Course organization

- Introduction (Week 1-2)
  - Course introduction
  - A brief introduction to molecular biology
  - A brief introduction to sequence comparison
- Part I: Algorithms for Sequence Analysis (Week 3 - 11)
  - Chapter 1-3, Models and theories
    - » Probability theory and Statistics (Week 4)
    - » Algorithm complexity analysis (Week 5)
    - » Classic algorithms (Week 6)
    - » Lab: Linux and Perl
  - Chapter 4, Sequence alignment (week 7)
  - Chapter 5, Hidden Markov Models (week 8)
  - Chapter 6. Multiple sequence alignment (week 10)
  - **Chapter 7. Motif finding (week 11)**
  - **Chapter 8. Sequence binning (week 11)**
- Part II: Algorithms for Network Biology (Week 12 - 16)

Chapter 8
Metagenomic sequence classification

Chaochun Wei

Fall 2014

Contents

1. Reading materials
2. Sequence composition
   - Metagenomic binning

Reading materials

Qi, J., Luo, H., Hao, B. “CVTree: a phylogenetic tree reconstruction tool based on whole genomes”, Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W45-7.

Liu, J., Wang, H., Yang, H., Zhang, Y., Wang, J., Zhao, F., Qi, J., “Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms ”, Nucleic Acids Res, 2012, Aug 31,

Brady, A., and Salzberg, S., “Phymm and PhymmBL: metagenomic phylogenic classification with interpolated Markov models”, Nat Methods. 2009 Sep;6(9):673-6. Epub 2009 Aug 2

Brady, A., and Salzberg, S., “PhymmBL expanded: confidence scores, custom databases, parallelization and more”, Nat Methods, 2011, May, 8(5):367

Microbes

- Microbes are ubiquitous, and they are essential to all life.
- Bacteria, archaea, and microeukaryotes dominate Earth’s habitats, compound recycling and nutrient sequestration.

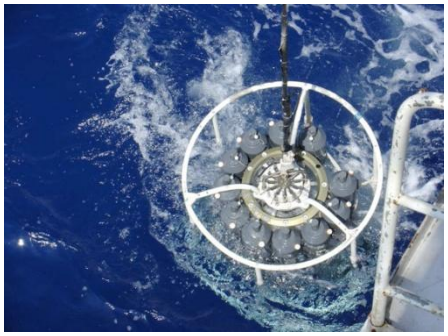
Metagenomics

- Definition:
  - Metagenomics is the study of metagenomes, genetic material recovered directly from environmental samples.

http://en.wikipedia.org/wiki/Metagenomics

Typical Sources of Metagenomes

- Soil samples
- Sea water samples
- Seabed samples
- Air samples
- Medical samples
- Ancient bones
- Human microbiome



Human contains not only the human genome

- In a healthy adult
  - Microbial cells are 10 times more than human cells
  - Most of them are in gut

http://nihroadmap.nih.gov/hmp/

Three basic questions

- Who are they?
  - Metagenome binning
- What are they doing?
  - Function annotation
- How do they compare?
  - Comparative metagenomics

Q1:WHO ARE THEY?

Two types of metagenomics methods

- Marker gene sequencing, 16S rRNA, 18S rRNA, …
- Whole genome shotgun sequencing

