Recibido: 6/6/2022 / Aceptado: 17/7/2022
doi: https://doi.org/10.26439/interfases2022.n015.5896

CLASIFICACIÓN DE ORGANISMOS EN REINOS UTILIZANDO FRECUENCIA DE CODONES DE ADN

Luis Beltrán Palma Ttito
luis.palma@unsaac.edu.pe / ORCID: 0000-0002-0950-5369
Departamento Académico de Ingeniería Informática
Universidad Nacional de San Antonio Abad del Cusco, Perú

Resumen

Este estudio de tiene por objetivo utilizar clasificadores de machine learning para predecir el reino al que pertenece un organismo por la frecuencia de uso de codones de ADN. Para ello se ha tomado 13 028 datos de organismos del GenBank distribuidos en once reinos y se los redujo a seis reinos (arqueas, bacterias, invertebrados, plantas, virus y vertebrados) con 9027 datos reagrupados. El proceso requirió la, depuración de atributos irrelevantes, el empleo de métricas de medición de clasificadores de exactitud, precisión, sensibilidad y puntuación, así como el ajuste de hiperparámetros de los modelos. Los algoritmos de clasificación fueron *voting*, *bagging*, *boosting* y *stacking*, usando KNN, AD, MLP, SVC y RF. La selección de atributos se hizo con random forest. El ensamble stacking, con sus modelos, predice mejor la clasificación de organismos en el presente estudio.

PALABRAS CLAVE: *machine learning* / ensambles / frecuencia de codones ADN / reino

CLASSIFICATION OF ORGANISMS INTO KINGDOMS USING DNA CODON FREQUENCY

Abstract

This study aims to use machine learning classifiers to predict the kingdom to which an organism belongs by the frequency of use of DNA codons. The study used 13,028 data from GenBank organisms distributed in eleven kingdoms and reduced them to six kingdoms (archaea, bacteria, invertebrates, plants, viruses, and vertebrates) with 9,027 regrouped data. The process required cleaning irrelevant attributes, using measurement metrics of accuracy, precision, sensitivity, and score classifiers, and the adjustment of hyperparameters of the models. The classification algorithms were voting, bagging, boosting, and stacking, using KNN, AD, MLP, SVC, and RF. Random forest was used in selecting the attributes. The stacking ensemble, with its models, better predicts the classification of organisms in the present study.

KEYWORDS: machine learning / ensembles / DNA codon frequency / kingdom /

1. INTRODUCCIÓN

Un enfoque novedoso para la clasificación de organismos y también de genes es la frecuencia de uso de codones en el ADN codificante. Se conoce que los diferentes organismos tienen la tendencia a utilizar determinados codones para la producción de aminoácidos, por lo tanto esta particularidad es innata a cada grupo de organismos (Im & Choi, 2017; Sharp, 2010), entonces ¿el análisis de la frecuencia de uso de un determinado codón será capaz de clasificar a los organismos en reinos? Para responder a la pregunta definimos como objetivo el predecir el reino al que pertenece un organismo por la frecuencia de uso de codones, utilizando clasificadores de *machine learning*.

Para esta investigación se utilizan las frecuencias de cada una de las 9027 secuencias de codificación proteica completas (CDS) que se han compilado a partir de las divisiones taxonómicas de la base de datos de secuencias de ADN del GenBank. Los archivos de datos se pueden obtener de los sitios ftp anónimos de DDBJ, Kazusa y EBI.

2. ESTADO DEL ARTE

Diversos genes y secuencias de ADN han sido utilizadas para realizar la clasificación taxonómica y filogenética de los organismos, así se utilizan secuencias de los genes ribosomales, rRNA 16S para los organismos procariotas y rRNA 18S para muchos organismos eucariotas, lo cual ha definido la clasificación actual en tres dominios de la naturaleza: bacteria, archaea y eucaria. Sin embargo, el uso de dichas secuencias no responde a la diferenciación para que pueda ser utilizado de forma universal.

Hay antecedentes de uso del *codon usage* como herramienta para predecir y clasificar características genómicas y evolutivas entre los tres dominios de la naturaleza. Por otro lado, también ha resultado útil para predecir el tipo de ADN, como, por ejemplo, ADN nuclear, mitocondrial o de cloroplastos, incluso se propuso su uso al interior de las secuencias de determinadas proteínas, de tal forma que ayuden a predecir anomalías. Utilizando el sesgo de uso de codones, se puede identificar orígenes evolutivos y composición genética, pero se tiene que tomar en cuenta que la alta data disponible dificulta su manipulación, por lo que métodos de clasificación de *machine learning* empiezan a cobrar importancia, como una herramienta que nos permita sistematizar y clasificar los datos genómicos.

