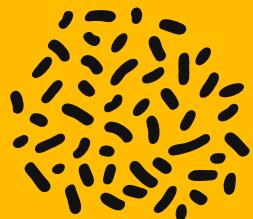


Modelo probabilístico para recrear un genoma.

Quintana Silva Brenda Paola



Antecedentes

Modificaciones genéticas



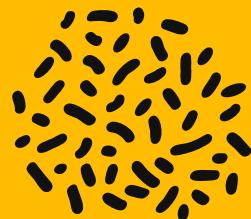
Bacterias y Levaduras

- ✗ Producir proteínas como la insulina o algunas hormonas.

Animales

- ✗ El estudio de enfermedades humanas y sus tratamientos.
- ✗ Transplante de órganos





Antecedentes



Humanos

- ✗ Modificar el tejido de un paciente que tiene una enfermedad debido a un gen defectuoso.



Plantas

- ✗ Incorporar caracteres de interés.



“Motivación

- ✗ Para hacer cualquier modificación genética es necesario disponer de la ubicación los genes.
- ✗ Los métodos experimentales pueden resultar costosos, el problema está en encontrar una forma de optimizar recursos y tiempo para encontrar una aproximación a un mapa genético.



Propuesta

Proyecto

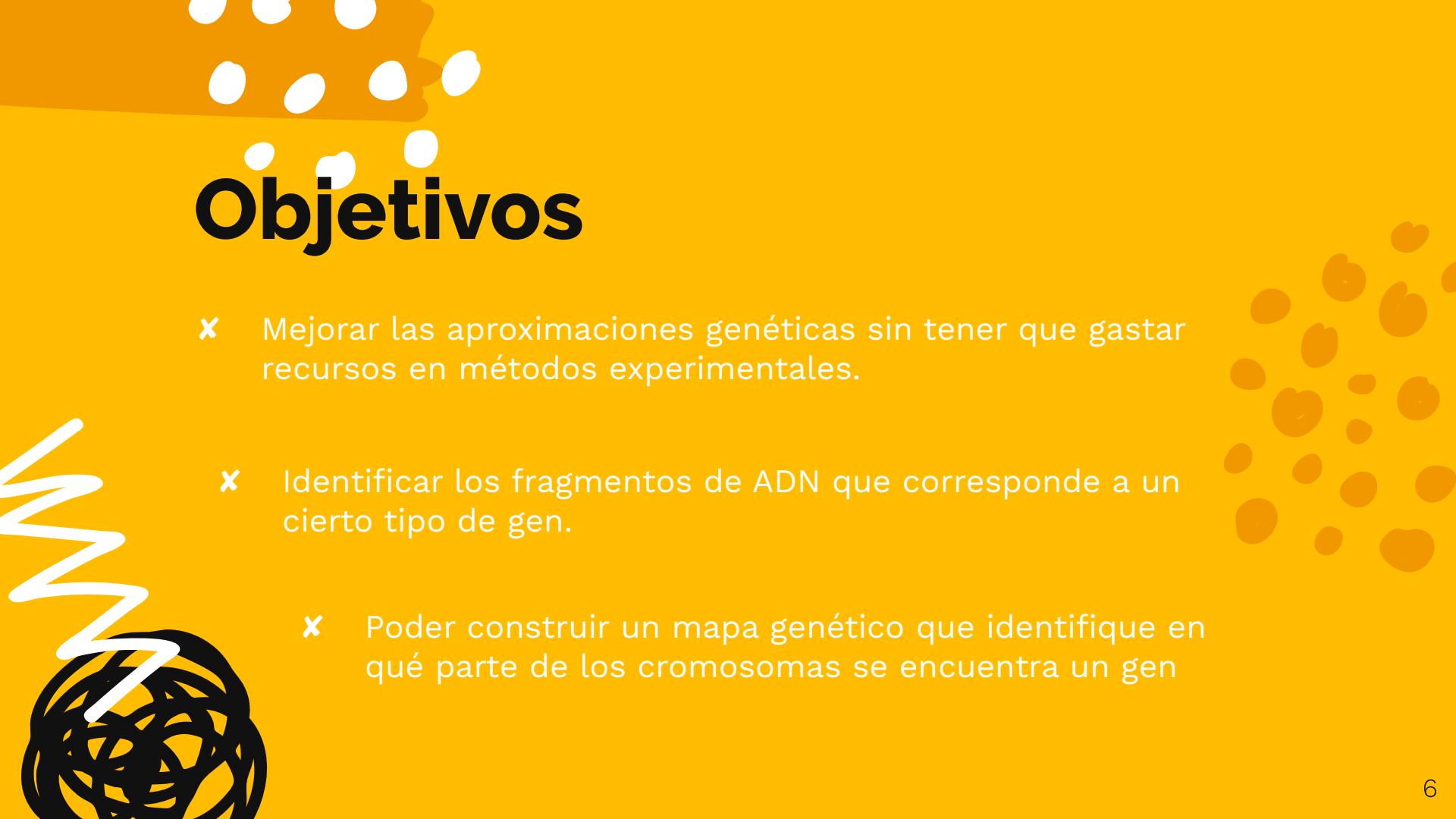
La intención es hacer una mejor primer aproximada de las regiones de codificación.

Se propone una forma distinta de aproximación a través de modelos probabilísticos.

¿Cómo?

Modelos Matemático,
estimación y optimización.





