Práctico 6

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.

Actividad

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.

	D. simulans vs. D. yakuba	
	Sustituciones	Polimorfismos
Reemplazo		
Sinónimos		

	D. melanogaster vs. D.yakuba	
	Sustituciones	Polimorfismos
Reemplazo		
Sinónimos		

	D. melanogaster vs. D. simulans	
	Sustituciones	Polimorfismos
Reemplazo		
Sinónimos		

TABLE 1 Variable nucleotides from the coding region of the *Adh* locus in *D. melanogaster*, *D. simulans* and *D. yakuba*

	Con.	*D. melanogaster* a b c d e f g h i j k l	*D. simulans* a b c d e f	*D. yakuba* a b c d e f g h i j k l		
781	G	T T T T T T T T T T T T	– – – – – –	– – – – – – – – – – – –	Repl.	Fixed
789	T	– – – – – – – – – – – –	– – – – – –	C C C C C C C C C C C C	Syn.	Fixed
808	A	– – – – – – – – – – – –	– – – – – –	G G G G G G G G G G G G	Repl.	Fixed
816	G	T T T T – – – – – – – T	T T T T T T	– – – – – – – – – – – –	Syn.	Poly.
834	T	– – – – – – – – – – – –	C C – – – C	– – – – – – – – – – – –	Syn.	Poly.
859	C	– – – – – – – – – – – –	– – – – – –	G G G G G G G G G G G G	Repl.	Fixed
867	C	– – – – – – – – – – – –	– – – – – –	G G G G G A G G G G G G	Syn.	2 Poly.
870	C	T T T T T T T T T T T T	– – – – – –	– – – – – – – – – – – –	Syn.	Fixed
950	G	– – – – – – – – – – – –	– A – – – –	– – – – – – – – – – – –	Syn.	Poly.
974	G	– – – – – – – – – – – –	T – T T T T	– – – – – – – – – – – –	Syn.	Poly.
983	T	– – – – – – – – – – – –	– – – – – –	C C C C C C C C C C C C	Syn.	Fixed
1019	C	– – – – – – – – – – – –	– – – – – –	– – – – A – – – – – – –	Syn.	Poly.
1031	C	– – – – – – – – – – – –	– – – – – –	– – – – – – – – A – – –	Syn.	Poly.
1034	T	– – – – – – – – – – – –	– – – – – –	– C C C C C – – C – C C	Syn.	Poly.
1043	C	– – – – – – – – – – – –	– – – – – –	– – – – A – – – – – – –	Syn.	Poly.
1068	C	T T – – – – – – – – – –	– – – – – –	– – – – – – – – – – – –	Syn.	Poly.
1089	C	– – – – – – – – – – – –	A A A A A A	– – – – – – – – – – – –	Repl.	Fixed
1101	G	– – – – – – – – – – – –	– – – – – –	A A A A A A A A A A A A	Repl.	Fixed
1127	T	– – – – – – – – – – – –	– – – – – –	C C C C C C C C C C C C	Syn.	Fixed
1131	C	– – – – – – – – – – – –	– – – – – –	– – – – T – – – – – – –	Syn.	Poly.
1160	T	– – – – – – – – – – – –	– – – – – –	C C C C C C C C C C C C	Syn.	Fixed
1175	T	– – – – – – – – – – – –	– – – – – –	C C C C C C C C C C C C	Syn.	Fixed
1178	C	– – – – – – – – – – – –	– – – – – –	– – – – – – – – – – A –	Syn.	Poly.
1184	C	– – – – – – – – – – – –	– – – – – –	G G G G G G G G G G G G	Syn.	Fixed
1190	C	– – – – – – – – – – – –	– – – – – –	– – A – – – – – – – – –	Syn.	Poly.
1196	G	– – – – – – – – – – – –	– – – – – –	T T T T – T T T – T – –	Syn.	Poly.
1199	C	– T – – – – – – – – – –	– – – – – –	– – – – – – – – – – – –	Syn.	Poly.
1202	T	– – – – – – – – – – – –	– – – – – –	C C C C C C C C C C C C	Syn.	Fixed
1203	C	– – – – – – – – – – – –	– – T – – –	– – – – – – – – – – – –	Syn.	Poly.
