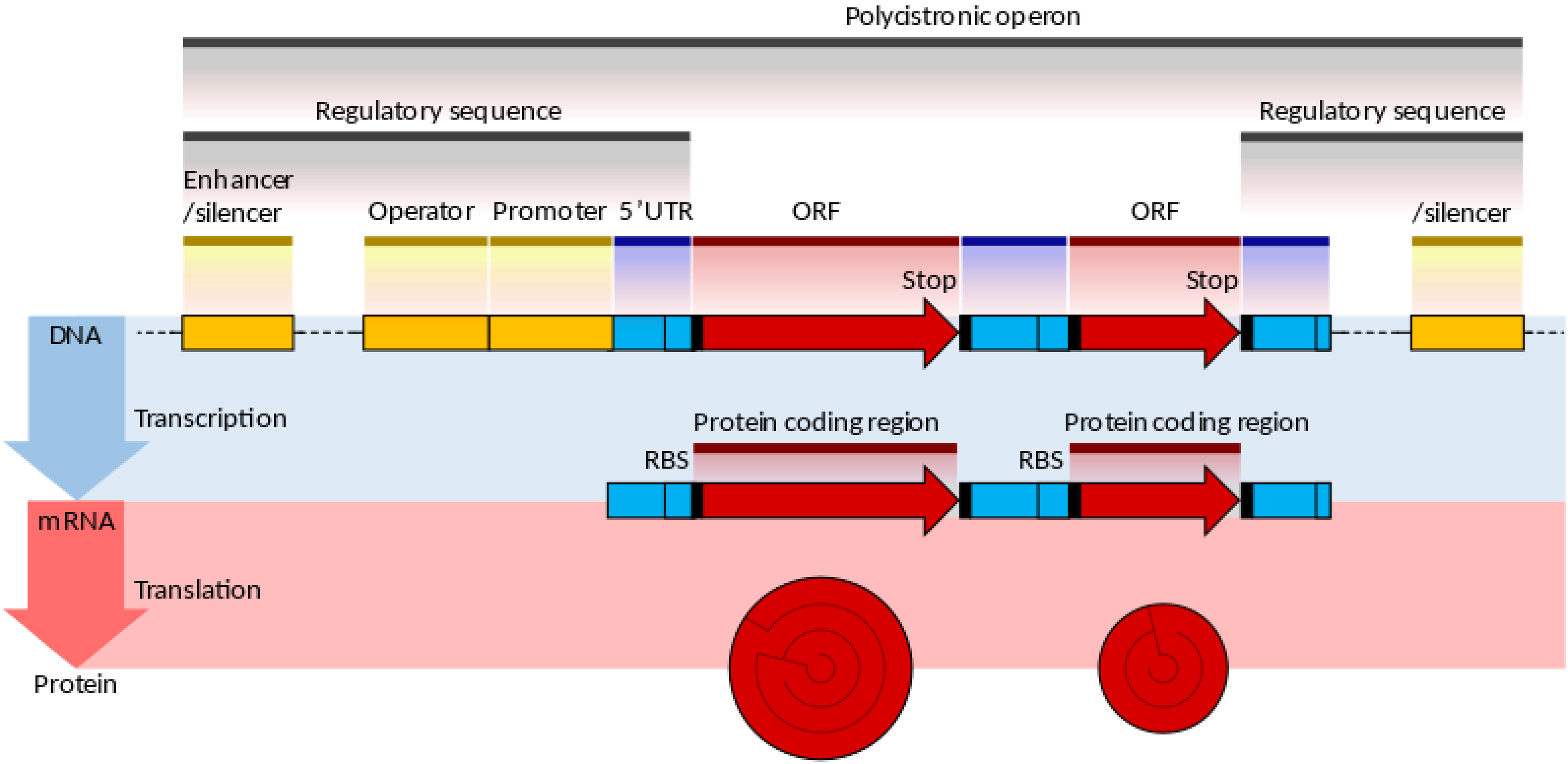
Team 1: Gene Prediction Background & Strategy

Aaron Pfennig, Priya Narayanan, Jessica Mulligan, Hira Anis, Winnie Zheng, Maria Ahmad

Prokaryotic gene structure



[1] Prokaryotic Gene Structure.

Gene Prediction Introduction

Gene prediction or gene finding is a process or identifying the regions of genomic DNA that encodes genes

Two classes of genes:

- Coding genes → proteins
- Non-coding genes → tRNAs, rRNAs etc.

