UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

IDENTIFICACIÓN DE RESIDUOS SENSIBLES A MUTACIONES MEDIANTE ANÁLISIS DE REDES ESTRUCTURALES DE PROTEÍNAS

TESIS

QUE PARA OBTENER EL TÍTULO DE

LICENCIADO

EN

INVESTIGACIÓN BIOMÉDICA BÁSICA

PRESENTA

MAURICIO CRUZ LOYA

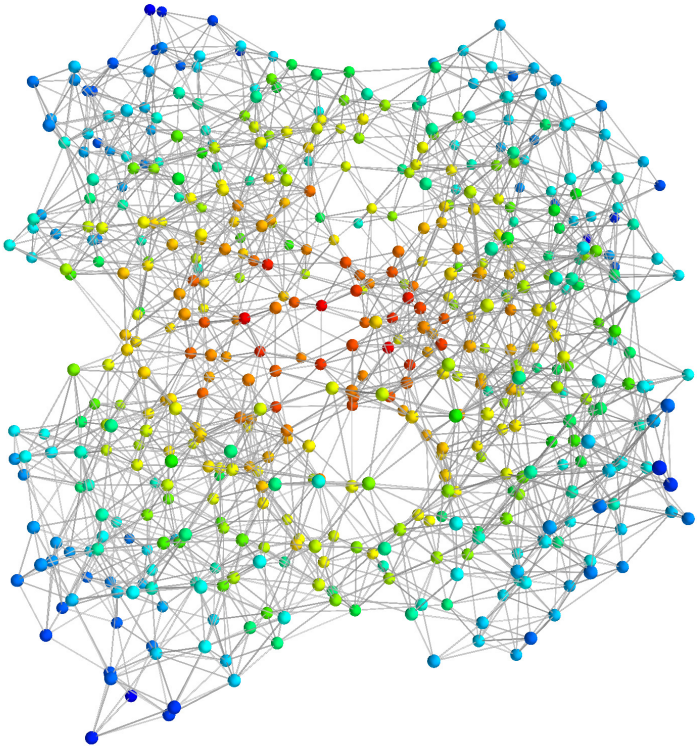
DIRECTOR DE TESIS

GABRIEL DEL RÍO GUERRA

CIUDAD UNIVERSITARIA, 2015

IDENTIFICACIÓN DE RESIDUOS SENSIBLES A MUTACIONES MEDIANTE ANÁLISIS DE REDES ESTRUCTURALES DE PROTEÍNAS

MAURICIO CRUZ LOYA



Explorando la relación entre la estructura y la función

Mauricio Cruz Loya: *Identificación de Residuos Sensibles a Mutaciones Mediante Análisis de Redes Estructurales de Proteínas*, Explorando la relación entre la estructura y la función, © Junio 2015

Dedicada a Roberto Cruz Ponce.
1958 – 2005

RESUMEN

Se ajustaron modelos de regresión logística regularizados para predecir los *residuos críticos* de una proteína (es decir, los más sensibles a mutaciones que eliminen la función de la proteína) a partir de su estructura tridimensional. Se usaron distintas combinaciones de atributos estructurales tradicionales (e.g. hidrofobicidad, estructura secundaria, área accesible al solvente) y atributos calculados a través de la red de contactos (entre residuos) de la proteína como variables predictivas. El mejor modelo ajustado tiene una precisión de 77.9 % y un coeficiente de correlación de Matthews de 0.42 (calculados por validación cruzada). Esto mejora sustancialmente la capacidad predictiva de `JAMMING`, un programa que fue desarrollado previamente en el laboratorio, y está basado exclusivamente en la red de contactos.

ABSTRACT

Regularized logistic regression models were fit to predict the *critical residues* of a protein (i.e. the most sensitive to mutations that eliminate protein function). Different combinations of traditional structural features (such as hydrophobicity, secondary structure and accesible solvent area) and features that depend on the contact network of the protein (between residues) were used to fit the models. The best model has a cross-validated accuracy of 77.9 % and Matthews correlation coefficient of 0.42. This substantially improves the predictive capability of the program `JAMMING`, which was previously developed in the lab, and is based exclusively in the contact network.

AGRADECIMIENTOS

En primer lugar quisiera agradecer a mi familia. A mi padre por ayudar a despertar en mí la curiosidad por el mundo que eventualmente me llevó a estudiar una carrera científica, a mi madre por sus consejos y su paciencia aún cuando parecía que nunca iba a terminar la tesis y a mi abuela por su apoyo incondicional.

Agradezco a la Universidad Nacional Autónoma de México, y a los mexicanos, la oportunidad que me proporcionaron de estudiar una carrera de alta calidad de manera prácticamente gratuita.

También quisiera agradecer a los maestros que tuve durante la licenciatura por enseñarme, además de la materia correspondiente, diferentes maneras de abordar la ciencia. Particularmente, me gustaría agradecer al Dr. Daniel Alejandro Fernández Velazco y al Dr. Alejandro Zentella Dehesa, maestros y personas excelentes de los que he aprendido muchas cosas y que han influenciado muchísimo mi manera de ver la ciencia.

La ciencia es una actividad práctica. No basta con adquirir conocimiento, sino es necesario saber aplicarlo para resolver problemas científicos reales. Por ello también agradezco a los tutores que he tenido a lo largo de la Licenciatura. He tenido el privilegio de rotar en los laboratorios de la Dra. Carolina Escobar Briones, el Dr. Luis Vaca Domínguez y el Dr. Gabriel del Río Guerra.

En los laboratorios en dónde he rotado también he tenido buenos compañeros de trabajo, de los que he aprendido muchas cosas. Esta es una lista larga, pero quisiera en particular agradecer a Katia Rodríguez y Jonathan Pacheco, y a la Dra. María Teresa Lara Ortiz por su apoyo técnico.

Agradezco también al Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT, proyecto número IN205911) que hizo posible la realización de este trabajo.

Finalmente, quisiera agradecer a mis compañeros de carrera. Como es costumbre en la LIBB, fuimos una generación relativamente pequeña y convivimos mucho durante los años que duró la carrera. Agradezco el ambiente que se fomentó entre nosotros, más cooperativo que competitivo, y haber tenido la oportunidad de conocer a personas tan excepcionales.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Parte I

INTRODUCCIÓN Y PLANTEAMIENTO DEL PROBLEMA

1

INTRODUCCIÓN

El trabajo descrito en esta tesis tiene como objetivo desarrollar y evaluar un método para identificar los residuos sensibles a mutaciones de una proteína, basado en información estructural. Para ello se usaron metodologías de dos áreas: la teoría de redes y el aprendizaje automático.

La organización de la introducción es la siguiente: inicio con un breve repaso de estructura de proteínas, y después incluyo una introducción a la teoría de redes y a la manera en que se utiliza para estudiar proteínas. Concluyo el capítulo con una introducción al aprendizaje automático, haciendo énfasis en el aprendizaje supervisado.

1.1 ESTRUCTURA DE PROTEÍNAS

Una proteína es una macromolécula que está constituída por una secuencia ordenada de aminoácidos unidos entre sí con enlaces peptídicos. A la secuencia de aminoácidos se le denomina la *estructura primaria* de la proteína. Como en la reacción química que forma un enlace peptídico se pierde una molécula de agua (de un grupo hidroxilo y un hidrógeno), a los aminoácidos que forman parte de una proteína se les denomina *residuos*.

Los residuos de una proteína tienen propiedades fisicoquímicas distintas derivadas de las cadenas laterales que los definen. Por los grupos químicos que tienen sus cadenas laterales, algunos residuos son hidrofílicos y otros hidrofóbicos; algunos están cargados eléctrostáticamente y otros no;

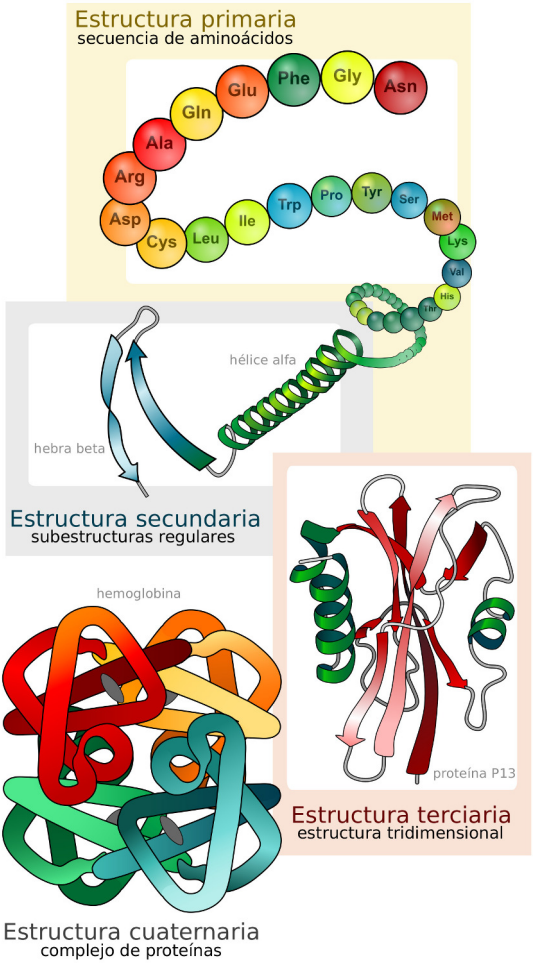


Figura 1.1: ESTRUCTURA DE UNA PROTEÍNA Se ilustran la estructura primaria, secundaria, terciaria y cuaternaria de una proteína. Adaptada de una imágen de dominio público creada por Mariana Ruiz Villarreal.

algunos tienen ciertos grupos funcionales y otros no. Por esta razón algunas conformaciones espaciales de los residuos son más estables que otras, pues los aminoácidos pueden arreglarse en el espacio de formas que son energéticamente (y/o entrópicamente) favorables o no.

Algunas partes contiguas de la estructura primaria de una proteína adquieren conformaciones espaciales características relativamente simples. A estas conformaciones se le denomina *estructura secundaria*. Las estructuras secundarias más comunes son las *hélices alfa* y las *hebras beta* (ver Figura 1.1). La diferencia en la estabilidad de las distintas conformaciones posibles de una proteína es tal que las proteínas tienden a *plegarse* y adquirir una estructura tridimensional característica, que corresponde a la más estable energéticamente (al menos de las que son cinéticamente accesibles en un ambiente dado). A ésta se le denomina la *estructura terciaria* de la proteína. La estructura terciaria no es estática: es modificada por las fluctuaciones térmicas, pero el ensamble estructural consiste de estructuras cercanas a la más estable. Algunas proteínas forman *complejos*, en los que varias cadenas protéicas distintas se asocian entre sí para dar lugar a la forma biológicamente activa. Este es el último nivel de estructura en las proteínas, al que se le denomina *estructura cuaternaria*.

1.2 TEORÍA DE REDES

La teoría de redes es una rama de las matemáticas que ha sido muy exitosa para modelar sistemas físicos, biológicos y de ciencias sociales.[1] Puede usarse para modelar cual-

1 La sección de teoría de redes está basada principalmente en Newman [28].

quier sistema que esté formado por partes que interactúan entre sí. A las partes del sistema se les representa como *nodos* o *vértices*, y a las interacciones entre ellos como *aristas*. Comunmente se representa a los nodos de una red como puntos, y a las *aristas* como líneas que unen a los puntos entre sí.

Por ejemplo, se puede representar a un grupo social como una red. Los individuos que forman parte del grupo social se representan como nodos, y las relaciones entre los individuos del grupo de algún tipo (por ejemplo, amistad) como aristas entre los nodos. También se pueden representar mecanismos de regulación genética como una red. En este caso, los nodos representarían a los genes y las aristas a las interacciones entre ellos (pueden existir varios tipos de aristas, algunas representando activación y otras inhibición).

Con el enfoque de la teoría de redes, el énfasis no es en los componentes del sistema en sí ni en la naturaleza de sus interacciones. Lo importante son los *patrones* que existen en las conexiones del sistema y qué cosas podemos inferir sobre el sistema a partir de ellos. Por ejemplo, nos puede interesar cómo es el flujo de información en un grupo social o encontrar los genes más importantes para la regulación de un sistema biológico.

1.2.1 *La red de contactos de una proteína*

Una proteína también puede ser representada como una red. Esto puede hacerse a través de la *red de contactos*, una representación en la que los residuos de la proteína se representan por nodos. Se considera que dos nodos interactúan si al menos un par de átomos de los residuos correspondien-

tes se encuentra a una distancia menor al *radio de contacto* R_c en la estructura tridimensional de la proteína.

Las redes de contacto de proteínas han sido usadas para estudiar problemas de plegamiento (Vendruscolo et al. [38]), de transmisión de la información (del Sol et al. [10]) y de predicción de residuos funcionales (Amitai et al. [1], Cusack et al. [5]), entre otros. En este trabajo las usamos para obtener atributos útiles para predecir residuos críticos, que incluyen a los residuos funcionales. En particular, calculamos *medidas de centralidad* de los nodos de la red (que corresponden a los residuos de la proteína).

1.2.2 *Medidas de centralidad*

Una pregunta que comunmente surge sobre un sistema que puede ser representado por una red es: ¿qué nodos son los más importantes? En una red social, tal vez nos interese identificar a los individuos más influyentes dentro del grupo social. En una red genética nos puede interesar predecir qué genes son indispensables para la supervivencia celular. En este trabajo, nos interesa identificar los residuos de una proteína que frecuentemente afectan la función de la misma al ser sustituídos por otros.

Una forma de abordar este tipo de preguntas es medir la *centralidad* (o importancia) de un nodo en la red. Existen distintas definiciones de importancia. Por ejemplo, en un grupo social los individuos pueden ser importantes porque tienen muchas conexiones, porque controlan el flujo de información o porque pueden diseminar un mensaje más rápidamente que el resto. Por esta razón existen varias medidas de centralidad, que corresponden a distintas definiciones de importancia. Las medidas de centralidad que utiliza-

mos en este trabajo se detallan en la sección de Materiales y Métodos.

1.3 APRENDIZAJE AUTOMÁTICO

El *aprendizaje automático* es una rama de las ciencias de la computación y de la estadística que consiste de una serie de métodos y algoritmos de análisis de datos para descubrir patrones en datos y usar estos patrones para hacer predicciones.

Existen dos tipos principales de aprendizaje automático: *supervisado* y *no supervisado*. En el aprendizaje supervisado, el objetivo es aprender a predecir una *variable de respuesta* a partir de un conjunto de *atributos*. En el aprendizaje no supervisado es aprender la "estructura interna" de los datos para encontrar una mejor representación de los mismos.

En este trabajo únicamente usaremos técnicas de aprendizaje automático supervisado, por lo que sólo profundizaremos en esta rama del aprendizaje automático.[2]

1.3.1 *Aprendizaje supervisado*

En los problemas de aprendizaje supervisado, el objetivo es deducir una función a partir de un conjunto de datos. En paticular, a partir de un *vector de atributos*

$$\mathbf{x} = (x_1, x_2, ..., x_N)$$

queremos predecir el valor de una *variable de respuesta* y. Para ello suponemos que existe una función desconocida f tal que

$$f(\mathbf{x}) = y$$

2 El material en esta sección está basado en Mostafa [27].

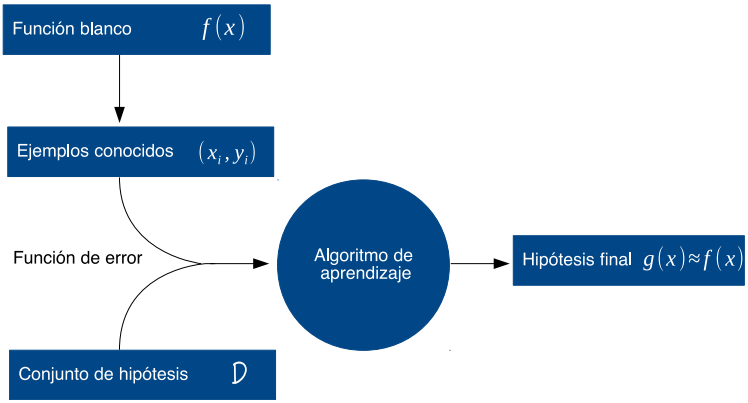


Figura 1.2: Diagrama simplificado del aprendizaje supervisado.

Idealmente quisiéramos conocer esta función para poder predecir la variable de respuesta para cualquier valor de **x**. Por eso a f se le llama *función blanco*.

Dependiendo de las características de la variable de respuesta, los problemas de aprendizaje supervisado pueden dividirse en dos grupos:

A. En los problemas de *regresión*, la variable de respuesta es continua ($y \in \mathbb{R}$).

B. En los problemas de *clasificación*, la variable de respuesta es discreta ($y \in \{0, 1, ..., C-1\}$, en dónde C es el número de *clases*). Al caso en el que $C = 2$ se le llama *clasificación binaria.*

La función blanco es completamente desconocida: sólo tenemos acceso a un conjunto de datos generados por ella. Más aún, se puede mostrar matemáticamente que es *imposible* obtener la función blanco de manera exacta (Mostafa [27]). Sin embargo, podemos buscar una función distinta que *aproxime* f en un *conjunto de hipótesis* $\mathcal{D}$ (que puede ser,

por ejemplo, el conjunto de modelos lineales). Esto pudiera parecer paradójico: si es imposible conocer la función blanco, ¿cómo podemos saber qué tanto se parece otra función a ella? Para hacerlo parece ser necesario saber cómo se comporta la función blanco en todo su dominio, lo cual es imposible si es desconocida.

La respuesta es que, aunque no podemos conocer nada sobre la función blanco de manera determinista, sí lo podemos hacer de manera probabilística. Bajo el supuesto de que el conjunto de datos que se usa para ajustar el modelo es una muestra independiente y aleatoria de todos los datos posibles de ese tipo, se puede calcular un límite probabilístico de la diferencia entre la proporción de datos correctamente clasificados por el modelo en la muestra (el conjunto de datos conocido) y en la población (todos los datos posibles). De esta forma se puede poner un límite probabilístico en el error de clasificación fuera del conjunto de datos usado (que llamaremos E_{out}) con base en el error dentro de él (E_{in}) y la complejidad de los modelos en el conjunto de hipótesis $\mathcal{D}$ (ver Mostafa [27] para mayores detalles). Entre menor sea E_{out}, más se acercará la función ajustada a la función blanco. Por eso podemos decir que la función ajustada *aproxima* a f.

Dentro de todas las $h \in \mathcal{D}$ buscamos una función g tal que $g(\mathbf{x}) \approx f(\mathbf{x})$. A g le llamamos *hipótesis final*, y, si la encontramos, se cumple que

$$g(\mathbf{x}) \approx f(\mathbf{x}) = y$$

y entonces

$$g(\mathbf{x}) \approx y$$

Para escoger la hipótesis final de entre todas las $h \in \mathcal{D}$ nos basamos en ejemplos conocidos: usamos un conjunto

de datos para los que se conozca tanto $\mathbf{x}$ como y. Al proceso de elegir esta hipótesis se le llama *entrenamiento* o *ajuste* del modelo y al conjunto de ejemplos que usamos para ello se le llama *conjunto de entrenamiento*. El objetivo es usar la hipótesis final para predecir los valores que toma la variable de respuesta en otros datos, para los que sólo conozcamos el vector de atributos $\mathbf{x}$.

Para ajustar el modelo es necesario tener un *algoritmo de aprendizaje*, basado en una *medida de error* $E(h(\mathbf{x}), f(\mathbf{x}))$ que cuantifica la calidad de la aproximación de cualquier hipótesis a la función blanco, según los datos del conjunto de entrenamiento. El rol del algoritmo de aprendizaje es encontrar la hipótesis de $\mathcal{D}$ que minimice este error.

El proceso de aprendizaje automático supervisado se resume en la Figura 1.2. La función blanco f es la función hipotética que nos gustaría obtener y que suponemos que generó los ejemplos conocidos. Buscamos una función en nuestro conjunto de hipótesis que aproxime el comportamiento de f en los datos. Esto se realiza a través de un algoritmo de aprendizaje. Al encontrar al mejor candidato dentro de nuestro conjunto de hipótesis, le llamamos hipótesis final (g). Una vez que obtenemos g, podemos usarla en nuevos datos para predecir y.

1.3.2 *Generalización y sobreajuste*

En los problemas de aprendizaje supervisado, el objetivo es tener el menor error posible al predecir la variable de respuesta. Desafortunadamente, el error de predicción dentro de los datos de entrenamiento (E_{in}) no siempre es representativo del error de predicción del modelo fuera de él (E_{out}).

Al ajustar el modelo a los datos de entrenamiento no sólo "se ajusta" a la función blanco, sino también al ruido presente en los datos de entrenamiento, que será distinto al ruido en datos nuevos. Aún en el caso de tener datos sin ruido, el número de ejemplos en el conjunto de entrenamiento puede no ser suficiente para ajustar correctamente la hipótesis final, o las funciones del conjunto de hipótesis pueden no poder reproducir correctamente algunos aspectos de la función blanco.

Por estas consideraciones, para saber el rendimiento real de la hipótesis final se debe medir el error de predicción *fuera del conjunto de datos de entrenamiento*. Por ello, frecuentemente se dividen los datos en dos grupos: un conjunto de *entrenamiento* y un conjunto de *prueba*. En este esquema, el modelo se ajusta únicamente a los datos de entrenamiento. Una vez que el modelo está completo, se obtiene el error en la predicción de los datos de prueba para conocer el rendimiento real del modelo.

Se puede definir el *error de generalización* E_{gen} como la discrepancia entre E_{in} y E_{out}. En la gran mayoría de los casos $E_{in} < E_{out}$, pues se está minimizando E_{in} explícitamente para ajustar el modelo. Entonces tenemos que

$$E_{gen} = E_{out} - E_{in}$$

Si E_{gen} es una cantidad grande, se dice que el modelo no se generaliza a datos fuera del conjunto de entrenamiento. A este fenómeno se le denomina *sobreajuste*, y lo ilustramos en la Figura 1.3. La función blanco normalmente es desconocida, pero en este ejemplo artificial se conoce: es un polinomio de orden 4 (línea roja). A partir de ella se generaron cinco puntos con ruido para usarse como ejemplos de entrenamiento. Como tienen ruido, los ejemplos no pasan exactamente por la línea roja (función blanco). Ajustamos

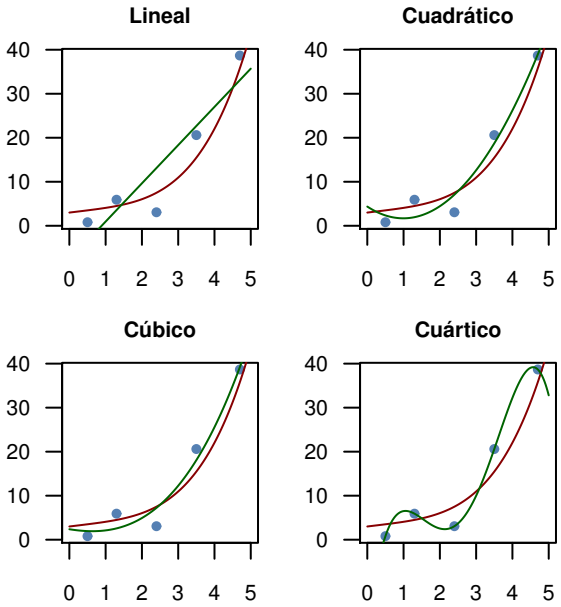


Figura 1.3: SOBREAJUSTE Se ajustaron polinomios a cinco puntos (azul) muestreados aleatoreamente de un polinomio de órden cuatro (rojo) con ruido normalmente distribuído. Se muestra el resultado de ajustar un modelo lineal, cuadrático, cúbico y cuártico (verde). El modelo cuártico está sobreajustando los datos, mientras que el lineal los está subajustando.

un modelo lineal, cuadrático, cúbico y cuártico (líneas verdes) a estos cinco ejemplos de entrenamiento con el fin de aproximar la función blanco.

Entre más se acerquen los modelos ajustados a la función blanco E_{out} será menor, pues los nuevos puntos serán generados a partir de la función blanco con ruido distinto al de los ejemplos de entrenamiento (en la figura, E_{out} para cada caso corresponde al área entre la línea roja y la línea verde). Observamos que, a pesar de que la función blanco es de orden 4, el mejor desempeño no se obtiene con el modelo cuártico, pues pasa exactamente por los puntos de entrenamiento y se aleja de la función blanco: los modelos cúbico y cuadrático tienen un menor E_{out} que el cuártico.

Para una combinación dada de una función blanco, un conjunto de hipótesis, un algoritmo de aprendizaje y una medida de error se puede graficar cómo varían E_{in} y E_{out} contra el número de ejemplos de entrenamiento N. A este gráfico se le denomina *curva de aprendizaje* (Figura 1.4). El error de generalización disminuye conforme aumenta el número de ejemplos de entrenamiento.

