UACM

Universidad Autónoma de la Ciudad de México

Nada humano me es ajeno

COLEGIO DE CIENCIAS Y HUMANIDADES

MAESTRÍA EN CIENCIAS DE LA COMPLEJIDAD

Sitios activos en macromoléculas biológicas mediante Cadenas de Markov y Teoría de Redes Complejas

TRABAJO RECEPCIONAL
PARA OBTENER EL TÍTULO DE
MAESTRO EN CIENCIAS DE LA COMPLEJIDAD

PRESENTA

GABRIEL ELOY AGUILAR PINEDA

DIRECTOR

Dr. Luis Olivares Quiroz

Ciudad de México, enero 2019

UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

CA Coordinación Académica

SISTEMA BIBLIOTECARIO DE INFORMACIÓN Y DOCUMENTACIÓN

UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO
COORDINACIÓN ACADÉMICA

RESTRICCIONES DE USO PARA LAS TESIS DIGITALES

DERECHOS RESERVADOS©

Agradecimientos

Este trabajo no habría sido posible sin el apoyo y el estímulo de mi asesor y amigo, Dr. Luis Olivares Quiroz, bajo cuya supervisión elegimos el tema y desarrollamos este trabajo de tesis.

Agradezco a la Universidad Autónoma de la Ciudad de México y sus autoridades, por el apoyo económico y los principios que rigen a la institución, bajo los cuales he podido ejercer la apasionada profesión de docente y al mismo tiempo dedicarme al proyecto de investigación que ha dado como producto esta tesis.

A mis compañeros de trabajo, docentes, administrativos y personal de apoyo, por hacer de las horas de trabajo, las mejores horas del día.

No puedo terminar sin agradecer a mi familia, en cuyo estímulo constante y amor he confiado todos estos años.

A mi hijo, cuya compañía es siempre mi mayor alegría.

Resumen

Presentamos un estudio sobre cómo ubicar sitios activos y de unión en la estructura de proteínas, utilizando Teoría de Redes Complejas y Cadenas de Markov. En el caso de Redes Complejas, cada aminoácido de la cadena proteíca representa un vértice en la red, y si los átomos de dos aminoácidos diferentes están a una distancia menor a un radio de corte que nosotros establecimos, se coloca un enlace entre ellos, lo que permitiría la comunicación o el transporte entre dichos vértices. En el caso de Cadenas de Markov, los aminoácidos son los estados del sistema, y se puede pasar de un estado a otro en un solo paso, con una probabilidad que depende de la cantidad de átomos de cada aminoácido y de cuántas parejas de átomos pueden establecerse entre dos de ellos, separados por una distancia menor al radio de corte. De esta manera pudimos calcular la probabilidad de ir de un residuo a otro, en un determinado número de pasos.

El Sitio Activo de una proteína está formado por los aminoácidos que realizan los procesos catalíticos. Otros residuos funcionales participan como sitios de unión para sustratos, cofactores o para iones. En este contexto, propusimos que los sitios activos y los de unión se relacionan con algunas medidas de centralidad de la teoría de redes, particularmente analizamos la centralidad de cercanía y la de intermediación.

En cuanto a Cadenas de Markov, buscamos una relación entre los sitios activos y de unión, y los estados a los que en promedio es posible llegar por primera vez en un proceso estocástico, desde cualquier otro, en la menor cantidad de pasos (Tiempo de Primeras Visitas).

Modelando así la estructura de las proteínas, ubicamos los residuos que cumplen con las características descritas, y comparamos nuestros resultados con otros reportados en la literatura.

Abstract

In this study we use Complex Network Theory and Markov Chain Processes, to locate and predict active sites in protein structures. In Network theory, each amino acid or residue is represented as a node or vertex in the network, and if the atoms of a couple of residues are closer than a cutoff distance, we assume there is a link between them, which allows comunication or transport whithin the protein network. As for the Markov Chain process, amino acids are states of the system and the communications between such states is possible with a given probability that depends on the atoms number of every residue and the total number of atom-atom contacts between amino acids based in the same cutoff distance.

Active Site residues in a protein correspond to the amino acids that carry out the catalytic activity, but there are also residues with other functional activity like binding of sustrates, cofactors and metals. In this context, we propose that the active sites and functional residues are related with some network centrality measures. Particularly, we analized clossness and betweness centrality. For the Markov Chain case, we propose there is a relationship between active sites and the states which the information takes minimal average number of steps to be transmitted for the first time.

This way of modeling the protein structure, allowed us to ubicate residues that satisfied certain characteristics that we could related with active sites reported in literature.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

En este capítulo, haremos una descripción básica de la estructura de las proteínas y los aminoácidos, que son los constituyentes fundamentales de las primeras. Discutiremos cómo las proteínas son construidas en el interior de las células, los diferentes niveles en los que suele describirse su estructura, y el llamado *problema del plegamiento de las proteínas* (*protein folding*), el cual se refiere a la manera como las proteínas después de ser construidas, logran adoptar rápidamente la configuración espacial correcta, es decir, aquella en la que son funcionales. Así mismo, discutiremos la relevancia del plegamiento correcto de las proteínas, y la existencia y descripción de los sitios activos, que son los lugares específicos en la estructura de algunas proteínas donde se llevan a cabo los procesos bioquímicos específicos que las hacen tan relevantes.

El propósito de esta introducción es colocar las bases biofísicas y dar un panorama fundamental sobre proteínas. Posteriormente en los siguientes capítulos, veremos cómo representar estas macromoléculas utilizando Teoría de Redes Complejas y cómo establecer un proceso estocástico en este sistema en términos de Cadenas de Markov.

1.1. Proteínas y Aminoácidos

Las proteínas se encuentran entre las moléculas biológicas más importantes en la Naturaleza debido a la gran variedad de funciones que realizan. Participan como catalizadores en prácticamente todas las reacciones bioquímicas en los organismos;

tienen una gran capacidad de reconocimiento de otras moléculas, lo que sirve para mandar "señales"de forma muy precisa, o facilitar el paso de compuestos a través de la membrana celular. Pueden incluso ser ellas mismas el conducto o el medio de transporte; contribuyen en la formación y degradación de otras sustancias, y además de todo, son parte estructural de muchos sistemas, desde sistemas macroscópicos como la queratina o el colágeno que forma parte del cabello y las uñas, o las histonas, proteínas que dan soporte y colaboran en el enrollamiento y compactación del ADN dentro de las células [1, 2]. Se puede decir de forma general, que todas las funciones que puede tener una proteína entran en alguna de las siguientes categorías: (a) *Regulación y señalización*, (b) *Catálisis*, (c) *Movimiento y transporte*, y (d) *Estructura*.

Los elementos atómicos básicos que constituyen a las proteínas son carbono, nitrógeno, oxígeno e hidrógeno, aunque también a veces contienen otros, como azufre, fósforo, magnesio o hierro. Las proteínas están formadas por cadenas de moléculas más pequeñas, cuya unidad o eslabón recibe el nombre de *aminoácido* o *residuo*, unidos entre sí mediante *enlaces covalentes*. Los átomos que forman este tipo de enlace comparten sus electrones exteriores, uno de ellos recibe uno o más electrones, y el otro los cede. De esta manera consiguen completar su configuración electrónica, lo que hace un enlace muy estable. Entre aminoácidos y aminoácidos con agua, existen además otras interacciones electrostáticas que contribuyen a mantener la estructura tridimensional de la proteína, y que se suelen clasificar en función de la naturaleza de la carga (iones, moléculas con polarización permanente, con polarización inducida, o interacciones producidas por efectos cuánticos [3]). La clasificación más común considera como las más importantes a las *interacciones de Van der Waals* y los *puentes de hidrógeno*.

Las interacciones de Van der Waals incluyen interacciones entre moléculas dipolares y con dipolos inducidos, además de interacciones cuánticas conocidas como fuerzas de dispersión. La interacción por puente de hidrógeno es un tipo especial de interacción dipolo-dipolo, que se forman cuando se produce una región de carga positiva en un átomo de H debido a que su electrón participa en un enlace covalente, por lo que este átomo sentirá atracción por elementos con carga negativa. Ambas interacciones son relativamente débiles, típicamente de 4 a 20 kJ/mol para las in-

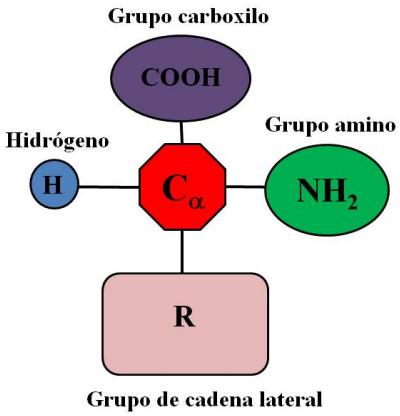


Figura 1.1: Estructura atómica de un aminoácido. Todos los aminoácidos tienen unidos al carbono central (C_α) un grupo amino (NH_2), un grupo carboxilo ($COOH$) y un átomo de hidrógeno (H). El grupo R llamado cadena lateral o residuo, es el que hace diferencia entre un aminoácido y otro.

teracciones de Van der Waals y de 5 a 30 kJ/mol los puentes de hidrógeno, por lo que son significativamente menores a la de los enlaces covalentes, que está entre 150 y 870 kJ/mol [3], sin embargo aparecen cientos de éstas entre los átomos de las proteínas y los átomos del agua. En conjunto, todas estas interacciones permiten a los átomos mantener diferentes posiciones relativas entre sí, y a las proteínas una estructura tridimensional muy variada, lo cual es de suma importancia, pues muchas propiedades y la funcionalidad de una proteína, dependen justamente de su estructura. Además ocurren también *enlaces iónicos*, en los cuales a diferencia de los enlaces covalentes, los electrones no se comparten sino se transfieren entre un átomo y otro, produciéndose así una atracción electrostática entre el par de átomos ionizados.

Los aminoácidos están formados químicamente por un grupo *amino* ($-NH_2$), un grupo *carboxilo* o *ácido* ($-COOH$); unidos ambos a un carbono "principal", cuyas otras dos valencias quedan cubiertas con un átomo de hidrógeno y con un grupo químico variable, llamado *residuo* ($-R$) o *cadena lateral* (figura (1.1)). Al carbono principal se le llama *carbono alfa*, C_α. Tridimensionalmente los aminoácidos son como tetraedros, con el C_α en el centro y los cuatro elementos que se unen a él, en los vértices.

Las proteínas están formadas por una secuencia de aminoácidos, esta secuencia es diferente para cada tipo de proteína y está indicada en el código genético del ADN. En el interior de las células existen unas macromoléculas llamadas ribosomas, que se encargan de traducir este código y unir los aminoácidos para sintetizar las proteínas.

De entre todos los aminoácidos que existen en la naturaleza, sólo 20 distintos se utilizan para la conformación de proteínas en los seres vivos, y cada uno de estos aminoácidos está caracterizado por su cadena lateral o grupo residual R.

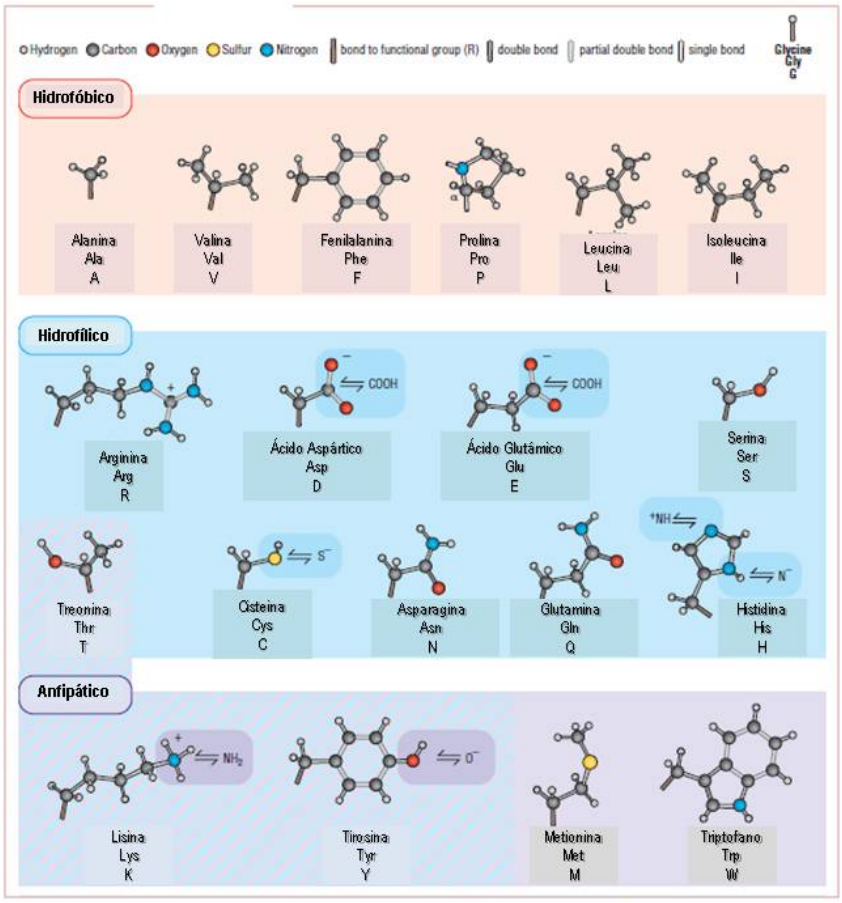


Figura 1.2: Existen 20 diferentes grupos residuales, mostrados en la figura (Imagen tomada de referencia [1]).

En la figura (1.2) se muestran el nombre, las abreviaturas clásicas de 3 letras y de una, y la estructura atómica de estos 20 residuos, clasificados de acuerdo a su

comportamiento en solución (hidrofóbico, hidrofílico o anfipático).

- **Aminoácidos hidrofóbicos**. En estos casos, el residuo de las proteínas evita entrar en contacto con el agua, por lo que tienden a ubicarse en el interior de la estructura de las proteínas. En esta categoría se incluye la alanina (Ala, A), valina (Val, V), leucina (Leu, L), isoleucina (Ile, I), fenilalanina (Phe, F) y prolina (Pro, P).

- **Aminoácidos hidrofílicos**. Las cadenas en esta categoría pueden tener carga eléctrica a ciertas condiciones de pH, lo que facilita su interacción y la formación de enlaces de H entre ellas o con otras moléculas, incluyendo el agua. El ácido glutámico (Glu, E) y el ácido aspártico (Asp, D), tienen carga negativa en condiciones de pH fisiológico, mientras la arginina (Arg R) y la lisina (Lys, K) se muestran positivas. La histidina (His, H) tiene el comportamiento más versátil, pues tiene dos zonas en las que puede formar enlaces de H. Además puede presentarse con carga positiva y aunque en muy pocos casos, también con carga negativa. Esta versatilidad puede ser la razón por la que sea el residuo que más participa en la actividad catalítica de las proteínas. La serina (Ser, S), treonina (Thr, T), glutamina (Gln, Q), y asparagina (Asn, N) pueden ser donadores o aceptores de puentes de H para formar enlaces.

- **Aminoácidos Anfipáticos**. Los aminoácidos en esta categoría tienen regiones con ambas características, polares y no polares. Esto hace que formen interfaces. La lisina (Lys, K), aunque tiene una región cargada, tiene también una extensa zona hidrofóbica. Tirosina (Tyr, Y) y triptofano (Trp, W) suelen no ionizar en condiciones de pH fisiológico, pero puede haber interacciones polares débiles con los anillos aromáticos de sus residuos. La metionina (met, M) es el menos polar de todos, aunque el azufre que lo compone puede participar en muchas interacciones.

El residuo glicina (Gly, G), está formada solamente por un átomo de H, así que es el más simple de los aminoácidos. Algunos autores lo consideran como una cuarta categoría, y otros lo incluyen en los hidrofóbicos [4].

Otra clasificación de los aminoácidos en el campo biológico es en función de si el organismo del que forman parte puede sintetizarlos por sí mismo o no. Cuando no es

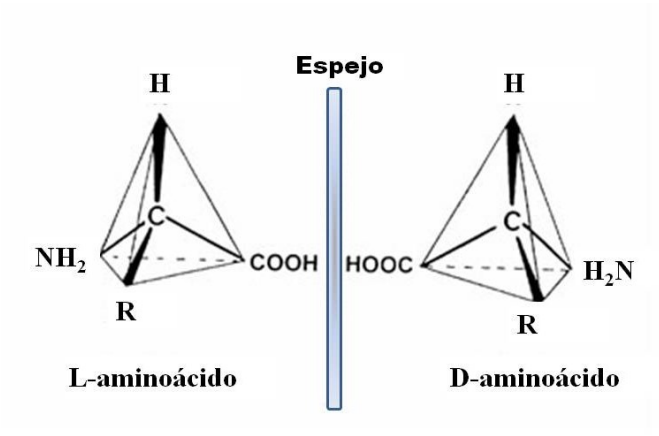


Figura 1.3: Las dos configuraciones isoméricas de los aminoácidos se llaman estereoisómeros, ya que cada uno es la imagen especular del otro. El isómero L es el único que se utiliza en la construcción de proteínas.

posible, es necesario introducir ciertas proteínas al organismo, las cuales son descompuestas para obtener estos aminoácidos y poder construir con ellos nuevas proteínas. Se les llama en este caso aminoácidos **esenciales**. En cambio, son aminoácidos **no esenciales** los que el propio organismo puede sintetizar. En el caso de los seres humanos se tiene

- **Aminoácidos Esenciales**: valina (Val), leucina (Leu), isoleucina (Ile), fenilalanina (Phe), triptófano (Trp), treonina (Thr), metionina (Met), histidina (His), lisina (Lys) y arginina (Arg).
- **Aminoácidos no Esenciales**: alanina (Ala), glicina (Gly), tirosina (Tir), serina (Ser), cisteína (Cys), glutamina (Glu), prolina (Pro), ácido aspártico (Asp), ácido glutámico (Glu) y asparagina (Asn).

Todos los aminoácidos, excepto la glicina, cuyo residuo es sólo un átomo de H, son *moléculas quirales*, es decir, pueden existir en dos diferentes formas o isómeros. Esto se debe a que el C_α está unido a 4 compuestos diferentes y estos no necesariamente están en el mismo orden. Para aminoácidos se acostumbran llamar forma **L** o forma **D**. Se reconoce una de otra con la *Regla del Maíz*, que va de la siguiente manera: Colocamos el aminoácido de forma que el H que va unido directamente al C_α quede

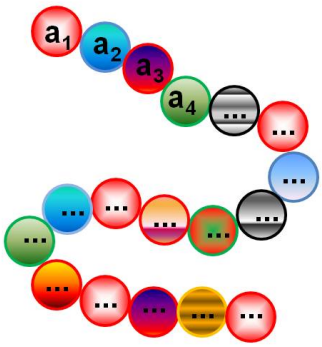


Figura 1.4: La estructura primaria de una proteína es la secuencia de aminoácidos que la forman.

en la parte superior, y los otros tres enlaces formando la base triangular del tetraedro. Le asignamos al grupo carboxilo las letras **CO**, al residuo la **R** y al grupo amino la **N**. Si leemos el orden de los grupos enlazados al C_α, comenzando por el carboxilo y siguiendo la dirección de las manecillas del reloj, se formará la palabra **CORN** (maíz en inglés) si el aminoácido es L y **CONR** si el aminoácido es D (fig. (1.3)). Los sistemas biológicos utilizan sólo la forma L, así que es la única que aparece en los aminoácidos que constituyen las proteínas.

1.2. Síntesis y Estructura de las Proteínas

Las proteínas están formadas por una cadena lineal de aminoácidos, de manera que la diferencia entre una proteína y otra es la cantidad y el orden de la secuencia en que aparecen estos aminoácidos. A esta secuencia se le conoce como **estructura primaria**, y es fundamental en la determinación de las propiedades de la proteína (figura (1.4)).

La secuencia de aminoácidos de cada proteína, está codificada en la secuencia de desoxinucleótidos que conforman el ADN (ácido desoxirribonucleico) del organismo que la produce. El ADN, junto con el ARN (ácido ribonucleico) son los *ácidos*

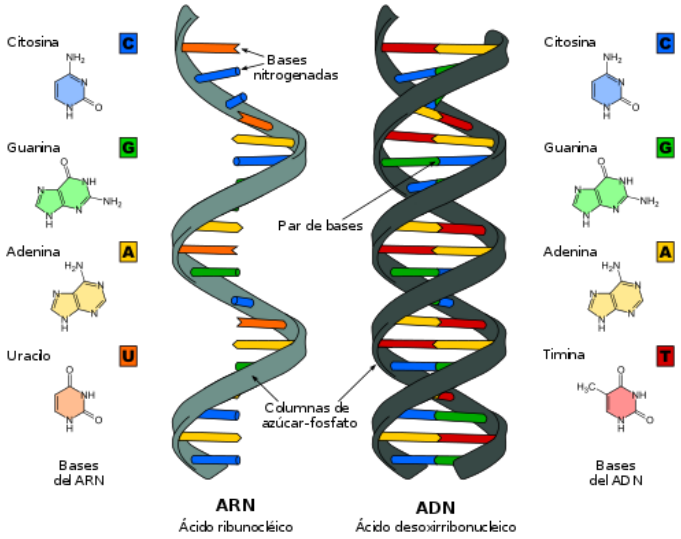


Figura 1.5: ADN y ARN. Se muestran los nucleótidos que las constituyen. El apareamiento en el ADN, sólo puede darse entre A y T, y entre C y G. Imagen tomada de https://www.lifeder.com/funciones-adn-arn.

nucleicos que forman parte de los organismos vivos. La molécula de ADN está formada por desoxinucleótidos, los cuales a su vez, están constituidos por un azúcar (2-desoxirribosa), una base nitrogenada que puede ser adenina (A), timina (T), citosina (C) o guanina (G) y un ácido fosfórico. Lo que distingue a un desoxinucleótido de otro, es la base nitrogenada, por ello la secuencia del ADN se especifica nombrando sólo la secuencia de sus bases. La disposición secuencial de estas cuatro bases a lo largo de la cadena codifica la información genética. En los organismos vivos el ADN se presenta como una doble cadena de desoxinucleótidos, en la que las dos ramas están unidas entre sí por enlaces de puente de hidrógeno (ver por ejemplo [5]).

A diferencia del ADN que contiene 2-desoxirribosa, el azúcar constitutivo del ARN es ribosa. Otra diferencia importante es que el ARN no contiene la base nucleotída timina, y en su lugar posee la base uracilo. Además el ARN generalmente es de una sola cadena (fig. (1.5)). Existen diferentes tipos de moléculas de ARN en la célula: el ARN ribosomal (ARN-r), el ARN de transferencia (ARN-t), el ARN mensajero (ARN-m) y otros más. Cada molécula de ARN-m contiene la información para la secuencia de aminoácidos de una proteína, mientras que las moléculas de ARN-r y de ARN-t forman parte de la maquinaria celular que traduce la información de los

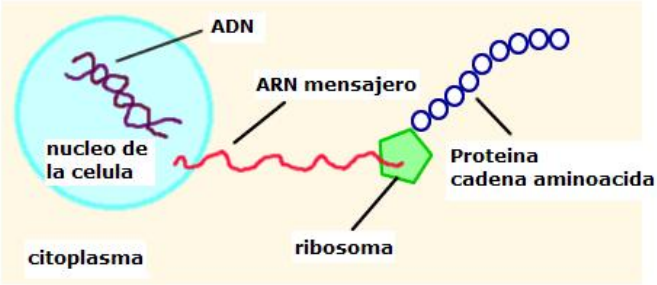


Figura 1.6: La transcripción del ADN a ARN ocurre en el núcleo de las células, mientras que la traducción a la cadena de aminoácidos que constituirá una proteína, se produce en los ribosomas.

ARN-m a proteínas.

El proceso para sintetizar proteínas involucra al ADN y al ARN, y se lleva a cabo a grandes rasgos, en dos pasos (ver fig. (1.6)).

Transcripción del ADN. A pesar de que el ADN contiene la información para configurar una proteína, no puede hacerlo por sí mismo, por lo que primero transcribe la información en el ARN. Durante la transcripción, dos enzimas llamadas *helicasa* y *girasa*, rompen los enlaces de puente de hidrógeno que unen las dos hélices de ADN, haciendo que éstas se separen. A continuación la enzima *ARN-polimerasa* sintetiza al ARN mensajero (ARN-m) a partir de una de las hebras de ADN, de manera que se mantenga en el ARN-m la información de la secuencia. Este proceso ocurre en el núcleo de las células y posteriormente el ARN-m sale, deja el núcleo y se mueve hacia los ribosomas.

Traducción. En esta parte del proceso, la cadena de ARN-m se une a una de las dos porciones que constituyen el *ribosoma*, un compuesto macromolecular formado por ARN-r y proteínas. La información para la secuencia de aminoácidos de una proteína está codificada en el ARN-m, en unidades independientes de tres bases llamadas *codones*.

En unidades de tres bases y dado que cada base tiene cuatro opciones, se pueden generar 64 codones. Los 64 codones y el significado de cada uno constituyen el código genético. Algunos aminoácidos están especificados por un único codón, otros

en cambio pueden obtenerse a través de hasta 6 diferentes, como se muestra en la tabla (1.1). Hay además otros codones para indicar en dónde comienza y dónde debe terminar la cadena de aminoácidos que se está construyendo.

AMINOÁCIDO	ABREVIATURAS		CODONES
Alanina	**Ala**	**A**	GCA, GCC, GCG, GCU
Valina	**Val**	**V**	GUA, GUC, GUG, GUU
Glicina	**Gli**	**G**	GGA, GGC, GGG, GGU
Leucina	**Leu**	**L**	UUA, UUG, CUA, CUC, CUG, CUU
Isoleucina	**Ile**	**I**	AUA, AUC, AUU
Metionina	**Met**	**M**	AUG
Prolina	**Pro**	**P**	CCA, CCC, CCG, CCU
Fenilalanina	**Phe**	**F**	UUC, UUU
Tirosina	**Tyr**	**Y**	UAC, UAU
Triptófano	**Trp**	**W**	UGG
Serina	**Ser**	**S**	AGC, AGU, UCA, UCC, UCG, UCU
Cisteína	**Cys**	**C**	UGC, UGU
Treonina	**Thr**	**T**	ACA, ACC, ACG, ACU
Asparagina	**Asg**	**N**	AAC, AAU
Glutamina	**Gln**	**Q**	CAA, CAG
Ácido Aspártico	**Asp**	**D**	GAC, GAU
Ácido Glutámico	**Glu**	**E**	GAA, GAG
Lisina	**Lys**	**K**	AAA, AAG
Arginina	**Arg**	**R**	AGA, AGG, CGA, CGC, CGG, CGU
Histidina	**His**	**H**	CAC, CAU

Cuadro 1.1: Cada aminoácido está representado en el ADN por un codón, es decir, una cadena de tres nucleótidos. Hay residuos que tiene más de una representación.

Finalmente, la secuencia transcrita en el ARN, es traducida a la secuencia de aminoácidos. Esta acción es llevada a cabo por una familia de ácidos ribonucleicos pequeños, los ARN-t, que de cierta manera actúan como intérpretes del código genético. Los ARN-t leen el mensaje expresado como codones en el ARN-m, y al mismo tiempo reconocen a los aminoácidos especificados por estos codones. Este paso fundamental de la traducción se lleva a cabo en sitios especiales de los ribosomas, que sostienen al ARN-m y lo mueven para que los codones se posicionen en dichos sitios, el ARN-t los lee y entrega el aminoácido que corresponde para formar la cadena. El ribosoma además juega el papel de catalizador en esta unión de aminoácidos.

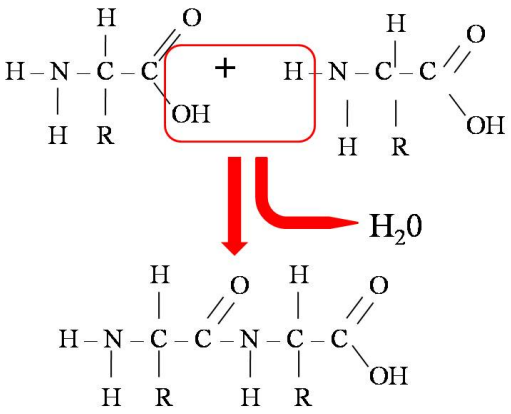


Figura 1.7: Los aminoácidos se unen mediante enlaces petídicos, que implican la deshidratación de los aminoácidos y la formación de un enlace covalente entre un grupo NH y el grupo carboxilo CO del siguiente residuo.

Los aminoácidos al formar esta cadena, se unen a través de *enlaces peptídicos*, que consisten en la reacción de un grupo carboxilo y el grupo amino del siguiente aminoácido, dándose un enlace covalente $CO - NH$ y la pérdida de una molécula de agua (figura (1.7)). El grupo carboxilo de este segundo aminoácido, reacciona con el grupo amino de un tercero, y así se forma la cadena o *esqueleto principal*, con el grupo amino NH, seguido del C_α, y el C del grupo carboxilo ($NH - C_\alpha - CO$), unido al grupo amino del siguiente aminoácido. Las cadenas laterales emergen de los C_α.

La **Estructura Secundaria** es la organización que la cadena lineal adopta, con la intención de conseguir una forma energéticamente más estable. Esta estructura se consigue gracias a los puentes de hidrógeno que se establecen entre los grupos COOH y NH_2 de residuos no necesariamente consecutivos. Las estructuras secundarias más comunes, son la *hélice* α, y las *hojas* β (figura (1.8)).

En la hélice α, el esqueleto de la cadena se enrolla de manera compacta alrededor del eje longitudinal de la molécula, cada aminoácido, se desplaza 0.15 nm a lo largo del eje con respecto al aminoácido anterior, y cada vuelta completa de la hélice implica una elevación de 0.54 nm. En este caso, los puentes de hidrógeno se forman casi siempre entre el grupo -C=O del n-ésimo aminoácido y el -NH del n+4-ésimo (es decir, cuatro aminoácidos más adelante en la cadena). Existen otros tipos de hélices,

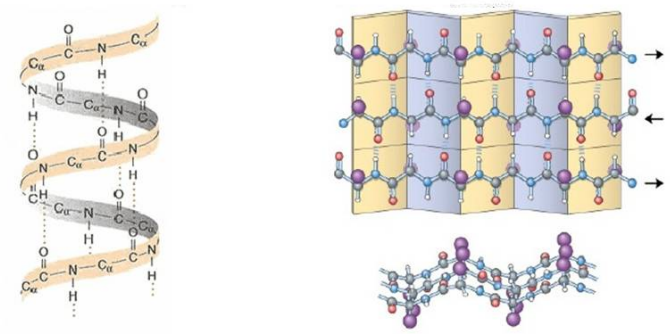


Figura 1.8: Estructura secundaria de una proteína. En general en (a), cada vuelta de la hélice está formado por 3.6 aminoácidos.

en las que los puentes de hidrógeno ocurren entre los aminoácidos n y $n+3$ o entre el n y el $n+5$, pero son mucho menos usuales. Así, cada vuelta sucesiva de la hélice se mantiene unida a las vueltas adyacentes mediante varios puentes de hidrógeno, que proporcionan a la estructura una considerable estabilidad.

En una hoja β, el esqueleto de la cadena polipeptídica se dispone en zig-zag con los grupos R de los distintos aminoácidos proyectándose alternativamente a uno y otro lado de dicho esqueleto. Los puentes de hidrógeno pueden ocurrir entre aminoácidos muy alejados entre sí en la secuencia primaria. Muchas de estas cadenas colocadas paralelamente unas a otras forman una estructura que recuerda a una hoja de papel plegada, en la que los grupos R de los aminoácidos se encuentran sobresaliendo por ambas caras de dicha hoja. La distancia entre dos aminoácidos adyacentes es en este caso de 0.35 nm. que es la que determina el enlace covalente.

