Informe 2015 - 2016

Estudio de la variabilidad estructural proteica

Resumen

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, que la divergencia evolutiva estructural ocurre principalmente a lo largo de los modos vibracionales de menor energía y que existe un core estructuralmente conservado. Aplicando un modelo puramente mutacional como es el Linearly Forced - Elastic Network Model (LF-ENM) se demostró que los patrones de cambio estructural se pueden reproducir muy bien sin recurrir a la selección natural. Sin embargo, es de esperar que la selección restrinja, aunque sea de forma muy leve, la divergencia estructural. Para estudiar esto, en primer lugar obtuvimos los alineamientos estructurales de 10 familias de proteínas. Luego, usando el LF-ENM generamos conjuntos de mutantes de una de las proteínas de cada familia. En uno de los conjuntos no consideramos a la selección natural (mutamos de forma aleatoria tantos sitios como corresponda al % de identidad secuencial del conjunto experimental) y en el otro conjunto consideramos a la selección natural (mutamos sitios mutados y sitios con gaps en la proteína de referencia. Luego, para el análisis, consideramos en un caso todo el alineamiento múltiple y, en otro caso, consideramos solo el core conservado (posiciones del alineamiento múltiples sin ningún gap). Por último, calculamos, para los conjuntos teóricos y experimentales de cada familia, perfiles de variabilidad estructural en coordenadas cartesianas (*MSDi*) y proyectadas en modos normales (*Pn*) y los comparamos utilizando el coeficiente de correlación de Pearson. Obtuvimos que, a nivel de coordenadas cartesianas, las mutantes simuladas considerando selección natural poseen perfiles más parecidos a los perfiles de los conjuntos experimentales, ya sea considerando o no el core. Sin embargo, a nivel de modos normales, no hay una clara evidencia de selección natural. Estos resultados podrían indicar que hay señal de selección natural a nivel estructural. Sin embargo, esto debe ser estudiado en mayor profundidad.

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, que la divergencia evolutiva estructural ocurre principalmente a lo largo de los modos vibracionales de menor energía y que existe un core estructuralmente conservado. 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" (LF-ENM), el cual predice el cambio en la posición de equilibrio de los sitios como consecuencia de mutaciones aleatorias, no sujetas a selección natural. 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.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, a nivel de secuencia diferentes sitios evolucionan a diferentes velocidades. Modelos de evolución puramente mutacionales, como el LF-ENM, 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.

Como dijimos, el modelo LF-ENM 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. Así por ejemplo si la estructura del sitio activo de una enzima es importante para la actividad enzimática, se esperaría que la 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. Encontrar esta evidencia, si existe, es el propósito orientador de esta parte de la investigación.

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.

Tabla 1: caracterización familias de proteínas.

Familia	Número de proteínas (N)	Proteína de referencia	Clase	% Identidad secuencial
Lipocalinas	15	1bj7	beta	21
Proteínas de unión a ácidos grasos	17	1hmt	beta	45
Globinas	38	1a6m	alfa	33
Serina-Treonina kinasas	15	1phk	alfa + beta	29
Fosfolipasas A2	18	1jia	alfa	54
Plastocyaninas	29	1bxv	beta	36
Proteínas de reconocimiento de RNA	20	1fxl	alfa + beta	25

Serin proteinasas	27	1mct	beta	46
Dominios homologos Src 3	20	1lck	chica	35
Toxinas de vivoras	20	1ntx	chica	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, **c = T**) y considerar todo el alineamiento **(c = F)**.

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:

$$H q_n = \lambda_n q_n$$

se obtienen los modos normales q_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.

Existen diferentes ENM que se diferencian en las constantes de fuerza de la red elástica. El modelo ENM específico que usamos en este caso es el Anisotropic Network Model (ANM), que usa la misma constante de fuerza, 1, para todos los pares de sitios cuya distancia es menor a un cierto “cut-off”, en este caso 10 Å, y 0 para el resto de los sitios.