Se pueden analizar variaciones intraespecies en cuanto al uso del codón, por ejemplo, se ha reportado que los niveles de producción de proteínas pueden ser predecidos de secuencias del genoma completo utilizando las bases de datos de proteínas ribosomales. No se necesita analizar todas las proteínas, el análisis de un solo grupo puede dar respuestas, sin embargo, hay otros datos que indican que el uso de determinados codones está aún relacionado a la expresión de determinadas proteínas,

lo cual lo complica más, sin embargo, es un modo interesante para ver la expresión de proteínas no solo a nivel de especies sino también a nivel de grupos de proteínas dentro de un mismo organismo (Nakamura et al., 2000)

En el mundo globalizado donde se comparten datos genómicos, se cuenta con bases de datos que compila los codones de proteínas completas, estos datos son utilizados para la anotación de genomas, y a su vez la secuencia de los genomas contiene información importante que puede ser utilizada en la probabilidad de uso como un modo de conseguir no solo clasificar a los organismos, sino identificarlos a través de los sesgos de un código genético degenerado.

3. DISEÑO DEL EXPERIMENTO

3.1 Descripción del conjunto de datos

El conjunto de datos del presente estudio consta de 13 028 datos de organismos, los que registran la frecuencia de usabilidad de cada uno de los codones de ADN, siendo la cantidad de atributos de 69, que detallamos a continuación:

- *Kingdom*: Reino al que pertenece el organismo, con una codificación de 3 caracteres de acuerdo al código CUTG.
- *DANtype*: Tipo o fuente de origen del ADN, expresado con valores numéricos.
- *SpeciesID*: Identificador único en números enteros.
- *Ncodons*: Cantidad de codones de la muestra.
- *SpeciesName*: Nombre descriptivo del organismo.
- *64 Codones*: Frecuencia relativa de los 64 codones de ADN, con 5 cifras significativas.

Podemos apreciar un porcentaje de los datos del estudio en la tabla 1.

Tabla 1

Conjunto de datos del estudio

Kingdom	DNAtype	SpeciesID	Ncodons	SpeciesName	UUU	.....	UGA
vrl	0	100220	1474	Bohle iridovirus	0,02714		0
vrl	0	100755	4862	Sweet potato feaf curl virus	0,01974		0,00144
vrl	0	100880	1915	Northern cereal mosaic virus	0,01775		0

(*continúa*)

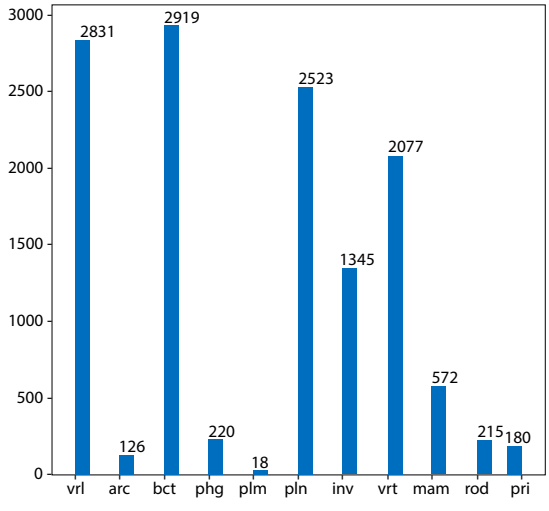
(*continuación*)

vrl	0	100887	22831	Soil-borne cereal mosaic virus	0,02816	0,00131
vrl	0	101029	5274	Human adenovirus type 7d	0,02579	0,00209
vrl	0	101688	3042	Apple latent spherical virus	0,04635	0
vrl	0	101764	2801	Aconitum latent virus	0,02285	0,00071
vrl	0	101947	2897	Pseudoarabies virus Ea	0,01105	0,00138
vrl	0	10249	61247	Vaccina virus Copenhagen	0,03411	0,00103
vrl	0	10253	55330	Vaccina virus Tlan Tan	0,03441	0,00112

Siendo el interés del presente estudio, la clasificación de organismos en reinos, la figura 1, muestra la distribución de los organismos en 11 reinos.

