

Práctico 7

Selección Natural: análisis a nivel molecular

Objetivo: Familiarizarse con el concepto de selección natural y algunos métodos para detectar su acción usando datos moleculares.

Introducción

Cuando la selección natural actúa sobre las poblaciones deja huellas que pueden ser reconocidas en el ADN. Para identificar esas huellas se han desarrollado diferentes pruebas aplicables a secuencias nucleotídicas codificantes de proteínas. Una aproximación robusta y sencilla desarrollada por McDonald y Kreitman es considerar que, bajo neutralidad, la relación entre la tasa de cambio nucleotídico sinónimo (dS) y no sinónimo (dN) o de reemplazo aminoacídico será la misma dentro y entre poblaciones. Cualquier desviación sugiere un apartamiento de la neutralidad, incluyendo algún tipo de selección positiva. Si no se cuenta con información poblacional, otra aproximación muy utilizada aunque exigente es considerar que, bajo neutralidad estricta, ambas tasas deberían ser iguales, por lo que dN/dS, también conocido como ω será 1. Si dN supera ampliamente dS, es decir si $\omega > 1$, se asume que actuó selección positiva (el caso inverso, $\omega < 1$ indicaría selección purificadora). En este práctico aplicaremos ambas aproximaciones.

Ejercicio 1 - Test de McDonald y Kreitman

Cuando McDonald y Kreitman en 1991 propusieron su test de neutralidad, lo aplicaron a un conjunto de secuencias de la enzima Alcohol deshidrogenasa (Adh), para tres especies diferentes.

La Tabla 1 muestra el resumen de los sitios variables de las secuencias de Adh incluidas en la base de datos.

- Interprete la Tabla 1: ¿qué hay en las filas y las columnas?
- Completar los siguientes cuadros a partir de los datos de la tabla.
- Utilizando los cuadros completados, realizar el test de McDonald y Kreitman (MK) y sacar conclusiones.

	<i>D. simulans vs. D. yakuba</i>	
	Polimorfismos	Sustituciones
Reemplazo		
Sinónimos		

	<i>D. melanogaster</i> vs. <i>D. yakuba</i>	
	Polimorfismos	Sustituciones
Reemplazo		
Sinónimos		

	<i>D. melanogaster</i> vs. <i>D. simulans</i>	
	Polimorfismos	Sustituciones
Reemplazo		
Sinónimos		

Tabla 1. Resumen de los sitios variables de las secuencias de Adh usadas por McDonald y Kreitman (1991) para proponer su test de neutralidad. La primera columna a la izquierda indica la posición del sitio en cuestión, la segunda es la secuencia consenso de referencia; el símbolo – indica un nucleótido idéntico a la secuencia consenso. Se muestran los estados para cada sitio para 12, 6 y 12 individuos de las especies *D. melanogaster*, *D. simulans* y *D. yakuba* respectivamente. Sitios subrayados indican individuos heterocigotas (portan la variante de consenso y la subrayada).

Ejercicio 2 - Variación en el ω entre sitios y entre linajes

En este ejercicio usaremos las secuencias de la Hemaglutinina (HA) del virus de la gripe. Esta proteína es una glucoproteína antigénica que se encuentra en la superficie del virus y es la mayor responsable de la unión del virus a la célula infectada. Esta proteína es muy estudiada en el diseño de vacunas, porque presenta una evolución asimétrica que sugiere una fuerte selección de aquellas variantes que son las que mejor escapan al sistema inmune del hospedero. [Podemos ver la estructura 3D en este link.](#) Además, el análisis de los cambios nucleotídicos sinónimos y no sinónimos muestra que muchos residuos aminoacídicos en la HA concentrados en el extremo distal y externo de la proteína (que son aquellos sitios que interactúan con el sistema inmune del hospedero) están siendo seleccionados positivamente (Fig. 1).

