2021

BIOMOLECULE INSPIRED DATA SCIENCE

Sambriddhi Mainali

BIOMOLECULE INSPIRED DATA SCIENCE

by

Sambriddhi Mainali

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Computer Science

The University of Memphis

August 2021

ACKNOWLEDGMENTS

Finding my own footing in the area of research has been a huge challenge for me in this dissertation. I am grateful to my advisor, Dr. Max H Garzon, for the kind of training I have received. I got an opportunity to learn a lot from him, in particular, how to behave professionally, conduct research, overcome challenges and judge myself so as not to repeat mistakes in the future. Looking back to the time when I started this research, I have found myself matured a lot, not only as a researcher but also as a person. I consider myself fortunate to have gotten such an opportunity to learn by examples from him.

I would also like to thank Drs Omar Skalli (Biology), Deepak Venugopal (Computer Science) and Russell Deaton (Electrical Computer Engineering) for serving on my committee. Their comments were very valuable and substantially helped to improve the presentation of the results in this dissertation. I would also like to acknowledge contributions from Dr. Fredy Alexander Colorado of the National University of Colombia. Without his guidance and input on sample selection and assessment of the results, I would not have been able to accomplish the goals in this research.

Finally, I express my gratitude towards my parents, my in-laws and my husband, Manu Bhandari, for their unfailing support and encouragement. Without them, this dissertation would not have been completed successfully. Thank you.

ABSTRACT

Mainali, Sambriddhi. Ph.D. The University of Memphis. August, 2021. Biomolecule inspired Data Science. Major Professor: Dr. Max Garzon

Our ability to generate data has far outdone our ability to analyze it in order to transform it into useful information. A major tool in addressing the problem is extraction or selection of informative features in the data. When the data is structured, dimensionality reduction and analyses can be much easier. However, structured data (beyond just superficial formatting and cleansing) is rare, hopeless in case of images and even text, particularly with DNA sequences. Deep networks resolve these issues to some extent but the question about explainability and timeliness of results still remains. Recent advancements in Genomic Information Systems (GenISs) have shown that the exquisite discriminating ability of DNA hybridization (double helix formation) can be leveraged for smarter data processing. Deep knowledge about the hybridization property of DNA (discovered by Watson and Crick in the 1950s) has enabled us to uncover some Euclidean embeddings of DNA oligonucleotides along with some interesting structural properties (like centroids, center of mass and so on) analogous to that of the planets in our solar system (like earth and saturn.) In this work, we develop a family of Genomic Information Systems (GenISs), based on novel coordinate systems (genomic and pmeric) for DNA sequences of arbitrary length obtained from their deep structural properties based on hybridization patterns, that can be leveraged to improve and develop new methods for data analytics of both biotic and abiotic data. We also assess the quality of these GenISs with a number of applications in the field of biology and computer science at large. The quality assessment of these results illustrate how DNA is capable of self-organizing unstructured data into semantic clusters meaningful to humans, in addition to supporting complex life processes for phenomics, metabolomics, species identification, pathogenicity and so on. Furthermore, these results hint at the tip of an iceberg about the capacity of DNA for not only encoding but also processing information that can be leveraged as a powerful tool in this era of big data and data science.

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1

Biomolecular Programming

Background

Our ability to generate enormous amounts of data through, for example, genomics (e.g., the human genome project, Next-Generation Sequencing (NGS)), proteomics and metabolomics, not to mention abiotic data, has raised an enormous challenge in processing this data to extract useful information. The purpose of this work is to further the development of Genomic Information Systems initiated in [5], [3] to create new tools to begin to tackle this challenge. This work was motivated by the field of DNA computing, inspired by the ideas of using DNA itself as a computational medium pioneered by Adleman [6] and as smart glue for self-assembly applications by Seeman [7] and Winfree [8].

Adleman [6] started the field of DNA computing by proposing to build computers using real DNA molecules. Eventually, it was realized that fundamental problems (such as Codeword Design (CWD) below) would need to be solved to get DNA molecules to do something they did not evolve for better than electronic computers. Finding a solution to this problem is **NP**-complete using any single reasonable metric that approximates the Gibbs energy, thus practically excluding the possibility of finding any procedure to find maximal sets exactly and efficiently [9]. The field then refocused on a potentially more impactful application, namely the self-assembly of complex nanostructures [7]. Rather than pursuing this line of research, in this work we are interested in using DNA and processes *in vivo* as an inspiration to develop new tools to solve problems in computer science and biology *in silico*.

It is well known that, as the blueprint of life, DNA encodes for critical information required to develop and sustain life in every living organism (e.g., protein synthesis and self-organization) due to its hybridization and self-organizing properties [10]. In particular,

[4], [11], [12] have demonstrated that DNA encodes enough information about an organism so that features about phenotype, environmental conditions of the natural habitat, taxonomic group and so on could be predicted.

CODEWORD DESIGN PROBLEM

INSTANCE: A positive integer n and a threshold τ

QUESTION: What is a largest set B of single DNA strands (of length n) that do not crosshybridize to themselves or to their complements (nxh set) under stringency τ, i.e., $|xy| > \tau$ for all $x \in B$?

A systematic attempt to develop such tools (described more in detailed below as Genomic Information Systems (GenISs)) was initiated in [9]; [3] to tackle a fundamental problem in our time, and particularly in bioinformatics

Research Objectives

This line of research can be simply described as biomolecule-inspired data science. The overarching goal is to probe into the questions, *what kind of information can be stored in DNA molecules? How much information can be stored in a single DNA molecule? How could it be extracted?*

My dissertation research builds up on the platform of GenISs along three distinct components. The first component is focused in extending the foundations of a truly universal GenIS for genomic analyses in biodata science.

The second component is focused on applying the platform developed through the fundamental research for biotic data analytics to answer some important questions in the field of biology. These questions include (but are not limited to) to:

- Does DNA encode enough information to enable *quantitative predictions* of phenotypic features in a biological organism, even though they depend on environmental factors presumably beyond DNA?

- Or, does DNA actually encode enough information to say something informative about environmental factors (e.g., latitude, longitude, temperature and so on) of the natural habitat where these organisms grew and lived?

- Is a principled and taxon-independent definition of the concept of *species* possible that is universal for biological taxonomies?

- Can we provide an objective taxon independent and systematic definition of pathogenicity shared between hosts (e.g., *homo sapiens*) and micro-organisms (e.g., bacteria and fungi) based on a computational approach?

Finally, the third component is likewise focused on further extending these capabilities to encoding and processing ordinary *abiotic data* in GenISs. We show that this can indeed be done by providing solutions to some challenging problems, such as Malware Classification (MC) and Semantic Image Segmentation (SIS) involving textual and image data.

In order to make inroads into these questions, we must take a very deep look into DNA sequences from the point of view of computer science.

Structural Properties of DNA Spaces

This chapter describes the foundational findings in the study of structural properties of DNA spaces. To begin with, we give formal definitions of concepts and terms elucidating these structural properties of DNA.

Definition 1 (DNA spaces)**.** *Given a positive integer n, a* DNA sequence *x is a string defined over the alphabet $\Sigma = \{a, c, g, t\}$. They will also be referred as n-*mers*, where $n = |x|$ is the length of string x. The* Watson-Crick (WC) complement *of x is the string y' obtained after first taking the reverse of x (i.e., x^r) and replacing every a (c) by t (g, respectively) and vice-versa, in x^r. A* pmer *(or $|x|$-pmer) is a pair of two WC complementary DNA*

sequences $\{x, WC(x)\}$ *(simply denoted* x/x'*, or just* x*, the lexicographically first element in the pair.) The* DNA space *of length* $n > 1$*,* $\mathbf{D}_n$*, consists of the set of all* n*-pmers.*

We remark that if x is a WC-palindrome and $WC(x) = x$, then the corresponding pmer is really a single string. For reasons that will become apparent to the goal of this research, we will thus exclude palindromes from consideration throughout.

The most fundamental property of DNA is hybridization, its *exquisite discriminating ability in forming double strands* (helices) as discovered by Watson and Crick [10].) Hybridization is determined by the familiar Gibbs energy, the chemical equivalent of the potential energy in physics, which depends on physical parameters (such as the internal energy, pressure, volume, temperature, and entropy) of the environment in which the duplex is formed. The more negative the Gibbs energy, the more stable the duplex formed. Unfortunately, the available models of biochemistry are approximations, and no gold standard exists to assess Gibbs energies other than accepted empirical approximations [13]. The most popular method to approximate the Gibbs energy is the so-called nearest-neighbor (NN) model, but this model does not offer a *metric* approximation. Further, the size and composition of nxh sets is very difficult to establish in this model due to the lack of intuition and tools as to the structure of the Gibbs energy landscapes [9].

In the field of DNA computing, many attempts have been made to address this issue. They have revolved about the CWD problem identified as a fundamental problem in the field. Adleman [6] emphasized the need of a good coding strategy for using DNA to process information. A coding scheme is crucial to experiments *in vitro*, for mutational analysis and for sequencing [14]. For a biologist, the obvious criteria for good choices were things like the GC content of the sequences since it is good indicator of the melting temperature of short oligonucleotides. [14] introduced a template method to generate a set of sequences of length l such that any of its member have approximately $l/3$ mismatches (based on the GC content) with other sequences, their complements and the overlaps of their concatenations. Some approaches also tested combinatorial design, random

generation and genetic algorithms [15, 16, 17]. A more refined method used to address the issue of undesirable changes being made to bits sent through a noisy communication channel. Thus CWD problem makes an analogy with error correcting codes in information theory. DNA sequences diffuse in solution (the channel) "looking" for a (WC) complementary sequence to hybridize to. If the probes attached to the chip are not carefully selected, we get the equivalent of a channel introducing errors in the intended hybridizations when a target encounters the wrong probe first (i.e., the hybridization affinity to the probe is not correct.) Thus, the CWD problem appears to be the equivalent of designing error-detecting/correcting codes. Shannon's solution for error-detecting codes was to use the Hamming distance in hypercubes to quantify the error detecting and correcting capabilities by separating codewords actually used to encode single bits so the noisy transmissions remain noncoding and can be detected and possibly corrected. The obvious choice for CWD is thus the Hamming distance [18]. Since the Hamming distance between any two aligned sequences counts the number of positions in which the two differ in a perfect alignment [19, 20], the ordinary Hamming distance must be modified so that *matching* now refers to Watson-Crick complementary pairs, i.e., a's and t's (c's and g's) occurring in aligned sequences should be considered as matches. [21] used this notion of Hamming distance to obtain sets of "orthogonal" sequences solving the CWD problem experimentally and theoretically for molecular recognition using microarrays. Although this a step in the right direction, Hamming distance between two DNA strands appears to be too crude as an estimate of the likelihood of hybridization because it seems to exclude the possibility of two strands hybridizing in shifted alignments [22], something much more likely to occur. To address this issue, an alternative was introduced in [5]. This h-distance turned out to be a reasonable choice for an approximation of the Gibbs Energy because it satisfies metric properties that the Gibbs energy does not, and more importantly, because hybridization decisions made using the h-distance agree with those

made using Gibb's Energy Nearest-Neighbor Model about 80% of the time [9], [23]. This distance is defined precisely as follows.

Definition 2 (Hybridization distance or h-distance [5])**.** *Given an integer $n > 0$, the* h-measure *between any two pmers x and y, $h(x, y)$ is defined as the minimum total number of WC complementary mismatches between facing nucleotide pairs in an optimal alignment. Precisely, it is computed as follows:*

- *align x and y^r (y reversed) in $2n - 1$ alignments shifted by k characters (left shift if $k < 0$; right if $k > 0$), $-n < k < n$;*
- *count the total number c_k of WC complementary mismatches between facing nucleotide pairs (single nucleotides are counted as mismatches);*
- *compute the h-measure*

$$h(x, y) = \min_k \ c_k \ .$$

The h-distance *(denoted just $|xy|$ hereforth) between two pmers x and y is defined as the minimum of the two h-measures $h(x, y)$ and $h(x, y')$ between x and y, where y' is the WC-complement of y.*

An example of the computation of the h-distance is shown in Fig. 1.

Theorem 1 ([5, 24])**.** *The h-distance is a metric, i.e., every triple $x, y, z \in \mathbf{D}_n$ satisfies*

- *(Reflexive) $|xy| = 0$ if and only if $x = y$;*
- *(Symmetry) $|xy| = |yx|$*
- *(Triangle Inequality) $|xz| \leq |xy| + |yz|$.*

Furthermore, hybridization decisions between two mers x and y made by comparing their h-distance against an appropriate threshold τ agree about 80% of the time with one made using the Gibbs' Energy Nearest Neighbor Model.

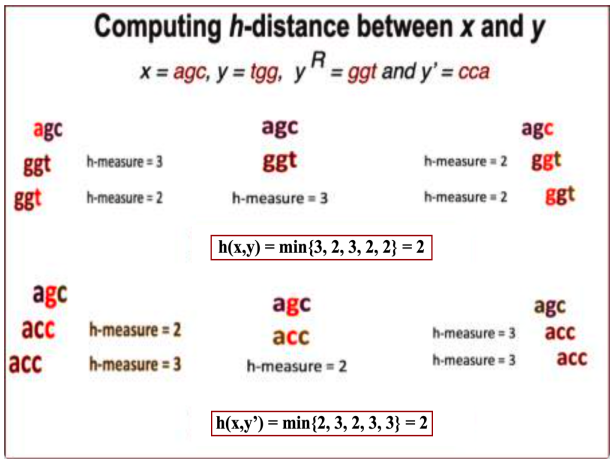

Figure 1: Workflow for computing the h-distance $|xy|$ between any two pmers x and y by optimizing the frameshift for WC-complementary pairs.

These metric properties of the h-distance can be used to solve the DNA CODEWORD DESIGN problem approximately, so that data representations can be built and reasoned about as though they were physical objects like mass, centroids and so on. Further, it reveals some deeper structure of DNA hybridization landscapes.

Definition 3 (Isometry)**.** *An* isometry ϕ *of a DNA space* $\mathbf{D}_n$ *is an h-distance preserving transformation* $\phi : \mathbf{D}_n \to \mathbf{D}_n$*, i.e., for every pair of pmers* $x, y \in \mathbf{D}_n$*,* $|\phi(x)\phi(y)| = |xy|$*.*

Theorem 2 (Isometries in $\mathbf{D}_n$)**.** *Every* $\mathbf{D}_n$ *space possesses the following properties:*

- *Every isometry* ϕ *of* $\mathbf{D}_n$ *must be injective and surjective. In particular, its inverse is also an isometry.*
- *An isometric image of every ball* $B_\tau[x] = \{y \in \mathbf{D}_n : |xy| \leq \tau\}$ *in* $\mathbf{D}_n$ *centered at* x *is also a ball* $B_\tau[\phi(x)]$ *centered at* $\phi(x)$*, for every radius* $\tau \geq 0$*.*
- $\mathbf{D}_n$ *has at least 16 isometries (shown in Table 1 for* $n \leq 8$*.)*

Proof. To prove ϕ is injective, let us assume that two arbitrary pmers $x, y \in \mathbf{D}_n$ have $\phi(x) = \phi(y)$ and $x \neq y$. Since ϕ is an isometry, $|xy| = |\phi(x)\phi(y)| = 0$ and the reflexive property implies that $x = y$, i.e., ϕ is injective. For surjectivity, we know $\phi(\mathbf{D}_n) \subseteq \mathbf{D}_n$ and

Table 1: Known isometries for DNA spaces for $n \leq 8$. They are homomorphic substitutions named bdef to indicate that the characters in acgt are mapped to those in $bdef$, respectively, in a poligo x/x'. Any composition of two isometries is also an isometry.

Name	**Isometry**	**Mapping**	**Transformation**
Identity/			a↔a, c↔c, g↔g, t↔t
WC Complement	acgt	acgt	Reverse+Complement
Polar ϕ_S^N	acgt	catg	a↔c, t↔g
Reverse ϕ^R	acgt	tgca	$\phi(x) = x^R$ (reverse oligo of x)
Polar + Reverse	acgt	gtac	
gc swap	acgt	agct	a↔a, c↔g, t↔t
at swap	acgt	tcga	a↔t, c↔c, g↔g
Polar2	acgt	gtac	a↔g, t↔c

both sets are of the same finite size by injectivity, so they must be equal, i.e., every $y \in \mathbf{D}_n$ must be $y = \phi(x)$ for some $x \in \mathbf{D}_n$, so ϕ is surjective. Thus, it is clear that the inverse is also an isometry.

By a similar argument, for the next property, it suffices to show that $\phi(B_\tau[x]) \subseteq B_\tau[\phi(x)])$. Let $z = \phi(y) \in \phi(B_\tau[x])$ i.e., $|yx| \leq \tau$. Since ϕ is an isometry,

$$|z\phi(x)| = |\phi(y)\phi(x)| = |xy| \leq \tau \ ,$$

so $z \in B_\tau[\phi(x)]$.

The isometries in Table 1 are obtained by homomorphic (character by character) substitutions as shown. Thus, WC matchings are preserved and the h-measure remains unaffected upon substitutions. Therefore, they preseve the h-distance as well. □

These isometries reveal an even more complete picture of the structure of the hybridization landscapes of oligomers of a given size (defined by the h-distance) through their images in DNA spaces. It is particularly interesting that this structure is something we humans are very familiar with the planets earth and Saturn.

Definition 4 (Geometric structures in DNA spaces)**.** $\mathbf{D}_n$ *can be fully described by the following geometric structures about ordinary Euclidean spheres:*

- *Two pmers x and y are an* antipodal *pair if and only if $|xy| = n$;*
- *The* north N *(*south S*) pole is the a^n/t^n-pmer (c^n/g^n, respectively.)*
- *The* northern ice cap P^{n-1} *is the set of pmers x satisfying $|xN| \leq 1$ and $|xS| = n$.*
- *The* equator E_n *is the set of all pmers x equidistant from the poles, i.e., satisfying $|xN| = |xS|$.*
- *The* northern hemisphere *(H^N) is the set of pmers x satisfying $|xN| < |xS|$.*
- *The images under the polar isometry ϕ of these objects are called the corresponding* southern ice cap P_{n-1} *and* southern hemisphere *(H_S).*

Thus, the north and south poles are an antipodal pair, and they partition the full DNA space $\mathbf{D}_n$. The northern and southern equators are identical to the equator.

Definition 5 (Parallels)**.** *For $1 \leq i \leq n$, the i^{th}* parallel *is the set P_i of all n-pmers x satisfying $|xN| = R - i + \epsilon_N$ and $|xS| = R + i - \epsilon_S$, where R is the maximum possible value of $|xN|$ and ϵ_N/ϵ_S are certain constants for each n.*

Theorem 3 (The Equator)**.** *The equator of $\mathbf{D}_n$ satisfies the following properties for arbitrary sizes n:*

- *It consists of (nearly) balanced n-pmers, i.e., the maximum number of occurrences of a's or t's is identical to the maximum number of c's or g's.*
- *The equator is closed under the polar and reversal isometries ϕ, i.e., $\phi(E_n) \subseteq E_n$.*
- $\mathbf{D}_n = H^N \cup E_n \cup H_S$.

