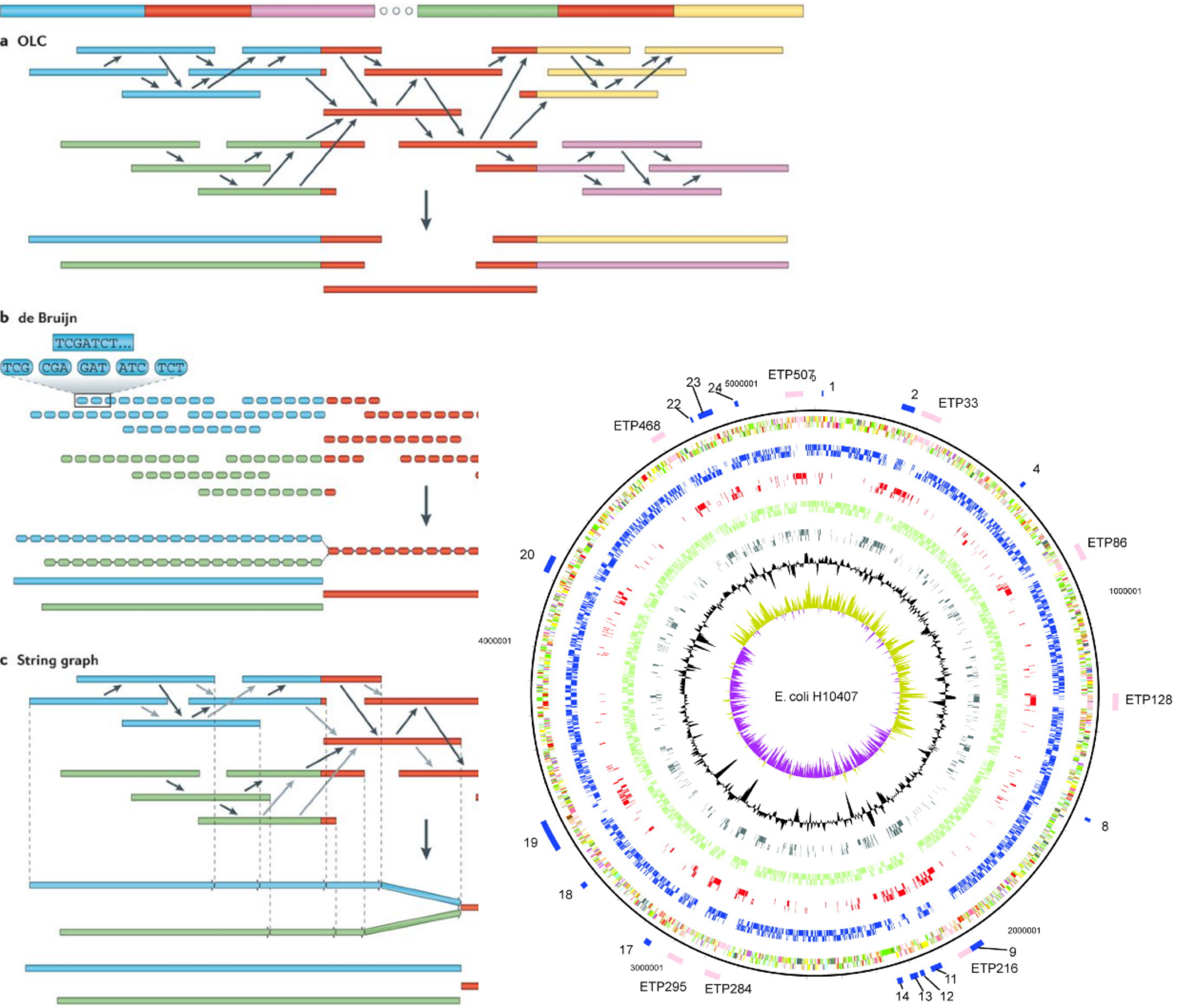
