Corvi, Javier Omar

Análisis de coevolución en alineamientos múltiples de secuencias de proteínas evolucionados artificialmente

Análisis de coevolución en alineamientos múltiples de secuencias de proteínas evolucionados artificialmente

TESIS DE MAESTRÍA

Javier Omar Corvi

javicorvi@gmail.com

Resumen

El objetivo general de este trabajo es estudiar y analizar la relación entre coevolución y conservación derivados de MSAs en cuanto a su capacidad para predecir contactos en estructuras terciarias de proteínas. En este estudio proponemos dicho análisis en MSAs naturales y derivados de simulaciones computacionales utilizando el software SCPE. Nuestra principal hipótesis es que el SCPE provee una adecuada simulación del proceso evolutivo bajo la conservación de una estructura con la cual contrastar el proceso coevolutivo en los alineamientos naturales, mucho más complejos y diversos.

Maestría en Bioinformática y Biología de Sistemas

Universidad Nacional de Quilmes

Tesis de Maestría

Análisis de coevolución en alineamientos múltiples de secuencias de proteínas evolucionados artificialmente

Aspirante: Javier Omar Corvi
Director: Cristina Marino Buslje [1].
Codirector: Gustavo Parisi [2].
1 Laboratorio de Bioinformática Estructural Fundación Instituto Leloir
2 Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Bernal, Argentina

1. Introducción

Existen varias formas para describir una proteína, según su composición química, su función, la familia a la que pertenece; y una muy importante: según su estructura. En esta última existen varios niveles, desde estructura primaria hasta cuaternaria (Figura 1.1). La estructura primaria: es la más básica y consiste en la secuencia ordenada de aminoácidos; la estructura secundaria: es el arreglo local de la estructura primaria estabilizado por puentes de hidrógeno. Los dos tipos más comunes de estructuras secundarias que mayormente se terminan generando son: Alfa Hélices y Hojas Betas. Estas estructuras se conectan unas con otras a través de Loops. La estructura terciaria es el arreglo espacial de las estructuras secundarias y lo que le da básicamente el cuerpo en el espacio a una proteína. La estructura cuaternaria se forma mediante la unión de varias cadenas con estructura terciaria para formar un complejo proteico.

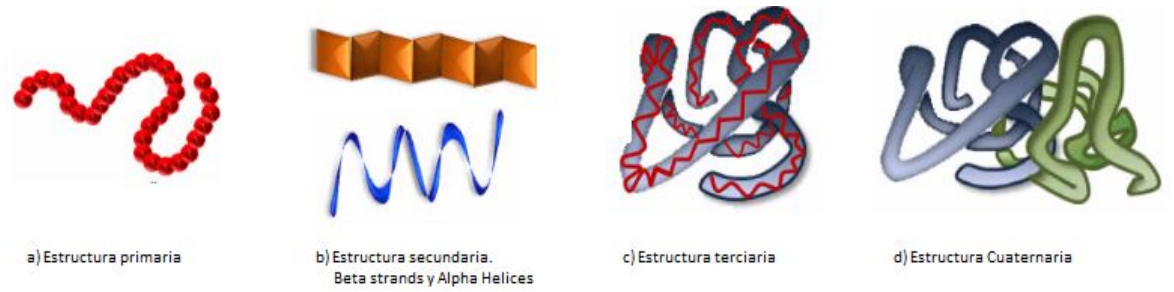


Figura 1.1. Categorías de estructuras de una proteína. Desde su estructura más simple, primaria, a la más compleja, cuaternaria.

Existen diferentes métodos que actualmente son utilizados para determinar la estructura de una proteína; dentro de los cuales se incluyen la cristalografía de rayos X (X-ray crystallography), la espectroscopia de resonancia magnética nuclear (nuclear magnetic resonance spectroscopy - NMR spectroscopy), y la microscopía de electrones (electron microscopy). Cada uno de los métodos tiene ventajas y desventajas.

En cada uno de ellos se utilizan elementos de información para crear el modelo atómico final. Esencialmente algún tipo de dato experimental acerca de la estructura de la molécula. En la cristalografía de rayos X, se utilizan los patrones de difracción de rayos X. Para la espectroscopia de resonancia magnética nuclear, es la información de la conformación local y la distancia entre los átomos que están cerca unos de otros. En la microscopía de electrones, es una imagen de la forma general de la molécula.

El resultado de cada uno de estos métodos es un archivo pdb (protein data bank) dentro del cual se incluye la información espacial de la proteína, es decir, se anotan con coordenadas espaciales la posición de cada uno de los átomos de los aminoácidos. En la Figura 1.2 se puede apreciar la estructura con código de pdb 2TRX de la proteína tioredoxina perteneciente a E. coli con código THIO_ECOLI.

La secuencia de una proteína determina su estructura tridimensional, la cual a su vez determina su función. Entonces, una forma de describir a la estructura espacial es indicando qué posiciones (aminoácidos) de la estructura primaria están en contacto entre sí. Por ejemplo, la posición 23 puede estar en contacto con la posición 54; diferentes contactos entre las posiciones terminan dándole estructura tridimensional a la proteína.

La definición de contacto puede ser, por ejemplo, cuando dos posiciones tienen algún átomo que se encuentra a una distancia menor a n ångströms (Å) entre ellos, en general entre 4-6 Å. El número n puede variar según el estudio que se realiza.

Por lo tanto al analizar el archivo pdb podemos verificar qué posiciones se encuentran a una distancia menor que n Å y de ahí indicar que dichas posiciones están en contacto.

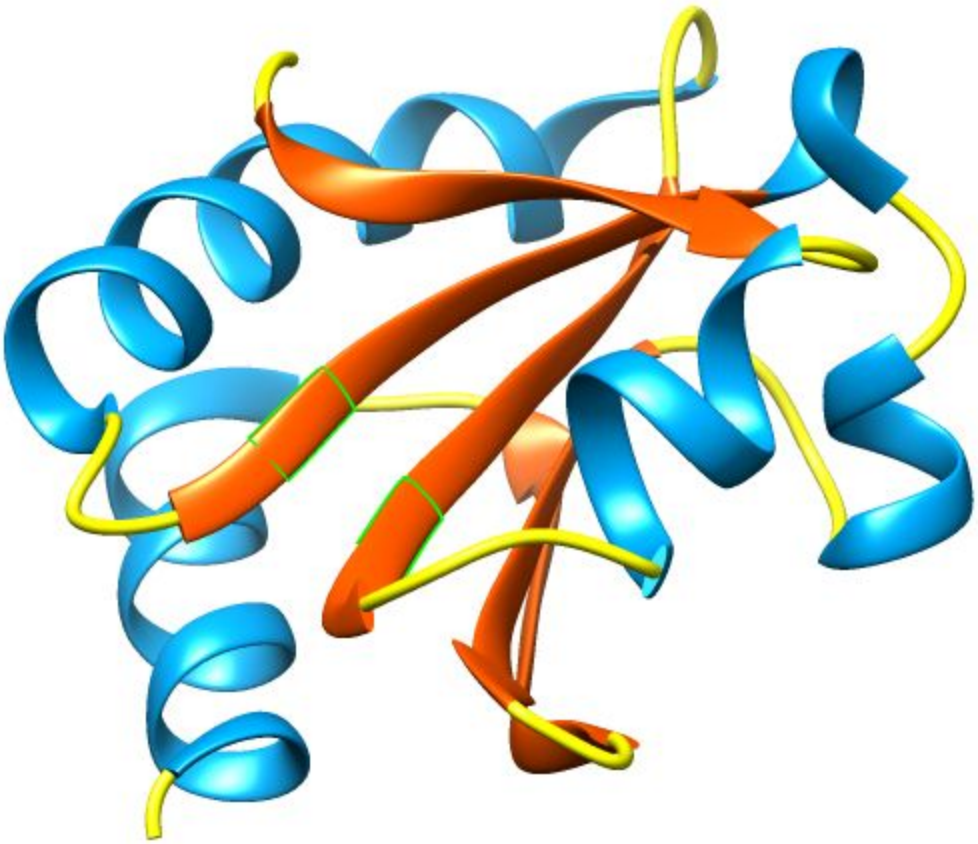


Figura 1.2. Estructura espacial de la proteína THIO_ECOLI, entrada de pdb 2TRX. Podemos apreciar las Alfa Hélices, Hojas Betas y Loops que conectan dichas estructuras. En la parte inferior se ve la secuencia primaria de aminoácidos de la proteína, tiene seleccionada dos posiciones (23 y 54). Se observa que dichas posiciones están a distancia de contacto.

Una proteína puede adoptar ligeramente diferentes disposiciones espaciales, denominados confórmeros. Es decir, la misma secuencia de aminoácidos tiene divergencia conformacional, esto se debe a la rotación que los enlaces sencillos puedan llegar a tener en el espacio. Un modelo de estructura, descrito en un pdb es un confórmero determinado; por ejemplo, para la proteína THIO_ECOLI, hemos descrito que una de sus estructuras conformacionales es la

2TRX, pero hay otras. La resolución de la estructura de la misma proteína, en diferentes instancias, puede resultar en diferentes confórmeros.

Una proteína puede pertenecer a una familia de proteínas, de la cual se obtiene un alineamiento múltiple de secuencias (MSA). Un MSA es un conjunto de secuencias alineadas, en este caso de proteínas, de las cuales se puede inferir homología y llevar adelante un análisis filogenético. Al tener un conjunto grande de secuencias (preferiblemente más de 400) se pueden realizar cálculos referidos a la conservación de los aminoácidos por posición; y también, para un posterior análisis de coevolución, puede, calcularse la información mutua (MI - por Mutual information) y de análisis de acoplamiento directo (DCA - por Direct Analysis) que existen entre las posiciones.

Esta información es importante porque nos brinda conocimiento acerca de la coevolución de los residuos: cuando un residuo es mutado, otros residuos deben cambiar para preservar o restaurar la estructura o la función de la proteína. La Figura 1.3 muestra un MSA de ejemplo en donde se describen cada una de estas definiciones.

Durante diferentes trabajos de investigación [1-5] se ha encontrado que a través de una alta señal de covariación entre dos posiciones se puede predecir que las mismas están en contacto en la estructura tridimensional, lo que se denomina coevolución por estructura. Por un lado obtenemos la matriz de contacto de una proteína cristalizada, dada según una definición de contacto. Por otro lado disponemos de varios métodos para calcular la covariación y realizar el análisis: MI, DCA y PSICOV. El desempeño predictivo de un método de coevolución (para predecir residuos en contacto) se puede evaluar a través del área bajo la curva ROC (AUC). Para más información acerca de covariación, coevolución y el método predictivo que se emplea durante el trabajo ver las secciones 3.4 y 3.11.

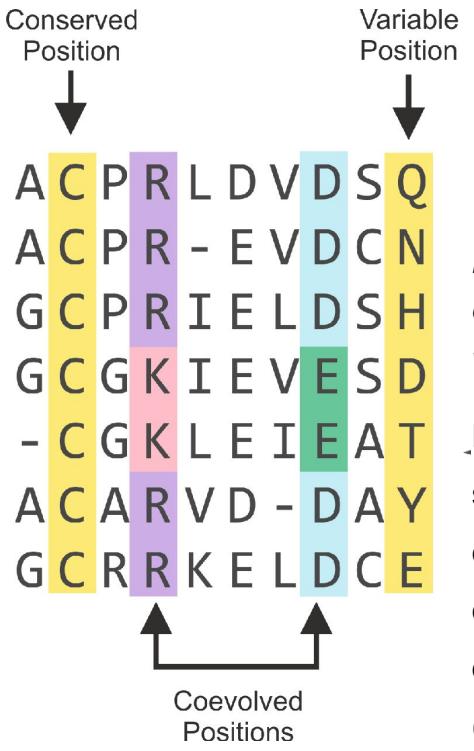


Figura 1.3. MSA de ejemplo para visualizar posiciones que coevolucionan, posiciones conservadas y posiciones variables.

En el campo de estudio de la diversidad secuencial y estructural de proteínas, se pueden encontrar diferentes algoritmos para simular su evolución, uno en particular y que se utilizará durante este proyecto es el software SCPE (Structurally Constrained Protein Evolutionary model); el cual dada la estructura de una proteína, extrae la secuencia de aminoácidos y realiza mutaciones sobre la misma, aceptando sólo aquellas mutaciones que sean compatibles con la estructura, es decir, manteniendo las restricciones estructurales de la proteína [6].

En este trabajo de investigación se estudiará y analizará la coevolución, predicción de contactos y conservación en MSAs resultantes de simular la evolución de proteínas utilizando el software SCPE, es decir, una evolución artificial, computacional, que no tiene en cuenta relaciones filogenéticas sino solo que se mantenga la estructura proteica. Por comodidad y para facilitar la lectura del manuscrito, al MSA resultante de este proceso in silico, lo denominaremos MSA simulado o evolucionado a lo largo del trabajo.

Se realizarán comparaciones entre los MSAs evolucionados y el MSA natural para analizar similitudes y diferencias en la información que contienen.

En la actualidad no se han realizado estudios sobre coevolución, predicción de contactos y conservación en MSAs generados artificialmente evolucionando

una proteína teniendo en cuenta su estructura. Durante la primera etapa, para el trabajo de investigación, se utilizará la proteína thiorredoxina de *E. Coli* (Uniprot accesión: THIO_ECOLI) perteneciente a la familia de Pfam[7] PF00085; la evolución artificial será realizada en base a la estructura 2TRX. Una vez conocido el sistema y puesto a punto el procedimiento y el algoritmo, se realizarán otros estudios: evolucionar diferentes confórmeros de la misma proteína. La proteína THIO_ECOLI fue cristalizada en varias oportunidades, por lo que disponemos de diferentes confórmeros. Evolucionar diferentes estructuras conformacionales, para luego realizar un análisis de la información conjunta y conocer si agregando más restricciones estructurales obtenemos mejores resultados y mayor similitud con el MSA natural, es otro de los estudios que se realizarán en este proyecto.

Para llevar a cabo el proyecto se realizaron desarrollos propios en los lenguajes de programación Python y Julia; se utilizaron librerías informáticas existentes como MIToS [8], MISTIC [9], HHPRED[10] y Seq2Logo [11].

El trabajo de investigación fue desarrollado de forma semipresencial en el laboratorio de Bioinformática Estructural de la Fundación Instituto Leloir que se encuentra dirigido por Doctora Cristina Marino Buslje.

2. Objetivos

El objetivo general de este trabajo es estudiar y analizar la relación entre coevolución y conservación derivados de MSAs en cuanto a su capacidad para predecir contactos en estructuras terciarias de proteínas. En este estudio proponemos dicho análisis en MSAs naturales y derivados de simulaciones computacionales utilizando el software SCPE. Nuestra principal hipótesis es que el SCPE provee una adecuada simulación del proceso evolutivo bajo la conservación de una estructura con la cual contrastar el proceso coevolutivo en los

alineamientos naturales, mucho más complejos y diversos. Esta comparación da las bases para formular los siguientes objetivos y preguntas que nos realizamos:

1. Estimar, evaluar y comparar los procesos de coevolución y conservación en ambos tipos de MSAs. Tiene el MSA evolucionado una señal de coevolución similar a la del MSA Natural ?. Cómo se comporta la conservación en el MSA evolucionado respecto al natural?
2. Evaluar la capacidad predictiva de distintos métodos de los contactos inter-residuos.
3. Investigar si la evolución a través de los años embebe otro tipo de información en las secuencias naturales además de la estructural, como por ejemplo la conservación de residuos con una finalidad funcional diferente de la de mantener la estructura.
4. Estudiar el efecto de la diversidad conformacional en la señal de coevolución y conservación. Evolucionar varios confórmeros de la misma proteína, nos brinda información complementaria de importancia ? Mejora la predicción de contactos ?

3. Procedimiento y Métodos

Durante esta primera introducción al procedimiento se describen los pasos a realizar para llevar a cabo el proceso de evolución simulado y su posterior análisis. Visualmente se puede apreciar en la Figura 3.1.

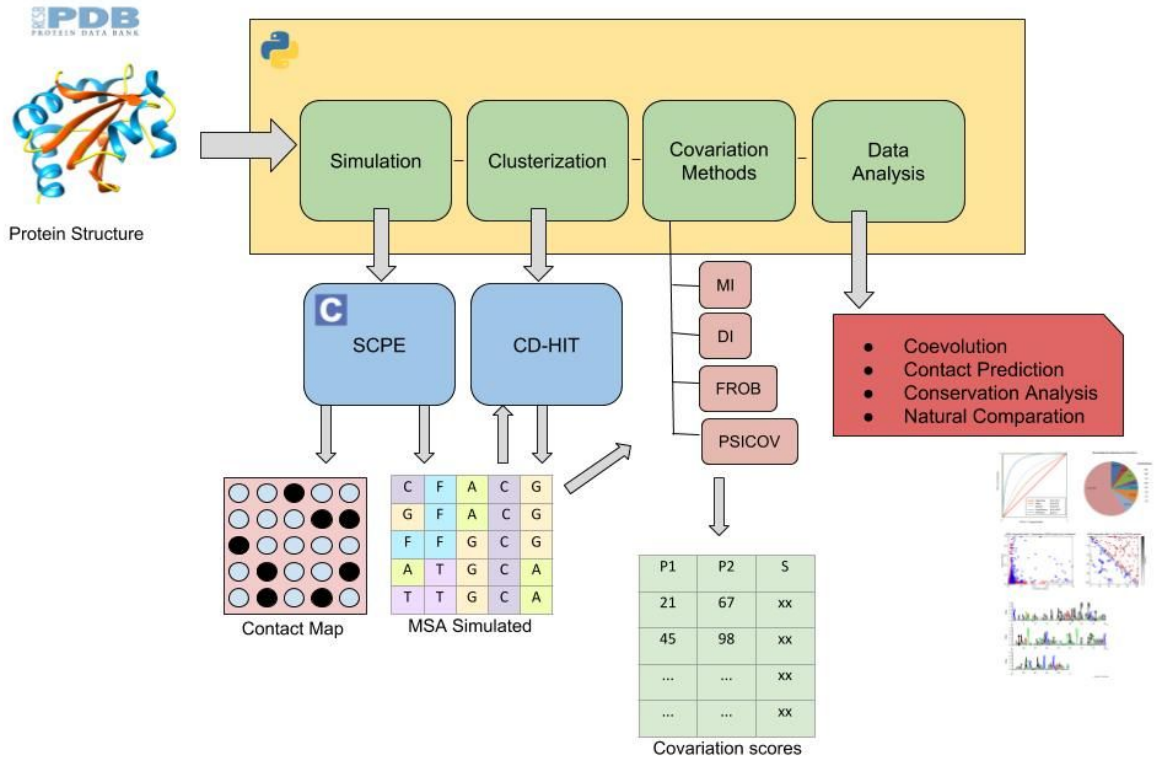


Figura 3.1. Procedimiento realizado para evolucionar una proteína y analizar los resultados.

El orden de los pasos es el siguiente:

- En primer lugar se realiza la evolución con SCPE, el cual recibe la estructura de la proteína a evolucionar (pdb). El resultado es la generación de un MSA con las secuencias evolucionadas (luego de haber aceptado mutaciones en los sucesivos ciclos de evolución); además, como resultado secundario, retorna la matriz de contacto en base a la estructura original.
- Luego en el MSA generado artificialmente se elimina la redundancia mediante un agrupamiento de las secuencias al 62% de identidad. Es decir, se agrupan las secuencias 62% idénticas o más.
- Una vez obtenido el MSA sin redundancia, se realiza el cálculo de covariación con diferentes métodos: MI, DCA(DI y FROB) y PSICOV.

- Por último, con los datos disponibles, MSA evolucionado y matriz de contacto, se realizan los cálculos de coevolución para cada uno de los métodos, predicción de contactos y conservación; finalmente los resultados son comparados con los obtenidos del MSA Natural.

Las secciones posteriores del capítulo brindan un marco teórico detallado de los procesos y métodos que se utilizaron durante el proyecto de investigación; como también el desarrollo del software bioinformático utilizado para llevarlo adelante.

3.1. SCPE (Structurally Constrained Protein Evolution)

El software SCPE [6] extrae la secuencia primaria de un archivo pdb y realiza mutaciones al azar teniendo en cuenta la estructura. Las mutaciones son aceptadas o rechazadas utilizando un *score* que mide la diferencia energética introducida por la mutación y un parámetro β que es la medida de presión selectiva aplicada para la conservación estructural: valores altos de β implican una mayor restricción estructural y por consiguiente una baja probabilidad de aceptación de las mutaciones. Por otro lado, valores bajos de β implican una relajación en cuanto a la conservación estructural y una mayor probabilidad de aceptación. El tiempo de divergencia, los ciclos de mutaciones, se basan en la definición de la cantidad de sustituciones no sinónimas por sitio permitidas en promedio (a partir de aquí nsus), dicha información también se encuentra parametrizada. Los parámetros más importantes que deben completarse para la ejecución del programa pueden apreciarse en la Tabla 3.1.

Por último, el número de simulaciones independientes (a partir de aquí runs); este parámetro indica la cantidad de secuencias en estado inicial que comenzarán a evolucionar. Por lo tanto, si agregamos un número de runs = 100 esto significa que al comienzo se realizarán 100 copias de la secuencia original, las cuales una vez comenzado el proceso serán mutadas a lo largo de la ejecución

hasta que se llegue a un número promedio de nsus. El tiempo de ejecución del software está ligado a la cantidad de runs y nsus definidos.

Nombre del parámetro	**Descripción**
PDB FILENAME	Archivo PDB de la proteína a evolucionar
PROTEIN CHAIN	Cadena de la proteína a evolucionar
BETA PARAMETER (β)	Medida de presión para la conservación estructural
NUMBER OF ACCEPTED SUSTITUTIONS PER SITE (nsus)	Número de sustituciones no sinónimas por sitio
NUMBER OF INDEPENDENT RUNS (runs)	Número de simulaciones independientes

Tabla 3.1. Parámetros, con sus respectivas descripciones, que acepta SCPE para la ejecución.

3.1.1. Adaptación SCPE: Nuevos requerimientos

El objetivo del software SCPE no era en un principio imprimir las secuencias evolucionadas para generar MSAs artificiales, consistía en realizar mutaciones sobre la secuencia de una estructura conocida manteniendo las restricciones estructurales para estudiar la relación entre la conservación de la estructura y la divergencia de la secuencia en la evolución de las proteínas; el resultado principal de la ejecución del programa es un conjunto de matrices de sustitución sitio específicas derivadas de realizar las mutaciones aceptadas en la evolución [8,13].

Para adaptar el software a los requerimientos propuestos se realizaron las siguientes tareas de desarrollo:

- Agregar la lógica correspondiente e imprimir las secuencias mutadas:
  Las secuencias generadas por el proceso son impresas en un archivo, indicando que parámetros del SCPE fueron utilizados para su ejecución. El resultado de cada evolución es un archivo fasta que contiene el MSA evolucionado con el nombre "sequences-betaXX-nsusXX-runsXXX.fasta", donde se puede destacar que parámetros del SCPE fueron utilizados para realizar la evolución. La primer secuencia que aparece en el archivo es la secuencia de la proteína sin ninguna mutación, la denominamos secuencia de referencia.

- Imprimir la matriz de contacto de la proteína:

Como resultado secundario, además del MSA resultante evolucionado, el software también retorna en un archivo denominado "contact_map.dat", el cual contiene la matriz de contacto resultante de analizar la estructura de la proteína evolucionada. El formato es 1-0 (Contacto-No Contacto). Esta matriz será luego utilizada para realizar los cálculos correspondientes a la predicción de contactos, entre otros.

- Agregar restricciones para obtener MSAs no redundantes:
  En base a pruebas realizadas, se agregan restricciones a la generación de los MSAs; por cada ciclo de mutación se imprimía la secuencia de cada ejecución independiente, la cual era exactamente igual a la anterior con una única mutación, esto generaba una redundancia sin sentido en el MSA resultante y afectaba a la performance tanto a nivel de tiempo de ejecución como de memoria utilizada por cada MSA. Por tal motivo, se realizó una modificación para que solamente se imprima la secuencia de una ejecución independiente una vez que la misma haya tenido una cantidad de mutaciones que la diferencie de la secuencia de esa ejecución independiente impresa anteriormente. Para ello se definió que el SCPE reciba otro parámetro más que indique la cantidad de mutaciones (print_sequences_screening) que debía sufrir una ejecución independiente antes de ser impresa. Por ejemplo, si el valor de print_sequences_screening = 10 esto significa que cada ejecución independiente será impresa luego de 10 mutaciones. A su vez se definió que si el valor ingresado era 0 automáticamente el sistema realiza el siguiente screening print_sequences_screening = (protsize*0.40)/3. Esta variable no genera ningún tipo de restricción o condicionamiento al resultado final del procedimiento; es utilizada simplemente para optimizar los tiempos de clusterización que se realiza una vez obtenido el MSA para eliminar redundancia.

- Agregar restricciones para obtener MSAs dentro de un rango de identidad a la secuencia de referencia:
  Otra restricción agregada antes de imprimir una secuencia es verificar que la misma tenga una identidad mayor al 25 % y menor al 62 % con respecto a la secuencia de referencia. Esta restricción, por un lado, es para no permitir que se impriman secuencias muy divergentes a la original (<25%); ni tampoco muy similares (>62%). Si la secuencia tiene un porcentaje de identidad menor al 25 % con respecto a la secuencia de referencia, la divergencia es muy grande y nos arriesgamos a perder la estructura

conformacional de la proteína de referencia (esto también depende del β y del nsus utilizado). De igual forma si la secuencia tiene una identidad > 62%, no estamos incluyendo mucha divergencia para realizar el análisis, obteniendo MSAs con secuencias redundantes para un análisis posterior.

A continuación se detalla el pseudocódigo 3.1.1, simple, obviando diferentes complejidades, de cómo sería la generación de los MSAs evolucionados.

```
Mientras siga realizando mutaciones{
    Para run_independientes desde 1 hasta cantidad de ejecuciones independientes {
        realizar_mutacion(run_independiente)
        Si (mutacion_aceptada y debo imprimir run independiente y identidad run
        independiente > 25 % y identidad run independiente < 62 %){
            Imprimir run independiente en archivo de secuencias
        }
    }
}
```

Pseudocódigo 3.1.1. Pseudocódigo intuitivo para describir la funcionalidad de SCPE e imprimir el MSA evolucionado. Mientras se deban realizar mutaciones, para cada ejecución independiente se realiza una mutación y luego se analiza si debe o no ser aceptada.

3.2. Matriz de Contacto

Una matriz de contacto indica que posiciones de la secuencia primaria de una proteína están en contacto entre sí. Si la proteína tiene un tamaño n entonces la matriz de contacto será de tamaño n*n; y en cada celda se indicará si esas dos posiciones se encuentran en contacto, esto puede definirse como 1 contacto y 0 no contacto. La matriz de contacto es una matriz de origen teórico y experimental que debe definirse en base al estudio que se desea realizar (Figura 3.2.1).

Cuando se ejecuta el software SCPE; unos de los resultados que retorna es la matriz de contacto de la estructura evolucionada, dicha matriz se define teniendo en cuenta los radios de Van der Waals de cada átomo, Tabla 3.2.1. Se define entonces que: existe contacto entre dos residuos si la distancia de cualquiera de sus átomos menos el radio de Van der Waals de los mismos es

menor a 1. Además, solo se tiene en cuenta las cadenas laterales R, obviando los átomos de la cadena primaria (main chain); y no se toman en cuenta como contactos los residuos vecinos (i-1 e i+1 respecto al residuo i) para no generar contactos triviales.

Átomo	Radio
Nitrógeno	1.7
Carbono	1.8
Oxígeno	1.4
Azufre	2
Fósforo	2

Tabla 3.2.1. Radios de Van der Waals utilizados para la definición de matriz de contacto.

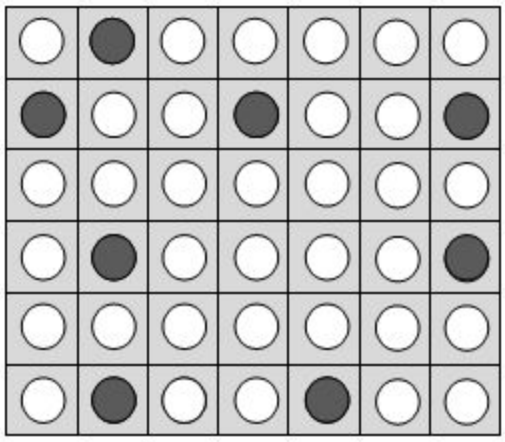


Figura 3.2.1. Esquematización de matriz de contacto de una proteína teórica de 7 aminoácidos. Normalmente se indica a través de un 1 o True las posiciones que se encuentran en contacto (círculos negros) y 0 o False las posiciones que no están en contacto (círculos blancos).

Luego de la lectura de todos los residuos del pdb se retorna la matriz de contacto resultante propuesta con esta definición. Dicha matriz será utilizada para realizar diferentes cálculos, siendo el principal la predicción de contactos mediante los métodos de covariación.

3.3. Cálculo de conservación. Logo de secuencias

En un alineamiento múltiple de secuencias, el grado de similitud entre los residuos que ocupan una posición concreta puede interpretarse como una medida aproximada de conservación. Los residuos conservados generalmente son posiciones que juegan un rol importante a nivel funcional, estructural o ambos. Existen varias métricas para medir la conservación, la más conocida refiere a la entropía de Shannon

$$S\ \ (p) = -\sum_{a} P_a log_2(p_a)$$

donde P(a) es la frecuencia del aminoácido a en la posición a.

Durante el proyecto de investigación utilizaremos lo que se conoce actualmente sobre la distribución background de los aminoácidos en la naturaleza; para ello existe una entropía relativa de Shannon, también denominada de Kullback-Leibler (KL). Para cada columna en el alineamiento, se calcula como

$$D(p||q) = \sum_{a} p_a log_2 \left(\frac{p_a}{q_a}\right)$$

donde P(a) es la frecuencia del aminoácido a en la posición a y Q(a) es la frecuencia que se conoce en la naturaleza (Cover and Thomas, 1991). En este enunciado, y durante todo el proyecto, las dos métricas miden sus valores en bits.

Una vez definidos los cálculos de conservación que utilizaremos sobre los alineamientos, tanto naturales como artificiales, incluimos el concepto-herramienta de logo de secuencias que se utilizará para la visualización clara de la conservación. La idea de logo de secuencias fue introducida en 1990 por

Schneider and Stephens [12]. El objetivo era visualizar de forma rápida y concisa las características más importantes de un alineamiento múltiple de secuencias, describiendo la información en cada una de las posiciones: conservación y frecuencia de los aminoácidos (Figura 3.3.1).

La herramienta utilizada para graficar los logos de secuencias es Seq2Logo [23], utiliza tanto Shannon como Kullbak-Leibler para realizar los cálculos. Cuando los resultados sean descritos se indicará claramente cuáles de las dos métricas fue utilizada.