Proof. For the statement, by definition, a pmer z in the equator E_n satisfies $|zN| = |zS|$. Now, from the definition of h-distance, it is easy to verify that $|zN| = n - max^N(z)$ and

$|zS| = n - max_S(z)$, where $log_b z$ is the number of occurrences of base b in z and $max^N, max_S, min^N, min_S$ are the functions defined by

$$max^N(z) = max\{log_a z, log_t z\}, \text{ and } max_S(z) = max\{log_c z, log_g z\} ,$$

$$min^N(z) = min\{log_a z, log_t z\}, \text{ and } min_S(z) = min\{log_c z, log_g z\} ,$$

respectively. Therefore, $max\{log_a z, log_t z\} = max_S\{log_c z, log_g z\}$

To prove the equator is closed under the polar isometry ϕ_S^N, let z $\in E_n$ so that $|zN| = |zS|$. Noting that $\phi_S^N(N) = S$ and *vice versa* and applying the polar isometry ϕ_S^N to both sides, we get

$$|\phi_S^N(z)S| = |\phi_S^N(z)\phi_S^N(N)| = |\phi_S^N(z)\phi_S^N(S)| = |\phi_S^N(z)N|$$

i.e., the polar isometric image of z also lies in E_n. For the reversal isometry ϕ^R, likewise,

$$|\phi^R(z)S| = |\phi^R(z)\phi^R(S)| = |zS| = |zN| = |\phi^R(z)\phi^R(N)| = |\phi^R(z)N|$$

i.e., $\phi^R(z)$ also lies in E_n. For the third statement, a pmer $z \in \mathbf{D}_n$ either satisfies $|zN| = |zS|$ (lies on the equator) or it does not. If not, either $|zN| < |zS|$, i.e., it is in the northern hemisphere, or $|zN| > |zS|$, i.e., it is in the southern hemisphere. □

Definition 6 (Ellipses)**.** *Given a* $\mathbf{D}_n$*, the* ellipse *with foci given by two n-pmers* f_1, f_2 *and a constant* $c \geq 0$*, is the set of n-pmers x satisfying the condition* $|xf_1| + |xf_2| = c$*. The* polar ellipse *(as shown in Fig. 2) has foci at the poles N, S and the maximum possible value for c.*

Theorem 4 (Properties of Polar ellipses)**.** *The nonempty ellipse with two poles* N, S *as foci (called the* polar ellipse*) with the largest c is at least* $2R$*, where R is the maximum possible distance of a n-pmer in the equator from the poles and includes the whole equator* E_n*.*

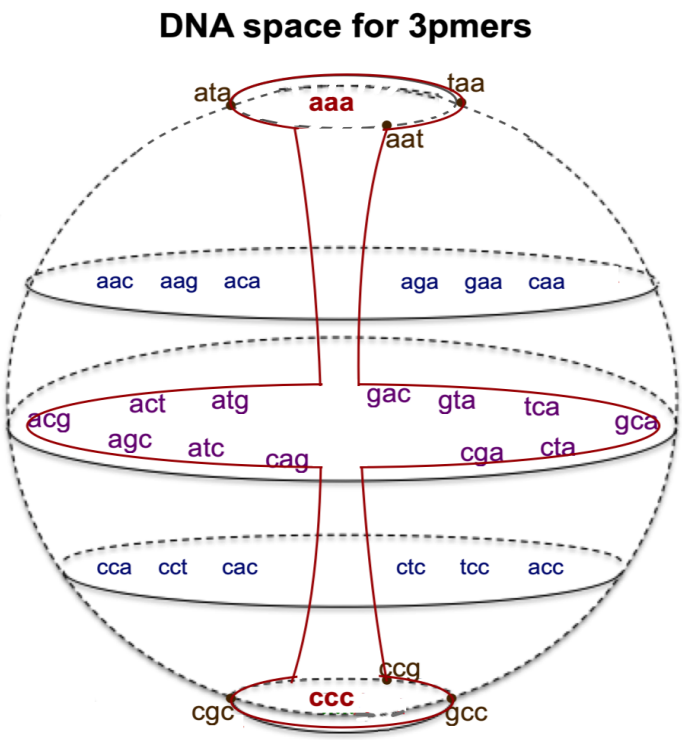

Figure 2: Rendition of the structure of an ellipse in DNA space for 3 pmers with the poles as foci (the actual 3D Euclidean distances are not identical to their h-distances.) The pmers on both icecaps and the equator form the polar ellipse (in red.)

Proof. Since $|NS| = n$, there follows from the triangle inequality that the radius of the $\mathbf{D}_n$ space $R = \max_{x \in E} |xN| \geq n/2$ where E is the equator and N represents the north pole and S represents the south pole. For all $x \in E, |xN| = |xS|$, $n = |NS| \leq |xN| + |xS|$ and hence, this particular ellipse has $c = 2R$. Moreover, this polar ellipse must include all pmers on the equator. □

Definition 7 (Centroid)**.** *Let $k > 0$ be an integer and S be a set of pmers in $\mathbf{D}_n$ of size $|S|$ and w_z ($z \in S$) be a set of real-valued weights for its elements. The (weighted) k^{th}* error function *$SE_w^k\ D_n \rightarrow R$, is defined as the average k^{th} powers of the h-distances from z to a pmer in S, i.e., $SE_w^k(z) = \frac{1}{|S|} \sum_{x \in S} w_z |zx|^k$. A pmer $z \in D_n$ is a k-*centroid *of S (or simply* centroid *if $k = 2$) if and only if it minimizes $SE_w^k(z)$, i.e. $a = \arg\min_z SE(z)$ across all z in $\mathbf{D}_n$.*

Theorem 5 (Properties of Centroids in $\mathbf{D}_n$)**.** *An isometric image of a centroid of* $\mathbf{D}_n$*, is also a centroid. There are several centroids in a* $\mathbf{D}_n$*.*

Proof. Let z is a centroid of $\mathbf{D}_n$. A squared error function of a pmer a can be defined as

$$SE(a) = 1/|\mathbf{D}_n| \sum\nolimits_{x \in \mathbf{D}_n} |ax|^2$$

Using the definition of a centroid, we know z minimizes SE. Since an isometry preserves h-distance, $|zx| = |\phi(z)\phi(x)|$, i.e.,

$$SE(z) = 1/|\mathbf{D}_n| \sum\nolimits_{x \in \mathbf{D}_n} |\phi(z)\phi(x)|^2 = SE(\phi(z))$$

.

If z is a centroid and minimizes $SE(z)$, so will $\phi(z)$ and therefore, it must be a centroid in $\mathbf{D}_n$. □

These findings hint to the shape of these spaces being very similar to that of planets in our solar system (like earth and Saturn) as shown in Figs 3 and 4.

Such structural features can be computed using brute-force method for small values of $n(\leq 8)$. But as the value of n increases, the size of $\mathbf{D}_n$ also increases *exponentially* (for e.g., D_3 has 32 pmers and D_5 has 512 pmers but D_7 has 8,192 pmers) causing combinatorial explosion. Beyond D_8, an exhaustive search of the space is practically impossible. Therefore, algorithms to generate all and only pmers satisfying the conditions of these properties would be useful for higher values of n. Two of them are given below. Similarly, these attributes for the lower (n up to 8) are shown in Tables 2, 3, and 4 below.

With these structures in place, we can proceed to define GenISs to tackle the problems described above.

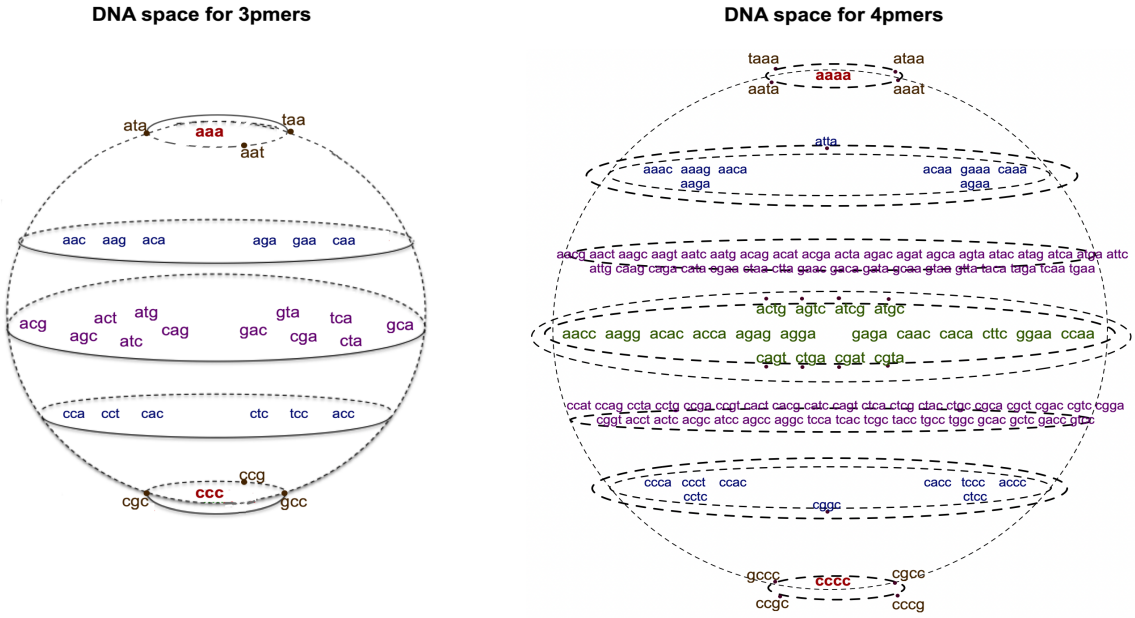

Figure 3: Rendition of the structure of DNA spaces, for left: 3 pmers (shape similar to earth) and right: 4pmers (shape similar to Saturn.) This representation is not isometric for $n = 3$, although the relative separation between the location of pmers in 3D Euclidean space is indicative of their actual h-distance in $\mathbf{D}_n$. (Watson-Crick palindromes have been excluded.)

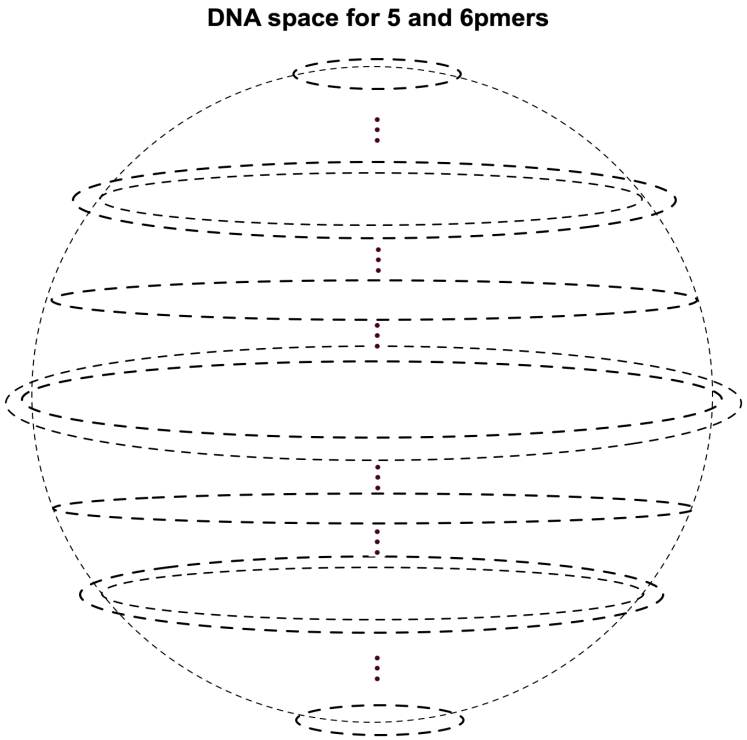

Figure 4: Rendition of the structures of DNA spaces 5 and 6 pmers. This representation is not isometric although the relative separation between the location of pmers is indicative of their actual h-distance in 3D Euclidean space. (WC palindromes have been excluded.)

Algorithm 1: Algorithm to Generate the Equator E_n of $\mathbf{D}_n$

Input: An integer $n > 1$

Output: A list containing all and only pmers in the equator E_n

begin

1. Set
   (a) $m = 0$
   (b) $temp = []$
   (c) $Eq = []$
2. For i in $(1, n)$:
   (a) $m = n–i$
   (b) $rems = `cgt'$
   (c) $remPms = []$
   (d) $j = n–m$
   (e) Generate all possible permutations of rems with a common length j and store them in $remPms$.
   (f) For each pm in $remPms$
      i. Eq.append(concatenate(a^n, pm))
   (g) Set $rems = \ 'acg'$, $remPms = []$, $j = n–m$
   (h) Set $rems = \ 'acg'$, $remPms = []$, $j = n–m$
   (i) Repeat step 3(e)
   (j) For each pm in $remPms$
      i. Eq.append(concatenate(t^n, pm))
   (k) Remove duplicates from Eq
   (l) If n is even, remove Watson-Crick Palindromes from Eq
3. Return Eq
4. End.

Algorithm 2: Algorithm to Generate the i^{th} Parallel of $\mathbf{D}_n$

Input: An integer $n > 1$

Output: A list containing all and only pmers in i^{th} Parallel

begin

1. Set $P_i, revP_i = [], []$
2. For each a and b (where $n > k > l > 0$ and $a + b = i$),
   (a) Set $seed = a^l c^{n-k}$ and $rems = `gt'$
   (b) Generate all possible permutation of rems so that a common length is $k - l$ and store them in a list, temp
   (c) For each pm in temp,
      i. $s = concatenate(seed, pm)$
      ii. append s to $tempP_i$
3. For each pm in $tempP_i$
   (a) permutate pm changing positions of characters in pm in all possible ways
   (b) append the resulting strings from 4(a) to P_i
4. Remove duplicates from P_i
5. For each pm in P_i, compute reverse of pm and append it to $revP_i$
6. Append $revP_i$ to P_i
7. Return P_i
8. End.

Table 2: Parallels P_i for the DNA spaces $\mathbf{D}_n$ (by their first lexicographic n-mers, e.g. *ac* stands for pmer *ac/gt*.)

n	$\lvert xN\rvert$	$\lvert xS\rvert$	R	i/ϵ_N, ϵ_S	$\lvert P_i\rvert$	P_i
3	2	2	2	$0/0,0$	12	$\mathrm{E} = \left\{ \begin{array}{c} \text{acg,act,agc,atc,} \\ \text{atg,cag,cga,cta,} \\ \text{gac,gca,gta,tca} \end{array} \right\}$
	1	2		$1/0,1$	6	$\mathrm{P}_1 = \left\{ \begin{array}{c} \text{aac,aag,aca,aga,} \\ \text{caa,gaa} \end{array} \right\}$
	1	3		$2/1,1$	3	$\mathrm{P}_2 = \{aat, ata, taa\}$
	0	3		$3/1,2$	1	$\mathrm{P}_3 = \{aaa\}$
4						$\lvert E_4\rvert = 20$ (see Fig 3.)
5						$\lvert E_5\rvert = 120$
6						$\lvert E_6\rvert = 580$
7						$\lvert E_7\rvert = 1,820$
8						$\lvert E_8\rvert = 5,832$

Table 3: Nxh bases from previous work [3, 4] for different DNA spaces.

Basis	**Length**	**Size**	τ	**Avg**	**Entropy**
3mE4b	3	4	1.1	1.09	0.45
4mP3-3	4	3	2.1	1	0
8mP10	8	10	4.1	1.1	0.57

Table 4: Centroids for the DNA spaces $\mathbf{D}_n$ of all pmers of length n (as given by their first lexicographical n-mers, e.g. *ac* stands for pmer *ac/gt*.)

n	h**-centroids**
3	aca, aga, cac, ctc
4	acca, agga, caac, cttc
5	accat, aggat, caacg, cgaag, gaagc, gcaac, tacca, tagga
6	acaagc, agaacg, atccac, atggag, caccta, cgaaca, gaggta,gcaaga
7	actccat, agtggat, cagaacg, cgaacag, gacaagc, gcaagac,tacctca, taggtga
8	actatccg, agtatggc, attcgctg, attgcgtc, cggtatga, ctgcgtta, gcctatca, gtcgctta

CHAPTER 2

Genomic Information Systems (GenISs)

This chapter describes two families of GenISs as a platform/framework to encode and extract useful information stored in/to DNA as described below.

Genomic Coordinate Systems

A family of nxh bases used by the proposed GenISs to extract information from data contained in DNA sequences, along with the means to assess their qualities, is described next.

Noncrosshybridizing (nxh) Bases

Microarrays have been the standard and popular tool to extract information from DNA sequences in biology. They are planar substrates such as glass, mica, plastic or silicon, where DNA strands are affixed to allow specific bindings of bio-samples collected from an organism [25]. During the early 1990s, the first microarray experiments were performed using complementary DNA (cDNA) affixed to the microarrays. The length of a typical cDNA is 500-2500 base pairs, and they are widely used in gene expression assays [25]. Since 1990s, microarrays have been refined to capture and mine genomic and metabolomic information. The information gathered by these tools has wide applications in the fields of biology, medicine, health and scientific research.

However, microarrays have a few serious disadvantages. First, the analysis relying on their redouts gives results that are hardly reproducible because of the high uncertainty of hybridization of targets to probes. The probes may not crosshybrize because they are affixed to the chip far apart, but the targets are floating in solution. No constraints are implemented in these chips to minimize crosshybridization between targets. As a result,

the results are not accurate and hence unreliable due to the lack of reproducibility of results, as argued in [26]. A second disadvantage of microarrays is that they might miss target strands if they do not hybridize to any probe on the microarray, and thus miss signals that could yield useful information.

Recent advances in next generation sequencing (NGS) have allowed us to move away from microarrays directly to DNA fragments coding for proteins that can be used for processing and analysis instead. Currently, a number of NGS platforms using different sequencing technologies are available. These platforms perform sequencing of millions of small fragments of DNA in parallel. Some bioinformatic analyses join these fragments by mapping the individual reads to the human reference genome [27]. However, analyzing the sequences generated using these platforms is a big challenge. Through the use of deep learning models on these sequences directly to extract useful predictors automatically, the disadvantages of microarray analyses could be avoided (no risk of unwanted hybridization as the phenomenon is not considered at all.) However the performance of such networks is highly dependent on the quality and/or relevance of the data as well as the size of the data. In particular, such models can pretty much memorize data when it is limited. In fact, there is an ongoing debate between two extreme approaches (i.e., feed raw data to a model without any processing to avoid bias vs manual selection of features that might be important) in the field of machine learning (ML) and researchers have concluded that there should be a trade-off [28]. Further, the results obtained using these methods are not explainable since it is not clear how the model is making decisions (e.g. non/cancerous) that would allow a human to rationalize and accept or reject the decisions.

In this work, we use an entirely different approach. By exploiting the structural properties described in Section 1, a selected set of nxh pmers (nxh bases) could be used to reduce the dimension of DNA sequences and thus extract more relevant information about the sequences based on the knowledge of Gibbs energy landscapes. These pmers will be referred to as *probes* (contrary to the standard use in biology where they are referred to as

targets.) [29] and [9] show how this problem is reduced to a popular and well-researched problem in geometry, a sphere-packing problem. There are several advantages of these bases over microarrays and NGS.

First, these bases can be used to transform any arbitrary sequence to numerical features. These vectors could be used to train any conventional statistical and machine learning models like regression models, support vector machines, random forests, decision trees, multilayer perceptrons and so on. The drawbacks of deep networks requiring abundant data for effective learning can be avoided with the use of these models based on nxh features. Furthermore, the results obtained using these bases can be rationalized because they reflect deep knowledge of the structural properties of DNA spaces. Hence, the results will be more explainable.

However, obtaining these bases is very difficult in general because CWD is computationally difficult (**NP**-complete) [29], [9]. Fortunately, the deep structural properties of DNA spaces (as discussed in Section 1) afford a method to obtain nxh bases of high quality, as discussed next.

Table 5 shows a number of such nxh bases along with the quantification of their quality. These bases were obtained using a judicious selection among the centroids of the parallels of $\mathbf{D}_n$.

With the design of nxh bases in hand, we can proceed to compute feature vectors, or genomic signatures, for target DNA sequences, as defined below.

We used a custom written Python code to shred the cleansed sequences into fragments of uniform length n (same as that of the probes in a nxh basis.) Once the shreds are obtained, they were tested for hybridization with the probes and a vector of total pmer counts present in the sequences was computed. The normalized vector obtained using the partition function will be referred to as a genomic signature of the given target DNA sequence. Perl scripts were used to compute the h-distances of these shreds in a sequence

Algorithm 3: Algorithm to compute nxh bases of $\mathbf{D}_n$

Input: An integer $n > 1$

Output: A centroidal nxh basis B of Dn

begin

1. Set $B = [a^n]$
2. Generate the first parallel, P_1 in the northern hemisphere and compute its h-centroids.
3. Set $i = 2$
   (a) Generate the $(n-i)^{th}$ parallel and compute its h-centroids and store them in $hSet$.
   (b) If there is any intersection between $hSet$ contains B
      i. Set $i = i + 1$
      ii. If $i < n$, go to step 3(a), else go to step 5;

   Else
      i. Choose one h-centroid randomly from $hSet$ and append it to B
      ii. Go to step 3(a)
4. for each probe p in B,
   (a) Compute its polar isometric image p'
   (b) Append p' to B
5. Remove any duplicates from B.
6. Output B.
7. End.

Table 5: New centroidal nxh Bases for larger DNA spaces.

Basis ID	τ	**Length**	**Size**	**Entropy**	**pmers hybridizing to** 0 1 2 **probes**
7miC4Sb	4.1	7	4	0.15	4 8,024 164
7miC4Sa	4.1	7	4	0.17	4 7,997 191
6miC4Sa	4.1	6	4	0.26	0 1,928 88
5miC3Spr2	3.1	5	3	0.31	0 483 29
5miC3Mg	3.1	5	3	0.34	0 479 33

to determine their hybridization affinity to the m probes in the nxh basis. This affinity can be expressed as an mD vector.

