



CURE

Centro Universitario
Regional del Este



UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Escuela Regional de Ecología Microbiana de Sistemas Acuáticos Edición 2023

Grupo de Ecología Microbiana de Sistemas Acuáticos,
Centro Universitario Regional del Este,
Universidad de la República

Bioinformática aplicada al análisis de datos de metabarcoding de ADN ambiental

Teórico 3

6 de diciembre 2023

Docentes: Paula Huber, Daiana Mir, Luciana Griffero, Cecilia Alonso, Juan Zanetti, Emiliano Pereira

- **Anotación taxonómica**
- **Normalización y distancias**

- **Anotación taxonómica**
- **Normalización y distancias**

Anotación taxonómica

Anotación taxonómica

- Taxonomía

La taxonomía: clasificación ordenada y jerárquica.



En función de las relaciones de parentesco entre organismos (mayoritariamente).

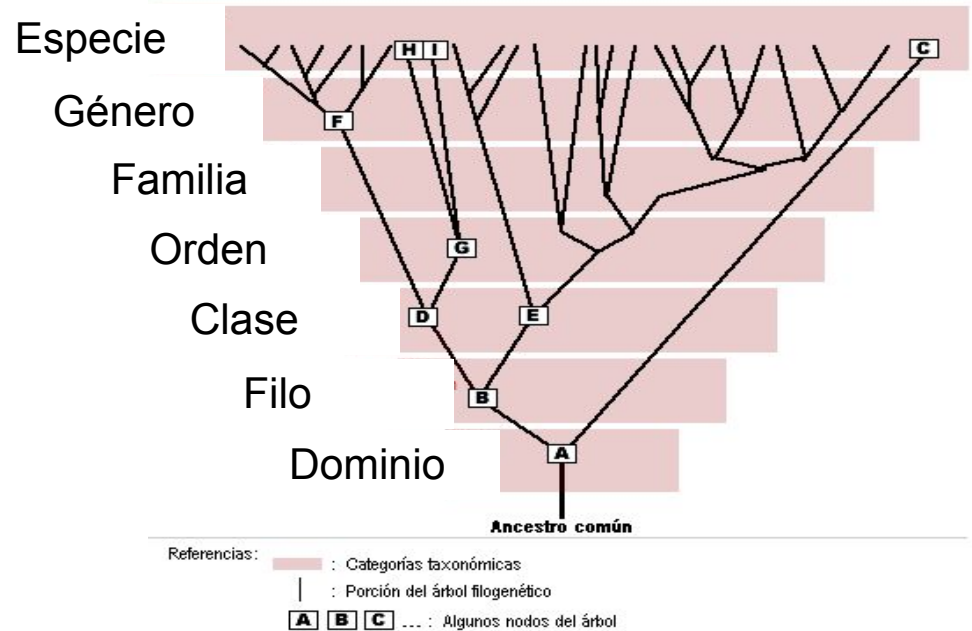


Necesidad de nombrar, ordenar y encontrar patrones para entender la realidad.



Anotación taxonómica

- Categorías taxonómicas



Anotación taxonómica

- Categorías taxonómicas

Canis lupus

Canis

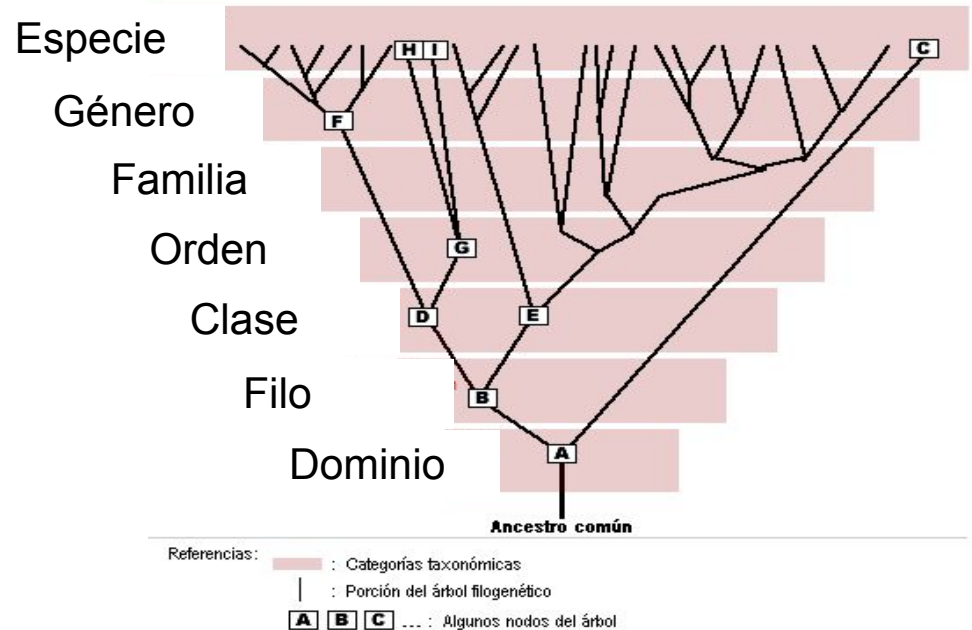
Canidae

Carnivora

Mammalia

Chordata

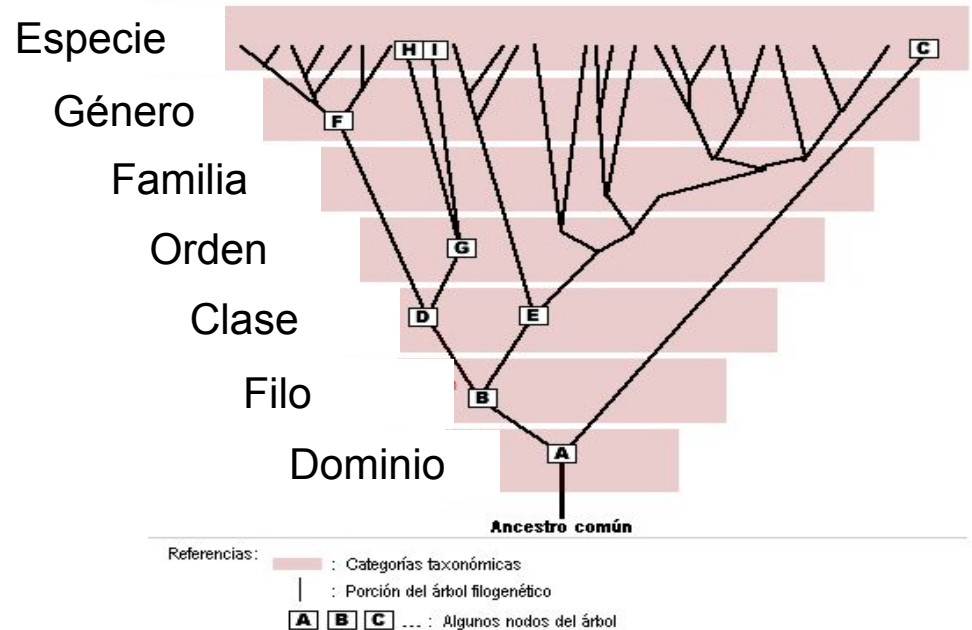
Eukarya



Anotación taxonómica

- Categorías taxonómicas

Escherichia coli
Escherichia
Enterobacteriaceae
Enterobacterales
Gammaproteobacteria
Proteobacteria
Bacteria



Anotación taxonómica

Reconocimiento de unidades taxonómicas: ¿Qué es una especie?

Anotación taxonómica

Especie biológica:

"Grupos de poblaciones naturales actual o potencialmente capaces de entrecruzamiento que se encuentran aisladas reproductivamente de otros grupos similares"

Anotación taxonómica

Especie biológica:

"Grupos de poblaciones naturales actual o potencialmente capaces de entrecruzamiento que se encuentran aisladas reproductivamente de otros grupos similares"



Basado en la reproducción sexual

Anotación taxonómica

Especie biológica:

"Grupos de poblaciones naturales actual o potencialmente capaces de entrecruzamiento que se encuentran aisladas reproductivamente de otros grupos similares"



No es universal

Anotación taxonómica

Especie biológica:

"Grupos de poblaciones naturales actual o potencialmente capaces de entrecruzamiento que se encuentran aisladas reproductivamente de otros grupos similares"



Vertebrados, Invertebrados (parcialmente) y Plantas (parcialmente).



Invertebrados (parcialmente), Plantas (parcialmente), Hongos, Líquenes, Algas y **Procariotas**.

Anotación taxonómica

- | | |
|----------------------------------|---------------------|
| 1. Agamospecies | 12. Hennigian |
| 2. Biological | 13. Internodal |
| 3. Cohesion | 14. Morphological |
| 4. Cladistic | 15. Non-dimensional |
| 5. Composite | 16. Phenetic |
| 6. Ecological | 17. Polythetic |
| 7. Evolutionary Significant Unit | 18. Phylogenetic |
| 8. Evolutionary | 19. Recognition |
| 9. Genealogical Concordance | 20. Successional |
| 10. Genotypic Cluster Definition | 21. Genetic |
| 11. Reproductive competition | 22. Taxonomic |

Anotación taxonómica

- **Conceptos más aplicados a procariotas**

1. **Concepto de especie evolutiva:** Linaje poblacional de antecesoros – descendientes que mantienen su identidad de otros linajes y tienen su propia tendencia evolutiva y destino histórico.
2. **Concepto filogenético de especie:** Grupo irreducible de organismos, diagnósticamente distinguibles de otros grupo semejantes y dentro del cual existe un patrón parental de ascendencia y descendencia.
3. **Concepto fenético de especie:** una especie es un grupo de organismos que son fenotípicamente similares y que parecen diferentes de otros grupos de organismos.

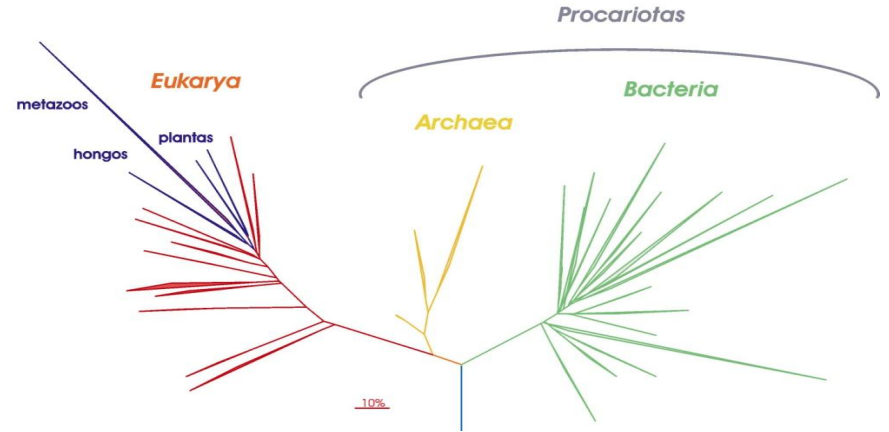
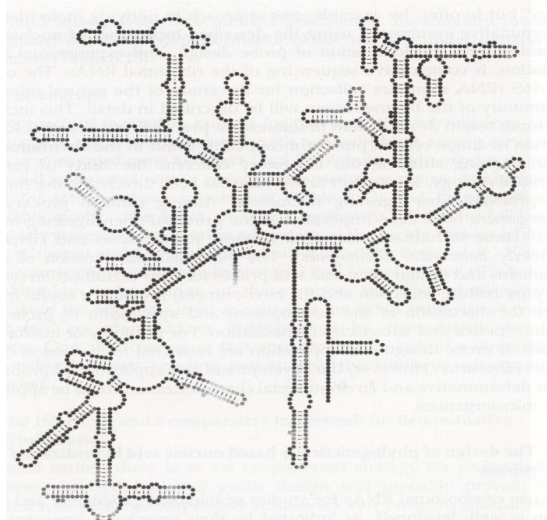
Anotación taxonómica

- **Conceptos más aplicados a procariotas**

Conceptos universales

1. Concepto de especie evolutiva: Linaje poblacional de antecesoros – descendientes que mantiene su identidad de otros linajes y tiene su propia tendencia evolutiva y destino histórico.
2. Concepto filogenético de especie: Grupo irreductible de organismos, diagnósticamente distinguibles de otros grupo semejantes y dentro del cual existe un patrón parental de ascendencia y descendencia.
3. Concepto fenético de especie: una especie es un grupo de organismos que son fenotípicamente similares y que parecen diferentes de otros grupos de organismos.

Anotación taxonómica

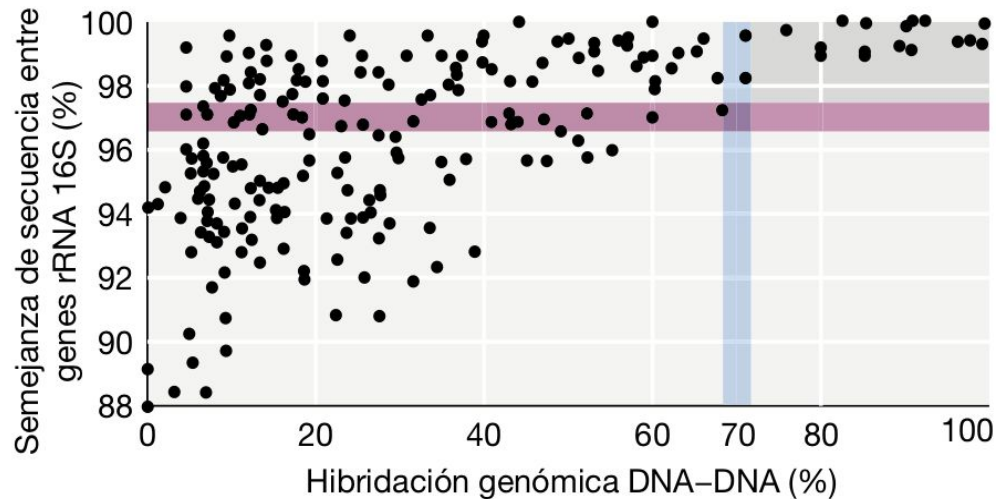


El RNAr 16S se ha convertido en la molécula de referencia para:

- Reconstruir la genealogía
- Construir el sistema de clasificación
- Identificar diversidad ambiental

Anotación taxonómica

En general dos organismos con >97% identidad pertenecen a especies distintas
Lo contrario no es cierto!!!



