNA tem duas fitas (+ e –)



6 quadros no total

DNA de bactéria

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

Um gene (CDS)

...
AGCTCGCGCTCCGCATCCATCCAGTAGGGTTCGGTGTCGACGAGCGTGCC
GTCCATATCCCAGAAGACGGCGGCCGGCATCGCGTGCGGAGTCAGTTCGG
TCACGGCTGACAAGTCTATCCCGGCGGCCCCGGGCCTATTCTTGAGGGAC
GGCGTCCTGACCGGTCGCCGGATGAAAGGACCAGAACGCCCCGTGACTGA
CGCGAACAGCATCCTCGGAGGGCGCATCCTCGTGGTGGCCTTCGAAGGGT
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 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 presente

Método (rudimentar) para achar genes em procariotos

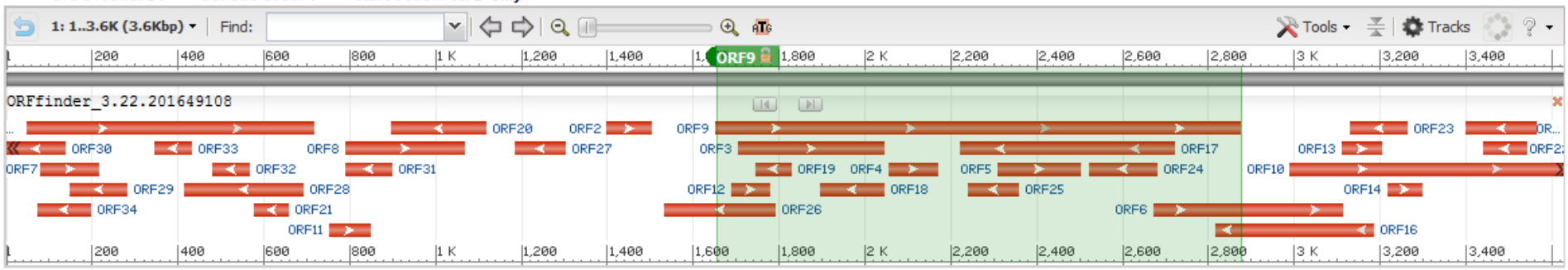
Ache todas as ORFs com pelo menos 900 bp

ORFfinder PubMed Search

Open Reading Frame Viewer

Sequence

ORFs found: 34 Genetic code: 1 Start codon: 'ATG' only



Add six-frame translation track

ORF9 (407 aa) Display ORF as... Mark

```
>lcl|ORF9
MLSPACPVIPGGHVCTVSCRSRILRTDRHAGLQPRAHRSM
HVVRLSIHRLRRFQTVELHPSSALNLLTGDNGAGKTSVLE
ALHLMAYGRSFRGRVRDGLIQQGANDLEVFVEWKEGGGAA
VERTRRAGLRHSGQEWTGRLDGEDVAQLGSLCAALAVVTF
EPGSHVLISGGGEPRRRFLDWGLFHVEPDFLTLWRRYARA
LKQRNALLKQGAQPRMLDAWDNELAESGETLTSRRMRYLE
RLQDRLVPVADAIAPALGLSALTFAPGWKRHEVSLADALL
LARERDRQNGYTSQGPHRADWMPSFHALPGKDALSRGQAK
LTALACLLAQAEDFAFERGEWPVIALDDLGSELDRHHQGR
VLQRLASAPAQVLITATETPPGLADAAALLQQFHVEHGQI
ARQATVN
```

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

Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF9	**+**	**2**	**1652**	**2875**	**1224 | 407**
ORF1	+	1	49	717	669 | 222
ORF10	+	2	2990	>3640	651 | 216
ORF17	-	1	2722	2222	501 | 166
ORF6	+	1	2671	3114	444 | 147
ORF16	-	1	3184	2816	369 | 122
ORF3	+	1	1705	2046	342 | 113
ORF8	+	2	791	1069	279 | 92
ORF28	-	2	693	415	279 | 92
ORF26		2	1701	1534	258 | 85

SmartBLAST ORF9

BLAST ORF9 BLAST marked set

BLAST Database:

UniProtKB/Swiss-Prot (swissprot)

Go back to the submitting page...