Ciertos aminoácidos se encuentran con más frecuencia formando parte de hélices α, y otros prefieren ser constituyentes de hojas β. Esta propiedad se conoce como *Propensión Helicoidal* [6]. Por ejemplo, leucina, metionina, glutamina y el ácido glutámico, se encuentran con frecuencia en hélices, en parte porque los residuos de

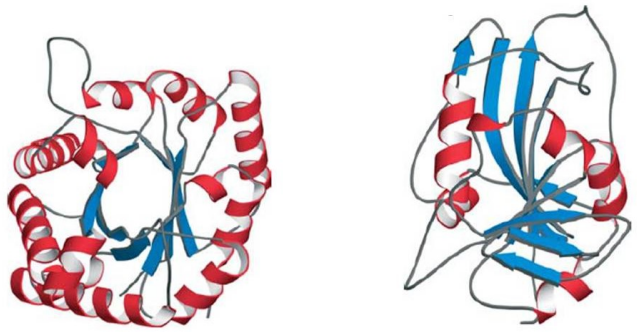


Figura 1.9: En la figura se muestra la estructura terciaria de dos proteínas. Aunque ambas están formadas por 8 hojas β, la ubicación de éstas es muy diferente, por lo que es de esperarse que tengan diferentes funciones (Imagen tomada del libro de Gregory Petsko, referencia [1]).

estos aminoácidos son cadenas largas y en esta configuración es posible extenderlas y mantenerlas lejos de la parte central, donde tendrían que interaccionar con el resto de las cadenas. La alanina, pese a no tener un residuo extenso, es el más fuerte "formador de hélices"[7]. La propensión helicoidal también depende de la temperatura, el pH de la solución e incluso de la ubicación de la hélice dentro de la estructura de la proteína [8]. En cambio, la valina, isoleucina, y fenilalanina, es más frecuente que formen parte de hojas β, pues son residuos ramificados por lo que les resulta más fácil extenderse en las caras opuestas de la hoja.

La **Estructura Terciaria** tiene que ver con la disposición espacial de todos los átomos que componen la proteína. Ésta puede ser de tipo fibroso, es decir, una de las dimensiones más grande que las otras, o bien de tipo globular, en donde no existe una dimensión predominante y la proteína asume forma redondeada.

Como resultado de la estructura terciaria, se genera una superficie compleja que le permita a la proteína interactuar específicamente ya sea con moléculas pequeñas o con otras macromoléculas. Un axioma fundamental de la biología es que la estructura tridimensional de las proteínas determina su función, pues esta estructura les permite interactuar de manera muy selectiva. La estructura tridimensional en la que una proteína es funcional, se conoce como *estructura nativa*.

En la figura (1.9) se muestran dos proteínas con estructuras secundarias parecidas, ya que en ambos casos hay 8 hojas β rodeadas por hélices α, sin embargo la estructura terciaria es diferente, por lo que no tienen la misma función. Entonces, si la estructura determina la función, la secuencia de aminoácidos, determina la estructura. Christian Anfinsen analizó en 1961 esta relación para la proteína Ribonucleasa A [9]. Anfinsen desplegaba (desnaturalizaba) esta proteína y luego permitía que volviera a plegarse. Observó que la proteína recuperaba al plegarse su funcionalidad, por lo que propuso que la estructura nativa tridimensional de una proteína en su medio fisiológico normal (condiciones de solvente, pH, fuerzas iónicas, presencia de otros componentes, temperatura, etc.), es aquella en la cual la energía libre del sistema es un mínimo, por lo tanto en la ecuación 1.1, ΔF debe ser negativo.

$$\Delta F = \Delta E - T\Delta S \tag{1.1}$$

El plegamiento implica una disminución en la entropía, lo que contribuye a un ΔS negativo, que no es favorable termodinámicamente [10, 11], sin embargo el reacomodo de las moléculas de agua que solvataban las regiones hidrofóbicas de los aminoácidos y que son expulsadas en el plegamiento, contribuyen positivamente a ΔS. En el proceso de plegamiento ocurre también un cambio de energía ΔE, por el reacomodo de los átomos de la proteína, variando de esta manera su interacción con el solvente, sin embargo este cambio es cercano a cero[12]. La contribución más significativa en la estabilización de la estructura terciaria proviene de la interacción hidrofóbica entre los aminoácidos no polares y las moléculas de agua [13]. La estructura final es una estructura estable y biológicamente activa. La propuesta de que el control termodinámico del plegamiento esté determinado por la secuencia de aminoácidos de la proteína misma, se conoce como el *Dogma de Anfinsen*. Este hecho lleva de inmediato a la búsqueda de mecanismos que nos sirvan para predecir la estructura terciaria de una proteína a partir de la secuencia de aminoácidos, problema aún abierto en la actualidad y que se relaciona con el proceso de plegamiento de las proteínas, que se discutirá más adelante.

La **Estructura Cuaternaria** no está presente en todas las proteínas, sólo en aquellas que constan de más de una cadena polipeptídica, pues se refiere a la forma como se acomodan las diferentes cadenas para formar un complejo proteico (figura

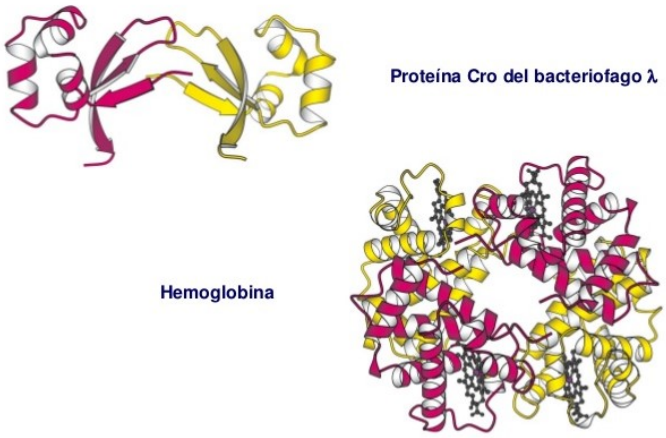


Figura 1.10: La estructura cuaternaria se refiere al ordenamiento espacial que adquieren las diferentes cadenas. Por ejemplo, la proteína Cro del bacteriófago λ, está formada por dos subunidades, mientras que la hemoglobina está constituida por cuatro cadenas.

1.10).

Esta unión permite la formación de importantes estructuras biológicas, como fibras, microtúbulos o complejos enzimáticos. La estructura cuaternaria le da funcionalidad a los complejos, aunque muchas veces sus componentes también tienen actividad de forma aislada.

1.3. El plegamiento de las proteínas

Como mencionamos previamente, la funcionalidad de una proteína depende fuertemente de su estructura tridimensional, por lo tanto, un problema fundamental aún en boga es el análisis del proceso por el cual una proteína alcanza de manera rápida y precisa, su estado nativo.

C. Levinthal planteó en 1969, que dado que una proteína se pliega en muy poco tiempo, incluso del orden de microsegundos, no es posible que el proceso ocurra de forma aleatoria, pues tendría que pasar por muchas posibles configuraciones, por lo que la localización de la estructura tridimensional correcta le tomaría una cantidad enorme de tiempo [14, 15]. Por ejemplo, si suponemos que cada enlace peptídico

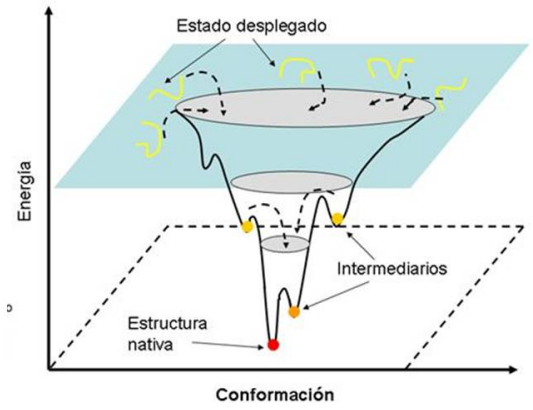


Figura 1.11: La energía libre del sistema disminuye en la escala vertical. El punto más bajo corresponde al estado nativo.

tiene sólo tres grados de libertad, entonces para una proteína de 100 aminoácidos existirán $3^{100} = 5 \times 10^{47}$ conformaciones, por lo que está fuera de toda posibilidad que las proteínas recorran todas ellas en el proceso de plegamiento [16]. La paradoja radica en que este plegamiento hacia la estructura de mínima energía, ocurre de forma espontánea y en un tiempo extremadamente corto. Por esta razón Levinthal concluyó que la búsqueda estocástica no es el mecanismo que este proceso sigue. Se puede considerar que este problema implica la conciliación de dos procesos: por un lado un proceso termodinámico por el hecho de considerar que el estado nativo del sistema es aquel que minimiza la energía libre, y por otro, el proceso cinético que ocurre al pasar por las diferentes configuraciones de la proteína hasta llegar a su configuración funcional [17, 18].

La propuesta de Bryngelson y Wolynes [19, 20] ante este problema, tiene que ver con lo que ha dado en llamarse *Teoría del Paisaje Energético* (Energy Landscape Theory), basada en la conformación de vidrios de spin. En este modelo, la superficie de energía libre del sistema, tiene forma de embudo, con el estado nativo en la parte inferior, es decir, en el lugar de menor energía, como se muestra en la fig. (1.11). Esto implica que no existe un camino preferente que deba seguir el sistema para, a

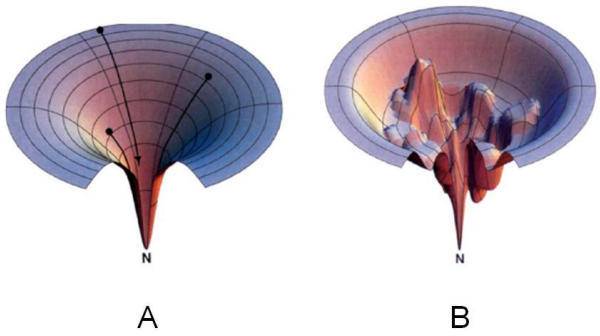


Figura 1.12: En (A) La superficie energética del embudo es suave, lo que facilita la evolución desde cualquier punto inicial a la parte inferior, que corresponde al estado nativo. En (B) en cambio, el sistema presenta muchos mínimos locales, lo que dificulta y frena el plegamiento del sistema. Imagen tomada de la referencia [23].

partir de un estado desplegado o desnaturalizado, llegar al plegamiento, más bien hay todo un conjunto de rutas que pueden llevar al sistema a la conformación de mínima energía. Conforme va disminuyendo la energía del sistema, hay menos configuraciones posibles, por lo que la entropía también disminuye, hasta llegar al punto más bajo del embudo, el punto de mínima energía, que es el estado nativo de la proteína [21, 22]. Así, cada proteína seguirá su propia vía. Este modelo también contempla la posibilidad de que el embudo no sea una superficie suave, sino que existan mínimos locales, en los cuales la proteína puede detenerse e incluso quedar atrapada [23], lo que impediría llegar al estado nativo, que es el único con función biológica. Este fenómeno es conocido como *frustración* y es un ingrediente fundamental de la teoría del paisaje energético (fig. 1.12).

La solución que se plantea ante este fenómeno de frustración, es apelar a la evolución. Es decir, las estructuras primarias que implicaban un proceso de plegamiento con mucha frustración o con mínimos locales muy abundantes o muy pronunciados, no fueron seleccionadas evolutivamente, de forma que las proteínas que conocemos son aquellas que consiguieron un paisaje energético considerablemente suave [23, 24]. La búsqueda de una respuesta definitiva al problema de plegamiento, se mantiene

como un problema activo de sumo interés.

1.4. Sitios Activos de Macromoléculas

Las *enzimas* son proteínas con funciones catalíticas. Esta función es realizada desde un lugar específico de su estructura llamado *sitio activo*, que suele estar formado por unos pocos aminoácidos cercanos geométricamente, aunque no necesariamente consecutivos en la estructura primaria. Algunos de estos residuos se unen, retienen e incluso orientan a la molécula con la que se llevará a cabo la reacción (sustrato) y a otras más que en ocasiones también actúan en el proceso químico (cofactores); mientras otros residuos participan propiamente en la función catalítica. Diremos que son parte del sitio activo todos los residuos que participan en alguna de estas funciones.

Estos sitios son consecuencia de la estructura tridimensional de la proteína [1]. En muchos casos, los residuos se ordenan de manera que forman una cavidad en la superficie, y debido a sus características, el medio ambiente alrededor del sitio y en la cavidad formada puede ser muy diferente al del resto del solvente, lo que favorecería la interacción de los residuos del sitio con algunas moléculas específicas del medio, como se ilustra en la fig. (1.13). Si por ejemplo, un sitio está formado por residuos cargados, puede éste ser atractor de iones metálicos. Si los residuos forman una cavidad cuyas paredes se constituyen de cadenas hidrofóbicas, se forma un sitio de unión de moléculas con características hidrofóbicas, como los lípidos.

Puesto que la unión de la proteína con alguna otra sustancia es un proceso selectivo, las propiedades de esta unión dependen también de las características del sustrato. Si el sustrato es un ion o una molécula pequeña, el sitio de unión suele tener forma de cavidad, pero cuando el sustrato es una macromolécula, el sitio de unión puede ser también una superficie plana o una concavidad en la superficie de la proteína. Incluso los sitios de unión pueden estar no sólo en la superficie, sino dentro de la proteína, lo que implicaría en estos casos, que debe haber al menos un camino para que el sustrato llegue al sitio activo [25].

Identificar los aminoácidos que forman el sitio activo es un problema de gran desarrollo teórico y experimental, y con importantes espectativas en cuanto a las aplicaciones médicas y tecnológicas implicadas en este reto [25–28]. Existen en la web

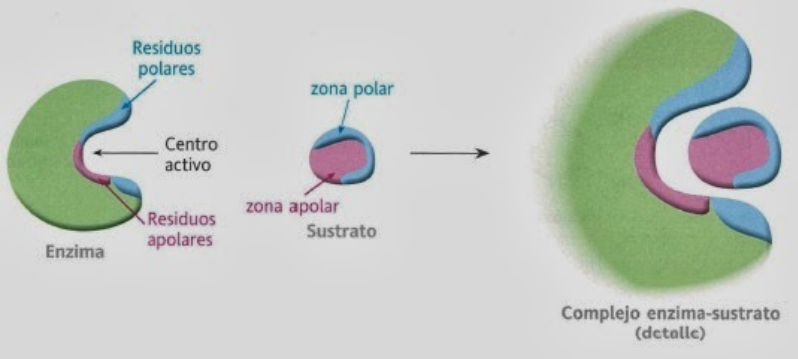


Figura 1.13: Esquema de la conformación de un sitio activo. Las propiedades de las paredes del sitios facilitan la interacción de éste con un sustrato específico. Imagen tomada del sitio http://apuntesbioquimicageneral.blogspot.mx.

muchas páginas que brindan información general sobre un conjunto bastante amplio de proteínas, y particularmente, otras que incluyen la ubicación en la cadena lineal de los residuos que forman los sitios activos de la proteína. Esta ubicación se hace en ocasiones de forma experimental, y otras veces comparando computacionalmente estructuras de diferentes proteínas.

Probablemente el sitio web más reconocido, es el *Protein Data Bank (PDB)* [29] (http://www.rcsb.org/), en el que se proporciona la estructura tridimensional de un conjunto muy amplio de proteínas y ácidos nucleicos. Estos datos son obtenidos mediante cristalografía de rayos X o resonancia magnética nuclear (RMN), pero no siempre se menciona la ubicación de los sitios activos.

También existe el *Universal Protein Resource* (*UniProt*) [30], que incluye información sobre la secuencia y funcionalidad de muchísimas proteínas. En esta página (http://www.uniprot.org/) sí es común que se indiquen los residuos que forman los sitios activos de la proteína, incluyendo las características de cada residuo, es decir, si se trata de un sitio de unión o de un sitio catalítico. Si se trata de un sitio de unión, se suele agregar si es un sitio de unión de metales y de qué tipo de metales (Mg, Zn, etc.), o si es un sitio para la unión del sustrato.

Otra base de datos especializada justamente en enzimas, es el Catalytic Site Atlas (CSA) [31] (http://www.ebi.ac.uk/thornton-srv/databases/CSA/), que contiene información sólo de sitios catalíticos, por lo que en este lugar no se incluye información de los sitios de unión. UniProt y CSA proporcionan información de sitios activos ubi-

cados de forma experimental, aunque en el CSA también se incluyen sitios ubicados de manera computacional. Esta forma de ubicación consiste en buscar y comparar la estructura de una proteína con otra en la que se conocen ya los sitios activos. Se dice que se buscan estructuras *homólogas*. Dos proteínas son homólogas si tienen una estructura similar en algunas partes de la estructura y en la secuencia de aminoácidos [28, 32–34].

Se sabe que las proteínas homólogas suelen tener la misma función, por lo tanto existen también algunos sitios web que contienen algoritmos que revisan y comparan la estructura de una proteína, obtenida usualmente del PDB ya que el formato en el que se presenta la información en este sitio, es conocido, aceptado y estandarizado por todo el mundo, con la estructura de sitios activos conocidos. Si en esta comparación se encuentra alguna similitud, se propone la existencia y ubicación de un sitio activo en la nueva proteína. La forma como se hace esta comparación, varía entre un sistema y otro. Por ejemplo en [35], se hace además de la búsqueda de regiones homólogas, un análisis estadístico de las curvas de titulación de los aminoácidos ionizables de la proteína, pues se sabe que los que participan en sitios activos, tienen un comportamiento anómalo en esta propiedad.

En [33, 36] se proponen Algoritmos Genéticos (AG) para ubicar los sitios activos. Los algoritmos genéticos se llaman así porque están basados en la teoría de la evolución, en donde se modifica al azar un conjunto de soluciones (población), y se rescata a las mejores. En el caso de proteínas, un individuo es un conjunto de residuos que incluye al sitio activo y sus vecinos cercanos. Estos individuos se comparan geométricamente con sitios activos conocidos, y al final el algoritmo señala los mejores candidatos.

Otra forma teórica de buscar sitios activos, es por medio de datos estructurales y teoría de redes. En estos métodos, generalmente cada aminoácido es un nodo y las aristas se definen por las interacciones entre ellos [25, 27, 37, 38]. Esta metodología ha mostrado buenos resultados, aunque en general sobreestima la cantidad de sitios activos conocidos en la proteína. En un estudio hecho con teoría de redes, sobre una base de 178 proteínas de las que se conocen en conjunto 615 sitios activos [25], se ha identificado que el 65 % de los sitios catalíticos, están formados por residuos cargados (Asp, Glu, Lys, Arg e His), y el 27 % por residuos polares (Gln, Asg,

Tre, Cys, Ser, Trp, y Try), lo cual no es extraño si consideramos que en el proceso catalítico son importantes las fuerzas eléctricas para conseguir el movimiento de cargas. Estos porcentajes no están relacionados con la aparición de dichos residuos en la conformación de las enzimas, pues por ejemplo, la histidina, que forma parte del 18 % de todos los sitios catalíticos, tiene un porcentaje de abundancia de apenas el 2.7 %.

Con estos antecedentes, en este trabajo se propone para estudiar la estructura de las proteínas, una metodología basada en el modelo de redes complejas y de cadenas de Markov para procesos estocásticos. La propuesta fundamental es que es posible establecer un caminante aleatorio sobre la estructura tridimensional de la proteína, que se mueva de acuerdo a una probabilidad que definiremos en función de la cercanía entre los átomos de diferentes aminoácidos. Se utilizarán cadenas de Markov para analizar el proceso de este caminante y ubicar los aminoácidos que en promedio requieren la menor cantidad de pasos para llegar a ellos, partiendo de cualquier otro residuo de la proteína. Suponemos que esta cantidad está relacionada con los sitios activos, ya que esto contribuiría a hacer más eficiente la transmisión de información física o química con el resto de la estructura[37]. Por otro parte, se compararán diferentes medidas de centralidad para saber cuál de ellas ubica mejor la posición de dichos sitios. En los siguientes capítulos, se muestran las bases teóricas de esta propuesta.

Capítulo 2

Procesos Estocásticos y Cadenas de Markov

Se inicia este capítulo con una introducción a los procesos estocásticos conocidos como cadenas de Markov. Se verán algunas propiedades importantes de esta descripción, junto con algunos ejemplos sencillos para ilustrar estas propiedades. Posteriormente se revisará cómo puede hacerse la representación de la estructura de una proteína como una cadena de Markov a tiempo discreto, y cómo esta representación permite calcular la matriz de primera visita promedio y cómo se conecta esta matriz con la ubicación de sitios activos en la estructura, que es el tema principal de este trabajo.

2.1. Cadenas de Markov a tiempo discreto

Suponga de manera general, un sistema que puede estar en cualquier estado dentro de un conjunto previamente especificado, y que después de un tiempo el sistema evoluciona o modifica su estado a otro dentro del mismo conjunto. Sea S_t el estado del sistema al tiempo t. Si este cambio se ha dado por razones no deterministas sino por un mecanismo azaroso, entonces puede considerarse que S_t representa el estado de la *variable aleatoria* S para cada valor del índice t, y que esta colección de estados es la definición de *proceso estocástico* [39].

De forma más precisa, si S es una variable aleatoria, un Proceso Estocástico para

esta variable, es la colección de estados $\{S_t : t \in \mathbf{T}\}$ parametrizada por un conjunto $\mathbf{T}$, llamado *espacio parametral*, en donde la variable toma valores de un conjunto $\mathbf{S}$ llamado espacio de estados.

Si el espacio parametral $\mathbf{T}$ es un conjunto discreto $\mathbf{T} = \{1, 2, ...\}$ y estos números se interpretan como tiempos, se dice entonces que el proceso es a *tiempo discreto*, y S_t es el estado del sistema al tiempo t. Existen muchos tipos de proceso estocásticos, que se diferencian de acuerdo a las distintas posibilidades para los espacios parametrales y los estados, y a las relaciones de dependencia que puede haber entre las variables aleatorias [39]. El proceso estocástico que será de interés en este trabajo es justo un *Proceso de Markov* o *Cadena de Markov*.

Una *Cadena de Markov* es un proceso estocástico en el que el valor del estado actual, S_t, sólo depende del estado inmediato anterior S_{t-1}. De la misma forma, la probabilidad de acceder al estado futuro S_{t+1} sólo dependerá del estado actual. Cada vez que el proceso se mueve de un estado a otro, se dice que ha dado un *paso*. De manera formal y en notación de probabilidad condicional, una *Cadena de Markov* a tiempo discreto es una sucesión de estados $\{S_t\} = \{S_0, S_1, S_{n-1}, S_n, ...\}$ que cumplen la igualdad

$$P(S_{n+1} = j | S_n = i, S_{n-1}, ..., S_0) = P(S_{n+1} = j | S_n = i) \tag{2.1}$$

Para toda n y cualquier conjunto $\{S_t\}$ de estados de la variable aleatoria. La expresión anterior significa que la probabilidad condicional de que el próximo estado del sistema sea $S_{n+1} = j$, dado que actualmente está en el estado $S_n = i$, no depende de los estados previos del sistema. Por eso se suele decir que las cadenas de Markov son procesos que no tienen memoria. Podemos sin perder generalidad, suponer que el conjunto de estados de la cadena de Markov es el conjunto $\mathbf{S} = \{1, 2, ...\}$. Si este conjunto es finito, se dice que la cadena también es finita.

Por lo tanto, si i y j son dos estados de una cadena de Markov, a la probabilidad condicional de que a partir del estado $S_n = i$, ocurra el estado $S_{n+1} = j$, se le denota p_{ij} y se obtiene

$$p_{ij} = P(S_{n+1} = j | S_n = i) \tag{2.2}$$

A p_{ij} se conoce como *probabilidad de transición*. Este término representa la pro-

babilidad de que a partir del estado S_n, ocurra el estado S_{n+1}. Las probabilidades de transición de los diferentes eventos de una cadena de Markov pueden representarse en forma matricial y a esta matriz se le denomina *Matriz de Transición* $\mathbf{M}$. Observe que, dado que los elementos m_{ij} de la matriz de transición representan una probabilidad, todos deben estar en el intervalo $0 \leq m_{ij} \leq 1$. Además, la suma de los elementos de un renglón es la suma de la probabilidad de pasar del estado i a todos los demás, por lo tanto $\sum_k m_{ik} = 1$. Toda matriz cuadrada que cumpla estas dos condiciones, es una *matriz estocástica*.

Con frecuencia es útil representar estos procesos estocásticos de forma gráfica como en la fig. (2.1). Estas representaciones se denominan *redes* o *grafos*. Cada estado del proceso es un *nodo* o *vértice*, y las líneas que los unen llamadas *aristas* o *enlaces*, representan la posibilidad de pasar de un estado a otro en la dirección indicada por la flecha. Si no hay flecha entonces el proceso puede ocurrir en ambas direcciones. En el ejemplo mostrado, la red es *ponderada* pues se indica en los *enlaces* la probabilidad de transición entre los estados. El paso del estado **1** al **2**, ocurre con una probabilidad igual a 1. El sistema puede permanecer en el estado **2** con una probabilidad de 2/3, o pasar al **3**, con probabilidad de 1/3. Observe que de acuerdo a las flechas, desde el estado **2** el sistema no puede regresar a **1**, en cambio desde el estado **3** sí es posible moverse a **2** o a **1**, con la probabilidad indicada en la figura.

Otra forma de interpretar un proceso estocástico es en términos de un *Caminante Aleatorio*, es decir un caminante que avanza un paso entre los estados i y j, cada vez que el sistema pasa del estado i al j. Así, un caminante que se encuentra en el estado i, se moverá al estado j en un solo paso, con una probabilidad p_{ij} dada por la ec. 2.2.

Para la cadena de Markov representada por la figura (2.1), la matriz de transición $\mathbf{M}$ es

$$\mathbf{M} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 2/3 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix}. \tag{2.3}$$

Es importante mencionar que esta forma de escribir una matriz de transición no es universal, pues algunos autores definen m_{ij} como la probabilidad de ir de j a i, y

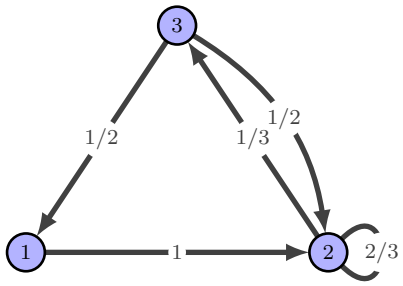


Figura 2.1: Los vértices en la figura representan los estados que el sistema puede tomar; y las aristas o enlaces, la probabilidad de ir de un estado a otro en un paso.

en este caso, es la suma de los elementos de cada columna de la matriz lo que debe ser igual a 1:

$$\sum_{k=1}^{r} m_{ki} = 1. \tag{2.4}$$

La expresión $m_{ij}^{(2)}$ denota la probabilidad de que una cadena de Markov que está en el estado i, pase al estado j en dos pasos. Si la cadena consta de r estados, entonces $m_{ij}^{(2)}$ es la suma de los productos $m_{ik}m_{kj}$ sobre todos los estados

$$m_{ij}^{(2)} = \sum_{k=1}^{r} m_{ik}m_{kj}, \tag{2.5}$$

es decir, m_{ij}^2 es el término *ij-ésimo* de la matriz $\mathbf{M}^2$. En general, se puede demostrar [40] que si $\mathbf{M}^n$ es la n-ésima potencia de la matriz de transición $\mathbf{M}$, los términos $m_{ij}^{(n)}$, representan la probabilidad de pasar del estado i al estado j en n pasos.

El proceso representado por una cadena de Markov, puede comenzar desde cualquier estado i, o bien, puede suponerse que comienza con una distribución inicial de probabilidad sobre todos los estados. A esta distribución se le llama *vector de probabilidad*, $\mathbf{u}^{(0)}$, y debe ser tal que sus componentes estén en $0 \leq u_k^{(0)} \leq 1$ y además $\sum_k u_k^{(0)} = 1$. La componente i-ésima de $\mathbf{u}^{(0)}$ es entonces la probabilidad de que la cadena comience en el estado i.

Si $\mathbf{u}^{(1)}$ es el vector de probabilidad cuya j-ésima entrada representa la probabilidad de alcanzar el estado j después de un paso, entonces

$$u_j^{(1)} = \sum_{i=1}^{n} u_i^{(0)} m_{ij} \tag{2.6}$$

Y en notación matricial

$$\mathbf{u}^{(1)} = \mathbf{u}^{(0)}\mathbf{M} \tag{2.7}$$

Y del mismo modo, si $\mathbf{u}^{(2)}$ es el vector de probabilidad de alcanzar los estados $1, 2, ..., n$ en dos pasos, entonces

$$\mathbf{u}^{(2)} = \mathbf{u}^{(1)}\mathbf{M} = \mathbf{u}^{(0)}\mathbf{M}^2 \tag{2.8}$$

Y generalizando para n pasos

$$\mathbf{u}^{(n)} = \mathbf{u}^{(n-1)}\mathbf{M} = \mathbf{u}^{(0)}\mathbf{M}^n \tag{2.9}$$

2.1.1. Ejemplo

Para ilustrar cómo interpretar los elementos de la matriz de transición y los vectores de probabilidad, planteamos el siguiente ejemplo sencillo. Suponga que en la ciudad se dan 3 posibles climas: Lluvia *(ll)*, Sol *(s)* y nieve *(n)*. Si hoy es día de nieve o lluvia, la probabilidad de que mañana ocurra el mismo clima es de 1/2, y de que cambie a cualquiera de los otros dos, de 1/4. En cambio si hoy es un día soleado, la probabilidad de que mañana haya nieve n o lluvia ll es 1/2, por lo que la probabilidad de que se repita otro día de sol es cero. Esquemáticamente, este problema está representado en la figura (2.2).

La matriz de transición que representa estas condiciones es, asumiendo que el orden en columnas y renglones es *ll, s, n*

$$\mathbf{M} = \begin{array}{c} \\ ll \\ s \\ n \end{array} \begin{array}{c} \begin{array}{ccc} ll & s & n \end{array} \\ \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix} \end{array} \tag{2.10}$$

Por lo tanto, si calculamos por ejemplo $\mathbf{M}^3$, cada término m_{ij}^3 indica la probabi-

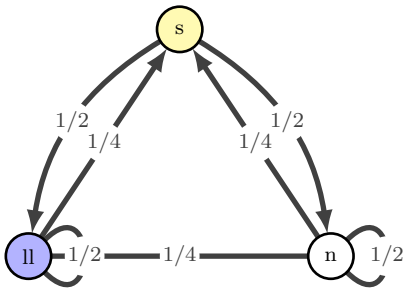


Figura 2.2: Los nodos son los tres posibles climas, y las aristas indican la probabilidad de pasar de uno a otro en un día.

lidad condicional de que en 3 días ocurra el clima j, dado que hoy tenemos el clima i.

En cambio, si queremos obtener la distribución de probabilidad para el clima en 3 días, calculamos $\mathbf{u}^{(3)} = (u_l^3, u_s^3, u_n^3)$. Para esto necesitamos una distribución inicial. Si asumimos que hoy todos los climas son igualmente probables, tenemos

$$\mathbf{u}^{(3)} = \mathbf{u}^0\mathbf{M}^3 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \begin{pmatrix} 0.406 & 0.203 & 0.391 \\ 0.406 & 0.188 & 0.406 \\ 0.391 & 0.203 & 0.406 \end{pmatrix} \tag{2.11}$$

se obtiene $\mathbf{u}^{(3)} = (0.401,\ 0.198,\ 0.401)$, cuyas entradas son la probabilidad de que haya lluvia, sol o nieve dentro de 3 días, respectivamente.

Y si quisiéramos obtener la probabilidad de que en tres días ocurra el clima j, dado que hoy el clima es i, tendríamos que utilizar como vector de probabilidad inicial, un vector con todas sus componentes igual a cero, excepto la componente i, que sería 1. Así, si queremos saber la probabilidad de tener un día soleado dentro de tres días, dado que hoy es un día soleado, el vector $\mathbf{u}^{(0)}$ es $(0, 1, 0)$, por lo tanto

$$\mathbf{u}^{(3)} = \mathbf{u}^{(0)}\mathbf{M}^3 = (0, 1, 0) \begin{pmatrix} 0.406 & 0.203 & 0.391 \\ 0.406 & 0.188 & 0.406 \\ 0.391 & 0.203 & 0.406 \end{pmatrix} = \begin{pmatrix} 0.406 \\ 0.188 \\ 0.406 \end{pmatrix} \tag{2.12}$$

y el resultado que buscamos es $u_s^3 = 0.188$, es decir, la componente m_{ss}^3 de la

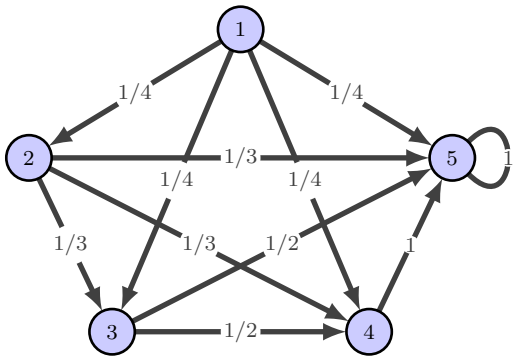


Figura 2.3: En este proceso, el sistema sólo puede moverse a un estado etiquetado con un número mayor. Al llegar a 5, el proceso ya no puede ir a otra parte, por lo tanto el estado 5 es absorbente.

matriz $\mathbf{M}^3$.

