UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Estabilidad y estructura: inferencia filogenética de un modelo evolutivo de proteínas basado en estabilidad termodinámica y desviación estructural simulada con modos normales torsionales

Máster Universitario en Bioinformática y Biología Computacional

Autor: LORCA ALONSO, Iván

Tutor: BASTOLLA BUFFALINI, Ugo
Departamento de Bioinformática, CBMSO
Ponente: MARTÍNEZ MUÑOZ, Gonzalo
Departamento de Ingeniería Informática, UAM

Junio, 2021

Resumen

Los modelos de sustitución de aminoácidos basados en estabilidad de plegamiento permiten inferir la evolución molecular de las proteínas teniendo en cuenta algunos vínculos impuestos por la biofísica, pero no pueden representar el efecto de las mutaciones sobre la estructura nativa. Para superar esta limitación el grupo con el cual he realizado el TFM ha propuesto predecir estos efectos a través del análisis de modos normales en el espacio de ángulos de torsión. El resultado principal de este trabajo muestra que la inclusión de este factor estructural en los modelos de sustitución mejora las predicciones de la entropía y la tasa de sustitución en cada posición de un alineamiento múltiple, y permite explicar la relación observada entre conservación de secuencia en cada posición y magnitudes estructurales como el número de contactos nativos . Dada la naturaleza biofísica del modelo, esta mejora apunta a posibles principios que rigen la evolución de proteínas y ofrece posibles mejoras en la inferencia filogenética respecto a los modelos empíricos que no tienen en cuenta la estructura.

Palabras clave: evolución molecular, estabilidad termodinámica, estructura nativa, análisis de modos normales.

Agradecimientos

Quiero agradecer a mi tutor, Ugo, por brindarme la oportunidad de trabajar en este proyecto, un trabajo interdisciplinar que me ha permitido adquirir conocimientos más allá de mi campo de estudio. También a mi familia por su apoyo durante mis estudios, especialmente durante un año singularmente incierto.

Índice de contenidos

Índice de Figuras

1

Introducción

1.1. Evolución molecular y modelos evolutivos

Las proteínas son la fuerza de trabajo de los organismos vivos, toman parte en procesos esenciales para la vida donde la estructura nativa de la proteína es esencial para su función. La evolución de esta estructura y la secuencia correspondiente están por tanto restringidas a formas que sean capaces de cumplir la función biológica dentro del organismo.

Este proceso de selección entre las conformaciones nativas posibles se refleja en los alineamientos múltiples de familias de proteínas. Siempre que su función no cambie, las proteínas suelen conservar su estructura nativa mientras la secuencia diverge. En base a este tipo de alineamientos se han obtenido diferentes modelos de sustitución empíricos consistentes en matrices que almacenan las tasas de sustitución entre aminoácidos. Estos modelos son la base de la inferencia filogenética para determinar las relaciones evolutivas entre secuencias y organismos. Sin embargo, estos modelos de sustitución empíricos reflejan sólo en parte la complejidad de los procesos selectivos en cuanto asumen que cada posición evoluciona de forma independiente del resto de la secuencia (una hipótesis necesaria para que los modelos matemáticos resultantes se puedan tratar computacionalmente) y que todas las posiciones evolucionan según el mismo proceso de sustitución, lo cual es desmentido por la evidencia observacional.

Estos modelos no son capaces de mostrar explícitamente porqué una posición evoluciona más rápida o lentamente ya que funcionan simplemente en base a la variabilidad capturada en los alineamientos múltiples y asunciones razonables que permiten la formulación de las tasas de sustitución. En ellos no se puede establecer si la evolución de una posición se ve modulada por su efecto en la estabilidad, estructura o función. Dicho de otra manera, los modelos de sustitución empíricos dejan de lado la interdependencia entre posiciones y no capturan información relativa a la estructura y la dinámica de la proteína a la que codifican, lo que puede mellar su precisión a la hora de realizar inferencia filogenética y hacerlos poco realistas para simular la evolución de una proteína.

1.2. Modelos biofísicos: la Física en la Evolución

En años recientes, se han presentado modelos que abordan la evolución de proteínas desde una perspectiva biofísica, utilizando herramientas matemáticas y conceptos de la física dentro del marco de la evolución. Estos modelos tienen el valor añadido de estar basados en principios y derivaciones físicas que buscan explicar la variabilidad observada en las secuencias. La formulación explicita de los fenómenos que resultan en las tasas de variación puede ayudar a determinar que parámetros son más influyentes (Bastolla et al., 2017).

Puede haber diferentes aproximaciones para desarrollar un modelo, entre estas, debida la importancia de la estructura nativa para el desarrollo de la función, se han desarrollado múltiples modelos que consideran la estabilidad de esta conformación, dada por la diferencia de energía libre (ΔG) entre la conformación plegada y sin plegar (revisados por ejemplo por Echave & Wilke (2017)).

Dado que romper la asunción de independencia entre posiciones conllevaría complicaciones computacionales, se pueden modelar procesos de sustitución para cada posición de la cadena peptídica teniendo en cuenta el cambio de estabilidad debido a la mutación en el contexto de la secuencia media generada por el modelo (Arenas et al. 2015). Esta estrategia, denominada modelo mean-field (MF) impone restricciones para mantener la estabilidad de la estructura nativa y genera distribuciones de aminoácidos específicas para cada posición que pueden utilizarse para producir matrices de sustitución para cada posición utilizables en los métodos de inferencia filogenética (Arenas et al., 2015). Sin embargo, para que el cálculo de la estabilidad sea factible, estos modelos asumen que las mutaciones no alteran la estructura nativa de las proteínas, una simplificación que se pretende superar con este proyecto.

Aunque el modelo MF consigue una mayor verosimilitud frente a los datos que los modelos empíricos, este sobreestima la tolerancia a mutaciones y predice que esta tolerancia alcanza un máximo para posiciones con un número intermedio de contactos en la estructura de la proteína nativa. Sin embargo, la relación entre tolerancia a mutaciones y número de contactos es monotónica para los datos empíricos de alineamientos múltiples (Jiménez et al., 2018), por lo que es evidente que el modelo de estabilidad restringe poco las mutaciones de estas posiciones. Esto se debe a que pueden estar ocupadas por aminoácidos tanto hidrofóbicos como polares, mientras que las posiciones con muchos contactos están enterradas en el interior de la proteína y ocupadas por aminoácidos hidrofóbicos y las posiciones con pocos contactos están expuestas al disolvente y ocupadas por aminoácidos polares. Si bien, es posible que la mutación de las posiciones con un número de contactos intermedio altere la estructura nativa a través de factores que no se tienen en cuenta apropiadamente en este modelo (por ejemplo, el tamaño de los aminoácidos y efectos estéricos). Todo esto implica que, al menos para posiciones con número de contactos intermedio, las restricciones estructurales pueden ser mayor que las de estabilidad.

Asimismo, es preciso señalar la existencia de modelos que imponen restricciones a través de la estructura en vez de la estabilidad, que predicen la relación mencionada anteriormente de manera correcta (Huang et al., 2014). Además, la diferencia de la conservación de estructura y secuencia entre proteínas homologas apunta a un rol tan importante o más de las restricciones estructurales, al ser mayor la conservación de la estructura y más lenta su evolución (Pascual-García et al., 2019). Por ende, es de interés la integración de estabilidad y estructura en un modelo evolutivo para mejorar la inferencia filogenética. Cabe destacar que, de momento, no se conocen maneras fiables de predecir el efecto estructural de una mutación concreta sobre la estructura nativa de la proteína. Por esta razón, en este trabajo se propone desarrollar un modelo predictivo que sea por lo menos capaz de predecir los cambios de manera estadística (o sea que exista una correlación entre el tamaño del cambio predicho y observado) con un coste computacional reducido, de forma que sea posible predecir el efecto de todos los posibles cambios de amino ácido de una proteína.

1.3. Análisis de modos normales

En este contexto, el análisis de modos normales (NMA) supone una técnica analítica que nos permite examinar los movimientos accesibles a la estructura nativa de la proteína. El NMA resulta en modos normales de vibración, descripciones de movimientos oscilatorios independientes caracterizados por un autovector (*eigenvector*) que define la dirección y amplitud del movimiento de cada átomo y un autovalor (*eigenvalue*) que define la frecuencia con la que todas las masas del sistema se mueven alrededor de su posición en el mínimo de energía. De entre estos modos normales, los modos de baja frecuencia suponen un menor coste energético y suelen consistir en desplazamientos lentos y más colectivos que los de alta frecuencia (Bauer et al., 2019). Son precisamente estos movimientos colectivos los que son de mayor interés para el estudio funcional de la estructura y los que han demostrado repetidamente ser robustos (Bahar & Rader, 2005) lo que sugiere que se trata de mociones intrínsecas a la estructura (Bahar et al., 2015).

1.4. Modelos de red elástica

Para simplificar los cálculos y aliviar la carga computacional asociada al NMA, Tirion M. (1996) simplificaba el potencial semi-empírico comúnmente utilizado por uno que instaura un solo parámetro, una constante de rigidez que rige la interacción entre los pares de partículas, de los cuales, solo se considera que contribuyen los que tienen una distancia menor a un límite (7.0-8.0 Å) y asumiendo que la estructura a analizar se encuentra ya en su forma de energía minimizada, eliminando la necesidad de realizar este proceso computacionalmente. Esto motivó la aparición de métodos bajo el mismo principio llamados Modelos de Red Elástica (ENM) ya que se puede visualizar las estructuras como sistemas simplificados de masas unidas por muelles con una misma constante de fuerza si se encuentran a una distancia por debajo del límite establecido. De esta manera, las fuerzas elásticas ejercidas por los muelles simulan las fuerzas interatómicas presentes en la estructura original.

Se conoce la conservación de los modos normales de baja frecuencia en estructuras similares y es frecuente la sospecha de que está relacionada con la conservación de la función en la estructura (Bahar et al., 2010; Ma, 2005; Maguid et al., 2008). Sin embargo, esta puede no ser la única razón pues modelando mutaciones aleatorias como una fuerza que perturba la estructura de la red elástica, Echave J. (2008) muestra que incluso bajo este efecto aleatorio los modos normales de baja frecuencia suelen mostrar una mayor conservación, sugiriendo que estos son más robustos de cara a perturbaciones en la estructura y esta conservación es una respuesta a esperar de cara a cualquier perturbación, incluso sin un efecto de selección.

