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

Outline

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

- **R**ed:
- **G**reen:
- **B**lue:

rgbRepresentation.png

One image using source

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

gibbsEnergy.png

Source: **?**

Two images

Result6n3tau.png

(a) 6-mers

Result8n4tau.png

(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 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

A big table

Table 2: Species description.

Specie	Scientific name	Common name	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

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 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?

Equations, equations and more equations

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

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

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

$$f_4 = tiempo_{envio} - tiempo_{generacion} \tag{4}$$

Equations, equations and more equations

$$f_5 = \left|\hat{\Theta} - f_4\right| \tag{5}$$

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

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

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

Equations, equations and more equations

$$f_{8.1} = P(Pago = x|Reclamante = empleado) \tag{8a}$$
$$f_{8.2} = P(Pago = x|Reclamante = familiar) \tag{8b}$$

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

$$f_{10} = P(Pago = x|Reclamante = ex\_empleado) \tag{10}$$

Equations, equations and more equations

$$f_{11} = P(Pago = x | Reclamante = empleado\_muerto) \tag{11}$$

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

$$f_{13} = P(Reclamante = x | Amparo = \{parcial\_daños, total\_hurto\}) \tag{13}$$

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