

Background and Strategy

Team 1 Gene Prediction

Genevieve Brandt, Victoria Caban, Yuntian He, Junyu Li, Yiqiuyi Liu, Yihao Ou,
Wenyi Qiu, Casey Smith, Mohit Thakur, Stephen Wist, Qinyu Yue

Content

Background

Tools and algorithms

Strategy

Content

Background

Tools and algorithms

Strategy

Our task

Background

Underlying question: What causes heteroresistance in *Klebsiella spp.*?

- Heteroresistance-a subset of a microbial population that is generally considered to be susceptible becomes resistant to certain antibiotics

Objective: given assembled genomes, predict genes for *Klebsiella spp.* that can be annotated to understand functionality

Klebsiella spp.

- Gram negative, non-motile, straight rods
- GC content: 57.35%



Klebsiella pneumoniae

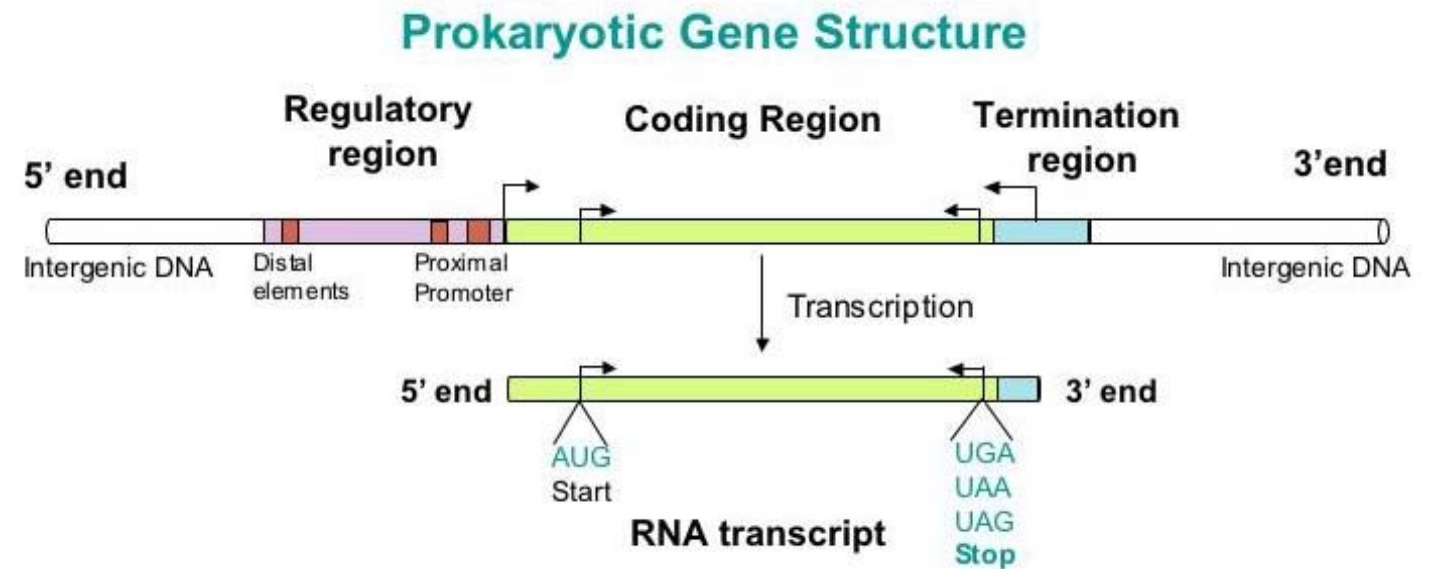
sciencesource.com

Gene Prediction

Background

- Gene prediction is the process of **identifying** the specific regions of genomic DNA that encode for **genes**
- After sequencing and assembly, gene prediction is one of the first steps in understanding the genome of a species

- General methods:
 - *Homology-based tools*
 - *Ab-initio tools*

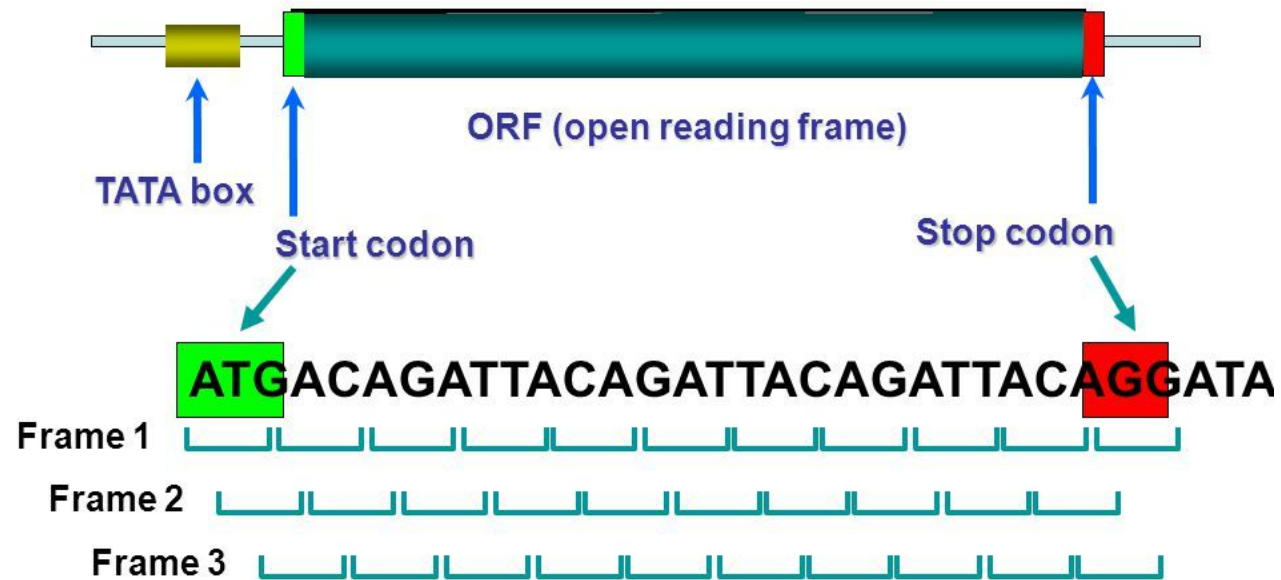


<https://iweb.langara.bc.ca/biology/mario/Biol2315notes/biol2315chap11.html>

Prokaryotic vs Eukaryotic

Background

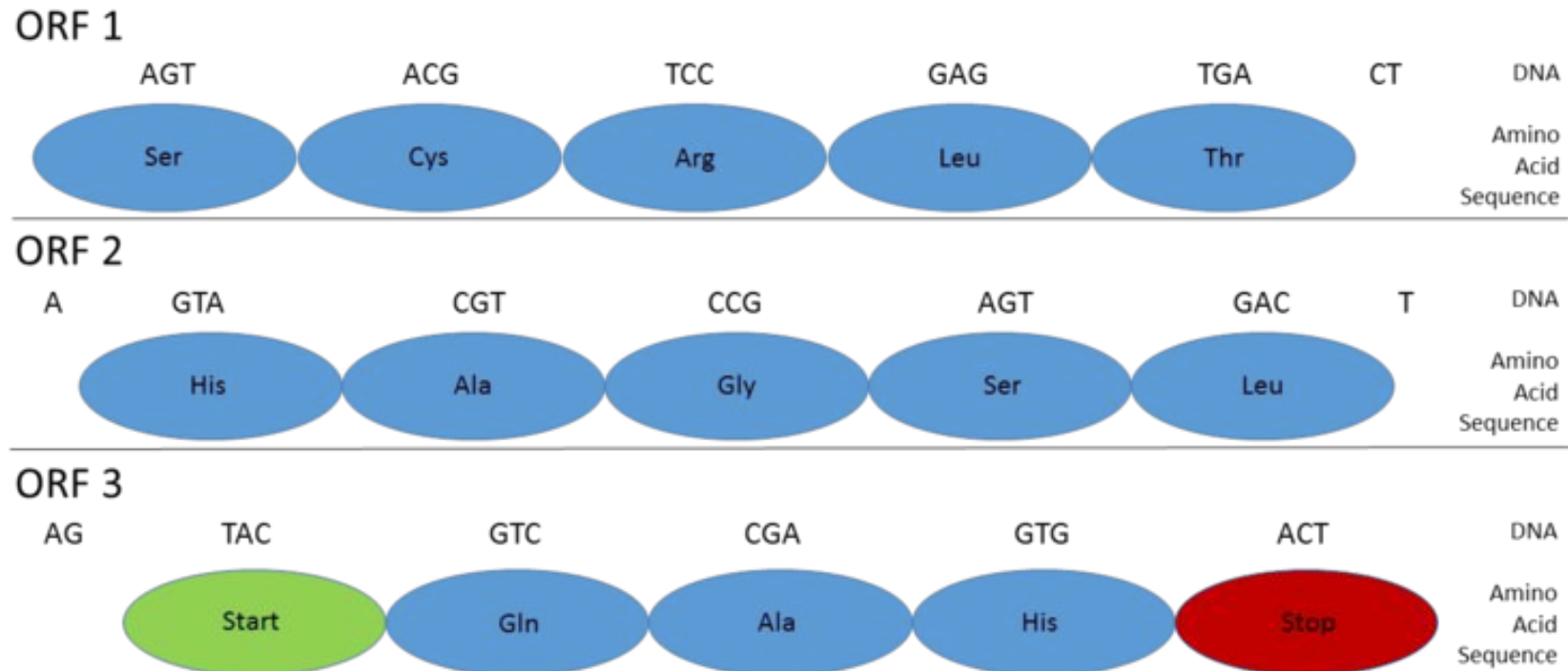
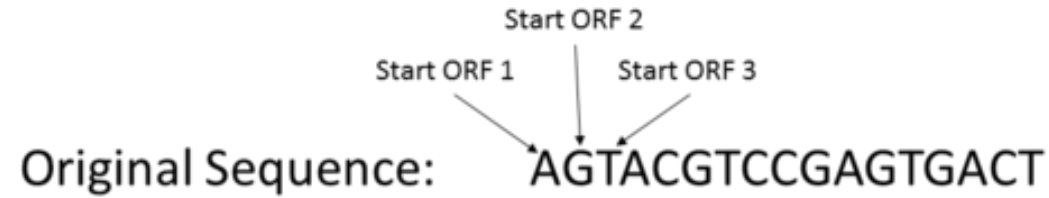
Prokaryotic Gene Structure*



