DNA visualization

Author: Jeisson Andres Prieto Velandia

Visual Computing
Universidad Nacional de Colombia, Bogotá D.C., Colombia

November 2017

Outline

Outline

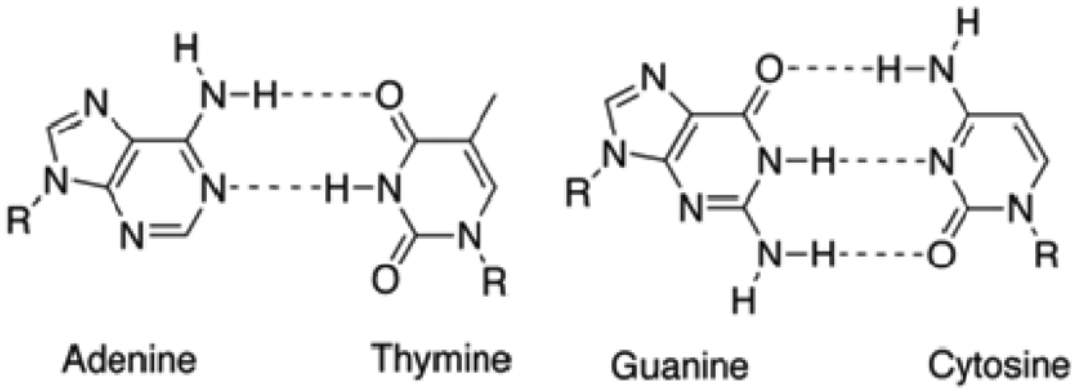
Introduction

DNA computing has originated novel ideas and uses for DNA as:

- **Self-assembly.** [Garzon et al. (2009); Qian & Winfree (2011); Seeman (2003)]
- **Natural Language Processing.** [Neel & Garzon (2006); Bobba et al. (2006)]
- **DNA-based memories.** [Garzon et al. (2003)]

Watson-Crick base pairs

A base pair is a unit consisting of two nucleobases bound to each other by hydrogen bonds.

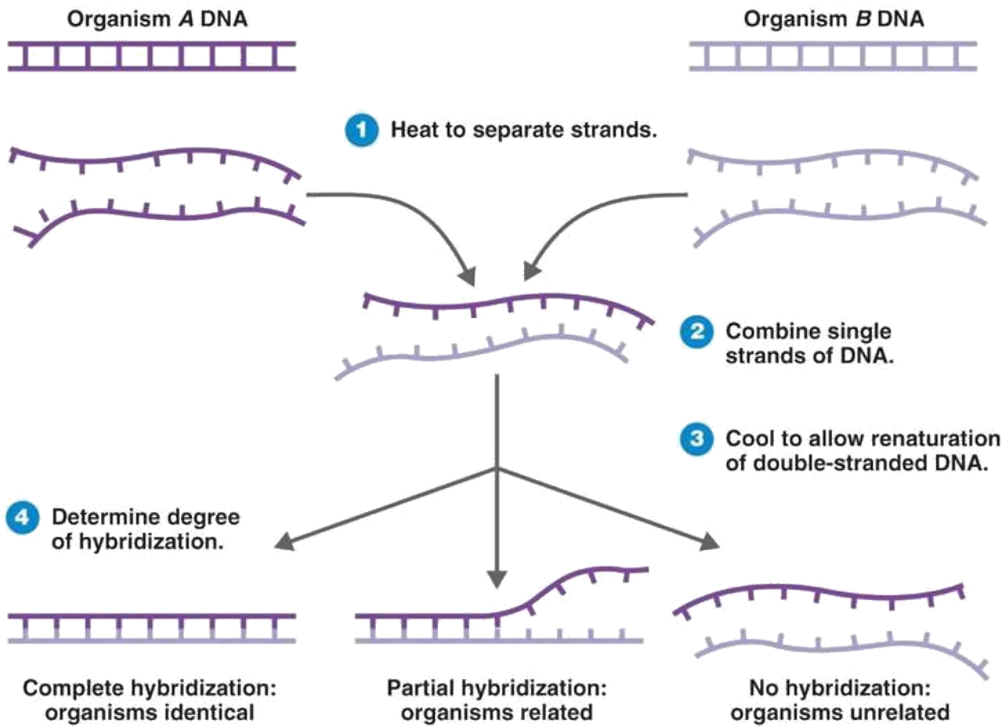


Source: wikipedia.com

The **Watson-Crick complement** y' of strand y is obtained by reversing it and swapping nucleotides within the pairs a, t and c, g. [Watson et al. (1953)]

Hybridization

Degree of genetic similarity between pools of DNA strands.



