Microorganism Population Analysis

Question: How can we conduct a census of the members of a bacterial community when the overwhelming number have never been sequenced?

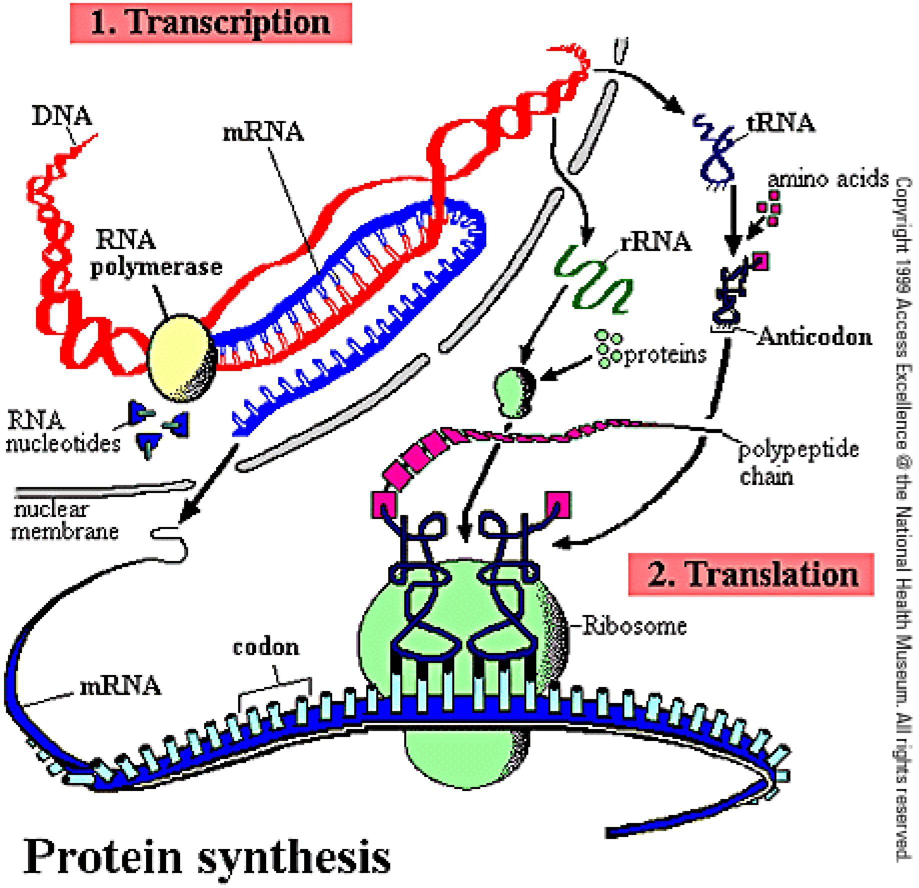
Motivation: Global warming

Input: Sequence data from random samples of one or more communities.

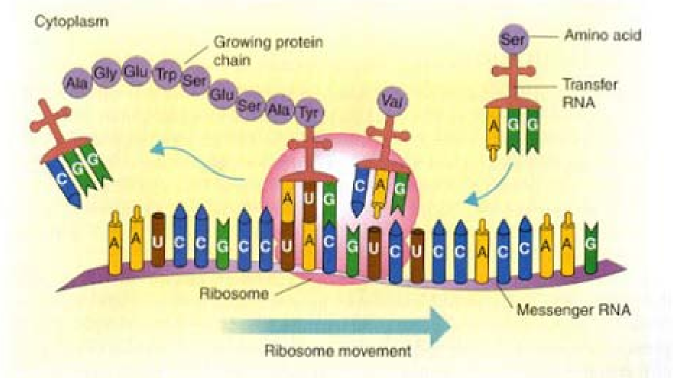
Output: What phylogenetic groups are there and how do they change?

General Introduction

Basic Functions



During translation, the genetic code in mRNA is read and converted into protein by means of the protein synthesizing machinery, which consists of ribosomes, tRNA, amino acids, and a number of enzymes.



