Algorithms and Methods for Cross-Species Forensics and Classification

Darryl Reeves, Ph.D.
Clinical and Research Genomics
04/11/2018

Outline

I. Introduction

II. An alternative view of biological sequences

III. A model for sequence identification

IV. Conclusion/Future plans

Outline

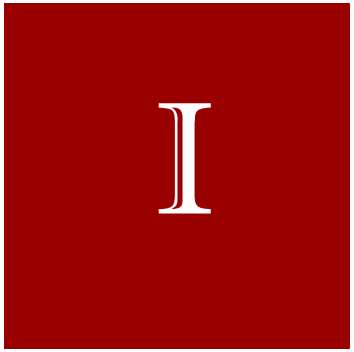
I. Introduction

II. An alternative view of biological sequences

III. A model for sequence identification

IV. Conclusion/Future Research Agenda

Why study communities of organisms (metagenomics)?



- Environment contains some interesting biological activity
- Interaction between community and external environment/host
- Understand how environmental changes impact organisms

I

Motivating examples

Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill

Olivia U Mason[1,2], Terry C Hazen[1,3], Sharon Borglin[1], Patrick SG Chain[4,5], Eric A Dubinsky[1], Julian L Fortney[1], James Han[4,5], Hoi-Ying N Holman[1], Jenni Hultman[1], Regina Lamendella[1], Rachel Mackelprang[5], Stephanie Malfatti[5,6], Lauren M Tom[1], Susannah G Tringe[5], Tanja Woyke[5], Jizhong Zhou[7,8], Edward M Rubin[5] and Janet K Jansson[1,5]

The Dynamics of the Human Infant Gut Microbiome in Development and in Progression toward Type 1 Diabetes

Aleksandar D. Kostic, Dirk Gevers, Heli Siljander, Tommi Vatanen, Tuulia Hyötyläinen, Anu-Maaria Hämäläinen, Aleksandr Peet, Vallo Tillmann, Päivi Pöhö, Ismo Mattila, Harri Lähdesmäki, Eric A. Franzosa, Outi Vaarala, Marcus de Goffau, Hermie Harmsen, Jorma Ilonen, Suvi M. Virtanen, Clary B. Clish, Matej Orešič, Curtis Huttenhower, Mikael Knip[23] on behalf of the DIABIMMUNE Study Group[22], Ramnik J. Xavier[23],

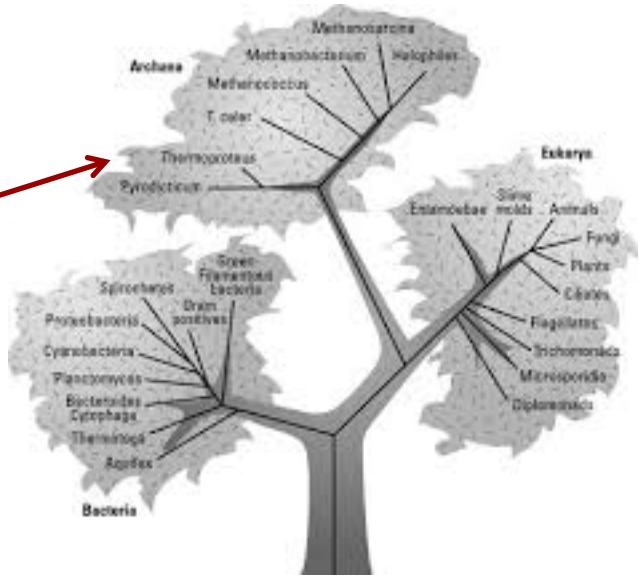
Metagenomics research

- Two very different research domains (environmental and medical)
- Same basic questions
- Given a heterogeneous community of organisms
  - Who is there?
  - What are they doing (function)?
  - What is the abundance of members?

The Goal

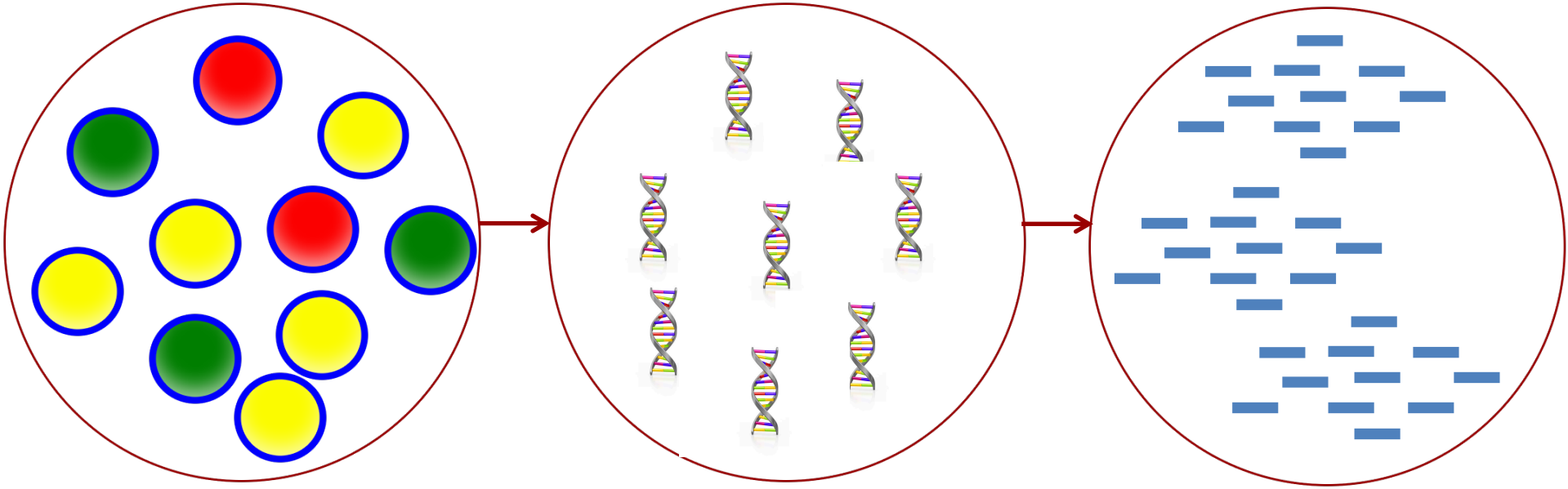
S

DNA Sequence



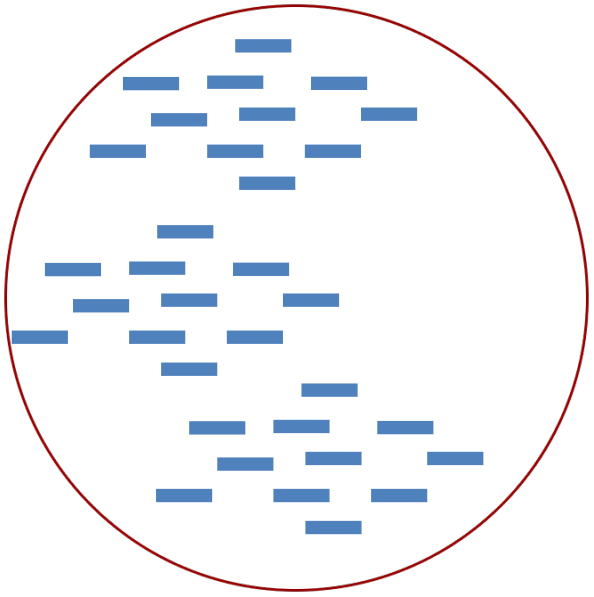
I

Metagenomics – data



I

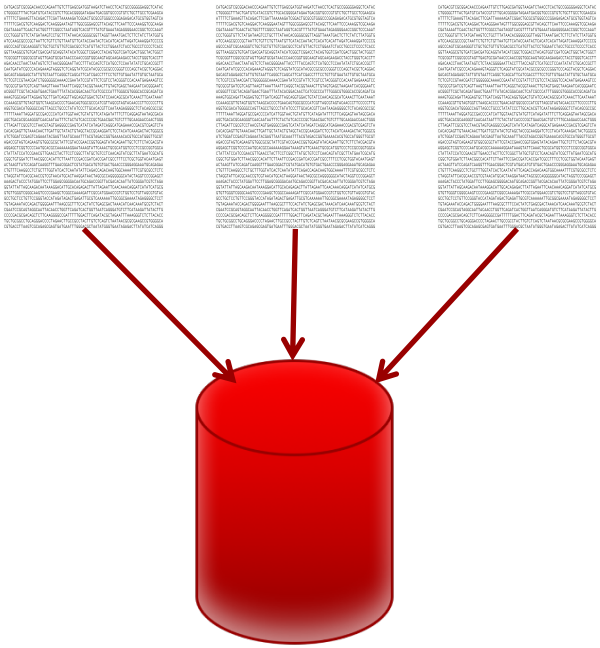
Metagenomics – data



Metagenomics – data

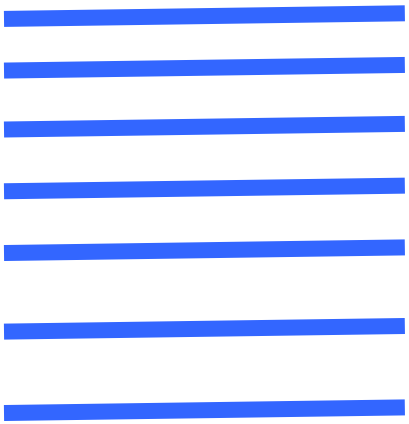
Metagenomic identification/classification