Source: pinterest.com

Codeword Design problem

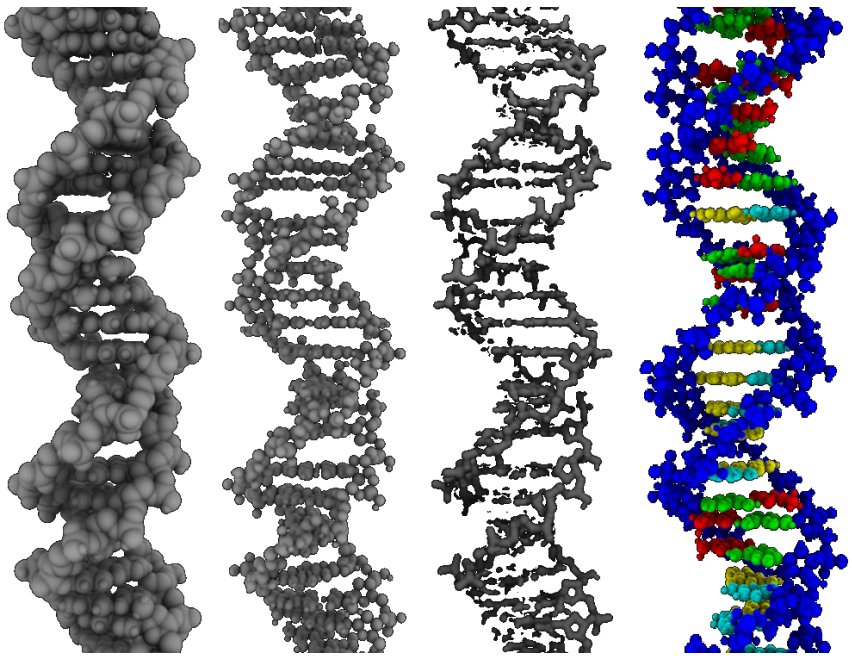
The **Codeword Design** problem calls for finding large sets of single DNA strands that do noncrosshybridize(nxh) to themselves and/or to their complements.

Theorem

*CODEWORD DESIGN is **NP**-complete for any measure of hybridization affinity, satisfying strict nonnegativity (i.e., such that* $d(x, y) \geq 0$ *and* $d(x, y) = 0$ *if and only if* $y = x'$*.)*

DNA visualization - Goal

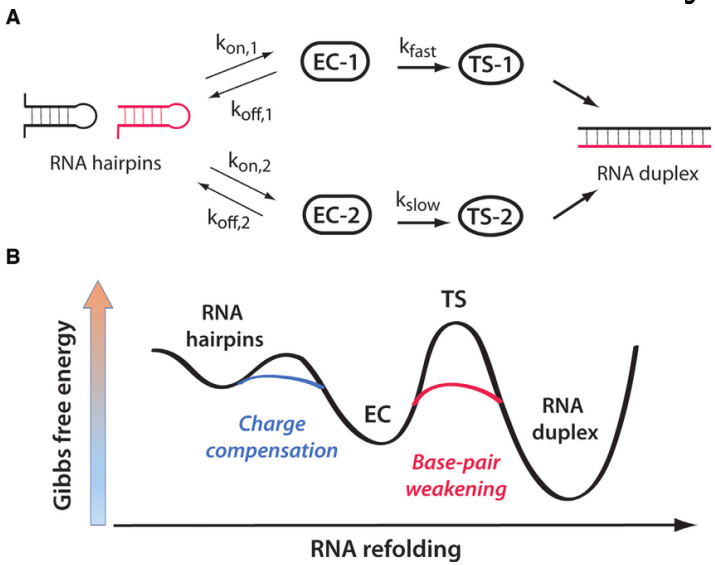
Visualize the DNA sequences using the Noncrosshybridizing set found using the Self-adaptive Evolutionary Algorithm.