1229	T	– – C C C C C C C C C C	– – – – – –	– – – – – – – – – – – –	Syn.	Poly.
1232	T	– – – – – – – – – – – –	– – – – – –	A A A A A A A A A A A A	Syn.	Fixed
1235	C	– – – – – – – – – – A –	– – – – – –	– – – – – – – – – – – –	Syn.	Poly.
1244	C	– – – – – – – – – – – –	– – – – – –	– – A – – – – – – – – –	Syn.	Poly.
1265	C	– – – – – – – – – – – –	– – – – – –	G G G G G G G G G G G G	Syn.	Fixed
1271	A	– – – – – – – – – – – –	– T – T – –	– – – – – – – – – – – –	Syn.	Poly.
1277	T	– – – – – – – – – – – –	– – – – – –	C C C C C C C C C C C C	Syn.	Fixed
1283	C	A A – – – – – – – – – –	– – – – – –	– – – – – – – – – – – –	Syn.	Poly.
1298	C	– – – – – – – – – – – –	– – – – – –	T T T T T T T T T T T T	Syn.	Fixed
1304	C	– – – – – – – – – – – –	– – – – T –	– – – – – – – – – – – –	Syn.	Poly.
1316	C	– – – – – – – – – – – –	– – T T – –	T T T T T T T T T T T T	Syn.	Poly.
1425	C	A A – – – – – – – – – –	– – – – – –	– – – – – – – – – – – –	Syn.	Poly.
1431	T	C C – – – – – – – – – –	– – – – – –	C C C C C C C C C C C C	Syn.	Poly.
1443	C	– – – – – – G G G G G G	– – – – – –	– – – – – – – – – – – –	Syn.	Poly.
1452	C	– – – – – T T T T T T T	– – – – – –	– – – – – – – – – – – –	Syn.	Poly.
1490	A	– – – – – – C C C C C C	– – – – – –	– – – – – – – – – – – –	Repl.	Poly.
1504	C	T T T T T T T T T T T T	– – – – – –	– – – – – – – – – – – –	Syn.	Fixed
1518	C	– – – – – T T T T T T T	– – – – – –	– – – – – – – – – – – –	Syn.	Poly.
1524	T	– – – – – – – – – – – –	– – – – – –	G G G G G G G G G G G G	Syn.	Fixed
1527	C	T T T T T T – – – – – –	– – – – – –	– – – – T – – – – – – –	Syn.	Poly.
1530	G	– – – – – – – – – – – –	– – – – – –	– – – A – – – – – – – –	Syn.	Poly.
1545	T	– – – – – – – – – – – –	– – – – – –	C C C C C C C C C C C C	Syn.	Fixed
1548	C	– – – – – – – – – – – –	– – – – – –	– – – – A – – – – – – –	Syn.	Poly.
1551	C	– – – – – – – – – – – –	– – – T – –	– – – – – – – – – – – –	Syn.	Poly.
1555	C	– – – – – – – – – – – T	– – – – – –	– – – – – – – – – – – –	Repl.	Poly.
1557	C	A A A A A – – – – – – –	– – – – – –	– – – – – – – – – – – –	Syn.	Poly.
1560	G	– – – – – – – – – – – –	– – – A – –	– – – – – – – – – – – –	Syn.	Poly.
1573	G	– – – – – – – – – – – –	– – – – – –	C C C C C C C C C C C C	Repl.	Fixed
1581	C	– – – – – – – – – – – –	– – – – – –	T T T T T T T T T T T T	Syn.	Fixed
1584	C	– – – – – – – – – – – –	– – – – – –	G G G G G G G G G G G –	Syn.	Poly.
1590	C	T T T T T T T T T T T T	T T T – – –	– – – – – – – – – – – –	Syn.	Poly.
1596	G	– – A A – A – – – – – A	– – – – – –	– – – – – – – – – – – –	Syn.	Poly.
1611	A	– – – – – – – – – – – –	– – – – – –	T T T T T T T T T T T T	Syn.	Fixed
1614	C	– – – – – – – – – – – –	– G – – – –	– – – – – – – T – – – –	Syn.	2 Poly.
1635	C	– – – – – – – – – – – –	– – – – – –	– – T – T – – – – – – –	Syn.	Poly.
1657	A	– – – – – – – – – – – –	– – – – – –	T T T T T T T T T T T T	Repl.	Fixed
1665	C	– – – – – – – – – – – –	– – – – – –	– – – – – – – – T – – –	Syn.	Poly.