En términos simplificados, el error de generalización es menor para modelos más simples (por ejemplo, un modelo lineal en relación a uno cuadrático). Sin embargo, frecuentemente un modelo más complejo puede acercarse más a la función blanco que un modelo simple (esto equivaldría a que la asíntota denotada por la línea punteada en la Figura 1.4 disminuya). Por ello, el mejor modelo a usar (es decir, el que minimiza E_{out}) depende de la forma de la función blanco y del número de ejemplos de entrenamiento. El error de aproximar la función blanco de la Figura 1.3 con un modelo cuártico sería menor al del resto de los modelos si hubiera un mayor número de ejemplos de entrenamiento.

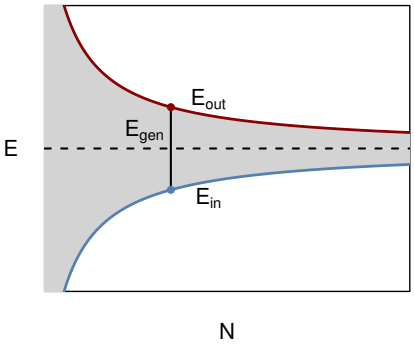


Figura 1.4: CURVA DE APRENDIZAJE Se ilustra una curva de aprendizaje típica. El error dentro del conjunto de entrenamiento (E_{in}) es casi siempre menor al error en datos nuevos (E_{out}). A la diferencia entre ambos se le denomina error de generalización (E_{gen}). Al aumentar el número de ejemplos de entrenamiento (N), disminuye el error de generalización. Adaptada de Mostafa [27].

1.3.3 *Regularización y validación*

Se han desarollado técnicas con el fin de combatir el problema del sobreajuste; las más usadas son la *regularización* y la *validación*.

1.3.3.1 *Regularización*

En la regularización se usa una medida de error en la que, además de penalizar el error de entrenamiento, se penaliza la complejidad del modelo usado. Se define entonces un *error aumentado*

$$E_{aug}(h, \lambda, \Omega) = E_{in} + \alpha\Omega(\mathrm{h})$$

El error aumentado depende de la complejidad del modelo Ω y del *parámetro de regularización* α. Este parámetro codifica qué tanto se va a penalizar a un modelo complejo en relación a uno sencillo.

Para el caso de los modelos lineales, en los que los parámetros son los coeficientes de las variables, se puede usar el tamaño de los coeficientes como una medida de la complejidad del modelo. Podemos definir el *error regularizado*

$$E_{reg} = E_{in} + \alpha \|\mathbf{w}\|$$

en dónde $\|\mathbf{w}\|$ es una *norma* de los coeficientes del modelo. En particular, frecuentemente se usan la norma euclideana (regularización l_2 o *ridge*) o la suma de los valores absolutos de los coeficientes (regularización l_1 o *lasso*). Se minimiza esta versión modificada del error en vez del error de entrenamiento.

Existe una manera de usar tanto la norma $l1$ como la norma $l2$ para penalizar el tamaño de los coeficientes. En los modelos de red elástica, se define

$$E_{reg} = E_{in} + \alpha\lambda \left\|\mathbf{w}\right\|_1 + \frac{1}{2}\alpha(1-\lambda) \left\|\mathbf{w}\right\|_2^2$$

en dónde $\alpha \geq 0$ determina la cantidad total de regularización a usar, y $\lambda \in [0,1]$ cómo se distribuye la regularización entre las normas $l1$ ($\left\|\mathbf{w}\right\|_1$) y $l2$ ($\left\|\mathbf{w}\right\|_2$).

1.3.3.2 *Validación*

La validación es un concepto muy parecido al de conjunto de prueba. De los N ejemplos de entrenamiento se separan K. A este conjunto de datos se le llama *conjunto de validación*. Se entrena al modelo en los $(N-K)$ datos restantes y se calcula el error en el conjunto de validación E_{val}. Este error sirve como un estimador de E_{out}, ya que el modelo no "ve" estos datos en el ajuste. Una vez que se tiene E_{val}, se entrena el modelo nuevamente con todos los N ejemplos.

Existe un conflicto al escoger el valor de K. Un valor grande tiene el efecto positivo de que E_{val} sea un mejor estimador de E_{out}, pues es calculado con más datos. Sin embargo, al escoger un valor grande de K perjudicamos al modelo, pues tiene menos datos de entrenamiento $(N-K)$. Esto ocasiona que E_{val}, a pesar de que es un muy buen estimador de E_{out} para el modelo entrenado con $N-K$ ejemplos, sea un mal estimador de E_{out} para el modelo entrenado con todos los N ejemplos.

Es decir, dependemos de la expresión

$$E_{out}(g) \approx E_{out}(g^-) \approx E_{val}(g^-)$$

en donde se denota con g la hipótesis final obtenida al entrenar con todos los N puntos y con g^- la hipótesis obtenida entrenando con $N-K$ puntos. Para la primera igualdad

aproximada es favorable tener una K pequeña, pues entonces el número de ejemplos con que se entrenan g y g^- es similar, pero para la segunda igualdad aproximada es favorable tener una K grande, pues el error de validación es un mejor estimador de E_{out} si se usan muchos datos para calcularlo.

1.3.3.3 *Validación cruzada*

La validación cruzada surgió como un intento de resolver este problema. Supongamos que, en el extremo, tomamos $K = 1$. Entonces tenemos que E_{val} es un mal estimador de $E_{out}(g^-)$, pero que $E_{out}(g^-)$ es un muy buen estimador de $E_{out}(g)$.

Sin embargo, ¿qué pasa si se repite este proceso, eligiendo cada vez un punto distinto para que sea el conjunto de validación, y se promedian los valores estimados de $E_{out}(g^-)$? El resultado es que, si esto se hace para cada punto del conjunto de entrenamiento, acabamos con un mejor estimador de $E_{out}(g^-)$. Los E_{val} son estimadores malos individualmente, pues se calculan con base en un sólo punto. Sin embargo, son estimadores no sesgados: al promediar muchos de ellos reducimos su varianza y acabamos con un estimador bueno. A este proceso se le llama LOOCV (*leave one out cross validation*).

En la práctica es frecuente que sea poco factible hacer esto, pues se necesitaría entrenar el modelo N veces (una por cada punto del conjunto de datos) y esto puede ser muy caro computacionalmente. Por ello, generalmente se usa otra variante de la validación cruzada: se divide el conjunto de datos en k partes, dejando cada parte como conjunto de validación en turno. El modelo se entrena k (en vez de N) veces, lo que requiere menos recursos computacionales. A

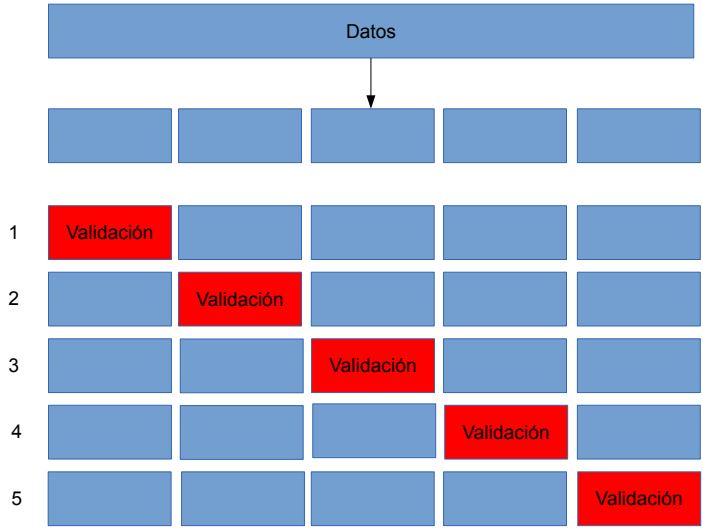


Figura 1.5: VALIDACIÓN CRUZADA k-FOLD Se ilustra un ejemplo de validación cruzada. Se separan los datos en $k = 5$ grupos. Cada grupo se usa en turno como conjunto de validación: se entrena el modelo con el resto de los datos y se calcula el error de validación E_{val}. Al final, se promedian todos los valores de E_{val} para obtener el error de validación cruzada E_{CV}, que es un estimador de E_{out}.

esto se le llama *validación cruzada k-fold* y lo ilustramos en la Figura 1.5.

2

EL PROBLEMA: PREDECIR RESIDUOS CRÍTICOS PARA LA FUNCIÓN

2.1 ANTECEDENTES

No todas las posiciones de una proteína producen un efecto similar al ser mutadas. Hay algunas que toleran un gran número de sustituciones sin que se afecte la función de la proteína. En cambio, hay otras que toleran muy pocas mutaciones (o incluso ninguna) sin que la función de la proteína se vea afectada. A estas posiciones se les ha denominado *residuos críticos para la función*, o simplemente *residuos críticos* (Loeb et al. [24], Wrobel et al. [39]).

En el grupo de los residuos críticos se incluyen *todos aquellos* que toleran pocas sustituciones sin que se afecte la función de la proteína en un ensayo experimental. Este un grupo heterogéneo de residuos, que incluye a los que son importantes por razones de estructura, de estabilidad y de catálisis. Como los residuos críticos son importantes para la función de la proteína, se podría esperar que estén muy conservados evolutivamente. Sin embargo se ha visto que, aunque esto sucede frecuentemente, no necesariamente es el caso. Existe un número importante de residuos que son críticos, pero no son conservados, y viceversa (Loeb et al. [24]).

La existencia de residuos críticos no conservados pudiera deberse en parte a mutantes compensatorias en un sitio distinto que reestablecen la función de una proteína no funcional. Esto pudiera suceder, por ejemplo, en el caso de

mutantes que afectan la estabilidad de la estructura de la proteína. Una mutante que desestabiliza la proteína puede contrarrestarse por otra que aumenta la estabilidad, por lo que no necesariamente estará conservada una posición en específico que estabiliza la proteína. Además, la existencia de funciones no reportadas en las proteínas dificulta evaluar la relación de conservación y función. Posiciones que están conservadas debido a su importancia para funciones no reportadas no serán detectadas como críticas.

La importancia del correcto plegamiento de una proteína para que ésta sea funcional y la mayor conservación de estructura que de secuencia son evidencia de que la estructura y la función de una proteína están relacionadas estrechamente. Por ello es interesante estudiar la relación que existe entre la estructura y los residuos críticos. En estudios de mutagénesis se ha encontrado una relación entre el *área accesible al solvente* de un residuo y su criticalidad. Los residuos que están expuestos al solvente tienden a tolerar una mayor cantidad de mutaciones comparados con los que no lo están (Loeb et al. [24], Rennell et al. [31], Huang et al. [18]). Aunque los residuos conservados no siempre son críticos, ni viceversa, también se ha encontrado una relación entre el área accesible al solvente, la densidad de contactos (que es equivalente al grado en una red de contactos) y la tasa evolutiva en un análisis de 938 proteínas de levadura de estructura conocida o modeladas por homología (Franzosa y Xia [13]). Esta relación es fuerte, lineal y positiva para el caso del área accesible al solvente (es decir, entre más accesible al solvente sea el residuo, evoluciona más rápido) y débil y negativa para la densidad de contactos (los residuos con alta densidad de contactos tienden a evolucionar más lento).

Algunos trabajos han buscado predecir residuos críticos para la función a partir de la red de contactos de una proteína. Amitai et al. [1] usaron la centralidad de cercanía (*closeness centrality*) calculada de una red de contactos para predecir sitios activos. Thibert et al. [37] evaluaron el uso de varias propiedades calculadas de la red de contacto como predictores de residuos críticos en tres proteínas y encontraron que el uso de algunos de ellos mejoraban la predicción basada en información evolutiva. En un trabajo subsecuente, Cusack et al. [5] encontraron que es posible predecir residuos críticos únicamente usando una medida de centralidad, la *conectividad dinámica*, basada en la centralidad de intermediación (*betweenness*). Este principio se usó para elaborar un programa de predicción de residuos críticos, `JAMMING`, que se compara con los modelos desarrollados en este trabajo.

En un trabajo reciente, Fajardo et al. [12] usaron *closeness*, *betweenness* y *pagerank* para predecir residuos catalíticos. Encontraron que las medidas de centralidad que usaron no fueron buenos predictores para este propósito en comparación al área accesible al solvente. Estos resultados sugieren que toda la capacidad predictiva de las medidas de centralidad para predecir residuos catalíticos pudiera estar ligada a su correlación con el área accesible al solvente.

2.2 PLANTEAMIENTO DEL PROBLEMA

Previamente se ha mostrado que algunas propiedades estructurales de un residuo, como su área accesible al solvente y algunas medidas de centralidad calculadas de una red de contactos, están relacionadas con la criticalidad del mismo. También se ha encontrado una alta correlación entre el área accesible al solvente y algunas medidas de centralidad,

por lo que incluso se ha propuesto que el poder predictivo de las medidas de centralidad (al menos para el caso de residuos catalíticos) sólo se debe a su correlación con el área accesible al solvente.

Uno de los fines de este trabajo es evaluar si las medidas de centralidad y el área accesible al solvente, en combinación con otras variables estructurales, tienen información redundante para la predicción de residuos críticos. También se evalúa si modelos que combinan varias medidas de centralidad entre sí y con atributos estructurales "tradicionales" son capaces de mejorar la predicción de `JAMMING`, que está basada en una sola medida de centralidad.

2.3 HIPÓTESIS

A. La predicción de residuos críticos con base a varios atributos estructurales mejorará la predicción de `JAMMING`, un método de predicción que está basado en una sóla medida de centralidad.

B. Las medidas de centralidad proveen información independiente al área accesible al solvente para predecir residuos críticos.

2.4 OBJETIVOS

A. Explorar si las medidas de centralidad corresponden a alguna propiedad estructural visualmente detectable en la estructura tridimensional de la proteína, distinta del área accesible al solvente.

B. Ajustar modelos que incorporen distintas combinaciones atributos estructurales dependientes de la estruc-

tura primaria, secundaria y terciaria de la proteína, incluyendo medidas de centralidad de la red de contactos.

c. Evaluar la predicción del modelo con una medida objetiva de validación, para evitar el sezgo presente en evaluarlo con los mismos datos ajustados.

Parte II

MATERIALES Y MÉTODOS

3

ANÁLISIS ESTRUCTURAL DE PROTEÍNAS

3.1 PROTEÍNAS, ESTRUCTURAS Y RESÍDUOS CRÍTICOS

Facts are the air of scientists.
Without them you can never fly.

— Linus Pauling

El objetivo de este trabajo es desarrollar un método para predecir los *residuos críticos* de una proteína (es decir, los que afectan la función más frecuentemente al ser mutados). Para ello nos basamos en datos publicados de seis proteínas de estructura conocida que han sido mutadas de manera extensa (Tabla 3.1). Nos enfocamos en trabajos en los que se obtuvieron mutantes en las que sólo se modificó un residuo de la proteína (a las cuales nos referimos como *mutantes sencillas*), con la única excepción de la beta-lactamasa, en el que se obtuvieron mutantes aleatorias de regiones de residuos contiguos (ver más abajo). Para cada proteína se escogió una estructura del Brookhaven Protein Data Bank representativa para el análisis estructural. Las estructuras se eligieron tomando en cuenta dos factores: a) cobertura de la secuencia y b) resolución.

Fue necesario tener una definición operacional clara de residuos crítico para poder clasificar las posiciones en críticas o no críticas según los datos experimentales. En algunos casos, los autores de la mutagénesis definieron los que para ellos son los residuos críticos de la proteína. En otros, únicamente se publicaron los datos crudos de las mutantes obte-

Tabla 3.1: PROTEÍNAS Y ESTRUCTURAS. Proteínas usadas para ajustar el modelo y estructuras del Brookhaven Protein Data Bank utilizadas para la elaboración de la red de contacto (ver abajo).

PROTEÍNA	ORGANISMO	PDB	REF.
Proteasa	VIH-1	2NPH	[7]
Retrotranscriptasa	VIH-1	2ZD1	[8]
Gen V	Bacteriófago f1	1GVP	[35]
Lisozima	Bacteriófago T4	3LZM	[26]
β-lactamasa TEM-1	*E. coli*	1ZG4	[34]
Represor Lac	*E. coli*	2P9H	[6]

nidas y sus fenotipos. Entonces nos fue necesario idear un criterio para poder clasificar los residuos de la proteína de acuerdo a los datos. Como cada grupo siguió una estrategia experimental distinta, no fue posible utilizar una definición única para todos los casos y tuvimos que usar criterios *ad hoc*. Sin embargo, hay una constante en las definiciones que usamos: los residuos críticos son los que toleran pocas mutaciones sin que se afecte la función. El criterio que usamos para cada proteína se detalla a continuación. La lista completa de los residuos críticos y no críticos para cada proteína se encuentra en el apéndice.

3.1.1 *Proteasa del VIH-1*

La proteasa del VIH-1 es una aspartil proteasa homodimérica de 99 aminoácidos necesaria para el ciclo infectivo del VIH. Esta proteína está codificada en el gen *pol* que, además de la proteasa, codifica una transcriptasa reversa, una

integrasa y una RNAsa. Los productos del gen *pol* forman una poliproteína inicialmente: la proteasa es responsable de procesarla para obtener los productos maduros.

Residuos críticos

Para definir los residuos críticos de la proteasa del VIH-1 nos basamos en el trabajo de Loeb et al. [24]. Ellos obtuvieron mutantes sencillas de la proteasa por medio de una mutagénesis aleatoria suficientemente extensa para que pudieran obtener al menos una mutante por cada posición de la proteína. Despues de ello evaluaron la actividad de las mutantes obtenidas detectando los productos del gen *pol* con Western Blots. Clasificaron a las mutantes en tres grupos: fenotipo silvestre si se detectaron únicamente los productos procesados, intermedio si se detectaron tanto productos procesados como no procesados, o mutante si sólo se detectaron productos no procesados.

Tabla 3.2: GRUPOS DE RESIDUOS Los veinte aminoácidos fueron agupados por Loeb et al. [24] en seis grupos. Las mutaciones por residuos del mismo grupo se consideran conservativas y las mutaciones por residuos de otros grupos no conservativas.

GRUPO	RESIDUOS
Polares sin carga real	ASTGP
Hidrofóbicos	VLIM
Aromáticos	FYW
Carga positiva	RKH
Carga negativa y polares	DENQ
Cisteína	C

Los autores de la mutagénesis agrupan a los aminoácidos en seis grupos, según sus características fisicoquímicas (Tabla 3.2), y definen una mutación conservativa como aquella en la que se sustituye un residuo por otro del mismo grupo. Consideran críticos a los residuos para los que todas las mutaciones no conservativas obtenidas dan lugar a un fenotipo mutante en el ensayo funcional descrito anteriormente.

En este trabajo usamos esta misma definición de residuo crítico para la proteasa del VIH-1, considerando sólo las posiciones para las que se haya obtenido al menos una sustitución no conservativa en las mutantes experimentales. Las posiciones que no cumplen con este requisito se consideraron de criticalidad desconocida y fueron excluídas del análsis.

3.1.2 *Retrotranscriptasa del VIH-1*

En la poliproteína del gen *pol* del VIH mencionada en la sección anterior también se incluye una transcriptasa reversa. Esta transcriptasa tiene dos actividades enzimáticas: a) una DNA polimerasa que puede copiar un molde de DNA o RNA y b) una RNAsa que degrada sólo si el RNA es parte de un duplex con DNA (Sarafianos et al. [33]). Combinando estas actividades, el resultado neto es transcribir RNA de cadena sencilla a DNA de doble cadena. El virus del VIH usa la retrotranscriptasa para copiar su genoma dentro del citosol de la célula infectada e integrarlo al genoma del huesped.

Estructuralmente, la retrotranscriptasa es un dímero asimétrico en el cual ambas cadenas se derivan de la misma porción de la poliproteína, pero una de las subunidades es más grande que la otra (66 kDa y 51 kDa).

Residuos críticos

Wrobel et al. [39] obtuvieron mutantes sencillas de dos subdominios de la transcriptasa reversa por mutagénesis aleatoria y evaluaron la capacidad de polimerizar DNA de las mutantes, así como su correcto procesamiento al heterodímero asimétrico.

Sin embargo, si hubiéramos usado la misma definición de residuo crítico que se usó para la proteasa habría una proporción mucho menor de residuos críticos que para el resto de las proteínas usadas en este trabajo. Para hacer los datos comparables a los de las demás proteínas, modificamos ligeramente la definición: se consideró un residuo crítico si al menos el 75 % de las mutantes no conservativas (con la misma definición de la Tabla 3.2) tuvieron un fenotipo mutante.

3.1.3 *Gen V del Bacteriófago f1*

Los productos de los genes II y V son necesarios para que el el bacteriófago f1 pueda replicar su DNA. La proteína del gen II es una endonucleasa necesaria para la síntesis fago-específica de DNA. La proteína del gen V es una proteína dimérica que se une a cadenas sencillas de DNA y RNA. Su función principal es evitar la síntesis de la cadena complementaria del genoma del bacteriófago (que es DNA de cadena sencilla) y ayudar a empaquetarlo para su introducción en la cápside del virus. También tiene funciones regulatorias: inhibe la traducción del RNA de los genes del hospedero y de la proteína II del fago.

La proteína del gen V es un homodímero. Sin embargo, la estructura tridimensional usada (Tabla 3.1) sólo contiene un monómero, y se analizó como tal.

Residuos críticos

Terwilliger et al. [36] definieron los residuos críticos de manera distinta a los grupos anteriores. Primero, modelaron la actividad esperada de una mutante con la siguiente ecuación:

$$A_{model} = \langle A \rangle + b_1(f - \langle f \rangle) + b_2(g - \langle g \rangle),$$

en donde

- A_{model} es la actividad esperada (una mutante inactiva se codifica como 0, una parcialmente activa como 1 y una mutante completamente activa como 2),
- $\langle A \rangle$ es la actividad promedio de todas las mutantes,
- f es la frecuencia de cambio del residuo original al residuo sustituyente según la matriz de Dayhoff et al. [9],
- $\langle f \rangle$ es la frecuencia de cambio promedio en los datos,
- g una variable que es 1 si el residuo sustituyente es glicina o prolina y 0 si es cualquier otro,
- $\langle g \rangle$ la proporción de residuos sustituyentes que es glicina o prolina, y
- b_1 y b_2 son parámetros del modelo que se ajustan a los datos.

Ajustaron el modelo a los datos experimentales de mutantes de la proteína del gen V para evaluar dos fenotipos: la inhibición del crecimiento de *E. coli* y la propagación del fago (el modelo se ajustó independientemente para cada fenotipo). Luego, calcularon la diferencia entre la actividad

esperada según el modelo y la observada para cada mutante

$$\Delta A = A_{observed} - A_{model}$$

y la tolerancia de un sitio a mutaciones como el promedio de esta diferencia para cada mutante del sitio

$$T_k = \frac{1}{n} \sum_{i=1}^{n} \Delta A_i \,.$$

Los valores negativos corresponden a posiciones en las que los fenotipos de las mutantes son menos activos que lo esperado. Consideramos críticas todas las posiciones con un valor de tolerancia $T_k \leq -0.4$ para cualquiera de los dos fenotipos, y no críticas las que tengan valores mayores, siempre y cuando se hubieran obtenido al menos 3 mutantes en la posición evaluada.

3.1.4 *Lisozima del Bacteriófago T4*

Las lisozimas son hidrolasas de glicósidos. Su función es degradar las paredes celulares bacteriales, pues hidrolizan las cadenas de carbohidratos presentes en ellas. Una función de la pared celular bacterial es resistir la diferencia entre la presión osmótica intracelular y extracelular. Debido a ello, si la diferencia de presión osmótica es alta, el daño en la pared celular ocasiona la lisis de la bacteria (de ahí el nombre).

Algunos organismos usan lisozimas como un mecanismo de defensa en contra de infecciones bacterianas. Sin embargo, esa no es la única función de esta familia de proteínas. Algunos bacteriófagos, como T4, usan una lisozima para

destruir a la bacteria huesped y expulsar las partículas virales maduras al final de su ciclo reproductivo. Además, algunas aves y mamíferos rumiantes las untilizan para digerir bacterias degradadoras de celulosa presentes en su tracto digestivo.

Residuos críticos

Rennell et al. [31] siguieron un procedimiento distinto al de los grupos anteriores para obtener las mutantes: en vez realizar mutagénesis aleatoria, utilizaron *mutantes ámbar*. En esta estrategia se obtiene una mutante sitio-dirigida en la que se modifica el codón de interés al codón de paro UAG, también conocido como ámbar (de ahí el nombre).

La construcción con el gen mutado se introduce a cepas de bacteria que tienen un tRNA modificado para insertar un aminoácido específico en codones UAG. De esta forma con una sóla mutante por posición a nivel de DNA se pueden obtener 11 o 12 mutantes a nivel de proteína (dependiendo de si el residuo silvestre en esa posición es parte del conjunto de residuos que las cepas de bacterias pueden incorporar en el codón ámbar o no). Este tipo de mutagénesis es mucho más homogénea que la mutagénesis aleatoria, pues todas las posiciones mutadas son cambiadas por los mismos residuos.