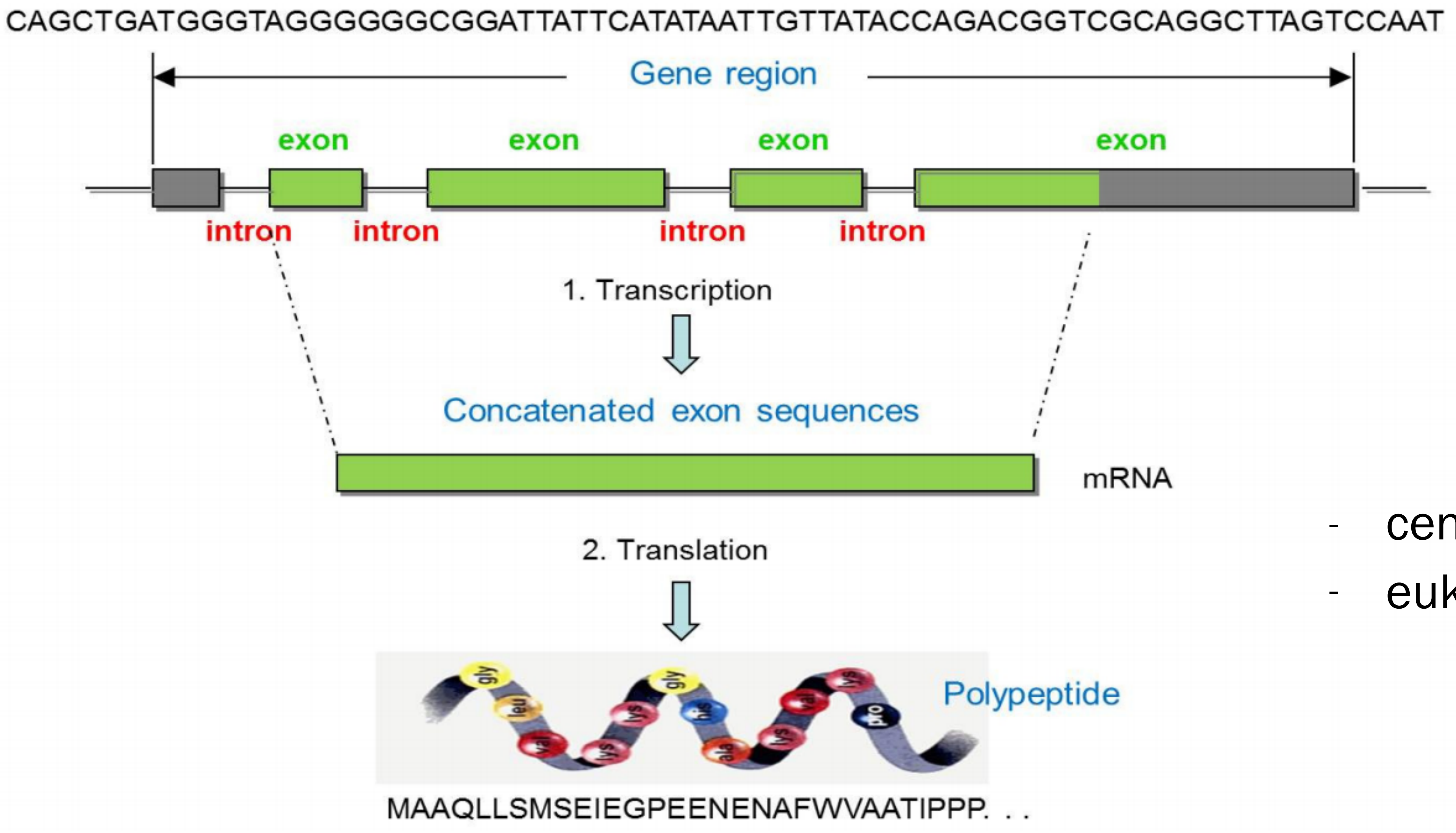
Genome Annotation