1.5. Objetivos y planteamiento

En este proyecto se persigue la composición de un modelo evolutivo de proteínas que integre el efecto de las mutaciones sobre la estabilidad (Arenas et al. 2015) y sobre la estructura de estas con la expectativa de que las restricciones estructurales solucionen la discrepancia con los datos empíricos, puesto que considerar la estabilidad de la estructura nativa no es suficiente para modelar las tasas de sustitución observadas en un alineamiento múltiple (Bastolla et al., 2017; Jiménez et al., 2018).

En primer lugar, de manera similar a Echave J. (2008), se analizan los efectos de diferentes mutaciones modeladas como fuerzas sobre la estructura, representada en este caso bajo un modelo de red elástica que utiliza los ángulos de torsión phi y psi de la cadena peptídica como grados de libertad, lo cual reduce la carga computacional y representa mejor los cambios conformacionales en las proteínas (Mendez & Bastolla, 2010). En este caso, se simulan todas las mutaciones posibles para todas las posiciones y se calculan las deformaciones resultantes mediante el modelo de red torsional (Torsional Network Model: TNM) utilizando las mismas estructuras utilizadas por Echave J. en un estudio más reciente (Echave & Fernández, 2010), en el que se muestra que las regiones más estructuralmente conservadas en las estructuras naturales se encontraban también conservadas en las estructuras sometidas a estas mutaciones simuladas. De esta manera se obtiene información sobre que posiciones tienen un mayor impacto en la estructura al ser mutadas.

En segundo lugar, se combinan las desviaciones estructurales producidas por el TNM con los modelos de estabilidad implementados en el programa Prot Evol (Arenas et al., 2015). Se estudia el efecto sobre los parámetros del modelo de la regularización de las frecuencias observadas y se evalúa la mejoría del modelo respecto a los modelos de estabilidad y las observaciones empíricas (alineamientos múltiples) en términos de verosimilitud (*likelihood*) de las frecuencias observadas en cada posición respecto al modelo, entropía de secuencia y tasa de sustitución predicha por el modelo en cada posición.

2

Materiales y Métodos

2.1. Obtención de desviaciones estructurales

2.1.1. Predicción del efecto de las mutaciones mediante el TNM

El programa TNM toma como input un archivo pdb que asocia una estructura tridimensional a una secuencia de aminoácidos e implementa un modelo que considera una mutación de una posición *p* de la secuencia que cambia a un aminoácido *a*, como una fuerza que perturba la estructura. Esta fuerza tiene componentes a lo largo de todos los contactos nativos formados en la estructura por el aminoácido original en la posición *p* y la magnitud de cada una de estas componentes se calcula teniendo en cuenta el cambio de tamaño del aminoácido, el cambio de estabilidad del contacto nativo, y el cambio de distancia óptima entre los residuos en contacto. A partir de la fuerza resultante, se calcula la desviación respecto la estructura nativa como la respuesta lineal de la red elástica representada en el modelo TNM. Para cada una de las 19*N* posibles mutaciones de la secuencia, donde 19 es el número de cambios de aminoácido no idénticos y *N* es el número de aminoácidos de la secuencia, el programa predice el cambio de energía elástica ΔE (DE) y la raíz de la desviación cuadrática media (RMSD),

$$RMSD_{pa} = \sqrt{\frac{\sum_i^N \delta_{pa,i}^2}{N}}$$

Donde $\delta_{pa,i}^2$ es el desplazamiento de la posición *i* entre la estructura nativa y la estructura predicha por el TNM al simular la mutación de $p \rightarrow a$. De manera que, $RMSD_{pa}$ representa la desviación estructural global de la estructura mutante respecto a la estructura nativa. Nótese el uso de *p* como posición en la secuencia e, *i* como posición en la estructura.

Los parámetros del modelo de mutación se obtuvieron mediante un análisis estadístico del Protein Data Bank, y los tres coeficientes C_SIZE, C_STAB y C_DIST se optimizaron en manera de maximizar la correlación entre RMSD predicho y observado en una base de datos de proteínas que difieren por un solo amino acido, filtrada de manera tal que sea razonable atribuir a la mutación la mayor parte del cambio de estructura observado.

Para probar la capacidad del TNM de simular las desviaciones estructurales causadas por mutaciones en la secuencia se realizaron todas las mutaciones posibles en todas las posiciones utilizando la estructura con identificador del Protein Data Bank (PDB id) 1a6m, la misma utilizada en (Echave & Fernández, 2010). A través del TNM, simulando las mutaciones con parámetros: C_SIZE=39, C_STAB=300 y C_DIST=52, se calculó el incremento en la energía elástica y el RMSD para todas las mutaciones posibles en todas las posiciones, y se computaron los promedios entre las 19 posibles mutaciones para cada posición, obteniendo una medida del efecto que tiene en la estructura global mutar esa posición, permitiendo comparar qué posiciones causan una mayor desviación al mutar.

Adicionalmente, el TNM produce perfiles de RMSD local similares a los presentados en el estudio mencionado anteriormente, los cuales ofrecen una medida de la variabilidad estructural de cada posición. Concretamente, se trata de un promedio de las desviaciones causadas en la posición *i* de la estructura para el conjunto de las estructuras producidas por todas las mutaciones posibles $p \rightarrow a$ de la secuencia:

$$RMSD_i^{local} = \frac{\sum_p \sum_a \sqrt{\delta_{pa,i}^2}}{N_{mut}}$$

Donde $\delta_{pa,i}^2$ de nuevo representa el desplazamiento respecto la estructura nativa de la posición *i* en la estructura simulada con la mutación $p \rightarrow a$. En este caso, el numerador representa la suma total de estos desplazamientos en la posición *i* para todos los cambios de aminoácido posibles en todas las posiciones y por tanto para todas las estructuras mutantes simuladas posibles. N_{mut} , es el total de mutaciones posibles ($19N$).

A diferencia del RMSD usual, esta medida nos da una visión local de la desviación estructural que refleja qué posiciones son más variables o robustas a nivel estructural respecto a cualquier mutación en la estructura y nos permite comparar con resultados anteriores.

2.1.2. Robustez de los parámetros de simulación

Para concluir el análisis de las deformaciones producidas por el TNM, se analizó también el efecto de los tres parámetros que simulan las mutaciones: C_SIZE, C_STAB y C_DIST, utilizando la misma estructura y configurando todas las combinaciones posibles de los parámetros usados, con uno o dos de los parámetros establecidos en cero para comprobar la robustez cualitativa de las desviaciones estructurales.

2.2. Análisis del modelo combinado

2.2.1. Computación de las desviaciones estructurales y el modelo combinado

El programa Prot Evol puede calcular la hidrofobicidad, entropía y tasa de sustitución para cada posición de la secuencia de una proteína en base a una variedad de modelos evolutivos. Actualmente hay implementados 2 modelos basados en estabilidad, el modelo mean field (MF) y el modelo wild-type (WT) y un modelo basado en la estructura que

utiliza las alteraciones en DE o RMSD producidas en la estructura a través de las mutaciones simuladas por el TNM. El último modelo y el motivo de este estudio es el modelo que combina los anteriormente mencionados. Este modelo combina automáticamente el modelo de estabilidad y el modelo estructural que mejor se adapta a los datos. Es decir, son posibles 4 combinaciones: MF-RMSD, MF-DE, WT-RMSD y WT-DE.

Previa a la computación con Prot Evol, se obtuvieron las deformaciones de la estructura necesarias para el modelo, RMSD y DE mediante el TNM utilizando los parámetros: C_SIZE=13.2, C_STAB=0 y C_DIST=6.3, optimizados a través de pares de proteínas wild-type y estructuras mutadas determinadas por cristalografía de rayos X en un trabajo anterior en el laboratorio anfitrión. Posteriormente, se utilizó Prot Evol para obtener la hidrofobicidad, entropía y tasa de sustitución bajo el modelo combinado para las 213 proteínas del set de datos presentado por Yeh et al. (2014). En dicho estudio se obtuvieron alineamientos múltiples de secuencias homólogas para 213 proteínas con estructuras previamente determinadas a través de cristalografía de rayos X, ambos argumentos de entrada necesarios para el procesamiento del modelo combinado.

Para la computación con Prot Evol, se utilizaron los parámetros predeterminados, a excepción del parámetro GET_FREQ=3 y el valor del parámetro REG, un factor de regulación que se aplica a las frecuencias de aminoácidos posición-específicas del alineamiento múltiple. Este parámetro es crucial dado que los modelos contienen un parámetro libre, la fuerza de selección para conservar la estructura (lambda) que se ajusta mediante la minimización de la divergencia de Kullback-Leibler (KL) simétrica entre las distribuciones de probabilidad del modelo y la distribución de aminoácidos que se deduce del alineamiento múltiple que se aporta junto a la estructura pdb como input al programa.

2.2.2. Razonamiento y optimización del parámetro REG

Dada la importancia del parámetro REG, merece la pena exponer porqué se requiere este parámetro y cómo influye en los resultados de Prot Evol. Si observamos la formulación de la divergencia KL del modelo respecto al alineamiento utilizada en la optimización del parámetro lambda, para cualquier posición i:

$$KL(P_i^{Model}|P_i^{MSA}) = \sum_a P_i^{Model}(a) \times \log \frac{P_i^{Model}(a)}{P_i^{MSA}(a)}$$

Aparece $P_i^{MSA}(a)$ en el denominador, algo problemático debido a que, mientras que el modelo siempre ofrece una probabilidad mínima para todos los aminoácidos, en los alineamientos frecuentemente hay posiciones en las que nunca aparecen algunos aminoácidos y por tanto $P_i^{MSA}(a) = 0$, precisamente el caso en el que la divergencia de Kullback-Leibler no está definida.