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
4. Resolva sobreposições

Na prática

- 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
- Técnica: modelos de markov de maior ordem

Programas mais usados

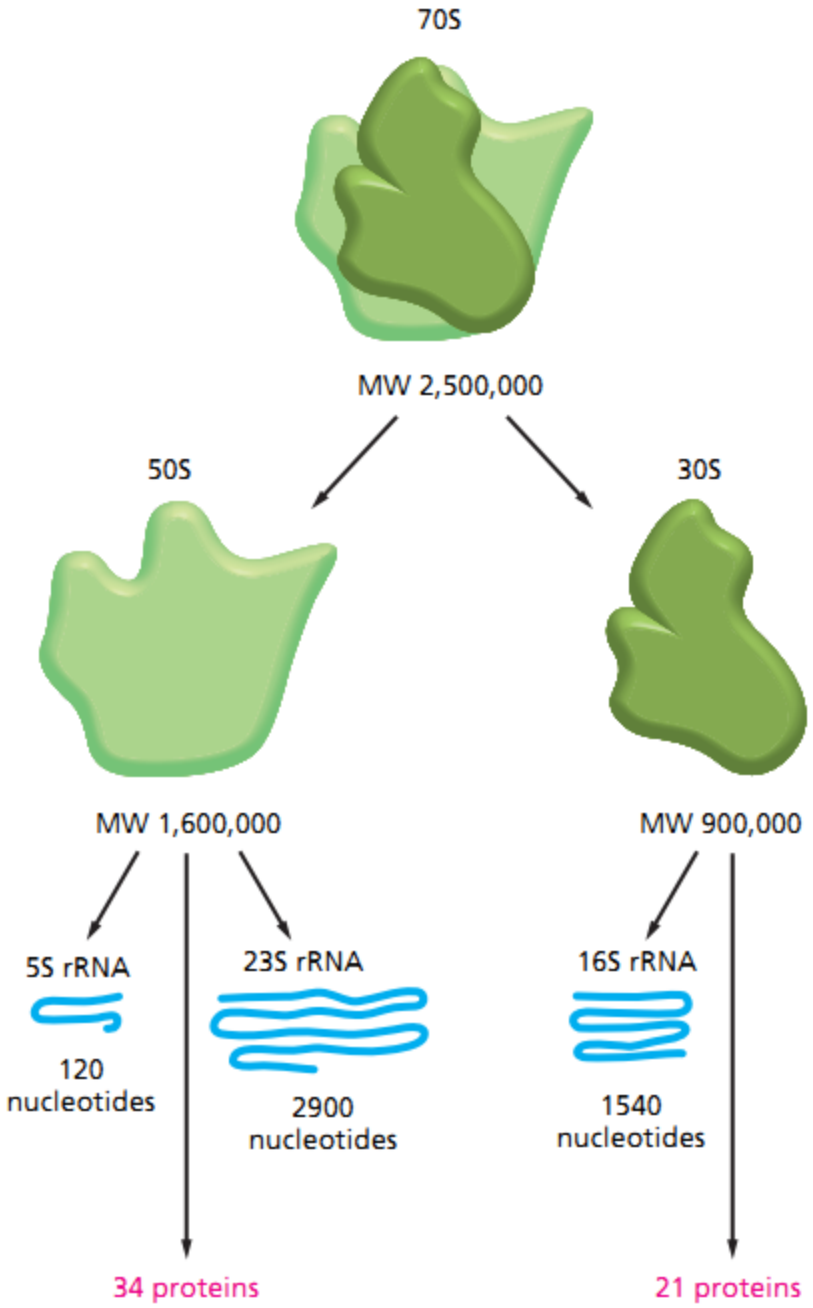
- Glimmer
  - http://ccb.jhu.edu/software/glimmer/index.shtml
- Prodigal
  - http://prodigal.ornl.gov/
- geneMark
  - http://exon.gatech.edu/
- Metagene (for metagenomics sequences)
  - http://weizhong-lab.ucsd.edu/metagenomic-analysis/server/metagene/