Presentation Overview

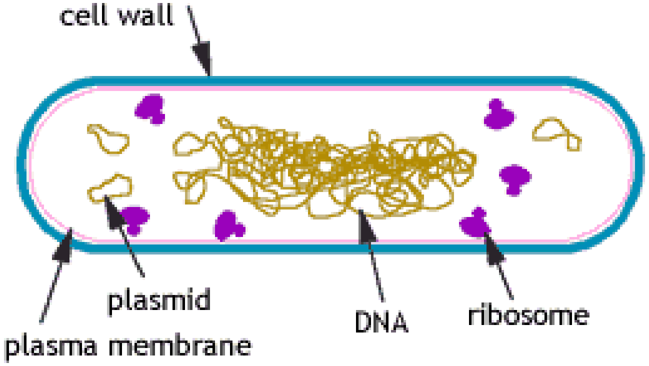
- General Introduction
- Sequence Analysis
  - k-mer classification
  - rRNA classification and population statistics
- Sequence Design
  - Novel sequence design
  - Gene overlapping
- Future Work
  - Sequence design tools
  - Classification and analysis framework

Sequence Analysis – k-mer classification

Introduction

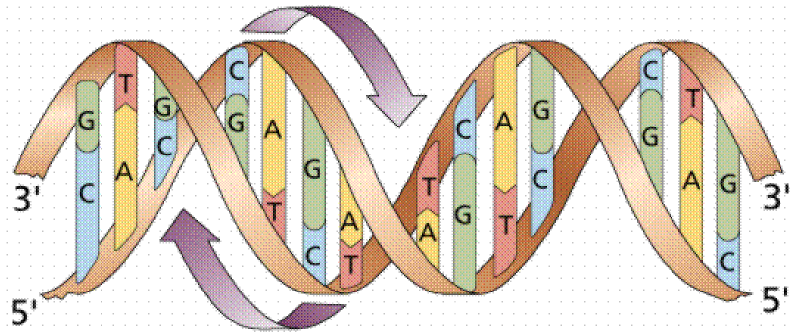
Bacteria

- Single cell organisms.
- Millions can fit into the eye of a needle.
- Can be found virtually everywhere (air, soil, water, in us).
- Our mouth is home to more than 500 species of bacteria.
- A teaspoon of soil contains about a billion of bacterial cells, representing thousands of bacterial types.



Sequence Analysis – k-mer classification

Introduction



For our purposes, bacterial DNA is a string over a four letter alphabet, {A, C, G, T}. We will call the letters of this alphabet "bases".

- Typical bacterial DNA sequence length ranges between 500,000 – 10,000,000 base pairs (bp).
- More than 500 genomes are fully sequenced today, since 1995. Discovery rates increased, but total number still small, because…
- … there are **millions** of different bacterial species in nature.
- Small percentage can grow in laboratory conditions (~1%).
- Environmental sample sequencing has only recently emerged.

Sequence Analysis – k-mer classification

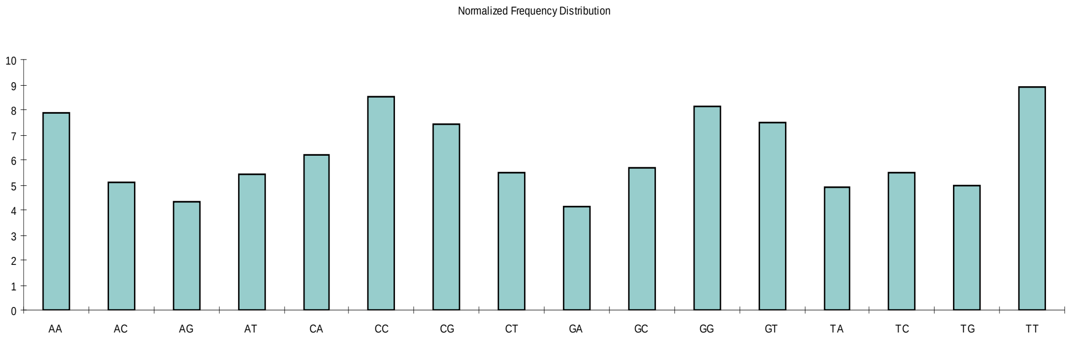
Our problem: Given an environmental sample, which bacterial species can be identified in it?

- Probability that *exact matching* will reveal already sequenced organisms is too small.
- We will use ***Genomic Signatures***, meaning the frequency distributions of oligonucleotides in a genomic sequence.

. . . TTGCAGTGTCGATCTAGCGTCGACTGATTTATCGCGGCGGATTGCGTACTACTAGCAGCTACGTA . . .

TG
GC
CA
AG
GT
TG . . .