Es debido a esto que se requiere una regularización de las frecuencias en el alineamiento mediante una distribución global, para lo cual Prot Evol sustituye P_i^{MSA} por:

$$P_i^{obs}(a) = \frac{f_{MSA_i}(a) + REG \times f_{glob}(a)}{Z_i}$$

Donde *obs* refiere a *observación* (ya que se trata de probabilidades basadas en frecuencias regularizadas y no directamente del MSA), Z_i es la condición de normalización para que la suma de probabilidades para cada posición sea 1 y la frecuencia global del aminoácido *a* en el alineamiento es:

$$f_{glob}(\mathrm{a}) = \sum_i \frac{f_{MSA}(a)}{L}$$

Se observa en la fórmula de frecuencias regularizadas que la inclusión de la frecuencia global evita probabilidades nulas, pero el peso de estas frecuencias dependerá del parámetro REG y un valor excesivamente alto conllevará la perdida de información posición-específica. El parámetro REG es lo que en la teoría de la regresión se denomina una regularización. En una versión anterior de Prot Evol se determinaba el coeficiente de selección Lambda minimizando la divergencia $KL(P_i^{MSA}|P_i^{Model})$ que se puede calcular también para columnas *i* donde faltan aminoácidos en el alineamiento múltiple. Sin embargo, de esta manera se obtiene un modelo que ajusta muy bien los datos disponibles en cada columna, pero tiene poca capacidad de generalización y es propenso al *overfitting*. Valores más altos de REG sesgan la distribución observada hacia una que no tiene información posicional y empeoran el ajuste entre el modelo y el alineamiento múltiple, pero mejoran la capacidad de generalización.

Para la optimización de este parámetro se decidió buscar el valor de REG para el cual la divergencia de Kullback Leibler es aproximadamente simétrica, KL(mod,obs)≈KL(obs,mod). Este valor cambia para cada proteína. Para elegir un valor óptimo, se utilizaron 8 proteínas con variedad de cualidades como la longitud de la proteína o la entropía y el número de secuencias en el alineamiento: 132L, 1DGK, 1FGH, 1IU4, 1K30, 1O98, 1QWN y 7ATJ. Tras ejecutar Prot Evol para estas proteínas y valores de REG 0.05, 0.07, 0.09, 0.11 y 0.13, un rango en el que se observó que se suele cruzar los valores de las KL en ambas direcciones, se pudo apreciar que existe una correlación entre el valor de REG para el cual KL es simétrica y la entropía del alineamiento. Debido a esto, internamente Prot Evol modula la fuerza de la regularización siguiendo:

$$REG = REG_{input} \times \left(1 - \frac{\langle Entropy \rangle}{\log 20}\right)$$

Donde ⟨Entropy⟩ representa la entropía promedio del alineamiento. De esta manera, se puede usar el mismo valor de REG para todas las proteínas de nuestro conjunto de prueba con sólo pequeñas vulneraciones de la simetría de la divergencia de KL.

2.2.3. Caracterización del modelo combinado y alineamientos

Prot Evol produce ficheros PDB_SSCPE_MOD_rate_profile.dat donde PDB es la estructura que se analiza y MOD el modelo utilizado. En estos ficheros se incluye el número de contactos, la hidrofobicidad media, entropía y tasa de sustitución posición para

cada posición. Para estos 3 atributos del modelo y para la hidrofobicidad media de los alineamientos múltiples también calculada por el programa, se computaron el promedio y el error estándar del promedio después de agrupar todas las posiciones de todas las proteínas analizadas en función del número de contactos. El mismo procedimiento fue aplicado a las tasas de sustitución y entropía ya calculadas por Jiménez et al. (2018) para los alineamientos del set de datos en uso.

3

Resultados

3.1. Simulación de mutaciones a través del TNM

En la figura 1 se presentan los resultados de las predicciones de desviación estructural computadas por el TNM sobre la estructura 1a6m para todas las mutaciones simuladas.

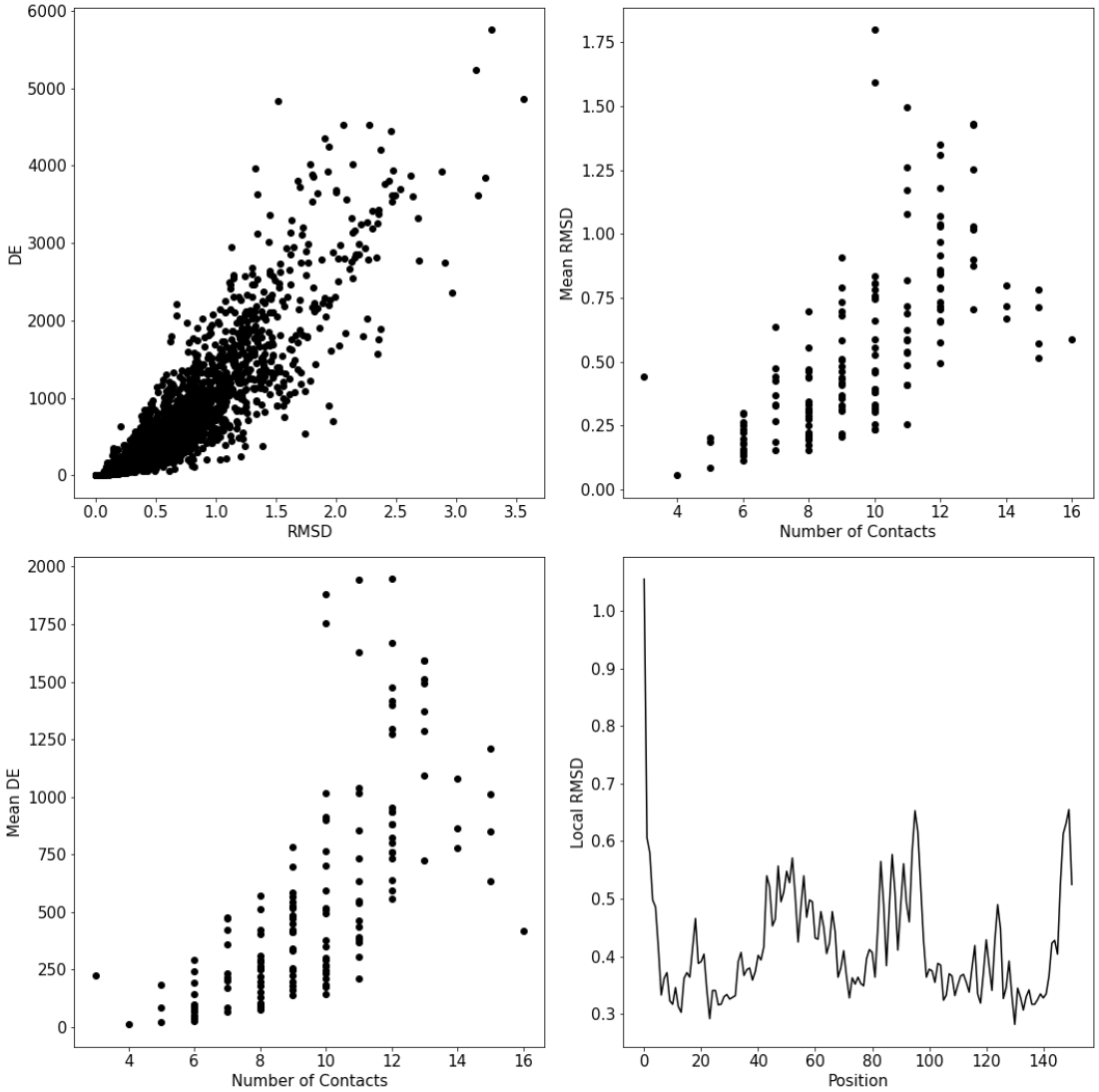

Figura 1. Caracterización de las deformaciones producidas por el TNM para 1a6m. Panel superior izq. Relación desviación estructural (RMSD) e incremento de la energía elástica (DE). Paneles superior dcho. e inferior izq. Gráficos de dispersión en función del número de contactos para RMSD y DE promedio de todas las mutaciones posibles en cada posición. Panel inferior dcho. RMSD local predicho en todas las posiciones de la estructura.

En el primer panel, se aprecia por lo general una correlación lineal entre la deformación estructural y el incremento de la energía elástica de la red. En los siguientes dos paneles, se examina la relación entre estas cantidades y el número de contactos de la posición donde se aplica la mutación y se calcula el promedio de todas las mutaciones posibles, mostrando un incremento en la deformación de la estructura global para posiciones con un mayor número de contactos, siendo la región entre 5 y 12 contactos la más poblada de ambos gráficos (no hay residuos con menos 3 contactos). En último lugar, el perfil de RMSD local por posición muestra como varía la deformación promedio a través de todas las mutaciones posibles de la proteína para cada posición, revela diferencias considerables en el efecto que tienen las mutaciones sobre diferentes posiciones, siendo algunas significativamente más variables que otras. Estas relaciones entre RMSD y DE, entre las mismas y el número de contactos, y los perfiles de RMSD local, se mantienen a pesar de variar los parámetros de la simulación de las mutaciones como se ve en las figuras suplementarias S1-S4, que presentan los paneles de la figura 1 para 6 combinaciones de parámetros diferentes.