Figura 1

Distribución de organismos por reino



3.2 Metodología

La figura 2 muestra la metodología empleada para el desarrollo del presente estudio, en la primera fase se selecciona un conjunto de muestras del conjunto de datos;

posteriormente, se aplica la selección de atributos utilizando *random forest*; en la tercera, se realiza entrenamiento de cuatro ensambles de clasificación, para finalmente realizar el comparativo de los modelos entrenados.

FIgura 2

Metodología del estudio

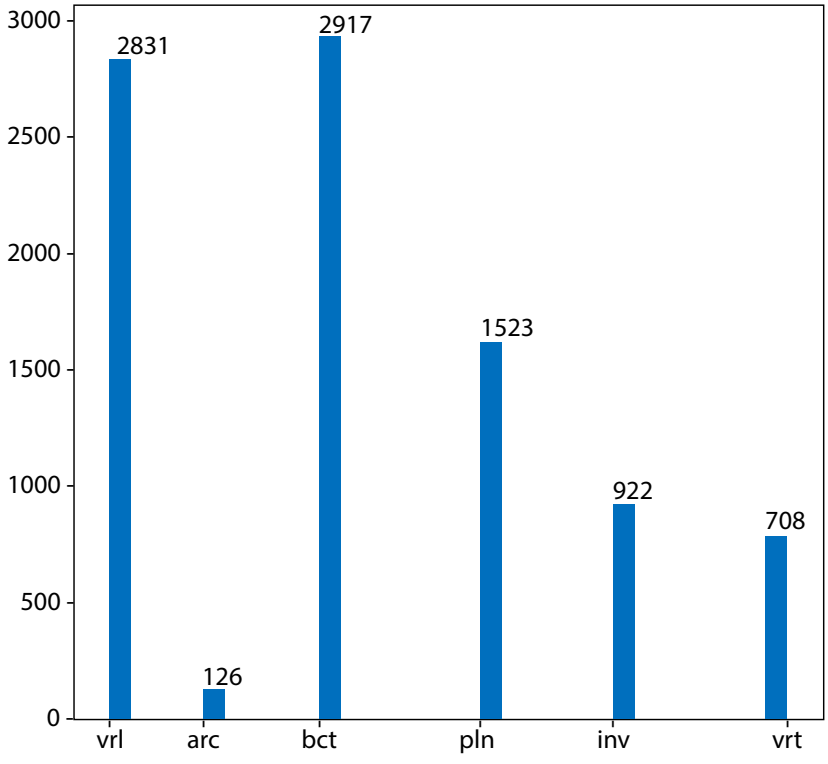
3.2.1 Selección de muestras

El conjunto de datos originales posee 11 clases en el atributo reino, de las cuales dos fueron retiradas: plásmidos (plm) y bacteriófagos (phg), ya que no fueron considerados en la clasificación del presente estudio. Luego se re-etiquetaron mamíferos (mam), primates (pri) y roedores (rod) por vertebrados (vrt). Quedándonos con 6 clases: arqueas (arc), bacterias (bct), invertebrados (inv), vertebrados (vrt), plantas (pln), virus (Vrl). También fueron eliminados los datos que no corresponden al ADN genómico.

Finalmente, la distribución de clases se aprecia en la figura 3, con un total de 9027 datos.

Figura 3

Distribución de reagrupamiento de clases



3.2.2 Cantidad de datos de entrenamiento y test

La cantidad de datos utilizados para el entrenamiento y test se aprecia en la tabla 2.

Tabla 2

Distribución de datos de entrenamiento y test

	Total	Train	Test
Arqueas	126	95	31
Bacterias	2917	2178	739
Invertebrados	922	709	213
Plantas	1523	1133	390
Virus	2831	2110	721
Vertebrados	708	545	162
∑	9027	6770	2256

3.2.3 Selección manual de atributos

Ya que el propósito del presente estudio es la identificación del reino al que pertenece un organismo, en función al uso de codones, fueron eliminados los atributos: DANtype, SpeciesID, NCodons y SpeciesName, por sugerencia del experto humano.