El gen para ARNr 16S NO tiene poder de discriminar perfectamente entre especies

Anotación taxonómica

Table 1 | **Taxonomic thresholds of bacteria and archaea***

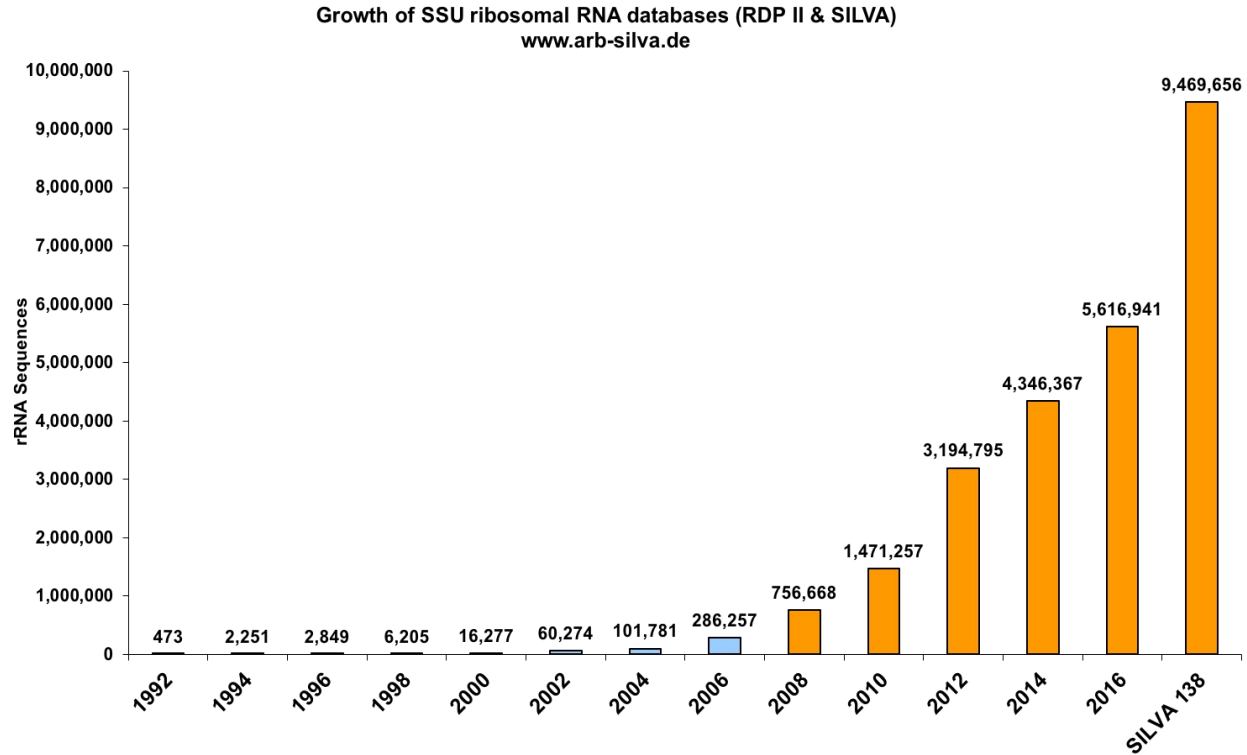
	Genus	Family	Order	Class	Phylum
Number of taxa	568	201	85	39	23
Median sequence identity	96.4% (96.2, 96.55)	92.25% (91.65, 92.9)	89.2% (88.25, 90.1)	86.35% (84.7, 87.95)	83.68% (81.6, 85.93)
Minimum sequence identity	94.8% (94.55, 95.05)	87.65% (86.8, 88.4)	83.55% (82.25, 84.8)	80.38% (78.55, 82.5)	77.43% (74.95, 79.9)
Threshold sequence identity	94.5%	86.5%	82.0%	78.5%	75.0%

Anotación taxonómica

Otros genes con información filogenética (diferenciar especies estrechamente emparentadas):

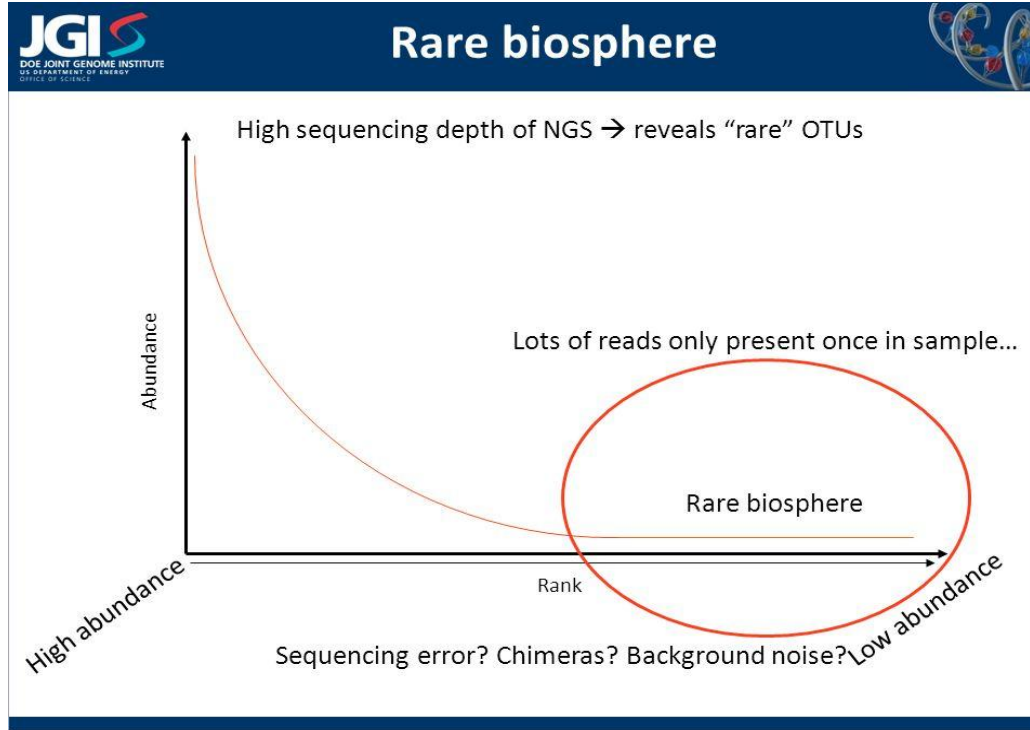
- . **RecA**, que codifica una proteína recombinasa
- . **gyrB**, que codifica una girasa del DNA

Anotación taxonómica

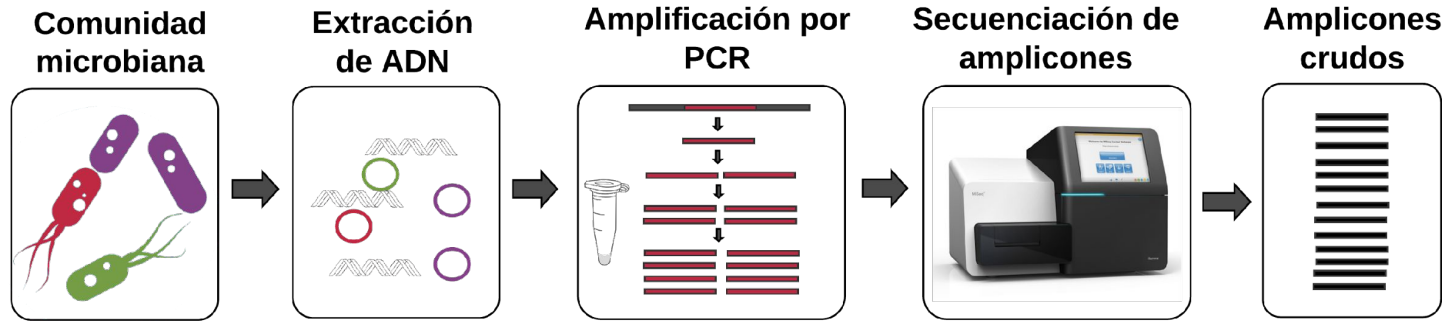


Anotación taxonómica

- Secuenciación masiva

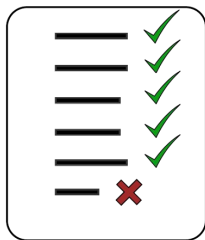


Anotación taxonómica

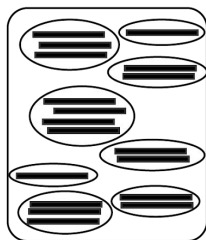


Anotación taxonómica

Control de
calidad



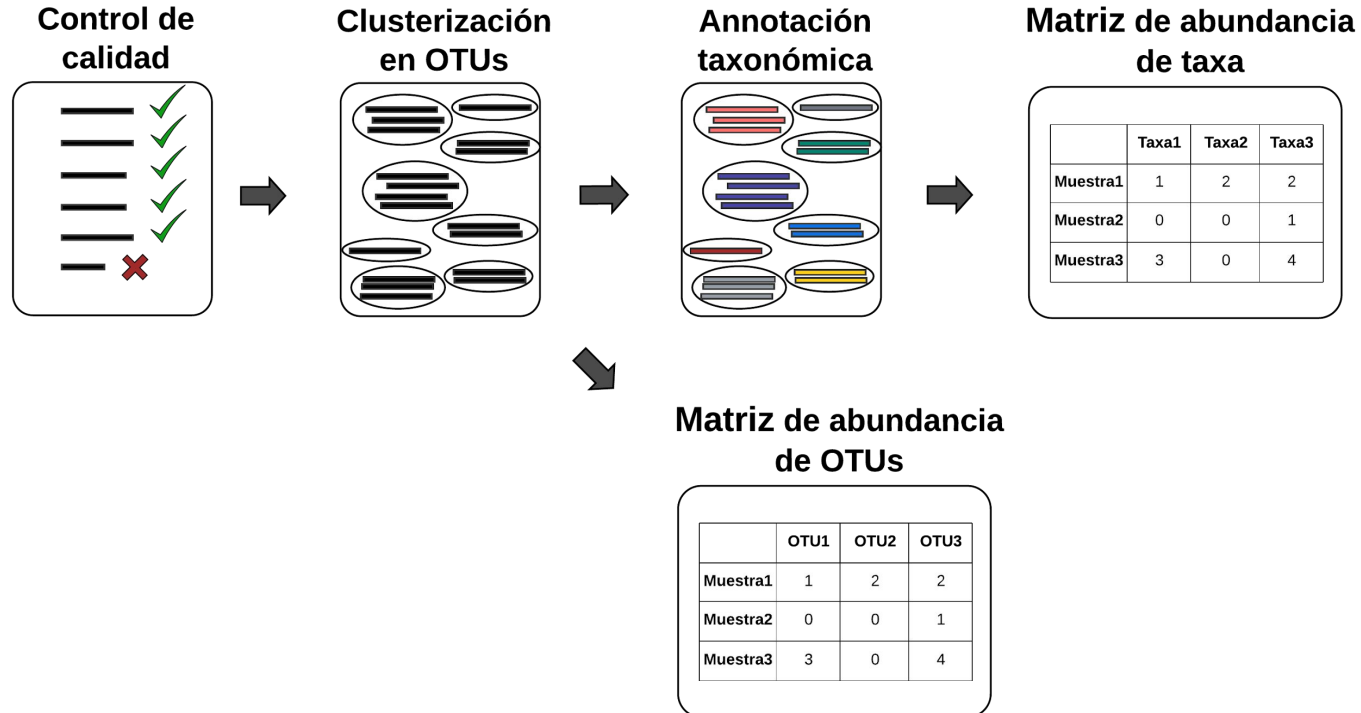
Clusterización
en OTUs



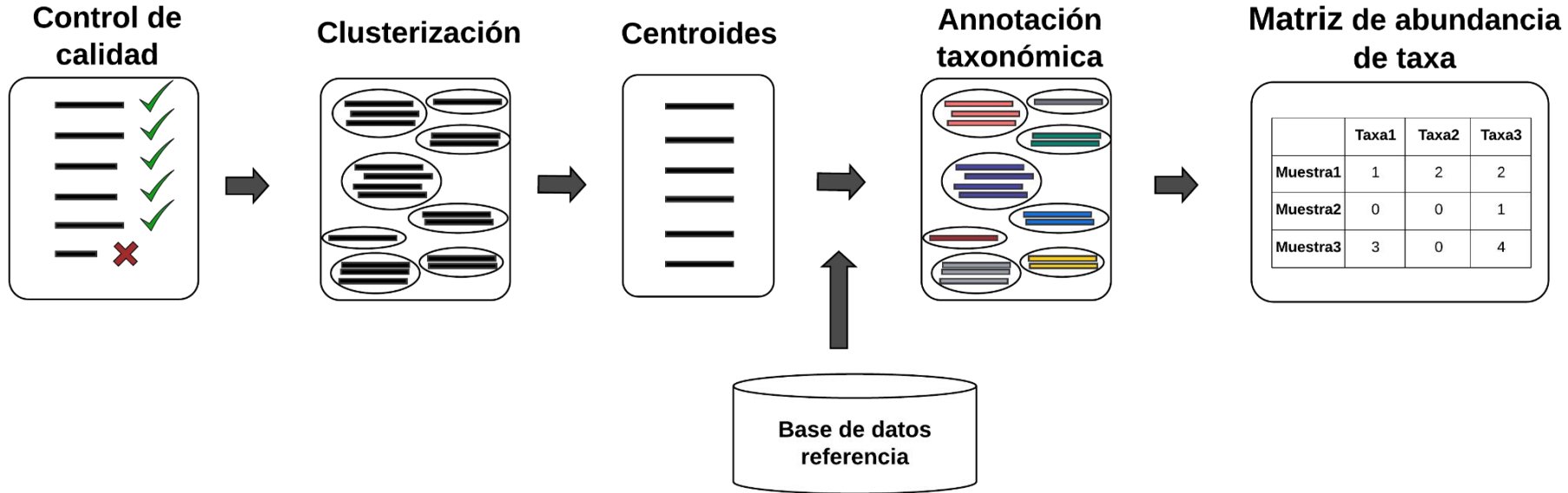
Matriz de abundancia
de OTUs

	OTU1	OTU2	OTU3
Muestra1	1	2	2
Muestra2	0	0	1
Muestra3	3	0	4

Anotación taxonómica

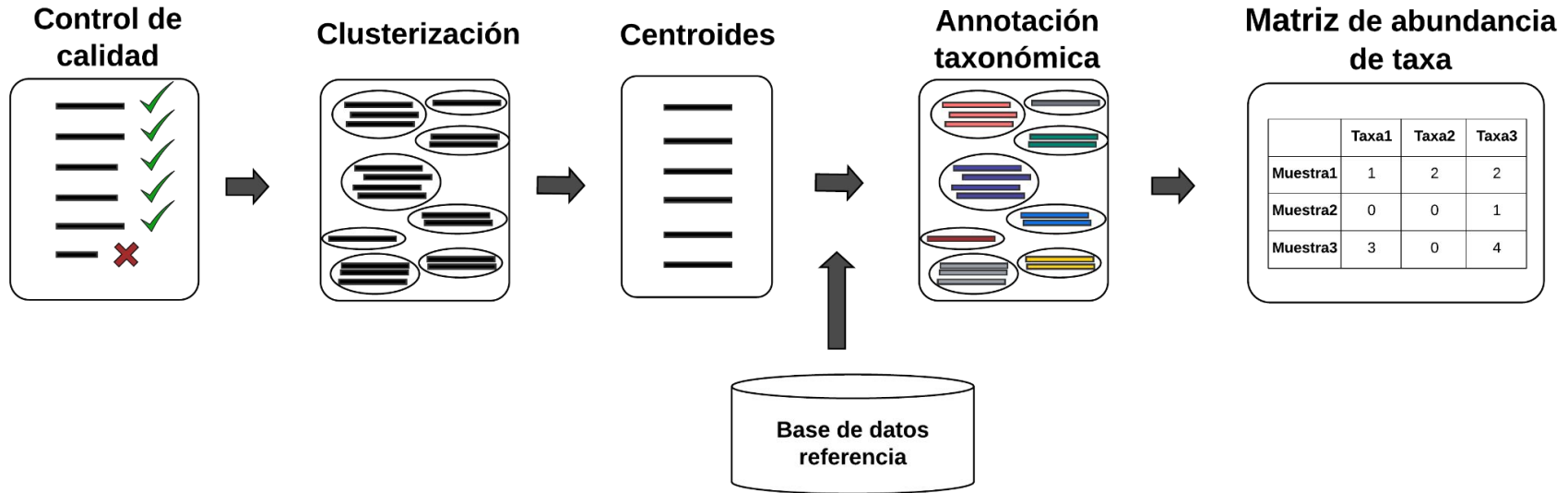


Anotación taxonómica



Anotación taxonómica

Comparación de centroides vs. secuencias referencia (con taxonomía conocida).



Ej., Silva o Greengenes

Anotación taxonómica



Comunmente, las anotaciones taxonómicas (ej., SILVA y Greengenes) son predicciones obtenidas mediante análisis computacionales y manuales, basadas en reconstrucciones filogenéticas.