It adopts two classes of methods:

- Similarity based (homology) searches
- *ab initio* prediction
    - Markov & Hidden Markov Model

Plans to assess the performance of the tools

It is possible to compute sensitivity, positive predictive value and specificity (only for start site) predictions based on annotations

$Sn = TP / (TP + FN)$

$PPV = TP / (TP + FP)$

$Sp = TN / (TN + FP)$ (only start site prediction)

-Sn=Sensitivity

-Sp=Specificity

-TP=True Positive ; TN=True Negative

-FP= False Positive ; FN=False Negative

-PPV= Positive Predictive Value

[2] Wang et al (2004).

Ab-initio approaches (CDS prediction)

Aims at predicting protein coding genes in a given genome based on certain features:

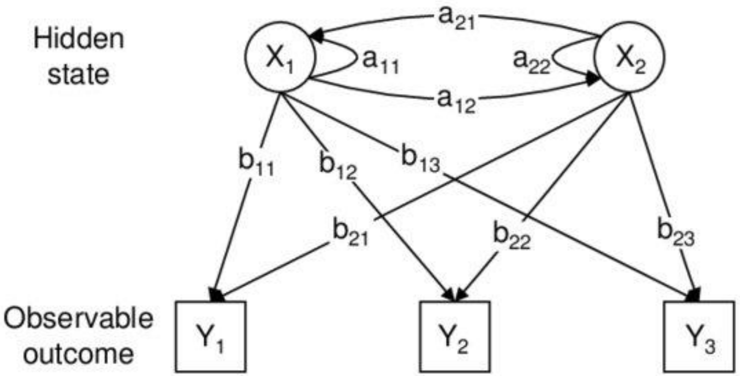
- ORFs
- GC content → codon usage bias
- regulatory motifs (SD, RBS etc.)

Highly popular methods rely on HMMs:

- GeneMarkS2
- Glimmer3

Another approach uses dynamic programming:

- Prodigal



[3] Toledo et. al. (2009)

Ab-initio approaches (CDS prediction)

GeneMarkS2:

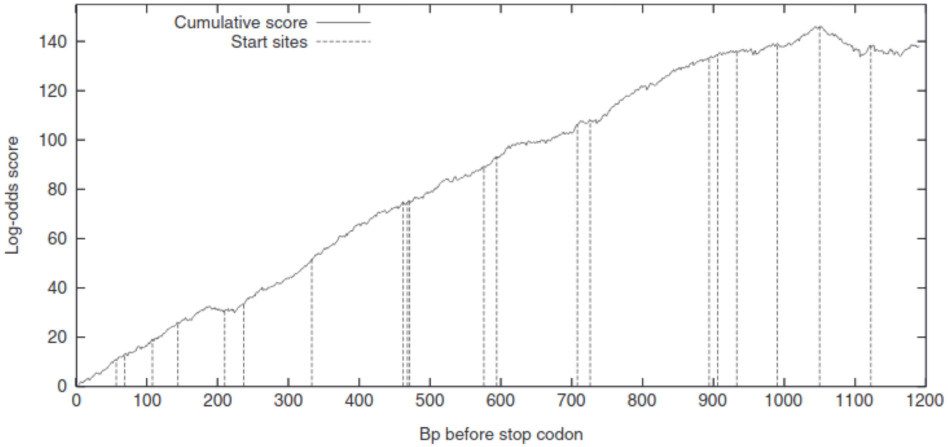
- Self-training algorithm based on a HMM
- Models transcription domain to predict gene start more accurately
- incl. heuristic model designed to predict horizontally transferred genes

Glimmer3:

- Interpolated Markov Model
- Reverse scoring → scoring relies on k-mer within coding region
- Trains on long ORFs

Prodigal:

- Identifies all ORFs and scores them using a dynamic programming approach
- Refines predictions after training on subset of ORFs



[4] Lomsadze et. al. 2017, [6] Delcher et .al. 2007, [7] Hyatt et. al. (2010)

Ab-initio approaches (CDS prediction)

GeneMarkS2:

- Highest sensitivity and specificity
- Works on diff. gene regulatory motifs
  - Leadered (Shine-Dalgarno +/-)
  - Leaderless

Glimmer3:

- Under performs by most metrics
- Predicts the least short genes (<150 nt)

Prodigal:

- Trained on *E.coli*
- Predicts the most gene starts correctly in *E.coli*

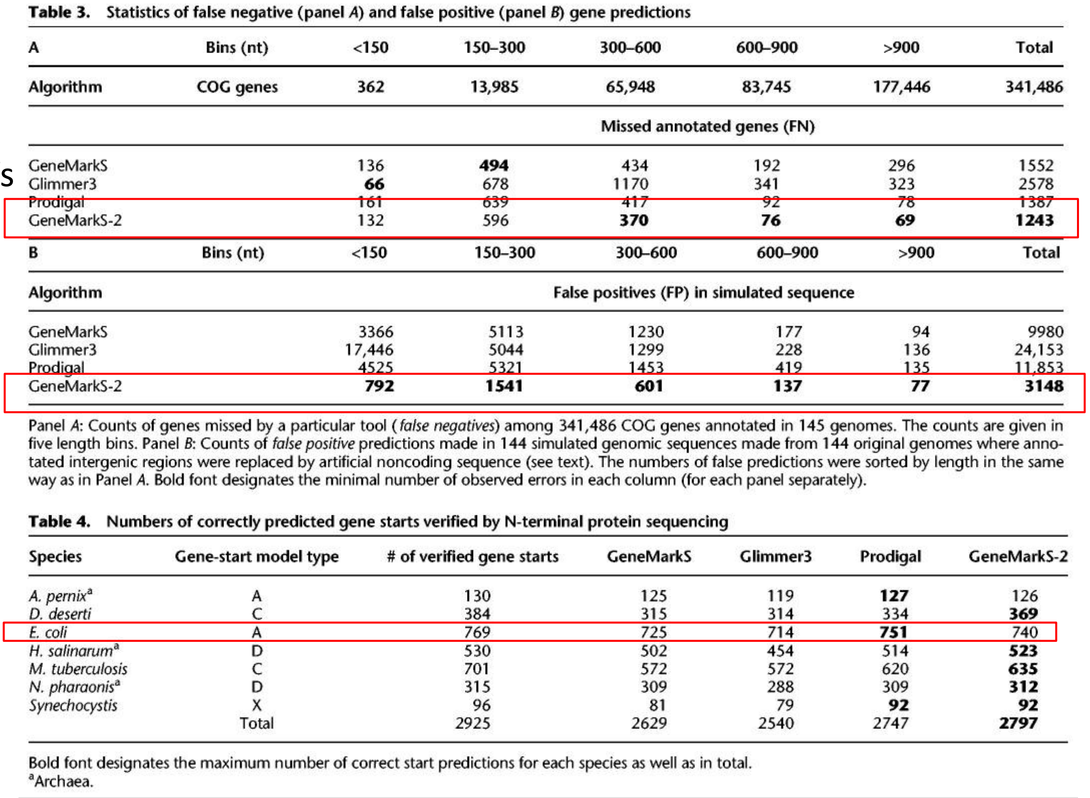


Table 3. Statistics of false negative (panel *A*) and false positive (panel *B*) gene predictions

A	Bins (nt)	<150	150–300	300–600	600–900	>900	Total
Algorithm	COG genes	362	13,985	65,948	83,745	177,446	341,486
				Missed annotated genes (FN)			
GeneMarkS		136	**494**	434	192	296	1552
Glimmer3		**66**	678	1170	341	323	2578
Prodigal		161	639	417	92	78	1387
GeneMarkS-2		132	596	**370**	**76**	**69**	**1243**

B	Bins (nt)	<150	150–300	300–600	600–900	>900	Total
Algorithm				False positives (FP) in simulated sequence			
GeneMarkS		3366	5113	1230	177	94	9980
Glimmer3		17,446	5044	1299	228	136	24,153
Prodigal		4525	5321	1453	419	135	11,853
GeneMarkS-2		**792**	**1541**	**601**	**137**	**77**	**3148**

Panel *A*: Counts of genes missed by a particular tool (*false negatives*) among 341,486 COG genes annotated in 145 genomes. The counts are given in five length bins. Panel *B*: Counts of *false positive* predictions made in 144 simulated genomic sequences made from 144 original genomes where annotated intergenic regions were replaced by artificial noncoding sequence (see text). The numbers of false predictions were sorted by length in the same way as in Panel *A*. Bold font designates the minimal number of observed errors in each column (for each panel separately).