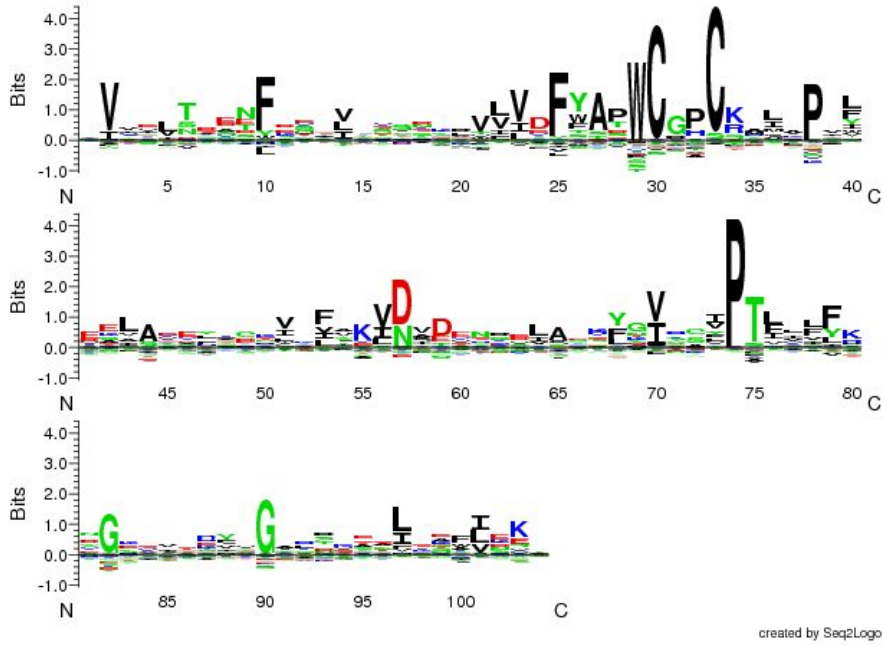


Figura 3.3.1. Logo de secuencias calculado con Seq2Logo, utilizando el algoritmo Kullbak-Leibler. Se visualiza de forma descriptiva los sitios conservados y sus aminoácidos correspondientes. Será utilizado durante la exposición de resultados de este trabajo. Proteína de ejemplo: THIO_ECOLI.

3.4. Coevolución

Existen varias definiciones de coevolución, pero el principio es el mismo, se define coevolución molecular como el cambio en un locus que afecta la presión de selección de otro locus, siendo este cambio recíproco [1]. La evolución de muchos genes y proteínas no es independiente sino que están ligadas a la evolución de

otros componentes. Lo mismo sucede a nivel de residuos individuales de una proteína, lo que se denomina coevolución sitio específica.

Yanofsky et al [13] y Fitch y Markowitz [14] propusieron que la coevolución, también denominada coadaptación por los autores, de secuencias es importante para mantener una estructura y función apropiada de la proteína. Estos dos grupos propusieron diferentes mecanismos para explicar esta coevolución. Específicamente durante los experimentos realizados en [13] se proponía realizar mutaciones de aminoácidos en un sitio específico de una proteína, obteniendo diferentes proteínas mutadas, donde constataron que algunas dejaron de ser funcionales enzimáticamente, pero otras lo continuaron siendo. En las mismas observaron que no solo el sitió mutado había sido modificado sino que también un sitio distante a 36 residuos había también acompañado dicha mutación (cambios concertados). Dieron cuenta luego, que este cambio en el segundo sitio reflejaba una adaptación de la proteína para continuar siendo funcional y que por consiguiente existía una relación entre estas dos posiciones para mantener la función de la misma.

La función y la conformación espacial de las proteínas imponen restricciones en las variaciones de residuos que pueden ser aceptados para que la proteína siga siendo funcional. Así, es esperable que los sitios que se encuentran próximos en la estructura tridimensional sean los que impongan restricciones en los cambios unos a otros, es decir, que cambien de manera concertada o coevolucionen para continuar manteniendo la función o la estructura de la proteína [3,4,7,9,14]. En la Figura 3.4.1 se esquematiza el concepto de coevolución entre residuos interactuantes.

Es importante, en este punto, distinguir los conceptos de coevolución y de covariación; la coevolución implica que el cambio sea recíproco y debido a un proceso evolutivo común. En cambio, la covariación describe un cambio simultáneo de aminoácidos observado inherente en una familia de proteínas, sin

importar las causas de los mismos [15]. La covariación observada en las familias de proteínas no implican coevolución, existen otros factores a tener en cuenta que veremos a continuación; pero está claro que las posiciones que coevolucionan producen posiciones que covarían en familias de proteínas [16].

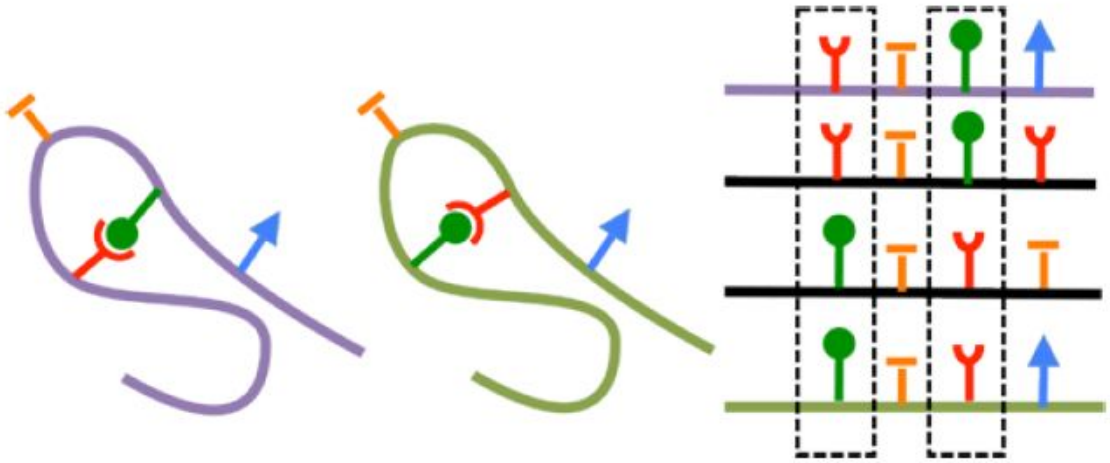


Figura 3.4.1. Representación de coevolución por estructura en un sitio específico. A la izquierda se representan aminoácidos con formas complementarias (verde y rojo), que han intercambiado sus posiciones en la cadena polipeptídica. A la derecha, el alineamiento múltiple de secuencias correspondiente donde se puede observar la covariación de residuos en las posiciones marcadas en línea punteada. Imagen extraída de http://gremlin.bakerlab.org/gremlin_faq.php.

Atchley et al. [1] sugirieron que la covariación observada entre las posiciones i y j en una proteína está compuesta de una señal proveniente de restricciones estructurales y funcionales, junto con ruido de fondo agregado por una filogenia compartida y eventos estocásticos. De esta forma, las señales estructurales y funcionales están superpuestas sobre el ruido filogenético y otros procesos aleatorios. Así, la covariación entre una posición i y j en un MSA puede ser descompuesta en diferentes términos como se muestra en la Ecuación 3.4.1

$$C_{ij} = C_{\text{filogenia}} + C_{\text{estructura}} + C_{\text{función}} + C_{\text{interacción}} + C_{\text{estocástica}}$$

Ecuación 3.4.1. Señales que forman parte de la covariación observada en una proteína.

Numerosos estudios se han realizado para desarrollar métodos y distinguir los diferentes tipos de coevolución expresados en la ecuación anterior. Existen enfoques filogenéticos utilizando, por ejemplo, matrices de distancia; y enfoques basados en la Información Mutua.

3.5. Información Mutua

El concepto de Información Mutua (Mutual Information) proveniente del área de Teoría de la Información, se encarga de medir el grado de relación (dependencia) entre dos variables aleatorias X, Y. MI (X,Y) es una medida de la información provista por los pares de símbolos (x, y), el resultado es la correlación entre las variables aleatorias. Si X e Y son independientes, entonces conocer X no da información sobre Y, y viceversa, no existe relación, por lo que su información mutua es cero, MI (X,Y) = 0.

A nivel biológico, dentro del marco de las familias de proteínas, este concepto puede ser utilizado para estimar el grado de la relación de coevolución entre dos posiciones de un MSA [5,17,18]. Para el análisis de secuencias, la MI entre dos posiciones (dos columnas de un MSA) refleja la disminución de la incertidumbre en una posición a partir del conocimiento de otra posición. Por ende, la MI mide la información compartida por dos columnas de un MSA. Dadas dos posiciones i y j cualesquiera en un MSA, la MI entre ellas está dada por la relación que se muestra en la ecuación 3.5.1.

$$MI_{ij} = \sum_{a}\sum_{b} P(a_i, b_j) \cdot log\left(\frac{P(a_i, b_j)}{P(a_i) \cdot P(b_j)}\right)$$

Ecuación 3.5.1. Calculo de Información Mutua. P(ai,bj) es la frecuencia de ocurrencia del aminoácido a en la posición i y del aminoácido b en la posición j, en la misma secuencia. P(ai) es la frecuencia del aminoácido a en la posición i y P(bj) es la frecuencia del aminoácido b en la posición j.

La ventaja del uso de MI para cuantificar la coevolución radica en la aplicabilidad del método sin requerir conocimientos acerca de la relación entre los residuos en el MSA y su dinámica evolutiva. Para los valores positivos de MI, su magnitud depende de la fuerza de covariación entre ambos sitios [19].

Existen varios puntos a tener en cuenta para el cálculo de coevolución utilizando MI, estos son:

- Dos posiciones que no varían (conservadas):

  Se asume independencia evolutiva al no tener evidencia de covariación entre ellas. Por ejemplo la posición *x* para todos sus valores es A; y la posición *y* para todos sus valores es C. Representarán un caso extremo de coaparición, no hay evidencia de correlación, y la MI entre ellas es cero.

- Relación con la conservación del MSA:

  El poder del método de MI para la predicción de coevolución está íntimamente relacionado con la conservación del MSA [18, 20].

- Afectación por la historia filogenética:

  Como se ha visto en la ecuación 3.4.1 de la sección anterior, el cálculo de coevolución se ve afectado por la señal filogenética $C_{filogenia}$. Por tal motivo, diferentes estudios [17,21-24], se centraron en corregir esta señal filogenética para disminuir su efecto en los cálculos de coevolución. Esto sucede porque las secuencias proteicas no son independientes; contienen información evolutiva histórica de sus ancestros. Estudios realizados [24], dan muestra de ejemplos donde la señal filogenética emula a la coevolución. En la Figura 3.5.1 podemos observar que las posiciones 1 y 2, dan un alta señal de MI pero en realidad dichas mutaciones fueron realizadas en sus ancestros y de forma totalmente independiente.

- Falta de un set de datos de referencia:

  No podemos evaluar los distintos métodos de coevolución ya que no tenemos acceso a los datos evolutivos reales. Por tal motivo, se asumen aproximaciones teóricas: dos residuos que están en contacto coevolucionan [21,24,25]. La definición de contacto que es utilizada durante este trabajo, se ha descrito en la sección 3.2. Esta aproximación es bastante intuitiva, para la mayoría de los pares de residuos que

coevolucionan puede asumirse que si se encuentran a una distancia menor a *n* Å están en contacto; ya que la modificación en uno de los residuos implicaría el cambio en el otro para que el contacto continúe y no se pierda la estructura.

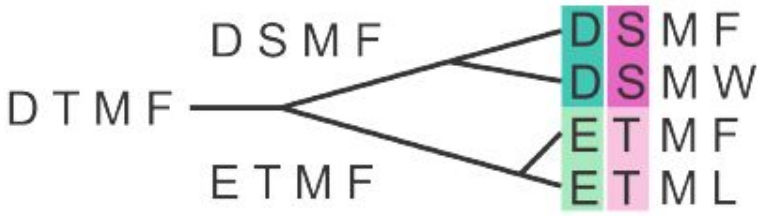


Figura 3.5.1. Covariación de la primer y segunda columna. Señal filogenética que imita a la coevolución. Figura obtenida de [24].

En conclusión para realizar un correcto análisis de coevolución utilizando MI es necesario que el alineamiento cumpla con ciertas características: la calidad del mismo, el número de secuencias que contiene y el nivel de divergencia de sus secuencias tienen un impacto en los resultados que se obtengan.

3.6. Corrección de redundancia : Clusterización

La redundancia es un problema muy común en los alineamiento múltiples de secuencias. Esta información redundante, frecuentemente, no ofrece información extra valiosa e inclusive puede tener un impacto negativo si realizamos cálculos con MI. La redundancia se da por ejemplo, debido a la secuenciación de múltiples cepas bacterianas o a la selección de especies a secuenciar (especies de mayor interés), por tal motivo es práctico considerar un conjunto representativo de secuencias del alineamiento y a su vez mejorar la eficiencia para análisis posteriores que se realicen sobre el MSA sin perder información relevante para el mismo.

Las restricciones para imprimir las secuencias que se agregaron al software SCPE no son suficientes para evitar redundancia, ya que en diferentes

ejecuciones independientes pueden darse secuencias que hayan evolucionado de forma similar, por tal motivo una vez obtenido los MSAs evolucionados se aplica lo que se denomina una clusterización. La misma se realiza definiendo un umbral (*threshold)* o corte de identidad, que puede ir de 0 a 100 %. Estudios realizados [5, 26] han dado cuenta que el valor de corte óptimo para trabajar con MSAs es de 62%.

El software CD-HIT [26] es una de las herramientas más utilizadas para eliminar la redundancia en alineamientos múltiples de secuencias.

CD-HIT utiliza el algoritmo incremental *greedy* implementado por Holm and Sander (1998). En primer lugar las secuencias son ordenadas de forma decreciente por longitud. La de mayor longitud se convierte en la secuencia representativa del primer *clúster*. Todas las demás secuencias son comparadas con las secuencias representativas de cada clúster, si existe una similitud con alguno de ellos, bajo un cierto *threshold*, es incluida dentro de ese clúster, en caso contrario un nuevo clúster se crea teniendo a dicha secuencia como representativa.

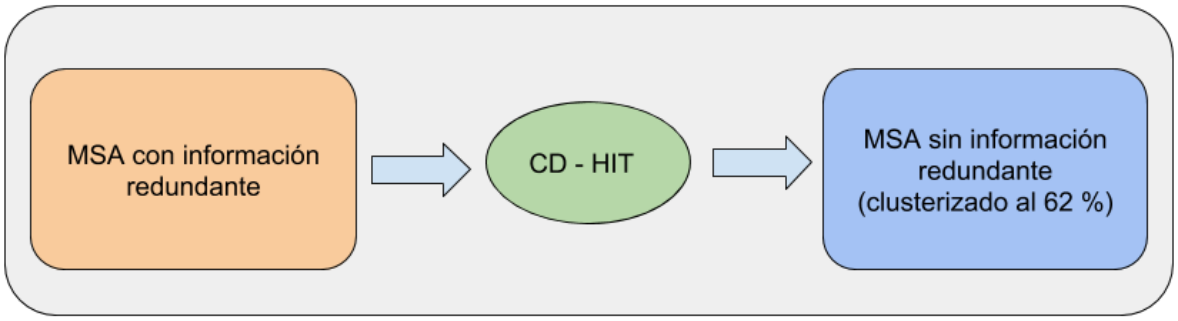


Figura 3.6.1. Dado un alineamiento múltiple de secuencias se realiza una clusterización para eliminar redundancia y obtener un alineamiento clusterizado al 62%. Todas las secuencias que tengan un 62% de identidad serán agregadas a un clúster que tendrá una única secuencia representativa la cual se incluye en el alineamiento resultante.

Una de las técnicas que utiliza CD-HIT es el uso de filtrado con palabras cortas (short word filters). Dos proteínas que comparten cierta identidad deben tener al menos cierto números idénticos de dipéptidos, tripéptidos, pentapéptidos,

etc. Por ejemplo, dos secuencias de 100 residuos de longitud que tiene el 85 % idénticos, van a tener como mínimo 70 de dipéptidos idénticos, 55 tripéptidos idénticos y 25 pentapéptidos idénticos. Entonces, directamente los pares de secuencias que no satisfacen estas condiciones no son necesarios que sean alineados, esto agiliza el proceso de clustering.

Existen convenciones en CD-HIT en cuanto al tamaño de palabra a utilizar y viene relacionado con el *threshold* que se quiere obtener, en el caso de este proyecto de investigación se utiliza una palabra de tamaño 4 para una clusterización al 62%.

3.7. Corrección por bajo número de secuencias: Corrected Mutual Information

En la realidad los MSAs que se encuentran disponibles pueden tener pocas secuencias para analizar, y por consiguiente el cálculo de ocurrencia de los aminoácidos se realiza con poca información, lo que genera diferentes inconvenientes para su análisis:

- Calculo de MI de baja calidad: al tener poca información la ocurrencia de los aminoácidos se realiza con pocos datos, lo que afecta el cálculo final de MI.
- Muestreo no suficiente: una combinación de aminoácidos en una posición i y j puede ni siquiera llegar a aparecer dentro del MSA por la simple cuestión que hay pocas observaciones.

Durante el desarrollo de este proyecto se utiliza la corrección a la información mutua desarrollada en el trabajo de C. M. Busjle [21], en donde se introducen varias correcciones a la fórmula original de MI para cuando se trabaja con un bajo número de secuencias, para borrar la señal filogenética y disminuir el sesgo por redundancia de secuencias.

Se introduce una corrección simple para lidiar con MSAs con pocas secuencias. La probabilidad de los aminoácidos, $P(a_i; b_j)$, se calcula a partir de N(a; b), el número de veces que el par (a; b) es observado en las posiciones i y j en el MSA más una constante λ, quedando la ecuación de la frecuencia del aminoácido a en la posición i y el aminoácido b en la posición j de la siguiente forma

$$P(a_i, b_j) = \frac{\lambda + N(a_i, b_j)}{N}$$

Donde

$$N = \sum_{a,b} (\lambda + N(a_i, b_j))$$

$$P(a_i) = \sum_{b} P(a_i, b_j)$$

$$P(b_j) = \sum_{a} P(a_i, b_j)$$

La nueva forma de realizar el cálculo $P(a_i; b_j)$ es luego agregada a la ecuación 3.5.1.

La introducción del parámetro λ responde a la necesidad de lidiar con el número bajo de observaciones (secuencias). Se asigna un valor inicial $N(a_i, b_j) = \lambda$ para todo par de aminoácidos posibles. En otras palabras, significa que todos los pares de de aminoácidos serán observados al menos λ veces. Por consiguiente, para MSAs con bajo número de observaciones en donde es probable que varios pares de aminoácidos no sean observados, el valor λ tendrá injerencia en el cálculo. Por el contrario, en el caso de MSAs que dispongan de una gran cantidad de secuencias, la mayoría de los pares de aminoácidos serán observados probablemente al menos una vez, en dicho caso la influencia del λ será menor.

3.8. Transformación a Z-Score: comparación de MI entre MSAs

Cada valor de MI entre un par dado de posiciones es comparado con la distribución de valores de MI obtenidos a partir de un grupo de MSAs aleatorizados. El Z-score es calculado como el número de desviaciones estándar que el valor de MI observado se aparta de la media obtenida con los MSAs aleatorios. Para cada MSA se realizaron 100 permutaciones, manteniendo los gaps fijos en sus posiciones originales. Se ensayaron dos métodos de permutación[21], uno basado en columnas (aleatorización vertical) y otro basado en secuencia (aleatorización horizontal). El primero desafía la hipótesis de que las secuencias son homólogas y están correctamente alineadas, pero que las columnas no están correlacionadas. Mientras que el segundo método desafía la hipótesis de que las secuencias no son homólogas. El mejor resultado fue obtenido con el segundo método de permutación, por lo que el Z-score no testea de manera adecuada la hipótesis nula (que las columnas no sean correlacionadas) por lo que debe ser interpretado solamente como un puntaje predictivo más, que permite la comparación entre familias de proteínas. Durante este trabajo se utiliza el concepto de Z-score de MI corregido, y de aquí en adelante, se referirá al mismo simplemente como puntaje de MI.

3.9. Análisis de acoplamiento directo. Inferencia Gaussiana

El análisis de acoplamiento directo, a partir de aquí DCA por sus siglas en inglés (Direct coupling Analysis), está basado en el principio de máxima entropía de Jaynes [27,28], que conduce a modelos estadísticos de familias de proteínas en términos de los llamados modelos Potts o campos aleatorios de Markov.

El concepto y objetivo para estudiar la coevolución a través de este método se basa en que las correlaciones entre los aminoácidos que se producen en dos

posiciones en una familia de proteínas, es decir, entre dos columnas del MSA, pueden resultar no solo de acoplamientos coevolutivos directos. Si una posición *a* tiene una alta señal de correlación con *b*, y a su vez *b* tiene una alta señal de correlación con *c,* entonces va a existir correlación entre *a* y *c* aunque en realidad no están realmente correlacionadas. La idea del modelo es determinar cuando estamos en presencia de una correlación directa y cuando es un correlación indirecta o transitiva, y descartar estas últimas para generar un modelo solo de correlación directa [29,30], de ahí el nombre Direct coupling analysis, al cual también se lo denomina Direct Information (DI).

Métodos como el análisis de acoplamiento directo [29] y la estimación de covarianza inversa dispersa (PSICOV) [31], han logrado un avance hacia el objetivo de evitar la correlación indirecta; y sus predicciones se han implementado con éxito en la predicción de contactos de la estructura terciaria y cuaternaria. Sin embargo, debido a la naturaleza discreta de los aminoácidos, la inferencia exacta requiere un tiempo exponencial, por tal motivo se necesitan aproximaciones eficientes para la aplicación práctica.

El segundo y tercer modelo de coevolución que será utilizado dentro del proyecto utiliza un modelo de inferencia gaussiana como una variante del DCA [32], puede encontrarse en la bibliografía como GaussDCA. En donde las variables de aminoácidos discretos son reemplazadas por variables aleatorias gaussianas continuas. Puede entenderse como una aproximación al modelo Potts de máxima entropía en el cual:

- Se libera la restricción de discretización, es decir, se permiten valores continuos para las variables que representan aminoácidos;
- se supone un modelo de interacción gaussiano; y
- se introduce una distribución previa para compensar el submuestreo de los datos.

Pueden utilizarse dos medidas (scores) con este método: Direct Information (DI) y Frobenius norm (FROB); el método de GaussDCA descrito con las medidas DI y FROB serán el segundo y tercer método de coevolución empleado para el análisis durante el proyecto. A partir de aquí nos referiremos a las mismas simplemente como DI y FROB. Los parámetros opcionales para la ejecución del método de coevolución son los que se indican por defecto en los estudios realizados por los autores [32].

3.10. Estimación de covarianza inversa dispersa (PSICOV)

La estimación de covarianza inversa dispersa (PSICOV)[31] es otro de los métodos de coevolución que han logrado un avance hacia el objetivo de evitar la correlación indirecta; y sus predicciones se han implementado con éxito en la predicción de contactos de la estructura terciaria y cuaternaria.

El punto de partida del método es considerar un alineamiento con m columnas y n filas, en donde cada fila representa una secuencia homóloga diferente y cada columna un conjunto conjunto de aminoácidos, considerando a los gaps como un tipo de aminoácido extra. Se puede realizar el cálculo de la matriz 21 x 21 de covarianza:

$$S_{ij}^{ab} = \frac{1}{n}\sum_{k=1}^{n}(x_i^{ak} - \bar{x}_i^a)(x_j^{bk} - \bar{x}_j^b)$$

Cualquier elemento individual de la matriz da la covarianza del aminoácido de tipo a en la posición i con el aminoácido b en la posición j. Al calcular la matriz inversa de covariación, la matriz de precisión o concentración (Θ) es obtenida, desde la cual una matriz de coeficientes de correlación parcial para todos los pares de variables puede ser calculada:

$$\rho_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}$$

La matriz de covarianza empírica producida en esta aplicación es singular debido a que no se observarán todos los aminoácidos en cada sitio, inclusive en familias con alta población de secuencias, y por lo tanto habrá más variables que observaciones. Este inconveniente también puede observarse en otras áreas, como por ejemplo en la reconstrucción de las redes de genes en donde el número de variables es a menudo menor que la dimensionalidad del problema.

Diferentes enfoques se han propuesto para permitir la estimación de covarianza inversa cuando la matriz de covarianza de la muestra no se puede invertir directamente; una de las técnicas más potentes es la estimación de covarianza inversa dispersa. Este problema ha sido estudiado por varios autores, durante el desarrollo de PSICOV se siguió la formulación conocida como el método gráfico de Lasso (en inglés, graphical Lasso) [33], y la implementación realizada por Friedman [34].

Si *S* es la matriz de covarianza empírica computada de una secuencia de d vectores bidimensionales, x1, …., xn, muestreada de una distribución de probabilidad fija pero desconocida. La matriz *S* puede ser computada como $S_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_i^k -$ para cada *i*, *j*=1,...,d, donde $\bar{x}$ es la media empírica. La gráfica de Lasso es un método estático que estima la covarianza minimizando la función objetivo:

$$\sum_{ij=1}^{d} S_{ij}\Theta_{ij} - \log\det\Theta + \rho\sum_{ij=1}^{d}|\Theta_{ij}|$$

Cuando nos encontramos con familias con pocas secuencias o con regiones altamente conservadas, el tiempo para alcanzar la convergencia puede ser problemático. Para acelerar la convergencia, particularmente en los peores

escenarios, se condiciona la matriz de covarianza de la muestra disminuyendo hacia un estimador insesgado altamente estructurado:

$$S' = \lambda F + (1 - \lambda) S$$

donde *F* es el estimador estructurado de la matriz y $\lambda \in [0, 1]$ es lo que se denomina el parámetro *shrinkage*. Se utiliza el simple enfoque de aumentar gradualmente λ hasta que la matriz de covarianza ajustada ya no sea singular. Habiendo condicionado la matriz de covarianza por reducción, el algoritmo gráfico de Lasso es utilizado para calcular inversión dispersa.

Como último paso para obtener una predicción de los residuos que están en contacto, para las columnas *i* y *j*, la ℓ_1-norm es calculada para la submatriz 20 x 20 de Θ correspondiente a los 20 x 20 tipos de aminoácidos observados en las dos columnas (los gaps son ignorados):

$$S_{ij}^{\text{contact}} = \sum_{ab} |\Theta_{ij}^{ab}|$$

Finalmente para calcular un *score* que reduzca la entropía y la señal filogenética, se utiliza APC (Average product correction). La ecuación final de PSICOV que da el *score* para la posiciones *i* y *j* está dada por

$$PC_{ij} = S_{ij}^{\text{contact}} - \frac{\Im_{(i-)}^{\text{contact}} \Im_{(-j)}^{\text{contact}}}{\Im^{\text{contact}}}$$

donde $\Im_{(i-)}^{contact}$ es la media entre la columna *i* y las demás columnas, $\Im_{(-j)}^{contact}$ es el equivalente para la columna *j*, y $\Im^{contact}$ es la media de todo el alineamiento. Finalmente el *score* corregido puede ser convertido fácilmente en un valor estimado predictivo ajustando una función logística a la distribución observada de los *scores*.

Dada la complejidad del procedimiento para la obtención de la ecuación en el cálculo de PSICOV, más detalles sobre el mismo puede obtenerse refiriéndose al trabajo dispuesto en [31].

3.11. Modelo predictivo de contactos. Desempeño: AUC

En las secciones anteriores se ha destacado que el cálculo MI, DCA o PSICOV, sobre un alineamiento de múltiples secuencias, nos brinda información sobre la covariación entre las posiciones del mismo y por consiguiente se puede predecir cuales son los residuos que están en contacto en la estructura tridimensional, coevolución por estructura.

El predictor que se diseña es de índole binario, los resultados pueden ser positivos (Contacto) o negativos (No contacto). El desempeño predictivo, se evaluará a través del área bajo la curva ROC (AUC, por su nombre en inglés Area Under the ROC Curve). Para calcular el AUC es necesario definir Verdaderos Positivos (TP, True Positive), Falsos Positivos (FP), Verdaderos Negativos (TN, True Negative) y Falsos Negativos (FN). La matriz de contacto descrita en 3.2, define cuales son los verdaderos y los falsos reales; mientras que el modelo predictivo se define a través de analizar la covariación, utilizando los métodos descritos anteriormente, de las posiciones del MSA. El modelo entonces define cuatro posibles resultados los cuales se describen a través de una matriz de confusión (Tabla 3.11.1).

	Valor en la realidad Matriz de Contacto	
Predictor Covariación MSA	Verdaderos Positivos (TP)	Falsos Positivos (FP)
	Falsos Negativos (FN)	Verdaderos Negativos(TN)

Tabla 3.11.1. Matriz de confusión del predictor desarrollado.

A partir de la matriz de confusión se definen otros valores, entre ellos pueden destacarse:

- La razón de éxitos es TP / TP + FN (TPR), también denominada sensibilidad.
- La razón de falsos positivos es TN / TN + FP (FPR).
- La especificidad se define como 1 - FPR.

La curva ROC es una forma visual de analizar la calidad del predictor, solo tiene en cuenta las razones de Verdaderos Positivos (TPR) y la de Falsos Positivos (FPR). Un valor de AUC 1 indica una predicción perfecta y un valor de 0.5 una predicción al azar. Valores superiores a 0.7 determinan una aceptable predicción de contactos, visualmente podemos observar estos valores en la Figura 3.11.1.

Este modelo predictivo será utilizado tanto para analizar los MSAs evolucionados artificialmente, como también para los MSAs naturales, obtenidos a través de Pfam, a los que pertenezcan las proteínas estudiadas.

En resumen, los elementos del proceso predictivo son:

- Un alineamiento múltiple de secuencias sobre el cual se realizará el cálculo de la covariación a través de MI, DI, FROB y PSICOV.
- La proteína de interés, particularmente seleccionando una estructura cristalográfica, la cual será utilizada para obtener la matriz de contacto.

Con estos elementos disponibles se procede a realizar un modelo predictivo de coevolución para cada uno de los métodos utilizados.

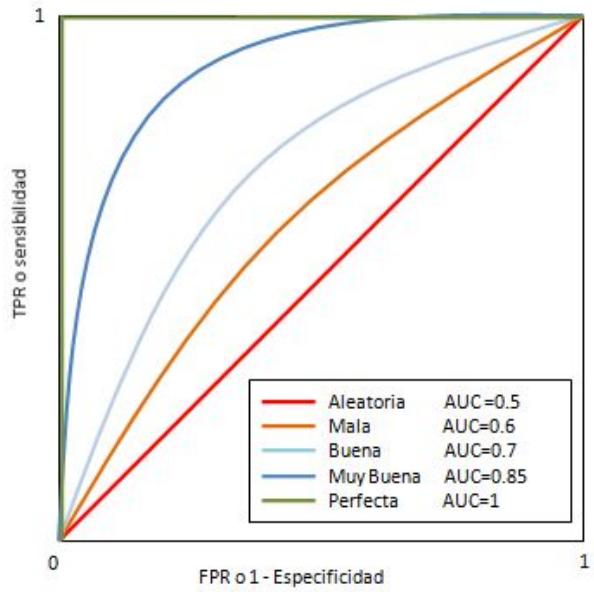


Figura 3.11.1. Diferentes curvas ROC con sus respectivos valores de performance predictivo. Una curva ROC con un AUC=0.5 indicaría un método predictivo totalmente al azar; en cambio con un AUC = 1 es un método predictivo perfecto.

4. Proteína THIO_ECOLI

Previamente a simular la evolución de una proteína y realizar el análisis sobre el MSA evolucionado resultante, debemos conocer cuáles son los valores de estudio de la misma en su contexto natural. La proteína que será de interés y será tomada como referencia es la THIO_ECOLI, y la estructura a evolucionar será la cadena A de la 2TRX.

La proteína THIO_ECOLI pertenece a la familia de Pfam PF00085. La base de datos Pfam es una gran colección de dominios de familias de proteínas [7]. Cada familia está representada por un MSA y un modelo oculto de Markov (HMM, del inglés Hidden Markov Model). A partir de aquí al MSA de la familia PF00085 lo denominaremos MSA natural.

La proteína THIO_ECOLI se encuentra presente dentro de la familia PF00085 desde la posición 4 a la 107; por tal motivo dentro del MSA de la familia la encontraremos anotada como THIO_ECOLI/4-107.

4.1. Procedimiento de Gapstrip

Para realizar el análisis del MSA natural, debemos en primer lugar, poner como referencia a la secuencia de la proteína de interés. Esto significa aplicar al MSA natural un procedimiento que denominaremos gapstrip, este proceso consiste en remover todas las columnas del MSA donde hay un gap en la secuencia de referencia. En la figura 4.1.1 vemos como la secuencia de referencia, la THIO_ECOLI/4-107, permanece inalterada, queda primera en el alineamiento y sin gaps. Adicionalmente, todas las columnas que contengan un porcentaje de gaps > 50% no son tenidas en cuenta en los cálculos; al igual que las secuencias con una cobertura < 50% del ancho del alineamiento.

