Gene Prediction: Background & Strategy

Team 2:
Danielle Temples, Kara Lee
Paarth Parekh, Shuting Lin

Thursday, February 13th
Computational Genomics
Spring 2020

Outline

- Project purpose & overview
- Our pathogenic organism
- Gene Prediction
  - Homology-based tools & comparisons
  - Ab initio tools & comparisons
  - Non-coding RNA prediction tools & comparisons
- Initial pipeline design
- Next steps
- References

Our Project

Purpose:

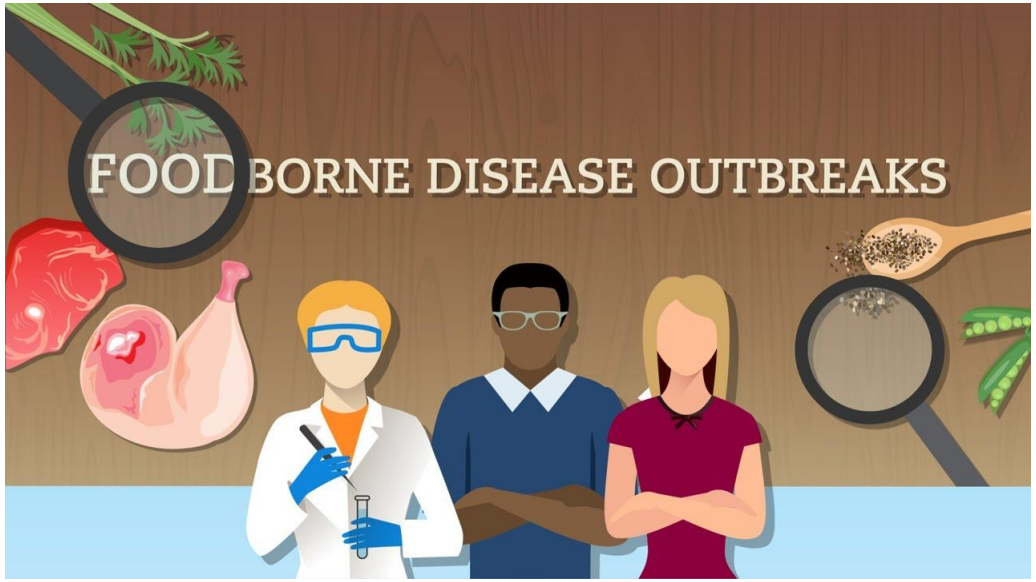
Investigate an unknown outbreak pathogen using raw genome sequence data from the Centers for Disease Control and Prevention (CDC) foodborne illness surveillance outbreak investigations

Goal:

Identify and characterize the pathogenic organism, make recommendations for the outbreak control, and build a public webserver that automates the computational steps

Objective for "Gene Prediction":

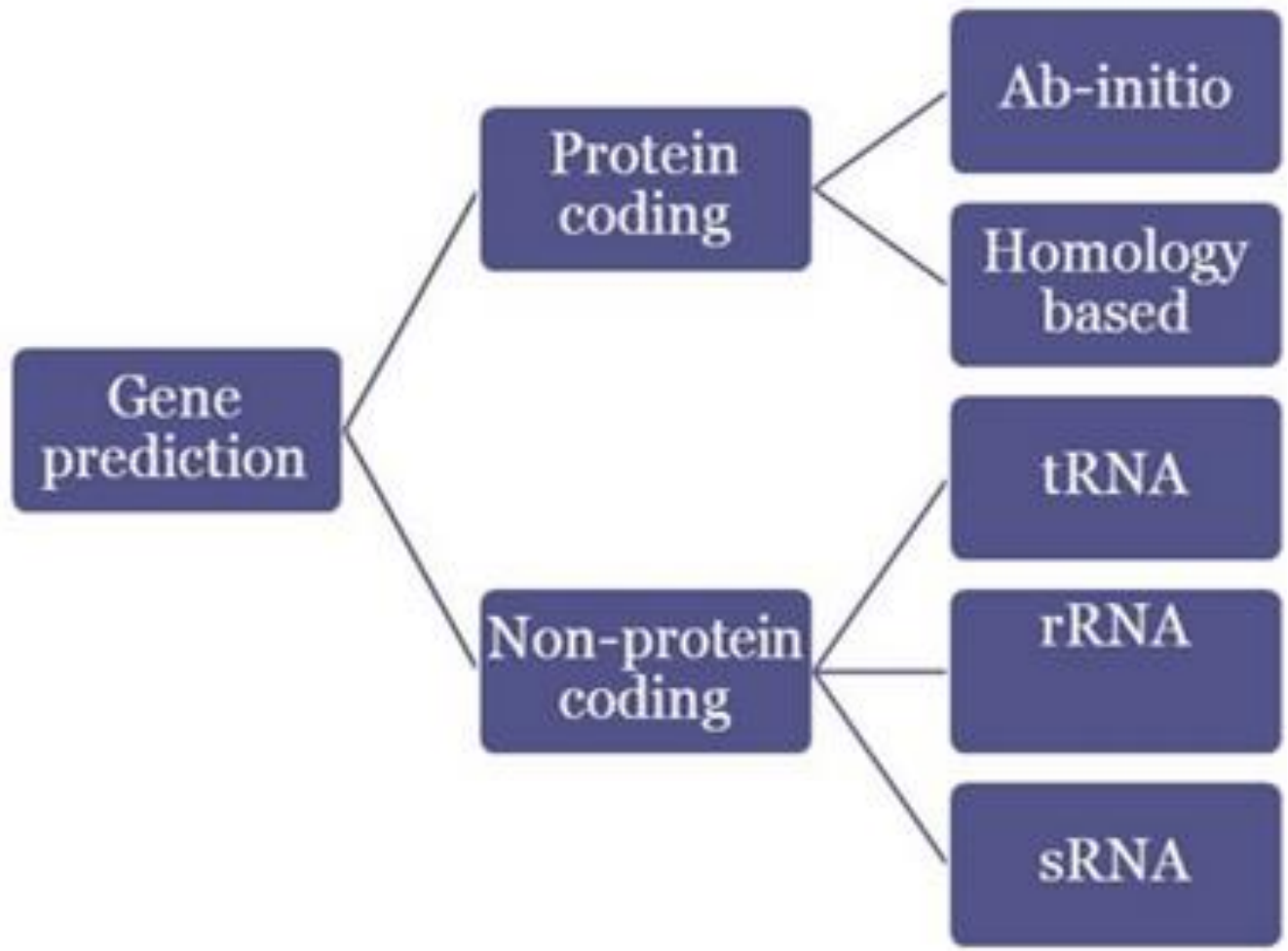
From assembled genomes, predict genes or features using different prediction methods and evaluate selected tools on their accuracy and performance



What is *Gene Prediction*?

