Gene Prediction

Background and Strategy

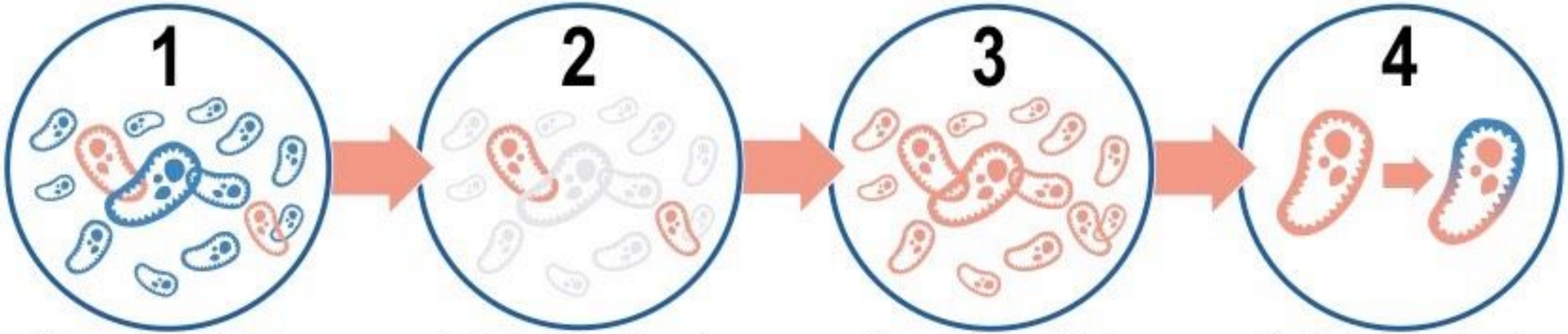
Team II

Beatriz E Saldana, Parisa Y Zowj, Ayush Semwal, Siu Lung Ng, Sini Nagpal,
Sarthak Sharma, Rong Jin, Jiani Long, Qi Zhang

Introduction - Project

- Initial data: 262 Klebsiella un-assembled genomes of unknown species

- Project goal: Use genetic determinants of antibiotic resistance to further understand heteroresistance



Introduction - Project

- From raw reads to biological knowledge:
    - Genome assembly
    - Genome annotation
    - Data analysis

Introduction - Project

- From raw reads to biological knowledge:
    - Genome assembly
    - **Genome annotation**
        - **Gene Prediction**
        - **Functional Annotation**
    - Data analysis

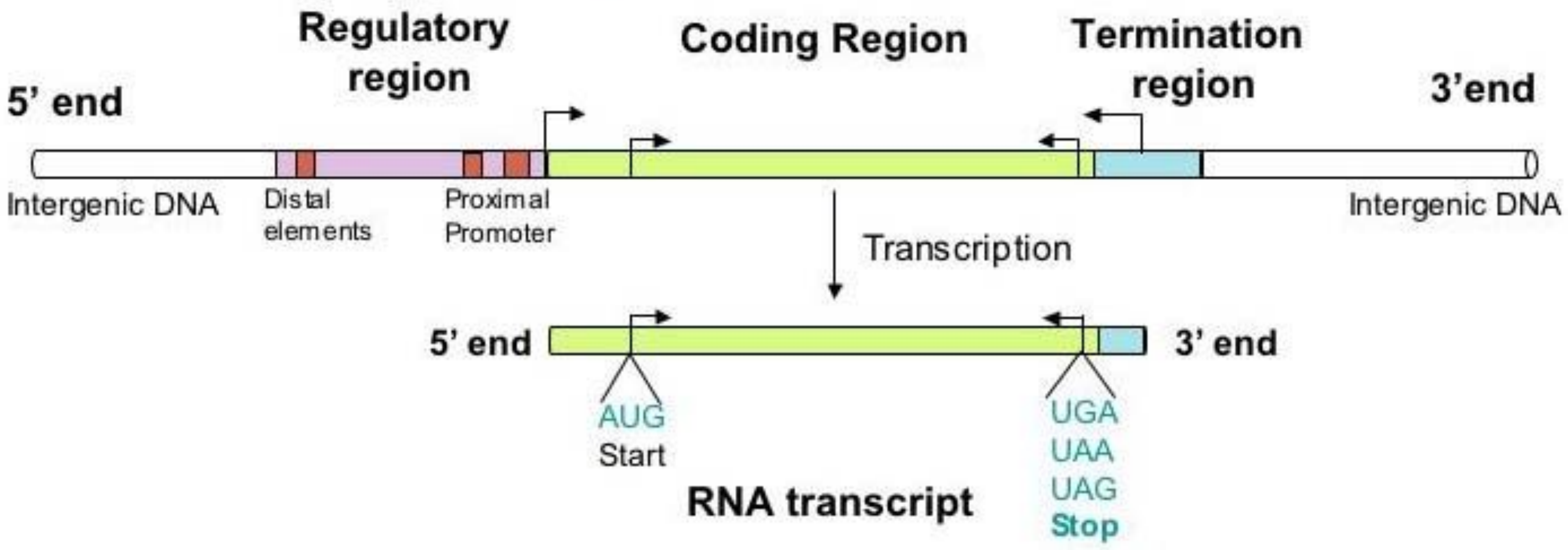
Introduction - Project

- From raw reads to biological knowledge:
    - Genome assembly
    - Genome annotation
        - **Gene Prediction**
        - Functional Annotation
    - Data analysis

