UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS BIOQUíMICAS

INSTITUTO DE BIOTECNOLOGíA

DEPARTAMENTO DE INGENIERÍA CELULAR Y BIOCATÁLISIS

“ANÁLISIS Y DISEÑO DE DOMINIOS ROSSMANN DE UNIÓN A DINUCLEÓTIDOS”

T E S I S

QUE PARA OPTAR POR EL GRADO ACADÉMICO DE

MAESTRO EN CIENCIAS

P R E S E N T A

JESÚS AGUSTÍN BANDA VÁZQUEZ

TUTOR PRINCIPAL:
DR. LORENZO PATRICK SEGOVIA FORCELLA (IBT-UNAM)

MIEMBROS DEL COMITÉ TUTOR:
DR. JOEL OSUNA QUINTERO (IBT-UNAM)
DR. ALEJANDRO SOSA PEINADO (FAC. DE MEDICINA, UNAM)

CUERNAVACA, MORELOS, AGOSTO 2013

El presente trabajo fue realizado gracias a la dirección del Dr. Lorenzo Patrick Segovia Forcella, cuyo laboratorio pertenece al Departamento de Ingeniería Celular y Biocatálisis del Instituto de Biotecnología (IBT) de la Universidad Nacional Autónoma de México. Muchas gracias Lorenzo por compartirme tu visión, confianza y paciencia, así como también apoyarme económicamente en los momentos en que no contaba con beca a finales de la licenciatura y finales de la maestría.

El proyecto fue financiado por la Dirección General de Asuntos del Personal Académico (DGAPA):IN213511, y el Consejo Nacional de Ciencia y Tecnología (CONACYT): 132580. Adicionalmente, agradezco al CONACYT por la beca otorgada con el No. de CVU: 388734.

Agradezco también a los miembros de mi comité tutoral por las recomendaciones otorgadas a lo largo del desarrollo de este trabajo: al Dr. Alejandro Sosa Peinado y al Dr. Joel Osuna Quintero, quien con anterioridad me otorgó un apoyo económico durante mis estudios de licenciatura, y a quien yo nunca me había dado la oportunidad de agradecer tal gesto y lo mucho que significó para mí esa ayuda para continuar con mi formación académica y comenzar la maestría.

Mucho he de agradecer también a mis compañeros de laboratorio y demás personas del IBT que me han acompañado y aconsejado a lo largo de este capítulo de mi vida que fue la Maestría en Ciencias Bioquímicas, en especial a la Dra. Claudia Martínez Anaya, quien me tuvo la paciencia para asesorarme en el aprendizaje de técnicas experimentales cuando recién llegué a este laboratorio.

Y por último, pero no menos importante, agradezco a mis padres, quienes hasta ahora han sido mis eternos patrocinadores.

Abreviaturas y palabras clave

- AntiSCA. Que no es SCA. A lo largo del texto este término se refiere a una molécula o un conjunto de moléculas cuyo diseño se hizo variando sólo a los residuos identificados como estadísticamente acoplados según SCA.
- BLASTp. Protein Basic Local Alignment Seach Tool. Es una herramienta informática de búsqueda por similitud de secuencia.
- CATH. Class Architecture Topology Homology. Es una base de datos de clasificación de proteínas.
- Eigenmodo. Representa una combinación ponderada de posiciones en secuencia.
- Eigenvalor. Es el múltiplo escalar de un eigenvector. En este trabajo representa qué tan estadísticamente significativo es su eigenvector asociado.
- Eigenvector. Es un vector diferente de cero que al ser multiplicado por la matriz de interés da lugar a un múltiplo escalar de tal vector. En este trabajo, los eigenvectores contienen los pesos para combinar linealmente las variables originales (grupos de posiciones de secuencia) en nuevas variables independientes (los eigenmodos).
- ICA. Independent Component Analysis. Es un método computacional para separar un conjunto en subconjuntos estadísticamente independientes entre sí.
- K. Kelvin, la unidad de temperatura. Equivale al valor en grados centrígrados (°C) más 273.15.
- MSA. Multiple Sequence Alignment. Un alineamiento múltiple de secuencias homólogas, en este caso de proteínas.
- NAD. Nicotinamide Adenine Dinucleotide. Es un cofactor.
- NADP. Nicotinamide Adenine Dinucleotide Phosphate. Es un cofactor.
- NCBI. National Center for Biotechnology Information.
- NPT. Moles (N), Presión y Temperatura. También llamado ensamble isotérmico-isobárico. Es un protocolo de restricciones a usar durante una simulación de dinámica molecular en la cual el número de moléculas presente en el sistema, la presión y la temperatura no cambiarán a lo largo del tiempo.
- NR. No redundante. Se refiere en específico a una base se datos que engloba en su totalidad a las bases de datos de proteínas no redundantes como son Genbank CDS translations, PDB, SwissProt, PIR (Protein Information Resource) y PRF (Protein Research Foundation). La NR de proteínas que usamos excluye a las muestras ambientales de proyectos de secuenciación de genomas completos.
- NVT. Moles (N), Volumen y Temperatura. También llamado ensamble canónico. Es un protocolo de restricciones a usar durante una simulación de dinámica molecular en la cual el número de moléculas presente en el sistema, el volumen y la temperatura no cambiarán a lo largo del tiempo.
- PDB. Protein Data Bank. Es una base de datos de proteínas cuya estructura tridimensional ha sido elucidada. A lo largo del texto también puede referirse a un identificador propio de esta base de datos.
- ps. Picosegundos. Un picosegundo equivale a $1x10^{-12}$ segundos.
- Rotámero. Isómero Rotacional. Es la conformación de una cadena lateral de un aminoácido representado como un conjunto de valores, uno por cada grado de libertad del ángulo diedro.
- SCA. Statistical Coupling Analysis. Es un método estadístico que analiza las posiciones de residuos en un alineamiento múltiple de secuencias de proteína. A lo largo del texto también es usado en algunas ocasiones para designar a una molécula o un conjunto de moléculas diseñadas donde se respetaron, además de los residuos muy conservados, a los residuos identificados como estadísticamente acoplados con este método.
- SCOP. Structural Classification Of Proteins. Es una base de datos de clasificación de proteínas.
- SDH. Shikimate Dehydrogenase. Es la enzima shikimato deshidrogenasa de *Escherichia coli*.

Índice

INTRODUCCIÓN

Las proteínas involucran una muy notable relación entre estructura y función a nivel molecular. Estas moléculas llevan a cabo diferentes tareas en bioquímica, las cuales son determinadas por su estructura tridimensional nativa, que a su vez está codificada en su cadena lineal compuesta de aminoácidos (Anfinsen, 1973; Dill & MacCallum, 2012). De hecho, las interacciones entre sus propios aminoácidos que favorecen la conformación nativa deben superar de forma colectiva las interacciones no nativas (Koga *et al.* 2012).

Las proteínas están compuestas de dominios, los cuales pueden definirse de diversas formas; para los fines de este proyecto un dominio se refiere a la parte de la proteína que tiene una historia evolutiva independiente con respecto al resto de la cadena polipeptídica en donde se encuentra (Ponting & Russell, 2002). Los dominios se clasifican según la base de datos de *Structural Classification Of Proteins* (SCOP) en un conjunto jerárquico de categorías que de forma adyacente son: familia (donde la identidad de secuencia por si sola implica una relación evolutiva), superfamilia (familias para las que la combinación de características estructurales y funcionales implica un origen evolutivo común, aunque el grado de identidad de secuencia sea menor al 30%), plegamiento (conjunto de superfamilias y/o familias que tienen las mismas estructuras secundarias principales en el mismo arreglo topológico) (Bashton & Chothia, 2002) y clase (composición similar de estructura secundaria y empacamiento) (Orengo & Thornton, 2005). Otra base de datos muy conocida para clasificar dominios es *Class Architecture Topology Homology* (CATH), en donde el nivel "Arquitectura" está basado en una inspección visual de la "Topología" (grupos de plegado similar) y referencias en la literatura, lo cual puede ser de particular importancia cuando se consideran nuevos plegamientos (Hadley & Jones, 1999). A menudo, los dominios constituyen unidades funcionales y de plegamiento independientes con más interacciones dentro de sí mismas que con cualquier otra parte de la proteína (Janin & Wodak, 1983). Las proteínas han evolucionado a partir de una amplia gama de duplicaciones de dominios e intercambio de los mismos; y es por la naturaleza modular de estos dominios que es posible el surgimiento de nuevas y a menudo complejas funciones a partir de la combinación de un número limitado de dominios (Han *et al.* 2007).

Una forma de estudiar la importancia de las combinaciones de dominios ha sido generando proteínas "quiméricas" por medio de ingeniería genética, lo cual ha tenido intereses tanto de investigación evolutiva como industrial (Banda; *Tesina de licenciatura,* 2011). Al realizar tales intercambios de dominios a nivel experimental se ha pretendido analizar por una parte qué tanto determinan éstos capacidades particulares en sus proteínas de origen, y qué tanto pueden combinarse tales dominios con otros y mantener esas capacidades. En algunas ocasiones se ha logrado el éxito en el plegamiento, aunque la eficiencia catalítica (en el caso de proteínas catalíticas) no siempre ha sido de lo más deseable (Kataoka *et al*, 1996; Thulasiram *et al*, 2007; Goihberg *et al*, 2010; Starkey *et al*, 2009; Segatori *et al*, 2004). Por esta razón, pensamos que un enfoque que tome en cuenta la coevolución entre sitios de una proteína multidominio sería de utilidad para entender qué regiones de un dominio necesitan regiones específicas del otro para poder combinarse.

El grupo del Dr. Lorenzo Segovia en el Instituto de Biotecnología, también estudia la importancia de las combinaciones de dominios en las enzimas, quienes son proteínas capaces de acelerar la velocidad con que suceden reacciones cruciales en los seres vivos.

La identificación de los aminoácidos críticos para un dominio es de suma importancia para comprender la forma en la que la información contenida en la secuencia actúa para dar lugar a un dominio correctamente plegado y funcional. Tal conocimiento puede ser aprovechado tecnológicamente para el diseño de proteínas (incluyendo enzimas) con propiedades biotecnológicas deseables, tanto para la industria, como para la biorremediación (Jez 2011).

En el presente trabajo (ver Diagrama 1), un método estadístico (SCA) fue utilizado para identificar los sitios que coevolucionan dentro del dominio Rossmann de la enzima shikimato deshidrogenasa de *Escherichia coli* y otras enzimas deshidrogenasas. Aquellos sitios cuya evolución es conjunta dentro del dominio Rossmann de la shikimato deshidrogenasa fueron respetados durante el rediseño de este dominio. Posteriormente, este rediseño fue analizado *in silico* en cuanto a su movimiento molecular, ésto para determinar si la información de los residuos estadísticamente acoplados era suficiente para que este dominio pudiera permanecer plegado y mantener unido al dinucleótido, en este caso NADP, en una dinámica similar a como lo hace el dominio silvestre.

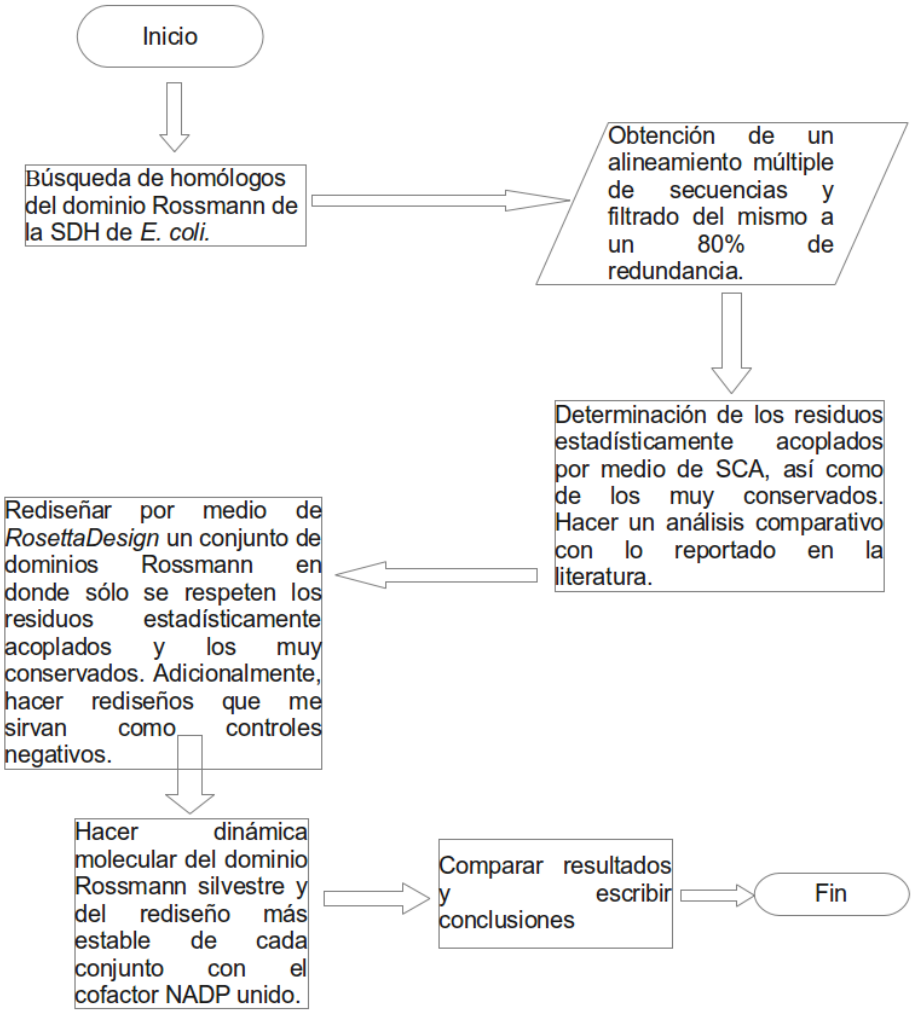
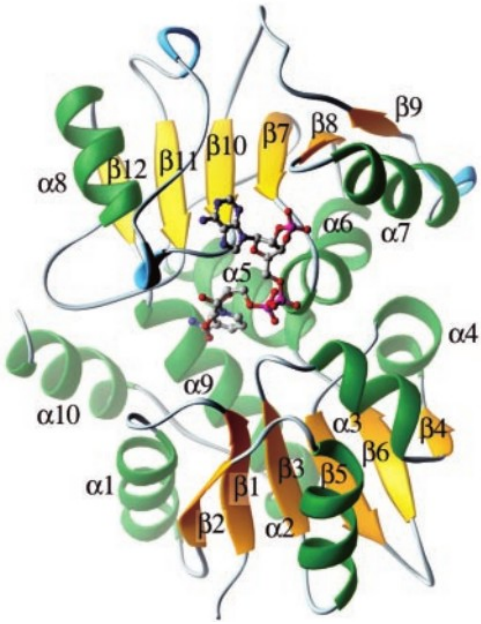


Diagrama 1. Estrategia que se utilizó durante el desarrollo del trabajo plasmado en esta tesis.

MODELO DE ESTUDIO: EL DOMINIO ROSSMANN DE UNIÓN A DINUCLEÓTIDOS

La enzima shikimato deshidrogenasa (E.C. 1.1.1.25) cataliza la reducción dependiente de NADP del 3-deshidroshikimato a shikimato. Tal reacción es el cuarto paso de la vía del shikimato, ruta esencial para la biosíntesis de aminoácidos aromáticos en plantas y microorganismos (Michel *et al.* 2003). En *E. coli* esta enzima, codificada por el gen *aroE,* se presenta de forma monomérica. Ésta molécula está compuesta por dos dominios (ver Esquema 1). A nivel de secuencia, el primer dominio está compuesto de dos segmentos discontinuos que, de acuerdo a la base de datos CATH, van de M1 a T101 y de A245 a S271; el segundo dominio (dominio Rossmann) comprende de D102 a Q244. A nivel tridimensional, ambos dominios tienen arquitectura α/β y están conectados por la hélice α5 y por el puente formado entre los residuos D102 y P118. El arreglo de estos dos dominios alrededor de las hélices conectoras da lugar a una profunda cavidad dentro de la cual se localiza el cofactor NADP (Michel *et al.* 2003).



Esquema 1. Estructura de la enzima Shikimato Deshidrogensa de *E. coli*. Se muestran en azul las estructuras α y en amarillo las estructuras β. El cofactor NADP se muestra en bastones. Imagen original de Michel *et al*. 2003

Al trabajar con la enzima shikimato deshidrogenasa (SDH) de *E. coli* y *Bacillus subtilis*, ha surgido el interés en profundizar en el estudio del dominio Rossmann, encargado de la unión de los cofactores dinucleótidos NAD y NADP. Este dominio es uno de los tres plegamientos más frecuentes en la base de datos del *Protein Data Bank* (PDB) y es también de los más poblados en α/β dentro de la misma (Bhattacharyya *et al.* 2012).

El plegamiento Rossmann une un mononucleótido y se compone de un motivo βαβαβ; es decir, las hélices α y las hebras β se van intercalando en la secuencia peptídica. Por lo tanto el dominio que une dinucléotidos como NAD/NADP (figura 1) involucra a dos motivos de unión a mononuclótidos relacionados por un eje de rotación, y es aquí donde los dos motivos βαβαβ forman una estructura de seis hebras β (donde la mayoría de las hebras N-terminales quedan adyacentes una con la otra) flanqueada por hélices (Bhattacharyya *et al.* 2012, Bottoms *et al.* 2002).

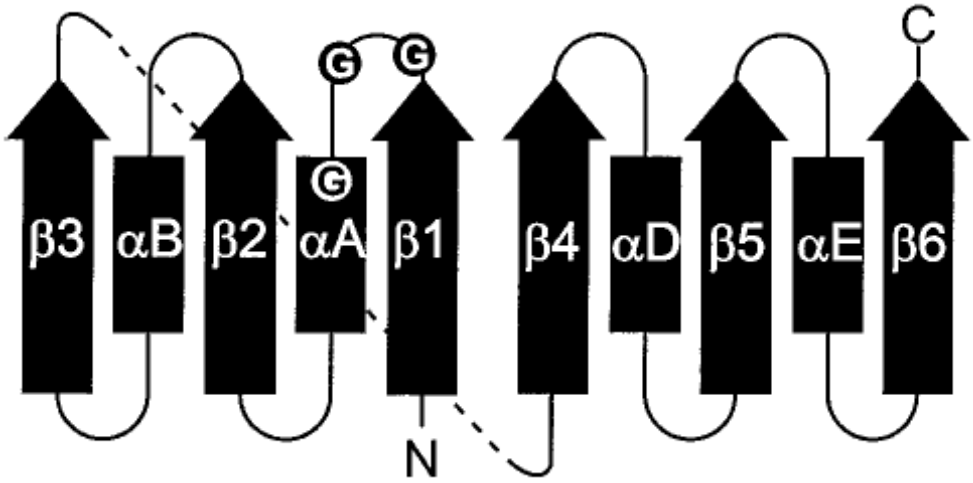


Figura 1. Topología clásica del dominio Rossmann. Las flechas son hebras β y los rectángulos α-hélices. Los círculos son glicinas conservadas. Imagen original de Bottoms *et al.* 2002

Se ha visto que la superfamilia de los dominios Rossmann se combina con un gran número de familias de dominios. De hecho, en el análisis realizado por Bashton & Chothia en el 2002, se determinó que los dominios catalíticos asociados al dominio Rossmann provienen de 17 familias de dominios diferentes que pertenecen a alguna de 7 superfamilias.

Debido a la versatilidad de este dominio nos interesó determinar los sitios esenciales para su plegamiento y función por métodos computacionales para en un futuro utilizar este conocimiento en el diseño de enzimas con propiedades deseables como novedosas (Röthlisberger *et al*, 2008).

Para la identificación de sitios importantes en el dominio Rossmann, utilizamos el análisis de acoplamiento estadístico (SCA), el cual es un método que se basa en determinar el nivel de covariación entre sitios de residuos en una familia de proteínas, ésto a partir de la información evolutiva plasmada en un alineamiento múltiple de secuencias (Lockless & Ranganathan, 1999). SCA ha demostrado poder de predicción en numerosos reportes (Bhattacharyya *et al*. 2012, Halabi *et al.* 2009, Lockless & Ranganathan 1999, McLaughlin *et al*. 2012, Reynolds *et* al. 2011, Russ *et al.* 2005, Smock *et al*. 2010, por mencionar algunos), sin embargo, dado que el dominio Rossmann puede presentarse en diferentes geometrías de acuerdo a Bashton & Chothia en el 2002, decidimos utilizar un miembro representativo de cada geometría para posteriormente contrastar los sitios identificados como estadísticamente acoplados en cada caso.

En el presente trabajo se realizaron análisis de SCA de las siguientes enzimas:

- La enzima lactato deshidrogenasa, quien cataliza la interconversión de lactato y piruvato en la vía glicolítica (Abad-Zapatero *et al.* 1987).
- La fenilalanina deshidrogenasa, es la responsable de la desaminación oxidativa de la L-fenilalanina para producir amonio y fenilpiruvato por medio de la reducción de NAD (Vanhook *et al.* 1999).

- La alanina deshidrogenasa, quien se encarga de catalizar la aminación reductiva del piruvato hacia L-alanina, utilizando como cofactor NADH (Baker *et al.* 1998).
- La enzima shikimato deshidrogenasa, quien como ya se mencionó, cataliza la reducción dependiente de NADP del 3-deshidroshikimato a shikimato.
- La enzima gliceraldehído 3-fosfato deshidrogenasa, es la catalizadora de la formación dependiente de NAD de glicerato 1,3-bifosfato a partir de gliceraldehído 3-fosfato (Lambeir *et al*. 1991).

Posteriormente se hicieron diseños *in silico* del dominio Rossmann de la SDH y algunos de ellos fueron sometidos a dinámica molecular para poner a prueba su estabilidad y capacidad de unión al ligando NADP.

SOBRE EL FUNCIONAMIENTO DE LOS MÉTODOS COMPUTACIONALES UTILIZADOS EN ESTE TRABAJO

SCA

Como ya se ha mencionado antes, la secuencia de aminoácidos de una proteína está muy relacionada a su estructura tridimensional y su función biológica (Anfinsen, 1973). A partir de esto surge el análisis de acoplamiento estadístico (SCA por *Statistical Coupling Analysis*), el cual es un método cuantitativo para entender el contenido informativo de una secuencia peptídica, ésto a través de una generalización del principio de la conservación evolutiva (McLaughlin Jr *et al.* 2012). La premisa medular del método es que, el patrón de acoplamientos energéticos entre residuos de una proteína (es decir, las restricciones funcionales entre aminoácidos) pueden ser identificados a través de un análisis de coevolución de residuos en sus respectivas posiciones en una familia de secuencias homólogas (Lockless, *et al*. 1999). Así, la principal conclusión del SCA es que sólo alrededor del 20% de los aminoácidos de una proteína están organizados en redes físicamente contiguas de posiciones que coevolucionan llamadas sectores de proteína (Halabi *et al.* 2009, Smock *et al.* 2010, Reynolds *et al.* 2011, McLaughlin Jr. *et al.* 2012), mientras que el resto de los residuos evolucionan casi independientemente, sin mucha influencia ni siquiera de su ambiente inmediato estructural. Los sectores de proteína se encuentran por lo general alrededor de los sitios activos de las proteínas, pero conectan residuos importantes de la superficie a través de caminos de interacción entre residuos con el núcleo de la proteína (Reynolds *et al.* 2011).

Dado que tales sectores se han encontrado en todas las familias de proteínas estudiadas hasta ahora y están relacionados con el plegamiento, estabilidad y/o actividades funcionales, se ha sugerido que esta característica estructural es una propiedad general de las proteínas (Halabi *et al.* 2009 y Smock *et al.* 2010).

SCA examina la conservación conjunta de todos los pares de posiciones de aminoácidos en una familia de proteínas, lo cual depende en un principio de un gran y diverso alineamiento múltiple de secuencias “MSA” por *Multiple Sequence Alignment* de una familia de proteínas compuesto de un número *M* de secuencias por un número *L* de posiciones, a partir del cual se calcula una matriz de peso de correlación de *LxL* (a quien llamaremos $\hat{C}$, la matriz SCA) que describe la coevolución de todos los pares de posiciones de secuencia. Cuando las secuencias son homogéneamente diversas a tal grado que las distintas subfamilias de origen no son evidentes en el alineamiento, la definición de los sectores alcanza a identificar posiciones que se agrupan en los eigenmodos (conjuntos con cierta independencia) más significativos de la matriz de correlación de SCA. Sin embargo, también es posible que distintas subfamilias sean identificadas en MSA heterogéneos gracias a métodos matemáticos que proveen un mapeo directo entre patrones de divergencia de secuencia y patrones de covariación posicional (Smock *et al.* 2010). A continuación se describirán los pasos claves del algoritmo SCA.

Paso 1. Definición de las matrices de alineamiento y correlación

En general, un MSA puede describirse como una matriz binaria de tres dimensiones X^{a}_{si} (compuesta por *M* x *L* x 20 aminoácidos) cuyos elementos son 1 si la secuencia contiene al aminoácido *a* en la posición *i* y 0 si no es así (figura 2A). Para usar métodos matemáticos

posteriores es necesario hacer una aproximación binaria de la matriz anterior, lo cual se logra reduciendo el MSA a una matriz binaria de dos dimensiones *M* x *L* (representada por X_{si}) quien ahora sólo incluye los términos en X^{a}_{si} que representan al aminoácido más prevalente en cada posición. Posteriormente se calcula un alineamiento normalizado de peso.

$$\ddot{X}_{si}=\phi_i(X_{si}-\check{R}_{si})$$

donde ϕ_i está relacionada con la conservación de la posición *i* en el MSA y es la función de peso usada en la última implementación de SCA (figura 2B), y $\check{R}_{si}$ representa al promedio de X_{si} sobre todas las secuencias. A partir de la matriz $\ddot{X}$, se pueden obtener dos matrices de correlación, $\hat{S}=(1/L)\ddot{X}\ddot{X}^T$, que es la matriz de correlación de secuencias sopesada por ϕ_i, y $\hat{C}=(1/M)\ddot{X}^T\ddot{X}$, la matriz de correlación de posiciones sopesada por ϕ_i (propiamente, $\hat{C}$ es la matriz SCA).

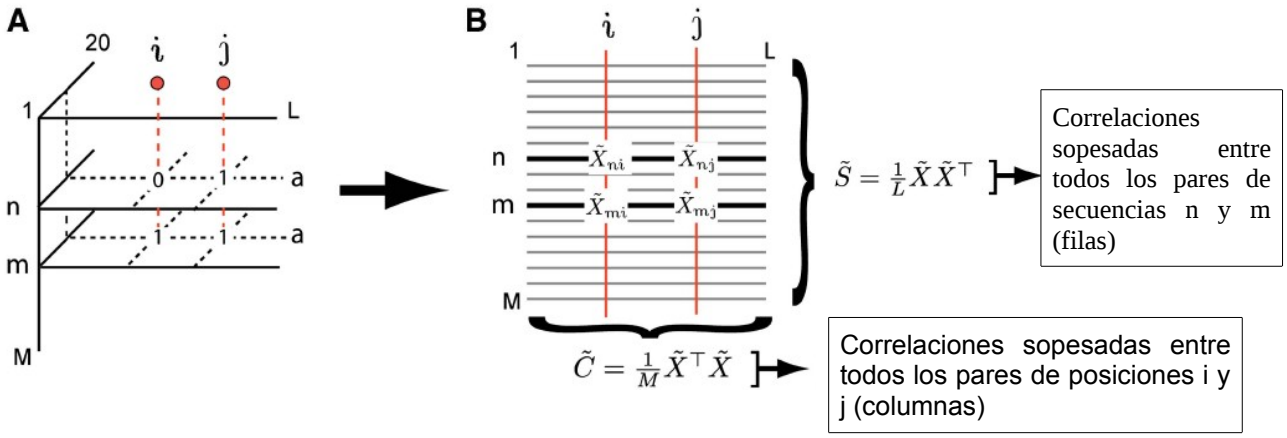


Figura 2. Construcción de las matrices de correlación en SCA. En A) se determina la presencia o ausencia de un aminoácido particular por secuencia y posición particulares, esta matriz se binariza en B) representando ahora al aminoácido más prevalente en una secuencia y posición particulares. Figura adaptada de Smock *et al*. 2010.

Paso 2. Mapeo de los modos de covariación de secuencia y covariación de posición.