Todas las secuencias ahora respetan la longitud y las posiciones de la secuencia de referencia; y a partir de allí la posibilidad de realizar los predicción de contactos, con el MSA resultante luego del gapstrip, y la matriz de contacto basada en la estructura de estudio.

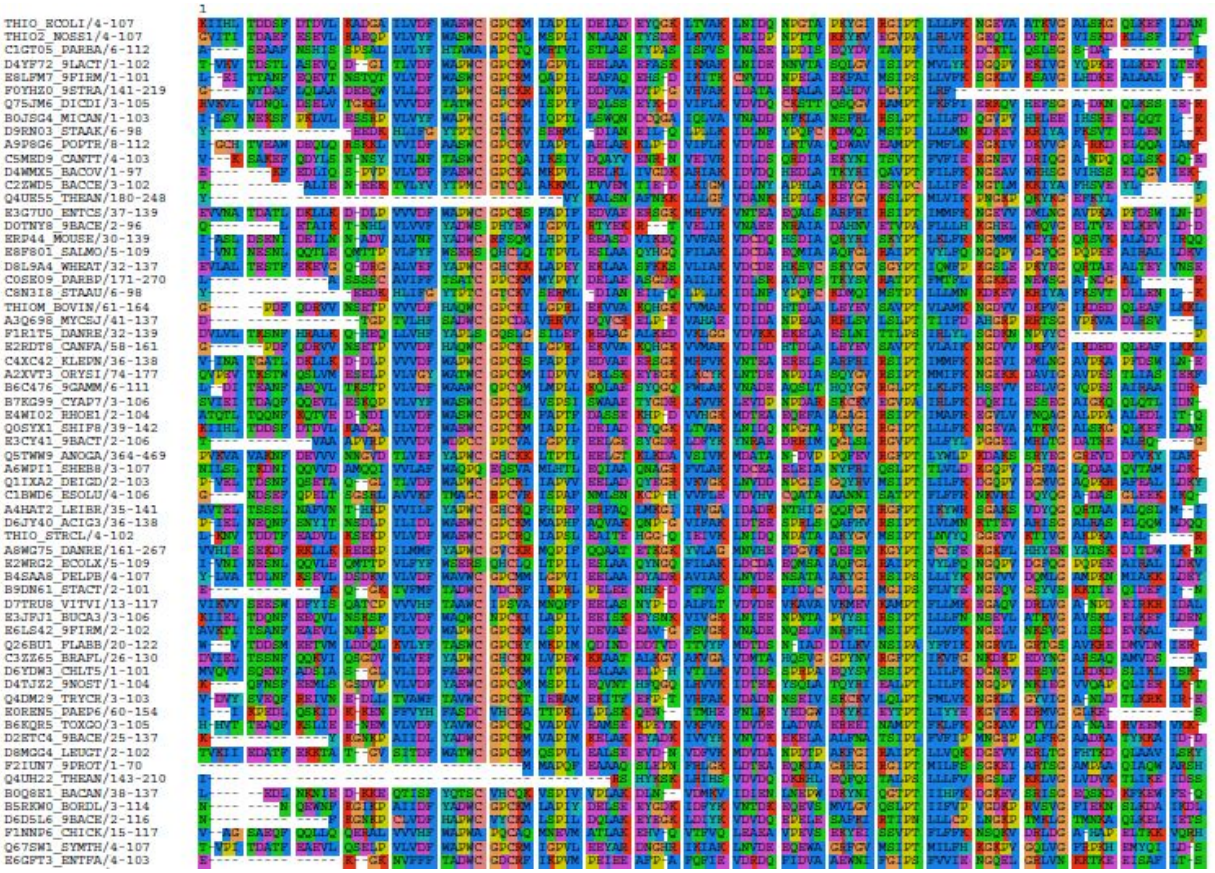


Figura 4.1.1. MSA natural de la familia PF00085 luego de realizar el gapstrip definiendo la secuencia de referencia THIO_ECOLI/4-107, que se encuentra en el primer lugar del alineamiento.

4.2. Predicción de contactos del alineamiento natural

El desempeño predictivo del algoritmo con el alineamiento natural para los métodos de covariación analizados da valores de AUC entre 0.83 y 0.9, lo cual indica un desempeño muy satisfactorio para cualquiera de los métodos. Las curvas ROCs pueden observarse en la Figura 4.2.1. Es el primer resultado elemental obtenido y totalmente esperado, en base a lo que indican los estudios sobre el campo [1-5] que hemos analizado en secciones anteriores: dada una alta señal de covariación entre dos posiciones de un alineamiento múltiple de secuencias se puede predecir que las mismas están en contacto en la estructura tridimensional.

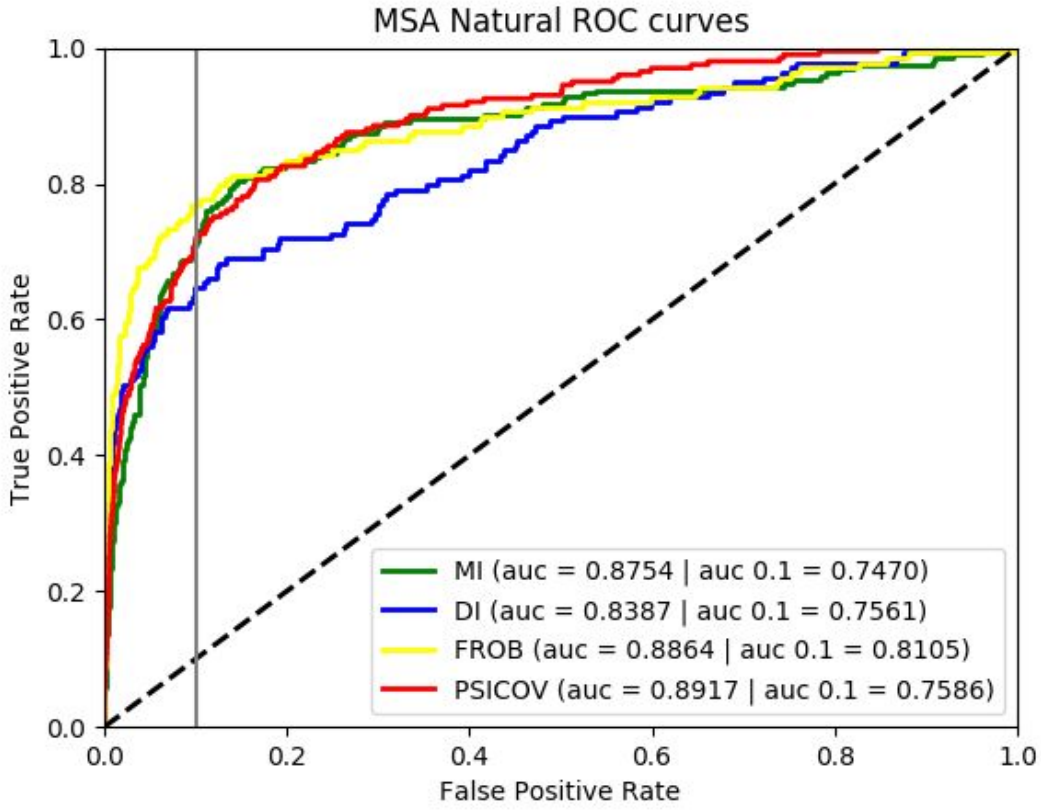


Figura 4.2.1. Curvas ROCs de los métodos de coevolución del alineamiento natural de la familia PF00085 de Pfam, con la proteína THIO_ECOLI/4-107 puesta como referencia luego del proceso de gapstrip.

Entre los resultados obtenidos se destacan los métodos de PSICOV y FROB; el primero obtiene el mejor resultado de AUC con 0.8917, pero el método FROB tiene un desempeño más alto en los primeros valores de la curva, con un AUC 0.1 de 0.81 obteniendo una diferencia apreciable por sobre los demás.

Estos resultados serán utilizados durante todo el proyecto de investigación para realizar las comparaciones correspondientes con los alineamientos evolucionados artificialmente en los capítulos posteriores.

5. Simulación: La proteína THIO_ECOLI

Durante las primeras ejecuciones y pruebas del procedimiento se obtuvieron resultados positivos: la simulación de evolución de la proteína THIO_ECOLI, utilizando la cadena A de la estructura 2TRX, ha generado MSAs artificiales, los cuales al analizar su desempeño predictivo de contactos dieron valores de AUC satisfactorios, entre 0.7 a 0.8. Al notar que los resultados eran distintos según los parámetros utilizados para la evolución con SCPE (β, nsus, runs), surgió el objetivo de conocer cuáles eran los valores óptimos de dichos parámetros, y de esta forma obtener el valor más alto de AUC para cada uno de los métodos.

5.1. Optimización de los parámetros de SCPE

Para poder definir cuáles son los parámetros óptimos del SCPE, durante la evolución y análisis de la estructura 2TRX, se realizó una comparación exhaustiva con un abanico de valores extenso, con el fin de detectar la combinación de valores de β, nsus y runs independientes que generen el MSA evolucionado con mejor desempeño predictivo para cada uno de los métodos de coevolución.

Al analizar los primeros resultados, se observó que si la presión para la restricción estructural, parámetro β, tiene un valor bajo (0.001, 0.01) entonces los valores de AUC no son satisfactorios, la predicción de contactos es aleatoria. Esto era de esperarse ya que estos valores implican una relajación estructural muy alta y por consiguiente los MSAs evolucionados con estos valores tienen una gran divergencia con la secuencia de referencia por lo que la mantención de la estructura se ve afectada. En la Figura 5.1.1 podemos apreciar las curvas ROCs, del método MI, con diferentes runs para el valor β=0.1 y nsus=3 que sirve como ejemplo. Resultados igualmente aleatorios se obtuvieron con los otros métodos de covariación.

A medida que la presión estructural comienza a aumentar los resultados de AUC tienden a mejorar, y se observa que el parámetro de ejecuciones independientes tiene una relación directa con el aumento de AUC. La cantidad de runs indican las secuencias en estado inicial que comenzarán a evolucionar de forma independiente, lo que genera que el MSA evolucionado contenga una población más grande de secuencias con divergencia secuencial pero siempre manteniendo la presión estructural indicada. Por ejemplo es el caso de β=0.5 y nsus=2, cuya información se encuentra en la Tabla 5.1.1. A mayor cantidad de ejecuciones independientes mayor será el tamaño del MSA obtenido, pero también mayor será el tiempo de ejecución del procedimiento. En la Figura 5.1.2 se observa la mejora de las curvas ROC para los métodos MI y DI, que sirven como ejemplos, a medida que aumenta el número de runs. En particular se

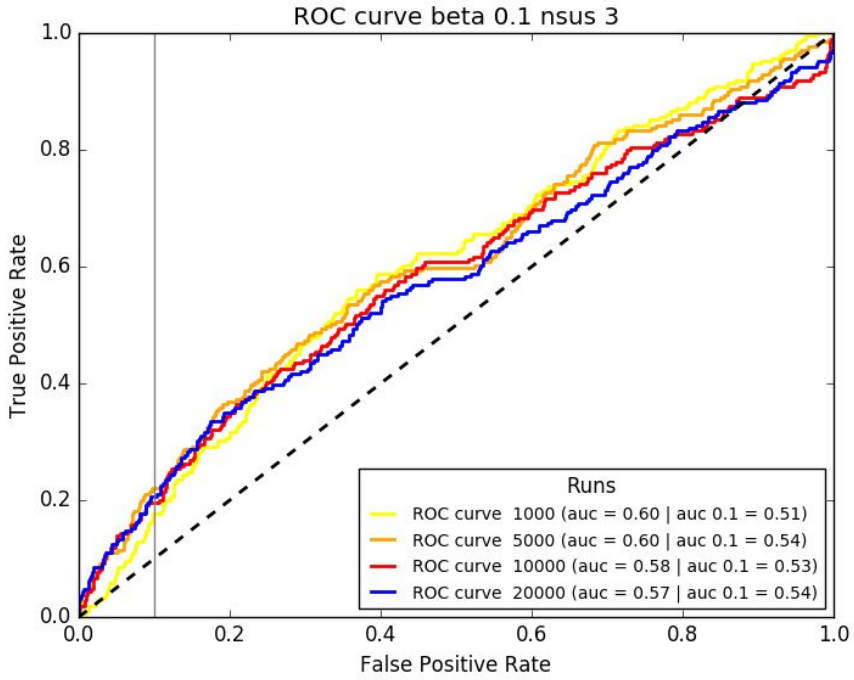


Figura 5.1.1. Valores de AUC y Curvas ROC de β=0.1 y nsus=3 para el método de coevolución MI. Muestran un resultado malo (al azar) en la predicción de contactos. Esto se debe a la relajación estructural al ser β muy chico.

observa que para β=0.5 y nsus=2 los resultados de MI son mejores que los de DI, pero esto no implica que sea así de forma general, cada uno de los métodos de coevolución tienen sus óptimos con diferentes parámetros del SCPE; esta figura es solo para verificar la relación con runs independientes; la comparación entre los métodos se realizará posteriormente.

beta	nsus	runs	auc_mi	auc_01_mi	auc_di	auc_01_di
0.5	2	1000	0.711880	0.520393	0.562470	0.517275
0.5	2	5000	0.732653	0.553852	0.668021	0.550492
0.5	2	10000	0.749799	0.571115	0.741602	0.591429
0.5	2	20000	0.793999	0.618839	0.778668	0.642821

Tabla 5.1.1. Extracto del "Resultados Optimización de 2TRX". Esta Información referida a la evolución de la proteína 2trx con β=0.5 y nsus=2 para diferentes runs. Se observa como el AUC para los dos métodos aumenta a medida que aumenta el número de ejecuciones independientes.

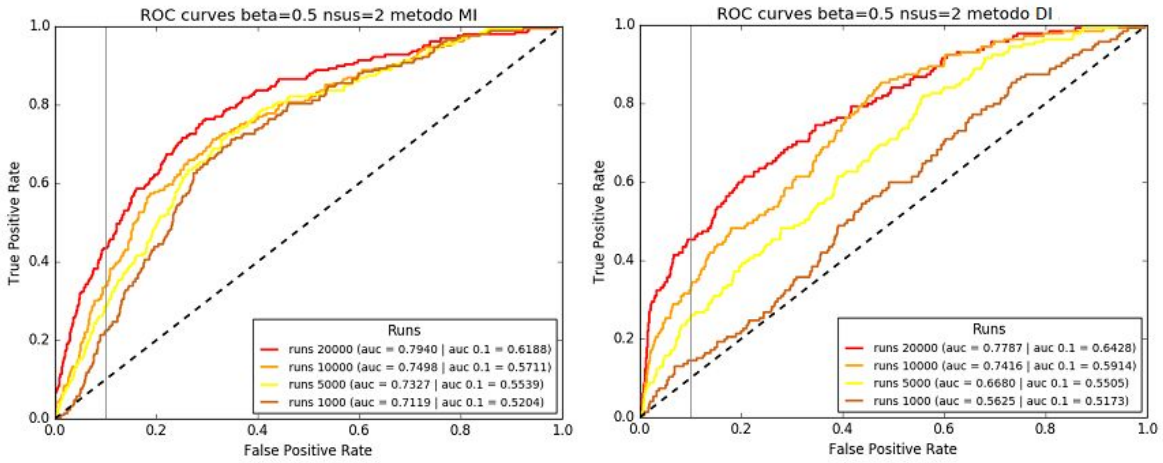


Figura 5.1.2. Valores de AUC y Curvas ROC con β=0.5 y nsus=2 para diferentes runs independientes. La figura izquierda el metodo MI y sobre la derecha el DI. A mayor cantidad de runs independientes aumenta el valor de AUC.

Los resultados completos del análisis exhaustivo realizado dan muestra de las diferentes combinaciones de parámetros que dan valores altos de AUC, gráficamente visualizados en las Figuras 5.1.3 a, b, c y d para los métodos de

coevolución MI, DI, FROB y PSICOV respectivamente. En dichas figuras se muestran las diferentes combinaciones: valores de β entre 0.1 y 20; y de nsus entre 1 y 20; dejando fijo la cantidad de runs en 20000; no se realizarán cómputos con valores más altos de runs por una cuestión de desempeño en el tiempo de ejecución, valores más altos fueron probados e implican un tiempo de ejecución mayor y una muy leve variación en los resultados.

Analizando las figuras, se destacan los diferentes valores óptimos que resultan para cada uno de los métodos de coevolución; MI tiene sus valores de máximo desempeño predictivo con los valores de β = 5-7 y nsus = 10-20; en cambio en el metodo DI están dados por un β mas chico de 0.5 con nsus = 5. FROB observa positivos con β = 0.5-5; todos ellos rondando entre 0.8 y 0.9; los resultados del método PSICOV se destacan por sobre los demás ya que a simple vista sus óptimos superan el valor de 0.9.

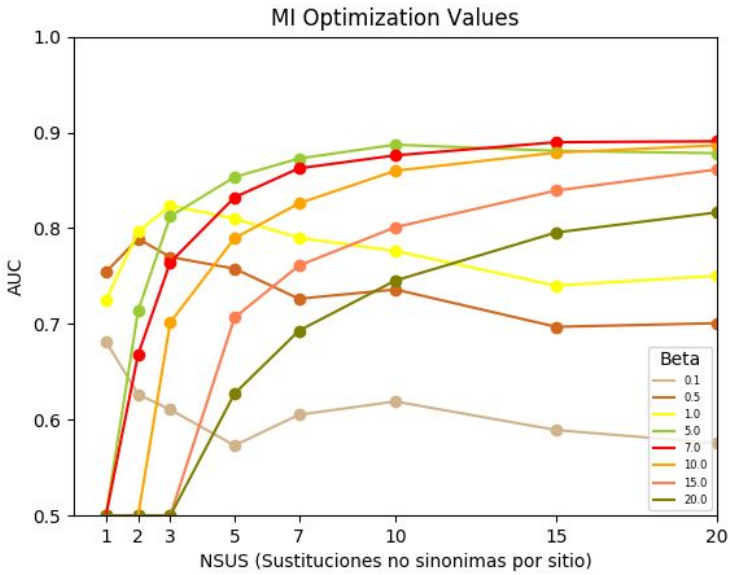


Figura 5.1.3.a Valores de AUC en base a los diferentes β y diferentes nsus (eje x); con 20000 runs independientes. Sobre el eje y el valor de AUC. Las figura corresponde al método de coevolución MI.

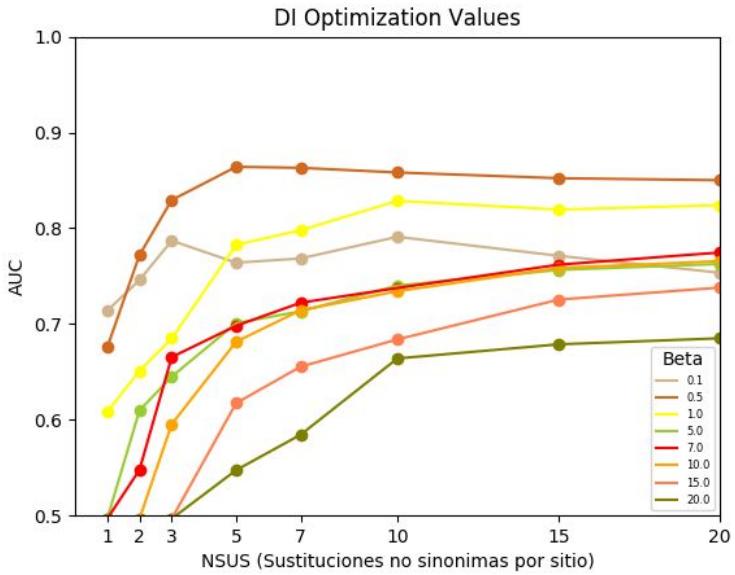


Figura 5.1.3.b Valores de AUC en base a los diferentes β y diferentes nsus (eje x); con 20000 runs independientes. Sobre el eje y el valor de AUC. Las figura corresponde al método de coevolución DI.

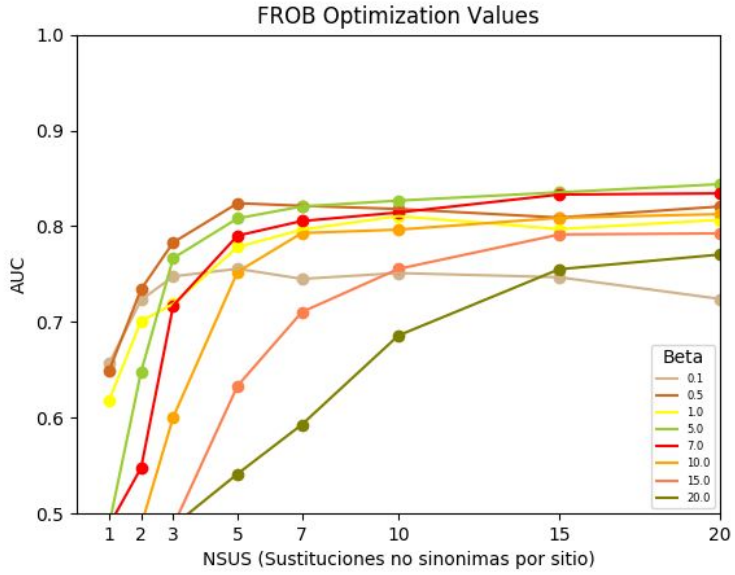


Figura 5.1.3.c Valores de AUC en base a los diferentes β y diferentes nsus (eje x); con 20000 runs independientes. Sobre el eje y el valor de AUC. Las figura corresponde al método de coevolución FROB.

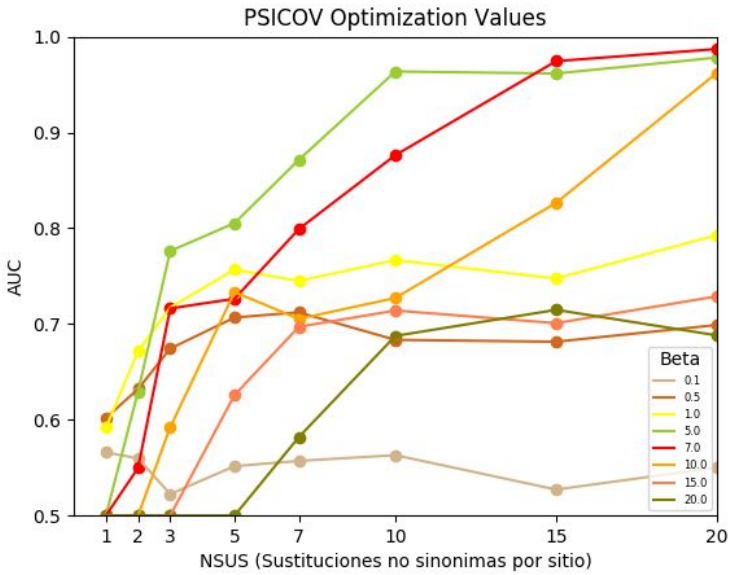


Figura 5.1.3.d Valores de AUC en base a los diferentes β y diferentes nsus (eje x); con 20000 runs independientes. Sobre el eje y el valor de AUC. Las figura corresponde al método de coevolución PSICOV.

El parámetro nsus, indica la cantidad en promedio de sustituciones no sinónimas por sitio, al aumentarlo, al igual que los runs, implican una divergencia y población en los MSAs, pero de igual forma un aumento desmedido genera altos costos en el tiempo de ejecución y no brinda resultados significativos más óptimos en cuanto a AUCs, por tal motivo se decidió que el límite fuera de 20 nsus.

Al contrario a lo que sucede con los valores pequeños de β, cuando incrementamos la presión estructural, β=15-20, el procedimiento es muy restringido, solo acepta muy pocas mutaciones en la secuencia, lo que genera una mínima divergencia dando como resultados MSAs con menor población. Pueden observarse los resultados, por ejemplo de β=20 con nsus=1-3, en donde se obtienen valores de 0.5 por tener poca información dentro de los MSAs. Este análisis es idéntico para todos los métodos de coevolución; con este tipo parametrización la evolución de la proteína no genera MSAs para un posible análisis porque la divergencia estructural es mínima y los MSAs generados carecen de una población que diverge de la secuencia de referencia.

Por último analizaremos con más detalle la información que nos brindan las curvas ROCs; para ello se estudiará a modo de ejemplo los resultados del método MI que se encuentran en la Tabla 5.1.2. Podría decirse que todos los valores que se visualizan en la misma tienen un desempeño predictivo muy satisfactorio, y que no existen diferencias entre ellos. Pero esto es así en realidad? Observemos que el valor de AUC en todos estas comparaciones es muy similar, oscila entre 0.88 y 0.89, lo cual no es una diferencia sustancial; pero en cambio existe una medida que será muy importante analizar, el AUC parcial 0.1, al mismo le daremos la notación AUC_01. El AUC_01 nos brinda información sobre cómo es nuestro predictor en los primeros valores de la curva ROC (es decir en la predicción de los primeros contactos). Aquí se pueden apreciar notorias diferencias en cuanto a su valor, por ejemplo, mientras que para β=5, nsus=15 y runs=20000 el valor de AUC_01 es 0.752202; para β=5, nsus=10 y runs=20000 es de 0.715675; y para β=10, nsus=15 y runs=20000 es de 0.658931. Para el método MI esto puede observarse en la Figura 5.1.4, en donde se puede distinguir claramente que en los primeros valores de la curva ROC, el evolucionado con β=5, nsus=15 y runs=20000 (curva roja) es superior a los demás. Por lo tanto, la evolución con los parametros β=5, nsus=15 y runs=20000 es el mejor resultado de las tres evoluciones y predice con más exactitud. Podemos concluir que nuestra función objetivo para realizar la optimización debe estar regida, no sólo por el AUC, sino también debe incorporarse al análisis el AUC parcial 0.1.

beta	nsus	runs	auc	auc_01
5	15	20000	0.899597	0.752202
5	10	20000	0.897741	0.715675
10	15	20000	0.882895	0.658931

Tabla 5.1.2. Resultados de desempeño predictivo con MI para MSAs evolucionados artificialmente. Ejemplo para analizar el rol del AUC_01.

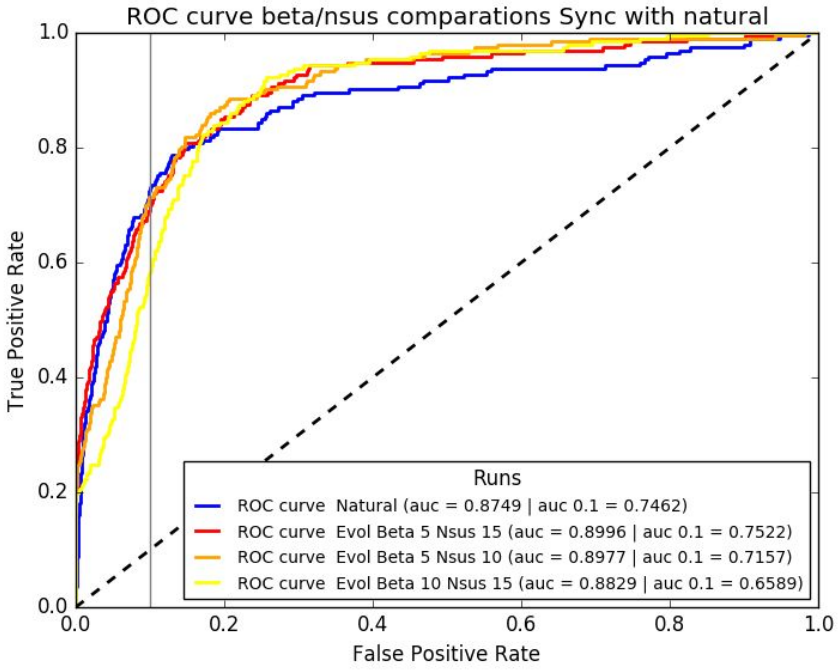


Figura 5.1.4. Curvas ROC con altos valores de AUC para el método MI. La curva en azul representa a la información obtenida del MSA natural, las demás curvas son de MSAs evolucionados con diferentes valores de β y nsus; todos los evolucionados son con runs=20000.

En esta sección obtuvimos los resultados con los mejores AUC y se analizó la relación de los mismos en base a las variables del SCPE; al obtener varias combinaciones con valores >=0.8 se realizarán comparaciones junto al MSA natural en la sección 5.5 para indicar cual es el resultado óptimo para cada uno de los métodos de coevolución.

5.2. Verificación del MSA evolucionado: HHPRED

Para realizar la verificación constatando que las secuencias que se generaron dentro de los MSAs evolucionados, continúan manteniendo las restricciones estructurales de la estructura de referencia, someteremos estas secuencias mutadas en distinto grado a un predictor de estructura. Se puede

utilizar el servidor HHPred para la predicción de estructuras en proteínas en base a una secuencia [10].

HHPred recibe como parámetro una secuencia y realiza la búsqueda en varias bases de datos, incluyendo la de pdb que es de nuestro interés; para ello realiza comparaciones utilizando un modelo oculto de Markov (HMM, del inglés *Hidden Markov Model*) y retorna una predicción estructural con un *score*.

La lista de aciertos (*hits*) incluye e-values y una probabilidad de ser verdaderos. Los alineamientos contienen anotaciones sobre las estructuras secundarias, secuencias consenso y confiabilidad en cada una de las posiciones (*position-specific reliability*). Además incluye cuáles son las estructuras de PDB que se encuentran relacionadas.

A modo de ejemplo se tomaron dos secuencias que se encuentran dentro de uno de los MSAs evolucionados con alto AUC, Figura 5.2.1; la primer secuencia es la secuencia de referencia 2TRX (sin mutaciones), la segunda es una secuencia evolucionada con 61 % de identidad con la secuencia de referencia y la tercera secuencia tiene un 29 % de identidad con respecto a la misma. El objetivo es validar que el proceso de evolucion *in silico* retorna como resultado una secuencia que tomará el plegamiento de la 2TRX.

Figura 5.2.1. En primer lugar la secuencia de referencia, 2TRX. En segundo y tercer lugar, secuencias evolucionadas con 61 y 29 % de identidad a la secuencia de referencia; utilizadas para realizar la validación con hhpred.

Los resultados son exitosos y pueden observarse en las Figuras 5.2.2 y 5.2.3 como las secuencias toman el plegamiento *trx-like* y dentro de los pdbs que otorga como *hits* se encuentra la estructura 2TRX_A.

La primera búsqueda, realizada con la secuencia que tiene 61% de identidad con la 2TRX, retorna el *hit* con un *score* de 95.75 y una probabilidad de 99.83% de ser verdadero. La segunda búsqueda fue realizada con una secuencia que tenía una identidad inferior, 29%; y el resultado fue positivo, con un *score* de 92.35 y una probabilidad de 99.8% de ser verdadero.

Esta validación, confirma que las secuencias incluidas dentro del MSA obtenido por evaluación *in silico* con SCPE, respetan la estructura salvaje; inclusive cuando el porcentaje de identidad es bajo, agregando asi diversidad secuencial al MSA y respetando la estructura.

Figura 5.2.2. Resultado de búsqueda de la secuencia evolucionada secuence_61_id con 61 % de identidad con respecto a la secuencia de referencia; se encuentra dentro del mismo, como pdb relacionado, la estructura 2TRX_A.

Figura 5.2.3. Resultado de búsqueda de la secuencia evolucionada secuence_29_id con 29 % de identidad con respecto a la secuencia de referencia; se encuentra dentro del mismo, como pdb relacionado, la estructura 2TRX_A.

5.3. Comparación entre el MSA Natural y los MSAs evolucionados

En esta sección continuaremos con el análisis de los valores obtenidos en el capítulo 5.1, enfocándonos en las combinaciones que obtuvieron un AUC >= 0.8, lo que puede considerarse un desempeño muy satisfactorio de los métodos predictivos utilizados.