- Prokaryotes: well understood promoter regions, continuous ORFs, high confidence with ab initio tools

Open Reading Frames (ORFs)

Background



GC content

Background

Gene prediction methods traditionally use stop codon frequency, and assume the GC content is about 50%

- High GC genomes contain fewer stop codons and more false ORFs
- Klebsiella has 57.35% GC content

GC Content Of Some Genomes

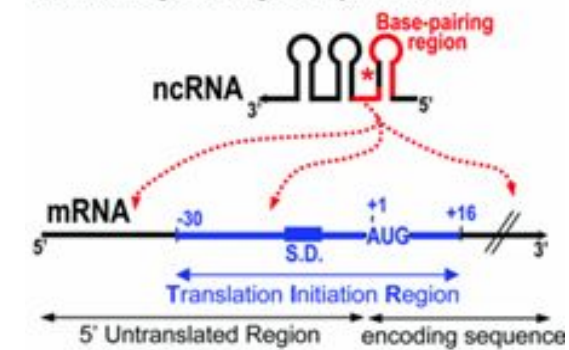
<u>Organism</u>	<u>% GC</u>
<i>Homo sapiens</i>	39.7 %
Sheep	42.4 %
Hen	42.0 %
Turtle	43.3 %
Salmon	41.2 %
Sea urchin	35.0 %
<i>E. coli</i>	51.7 %
<i>Staphylococcus aureus</i>	50.0 %
Phage λ	55.8 %
Phage T7	48.0 %

Non-coding region in Bacteria

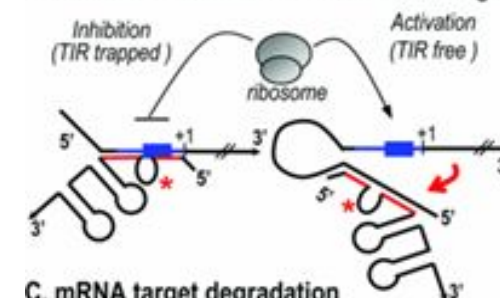
Background

- Function:
 - Improve resistance
 - Coordinate stress response
 - Regulate protein synthesis
 - Type: tRNA, rRNA, ncRNA...

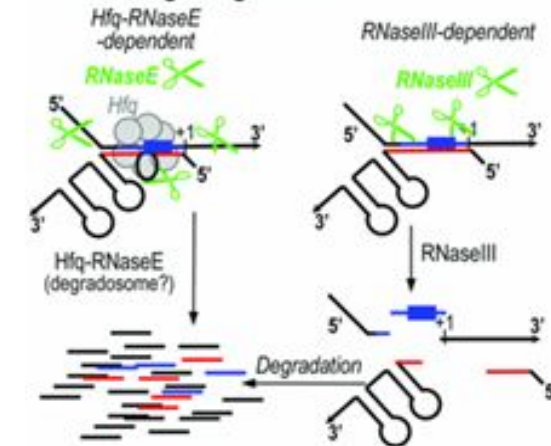
A. mRNA regions targeted by a ncRNA



B. General mechanisms of translation regulation



C. mRNA target degradation



Content

Background

Tools and algorithms

Strategy

Tools: Reference VS ab initio

Tools and algorithms

- **Reference-based method**

Target genome is searched through database for sequences that are similar to extrinsic evidence in the form of known mRNA, protein products, homologous or orthologous sequences.

- **Ab initio method**

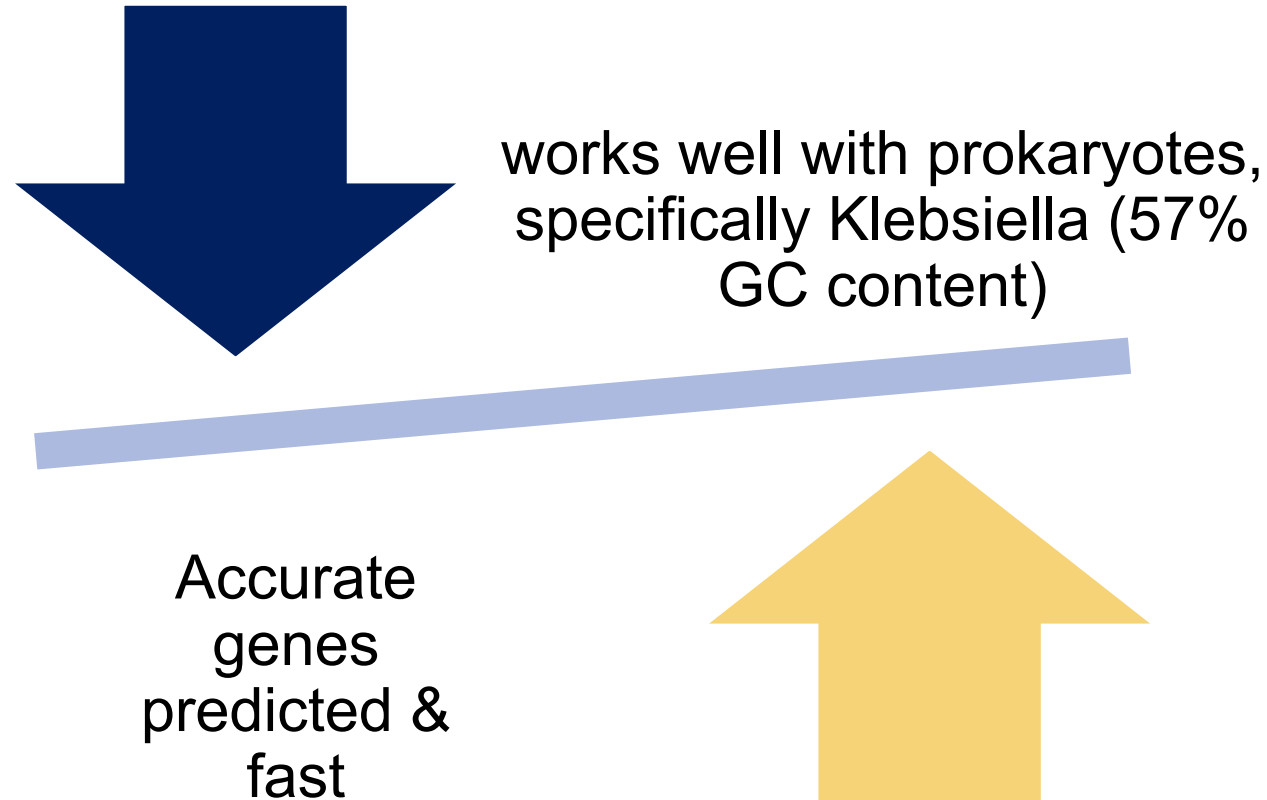
Intrinsic method based on gene content and signal detection

What makes a gene a gene?