Anotación taxonómica

Herramientas:

- BLAST (SILVAngs)
- VSEARCH
- Clasificador de Bayes Ingenuo (Naive Bayes Classifiers (NBC))

Anotación taxonómica

- **BLAST (SILVAngs)**

Anotación taxonómica

- **BLAST (SILVAngs)**

Un ejemplo de aplicación de BLAST para la anotación taxonómica es SILVAngs (<https://ngs.arb-silva.de/silvangs/>).

SILVAngs es un servicio de análisis de amplicones del gen de ARN ribosomal.

Anotación taxonómica

- **BLAST (SILVAngs)**

SILVAngs incluye procesamiento completo de datos de amplicones:


- Alineamiento
- Control de la calidad
- Dereplicación
- Clusterización
- Anotación taxonómica

Anotación taxonómica

- **BLAST (SILVAngs)**

SILVAngs incluye procesamiento completo de datos de amplicones:

- Alineamiento
- Control de la calidad
- Dereplicación
- Clusterización
- Anotación taxonómica



Permite un análisis completo sin correr un comando!

Anotación taxonómica

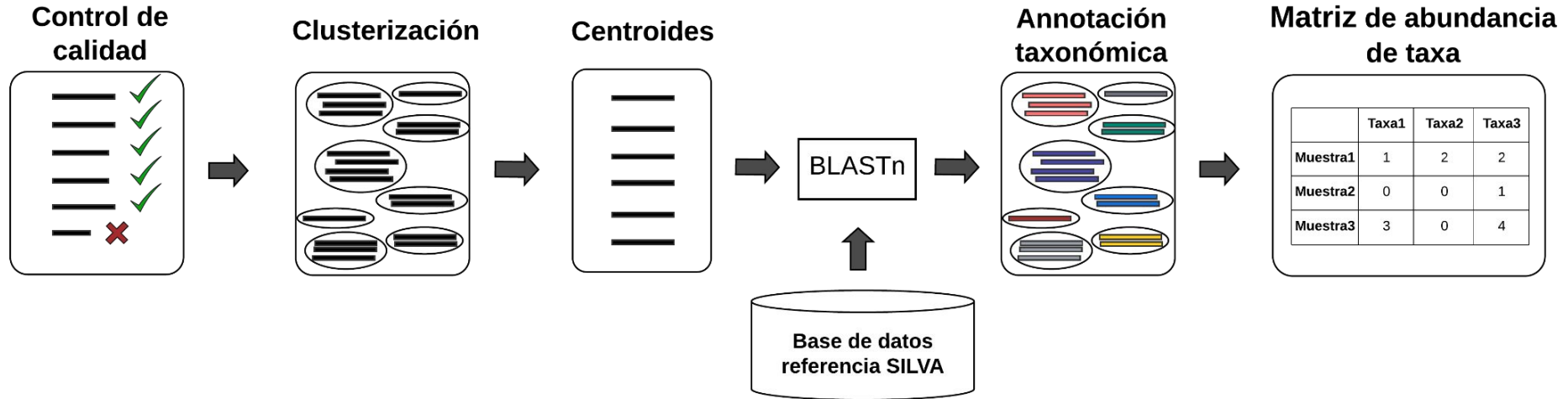
- **BLAST (SILVAngs)**

SILVAngs incluye procesamiento completo de datos de amplicones:

- Alineamiento
- Control de la calidad
- Dereplicación
- Clusterización
- **Anotación taxonómica**

Anotación taxonómica

- **BLAST (SILVAngs)**



Anotación taxonómica

- **BLAST (SILVAngs)**

Al ser SILVA es un base de datos muy completa y curada, aplicar una metodología basada en el primer *hit* BLAST es suficiente (según los autores de SILVAngs).

- Sólo los *hits* significativos son considerados.
- Todo lo demás pasa a ser “*No relative*”.

Anotación taxonómica

- **BLAST (SILVAngs)**

Umbral determinado empíricamente:

$$\frac{(\text{identidad de secuencia} + \text{cobertura de alineación})}{2} \geq 93$$

Esto permite filtrar *hits* que tienen baja identidad o que sólo se alinean parcialmente con la secuencia referencia.

Anotación taxonómica

- **VSEARCH**

Anotación taxonómica

- **VSEARCH**

SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences

Robert C. Edgar

Independent Investigator

Tiburon, California, USA.

robert@drive5.com

Anotación taxonómica

- **VSEARCH**

Algoritmo SINTAX: un clasificador taxonómico simple no-Bayesiano para genes 16S de ARNr y secuencias ITS.

Anotación taxonómica

- **VSEARCH**

Algoritmo SINTAX: un clasificador taxonómico simple no-Bayesiano para genes 16S de ARNr y secuencias ITS.



Basado en similitud de k-mers (sin alineamientos entre secuencias).

Anotación taxonómica

- **VSEARCH** (SINTAX)

Secuencia problema (*query*) Q.

A	G	G	T	A	C	C	A	T	G	T	A	C	C	G	T	T	A	G	G	A	A	T	T	C	G	A	C	T	A	G	C
G	G	T	A	C	C	A	T																								
	G	T	A	C	C	A	T	G																							
		T	A	C	C	A	T	G	T																						
			A	C	C	A	T	G	T	A																					
				C	C	A	T	G	T	A	C																				

...

A	A	T	T	C	G	A	C																								
	A	T	T	C	G	A	C	T																							
		T	T	C	G	A	C	T	A																						
			T	C	G	A	C	T	T	G																					
				C	G	A	C	T	T	G	C																				

W(Q): todos los k-mers de Q.

Anotación taxonómica

- **VSEARCH** (SINTAX)

Secuencia problema (*query*) Q.

A	G	G	T	A	C	C	A	T	G	T	A	C	C	G	T	T	A	G	G	A	A	T	T	C	G	A	C	T	A	G	C
G	G	T	A	C	C	A	T																								
	G	T	A	C	C	A	T	G																							
		T	A	C	C	A	T	G	T																						
			A	C	C	A	T	G	T	A																					
				C	C	A	T	G	T	A	C																				

...

A	A	T	T	C	G	A	C																							
	A	T	T	C	G	A	C	T																						
		T	T	C	G	A	C	T	A																					
			T	C	G	A	C	T	T	G																				
				C	G	A	C	T	T	G	C																			

W(Q): todos los
8-mers de Q.

Anotación taxonómica

- **VSEARCH (SINTAX)**

W(Q)

G	G	T	A	C	C	A	T
G	T	A	C	C	A	T	G
T	A	C	C	A	T	G	T
A	C	C	A	T	G	T	A
C	C	A	T	G	T	A	C

■ ■ ■

A	A	T	T	C	G	A	C
A	T	T	C	G	A	C	T
T	T	C	G	A	C	T	A
T	C	G	A	C	T	T	G
C	G	A	C	T	T	G	C

Anotación taxonómica

- **VSEARCH (SINTAX)**

W(Q)

G	G	T	A	C	C	A	T
G	T	A	C	C	A	T	G
T	A	C	C	A	T	G	T
A	C	C	A	T	G	T	A
C	C	A	T	G	T	A	C

...

A	A	T	T	C	G	A	C
A	T	T	C	G	A	C	T
T	T	C	G	A	C	T	A
T	C	G	A	C	T	T	G
C	G	A	C	T	T	G	C

Seleccionamos
s k-mers tomados al
azar.

Anotación taxonómica

- **VSEARCH (SINTAX)**

W(Q)

G	G	T	A	C	C	A	T
G	T	A	C	C	A	T	G
T	A	C	C	A	T	G	T
A	C	C	A	T	G	T	A
C	C	A	T	G	T	A	C

...

A	A	T	T	C	G	A	C
A	T	T	C	G	A	C	T
T	T	C	G	A	C	T	A
T	C	G	A	C	T	T	G
C	G	A	C	T	T	G	C

Seleccionamos
s k-mers tomados al
azar.

w_s(Q)

G	T	A	C	C	A	T	G
T	A	C	C	A	T	G	T
C	C	A	T	G	T	A	C

...

A	A	T	T	C	G	A	C
T	T	C	G	A	C	T	A
C	G	A	C	T	T	G	C

Anotación taxonómica

- **VSEARCH (SINTAX)**

W(Q)

G	G	T	A	C	C	A	T
G	T	A	C	C	A	T	G
T	A	C	C	A	T	G	T
A	C	C	A	T	G	T	A
C	C	A	T	G	T	A	C

...

A	A	T	T	C	G	A	C
A	T	T	C	G	A	C	T
T	T	C	G	A	C	T	A
T	C	G	A	C	T	T	G
C	G	A	C	T	T	G	C

Seleccionamos
32 8-mers tomados
al azar.

w₃₂(Q)

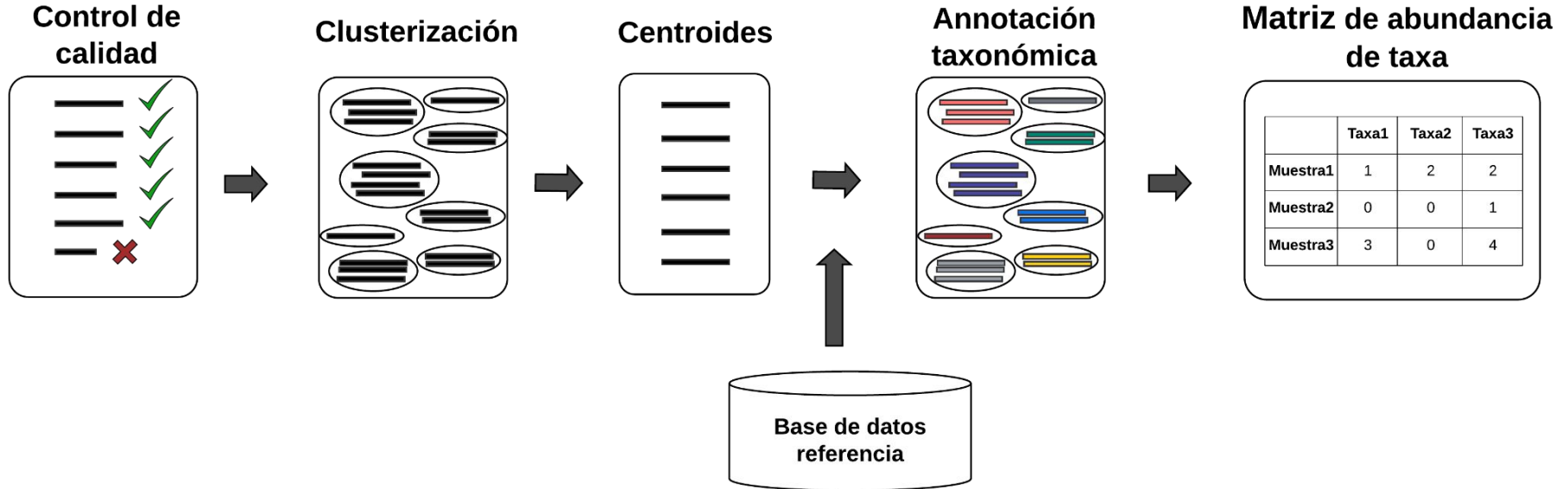
G	T	A	C	C	A	T	G
T	A	C	C	A	T	G	T
C	C	A	T	G	T	A	C

...

A	A	T	T	C	G	A	C
T	T	C	G	A	C	T	A
C	G	A	C	T	T	G	C

Anotación taxonómica

- **VSEARCH (SINTAX)**



Anotación taxonómica

- **VSEARCH (SINTAX)**



R secuencias referencia.

Secuencia referencia 1	<table><tr><td>A</td><td>G</td><td>G</td><td>T</td><td>A</td><td>C</td><td>C</td><td>A</td><td>T</td><td>G</td><td>T</td><td>A</td><td>C</td><td>C</td><td>A</td><td>T</td><td>T</td><td>A</td><td>G</td><td>G</td><td>A</td><td>A</td><td>G</td><td>T</td><td>C</td><td>G</td><td>A</td><td>T</td><td>T</td><td>A</td><td>G</td><td>C</td></tr></table>	A	G	G	T	A	C	C	A	T	G	T	A	C	C	A	T	T	A	G	G	A	A	G	T	C	G	A	T	T	A	G	C	Taxon1 (Especie - Genero - Familia - Clase - Orden - Filum)
A	G	G	T	A	C	C	A	T	G	T	A	C	C	A	T	T	A	G	G	A	A	G	T	C	G	A	T	T	A	G	C			
Secuencia referencia 2	<table><tr><td>G</td><td>G</td><td>G</td><td>A</td><td>A</td><td>C</td><td>C</td><td>G</td><td>T</td><td>G</td><td>T</td><td>A</td><td>C</td><td>C</td><td>G</td><td>T</td><td>T</td><td>A</td><td>G</td><td>G</td><td>A</td><td>A</td><td>T</td><td>G</td><td>C</td><td>G</td><td>A</td><td>C</td><td>T</td><td>A</td><td>G</td><td>C</td></tr></table>	G	G	G	A	A	C	C	G	T	G	T	A	C	C	G	T	T	A	G	G	A	A	T	G	C	G	A	C	T	A	G	C	Taxon2
G	G	G	A	A	C	C	G	T	G	T	A	C	C	G	T	T	A	G	G	A	A	T	G	C	G	A	C	T	A	G	C			
Secuencia referencia 3	<table><tr><td>T</td><td>G</td><td>G</td><td>T</td><td>C</td><td>C</td><td>C</td><td>A</td><td>T</td><td>T</td><td>T</td><td>A</td><td>C</td><td>C</td><td>G</td><td>T</td><td>T</td><td>A</td><td>G</td><td>G</td><td>A</td><td>A</td><td>T</td><td>T</td><td>C</td><td>G</td><td>G</td><td>C</td><td>T</td><td>A</td><td>G</td><td>C</td></tr></table>	T	G	G	T	C	C	C	A	T	T	T	A	C	C	G	T	T	A	G	G	A	A	T	T	C	G	G	C	T	A	G	C	Taxon3
T	G	G	T	C	C	C	A	T	T	T	A	C	C	G	T	T	A	G	G	A	A	T	T	C	G	G	C	T	A	G	C			
...																																		
Secuencia referencia 3	<table><tr><td>A</td><td>G</td><td>T</td><td>T</td><td>T</td><td>C</td><td>C</td><td>A</td><td>T</td><td>G</td><td>T</td><td>A</td><td>C</td><td>C</td><td>G</td><td>T</td><td>T</td><td>A</td><td>T</td><td>G</td><td>A</td><td>A</td><td>T</td><td>T</td><td>C</td><td>G</td><td>A</td><td>C</td><td>T</td><td>G</td><td>G</td><td>C</td></tr></table>	A	G	T	T	T	C	C	A	T	G	T	A	C	C	G	T	T	A	T	G	A	A	T	T	C	G	A	C	T	G	G	C	TaxonN
A	G	T	T	T	C	C	A	T	G	T	A	C	C	G	T	T	A	T	G	A	A	T	T	C	G	A	C	T	G	G	C			

Anotación taxonómica

- **VSEARCH (SINTAX)**

Secuencia problema referencia $r \in R$.

A	G	G	T	A	C	C	A	T	G	T	A	C	C	G	T	T	A	G	G	A	A	T	T	C	G	A	C	T	A	G	C
G	G	T	A	C	C	A	T																								
	G	T	A	C	C	A	T	G																							
		T	A	C	C	A	T	G	T																						
			A	C	C	A	T	G	T	A																					
				C	C	A	T	G	T	A	C																				

...

A	A	T	T	C	G	A	C																								
	A	T	T	C	G	A	C	T																							
		T	T	C	G	A	C	T	A																						
			T	C	G	A	C	T	T	G																					
				C	G	A	C	T	T	G	C																				

W(r): todos los k-mers de r.