Source: NCSA'S Advanced Visualization Laboratory

Gibbs energy

DNA formation of two strands is determined by the **Gibbs energy**.



Source: Rennella et al. (2017)

The threshold for duplex formation of strands can be considered in $-6Kcal/mol$. The more negative the energy, the more stable the duplex formed.

Approximation to the Gibbs Energy

The ***h-distance*** provides a computationally efficient approximation of the Gibbs energy based solely on composition and sequence. [Garzon et al. (1997)]

$$h(x,y) = \min_{-n<k<n} \{|k| + H(x, \sigma^k(y'))\} \tag{1}$$

where $\sigma^k(y')$ is the shift of y' by k positions from a perfect alignment with x (right-shift if $k > 0$; left-shift if $k < 0$), y' is the Watson-Crick complement of y, and the Hamming distance H measures the number of mismatched base pairs in the overlap of x and y' in the specified frame shift $\sigma^k(y')$.

h-distance

For example, if:

$$x = agc, y = tgg \text{ (and so } y' = cca)$$

- at shift $k = -2$, $\overset{agc}{cca}$ the distance is: $2 + H(a, a) = 2$
- at shift $k = -1$, $\overset{agc}{cca}$ the distance is: $1 + H(ag, ca) = 3$
- at shift $k = 0$, $\overset{agc}{cca}$ the distance is: $0 + H(agc, cca) = 3$
- at shift $k = 1$, $\overset{agc}{cca}$ the distance is: $1 + H(gc, cc) = 2$
- at shift $k = 2$, $\overset{agc}{cca}$ the distance is: $2 + H(c, c) = 2$

Thus:

$$h(agc, tgg) = 2$$

DNA codes

The metric space D_n has the following properties for all $n \geq 1$ [Phan & Garzon (2008)]

1. There are $|P| = 4^{n/2}$ n-mers consisting of a single palindromic DNA strand that are their own reverse complements (i.e., $|X| = 1$) for n even, and 0 for n odd.
2. There are $|D_n| = \frac{4^n - |P|}{2}$, nonpalindromic n-mers.
3. There are $|D_n| = \frac{4^n + |P|}{2}$, n-mers in total.

Codeword Design in terms of the *h-distance*

Given a set of n-mers S, i.e., D_n, possible with a lot of crosshybridization, and a reaction stringency threshold τ, find a subset of S with no crosshybridization, i.e., find the largest (n, τ)-code in S.

Codeword Design in DNA Spaces

Input: A set S of n-mers, a threshold τ and an integer K;
Output:Is there an (n, τ)-code subset of S of cardinality at least K, i.e., where every two distinct words are at a distance at least τ from each other?

Self-adaptive Evolutionary Algorithm - nxh set

Table 1: Noise quality of various nxh bases founded using SaEA, quantified using the expected number of hybridizations of a random pmer and Shannon Entropy of the corresponding distribution (Expected value/Shannon entropy)

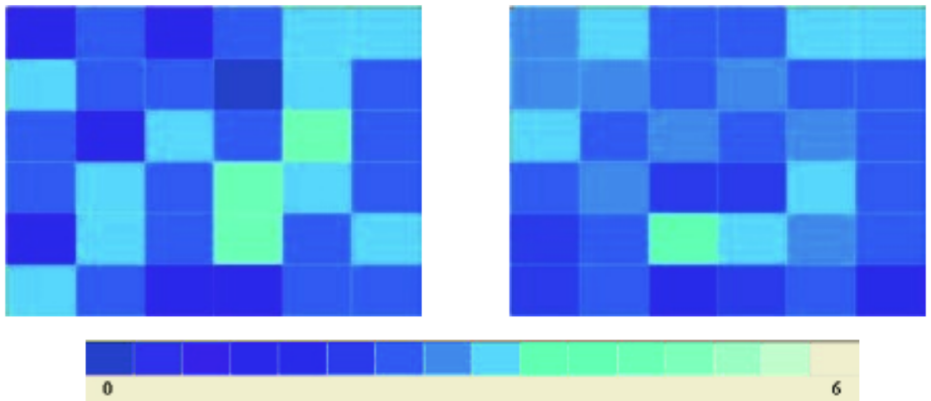
Length	$\tau = 50\%$
4-mers	0.97 / 0.88
6-mers	0.92 / 0.56
8-mers	0.89 / 0.61