Objetivos

- ✗ Mejorar las aproximaciones genéticas sin tener que gastar recursos en métodos experimentales.
- ✗ Identificar los fragmentos de ADN que corresponde a un cierto tipo de gen.
- ✗ Poder construir un mapa genético que identifique en qué parte de los cromosomas se encuentra un gen

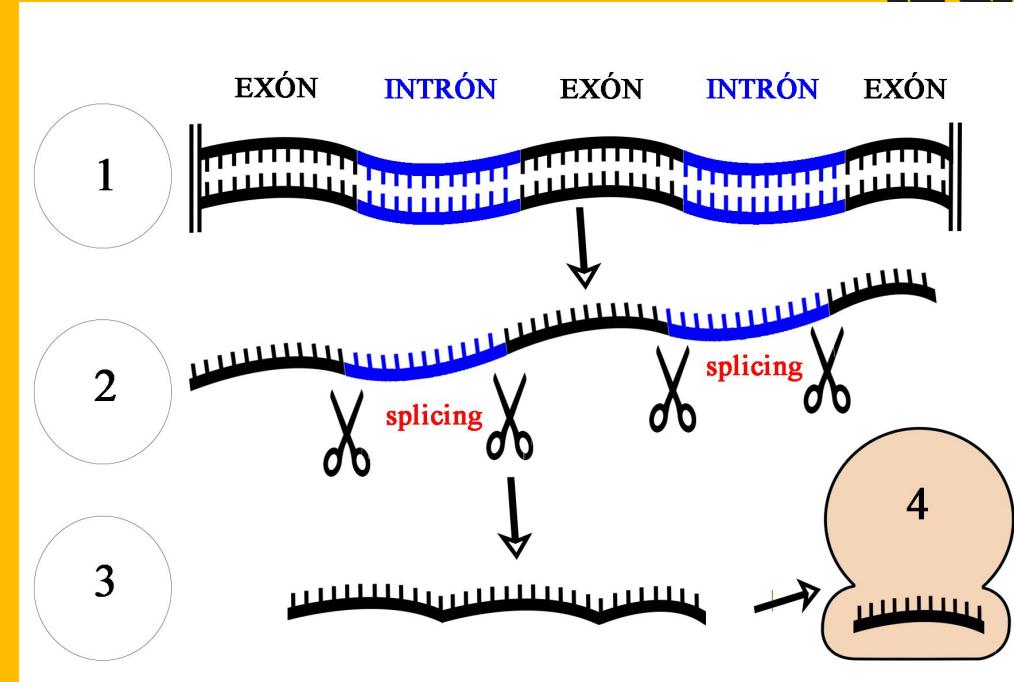
Características del problema

Clasificación

El ADN es una secuencia de nucleótidos muy larga; contiene subsecuencias específicas llamadas exones e intrones

El número de intrones en un genoma varía de acuerdo a la especie.

En los exones es donde se codifican los genes; estos se separan en tipos de genes.



Stops y Start

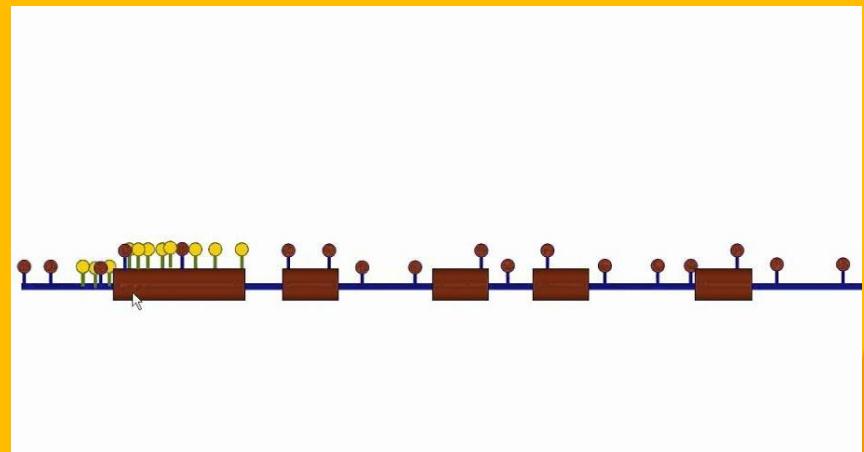
Son aminoácidos que tiene la función de indicar dónde inician y terminan las regiones codificantes (start y Stops)

Start: ATG

Stops: TAG, TGA y TAA .

Islas CpG

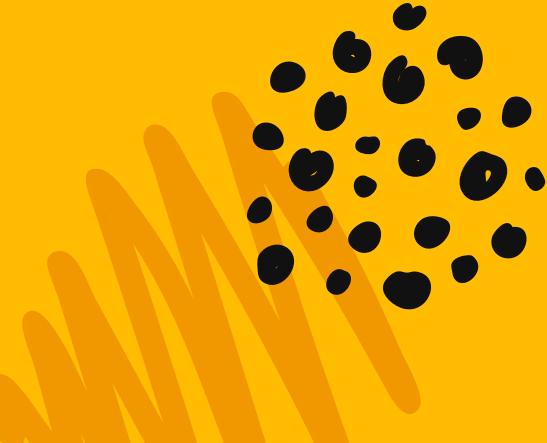
- ✖ Las islas CpG son regiones de ADN donde se encuentra esta secuencia.
- ✖ Entre el 50% y el 70% de los genes tienen una isla CpG asociada a sus promotores