Definition 8 (genomic signature)**.** *Let B be a nxh basis of probe length $n > 1$ with m probes, $\tau > 0$ and x be a DNA sequence. The m-dimensional (mD)* genomic signature *of x on B for h-threshold τ is defined as follows:*

- *shred x to nonoverlapping fragments of size n (ignoring any shorter leftover shreds, if any);*
- *for each probe $z_i \in B$, compute the total number of shreds in x that hybridizes with z_i for the given threshold τ;*
- *normalize the mD vector obtained from the previous step using the partition function (i.e., dividing by the total number of shreds.)*

Quality Assessment

There are two kinds of assessments of the quality of nxh bases. The first one is a principled inherent metric where the quality of the information extracted by these bases is quantified regardless of their application. The second one is by quantifying the quality of solution models (by standard quantitative metrics discussed below) to problems arising in applications based on the features extracted by nxh bases from genomic sequences.

The first metric requires the standard concepts from probability theory, namely, a sample space Ω (the set of all possible outcomes of a random experiment), a (discrete) probability distribution on it, random variables (RVs, observation on all possible outcomes in Ω), and the expected value of a RV.

The metric is the (Shannon) Entropy quantifying the degree of uncertainty of a random process [30] for the appropriate random variable counting the number of probes in B that a random pmer hybridizes to under a given stringency τ.

Definition 9 (Shannon entropy)**.** *Let X be random variable taking on a finite number of values $x_1, x_2, \cdots, x_n$ with corresponding probabilities $p_1, p_2, \cdots, p_n$. The Shannon entropy $H(X)$ of X is the average uncertainty or average information content of the underlying probability distribution of X given by*

$$H(X) = -\Sigma_i x_i \log(x_i) \ .$$

An ideal nxh basis (like B=4mP3-3 at $\tau = 2.1$, obtained through an exhaustive search of $\mathbf{D}_4$ shown in Table 3) will produce a noise-free genomic signature [31] with $H(\,B\,) = 0$. Table 5 shows the proposed new nxh bases having entropies less than 0.5. As we increase the value of n, the value of the entropy comes closer to 0. Thus, as the length of the probes grows, the quality of information extracted also increases for these nxh bases obtained by the centroid methods.

We also performed a control for the quality of the process to obtain them, i.e., how noncrosshybridizing they are as bases (in terms of separation and coverage of $\mathbf{D}_n$ by all the balls of radius τ centered at them.) For 6-pmers, we selected a random set of pmers containing the same number and the length of pmers as in the nxh basis in Table 5, then repeated the same procedure 16 times for 6-pmers and 7-pmers. The test is the comparison of their quality metrics (the average of the expected number of hybridizations to the given probes and their entropies) to the nxh bases. We also performed two t-tests each with a null hypothesis "the mean entropy of the sample is the same as the entropy of our corresponding nxh basis", for entropy for example. For $\alpha = 0.05$ and one-tailed test, the critical value is 1.746. Our computed t-values for two bases (11.748 for 7miC4Sb; 11.475 for 7miC4Sa and 14.159 for 6miC4Sa) are greater than the critical value. Thus, the null hypotheses are rejected, i.e., the quality of the information extracted by these bases should be statistically significantly better than that by a random set of pmers.

Pmeric (pmc) Coordinate Systems

Another family of GenISs can be obtained by using pmeric (pmc) coordinates to extract information from DNA sequences. The coordinate systems along with the means to assess their qualities are described next. They use the patterns of hybridization affinity to all h-centroids (2^{nd} power error function defined in Section 1) to represent random pmers.

h-centroids

Proceeding in analogy with the genomic signatures, we used a Python script to shred these sequences into uniform length pmers of size n. Each pmer in $\mathbf{D}_n$ can be viewed as a point with certain weights given by its h-distances from the h-centroids of $\mathbf{D}_n$. However, due to the symmetries of $\mathbf{D}_n$, there is no unique h-centroid, unlike on earth where all objects are attracted towards a unique center of mass due to gravitational forces. Further, more than one pmers might share the same coordinates. However, the number of appearances of these pmers in genomic sequences of different organisms are likely to be different if we place masses at a pmer of size equal to the ratio of the total number of times the pmer occurs in x to the total number of n-pmers shreds in x. Thus, distinguishing several organisms/abiotic datapoints (e.g., images) based on these vectors is still possible. These vectors will be used as pmeric signatures for the respective organisms.

Definition 10 (pmeric signature)**.** *Let m be the number of centroids in D_n and x be a DNA sequence of shred size $n > 1$. The m-dimensional (mD)* pmeric signature *of x is obtained as follows:*

- *shred x into nonoverlapping fragments of size n (ignoring any shorter leftover shreds, if any);*
- *for each centroid $z_i \in \mathbf{D}_n$ and for each unique shred x_j, compute $y_{ij} = w_j|z_i x_j|$,*

where w_j is the fraction of the number of occurrences of x_j to the total number of shreds in x;

- *the i_{th} component of a pmeric signature of x is given by the average of the y_{ij} across all shreds x_j.*

Quality Assessment

An entropic quality assessment of pmeric coordinates cannot be done because these centroids are very close to each other in the expanse of the entire $\mathbf{D}_n$. We must resort to the second option, using the same quantitative metrics for ML solutions based on them using some application problems.

The h-centroids (up to D_{12}) were computed and are shown in Table 4. Computing such centroids requires a brute-force search of an entire space. As we mentioned earlier, with the increment of the length of pmers, the size of such spaces explodes combinatorically. Thus, it is impossible to perform such a search of the space beyond $\mathbf{D}_8$.

Genomic Information Systems (GenISs)

GenISs are analogous to a Geographic Positioning System (GPS) for positional information on planet earth. Methods developed for computer networks (such as the internet, the web, and wireless communication) have enabled billions of people on the planet to use a cell phone to communicate. This requires, in particular, the ability of the systems to determine the location of the phone anywhere on the planet so as to quickly establish paths to send messages through. That is similar to what biological organisms do (e.g., living cells and brains), where location, physical proximity and obstruction represent hard anchoring constraints that are exploited for biological function, such as cell membranes, organs and organisms. Without them, biological reality, in particular organs and living organisms as we know them, would be impossible. A GenIS is aimed at developing a similar system for

abiotic/biotic information processing where planet earth is replaced by the entire biome on planet earth. The development of these systems was initiated in [3] and is refined through several iterations [11], [12], [4].

A GenIS for genomic information processing is an integrated software platform comprising a coordinate system (e.g., genomic or pmeric) to transform an arbitrary DNA sequence into a numeric vector and conventional statistical or ML models designed to solve a data science problem using these coordinates as input features.

Methods

Sections 2 and 2 described the process of transforming a genomic sequence into numerical feature vectors. These numerical vectors representing a group of DNA sequences (proxies for a group of organisms, or taxon) can be used as feature vectors to train some conventional machine learning models. Here, we summarize several machine learning models used in this dissertation as components of the integrated platform provided by our GenISs.

Decision trees

A Decision Tree (DT) is a decision support classifier that uses a comparison between features in a feature vector to successively split the data into pieces (e.g., halves) and other possible criteria to assign a category to the feature vector [32].

Random Forests

A Random Forest (RF) [33] is an ensemble classifier that combines multiple decision trees using a bootstrap aggregation (also called bagging) technique. Specifically, a RF learns multiple decision tree classifiers that have low correlations with one another. Increasing the number of uncorrelated decision tree classifiers in the ensemble reduces the variance of the RF classifier.

k-Nearest Neighbor

A k-Nearest Neighbor (kNN) classifier is an instance-based classifier that does not

construct an explicit model for classification. Instead, it assigns a classification of a particular data point based on that of k of its neighboring points in the training data.

Neural networks

Neural Networks (NNs) are machine learning models consisting of a number of simple components (called neurons) [34]. They were inspired by the way our mammalian brains process data to solve problems. The neurons interact with each other via some synaptic connections with some predetermined weights. A neuron outputs either a certain input value to the model (input neurons) or a linear combination of the inputs sum of all its inputs from other neurons weighted by its characteristic weights. The neurons arranged in several layers (the first is *input layer* and the last is the *output layer*) in a feed-forward fashion to extract hidden patterns in data (e.g., [4 8 3 1] describes a MLP neural network architecture with 4 input features, two hidden layers with 8 and 3 neurons respectively and 1 output layer with a single neuron coding an answer.) Several learning algorithms are available to train these networks for specific tasks. Some of them require well designed features through some manual process (for e.g., multilayer perceptrons (MLP), feed forward networks (FFN) and so on). However, a recent research has shown that so-called Deep networks, DNN (networks based on deep learning algorithms) can extract abstract and fine-grained features from a raw dataset.

Adaboost classifier

Adaboost classifier (AB) [35] is an ensemble boosting classifier that combines multiple poorly performing classifiers to yield a strong classifier. The core idea behind this classifier is to set the weights of multiple classifier and train them on a data sample such that in each iteration, more accurate predictions of unusual observations are ensured.

Support Vector Machines (SVM)

Support vector machines (SVM) [36] are supervised learning algorithms that find a hyperplane (or a set of hyperplanes in mSVMs) in a high-dimensional space that can be used to separate input data points under consideration into two or several categories. We

used a nonlinear SVM classifier, so-called radial basis function (RBF) by applying a kernel trick to maximum-margin hyperplanes [37].

Self-Organizing Maps (SOMs)

Self-Organizing Maps (SOMs) belong to the family of neural networks trained by unsupervised algorithms (using unlabeled data, with no categories assigned *a priori*) to produce a low-dimensional discretized representation of the input space of training dataset [38]. They use competitive learning, i.e., a neighborhood function is used to preserve the topological properties of input examples.

Voronoi Diagrams

In dD Euclidean space $\mathbf{R}^d$, k points (called the centroids) define a nearest neighbor partition (classification) of the space into k classes, i.e. a random point x belongs to the class defined by its nearest centroid. The categories thus defined are polygonal regions with boundaries determined by the geometric perpendicular bisectors joining two points. This partition is called the *Voronoi diagram* of the given centroids [39].

kMeans clustering

kMeans clustering [40] is an unsupervised learning algorithm that aims to find an ideal set of k *centroids* to determine a partition of the datapoints into k clusters in such a way that each point belongs to the cluster with the nearest centroid. We used a careful seeding of the centroids suggested by [41] to speed up the convergence to a stable set of centroids for our samples.

Quality Assessment

As a platform (software package) for solving problems, an assessment of a GenIS can only to be based on the problem it solves. This problem usually falls into three categories: classification, clustering and prediction problems. A *classification* problem calls for assigning a pre-defined category to a given data point, and for a solution based on supervised learning so labels for a dataset is already available. A *clustering* problem calls

Table 6: Several quantitative metrics

Metric Problem	Classification Problems	Clustering Problems	Prediction Problems
Accuracy	Yes	N/A	Possible with preprocessing
Precision	Yes	N/A	
Recall	Yes	N/A	
F1 score	Yes	N/A	
Silhouette	Possible with preprocessing	Yes	
Relative Error (RE)		N/A	Yes

for a solution based on unsupervised learning and hence, such labels might not be handy. A *prediction* problem calls for an estimation of the value of a function f defined on the data set Ω (e.g., a response variable.) Hence, the choice of a metric to quantify the quality of a solution highly depends on the nature of the problem under consideration. Table 6 summarizes different available metrics suitable for an application.

We have described two families of GenISs, along with the types of analyses done to quantify their quality. These GenISs serve as tools for analyzing both biotic and unstructured abiotic data (like texts and images.) They originate in the deep structural properties of DNA spaces and will prove capable of large reductions in the dimensionality of large data sets to very few dimensional feature vectors, as will become evident in the following two chapters.

CHAPTER 3

Biotic Applications of GenISs

This chapter describes how GenISs can be used to solve some fundamental problems in biology.

Phenotypic Feature Prediction

A classical problem in biology is to decouple nurture from nature i.e., determine to what extent do genomic sequences generally and causally determine phenotypic features (i.e., the physical and biochemical traits [42]) of a biological organism, environmental conditions aside; likewise, how do environmental conditions modify the phenotype determined by a genotype. Estimating these features in a biological organism from their genomic sequence alone has major applications in anthropological paleontology and criminal forensics, for example. The standard procedure to do so relies on *-omics (i.e., genomic, proteomic, metabolomics and so on) analyses as described in Section 3. The common answers in biology are usually qualitative (e.g., the offspring looks like the parent.) In this work, we use GenISs (as described in Section 2 and presented in [29]) to make more refined quantitative predictions using DNA sequences alone (e.g., partial subsequences of the Cytochrome C Oxidase subunit I (COIs) and whole genome) bypassing the complex process of protein synthesis [4].

For this purpose, we define the problem precisely as follows. We are given a DNA sequence of an organism s in a certain taxonomic group (taxon) and a definition of a phenotypic feature of s to predict quantitatively.

PHENOTYPIC PREDICTION(T,F)

INSTANCE: A DNA sequence x from an organism s in taxon T

QUESTION: What is the quantitative measurement of F?

Example features are the area of cephalic apotome/postgenal clef or body color in cephalic apotome.

To show an example of how GenISs enable us to solve this problem, we use two model organisms. For the first one we use blackfly because species in this group have high degree of intra- and inter-specific morphological variation that usually places limits on their taxonomic determination [43], [44], [45]. The phenotypic features we chose were the areas and body pattern of the cephalic apotome and postgenal cleft in 20 specimens from the genus *Simulium* [1]. For the second organism we chose *Arabidopsis thaliana* because it is of exceptional interest, e.g., for food production. The phenotypic features, namely life span and rosette dry mass, were selected from 120 specimens [46]. (The precise definitions of these features are given below in Section 3.)

Prior Work

The major problem to tackle is to infer certain components of phenotype (a set of physical and biochemical traits that characterize a given organism both through space and time [47]) as a function of DNA sequence alone, leveraging the relationship between the genome and the phenotype of an organism. This complex relationship remains a major challenge in understanding biological morphogenesis despite enormous progress in the last half century –as seen, for example, in the survey of the field in [48], [47] or the whole cell model in [49] for the human pathogen *Mycoplasma genitalium*. As the blueprint of life [10], it is plausible that DNA deeply encodes for some of these traits in an organism, but the analysis and quantification of the degree to which the genotype of an organism determines its phenotype remains unearthed in full. This is a question of critical importance in matters such as "yield improvement in food and energy crops, environmental remediation using microbes and plants and understanding complex networks that control fundamental life processes," as well as in fundamental, translational and applied research [47][p. 3]. This question has been partially addressed in [50], where a software called Traitar was introduced as a

microbial trait analyzer for deriving phenotypes from a genomic sequence. They described 67 microbial traits, including gram positive, gram negative, bacillus/coccobacillus, coccus, motile, and pigmentation (e.g., yellow.) Traitar reliably assigned phenotypic features to bacterial genomes of 572 species in 8 phyla based on incomplete single-cell genomes and simulated draft genomes. Currently, Traitar predicts the presence or absence of these phenotypic features for a given input DNA sequence by converting them to associated proteins, which are then used as predictors to classify a phenotype, with accuracy about 75%. Further, a good survey of conventional phenomics theory and practice today can be found in [48], [51], [47]. They require large samples with representative genetic variants of the actual population. The major goal of these analyses is to either identify causal relationships between genes and phenotype (for e.g., relationship between certain genes and eyes color in a human) or to reveal correlations between seemingly unrelated phenotypic features (e.g., symmetry in the wings of a butterfly.)

Data

For phenotype encoding of *Simulium*, 20 larva specimens collected and curated for the work in [1] were used as sample data. The specimens belong to two closely related species, *Simulium ignescens* and *S. tunja*. A critical subsequence of the mitochondrial marker COI of these specimens (of length about 1,500 bps) were amplified and sequenced by the authors of [1], using the primer suggested by [52]. These segments from the 5' terminal markers region (length about 658 bps) was selected to compute their genomic signatures because they have proven to be very suitable for distinguishing closely related species of animals [53], [54], [55] despite their high evolutionary rate. Since these two species also exhibit high geographical dispersal, they were collected from several separate geographical regions in the Colombian Andes mountains [1] and offer an appropriate dataset to address the target question. Two morphological structures in blackfly larvae (the cephalic apotome

and the postgenal cleft) were used as phenotypic features, as shown in Fig 5. These features were precisely coded as follows:

Cephalic Apotome (CA) Area

Area of a trapezoid (made of a rectangle and two adjacent triangles);

Spot Pattern (CA Spots)

A vector containing the lengths of the arms of the latin cross determined by opposing points (in white diamonds), plus the rotation angle of the cross with respect to the symmetry axis of the apotome;

Postgenal Cleft (PG) Area

Area of the region under the mandible of the fly;

Body Color

One of four colors for a specimen's thorax and abdomen, as curated in [1] (I: Yellowish with green in the middle; II: Blue pigmentation, III: Intensive yellow and green bodies; and IV: Pale yellowish bodies.)

For *A. thaliana*, data from the 1001 Genome project [56] were obtained from their Supplementary materials [46], available online. Sequences and corresponding phenotypic measurements were selected from 120 specimens, as follows:

Life Span (LS)

Estimate of the days between the first day of growth (d_0) and plant maturity (end of reproduction) at which each specimen was harvested [46].

Rosette Dry Mass (RDM)

The weight of the replicate's "harvest at maturity, at the end of reproduction when the first fruits have started to senesce." Rosettes were dried for at least 3 days at $65°C$ and then weighed with a microbalance [46].

In order to develop a scalable machine learning (ML) model, sufficient data representative of the variability of the features in the population are required. However, in the preparation of the original data in [1], the morphology of between 20–35 of specimens

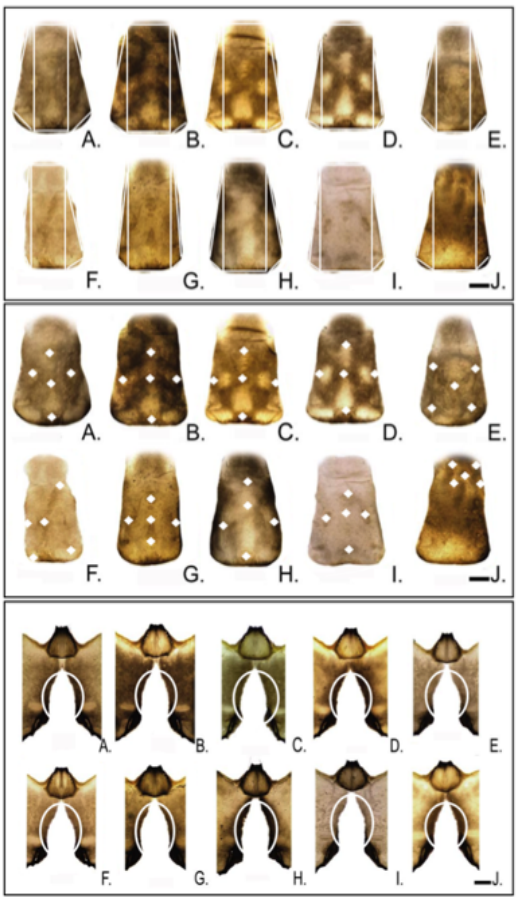

Figure 5: Top: Ten morphoforms of the cephalic apotome in [1] were coded as approximate areas or lengths of the appropriate regions for the phenotypic feature (s): Top: cephalic apotome area; Middle: spot pattern; Bottom: postgenal cleft. (Figures reprinted with permission from [1].)

per morphoform were analyzed to establish its typical intra- and inter-specific shape, but unfortunately, the full set of data set of specific morphoforms for each corresponding specimens was not captured. (The few precious data points of the 20 specimens for which sequences and corresponding paired measurements in the same specimens were utterly insufficient.) To address the problem, a group of data for each of the 20 specimens from [1] and each feature was obtained by generating 32 different values under a normal distribution with mean μ equal to the actual measurement of the specimen's feature and common standard deviation equal to that of the data set ($\sigma = 0.08$ for blackfly.) The size of the data set is thus $20 * 32 = 640$ points. Furthermore, each of the corresponding COI sequences in a group were mutated 32 times according to a biological model, namely with a probability of 0.0067 for nucleotides for the second codon position and 0.0333 for the third position, so that the overall rate of mutation of the sequence was about 4%, which is the estimated mutation rate for these species [1]. A point in the data corpus thus consisted of a genomic signature on a basis (obtained from the original COIs as predictors) plus one (five for spot patterns) feature(s) coding for the phenotypic feature for the dependent variable(s), as described above.