Introduction - Gene Prediction

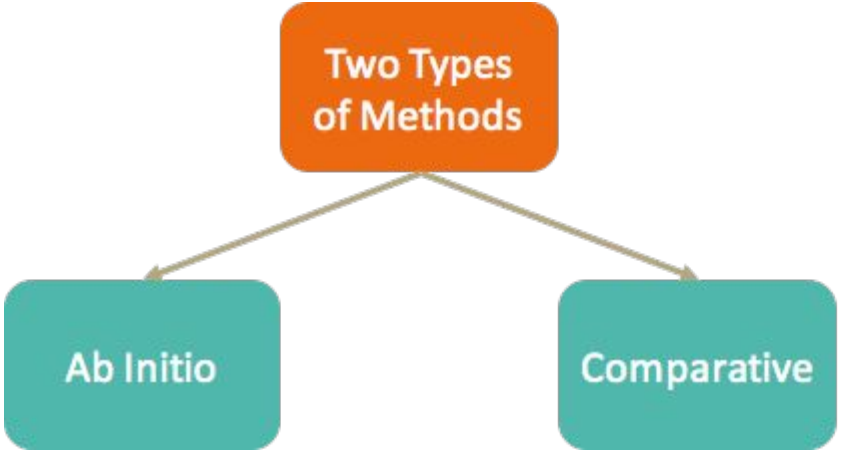
What is Gene Prediction?

- The process of finding regions of DNA that encode genes



Introduction - Gene Prediction: Our Plan

- Divide into three groups
    - Comparative / Similarity-Based
    - Ab Initio
    - Non Coding RNA
- Each group will:
    - Explore their specific task
    - Find tools
        - Specific to our data
    - Test the tools
    - Compare the tools


Two Types of Methods

Ab Initio

Comparative

Comparative Approach

Description, Tools, and Strategy

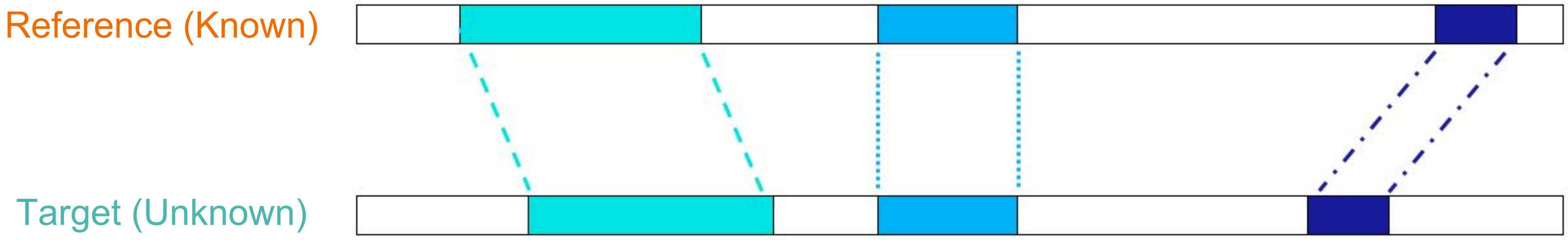
Comparative Methods

- Comparative or *similarity* based gene prediction
- Using **Known Genes** to predict **New Genes**

- Motivation:
    - Recently, the number of sequenced genomes has increased drastically
    - 99% of genes have homologous partner
    - 80% have orthologous partner
    - 85 % identity (protein coding DNA) versus 69 % identity (intronic DNA)

Problem

Given a known gene and an unannotated genome sequence, find a set of substrings in the genomic sequence whose concatenation best matches the known gene



Comparing genes in two genomes

- Since klebsiella is a prokaryote (does not have introns)
- We won't have splice alignment problem

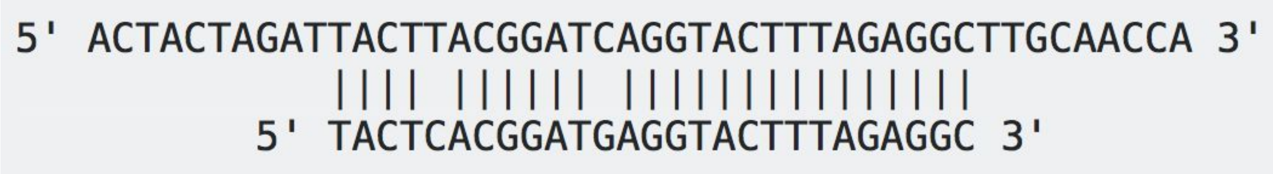
Sequence alignment

- Sequence alignment is a way of arranging the sequences to identify regions of similarity that may be results of:
    - Functional
    - Structural
    - Evolutionary relationships

- Two methods based on similarity research are:
    - Local alignment
    - Global alignment

Local Alignment

- Try to match your query with a <u>substring</u> of your reference
- Smith–Waterman algorithm