(Promoter sequence, start/stop codon, GC ratio)

What are we looking for in the tools?

Tools and algorithms



Gene Prediction tools

Tools and algorithms

ab initio

homology
based

Tools	Algorithms
Prodigal	Dynamic programming gene finding
EasyGene	Hidden Markov Model
GeneMarkS	Hidden Markov Model
GeneMark HMM	Hidden Markov Model
Infernal	Hidden Markov Model
Glimmer	Interpolated Markov Model
RNAmmer	Markov Models
ChemGenome	Linear Discriminant Analysis
RescueNet	Synonymous codon usage
RNAScan	Covariance model
BLAST	BLAST
Aragorn	Heuristic tRNA detection

Homology-based Gene Finding

Tools and algorithms

- Local alignment tools are used to find complete or partial matches in databases
- ***Protein coding genes VS non-coding regions:***
- These tools are especially helpful to find non-coding regions that are transcribed into tRNAs or rRNAs

Algorithms, algorithms, algorithms!

Tools and algorithms

- **Prodigal**
- **Markov Models (HMM/IMM)**
- **Linear Discriminant Analysis**

Prodigal

Tools and algorithms

Prokaryotic Dynamic-Programming Gene finding Algorithm

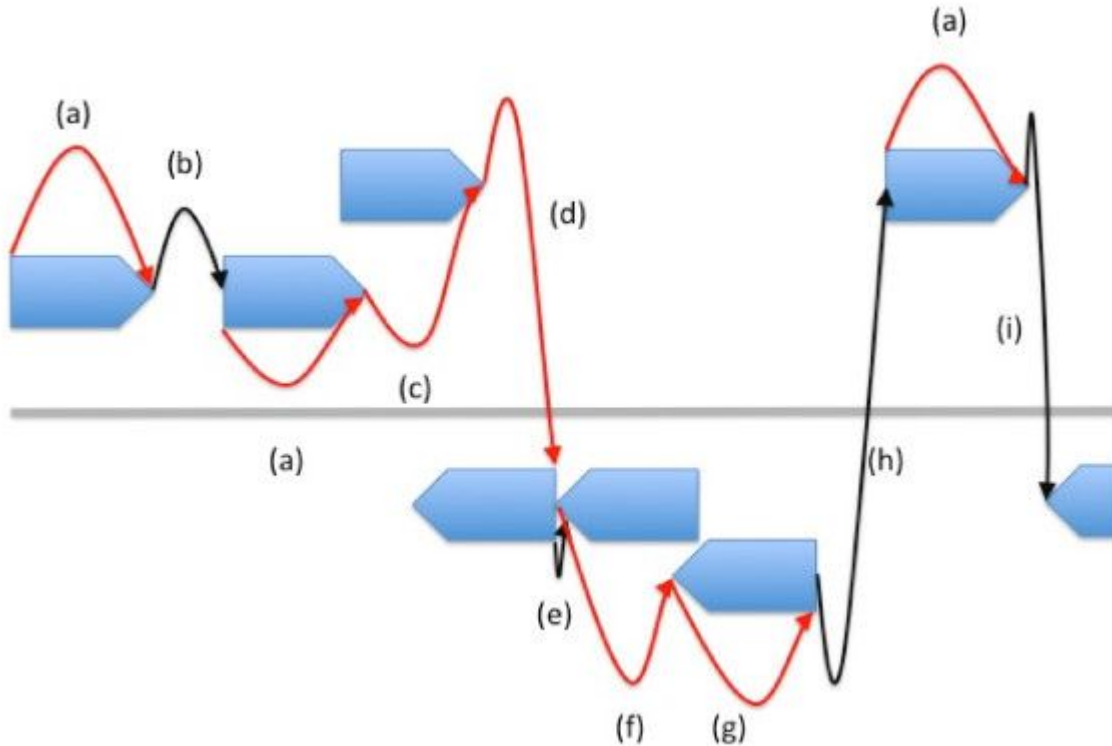
Goal:

- Attain greater sensitivity in identifying existing genes
- Predict start codon more accurately
- Minimize the number of false positive predictions

Basic steps:

- Constructing a training set for protein coding: *GC content matters!!!*
- Building log-likelihood coding statistics from the training data
- Sharpening coding scores
- Length factor to coding
- Iterative start training
- Final dynamic programming

Dynamic programming connection



Red Arrows: Gene Connections
 Black Arrows: Intergenic Connections
 Blue Pieces: Potential Genes

Dynamic Programming Connections in Prodigal

Left Node	Right Node	Connection Type	Connection Score
5' forward	3' forward	Gene	Start+coding score
3' reverse	5' reverse	Gene	Start+coding score
3' forward	5' forward	Intergenic Space	Distance modifiers
3' forward	3' reverse	Intergenic Space	Distance modifiers
5' reverse	3' reverse	Intergenic Space	Distance modifiers
5' reverse	5' forward	Intergenic Space	Distance modifiers
3' forward	3' forward	Overlapping Genes	Score of 2 nd gene
3' reverse	3' reverse	Overlapping Genes	Score of 2 nd gene
3' forward	5' reverse	Opposite Strand Overlap	Score of 2 nd gene

Gene Prediction tools

Tools and algorithms

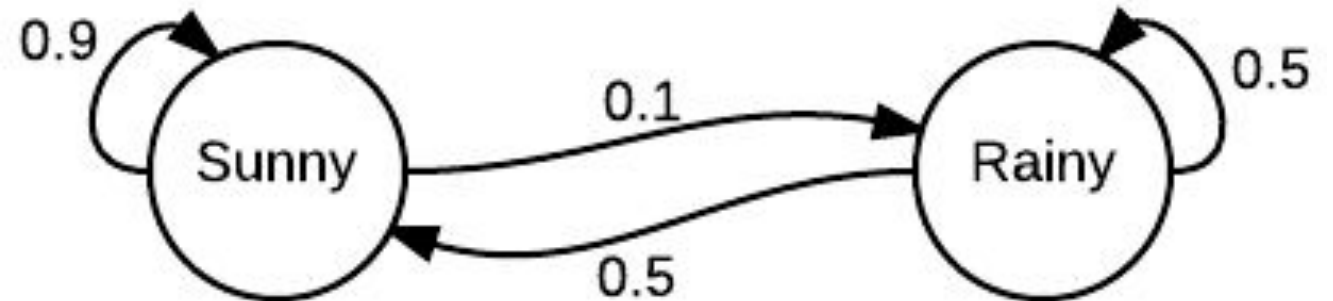
homology
based

ab initio

Tools	Algorithms
Prodigal	Dynamic programming gene finding
EasyGene	Hidden Markov Model
GeneMarkS	Hidden Markov Model
GeneMark HMM	Hidden Markov Model
Infernal	Hidden Markov Model
Glimmer	Interpolated Markov Model
RNAmmer	Markov Models
ChemGenome	Linear Discriminant Analysis
RescueNet	Synonymous codon usage
BLAST	BLAST
Aragorn	Heuristic tRNA detection