2.1.2. Cadenas de Markov Absorbentes

Si un estado i de una cadena de Markov, tiene una probabilidad de transición $p_{ii} = 1$, significa que la probabilidad de evolucionar a otro estado es cero, por lo tanto es imposible que el sistema abandone el estado i. Se dice que dicho estado es *absorbente*. Un estado se denomina *transitorio* cuando no es absorbente.

Una cadena de Markov es *absorbente* si tiene al menos un estado absorbente y si además es posible ir de cualquier otro estado de la cadena al estado absorbente. A continuación discutiremos un ejemplo, con el que analizaremos también algunas propiedades de las cadenas absorbentes.

Sea un proceso que se mueve sobre los enteros $1, 2, 3, 4$ y 5. El proceso en cada paso, se mueve con igual probabilidad sólo a un entero mayor. Gráficamente, este proceso está representado en la fig. (2.3).

Se tiene para este problema la siguiente matriz de transición

$$\mathbf{M} = \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.13}$$

En general, cuando el o los estados transitorios están colocados al principio y los absorbentes al final en la matriz, se dice que está en su forma *canónica*, lo que nos permite agruparla de la siguiente manera

$$\mathbf{M} = \left(\begin{array}{c|c} \mathbf{Q} & \mathbf{R} \\ \hline \mathbf{0} & \mathbf{I} \end{array}\right). \tag{2.14}$$

En este ejemplo particular

$$\mathbf{Q} = \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{3} \\ \frac{1}{2} \\ 1 \end{pmatrix}. \tag{2.15}$$

0 es una matriz formada por ceros, y la matriz **I** la matriz identidad. En general, si hay t estados transitorios y r estados absorbentes, **M** está formada por la matriz **Q** de $t \times t$, la matriz **R** de $t \times r$, la matriz **0** de $r \times t$, y la matriz identidad **I** de $r \times r$.

Una de las ventajas de escribir **M** en su forma canónica, es que es fácil obtener sus potencias. Puede verse que

$$\mathbf{M}^2 = \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{Q}^2 & (\mathbf{I}+\mathbf{Q})\mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}. \tag{2.16}$$

Y de la misma forma se obtiene $\mathbf{M}^3$

$$\mathbf{M}^3 = \begin{pmatrix} \mathbf{Q}^3 & (\mathbf{I}+\mathbf{Q}+\mathbf{Q}^2)\mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}. \tag{2.17}$$

Así que en general

$$\mathbf{M}^n = \begin{pmatrix} \mathbf{Q}^n & (\mathbf{I}+\mathbf{Q}+\mathbf{Q}^2+...+\mathbf{Q}^{n-1})\mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}. \tag{2.18}$$

Y puesto que es posible ir de cualquier estado transitorio a uno absorbente, $0 \leq q_{ij} < 1$ para todos los términos de **Q**, por lo tanto $\mathbf{Q}^n \to \mathbf{0}$ cuando $n \to \infty$. Esto significa que la probabilidad de continuar indefinidamente en los estados transitorios es cero, es decir, en una cadena de Markov absorbente el proceso llega siempre a

alguno de los estados absorbentes.

Por otra parte, para cadenas de Markov absorbentes, se llama $\mathbf{N}$ a la *Matriz Fundamental* para $\mathbf{M}$, definida como

$$\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1} \tag{2.19}$$

Note que

$$(\mathbf{I} - \mathbf{Q})(\mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + ... + \mathbf{Q}^n) = \mathbf{I} - \mathbf{Q}^{n+1}$$

y al multiplicar por $\mathbf{N}$ ambos lados de la expresión:

$$\mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + ... + \mathbf{Q}^n = \mathbf{N}(\mathbf{I} - \mathbf{Q}^{n+1})$$

Así que cuando $n \to \infty$, se obtiene

$$\mathbf{N} = \mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + ... \tag{2.20}$$

Por lo tanto, los términos n_{ij} de $\mathbf{N}$ son la suma de la probabilidad de ir del estado i al j en $1, 2, ..., n$ pasos. Esto también puede interpretarse como el tiempo o el número esperado de veces que el proceso pasó por el estado j dado que comenzó en el estado i, antes de llegar a un estado absorbente. Esta interpretación es muy importante para los procesos que analizaremos en este trabajo, así que la usaremos en ejemplos posteriores.

En función de $\mathbf{N}$, cuando $n \to \infty$, la matriz de transición puede escribirse

$$\lim_{n\to\infty} \mathbf{M}^n = \begin{pmatrix} \mathbf{0} & \mathbf{NR} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}. \tag{2.21}$$

Se suele llamar *matriz de probabilidad de absorción*, a $\mathbf{B} = \mathbf{NR}$. Se puede demostrar que los elementos b_{ij} de esta matriz, representan la probabilidad de que un sistema termine absorbido en el estado absorbente j, dado que comenzó en el estado transitorio i.

Para el ejemplo de esta sección

$$\mathbf{I\text{-}Q} = \begin{pmatrix} 1 & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ 0 & 1 & -\frac{1}{3} & -\frac{1}{3} \\ 0 & 0 & 1 & -\frac{1}{2} \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{2.22}$$

por lo tanto la matriz fundamental es

$$\mathbf{N} = (\mathbf{I\text{-}Q})^{-1} = \begin{pmatrix} 1 & \frac{1}{4} & \frac{1}{3} & \frac{1}{2} \\ 0 & 1 & \frac{1}{3} & \frac{1}{2} \\ 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.23}$$

Obtenemos una matriz diagonal con $n_{ij} = 0$, para $i > j$, lo cual tiene sentido pues de acuerdo al problema, el sistema no puede evolucionar a un estado j tal que $i > j$. Además la diagonal de la matriz es 1, pues el sistema sólo permanece un paso en ese estado, ya que después de la primera evolución no puede regresar. Por otra parte, el producto **NR** es

$$\mathbf{NR} = \mathbf{B} = \begin{pmatrix} 1 & \frac{1}{4} & \frac{1}{3} & \frac{1}{2} \\ 0 & 1 & \frac{1}{3} & \frac{1}{2} \\ 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{4} \\ \frac{1}{3} \\ \frac{1}{2} \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}. \tag{2.24}$$

Así que si $n \to \infty$, obtenemos

$$\lim_{n\to\infty} \mathbf{M}^n = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.25}$$

Lo que muestra, como era de esperarse, que como el sistema sólo avanza hacia estados de número más grande, después de un tiempo termina en el estado 5, independientemente del estado inicial.

En general, si sumamos $\sum_j n_{ij}$, obtenemos el número de veces que se espera que

el sistema pase por todos los estados transitorios antes de ser absorbido, dado que empezó en el estado i. A esta cantidad se le llama *tiempo de absorción del estado i,* t_i. En notación vectorial, el vector tiempo de absorción $\mathbf{t}$, es

$$\mathbf{t} = \mathbf{Nc}, \tag{2.26}$$

donde $\mathbf{c}$ es un vector columna cuyas entradas son todas igual a 1. Para el ejemplo que nos ha ocupado se obtiene

$$\mathbf{t} = \begin{pmatrix} \frac{25}{12} \\ \frac{11}{6} \\ \frac{3}{2} \\ 1 \end{pmatrix}. \tag{2.27}$$

2.1.3. Cadenas de Markov Ergódicas

Una cadena de Markov *ergódica* o *irreducible* es aquella en la que es posible ir de un estado a cualquier otro, no necesariamente en un paso. Note que una cadena absorbente no es ergódica, pues como se mencionó, no es posible ir del estado absorbente a ningún otro estado.

Se dice que una cadena de Markov es *regular*, si alguna potencia n de su matriz de transición tiene únicamente elementos $0 < m_{ij}^n < 1$. Esto significa que para alguna n, es posible ir de un estado i a cualquier otro estado j, en exactamente n pasos. Todas las cadenas regulares son ergódicas, pero no todas las ergódicas son regulares. De igual manera, una matriz de transición con todos sus elementos diferentes de cero representa a una cadena de Markov regular, sin embargo puede haber cadenas de Markov regulares, que tengan matriz de transición con elementos iguales a cero, siempre y cuando para alguna potencia n desaparezcan los términos cero. Una cadena de Markov absorbente es un ejemplo de cadena no regular, pues dado que el sistema termina por llegar a los estados absorbentes a cualquier potencia de la matriz de transición, los elementos del renglón del estado absorbente, serán siempre cero.

Una propiedad importante de las cadenas de Markov regulares, tiene que ver justamente con las potencias de la matriz de transición. Si $\mathbf{M}$ es la matriz de transición

de una cadena de Markov regular, entonces

$$\mathbf{W} = \lim_{n\to\infty} \mathbf{M}^n$$

Es decir, si $n \to \infty$, $\mathbf{M}^n$ converge a la matriz $\mathbf{W}$, donde todos los renglones de $\mathbf{W}$ son el mismo vector $\mathbf{w}$. Además todos los elementos w_{ij} de $\mathbf{W}$, satisfacen que $0 \leq w_{ij} \leq 1$ y $\sum_j w_{ij} = 1$ para todos los valores de i. Por lo tanto, cada renglón de $\mathbf{W}$ es un vector de probabilidad.

Además, como $\lim_{n\to\infty} \mathbf{M}^n \to \mathbf{W}$, entonces

$$\lim_{n\to\infty} \mathbf{M}^{n+1} = \lim_{n\to\infty} \mathbf{M}^n\mathbf{M} \to \mathbf{WM}$$

.

Pero también $\lim_{n\to\infty} \mathbf{M}^{n+1} \to \mathbf{W}$, así que

$$\mathbf{W} = \mathbf{WM}$$

.

Por lo tanto para cada renglón de $\mathbf{W}$ se cumple $\mathbf{w} = \mathbf{wM}$.

Al vector $\mathbf{w}$ se le llama *vector renglón fijo* de $\mathbf{M}$. Podemos ver que $\mathbf{w}$ es un eigenvector de la matriz de transición $\mathbf{M}$, al que le corresponde el eigenvalor 1.

Regresando al ejemplo analizado en la sección 2.1.1 respecto a los tres tipos de clima en una ciudad. Vemos que en la matriz de transición (2.10) el elemento $m_{22} = 0$, pero al calcular $\mathbf{M}^2$

$$\mathbf{M}^2 = \begin{pmatrix} \frac{7}{16} & \frac{3}{16} & \frac{3}{8} \\ \frac{3}{8} & \frac{1}{4} & \frac{3}{8} \\ \frac{3}{8} & \frac{3}{16} & \frac{7}{16} \end{pmatrix} \tag{2.28}$$

Ya ningún elemento es igual a cero, por lo que la cadena de Markov del sistema es regular. Si continuamos calculando las potencias de $\mathbf{M}$, obtenemos por ejemplo:

$$\mathbf{M}^5 = \begin{pmatrix} 0.400 & 0.200 & 0.399 \\ 0.400 & 0.199 & 0.400 \\ 0.399 & 0.200 & 0.400 \end{pmatrix}, \mathbf{M}^6 = \begin{pmatrix} 0.400 & 0.200 & 0.400 \\ 0.400 & 0.200 & 0.400 \\ 0.400 & 0.200 & 0.400 \end{pmatrix}. \tag{2.29}$$

Notamos que a la sexta potencia, los renglones de $\mathbf{M}$ ya son iguales, y el vector renglón fijo es $\mathbf{w} = (0.4, 0.2, 0.4)$.

Evidentemente, esta forma de calcular el vector renglón fijo no es la mejor. Otra forma sería precisamente calculando el eigenvector de $\mathbf{M}$ al que le corresponde el eigenvalor 1, y utilizando el hecho de que es un vector de probabilidad, es decir $\sum_j w_j = 1$. Continuando con el ejemplo anterior, debemos resolver

$$(w_1, w_2, w_3) = (w_1, w_2, w_3) \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix} \tag{2.30}$$

Lo cual arroja el sistema de ecuaciones

$$\begin{aligned} w_1 + w_2 + w_3 &= 1 \\ \tfrac{1}{2}w_1 + \tfrac{1}{2}w_2 + \tfrac{1}{4}w_3 &= w_1 \\ \tfrac{1}{4}w_1 + \tfrac{1}{4}w_3 &= w_2 \\ \tfrac{1}{4}w_1 + \tfrac{1}{2}w_2 + \tfrac{1}{2}w_3 &= w_3 \end{aligned} \tag{2.31}$$

Y obtenemos el mismo resultado para $\mathbf{w}$ que al calcular las potencias de $\mathbf{M}$, es decir en este caso $\mathbf{w} = (0.4, 0.2, 0.4)$.

Se puede probar que para una cadena de Markov ergódica, existe un vector renglón fijo único, y cualquier otro vector renglón $\mathbf{v}$ que satisfaga $\mathbf{vM} = \mathbf{v}$, debe ser múltiplo de $\mathbf{w}$. Esto es importante porque nos indica que para una n suficientemente grande $m_{1j}^n = m_{2j}^n = ... = m_{ij}^n$, es decir, la probabilidad de que ocurra el evento j es independiente de la condición inicial.

2.2. Tiempo Medio de Primer Visita (Mean First Passage Time) para cadenas ergódicas

Un importante concepto dentro de los procesos estocásticos, que será fundamental en este trabajo de investigación, es el *tiempo medio de primer visita* (*mean first passage time* en inglés), el cual se refiere al número de pasos que en promedio debe dar un sistema para que, partiendo del estado i, llegue por primera vez al estado j. A esta cantidad la llamamos h_{ij} y $\mathbf{H}$ a la matriz de todos estos elementos, es decir la *matriz de primeras visitas*. Por convención $h_{ii} = 0$. No obstante, para estar acorde con la notación que sigue nuestra principal referencia [38], llamaremos $H(j, i)$ a los elementos de $\mathbf{H}$. Así, $H(j, i)$ es el número de pasos que en promedio da un sistema para ir del estado i por primera vez al estado j.

Una forma de calcular el tiempo de primer visita al estado j partiendo del estado i en una cadena de Markov ergódica, será construir una nueva cadena de Markov, en todo semejante a la original, salvo que ahora el estado j es absorbente. Esto se consigue haciendo $m_{jj} = 1$ en la matriz de transición. El comportamiento de este nuevo sistema será en todo similar al original, y puesto que la cadena original es ergódica, el sistema necesariamente tendrá que llegar en algún momento al estado j.

Podemos entonces calcular para la nueva cadena, la matriz fundamental $\mathbf{N}$, definida en la ecuación (2.19), cuyos términos como hemos dicho, nos dan el tiempo que en promedio pasa el sistema en cada estado. Con esta información, calculamos el vector de tiempo de absorción $\mathbf{t}$ dado por la ec. (2.26), lo cual significa calcular para la nueva cadena, el tiempo que en promedio tarda el sistema que comenzó en el estado i en ser absorbido. Este tiempo para la cadena original es justamente el tiempo medio de primera visita al estado j.

A continuación resolveremos un ejemplo para ilustrar los conceptos anteriores. Supongamos que un ratón es colocado en una caja en la que hay 6 espacios conectados entre sí, como se ilustra en la figura (2.4). El ratón se mueve al azar por los espacios, de forma que la probabilidad de pasar de uno a otro depende de la cantidad de entradas que tenga cada compartimento. El grafo que ilustra este sistema se muestra en la fig. (2.5)

La matriz de transición para este problema es

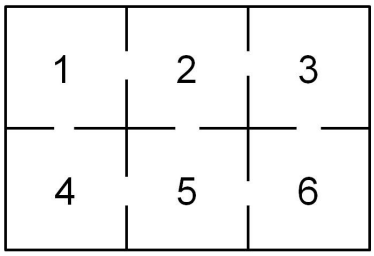


Figura 2.4: Conjunto de espacios en los cuales, un ratón puede moverse al azar. El número de entradas en cada espacio, indica la probabilidad de moverse de ese estado a los demás.

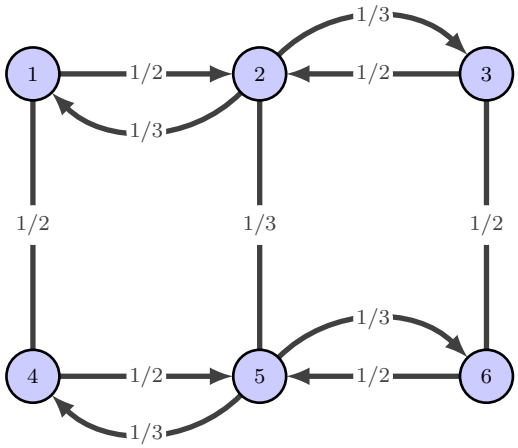


Figura 2.5: Grafo de la cadena que representa a un ratón en el interior de una caja con 6 espacios.

$$\mathbf{M} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix} \end{array} \tag{2.32}$$

Este proceso es ergódico pues siempre es posible ir a cualquier espacio a partir de cualquier otro, aunque no es regular pues por ejemplo, no es posible ir de un compartimento marcado con un número impar, a otro impar, en un número impar de pasos. El vector renglón fijo del sistema $\mathbf{w}$, se obtiene calculando el eigenvector de $\mathbf{M}$ al que le corresponde el eigenvalor 1 y utilizando el hecho de que $\mathbf{w}$ es un vector de probabilidad. Se plantea un sistema de 6 ecuaciones con 6 incógnitas, del cual obtenemos

$$\mathbf{w} = \frac{1}{14}(2, 3, 2, 2, 3, 2)$$

Supongamos ahora que queremos calcular el tiempo de primer visita al espacio marcado con el número 5, partiendo de cualquier otro estado, es decir, vamos a calcular los elementos $H(5, i)$ de la matriz $\mathbf{H}$. Para eso modificamos el sistema original y ahora consideramos que el estado 5 es absorbente, digamos que se ha colocado en ese compartimiento un queso (figura (2.6)) de tal manera que el ratón al llegar ahí, no desea salir. El grafo (2.7) ilustra el problema.

La matriz de transición escrita en forma canónica es ahora

$$\mathbf{M} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 6 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 6 & 5 \end{array} \\ \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \hline 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \tag{2.33}$$

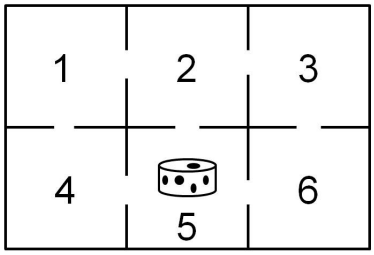


Figura 2.6: Al hacer que el estado 5 se vuelva absorbente, podemos calcular el tiempo medio que el ratón está en el resto de los estados antes de ser absorbida.

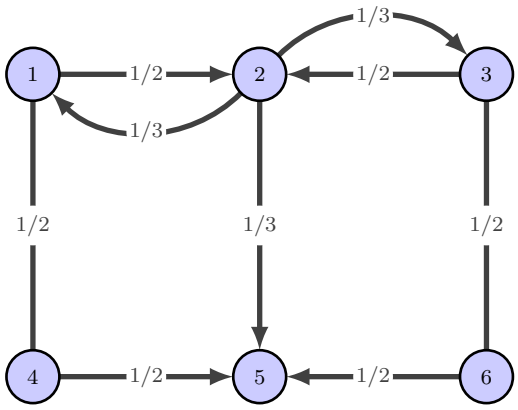


Figura 2.7: Grafo de la cadena que representa a un ratón en el interior de una caja en la que ahora, el espacio 5 es absorbente.

Por lo que

$$\mathbf{I\text{-}Q} = \begin{pmatrix} 1 & -\frac{1}{2} & 0 & -\frac{1}{2} & 0 \\ -\frac{1}{3} & 1 & -\frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{2} & 1 & 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{2} & 0 & 1 \end{pmatrix}, \tag{2.34}$$

e invirtiendo la matriz anterior, encontramos la matriz fundamental $\mathbf{N}$

$$\mathbf{N} = (\mathbf{I\text{-}Q})^{-1} = \begin{pmatrix} \frac{28}{15} & \frac{6}{5} & \frac{8}{15} & \frac{14}{15} & \frac{4}{15} \\ \frac{4}{5} & \frac{9}{5} & \frac{4}{5} & \frac{2}{5} & \frac{2}{5} \\ \frac{8}{15} & \frac{6}{5} & \frac{28}{15} & \frac{4}{15} & \frac{14}{15} \\ \frac{14}{15} & \frac{3}{5} & \frac{4}{15} & \frac{22}{15} & \frac{2}{15} \\ \frac{4}{15} & \frac{3}{5} & \frac{14}{15} & \frac{2}{15} & \frac{22}{15} \end{pmatrix}. \tag{2.35}$$

Así que el vector de tiempo medio de absorción es

$$\mathbf{t} = \mathbf{Nc} = \begin{pmatrix} 4.8 \\ 4.2 \\ 4.8 \\ 3.4 \\ 3.4 \end{pmatrix}. \tag{2.36}$$

Por lo tanto, los elementos $H(5, i)$ son

$$H(5, i) = (4.8, 4.2, 4.8, 3.4, 0, 3.4) \tag{2.37}$$

Vemos que si el ratón comienza en el espacio 1, le toma en promedio 4.8 pasos llegar a la comida. Por la simetría de la caja, era de esperar que le tome el mismo número de pasos si comienza en el espacio 3, y también que le tome menos pasos si comienza en el espacio 2, y aún menos, si comienza en los espacios 4 o 6, pues al tener menos entradas estos compartimientos que el 2, tiene el ratón una mayor probabilidad de llegar al queso a partir de estos que partiendo de 2.

2.2.1. Cálculo de la matriz H

El método anterior nos permite calcular los elementos $H(j,i)$ para una j fija, es decir, calcular el tiempo medio de primeras visitas para un estado del sistema en particular, así que para obtener la matriz completa **H** debemos aplicarlo j veces, lo cual puede resultar poco práctico.

Otra manera de calcular los elementos $H(j,i)$ será usando una fórmula recursiva que consiste en suponer el paso de i a j, a través de un nodo intermedio k. Se considera la probabilidad de ir de i a k en un paso, más los pasos que en promedio son necesarios para pasar de k a j asumiendo que se conoce $H(j,k)$. Esto se suma sobre todos los estados del sistema

$$H(j,i) = \sum_{k=1}^{n} (1 + H(j,k)) m_{ik} \tag{2.38}$$

desarrollando la suma

$$H(j,i) = \sum_{k=1}^{n} m_{ik} + \sum_{k=1}^{n} H(j,k) m_{ik} = 1 + \sum_{k=1}^{n} H(j,k) m_{ik} \tag{2.39}$$

Como $H(j,j) = 0$, quitamos este término de la suma, obteniendo

$$H(j,i) = 1 + \sum_{k=1, k\neq j}^{n} H(j,k) m_{ik} \tag{2.40}$$

Esta expresión también puede escribirse

$$(1 - m_{ii})\, H(j,i) - \sum_{k=1, k\neq i,j}^{n} m_{ik} H(j,k) = 1. \tag{2.41}$$

Si suponemos que el sistema tiene necesariamente que cambiar de estado en cada paso, tenemos que $m_{ii} = 0$ para toda i, por lo tanto la ecuación queda

$$H(j,i) = 1 + \sum_{k=1, k\neq i,j}^{n} m_{ik} H(j,k). \tag{2.42}$$

La matriz de transición no es simétrica, por lo que **H** tampoco, así que en general

$H(j,i) \neq H(i,j)$.

Por convención $H(i,i) = 0$, lo cual significa que no es necesario dar ningún paso para ir del estado s_i a s_i. No obstante, es posible calcular el número de pasos que en promedio deben darse para regresar por primera vez al estado inicial. A esta cantidad se le llama *tiempo de recurrencia media* r_i. A la matriz $\mathbf{D}$ con todas sus entradas igual a cero, salvo la diagonal, en donde $d_{ii} = r_i$, se le llama *matriz de recurrencia media*. El proceso para obtener de forma recursiva r_i es semejante al descrito para los elementos de $\mathbf{H}$, es decir, suponemos la probabilidad de llegar en un paso al estado k y a continuación el promedio de pasos necesarios para ir de k al estado original i

$$r_i = 1 + \sum_k m_{ik} H(i,k) \tag{2.43}$$

Las ecuaciones (2.42) y (2.43) pueden unirse en una representación matricial, usando $\mathbf{C}$ definida como una matriz en la que todos sus elementos son 1

$$\mathbf{H} = \mathbf{MH} + \mathbf{C} - \mathbf{D} \tag{2.44}$$

o bien

$$(\mathbf{I} - \mathbf{M})\,\mathbf{H} = \mathbf{C} - \mathbf{D} \tag{2.45}$$

Una propiedad importante del tiempo de recurrencia medio, se puede deducir si multiplicamos la ec. (2.45) por el vector $\mathbf{w}$

$$\mathbf{w}\,(\mathbf{I} - \mathbf{M})\,\mathbf{H} = \mathbf{wC} - \mathbf{wD}$$

pero

$$\mathbf{wI} - \mathbf{wM} = \mathbf{w} - \mathbf{w} = 0,$$

por lo tanto

$$\mathbf{wC} = \mathbf{wD}$$

es decir

$$(1, 1, ..., 1) = (w_1 r_1, w_2 r_2, ...)$$

así que para todos los estados en una cadena de Markov ergódica, el tiempo medio de recurrencia es el inverso del vector de probabilidad fijo

$$r_i = \frac{1}{w_i}. \tag{2.46}$$

Podemos a continuación, utilizar la ecuación (2.42) para resolver el problema del ratón en la caja y encontrar el tiempo medio de primer visita al espacio 5.

$$H(5,i) - \sum_{k=1,k\neq i,5}^{6} m_{ik}H(5,k) = 1. \tag{2.47}$$

Esta ecuación representa un conjunto de solamente 5 ecuaciones simultáneas, pues ya sabemos que $H(5,5) = 0$. Tenemos entonces

$$\begin{aligned}
&H(5,1) - m_{12}H(5,2) - m_{13}H(5,3) - m_{14}H(5,4) - m_{16}H(5,6) = 1,\\
&H(5,2) - m_{21}H(5,1) - m_{23}H(5,3) - m_{24}H(5,4) - m_{26}H(5,6) = 1,\\
&H(5,3) - m_{31}H(5,1) - m_{32}H(5,2) - m_{34}H(5,4) - m_{36}H(5,6) = 1,\\
&H(5,4) - m_{41}H(5,1) - m_{42}H(5,2) - m_{43}H(5,3) - m_{46}H(5,6) = 1,\\
&H(5,6) - m_{61}H(5,1) - m_{62}H(5,2) - m_{63}H(5,3) - m_{64}H(5,4) = 1.
\end{aligned} \tag{2.48}$$

donde los elementos m_{ik} están dados por la matriz de transición del problema original, es decir, la expresión (2.32), no la que resulta de convertir la cadena a una absorbente. Resolviendo el sistema, encontramos efectivamente, los mismos valores que fueron hallados con la matriz fundamental del sistema absorbente, ecuación (2.36).

$$H(5,i) = (4.8, 4.2, 4.8, 3.4, 0, 3.4)$$

2.2.2. Otros ejemplos analíticos simples: 3 estados

Para ilustrar el cálculo de **H**, su significado físico e interpretación, resolveremos analíticamente los casos sencillos de un sistema de 3 y de 4 estados.

Primero consideramos un sistema con 3 estados conectados formando una línea recta (figura (2.8)), con la misma probabilidad de moverse de un estado a otro.

Si el sistema comienza en el estado 1 o en el 3, sólo puede trasladarse al estado 2. En cambio, si está en el estado 2, supondremos que tiene la misma probabilidad de cambiar al 1 o al 3. La matriz de transición **M** en este caso es

Figura 2.8: 3 estados en línea recta.

$$\mathbf{M} = \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ 1 & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{pmatrix} \tag{2.49}$$

Estamos ahora siguiendo la notación en donde los elementos de $\mathbf{M}$ representan la probabilidad de ir del estado j al i. La suma de los términos por columna es 1.

Y ahora podemos obtener los elementos de la matriz $H(j,i)$, aplicando la ecuaciones de recurrencia (2.40). Recordemos que los elementos de la diagonal de $\mathbf{H}$ son cero, así que no los consideramos en los sistemas de ecuaciones. Para el primer renglón de la matriz $\mathbf{H}$ obtenemos:

$$\begin{aligned} H(1,2) &= 1 + H(1,2)m_{22} + H(1,3)m_{32} \\ H(1,3) &= 1 + H(1,2)m_{23} + H(1,3)m_{33} \end{aligned} \tag{2.50}$$

Y sustituyendo los valores de $\mathbf{M}$

$$\begin{aligned} H(1,2) &= 1 + \tfrac{1}{2}H(1,3) \\ H(1,3) &= 1 + H(1,2) \end{aligned} \tag{2.51}$$

Por lo tanto $H(1,2) = 3$, y $H(1,3) = 4$.

Análogamente, se plantea el sistema de ecuaciones para el segundo renglón de $H(j,i)$

$$\begin{aligned} H(2,1) &= 1 + H(2,1)m_{11} + H(2,3)m_{31} \\ H(2,3) &= 1 + H(2,1)m_{13} + H(2,3)m_{33} \end{aligned} \quad , \quad \begin{aligned} H(2,1) &= 1 \\ H(2,3) &= 1 \end{aligned} \tag{2.52}$$

y para el tercer rengón:

$$\begin{aligned} H(3,1) &= 1 + H(3,1)m_{11} + H(3,2)m_{21} \\ H(3,2) &= 1 + H(3,1)m_{12} + H(3,2)m_{22} \end{aligned} \quad , \quad \begin{aligned} H(3,1) &= 1 + H(3,2) \\ H(3,2) &= 1 + \tfrac{1}{2}H(3,1) \end{aligned} \quad , \tag{2.53}$$

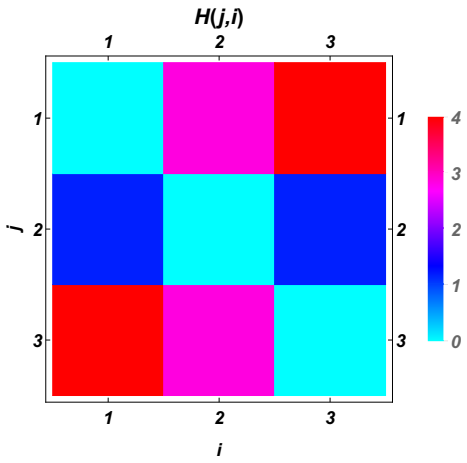


Figura 2.9: Matriz de primeras visitas **H** para 3 elementos en línea recta.

y en este caso: $H(3,1) = 4$, y $H(3,2) = 3$.

Así que la matriz **H** es

$$\mathbf{H} = \begin{pmatrix} 0 & 3 & 4 \\ 1 & 0 & 1 \\ 4 & 3 & 0 \end{pmatrix} \tag{2.54}$$

En este caso se resolvieron 3 sistemas de ecuaciones de 2 variables cada uno de forma recursiva, lo que nos da 6 elementos de la matriz, los tres restantes son los de la diagonal que en todos los casos por definición valen 0.

En este sistema $H(2,1) = H(2,3) = 1$, pues hay sólo un camino para ir de 1 a 2, o de 3 a 2. Por otro lado, con el método iterativo se obtuvo que para ir de un extremo del sistema al otro, se necesitan en promedio 4 pasos ($H(1,3) = H(3,1) = 4$), y que ir del estado central a un extremo, requiere en promedio 3 pasos ($H(1,2) = H(3,2) = 3$).

Una forma diferente de ilustrar la matriz **H** es usando un mapa de colores, como se muestra en la fig. (2.9). La escala de colores muestra al azul claro como el valor mínimo de $H(j,i)$ y se va modificando hacia tonalidades rojas conforme $H(j,i)$ crece. En este caso no hay tantos tonos, pues no hay muchos valores diferentes en las entradas de **H**, pero esta forma de visualizar la matriz de primeras visitas, es muy útil sobre todo para sistemas con muchos más estados. Podemos notar por ejemplo en los colores, que la matriz no es simétrica. Los extremos $H(3,1)$ y $H(1,3)$ se ven

en rojo, lo que indica que son los valores más grandes de la matriz, pues son el número de pasos que en promedio se necesitan para recorrer toda la red y en este caso, $H(1,3) = H(3,1)$.

