Bioinformatics
A Communication/Signal Processing Perspective

Khalid Sayood
Occult Information Lab
Department of Electrical and Computer Engineering
University of Nebraska-Lincoln

Outline

- Beginning
- Middle
- End

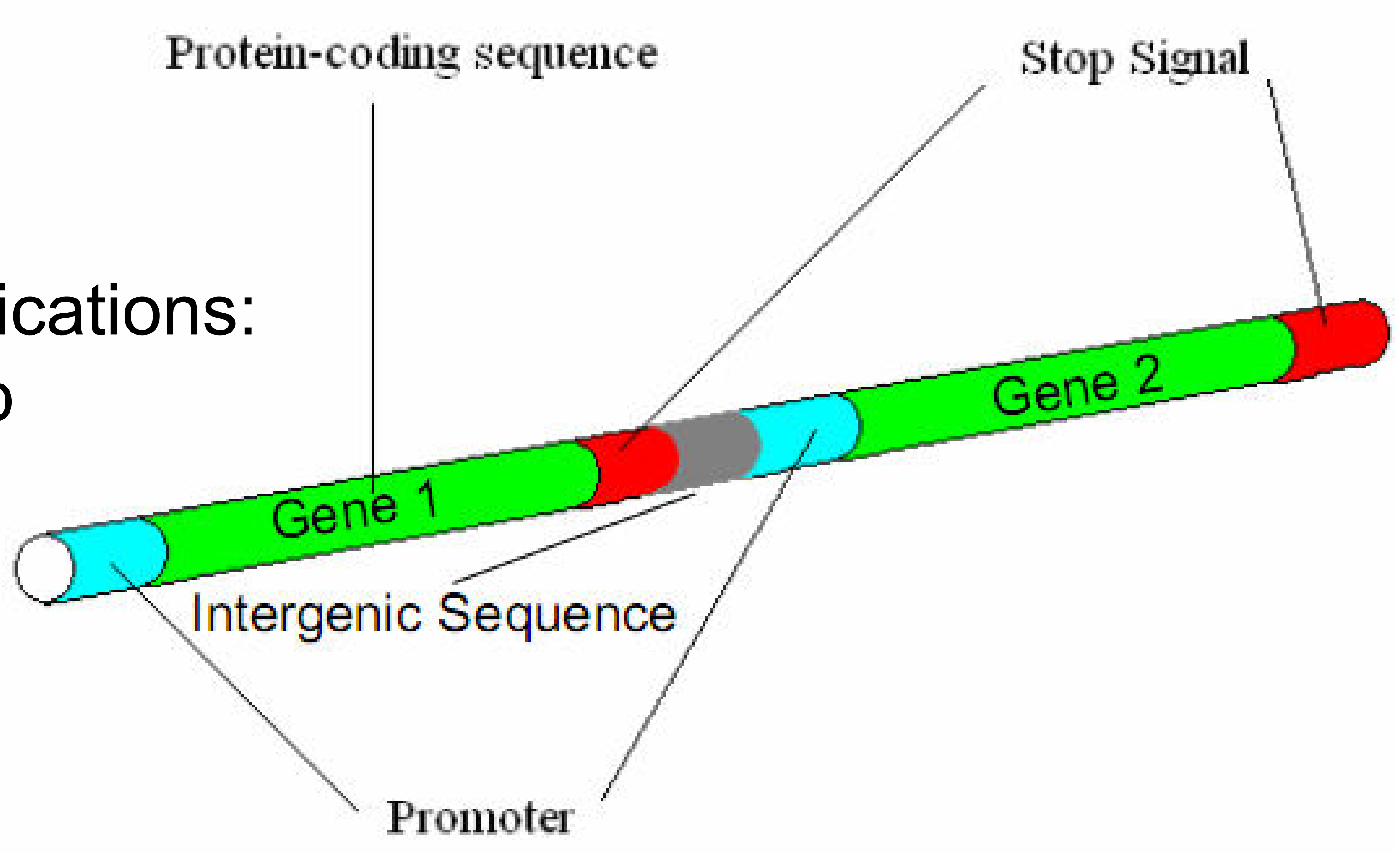
My roots

- Signal Processing
- Communication
- Information Theory
- Data Compression

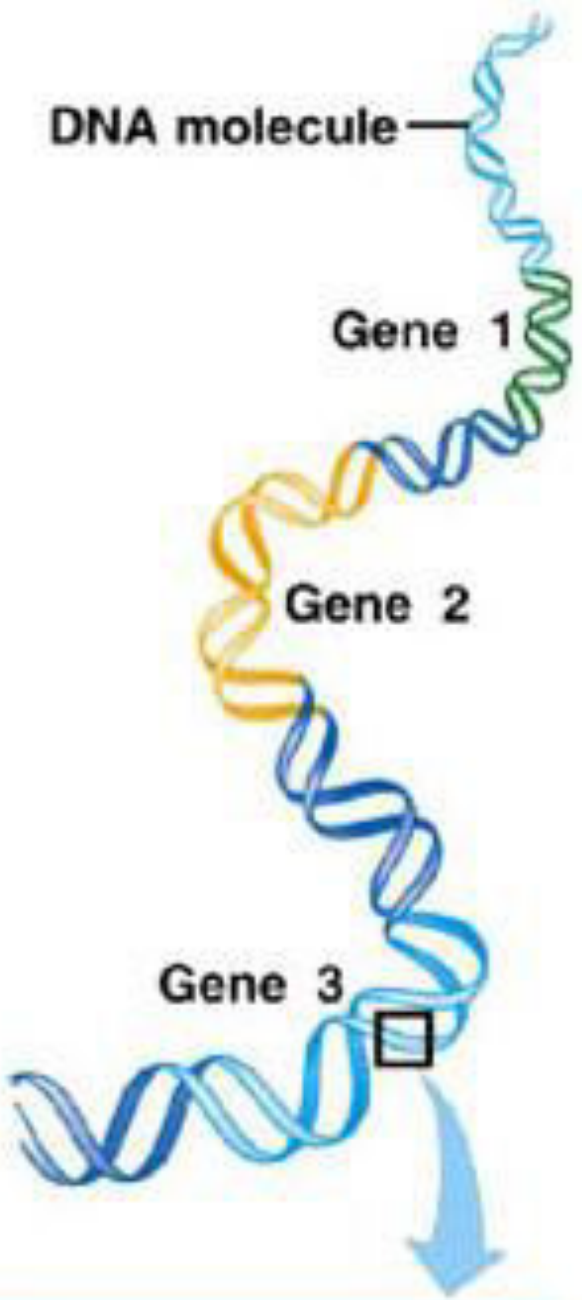
My roots

- Data Compression - The art or science of finding compact representations of information in data
  - Speech → Images → Video → DNA

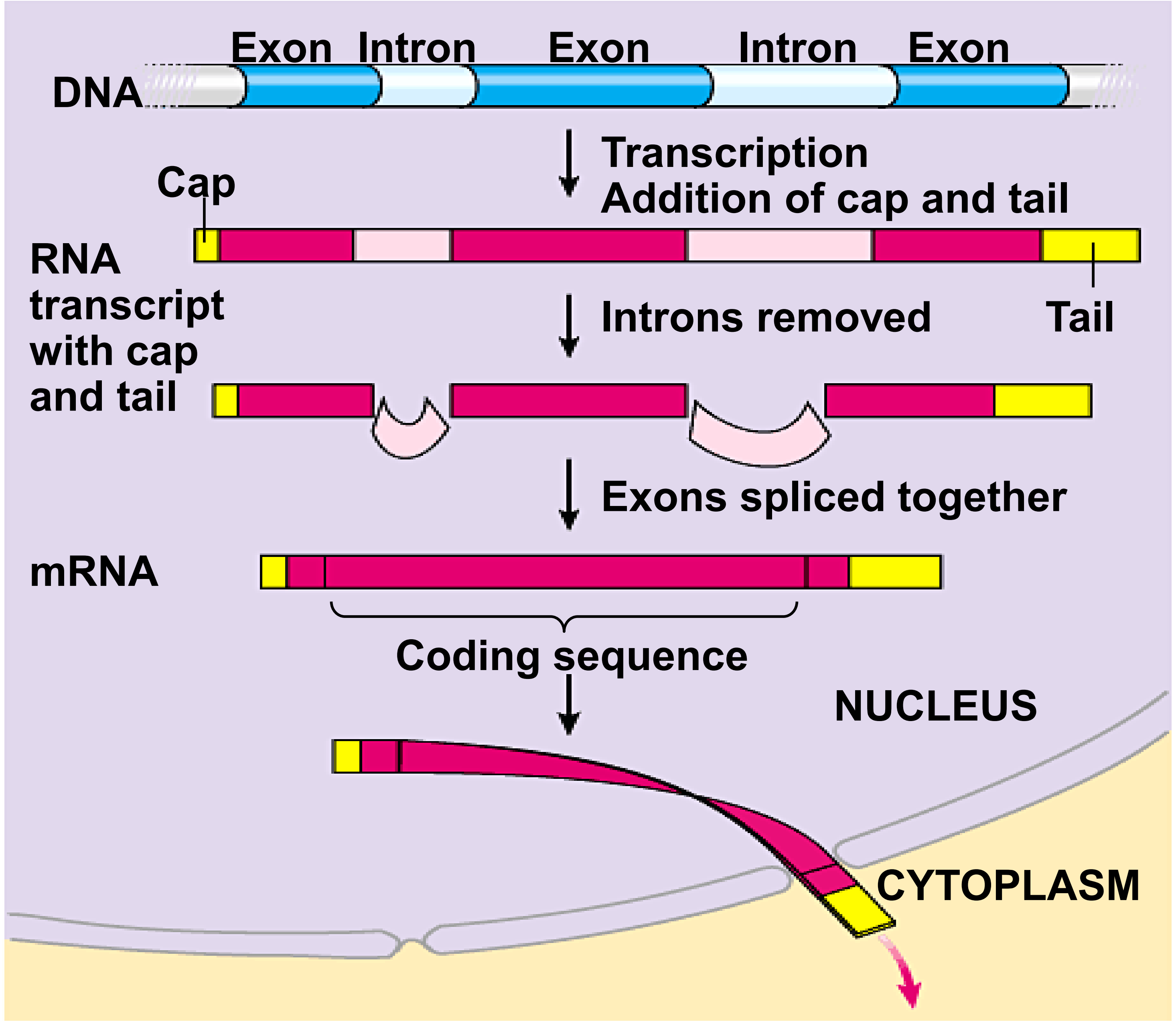
Prokaryotes

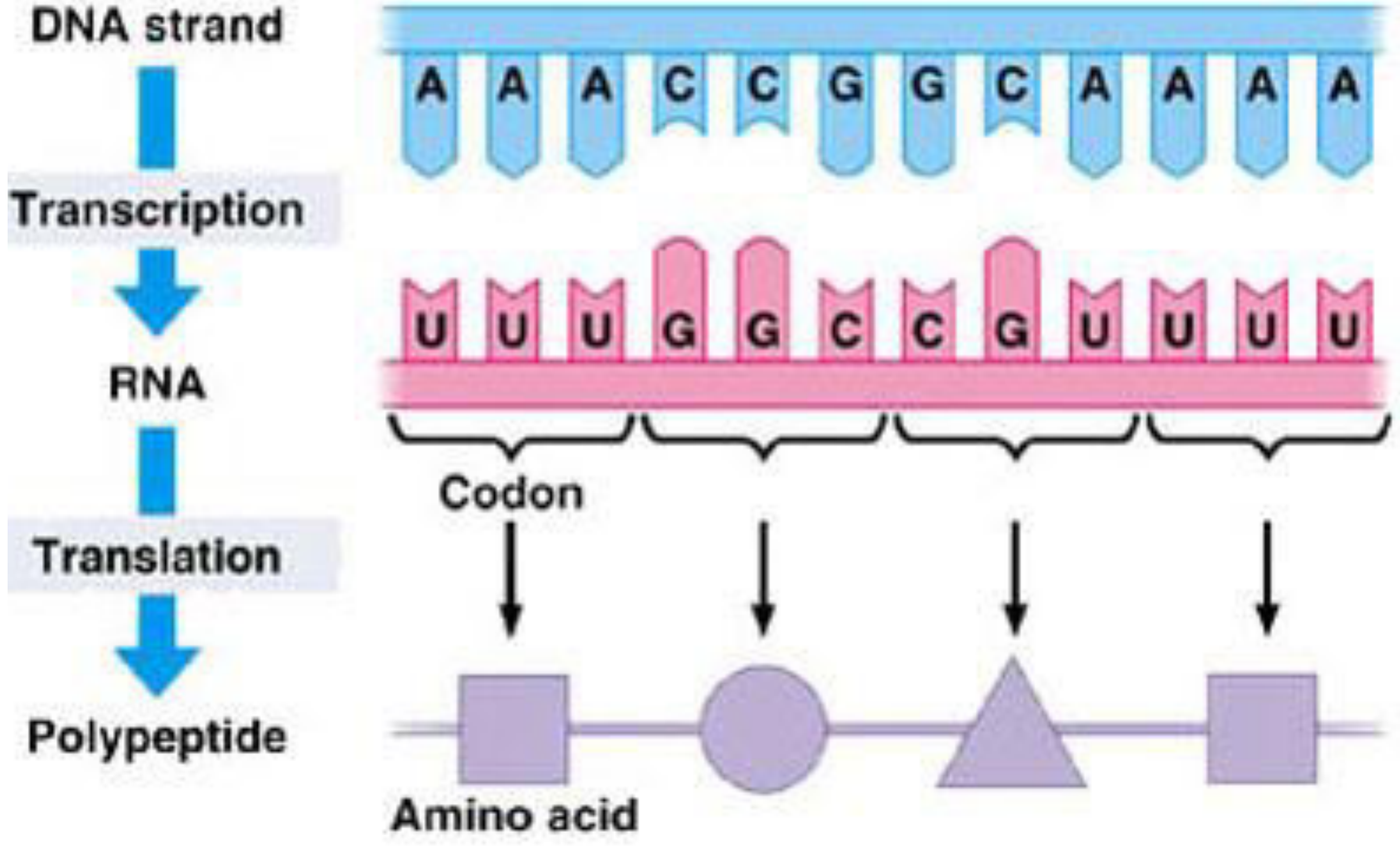


Analogy to Communications:
Start, Message, Stop



Eukaryotes





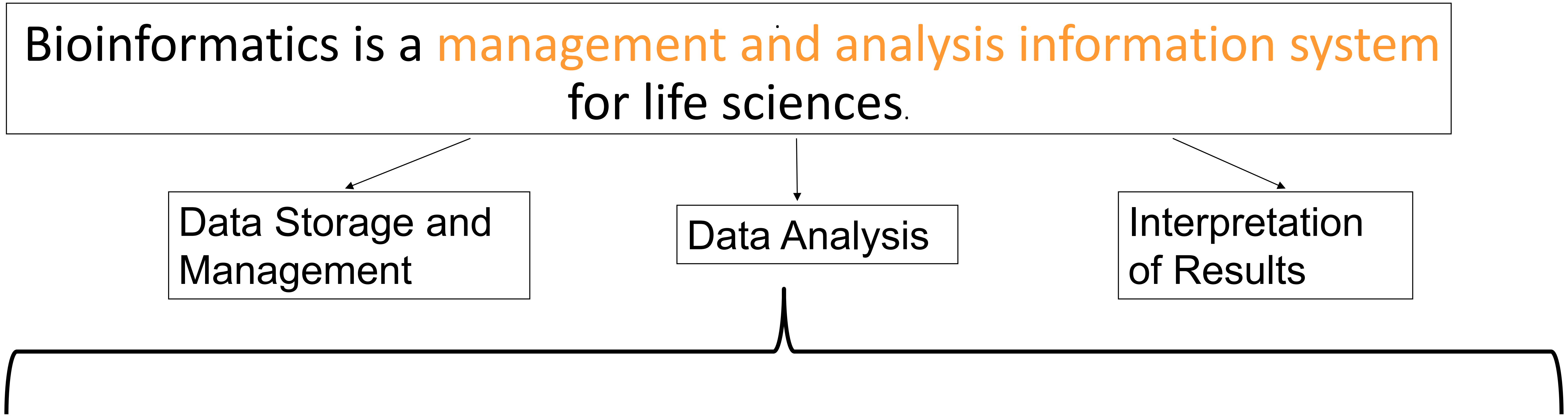
Definitions of Bioinformatics

- Original definition: Study of informatic processes in biotic systems – Pauline Hogeweg and Ben Hesper

Definitions of Bioinformatics

- Original definition: Study of informatic processes in biotic systems – Pauline Hogeweg and Ben Hesper

- Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **“informatics” techniques** (derived from disciplines such as applied math, CS, and statistics) to understand and **organize the information associated** with these molecules, **on a large-scale**. (Mark Gerstein, 1999)



Bioinformatics is a management and analysis information system for life sciences.

Data Storage and Management

Data Analysis

Interpretation of Results

Protein Structure Prediction

- Protein/RNA tertiary structure
- Docking
- Drug Design

Molecular Sequence Analysis

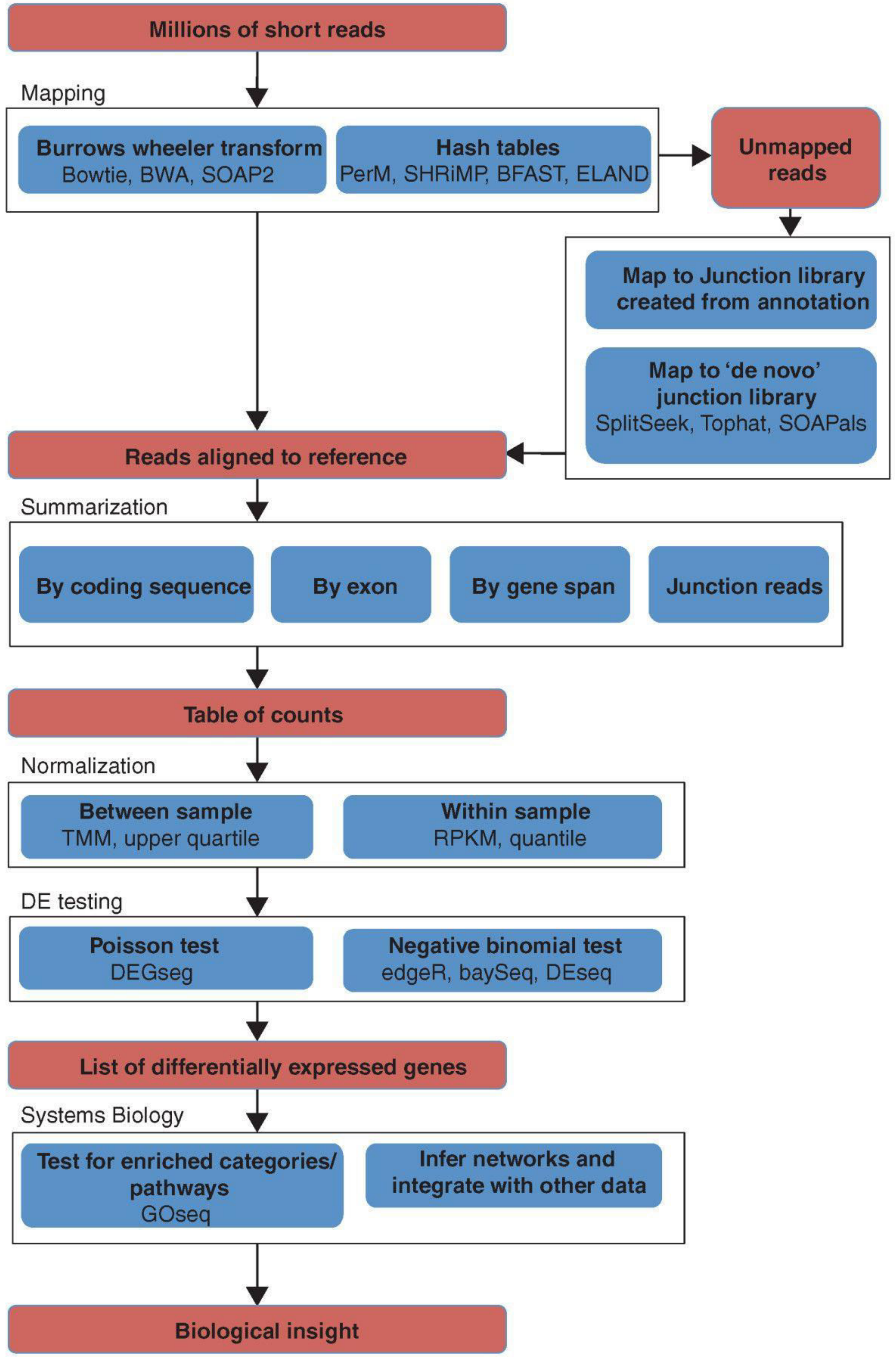
- Homology Search
- Phylogeny Construction
- Whole Genome Sequencing
- Gene Finding

Functional Genomics and Proteomics

- Microarrays
- Biomarker Discovery

Systems Biology

- Pathways
- Network based wholistic approach



Millions of short reads

Mapping

Burrows wheeler transform
Bowtie, BWA, SOAP2

Hash tables
PerM, SHRiMP, BFAST, ELAND

Unmapped reads

Map to Junction library created from annotation

Map to 'de novo' junction library
SplitSeek, Tophat, SOAPals

Reads aligned to reference

Summarization

By coding sequence

By exon

By gene span

Junction reads

Table of counts

Normalization

Between sample
TMM, upper quartile

Within sample
RPKM, quantile

DE testing

Poisson test
DEGseg

Negative binomial test
edgeR, baySeq, DEseq

List of differentially expressed genes

Systems Biology

Test for enriched categories/ pathways
GOseq

Infer networks and integrate with other data

Biological insight

From RNA-seq reads to differential expression results

- Alicia Oshlack, Mark D Robinson andMatthew D Young

Genome Biology2010**11:**220

- Preprocessing:
  - *FastQC*
  - *Trimmomatic*
- Indexing the Reads:
  - *Bowtie 2*
- Aligning to hg19/KSHV
  - *TopHat 2*
    - Local alignment
- Extracting Features
  - *CuffDiff*
    - $P < 0.01$
    - FPKMs values
    - Differentially Expressed Genes

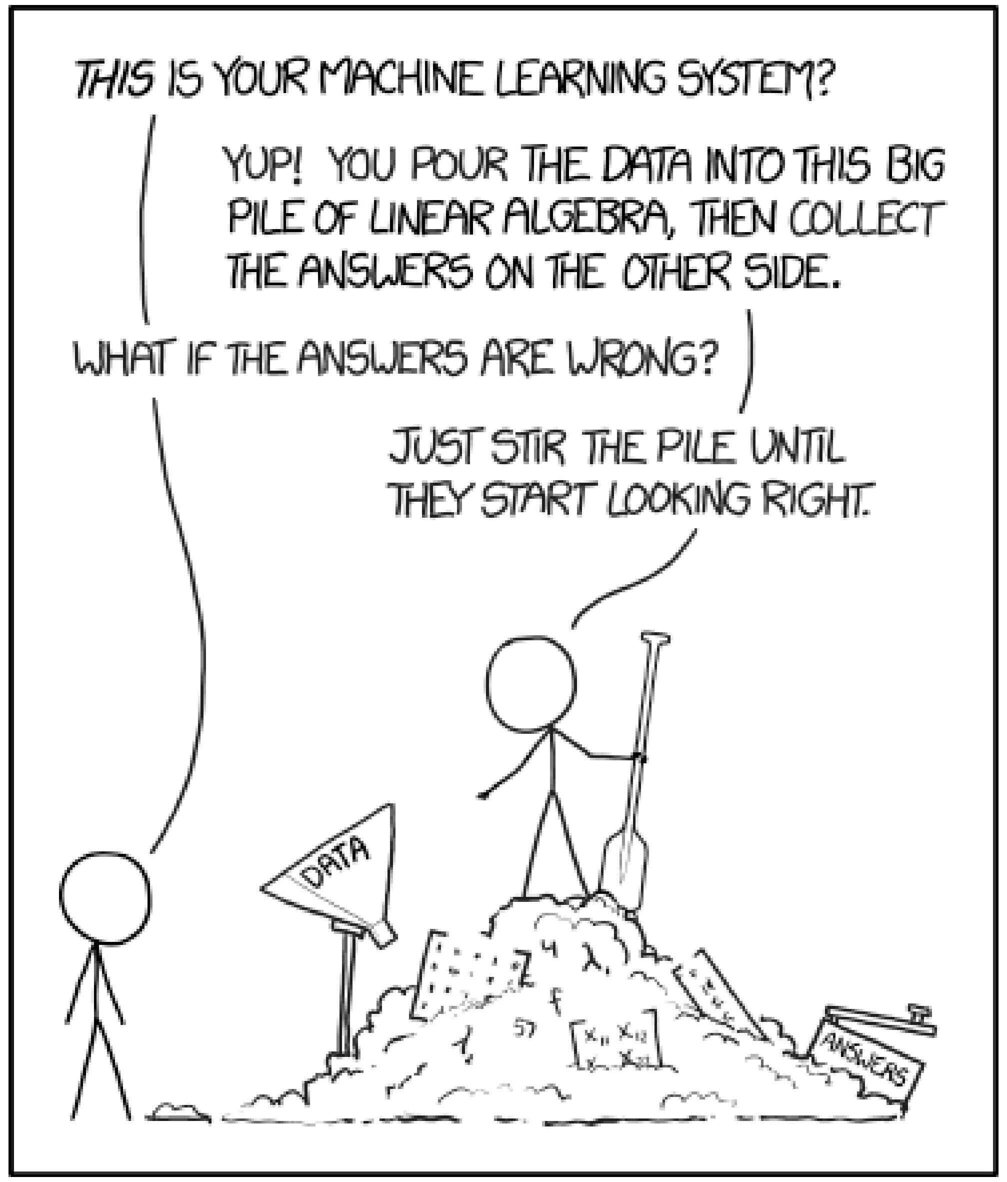
UniProt
Universal Protein Resource
prosite
Fasta
Similarity and Homology Searching
WU-Blast2
Basic Local Alignment Search Tool
IntEnz
SeqVISTA
repeat region
Pfam
PRINTS
Protein Fingerprints Database
InterPro
CAST
GENOMES
ClustalW
Multiple Sequence Alignment Tool
emboss
ProDom
SMART
Alternative Splicing Database
Integr8
CSA
Catalytic Site Atlas
PDBj
Protein Data Bank Japan
TIGR
tigr fams
PANTHER
Classification System
caBIG
SeqHound
PIRSF
Gene3D
BioLayout
NCBI
LION
Superfamily
SCOP
DALI
GeneQuiz
PubMed.gov
U.S. National Library of Medicine
National Institutes of Health
caBIO
MOUNT SINAI HOSPITAL
CATH
Protein Structure Classification
MSD
e! Ensembl
Parasite Blast
DDBJ
DNA Data Bank of Japan
wellcome trust sanger institute
SWISS-MODEL
ModBase
GENOME REVIEWS
IUPHAR
KEGG
Kyoto Encyclopedia of Genes and Genomes
PathPort
The Pathogen Portal Web Project