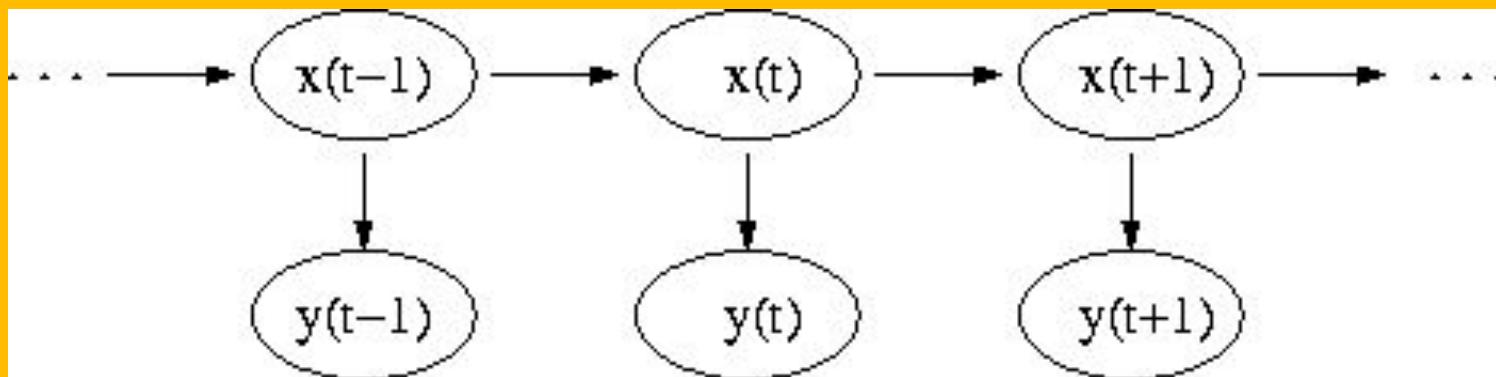
Modelación y Simulación

HMM y algoritmos MCMC



Modelos Ocultos de Markov

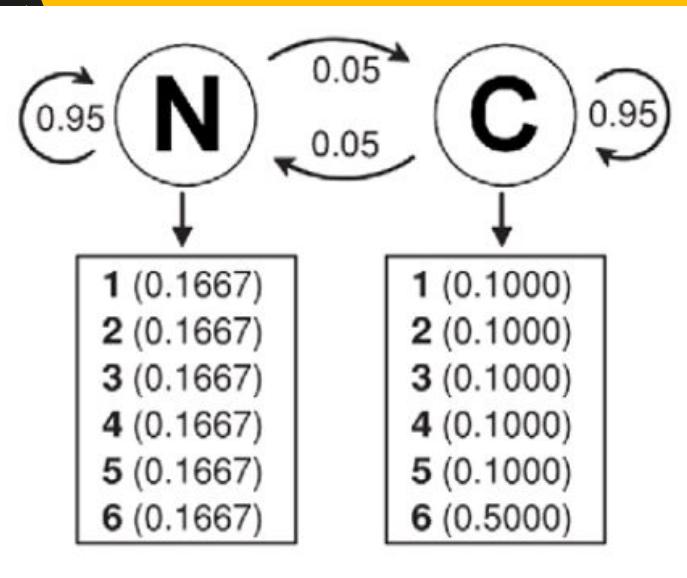
Un doble proceso



Modelos Ocultos de Markov

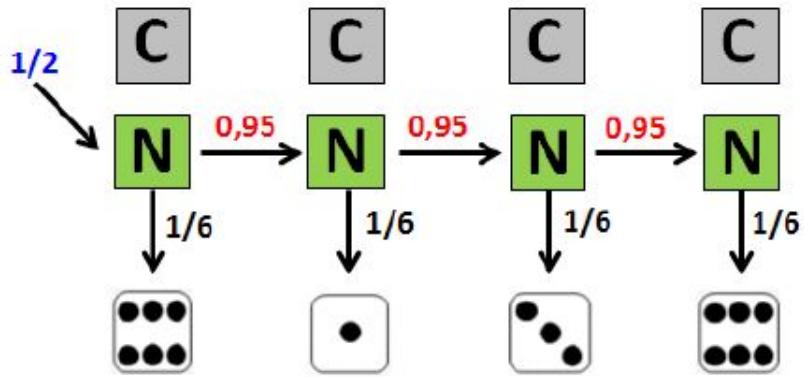
- ✗ El **número de estados** que aparecen en la secuencia **observable** m. Espacio de Estados $O = \{A, T, G, C\}$.
- ✗ Las **probabilidades de transición** de la cadena observable: es una matriz $r \times m$
- ✗ El **número de estados ocultos** r, $r \geq 2$. Espacio de estados $G = \{l, g_1, \dots, g_{r-1}\}$
- ✗ Las **probabilidades de transición** de la cadena oculta. Matriz de $r \times r$.
- ✗ Las probabilidades del **estado inicial**.

Ejemplo



Casino

dados normales (N, no cargados) la probabilidad de que salga cada uno de los 6 números es idéntica ($1/6$). Si el dado está cargado (C, cargado) la mitad de las veces sale el 6 (probabilidad = $1/2$) y el resto de los números aparecen con igual probabilidad ($1/10$).



Secuencia Oculta: NNNN

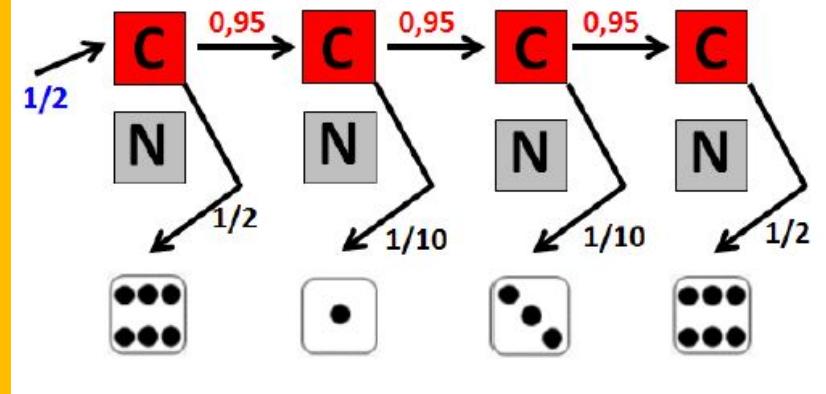
Secuencia Visible: 6136

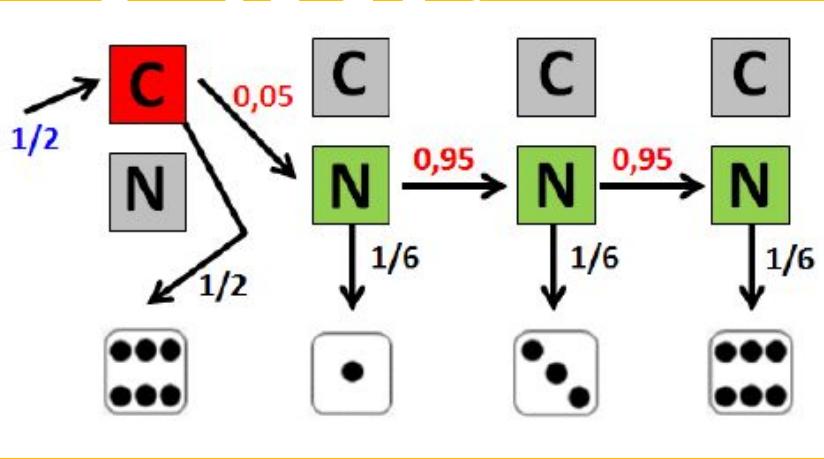
$$P(s,h) = \frac{1}{2} \times \left(\frac{1}{6}\right)^4 \times 0.95^3 = 3,31 \times 10^{-4}$$