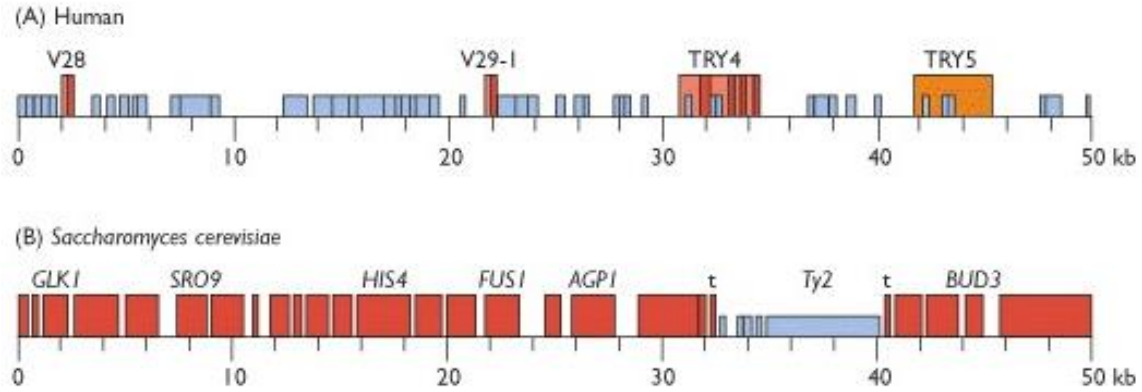
Identification of the regions of genomic DNA that encode genes, which are fragments of DNA that encodes a functional molecule:

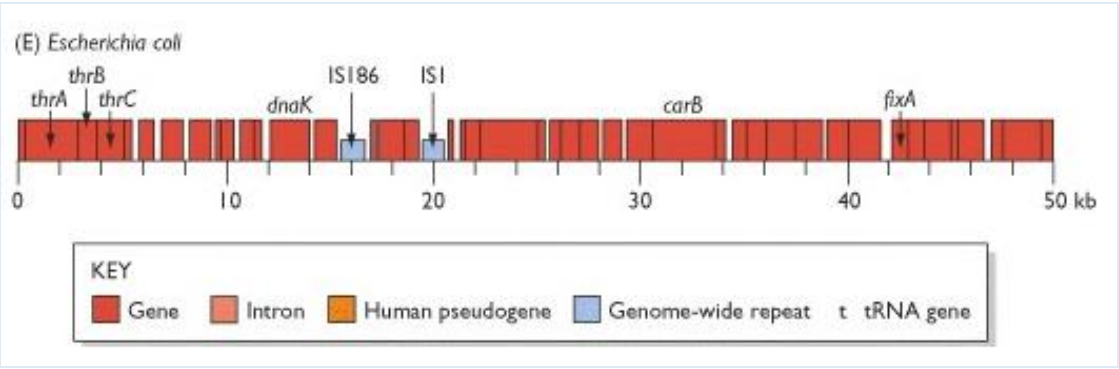
- Protein-coding genes
- RNA genes
- May also include other functional elements (i.e. regulatory regions)



Prokaryotic Genome

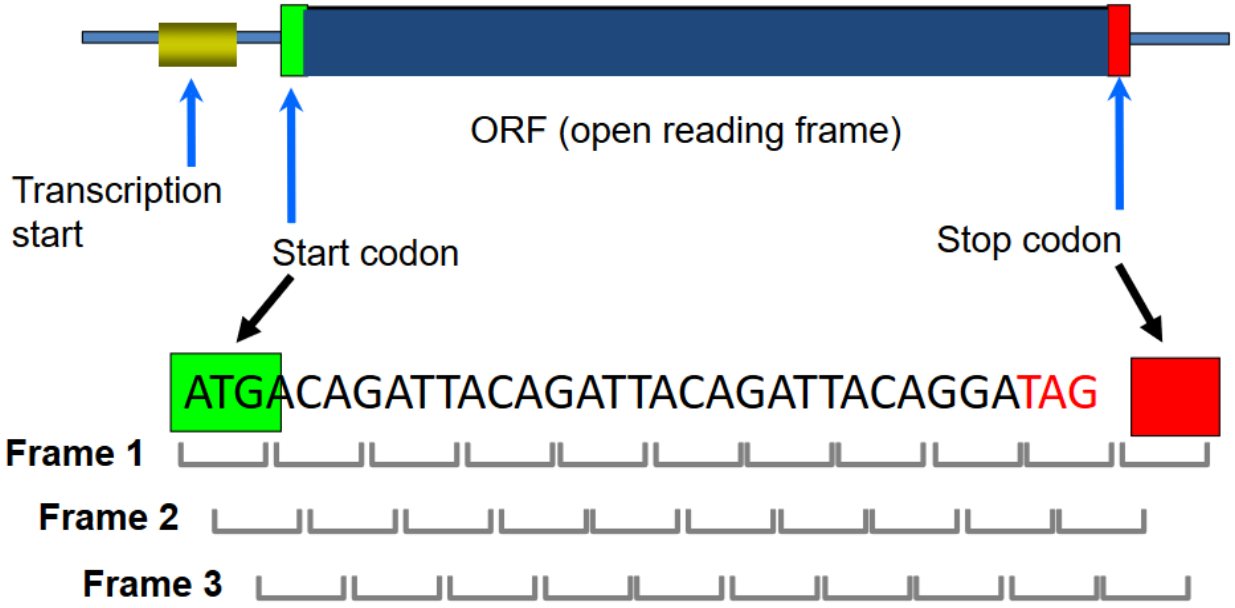
- Have a high gene density and do not contain introns in their protein coding regions
- Genes are called Open Reading Frames or “ORFs” (include start & stop codon)





Prokaryotic Genome (cont'd)

- Prediction of prokaryotic genes tends to be relatively simpler with contiguous ORFs
- However, overlapping ORFs and short genes can cause issues
- *Each gene is an ORF, but not every ORF is a gene*



Characteristics of *Campylobacter spp.*

- Domain: Bacteria
- Phylum: Proteobacteria
- Class: Epsilonproteobacteria
- Family: Campylobacteraceae
- Low G+C content (guanine-cytosine content) - GC ration is about 30 percent
- DNA ranges between 1.6-1.7 Mbps and contains a high content of adenine and thymine
- *Campylobacter jejuni* is the leading cause of bacterial diarrhea as well as the causative agent of gastroenteritis among human beings and animals.