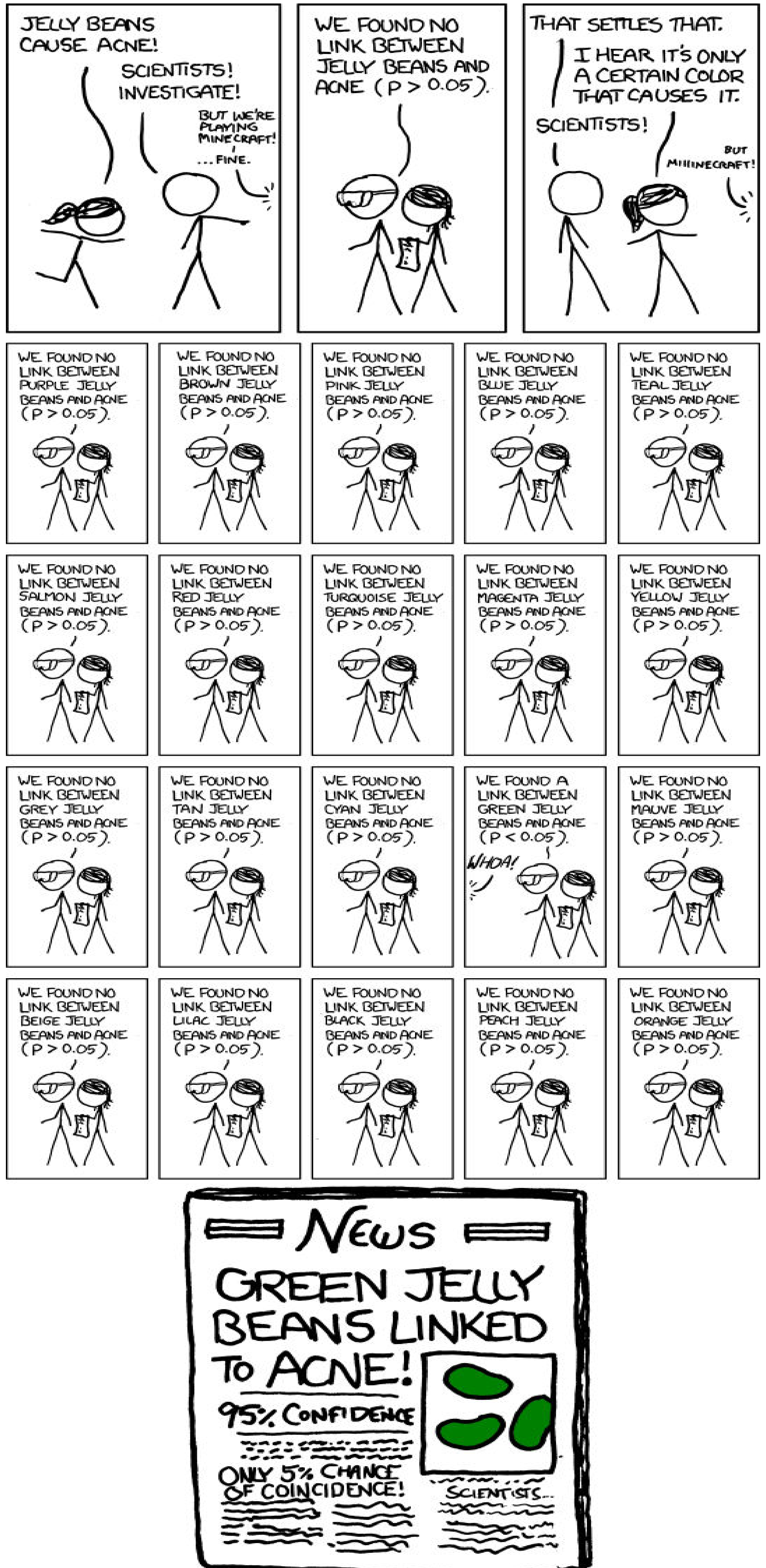
Apache Taverna

Wageningen University and Research

So what is the problem?

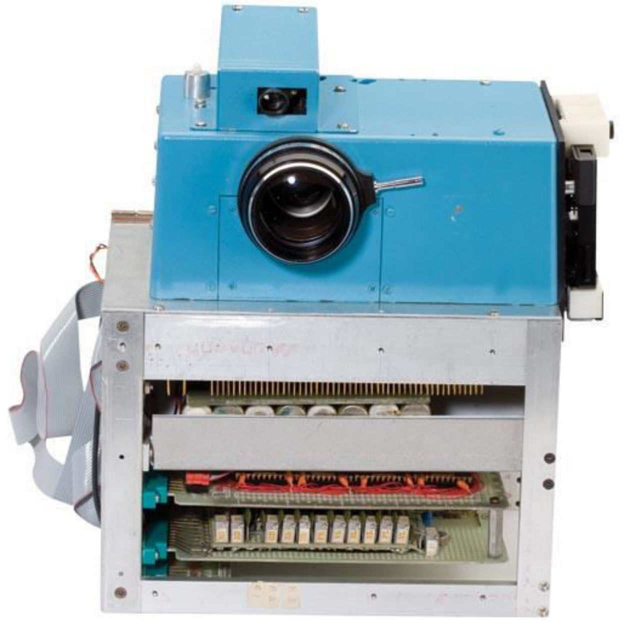
- What is actually going on.


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Is there hope (for me)?

Hope springs eternal



Digital Camera (1975)

Genome Sequencer (2018)

DON'T WORRY - TECHNOLOGY WILL SAVE YOU.
MUELLER

What do we mean by a Communication Theory Perspective

What do [redacted] I mean by a Communication Theory Perspective

Information exists in the form of a stochastic process

How do we deal with stochastic processes?

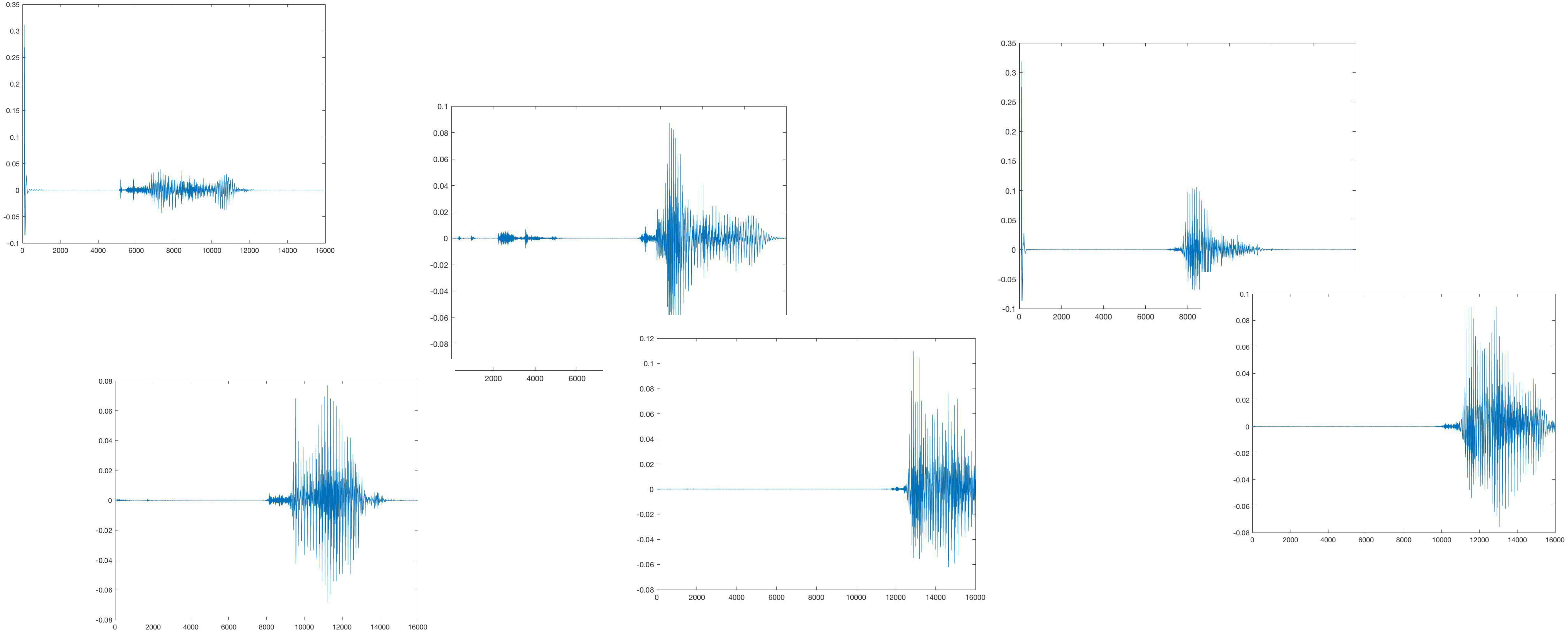
- Look at the signal using different basis sets – frequency domain processing.
- Look at correlation structures.
- Look at models.

All these involve averaging of some sort

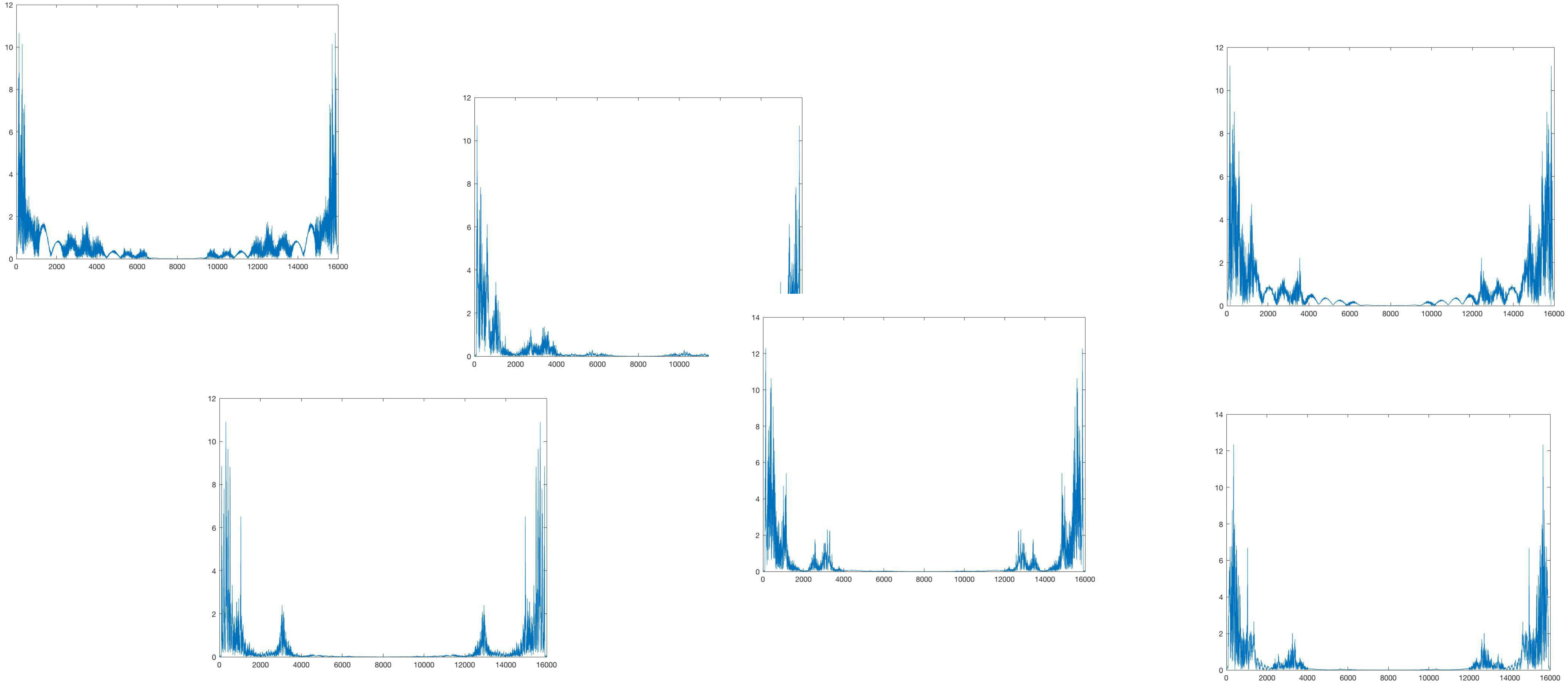
All these result in the discovery of underlying structure

They can also result in dimensionality reduction.

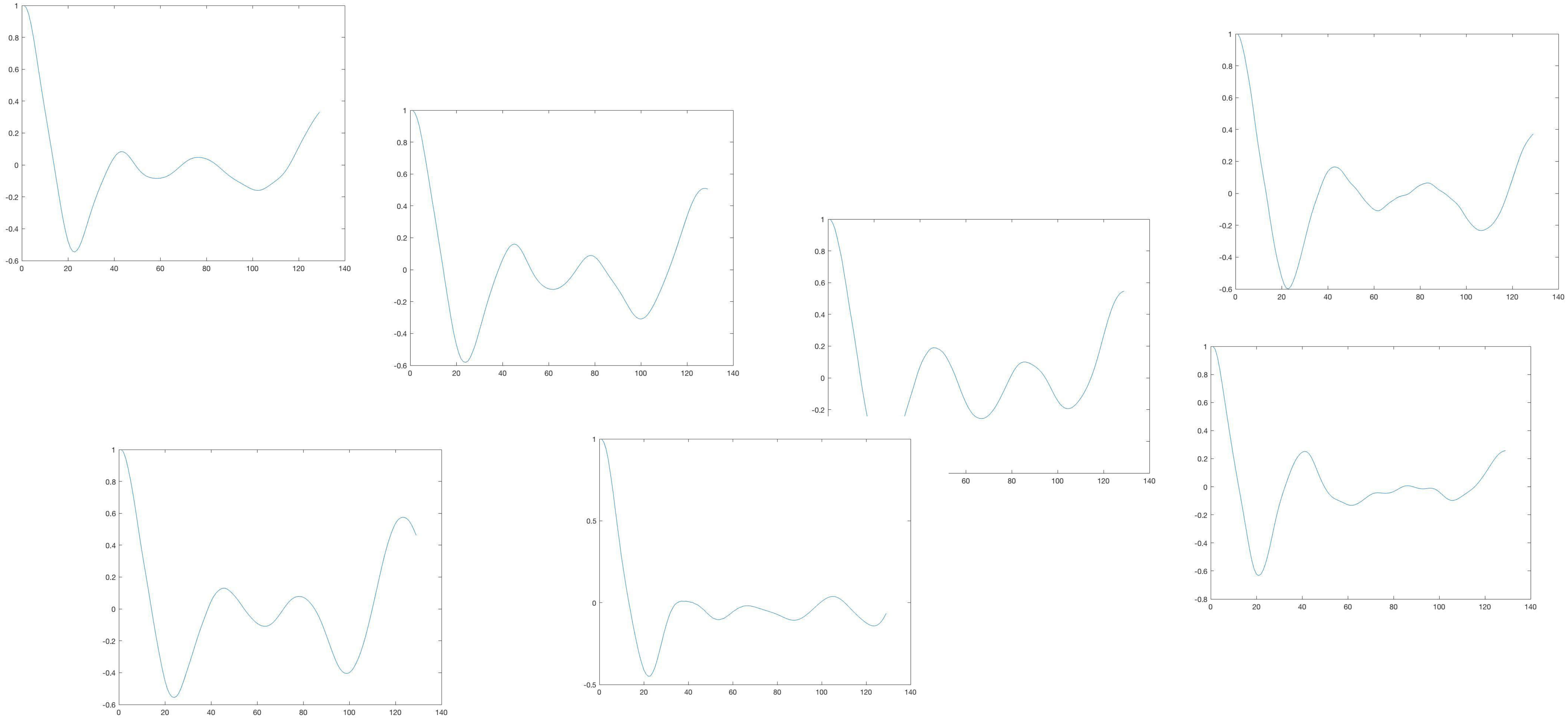
Realizations of a stochastic process



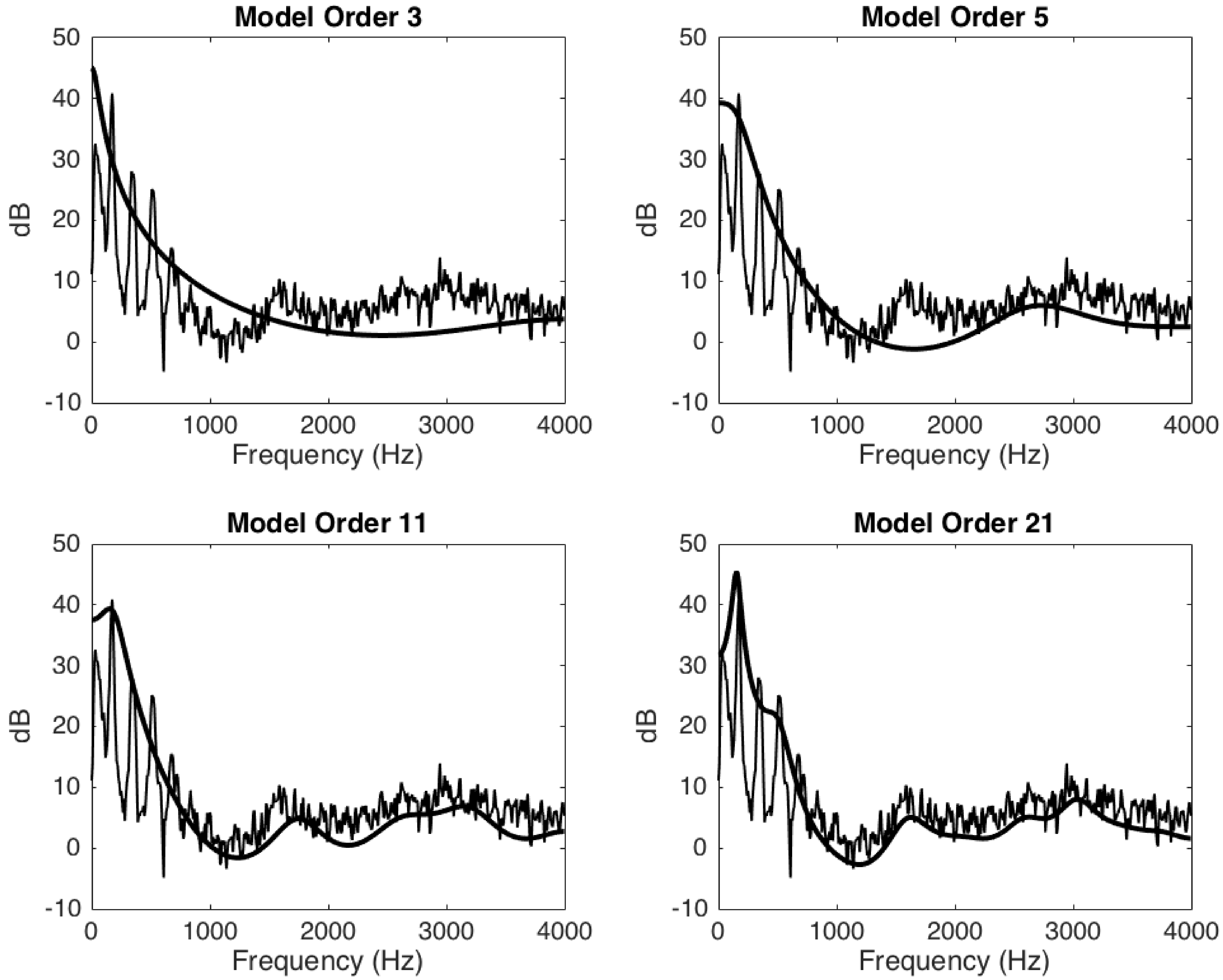
Frequency profile



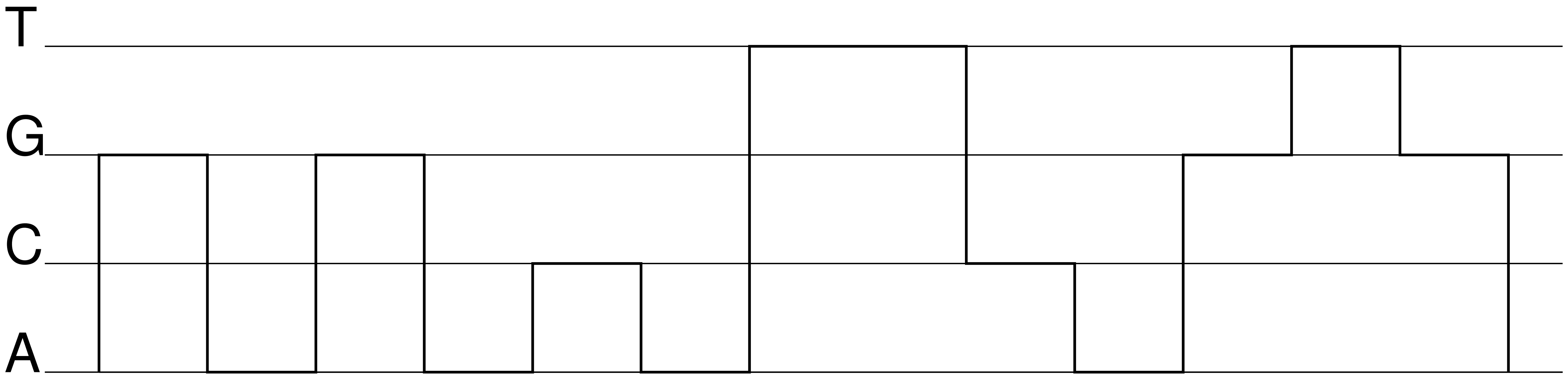
Statistical Profile



Models

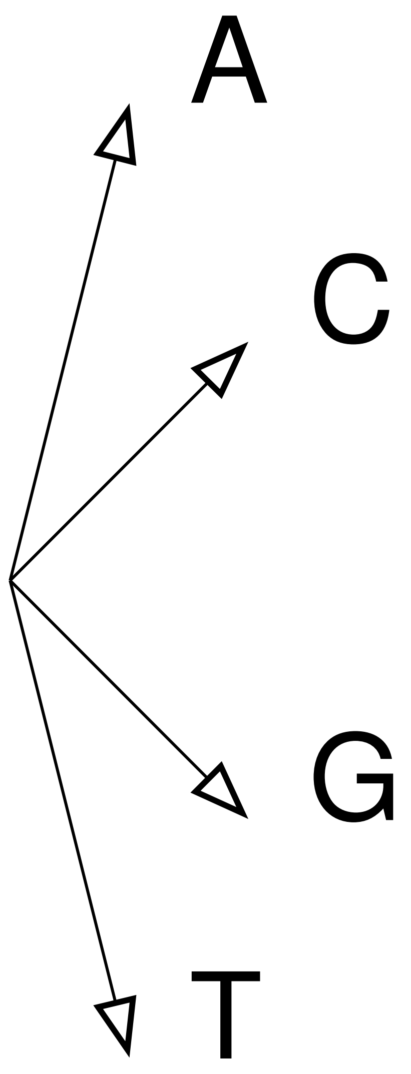

Model Order 3
Model Order 5
Model Order 11
Model Order 21
dB
50
40
30
20
10
0
-10
0
1000
2000
3000
4000
Frequency (Hz)