HTS Workshop Genomics & Transcriptomics KAUST 2019

Robert Lehmann
Octavio Salazar

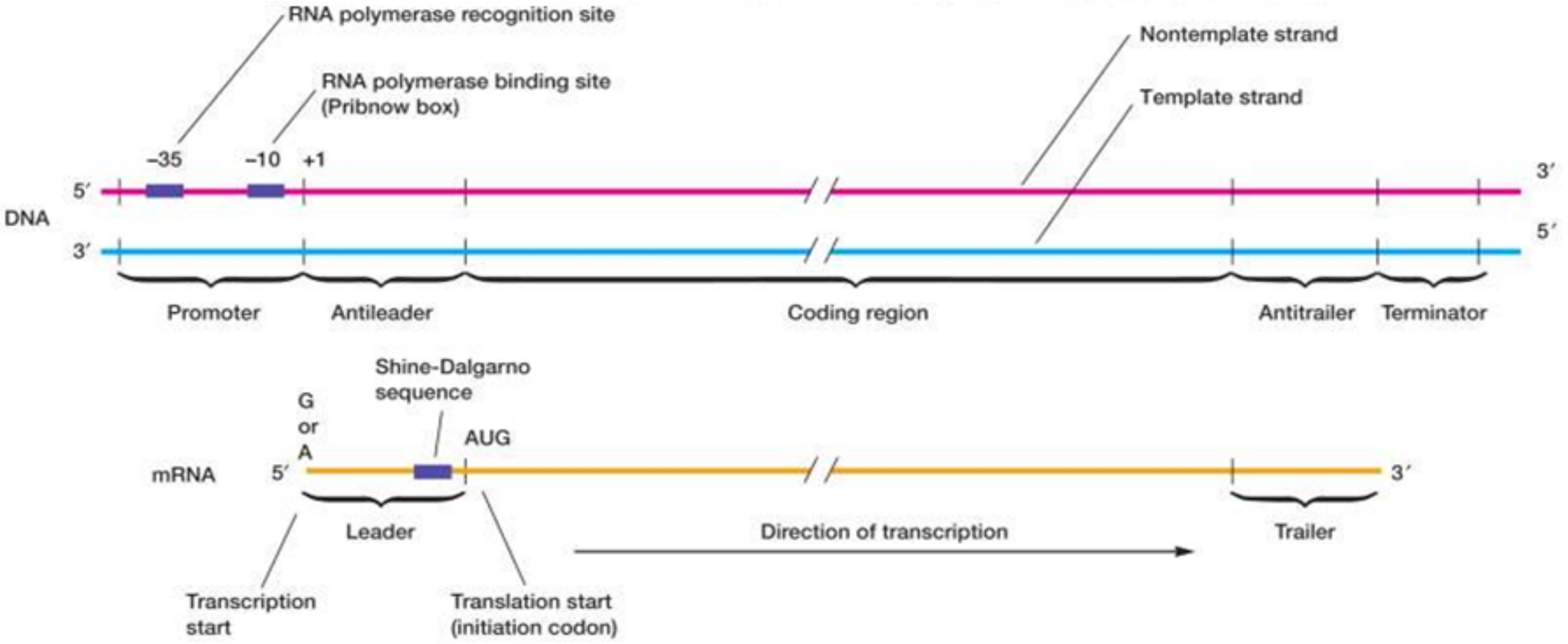
How to find genes on the assembly



- central dogma
- eukaryotic genes -> splicing

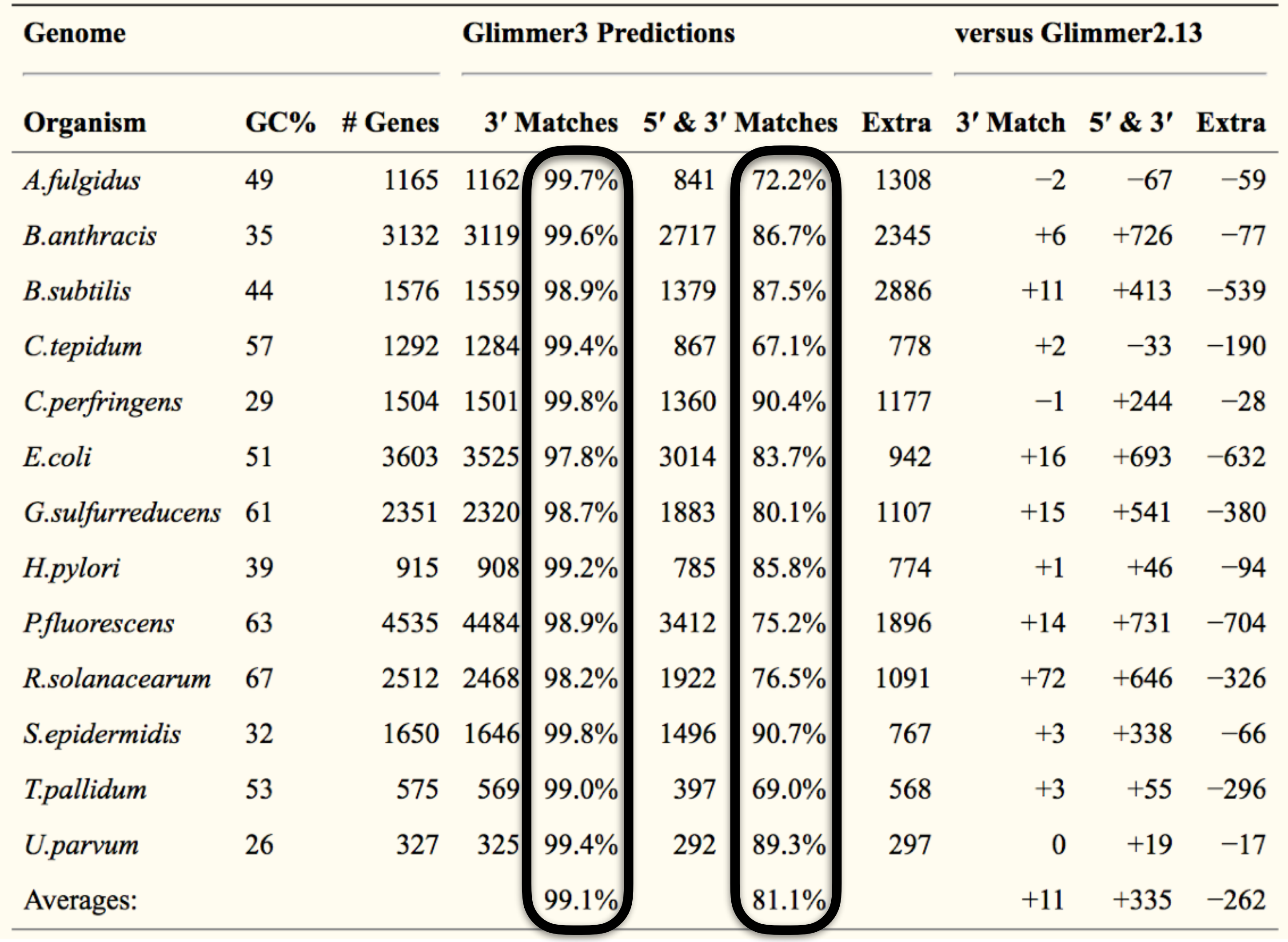
Bacterial gene structure

- no splicing
- strong sequence motifs indicating gene locations



https://nptel.ac.in/courses/102103015/module7/lec3/2.html

Bacterial gene annotation

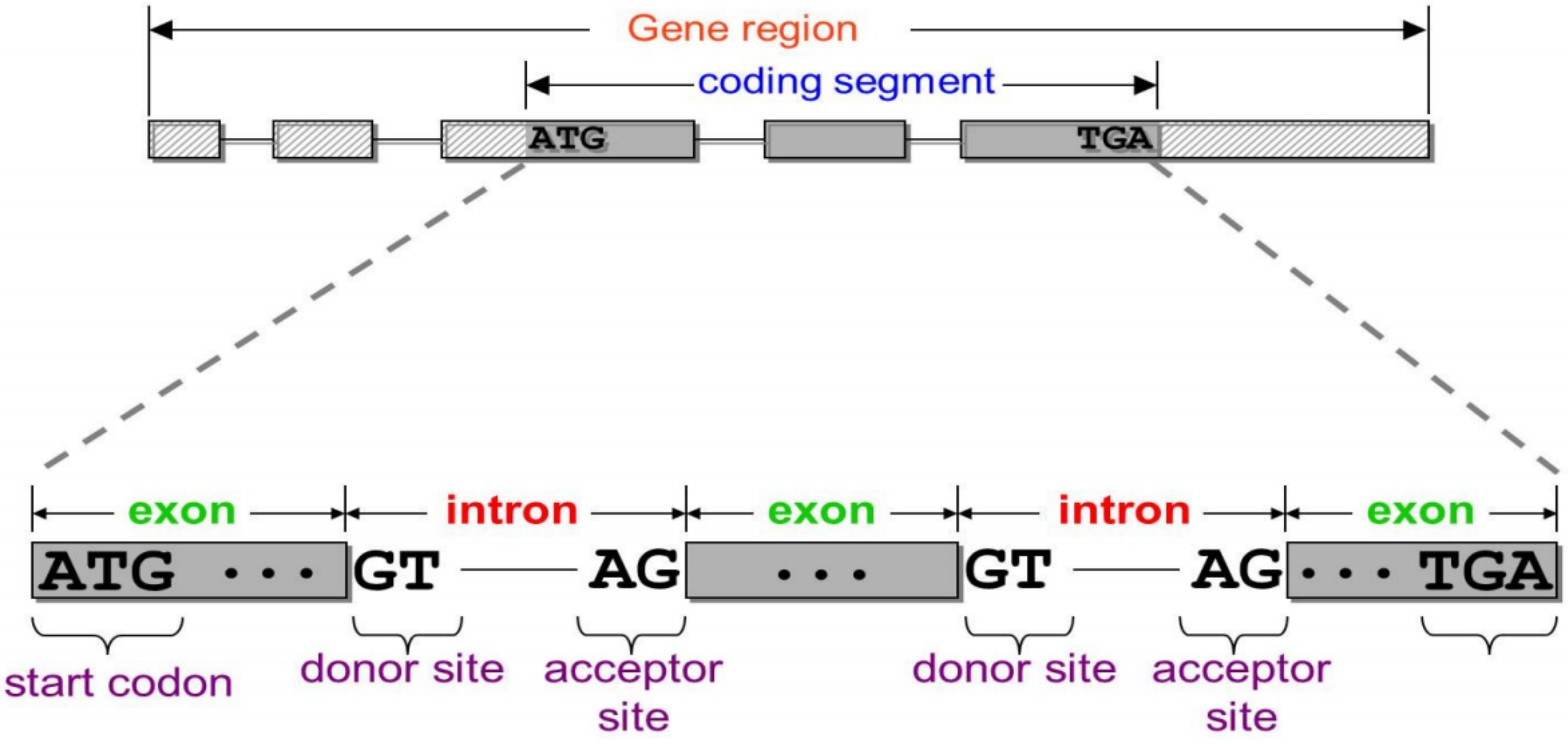


Genome			Glimmer3 Predictions					versus Glimmer2.13		