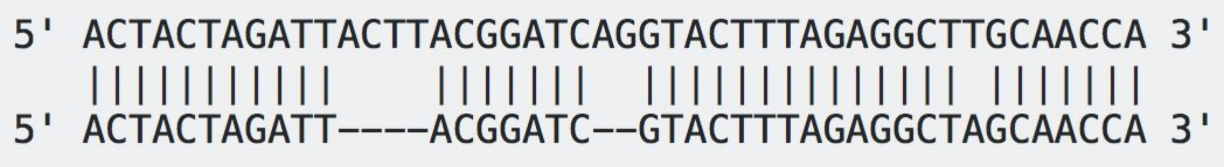


```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| |||||||||||||||
          5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

2 mismatch , 0 gaps

Global Alignment

- Forces the alignment to span the entire length of all query sequences
- Most useful when the sequences are similar and roughly equal size
- May end up with a lot of gaps
- *Needleman–Wunsch algorithm*
- Based on Dynamic programing



```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

1 mismatch , 2 gaps of length 4 and 2

Strategy

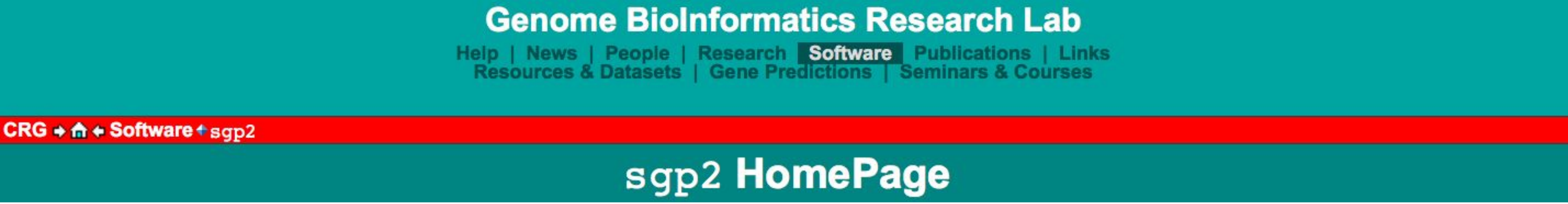
Comparative Based Exploratory Research

Find species of the samples

BLAST

Testing Tools

SPG2
TwinScan
GenomeScan



Genome BioInformatics Research Lab

Help | News | People | Research | Software | Publications | Links
Resources & Datasets | Gene Predictions | Seminars & Courses

CRG → ⌂ ← Software ← sgp2

sgp2 HomePage

- Input format is FastA
- Output format is geneid, gff, XML
- It takes one DNA sequence (target) and several DNA sequences (references) which have partial Tblastx matches to it (i.e. protein level)
- Very efficient in terms of speed and memory usage

Twinscan/N-SCAN

The Brent Lab

Washington University in St. Louis

- Begins with local alignments between a unknown genome and a database of reference sequences
- Twinscan is currently available for Mammals, Caenorhabditis (worm), Dicot plants, and Cryptococci

- Predicting the locations and exon-intron structures of genes in genomic sequences
- Input:
  - Unknown DNA sequence
  - Reference sequence/s (as proteins) in FastA format
- Predicts gene structure which corresponds to maximum probability conditional on similarity information

Comparative methods Pros / Cons

- Fast implementation
- High accuracy
- Efficient in terms of memory usage

- Reference dependent
- Does not guarantee optimal alignment
- Returns only one best alignment

Ab Initio

Description, Tools, and Strategy

Ab-Initio Methods

Predict gene based on given sequence alone

Rely on two major features:

1. Gene signals (start and stop codon, intron splice signals, codon structure, etc.)
2. Statistical description of coding regions.

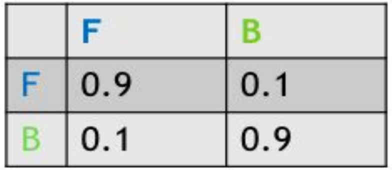
Hidden Markov Model (HMM)

- Machine with k **hidden states (F and B)** proceeding in a sequence of steps
- In each step emission of a symbol (H or T) while being in one of its hidden states
- In a certain state makes two decisions:
  1. Which symbol to emit
  2. Which hidden state to move next

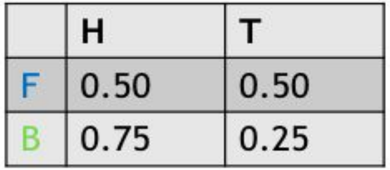
Hidden Markov Model (HMM)

Transition : changing from hidden state l to hidden state k

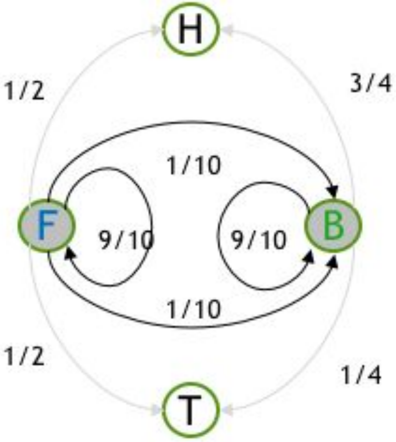
Emission : emission of symbol when the HMM is in state k