3.2.4 Selección de métricas de medición de clasificadores

Por tener una distribución de clases desequilibrada, se propone utilizar diferentes métricas: exactitud, precisión, sensibilidad y puntuación F1, las que se calculan por macro promedio:

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN} \quad (1)$$

$$Sensibilidad = \frac{VP}{VP + FN} \quad (2)$$

$$Precisión = \frac{VP}{VP + FP} \quad (3)$$

$$Puntuación\ F1 = 2 * \frac{Precisión * Sensibilidad}{Precisión + Sensibilidad} \quad (4)$$

Donde: *VP* es verdadero positivo; *VN* es verdadero negativo; *FP* es falso positivo; *FN* es falso negativo.

3.2.5 Estrategia de validación y ajuste de hiperparámetros

Para la búsqueda de ajuste de hiperparámetros de los diferentes modelos, se utiliza la búsqueda aleatoria en cuadrícula con validación cruzada, con un máximo de 100 iteraciones y una validación cruzada de 3.

3.2.6 Algoritmos de clasificación

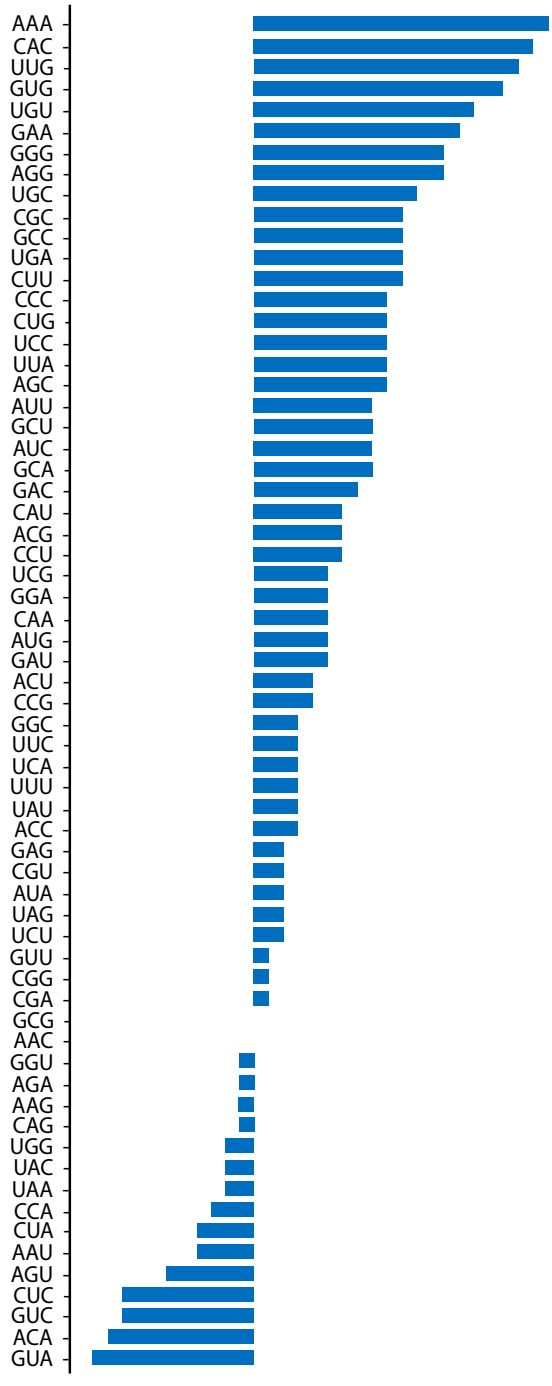
Los algoritmos de clasificación a utilizar corresponden a cuatro arquitecturas de ensamble: *voting*, *bagging*, *boosting* y *stacking*, en las que se hace uso de k-vecinos más cercanos (KNN), árbol de decisión (AD), red neuronal perceptrón multicapa (MLP), máquina de vector de soporte para clasificación (SVC) y ensamble, bosque aleatorio (RF).

3.2.7 Selección de atributos mediante random forest

Se hace uso de bosque aleatorio (*random forest*) para determinar variables predictoras, las que podemos apreciar en la figura 4, esta información es utilizada en los experimentos de diferentes modelos de clasificación.

Figura 4

Importancia de variables en la predicción del reino



4. EXPERIMENTACIÓN Y RESULTADOS

La tabla 3 muestra los mejores modelos encontrados para cada ensamble.