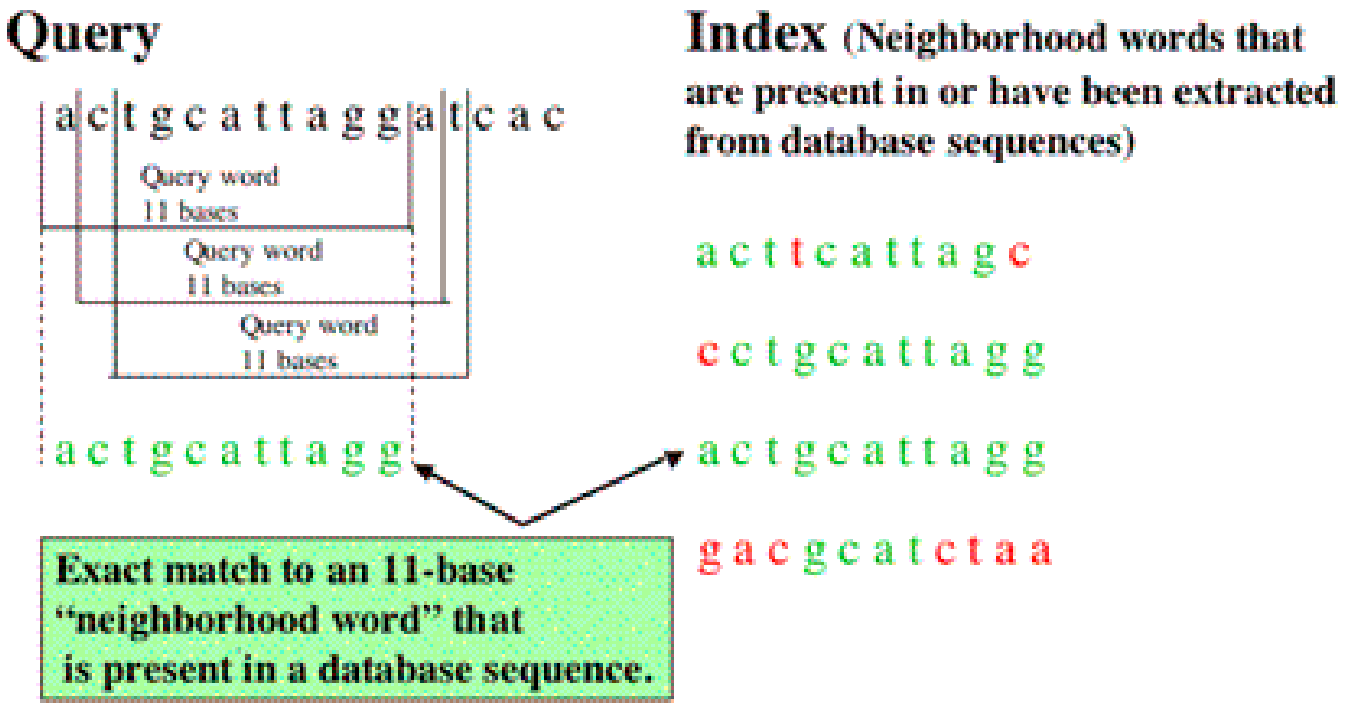
Homology Methods

- Makes predictions via comparisons with sequences of previously known genes
- Extrinsic information
- Can be used to validate/support Ab Initio findings
- Limited by the use of no new knowledge

BLAST+
Homology Tool #1

Identifying species, locating domains, establishing phylogeny, DNA mapping, comparison

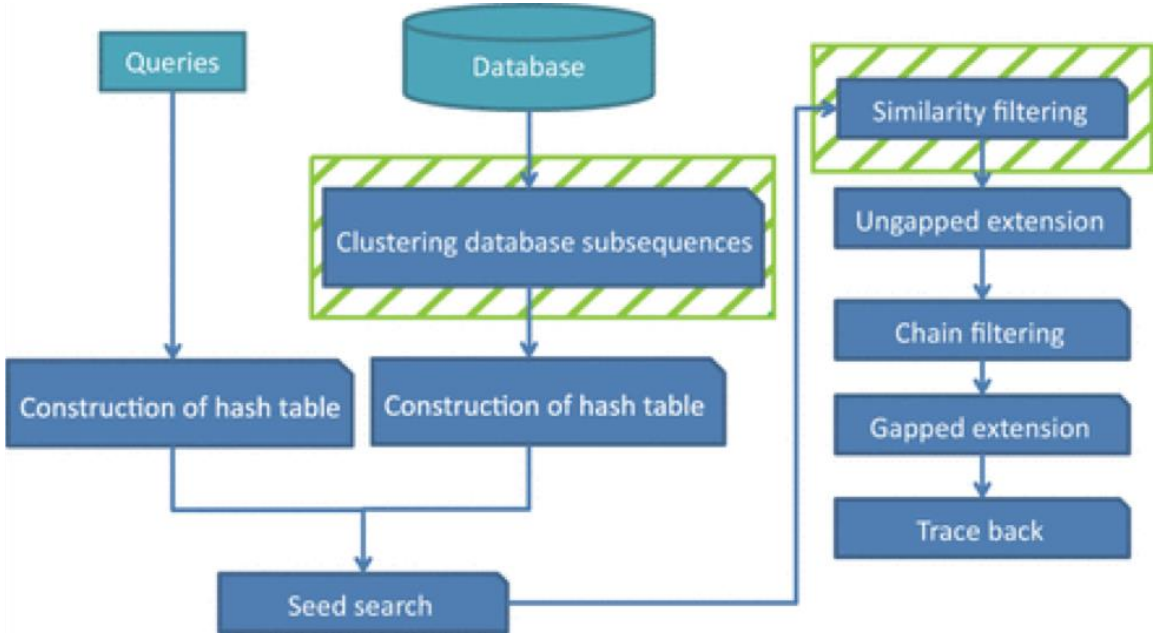
1. Break query into words of length **W**
2. Align words with sequence in database & identify matches
3. Calculate **T** score for matches
4. Extend sequence in both directions until score falls below cutoff (HSPs)
5. Report hits that meet or exceed BLAST cutoff for statistically significant hits

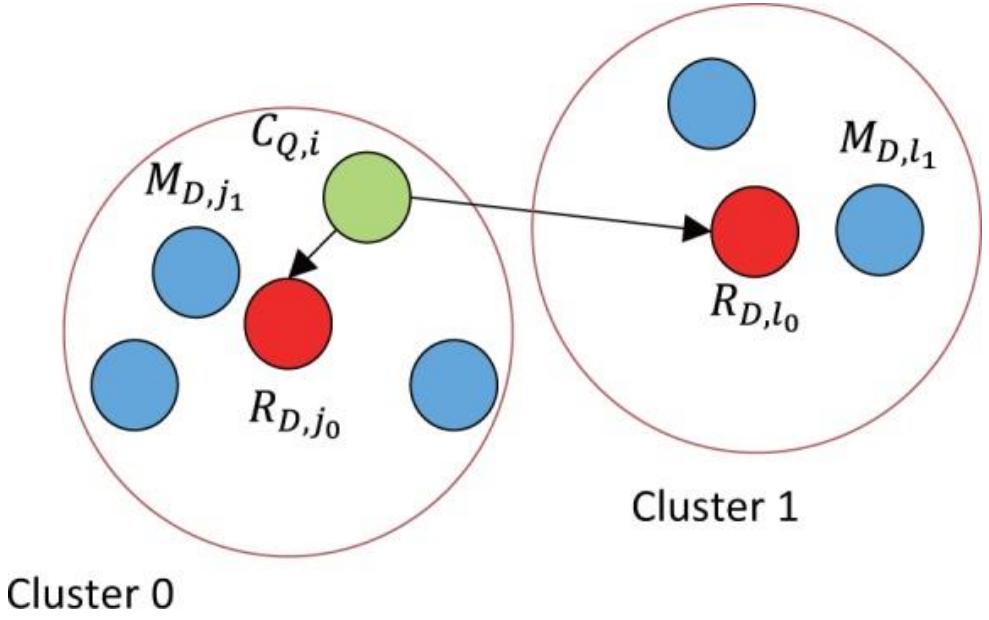


GHOSTZ
Homology Tool #2

A new faster homology search method using database subsequence clustering

1. Sequences are extracted from a database & similar ones are clustered
2. Construct into hash tables
3. Use hash tables to select seeds for the alignments from representative sequences in the clusters
4. Distance between a query subsequence and cluster representative is calculated
5. Lower bounds calculated
6. Similarity Filtering – if computed lower bound is less than or equal to distance threshold, continue