Organism	GC%	# Genes	3′ Matches		5′ & 3′ Matches		Extra	3′ Match	5′ & 3′	Extra
A.fulgidus	49	1165	1162	99.7%	841	72.2%	1308	−2	−67	−59
B.anthracis	35	3132	3119	99.6%	2717	86.7%	2345	+6	+726	−77
B.subtilis	44	1576	1559	98.9%	1379	87.5%	2886	+11	+413	−539
C.tepidum	57	1292	1284	99.4%	867	67.1%	778	+2	−33	−190
C.perfringens	29	1504	1501	99.8%	1360	90.4%	1177	−1	+244	−28
E.coli	51	3603	3525	97.8%	3014	83.7%	942	+16	+693	−632
G.sulfurreducens	61	2351	2320	98.7%	1883	80.1%	1107	+15	+541	−380
H.pylori	39	915	908	99.2%	785	85.8%	774	+1	+46	−94
P.fluorescens	63	4535	4484	98.9%	3412	75.2%	1896	+14	+731	−704
R.solanacearum	67	2512	2468	98.2%	1922	76.5%	1091	+72	+646	−326
S.epidermidis	32	1650	1646	99.8%	1496	90.7%	767	+3	+338	−66
T.pallidum	53	575	569	99.0%	397	69.0%	568	+3	+55	−296
U.parvum	26	327	325	99.4%	292	89.3%	297	0	+19	−17
Averages:				99.1%		81.1%		+11	+335	−262

- up to 92% precise gene prediction possible (>98% for gene ends)

Identifying bacterial genes and endosymbiont DNA with Glimmer. AL Delcher, KA Bratke, EC Powers, SL Salzberg. Bioinformatics. 2007 Mar 15;23(6):673-9.