Se puede relacionar la divergencia de subfamilias de secuencias a la evolución correlacionada de grupos de posiciones por medio de la "descomposición en valores singulares" (*singular value decomposition*). En este método, la matriz *M* x L binaria $\ddot{X}$ puede escribirse como un producto de tres matrices: $\ddot{X}=U\Sigma V^T$, donde *U* es una matriz *M* x *M* cuyas columnas contienen a los eigenvectores de $\hat{S}$, quien es la matriz de correlación de secuencia, y *V* es una matriz *L* x *L* cuyas columnas contienen a los eigenvectores de $\hat{C}$, quien es la matriz SCA de correlación posicional. Σ es una matriz diagonal *M* x *L* de "valores singulares", los cuales están relacionados con los eigenvalores de las matrices $\hat{S}$ y $\hat{C}$.

$$\tilde{X} = U \times \Sigma \times V^{\mathrm{T}}$$

$$\begin{bmatrix} \tilde{x}_{11} & \dots & \tilde{x}_{1L} \\ \tilde{x}_{21} & \dots & \tilde{x}_{2L} \\ \vdots & & \vdots \\ \tilde{x}_{M1} & \dots & \tilde{x}_{ML} \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1M} \\ u_{21} & u_{22} & \dots & u_{2M} \\ \vdots & & \ddots & \\ u_{M1} & & & u_{MM} \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \cdots & 0 \\ 0 & \ddots & 0 \\ \vdots & & \Sigma_{LL} \\ 0 & & 0 \end{bmatrix} \begin{bmatrix} v_{11} & \cdots & v_{L1} \\ \vdots & \ddots & \\ v_{1L} & & v_{LL} \end{bmatrix}$$

Si un eigenmodo de la matriz de correlación de secuencias (una columna en *U*) revela una separación de dos clases de secuencias, entonces el eigenmodo correspondiente de la matriz de correlación de posiciones (una columna en *V*) revelará las posiciones que más contribuyen a esta divergencia de secuencias.

Si se desea representar de una mejor manera la divergencia de subfamilias, SCA permite el uso del análisis de componentes independientes (ICA por sus siglas en inglés), el cual está específicamente diseñado para transformar *K* eigenmodos principales (los principales son los más informativos) de una matriz de correlación en *K* componentes máximamente independientes (Smock *et al.* 2010). En los tutoriales del algoritmo donde se aplica esta transformación se definen 3 sectores, uno por cada uno de los 3 primeros componentes independientes.

Al final, para definir los residuos estadísticamente acoplados arriba de un punto de corte (0.8 o superior según los ejemplos de los tutoriales del algoritmo, lo cual refleja lo ya mencionado, con respecto a que alrededor 20% de los residuos de una proteína coevolucionan de forma crítica) el mejor modelo estadístico a usar, en el caso de los datos separados en *K* componentes máximamente independientes por ICA, es la distribución *t* de Student. En caso de únicamente definir un sector, tal grupo es analizado con una distribución Log-normal, a partir del primer eigenmodo de la matriz de correlación de posiciones.

Un aspecto importante a entender es por qué el diseño natural de las proteínas debiera parecer una arquitectura de sectores, el cual es un conjunto de redes dispersas de residuos que actúan de forma cooperativa sumergidos en un ambiente de aminoácidos débilmente acoplados (McLaughlin Jr *et al.* 2012) y que pueden elucidarse por SCA. Como han propuesto McLaughlin Jr y sus colaboradores en el 2012: el sector es una consecuencia natural de las restricciones evolutivas que rara vez son consideradas en la ingeniería de proteínas y los modelos biofísicos. Tales sectores surgen principalmente de la necesidad de variación adaptativa en respuesta a condiciones fluctuantes de la selección natural.

Así, al poner las restricciones del plegamiento nativo y la función en las posiciones que conforman a los sectores, esta arquitectura permitiría la capacidad de una variación adaptativa rápida a través de mutaciones de unos cuantos residuos que actúen cooperando entre sí. De ser así, la gran mayoría de las posiciones que no forman parte de ningún sector, a pesar de su localización estructural, debieran mostrar mucha más tolerancia a las mutaciones y un menor potencial adaptativo.

ROSETTADESIGN

Este algoritmo forma parte de la herramienta bioinformática *Protein Modeling Rosetta Suite*, y conlleva un proceso iterativo que optimiza energéticamente tanto la estructura como la secuencia de una proteína. RosettaDesign alterna entre rondas de optimización de secuencia de esqueletos (*backbone*) rígidos y minimización energética de esqueletos flexibles. Durante el paso de la optimización de secuencia se usa un algoritmo de búsqueda por "realineamiento simulado" (*Simulated Annealing)* por medio del método estadístico numérico Monte Carlo para muestrear el "espacio" (conjunto de aminoácidos con respectivos rotámeros posibles) de secuencia. Cada aminoácido es considerado en cada posición y los rotámeros son aquellos que forman parte de la libreria de Dunbrack (Dunbrack & Karplus, 1993), la cual presenta conformaciones o frecuencias de rotámeros dependientes de la conformación local del esqueleto. Después de cada ronda de optimización de secuencia según Monte Carlo, el esqueleto es relajado para acomodar a los aminoácidos del diseño. El algoritmo de RosettaDesign puede utilizarse para el diseño de nuevos plegamientos, rediseño de proteínas ya existentes, diseño de interfaces de proteínas, diseño de enzimas, y la predicción de regiones formadoras de fibras en proteínas (Kaufmann *et al.* 2010).

DINÁMICA MOLECULAR

La dinámica de una proteína se define como cualquier cambio en las coordenadas atómicas dependiente del tiempo (Henzler-Wildman & Kern 2007). Aunque los subestados conformacionales de las proteínas y sus tasas de interconversión pueden ser detectados experimentalmente, una descripción estructural con resolución atómica del cambio de un subestado a otro está fuera del alcance experimental. Esta limitación es razonablemente superada por métodos computacionales dependiendo de qué tan buena sea la descripción del sistema proteína-solvente de acuerdo al campo de fuerza utilizado, el cual es un conjunto de parámetros que describe el potencial de energía de todos los átomos (Henzler-Wildman & Kern 2007).

ANTECEDENTES

Como ya se ha mencionado antes, la SDH de *E. coli* es codificada por el gen *aroE.* Se sabe que la estructura 3D de esta enzima contiene al dominio Rossmann insertado en el dominio catalítico, pues las últimas dos α-hélices del C-terminal se encuentran muy cerca del primer dominio (catalítico). Esto concuerda con análisis previos y aún no publicados de SCA en la SDH realizados en el laboratorio del Dr. Segovia, ya que la SDH muestra conjuntos de residuos acoplados involucrados mayoritariamente en el reconocimiento del sustrato en el dominio catalítico, y otros pocos que, aunque se encuentran en el dominio C-terminal (precisamente en las últimas dos α-hélices antes mencionadas), interactúan físicamente con el dominio N-terminal, tal vez para brindarle una mejor estabilidad y/o empacamiento a la proteína en su conjunto (Banda; *Tesina de Licenciatura*, 2011). El conocimiento de estos residuos importantes en el extremo C-terminal fue determinante para el diseño de quimeras en las que se intercambiaron los dominios Rossmann entre proteínas homólogas (Banda; *Tesina de Licenciatura*, 2011).

Las últimas versiones del algoritmo de SCA son capaces, no sólo de localizar residuos estadísticamente acoplados, sino de determinar sectores de residuos dentro de una proteína en cuestión. Se ha visto que cada uno de estos sectores puede tener un papel funcional distinto de los demás e incluso un modo independiente de divergencia en secuencia en una familia proteica (Halabi *et al.* 2009); así como de delatar a los residuos responsables de efectos alostéricos (Smock *et al.* 2010 & Reynolds *et al.* 2011).

Actualmente, la información del SCA está siendo utilizada en el grupo del Dr. Segovia, para rediseñar el núcleo de plegamiento de la proteína fosforibosil antranilato isomerasa (PRAI) por medio de RosettaDesign.

Por otra parte, el dominio Rossmann de la shikimato deshidrogenasa de *E. coli* fue clonado de forma aislada en un plásmido, y de acuerdo a análisis experimentales de nuestro laboratorio, es capaz de plegarse y presentarse de forma soluble dentro de *E. coli* (Chenge; *Tesis de Licenciatura*, 2010).

JUSTIFICACIÓN

La identificación por métodos *in silico* de los sitios más importantes para un dominio contribuiría de manera importante a la rápida comprensión de la información contenida en la secuencia de aminoácidos previo a los análisis experimentales. Además de ayudarnos al mejor entendimiento de la evolución de los dominios y las proteínas, este aprendizaje puede ser aplicado al diseño con un importante impacto biotecnológico, farmacológico, la industria (Kiss *et al*. 2010) o la biorremediación (Jez 2011).

Basándonos en la idea de McLauglin Jr y sus colaboradores en 2012 de que la información obtenida a partir del SCA coincide de forma casi absoluta con la de los sitios “mutacionalmente significativos”, el hecho de que RosettaDesign es una buena herramienta de diseño (Kaufmann *et al.* 2010) y de que las funciones de las proteínas se gobiernan por sus características dinámicas (Henzler-Wildman & Kern 2007) propusimos lo siguiente:

HIPÓTESIS

Es posible rediseñar dominios Rossmann estables y capaces de unir a un dinucleótido al utilizar la información de los residuos estadísticamente acoplados y los muy conservados.

OBJETIVO GENERAL

Sentar las bases para el diseño de dominios Rossmann de unión a dinucleótidos estables.

Objetivos particulares:

1. Determinar los sectores de residuos estadísticamente acoplados en proteínas completas que utilicen el dominio Rossmann usando como modelo a la SDH de *E. coli*.
2. Determinar los sectores de residuos estadísticamente acoplados en el dominio Rossmann cuando está asociado a dominios distintos en diferentes geometrías.
3. Diseñar dominios Rossmann en donde a nivel de secuencia:
    a) Se varíen todos los residuos de un dominio Rossmann original excepto los estadísticamente acoplados.
    b) Se varíen sólo los residuos estadísticamente acoplados.
    c) Se varíen todos los residuos.
4. Analizar la estabilidad y funcionalidad de las dinámicas moleculares *in silico* de algunos de estos diseños por medio de la paquetería Maestro 9.3.

MÉTODOS

Obtención de los alineamientos múltiples

Buscando con la herramienta en línea de *Protein Basic Local Alignment Search Tool*, mejor conocido por sus siglas*:* BLASTp (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome), versión 2.2.25+, con un valor E máximo de 0.001 y la base de datos NR de secuencias de proteínas, actualizada al día 27 de Agosto del 2011 a las 2:12 pm, se obtuvieron dos alineamientos locales para cada proteína analizada: Uno utilizando como secuencia de búsqueda a la secuencia completa de la proteína en cuestión y otro donde la secuencia de búsqueda fue únicamente el segmento correspondiente al dominio Rossmann de cada caso de acuerdo a la base de datos para la clasificación de estructuras de proteínas CATH para el día 29 de Agosto de 2011.

Procesamiento de los alineamientos:

Por medio de la versión 4.1 de la herramienta bioinformática CD-HIT (http://weizhong-lab.ucsd.edu/cd-hit/) se hizo un filtrado de cada alineamiento para quedarse finalmente con las secuencias que mantenían un máximo de 80% de identidad entre si y una longitud en residuos mínima del 80% de la longitud de la secuencia a analizar para reducir la proporción de gaps y asegurar que el cómputo con SCA 5.0 contemplara la mayor cantidad posible de sitios.

Análisis de acoplamiento estadístico (SCA):

Se alimentó con cada alineamiento y el archivo de las coordenadas cristalográficas (archivo pdb) correspondiente a la secuencia de búsqueda al algoritmo de SCA 5.0 en la paquetería MATLAB R2010B. Es importante mencionar que la estructura cristalográfica es usada en el algoritmo únicamente para obtener la numeración de los sitios, ya que en ningún punto se toma en cuenta la información tridimensional para realizar los cálculos. Así pues, únicamente a partir del alineamiento, el algoritmo SCA despliega diferentes gráficos en cada uno de sus pasos, a partir de los cuales el usuario puede (si lo considera pertinente) hacer análisis opcionales de acuerdo a la información mostrada en pasos previos.

Diseños con RosettaDesign

Utilizando el algoritmo fixbb de la herramienta RosettaDesign, quien forma parte de la suite de Rosetta versión 3.3, se generaron 3 conjuntos (SCA, con 40 residuos respetados; AntiSCA, con 40 residuos a alterar, y Aleatoria, donde se permitió cambiar toda la secuencia del dominio Rossmann; para mayores detalles sobre la construcción de estos conjuntos ver *Diseños con RosettaDesign* en la sección de **Resultados**) de 999 secuencias cada uno,

donde siempre y cuando se mantuviera una estructura equivalente al esqueleto del dominio Rossmann (residuos 102 al 244) de la SDH. Los cambios en secuencia con respecto al templado siempre fueron simultáneos. En lo sitios donde se permitió la variación no se excluyó la posibilidad de incluir al aminoácido original. A la secuencia del diseño con el score más significativo de cada conjunto (menor energía) se le añadió el cofactor NADP en la misma posición y con la misma orientación atómica que en el cristal de la SDH de *E. coli* (PDB ID: 1NYT) ya que los esqueletos de los diseños se encuentran en exactamente las mismas coordenadas que el cristal templado.

Dinámicas Moleculares

Se hizo uso del programa Desmond de la paquetería Maestro 9.3 para realizar y analizar las dinámicas moleculares de los diseños ya mencionados. Se utilizó un ambiente de agua SPC con suficientes iones Na^+ para mantener un entorno neutro bajo el protocolo *Simulated Annealing* con 6 etapas iniciales (la primera fue a 10 K de temperatura durante un tiempo de 30 ps, la segunda de 100 K durante 100 ps, la tercera de 300 K durante 200 ps, la cuarta de 400 K por 300 ps, la quinta de 400 K por 500 ps y la sexta de 300 K durante 1000 ps) para finalmente mantener una temperatura constante de 300 K durante 50 nanosegundos. Se capturaron los datos cada 1.2 picosegundos. El campo de fuerza utilizado fue el OPLS-2005. Se hicieron las dinámicas con dos clases de ensamble por separado, la NPT (con una presión isobárica de 1.01325 bar) y la NVT. En todos los casos se incluyó al cofactor NADP colocado exactamente en el mismo lugar y en la misma orientación atómica de acuerdo al cristal del caso silvestre (PDB ID: 1NYT).

RESULTADOS

Determinación de los sectores de residuos estadísticamente acoplados en proteínas completas que utilizan el dominio Rossmann usando como modelo a la SDH de E. coli

En el caso de usar a la SDH completa de *E. coli* como secuencia de búsqueda, después del tratamiento mencionado en la sección de **MÉTODOS**, el alineamiento se redujo de 4280 a un total 1637 secuencias finales. De acuerdo al análisis de clustering jerárquico (figura 3) se infiere que solamente hay un sector, el cual comprende a varios residuos considerados significativos (con un punto de corte de 0.8 de acuerdo al ajuste con la distribución Log-normal, por lo que sólo se tomó en cuenta al 20% más significativo).

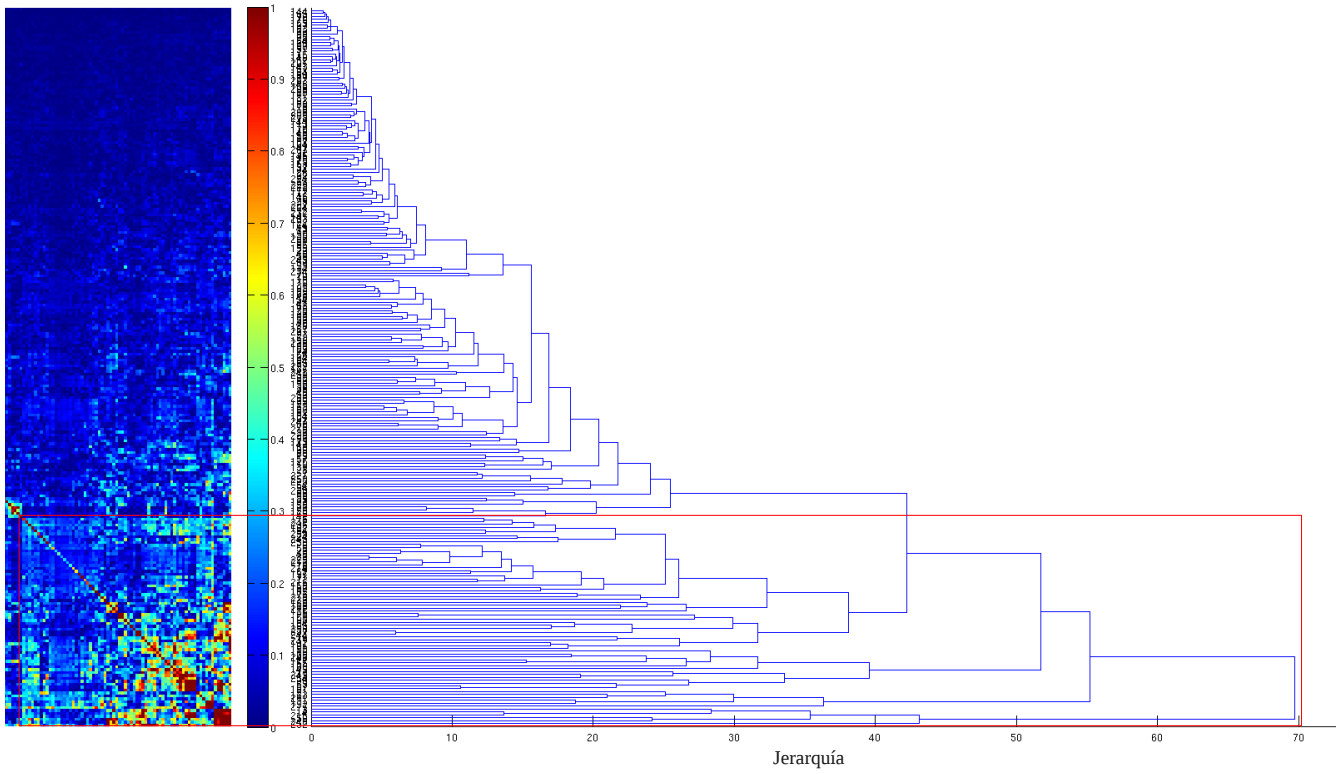


Figura 3. Clustering jerárquico. Izquierda. La significancia del acoplamiento estadístico aumenta del azul al rojo (en el rectángulo rojo se muestra la región correspondiente al cluster que da lugar al sector). Derecha. Dendograma de la matriz de la izquierda involucrando a todos los sitios de la SDH completa de *E. coli* (eje Y). El eje X del dendograma es una medida de la jerarquía entre grupos.

El dendograma de la figura 3 refleja la forma en la que la coevolución es capaz de descomponerse en jerarquías que relacionan a los diversos sitios y grupos de sitios con otros por medio del acoplamiento de pares de posiciones. A este nivel es posible apreciar la cooperatividad entre aminoácidos de la proteína particular para algún papel específico.

Una vez establecido que sólo habría un sector de acuerdo a lo que se visualiza en el rectángulo rojo de la figura 3, el primer eigenmodo (figura 4) revelará cuáles son las posiciones de la proteína que presentan la coevolución más significativa de acuerdo a un punto de corte.

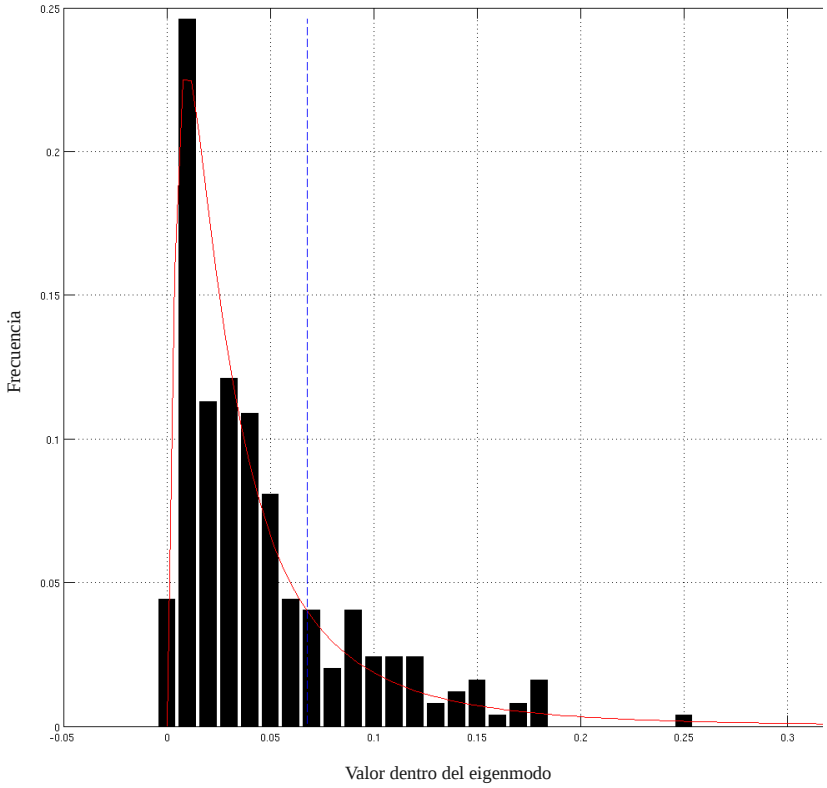


Figura 4. Definición del sector en SDH de *E. coli*. Una distribución Log-normal ajustada (línea roja) a un histograma del primer eigenmodo (donde las combinaciones entre residuos ya están ponderadas, lo que se grafica es justo el valor del peso) revela que a un punto de corte de 0.8 los datos más significativos se encuentran a la derecha de la línea vertical azul punteada.

Los residuos identificados como significativos de acuerdo a la figura anterior se visualizan en la figura 5 en color azul a lo largo de la estructura de la SDH (PDB ID: 1NYT, cadena A), atrapando principalmente a los residuos que apuntan a la cavidad de la proteína o están en el núcleo del dominio catalítico.

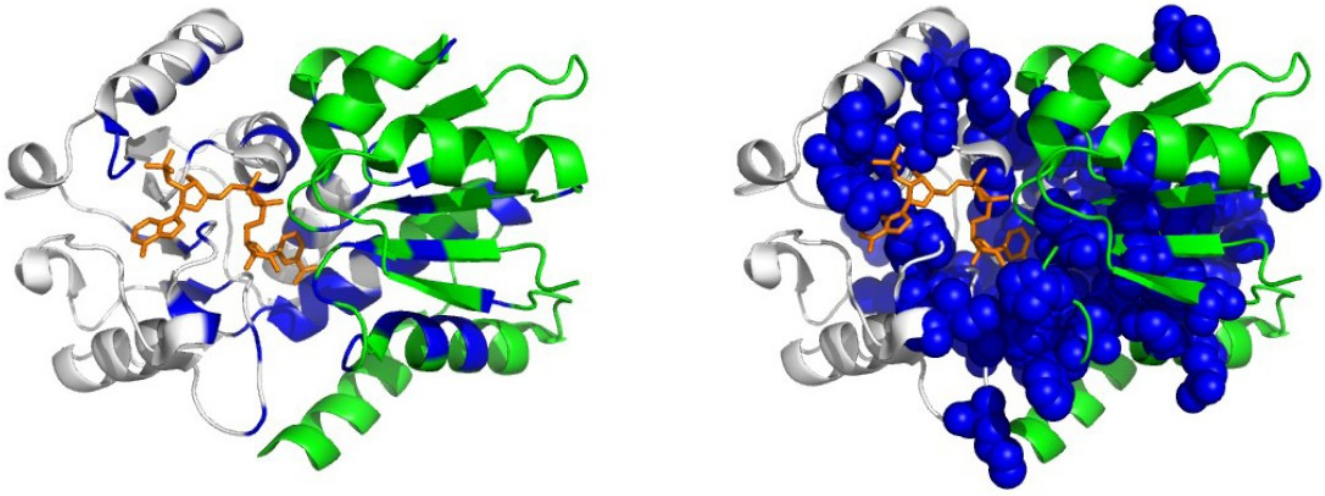


Figura 5. Residuos destacados a un punto de corte de 0.8 según SCA en la SDH completa de *E. coli*. Del lado izquierdo, se muestran los residuos acoplados en color azul y, en el lado derecho los residuos acoplados como esferas azules, ésto para enfatizar el alcance de las nubes electrónicas de los mismos por medio de sus cadenas laterales. En ambos casos el dominio Rossmann se muestra de color gris, el dominio catalítico en verde y el cofactor NADP se muestra en bastones naranjas.

Dado que la cantidad de residuos acoplados considerados a un punto de corte de 0.8 ocupa un espacio tal de la proteína que impide su adecuado análisis visual, se consideró que subir el punto de corte podía facilitar tal análisis. Puntos de corte más estrictos son mostrados en la figura 6, donde aquellos residuos que sobreviven a un punto de corte de 0.95 comprende a quienes tienen contacto con el cofactor NADP y a aquellos que forman parte del núcleo del dominio catalítico.

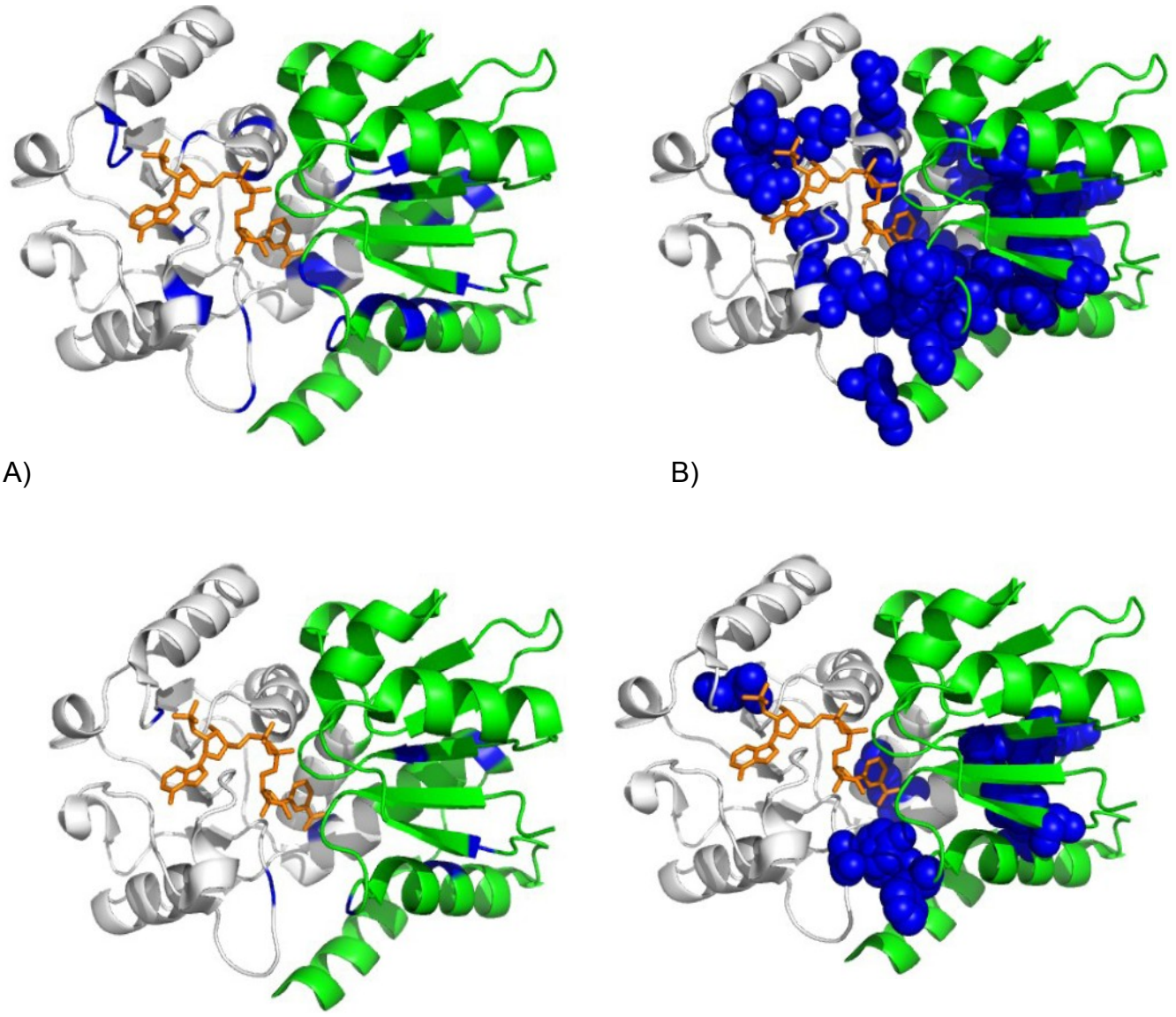


Figura 6. Residuos destacados según SCA en la SDH de *E. coli*. En gris el dominio Rossmann y en verde el dominio catalítico; en azul los residuos estadísticamente acoplados y en bastones naranjas el cofactor NADP. A) Residuos que corresponden al 10% más significativo. B) Residuos que pertenecen al 5% más significativo. En ambos casos se muestran del lado izquierdo los residuos significativos resaltados sólo de color azul, mientras que en el lado derecho se resaltan en azul y esferas, ésto para apreciar el alcance de las nubes electrónicas.