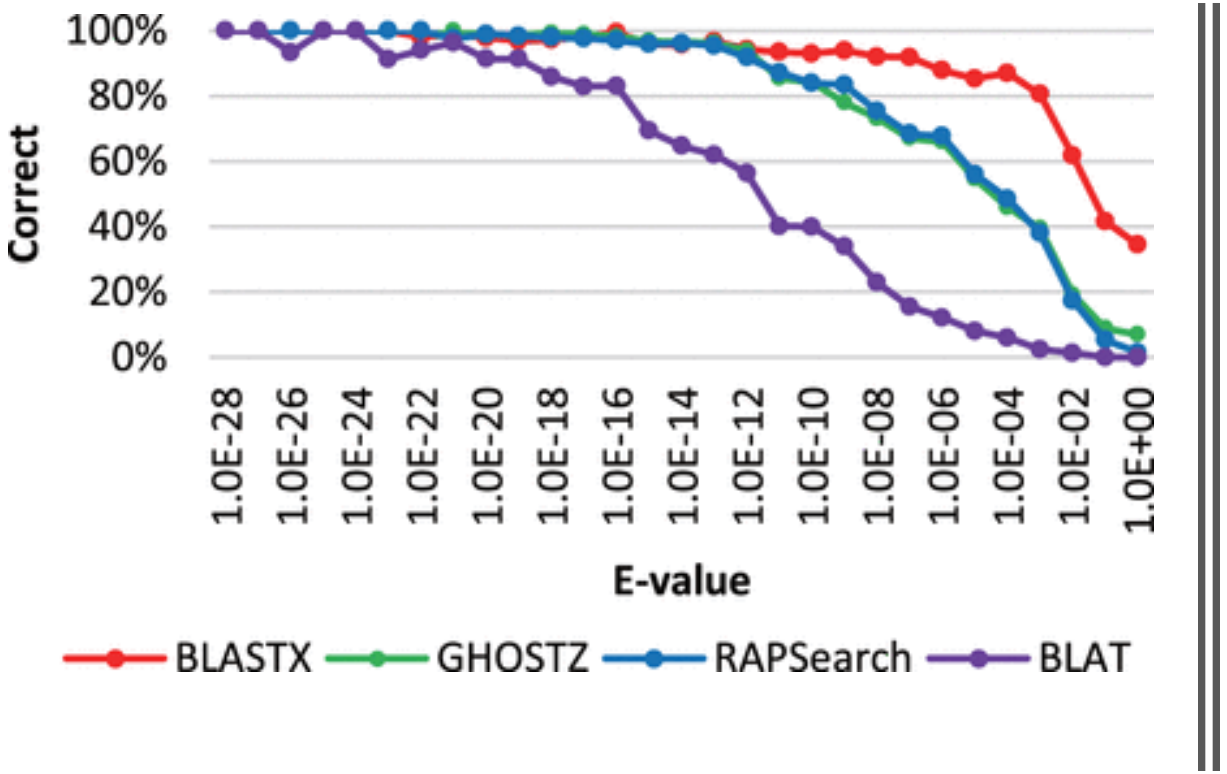




Homology Tools Comparison

	BLAST	GHOSTZ
PROS	• 50 times faster than dynamic programming • Computer storage efficient • Allows for gapped matches	• 200 times more efficient than BLAST • Does not depend on search sensitivity
CONS	• Less accurate than Smith-Waterman • May have low sensitivity	• Requires more computer storage

Homology Tools Comparison (cont'd)



	Computation time (s)	Acceleration ratio
GHOSTZ	460.8	261.3
RAPSearch	1285.5	93.7
BLAT	2514.9	47.9
BLASTX	120395.2	1.0

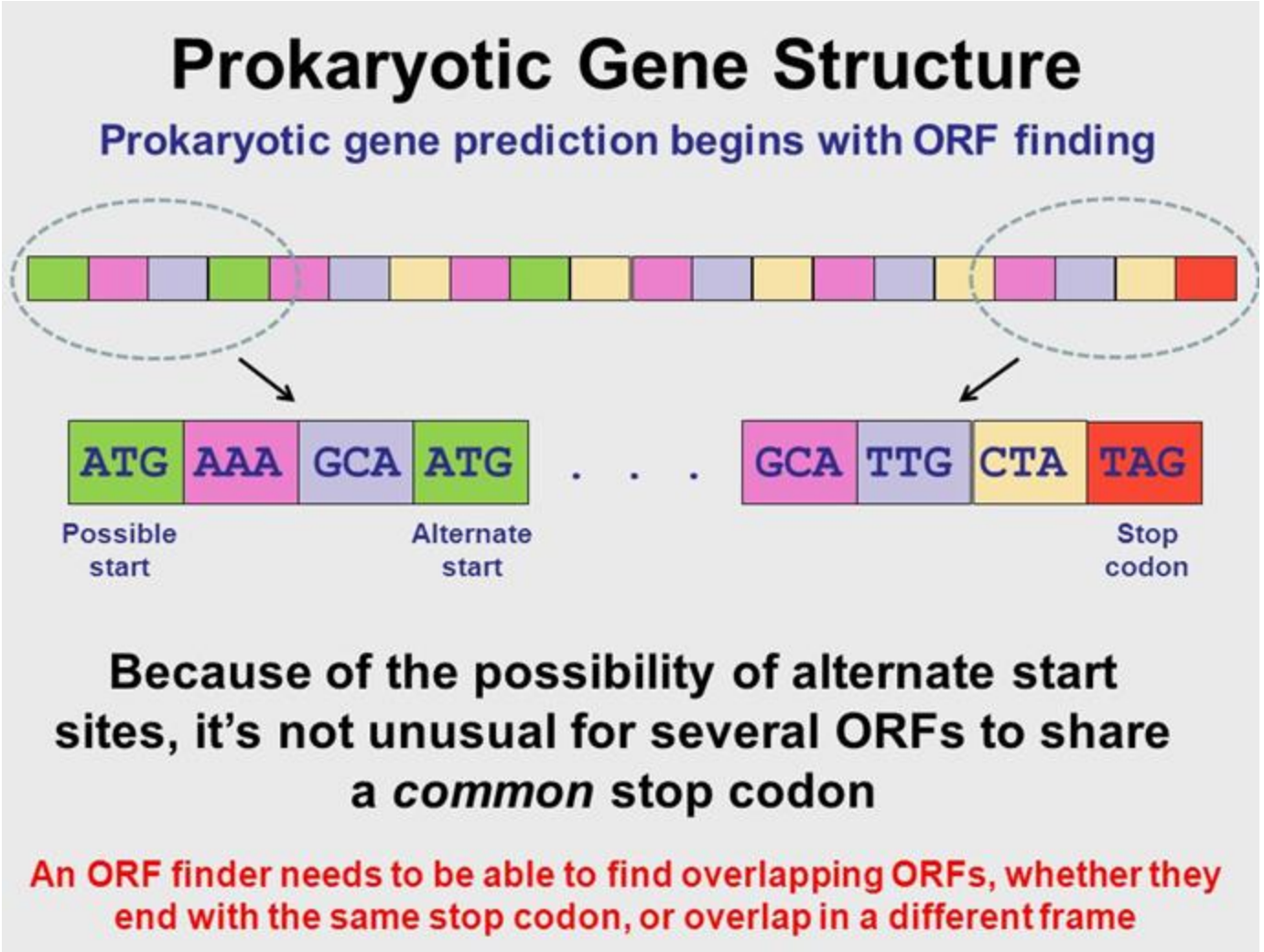
Suzuki, S., Kakuta, M., Ishida, T., & Akiyama, Y. (2015). Faster sequence homology searches by clustering subsequences. *Bioinformatics (Oxford, England)*, *31*(8), 1183–1190. doi:10.1093/bioinformatics/btu780

Homology Tools Pipeline

Reference Genome: Campylobacter jejuni from NCBI (NC_002163.1)
Query FASTA File: NCBI Reference Sequence NR_041834.1

Ab Initio Methods

- Inspect the input sequence and searches for traces of gene presence
- Simplest method is to inspect ORFs
- Relies on:
  - Probability models
  - Specific DNA motifs (signals)
- Markov Models and Dynamic Programming



Hidden Markov Models

- **Markov Model is** a chain structured process where future states depends only on the present state, not on the sequence of events that preceded it.
- Used to model randomly changing systems.
- **Hidden Markov Model (HMM**) is a statistical Markov model with hidden states
- Viterbi Algorithm used to find the most likely sequence of hidden paths.