3.2. Caracterización del modelo combinado

3.2.1 Optimización del parámetro REG

El parámetro REG acepta valores entre 0-1 y es de crucial importancia para el modelo combinado. Utilizando proteínas con diferentes longitudes y número de secuencias en el

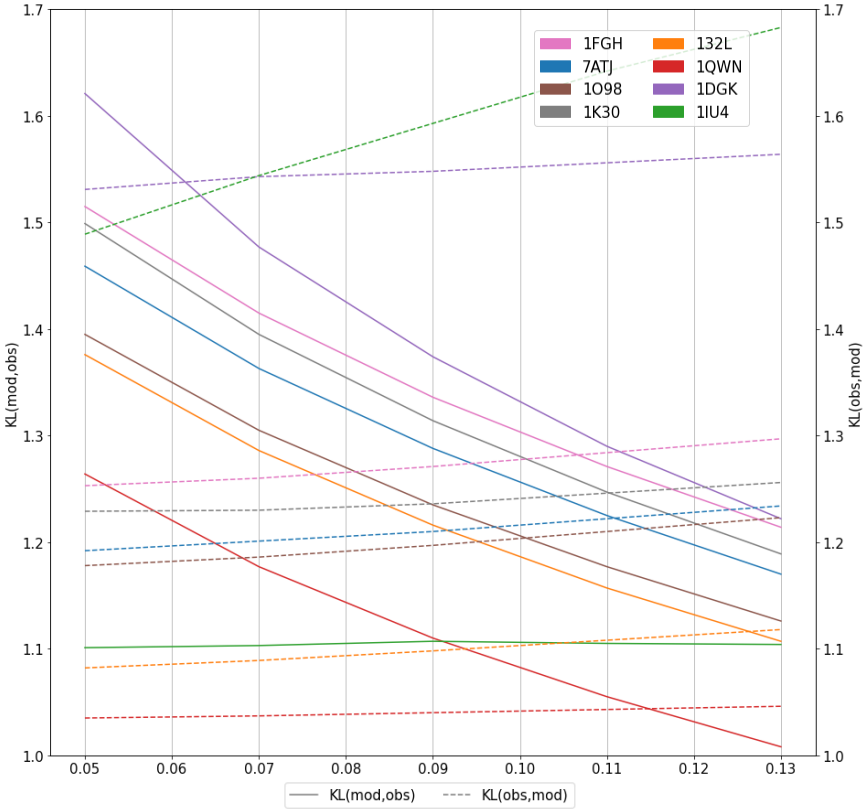

Figura 2. Optimización del parámetro REG a través de la divergencia Kullback Leibler (KL) simétrica del modelo combinado. La divergencia KL de las distribuciones del modelo hacia las de los alineamientos regularizados se muestran en el eje izquierdo con línea continua y la KL de las distribuciones de los alineamientos regularizados hacia las del modelo en el eje derecho con línea discontinua, ambos en función del valor del parámetro REG para 8 proteínas del set de datos.

alineamiento múltiple, se observó que el valor óptimo para muchas de las proteínas se encontraba entre 0.10-0.13 como se puede apreciar en la figura 2. Sin embargo, mientras que la divergencia de las frecuencias modelo hacia las frecuencias regularizadas del alineamiento, KL(mod,obs), desciende con una pendiente significativa a medida que se incremente el valor REG, la divergencia en la otra dirección aumenta muy lentamente por lo que se escogió un valor alto en este rango, 0.12, para conseguir una divergencia simétrica total un poco más baja.

3.2.2. Desempeño del modelo combinado

En todos los casos examinados, el modelo con selección sobre la estructura basado en RMSD alcanza valores de divergencia de KL menores que el mejor modelo basado en estabilidad, y el modelo combinado estabilidad-estructura muestra los mejores resultados.

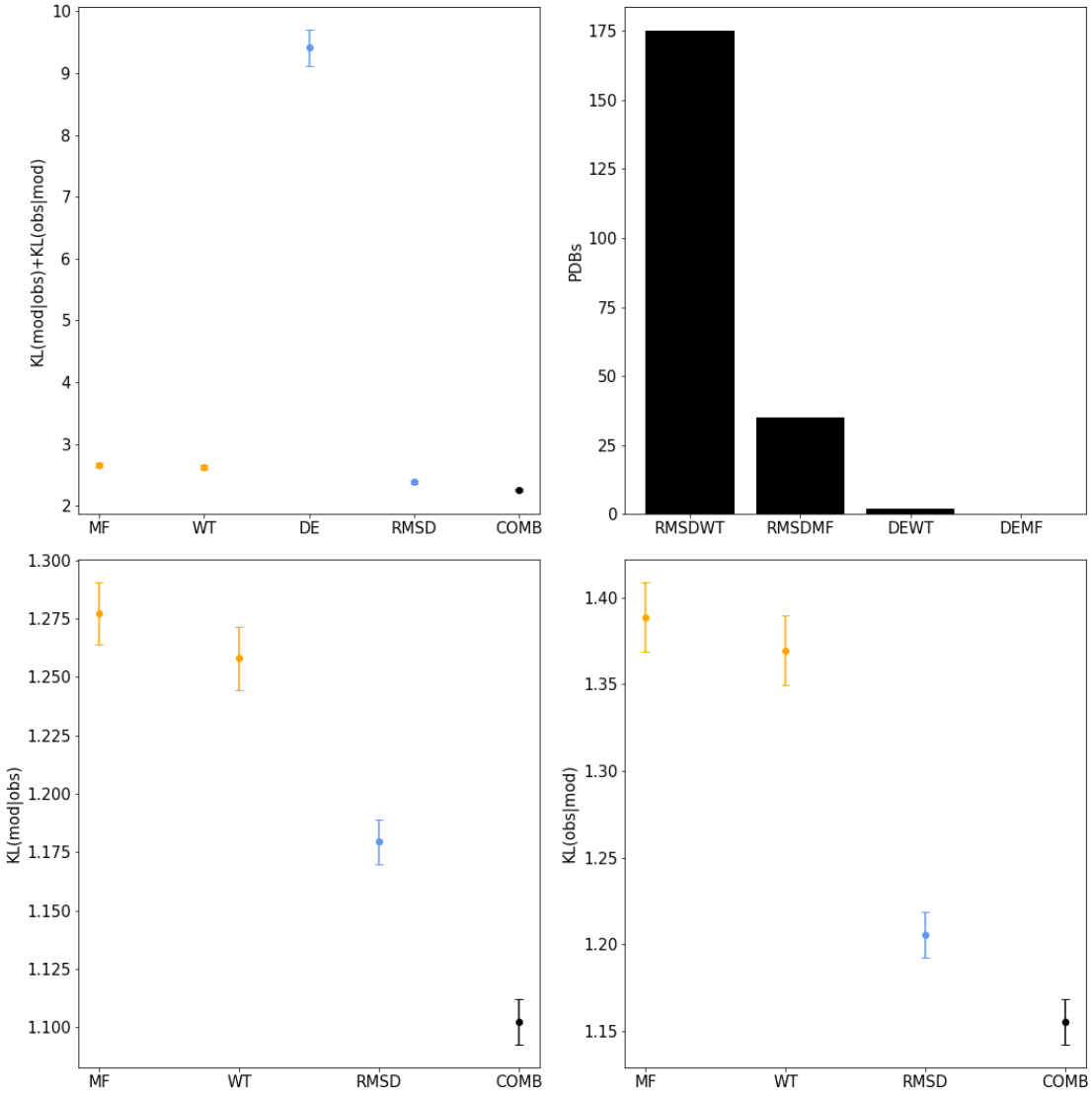

Figura 3. Divergencias Kullback Leibler de los modelos y frecuencia de las combinaciones de modelos. Panel superior izq. suma los promedios de las divergencias Kullback-Leibler y error estándar del promedio para todas las proteínas para los modelos de estabilidad (MF,WT: naranja), modelos estructurales (DE, RMSD: cian) y el modelo combinado. Panel superior dcho. gráfico de barras del número de PDBs para los que cada una de las combinaciones de modelos resulta óptima (menor divergencia). Paneles inferiores: promedio de divergencias KL en ambas direcciones y error estándar para los modelos MF, WT, RMSD y el modelo combinado.

En la figura 3 se muestran los valores promedio de las divergencias KL para los diferentes modelos. Al contrario que en el caso del modelo basado en RMSD, la suma de las divergencias KL es sustancialmente mayor para el modelo DE respecto a todos los demás. Esto se debe principalmente a la divergencia de las observaciones regularizadas hacia el modelo, es decir, KL(obs,mod), que en modelo DE es extraordinariamente alta. Consecuentemente, el modelo DE es elegido escasas veces para el modelo combinado, tal y como se aprecia en el panel superior derecho de la figura 3. De los cuatro posibles modelos combinados, entre las 213 estructuras utilizadas, la combinación RMSDWT fue la óptima en la mayoría de los casos (175), seguida por RMSDMF (35) y DEWT (2). La combinación DEMF no se utilizó en ninguna de las proteínas y una de las estructuras (1I1E) fue eliminada debido a un error durante la computación.

Para las 212 proteínas que pudieron ser computadas, se calculó el promedio de varias cantidades predichas por el modelo para todas las posiciones con un número determinado de contactos. Los resultados se muestran en la figura 4, comparando el comportamiento de los diferentes modelos frente a el comportamiento observado en los alineamientos.

En la primera columna de paneles se puede apreciar una aparente tendencia general del modelo combinado a sumar el comportamiento de los modelos individuales. Adicionalmente, es notable en todos los atributos que para las posiciones con un número de contactos mayor que 13 el error estándar del promedio se incrementa considerablemente respecto a las demás posiciones con menor número de contactos. Además, se debe considerar que, los modelos de estabilidad ambos representan valores sobre 212 proteínas, pero los modelos estructurales de RMSD y DE representan valores para 210 y 2 proteínas respectivamente, razón por la que no hay datos para el comportamiento del modelo DE respecto a posiciones con un número de contactos mayor de 11.

La hidrofobicidad muestra una función monótona creciente respecto al número de contactos para casi todos los modelos, siendo para el modelo combinado la más similar a la observada en los alineamientos cualitativa y cuantitativamente. La entropía del modelo, y tasa de sustitución del modelo combinado, ambas representan una función monótona decreciente hasta las posiciones con 14 contactos y son cualitativamente similares a las que se observan para los alineamientos. Sin embargo, aunque la variación cuantitativa es similar, ambas, son significativamente mayores para todos los modelos respecto a los alineamientos. Asimismo, los modelos basados en estabilidad producen curvas con un máximo de entropía alrededor de 4-5 contactos a pesar de que esto no ocurre para los alineamientos.