Cada posición de la lisozima (excepto el codón de inicio) fue sustituída por un codón ambar por Renell et al. El fenotipo que evaluaron fue la formación de placas, que sucede cuando las células son lisadas por lisozima activa, a dos temperaturas: 25°C y 37°C. Consideramos críticas las posiciones de la proteína para las que sólo el 25 % o menos de las mutantes ámbar resultaron en un fenotipo silvestre para ambas temperaturas.

3.1.5 β-*lactamasa de tipo TEM-1*

Las β-lactamasas son enzimas bacteriales que degradan antibióticos β-lactámicos (como las penicilinas y cefalosporinas) y confieren resistencia a los mismos. Frecuentemente se secretan al espacio extracelular para degradar los antibióticos antes de que entren a la bacteria. La β-lactamasa más común en las bacterias Gram negativas es la de tipo TEM-1. El gen de las β-lactamasas de tipo TEM casi siempre se encuentra en plásmidos.

Residuos críticos

Nos basamos en los datos de Huang et al. [18] para obtener los residuos críticos de la β-lactamasa. Ellos usaron una estrategia de mutagénesis aleatoria para obtener un gran número de mutantes para cada posición evaluada. En esta estrategia, primero dividieron el gen en grupos de codones contiguos que comprenden desde la posición 22 hasta el carboxilo terminal de la proteína. La mayoría de los grupos fueron de tres codones de longitud, pero hubo algunos más grandes (de hasta 6 codones contiguos). Para cada uno de los grupos generaron una mutante en la que sustituyeron los codones correspondientes por un inserto con un sitio de restricción *SalI*, que no estaba presente en ningún otro lugar del plásmido utilizado.

Después sintetizaron un oligonucleótido con el fin de reemplazar al sitio *SalI* por DNA de secuencia aleatoria (dentro del marco de lectura del gen). De esta forma generaron un gran número de mutantes del grupo de codones correspondientes del gen de la β-lactamasa. Para cada grupo, los plásmidos recombinantes fueron mezclados y se usaron para transformar *E. coli* (después de digerir los plásmidos con

SalI para evitar la transformación plásmidos no mutagenizados).

Huang et al. repitieron este procedimiento para cada uno de los grupos de codones contiguos mencionados anteriormente. Las colonias transformadas fueron puestas en presencia de ampicilina para evaluar el fenotipo. Como sólo las colonias con una β-lactamasa funcional pueden crecer en estas condiciones, *únicamente se recuperaron las mutantes con fenotipo silvestre*. Secuenciaron la β-lactamasa en las colonias que crecieron, y de esa forma obtuvieron una lista de mutantes con fenotipo silvestre para cada una de las posiciones evaluadas de la proteína.

Los autores consideraron críticos a los residuos para los que no se encontró ninguna mutante WT. Sin embargo, nosotros modificamos ligeramente esta definición para tratar de ser un poco más consistentes con la usada para las otras proteínas. Bajo nuestra definición, toleramos la existencia de mutantes conservativas con fenotipo silvestre en los residuos críticos, pero no mutantes con fenotipo mutante. Es decir, consideramos críticas a aquellas posiciones para las que no obtuvieron mutantes no conservativas de fenotipo silvestre (según los mismos grupos de la Tabla 3.2).

3.1.6 *Represor Lac de E. coli*

El represor Lac es un homotetrámero que funciona como regulador del operón lac. En ausencia de lactosa, el represor se une al operador de Lac, que codifica para proteínas necesarias para metabolizar este azúcar. Al hacerlo, inhibe la transcripción del gen y evita la producción innecesaria de sus productos. La alolactosa, un derivado de la lactosa que se obtiene en el metabolismo, se une al represor Lac de manera alostérica. Cuando esto ocurre, el represor sufre un

cambio conformacional que disminuye su afinidad por el operador, lo que permite la expresión del gen.

Existen dos trabajos publicados en los se obtuvieron mutantes del represor Lac (Kleina y Miller [21], Markiewicz et al. [25]). Entre ambos trabajos, hicieron mutantes ámbar para las posiciones 2 a 329 de la secuencia de la proteína. Los represores mutantes pueden tener varios fenotipos distintos. Al fenotipo mutante más común se le denomina I^-, estas mutantes expresan los productos de manera constitutiva (es decir, en ausencia de lactosa). Este fenotipo puede ocasionarse por no unirse al operador del gen, mal plegamiento o agregación. Hay otro fenotipo, menos común, que se denomina I^s en el cual los productos del gen no se expresan aún en presencia de lactosa. Algunos de estos fenotipos se causan por falta de la unión alostérica de alolactosa y otros por una unión demasiado fuerte al operador. A estos últimos se les denomina I^{tb}. Todos los fenotipos distintos a WT se consideraron fenotipos mutantes para fines del análisis de criticalidad.

Residuos críticos

En los trabajos de mutagénesis también evaluaron los fenotipos a distintas temperaturas (30°C, 37°C y 42°C). Sin embargo, los resultados fueron reportados de manera distinta en el trabajo de Kleina y Miller que en el de Markiewicz et al. El primer trabajo reporta los fenotipos individuales de las mutantes. En el segundo trabajo se hicieron muchas más mutantes, pero sólo se reportó el número de mutantes totales por posición de la proteína de cada fenotipo en una figura.

De manera similar al caso de la lisozima, se consideraron críticas las posiciones para las que el 25 % o menos de

las mutantes tuvieron fenotipo WT. Nos basamos principalmente en Markiewicz et al. para determinar las posiciones críticas y, en los casos ambiguos, comprobamos los fenotipos de las mutantes individuales en Kleina y Miller.

3.2 REDES DE CONTACTOS

La predicción de residuos críticos se hace a través de información estructural. Algunos de los atributos usados para la predicción se calculan de la *red de contactos* de la proteína. Para construir la red de contactos a partir de la estructura correspondiente se escribió un programa en el lenguaje Python, basado en las librerías `Biopython` (Cock et al. [4], Hamelryck y Manderick [17]), `scipy` (Jones et al. [19]) y `networkx` (Hagberg et al. [15]). Se detalla el algoritmo para la construcción de la red a continuación. Para construirla se requiere de un parámetro, el radio de contacto R_c.

Algoritmo

1. Extraer las coordenadas espaciales (x, y, z) de cada átomo de la proteína presente en el archivo del Protein Data Bank.

2. Encontrar la lista de todos los pares de átomos que están a una distancia menor o igual a R_c.

3. Construir una red en la cual los nodos sean los residuos de la proteína.

4. Unir los residuos que tengan al menos un par de átomos (uno de cada residuo) a una distancia menor o igual a R_c con una arista en la red.

El usuario puede especificar qué cadenas de la proteína usar para construir la red. En este trabajo, siempre construimos la red con todas las cadenas idénticas de interés presentes en el archivo PDB (por ejemplo, si la proteína es un homodímero, incluimos ambas cadenas en la red). A las aristas entre residuos se les denomina *contactos*. Las redes construidas de esta forma son usadas para calcular algunos atributos, como medidas de centralidad (ver más adelante).

Visualización

Para visualizar redes de contactos se escribió un programa en el lenguaje Python utilizando la librería `mayavi` (Ramachandran y Varoquaux [30]). Las visualizaciones se realizaron en tres dimensiones, colocando cada nodo en la posición correspondiente al centro de masa del residuo que representa, y dibujando una arista entre cada nodo y sus vecinos en la red de contactos.

Los nodos pueden colorearse de acuerdo a cualquier propiedad, lo que hace posible la exploración visual de medidas de centralidad u otros atributos. Visualizar las redes de contactos de esta forma permite observar si la propiedad evaluada depende de la estructura terciaria y/o cuaternaria de la proteína o es independiente a ella.

3.3 ATRIBUTOS ESTRUCTURALES

Las estructuras de proteínas mencionadas anteriormente (Tabla 3.1) fueron analizadas para extraer atributos que dependen de varios niveles de estructura. Algunos sólo dependen de la estructura primaria (secuencia) de la proteína. Hay otros que dependen de la estructura secundaria y/o terciaria, y algunos más que dependen de la combinación

de varios niveles de estructura. Se describe cada uno de ellos a continuación.

3.3.1 *Estructura primaria*

Algunos de los atributos que usamos dependen únicamente del tipo de aminoácido presente en la posición de la proteína a evaluar. Estos atributos se enlistan en la Tabla 3.3 y se detallan a continuación.

Tabla 3.3: ATRIBUTOS DE ESTRUCTURA PRIMARIA Atributos dependientes de la secuencia primaria de la proteína.

ABREVIACIÓN	ATRIBUTO
Hyd_{KD}	Hidrofobicidad (Kyte y Doolittle)
Hyd_{Rose}	Hidrofobicidad (Rose)
SA	Área superficial
Exch	Intercambiabilidad
Chrg	Residuo cargado
Arom	Residuo aromático

Hidrofobicidad

Las moléculas no polares tienden a sufrir una repulsión aparente de solventes acuosos. La hidrofobicidad es una propiedad física de una molécula o parte de ella que se refiere al grado en que esto ocurre. Existen varias escalas distintas de hidrofobicidad para los residuos de una proteina. Se tomaron como atributos los valores de dos de las más comunmente usadas: la de Kyte y Doolittle [22] y la de Rose et al. [32]. Estas escalas se construyeron en base a

propiedades distintas: la de Kyte y Doolittle se fundamenta principalmente en propiedades estructurales y fisicoquímicas de los aminoácidos, mientras que la de Rose lo hace en la propensidad empírica que tienen de estar en regiones poco accesibles al solvente (calculada a través de un análisis de estructuras cristalográficas).

Área superficial

Se usó como atributo el área superficial de cada residuo accesible en un tripéptido de la forma Gly-R-Gly, calculada por Chothia [3]. Los residuos con mayor área superficial tienden a tener más vecinos en la red de contactos de la proteína.

Intercambiabilidad

Las matrices de sustitución de aminoácidos describen la probabilidad evolutiva de sustitución de un aminoácido a otro en una proteína. Aminoácidos con propiedades similares suelen tener una probabilidad de sustitución más alta que aquellos con propiedades distintas. Nos basamos en la matriz de sustitución de Le y Gascuel [23] para obtener una medida de qué tan similar es un aminoácido en promedio a todos los demás, que nombramos *intercambiabilidad*. Para obtenerla calculamos la media intercuartil (es decir, la media desechando el 25 % menor y 25 % mayor de los datos) para cada aminoácido con respecto a todos los demás en esta matriz con el lenguaje de programación R (R Core Team [29]).

Residuo cargado

Algunos aminoácidos tienen grupos amino o carboxilo libres que son capaces de ionizarse (aspartato, glutamato,

lisina, arginina e histidina). Incluimos una variable binaria que es 1 si el aminoácido en cuestión pertenece a este grupo y 0 de lo contrario.

Residuo aromático

La fenilalanina, la tirosina, el triptofano y la histidina son aminoácidos que tienen grupos funcionales aromáticos. Incluimos un atributo binario que tiene un valor de 1 si el aminoácido en cuestión es alguno de los mencionados y 0 de lo contrario.

3.3.2 *Estructura secundaria*

También se tomó en cuenta la estructura secundaria de la proteína en la posición a evaluar en el modelo. Ésta se obtuvo con el programa `DSSP` (Kabsch y Sander [20]), que asigna una de siete estructuras secundarias posibles a cada residuo de la proteína a partir de las coordenadas espaciales en el archivo PDB.

Algunas de las estructuras secundarias son poco comunes, y, como sólo fueron analizadas seis estructuras, hay ejemplos inexistentes o escasos de algunas de ellas. Además, es conveniente limitar el número de atributos en el modelo para evitar el sobreajuste. Por estas razones decidimos simplificar los resultados de `DSSP`, agrupando las estructuras secundarias en tres grupos: hélices (incluyendo hélices α, π y 3_{10}), estructuras beta (puentes y hojas β) y otros.

Para fines del modelo, codificamos la estructura secundaria simplificada en dos variables binarias: *Hel* y *Beta*. Estas variables se enlistan en la Tabla 3.4. Los valores de *Hel* y *Beta* correspondientes a cada estructura secundaria se detallan en la Tabla 3.5.

Tabla 3.4: ATRIBUTOS DE ESTRUCTURA SECUNDARIA Atributos dependientes de la estructura secundaria de la proteína.

ABREVIACIÓN	ATRIBUTO
Hel	El residuo es parte de una hélice
Beta	El residuo es parte de un puente β u hoja β

Tabla 3.5: VALORES DE ESTRUCTURA SECUNDARIA Estructura secundaria que corresponde a los valores posibles de las variables de la Tabla 3.4.

$(Hel, Beta)$	ESTRUCTURA SECUNDARIA
$(1,0)$	Hélice α, hélice π o hélice 3_{10}
$(0,1)$	Puente u hoja beta
$(0,0)$	Otra

3.3.3 *Estructura terciaria y cuaternaria*

Finalmente, incorporamos atributos sobre los aminoácidos que dependen de la estructura terciaria de la proteína. Estos atributos incluyen el área accesible al solvente relativa del residuo en cuestión y otros atributos obtenidos a partir del análisis de la red de contactos de la proteína. La contrucción de la red de contactos se detalla en la Sección 3.2.

Área accesible al solvente relativa

El área accesible al solvente relativa ($RASA$) es la proporción del área superficial del aminoácido en cuestión que

Tabla 3.6: Atributos dependientes de la estructura terciaria y cuaternaria de la proteína.

ABREVIACIÓN	ATRIBUTO
RASA	Área accesible al solvente relativa
Deg	Grado
Eig	Centralidad de eigenvector
Pag	Pagerank
Cmn	Comunicabilidad
Cls	Cercanía
Cfc	Cercanía (flujo de corriente)
Bet	Intermediación
Cmb	Intermediación (comunicabilidad)
Cfb	Intermediación (flujo de corriente)
N_{Hyd}	Hidrofobicidad de los vecinos
C_{Frac}	Fracción de contactos

es accesible por moléculas de agua, ya sea por que el residuo está cerca de la superficie de la proteína o por que se encuentra cerca de alguna cavidad. El cálculo de esta propiedad se realizó utilizando el programa `DSSP` (Kabsch y Sander [20]). Este es el único atributo dependiente de la estructura terciaria de la proteína que usamos que se calcula sin usar la red de contactos.

Medidas de centralidad

Las medidas de centralidad pretenden medir la importancia de un nodo en una red (Newman [28]). Puede haber distintas definiciones de importancia, que corresponden a

distintas medidas de centralidad. Se calcularon varias de ellas a partir de las redes de contactos de las proteínas usando funciones de la librería `networkx` del lenguaje de programación Python. Las funciones usadas se detallan en la Tabla 3.7. Los detalles de los métodos que se usan el para cálculo de cada centralidad se encuentran en el manual de `networkx` (Hagberg et al. [16]).

Tabla 3.7: CÁLCULO DE CENTRALIDADES Funciones de `networkx` con las que se calcularon las medidas de centralidad usadas como atributos.

ABREVIACIÓN	FUNCIÓN
Deg	`degree`
Eig	`eigenvector_centrality_numpy`
Pag	`pagerank_scipy`
Cmn	`communicability_centrality`
Cls	`closeness_centrality`
Cfc	`current_flow_closeness_centrality`
Bet	`betweenness_centrality`
Cmb	`communicability_betweenness_centrality`
Cfb	`current_flow_betweenness_centrality`

La distribución que toman los valores de algunas de las medidas de centralidad depende del tamaño de la red (es decir, del número de residuos de la proteína, ver Subsección A.1.1). Por ello nos fue necesario transformar los valores de distintas proteínas a una escala común. Para hacerlo, ordenamos los valores de cada centralidad de una proteina por magnitud y los convertimos a *rangos* (número de órden). Es decir, para una estructura con n residuos, $R = 1$ para el

valor más bajo de centralidad y $R = n$ para el valor más alto. Los rangos se normalizaron con la siguiente fórmula:

$$C_{norm} = \frac{R-1}{n-1}$$

en dónde la *centralidad normalizada* $C_{norm} \in [0,1]$ tiene una distribución uniforme para todas las proteínas.

Grado

El grado (*Deg*) de un nodo es la medida de centralidad más sencilla: es simplemente el número de vecinos que tiene en la red. Esta es una medida de centralidad *local*, pues sólo depende de la estructura de la red alrededor del nodo. Esto es en contraste con las medidas de centralidad globales, que dependen de la estructura completa de la red.

En el caso de la red de contactos, el grado de un residuo corresponde al número de residuos con los que hace contacto (es decir, al número de residuos que están a una distancia menor a R_C).

Eigenvector y Pagerank

La centralidad de eigenvector (*Eig*) puede considerarse una versión extendida de la centralidad de grado. La diferencia es que se considera que no todos los vecinos del nodo son igualmente importantes. A cada nodo se le da una puntuación que depende del número de vecinos y de la importancia de los mismos. Se puede usar un método iterativo, en el que inicialmente se le da la misma importancia a todos los nodos, y en cada paso los nodos mejor conectados "roban" importancia a los menos conectados. Este proceso se repite hasta que se llega a valores estables de importancia para cada nodo.

Existen variantes de la centralidad de eigenvector, como *Pagerank* (*Pag*), desarrollada por Google para ordenar los resultados de búsquedas en internet. Esta variante está adaptada para simular la probabilidad de que un usuario siga enlaces en una página de internet o no.

Comunicabilidad

Un camino cerrado en una red es aquel que empieza y termina en el mismo nodo. La comunicabilidad (*Cmn*), también conocida como centralidad de subgrafo, es el número de caminos cerrados dentro de la red (de todas longitudes) que inician y terminan en el nodo de interés.

Cercanía

La cercanía (*Cls*) de un nodo es el inverso de la distancia promedio entre el nodo y todos los demás nodos de la red. La distancia entre dos nodos se puede calcular de dos maneras distintas: a) distancia del camino más corto o b) caminos aleatorios. Esto da origen a dos medidas de centralidad distintas: cercanía de caminos más cortos (o simplemente cercanía) y cercanía de flujo de corriente.

Intermediación

La intermediación (*betweenness*) de un nodo es una medida que depende del número de veces que el nodo se encuentra dentro de los caminos entre todos los otros pares de nodos de la red. Al igual que en el caso de la cercanía, existen variantes dependiendo de si se toman los caminos más cortos (intermediación, *Bet*) o caminos aleatorios (intermediación de flujo de corriente, Cfb).

Más recientemente, Estrada et al. [11] desarrollaron otra medida de la intermediación de un nodo: se toman en cuen-

ta caminos de todas longitudes, pero se les da un mayor peso a los caminos cortos. A esta medida de centralidad se le llama *intermediación de comunicabilidad* (Cmb).

Hidrofobicidad de los vecinos

Las redes de contacto también se usaron para calcular la hidrofobicidad promedio de los vecinos del residuo en cuestión en la red, según la escala de Kyte y Doolitle. Esta es una medida aproximada de la hidrofobicidad del ambiente alrededor del residuo.

Fracción de contactos

También calculamos una medida que llamamos la *fracción de contactos* de un residuo a partir de su red de contactos. Esta medida corresponde a la proporción de los contactos de la cadena de la proteína en donde se encuentra el residuo en los que éste está involucrado. Para obtenerla, primero se obtiene una subred que involucra sólo a los residuos de la cadena en la que está el residuo en cuestión. Luego se calcula la fracción de contactos

$$C_{Frac} = \frac{Deg_s}{N_c},$$

en dónde

- Deg_s es el grado (número de vecinos) del nodo en la cadena (subred), y
- N_c es el número total de contactos en la cadena.

3.3.4 *Atributos derivados / interacciones*

Además de los atributos previamente mencionados, se calcularon otros atributos derivados de ellos. Algunos de-

penden de la interacción entre la estructura primaria, secundaria y terciaria de la proteína, mientras que otros se derivan de la interacción entre dos propiedades dependientes de la estructura terciaria. Estos atributos se enlistan en la Tabla 3.8 y se describen a continuación.

Tabla 3.8: ATRIBUTOS DERIVADOS Atributos "compuestos" que fueron derivados de los atributos anteriores.

ABREVIACIÓN	ATRIBUTO
$Hel \times H_p$	Propensidad del residuo de formar hélices α
$Beta \times B_p$	Propensidad del residuo de formar hebras β
$Hyd_{KD} \times RASA$	Int. hidrofobicidad (KD) y RASA
$Hyd_{Rose} \times RASA$	Int. hidrofobicidad (Rose) y RASA
$Chrg \times RASA$	Int. carga y área accesible al solvente
$Arom \times RASA$	Int. residuo aromático y RASA
Deg/SA	Grado normalizado por área del residuo
$Hel \times RASA$	Int. hélice y RASA
$Beta \times RASA$	Int. hebra u hoja β y RASA
$Cmb \times Cls$	Int. *Cmb* y *Cls*

Interacciones

A los términos en los que una variable se multiplica por otra en un modelo lineal se les llama *interacciones*. Si consideramos la interacción entre las variables x_i y x_j cuando las otras variables se mantienen constantes en un modelo lineal obtenemos:

$$y = w_i x_i + w_j x_j + w_{ij} x_i x_j + c,$$

en dónde la constante c depende de los valores en que se dejan constantes las otras variables.

Podemos reajustar los términos para obtener

$$y = w_i x_i + (w_j + w_{ij} x_i) x_j + c,$$

un modelo que es lineal en x_j a valores fijos de x_i. Viendo al modelo desde esta prespectiva, podemos interpretar que el coeficiente w_i cambia la *ordenada aparente* a distintos valores de x_i, mientras que el coeficiente de la interacción w_{ij} cambia la *pendiente aparente* de x_j. De esta forma w_{ij} puede interpretarse como el cambio en el coeficiente de x_j por unidad de x_i.

Incluímos en el modelo varios atributos que son interacciones de diversas propiedades de la proteína con el área accesible al solvente. Entre estas propiedades están la hidrofobicidad, carga, aromaticidad y estructura secundaria del residuo. Estas interacciones pueden interpretarse como un cambio en el coeficiente correspondiente al área accesible al solvente cuando se modifica el valor de la variable correspondiente a la propiedad de interés.

Propensidad de hélice α / hebra β

Tambien incluímos interacciones que involucran atributos que no habíamos usado anteriormente y que merecen una mención especial. Fujiwara et al. [14] calcularon la propensidad que tiene cada aminoácido de pertenecer a hélices alfa y hebras beta para residuos expuestos ($RASA > 0.2$) y no expuestos ($RASA \leq 0.2$).

Calculamos las variables

$$\begin{aligned} Hp &= f \cdot H_{pexposed} + (1-f) \cdot H_{pburied} \\ Bp &= f \cdot B_{pexposed} + (1-f) \cdot B_{pburied}, \end{aligned}$$

en dónde $f = \frac{1}{1+ae^{-b(RASA)}}$ es una función que se usó para interpolar las propensidades de los residuos expuestos y no expuestos. Los parámetros $a = 99$ y $b = 5\ln 99$ se escogieron para que $f(0) \approx 0$, $f(1) \approx 1$ y $f(0.2) = 0.5$ (ver Figura 3.1).

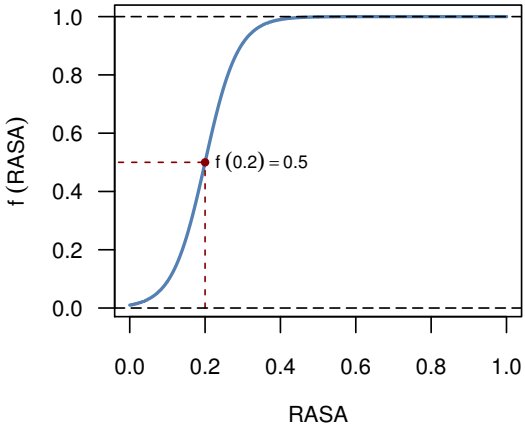


Figura 3.1: FUNCIÓN INTERPOLADORA Se muestra la gráfica de la función $f(RASA)$, con la que se interpola la propensidad de hélice/hebra a distintos valores de $RASA$. A valores bajos, $H_p \approx H_{pburied}$ y $B_p \approx B_{pburied}$. A valores altos, $H_p \approx H_{pexposed}$ y $B_p \approx B_{pexposed}$.

Se multiplica $Hel \times H_p$ y $Beta \times B_p$ para obtener los atributos finales de propensidad de hélice/beta. Esto se hace para que la variable sea H_p o B_p cuando el residuo tiene la estructura secundaria correspondiente (hélice o beta, respectivamente) y 0 de lo contrario.

Grado normalizado

El grado de un residuo Deg es el número de vecinos que tiene en la red de contactos. Sin embargo, el número de

vecinos es proporcional al área superficial del residuo. Por eso calculamos también el grado normalizado Deg/SA y lo incluimos como atributo en el modelo.

4

MODELADO

Essentially, all models are wrong, but some are useful.