Por otra parte, presentar en forma de red cómo es que se acoplan los sitios que sobreviven a un punto de corte tan estricto como el 5% puede ser muy útil, como es el caso de la figura 7, donde es apreciable que puede haber coevolución entre residuos de distinto dominio sin importar la distancia entre ellos, como es el caso de N149 y W252.

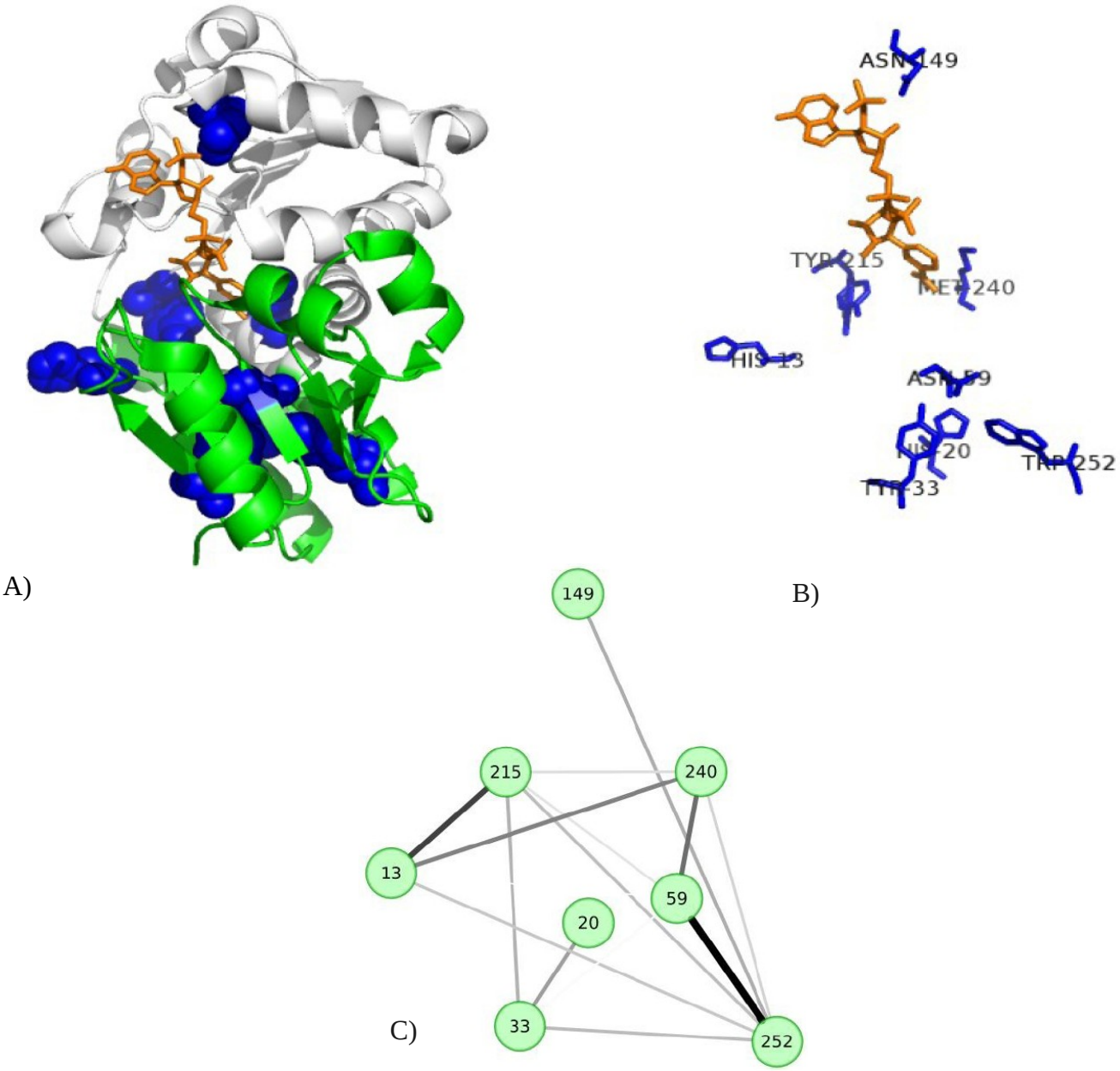
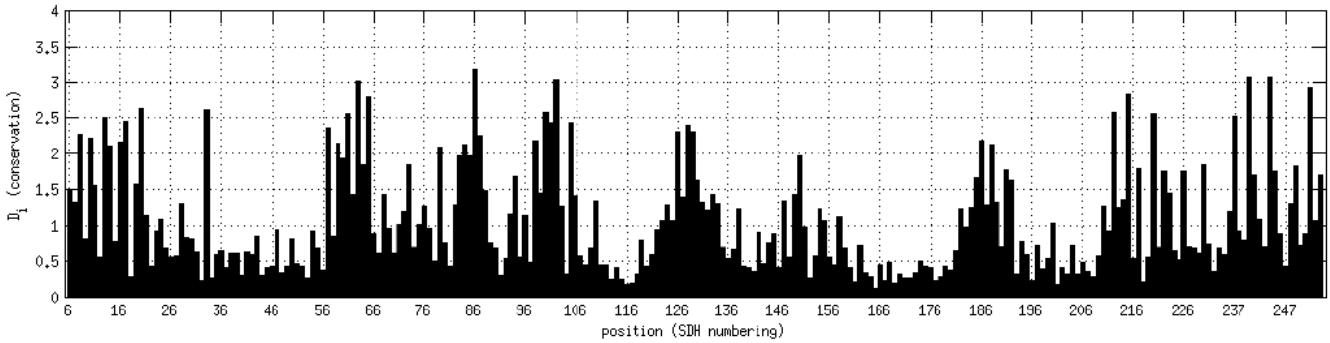


Figura 7. Residuos más acoplados. Otras vistas de la SDH donde se resaltan en A) los 8 residuos que se mantienen como significativos a un punto de corte del 95%, en B) únicamente de ellos y el cofactor. En C), las mismas posiciones en forma de red y acomodados en una topología similar a la de las otras imágenes de este apartado. En este último caso, una línea de conectividad es más gruesa y oscura entre mayor sea el grado de acoplamiento entre dos sitios.

Por otra parte, consideramos a los residuos muy conservados según el alineamiento como aquellos que según SCA tienen un valor de entropía relativa de Kullback-Leibler superior a 2.5. Tal entropía es calculada en los primeros pasos de SCA. Esta entropía captura la divergencia entre las frecuencias observadas de aminoácidos por posición en el alineamiento múltiple con respecto a las frecuencias de fondo en la base de datos no redundante de proteínas. En otras palabras, la entropía relativa de Kullback-Leibler es una medida estadística de qué tan inverosímil es que cierto aminoácido observado se presente en una posición particular (frecuencia en el alineamiento) dada la probabilidad con que ese aminoácido aparecería de acuerdo a una distribución ya conocida (frecuencia en la base de datos no redundante). La conservación de todos los residuos considerados está plasmada en la figura 8.

A)



B)

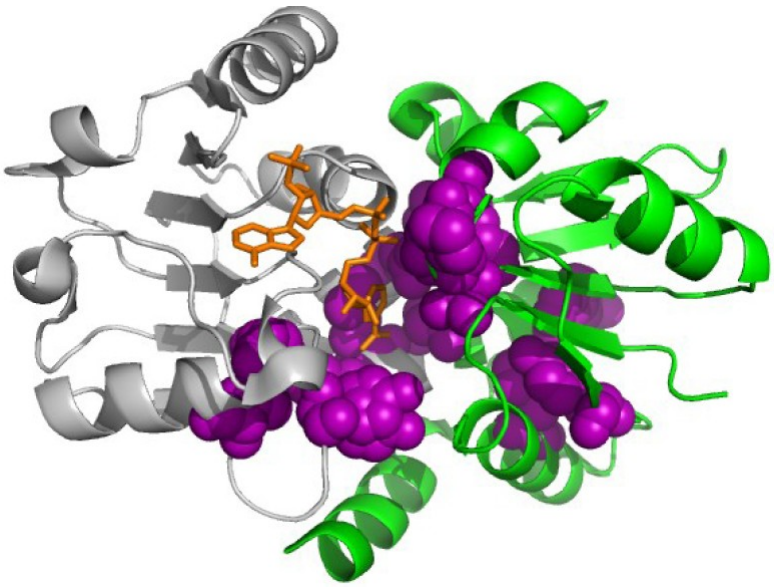


Figura 8. Residuos muy conservados en la SDH completa. A) Conservación por residuo de la SDH completa. B) Los residuos con una entropía relativa superior a 2.5 en A) se muestran en esferas moradas. En verde el dominio catalítico, en gris el dominio Rossmann y en sticks naranjas el cofactor NADP.

De acuerdo a la figura 8B, los residuos más conservados tienden a agruparse mayoritariamente en la región intermedia de la enzima con un sesgo hacia el sitio catalítico en la zona que plausiblemente une al shikimato; adicionalmente hay varios residuos conservados en la región que contactaría con el anillo nicotinamida del NADP.

Ahora bien, ya se mencionó antes que es posible separar en sectores la información de la covariación entre residuos (ver la parte final del apartado de *SCA* en *Sobre el funcionamiento de los métodos computacionales utilizados en este trabajo*) de una proteína. Por lo tanto, se hizo el ejercicio de determinar si el análisis se modifica sustancialmente si se asumían tres sectores (ya que las publicaciones que hemos revisado publican uno o tres sectores, lo cual es arbitrario, pero es lo que se ha manejado), por lo que se ignoró el hecho de que el clustering jerárquico de la figura 3 sugiriera la existencia de un único sector. Los eigenmodos principales (9 en total) de acuerdo a la figura 9 fueron sometidos a ICA (ver la parte final del apartado de *SCA* en *Sobre el funcionamiento de los métodos computacionales utilizados en este trabajo*) para posteriormente obtener un sector por cada uno de los primeros tres componentes independientes (figura 10).

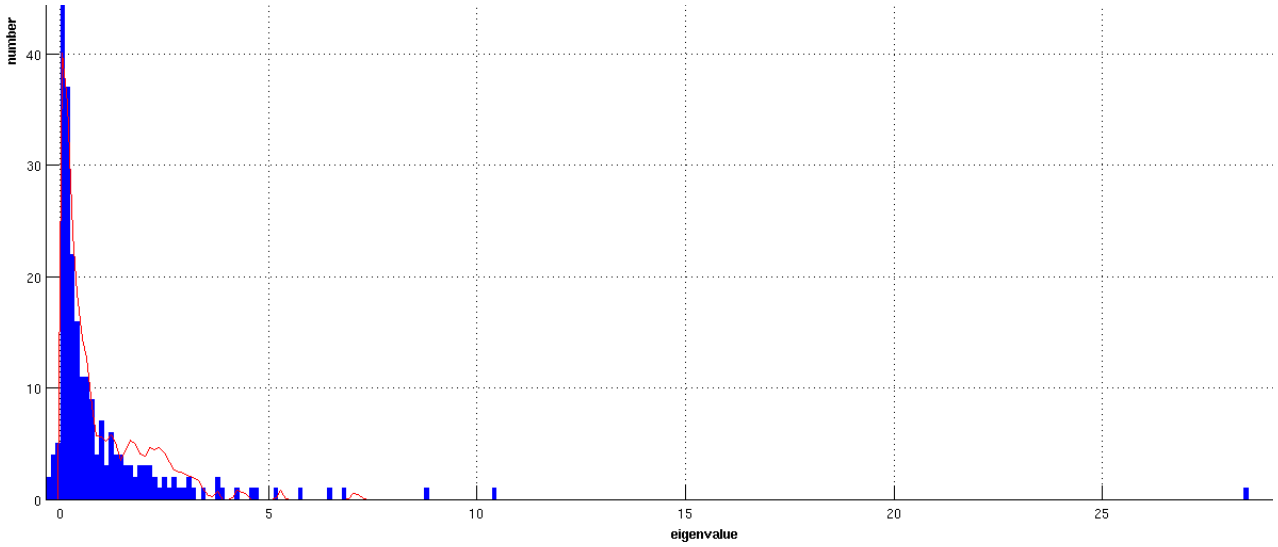


Figura 9. Eigenvalores de la matriz de correlaciones de SDH de *E. coli*. Un eigenvalor es más informativo entre mayor sea su posición en el eje X. En azul se muestra el histograma de los eigenvalores originales y en rojo la línea espectral, quien representa a un histograma de eigenvalores determinados a partir de las matrices de correlación de 100 aleatorizaciones del alineamiento original. Según los tutoriales del algoritmo, aquellos eigenvalores que no sobrelapen con la línea espectral son estadísticamente significativos. Se consideró que sólo los 9 eigenvalores que se encuentran más a la derecha son los candidatos a considerar para el posterior análisis de ICA.

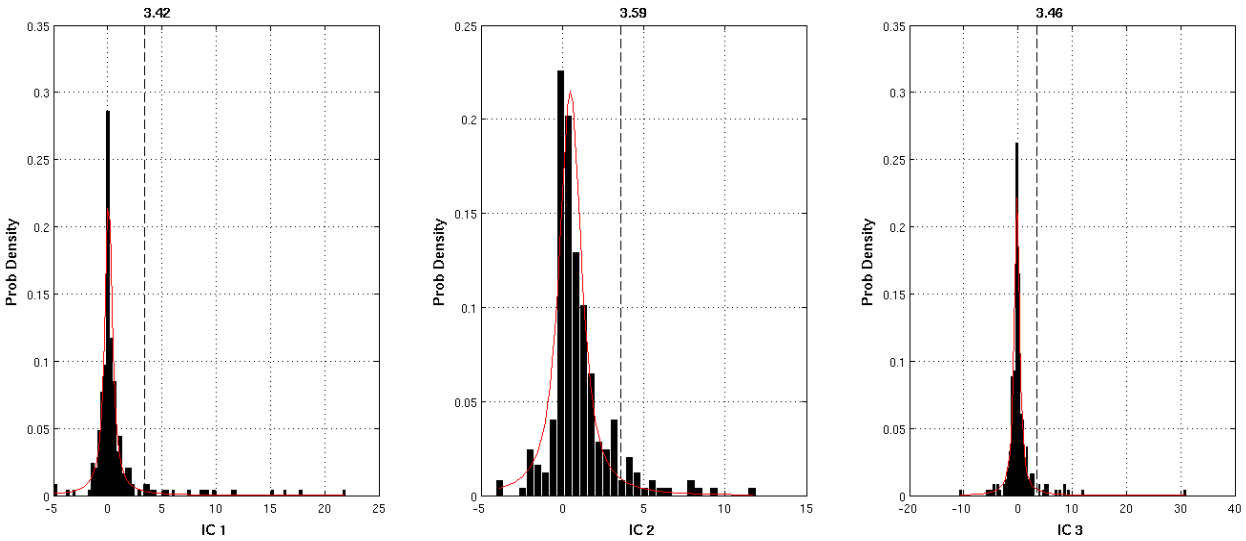


Figura 10. Definición de 3 sectores en SDH de *E. coli*. Por cada componente independiente se ajusta una distribución *t* (línea roja) y se determinan los residuos de cada sector según el punto de corte (línea punteada) de 0.96, por lo que sólo se consideró el 4% más significativo de cada componente.

De acuerdo a los residuos identificados como estadísticamente significativos durante los diferentes análisis con el alineamiento de la SDH completa se procedió a condensar la información final en la siguiente tabla de acuerdo a los residuos y numeración de la SDH:

Proteína y punto de Corte	Sector1	Sector2	Sector3	Conservados
C-R-C (SDH) al 0.8	V6, G8, P10, H13, S14, S16, P17, H20, Q21, Y33, G54, N59, T61, D73, A84, T87, N100, T101, G103, L106, A127, A130, R132, G133, N149, R150, T151, A155, L158, F162, D182, N186, T188, G191, I192, C210, Y211, D212, Y215, K217, D236, G237, L238, G239, M240, L241, V242, Q244, A245, A246, F249, W252, H253, G254	*ND	ND	H20, Y33, T61, P63, K65, N86, N100, D102, D212, Y215, T220, G237, M240, Q244, W252
C-R-C (SDH) al 0.9	P10, H13, S14, S16, P17, H20,Y 33, N59, T87, N100, T101, A127, A130, R132,	ND	ND	H20, Y33, T61, P63, K65, N86, N100, D102, D212, Y215, T220, G237, M240,

	N149, R150, T151, N186, G191, I192, Y215, K217, M240, L241, F249, W252			Q244, W252
C-R-C (SDH) al 0.95	H13, H20, Y33, N59, N149, Y215, M240, W252	ND	ND	H20, Y33, T61, P63, K65, N86, N100, D102, D212, Y215, T220, G237, M240, Q244, W252
C-R-C (SDH) al 0.96 TRES SECTORES	S14, S16, P17, H20, N28, Y33, G57, T87, N100, T101, Y215, C226, G237, Q244, A245, F249, L251, H253, G254	Q21, A24, F50, G54, N59, V62, A82, G103, R121, G133, D182, P197, Y211, K217, D236, G239, V242, A245, A246, H253	Y33, N59, A84, A130, R132, N149, R150, T151, F162, Y211, D236, L238, L241, W252	H20, Y33, T61, P63, K65, N86, N100, D102, D212, Y215, T220, G237, M240, Q244, W252

Tabla1. Residuos acoplados por sector de la SDH completa a diferentes puntos de corte. *ND (no determinado).

De acuerdo a la tabla 1 podemos observar que el SCA con tres sectores a un punto de corte de 0.96 comprende un total de 53 residuos acoplados; de los cuales se comparten 40 con el análisis de un sector al 0.8 de punto de corte (segunda fila), 20 con el análisis de un sector al 0.9 (tercera fila) y 7 con el de un sólo sector al 0.95 (cuarta fila). Aunque es de esperarse que se compartan más residuos acoplados en el análisis anterior con un punto de corte más relajado, se encuentran a otros distintos, a pesar de que en todos los casos se utilizó el mismo alineamiento. Esos residuos que no se encontraron en el análisis de un sólo sector de 0.8 como punto de corte fueron A24, F50, G57, P197 y L251, todos ellos clasificados como neutros, y a excepción de A24, usualmente como hidrofóbicos (ver figura 11). Sin embargo, hasta donde pudimos observar, ninguno de ellos tiene un papel asociado en la literatura.

La determinación del punto de corte va de la mano a lo ya comentado en trabajos previos de SCA, de que sólo alrededor del 20% de los residuos son los que coevolucionan como para convertirse en “mutacionalmente significativos” (McLaughlin *et al*. 2012). Así, si consideramos el ejemplo de que al buscar los residuos acoplados, donde por cada sector el punto de corte será de 0.95, por lo que sólo el 5% de los residuos será considerado, entonces esperaríamos obtener un total del 15% de los residuos como estadísticamente acoplados.

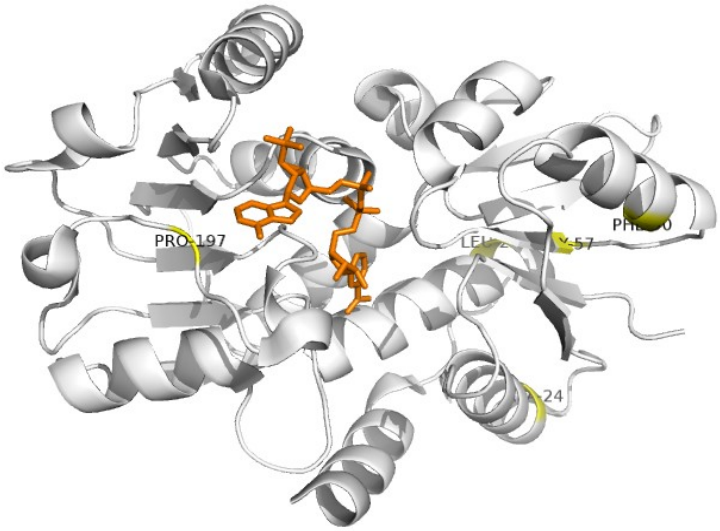


Figura 11. Residuos acoplados únicamente en SCA de SDH de 3 sectores. Se muestran en amarillo aquellos residuos que se presentan exclusivamente en la quinta fila de la tabla 1.

De forma inversa, aquellos que se recuperan con un sólo sector al 0.8 de corte que no se presentan en el análisis de tres sectores son V6, G8, P10, H13, T61, D73, L106, A127, A155, L158, N186, T188, G191 y I192. A excepción de D73, todos los demás aminoácidos presentan carga neutra a pH 7; y a excepción de N186, A127 y A155, todos los demás suelen considerarse hidrofóbicos. Estos residuos se presentan identificados en la estructura de la figura 12. Hasta donde pudimos observar, en la literatura sólo se menciona que T61 forma parte de la cavidad de unión del sitio activo de la SDH (Michel *et al*. 2003), y que incluso pudiera unir al hidroxilo C3 del shikimato (Peek *et al*. 2011). Por otra parte, A127 es uno de lo sitios que de acuerdo al sitio www.pdb.org unen al cofactor NADP.

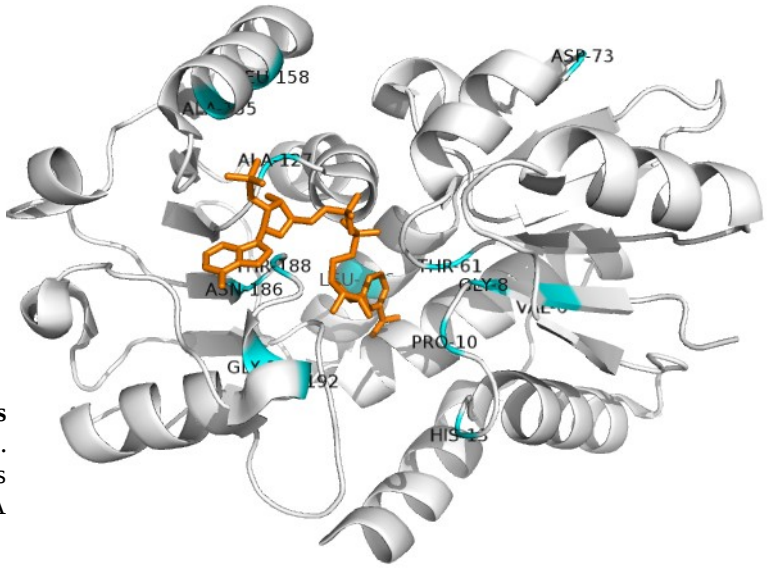


Figura 12. Residuos acoplados únicamente en SCA con un sólo sector. Se muestran en color cian aquellos residuos que no caen en el análisis de SCA con tres sectores.

Determinación de los sectores de residuos estadísticamente acoplados en el dominio Rossmann de la SDH

Se hicieron dos análisis del dominio Rossmann de la SDH. Por un lado, se utilizó el mismo alineamiento de la proteína completa, recortando y recuperando únicamente la parte que correspondía al Rossmann de SDH (sitios 102 al 244) y aquello que alineara con él. A este alineamiento se le llamó “podado”.

Por otra parte, utilizando la herramienta en línea BLAST, se sometió al dominio Rossmann de la SDH de *E. coli* (residuos D102 al Q244) como secuencia de búsqueda y se recuperó el alineamiento local. Era de esperarse que los homólogos recuperados no fuesen los mismos que en el alineamiento donde la secuencia de búsqueda fue el de la proteína completa. Esto debido a que, como ya se mencionó antes en nuestra definición de un dominio, el segmento correspondiente al Rossmann presenta una historia evolutiva diferente al resto de la cadena polipeptídica en donde se haya inmerso, por lo que los resultados del BLAST variaron con respecto a una búsqueda con la SDH completa.

Análogamente, analizamos al dominio Rossmann cuando esta combinado con otros dominios. A continuación la explicación.

Determinación de los sectores de residuos estadísticamente acoplados en el dominio Rossmann cuando está asociado a dominios distintos

Como ya se mencionó en el apartado de *Modelo de Estudio*, el dominio Rossmann ha mostrado ser de los más versátiles de la naturaleza al encontrarse combinado con dominios de distintas familias, y de acuerdo al reporte de Bashton y Chothia en 2002; el dominio Rossmann puede presentarse combinado con otro dominio en una misma proteína en 4 geometrías distintas (R-C, C-R, R/C/R, y C/R/C; donde R hace referencia a Rossmann y C a catalítico), y para cada caso existen estructuras cristalográficas representativas, las cuales se muestran en la tabla 2. Estas estructuras nos sirvieron para visualizar sectores a partir de

alineamientos de la secuencia correspondiente a cada cristal. Por cierto, la SDH presenta a su dominio Rossmann en la geometría C/R/C.

PDB ID	Proteína	Familia	Geometría
1LDM	Apo-Lactato Deshidrogenasa de *Squalus acanthius*	Lactato y malato deshidrogenasas, dominio C-terminal	R-C
1BW9	Fenilalanina deshidrogenasa de *Rhodococcus sp.* M4	Aminoácido deshidrogenasas	C-R
2X0N	Gliceraldehído-3-fosfato deshidrogenasa de *Trypanosoma Brucei*	Gliceraldehído-3-fosfato deshidrogenasas-like	R/C/R
1PJC	L-alanina deshidrogenasa de *Phormidium lapideum*	L-alanina deshidrogenasas	C/R/C

Tabla 2. Cristales representativos de las diferentes geometrías en las que se presenta un dominio Rossmann de acuerdo a Bashton y Chothia en el 2002.

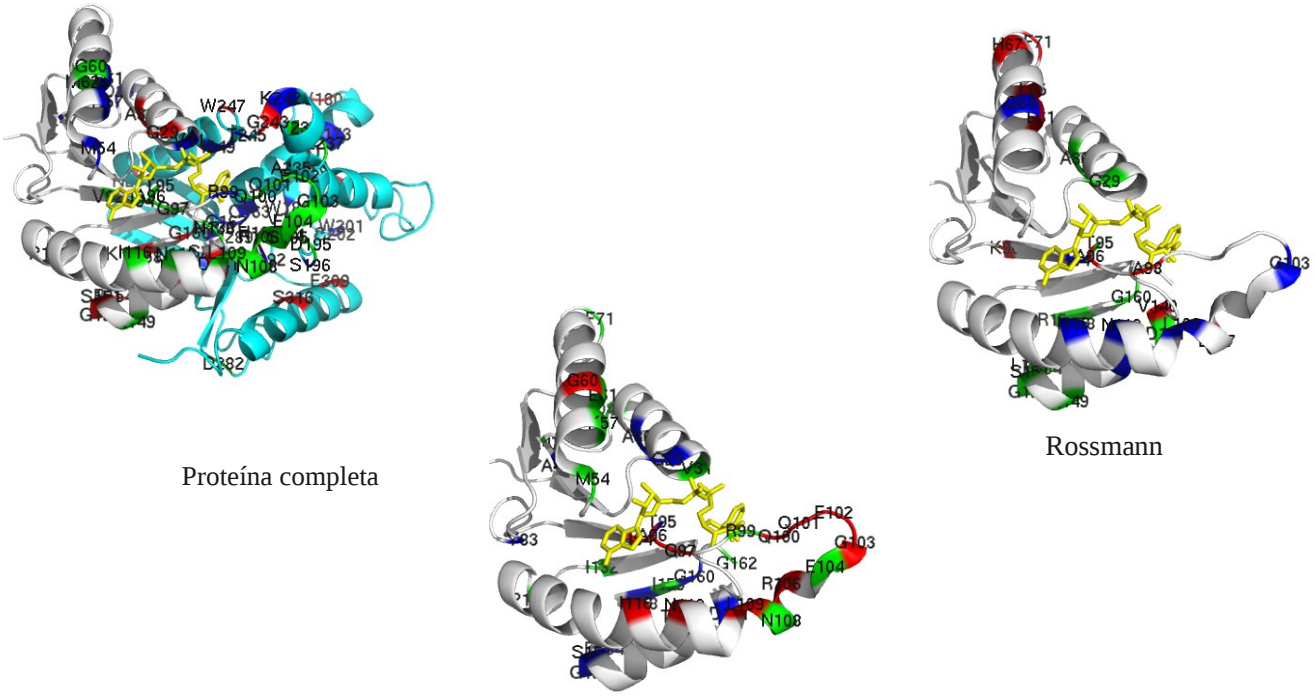
Para cada caso mencionado en la Tabla 2 se hicieron dos análisis de SCA distintos: Uno utilizando un alineamiento contra la proteína completa y otro con un alineamiento contra únicamente la secuencia correspondiente al dominio Rossmann, de acuerdo a las anotaciones de la base de datos de CATH. Como ya se mencionó ambos alineamientos difieren tanto en tamaño como en homólogos encontrados. Por tal motivo, se decidió hacer también el análisis con el alineamiento “podado” correspondiente a cada cristal en analogía al caso de la SDH mencionado anteriormente.