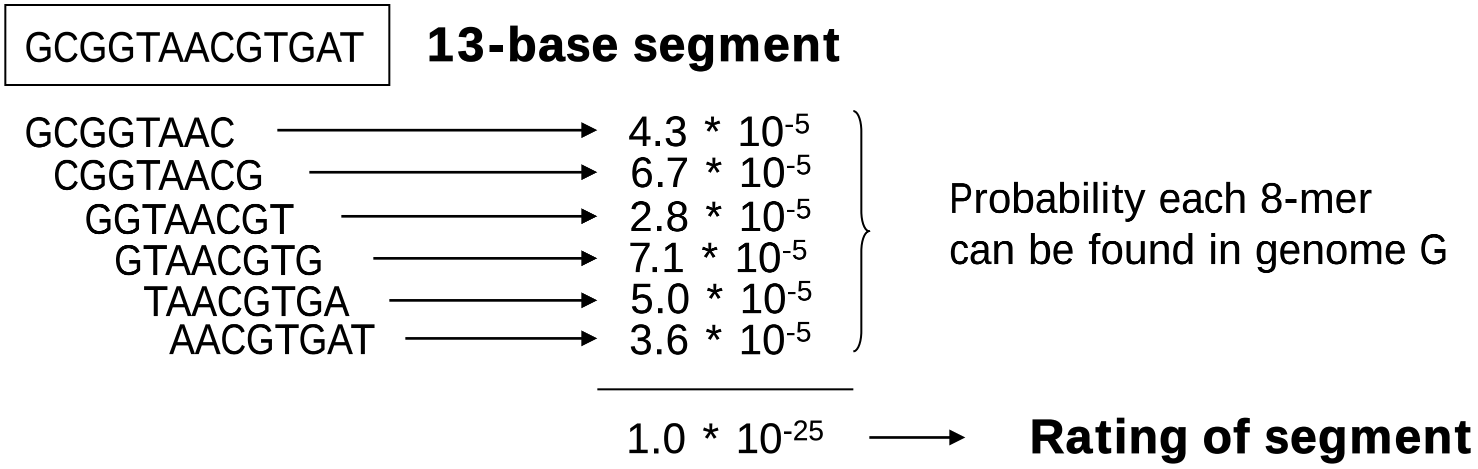
Dinucleotide Example



Sequence Analysis – k-mer classification

Identifying bacteria using genomic signatures

- Using a ***Naïve Bayesian Classifier,*** Sandberg et al. identified 400bp segments with 85% probability from a pool of 25 known unrelated fully sequenced microbes.
- A naïve bayesian classifier calculates the probability of finding a sequence S of length N in a genome G as the product of the individual probabilities of k-mers constituting S, in G.



GCGGTAACGTGAT **13-base segment**

GCGGTAAC → $4.3 * 10^{-5}$

CGGTAACG → $6.7 * 10^{-5}$

GGTAACGT → $2.8 * 10^{-5}$

GTAACGTG → $7.1 * 10^{-5}$

TAACGTGA → $5.0 * 10^{-5}$

AACGTGAT → $3.6 * 10^{-5}$

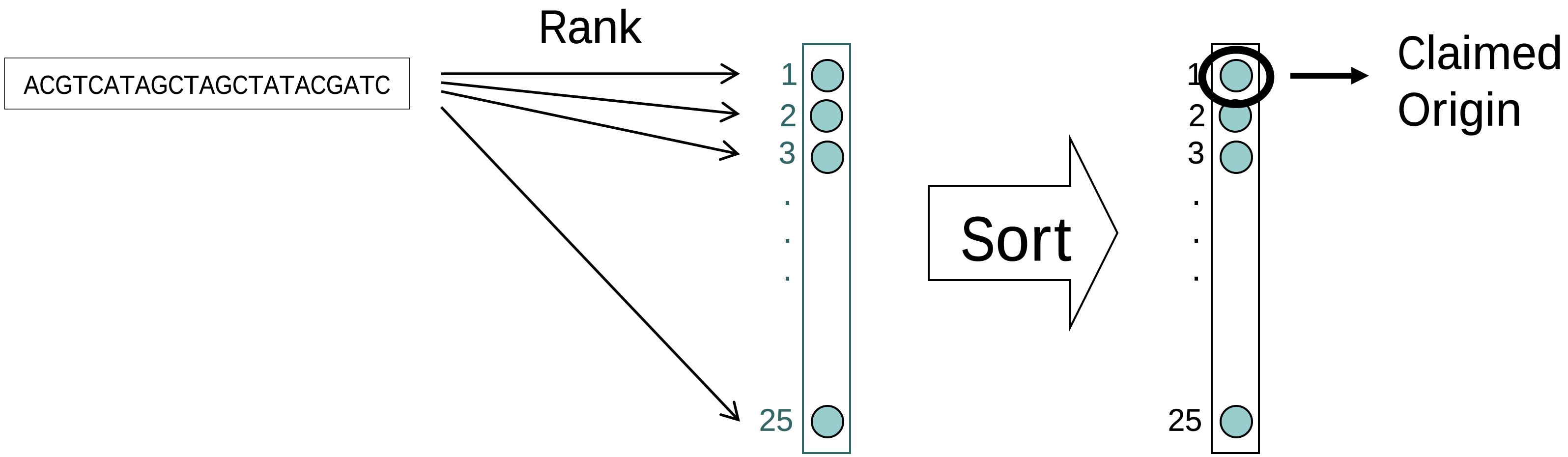
Probability each 8-mer can be found in genome G

$1.0 * 10^{-25}$ → **Rating of segment**

Sequence Analysis – k-mer classification

Identifying bacteria using genomic signatures

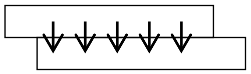
- Based on segment ratings, a sequence fragment is scored against all known genome signatures. The origin is claimed as the highest scoring genome.