— George Box

Predecir los residuos críticos de una proteína es esencialmente un problema de *clasificación binaria*, pues queremos clasificar los residuos en críticos o no críticos. Este tipo de problemas pueden abordarse por medio del aprendizaje automático supervisado. En esta sección describo a detalle el modelo usado en este trabajo y la estrategia que se siguió para ajustarlo a los datos.

4.1 DEFINICIONES

Para una proteína con n residuos de criticalidad conocida definimos un vector

$$\mathbf{y}_{prot} = (y_1, y_2, ..., y_n)$$

en el que $y_i = 1$ si el residuo i es crítico y $y_i = 0$ si no lo es (según las definiciones en la Sección 3.1). Este vector contiene los valores observados de la *variable de respuesta* que queremos predecir con el modelo.

El objetivo es predecir esta variable a partir de algunos de los atributos estructurales descritos en la Sección 3.3. Para ello definimos el *vector de atributos*

$$\mathbf{x}_i = (x_1, x_2, ..., x_m)_i = (x_{i1}, x_{i2}, ..., x_{im})$$

en el que se enlistan los valores de cada atributo para el residuo i. Por ejemplo, si x_1 corresponde a Hyd_{KD} y el residuo i es una lisina en la secuencia silvestre de la proteína, entonces $x_{i1} = -3.9$, el valor de la hidrofobicidad de Kyte y Doolittle para la lisina. Para cada proteina podemos definir la *matriz de atributos*

$$\mathbf{X}_{prot} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

que contiene los valores del vector de atributos para cada residuo de la proteína.

4.2 MODELO: REGRESIÓN LOGÍSTICA

En el aprendizaje automático supervisado suponemos que existe una función desconocida f que nos lleva del vector de atributos $\mathbf{x}_i$ a la variable de respuesta y_i (es decir, tal que $\mathrm{f}(\mathbf{x}_i) = y_i$). Esta función es la que nos gustaría obtener idealmente en el modelo, por lo que se le llama *función blanco.* Como se hace comunmente en el área del aprendizaje automático, consideramos la criticalidad observada de los residuos como si fueran valores generados por esta función. No tenemos manera de conocer f, pero podemos aproximarla con otra función de forma conocida.

Para aproximar la función blanco usamos modelos de *regresión logística*, que son de la forma

$$h(\mathbf{x}_i) = \theta(w_0 + w_1 x_{i1} + w_2 x_{i2} + ... + w_m x_{im})$$

en dónde θ es la *función logística*

$$\theta(s) = \frac{e^s}{1+e^s}$$

para escoger la función que usaremos para aproximar la función blanco. Si agregamos una coordenada artificial $x_{i0} = 1$ a cada vector de atributos podemos expresar el modelo de una forma más compacta:

$$h(\mathbf{x}_i) = \theta(w_0 x_{i0} + w_1 x_{i1} + w_2 x_{i2} + ... + w_m x_{im}) = \theta(\mathbf{w} \cdot \mathbf{x}_i) \tag{4.1}$$

en dónde $\mathbf{w} = (w_0, w_1, w_2, ..., w_m)$ y $\mathbf{x}_i = (x_{i0}, x_{i1,} x_{i2}, ..., x_{im})$.

La función logística transforma su parámetro $s \in \mathbb{R}$ a $\theta(s) \in (0,1)$. Es decir, transforma cualquier valor de entrada a un número entre cero y uno. Al pasar el modelo por esta función, $h(\mathbf{x}_i)$ tendrá valores entre cero y uno, que pueden ser interpretados como la *probabilidad predicha* por el modelo de que el residuo sea crítico según los valores de los atributos estructurales del residuo.

4.2.1 *Función de error: log loss*

Para ajustar cualquier modelo a datos experimentales es necesario minimizar una *función de error,* que es una medida de la calidad de la predicción que es una función de los parámetros del modelo (en este caso los coeficientes). El ajuste consiste en encontrar los parámetros para los que se minimize el error. Existe una función de error llamada *log loss* que es la correspondiente al modelo de regresión logística. El *log loss* de la predicción de un residuo i es

$$E_i = -(y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i))$$

en dónde

- y_i es el valor observado de la criticalidad del residuo en los datos, y
- $\hat{y}_i = \mathrm{h}(\mathbf{x}_i)$ es la probabilidad de que el residuo sea crítico, según la predicción del modelo.

Podemos definir el error *log loss* de la predicción de los n residuos de una proteína como el promedio del de cada residuo. Es decir,

$$E_{prot} = \frac{1}{n}\sum_{i=1}^{n} E_i = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)) \tag{4.2}$$

Hasta ahora hemos definido el error que usamos para residuos individuales y proteínas. Sin embargo, queremos ajustar el modelo a datos de varias proteínas. Entonces debemos definir el error para N proteínas distintas. De manera analoga, lo definimos como el error promedio de cada proteína

$$E_{total} = \frac{1}{N}\sum_{j=1}^{N} E_j = -\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i=1}^{n_j}(y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i))$$

en dónde E_j es el error de la proteína j y n_j su número de residuos de criticalidad conocida.

4.2.2 *Error aumentado*

Con la definición de E_{total} ya es posible ajustar un modelo de regresión logística a varias proteínas. Sin embargo para evitar el sobreajuste es conveniente agregarle otros términos para penalizar modelos complejos (a esto se le llama

regularización, ver la Subsección 1.3.3). Definimos el *error aumentado* como

$$E_{aug} = E_{total} + \alpha\lambda \|\mathbf{w}\|_1 + \frac{1}{2}\alpha(1-\lambda)\|\mathbf{w}\|_2^2 \tag{4.3}$$

en dónde

- $\|\mathbf{w}\|_1 = \sum_{i=0}^{m} |w_i|$ y $\|\mathbf{w}\|_2 = \sqrt{\sum_{i=0}^{m} w_i^2}$ son la norma $l1$ y $l2$ de $\mathbf{w}$, el vector de los coeficientes, respectivamente,
- α es el *parámetro de regularización*, y
- λ es la *proporción de regularización* $l1$.

La función de los dos términos que agregamos es penalizar modelos que tengan coeficientes elevados. Para dos modelos con el mismo E_{total} en los datos se va a preferir al modelo con los coeficientes más pequeños (que en cierto sentido es el modelo *más simple*) al minimizar E_{aug}.

El tamaño de los coeficientes se mide con una *norma* del vector que los contiene. Los términos que usamos incluyen dos normas distintas: la norma $l1$ que corresponde a la suma de los valores absolutos de los coeficientes, y la norma $l2$ que corresponde a la raíz de la suma de cuadrados de los mismos.

La norma $l1$ penaliza los coeficientes de las variables que son poco importantes para la predicción de tal forma que tienden a volverse exactamente 0. Por ello funciona como un proceso de *selección de variables*, pues ponerle un coeficiente de 0 a una variable equivale a excluirla del modelo (ver Ecuación 4.1).

El segundo término de regularización (norma $l2$) también "prefiere" coeficientes pequeños, pero éstos no tienden a volverse 0 exactamente. Por ello, penalizar el tamaño de la

norma $l2$ no selecciona variables. Sin embargo, los modelos regularizados con norma $l2$ tienden a tener una mejor capacidad predictiva que los regularizados con norma $l1$ en la mayoría de los casos (especialmente cuando el número de variables en el modelo no es muy elevado).

En este trabajo utilizamos ambas normas, en un proceso que se denomina *red elástica* (Zou y Hastie [40]). El tamaño de los términos de regularización en relación al error total depende de dos *hiperparámetros*: α y λ. Se les denomina hiperparámetros para contrastarlos con los parámetros del modelo, que son los coeficientes $\mathbf{w}$. Los hiperparámetros son constantes durante la minimización del error, y deben escogerse antes.

El valor de α determina el tamaño total de los términos de regularización (es decir, cuánta regularización se utilizará en total, sin importar el tipo). El valor de $\lambda \in [0, 1]$ determina qué proporción de regularización de cada tipo ($l1$ o $l2$) se usará. Por ejemplo, cuando $\lambda = 1$ tenemos sólo regularización $l1$, mientras que cuando $\lambda = 0$ sólo tenemos $l2$. Valores intermedios de λ resultan en ambos tipos de regularización en distintas proporciones (por ejemplo, $\lambda = 0.5$ les da el mismo peso a la regularización $l1$ y $l2$, mientras que $\lambda = 0.9$ le da mucho más peso a la regularización $l1$).

Para usar un esquema de regularización es conveniente normalizar las variables de tal forma que todas tengan la misma desviación estándar (pues de otra forma algunas variables serán penalizadas más que otras). Por ello cada variable se normalizó como

$$x_{ijk}^{norm} = \frac{x_{ijk} - \bar{x_i}}{s_{x_i}} \tag{4.4}$$

antes de minimizar el error aumentado.

En esta expresión x_{ijk} representa el valor de la variable i para el residuo k de la proteína j. También aparecen

$$\bar{x_i} = \frac{1}{N}\sum_{j=1}^{N}\bar{x_{ij}} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{k=1}^{n_j}x_{ijk}$$

y

$$s_{x_i} = \frac{1}{N}\sum_{j=1}^{N}\sqrt{\frac{1}{n_j - 1}\sum_{k=1}^{n_j}(x_{ijk} - \bar{x_{ij}})^2}$$

que son el promedio entre todas las proteínas del conjunto de entrenamiento de la media de la variable y de su desviación estándar, respectivamente.

4.3 AJUSTE DEL MODELO

4.3.1 *Validación cruzada*

En el aprendizaje automático se hace una distinción entre el *conjunto de entrenamiento*, usado para ajustar el modelo, y un *conjunto de validación*, que sirve para estimar el rendimiento del modelo en datos nuevos y consiste de datos que no han sido "vistos" por el modelo durante el proceso de ajuste.

La *validación cruzada* es una manera de usar los mismos datos como conjunto de entrenamiento y de validación. En este esquema se dividen los datos en k partes: una de ellas se separa para usarse como conjunto de validación y se entrena el modelo con el resto. La parte separada se usa para estimar el error fuera de los datos. Este proceso se repite, usando cada una de las k partes como conjunto de validación y entrenando el modelo en el resto. Al final se pro-

median los errores obtenidos en las partes separadas para obtener el estimador de error de validación cruzada.

Lo más común es dividir los datos aleatoriamente en grupos iguales. En nuestro caso, esto corresponde a dividir los residuos de todas las proteínas al azar en k grupos. Sin embargo, algunos de los atributos que usamos dependen de los valores que haya tomado el mismo atributo en los otros residuos de la proteína (como las centralidades normalizadas). Cuando el modelo se use en datos nuevos no será enfrentado a residuos distintos de las proteínas usadas para ajustarlo, sino a proteínas distintas. Por ello dividir estos datos de la forma convencional probablemente subestimaría el error del modelo en datos nuevos.

Para evitar este problema, la división de los datos se hizo *por proteína*. Cada proteína en turno se escoge como *conjunto de validación* (Figura 4.1, panel A), y se usa para calcular el error del modelo después de ser ajustado a las otras cinco proteínas.

La validación cruzada también puede usarse para escoger los hiperparámetros de un modelo. Para hacerlo, se prueban valores candidatos de ellos al entrenar los modelos, y se eligen los valores para los que el error de validación cruzada sea menor. En nuestro modelo existen tres hiperparámetros que hay que seleccionar (α, λ y el radio de contacto R_c).

Para seleccionar los valores candidatos, usamos 300 iteraciones del algoritmo de *simulated annealing* con la librería de Python `hyperopt` (*hyperparameter optimization*) (Bergstra et al. [2]). Los rangos de optimización escogidos para los hiperparámetros se detallan en la Tabla 4.1.

Los valores candidatos escogidos por `hyperopt` fueron evaluados mediante una validación cruzada. Aunque es posible usar *log loss* como medida de error, encontramos mejo-

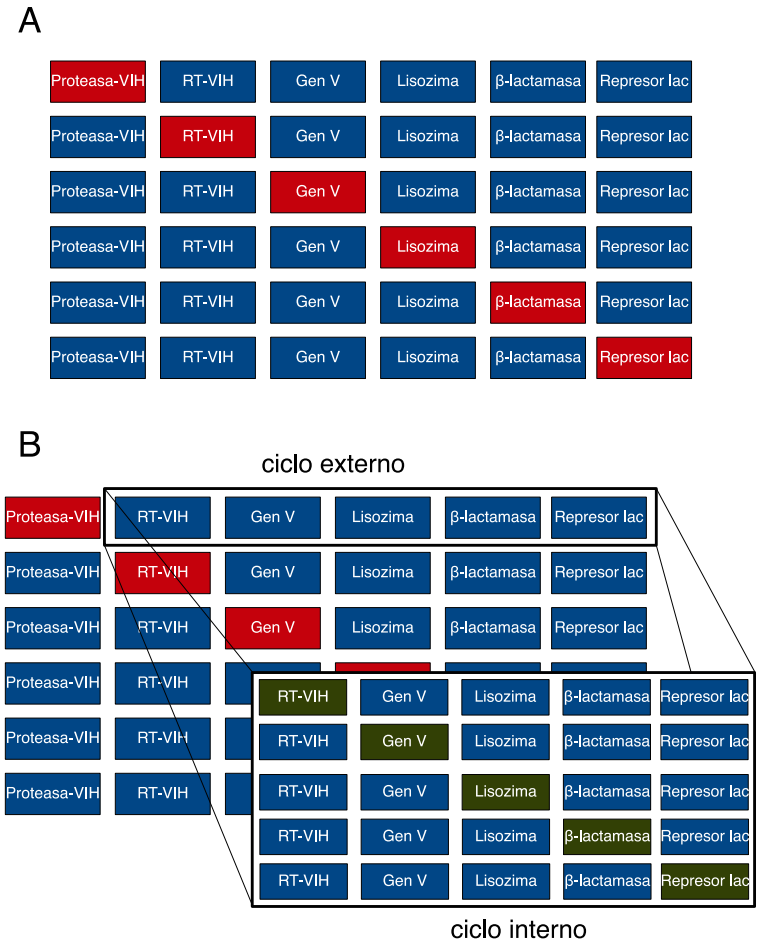


Figura 4.1: VALIDACIÓN CRUZADA ANIDADA A. Validación cruzada por proteína. Se hicieron seis ciclos de ajuste del modelo. La proteína de validación se muestra en rojo. B. Validación cruzada anidada. En cada ciclo se separa una *proteína de validación externa* (roja), que se usa para obtener un estimador no sesgado de E_{out}. Se hace un ciclo de validación cruzada *interno* con el resto: cada proteína en turno se toma como *proteína de validación interna* (verde). La validación interna se usa para escoger los valores de los hiperparámetros del modelo.

Tabla 4.1: Se muestra el rango de valores entre los que se optimizaron los hiperparámetros R_c, α y λ, y la distribución con la que se muestrearon valores aleatorios con `hyperopt`. El radio de contacto sólo lo optimizamos entre valores que difieren en 0.25 Å.

HIPERPARÁMETRO	RANGO	DISTRIBUCIÓN
R_c	$(3, 9)$Å	Unif. discreta (paso de 0.25Å)
α	$(10^{-8}, 1)$	Log uniforme
λ	$(0, 1)$	Unif. contínua

res resultados usando el *coeficiente de correlación de Matthews (MCC):*

$$MCC = \frac{TP \times TN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.5}$$

Este coeficiente es una medida de la calidad de una predicción binaria que va de 0 a 1. Toma en cuenta los verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN) de la predicción del modelo. Los hiperparámetros de los modelos finales fueron elegidos como se describe a continuación:

Algoritmo para escoger hiperparámetros y ajustar modelo final

1. Escoger valores candidatos de α, λ y R_c. Esto lo hace `hyperopt` con el algoritmo de *simulated annealing*.

2. Para cada proteína:

   a) Llamarla *proteína de validación*.

b) Llamar al resto de las proteínas *conjunto de entrenamiento*.

c) Normalizar las variables del conjunto de entrenamiento con la Ecuación 4.4.

d) Ajustar un modelo de regresión logística de red elástica al conjunto de entrenamiento, minimizando el error aumentado (Ecuación 4.3). Los valores de α, λ y R_c se consideran fijos; lo que se ajusta son los coeficientes del modelo ($\mathbf{w}$).

e) Usar el modelo ajustado para predecir la criticalidad de los residuos de la proteína de validación. Se predicen como críticos si $h(\mathbf{x}_i) \geq 0.5$ y como no críticos si $h(\mathbf{x}_i) < 0.5$.

f) Calcular el coeficiente de correlación de Matthews de la predicción de los residuos de la proteína de validación (Ecuación 4.5).

3. Promediar los valores obtenidos del coeficiente de correlación de Matthews para todas las proteínas. Llamarlo MCC_{CV}, y calcular el *error de validación cruzada* $E_{CV} = 1 - MCC_{CV}$.

4. Pasar E_{CV} a `hyperopt` como medida de error. En base al error recibido, hyperopt calcula nuevos valores candidatos de α, λ y R_c.

5. Repetir el proceso 300 iteraciones.

6. Escoger los valores de α, λ y R_c que den lugar al menor E_{CV} (esto es equivalente a maximizar MCC_{CV}). No se maximizó MCC_{CV} directamente, pues `hyperopt` está diseñado para minimizar funciones.

Ya que se escogieron los valores de los hiperparámetros, se ajusta el modelo final, usando todas las proteínas como conjunto de entrenamiento.

Sin embargo, existe un problema: con este algoritmo no tenemos un estimador del rendimiento del mismo en datos nuevos (distintos a los de entrenamiento). Aunque es posible hacer otro procedimiento de validación cruzada ya habiendo escogido los hiperparámetros, hacerlo subestimaría el error real, pues estos valores fueron escogidos precisamente porque tienen buen rendimiento en los datos de la validación cruzada.

Por ello, fue necesario encontrar una manera distinta de estimar el error. Lo que hicimos para solucionar este problema fue seguir un procedimiento de *validación cruzada anidada*, en el que hay una *validación cruzada externa* y una segunda *validación cruzada interna* dentro de ella (Figura 4.1, panel B). Aunque aquí lo describimos antes por que esto hace la exposición de las ideas de este trabajo más sencilla, el procedimiento para ajustar el modelo final no fue usado inicialmente: primero se hizo la validación cruzada anidada detallada a continuación.

4.3.2 *Validación cruzada anidada*

En la validación cruzada anidada separamos cada proteína en turno. A la proteína separada la llamamos *proteína de validación externa* (esta es la proteína roja en la Figura 4.1, B). En el resto de las proteínas seguimos un procedimiento idéntico al anterior para escoger los hiperparámetros y ajustar el modelo con una diferencia importante: no se usa la proteína de validación externa en ningún paso. Después evaluamos el rendimiento del modelo con tres métricas distintas en la proteína de validación externa:

A. *Log loss* (con la Ecuación 4.2),

B. La precisión de predicción, calculada como

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

(en dónde TP: verdaderos positivos, TN: verdaderos negativos, FP: falsos positivos, y FN: falsos negativos), y

C. El coeficiente de correlación de Matthews (con la Ecuación 4.5).

Este procedimiento se repite para cada proteína de validación externa, y al final promediamos los valores obtenidos para cada proteína. Como los parámetros e hiperparámetros del modelo no se ajustan a las proteínas de validación externa, ésta sirve como un estimador del error fuera de los datos de entrenamiento. El algoritmo completo para obtener los estimadores entonces es:

Algoritmo para estimar rendimiento del modelo

1. Para cada proteína:
   a) Llamarla *proteína de validación externa* (Figura 4.1, B, roja). Llamar al resto de las proteínas *conjunto de entrenamiento externo*.
   b) Escoger valores candidatos de α, λ y R_c. Esto lo hace `hyperopt` con el algoritmo de *simulated annealing*.
   c) Para cada proteína del conjunto de entrenamiento externo:
      1) Llamarla *proteína de validación interna* (Figura 4.1, B, verde).

2) Llamar al resto *conjunto de entrenamiento interno*.

3) Normalizar las variables del conjunto de entrenamiento interno (Ecuación 4.4).

4) Ajustar un modelo de regresión logística de red elástica al conjunto de entrenamiento interno, minimizando el error aumentado (Ecuación 4.3).

5) Usar el modelo ajustado para predecir la criticalidad de los residuos de la proteína de validación. Se predicen como críticos si $h(\mathbf{x}_i) \geq 0.5$ y como no críticos si $h(\mathbf{x}_i) < 0.5$.

6) Calcular el coeficiente de correlación de Matthews de la predicción de los residuos de la proteína de validación interna (Ecuación 4.5).

d) Promediar los valores obtenidos del coeficiente de correlación de Matthews para todas las proteínas. Llamarlo MCC_{int}, y calcular el *error de validación cruzada interna* $E_{int} = 1 - MCC_{int}$.

e) Pasar E_{int} a `hyperopt` como medida de error. En base al error recibido, hyperopt calcula nuevos valores candidatos de α, λ y R_c.

f) Repetir el proceso 300 iteraciones.

g) Escoger los valores de α, λ y R_c que den lugar al menor E_{int}.

h) Normalizar las variables del conjunto de entrenamiento externo.

i) Ajustar un modelo de regresión logística de red elástica con los valores escogidos de α, λ y R_c al

conjunto de entrenamiento *externo*, minimizando el error aumentado (Ecuación 4.3).

j) Evaluar medidas de error (*log loss*, *Acc* y *MCC*) en la proteína de validación externa.

2. Promediar los valores de las medidas de error calculados para cada proteína en la validación externa para obtener los estimadores finales.

Primero obtuvimos los estimadores del rendimiento del modelo con este algoritmo, y sólo una vez que se tienen los estimadores se ajusta el modelo final con el algoritmo anterior, que es equivalente al ciclo interno de la validación cruzada anidada, pero usando todas las proteínas.

Parte III

RESULTADOS Y ANÁLISIS

5

RESULTADOS

En este trabajo desarrollamos modelos para predecir los *residuos críticos* de una proteína (es decir, los que son más sensibles a mutaciones, en el sentido de cambiar más frecuentemente la función de la proteína al ser mutados). Basamos las predicciones de los modelos en distintas combinaciones de atributos que dependen de la estructura primaria, secundaria, terciaria y cuaternaria de la proteína (Figura 5.1).

La hidrofobicidad del residuo, su estructura secundaria y el área accesible al solvente son ejemplos del tipo de variables que utilizamos. La descripción completa de los atributos estructurales usados en los modelos se encuentra en la Sección 3.3, en el capítulo de Materiales y Métodos.

Algunos de los atributos más interesantes fueron calculados a partir de la *red de contactos* de la proteína. Esta red se construye con una estructura tridimensional: cada residuo se representa por un nodo y los residuos que tienen al menos un par de átomos a una distancia menor a R_c, el *radio de contacto* son unidos por una arista. Se dice que los residuos que cumplen con este criterio están *en contacto*.

La red de contactos se usa para calcular medidas de centralidad, que tienen como objetivo encontrar los nodos más importantes de la red (para distintas definiciones de importancia). En trabajos previos [1, 37, 5] se ha visto que algunas medidas de centralidad pueden usarse para predecir residuos críticos.

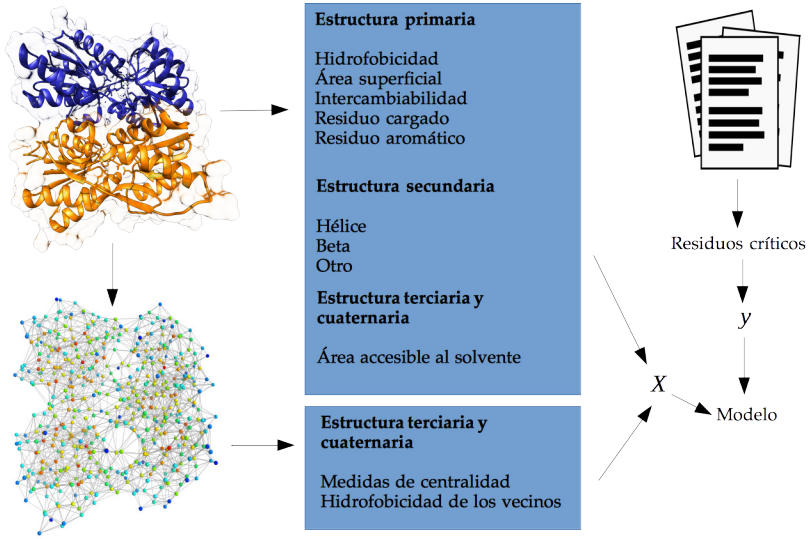


Figura 5.1: RESUMEN DE LA ELABORACIÓN DE LOS MODELOS
Se obtuvieron atributos estructurales dependientes de la estructura primaria, secundaria y terciaria/cuaternaria de seis proteínas. Algunos de ellos se obtuvieron directamente de la estructura y otros a través de la red de contactos. Los atributos fueron usados para ajustar un modelo de regresión logística: se codificaron en la matriz X, que se usa junto con el vector de residuos críticos y (obtenidos de trabajos de mutagénesis) para ajustar el modelo.