	F	B
F	0.9	0.1
B	0.1	0.9



	H	T
F	0.50	0.50
B	0.75	0.25



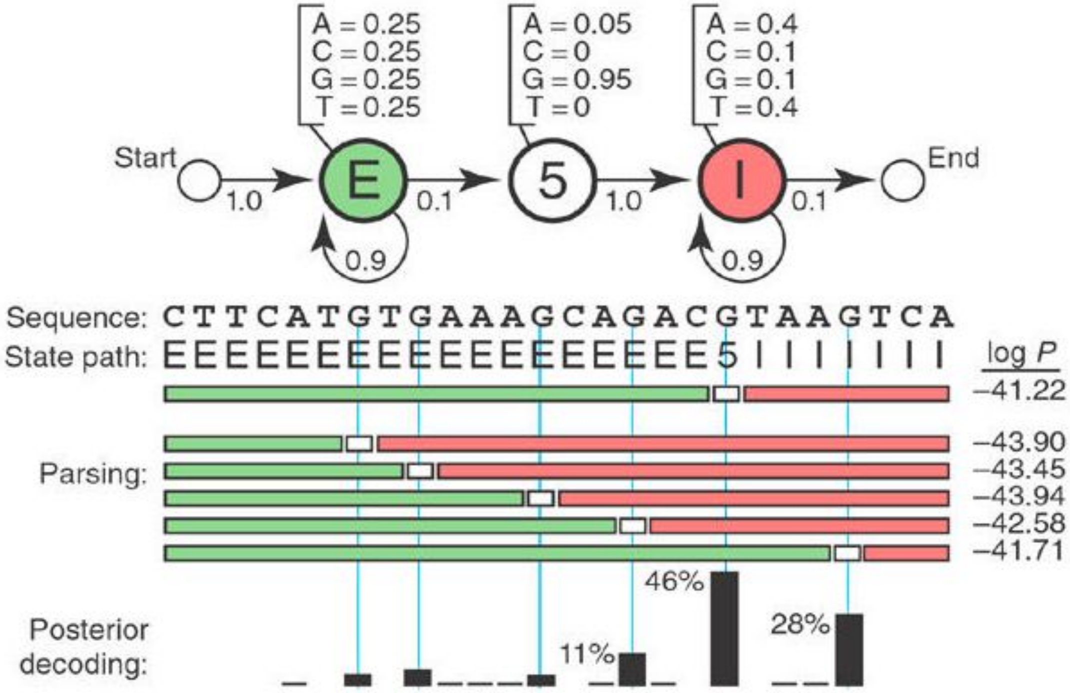
Central Issues in HMM

Evaluation Problem: Given, sequence of visible symbols V^T, what is the probability that this V^T was generated by Θ (HMM)? ($P(V^T | \Theta)$ to be calculated)

Decoding problem: What's the most likely sequence of hidden states which led to the generation of V^T?

Learning Problem: Using large number of training sequences, estimate transition probabilities (both – between hidden states as well as emission symbols)

Gene prediction using HMM



Gene prediction using HMM

- **kth order model -** in which the conditional probability of a particular sequence position depends on *k* previous positions.
- A zero-order Markov model assumes each base occurs independently with a given probability.
- A second-order model looks at the preceding two bases to determine which base follows, which is more characteristic of codons in a coding sequence.
- the higher the order of a Markov model, the more accurately it can predict a gene.