Previamente, a los MSAs evolucionados se les realiza una sincronización para que puedan compararse con los resultados del MSA natural, obtenidos en la sección 4.2. Los residuos de la cadena A de la estructura 2TRX van desde la posición 1 a la 108 de la proteína THIO_ECOLI; la proteína THIO_ECOLI que se encuentra dentro de la familia de Pfam PF00085 se encuentra mapeada a la estructura 2TRX desde la posición 3 a la 106. Por tal motivo, se realiza un recorte en los MSAs evolucionados eliminando las columnas sobrantes al principio y al

final, para que tanto el MSA natural como el evolucionado tengan la misma longitud, se respeten las posiciones de las columnas y puedan compararse. La misma operatoria se realiza con la matriz de contacto, eliminando las filas y las columnas sobrantes.

Una vez realizada la sincronización correspondiente, se vuelven a realizar los cálculos de cada método de coevolución con el objetivo de detectar cual es el óptimo para cada uno ellos, es decir cuales son los parámetros de SCPE que generan un MSA evolucionado sobre el cual se encuentra el valor más alto de AUC_01. La información puede apreciarse en la Tabla 5.3.1 y su Figura 5.3.1 correspondiente.

Al igual que se había observado en la Figura 5.1.3, se destaca por sobre el resto los resultados obtenidos con el método PSICOV, el cual tiene una predicción casi perfecta con un AUC de 0.9875 y un AUC_01 de 0.9515. En segundo lugar MI, observa una mejor predicción que DI y FROB, sobre todo en los primeros valores de la curva con un valor 0.77 contra un 0.72 de DI y un bajo 0.67 de FROB.

method	auc_01	auc	beta	nsus	runs
mi	0.7746860315	0.8957238905	5	20	20000
di	0.7237858261	0.8597518576	0.5	5	20000
frob	0.6748343426	0.837700324	5	20	20000
psicov	0.951470251	0.9875180328	7	20	20000

Tabla 5.3.1. Información de los parámetros de SCPE que generan los valores óptimos de AUC para cada uno de los métodos de coevolución.

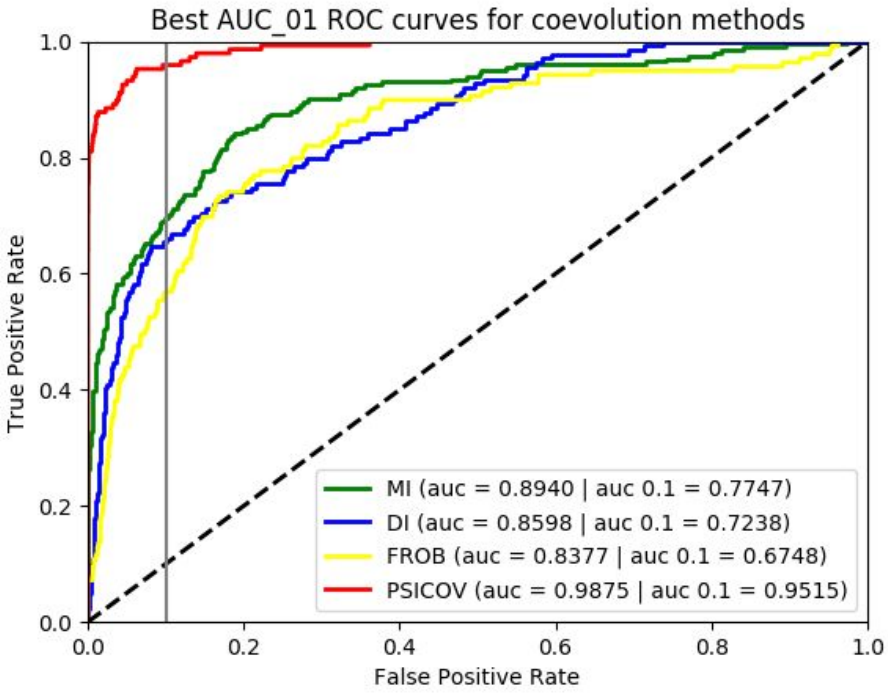


Figura 5.3.1. Curvas ROC con los óptimos de cada uno de los métodos de coevolución.

A continuación analizaremos el comportamiento de cada uno de los métodos de coevolución, comparando el MSA natural con el MSA evolucionado (con los parámetros óptimos para cada uno de ellos). Esto puede observarse en las Figuras 5.3.2 a, b, c y d correspondientes a los métodos MI, DI, Frob y PSICOV respectivamente. En las mismas pueden observarse diferentes resultados, mientras que con MI se obtienen mejores valores con el MSA evolucionado por sobre el natural; con el metodo DI sucede lo mismo pero solamente con el resultado general de AUC, en cambio con los primeros resultados de la curva AUC_01 puede observarse que el natural obtiene un mejor resultado que el evolucionado. En el caso del método de convolución FORB el natural supera al evolucionado claramente. Por último, el método PSICOV obtiene un resultado en el evolucionado casi perfecto, superando holgadamente al natural.

Para profundizar el análisis podemos comparar cuales son los pares de posiciones con alto *score* para cada uno de los métodos de covariación para el MSA evolucionado y para el MSA natural, indicando, además, cuales son contactos y cuáles no. Esta información la podemos observar en las Figuras 5.3.3 a, b, c y d para los métodos de covariación MI, DI, Frob y PSICOV respectivamente, las mismas describen la información obtenida de la evolución óptima de cada uno; cada punto en el gráfico es un par de posiciones del alineamiento; en azul posiciones que son contacto y en rojo posiciones que no son contactos. Sobre el eje x el valor del *score* del método de variación correspondiente para el MSA generado artificialmente; y en el eje y la información del *score* con respecto al MSA natural.

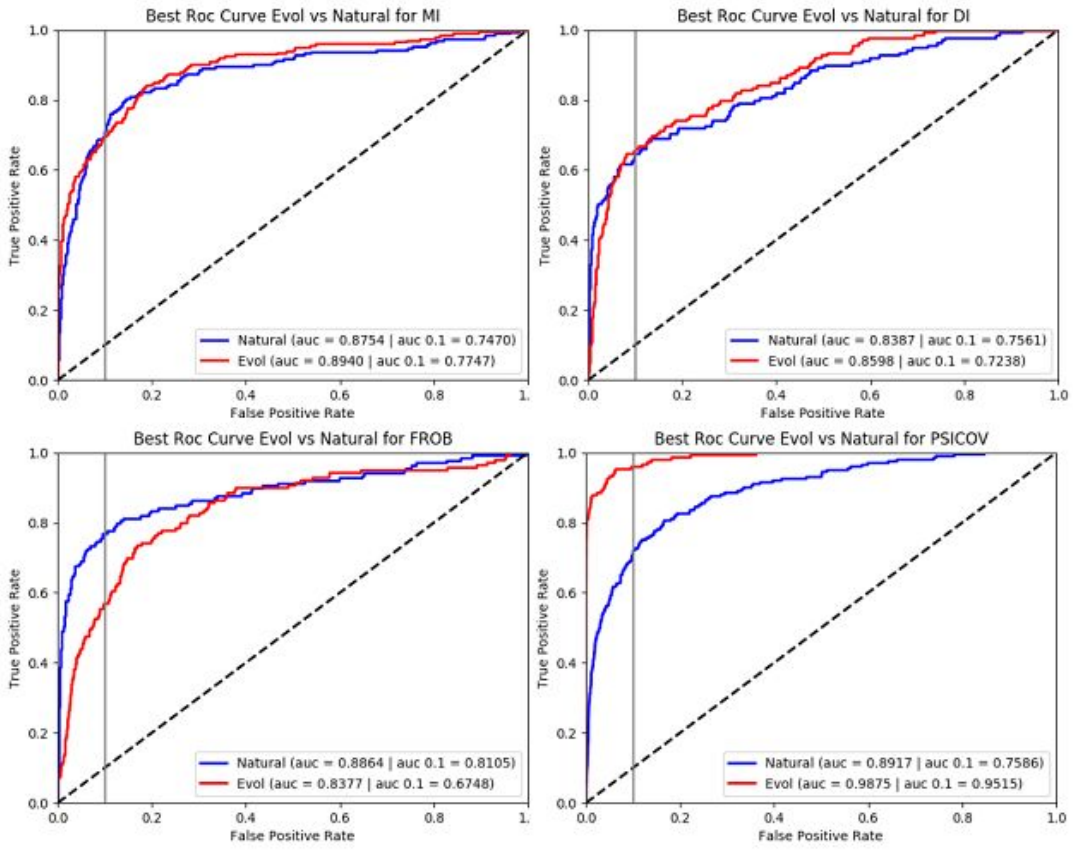


Figura 5.3.2. Curvas Rocs de los métodos de coevolución MI, DI, Frob y PSICOV del evolucionado (rojo) comparados con sus respectivos valores naturales (azul).

El primer gráfico de la Figura 5.3.3 nos indica que los pares de posiciones que tienen alto valor de MI en el evolucionado son contactos, mientras que para el natural existen valores con alto MI que no lo son. Pero no puede visualizarse una distribución que indique claramente que todos los valores con alto MI son contactos; a su vez se aprecian valores con alto MI del natural que son contactos y que no tienen un alto MI en el evolucionado. En el segundo y tercer gráfico referido al método DI y Frob, la distribución es más dispersa, existen pares con alto *score* en el evolucionado que no son contactos. Por último, y continuando con los resultados anteriores, el método PSICOV observa una distribución en donde se percibe que los pares de posiciones que tienen alto *score* son contactos, e inclusive que existen más pares de posiciones que tienen un alto *score* en el natural y también lo tienen en el evolucionado.

Para los métodos de covariación analizados, excepto para FROB, los resultados de AUC y AUC_01 tienen una alta similitud entre el evolucionado óptimo y el natural, inclusive en muchos casos el evolucionado supera al natural; pero es evidente, en base a las Figuras 5.3.3, que existen diferencias respecto al natural. A que se deben esas diferencias? Cómo podemos detectarlas? Cuáles son los pares de posiciones que comparten un alto *score* tanto en el natural como en el evolucionado para cada uno de los métodos? Para ello definimos el concepto de top n % de covariación. En primer lugar se ordena de mayor a menor los pares de posiciones según el valor de covariación, luego el top n % son los primeros n*total_pp/100, donde total_pp es la cantidad total de pares de posiciones. Obtenemos así los top n % de MI, DI, Frob y PSICOV.

La información referida a los diferentes tops de covariación con parámetros óptimos puede observarse en la Tabla 5.3.2. En las misma, se describen el porcentaje de pares de posiciones que son contactos en el evolucionado y en el natural para los diferentes tops analizados, columnas nat_ y evol_ seguidas por el nombre del método de covariación. Además, se encuentra presente la columna

match_ la cual indica el porcentaje de pares de contactos que tienen en común los MSA evolucionado y el natural para cada uno de los métodos.

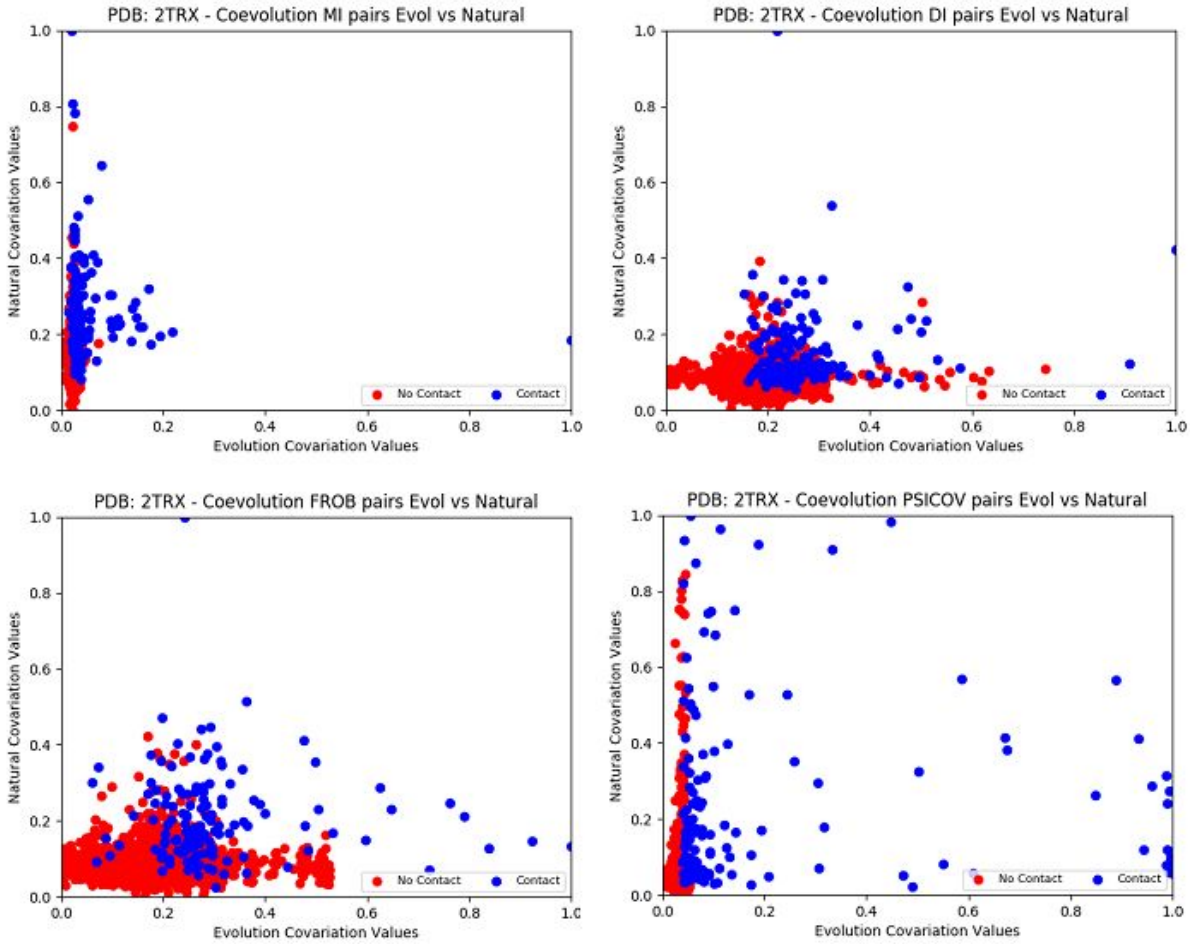


Figura 5.3.3. Gráficos de puntos obtenido de la evolución óptima de los métodos de coevolución MI, DI, Frob y PSICOV. Cada punto es un par de posiciones; sobre el eje x el valor de covariación del evolucionado que tiene el par; sobre el eje y se encuentra el valor de MI del natural. Además para cada par de posiciones se indica en azul los que son contactos y en rojos los que no lo son.

El mejor resultado es el obtenido con el método PSICOV, en donde predice de forma perfecta para el top 0.5 y 1 %, y casi perfecta (99%) para el top 2 %. Supera a los porcentajes del natural holgadamente, y es el método que más pares de posiciones con alto *score* comparte con el natural.

En los resultados de MI, la predicción de contactos en el evolucionado es mejor que en el natural; dando como resultado, por ejemplo, que en el top 0.5, 1 y 2, el evolucionado obtuvo 96, 92 y 69 % de contactos, contra un 72, 62 y 54 % del natural. Tanto en el natural como en el MSA evolucionado, a medida que el top se incrementa, se van incorporando falsos positivos, pero el evolucionado continua por delante en la predicción. La cantidad de pares de posiciones que son contactos y que son compartidos entre el natural y el evolucionado con MI para los diferentes tops es muy baja. Esto demuestra que los pares de posiciones que tienen un alto MI son diferentes entre los MSAs.

Por su parte los métodos DI y FROB están muy por debajo de los obtenidos con PSICOV y MI; inclusive en todos los tops el porcentaje de contactos encontrados con el natural supera al evolucionado.

top %	nat_MI %	evol_MI%	match_MI %	nat_DI %	evol_DI %	match_DI %	nat_FROB %	evol_FROB %	match_FROB %	nat_PSICOV %	evol_PSICOV %	match_PSICOV%
0.5	72	96	4	66.66	29.16	8.33	70.83	41.66	0	65.38	100	7.69
1	62	92	17	71	30	14	73	26	4	63	100	23
2	54	69	18	58	34	13	64	21	10	60	99	34
3	44	59	19	43	33	13	50	22	14	50	79	31
4	39	48	20	35	29	14	41	24	15	45	61	30
5	34	42	19	29	26	13	36	22	16	39	50	25

Tabla 5.3.2. Descripción de la información de los tops para los métodos de covariación. La columnas nat_MI y evol_MI describen el porcentaje de los pares que son contactos para el MSA natural y el evolucionado con el método MI, respectivamente; la columna match_MI indica el % de pares de posiciones, que son contactos, y que se encuentran en coincidencia en el MI natural y en el MI evolucionado. De forma idéntica se describe la información para los métodos de covariación DI, FROB y PSICOV.

Para tener un acercamiento visual del comportamiento de los diferentes tops, podemos observar las Figuras 5.3.4 a, b, c y d que describe sobre la matriz

de contactos el top 1% para los métodos de MI, DI, Frob y PSICOV respectivamente. En las mismas se pueden apreciar las matrices de contacto simétricas, en donde los puntos en azul que se encuentran en la diagonal inferior son los pares de posiciones con alto score del MSA natural; por encima, en la diagonal superior y de color rojo, los pares de posiciones con alto score del evolucionado. Los métodos DI y FROB observan varios falsos positivos, como se apreciaba anteriormente en el bajo porcentaje de contactos en la Tabla 5.3.2. Inclusive pueden observarse señales verticales y horizontales con alto score en estos métodos, estas zonas indican posiciones altamente conservadas. En cambio, los métodos de MI y PSICOV tienen una predicción de los contactos más eficiente, teniendo en cuenta el problema de la conservación de las columnas; al observar las figuras notamos que los puntos con alto score están sobre las zonas marcadas como contactos.

Como se ha descrito en la Tabla 5.3.2 a medida que avanza el top analizado la cantidad de falsos positivos es mayor, sobre todo en los métodos de DI y FROB; no es el caso del método PSICOV el cual inclusive en los tops 2 y 3 % tiene un alto porcentaje de aciertos de contactos (99 y 79 %), visualmente lo podemos ver en las Figuras 5.3.5 a y b.

Podemos concluir que el método de PSICOV es superior a los demás y brinda valores muy similares a la información natural. Los falsos positivos del MSA evolucionado, pueden desprenderse de posiciones que bajo la circunstancia del tipo de evolución, terminen siendo conservadas y se vislumbre una falsa coevolución entre las mismas, los métodos DI y FROB vieron afectada su señal de coevolución por esta situación. Además, debemos recordar que durante el procedimiento se definió, como vimos en la sección 3.2, una matriz de contactos teniendo en cuenta los radios de Van der Waals: existe contacto entre dos residuos si la distancia de cualquiera de sus átomos menos el radio de Van der Waals de los mismos es menor a 1; sumando la restricción de que solamente se tiene en cuenta la cadena lateral R de la proteína. Trabajos a futuro pueden

involucrar evaluar diferentes matrices de contactos, en base a otras restricciones, por ejemplo, que cualquier par de posiciones que se encuentren a una distancia menor que 6.05 Å (medida utilizada normalmente), estén en contacto, de esta forma podríamos analizar las diferencias e inclusive mejorar el procedimiento evolutivo del SCPE.

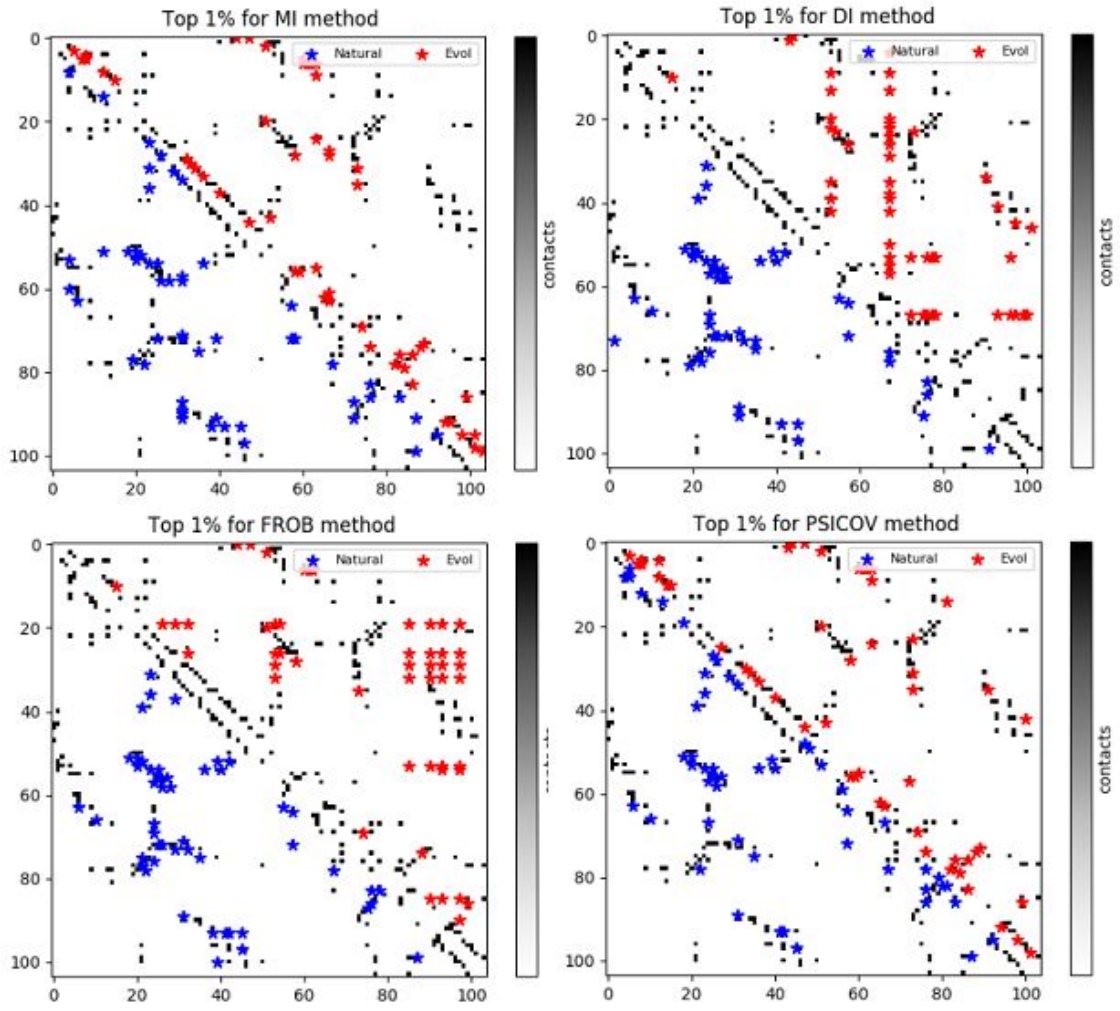


Figura 5.3.4. Matrices de contacto simétricas. Los contactos reales son los puntos sombreados negros. La información describe el top 1% de MI, DI, FROB y PSICOV respectivamente; los puntos en azul que se encuentran en la diagonal de abajo son los pares de posiciones con alto score del MSA natural; por encima en la diagonal superior y de color rojo, los pares de posiciones con alto score del evolucionado.

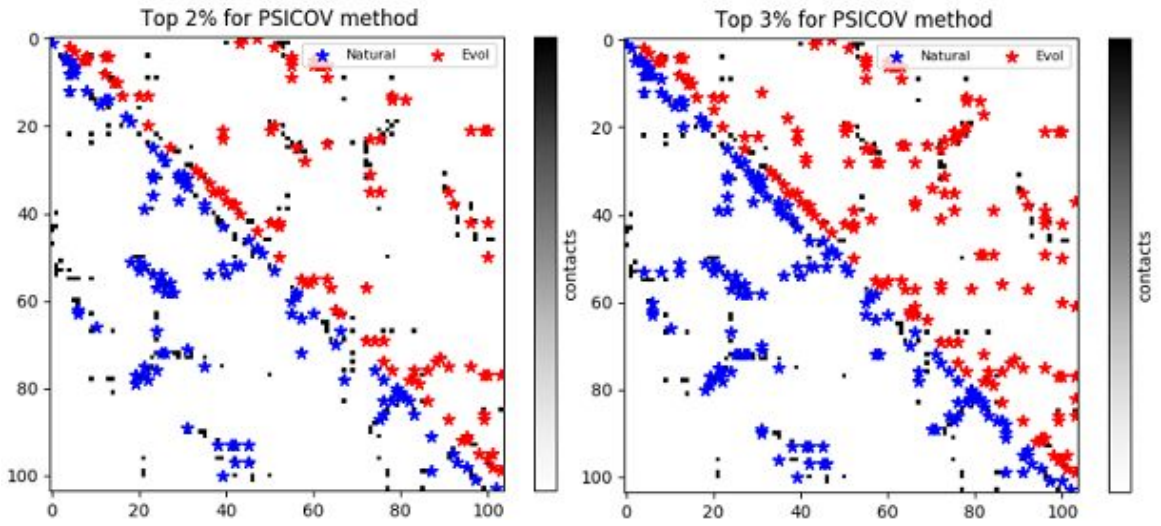


Figura 5.3.5. Matrices de contacto simétricas. Los contactos reales son los puntos sombreados negros. La información describe el top 2 y 3 % del método PSICOV; los puntos en azul que se encuentran en la diagonal de abajo son los pares de posiciones con alto score del MSA natural; por encima en la diagonal superior y de color rojo, los pares de posiciones con alto score del evolucionado.

La razón de que se encuentren pocos contactos compartidos, como también la superioridad del evolucionado con el natural en el porcentaje de contactos encontrados puede deberse a que la evolución in silico solo tiene en cuenta las restricciones estructurales de la estructura 2TRX y, en cambio, la evolución natural contiene, además, otro tipo de señales que resultan en un alto score de coevolución dentro de su alineamiento; la información filogenética incorporada, la información codificada involucrada en la interacción con otras proteínas y otros sitios activos en donde exista una coevolución entre pares de posiciones que tengan que ver única y exclusivamente con lo funcional y no tenga restricciones estructurales, son algunos de los casos que pueden describir este resultado. Inclusive puede deberse a información estructural que se encuentra codificada pero que pertenece a otras proteínas de la familia o a otros confórmeros de la misma proteína, ya que en el análisis se toma como matriz de contacto teniendo en cuenta solamente la estructura del confórmero 2TRX. Sobre esto último, veremos en los puntos 6.2 y 6.3, que las matrices de contactos de los diferentes confórmeros de la proteína son diferentes.

5.4. Conservación: Comparación entre el MSA evolucionado y el natural

El análisis de la conservación debe realizarse sobre lo que denominamos los MSA óptimos para cada método de covariación. Es decir, los MSAs que al analizar su coevolución dieron sus mejores resultados para cada uno de los métodos utilizados. La Figura 5.4.1 muestra la conservación de los diferentes MSA óptimos para cada método, junto con la conservación del MSA natural.

Una de las primeras observaciones que podemos realizar es que los MSAs evolucionados contienen mayor número de posiciones conservadas que el natural. Esto se debe a la naturaleza de la simulación estructural que estamos realizando, por lo cual se mantienen posiciones conservadas que tienen una lógica estructural para la simulación. Es evidente que en el MSA natural, no se necesitan fijar tantas posiciones a nivel estructural para su funcionamiento, y si en cambio tiene otras posiciones conservadas que cumplen un rol funcional, por ejemplo, catalítico.

Otra observación es que los MSAs generados con los valores óptimos de SCPE para los distintos métodos tienden a conservar las mismas posiciones. Lo que hace pensar que en un rango grande de parámetros de nsus y beta las posiciones que terminan siendo conservadas son las mismas.

Para realizar un análisis en profundidad y conocer los aminoácidos que aparecen en las posiciones conservadas podemos observar la Figura 5.4.2 que muestra el logo de secuencias de la familia PF00085 a la que pertenece la proteína THIO_ECOLI luego de realizar el proceso de gapstrip descrito en la sección 4.1. Se puede apreciar el sitio activo de la proteína, en base a la información obtenida a través de UniProt (Entrada de UniProt: P0AA25). Las Cysteinas (C) de las posiciones 30 y 33 se encuentran conservadas, al igual que la Glicina (G) y Prolina (P), en menor medida, que están en las posiciones 31 y 32

respectivamente. Puede notarse también la conservación en el Ac. Aspártico (D) de la posición 24. Otros aminoácidos importantes son las Fenilalaninas (F) 10 y 25, las P 38 y 74 y las G 82 y 90. La numeración está dada por el número en el pdb.

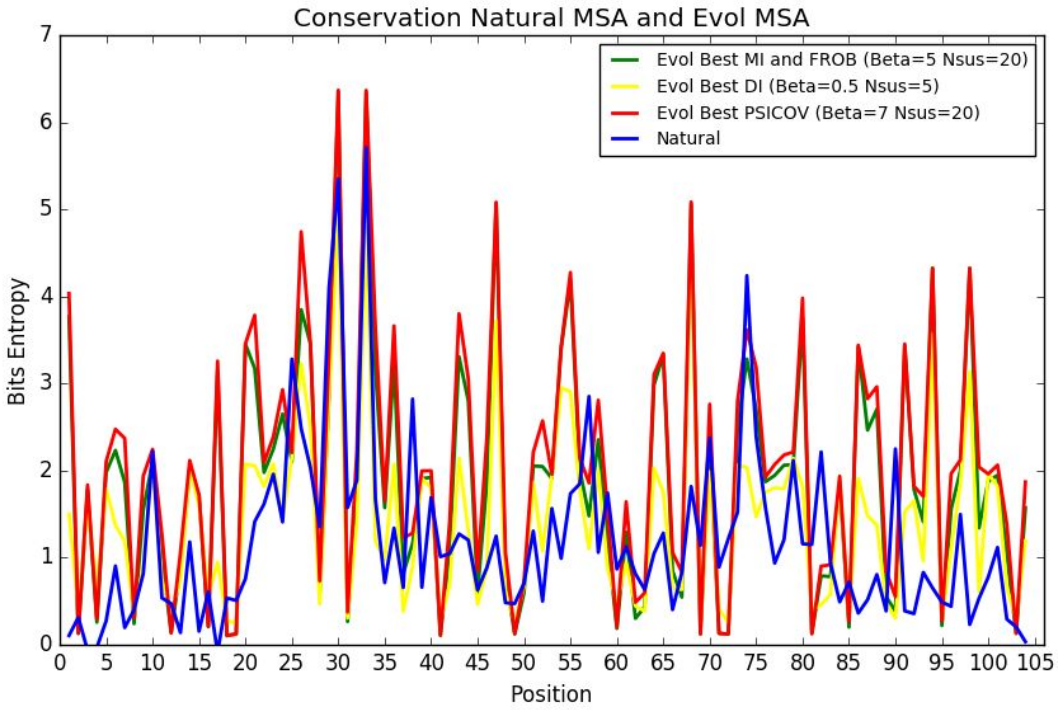


Figura 5.4.1. Conservación de los MSAs óptimos para cada metodos de covariación, junto con la conservación del MSA natural.