Secuencia Oculta: CCCC

Secuencia Visible: 6136

$$P(s,h) = \frac{1}{2} \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{10}\right)^2 \times 0.95^3 = 1,07 \times 10^{-3}$$





Secuencia Oculta: CNNN
Secuencia Visible: 6136

$$P(s,h) = \frac{1}{2} \times \left(\frac{1}{2}\right) \times 0,05 \times \left(\frac{1}{6}\right)^3 \times 0,95^2 = 5,22 \times 10^{-5}$$

#	start a estado 1	estado 1	emisión 1	estado 1 a estado 2	estado 2	emisión 2	estado 2 a estado 3	estado 3	emisión 3	estado 3 a estado 4	estado 4	emisión 4	probabilidad
1	1/2	N	1/6	19/20	N	1/6	19/20	N	1/6	19/20	N	1/6	3,31E-04
2	1/2	N	1/6	19/20	N	1/6	19/20	N	1/6	1/20	C	1/2	5,22E-05
3	1/2	N	1/6	19/20	N	1/6	1/20	C	1/10	1/20	N	1/6	5,50E-07
4	1/2	N	1/6	1/20	C	1/10	1/20	N	1/6	19/20	N	1/6	5,50E-07
5	1/2	C	1/2	1/20	N	1/6	19/20	N	1/6	19/20	N	1/6	5,22E-05
6	1/2	C	1/2	19/20	C	1/10	1/20	N	1/6	19/20	N	1/6	3,13E-05
7	1/2	C	1/2	1/20	N	1/6	1/20	C	1/10	1/20	N	1/6	8,68E-08
8	1/2	C	1/2	1/20	N	1/6	19/20	N	1/6	1/20	C	1/2	8,25E-06
9	1/2	N	1/6	1/20	C	1/10	19/20	C	1/10	1/20	N	1/6	3,30E-07
10	1/2	N	1/6	1/20	C	1/10	1/20	N	1/6	1/20	C	1/2	8,68E-08
11	1/2	N	1/6	19/20	N	1/6	1/20	C	1/10	19/20	C	1/2	3,13E-05
12	1/2	N	1/6	1/20	C	1/10	19/20	C	1/10	19/20	C	1/2	1,88E-05
13	1/2	C	1/2	1/20	N	1/6	1/20	C	1/10	19/20	C	1/2	4,95E-06
14	1/2	C	1/2	19/20	C	1/10	1/20	N	1/6	1/20	C	1/2	4,95E-06
15	1/2	C	1/2	19/20	C	1/10	19/20	C	1/10	1/20	N	1/6	1,88E-05
16	1/2	C	1/2	19/20	C	1/10	19/20	C	1/10	19/20	C	1/2	1,07E-03
													1,63E-03

Diagrama

Intron	Gen 1	Gen 2	Gen R
Dist I	Dist g1	Dist g2		Dist gr
%AI	%Ag1	%Ag2		%Agr
%CI	%Cg1	%Cg2		%Cgr
%TI	%Tg1	%Tg2		%Tgr
%GI	%Gg1	%Gg2		%Ggr

- Sea $X_t = (X_1, \dots, X_N)$ una cadena de Markov a tiempo discreto, que corresponde a los estados observables, $S = \{A, T, G, C\}$ Espacio de Estados.
- sea $Y = (Y_1, \dots, Y_N)$ una cadena de Markov, está jugará el rol de la cadena oculta de Markov y representa: dada una secuencia de nucleótidos observados el posible gen correspondiente.

Ejemplo:

ti	i+1	i+2	i+3	i+4	i+5	i+6	i+7....
Xt	g1	g1	g1	I	g3	g3	g4	I
YtC	A	T	G	T	T	A	C....

$$(Y^*1, \dots, Y^*n) = \operatorname{argmax} \{P(Y_1, \dots, Y_n | X_1, \dots, X_n)\}$$

$$P(r, P, \Lambda) = P(r)P(\lambda|r)P(P|r)$$

Si X^* son los datos observados

$$P(r, P, \Lambda | X^*) \quad (1)$$



$$\pi(\mathbf{y}|\boldsymbol{\Lambda}, \mathbf{P}, \mathbf{r}) \propto f = \prod_{i \in G} \prod_{j \in G} \lambda_{ij}^{m_{ij}} \prod_{K \in G} \prod_{l \in O} P_{kl}^{n_{kl}}$$

$$m_{ij} = \sum_t \mathbb{1}(x_{t-1} = i, x_t = j) \quad n_{kl} = \sum_t \mathbb{1}(y_t = l, x_t = k)$$

- ✖ La anterior es la función de verosimilitud, el problema se reduce al calcular el argumento donde le maximiza la verosimilitud dado los datos.

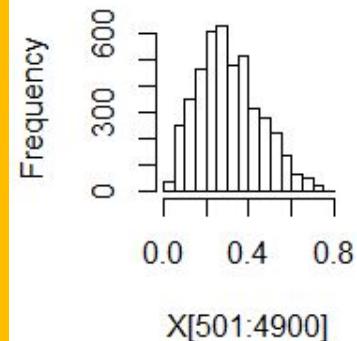
- ✖ Obtener muestras aleatorias de (1) para poder maximizar la función de verosimilitud resulta ser difícil, por esta razón se usará algoritmo de simulación de tipo MCMC

Método MCMC

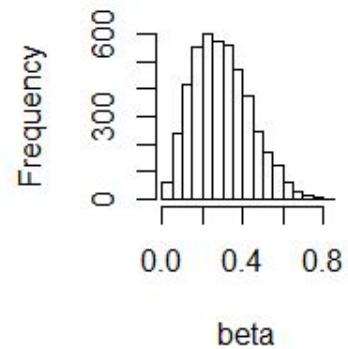
Es cualquier método que genere una cadena de Markov ergódica X_t con distribución estacionaria f .