Es importante mencionar que existen residuos que pueden compartirse entre sectores, esto ya se ha observado por Halabi y colaboradores en el 2009 con un par de posiciones que formaban parte tanto del sector rojo como del verde (Halabi *et al*. 2009) en la interfase de ambos sectores en la tripsina de rata. En un esfuerzo por entender el papel de los sectores encontrados en las diferentes moléculas analizadas en nuestro proyecto, recurrimos a la literatura en búsqueda de papeles para ciertos aminoácidos que coincidieran con nuestros resultados, pues, como Halabi y sus colaboradores mencionaron en 2009, la interpretación funcional de los sectores no es obvia en ausencia de información adicional. No hay sector que corresponda a ninguna subdivisión de proteínas como segmentos de estructura primaria, elementos de estructura secundaria ni arquitectura de dominio. Los sectores tampoco son distinguibles por grado de exposición al solvente ni conservación de posiciones tomada de forma independiente (Halabi *et al*. 2009).

A continuación se mencionan los papeles reportados para para algunos de los residuos identificados según el análisis de acoplamiento estadístico para cada una de las proteínas utilizadas en este trabajo; tal información se resume en las tablas 5, 7, 9, 11 y 13. Para facilitar la visualización, las imágenes correspondientes a cada geometría son únicamente de altos puntos de corte (superiores a 0.9) en SCA. Adicionalmente se muestra en forma de tablas (tablas 3, 6, 8, 10 y 12) las diferentes posiciones identificadas como significativas de acuerdo a tales puntos altos de corte.

Geometría R-C (1LDM): Lactato Deshidrogenasa

A)



B)

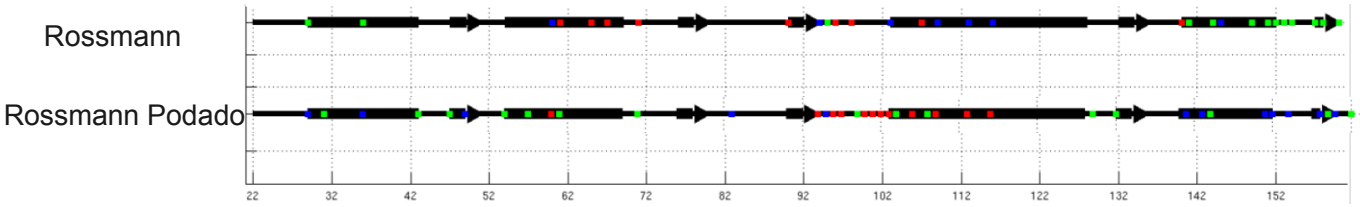


Figura 13. Residuos estadísticamente acoplados en la enzima Lactato deshidrogenasa. Los residuos de los sectores 1, 2 y 3 se resaltan en colores azul, rojo y verde respectivamente. En A) se muestra en gris al dominio Rossmann (residuos 1-162), en cian el dominio catalítico y en bastones amarillos el cofactor NAD. En B) se muestran los residuos acoplados en la estructura secundaria del dominio Rossmann.

Polipéptido y punto de corte	Sector1, azul	Sector2, rojo	Sector3, verde	Conservados
R-C (1LDM) al 0.96	31, 43, 47, 54, 57, 61, 99, 104, 108, 129, 162, 163, 173, 188, 189, 192, 196, 201, 202, 237, 242,	29, 36, 95, 141, 144, 151, 152, 154, 158, 160, 170, 171, 180, 193, 194, 195, 208, 210, 238, 243, 247,	60, 62, 92, 94, 95, 96, 97, 100, 101, 102, 103, 104, 105, 106, 108, 109, 113, 116, 138, 149, 157,	52, 64, 106, 138, 139, 141, 166, 169, 191, 193, 289, 296

	245, 249, 270, 271	263, 289, 309, 316	195, 237, 238, 282	
Rossmann 1LDM al 0.96	60, 94, 95, 96, 103, 109, 113, 116, 145, 149	61, 65, 67, 71, 90, 96, 98, 107, 140, 151	29, 36, 95, 141, 144, 149, 151, 152, 153, 154, 157, 158, 160	27, 32, 52, 64, 106, 138, 139, 141
Rossmann Podado 1LDM al 0.96	29, 36, 49, 83, 95, 141, 143, 144, 151, 152, 154, 158, 160	60, 94, 96, 97, 100, 101, 102, 103, 104, 106, 108, 109, 113, 116	31, 43, 47, 54, 57, 61, 71, 99, 104, 108, 129, 132, 144, 159, 162	52, 64, 106, 138, 139, 141

Tabla 3. Posiciones de residuos estadísticamente acoplados para la enzima lactato deshidrogenasa (PDB ID: 1LDM).

En cuanto a los residuos estadísticamente acoplados en la enzima lactato deshidrogenasa completa (Figura 13), destaca la presencia de H193, quien es esencial para la catálisis y se presenta flanqueada por G191 y G194, quienes también están acoplados y permiten que existan cambios conformacionales durante la catálisis. Tal H193 presenta enlaces de hidrógeno con N138 y D166, quienes se unieron a un ion sulfato durante la obtención del cristal. Otro sulfato se une, entre otros residuos, a R171. T245 está cerca de este sitio de unión (Abad-Zapatero *et al.* 1987).

Por otra parte, R106 tiene el mayor cambio conformacional y está involucrado en catálisis, presentando el mayor factor de temperatura. Dado que G103 también tiene un alto factor de temperatura, podría además ayudar en la flexibilización del asa que cubre el centro activo de la lactato deshidrogenasa en presencia del sustrato y el cofactor unidos a ella. En analogía a otras enzimas, la Y236 podría fosforilarse inhibiendo a la enzima. Entre otras cosas, V26 e I94 comparten un enlace de hidrógeno (Abad-Zapatero *et al.* 1987).

En la comparación de las estructuras de la apo-enzima y su complejo ternario, Hackert y sus colaboradores en 1976 mencionaron que los mayores movimientos están asociados con la región que va de la A96 a K118 y con la última hélice. Esta última hélice hace contacto, entre otros residuos, con L108. La Q192 (vecina de la histidina esencial 193) forma un enlace de hidrógeno con S316 y un contacto hidrofóbico con L320, quien no está estadísticamente acoplado ni muy conservado. Esta región tiene interacciones hidrofóbicas con otros residuos, como L287, W188 y P268. En sí, varios movimientos que involucran residuos como L108, T141, Q192, entre otros, están alrededor del centro activo, y el movimiento del asa no sólo presenta residuos importantes para la coenzima y el sustrato, sino que también es responsable de cambios conformacionales de la H193 esencial. Los autores de tal reporte también mencionaron que es notable que la parte de la molécula que sufre alteraciones conformacionales no sólo está asociada con el centro activo, sino también con la parte externa de la enzima, mientras que mucho del resto de la subunidad está asociado con estructuras involucradas en contactos subunidad-subunidad. En cuanto al complejo abortivo (lactato deshidrogenasa-NAD-piruvato), el piruvato se une covalentemente a la posición C4 de la nicotinamida y forma un enlace de hidrógeno con H193 (Hackert *et al*. 1976).

R99 une el asa con los fosfatos de NAD. Se piensa que E138 empuja el asa en la región del residuo Q100 (Hackert *et al*. 1976).

Por otra parte, Beek y sus colaboradores reportaron en 1997 que los residuos R99, V136, N138, D143, S161, D166, R171, H193 están a una cercanía importante del cofactor NAD (Figura 13A). R99 es un residuo del asa del sitio activo (que va del sitio 96 al 108), D166 forma un par iónico con la histidina del sitio activo (H193) y R169, ésta última arginina es en parte responsable de unir al sustrato. El NH_2 del cofactor forma un puente de hidrógeno con V136 y con la cadena lateral de N138.

Los residuos conservados G27 y D52 parecen tener cierta interacción con el cofactor (Vincent *et al.* 1997).

M62 es uno de los residuos que en especies de peces barracuda parece ser responsable del cambio de K_m entre especies que viven a diferentes temperaturas a pesar de pertenecer a una región de muy baja movilidad en la enzima y estar alejado del sitio activo (Holland *et al.* 1997).

Existe reportado un listado de residuos del centro activo y/o de catálisis de la lactato deshidrogenasa (Eventoff *et al.* 1977). A continuación se menciona en la tabla 4 aquellos que fueron detectados como acoplados (en negritas):

Región del ligando	Residuos
Adenina	**V26, V51, D52, V53, M54, Y83, I94, A96, I116, I120**.
Ribosa de la adenina	G27, **G29, D52, K57**
Pirofosfato	A30, **K57, G97, R99**, Y244
Ribosa de la nicotinamida	**V31, T95, A98**, S137
Nicotinamida	**V31, V136, N138, L165, T245, I249**
Sustrato	**R106, H193**
Anillo negativo	**E104, S105, G194, D195, D230**, D233, S234, **Y236, E237**
Rearreglo del asa	**Q100**

Tabla 4. Residuos del centro activo y/o catálisis según Eventoff *et al*. 1977 que coinciden con los residuos acoplados (negritas) de este trabajo.

El espacio entre el asa y la parte rígida de la enzima lactato deshidrogenasa parece estar alineado por un anillo de cargas negativas provisto por D195 y E237, además de D233, quien no está acoplado. Cuando el asa se cierra, E104 se añade al anillo descrito. La carga positiva del grupo guanidino de R106 debe pasar por este anillo durante el cambio conformacional entre las estructuras de la apo-enzima y el complejo ternario. Eventoff y colaboradores pensaron que esta carga positiva en el centro activo puede ser un factor que contribuya a la unión débil del NAD+ en comparación al NADH. La Y236 también está dentro de este anillo de cargas negativas, además de estar asociada con una pérdida de actividad. Una inspección de los mapas de densidad electrónica muestra posibilidades de enlaces hidrógeno del grupo carboxiamida de la nicotinamida con la S161. La revisión de la secuencia también ha hecho pensar que el sitio de unión del lado B de la nicotinamida sea

hidrofóbico debido a la presencia de V136 e I249. Q100 hace un enlace de hidrógeno con N140 cuando el asa está abajo, mientras que hace un puente de hidrógeno con el carbonilo de la cadena principal de G103 cuando el asa está arriba (Eventoff *et al.* 1977).

Dado que la lactato deshidrogenasa del músculo de *S. acanthius* tiene una isoenzima en el corazón de organismos como pollo o cerdo; se piensa que los cambios en la hidrofobicidad de la cavidad de unión de adenosina (dada por los residuos I94, I116, M54), en la isoenzima de corazón, es debido a una disminución en el tamaño de las cadenas de aminoácidos, por lo que al inicio de la unión del cofactor, la adenosina se pegará con una afinidad reducida a esta isoenzima en comparación con la versión de la enzima de músculo (Eventoff *et al.* 1977).

Por último, la página www.pdb.org reporta para el cristal 1LDM, en la sección "Sequence" que los sitios de unión a NAD por evidencia del software GetSite (El autor de esta tesis supo de tal programa sólo gracias a una comunicación personal con la Dra. Rachel Kramer Green, una persona íntimamente involucrada en el mantenimiento de la base de datos PDB, durante el 2013) son los siguientes: D52, V53, M54, Y83, T95, A96, A98, R99, V136, S137, N138, S161, H193. Todos ellos identificados como acoplados y/o muy conservados en los análisis realizados a la lactato deshidrogenasa de *S. acanthius* (excepto S137).

Residuos Estadísticamente Acoplados	Papel	Referencia
H193	Esencial, forma enlace de hidrógeno con el piruvato.	Abad-Zapatero *et al.* 1987 Hackert *et al*. 1976
G191, G149	Cambios conformacionales durante la catálisis.	Abad-Zapatero *et al.* 1987
N138, D166	Enlace de hidrógeno con H193 y unión a ión sulfato. El cofactor forma un puente de hidrógeno con N138. D166 forma un par iónico con H193.	Abad-Zapatero *et al*. 1987 Beek *et al.* 1997
R106	Catálisis.	Abad-Zapatero *et al.* 1987
Y236	Fosforilación que inhibiría a la enzima.	Abad-Zapatero *et al*. 1987 Eventoff *et al*. 1977
V26, I94	Comparten enlace de hidrógeno.	Abad-Zapatero *et al*. 1987
A96-K118	Región de alto movimiento.	Hackert *et al*. 1976
Q192, S316	Comparten enlace de hidrógeno.	Hackert *et al*. 1976
R169	En parte responsable de unir al sustrato.	Beek *et al.* 1997
G27, D52	Cierta interacción con el cofactor.	Vincent *et al*. 1997
L287, W188 y P268	Tienen interacciones hidrofóbicas con Q192 y S316.	Hackert *et al*. 1976
L108, T141 y Q192	Se encuentran alrededor del sitio	Hackert *et al*. 1976

	activo y presentan movimiento, los cuales provocan cambios conformacionales en H193.	
R99, E138 y E138	R99 une el asa donde se encuentra con los fosfatos de NAD. Se piensa que E138 empuja esa asa en la región de Q100.	Hackert *et al*. 1976
V136	Forma enlace de hidrógeno con el cofactor.	Beek *et al.* 1997
M62	Parece ser el responsible del cambio en K_m entre especies de peces barracuda que viven a diferentes temperaturas.	Holland *et al*. 1997
D195, E237, E104, R106.	Proveen de un anillo de cargas negativas que parece alinear un asa de la enzima y su parte rígida. Cuando el asa se cierra, E104 se añade al anillo descrito. La carga positiva del guanidino de R106 debe pasar por este anillo durante el cambio conformacional entre apoenzima y complejo ternario. Lo cual puede influir en unir débilmente al NAD+ en comparación al NADH.	Eventoff *et al*. 1977
I94, I116, M54	Cambian la hidrofobicidad de la cavidad de unión a adenosina en la isoenzima del corazón de pollo o cerdo, ya que los residuos de la isoenzima de corazón tienen cadenas laterales más cortas, lo cual provoca que la adenosina se una con menor afinidad que en la enzima de músculo.	Eventoff *et al*. 1977
D52, V53, M54, Y83, T95, A96, A98, R99, V136, N138, S161, H193.	Sitios de unión a NAD.	http://www.pdb.org/pdb/explore/remediatedSequence.do?structureId=1LDM

Tabla 5. Papeles de algunos de los residuos acoplados en la enzima lactato deshidrogenasa según la literatura.

Geometría C-R (1BW9): Fenilalanina deshidrogenasa

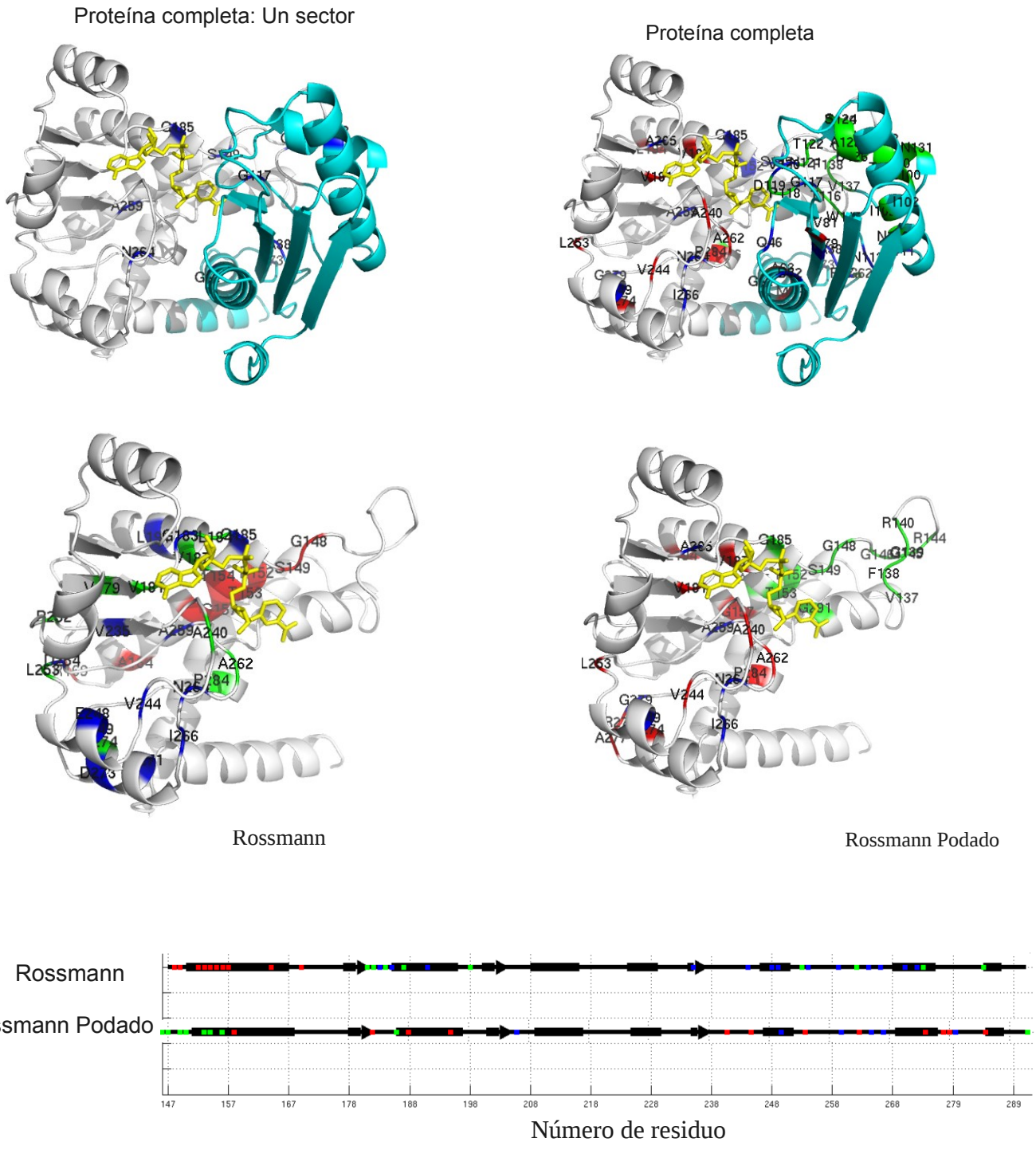


Figura 14. Residuos acoplados en la enzima fenilalanina deshidrogenasa. Los residuos del sector 1, 2 y 3 se resaltan en colores azul, rojo y verde respectivamente. En A) se muestra en blanco al dominio Rossmann (residuos 137-338), en cian el dominio catalítico y en bastones amarillos el cofactor NAD. En B) se muestran los residuos acoplados según la estructura secundaria del dominio Rossmann.

Polipéptido y punto de corte	Sector1, azul	Sector2, rojo	Sector3, verde	Conservados
C-R (1BW9) al 0.95	38, 62, 73, 117, 128, 139, 149, 185, 259, 264	ND	ND	36, 39, 40, 42, 63, 66, 75, 76, 78, 118, 126, 182, 184, 233, 237, 262, 283, 284, 288, 290
C-R (1BW9) al 0.96 tres sectores	37, 38, 46, 52, 73, 79, 81, 111, 113, 117, 120, 125, 139, 149, 152, 185, 194, 205, 249, 259, 264, 266, 279	63, 64, 81, 116, 138, 140, 181, 187, 194, 240, 244, 253, 262, 274, 284	36, 62, 100, 102, 106, 115, 116, 118, 119, 121, 122, 124, 125, 126, 127, 128, 129, 130, 131, 137, 138, 286	36, 39, 40, 42, 63, 66, 75, 76, 78, 118, 126, 182, 184, 233, 237, 262, 283, 284, 288, 290
Rossmann 1BW9 al 0.96	149, 152, 183, 185, 191, 235, 244, 248, 249, 254, 259, 264, 266, 271, 273	148, 149, 152, 153, 154, 155, 156, 157, 164, 169	179, 181, 184, 187, 232, 240, 253, 262, 274, 284	181, 182, 184, 233, 237, 261, 262, 283, 284, 288, 290
Rossmann extraído 1BW9 al 0.96	139, 149, 152, 185, 194, 205, 249, 259, 264, 266, 278, 279	138, 157, 181, 187, 194, 240, 244, 253, 262, 274, 277, 278, 284	137, 138, 139, 140, 144, 145, 146, 148, 149, 152, 153, 155, 185, 291	182, 184, 233, 237, 262, 283, 284, 288, 290

Tabla 6. Posiciones de residuos estadísticamente acoplados en la enzima Fenilalanina deshidrogenasa (PDB ID: 1BW9). ND = no determinado.

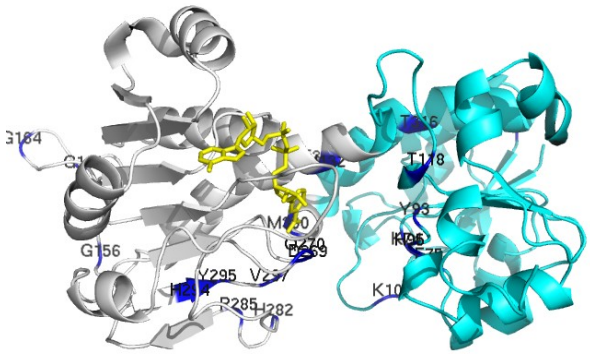
En el artículo de la obtención de las coordenadas atómicas del PDB ID: 1BW9 de Vanhooke y sus colaboradores en 1999, se menciona que los residuos S255, N262, S306 fueron modelados en una conformación alternativa para tener sentido con su densidad electrónica. Adicionalmente, el grupo amida del esqueleto de N262 y el grupo carbonilo de la A239 están a distancia de enlace de hidrógeno con los grupos hidroxilo 2' y 3' de la ribosa de la nicotinamida del cofactor NAD respectivamente (Figura 14A). Existen enlaces de hidrógeno entre los oxígenos del fosforil de NAD y los grupo amida del esqueleto de A185. En un caso donde se cristalizó el complejo inhibidor enzima*NAD*β-fenilpropionato, se observó que tanto el residuo T153 como el S149 forman puentes de hidrógeno a partir de sus respectivos O^{γ} con el oxígeno carboxiamida del dinucleótido. El D118 tiene un papel activo durante la catálisis. D205 es candidato a formar hasta dos enlaces de hidrógeno con el NAD, mientras M63, P117 y G184 simplemente forman parte del sitio activo y están muy cerca del dinucleótido (Vanhook *et al.* 1999).

Por otra parte, el sitio www.pdb.org indica para la secuencia del cristal 1BW9, en la sección “Sequence”, que los residuos que unen al NAD según evidencia de software (en negritas quienes adicionalmente son acoplados y/o muy conservados) son K66, K78 (quien es uno de los catalíticos), **P115**, **V119**, **G182**, **G184**, **A185**, V186, A204, **D205**, T206, R210, L224,C238, **A239**, **M240**, A260, **A261**, **N262**.

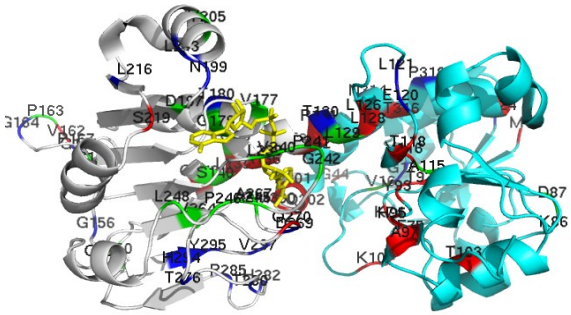
Residuos estadísticamente acoplados	Papel	Referencia
S255, N262 y S306	Fueron modelados dentro de la densidad electrónica en conformaciones alternativas. El grupo amida del esqueleto de N262 está a una distancia de enlace de hidrógeno del grupo hidroxilo 2' de la ribosa de la nicotinamida del NAD.	Vanhooke *et al*. 1999
A239	Está a una distancia de enlace de hidrógeno del grupo hidroxilo 3' de la ribosa de la nicotinamida del cofactor NAD.	Vanhooke *et al*. 1999
A185	Hay enlaces de hidrógeno entre los oxígenos del fosforil de NAD y el grupo amida del esqueleto de A185.	Vanhooke *et al*. 1999
T153, S149	Forman puentes de hidrógeno a partir de sus respectivos O^{γ} con el oxígeno carboxiamida del dinucleótido.	Vanhooke *et al*. 1999
D118	Papel activo durante la catálisis.	Vanhooke *et al*. 1999
D205	Es candidato a formar hasta dos enlaces de hidrógeno con el NAD.	Vanhooke *et al*. 1999
M63, P177 y G184	Forman parte del sitio activo y están muy cerca del dinucleótido.	Vanhooke *et al*. 1999
P115, V119, G182, G184, A185, D205, A239, M240, A261, N262	Sitios de unión a DNA.	http://www.pdb.org/pdb/explore/remediatedSequence.do?structureId=1BW9

Tabla 7. Residuos estadísticamente acoplados en la enzima fenilalanina deshidrogenasa que tienen un papel asociado en la literatura.

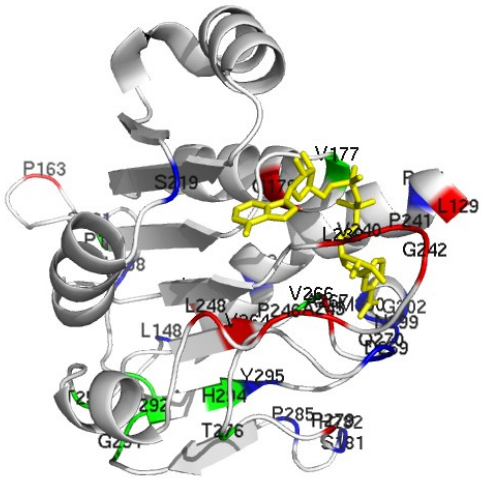
Geometría C-R-C (1PJC): Alanina deshidrogenasa



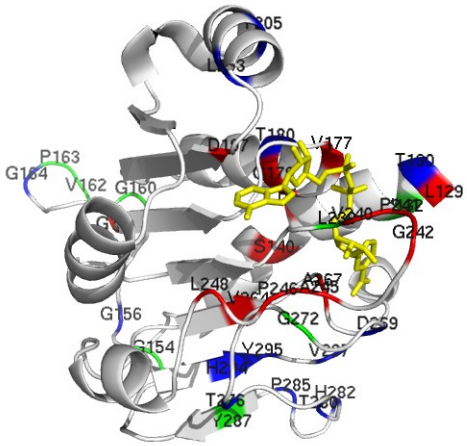
Proteína completa: Un sector



Proteína completa



Rossmann



Rossmann Podado

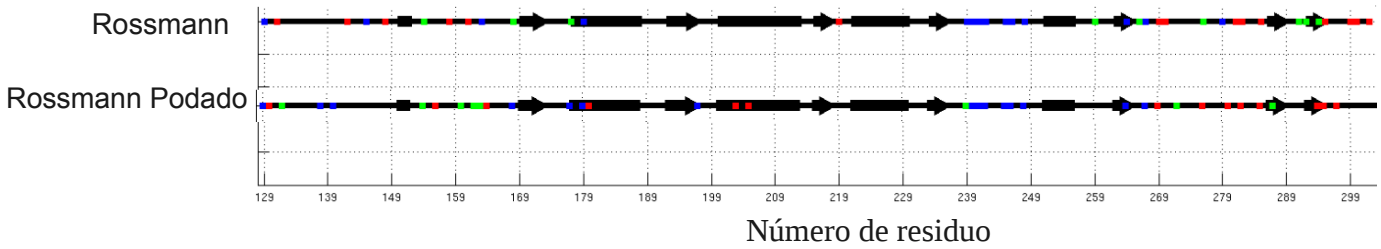


Figura 15. Residuos estadísticamente acoplados en la enzima Alanina deshidrogenasa. Los residuos correspondientes a los sectores 1, 2 y 3 se muestran en colores azul, rojo y verde respectivamente. En A) el dominio Rossmann (residuos 129-304) se muestra en gris, mientras que el catalítico en cian. El cofactor NAD se muestra en bastones amarillos. En B) se muestran los residuos acoplados en la estructura secundaria del dominio Rossmann.