Limitações

- Genes pequenos (menores do que 150 bp) geralmente são perdidos
  - Se se aumenta a sensibilidade, vem muitos falsos positivos
- Início de tradução nem sempre é correto

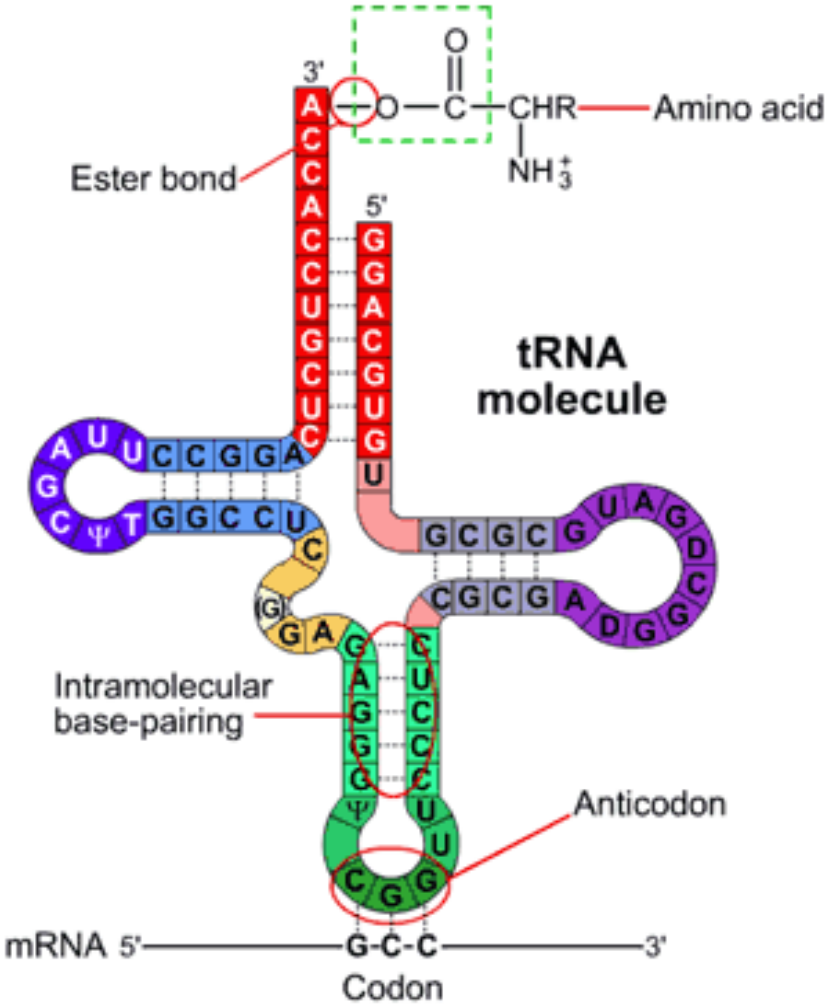
Achar genes de RNA

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



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
  - BLASTN, RNAmmer
  - Fronteiras exatas?
- tRNA
  - tRNAscan-SE
  - Aragorn
- Outros RNAs
  - RFAM

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**.