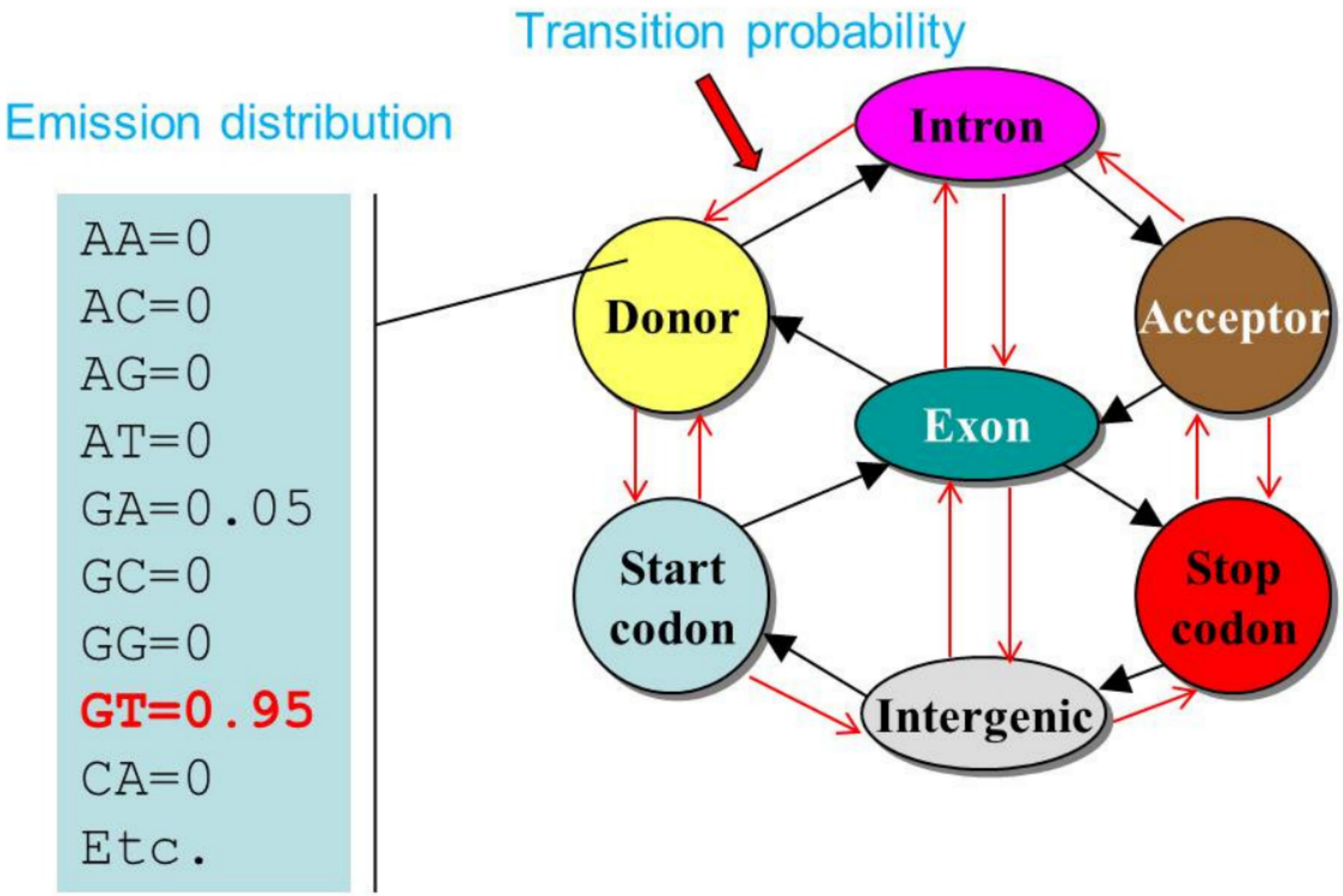
Eukaryotic gene structure



- splicing makes prediction more complex

Evaluating High-Throughput Ab Initio Gene Finders to Discover Proteins Encoded in Eukaryotic Pathogen Genomes Missed by Laboratory Techniques. SJ Goodswen, PJ Kennedy, JT Ellis, PLoS ONE 7(11): e50609.

HMMs for gene annotation



Transition probability

Emission distribution

AA=0
AC=0
AG=0
AT=0
GA=0.05
GC=0
GG=0
GT=0.95
CA=0
Etc.