Likewise, for 29 *A. thaliana* specimens, relevant genes for the target phenotypic features were extracted from the data in the 1001 Genome Project [56] and the corresponding phenotypic features from their Supplementary data [46], available online. Again, a distribution of $29 * 32 = 928$ was generated based on a normal distribution with the same mean of the actual measurements and standard deviation σ of the sample and the same mutation rates. In order to gauge the effect of the randomization, a second sample of 120 points was extracted from the Supplementary data, available online, in [46] so that the dataset was big enough and no randomization was deemed necessary for this data set.

Results

We used NNs described in Section 2 above as solution models. The networks were trained using genomic signatures on nxh bases described in Section 2 on 70% of the data corpus. These networks were then tested and validated using the remaining 30% of the data points, held out from the networks during training. The performance of these networks was assessed using accuracy in the prediction, as measured by their relative error (RE) to the actual observed values in the specimens, on both training and testing datasets, i.e.,

$$RE = |Observed - Predicted|/|Observed|$$

The results are shown in Figs. 6 and 7 for blackfly. In other words, for a given feature, a model was considered to be good enough if the average RE made by the model is at most 16% ($2 * \sigma(= 0.08) = 0.16$) on the testing set. We used this threshold to assess the quality of our models because there is no prior work predicting phenotypic features quantitatively based on DNA sequence alone, to the best of our knowledge. As shown in Fig. 6: Top, the performance of deep neural networks for predicting the size of the cephalic apotome in blackfly larvae on the bases 3mE4b-2 and 4mP3-3 gave relative errors below 15% on both training and testing dataset, on the average (only results on the first basis are shown throughout.) These values are less than 2σ for all features, so that the models on these bases can be considered acceptable. Similarly, the performance of the deep neural networks for predicting the spot pattern in cephalic apotomes for the blackfly larvae on both bases gave the relative errors around 0.11, i.e., 11% on both training and testing datasets, as shown in Fig. 6: Bottom. Since these values are less than 2σ for all features, the models on these bases are also good enough to be considered of acceptable quality. However, these models exhibit higher values than 2σ for some groups (3b, 9, 8a) so the models tend to make an unreliable prediction if exact values are required specimenwise for a given sequence based on this partial COI alone.

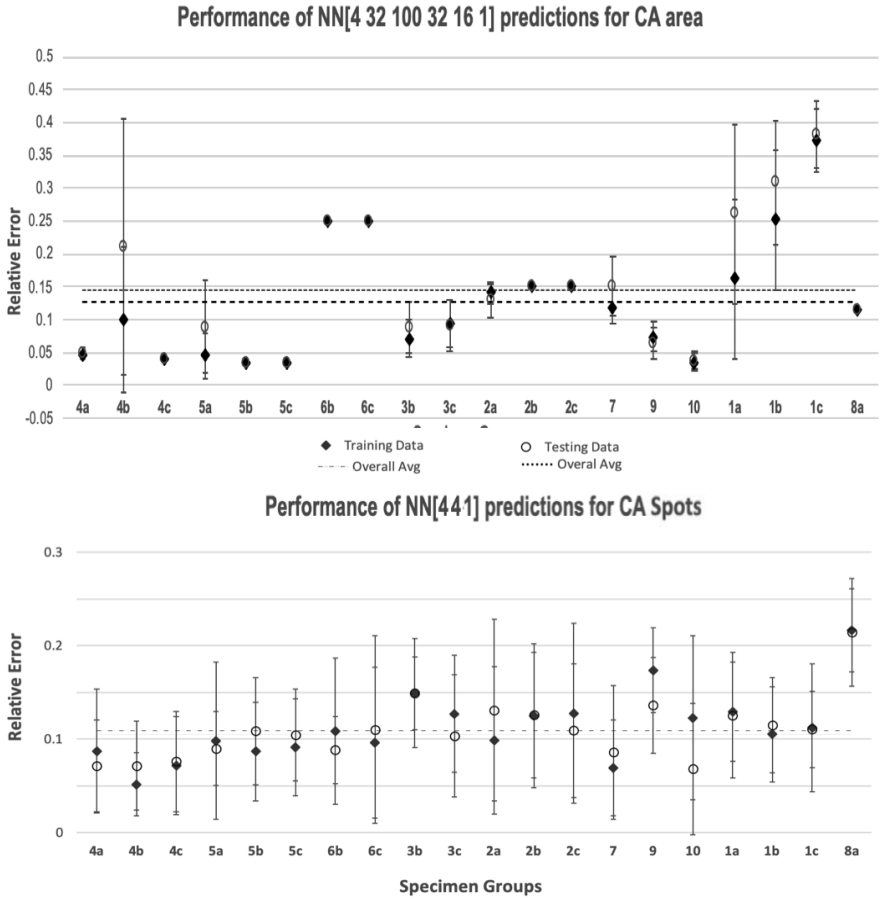

Figure 6: The performance assessment of two deep neural networks for predicting various phenotypic features of blackfly larvae on nxh DNA chip 3mE4b-2. Top: a NN [4 32 100 32 16 1] for cephalic apotome area; Bottom: NN[4 4 1] for CA spot pattern.)

Moreover, the performance of the neural networks for predicting the area of the postgenal cleft on both bases led to an RE below 11% for both training and testing datasets, as shown in Fig. 7: Top, although the value was not below 2σ for three groups of data (4b, 4c, 8a). Thus, the models based on this data are considered to be of acceptable quality for the entire dataset. Similarly, the performance of the neural networks on both bases provided 65% accuracy for predicting body colors of the blackfly larvae, as shown in Fig. 7: Bottom. This implies that the models are not good enough to be considered acceptable for color prediction of the corresponding morphoforms, as they produced results only slightly better than the predictions that could be achieved with just coin flips, despite the fact that the predictions are perfect for most of the groups.

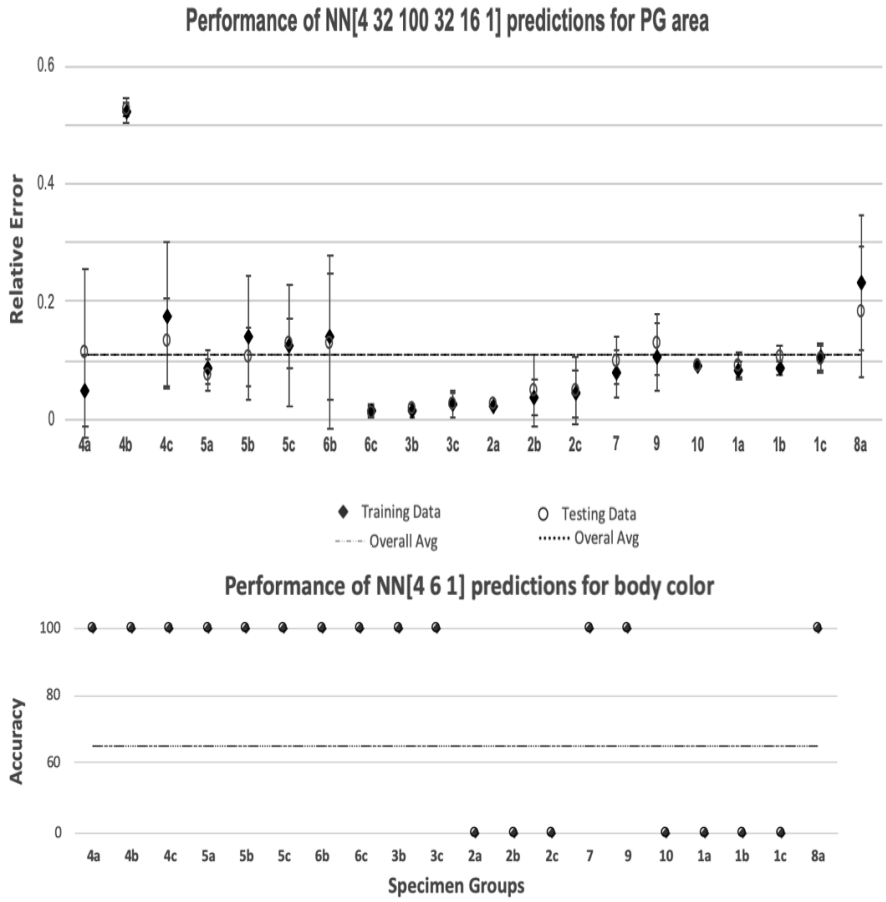

Figure 7: The performance assessment of two more deep neural networks for predicting various phenotypic features of blackfly larvae on nxh DNA chip 3mE4b-2. Top: a NN [4 32 100 32 16 1] for postgenal cleft area; Bottom: a NN[4 6 1] for body color. It may seem strange that this neural network performs better for some groups in the testing phase than in training. This is likely to be since a single network is being trained to predict phenotypic features across all groups of specimens, where clearly they may be more predictable for one group than for another.)

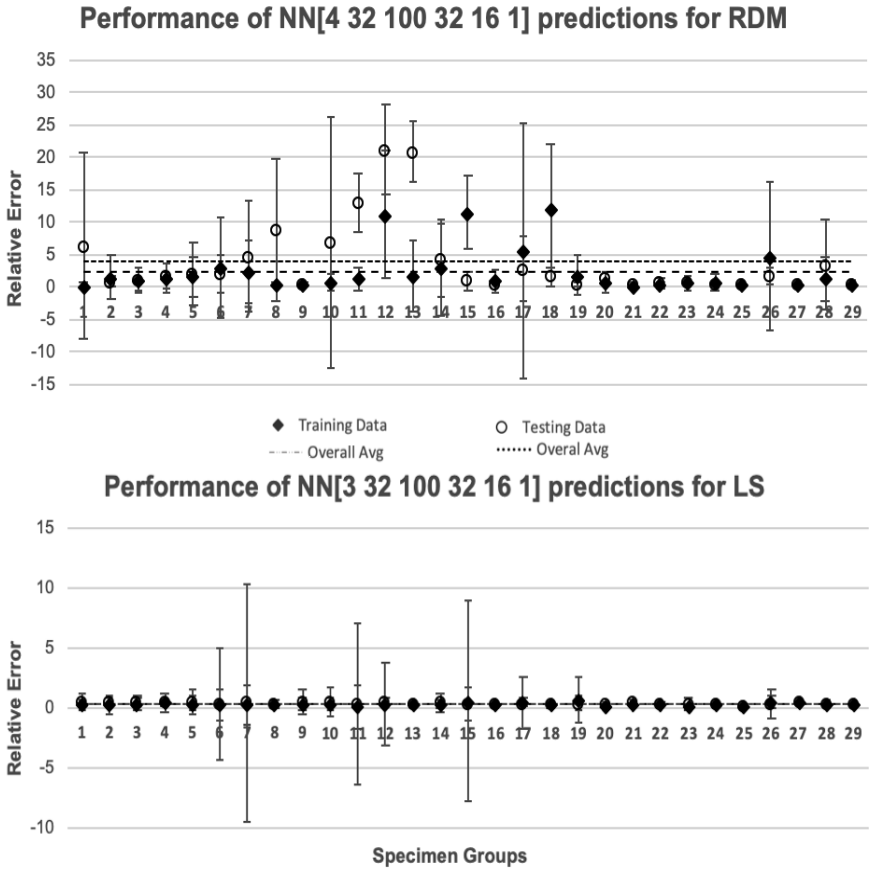

Figure 8: The performance assessment of four neural networks for predicting two phenotypic features of *A. thaliana* on nxh DNA chip 4mP3-3. Top: for rosette dry mass (RDM); Bottom: for life span (LS) on nxh DNA chip 3mE4b-2.

A similar procedure was followed for *A. thaliana* and the results are shown in Figs. 8 and 9. The choice of threshold for acceptance of model was selected to be $2 * 0.17 = 0.34$ for LS and 11.25 for RDM. As shown in Fig. 8: Bottom, the performance of deep neural networks for predicting life span (LS) on randomized data are of nearly acceptable quality with relative errors below 37% barely missing the threshold (34%). Nevertheless, for RDM on randomized data, the models show REs below 3.82, well within the threshold of $2 * \sigma = 11.25$, as shown in Fig. 8: Top. On the other hand, on nonsynthetic data, the models for both features LS and RDM are of acceptable quality, with even better scores, as shown in Fig. 9.

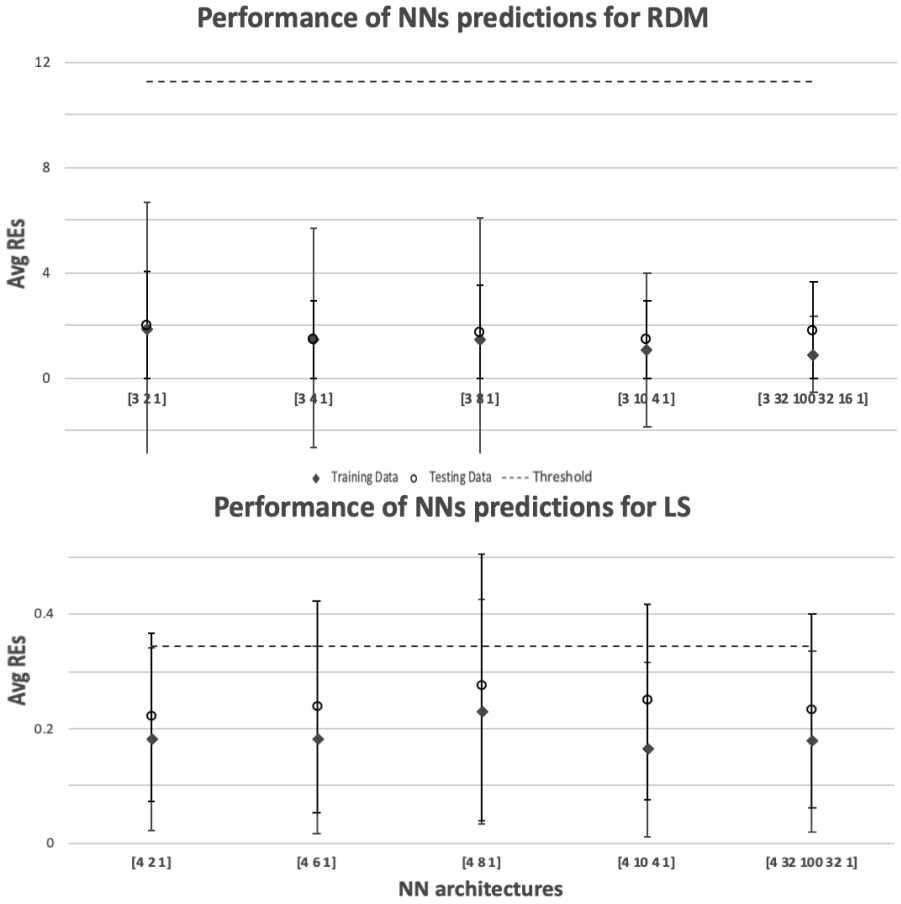

Figure 9: The performance assessment of four neural networks for predicting two phenotypic features of *A. thaliana* on nxh DNA chip 4mP3-3. Top: for rosette dry mass (RDM); Bottom: for life span (LS) on nxh DNA chip 3mE4b-2.

We were able to obtain these results using DNA sequences alone, without providing our GenISs any other information. Therefore, these results establish that a substantial component of these phenotypic features (over 75%) are at least logically inferable, if not causally determined, by genomic fragments alone, despite the fact that these phenotypic features are not 100% determined entirely by genetic traits. They suggest that it is possible to infer the genetic contribution in the determination of specific phenotypic features of a biological organism, *without recourse to the causal chain of metabolomics and proteomic events leading to them from genomic sequences*.

We demonstrated that that it is possible to make quantitative predictions about phenotypic features of two fairly distant model organisms from their genomic sequences using machine learning (deep neural network) models, for two closely related species of animal blackfly larvae (*Simulium ignescens* and *S. tunja*) and a number of strains of plant *A. thaliana*. The question arises whether predictions with these margins of error can be considered accurate enough to be significant. The fact that, environmental conditions have some influence on the phenotypic features of organisms, puts an upper bound on how accurately any model could possibly predict the features based on sequence alone. Under the assumption that the environment can affect at least 20% of the outcome, the results herein would show that the model's prediction can be as good as actually allowed by the imponderable influence of environmental factors.

Further, there are several other factors affecting the performance of the proposed models as pointed in [11]. The sequences used in this research were not the full COIs or genomes. Moreover, the nxh bases used in this research have very short probe lengths. Hence, it is possible that with longer sequences such as coding sequences in Open Reading Frames (ORFs) or whole genomes or mitochondrial genomes, and/or on bases containing longer probes, better results could be obtained. Nevertheless, for each phenotypic feature considered, a single universal neural net model was constructed to predict each phenotype for all data groups in all samples for each phenotypic feature. Hence, a model might also

perform better if it is constructed for predicting a phenotype feature for an individual species.

Habitat Prediction

Another problem in biology is to determine to what extent do genomic sequences determine environmental features present in the natural habitat of an organism. We demonstrated that some phenotypic features can be predicted using DNA sequences alone in the previous section. However, for environmental conditions, the conventional assumption is that they are too random and ephemeral to be encoded in the DNA of an organism. This section provides evidence to the contrary, that DNA does encode sufficient information about certain environmental features of an organism's habitat for a machine learning model to reveal them.

For this purpose, we define the problem precisely as follows. We are given a DNA sequence of an organism s in a certain taxonomic group (taxon) and a definition of an environmental feature in the natural habitat of s to predict quantitatively.

HABITAT PREDICTION(T,F)

INSTANCE: A DNA sequence x from an organism s in taxon T

QUESTION: Where was x grown (F = latitude and longitude)?

To show an example of how GenISs enable us to solve this problem, we use the same model organisms as in Section 3 (i.e., blackfly and *A. thaliana*.) The environmental conditions, namely latitude, longitude and temperature were selected for both organisms.

Prior Work

A solution to this problem appears to be somewhat impossible since, according to the Darwinian theory of evolution, life on earth appears to be essentially determined by the occurrence of random phenomena, such as mutations and their consequent changes given

by phenotype and environmental conditions. However, as the blueprint of life, it plays a major role in determining the phenotypic features (as demonstrated in Section 3) and metabolic behaviors [57] an organism. The relative contributions of environmental conditions in determining these behaviors of a given organism have remained a matter of debate since the discovery of DNA [10], but one cannot deny the fact that interaction between any given organism and its environment might be the outcome of rapid evolution (punctuated equilibrium) or the result of long-time evolution (gradualism) [58], [59], [60] as restricted by biological, chemical and physical conditions [61]. All these processes point towards the exceptional capacity of DNA as a memory structure of past and even present events witnessed by any given organism. Even more surprisingly, DNA molecules might be used to predict (at least with reasonable probability) the location of a species using Species Distribution Models (SDM) (hypothesizing the occurrence of species at unexplored areas on top of previously sampled areas.) This is an important advance given that field work is expensive [62], [63], not to mention that such models can become a powerful tool to predict migrations and new habitats of species in a dynamic planet, where global warming is a major new player to reckon with [64]. Thus, it is at least conceivable that DNA might also keep temporal and spatial information like a "living storage device".

Data

Three data samples were selected to assess the quality of the models, as shown in Table 7. The first sample was designed targeting *A. thaliana*. We extracted partial genomes of 83 specimens from the 1001 genome project [56] and the respective information about their habitat from the supplementary materials provided in [46]. The second sample was designed by mutating the genomic sequences from the first sample and generating synthetic observations using the same mutation rates and the similar probability distribution as discussed in Section 3. Similarly, the third sample contained mutated sequences and synthetic datapoints from 46 larvae specimens in blackfly (that were curated for [1], along

Table 7: Description of data samples for predicting environmental features consisting of DNA for *A. thaliana* and blackfly in *Simuliidae*. The field observations about environmental conditions for Simuliidae were recorded by the authors for the research reported in the source.

ID	**Organism**	**No of Specimens**	Sequences	**Source**
A83	*A. thaliana*	83	Partial genomes	[56]
A2656	*A. thaliana*	2656	Partial genomes	[56]
S46	*Simuliidae* (blackfly)	46	Partial COIs	[1]
S1216	*Simuliidae* (blackfly)	1216	Partial COIs	[1]

Table 8: Description of environmental features defining the habitat of the specimens along with the working definition used in their collection.