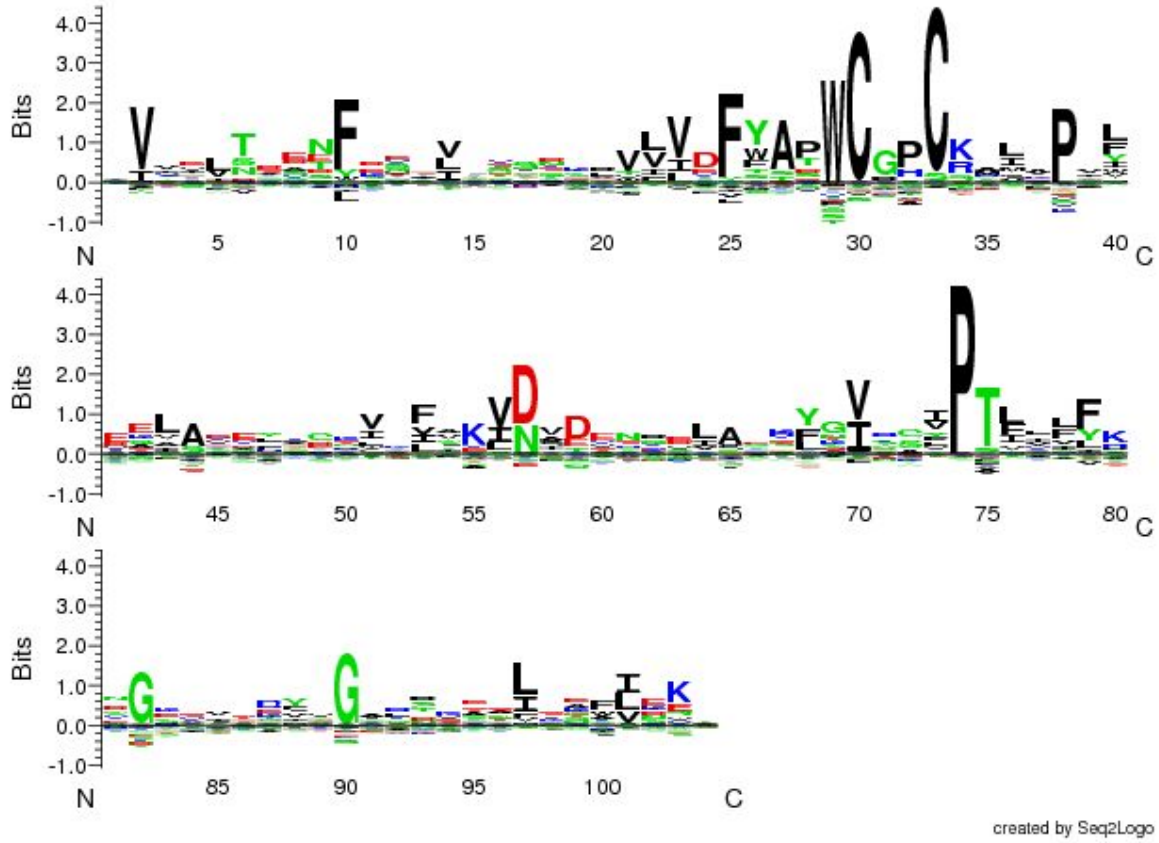


Figura 5.4.2. Logo de secuencias, calculado con Seq2Logo, de la familia natural PF00085 luego de realizar el proceso de gapstrip con la proteína THIO_ECOLI puesta de referencia, utilizando el algoritmo Kullbak-Leibler. Se visualiza de forma descriptiva los sitios conservados y sus aminoácidos correspondientes.

Se visualiza a continuación en la Figura 5.4.3 el logo de secuencias del MSA evolucionado óptimo de los métodos MI y FROB, a modo de ejemplo ya que no existen diferencias con el MSA evolucionado óptimo de método PSICOV y DI. Vemos que se conservan los principales residuos del sitio activo de la proteína, las C de las posiciones 30 y 33, al igual que la Prolina (P) de la posición 32. En cambio, la G de la posición 31 no se encuentra conservada, es una diferencia que existe con el alineamiento natural que debemos analizar. Las posiciones del sitio activo que se conservan, además de ser importantes para la catálisis, probablemente también tengan restricciones estructurales, esto explicaría el porqué la conservación en el MSA evolucionado. Con las Glicinas, no solo la del sitio 31 no fue conservada, también ocurrió lo mismo con las posiciones 82 y 90.

Esto puede deberse a que el algoritmo de evolución presentado no favorece su conservación al ser el aminoácido más pequeño. Con las F presentes en el natural en las posiciones 10, 25 y 79 sucedió lo mismo.

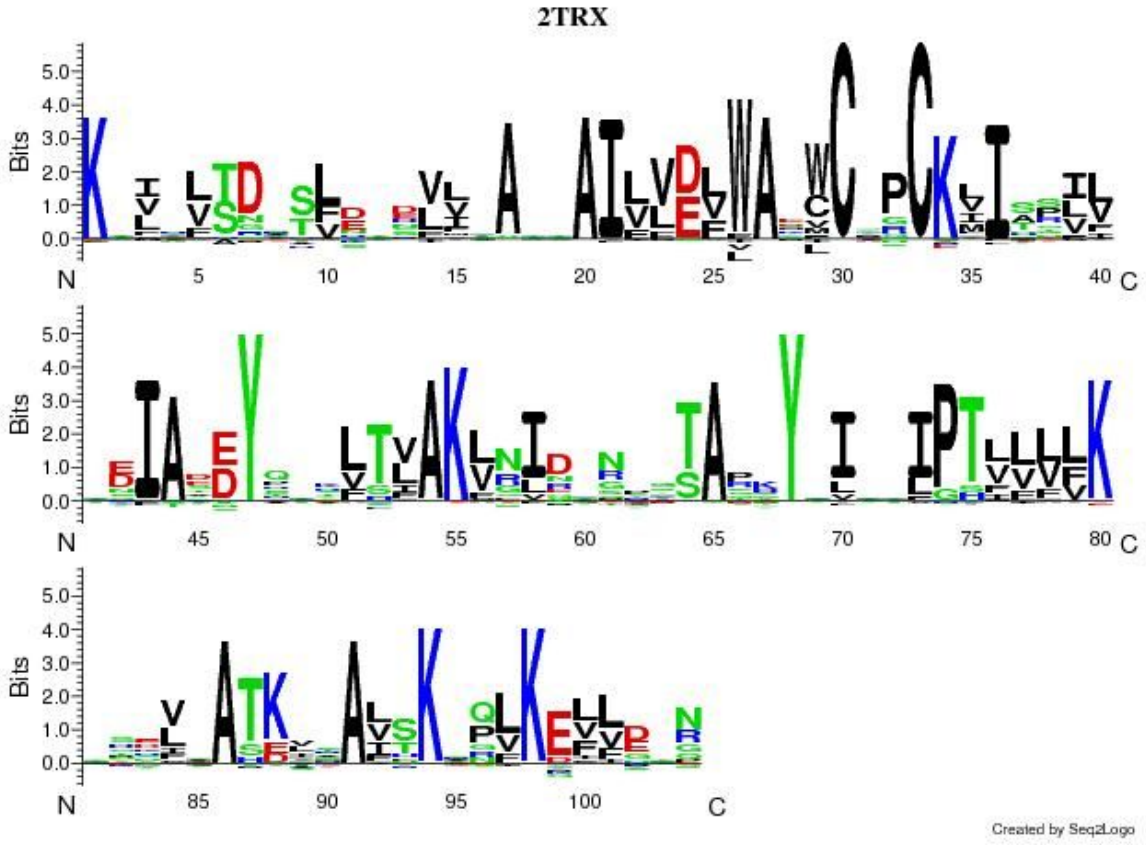


Figura 5.4.3. Logo de secuencia, con Seq2Logo, del MSA evolucionado, utilizando el algoritmo Kullbak-Leibler.

En el MSA evolucionado puede notarse la aparición de otros sitios conservados, y llamativamente son muchos más. Es criterioso pensar que al buscar conservar sólo la estructura, el SCPE admite menos cambios en posiciones claves, esto tiene que ver con la naturaleza del algoritmo usado para hacer el modelo de evolución, que busca preservar la estructura de menor energía; y que exceden a la realidad en un alineamiento natural. Los casos más claros son la introducción de Alaninas (A), Lisinas (K), Tirosinas (Y) y Treoninas (T).

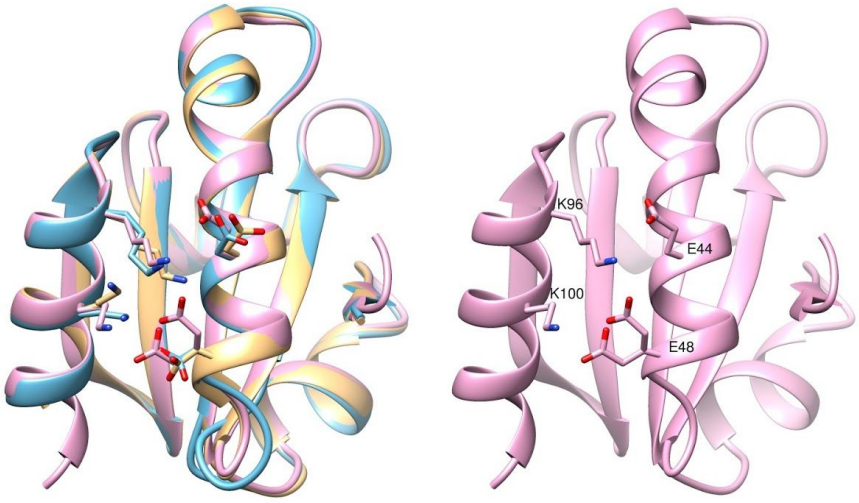
Entonces la pregunta que surge: ***Por qué se conservan ciertos aminoácidos en el MSA evolucionado que no lo están en el MSA natural?***

Para comprobarlo, se realizó un modelo estructural de dos secuencias tomadas del MSA evolucionado de 33 y 61% de identidad con la 2TRX (inicial del SCPE). El modelo fue realizado con Modeller [35] a través de la interface que proporciona Chimera [36]. En la Figura 5.4.4, se muestra el alineamiento utilizado para realizar el modelado por homología. Se puede observar también la conservación en las secuencias de ciertos aminoácidos que están conservados en el logo del MSA evolucionado y no en el del MSA natural.

Figura 5.4.4: Alineamiento utilizado para hacer el modelo por homología.

Observamos que las K tienden a no ser reemplazadas. Esto puede ser debido a que dado que en la secuencia de la proteína inicial para correr el SCPE el residuo en esa posición es K, y en la estructura, generalmente está enfrentada con una D/E, es difícil cambiar este enlace iónico para el SCPE (Figura 5.4.5 a). La numeración en esta sección está basada en la del pdb (+ 2 residuos respecto a la secuencia de referencia). Así las K96 y 100 de esta sección corresponden a las K94 y 98 en la secuencia. Idem para todos los residuos nombrados en esta sección.

A)



B)

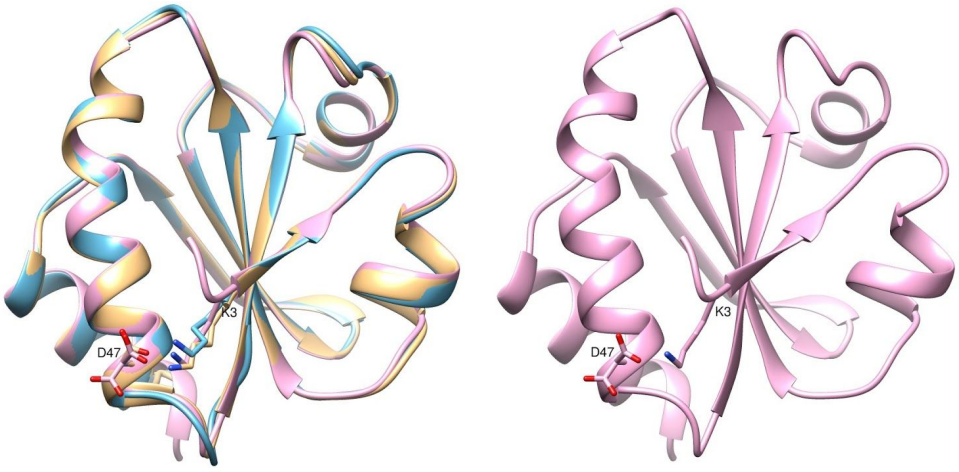


Figura 5.4.5: representación de cinta. Rosa: 2TRX; celeste: modelo seq_33; dorado: seq_61. Panel izquierdo, las tres estructuras; Panel derecho, solo la 2TRX para facilitar la vista. Las estructuras son rotadas según conveniencia para la visualización. A) Los residuos K96 y K100 y sus opuestos E44 y E48 están representados con sticks y coloreados por heteroátomos. B) K3 y D47 están representados con sticks y coloreados por heteroátomos.

En la Figura 5.4.6 la Y70 está rodeada de un ambiente de residuos aromáticos e hidrofóbicos, también, siendo este un entorno muy favorable para la

estabilización de la estructura, por lo tanto difícilmente reemplazable. Con los aminoácidos tipo A, L, I, V, probablemente están formando core hidrofóbicos.

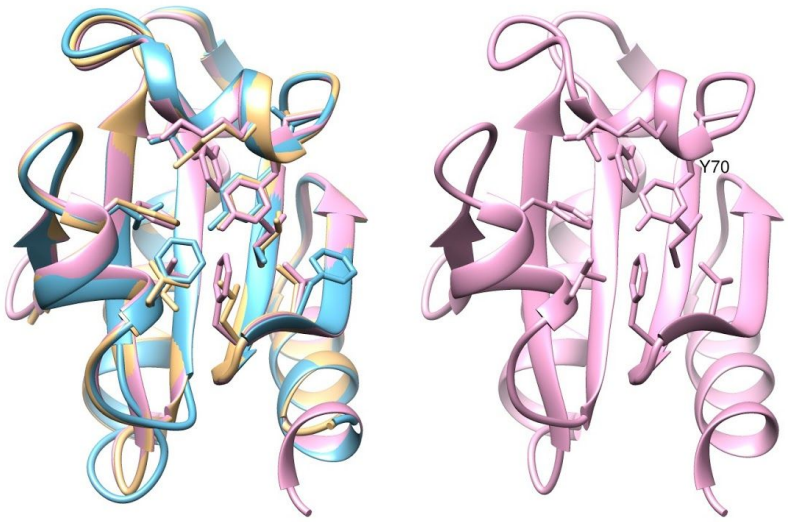


Figura 5.4.6: representación de cinta. Rosa: 2TRX; celeste: modelo seq_33; dorado: seq_61. Panel izquierdo, las tres estructuras; Panel derecho, solo la 2TRX para facilitar la vista. Las estructuras son rotadas según conveniencia para la visualización. La Y70 se encuentra representada.

5.5. Conclusión

En este capítulo hemos encontrado que al simular la evolución de la proteína THIO_ECOLI manteniendo las restricciones estructurales, basándonos en la cadena A de la estructura 2TRX, se generan MSAs, que al analizar su predicción de contactos, alcanzan desempeños predictivos muy altos. Se han utilizado para el análisis cuatro métodos de covariación: MI, DI, FORB y PSICOV; se obtuvo la optimización de los parámetros de SCPE para cada uno de ellos, luego de realizar una ejecución exhaustiva con un abanico amplio de parámetros. Se observó que si la evolución se realizaba con una presión estructural baja (β chico) la predicción de contactos no era tan eficiente; y por el contrario una presión estructural alta (β grande), no generaba suficiente divergencia secuencial y los MSAs quedaban poco poblados, dando luego resultados predictivos menos satisfactorios.

En una observación más precisa, se destacó la injerencia del AUC parcial 0.1 al comparar MSAs evolucionados, que brindan resultados similares en el valor de AUC, pero que al ser analizados en profundidad contenían diferencias en la predicción de los primeros valores de la curva. Concluyendo, que además del AUC, valor de desempeño predictivo usado comúnmente, debía analizarse el AUC parcial 0.1 como función objetivo para detectar el óptimo.

Posteriormente, se analizaron las similitudes y diferencias encontradas entre el MSA evolucionado y el natural. Los valores de AUC obtenidos en la Tabla 5.3.1 y la predicción de contactos observada en la Tabla 5.3.2 y en las Figuras 5.3.4 nos indica que el MSA evolucionado tiene mayor precisión en la predicción de contactos analizando los diferentes tops para los métodos de MI y PSICOV, este último siendo el método con mejor desempeño predictivo, con valores casi perfectos de AUC con 0.9875 y un AUC_01 de 0.9514 en el óptimo encontrado. Es destacable los resultados del método PSICOV a nivel de predicción de contactos, incluyendo el top 2 y 3 %, en donde se continúan obteniendo valores altos de predicción con un 99 y 79 % respectivamente.

Estos resultados son acorde a los esperados por las características del experimento realizado, el evolucionado brinda un mejor resultado por su naturaleza de simular la evolución solamente teniendo en cuenta su estructura; mientras que el natural contiene señales altas de coevolución que no se pueden explicar solamente por restricciones estructurales; señales filogenéticas, coevolución puramente funcional y demás señales que se encuentran codificadas en la evolución de millones de años dentro de la proteína. Además, en este procedimiento solamente se tuvo en cuenta la información evolucionada de la cadena A de la estructura 2TRX; capítulos posteriores harán hincapié en analizar si sumando mas información referida a la evolución *in silico* partiendo de distintas estructuras del MSA, cambian los resultados.

Un enfoque similar pudo observarse a nivel de conservación, el MSA evolucionado mantuvo los residuos del sitio activo conservados, pero otros fueron desfavorecidos por el tipo de evolución, el caso de las glicinas fue uno de los observados. Además, surgieron posiciones conservadas que en el alineamiento natural no se encontraban y que están totalmente relacionadas con las restricciones estructurales. El MSA natural, no necesita fijar tantas posiciones a nivel estructural para su funcionamiento, y si en cambio tiene otras posiciones conservadas que cumplen un rol funcional, por ejemplo, catalítico.

La generación de secuencias que respeten las propiedades estructurales de una proteína pueden ser utilizadas para realizar estudios referidos a sumar divergencia secuencial manteniendo la misma estructura conformacional. Se podrían poblar alineamientos naturales que poseen escasas secuencias, sumándole secuencias que respetan las restricciones estructurales de varias de las proteínas que se encuentren cristalizadas.

6. Simulación: confórmeros de la proteína THIO_ECOLI

En el capítulo anterior se analizó simular la evolución de la proteína THIO_ECOLI basándonos en la cadena A de la estructura 2TRX. En este capítulo nos enfocaremos en la misma proteína pero iniciando el software de simulación de evolución con diferentes confórmeros de la misma. Al tener igual secuencia y ser diferente la estructura inicial, podremos evaluar sólo el efecto de iniciar la evolución con una variante estructural, se evaluarán un total de 8 estructuras.

En un paso preliminar se analizarán los resultados de cada estructura obteniendo las similitudes y diferencias; luego la información será analizada de forma conjunta, realizando la metodología explicada en la sección 6.4, para

conocer si agrupando la información de covariación obtenemos mejores resultados. Para ello se suman cuantas veces los pares de posiciones aparecen en los tops, procedimiento que se realiza para cada uno de los métodos de covariación estudiados.

La última sección del capítulo está orientada a generar un MSA conjunto realizando un bootstrap de diferentes cantidades de secuencias de cada MSA obtenido luego de la simulación de un confórmero; y con ellas generar un MSA conjunto para luego aplicar el procedimiento predictivo y analizar los resultados.

6.1 Selección de los confórmeros

Existen muchas estructuras cristalizadas de la proteína THIO_ECOLI; en este primer paso, se seleccionarán 8 estructuras, incluyendo la estructura de referencia 2TRX, para realizar la evolución de las mismas y analizar sus resultados.

Es preciso seleccionar las estructuras con mayor divergencia conformacional entre ellas, esto es para sumar diferentes contactos y dar cierta diversidad a la evolución. Si no se tiene en cuenta este punto, se podría caer en el error de seleccionar estructuras muy similares, lo que implicaría poca diversidad conformacional durante el análisis. Para realizar dicha selección se utilizará la base de datos CoDNaS [37]. Para cada proteína representada, la base de datos CoDNaS contiene la colección de todas las estructuras cristalizadas para una misma proteína en diferentes condiciones. En consecuencia, pueden considerarse como diferentes instancias del dinamismo de la proteína; estas estructuras son denominadas confórmeros. Como medida de la diversidad conformacional proteica (PCD), se utiliza el RMSD (por su nombre en inglés: root mean square deviation) de los CA (Carbonos Alfa) que surge de la superposición de todas las estructuras depositadas para cada proteína.

El dendograma obtenido con CoDNas de la Figura 6.1.1 muestra todos los confórmeros, para la proteína THIO_ECOLI, describiendo la distancia estructural que existe entre ellos.

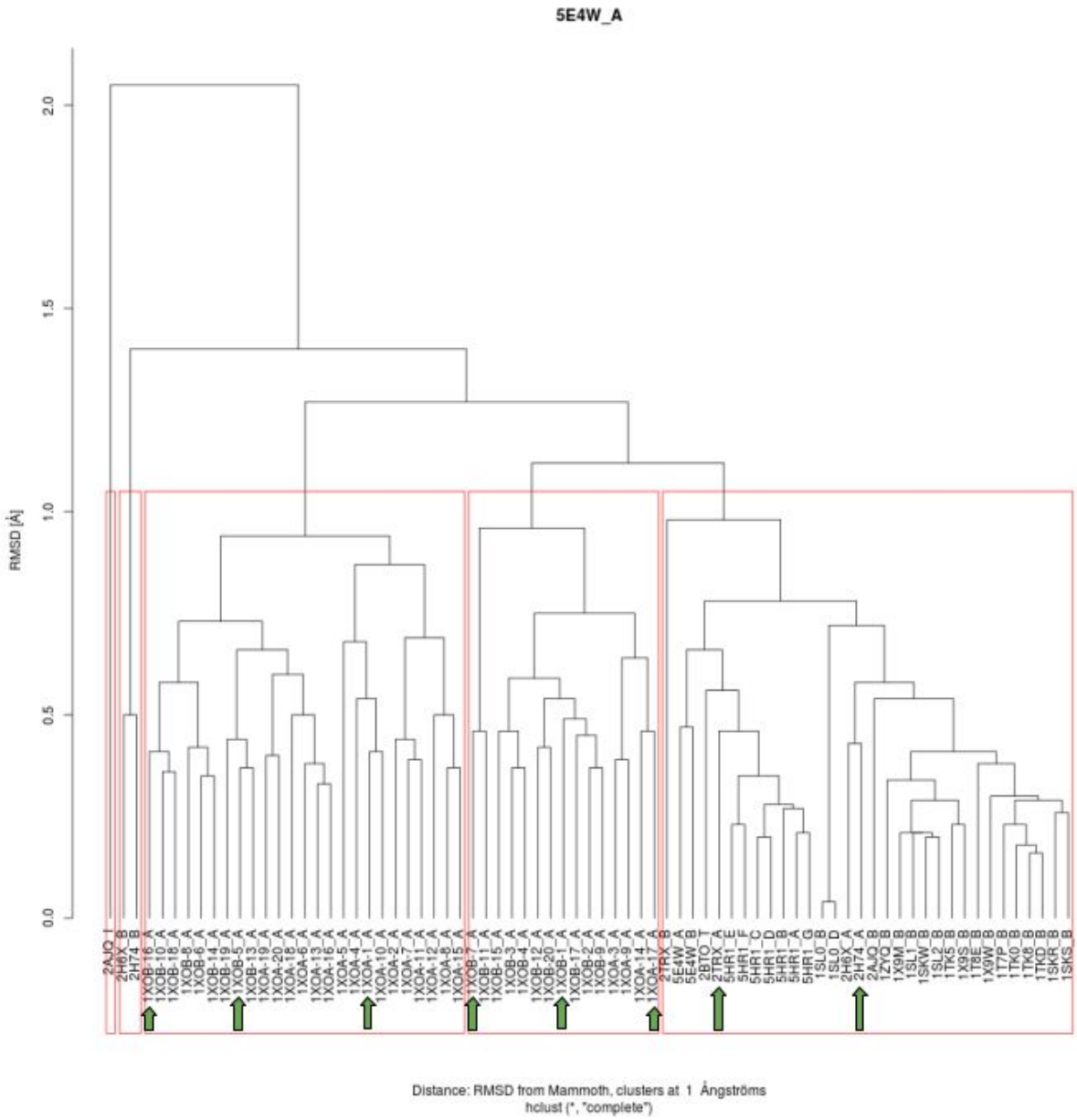


Figura 6.1.1. Dendrograma de distancias entre los confórmeros de la proteína THIO_ECOLI. En verde se encuentran seleccionados los confórmeros a evolucionar. Los cuadros rojos indican que esos pdb pertenecen al mismo cluster donde los confórmeros no difieren más de 0.4 RMSD entre ellos.

Los confórmeros seleccionados se encuentran marcados en color verde dentro de la figura. Hemos seleccionado los confórmeros de manera tal de tener una buena representación del espacio estructural de la proteína. Los confórmeros cumplen con los requisitos de ser de la cadena A y; para mayor simplicidad, contener la misma cantidad de residuos cristalizados que la estructura de referencia 2TRX. En la Tabla 6.1.1 se enumeran los confórmeros a evolucionar junto con el método utilizado para su obtención y la resolución (en el caso de la cristalización). En la columna pdb_model se incluye también el modelo; si el mismo no está presente se se da por entendido que nos referimos al modelo 1 de la estructura.

pdb_model	method	resolution
2TRX	X-ray	1.68
1XOB	NMR	--
1XOA	NMR	--
2H74	X-ray	2.4
1XOB_M5	NMR	--
1XOB_M7	NMR	--
1XOB_M16	NMR	--
1XOA_M17	NMR	--

Tabla 6.1.1 Estructuras pertenecientes a la proteína THIO_ECOLI que se evolucionaron. Puede apreciarse el método utilizados para la obtención de la estructura y la resolución del mismo. En la columna pdb_model se indica la estructura y el modelo particular que se evolucionara.

6.2. Análisis de la predicción de contactos de los confórmeros

Al realizar el procedimiento sobre las distintas conformaciones de la proteína THIO_ECOLI, se han obtenido resultados similares a los del capítulo 5. Los resultados de AUCs óptimos obtenidos para cada uno de los confórmeros pueden observarse en la Tabla 6.2.1. El método de covariación PSICOV presenta un mejor desempeño predictivo que los demás, con valores muy altos, mayores a

0.9, tanto para el AUC como para el AUC_01; seguido por el método MI con valores de AUC cercanos a 0.9 y de AUC_01 cercanos a 0.8 en promedio; finalmente, con resultados más alejados los métodos de coevolución DI y FROB.

La información de la Tabla 6.2.1 se complementa visualmente con las curvas ROCs de las Figuras 6.2.1, donde pueden apreciarse las similitudes y diferencias entre los resultados de cada uno de los métodos de coevolución para el MSA simulado. En las figuras se destaca nuevamente que el método de coevolución PSICOV presenta un mejor desempeño predictivo con respecto de los demás método de coevolución, vemos como las curvas ROC de color rojo, correspondientes al método PSICOV del evolucionado óptimo, tienen un desempeño casi perfecto.

pdb_name	mi	mi_01	di	di_01	frob	frob_01	psicov	psicov_01
1XOA	0.9132	0.8070	0.8394	0.7147	0.8526	0.6942	0.9789	0.9456
1XOA_M17	0.9007	0.7906	0.8346	0.7188	0.8465	0.6789	0.9673	0.9307
1XOB	0.9056	0.8172	0.8734	0.7407	0.8612	0.7125	0.9643	0.9263
1XOB_M5	0.9047	0.7831	0.8493	0.7243	0.8494	0.6694	0.9668	0.9159
1XOB_M7	0.9018	0.8085	0.8531	0.7174	0.8828	0.6996	0.9677	0.9264
1XOB_M16	0.9097	0.7911	0.8266	0.7105	0.8628	0.7019	0.9624	0.9283
2H74	0.8991	0.7624	0.8393	0.7154	0.8322	0.6476	0.9787	0.9453
2TRX	0.8930	0.7695	0.8623	0.7239	0.8375	0.6703	0.9787	0.9492

Tabla 6.2.1 AUCs obtenidos al analizar los MSAs simulados a partir de diferentes confórmeros de la proteína THIO_ECOLI. Se indica el confórmero y modelo (pdb), y los valores de los auc para cada uno de los métodos de coevolución.

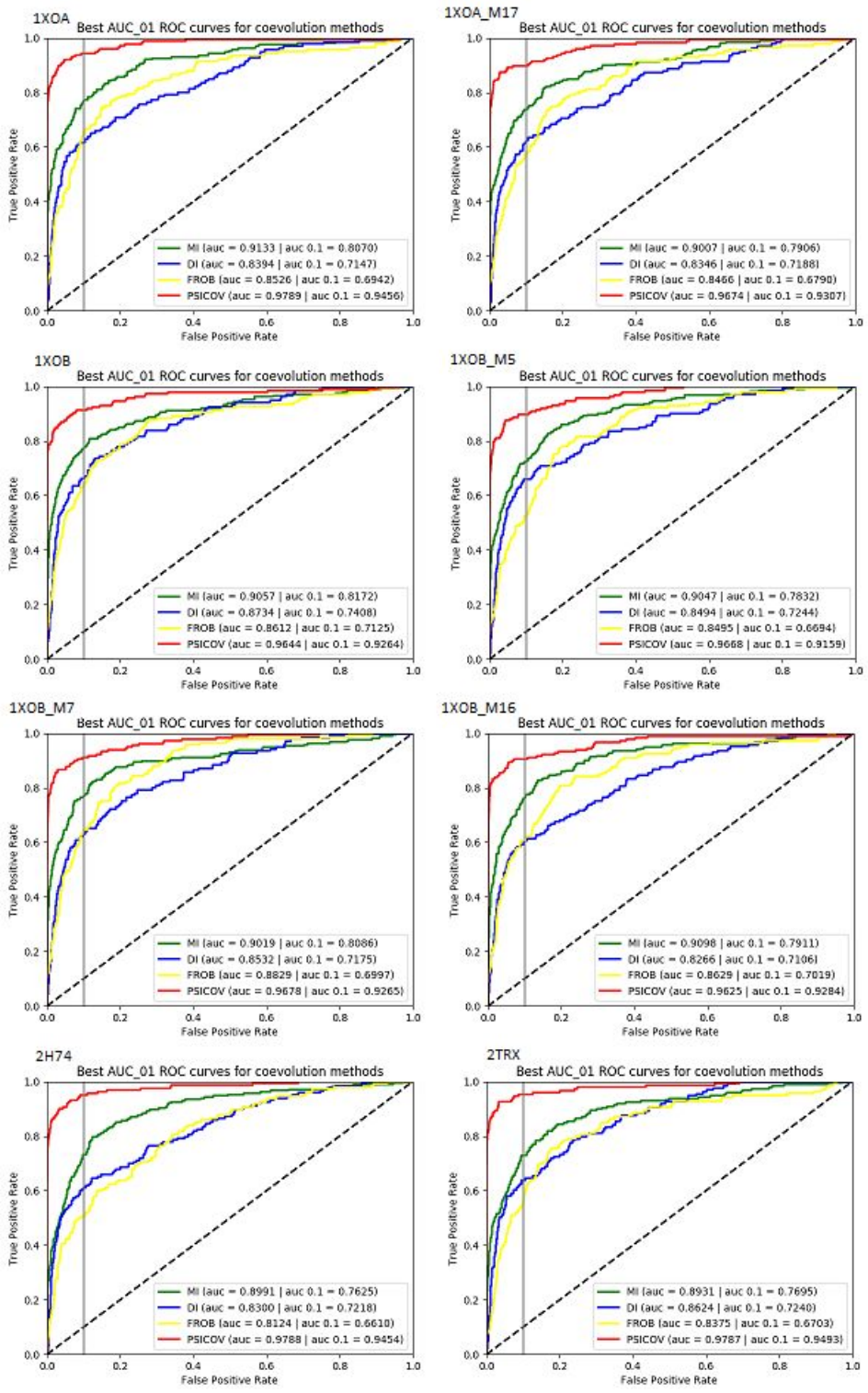


Figura 6.2.1. Curvas ROCs del evolucionado óptimo para las 8 estructuras analizadas. En diferentes colores se puede apreciar el método de covariación.