Podemos verificar algunas de las entradas de **H** haciendo el cálculo directo del número de pasos promedio. Por ejemplo $H(3,1)$ es el número de pasos que en promedio son necesarios para ir del estado 1 al 3 por primera vez. Para hacer esto existen un número infinito de trayectorias, algunas de las cuales se muestran a continuación, junto con los pasos necesarios y la probabilidad de cada una de ellas:

Trayectoria	*Pasos*	*Probabilidad*
$1 \to 2 \to 3$	2	0,5
$1 \to 2 \to 1 \to 2 \to 3$	4	$(0{,}5)^2$
$1 \to 2 \to 1 \to 2 \to 1 \to 2 \to 3$	6	$(0{,}5)^3$
...		

Por lo tanto, $H(3,1)$ se obtiene sumando el número de pasos de cada trayectoria por su probabilidad:

$$H(3,1) = 2(0{,}5) + 4(0{,}5)^2 + 6(0{,}5)^3 + ... = 2\sum_{i=1}^{\infty} i \left(\frac{1}{2}\right)^i = 4. \tag{2.55}$$

Análogamente, se puede cacular $H(1,2)$. Mostramos los pasos y la probabilidad de algunas de las trayectorias

Trayectoria	*Pasos*	*Probabilidad*
$2 \to 1$	1	0,5
$2 \to 3 \to 2 \to 1$	3	$(0{,}5)^2$
$2 \to 3 \to 2 \to 3 \to 2 \to 1$	5	$(0{,}5)^3$
...		

Sumamos otra vez:

$$H(1,2) = (0{,}5) + 3(0{,}5)^2 + 5(0{,}5)^3 + ... = \sum_{i=1}^{\infty} (2i-1) \left(\frac{1}{2}\right)^i = 3. \tag{2.56}$$

Obtenemos que se necesitan 3 pasos en promedio para ir del estado 2 al 1 por primera vez, exactamente como se obtuvo resolviendo el sistema de ecuaciones recurrentes. Así ilustramos el significado de la matriz de primeras visitas **H**. El promedio de pasos que debe dar el sistema antes de llegar por primera vez al estado j, par-

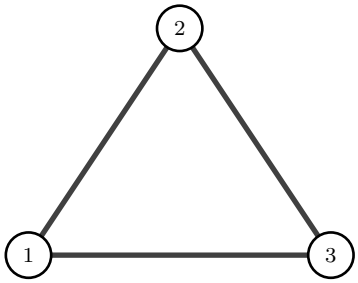


Figura 2.10: En este caso, los 3 estados están conectados entre sí de manera que cada uno de ellos puede evolucionar a los otros dos estados, con la misma probabilidad.

tiendo del estado i. Esta definición no tiene que ver con el camino más corto ni con otras propiedades que pueda tener el sistema. Simplemente se consideran todas las posibles trayectorias y se ponderan de acuerdo a la probabilidad asignada.

Si consideramos ahora un sistema en el que los estados se pueden representar de manera esquemática formando un triángulo equilátero como en la fig. (2.10). En este caso, el sistema es completamente simétrico, todos los estados están en contacto entre sí, por lo que debe tomar en promedio el mismo número de pasos ir de un nodo a cualquier otro. La matriz de transición $\mathbf{M}$ refleja la simetría del sistema, esto es:

$$\mathbf{M} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \tag{2.57}$$

Aplicando la ecuaciones de recurrencia (2.40) para obtener el primer renglón de $\mathbf{H}$, obtenemos nuevamente las ecs. (2.50), pero en este caso al sustituir los valores de m_{ij}:

$$\begin{aligned} H(1,2) &= 1 + \tfrac{1}{2}H(1,3) \\ H(1,3) &= 1 + \tfrac{1}{2}H(1,2) \end{aligned} \tag{2.58}$$

Por lo tanto $H(1,2) = 2$, y $H(1,3) = 2$.

Y al resolver los otros renglones obtenemos los mismos sistemas de ecuaciones, por lo tanto la matriz $\mathbf{H}$ de este sistema es

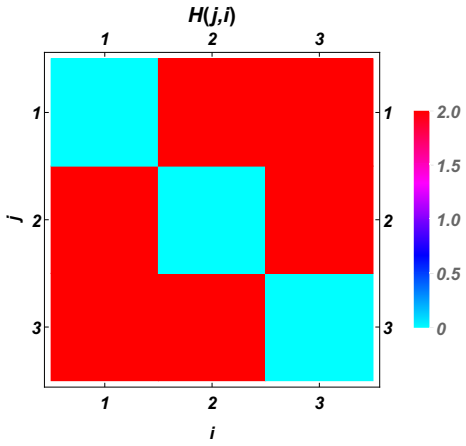


Figura 2.11: Matriz de primeras visitas **H** para 3 elementos unidos en triángulo.

$$\mathbf{H} = \begin{pmatrix} 0 & 2 & 2 \\ 2 & 0 & 2 \\ 2 & 2 & 0 \end{pmatrix} \tag{2.59}$$

Se muestra que para este caso, se requieren en promedio 2 pasos para ir por primera vez de un estado a cualquier otro. En la fig. (2.11) se muestra la matriz en colores y también es claro que todos los elementos tienen el mismo valor, salvo la diagonal que por definición es cero.

Si calculamos otra vez el número de pasos necesarios para ir por ejemplo, del estado 1 al 2, a través de las infinitas trayectorias que existen, obtenemos

Trayectoria	*Pasos*	*Probabilidad*
$1 \to 2$	1	0,5
$1 \to 3 \to 2$	2	$(0{,}5)^2$
$1 \to 3 \to 1 \to 2$	3	$(0{,}5)^3$
$1 \to 3 \to 1 \to 3 \to 2$	4	$(0{,}5)^4$
...		

Sumando el número de pasos de cada trayectoria por su probabilidad, obtenemos $H(2,1)$:

$$H(2,1) = (0{,}5) + 2(0{,}5)^2 + 3(0{,}5)^3 + ... = \sum_{i=1}^{\infty} i \left(\frac{1}{2}\right)^i = 2. \tag{2.60}$$

Figura 2.12: 4 estados en línea recta.

Que es efectivamente, el valor calculado al resolver el sistema de ecuaciones. Por la simetría del sistema el resultado es el mismo para cualquier otro par de estados.

2.2.3. Ejemplos Analíticos simples: 4 estados

Apliquemos ahora este método iterativo para 4 estados conectados en línea, como se muestra en la fig. (2.12).

La matriz de transición es ahora

$$\mathbf{M} = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & 1 \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \tag{2.61}$$

Para calcular el primer renglón de $\mathbf{H}$, el sistema de ecuaciones es:

$$\begin{aligned} H(1,2) &= 1 + H(1,2)m_{22} + H(1,3)m_{32} + H(1,4)m_{42} \\ H(1,3) &= 1 + H(1,2)m_{23} + H(1,3)m_{33} + H(1,4)m_{43} \\ H(1,4) &= 1 + H(1,2)m_{24} + H(1,3)m_{34} + H(1,4)m_{44} \end{aligned} \tag{2.62}$$

y sustituyendo los valores de m_{ij}:

$$\begin{aligned} H(1,2) &= 1 + \tfrac{1}{2}H(1,3) \\ H(1,3) &= 1 + \tfrac{1}{2}H(1,2) + \tfrac{1}{2}H(1,4) \\ H(1,4) &= 1 + H(1,3) \end{aligned} \tag{2.63}$$

al resolver el sistema se obtiene: $H(1,2) = 5$, $H(1,3) = 8$, y $H(1,4) = 9$.

De forma análoga calculamos los términos $H(2,i)$:

$$\begin{aligned} H(2,1) &= 1 + H(2,1)m_{11} + H(2,3)m_{31} + H(2,4)m_{41} \\ H(2,3) &= 1 + H(2,1)m_{13} + H(2,3)m_{33} + H(2,4)m_{43} \\ H(2,4) &= 1 + H(2,1)m_{14} + H(2,3)m_{34} + H(2,4)m_{44}, \end{aligned} \tag{2.64}$$

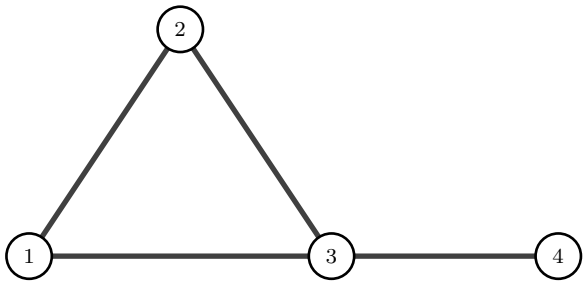


Figura 2.13: En este caso, los 3 primeros estados están conectados entre sí y el 4 aparece a continuación del estado 3.

Por lo tanto $H(2,1) = 1$, $H(2,3) = 3$, y $H(2,4) = 4$.

Planteando el conjunto de ecuaciones para los siguientes renglones, completamos la matriz $\mathbf{H}$, obteniendo para este sistema:

$$\mathbf{H} = \begin{pmatrix} 0 & 5 & 8 & 9 \\ 1 & 0 & 3 & 4 \\ 4 & 3 & 0 & 1 \\ 9 & 8 & 5 & 0 \end{pmatrix} \tag{2.65}$$

Por último, resolvemos el sistema de 4 estados representado en la figura (2.13). En este caso algunos términos de $\mathbf{H}$ pueden ser ya intuitivos. Por ejemplo, dado que estamos calculando primeras visitas, $H(3,1)$ y $H(3,2)$ deben valer 2, como en el caso de 3 sitios en triángulo, pues si se inicia en 1 o en 2, no es posible ir al sitio 4 sin pasar antes por el 3. De igual manera, es claro que $H(3,4) = 1$.

La matriz de transición es en este ejemplo

$$\mathbf{M} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 1 \\ 0 & 0 & \frac{1}{3} & 0 \end{pmatrix} \tag{2.66}$$

Para calcular cada renglón de $\mathbf{H}$, seguimos el mismo procedimiento que en el ejemplo anterior. Obtenemos en este caso

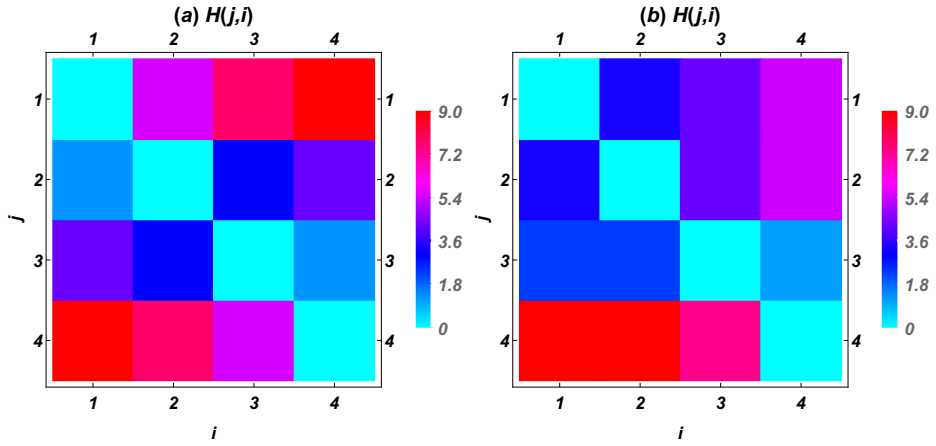


Figura 2.14: Matriz de primeras visitas **H** para 4 elementos. En (a) están todos en línea recta. En (b) hay 3 en triángulo y el cuarto colocado a un costado del vértice 3.

$$\mathbf{H} = \begin{pmatrix} 0 & \frac{8}{3} & \frac{10}{3} & \frac{13}{3} \\ \frac{8}{3} & 0 & \frac{10}{3} & \frac{13}{3} \\ 2 & 2 & 0 & 1 \\ 9 & 9 & 7 & 0 \end{pmatrix} \tag{2.67}$$

Estos ejemplos, dejan más claro el procedimiento algebraico para calcular el renglón j de **H**: es necesario resolver un sistema de $n-1$ ecuaciones, para las incógnitas $H(j,i)$, con $i = 1$ hasta n, excepto $i = j$.

En la fig. (2.14) se muestran las representaciones en colores de las dos matrices **H** calculadas en esta subsección. En (a) vemos el sistema formado por 4 estados en línea recta. Como en el caso de sólo 3 estados, ir de un extremo del sistema al otro, es el recorrido que en promedio requiere una mayor cantidad de pasos, en ambas direcciones. Los recorridos más cortos son $H(2,1)$ y $H(3,4)$ pues en ambos casos se necesita sólo un paso ya que el sistema tiene un único camino. Estos valores no son iguales a $H(1,2)$ y $H(4,3)$ respectivamente, pues existen más caminos para recorrer esos caminos, lo que hace que **H** no sea simétrica. En (b) la situación es muy diferente, ya que aquí si sucede que $H(2,1) = H(1,2)$ por la simetría del sistema. Vemos también que en este caso, llegar al estado 4 requiere en promedio más pasos que llegar a cualquier otro.

2.3. Matriz Fundamental para cadenas ergódicas y su relación con la matriz de primeras visitas

Veremos a continuación una matriz que juega el mismo papel en las cadenas ergódicas, que la matriz fundamental $\mathbf{N}$ definida para cadenas absorbentes. Esta matriz será utilizada para calcular la matriz de primeras visitas $\mathbf{H}$ para cadenas ergódicas, de una manera diferente a la que se desarrolló en la sección anterior.

Recordemos que la matriz fundamental $\mathbf{N} = (\mathbf{I}-\mathbf{Q})^{-1}$ para cadenas absorbentes, puede escribirse también como

$$\mathbf{N} = \mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + ...$$

Esta serie de potencias converge, pues $\mathbf{Q}^n \to 0$, así que pensando en esta propiedad, se propone en el caso de cadenas ergódicas, usar una serie convergente similar

$$\mathbf{I} + (\mathbf{M} - \mathbf{W}) + (\mathbf{M} - \mathbf{W})^2 + ...$$

Se puede demostrar que $(\mathbf{M} - \mathbf{W})^n = \mathbf{M}^n - \mathbf{W}$, y como habíamos visto, para cadenas ergódicas regulares $\mathbf{M}^n \to \mathbf{W}$ cuando n crece, por lo tanto esta serie también converge. Además la expresión $\mathbf{I} - (\mathbf{M} - \mathbf{W})$ tiene inversa, por lo que se define la *matriz fundamental* $\mathbf{Z}$ asociada a una cadena ergódica como

$$\mathbf{Z} = (\mathbf{I} - \mathbf{M} + \mathbf{W})^{-1} \tag{2.68}$$

La matriz $\mathbf{Z}$ existe incluso aunque la cadena no sea regular, ya que el término $\mathbf{I} - \mathbf{M} + \mathbf{W}$ tiene inversa.

Algunas propiedades útiles de $\mathbf{Z}$ son

$$\begin{aligned} \mathbf{Zc} &= \mathbf{c}, \\ \mathbf{wZ} &= \mathbf{w}, \\ \mathbf{Z}\,(\mathbf{I} - \mathbf{M}) &= \mathbf{I} - \mathbf{W} \end{aligned} \tag{2.69}$$

donde $\mathbf{c}$ es un vector columna de entradas igual a 1.

Por lo tanto, si la ecuación (2.45) es multiplicada por $\mathbf{Z}$

$$\mathbf{Z}(\mathbf{I}-\mathbf{M})\mathbf{H} = \mathbf{ZC} - \mathbf{ZD}$$

utilizando las propiedades anteriores

$$(\mathbf{I}-\mathbf{W})\mathbf{H} = \mathbf{C} - \mathbf{ZD}$$

$$\mathbf{H} = \mathbf{C} - \mathbf{ZD} + \mathbf{WH}$$

En término de las componentes

$$H(j,i) = 1 - z_{ij}r_j + (\mathbf{wH})_j \tag{2.70}$$

y para el caso $i = j$

$$0 = 1 - z_{jj}r_j + (\mathbf{wH})_j \tag{2.71}$$

por lo tanto, restando las ecuaciones anteriores para las componentes

$$H(j,i) = (z_{jj} - z_{ij})r_j \tag{2.72}$$

Y como habíamos visto $r_j = 1/w_j$, por lo tanto

$$H(j,i) = \frac{z_{jj} - z_{ij}}{w_j}. \tag{2.73}$$

Usando nuevamente el ejemplo del ratón en la caja, podemos calcular $\mathbf{I}-\mathbf{M}+\mathbf{W}$ utilizando la matriz de transición dada en la ec. (2.32) y la matriz $\mathbf{W}$ que se construye con el vector fijo, para obtener

$$\mathbf{I}-\mathbf{M}+\mathbf{W} = \begin{pmatrix} \frac{8}{7} & -\frac{2}{7} & \frac{1}{7} & -\frac{5}{14} & \frac{3}{14} & \frac{1}{7} \\ -\frac{4}{21} & \frac{17}{14} & -\frac{4}{21} & \frac{1}{7} & -\frac{5}{42} & \frac{1}{7} \\ \frac{1}{7} & -\frac{2}{7} & \frac{8}{7} & \frac{1}{7} & \frac{3}{14} & -\frac{5}{14} \\ -\frac{5}{14} & \frac{3}{14} & \frac{1}{7} & \frac{8}{7} & -\frac{2}{7} & \frac{1}{7} \\ \frac{1}{7} & -\frac{5}{42} & \frac{1}{7} & -\frac{4}{21} & \frac{17}{14} & -\frac{4}{21} \\ \frac{1}{7} & \frac{3}{14} & -\frac{5}{14} & \frac{1}{7} & -\frac{2}{7} & \frac{8}{7} \end{pmatrix} \tag{2.74}$$

Por lo tanto, la matriz fundamental $\mathbf{Z} = (\mathbf{I}-\mathbf{M}+\mathbf{W})^{-1}$ es

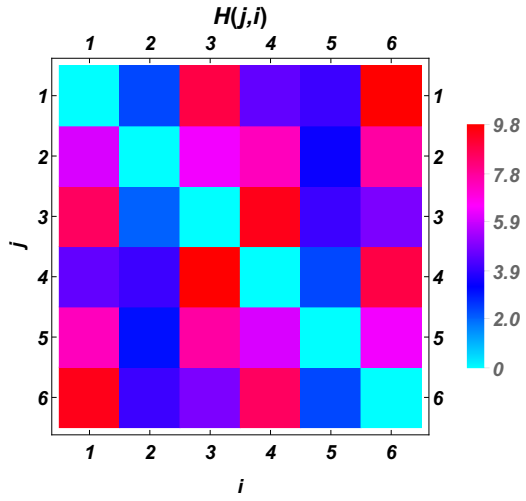


Figura 2.15: Representación en colores de la matriz **H** de primeras visitas para el problema del ratón en una caja con 6 espacios.

$$\mathbf{Z} = \begin{pmatrix} \frac{1619}{1470} & \frac{177}{980} & -\frac{341}{1470} & \frac{541}{1470} & -\frac{117}{980} & -\frac{439}{1470} \\ \frac{59}{490} & \frac{891}{980} & \frac{59}{490} & -\frac{39}{490} & \frac{9}{980} & -\frac{39}{490} \\ -\frac{341}{1470} & \frac{177}{980} & \frac{1619}{1470} & -\frac{439}{1470} & -\frac{117}{980} & \frac{541}{1470} \\ \frac{541}{1470} & -\frac{177}{980} & -\frac{439}{1470} & \frac{1619}{1470} & -\frac{177}{980} & -\frac{341}{1470} \\ -\frac{39}{490} & \frac{9}{980} & -\frac{39}{490} & \frac{59}{490} & \frac{891}{980} & \frac{59}{490} \\ -\frac{439}{1470} & -\frac{117}{980} & \frac{541}{1470} & -\frac{341}{1470} & \frac{177}{980} & \frac{1619}{1470} \end{pmatrix} \tag{2.75}$$

Y con la ec. (2.73) obtenemos las componentes de la matriz **H**, representada en colores en la fig.(2.15).

$$\mathbf{H} = \begin{pmatrix} 0 & 3{,}4 & 9{,}3333 & 5{,}1333 & 4{,}8 & 9{,}8 \\ 6{,}8667 & 0 & 6{,}8667 & 8{,}2667 & 4{,}2 & 8{,}2667 \\ 9{,}3333 & 3{,}4 & 0 & 9{,}8 & 4{,}8 & 5{,}1333 \\ 5{,}1333 & 4{,}8 & 9{,}8 & 0 & 3{,}4 & 9{,}3333 \\ 8{,}2667 & 4{,}2 & 8{,}2667 & 6{,}8667 & 0 & 6{,}8667 \\ 9{,}8 & 4{,}8 & 5{,}1333 & 9{,}3333 & 3{,}4 & 0 \end{pmatrix} \tag{2.76}$$

2.4. Aplicación de cadenas de Markov a proteínas

A continuación se aplicará este formalismo para representar la estructura de una proteína como una cadena de Markov, en la que cada aminoácido es un estado del sistema. Un proceso de caminante aleatorio sobre este sistema representará la transmisión de algún tipo de información entre los residuos que constituyen la red.

La propagación de información por medio de este caminante no es únicamente a través de la estructura primaria de la proteína, es decir entre residuos consecutivos unidos con enlaces covalentes; se considera más bien que la transmisión puede darse entre dos residuos si éstos están suficientemente cercanos físicamente. Definimos esta cercanía mediante la *Matriz de Afinidad*, $\mathbf{F}$, cuyos elementos son

$$f_{ij} = \frac{R_{ij}}{\sqrt{N_i N j}} \tag{2.77}$$

R_{ij} es el *número de contactos* átomo-átomo entre los residuos i y j, es decir, el resultado de contar para todos los átomos del residuo i, cuántos átomos del residuo j están de ellos, a una distancia menor que la distancia de corte R_c, exceptuando para ambos residuos, los átomos de hidrógeno.

$$R_{ij} = \sum_{k}^{N_i} \sum_{h}^{N_j} c_{kh}, \tag{2.78}$$

donde

$$c_{kh} = \begin{cases} 1, & r_{kh} \leq R_c \\ 0, & r_{kh} > R_c \end{cases} \tag{2.79}$$

r_{kh} es la distancia entre el k-esimo átomo del residuo i y el h-ésimo átomo del residuo j; N_i y N_j son el número de átomos que forman a los residuos i y j respectivamente, sin tomar en cuenta los átomos de H.

De esta manera, se está considerando la cercanía entre átomos como la única razón que define si existe interacción entre aminoácidos y si se debe establecer un enlace entre ellos. La cantidad de átomos de un aminoácido que cumplen con estar a una distancia menor que R_c de los átomos de otro aminoácido, cuantifican esta interacción. Dividir entre el número de átomos de cada residuo permite hacer un escalamiento para no sobreestimar el contacto entre residuos grandes, es decir,

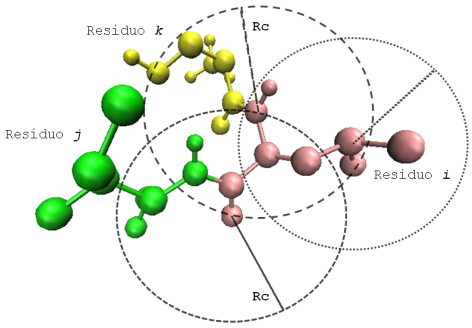


Figura 2.16: Esquema para mostrar cómo calcular los elementos de la matriz de afinidad f_{ij}. Se toma un átomo del residuo i y se cuenta cuántos átomos del residuo j están a una distancia menor a R_c. Esto se hace para todos los átomos de i, se suman todos los resultados y se divide entre $\sqrt{N_i N_j}$.

formados por muchos átomos.

La *Densidad de Interacción Local,* d_j, del residuo j es la suma de la afinidad de j con el resto de residuos de la red, es decir:

$$d_j = \sum_{i=1}^{n} f_{ij} \tag{2.80}$$

Si $\mathbf{F}$ es la matriz de afinidad y $\mathbf{D}$ la matriz diagonal que se forma con los elementos d_j, entonces, se puede definir la matriz $\mathbf{\Gamma}$ como

$$\mathbf{\Gamma} = \mathbf{D} - \mathbf{F} \tag{2.81}$$

En teoría de grafos $\mathbf{\Gamma}$ se conoce como *Matriz Laplaciana* o *Matriz de Kirchhoff.*

Para poner al sistema en el contexto de un modelo Markoviano, la matriz de transición en este caso definirá la probabilidad condicional de comunicación entre un par de residuos en función de la afinidad, es decir, los elementos m_{ij} representan la probabilidad condicional de que un caminante aleatorio llegue en un solo paso al residuo i, dado que ésta estaba inicialmente en el residuo j.

$$m_{ij} = \frac{f_{ij}}{d_j} \tag{2.82}$$

Note que efectivamente $\sum_{k=1}^{n} m_{kj} = 1$, esto es, la suma de los elementos *por columna* es igual a 1. La matriz de transición $\mathbf{M}$ es entonces $\mathbf{M} = \mathbf{F}\mathbf{D}^{-1}$.

Los términos de la *Matriz de primeras visitas,* $H(j,i)$, serán entonces el número de pasos que en promedio, le toma al caminante aleatorio que está en el residuo i (*residuo emisor*), llegar por primera vez al residuo j (*receptor*), tomando en cuenta todas las trayectorias posibles entre estos aminoácidos y la probabilidad de cada una de estas trayectorias.

Un procedimiento diferente para encontrar $\mathbf{H}$ utilizando las matrices que acabamos de definir, es el siguiente. Sea $\vec{H}$ el vector de $n-1$ términos que representa al j-ésimo renglón de la matriz $\mathbf{H}$, sin el término H_{jj}. Podemos entonces escribir la ec. (2.40) como

$$\vec{H} = \vec{1} + \vec{H}\hat{M} \tag{2.83}$$

Donde $\hat{M}$ es la matriz $\mathbf{M}$ truncada, sin el renglón ni la columna j. Observe que

$$\vec{H} - \vec{H}\hat{M} = \vec{H}\left(\hat{1} - \hat{M}\right) = \vec{1} \tag{2.84}$$

Podemos considerar matrices $\hat{F}$ y $\hat{D}$ truncadas, de manera que $\hat{M} = \hat{F}\hat{D}^{-1}$, por lo tanto

$$\vec{H}\left(\hat{D} - \hat{F}\right) = \vec{1}\hat{D} \tag{2.85}$$

Análogamente, podemos definir una matriz de Kirchhoff truncada $\hat{\Gamma} = \hat{D} - \hat{F}$, así que finalmente

$$\vec{H} = \vec{1}\hat{D}\hat{\Gamma}^{-1}. \tag{2.86}$$

Tenemos así, que otra manera de obtener cada renglón de la matriz $\mathbf{H}$, requiere sólo de multiplicar las matrices $\hat{D}$ y $\hat{\Gamma}^{-1}$ correspondientes.

Si regresamos al ejemplo de 3 estados acomodados de forma lineal, vemos que la matriz $\mathbf{\Gamma}$ es

$$\mathbf{\Gamma} = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \tag{2.87}$$

Por lo tanto, para resolver el primer renglón de la matriz truncada $\hat{H}$, hay que resolver el sistema de ecuaciones (2.86), así que haciendo las respectivas sustituciones:

$$(H(1,2), H(1,3)) = (1,1) \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}^{-1}, \tag{2.88}$$

invirtiendo Γ y haciendo las multiplicaciones correspondientes, obtenemos efectivamente $(H(1,2), H(1,3)) = (3,4)$.

Para resolver el segundo renglón de $\mathbf{H}$, eliminamos el segundo renglón y la segunda columna de $\mathbf{D}$ y de Γ. En este caso resulta

$$(H(2,1), H(2,3)) = (1,1) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{-1} = (1,1)\,, \tag{2.89}$$

por lo tanto, $(H(2,1), H(2,3)) = (1,1)$.

Resolviendo de igual forma el tercer renglón de $\mathbf{H}$ y agregando los términos $H(j,j)$, obtenemos el mismo resultado que con el sistema de ecuaciones recurrentes, expresado con la matriz (2.54).

en el siguiente capítulo se verán algunos conceptos propios de la Teoría de Redes, que también se usarán para modelar proteínas. De esta manera se podrán comparar ambos métodos entre sí y con resultados experimentales conocidos.

Capítulo 3

Teoria de Grafos y Redes

En la *Teoría de Grafos (Graph Theory)* o *Teoría de Redes (Network Theory)*, la representación de un sistema se hace de forma abstracta y sencilla, y no obstante, el análisis del mismo puede ser profundo y muy completo. Un sistema se representa mediante *nodos* o *vértices*, para indicar las partes o los diferentes estados que el sistema puede tener, y *enlaces (links)* o *aristas* para representar la relación o interacción entre estos vértices.

Esta representación simple, permite sin embargo utilizar la teoría de redes en muy diferentes áreas de investigación [41, 42], que van desde modelos de comportamiento social [43, 44], sistemas físicos [45], tecnológicos [46, 47], de comunicación [48], o como en nuestro caso, de conformación de estructuras biológicas [49–51].

En este capítulo describiremos algunos conceptos fundamentales de la teoría de redes, así como ciertas características que se pueden analizar en un sistema mediante este modelo, como medidas de conectividad o centralidad entre las componentes, las distancias y trayectorias que pueden establecerse, etc. En particular mostraremos cómo modelar la estructura terciaria de una proteína con esta metodología y cómo relacionar las características mencionadas de una red, con la ubicación de los sitios activos de la proteína.

3.1. Conceptos Básicos

Una *Red* o *Grafo* G está definida por un par de conjuntos, $G(\nu, \varepsilon)$, donde ν es el conjunto no vacío y contable de *vértices* o *nodos* $\{\nu_1, \nu_2, ..., \nu_N\}$, y ε el de *enlaces* representados por el conjunto $\{(\nu_i, \nu_j), (\nu_k, \nu_l), ...\}$, donde el par (ν_i, ν_j) indica una relación o interacción entre los vértices i y j, que se representa en el grafo con una línea que los une. El *orden* del grafo, N, es el número de vértices o nodos que la forman. Comúnmente este número se identifica con el tamaño de la red, aunque en algunos sistemas, se prefiere relacionar el tamaño con el número de enlaces L [52].

Un ejemplo estándar de un grafo G puede ser una red social, donde los nodos son cada una de las personas que participan, y los enlaces se establecen si existe una relación entre cada par de personas. Incluso sobre el mismo conjunto de nodos puede establecerse otro tipo de relación. Puede por ejemplo establecerse un enlace si son personas que viven a menos de una cierta distancia fija, aunque los individuos no se conozcan en persona [41, 46, 52]. Los vértices pueden ser ciudades y los enlaces las vías de comunicación existentes entre las comunidades [48]. O bien, podemos pensar que los vértices son las neuronas que forman un cerebro, y se establece un enlace si entre un par de neuronas hay flujo eléctrico o de sustancias bioquímicas [53], en fin. Las aplicaciones de la teoría de redes a sistemas físicos, matemáticos, sociales o biológicos, son extensísimas [54, 55].

Se dice que un par de vértices son *vecinos* o *adyacentes* si están unidos por un enlace. Se llama *Red* o *Grafo simple*, cuando sólo puede existir a lo más un enlace entre cualquier par de vértices (es decir, no hay multienlaces), y cuando en la red no existen enlaces entre un nodo consigo mismo (bucle).

