Life at the Extremes: Harnessing Machine Learning for Biodiversity Informatics and Extremophile Genomics

Lila Kari

School of Computer Science

Faculty of Mathematics

University of Waterloo

Canada

Mathematical structures in genomes

- **Question:** Does **biological kinship** induce a detectable mathematical signature in genomes?

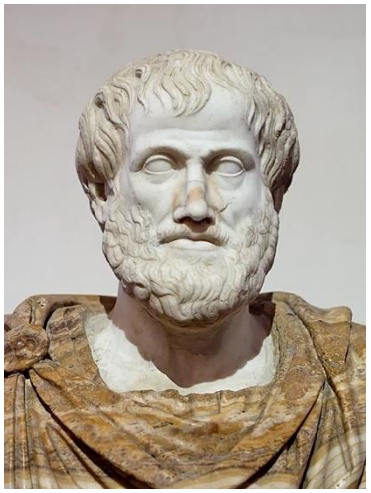
- **Question:** Can **the environment** induce a detectable, *kinship-independent*, mathematical signature in genomes?

Contents

- Mathematical representations of DNA
- Supervised machine learning for *taxonomic* classification (ML-DSP)
- Unsupervised clustering for *taxonomic* identification (iDeLUCS)
- Test the hypothesis of an *environmental signal* in extremophile genomes

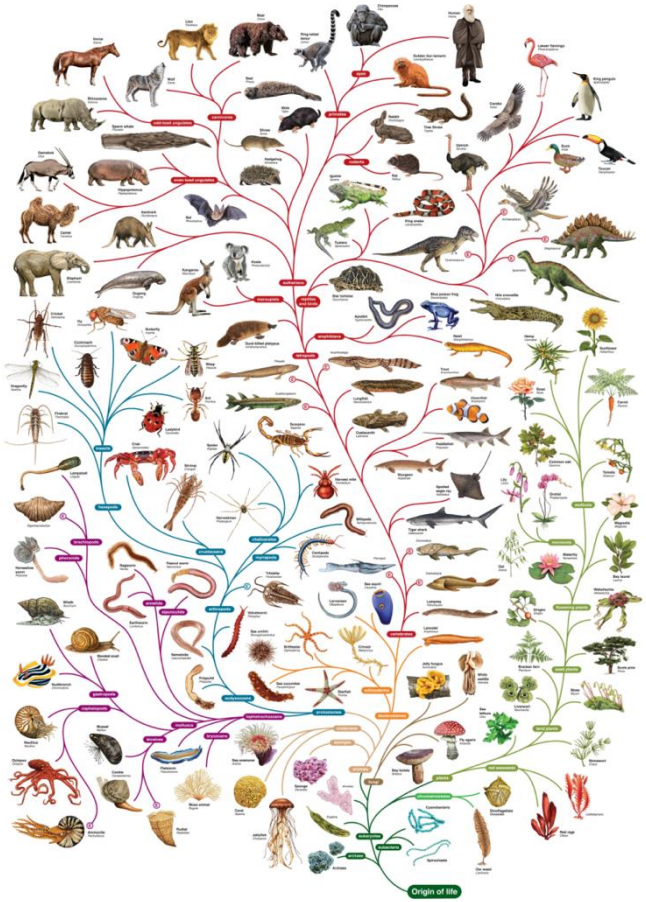
Taxonomy in Biology

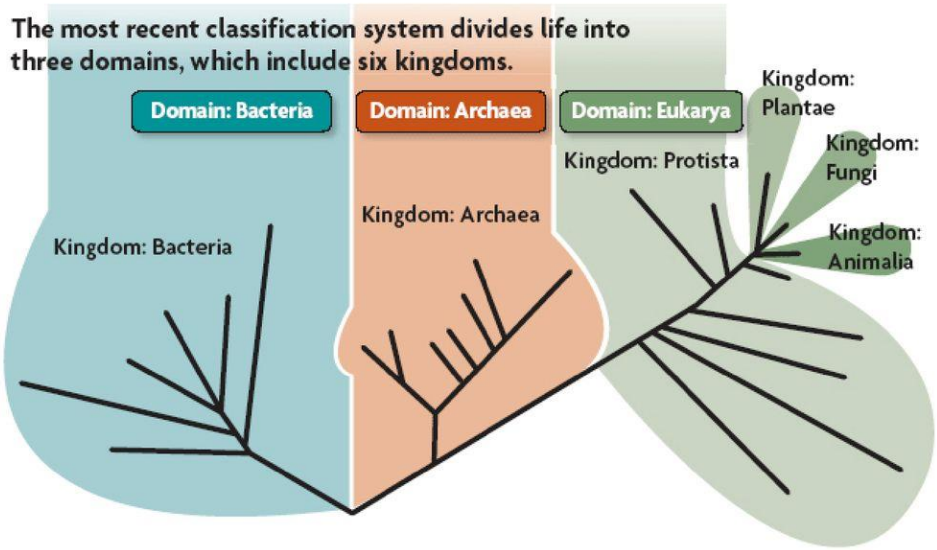
- Taxonomy – the branch of science that *groups* and *names* living organisms based on *relationships* inferred by shared characteristics



- Aristotle was the first scientist who attempted to classify organisms. He subdivided plants (into shrubs, herbs, trees) and animals (into air, water, land)

Darwinian evolution and the Tree of Life





Currently, DNA-based techniques such as **gene alignment** are used for taxonomic classifications

Lila Kari, University of Waterloo

Earth's biodiversity

- 95% of multicellular species on Earth do not yet have a scientific name

One of the newly named frogs, *Guibemantis ambakoana*. Ambakoana means 'living within Pandanus' in Malagasy. Image courtesy of Hugh Gabriel.

A Chaunacops, a genus of bony fish in the sea toad family, seen at a depth of nearly 1,400 m (4,560 ft) on Seamount SF2 inside Nazca-Desventuradas Marine Park. Image courtesy of Schmidt Ocean Institute. CC BY-NC-SA

This 'blob-headed' fish (Chaetostoma sp.), is new to science and was a shocking discovery due to its enlarged blob-like head, a feature that the fish scientists have never seen before, even though this species is already familiar to the Indigenous Awajun people who worked with scientists. It is a type of bristlemouth armored catfish. Photo courtesy of Conservation International / Robinson Olivera.

- Biodiversity Grand Challenge:

  Map all life on Earth by 2045!

Contents

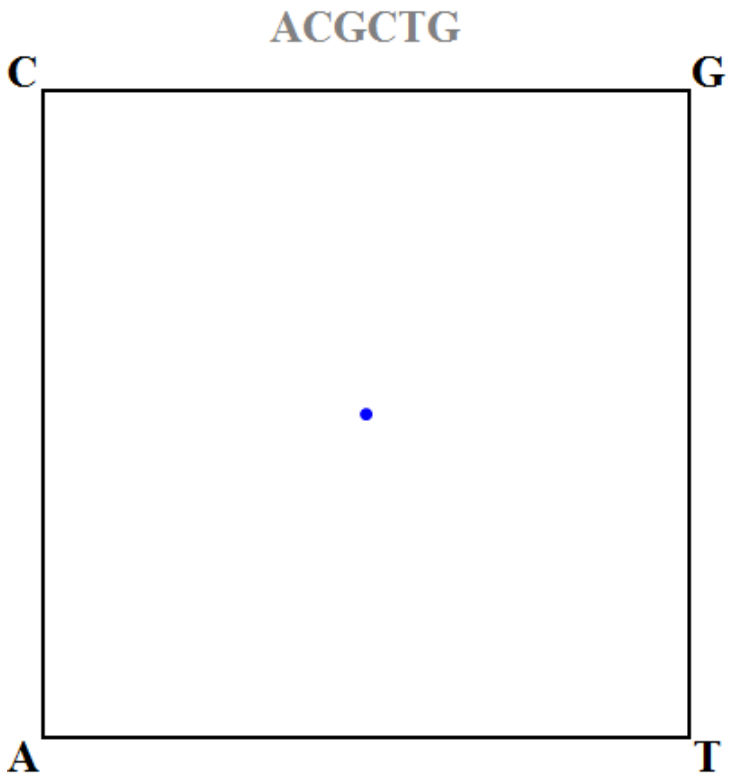
- Mathematical representations of DNA
- Supervised machine learning for *taxonomic* classification (ML-DSP)
- Unsupervised clustering for *taxonomic* identification (iDeLUCS)
- Test the hypothesis of an *environmental signal* in extremophile genomes

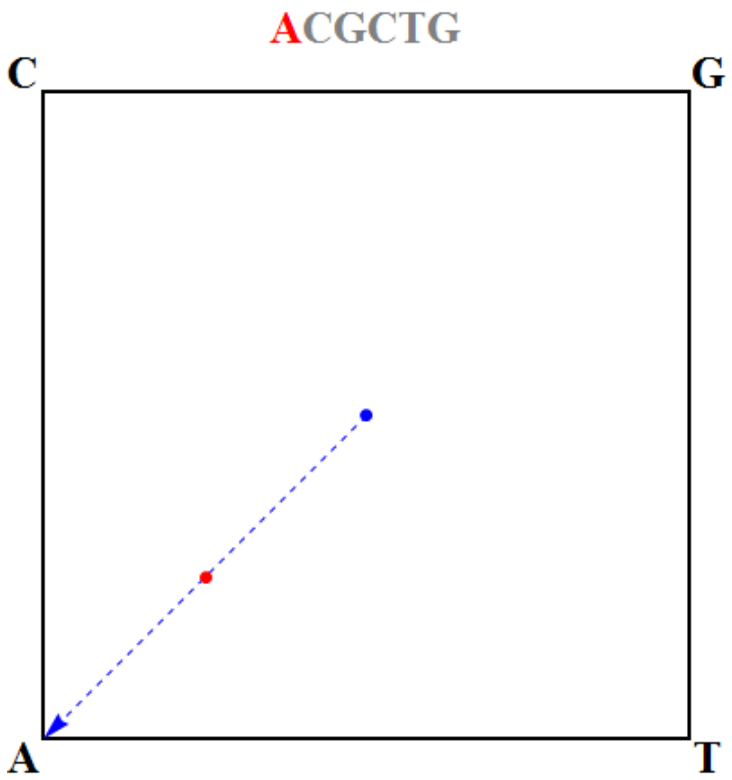
Chaos Game Representation (CGR) of DNA sequences [Jeffrey'90]

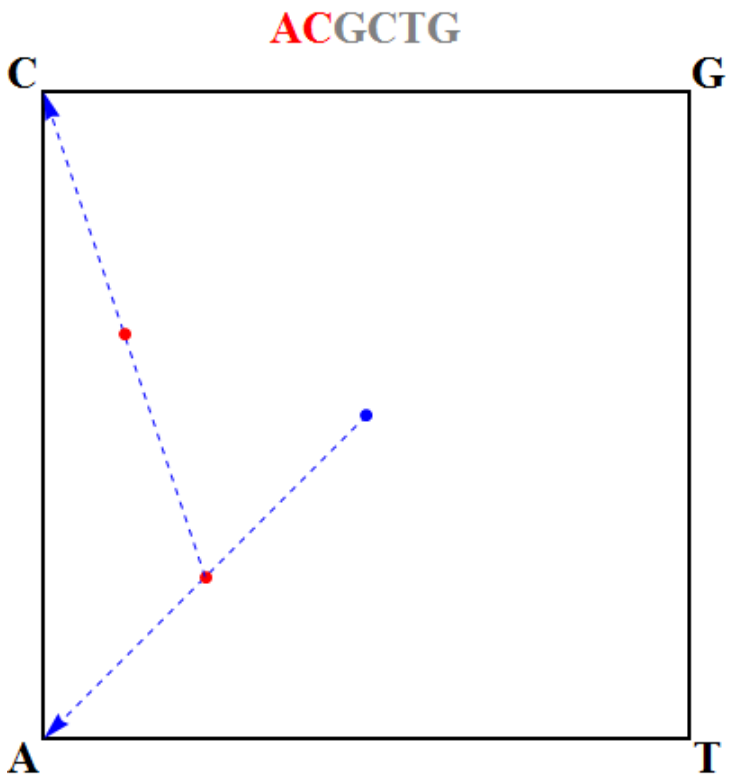
Start with a square with corners labelled *A, C, G, T*

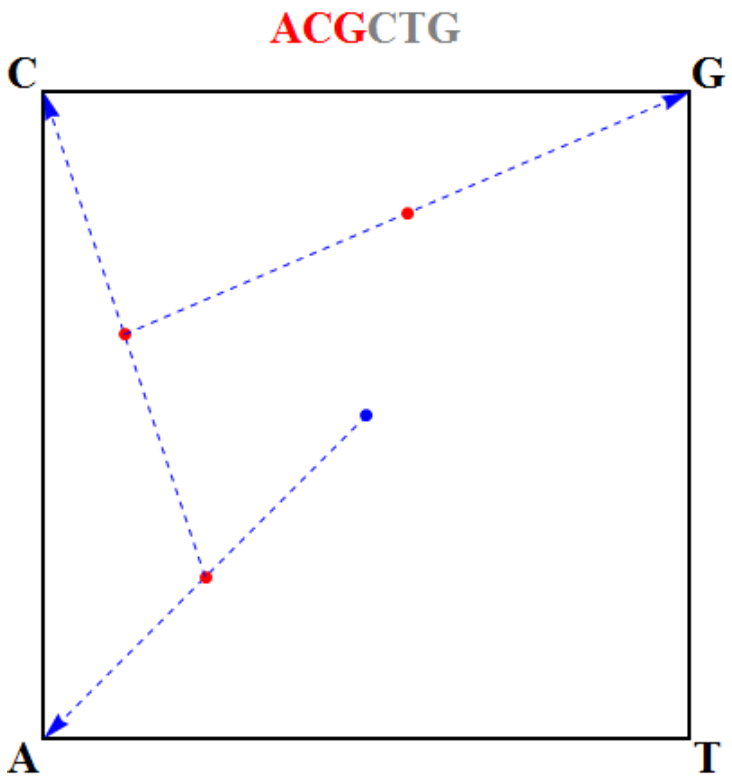
- The first point of the CGR is the square's center
- The DNA sequence is read left-to-right
- A new point is plotted midpoint between
  - * the current point, and
  - * the corner labelled by the DNA letter that is being read