Gene prediction using HMM

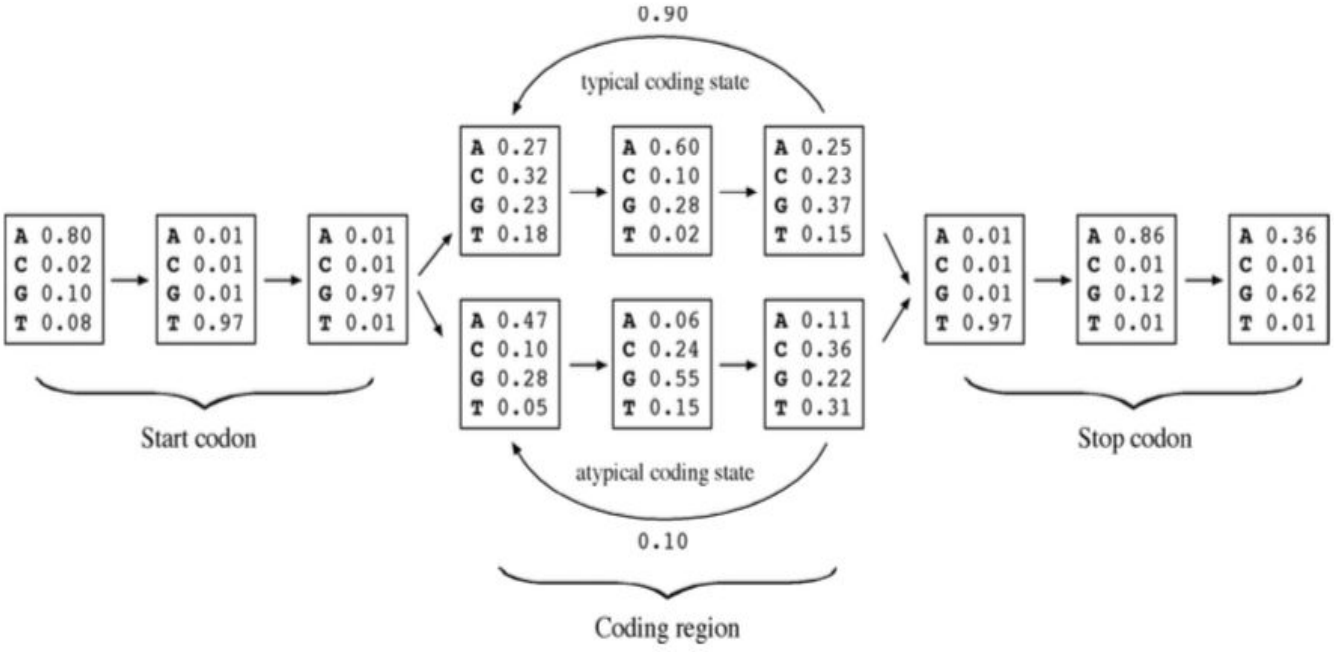


Fig. *A Simplified second-order HMM for prokaryotic gene prediction*

- More effective Markov models built in sets of three nucleotides, describing non-random distributions of trimers or hexamers, and so on.
- The parameters of a Markov model have to be trained

Gene prediction using HMM

- Statistical analyses have shown that pairs of codons tend to correlate.
- Frequency of six unique nucleotides appearing together in a coding region is much higher than by random chance.
- Therefore, a fifth-order Markov model, can detect nucleotide correlations found in coding regions more accurately.
- Drawback – method’s efficacy is limited (in case of short gene sequences – not enough hexamers)
- Overcome using **Interpolated Markov Model (IMM).**

GeneMark

- A suite of gene prediction programs based on the fifth-order HMMs
- The main program – **GeneMark.hmm** – trained on a number of complete microbial genomes
- If the sequence to be predicted is from a non-listed organism, the most closely related organism can be chosen as the basis for computation.
- If new organism – **GeneMarkS** can be used (self-trained program). Longer than 50kb sequences to be provided.
- If shorter sequences – GeneMark heuristic program can be used with loss of some accuracy.

Glimmer

- **G**ene **L**ocator and **I**nterpolated **M**arkov **M**odel**er**
- Developed at 'The Institute of Genomic Research (TIGR)'
- UNIX program that uses the IMM algorithm to predict potential coding regions
- Two Steps –
  1. Model Building
  2. Computation

Gene Prediction Using Log-likelihood

A Simplistic Explanation:

- For a random sequence $N_1N_2N_3N_4N_5N_6N_7$, *$P(N_i)$ = ¼ where $N_i \in \{A, T, C, G\}$*
- For a putative coding sequence, assume the following probabilities:

	1	2	3	4	5	6	7
A	0.3	0.6	0.1	0.00	0.00	0.6	0.7
C	0.2	0.2	0.1	0.00	0.00	0.2	0.1
G	0.1	0.1	0.7	1.00	0.00	0.1	0.1
T	0.4	0.1	0.1	0.00	1.00	0.1	0.1

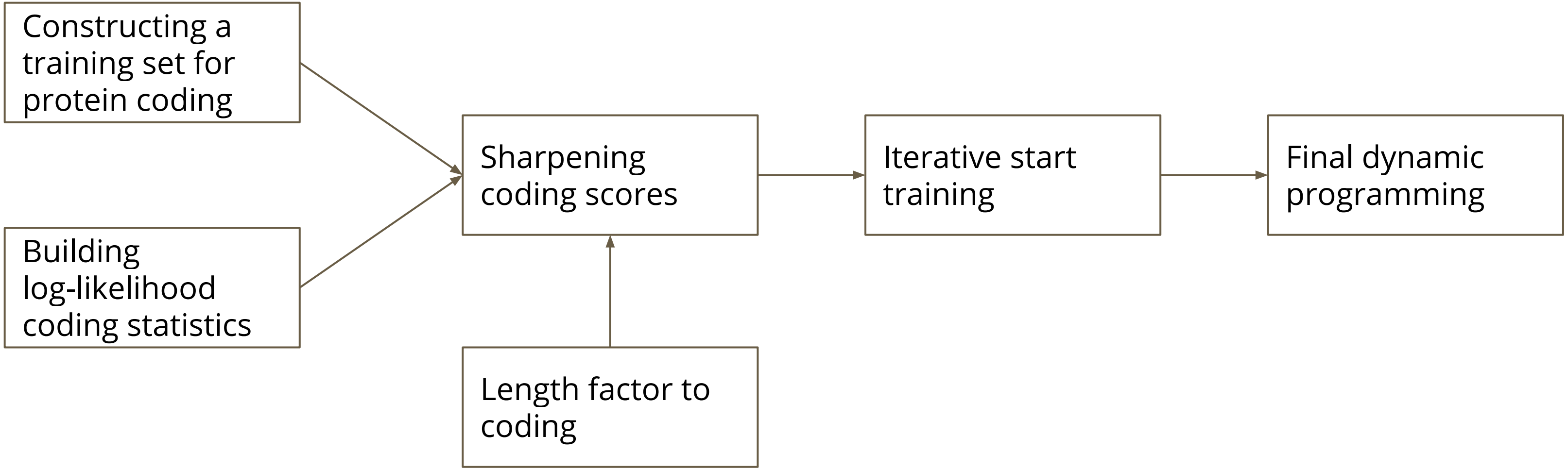
- *P(random sequence) = $(¼)^7$ = 0.00006103515*
- *P(coding sequence,say ATGGTTC) = 0.3*0.1*0.7*1.0*1.0*0.1*0.1 = 0.00021*

