Estudio de la variabilidad estructural proteica

Resumen

Introducción

Las proteínas divergen durante el proceso de evolución biológica, lo cual se hace evidente en la variación de la secuencia de aminoácidos y los consiguientes cambios estructurales, dinámicos y funcionales.

Se sabe que la estructura diverge mucho más lentamente que la secuencia y que la divergencia evolutiva estructural ocurre principalmente a lo largo de los modos vibracionales de menor energía de las proteínas. Este hecho es difícil de interpretar debido a que todos los estudios realizados son puramente empíricos. Para avanzar en este sentido, se desarrolló el modelo "Linearly Forced Elastic Network Model" (LFENM), el cual predice el cambio en la posición de equilibrio de los sitios como consecuencia de mutaciones aleatorias, no sujetas a selección natural[3]. Aplicando este modelo se demostró que los patrones de cambio estructural (mayor contribución de los modos normales de menor energía y existencia de un "core" estructuralmente conservado) se pueden reproducir muy bien sin recurrir a la selección natural [3,4]. Todos estos resultados ponen en cuestión interpretaciones basadas en la suposición de que todo lo que se conserva o varía está relacionado con la conservación o variación de la función biológica.

Si bien la selección natural aparentemente afecta poco a los patrones de divergencia estructural y dinámica, a nivel de secuencia diferentes sitios evolucionan a diferentes velocidades. Modelos de evolución puramente mutacionales, como el LFENM, no pueden dar cuenta de este hecho. Para explicar tales patrones de variación secuencial se debe modelar la selección natural. Recientemente, hemos propuesto un modelo mecanístico ("Stress Model") en que una mutación se acepta con una probabilidad proporcional a la probabilidad de que el mutante adopte la “estructura activa”. Este modelo ha servido para dar cuenta de la variación de la velocidad promedio de evolución de un sitio a otro [6].

Como dijimos, el modelo LFENM fue usado exitosamente para explicar los patrones observados de divergencia estructural en ausencia de selección natural. Sin embargo, es de esperar que la selección restrinja la divergencia estructural y/o dinámica. Así por ejemplo si la estructura del sitio activo de una enzima es importante para la actividad enzimática, se esperaría que la su estructura se conserve evolutivamente significativamente más que lo esperado por un modelo puramente mutacional. Por lo tanto, por débil que sea, se esperaría alguna evidencia de selección natural a nivel de la divergencia estructural y funcional. Encontrar esta evidencia, si existe, es el propósito orientador de este informe.

Materiales y Métodos

Familias de proteínas:

Se seleccionaron 10 familias de proteínas de la base de alineamientos estructurales múltiples de homólogos HOMSTRAD (http://mizuguchilab.org/homstrad/). Las familias seleccionadas se muestran en la tabla 1. Se eligieron a estas familias ya que las mismas poseen alineamientos múltiples con más de 15 proteínas y con una longitud de alineamiento mayor a 50 sitios.

Familia	Número de proteínas	Proteína de referencia	Clase	% Identidad secuencial
lipocalin	15	1bj7	all beta	21
fabp	17	1hmt	all beta	45
globins	38	1a6m	all alpha	33
kinase	15	1phk	alpha + beta	29
phospholip	18	1jiaa	all alpha	54
plastocyanins	29	1bxv	all beta	36
rrm	20	1fxla2	alpha + beta	25
serinProteases	27	1mct	all beta	46
sh3	20	1lcka	small	35
snakeToxins	20	1ntx	small	46

Elección de las proteínas de referencia: Para cada familia de proteínas se seleccionó a una proteína de referencia. Para esto, en primer lugar, se calculó la estructura promedio de cada uno de los alineamientos. Luego, se calculó la desviación cuadrática promedio entre la estructura de cada proteína del alineamiento y la estructura promedio obtenida. Por último, se seleccionó a la proteína con menor desviación cuadrática promedio como proteína de referencia.

Análisis de alineamientos:

Para analizar a los alineamientos múltiples de las distintas familias de proteínas se utilizó dos estrategias, considerar solo el core conservado del alineamiento (sitios sin ningún gap en todo el alineamiento) y considerar todo el alineamiento.