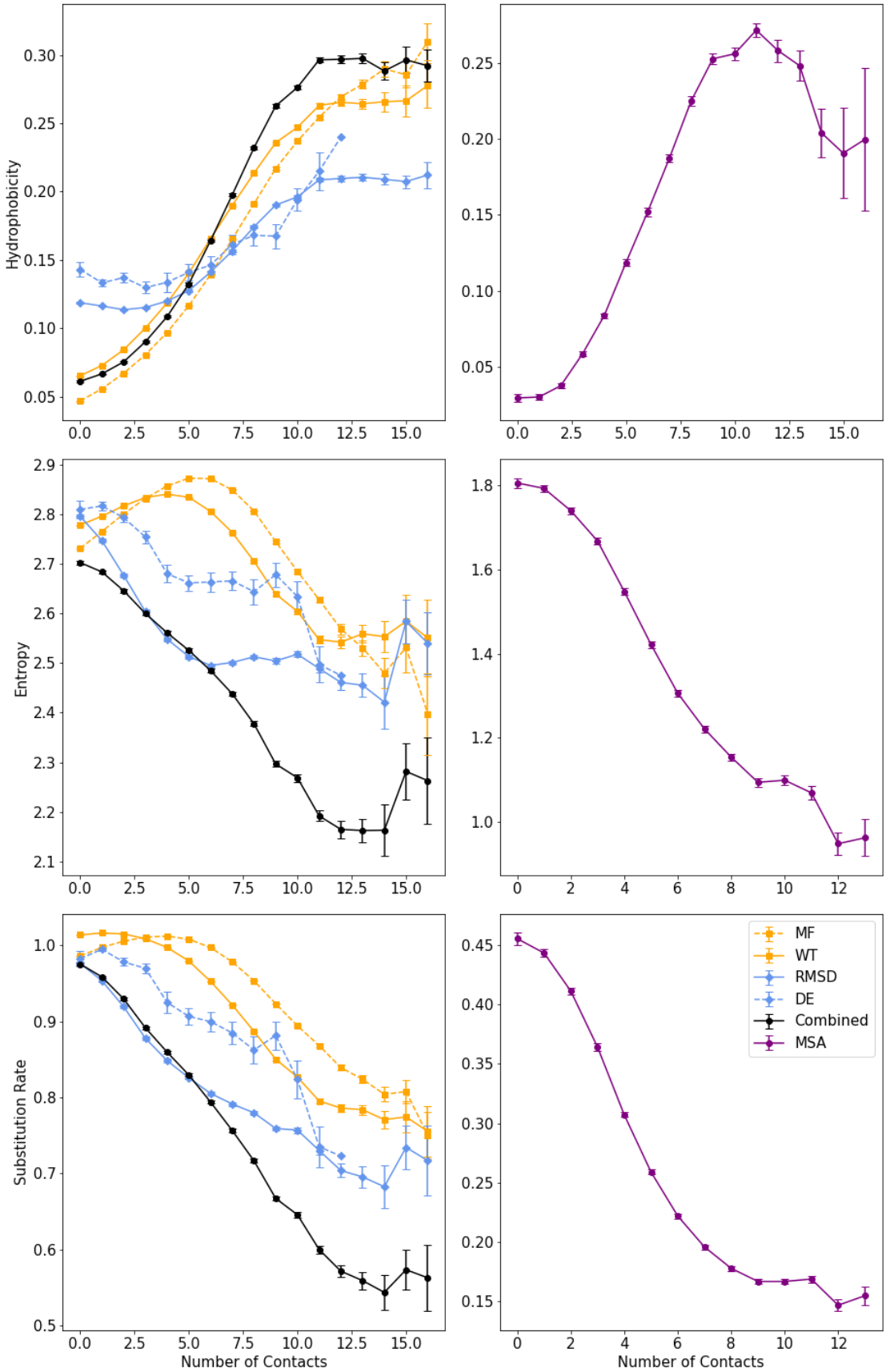

Figura 4. Comparación de las predicciones de los modelos y los alineamientos múltiples. Columna izq. hidrofobicidad, entropía y tasa de sustitución promedio por número de contactos y error estándar del promedio para los modelos MF (amarillo, línea discontinua), WT (amarillo, línea continua), RMSD (azul, línea continua), DE (amarillo, línea discontinua) y el modelo combinado (negro, línea continua). Columna dcha. hidrofobicidad, entropía y tasa de sustitución promedio por número de contactos y error estándar del promedio para los alineamientos múltiples (morado, línea continua).

4

Discusión

Los modelos basados en estabilidad utilizan conceptos de genética de poblaciones como puente para introducir procedimientos de la física en los modelos de evolución. De un modo u otro se formula una aptitud (*fitness*), concepto alrededor del que se organizan la mayoría de los modelos evolutivos, en función de la estabilidad termodinámica (Echave & Wilke, 2017). Es evidente que hay multitud de factores interrelacionados más allá de la estabilidad que pueden influir en la aptitud, pero esta resolución no carece de respaldo ni teórico ni empírico. De hecho, la selección en base a la estabilidad para limitar mutaciones que desfavorezcan (y permitir las que favorezcan) el plegamiento de la cadena peptídica en la estructura nativa, es una de las fuerzas de selección que más impacto puede tener en la modulación de la evolución de cada posición (Bastolla et al., 2017; Echave et al., 2016; Lobkovsky et al., 2010). Aun así, a pesar de que muchos de los múltiples factores que podrían afectar a la aptitud se pueden relacionar con la estabilidad, los modelos basados en estabilidad WT y MF, presentados por Jiménez et al. (2018), mostraban una excesiva tolerancia a mutaciones en posiciones con un número de contactos intermedio en comparación con datos de alineamientos múltiples. Se expone también que estas posiciones son las que están menos restringidas por el modelo e hipotetiza que se debe a que las mutaciones en estas posiciones tienen un mayor efecto en la estructura del que tienen en la estabilidad, y que la inclusión del efecto que tiene la mutación sobre la estructura nativa resultaría en un modelo más realista.

Es precisamente este el cometido de la capacidad de simulación del TNM en este proyecto. Dada la robustez de las ENM para representar las dinámicas de las estructuras (Bahar & Rader, 2005; Tama & Sanejouand, 2001) y la posibilidad de simular mutaciones en la red elástica a través de perturbaciones en el potencial de la red (Echave, 2008; Echave & Fernández, 2010; Huang et al., 2014), que reflejan la deslocalización de los efectos estructurales de las mutaciones respecto a la posición en la que ocurren (Sinha & Nussinov, 2001), debería ser posible obtener desviaciones globales de la estructura nativa causadas por las mutaciones simuladas que respeten estos movimientos intrínsecos a la estructura. El TNM ha sido utilizado con éxito previamente para representar cambios de conformación en el espacio de ángulos de torsión (Bastolla & Dehouck, 2019), estudiar los movimientos coordenados entre residuos funcionales en los modos normales

(Alfayate et al., 2019) e investigar la correlación entre los movimientos funcionales y los modos normales de baja frecuencia (Bastolla & Dehouck, 2019; Dos Santos et al., 2013). Con la adición de la capacidad de simular mutaciones de manera similar a estudios anteriores (Echave, 2008; Echave & Fernández, 2010), pero de forma aminoácido especifica (sin publicar), es posible predecir el efecto de una mutación específica sobre la estructura nativa para todas las mutaciones posibles, pudiendo conseguir perfiles exhaustivos de qué posiciones y aminoácidos alteran más la estructura al mutar.

Si las estructuras evolucionan en torno a los modos normales por selección (Bastolla, 2014; Bastolla & Dehouck, 2019; Dos Santos et al., 2013), por la propia configuración de la estructura (Echave, 2012; Echave & Fernández, 2010; Maguid et al., 2008), o una combinación de ambos (Marcos & Echave, 2020), supone aún una cuestión debatible. Sin embargo, irrespectivamente de la causa de estas observaciones, para la tarea en cuestión solo se requiere que el TNM represente estos fenómenos de la manera más fiel posible.

En base a los resultados presentados en los paneles de la figura 1 podemos afirmar que se dan las condiciones necesarias para poder utilizar las desviaciones globales para informar al modelo de estabilidad. El primer panel muestra una correlación lineal entre, RMSD y DE demostrando que los efectos de las mutaciones simuladas por el modelo son coherentes ya que mayores desviaciones de la estructura nativa, supuesta la estructura con energía mínima, deberían requerir mayores incrementos en la energía elástica. Adicionalmente, ambas medidas pueden representar la magnitud de la desviación de la estructura nativa y como se puede apreciar en los paneles B y C, ambos se incrementan con respecto al número de contactos del residuo, uno de los mejores predictores estructurales de la tasa de evolución (Echave et al., 2016; Huang et al., 2014; Yeh et al., 2014). En último lugar, el perfil de RMSD local es visiblemente irregular para las diferentes posiciones, mostrando que algunas acumulan mayores desviaciones irrespectivamente de donde se origine la mutación (Sinha & Nussinov, 2001) y además es similar al que muestran Julián Echave & Fernández (2010) para la misma estructura. Adicionalmente, en las figuras suplementarias S1-S4, se muestra que los resultados mostrados en la figura 1 son robustos respecto a los cambios de parámetros que simulan las mutaciones. En conjunto, estos resultados validan la capacidad del TNM para simular mutaciones y predecir la magnitud de la desviación causada en la estructura respecto a la conformación nativa, en concreto la desviación global, que es la que utiliza Prot Evol.

De cara a la utilización de Prot Evol, la optimización del parámetro REG es compleja y por el momento no hay una manera ideal evidente de llevarla a cabo. En esta ocasión, se buscó un valor donde la divergencia Kullback-Leibler simétrica entre las frecuencias del modelo y las frecuencias regularizadas de los alineamientos fuera la menor posible. El valor utilizado, 0.12, es un valor aproximado que produce divergencias bajas y supone una regularización moderada. En el rango cercano a este valor no se apreciaron diferencias cualitativas en los resultados.

Respecto a los resultados de los modelos, se muestra una mejoría cualitativa del modelo combinado frente a los modelos de estabilidad, en particular con respecto a su capacidad de ajustar la frecuencia de aminoácidos observada en cada columna de los alineamientos

múltiples con un solo parámetro libre que representa la fuerza de la selección natural. Como se observa en la figura 3, aunque el modelo DE no representa correctamente la variabilidad observada en los alineamientos, el modelo RMSD sí que es superior a los modelos de estabilidad dado que se obtienen menores valores de divergencia KL en ambas direcciones. Esta ventaja se transmite también al modelo combinado, que incluye en casi todos los casos el modelo RMSD y que consigue valores de divergencia aún más bajos. En consecuencia, es de esperar que las distribuciones producidas por el modelo combinado resulten en una mayor verosimilitud al construir árboles filogenéticos.