Anotación taxonómica

- **VSEARCH (SINTAX)**

W(r)

G	G	T	A	C	C	A	T
G	T	A	C	C	A	T	G
T	A	C	C	A	T	G	T
A	C	C	A	T	G	T	A
C	C	A	T	G	T	A	C

■ ■ ■

A	A	T	T	C	G	A	C
A	T	T	C	G	A	C	T
T	T	C	G	A	C	T	A
T	C	G	A	C	T	T	G
C	G	A	C	T	T	G	C

Anotación taxonómica

- **VSEARCH (SINTAX)**

Para cada $r \in R$ secuencias referencia, en número k-mers en común entre $w_s(Q)$ y $W(r)$ es $U^{\text{subset}}(r) = |w_s(Q) \cap W(r)|$.

Anotación taxonómica

- VSEARCH (SINTAX)

$w_s(Q)$

G	T	A	C	C	A	T	G
T	A	C	C	A	T	G	T
C	C	A	T	G	T	A	C

...

A	A	T	T	C	G	A	C
T	T	C	G	A	C	T	A
C	G	A	C	T	T	G	C

$W(r)$

G	G	T	A	C	C	A	T
G	T	A	C	C	A	T	G
T	A	C	C	A	T	G	T
A	C	C	A	T	G	T	A
C	C	A	T	G	T	A	C

...

A	A	T	T	C	G	A	C
A	T	T	C	G	A	C	T
T	T	C	G	A	C	T	A
T	C	G	A	C	T	T	G
C	G	A	C	T	T	G	C

$$U^{\text{subset}}(r) = |w_s(Q) \cap W(r)|$$

G	T	A	C	C	A	T	G
T	A	C	C	A	T	G	T

...

A	A	T	T	C	G	A	C
---	---	---	---	---	---	---	---

Anotación taxonómica

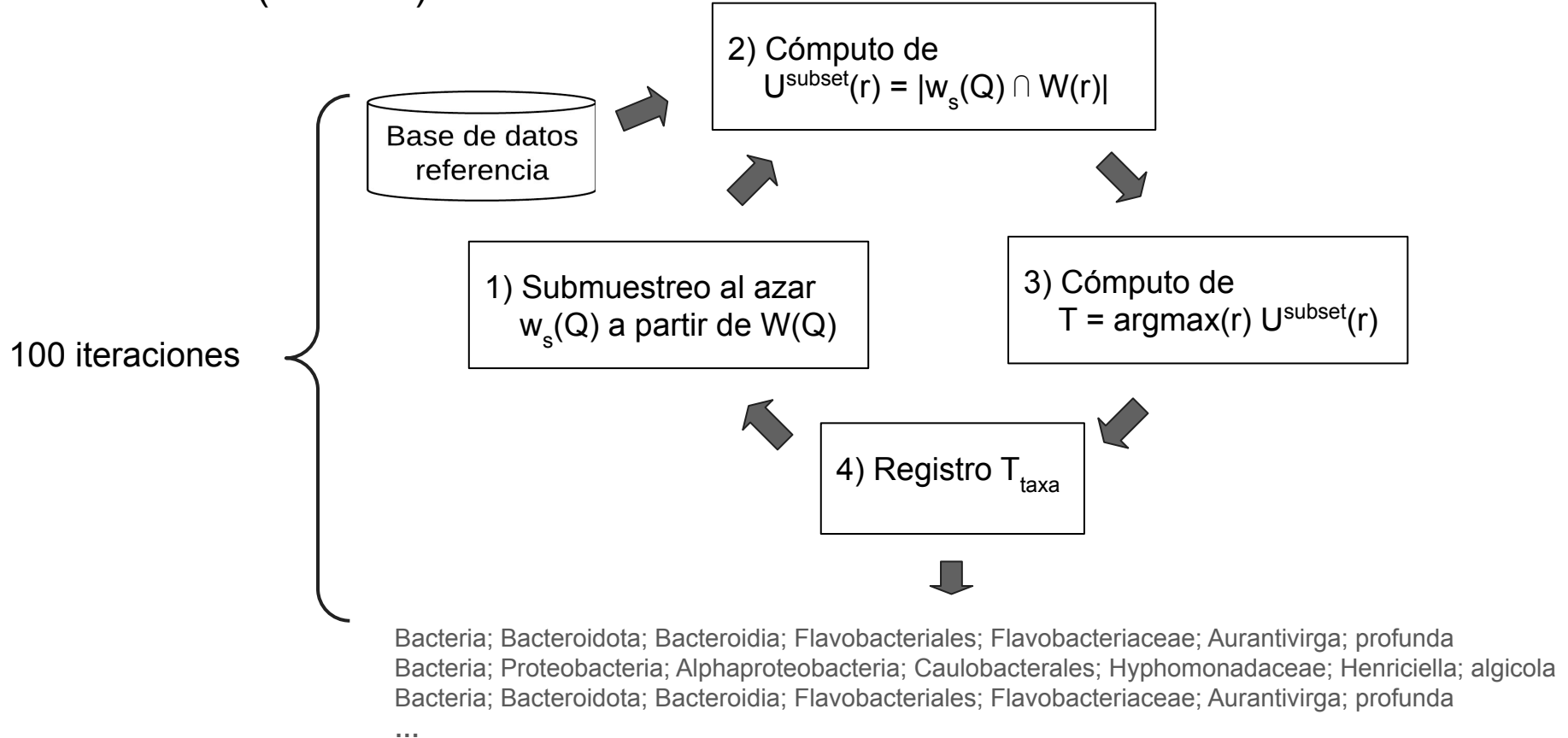
- **VSEARCH** (SINTAX)

$T = \operatorname{argmax}(\mathbf{r}) \mathbf{U}^{\text{subset}}(\mathbf{r})$: secuencia $T \in R$ de mayor similitud en términos de k-mers compartidos con Q (*top hit*).

T_{taxa} : anotación taxonómica de las secuencias de T (i.e., Especie, Genero, Familia, Clase, Orden y Filum).

Anotación taxonómica

- VSEARCH (SINTAX)



Anotación taxonómica

- **VSEARCH (SINTAX)**

Anotación taxonómica de Q:

Nivel taxonómico	Taxón	Bootstrap
Dominio	Bacteria	100%
Orden	Bacteroidota	100%
Clase	Flavobacteriales	85%
Familia	Flavobacteriaceae	85%
Genero	Aurantivirga	60%
Especie	profunda	55%

Anotación taxonómica

- **VSEARCH (SINTAX)**

¿Por qué un submuestreo de k-mers con un $n = 32$ fijo?

Si $U^{all}(r1) \gg U^{all}(r2)$ entonces $U^{subset}(r1)$ será mayor que $U^{subset}(r2)$ en la mayoría o en todas las iteraciones y, por lo tanto, $T1_{taxa}$ tendrá una alta confianza.

Si $U^{all}(r1) \sim U^{all}(r2)$ entonces $U^{subset}(r1)$ será mayor que $U^{subset}(r2)$ aproximadamente la mitad (o poco más) iteraciones y, por lo tanto, $T1_{taxa}$ tendrá una baja confianza.

Anotación taxonómica

- **VSEARCH** (SINTAX)

¿Por qué un submuestreo de k-mers con un $n = 32$ fijo?

Dado un ranqueo de todas las secuencias referencias utilizando todos los k-mers de la secuencia problema (*query*) Q :

$U^{all}(r1) > U^{all}(r2) > \dots > U^{all}(rm)$, con $U^{all}(r) = |W(Q) \cap W(r)|$.

Y las taxonomías asociadas a cada secuencias referencia también ordenadas: $T1_{taxa}, T2_{taxa}, T3_{taxa}, \dots, Tm_{taxa}$.

Anotación taxonómica

- **VSEARCH (SINTAX)**

¿Por qué un submuestreo de k-mers con un $n = 32$ fijo?

Si n es grande, entonces $\mathbf{U}^{\text{subsetl}}(\mathbf{r}) \rightarrow \mathbf{U}^{\text{all}}(\mathbf{r})$, por lo que los valores *bootstrap* dejan de ser informativos.

Anotación taxonómica

- **Bootstrapping:** proviene de la frase (imposible) de empujarse hacia arriba tirando uno mismo de sus propios *bootstraps*.



Anotación taxonómica

- **Clasificador de Bayes Ingenuo (NBC)**

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Aug. 2007, p. 5261–5267
0099-2240/07/\$08.00+0 doi:10.1128/AEM.00062-07

Vol. 73, No. 16

Copyright © 2007, American Society for Microbiology. All Rights Reserved.

Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy^{∇†}

Qiong Wang,¹ George M. Garrity,^{1,2} James M. Tiedje,^{1,2} and James R. Cole^{1*}

*Center for Microbial Ecology¹ and Department of Microbiology and Molecular Genetics,² Michigan State University,
East Lansing, Michigan 48824*

Anotación taxonómica

- **Clasificador de Bayes Ingenuo (NBC)**

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Aug. 2007, p. 5261–5267
0099-2240/07/\$08.00+0 doi:10.1128/AEM.00062-07
Copyright © 2007, American Society for Microbiology. All Rights Reserved.

Vol. 73, No. 16

Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy^{∇†}

Qiong Wang,¹ George M. Garrity,^{1,2} James M. Tiedje,^{1,2} and James R. Cole^{1*}

Center for Microbial Ecology¹ and Department of Microbiology and Molecular Genetics,² Michigan State University, East Lansing, Michigan 48824

Uno de los artículos más citados en microbiología!

<http://archive.sciencewatch.com/dr/erf/2011/11decerf/11decerfCole>

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)



Implementado en el paquete de R DADA2.

DADA2: Fast and accurate sample inference from amplicon data with single-nucleotide resolution



Anotación taxonómica

- **Clasificador de Bayes Ingenuo (NBC)**

Clasificador taxonómico “Bayesiano” para genes 16S de ARNr (aplicable también a otros genes).

Anotación taxonómica

- **Clasificador de Bayes Ingenuo (NBC)**

Clasificador taxonómico “Bayesiano” para genes 16S de ARNr (aplicable también a otros genes).



Basado en similitud de k-mers (sin alineamientos entre secuencias).

Anotación taxonómica

- Teorema de Bayes

Probabilidad condicional: Evento A dado evento B es: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Probabilidad condicional: Evento B dado evento A es: $P(B|A) = \frac{P(B \cap A)}{P(A)}$

Sabemos que: $P(B \cap A) = P(A \cap B)$

Entonces: $P(A \cap B) = P(A|B)P(B) = P(B \cap A) = P(B|A)P(A)$

Sustituyendo obtenemos: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Anotación taxonómica

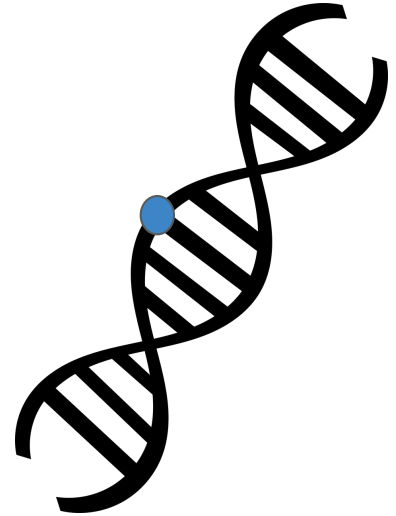
- **Teorema de Bayes**

Un test de determinada variante genética es da positivo.

El test tiene una precisión del 99%.

La variante tiene una prevalencia del 0.1%

¿Cuál es la probabilidad de tener la variante?



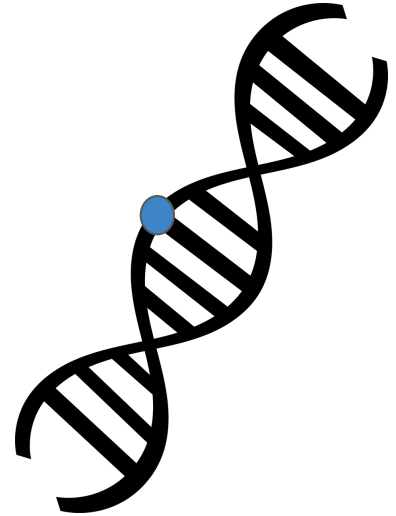
Anotación taxonómica

- Teorema de Bayes

$$P(V+|T+) = \frac{P(T+|V+)*P(V+)}{P(T+)}$$

V+: realmente tener la variante

T+: test dio positivo



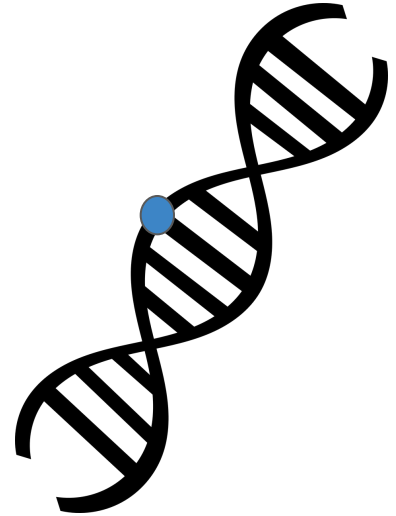
Anotación taxonómica

- Teorema de Bayes

$$P(V+|T+) = \frac{P(T+|V+)*P(V+)}{P(V+)*P(T+|V+) + P(V-)*P(T+|V-)}$$

V+: realmente tener la variante

T+: test dio positivo

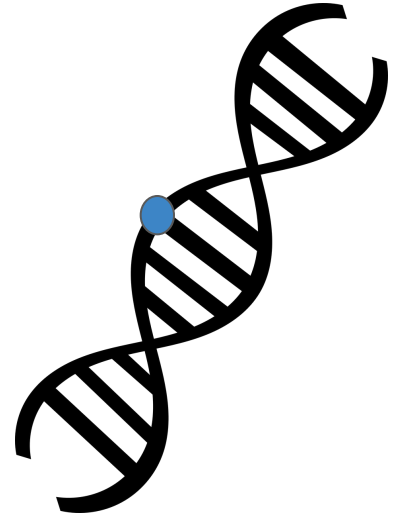


Anotación taxonómica

- Teorema de Bayes

$$P(V+|T+) = \frac{0.99 \cdot 0.001}{0.001 \cdot 0.99 + 0.999 \cdot 0.01} = 0.0901$$

La prob. de V+, dado T+ es de 9%!



Anotación taxonómica

- **Clasificador de Bayes Ingenuo (NBC)**

Ingenuo: se refiere a la suposición (ingenua) de que los atributos de los datos (palabras de secuencia de ADN) son independientes.

NBC viola este supuesto, pero se ha demostrado que tal violación no afecta sustancialmente la performance.

Anotación taxonómica

- **Clasificador de Bayes Ingenuo (NBC)**

Lo que se hace en este caso es generar modelos a partir de una base de datos de referencia de entrenamiento.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Secuencia referencia “r” de entrenamiento

A	G	G	T	A	C	C	A	T	G	T	A	C	C	G	T	T	A	G	G	A	A	T	T	C	G	A	C	T	A	G	C
G	G	T	A	C	C	A	T																								
	G	T	A	C	C	A	T	G																							
		T	A	C	C	A	T	G	T																						
			A	C	C	A	T	G	T	A																					
				C	C	A	T	G	T	A	C																				

...