La comparación entre los AUCs correspondientes al cálculo de los MSA evolucionados con la información del MSA natural para cada uno de los métodos de coevolución también presenta similitudes a la información analizada en el capítulo 5 sobre la estructura 2TRX; y se comporta de igual forma para todos los confórmeros. Por tal motivo solamente se presenta la comparación del confórmero 1XOA. En la siguiente Figura 6.2.2 puede apreciarse que lo métodos PSICOV y MI obtienen mejores resultados con el evolucionado en comparación con la información natural; no así es el caso de los métodos DI y FROB.

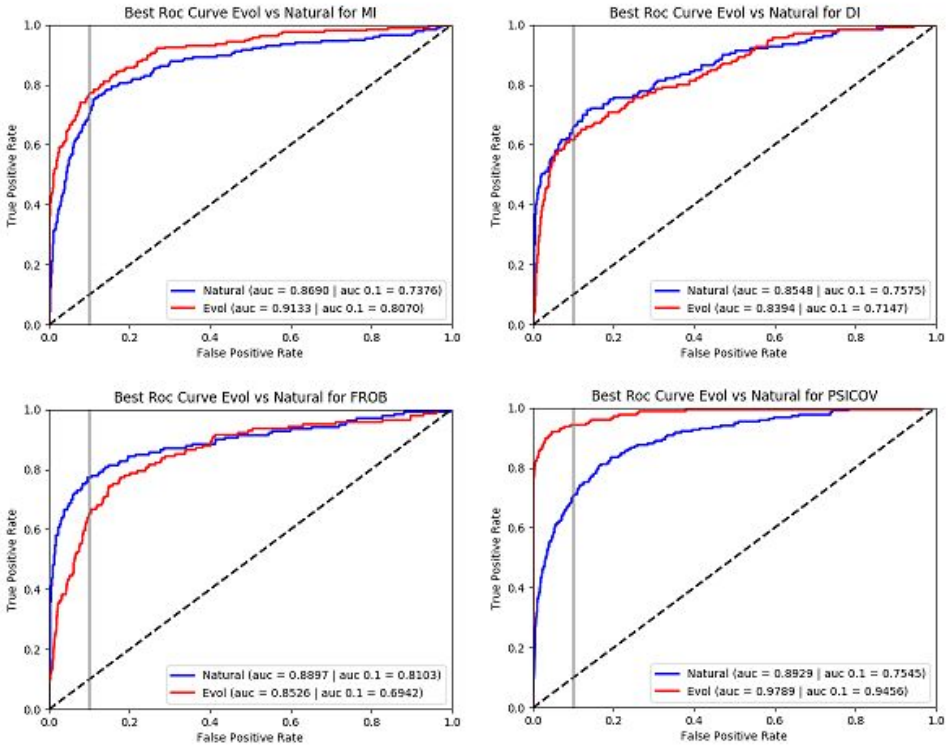


Figura 6.2.2. Curvas ROCs del confórmero 1XOA por método de coevolución para comparar los resultados naturales con los evolucionados. Los demás confórmeros obtienen resultados similares.

La comparación por cada par de posiciones y sus respectivos *scores* en el evolucionado y en el natural para cada uno de los métodos de coevolución pueden observarse en las Figuras 6.2.3 y 6.2.4 respectivamente. Solamente se representan los resultados de los confórmeros 1XOA y 1XOB ya que los resultados de los demás confórmeros son similares. Nuevamente se observa que

el método PSICOV tiene una mejor distribución de los pares de posiciones con alto *score*, los cuales se presentan como contactos.

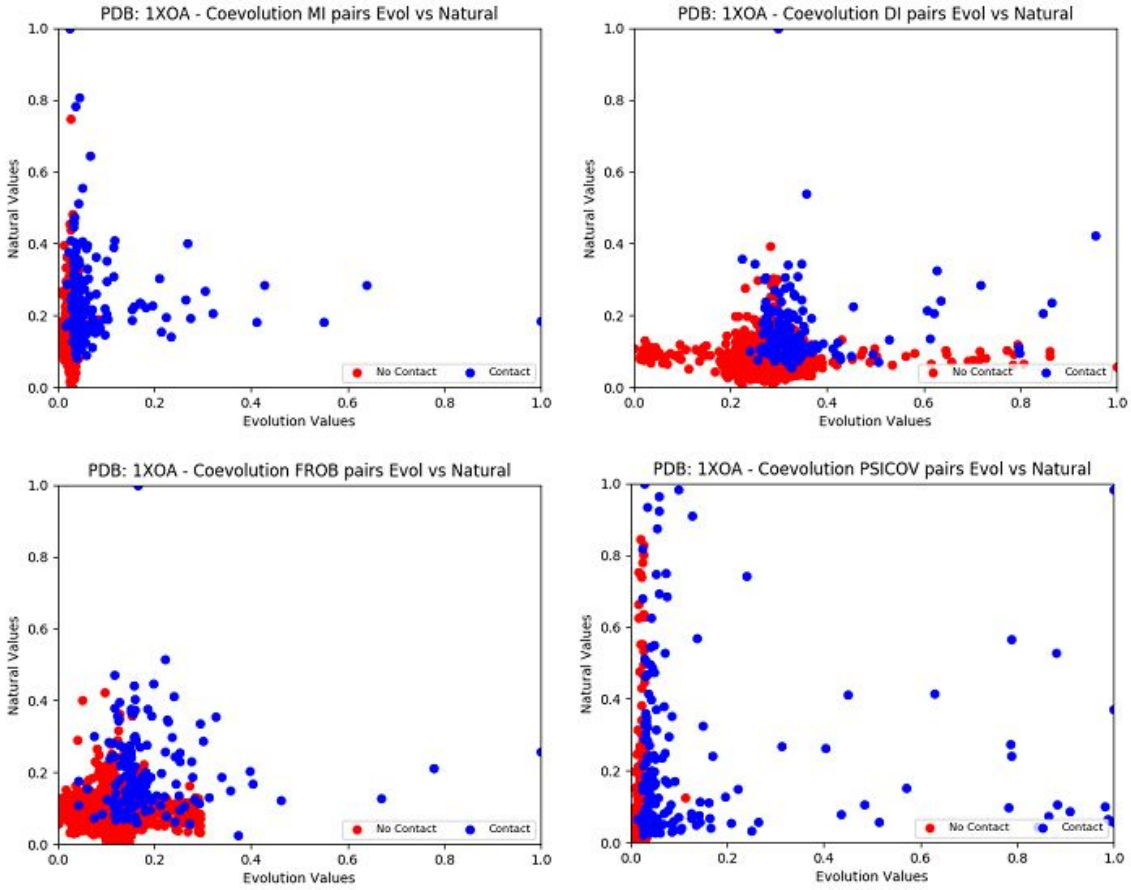


Figura 6.2.3 Gráficos de pares de posiciones y sus scores para cada uno de los métodos de coevolución estudiados; la información corresponde al confórmero 1XOA. Cada punto es un par de posiciones en donde en el eje x se observa el score del evolucionado, y en el eje y el score del natural. Los puntos en azul indican contactos y los rojos no contactos.

Al analizar el porcentaje de contactos por tops de los distintos confórmeros evolucionados, se observó que se obtienen resultados similares a los descritos en el capítulo 5, Tabla 5.3.2, con la estructura 2TRX; solo a modo de ejemplo en la Tabla 6.2.2 se presenta los resultados obtenidos para el confórmero 1XOB. En la tabla se describen el porcentaje de pares de posiciones que son contactos en el evolucionado y en el natural para los diferentes tops, columnas nat_ y evol_ seguidas por el nombre del método de covariación. Además, se encuentra

presente la columna match_ la cual indica cuántos pares de contactos en común existen entre el evolucionado y el natural para cada uno de los métodos.

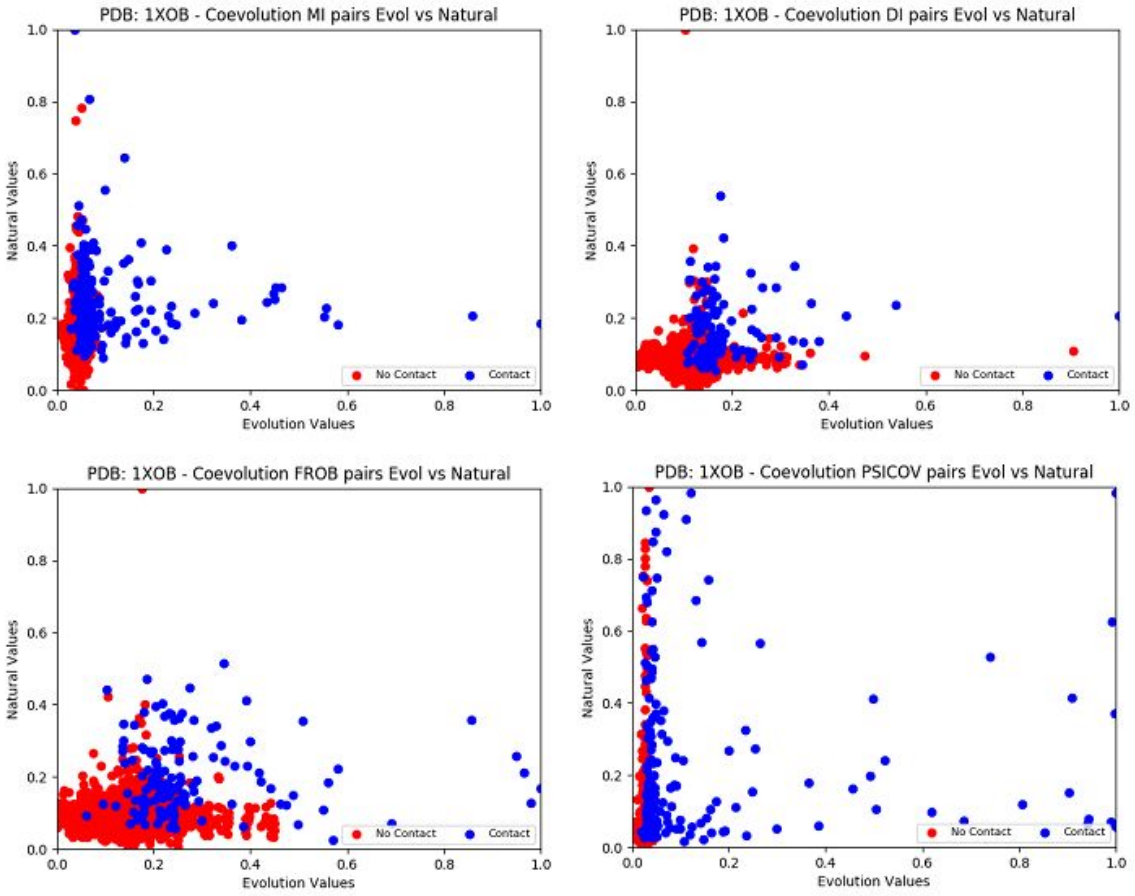


Figura 6.2.4 Gráficos de pares de posiciones y sus scores para cada uno de los métodos de coevolución estudiados; la información corresponde al confórmero 1XOB. Cada punto es un par de posiciones en donde en el eje x se observa el score del evolucionado, y en el eje y el score del natural. Los puntos en azul indican contactos y los rojos no contactos.

top %	nat_ MI %	evol_ M %I	match_ MI %	nat_ DI %	evol_ DI %	match_ DI %	nat_ FROB %	evol_ FROB %	match_ FROB %	nat_ PSICOV %	evol_ PSICOV %	match_ PSICOV %
0.5	64	100	4	70.83	41.66	4.16	79.16	62.5	8.33	65.38	100	3.84
1	60	98	11	75	38	18	77	42	8	63	100	17
2	52	82	17	59	34	18	63	31	16	61	98	31
3	42	68	20	45	36	19	51	31	18	51	90	34

4	38	56	19	35	32	19	42	26	18	46	70	33
5	34	49	19	29	29	17	36	24	16	40	58	30

Tabla 6.2.2. Descripción de la información de los tops para los métodos de covariación para el confórmero 1XOB. La columnas nat_MI y evol_MI describen el porcentaje de los pares que son contactos para el MSA natural con el método mi y el porcentaje de pares que son contactos para el evolucionado de MI optimo respectivamente; la columna match_MI indica el % de pares de posiciones, que son contactos, y que se encuentran en coincidencia en el MI natural y en el MI evolucionado. De forma idéntica se describe la información para los métodos DI, FROB y PSICOV.

Los resultados demuestran que el porcentaje de contactos encontrados en los confórmeros evolucionados con los métodos MI y PSICOV, sobre todo este último, es mucho mayor al del natural; puede apreciarse que el método PSICOV tiene una predicción perfecta para los tops 0.5 y 1 %, de 99% para el top 2 %, es destacable que incluso en el top 3 %, en donde se tienen en cuenta 158 pares de posiciones, continúe con un porcentaje del 90 %; punto en donde comienza a notarse la diferencia con MI, el segundo mejor método, el cual obtiene un 68 %.

A modo de ejemplo, para visualizar claramente cómo a medida que avanzamos en el análisis de los tops, de menor a mayor, los pares de posiciones que tienen alto *score* son los mismos que los contactos podemos analizar la siguiente Figura 6.2.5. En la misma se analizan los tops 0.5, 1, 2 y 3 % para el método PSICOV y el confórmero 1XOB.

Los resultados obtenidos en este capítulo afirman, que el procedimiento, junto con la optimización realizada sobre los parámetros de SCPE, sirven para diferentes estructuras de la misma proteína THIO_ECOLI. Siendo el método PSICOV el que mejor desempeño predictivo posee. En los siguientes capítulos se realizará el análisis de la información de forma conjunta.

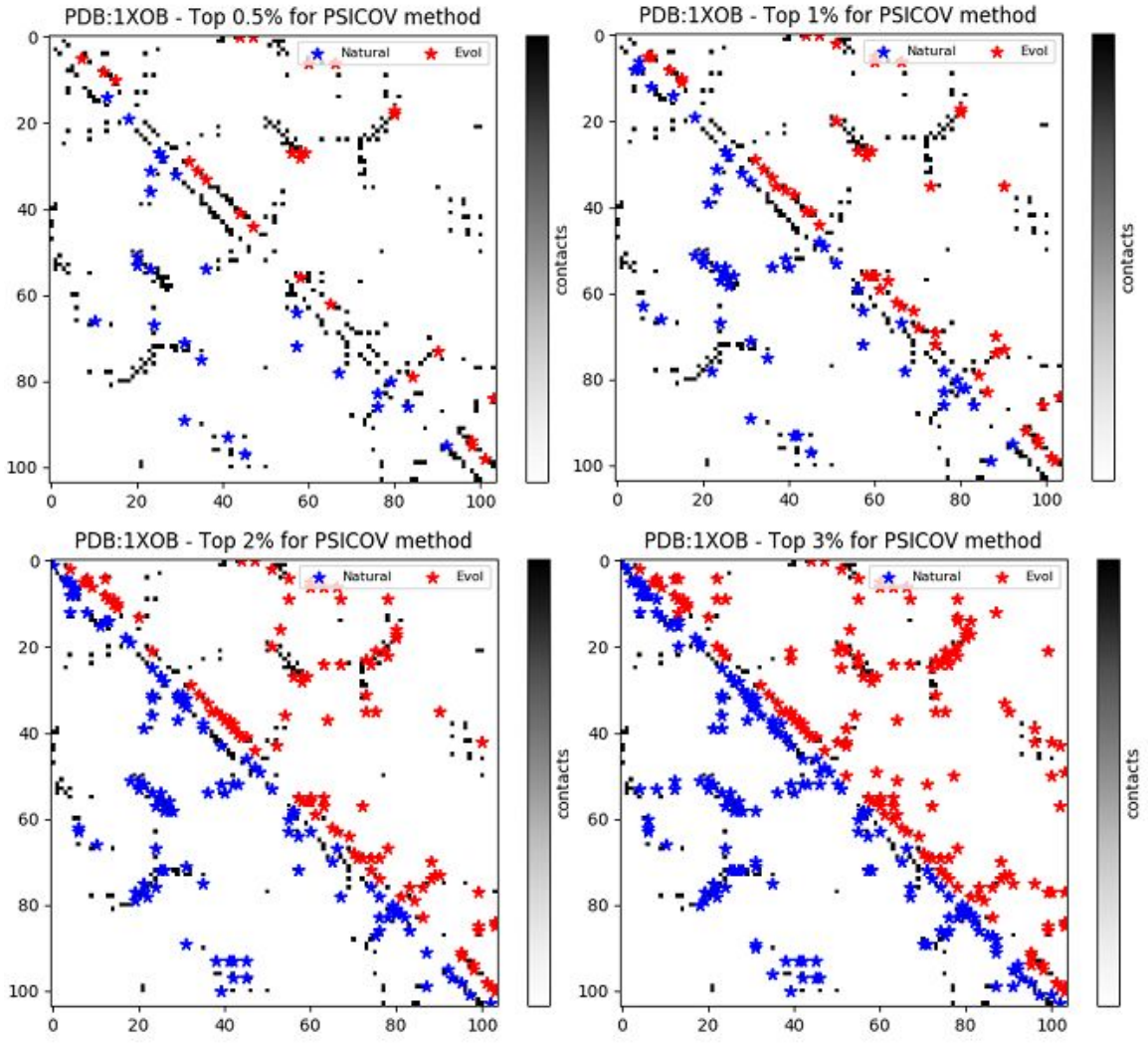


Figura 6.2.5. Matrices de contacto simétricas. Los contactos reales son los puntos sombreados negros. La información describe los tops 0.5, 1, 2 y 3 % respectivamente para el método PSICOV y el confórmero 1XOB; los puntos en azul que se encuentran en la diagonal de inferior son los pares de posiciones con alto score del MSA natural; por encima en la diagonal superior y de color rojo, los pares de posiciones con alto score del evolucionado.

6.3. Matriz de Contacto Conjunta: similitudes y diferencias entre estructuras

En el capítulo 5, los cálculos predictivos y el análisis se realizaron tomando como referencia una única estructura, la 2TRX, en consecuencia se utilizaba una única matriz de contacto. De la misma forma, fue el proceder de la sección

anterior, los cálculos predictivos para cada evolución de confórmero fueron realizados con su propia matriz de contacto. Entre los confórmeros analizados es esperable que existan muchas coincidencias de contactos, y haya en menor medida contactos específicos de cada confórmero.

Cualitativamente en la Figura 6.3.1, se puede observar en la superposición de las estructuras analizadas, que aunque son muy similares, pueden verse diferencias en algunos lugares (la superposición no es perfecta). Además, la matriz de contacto está calculada teniendo en cuenta solo las cadenas laterales, por lo que aunque veamos una muy buena superposición en el backbone (esqueleto, o carbonos alfa) de las estructuras, las diferencias de contactos formados por las cadenas laterales puede ser muy amplia.

Figura 6.3.1. Superposición de las 8 estructuras de la THIO_ECOLI analizadas en esta sección.

La Figura 6.3.2 indica cuantitativamente el porcentaje de contactos que comparten los confórmeros entre sí, tomando como el total de contactos la sumatoria de todos los contactos de todos los confórmeros. Se observa porcentajes de contactos compartidos entre 66.28 al 80.46, siendo el promedio 74.53.

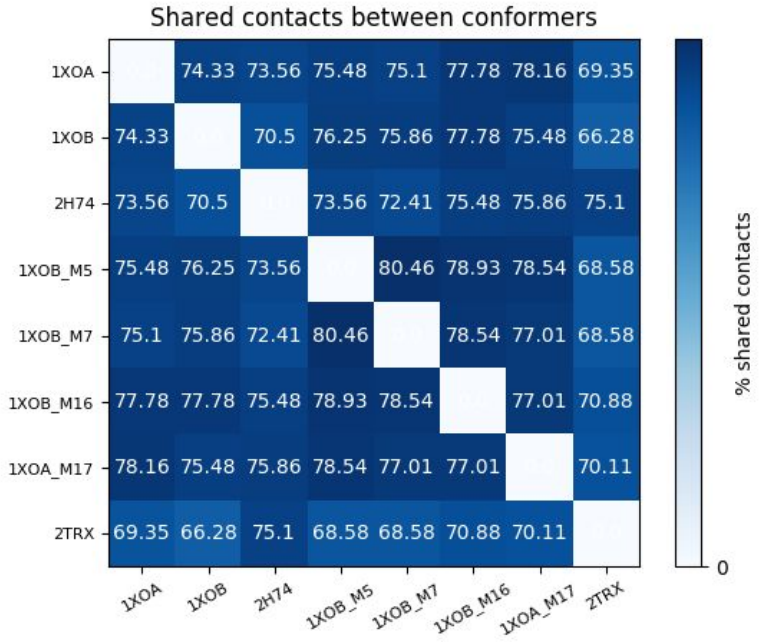


Figura 6.3.2. Porcentaje de contactos compartidos entre los confórmeros con respecto al total de los contactos de todos los confórmeros. Sobre los ejes x e y los confórmeros analizados, la información numérica se refiere al porcentaje de contactos compartidos.

Para conocer la cantidad de veces que un contacto ocurre en los diferentes confórmeros se realizará la suma de las matrices de contacto de cada uno de ellos, generando así una matriz denominada *Matriz de Contacto Conjunta* (Figura 6.3.3 - 6.3.4). Los valores de las celdas de dicha matriz varían entre 0 y 8, indicando la cantidad de estructuras que tienen como contacto ese par de posiciones. Esto equivale a la probabilidad de ser contacto; el valor 0 indica que ninguna estructura tiene a ese par de posiciones como contacto, el valor 1 indica que solo una estructura lo tiene como contacto; así hasta llegar al máximo que es

8, lo que significa que todas las estructuras poseen ese contacto, cuando esto último ocurre, el contacto se define como totalmente conservado.

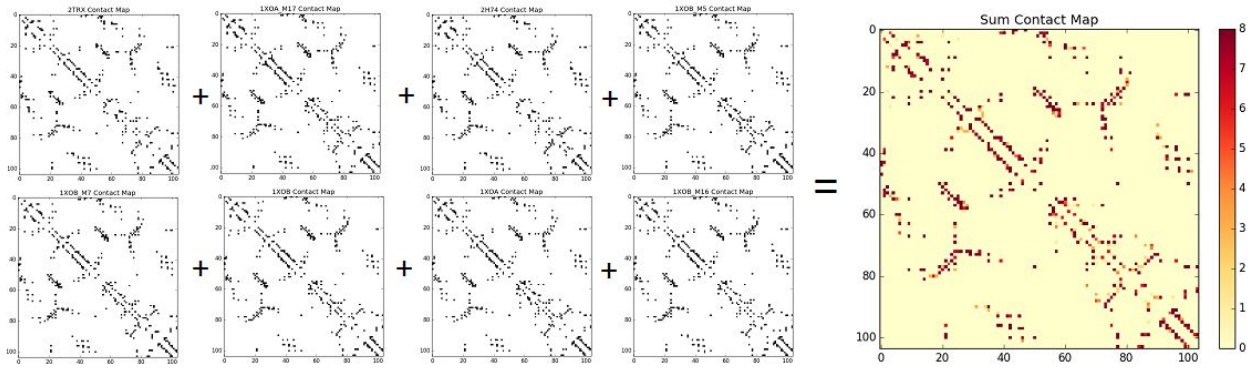


Figura 6.3.3. Esquematización de la suma de las matrices de los confórmeros analizados para la obtención de la matriz de contacto conjunta. Las matrices a la izquierda de la ecuación pertenecen a cada uno de los confórmeros analizados de la proteína THIO_ECOLI; la matriz resultante contiene la suma de las mismas.

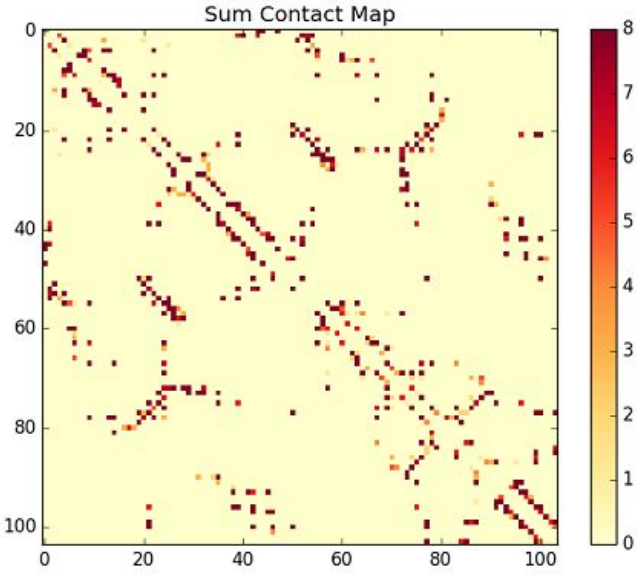


Figura 6.3.4. Matriz simétrica resultante de realizar la suma de las matrices de contactos de los confórmeros de la proteína THIO_ECOLI analizados. En rojo puede apreciarse las zonas más conservadas, mientras que en amarillo las de menor conservación.

En la Figura 6.3.4 podemos apreciar la matriz de contactos conjunta. El análisis de esta matriz nos permite conocer el grado de conservación de los contactos y las zonas donde los mismos ocurren, la escala de colores es amarillo-rojo, siendo rojo los contacto más conservados.

En total existen 542 contactos en la matriz de contacto conjunta, para completar el análisis se analiza la información dispuesta en la Figura 6.3.5 que indica la distribución de los mismos, es decir cuantos contactos aparecen en una

única estructura, cuantos en dos, y siguiendo hasta ver cuantos aparecen en las todas las estructuras analizadas. La cantidad de estructuras se indica en #structures, mientras que el número de contactos y el porcentaje del total, se describe sobre el gráfico. Por ejemplo, existen 22 contactos que solo se encuentran en dos estructuras (color bordo), siendo estos el 4.04% del total de los contactos.

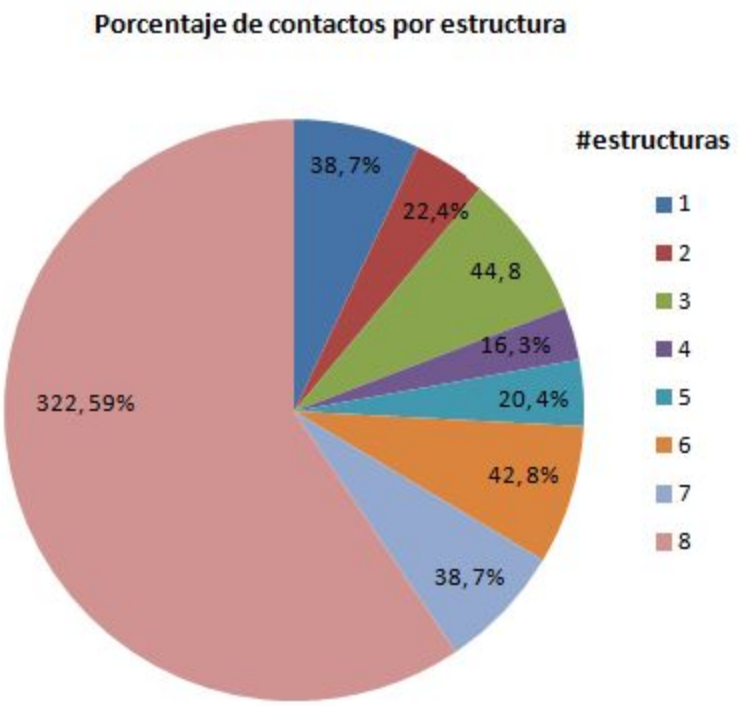


Figura 6.3.5. Distribución de los contactos dentro de la matriz de contactos conjunta por cantidad de estructuras en las que aparecen, #structures describe el número de estructuras, mientras que sobre el gráfico se visualiza la cantidad de contactos y el porcentaje (separados por ,)

Del total de 542 contactos hay 322, un 59.4%, que aparecen en las 8 estructuras y que son totalmente conservados en nuestro análisis. Esto es de esperar ya que si bien los confórmeros son diferentes, las similitudes de las estructuras van a ser mucho mayor que las diferencias. Inclusive, si realizamos el corte en contactos que se encuentren en 4 o más estructuras, lo que denominamos medianamente conservados, el resultado es de 438, un 80.81 % del total. Si bien casi el 60% de los pares en contacto, lo están en todas las estructuras, hay un 40% que puede variar según la estructura elegida.

La matriz de contacto conjunta será la utilizada a lo largo del capítulo para realizar los cálculos de la información conjunta; para ello cuando nos refiramos a la misma y quisiéramos diferenciar teniendo en cuenta la conservación de los contactos, se utilizará la notación *contacts* >= *n* para indicar que solo se tendrán en cuenta como contactos los que aparecen n o más veces en los confórmeros. Siendo contacts >= 1 todos los contactos de la matriz conjunta, contacts >= 4 los medianamente conservados y contacts = 8 los totalmente conservados.

6.4. Sumatoria de puntajes de covariación

El método para analizar la covariación de forma conjunta es realizando la sumatoria de los top de covariación resultantes de la evolución de los confórmeros para cada uno de los métodos. Para ello en una primera etapa se obtienen los top 1% de covariación de todos los confórmeros; esto es los 51 pares de posiciones con más alto *score* de cada MSA evolucionado. Luego, la agrupación se lleva a cabo a través de la sumatoria de estos 8 resultados de top, generando así lo que denominaremos sumatoria del top 1 % de la información mutua. Esta operación nos permite obtener cuales son los pares de posiciones que aparecen en la mayoría de los tops de los confórmeros. En la Tabla 6.4.1 puede apreciarse la sumatoria de la covariación, donde la columna Sum top COV indica la cantidad de veces que el par de posiciones (Position 1 y Position 2) aparecen en los top 1% de covariación de los MSA de los confórmeros evolucionados. El procedimiento es realizado para los tops 0.5, 1, 2, 3, 4 y 5 y para cada método de covariación estudiado.

Posición 1	Posición 2	Sum top COV
29	59	8
45	48	8
84	87	8

57	59	8
30	33	8
32	35	8
9	13	8
70	75	8
21	52	8
93	96	8
7	61	8
.....	.....	.....
10	64	7
....	....	....
7	67	5
....	....	.....

Tabla 6.4.1 Ejemplo: Sumatoria de top 1 % de la MI. En la columna Sum top COV se indica en cuantos tops aparece. Se muestra la información acotada en forma de ejemplo.

Además de disponer de la sumatoria de tops, previamente en la sección 6.3, se realizó el cálculo de la matriz de contacto conjunta, la cual se definió como la sumatoria de todas las matrices de contactos pertenecientes a cada uno de los confórmeros evolucionados. Esta información puede relacionarse para obtener la correlación entre la suma de los top de score de covariación y la suma de las matrices de contacto; es de esperar que si un par de posiciones aparece en los 8 tops de covariación también aparezca como contacto en las 8 matrices. La información de la sumatoria de contactos que se aprecia en la Tabla 6.4.2, corresponde a la sumatoria de top 1% MI. La información se encuentra ordenada primero por Sum top COV y luego por la cantidad de veces que aparece como contacto, columna Contacts.

Posición 1	Posición 2	Sum top COV	Contacts
29	59	8	8
45	48	8	8
84	87	8	8
57	59	8	8
...	....	...	...
57	60	7	8
80	85	7	8
...	...	...	...
42	46	7	7
12	16	7	7
11	16	7	7
.....	....	...	...
77	87	6	8
....	...	...	...

Tabla 6.4.2. Ejemplo: Sumatoria de top 1 % de MI junto con la matriz de contactos conjunta. Los pares de posiciones se indican en las columnas Posición 1 y Posición 2, mientras que la sumatoria de top MI en Sum top COV y la Sumatoria de la matriz conjunta en Contacts. El mismo procedimiento se realiza para obtener la sumatoria de top 1 % para los métodos DI, FROB y PSICOV. Se muestra la información acotada en forma de ejemplo.