80 % sunny
15% cloudy
5% rainy

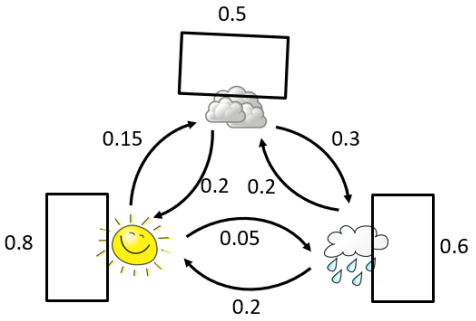
State : sunny cloudy rainy sunny

State transition probability (table/graph)

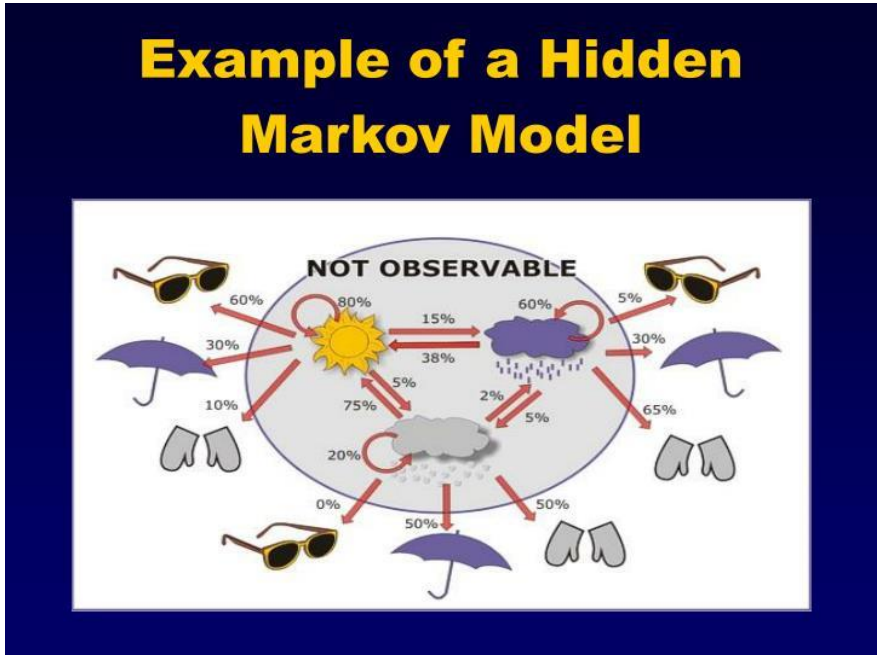
Output format 1:

Today	Tomorrow	Probability
sunny	sunny	0.8
sunny	rainy	0.05
sunny	cloudy	0.15
rainy	sunny	0.2
rainy	rainy	0.6
rainy	cloudy	0.2
cloudy	sunny	0.2
cloudy	rainy	0.3
cloudy	cloudy	0.5

Output format 3:



Weisstein et al. A Hands-on Introduction to Hidden Markov Models



Ab Initio Tools

Gene Finder	# Genes	# Genes on the + Strand	# Genes on the - Strand	#Correct Genes	% Correct Genes (compared to the Original)	% Correct Genes from (from all found genes)
Original	6061	2993	3067	6061	100,00%	100,00%
Prodigal	6055	3014	3041	5286	89,14%	87,30%
FGenesB	6197	3094	3103	5070	85,50%	81,81%
Glimmer3.0	6276	3100	3176	5043	85,04%	80,35%
GeneMarkS	6100	3043	3057	5006	84,42%	82,07%
JCVI	6270	3098	3172	5036	83,10%	80,32%
GeneMarkHMM	6129	3055	3074	4920	82,97%	80,27%
Rast	6297	3116	3181	4940	81,52%	78,45%
MED	7475	3708	3767	4747	80,05%	63,51%
Maker with model	6149	3065	3084	4588	75,71%	74,61%
Maker	5884	2904	2980	4370	72,11%	74,27%
Augustus	5268	2587	2681	3529	59,51%	66,99%
AMIGene	6154	3077	3077	2967	50,03%	48,21%
EasyGene	3150	0	3150	2570	43,34%	81,59%

A

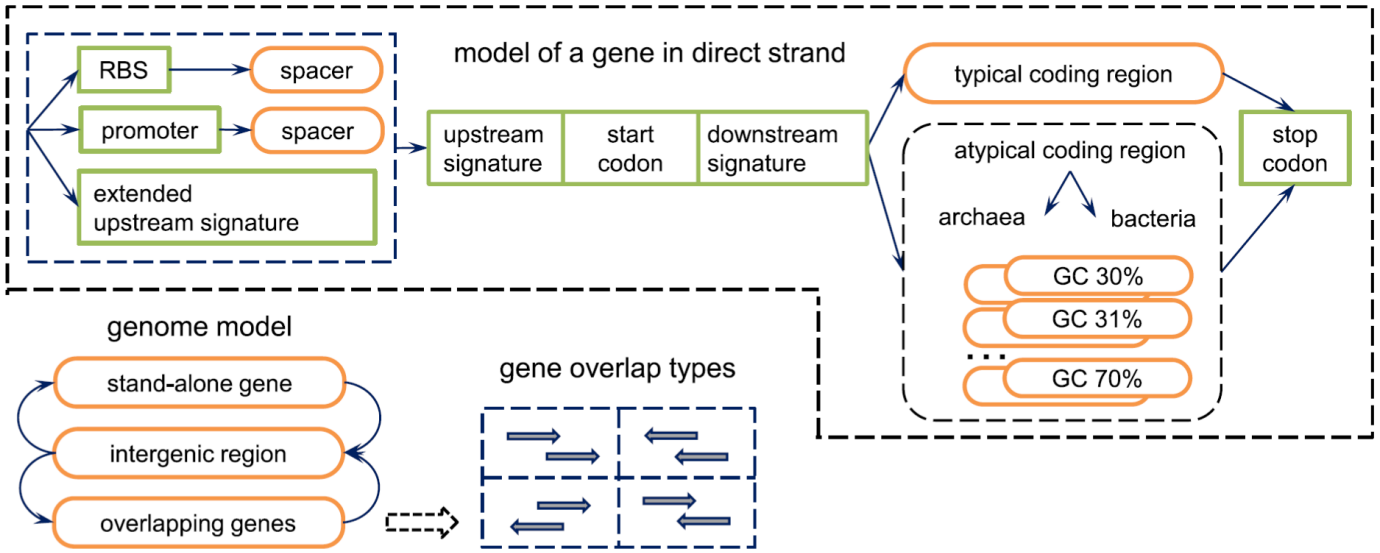
Algorithm	Missed MS confirmed genes (from 89,466)	Missed COG genes (not MS) (from 287,237)
GeneMarkS	376	1467
Glimmer3	496	1990
Prodigal	217	1389
GeneMarkS-2	181	1147

B

Algorithm	False predictions overlapping MS-confirmed genes	False predictions overlapping COG genes (not MS)
GeneMarkS	352	2046
Glimmer3	921	6435
Prodigal	211	1339
GeneMarkS-2	114	932