Outline

Case of Study

Signature of a DNA sequence. [Garzon et al. (2004)]

The signature of a string x is represented as a vector V based on how x hybridize in some stringency parameter τ with a nxh set. V could be visualize as 1D or 2D signature.

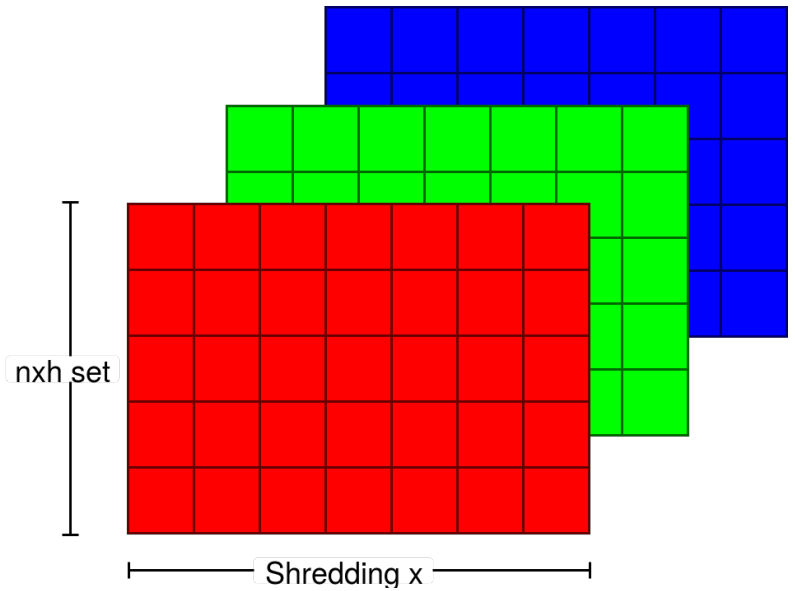


Source: Garzon et al. (2004)

Methodology

Represent the DNA sequence as a RGB color model

- **R**ed: nxh set with $n = 8$ and $\tau = n/2$
- **G**reen: nxh set with $n = 6$ and $\tau = n/2$
- **B**lue: Leave the light at 0(Noisy).



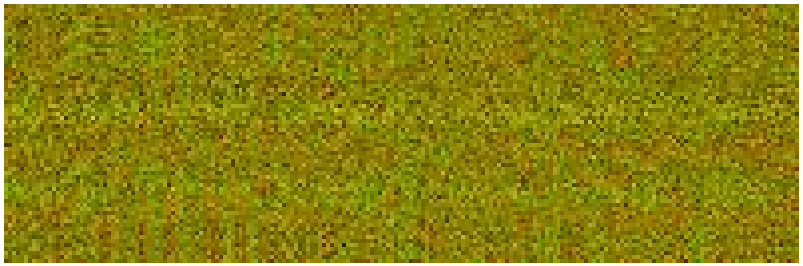
Outline

Dataset

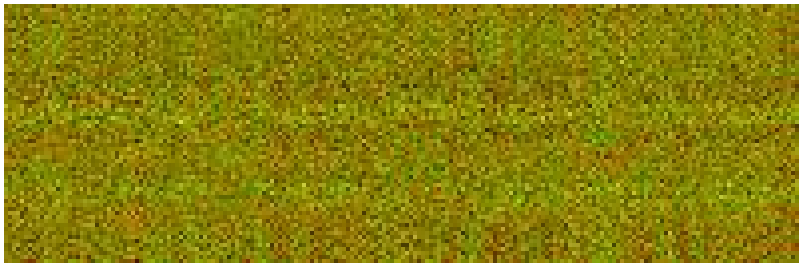
Table 2: Species description.