En este trabajo desarrollamos modelos usando más de una medida de centralidad simultáneamente, y en combinación con otros atributos estructurales para evaluar si las medidas de centralidad proveen información independiente a la de otros atributos estructurales. Distintas combinaciones de atributos se utilizaron para ajustar modelos de regresión logística regularizados (como se describe en el Capítulo 4), con el fin de predecir los residuos críticos de seis proteínas que han sido extensamente mutadas.

Primero mostramos algunos resultados sobre la correlación de las medidas de centralidad y otros atributos estructurales. Procedemos a describir el rendimiento de los modelos ajustados y a compararlos con `JAMMING`, un programa diseñado para predecir residuos críticos con una medida de centralidad.

5.1 CORRELACIONES ENTRE ATRIBUTOS ESTRUCTURALES

Se sabe que los residuos poco accesibles al solvente tienden a ser más sensibles a mutaciones que los expuestos, y que algunas medidas de centralidad calculadas a través de la red de contactos, como la cercanía *(closeness)* y la intermediación (*betweenness*) sirven para predecir residuos críticos. En trabajos previos se ha encontrado una correlación elevada entre algunas de las variables que usaremos en el modelo (por ejemplo, Fajardo y Fiser [12] encontraron correlaciones altas entre el área accesible al solvente y *closeness*, *betweenness* y *pagerank*). Por ello evaluamos las correlaciones existentes entre el área accesible al solvente y las medidas de centralidad para las estructuras de las proteínas que usaremos para construir el modelo (Figura 5.2).

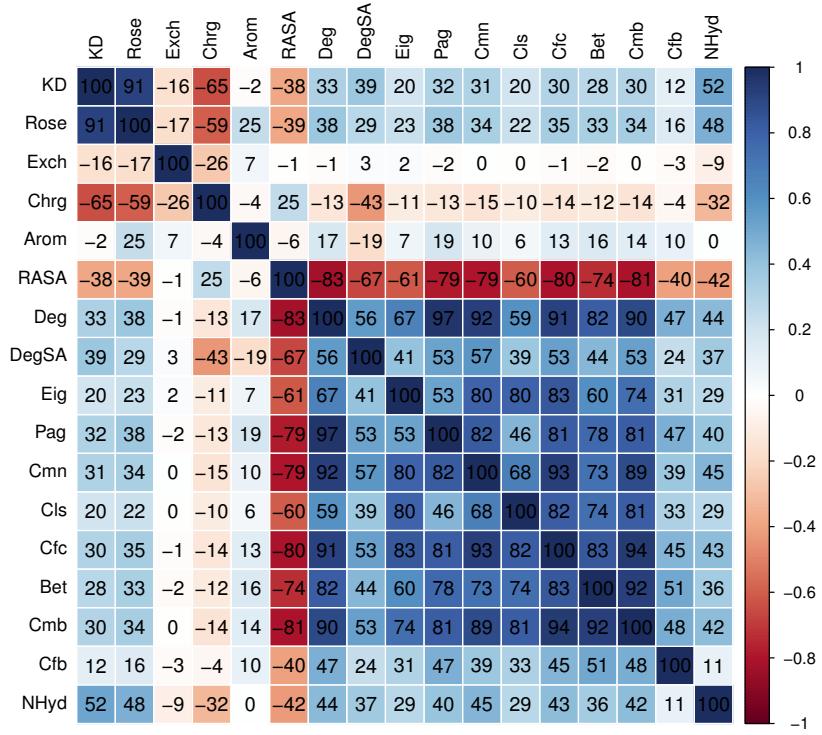


Figura 5.2: CORRELACIONES Se muestra la correlación entre el área accesible al solvente ($RASA$), medidas de centralidad normalizadas (Deg-Cfb), parámetros dependientes de la estructura primaria (KD-Arom), y la hidrofobicidad de los vecinos en la red de contacto (N_{Hyd}). Se muestra la correlación promedio para los residuos de las seis proteínas evaluadas. Las correlaciones positivas se muestran en azul y las negativas en rojo.

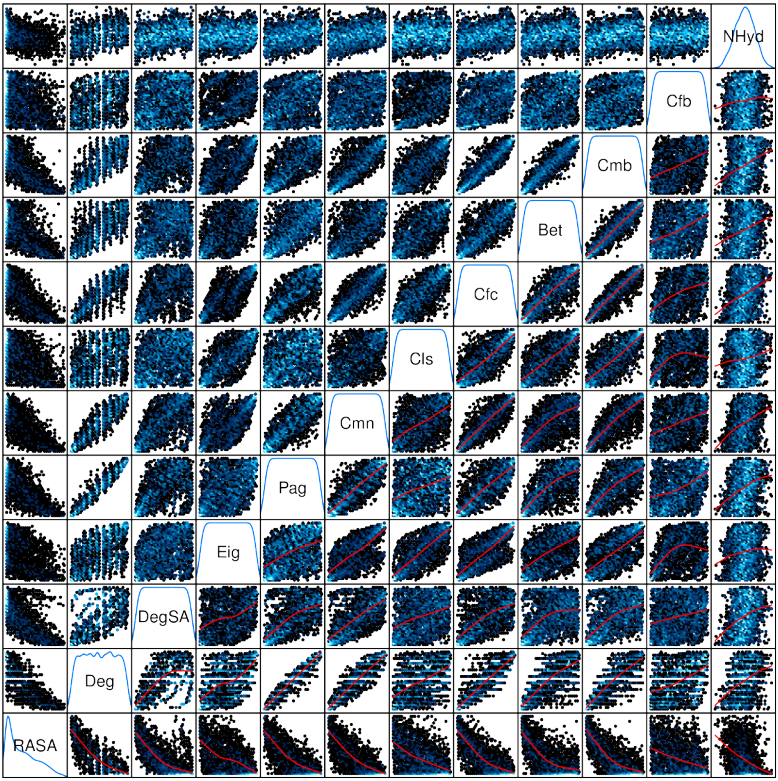


Figura 5.3: MATRIZ DE GRÁFICAS DE DISPERSIÓN Se muestra la gráfica de dispersión para cada par de variables en el cuadrante correspondiente. Las regiones con alta densidad de puntos se grafican de color azul. En los cuadrantes inferiores se grafica una regresión local *loess* para observar la tendencia de los datos.

Al igual que Fajardo y Fiser, encontramos una correlación elevada entre el área accesible al solvente y medidas de centralidad. Esto es cierto no sólo para *pagerank*, *closeness* y *betweenness*, sino para todas las medidas de centralidad evaluadas. De hecho, encontramos correlaciones aún más elevadas que ellos, probablemente debido a que normalizamos las medidas de centralidad para cada proteína por rango (ver la Subsección 3.3.3). Sin embargo, el que dos variables tengan una alta correlación no significa necesariamente que sean redundantes. Las diferencias entre variables altamente correlacionadas pueden ser significativas y dar información importante para la predicción.

Para explorar la relación entre el área accesible al solvente y las centralidades a un mayor detalle, graficamos una matriz de gráficas de dispersión (Figura 5.3) de los datos de las seis proteínas. Podemos observar que a pesar de las correlaciones elevadas existe un gran número de puntos en los que los valores de las centralidades son muy distintos a lo que se esperaría por el área accesible al solvente. En la siguiente sección, exploramos algunas medidas de centralidad visualmente para tratar de determinar si existe algún patrón visible en la estructura tridimensional de la proteína.

5.2 EXPLORACIÓN VISUAL DE MEDIDAS DE CENTRALIDAD EN REDES DE CONTACTOS

La razón por la que las medidas de centralidad sirven como predictores de residuos críticos es poco clara. Por ello, y para explorar sus diferencias con el área accesible al solvente ($RASA$) decidimos explorar las redes de contacto en una visualización tridimensional en la que cada nodo de la red es colocado en el centro de masa del residuo corresponiente (como se describe en la Sección 3.2). El objetivo de ello es

buscar si existe algún patrón discernible en la posición de los residuos con centralidades altas en la proteína.

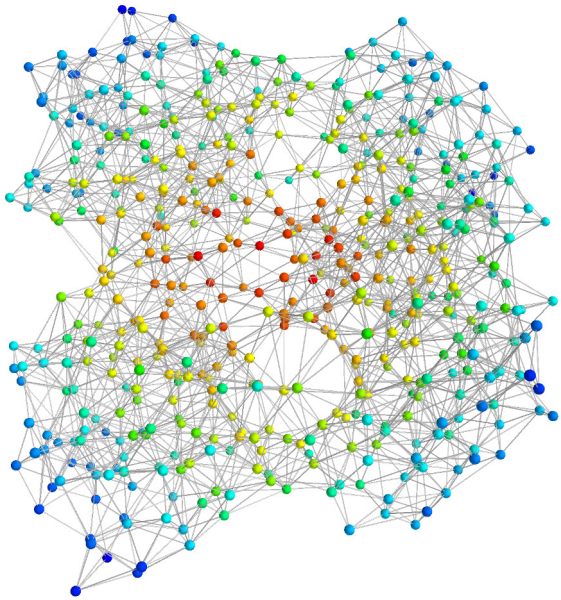


Figura 5.4: CENTRALIDAD DE CERCANÍA Se muestra la red de contactos del represor Lac. El valor de la *centralidad de cercanía Cls* de cada residuo se representa con un gradiente de color que va desde azul (valores bajos) a rojo (valores altos).

Por tratarse de un numero relativamente elevado de imágenes, no se reproducen aquí las visualizaciones de todas las medidas de centralidad. Nos enfocamos en tres de ellas en esta sección por su importancia en los modelos (ver más adelante): la cercanía (Cls), la intermediación de comunicabilidad (Cmb) y el grado normalizado (Deg/SA). El resto de las visualizaciones en esta red se encuentran en el apéndice de este documento (Subsección A.1.2).

La medida de centralidad para la que se observa un patrón más claro es la cercanía (Figura 5.4). Los residuos con cercanía elevada parecen ser los que están más cerca del centro (geométrico) de la proteína. Esta propiedad no es equivalente a la que mide el área accesible al solvente (Figura 5.5); pueden existir residuos muy accesibles o muy poco accesibles al solvente cercanos al centro de la proteína. Sin embargo, la alta correlación entre ambas variables puede entenderse: el centro de la proteína no está expuesto al solvente, y por lo tanto, los residuos cercanos al centro tenderán a estar menos expuestos al solvente que los lejanos en promedio.

La intermediación de comunicabilidad (Cmb) también tiene cierta preferencia por residuos cercanos al centro de la proteína (Figura 5.6), aunque en menor grado que Cls. En cambio, parece dar un mayor énfasis a los residuos en el "centro" de partes de la proteína densamente conectadas (Figura 5.7) o que funcionan como "puentes" entre ellas. Esto tiene sentido considerando la definición de la intermediación de comunicabilidad, que da puntuaciones altas a nodos que están en caminos cortos entre todos los pares de nodos de la red.

El grado de un nodo es el número de vecinos que tiene en la red. En nuestro caso, el grado de un residuo (Deg) es el número de residuos con los que hace contacto. Sin embargo, algunos residuos son más grandes que otros, y por esa razón tenderán a tener más vecinos que los residuos más pequeños. Para eliminar esta dependencia del grado en el tamaño del residuo, se dividió el valor del grado entre el área superficial del residuo antes de ser normalizado entre cero y uno (Deg/SA). Se muestra la visualización de esta medida de centralidad en la Figura 5.8.

En relación a las medidas anteriores, existe una mucho menor preferencia al centro de la proteína (pues hay una cavidad), y se tiende a preferir residuos lejanos a la superficie de la proteína en partes densamente empacadas (en dónde hay muchos contactos entre residuos). También se observa que, en comparación con el área accesible al solvente, tiene un mayor rango dinámico. Es decir, hay mayores diferencias entre los valores de Deg/SA que de $RASA$ entre diferentes residuos.

5.3 MODELOS DE REGRESIÓN LOGÍSTICA

Incluímos diferentes combinaciones de atributos que dependen de la estructura primaria, secundaria y terciaria de la proteína en tres modelos de regresión logística regularizados por red elástica.

En primer lugar quisimos confirmar si la red de contactos, y en particular las medidas de centralidad, proveen información para predecir residuos críticos. Para ello ajustamos el *modelo de centralidades* (CEN), en el que únicamente incluímos atributos que corresponden a medidas de centralidad calculadas de la red de contactos de las proteínas. Los atributos incluidos fueron el grado y grado normalizado (Deg, Deg/SA), la centralidad de eigenvector (Eig), la comunicabilidad (Cmn), las cercanías (Cls y Cfc) y las intermediaciones (Bet, Cmb y Cfc). Este modelo logra una precisión de predicción de 74.4 % y un MCC de 0.364, lo que es mejor que el azar (Tabla 5.1).

También ajustamos un *modelo sin redes de contacto* (SRC), en el que omitimos todos los atributos que se calculan mediante la red de contactos de la proteína (esto es, todas las medidas de centralidad y la hidrofobicidad de los vecinos). En este modelo incluímos todos los demás atributos men-

Tabla 5.1: CALIDAD DEL AJUSTE DE LOS MODELOS La calidad del ajuste de los tres modelos (CEN: sólo medidas de centralidad; SRC: sin redes de contacto; Completo: modelo que incluye los atributos de los modelos CEN y SRC) se evaluó en una validación cruzada anidada con tres métricas: *log loss*, la precisión de la predicción (*Acc*) y el coeficiente de correlación de Matthews (*MCC*). Se muestran los errores calculados en la validación externa.

MÉTRICA	CEN	SRC	COMPLETO
Log loss	0.787	0.578	0.508
Acc	0.744	0.747	0.779
MCC	0.364	0.361	0.421

cionados en la Sección 3.3, que son principalmente atributos estructurales más tradicionales, como la hidrofobicidad del residuo, la estructura secundaria y el área accesible al solvente (*RASA*). El modelo SRC tuvo una precisión y un MCC muy similares a los de CEN. Sin embargo, tuvo un error *log loss* menor (0.578, comparado con 0.787 para CEN), lo que indica que las probabilidades predichas están mejor calibradas (es decir, son más cercanas a las reales en promedio) en el modelo SRC que en CEN.

Para evaluar si los atributos dependientes de la red de contacto dan información independiente a la del modelo SRC para la predicción de residuos críticos se ajustó un tercer modelo, el *modelo completo*, en el que se incluyeron todos los atributos de la Sección 3.3. Esto incluye todos los atributos que se usaron en los modelos CEN y SRC. Tanto la precisión como el MCC mejoraron, mientras que el *log*

Tabla 5.2: HIPERPARÁMETROS Se muestran los valores finales de los hiperparámetros para los tres modelos.

HIPERPARÁMETRO	CEN	SRC	COMPLETO
R_c [Å]	4.5	–	4.5
α	7.235	3.916	1.490
λ	0.317	0.290	0.250

loss del modelo disminuyó con respecto a ambos modelos anteriores, lo que denota probabilidades mejor calibradas.

El radio de contacto elegido fue el mismo para los modelos CEN y completo (4.5 Å, Tabla 5.2). En todos los modelos, se eligió una regularización mixta de tipo $l1$ y $l2$, con un predominio de regularización $l2$. Sin embargo, la regularización $l1$ tuvo un efecto importante, como puede observarse en que los coeficientes ajustados fueron exactamente cero en algunos casos (ver más adelante).

5.3.1 *Modelo de medidas de centralidad*

Se escogieron valores más altos de α y λ en el modelo CEN en comparación con los otros dos modelos, lo que corresponde a una mayor regularización total y a una mayor proporción de tipo $l1$. Esto sugiere que existen más variables redundantes entre las centralidades que entre el resto de los atributos. De hecho, a cuatro medidas de centralidad (Deg, Pag, Cmn y Cfc) se les asignó un coeficiente de cero (Tabla 5.3), por lo que no tienen efecto en el modelo. Otras tres (Cls, Bet y C_{Frac}) tienen un coeficiente estandarizado menor a 0.1.

Tabla 5.3: COEFICIENTES DEL MODELO DE MEDIDAS DE CENTRALIDAD Se muestran los coeficientes ajustados originales, así como los coeficientes normalizados por la desviación estándar de la variable.

VARIABLE	COEFICIENTE	C. ESTANDARIZADO
ordenada	-3.265	-1.278
Deg	0.000	0.000
Eig	-0.194	-0.056
Pag	0.000	0.000
Cmn	0.000	0.000
Cls	0.257	0.074
Cfc	0.000	0.000
Bet	0.318	0.092
Cmb	1.170	0.338
Cfb	-0.410	-0.118
C_{Frac}	16.631	0.057
Deg/SA	1.342	0.388
$Cmb \times Cls$	1.696	0.477

Los efectos principales en este modelo se encuentran mediados por la la intermediación de comunicabilidad (Cmb), el grado normalizado (Deg/SA) y la interacción entre Cmb y la centralidad de cercanía ($Cmb \times Cls$). La intermediación de comunicabilidad es una medida que se calcula removiendo el nodo correspondiente en la red y determinando el cambio de conectividad causado por ello. Puede interpretarse como la sensibilidad de la red a perturbaciones pequeñas a las aristas del nodo (Estrada et al. [11]). En las redes de interacción de residuos esto corresponde a perturbaciones en los contactos entre residuos.

El grado normalizado es simplemente el número de vecinos en la red del residuo dividido entre su área superficial. Como incluímos aristas en la red dependiendo de la distancia entre pares de átomos de los dos residuos correspondientes, los residuos con mayor área superficial tenderán a tener más contactos en la red simplemente por virtud de su tamaño. El grado normalizado tiene como objetivo remover esta dependencia, y es un mejor predictor que los valores directos del grado del residuo (comparar los coeficientes de Deg y Deg/SA).

En las visualizaciones observamos que la centralidad de cercanía está relacionada con la posición del residuo con respecto al centro geométrico de la proteína. En nuestros resultados, la centralidad no fue muy importante por sí misma, pero sí en combinación con Cmb. Los residuos que se predicen con una mayor probabilidad de ser críticos son aquellos que tienen valores altos para tanto Cmb como Cls. El modelo CEN se visualiza en la Figura 5.9.

5.3.2 *Modelo sin atributos de la red de contactos*

En este modelo se omitieron todos los atributos calculados a partir de la red de contactos. Por ello no fue necesario calcular un valor de R_c. Hubo menos variables redundantes en el modelo en comparación al modelo CEN: únicamente se le asignó un coeficiente de cero a dos interacciones (Tabla 5.4).

Aunque el modelo SRC ajusta 17 atributos en total, no son todos independientes entre sí. De hecho, los atributos dependen únicamente del tipo de residuo (Hyd_{KD}, Hyd_{Rose}, SA, $Exch$, $Chrg$, $Arom$), estructura secundaria (Hel, $Beta$), área accesible al solvente ($RASA$) y combinaciones entre ellos ($H \times Hp$, $B \times Bp$ y el resto de las interacciones). Esto hizo posible graficar las predicciones del modelo en la Figura 5.11. Aunque no usamos la red de contactos en este modelo, la visualizamos también para poder comparar este modelo con los demás en la Figura 5.10.

En el modelo SRC se confirma el efecto del área accesible al solvente ($RASA$) observado experimentalmente: residuos menos expuestos tienden a ser menos tolerantes a mutaciones. Esta fue la variable con mayor efecto en la probabilidad predicha. Sin embargo, sus interacciones con otras variables mostraron efectos importantes. En particular, la probabilidad de que un residuo no expuesto al solvente sea crítico aumenta considerablemente si el residuo es cargado o aromático.

La hidrofobicidad de Rose et al. [32] tuvo un efecto mayor en el modelo que la hidrofobicidad de Kyte y Doolittle [22]. Esta diferencia es más importante en el contexto del área accesible al solvente: la escala de Kyte y Doolittle no tiene un efecto en interacción con $RASA$, mientras que residuos

Tabla 5.4: COEFICIENTES DEL MODELO SRC Se muestran los coeficientes originales, así como normalizados por la desviación estándar de la variable correspondiente.

VARIABLE	COEFICIENTE	C. ESTANDARIZADO
ordenada	4.419	-1.868
Hyd_{KD}	-0.077	-0.240
Hyd_{Rose}	-3.394	-0.370
SA	-0.340	-0.140
$Exch$	-1.150	-0.267
$Chrg$	0.729	0.309
$Arom$	0.844	0.240
Hel	-0.296	-0.121
$Beta$	1.230	0.515
$RASA$	-2.922	-0.746
$Hel \times H_p$	0.000	0.000
$Beta \times B_p$	-0.761	-0.416
$Hyd_{KD} \times RASA$	0.000	0.000
$Hyd_{Rose} \times RASA$	-3.886	-0.672
$Chrg \times RASA$	-5.959	-1.201
$Arom \times RASA$	-5.199	-0.447
$Hel \times RASA$	-3.694	-0.644
$Beta \times RASA$	-3.971	-0.504

hidrofílicos según la escala de Rose tienden a ser críticos cuando estan poco expuestos al solvente.

La estructura secundaria del residuo tiene efectos distintos a distinta *RASA*. Se predice una mayor probabilidad de ser críticos para los residuos poco accesibles que forman parte de una hoja u hebra β. En cambio, para residuos expuestos la mayor probabilidad se predice para los que no forman parte de una hélice o hébra. Los residuos que participan en una hélice son predichos como críticos con una menor probabilidad que los demás a todos los niveles de *RASA*.

El efecto de la estructura secundaria en la predicción no sólo depende de *RASA*: también es dependiente del tipo de residuo. Se observan esencialmente tres comportamientos. En residuos cargados el efecto de la estructura secundaria es menor que en los demás, pues el efecto del área accesible al solvente es mucho más fuerte. Los residuos cargados (D, E, R, K y H) son predichos como críticos cuando son poco accesibles, y como no críticos cuando son muy accesibles al solvente sin importar su estructura secundaria.

Sin embargo, las diferencias entre estructuras secundarias pueden observarse a niveles intermedios de *RASA* (entre 0.1 y 0.3). Un efecto similar se observa en residuos aromáticos (F, Y, W y H), en dónde el efecto de *RASA* también es mucho más importante que el de la estructura secundaria. En el caso de la histidina, que es un residuo cargado y aromático, este comportamiento es aún más marcado.

En los demás residuos, el efecto de la estructura secundaria es más importante. La mayor diferencia en la predicción para distintas estructuras secundarias se encuentra en residuos relativamente hidrofílicos (por ejemplo N, Q, S, T, G). Los residuos hidrofóbicos también muestran diferencias entre estructuras secundarias, pero en general las probabi-

lidades predichas son bastante menores para los residuos poco expuestos. Mientras que se predice que la probabilidad para residuos hidrofílicos poco expuestos es cercana a 0.7, para los residuos hidrofóbicos sólo llega a valores cercanos a 0.5. De hecho, el residuo más hidrofóbico, la valina, no es predicho como crítico por este modelo en ninguna situación a un valor de corte de la probabilidad de 0.5.

5.3.3 *Modelo completo*

Los coeficientes ajustados del modelo completo se encuentran en la Tabla 5.5. Se incluyeron todos los atributos mencionados en la Sección 3.3, que contienen a todos los usados en los modelos SEC y CEN.

Tabla 5.5: COEFICIENTES DEL MODELO COMPLETO Se muestran los coeficientes ajustados del modelo completo, así como los coeficientes normalizados por la desviación estándar de la variable.

VARIABLE	COEFICIENTE	C. ESTANDARIZADO
ordenada	2.629	-4.834
Hyd_{KD}	-0.034	-0.105
Hyd_{Rose}	-15.136	-1.650
SA	2.035	0.839
Exch	-2.436	-0.566
Chrg	2.903	1.230
Arom	5.145	1.461
Hel	-0.027	-0.011
Beta	3.153	1.320
RASA	-5.640	-1.441

VARIABLE	COEFICIENTE	C. ESTANDARIZADO
Deg	-4.719	-1.362
Eig	-1.683	-0.486
Pag	-0.845	-0.244
Cmn	4.924	1.422
Cls	0.000	0.000
Cfc	-6.215	-1.794
Bet	0.000	0.000
Cmb	8.211	2.370
Cfb	1.082	0.312
N_{Hyd}	-0.432	-0.469
C_{Frac}	-7.807	-0.027
$Hel \times H_p$	-2.668	-1.204
$Beta \times B_p$	-2.521	-1.378
$Hyd_{KD} \times RASA$	-0.353	-0.401
$Hyd_{Rose} \times RASA$	-2.489	-0.430
$Chrg \times RASA$	-23.608	-4.756
$Arom \times RASA$	-8.987	-0.772
Deg/SA	7.609	2.196
$Hel \times RASA$	-9.263	-1.616
$Beta \times RASA$	-7.337	-0.932
$Cmb \times Cls$	9.637	2.711

Al igual que en el modelo SEC, existe un efecto muy grande del área accesible al solvente en residuos cargados (ver el coeficiente $Chrg \times RASA$). Sin embargo, el efecto de los residuos aromáticos es menor. Esto podría deberse a que los

residuos aromáticos suelen tener un área superficial grande, por lo que tienden a tener más vecinos en la red de contactos y centralidades más elevadas que el promedio. Como en el modelo CEN, hubo un efecto grande de *Cmb*, y de la interacción $Cmb \times Cls$. En contraste con el modelo SRC, la propensidad de que un residuo forme hélice y de hoja beta afectaron de forma importante la predicción.

