

# Asociación de variantes en regiones codificantes de genes con datos clínicos en pacientes colombianos usando minería de datos.

Author: Jennifer Vélez Segura

Grupo de Investigación – MIDAS  
Universidad Nacional de Colombia, Bogotá D.C., Colombia

Mayo 2018

# Outline

- 1 Introducción
- 2 Images
  - One image
  - More than one image
- 3 Tables and Equations
  - Tables
  - Equation
- 4 Conclusions and Future Work

# Outline

## 1 Introducción

## 2 Images

- One image
- More than one image

## 3 Tables and Equations

- Tables
- Equation

## 4 Conclusions and Future Work

# Introducción


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

- **Self-assembly.** [???
- **Natural Language Processing.** [??]
- **DNA-based memories.** [?]

# One image from experiments

Represent the DNA sequence as a RGB color model

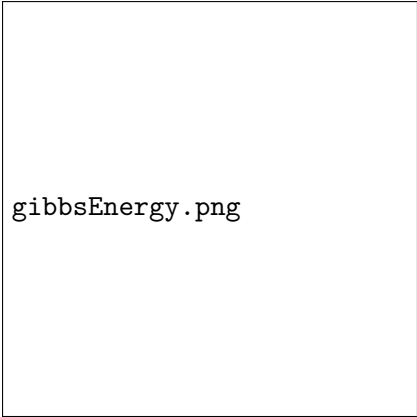
- **Red:**
- **Green:**
- **Blue:**



rgbRepresentation.png

## One image using source

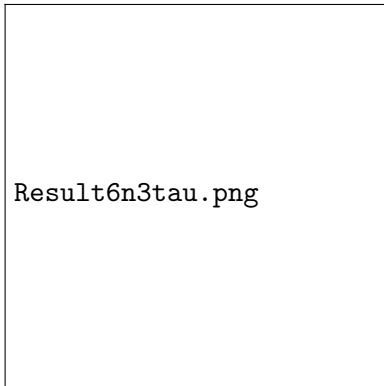
DNA formation of two strands is determined by the **Gibbs energy**.



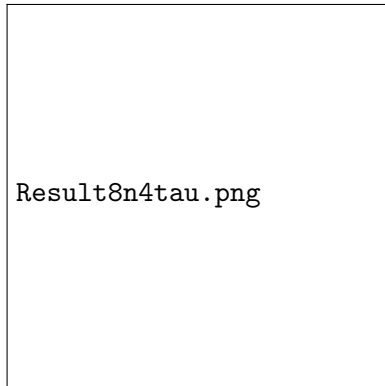
`gibbsEnergy.png`

Source: ?

## Two images



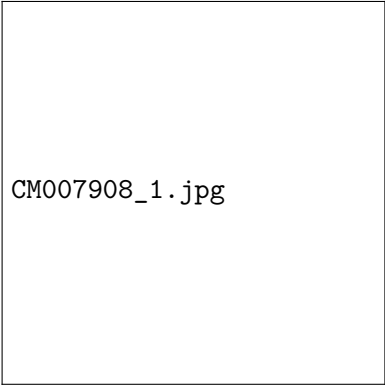
(a) 6-mers



(b) 8-mers

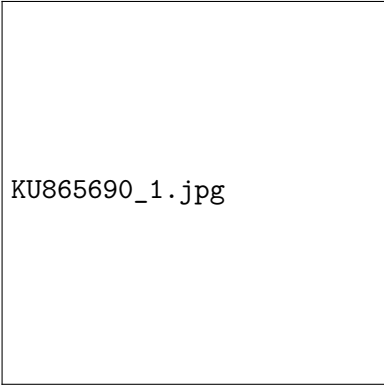
**Figure 1:** Average of probabilities over 10 runs using a population of 400 individuals and 100 generations for (a) 6-mers, (b) 8-mers with  $\tau = 50\%$  .