I



Metagenomic identification/classification

Species X



Sequence alignment

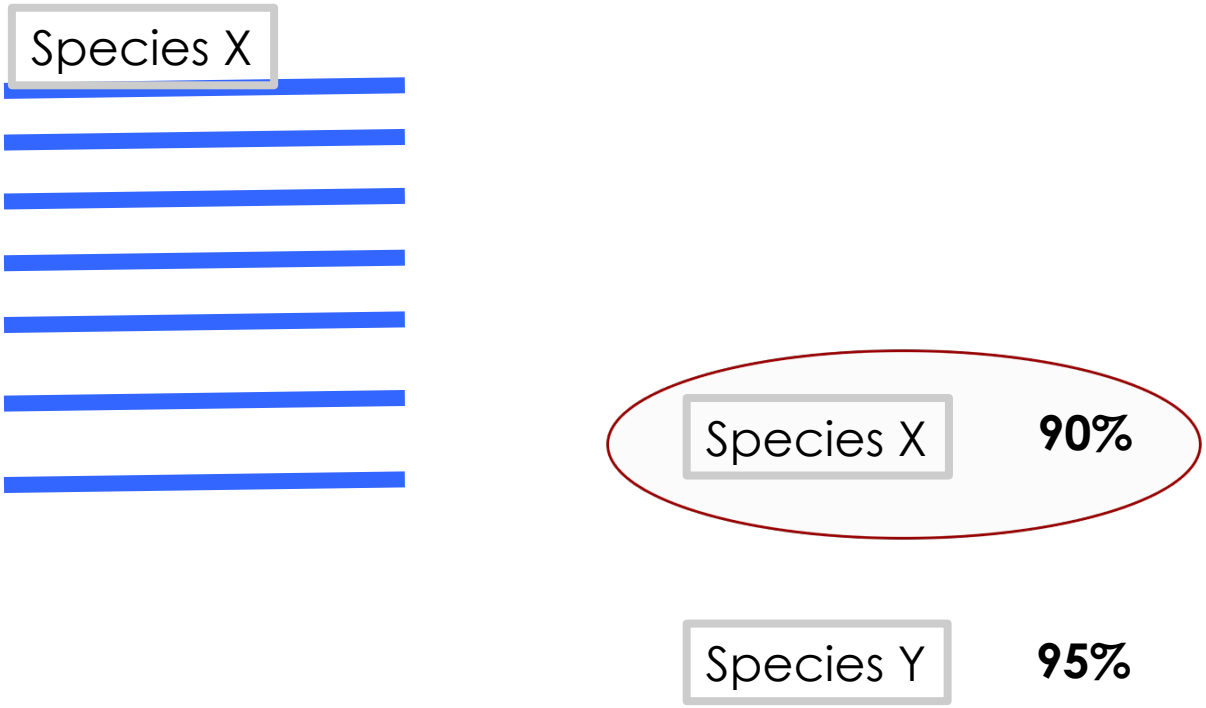
Sub-sequence markers

Sub-sequence composition

Hybrid approaches

Common Limitation

Metagenomic identification/classification



Metagenomic identification/classification

Species X

Species Y

Species Z

Species 1

Species 2

Species 3

Hypothesis

Current methods fail to utilize all available information

Results in a local view of community composition

Statistical learning can be used to develop a global view of communities

Outline

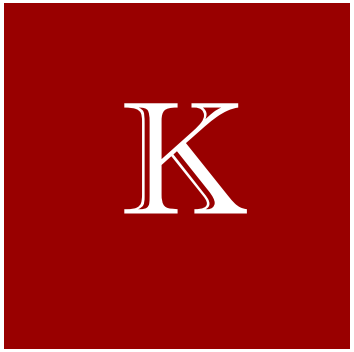
I. Introduction

II. An alternative view of biological sequences

III. A model for sequence identification

IV. Conclusion/Future plans