Specie	Scientific name	Common name	DNA length
Plant	*Helianthus annuus*	Sunflower	301004
	Hordeum vulgare	Barley	416675
	Triticum aestivum	Wheat	452526
Virus	*Camelpox virus*	Camels disease	205719
	Canarypox virus	Birds disease	359853
	Variola major virus	Smallpox	186103
Fungi	*Ganoderma lucidum*	Lingzhi mushroom	60635
	Lentinula edodes	Shiitake	121394
	Pleurotus ostreatus	Oyster mushroom	73242
Bacterium	*Anaplasma phagocytophilum*	Tick-borne fever	1471282
	Neisseria gonorrhoeae	Gonorrhea	942943
	Streptococcus pyogenes	Mastitis	1750832

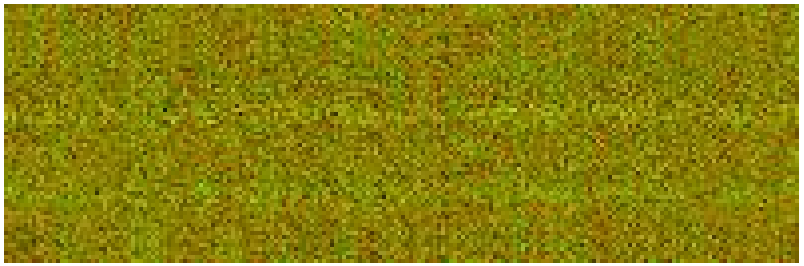
Plant



(a) Helianthus annuus



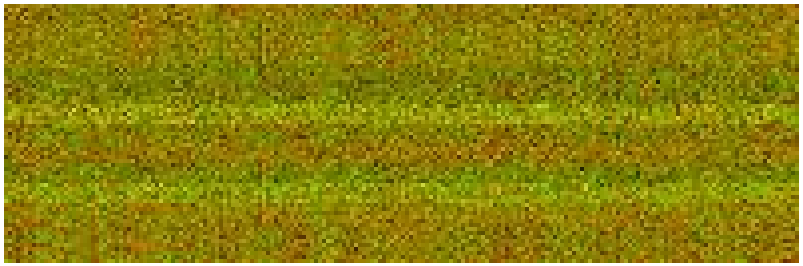
(b) Hordeum vulgare



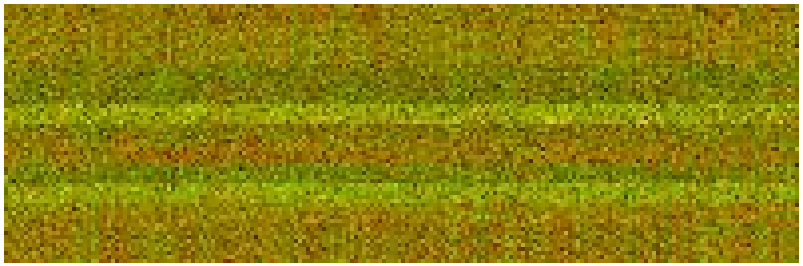
(c) Triticum aestivum

Figure 1: RGB representation of the DNA sequences of Plant specie.