La figura (3.1) muestra algunos ejemplos sencillos de grafos. 3.1**(a)** representa un grafo *nulo*, es decir uno en el que no existen enlaces. En cambio 3.1**(b)** es un grafo *completo*, pues todos los vértices están unidos entre sí. Si un grafo completo está formado por N vértices, entonces el número total de enlaces es:

$$L = \begin{pmatrix} N \\ 2 \end{pmatrix} = \frac{N(N-1)}{2}$$

Se define la *densidad,* D de una red, como el número de enlaces que la forman,

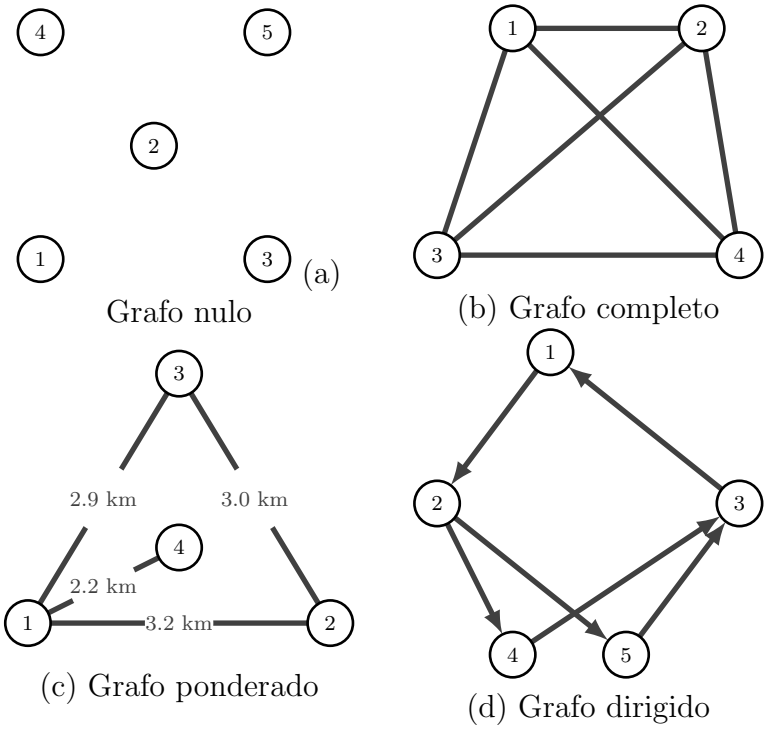


Figura 3.1: Diferentes tipos de grafos en función de su conectividad.

entre el número máximo posible de enlaces es decir,

$$D = \frac{2L}{N(N-1)}$$

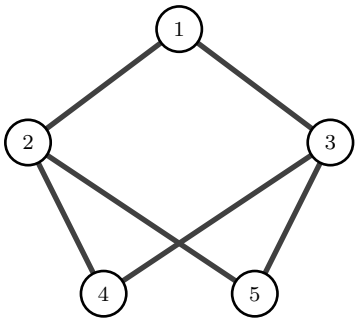
Si el grafo incluye algún tipo de información para los enlaces, como en 3.1**(c)**, se dice que es *etiquetado* o *ponderado*. Cuando los enlaces están direccionados con flechas significa que hay una relación sólo en la dirección marcada por la flecha y no en sentido opuesto. En este caso se habla de grafos *dirigidos* (fig. 3.1**(d)**).

Existe una relación biunívoca entre un grafo y la *matriz de adyacencia* $N \times N$, cuyos términos A_{ij} indican la relación entre los vértices i y j, y se definen

$$A_{ij} = \begin{cases} 1, & \text{si i, j son vértices conectados} \\ 0, & \text{si i, j son vértices no conectados} \end{cases} \tag{3.1}$$

En el caso de grafos dirigidos, $A_{ij} = 1$ indica que hay una interacción que va de i a j, pero no necesariamente $A_{ij} = A_{ji}$. Si el grafo es no dirigido, se satisface

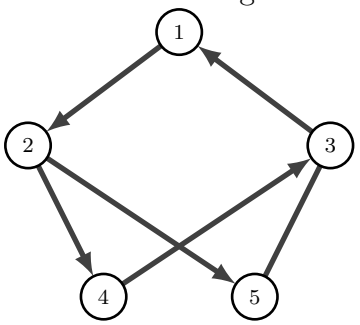
la igualdad y la matriz de adyacencia será siempre simétrica, como se ilustra en la figura (3.2).



Grafo no dirigido

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix}$$

Matriz de adyacencia



Grafo dirigido

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Matriz de adyacencia

Figura 3.2: Diferencias entre la matriz de adyacencia de dos grafos cuando uno es no dirigido y el otro dirigido. En el primer caso, la matriz es simétrica.

Para grafos *no dirigidos*, se define el *grado* k_i del vértice i, como el número de enlaces que inciden en él. Si $\mathbf{A}$ es la matriz de adyacencia de una red no dirigida, es simétrica, y el grado del vértice k_i se obtiene sumando todos los elementos del i-ésimo renglón

$$k_i = \sum_j A_{ij} \tag{3.2}$$

El número total de enlaces que contiene una red L, se obtiene sumando el grado de todos los nodos de la misma y dividiendo entre 2, pues como cada enlace tiene dos bordes es contado dos veces.

$$L = \frac{1}{2}\sum_i^N k_i = \frac{1}{2}\sum_{i,j} A_{ij} \tag{3.3}$$

El *grado promedio de una red no dirigida,* $\langle k \rangle$, se define como el valor promedio de k sobre todos los nodos de la red

$$\langle k \rangle = \frac{1}{N} \sum_i k_i = \frac{2L}{N} \tag{3.4}$$

Es decir, el grado promedio de la red se obtiene de dividir el doble del total de enlaces entre el número de nodos, pues cada enlace participa dos veces al definir el grado de los vértices.

Por otra parte, el *coeficiente de agrupación* C_i (o *clustering coefficient*), es una medida de la probabilidad de que dos vértices que están unidos a ν_i, sean adyacentes entre sí. Si el nodo i tiene k_i enlaces, los nodos a los que está unido pueden a su vez tener entre ellos un máximo de $\frac{k_i(k_i-1)}{2}$ enlaces. Por lo tanto el coeficiente de agrupación para el vértice i, se define como la razón entre los enlaces que en verdad existen L_i entre los vecinos del nodo i, y la cantidad total que podrían formarse

$$C_i = \frac{2L_i}{k_i(k_i - 1)}. \tag{3.5}$$

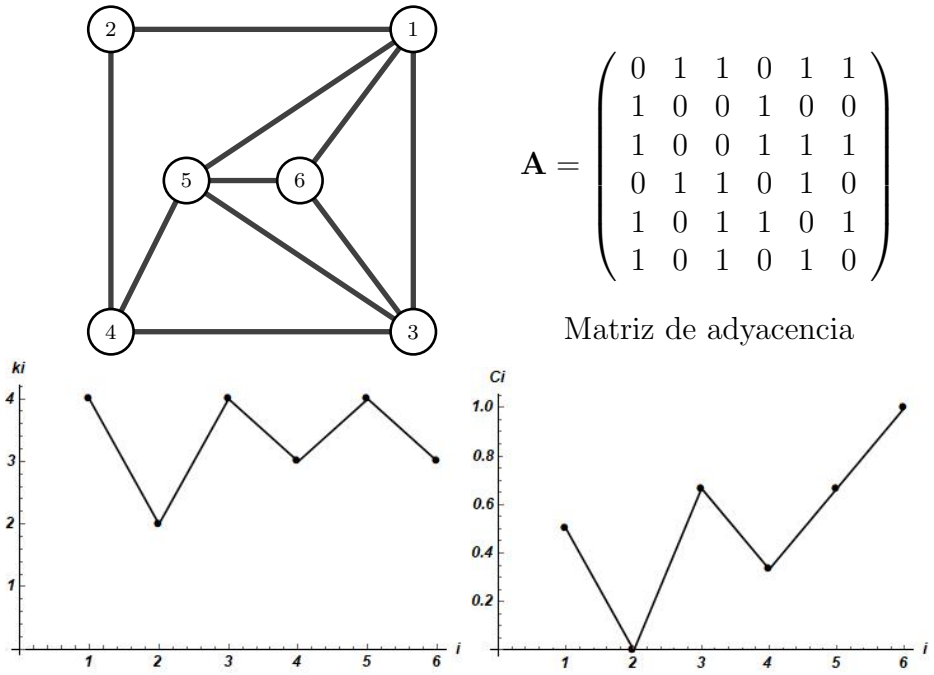
En la fig. (3.3) se muestra como ejemplo el grado y el coeficiente de agrupación para una red de 6 nodos.

El coeficiente de agrupación de la red, C, es el promedio de los coeficientes de agrupación de todos los nodos.

Otra matriz de mucha utilidad en el análisis estructural de redes, es la *matriz Laplaciana* o *matriz de Kirchoff* mencionada en el capítulo anterior. La matriz Laplaciana $\mathbf{L}$ de una red se define [56]

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \tag{3.6}$$

donde $\mathbf{D}$ es la *matriz de Grado*, una matriz diagonal cuyos términos d_{ii} son el grado del vértice i. Por ejemplo, para la red mostrada en la figura(3.3), la matriz de Grado



$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Matriz de adyacencia

Figura 3.3: Se muestra una red de 6 nodos, la matriz de adyacencia, la gráfica del grado de cada vértice, dado por la ec. (3.2) y el coeficiente de agrupación, ec. (3.5).

es

$$\mathbf{D} = \begin{pmatrix} 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

por lo tanto la matriz Laplaciana

$$\mathbf{L} = \begin{pmatrix} 4 & -1 & -1 & 0 & -1 & -1 \\ -1 & 2 & 0 & -1 & 0 & 0 \\ -1 & 0 & 4 & -1 & -1 & -1 \\ 0 & -1 & -1 & 3 & -1 & 0 \\ -1 & 0 & -1 & -1 & 4 & -1 \\ -1 & 0 & -1 & 0 & -1 & 3 \end{pmatrix}$$

Esta matriz recibe este nombre, porque si se establece un proceso de difusión de un campo escalar ψ, dentro de la estructura de una red, es posible establecer una ecuación de difusión semejante a la ecuación de difusión de gases, en donde la matriz $\mathbf{L}$ juega el mismo papel que el operador laplaciano [41].

Otra característica importante de las redes es que la numeración de los nodos es en general arbitraria, por lo que las propiedades que son de interés en teoría de redes, son aquellas relacionadas con la *invarianza* de las matrices $\mathbf{A}$ y $\mathbf{L}$ bajo permutaciones de renglones y columnas.

Puesto que las entradas de $\mathbf{A}$ y $\mathbf{L}$ son números reales, y además son matrices simétricas para redes no dirigidas, un resultado conocido del álgebra lineal es que sus eigenvalores deben ser números reales, y en el caso de la matriz de adyacencia, dado que todos los elementos de la diagonal son cero, la suma de los eigenvalores debe ser cero. Para la matriz laplaciana en cambio, todos los eigenvalores θ_i son $\theta_i \leq 0$. (vea por ejemplo [57]). Los eigenvalores y eigenvectores de las matrices asociadas a una red, proporcionan información sobre la estructura de dicha red [49, 51, 56]. Al análisis de estas cantidades se le llama *Análisis Espectral* de la red.

En particular, el polinomio característico de la matriz $\mathbf{A}$ de la red G es $p_G(\lambda) = det(\lambda I - A)$, donde λ representa los eigenvalores de $\mathbf{A}$, y por tanto las raices del polinomio. Si $\mathbf{A}$ tiene s eigenvalores distintos, tales que $\lambda_1 > \lambda_2 > ... > \lambda_s$ con multiplicidad $m(\lambda_1), ..., m(\lambda_)$, el *espectro de la red* es la matriz $2 \times s$, donde el primer renglón está formado por los eigenvalores dispuestos en orden decreciente, y el segundo con la multiplicidad de estos eigenvalores

$$Esp(G) = \begin{pmatrix} \lambda_1 & \lambda_2 & ... & \lambda_s \\ m(\lambda_1) & m(\lambda_2) & ... & m(\lambda_s) \end{pmatrix} \tag{3.7}$$

Por ejemplo, en la figura (3.3) se muestra una red formada por 6 vértices, se indica la matriz de adyacencia de la cual se obtiene el polinomio característico

$$p(\lambda) = \lambda^6 - 10\lambda^4 - 10\lambda^3 + 6\lambda^2 + 6\lambda - 1$$

de cuyas raíces se obtienen los eigenvalores de la matriz de adyacencia, por lo que para este ejemplo, el espectro de la red es

$$Espect(G) \begin{pmatrix} 3{,}49785 & 0{,}72992 & 0{,}15054 & -1 & -1{,}18763 & -2{,}19069 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Se puede demostrar que en general, para una red G de N vértices y L enlaces, el polinomio característico de la matriz de adyacencia $\mathbf{A}$

$$p_G(\lambda) = \lambda^N + a_1\lambda^{N-1} + ... + a_{N-1}\lambda + a_N$$

satisface siempre que $a_1 = 0$ y $a_2 = -L$ [58].

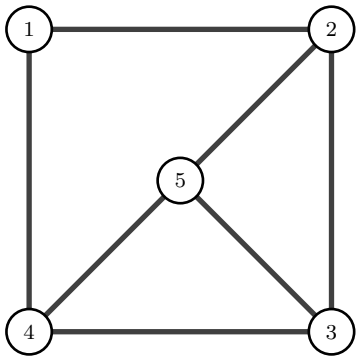
3.1.1. Trayectorias y Conectividad

Uno de los temas fundamentales en el estudio de redes, es determinar la accesibilidad de los nodos, es decir, qué tan factible es ir de uno a otro siguiendo los enlaces que conectan la red. Este concepto es importante porque puede utilizarse para estudiar variables dinámicas en redes, como flujos de información o de energía [50], y procesos de difusión [59].

Para analizar la conectividad de una red, se define una *trayectoria* o *camino* (*path* en inglés) P_{i_0,i_n} en el grafo $G(\nu, \varepsilon)$, como el conjunto ordenado de $n+1$ vértices $\nu_P = \{\nu_0, \nu_1, ..., \nu_n\}$ y el conjunto de n enlaces $\varepsilon_P = \{(\nu_0, \nu_1), (\nu_1, \nu_2), ..., (\nu_{n-1}, \nu_n)\}$, tales que $\nu_\alpha \in \nu$ y $(\nu_{\alpha-1}, \nu_\alpha) \in \varepsilon$ para todo α. Dicho en palabras, el camino del vértice ν_0 al ν_n es la secuencia ordenada de vértices tales que, comenzando en ν_0, siempre sea posible encontrar dos consecutivos conectados por un enlace, hasta llegar a ν_n.

Un grafo es *conectado* o *ergódico* si existe una trayectoria que conecte a cualquier par de vértices del grafo. Un *ciclo* es una trayectoria cerrada, es decir ($\nu_0 = \nu_n$), y todos los demás vértices y enlaces son diferentes.

Estrictamente, al recorrer una trayectoria uno podría regresar a un vértice en cualquier momento las veces que sea. Entonces, la *longitud de un camino,* t_i entre dos vértices, es el número de enlaces por el número de veces que fueron cruzadas para ir de uno al otro. En cambio, la *distancia* d_{ij} entre los vértices i y j, se define como el número de aristas que hay que cruzar por *el camino más corto* que conecte a



$$t_1 = \{(\nu_1, \nu_4), (\nu_4, \nu_3), (\nu_3, \nu_2)\} = 3$$

$$t_2 = \{(\nu_1, \nu_4), (\nu_4, \nu_3), (\nu_3, \nu_5), (\nu_5, \nu_2)\} = 4$$

$$d_{12} = 1$$

Figura 3.4: Entre los vértices 1 y 2 pueden definirse muchos caminos. En la figura se muestran t_1 y t_2, ambos de diferente longitud, pero la distancia entre estos vértices es $d_{12} = 1$.

los nodos i y j, es decir, por el camino que implique pasar por la menor cantidad de aristas, como se ilustra en la figura (3.4). En un grafo no dirigido $d_{ij} = d_{ji}$, mientras que en una dirigida, lo anterior no siempre es cierto. El *diámetro D de una red*, es la distancia más grande que puede establecerse entre todos los pares de nodos de dicha red. La *distancia media, d*, es el valor medio de las distancias entre todos los pares de nodos, y como con N nodos se pueden formar $N(N-1)$ parejas

$$d = \frac{1}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j \neq i}^{N} d_{ij} \tag{3.8}$$

Es posible obtener el número total de caminos de longitud l que hay en una red. Por ejemplo, para saber si existe un camino de longitud 2 entre los vértices i y j, que pasan por el vértice k, habrá que hacer el producto $A_{ik}A_{kj}$. Si el resultado es 1, es que sí están conectados i con k y k con j. Por lo tanto el número total de trayectorias de longitud 2 entre i y j, definido como $N_{ij}^{(2)}$, es

$$N_{ij}^{(2)} = \sum_{k}^{N} A_{ik}A_{kj} = [\mathbf{A}^2]_{ij}$$

Podemos generalizar este resultado a trayectorias de cualquier longitud r, esto es,

$$N_{ij}^{(r)} = [\mathbf{A}^r]_{ij} \tag{3.9}$$

Por ejemplo, para la red mostrada en la figura (3.4), se tiene la matriz de adyacencia

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

por lo tanto

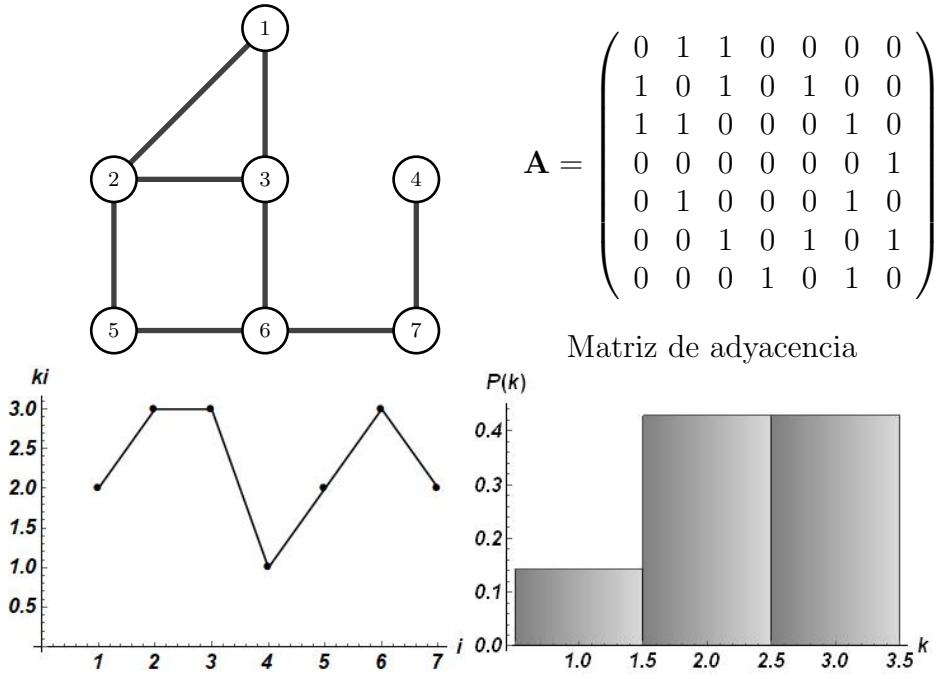
$$\mathbf{A}^2 = \begin{pmatrix} 2 & 0 & 2 & 0 & 2 \\ 0 & 3 & 1 & 3 & 1 \\ 2 & 1 & 3 & 1 & 2 \\ 0 & 3 & 1 & 3 & 1 \\ 2 & 1 & 2 & 1 & 3 \end{pmatrix}$$

lo que significa que por ejemplo, $N_{11}^{(2)} = 2$, hay entonces dos caminos de longitud 2 para ir y regresar al vértice 1. Efectivamente, podemos ir de 1 a 2 y regresar por el mismo enlace, o de 1 a 4 y regresar igual. En cambio, $N_{24}^{(2)} = 3$, indica 3 caminos de longitud 2 entre los nodos 2 y 4, que son $(2,1),(1,4)$, o $(2,3),(3,4)$, y por último $(2,5),(5,4)$.

3.2. Medidas de Centralidad

En muchas aplicaciones es útil determinar y cuantificar qué tan importante es cada vértice o cada enlace dentro de la estructura de la red, no precisamente en función de las propiedades intrínsecas del nodo o del enlace, sino por su ubicación, por la cantidad de conexiones o la forma de la red. En teoría de redes, las *Medidas de Centralidad* son justamente las variables que se usan para determinar esta importancia. Existen muchas medidas de centralidad que cuantifican diferentes características de la red [41, 52]. Mencionaremos algunas.

- *Centralidad de grado (Degree centrality)*: La forma más básica de medir la centralidad de un vértice es por su *grado*, definido en la ecuación (3.2), es decir,



$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Matriz de adyacencia

Figura 3.5: Red formada por 7 nodos. Se muestra la matriz de adyacencia, la centralidad de grado y la distribución de grado. En este caso, $P(k) = 0$ para $k \geq 4$.

por su número de conexiones con otros vértices. Mientras más conexiones tenga un nodo, puede ser más eficiente al compartir o propagar información, puede también ser parte de una mayor cantidad de trayectorias a lo largo de la red. En grafos dirigidos es conveniente hacer un refinamiento en esta definición, pues habrá que separar los casos en los que el enlace se dirige al vértice en cuestión, de aquellos en los que parte de dicho vértice. Llamamos *grado de entrada* $k_{in,i}$ en una grafo dirigido, al número de enlaces que se dirigen al vértice i. Análogamente, el *grado de salida* $k_{out,i}$, es el número de enlaces que salen de dicho nodo.

$$\begin{aligned} k_{in,i} &= \sum_{j=1}^{N} A_{ji}, \\ k_{out,i} &= \sum_{j=1}^{N} A_{ij}. \end{aligned} \tag{3.10}$$

Y el grado total del vértice en el caso de grafos dirigidos es $k_i = k_{in,i} + k_{out,i}$.

La *distribución de grado, P(k)*, de un grafo no dirigido, es la probabilidad de que al elegir al azar un vértice, resulte ser de grado k. En la fig. (3.5) se muestra una red de 7 nodos, la matriz de adyacencia correspondiente, la centralidad de grado y la gráfica de $P(k)$, es decir, la *distribución de grado* de la red.

La distribución de grado es una cantidad muy utilizada en la descripción de las propiedades y de la estructura o topología de las redes [41, 42]. Por ejemplo, muchas redes aleatorias, es decir, aquellas en donde los enlaces son asignados al azar, tienen una distribución de grado de Poisson. Uno de los resultados más importantes en el estudio de redes complejas, fue el descubrimiento de que para muchas redes de sistemas reales grandes, como internet, la distribución de grado sigue una ley de potencias [59].

- *Centralidad de cercanía (Closeness centrality)*: Se define como el inverso del promedio de las distancia de un vértice a todos los demás.

$$g_i = \frac{N}{\sum_{i \neq j} d_{ij}}. \tag{3.11}$$

Es de suponer que mientras menor sea la distancia de un vértice al resto, puede participar más y por lo tanto, ser más importante en las interacciones

que ocurran en la red, por lo que con este criterio los vértices más importantes deben ser los que tengan el menor valor promedio de distancia a los demás. Por eso se prefiere definir g_i como el inverso de la distancia, de esta manera g_i es más grande para los nodos que están a menor distancia del resto. En la fig. (3.6) ilustramos la centralidad de cercanía de la red de la fig. (3.5). En esta red, los vértices 2, 3 y 6 son de grado 3, sin embargo la suma de las distancias de todos los nodos al 6, es menor que la del resto, por lo tanto este vértice es el de máxima cercanía.

- *Centralidad de intermediación (Betweenness centrality)*: En este caso se cuentan las distancias en las que participa cada vértice, es decir esta medida cuantifica el que un nodo haga el papel de puente o intermediario entre otros vértices de la red. Sea σ_{hj} el número de trayectorias más cortas o de distancias que pueden encontrarse entre los vértices h y j, y $\sigma_{hj}(i)$ el número de estas trayectorias que pasan por el vértice i [52]. Entonces, la intermediación de i se define:

$$b_i = \sum_{h \neq j \neq i} \frac{\sigma_{hj}(i)}{\sigma_{hj}}. \tag{3.12}$$

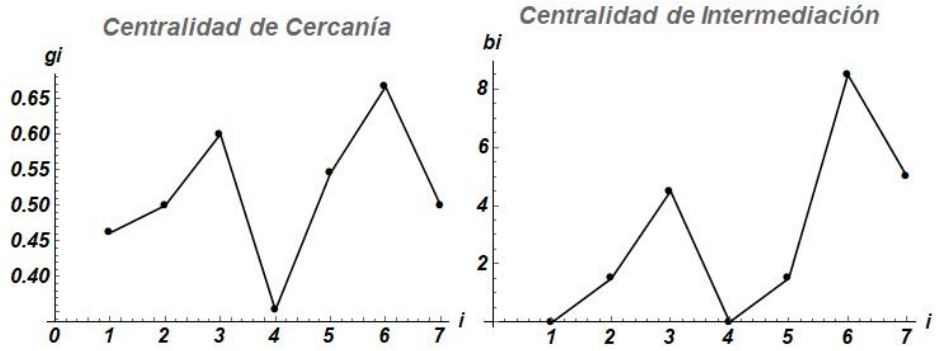


Figura 3.6: Centralidad de Cercanía y de Intermediación para la red de 7 nodos de la fig. (3.5). En el caso de la centralidad de cercanía, vemos que aunque los vértices 2, 3 y 6 tienen el mismo grado (3), el vértice 6 tiene una mayor centralidad de cercanía. De igual forma, la Centralidad de Intermediación de el vértice 6 es la más grande de la red, lo que indica que es por el que pasa una mayor cantidad de caminos más cortos.

En la fig. (3.6) se muestra también la centralidad de intermediación de la red de la fig. (3.5). Vemos que el vértice 6 es el que tiene una mayor b_i, es decir, por él pasan la mayor cantidad de caminos más cortos.

- *Centralidad de vector propio (Eigenvector centrality)*: Mide la influencia de un vértice en una red considerando no sólo la cantidad de enlaces que tiene, sino también la importancia de los vértices con los que está enlazado [41]. Esta importancia se mide por la cantidad de enlaces que tienen dichos vértices vecinos, por lo tanto los vértices con una centralidad de vector propio alta, están conectados a vértices con alta centralidad de grado, por lo que son buenos candidatos para transmitir de manera más eficiente la información a través de la red.

  Sea x_i la centralidad de vector propio. Este valor se obtiene sumando los elementos del renglón i de la matriz de adyacencia, pero multiplicados por su respectivo x_j, es decir

$$x_i = \frac{1}{\lambda}\sum_j A_{ij}x_j$$

  donde λ es una constante. Definiendo el vector de centralidad $\mathbf{x} = (x_1, x_2, ...)$, podemos escribir la ecuación anterior en forma matricial

$$\lambda\mathbf{x} = \mathbf{A}\mathbf{x} \tag{3.13}$$

  Así que $\mathbf{x}$ es el vector propio de $\mathbf{A}$ cuyo valor propio es λ, y para que $\mathbf{x}$ no tenga componentes negativas, se puede demostrar [41] que se debe elegir el valor propio más grande de la matriz $\mathbf{A}$.

  Por ejemplo, para la red mostrada en la figura (3.3), obtuvimos el polinomio característico y el espectro de la red. El más grande de los eigenvectores es $\lambda_1 = 3.49785$, al que le corresponde el vector

$$\mathbf{x_1} = (0.4554, 0.2273, 0.4804, 0.3397, 0.4804, 0.4049)$$

Podemos ver en la figura que los vértices 1, 3 y 5 son todos de grado 4, sin embargo la centralidad de vector propio es mayor para los vértices 3 y 5 que para el 1. Esto es así porque los nodos adyacentes a 3 y 5 tienen en conjunto un mayor grado que los nodos adyacentes a 1 por lo que disminuye esta medida de centralidad para este último vértice.

En el siguiente capítulo usaremos algunos de estos conceptos para interpretar las redes construidas a partir de la estructura terciaria de proteínas. Cada aminoácido representará un vértice, y colocaremos una arista en función a la matriz de afinidad, cuyos elementos fueron definidos por la ecuación (3.1). Calcularemos algunas medidas de centralidad para los vértices de estas redes y analizaremos la relación entre los valores de centralidad y la existencia de sitios activos en la estructura de la proteína.

Capítulo 4

Matriz de Primeras Visitas y Medidas de Centralidad Calculadas para Macromoléculas Biológicas

Como se expuso en el capítulo 2, se puede representar la estructura tridimensional de una macromolécula como una red y calcular la matriz de afinidad **F** asociada, utilizando las coordenadas de los átomos que la conforman. En este capítulo se obtendrá **F**, y se usará para calcular la matriz de transición **M**, que a su vez nos permitirá calcular la matriz de primeras visitas **H** de todos los residuos de la macromolécula. Para el análisis, es conveniente definir el número de pasos promedio por cada vértice como receptor y como emisor:

$$\begin{aligned} \langle H^r(j)\rangle &= \textstyle\sum_i H(j,i)/n, \\ \langle H^e(i)\rangle &= \textstyle\sum_j H(j,i)/n, \end{aligned} \tag{4.1}$$

Donde n es el número total de residuos de la proteína y corresponde al número de vértices de la red, así que $\langle H^r(j)\rangle$ es el promedio de los elementos del renglón j. De esta manera obtenemos el número promedio de pasos que necesita un caminante aleatorio para llegar por primera vez a j, desde cualquier residuo del que provenga. En cambio $\langle H^e(i)\rangle$ es un promedio sobre la columna i y nos da el número de pasos

que recorre un caminante aleatorio que sale del residuo i, hasta llegar por primera vez a los demás vértices. El caminante aleatorio representa un proceso de transferencia de información a lo largo de la red de residuos.

En la referencia [38] se calculó la matriz de primeras visitas **H** y los promedios obtenidos con las ecs. (4.1). Se encuentra que los residuos que constituyen los sitios activos coinciden con los mínimos de $\langle H^r(j)\rangle$, es decir, los sitios activos son los lugares a donde la información puede llegar en promedio en la menor cantidad de pasos o de tiempo a través de la red, lo que permite suponer que los sitios activos deben tener la propiedad de recibir de manera eficaz la información que se genera en el resto de la macromolécula. En este trabajo se hará un análisis similar al de la referencia mencionada, para un conjunto diferente de proteínas, y además utilizando la misma red se calculará la matriz de adyacencia **A** a partir de **F** mediante la siguiente relación:

$$A_{ij} = \begin{cases} 0, & \text{si } F_{ij} = 0 \\ 1, & \text{si } F_{ij} \neq 0 \end{cases} \tag{4.2}$$

de esta manera será posible calcular las diferentes medidas de centralidad definidas en el capítulo 3, y analizar la relación entre estas medidas y los sitios activos de las proteínas.

Este análisis se hará sobre un conjunto amplio de proteínas de diferentes características bioquímicas. La elección de éstas fue con base en la disponibilidad de información experimental de los sitios activos de las mismas. Primero se verificaron los códigos desarrollados para resolver matemáticamente las matrices y ecuaciones involucradas, comparando los resultados de la matriz **H** con los reportados en la referencia [38]. Una vez hecho esto, se procede a ampliar el estudio incluyendo las medidas de centralidad y el conjunto nuevo de proteínas.

4.1. Comparación de resultados de la matriz H y cálculo de medidas de centralidad

De acuerdo a [38], se calcula $H(j,i)$, y $\langle H^r(j)\rangle$ para 3 enzimas analizadas en esta referencia, Fosfolipasa A2 de veneno de serpiente, cuya ID en el Protein Data Bank

[29] es *1BK9*, HIV-1 proteasa (*1A30*) y Proteasa 3C de rinovirus humano (*1CQQ*).

Se muestran en la figura (4.1) la estructura de dichas proteínas, de acuerdo al archivo PDB obtenido del *Protein Data Bank*, obtenidas con el programa *Visual Molecular Dynamics (VMD)*. Junto a ellas se indica la red que se obtiene con la metodología mencionada, utilizando *Mathematica* 11.0. Para ilustrar estas redes, los vértices se colocaron en las coordenadas del C_α de cada residuo. En ambas representaciones se han destacado los residuos que forman los sitios activos de las proteínas.

La matriz de afinidad se obtiene a partir de la ecuación (2.77). Se tiene un código escrito en Fortran que lee las coordenadas atómicas dadas en los archivos PDB y calcula la distancia entre todos los pares de átomos de residuos diferentes, excluyendo los átomos de hidrógeno. El radio de corte utilizado en todos los casos es de 4.0 Å y en función de éste, el programa decide si la distancia entre un par de átomos de residuos distintos i y j, se considera para el correspondiente término f_{ij} de la matriz de afinidad o no. Una vez calculada $\mathbf{F}$, ésta se utiliza para obtener las matrices $\mathbf{D}$ y $\mathbf{M}$. Esta información se exporta al programa *Mathematica* en su versión 11.0 donde se usa para resolver los sistemas de ecuaciones recursivas dados por las ecuaciones (2.42), y obtener la matriz de primeras visitas $\mathbf{H}$, así como los promedios por renglón y columna para obtener $\langle H^r(j)\rangle$ y $\langle H^e(i)\rangle$.