Para calcular la correlación entre la cantidad de veces que los pares de posiciones aparecen en los tops y la cantidad de veces que dichos pares son contactos, analizaremos los coeficientes de correlación de pearson, kendall y spearman. La tabla 6.4.4 detalla los valores obtenidos para los diferentes tops por cada uno de los métodos.

top	mi_ pearson	mi_ kendall	mi_ spear	di_ pearson	di_ kendall	di_ spear	frob_ pearson	frob_ kendall	frob_ spear	psicov_ pearson	psicov_ kendall	psicov_ spear
0.5	0.2534	0.1334	0.1612	0.046	0.0809	0.0972	-0.0433	0.014	0.0228	0.3352	0.2306	0.2886
1	0.428	0.3459	0.4255	0.2202	0.171	0.2021	0.0242	0.0149	0.0197	0.2554	0.186	0.226
2	0.6007	0.5208	0.6101	0.246	0.2018	0.2361	0.1304	0.1003	0.1147	0.3065	0.2157	0.263
3	0.6042	0.4841	0.562	0.2869	0.2194	0.258	0.1343	0.1043	0.1198	0.6173	0.5113	0.6142
4	0.5976	0.4531	0.5211	0.2637	0.1855	0.2192	0.2386	0.1932	0.2182	0.8078	0.7163	0.8155
5	0.5932	0.4262	0.4889	0.2978	0.2249	0.2602	0.2578	0.1893	0.2148	0.849	0.7317	0.8035

Tabla 6.4.4. Coeficientes de correlación de pearson, kendall y spearman (columna spear) para cada uno de los tops analizados y métodos de covariación.

Tomando como parámetro la correlación de spearman, a partir del 2 % la los valores obtenidos en el método MI no mejoran; en cambio PSICOV obtiene a partir del 3 % una correlación de 0.61 que aumenta a 0.81 en el top 4 % y por último el resultado de mejor correlación 0.80 en el top 5%.

La Figura 6.4.1, gráfica la correlación entre la variables Contacts y Sum top COV para diferentes tops analizados para el método PSICOV. El tamaño de cada uno de los puntos indica la cantidad de pares de posiciones; por ejemplo vemos en los tops 1 y 2 % que en la coordenadas (x, y) = (8, 8) el punto es de mayor tamaño; esto significa que existen varios pares de posiciones que se encuentran en los 8 tops 1 y 2 % y que a su vez son contactos en las 8 estructuras; es decir a los contactos más conservados les corresponde una alta señal de covariación.

Del gráfico se puede extraer mucha información interesante, como por ejemplo, que los pares con scores más altos (que aparecen en el 0.5%) no responden a una tendencia, es decir pueden aparecer en el top 0.5% entre 1 y 8 veces y ser contacto en 1 a 8 estructuras (los puntos están esparcidos por todo el espacio). Conforme vamos considerando más pares de posiciones (top 1%), comienza a hacerse más evidente que los pares de posiciones que son contacto

en 8 estructuras, aparecen en el top 1% más veces (x,y = 8,8). Sin embargo comienza a notarse que pares en contacto en todas las estructuras (y = 8) pueden salir en 1, 2, 3,...7 tops 1% (x = 1 a 7). Osea, que calcular covariación en diferentes confórmeros, revela diferentes pares en contacto. Por último, algo sorprendente es que al considerar el top 3% de los pares, aparecen muchos pares que no son contacto (y = 0) y que aparecen en uno de los tops 3% (x = 1). Estos son falsos positivos, por lo cual a la vez que aumentamos el número de pares presentes en todos los confórmeros (conservados), aumentamos la cantidad de falsos positivos que aparecen al calcular covariación en alguno (uno) de los confórmeros.

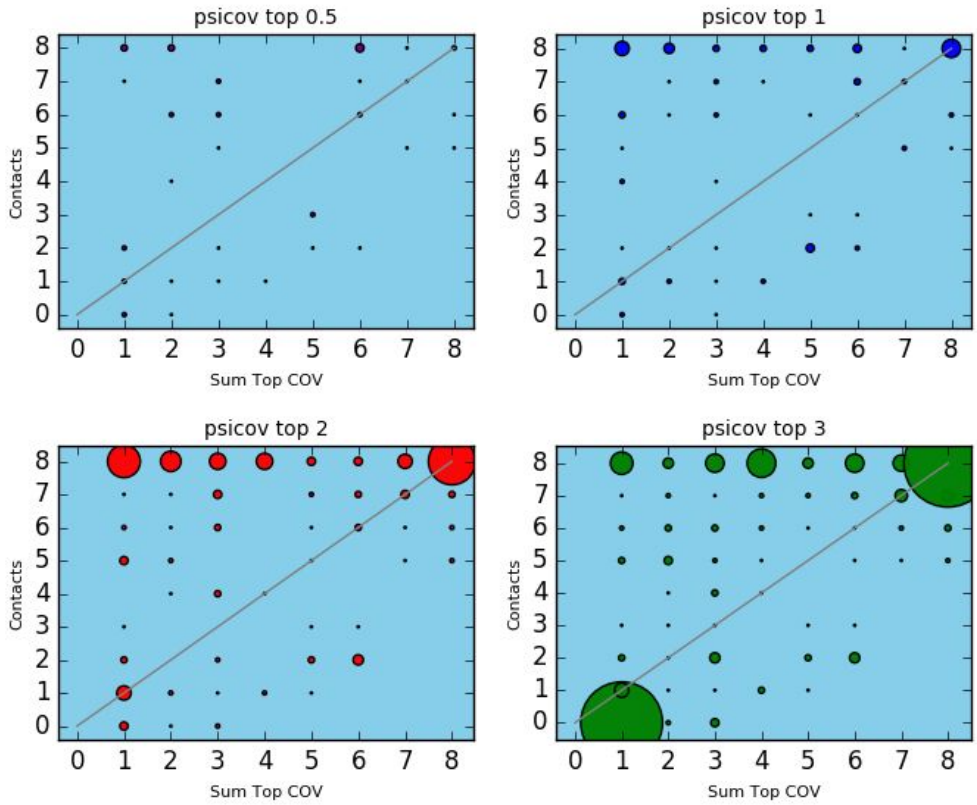


Figura 6.4.1. Correlación de las variables Contacts y Sum top COV. Es decir, entre la cantidad de veces que los pares de posiciones aparecen en los tops n% y la cantidad de veces que dichos pares con contactos. Se analizan de izquierda a derecha los tops 0.5, 1, 2 y 3 %. El tamaño del punto en la coordenada indica la cantidad de pares de posiciones.

Esto sugiere que al contar con más confórmeros de una estructura, si nos quedamos con los pares que aparecen en los tops 3% de las n estructuras, estamos asegurando que esos son contactos conservados. Nos estaríamos quedando con los pares que se encuentran en x,y = 8,8. A la vez, si desechamos los pares que aparecen en un sólo 3%, estaríamos eliminando la mayor parte de todos los falsos positivos. La desventaja es que eliminamos también algunos verdaderos positivo (los de la línea x = 0, y = 1-8).

Los gráficos que se analizan a continuación tratan de profundizar más e intentan responder si los pares de posiciones que comparten los diferentes confórmeros con alto *score* de MI, DI, FROB y PSICOV son los mismos que los confórmeros comparten en su estructura? Para ello primero observemos la Figura 6.4.2, en donde visualizamos las cuatro matrices, una por cada método de covariación; en la diagonal superior se encuentra la probabilidad de contactos; y en la matriz inferior la probabilidad de top 1 %, es decir, la probabilidad de los pares de aparecer en el top 1 % en los confórmeros. Los pares de posiciones que se encuentre más conservados ya sea en los contactos (diagonal superior) como en las veces que se encuentra en el 1% (diagonal inferior) estarán coloreados más oscuros sobre el heatmap.

Al igual que pudimos apreciar durante los capítulos anteriores la señal de top en los métodos de DI y FROB se ve afectada por la conservación de columnas en el MSA evolucionado, por tal motivo se destacan posiciones horizontales y verticales; no ocurre lo mismo con MI y PSICOV, en donde levemente se aprecia un parecido a los contactos observados en la diagonal superior. Pero la señal del top 1 % es muy leve y no se distingue claramente. En la Figura 6.4.3 se observa la matriz de sumatoria del top 3 % para el método PSICOV, la cual está acompañada de dos dendrogramas, en el superior se describe la distancia de los confórmeros a nivel estructural; y en la parte derecha el dendrograma que describe la clusterización de la sumatoria del top 3 %.

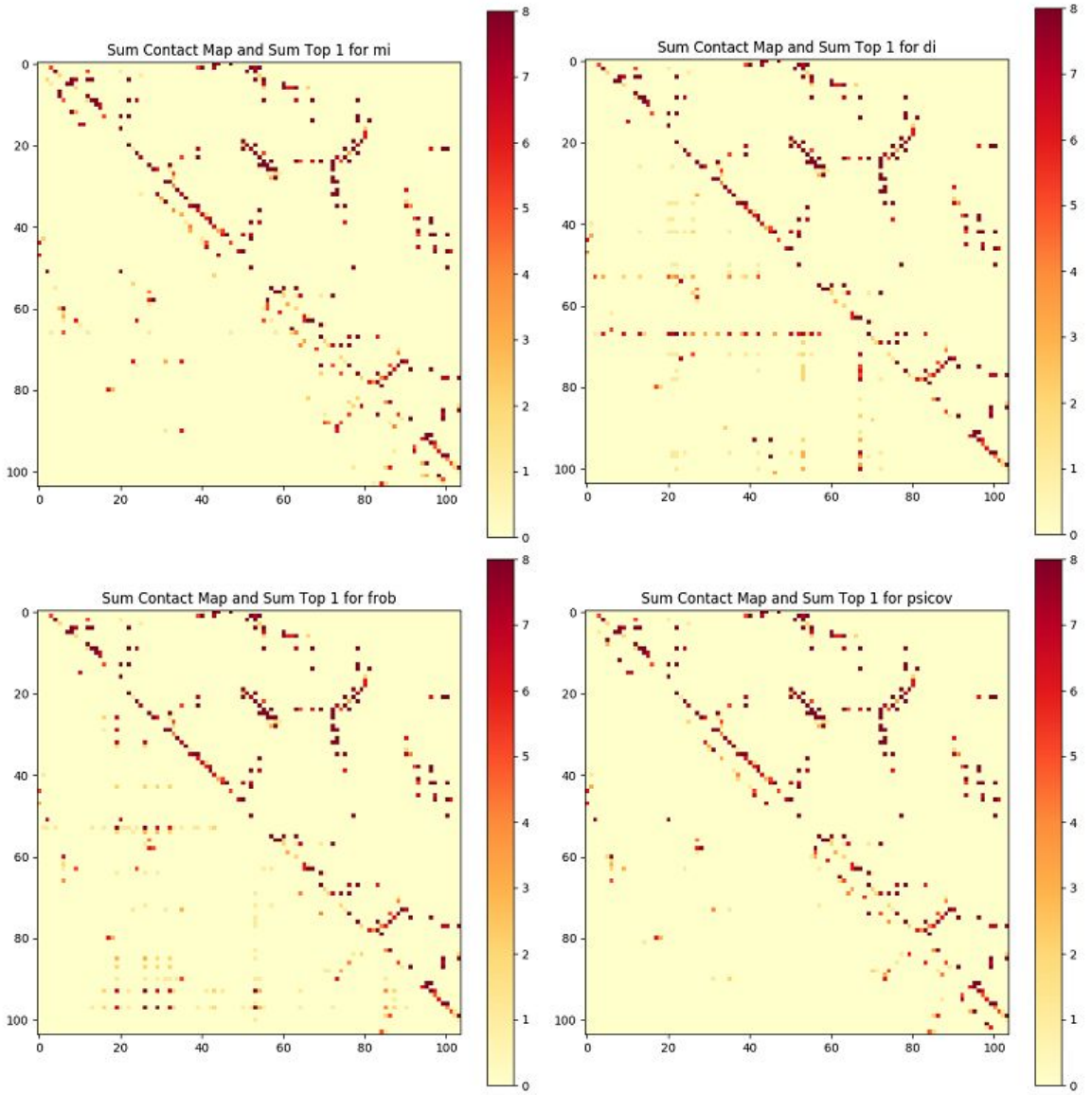


Figura 6.4.2. Se visualizan cuatro matrices, una por cada método de covariación; en la diagonal superior se encuentra la probabilidad de contactos; y en la matriz inferior la probabilidad de los pares de aparecer en el top 1 % en los confórmeros.

El método de clusterización para medir la distancia tanto en el dendrograma de contactos como en la sumatoria de top de covariación está dado por la ecuación:

$$d(u,v) = \min(dist(u[i],v[j]))$$

para todos los puntos *i* en el cluster *u* y para todos los puntos *j* en el cluster *v.* Este método de clusterización es conocido como algoritmo del vecino más cercano, en inglés Nearest Point Algorithm.

Al analizar la Figura 6.4.3, centrándonos primero en la información que vemos sobre la matriz, observamos una similitud entre la probabilidad de contactos y la señal conjunta de la sumatoria del top 3 % del método PSICOV. Este gráfico nos brinda una visualización cualitativa de la correlación calculada en la Tabla 6.4.4. En la misma se puede observar que la señal de lo pares de posiciones que se encuentran en la sumatoria del top 3 % se asemeja a la información de contactos en la matriz superior.

El dendograma superior describe la distancia que existe entre los confórmeros a nivel estructural; es decir, los confórmeros que compartan más pares de posiciones que están en contacto se encontraran cercanos en el dendograma. Del mismo modo, el dendograma de la derecha describe la distancia entre confórmeros a nivel de top 3 % de scores del método PSICOV; osea se encontrarán más cercanos en el dendograma los confórmeros que compartan más pares de posiciones en el top 3 %. Estos dendrogramas nos permiten analizar si los confórmeros que se encuentran cercanos a nivel estructural (comparten más contactos) también se encuentran cercanos a nivel de señal de evolución (comparten más pares de posiciones dentro del top 3 %); vemos por ejemplo que los confórmeros 2TRX-2H74 se corresponden a nivel estructural y a nivel de señal en el top 3 % ya que en los dos dendrogramas aparecen como vecinos inmediatos. De igual forma podemos observar otros ejemplos, como los confórmeros 1XOA-1XOA_17 y 1XOB_5-1XOB_7.

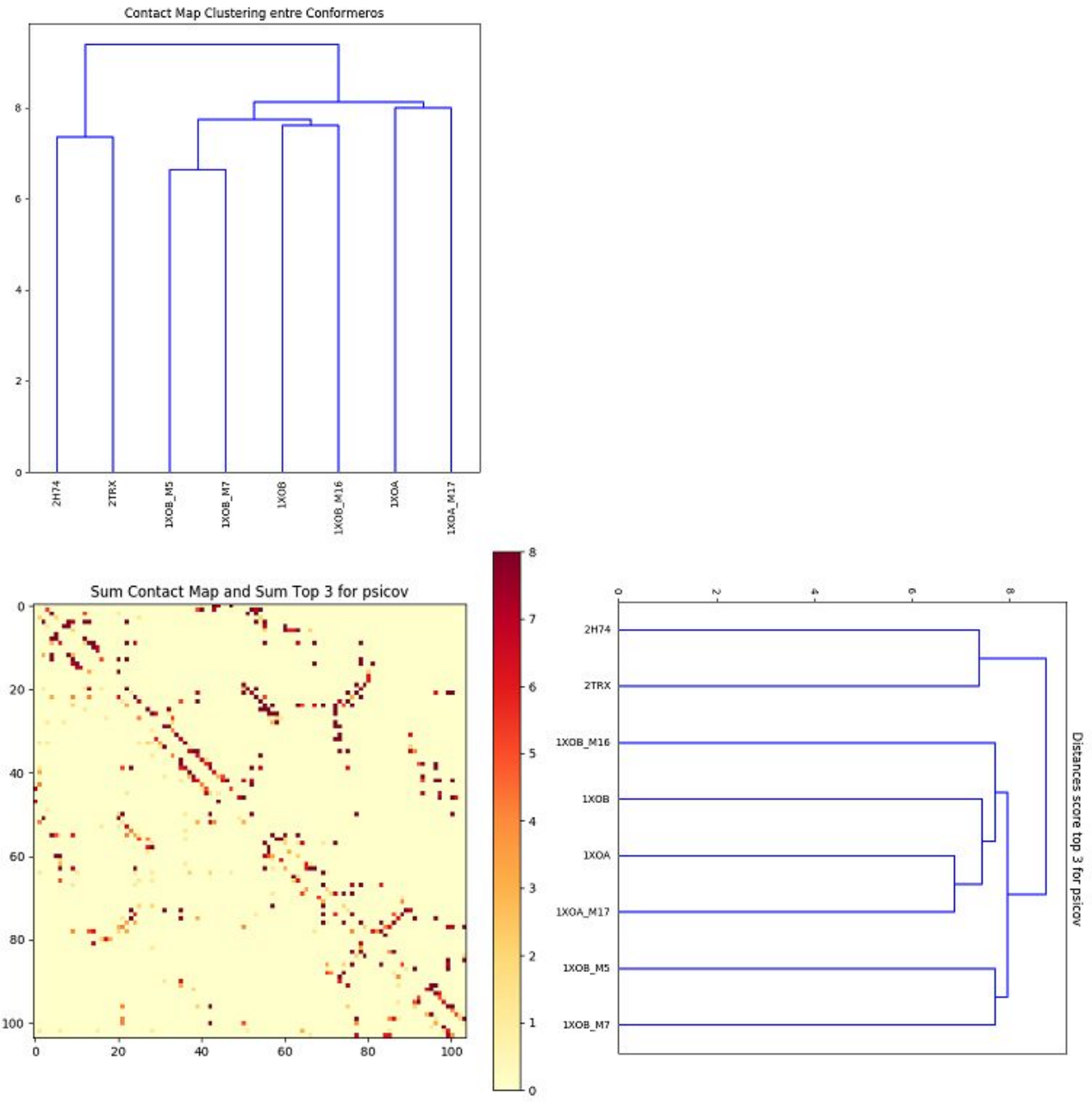


Figura 6.4.3. En la diagonal superior de la matriz se encuentra la probabilidad de ser contactos; y en la matriz inferior la probabilidad de los pares de aparecer en el top 3 % en los confórmeros. El dendograma superior describe la distancia a nivel de estructura entre los confórmeros; y el dendograma de la derecha la distancia a nivel de top 3 %.

6.5. Comparación de la Información Mutua Conjunta y Matriz de Contacto Conjunta con los resultados del MSA Natural

En el capítulo anterior se definió el concepto de sumatoria de top, resultado que contiene cuales son los pares de posiciones que más veces aparecen en los diferentes tops; por ejemplo, la sumatoria de tops 1 % contiene ordenados de forma descendente los pares de posiciones que más veces aparecen en el top 1 % para cada método de covariación. En este capítulo nos disponemos a comparar y analizar la información natural con la información de la sumatoria de tops, para ello se realiza el siguiente procedimiento:

- Se realiza el top 1 % del resultado de MI para la información natural, en donde obtenemos los 51 pares de posiciones con mas alto MI.
- Se obtienen los primeros 51 pares de posiciones de la sumatoria de top 1 % de la información obtenida de los confórmeros; al estar ordenados primero por cantidad de veces que aparecen en el top 1 % y luego por la cantidad de veces que son contactos en la matriz de contacto conjunta, nos retornan los mejores pares de posiciones en cuanto a *score.*
- Se calcula que porcentaje de esos pares de posiciones, tanto del natural como del evolucionado, son contactos; y cuantos pares de posiciones tienen en común entre el natural y el evolucionado.
- El procedimiento se repite aplicando un *threshold* de contactos, esto es para tener en cuenta la conservación de los mismos; definición que fue realizada en el capítulo 6.3. Indicando contacts >= n; siendo contacts >= 1 todos los contactos de la matriz conjunta, contacts >= 4 los medianamente conservados y contacts = 8 los totalmente conservados.

- Repetimos el procedimiento para la sumatoria de tops de 1 % para los demás métodos de covariación; por último
- Repetimos el procedimiento para los diferentes tops analizados: 0.5, 2, 3, 4 y 5 %.

El resultado de realizar este procedimiento puede observarse en la Tabla 6.5.1. Para cada método de covariación se agregaron 3 columnas, la primera (nat) indica el porcentaje de contactos del natural, la segunda (evol) del evolucionado; y la última, la cantidad de pares de posiciones que comparten el evolucionado con el natural. Adicionalmente, y para estudiar la relación con la conservación de los contactos, se incorpora la columna contact_threashold, la cual indica el grado de conservación que se tiene en cuenta para realizar la predicción.

top%	contact_ threashold	nat_ mi%	evol_ mi%	match _mi%	nat_ di%	evol_ di%	match _di%	nat_ frob%	evol_ frob%	match _frob%	nat_ psicov%	evol_ psicov%	match_ psicov%
0.5	1	72	100	4	75	45.83	12.5	83.33	62.5	8.33	69.23	100	3.84
0.5	4	72	76	4	70.83	37.5	12.5	79.16	45.83	4.16	69.23	73.07	3.84
0.5	8	64	32	4	62.5	16.66	8.33	58.33	20.83	4.16	57.69	26.92	3.84
1	1	64	100	11	79	48	22	81	48	8	67	100	17
1	4	64	78	11	75	42	18	79	38	6	65	78	15
1	8	54	49	11	59	20	10	61	22	6	55	44	15
2	1	59	95	21	69	47	19	71	37	19	66	99	30
2	4	56	82	21	65	44	17	68	32	17	62	83	28
2	8	43	55	18	47	25	9	52	21	15	48	51	23
3	1	49	81	24	52	44	20	59	35	20	56	97	34
3	4	46	70	22	48	41	18	54	31	18	53	84	31
3	8	34	46	18	35	25	12	39	20	14	38	56	25
4	1	44	66	23	42	37	19	50	31	20	53	88	38
4	4	41	57	22	38	34	18	45	28	19	48	76	35

4	8	29	37	18	27	22	12	32	18	13	33	48	25
5	1	39	58	24	35	32	17	43	28	19	46	72	34
5	4	36	50	23	31	29	15	39	25	18	42	62	31
5	8	25	33	18	23	20	11	27	17	14	29	40	23

Tabla 6.5.1. Comparación de los tops de la sumatoria de los métodos de covariación MI, DI, FROB y PSICOV de la información evolucionada y de la información natural.

Los resultados obtenidos teniendo en cuenta todos los contactos en conjunto, que se visualizan en la Figura 6.5.1, demuestran que los métodos de PSICOV y MI tienen un desempeño predictivo superior a los métodos de FROB y DI, cómo se ha podido observar durante todo el proyecto. PSICOV supera a los resultados de MI a partir del top 2 %, teniendo una predicción muy alta; inclusive para el top 3 % obtiene una predicción de contactos en conjunto de 97 % teniendo en cuenta todos los contactos de la matriz.

Una de las preguntas realizadas era si al evolucionar varios confórmeros de la misma proteína, nos brindaba información complementaria de importancia; y si en base a esa información se mejoraba la predicción de contactos ? Al analizar nuevamente la Tabla 5.3.2 en donde se muestran los resultados al evolucionar solamente la estructura 2TRX y comparando dichos resultados con los de la Tabla 6.5.1 podemos afirmar que es así; si por ejemplo al resultado de PSICOV para el top 3 %, con un valor de 97% de predicción de contactos, lo comparamos con la predicción del 3 % del método PSICOV cuando solamente se analizaba la estructura 2TRX se obtiene una predicción del 79 %.

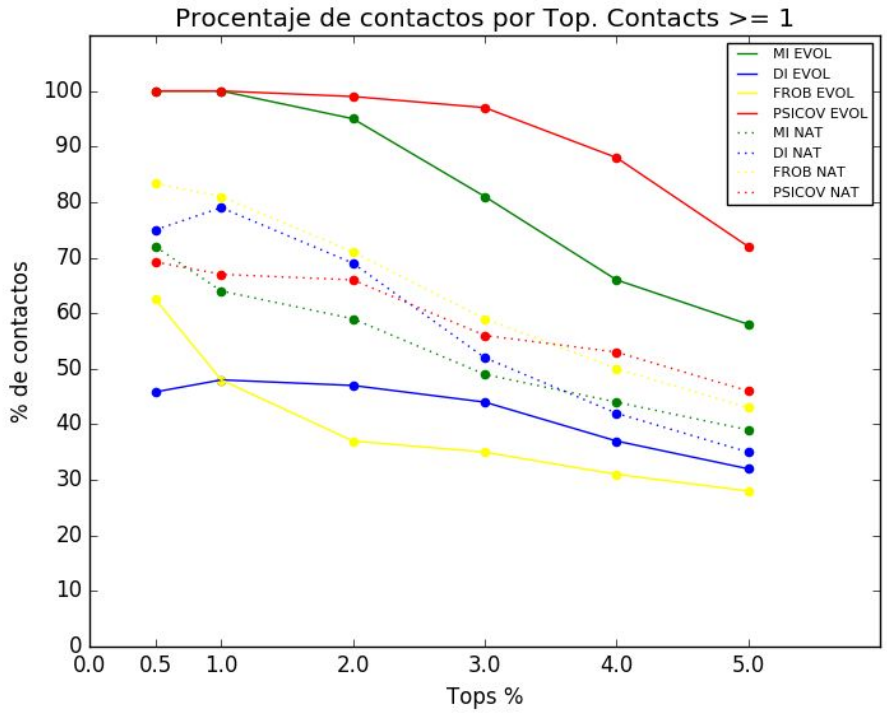


Figura 6.5.1 Comparación de los tops de la sumatoria de los métodos de coevolución MI, DI, FROB y PSICOV de la información evolucionada y de la información natural. La información se refiere a todos los contactos de la matriz conjunta contacts >= 1.

El método MI también observa una mejor predicción analizando los contactos de forma conjunta; en el top 3 % del análisis de la estructura 2TRX daba una predicción del 59 % y en el análisis conjunto del 81 %. Esta diferencia puede observarse a partir del top 2% ya que para el top 0.5 y 1 % la predicción de contactos de PSICOV y MI es perfecta tanto para el análisis de la información de los confórmeros de forma conjunta como para el análisis teniendo en cuenta únicamente la estructura 2TRX.

Las Figuras que veremos a continuación se centran en realizar el análisis teniendo en cuenta cierto grado de conservación, para la predicción de contactos medianamente conservados (contact >= 4) se puede visualizar la Figura 6.5.2; y para el análisis de la predicción de los contactos conservados (contacts = 8) la Figura 6.5.3.

Para los contactos semi conservados, puede observarse que el método PSICOV obtiene su mejor resultado en el top 3 %, siendo el mismo el mejor para la predicción seguido por el de MI y muy por debajos los demás métodos.

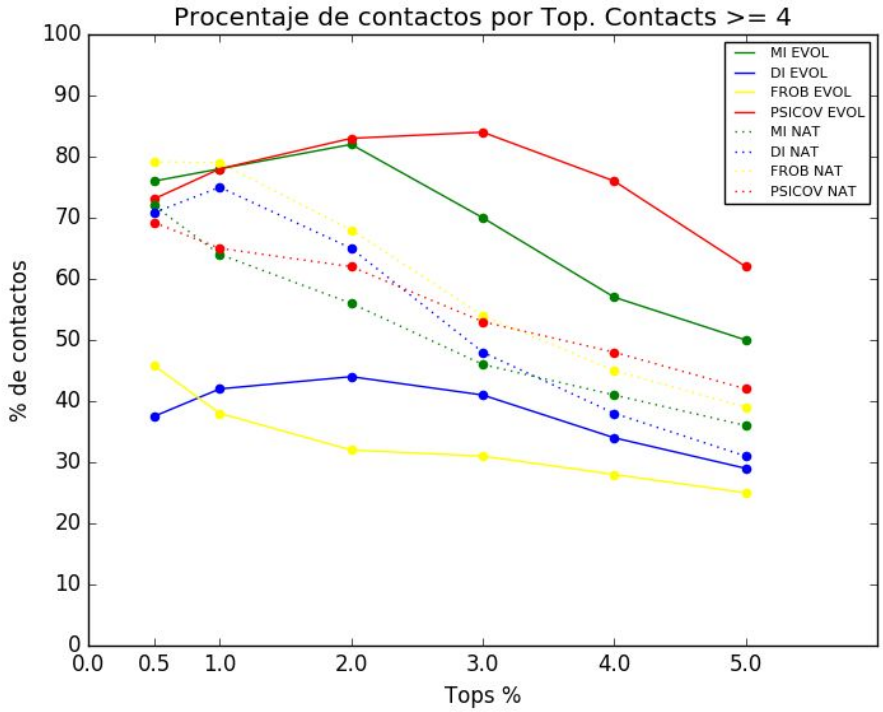


Figura 6.5.2 Comparación de los tops de la sumatoria de los métodos de covariación MI, DI, FROB y PSICOV de la información evolucionada y de la información natural. La información se refiere a todos los contactos semi conservados de la matriz conjunta, contacts >= 4.

Sobre los resultados de la información natural se destaca, que tanto para todos los contactos como para la matriz de contactos conjunta semi conservados, su desempeño se encuentra por debajo de la información evolucionada en los métodos de PSICOV y MI; y por encima de la información evolucionada para los métodos FROB y DI. En otras palabras, los métodos PSICOV y MI obtienen mejores resultados con la información evolucionada por sobre la natural; y en cambio los métodos de FROB y DI obtienen mejores resultados con la información natural.

El método FROB obtiene mejores resultados en cuanto a la información natural, con valores de 83.37, 81 y 71 para los tops 0.5, 1 y 2 % respectivamente

para todos los contactos de la matriz conjunta. En el Capítulo 5, en donde solamente se tenía en cuenta los contactos de la estructura 2TRX, el mismo método obtiene los mejores resultados en comparación con los demás, con 70.83, 73 y 64 % de contactos para los tops 0.5, 1 y 2 respectivamente. Esto era de esperarse ya que si agregamos la información estructural de más confórmeros al análisis, la predicción de contactos con el MSA natural va a ser más alta porque en el alineamiento natural se encuentran codificadas las señales de más estructuras, no solo de la 2TRX.

Por último, el análisis de la matriz de contactos conservada (contacts=8), visualizado en la Figura 6.5.3, se observan mejores resultados para la información natural por sobre la evolucionada para los tops 0.5 y 1 % para todos los métodos de covariación. El método PSICOV, nuevamente en el top 3 %, da la mejor predicción para los MSA evolucionados con un 56% de contactos bien predichos.

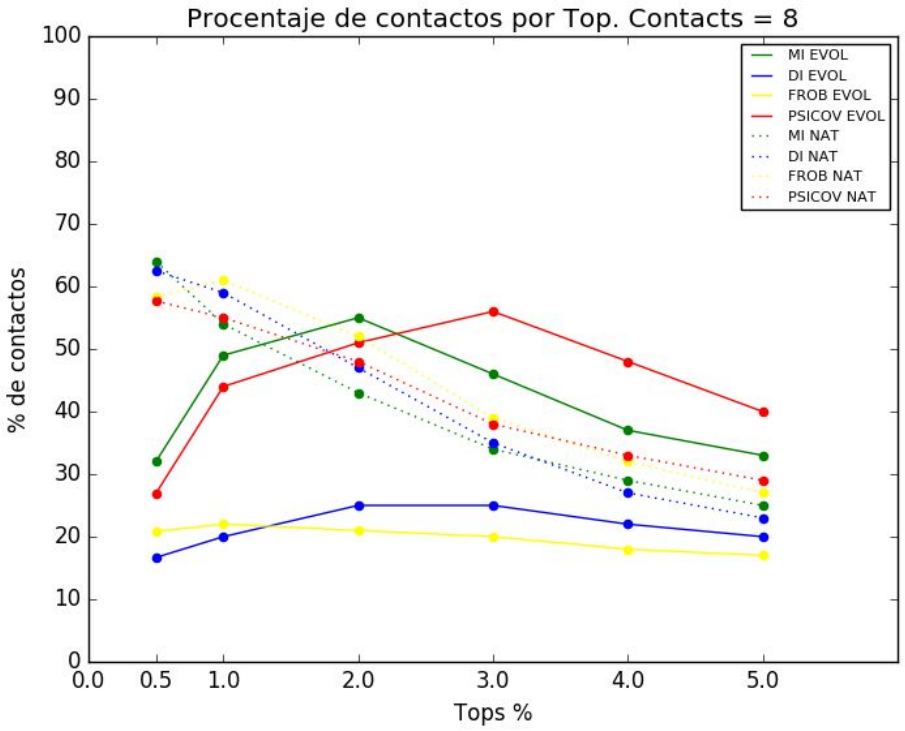


Figura 6.5.3 Comparación de los tops de la sumatoria de los métodos de covariación MI, DI, FROB y PSICOV de la información evolucionada y de la información natural. La información se refiere a todos los contactos conservados de la matriz conjunta, contacts = 8.