Polipéptido y punto de corte	Sector1, azul	Sector2, rojo	Sector3, verde	Conservados
C-R-C (1PJC) al 0.9	10, 31, 74, 75, 93, 95, 118, 132, 156, 161, 164, 269, 270, 282, 285, 294, 295, 297, 300, 307, 316	ND	ND	1, 8, 13, 15, 20, 40, 47, 50, 132, 137, 174, 176, 179, 182, 197, 232, 246, 257, 265, 271
C-R-C (1PJC) al 0.96 tres sectores	17, 31, 87, 120, 121, 130, 156, 161, 164, 177, 180, 199, 203, 205, 216, 269, 276, 280, 282, 285, 294, 295, 297, 307, 314, 318	1, 9, 10, 44, 74, 75, 92, 93, 95, 97, 103, 116, 118, 126, 128, 129, 131, 132, 136, 139, 142, 162, 163, 219, 239, 270, 300, 301, 302, 316, 324	16, 87, 115, 129, 138, 140, 150, 163, 168, 177, 179, 197, 205, 239, 240, 241, 242, 245, 246, 248, 260, 264, 267	1, 8, 13, 15, 20, 40, 47, 50, 132, 137, 174, 176, 179, 182, 197, 232, 246, 257, 265, 271
Rossmann 1PJC al 0.96	131, 142, 148, 158, 161, 219, 269, 270, 281, 282, 285, 295, 299, 300, 302	129, 145, 163, 168, 177, 179, 239, 240, 241, 242, 245, 246, 248, 264, 267, 279	154, 168, 177, 259, 266, 276, 291, 292, 294	132, 137, 174, 176, 179, 182, 197, 232, 246, 257, 265, 271, 299
Rossmann extraído 1PJC al 0.96	130, 156, 164, 168, 177, 180, 203, 205, 269, 276, 280, 282, 285, 294, 295, 297	129, 138, 140, 168, 177, 179, 197, 240, 241, 242, 245, 246, 248, 264, 267	132, 154, 160, 162, 163, 239, 272, 287	132, 137, 174, 176, 179, 182, 197, 232, 246, 257, 265, 271

Tabla 8. Posiciones de residuos estadísticamente acoplados en la enzima Alaninda deshidrogenasa (PDB ID: 1PJC). ND= no determinado.

Investigando en la literatura sobre el papel de los residuos acoplados encontrados (Figura 15) en la alanina deshidrogenasa, se ha encontrado que en el momento de formar la estructura cuaternaria, que en este caso consiste en un hexámero, los residuos acoplados K184, G161 y R138 de una subunidad se empacan con E207 (quien no está acoplado) de la siguiente subunidad formando un parche hidrofílico. También los residuos P163,G164,V165 y P167 forman un asa y ésta interactúa con otra subunidad de enzima para formar una interface; mientras que los residuos K169 y E194 se localizan en el centro de esa interface (Baker *et al.* 1998).

Con respecto a la unión del dinucleótido, el anillo de adenina se sitúa en una cavidad hidrofóbica formada (entre otros residuos) por V238 y L248, además que S219 forma un puente de hidrógeno con tal anillo. Se forman enlaces de hidrógeno entre los hidroxilos 2' y 3' de la ribosa de la adenina y el residuo D197, quien también es conservado. También se forma un enlace de hidrógeno mediado por una molécula de agua entre el hidroxilo 2' y el nitrógeno del grupo amida de N199. La región que comprende al pirofosfato forma algunos enlaces de hidrógeno ya sean directos o mediados por agua con el asa de G176, V177, V178 y N179. Otro enlace más con S133. El hidroxilo 3' de la ribosa de nicotinamida forma un enlace de hidrógeno con V238, mientras que una cara del anillo de nicotinamida se empaca contra una parte de la superficie de la enzima formada por los residuos A136, V266, V178, y M300. Este patrón de enlaces de hidrógeno es completado por las interacciones entre la región carboxiamida del anillo nicotinamida el grupo amino de M300 y el carbonilo de V297, el cual provoca que el enlace glicosídico quede en conformación *anti* entre el anillo nicotinamida y su ribosa asociada. De hecho, la conformación *syn* de este enlace está

estéricamente impedida por M132 y S133 (Baker *et al.* 1998).

Con respecto al rearreglo de ciertas asas entre la enzima libre y la enzima asociada a su cofactor resalta lo siguiente: El asa que contiene a los residuos V238, L239, V240, P241, G242, A245 y P246 se mueve hasta 6 Å mientras permitiendo que en la cavidad se una el anillo de adenina del NAD. En conjunto con este reposicionamiento, el asa donde se encuentran los residuos V266, A267, V268, D269 y Q270, aquella donde se encuentran T280, S281, H282 y P285, y H294, Y295, V297, P298 y la que tiene al residuo N299 se mueven en la misma dirección hasta 2 Å. Estos movimientos optimizan el acomodo del cofactor en la enzima y permite que la región carboxiamida del anillo nicotinamida forme enlace de hidrógeno con el grupo amino de la cadena principal de M300 y el carbonilo de V297 (Baker *et al.* 1998).

Por el lado de la unión del piruvato, el grupo metilo se empaca contra los residuos hidrofóbicos Y93, M132 y L129. El grupo amino de la cadena lateral de K74 interactúa tanto con el oxígeno del carbonilo del piruvato como con uno de los oxígenos de su carboxilo. También se forma un enlace de hidrógeno entre el oxígeno del carbonilo del piruvato y la cadena lateral de H95, mientras que uno de los nitrógenos de la cadena lateral de N299 interactúa con uno de los oxígenos del carboxilato del piruvato (Baker *et al.* 1998).

Ya adentrándose en el mecanismo catalítico, dada la proximidad con que el anillo imidazol de H95 se encuentra del grupo carbonilo del piruvato, se sugiere que tal histidina actúa como el catalizador ácido/base. Es posible que E117 y D269 ayuden a H95 durante la catálisis (Baker *et al.* 1998).

Ågren y colaboradores mencionaron en el 2008 que en el homólogo de *Mycobacterium tuberculosis* los residuos equivalentes a los ya mencionados D269 y H95 son potencialmente catalíticos; mientras que K74 es uno de los residuos que une al piruvato.

La página www.pdb.org menciona para la secuencia del cristal 1PJC, en la sección “Sequence”, que por evidencia de software los residuos que se unen a NAD (en negritas quienes adicionalmente están acoplados y/o muy conservados) son **S133, A136,** G174, **G176, V177, V178, D197,** I198, **S219,** A237, **V238, L239, V266, A267, Q270, N299, M300** y **P301**.

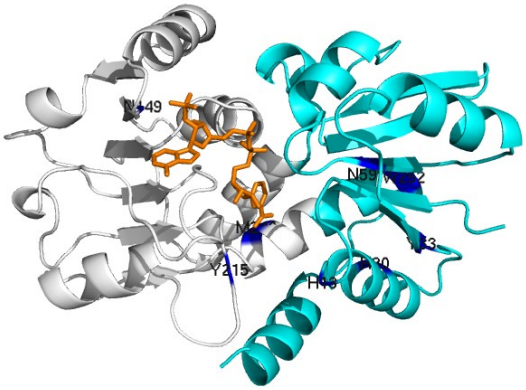
Residuos estadísticamente acoplados	Papel	Referencia
K184, G161, R138, P163, G164, V165, P167, K169, E194	En el momento de formar la estructura cuaternaria (un hexámero), K184, G161, y R138 de una subunidad se empacan con E207 (quien no está acoplado) de la siguiente subunidad formando un parche hidrofílico. P163, G164, V165 y P167 forman un asa y ésta interactúa con otra subunidad de enzima para formar una interfase; mientras que los residuos K169 y E194 se localizan en el centro de esa interfase.	Baker *et al*. 1998

V238, L248, S219.	El anillo de adenina del dinucleótido se sitúa en una cavidad formada, entre otros residuos, de V238 y L248, mientras que S219 forma un puente de hidrógeno con tal anillo.	Baker *et al*. 1998
D197, N199	D197 (quien también es conservado) forma enlaces de hidrógeno con los hidroxilos 2' y 3' de la ribosa de la adenina. Mientras que el grupo amida N199 forma un enlace de hidrógeno mediado por una molécula de agua con el hidróxilo 2' de la ribosa de adenina también.	Baker *et al*. 1998
G176, V177, V178, N179, S133	El loop que va de G176 a N179 forma enlaces de hidrógeno con la región del pirofosfato del NAD, además de otro con S133. Esta S133 impide (junto con M132) que el enlace glicosídico entre el anillo nicotinamida y su ribosa asociada quede en conformación *syn*.	Baker *et al*. 1998
V238	Forma un enlace de hidrógeno con con el hidroxilo 3' de la ribosa de nicotinamida.	Baker *et al*. 1998
A136, V266, V178, M300, V297, M132	Forman parte de la superficie de la enzima en donde se empaca una cara del anillo de nicotinamida. La región carboxiamida del anillo nicotinamida tiene interacciones con el grupo amino de M300 y el carbonilo de V297. Este carbonilo provoca que el enlace glicosídico quede en conformación *anti* entre el anillo nicotinamida y su ribosa asociada.	Baker *et al*. 1998
V238, L239, V240, P241, G242, A245, P246	Forman parte un asa que se mueve hasta 6 Å para formar una región en la cavidad en la cual se unirá el anillo de adenina.	Baker *et al*. 1998
(V266, A267, V268, D269, Q270), (T280, S281, H282, P285), (H294, Y295, V297, P298, N299)	Estas asas se mueven en la misma dirección hasta 2 Å, lo cual optimiza el acomodo del cofactor en la enzima y permite que la región carboxiamida del anillo nicotinamida forme un enlace de hidrógeno con el amino de la cadena principal de M300 y el carbonillo de V297.	Baker *et al*. 1998
Y93, M132, L129, K74, N299	El grupo metilo del piruvato se empaca contra Y93, M132 y L129.	Baker *et al*. 1998 Ågren *et al*. 2008

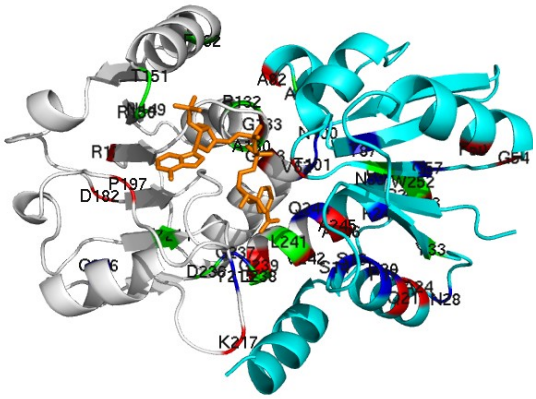
	El grupo amino de la cadena lateral de K74 interactúa tanto con el oxígeno del carbonilo como como con uno de los oxígenos del carboxilo del piruvato. Uno de los nitrógenos de la cadena lateral de N299 interactúa con uno de los oxígenos del carboxilato del piruvato.	
H95, E117 y D269	La cadena lateral de H95 forma un enlace de hidrógeno entre el oxígeno del carbonilo del piruvato. Dada la proximidad con que el anillo imidazol se encuentra del grupo carbonilo del piruvato, se sugiere que ésta histidina actúa como catalizador ácido/base. Es posible que E117 y D269 la ayuden durante la catálisis.	Baker *et al*. 1998 Ågren *et al*. 2008
S133, A136, G176, V177, V178, D197, S219,V238, L239, V266, A267, Q270, N299, M300, P301	Residuos que por evidencia de software unen al cofactor NAD.	http://www.pdb.org/pdb/explore/remediatedSequence.do?structureId=1PJC

Tabla 9. Algunos residuos estadísticamente acoplados en la enzima alanina deshidrogenasa y su papel asociado en la literatura.

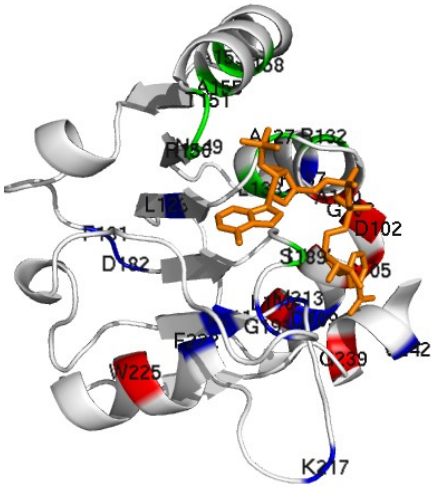
Geometria C-R-C (SDH): Shikimato Deshidrogenasa



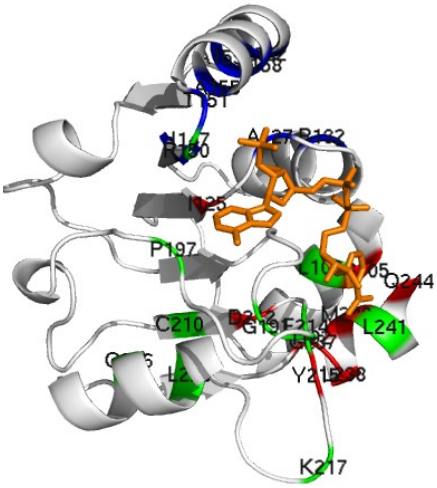
Proteína completa



Proteína completa: Un sector



Rossmann



Rossmann Podado

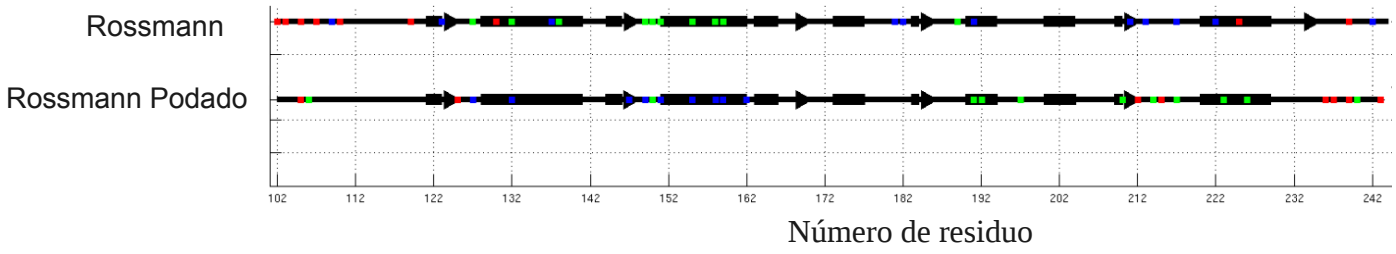


Figura 16. Residuos estadísticamente acoplados en la enzima shikimato deshidrogenasa. Se muestran los residuos correspondientes a los sectore 1, 2 y 3 en colores azul, rojo y verde respectivamente. En A) se muestra en gris al dominio Rossmann (residuos 102-244), en cian al dominio catalítico y en bastones naranja el cofactor NADP. En B) los residuos de los diferentes sectores en la estructura secundaria del dominio Rossmann.

Polipéptido y punto de corte	Sector1, azul	Sector2, rojo	Sector3, verde	Conservados
C-R-C (SDH) al 0.95	13, 20, 33, 59, 149, 215, 240, 252	ND	ND	20, 33, 61, 63, 65, 86, 100, 102, 212, 215, 220, 237, 240, 244, 252
C-R-C (SDH) al 0.96 tres sectores	14, 16, 17, 20, 28, 33, 57, 87, 100, 101, 215, 226, 237, 244, 245, 249, 251, 253, 254	21, 24, 50, 54, 59, 62, 82, 103, 121, 133, 182, 197, 211 , 217, 236, 239, 242, 245, 246, 253	33, 59, 84, 130, 132, 149, 150, 151, 162, 211, 236, 238, 241, 252	20, 33, 61, 63, 65, 86, 100, 102, 212, 215, 220, 237, 240, 244, 252
Rossmann SDH al 0.96	109, 123, 137, 181, 182, 191, 211, 213, 217, 222, 242	102, 103, 105, 107, 110, 119, 130, 225, 239	127, 132, 138, 149, 150, 151, 155, 158, 159, 189	102, 126, 128, 129, 149, 150, 186, 188, 212, 215, 237, 240, 244
Rossmann extraído de SDH al 0.96	106, 127, 132, 147, 149, 150, 151, 155, 158, 159, 162	105, 125, 150, 212, 215, 237, 238, 240, 241, 244	106, 150, 191, 192, 197, 210, 214, 217, 223, 226, 241	102, 212, 215, 220, 237, 240, 244

Tabla 10. Posiciones de residuos estadísticamente acoplados en la Shikimato deshidrogenasa (PDB ID: 1NYT). ND = no determinado.

En la estructura correspondiente al PDB ID: 1NYT (Figura 16A), el grupo amida N-7 del anillo nicotinamida comparte puentes de hidrógeno con el grupo carbonilo de M213 y G237; mientras que la ribosa forma contactos de van der Waals únicamente con las cadenas laterales hidrofóbicas. La parte del pirofosfato forma enlaces de hidrógeno con los átomos de nitrógeno de A130 y otro G129 (no acoplado) (Michel *et al.* 2003).

N149 y R150 están involucrados en el reconocimiento de la parte de la adenosina. Específicamente, el O-3' del grupo hidroxilo de la ribosa de adenosina tiene un enlace de hidrógeno con N149 y con el NH de la cadena principal de A127. Además, la amida de N149 forma un enlace de hidrógeno con el átomo O-1 del fosfato 2'. R150 forma dos enlaces de hidrógeno con los otros átomos de oxígeno del fosfato, mientras que su grupo guanidino se apila contra la cara A del anillo de adenina. Este fosfato es posteriormente estabilizado por interacciones electrostáticas con R154 y por un enlace de hidrógeno con el hidroxilo de T151. En cambio, la cara B de la adenina contacta con la cadena lateral de T188 y S190. Las argininas 150 y 154 juegan un rol crucial en la unión del fosfato de adenosina al formar una “abrazadera electrostática” que toma al fosfato como si fuera el relleno de un emparedado (Michel *et al.* 2003).

Un ion sulfato o fosfato está presente en la cavidad de unión del sitio activo, el cual mantiene enlaces de hidrógeno con los hidroxilos de S14, S16, T61, Y215. También se encontró una molécula de DTT, con un enlace de hidrógeno con Q244, D102 y T61 (Michel *et al.* 2003).

En la conformación abierta de esta SDH, los siguientes residuos se unen por enlace de hidrógeno entre sus cadenas laterales N59-T87-N86(no acoplado), N86(no acoplado)-T101-Q244-N59. Esta red circular de enlaces hidrógeno se rearregla en la conformación cerrada de la enzima, dado que Q244 no presenta puente de hidrógeno con las cadenas laterales de N59 y T101, sino que esta glutamina se enlaza con N86(no acoplado) y con el

hidroxilo principal de T101 (Michel *et al.* 2003).
Con respecto al mecanismo de reacción, se piensa que D102 es un potencial residuo catalítico (Michel *et al.* 2003).

En el estudio realizado por Peek y colaboradores en 2011 se detalla información sobre algunos residuos de una proteína AroE-like1 (Ael1) que ellos mismos cristalizaron. Utilizando un alineamiento para traducir esa información para el cristal 1NYT es posible destacar que A127, G129 y A130 son parte del asa rica en glicinas responsable de la unión a la región pirofosfato de NAD(P). Por otra parte, R150 es la indicada para discriminar entre NAD y NADP. Vuelve a destacarse el hecho de que S14 y S16 son los responsables de unir el carboxilo del shikimato. Es posible que Y215 oriente al sustrato en una posición favorable para la catálisis. Se sugiere también que T61 puede unir al hidroxilo C3 del shikimato, mientras que Q244 une al del C5. Nuevamente se dice que D102 puede ser importante catalíticamente, además de orientar a otros residuos como Q244.

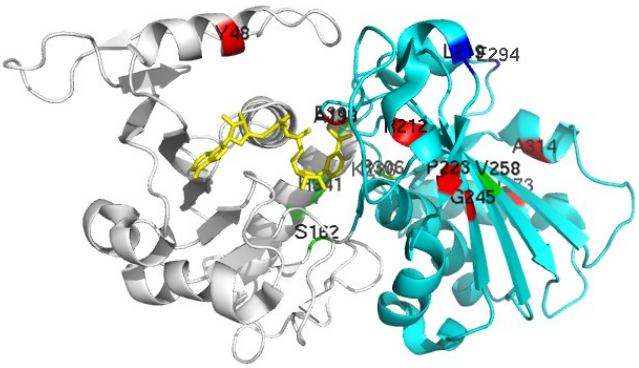
De acuerdo a www.pdb.org, en la sección “Sequence”, los sitios que unen al cofactor NADP en el cristal 1NYT según evidencia de software (en negritas quienes además son acoplados y/o muy conservados) son los siguientes: **D102**, G126, **A127**, G128, G129, **A130, N149, R150, T151, R154**, N187, **T188, S189, S190, M213, Y215, G237, M240 y L241**.

Residuos Estadísticamente Acoplados	Papel	Referencia
M213, G237	El grupo amida N7 del anillo nicotinamida del NADP comparte puentes de hidrógeno con estos residuos, mientras que la ribosa forma contactos de van der Waals con las cadenas laterales hidrofóbicas.	Michel *et al*. 2003
A130	Sus átomos de nitrógeno forman enlaces de hidrógeno con la parte del pirofosfato del NADP.	Michel *et al*. 2003ramer
N149, R150, A127, R154, T151	Están involucrados en el reconocimiento de la parte de la adenosina del cofactor. El O-3' del grupo hidroxilo de la ribosa de adenosina tiene un enlace de hidrógeno con N149 y con el NH de la cadena principal de A127. Además, la amida de N149 forma un enlace de hidrógeno con el átomo O-1 del fosfato 2'. R150 forma dos enlaces de hidrógeno con los átomos de oxígeno del fosfato, mientras que su grupo guanidino se apila contra la cara A del anillo de adenina. Este fosfato es posteriormente estabilizado por interacciones electrostáticas con	Michel *et al*. 2003 Peek *et al*. 2011

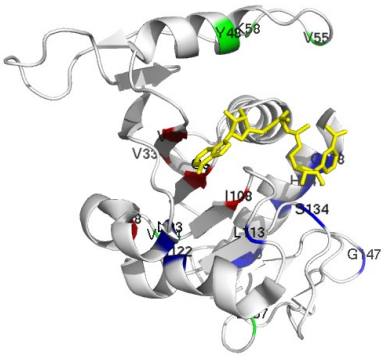
	R154 y por un enlace de hidrógeno con el hidroxilo de T151. R150 y R154 juegan un rol crucial en la unión del fosfato de adenosina al formar una abrazadera electrostática. R150 parece la indicada para discriminar entre NAD y NADP.	
T188, S190	Las cadenas laterales de estos residuos contactan con la cara B de la adenina del cofactor.	Michel *et al*. 2003
S14, S16, T61, Y215	Un ión sulfato o fosfato mantiene enlaces de hidrógeno con los hidroxilos de estos residuos. S14 y S16 son los responsables de unir al carboxilo del shikimato.	Michel *et al*. 2003 Peek *et al*. 2011
N59, T87, T101, Q244, N59, D102, T61	En la conformación abierta, N59-T87-N86(no acoplado), N86(no acoplado)-T101-Q244-N59 se unen por enlace de hidrógeno entre sus cadenas laterales. Esta red circular de enlaces de hidrógeno se rearregla en la conformación cerrada de la enzima, dado que Q244 no presenta puentes de hidrógeno con las cadenas laterales de N59 y T101, sino que ahora, Q244 se enlaza con N86(no acoplado) y con el hidroxilo principal de T101. Una molécula de DTT forma enlace de hidrógeno con estos residuos. D102 es un potencial residuo catalítico, que de hecho parece orientar a otros residuos, como Q244. T61 puede unir al hidroxilo C3 del shikimato.	Michel *et al*. 2003 Peek *et al*. 2011
A127, G129, A130	De acuerdo a un alineamiento con la proteína Ael1, éstos residuos son parte del asa rica en glicinas responsible de la unión a la región del pirofosfato de NADP.	Peek *et al*. 2011
D102, A127, A130, N149, R150, T151, R154, T188, S189, S190, M213, Y215, G237, M240, L241	Sitios de unión al cofactor NADP según evidencia de software.	http://www.pdb.org/pdb/explore/remediatedSequence.do?structureId=1NYT

Tabla 11. Algunos residuos estadísticamente acoplados de la enzima shikimato deshidrogenasa y su papel asociado en la literatura.

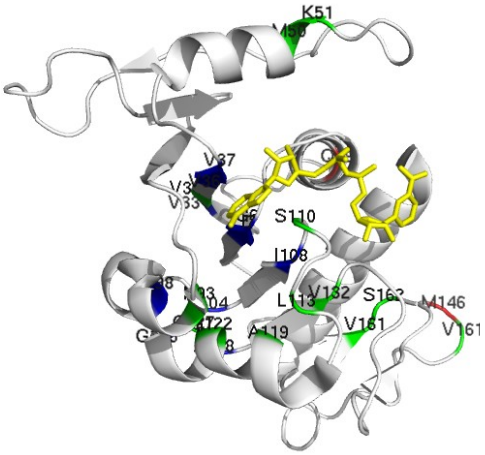
Geometría R-C-R (2X0N): Gliceraldehído 3-fosfato deshidrogenasa



Proteína completa



Rossmann



Rossmann Podado

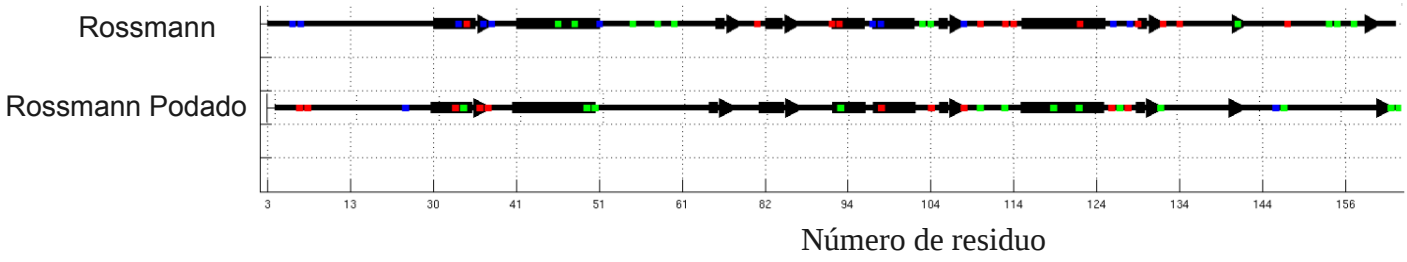


Figura 17. Residuos estadísticamente acoplados en la enzima gliceraldehído 3-fosfato deshidrogenasa. Los residuos correspondientes a los sectores 1, 2 y 3 se muestran en colores azul, rojo y verde respectivamente. En A) color gris se muestra al dominio Rossmann (residuos 1-164 y 336-359), en cian al dominio catalítico y en bastones amarillos el cofactor NAD. En B) se colorearon los residuos acoplados en la estructura secundaria del dominio Rossmann.

Polipéptido y punto de corte	Sector1, azul	Sector2, rojo	Sector3, verde	Conservados
R-C-R (2X0N) al 0.99	219, 294, 306, 336, 341	48, 198, 212, 223, 245, 273, 314	162, 258, 286, 306, 336, 341, 344	7, 8, 10, 11, 12, 13, 14, 52, 55, 98, 110, 111, 122, 130, 135, 144, 148, 164, 165, 166, 167, 168, 169, 172, 189, 191, 193, 196, 199, 203, 211, 214, 222, 225, 236, 248, 249, 250, 296, 297, 299, 310, 331, 332, 333, 334, 335, 338
Ross 2X0N al 0.99	93, 113, 122, 134, 147, 338, 341, 344, 349	6, 7, 33, 36, 98, 108	48, 55, 58, 104, 157	7, 8, 9, 10, 11, 12, 13, 14, 52, 55, 98, 110, 111, 122, 130, 135, 144, 148, 338
Ross extraído 2X0N al 0.96	6, 7, 33, 36, 37, 51, 98, 104, 108, 126, 128	19, 113, 146, 162	34, 50, 51, 93, 110, 113, 119, 122, 127, 132, 147, 161, 162	7, 8, 10, 11, 12, 13, 14, 52, 55, 98, 110, 111, 122, 130, 135, 144, 148

Tabla 12. Posiciones de residuos estadísticamente en la Gliceraldehído 3-fosfato deshidrogenasa glicosomal (PDB ID: 2X0N).

Las coordenadas que decidimos usar en este caso son las correspondientes a la enzima glicosomal (PDB ID: 2X0N). Ver Figura 17.

En cuanto a lo reportado en la literatura con respecto al cristal 2X0N, los residuos I132, S133, A134, A136, S137, G139, A140, K141, T142, F143 forman parte del segmento involucrado en la localización de esta enzima en el glicosoma de *T. brucei* (Michels *et al.* 1991).

De acuerdo a otro estudio de esta enzima, la A89 puede contribuir a la reducción de la afinidad por el NAD en comparación con la afinidad que presenta la isoenzima citosólica por el mismo compuesto (Lambeir *et al.* 1991).