Estos métodos se pueden usar para optimizar

Histogram of X[501:49]



Histogram of beta



Muestreo de Gibbs

Para Generar

$$p(\theta_1, \theta_2, \dots, \theta_k | x)$$

Se necesita conocer

$$p(\theta_1 | \theta_2, \dots, \theta_k, x)$$

$$p(\theta_1 | \theta_2, \theta_3, \dots, \theta_k, x)$$

$$p(\theta_2 | \theta_1, \theta_3, \dots, \theta_k, x)$$

....

$$p(\theta_k | \theta_1, \theta_2, \dots, \theta_{k-1}, x)$$

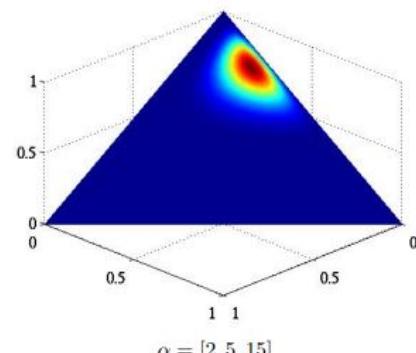
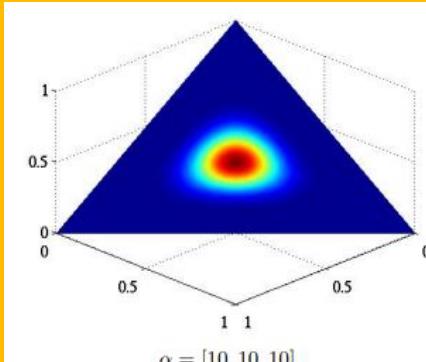
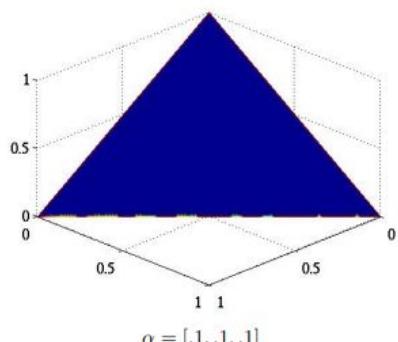
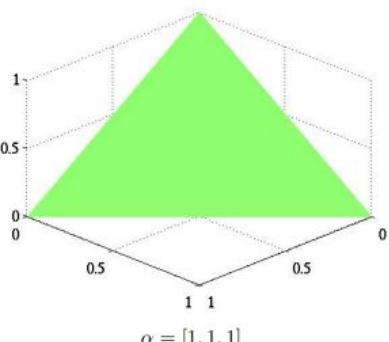
El muestreo de gibbs es un caso particular de metropolis hasting, este no necesita una distribución instrumental.

Solo se necesita poder generar muestras de la distribución posterior condicional de cada uno de los parámetros individuales.

Distribución de Dirichlet

Sea $Q = [Q_1, Q_2, \dots, Q_k]$ tal que $0 \leq Q_i$ para $i = 1, \dots, n$ y $\sum_{i=1}^k Q_i = 1$ y $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_k]$ con $\alpha_i > 0$ para cada i y defina $\alpha_0 = \sum_{i=1}^k \alpha_i$, entonces se dice $Q \sim Dirichlet(\alpha)$

$$f(q; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k q_i^{\alpha_i-1}$$



Muestreo de Gibbs

Condicionales 'fáciles'

Núm. estados
ocultos

$r | \text{Datos} \sim \text{poisson}(a), a \in \mathbb{R}$

Probabilidades
de transición
de los estados
ocultos

$\Lambda | \text{Datos} \sim \text{Dirichlet}(\Lambda), \Lambda \in \mathbb{R}^r$

Probabilidades
de transición
de los estados
observables.

$P | \text{datos} \sim \text{Dirichlet}(D), D \in \mathbb{R}^4$

Algoritmo a implementar

1. Inicializar con la base de datos X^* , Y^* y r .
2. Elegir r , X^* o Y^* con probabilidad $\frac{1}{3}$, al variable obtenido le llamamos θ .
3. Simular $\theta|$ datos al nuevo valor le llamamos θ^* y Θ a los otros dos parámetros.
4. Valuar en f , si $f(\theta^*, \Theta) > f(X^*, Y^*, r)$ ent $\theta = \theta^*$, si no hacemos $\theta = \theta^*$ con probabilidad $f(\theta^*, \Theta) / f(X^*, Y^*, r)$ de lo contrario $\theta = \theta$.
5. Volver a 2.

Estimación de parámetros de las distribuciones condicionales

$D_i \in R^4, i = \{1, \dots, r\}$

Media empírica de A, T, C, G correspondiente al Gen estimado.

$a \in R$

A el número de genes encontrados de forma empírica.

$A_i \in R^r, i = \{1, \dots, r\}$

Media empírica del número de exones correspondiente a un gen y distribución empírica del número de intrones.

Estimación de parámetros

Base de datos

Se obtendrá de NCBI (*National Center for Biotechnology Information*).

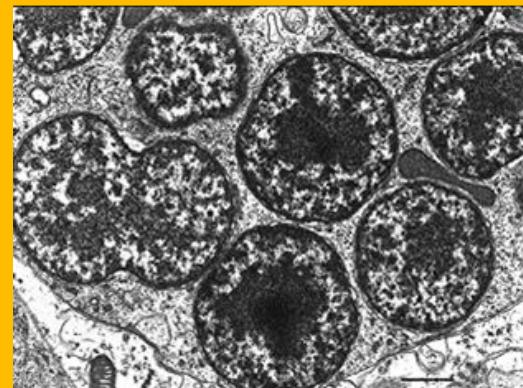
Almacena y constantemente actualiza la información referente a secuencias genómicas en GenBank