Table 4. Numbers of correctly predicted gene starts verified by N-terminal protein sequencing

Species	Gene-start model type	# of verified gene starts	GeneMarkS	Glimmer3	Prodigal	GeneMarkS-2
A. pernix[a]	A	130	125	119	**127**	126
D. deserti	C	384	315	314	334	**369**
E. coli	A	769	725	714	**751**	740
H. salinarum[a]	D	530	502	454	514	**523**
M. tuberculosis	C	701	572	572	620	**635**
N. pharaonis[a]	D	315	309	288	309	**312**
Synechocystis	X	96	81	79	**92**	**92**
	Total	2925	2629	2540	2747	**2797**

Bold font designates the maximum number of correct start predictions for each species as well as in total.
[a]Archaea.

[4,5] Lomsadze et al. (2017)

Homology based approaches

Compare sequence with known genes

Relies on extrinsic information (eg. known expressed sequence tags, messenger RNA, protein products, and homologous or orthologous sequences)

Important to consider *horizontal gene transfer* in prokaryotes, and to find tools which consider this

[8] Gene Prediction. (2019).

Homology based approaches

- BLASTX
  - Translated nucleotide → protein
  - Produces more reliable and accurate results than BLASTn & BLASTp when dealing with coding DNA
- DIAMOND
  - Protein alignment algorithm which uses double indexing
  - Intended for replacing BLASTx in high-throughput setting
  - Aligns short sequence reads 20k X faster than BLASTx, with similar level of sensitivity
- HMMER
  - HMM
  - Designed to detect remote homologs as sensitively as possible → horizontal gene transfer
  - As fast as BLAST

[9] Altschul et. al. (1990), [10] Buchfink et. al. (2015),
[11] HMMER.org

Non coding gene prediction (tRNA)

tRNAs play an important role in cellular transcription by being able to recognize codons within mRNA and attaching the corresponding amino acids to amino acid chains

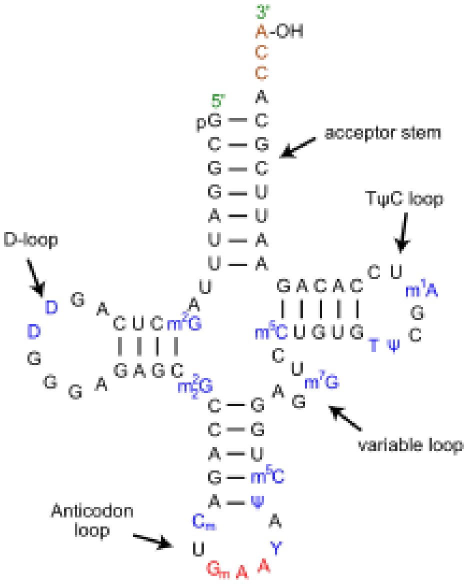
They also have regulatory and synthesis functions outside of translation

tRNAscan-SE

- Uses covariance model
- Must exceed similarity threshold + potential to form T-loop

Aragorn

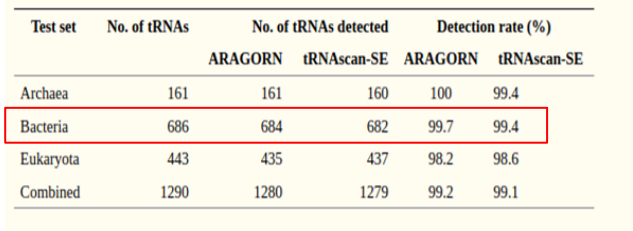
- Predicts tRNA and tmRNA genes
- Attempts to find subset of the B box (GTTC)
- Expands around hits in order to find characteristic motifs



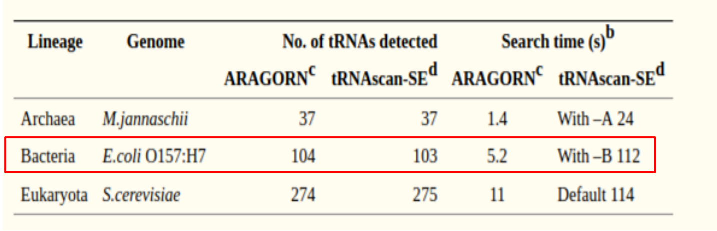
[12] Lowe et. al. (1997), [13] Laslett et. al. (2004),
[14] Transfer RNA (2020)