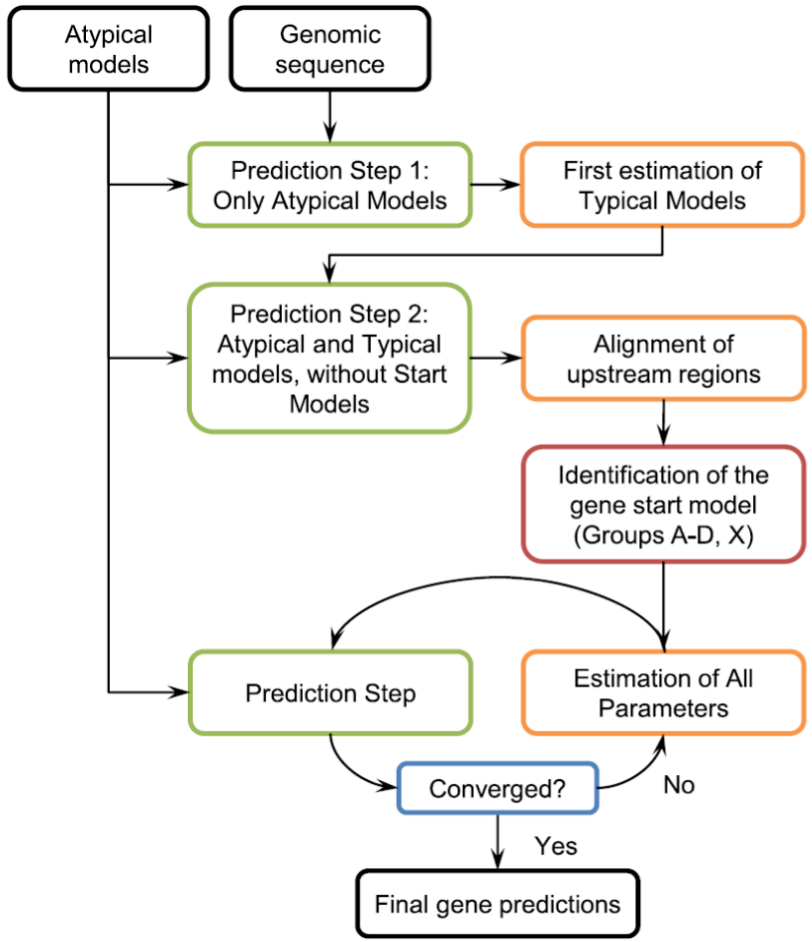
GeneMarkS-2/ Gene Mark S Ab Initio Tool #1

- Uses HMM and a self training algorithm (non supervised) to predict genes.
- 5th Order HMM for coding and 2nd order for non-coding regions.
- Uses a complex model to predict the prokaryotic gene.
- Identifies several different types of distinct sequence patterns.
- The model which yields the highest log-odds score is selected



GeneMarkS-2 (cont'd)

- Classifies the genome into 4 distinct groups:
  - Group A: Typical Model of Prokaryotes having RBS sites having (SD)Consensus
  - Group B: Atypical Model having RBS sites not having SD consensus
  - Group C and D: Represent Bacterial and Archeal Genomes (Leaderless Transcription).
  - Group X: Weak, Hard to classify regulatory signal patterns
- Algorithm stops after 10 iterations in the final prediction step, if it doesn't converge



PRODIGAL
Ab Initio Tool #2

- Prokaryotic Dynamic Programming Genefinding Algorithm.
- Looks at GC bias for each of three codon positions and chooses the one with highest GC content.
- Prodigal scores every start-stop pair above 90 bp in the entire genome based on simple GC codon statistics.
- Penalizes or gives bonus to intergenic spaces according to gene distance.
- Then uses Dynamic Programming to force the program to choose between two heavily overlapping ORFS.
- Sacrifices some genuine predictions to eliminate a much larger number of false identifications.

PRODIGAL (cont'd)

PROS	CONS
• Provide fast, accurate protein-coding gene prediction. • Runs unsupervised. • Handles gaps and partial genes. • Identifies translation initiation sites. • Open Source. • Higher accuracy in GC rich genomes. • Predicts Genes in 3 Formats. (GFF/GenBank/Sequin)	• The results could be biased. • To minimize false positives, sacrifices some genuine predictions. • Cannot Handle Introns (Works only on Prokaryotes).

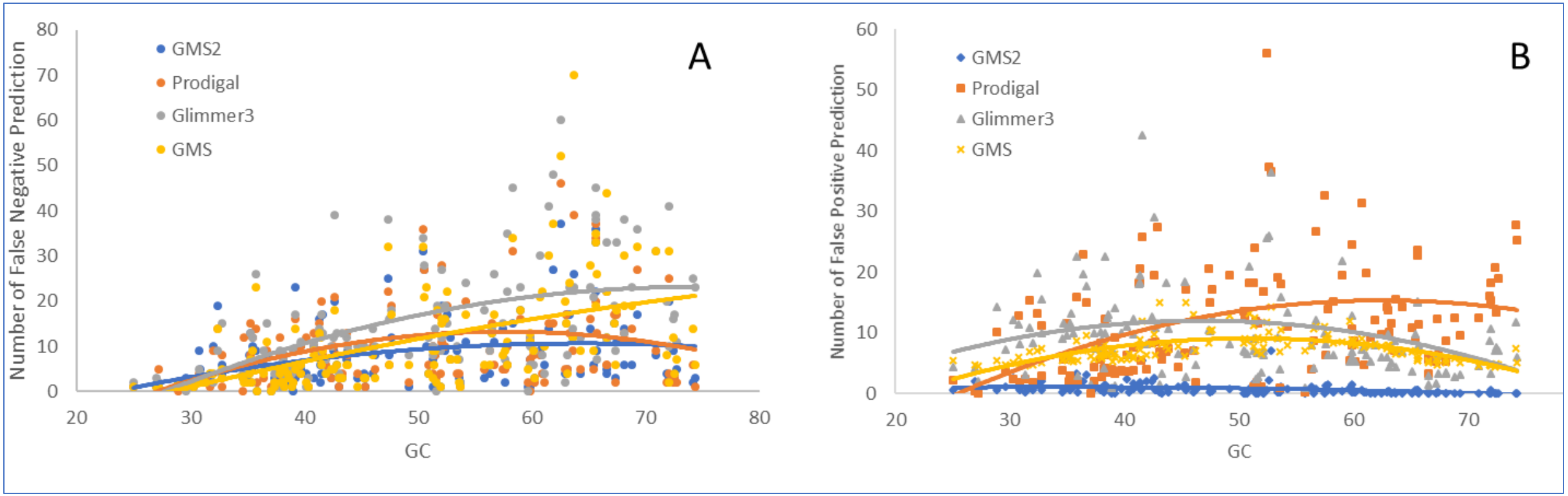
Glimmer3
Ab Initio Tool #3

- Identify genes within microbial DNA sequences (bacteria, archaea, and viruses)
- Requires training of samples genes
- Uses a dynamic programming algorithm to choose the highest-scoring set of orfs and start sites.
- Glimmer extracts every sufficiently long ORF from the sequence and scores it by the log-likelihood ratio of generating the ORF between models trained on coding versus non-coding sequence.

Glimmer3 (cont'd)

- Utilizes an **Interpolated Markov Model (IMM)**
  - Combines 1st through 8th order Markov models
- In Glimmer3 orfs are scored from 3’ end to 5’ end, *i.e.*, from stop codon back toward start codon, which helps find the start site.
- Builds Interpolated Context Model
- For each ORF:
  - calculate the probability of the ORF sequence in each of the 6 possible reading frames
  - if the highest scoring frame corresponds to the reading frame of the ORF, mark the ORF as a gene
- However, it does not work as well on high-GC genomes because it trains on long ORFs

Ab-initio Tool Comparison: CG Content



Ab-initio Tool Comparison: Gene Length

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).