6.6. Análisis de la conservación en la evolución de los confórmeros

En la sección 5.4 se realizó el análisis sobre la conservación entre el MSA natural y el del MSA simulado del confórmero 2TRX, en dichos logos de secuencia se observan las similitudes; tanto en el natural como en el evolucionado se conservan los sitios activos; como las diferencias, en donde se destaca en el evolucionado la conservación de más posiciones que se encuentran relacionadas con las restricciones estructurales de la simulación. En esta sección extendemos el análisis y nos enfocamos en comparar la conservación de los MSAs de los confórmeros simulados. En la Figura 6.6.1 podemos apreciar la conservación media de los MSAs simulados de los 8 confórmeros, en color rojo, en comparación con la conservación del MSA natural, en color azul.

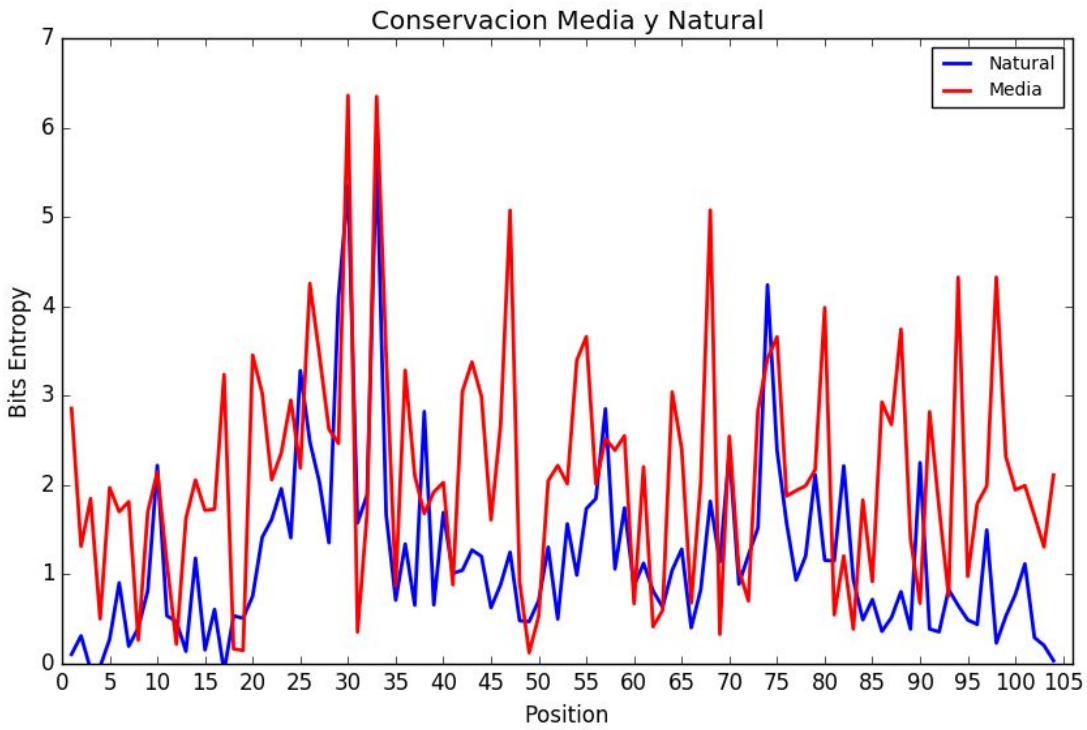


Figura 6.6.1. En rojo la conservación media por columna de los confórmeros analizados. En azul la conservación del MSA natural.

Se advierte una diferencia sustancial de posiciones conservadas en la media de los MSAs simulados sobre el MSA natural, esto es debido a la naturaleza de la simulación, en la cual se mantienen las restricciones estructurales, generando asi la conservacion de posiciones que cumplan un rol en base a la estructura de los confórmeros.

Podemos preguntarnos si las posiciones que se visualizan como conservadas en la media de los MSAs simulados tienen el mismo aminoácido conservado, o varia entre los diferentes MSAs ? Para responder esta pregunta observemos la Figura 6.6.2 en donde se encuentran los logos de secuencias de los MSAs simulados de los 8 confórmeros, junto con el del natural que se encuentra en la última posición.

Al igual que en la sección 5.4 se observa que el sitio activo, posiciones 31 y 33 de Cisteínas, se mantienen conservadas en todos los MSAs simulados. La aparición de otros sitios conservados que tienen que ver con la naturaleza de la simulación; los casos más claros son la introducción de Alaninas (A), Lisinas (K), Tirosinas (Y) y Treoninas (T) es compartida por todos los MSAs simulados de todos los confórmeros, por ejemplo observamos que las K de las posiciones 88, 94 y 98 se encuentran conservadas en los MSAs de todos los confórmeros; otro caso es el de las Y de las posiciones 47 y 68, por nombrar algunos. Pero existen algunas posiciones en donde el aminoácido es conservado en varios de ellos pero no en todos los MSAs simulados, por ejemplo la K en la posición 1 se encuentra conservada en todos los MSAs simulados excepto en los MSAs simulados de los confórmeros 1XOB_M5 y 1XOB_M16; por lo cual es probable que en estos confórmeros la posición 1 no se encuentre enmarcada en una relación estructural con otra posición y por tal motivo no sea necesaria su conservación. No se visualiza, en ningún caso, una posición que en un MSA simulado de un confórmero se encuentra conservada con un aminoácido en particular y que luego en otro MSA simulado se encuentra conservado con otro aminoácido.

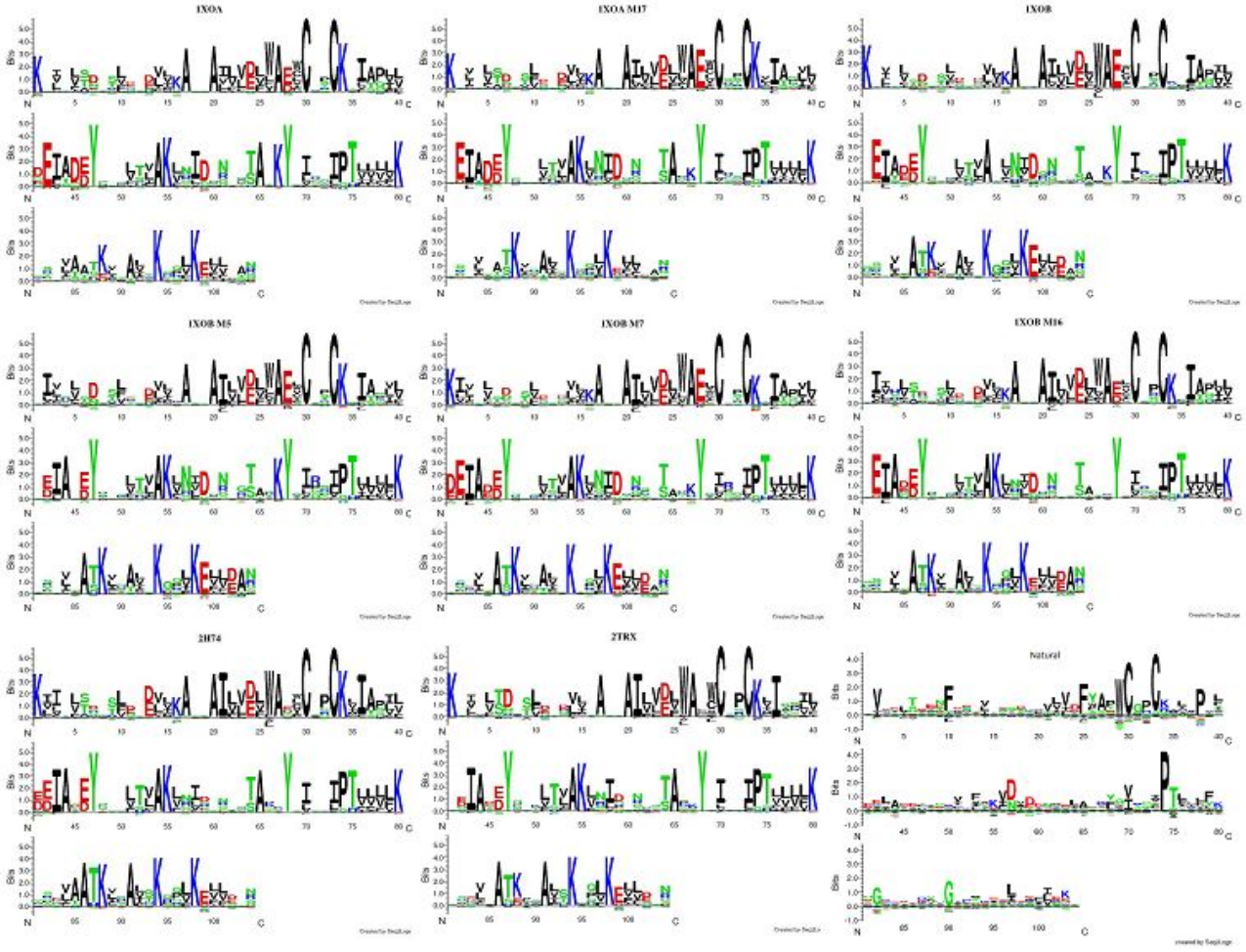


Figura 6.6.2. Panel con los logos de secuencia de los MSAs simulados de los confórmeros, el último logo corresponde al del MSA natural. Herramienta Seq2Logo. Algoritmo Kullbak-Leibler.

Se detectan algunas posiciones conservadas en el natural que no se conservan en ninguno de los MSAs simulados, probablemente, como ya lo hemos comentado en el capítulo 5, se deben a que son posiciones con información funcional y por tal motivo se pierde su conservación durante la simulación de evolución de los confórmeros.

6.7 Generación de MSA combinado

Hemos observado que analizar los resultados de la evolución de los confórmeros en forma conjunta, especialmente a través de los método PSICOV y

MI, permite obtener mejores resultados que en forma individual; encontrando porcentajes superiores de contactos y más similitudes con la información referida al MSA natural. Pero en la sección 6.5 evolucionamos los confórmeros generando MSAs artificiales a los que luego se les realizan los cálculos referidos a la covariación por separado para luego agruparlos. No se ha generado un MSA conjunto con la evolución de los 8 confórmeros; esta tarea se abordará en esta sección. El procedimiento es generar un MSA conjunto realizando diferentes *bootstraps* de 1500, 3000, 5000, 10000 y 15000 secuencias que se encuentran dentro de cada MSA simulado de cada confórmero. Al referirnos al término *bootstrap* estamos indicando que se tomarán, por ejemplo, 1500 secuencias al azar de cada uno de los MSA simulados; obteniendo así un MSA combinado con secuencias evolucionadas de cada confórmero.

Continuando con el procedimiento, los MSAs combinados son clusterizados al 62%, lo que resulta en una pequeña reducción en cada uno de ellos, ya que al partir de diferentes evoluciones y confórmeros no existe una redundancia manifiesta; luego aplicamos el procedimiento de predicción utilizando la matriz de contactos conjunta y nos disponemos a analizar los resultados.

Los resultados, descritos en la Tabla 6.7.1, no permiten indicar que este método para analizar la evolución de más estructuras muestre una mejora respecto al desempeño del método para la predicción de contactos en relación a los obtenidos en los capítulos anteriores; probablemente esto se deba a que en la evolución partiendo de los 8 confórmeros, las señales de coevolución de cada confórmero, dada por restricciones y contactos particulares de cada uno de ellos, se diluyen frente a las restricciones y contactos particulares de otro confórmero. Es decir la señal que pueda generar una estructura (confórmero) está presente en un subset de secuencias y otro subset tendrá otra señal involucrando seguramente otras posiciones. Sin embargo, las restricciones comunes a todas las estructuras (contactos conservados), están impresas en el MSA; para convalidar esta teoría podemos observar las Figuras 6.7.1 y 6.7.2, en la primera se

encuentran las curvas ROCs para el bootstrap 15000 analizando todos los contactos de la matriz; y en la segunda solamente los contactos semi-conservados y conservados.

bootstrap#	auc_ mi_01	auc_mi	auc_ di_01	auc_di	auc_ frob_01	auc_frob	auc_psicov_ 01	auc_ psicov
1500	0.591212	0.782223	0.599005	0.72394	0.597227	0.744305	0.647487	0.781598
3000	0.6155	0.79924	0.64339	0.763118	0.595167	0.772534	0.712708	0.830038
5000	0.626291	0.806177	0.670191	0.79101	0.604285	0.776078	0.748325	0.860518
10000	0.631966	0.805933	0.681828	0.800011	0.617016	0.789138	0.763905	0.869995
15000	0.633409	0.802572	0.696337	0.805977	0.615148	0.78652	0.771911	0.881575

Tabla 6.7.1. Resultados de AUCs para los diferentes bootstraps.

Se observa una mejora en el desempeño predictivo con respecto al análisis con todos los contactos; el método PSICOV, por ejemplo, obtiene un desempeño predictivo de AUC 0.88 para todos los contactos y de AUC 0.92 teniendo en cuenta solamente los contactos conservados.

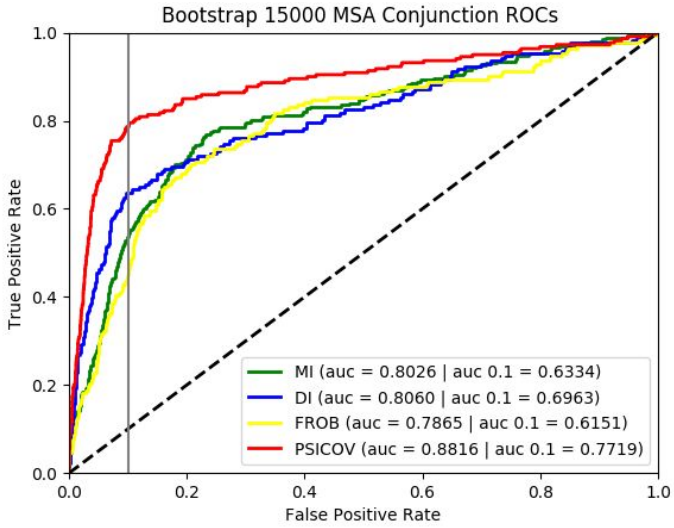


Figura 6.7.1. Curvas ROCs para los métodos de covariación con un bootstrap de 15000 teniendo en cuenta todos los contactos.

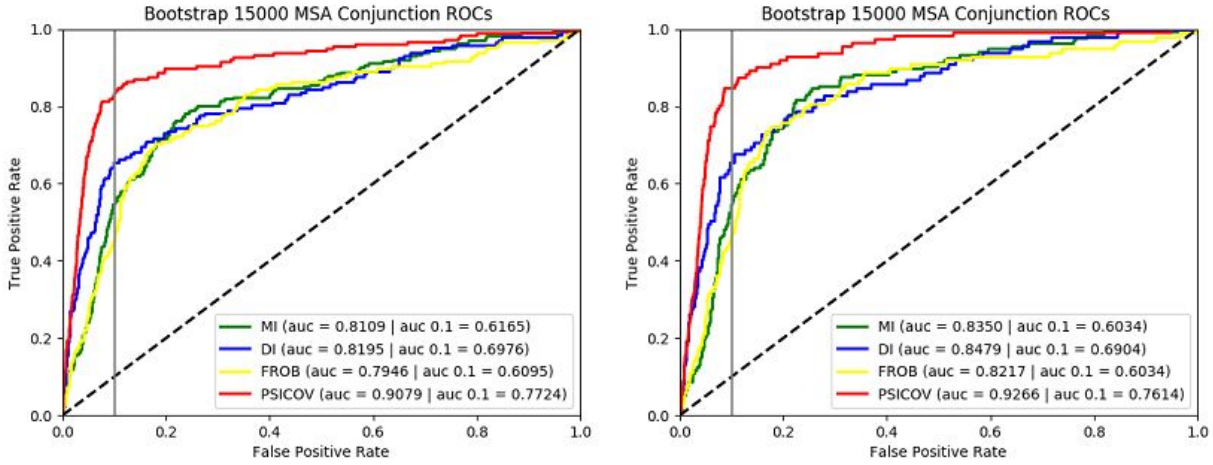


Figura 6.7.2. Curvas ROCs para los métodos de covariación teniendo en cuenta los contactos semi-conservados (izquierda) y los contactos conservados (derecha), es decir los contacts>=4 y contacts=8 de la matriz de contacto conjunta.

En la Tabla 6.7.2 se completa el análisis; en la misma se observan los diferentes valores de AUC a medida que se aumenta la restricción de conservación de contactos los van mejorando.

contact_ threashold	auc_mi_01	auc_mi	auc_di_01	auc_di	auc_ frob_01	auc_frob	auc_ psicov_01	auc_ psicov
1	0.63341	0.802573	0.696338	0.805977	0.615149	0.786521	0.771911	0.881576
2	0.629034	0.807255	0.700801	0.816707	0.613372	0.791941	0.779579	0.900158
3	0.619236	0.814281	0.699413	0.818973	0.606876	0.794823	0.778865	0.908641
4	0.616527	0.810872	0.697624	0.819516	0.609549	0.794568	0.772434	0.907857
5	0.617586	0.812199	0.700197	0.8219	0.610827	0.796703	0.777057	0.906187
6	0.61774	0.82942	0.711531	0.844361	0.61861	0.814833	0.785382	0.923467
7	0.620111	0.837159	0.707572	0.84492	0.621627	0.816303	0.781973	0.931045
8	0.6034	0.835014	0.690379	0.847914	0.603374	0.821651	0.761432	0.926593

Tabla 6.7.2. Valores de AUC para las diferentes restricciones de conservación de contactos. Los valores pertenecen al bootstrap 15000.

6.8. Conclusión

En primer lugar pudimos observar que los resultados obtenidos certifican que el procedimiento, junto con la optimización realizada sobre los parámetros de SCPE, sirven para diferentes estructuras de la misma proteína. La simulación de los diferentes confórmeros brinda resultados similares a los obtenidos en el capítulo 5; y se generaliza además, que el método PSICOV es el que mejor desempeño predictivo obtiene para los confórmeros analizados, seguido por el método MI y en tercer orden los métodos de DI y FROB. Es destacable la Figura 6.2.4 para visualizar claramente cómo a medida que avanzamos en el análisis de los tops, de menor a mayor, los pares de posiciones que tienen alto score en el método PSICOV son los mismos que los contactos del confórmero.

Sobre el análisis de la información de covariación conjunta de los confórmeros hemos analizado en primera instancia la correlación entre sumatoria de covariación de los diferentes tops con la sumatoria de la matriz de contactos conjunta; obteniendo como resultado que si consideramos el top 1%, comienza a hacerse evidente que los pares de posiciones que son contacto en 8 estructuras, aparecen en el top 1% más veces (x,y = 8,8), Figura 6.5.1. Sin embargo comienza a notarse que pares en contacto en todas las estructuras (y = 8) pueden salir en 1, 2, 3,...7 tops 1% (x = 1 a 7). Osea, que calcular covariación en diferentes confórmeros, revela diferentes pares en contacto. Al considerar el top 3% de los pares, aparecen muchos pares que no son contacto y que aparecen en uno de los tops 3%, estos son falsos positivos. Podemos concluir entonces que si nos quedamos con los pares que aparecen en los tops 3% de covariación de las n estructuras, estamos asegurando que esos son contactos conservados; y si desechamos los pares que aparecen en un sólo top 3%, estaríamos eliminando la mayor parte de todos los falsos positivos.

En la Figura 6.5.4 observamos cualitativamente para el método PSICOV (método que mejor resultado obtuvo durante todo el trabajo) una similitud entre la probabilidad de contactos y la señal conjunta de la sumatoria del top 3 %. Luego, los dendrogramas que acompañaron la matriz de probabilidad de contactos junto con la probabilidad del top 3%, nos permitieron analizar si los confórmeros que se encuentran cercanos a nivel estructural (comparten más contactos) también se encuentran cercanos a nivel de señal de covariación (comparten más pares de posiciones dentro del top 3 %); para lo cual se destacaron algunos ejemplos, entre ellos el de los confórmeros 2TRX-2H74, en donde se pudo verificar que hay una correspondencia de cercanía a nivel estructural y a nivel de señal en el top 3 % de covariación.

Al analizar el porcentaje de contactos encontrados de la simulación de los confórmeros, teniendo en cuenta todos los contactos, en forma conjunta pudimos obtener mejores resultados que de forma individual; esto es, más probabilidad de contactos encontrados y más similitudes con la información referida al MSA natural. Los resultados de la sumatoria de covariación son superiores a los obtenidos solamente teniendo en cuenta una única estructura, tomando como referencia la 2TRX. Por ejemplo, el método PSICOV obtuvo para el top 3 % un 79 % de contactos encontrados contra un 97 % cuando analizamos la información de forma conjunta. Esto era esperable ya que la 2TRX solamente fue evolucionada con sus propios contactos; en cambio el análisis conjunto contiene información estructural de todos los confórmeros analizados. Concluyendo entonces que simular la evolución de varios confórmeros de la misma proteína nos brinda información complementaria de importancia para la predicción de los contactos.

En cuanto al análisis de la conservación, se obtuvo la conservación media de los MSAs simulados de los confórmeros en donde se advirtió una diferencia sustancial de posiciones conservadas en la media de los MSAs simulados sobre el MSA natural, esto es debido a la naturaleza de la simulación, en la cual se

mantienen las restricciones estructurales, generando así la conservación de posiciones que cumplan un rol en base a la estructura de los confórmeros. Se pudo verificar, a través de los logos de secuencias, que los sitios conservados siempre tienen el mismo aminoácido entre los diferentes MSAs simulados. En algunos casos podría darse la situación de que el sitio no fuera conservado para algún MSA simulado de algún confórmero en particular, en esas situaciones la posición dentro de ese confórmero no estaría cumpliendo una función estructural por tal no es necesaria su conservación durante la simulación.

Pudo observarse en forma generalizada, que el sitio activo de la proteína, posiciones 31 y 33 de Cisteínas, se mantienen conservadas en todos los MSAs simulados. Nuevamente se detectaron algunas posiciones conservadas en el natural que no se conservan en ninguno de los MSAs simulados, probablemente, como lo hemos comentado, se deben a que son posiciones con información funcional y por tal motivo se pierde su conservación durante la simulación de evolución de los confórmeros.

La generación de un MSA combinado a partir de secuencias que se encontraban en MSAs simulados de los confórmeros tuvo como resultado un desempeño predictivo más bajo teniendo en cuenta todos los contactos de la matriz conjunta de los que se venían obteniendo en secciones anteriores. Esto se debe a que en la simulación partiendo de los 8 confórmeros, las señales de coevolución de cada confórmero, dada por restricciones y contactos particulares de cada uno de ellos, se diluyen frente a las restricciones y contactos particulares de otro confórmero. En cambio, las restricciones comunes a todas las estructuras (contactos conservados), dejan una señal más fuerte en el MSA. Por lo cual, cuando realizamos el análisis del desempeño predictivo teniendo en cuenta la conservación de los contactos, notamos que los valores de AUC van obteniendo una mejora a medida que incrementamos el corte de conservación de la matriz conjunta.

Al concluir esta tesis, se abren innumerables caminos para continuar con la investigación, entre los que se destacan; profundizar el análisis de los contactos compartidos con el MSA natural; realizar mejoras al SCPE, por ejemplo, parametrizar la definición de contacto a utilizar en la evolución, o agregar características de evolución que hagan hincapié en diferentes tipos enlaces; evolucionar toda una familia de proteínas para generalizar aún más los resultados; y desarrollar un nuevo método de evolución que permita ingresar más de una estructura para poder generar MSAs con información mas información estructural. Así mismo, se abren posibilidades de testear nuevas hipótesis biológicas como: podremos separar la señal estructural de la funcional que hace que un grupo de residuos evolucionen en forma concertada? eso serviría para predecir la función de proteínas para las que no se conoce. Otra pregunta interesante seria, cuantos contactos son necesarios para mantener la estructura, alcanza solamente con los conservados?.

El trabajo desarrollado para esta tesis es un comienzo, se continuará con la investigación, analizando nuevos caminos que vayan surgiendo e intentando contestar preguntas relevantes para el campo de la evolución de las proteínas y su estructura conformacional.

Bibliografía

[1] W.R. Atchley, K.R. Wollenberg, W.M. Fitch, W. Terhalle, and A.W. Dress, "Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis." Molecular Biology and Evolution, vol. 17, no. 1, pp. 164-178, jan 2000.

[2] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic acids research, vol. 25, no. 17, pp. 3389-3402, sep 1997.

[3] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, "Protein sectors: evolutionary units of three-dimensional structure." Cell, vol. 138, no. 4, pp. 774-786, aug 2009.

[4] D. Aguilar, B. Oliva, and C.M. Buslje, "Mapping the mutual information network of enzymatic families in the protein structure to unveil functional features." PloS one, vol. 7, no. 7, p. e41430, jan 2012.

[5] G.B. Gloor, L.C. Martin, L. M. Wahl, and S.D. Dunn, "Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions." Biochemistry, vol. 44, no. 19, pp. 7156-7165, may 2005.

[6] G. Parisi, and J. Echave, “Structural Constraints and Emergence of Sequence Patterns in Protein Evolution.” Molecular Biology and Evolution, vol. 18, no. 5, pp. 750–756, may 2001.

[7] R.D. Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A.L. Mitchell, S.C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman, "The Pfam protein families database: towards a more sustainable future." Nucleic Acids Research, vol. 44, no. D1, pp. D279-D285, jan 2016.

[8] D.J. Zea, D. Anfossi, M. Nielsen, and C.M. Buslje, “MIToS.jl: mutual information tools for protein sequence analysis in the Julia language.” Bioinformatics, vol. 33, no. 4, pp. 564–565, feb 2017.

[9] F.L. Simonetti, E. Teppa, A. Chernomoretz, M. Nielsen, and C.M. Buslje, "MISTIC: mutual information server to infer coevolution." Nucleic Acids Research, vol. 41, no. W1, pp. W8–W14, jul 2013.

[10] J. Söding, A. Biegert, and A.N. Lupas, “The HHpred interactive server for protein homology detection and structure prediction.” Nucleic Acids Research, vol. 33, no. suppl_2, pp. W244-W248, jul 2005.

[11] M.C.F. Thomsen, and M. Nielsen, "Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion." Nucleic acids research, vol. 40, no. Web Server issue, pp. W281-287, jul 2012.

[12] T.D. Schneider, and R.M. Stephens, “Sequence logos: a new way to display consensus sequences.” Nucleic Acids Res, vol. 18, no. 20, pp. 6097–6100, oct 1990.

[13] C. Yanofsky, V. Horn, and D. Thorpe, "Protein structure relationships revealed by mutational analysis." Science, vol. 146, no. 3651, pp. 1593-1594, dec 1964.

[14] W.M. Fitch, and E. Markowitz, "An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution." Biochemical genetics, vol. 4, no. 5, pp. 579-93, oct 1970.

[15] F.M. Codoñer, and M.A. Fares, "Why should we care about molecular coevolution?" Evolutionary bioinformatics online, vol. 4, pp. 29-38, 2008.

[16] F. Pazos, and A. Valencia, "Protein co-evolution, co-adaptation and interactions." The EMBO journal, vol. 27, no. 20, pp. 2648-55, oct 2008.

[17] E.R. Tillier, and T.W. Lui, "Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments." Bioinformatics, vol. 19, no. 6, pp. 750-755, apr 2003.

[18] L.C. Martin, G.B. Gloor, S.D. Dunn, and L.M. Wahl, "Using information theory to search for co-evolving residues in proteins." Bioinformatics (Oxford, England), vol. 21, no. 22, pp. 4116-4124, nov 2005.

[19] R.E. Blahut, “Principles and practice of information theory.” Addison-Wesley, jan 1987.

[20] A.A. Fodor, and R.W. Aldrich, "Influence of conservation on calculations of amino acid covariance in multiple sequence alignments." Proteins, vol. 56, no. 2, pp. 211-21, aug 2004.

[21] C.M. Buslje, J. Santos, J.M. Delfino, and M. Nielsen, "Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information." Bioinformatics (Oxford, England), vol. 25, no. 9, pp. 1125-1131, may 2009.

[22] J.Y. Dutheil, "Detecting coevolving positions in a molecule: why and how to account for phylogeny." Briefings in bioinformatics, vol. 13, no. 2, pp. 228-243, mar 2012.

[23] S.D. Dunn, L.M. Wahl, and G.B. Gloor, "Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction." Bioinformatics (Oxford, England), vol. 24, no. 3, pp. 333-340, feb 2008.

[24] R. Gouveia-Oliveira, and A.G. Pedersen, "Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation." Algorithms for molecular biology, vol. 2, p. 12, jan 2007.

[25] A. Del Sol, M.J. Araúzo-Bravo, D. Amoros, and R. Nussinov, "Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages." Genome biology, vol. 8, no. 5, p. R92, 2007.

[26] W. Li, L. Jaroszewski, and A. Godzik, “Clustering of highly homologous sequences to reduce the size of large protein databases.” Bioinformatics, vol. 17, no. 3, pp. 282–283, mar 2001.

[27] E.T. Jaynes, “Information Theory and Statistical Mechanics.” Physical Review Series II, vol. 106, no. 4, pp. 620-630, mar 1957.

[28] E.T. Jaynes, “Information Theory and Statistical Mechanics II.” Physical Review Series II, vol. 108, no. 2, pp. 171-190, oct 1957.

[29] M. Weigt, R.A. White, H. Szurmant, J.A. Hoch, and T. Hwa, “Identification of direct residue contacts in protein-protein interaction by message passing.” Proceedings of the National Academy of Sciences, vol. 106, no. 1, pp. 67–72, jan 2009.

[30] T. Mora, A.M. Walczak, W. Bialek, and C.G. Callan, “Maximum entropy models for antibody diversity.” Proceedings of the National Academy of Sciences, vol. 107, no. 12, pp. 5405–5410, mar 2010.

[31] D.T. Jones, D.W.A. Buchan, D. Cozzetto, and M. Pontil, “PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments.” Bioinformatics, vol. 28, no. 2, pp. 184-190, jan 2012.

[32] C. Baldassi, M. Zamparo, C. Feinauer, A. Procaccini, R. Zecchina, and M. Weigt, “Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners.” PLoS one, vol. 9, no. 3, pp. e92721, mar 2014.

[33] O. Banerjee, L.E. Ghaoui, and A. d’Aspremont, “Model selection through sparse maximum likelihood estimation for Multivariate Gaussian or Binary Data. “ Journal Machine Learning Research, vol. 9, pp. 485-516, mar 2008.

[34] J. Friedman, T. Hastie, and R. Tibshirani, “Sparse inverse covariance estimation with the graphical Lasso.” Biostatistics (Oxford, England), vol. 9, no. 9, pp. 432-441, jul 2008.

[35] N. Eswar, D. Eramian, B. Webb, M.Y. Shen, and A. Sali, “Protein Structure Modeling with MODELLER.” Methods in Molecular Biology, vol. 426, pp. 145-59, 2008.

[36] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin, ”UCSF Chimera--a visualization system for exploratory research and analysis.” J Comput Chem, vol. 25, no. 13, pp. 1605-12, oct 2004.

[37] A.M. Monzon, C.O. Rohr, M.S. Fornasari, and G. Parisi, “CoDNaS 2.0: A comprehensive database of protein conformational diversity in the native state.” Database (Oxford Journals), vol. 2016, no. baw038, jan 2016.