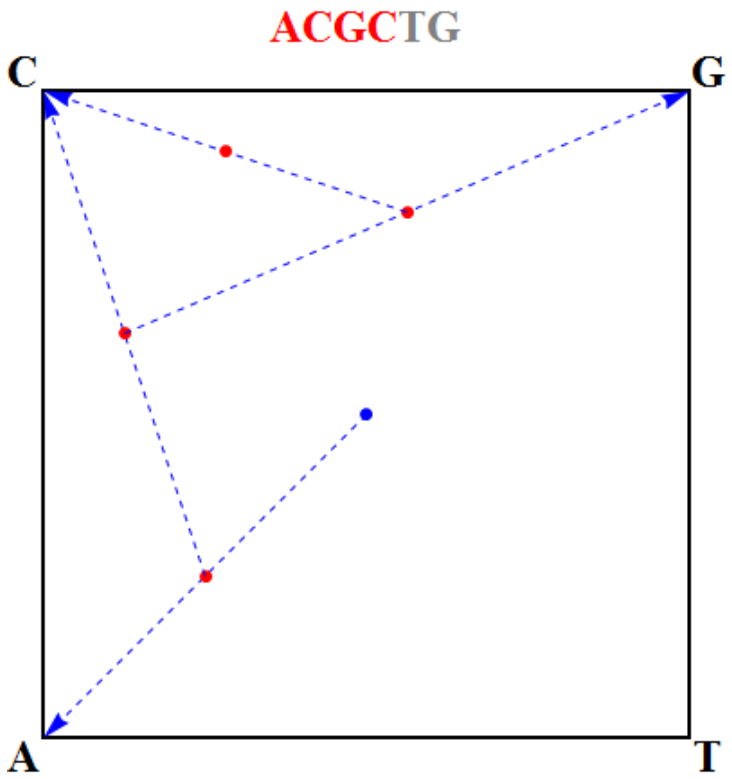
CGR of the DNA sequence *ACGCTG*

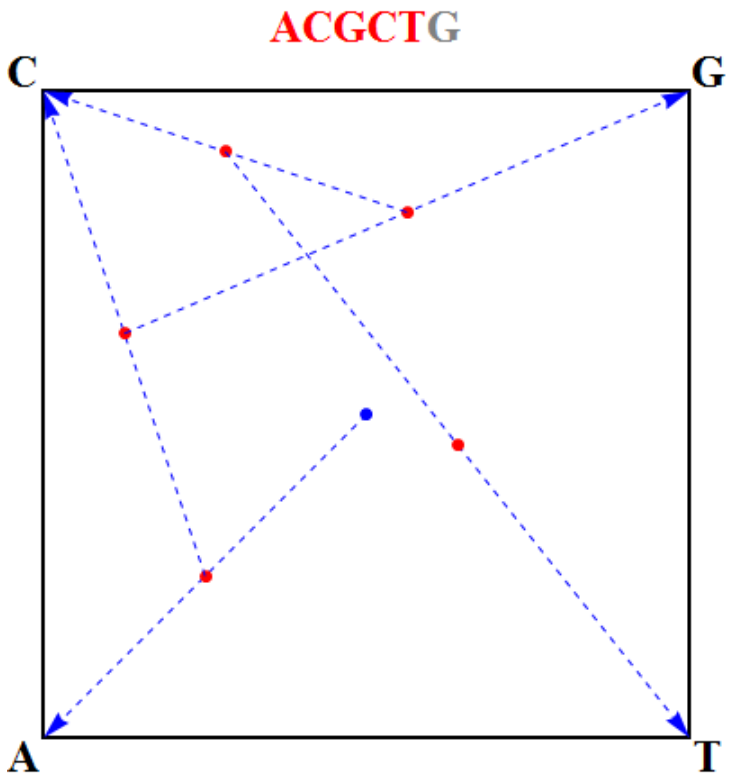


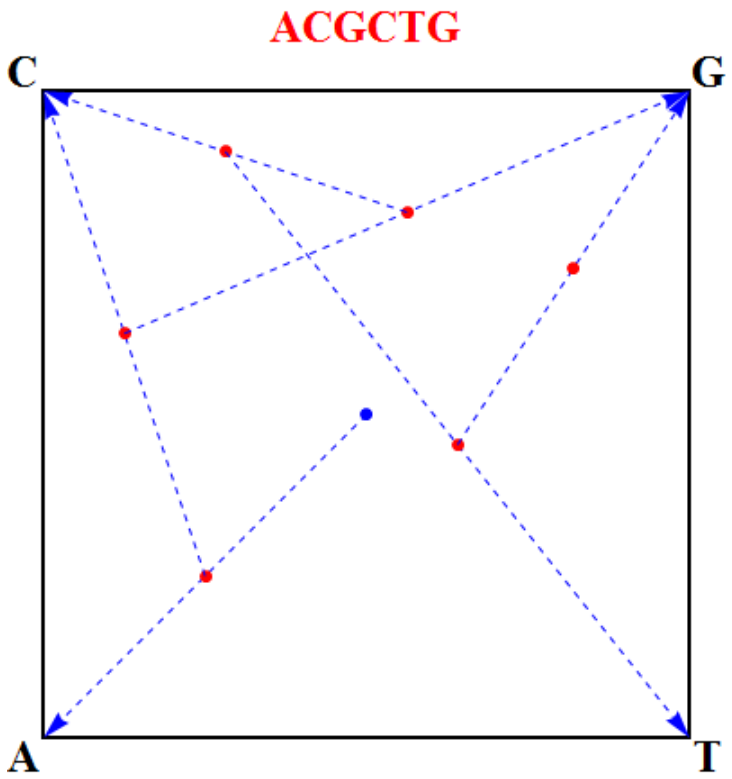

ACGCTG
C
G
A
T


ACGCTG
C
G
A
T

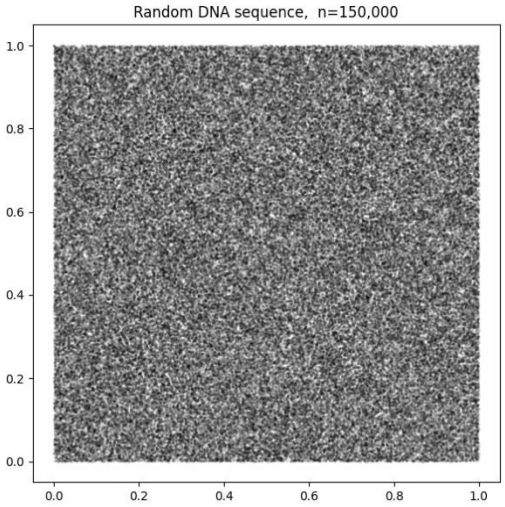

ACGCTG
C
G
A
T

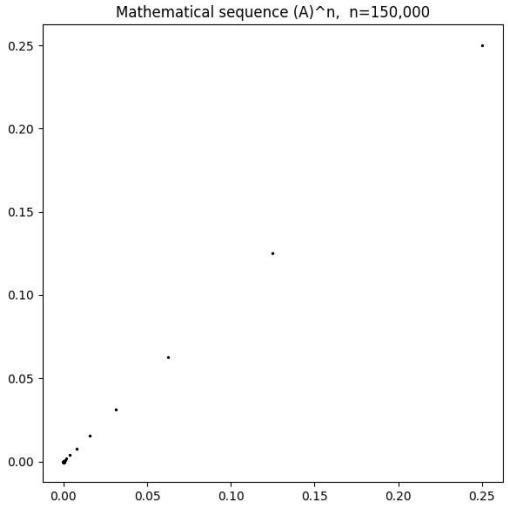

ACGCTG
C
G
A
T

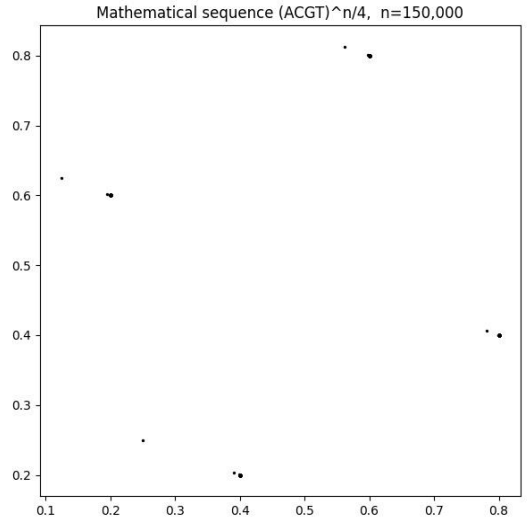

ACGCTG
C
G
A
T


ACGCTG
C
G
A
T

CGRs of computer-generated DNA sequences







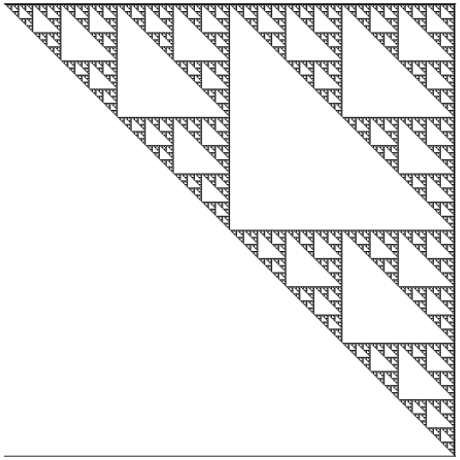


Figure 5: Avoided pattern: A

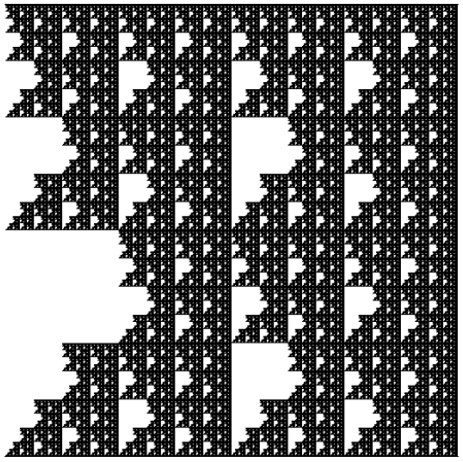


Figure 7: Avoided pattern: CA

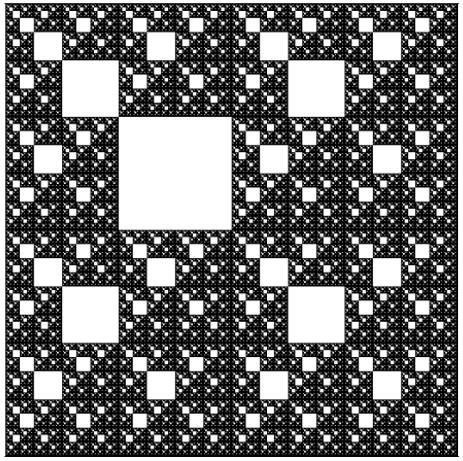
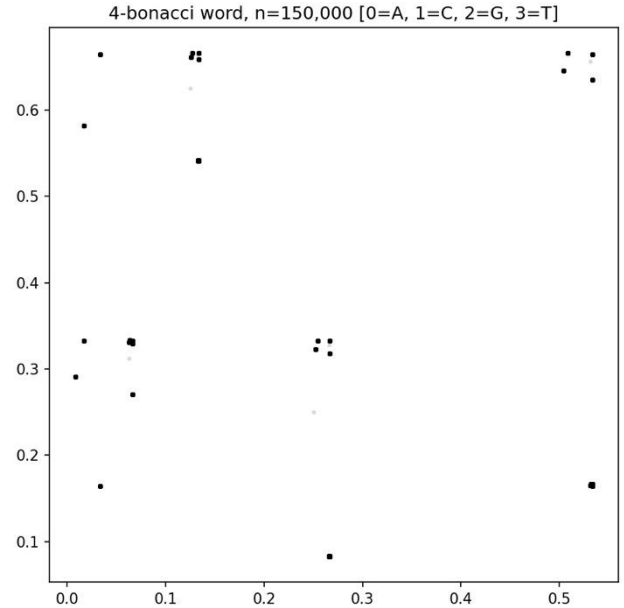
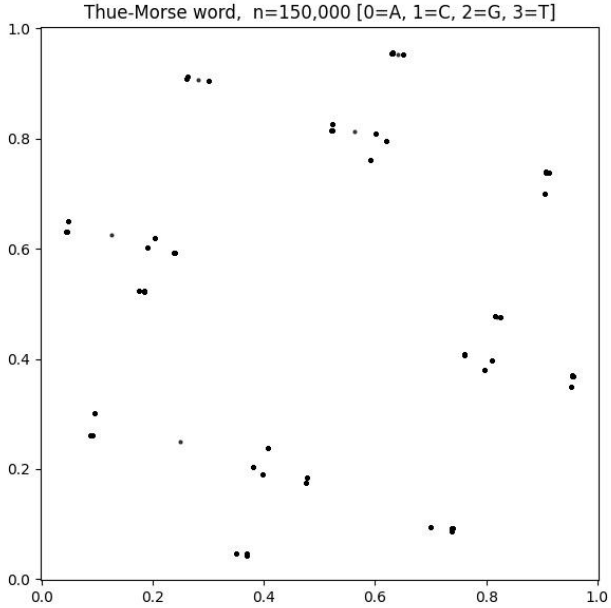


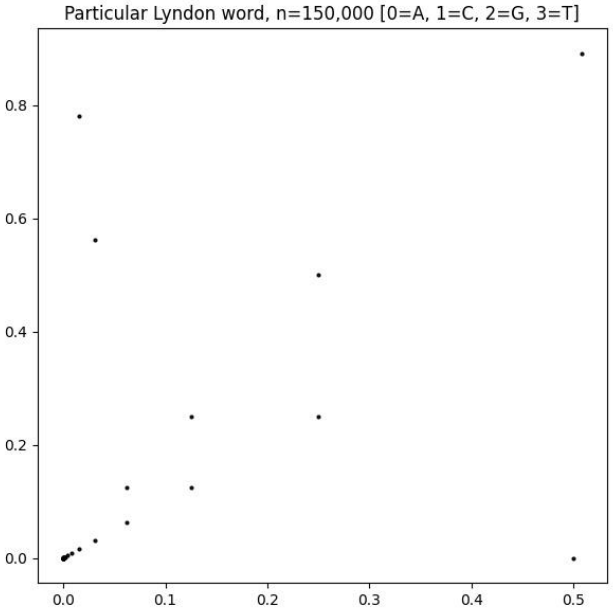
Figure 8: Avoided pattern: TC

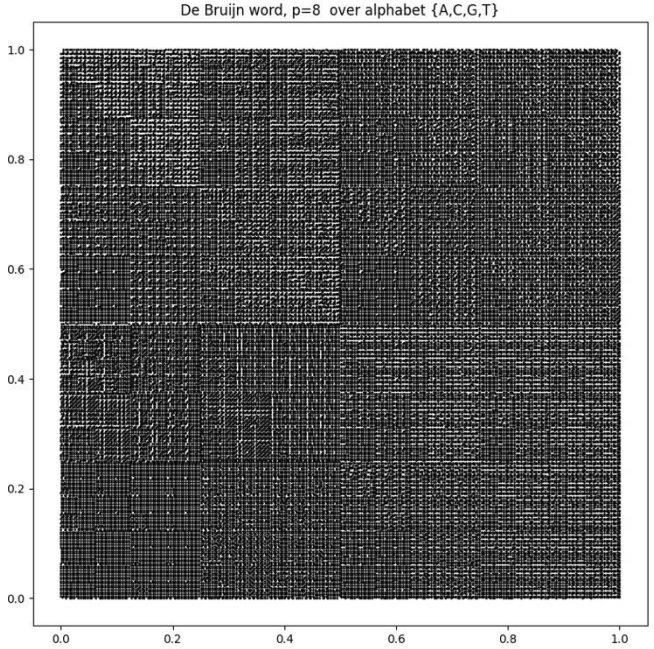
CGRs of mathematical sequences

- Fibonacci words $F_0 = 0, F_1 = 01, F_n = F_{n-1} F_{n-2},$
  (or fixed point of morphism *0-> 01, 1-> 0*)
- *M*-bonacci words over *{0, 1,2, 3}* - fixed point of morphism *0->01, 1-> 02, 2 -> 03, 3-> 0*
- Thue-Morse word over *{0, 1, 2, 3}* - *f*ixed point of morphism *0 ->012, 1-> 123, 2-> 230, 3-> 301*
- Lyndon word over *{0, 1, 2, 3}* - a non-empty word that is smaller (in lexicographic order) than all of its rotations
- De-Bruijn word of order *p* - concatenate, in lexicographic order, all Lyndon words whose length divides *p*