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. 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 para cambiar (Fig. 1).

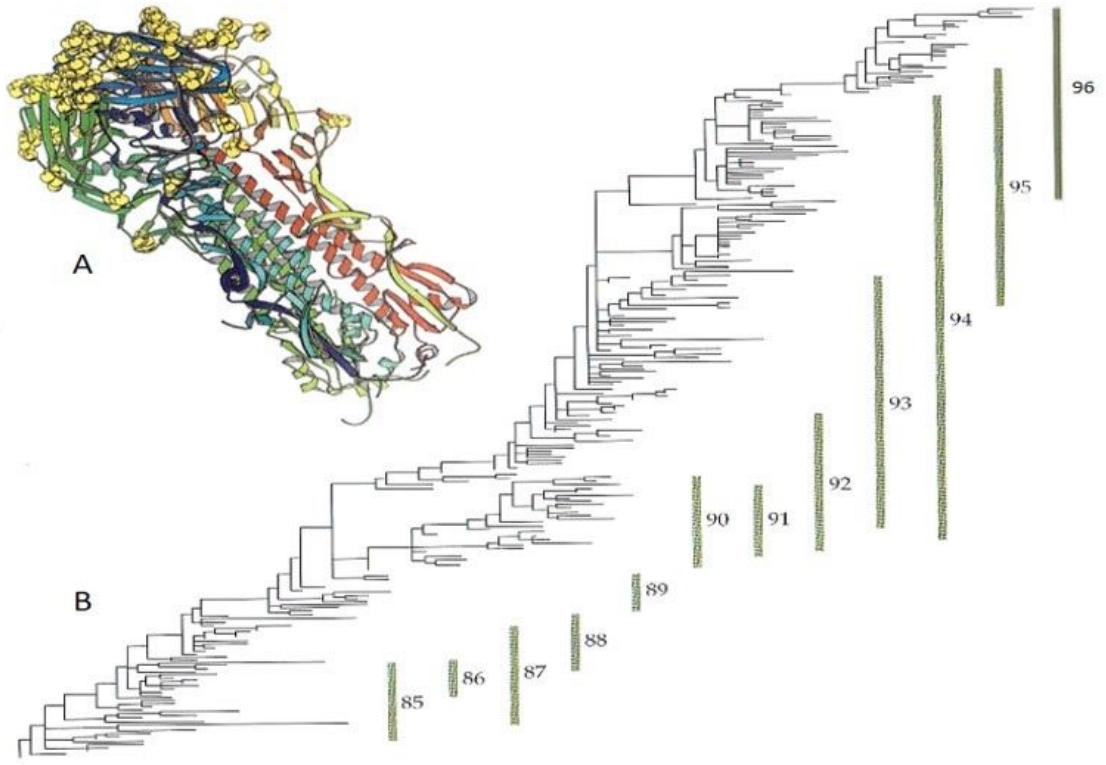

Fig 1. A) Modelo tridimensional de la Hemaglutinina del virus de la gripe, mostrando los sitios aminoacídicos seleccionados positivamente para cambiar. B) Filogenia de cepas del virus aisladas desde 1985 a 1996, basada en el análisis de las secuencias nucleotídicas de ese gen (Tomado de Hillis, 2009).

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 de los nodos ancestrales. Utilizaremos una aproximación asociada al MEGA llamada SLAC (Single-Likelihood Ancestor Counting) implementada en el servidor Datamonkey (https://www.datamonkey.org/) donde se pueden implementar otros algoritmos.

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. En este caso, cargaremos al programa la información de las secuencias, y el árbol será estimado por el propio programa. 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. Así, compara 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). La significancia estadística se determina en cada sitio utilizando una distribución binomial extendida. Es importante destacar que, debido a su enfoque basado en el recuento, es posible que SLAC no sea preciso para conjuntos de datos con altos niveles de divergencia. 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. Existen otros métodos más flexibles y sensibles que pueden ser explorados en el servidor Datamonkey.

Actividad

- Abrir el programa MEGAX, y dentro del programa abrir el archivo SecuenciasHA.meg.

- Ir a la opción "SELECTION" en el menú principal, donde se pueden implementar algunos tests para identificar sitios bajo selección.

- Marcar la opción DATAMONKEY (donde se aplicará la opción "codon by codon selection by SLAC"). Una vez en esta opción no hay que especificar el archivo (porque ya fue cargado) pero sí algunas opciones, como el código genético (Universal) y los sitios a usar (all sites).

El resultado demorará unos minutos. En caso que haya dificultades puede ver los resultados tal y como se verían en MEGAX, en el siguiente enlace: http://datamonkey.org/slac/5f3d5c21503f2f234815f7ff

- Interprete los resultados. ¿Existe algún sitio y/o linaje seleccionado? ¿Qué sitio presenta 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?

- ¿Qué similitudes y diferencias encuentra entre esta aproximación y el test de MK?

- Explore otros programas implementados en Datamonkey.org. En este enlace verá el análisis de el mismo conjunto de secuencias pero habiendo aplicado otro método para inferir selección a nivel molecular denominado MEME (Mixed Effects Model of Evolution), que permite estimar sitos y ramas bajo selección simultáneamente (https://www.datamonkey.org/meme/5f3d4598503f2f234815f3d1).

Referencias

McDonald, J., Kreitman, M. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351, 652–654 (1991). https://doi.org/10.1038/351652a0

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.

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).

Murrell, B et al. "Detecting individual sites subject to episodic diversifying selection." PLoS Genetics 8, e1002764 (2012).