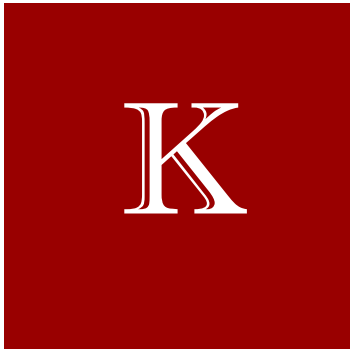
An alternative view



- Sequence identification fundamentally a comparison problem

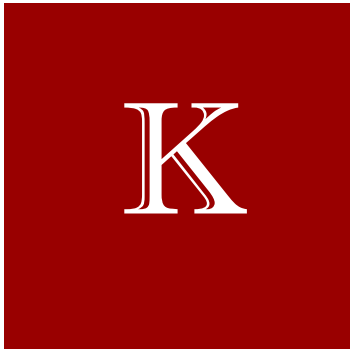
- Comparing text vs. numbers

K-mers



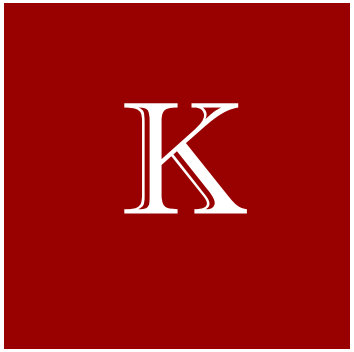
ATGCTAGGAACCCTAGCTTACAGAGCAGTTGCAGG

K-mers

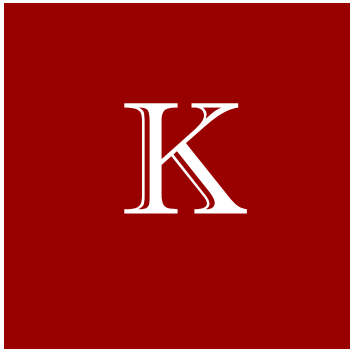


ATGCTAGGAACCCTAGCTTACAGAGCAGTTGCAGG

K-mers

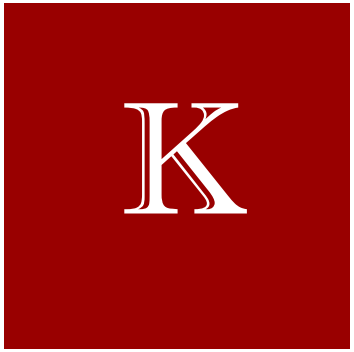


ATGCTAGGAACCCTAGCTTACAGAGCAGTTGCAGG



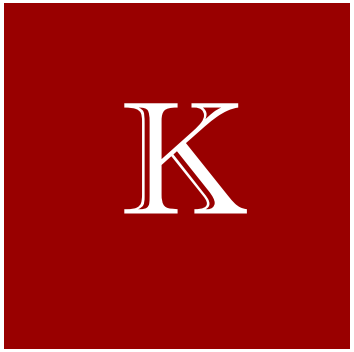
K-mers

AT**GCT**AGGAACCCTAGCTTACAGAGCAGTTGCAGG



K-mers

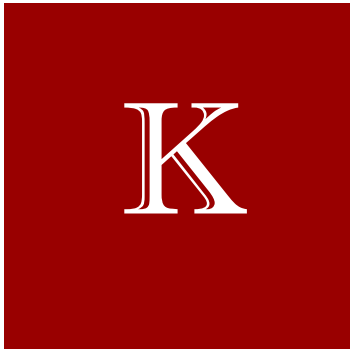
ATGCTAGGAACCCTAGCTTACAGAGCAGTTGCAGG



K-mers

ATGCTAGGAACCCTAGCTTACAGAGCAGTTGCAGG

K-mers



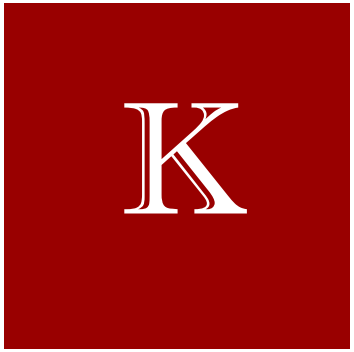