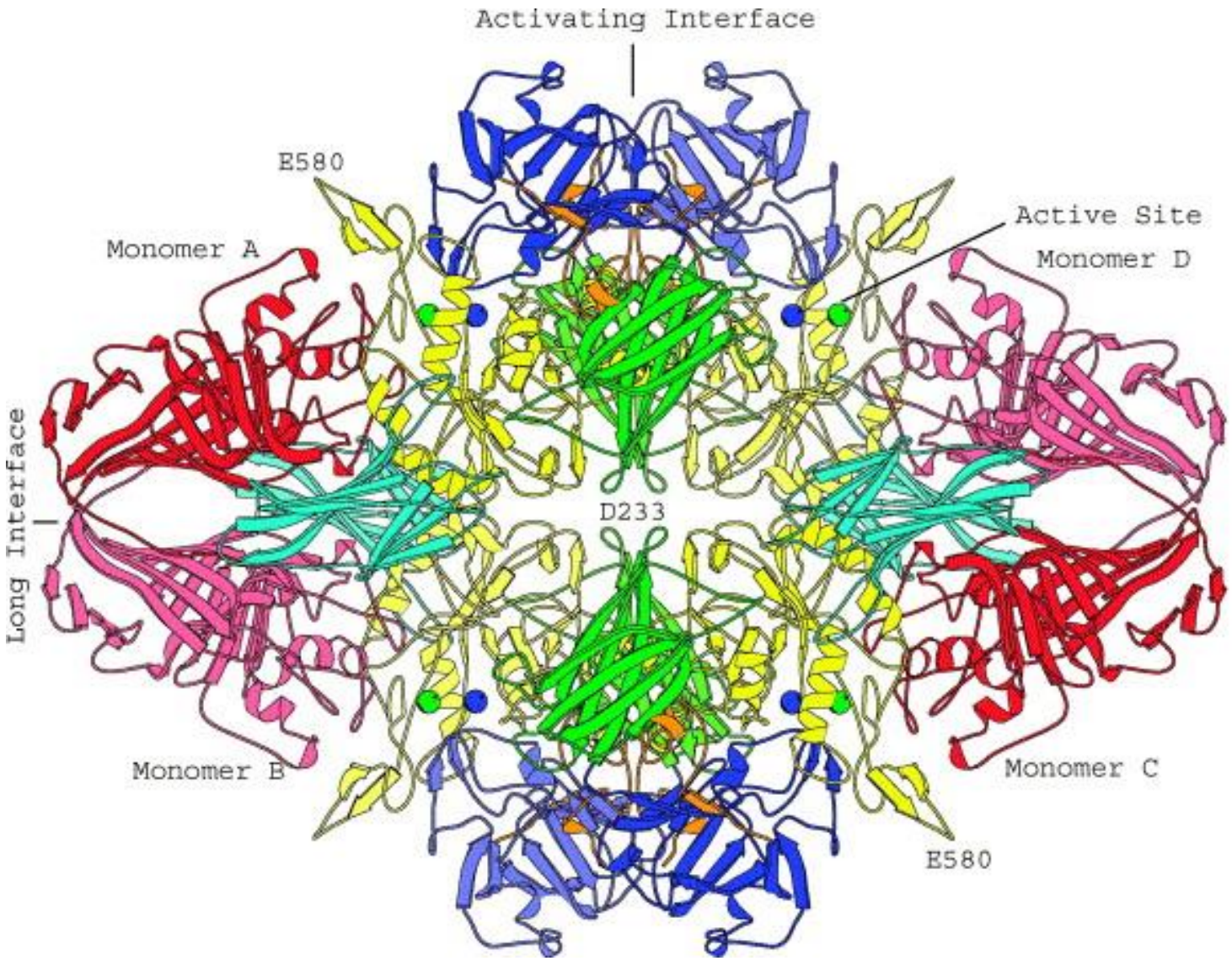
Further reference:

- **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.