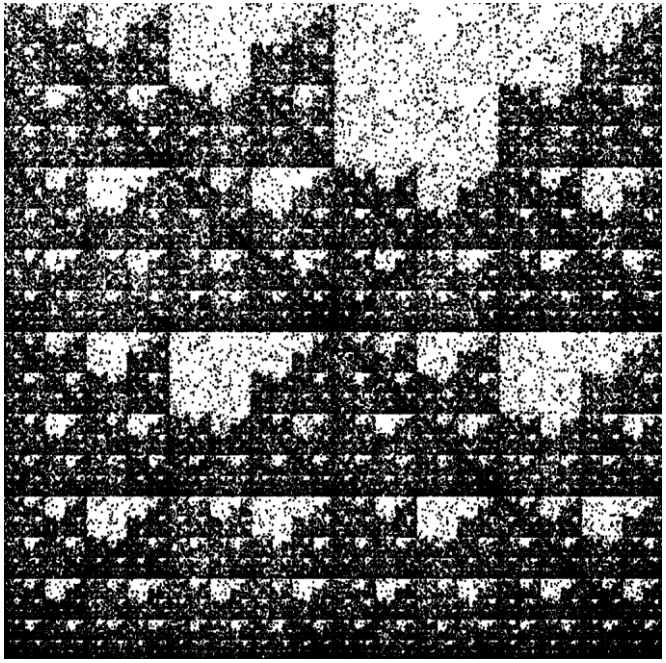

4-bonacci word, n=150,000 [0=A, 1=C, 2=G, 3=T]
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.1
0.2
0.3
0.4
0.5


Thue-Morse word, n=150,000 [0=A, 1=C, 2=G, 3=T]
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

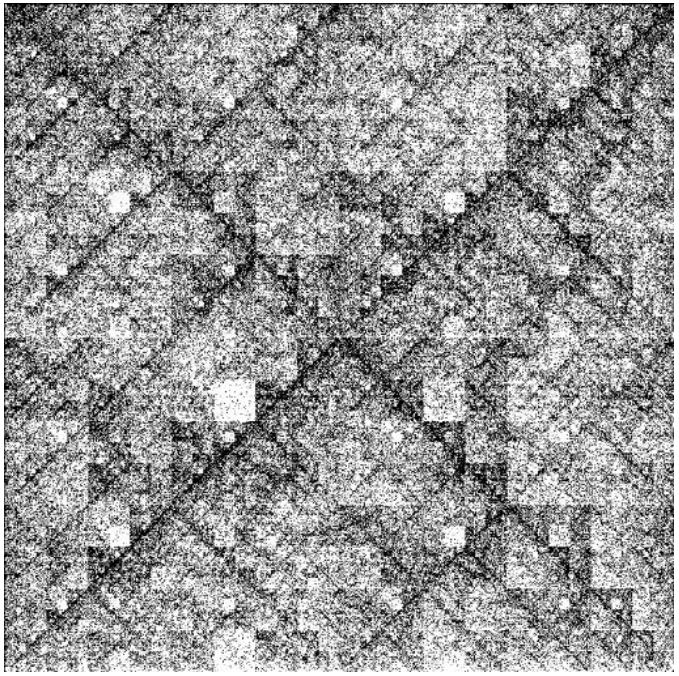

Particular Lyndon word, n=150,000 [0=A, 1=C, 2=G, 3=T]
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5


De Bruijn word, p=8 over alphabet {A,C,G,T}
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

CGRs of (natural) genomic DNA



(i) *H. sapiens* (nuclear) genome



(ii) *A. fulgidus* genome

Lila Kari, University of Waterloo

From *CGR* to $FCGR_k$ (frequency *CGR* of order *k*)

- A *k*-mer is a DNA word of length *k*
- There are 4^k possible distinct *k*-mers
- In a 2^k x 2^k grid, each *k*-mer falls into a particular grid cell
- The value of *k* determines the resolution of the CGR

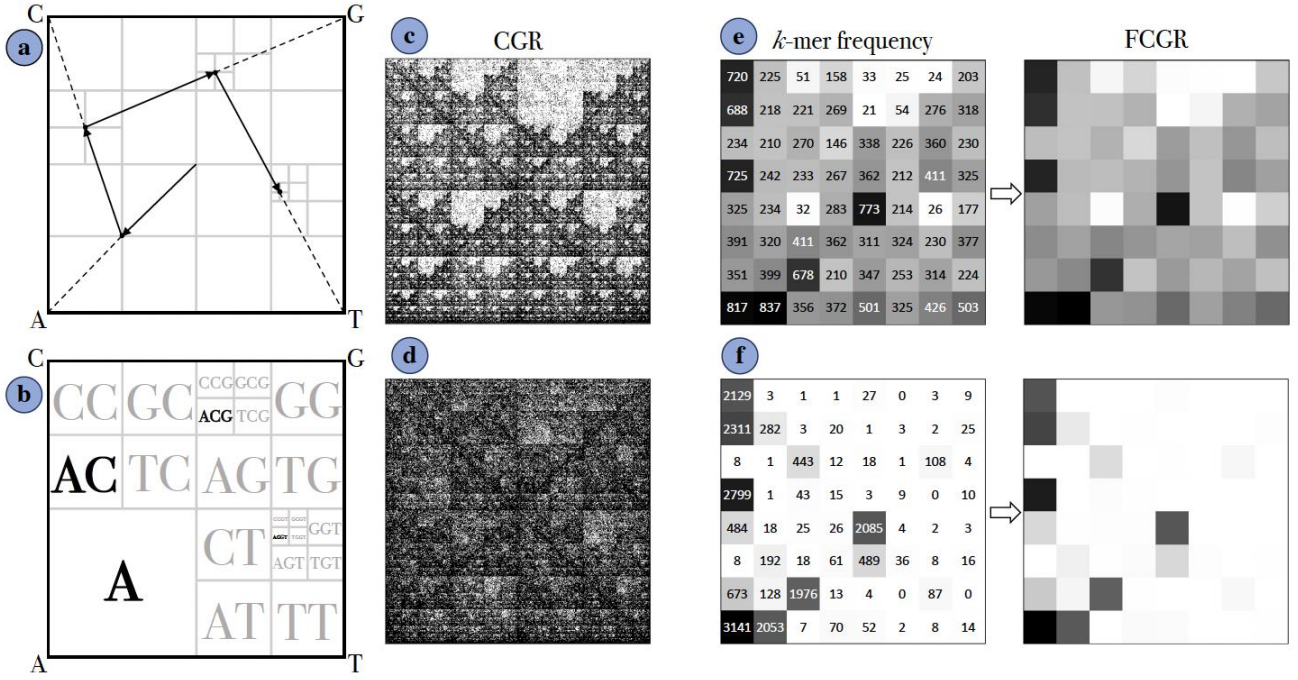
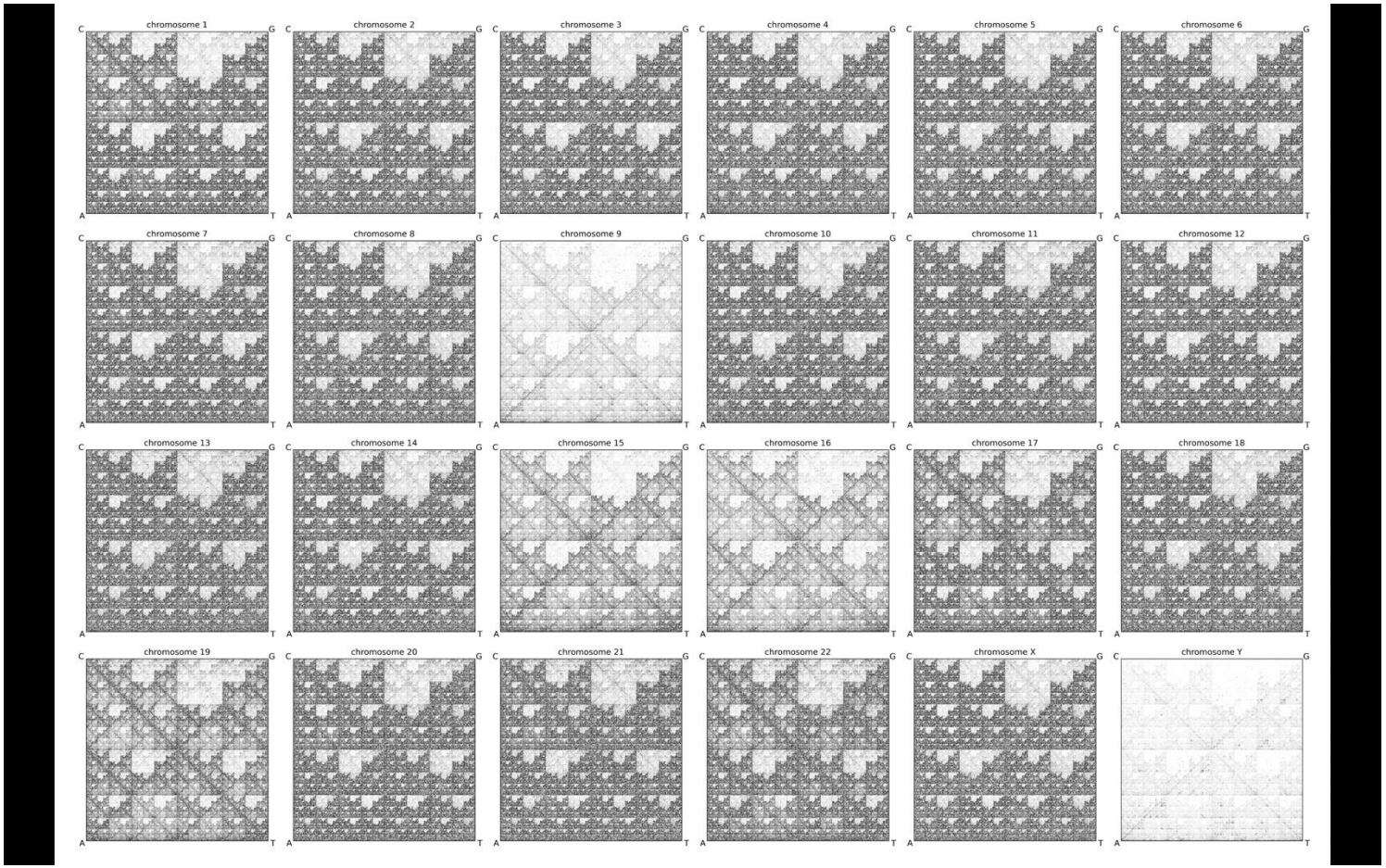


Figure 2. CGR/FCGR image generation. a. A schematic of CGR image generation from a DNA sequence. **b.** Mapping of *k*-mers to specific positions in the CGR. **c, d.** Examples of CGR images (512×512, $k = 9$) for human (panel c) and maize (panel d). **e, f.** Generating FCGR images by counting *k*-mer frequencies ($k = 3$) for human and maize, respectively.

CGR as ``*genomic signature*”

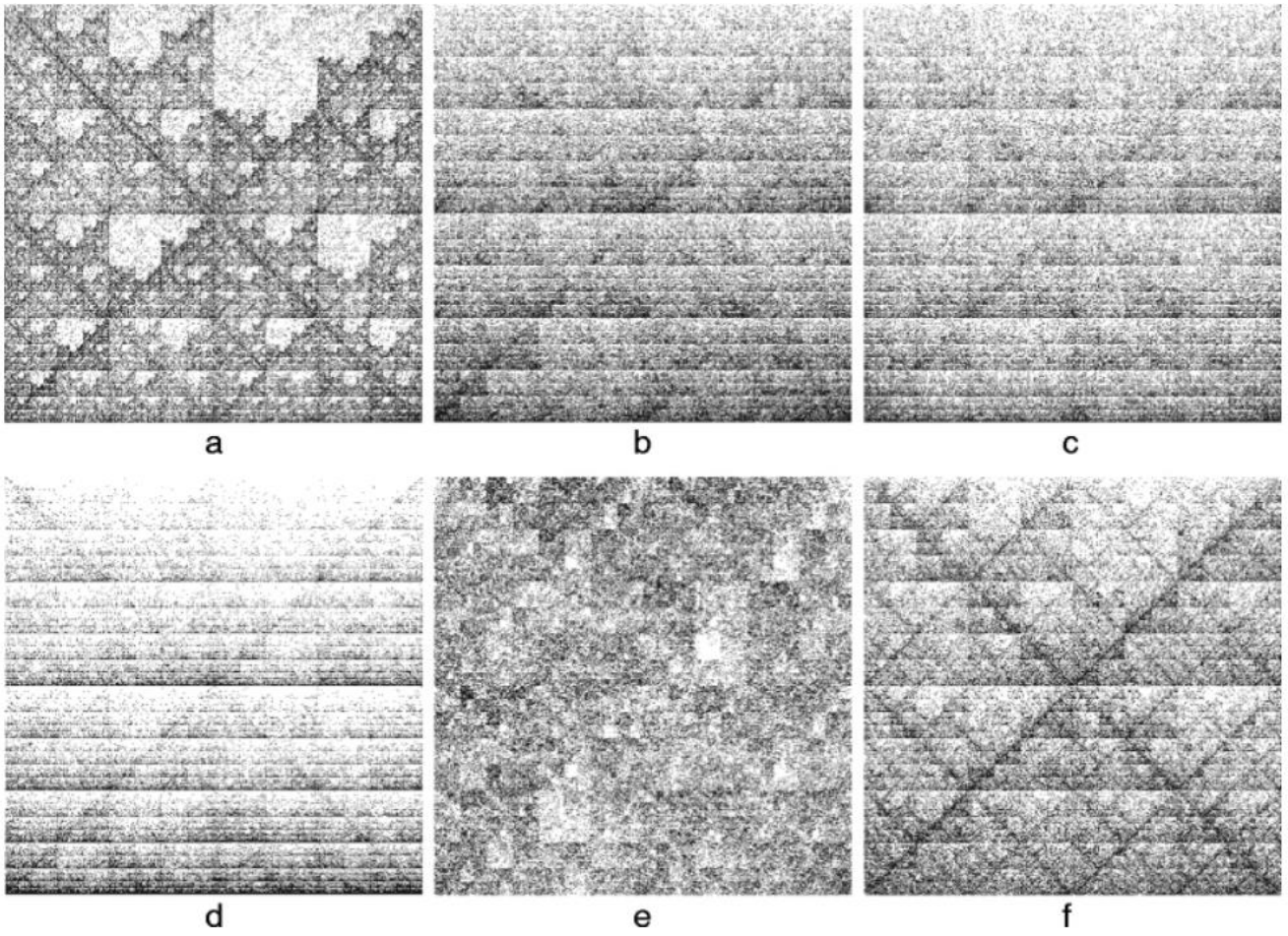
- Real life genomic CGRs are species-specific
- Patterns are *similar* for DNA sequences from the *same* genome
- Patterns are *different* for DNA sequences from *different* species [Deschavanne+99]
- Patterns are preserved regardless of the length and location of the DNA sequence
- This qualifies CGR as a *genomic signature* [KarlinBurge95]