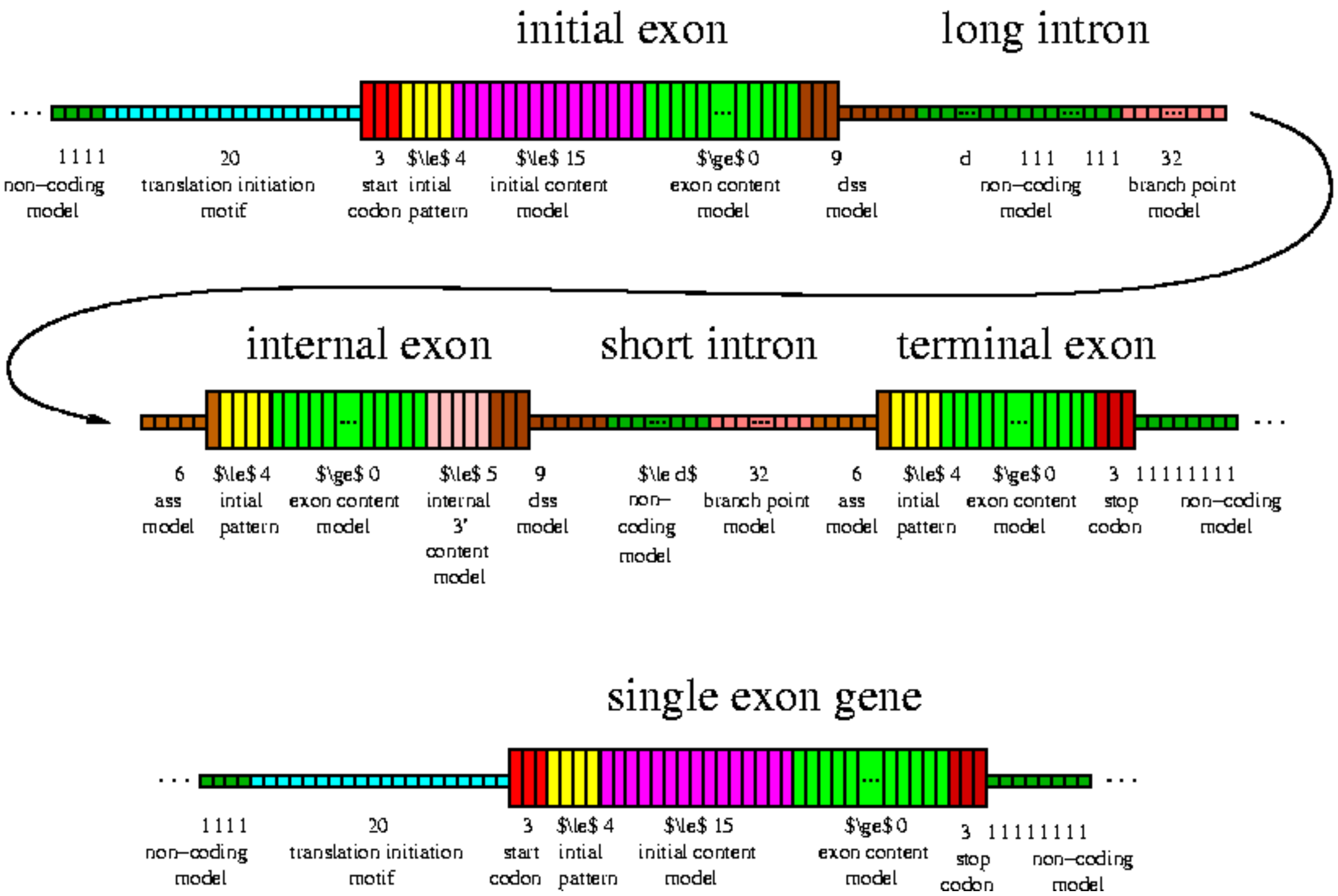
Hidden Markov Models (HMMs)

- statistical model
- each node is a state
- state can generate nucleotide sequence
- different stats generate sequences according to different probability distributions