ATGCTAGGAACCCTAGCTTACAGAGCAGTTGCAGG

K-mers

ATGCTAGGAACCCTAGCTTACAGAGCAGTTGCAGG

3-mer	Count
ATG	1
TGC	2
CTA	2
TAG	2
AGG	2
GGA	1
GAA	1

***K*-mer Frequency Distribution**



K-mers

ATGCTAGGAACCCTAGCTTACAGAGCAGTTGCAGG

3-mers, 64 combinations

5-mers, 1024 combinations

23-mers, 70,368,744,177,664 combinations

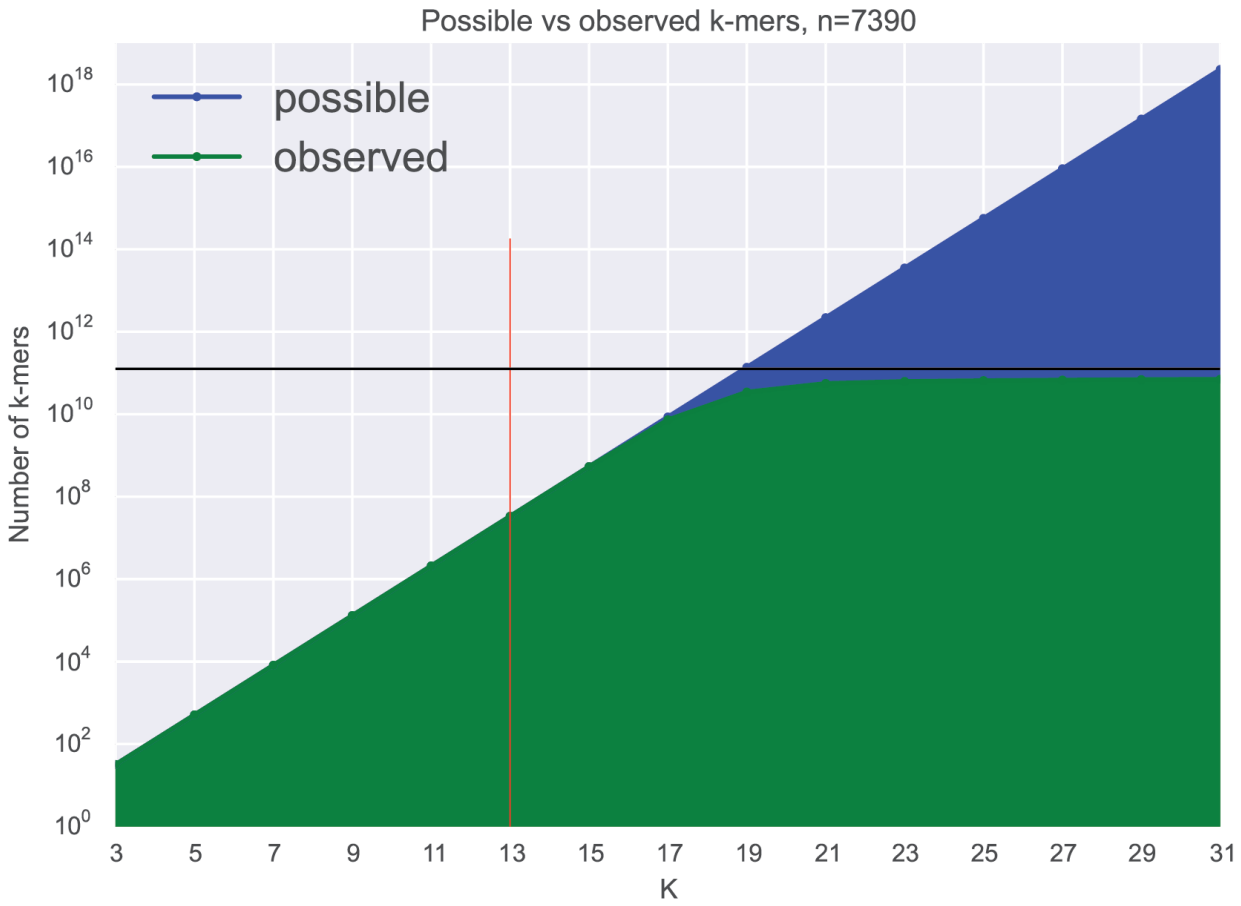
K-mers

K

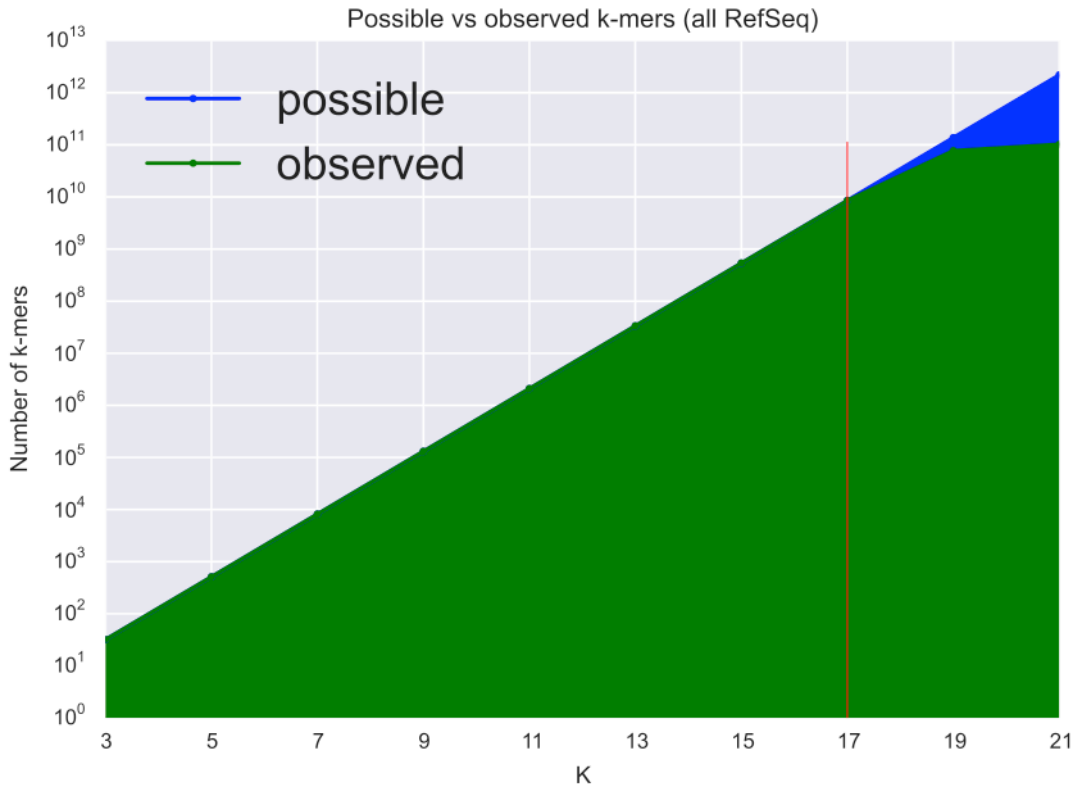
ATGCTAGGAACCCTAGCTTACAGAGCAGTTGCAGG

$$4^K$$

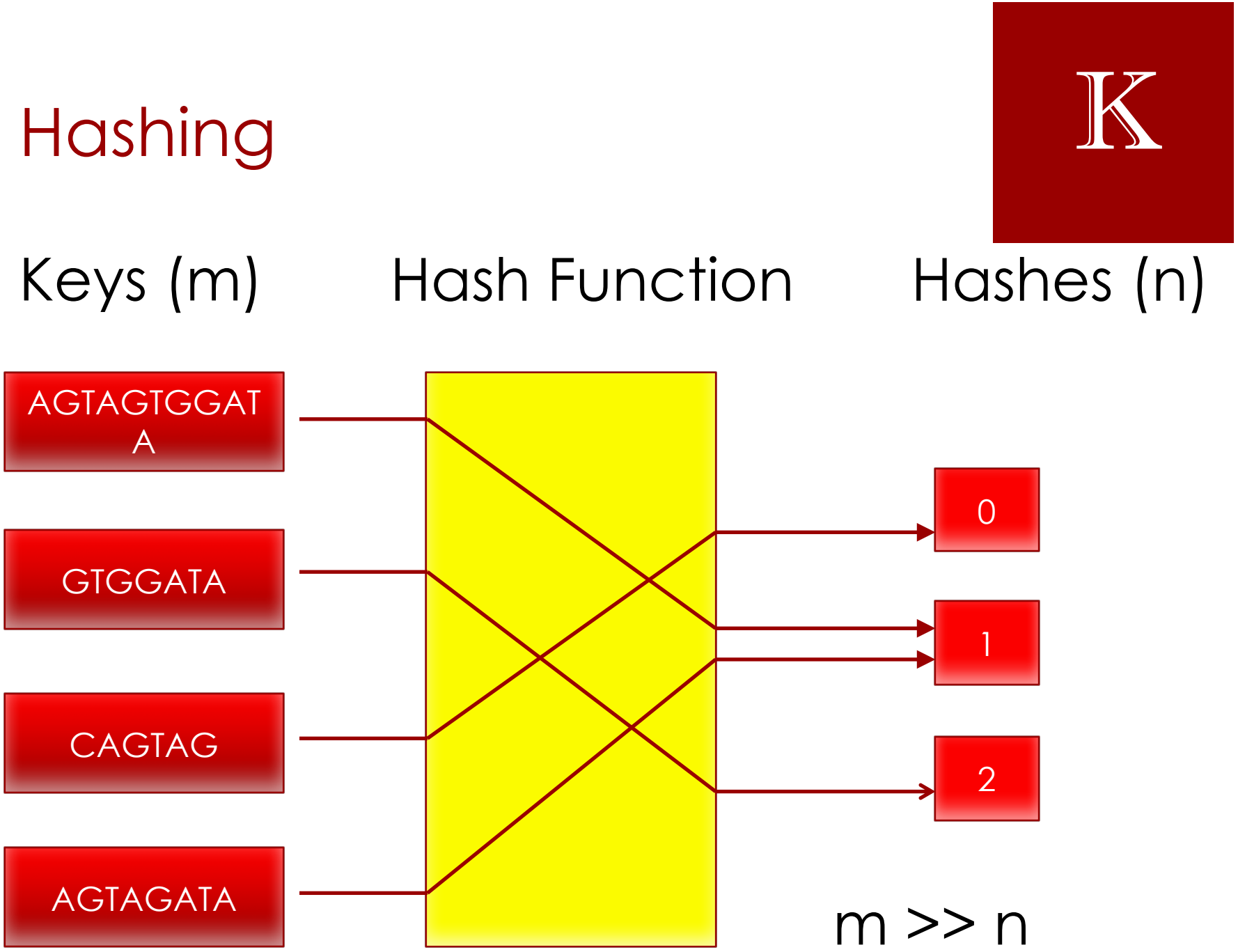
K-mer diversity analysis – excluding *k*-mers



K-mer diversity analysis – excluding *k*-mers



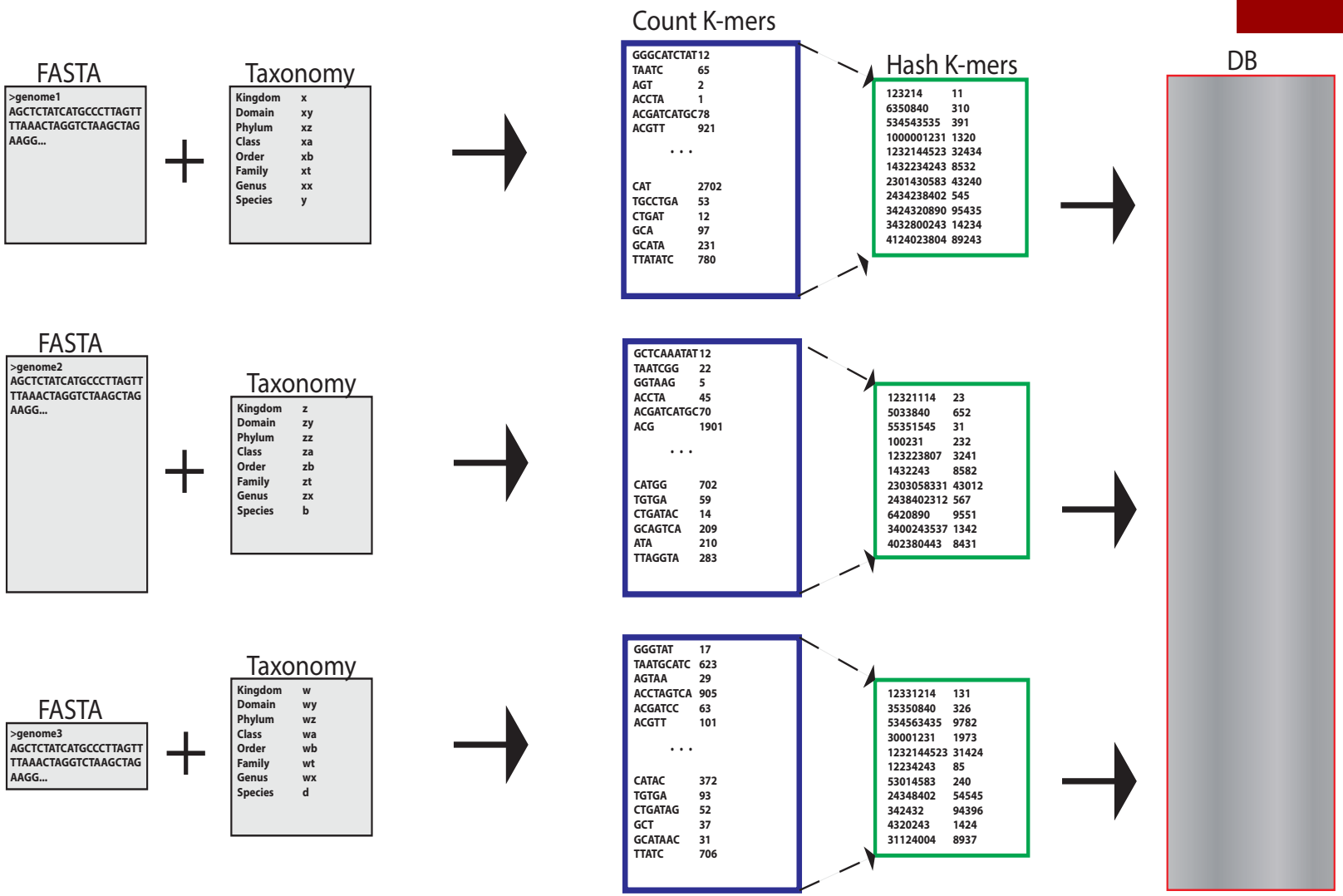
Hashing


K
Keys (m)
Hash Function
Hashes (n)
AGTAGTGGATA
GTGGATA
CAGTAG
AGTAGATA
0
1
2
m >> n