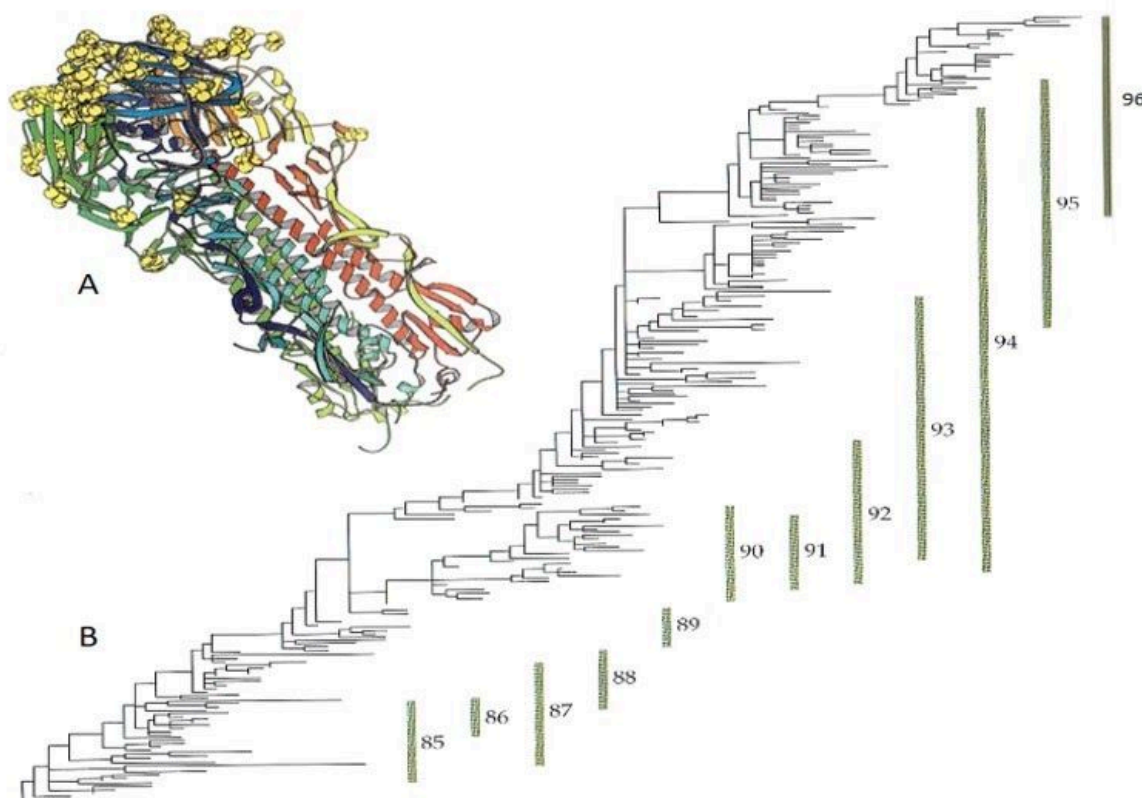


Fig 1. A) Modelo tridimensional de la Hemaglutinina del virus de la gripe, mostrando los sitios aminoacídicos seleccionados positivamente para cambiar. B) Filogenia de cepas del virus aisladas desde 1985 a 1996, basada en el análisis de las secuencias nucleotídicas de ese gen (Tomado de Hillis, 2009).

A continuación, estimaremos la tasa de cambio nucleotídico sinónimo (dS) y no sinónimo (dN) y su relación mediante máxima verosimilitud (ML). Estas estimaciones pueden ser realizadas para cada codón y/o linaje, considerando una filogenia que permita estimar las secuencias nucleótidos en cada uno de los nodos en la filogenia. Dado la gran cantidad de cálculos necesarios para esta tarea, esta estrategia es preferible hacerla en servidores online.

Utilizaremos un servidor llamado Datamonkey (<https://www.datamonkey.org/>) y usaremos el método SLAC (Single-Likelihood Ancestor Counting). SLAC es una de las tantas formas de evaluar sitios bajo selección. Utiliza una combinación de enfoques de ML y conteos para inferir dN y dS por sitio para un conjunto de secuencias nucleotídicas codificantes de proteínas alineadas y la filogenia que las vincula. SLAC comienza optimizando las longitudes de las ramas y los parámetros de sustitución de nucleótidos bajo un modelo complejo (denominado MG94xREV). Sin embargo, en lugar de usar ML para ajustar los parámetros dN y dS específicos del sitio, SLAC usa ML para inferir la secuencia ancestral más probable en cada nodo de la filogenia. De esta forma se comparan las secuencias nucleotídicas entre nodos adyacentes, y luego cuenta directamente el número total de cambios no sinónimos y sinónimos que se han producido en cada sitio (para eso emplea una

versión modificada del método de recuento Suzuki-Gojobori). A su vez, esta aproximación asume que la presión de selección para cada sitio es constante a lo largo de toda la filogenia, algo poco realista.