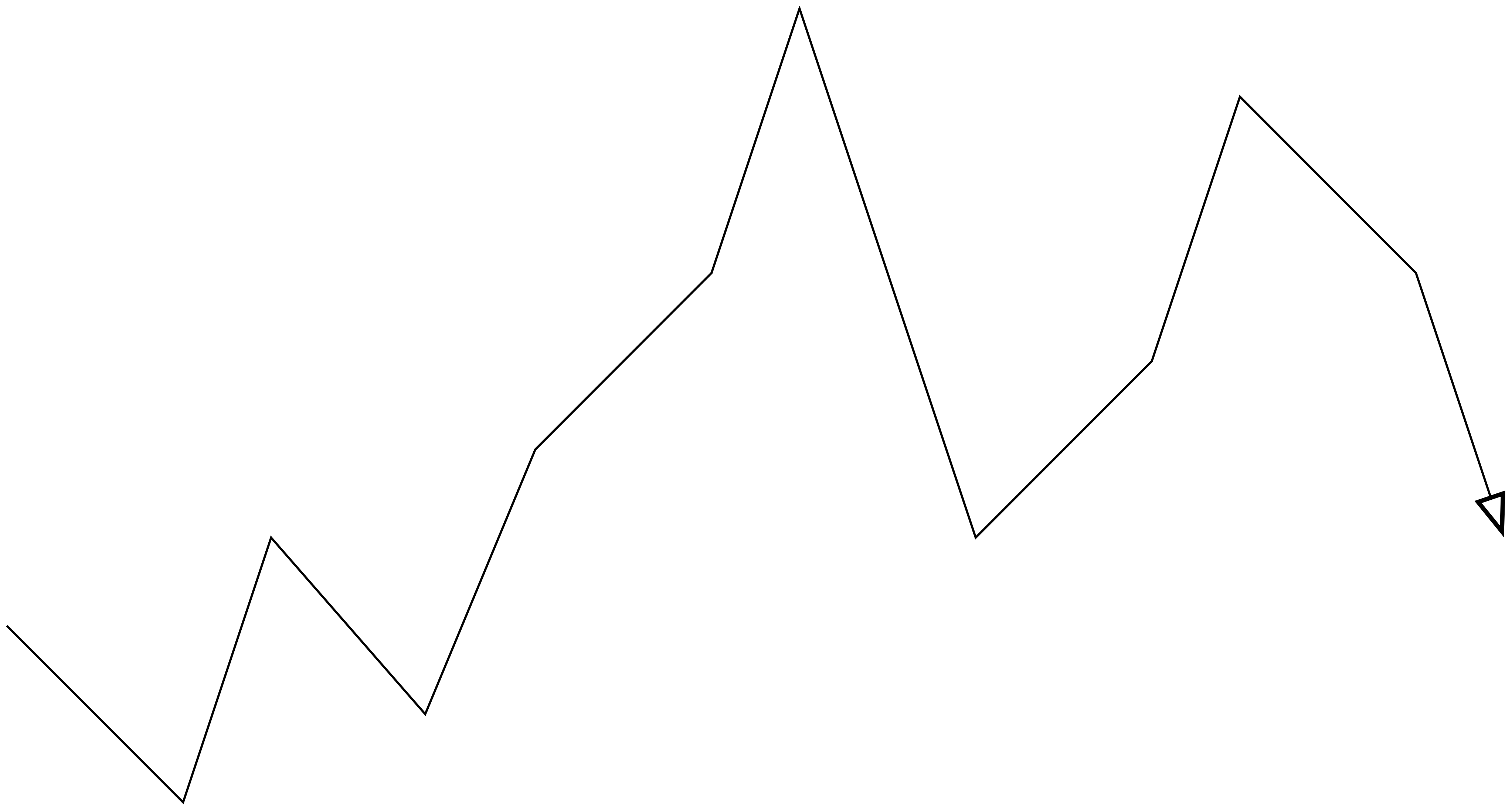
But we have sequences – we don’t have numbers



GAGACATTCAGTG

But we have sequences – we don’t have numbers





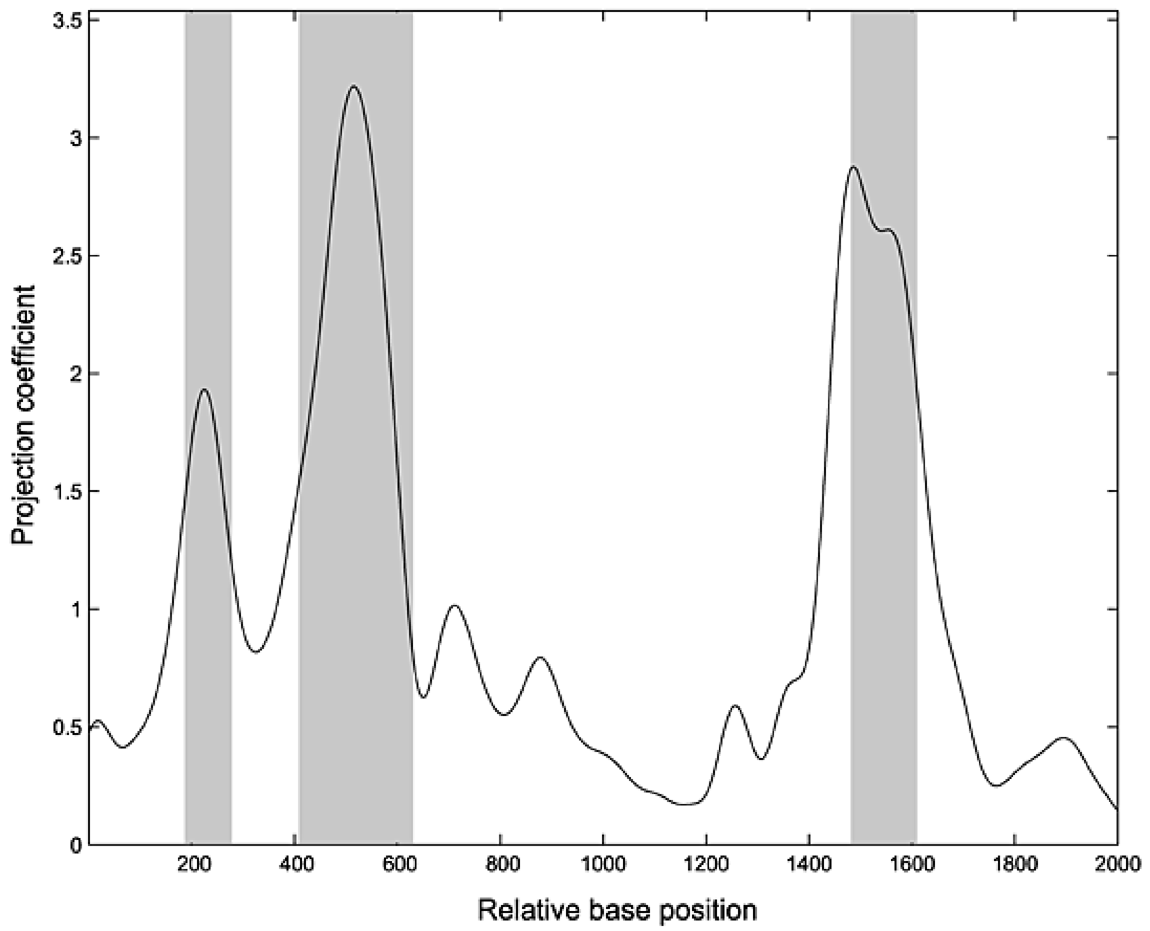
GAGACATTCAGT

But we have sequences – we don’t have numbers

	G	A	G	A	C	A	T	T	C	A	G	T	G
A	0	1	0	1	0	1	0	0	0	1	0	0	0
C	0	0	0	0	1	0	0	0	1	0	0	0	0
G	1	0	1	0	0	0	0	0	0	0	1	0	1
T	0	0	0	0	0	0	1	1	0	0	0	1	0

Identification of Protein Coding Regions Using the Modified Gabor-Wavelet Transform

Mena-Chalco et al. IEEE/ACM Trans. Comp. Bio



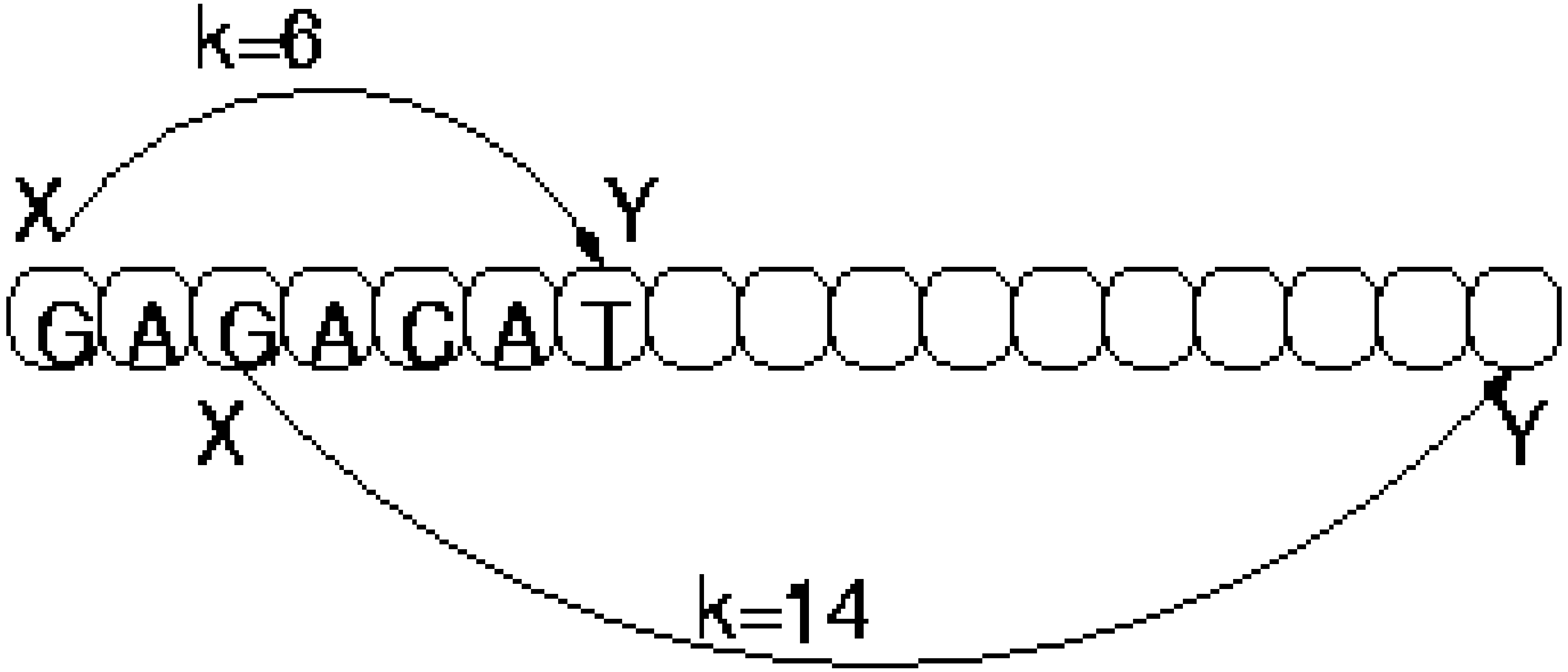
But we have sequences – we don’t have numbers

G	A	G	A	C	A	T	T	C	A	G	T	G

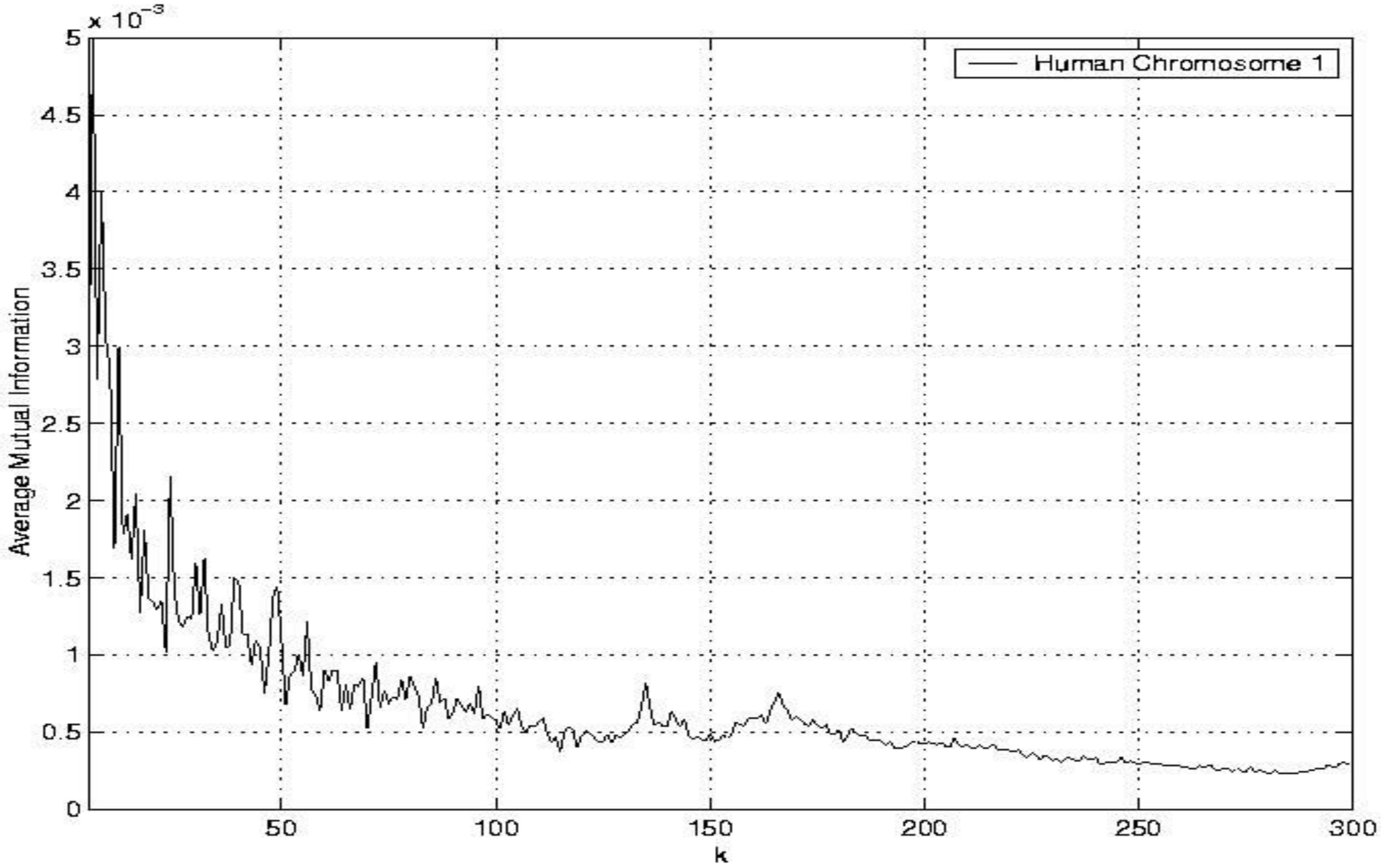
AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	1	2	1	2	0	0	0	2	0	0	1	0	1	1	1

But we have sequences – we don’t have numbers

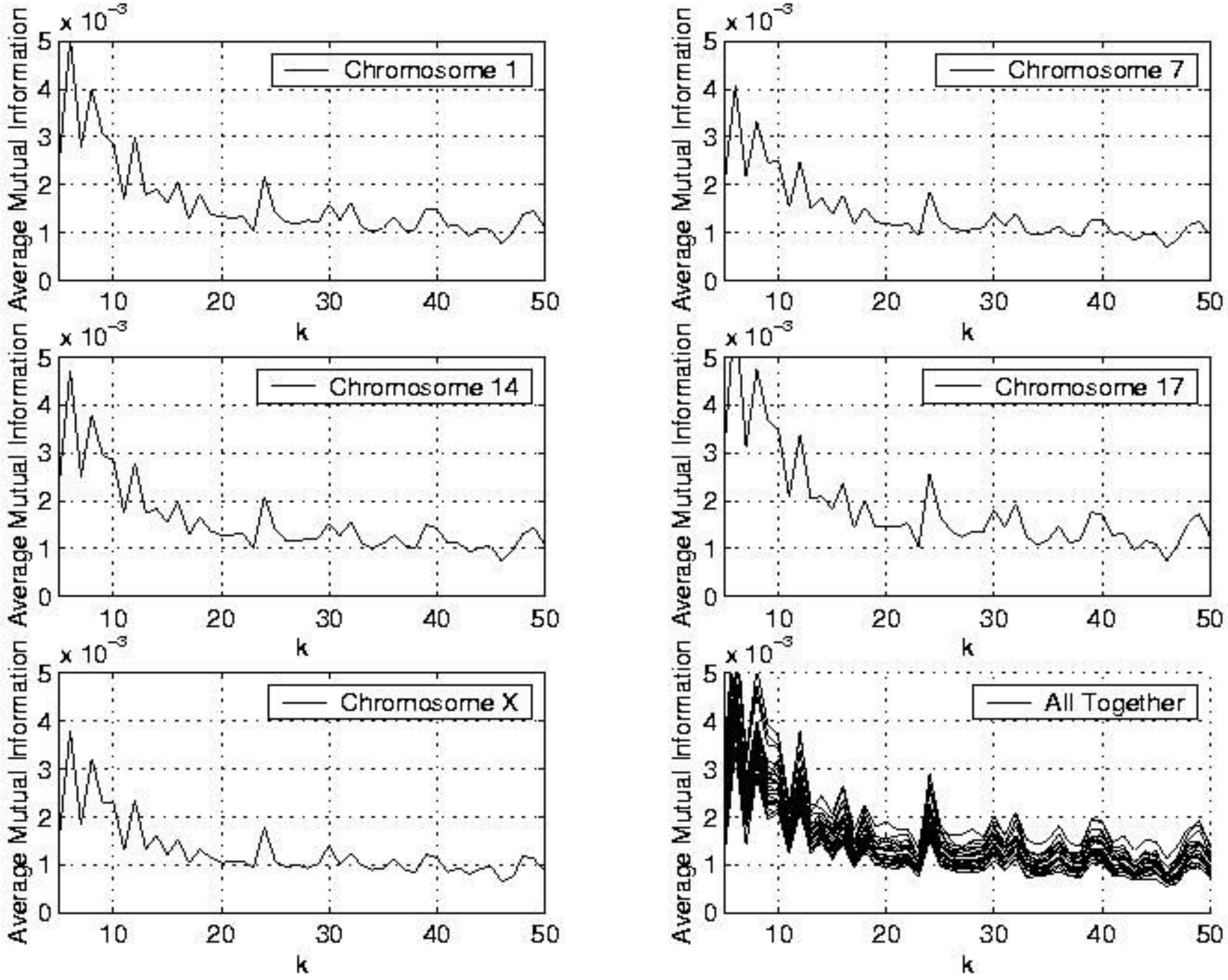
$$I_k = \sum_{X \in A} \sum_{Y \in A} p_k(X,Y) \log\left(\frac{p_k(X,Y)}{p(X)\,p(Y)} \right)$$