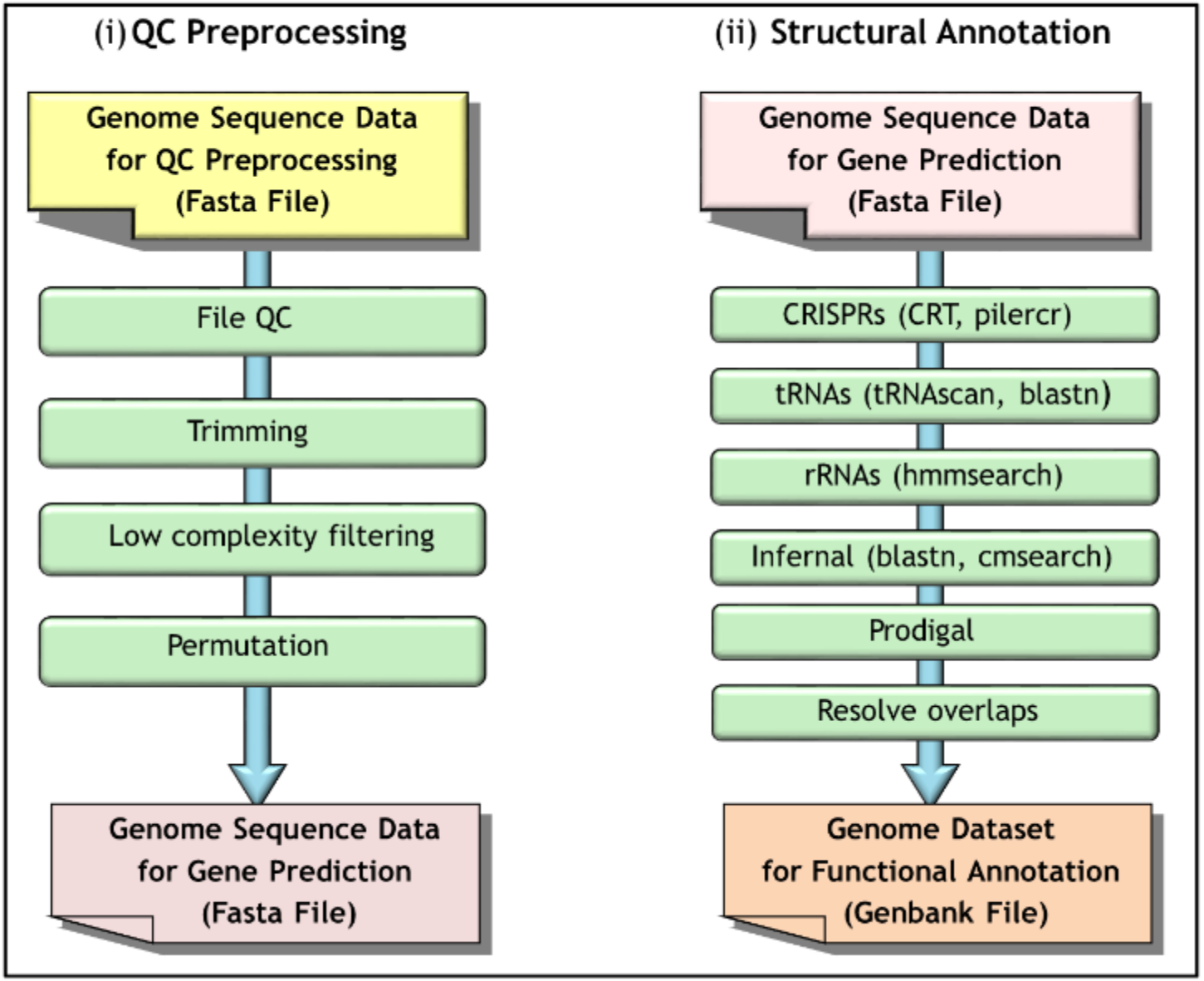


Figure 1. Genome sequence data preprocessing and structural annotation steps.

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].