Para ambas estrategias, se alineó la proteína de referencia con cada una de las proteínas del conjunto y se obtuvieron los sitios alineados y los no alineados. Luego, en el caso del análisis del core, de los sitios alineados, seleccionamos solo los que se encuentran dentro del core conservado.

Modelos de red elástica:

Utilizamos el Elastic Network Model (ENM) para representa a cada proteína. Este modelo considera a la proteína plegada como una red de sitios (por lo general los carbonos alfa

de los aminoácidos) conectados por resortes. El potencial de la red elástica de cada proteína puede aproximarse de la siguiente forma:

$$V(r) = \frac{1}{2}(r - r^0)^T H(r - r^0)$$

donde H es el Hesiano: la matriz de derivadas segundas del potencial, *r* es el vector clumna de la posición de los sitios y r^0 el vector columna posición de equilibrio de los sitios. Diagonalizando H:

$$Hu_n = \lambda_n u_n$$

se obtienen los modos normales u_n, que son combinaciones de coordenadas cartesianas que representan las vibraciones independientes de la molécula. Los autovalores correspondientes λ_n representan las energías de deformación (correspondientes a un desplazamiento unitario en la dirección del modo normal). En general, los modos normales de menor energía combinan las coordenadas cartesianas de muchos átomos: movimientos globales lentos y coherentes.

Conjunto de proteínas simuladas:

Para simular proteínas de cada familia se utilizó el modelo mutacional LF - ENM (Linearly Forced – Elastic Network Model). Este modelo permite generar proteínas mutantes modelando a una mutación puntual aplicando fuerzas a lo largo de los contactos del sitio mutado. En este caso generamos mutantes de múltiples sitios considerando a las mutaciones puntuales aditivas entre sí.

Para generar las mutantes, en un caso consideramos a la selección natural a nivel de la secuencia y en otro caso no la consideramos:

- **Mutantes con selección natural (ns = T):** para determinar qué sitios mutar, se alineó la proteína de referencia con cada una de las demás proteínas del conjunto, se obtuvieron los índices de los sitios alineados pero mutados y de los sitios con gaps y se simularon 10 mutantes por cada par de proteínas. Luego, para el análisis, se tomaron como alineados los sitios alineados de la proteína de referencia para cada par de proteínas.
- **Mutantes sin selección natural (ns = F):** se mutaron al azar la cantidad de sitios correspondientes al % de identidad secuencial del conjunto experimental y luego, para el análisis, se tomaron como alineados todos los sitios de la proteína de referencia. Se generaron 10 * (número de proteínas experimentales) mutantes.

Análisis de variabilidad estructural

Para ambos conjuntos se obtuvieron las coordenadas de los sitios alineados y no alineados de cada proteína. Luego, se calcularon medidas de variabilidad estructural en coordenadas cartesianas o en modos normales.

Coordenadas cartesianas: MSDi

Para los conjuntos teóricos y experimentales se calculó la variación estructural de cada proteína con respecto a la proteína de referencia en los sitios alineados y luego se calculó, para cada sitio, la desviación cuadrática:

$$||\Delta\bar{\mathbf{r}}_i||^2 = \Delta\bar{x}_i^2 + \Delta\bar{y}_i^2 + \Delta\bar{z}_i^2$$

Siendo $\Delta\bar{\mathbf{r}}_i = (\Delta\bar{x}_i\ \Delta\bar{y}_i\ \Delta\bar{z}_i)^T$ el vector columna de desplazamiento cartesiano del C_α *i* con respecto a la proteína de referencia.

Modos normales: Pn

El análisis del cambio estructural de modos normales se calculó proyectando las diferencias estructurales de los sitios alineados sobre los modos normales de la proteína de referencia. Los modos normales fueron obtenidos resolviendo la ecuación:

$$\mathbf{K}\mathbf{q}_n = \lambda_n \mathbf{q}_n$$

En el caso de proteínas que no alinean en todos los sitios de la referencia, en lugar de K, se usó Keff, la cual permite obtener los modos normales que describen los movimientos de los sitios alineados.