En las figuras (4.2), se ilustra la matriz $\mathbf{H}$ para las proteínas mencionadas. La primer figura corresponde a la reportada en [38] para la proteína $1BK9$ y a su lado la calculada en este trabajo. Si bien el código de colores no es el mismo, la información es idéntica. En todas las matrices reportadas se forman bandas horizontales de color uniforme, lo que indica que $H(j,i)$ es prácticamente constante para una j dada. Esto quiere decir que el número de pasos para que un caminante aleatorio llegue por primera vez al residuo j es casi el mismo, independientemente del residuo i del que proviene; a menos que sean residuos contiguos, como lo indica la línea diagonal. Vemos en todos los casos una línea diagonal en azul claro, que es el color que indica los valores mínimos de $H(j,i)$, la cual era de esperarse pues por definición $H(j,j)=0$, y además si se toman dos residuos consecutivos en la cadena, se requieren en promedio muy pocos pasos para pasar por primera vez de uno al otro.

De acuerdo a la escala de color, se concluye que hay residuos o grupos de residuos vecinos, a los cuales la información del caminante puede llegar por primera vez en un

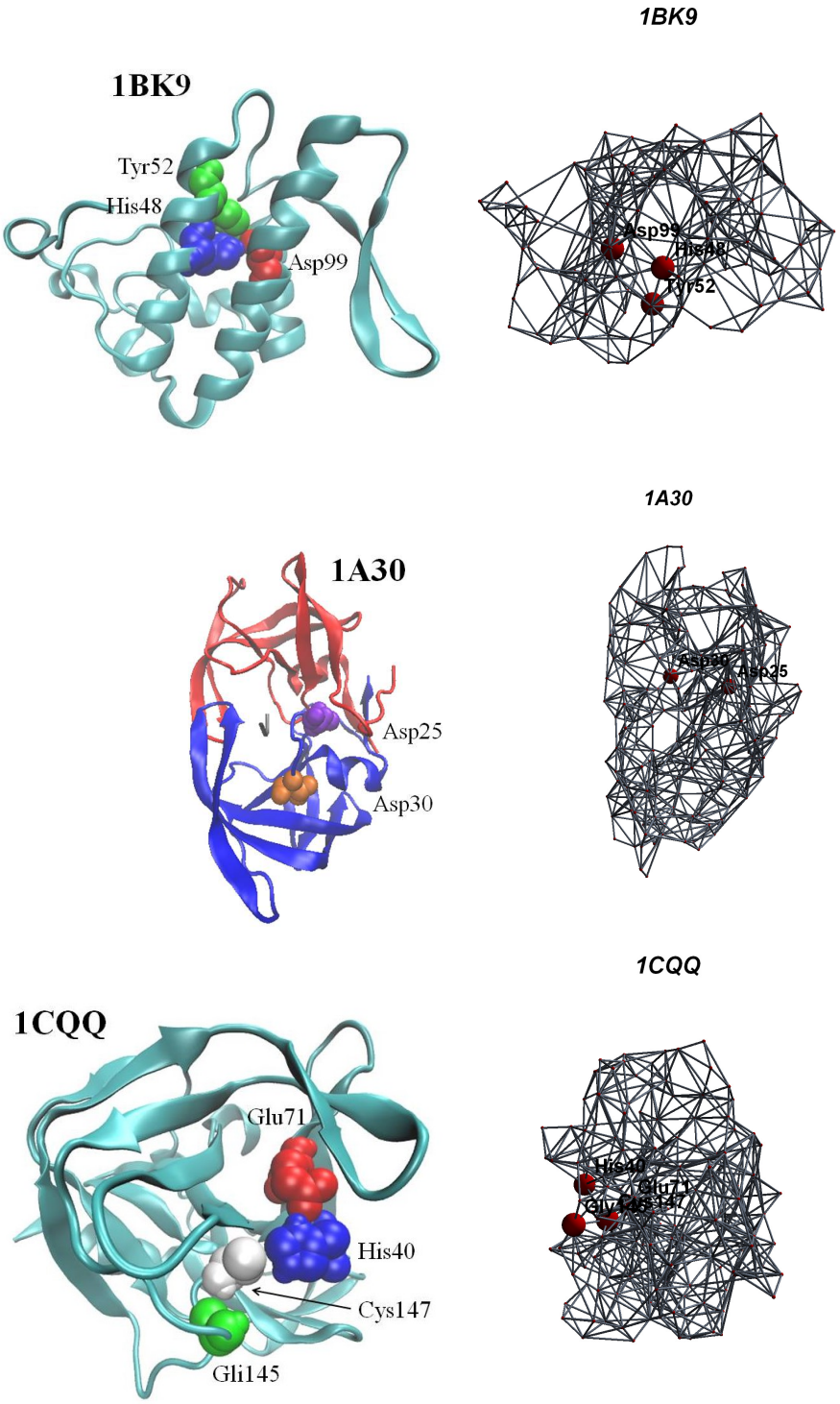


Figura 4.1: Estructura de las proteínas $1BK9$, $1A30$ y $1CQQ$, de acuerdo al archivo PDB, obtenidas con el programa *Visual Molecular Dynamics (VMD)* junto con la estructura de redes, calculada con el programa *Mathematica 11*. Se muestran los residuos señalados como sitios activos en los dos esquemas. En $1A30$, los colores sirven para diferenciar las dos cadenas que forman la proteína.

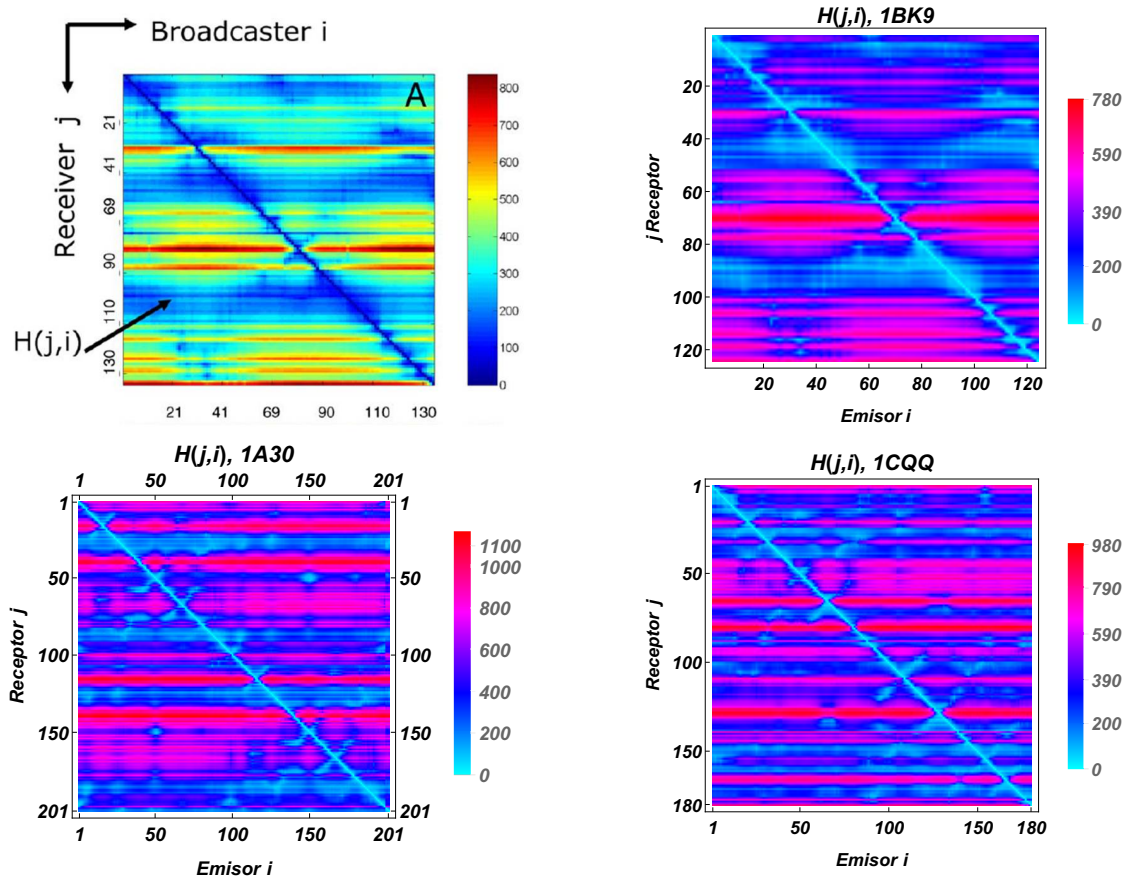


Figura 4.2: Matriz de primeras visitas, $H(j,i)$ para las proteínas 1BK9, 1A30 y 1CQQ. La primer figura es la del artículo [38] correspondiente a $1BK9$. Tal como se muestra en la escala de color, las zonas rojas representan sitios que requieren en promedio una mayor cantidad de pasos en la red que representa a la enzima para recibir información de un caminante aleatorio por primera vez. Las líneas en azul claro en cambio, muestran las zonas con los valores mínimos de $H(j,i)$. Las gráficas se elaboraron con *Mathematica* 11.0.

número pequeño de pasos, comparado con los demás. Por ejemplo, para la proteína *1BK9*, se notan regiones con valores pequeños de **H**, alrededor de los residuos que ocupan los lugares 48 y 90. Lo mismo ocurre en las figuras de las otras proteínas. Hay regiones horizontales en las cuales predomina el color que indica una pequeña cantidad de pasos para acceder a estos residuos.

En la figura (4.3) se muestra $\langle H^r(j)\rangle$, que como se ha dicho, representa el promedio sobre cada renglón de **H**, es decir, el número de pasos que en promedio requiere cada residuo para recibir por primera vez información proveniente de los demás aminoácidos de la red. Los puntos azules en la gráfica de [38] y los puntos rojos de las figuras elaboradas en este trabajo, marcan los residuos del sitio activo reportado. En primer lugar, se ve que las dos primeras figuras que corresponden a *1BK9*, tienen la misma estructura; la primera es la del artículo de referencia y la segunda la que resulta de este trabajo. Por otra parte, experimentalmente se sabe que 9 residuos conforman los sitios activos de las tres proteínas mostradas, y con esta metodología encontramos que 5 de ellos, coinciden con mínimos locales de $\langle H^r(j)\rangle$. De hecho en $1BK9$ y en $1A30$, uno de los residuos reportados es el mínimo global. Se ha agregado en las gráficas una región en azul que abarca el 15 % de los valores más bajos de $\langle H^r(j)\rangle$ y como podemos observar, 8 de los 9 residuos en las tres proteínas están dentro de esta región.

Verticalmente, la barra azul de las gráficas comienza en el valor mínimo de $\langle H^r(j)\rangle$, llamado m, y termina en h definida como

$$h = m + 0{,}15(M - m)$$

donde $M =$ *Valor máximo de* $\langle H^r(j)\rangle$.

Es importante aclarar que las gráficas elaboradas en este trabajo, muestran una numeración de residuos consecutiva, a diferencia de algunos archivos PDB, en los que no siempre es así, por ejemplo para la enzima *1BK9* no aparecen 10 residuos, el 15, 57, 58, 60, 62, 63, 64, 65 66 y 87, así que aunque el archivo indica que *1BK9* está formada por 134 aminoácidos, sólo reporta coordenadas para los átomos de 124 de ellos, por lo que en los cálculos con el método aquí resuelto, las matrices de esta proteína son de 124×124. Esto hace que el residuo marcado como $Asp99$ no corresponda con el punto 99 del eje horizontal, sino con el 89. En el caso de la

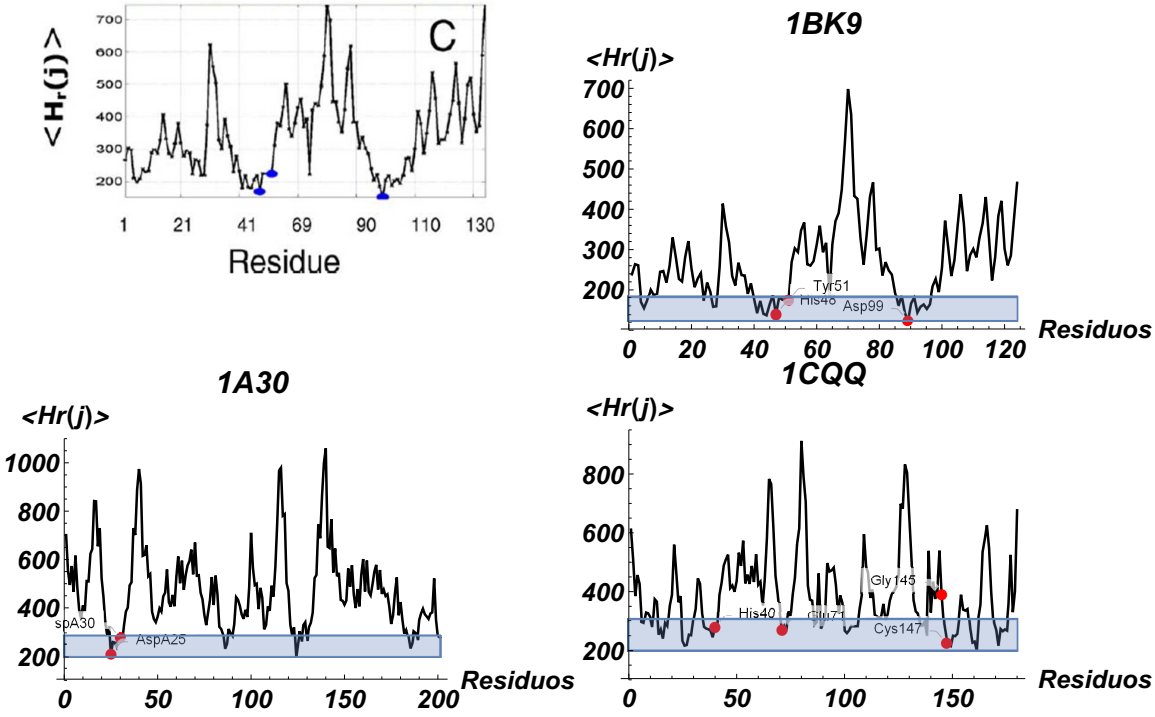


Figura 4.3: $\langle H^r(j)\rangle$ para el mismo conjunto de proteínas. La primer figura es la reportada en la referencia [38] para $1BK9$, y a su lado, la que se obtuvo en este trabajo. Los puntos azules en la primer gráfica y los rojos en las restantes, indican los residuos que forman el sitio activo conocido en cada caso. En todas las gráficas, estos puntos están cerca de mínimos locales, o ellos mismos son un mínimo local, lo cual es buen indicativo de que efectivamente la matriz de primeras visitas **H**, está relacionada con los sitios activos de las proteínas.

enzima *1A30*, en el archivo PDB los residuos están numerados del 1 al 99 para la cadena A, luego la numeración comienza de nuevo para la cadena B con algunos residuos faltantes, y comienza nuevamente en una pequeña cadena C. En cambio, en las figuras mostradas, los residuos siguen una numeración consecutiva.

En el cálculo de $\langle H^e(i)\rangle$, es decir, en el promedio por columnas, o en el promedio de pasos que un caminante aleatorio da para que un residuo *transmita* información por primera vez al resto de la red, se obtiene aproximadamente un valor constante para todos los residuos de cada proteína, lo cual indica que en este modelo no hay residuos en la red que sean mejores emisores que otros.

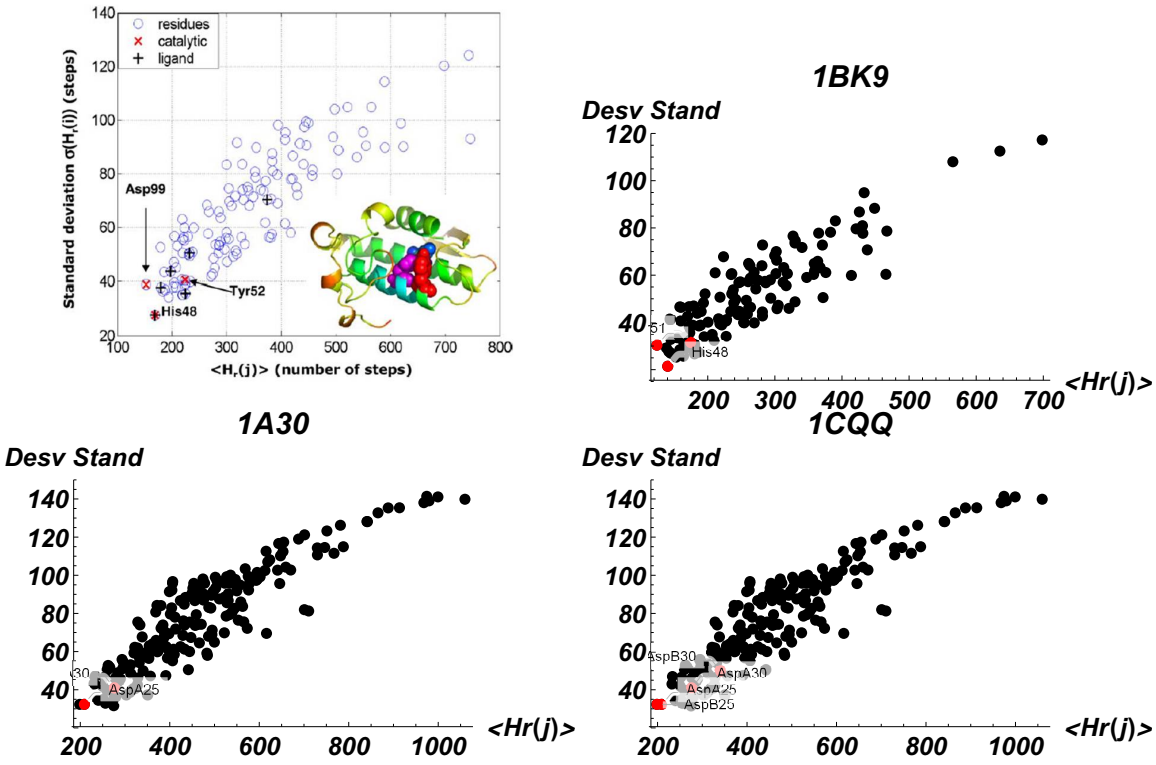


Figura 4.4: Análisis de $\langle H^r(j)\rangle$ vs $\sigma(\langle H^r(j)\rangle)$. Los residuos que forman un sitio activo, tienen en general una desviación estándar pequeña, lo que indica que además de recibir pronto la información, ésta es poco dispersa.

En la figura (4.4) se muestra la desviación estándar σ de los promedios $\langle H^r(j)\rangle$. De nueva cuenta, las figuras correspondientes a $1BK9$ reportada en la referencia [38] y la calculada, son muy similares. Se observa que los residuos que forman los sitios activos, indicados con puntos rojos, están en el conjunto de valores con desviación estándar pequeña, lo cual indica que el caminante o la información que porta, no solo requiere pocos pasos para llegar por primera vez a estos residuos, sino que además este resultado es muy uniforme, es decir, no ocurre que haya residuos con valores muy alejados del promedio. Este resultado está relacionado con la descripción cualitativa de las bandas horizontales de color uniforme mencionada para las gráficas de **H**.

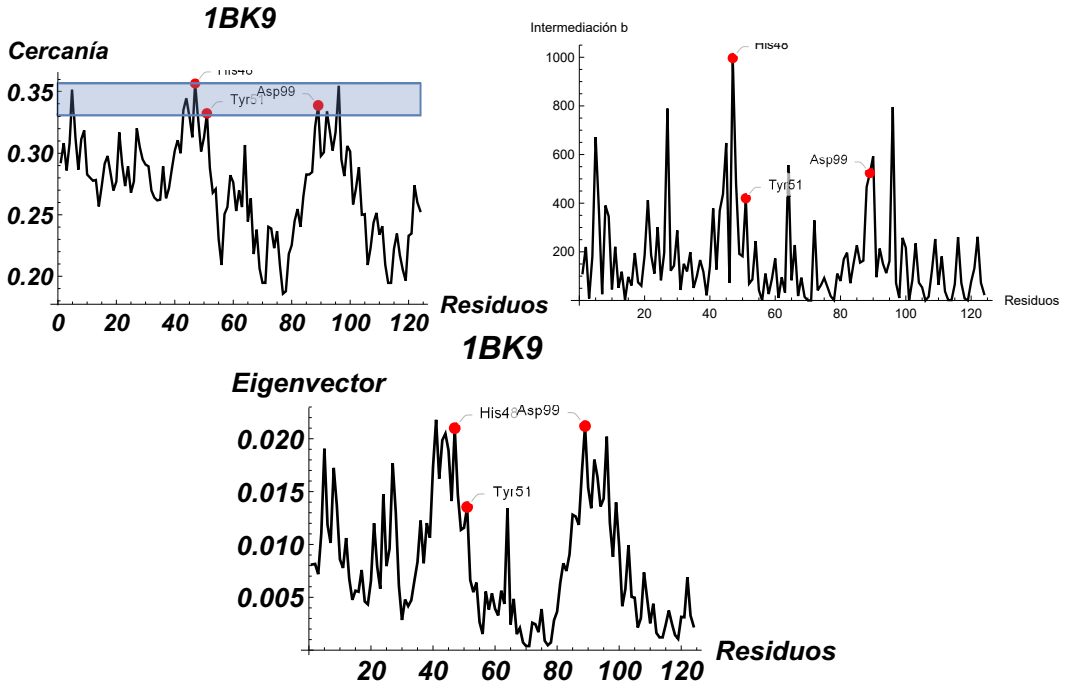


Figura 4.5: Medidas de Centralidad para la proteína $1BK9$. Se muestra la centralidad por *cercanía*, de *intermediación* y de *eigenvector*.

4.1.1. Resultados para las medidas de centralidad

La comparación entre los resultados calculados en este trabajo con las de la referencia mencionada, permite mucha certeza respecto a los cálculos de las matrices de afinidad y de adyacencia de las proteínas. Hecha esta revisión, se muestra ahora un análisis sobre el mismo conjunto de proteínas en función de las medidas de centralidad.

La figura (4.5) muestra las medidas de centralidad por cercanía (closeness), de intermediación (betweenness) y de eigenvector de la proteína fosfolipasa A2 ($1BK9$). Las fosfolipasas A2 constituyen un grupo de enzimas que juegan un papel importante en una variedad de procesos celulares, incluyendo la digestión y metabolismo de fosfolípidos, así como la producción de precursores para reacciones inflamatorias [60]. Los valores más altos en la centralidad por cercanía indican los residuos cuya distancia promedio a los demás, es mínima. Los residuos que forman el sitio activo, ocupan picos que se encuentran en la región sombreada, que representa el 15 % de los valores más altos. Incluso $His48$ es el máximo global, o el residuo que está a

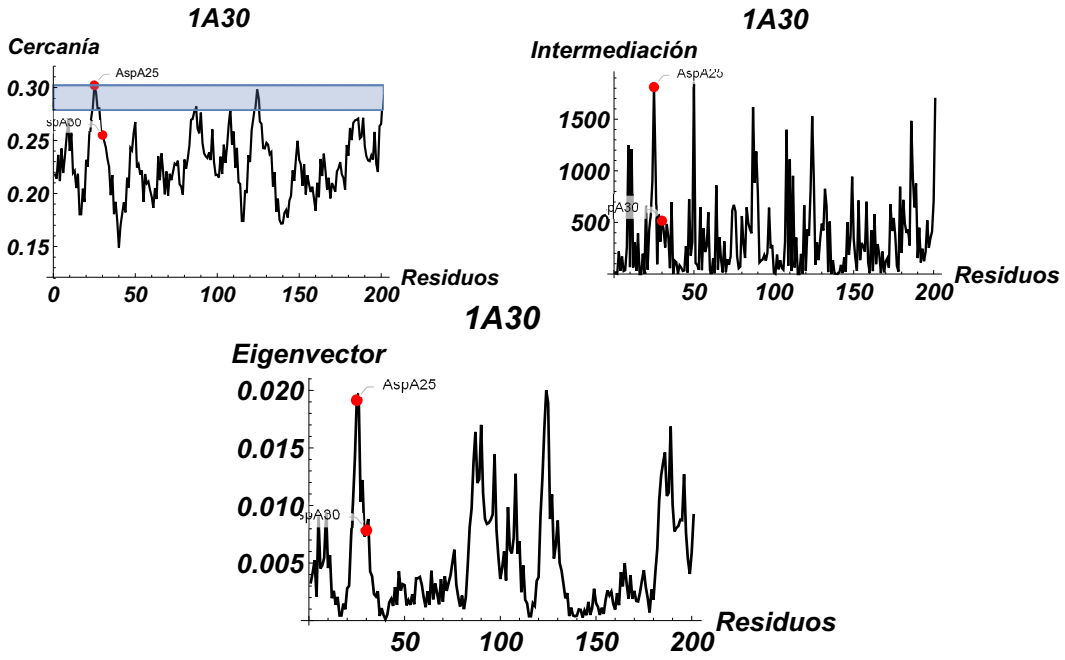


Figura 4.6: Medidas de Centralidad para la proteína 1A30.

menor distancia de todos los demás. La centralidad de intermediación del residuo i, indica la cantidad de trayectorias entre dos residuos h, j en las que i participa. En este caso otra vez $His48$ es el máximo global, lo cual significa que es el residuo por el que pasan la mayor cantidad de caminos que comunican al resto. La centralidad de eigenvector es una centralidad de grado más elaborada, pues toma en cuenta no sólo la cantidad de vecinos de un residuo dado, sino también el grado de conectividad de dichos vecinos. En este caso $His48$ y $Asp99$ tienen casi el mismo valor de centralidad y están entre los 3 más altos. El residuo $Tyr52$ no parece responder a ninguna de las tres medidas, lo cual no significa que no forme parte del sitio activo, sino que no es relevante para estas medidas de centralidad. Por otra parte, en las tres gráficas e incluso en la fig. (4.3), se ven picos destacados en residuos que no están reportados experimentalmente como parte de sitios activos. Con esta metodología no se puede asegurar que lo sean, sin embargo, esta formulación puede representar un punto de arranque para explorar con otras técnicas, la posibilidad de que esas regiones tengan alguna participación activa en la función de la proteína.

La fig. (4.6) muestra las mismas medidas de centralidad de la figura anterior (cercanía, intermediación y eigenvector), pero ahora para la proteasa $HIV-1$ (1A30).

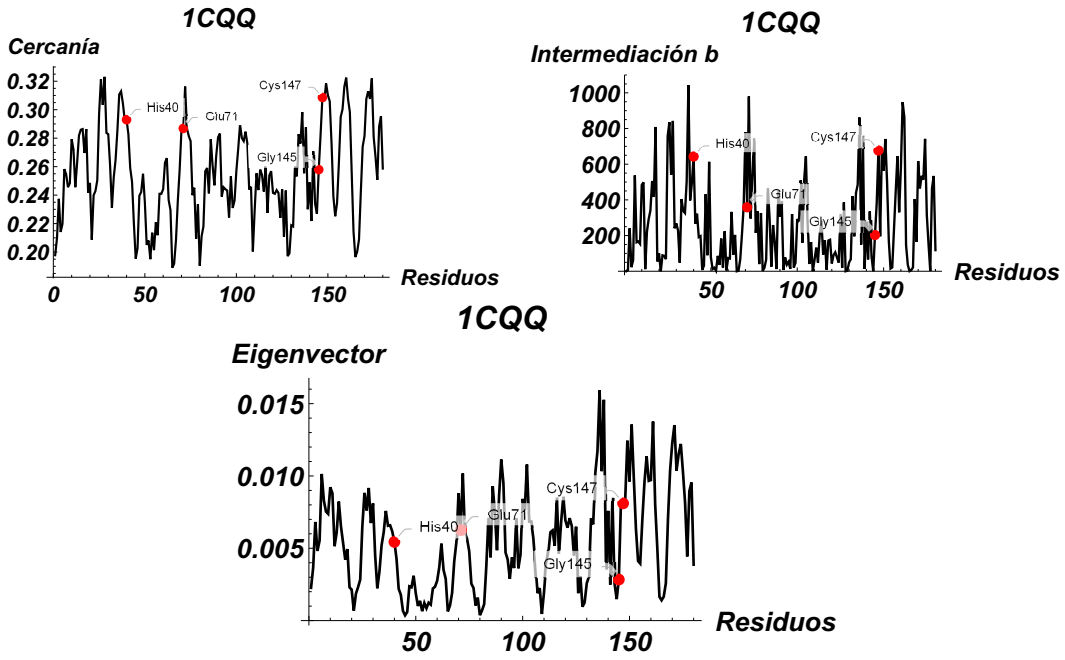


Figura 4.7: Medidas de Centralidad para la proteína $1CQQ$.

Esta enzima pertenece al virus de inmunodeficiencia humana (VIH). Se trata de una enzima de la familia de las aspartil proteasas también conocidas como ácido proteasas. Esta proteína está formada por dos cadenas semejantes y en la gráfica (4.6) relativa a la centralidad de cercanía, se observa que el residuo $Asp25$ de la cadena A ($AspA25$) es el pico más alto. El segundo más alto corresponde al aminoácido $AspB25$, pero sólo el de la cadena A está reportado en [38] como sitio activo. Este residuo muestra también un alto valor de centralidad de intermediación y de eigenvector. En cambio $AspA30$ no tiene valores máximos locales en ninguna de las tres medidas de centralidad, aunque en la fig. (4.3) sí se encuentra en la región de valores mínimos para $\langle H^r(j)\rangle$.

Por otra parte, los rhinovirus son los patógenos más comunes en humanos; son los agentes causantes del resfriado común. La proteasa 3C ($1CQQ$) del rhinovirus participa en el proceso de infección [61]. Esta proteína se muestra en la fig. (4.7), se tienen 4 residuos en el sitio activo reportado. Los aminoácidos señalados se encuentran cerca de puntos máximos en el caso de la centralidad de cercanía, pero para las otras dos medidas de centralidad, no es evidente la relación de estos residuos con valores extremos.

4.2. Cálculo de H y medidas de centralidad para proteínas diferentes

La metodología anterior fue aplicada a 20 proteínas diferentes a las reportadas en [38], las cuales cubren un espectro amplio de enzimas biológicas. Se incluye en la tabla (4.1) un resumen con las proteínas analizadas, de los residuos que constituyen los sitios activos reportados en diferentes fuentes y de aquellos que fueron bien ubicados por los métodos revisados. No se incluye en la tabla los resultados de la centralidad de eigenvector pues se constató que esta medida ubicó menos residuos de sitios activos dentro de sus valores extremos. Es notorio en esta tabla que la medida que ubica una mayor cantidad de residuos es $\langle H^r(j)\rangle$, y entre las medidas de centralidad, la de cercanía es la que logra ubicar una mayor cantidad.