Por otro lado, la dependencia entre el número de contactos nativos de la posición examinada y el promedio de los atributos computados se asemeja notablemente a los de los alineamientos, aunque también el nuevo modelo predice una mayor variabilidad de aminoácidos respecto a lo que se observa en los alineamientos múltiples. La tendencia de residuos hidrofóbicos a ocupar las posiciones interiores de las proteínas es de las observaciones más antiguas de la biología estructural (Perutz et al., 1965). Debido a la proximidad de tantos otros residuos, es de esperar que estas posiciones sean también las que conllevan un mayor número de contactos. Es precisamente esto lo que se observa en la primera fila de paneles de la figura 4. Es notable que el modelo basado sólo en estabilidad predice la hidrofobicidad respecto al número de contactos remarcablemente bien por si solo pero el modelo combinado se asemeja aún más a la hidrofobicidad mostrada en los alineamientos. En cuanto a la entropía por número de contactos, el modelo combinado muestra un comportamiento mucho más similar al del MSA que los modelos basados en estabilidad. Esto sugiere que la inclusión de las restricciones estructurales causa que mutaciones que modifican la estructura, pero conservan la estabilidad (las cuales son aceptables bajo los modelos basados únicamente en estabilidad), se vean desfavorecidas. Al reducir la probabilidad de encontrar aminoácidos que causan la deformación de la estructura se reduce la incertidumbre respecto a qué aminoácido puede ocupar una posición. Esta reducción es apreciable para todas las posiciones, restringiendo las mutaciones tanto o incluso más que los modelos de estabilidad individualmente para posiciones con un número de contactos alto. Es especialmente destacable el cambio cualitativo para posiciones con un número de contactos intermedio, donde los modelos de estabilidad toleraban demasiadas mutaciones, favoreciendo la hipótesis de que estas posiciones (usualmente anfifílicas), tienen una mayor limitación estructural que de estabilidad (Jiménez et al., 2018).

Por último, la tasa de sustitución tiene un comportamiento similar al de la entropía, siendo menor para mayores números de contactos. Algo que resulta intuitivo pues al mutar estas posiciones, es muy probable romper interacciones importantes para la estabilidad de la estructura además de las desviaciones en la estructura que puede provocar la introducción de un aminoácido de diferente tamaño en estas posiciones con alta densidad atómica en su proximidad. Tal y como exponen más formalmente Yeh et al. (2014), es probable que la relación entre el número de contactos y medidas similares y las tasas de sustitución se deba a que las posiciones con menos contactos y usualmente en el exterior pueden acomodar más fácilmente mutaciones debido a una mayor flexibilidad. Cabe mencionar sin embargo, que en otro estudio (Huang et al., 2014) se muestra que las tasas de

sustitución se correlacionan mejor con el estrés local introducido por la mutación en los muelles de la red elástica que modela la estructura activa (el mínimo de energía por defecto), una medida que involucra la flexibilidad y la densidad atómica local. Es más, una vez se ha tenido en cuenta este estrés local, la contribución de la flexibilidad es mínima.

En cualquier caso, la evidencia apoya que la corrección de la entropía que se aprecia en los resultados se debe a la inclusión de la información estructural y se hace evidente que las posiciones del interior son, en efecto las que conllevan una mayor restricción y están altamente restringidas en ambos aspectos de estabilidad y estructura.

Globalmente, es remarcable la semejanza de los valores de hidrofobicidad entre el modelo combinado y los alineamientos. Sin embargo, hay grandes diferencias entre los valores de la entropía y las tasas de sustitución siendo ambas mucho más altas para todos los modelos. Esto puede deberse en gran medida a que las secuencias en los alineamientos no son secuencias independientes y no representan necesariamente la variedad completa de secuencias viables, lo que significa que se infraestima la entropía de secuencia y tasa de sustitución. Por otro lado, aún puede haber factores que el modelo no tiene en cuenta, puesto que la combinación de las probabilidades posición específicas de los modelos lo que nos indican es, qué aminoácidos son más o menos factibles en las diferentes posiciones bajo los principios que rigen el modelo. La falta de la inclusión de estos factores que restringirían aún más las probabilidades incrementa también la entropía y las tasas de sustitución del modelo.

El principal problema de la predicción de tasas de sustitución es que hay multitud de factores que afectan a la tasa de sustitución de una posición. Dichos factores influyen en cada posición de manera diferente y además pueden ser en sí mismos dependientes de las demás posiciones (Bastolla et al., 2017; Echave et al., 2015, 2016; Pollock et al., 2012; Sikosek & Chan, 2014). Esto invoca la necesidad de las tasas de sustitución específicas para cada posición y resulta en un entramado de efectos difícil de resolver donde las mediciones que podemos obtener suelen estar sujetas a múltiples efectos, pero sin embargo ninguna es capaz por si sola de capturar toda la influencia de estos factores en las tasas de sustitución. En este contexto, el modelo combinado consigue abarcar dos grandes factores, la estabilidad y desviación estructural, poniendo el foco en la conformación nativa. Como se ha expuesto, esta combinación corrige defectos del modelo basado sólo en estabilidad y efectivamente se observa la emergencia de un comportamiento similar al observado empíricamente a partir de principios biofísicos, mientras que se mantiene un coste computacional bajo.

5

Conclusión

El análisis de las desviaciones estructurales producidas a través del TNM muestra que este puede representar mutaciones de forma coherente con principios teóricos y observaciones en estudios previos y que estos resultados son persistentes respecto a cambios a los parámetros de simulación. Consecuentemente, estos resultados se pueden utilizar como información estructural para integrarla con el modelo de estabilidad. El resultado de esta combinación muestra que la información estructural mejora el ajuste entre el modelo y los aminoácidos observados en cada posición (menor divergencia KL). Además, corrige la sobreestimación de la tolerancia a mutaciones que se observaba para posiciones con un número de contactos intermedio en el modelo de estabilidad y se asemeja más en todos los parámetros estudiados a la evidencia empírica, por lo que además de implicar directa y explícitamente factores estructurales en la variabilidad de las posiciones con un número de contactos intermedio, cabe esperar un mejor desempeño de este modelo en la reconstrucción de matrices de sustitución para inferencia filogenética.

Bibliografía

Alfayate, A., Rodriguez Caceres, C., Gomes Dos Santos, H., & Bastolla, U. (2019). Predicted dynamical couplings of protein residues characterize catalysis, transport and allostery. *Bioinformatics (Oxford, England)*, *35*(23), 4971-4978. https://doi.org/10.1093/bioinformatics/btz301

Arenas, M., Sánchez-Cobos, A., & Bastolla, U. (2015). Maximum-Likelihood Phylogenetic Inference with Selection on Protein Folding Stability. *Molecular Biology and Evolution*, *32*(8), 2195-2207. https://doi.org/10.1093/molbev/msv085

Bahar, I., Cheng, M. H., Lee, J. Y., Kaya, C., & Zhang, S. (2015). Structure-Encoded Global Motions and Their Role in Mediating Protein-Substrate Interactions. *Biophysical Journal*, *109*(6), 1101-1109. https://doi.org/10.1016/j.bpj.2015.06.004

Bahar, I., Lezon, T. R., Bakan, A., & Shrivastava, I. H. (2010). Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins. *Chemical Reviews*, *110*(3), 1463-1497. https://doi.org/10.1021/cr900095e

Bahar, I., & Rader, A. (2005). Coarse-grained normal mode analysis in structural biology. *Current Opinion in Structural Biology*, *15*(5), 586-592. https://doi.org/10.1016/j.sbi.2005.08.007

Bastolla, U. (2014). Detecting Selection on Protein Stability through Statistical Mechanical Models of Folding and Evolution. *Biomolecules*, *4*(1), 291-314. https://doi.org/10.3390/biom4010291

Bastolla, U., & Dehouck, Y. (2019). Can Conformational Changes of Proteins Be Represented in Torsion Angle Space? A Study with Rescaled Ridge Regression. *Journal of Chemical Information and Modeling*, *59*(11), 4929-4941. https://doi.org/10.1021/acs.jcim.9b00627

Bastolla, U., Dehouck, Y., & Echave, J. (2017). What evolution tells us about protein physics, and protein physics tells us about evolution. *Folding and binding • Proteins: Bridging theory and experiment*, *42*, 59-66. https://doi.org/10.1016/j.sbi.2016.10.020

Bauer, J. A., Pavlović, J., & Bauerová-Hlinková, V. (2019). Normal Mode Analysis as a Routine Part of a Structural Investigation. *Molecules*, *24*(18), 3293. https://doi.org/10.3390/molecules24183293

Dos Santos, H. G., Klett, J., Méndez, R., & Bastolla, U. (2013). Characterizing conformation changes in proteins through the torsional elastic response. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1834*(5), 836-846. https://doi.org/10.1016/j.bbapap.2013.02.010

Echave, J. (2008). Evolutionary divergence of protein structure: The linearly forced elastic network model. *Chemical Physics Letters*, *457*(4), 413-416. https://doi.org/10.1016/j.cplett.2008.04.042

Echave, J. (2012). Why are the low-energy protein normal modes evolutionarily conserved? *Pure and Applied Chemistry*, *84*(9), 1931-1937. https://doi.org/10.1351/PAC-CON-12-02-15

Echave, J., & Fernández, F. M. (2010). A perturbative view of protein structural variation. *Proteins: Structure, Function, and Bioinformatics*, *78*(1), 173-180. https://doi.org/10.1002/prot.22553

Echave, J., Jackson, E. L., & Wilke, C. O. (2015). Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. *Physical Biology*, *12*(2), 025002. https://doi.org/10.1088/1478-3975/12/2/025002

Echave, J., Spielman, S. J., & Wilke, C. O. (2016). Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics*, *17*(2), 109-121. https://doi.org/10.1038/nrg.2015.18