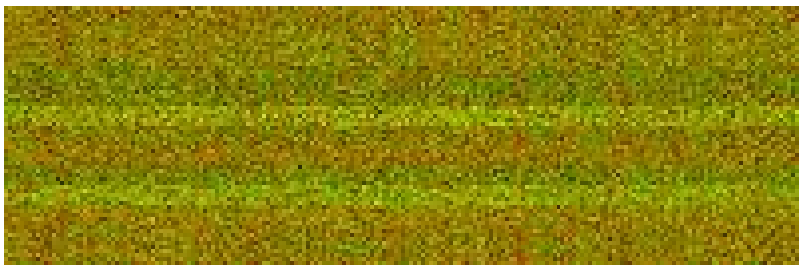
Virus



(a) Camelpox virus



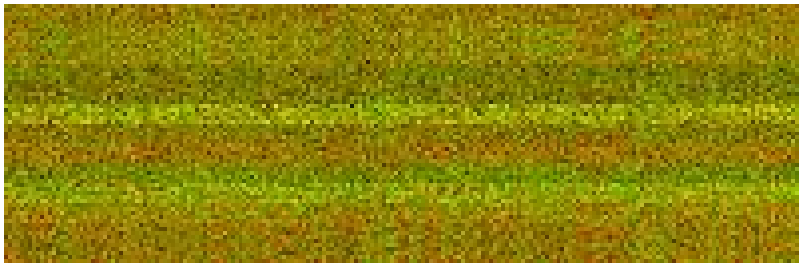
(b) Canarypox virus



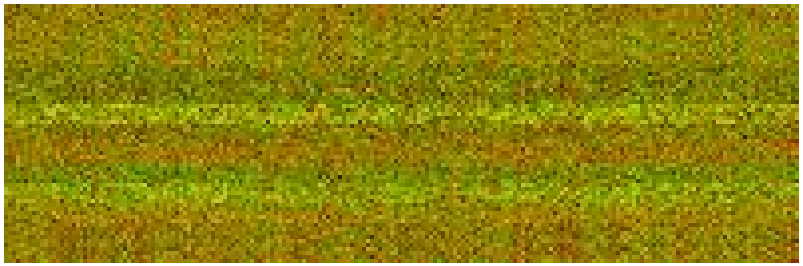
(c) Variola major virus

Figure 2: RGB representation of the DNA sequences of Virus specie.

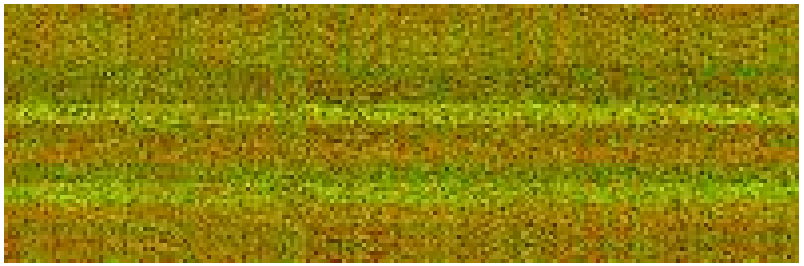
Fungi



(a) Ganoderma lucidum



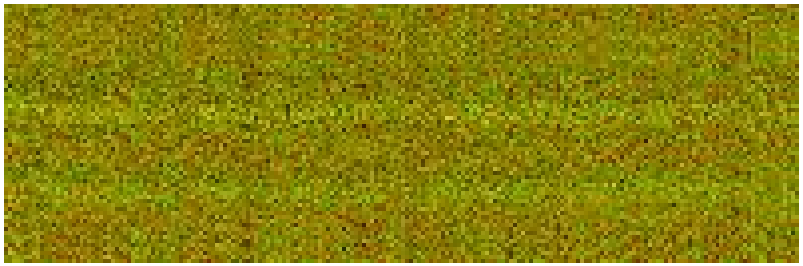
(b) Lentinula edodes



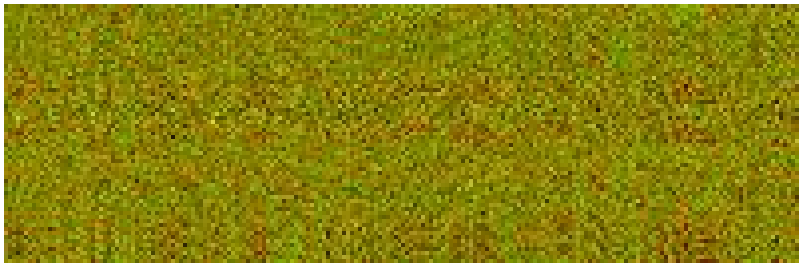
(c) Pleurotus ostreatus

Figure 3: RGB representation of the DNA sequences of Fungi specie.