PROT	SITIOS ACTIVOS	$\langle H^r(j)\rangle$	C_c	C_b
1YAZ 153	**CSA:** H63, R143 **Uniprot:** E42, H46, H48, H63, H71, H80, D83, H120, R143 **PDBSite:** H46, H48, H63, H120	46, 48, 63, 71 80, 83, 120, 143	46, 48	46, 48
1JUK 248	**CSA:** E51, K53, K110, E159, N180, S211	51, 110, 159, 180, 211	51, 110	51, 110
2CI7 284	**Uniprot:** L29, D72, E77, D78, R97, R144, H172, V267, C273 **PDBSite:** L29, D72, F75, R97, R144, H172, C273	72, 77, 78, 97, 144, 172, 273	77, 78, 172	77, 78, 144, 172
1KIR 352	**CSA:** EC35, DC52 **Uniprot:** EC35, DC52, DC101	B31, B32, B52, B53, B99,	B52, B99, B100, B101,	B52, B100,

PROT	SITIOS ACTIVOS	$\langle H^r(j)\rangle$	C_c	C_b
	PDBSite: GB31, YB32, WB52, GB53, DB54, RB99, D100, YB101, RB102, GC22, YC23, SC24, NC27 GC102, NC103, KC116, GC117, TC118, DC119, VC120, QC121	B100, B101, B102, C22, C23, C24, C27, C116, C117, C118, C119, C120, C121	B102, C22, C23, C121	C22, C121
1PRH 1108	**CSA:** Q203, H207, Y385 **Uniprot:** H207, Y385, H388, S530 **PDBSite:** R120, Q203, H207, Y385, H388, E524, S530	203, 207, 385 524, 530		
2C14 337	**CSA:** K205, K260 **Uniprot:** K205, R215, K229, E245, K260, S286, Y324 **SABER:** D131, K229	205, 215, 229 245, 260, 286, 324	286, 324	205
1IMA 277	**CSA:** EA70, TA95 **Uniprot:** E70, D90, I92, D93, E213, D220 **PDBSite:** E70, D90, P91, I92, D93, G94, T95, D220 en cadenas A y B	70, 90, 91, 92, 93, 94, 95, 213, 220 en cadenas A y B		
1AHA 246	**CSA:** E160, R163 **Uniprot:** R163 **PDBSite:** Y70, E160, R163	160, 163	160, 163	
1BIB 321	**CSA:** R118, K183, R317 **Uniprot:** Q112, K183 **PDBSite:** S89, T90, N91, Q112, G115, R116, G117, R118, Y132, K183, I187, L188, G204, A205	89, 90, 91, 112, 132, 183, 187, 188, 204, 205	132, 204, 205	132

PROT	SITIOS ACTIVOS	$\langle H^r(j)\rangle$	C_c	C_b
1CC1 283	**CSA:** TS19, EL23 **Uniprot:** EL51, CL70, CL73, IL444, UL492, CL495, HL498 **PDBSite:** CL70, CL73, CL495	S19, L23 L51, L70, L73, L493, L495, L499	S19, L23 L70, L73, L492, L495	
1C50 830	**CSA:** H377, H680 **Uniprot:** Y75, C108, C142, Y155 **PDBSite:** R60, V64, K191	60, 64, 75, 108, 142, 155, 191, 377, 680	143, 377	
1DZE 248	**Uniprot:** D85 **PDBSite:** W86, T89, T90, M118, W138, S141, T142, W182, Y185, P186, L216	85, 86, 89, 90 118, 138, 141, 142, 182, 185, 182, 186, 216	85, 86, 89, 90, 185, 216	182, 185
1DZY 215	**CSA:** E73, A117 **Uniprot:** E73, H92, H94, H155 **PDBSite:** E73, H92, H94, H155	73, 92, 94, 155	73, 92	73, 92, 155
1ELC 240	**CSA:** H60, D108, G201, S203, G204 **Uniprot:** H60, E74, Q79, E84, D108, S203 **PDBSite:** H60, D108, S203	60, 108, 201, 203, 204	203, 204	
1NEL 436	**CSA:** E168, E211, K345, K396 **Uniprot:** H159, E168, E211, D246, E295, D320, K345, K396 **PDBSite:** A38, H159, E168, E211, D246, E295, D320, K345, H373, S375, K396	168, 211, 246, 295, 320, 345, 373, 375, 396	211, 295, 345, 373, 396	295, 320, 345, 373, 396
1X0L 333	**CSA:** YA125, KA171, DA204, YB125, KB171, DB204	A98, A118, A228, A232	A98, A171, A204, B204	

PROT	SITIOS ACTIVOS	$\langle H^r(j)\rangle$	C_c	C_b
	Uniprot: RA85, RA88, RA98, RA118, YA125, KA171, DA204, DA228, DA232 **SABER:** KA171, DA204	B171, B204		
2BEO 236	**PDBSite:** QA61, YA62, KA64, AA66, FA67, YA126, QA146, LA150, QB6a, YB62, QB123, YB126	A61, A62, A64, A66, A67, A126, A146, A150, B61, B62, B123, B126	A61, A126, B61, B123, B126	A126, B126
1E5M 416	**CSA:** C167, H307, E318, K339, H344, F403 **PDBSite:** C167	167, 307, 318, 339, 344	167, 307, 339, 344	307, 339, 344
2CIX 299	**CSA:** C29, H105, D106, E183 **Uniprot:** C29, E104, H105, S108	29, 104, 105, 106, 108, 183	108	
3PGK 416	**CSA:** R38, K213, G371, G394 **Uniprot:** R38, R121, R168, K217, G310, N334, E341	121, 168, 213, 217, 310, 334, 371, 394	394	

PROT	SITIOS ACTIVOS	$\langle H^r(j)\rangle$	C_c	C_b

Cuadro 4.1: Se muestran los residuos que forman sitios activos en un conjunto de 20 proteínas. La primera columna muestra el ID asociado a cada proteína y el número de residuos que la constituyen de acuerdo al archivo PDB. En la siguiente columna se indican los aminoácidos reportados como sitios activos en diferentes bases de datos. **CSA** es el *Catalytic Site Atlas* [31], **Uniprot** (*Universal Protein Resource*) [30], **PDBSite** [32], **SABER** (*Selection of Active/Binding Sites for Enzyme Redesing*) [28] La primer letra es la abreviatura del aminoácido correspondiente, y en los casos donde existe una segunda letra, ésta se refiere a la cadena en la que se encuentra dicho aminoácido. La tercer columna, señalada con $\langle H^r(j)\rangle$ contiene a los residuos reportados en las bases de datos cuyo número de pasos promedio como receptor, está entre el 15 % de los más bajos. En ésta y en las columnas siguientes, ya no se incluye la abreviatura del aminoácido, pero sí la de la cadena correspondiente. En la columna C_c se señalan los residuos del sitio activo que son o están cercanos a algún máximo local de la centralidad de cercanía, y la columna C_b lo mismo, pero en relación a la centralidad de intermediación.

De los archivos PDB, se obtiene la información relativa a las coordenadas de todos los átomos de la proteína, exceptuando los átomos de hidrógeno. Se utilizan estas coordenadas para calcular la matriz de afinidad y la de adyacencia, con la metodología descrita en los capítulos anteriores. A continuación se da una descripción breve de los resultados para algunas de la proteínas analizadas.

4.2.1. 1YAZ

En primer lugar se muestra la proteína $1YAZ$ (*Superoxide dismutase Cu/Zn*). De las analizadas es la más pequeña, pues tiene sólo 153 residuos formando una cadena. Esta enzima es importante por la protección antioxidante que brinda a las células de una gran cantidad de seres vivos [62]. En la fig. (4.8) se muestra la estructura terciaria a partir del archivo PDB, tanto en la representación 3D, como en forma de red. En las gráficas de esta sección, los vértices de la red están ubicados en el plano, y no representan la posición real de los residuos. En CSA y Uniprot se menciona que el sitio activo realiza funciones catalíticas y además es un sitio de unión de metal Cu/Zn. La gráfica de la matriz **H** da un panorama general acerca de la cantidad de pasos que en promedio requiere un caminante aleatorio en la estructura de la proteína para visitar por primera vez a cualquier aminoácido. La existencia de franjas horizontales de color uniforme, indica residuos para los que en promedio, el número de pasos necesarios para ser visitados por primera vez desde cualquier otro residuo, es constante. Líneas verticales de color uniforme indicarían residuos con valor constante para el número promedio de pasos que el caminante aleatorio que sale de ellos necesita para llegar por primera vez a los demás residuos.

Se pueden ver algunas franjas en azul claro horizontales, por ejemplo, en los residuos previos al residuo 50. Esto significa que en la red se necesitan en promedio pocos pasos para acceder a residuos localizados en esa zona, lo que coincide con los datos experimentales de las bases de datos, pues efectivamente, Uniprot y PDBSite reportan a $His46$ y $His48$ como residuos del sitio activo. La figura que representa $\langle H^r(j)\rangle$ cuantifica con mayor precisión esta observación, ya que para esta cantidad, $His46$ es el mínimo global, y salvo $Glu42$, el resto de los residuos reportados como parte del sitio activo, están en la región sombreada que requiere menos pasos para recibir información.

En la gráfica de la centralidad de cercanía de esta proteína, también se ve que $His46$ y $His48$ están entre los residuos con una mayor cercanía. El primero vuelve a ser un máximo global y aunque los otros residuos reportados no están en la franja de máximos valores, son máximos locales. En la centralidad de intermediación $His46$ no es el máximo global, pero permanece cercano a él. En ambas gráficas, $Glu42$ parece no responder a estas medidas de centralidad.

4.2.2. 1JUK

La proteína 1JUK (*Indole-3-glycerol phosphate synthase*) está formada por una cadena de 248 residuos. Es una enzima que participa como catalizador en la ruptura de enlaces, particularmente de los enlaces C-C. Participa además en la biosíntesis de aminoácidos como triptofano y tirosina [30]. Los resultados se muestran en las figuras (4.9). Las primeras dos son representaciones de la proteína, la estructura terciaria en la primer figura, y en la segunda la red que se obtiene con esta metodología. En ambas se muestran también los residuos del sitio activo reportado en CSA. Notamos que de los 6 residuos señalados, 5 están en la región sombreada que representa al 15 % de los valores mínimos de $\langle H^r(j)\rangle$, siendo $Glu51$ el mínimo global. En el caso de la centralidad de cercanía, 4 de los 6 residuos están en la vecindad de un máximo local, aunque en este caso $Glu51$ pasa a ser el segundo residuo con más altos valores de centralidad, y ahora $Lys110$ es el máximo total. En la centralidad de intermediación también sucede que $Lys110$ es el máximo total y $Glu51$ el segundo residuo con el valor más alto de intermediación, ambos tienen un valor de intermediación considerablemente más grande que el resto de los vértices de la red.

Justo en relación con el comentario anterior, hay otro residuo con alto valor de intermediación, que es $Glu210$ y que no está reportado como parte del sitio activo. De hecho, también tiene un valor alto de cercanía pues aparece en la zona sombreada de la gráfica, e incluso es uno de los picos mínimos locales para $\langle H^r(j)\rangle$. Por lo tanto $Glu210$, desde este punto de vista estructural, cumpliría los requisitos de estar incluido en el sitio activo. En cambio su vecino, el residuo $Ser211$, es reportado como parte del sitio activo, aunque en ninguna de las representaciones gráficas mostradas es un valor crítico local. Esto nos indica que la estructura de la proteína favorece a $Glu210$, pero las propiedades fisicoquímicas del medio y de los propios residuos, determinan que el residuo contiguo, el $Ser211$ sea el que integre al sitio activo.

4.2.3. 2CI7

La enzima $2CI7$ (*N(G),N(G)-dimethylarginine dimethylaminohydrolase 1*), es una enzima que se encuentra en los mamíferos y que participa en la degradación de ciertos compuestos que inhiben la producción de óxido nítrico (NO), compuesto

importante para mantener vasos sanguíneos fortalecidos [63]. Está formada por 284 aminoácidos en una sola cadena, sin embargo en el archivo PDB, no se muestran las coordenadas de los átomos de los primeros 7 aminoácidos y de los 4 últimos, por esta razón en las figuras (4.10) correspondientes, el eje horizontal sólo indica hasta 273 residuos.

En la figura que representa $H(j,i)$, son claras 5 líneas horizontales azules, lo que indica la presencia del sitio activo alrededor de los residuos 80, 120, 170, 210 y 270, lo cual es acertado en parte, pues en Uniprot se indican como parte del sitio activo a $His172$ y $Cys273$, y además marca a $Asp72$, $Glu77$, $Asp78$ y al mismo $Cys273$ como sitio de unión del sustrato y de Zn. PDBSite incluye además a $Phe75$, aunque en ninguna de estas bases se reportan residuos cercanos al 210. La posibilidad de que el residuo 210 sea un residuo con alguna funcionalidad, se ve reforzada en las siguientes gráficas, pues en la de $\langle H^r(j)\rangle$ existe un mínimo local en el aminoácido $Asn220$, casi tan marcado como el mínimo global que corresponde al residuo $Glu77$. También hay un máximo local en la centralidad de cercanía en esa misma posición. Estos resultados conocidos estadísticamente como *falsos positivos*, aparecen con cierta frecuencia en el análisis de proteínas con esta metodología [37]. Estrictamente indican sitios en la red que en promedio requieren pocos pasos para recibir información, o valores extremos en las medidas de centralidad, por lo que podrían representar regiones en la estructura de la proteína, candidatas a constituir un sitio activo, por lo que a nivel experimental podría ser útil hacer un estudio previo de este tipo, para orientar o como antecedente a otros métodos de investigación. Desafortunadamente no se encontró información respecto a alguna posible actividad alostérica de este residuo.

4.2.4. 1KIR

$1KIR$ es una proteína compuesta formada por tres cadenas, dos provienen de anticuerpo monoclonal de ratón, completada con lisozima de huevo de gallina. Este tipo de compuestos son herramientas esenciales en el ámbito clínico, para el diagnóstico y tratamiento de enfermedades infecciosas e inmunológicas [64, 65]. En la fig. (4.11) se presenta el sistema y los resultados. En la representación de la estructura terciaria y la red, se destacan los residuos del sitio activo de la lisozima. Las otras dos cadenas no tienen sitio activo, pero se destacan en la figura algunos de los residuos reporta-

dos en PDBSite como sitios de interacción proteína - proteína. En la gráfica de la matriz **H** se pueden reconocer las tres cadenas, pues para una i dada, hay cambios repentinos entre $H(107, i)$ y $H(108, i)$ y entre $H(213, i)$ y $H(214, i)$, justamente al pasar de una cadena a otra. Observamos que el intercambio de información entre residuos de la misma cadena es más rápido que cuando los residuos son de cadenas diferentes. Incluso por los colores se puede ver que la comunicación entre las cadenas A y B es más rápida que con la cadena C.

En este caso, del modelo se obtienen valores mínimos de $\langle H^r(j)\rangle$ alrededor de los residuos localizados en las posiciones 200 y 350. Se ve en la gráfica de **H** una línea azul que abarca las tres cadenas en las cercanía de dichas posiciones y además, los mínimos más profundos de $\langle H^r(j)\rangle$ corresponden a los residuos $ArgB102$ y $LysC116$. En este caso, ambos residuos forman parte de los sitios de interacción proteína - proteína, no del sitio catalítico. La centralidad de cercanía tiene como máximo global a $TyrB101$, que también aparece como sitio de interacción en PDBSite, además este residuo es uno de los máximos más grandes en la gráfica de la centralidad de intermediación

Como en las anteriores proteínas analizadas, los valores promedio $\langle H^r(j)\rangle$ se adecúan mejor para resaltar los sitios de interacción, que usando medidas de centralidad.

4.2.5. 1PRH

Se muestra a continuación la proteína $1PRH$ (*Prostaglandin H2 synthase-1*). Es una proteína humana, de las más importantes en la síntesis de prostaglandinas, sustancias que afectan y actúan sobre diferentes sistemas del organismo, incluyendo el sistema nervioso, la sangre y el sistema reproductor [66]. En conjunto son 1108 residuos, formando parte de dos cadenas de 554 cada una. Esta es la proteína más larga que analizamos. El tamaño de la red no hace ninguna diferencia teórica, únicamente las matrices son más grandes y en consecuencia, más lenta la resolución numérica de los sistemas de ecuaciones recurrentes para obtener la matriz de primeras visitas **H**.

Los residuos del sitio activo reportados por CSA y Uniprot son $Gln203$, $His207$ y $Tyr385$. Además Uniprot incluye a $His388$ como sitio de unión de Fe. En este caso, estos residuos no coinciden con mínimos locales en ninguna de las medidas hechas, aunque sí es notorio que todos los residuos reportados experimentalmente, a excep-

ción de $Arg120$, están en la región sombreada en la gráfica de $\langle H^r(j)\rangle$. Nuevamente esta cantidad describe mejor la existencia de sitios activos en la estructura de la red que las medidas de centralidad.

Evidentemenmte, si cambiamos el radio de corte R_c, es decir, la distancia que determina que haya o no interacción entre átomos de residuos diferentes, la red se modificará, pues aumenta la cantidad de enlaces y por lo tanto de caminos. Sin embargo pudimos constatar que esta modificación hace que las distancias sean más cortas y que en promedio se necesiten menos pasos para ir por primera vez de un residuo a otro, sin embargo, la forma cualitativa de las gráficas no cambia.

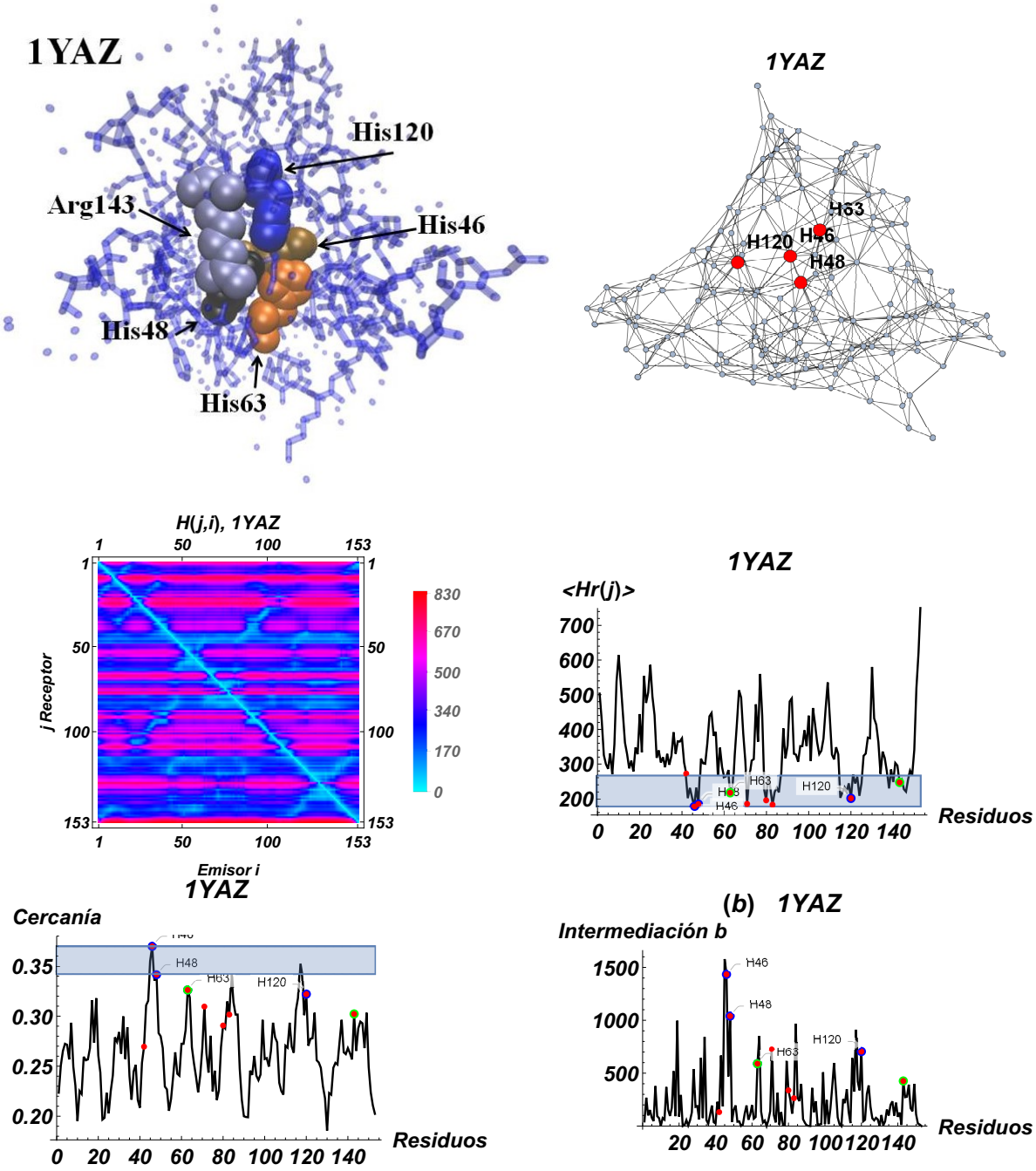


Figura 4.8: Resultados para $1YAZ$ (*Superoxide dismutase Cu/Zn*), proteína formada por 153 aminoácidos en una sola cadena. Se muestra la estructura terciaria resaltando algunos de los residuos reportados, además la estructura en forma de red, La matriz $\mathbf{H}$, $\langle H^r(j)\rangle$, la centralidad de cercanía y la de intermediación. En las últimas 3, los puntos verdes corresponden a los residuos reportados por CSA, los rojos a los obtenidos en Uniprot, y los azules son los de PDBSite.

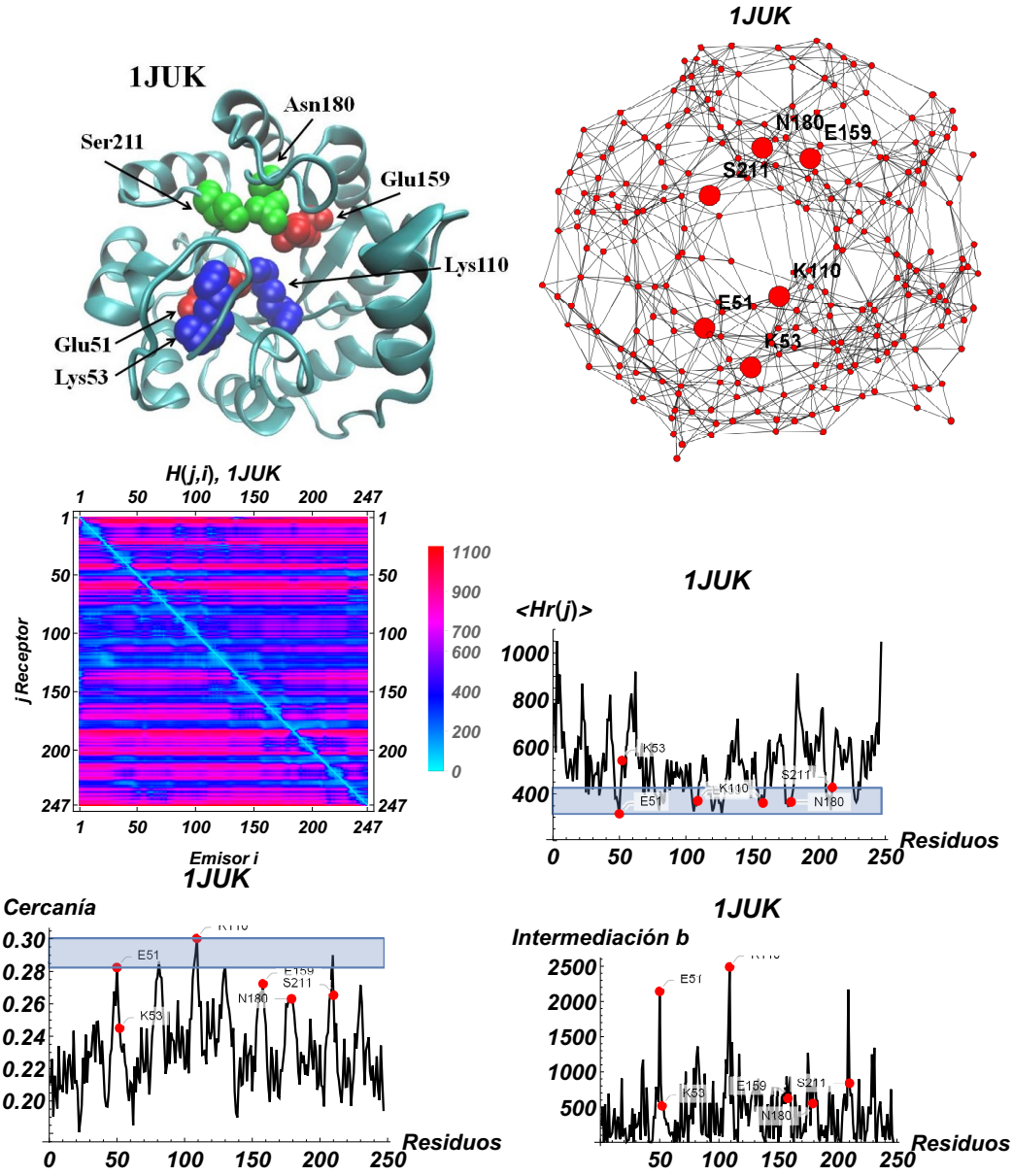


Figura 4.9: Resultados para $1JUK$ (*Indole-3-glycerol phosphate synthase*). Se muestra en las dos primeras figuras la estructura terciaria y la red que se obtiene; la matriz $\mathbf{H}$, $\langle H^r(j)\rangle$ y las medidas de centralidad de cercanía y de intermediación en las siguientes. Esta proteína está formada por una cadena de 248 residuos. Los puntos rojos en las últimas 3 gráficas, indican los residuos que pertenecen al sitio activo reportado en CSA.

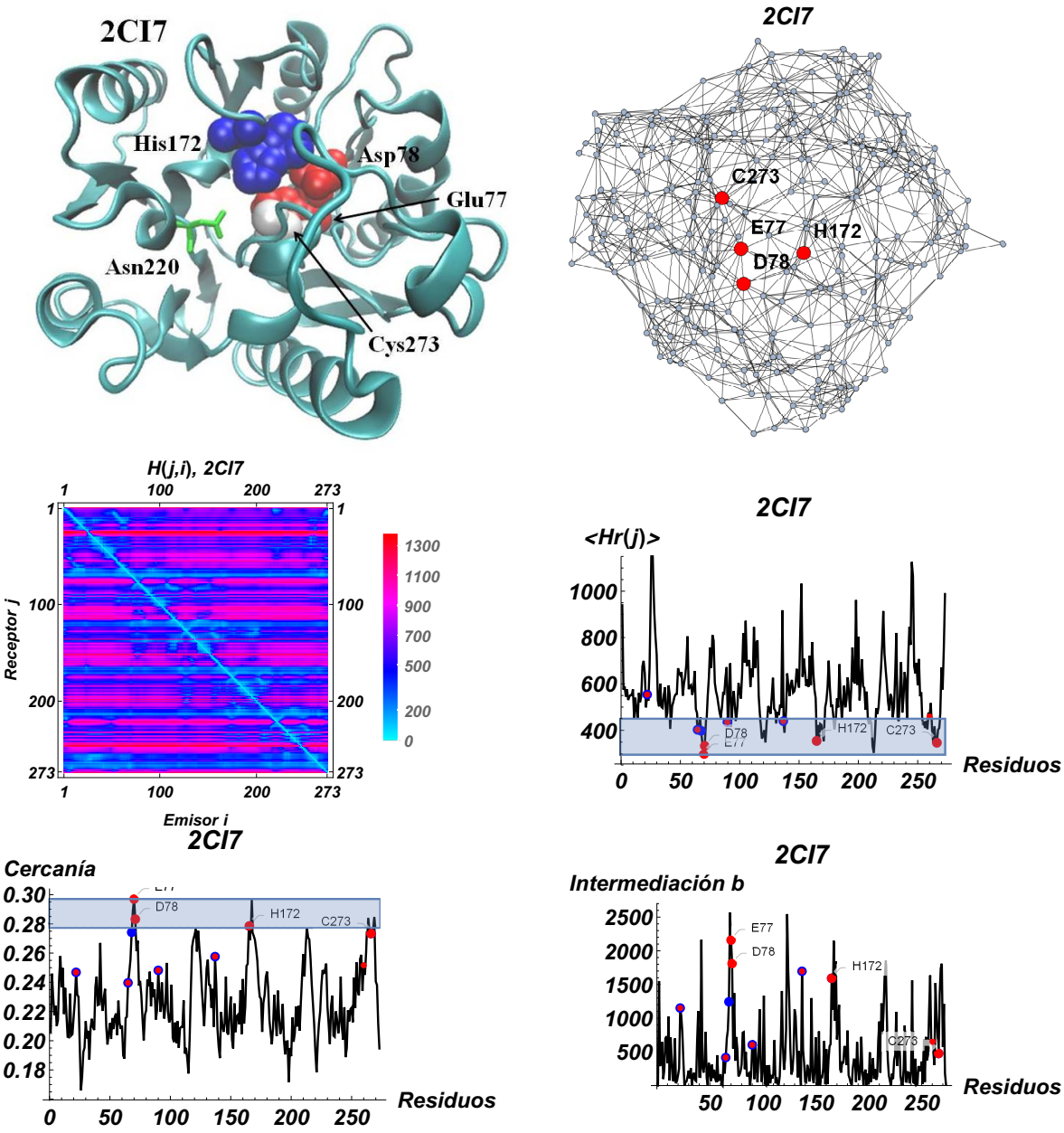


Figura 4.10: Resultados para 2*CI7* (*N(G),N(G)-dimethylarginine dimethylaminohydrolase 1*). Está formada por 284 aminoácidos en una sola cadena. Se muestra la estructura terciaria y los residuos del sitio activo cuyos valores de $\langle H^r(j)\rangle$ son mínimos, incluido el residuo *Asn*220, que aunque no está reportado como parte del sitio activo, es un mínimo local. Se muestra además la estructura en forma de red, La matriz $\mathbf{H}$, $\langle H^r(j)\rangle$, la centralidad de cercanía y la de intermediación. Los puntos rojos de las últimas tres gráficas, son los residuos reportados en Uniprot, y los puntos azules los mencionados en PDBSite.

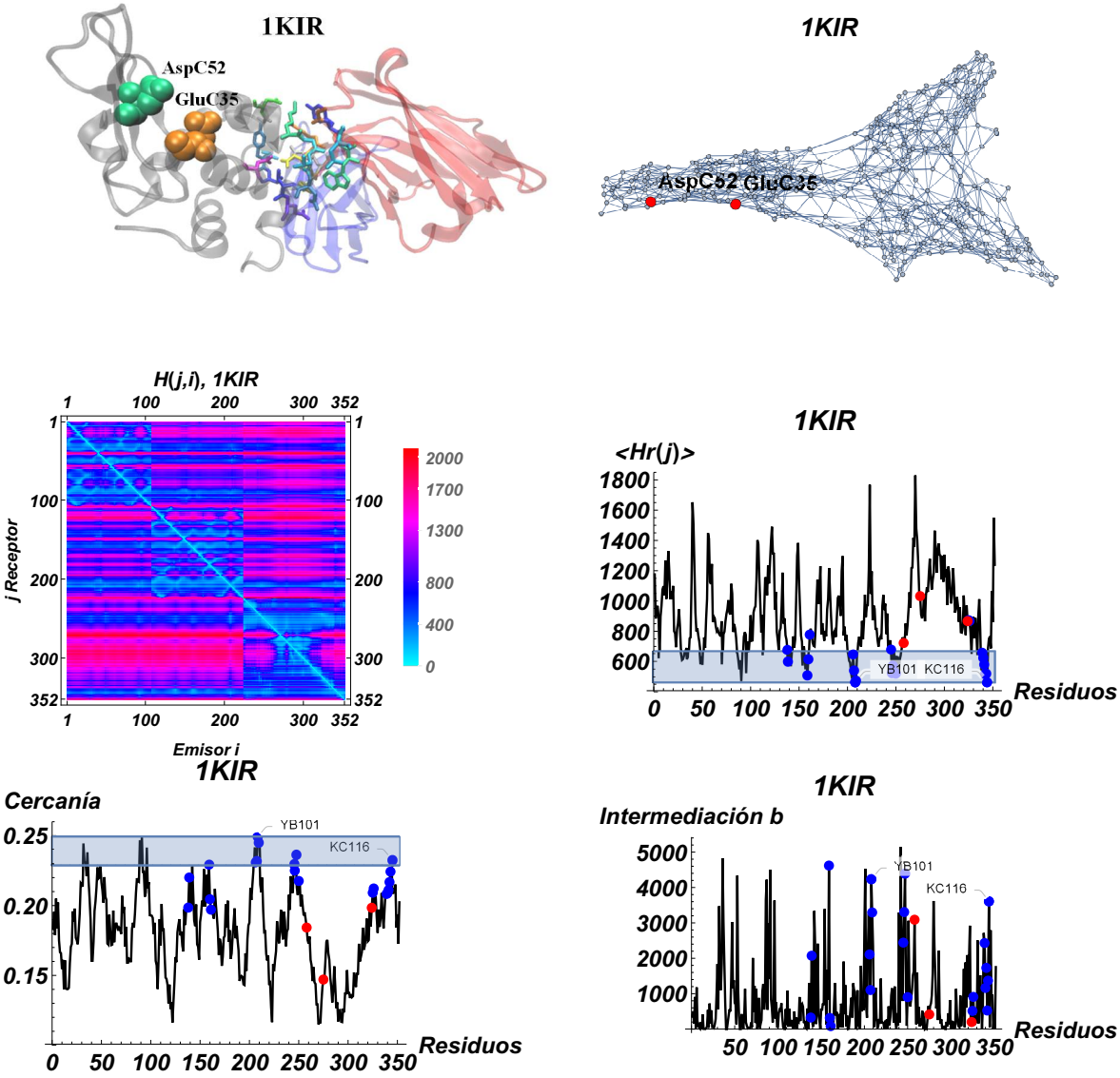


Figura 4.11: Resultados para $1KIR$, una proteína compuesta formada por tres cadenas, dos provienen de anticuerpo monoclonal de ratón, completada con lisozima de huevo de gallina. Las cadenas son de 107, 116 y 129 residuos respectivamente. Las dos primeras no tienen sitio catalítico, pero se incluyen en la estructura algunos residuos señalados en PDBSite como sitios de unión proteína-proteína. Además se resaltan los 2 residuos reportados por Uniprot como parte del sitios activo de la tercer cadena. Se muestra además la estructura en forma de red, La matriz $\mathbf{H}$, $\langle H^r(j)\rangle$, la centralidad de cercanía y la de intermediación. Los puntos azules son residuos de los sitios de unión reportados por PDBSite, mientras que los puntos rojos, son los residuos reportados en Uniprot como partes del sitio activo ($GluC35$, $AspC52$) y como sitio de unión de sustrato ($AspC101$).