Gene Prediction Using Log-likelihood

- The ratio between the probabilities of the putative coding sequence and the random sequence is the likelihood ratio.
- The logarithm of this ratio is the log-likelihood ratio □
- In this case, log(P(c)/P(r)) = 1.78
- This score is
  - **0**, if both the sequences are **equally** likely
  - **>0**, if the sequence is **more** likely to be a **coding region** than a random sequence
  - **<0**, if the sequence is **less** likely to be in a **coding region** than a random sequence
- A more advanced modification of the above, combined with a lot of heuristics is what PRODIGAL implements

PRODIGAL - in a nutshell

- PROkaryotic DYnamic programming Gene-finding ALgorithm



Constructing a training set for protein coding

Building log-likelihood coding statistics

Sharpening coding scores

Length factor to coding

Iterative start training

Final dynamic programming

PRODIGAL

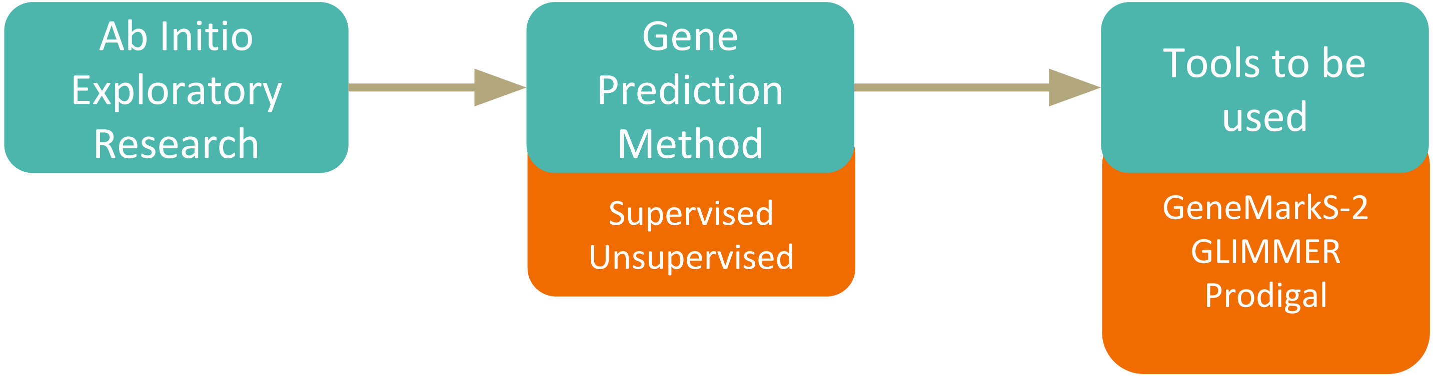
Advantages:

- Extremely fast and lightweight
- Highly Specific - False positive rate < 5%
- A distinct advantage of Prodigal over other gene-finders:
  - Performs well with high GC content genomes

Disadvantages:

- The results from Prodigal could be biased, because it was developed using results from GenBank annotation and using a small set of initial genomes
- Recognition of short and atypical genes needs improvement

Ab Initio - Proposed Strategy



Ab Initio Exploratory Research

Gene Prediction Method

Supervised
Unsupervised

Tools to be used

GeneMarkS-2
GLIMMER
Prodigal

Non-Coding RNA

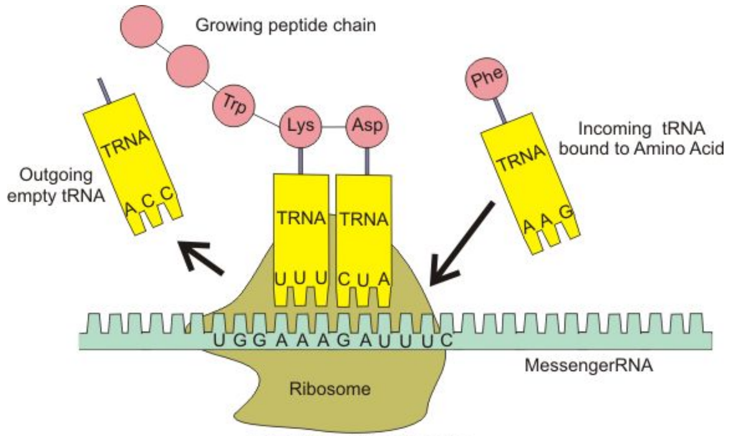
Description, Tools, and Strategy

Non Coding RNA

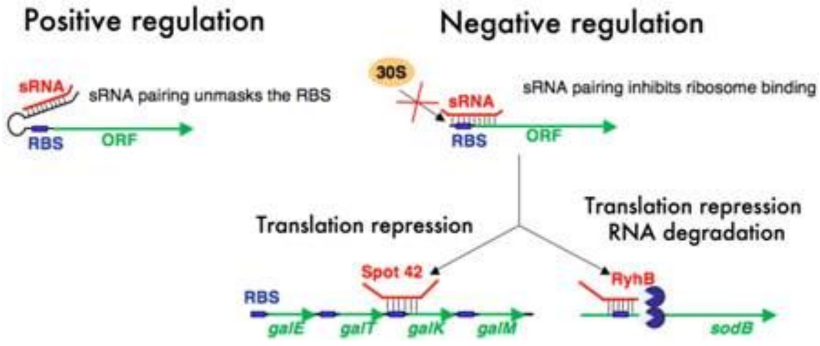
- RNA that gets transcribed from a DNA template but not translated into a protein
- Secondary Structure plays a key role
- Three main classes in bacteria:
    - tRNA/tmRNA
    - rRNA
    - sRNA