Luego, para una proteína con variación estructural de los sitios alineados $\Delta\bar{\mathbf{r}}$, se calculó la proyección sobre los modos normales de la siguiente forma:

$$P_n \equiv \frac{(\mathbf{q}_n^T \Delta\bar{\mathbf{r}})^2}{\sum_n (\mathbf{q}_n^T \Delta\bar{\mathbf{r}})^2}$$

Comparaciones de perfiles:

Para realizar comparaciones entre las medidas de variabilidad teóricas y experimentales se prosiguió de la siguiente forma:

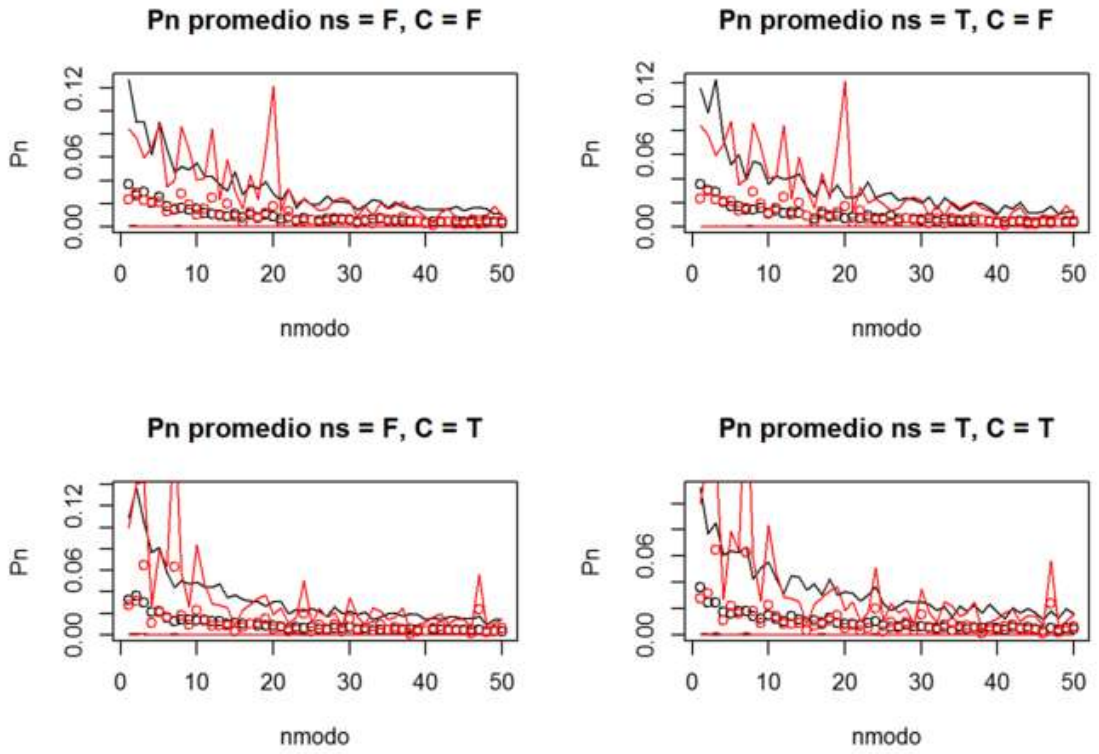
- Medidas teóricas: Los (10 * número de proteínas experimentales) perfiles de Pn y MSDi teóricos se reagruparon en 10 sub - conjuntos de modo que en todos haya una mutante que corresponda a un proteína teórica. Luego se calculó, para estos 10 conjuntos, el promedio y los cuantiles 0.05 y 0.95. Luego, se calculó el promedio de promedios y el promedio de cuantiles.