A	A	T	T	C	G	A	C																								
	A	T	T	C	G	A	C	T																							
		T	T	C	G	A	C	T	A																						
			T	C	G	A	C	T	T	G																					
				C	G	A	C	T	T	G	C																				

$W(r)$: todos los
k-mers de r.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Secuencia referencia “r” de entrenamiento

A	G	G	T	A	C	C	A	T	G	T	A	C	C	G	T	T	A	G	G	A	A	T	T	C	G	A	C	T	A	G	C
G	G	T	A	C	C	A	T																								
	G	T	A	C	C	A	T	G																							
		T	A	C	C	A	T	G	T																						
			A	C	C	A	T	G	T	A																					
				C	C	A	T	G	T	A	C																				

...

A	A	T	T	C	G	A	C																								
	A	T	T	C	G	A	C	T																							
		T	T	C	G	A	C	T	A																						
			T	C	G	A	C	T	T	G																					
				C	G	A	C	T	T	G	C																				

W(r): todos los
8-mers de r.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Sea $D = 4^k$ (i.e., todos los posibles k-mers).

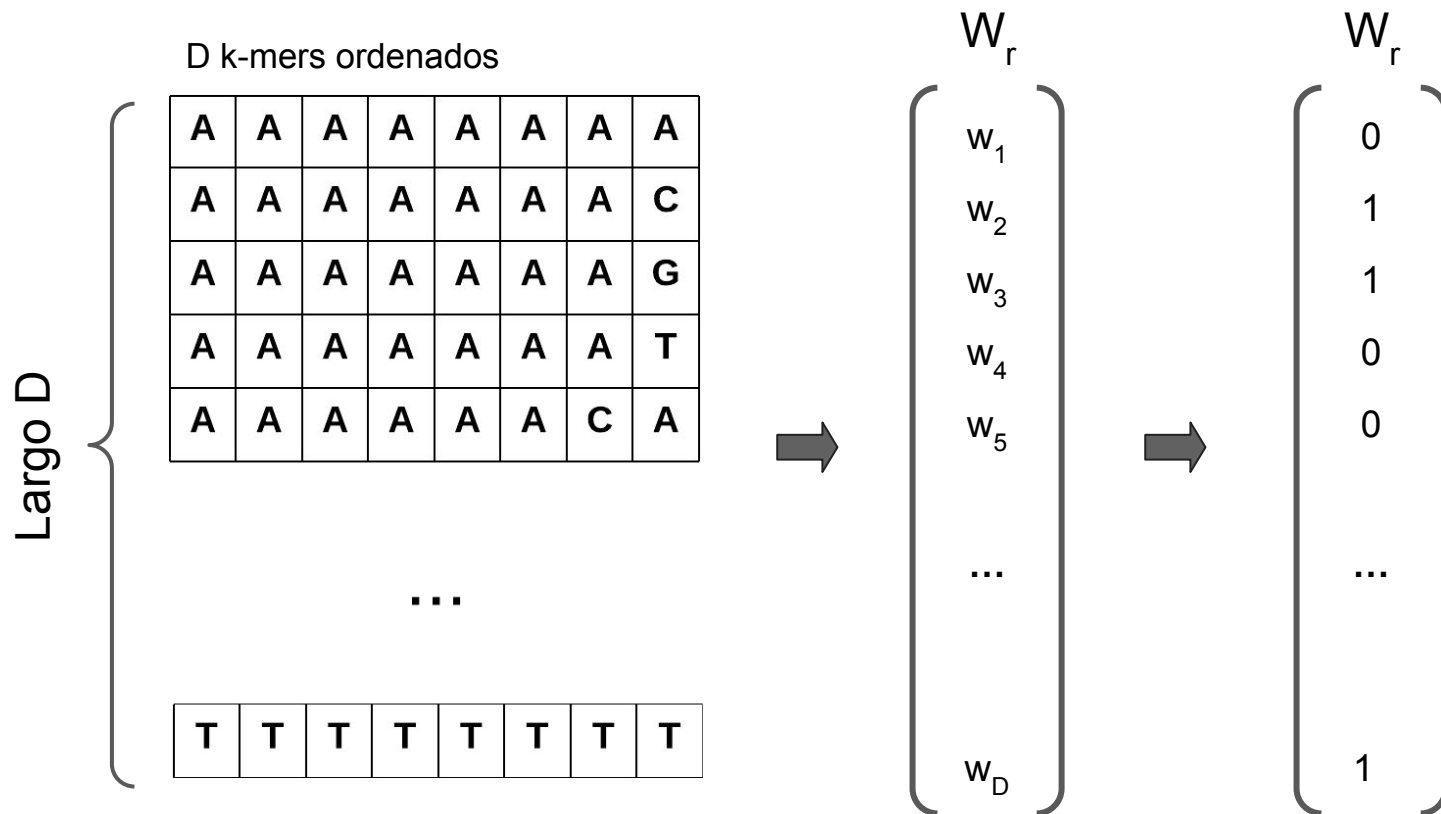
Para cada secuencia “r”, creamos un vector W_r de longitud D.

Cada posición (w_i) corresponde un $k\text{-mer}_i$, y se incluyen todos los k-mers posibles ordenados alfabéticamente.

$w_i = 1$ si $k\text{-mer}_i$ está presente en r y 0 si no lo está.

Anotación taxonómica

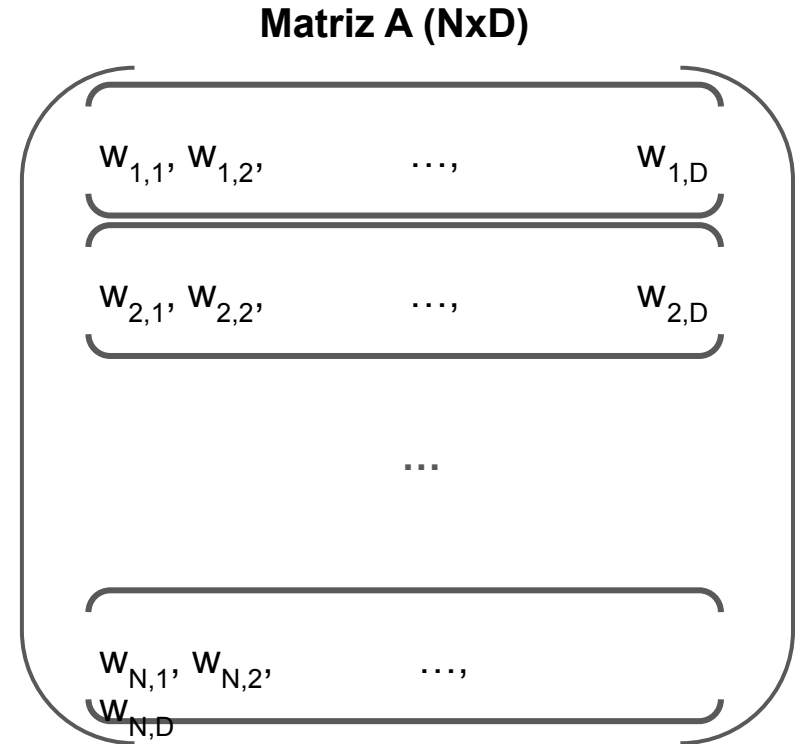
- Clasificador de Bayes Ingenuo (NBC)



Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Para cada secuencias “r” referencia (de entrenamiento) de las N secuencias totales, computamos un vector W, los cuales son ordenados como filas en una matriz A (NxD).



Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Primera parte: Probabilidad de cada k-mer_i independientemente de su taxonomía.

Sumando todas las filas de A obtenemos un vector (n_1, n_2, \dots, n_D) , donde $n_j = \sum_{i=1}^N w_{i,j}$.
Esto es, la suma de todas las secuencias con el k-mer_j.

Matriz A (NxD)

$w_{1,1}$	$w_{1,2}$...	$w_{1,D}$
$w_{2,1}$	$w_{2,2}$...	$w_{2,D}$
...			
$w_{N,1}$	$w_{N,2}$...	$w_{N,D}$

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Primera parte: Probabilidad de cada k-mer_i independientemente de su taxonomía.

Sumando todas las filas de A obtenemos un vector (n_1, n_2, \dots, n_D) , donde $n_j = \sum_{i=1}^N w_{i,j}$.
Esto es, la suma de todas las secuencias con el k-mer_j.

Matriz A (NxD)

$w_{1,1}$	$w_{1,2}$...	$w_{1,D}$
$w_{2,1}$	$w_{2,2}$...	$w_{2,D}$
...			
$w_{N,1}$	$w_{N,2}$...	$w_{N,D}$
n_1	n_2	...	n_D

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Primera parte: Probabilidad de cada k-mer_j independientemente de su taxonomía.

$$\text{Esto es: } Pr(w_j) = \frac{n_j + 0.5}{N + 1}$$

0.5 y 1 son sumados para garantizar que $0 < pr(w_j) < 1$.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Segunda parte: probabilidad de obtener $k\text{-mer}_j$ dado que el género g .

Consideramos A_g , sub-matriz de A , con M_g filas correspondientes al género g .

Matriz $A_g(M_g \times D)$

$w_{1,1}$	$w_{1,2}$...	$w_{1,D}$
$w_{2,1}$	$w_{2,2}$...	$w_{2,D}$
...			
$w_{M_g,1}$	$w_{M_g,2}$...	$w_{M_g,D}$

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Segunda parte: probabilidad de obtener $k\text{-mer}_j$ dado que el género g .

Sumando todas las filas de A_g obtenemos un vector $(m_{g1}, m_{g2}, \dots, m_{gD})$,

donde
$$m_{gj} = \sum_{i=1}^{M_g} w_{i,j}.$$

Esto es, la suma de todas las secuencias con el $k\text{-mer}_j$ dentro del género g .

Matriz $A_g (M_g \times D)$

$w_{1,1}$	$w_{1,2}$	\dots	$w_{1,D}$
$w_{2,1}$	$w_{2,2}$	\dots	$w_{2,D}$
\dots			
$w_{M_g,1}$	$w_{M_g,2}$	\dots	$w_{M_g,D}$
m_{g1}	m_{g2}	\dots	m_{gD}

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Segunda parte: probabilidad de obtener $k\text{-mer}_j$ dado que el taxón g .

$$Pr(w_j|g) = \frac{m_{gj} + pr(w_j)}{M_g + 1}$$

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Segunda parte: probabilidad de obtener k-mer_j dado que el taxón g.

$$Pr(w_j|g) = \frac{m_{gj} + pr(w_j)}{M_g + 1}$$

The screenshot shows the usearch v11 website. The main heading is "NBC calculation of genus-specific conditional probability". Below this, there is a section titled "See also" with a link to "RDP Naive Bayesian Classifier algorithm". The main text explains the simplest estimate for the frequency observed in the training set, $m(w_j)/M$, and mentions the addition of pseudo-counts to model unobserved sequences. It then discusses "Genus-specific conditional probabilities" for genus G with a training set of M sequences, where $m(w_j)$ is the number of sequences containing word w_j . The conditional probability that a member of G contains w_j is estimated with the equation $P(w_j|G) = [m(w_j) + P_j]/(M + 1)$. The text notes that this ignores the dependency between words in an individual sequence and the joint probability of observing from genus G a (partial) sequence, S , containing a set of words, $V = \{v_1, v_2, \dots, v_j\}$ ($V \subseteq W$), which was estimated as $P(S|G) = \prod P(v_i|G)$. A reference is provided at the bottom: Wang, Q. et al. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *AEM* 73, 5261-7.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Segunda parte: probabilidad de obtener k-mer_j dado que el taxón g.

Decimos que $q_{g,i} = Pr(w_j|g) = \frac{m_{gj} + pr(w_j)}{M_g + 1}$.

Los $q_{g,j}$ forman una matriz Q (GxD), siendo G el número total de géneros en nuestro se de datos de entrenamiento.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Modelo entrenado: matriz Q ($G \times D$) con todas las probabilidades $q_{g,j}$ para cada k-mer de cada genero.

Matriz Q ($G \times D$)

$q_{1,1}$, $q_{1,2}$, ..., $q_{1,D}$
$q_{2,1}$, $q_{2,2}$, ..., $q_{2,D}$
...
$q_{G,1}$, $q_{G,2}$, ..., $q_{G,D}$

Anotación taxonómica

- **Clasificador de Bayes Ingenuo (NBC)**

Clasificación de una nueva secuencia “a”.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Clasificación de una nueva secuencia “a”.

Lo que queremos obtener son los $P(g|a)$ para todo $g \in G$.

Es decir, las probabilidades de todos los géneros en nuestro set de entrenamiento, dado que observamos la secuencia “a”.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Clasificación de una nueva secuencia “a”.

Lo que queremos obtener son los $P(g|a)$ para todo $g \in G$.

Es decir, las probabilidades de todos los géneros en nuestro set de entrenamiento, dado que observamos la secuencia “a”.



La secuencias “a” es clasificada en el género con mayor $P(g|a)$.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Clasificación de una nueva secuencia “a”.

Lo que podemos obtener con nuestra matriz Q es $P(\mathbf{a}|\mathbf{g})$ para todo $\mathbf{g} \in \mathbf{G}$.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Clasificación de una nueva secuencia “a”.

Lo que podemos obtener con nuestra matriz Q es $\mathbf{P}(\mathbf{a}|\mathbf{g})$ para todo $\mathbf{g} \in \mathbf{G}$.

Acá es donde interviene el teorema de Bayes.

$$Pr(g|a) = \frac{Pr(g) * Pr(a|g)}{Pr(a)}$$

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Clasificación de una nueva secuencia “a”.

Lo que podemos obtener con nuestra matriz Q es $\mathbf{P(a|g)}$ para todo $\mathbf{g \in G}$.

Acá es donde interviene el teorema de Bayes.

$$Pr(g|a) = \frac{Pr(g) * Pr(a|g)}{Pr(a)}$$



Veamos cómo podemos computar cada uno de estos términos.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Clasificación de una nueva secuencia “a”.

Lo que podemos obtener con nuestra matriz Q es $\mathbf{P(a|g)}$ para todo $\mathbf{g \in G}$.

Acá es donde interviene el teorema de Bayes.

$$Pr(g|a) = \frac{Pr(g) * Pr(a|g)}{Pr(a)}$$



Esta probabilidad la podemos computar a partir de nuestra matriz Q (modelo entrenado).

Anotación taxonómica

- **Clasificador de Bayes Ingenuo (NBC)**

Clasificación de una nueva secuencia “a”.

Cómputo de $\Pr(a|g)$.

Anotación taxonómica

- **Clasificador de Bayes Ingenuo (NBC)**

Clasificación de una nueva secuencia “a”.