Markov Model

Tools and algorithms

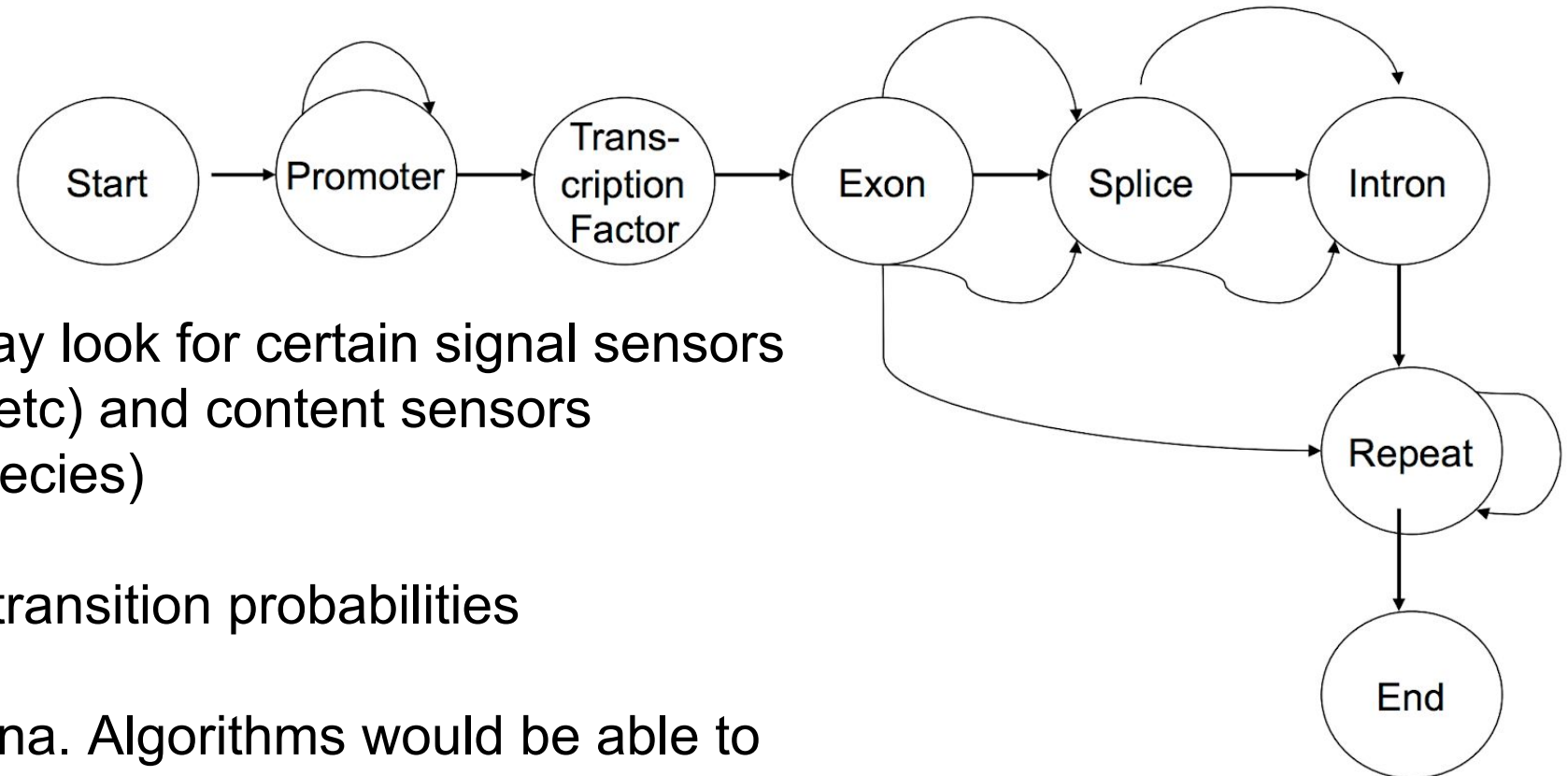


- Simplest form: certain states and the probability to switch between states
- Problem is we don't **know** states we're in - all we have is the sequence of nucleotides

http://ece.drexel.edu/gailr/ECE-S690-503/markov_models.ppt.pdf

Hidden Markov Model

Tools and algorithms



- Ab initio algorithm may look for certain signal sensors (splice sites, codons, etc) and content sensors (patterns unique to species)
- Start, emission, and transition probabilities
- General flow of the dna. Algorithms would be able to predict which state is most likely to come next and when the current state switches.

Interpolated Markov Model by GLIMMER

Tools and algorithms

- Glimmer3
 - Optimize order of markov model (1st through 8th), i.e. amount of context based on sequence information

Gene Prediction tools

Tools and algorithms

homology
based

ab initio

Tools	Algorithms
Prodigal	Dynamic programming gene finding
EasyGene	Hidden Markov Model
GeneMarkS	Hidden Markov Model
GeneMark HMM	Hidden Markov Model
Infernal	Hidden Markov Model
Glimmer	Interpolated Markov Model
RNAmmer	Markov Models
ChemGenome	Linear Discriminant Analysis
RescueNet	Synonymous codon usage
BLAST	BLAST
Aragorn	Heuristic tRNA detection

Linear Discriminant Analysis

Tools and Algorithms

- Linear classification
- Tool: ChemGenome
- Open reading frames → gene vs. non-gene
- Physicochemical features:
 - Watson-Crick hydrogen-bonding energy
 - Base-pair stacking energy
 - Propensity for protein-nucleic acid interactions
- Knowledge-based screening

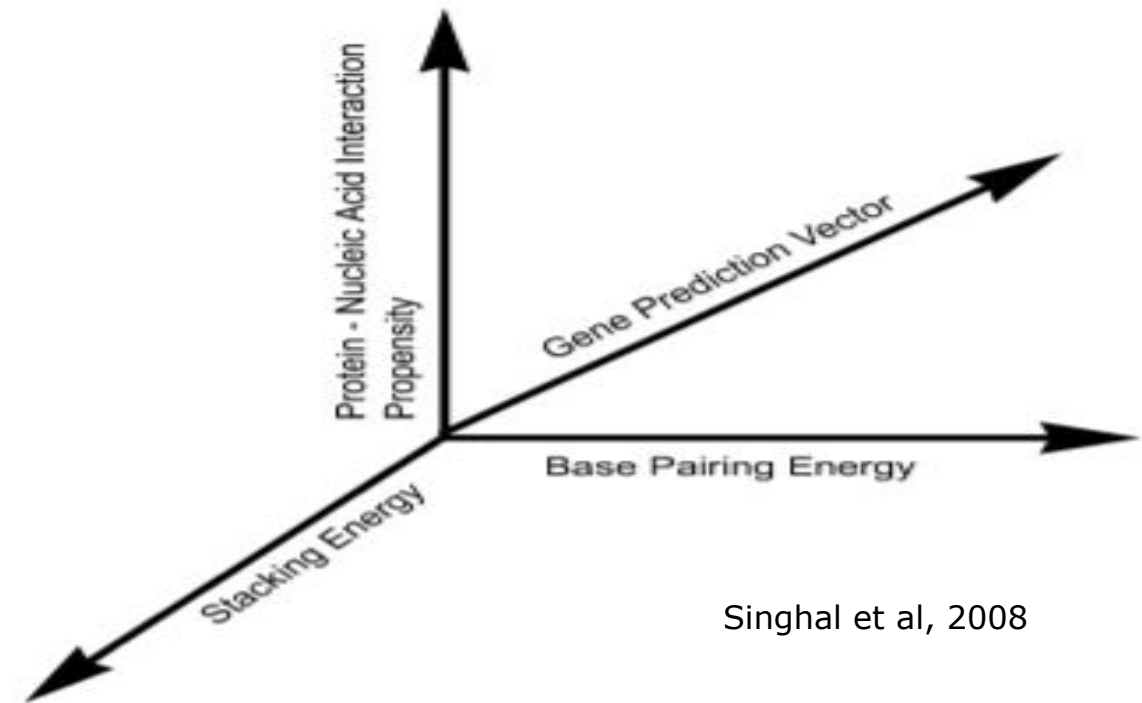
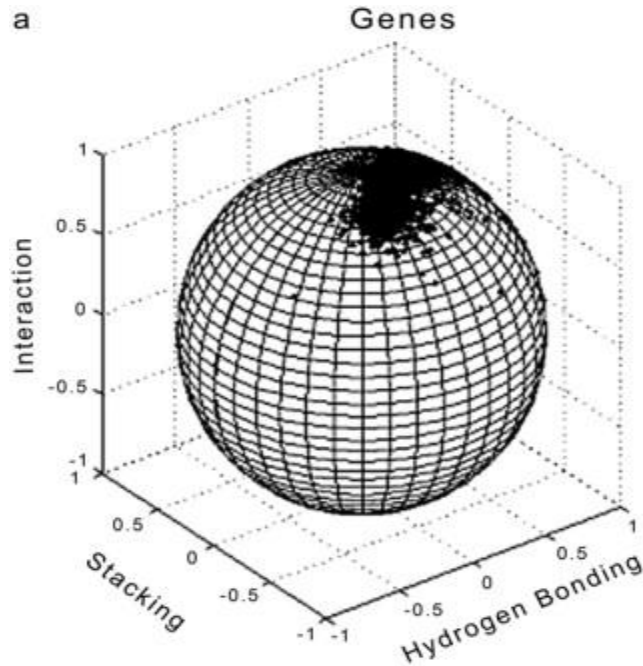


TABLE 1 The x (hydrogen-bonding energy), y (stacking energy), and z (protein-nucleic acid interaction propensity parameter) values assigned for each of the 64 codons