Variant regions of 16S rRNA gene

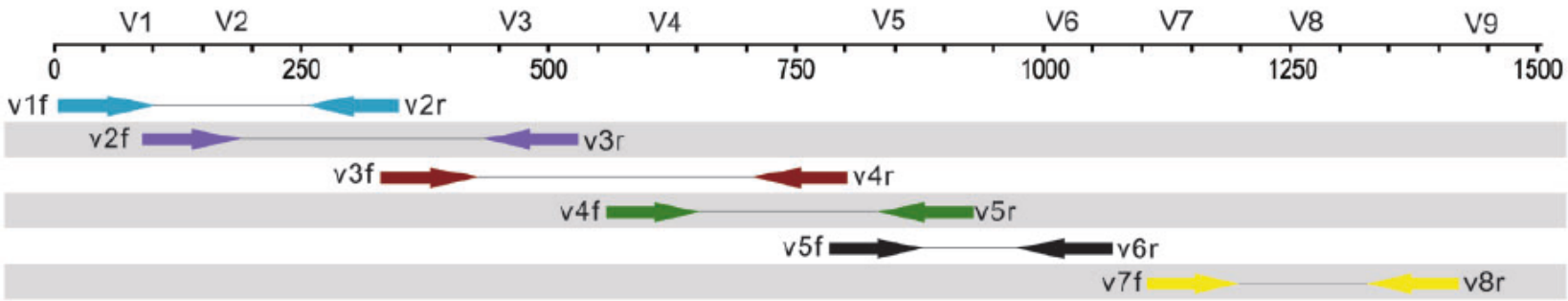
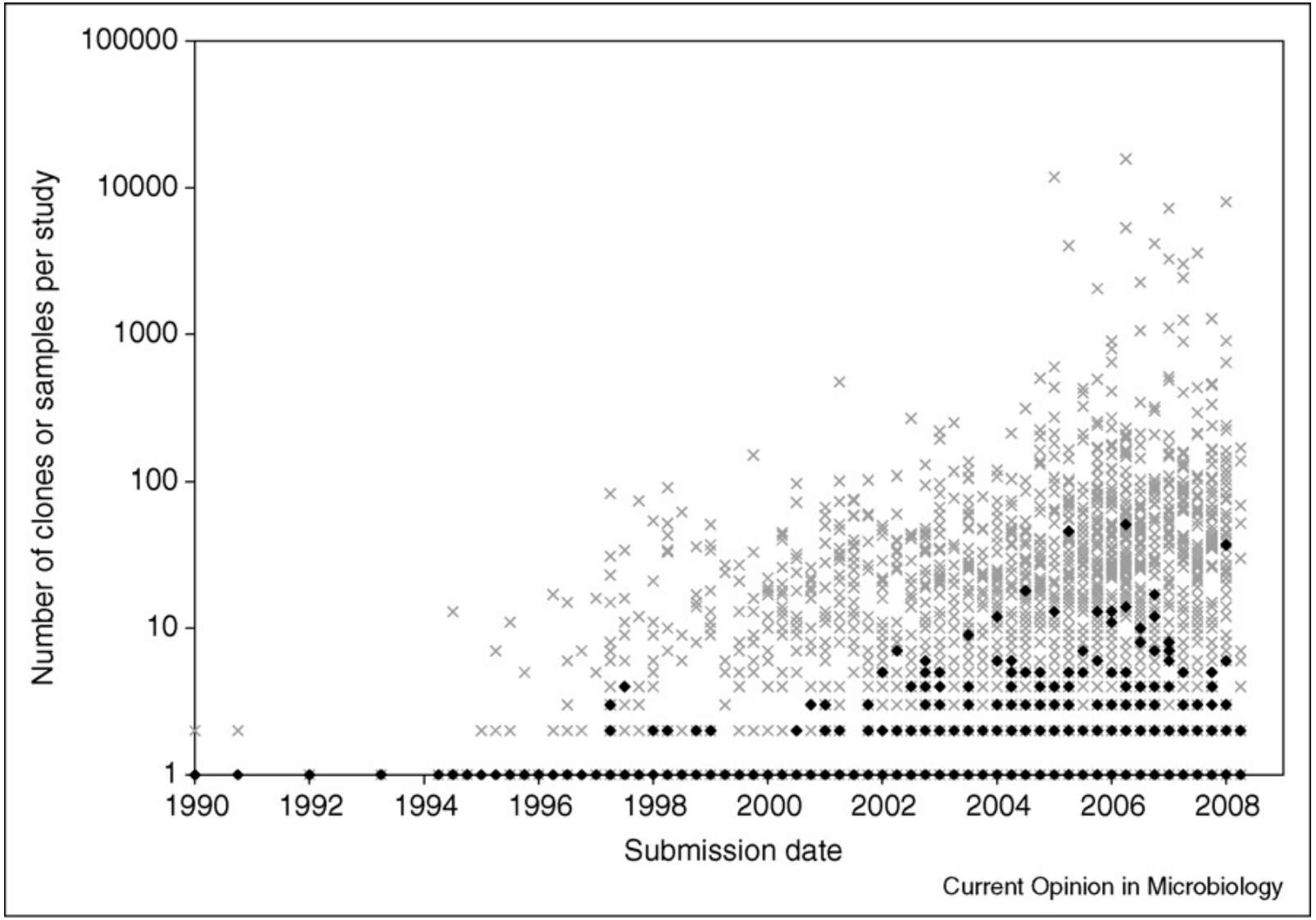


Figure 1. Positions of primer sequences and tandem regions used in this work for 454 titanium and Illumina, mapped along 16S rRNA gene (co-ordinates based on the *Escherichia coli* 16S rRNA gene sequence). The arrows (~100 bases) show approximately Illumina sequence read length.