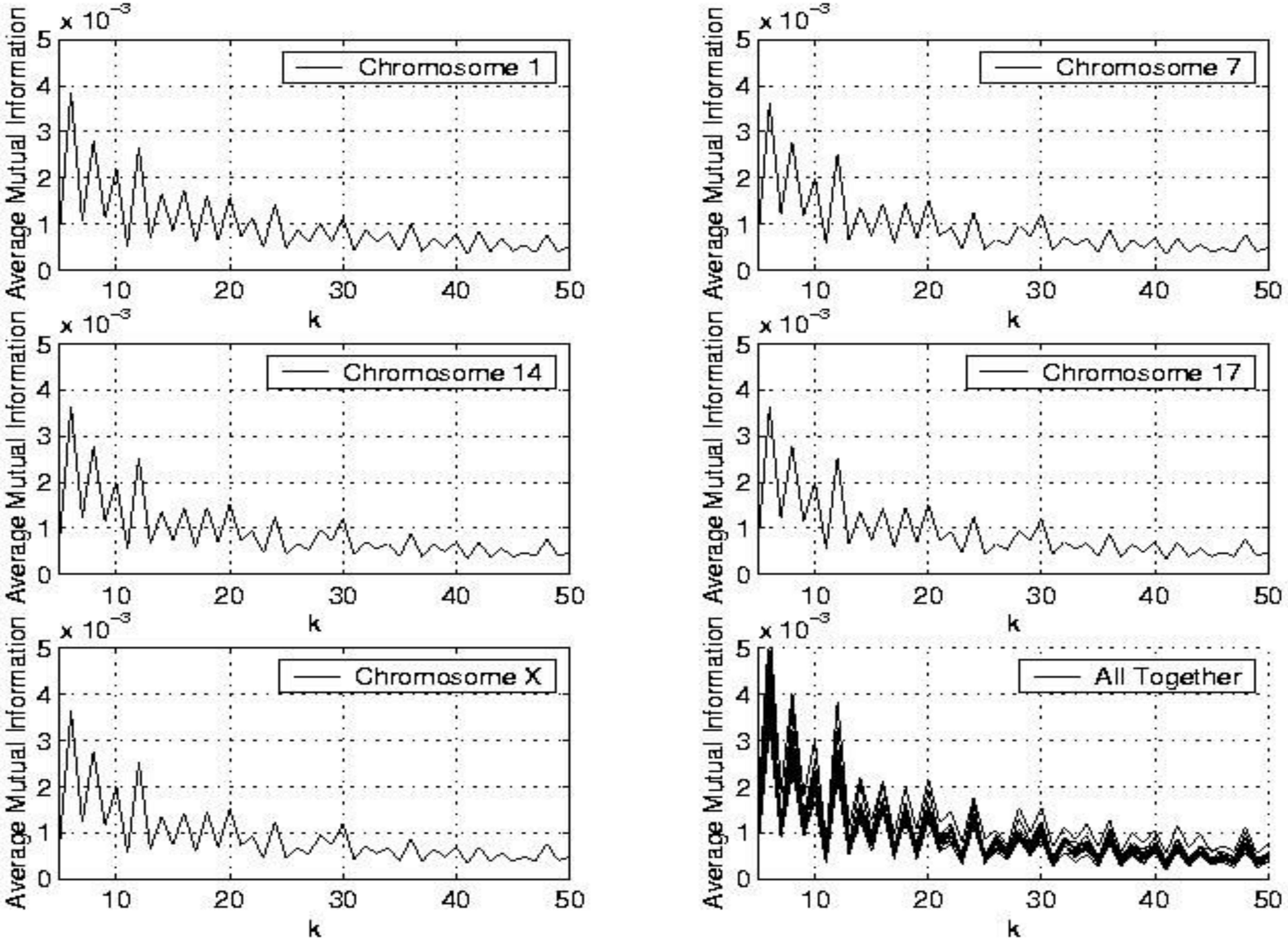
AMI Profile for Human Chromosome 1



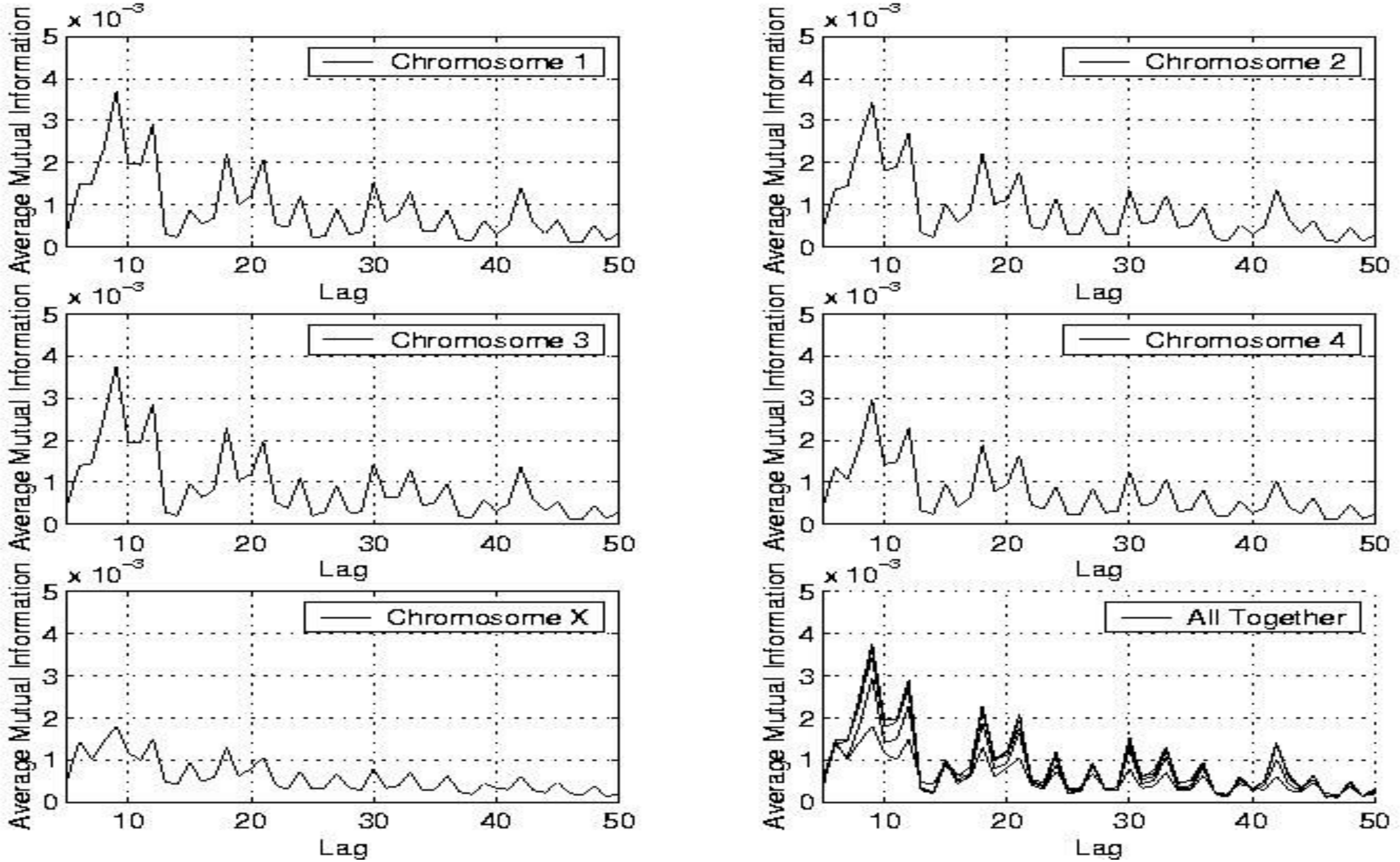
Human chromosomes

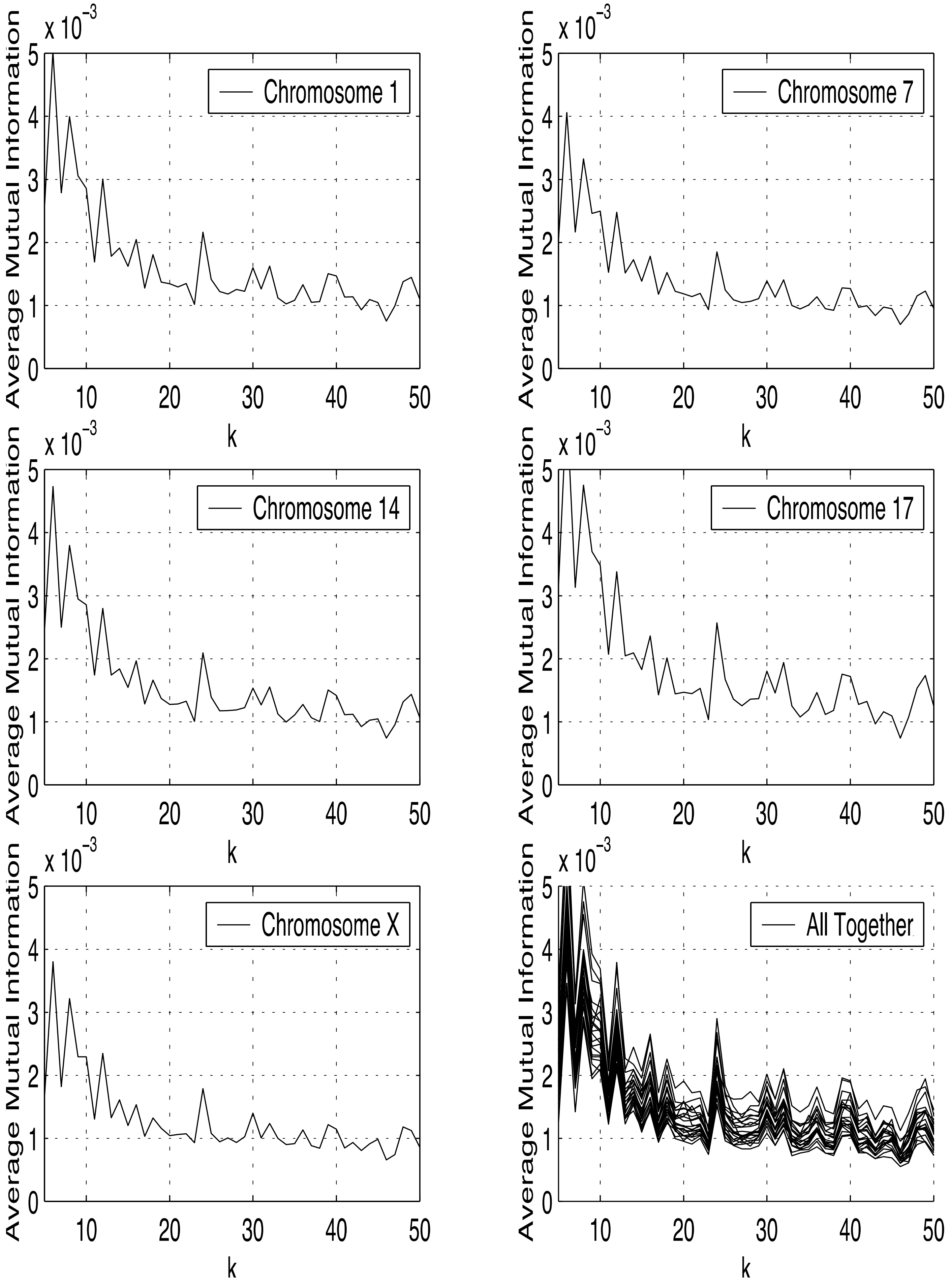


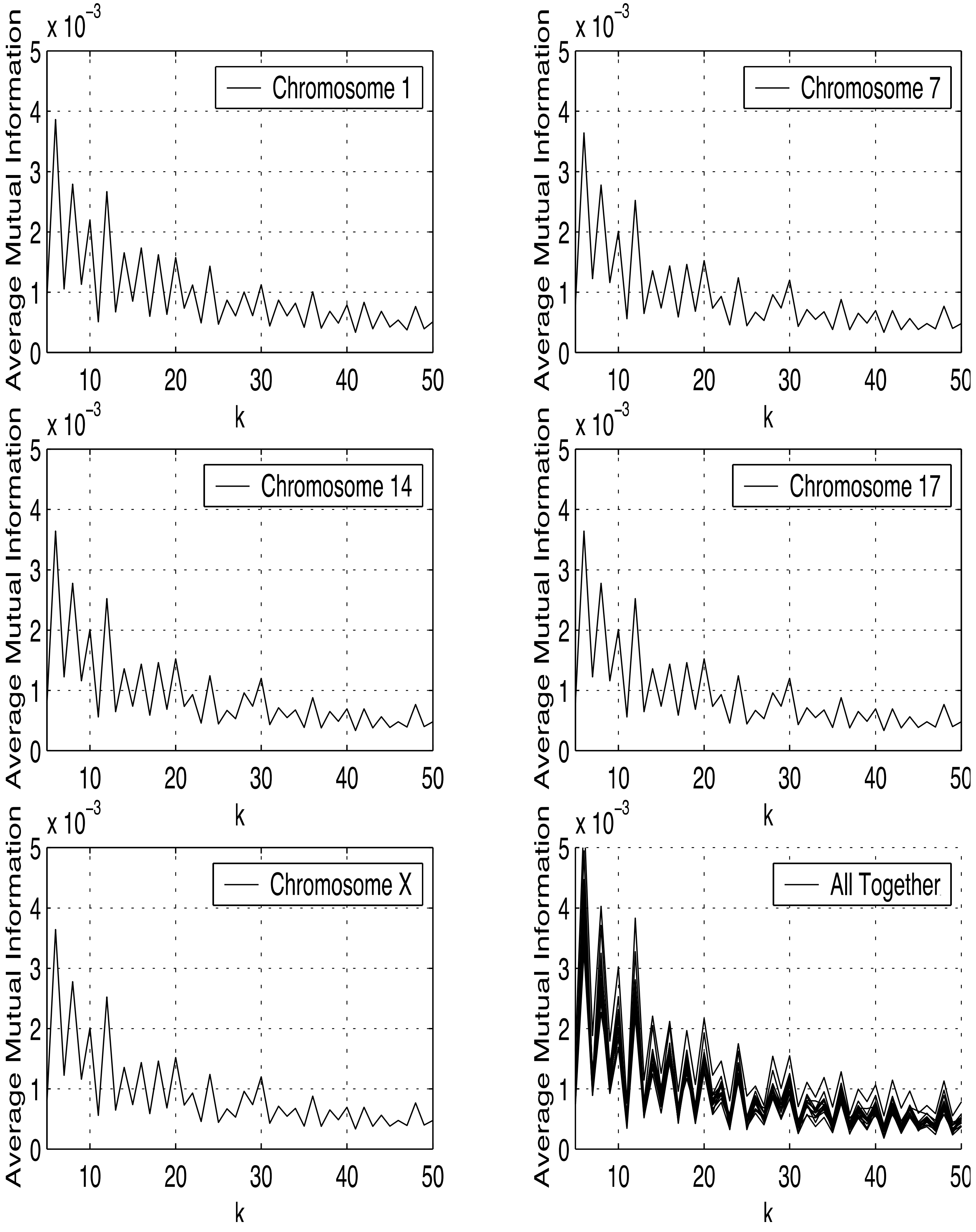
Mouse Chromosome



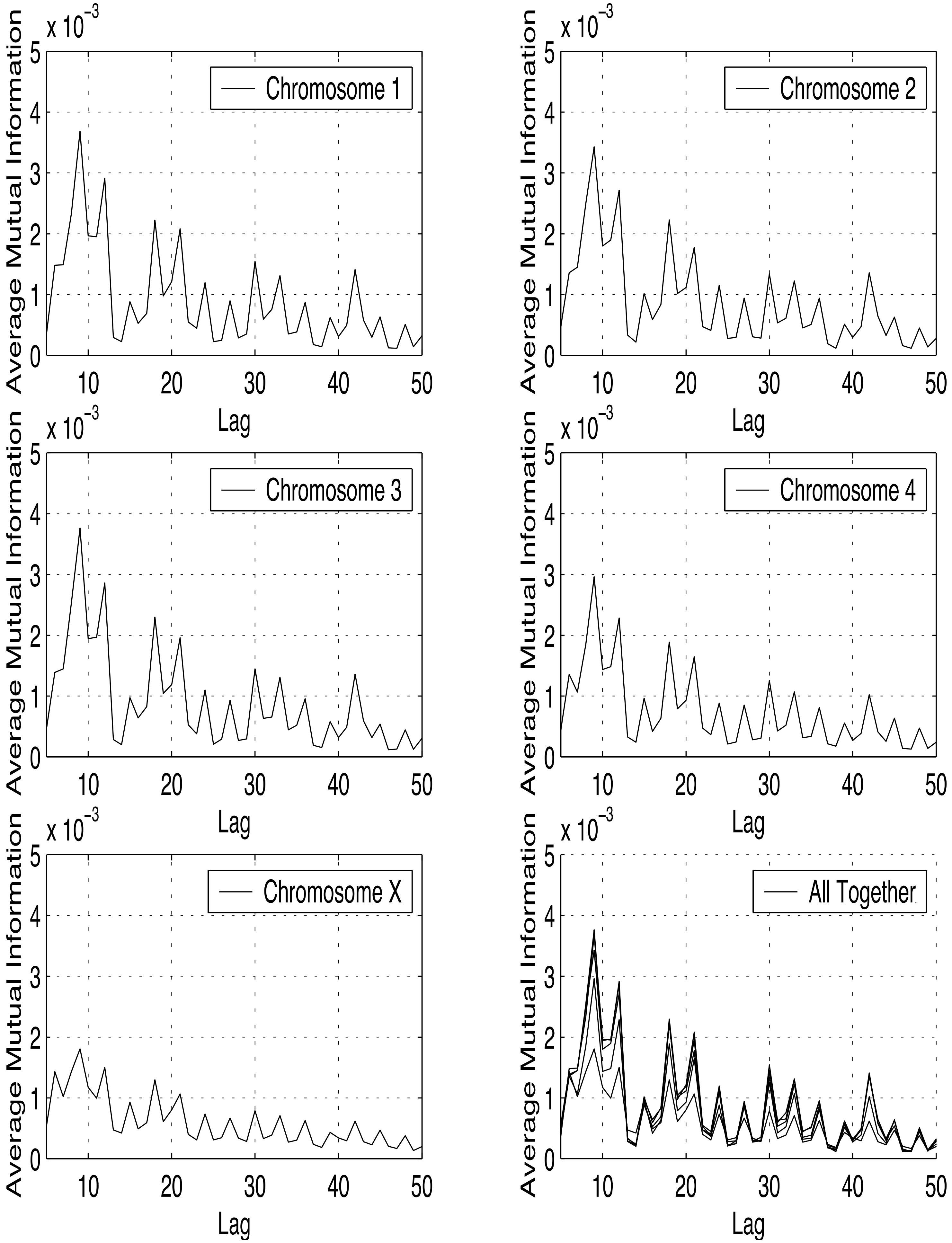
C. Elegans Chromosome



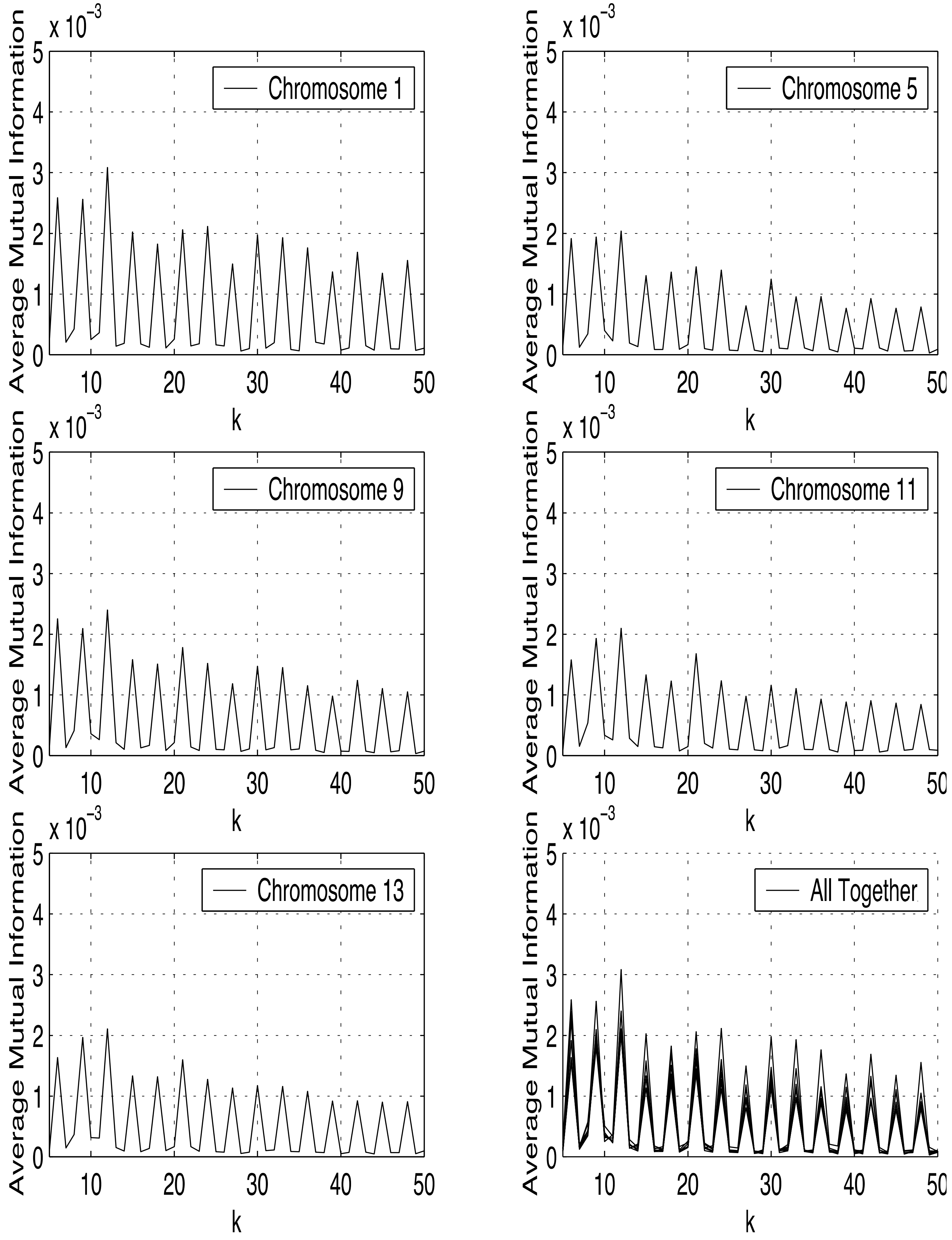

Chromosome 1
Chromosome 7
Chromosome 14
Chromosome 17
Chromosome X
All Together
Average Mutual Information
x 10^-3
0
1
2
3
4
5
10
20
30
40
50
k

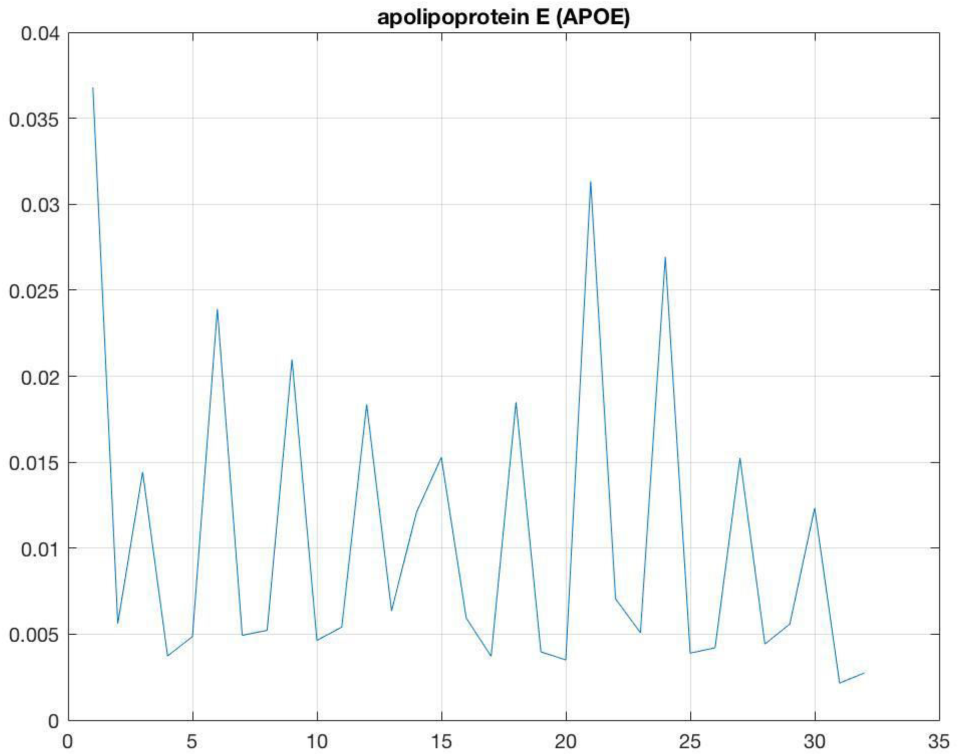

Chromosome 1
Chromosome 7
Chromosome 14
Chromosome 17
Chromosome X
All Together
Average Mutual Information
x 10^-3
0
1
2
3
4
5
10
20
30
40
50
k