Features	**Sample**	**Feature Description**
Latitudes (degrees) (Standard deviation (std) on A83 = 5.58, S46 = 1.36)	A83, A2656, S1216	the angular distance of the location (where a specimen grew) north of the earth's equator
Longitudes (degrees) (Std on A83 = 16.63, S46 = 0.83)		the angular distance of the location (where a specimen grew) west of the prime meridian
Annual Mean temperature (°C) (std= 3.19)	A83, A2656	annual mean temperature of the location where a specimen grew
Temperature (°C) (std = 1.52)	S1216	temperature of the location at the time and day where a specimen of blackfly was found
Isothermal (std = 0.42)	A83, A2656	a derived feature obtained using the equation given below.
Annual Precipitation (AnPn) (mm) (std = 253.46)		annual precipitation of the location where a specimen grew

with their respective habitat information.) The sequences we used in this sample were partial Cytochrome c oxidase subunit I (COI). Table 8 shows the corresponding features of the habitat where these specimens were grown.

$$100\% * \frac{(\text{Monthly Mean(maximum temperature – minimum temperature))}}{(\text{maximum temperature warmest month – minimum temperature coldest month})}$$

Results

For a predictive model, we trained several neural networks using both genomic and pmeric signatures on 70% of the data corpus and remaining data corpus was used for the validation of the networks. These networks gave comparable predictions on the testing dataset, so we are reporting the results for the networks trained using genomic signatures only in this dissertation. The performance of these networks was assessed using their RE. A predictive model was considered to be of an acceptable quality if its average RE is less than the standard deviation of the observations about an environmental condition (i.e., latitude or longitude or temperature.) Again, we used these thresholds to assess the quality of our models because there is no prior work predicting environmental features of the natural habitat of living organisms quantitatively based on DNA sequence alone, to the best of our knowledge.

The results on *A. thaliana* are shown in Figs 10, 11, 13 and 14 and those on blackfly is shown in Fig 12.

We found that these models predicted the latitudes with average relative error about 0.09 (9%) on the training dataset and about 0.10 (10%) on the testing dataset for sample A83. These scores are below the standard deviation in the actual observations, which is about 5.58, as shown in Fig. 10 (left side.) Similar kinds of results were obtained for the networks predicting the features longitudes and annual mean temperatures, as shown in Fig. 10 (middle) and (right). In this sample, there were only 83 specimens in the sample, so it was still possible that these networks may not be robust enough, i.e. they might lack generalization ability. However, the scores shown in 11 to address this issue to an extent. The standard deviation of the predictions in longitude on both bases decreased

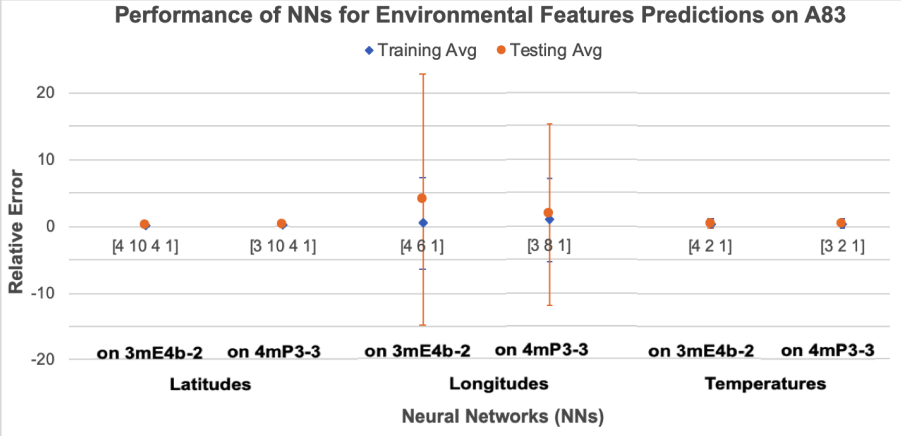

Figure 10: The performance assessment of two best performing deep networks for predicting latitudes (left), longitudes (middle) and annual mean temperature (right) of *A. thaliana* from sample A83 on two GenISs based on nxh bases. (The standard deviations between all relative errors are so small that they are hardly visible.)

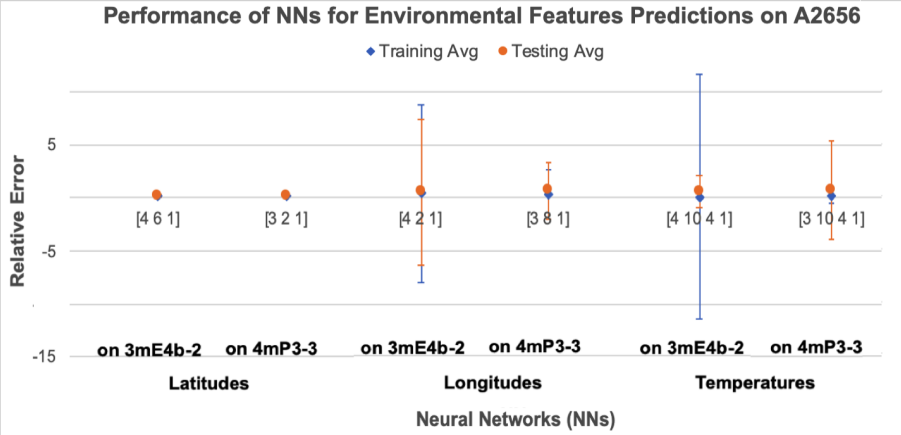

Figure 11: The performance assessment of two best performing deep networks for predicting latitudes (left), longitudes (middle) and annual mean temperature (right) of *A. thaliana* from sample A2656 in two GenISs based on nxh bases. The standard deviations between all relative errors while making predictions for latitude are so small that they are not distinguishable.

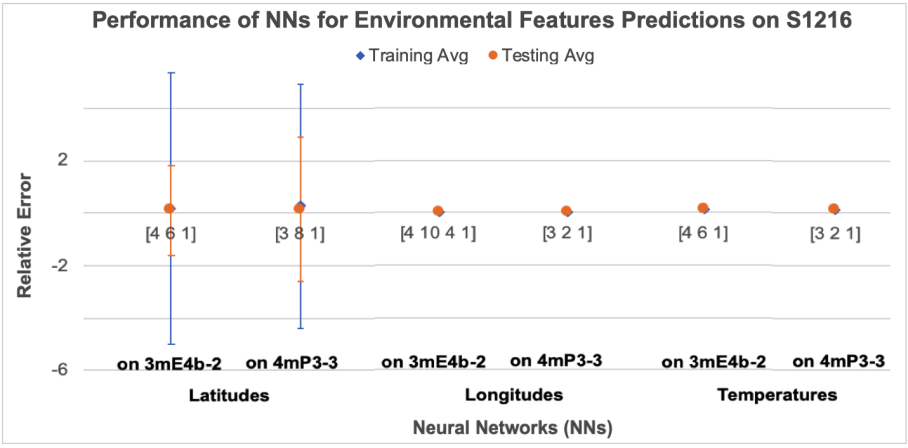

Figure 12: The performance assessment of two best performing deep networks for predicting latitude (left), longitude (middle) and annual mean temperature (right) of blackfly in sample S1216 on two nxh bases. The standard deviations between all relative errors while making predictions for longitude and temperature are so small that they are not distinguishable.

substantially on both the training and testing datasets (roughly, from 23 to 9 and from 15 to 7) while maintaining the quality of the predictions for latitude. On the other hand, the standard deviation of the predictions for temperature increased in the training phase for basis 3mE4b-2 and in the testing phase for basis 4mP3-3, although they remained about the same for the other phases.

We were also interested in trying our approach on an entirely different type of organism to gauge the scalability of our models. So, we trained the same type of model using the genomic signatures of specimens in sample S1216. The trained models were able to predict these features within an acceptable margin of relative error, as shown in Fig. 12.

The same method produces results of comparable quality for feature Isothermal, the mean of the monthly range of temperatures compared to the annual range of temperature during growth for samples A83 and A2656. Interestingly, despite efforts, we were unable to train a model to make a prediction for feature Annual Precipitation of high enough quality. Therefore, only information about certain environmental factors seems to be encoded into DNA. Nevertheless, they are somewhat informative when used as features for better predictions of other features such as longitude, as can be observed by Figs 3 and 6 (roughly

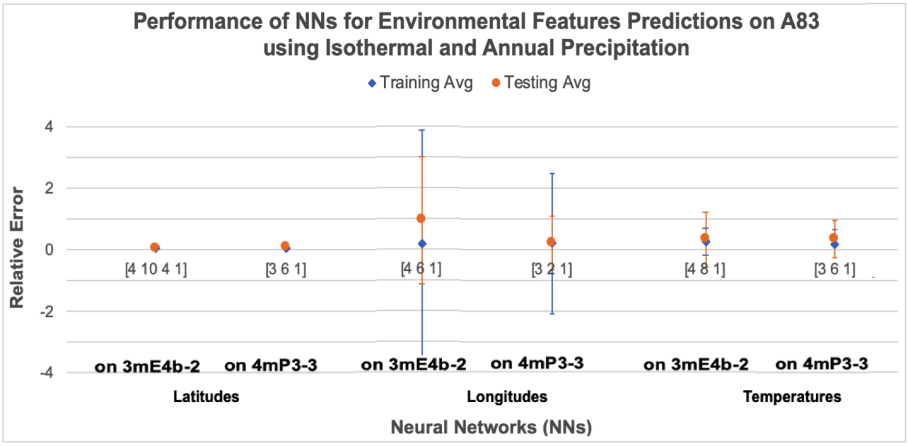

Figure 13: The performance assessment of two best performing deep networks calibrated with Isothermal and Annual Precipitation for predicting latitudes (left), longitudes (middle) and annual mean temperature (right) of *A. thaliana* from sample A83 on two GenISs based on nxh bases.

from 22 to 4 in statistically significant units) and Figs 4 and 7 (roughly from 9 to 4 in statistically significant units), where the model used Isothermal and Annual Precipitation as predictors along with genomic signatures. The corresponding improvement for latitude was not as significant, however.

We have demonstrated that, contrary to conventional wisdom that the influence of environmental conditions is too random to actually be encoded in genomic DNA, it contains enough information to make possible some determination of the environmental conditions of the habitat where an organism grows, such as location (latitude and longitude) and average temperature. Again, we were able to obtain these results using DNA sequences alone, without providing our GenISs any other information. Regardless of our effort, some features (like Annual Precipitation) could not be predicted, unless our methods miss that information. Nevertheless, we show that these features can be used to train better models to make better predictions of other environmental factors. Further, these results are consistent with the results in [65] demonstrating that DNA sequences can act as means to estimate the spatial distributions for specimens implying that there are some features that describe the environment of the specimens but are not encoded by DNA.

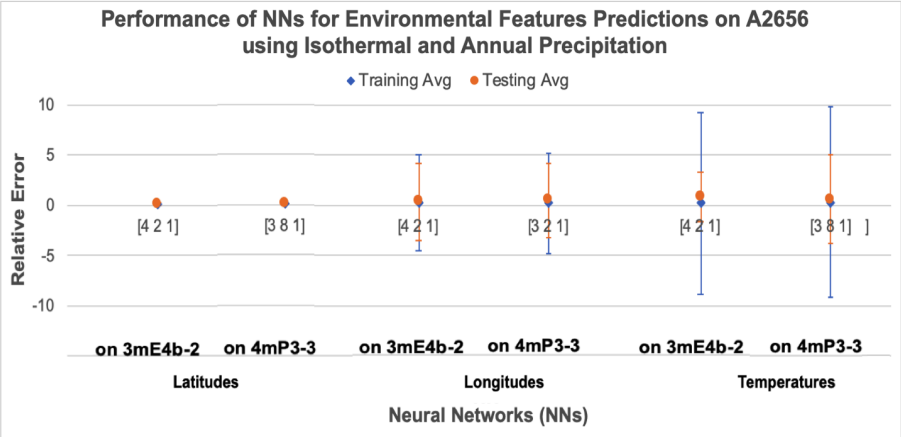

Figure 14: The performance assessment of two best performing deep networks calibrated with Isothermal and Annual Precipitation for predicting latitudes (left), longitudes (middle) and annual mean temperature (right) of *A. thaliana* from sample A2656 on two GenISs based on nxh bases.

According to Darwin's theory of natural selection, only organisms capable of adapting themselves to their environment can survive. Much later, Maturana and Varela went one step further and introduced the concept of autopoietic systems to characterize the kind of interactions that living organisms' effect internally and possibly with their environment in order to maintain themselves and reproduce (survive)[66]. Our results seem to be well aligned with this principle of autopoiesis. If so, DNA may be accumulating information regarding ecological changes somehow similar to the way the layers of soil and ice record certain environmental conditions and geological time on earth, including fossil records. Therefore, DNA may actually be, in addition to its genetic role, a living repository of information about the environment where its lineage has developed over evolutionary history.

Species Delimitation

Another classical problem in biology is the problem of defining the concept of species precisely, i.e., "What exactly is a biological species?" A solution to this problem is very important in cataloging biodiversity for wildlife preservation and new species identification.

The standard procedure relies on integrative studies focusing on geographic distribution, geological calibration, molecular characterization, ecological behavior and so on (as discussed in Section 3.) In this work, we use GenISs to provide a more objective and systematic solution to the problem based on DNA sequences alone.

For this purpose, we define the problem precisely as follows. We are given a DNA sequence of an organism s in a certain taxonomic group (taxon) T consisting of several species, and ask to identify the species in T that s belongs to.

SPECIES DELIMITATION(T)

INSTANCE: a DNA sequence x from organism s in T

QUESTION: Which species in T does s belong to?

To show an example of how GenISs enable us to solve this problem, we use three samples. For the first sample, we chose two closely related species in blackfly. For the second and third, we chose three species in *Arabidopsis* and bacteria responsible for hospital acquired infections [67] respectively. For the last sample, we chose a wide variety of organisms spread almost uniformly across three domains of life [31].

Prior Work

Ever since [68] initial proposal to standardize, organize, and rank the biome (all living organisms on earth) into a universal system (known as a taxonomy) to catalog all the biodiversity of life, biologists have discovered that delimiting species boundaries is quite a difficult task that demands years of research to get a meaningful comprehension even of a relatively small group of organisms [55]. A general definition of the concept of species is currently a challenge in theoretical biology because it is the most relevant taxonomic category in areas such as conservation, genetics, evolution and phylogenetics. More than 25 definitions have been proposed in the last century ([69]; [2]; [70]; [71]; [72]. Figure 15 illustrates the range of criteria that could be used.) As a result, some taxonomists have reached the conclusion that the ideal of establishing a single species definition applicable to

Species concept	Property(ies)
Biological	Interbreeding (natural reproduction resulting in viable and fertile offspring)
Isolation	*Intrinsic reproductive isolation (absence of interbreeding between heterospecific organisms based on intrinsic properties, as opposed to extrinsic [geographic] barriers)
Recognition	*Shared specific mate recognition or fertilization system (mechanisms by which conspecific organisms, or their gametes, recognize one another for mating and fertilization)
Ecological	*Same niche or adaptive zone (all components of the environment with which conspecific organisms interact)
Evolutionary	Unique evolutionary role, tendencies, and historical fate
(some interpretations)	*Diagnosability (qualitative, fixed difference)
Cohesion	Phenotypic cohesion (genetic or demographic exchangeability)
Phylogenetic	Heterogeneous (see next four entries)
Hennigian	Ancestor becomes extinct when lineage splits
Monophyletic	*Monophyly (consisting of an ancestor and all of its descendants; commonly inferred from possession of shared derived character states)

Figure 15: A summary in [2] alternative criteria for contemporary species concepts, extracted from a summary according to.

all (present and extinct) species that inhabit(ed) planet earth, as desirable as it may be, might be practically unattainable [73].

Conventional methods classify different organisms by grouping them into a different taxa to illustrate the degree of difference between living organisms, i.e. “species”, “genus”, “family”, “orders”, “class”, “phylum”, “kingdom” or “domain”, as in [74] or [75]. In this work, our major aim is to present evidence that it is indeed possible to create a new atlas of the biome that would encompass the vastness of biological diversity in a geometric representation that groups together biological organisms by their molecular characteristics, as follows.

Data

In order to test the soundness of this definition, we selected genomic sequences representing a number of species (as shown in Table 9), constituting four samples (first containing only

Table 9: Sample data for estimating the taxon definition of several species.

ID	Taxa	(Specimens)/Taxa	Type	Source
S20	Simuliidae	20/2 species	Partial COIs	[1]
A17	Arabidopsis	17/3 species	rRNA, mt COIs	[56]
B80	Bacteria	80/16 species	Whole Genome	[67]
BT249	Entire three domains	249/21 species	cytochrome Oxidase genes COI, COII, COIII and CytB	[31]

blackfly, second containing Arabidopsis [12], the third containing 16 species of bacteria and the last containing 21 species from [31]).

Results

We used kMeans clustering algorithm along with Voronoi diagrams to compute and visualize 2D species map as described in Section 2. An assessment of the quality of the OTU was done using the standard biological taxonomy as ground truth. The standard quantitative metrics namely, accuracy, precision, recall and f1-score were used to quantify the quality of these maps. The maps at the species level with accuracy above 92.6% were considered to be of acceptable quality because this is the average score for different methods solving similar problems [76].

The results for these samples are shown in Table 10 using standard measures of accuracy, precision, recall and F1-score. The quality appears relatively low for S20, probably because the COIs were probably too small to contain enough information about the specimens. The same procedure was applied to sample A17 consisting of 17 specimens of *Arabidopsis* distributed across three species *A. lyrate*, *A. helleri* and *A. thaliana*. Table 10 also shows the quality of the species definitions. They appear to be much better, with perfect accuracy on 3-pmers but above 0.9 overall for all metrics considered, including the F1-score. The similar type of results was obtained for samples B80 and BT249.

Table 10: Quality assessment of the maps for the OTUs of the two species in S20, three species in A17, 16 species in B80 and 21 species in BT249

n	**Sample**	**Accuracy**	**Precision**	**Recall**	**F1-score**
3-pmers	S20	0.700	0.531	0.531	0.531
	A17	1.000	1.000	1.000	1.000
	B80	1.000	1.000	1.000	1.000
	BT249	0.949	0.938	0.950	0.928
4-pmers	S20	0.500	0.438	0.406	0.405
	A17	0.941	0.952	0.933	0.937
	B80	1.000	1.000	1.000	1.000
	BT249	0.918	0.913	0.922	0.893

These results lead to an interesting question – what are the precise locations of these organisms with respect to their centroids in an Euclidean space? Unfortunately, these spaces are 4D spaces and it is very difficult for a human eye to capture the sense of their locations graphically. So, we used the basis 4mP3-3 as introduced in [3] to get their locations in 3D Euclidean space. Then, we rotated these signatures to fall onto a 2D plane. Fig. 16 shows the graphical representation of the genomic signatures of these organisms and their arrangement with respect to their centroids defining species for sample B80.

We also performed an experimental control to address a critical question - whether the choice of the full set of h-centroids is better than any other set of pmers chosen randomly? We randomly selected 32 different batches of k-pmers for each case of 3- and 4-pmers and repeated exactly the same procedure for kMeans clustering. Then, we averaged the scores for each batch consisting of 32 batches of pmers for each of the three samples. The averages for S20 and A17 are reported in Table 11. There was a huge difference with corresponding scores for the full set of h-centroids. To test the statistical significance of the difference, we ran a hypothesis z-test for each sample. In all cases, the

Table 11: Comparison of quality scores for OTUs obtained from the pmeric signatures on the full set of h-centroids and those from random sets of k-mers of the same size. The choice of h-centroids is significantly better since the p-values obtained from hypothesis tests, with the rejected null hypotheses being equality between the pairs of scores (here $C = e10$ and $E = e16$.)

n	**S20**				**A17**			
	A	**P**	**R**	**F**	**A**	**P**	**R**	**F**
3pmeric	0.700	0.531	0.531	0.531	1.000	1.000	1.000	1.000
3pm-rand	0.502	0.267	0.209	0.227	0.415	0.400	0.300	0.271
p-value	$1.51C$	$< 2.2E$	$< 2.2E$	$< 2.2E$	$< 2.2E$	$< 2.2E$	$< 2.2E$	$< 2.2E$
4pmeric	0.500	0.438	0.406	0.405	0.941	0.952	0.933	0.937
4pm-rand	0.325	0.248	0.135	0.168	0.638	0.496	0.492	0.493
p-value	1.71C	$< 2.2E$	$< 2.2E$	$< 2.2E$	$< 2.2E$	$< 2.2E$	$< 2.2E$	$< 2.2E$

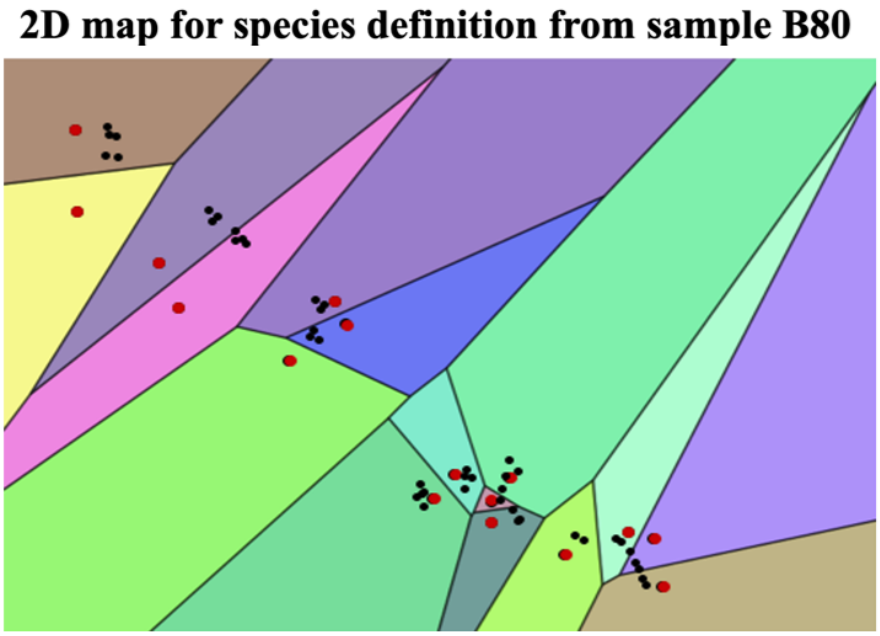

Figure 16: 2D map for species definition from sample B80 containing specimens from the domain of bacteria across 16 different genera prevalent in hospital acquired infections on the nxh basis 4mP3-3 using genomic signatures.

results of the tests confirmed that the null hypothesis (equality between the average score and to the score for full h-centroids) should be rejected. We have only reported p-values for samples S20 and A17 but p-values on all performance scores for the remaining samples on both 3 and 4-pmers were less than $2.2e16$.