- We find that different taxonomic assignment methods vary radically in their ability to recapture the taxonomic information in full-length 16S rRNA sequences: most methods are sensitive to the region of the 16S rRNA gene that is targeted for sequencing, but many combinations of methods and rRNA regions produce consistent and accurate results. To process large datasets of partial 16S rRNA sequences obtained from surveys of various microbial communities, including those from human body habitats, we recommend the use of Greengenes or RDP classifier with fragments of at least 250 bases, starting from one of the primers R357, R534, R798, F343 or F517.

Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers.
Liu Z, DeSantis TZ, Andersen GL, Knight R.
Nucleic Acids Res. 2008 Oct;36(18):e120. Epub 2008 Aug 22.

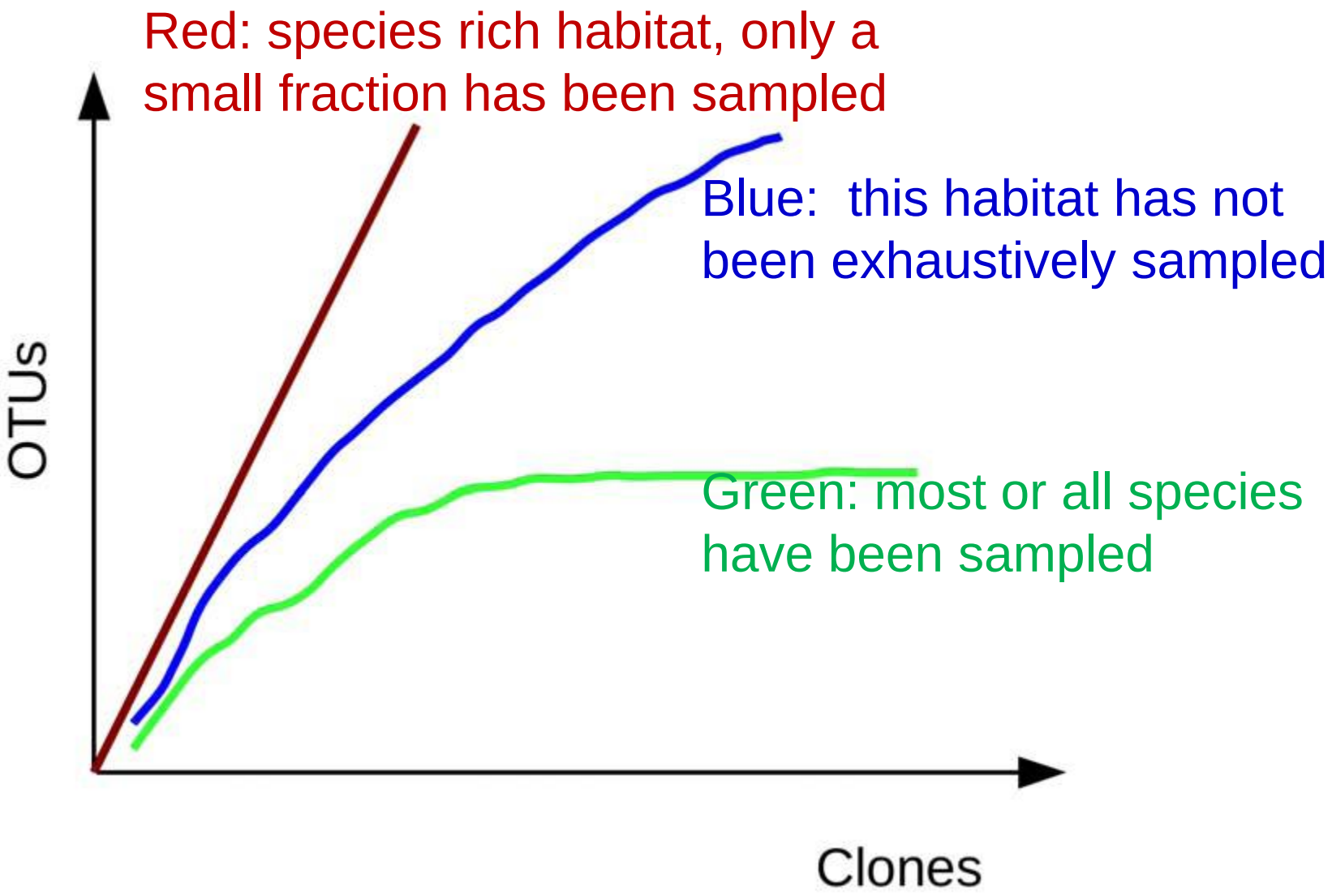
16S rRNA genes



DNA Sequencing

- First generation
  - Sanger
- Second generation
  - 454, Illumina, SOLiD
- Third generation