Actividad

Usamos el archivo fasta SecuenciasHA.fas (en la carpeta de Prácticos). Vamos a ver los resultados en el siguiente link:

<https://www.datamonkey.org/slac/68e5640217c96c384b69e1fe>

Interprete los resultados tomando un nivel de significancia de $p\text{-value} = 0,2$. Ordene los resultados por $P[dN/dS > 1]$ y $P[dN/dS < 1]$.

¿Existe algún sitio y/o linaje seleccionado? ¿Qué sitio/s presenta/n una fuerte evidencia de selección positiva? ¿Qué valores de dN y dS tiene ese sitio? ¿Qué cambios aminoacídicos se registran en ese sitio? ¿Dónde cree que se ubicará ese sitio en la proteína dados los antecedentes planteados.

Ubicar los sitios bajo selección en el esquema "SLAC Site Graph". Acá podemos ordenar la tabla viendo la evidencia de selección positiva [$P [dN/dS > 1]$] y purificadora [$P [dN/dS < 1]$].

¿Qué similitudes y diferencias encuentra entre SLAC y el test de MK?

Ahora probemos otro enfoque. Podemos usar MEME (Mixed Effects Model of Evolution), que permite estimar selección donde sólo algunas de las ramas han experimentado presiones selectivas. Aca permitimos detectar sitios individuales bajo selección episódica diversificadora. Vemos los resultados en el siguiente link:

<https://www.datamonkey.org/meme/68e5644517c96c384b69e38e>.

En los resultados, en Figura 1. Primero visualizo los p-valor de los test de selección.

Alignment-wide results

Plot type

Figure 1. P-values derived from the asymptotic mixture model.
Solid line = user selected significance threshold.

Tabla 1 en resultados. Ordenar los resultados por p-value.

Identificar y discutir qué tipo de selección se infiere acá. Se puede visualizar un árbol con los cambios en los codones bajo selección.

Tree to view

Codon 226

Branch length

Global MG94xREV

Tree labels

☐ amino-acids ☐ codons ☐ show internal

☐ show only multiple hits ☒ show only non-synonymous changes

☐ sequence names ☐ align tips

Color branches

Support for selection

De forma similar probemos FUBAR (Fast, Unconstrained Bayesian AppRoximation). que emplea un enfoque bayesiano para inferir tasas de sustitución no sinónimas (dN) y sinónimas (dS) por sitio para un alineamiento y la filogenia correspondiente. Este método supone que la presión de selección en cada sitio es constante a lo largo de toda la filogenia. Vemos los resultados en el siguiente link:

<https://www.datamonkey.org/fubar/68e5644a17c96c384b69e4c1>.

Ordenar los resultados por BayesFactor[$\alpha < \beta$]

Si tenemos en cuenta distintos modelos.. ¿Tenemos resultados congruentes?

Referencias

1. McDonald, J., Kreitman, M. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351, 652–654 (1991). <https://doi.org/10.1038/351652a0>
2. Hillis, DM. (2009). Phylogenetic Progress and Applications of the Tree of Life. En: Evolution since Darwin: The First 150 Years, pp. 421-449. Eds: MA Bell, DJ Futuyma, WF Eanes, JS Levinton. Sinauer Associates, Inc. • Publishers Sunderland, Massachusetts U.S.A.
3. Kosakovsky Pond, SL and Frost, SDW. "Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection." Mol. Biol. Evol. 22, 1208--1222 (2005).
4. Murrell, B et al. "Detecting individual sites subject to episodic diversifying selection." PLoS Genetics 8, e1002764 (2012).