Echave, J., & Wilke, C. O. (2017). Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence. *Annual Review of Biophysics*, *46*(1), 85-103. https://doi.org/10.1146/annurev-biophys-070816-033819

Huang, T.-T., del Valle Marcos, M. L., Hwang, J.-K., & Echave, J. (2014). A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. *BMC Evolutionary Biology*, *14*(1), 78. https://doi.org/10.1186/1471-2148-14-78

Jimenez, M. J., Arenas, M., & Bastolla, U. (2018). Substitution Rates Predicted by Stability-Constrained Models of Protein Evolution Are Not Consistent with Empirical Data. *Molecular Biology and Evolution*, *35*(3), 743-755. https://doi.org/10.1093/molbev/msx327

Lobkovsky, A. E., Wolf, Y. I., & Koonin, E. V. (2010). Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proceedings of the National Academy of Sciences*, *107*(7), 2983-2988. https://doi.org/10.1073/pnas.0910445107

Ma, J. (2005). Usefulness and Limitations of Normal Mode Analysis in Modeling Dynamics of Biomolecular Complexes. *Structure*, *13*(3), 373-380. https://doi.org/10.1016/j.str.2005.02.002

Maguid, S., Fernandez-Alberti, S., & Echave, J. (2008). Evolutionary conservation of protein vibrational dynamics. *Gene*, *422*(1), 7-13. https://doi.org/10.1016/j.gene.2008.06.002

Marcos, M. L., & Echave, J. (2020). The variation among sites of protein structure divergence is shaped by mutation and scaled by selection. *Current Research in Structural Biology*, *2*, 156-163. https://doi.org/10.1016/j.crstbi.2020.08.002

Mendez, R., & Bastolla, U. (2010). Torsional Network Model: Normal Modes in Torsion Angle Space Better Correlate with Conformation Changes in Proteins. *Physical Review Letters*, *104*(22), 228103. https://doi.org/10.1103/PhysRevLett.104.228103

Pascual-García, A., Arenas, M., & Bastolla, U. (2019). The Molecular Clock in the Evolution of Protein Structures. *Systematic Biology*, *68*(6), 987-1002. https://doi.org/10.1093/sysbio/syz022

Perutz, M. F., Kendrew, J. C., & Watson, H. C. (1965). Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. *Journal of Molecular Biology*, *13*(3), 669-678. https://doi.org/10.1016/S0022-2836(65)80134-6

Pollock, D. D., Thiltgen, G., & Goldstein, R. A. (2012). Amino acid coevolution induces an evolutionary Stokes shift. *Proceedings of the National Academy of Sciences*, *109*(21), E1352-E1359. https://doi.org/10.1073/pnas.1120084109

Sikosek, T., & Chan, H. S. (2014). Biophysics of protein evolution and evolutionary protein biophysics. *Journal of The Royal Society Interface*, *11*(100), 20140419. https://doi.org/10.1098/rsif.2014.0419

Sinha, N., & Nussinov, R. (2001). Point mutations and sequence variability in proteins: Redistributions of preexisting populations. *Proceedings of the National Academy of Sciences*, *98*(6), 3139-3144. https://doi.org/10.1073/pnas.051399098

Tama, F., & Sanejouand, Y.-H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein Engineering, Design and Selection*, *14*(1), 1-6. https://doi.org/10.1093/protein/14.1.1

Tirion, M. M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical Review Letters*, *77*(9), 1905-1908. https://doi.org/10.1103/PhysRevLett.77.1905

Yeh, S.-W., Liu, J.-W., Yu, S.-H., Shih, C.-H., Hwang, J.-K., & Echave, J. (2014). Site-Specific Structural Constraints on Protein Sequence Evolutionary Divergence: Local Packing Density versus Solvent Exposure. *Molecular Biology and Evolution*, *31*(1), 135-139. https://doi.org/10.1093/molbev/mst178

Material Suplementario

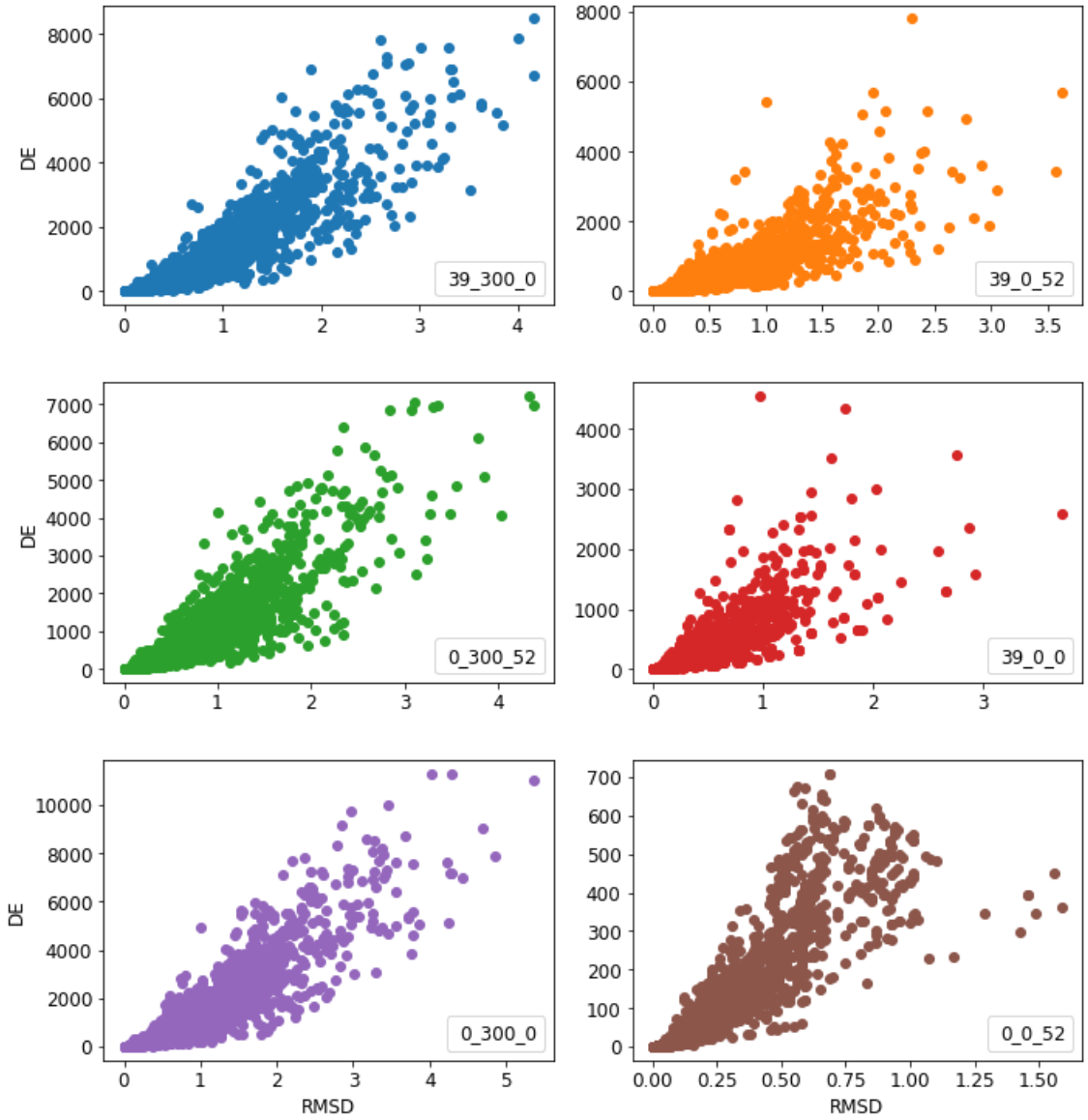

Figura S1. Relación de la desviación estructural (RMSD) e incremento de la energía elástica (DE) predichas por el TNM al introducir las deformaciones producidas por las mutaciones simuladas en la estructura 1a6m para seis combinaciones de parámetros C_SIZE, C_STAB y C_DIST.

