Advancing Science with DNA Sequence

Finding the genes in microbial genomes

Natalia Ivanova

MGM Workshop

January 7, 2008

Advancing Science with DNA Sequence

Finding the ~~genes~~ features in microbial genomes

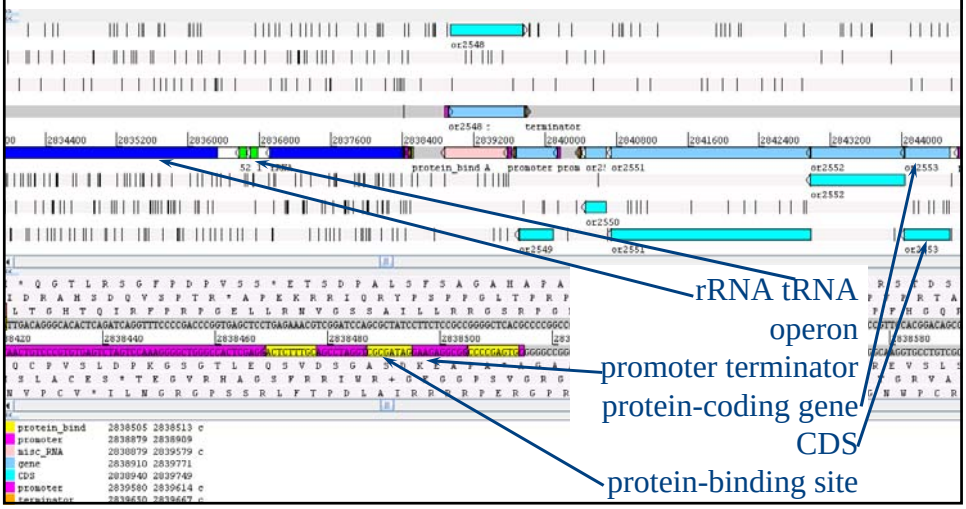
Sequence features in prokaryotic genomes:

- **stable RNA-coding genes (**rRNAs, tRNAs, RNA component of RNaseP, tmRNA**)**
- **protein-coding genes (**CDSs**)**
- **transcriptional features (**mRNAs, operons, promoters, terminators, protein-binding sites, DNA bends**)**
- **translational features (**RBS, regulatory antisense RNAs, mRNA secondary structures, translational recoding and programmed frameshifts, inteins**)**
- **pseudogenes (**tRNA and protein-coding genes**)**
- …

Advancing Science with DNA Sequence

Finding the genes in microbial genomes features

Well-annotated bacterial genome in Artemis genome viewer:



Advancing Science with DNA Sequence

Servers for microbial genome annotation

- **IMG-ER**

http://img.jgi.doe.gov/er

Output: stable RNA-encoding genes, CDSs, functional annotations

- **RAST**

http://rast.nmpdr.org/

Output: stable RNA-encoding genes, CDSs, functional annotations

- **REGANOR**

https://www.cebitec.uni-bielefeld.de/groups/brf/software/reganor/

Output: stable RNA-encoding genes, CDSs; file in gff format

- **RefSeq**

http://www.ncbi.nlm.nih.gov/genomes/MICROBES/genemark.cgi
http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi

Output: CDSs; file in tbl format

- **EasyGene**

http://www.cbs.dtu.dk/services/EasyGene/

Output: CDSs; sequence size restriction - <1 MB

Advancing Science with DNA Sequence

Finding stable RNAs - I

- Stable RNAs: large RNAs (16S and 23S rRNAs) and small RNAs (5S rRNA, tRNAs, tmRNA, RNase P component, riboswitches)
- For small RNAs statistical models can be generated and used to identify them in newly sequenced genomes
- Large RNAs are found by sequence similarity search (BLASTn) => there is no universally accepted tool; many errors in defining the boundaries
- search_for_rnas by Niels Larsen, rRNA database – used by all 3 servers predicting rRNAs

Genome	Sequencing center	16S rRNA, nt
Synechococcus sp. CC9311	UCSD, TIGR	1477
Synechococcus sp. CC9605	JGI	1440
Synechococcus elongatus PCC 7942	JGI	1490
Synechococcus sp. JA-2-3BA(2-13)	TIGR	1323
Synechococcus sp. JA-3-3Ab	TIGR	1324
Synechococcus sp. RCC307	Genoscope	1498
Synechococcus sp. WH7803	Genoscope	1497, 1464