Ab-initio Tool Comparison: Gene Length

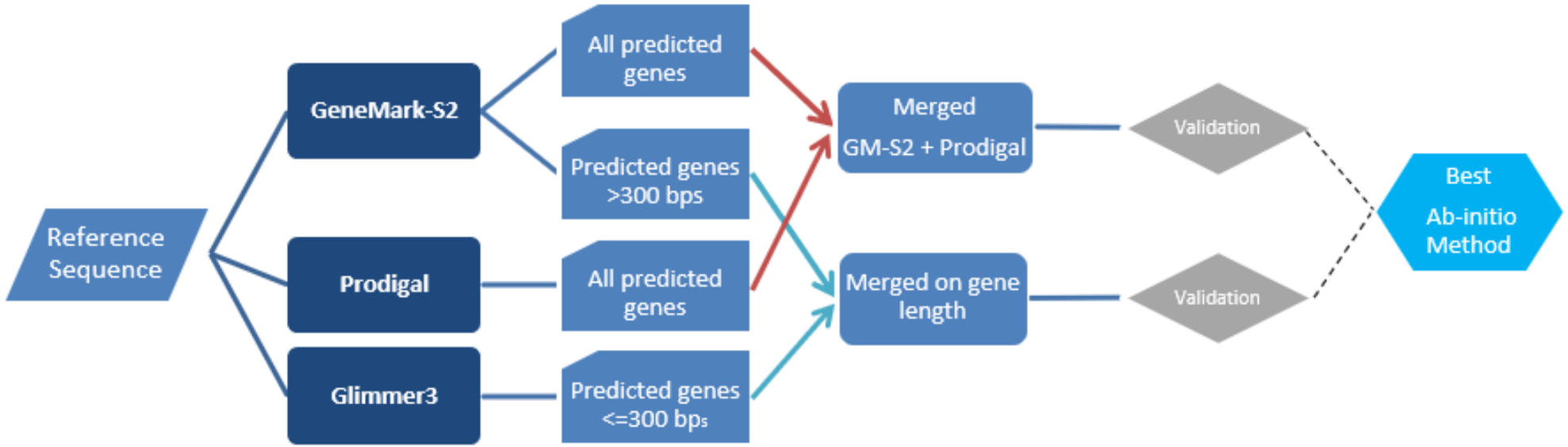
Table 3. Statistics of false negative (panel *A*) and false positive (panel *B*) gene predictions

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).

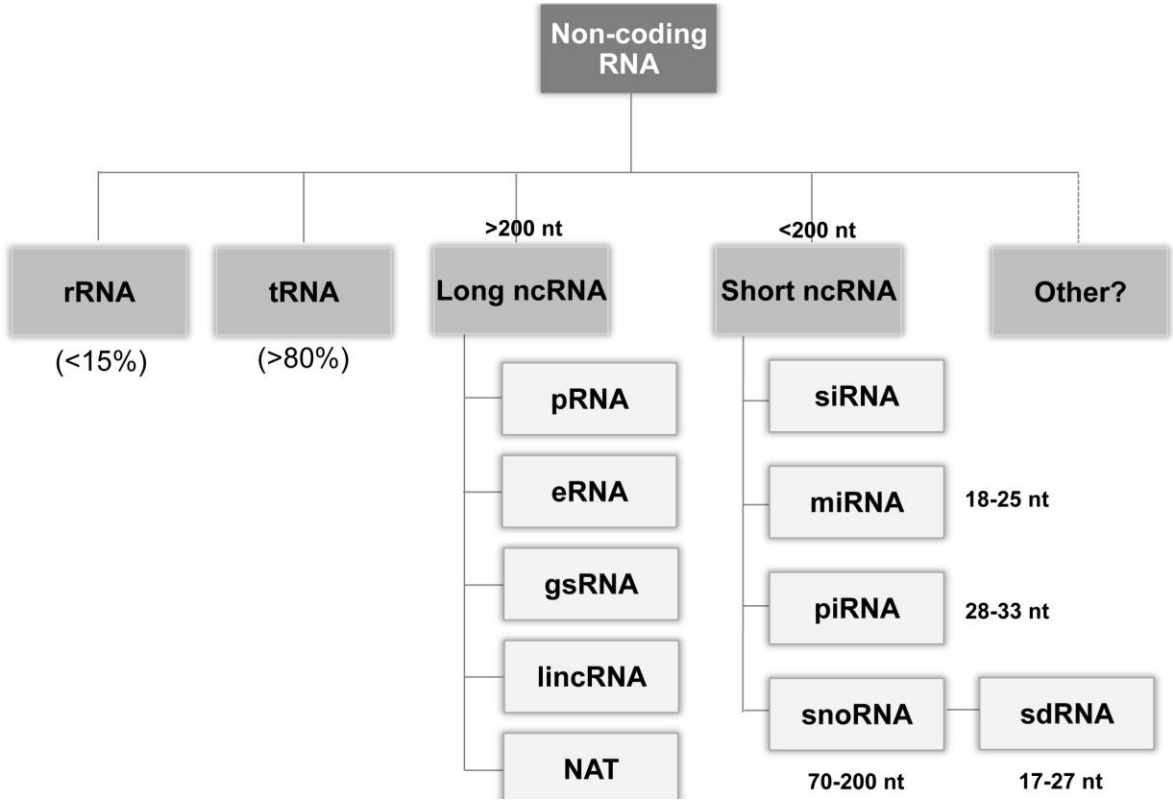
Tool Evaluation Plan

1. Use reference sequence on GeneMark-S2, Prodigal and Glimmer3
2. Prepare MERGED data in 2 ways:
    1. All predicted genes: Prodigal + GeneMark-S2
    2. Genes by gene length:
        1. >300 bps: GeneMark-S2
        2. <=300 bps: Glimmer3
3. Validate using “best” homology method
    1. Check for sensitivity, specificity, etc.
4. Select the best Ab-initio method and proceed with our data



Non-coding RNA

- A non-coding RNA (ncRNA) is an RNA molecule that is not translated into a protein
- transfer RNAs (tRNAs), ribosomal RNAs (rRNAs) and small RNAs(sRNAs)
- Role of ncRNA in bacterial genomes:
  - Protein synthesis/Translation (tRNA and rRNA)
  - Gene regulation (sRNA)
  - Related to antibiotic resistance



ARAGORN tRNA Tool

- Homology based tool
- Uses the heuristic algorithms that score the tRNA and tmRNA genes based on their sequence and secondary structure similarities.
- an effective tRNA search program, with sensitivity better than other current heuristic tRNA search algorithms.

Lineage	Genome	No. of tRNAs detected		Search time (s)[b]	
		ARAGORN[c]	tRNAscan-SE[d]	ARAGORN[c]	tRNAscan-SE[
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

[a]tRNAscan-SE version 1.23.
[b]Tested on an AMD Athlon, 1.6 GHz, 1024 Mb RAM with Linux.
[c]ARAGORN run with a maximum intron size of 100 nucleotides and the –t switch (tRNA detection only) The intron size roughly corresponds to the default used by tRNAscan-SE.