Non coding gene prediction (tRNA)

- ARAGORN vs tRNAscan-SE
  - ARAGORN is much faster and as sensitive



Test set	No. of tRNAs	No. of tRNAs detected		Detection rate (%)	
		ARAGORN	tRNAscan-SE	ARAGORN	tRNAscan-SE
Archaea	161	161	160	100	99.4
Bacteria	686	684	682	99.7	99.4
Eukaryota	443	435	437	98.2	98.6
Combined	1290	1280	1279	99.2	99.1



Lineage	Genome	No. of tRNAs detected		Search time (s)[b]	
		ARAGORN[c]	tRNAscan-SE[d]	ARAGORN[c]	tRNAscan-SE[d]
Archaea	*M.jannaschii*	37	37	1.4	With –A 24
Bacteria	*E.coli* O157:H7	104	103	5.2	With –B 112
Eukaryota	*S.cerevisiae*	274	275	11	Default 114

[13] Laslett et al. (2004)

Non coding gene prediction (rRNA)

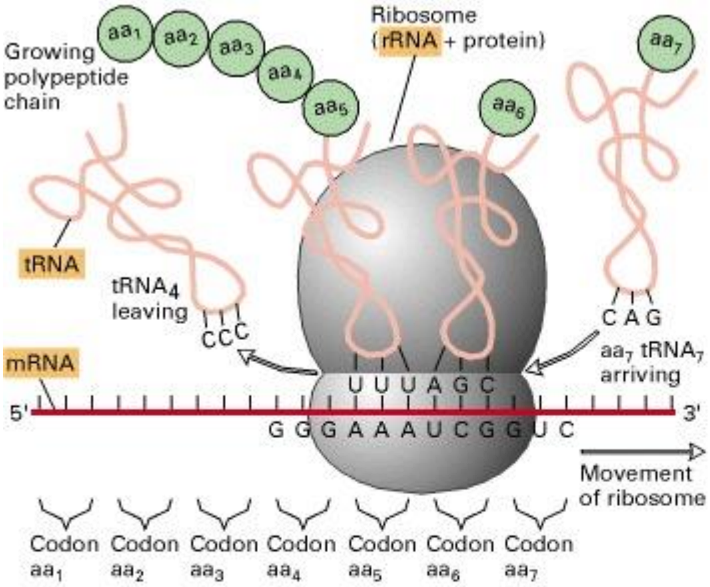
rRNAs are highly conserved due to their role in protein synthesis

RNAmmer

- Predicts rRNA genes
- HMM from structural alignment
- Allows variation in rRNA genes

barrnap

- Uses a HMM for each rRNA gene
- Built from full length seed alignments



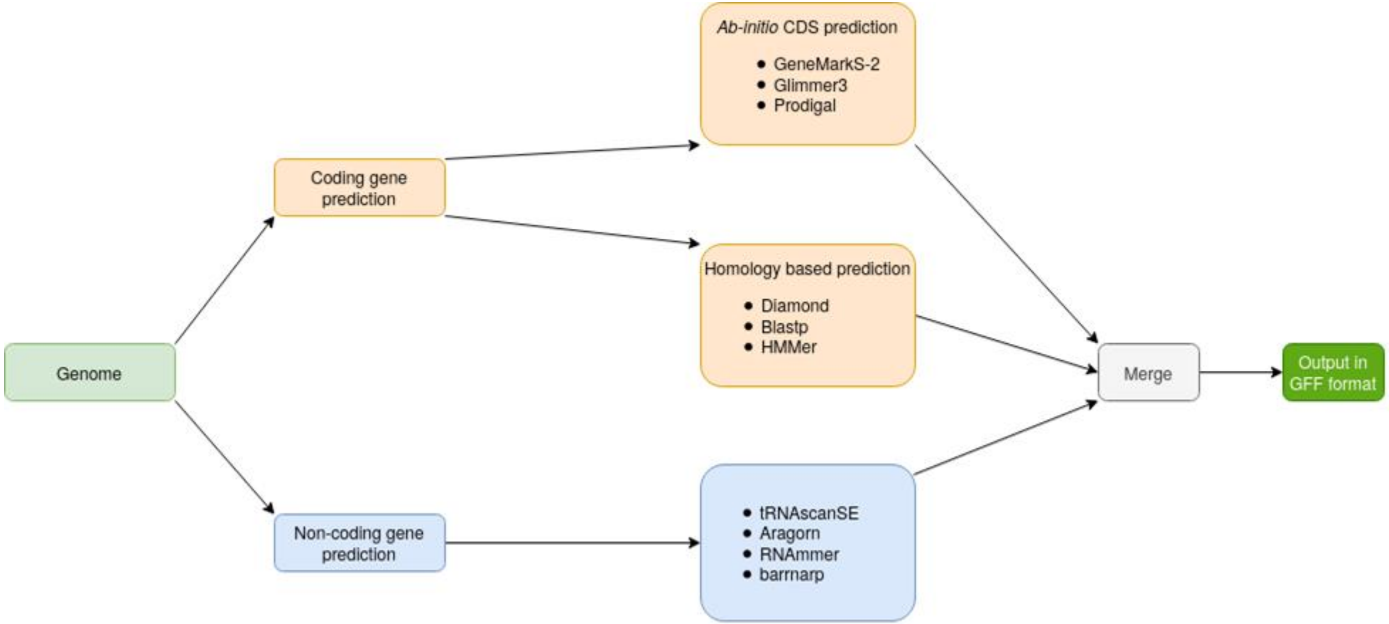
[15] Lagasen et. al. (2007), [16] Seemann (2018),
[17] Kate (2017)