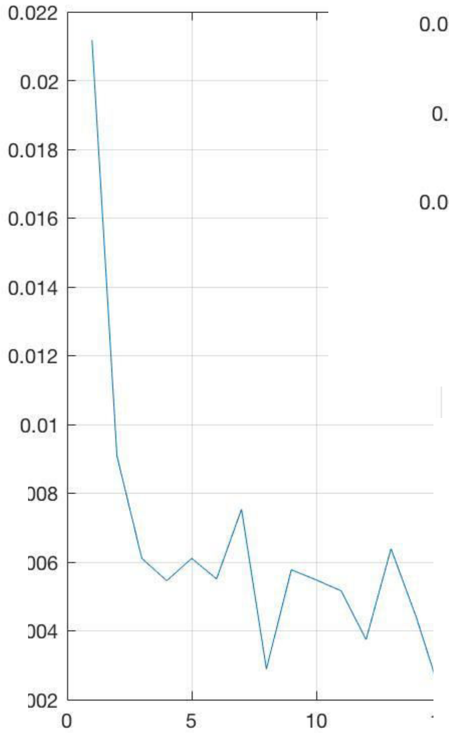
a) C. Elegans Chromosomes

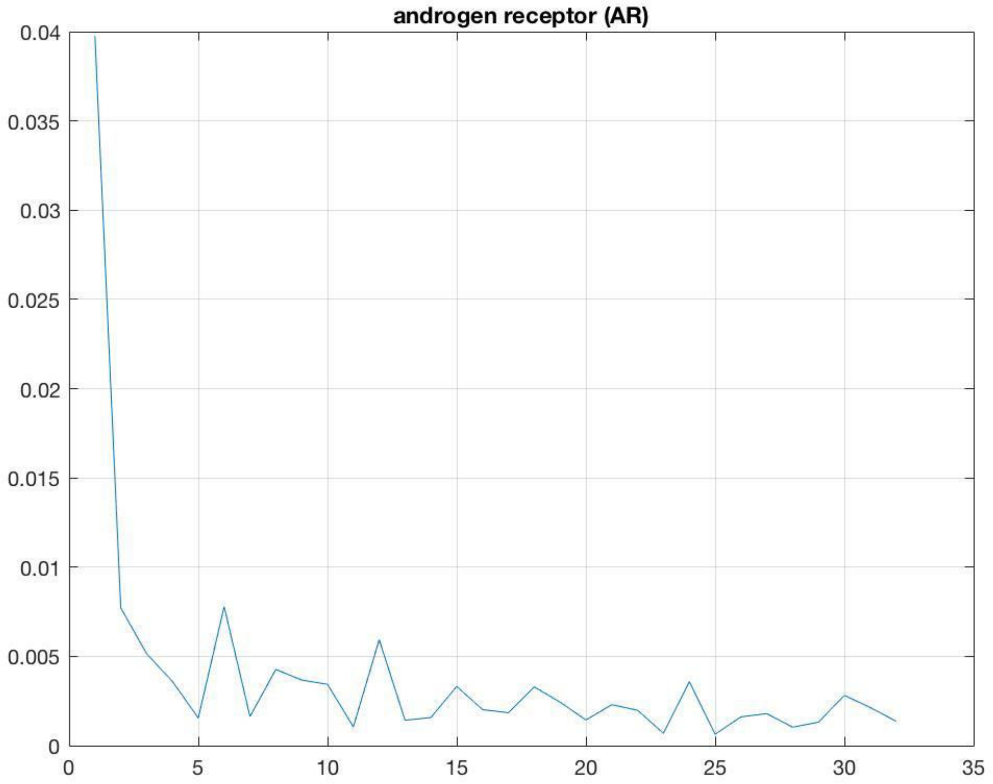


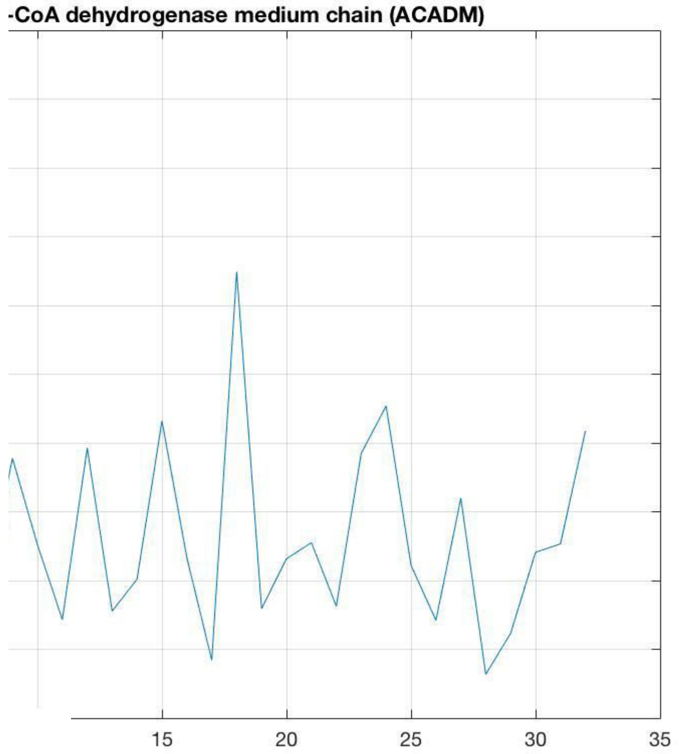
b) S. Cerevisae Chromosomes

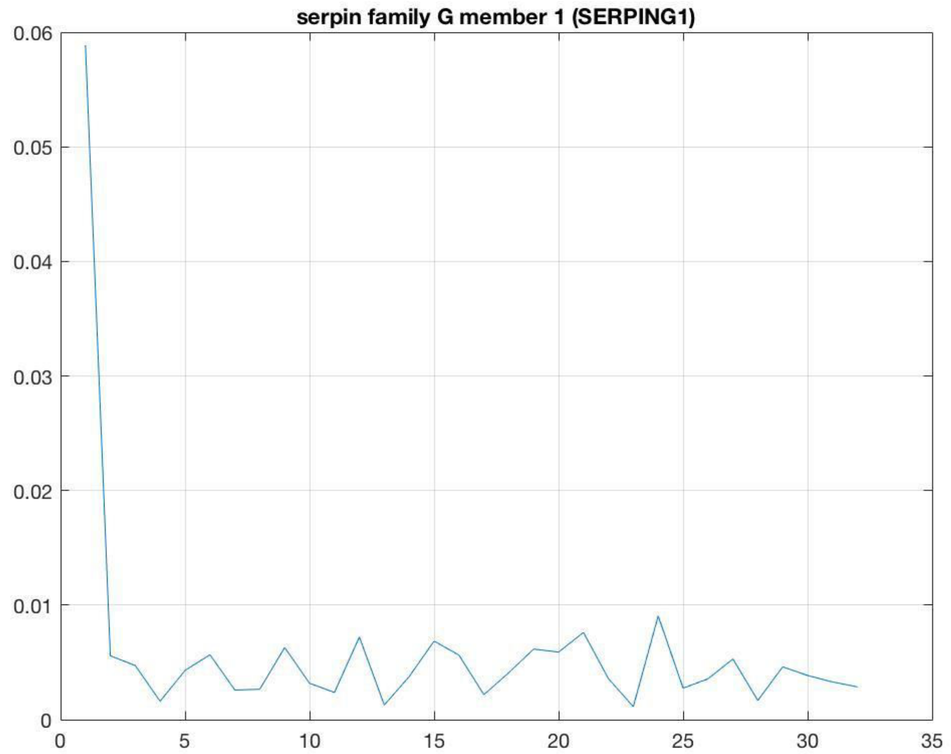


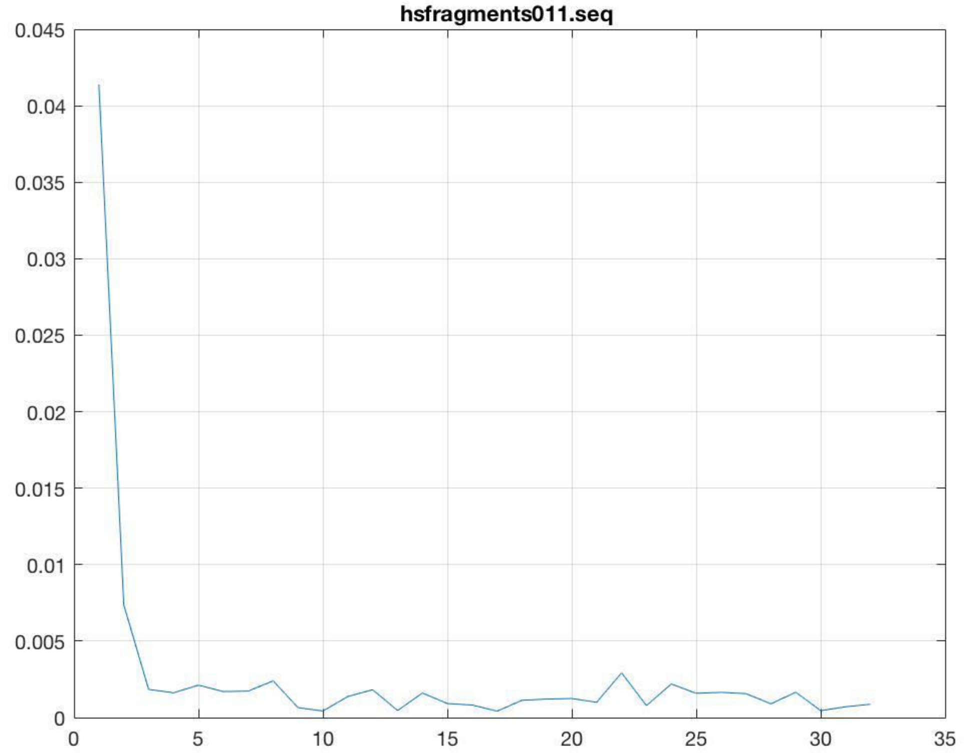

apolipoprotein E (APOE)

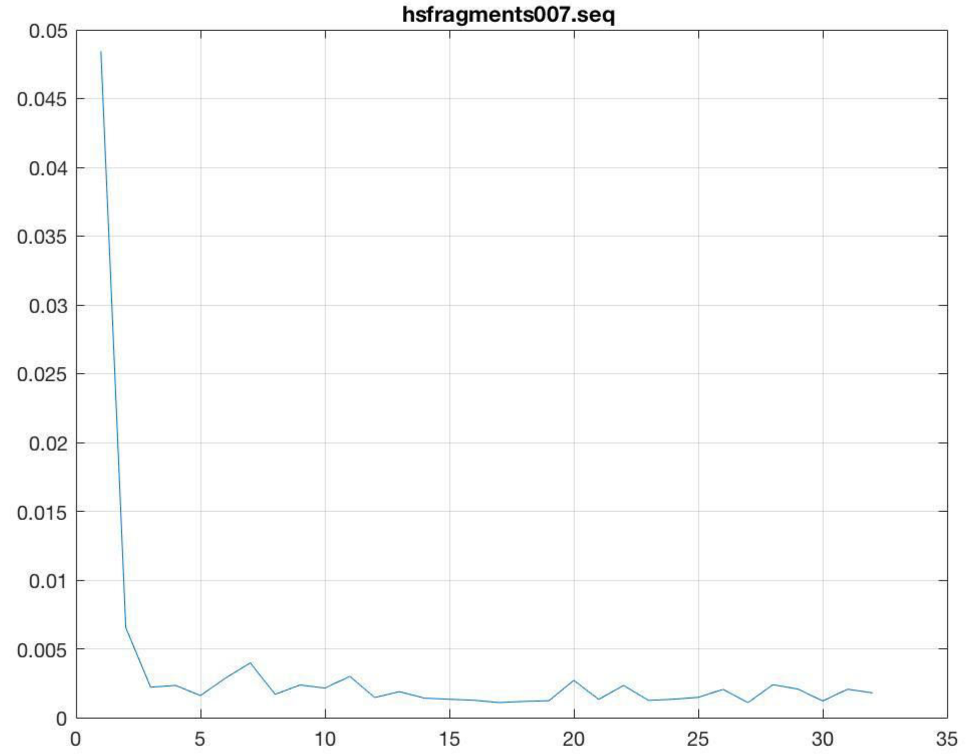



androgen receptor (AR)

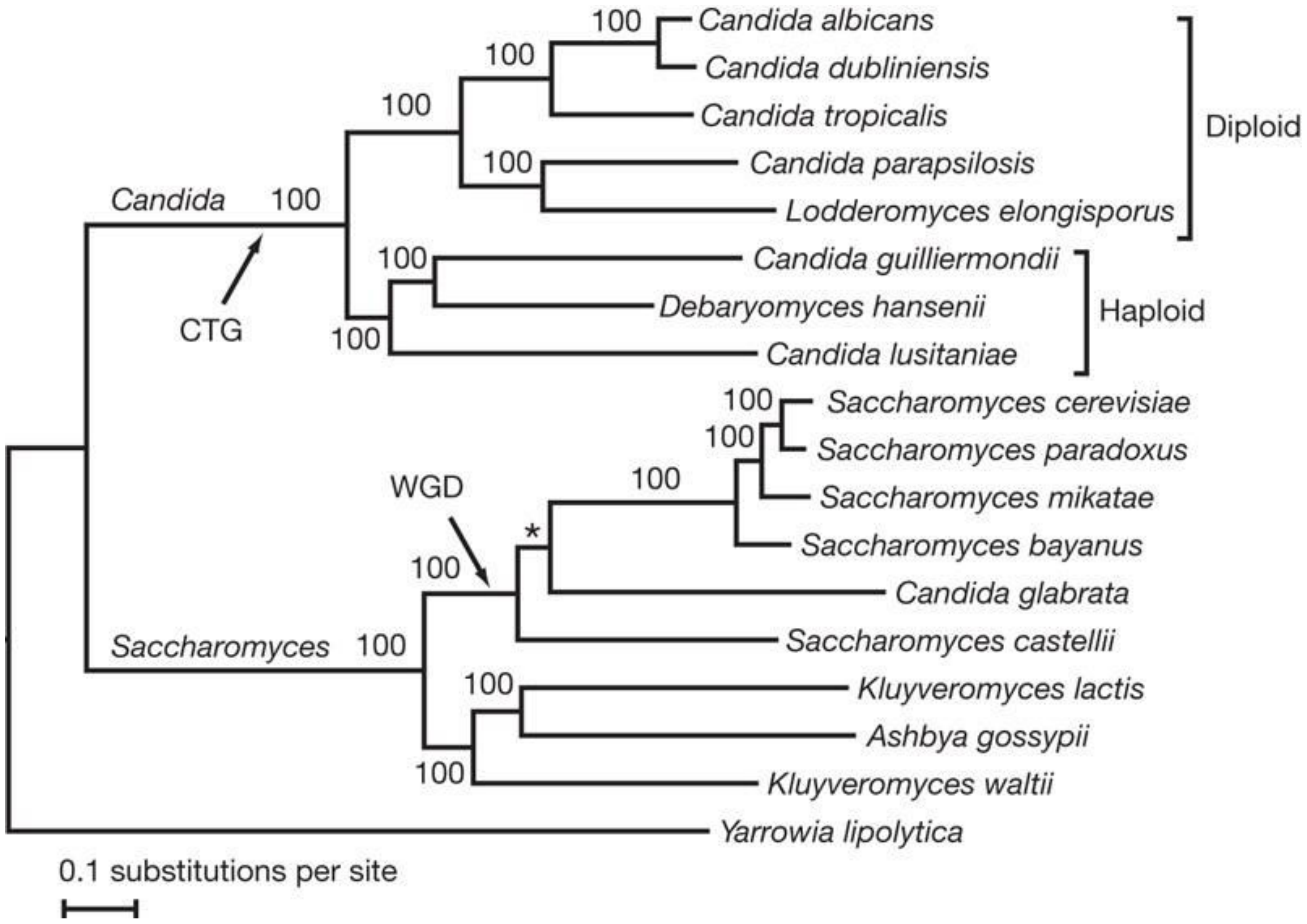

-CoA dehydrogenase medium chain (ACADM)


serpin family G member 1 (SERPING1)


hsfragments011.seq


hsfragments007.seq

- Phylogeny for *Candida* and *Saccharomyces* clades based on multiple sequence alignment of 706 orthologous genes
- Posterior probabilities shown
- WGD: Whole Genome Duplication
- CTG: Translation of CTG codons as serine rather than leucine



- Ribosomal DNA (rDNA) is commonly used to evaluate species relatedness

- The rDNA gene complex contains 3 genes, each of which are ribosomal components once transcribed

- Internal transcribed spacer (ITS) 1 and ITS2 separate these genes

- ITS regions have 2 benefits:

1. Easy to design primers (ribosome genes highly conserved, many copies)

2. Spacers diverge more quickly than ribosome genes

- Distance matrix *D* generated by calculating pairwise distance d_{ij} between AMI profiles $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$

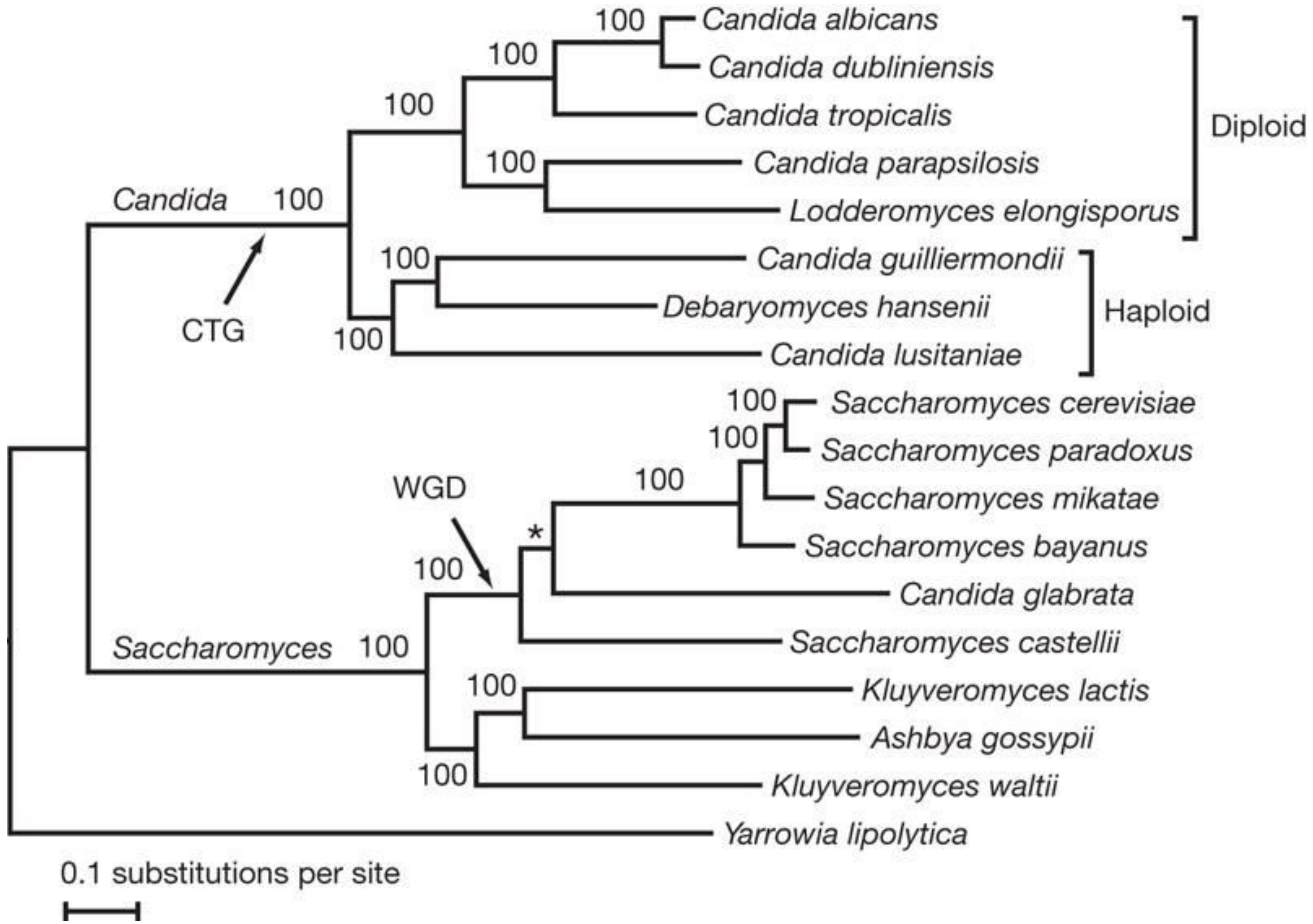
- Distance defined in two ways:
  1. Correlation distance (angle between profiles)

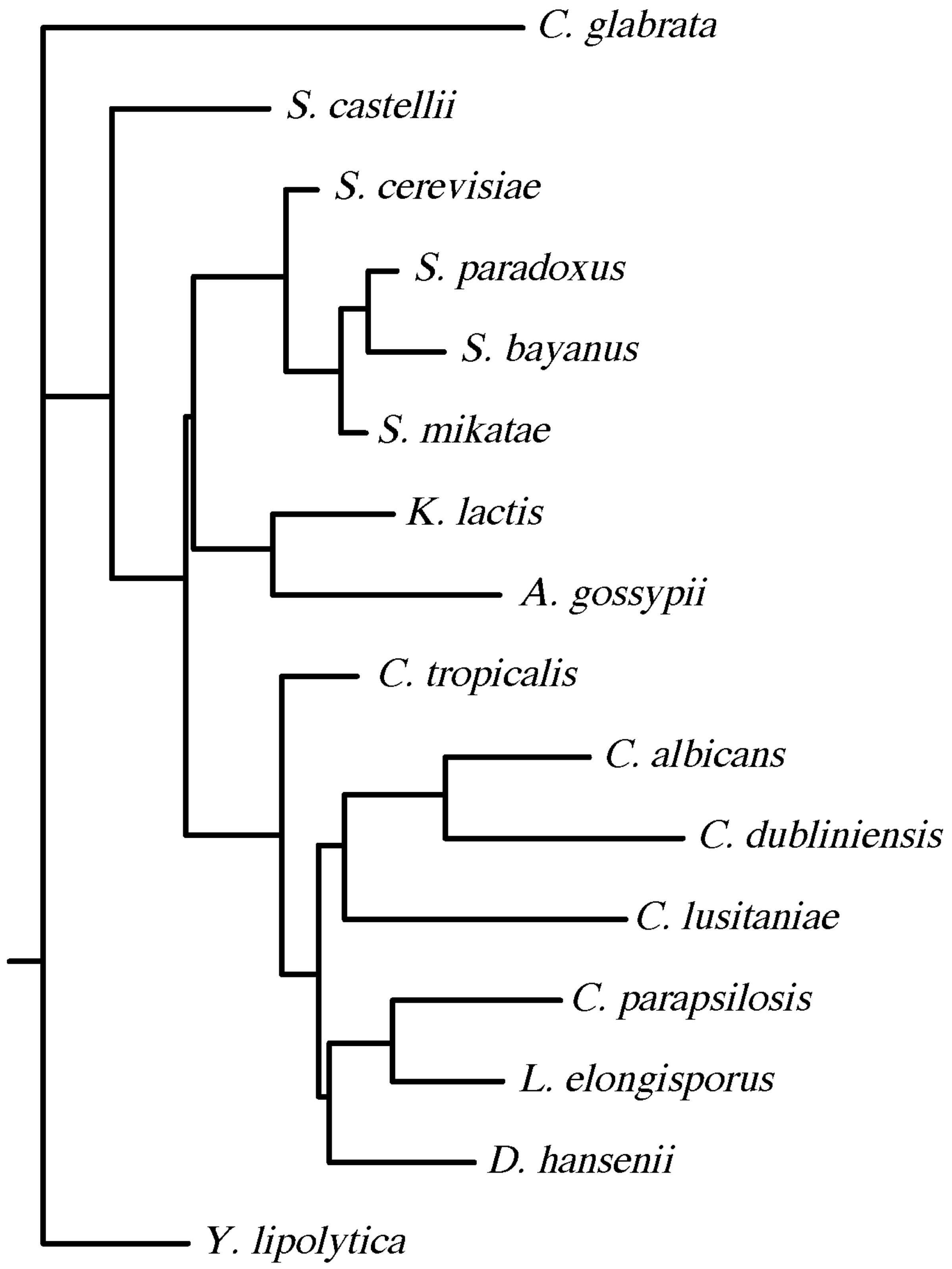
$$d_{ij} = 1 - \cos\theta = 1 - \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{\|\boldsymbol{x}_i\| \, \|\boldsymbol{x}_j\|}$$

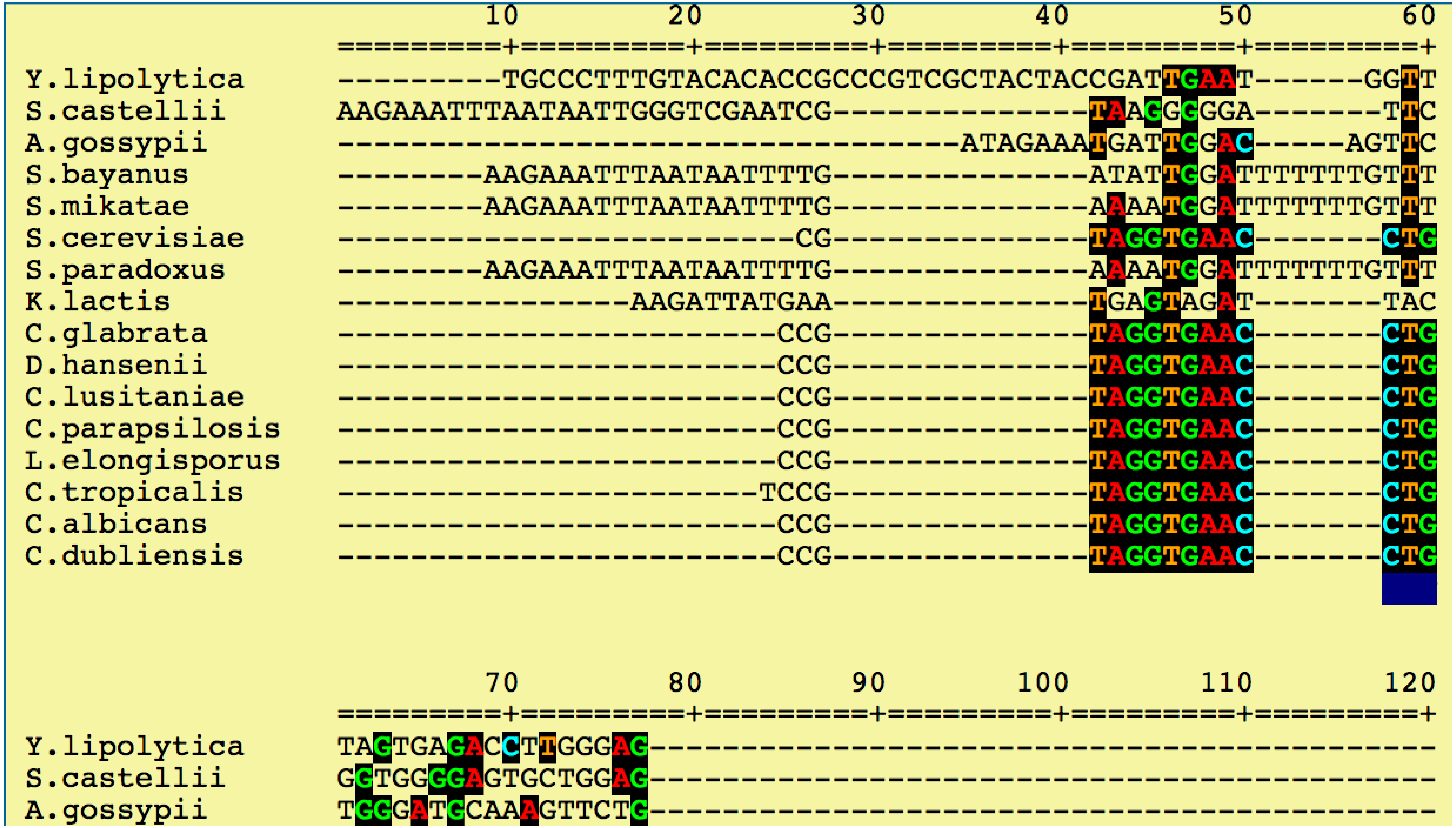
  2. Euclidean distance

$$d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$$

- Phylogenetic trees generated using PHYLIP (neighbor joining)