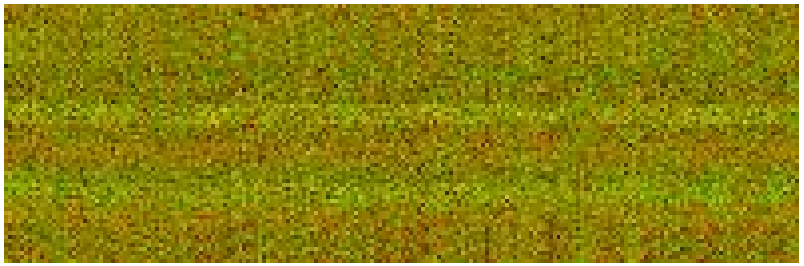
Bacterium



(a) Anaplasma phagocytophilum



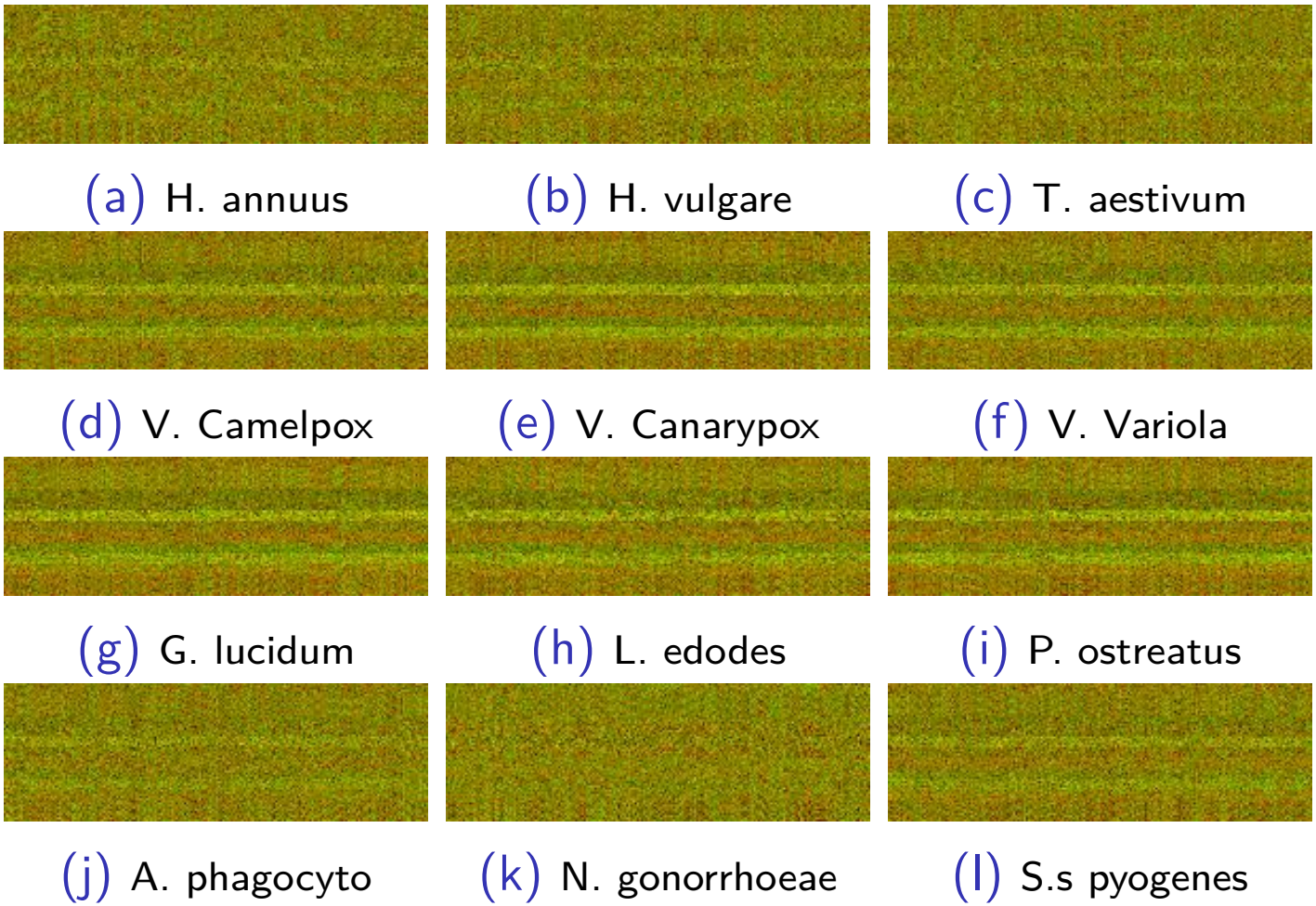
(b) Neisseria gonorrhoeae



(c) Streptococcus pyogenes

Figure 4: RGB representation of the DNA sequences of Bacterium specie.

Cmparison between species



(a) H. annuus (b) H. vulgare (c) T. aestivum

(d) V. Camelpox (e) V. Canarypox (f) V. Variola

(g) G. lucidum (h) L. edodes (i) P. ostreatus

(j) A. phagocyto (k) N. gonorrhoeae (l) S.s pyogenes

Figure 5: Comparison between species using the RGB representation.

Outline

Conclusions and Future Work

A new methodology for find a image representation of a DNA sequence has been described which uses the Noncrosshybridizing sets in the DNA spaces to get the best.

The new DNA-based technique for species identification offers several other advantages.

- More effectively in terms of cost and time than by traditional methods
- Can be readily extended to whole-genomes and thus applicable to arbitrary organisms.

Conclusions and Future Work

Using a larger space (10-mers), the images can be take more advantages of the hybridizations properties in the DNA space.

Additional ways to make the species identification. (Hybrid model)

Machine learning algorithms to image classification.

THANKS!

References I

Bobba, K. C., Neel, A. J., Phan, V., & Garzon, M. (2006). "reasoning" and "talking" dna: Can dna understand english? In C. Mao & T. Yokomori (Eds.), *Dna computing: 12th international meeting on dna computing, dna12, seoul, korea, june 5-9, 2006, revised selected papers* (pp. 337–349). Berlin, Heidelberg: Springer Berlin Heidelberg.