La página www.pdb.org reporta en la sección “Sequence” que los sitios de unión a NAD según evidencia de software (en negritas los residuos que además son acoplados y/o muy conservados) son: G8, G10, R11, **I12, V36, D37, M38**, A89, **Q90, S109, T110,** G111, **S133**, N334 y Y338.

Residuos Estadísticamente Acoplados	Papel	Referencia
I132, S133, A134, A136, S137, G139, A140, K141, T142, F143.	Forman parte del segmento involucrado en la localización de esta enzima en el glicosoma de *T. brucei*.	Michels *et al*. 1991
A89	Puede contribuir a la reducción de la afinidad por el NAD en	Lambeir *et al*. 1991

	comparación con la afinidad que presenta la isoenzima citosólica por el mismo compuesto.	
I12, V36, D37, M38, Q90, S109, T110, S133	Sitios que unen al NAD según evidencia de software.	http://www.pdb.org/pdb/explore/remediatedSequence.do?structureId=2X0N

Tabla 13. Algunos residuos estadísticamente acoplados de la enzima gliceraldehído 3-fosfato deshidrogenasa y su papel asociado en la literatura.

Después de todos los datos recabados es la revisión bibliográfica para encontrar papeles reportados para los residuos estadísticamente acoplados, y la coincidencia de algunos de estos con los que forman parte de la cavidad de unión a sustrato y/o cofactor en cada cristal, así como movimiento o reacomodo de segmentos de la enzima, es posible ver el poder que tiene el método SCA para revelar, a partir de información lineal y en unos cuantos minutos, grupos de residuos importantes de la proteína, incluso los que tienen un papel en cuanto a reconocimiento de sustrato. Aunque esto también se visualiza en la estructura tridimensional, se necesita de días, semanas o meses elucidar la estructura de un proteína. En otras palabras, en un escenario en donde se deseara hacer un análisis rápido sobre qué residuos son importantes en una proteína, SCA es una buena herramienta a usar cuando no contamos con la estructura cristalográfica de una proteína de interés y sólo conocemos su secuencia y la secuencia de proteínas homólogas contra las cuales alinear.

Más adelante, en la sección de discusión se retomará mucha de la información sobre los análisis de SCA ya presentados.

Diseños con RosettaDesign

Después de realizar los diseños (sólo del dominio Rossmann de SDH) como se describió en la sección de **MÉTODOS**, se tomó a la estructura con la menor calificación de cada conjunto (la más estable) y se le asignó un nombre significativo. A continuación se revelan detalles sobre los diseños utilizados:

- SCA: El diseño con menor energía fue el número 981 y presentó una identidad del 55% con respecto a la secuencia del dominio Rossmann original de SDH. Se permitió la variación de todos los residuos a excepción de los estadísticamente acoplados y los muy conservados de acuerdo a un análisis de SCA del alineamiento del dominio Rossmann. El análisis dio lugar a tres sectores con un punto de corte de 0.96 por sector. En total se respetaron 40 residuos, en donde 10 de ellos eran exclusivamente muy conservados. El tiempo utilizado para generar a las 999 secuencias fue de 71.9 horas utilizando una computadora con un procesador Intel® Core™2 Quad CPU Q6600 @ 2.40GHz × 4 y 4 Gb de memoria RAM.

- AntiSCA: El mejor diseño a nivel energético fue el 794, presentando una identidad del 87% con el dominio Rossmann original de SDH. En este caso se permitió la variación de únicamente los sitios que se encontraban estadísticamente acoplados y los muy conservados. El tiempo utilizado para generar a las 999 secuencias fue de 17.8 horas utilizando una computadora con un procesador Intel® Core™2 Quad CPU Q6600 @ 2.40GHz × 4 y 4 Gb de memoria RAM.

- Aleatorio: Quien obtuvo la menor energía fue el diseño 16, con una identidad del 63% con respecto al Rossmann de SDH. En estos diseños se permitió la variación en secuencia de absolutamente todos los sitios, siempre y cuando la estructura tridimensional fuese según RosettaDesign la del dominio Rossmann de SDH a nivel de esqueleto. El tiempo invertido para generar a las 999 secuencias fue de 144.6 horas utilizando una computadora con un procesador Intel® Core™2 Quad CPU Q6600 @ 2.40GHz × 4 con 4 Gb de memoria RAM.

En la tabla 14 se hacen las comparaciones en cuanto a porcentaje de identidad de las secuencias de los distintos diseños antes descritos entre sí. Los porcentajes de identidad con respecto al dominio silvestre son más altos de lo que se esperaría debido a que nunca se excluyó la posibilidad de incluir al aminoácido original en los sitios donde se permitió la variación, puesto que se consideró que era muy posible que, por motivos de satisfactibilidad, sólo el residuo original fuera el único posible para una estructura del esqueleto equivalente a la del templado en una posición específica.

[redacted]	SDH (natural)	SCA	AntiSCA
SCA	55	[redacted]	[redacted]
AntiSCA	87	42	[redacted]
Aleatorio	63	69	43

Tabla 14. Porcentajes de identidad entre las secuencias de las diversas moléculas diseñadas más estables.

Era de esperarse que el diseño AntiSCA fuera quien más se pareciera al dominio silvestre, ya que sólo se variaron los 30 residuos acoplados y los 10 muy conservados; es decir, 40 residuos de 144 totales.
Sin embargo, a un nivel más detallado de análisis entre secuencias, se encontró que la secuencia del diseño AntiSCA y el Aleatorio comparten 24 residuos idénticos a pesar de pertenecer a conjuntos de diseños distintos. De estos 24 residuos compartidos, 16 de ellos son idénticos a las secuencia original del dominio Rossmann de SDH, y de estos últimos residuos, 11 pertenecen a los sitios definidos como estadísticamente acoplados y los 5 restantes a los muy conservados.

De acuerdo al reporte de Martínez-Castilla & Rodríguez-Sotres en 2010, el hecho de que exista cierta convergencia entre las secuencias silvestre y los diseños de RosettaDesign mencionados no es de llamar la atención, ya que como ellos proponen, el esqueleto de la

proteína templado puede ser informacionalmente suficiente para capturar la esencia del plegado e incluso de la función original al momento de generar poblaciones de secuencias distintas que de acuerdo al programa adquieran una estructura tridimensional igual a la del templado. Este efecto tiene la peculiaridad de que modelos ocultos de Markov generados a partir de los alineamientos de las secuencias diseñadas guiadas por un mismo templado son capaces de capturar a la secuencia templado y homólogos cercanos al buscar en bases de datos del NCBI.

Ahora bien, dado que Martínez-Castilla & Rodríguez-Sotres obtuvieron esos resultados con secuencias a las que se les permitió variar de forma aleatoria todos y cada uno de los residuos de las secuencias de diseño (la única condición es que cada secuencia en su totalidad se plegara como el templado de acuerdo a RosettaDesign), el hecho de que en este trabajo coincidieran ciertos sitios entre el diseño AntiSCA y el natural no es del todo sorprendente. Aun así, lo que sí es sorprendente es que precisamente los 16 residuos que comparten las 3 secuencias (Silvestre, AntiSCA y Aleatoria) puedan dividirse (sin excepción) en dos grupos: estadísticamente acoplados (G103, G105, L110, G119, L123, L138, A155, A159, F181, M213 y W225) y muy conservados (G126, G128, T188, D212 y G237). De hecho, de los residuos estadísticamente acoplados, sólo M213 tiene contacto con el ligando, mientras que los muy conservados se concentran en el centro y a excepción de D212, todos los demás tienen contacto con el ligando. Ver figura 18.

Se pensó que la importancia de estos residuos pudiera verse reflejada en que sus valores fueran más extremos dentro de los conjuntos de los cuales se obtuvieron, sin embargo, al observar los valores de los 11 residuos estadísticamente acoplados se encontró que éstos no se localizan como los más extremos dentro de la distribución *t* a partir de la cual se definió cada sector.

Con respecto a los 5 residuos muy conservados compartidos entre las 3 secuencias referidas en este apartado, éstos tampoco presentan un valor de entropía extrema dentro del grupo de los 13 originalmente considerados como muy conservados (Ver figura 18).

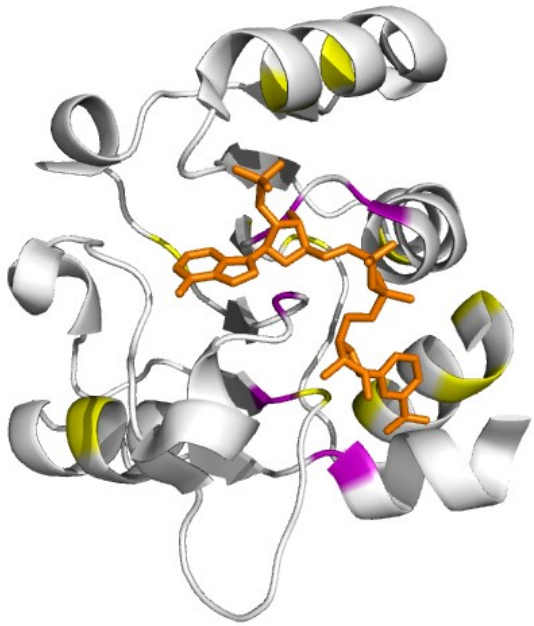


Figura 18. Residuos coincidentes entre el Rossmann silvestre de SDH y los diseños Aleatorio y AntiSCA. En amarillo se muestran los 11 residuos estadísticamente acoplados, en magenta los 5 fuertemente conservados y en sticks color naranja el cofactor NADP.

Dinámica Molecular

Dado que SCA puede determinar aquellos residuos del dominio Rossmann de la SDH más importantes, se esperaba que un diseño *in silico* donde se respetasen tales residuos tuviera una movilidad más parecida a la del dominio silvestre que diseños donde no se tomó en cuenta tal restricción. Es por eso que se decidió someter a dinámica molecular en algoritmo de “recocido simulado” (simulated annealing) al dominio Rossmann de la SDH y lo siguiente: un diseño donde se respetaron únicamente los sitios estadísticamente acoplados y los muy conservados (SCA), un diseño donde sólo se variaron los sitios acoplados y los muy conservados (AntiSCA) y otro donde se variaron todos los residuos del dominio (Aleatorio). Se realizaron las dinámicas bajo el protocolo de ensamble NPT (moles, presión y temperatura constantes) ya que este corresponde de forma más cercana a las condiciones de laboratorio. También se hicieron las dinámicas usando el protocolo NVT (moles, volumen y temperatura constantes) con fines comparativos.

Posteriormente, se procedió a comparar el RMSD de cada proteína con respecto al estado inicial a lo largo del tiempo (figuras 19, 20 y 21) para corroborar que la movilidad de cada modelo se comportara de acuerdo a lo esperado, es decir, que el diseño SCA se pareciera más al Silvestre, mientras que el AntiSCA y Aleatorio se parecieran poco. De hecho, se esperaba que el diseño AntiSCA se pareciera mucho menos en analogía a los resultados obtenidos por Russ y colaboradores en el 2005 cuando se probaron dominios artificiales WW basados en la información de SCA.

Sin embargo, al observar los videos de las dinámicas, ni el Rossmann Silvestre ni los diseños se desnaturalizaron a lo largo de los 50 ns de simulación. Aún más, a excepción del diseño Aleatorio, ninguno de ellos mostró señales de perder al NADP.

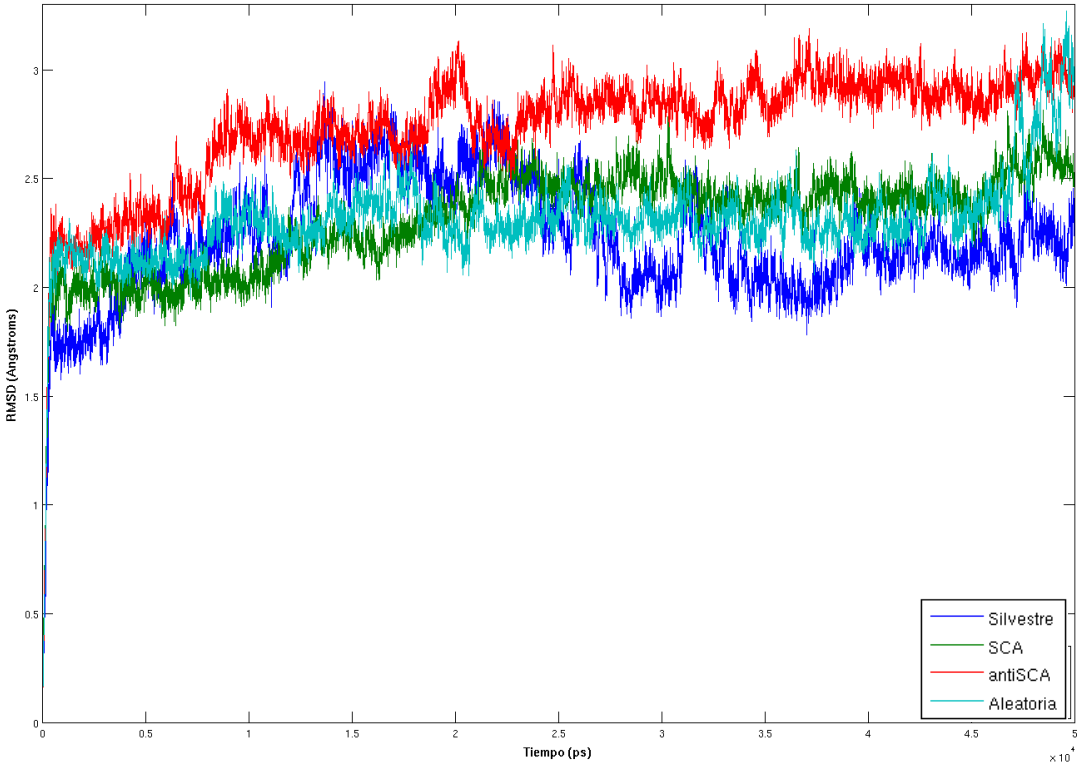


Figura 19. RMSD en NPT. Se sobrelapan los diversos espectros de RMSD (Root Mean Square Deviation) de las diferentes moléculas simuladas.

En la figura 19 se puede apreciar que el diseño AntiSCA es quien tiende a alejarse más en su conformación con respecto al estado inicial y en general su movimiento es mayor que el de cualquier otro caso, sin embargo, aunque el comportamiento del diseño SCA es similar al del Silvestre, el Aleatorio es aún más similar. Con la mira a corroborar de una mejor forma esta observación, se procedió a graficar la diferencia absoluta de RMSD por punto entre cada diseño y el caso Silvestre, generando de esta manera la figura 20.

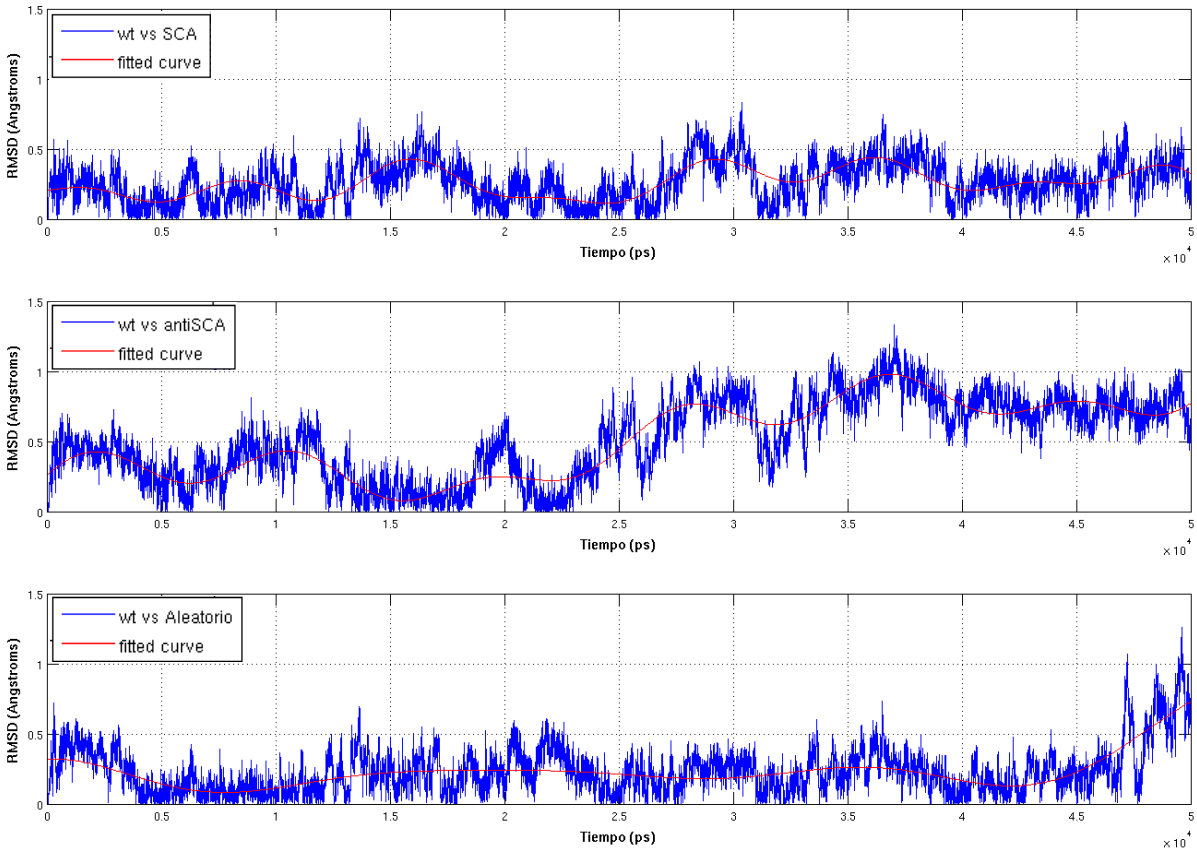


Figura 20. Silvestre vs todas NPT. Diferencias absolutas con respecto al RMSD del dominio Rossmann silvestre. En cada caso se muestra en rojo la linea de tendencia de cada "espectro absoluto".

En la figura 20 un valor más cercano a cero en el eje Y indica una mayor similitud en cuanto a movilidad en los casos comparados. Nuevamente, el espectro de la comparación entre el Rossmann Silvestre y el diseño Aleatorio muestra una mayor tendencia a mantenerse cercano al 0 que los demás casos.

Para ser aún más explícitos en la forma de presentar las diferencias absolutas de RMSD por punto entre los diseños y el caso Silvestre, se graficó únicamente la linea de tendencia de la figura anterior, dando lugar a la figura 20.

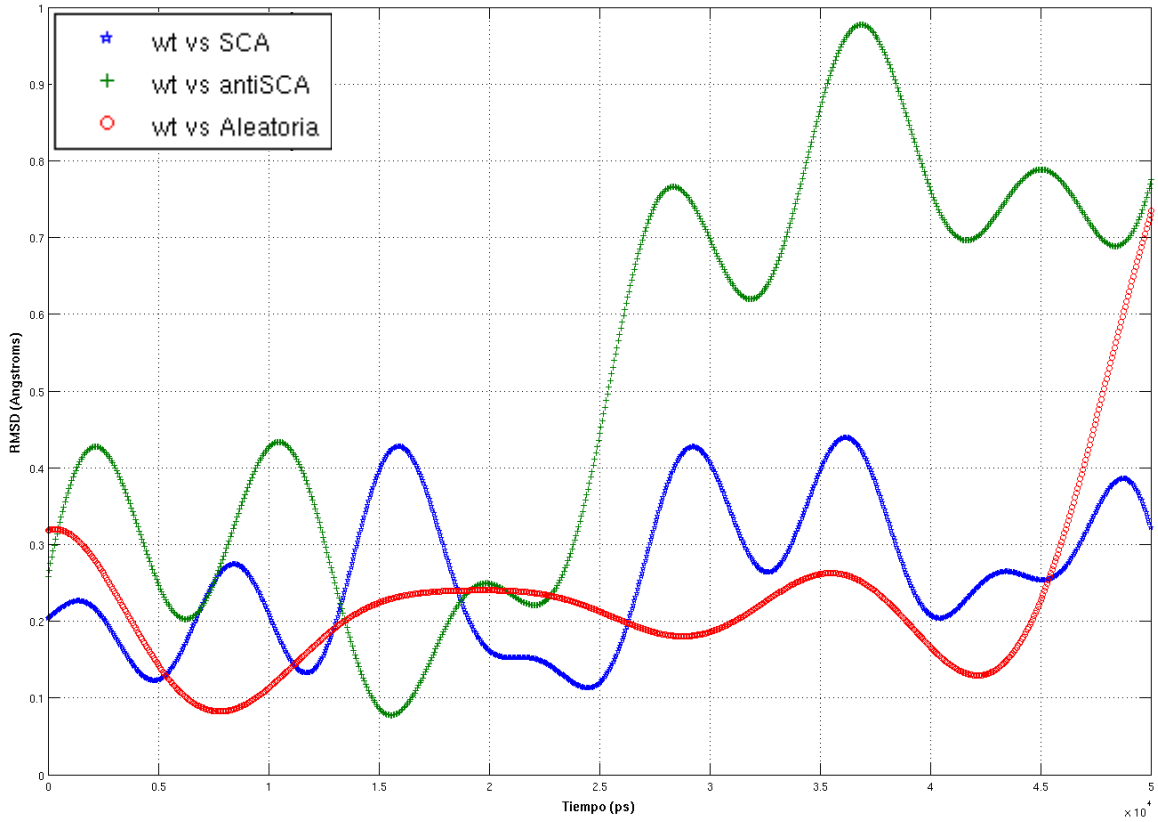


Figura 21. Curvas de ajuste con respecto a la dinámica silvestre NPT. Se muestran sólo las líneas de tendencia de la gráfica anterior.

Una vez más, por la figura 21 es más claro el hecho de que el diseño aleatorio (curva roja) es quien se mueve de forma más parecida al dominio Rossmann de la SDH de *E. coli* a lo largo del tiempo que cualquiera de las otras moléculas diseñadas, aunque a los 45 nanosegundos esa diferencia se eleva significativamente. Por otra parte, el diseño SCA , en azul, aunque muestra diferencias fluctuantes a lo largo del tiempo, éstas siempre son menores a 0.45 Å; similitud que no es despreciable.

De acuerdo a lo esperado, el caso más pronunciado en cuanto a desviación del dominio natural es el diseño AntiSCA, en verde, quien muestra mayor desviación general desde el principio, disparándose a partir de los 25 nanosegundos de simulación.

Por otra parte, al observar los videos de cada simulación en protocolo NPT, es precisamente el diseño Aleatorio quien comienza a desprenderse del cofactor. De hecho, la parte de la adenina es quien desde el principio comienza a despegarse y al final está completamente fuera de su posición inicial, por lo que la molécula de NADP queda en una conformación donde sólo la nicotinamida y su ribosa son las únicas partes que siguen unidas al dominio Rossmann. Es posible que de haber extendido el tiempo de la simulación el NADP se desprendiera por completo.

Para terminar con la sección de dinámica molecular, se comparó la movilidad por diseño entre protocolos en la figura 22.

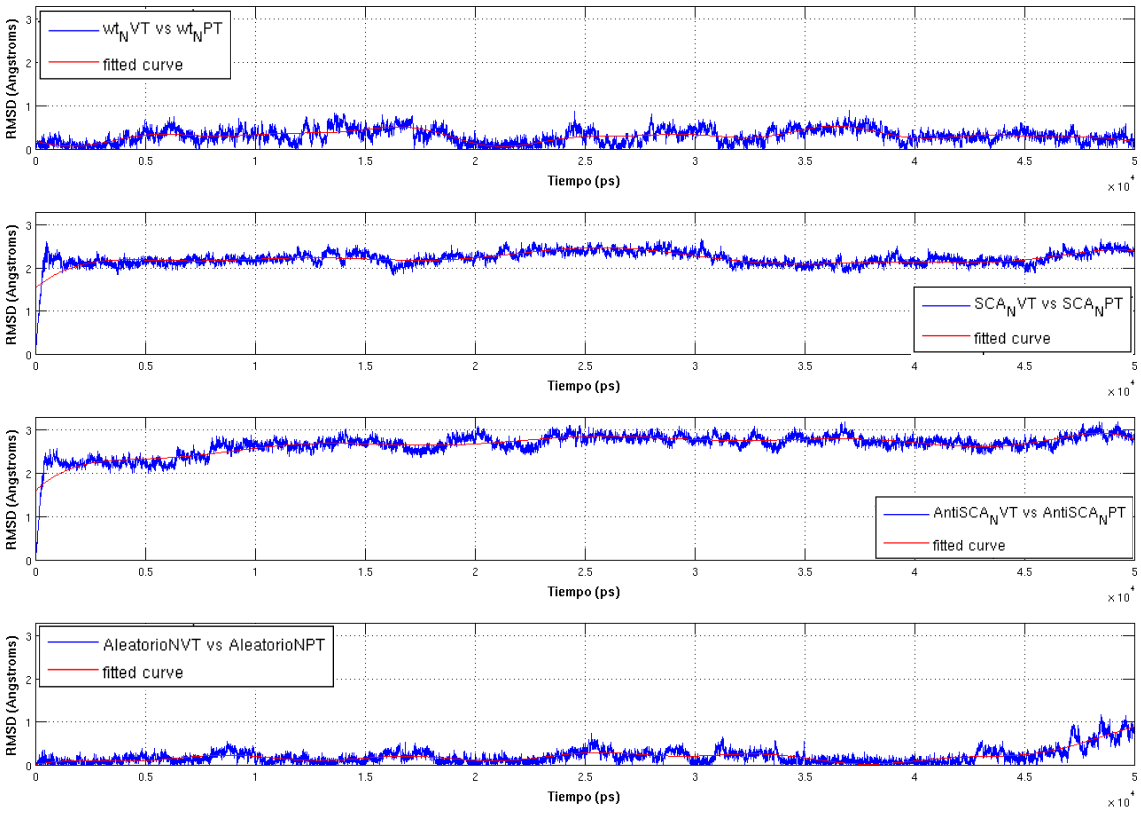


Figura 22. Diferencia entre NVT y NPT. Espectros de la diferencia absoluta entre ambos protocolos en una misma molécula.

Lo inesperado en el caso de la figura 22, es que sólo los diseños SCA y AntiSCA muestran diferencias significativas a lo largo del tiempo al comparar ambos protocolos. Parece ser que la permisividad en cuanto a la variación del volumen permite una mayor flexibilidad sólo en estos diseños. No podemos asumir que esa diferencia se deba a que una variación en el volumen no forme parte de un escenario natural ante el cual la evolución no ha podido actuar, puesto que el dominio Rossmann Silvestre no muestra grandes diferencias (no llega nunca a 1 Å de diferencia a lo largo del tiempo que comprendió la simulación), así como tampoco podemos atribuir ese comportamiento a que los diseños (necesariamente no naturales) sean más propensos a moverse de diferente forma ante protocolos distintos porque el diseño Aleatorio tampoco presenta grandes diferencias (las más grandes están alrededor de 1 Å) a lo largo del tiempo. En este último respecto sería interesante analizar dinámicas temporalmente más extensas.

Con el fin de saber si a nivel poblacional los conjuntos de diseños de donde se obtuvieron la moléculas que se simularon mantenían una estabilidad coherente con lo observado durante la dinámica molecular, se procedió a graficar en la figura 23 las calificaciones energéticas de cada conjunto. Así, al conjunto en donde se respetaron

únicamente los sitios estadísticamente acoplados y los muy conservados se le llamó conjunto SCA, aquel de donde obtuvimos el diseño AntiSCA lo llamamos conjunto AntiSCA y finalmente siguiendo la misma temática tenemos al conjunto Aleatorio.

Adicionalmente, se realizaron nuevas rondas de diseño de cada conjunto, pero de una forma completamente distinta para comparar con los datos originales. Por una parte se escogieron 40 residuos al azar del dominio Rossmann de la SDH y se armó un conjunto de 100 diseños al que simbólicamente llamamos “SCA-toy” en donde se respetarían estos residuos. Este conjunto poco o nada tiene que ver con el conjunto SCA original, ya que la proporción de residuos escogidos que coincidieran con los estadísticamente acoplados no es superior a la esperada al azar (alrededor de 20).

Análogamente, se hicieron los “contradiseños” del conjunto anterior, es decir; se armó un conjunto “AntiSCA-toy” con la restricción de únicamente variar aquellos residuos que se respetaron para generar el conjunto SCA-toy.

Finalmente, se generó un conjunto Aleatorio-toy, para el cual se permitió variar a todos los residuos del Rossmann de la SDH.