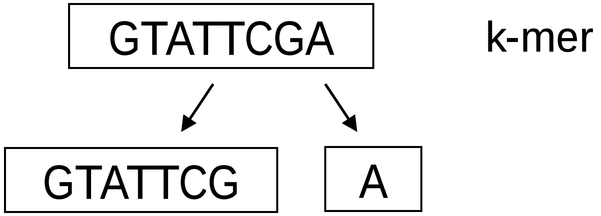
Sequence Analysis – k-mer classification

Our improvement

Since k-mers are not really ***independent***, we calculate the conditional probability of a k-mer in a sequence as the probability of the last base appearing after the k-1 bases of the prefix.



Example:



GTATTCGA

k-mer

GTATTCG

A

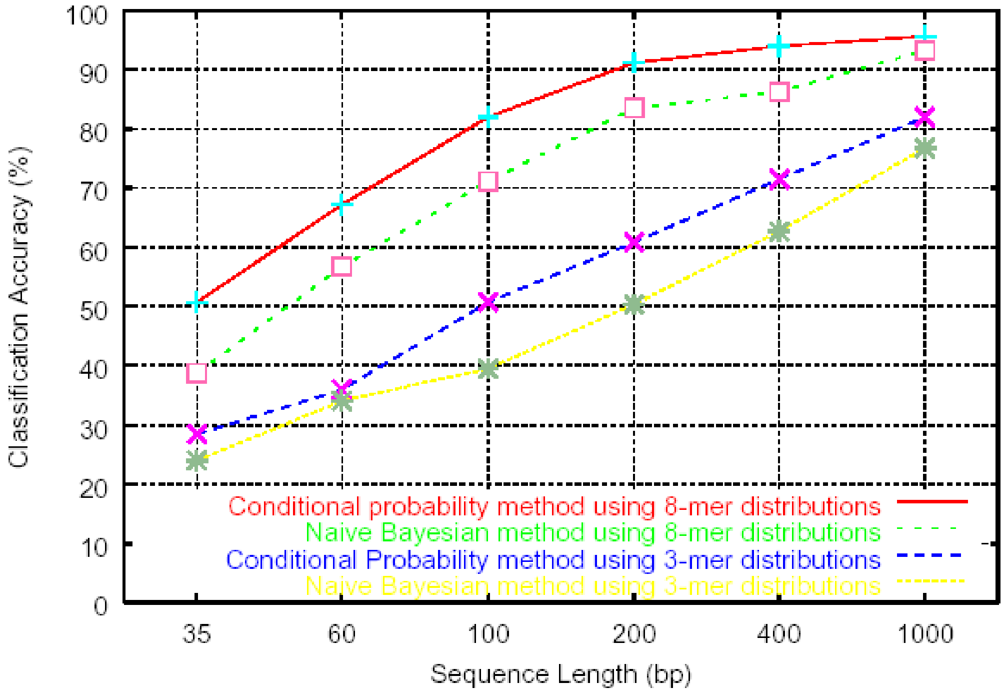
After a GTATTCG , probability of a

A	: 2/7
C	: 4/7
G	: 1/7
T	: 0

Sequence Analysis – k-mer classification

Our improvement

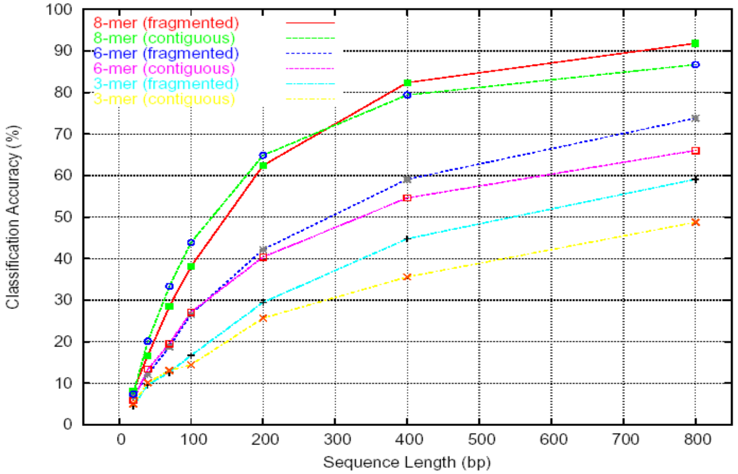
The resulting classifier using the conditional probabilities outperforms the naïve one.