Genoma de prueba

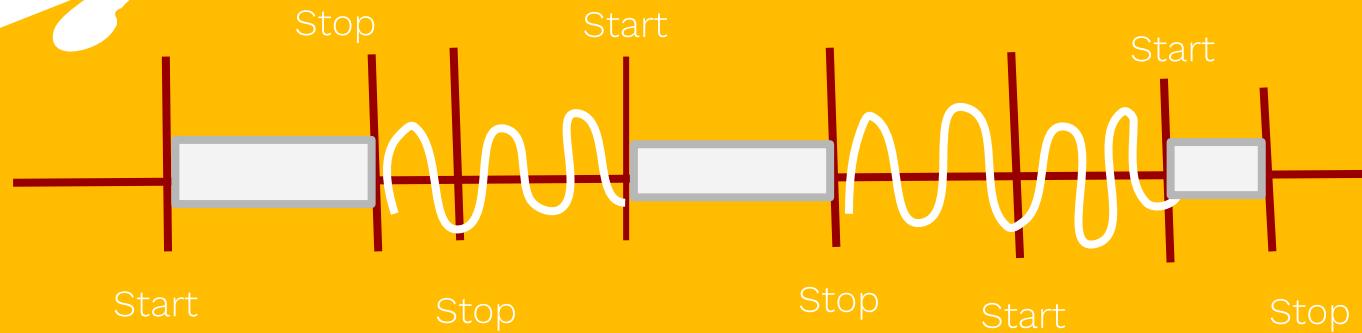
El genoma que se usará es de la bacteria *Buchnera aphidicola* pues tiene un genoma ‘pequeño’.

Núm. de nucleotidos:
655725

Num. de protein genes: 574
Num. de RNA genes: 36



Estimación de matriz de transición de estados observable

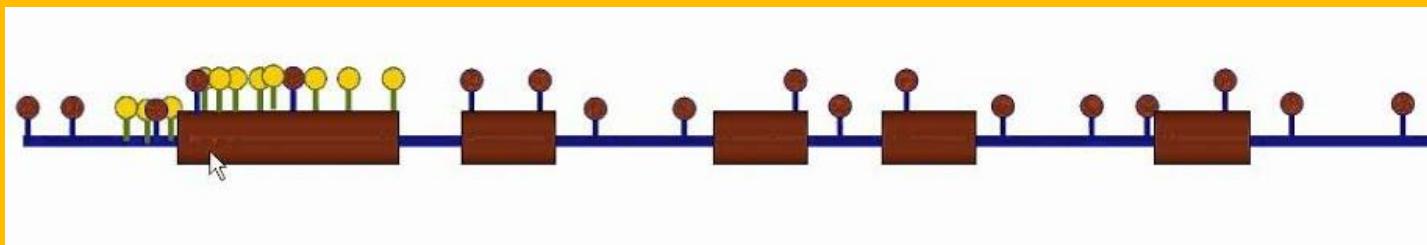


Intron



Exones

Estimación de matriz de transición de estados ocultos y el número de estados ocultos



Exones

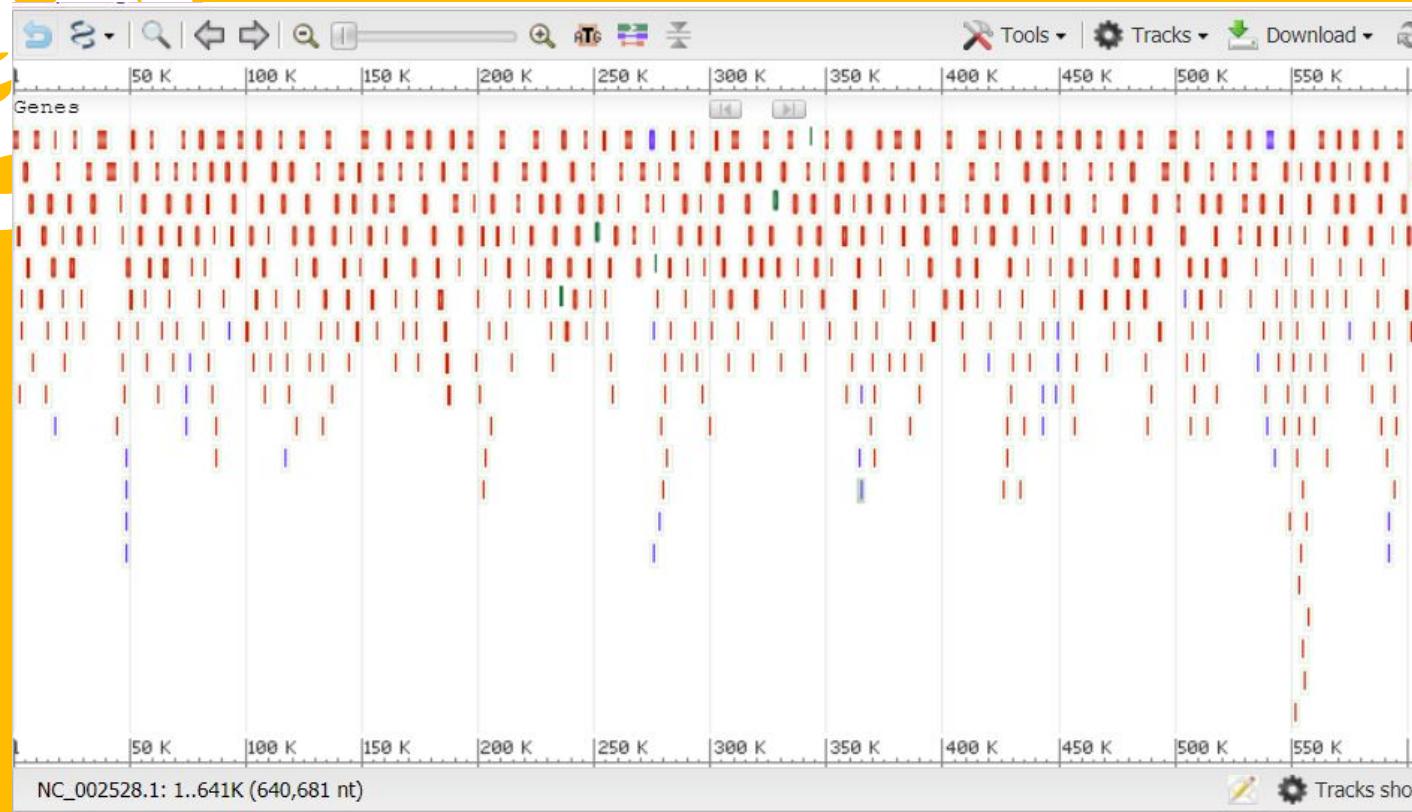


Intrones

$r =$ número de genes
encontrados

Resultados

Whoa!