Gene annotation

Augustus

http://augustus.gobics.de/

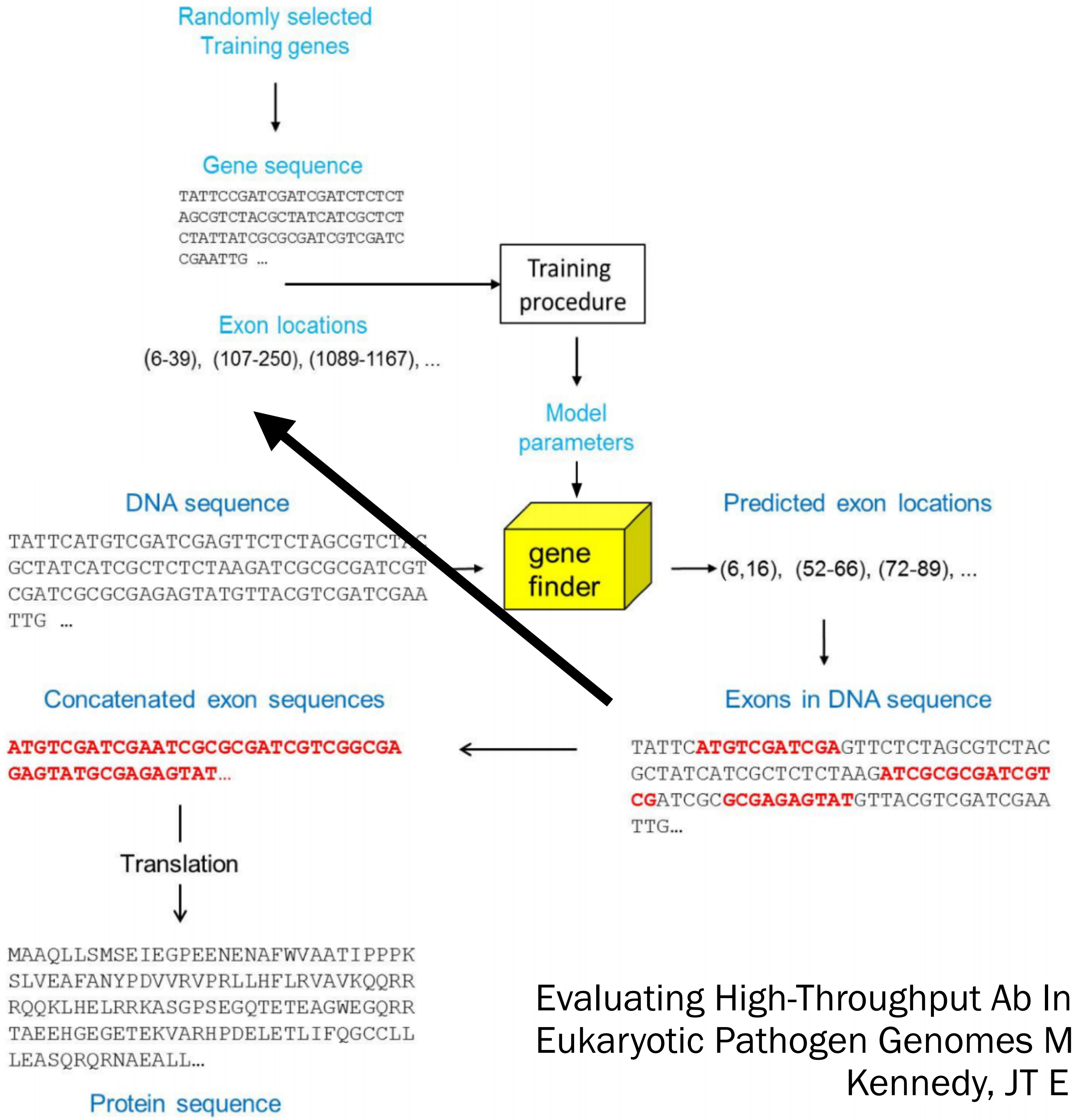


- eukaryotic gene finder
- very good performance
- requires good gene models for training
- requires extensive training / multiple rounds

Augustus

http://augustus.gobics.de/

Gene finder training



Randomly selected Training genes

Gene sequence

TATTCCGATCGATCGATCTCTCT
AGCGTCTACGCTATCATCGCTCT
CTATTATCGCGCGATCGTCGATC
CGAATTG ...

Exon locations

(6-39), (107-250), (1089-1167), ...

Training procedure

Model parameters

DNA sequence

TATTCATGTCGATCGAGTTCTCTAGCGTCTAC
GCTATCATCGCTCTCTAAGATCGCGCGATCGT
CGATCGCGCGAGAGTATGTTACGTCGATCGAA
TTG ...

gene finder

Predicted exon locations

(6,16), (52-66), (72-89), ...

Exons in DNA sequence

TATTC**ATGTCGATCGA**GTTCTCTAGCGTCTAC
GCTATCATCGCTCTCTAAG**ATCGCGCGATCGT**
CGATCGC**GCGAGAGTAT**GTTACGTCGATCGAA
TTG...

Concatenated exon sequences

ATGTCGATCGAATCGCGCGATCGTCGGCGA
GAGTATGCGAGAGTAT...

Translation

MAAQLLSMSEIEGPEENENAFWVAATIPPPK
SLVEAFANYPDVVRVPRLLHFLRVAVKQQRR
RQQKLHELRRKASGPSEGQTETEAGWEGQRR
TAEEHGEGETEKVARHPDELETLIFQGCCLL
LEASQRQRNAEALL...

Protein sequence

- often iterative training necessary
- A) start with default parameters (from closest species with available parameter set)
- B) run gene finding
- C) select best annotations
- D) train parameters on best annotations -> back to B)

Evaluating High-Throughput Ab Initio Gene Finders to Discover Proteins Encoded in Eukaryotic Pathogen Genomes Missed by Laboratory Techniques. SJ Goodswen, PJ Kennedy, JT Ellis, PLoS ONE 7(11): e50609.

Questions?