CGRs of human chromosomes



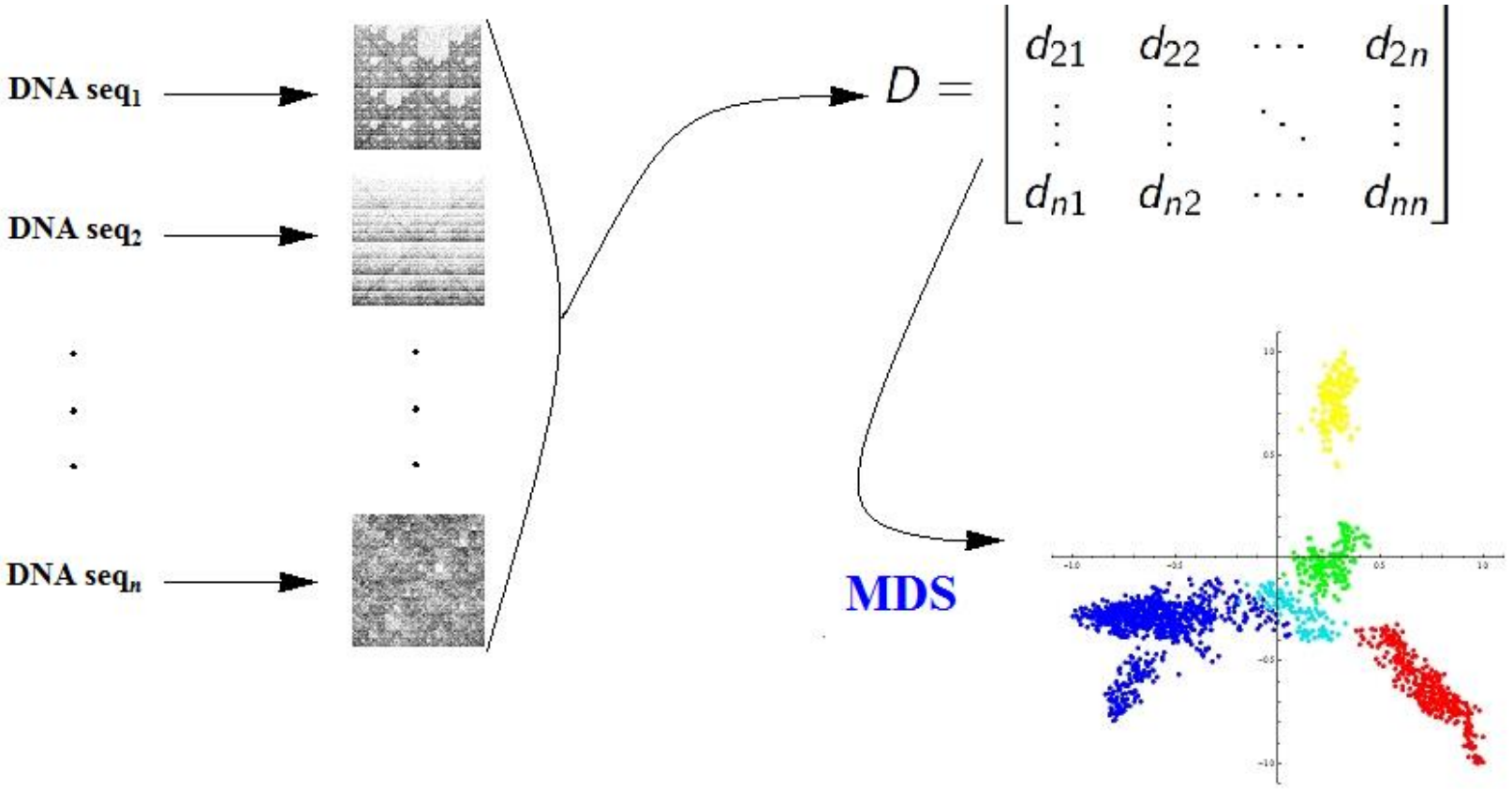
[N. Sadjadi, C. de Souza, G. Randhawa, K. Hill, L.Kari. *Scientific Reports* 2026]

CGRs of genomic DNA from organisms in the 6 Kingdoms of life



CGRs of 150,000 long DNA genomic fragments of organisms from Kingdoms
a: Animalia, b: Fungi, c: Plantae, d: Protista, e: Bacteria, f: Archaea

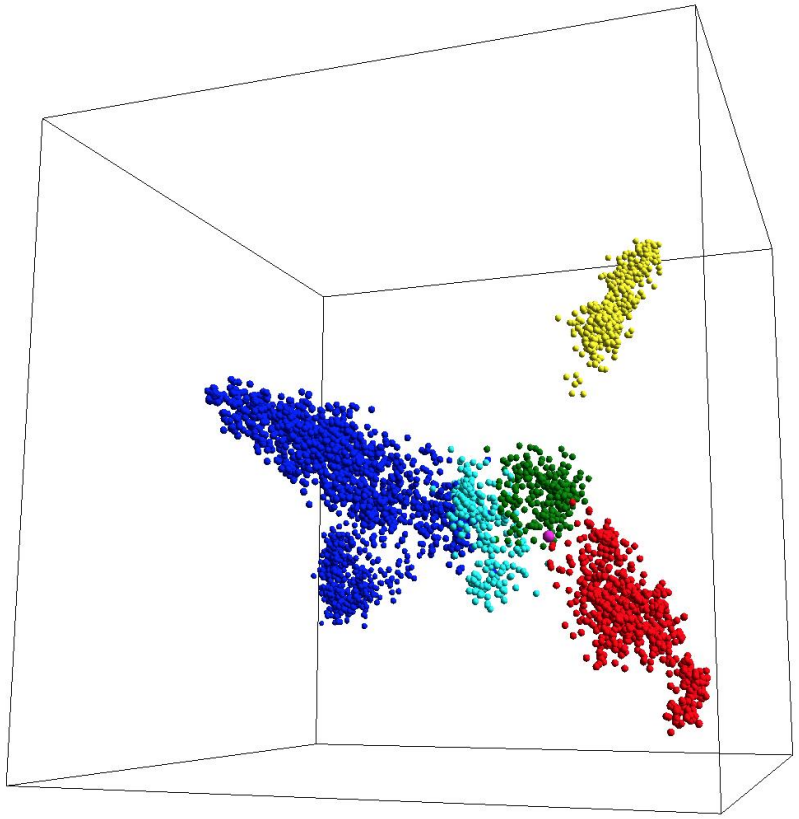
CGRs ➡ distance ➡ classification



Lila Kari, University of Waterloo

MoDMap interactive webtool

https://moleculardistancemaps.github.io/MoDMaps3D/



4,322 vertebrate mitochondrial DNA genomes (~16,000 bp)

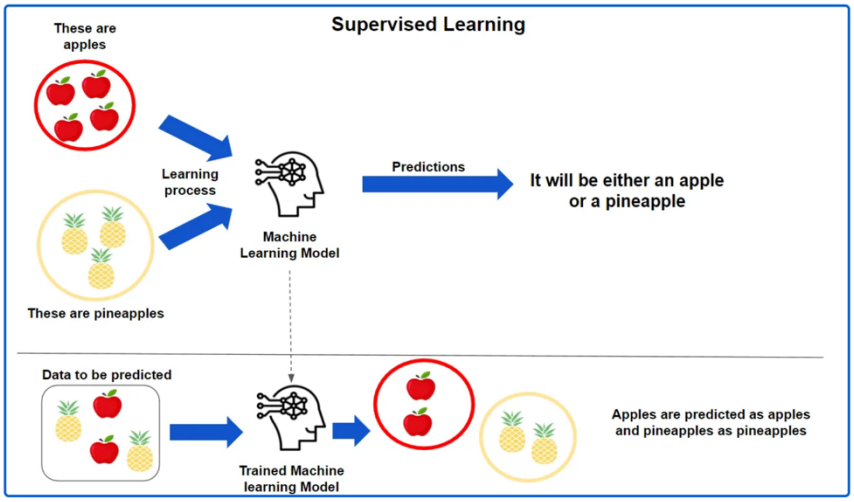
Blue = fish; Aqua = reptiles, Green = amphibians, Yellow= birds, Red = mammals

[Karamichalis, Kari, *Bioinformatics*, 2017]

Contents

- Mathematical representations of DNA
- Supervised machine learning for *taxonomic* classification (ML-DSP)
- Unsupervised clustering for *taxonomic* identification (iDeLUCS)
- Test the hypothesis of an *environmental signal* in extremophile genomes

Supervised Machine Learning

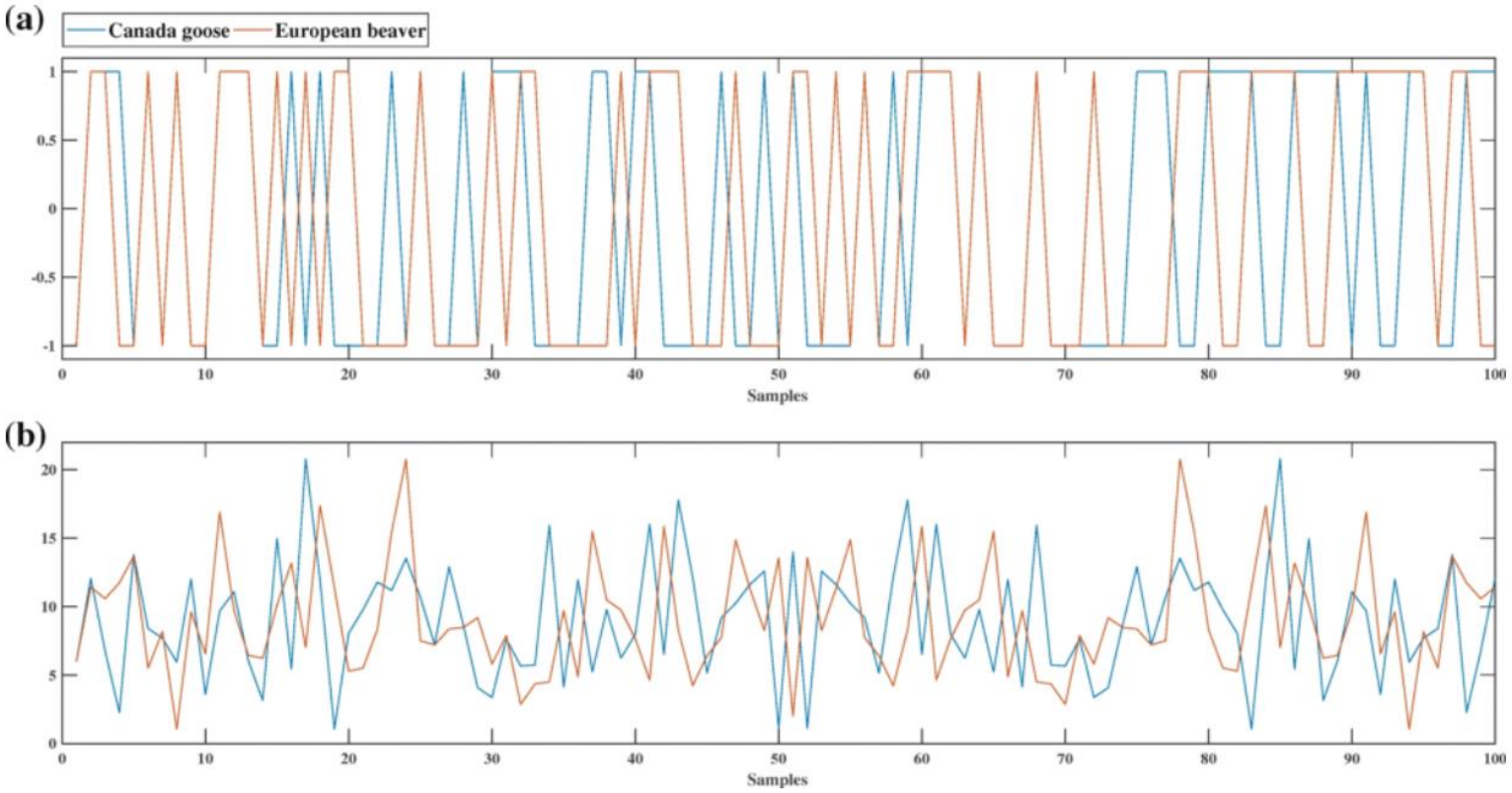


- Train the algorithm on a training set of pairs: (DNA sequence representation, species name)
- Test its power to *predict* the species of an *unknown* DNA sequence representation

Machine Learning with Digital Signal Processing: *ML-DSP*

- Convert each DNA sequence into a numerical vector (discrete digital signal)
- Compute the Discrete Fourier Transform (DFT) of the discrete digital signals
- Compute pairwise distances between magnitude spectra of DFTs (Pearson Correlation Coefficient)
- Use distances between a given item and all other items in the training set as its feature vector to train a supervised machine learning classifier