Advancing Science with DNA Sequence

Finding stable RNAs - II

- Small RNAs (also called non-coding or ncRNAs) are found by search against Rfam covariance models using INFERNAL software suite and Rfam collection of models – see
  http://infernal.janelia.org/
  http://www.sanger.ac.uk/Software/Rfam/
  http://rfam.janelia.org/
- Both Rfam servers provide pre-calculated lists of short ncRNAs; Sanger center also provides web search facility for short DNA sequences

Other (less popular) tools:

- Pipeline for discovering cis-regulatory ncRNA motifs:

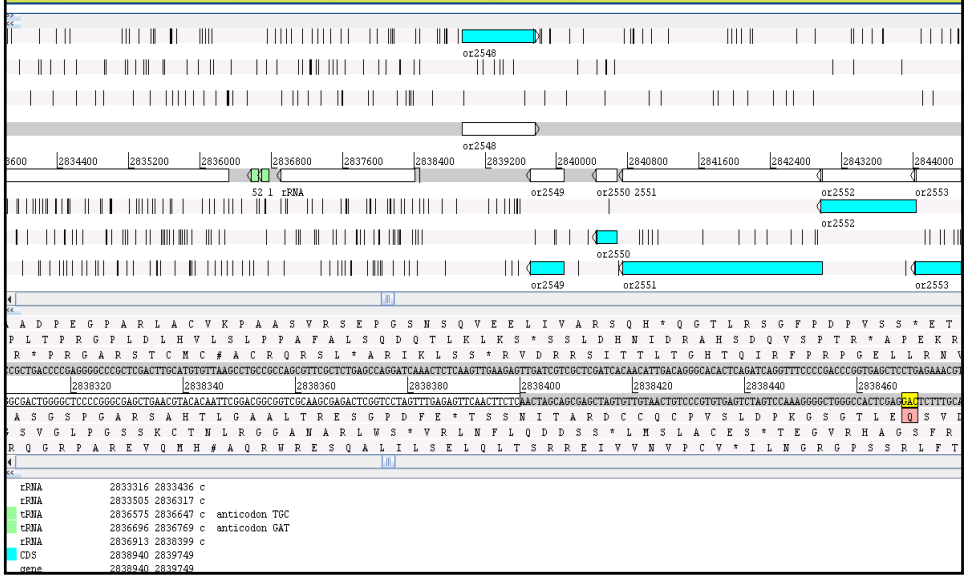
http://bio.cs.washington.edu/supplements/yzizhen/pipeline/

- RNAz

http://www.tbi.univie.ac.at/~wash/RNAz/

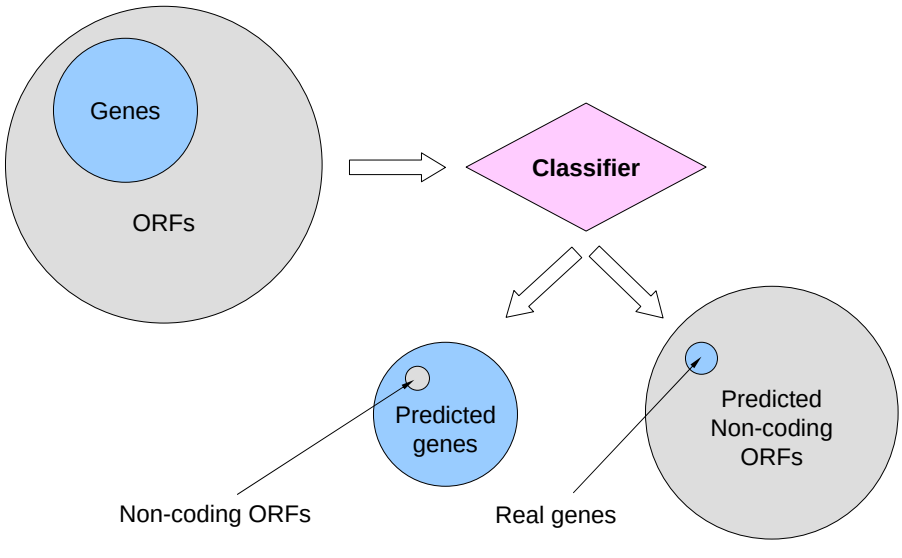
Advancing Science with DNA Sequence

Finding protein-coding genes



Advancing Science with DNA Sequence

Finding of protein-coding genes as classification problem



Advancing Science with DNA Sequence

Evidence-based vs ab initio algorithms

Two major approaches:

- **"*evidence-based*" (ORFs with translations homologous to the known proteins are CDSs)**

Advantages: finds the "unusual" genes (e. g. horizontally transferred); relatively low rate of false positive predictions

Limitations: cannot find "unique" genes; low sensitivity towards short genes; prone to propagation of false positive results of ab initio annotation tools

- ***ab initio* (ORFs with nucleotide composition similar to CDSs are also CDSs)**

Advantages: finds "unique" genes; high sensitivity

Limitations: often misses "unusual" genes; high rate of false positives

Advancing Science with DNA Sequence

Most popular CDS-finding tools

- CRITICA
- Glimmer family (Glimmer2, Glimmer3, RBS finder)
- GeneMark family (GeneMark-hmm, GeneMarkS)
- EasyGene

Combinations and variations of the above

- REGANOR (CRITICA + Glimmer3 + pre-processing)
- ORNL pipeline (CRITICA + Glimmer3)
- RAST (Glimmer2 + pre- and post-processing)

Advancing Science with DNA Sequence

Features and differences between gene finding tools

- Training set selection (evidence-based vs purely ab initio)
- Statistical model of coding and non-coding regions (codon frequencies, dicodon frequencies, hidden Markov models)
- Statistical model architecture (i. e. which parts of the CDS are explicitly modeled – may include RBS, spacer region, start codon, second codon, internal codons, stop codon, etc.)
- Additional algorithms for refinement of predictions (RBS finder, overlap resolution, estimation of statistical significance)

Advancing Science with DNA Sequence

Examples - I

- CRITICA and EasyGene use evidence-based training sets (BLASTn with counting synonymous/non-synonymous codons in CRITICA, BLASTx in EasyGene)
- Glimmer and GeneMark use ab initio training sets (Glimmer uses long non-overlapping ORFs, GeneMark uses heuristic model)
- Tools using ab initio training sets run much faster than tools using evidence-based training sets

Advancing Science with DNA Sequence

Examples – II

- CRITICA uses dicodon frequencies to model coding regions
- Glimmer uses interpolated Markov models (IMM) of up to 5-th order; GeneMark uses order 2 hmm for coding regions, order 0 hmm for non-coding regions; EasyGene uses order 4 hmm for coding regions, order 0 hmm for non-coding regions
- CRITICA is the least sensitive
- Order of the Markov model will determine the minimal size of the training set => application to metagenomes

Markov model order	0	1	2	3	4	5
Minimal size of the training set (kb)	1.6	6.4	25.6	102.4	409.6	1638.4

Advancing Science with DNA Sequence

Different gene prediction tools applied to the same genome

	total features	total CDSs	non-pseudo CDSs	pseudo	total rRNA	total tRNA	total misc RNA
manual	7124	7042	6699	343	18	63	1
GeneMark	7059	6974	6974	0	18	62	2
ORNL	7076	6994	6994	0	18	63	1
RAST	5503	5422	5422	0	18	63	0
REGANOR	6420	6339	6339	0	18	63	0
Glimmer3	8218	8218	8218	0	0	0	0

	missed by automated annotation		false positive (deleted by manual curation)		too short (extended by manual curation)		too long (truncated by manual curation)		total modifications	
	#	% CDSs	#	% CDSs	#	% CDSs	#	% CDSs	#	% CDSs
GeneMark	347	4.9	282	4.0	846	12.0	106	1.5	1581	22.4
ORNL	610	8.6	560	7.9	300	4.2	1152	16.3	2622	37.2
RAST	1783	25.3	167	2.3	83	1.1	2228	31.6	4261	60.5
REGANOR	904	12.8	203	2.8	207	2.9	1059	15.0	2373	33.6
Glimmer3	237	3.3	1408	19.9	500	7.1	542	7.6	2687	38.1

JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

Conclusions

- There are several tools for automated annotation of microbial genomes
- These tools identify a limited range of features and development of tools for identification of operons, promoters, terminators etc. is highly desirable
- But this development requires significant experimental input
- Different automated gene finders have different advantages and limitations; the best strategy is using any of them or a combination followed by evidence-based manual curation
- => talk on Wednesday by Thanos Lykidis