Garzon, M., Bobba, K., & Hyde, B. (2004). Digital information encoding on dna. In N. Jonoska, G. Păun, & G. Rozenberg (Eds.), *Aspects of molecular computing: Essays dedicated to tom head, on the occasion of his 70th birthday* (pp. 152–166). Berlin, Heidelberg: Springer Berlin Heidelberg.

References II

Garzon, M., Bobba, K., & Neel, A. (2003). Efficiency and reliability of semantic retrieval in dna-based memories. In *International workshop on dna-based computers* (pp. 157–169).

Garzon, M., Neathery, P., Deaton, R., Murphy, R. C., Franceschetti, D. R., & Stevens Jr, S. E. (1997). A new metric for DNA computing. *Proceedings of the 2nd Genetic Programming Conference*, 278–472.

Garzon, M., Phan, V., & Neel, A. (2009). Optimal DNA Codes for Computing and Self-Assembly. *Journal of Nanotechnology and Molecular Computation*, *1*(1), 1–17.

Neel, A., & Garzon, M. (2006). Semantic retrieval in dna-based memories with gibbs energy models. *Biotechnology Progress*, *22*(1), 86-90.

References III

Phan, V., & Garzon, M. (2008, Jun 25). On codeword design in metric dna spaces. *Natural Computing*, *8*(3), 571.

Qian, L., & Winfree, E. (2011). Scaling up digital circuit computation with dna strand displacement cascades. *Science*, *332*(6034), 1196–1201.

Rennella, E., Sára, T., Juen, M., Wunderlich, C., Imbert, L., Solyom, Z., ... others (2017). Rna binding and chaperone activity of the e. coli cold-shock protein cspa. *Nucleic acids research*, *45*(7), 4255–4268.

Seeman, N. C. (2003). Dna in a material world. *Nature*, *421*(6921), 427–431.

Watson, J. D., Crick, F. H., et al. (1953). Molecular structure of nucleic acids. *Nature*, *171*(4356), 737–738.