Para tener más información con la cual contrastar, se hizo otro conjunto de 100 diseños como el SCA-toy pero con otros residuos respetados al azar, este nuevo conjunto fue llamado SCA-toy2, mientras que el conjunto de contradiseños fue llamado AntiSCA-toy2.

Dado que el conjunto Aleatorio-100 sólo era de control y demandaba decenas de horas de cómputo, decidimos no generar nuevos.

Por último, para explorar más espacio de búsqueda, algo adecuado sería generar 100 conjuntos de alrededor de 100 diseños cada uno donde se respetaran al azar 40 residuos por conjunto y sus respectivos contradiseños. Sin embargo, dado que generar tantas secuencias (10000 en total) nos consumiría mucho tiempo, se optó por generar 100 diseños donde se respetaran 40 residuos al azar distintos para cada diseño. A esta nueva población se le llamó “SCA-toyPOP”, mientras que por cada uno de estos diseños se generó el diseño complementario para generar a la población “AntiSCA-toyPOP”. La distribución energética de todos estos diseños se condensa en la figura 23.

A)

B)

C)

D)

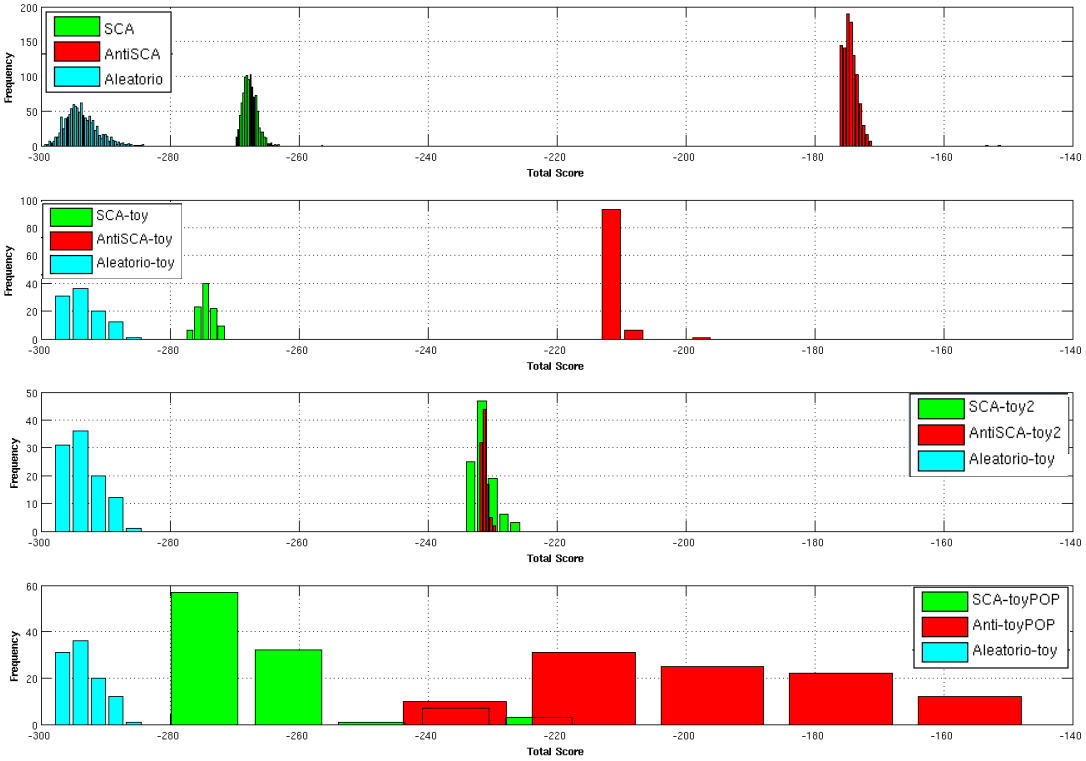


Figura 23. Distribución energética de los diseños en RosettaDesign. En todos los casos se muestran histogramas de la calificación total (Score) de los diferentes diseños. En A) los conjuntos SCA, AntiSCA y Aleatorio son de 999 elementos cada uno, mientras que para B), C) y D) todos los conjuntos son de únicamente 100 elementos. Un diseño es energéticamente más estable entre más baja sea su calificación total.

De acuerdo a lo que se presenta en la figura 23A, los conjuntos SCA, AntiSCA y Aleatorio ocupan rangos muy bien definidos, siendo la estabilidad del conjunto SCA cercana a la alcanzada por los elementos del conjunto Aleatorio, mientras que como era de esperarse, el conjunto AntiSCA es quien presenta las peores calificaciones y se encuentra muy alejada de las demás. Aunque en B) y en D) la posición sea similar a la de A), la diferencia entre el conjunto SCA-toy# y el AntiSCA-toy# no es tan grande como en sus homólogos de A), de hecho el experimento realizado con los conjuntos SCA-toyPOP y AntiSCA-toyPOP da como resultado distribuciones amplias que ya comienzan a sobrelaparse, mientras que en C) las poblaciones SCA-toy2 y AntiSCA-toy2 están definitivamente sobrelapadas siendo la segunda mucho más delgada. Es muy posible que la forma que adquieren las distribuciones y las posiciones que asumen en los incisos B), C) y D) se deba a un mero efecto de proporción de residuos respetados dependiendo únicamente de la forma de operar RosettaDesign. Concluimos que tal diferencia entre A) y los demás incisos de la figura 23 se debe a los residuos específicos en respetarse, es decir, que los residuos estadísticamente acoplados son cruciales para la movilidad que da lugar a la función gracias a la estabilidad estructural de la proteína.

DISCUSIÓN

Al comparar entre dos análisis de una misma proteína completa (los casos donde se obtuvo un sector contra los de tres sectores) las diferencias en cuanto a residuos detectados son visiblemente escasas. Caso similar en cuanto a los análisis de sus dominios Rossmann a partir de alineamientos podados y no podados.

En todos los casos donde se analizó a la proteína completa se obtuvieron residuos estadísticamente acoplados en ambos dominios. En la mayoría de los casos, tales residuos acoplados de un mismo sector se distribuyen homogéneamente entre ambos dominios; por lo que el caso particular de 1BW9 llama la atención al concentrar al sector rojo en el dominio Rossmann y al sector verde en el catalítico; teniendo al azul distribuido alrededor del centro de la enzima. En analogía a los hallazgos presentados por Halabi y colaboradores en el 2009 para las serina proteasas S1A, es posible que tal distribución de sectores en nuestra fenilalanina deshidrogenasa ya esté delatando que el sector rojo tenga que ver con el plegamiento y/o función (unión a cofactor) del dominio Rossmann en la geometría C-R, el plegamiento y/o función (unión a sustrato) del dominio catalítico mediado en gran manera por el sector verde, mientras que el sector azul tenga una tarea ya directa en la catálisis.

En algunas proteínas existen residuos acoplados en algunas de las β centrales (en especial al final de las mismas) del dominio Rossmann. Esta presencia de acoplamiento es más pronunciada en los análisis de dominio Rossmann únicamente, sea este podado o no podado. Tal observación de acoplamiento al final de las β centrales coincide con la de los análisis que Bhattacharyya y colaboradores realizaron en redes en el 2012; así como también el hecho de que se encuentran muchos residuos importantes en las asas de la proteína y otras regiones accesibles al solvente en todos los casos estudiados y que ellos propusieron como *fold-specific hot spots*.

Por otra parte, es de particular interés la coincidencia en la información reportada en la literatura para diversos residuos de cada proteína analizada en este trabajo y aquellos reconocidos como significativos de acuerdo al análisis de SCA. Más que nada, la coincidencia casi absoluta de los sitios que tienen contacto con el cofactor según www.pdb.org y los estadísticamente acoplados, quienes adicionalmente presentan enlaces de hidrógeno entre sí y en ocasiones otro tipo de interacciones débiles, como las hidrofóbicas o las fuerzas de van der Waals, que en conjunto delatan la importancia de nuestros residuos significativos para armar una telaraña cooperativa que puede ser importante dentro de la estructura y/o funcionalidad de cada proteína. A este respecto, vale la pena comentar la imperiosa necesidad de la movilidad intrínseca de la enzima, mencionada en múltiples ocasiones en asas y en el contexto del cambio de conformación entre estado abierto y cerrado para la SDH mediada por un rearreglo de enlaces de hidrógeno. Esta movilidad está íntimamente ligada con la función de la enzima, y de acuerdo a Russ y colaboradores en 2005, es propiamente la función y no tanto la estructura de una proteína quien está sometida a selección, por lo que los análisis de dinámica molecular de un dominio Rossmann Silvestre y de distintos diseños de RosettaDesign nos pareció adecuado.

En otros contextos, SCA nos ha demostrado la capacidad del análisis coevolutivo para capturar, a partir de la información lineal plasmada en los MSA residuos catalíticos (como el D102 en la SDH, así como M62, H95, E117, Y236 y D269 en la alanina deshidrogenasa), afinidad por algún ligando (I94, I116 y M54 para el cofactor en lactato deshidrogenasa y R150

como un sospechoso en cuanto a la capacidad de discriminar entre NAD y NADP en SDH, así como Y215 para orientar al sustrato dentro de esta última enzima) , oligomerización para la estructura cuaternaria (caso de la alanina deshidrogenasa mediada por K184, G161, R138, P163, G164, V165, P167, K169 y E194) e incluso localización dentro de una célula, refiriéndonos particularmente al caso en el que los residuos I132,S133, A134, A136, S137, G139, A140, K141, T142 y F143 forman parte del segmento responsable de que la gliceraldehído 3-fosfato deshidrogenasa analizada se encuentre en el glicosoma y no en el citosol (ver sección de Resultados).

Por lo anterior, podemos tener una mayor confiabilidad en que el enfoque de análisis coevolutivo basado en estadística fina sobre la información lineal contenida en secuencias de proteínas es una buena herramienta para anteceder los análisis experimentales.

Retomando los análisis de dinámica molecular, es evidente que en movilidad quien más se aleja del comportamiento del dominio Rossmann silvestre es el diseño AntiSCA, mientras que el diseño Aleatorio es más constante en su movimiento hasta casi el final del tiempo de simulación, ya que alrededor de los 47 nanosegundos en protocolo NPT dispara su movilidad oscilando entre 2.5 y 3 Å (ver figura 21). La constancia del movimiento en el diseño aleatorio debe obedecer a la forma de operar de RosettaDesign en cuanto a su tendencia de estabilizar a la molécula diseñada, y dado que en este diseño no pusimos ninguna restricción salvo el hecho de respetar la estructura del templado, obtuvimos una proteína muy estable, por lo que el hecho de que su movilidad se parezca más al diseño silvestre no significa más que una simple casualidad producto de la forma de operar de RosettaDesign. Adicionalmente a esto, es importante mencionar que el tiempo de cómputo invertido en generar 999 diseños completamente aleatorios es aproximadamente 2 veces mayor que el invertido para generar los 999 diseños del conjunto SCA y hasta 8 veces con respecto al utilizado para el conjunto AntiSCA; por lo que es obvio que RosettaDesign trabajará más rápido entre más residuos le indiquemos respetar; y dado que aun en un diseño Aleatorio, cuya determinación toma valioso tiempo de cómputo existe una tendencia a converger en ciertos sitios con el conjunto de los residuos estadísticamente acoplados, darle la información *a priori* de cuáles son los residuos que debe respetar nos encamina a una forma más eficiente de hacer diseño de proteínas.

CONCLUSIÓN

A la luz de lo obtenido y discutido a lo largo de este trabajo tenemos suficientes motivos para pensar que un diseño basado en la información adquirida a partir del SCA es más eficiente y presenta mayores oportunidades de funcionar en comparación con otro donde no se realicen análisis sobre la coevolución de los residuos de la proteína templado, ignorándose de esta forma las redes de cooperación entre sitios sobre las que ha actuado la selección natural. Dado que la funcionalidad del diseño depende necesariamente de la estabilidad estructural de la molécula, la exploración del movimiento atómico *in silico* por medio de dinámica molecular ofrece una estrategia lógica previa a los análisis experimentales, dando lugar a un panorama que ahorre tiempo y recursos de laboratorio. Esto nos encamina con mayor seguridad a la meta del diseño exitoso de proteínas.

PERSPECTIVAS

1. La síntesis de los genes correspondientes a los diseños y su expresión serían de mucha utilidad para confirmar o refutar las ideas ya planteadas con respecto a los resultados obtenidos en este estudio. En particular con la fenilalanina deshidrogenasa, la cual, como se comentó en la discusión, presenta sectores muy bien localizados en cada dominio (rojo en el Rossmann, verde en el catalítico y azul en el centro), lo cual pudiera indicar papeles bioquímicos fáciles de identificar.
2. Sería interesante en caso de obtener dominios Rossmann plegados y estables (y adicionalmente que puedan unir al ligando) hacer mutaciones de sitios acoplados y así determinar su importancia en el laboratorio.
3. Con respecto al punto anterior, mutaciones en sitios correspondientes a distintos sectores revelarían cual es el papel de cada uno de ellos *in silico*. Nuevamente, enfatizamos que, además de los diseños con el Rossmann de la SDH, los sitios de los sectores de la fenilalanina deshidrogenasa convierten a esta última enzima en un buen modelo de estudio.
4. Sería interesante confirmar si aquellos residuos convergentes entre los diseños AntiSCA, Aleatorio y la secuencia original son críticos a nivel experimental haciendo algunas mutaciones puntuales.

REFERENCIAS:

1) Abad-Zapatero, C., Griffith, J. P., Rossmann, M. G., Lafayette, W., Sussman, J. L., Rossmann, M. G. (1987). Refined Crystal Structure of Dogfish M ,. JMB, Vol. (198): 445–467.

2) Agren, D., Stehr, M., Berthold, C. L., Kapoor, S., Oehlmann, W., Singh, M., Schneider, G. (2008). Three-dimensional structures of apo- and holo-L-alanine dehydrogenase from Mycobacterium tuberculosis reveal conformational changes upon coenzyme binding. JMB, Vol. (377) Issue (4):1161-73. doi: 10.1016/j.jmb.2008.01.091

3) Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. Science. Vol. (181) Issue (4096):223-30. doi::10.1126/science.181.4096.223

4) Baker, P. J., Sawa, Y., Shibata, H., Sedelnikova, S. E., Rice, D. W. (1998). Analysis of the structure and substrate binding of Phormidium lapideum alanine dehydrogenase. Nature Structural Biology. Vol. (5) Issue (7):561-7. doi:10.1038/817

5) Banda, J. (2011). Análisis de proteínas quiméricas entre AroE y YdiB de Escherichia coli. (Tesina de Licenciatura).

6) Bashton, M., & Chothia, C. (2002). The geometry of domain combination in proteins. JMB, Vol. (315) Issue (4): 927–39. doi:10.1006/jmbi.2001.5288

7) Bhattacharyya, M., Upadhyay, R., & Vishveshwara, S. (2012). Interaction signatures stabilizing the NAD(P)-binding Rossmann fold: a structure network approach. PloS one, 7(12), e51676. doi:10.1371/journal.pone.0051676

8) Bottoms, C. A., Smith, P. E., Tanner, J. J. (2002). A structurally conserved water molecule in Rossmann dinucleotide-bindig domains. Protein Science: a publication of the Protein Society., Vol. (11) Issue (9):2125-37. doi:10.1110/ps.0213502

9) Dill, K. a, & MacCallum, J. L. (2012). The protein-folding problem, 50 years on. Science, Vol. (338) Issue (6110): 1042–6. doi:10.1126/science.1219021

10) Eventoff, W., Rossmann, M. G., Taylor, S. S., Torff, H. J., Meyer, H., Keil, W., & Kiltz, H. H. (1977). Structural adaptations of lactate dehydrogenase isozymes. Proceedings of the National Academy of Sciences of the United States of America, Vol. (74) Issue 7: 2677–81. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=431242&tool=pmcentrez&rendertype=abstract

11) Goihberg, E., Peretz, M., Tel-or, S., Dym, O., Shimon, L., Frolow, F., Burstein, Y. (2010). Biochemical and structural properties of chimeras constructed by exchange of cofactor-binding domains in alcohol dehydrogenases from thermophilic and mesophilic microorganisms. Biochem., Vol. (49) Issue (9): 1943-1953. doi:10.1021/bi901730x

12) Hadley, C., Jones, D. T. (1999). A systematic comparison of protein structure classsifications: SCOP, CATH and FSSP. Cell, Vol. (7):1099-1112.

13) Halabi, N., Rivoire, O., Leibler, S., & Ranganathan, R. (2009). Theory Protein Sectors : Evolutionary Units of Three-Dimensional Structure. Cell, Vol. (138) Issue (4): 774–786. doi:10.1016/j.cell.2009.07.038

14) Han, J.H., Batey, S., Nickson, A. A., Teichmann, S. A., Clarke, J. (2007). The folding and evolution of multidomain proteins. Nature reviews: Molecular Cell Biology., Vol. (8) Issue (4):319-30. doi:10.1038/nrm2144

15) Henzler-Wildman, K. & Kern, D. (2007). Dynamic personalities of proteins. Nature. Vol. (450) Issue (7172):964-72. doi:10.1038/nature06522

16) Holland, L. Z., McFall-Ngai, M., Somero, G. N. (1997). Evolution of lactate dehydrogenase-A homologs of barracuda fishes (genus Sphyraena) from different thermal environments: differences in kinetic properties and thermal stability are due to amino acid substitutions outside the active site. Biochemistry. Vol. (36) Issue (11):3207-15. doi: 10.1021/bi962664k

17) Janin, J. & Wodak, S. J. (1983). Structural domains in proteins and their role in the dynamics of protein function. Prog. Biophys. Mol. Biol., Vol. (42) Issue (1):21-78.

18) Jez, J. (2011). Toward Protein Engineering for Phytoremediation: Possibilities and Challenges. International Journal of Phytoremediation, Vol. (13) Issue (S1):77-89. doi:10.1080/15226514.2011.568537

19) Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H., & Meiler, J. (2010). Practically useful: what the Rosetta protein modeling suite can do for you. Biochemistry, Vol. (49) Issue (14): 2987–98. doi:10.1021/bi902153g

20) Kataoka, K. & Takada, H. (1994). Construction and characterization of chimeric enzyme consisting of an amino-terminal domain of phenylalanine dehydrogenase and a carboxy-terminal domain of leucine dehydrogenase. Biochem. Vol. (936): 931-936.

21) Kiss, G., Röthlisberger, D., Baker, D., Houk, K. N. (2010). Evaluation and ranking of enzyme designs. Protein Science, Vol. (19). Issue (9): 1760-1773. doi:10.1002/pro.462

22) Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T., Baker, D. (2012). Principles for designing ideal protein structures. Nature, Vol. (491): 222-229. doi:10.1038/nature11600

23) Lambeir, A. M., Loiseau, A. M., Kuntz, D. A., Vellieux, F. M., Michels, P. A. M. (1991). The cytosolic and glycosomal glyceraldehyde-3-phosphate dehydrogenase from Trypanosoma brucei: Kinetic properties and comparison with homologous enzymes. Eur. J. Biochem. Vol. (198):429-435.

24) Lockless, S. W., & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. Science (New York, N.Y.), Vol. (286) Issue (5438): 295–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10514373

25) Martínez-Castilla, L. P., & Rodríguez-Sotres, R. (2010). A Score of the Ability of a Three-Dimensional Protein Model to Retrive Its Own Sequence as a Quantitative Measure of Its Quality and Appropriateness. PLoS ONE, Vol. (5) Issue (9): e12483, 1-19. doi:10.1371/journal.pone.0012483

26) McLaughlin, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S., & Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature, Vol. (491) Issue(7422): 138–42. doi:10.1038/nature11500

27) Michel, G., Roszak, A., Sauvé, V., Maclean, J., Matte, A., Coggins, J., Cygler, M., et al. (2003). Structures of Shikimate Dehydrogenase AroE and Its Paralog YdiB. Biochemistry, Vol. (278) Issue (21): 19463–19472. doi:10.1074/jbc.M300794200

28) Michels, P. a, Marchand, M., Kohl, L., Allert, S., Wierenga, R. K., & Opperdoes, F. R. (1991). The cytosolic and glycosomal isoenzymes of glyceraldehyde-3-phosphate dehydrogenase in Trypanosoma brucei have a distant evolutionary relationship. European journal of biochemistry / FEBS, Vol. (198) Issue (2): 421–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2040303

29) Orengo, C. a, & Thornton, J. M. (2005). Protein families and their evolution-a structural perspective. Annual review of biochemistry, Vol. (74): 867–900. doi:10.1146/annurev.biochem.74.082803.133029

30) Peek, J., Lee, J., Hu, S., Senisterra, G., & Christendat, D. (2011). Structural and mechanistic analysis of a novel class of shikimate dehydrogenases: evidence for a conserved catalytic mechanism in the shikimate dehydrogenase family. Biochemistry, Vol. (50) Issue (40): 8616–27. doi:10.1021/bi200586y

31) Ponting, C. P. & Russell, R. R. (2002). The natural history of protein domains. Annual Review of Biophysics and Biomolecular Structure, Vol. (31):45-71. doi:10.1146/annurev.biophys.31.082901.134314

32) Reynolds, K. A., Mclaughlin, R. N., & Ranganathan, R. (2011). Hot Spots for Allosteric Regulation on Protein Surfaces. Cell, Vol. (147) Issue (7): 1564–1575. doi:10.1016/j.cell.2011.10.049

33) Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S., Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. Nature, Vol. (453): 190-195. doi:10.1038/nature06879

34) Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B., Ranganathan, R. (2005). Natural-

Like function in artificial WW domains. Nature, Vol. (437) Issue (7058): 579-583. doi:10.1038/nature03990

35) Segatorio, L. Paukstelis, P. J., Gilbert, H. F., Georgiou, G. (2004). Engineered DsbC chimeras catalyze both protein oxidation and disulfide-bond isomerization in *Escherichia coli*: Reconciling two competing pathways. PNAS of USA, Vol. (101) Issue (27): 10018-10023.

36) Smock, R. G., Rivoire, O., Russ, W. P., Swain, J. F., Leibler, S., Ranganathan, R., & Gierasch, L. M. (2010). An interdomain sector mediating allostery in Hsp70 molecular chaperones. Molecular Systems Biology, Vol. (6) Issue (414): 1–10. doi:10.1038/msb.2010.65

37) Sharkey, M. A. & Engel, P. C. (2009). Modular coenzyme specificity: a domain-swopped chimera of glutamate dehydrogenase. Proteins, Vol. (77) Issue (2):268-278. doi: 10.1002/prot.22433.

38) Vanhooke, J. L., Thoden, J. B., Brunhuber, N. M., Blanchard, J. S., & Holden, H. M. (1999). Phenylalanine dehydrogenase from Rhodococcus sp. M4: high-resolution X-ray analyses of inhibitory ternary complexes reveal key features in the oxidative deamination mechanism. Biochemistry, Vol. (38) Issue (8): 2326–39. doi:10.1021/bi982244q

39) Vincent, S. J., Zwahlen, C., Post, C. B., Burgner, J. W., Bodenhausen, G. (1997). The conformation of NAD+ bound to lactate dehydrogenase determined by nuclear magnetic resonance with suppression of spin diffusion. PNAS of USA. Vol. (94) Issue (9):4383-8.

40) Thulasiram, H. V., Erickson, H. K., Poulter, C. D. (2007). Chimeras of two isoprenoid synthases catalyze all four coupling reactions in isoprenoid biosynthesis. Science. Vol. (316) Issue (5821):73-76.

Material Suplementario

COMANDOS PARA MATLAB

Las líneas que comienzan con % son interpretados como comentarios. Para reducir la redundancia, en este apartado sólo se usarán en los comandos referentes al caso de la obtención de un único sector y de tres sectores en la SDH de *E. coli*. El algoritmo y la mayor parte de los comentarios fueron tomados de la libreria de SCA5 que se encuentra disponible en http://systems.swmed.edu/rr_lab/

Se decidió dejar la coloración original para facilitar la interpretación visual entre comentarios, instrucciones, variables, etcétera.

El resto de los scripts (archivos *.m) comentados para cada uno de los análisis realizados para este trabajo, se encuentran en la carpeta comprimida que acompaña a la tesis en la base de datos de la biblioteca de la UNAM.