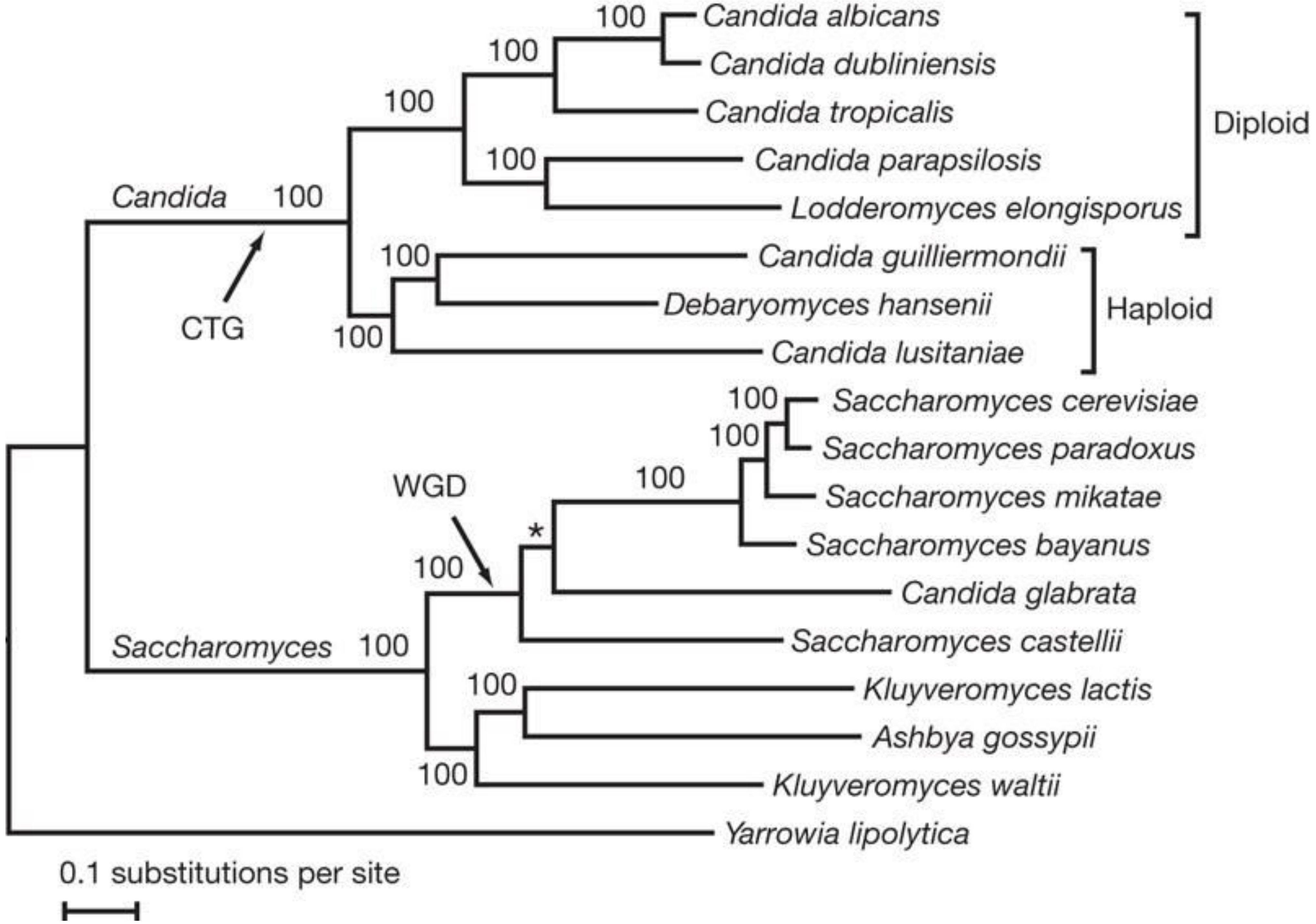

100
Candida albicans
100
Candida dubliniensis
100
Candida tropicalis
100
Candida parapsilosis
Lodderomyces elongisporus
Diploid
Candida
100
CTG
100
Candida guilliermondii
Debaryomyces hansenii
100
Candida lusitaniae
Haploid
100
Saccharomyces cerevisiae
100
Saccharomyces paradoxus
100
Saccharomyces mikatae
Saccharomyces bayanus
WGD
*
100
Candida glabrata
Saccharomyces
100
Saccharomyces castellii
100
Kluyveromyces lactis
Ashbya gossypii
100
Kluyveromyces waltii
Yarrowia lipolytica
0.1 substitutions per site

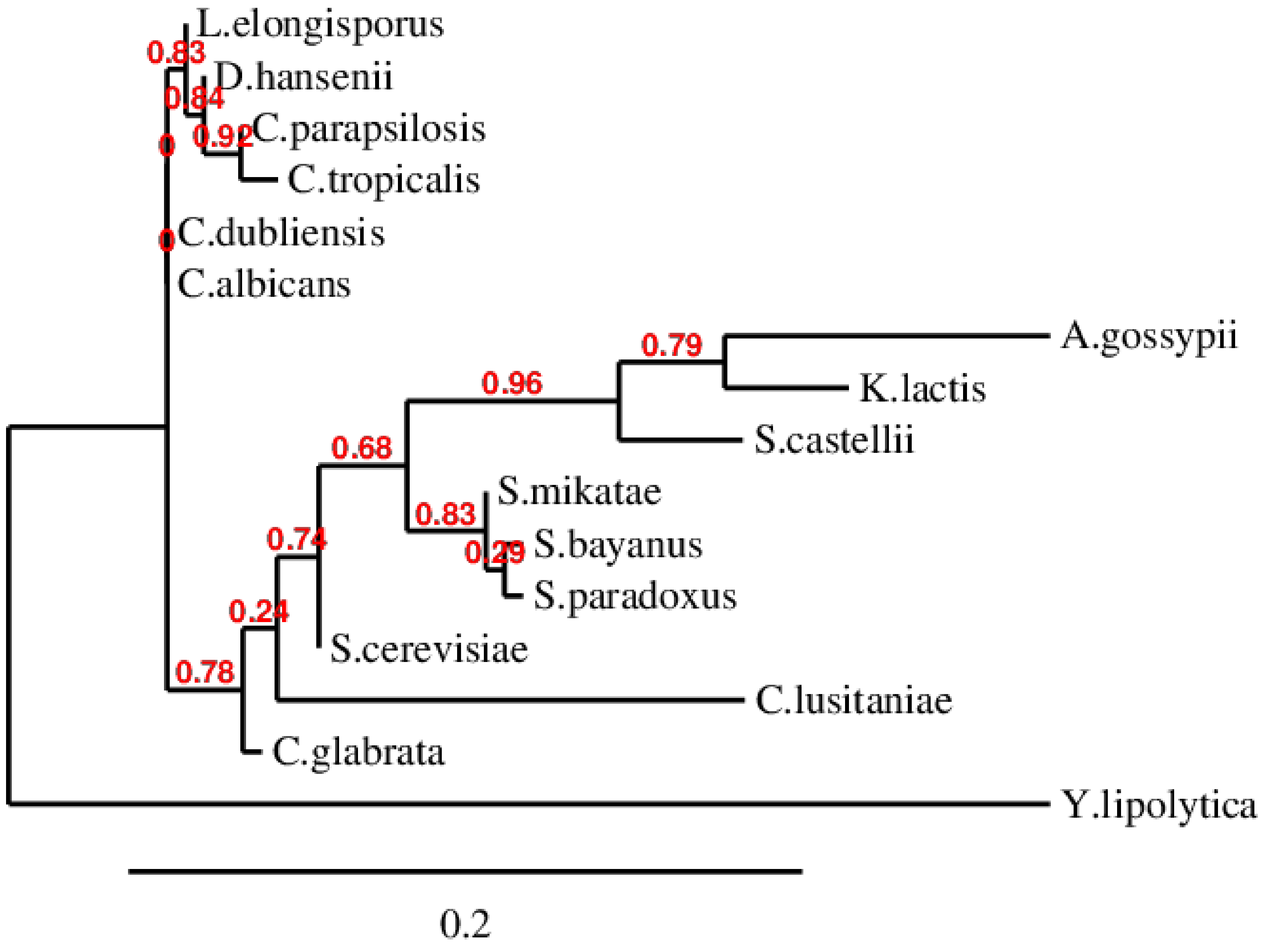

C. glabrata
S. castellii
S. cerevisiae
S. paradoxus
S. bayanus
S. mikatae
K. lactis
A. gossypii
C. tropicalis
C. albicans
C. dubliniensis
C. lusitaniae
C. parapsilosis
L. elongisporus
D. hansenii
Y. lipolytica



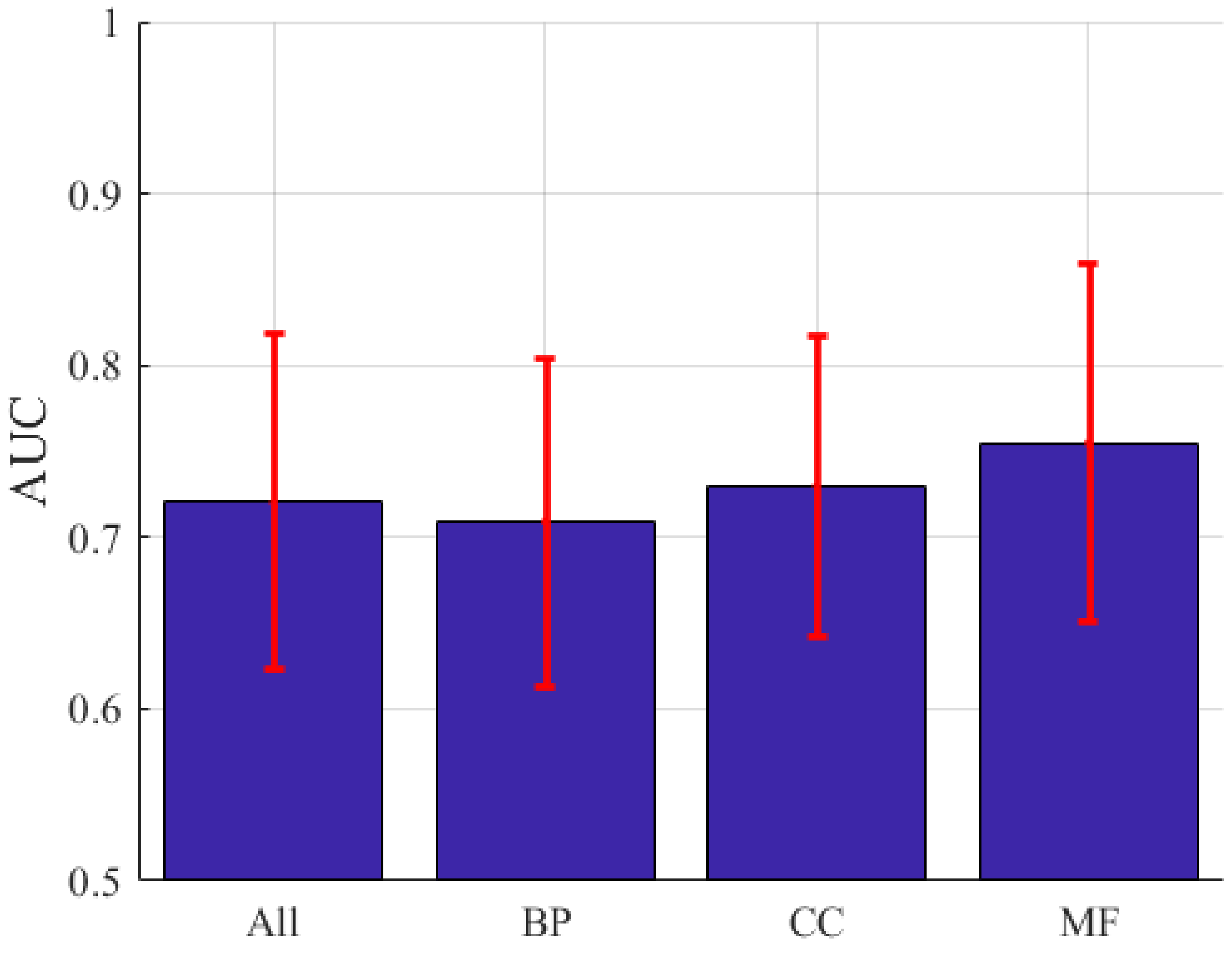
```
                         10        20        30        40        50        60
                 =========+=========+=========+=========+=========+=========+
Y.lipolytica     ---------TGCCCTTTGTACACACCGCCCGTCGCTACTACCGATTGAAT------GGTT
S.castellii      AAGAAATTTAATAATTGGGTCGAATCG--------------TAAGGGGA--------TTC
A.gossypii       ----------------------------------ATAGAAATGATTGGAC-----AGTTC
S.bayanus        --------AAGAAATTTAATAATTTTG--------------ATATTGGATTTTTTTGTTT
S.mikatae        --------AAGAAATTTAATAATTTTG--------------AAAATGGATTTTTTTGTTT
S.cerevisiae     -------------------------CG--------------TAGGTGAAC-------CTG
S.paradoxus      --------AAGAAATTTAATAATTTTG--------------AAAATGGATTTTTTTGTTT
K.lactis         ----------------AAGATTATGAA--------------TGAGTAGAT-------TAC
C.glabrata       ------------------------CCG--------------TAGGTGAAC-------CTG
D.hansenii       ------------------------CCG--------------TAGGTGAAC-------CTG
C.lusitaniae     ------------------------CCG--------------TAGGTGAAC-------CTG
C.parapsilosis   ------------------------CCG--------------TAGGTGAAC-------CTG
L.elongisporus   ------------------------CCG--------------TAGGTGAAC-------CTG
C.tropicalis     -----------------------TCCG--------------TAGGTGAAC-------CTG
C.albicans       ------------------------CCG--------------TAGGTGAAC-------CTG
C.dubliensis     ------------------------CCG--------------TAGGTGAAC-------CTG


                         70        80        90       100       110       120
                 =========+=========+=========+=========+=========+=========+
Y.lipolytica     TAGTGAGACCTTGGGAG-------------------------------------------
S.castellii      GGTGGGGAGTGCTGGAG-------------------------------------------
A.gossypii       TGGGATGCAAAGTTCTG-------------------------------------------
```

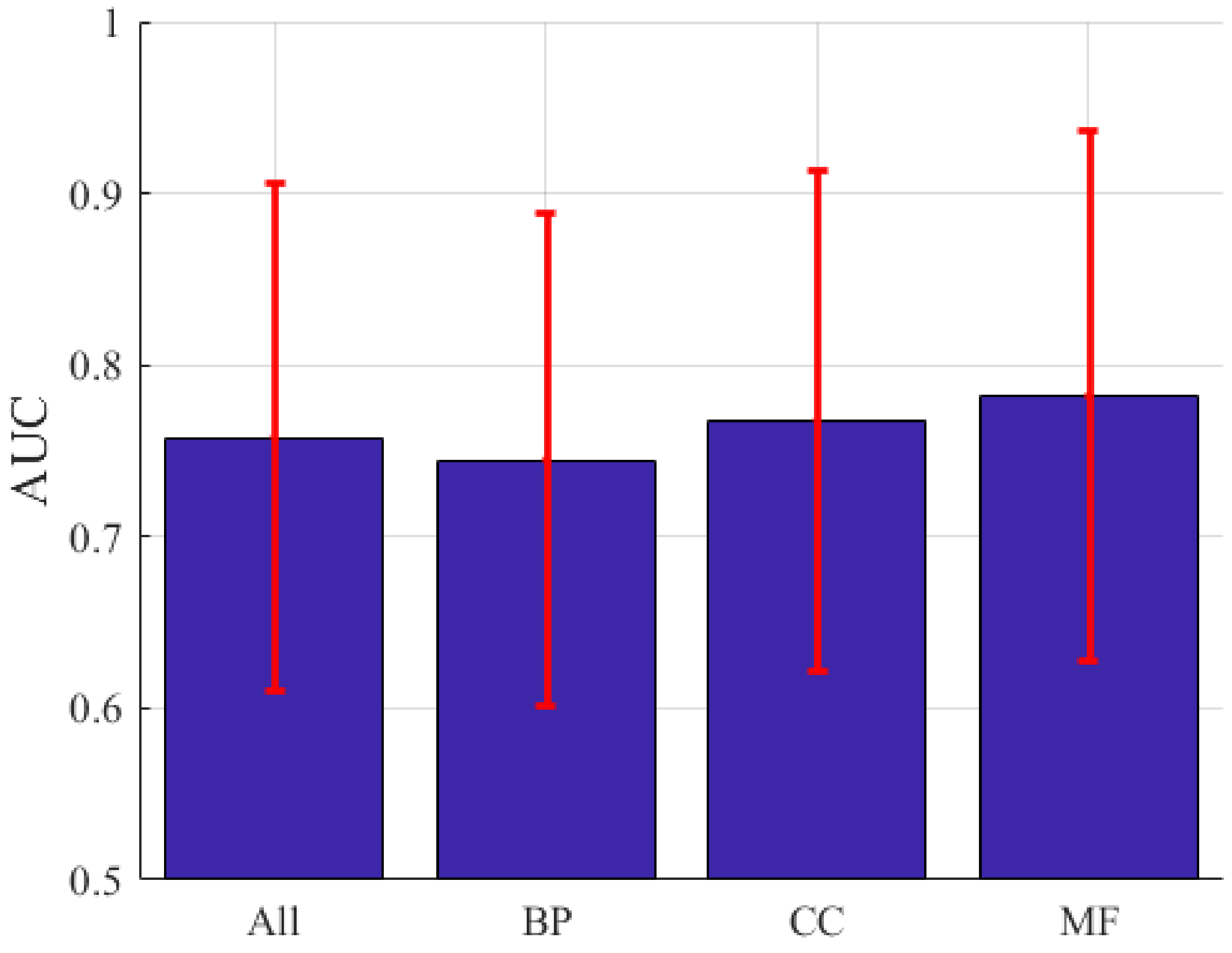

100
Candida albicans
Candida dubliniensis
Candida tropicalis
Candida parapsilosis
Lodderomyces elongisporus
Diploid
100
100
100
Candida
100
CTG
Candida guilliermondii
Debaryomyces hansenii
Candida lusitaniae
Haploid
100
100
Saccharomyces cerevisiae
Saccharomyces paradoxus
Saccharomyces mikatae
Saccharomyces bayanus
Candida glabrata
Saccharomyces castellii
100
100
100
WGD
*
100
Saccharomyces
100
100
Kluyveromyces lactis
Ashbya gossypii
Kluyveromyces waltii
100
Yarrowia lipolytica
0.1 substitutions per site


L.elongisporus
0.83
D.hansenii
0.84
C.parapsilosis
0
0.92
C.tropicalis
C.dubliensis
0
C.albicans
0.79
A.gossypii
0.96
K.lactis
S.castellii
0.68
S.mikatae
0.83
S.bayanus
0.29
0.74
S.paradoxus
0.24
S.cerevisiae
0.78
C.lusitaniae
C.glabrata
Y.lipolytica
0.2