# Three images



CM007908\_1.jpg

(a) *Helianthus annuus*

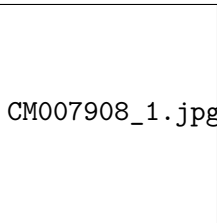


KU865690\_1.jpg

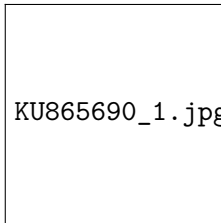
(b) *Hordeum vulgare*



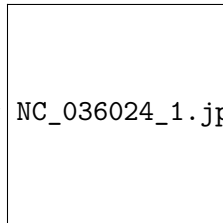
# A lot of images



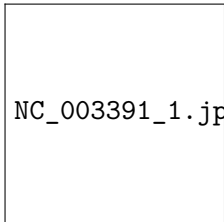
CM007908\_1.jpg

(a) *H. annuus*

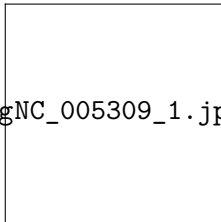
KU865690\_1.jpg

(b) *H. vulgare*

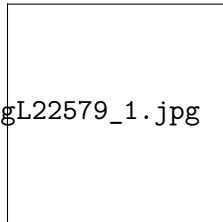
NC\_036024\_1.jpg

(c) *T. aestivum*

NC\_003391\_1.jpg

(d) *V. Camelpox*

NC\_005309\_1.jpg

(e) *V. Canarypox*

L22579\_1.jpg

(f) *V. Variola*

## A little table

**Table 1:** Noise quality of various  $n$ xh 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

# A big table

Table 2: Species description.

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

# An equation an his explanation

The ***h-distance*** provides a computationally efficient approximation of the Gibbs energy based solely on composition and sequence. [?]

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

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  $\sigma^k(y')$ .

# Making a example of a given equation

For example, if:

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

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

Thus:

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

## Enumerate equations

The metric space  $D_n$  has the following properties for all  $n \geq 1$  [?]

- 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.

# A box with some theorem

## Codeword Design in DNA Spaces

**Input:** A set  $S$  of  $n$ -mers, a threshold  $\tau$  and an integer  $K$ ;

**Output:** Is there an  $(n, \tau)$ -code subset of  $S$  of cardinality at least  $K$ , i.e., where every two distinct words are at a distance at least  $\tau$  from each other?

# Equations, equations and more equations

$$f_1 = Correspondencia(\text{Nro cuenta}, \text{id cuenta}) \quad (1)$$

$$f_2 = \frac{P(Pago = x)}{x} * f_1 \quad (2)$$

$$f_3 = distancia(tiempo, \$\$) \text{ al vecino mas cercano} \quad (3)$$

$$f_4 = tiempo_{envio} - tiempo_{generacion} \quad (4)$$



# Equations, equations and more equations

$$f_5 = \left| \hat{\Theta} - f_4 \right| \quad (5)$$

$$f_{6.1} = P(\textit{Proveedor} = x | \textit{Tipo\_Reclamante} = 10) \quad (6a)$$

$$f_{6.2} = P(\textit{Tipo\_Reclamante} = x) \quad (6b)$$

$$f_7 = \textit{Correspondencia}(\text{Nro cuenta}, \text{id cuenta}) \quad (7)$$

## Equations, equations and more equations

$$f_{8.1} = P(\textit{Pago} = x | \textit{Reclamante} = \textit{empleado}) \quad (8a)$$

$$f_{8.2} = P(\textit{Pago} = x | \textit{Reclamante} = \textit{familiar}) \quad (8b)$$

$$f_9 = P(\textit{Sucursal} = x) | x \in \{\textit{bajo}, \textit{medio}, \textit{alto}\} \quad (9)$$

$$f_{10} = P(\textit{Pago} = x | \textit{Reclamante} = \textit{ex-empleado}) \quad (10)$$

# Equations, equations and more equations

$$f_{11} = P(Pago = x | Reclamante = empleado_muerto) \quad (11)$$

$$f_{12} = P(Reclamante = x | Ramo = Autos) \quad (12)$$

$$f_{13} = P(Reclamante = x | Amparo = \{parcial_daños, total_hurto\}) \quad (13)$$

# Outline

- 1 Introducción
- 2 Images
  - One image
  - More than one image
- 3 Tables and Equations
  - Tables
  - Equation
- 4 Conclusions and Future Work

## 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