Para comparar más a detalle el modelo completo con el modelo SRC se hizo un gráfico equivalente al mostrado anteriormente, en el que la probabilidad predicha se grafica contra el área accesible al solvente para cada tipo de residuo y estructura secundaria (Figura 5.14). Sin embargo, en el modelo completo estos no son los únicos parámetros del modelo: también lo son las medidas de centralidad y la hidrofobicidad de los vecinos.

Por ello se graficó dejando estos parámetros constantes. La hidrofobicidad de los vecinos se dejó al valor promedio en todos los casos, mientras que se graficó cómo varía la probabilidad cuando todas las centralidades se dejan simultáneamente constantes en el cuantil 25, 50 (que corresponde a la mediana) y 75, respectivamente.

Se observa que para valores bajos de las centralidades (cuantil 25) sólo los residuos cargados (y algunos aromáticos) poco accesibles al solvente son predichos como críticos. Existe una preferencia hacia los residuos con una estructura secundaria distinta a la de hélice. Para valores medios (cuantil 50, la mediana), los residuos fuertemente hidrofílicos (como N y Q) y todos los aromáticos también se predicen como críticos si están poco expuestos al solvente, pero esto no es el caso para los residuos hidrofóbicos o sólo moderadamente hidrofílicos.

Para comparar las diferencias entre los modelos SRC y completo debidas a la inclusión de las medidas dependien-

tes de la red de contactos se visualizó la diferencias en la probabilidad predicha entre ambos modelos (Figura 5.13). En comparación con el modelo SRC, el modelo completo tiende a predecir como críticos los residuos que se encuentran dentro de regiones densamente conectadas, los que se encuentran conectando regiones densamente conectadas entre sí y los que se encuentran cercanos al centro de la proteína. En contraste, el modelo SRC tiene una mayor preferencia por residuos poco accesibles al solvente en regiones alejadas del centro y/o que no están conectadas densamente.

5.4 COMPARACIÓN CON JAMMING

Uno de los objetivos de este trabajo fue mejorar la predicción de JAMMING, un programa previamente desarrollado en el grupo del Dr. Gabriel Del Río para predecir residuos críticos. Este programa está basado en una sóla medida de centralidad: la conectividad dinámica, estrechamente relacionada con la intermediación de caminos cortos (Bet).

Para comparar los modelos desarrollados en este trabajo con JAMMING, la sensibilidad, especificidad y el coeficiente de correlación de Matthews (MCC, Ecuación 4.5) de las predicciones de cada modelo fueron evaluadas en el conjunto de seis proteínas usado en este trabajo. Comparamos los resultados de la predicción de JAMMING con los valores equivalentes obtenidos en la validación cruzada externa para los modelos CEN, SRC y completo (Tabla 5.6 y Subsección A.2.1).

Observamos que se mejoró la predicción de JAMMING sustancialmente en los modelos desarrollados en este trabajo, evaluada tanto con presición como con el coeficiente de co-

Tabla 5.6: COMPARACIÓN DE JAMMING Y LOS MODELOS DESARROLLADOS EN ESTE TRABAJO Los valores de sensibilidad, especificidad y MCC fueron calculados para JAMMING y los modelos CEN, SRC y completo (en la validación cruzada externa para estos últimos). Se calcularon prediciendo como residuos críticos a aquellos con una probabilidad predicha mayor a 0.5.

MODELO	SENSIBILIDAD	ESPECIFICIDAD	MCC
JAMMING	0.275	0.873	0.200
CEN	0.514	0.847	0.364
SRC	0.478	0.857	0.361
Completo	0.453	0.919	0.421

rrelación de Matthews (como JAMMING no hace predicciones probabilísticas no se puede evaluar con *log loss*).

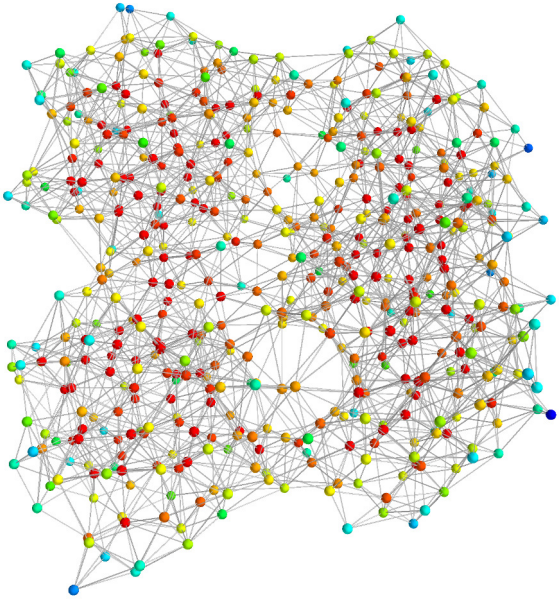


Figura 5.5: ÁREA ACCESIBLE AL SOLVENTE Se muestra la red de contactos del represor Lac. El valor del área accesible al solvente ($RASA$) se representa con un gradiente de color. Como la correlación de $RASA$ con las medidas de centralidad es negativa, para que esta visualización sea comparable la escala es inversa a la de las demás figuras: el gradiente va del color rojo (valores de $RASA$ bajos) al azul (valores elevados).

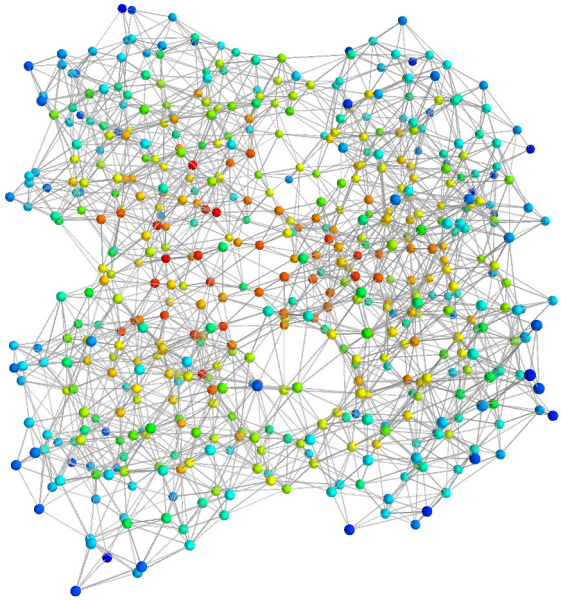


Figura 5.6: CENTRALIDAD DE INTERMEDIACIÓN DE COMUNICABILIDAD Se muestra la red de contactos del represor Lac. El valor de la *centralidad de intermediación de comunicabilidad Cmb* se representa con un gradiente de color que va desde azul (valores bajos) a rojo (valores altos).

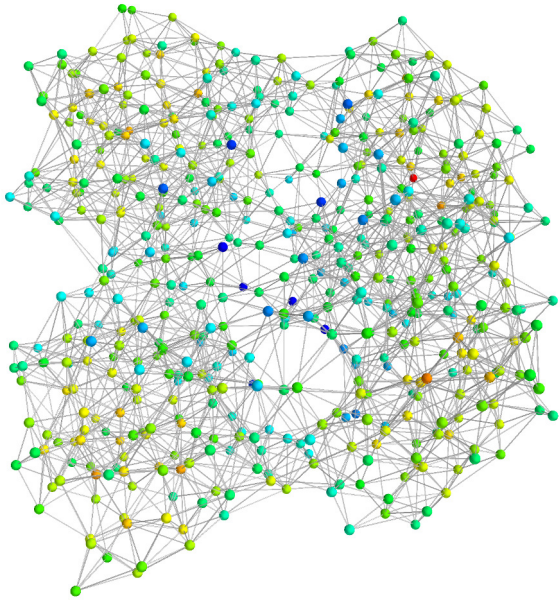


Figura 5.7: DIFERENCIA ENTRE CMB Y CLS Se muestra la red de contactos del represor Lac. El valor de $Cmb - Cls$ se representa con un gradiente de color que va desde azul (valores bajos) a rojo (valores altos).

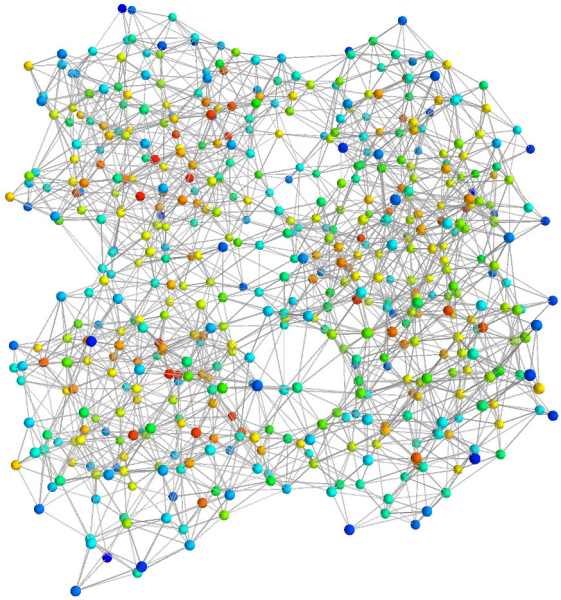


Figura 5.8: GRADO NORMALIZADO POR EL ÁREA DEL RESIDUO
Se muestra la red de contactos del represor Lac. El color del nodo representa el valor de Deg/SA: el gradiente de color va del azul (valores bajos) al rojo (valores altos).

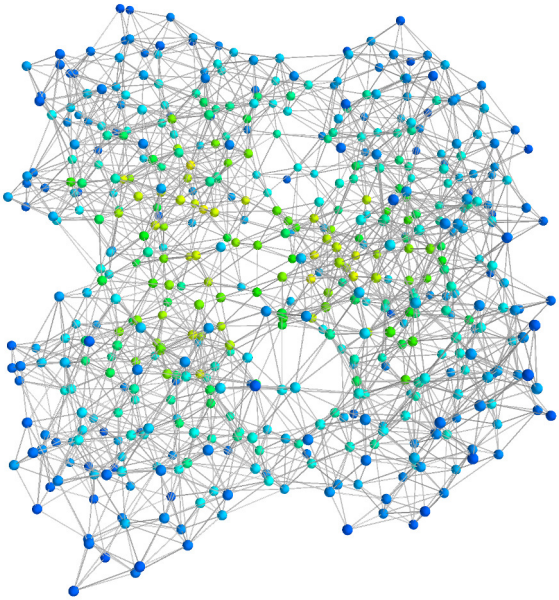


Figura 5.9: MODELO CEN Se muestra la red de contactos del represor Lac. Cada nodo de la red representa un residuo de la proteína. El color del nodo representa la probabilidad predicha por el modelo CEN de que el residuo sea crítico. La escala va de 0 (azul) a 1 (rojo). Los nodos verdes representan valores cercanos a 0.5.

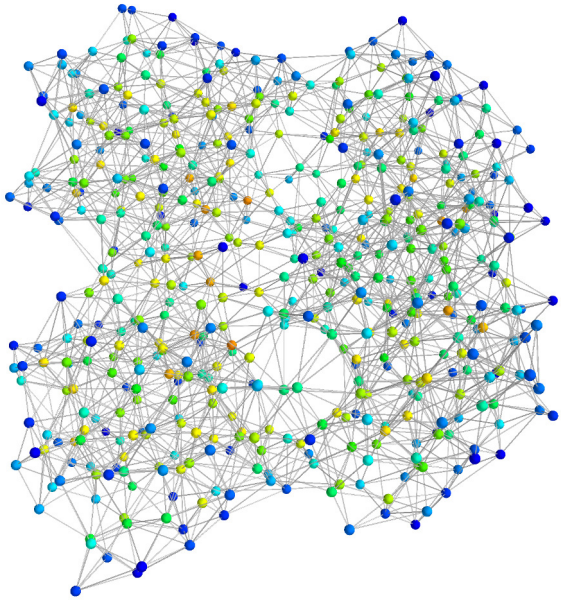


Figura 5.10: MODELO SRC Se muestra la red de contactos del represor Lac. El color del nodo representa la probabilidad predicha de que el residuo sea crítico. La escala va de 0 (azul) a 1 (rojo), los nodos verdes representan valores cercanos a 0.5.

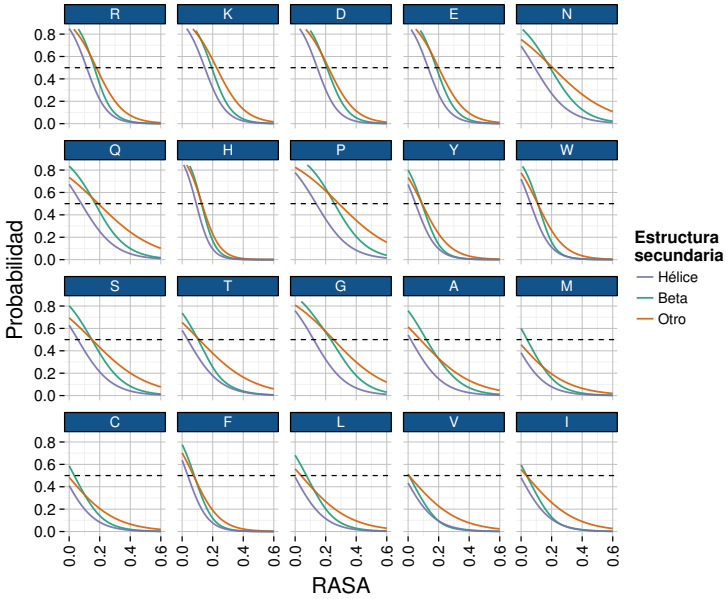


Figura 5.11: PROBABILIDADES DEL MODELO SRC Se grafica la probabilidad predicha por el modelo SRC de que un residuo sea crítico como función del área accesible al solvente ($RASA$) para cada tipo de residuo (código de una letra) y estructura secundaria (color). La línea punteada marca una probabilidad de 0.5, que es el valor de corte estándar en el modelo para predecir un residuo como crítico.

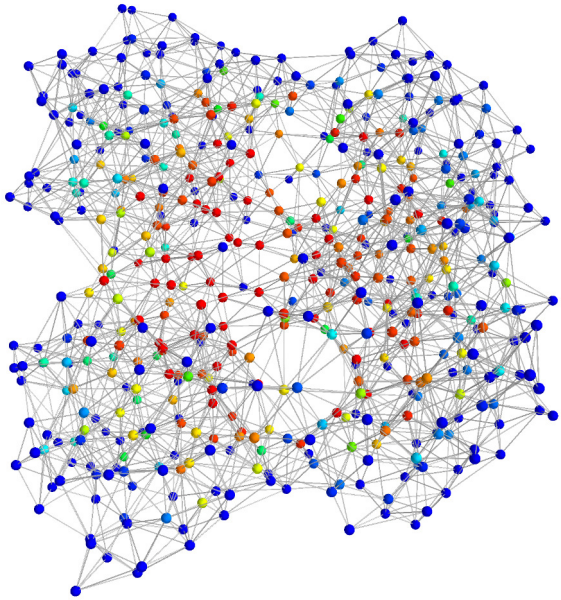


Figura 5.12: MODELO COMPLETO Se muestra la red de contactos del represor Lac. El color del nodo representa la probabilidad predicha de que el residuo sea crítico. La escala va de 0 (azul) a 1 (rojo), los nodos verdes representan valores cercanos a 0.5.

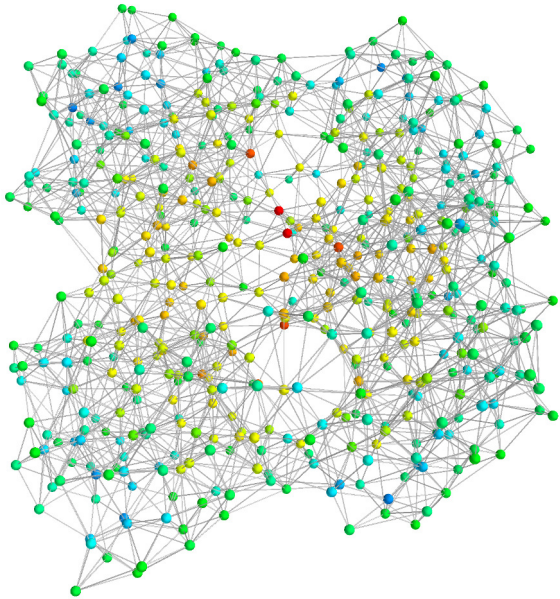


Figura 5.13: DIFERENCIA ENTRE LAS PROBABILIDADES PREDICHAS POR EL MODELO COMPLETO Y SRC El valor máximo de la diferencia corresponde a 0.726 (rojo intenso). Diferencias cercanas a 0 corresponden a un color verde.

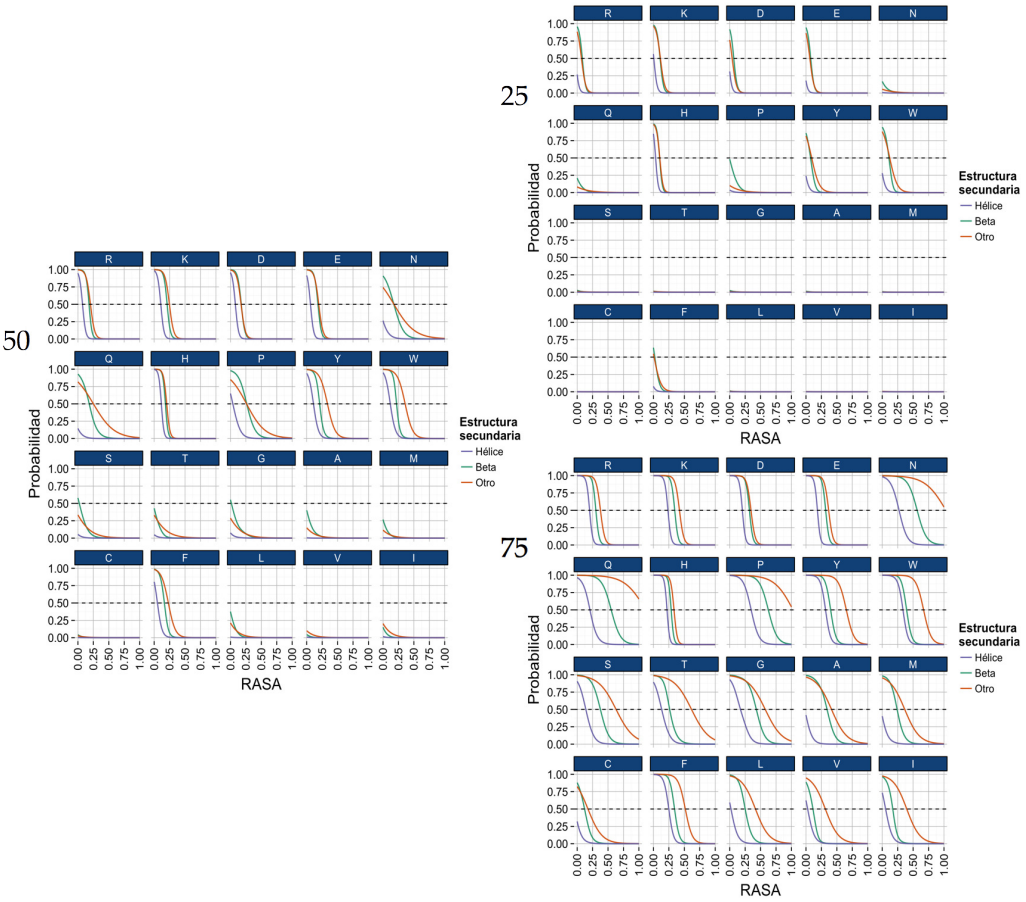


Figura 5.14: PROBABILIDADES DEL MODELO COMPLETO Se grafica la probabilidad de ser critico predicha como función de $RASA$ para cada tipo de residuo y estructura secundaria. Se dejan constantes la hidrofobicidad de los residuos (N_{Hyd}) en el valor promedio, y todas las centralidades, en el cuantil 25, 50 y 75.

6

ANÁLISIS DE RESULTADOS Y CONCLUSIONES

The purpose of computing is insight, not numbers.

— Richard Hamming

6.1 ANÁLISIS DE RESULTADOS

6.1.1 *Medidas de centralidad*

A pesar de la alta correlación entre las medidas de centralidad calculadas de la red de contacto de una proteína con el área accesible al solvente (Figura 5.2), al menos algunas medidas de centralidad parecen tener información independiente al área accesible al solvente y otros parámetros dependientes de la estructura primaria y secundaria de la proteína para la predicción de residuos críticos.

En primer lugar, tanto la exploración visual de redes de contacto (Sección 5.2 y A.1.2) como la gráfica de dispersión de las medidas de centralidad (Figura 5.3) sugieren que el área accesible al solvente y las medidas de centralidad no son equivalentes, aunque están fuertemente correlacionadas.

En segundo lugar, en el modelo completo tanto el área accesible al solvente ($RASA$) como algunas medidas de centralidad (como Cmb) tuvieron coeficientes estandarizados de tamaño relativamente elevado, lo que indica que tienen un efecto importante en la predicción del modelo (Ta-

bla 5.5). El tamaño de los coeficientes de los modelos es penalizado como parte del error aumentado en la regularización. Por ello, los coeficientes de variables redundantes o poco importantes tienden a ser reducidos, o incluso a volverse cero exactamente[1], lo que no sucedio para estas variables. En cambio, otras variables, como la centralidad de intermediación (*Bet*) tuvieron un coeficiente de cero, lo que sugiere que es redundante cuando se toman en cuenta los otros atributos.

Además, el modelo completo mejoró la predicción de los modelos CEN y SRC en todas las métricas evaluadas (Tabla 5.1). Esto sugiere que el modelo completo tiene más información para predecir residuos críticos de la que tienen los modelos CEN y SRC por separado, y que, por lo tanto, ambos conjuntos de atributos son relevantes y proveen cierta información independiente.

En conjunto la evidencia apunta a que las medidas de centralidad proveen información independiente al área accesible al solvente y el resto de los parámetros de estructura primaria y secundaria incluidos en el modelo SRC. Esto no concuerda con las observaciones hechas por Fajardo y Fiser [12] para la predicción de residuos catalíticos. Observamos también mayores correlaciones entre el área accesible y las medidas de centralidad que Fajardo y Fiser, probablemente debidas a que normalizamos las centralidades por rangos.

Una posible explicación a estas diferencias en nuestros resultados y los de Fajardo y Fiser es que los valores de algunas medidas de centralidad dependen del número de residuos de la estructura (Subsección A.1.1). Como tal, los valores crudos de las centralidades de distintas proteínas no

1 El que algunos coeficientes se vuelvan exactamente cero se debe el componente de penalización de la norma $l1$ de los coeficientes en el error aumentado.

son directamente comparables entre sí. Al ajustar un modelo con los datos sin normalizar se tenderán a predecir como críticos los residuos de algunas proteínas más que de otras, dependiendo del número de residuos que tengan. Esto pudiera disminuir el poder predictivo de las medidas de centralidad considerablemente, lo que explicaría que Fajardo y Fiser encontraran que no proveen información importante para la predicción independiente al área accesible al solvente. Por otra parte, el conjunto de residuos catalíticos es un subconjunto de los residuos críticos. Las diferencias entre los resultados también pudieran deberse a que las medidas de centralidad no sean buenos predictores de residuos catalíticos, pero sí de otro subconjunto de los residuos críticos.

6.1.2 *Variables estructurales tradicionales*

Los residuos poco expuestos al solvente son predichos con una mayor probabilidad de ser críticos que los expuestos en ambos modelos en los que se incluyó la accesibilidad al solvente como atributo (5.14 y Figura 5.11). Esta dependencia en el área accesible es más importante para los residuos cargados, en los que se observa un mayor efecto de $RASA$ en la probabilidad que el de cualquier otra variable en los modelos. En el resto de los residuos la dependencia es más suave, aunque es más importante para los residuos aromáticos y polares que para los no polares.

En términos generales, la probabilidad predicha para los residuos que forman parte de una hélice tiende a ser menor que la de los residuos que tienen otra estructura secundaria. En el modelo SRC la estructura secundaria tiene un papel menor que en el modelo completo. Las diferencias de la probabilidad predicha entre distintas estructuras secundarias puede ser muy grande, especialmente para residuos

polares (pero sin carga) con valores altos de centralidades (por ejemplo N y Q).

La diferencia en la naturaleza de las predicciones del modelo completo y del modelo SRC se confirma en la visualización de una red de contactos (Figura 5.12 y 5.10, respectivamente). Mientras que el modelo SRC predice como críticos a los residuos poco expuestos al solvente sin importar la región de la proteína en que se encuentren, el modelo completo predice preferentemente los residuos cercanos al centro geométrico de la misma, y los que están dentro de regiones densamente conectadas o conectándolas entre sí. Esto se observa claramente en la Figura 5.13, en dónde visualizamos la diferencia entre las predicciones entre estos modelos.