Non coding gene prediction (rRNA)

- RNAmmer vs barrnap
    - RNAmmer is **more sophisticated** and **accurate**
        - Uses HMMer 2.x in ‘glocal’ alignment mode
        - barrnap uses nHMMer in local alignment mode
    - barrnap is available without license

Lagasen et. al. (2007).
Torsten Seemann. (2018)

Workflow



Genome

Coding gene prediction

Non-coding gene prediction

Ab-initio CDS prediction

- GeneMarkS-2
- Glimmer3
- Prodigal

Homology based prediction

- Diamond
- Blastp
- HMMer

- tRNAscanSE
- Aragorn
- RNAmmer
- barrnarp

Merge

Output in GFF format

Reference

[1] Prokaryotic Gene Structure. *Wikipedia*. https://en.wikipedia.org/wiki/Template:Prokaryote_gene_structure

[2] Wang, Z., Chen, Y., & Li, Y. (2004). A brief review of computational gene prediction methods. *Genomics, proteomics & bioinformatics*, *2*(4), 216–221. doi:10.1016/s1672-0229(04)02028-5

[3] Toledo, Tomer & Katz, Romina. (2009). State Dependence in Lane-Changing Models. Transportation Research Record. 2124. 81-88. 10.3141/2124-08.

[4] Lomsadze Alexandre. et al. (2017) Improved prokaryotic gene prediction yields insights into transcription and translation mechanisms on whole genome scale. 1–24. doi: 10.1101/193490.

[5] Lomsadze, Alexandre et al. (2017) “GeneMarkS-2 : Raising Standards of Accuracy in Gene Recognition.” .

[6] A.L. Delcher, K.A. Bratke, E.C. Powers, and S.L. Salzberg. Identifying bacterial genes and endosymbiont DNA with Glimmer, Bioinformatics 23:6 (2007), 673-679.

[7] Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, *11*, 119. doi:10.1186/1471-2105-11-119

[8] Gene Prediction. *Wikipedia*. 2019. https://en.wikipedia.org/wiki/Gene_prediction

[9] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. PubMed

[10] Benjamin Buchfink, Chao Xie & Daniel H. Huson, Fast and Sensitive Protein Alignment using DIAMOND, Nature Methods, 12, 59–60 (2015) doi:10.1038/nmeth.3176.

[11] HMMER: biosequence analysis using profile hidden Markov models. http://hmmer.org/

[12] Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research*, *25*(5), 955–964. doi:10.1093/nar/25.5.955

[13] Laslett, D., & Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic acids research*, *32*(1), 11–16. doi:10.1093/nar/gkh152

[14] Transfer RNA. *Wikipedia.* (2020).

[15] Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research*, *35*(9), 3100–3108. doi:10.1093/nar/gkm160

[16] Torsten Seemann. (2018). BAsic Rapid Ribosomal RNA Predictor. https://github.com/tseemann/barrnap

[17] Kate. M. (2017). How do mRNA, tRNA work together in translation to build protein? *Socratic Q&A.* https://socratic.org/questions/how-do-mrna-trna-and-rrna-work-together-in-translation-to-build-protein