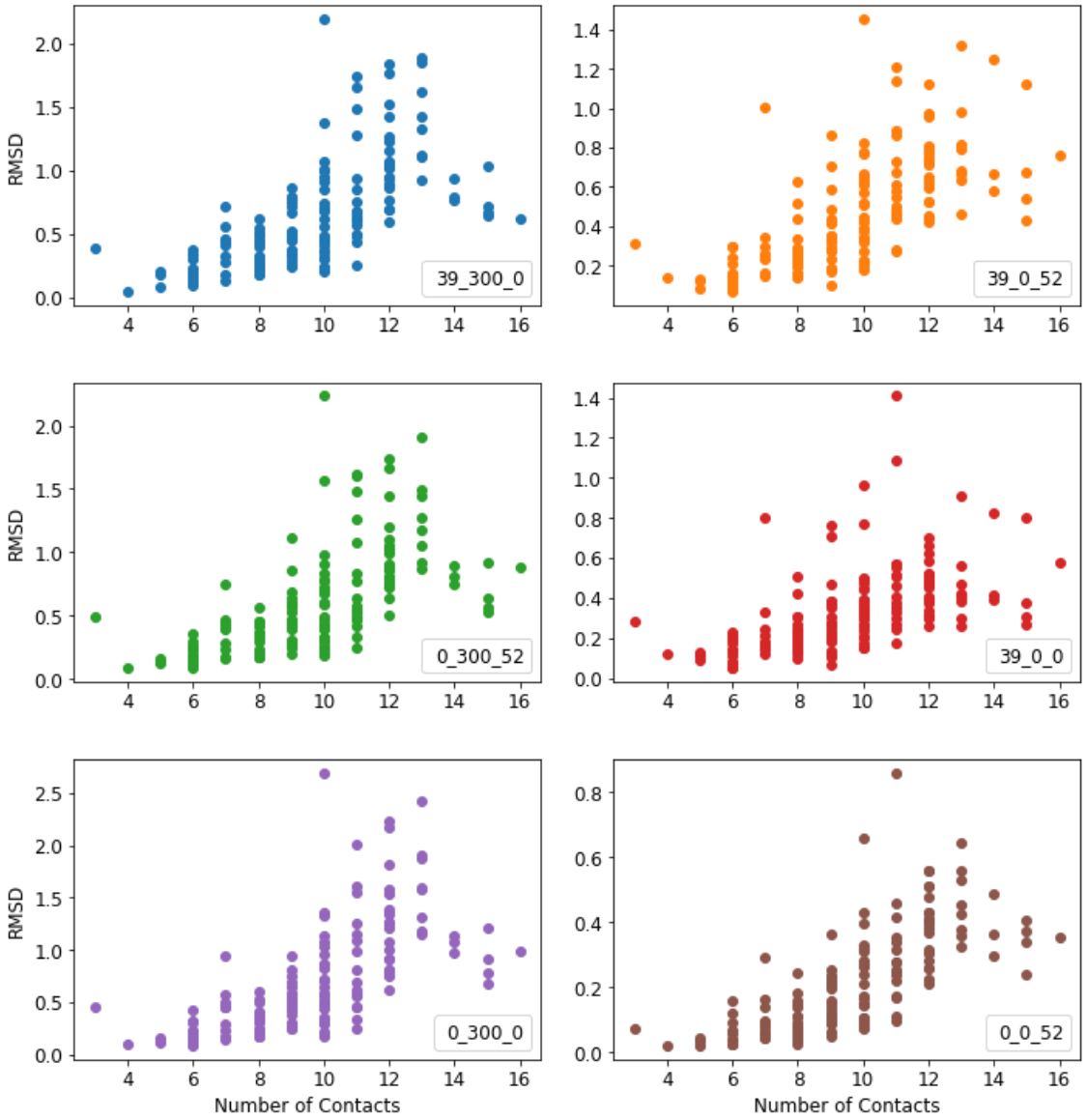

Figura S2. Gráficos de dispersión en función del número de contactos del promedio de todas las mutaciones posibles en cada posición del RMSD predicho por el TNM en la estructura 1a6m para seis combinaciones de parámetros C_SIZE, C_STAB y C_DIST

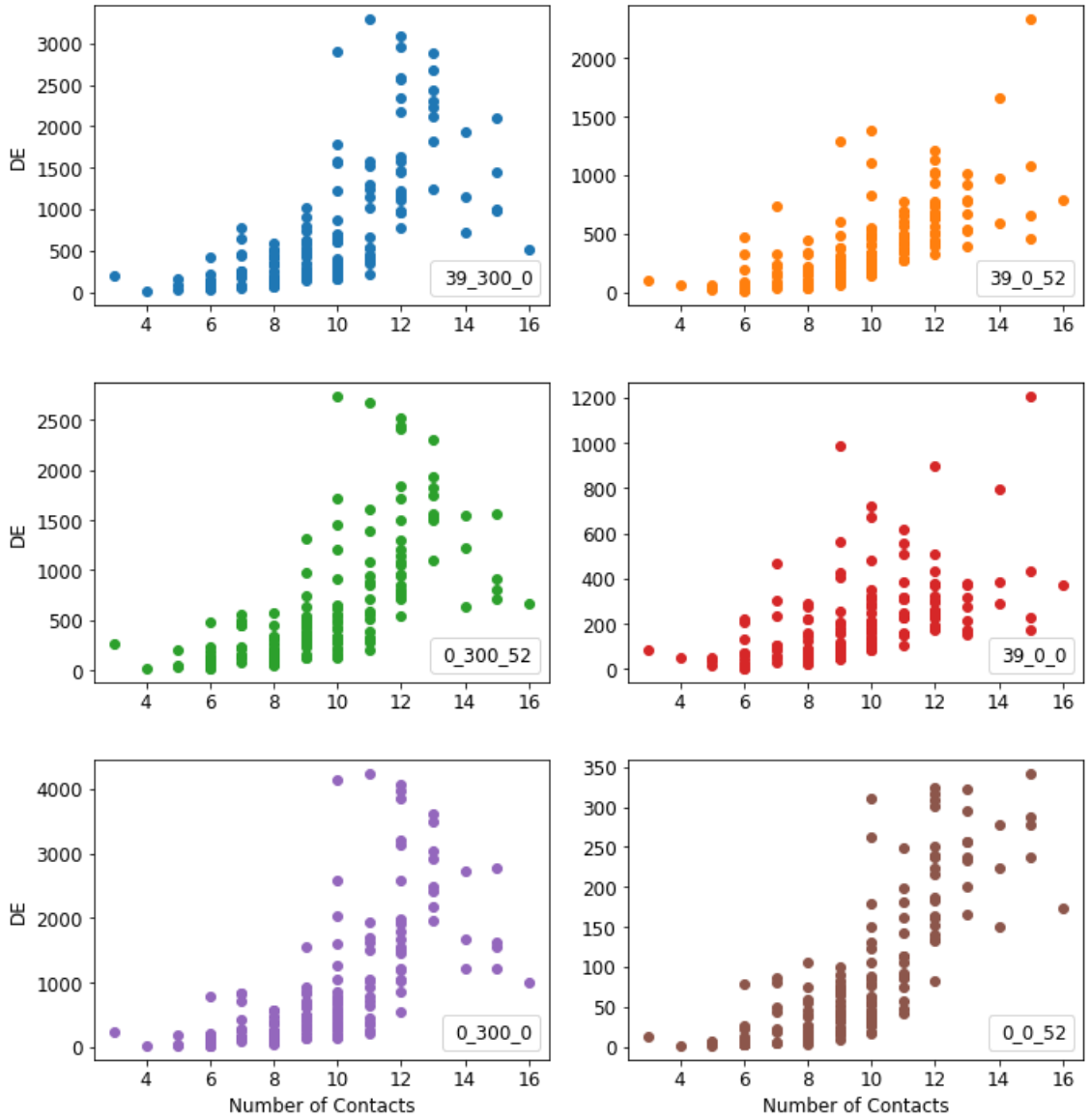

Figura S3. Gráficos de dispersión en función del número de contactos del promedio de todas las mutaciones posibles en cada posición del DE predicho por el TNM en la estructura 1a6m para seis combinaciones de parámetros C_SIZE, C_STAB y C_DIST

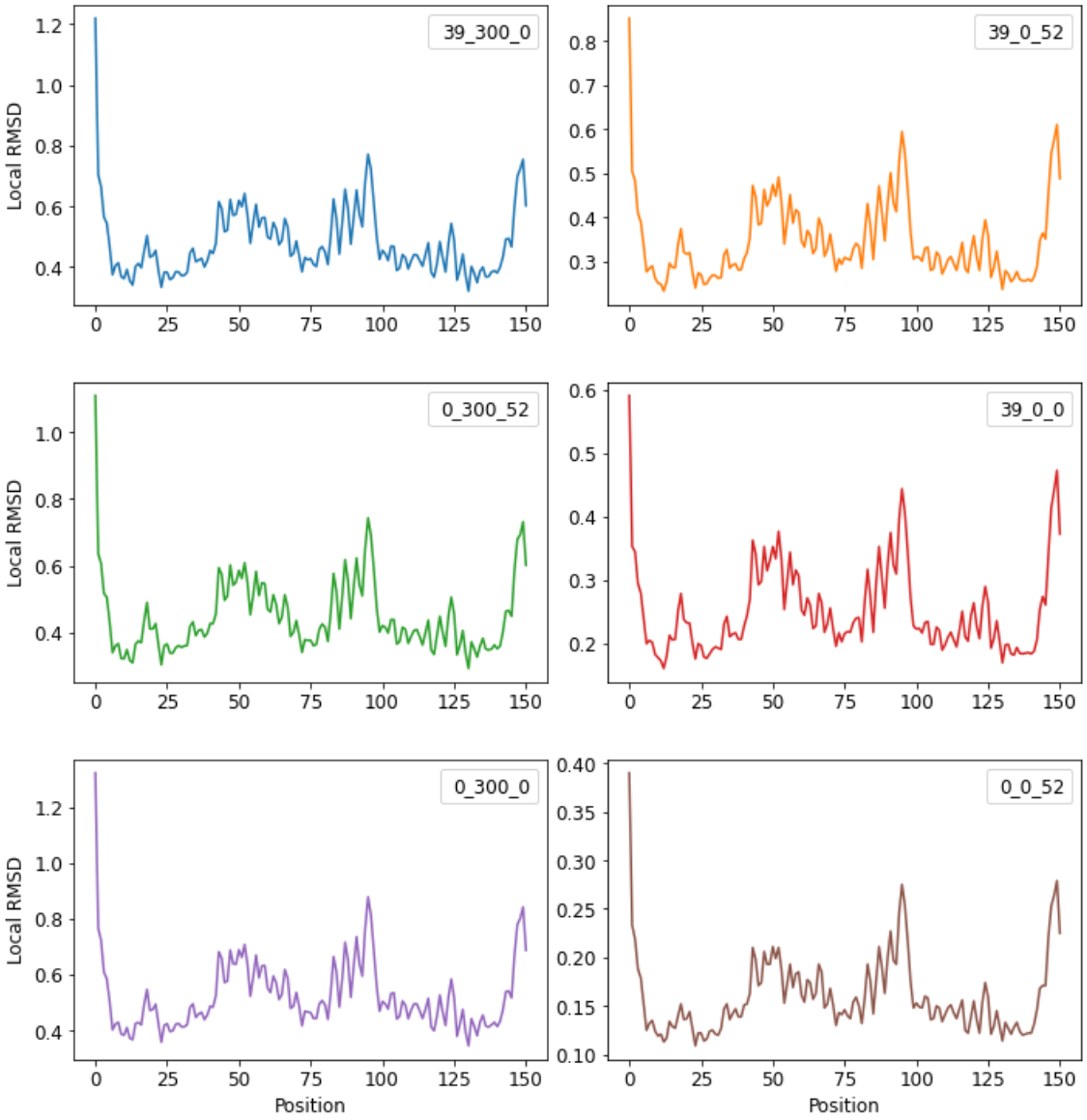

Figura S4. RMSD local predicho por el TNM en todas las posiciones de la estructura 1a6m para seis combinaciones de parámetros C_SIZE, C_STAB y C_DIST