We have presented an alternative coordinate system, pmeric coordinates, to a known GenIS in [3] based on genomic signatures (obtained using DNA sequences alone), again exploiting other structural properties of DNA spaces. Further steps to bring this program to fruition include at least two important choices. First, the specific selection of a common choice of genomic sequence for all organisms. Second,

Finally, regardless of the choices, it is truly remarkable that such as simple-minded definition of a species, purely based on geometric distance of the feature vectors from the centroids afforded by pmeric coordinates, could capture so much of the complexity of the taxon, as given by the standard biological classification. In the final analysis, this fact illustrates the power of the deep structure in our selection of nxh bases upon which features these predictions are based.

A Computational Approach to Pathogenicity

Another classical problem in biology is presented by the concept of a *pathogen*, usually meant to humans. The problem is of high significance in the fields of medical pathology and immunology. The global COVID-19 pandemic has raised the urgency for methodologies to predict new strains for potential pathogenicity in short time to control or mitigate the spread of a disease in its early phases [77]. The standard procedure involves the decisions about the pathogenicity of microbes are currently made from harm they have caused in hosts (e.g. sickness or death [78, 79].)

A recent important development in the evolution of the concept is that an absolute

notion of pathogen is not really meaningful, it is rather about a relationship to another organism (a host), where the former is responsible for causing disturbances in the homeostasis of the latter [79].) Thus, the problem really becomes to provide a general, objective and operational definition of the concept. Hence, we rather define the problem precisely as follows.

PATHOGENICITY

INSTANCE: Two DNA sequences representing a host (H) and a microbe (P)

QUESTION: Is P pathogenic to H?

An example of how GenISs enable us to solve this problem can be given with bacteria and fungi if we obtain proxies for microbes and *homo sapiens* for a host. We chose bacteria because they are abundant and are well researched in the pathogen literature, and also fungi, because the structure of their cells is much more complex than bacteria and hence require a more rigorous analysis to decide about their pathogenicity [80, 81, 77, 82, 83].

Further, the problem of PATHOGENICITY does not really make sense unless we characterize the concept of a 'pathogenic relationship'. There is little in the literature about such a definition. To get started, we can characterize it as follows. A specimen P has a *pathogenic relationship with a species H over a given period of time* if and only if

- P interacts with any specimen in H and begins to reproduce;
- H produces a defense in response to counteract the resulting colony of Ps;
- Ps may push back, and H may counteract, until H reaches a stable condition that may be different from the condition prior to interaction with P;
- All three conditions remain true with at least 32 other specimens in H, in the absence of any other such P*.

There are situations where two pathogens can attack the host simultaneously and may be successful jointly, but not individually. Therefore, a general definition should allow for

multi-way pathogenic relationships. However, in this first attempt towards a general definition of pathogenic relationship, we will simply assume the relationship to be binary.

Prior Work

The term pathogen (borrowed from Greek; *pathos* meaning disease and *genos* meaning kind, race, family, birth or origin) has been in use since the 1880s to refer to infectious microorganisms including virus, bacterium, protozoan, prion, viroid and fungus [78], [79]. The study of diseases caused by these organisms, pathogens, does not have a distinctive root but could be traced back to the documentation of disease with Egyptian medicine, for instance, Edwin Smith Papyrus (17th century BC) and Papyrus Ebers (about 1550 BC.) Since then, a number of characterizations have been proposed to give a general overview of what pathogens are. Earlier views were primarily based on microorganisms and their intrinsic properties only, although it was also known that pathogenicity was neither invariant nor absolute [84]. In early 20th century, Bail proposed aggressins and Rosenow proposed virulins as microbial products ushering pathogens themselves into the host [85]. [86] pointed out that pathogens were deemed to have "offensive" and "defensive" functions separating themselves from nonpathogenic microbes and determining the type and outcome of the host-pathogen interaction. Later, in 1914, Zinsser grouped microorganisms into three different categories i.e., a) saprophytes that were unable to establish themselves in living tissue; b) pure parasites that were able to establish themselves easily in normal hosts; and c) half parasites having low invasive power and causing infection only in certain circumstances [85]. Similarly, Watson and Brandly noted that the term pathogenicity was used for defining the degree of involvement for microbes that did not cause rapidly fatal infections [87]. Later, in 1990s, the definition of pathogens had solely focused on the ability to cause disease. The chemistry of the microbial surface is an critical ingredient in the ability of a microorganism to cause disease. Similarly, Falkwo proposed "Molecular Koch's Postulates" as a conceptual framework to identify the genes causing diseases [88] and also

noted that a pathogen has an intrinsic ability to breach cell barriers of a host [89]. These several definitions were reviewed in [79] and are summarized next.

- A microbe capable of causing disease [82], [90]
- A microorganism that can increase in living tissue and produce disease [91]
- Any microorganism whose survival is dependent upon its capacity to replicate and persist on or within another species by actively breaching or destroying a cellular or humoral host barrier that ordinarily restricts or inhibits other microorganisms [89]
- A parasite capable of causing or producing some disturbance in the host [92]

These characterizations cannot really be regarded as logically satisfactory general definitions because they still rely on "someone dying/getting sick" to prove that some microbe is pathogenic. In addition, all these approaches place the ability to cause diseases solely on a microorganism, regardless of the affected host. On the other hand, it is evident that the ability of some microorganisms to cause disease depends on a specific host. For example, an influenza or covid-19 virus may cause death in some hosts but no effect whatsoever in others. For this reason, recent studies are shifting their focus on two-way relationship between host and microorganisms in the form of pathogenicity to ditch the term "pathogens" [84]. Furthermore, all these research point towards a single conclusion, that although there are several approaches to make a distinction between pathogens and nonpathogens, there is no general and operational and acceptable definition of pathogens.

Data

Two samples were collected to obtain datasets for the assessment of the proposed definition below. The first sample was designed to contain 107 pathogenic and 109 nonpathogenic bacteria for *Homo sapiens* in general. A similar selection was made for the second sample for fungi (25 pathogens and 25 nonpathogens.) Once these datasets were designed, we

downloaded coding sequences of whole genome for bacteria and coding sequences of mitochondrial genome for fungi. We also downloaded four mitochondrial genes (COI, COII, COIII, CytB) for homo sapiens. We computed genomic and pmeric signatures of these sequences, but are reporting the scores for genomic GenISs only because they yield the better scores.

Results

The results for the quantitative assessments of the ML models are shown in Fig. 17. A solution model is being deemed acceptable only if its specificity is at least 80% and its sensitivity 70%, which are in the order of the average scores for similar problems in the literature [80], [93], [94], [95].)

We trained all machine learning models discussed in Section 2 using DNA sequences alone. We report the performance scores of four models trained on the features obtained using optimal combination of nxh bases. For bacteria, kNN, RBF and MLP models produced perfect scores for all performance metrics on the basis 3mE4b-2. All scores were above 80%, but ML models DT and AB yielded the best scores. We assessed the performance of our GenISs combining the two datasets of bacteria and fungi. We were able to obtain the machine learning models, namely RBF, DR and AB with 100% sensitivity (as shown in Fig. 17), which is very good for diagnostic methods in pathology and immunology. However, only GenIS using AB was able to yield all scores above 80%. Thus, some models are capable of leveraging patterns from the information extracted by genomic signatures in our GenISs to provide a general and uniform definition of pathogenicity (regardless of the choice of the taxonomic group of a microorganism) based on molecular data only. These definitions could be used as alternative *hypotheses* estimating the degree of pathogenicity between microbes (e.g., bacteria and fungi) and a host (e.g., humans.)

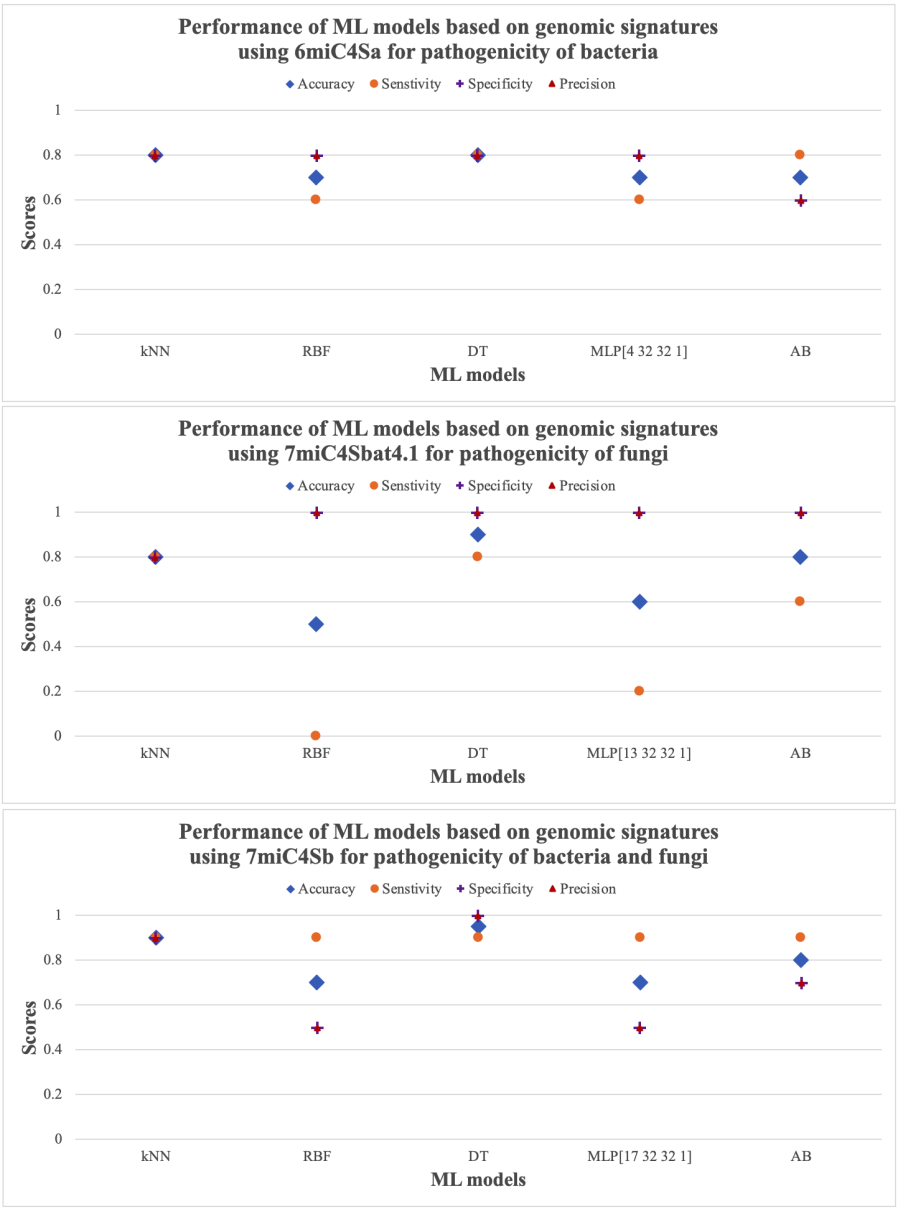

Figure 17: The performance assessment of the definition of pathogenicity of bacteria (top), fungi (middle) and both (bottom) obtained using machine learning models trained on genomic signatures. Interestingly, when both bacteria and fungi are combined, DT gave the best scores (accuracy = 0.95, sensitivity = 0.9, specificity = 1 and precision = 1.)

Early genetic studies made obvious [57] that DNA is responsible for the preservation of phenotypes from parents (specifying their structure and function) onto their offspring. Further studies have probed into the role of epigenetics in disease [96] and various life processes for the prediction of psychological disorders (e.g., depression, schizophrenia and so on [97].) A common denominator in these studies is a deeper analysis of the causal chain of events that leads from the sequences to the metabolomics and proteomics that constitute the phenotype. By contrast, our GenISs *only* require DNA inputs and bypass the entire metabolomic and proteomic analyses of the events leading to the predictions and results. To our knowledge, no approach has been proposed to predict quantitative phenotypic or environmental features concerning a living organism, let alone a versatile and universal platform to obtain such results.

CHAPTER 4

Abiotic Applications of GenISs

This chapter shows how the same GenISs used in the previous chapter may just as effectively be used to solve problems with abiotic data with some fairly straightforward encodings of the ordinary data inputs into DNA. We illustrate with two important problems in computer science, namely Malware Classification (MC) and Image Segmentation (IS).

Malware Classification

A classical problem in computer science is to identify the type of a malware, i.e., a Malware Classification problem. A solution to this problem is very critical in cyber security, where new malware is emerging at an alarming rate (for instance, more than 4.62 million new instances of malicious code were detected from June 2019 to July 2019 [98].)

The problem is defined precisely as follows. It assumes a partition of all malwares into a finite number of distinct categories (or classes) C is given in advance.

MALWARE CLASSIFICATION(C)

INSTANCE: A piece of malware x of classes available in M

QUESTION: Which type/category in C does x belong to?

To show an example of how GenISs enable a solution this problem, we use dataset provided by Microsoft for the Malware Challenge [99].

Prior Work

Most real-life data, such as text and images, are unstructured in the sense that they are not generated by a well-defined model or even organized in tabular format. Advances in modern internet technologies have enabled us to generate and/or store huge amounts of

data (in the scale of exabytes per day) but we still lack methods to analyze them in the same scale at speed. A major roadblock is this lack of structure to select or extract useful critical information to solve problems at humanly meaningful (semantic) scale. Deep learning methods have handled these difficulties to an extent, but it is very difficult to explain how these methods learn significant features for such data analytics.

The emergence of malwares at an alarming rate has created a serious threat in cyber security. A rigorous analysis to this problem is critical to study the evolution of the malware and tracing cybercrime. Such an analysis can be either static or dynamic. A dynamic analysis depends on the execution of malwares in a controlled environment [100], [101] and is costly effortwise [102]. On the other hand, a static analysis relies on decompilation tools (like IDA Pro) and is more effective and efficient [103, 104]. However, static analysis suffers from a major information retrieval issue since information in the source code could be lost in the compilation process. Furthermore, encryption and obfuscation techniques can easily forfeit solutions to these issues [98].

Many works have been proposed to avoid these drawbacks. For instance, [105] proposed a method based on byte n-grams as features to train gradient-boosting decision tree classifiers for malicious code. This approach is widely used to solve the problem, but it ignores the fact that not all static and dynamic attributes are related to each other [106]. There have been many more attempts to solve this problem with the advent of machine and deep learning methods. For example, [107] proposed four kinds of LSTM-based deep networks based on input opcode sequences as features for training and testing. Later on, [108] combined these opcode sequences with grayscale images of malware binary files and used CNN and LSTM networks to learn features. Similarly, [109] proposed a new image-based malware classification using a CNN architecture that showed higher accuracy when compared with traditional ML models. These findings suggest that some kind of encoding of these malwares by an automated process might lead to better performance in feature extraction rather than using a manual process.

Table 12: Summary of the dataset for Microsoft's Malware Classification Challenge

Malware Class	Original size	Size used	Type of Malware
Ramnit	1,541	66	Worm
Lollipop	2,478	127	Adware
Kelihos_ver3	2,942	164	Backdoor
Vundo	475	5	Trojan
Simda	42	1	Backdoor
Tracur	751	33	TrojanDownloader
Kelihos_ver1	398	8	Backdoor
Obfuscator.ACY	1,228	62	Any kind of obfuscated malware
Gatak	1,013	45	Backdoor

Data and DNA encodings

We first encode a piece of malware into a DNA sequence. A malware program can be regarded as an ordered set of hexcodes representing a string of hexadecimal characters. Thus, a DNA encoding can be simply obtained by converting each hexadecimal character to its binary equivalent, and then concatenating these binaries into a sequence in DNA form. Two bits encode four possible strings that can be regarded as a,c,g, or t, i.e.,

$$00 \to a;\ 01 \to c;\ 10 \to g;\ 11 \to\ t.$$

We used the Microsoft Malware Classification Challenge [99] dataset that was released in 2015 and is publicly available through Kaggle. The dataset is summarized in Table 12.

Results

We computed the genomic and pmeric signatures of these straightforward malware encodings and trained and tested some conventional ML models. We assessed the quality of these models using the standard metrics, namely accuracy, precision, recall and F1-score but are reporting the performance of the models giving top two scores. A model was

Table 13: Comparing performance of our GenISs with Common Machine learning Models and similar work as the state-of-the-art methods.

Group	**Method**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
GenIS-8pmc	RF	95.83	95.75	96.62
	KNN	95.41	95.24	96.38
GenIS-3pmc	RF	94.46	94.17	95.62
	KNN	93.97	93.33	95.24
GenIS-8mP10	RF	73.22	74.79	73.34
	KNN	71.08	70.85	69.86
GenIS-3mE4b	RF	99.33	99.05	99.13
	KNN	94.42	95.19	94.11
Related Work	Kaggle Winner [110]	99.63	99.07	99.35
	SNNMAC [98]	99.21	99.18	99.19
	MalNet [108]	99.14	97.96	98.55
	Hanqi Zhang [111]	92.13	90.64	91.38
Common Machine Learning Models[98]	Random Forest	84.46	82.34	83.38
	Xgboost	85.13	72.02	78.02
	Naive Bayes	70.21	70.06	70.13
	Logistic Regression	71.42	67.38	69.34
	Support Vector Machine	54.84	28.75	37.72

considered to be of an acceptable quality if its performance (accuracy / F1-) score is within 1 standard deviation from Kaggle Winner (the standard deviation was computed among each set of performance scores for other methods in literature in Table 13.) A comparison of our solutions to other solutions available in the literature is shown in Table 13.

From 13, the performance scores of GenIS based on genomic signature using 3mE4b are almost equal to the ones of Kaggle Winner. All performance scores (except for the GenIS based on genomic signature using 8mP10) are greater than the difference between the corresponding performance score of the Kaggle Winner and the standard deviation of the scores from the literature (i.e., the threshold for $precision = 83.93$, $recall = 76.42$ and $F1 - score = 79.27$.) Therefore, these models are of more than acceptable quality. In fact,

the results for MC are competitive with the performance of Kaggle winner of the Microsoft MC challenge.

We have presented two Genomic Information Systems (GenISs) leveraging structural properties of DNA spaces as an alternative solution to the malware classification problem. Most of our GenISs produced comparable results with the benchmark models for malware classification problem with the scores of greater than 95% for precision, recall and F1-score. Most of these deep networks require at least a few hours for training. However, our GenISs can produce these results in the order of minutes. Moreover, these results provide strong evidence that DNA can be very helpful in text analysis of malware at a deep level, with very low dimensional vectors and performance when other methods require at least hundreds of features to do.

Image Segmentation

Another classical problem in computer science is to find a partition of an image into multiple segments identifying semantically meaningful whole objects present in the image. A solution to this problem is very critical in applications for autonomous driving, medical image analysis, health care and so on. The standard procedure relies on the annotation of images and the use of deep learning models (the majority of the solutions are obtained by supervised learning approach.) In this work, we use GenISs based on unsupervised learning approaches to produce explainable results.