Non Coding RNA - Bacteria

- Role of ncRNA in bacterial genomes:
    - Protein synthesis/Translation (tRNA and rRNA)
    - Gene regulation (sRNA)
    - Both of them can be related to antibiotic resistance



Peptide Synthesis

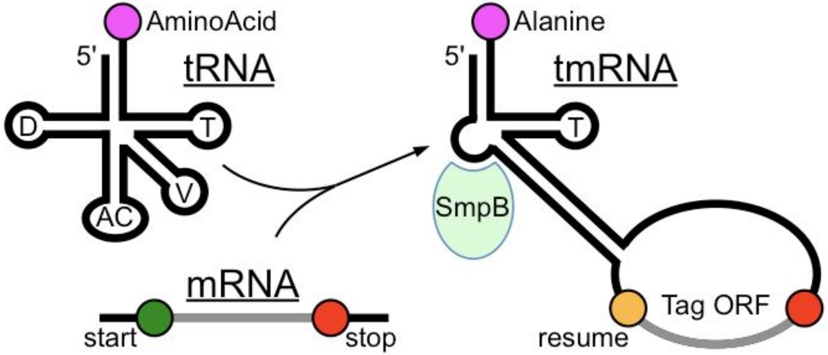


Non Coding RNA - Tools: Tool Selection

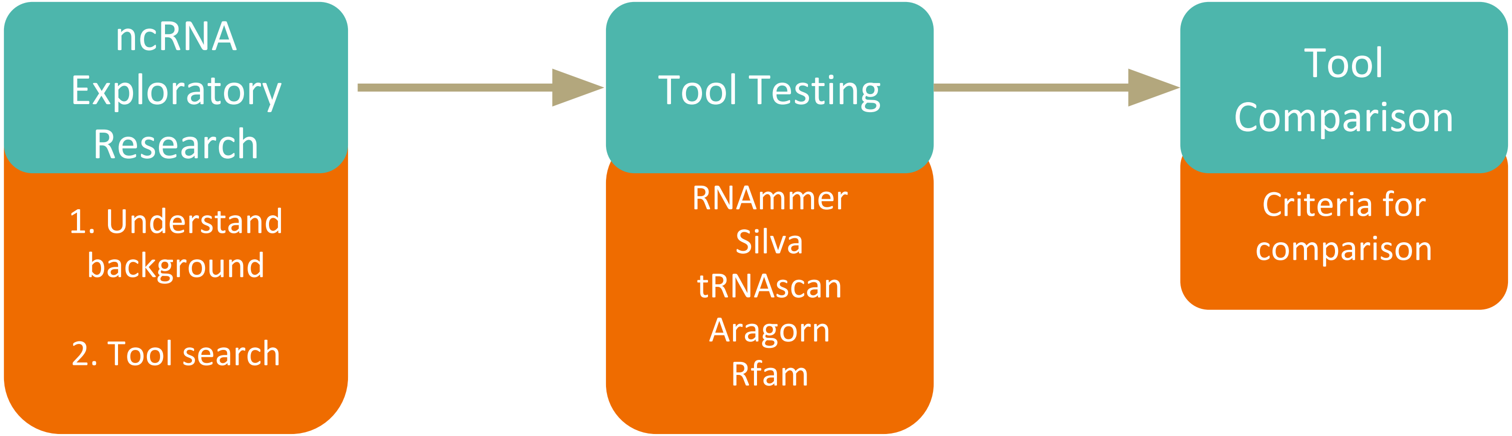
- Data: 260+ assembled *Klebsiella* genomes (unknown species)
- Needs:
    - Speed
    - Accuracy
    - Specific to ncRNAs in Prokaryotic genomes
    - Preferably no need for reference genome

Non Coding RNA - Tools

- **rRNA**
  - RNAmmer
    - Using data from rRNA database
    - Higher Novelty and <1 min/genome
    - Online tool has limitation
  - Silva
    - Using data from rRNA database
    - Many features online
- **tRNA**
  - tRNAscan-SE 2.0
    - Better at finding weird tRNAs
    - Accurate, low error rate and ~1.8 mb/min
  - Aragorn
    - tRNA and tmRNA
    - Error and speed are CG content dependent
    - 5X faster with 40-60% CG
- **sRNA**
  - Rfam
    - Database of ncRNA
    - Group ncRNA into families using multiple sequence alignments and covariance models