Sin embargo, la cercanía, que claramente depende de la distancia al centro geométrico de la proteína, tiene un coeficiente de cero por sí sola en el modelo completo (Tabla 5.5). Es importante sólo en combinación con la intermediación de comunicabilidad *Cmb*, que tiene un coeficiente elevado en el modelo tanto por sí misma como en interacción con *Cls*. Esto parece indicar que la cercanía en sí no es la propiedad más importante. Los residuos más sensibles a mutaciones parecen ser los que tienen una *Cmb* alta (que, siguiendo la definición de *Cmb* serían los que participan en la transmisión de información entre distintas partes de una proteína). La probabilidad de que este subconjunto de residuos sea crítico aumenta conforme más cerca estén del centro de la proteína.

6.1.3 *Limitaciones de los modelos*

Todas las proteínas que usamos en este trabajo son globulares y relativamente pequeñas, por lo que sería díficil

pensar que el modelo puede generalizarse a otro tipo de proteínas, como podrían ser proteínas membranales o cuya actividad depende de complejos. En el conjunto de datos sólo se incluyen proteínas virales y bacteriales, por lo que también se desconoce si el modelo generaliza o no a proteínas eucariontes (que suelen tener modificaciones postraduccionales).

Los datos de mutagénesis que usamos no son homogéneos: provienen de grupos que usaron estrategias experimentales diferentes. Además, para algunos residuos se cuenta con más datos de mutagénesis que para otros. Esto hizo imposible tener una definición de residuo crítico única para todos los casos, y se tuvo que recurrir a definiciones *ad hoc* para cada proteína.

Aunque en todos los casos la definición se escogió para que los residuos menos tolerantes a mutaciones fueran críticos, esto vuelve nuestra lista de residuos críticos subjetiva hasta cierto punto, y la clasificación de los residuos puede no ser completamente consistente entre las distintas proteínas debido a ello. También es probable que existan algunos errores experimentales en los datos de mutagénesis. Un residuo que actualmente es clasificado como crítico podría ser reclasificado dependiendo de los errores experimentales o datos faltantes del fenotipo de las mutantes en esa posición. Estas inconsistencias dan lugar a ruido en la variable de respuesta, que es imposible de predecir en el modelo y probablemente reduzca el rendimiento del mismo.

Además de lo anterior, en el caso de la proteína del gen V estamos analizando una proteína que funciona como dímero como si fuera un monómero, pues la estructura del complejo no ha sido resuelta. La asimetría de las cadenas de la retrotranscriptasa puede significar que el rol de un residuo es distinto en ambas cadenas. Con los datos de mu-

tagénesis no se puede saber si se afectó la función de ambas cadenas o de sólo una. Para propósitos de ajustar y evaluar los modelos, estamos suponiendo que se afectaron ambas en todos los casos. Estas consideraciones pudieran explicar el menor rendimiento del modelo en estas dos proteínas en comparación a las demás (ver la Subsección A.2.1 en el Apéndice).

A pesar de estas consideraciones, se fue capaz de clasificar correctamente un 77.9 % de los residuos en los datos de la validación cruzada externa (para el modelo completo). Esta precisión se obtiene al predecir los residuos críticos de cada una de las proteínas con un modelo ajustado *sin los datos de esa proteína*. El hecho de que esto suceda muestra que el modelo se generaliza a datos distintos de los que se usaron para ajustarlo. Sin embargo, estrictamente sólo sabemos que lo hace para proteínas similares a las del conjunto de datos usado.

6.2 CONCLUSIONES

- Algunas medidas de centralidad de la red de contactos de una estructura proteíca, como las centralidades de intermediación de comunicabilidad y de cercanía, son atributos estructurales útiles para la predicción de residuos críticos.
- El área accesible al solvente está altamente correlacionada con las medidas de centralidad; sin embargo no son equivalentes y proveen información distinta útil para la predicción de residuos críticos.
- Un modelo que incorpora varias medidas de centralidad (CEN) fue capaz de mejorar la capacidad predictiva de un modelo basado en una sóla (`JAMMING`).

- Si además de usar medidas de centralidad se incorporan otras variables dependientes de la estructura primaria, secundaria, terciaria y cuaternaria de la proteína, la predicción del modelo mejora (modelo completo).
- Los residuos cargados en regiones poco accesibles al solvente frecuentemente son críticos.
- Lo mismo sucede para residuos que se encuentran en áreas densamente conectadas de una proteína (o conectándolas entre sí), especialmente si están cerca del centro geométrico de la proteína.

6.3 PERSPECTIVAS

6.3.1 *Medidas de centralidad como atributos estucturales*

Las centralidades de intermediación fueron desarrolladas para encontrar los nodos que influyen más en la *transmisión de información* en una red. En particular, la intermediación de comunicabilidad mide el número de veces que un nodo aparece en todos los caminos posibles entre todos los pares de nodos en la red, dando un mayor peso a los caminos cortos.

En una red de contactos estos residuos pudieran estar involucrados en la transmisión de información de una parte a la proteína a otra (por ejemplo, al unirse un ligando).[2] Por ello sería interesante evaluar si los residuos con una intermediación de comunicabilidad alta tienden a participar en procesos alostéricos.

2 Aunque esto sólo es cierto en la medida en que la red de contactos aproxime las rutas verdaderas por las que pasa la información en una proteína.

Los residuos que tienen una *cercanía* alta tienden a estar cerca del centro geométrico de la proteína, en dónde frecuentemente hay sitios de unión. También sería interesante evaluar si la cercanía, en combinación con otros atributos estructurales es capaz de predecir estos sitios.

Algunas medidas de centralidad fueron relativamente importantes en el modelo completo, aunque no encontramos una interpretación clara de su significado (aunque se observa cierto patrón geométrico en las visualizaciones). Este es el caso de la comunicabilidad Cmn y la centralidad de flujo de corriente Cfc (ver Subsección A.1.2). Estas centralidades no fueron importantes en el modelo CEN, que incluye sólo medidas de centralidad, pero sí lo fueron al tomar en cuenta el resto de los atributos estructurales, lo que sugiere que proveen información importante para la predicción sólo en combinación con algunos atributos tradicionales. Su rol puede evaluarse en trabajos subsecuentes.

6.3.2 *Modelos de este trabajo*

La mayor limitante de los modelos es que están basados en un conjunto de sólo seis proteínas, todas globulares y relativamente pequeñas, lo que limita su aplicación generalizada. En el futuro se pueden ajustar modelos similares a un mayor número de proteínas y/o a distintos conjuntos de proteínas que no están representadas en estos datos (como proteínas membranales).

Aún está pendiente poner los modelos desarrollados en este trabajo a disposición de la comunidad científica. Para hacerlo, se elaborará un programa para usar los modelos para predecir residuos críticos en cualquier estructura tridimensional, y se pondrá en un servidor de libre acceso.

Las medidas de centralidad y otros atributos usados en este trabajo pudieran también usarse para predecir otros conjuntos de residuos, como los catalíticos o los que están evolutivamente conservados. Otra aplicación interesante sería intentar usarlos para la predicción del fenotipo de mutantes individuales, en vez de la criticalidad de la posición.

Parte IV

APÉNDICE

A

APÉNDICE

En este capítulo incluyo las figuras que se omitieron del texto principal. La razón para hacerlo es que no son esenciales para la comprensión del trabajo y que su presencia dentro del texto volvería la presentación de las ideas principales más difusa. Sin embargo, se incluyen aquí para el lector interesado.

A.1 MEDIDAS DE CENTRALIDAD

A.1.1 *Dependencia con el tamaño de la red*

Algunas medidas de centralidad dependen del tamaño de la red, por lo que es difícil comparar los valores de centralidad entre distintas proteínas. Como ejemplo, graficamos la gráfica de caja de las distribuciones de la centralidad de cercanía y la de intermediación de comunicabilidad contra el número de residuos (Figura A.1 y A.2). Por ello fue necesario normalizar las centralidades para hacerlas comparables entre proteínas. Lo hicimos mediante una normalización por rangos, como se describe en la Subsección 3.3.3.

En las figuras se muestra la proteína correspondiente en la parte superior de la gráfica de caja correspondiente. Las abreviaturas usadas fueron GVP (proteína del gen V), T4L (lisozima T4), HIVP (proteasa del VIH), BL (beta lactamasa), LR (represor lac) y HIVR (retrotransciptasa del VIH).

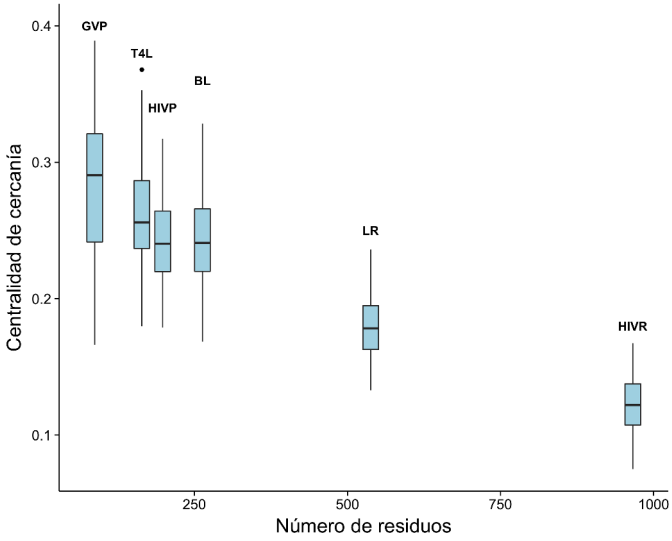


Figura A.1: Centralidad de cercanía.

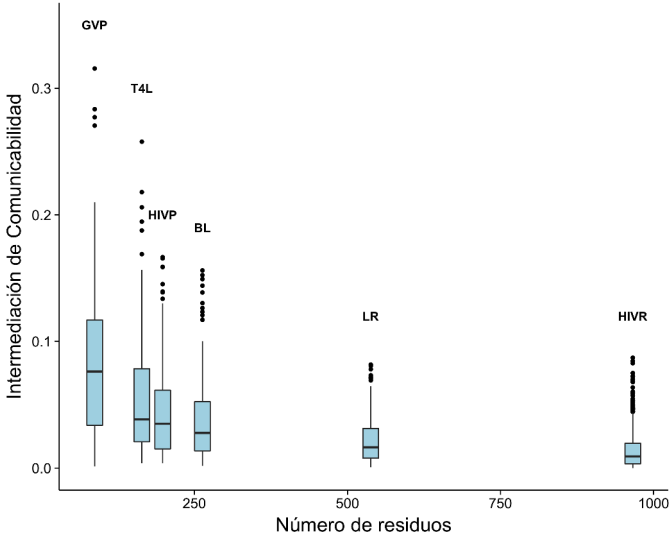


Figura A.2: Centralidad de intermediación de comunicabilidad.

A.1.2 *Visualización de medidas de centralidad en redes de contactos*

Para explorar si las medidas de centralidad tienen un significado en la estructura de la proteína, se elaboraron visualizaciones tridimensionales de la red de contactos del represor Lac, en la que los nodos se colorearon por los valores de centralidad normalizados.

En el texto principal se incluyen las visualizaciones para la cercanía (*Cls*), la intermediación de comunicabilidad (*Cmb*) y el grado normalizado (*Deg/SA*). Aquí se incluyen las visualizaciones para el resto de las centralidades (Figuras A.3-A.9).

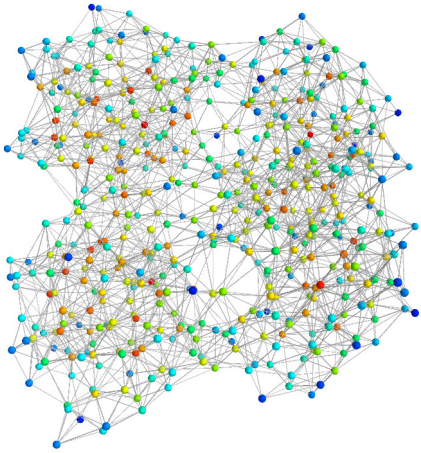


Figura A.3: GRADO El valor de *Deg* se representa con un gradiente de color que va desde azul (valores bajos) a rojo (valores altos).

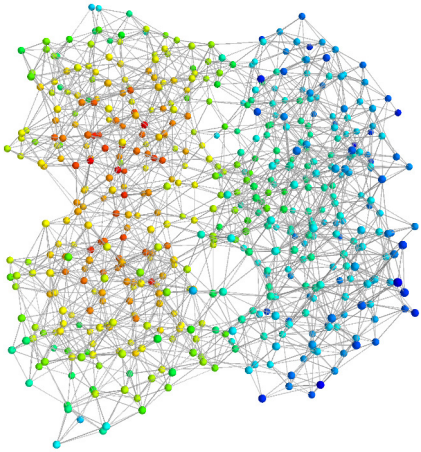


Figura A.4: CENTRALIDAD DE EIGENVECTOR El valor de *Eig* se representa con un gradiente de color que va desde azul (valores bajos) a rojo (valores altos).

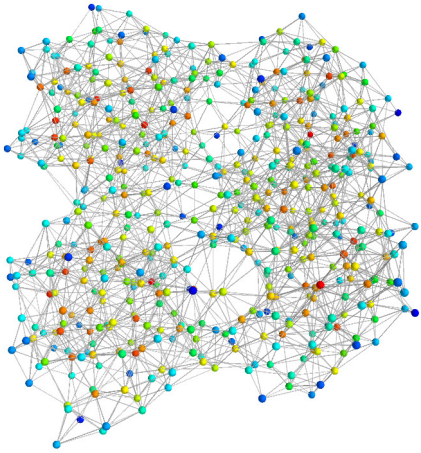


Figura A.5: PAGERANK El valor de *Pag* se representa con un gradiente de color que va desde azul (valores bajos) a rojo (valores altos).

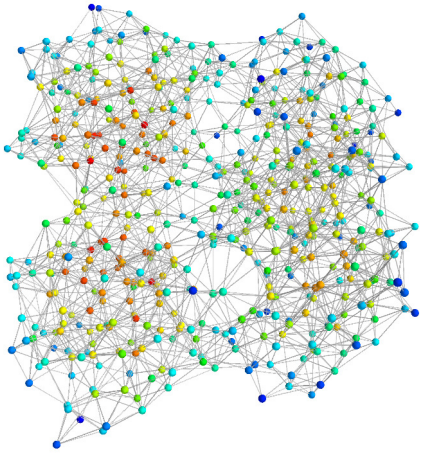


Figura A.6: COMUNICABILIDAD El valor de Cmn se representa con un gradiente de color que va desde azul (valores bajos) a rojo (valores altos).

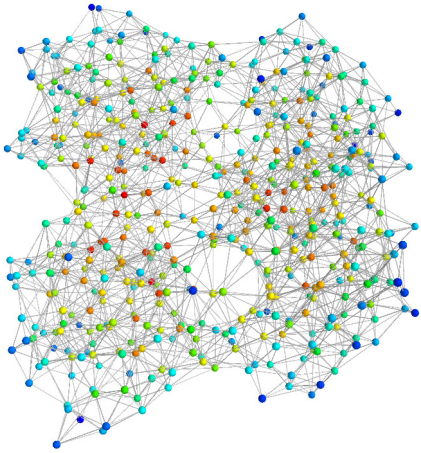


Figura A.7: CERCANÍA DE FLUJO DE CORRIENTE El valor de Cfc se representa con un gradiente de color que va desde azul (valores bajos) a rojo (valores altos).

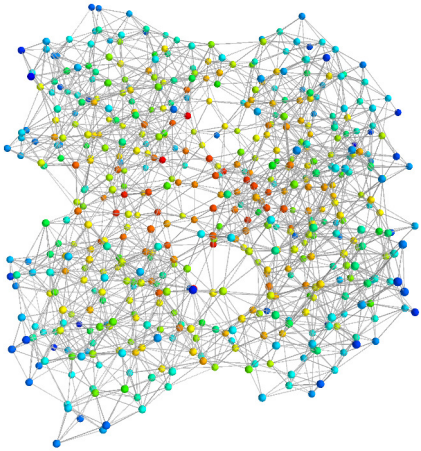


Figura A.8: INTERMEDIACIÓN El valor de *Bet* se representa con un gradiente de color que va desde azul (valores bajos) a rojo (valores altos).

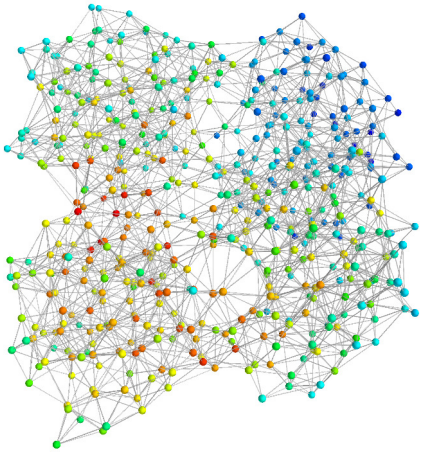


Figura A.9: INTERMEDIACIÓN DE FLUJO DE CORRIENTE El valor de Cfb se representa con un gradiente de color que va desde azul (valores bajos) a rojo (valores altos).

A.2 MODELOS

A.2.1 *Comparación entre* JAMMING *y los modelos desarrollados en este trabajo*

En esta sección se grafican los resultados individuales de precisión, *MCC* y *log loss* para cada proteína, calculadas en el paso de validación cruzada externa.

En el caso de JAMMING simplemente se grafican los resultados de predecir los residuos de las proteínas de nuestro conjunto de datos. No se calculó el *log loss* para JAMMING, pues no hace predicciones probabilísticas.

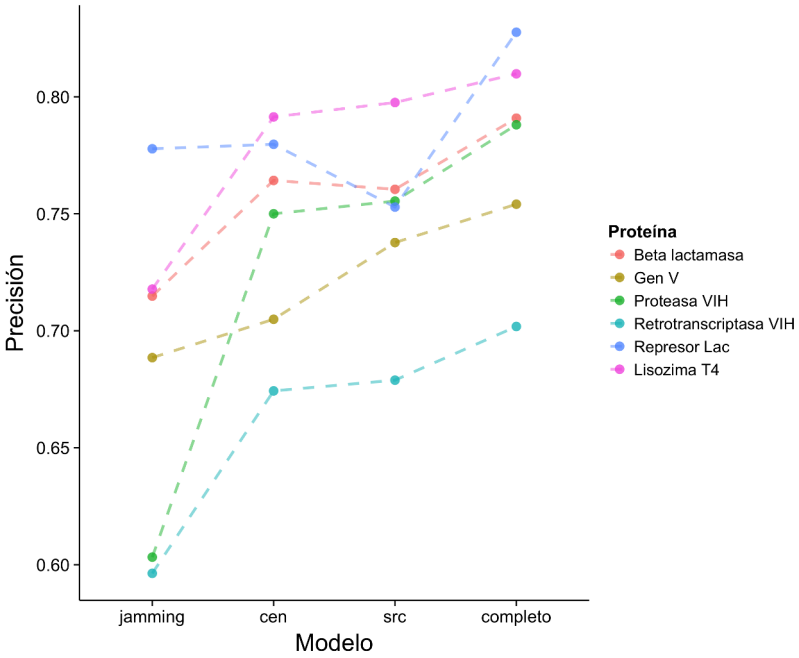


Figura A.10: PRECISIÓN Se muestra la precisión (*Acc*) de la predicción de cada modelo (eje x) para cada proteína (color). En el caso de JAMMING simplemente se muestra el resultado de predecir el conjunto de datos, para el resto de los modelos se muestran la precisión de la validación cruzada externa.

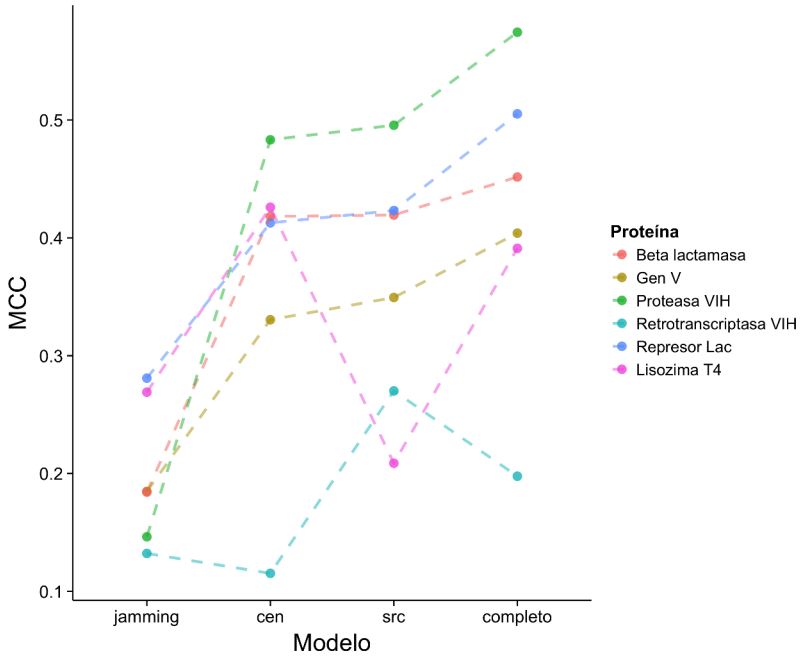


Figura A.11: COEFICIENTE DE CORRELACIÓN DE MATTHEWS Se muestra el coeficiente de correlación de Matthews (*MCC*) de la predicción de cada modelo (eje x) para cada proteína (color). En el caso de JAMMING simplemente se muestra el resultado de predecir el conjunto de datos, para el resto de los modelos se muestran la precisión de la validación cruzada externa.

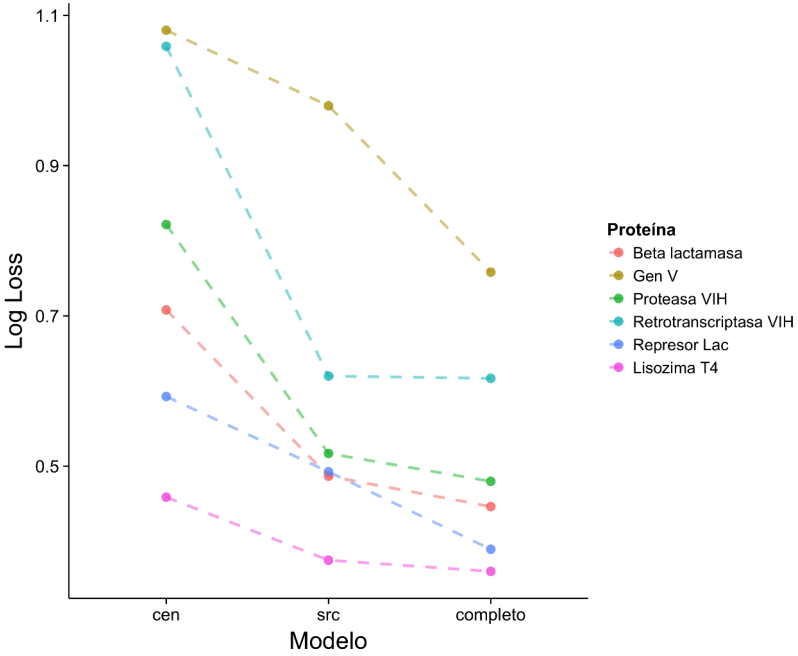


Figura A.12: LOG LOSS Se muestra el error *log loss* de la predicción de cada modelo (eje x) para cada proteína del conjunto de datos (color). Se muestran la precisión de la validación cruzada externa. JAMMING no hace predicciones probabilísticas, así que no se incluye en la comparativa.

A.2.2 *Sensibilidad y especificidad*

Una ventaja de los modelos de regresión logística con respecto a otros algoritmos de clasificación es que predicen la *probabilidad* de que datos nuevos pertenezcan a una clase, en vez de sólo predecir la clase más probable. Esto permite modificar el valor de corte de la probabilidad que se usa para predecir, dependiendo de la importancia relativa de los falsos positivos y falsos negativos.

En esta sección se incluyen tablas en las que se muestra la sensibilidad y especificidad de los modelos al variar la probabilidad de corte. Los residuos con una probabilidad mayor a ella son predichos como críticos y los que tengan una probabilidad menor o igual como no críticos.

Tabla A.1: Sensibilidad y especificidad del modelo CEN al variar la probabilidad de corte para predecir un residuo como crítico.

PROBABILIDAD	SENSIBILIDAD	ESPECIFICIDAD
0.05	0.782	0.556
0.1	0.723	0.643
0.15	0.691	0.693
0.2	0.666	0.731
0.25	0.645	0.766
0.3	0.626	0.785
0.35	0.586	0.809
0.4	0.552	0.824
0.45	0.531	0.839
0.5	0.514	0.847
0.55	0.484	0.856
0.6	0.445	0.862
0.65	0.398	0.877
0.7	0.386	0.898
0.75	0.349	0.918
0.8	0.297	0.927
0.85	0.25	0.951
0.9	0.206	0.97
0.95	0.122	0.986

Tabla A.2: Sensibilidad y especificidad del modelo SRC al variar la probabilidad de corte para predecir un residuo como crítico.