Discrete Digital Signals



Canada Goose **(blue)** vs. European beaver **(red)** 100bp mtDNA genome:
(a) Discrete digital signals (b) DFT magnitude spectra
(PP representation, *T/C = 1, A/G = -1*)

Lila Kari, University of Waterloo

ML-DSP classification results

- Primates (148), Protists (159), Fungi (226)
- Plants (174), Amphibians (290), Mammals (830)
- Insects (898), three Vertebrates Classes (2170), Vertebrates (4322)
- 6 Classifiers: Linear Discriminant, Linear SVM, Quadratic SVM, Fine KNN, Subspace Discriminant, Subspace KNN

Best digital signal representation of DNA sequences:
PP (Purine/Pyrimidine, *T/C =1, A/G = -1*)

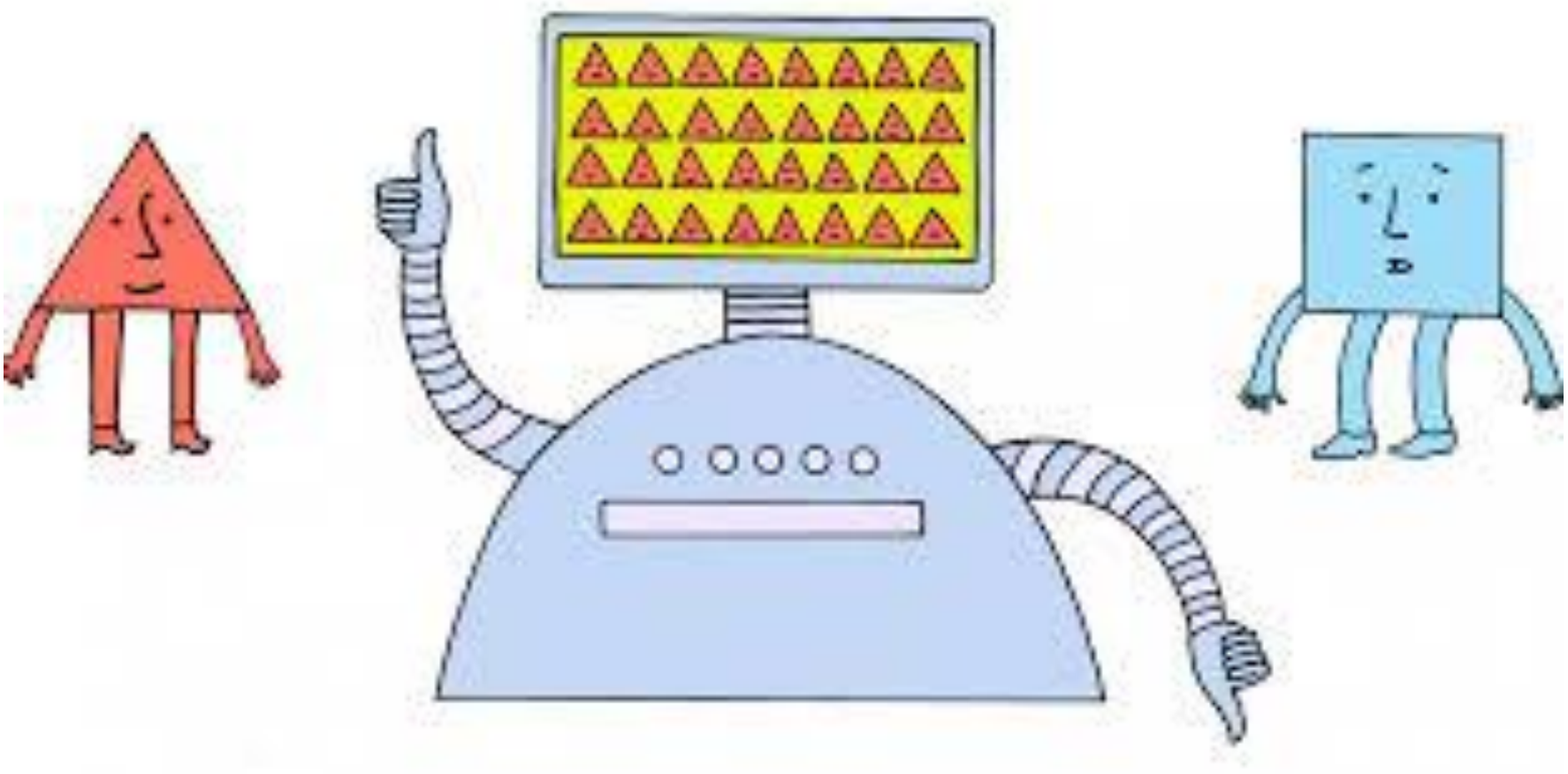
DataSet/	Numerical representation												
classification model	Integer	Integer (Other)	Real	Atomic	EIIP	PP	Paired Num.	NN based doublet	Codon	Just-A	Just-C	Just-G	Just-T
Table average	90.0%	88.7%	91.6%	79.4%	81.3%	92.3%	90.5%	90.7%	89.4%	91.9%	90.5%	91.5%	90.7%

Accuracy and time comparison with other *alignment-free* taxonomic classification algorithms

DataSet	Parameter	MEGA7 (MUSCLE)	MEGA7 (CLUSTALW)	FFP	ML-DSP
Influenza Virus	Maximum Classification Accuracy	97.4%	97.4%	68.4%	100%
(38 sequences)	Average Classification Accuracy	93.4%	95.6%	57.0%	94.7%
Average Length: 1407bp	Processing Time	7.44 sec	2 min 14 sec	0.2 sec	0.2 sec
Mammalia	Maximum Classification Accuracy	95.1%	95.1%	49.6%	92.7%
(41 sequences)	Average Classification Accuracy	89.7%	90.7%	41.5%	87.8%
Average Length: 16647bp	Processing Time	11 min 15sec	5 hr 38 min	0.3 sec	0.3 sec
Vertebrates	Maximum Classification Accuracy	——	——	61.7%	99.7%
(4322 sequences)	Average Classification Accuracy	——	——	48.3%	98.3%
Average Length: 16806bp	Processing Time	>2 h	>6 h	94 sec	28 sec

[Randhawa, Hill, Kari, *Bioinformatics*, 2020]

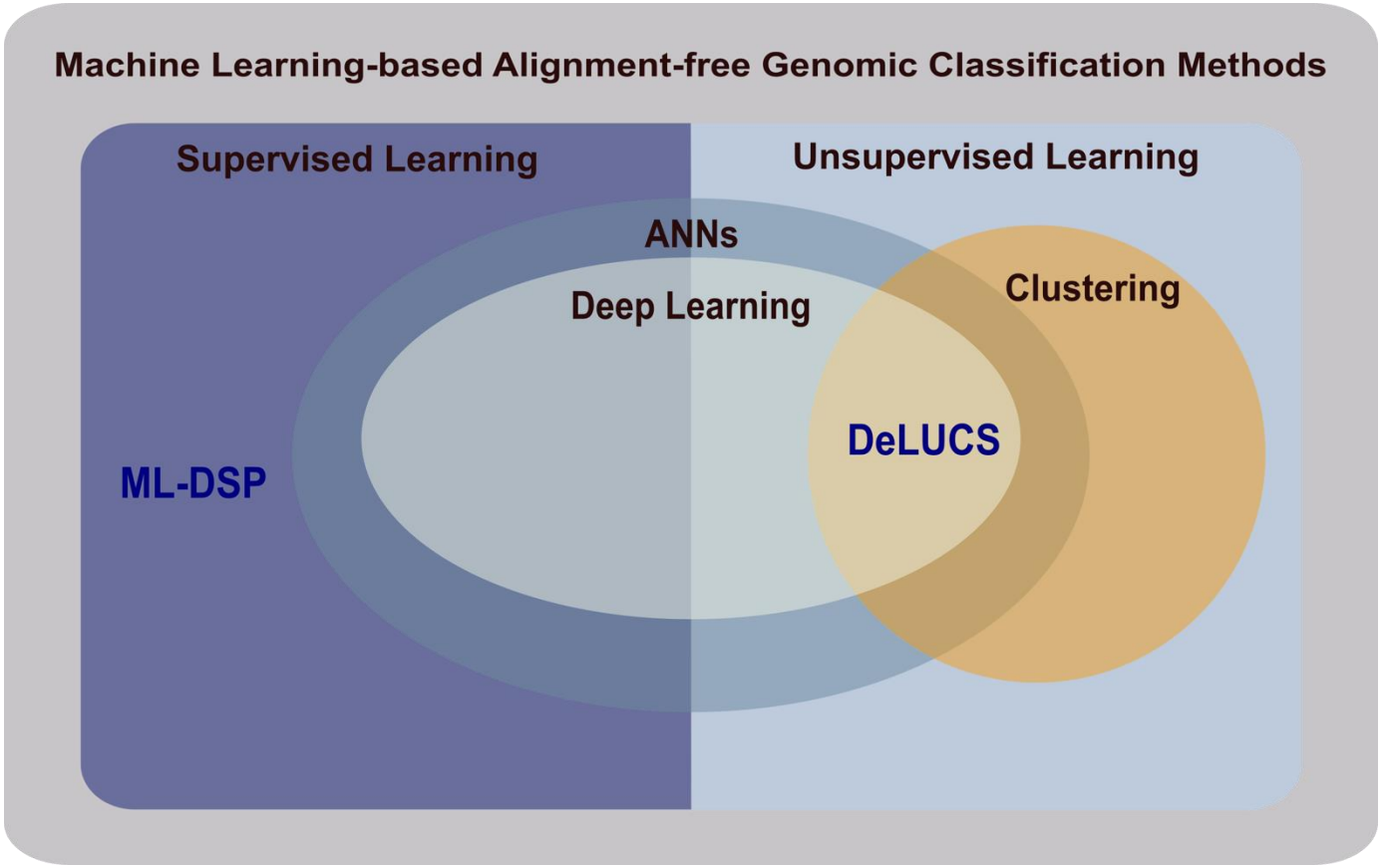
Supervised machine learning limitations



Contents

- Mathematical representations of DNA
- Supervised machine learning for *taxonomic* classification (ML-DSP)
- Unsupervised clustering for *taxonomic* identification (iDeLUCS)
- Test the hypothesis of an *environmental signal* in extremophile genomes

Machine learning for genomics

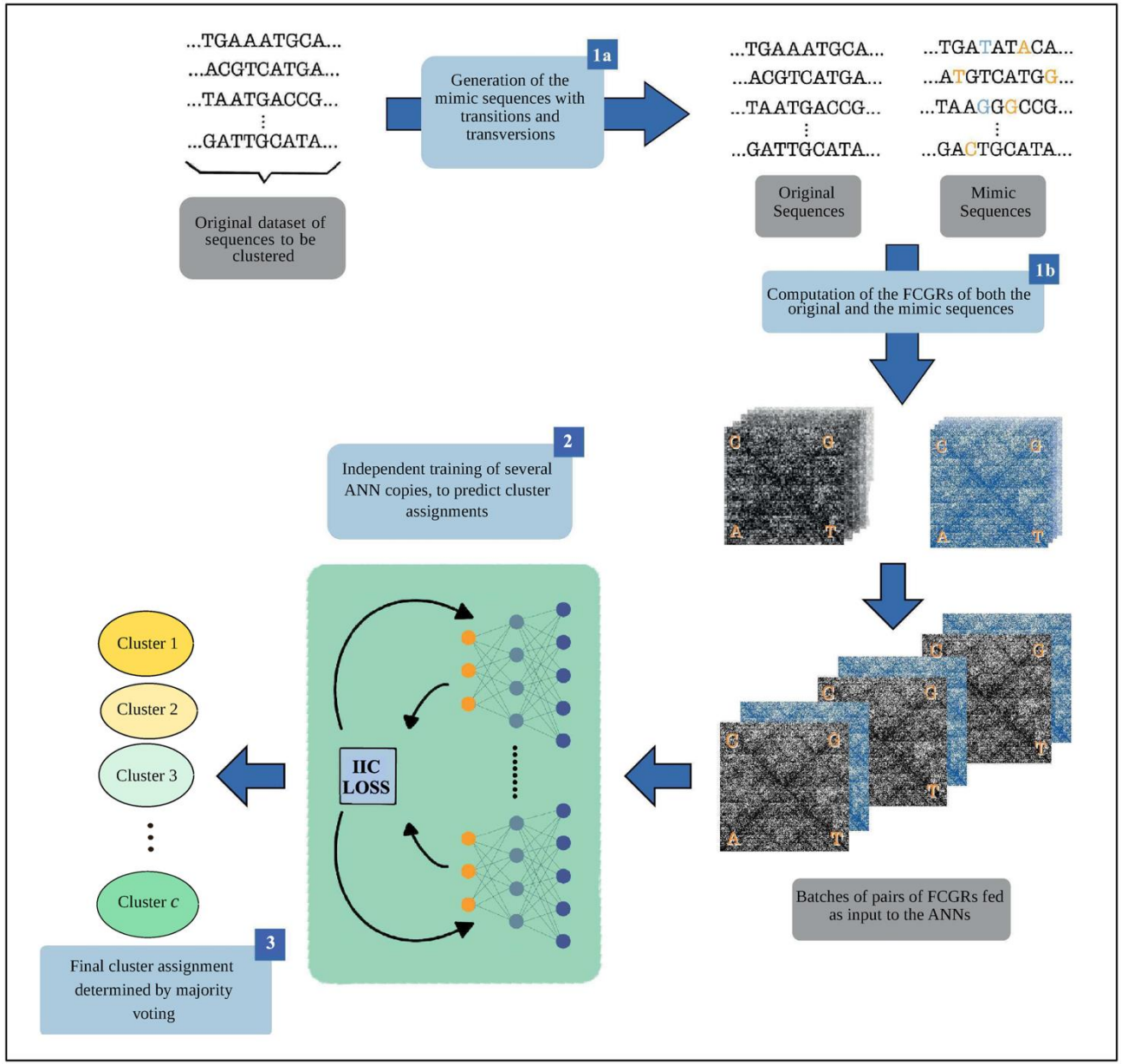


Deep Learning for Unsupervised Clustering of DNA Sequences: *iDeLUCS*

- Mathematical representation of DNA: ***FCGR***
- Generate "mimic" sequences
- Use mimic sequences to self-learn data patterns (unsupervised clustering)
- Use a majority voting scheme to determine the final cluster assignment for each sequence