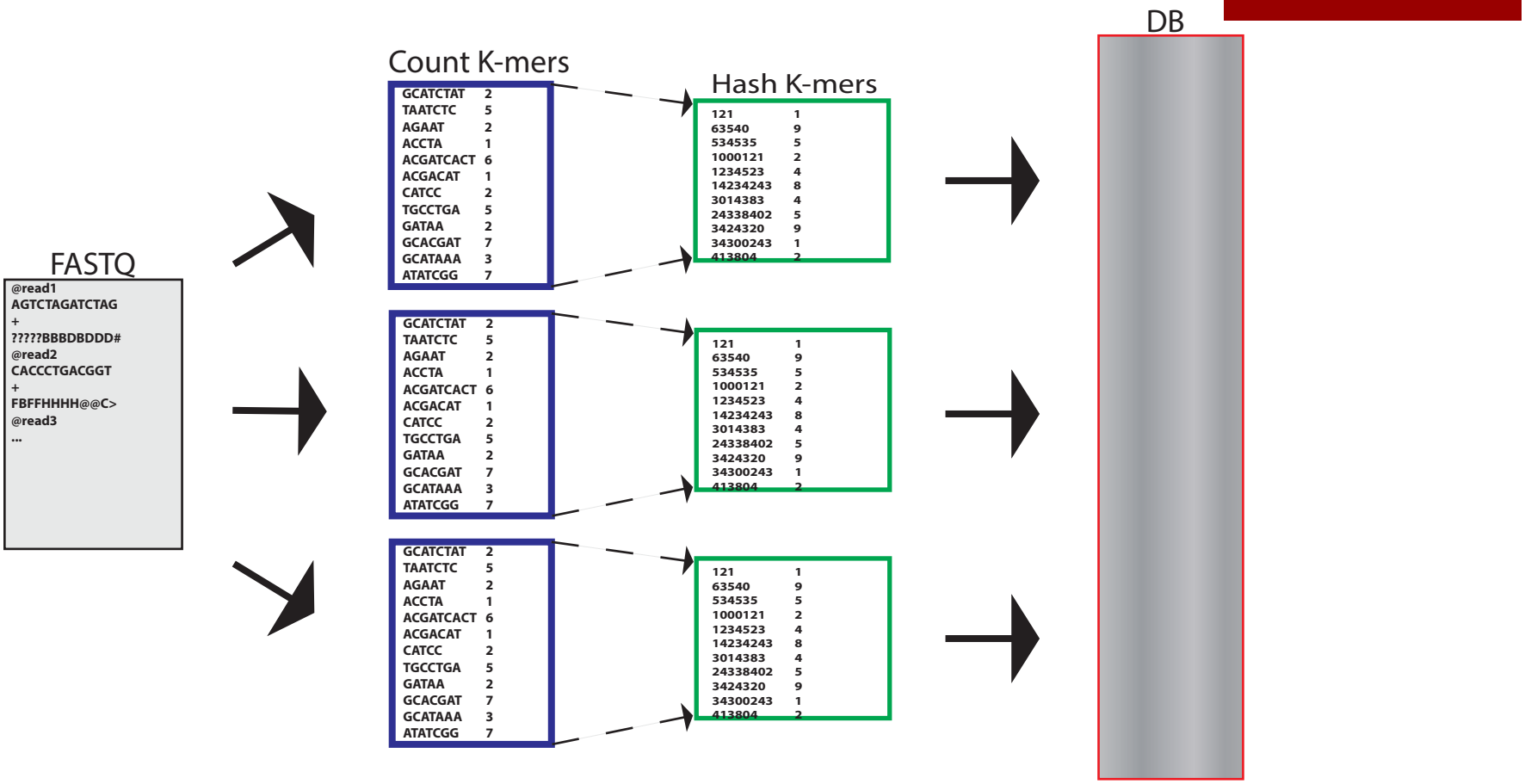
Data from KMerge

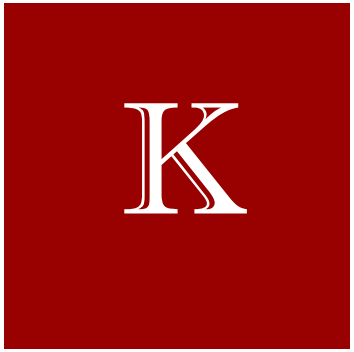
Sequence S_i	
Hash Val	**Count**
0	4
1	0
2	9
⋮	⋮
|H|-3	13
|H|-2	1
|H|-1	2