PROBABILIDAD	SENSIBILIDAD	ESPECIFICIDAD
0.05	0.906	0.374
0.1	0.88	0.494
0.15	0.86	0.569
0.2	0.833	0.614
0.25	0.791	0.659
0.3	0.718	0.693
0.35	0.654	0.724
0.4	0.597	0.77
0.45	0.55	0.815
0.5	0.478	0.857
0.55	0.351	0.902
0.6	0.253	0.932
0.65	0.182	0.95
0.7	0.121	0.973
0.75	0.065	0.984
0.8	0.034	0.995
0.85	0.017	0.999
0.9	0.003	1
0.95	0	1

Tabla A.3: Sensibilidad y especificidad del modelo completo al variar la probabilidad de corte para predecir un residuo como crítico.

PROBABILIDAD	SENSIBILIDAD	ESPECIFICIDAD
0.05	0.942	0.246
0.1	0.905	0.43
0.15	0.851	0.556
0.2	0.811	0.643
0.25	0.763	0.711
0.3	0.725	0.758
0.35	0.658	0.807
0.4	0.624	0.855
0.45	0.553	0.892
0.5	0.453	0.919
0.55	0.387	0.946
0.6	0.333	0.966
0.65	0.251	0.981
0.7	0.176	0.988
0.75	0.095	0.99
0.8	0.069	0.993
0.85	0.043	0.995
0.9	0.011	0.996
0.95	0	1

A.3 RESIDUOS CRÍTICOS

En esta sección se encuentra la lista completa de residuos críticos para cada proteína del conjunto de datos usados, según los criterios mencionados en Materiales y Métodos.

Tabla A.4: RESIDUOS CRÍTICOS DE LA PROTEASA DEL VIH Se muestra la clasificación que se hizo de los residuos de la proteasa del VIH en críticos y no críticos. Se muestra entre paréntesis el número de residuos de cada tipo.

GRUPO	RESIDUOS
Críticos (39)	LEU5, PRO9, ILE15, ALA22, LEU23, LEU24, ASP25, THR26, GLY27, ALA28, ASP29, THR31, VAL32, LEU33, MET36, GLY40, ILE47, GLY49, ILE50, GLY51, GLY52, VAL56, ARG57, TYR59, GLU65, GLY68, THR74, VAL75, VAL77, GLY78, PRO79, PRO81, ASN83, ILE84, ILE85, GLY86, ARG87, LEU90, LEU97
No críticos (53)	PRO1, VAL3, THR4, TRP6, GLN7, ARG8, LEU10, VAL11, THR12, LYS14, GLY16, GLY17, GLN18, LEU19, LYS20, GLU21, ASP30, GLU34, GLU35, SER37, LEU38, PRO39, ARG41, TRP42, LYS43, PRO44, LYS45, GLY48, PHE53, ILE54, LYS55, GLN58, ASP60, GLN61, LEU63, ILE64, ILE66, CYS67, HIS69, LYS70, ALA71, ILE72, GLY73, VAL82, ASN88, THR91, GLN92, ILE93, GLY94, CYS95, THR96, ASN98, PHE99
Desconocidos (7)	GLN2, ILE13, MET46, ILE62, LEU76, THR80, LEU89

Tabla A.5: RESIDUOS CRÍTICOS DE LA RETROTRANSCRIPTASA DEL VIH Se muestra la clasificación que se hizo de los residuos de la proteasa del VIH en críticos y no críticos. Se muestra entre paréntesis el número de residuos de cada tipo.

GRUPO	RESIDUOS
Críticos (34)	PRO95, PRO97, GLY99, LEU109, ASP110, VAL111, ALA114, SER117, LEU120, ARG125, ALA129, PHE130, ILE132, ASN136, PRO140, ARG143, TYR144, TYR146, LEU149, PRO150, GLY152, SER156, PRO157, ALA158, LEU168, GLN182, TYR183, MET184, ASP185, ASP186, VAL189, SER191, HIS198, ILE202
No críticos (75)	HIS96, ALA98, LEU100, LYS101, LYS102, LYS103, LYS104, SER105, VAL106, THR107, VAL108, GLY112, ASP113, TYR115, PHE116, VAL118, PRO119, ASP121, GLU122, ASP123, PHE124, LYS126, TYR127, THR128, THR131, PRO133, SER134, ILE135, ASN137, GLU138, THR139, GLY141, ILE142, GLN145, ASN147, VAL148, GLN151, TRP153, LYS154, GLY155, ILE159, PHE160, GLN161, SER162, SER163, MET164, THR165, LYS166, ILE167, GLU169, PRO170, PHE171, ALA172, ALA173, GLN174, ASN175, PRO176, ASP177, ILE178, VAL179, ILE180, TYR181, LEU187, TYR188, GLY190, ASP192, LEU193, GLU194, ILE195, GLY196, GLN197, ARG199, THR200, LYS201, GLU203

Desconocido (445)	Por su longitud, se omite esta lista de residuos. Consiste de todos los residuos de la proteína no presentes en las listas anteriores.

Tabla A.6: RESIDUOS CRÍTICOS DE LA PROTEÍNA DEL GEN V
Se muestra la clasificación que se hizo de los residuos de la proteína del gen V en críticos y no críticos. Se muestra entre paréntesis el número de residuos de cada tipo.

GRUPO	RESIDUOS
Críticos (20)	ILE6, ARG16, VAL19, TYR26, ASN29, GLU30, CYS33, LEU37, LEU49, ALA55, GLY59, TYR61, SER67, PHE68, LEU76, ARG80, LEU81, ARG82, LEU83, PRO85
No críticos (41)	GLU5, PRO8, SER9, GLN12, PHE13, THR15, GLY18, LYS24, LEU28, LEU32, TYR34, VAL35, ASP36, GLY38, ASN39, GLU40, TYR41, VAL43, LEU44, VAL45, ILE47, THR48, ASP50, GLY52, GLN53, PRO54, LEU60, THR62, VAL63, HIS64, LEU65, LYS69, VAL70, GLY71, GLY74, MET77, ILE78, ASP79, VAL84, ALA86, LYS87
Desconocidos (26)	MET1, ILE2, LYS3, VAL4, LYS7, GLN10, ALA11, THR14, SER17, SER20, ARG21, GLN22, GLY23, PRO25, SER27, GLN31, PRO42, LYS46, GLU51, TYR56, ALA57, PRO58, SER66, GLN72, PHE73, SER75

Tabla A.7: RESIDUOS CRÍTICOS DE LA β-LACTAMASA Se muestra la clasificación que se hizo de los residuos de la β-lactamasa de tipo TEM en críticos y no críticos. Se muestra entre paréntesis el número de residuos de cada tipo.

GRUPO	RESIDUOS
Críticos (77)	LEU30, ALA42, VAL44, GLY45, TYR46, GLY54, GLU64, PHE66, PRO67, SER70, THR71, LYS73, VAL74, LEU75, LEU76, MET81, ILE95, LEU102, PRO107, LEU122, ALA125, ALA126, SER130, ASP131, ASN132, THR133, ALA134, ALA135, LEU148, LEU152, ASP157, THR160, ARG164, GLU166, ASP176, ASP179, THR180, THR181, PRO183, MET186, LEU207, TRP210, ASP214, VAL216, ALA217, LEU220, ARG222, LEU225, PRO226, TRP229, ALA232, ASP233, LYS234, SER235, GLY236, ALA237, GLY238, GLY242, ARG244, GLY245, ILE246, ILE247, LEU250, GLY251, ARG259, ILE260, VAL261, ILE263, TYR264, THR271, MET272, ASP273, ASN276, ILE279, ILE282, GLY283, TRP290

No críticos (190)

PRO22, SER23, LEU24, ALA25, HIS26, GLY27, GLU28, THR29, VAL31, LYS32, VAL33, SER34, ASP35, ALA36, GLU37, LEU38, LEU39, LEU40, GLY41, ARG43, LEU47, THR48, LEU49, GLU50, LEU51, ASN52, SER53, LYS55, ILE56, LEU57, ASP58, VAL59, HIS60, ARG61, GLU62, TYR63, ARG65, LEU68, MET69, PHE72, CYS77, GLY78, CYS79, VAL80, SER82, ARG83, VAL84, ASP85, ALA86, GLY87, HIS88, GLU89, GLN90, LEU91, GLY92, ARG93, LEU94, HIS96, TYR97, SER98, ARG99, PRO100, ASP101, VAL103, LYS104, TYR105, SER106, SER108, VAL109, LYS110, LYS111, HIS112, ALA113, THR114, GLU115, GLY116, MET117, THR118, VAL119, LYS120, GLU121, CYS123, THR124, ILE127, THR128, MET129, ASN136, LEU137, LEU138, LEU139, SER140, THR141, ILE142, GLY143, GLY144, PRO145, TYR146, GLU147, THR149, ALA150, PHE151, HIS153, SER154, MET155, GLY156, HIS158, ALA159, ARG161, LEU162, ASP163, TRP165, THR167, ASP168, LEU169, ASN170, GLU171, ALA172, ILE173, PRO174, ASN175, GLU177, ARG178, GLY182, ALA184, ALA185, ALA187, THR188, THR189, LEU190, ARG191, LYS192, LEU193, LEU194, THR195, GLY196, GLU197, LEU198, LEU199, THR200, ILE201, ALA202, SER203, ARG204, GLN205, GLU206, ILE208, ASP209, MET211, GLU212, ALA213, MET215, GLY218, PRO219, LEU221, SER223, ALA224, HIS227, GLY228, TYR230, ILE231, GLU240, ARG241, SER243, ALA248, ALA249, PRO252, ASP254, GLY255, LYS256, PRO257, SER258, VAL262, THR265, THR266, GLY267, SER268, GLN269, ALA270, ALA274, ARG275, ARG277, GLN278, ALA280, MET281, ALA284, SER285, LEU286, VAL287, ALA288, HIS289

Desconocidos (19)	MET3, SER4, ILE5, GLN6, HIS7, PHE8, ARG9, VAL10, ALA11, LEU12, ILE13, PRO14, PHE15, PHE16, ALA17, ALA18, PHE19, CYS20, LEU21

Tabla A.8: RESIDUOS CRÍTICOS DEL REPRESOR LAC Se muestra la clasificación que se hizo de los residuos del represor Lac en críticos y no críticos. Se muestra entre paréntesis el número de residuos de cada tipo.

GRUPO	RESIDUOS
Críticos (100)	THR5, LEU6, TYR7, ASP8, VAL9, ALA10, ALA13, VAL15, SER16, TYR17, GLN18, THR19, VAL20, SER21, ARG22, VAL23, ASN25, HIS29, VAL30, LYS33, THR34, VAL38, ALA41, MET42, LEU45, TYR47, ILE48, PRO49, ASN50, ARG51, ALA53, GLN54, LEU56, ALA57, GLY58, LYS59, GLY65, VAL66, LEU71, LEU73, ALA75, PRO76, SER77, ILE83, LYS84, ALA87, ASP88, VAL95, VAL96, MET98, ASN125, PRO127, LEU128, ASP149, VAL150, ILE159, PHE161, GLY166, HIS179, LEU184, LEU185, GLY187, SER191, VAL192, SER193, ALA194, ARG195, ARG197, TRP220, GLY225, ALA241, LEU243, ALA245, ASN246, ASP247, GLN248, MET249, ALA250, LEU251, GLY252, ALA253, ILE268, VAL271, GLY272, TYR273, ASP274, ASP275, THR276, SER279, TYR282, PRO284, LEU286, THR287, ILE289, PHE293, GLY297, SER300, LEU319, ARG326, THR328

No críticos (220)

LYS2, PRO3, GLU11, TYR12, GLY14, VAL24, GLU26, ALA27, SER28, SER31, ALA32, ARG35, GLU36, LYS37, GLU39, ALA40, ALA43, GLU44, ASN46, VAL52, GLN55, GLU60, SER61, LEU62, LEU63, ILE64, ALA67, SER69, SER70, ALA72, HIS74, ALA82, SER85, ARG86, GLN89, LEU90, GLY91, SER93, VAL94, SER97, VAL99, GLU100, ARG101, SER102, GLY103, VAL104, GLU105, ALA106, CYS107, LYS108, ALA109, ALA110, VAL111, HIS112, ASN113, LEU114, LEU115, ALA116, GLN117, ARG118, SER120, GLY121, LEU122, ILE123, ILE124, TYR126, ASP129, ASP130, GLN131, ASP132, ALA133, ILE134, ALA135, VAL136, GLU137, ALA138, ALA139, CYS140, THR141, ASN142, VAL143, PRO144, ALA145, LEU146, PHE147, LEU148, SER151, ASP152, GLN153, THR154, PRO155, ILE156, ASN157, SER158, ILE160, SER162, HIS163, GLU164, ASP165, THR167, ARG168, LEU169, GLY170, VAL171, GLU172, HIS173, LEU174, VAL175, ALA176, LEU177, GLY178, GLN180, GLN181, ILE182, ALA183, ALA186, PRO188, LEU189, SER190, LEU196, LEU198, ALA199, GLY200, TRP201, HIS202, LYS203, TYR204, LEU205, THR206, ARG207, ASN208, GLN209, ILE210, GLN211, PRO212, ILE213, ALA214, GLU215, ARG216, GLU217, GLY218, ASP219, SER221, ALA222, MET223, SER224, PHE226, GLN227, GLN228, THR229, MET230, GLN231, MET232, LEU233, ASN234, GLU235, GLY236, ILE237, VAL238, PRO239, THR240, MET242, VAL244, MET254, ARG255, ALA256, ILE257, THR258, GLU259, SER260, GLY261, LEU262, ARG263, VAL264, GLY265, ALA266, ASP267, SER269, VAL270, GLU277, ASP278, SER280, CYS281, ILE283, PRO285, THR288, LYS290, GLN291, ASP292, ARG294, LEU295, LEU296, GLN298, THR299, VAL301, ASP302, ARG303, LEU304, LEU305, GLN306, LEU307, SER308, GLN309, GLY310, GLN311, ALA312, VAL313, LYS314, GLY315, ASN316, GLN317, LEU318, PRO320, VAL321, SER322, LEU323, VAL324, LYS325, LYS327, THR329

Desconocidos (40)	MET1, VAL4, THR68, GLN78, ILE79, VAL80, ALA81, ALA92, VAL119, LEU330, ALA331, PRO332, ASN333, THR334, GLN335, THR336, ALA337, SER338, PRO339, ARG340, ALA341, LEU342, ALA343, ASP344, SER345, LEU346, MET347, GLN348, LEU349, ALA350, ARG351, GLN352, VAL353, SER354, ARG355, LEU356, GLU357, SER358, GLY359, GLN360

Tabla A.9: RESIDUOS CRÍTICOS DE LA LISOZIMA T4 Se muestra la clasificación que se hizo de los residuos de la lisozima T4 en críticos y no críticos. Se muestra entre paréntesis el número de residuos de cada tipo.

GRUPO	RESIDUOS
Críticos (33)	ASP10, GLU11, TYR18, ASP20, THR26, ILE27, GLY28, GLY30, LEU46, GLY51, ILE58, LEU66, ILE78, LEU84, ARG95, ALA98, ASN101, PHE104, GLN105, GLY107, PHE114, SER117, LEU121, ALA129, LEU133, SER136, TRP138, THR142, ARG145, ALA146, VAL149, GLY156, TYR161

No críticos (130)	ASN2, ILE3, PHE4, GLU5, MET6, LEU7, ARG8, ILE9, GLY12, LEU13, ARG14, LEU15, LYS16, ILE17, LYS19, THR21, GLU22, GLY23, TYR24, TYR25, ILE29, HIS31, LEU32, LEU33, THR34, LYS35, SER36, PRO37, SER38, LEU39, ASN40, ALA41, ALA42, LYS43, SER44, GLU45, ASP47, LYS48, ALA49, ILE50, ARG52, ASN53, CYS54, ASN55, GLY56, VAL57, THR59, LYS60, ASP61, GLU62, ALA63, GLU64, LYS65, PHE67, ASN68, GLN69, ASP70, VAL71, ASP72, ALA73, ALA74, VAL75, ARG76, GLY77, LEU79, ARG80, ASN81, ALA82, LYS83, LYS85, PRO86, VAL87, TYR88, ASP89, SER90, LEU91, ASP92, ALA93, VAL94, ARG96, CYS97, LEU99, ILE100, MET102, VAL103, MET106, GLU108, THR109, GLY110, VAL111, ALA112, GLY113, THR115, ASN116, LEU118, ARG119, MET120, GLN122, GLN123, LYS124, ARG125, TRP126, ASP127, GLU128, ALA130, VAL131, ASN132, ALA134, LYS135, ARG137, TYR139, ASN140, GLN141, PRO143, ASN144, LYS147, ARG148, ILE150, THR151, THR152, PHE153, ARG154, THR155, THR157, TRP158, ASP159, ALA160, LYS162, ASN163, LEU164
Desconocidos (1)	MET1

BIBLIOGRAFÍA

[1] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, y S. Pietrokovski. Network analysis of protein structures identifies functional residues. *J. Mol. Biol.*, 344(4):1135–1146, Dec 2004.

[2] J. Bergstra, D. Yamins, y D. D. Cox. Making a science of model search : Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th Innternational Conference on Machine Learning*, 2013.

[3] C. Chothia. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.*, 105(1):1–12, Jul 1976.

[4] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, y M. J. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, Jun 2009.

[5] M. P. Cusack, B. Thibert, D. E. Bredesen, y G. Del Rio. Efficient identification of critical residues based only on protein structure by network analysis. *PLoS ONE*, 2(5):e421, 2007.

[6] R. Daber, S. Stayrook, A. Rosenberg, y M. Lewis. Structural analysis of lac repressor bound to allosteric effectors. *J. Mol. Biol.*, 370(4):609–619, Jul 2007.

[7] A. Das, V. Prashar, S. Mahale, L. Serre, J. L. Ferrer, y M. V. Hosur. Crystal structure of HIV-1 protease in situ product complex and observation of a low-barrier hydrogen bond between catalytic aspartates. *Proc. Natl. Acad. Sci. U.S.A.*, 103(49):18464–18469, Dec 2006.

[8] K. Das, J. D. Bauman, A. D. Clark, Y. V. Frenkel, P. J. Lewi, A. J. Shatkin, S. H. Hughes, y E. Arnold. High-resolution structures of HIV-1 reverse transcriptase/TMC278 complexes: strategic flexibility explains potency against resistance mutations. *Proc. Natl. Acad. Sci. U.S.A.*, 105(5):1466–1471, Feb 2008.

[9] M. O. Dayhoff, R. M. Schwartz, y B. C. Orcutt. Chapter 22: A model of evolutionary change in proteins. In *in Atlas of Protein Sequence and Structure*, 1978.

[10] A. del Sol, H. Fujihashi, D. Amoros, y R. Nussinov. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol. Syst. Biol.*, 2:2006.0019, 2006.

[11] E. Estrada, D. J. Higham, y N. Hatano. Communicability betweenness in complex networks. *Physica A*, 388: 764–774, 2009.

[12] J. E. Fajardo y A. Fiser. Protein structure based prediction of catalytic residues. *BMC Bioinformatics*, 14:63, 2013.

[13] E. A. Franzosa y Y. Xia. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.*, 26(10):2387–2395, Oct 2009.

[14] K. Fujiwara, H. Toda, y M. Ikeguchi. Dependence of alpha-helical and beta-sheet amino acid propensities

on the overall protein fold type. *BMC Struct. Biol.*, 12: 18, 2012.

[15] Aric A. Hagberg, Daniel A. Schult, y Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, páginas 11–15, Pasadena, CA USA, Agosto 2008.

[16] Aric A. Hagberg, Daniel A. Schult, Pieter J. Swart, et al. NetworkX reference guide, 2015. URL http://networkx.github.io/documentation/latest/reference/index.html.

[17] T. Hamelryck y B. Manderick. PDB file parser and structure class implemented in Python. *Bioinformatics*, 19(17):2308–2310, Nov 2003.

[18] W. Huang, J. Petrosino, M. Hirsch, P. S. Shenkin, y T. Palzkill. Amino acid sequence determinants of beta-lactamase structure and activity. *J. Mol. Biol.*, 258(4): 688–703, May 1996.

[19] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL http://www.scipy.org/.

[20] W. Kabsch y C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12): 2577–2637, Dec 1983.

[21] L. G. Kleina y J. H. Miller. Genetic studies of the lac repressor. XIII. Extensive amino acid replacements generated by the use of natural and synthetic nonsense suppressors. *J. Mol. Biol.*, 212(2):295–318, Mar 1990.

[22] J. Kyte y R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157(1):105–132, May 1982.

[23] S. Q. Le y O. Gascuel. An improved general amino acid replacement matrix. *Mol. Biol. Evol.*, 25(7):1307–1320, Jul 2008.

[24] D. D. Loeb, R. Swanstrom, L. Everitt, M. Manchester, S. E. Stamper, y C. A. Hutchison. Complete mutagenesis of the HIV-1 protease. *Nature*, 340(6232):397–400, Aug 1989.

[25] P. Markiewicz, L. G. Kleina, C. Cruz, S. Ehret, y J. H. Miller. Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers"which do not require a specific sequence. *J. Mol. Biol.*, 240(5):421–433, Jul 1994.

[26] M. Matsumura, J. A. Wozniak, D. P. Sun, y B. W. Matthews. Structural studies of mutants of T4 lysozyme that alter hydrophobic stabilization. *J. Biol. Chem.*, 264 (27):16059–16066, Sep 1989.

[27] Yaser Abu Mostafa. *Learning from data : a short course*. AMLBook.com, United States, 2012. ISBN 1600490069.

[28] M. E. J. Newman. *Networks : an introduction*. Oxford University Press, Oxford New York, 2010. ISBN 0199206651.

[29] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

[30] P. Ramachandran y G. Varoquaux. Mayavi: 3D Visualization of Scientific Data. *Computing in Science & Engineering*, 13(2):40–51, 2011. ISSN 1521-9615.

[31] D. Rennell, S. E. Bouvier, L. W. Hardy, y A. R. Poteete. Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.*, 222(1):67–88, Nov 1991.

[32] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, y M. H. Zehfus. Hydrophobicity of amino acid residues in globular proteins. *Science*, 229(4716):834–838, Aug 1985.

[33] S. G. Sarafianos, B. Marchand, K. Das, D. M. Himmel, M. A. Parniak, S. H. Hughes, y E. Arnold. Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition. *J. Mol. Biol.*, 385(3):693–713, Jan 2009.

[34] B. Stec, K. M. Holtz, C. L. Wojciechowski, y E. R. Kantrowitz. Structure of the wild-type TEM-1 beta-lactamase at 1.55 A and the mutant enzyme Ser70Ala at 2.1 A suggest the mode of noncovalent catalysis for the mutant enzyme. *Acta Crystallogr. D Biol. Crystallogr.*, 61(Pt 8):1072–1079, Aug 2005.

[35] S. Su, Y. G. Gao, H. Zhang, T. C. Terwilliger, y A. H. Wang. Analyses of the stability and function of three surface mutants (R82C, K69H, and L32R) of the gene V protein from Ff phage by X-ray crystallography. *Protein Sci.*, 6(4):771–780, Apr 1997.

[36] T. C. Terwilliger, H. B. Zabin, M. P. Horvath, W. S. Sandberg, y P. M. Schlunk. In vivo characterization

of mutants of the bacteriophage f1 gene V protein isolated by saturation mutagenesis. *J. Mol. Biol.*, 236(2): 556–571, Feb 1994.

[37] B. Thibert, D. E. Bredesen, y G. del Rio. Improved prediction of critical residues for protein function based on network and phylogenetic analyses. *BMC Bioinformatics*, 6:213, 2005.

[38] M. Vendruscolo, N. V. Dokholyan, E. Paci, y M. Karplus. Small-world view of the amino acids that play a key role in protein folding. *Phys Rev E Stat Nonlin Soft Matter Phys*, 65(6 Pt 1):061910, Jun 2002.

[39] J. A. Wrobel, S. F. Chao, M. J. Conrad, J. D. Merker, R. Swanstrom, G. J. Pielak, y C. A. Hutchison. A genetic approach for identifying critical residues in the fingers and palm subdomains of HIV-1 reverse transcriptase. *Proc. Natl. Acad. Sci. U.S.A.*, 95(2):638–645, Jan 1998.

[40] H. Zou y T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67:301–320, 2005.

COLOFÓN

Este documento fue elaborado en el programa LyX usando el estilo tipográfico `classicthesis` desarrollado por André Miede. Las figuras se elaboraron con una combinación de scripts de Python, librerías gráficas del lenguaje de programación R y LibreOffice Draw.