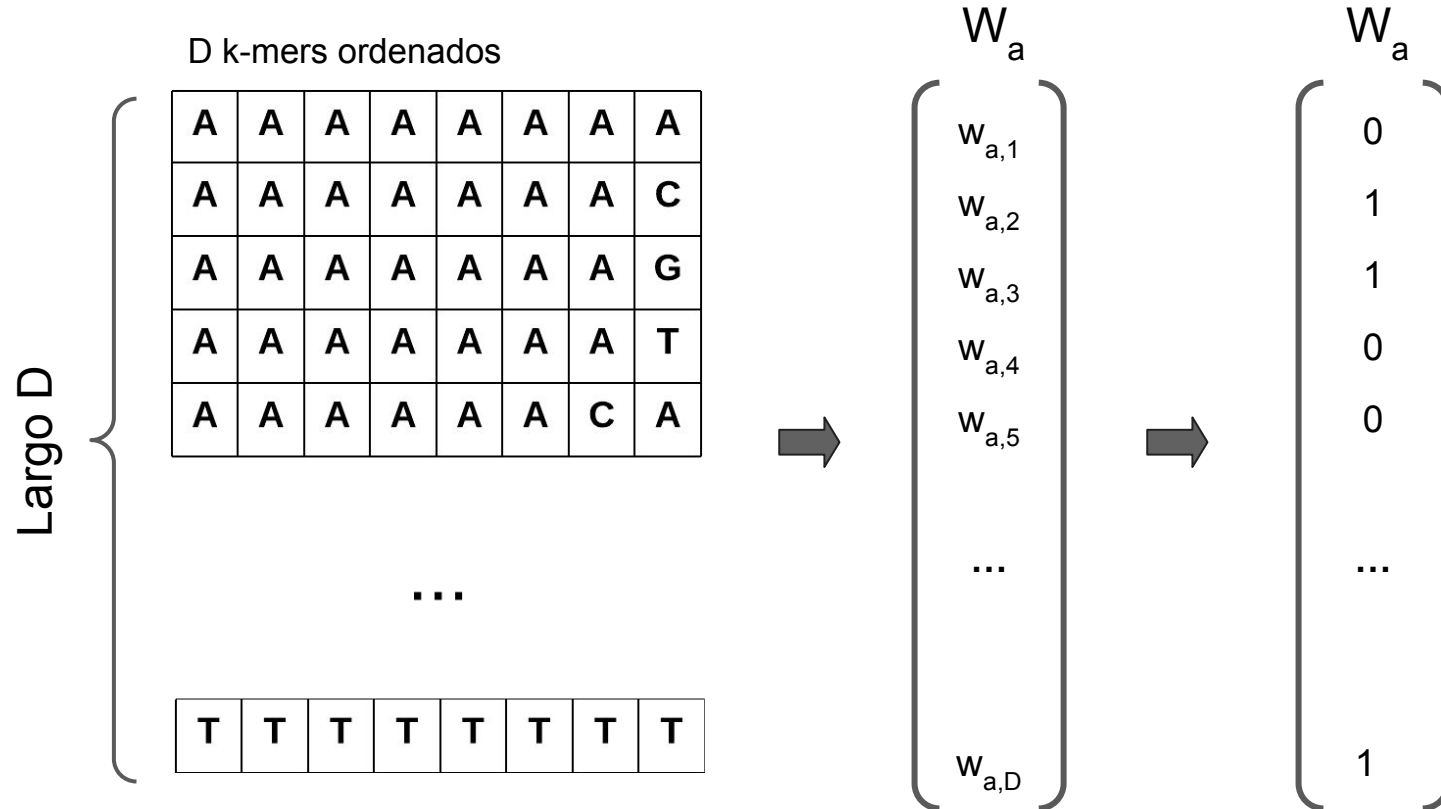
Cómputo de $\Pr(a|g)$.



Primero computamos el vector W_a .

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)



Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Clasificación de una nueva secuencia “a”.

Luego de obtener W_a podemos computar **$\Pr(a|g)$** de la siguiente forma:

$$\Pr(a|g) = \prod_{j=1}^D \Pr(w_{aj}|g)$$

Donde $\Pr(w_{aj}|g) = \frac{m_{gj} + \Pr(w_{aj})}{M_g + 1}$

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Clasificación de una nueva secuencia “a”.

Luego de obtener W_a podemos computar $\mathbf{Pr(a|g)}$ de la siguiente forma:

$$Pr(a|g) = \prod_{j=1}^D Pr(w_{aj}|g)$$



Esto implica asumir que los k-mers de “a” independientes (i.e, **ingenuo**).

Donde $Pr(w_{aj}|g) = \frac{m_{gj} + Pr(w_{aj})}{M_g + 1}$

Esto no se cumple, pero el método igual funciona satisfactoriamente.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Clasificación de una nueva secuencia “a”.

Luego de obtener W_a podemos computar $\mathbf{Pr(a|g)}$ de la siguiente forma:

$$Pr(a|g) = \prod_{j=1}^D Pr(w_{aj}|g)$$



De igual modo a cómo computamos $Pr(w_j|g)$ en el set de entrenamiento.

Donde $Pr(w_{aj}|g) = \frac{m_{gj} + Pr(w_{aj})}{M_g + 1}$

$$Pr(w_j|g) = \frac{m_{gj} + pr(w_j)}{M_g + 1}$$

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)


Clasificación de una nueva secuencia “a”.

$$Pr(g|a) = \frac{Pr(g) * Pr(a|g)}{Pr(a)} \Rightarrow Pr(a|g) = \prod_{j=1}^D Pr(w_{aj}|g)$$

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Clasificación de una nueva secuencia “a”.

$$Pr(g|a) = \frac{Pr(g) * Pr(a|g)}{Pr(a)}$$


Lo que buscamos que maximizar $Pr(g|a)$, para todo $g \in G$ en nuestro set de entrenamiento.



$Pr(a)$ es un constante que puede ser ignorada.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Clasificación de una nueva secuencia “a”.

$$Pr(g|a) = \frac{Pr(g) * Pr(a|g)}{\cancel{Pr(a)}} \rightarrow$$

Lo que buscamos que maximizar $Pr(g|a)$,
para todo $g \in G$ en nuestro set de
entrenamiento.




$Pr(a)$ es un constante que puede ser
ignorada.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Clasificación de una nueva secuencia “a”.

$$Pr(g|a) = \frac{Pr(g) * Pr(a|g)}{Pr(a)}$$


Todos los géneros se asumen igualmente probables.

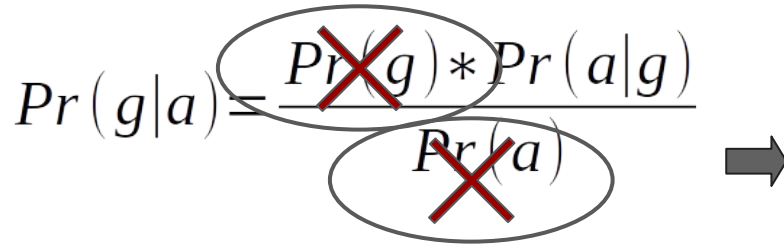


$Pr(g)$ es un constante que puede ser ignorada.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Clasificación de una nueva secuencia “a”.

$$Pr(g|a) = \frac{\cancel{Pr(g)} * Pr(a|g)}{\cancel{Pr(a)}}$$


Todos los géneros se asumen igualmente probables.



$Pr(g)$ es un constante que puede ser ignorada.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Clasificación de una nueva secuencia “a”.

Se reduce a encontrar un $g \in G$ que maximice $Pr(a|g) = \prod_{j=1}^D Pr(w_{aj}|g)$.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Estimación de confianza por *bootstrap*.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Estimación de confianza por *bootstrap*.

Sequencia query “a”.

A	G	G	T	A	C	C	A	T	G	T	A	C	C	G	T	T	A	G	G	A	A	T	T	C	G	A	C	T	A	G	C
G	G	T	A	C	C	A	T																								
	G	T	A	C	C	A	T	G																							
		T	A	C	C	A	T	G	T																						
			A	C	C	A	T	G	T	A																					
				C	C	A	T	G	T	A	C																				

...

A	A	T	T	C	G	A	C																								
	A	T	T	C	G	A	C	T																							
		T	T	C	G	A	C	T	A																						
			T	C	G	A	C	T	T	G																					
				C	G	A	C	T	T	G	C																				

Realizamos 100 muestreos
de $n = 1/8$ de todos 8-mers.



En correspondencia con 8-mers
independientes, es decir, no
superpuestos (aunque no lo
son).

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Estimación de confianza por *bootstrap*.

Sequencia query “a”.

A	G	G	T	A	C	C	A	T	G	T	A	C	C	G	T	T	A	G	G	A	A	T	T	C	G	A	C	T	A	G	C
G	G	T	A	C	C	A	T																								
	G	T	A	C	C	A	T	G																							
		T	A	C	C	A	T	G	T																						
			A	C	C	A	T	G	T	A																					
				C	C	A	T	G	T	A	C																				

...

A	A	T	T	C	G	A	C																								
	A	T	T	C	G	A	C	T																							
		T	T	C	G	A	C	T	A																						
			T	C	G	A	C	T	T	G																					
				C	G	A	C	T	T	G	C																				

Realizamos 100 muestreos
de $n = 1/8$ de todos 8-mers.



Determinamos el g con el
máximo valor $P(g|a)$ en cada
caso.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Estimación de confianza por *bootstrap*.

Sequencia query “a”.

Realizamos 100 muestreos
de $n = 1/8$ de todos 8-mers.



Registramos el % de veces que
aparece cada género y grupos
taxonómicos superiores.

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Estimación de confianza por *bootstrap*.

Secuencia query “a”.

Realizamos 100 muestreos
de $n = 1/8$ de todos 8-mers.



¿Qué pasa si la secuencia a es
muy larga?

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Estimación de confianza por *bootstrap*.

Sequencia query “a”.

Nivel taxonómico	Taxón	Bootstrap
Dominio	Bacteria	100%
Orden	Bacteroidota	100%
Clase	Flavobacteriales	85%
Familia	Flavobacteriaceae	85%
Genero	Aurantivirga	60%
Especie	profunda	55%

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Validación utilizando *Leave-One-Out Cross Validation*

Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Validación utilizando *Leave-One-Out Cross Validation*.

Corpus de Bergey: 5,014 secuencias de ARNr.

Corpus NCBI: 23.095 secuencias de ARNr.

TABLE 1. Number of taxa at different ranks

Taxonomy	No. of sequences in corpus	No. of:					
		Domains	Phyla	Classes	Orders	Families	Genera
Bergey's	5,014	1	24	33	79	211	988
NCBI	23,095	1	24	31	82	209	1,187


Anotación taxonómica

- **Clasificador de Bayes Ingenuo (NBC)**

Validación utilizando *Leave-One-Out Cross Validation*.

Corpus de Bergey: 5,014 secuencias de ARNr.

Corpus NCBI: 23.095 secuencias de ARNr.

- 
1. Secuencias completas y segmentos de 400, 200, 100, y 50 pb seleccionados al azar.
 2. Regiones de 100 pb recorriendo todas la regiones.

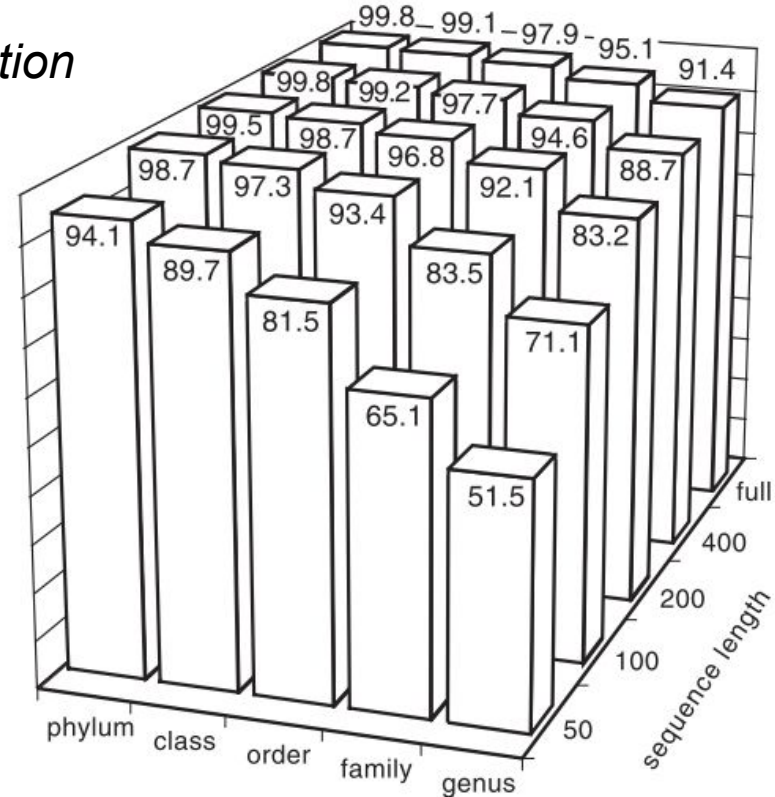
Anotación taxonómica

- Clasificador de Bayes Ingenuo (NBC)

Validación utilizando *Leave-One-Out Cross Validation*

1. Secuencias completas y segmentos de 400, 200, 100, y 50 pb seleccionados al azar.

Precisión de clasificación general por tamaño de secuencia. Los números son porcentajes de clasificaciones correctas.



Anotación taxonómica

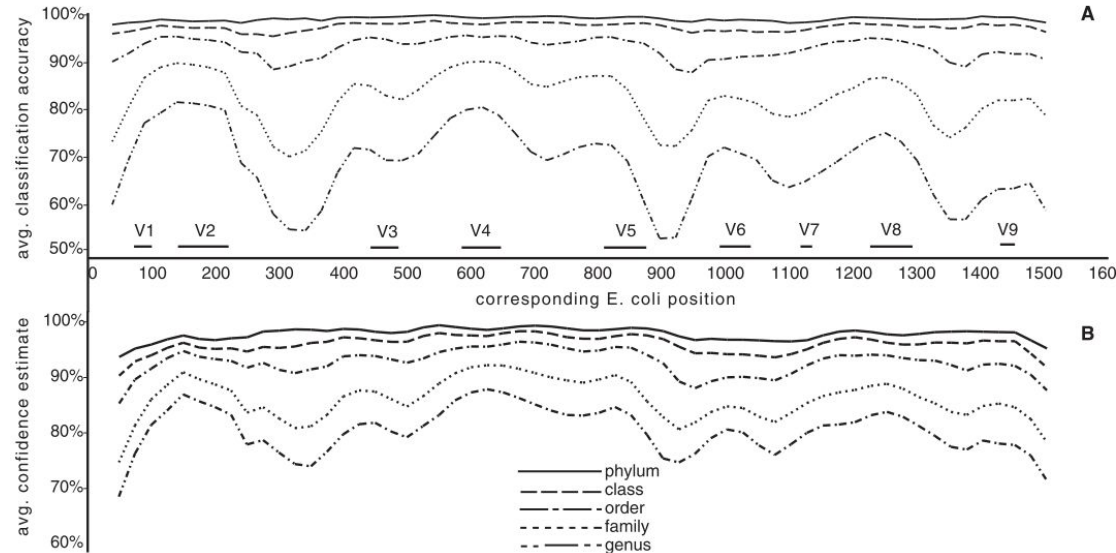
- Clasificador de Bayes Ingenuo (NBC)

Validación utilizando *Leave-One-Out Cross Validation*

2. Regiones de 100 pb
recorriendo todas la regiones.

(A) Precisión de clasificación.

(B) Estimación de confianza.



- **Anotación taxonómica**
- Normalización y distancias

- **Anotación taxonómica**
- **Normalización y distancias**

Normalización

Normalización

Transformación de datos para permitir una comparación precisa entre diferentes mediciones eliminando sesgos derivados de artefactos técnicos (e.g., en la colección de muestras, preparación de librerías, y secuenciado).

Normalización

Transformación de datos para permitir una comparación precisa entre diferentes mediciones eliminando sesgos derivados de artefactos técnicos (e.g., en la colección de muestras, preparación de librerías, y secuenciado).



Dilucidar diferencias o similitudes biológicas.

Normalización

- **Por esfuerzo de muestreo.**

Abundancia de OTUs

	OTU1	OTU2	OTU3	OTU4	OTU5
s1	10	20	20	50	50
s2	150	400	300	200	2000
s3	20	80	10	30	15

¿Ven algo potencialmente problemático en esta tabla?

Normalización

- Por esfuerzo de muestreo.

Abundancia de OTUs

	OTU1	OTU2	OTU3	OTU4	OTU5	Suma
s1	10	20	20	50	50	150
s2	150	400	300	200	2000	3050
s3	20	80	10	30	15	155



~x20

Normalización

- **Escalado**

Suma total (*Total-Sum Scaling* (TSS))

Normalización

- **Escalado**

Suma total (*Total-Sum Scaling* (TSS)).

$$tss(x) = \left[\frac{x_1}{s(X)}, \frac{x_2}{s(X)}, \frac{x_3}{s(X)}, \dots, \frac{x_D}{s(X)} \right]$$

Donde $s(X) = \sum_{i=1}^D x_i$.