KMerge - reference

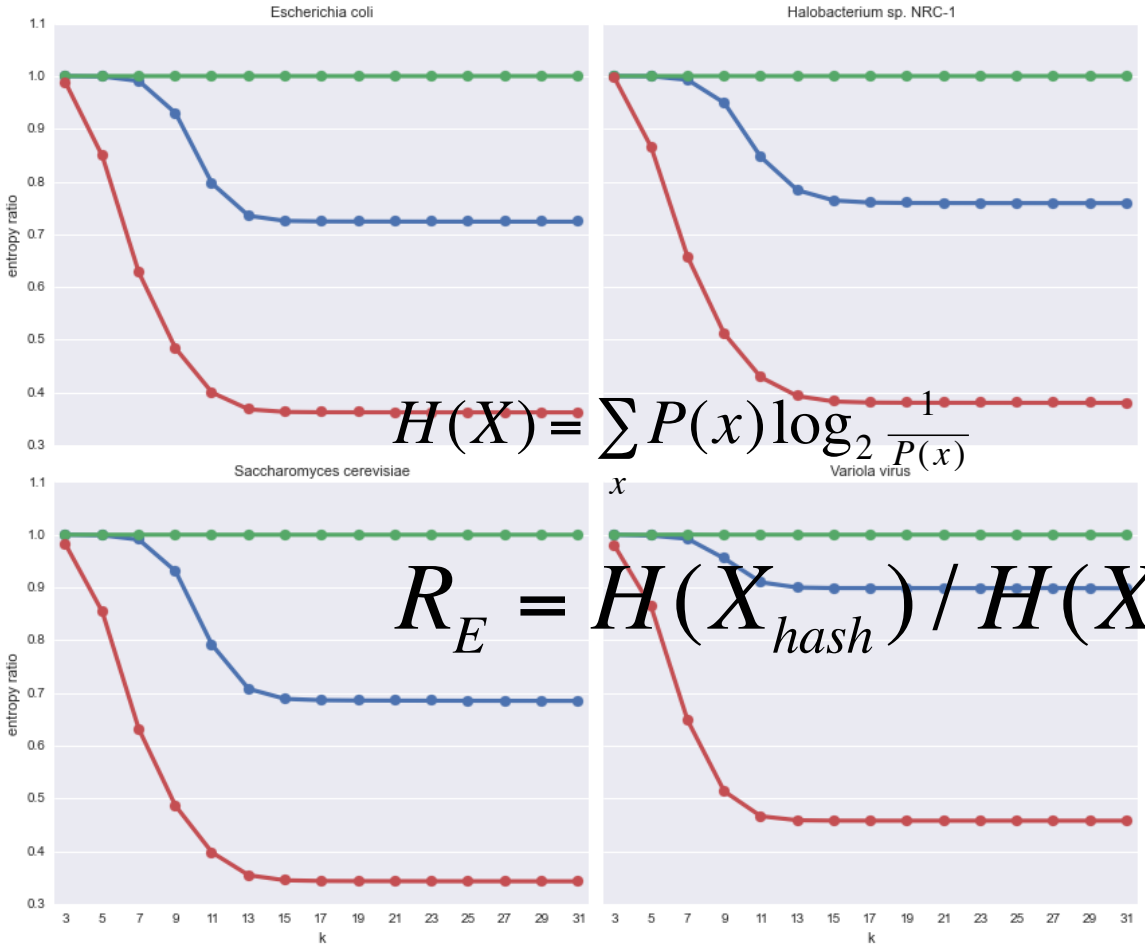


KMerge - sample





K-mer information content



$$H(X) = \sum_x P(x) \log_2 \frac{1}{P(x)}$$

8-bit
16-bit
32-bit

$$R_E = H(X_{hash}) / H(X_{raw})$$

Outline

I. Introduction

II. An alternative view of biological sequences

III. A model for sequence identification

IV. Conclusion/Future plans

Flipping Coins

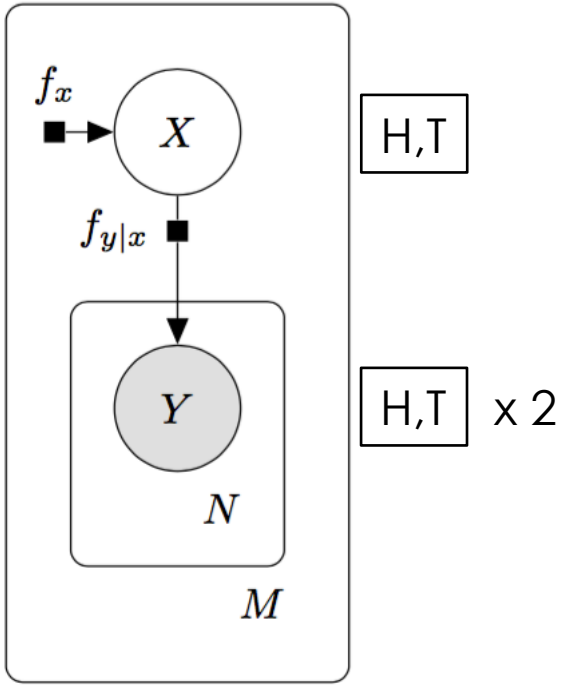
H = Flip Coin A

T = Flip Coin B

Flipping Coins

1	H	T	T	T	H	H	T	H	T	H
2	H	H	H	H	T	H	H	H	H	H
3	H	T	H	H	H	H	H	T	H	H
4	H	T	H	T	T	T	H	H	T	T
5	T	H	H	H	T	H	H	H	T	H

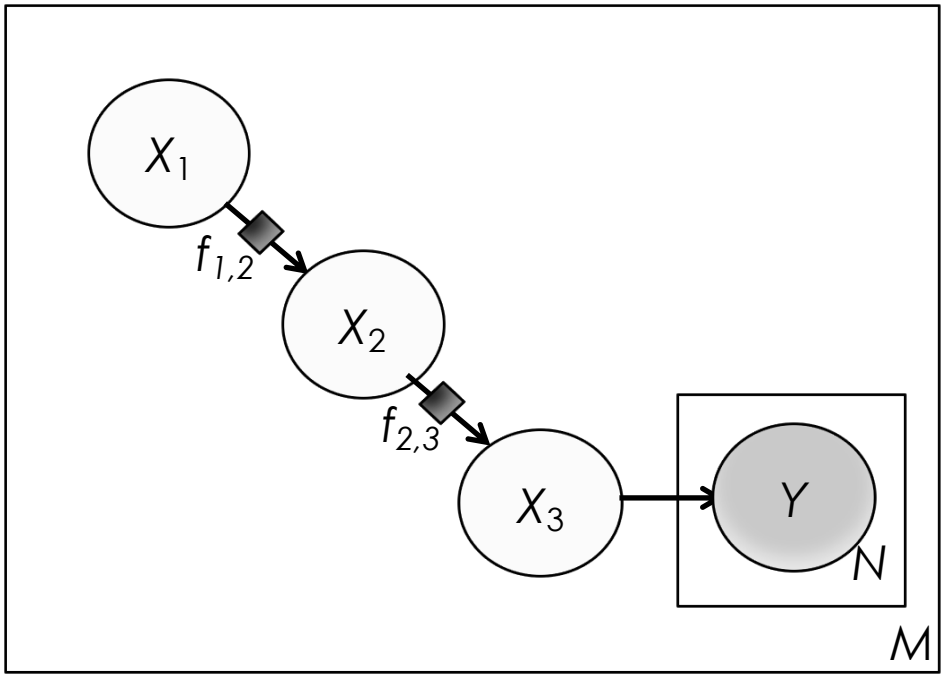
Flipping Coins



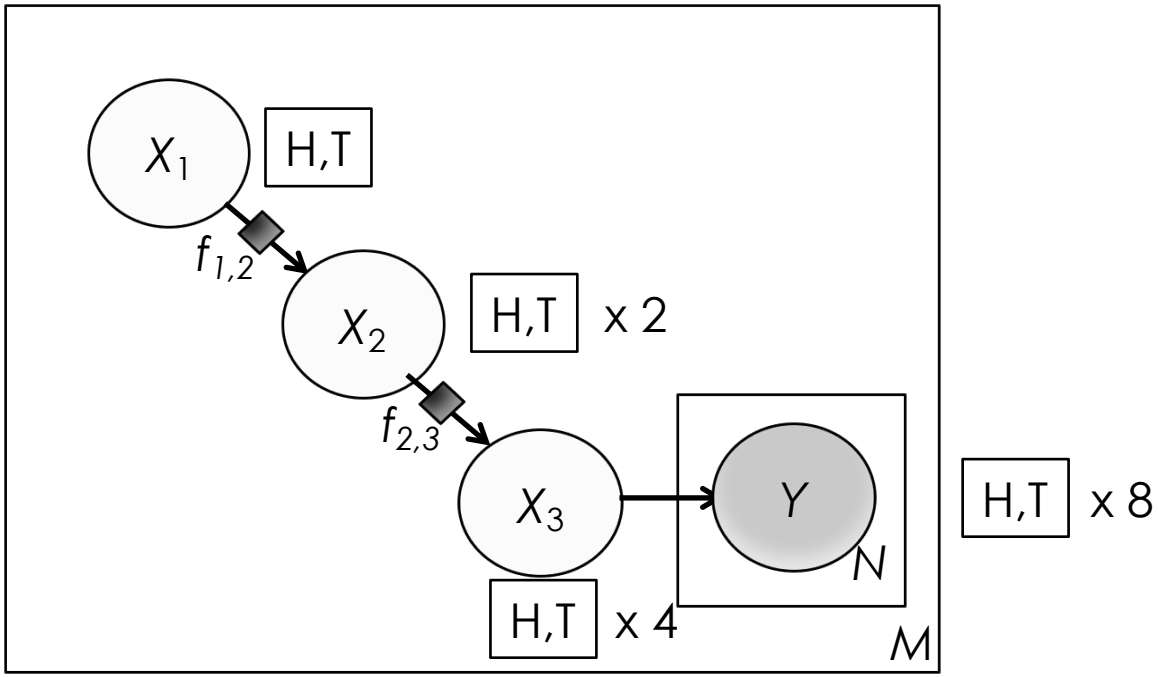
H,T

H,T x 2

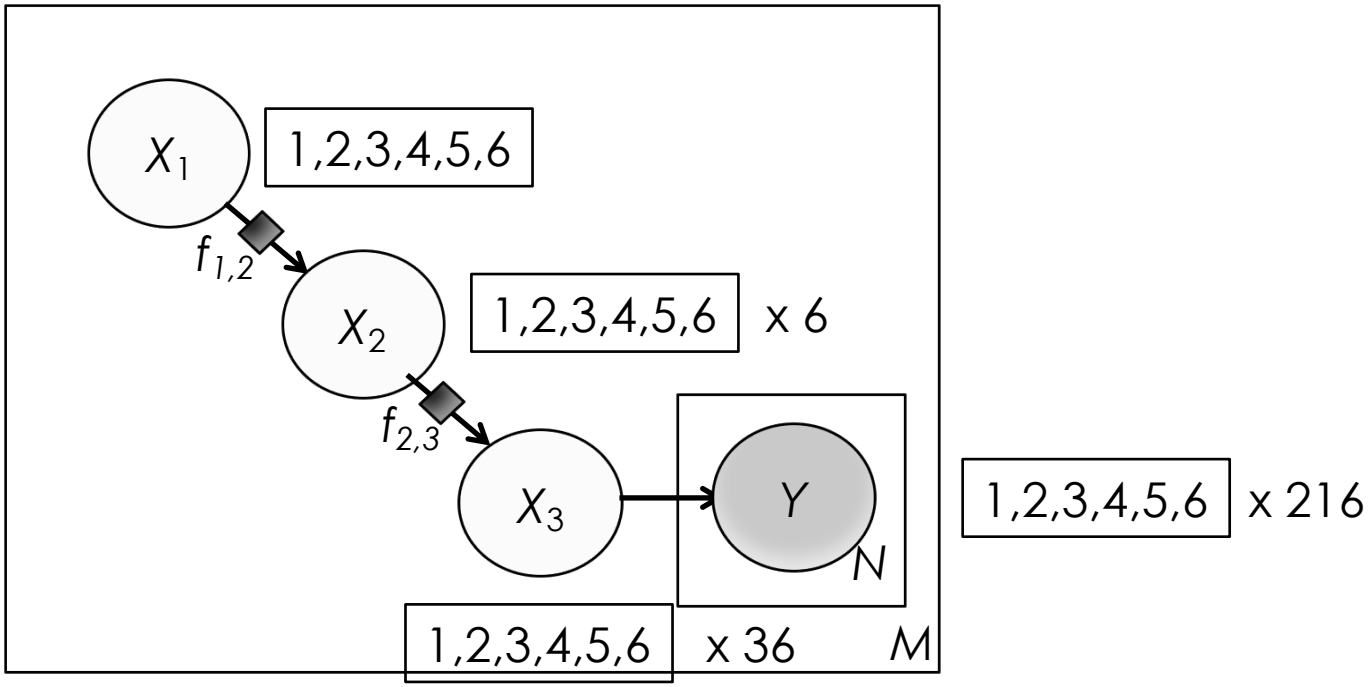
Flipping Multiple Coins



Flipping Multiple Coins



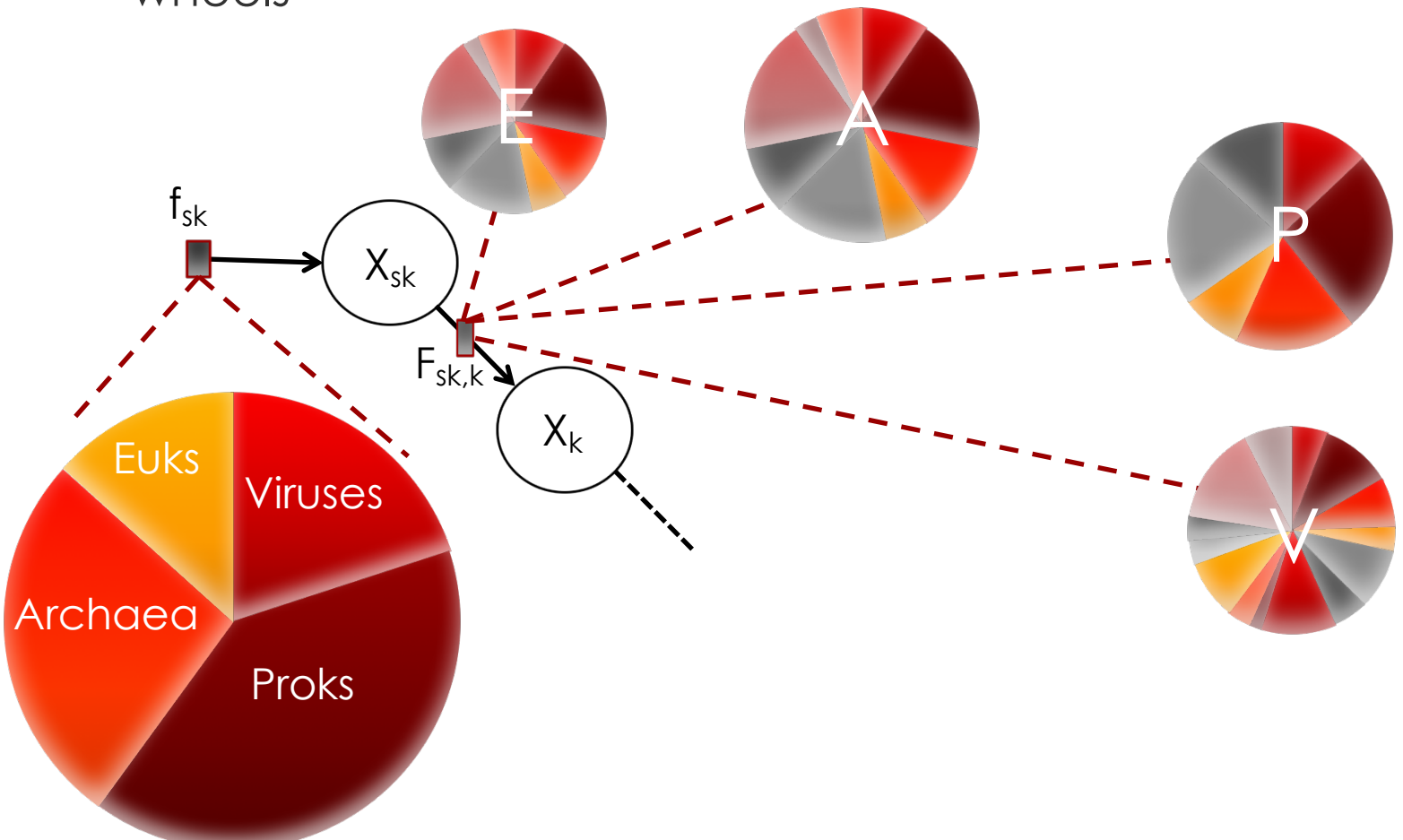
Rolling Dice



Generative model

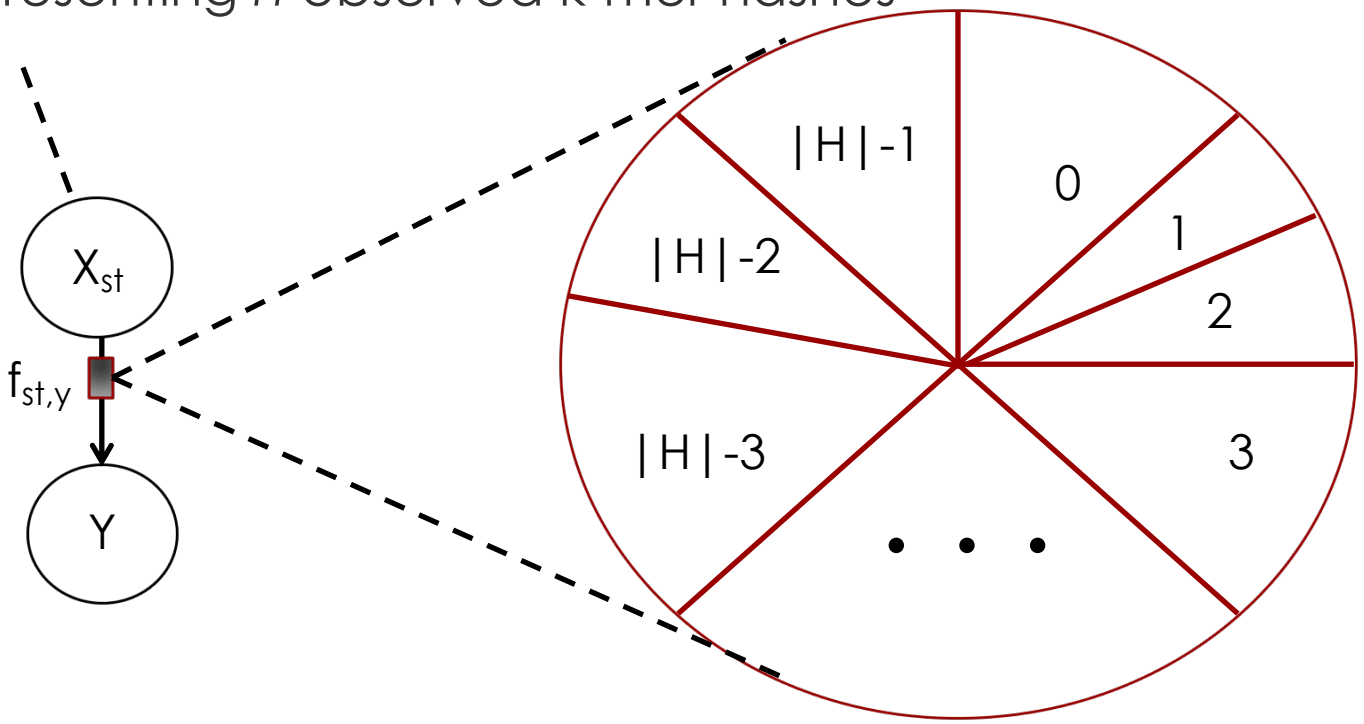
S

- Each factor can be envisioned as one or more sub-rank wheels

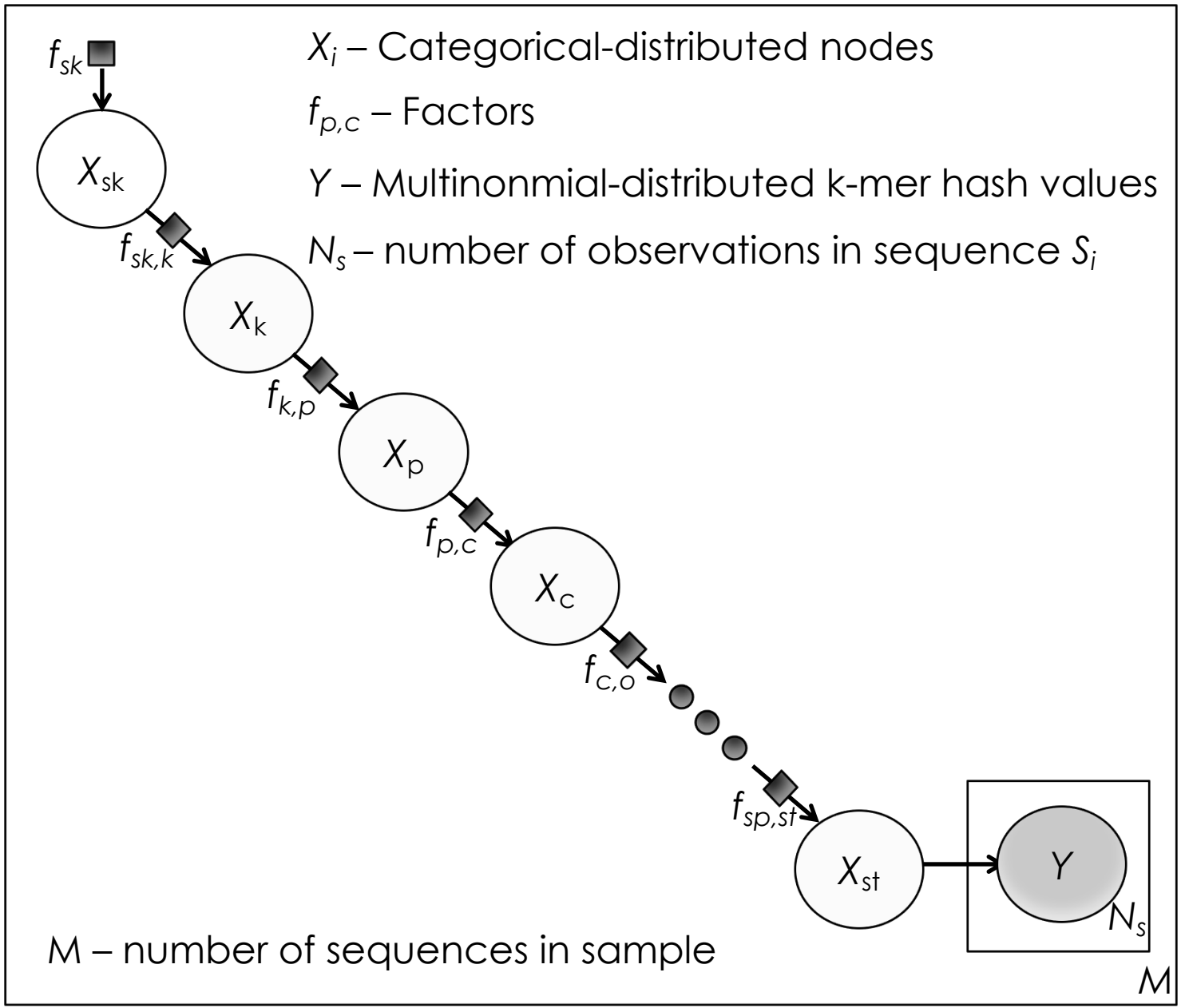