Tabla 3

Ensambles encontrados en la clasificación de seis reinos a partir de frecuencias de codones de ADN

	Hiperparámetro de ensamble	Modelo	Hiperparámetro de modelo	Exactitud
Voting	pesos de modelos = 1	AD	criterio separación = entropia, prof. árbol = 20, vc =5	89,46%
	votación mayoría simple	KNN	k vecinos = 3, métrica distancia = euclidiana, vc =5	
	cv folds = 5	SVC	gamma = 5, kernel = función de base radial, costo = 300, cv = 5	
	repeticiones = 10	MLP	capas ocultas = (1000,500,200,80,10), alfa = 0.01, fun. activación = tanh, cv =5	
Bagging	cantidad svc = 30 cv folds = 5 repeticiones = 10	SVC	Costo = 1, kernel = función de base radial, degree = 3, gamma = scale,	90,17%
Boosting	cv folds = 5 repeticiones = 10		cant árboles = 90, prof. Árbol = 30, coef. entrenamiento = 0.2	88,29%
Stacking	cv folds = 5		k vecinos = 3, métrica distancia = euclidiana	90,97%
	repeticiones = 10		prof. árbol = 50, criterio separación = entropia	
			gamma = 5, costo = 300, kernel = función de base radial	
			capas ocultas = (500,200,100,50), alfa = 0.01, activación = thanh	
			n.º árboles = 300	

Tabla 4

Presencia de algunos de los datos mal clasificados

Kingdom	Pred	UUU	UUC	UUA	...	UAA	UAG	UGA
inv	vrt	0,00717	0,01402	0,00163		0,00049	0,00016	0,00033
pln	vrl	0,03234	0,00667	0,029		0,00128	0,00205	0,00051
vrt	inv	0,01131	0,01654	0,00522		0	0,00087	0,00087
arc	bct	0,02566	0,01537	0,01202		0,00266	0,0015	0,00162
inv	vrt	0,01782	0,02858	0,00928		0,00074	0	0,00148
arc	bct	0,00089	0,022	0,00268		0,00089	0,00059	0,00119
vrl	pln	0,02682	0,02778	0,00479		0	0,00096	0,00287
inv	vrt	0,01828	0,02611	0		0,00261	0	0
vrl	bct	0,00488	0,03245	0,00096		0,00188	0,00092	0,00218
pln	bct	0,01427	0,02379	0,00666		0	0,00095	0
bct	pln	0,02274	0,01005	0,01546		0,0022	0,0004	4,00E-05
pln	vrl	0,01279	0,01918	0,0024		0	0,0024	0
bct	vrl	0,01004	0,03161	0		0	0,00355	0,00059
bct	inv	0,01495	0,02415	0,0046		0,00173	0,00115	0,00115

Nota. La tabla muestra algunos de los datos mal clasificados. La columna *Kingdom* registra la etiqueta correcta y la columna *Pred* registra la predicción del mejor modelo.

En la tabla 4 se muestran las predicciones incorrectas, cuya frecuencia relativa se aprecia en la tabla 5.

Tabla 5

Frecuencia relativa de organismos mal clasificados por el mejor modelo encontrado

	Total de predicciones	Predicciones incorrectas	Frecuencia %
Arqueas	31	5	16,13
Bacterias	739	26	3,52
Invertebrados	213	40	18,78
Plantas	390	8	2,05
Virus	721	17	2,36
Vertebrados	163	16	9,82
∑	2257	112	

5. DISCUSIÓN DE LA EXPERIMENTACIÓN Y RESULTADOS

Para realizar una aproximación más estricta de la clasificación de los organismos en estudio, se descartaron algunos datos de la matriz inicial, porque no cumplían con la posición taxonómica de reino, otros datos se reclasificaron para conseguir una mejor agrupación y únicamente se trabajó con datos de ADN genómico, para evitar el sesgo de excepcionalidades que tiene el código genético en estructuras intracelulares como los plásmidos, mitocondrias y/o cloroplastos. Para conocer la capacidad predictora de los 64 codones, estos se evaluaron utilizando RF y se determinaron 17 codones que no tenían capacidad predictora por lo que se realizaron ensayos sin utilizarlos (datos no mostrados). Los resultados indican que el uso de estos 17 codones no incrementa mayor predicción en la clasificación de reinos, por lo que se puede prescindir de su uso. Por otro lado, será necesario un análisis más profundo para conocer las razones biológicas de su poca o ninguna capacidad de clasificación en grupos taxonómicos.