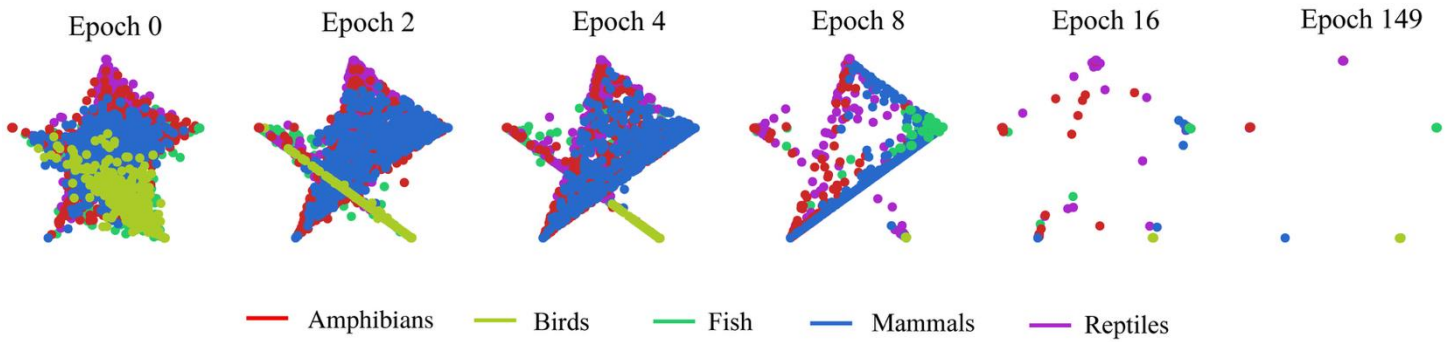
[Millan Arias, Alipour, Hill, Kari. *PLOS ONE,* 2022,
Millan Arias, Hill, Kari, *Bioinformatics*, 2023]

iDeLUCS pipeline



iDeLUCS clustering results

- Learning process for the clustering of 2,500 vertebrate mtDNA genomes into 5 clusters
- Position of a point indicates the probability that it is assigned to different clusters



Test #	Classification	Number of Mimics	Supervised	Unsupervised		
				GMM	K-Means++	DeLUCS
1	Subphylum	3	99%	72%	81%	93%
2	Class to Subclass	8	98%	92%	96%	100%
3	Subclass to Superorder	3	99%	70%	82%	85%
4	Superorder to Order	8	100%	68%	73%	94%
5	Order to Family	8	87%	66%	77%	79%
6	Family to Genus	8	80%	84%	87%	91%

For unsupervised learning, reported accuracy values are the average over 10 runs of the algorithm. For supervised learning, the accuracy is that of classifying the test set.

https://doi.org/10.1371/journal.pone.0261531.t004

Mathematical structures in genomes

- **Question:** Does **biological kinship** induce a detectable mathematical signature in genomes?

  YES (always)

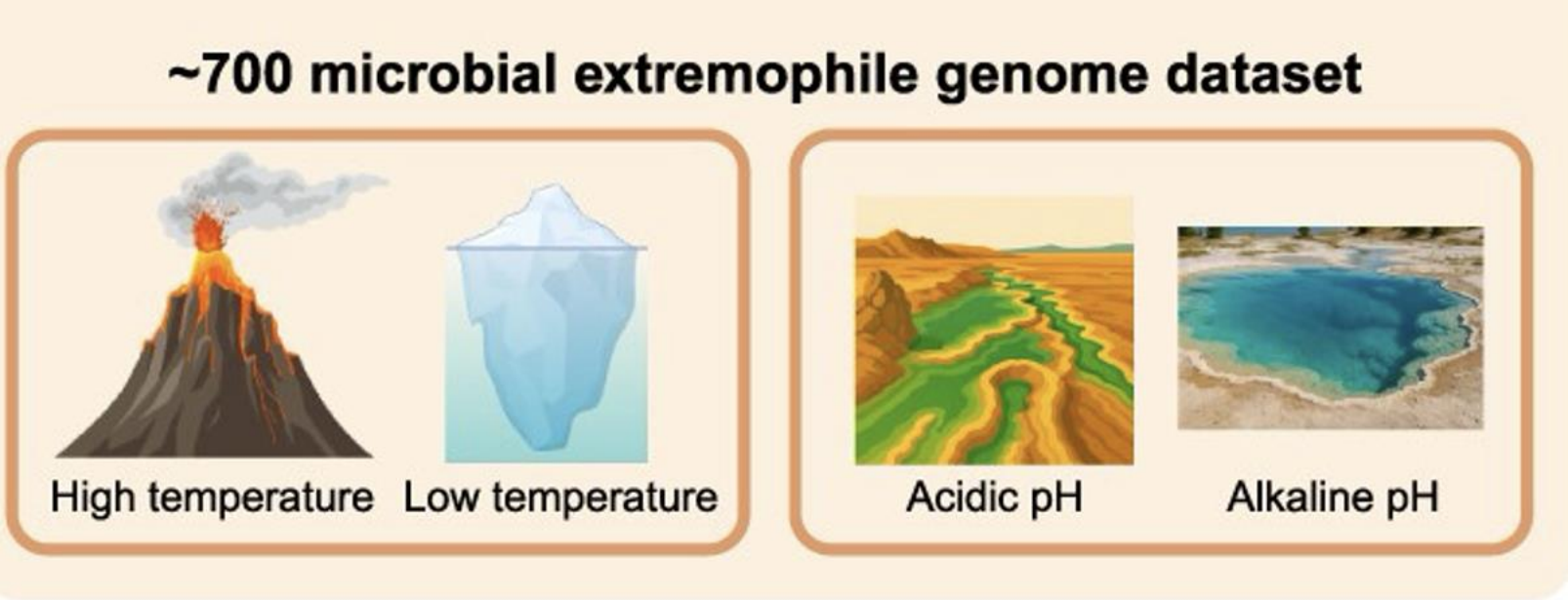
- **Question:** Can **the environment** induce a detectable, *kinship-independent*, mathematical signature in genomes?

Contents

- Mathematical representations of DNA
- Supervised machine learning for *taxonomic* classification (ML-DSP)
- Unsupervised clustering for *taxonomic* identification (iDeLUCS)
- Test the hypothesis of an *environmental signal* detectable in extremophile genomes

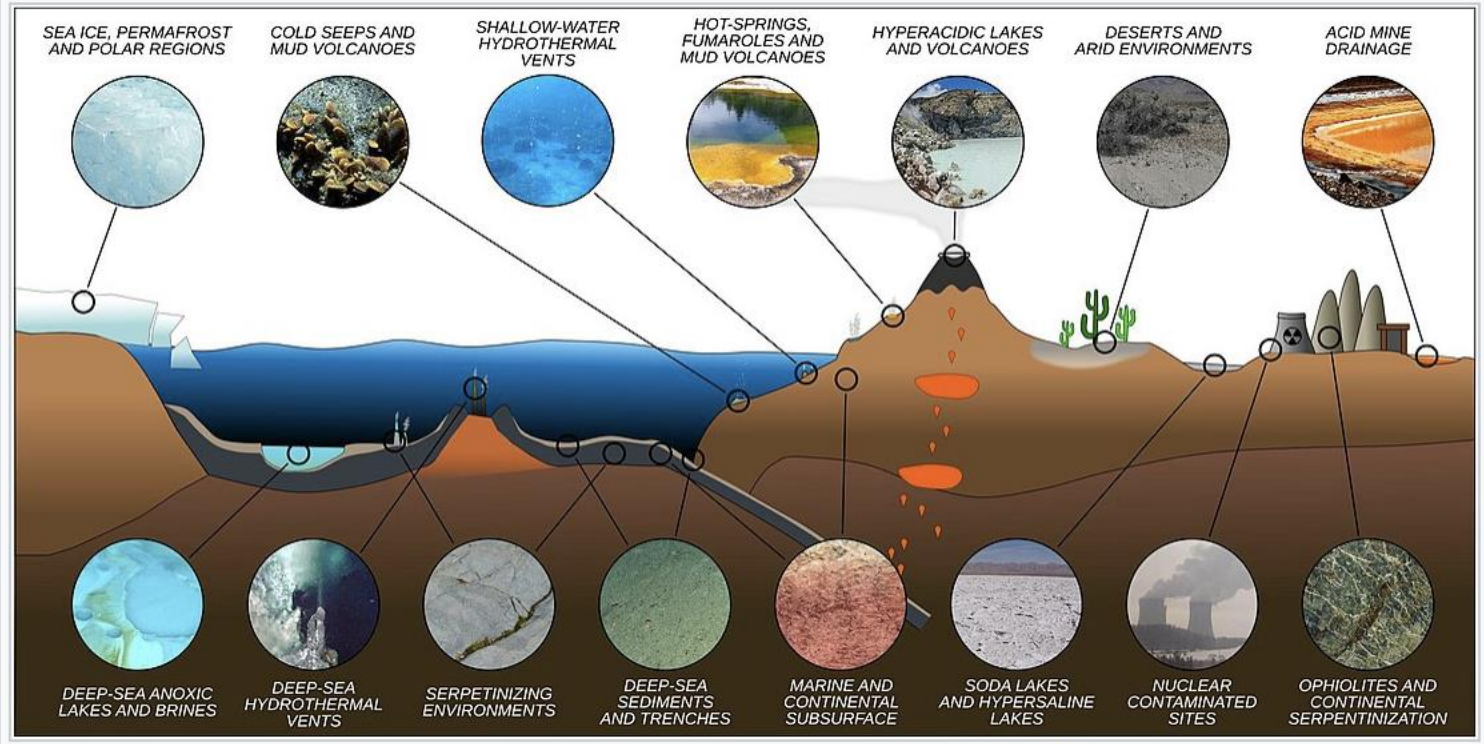
Hypothesis & Dataset

- Traditional view: DNA exclusively contains ancestry (phylogenetic) information
- Hypothesis: If they live in similar extreme environments, evolutionarily distant microbes have similar DNA patterns



Extremophiles

Organisms able to live (or thrive) in extreme environments
e.g., extreme temperature, radiation, salinity, or pH (acidity)



Diversity of extreme environments on Earth[5]

Lila Kari, University of Waterloo

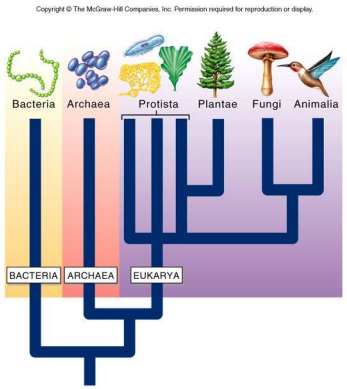
Extremophile microbes: Temperature and pH

- Psychrophiles
  OGT $< 20^{o}C$
- Mesophiles
  OGT $20^{o}C - 45^{o}C$
- Thermophiles
  OGT $45^{o}C - 80^{o}C$
- Hyperthermophiles
  OGT $> 80^{o}C$

OGT: optimal growth temperature
(human body temperature = 36.5C)

- Acidophiles
  OGpH < 5
- Alkaliphiles
  OGpH > 9

OGpH: optimal growth pH
(water has a neutral pH of 7)

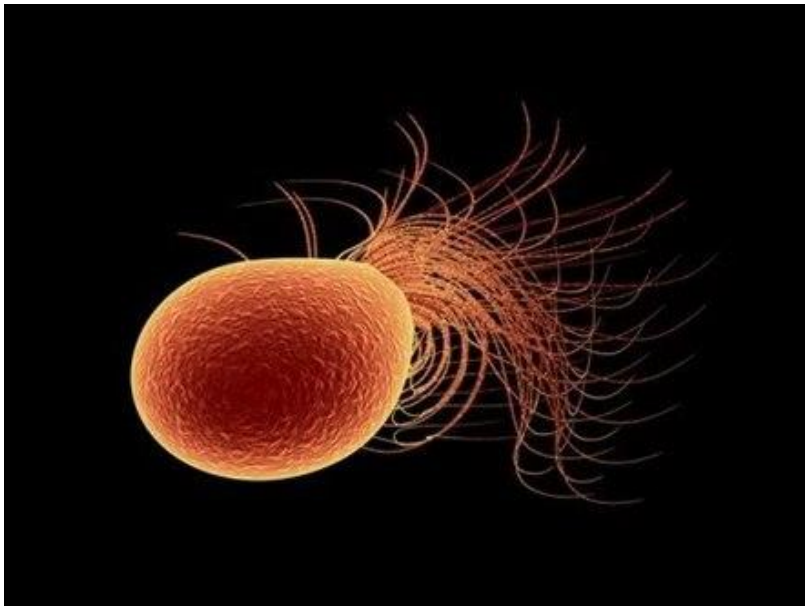


Two extremophile microbial datasets (700 genomes)

Domain	Temperature Category	# Phyla	# Classes	# Orders	# Families	# Genera	# Species
Archaea	Psychrophiles	2	4	4	5	7	8
	Mesophiles	4	6	7	20	45	84
	Thermophiles	6	11	14	21	41	67
	Hyperthermophiles	5	6	8	15	31	70
Bacteria	Psychrophiles	4	4	6	13	19	140
	Mesophiles	3	3	6	10	14	106
	Thermophiles	15	19	24	27	47	116
	Hyperthermophiles	5	5	5	5	5	7

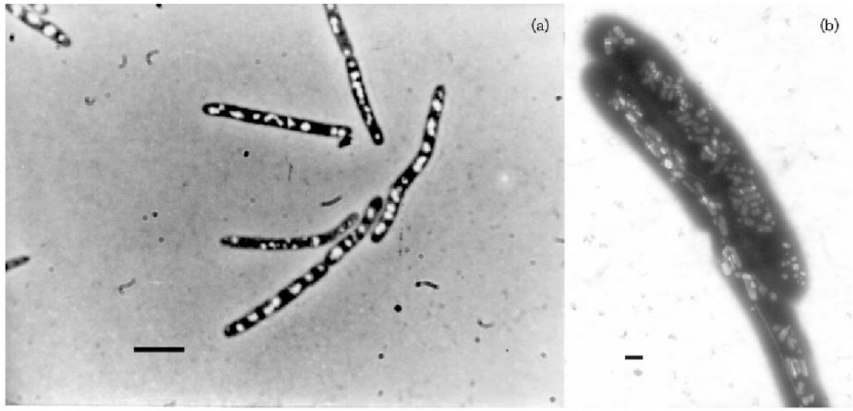
Domain	pH Category	# Phyla	# Classes	# Orders	# Families	# Genera	# Species
Archaea	Acidophiles	4	5	7	11	24	39
	Alkaliphiles	2	5	5	9	18	30
Bacteria	Acidophiles	10	12	13	13	32	61
	Alkaliphiles	12	14	25	30	36	56