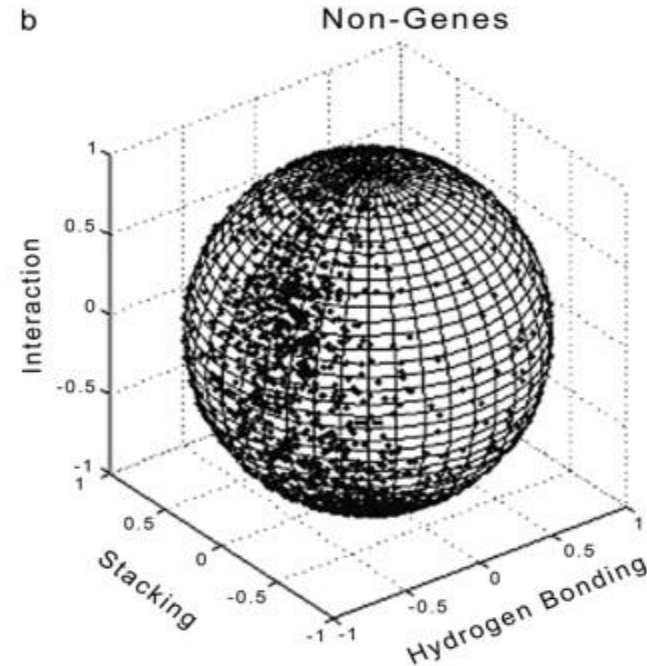
Codon	x	y	z	Codon	x	y	z
CCC	-1.0	0.97	-1	TCC	-0.85	0.66	-1
CCG	-0.85	0.14	1	TCG	-0.41	-0.10	-1
CCT	-0.03	1.00	1	TCT	-0.15	0.74	-1
CCA	-0.02	0.81	-1	TCA	-0.18	0.23	-1
CGC	-0.98	-1.00	-1	TGC	-0.49	-0.38	-1
CGG	-0.85	0.14	1	TGG	-0.02	0.81	-1
CGT	-0.30	-0.71	1	TGT	-0.13	0.07	-1
CGA	-0.41	-0.10	-1	TGA	-0.18	0.23	-1

Linear Discriminant Analysis

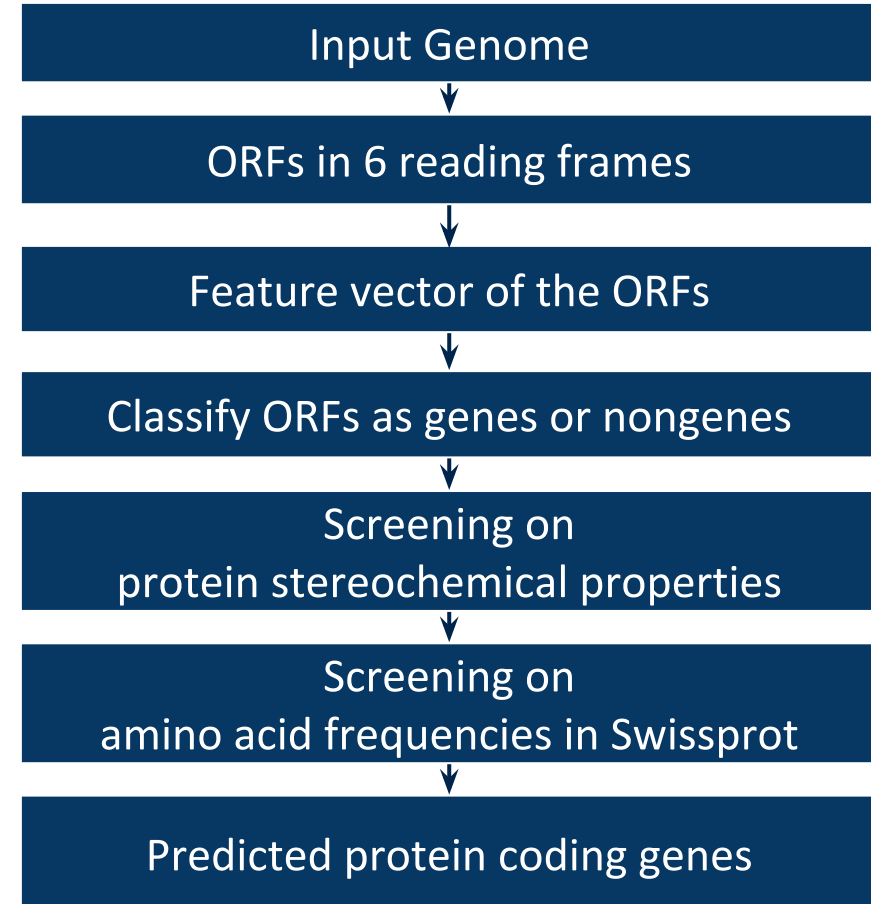
Tools and Algorithms



854 experimentally
verified *E.coli* genes



Corresponding non-gene
from frameshift



Gene Prediction tools

Tools and algorithms

homology
based

ab initio

Tools	Algorithms
Prodigal	Dynamic programming gene finding
EasyGene	Hidden Markov Model
GeneMarkS	Hidden Markov Model
GeneMark HMM	Hidden Markov Model
Infernal	Hidden Markov Model
Glimmer	Interpolated Markov Model
RNAmmer	Markov Models
ChemGenome	Linear Discriminant Analysis
RescueNet	Synonymous codon usage
BLAST	BLAST
Aragorn	Heuristic tRNA detection