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 * N mutantes.

Medidas de variabilidad estructural

Para ambos conjuntos, teóricos y experimentales, 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 proyectadas sobre los modos normales:

- **Coordenadas cartesianas:**

Para los conjuntos teóricos y experimentales de cada familia 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. De esta forma obtuvimos, para cada familia de proteínas, 4 conjuntos teóricos (c = F y ns = F, c = F y ns = T , c = T y ns = F , c = T y ns = T) con 10 x N perfiles y un conjunto experimental con N perfiles.

- **Modos normales:**

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:

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

De esta forma obtuvimos, para cada familia de proteínas, 4 conjuntos teóricos (c = F y ns = F, c = F y ns = T , c = T y ns = F , c = T y ns = T) con 10 x N perfiles y un conjunto experimental con N perfiles.

Comparaciones de perfiles:

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

- **Medidas teóricas:** Los (10 x N) perfiles de Pn y MSDi teóricos de cada familia se reagruparon en 10 sub - conjuntos de modo que en todos haya una mutante que corresponda a una proteína experimental. Luego se calculó, para estos 10 conjuntos, el promedio y los cuantiles 0.05 y 0.95. Por último, 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 de cada familia.

Posteriormente, se fiteó los perfiles promedio teóricos con los perfiles experimentales y se calculó la correlación y R^2 entre estos perfiles y los perfiles experimentales promedio de cada familia.

Resultados

Los perfiles de Pn promedio vs número de modo y de MSDi vs número de sitio de la familia de las globinas se muestran la figura 1 y figura 2 respectivamente.