Khomtchouk (2020), con la misma base de datos, en el trabajo "Los niveles de sesgo en el uso de codones predicen la identidad taxonómica y la composición genética", muestra que los niveles de sesgo de uso de codones de un organismo pueden servir como predictor y clasificador de varias características genómicas y evolutivas en los tres dominios de la vida (arqueas, bacterias, eucaria). Teniendo los siguientes resultados:

Tabla 6

Reultados de la clasificación del reino según Khomtchouk

Model	*Precision*	*Recall*	*Micro F1-Score*	*Macro F1-Score*	*Accuracy*	*AUC*
k-Nearest Neighbors	0,9660	1	0,9827	0,9293	0,9669	0,9792
Random Forests	0,9298	1	0,9636	0,8611	0,9298	0,9954
Extreme Gradient Boosting	0,9502	1	0,9745	0,8846	0,9502	0,9970
Artificial Neural Networks	0,9132	1	0,9546	0,8425	0,9132	0,9901
Naive Bayes	0,7200	0,3529	0,4737	0,5487	0,6561	0,8410

Nota. Resultado obtenidos de la clasifiación del reino realizada por Khomtchouk (2020)

Donde su mejor clasificador es KNN el cual tiene una precisión de 0,9660 para una clasificación de 3 dominios, lo cual significa su capacidad de discernimiento en una clasificación taxonómica superior a reino. En comparación con nuestra investigación, con el ensamble de *stacking*, tiene una precisión de 0,9097 para la clasificación de 6 reinos, lo cual es muy buena.

A pesar de alcanzar unos índices de precisión y exactitud altos, se observa que con los modelos empleados no logramos la clasificación de la totalidad de los organismos ensayados, han sido 135 organismos que no son correctamente clasificados, una posible explicación es la distribución de datos, que se aprecia en la figura 3, y es la disponibilidad de datos desbalanceada, que podría inducir a los modelos de aprendizaje y aplicación, una predicción incorrecta, probablemente con un mejor balance de distribución de la frecuencia de datos podrían mejorar los resultados obtenidos. Por otro lado, la explicación podría encontrarse en la característica del código genético degenerado, particularidades que deben de ser analizadas con cuidado.

6. CONCLUSIONES

En primer lugar, las técnicas de *machine learning* no responden de forma adecuada a la heterogeneidad de los datos o desbalance en el número de datos. En segundo lugar, se ha comprobado la existencia de 17 codones que tienen menos importancia en el agrupamiento en reinos. Finalmente, se necesita una corrección de excepcionalidades del código genético universal previo a la aplicación de la técnicas de *machine learning*.

7. TRABAJOS FUTUROS

Para trabajos futuros proponemos el uso de *deep learning* y de redes convolucionales, utilizando optimizaciones de *simple gradient descent update* y *mini-batch gradient descent*. Además, el de incrementar datos del reino *arqueas*, utilizando las herramientas de generación de datos sintéticos como DataProf y Iri Rougen, y comprobar estos datos con el algoritmo de generación de datos sintéticos basados en reglas.

REFERENCIAS

Khomtchouk, B. B. (2020). Codon usage bias levels predict taxonomic identity and genetic composition. *BioRxiv. The Preprint Server for Biology*. https://doi.org/10.1101/2020.10.26.356295.

Im, E.-H., & Choi, S. S. (2017). Synonymous codon usage controls various molecular aspects. *Genomic & Informatics*, *15*(4), 123-127. https://doi.org/10.5808/GI.2017.15.4.123.

Nakamura, Y, Gojobori, T, & Ikemura, T. (2000). Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Research*, 28(1), 292. https://doi.org/10.1093/nar/28.1.292

Parvathy, S. T., Udayasuriyan, V., & Bhadana, V. (2021). Codon usage bias. *Molecular Biology Reports*, *49*, 539-565. https://doi.org/10.1007/s11033-021-06749-4

Sharp, P. M., Emery, L. R., & Zeng, K. (2010). Forces that influence the evolution of codon bias. *Philosophical Transactions of the Royal Society B. Biological Sciences*, *365*(1544), 1203-1212. https://doi.org/10.1098/rstb.2009.0305 Wang, F.-P., & Li, H. (2009). Codon-pair usage and genome evolution. Gene, *433*(1-2), 8-15. https://doi.org/10.1016/j.gene.2008.12.016