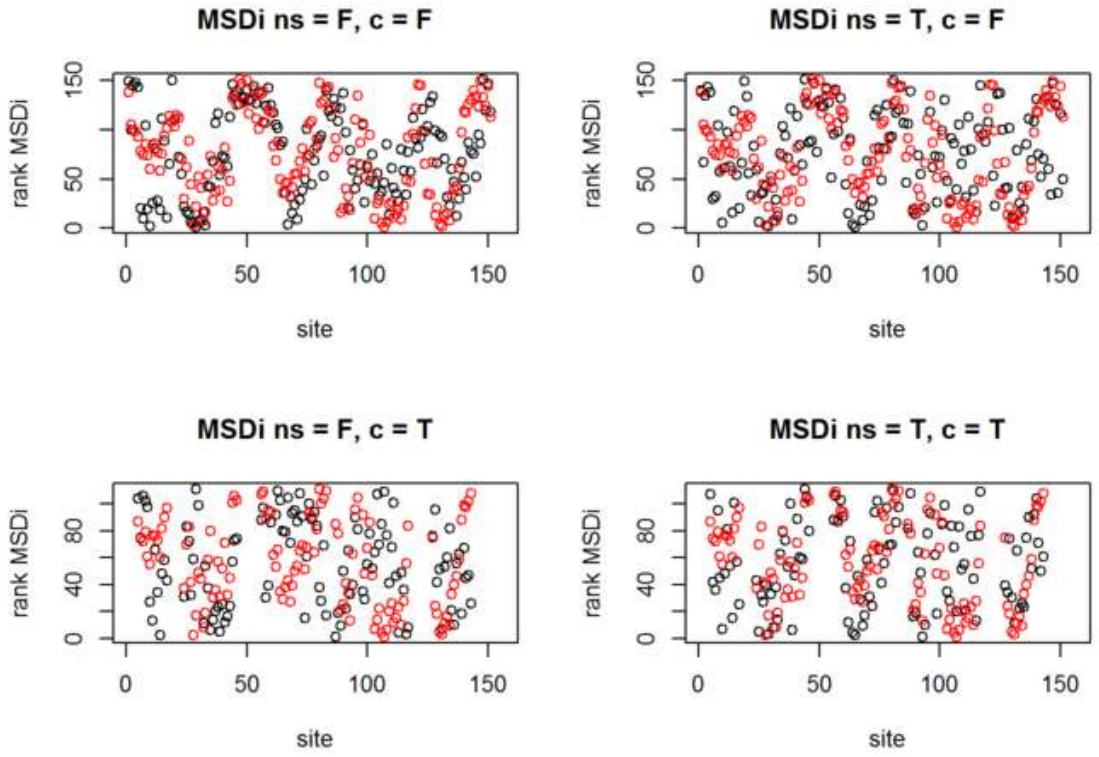
- Medidas experimentales: se calculó el promedio y los cuantiles 0.05 y 0.95 de los perfiles de Pn y MSDi.

Posteriormente, se calculó la correlación de Pearson y el MSE entre perfiles Pn y MSDi promedio teóricos y experimentales promedio.

Resultados:

Si bien para todas las familias de proteínas las similitudes entre Pn y MSDi teóricas y experimentales es alta con todos los conjuntos de mutantes teóricas se observa que, para la mayor parte de los casos, la similitud aumenta al analizar el core y al considerar a la selección natural. La contribución de la selección natural es más evidente en los perfiles de MSDi que en los de Pn.





Familia	Correlación Pn ns = F c = F	Correlación Pn ns = T c = F	Correlación Pn ns = F c = T	Correlación Pn ns = T c = T
lipocalin	0.75	0.74	0.73	0.66
fabp	0.75	0.79	0.75	0.76
globins	0.67	0.66	0.42	0.47
kinase	0.78	0.8	0.88	0.88
phospholip	0.54	0.64	0.56	0.62
plastocyanins	0.4	0.52	0.76	0.74
rrm	0.91	0.94	0.78	0.84
serinProteases	0.62	0.61	0.54	0.64
sh3	0.65	0.53	0.78	0.89
snakeToxins	0.84	0.82	0.8	0.79

Familia	Correlación MSDi ns = F c = F	Correlación MSDi ns = T c = F	Correlación MSDi ns = F c = T	Correlación MSDi ns = T c = T
lipocalin	0	0.24	0	0.22
fabp	0.5	0.49	0	0.59
globins	0.19	0.23	0	0.22
kinase	0.01	0.21	0	0.37
phospholip	0.45	0.49	0.07	0.48

plastocyanins	0.13	0.2	0.05	0.19
rrm	0.34	0.7	0.01	0.45
serinProteases	0.34	0.44	0	0.23
sh3	0.5	0.53	0.03	0.5
snakeToxins	0.2	0.51	0.01	0.64