Normalización

- **Escalado**

Suma total (*Total-Sum Scaling* (TSS)).

Abundancia de OTUs

	OTU1	OTU2	OTU3	OTU4	OTU5
s1	10	20	20	50	50
s2	150	400	300	200	2000
s3	20	80	10	30	15



Abundancia relativa de OTUs

	A	B	C	D	E
s1	0.067	0.133	0.133	0.333	0.333
s2	0.049	0.131	0.098	0.066	0.656
s3	0.129	0.516	0.065	0.194	0.097

Normalización

- **Escalado**

Suma total (*Total-Sum Scaling* (TSS)).

Problemáticas con matrices dispersas.

Abundancia de OTUs

	OTU1	OTU2	OTU3	OTU4	OTU5
s1	0	20	0	50	0
s2	0	400	0	200	0
s3	0	0	10	0	15

Normalización

- Rarefacción

Normalización

- **Rarefacción**

Normalizamos por esfuerzo de muestreo.

Generamos sub muestras al azar de tamaño “m”.

Donde $m = \min(\text{sum}(s1), \text{sum}(s2), \text{sum}(s3))$.

Normalización

- **Rarefacción**

Normalizamos por esfuerzo de muestreo.

Generamos sub muestras al azar de tamaño “m”.

Donde $m = \min(\text{sum}(s1), \text{sum}(s2), \text{sum}(s3))$.

Abundancia de OTUs

	OTU1	OTU2	OTU3	OTU4	OTU5
s1	10	20	20	50	50
s2	150	400	300	200	2000
s3	20	80	10	30	15



Abundancia de OTUs rarificada

	OTU1	OTU2	OTU3	OTU4	OTU5
s1	10	20	20	50	50
s2	4	25	15	11	95
s3	20	79	9	29	13

Normalización

- **Rarefacción**

Normalizamos por esfuerzo de muestreo.

Generamos sub muestras al azar de tamaño “m”.

Donde $m = \min(\text{sum}(s1), \text{sum}(s2), \text{sum}(s3))$.

Abundancia de OTUs

	OTU1	OTU2	OTU3	OTU4	OTU5	Suma
s1	10	20	20	50	50	150
s2	150	400	300	200	2000	3050
s3	20	80	10	30	15	155



Abundancia de OTUs rarificada

	OTU1	OTU2	OTU3	OTU4	OTU5	Suma
s1	10	20	20	50	50	150
s2	4	25	15	11	95	150
s3	20	79	9	29	13	150

- Rarefacción

OPEN  ACCESS Freely available online

 **PLOS** | COMPUTATIONAL
BIOLOGY

Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

Paul J. McMurdie, Susan Holmes*

Statistics Department, Stanford University, Stanford, California, United States of America

- Rarefacción

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

Paul J. McMurdie, Susan Holmes*

Statistics Department, Stanford University, Stanford, California, United States of America



Normalización

- Rarefacción

Weiss *et al. Microbiome* (2017) 5:27
DOI 10.1186/s40168-017-0237-y

Microbiome

RESEARCH

Open Access



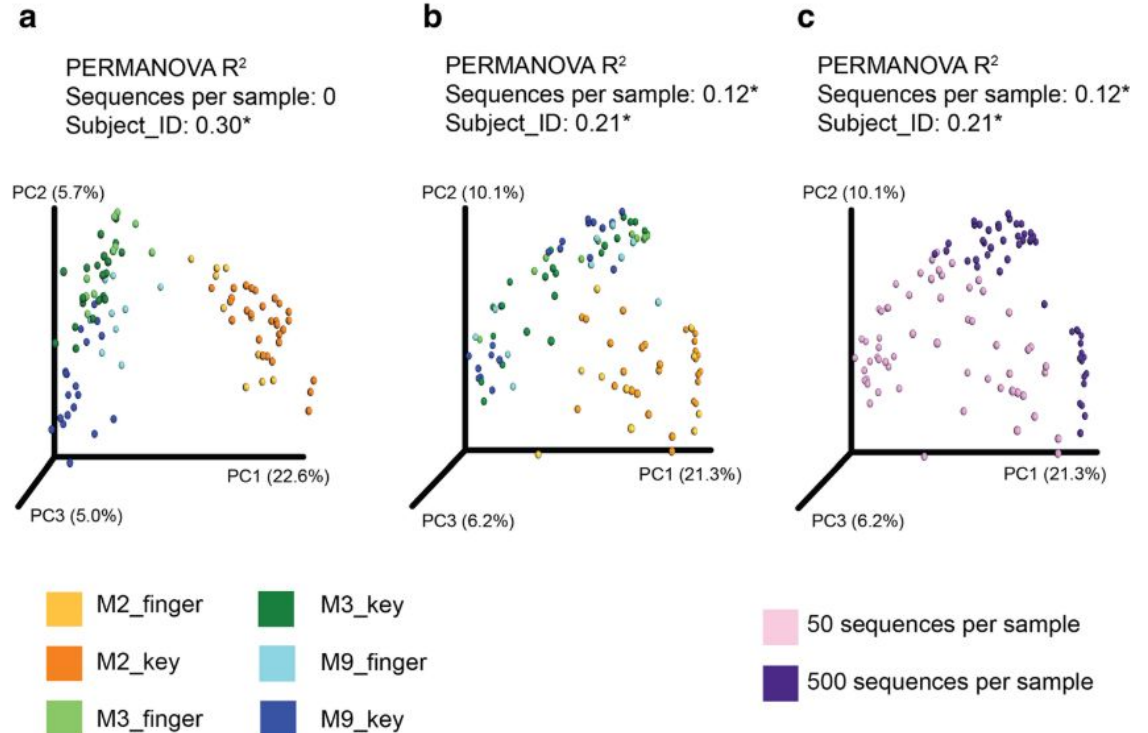
Normalization and microbial differential abundance strategies depend upon data characteristics

Sophie Weiss¹, Zhenjiang Zech Xu², Shyamal Peddada³, Amnon Amir², Kyle Bittinger⁴, Antonio Gonzalez², Catherine Lozupone⁵, Jesse R. Zaneveld⁶, Yoshiki Vázquez-Baeza⁷, Amanda Birmingham⁸, Embriette R. Hyde² and Rob Knight^{2,7,9*}

Normalización

● Rarefacción vs TSS

a. El estudio forense donde se relaciona los dedos del sujeto con los teclados que tocaban, rarificado a 500 secuencias por muestra. **b, c** Datos no normalizados, con una mitad aleatoria de las muestras submuestreadas a 500 secuencias por muestra y la otra mitad a 50 secuencias por muestra. **b** Coloreado por sujeto_ID. **c** Coloreado por secuencias por muestra.



Normalización

- Composicionalidad

Normalización

- **Composicionalidad**

Datos que representan proporciones relativas de un todo.

Normalización

- **Composicionalidad**

Datos que representan proporciones relativas de un todo.

**Matriz de abundancia
de OTUs**

	OTU1	OTU2	OTU3
Muestra1	1	2	2
Muestra2	0	0	1
Muestra3	3	0	4



$$X = [x_1, x_2, x_3, \dots, x_D] \in \mathbb{R}^D : x_i > 0, i = 1, \dots, D; \sum_{i=1}^D x_i = K$$

Normalización

- **Composicionalidad**

Datos que representan proporciones relativas de un todo.

**Matriz de abundancia
de OTUs**

	OTU1	OTU2	OTU3
Muestra1	1	2	2
Muestra2	0	0	1
Muestra3	3	0	4



$$X = [x_1, x_2, x_3, \dots, x_D] \in \mathbb{R}^D : x_i > 0, i = 1, \dots, D; \sum_{i=1}^D x_i = K$$



Ningún componente puede considerarse independiente de otros.

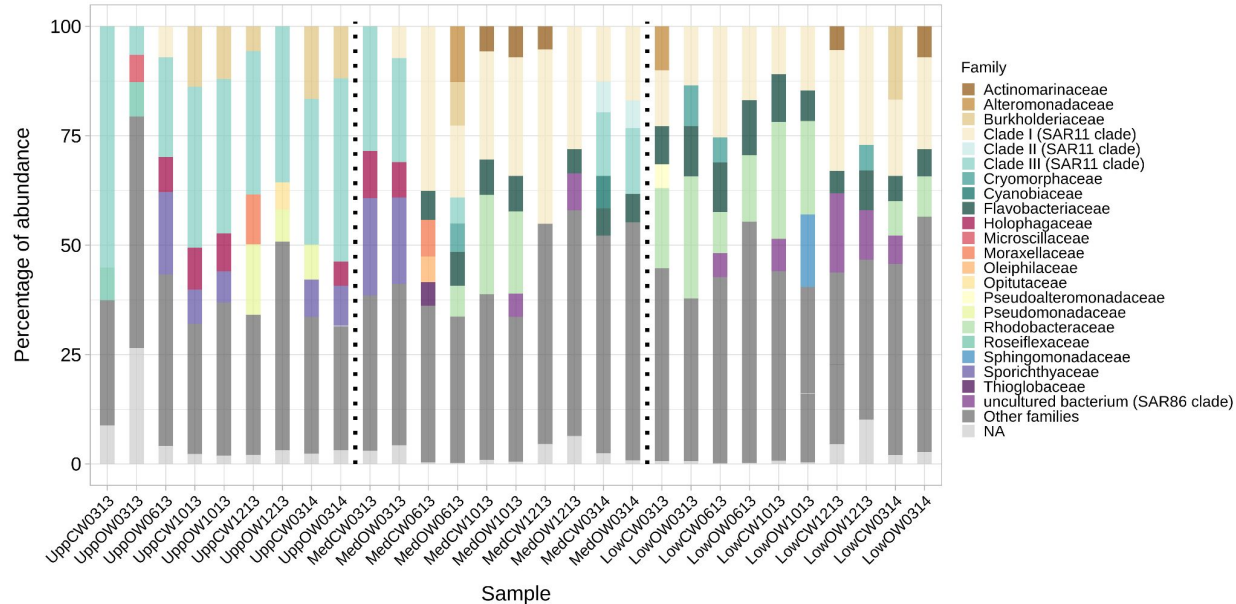
Normalización

- Composicionalidad

Datos que representan proporciones relativas de un todo.

Matriz de abundancia de taxa

	Taxa1	Taxa2	Taxa3
Muestra1	1	2	2
Muestra2	0	0	1
Muestra3	3	0	4



Normalización

- **Composicionalidad**

Datos que representan proporciones relativas de un todo.

Su representación espacial se encuentra dentro de un simplex.

$$S^D = \left\{ X = [x_1, x_2, x_3, \dots, x_D] \in \mathbb{R}^D : x_i > 0, i = 1, \dots, D; \sum_{i=1}^D x_i = K \right\}$$

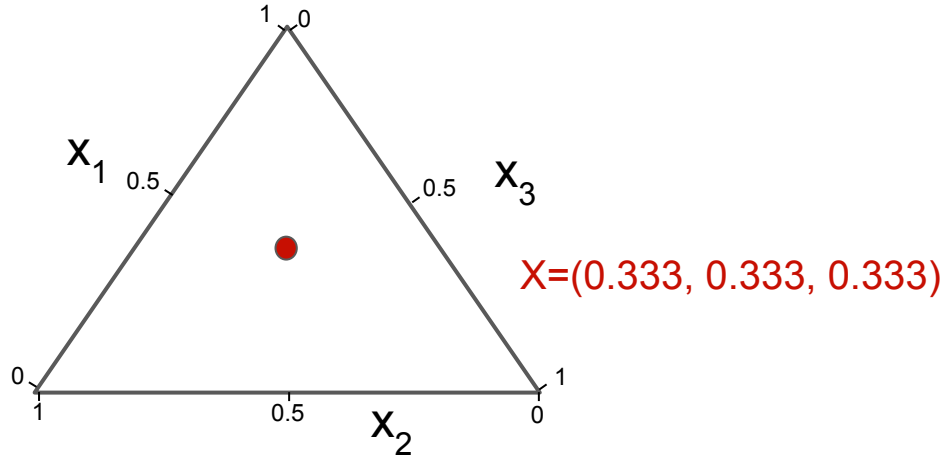
Normalización

- **Composicionalidad**

Datos que representan proporciones relativas de un todo.

Su representación espacial se encuentra dentro de un simplex.

$$S^3 = \{ X = [x_1, x_2, x_3] \in \mathbb{R}^3 : x_i > 0, i=1,2,3; \sum_{i=1}^3 x_i = K \}$$



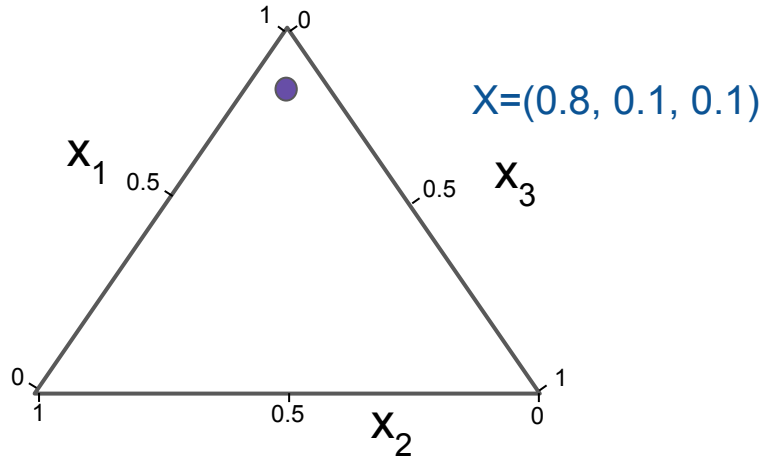
Normalización

- **Composicionalidad**

Datos que representan proporciones relativas de un todo.

Su representación espacial se encuentra dentro de un simplex.

$$S^3 = \{ X = [x_1, x_2, x_3] \in \mathbb{R}^3 : x_i > 0, i=1,2,3; \sum_{i=1}^3 x_i = K \}$$



Normalización

- **Composicionalidad**

Los datos de secuenciación de microbiomas son composicionales: y esto no es opcional!



Normalización

- **Composicionalidad**

Los datos de secuenciación de microbiomas son composicionales: y esto no es opcional!



Los instrumentos de secuenciación tienen una capacidad limitada de *reads*.

Normalización

- **Composicionalidad**

Los datos de secuenciación de microbiomas son composicionales: y esto no es opcional!



Los instrumentos de secuenciación tienen una capacidad limitada de *reads*.

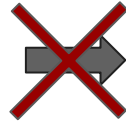


Muestra de aleatoria de abundancias relativas de moléculas.

Normalización

- Composicionalidad

Muestra de aleatoria de abundancias
relativas de moléculas.



Número absoluto de moléculas
en la muestra.

Normalización

- **Composicionalidad**

Varios problemas con los datos composicionales.

Normalización

- **Composicionalidad**

Varios problemas con los datos composicionales.

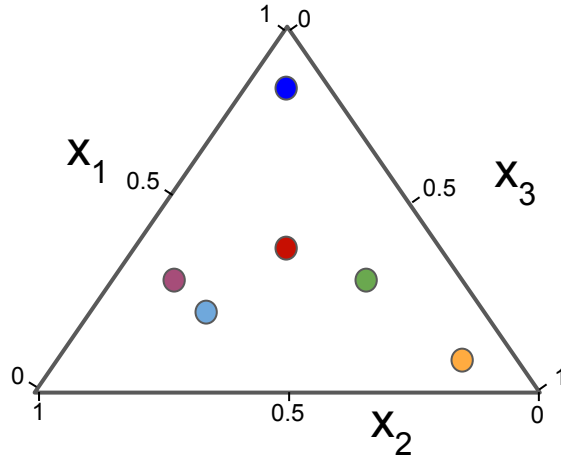
Restringidos al simplex y no libremente distribuidos en el espacio euclidiano.