Non Coding RNA - Proposed Strategy



ncRNA Exploratory Research

1. Understand background

2. Tool search

Tool Testing

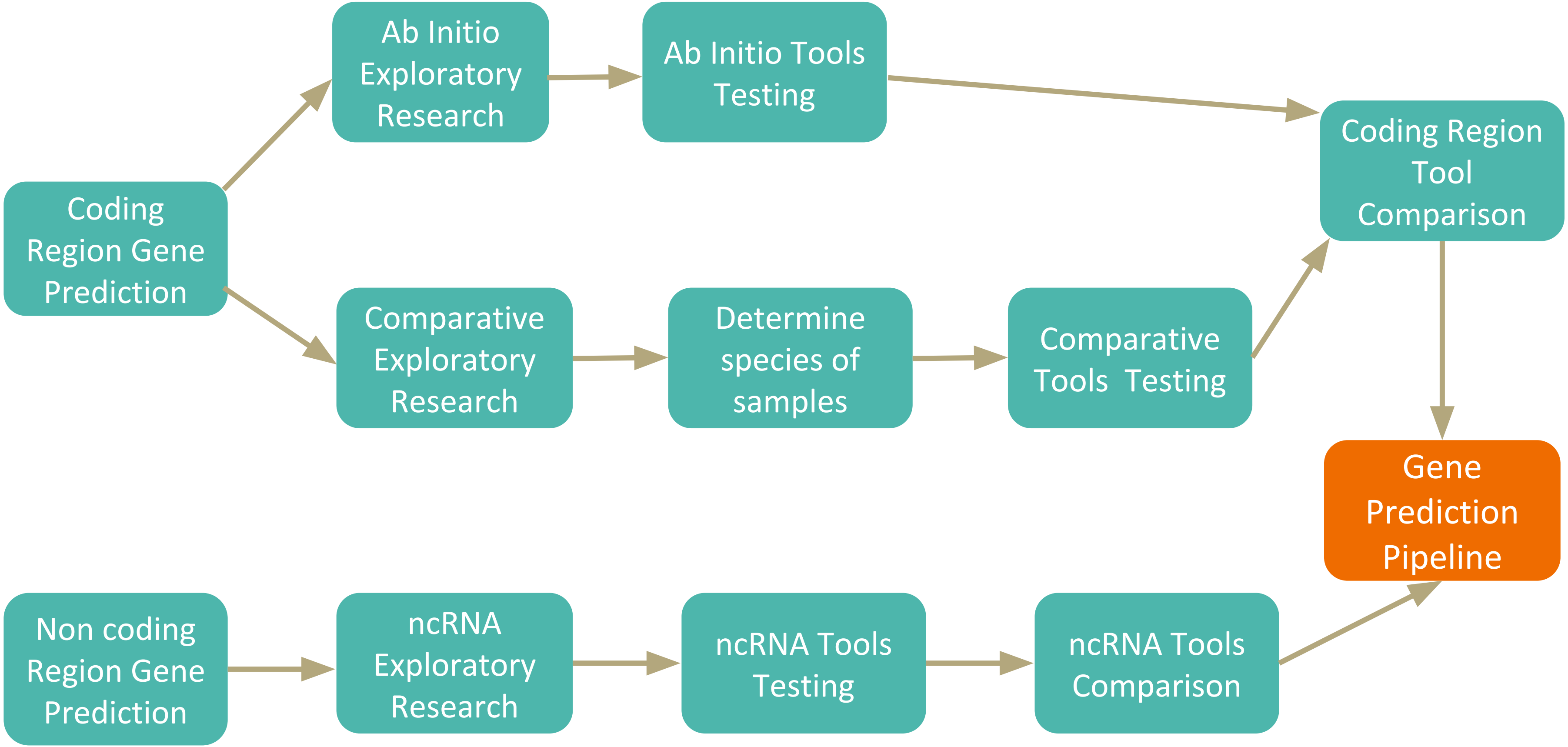
RNAmmer
Silva
tRNAscan
Aragorn
Rfam

Tool Comparison

Criteria for comparison

Proposed Strategy Overview

Workflow Diagram



Coding Region Gene Prediction

Ab Initio Exploratory Research

Ab Initio Tools Testing

Comparative Exploratory Research

Determine species of samples

Comparative Tools Testing

Coding Region Tool Comparison

Gene Prediction Pipeline

Non coding Region Gene Prediction

ncRNA Exploratory Research

ncRNA Tools Testing

ncRNA Tools Comparison

Questions?

References

- Wahlestedt, Claes. "Targeting long non-coding RNA to therapeutically upregulate gene expression." *Nature reviews Drug discovery* 12.6 (2013): 433-446.
- Nadya Dimitrova, and Thales Papagiannakopoulos. *7.345 Non-coding RNAs: Junk or Critical Regulators in Health and Disease?.* Spring 2012. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.
- Wilusz, Jeremy E., Hongjae Sunwoo, and David L. Spector. "Long noncoding RNAs: functional surprises from the RNA world." *Genes & development* 23.13 (2009): 1494-1504.
- Flicek P, Keibler E, Hu P, Korf I, Brent MR. Leveraging the Mouse Genome for Gene Prediction in Human: From Whole-Genome Shotgun Reads to a Global Synteny Map. *Genome Research*. 2003;13(1):46-54.
- Parra, Genís et al. “Comparative Gene Prediction in Human and Mouse.” *Genome Research* 13.1 (2003): 108–117. PMC. Web. 22 Feb. 2018.