GO Prediction

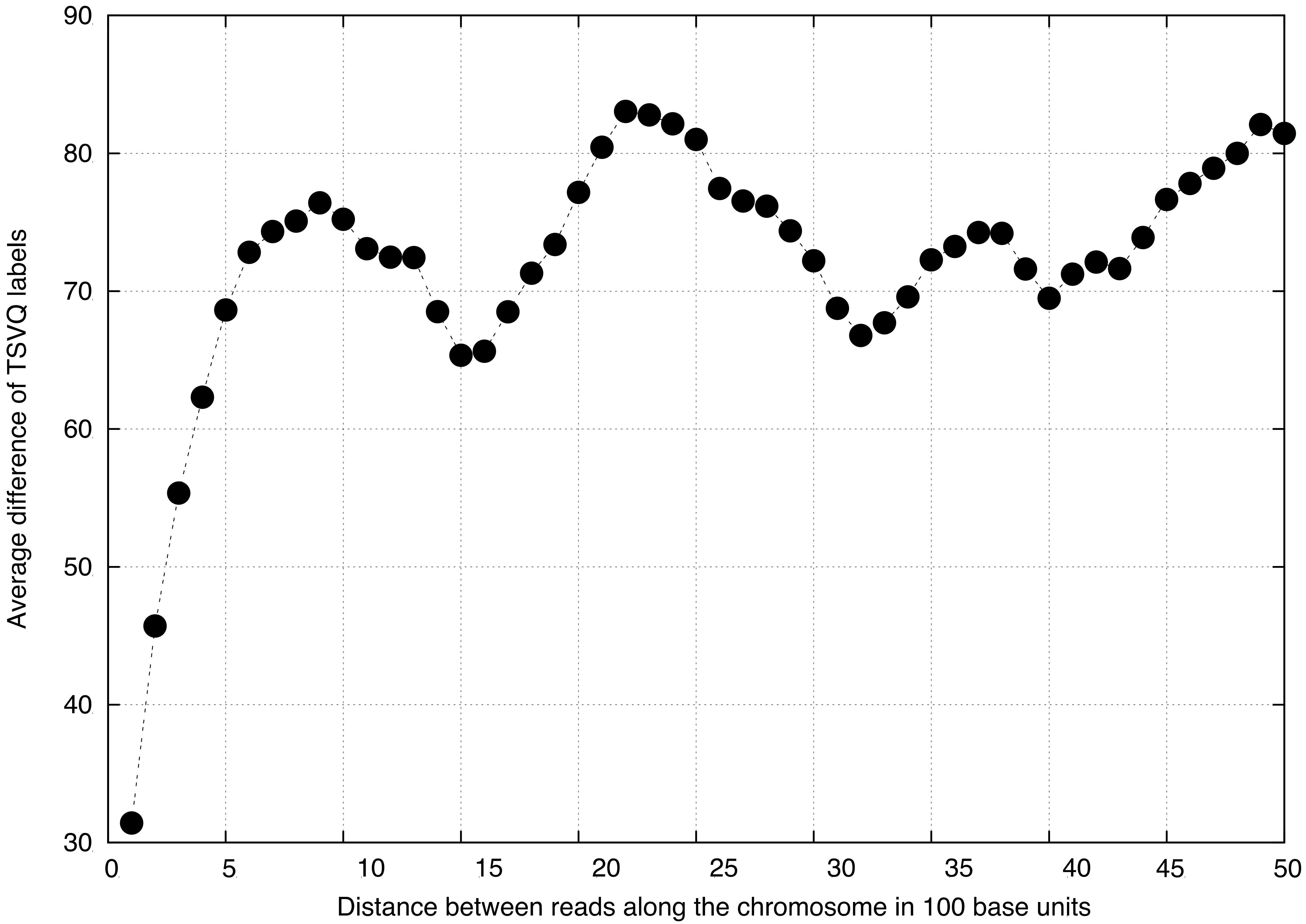


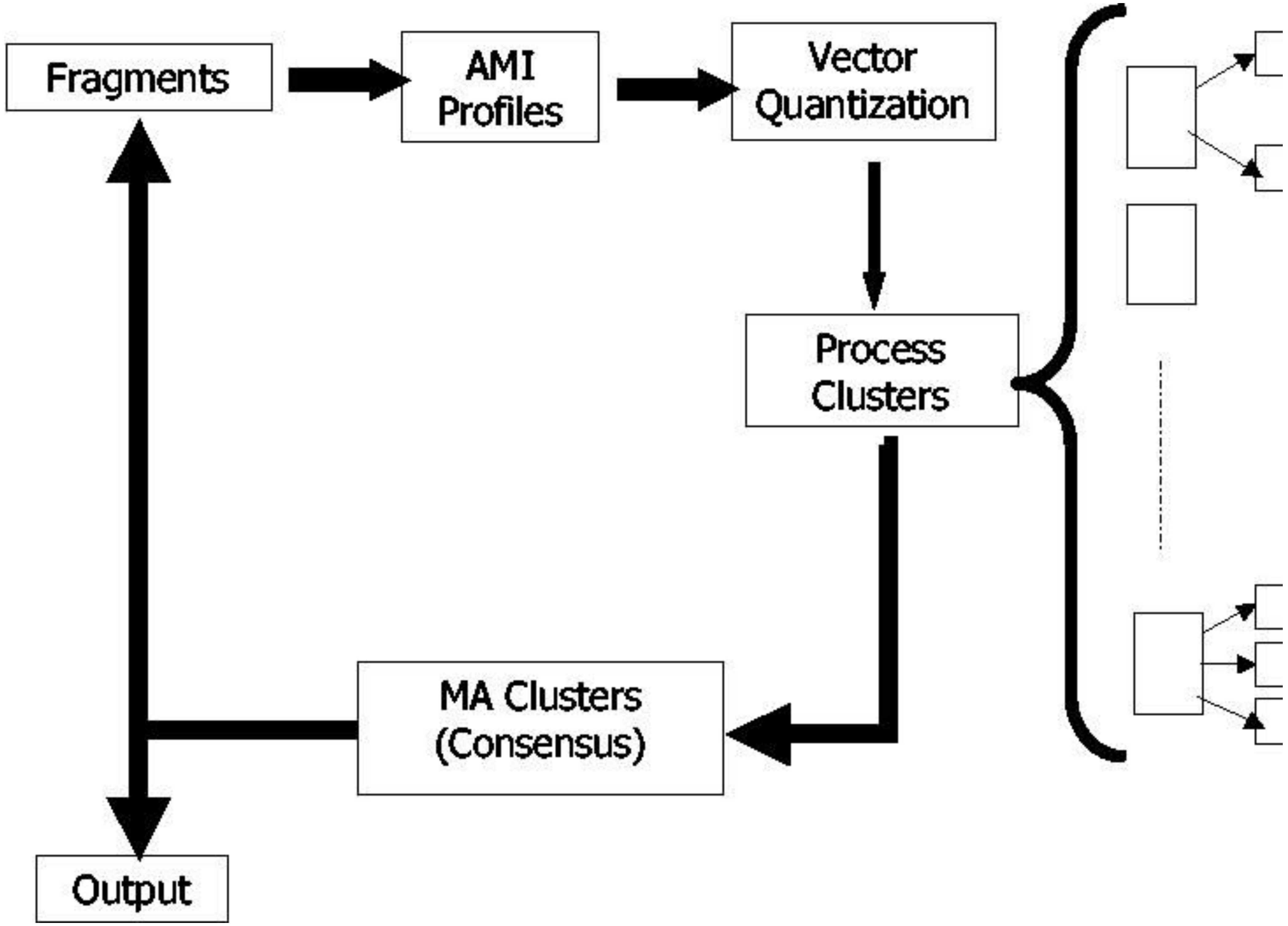
“High Abundance" GO terms



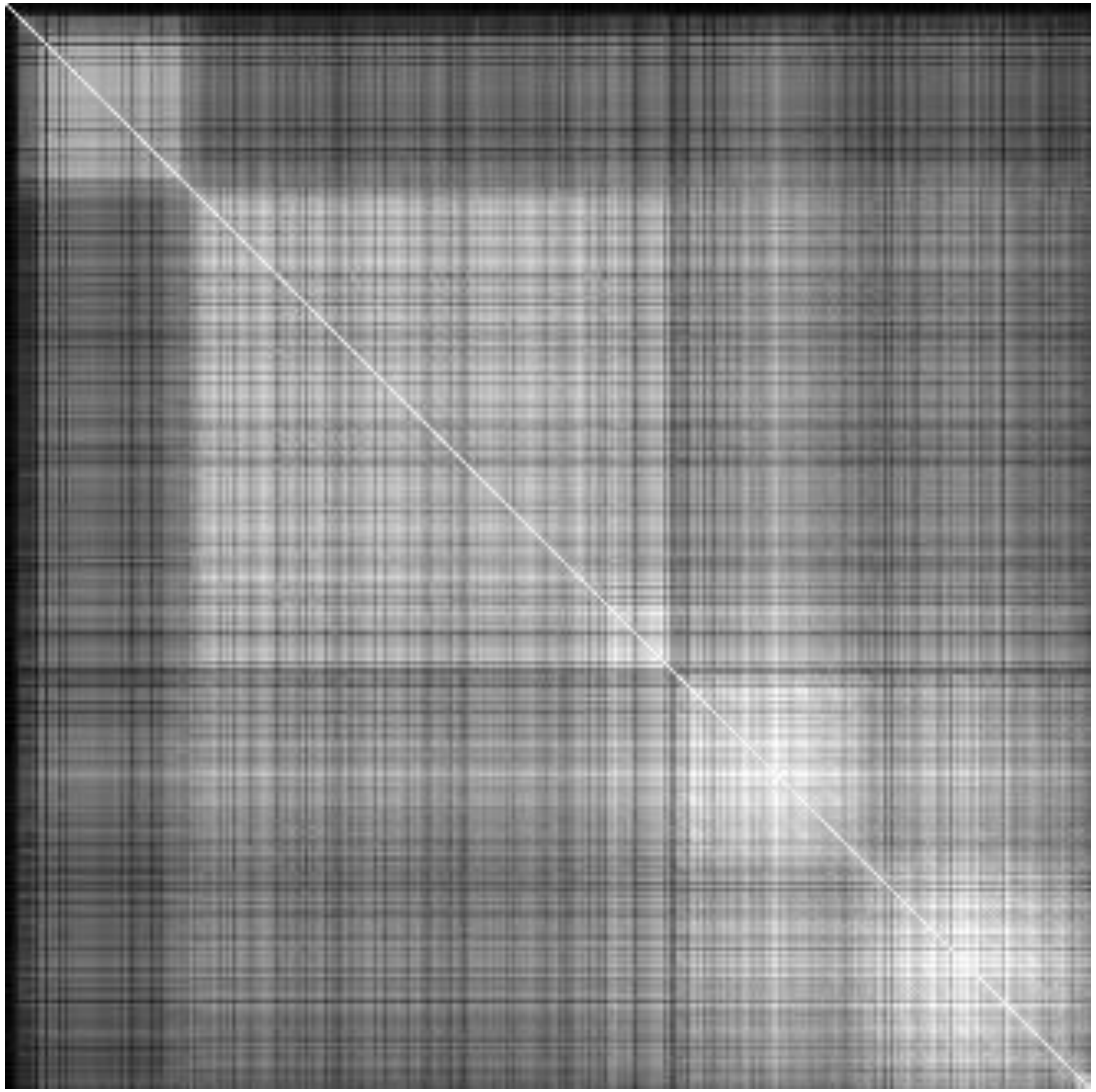
“Low Abundance” GO terms

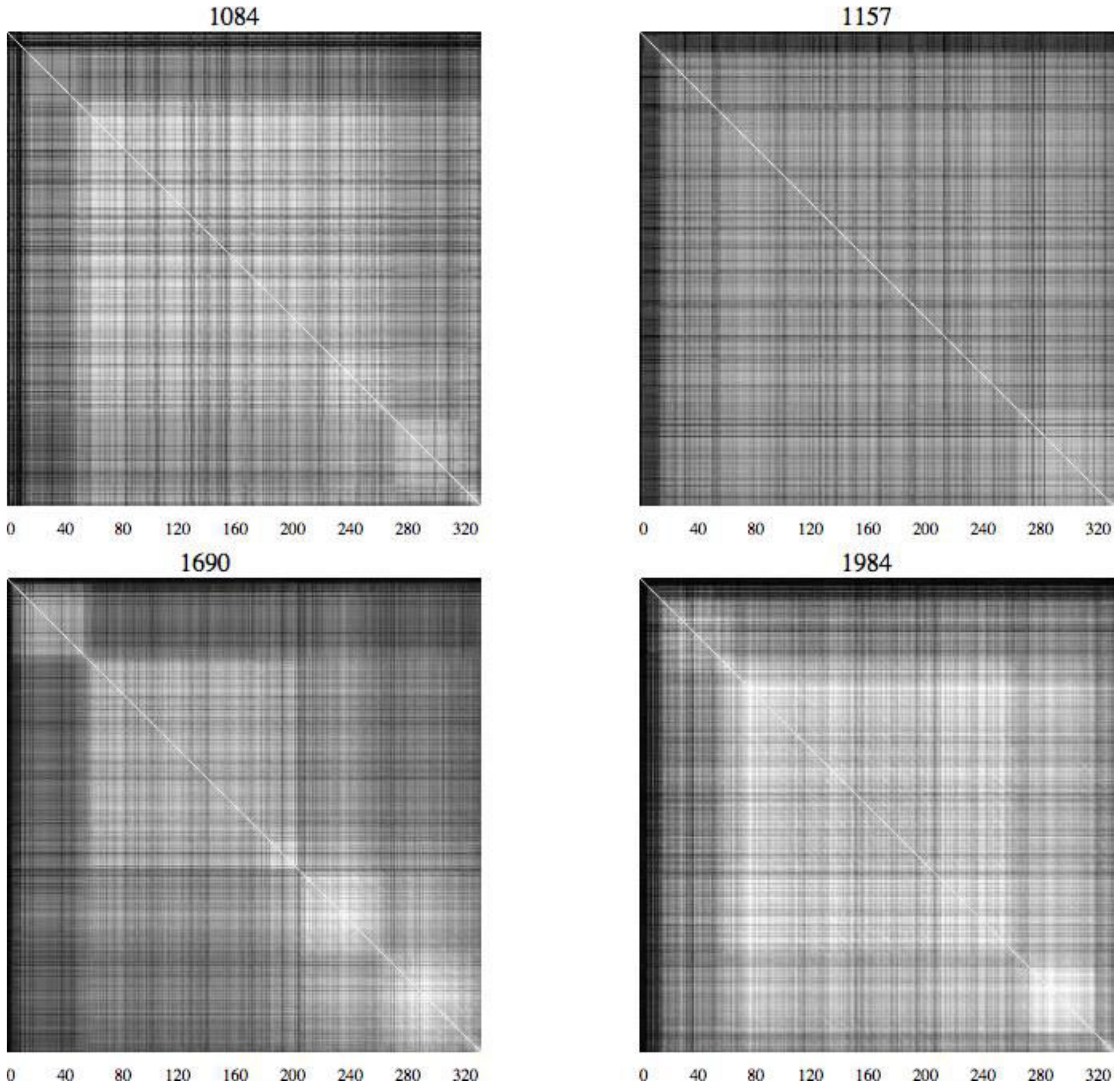
BP: Biological Processes, CC: Cellular Component, MF: Molecular Function


Average difference of TSVQ labels
90
80
70
60
50
40
30
0
5
10
15
20
25
30
35
40
45
50
Distance between reads along the chromosome in 100 base units

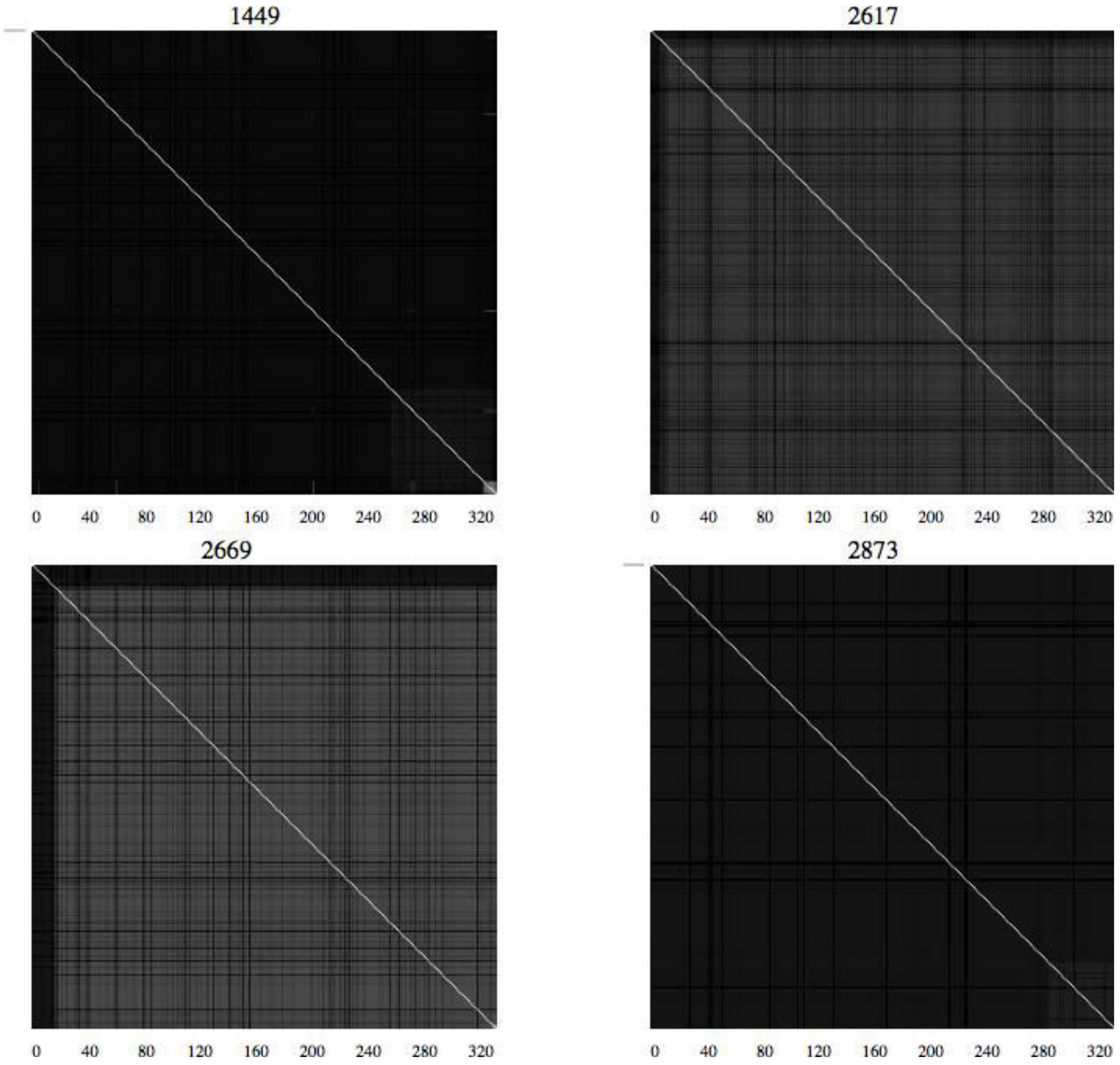

Fragments
AMI Profiles
Vector Quantization
Process Clusters
MA Clusters (Consensus)
Output

								k							m				
M	D	C	S	I	N	A	N	Q	M	K	L	G	H	Q	R	I	H	L	K
M	R	C	S	T	M	D	N	Q	M	N	L	G	R	Q	H	I	H	F	K
M	D	G	S	E	N	A	K	H	I	K	L	D	Q	H	R	I	Q	L	N
M	Q	C	S	I	N	A	N	H	K	K	F	G	Q	Q	R	T	H	L	K
M	D	G	S	I	N	A	N	Q	K	I	L	G	H	H	R	I	Q	L	R
M	E	C	S	E	N	A	K	R	M	K	S	G	H	Q	H	I	H	F	K
M	D	C	S	I	N	A	N	Q	I	K	F	A	Q	Q	R	M	H	L	N
M	D	W	S	I	N	A	N	H	M	K	L	D	R	P	Q	I	H	L	T
M	E	C	S	I	N	A	N	Q	R	N	L	G	H	Q	R	I	Y	L	K
M	D	C	S	E	N	A	K	Y	I	K	L	A	Q	H	R	I	H	L	N
M	D	G	S	I	N	A	N	Q	M	N	F	G	H	Q	R	I	H	L	K
M	D	C	F	I	R	S	K	H	L	K	L	G	H	Q	H	I	H	L	N
M	E	W	S	I	N	A	N	Q	M	Q	L	D	Y	Q	R	T	R	L	K