Generative process

- |H|- number of possible hash values
- Hash wheel for St^* is the distribution of hashes in genome
- Hash wheel spun $N_s = n$ times
- $Y_0,\ldots Y_n$ representing n observed k-mer hashes



Graphical model



X_i – Categorical-distributed nodes

$f_{p,c}$ – Factors

Y – Multinonmial-distributed k-mer hash values

N_s – number of observations in sequence S_i

M – number of sequences in sample

Generative process

S

Sequence S	
Hash Val	**Count**
0	4
1	0
2	9
⋮	⋮
|H|-3	13
|H|-2	1
|H|-1	2

Model inference

$$Z = \{X, f\}$$

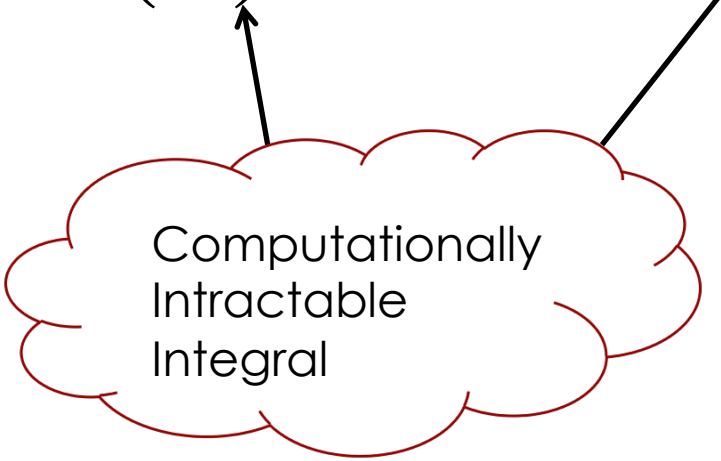
Z: hidden variables
X: rank variables
f: conditional probabilities

$$P(Z \mid Y) = ?$$

$$P(Z \mid Y) = \frac{P(Y \mid Z)P(Z)}{P(Y)}$$

Parameter Learning:

variational approximation



Outline

I. Introduction

II. An alternative view of biological sequences

III. A model for sequence identification

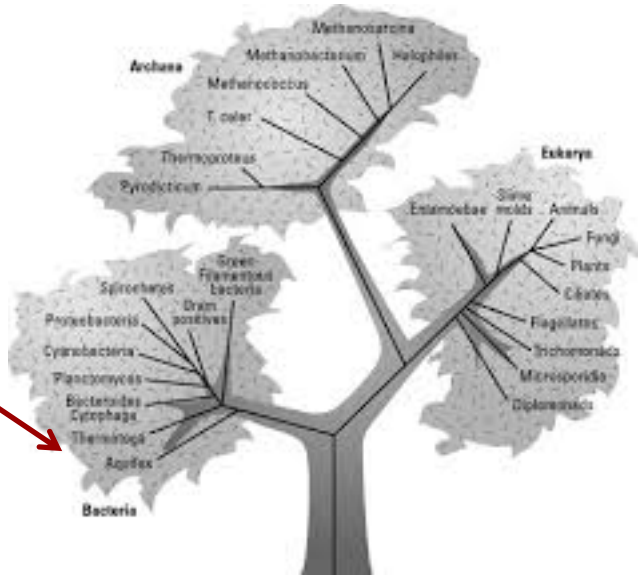
IV. Conclusion/Future plans

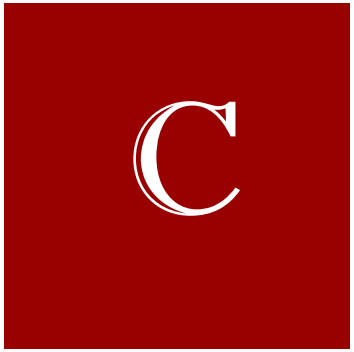