Pyrocossus furiosus



- *Archaeon* isolated from heated sediments on Vulcano Island, Italy
- *Hyperthermophile* Optimal growth temperature of 100°C
- *Pyrococcus* (Greek) - fireball
- *Furiosus* (Latin) - refers to its rapid swimming

Psychromonas ingrahamii

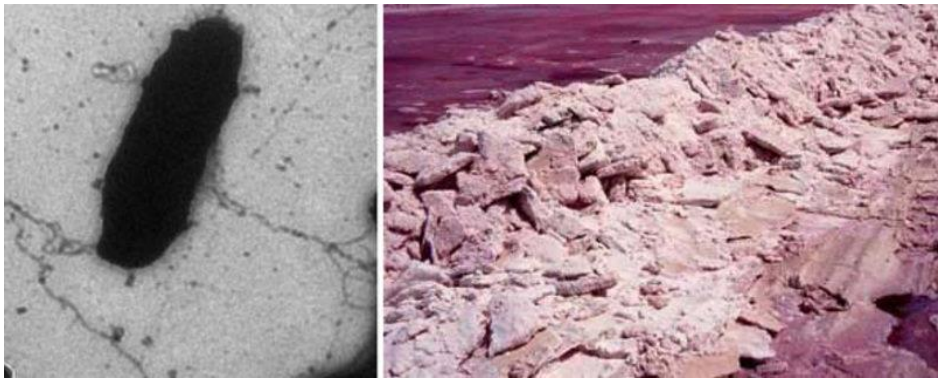


- *Bacterium*, isolated from Arctic polar sea ice
- *Psychrophile* – optimal growth temperature -12°C lowest recorded growth temperature
- *Psychros* (Greek) - cold, frozen

Picrophilus torridus

- *Archaeon*, isolated in a hot spring, Hokkaido, Japan
- *Acidophile,* pH < 0.5
- *Picro* (Greek) - bitter
- *Torridus* (Latin) – hot (optimal growth 60°C)
- *Polyextremophile*

Natromonas pharaonis



- *Archaeon*, found in salt lake Zuf , Wadi Natrun, Egypt
- *Alkaliphile (halophile),* thrives in extremely high salt concentration, pH 11 (300 g of salt per liter)
- Similar to lye soap, or the salt content of Dead Sea

Representative DNA fragment

A 100,000 bp genome fragment was selected from each genome, as the ``representative DNA fragment'' to act as *genome proxy*

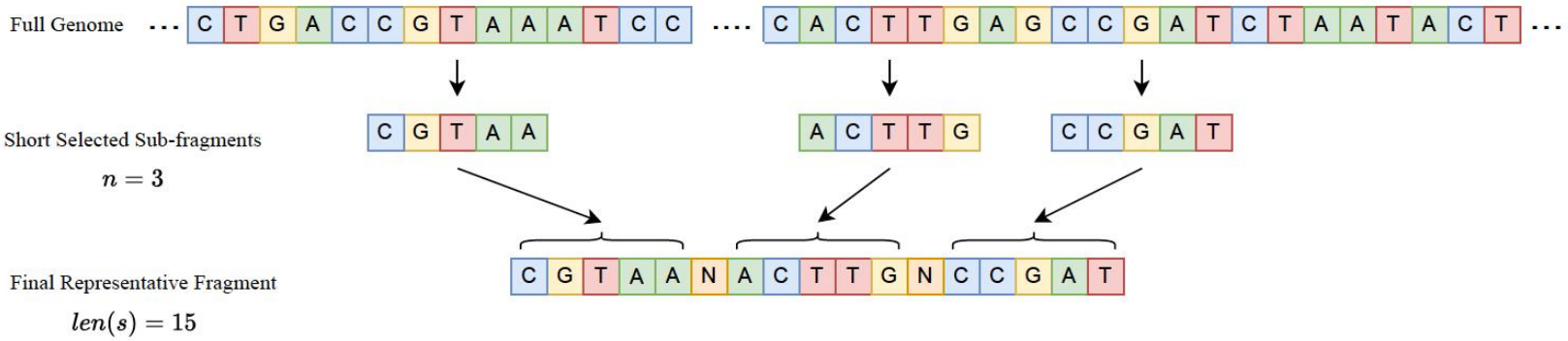


Fig. 2 The selection process of a representative DNA fragment s, comprising $n = 3$ non-overlapping sub-fragments, and with total length $len(s) = 15$. Top: Full genome, consisting of only one contig. Middle: n non-overlapping sub-fragments (here $n = 3$) randomly selected from the genome. Bottom: The representative DNA fragment of length $len(s) = 15$ obtained by pseudo-concatenating the sub-fragments.

* 100,000bp = ~ 3% of the average genome length (3 million bp)

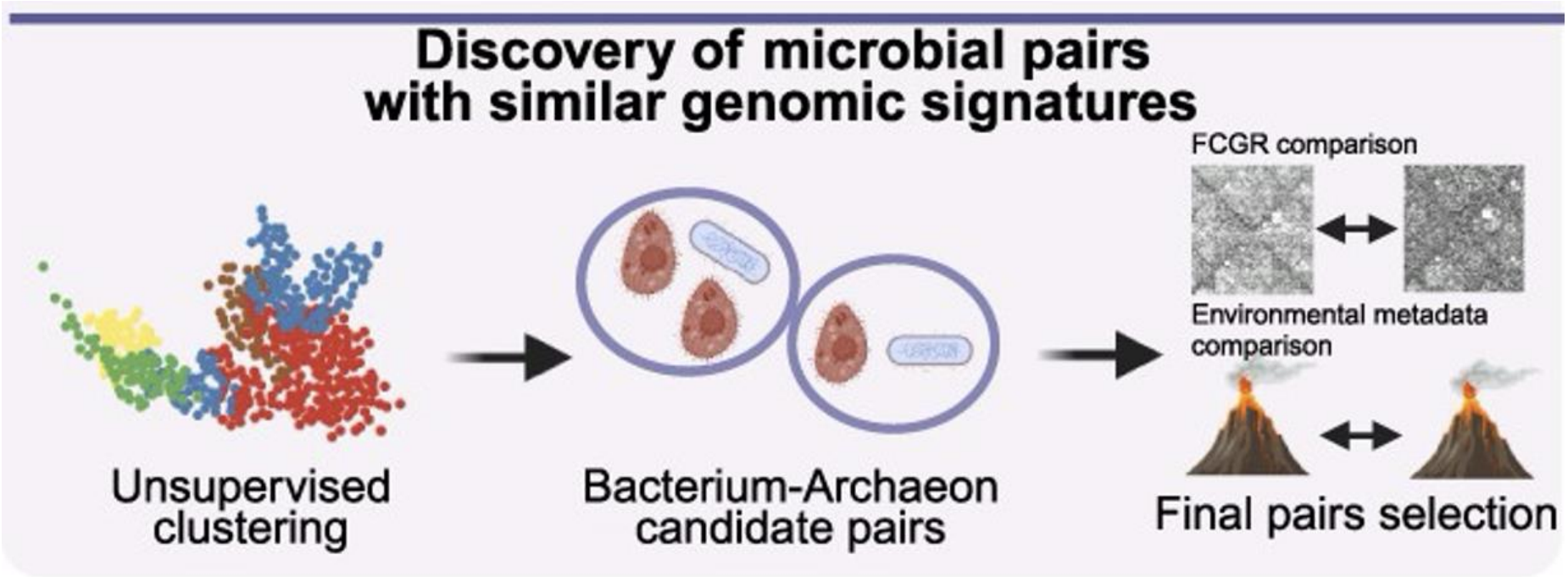
Supervised learning *classification* of the Temperature dataset

Table 3. Classification accuracies of six supervised learning classifiers trained on the Temperature Dataset and pH Dataset, in the *restriction-free* scenario, for three different label assignments (taxonomy, environment category, and random label assignment), and values of $1 \leq k \leq 6$. The classification accuracy in each cell is calculated using standard stratified 10-fold cross-validation.

Dataset	k-value	Class Labelling Type	Classification Model Accuracy (%)					
			RBF SVM	Random Forest	ANN	MLDSP-1	MLDSP-2	MLDSP-3
Temperature	$k = 1$	Taxonomy	62.88	53.87	62.21	47.99	54.85	59.03
		by Environment	39.97	35.29	38.65	26.92	32.27	31.44
		Random	22.26	29.42	31.77	27.59	26.92	27.59
	$k = 2$	Taxonomy	96.65	95.14	96.14	86.79	92.64	86.79
		by Environment	74.58	76.91	74.42	46.49	68.06	46.32
		Random	23.25	28.10	27.09	26.42	25.08	25.75
	$k = 3$	Taxonomy	98.82	97.99	97.32	92.64	96.82	92.64
		Environment	82.11	81.59	75.41	71.91	74.58	71.24
		Random	23.58	25.08	27.76	25.59	26.09	24.58
	$k = 4$	Taxonomy	99.50	98.33	98.66	98.16	97.16	98.16
		Environment	83.29	84.11	82.28	78.43	75.08	80.43
		Random	25.06	23.74	27.59	25.42	26.92	23.58
	$k = 5$	Taxonomy	99.50	98.16	99.33	97.32	97.32	98.16
		Environment	83.27	84.76	83.29	69.23	77.26	81.77
		Random	24.08	20.23	23.07	26.09	25.42	24.25
	$k = 6$	Taxonomy	99.50	98.50	99.33	99.16	97.49	98.83
		Environment	83.46	83.94	84.12	79.60	77.59	82.44
		Random	27.24	22.91	26.58	28.09	25.59	24.25

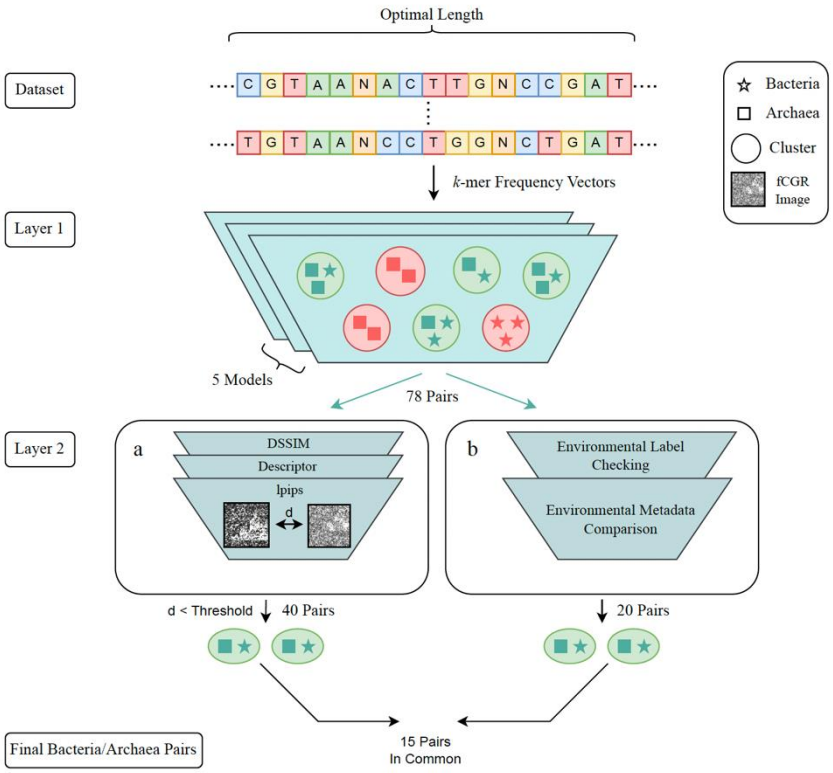
Lila Kari, University of Waterloo

Unsupervised machine learning to find organisms that are genomically similar



Lila Kari, University of Waterloo

Finding "peculiar" bacteria/archaea (unsupervised) clusters



Input: Dataset of bacteria and archaea DNA representative fragments

Layer 1): Unsupervised clustering produces 78 candidate bacteria/archaea pairs with similar FCGRs

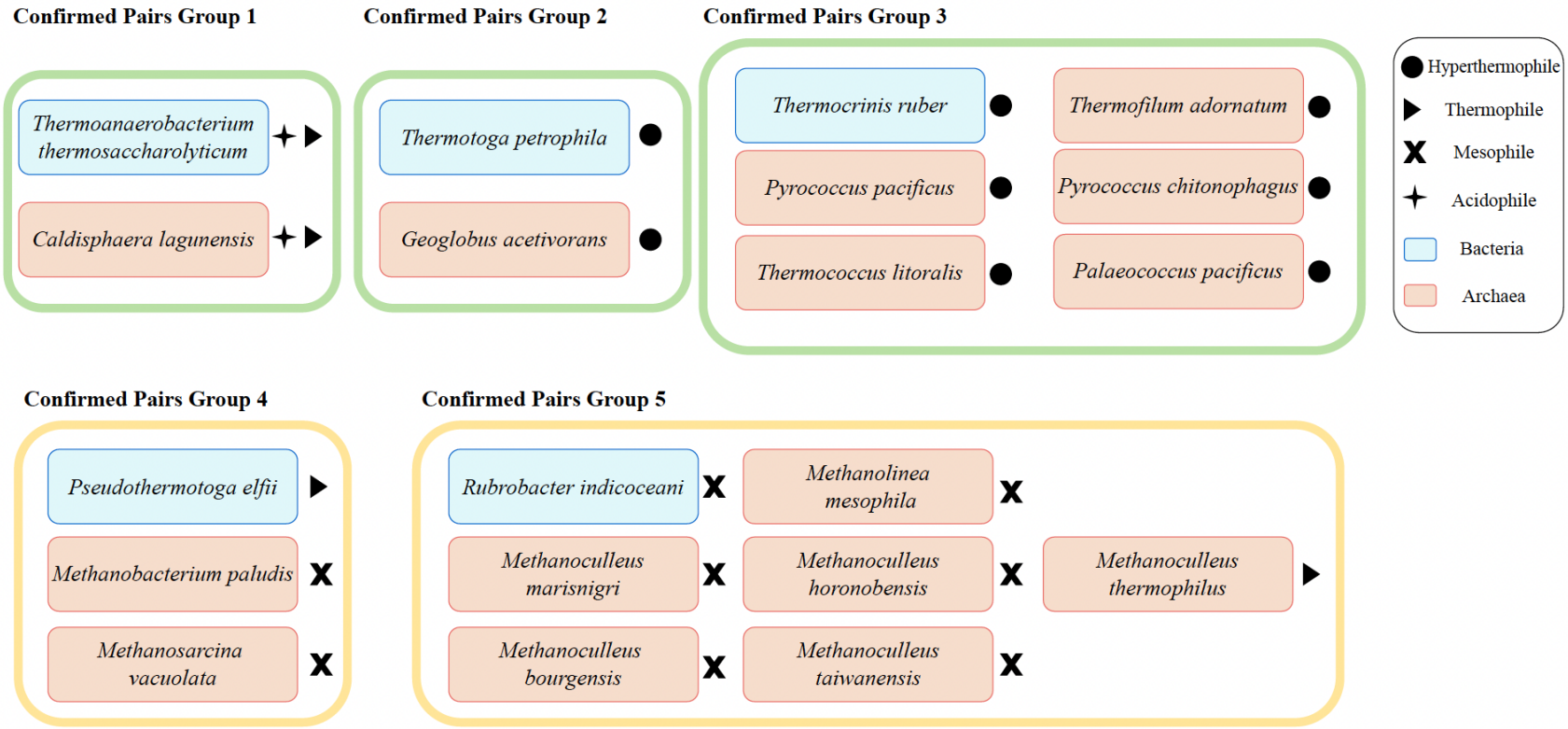
Layer 2a): FCGR distance calculations of members of candidate pairs (produces 40 pairs with distances less than the genus threshold)