Content

Background

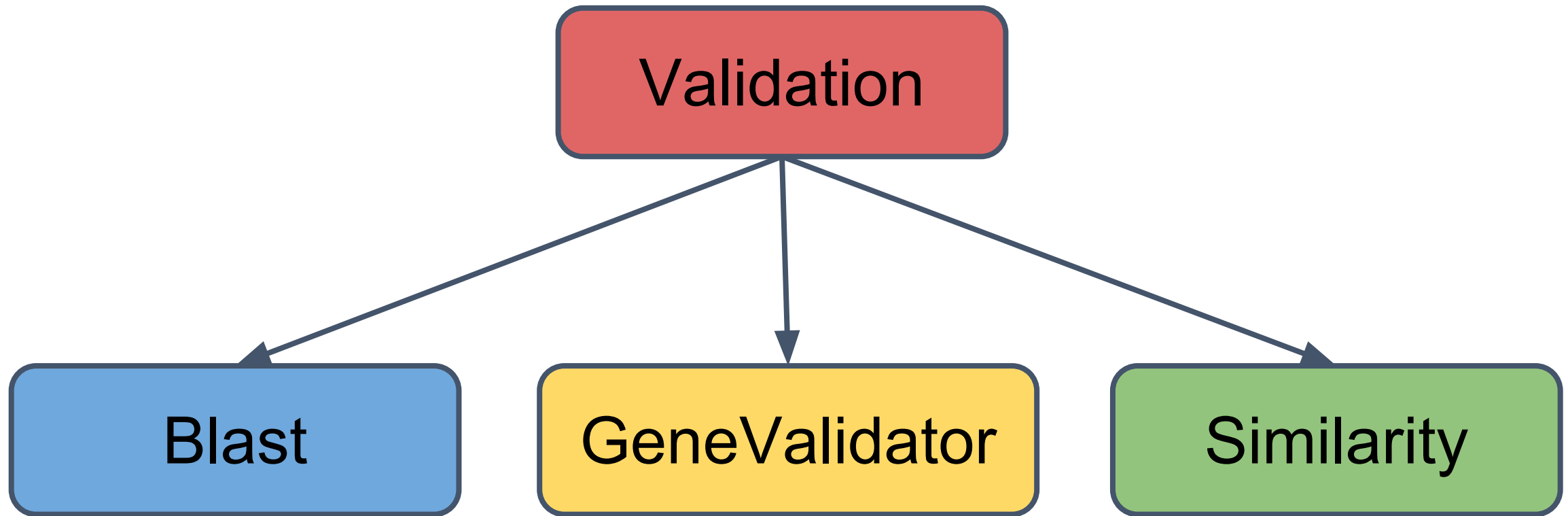
Tools and algorithms

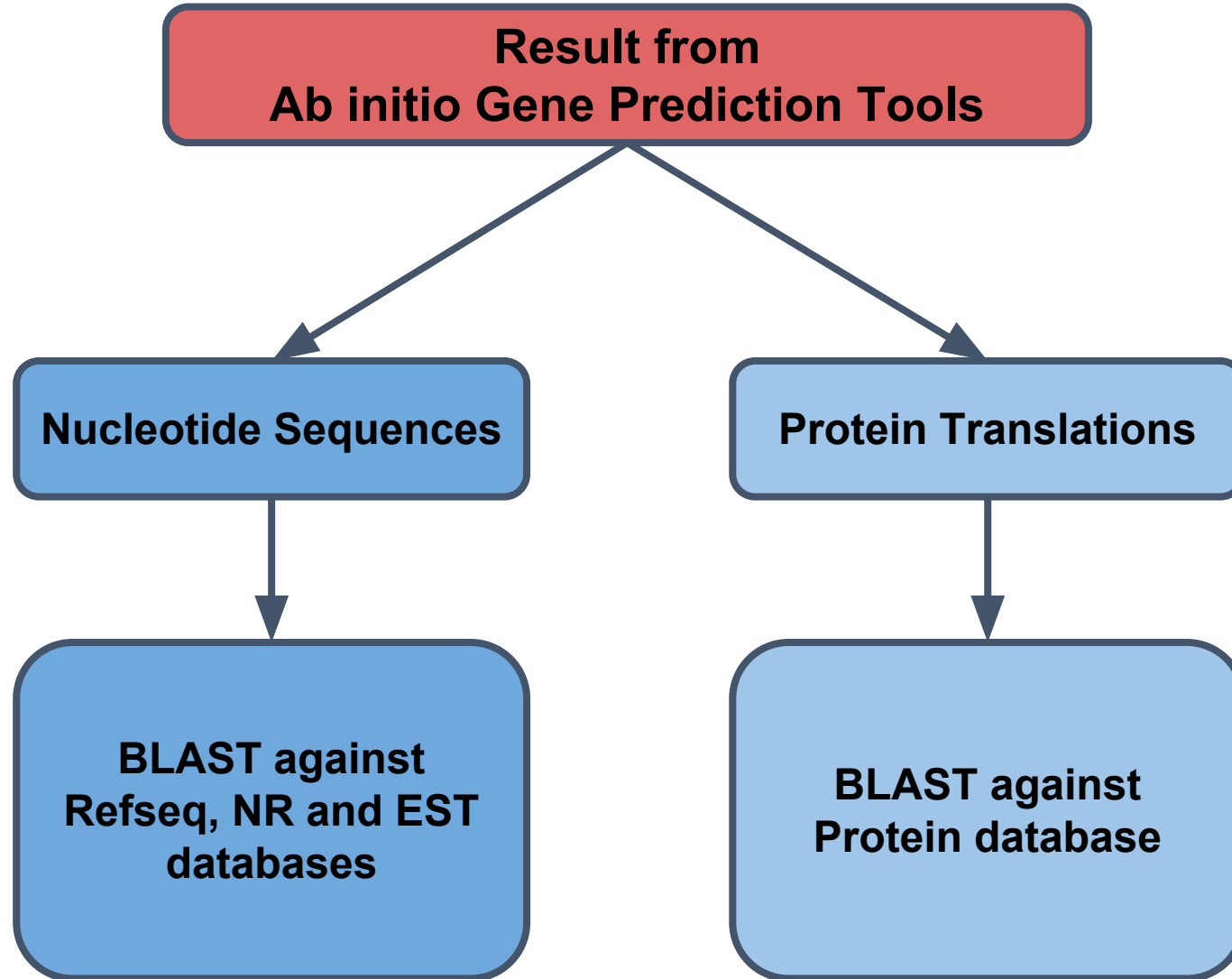
Strategy

ARE THEY REALLY GENES?