The Goal

S

DNA Sequence

KMerge
+
Model





Next Steps

- Implement complete model
- Scaling, tuning, and benchmarking method
- Web application/visualization platform
- Publish results

Acknowledgements

Mason Lab
Chris Mason, Ph.D.*
Dhruva Chandramohan, Ph.D.*
Cem Meydan, Ph.D.*
Elizabeth Hénaff,Ph.D.*
Francine Garrett-Bakelman, MD, Ph.D.
Virginia Wagatsuma, Ph.D.
Niamh O'Hara, Ph.D.
Bharath Prithiviraj, Ph.D.
Marjan Bozinoski, Ph.D.
Lenore Pipes, Ph.D.
Priya Vijay, Ph.D.
Pradeep Ambrose
Jake Reed
Alexa McIntyre
Noah Alexander
Sofia Ahsanuddin
Ebrahim Afshinnekoo
Chou Chou
Maureen Milici

PhD Survival Team
Noah Dukler
Tom Vincent, Ph.D.
Andre Kahles, Ph.D.
Jaakko Luttinen, Ph.D.
PBTech

Tri-I CBM
David Christini, Ph.D.
Christina Leslie, Ph.D.
Margie Hinonangan-Mendoza
Kathleen Pickering
Francine Collazo-Espinell

Thesis Committee
Olivier Elemento, Ph.D.*
Gunnar Rätsch, Ph.D.
Jason Mezey, Ph.D.

Funding
NIH Ruth L. Kirschtein
National Research Service
Award (F31GM111053)