Normalización

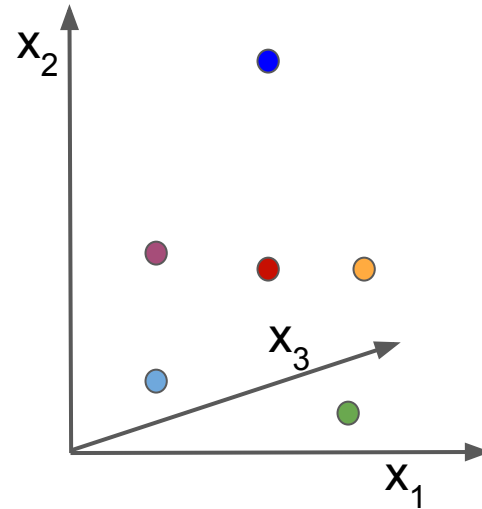
- **Composicionalidad**

Varios problemas con los datos composicionales.

Restringidos al simplex y no libremente distribuidos en el espacio euclidiano.



Vs.



Normalización

- **Composicionalidad**

Varios problemas con los datos composicionales.

Restringidos al simplex y no libremente distribuidos en el espacio euclidiano.



Análisis estadísticos comúnmente utilizados no son aplicables.

La correlación o covariación se ven afectadas, las cuales son clave para la ordenación, la clustering, el análisis de redes y tests abundancia diferencial.

Normalización

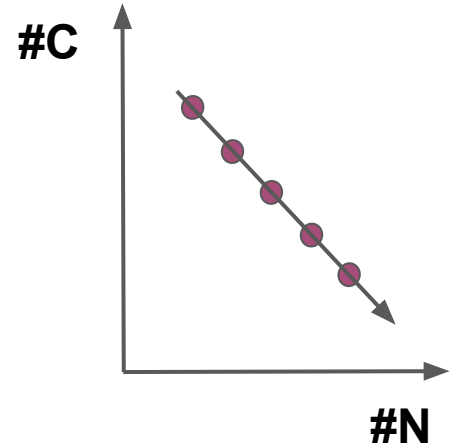
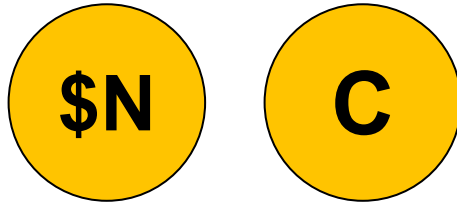
- **Composicionalidad**

Varios problemas con los datos composicionales.

Presentan un sesgo hacia las correlaciones negativas.

Ej., El recuento de caras (C) y números (N) al tirar una moneda 100 veces son datos composicionales.

$$\#C = 100 - \#N.$$



Normalización

- **Composicionalidad**

Varios problemas con los datos composicionales.

Pueden Presentar correlaciones espúreas al agregar o eliminar variables.

Tabla 1. Conteo absoluto

	A	B	C	D	E
s1	10	20	20	50	50
s2	15	40	30	20	200
s3	20	80	10	30	15

Tabla 2. Conteo relativo

	A	B	C	D	E
s1	0.067	0.133	0.133	0.333	0.333
s2	0.049	0.131	0.098	0.066	0.656
s3	0.129	0.516	0.065	0.194	0.097

Tabla 3. Conteo relativo sin E

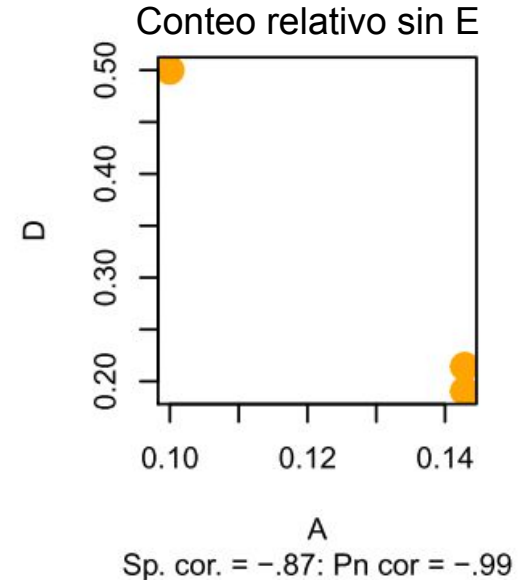
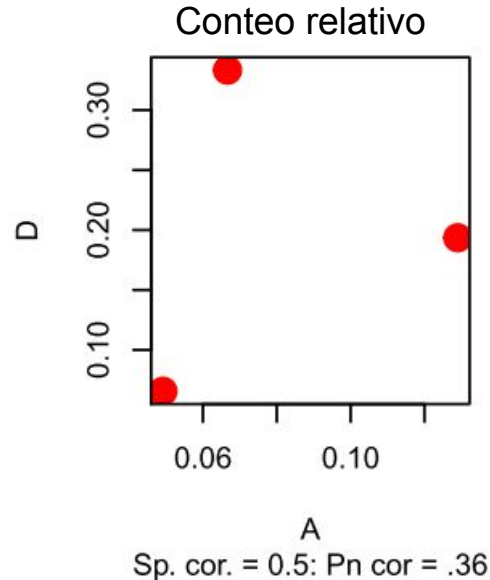
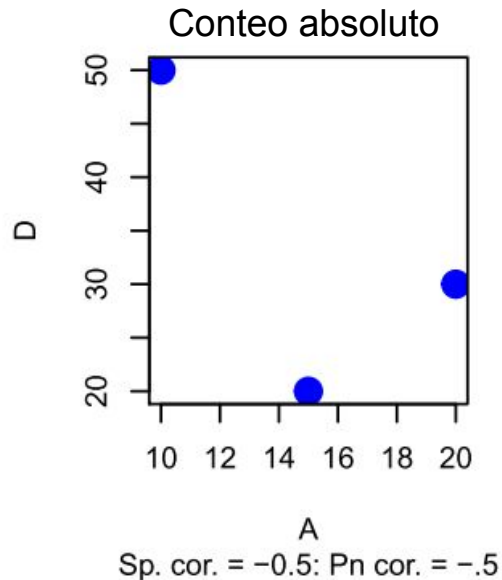
	A	B	C	D
s1	0.100	0.200	0.200	0.500
s2	0.143	0.381	0.286	0.190
s3	0.143	0.571	0.071	0.214

Normalización

- Composicionalidad

Varios problemas con los datos composicionales.

Pueden Presentar correlaciones espúreas al agregar o eliminar variables.



Normalización

- **Composicionalidad**

Las proporciones permanecen intactas.

Tabla 1. Conteo absoluto

	A	B	C	D	E
s1	10	20	20	50	50
s2	15	40	30	20	200
s3	20	80	10	30	15

$$s1 \quad A/B = 10/20 = 0.5$$

Tabla 2. Conteo relativo

	A	B	C	D	E
s1	0.067	0.133	0.133	0.333	0.333
s2	0.049	0.131	0.098	0.066	0.656
s3	0.129	0.516	0.065	0.194	0.097

$$s1 \quad A/B = 0.067/0.133 = 0.5$$

Tabla 3. Conteo relativo sin E

	A	B	C	D
s1	0.100	0.200	0.200	0.500
s2	0.143	0.381	0.286	0.190
s3	0.143	0.571	0.071	0.214

$$s1 \quad A/B = 0.1/0.2 = 0.5$$

Normalización

- Transformación de relación logarítmica centrada (*centered log-ratio* (clr)).

Normalización

- Transformación de relación logarítmica centrada (*centered log-ratio* (clr)).

$$clr(x) = \left[\log\left(\frac{x_1}{g(x)}\right), \log\left(\frac{x_2}{g(x)}\right), \log\left(\frac{x_3}{g(x)}\right), \dots, \log\left(\frac{x_D}{g(x)}\right) \right]$$

Donde

$$g(x) = \sqrt[D]{x_1 \cdot x_2 \cdot x_3 \dots x_D}.$$

Normalización

- **Transformación de relación logarítmica centrada (*centered log-ratio* (clr)).**

Problemáticas con matrices dispersas:

1. **Ceros de muestreo:** taxa presente de muy baja abundancia
2. **Ceros estructurales:** taxa que no está presente.

Normalización

- **Transformación de relación logarítmica centrada (*centered log-ratio* (clr)).**

Problemáticas con matrices dispersas:

1. **Ceros de muestreo:** taxa presente de muy baja abundancia
2. **Ceros estructurales:** taxa que no está presente.




Normalización

- **Transformación de relación logarítmica centrada (*centered log-ratio* (clr)).**

Dependiente del tamaño de la muestras (i.e., número de reads por muestra).

Tabla 1. Conteo absoluto

	A	B	C	D	E	Suma
s1	10	20	20	50	50	150
s2	150	400	300	200	2000	3050
s3	20	80	10	30	15	155



~x20

Distancias

- **Distancia Euclidea**

Dados dos vectores X e Y:

$$X = (x_1, x_2, x_3, \dots, x_D); Y = (y_1, y_2, y_3, \dots, y_D)$$

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + \dots + (x_D - y_D)^2}$$

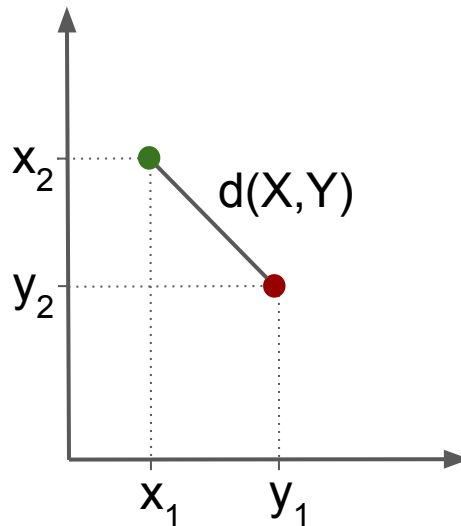
Distancias

- **Distancia Euclidea**

Dados dos vectores X e Y:

$$X = (x_1, x_2); Y = (y_1, y_2)$$

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$



Distancias

- **Distancia Euclidea**

Problema del doble cero (muy común en matrices dispersas).

Tabla 1. Conteo absoluto

	A	B	C	D	E
s1	0	0	20	50	50
s2	0	0	30	0	200
s3	20	80	10	0	15

Distancias

- Distancia de Aitchison

$$d(X, Y) = \sqrt{\left(\log\left(\frac{x_1}{g(X)}\right) - \log\left(\frac{y_1}{g(Y)}\right)\right)^2 + \left(\log\left(\frac{x_2}{g(X)}\right) - \log\left(\frac{y_2}{g(Y)}\right)\right)^2 + \dots + \left(\log\left(\frac{x_D}{g(X)}\right) - \log\left(\frac{y_D}{g(Y)}\right)\right)^2}$$

Disimilitud

- **Disimilitud Bray-Curtis**

$$BC_{xy} = 1 - \frac{2C_{xy}}{S_x + S_y}$$

Donde C_{xy} es la suma de especies con menor valor compartidas.

S_x y S_y es la suma de todas las especies en X e Y, resp.

Disimilitud

- **Disimilitud Bray-Curtis**

$$BC_{xy} = 1 - \frac{2C_{xy}}{S_x + S_y}$$

Donde C_{xy} es la suma de especies con menor valor compartidas.

S_x y S_y es la suma de todas las especies en X e Y, resp.



Conveniente para trabajar con datos de comunidades.

No tenemos el problema del doble cero.

Disimilitud

- **Disimilitud Bray-Curtis**

$$BC_{xy} = 1 - \frac{2C_{xy}}{S_x + S_y}$$

Donde C_{xy} es la suma de especies con menor valor compartidas.

S_x y S_y es la suma de todas las especies en X e Y, resp.



Notar que Bray-Curtis no es una distancia, sino una disimilitud.

No cumple la desigualdad triangular.

Distancias

- **Distancia UniFrac**

Mide la diferencia entre dos muestras como la cantidad de historia evolutiva que es exclusiva de cualquiera de estas.

Distancias

- **Distancia UniFrac**

Mide la diferencia entre dos muestras como la cantidad de historia evolutiva que es exclusiva de cualquiera de estas.

$$\frac{(\text{Distancia de largos de rama no compartidos})}{(\text{Suma de todos los largos de rama})} = \text{Fracción de largos de rama no compartidos}$$

Distancias

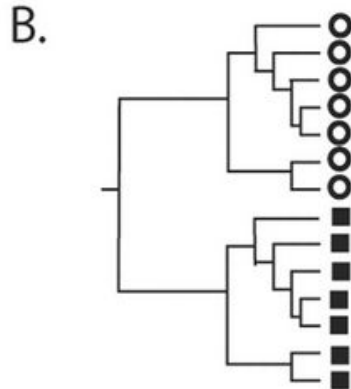
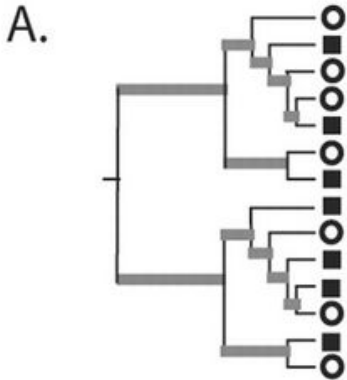
- **Distancia UniFrac**

Mide la diferencia entre dos muestras como la cantidad de historia evolutiva que es exclusiva de cualquiera de estas.

(Distancia de largos de rama no compartidos)

(Suma de todos los largos de rama)

= Fracción de largos de rama no compartidos



Distancias

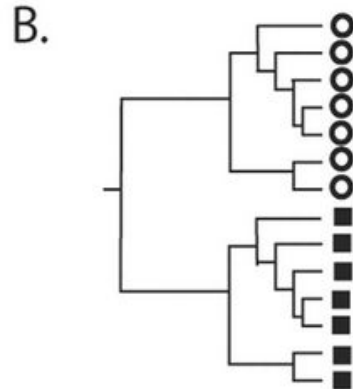
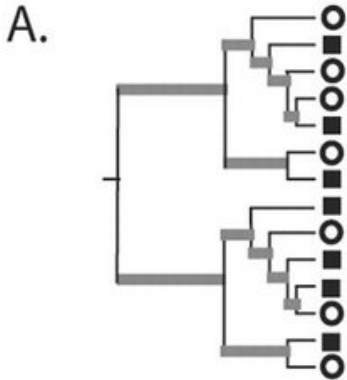
- **Distancia UniFrac**

Mide la diferencia entre dos muestras como la cantidad de historia evolutiva que es exclusiva de cualquiera de estas.

(Distancia de largos de rama no compartidos)

(Suma de todos los largos de rama)

= Fracción de largos de rama no compartidos



A) Comunidades similares

B) Comunidades distintas

Bibliografía recomendada

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol*;8:2224. DOI: 10.3389/fmicb.

Robert C. Edgar. (2016). SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *BioRxiv*. DOI: 10.1101/074161.

Vinje, H., Liland, K.H., Almøy, T. et al. (2015). Comparing K-mer based methods for improved classification of 16S sequences. *BMC Bioinformatics* **16**, 205 . DOI: 10.1186/s12859-015-0647-4.

Weiss, S., Xu, Z.Z., Peddada, S. et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5, 27. DOI:10.1186/s40168-017-0237-y.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol*. 73(16):5261-7. DOI: 10.1128/AEM.00062-07.

Yarza, P., Yilmaz, P., Priesse, E. et al. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* 12, 635–645. DOI: 10.1038/nrmicro3330