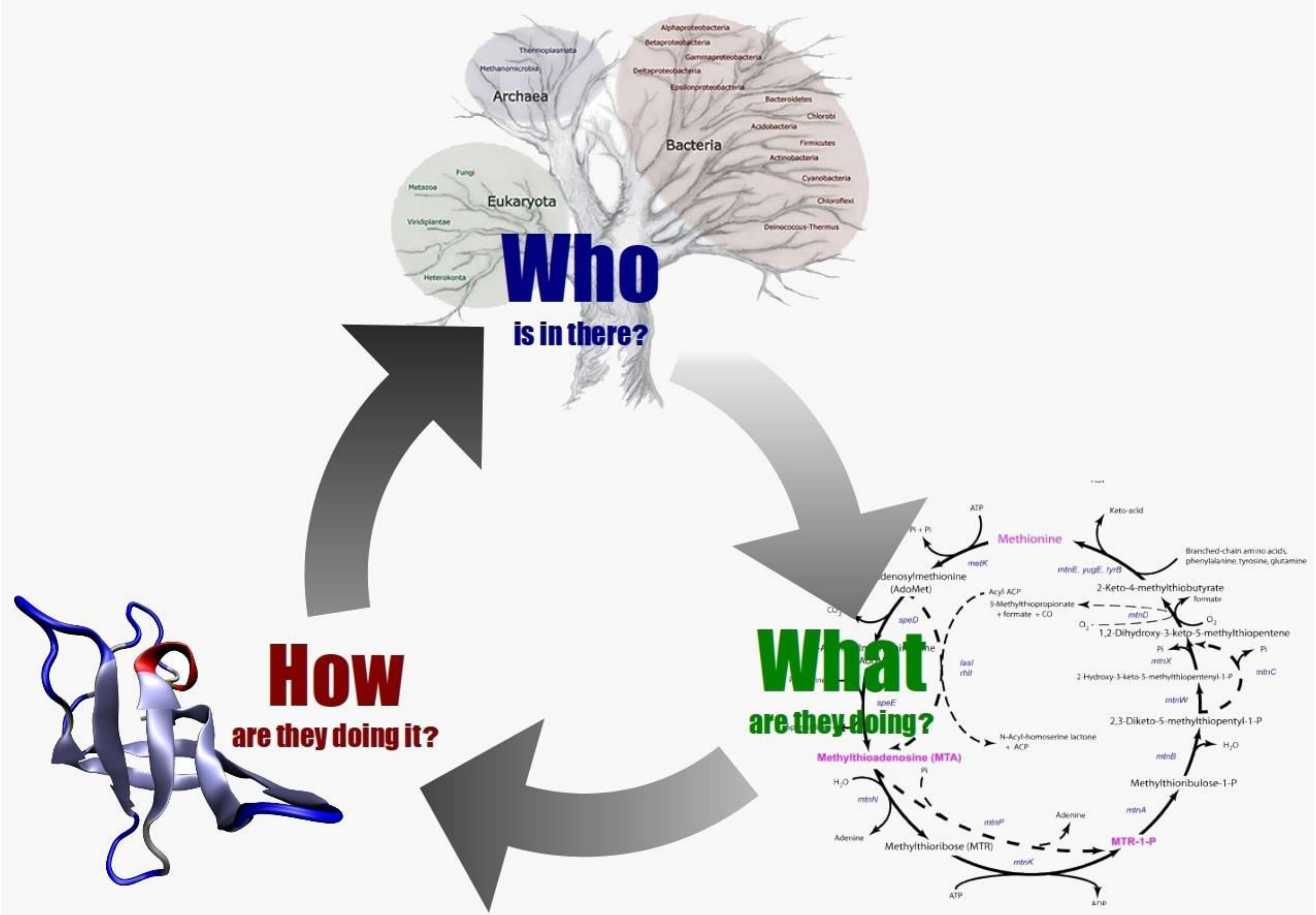


Slow progressor populations

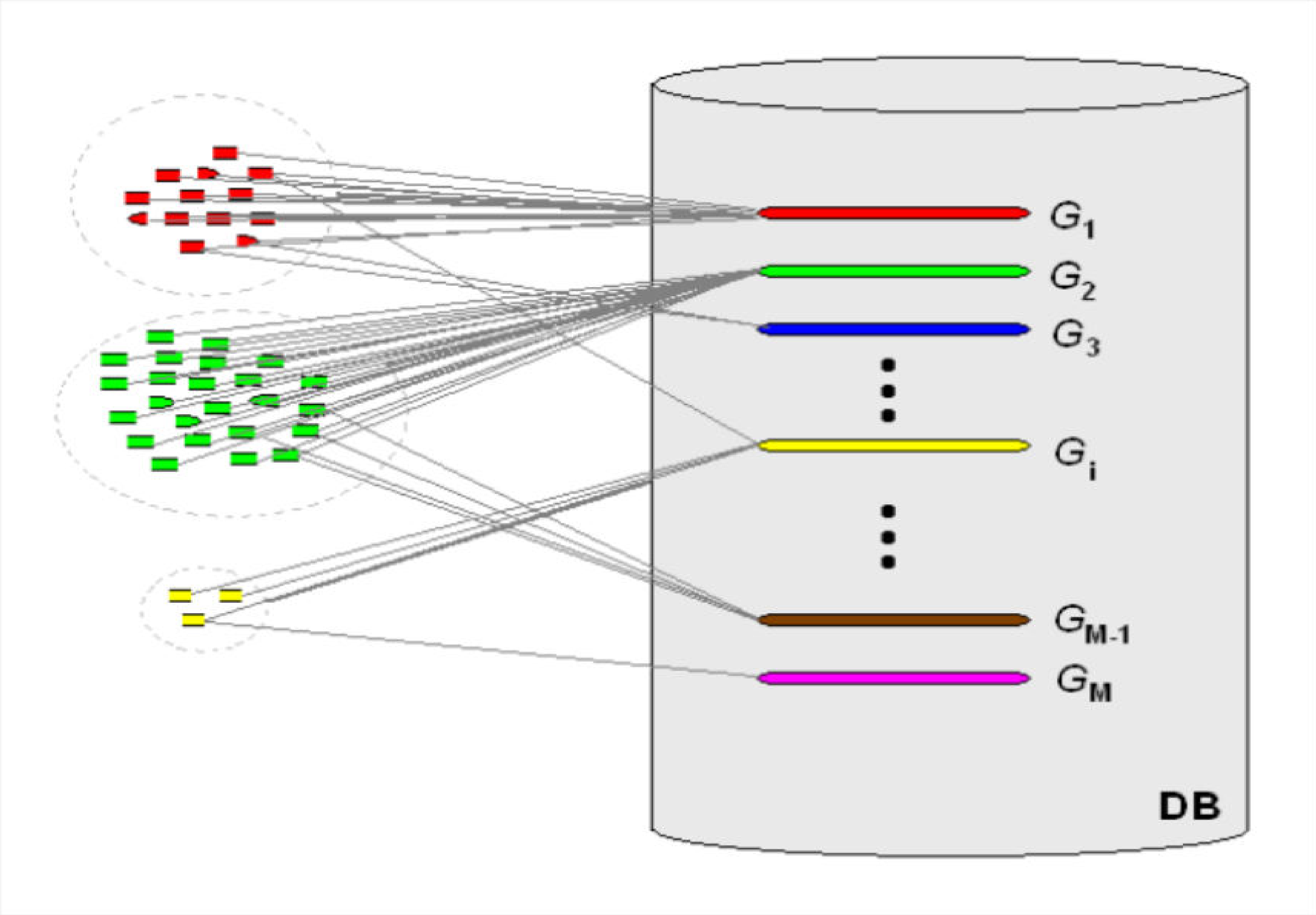


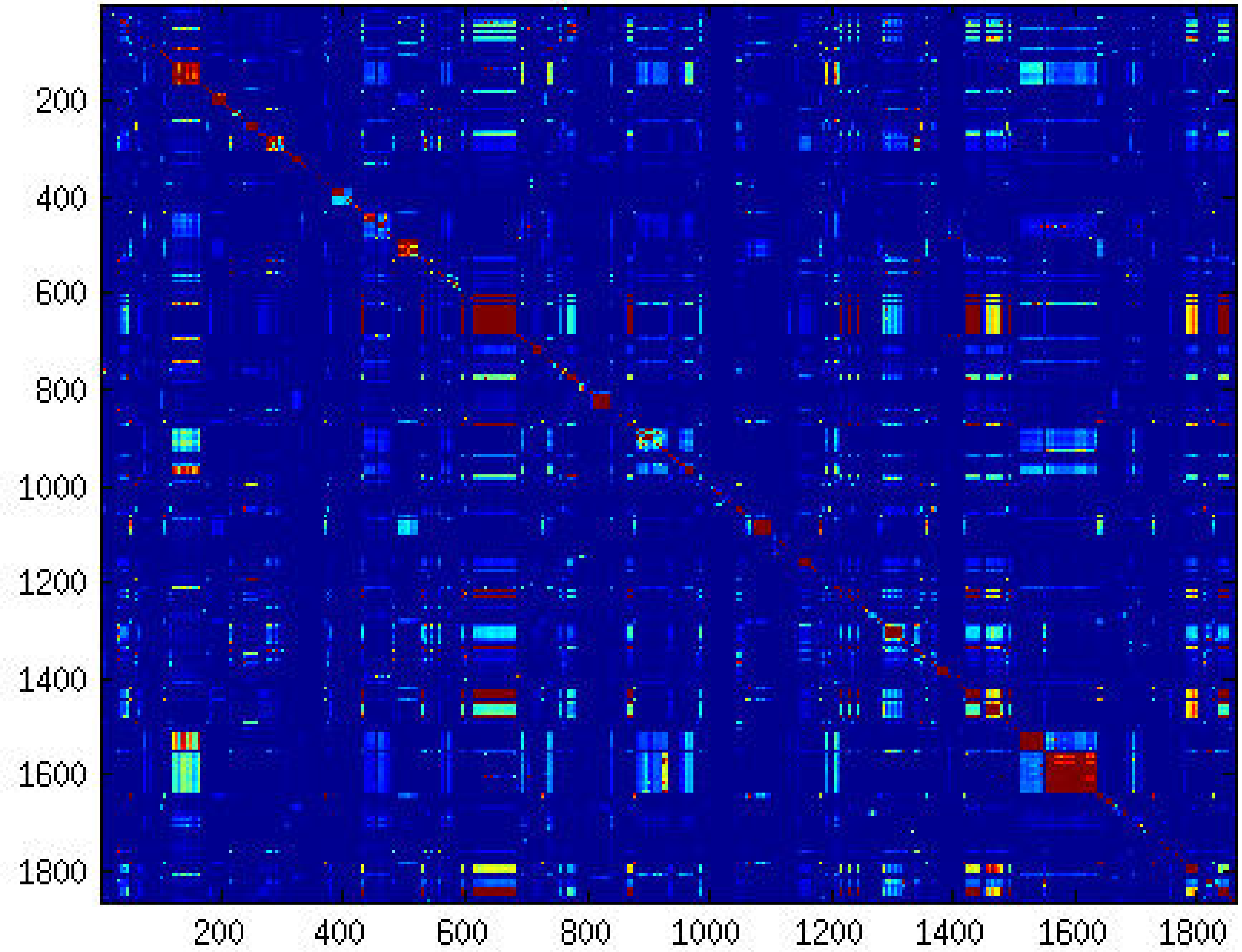
Rapid progressor populations

Metagenomics



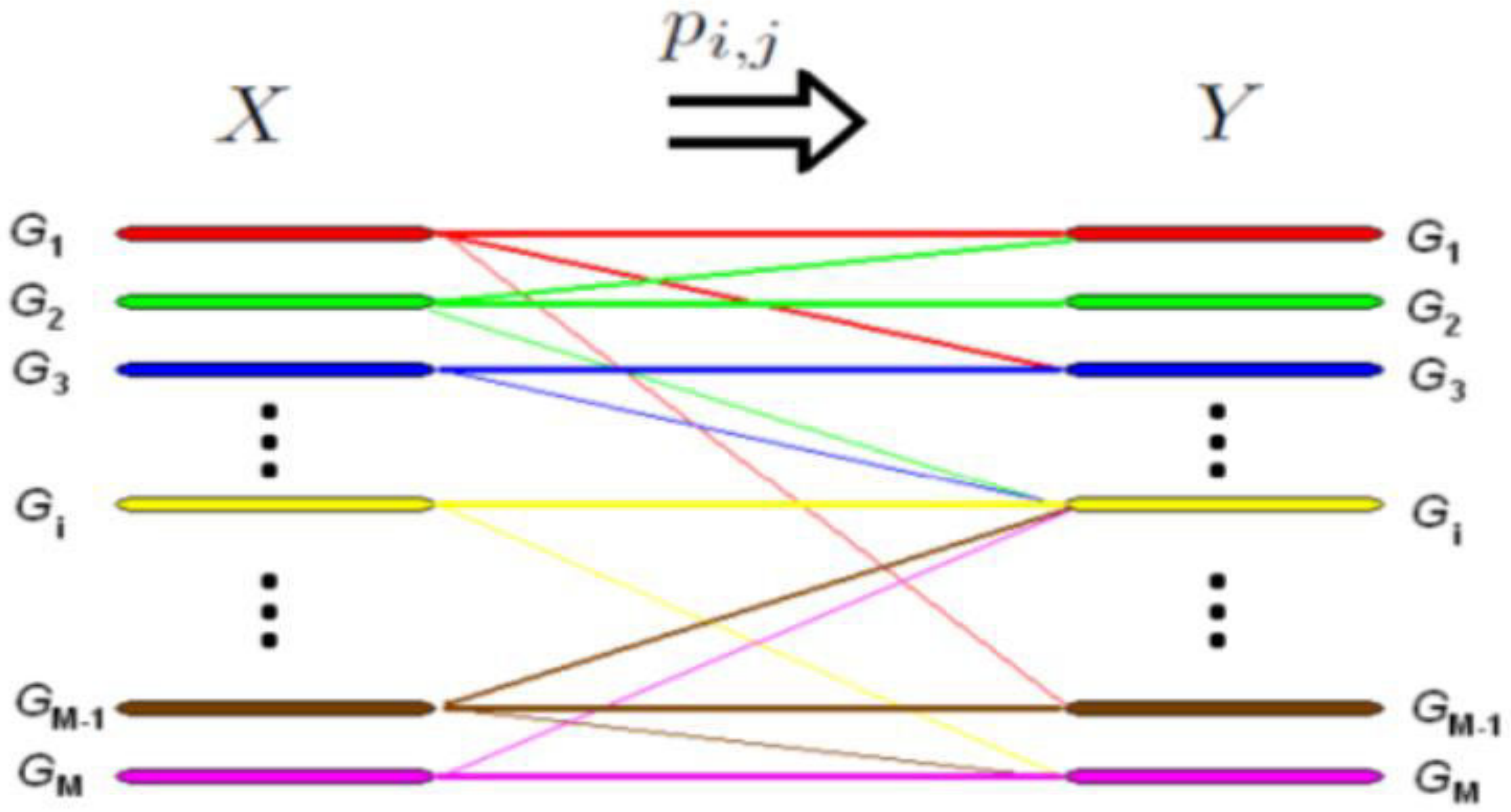
Similarities between genomes + single read processing = detection errors



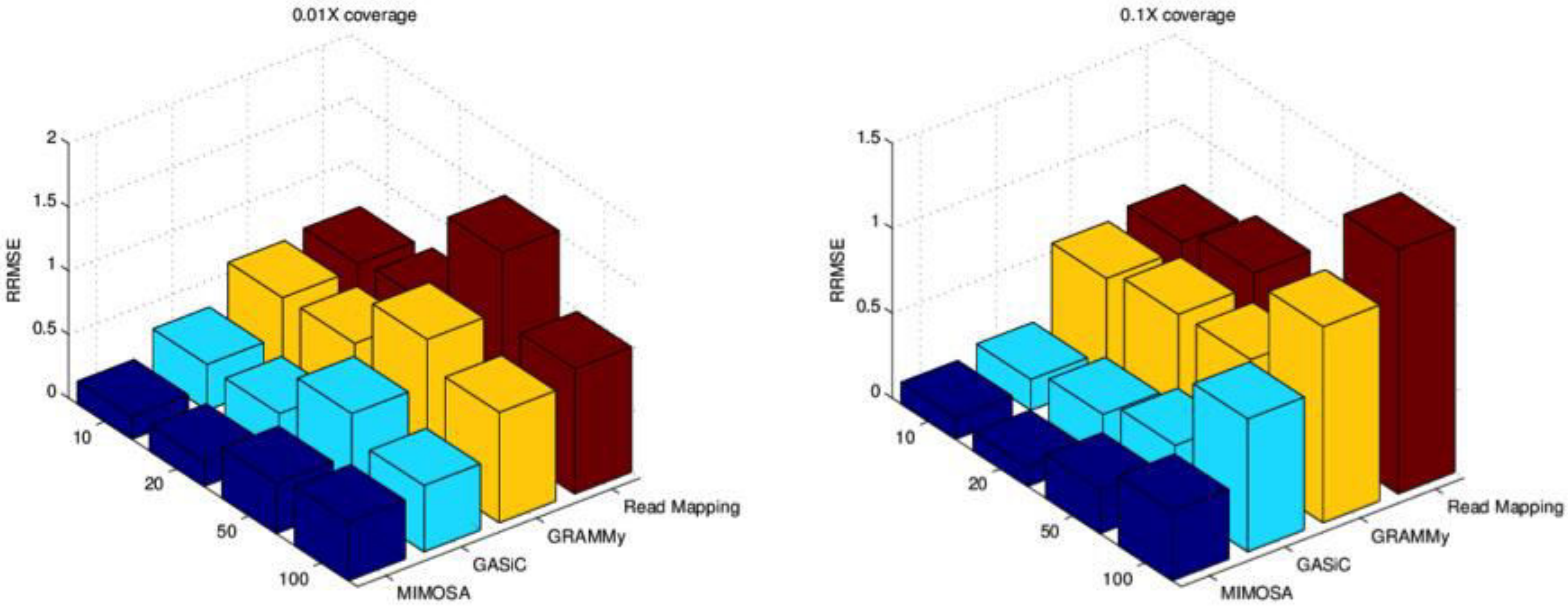


200
400
600
800
1000
1200
1400
1600
1800
200
400
600
800
1000
1200
1400
1600
1800

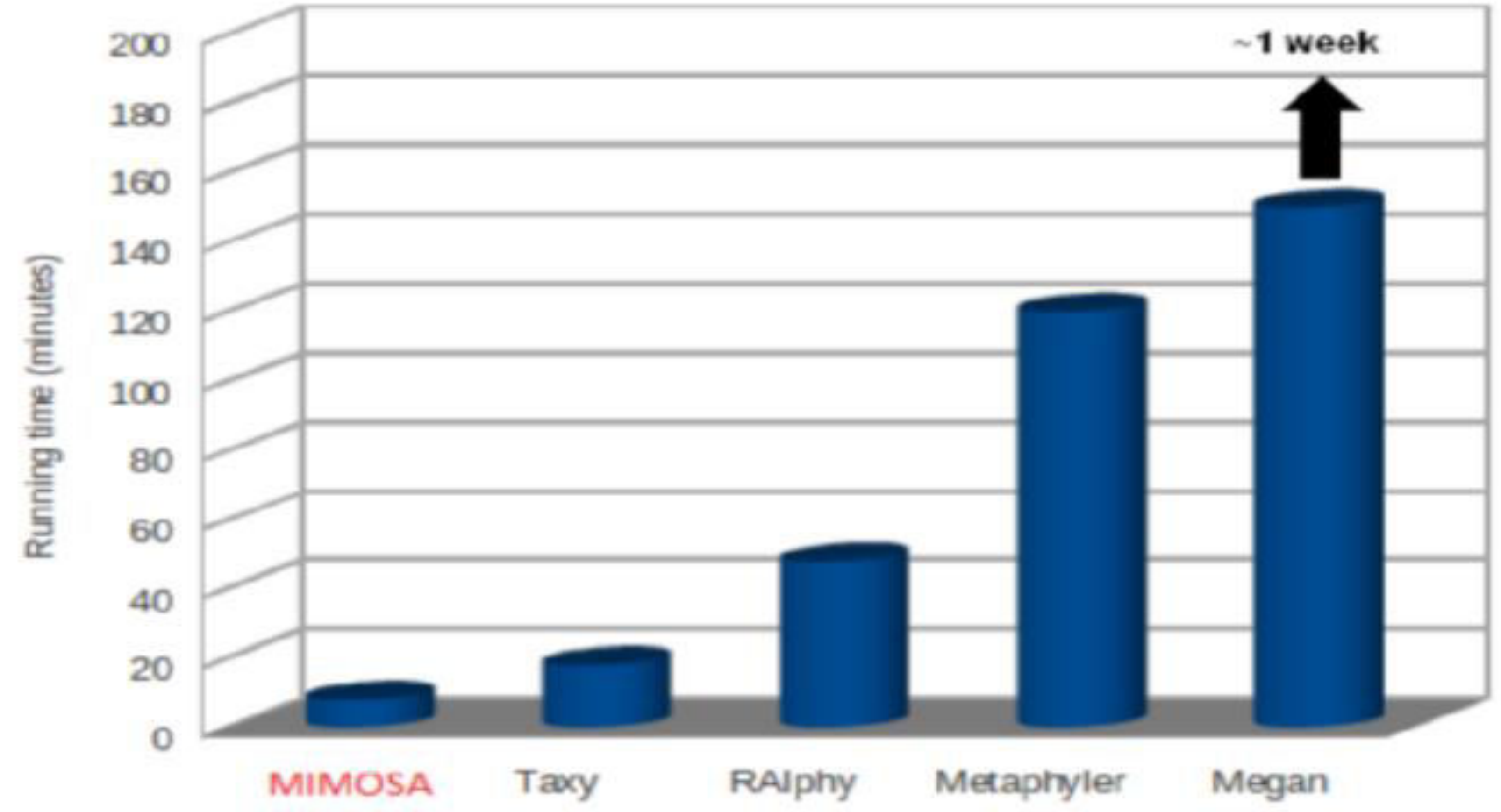
Learn channel statistics to estimate true abundance given the observed detection.



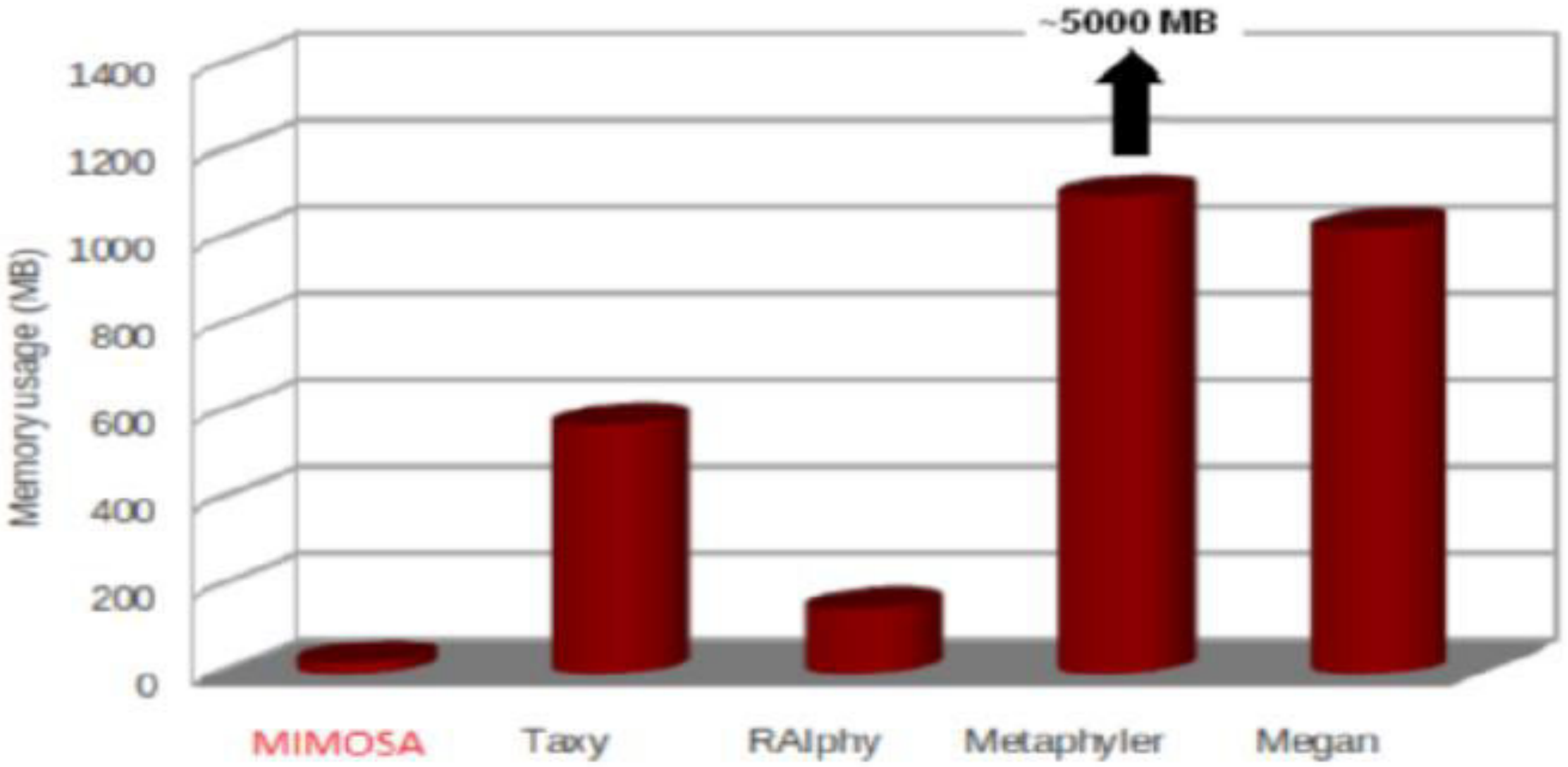
$$p_{i,j} = \frac{|G_i \cap G_j|}{|G_j|}$$



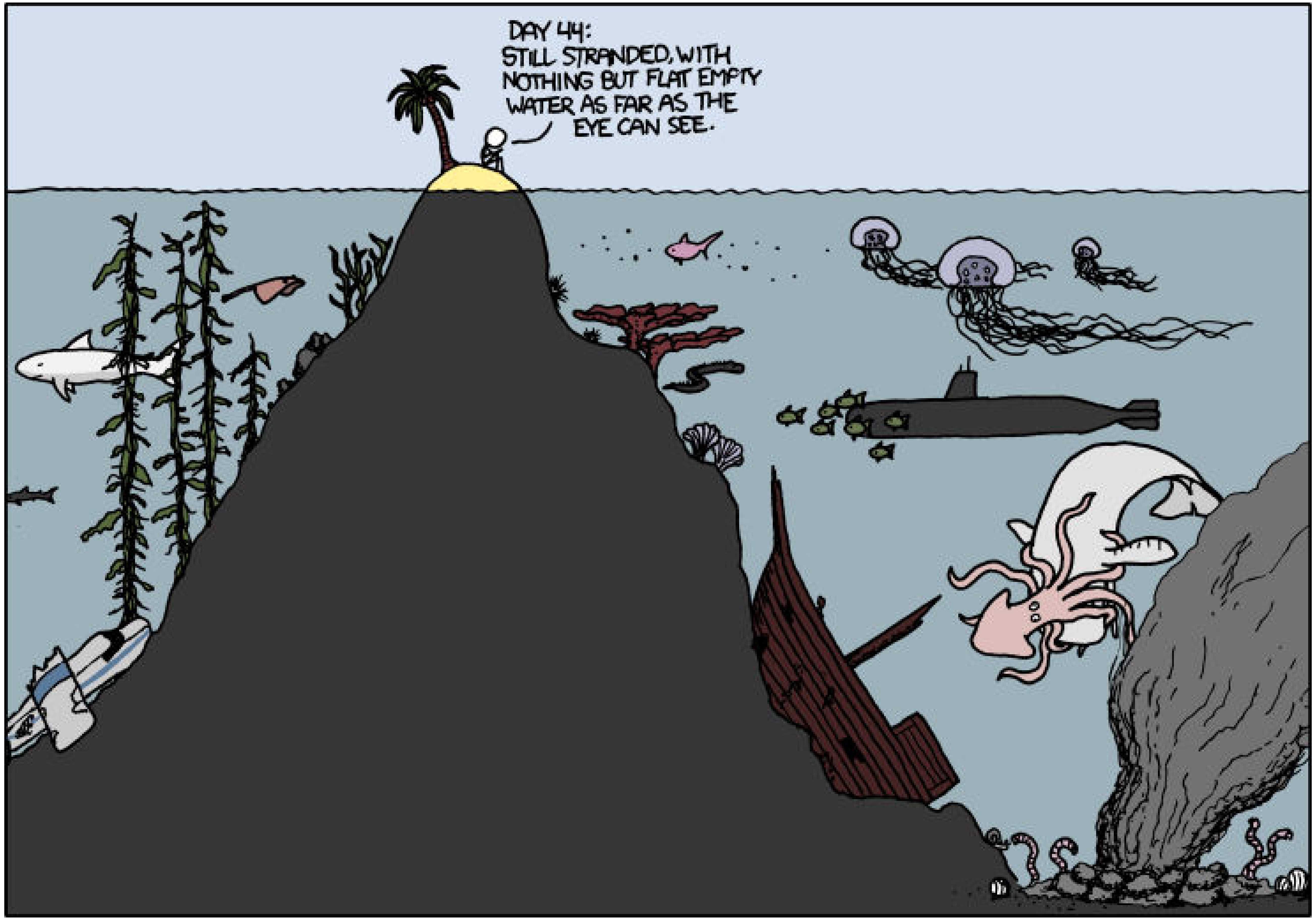
RRMSE for the simulated metagenomes corresponding to a mixture of 10, 20, 50, and 100 randomly selected organisms for 0.01X and 0.1X average genome coverages.



Running time (minutes)
0
20
40
60
80
100
120
140
160
180
200
~1 week
MIMOSA
Taxy
RAIphy
Metaphyler
Megan



Memory usage (MB)
0
200
400
600
800
1000
1200
1400
~5000 MB
MIMOSA
Taxy
RAIphy
Metaphyler
Megan


DAY 44:
STILL STRANDED, WITH NOTHING BUT FLAT EMPTY WATER AS FAR AS THE EYE CAN SEE.

Occult Information Lab

Bioinformatics

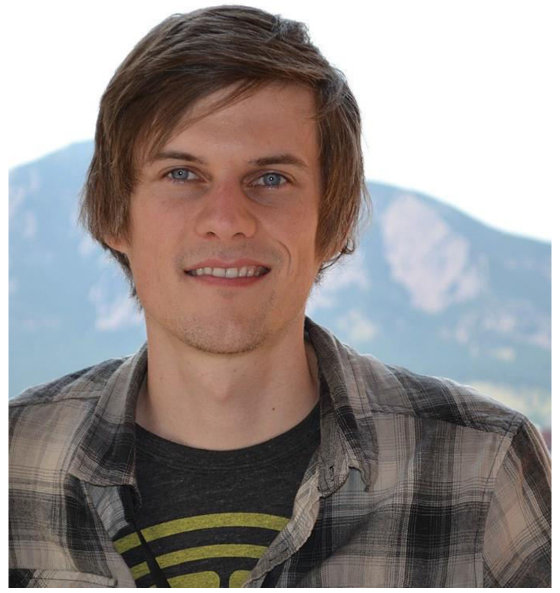
University of Nebraska Lincoln

Mark Bauer

Hasan Otu

David Russell

Ufuk Nalbantoglu



Sam Way

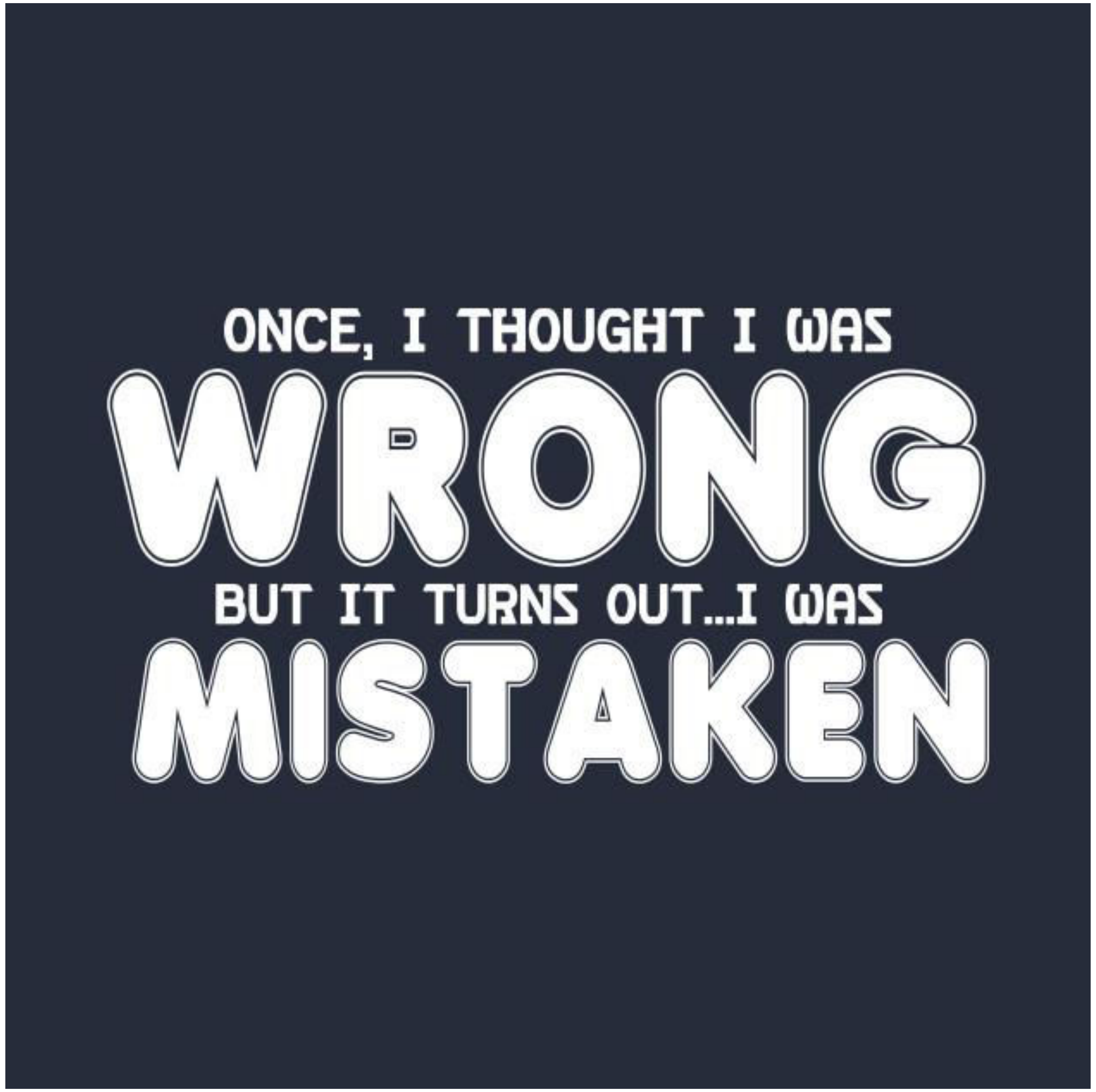
Garin Newcomb

Amirsalar Mansouri

Dicle Yalcin

Keith Murray

Jacob Bohac



ONCE, I THOUGHT I WAS

WRONG

BUT IT TURNS OUT...I WAS

MISTAKEN