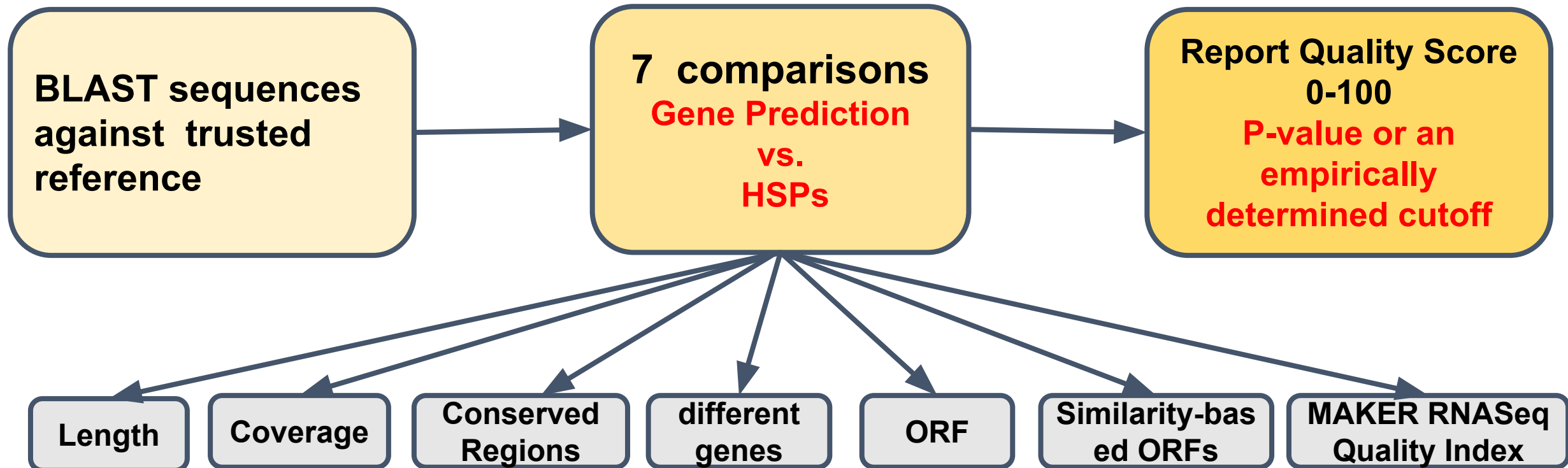
Validation

Strategy



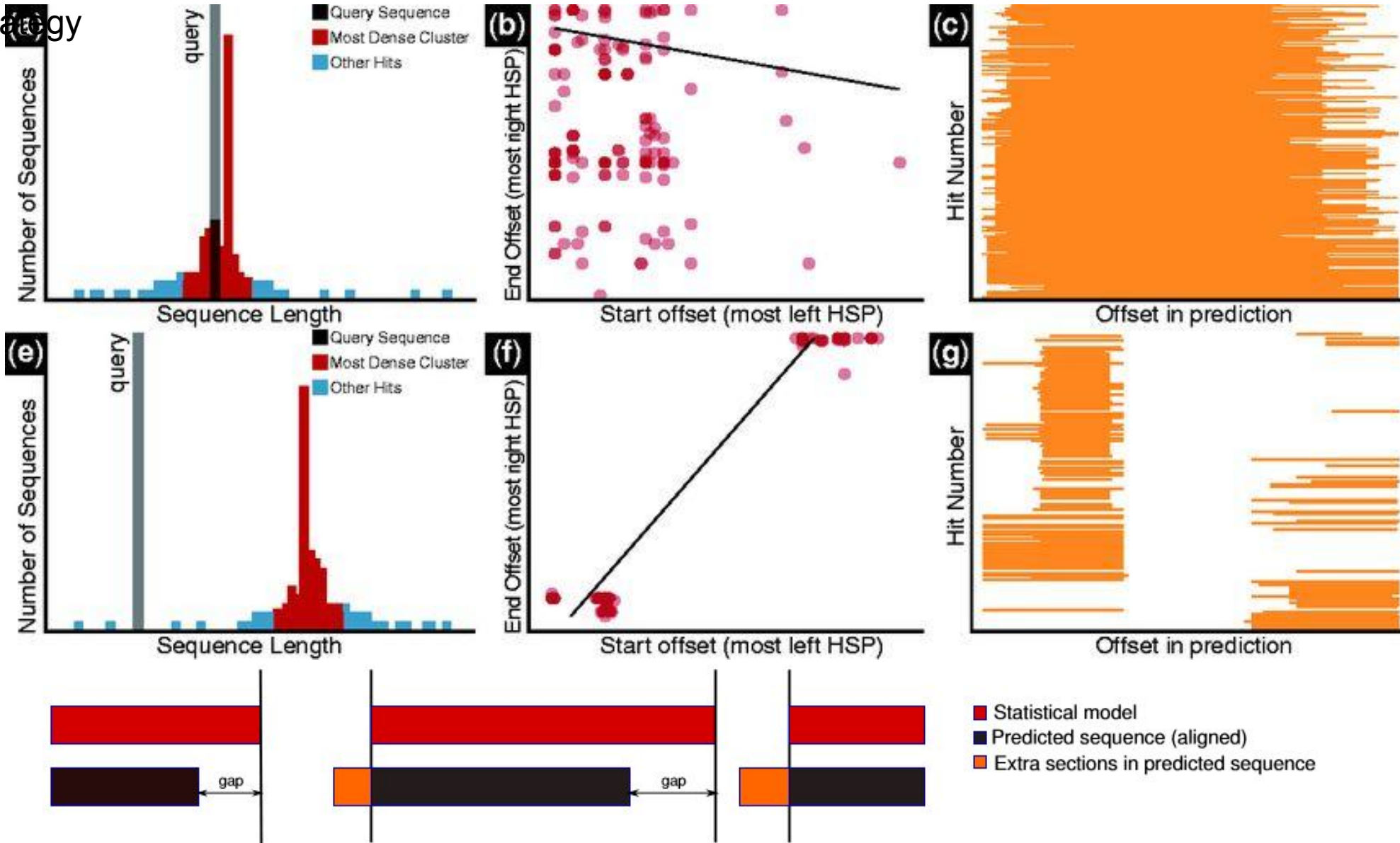


How does it work?