For this purpose, we define the problem precisely as follows. We are given an image.

IMAGE SEGMENTATION(x)

INSTANCE: A 2D image x

QUESTION: What is a partition of x into recognizable objects?

To show an example of how GenISs enable us to solve this problem, we use two benchmark datasets, i.e., CamVid and KITTI datasets.

Prior Work

Recent deep learning solutions solve this problem by learning significant patterns from the annotations of large amounts of pixels in images, made possible by the introduction of Convolutional Neural Networks for feature learning and large dataset annotation. Recent literature [112] argues that the performance of these trained models are highly sensitive to the quality of the annotation, which is taxing time and effortwise, not to mention subjective. For instance, it takes an average of 1.5 hours to annotate all pixels in an image of size $1024 * 2048$ in the Cityscapes dataset [113]. Further, some datasets are solely based on images captured from continuous video frame sequences at regular time intervals. Therefore, several works [114, 115, 116, 117, 118] have leveraged temporal constraints to propagate ground truth labels from labeled to unlabeled frame using two major approaches, namely optical flow [114, 115] and patch matching [116, 117, 118]. Patch matching methods are generally sensitive to patch size and threshold values and occasionally assume *a priori* knowledge of class statistics. On the other hand, optical flow methods rely on very accurate optical flow estimation. More recent works (e.g., [112]) use motion vectors from video prediction models to obtain such propagation and the learned vectors to handle occlusion.

Another approach to image (semantic) segmentation is to incorporate edge cues as constraints to handle boundary pixels [119, 120]. However, these methods might propagate error from edge estimation or lead to over-fitting due to extremely hard boundary cases. In order to resolve these issues, other methods have been attempted, such as affinity field [121], random walk [122], relaxation labelling [123]. The problem with these alternatives is that, instead of handling boundary pixels directly, they attempt to emulate the interactions between segments and object boundaries [112]. Another issue with all

these models is that they impose an additional burden to annotate images for training and testing purposes. Worse yet, that entails a human in the loop.

A common denominator among all these models is supervised learning algorithms. This suggests that the use of unsupervised learning algorithms like Self Organizing Maps (SOMs) may avoid these burdens. The approach has been explored with several variants. For instance, [124] proposed a network called a Local Adaptive Receptive Field Self Organizing Map (LARFSOM) for color image segmentation. Later, [125] integrated SOM with kMeans and saliency map to perform image segmentation via clustering. Most of these algorithms are solely based on SOM derived prototype parameters [126, 127, 128]. Much later, [129] integrated SOMs with extended fuzzy clustering. On average, a performance of 84.5% of sensitivity and 88.1% of accuracy was reported. Furthermore, qualitative evaluation also indicated improvement in the performance.

Data and pmeric encodings

We first calculate the pmeric coordinates of the full set of centroids plus one point in the north casket (we use *aaat*.) Second, we compute the convex hull of these numerical vectors in the corresponding Euclidean spaces to produce an encoding region for arbitrary images. Each digital image can now be represented by a set of pixels. Each pixel constitutes a 3D vector containing values for a Red, Green and Blue colors (R, G, B) combination (each value ranging between 0 and 255.) We used a 4D vector containing these color values and the average of these values (i.e. (R, G, B, avg(R, G, B)) to represent such a pixel. Then the pixel vectors are mapped by an affine transformation to points inside the convex hull of the encoding region. As shown in Fig 18, these points can then be represented by their barycentric coordinates. Each vector contains the pmeric coordinates of the original image. They are then used as feature vectors to a SOM that will segment the image with pmeric encodings.

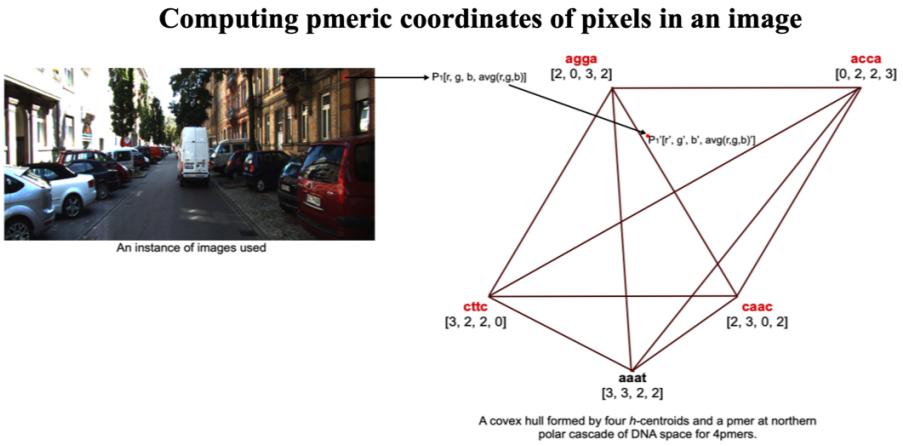

Figure 18: Workflow to compute pmeric coordinates encoding an image in a DNA space of 4-pmers. A pixel in an image can be represented by a 4D vector containing RGB values and an average of these three values. Then, the 4D vector can be mapped to a point in the convex hull spanned by the pmeric coordinates of the four h-centroids and $aaat$ as corners. The raw pixel vectors are then the mapped point to the convex hull. The process is repeated for all pixels in the image to obtain an encoding into 4D feature vectors using frequencies as weights

Table 14: Description of data samples for image segmentation problem.

Name	**Resolution of the images**	**Size (training, testing and validation)**	**Source**
CamVid	$720 * 960$	701 (367, 101, 233)	[130]
KITTI	$375 * 1242$	400 (200, 200, 0)	[131]

We used CamVid and KITTI datasets as samples to assess the quality of our solution (the datasets are described in Table 14 below.)

Results

We used SOMs (as described in Section 2) as solution models for this problem. Both quantitative and qualitative analyses were done to assess the quality of the proposed solutions. The qualitative analysis for two instances on CamVid and KITTI datasets is shown in Figs. 19, 20. We used mean Intersection Over Union (IOU) for quantitative analysis. We also performed an experimental control on the CamVid dataset an alternative assessment of the performance of the proposed solutions. The comparison of our solutions to the others in the literature based on mean IOU are shown in Tables 15 and 16.

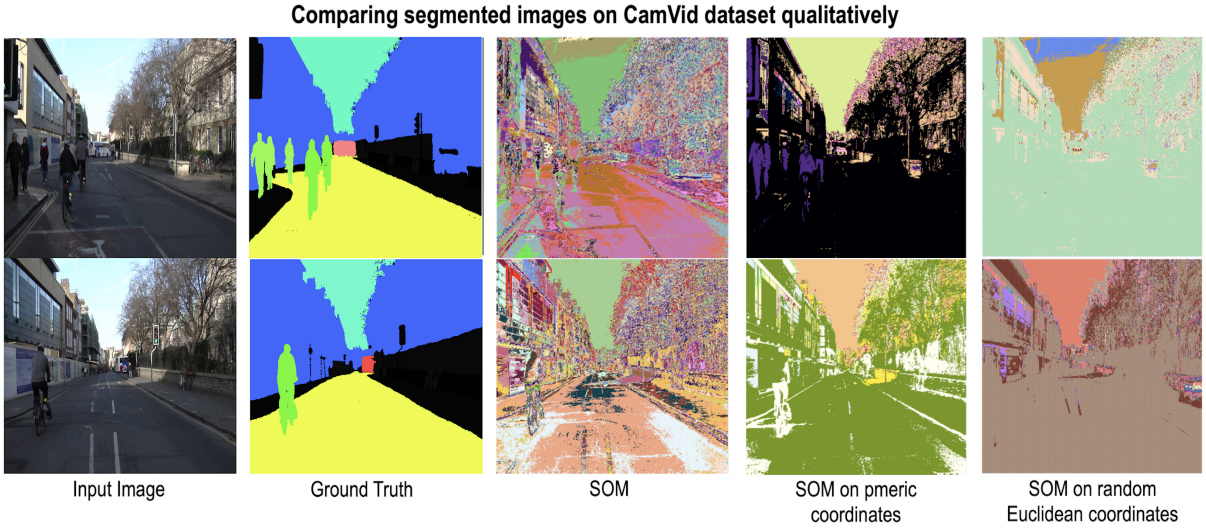

Figure 19: Typical qualitative performance of two datapoints (rows) in the CamVid dataset of three solutions, without (third column) and with pmeric (fourth column) and random (fifth column) encodings. Compared to the ground truth (second column), the quality of clustering by plain SOMs is poor (too many clusters), while that on the pmeric encodings is evidently better (semantically more meaningful fewer clusters.) On the other hand, the clustering based on random encodings is evidently not semantically meaningful (too fewer clusters to identify objects in the original image)

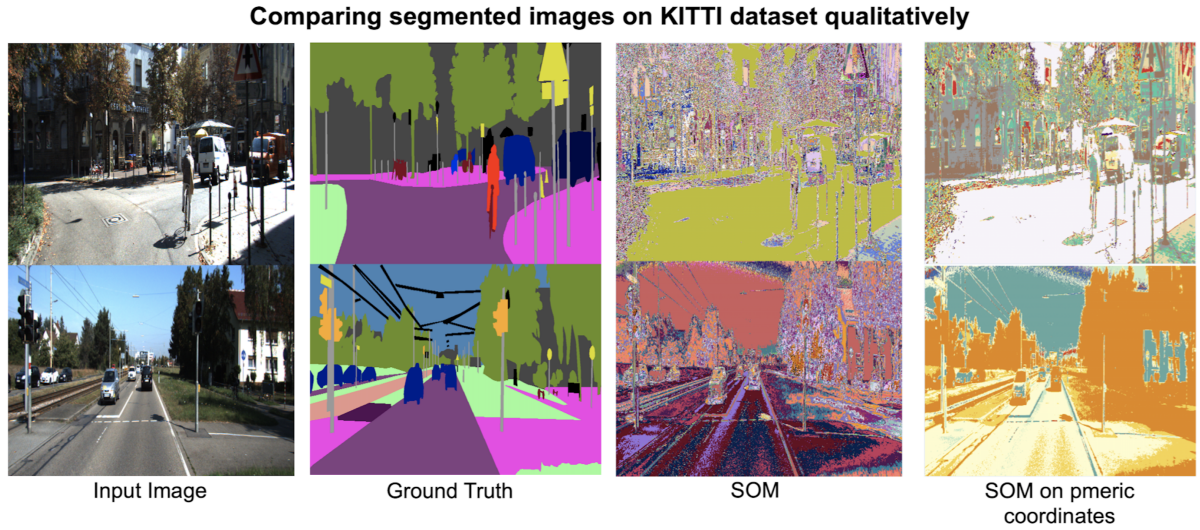

Figure 20: Typical qualitative performance of two datapoints (rows) in the KITTI dataset of two solutions, without (third column) and with pmeric (fourth column) encodings. The results are similar to those in the CamVid dataset (see Fig. 19), although the quality of segmentation on the pmeric encoding is not as good as for the CamVid dataset compared to the ground truth (second column)

Table 15: Performance of solutions to image segmentation on the CamVid dataset

Method	**Encoder**	**mean IOU (%)**
GenIS (4pmc)	-	90.4
SOM	-	85.5
VPred+LP [112]	WideResNet38	82.9
VideoGCRF [132]	ResNet101	75.2
BiSeNet [133]	ResNet18	68.7
Dilate8 [134]	Dilate	65.3
RTA [119]	VGG16	62.5
SegNet [135]	VGG16	60.1
Control	-	3.3E-04

Table 16: Performance of solutions to image segmentation on the KITTI dataset

Method	**mean IOU (%)**
GenIS (4pmc)	84.90
SOM	83.67
VPred+LP [112]	72.83
LDN2 [136]	63.51
AHiSS [137]	61.24
MapilaryAI [138]	69.56
SegStereo [139]	59.10
APMoE$_seg$ [140]	47.96

As shown in Tables 15 and 16, the average IOU scores show a huge improvement on results given by our GenISs when compared with the *state-of-the-art* methods. In order to test the statistical significance of this difference in performance, we ran a z test on both datasets with the null hypothesis being, "There is no significant difference in average IOU scores given by SOM trained on raw pixels with that trained on encoded pixels". The test was performed on all 101 images from CamVid dataset. We computed z-value to be 3.804 which is greater than the critical z-value (1.96). Therefore, the null hypothesis was rejected and there is enough evidence in the dataset to prove that the difference is statistically significant. But the data did not have sufficient evidence to support our claim of this difference being significant as we computed z-value to be 0.5804 (which is lesser than 1.96) on KITTI dataset.

Similarly, as shown in Figs. 19 and 20, with the same architecture, SOMs trained on raw pixels tend to group these pixels in too many clusters producing noise in the segmentation for both datasets. However, mapping to a DNA space helps reducing the number of unwanted clusters and the resulting segmented images are more refined. Therefore, although we could not find statistical evidence, these differences have some impact semantically on real images for the KITTI dataset.

At last, an experimental control on CamVid dataset demonstrates that pmeric coordinates make a significant contribution in construing the semantics in images, as illustrated in Fig. 19. The segmented images do not contain enough clusters to properly distinguish objects/semantics. This was also reflected in average IOU as shown in Table 15.

We have presented pmeric GenISs leveraging the structural properties of DNA spaces for ordinary image processing. Current solutions rely on pixels to segment different regions in images. In reality, pixel intensities of objects in an image do not contain full information to do so. For example, an object/segment in an image might interfere with another (e.g. a building might cast a shadow on the road causing a difference in the pixel intensities of the different parts of the road) and lead to misclustering. The performance of our proposed GenIS for image segmentation is comparable (if not better) than the state-of-the-art deep learning models and SOM with the average IOU score of about 90% on CamVid and about 84% on KITTI dataset. Our methods seem to be able to address these issue with the semantic segmentation problem (SIS) since a control run with random encodings produce results of significantly lower quality.

CHAPTER 5

Conclusion and Future Work

This dissertation presents a novel approach to designing Genomic Information Systems (GenISs) aimed at solving some challenging problems in biology and in computer science at large. They are based on novel coordinate systems for DNA sequences (genomic and pmeric) that leverage fundamental research on the deep metric structure of DNA oligomers up to 12-pmers. These GenISs offer multiple advantages over the current tools and techniques used in data science to handle large biodata sets. First, these GenISs allow extraction of few very informative features from long DNA sequences anchored in biological reality that could be used to train machine learning models on large data sets (e.g., whole genomes of humans.) Second, recent advancements in data science, especially Convolutional and Deep networks, have produced remarkable results in analyzing DNA sequences (for example in predicting EC number on enzyme function prediction, motif recognition in Polydentilation in RNA maturation and promoter recognition of DNA fragments in the human genome), but these networks require a lot of computational and processing time to get better results. Further, they provide black-box answers lacking explanations that would allow a human to rationalize decisions or answers. By contrast, we have demonstrated that GenISs can process biotic data efficiently and furthermore, they can be extended to process abiotic data just as effectively and efficiently. Third, in a nutshell, all these findings reveal that *DNA molecules are capable of encoding a great variety of information beyond what is currently known (e.g., of the habitat where a living organisms grew or lived), not only for living organisms, but also for ordinary abiotic data, and that this information has a structure that can be decoded through well-known methods in data science and machine learning.*

This line of research opens up several possibilities for the further exploration. First, these GenISs offer a transformation of DNA sequences of higher dimensions (e.g., up to

millions of base pairs as in the case of whole genome, mitochondrial genome and malware features) into exponentially lower and very few dimensions, essentially preserving most significant information. Upon reflection, these reductions appear hard to believe. Therefore, these GenISs could be further explored as new dimensionality reduction in other domains. How scalable are these methods to other domains? Are they close to a theoretical limit?

Further, a second possibility involves domain knowledge. Recent trends in machine and deep learning methods are beginning to explore the role of domain knowledge integration in the quality of a solution model. The traditional view of domain knowledge integration requires an expert to provide knowledge of the field to select or extract features to produce or improve good solution models. This process is thus manual and subjective, e.g., they are subject to biases. These biases could be avoided to an extent by integrating the knowledge of several experts. However, such an integration does not guarantee to removal of such biases 100%. On the other hand, GenISs are based on hybridization affinity, an objective property of DNA. Therefore, GenISs provide a means to meaningfully compare the inherent difficulty of problems and the performance of their solutions across domains, currently a challenge in data science and machine learning.

Finally, this research also points to the possibility of that DNA has a built-in capability for unsupervised learning, e.g. in the choice of a convex hull for pmeric coordinate to encode images for semantic clustering. A further exploration of deeper geometric properties (using h-distance or more accurate refinements) might lead to deeper connections to the structure of ordinary Euclidean spaces built into DNA. Such connections might lead to a next generation of machine learning methods.

Bibliography

[1] F. A. Colorado-Garzón, P. H. Adler, L. F. García, P. Muñoz de Hoyos, M. L. Bueno, and N. E. Matta, "Estimating diversity of black flies in the simulium ignescens and simulium tunja complexes in colombia: chromosomal rearrangements as the core of integrative taxonomy," *Journal of Heredity*, vol. 108, no. 1, pp. 12–24, 2017.

[2] K. De Queiroz, "Species concepts and species delimitation," *Systematic biology*, vol. 56, no. 6, pp. 879–886, 2007.

[3] M. H. Garzon and S. Mainali, "Towards a universal genomic positioning system: phylogenetics and species identification," in *International Conference on Bioinformatics and Biomedical Engineering*. Springer, 2017, pp. 469–479.

[4] S. Mainali, F. A. Colorado-Garzon, and M. Garzon, "Foretelling the phenotype of a genomic sequence," *IEEE/ACM transactions on computational biology and bioinformatics*, 2021.

[5] M. Garzon, P. Neathery, R. Deaton, R. C. Murphy, D. R. Franceschetti, and S. Stevens Jr, "A new metric for dna computing," in *Proceedings of the 2nd Genetic Programming Conference*, vol. 32, no. 1. Morgan Kaufman, 1997, pp. 636–638.

[6] L. M. Adleman, "Molecular computation of solutions to combinatorial problems," *Science*, vol. 266, no. 5187, pp. 1021–1024, 1994.

[7] N. C. Seeman, "Dna in a material world," *Nature*, vol. 421, no. 6921, pp. 427–431, 2003.

[8] E. Winfree, F. Liu, L. A. Wenzler, and N. C. Seeman, "Design and self-assembly of two-dimensional dna crystals," *Nature*, vol. 394, no. 6693, pp. 539–544, 1998.

[9] M. H. Garzon and K. C. Bobba, "A geometric approach to gibbs energy landscapes and optimal dna codeword design," in *International Workshop on DNA-Based Computers*. Springer, 2012, pp. 73–85.

[10] J. D. Watson and F. H. Crick, "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[11] S. Mainali, M. H. Garzon, and F. A. Colorado, "Profiling environmental conditions from dna," in *International Work-Conference on Bioinformatics and Biomedical Engineering*. Springer, 2020, pp. 647–658.

[12] ——, "New Genomic Information Systems (GenISs): Species Delimitation and IDentification," in *International Work-Conference on Bioinformatics and Biomedical Engineering*. Springer, 2020, pp. 163–174.

[13] J. G. Wetmur, "Dna probes: applications of the principles of nucleic acid hybridization," *Critical reviews in biochemistry and molecular biology*, vol. 26, no. 3-4, pp. 227–259, 1991.

[14] M. Arita and S. Kobayashi, "Dna sequence design using templates," *New Generation Computing*, vol. 20, no. 3, pp. 263–277, 2002.

[15] A. Ben-Dor, R. Karp, B. Schwikowski, and Z. Yakhini, "Universal dna tag systems: a combinatorial design scheme," *Journal of computational Biology*, vol. 7, no. 3-4, pp. 503–519, 2000.

[16] R. Deaton, M. Garzon, R. Murphy, J. Rose, D. Franceschetti, and S. E. Stevens Jr, "Reliability and efficiency of a dna-based computation," *Physical Review Letters*, vol. 80, no. 2, p. 417, 1998.

[17] A. G. Frutos, Q. Liu, A. J. Thiel, A. M. W. Sanner, A. E. Condon, L. M. Smith, and R. M. Corn, "Demonstration of a word design strategy for dna computing on surfaces," *Nucleic Acids Research*, vol. 25, no. 23, pp. 4748–4757, 1997.

[18] G. F. Joyce, "Directed evolution of nucleic acid enzymes," *Annual review of biochemistry*, vol. 73, no. 1, pp. 791–836, 2004.