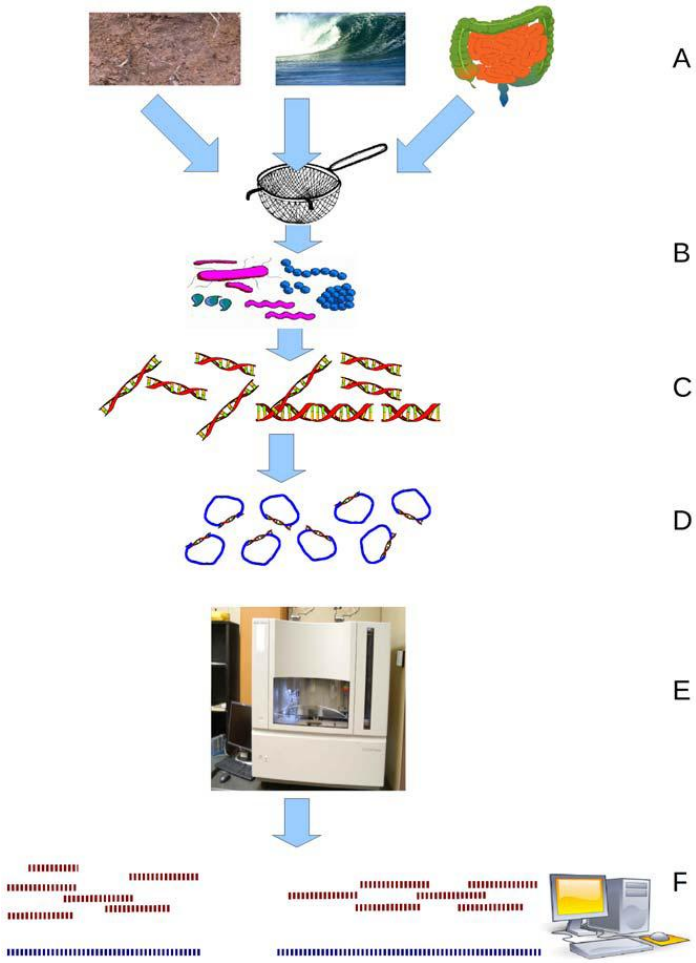
Rarefaction curves.



Limitations of 16S classification

- The copy number of 16S rRNA gene can vary by an order of magnitude between bacterial species
- PCR-induced biases.

Environmental Shotgun Sequencing



A Sampling from habitat;

B Filtering particles, typically by size;

C DNA extraction and lysis;

D Cloning and library;

E Sequence the clones;

F Computational analysis

A primer on metagenomics. Wooley JC, Godzik A, Friedberg I. PLoS Comput Biol. 2010 Feb 26;6(2):e1000667. Review.

Binning

- Alignment based methods
- Non-alignment based methods
  - K-mer frequency (Codon usage)
  - High-order Markov Model
    - HMM/IMM

Non-alignment based methods: component composition method (1)

- K-mer frequency
  - Input: query sequences, a reference database
  - Output: binning result of query sequences

1. Count k-mer frequencies for genomes in the reference database (can be pre-built)
2. Count k-mer frequencies for the query sequences
3. Compare the k-mer frequencies to the reference database, pick the genome with closest distance for each query

Non-alignment based methods: component composition method (2)

- High-order Markov model method
  - Input: query sequences, a reference database
  - Output: binning result of query sequences

1. Count transition probability of k^{th}-order Markov models for genomes in the reference database (can be pre-built)
2. Calculate a k^{th}-order Markov model score for the query sequences for each genome in the reference database
3. Assign the query sequence to the genome with highest score

Non-alignment based methods: component composition method (2)

- The transition probabilities

$$kMM_{i,mn} = P_i(O_m \mid O_n) = \frac{F_i(O_m \mid O_n)}{F_i(O_m)}$$

where O_m and O_n are oligonucleotides of length k,

$P(O_m \mid O_n)$ represents the transition probability from O_m to O_n ,

$F(O_m \mid O_n)$ represents observed count of transitions from O_m to O_n in a genomic sequence i and $F(O_m)$ is the observed count of O_m.