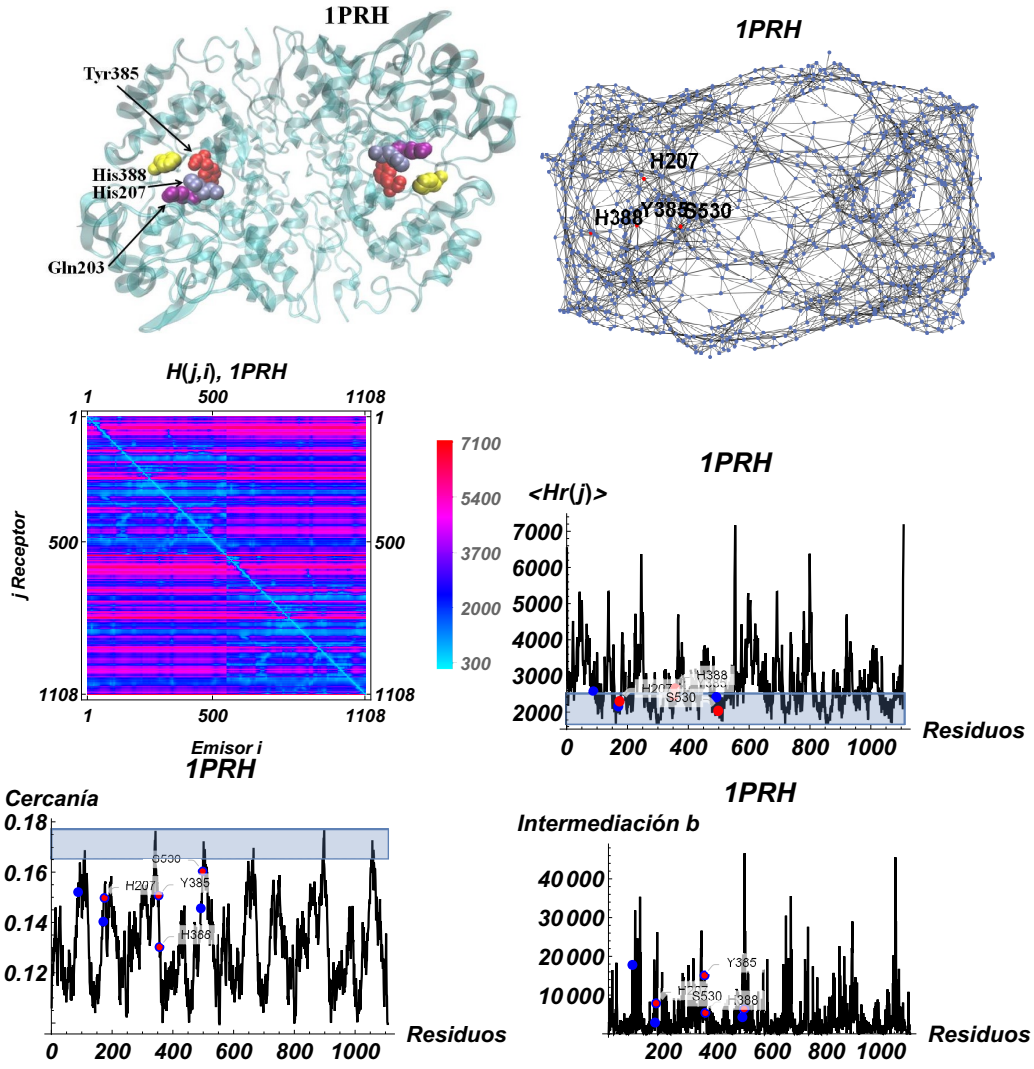


Figura 4.12: Resultados para $1PRH$ (*Prostaglandin H2 synthase-1*). Esta proteína está formada por dos cadenas y 1108 residuos en total. Se muestra la estructura terciaria, la red que se forma y en ambas figuras los 4 residuos con valores mínimos de $\langle H^r(j)\rangle$. Se incluye además la matriz $\mathbf{H}$, $\langle H^r(j)\rangle$, la centralidad de cercanía y de intermediación. Los puntos azules en las últimas 3 gráficas corresponden a los residuos del sitio activo reportados por PDBSite, y los puntos rojos por Uniprot.

Capítulo 5

Conclusiones y Perspectivas

El estudio de proteínas es en la ciencia, una de las áreas que más crecimiento ha experimentado a nivel mundial, en parte debido a los avances tecnológicos y computacionales de las últimas décadas, pues gracias a estos hay ahora una mayor cantidad de herramientas y de mejor calidad, para observar y experimentar a nivel microscópico. Esto ha permitido conocer muchas más estructuras de macromoléculas biológicas con una mejor precisión. Pero además, este auge está motivado por la gran cantidad de aplicaciones que se vislumbran al comprender cómo se construyen y cómo funcionan estas macromoléculas. Las proteínas en particular intervienen prácticamente en todas las funciones de los organismos vivos, por lo que es importante entender cómo es su estructura, cómo se forman, cómo interactúan entre sí y con el medio, y cómo se podrían modificar o diseñar nuevas.

Nuestra contribución en este campo consiste en el estudio de procesos estocásticos y teoría de redes complejas, aplicadas en la estructura tridimensional de proteínas, con la finalidad de ubicar en ellas los sitios activos y los sitios de unión, es decir, en general los lugares que interaccionan con los sustratos, cofactores y con otras macromoléculas, y donde se llevan a cabo las funciones catalíticas de la proteína.

Empleamos Cadenas de Markov para modelar un proceso de transmisión de información de manera estocástica a través de la proteína. Obtuvimos así la matriz de primeras visitas $\mathbf{H}$, que nos permite calcular el tiempo medio que le toma a un caminante aleatorio, llegar por primera vez a un residuo j, cuando proviene del residuo i. Asociamos a este caminante, la transmisión de información sobre la estructura de

la proteína. Con la matriz de primeras visitas calculamos el número de pasos que en promedio necesita cada residuo para recibir o mandar información a cualquier otro.

La primer conclusión general de esta metodología, es que para una proteína no hay diferencia entre los residuos al considerarlos *emisores*, pues el número promedio de pasos para mandar información por primera vez a cualquier otro residuo, $\langle H^e(i)\rangle$ es prácticamente constante para todos. En cambio, como *receptores*, $\langle H^r(j)\rangle$, sí existe una marcada diferencia entre aminoácidos. Esta cantidad es justamente la que relacionamos con los sitios activos y de unión de la proteína, puesto que en este análisis encontramos que en muchos casos los residuos reportados con actividad catalítica o de unión, tuvieron un valor pequeño de este promedio, comparado con el resto de aminoácidos.

Desde un punto de vista estructural, esto quiere decir que aquellos residuos a los que se pude llegar en promedio rápidamente, juegan un papel importante en la funcionalidad de la proteína. Dada la complejidad de reacciones y otro tipo de eventos fisicoquímicos que ocurren en el ambiente real, al comparar los residuos o las regiones con valores pequeños de $\langle H^r(j)\rangle$, con los sitios activos conocidos, se encuentra que esta condición no es suficiente ni necesaria. Podemos ver en las proteínas analizadas, residuos que forman parte de sitios activos conocidos, que no corresponden con valores mínimos de $\langle H^r(j)\rangle$, y también aminoácidos con valores pequeños de esta cantidad, que no son parte de los sitios activos. No obstante los resultados muestran que existe una relación, por lo que creemos que puede ser útil desde un punto de vista experimental, encontrar con esta metodología residuos con características de sitios activos, ya que podrían usarse como punto previo a la utilización de otros métodos o técnicas fisicoquímicas experimentales.

La implementación de este método, es rápida en términos computacionales, pues un código que realice los cálculos que este estudio requiere, no debe tardar más de 1 minuto en una computadora personal común con la versión 11 de Mathematica, para una proteína con unos 600 residuos.

Una dificultad, como se ha discutido en el texto, es que se requiere conocer la posición de todos los átomos de la proteína, exceptuando los átomos de hidrógeno, lo cual representa en ocasiones una complicación extra, pues hay proteínas en el *Protein Data Bank* que no tienen esta información completa. Hay residuos cuya movilidad

impide ubicarlos de forma precisa parcial o completamente, por lo que a veces se omiten, o bien, se presentan dos o incluso más opciones para sus coordenadas. En estos casos, las redes se hicieron sin considerar los aminoácidos faltantes o tomando en cuenta la primer opción de coordenadas presentada en el archivo PDB, y a pesar de eso, los resultados obtenidos son bastante aceptables. Incluso, se podría utilizar información de una proteína que esté cristalizada experimentalmente.

Existe un grupo de proteínas a las que se les ha denominado como *proteínas intrínsecamente desordenadas* (PIDs o IDPs, del inglés Intrinsically Disordered Proteins) o *proteínas no estructuradas* (PINEs o IUPs, de Intrinsically Unstructured Proteins), que son en realidad muy abundantes en la naturaleza, y que son funcionales a pesar de que presentan una estructura de equilibrio flexible o azarosa en toda su estructura o en parte de ella [67, 68]. El modelo que se desarrolló podría utilizarse únicamente si se conociera alguna de las estructuras adoptada por estas proteínas, y si la estructura se modifica se tendría que volver a calcular la red con la nueva disposición de los átomos de la proteína.

Por otra parte, siguiendo un procedimiento similar en cuanto al uso de las coordenadas de los átomos de los residuos, se construyó la matriz de adyacencia **A** correspondiente. Con ella se analizaron los sistemas con la teoría de redes complejas. En particular, se calcularon algunas medidas de centralidad buscando una relación entre éstas y la ubicación de sitios activos. Se probó con la centralidad de cercanía, de intermediación y de vector propio o eigenvector. La *centralidad de cercanía* se refiere a la distancia promedio de un vértice a los demás. Esta cantidad es máxima para los residuos mejor comunicados, para los que están en promedio más cerca del resto. Por otra parte, la *centralidad de intermediación* cuantifica la importancia de un vértice como puente o como parte de caminos dentro de la red. Los residuos con alta intermediación, intercambia información con una buena cantidad de elementos dentro de la estructura de la proteína. La *centralidad de vector propio* es parecida a la centralidad de grado pues también cuantifica la cantidad de vértices adyacentes que tiene cada uno de ellos, pero en este caso, esta medida es pesada por la cantidad de nodos que a su vez tienen los que son adyacentes al primero.

Una conclusión importante en este tema, se obtiene en función de la tabla (4.1), en la que podemos ver que la centralidad de cercanía funciona mejor para ubicar

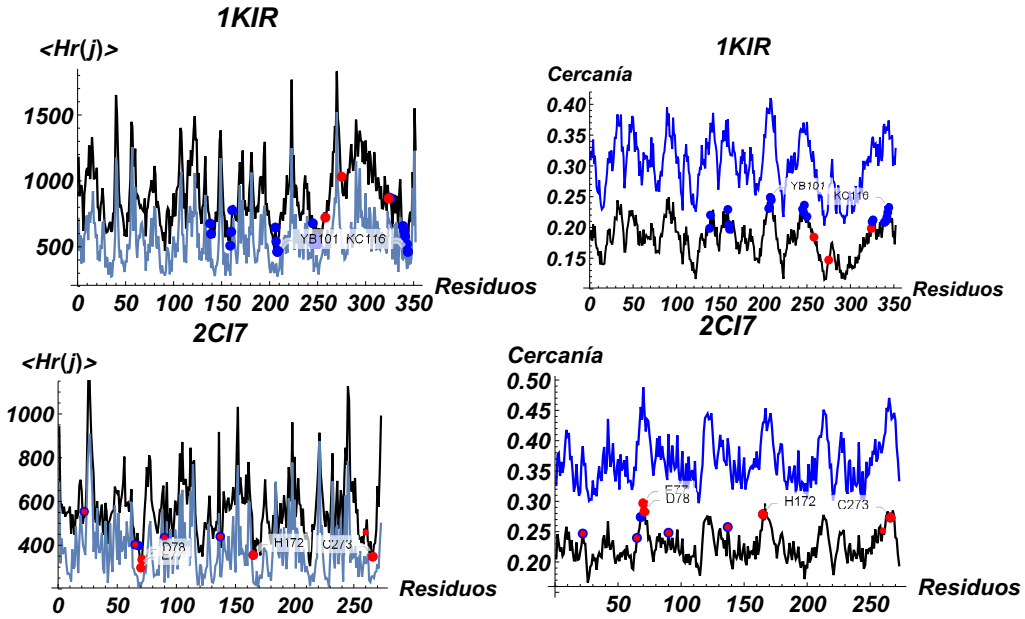


Figura 5.1: Se muestran las gráficas de $\langle H^r(j)\rangle$ y la centralidad de cercanía para las proteínas $1KIR$ y $2CI7$. En todos los casos, las líneas negras son para un $R_c = 4$ Å, y las líneas azules para $R_c = 8$ Å.

residuos funcionales en la estructura de la proteína. En todos los casos analizados, hay más residuos cercanos a los máximos locales de la centralidad de cercanía, que a los de intermediación. No incluimos en la tabla ni en las gráficas los resultados de la centralidad de eigenvector, pues en general obtuvimos una menor similitud entre los máximos o mínimos de esta cantidad con los residuos que forman los sitios activos, que la encontrada usando las otras dos medidas de centralidad.

Los sitios activos y de unión entonces, dentro de la estructura están formados por vértices preferentemente cercanos a los demás residuos, es decir, en promedio son los residuos que requieren menos pasos para comunicarse con el resto. Es importante mencionar que $\langle H^r(i)\rangle$ representa un mejor indicador de sitios activos y de unión, que las medidas de centralidad analizadas. La centralidad de cercanía tiene cierta similitud con este promedio, pues mientras el camino entre dos vértices sea más corto, también será pequeño el número promedio de pasos para ir de uno al otro, tomando en cuenta en el promedio todos los caminos posibles, como se hace en este caso con las cadenas de Markov.

Por otra parte, se encontró que al modificar el radio de corte R_c no se obtiene

un cambio cualitativo en los resultados. Al aumentar R_c aumentan los caminos o la cantidad de enlaces en la red, se reduce por lo tanto el número de pasos para transmitir información y la distancia entre vértices, pero no hay un cambio en la forma de las gráficas de $\langle H^r(i)\rangle$ ni en las medidas de centralidad. En la fig. (5.1) se muestra una comparación en las figuras de $\langle H^r(j)\rangle$ y la centralidad de cercanía para dos diferentes radios de corte, en las proteínas $1KIR$ y $2CI7$. Se observa que los valores de $\langle H^r(i)\rangle$ disminuyen cuando R_c aumenta, mientras que la centralidad de cercanía aumenta al aumentar R_c, sin embargo, la forma cualitativa de las gráficas no s modifica.

Dentro de la teoría de redes complejas, existen otras herramientas de análisis que pueden utilizarse para describir y caracterizar las redes que representan a una proteína [25, 50, 51]. Este puede ser el camino a seguir en esta área de investigación, lo cual dará información no sólo de la estructura funcional de la proteína, sino también de la manera como se organiza y comunica con otras sustancias y macromoléculas de su entorno.

La única información de la proteína que utilizamos para calcular la matriz de transición, es la ubicación de los átomos de todos los residuos. Se podría pensar en incluir además algún factor que tome en cuenta la naturaleza de los aminoácidos, es decir, darle un peso mayor a la probabilidad de transición a aquellos que por sus características fisicoquímicas sean frecuentemente constituyentes de sitios activos. Esto es como una retroalimentación, pues el método proporciona lugares con características geométricas de sitios funcionales, pero también podría considerar con una mayor probabilidad, aquellos residuos que se sabe son generalmente parte de sitios activos. En este trabajo hemos encontrado que la metodología matemática desarrollada puede aplicarse de manera general sobre la estructura de muchas proteínas, aunque no involucramos en el análisis ninguna propiedad fisicoquímicas en especial. Por lo tanto las perspectivas de aplicación son bastante amplias, ya que se pueden analizar conjuntos de proteínas con ciertas características y propiedades particulares.

Por mencionar sólo un ejemplo, se puede hacer un análisis de sitios activos en proteínas *multifuncionales* también conocidas como *moonlight* [69], que son proteínas no estructuradas que realizan más de una función, generalmente no relacionadas. En algunos casos estas actividades ocurren en el mismo sitio activo, pero otras veces

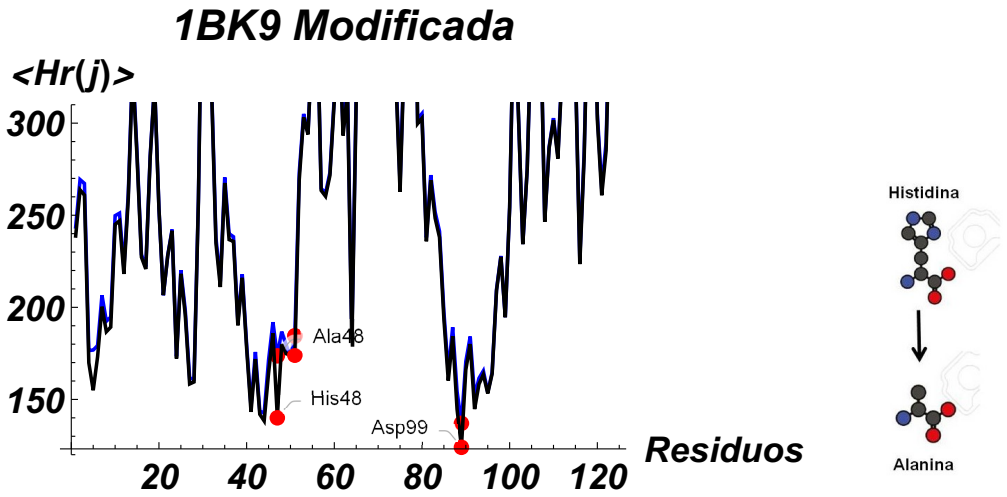


Figura 5.2: Se modificó la proteína $1BK9$ sustituyendo His por Ala en la posición 48 de la estructura primaria. La línea negra de la gráfica muestra la curva de $\langle H^r(j)\rangle$ original, y la línea azul los resultados de la proteína modificada

no. A veces la adquisición de una segunda función deriva en enfermedades o en la aparición de resistencia a antibióticos. Si se conoce la estructura en la cual adquiere la proteína alguna funcionalidad, sería posible aplicar toda la metodología desarrollada.

Otra perspectiva interesante sobre la aplicación de los modelos propuestos, es utilizarlos como parte del diseño de proteínas, estudiando la ubicación de sitios activos en las estructuras nuevas. Esta actividad es muy importante en las industrias biotecnológica y farmaceútica, en donde se busca proteínas que puedan mantener una estructura estable y además tengan mejores propiedades [70], o anticuerpos que contribuyan con el sistema inmune [65]. Para ilustrar este punto, en la fig. (5.2) se muestra el resultado de sustituir el residuo $His48$ del sitio activo de la proteína $1BK9$ analizada, por el residuo Ala, eliminando del archivo PDB los átomos que hacen la diferencia entre uno y otro aminoácido. Se sabe que la alanina no suele participar en sitios activos y se observa al comparar las gráficas de $\langle H^r(j)\rangle$, que para la proteína modificada, el residuo que aparece en la posición 48, ya no está entre los mínimos más importantes.

Por último, debemos mencionar que otro tema que puede ser abordado con esta metodología, es la localización de *sitios alostéricos* (*Allosteric sites*). El alosterismo es una propiedad intrínseca de la proteína, y es el mecanismo por el cual la inter-

acción de una molécula en una parte de ella, afecta a otro sitio localizado en una región diferente de la misma [71, 72], pudiendo incluso ocasionar que la actividad catalítica de los sitios activos desaparezca. Los métodos que hemos descrito en este trabajo, tienen que ver con la transmisión de información a través de la proteína, en consecuencia se espera poder ubicar los sitios que intervienen en esta propiedad.

Bibliografía

[1] Gregory Petsko and Dagmar Ringe. *Protein Structure and Function: Primers in Biology*. Wiley, 2003.

[2] Charles L. Brooks, Martin Karplus, B. Montgomery Pettitt, Carl Branden, and John Tooze. *Proteins: a theoretical perspective of dynamics, structure, and thermodynamics. (Advances in chemical physics; v. 71)*. Wiley, 1988.

[3] J.N. Israelachvili. *Intermolecular and Surface Forces*. Academic Press, 1994.

[4] Carl Branden and John Tooze. *Introduction to Protein Structure*. Garland Publishing, 1999.

[5] Luis Felipe Jiménez and Horacio Merchant. *Biología Celular y Molecular*. Pearson Educación, 2003.

[6] Kazuo Fujiwara, Hiromi Toda, and Masamichi Ikeguchi. Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type. *BMC Structural Biology*, 12:650 – 668, 2012.

[7] Avijit Chakrabartty, Tanja Kortemme, and Robert L. Baldwin. Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Science*, 3:843 – 852, 1994.

[8] Robert J. Moreau, Christian R. Schubert, Khaled A. Nasr, Marianna Török, Justin S. Miller, Robert J. Kennedy, , and Daniel S. Kemp. Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *J. Am. Chem. Soc.*, 131:13107 – 13116, 2009.

[9] Christian B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[10] Leopoldo García Colín S. *Introducción a la Termodinámica Clásica*. Ed. Trillas, 4a. Edición, 1989.

[11] Philip Nelson. *Física Biológica*. Ed. Reverté, 2004.

[12] Alexei V. Finkelstein and Oleg B. Ptitsyn. *Protein Physics. A Course of Lectures*. Academic Press, 2002.

[13] Luis Olivares-Quiroz and Leopoldo García-Colín Scherer. Plegamiento de las proteínas: Un problema interdisciplinario. *Rev. Soc. Quím. Méx.*, 48:95–105, 2004.

[14] C. Levinthal. Are there pathways for protein folding? *Journal de Chimie Physique et de Physico Chimie Biologique*, 65, 1968.

[15] C. Levinthal. How to fold graciously. *Mössbaun Spectroscopy in Biological Systems Proceedings*, 67:22 – 24, 1969.

[16] Alvaro Sebastián and Alberto Pascual-García (Coordinadores). *Bioinformática con $\tilde{N}$*. Libro autoeditado e impreso por CreateSpace, 2014.

[17] Cruzeiro-Hansson L. Protein folding: thermodynamic versus kinetic control. *Journal of Biological Physics*, 27:S6, 2001.

[18] Luis Olivares Quiroz. *El Modelo Extendido de Zwanzig y la Teoría de Adam-Gibbs en el Plegamiento y Desnaturalización de Proteínas*. Tesis de Doctorado, UAM-I, 2007.

[19] Joseph D. Bryngelson and Peter G. Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA*, 84:7524 – 7528, 1987.

[20] Jose N. Onuchic and Peter G. Wolynes. Theory of protein folding. *Current Opinion in Structural Biology*, 14:70 – 75, 2004.

[21] Peter G. Wolynes, Jose N. Onuchic, and D. Thirumalai. Navigating the folding routes. *Science*, 267:1619 – 1620, 1995.

[22] M. Karplus. Behind the folding funnel diagram. *Nature Chemical Biology*, 7:650 – 668, 2011.

[23] K. Dill and H. S. Chan. From levinthal to pathways to funnels. *Nature Structural Biology*, 4:10 – 19, 1997.

[24] Ken A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24:1501 – 1509, 1985.

[25] Gail J. Bartlett, Craig T. Porter, Neera Borkakoti, and Janet M. Thornton. Analysis of catalytic residues in enzyme active sites. *Journal of Molecular Biology*, 324:105–121, 2002.

[26] Gemma L. Holliday, John B. Mitchell, and Janet M. Thornton. Understanding the functional roles of amino acid residues in enzyme catalysis. *Journal of Molecular Biology*, 390:560–577, 2009.

[27] Alex Gutteridge, Gail J. Bartlett, and Janet M. Thornton. Using a neural network and spacial clustering to predict the location of active sites in enzymes. *Journal of Molecular Biology*, 330:719–734, 2003.

[28] Geoffrey R. Nosrati and K. N. Houk. Saber: A computational method for identifying active sites for new reactions. *Protein Science*, 21:697–706, 2012.

[29] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[30] The Uniprot Consortium. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 45:D158–D169, 2017.

[31] Nicholas Furnham, Gemma L. Holliday, Tjaart A. P. Beer, Julius O. B. Jacobsen, William R. Pearson, and Janet M. Thornton. The catalytic site atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Research*, 42:D485–D489, 2013.

[32] Vladimir A. Ivanisenko, Sergey S. Pintus, Dmitry A. Grigorovich, and Nickolay A. Kolchanov. Pdbsite: a database of the 3d structure of protein in functional sites. *Nucleic Acids Research*, 33:183–187, 2005.

[33] Sandro C. Izidoro, Raquel C. de Melo-Minardi, and Gisele L. Pappa. Gass: Identifying enzyme active sites with genetic algorithms. *Bioinformatics*, 31:864–870, 2015.

[34] Jun Gao, Qingchen Zhang, Min Liu, Lixin Zhu, Dingfeng Wu, Zhiwei Cao, and Ruixin Zhu. bsitefinder, an improved protein-binding sites prediction server based on structural alignment: more accurate and less time-consuming. *Journal of Cheminformatics*, 8, 2016.

[35] Jaeju Ko, Leonel F. Murga, Ying Wei, and Mary Jo Ondrechen. Prediction of active sites for protein structures from computed chemical properties. *Bioinformatics*, 21:258–265, 2005.

[36] Joao P. A. Moraes, Gisele L. Pappa, and Sandro C. Izidoro. Gass-web: a web server for identifying enzyme active sites based on genetic algorithms. *Nucleic Acids Research*, 45:315–319, 2017.

[37] Gil Amitai, Arye Shemesh, Einat Sitbon, Maxim Shklar, Dvir Netanely, Llya Venger, and Shmuel Pietrokovski. Network analysis of proteins structures identifies functional residues. *Journal Of Molecular Biology*, 344:1135–1146, 2004.

[38] Chakra Chennublota and Ivet Bahar. Signal propagation in proteins and relation to equilibrium fluctuations. *PLOS Computational Biology*, 3:1716–1726, 2007.

[39] Luis Rincón. *Introducción a los Procesos Estocásticos.* Facultad de Ciencias, UNAM. México, 2012.

[40] J. R. Norris. *Markov Chains.* Cambridge University Press, 1997.

[41] M. E. J. Newman. *Networks. An Introduction.* Oxford University Press, 2010.

[42] Steven H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.

[43] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, 1994.

[44] J. Leónidas Aguirre. Introducción al análisis de redes sociales: Documentos de trabajo. *Centro Interdisciplinario para el Estudio de Políticas Públicas*, 2011.

[45] S.N. Dorogovtsev and J.F.F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, 2003.

[46] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[47] L. A. Adamic and B. Hubeerman. Power-law distribution of the word wide web. *Science*, 287:2115 – 2116, 2000.

[48] J. A. de la Peña. Sistemas de transporte en méxico: un análisis de centralidad en teoría de redes. *Realidad, Datos y Espacio. Revista Internacional de Estadística y Geografía*, 3:72 – 91, 2012.

[49] Broto Chakrabarty and Nita Parekh. Naps: Network analysis of protein structures. *Nucleic Acids Research*, 44:W375–W382, 2016.

[50] Ganesh Bagler and Somdatta Sinha. Network properties of proteins structures. *PHysica A*, 346:27–33, 2005.

[51] Saraswathi Vishveshwara, K. V. Brinda, and N. Kannan. Protein structure: Insights from graph theory. *Journal of Theoretical and Computational Chemistry*, 1:1–25, 2002.

[52] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[53] David Kriesel. *A Brief Introduction to Neural Networks*. Localización: *http* : $//www.dkriesel.com/en/science/neural_networks$, 2005.

[54] David Papo, Javier M. Buldú, Stefano Boccaletti, and Edward T. Bullmore. Complex network theory and the brain. *Phil. Trans. R. Soc. B*, 369, 2014.

[55] J.M. Buldú, J. Busquets, J.H. Martínez, J.L. Herrera-Diestra, I. Echegoyen, J. Galeano, and J. Luque. Using network science to analyse football passing networks: dynamics, space, time and the multilayer nature of the game. *Frontiers in Psychology*, 9, 2018.

[56] Fan R. K. Chung. *Spectral Graph Theory*. Conference Board of the Mathematical Sciences, 1997.

[57] Howard Anton. *Introducción al Álgebra Lineal*. Limusa Wiley, 2010.

[58] Norman Biggs. *Algebraic Graphs Theory*. Cambridge University Press, 1974.

[59] Reka Albert and Albert-Laszlo Barabasi. Statistical mechanics of complex networks. *Review of Modern Physics*, 74:47–97, 2002.

[60] Grégory A. García and Ananías G. Cardona. Fosfolipasas a2: Grandes familias y mecanismos de acción. *Repertorio de Medicina y Cirugía*, 18:199 – 209, 2009.

[61] Sarah N. Croft, Erin J. Walker, and Reena Ghildyal. Human rhinovirus 3c protease cleaves ripk1, concurrent with caspase 8 activation. *Scientific Reports*, 8:1 – 11, 2018.

[62] Joe M. McCord and Irwin Fridovich. Superoxide dismutase. an enzymic function for erythrocuprein (hemocuprein). *J. of Biological Chemistry*, 244:6049 – 6065, 1969.

[63] Raymond J. MacAllister, Heather Parry, Masumi Kimoto, Tadashi Ogawa, Rachel J. Russell, Harold Hodson, Guy St.J. Whitley, and Patrick Vallance. Regulation of nitric oxide synthesis by dimethylarginine dimethylaminohydrolase. *Brit. J. Pharmacology*, 119:1533 – 1540, 1996.

[64] B. A. Fields, F. A. Goldbaum, W. DallÁcqua, E. L. Malchiodi, A. A. Cauerhff, F. P. Schwarz, X. Ysern, R. J. Poljak, and R. A. Mariuzza. Fv mutant y(a 50)s (vl domain) of mouse monoclonal antibody d1.3 complexed with hen egg white lysozyme. *Biochemistry*, 35:15494 – 15503, 1996.

[65] Nina P. Machado, Germán A. Téllez, and John C. Castaño. Anticuerpos monoclonales: desarrollo físico y perspectivas terapéuticas. *SciELO*, 10:186 – 197, 2006.

[66] Y. Ren, D. S. Loose-Mitchell, and R. J. Kulmacz. Prostaglandin h synthase-1: evaluation of c-terminus function. *Arch. Biochem. Biophys.*, 316:751 – 757, 1995.

[67] César L. Cuevas-Velázquez and Alejandra A. Covarrubias-Robles. Las proteínas desordenadas y su función: Una nueva forma de ver la estructura de las proteínas y la respuesta de las plantas al estrés. *Revista Especializada en Ciencias Químico-Biológicas*, 14:97 – 105, 2011.

[68] Vladimir N. Uversky. Introduction to intrinsically disordered proteins (idps). *Chemical Review*, 114:6557 – 6560, 2014.

[69] Sergio I. Hernández Ranzani. *Análisis Bioinformático de las proteínas multifuncionales (Moonlighting)*. Tesis Doctoral, Universitat Autonoma de Barcelona, 2016.

[70] E. Marcos, B. Basanta, T. M. Chidyausiku, Y. Tang, G. Oberdorfer, G. Liu, G. V. T. Swapna, R. Guan, D. Silva, J. Dou, J. H. Pereira, R. Xiao, B. Sankaran, P. H. Zwart, G. T. Montelione, and D. Baker. Principles for designing proteins with cavities formed by curved β sheets. *Science*, 355:201 – 206, 2017.

[71] B. R. C. Amor, M. T. Schaub, S. N. Yaliraki, and M. Barahona. Prediction of allosteric sites and mediating interactions through bond-to-bond propensities. *Nature Communications*, 7, 2016.

[72] A. U. Rehman, S. Saud, N. Ahmad, A. Wadood, and R. Hamid. Allosteric regulation in drug design. *Curr. Trends Biomedical Eng. & Biosci.*, 4, 2017.