Layer 2b): Biological analysis of candidate pairs (produces 20 pairs)

Output: Intersection of 2a) and 2b) = 15 confirmed bacteria/archaea "peculiar" pairs

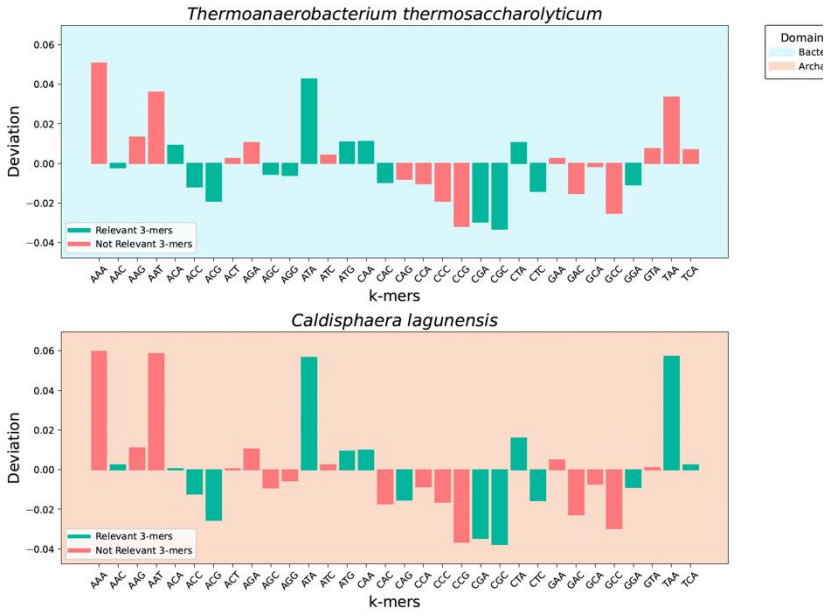
Lila Kari, University of Waterloo

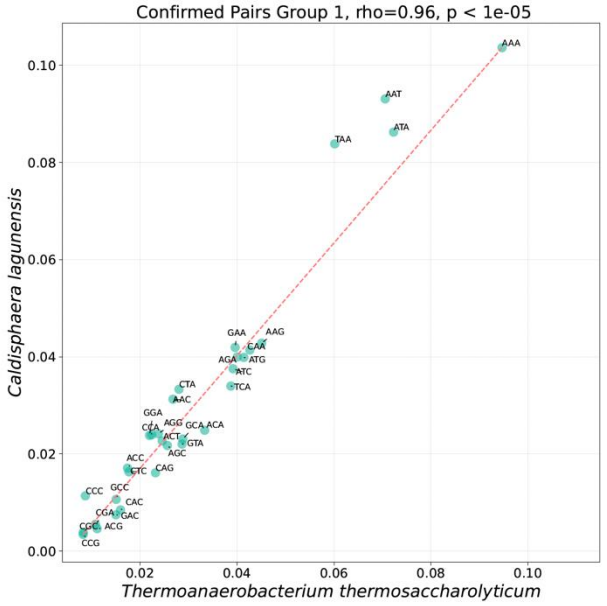
4 confirmed “peculiar” clusters



Lila Kari, University of Waterloo

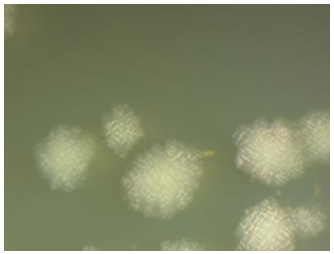
Additional statistical analysis



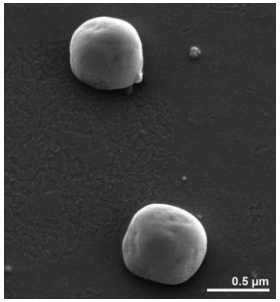


Lila Kari, University of Waterloo

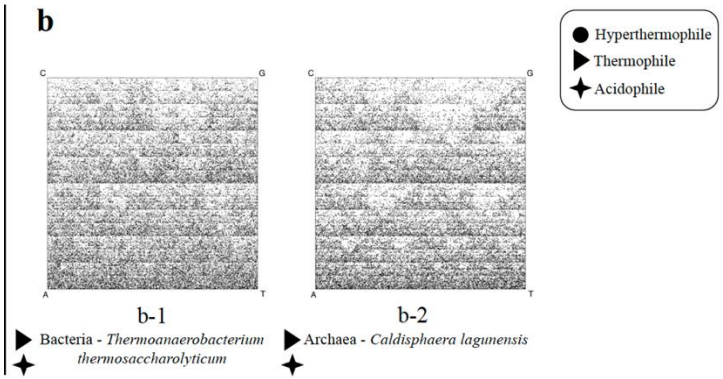
Co-occurring bacterium/archaeon pair



Thermoanaerobacterium thermosaccharolyticum
thermophile acidophile bacterium

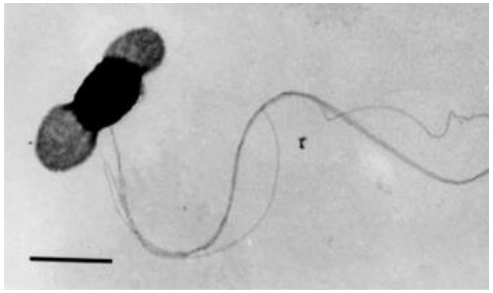


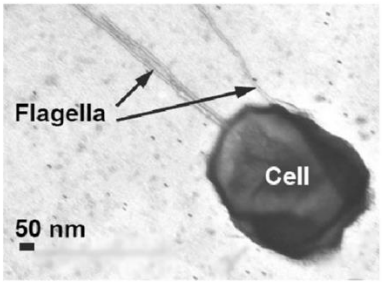
Caldisphaera lagunensis
thermophile acidophile archaeon



Co-occurring in Washburn Hot Springs, Yellowstone Provincial Park, Wyoming

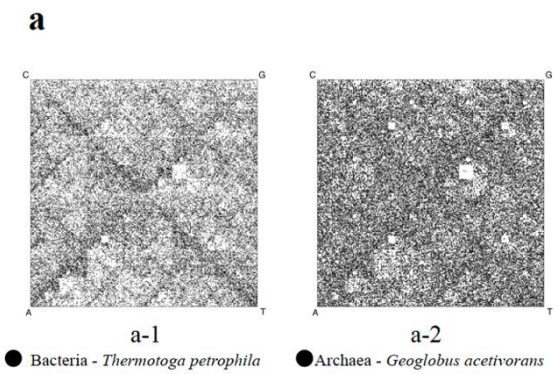
Co-occurring bacterium/archaeon pair

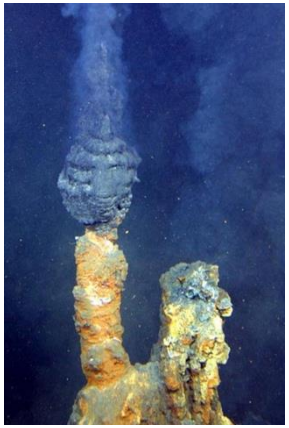




Thermotoga petrophila
hyperthermophilic bacterium

Geoglobus acetivorans
hyperthermophilic archaeon.



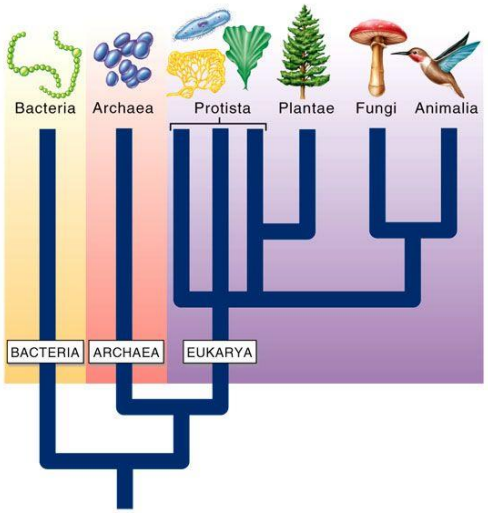


An octopus defending its garden in the Endeavour hydrothermal vent area of the Juan de Fuca Ridge (depth 2200m) PNG

Co-occurring in Brothers submarine volcano, New Zealand (left)
Juan de Fuca mid-ocean ridge flank near Vancouver Island (right)

Profoundly distant species with striking degree of DNA similarity

- Mammals split from fish *400 million* years ago
- Bacteria and archaea split from a common ancestor *4 billion years ago* (10x further back)



- There was no expectation of genetic similarity between bacteria and archaea, so any DNA similarity is **startling!**

Profoundly distant species with striking degree of DNA similarity

- Additionally, for the clustered-together bacterium archaeon pairs, the DNA pattern similarity is very strong
- Pattern (*k*-mer frequency profile) similarity is at the *genus level*, i.e., the same level of similarity as a brown bear *vs.* a polar bear (genus *Ursidae*)

Mathematical structures in genomes

- **Question:** Does **biological kinship** induce a detectable mathematical signature in genomes?

  YES (always)

- **Question:** Can **the environment** induce a detectable, *kinship-independent*, mathematical signature in genomes?

  YES (sometimes)

Conclusions

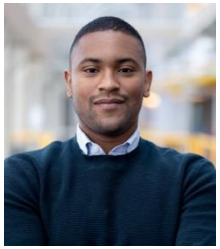
- DNA was traditionally thought to be a "family portrait" determined solely by ancestry/phylogeny
- Our findings show that, surprisingly, extremophile DNA also carries environment-type patterns, that sometimes are so strong that they override the "family resemblance"
- The environment-type DNA patterns are not only present in genes; like a watermark, these patterns are detectable across the entire genome

[Safari, Butler, Randhawa, Hill, Kari. *NAR Genomics and Bioinformatics*, 2025]

Research collaborators

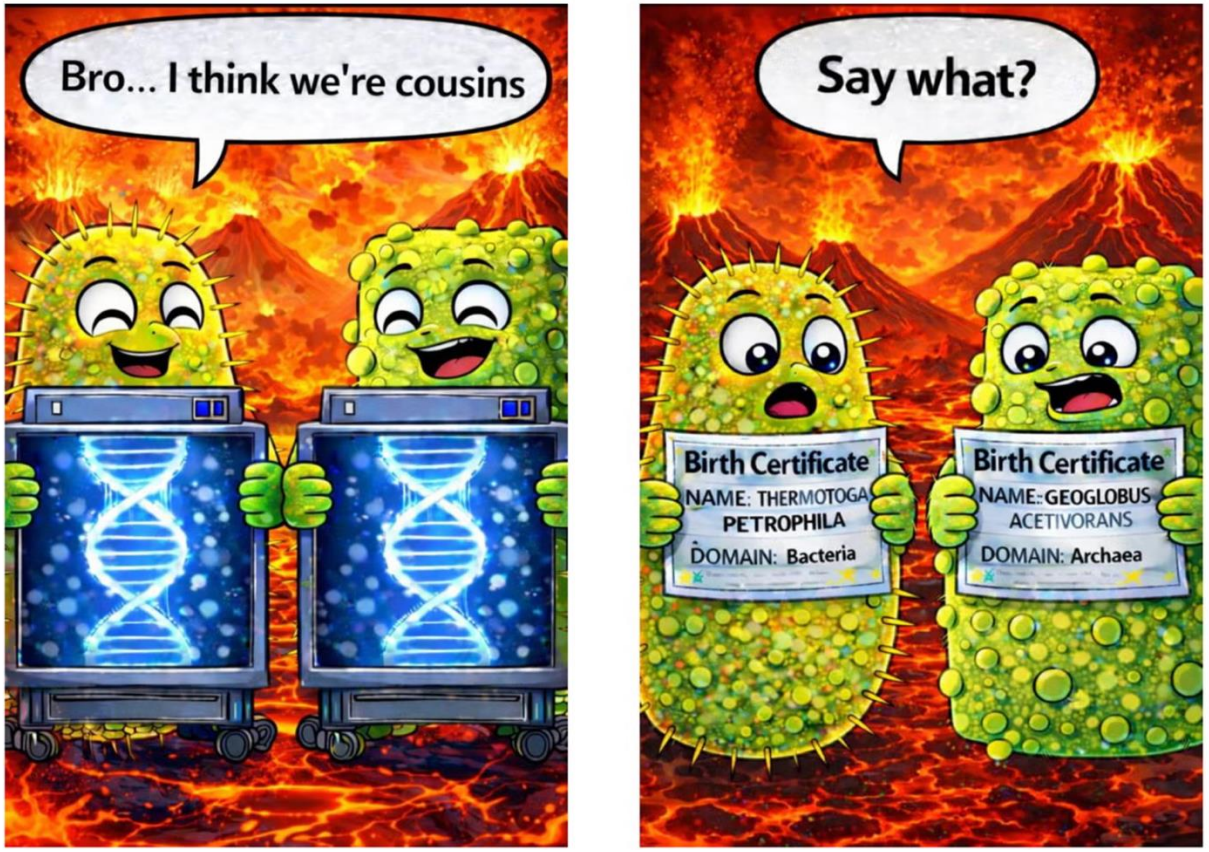


- Monireh Safari (U Waterloo)
- Niousha Sadjadi (U Waterloo)
- Prof. Kathleen Hill (U Western Ontario, Biology)
- Dr. Pablo Millan Arias (U Waterloo)
- Prof. Camila de Souza (U Western Ontario, Statistics and Actuarial Sciences)
- Haoze He (EPFL Lausanne)
- Rallis Karamichalis (U Western Ontario)
- Joseph Butler (U Western Ontario, Biology)
- Prof. Gurjit Randhawa (U Guelph)






Thank you!



Lila Kari, University of Waterloo