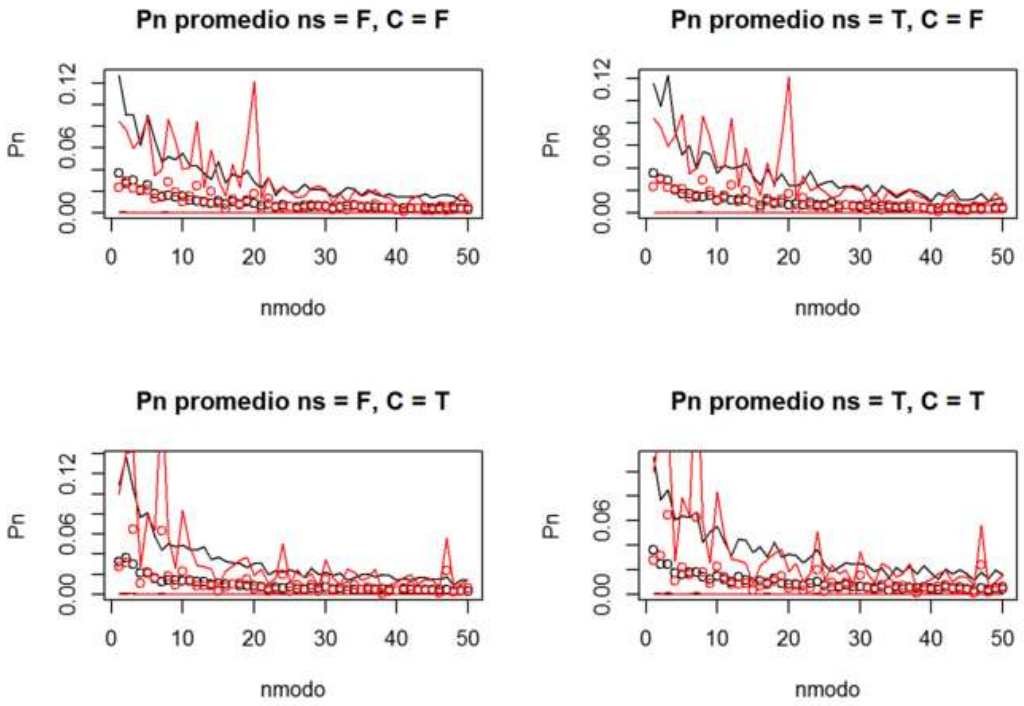

Figura 1: Pn vs. número de modo de la familia de las Globinas. a) Mutantes sin selección natural y análisis sin considerar el core. b) Mutantes con selección natural y análisis sin considerar el core. c) Mutantes sin selección natural y análisis considerando el core. d) Mutantes con selección natural y análisis considerando el core. Los puntos negros corresponden a promedio de promedios de 10 grupos de mutantes teóricas. Las líneas negras corresponden a promedios de cuantiles 0.05 y 0.95 de los 10 grupos de mutantes teóricas. Los puntos rojos corresponden a promedio de proteínas experimentales. Las líneas rojas corresponden a cuantiles 0.05 y 0.95 de proteínas experimentales.

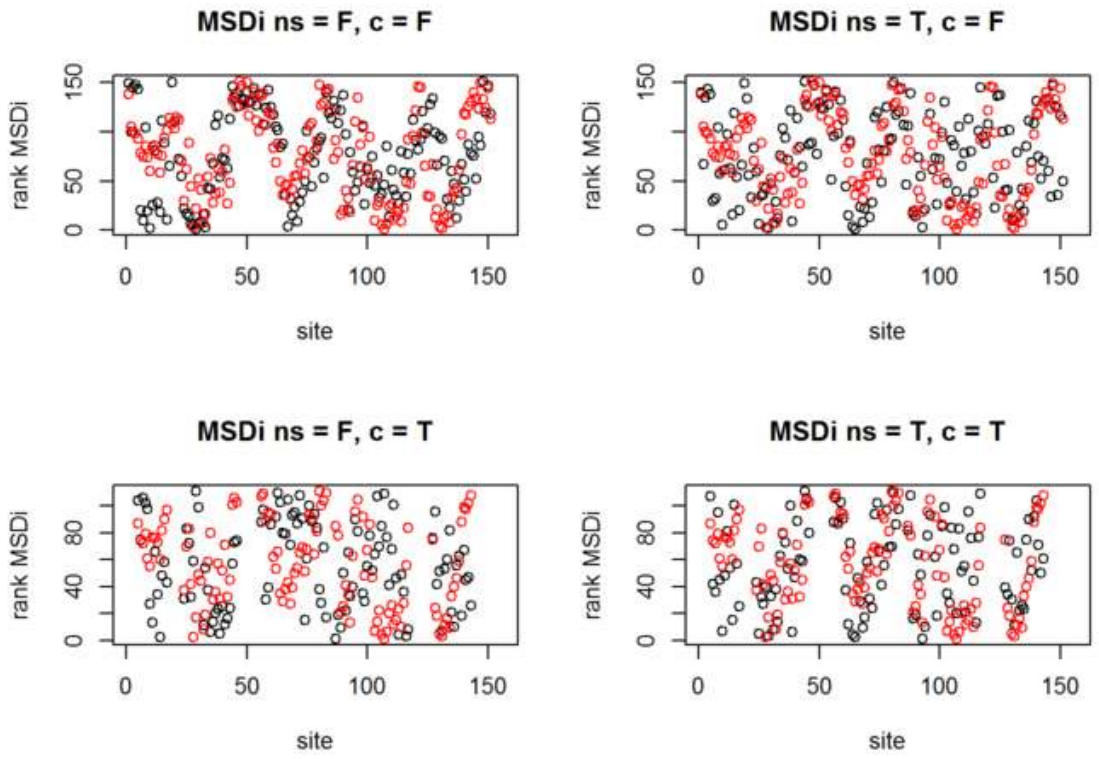

Figura 2: MSDi vs. número de sitio de la familia de las Globinas. a) Mutantes sin selección natural y análisis sin considerar el core. b) Mutantes con selección natural y análisis sin considerar el core. c) Mutantes sin selección natural y análisis considerando el core. d) Mutantes con selección natural y análisis considerando el core. Los puntos negros corresponden a promedio de promedios de 10 grupos de mutantes teóricas. Los puntos rojos corresponden a promedio de proteínas experimentales. Los perfiles promedio fueron ranqueados con la función de R rank().