Prueba 1, Tipo 1

Largo de la base: 655 725

Genes: 574

Porción de ADN analizada:
9216

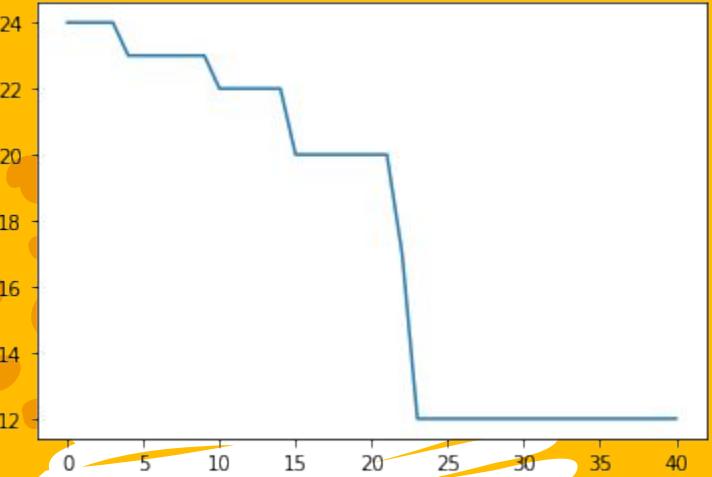
Genes Esperados:

8.06738

Genes Estimados:

14

Iteraciones: 40



Prueba 1.2, Tipo 1

Largo de la base: 655

725

Genes: 574

**Porción de ADN
analizada:** 9216

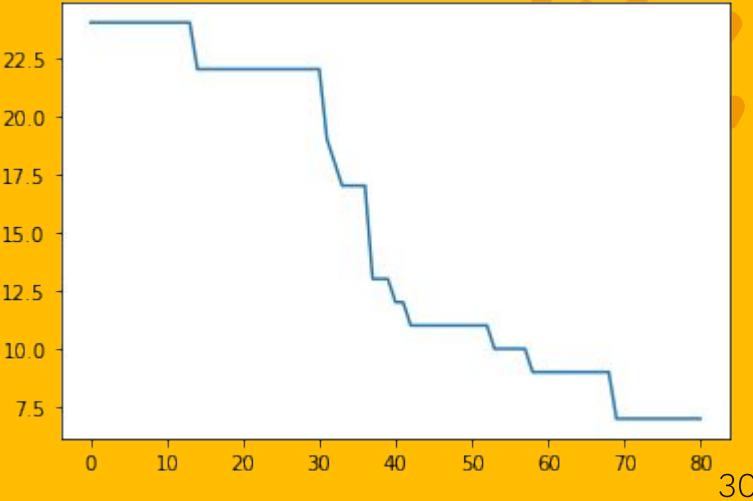
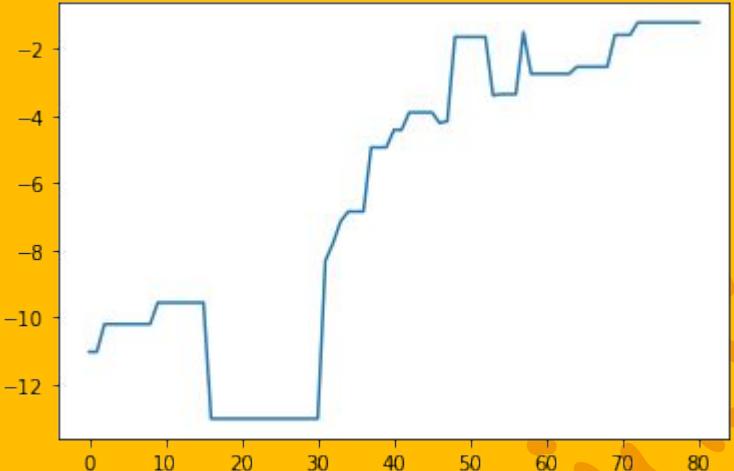
Genes Esperados:

8.06738

Genes Estimados:

7

Iteraciones: 80



Prueba 1.2, Tipo 2

Largo de la base: 655 725

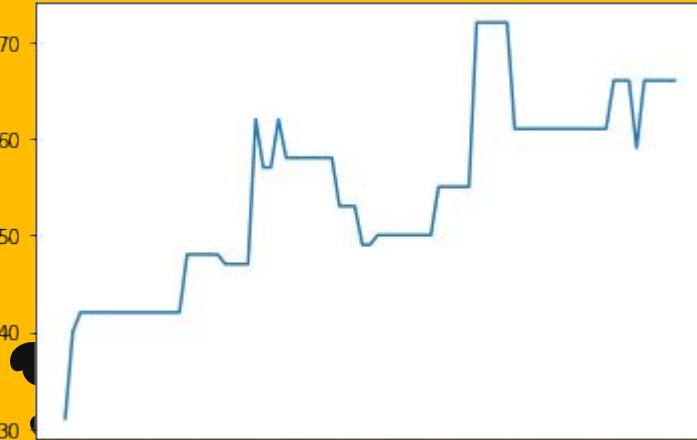
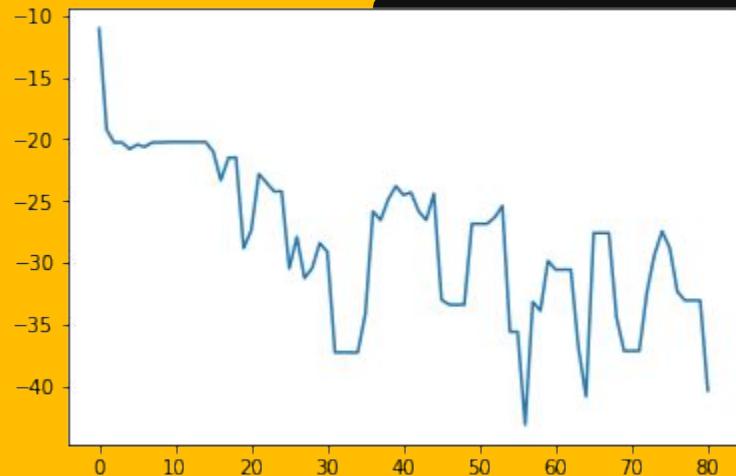
Genes: 574