- The score for a query sequence to a reference genome i

$$S_i = -\sum_{j=0}^{l-k-1} ln(p_i(O_j \mid O_{j+1}))$$

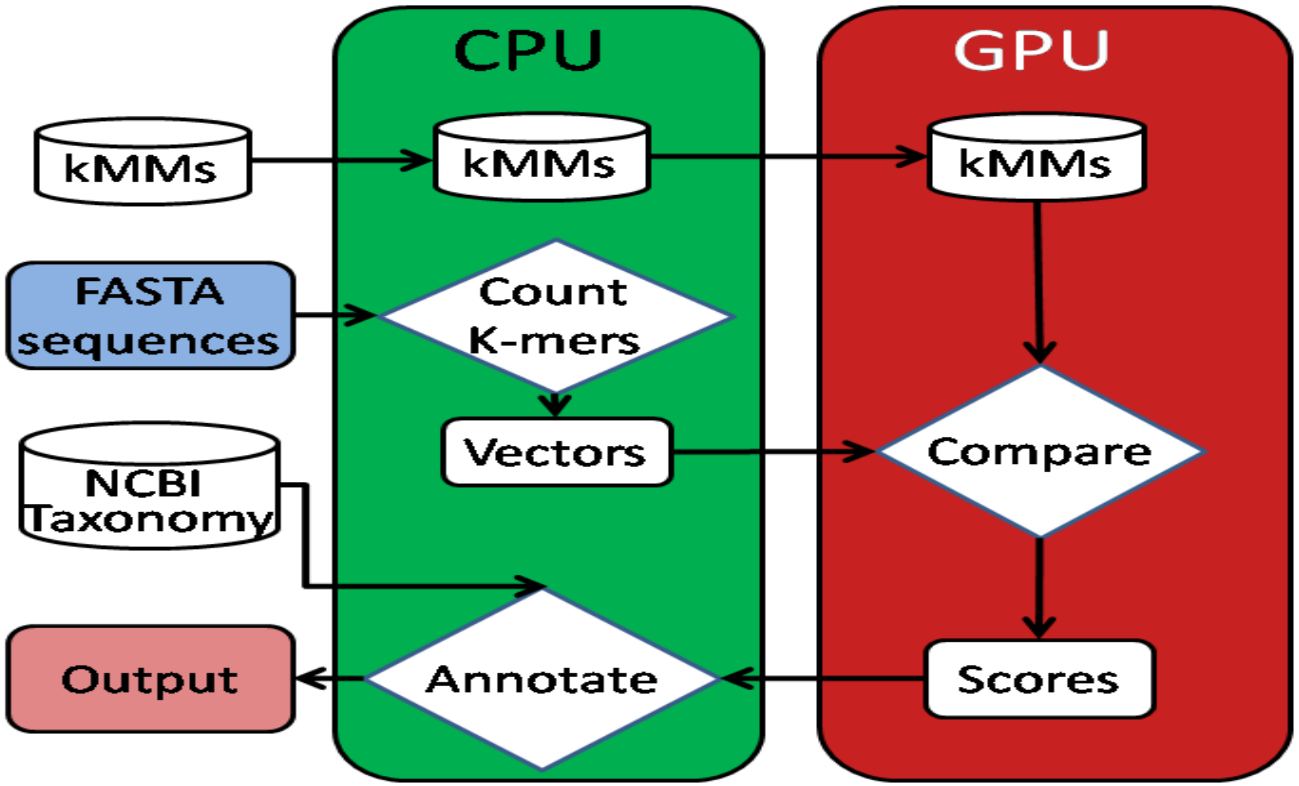
where O_j and O_{j+1} are two oligonucleotides of length k, and $P(O_j|O_{j+1})$ is the transition probability from O_j to O_{j+1} observed in the i-*th* genome. l is the length of the query sequence.

Non-alignment based methods

- Compared to alignment based methods
  - Advantages
    - Simple
    - Faster
      - Can be 2-5 orders faster
  - Disadvantages
    - Less accurate

Non-alignment based methods: component composition method (2)

- MetaBinG: a High-order Markov model method using GPUs



Jia, P., Xuan, L., Liu, L., Wei., C., PLoS ONE, 2011

Comparison of Phymm and MetaBinG.

Sequence Length (bps)	Phymm		MetaBinG		Speedup
	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)	
100	53.62	573	50.61	4	143
200	64.21	880	60.82	5	176
300	70.71	1262	67.66	6	210
400	73.36	1652	71.56	6	275
500	76.02	1949	74.48	8	244
600	78.47	2330	77.24	8	291
700	79.89	2632	79.21	9	292
800	81.86	3006	80.25	10	301
900	82.40	3403	80.77	12	284
1000	84.18	3795	82.35	13	292

Jia, P., Xuan, L., Liu, L., Wei., C., PLoS ONE, 2011

Real examples

- biogas metagenome (~130 Mb)
  - 616,072 454 reads (~230bp)
  - Phymm: 4 days 5 hours 57 minutes and 56 seconds
  - MetaBinG: 248 seconds
- Sponge metagenome (~70 Gb)
  - ~half billion reads
  - MetaBinG: 40 hours

End of Part I

- Practice
- Practice
- Practice