En estas figuras se observa que, al menos cualitativamente, los perfiles de Pn promedio y de MSDi de todos los conjuntos teóricos son similares al perfil correspondiente del conjunto experimental. Solo en el caso de MSDi ns = F y c = T se observa que los perfiles no son similares. Resultados similares se observaron para el resto de las familias (no se muestran).

Para evaluar cuantitativamente la similitud entre perfiles promedio, se calculó el coeficiente de correlación y R^2 para Pn (Tabla 2) y para MSDi (Tabla 3).

Tabla 2: R^2 entre perfiles Pn promedio fiteados experimentales y teóricos.

Familia	R^2 Pn ns = F c = F	R^2 Pn ns = T c = F	R^2 Pn ns = F c = T	R^2 Pn ns = T c = T
Lipocalinas	0.75	0.74	0.73	0.66
Proteínas de unión a ácidos grasos	0.75	0.79	0.75	0.76
Globinas	0.67	0.66	0.42	0.47
Serina-Treonina kinasas	0.78	0.8	0.88	0.88
Fosfolipasas A2	0.54	0.64	0.56	0.62
Plastocyaninas	0.4	0.52	0.76	0.74
Proteínas de reconocimiento de RNA	0.91	0.94	0.78	0.84
Serin proteinasas	0.62	0.61	0.54	0.64
Dominios homologos Src 3	0.65	0.53	0.78	0.89
Toxinas de vivoras	0.84	0.82	0.8	0.79

Tabla 3: R^2 entre perfiles promedio fiteados de MSDi experimentales y teóricos.

Familia	R^2 MSDi ns = F c = F	R^2 MSDi ns = T c = F	R^2 MSDi ns = F c = T	R^2 MSDi ns = T c = T
Lipocalinas	0	0.24	0	0.22
Proteínas de unión a ácidos grasos	0.5	0.49	0	0.59
Globinas	0.19	0.23	0	0.22
Serina-Treonina kinasas	0.01	0.21	0	0.37
Fosfolipasas A2	0.45	0.49	0.07	0.48
Plastocyaninas	0.13	0.2	0.05	0.19
Proteínas de reconocimiento de	0.34	0.7	0.01	0.45

RNA				
Serin proteinasas	0.34	0.44	0	0.23
Dominios homologos Src 3	0.5	0.53	0.03	0.5
Toxinas de vivoras	0.2	0.51	0.01	0.64

En la Tabla 2 se observa que, en general, las correlaciones entre perfiles teóricos y experimentales de Pn son buenas para todas las familias de proteínas, tanto considerando como no considerando el core del alinemaineto. No parecería haber un efecto claro de selección natural a nivel de proyección del cambio estructural sobre modos normales.

En la Tabla 3 se observa que, para la mayor parte de los casos, la similitud entre perfiles teóricos y experimentales, considerando o no el core del alinemiento, aumenta al considerar a la selección natural. Esto podría ser evidencia de contribución de selección natural a nivel estructural.

Conclusiones

Como previamente se ha demostrado, no parece haber rastros de selección natural en los cambios estructurales proteicos proyectados sobre los modos normales. De todas formas, parecía haber una cierta evidencia de selección natural a nivel de cambios estructurales proteicos en coordenadas cartesianas. Sin embargo, para poder afirmar esto, la variabilidad estructural debe seguir siendo estudiada.

Presentaciones a congresos 2015

Congreso Argentino de Fisicoquímica y Química Inorgánica de la Asociación Argentina de Investigación Fisicoquímica (AAIFQ) de 2015.

Cursos realizados 2015

"Transporte óptimo y análisis de datos" Prof. Esteban Tabak (Courant Institute New York) Instituto de Cálculo, FCEyN-UBA, Buenos Aires.