%%SDH, un sólo sector al 0.8 como punto de corte. Dentro de la SDH, éste script es aplicable a cualquier punto de corte siempre y cuando se éste el único parámetro que se altere.

```
% Tutorial de SCA 5.0 para analizar un alineamiento de homólogos de la SDH.

addpath sca5
clear; close all

%% Paso 1. Cargando y acondicionando el alineamiento

% En este caso, cargamos un alineamiento que se encuentra en formato libre (*.ran)

[labels_seq,algn_full]=get_seqs('/home/jbanda/Desktop/Deshidro_rangas_090112/1nyt_80.ran');
N_seq=size(algn_full,1);

% Una estrategia práctica es truncar al alineamiento a posiciones de secuencia
% con una frecuencia de gaps no mayor a 20%. Esto previene la sobrerrepresentación
% trivial de gaps en el alineamiento y asegura que los cálculos se realicen únicamente en
% posiciones que en su mayoría no contengan gaps.

cut_off=.2;
fraq_gaps=sum(isletter(algn_full)==0)/N_seq;
algn=algn_full(:,fraq_gaps<cut_off);
N_pos=size(algn,2);


% Es muy útil tener a las posiciones de los residuos numeradas de acuerdo a
% un miembro específico de la familia de proteínas (más que por numeración
% en el alineamiento). Ésto para facilitar el mapeo de los datos de correlación
% con los de la estructura de la proteína (si es que ya se ha elucidad la estructua
% por algún otro método. Así, dado un archivo de coordenadas PDB
% es posible alinear la secuencia correspondiente con cada miembro
% del alineamiento y de esa forma obtener la numeración de acuerdo al mejor
% resultado. Otra opción es indicar a priori
% que sea una secuencia particular en el alineamiento la que sea usada como guía.
```

```
pdb_id='1NYT'; chain='A';
pdb=pdbread(['/home/jbanda/Desktop/Deshidro_rangas_090112/' pdb_id '.pdb']);
[strseqnum,ats,best_align]=MSAsearch(pdb,chain,algn);



%% Paso 2. Matriz de similitud de secuencia

% La manera en la que el alineamiento representará una muestra uniforme”
% u homogénea del espacio de secuencia, puede ser juzgado a partir de la
% estructura de las correlaciones (o distancias) entre secuencias. En un
% escenario ideal, las secuencias debieran equivalentemente dissimilares,
% sin clusters distintos de correlaciones de secuencias. En tal caso, la
% identificación del sector radicará en examinar el patrón de correlaciones
% entre posiciones en los primeros eigenmodos de la matriz de ocrrelación
% posicional de SCA.


% Para examinar el espacio de secuencia, computamos a la matriz S de
% similitud entre pares de secuencias.

[S]=sim_seq(algn);

% Hacemos un histograma:

listS=nonzeros(triu(S,1));
h_seqsim=figure; clf;
set(h_seqsim,'Units','normalized','Position',[0 0.3 0.9 0.5],'Name','Sequence Correlations: SDH');
subplot(1,2,1);hist(listS,N_pos/2);
xlabel('Pairwise SeqID','FontSize',14,'FontWeight','bold');
ylabel('number','FontSize',14,'FontWeight','bold'); grid on

% Adicionalmente podemos visualizar directamente a la matriz de similitud
% de secuencias.

figure(h_seqsim);
subplot(1,2,2); imshow(S,[0 1],'InitialMagnification','fit'); colormap(jet); colorbar;
title('SeqID', 'FontSize',12,'FontWeight','bold');

%% Paso 3. Conservación posicional.

[D_glo]=cons(algn);

h_D=figure; set(h_D,'Units','normalized','Position',[0 0.6 0.4 0.4],'Name','Positional Conservation');clf
subplot(2,1,1);hist(D_glo,25); grid on;
xlabel('D (conservation)','FontSize',10,'FontWeight','bold');
ylabel('number','FontSize',10,'FontWeight','bold');
subplot(2,1,2);bar([1:numel(ats)],D_glo,'k'); grid on;
axis([0 numel(ats)+1 0 4]);
set(gca,'XTick',[1:10:numel(ats)]);
set(gca,'XTickLabel',ats([1:10:numel(ats)]));
xlabel('position (1BE9 numbering)','FontSize',10,'FontWeight','bold');
ylabel('D_i (conservation)','FontSize',10,'FontWeight','bold');


%% Paso 4. Cálculos de SCA

% La versión 5 de SCA toma un alineamiento múltiple de secuencias de
```

```
% proteína como entrada y regresa dos matrices: (1) una matriz de correlación
% posicional (Cp), la cual cuantitativamente indica la evolución correlacionada
% de todos los pares de posiciones en el alineamiento, y (2) una matriz de corre-
% lación de secuencias, la cual indica el patrón de similitud entre todos los
% pares de secuencias. Ambas matrices están ponderadas por el grado de
% conservación entre posiciones de aminoácidos. Así, la matriz Cp contiene
% información acerca de las correlaciones conservadas entre pares de
% posiciones, y la matriz Cs contiene información acerca de la similitud entre pares
% de secuencias, sesgada hacia las posiciones más conservadas.


[pdzsca]=sca5(algn);

%% Paso 5. Descomposición espectral (o en eigenvalores)

% Para analizar la matriz de correlación posicional, se lleva a cabo una
% técnica matemática de descomposición espectral (o en eigenvalores),
% la cual se aplica para demostrar la existencia de correlaciones no triviales
% entre posiciones que indican que tratar a los aminoácidos como las
% unidades básicas de las proteínas, no es la representación más informativa.
% En lugar de eso, buscamos una reparametrización de la proteína, en la
% cual las unidades de las proteínas son grupos colectivos de aminoácidos
% que coevolucionan por matriz de correlación posicional. Estos grupos de
% residuos que coevolucionan son llamados “sectores”, y se ha propuesto que
% tales sectores son las unidades evolutivas fundamentales de las proteínas
% (Halabi et al. 2009).

% La descomposición en eigenvalores es la forma más simple de lograr tal
% reparametrización, Tal descomposición es siempre posible para cualquier
% matriz positiva cuadrada semi-definida (como el caso de una matriz de
% correlación; y matemáticamente, la descomposición transforma la
% representation original de un sistema en el cual las variables están
% correlacionadas (en este caso, las posiciones de secuencia) en nuevas
% variables que ahora tienen la propiedad de estar NO correlacionadas
% entre si. Estas nuevas variables son combinaciones lineales de la
% variables originales (en nuestro caso, grupos de posiciones de secuencia)
% y  representan la parametrización más informativa del sistema. Adiconalmente,
% la descomposición en eigenvalores ofrece el ordenamiento de las nuevas
% variables transformadas por medio de la magnitud de captura de contenido
% informacional de la matriz de correlación original.


% Para checar qué eigenvalores son explicados simplemente por ruido estadístico,
% proveniente del muestreo de secuencias, se compara la descomposición
% espectral del alineamiento original con aquel de la conjunción de
% alineamientos aleatorizados en los cuales los aminoácidos son intercambiados
% de forma independiente a lo largo de cada columna. Esta manipulación
% remueve todas las correlaciones funcionales y retiene solamente las
% correlaciones espurias que posiblemente se deben a muestreo finito.

[spect]=spectral_decomp(pdzsca,100);

%% Paso 6a. Estructura de los eigenmodos principales

% Los sectores son empíricamente definidos a través de la examinación del
% patrón de contribuciones posicionales a los poco eigenvectores principales.
% Es por esta razón que usualmente se examina la estructura de únicamente
% los 3 eigenmodos principales, quienes son de hecho, los más informativos.
```

```
% 3-D plots de los tres eigenvectores principales
h_3Dtopmodes=figure; set(h_3Dtopmodes,'Units','normalized','Position',[0 0.7 0.3 0.4],'Name','Top Eigenmodes -
3D'); clf;
scatter3(spect.evpos(:,1),spect.evpos(:,2),spect.evpos(:,3),'ko','SizeData', 50, 'MarkerFaceColor','b');
hold on;for i=1:numel(ats);text(spect.evpos(i,1)+.01,spect.evpos(i,2)+.01,spect.evpos(i,3)+.01,ats(i));end;hold off
az=136;el=20;view(az,el);
xlabel('ev 1','FontSize',12,'FontWeight','b');
ylabel('ev 2','FontSize',12,'FontWeight','b');
zlabel('ev 3','FontSize',12,'FontWeight','b');


% 2-D plots de los tres eigenvectores principales
h_2Dtopmodes=figure; set(h_2Dtopmodes,'Units','normalized','Position',[0 0 1.0 0.4],'Name','Top Eigenmodes-
2D'); clf;
subplot(1,3,1);
scatter(spect.evpos(:,1),spect.evpos(:,2),'ko','SizeData', 50, 'MarkerFaceColor','b');
hold on;for i=1:numel(ats);text(spect.evpos(i,1)+.01,spect.evpos(i,2)+.01,ats(i));end;hold off;grid on
xlabel('ev 1','FontSize',12,'FontWeight','b');ylabel('ev 2','FontSize',12,'FontWeight','b');
subplot(1,3,2);
scatter(spect.evpos(:,1),spect.evpos(:,3),'ko','SizeData', 50, 'MarkerFaceColor','b');
hold on;for i=1:numel(ats);text(spect.evpos(i,1)+.01,spect.evpos(i,3)+.01,ats(i));end;hold off;grid on
xlabel('ev 1','FontSize',12,'FontWeight','b');ylabel('ev 3','FontSize',12,'FontWeight','b');
subplot(1,3,3);
scatter(spect.evpos(:,2),spect.evpos(:,3),'ko','SizeData', 50, 'MarkerFaceColor','b');
hold on;for i=1:numel(ats);text(spect.evpos(i,2)+.01,spect.evpos(i,3)+.01,ats(i));end;hold off;grid on
xlabel('ev 2','FontSize',12,'FontWeight','b');ylabel('ev 3','FontSize',12,'FontWeight','b');


% En el caso del alineamiento para la proteína SDH, el análisis de los modos
% principales muestra tres cosas: (1) la mayoría de las posiciones contribuyen
% débilmente (pesan casi cero), y por lo tanto se predice que evolucionan casi
% independientemente, (2) unas cuantas posiciones (~20%) se distribuyen en
% las colas de las distribuciones de los eigenvectroes, y (3) la contribución de
% las posiciones forma un patrón disperso de peso posicional que no indica
% de forma obvia patrones de varios sectores independientes. En lugar de eso,
% los datos son consistentes con un sólo sector en el cual las posiciones
% que lo componen están involucradas en un patrón jerárquico de correlaciones
% con cada uno.


%% Paso 6b. Análisis por clustering jerárquico [OPCIONAL]

% Para intuir de una forma más visula lo anterior, podemos hacer
% un “clustering” jerárquico de la matriz Cp. Con esto, dos posiciones
% se yuxtaponen si muestran un perfil similar de correlación con
% otras posiciones.

[p_pos,l_pos,sort_pos,Csorted]=SCAcluster(pdzsca.Cp,ats,1);
figure(gcf);

% De nuevo, el clustering muestra que la estructura estadística de la matriz
% Cp parece estar dominada por un grupo de posiciones jerárquicamente
% correlacionadas.

%% Paso 6c. Un mapeo entre correlaciones posicionales y correlaciones
%     en secuencia

[U,sv,V]=svd(pdzsca.pwX);
```

```
% La matriz Pi=U*V' provee un mapeo mapeo matemático entre las matrices
% de correlación posicional y de correlación en secuencias.

N_min=min(N_seq,N_pos);
Pi=U(:,1:N_min)*V(:,1:N_min)';
U_p=Pi*spect.evpos;


%  Así, si un eigenmodo de la matriz de correlación posicional
% (una columna en la variable spect.evpos) describe un sector
% (un grupo de posiciones de aminoácidos que coevolucionen),
% entonces la columna correspondiente de la variable U_p
% contendrá el patrón de divergencia en secuencia (si es que
% existe) que define al sector.

h_SectSeq=figure; set(h_SectSeq,'Units','normalized','Position',[0 0.1 0.6 0.4],'Name','Mapping Seq Correlations
by Positional Correlations'); clf;

h_SectSeq(1)=subplot(1,2,1)
scatter3(spect.evpos(:,1),spect.evpos(:,2),spect.evpos(:,3),'ko','SizeData', 50, 'MarkerFaceColor','b');
hold on;for i=1:numel(ats);text(spect.evpos(i,1)+.01,spect.evpos(i,2)+.01,spect.evpos(i,3)+.01,ats(i));end;hold off
az=58;el=30;view(az,el);
xlabel('ev 1','FontSize',12,'FontWeight','b');
ylabel('ev 2','FontSize',12,'FontWeight','b');
zlabel('ev 3','FontSize',12,'FontWeight','b');

h_SectSeq(2)=subplot(1,2,2)
scatter3(U_p(:,1),U_p(:,2),U_p(:,3),'ko','SizeData', 50, 'MarkerFaceColor','b');
az=58;el=30;view(az,el);
xlabel('Seq 1','FontSize',12,'FontWeight','b');
ylabel('Seq 2','FontSize',12,'FontWeight','b');
zlabel('Seq 3','FontSize',12,'FontWeight','b');


% Para el caso particular del alieamiento con la familia SDH, este
% mapeo es algo aburrido. No hay estructuras particulares para las
% correlaciones posicionales, como tampoco las hay para el espacio
% de secuencia. En lugar de eso, las secuencias parecen ocupar de
% forma homogénea el espacio definido por los principales eigenmodos
% de la matriz de correlación de secuencia. Por lo tanto, concluimos que
% sólo existe un sector jerárquico en esta familia de proteínas.

%% Paso 6d. Definición del Sector

% Se define al sector basándonos en la distribución de los pesos
% posicionales del primer eigenmodo de la matriz Cp.
% El análisis empírico muestra que esta distribución se puede
% ajustar a una distribución Log-normal. Usamos este ajuste
% simplemente para proveer de una base lógica y sistemática
% definir un sector de posiciones.


% Ajustamos al primer eigenvector a la distribución Log-normal y definimos
% al sector por medio de una función de densidad acumulada (cdf) a partir
% de la distribución ajustada y tomando un punto de corte de la cola (a 0.8,
% por lo que elegimos al 20% de la cdf).


h_secdef=figure;
```

```
set(h_secdef,'Units','normalized','Position',[0 1 .5 0.3],'Name','Top Eigenmode'); clf;
p_cutoff=0.8; % punto de corte
secpos = [];

% histograma
xhist=[0:.01:.4];
[yhist]=hist(spect.evpos(:,1),xhist); bar(xhist,yhist./N_pos,'k');hold on;grid on

% ajuste de distribución
pd=fitdist(spect.evpos(:,1),'lognormal');
x_dist=[min(xhist):(max(xhist)-min(xhist))/100:max(xhist)];
area_hist=N_pos*(xhist(2)-xhist(1));
pdf_jnk=pdf(pd,x_dist);
scaled_pdf=area_hist.*pdf_jnk;
plot(x_dist,scaled_pdf./N_pos,'r-','LineWidth',1.5);

% CDF y definición del sector
cdf_jnk=cdf(pd,x_dist);
clear sec cutoff_ev
[jnk,x_dist_pos_right]=min(abs(cdf_jnk-(p_cutoff)));
cutoff_ev = x_dist(x_dist_pos_right)';
% extraemos las posiciones dadas el punto de corte
[sec.def] = find(spect.evpos(:,1)>cutoff_ev);
sprintf('%g+',str2num(char(ats(sec.def))))
sec.cutoff=cutoff_ev;
figure(h_secdef); line([cutoff_ev cutoff_ev],[0 max(yhist)/N_pos],'LineWidth',1,'LineStyle','--','Color','b');
sec.col=2/3;

% comparación de la distribución de los eigenvectores de los alineamientos
% aleatorizados, escalados para comparación con el alineamiento original.
clear tmp
N_samples=100;
for i=1:N_samples;tmp(i,:)=squeeze(spect.lbdposrnd(i,1).*spect.evposrnd(i,:,1));end
[yhist_rnd]=hist(tmp(:)./spect.lbdpos(1),xhist);
plot(xhist,yhist_rnd/(N_samples*N_pos),'g','LineWidth',1.5);
hold off

%% Paso 10. Análisis estructural de sectores.

% Organización por estructura primaria/secundaria:

h_PriSecStr=figure;
set(h_PriSecStr,'Name','Sectors: Primary and Secondary Structure','Position',[41 37 841 266]);
clf
subplot(2,1,1);
show_vect(D_glo,sec,ats);
pdzss.Sheet=[1:12]; pdzss.Helix=[1:13];
figure(h_PriSecStr); hsub=subplot(2,1,2);hold on;drawSS(pdb,ats,pdzss,sec);hold off


% Residuos del sector
sprintf('%g+',str2num(char(ats(sec.def))))

%MOST CONSERVED RESIDUES
 sprintf('%g+',str2num(char(ats(D_glo>2.5))))
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

%% SDH, tres sectores, al 0.96 como punto de corte por sector. Dentro de la SDH, este script es aplicable a cualquier punto de corte siempre y cuando sea éste el único parámetro que se altere.

```
addpath sca5
clear; close all

%% Paso 1. Cargado del alineamientos y acondicionamiento del mismo

[labels_seq,algn_full]=get_seqs('/home/jbanda/Desktop/Deshidro_rangas_090112/1nyt_
80.ran');
N_seq=size(algn_full,1);

cut_off=.2;
fraq_gaps=sum(isletter(algn_full)==0)/N_seq;
algn=algn_full(:,fraq_gaps<cut_off);
N_pos=size(algn,2);

pdb_id='1NYT'; chain='A';
pdb=pdbread(['/home/jbanda/Desktop/Deshidro_rangas_090112/' pdb_id '.pdb']);
[strseqnum,ats,best_align]=MSAsearch(pdb,chain,algn);

%% Paso 2. Matriz de similitud de secuencia.

[S]=sim_seq(algn);


listS=nonzeros(triu(S,1));
h_seqsim=figure; clf;
set(h_seqsim,'Units','normalized','Position',[0 0.3 0.9 0.5],'Name','Sequence
Correlations: SDH');
subplot(1,2,1);hist(listS,N_pos/2);
xlabel('Pairwise SeqID','FontSize',14,'FontWeight','bold');
ylabel('number','FontSize',14,'FontWeight','bold'); grid on


figure(h_seqsim);
subplot(1,2,2); imshow(S,[0 1],'InitialMagnification','fit'); colormap(jet);
colorbar;
title('SeqID', 'FontSize',12,'FontWeight','bold');

%% Paso 3. Conservación posicional.

[D_glo]=cons(algn);

h_D=figure; set(h_D,'Units','normalized','Position',[0 0.6 0.4
0.4],'Name','Positional Conservation');clf
subplot(2,1,1);hist(D_glo,25); grid on;
xlabel('D (conservation)','FontSize',10,'FontWeight','bold');
ylabel('number','FontSize',10,'FontWeight','bold');
```

```
subplot(2,1,2);bar([1:numel(ats)],D_glo,'k'); grid on;
axis([0 numel(ats)+1 0 4]);
set(gca,'XTick',[1:10:numel(ats)]);
set(gca,'XTickLabel',ats([1:10:numel(ats)]));
xlabel('position (SDH numbering)','FontSize',10,'FontWeight','bold');
ylabel('D_i (conservation)','FontSize',10,'FontWeight','bold');

%% Paso 4. Cálculos de SCA

[pdzsca]=sca5(algn);


%% Paso 5. Descomposición espectral (o eigenvalores)

[spect]=spectral_decomp(pdzsca,100);

%% Paso 6a. Estructura de los eigenmodos principales.

% 3-D plot de los tres eigenmodos principales.
h_3Dtopmodes=figure; set(h_3Dtopmodes,'Units','normalized','Position',[0 0.7 0.3
0.4],'Name','Top Eigenmodes - 3D'); clf;
scatter3(spect.evpos(:,1),spect.evpos(:,2),spect.evpos(:,3),'ko','SizeData', 50,
'MarkerFaceColor','b');
hold on;for
i=1:numel(ats);text(spect.evpos(i,1)+.01,spect.evpos(i,2)+.01,spect.evpos(i,3)+.01
,ats(i));end;hold off
az=136;el=20;view(az,el);
xlabel('ev 1','FontSize',12,'FontWeight','b');
ylabel('ev 2','FontSize',12,'FontWeight','b');
zlabel('ev 3','FontSize',12,'FontWeight','b');

% 2-D plots de los tres eigenmodos principales.
h_2Dtopmodes=figure; set(h_2Dtopmodes,'Units','normalized','Position',[0 0 1.0
0.4],'Name','Top Eigenmodes-2D'); clf;
subplot(1,3,1);
scatter(spect.evpos(:,1),spect.evpos(:,2),'ko','SizeData', 50,
'MarkerFaceColor','b');
hold on;for
i=1:numel(ats);text(spect.evpos(i,1)+.01,spect.evpos(i,2)+.01,ats(i));end;hold
off;grid on
xlabel('ev 1','FontSize',12,'FontWeight','b');ylabel('ev
2','FontSize',12,'FontWeight','b');
subplot(1,3,2);
scatter(spect.evpos(:,1),spect.evpos(:,3),'ko','SizeData', 50,
'MarkerFaceColor','b');
hold on;for
i=1:numel(ats);text(spect.evpos(i,1)+.01,spect.evpos(i,3)+.01,ats(i));end;hold
off;grid on
xlabel('ev 1','FontSize',12,'FontWeight','b');ylabel('ev
3','FontSize',12,'FontWeight','b');
subplot(1,3,3);
scatter(spect.evpos(:,2),spect.evpos(:,3),'ko','SizeData', 50,
'MarkerFaceColor','b');
hold on;for
i=1:numel(ats);text(spect.evpos(i,2)+.01,spect.evpos(i,3)+.01,ats(i));end;hold
off;grid on
xlabel('ev 2','FontSize',12,'FontWeight','b');ylabel('ev
3','FontSize',12,'FontWeight','b');
```

```
%% Paso 6b. Análisis por clusterir jerárquico.

[p_pos,l_pos,sort_pos,Csorted]=SCAcluster(pdzsca.Cp,ats,1);
figure(gcf);


%% Paso 7. Análisis de Componentes Independientes



% En principio, el proceso de ICA debería ayudar a definir de una mejor manera
% sectores independientes, como grupos de posiciones proyectándose específicamente
% a lo largo de los ejes transformados -los componentes independientes.

kmax =9;
learnrate=.0001; iterations=20000;
[W,changes_s]=basic_ica(spect.evpos(:,1:kmax)',learnrate,iterations);
ic_P=(W*spect.evpos(:,1:kmax)')';

h_ICA3D=figure; set(h_ICA3D,'Units','normalized','Position',[0 0.1 0.25
0.4],'Name','Independent Components - 3D'); clf;
scatter3(ic_P(:,1),ic_P(:,2),ic_P(:,3),'ko','SizeData', 50,
'MarkerFaceColor','b');
hold on;for
i=1:numel(ats);text(ic_P(i,1)+.05,ic_P(i,2)+.05,ic_P(i,3)+.05,ats(i));end;hold off
az=125;el=42;view(az,el);
xlabel('IC 1','FontSize',12,'FontWeight','b');
ylabel('IC 2','FontSize',12,'FontWeight','b');
zlabel('IC 3','FontSize',12,'FontWeight','b');

% 2-D plots de los primeros tres eigenvectores.
h_ICA2D=figure; set(h_ICA2D,'Units','normalized','Position',[0 1 0.75
0.3],'Name','Top Eigenmodes-2D'); clf;
subplot(1,3,1);
scatter(ic_P(:,1),ic_P(:,2),'ko','SizeData', 50, 'MarkerFaceColor','b');
hold on;for i=1:numel(ats);text(ic_P(i,1)+.01,ic_P(i,2)+.01,ats(i));end;hold
off;grid on
xlabel('ev 1','FontSize',12,'FontWeight','b');ylabel('ev
2','FontSize',12,'FontWeight','b');
subplot(1,3,2);
scatter(ic_P(:,1),ic_P(:,3),'ko','SizeData', 50, 'MarkerFaceColor','b');
hold on;for i=1:numel(ats);text(ic_P(i,1)+.01,ic_P(i,3)+.01,ats(i));end;hold
off;grid on
xlabel('ev 1','FontSize',12,'FontWeight','b');ylabel('ev
3','FontSize',12,'FontWeight','b');
subplot(1,3,3);
scatter(ic_P(:,2),ic_P(:,3),'ko','SizeData', 50, 'MarkerFaceColor','b');
hold on;for i=1:numel(ats);text(ic_P(i,2)+.01,ic_P(i,3)+.01,ats(i));end;hold
off;grid on
xlabel('ev 2','FontSize',12,'FontWeight','b');ylabel('ev
3','FontSize',12,'FontWeight','b');


%% Paso 6c. Mapeo entre correlaciones posicionales y de secuencia.


[U,sv,V]=svd(pdzsca.pwX);
```

```
N_min=min(N_seq,N_pos);
Pi=U(:,1:N_min)*V(:,1:N_min)';
U_p=Pi*spect.evpos;


h_SectSeq=figure; set(h_SectSeq,'Units','normalized','Position',[0 0.1 0.6
0.4],'Name','Mapping Seq Correlations by Positional Correlations'); clf;

h_SectSeq(1)=subplot(1,2,1)
scatter3(spect.evpos(:,1),spect.evpos(:,2),spect.evpos(:,3),'ko','SizeData', 50,
'MarkerFaceColor','b');
hold on;for
i=1:numel(ats);text(spect.evpos(i,1)+.01,spect.evpos(i,2)+.01,spect.evpos(i,3)+.01
,ats(i));end;hold off
az=58;el=30;view(az,el);
xlabel('ev 1','FontSize',12,'FontWeight','b');
ylabel('ev 2','FontSize',12,'FontWeight','b');
zlabel('ev 3','FontSize',12,'FontWeight','b');

h_SectSeq(2)=subplot(1,2,2)
scatter3(U_p(:,1),U_p(:,2),U_p(:,3),'ko','SizeData', 50, 'MarkerFaceColor','b');
az=58;el=30;view(az,el);
xlabel('Seq 1','FontSize',12,'FontWeight','b');
ylabel('Seq 2','FontSize',12,'FontWeight','b');
zlabel('Seq 3','FontSize',12,'FontWeight','b');

%% Paso 9. Definición del sector usando los primeros tres componentes.

h_ICAfit=figure;
set(h_ICAfit,'Units','normalized','Position',[0 1 1 0.5],'Name','IC
distributions'); clf;
clear sec cutoffs

p_cutoff=0.96;  % Cutoff
nfit=3;
cutoffs = zeros(nfit,1);

for i=1:nfit
    pd=fitdist(ic_P(:,i),'tlocationscale');
    subplot(1,nfit,i);
    binwidth=2*iqr(ic_P(:,i))*(numel(ic_P(:,i))^-0.33); %regla de Freedman-
Diaconis
    nbins=round(range(ic_P(:,i))/binwidth);

% histograma de pesos de IC como densidades de probabilidad.

    [yhist,xhist]=hist(ic_P(:,i),nbins); bar(xhist,yhist/N_pos,'k');hold on;grid
on
    % distribución ajustada
    x_dist=[min(xhist):(max(xhist)-min(xhist))/100:max(xhist)];
    area_hist=N_pos*(xhist(2)-xhist(1)); % for proper scaling of the pdf
    pdf_jnk=pdf(pd,x_dist);
    scaled_pdf=area_hist.*pdf_jnk;
    plot(x_dist,scaled_pdf./N_pos,'r-','LineWidth',1.5);
```

```
    cdf_jnk=cdf(pd,x_dist);
    % here, we identify the direction of the tail (the sign of independent
    % components is ar0.96bitrary), the cutoff, and the sector positions based
    % on the fitted cdf:
% identificamos la dirección de la cola (el signo de los componentes
% independientes es arbitrario), el punto de corte, y las posiciones de cada
% sector basados en la función de densidad acumulada ajustada.
    [~,maxpos]=max(pdf_jnk);
    tail=zeros(1,numel(pdf_jnk));
    if abs(max(ic_P(:,i)))>abs(min(ic_P(:,i)))
        tail(maxpos:end)=cdf_jnk(maxpos:end);
    else
        cdf_jnk=1-cdf_jnk;
        tail(1:maxpos)=cdf_jnk(1:maxpos);
    end
    [~,x_dist_pos]=min(abs(tail-p_cutoff));
    cutoffs(i) = x_dist(x_dist_pos);

    y_lim=get(gca,'ylim');
    line([cutoffs(i)
cutoffs(i)],y_lim,'Color','k','LineWidth',1,'LineStyle','--');
    text(cutoffs(i),1.03*(y_lim(2)-
y_lim(1)),num2str(cutoffs(i),'%0.2f'),'FontWeight','bold','FontSize',11);
    xlabel(['IC ' num2str(i)],'FontSize',12,'FontWeight','b');
    ylabel('Prob Density','FontSize',12,'FontWeight','b');

    % obtenemos la numeración final de cada sector
    if abs(max(ic_P(:,i)))>abs(min(ic_P(:,i)))
        sec(i).def = find(ic_P(:,i)>cutoffs(i));
    else
        sec(i).def = find(ic_P(:,i)<cutoffs(i));
    end
end
cutoffs = x_dist(x_dist_pos);
sec(1).col=2/3;
sec(2).col=0;
sec(3).col=1/3;

%% Organización en estructura primaria y secundaria

h_PriSecStr=figure;
set(h_PriSecStr,'Name','Sectors: Primary and Secondary Structure','Position',[41
37 841 266]);
clf
subplot(2,1,1);
show_vect(D_glo,sec,ats);
pdzss.Sheet=[1 2 3 4 5 6 7 8 9 10 11 12]; pdzss.Helix=[1 2 3 4 5 6 7 8 9 10 11 12
13];
figure(h_PriSecStr); hsub=subplot(2,1,2);hold on;drawSS(pdb,ats,pdzss,sec);hold
off


%% Paso 10. Análisis estructural de sectores

sprintf('%g+',str2num(char(ats(sec(1).def))))
sprintf('%g+',str2num(char(ats(sec(2).def))))
sprintf('%g+',str2num(char(ats(sec(3).def))))
sprintf('%g+',str2num(char(ats(D_glo>2.5))))
```

COMANDOS PARA EL REDISEÑO DE LOS DOMINIO ROSSMANN DE LA SDH PARA ROSETTADESIGN

Los siguiente es un ejemplo de cómo luce un archivo de instrucciones para ROSETTADESIGN. A este archivo le llamamos SDH_ROSS_SCA.resfile. Aquello que se encuentre después del signo # es interpretado como comentario.

EX 1

start

101 A PIKAA M #Metionina de inicio
102 A NATAA
103 A NATAA
105 A NATAA #Dejar al aminoácido natural
107 A NATAA
109 A NATAA
110 A NATAA
119 A NATAA
123 A NATAA
126 A NATAA
127 A NATAA
128 A NATAA
129 A NATAA
130 A NATAA
132 A NATAA
137 A NATAA
138 A NATAA
149 A NATAA
150 A NATAA
151 A NATAA
155 A NATAA
158 A NATAA
159 A NATAA
181 A NATAA
182 A NATAA
186 A NATAA
188 A NATAA
189 A NATAA
191 A NATAA
211 A NATAA
212 A NATAA
213 A NATAA
215 A NATAA
217 A NATAA
222 A NATAA
225 A NATAA

237 A NATAA
239 A NATAA
240 A NATAA
242 A NATAA
244 A NATAA

(Esto es un comentario) Lo siguiente es una sóla línea de comando, en donde sdh_A_ross2.pdb es un archivo de coordenadas de únicamente el dominio Rossmann de las SDH de E. coli, y SDH_ROSS_SCA.resfile es el archivo que dicta qué residuos serán inalterables y el -nstruct indica que deseamos generar 999 secuencias:

```
nohup /usr/local/bin/fixbb.linuxgccrelease -s
/home/jbanda/rossetta/SDH_ROSS_MET2/sdh_A_ross.pdb -resfile
/home/jbanda/rossetta/SDH_ROSS_MET2/SDH_ROSS_SCA.resfile  -database
/home/jbanda/rosetta_database/ -nstruct 999 &
```

(Esto es un comentario) Los demás archivos *.resfile (lo único críticamente variable en la línea anterior) para todos los demás conjuntos de diseños se encuentran en la carpeta comprimida asociada a la tesis.

COMANDO PARA LA DINÁMICA MOLECULAR

Ejemplo, la dinámica molecular para el diseño Aleatorio. El comando es una sóla línea:

$SCHRODINGER/utilities/multisim -JOBNAME desmond_sa_NPT_random -HOST compute-0-0 -maxjob 0 -cpu "1 2 6" -m desmond_sa_NPT_random.msj -c desmond_sa_NPT_random.cfg desmond_sa_NPT_random.cms

(Esto es un comentario) Estamos indicando que deseamos usar 6 procesadores de cada uno de 2 núcleos. El archivo *msj contiene el protocolo de la simulación, el archivo *.cfg contiene todas las condiciones propias del proceso completo, en este caso NPT; y el archivo *.cms contiene los parámetros correspondientes al campo de fuerza que se decidió usar (OPLS-2005) para la molécula. Todos estos archivos son específicos por simulación. Todos ellos vienen en la carpeta comprimida asociada a la tesis.