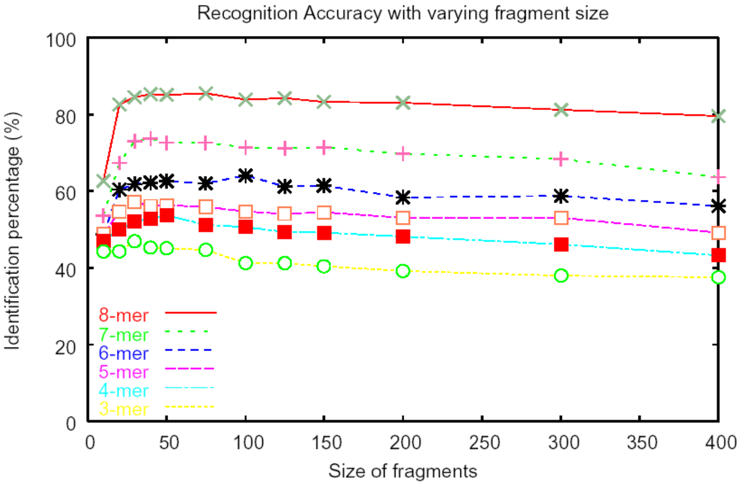
Sequence Analysis – k-mer classification

Results

- Fragmentation of the sequence also results in more accurate classification (which seems counter-intuitive, since less k-mers are produced)



- The optimal size fragment depends on the k-mer size

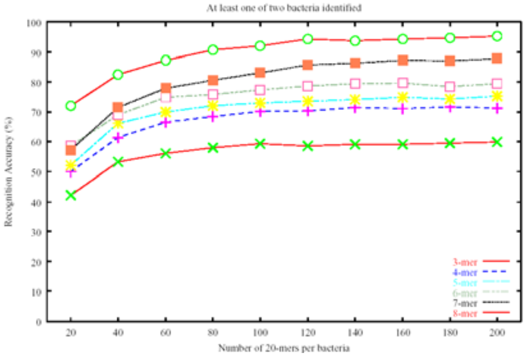


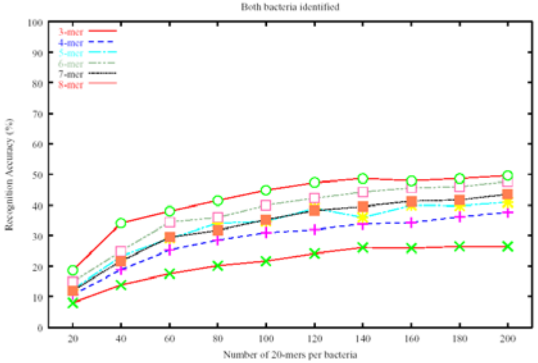
Sequence Analysis – k-mer classification

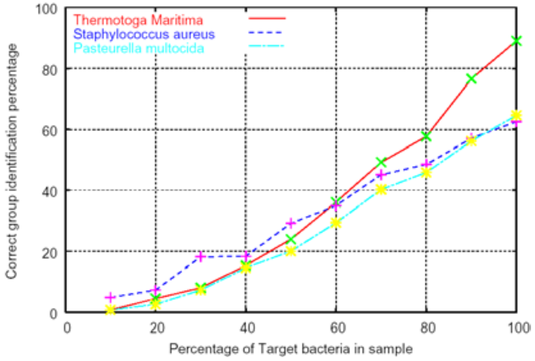
Results

The conditional classifier can also:

- Accurately identify phylotypes of sequence fragments from sequences resembling ones in database.
- Recognize accurately one of two bacteria in a equi-probable mixed sample or even both with 50% probability. Also identify the majority bacterium in a sample and approximate its frequency.


At least one of two bacteria identified
Recognition Accuracy (%)
Number of 20-mers per bacteria
3-mer
4-mer
5-mer
6-mer
7-mer
8-mer


Both bacteria identified
Recognition Accuracy (%)
3-mer
4-mer
5-mer
6-mer
7-mer
8-mer
Number of 20-mers per bacteria


Thermotoga Maritima
Staphylococcus aureus
Pasteurella multocida
Correct group identification percentage
Percentage of Target bacteria in sample