GeneValidator

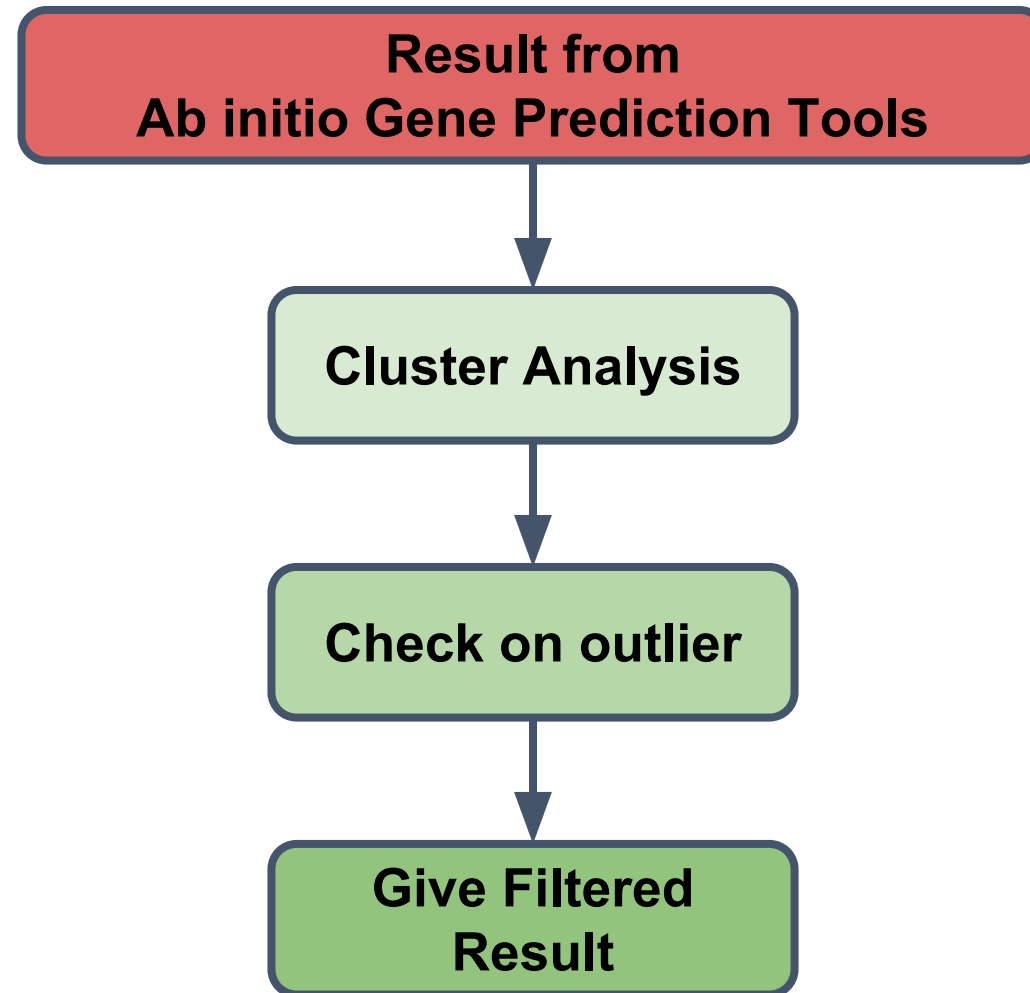
Strategy



Problem!
**Both of these methods rely on
reference**

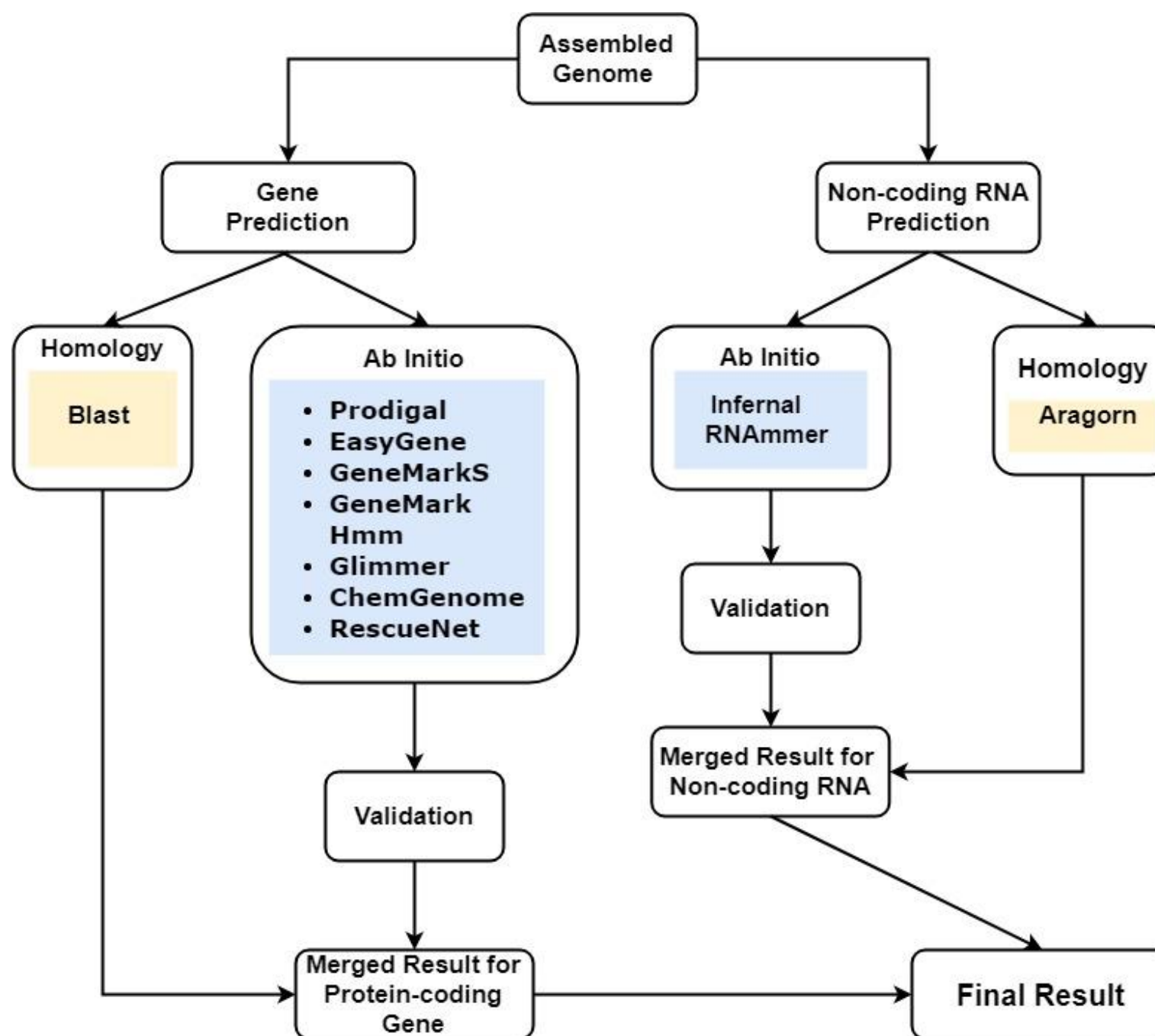
Similarity

Strategy



Workflow

Strategy



References

<https://ghr.nlm.nih.gov/primer/basics/gene>

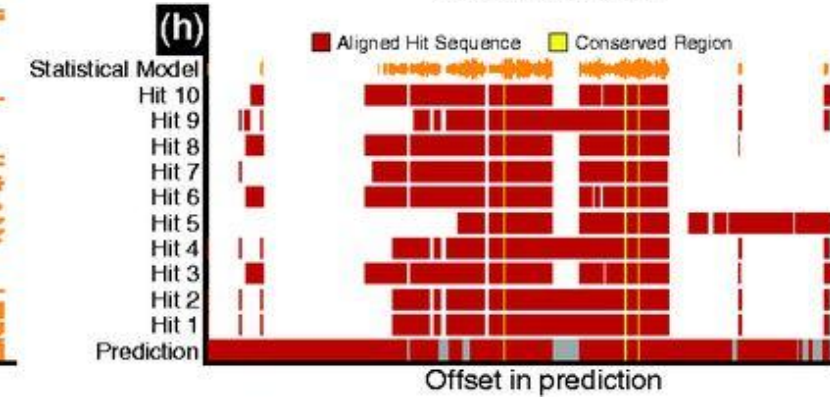
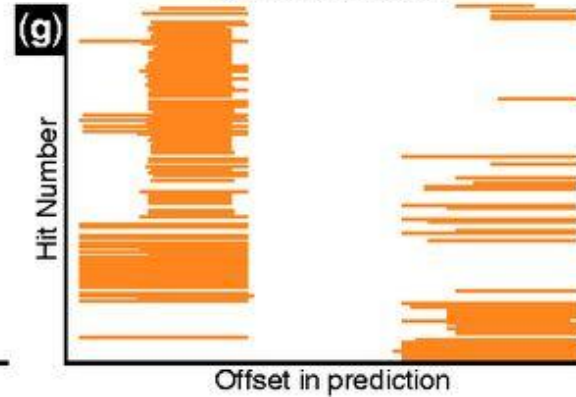
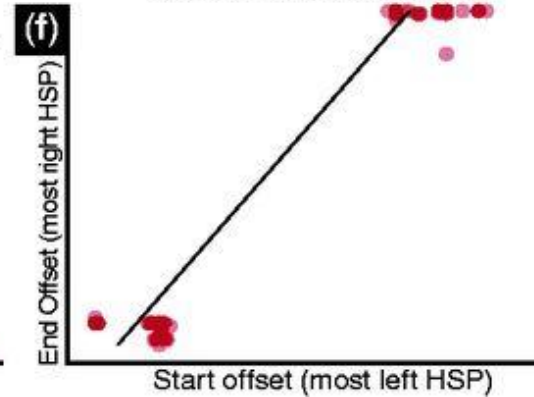
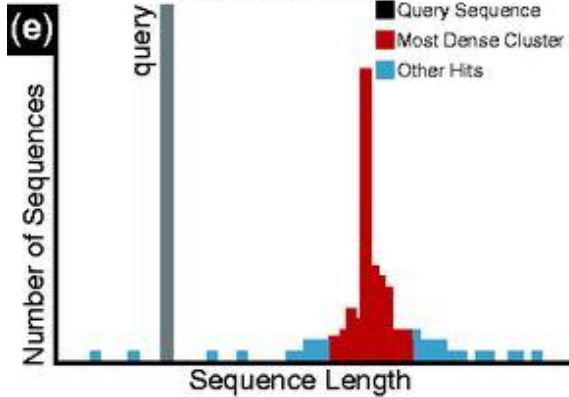
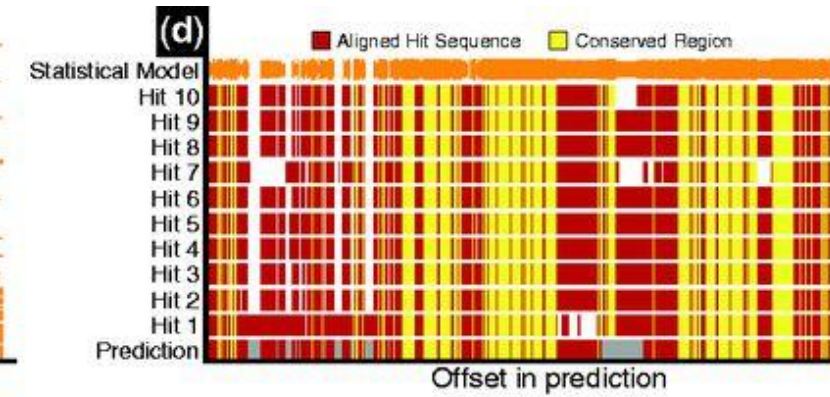
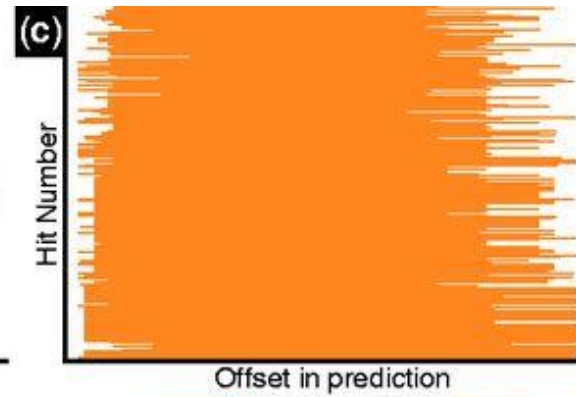
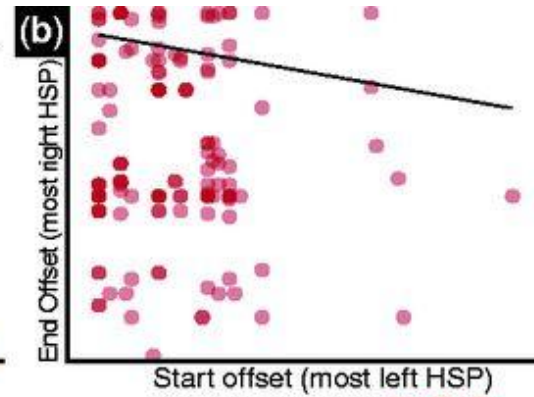
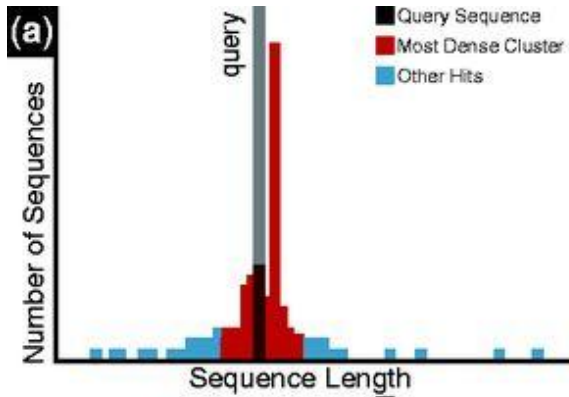
<https://www.biostat.wisc.edu/bmi776/spring-15/lectures/IMMs.pdf>

http://ece.drexel.edu/gailr/ECE-S690-503/markov_models.ppt.pdf

<http://onlinelibrary.wiley.com/doi/10.1042/BC20070137/full#footer-citing>

<https://iweb.langara.bc.ca/biology/mario/Biol2315notes/biol2315chap11.html>

Questions?



Homology-based Gene Prediction (Finding)

Example: BLAST, Diamond, Aragon

- Use local alignment tools to find complete or partial matches in:
 - Expressed sequence tags (short sub-sequences of a cDNA sequence)
 - mRNA
 - Protein products (non-redundant)
 - Homologous or orthologous sequences
- Databases
 - RefSeq reference genomes
 - Prokaryotic RefSeq genomes
 - RefSeq non-redundant proteins
 - Expressed sequence tags database
 - Non-coding RNA database
- Horizontal gene transfer
- Depend on the contents and accuracy of the database

RNA Prediction Tools

- RNAScan 2.0
 - Covariance model, replacing COVE (still available as backward compatible option)
 - Better recognition of atypical tRNA, high sensitivity / specificity
- RNAmmer
 - predicts ribosomal RNA genes in full genome sequences by utilising two levels of Hidden Markov Models (spotter model--full model)
 - Fast