Porción de ADN analizada:
9216

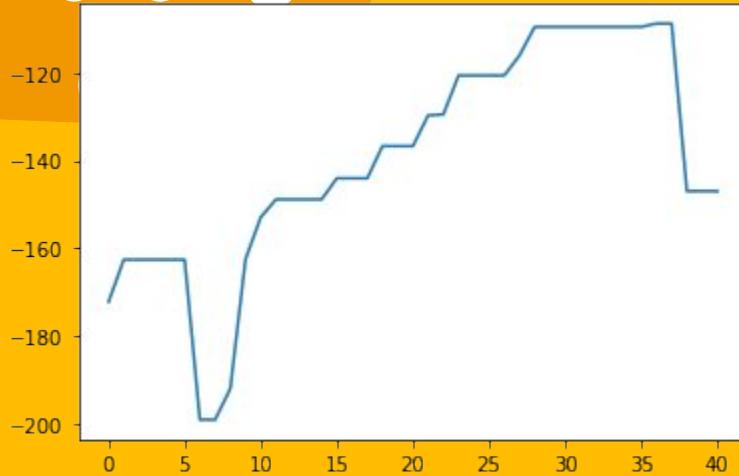
Genes Esperados:
8.06738

Genes Estimados:
68

Iteraciones: 80



Prueba 2, tipo 1



Largo de la base: 655 725

Genes: 574

Porción de ADN analizada:

149583

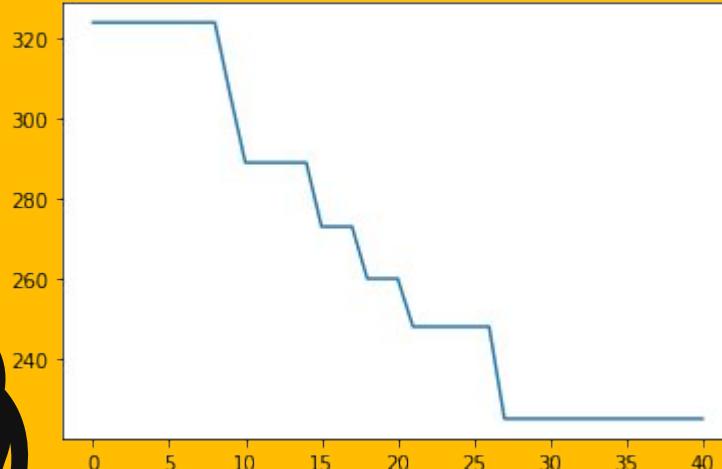
Genes Esperados:

130

Genes Estimados:

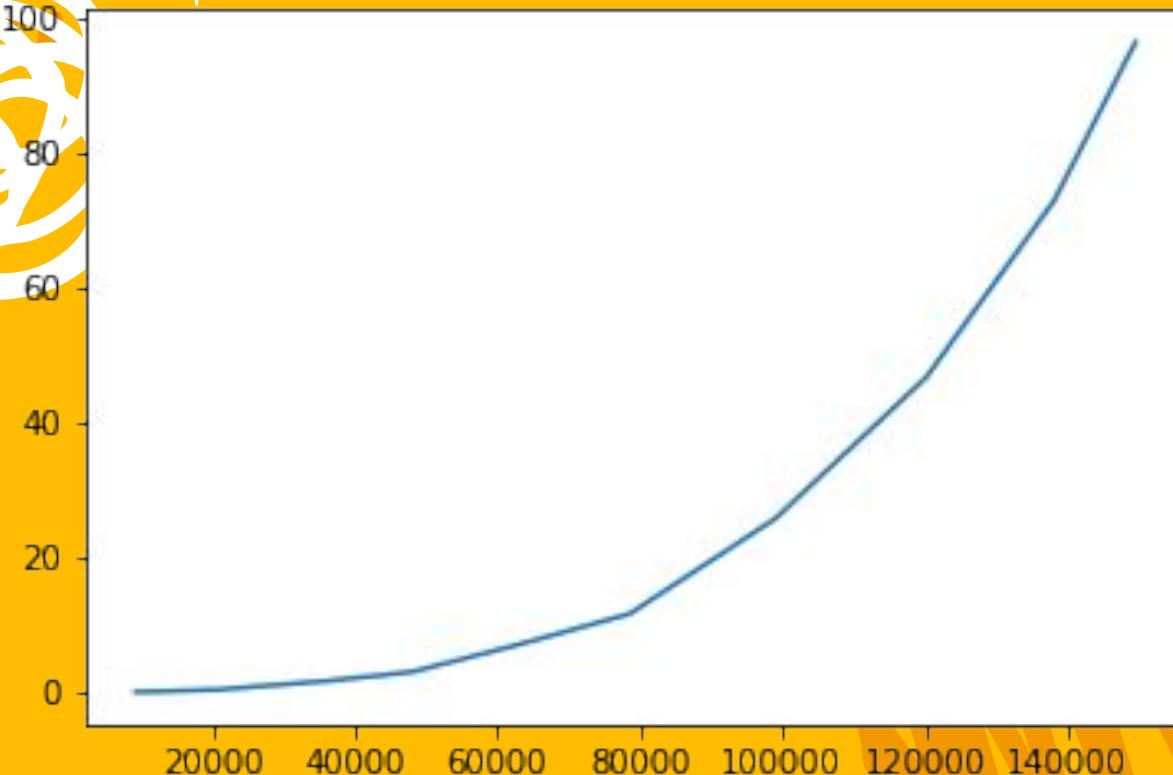
225

Iteraciones: 40



Tiempo

Tiempo de ejecucion para recalcular
los parametros





Algoritmo 1

Es bueno, pero le falta el aumento



Algoritmo 2

Presenta mucha variabilidad

Optimidad

Se necesita un gran poder de cómputo para hacer un análisis completo.



Gracias!