[19] R. W. Hamming, "Error detecting and error correcting codes," *The Bell system technical journal*, vol. 29, no. 2, pp. 147–160, 1950.

[20] ——, *Coding and information theory*. Prentice-Hall, Inc., 1986.

[21] M. Mohammadi-Kambs, K. Holz, M. M. Somoza, and A. Ott, "Hamming distance as a concept in dna molecular recognition," *ACS omega*, vol. 2, no. 4, pp. 1302–1308, 2017.

[22] V. Phan, "A method for constructing large dna codesets," *Advanced Computational Methods for Biocomputing and Bioimaging. Nova Science Publishers, New York*, 2006.

[23] M. Garzon, V. Phan, K. Bobba, and R. Kontham, "Sensitivity analysis of microarray data: A new approach," in *Proc. IBE Conference, Athens GA*, 2005.

[24] V. Phan and M. H. Garzon, "On codeword design in metric DNA spaces," *Natural Computing*, vol. 8, no. 3, p. 571, 2009.

[25] M. Schena, *Microarray analysis*. Wiley-Liss,, 2003.

[26] M. H. Garzon and S. Mainali, "Towards reliable microarray analysis and design," in *9th International Conference on Bioinformatics and Computational Biology, ISCA, 6p*, 2017.

[27] S. Behjati and P. S. Tarpey, "What is next generation sequencing?" *Archives of Disease in Childhood-Education and Practice*, vol. 98, no. 6, pp. 236–238, 2013.

[28] G. Marcus, "Innateness, alphazero, and artificial intelligence," *arXiv preprint arXiv:1801.05667*, 2018.

[29] M. H. Garzon, "Dna codeword design: Theory and applications," *Parallel Processing Letters*, vol. 24, no. 02, p. 1440001, 2014.

[30] C. E. Shannon, "A note on the concept of entropy," *Bell System Tech. J*, vol. 27, no. 3, pp. 379–423, 1948.

[31] S. Mainali, M. Garzon, D. Venugopal, K. Jana, C.-C. Yang, N. Kumar, D. Bowman, and L.-Y. Deng, "An information-theoretic approach to dimensionality reduction in data science," *International Journal of Data Science and Analytics*, pp. 1–19, 2021.

[32] O. Z. Maimon and L. Rokach, *Data mining with decision trees: theory and applications*. World scientific, 2014, vol. 81.

[33] A. Liaw, M. Wiener, *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[34] S. S. Haykin *et al.*, "Neural networks and learning machines/simon haykin." 2009.

[35] Y. Freund, R. E. Schapire, *et al.*, "Experiments with a new boosting algorithm," in *icml*, vol. 96. Citeseer, 1996, pp. 148–156.

[36] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[37] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.

[38] T. Honkela, "Kohonen network," *Scholarpedia*, vol. 2, no. 1, p. 1568, 2007.

[39] H. Imai, M. Iri, and K. Murota, "Voronoi diagram in the laguerre geometry and its applications," *SIAM Journal on Computing*, vol. 14, no. 1, pp. 93–105, 1985.

[40] M. A. Wong and J. Hartigan, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[41] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," Stanford, Tech. Rep., 2006.

[42] F. B. Churchill, "William johannsen and the genotype concept," *Journal of the History of Biology*, vol. 7, no. 1, pp. 5–30, 1974.

[43] Z. Usova, "Ph adler, dc carry, and dm wood, the black flies (simuliidae) of north america (ithaca, london, cornell university press, 2004)," 2007.

[44] J. Rivera and D. C. Currie, "Identification of nearctic black flies using dna barcodes (diptera: Simuliidae)," *Molecular Ecology Resources*, vol. 9, pp. 224–236, 2009.

[45] A. J. Shelley, *Blackflies (Diptera: Simuliidae) of Brazil*. Pensoft Publishers, 2010.

[46] F. Vasseur, M. Exposito-Alonso, O. J. Ayala-Garay, G. Wang, B. J. Enquist, D. Vile, C. Violle, and D. Weigel, "Adaptive diversification of growth allometry in the plant arabidopsis thaliana," *Proceedings of the National Academy of Sciences*, vol. 115, no. 13, pp. 3416–3421, 2018.

[47] R. Vinces, "Phenomics: genotype to phenotype. a report of the usda/nsf phenomics workshop (2011)," 2020.

[48] R. Fritsche-Neto and A. Borém, *Phenomics: how next-generation phenotyping is revolutionizing plant breeding*. Springer, 2015.

[49] J. R. Karr, J. C. Sanghvi, D. N. Macklin, M. V. Gutschow, J. M. Jacobs, B. Bolival Jr, N. Assad-Garcia, J. I. Glass, and M. W. Covert, "A whole-cell computational model predicts phenotype from genotype," *Cell*, vol. 150, no. 2, pp. 389–401, 2012.

[50] A. Weimann, K. Mooren, J. Frank, P. B. Pope, A. Bremges, and A. C. McHardy, "From genomes to phenotypes: Traitar, the microbial trait analyzer," *MSystems*, vol. 1, no. 6, 2016.

[51] J. M. Hancock, *Phenomics*. CRC Press, 2014.

[52] O. Folmer, M. Black, W. Hoeh, R. Lutz, and R. Vrijenhoek, "Dna primers for amplification of mitochondrial cytochrome c oxidase subunit i from diverse metazoan invertebrates 3, 294–299," 1994.

[53] M. E. Carew, V. Pettigrove, R. L. Cox, and A. A. Hoffmann, "Dna identification of urban tanytarsini chironomids (diptera: Chironomidae)," *Journal of the North American Benthological Society*, vol. 26, no. 4, pp. 587–600, 2007.

[54] A. Cywinska, F. Hunter, and P. D. Hebert, "Identifying canadian mosquito species through dna barcodes," *Medical and veterinary entomology*, vol. 20, no. 4, pp. 413–424, 2006.

[55] P. D. Hebert, A. Cywinska, S. L. Ball, and J. R. Dewaard, "Biological identifications through dna barcodes," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. 1512, pp. 313–321, 2003.

[56] D. Weigel and R. Mott, "The 1001 genomes project for arabidopsis thaliana," *Genome biology*, vol. 10, no. 5, pp. 1–5, 2009.

[57] R. Cook-Deegan, C. DeRienzo, J. Carbone, S. Chandrasekharan, C. Heaney, and C. Conover, "Impact of gene patents and licensing practices on access to genetic testing for inherited susceptibility to cancer: comparing breast and ovarian cancers with colon cancers," *Genetics in Medicine*, vol. 12, no. 1, pp. S15–S38, 2010.

[58] P. Darlington, "The cost of evolution and the imprecision of adaptation," *Proceedings of the National Academy of Sciences*, vol. 74, no. 4, pp. 1647–1651, 1977.

[59] R. E. Ricklefs, "Phyletic gradualism vs. punctuated equilibrium: applicability of neontological data," *Paleobiology*, pp. 271–275, 1980.

[60] E. Sober, "What is wrong with intelligent design?" *The Quarterly Review of Biology*, vol. 82, no. 1, pp. 3–8, 2007.

[61] I. Chuine, "Why does phenology drive species distribution?" *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1555, pp. 3149–3160, 2010.

[62] J. Elith and J. R. Leathwick, "Species distribution models: ecological explanation and prediction across space and time," *Annual review of ecology, evolution, and systematics*, vol. 40, pp. 677–697, 2009.

[63] A. Guisan, R. Tingley, J. B. Baumgartner, I. Naujokaitis-Lewis, P. R. Sutcliffe, A. I. Tulloch, T. J. Regan, L. Brotons, E. McDonald-Madden, C. Mantyka-Pringle, *et al.*, "Predicting species distributions for conservation decisions," *Ecology letters*, vol. 16, no. 12, pp. 1424–1435, 2013.

[64] O. Hoegh-Guldberg, L. Hughes, S. McIntyre, D. Lindenmayer, C. Parmesan, H. P. Possingham, and C. Thomas, "Ecology. assisted colonization and rapid climate change." *Science (New York, NY)*, vol. 321, no. 5887, pp. 345–346, 2008.

[65] A. Barberán, K. S. Ramirez, J. W. Leff, M. A. Bradford, D. H. Wall, and N. Fierer, "Why are some microbes more ubiquitous than others? predicting the habitat breadth of soil bacteria," *Ecology Letters*, vol. 17, no. 7, pp. 794–802, 2014.

[66] H. R. Maturana and F. J. Varela, *Autopoiesis and cognition: The realization of the living.* Springer Science & Business Media, 2012, vol. 42.

[67] M. H. Garzon and D. T. Pham, "Genomic solutions to hospital-acquired bacterial infection identification," in *International Conference on Bioinformatics and Biomedical Engineering.* Springer, 2018, pp. 486–497.

[68] C. Linnaeus, *Systema naturae.* Stockholm Laurentii Salvii, 1758, vol. 1, no. part 1.

[69] M. Valan, K. Makonyi, A. Maki, D. Vondráček, and F. Ronquist, "Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks," *Systematic Biology*, vol. 68, no. 6, pp. 876–895, 2019.

[70] L. Van Valen, "Ecological species, multispecies, and oaks," *Taxon*, pp. 233–239, 1976.

[71] R. R. Sokal and T. J. Crovello, "The biological species concept: a critical evaluation," *The American Naturalist*, vol. 104, no. 936, pp. 127–153, 1970.

[72] E. Mayr, *Systematics and the origin of species, from the viewpoint of a zoologist.* Harvard University Press, 1999.

[73] K. De Queiroz, "Ernst mayr and the modern concept of species," *Proceedings of the National Academy of Sciences*, vol. 102, no. suppl 1, pp. 6600–6607, 2005.

[74] W. Hennig, "Phylogenetic systematics,(university of illinois press: Urbana, il, usa)," 1966.

[75] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: the primary kingdoms," *Proceedings of the National Academy of Sciences*, vol. 74, no. 11, pp. 5088–5090, 1977.

[76] C.-H. Yang, K.-C. Wu, L.-Y. Chuang, and H.-W. Chang, "Deepbarcoding: Deep learning for species classification using dna barcoding," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021.

[77] J. J. Credle, M. L. Robinson, J. Gunn, D. Monaco, B. Sie, A. Tchir, J. Hardick, X. Zheng, K. Shaw-Saliba, R. E. Rothman, *et al.*, "Highly multiplexed oligonucleotide probe-ligation testing enables efficient extraction-free sars-cov-2 detection and viral genotyping," *Modern Pathology*, pp. 1–11, 2021.

[78] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, "Introduction to pathogens," in *Molecular Biology of the Cell. 4th edition.* Garland Science, 2002.

[79] A. Casadevall and L.-a. Pirofski, "Microbiology: ditch the term pathogen," *Nature News*, vol. 516, no. 7530, p. 165, 2014.

[80] S. Cosentino, M. V. Larsen, F. M. Aarestrup, and O. Lund, "Pathogenfinder-distinguishing friend from foe using bacterial whole genome sequence data," *PloS one*, vol. 8, no. 10, p. e77302, 2013.

[81] C. Deneke, R. Rentzsch, and B. Y. Renard, "Paprbag: A machine learning approach for the detection of novel pathogens from ngs data," *Scientific reports*, vol. 7, no. 1, pp. 1–13, 2017.

[82] P. D. Hoeprich, "Host-parasite relationships and the pathogenesis of infectious disease," *Infectious diseases, PD Hoeprich and M. C. Jordan (eds.). Lippincott, Philadelphia, Pennsylvania*, pp. 41–53, 1989.

[83] C. for Disease Control, Prevention, *et al.*, "Fourth report on human exposure to environmental chemicals, updated tables," *Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention*, 2019.

[84] A. Casadevall and L.-a. Pirofski, "Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity," *Infection and immunity*, vol. 67, no. 8, p. 3703, 1999.

[85] H. Zinsser, "Infection and the problem of virulence," *Infection and resistance. The Macmillan Company, New York, NY*, pp. 1–27, 1914.

[86] T. Smith, "An attempt to interpret present-day uses of vaccines," *Journal of the American Medical Association*, vol. 60, no. 21, pp. 1591–1599, 1913.

[87] D. W. Watson and C. A. Brandly, "Virulence and pathogenicity," *Annual Review of Microbiology*, vol. 3, no. 1, pp. 195–220, 1949.

[88] S. Falkow, "Molecular koch's postulates applied to microbial pathogenicity," *Reviews of infectious diseases*, pp. S274–S276, 1988.

[89] ——, "What is a pathogen?" *ASM news*, vol. 63, p. 359, 1997.

[90] S. T. Shulman, *The Biologic and Clinical Basis of Infectious Diseases*. W.B. Saunders Company, 1997.

[91] W. W. Ford *et al.*, "Text-book of bacteriology," 1927.

[92] T. Smith, *Parasitism and disease*. Princeton, 1934.

[93] A.-E. Saliba, S. C. Santos, and J. Vogel, "New rna-seq approaches for the study of bacterial pathogens," *Current opinion in microbiology*, vol. 35, pp. 78–87, 2017.

[94] W. Gu, X. Deng, M. Lee, Y. D. Sucu, S. Arevalo, D. Stryke, S. Federman, A. Gopez, K. Reyes, K. Zorn, *et al.*, "Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids," *Nature Medicine*, vol. 27, no. 1, pp. 115–124, 2021.

[95] W. Liu, Z. Fan, Y. Zhang, F. Huang, N. Xu, L. Xuan, H. Liu, P. Shi, Z. Wang, J. Xu, *et al.*, "Metagenomic next-generation sequencing for identifying pathogens in central nervous system complications after allogeneic hematopoietic stem cell transplantation," *Bone marrow transplantation*, pp. 1–6, 2021.

[96] G. M. Cooper, R. E. Hausman, and R. E. Hausman, *The cell: a molecular approach*. ASM press Washington, DC, 2007, vol. 4.

[97] R. Plomin, *Blueprint: How DNA makes us who we are*. Mit Press, 2019.

[98] P. Yang, H. Zhou, Y. Zhu, L. Liu, and L. Zhang, "Malware classification based on shallow neural network," *Future Internet*, vol. 12, no. 12, p. 219, 2020.

[99] R. Ronen, M. Radu, C. Feuerstein, E. Yom-Tov, and M. Ahmadi, "Microsoft malware classification challenge," *CoRR*, vol. abs/1802.10135, 2018. [Online]. Available: http://arxiv.org/abs/1802.10135

[100] A. Fattori, A. Lanzi, D. Balzarotti, and E. Kirda, "Hypervisor-based malware protection with accessminer," *Computers & Security*, vol. 52, pp. 33–50, 2015.

[101] H. Hashemi, A. Azmoodeh, A. Hamzeh, and S. Hashemi, "Graph embedding as a new approach for unknown malware detection," *Journal of Computer Virology and Hacking Techniques*, vol. 13, no. 3, pp. 153–166, 2017.

[102] A. Pektaş and T. Acarman, "Classification of malware families based on runtime behaviors," *Journal of information security and applications*, vol. 37, pp. 91–100, 2017.

[103] C.-I. Fan, H.-W. Hsiao, C.-H. Chou, and Y.-F. Tseng, "Malware detection systems based on api log data mining," in *2015 IEEE 39th annual computer software and applications conference*, vol. 3. IEEE, 2015, pp. 255–260.

[104] I. Santos, F. Brezo, X. Ugarte-Pedrero, and P. G. Bringas, "Opcode sequences as representation of executables for data-mining-based unknown malware detection," *Information Sciences*, vol. 231, pp. 64–82, 2013.

[105] J. Reimann and G. Vachtsevanos, "Uavs in urban operations: Target interception and containment," *Journal of Intelligent and Robotic Systems*, vol. 47, no. 4, pp. 383–396, 2006.

[106] Ö. A. Aslan and R. Samet, "A comprehensive review on malware detection approaches," *IEEE Access*, vol. 8, pp. 6249–6271, 2020.

[107] H. HaddadPajouh, A. Dehghantanha, R. Khayami, and K.-K. R. Choo, "A deep recurrent neural network based approach for internet of things malware threat hunting," *Future Generation Computer Systems*, vol. 85, pp. 88–96, 2018.

[108] J. Yan, Y. Qi, and Q. Rao, "Detecting malware with an ensemble method based on deep neural network," *Security and Communication Networks*, vol. 2018, 2018.

[109] D. Vasan, M. Alazab, S. Wassan, B. Safaei, and Q. Zheng, "Image-based malware classification using ensemble of cnn architectures (imcec)," *Computers & Security*, vol. 92, p. 101748, 2020.

[110] X. Wang, J. Liu, and X. Chen, "Microsoft malware classification challenge (big 2015) first place team: say no to overfitting," *no. Big*, 2015.

[111] H. Zhang, X. Xiao, F. Mercaldo, S. Ni, F. Martinelli, and A. K. Sangaiah, "Classification of ransomware families with machine learning based on n-gram of opcodes," *Future Generation Computer Systems*, vol. 90, pp. 211–221, 2019.

[112] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro, "Improving semantic segmentation via video propagation and label relaxation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8856–8865.

[113] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[114] V. Badrinarayanan, F. Galasso, and R. Cipolla, "Label propagation in video sequences," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3265–3272.

[115] I. Budvytis, P. Sauer, T. Roddick, K. Breen, and R. Cipolla, "Large scale labelled video data augmentation for semantic segmentation in driving scenarios," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 230–237.

[116] S. K. Mustikovela, M. Y. Yang, and C. Rother, "Can ground truth label propagation from video help semantic segmentation?" in *European Conference on Computer Vision*. Springer, 2016, pp. 804–820.

[117] R. Gadde, V. Jampani, and P. V. Gehler, "Semantic video cnns through representation warping," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4453–4462.

[118] D. Nilsson and C. Sminchisescu, "Semantic video segmentation by gated recurrent flow propagation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6819–6828.

[119] P.-Y. Huang, W.-T. Hsu, C.-Y. Chiu, T.-F. Wu, and M. Sun, "Efficient uncertainty estimation for semantic segmentation in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 520–535.

[120] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 158–172, 2018.

[121] T.-W. Ke, J.-J. Hwang, Z. Liu, and S. X. Yu, "Adaptive affinity fields for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 587–602.

[122] G. Bertasius, L. Torresani, S. X. Yu, and J. Shi, "Convolutional random walk networks for semantic image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 858–866.

[123] R. Vieux, J. Benois-Pineau, J.-P. Domenger, and A. Braquelaire, "Segmentation-based multi-class semantic object detection," *Multimedia Tools and Applications*, vol. 60, no. 2, pp. 305–326, 2012.

[124] A. R. Araújo and D. C. Costa, "Local adaptive receptive field self-organizing map for image color segmentation," *Image and Vision Computing*, vol. 27, no. 9, pp. 1229–1239, 2009.

[125] D. Chi, "Self-organizing map-based color image segmentation with k-means clustering and saliency map," *International Scholarly Research Notices*, vol. 2011, 2011.

[126] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on neural networks*, vol. 11, no. 3, pp. 586–600, 2000.

[127] K. Tasdemir and E. Merényi, "Exploiting data topology in visualization and clustering of self-organizing maps," *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 549–562, 2009.

[128] D. Brugger, M. Bogdan, and W. Rosenstiel, "Automatic cluster detection in kohonen's som," *IEEE Transactions on Neural Networks*, vol. 19, no. 3, pp. 442–459, 2008.

[129] E. Aghajari and G. D. Chandrashekhar, "Self-organizing map based extended fuzzy c-means (seefc) algorithm for image segmentation," *Applied Soft Computing*, vol. 54, pp. 347–363, 2017.

[130] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *The journal of machine learning research*, vol. 13, no. 1, pp. 27–66, 2012.

[131] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "The kitti vision benchmark suite," *URL http://www. cvlibs. net/datasets/kitti*, vol. 2, 2015.

[132] S. Chandra, C. Couprie, and I. Kokkinos, "Deep spatio-temporal random fields for efficient video segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8915–8924.

[133] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341.

[134] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[135] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[136] I. Kreso, S. Segvic, and J. Krapac, "Ladder-style densenets for semantic segmentation of large natural images," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 238–245.

[137] P. Meletis and G. Dubbelman, "Training of convolutional networks on multiple heterogeneous datasets for street scene semantic segmentation," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1045–1050.

[138] S. R. Bulo, L. Porzi, and P. Kontschieder, "In-place activated batchnorm for memory-optimized training of dnns," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5639–5647.

[139] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "Segstereo: Exploiting semantic information for disparity estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 636–651.

[140] S. Kong and C. Fowlkes, "Pixel-wise attentional gating for parsimonious pixel labeling," *arXiv preprint arXiv:1805.01556*, 2018.