RNAmmer rRNA Tool

- Ab Initio based tool
- It uses Hidden Markov Models trained on data from 5s rRNA database.
- fast with little loss of sensitivity, enabling the analysis of a complete bacterial genome in less than a minute.
- the location of rRNAs can be predicted with a very high level of accuracy.

OPEN ACCESS Freely available online

PLOS ONE

Comprehensive Genomic Characterization of *Campylobacter* Genus Reveals Some Underlying Mechanisms for its Genomic Diversification

Yizhuang Zhou[1], Lijing Bu[2], Min Guo[1], Chengran Zhou[3], Yongdong Wang[4], Liyu Chen[5]*, Jie Liu[6]*

1 BGI-Shenzhen, Shenzhen, Guangdong Province, China, **2** Biology Department of University of New Mexico, Albuquerque, New Mexico, United States of America, **3** Department of Biology, Sichuan University, Chengdu, Sichuan Province, China, **4** Key Discipline Laboratory for National Defense for Biotechnology in Uranium Mining and Hydrometallurgy, University of South China, Hengyang, Hunan Province, China, **5** Department of Microbiology, Xiangya School of Medicine, Central South University, Changsha, Hunan Province, China, **6** Translational Center for Stem Cell Research, Tongji Hospital, Stem Cell Research Center, Tongji University School of Medicine, Shanghai, China

Infernal ncRNA Tool

- an implementation of covariance models (CMs)
- RNA homology search based on accelerated profile hidden Markov model (HMM) methods and HMM-banded CM alignment methods
- 100-fold faster RNA homology searches and ~10 000-fold acceleration over exhaustive non-filtered CM searches.

Rfam: an RNA family database

Sam Griffiths-Jones*, Alex Bateman, Mhairi Marsha

The Wellcome Trust Sanger Institute, Wellcome Trust Genome Ca
[1]Howard Hughes Medical Institute and Department of Genetics, W
St Louis, MO 63110, USA

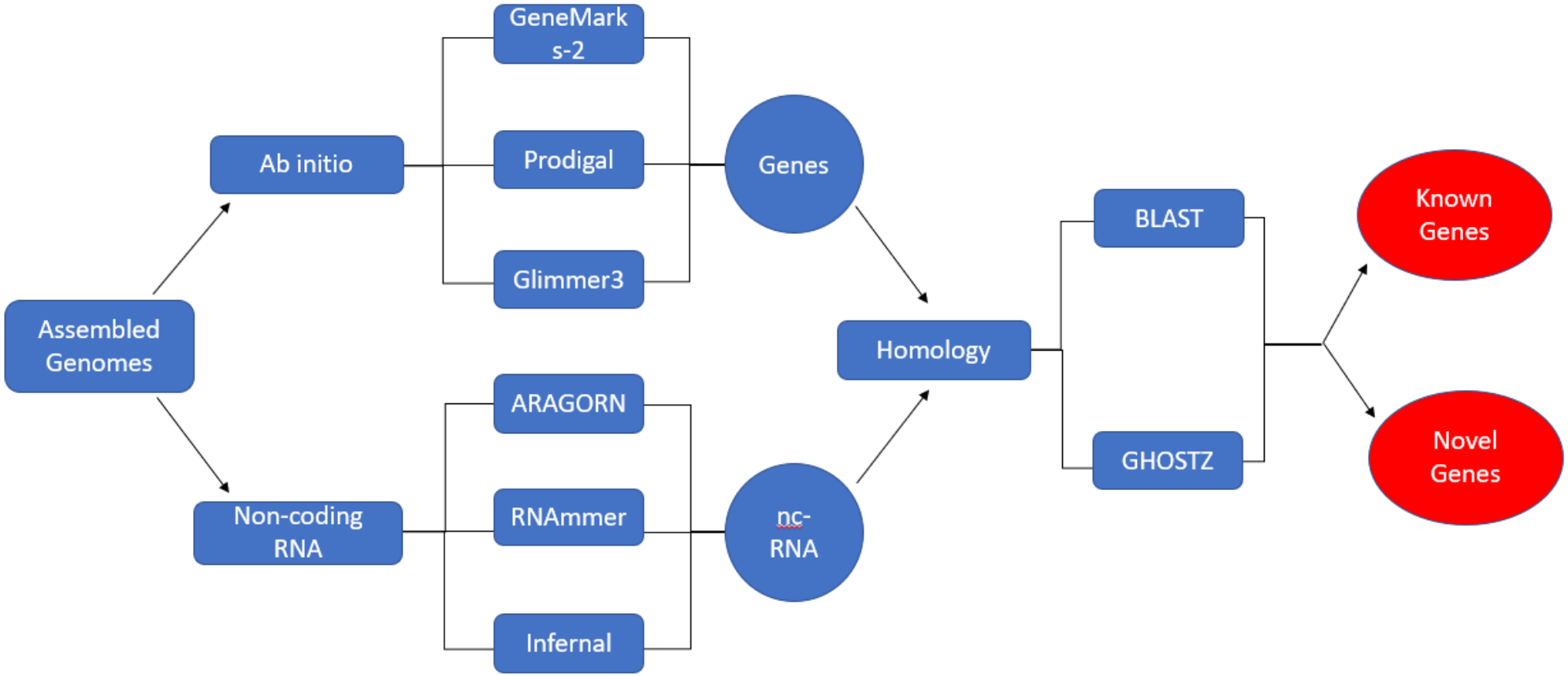
Received August 15, 2002; Accepted September 1, 2002

ABSTRACT

Rfam is a collection of multiple sequence alignments and covariance models representing non-coding RNA families. Rfam is available on the web in the UK at http://www.sanger.ac.uk/Software/Rfam/ and in the US at http://rfam.wustl.edu/. These websites allow the user to search a query sequence against a library of covariance models, and view multiple sequence alignments and family annotation. The database can also be downloaded in flatfile form and searched locally using the INFERNAL package (http://infernal.wustl.edu/). The first release of Rfam (1.0) contains 25 families, which annotate over 50 000 non-coding RNA genes in the taxonomic divisions of the EMBL nucleotide database.

RNA s
RNA st
curation
structur
called
led us t
a databa
the use
consens
of prote
Sever
and inf
Riboson
uRNA
others (
large ar
vary gr
There a

Initial Pipeline



Next Steps

- Test out best tools for Homology method
- Perform Ab-initio tool evaluation and merge the results with non-coding RNA prediction results
- Validate using the selected homology-based method
- Output data in GFF format for the Functional Annotation group

References

1. Angelova, Mihaela & Kalajdziski, Slobodan & Kocarev, Ljupco. (2010). Computational Methods for Gene Finding in Prokaryotes. ICT Innovations. 1. 1857-7288.
2. Hyatt, D., Chen, G., LoCascio, P., Land, M., Larimer, F. and Hauser, L. (2020). *Prodigal: prokaryotic gene recognition and translation initiation site identification*.
3. Lomsadze, A., Gemayel, K., Tang, S. and Borodovsky, M. (2018). Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Research*, 28(7), pp.1079–1089.
4. https://www.ndsu.edu/pubweb/~mcclean/plsc411/Blast-explanation-lecture-and-overhead.pdf
5. https://www.ccg.unam.mx/~vinuesa/tlem/pdfs/Bioinformatics_explained_BLAST.pdf
6. https://academic.oup.com/bioinformatics/article/31/8/1183/212151
7. https://www.semanticscholar.org/paper/ARAGORN%2C-a-program-to-detect-tRNA-genes-and-tmRNA-Laslett-Canback/
8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3810854/
9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3214069/#pone.0027121.s012
10. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4450053/

QUESTIONS?