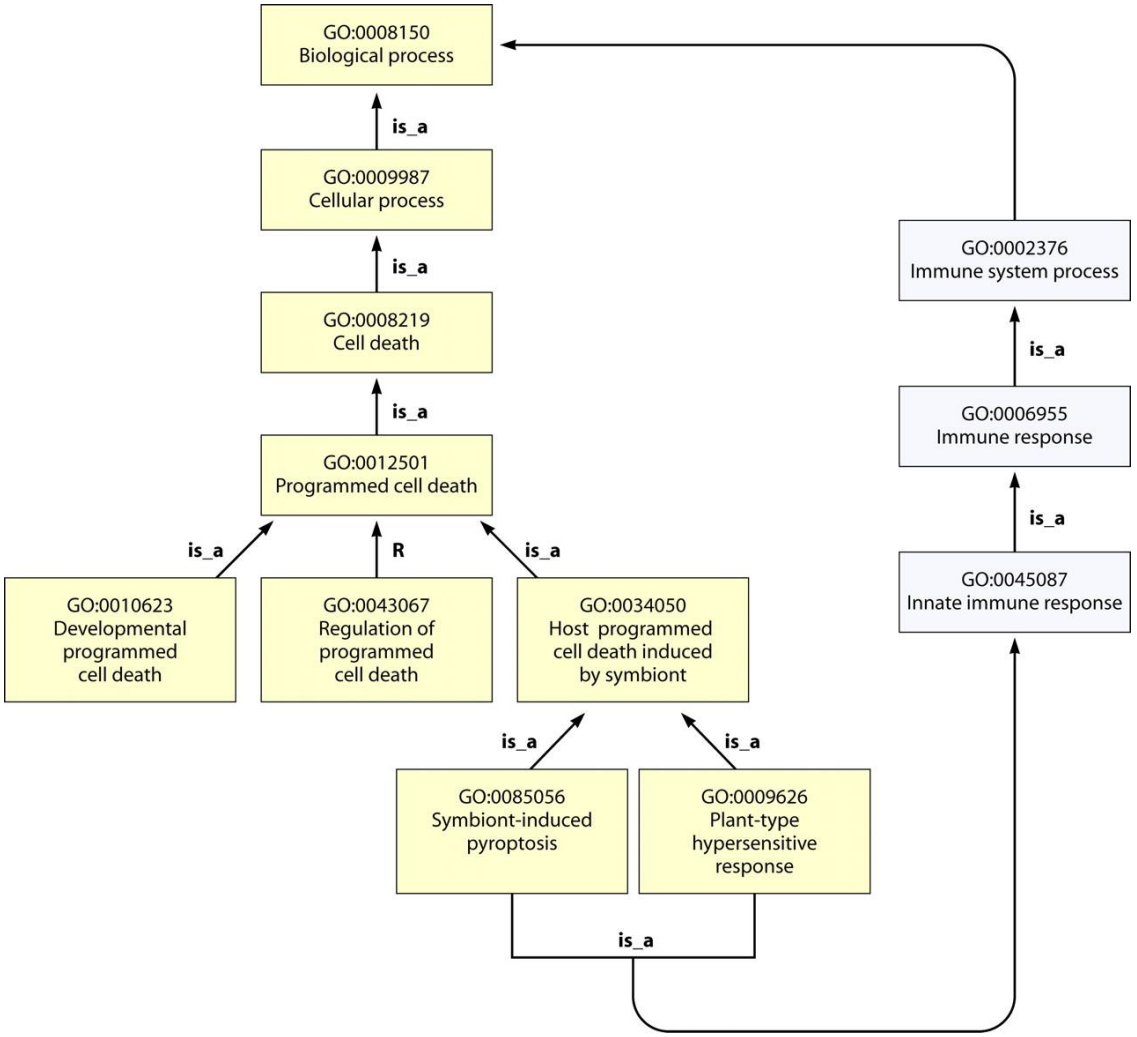
O problema dos termos

- Diferentes pessoas usam diferentes palavras para descrever a mesma função
- Diferentes pessoas usam as mesmas palavras para descrever funções diferentes
- É necessário uma padronização
  - Gene Ontology

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

Códigos de evidência

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

Use of an experimental evidence code in a GO annotation indicates that the cited paper displayed results from a physical characterization of a gene or gene product that has 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)

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)

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

- lacZ
- Ou mesmo…
- A frase curta que supostamente descreve a função dos genes
- Então alguns problemas babélicos continuam

Propagação de erros

- Tsunami de sequências
- => propagação automática de anotações
- Mas toda anotação precisa estar ancorada em dados experimentais
  - Estes são escassos
- Resultam muitos erros por propagação

Análise de enriquecimento

- Padronização de termos permite análise de enriquecimento
  - Exemplo típico é 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 d.e.?
  - Super-representação
  - Sub-representação

Resultado final