



# **UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO**

**PROGRAMA DE MAESTRÍA Y DOCTORADO EN INGENIERÍA  
INGENIERÍA ELÉCTRICA - PROCESAMIENTO DIGITAL DE SEÑALES**

**IDENTIFICACIÓN DE PARTÍCULAS MAL ALINEADAS DURANTE LA  
RECONSTRUCCIÓN DE MAPAS 3D EN CRIO-MICROSCOPIA ELECTRÓNICA**

## **TESIS**

**QUE PARA OPTAR POR EL GRADO DE:**

**DOCTOR EN INGENIERÍA**

**PRESENTA:**

**JEISON MÉNDEZ GARCÍA**

**TUTOR PRINCIPAL:**

**Dr. Edgar Garduño Ángeles, IIMAS-UNAM**

**COMITÉ TUTOR:**

**Dr. Boris Escalante Ramirez, Facultad de Ingeniería-UNAM**

**Dr. Fernando Arámbula, IIMAS-UNAM**

**Ciudad Universitaria, CD. MX. - noviembre del 2021**



Universidad Nacional  
Autónoma de México



**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

**JURADO ASIGNADO:**

**Presidente: Dr. Caleb A. Rascón Estebané**

**Secretario: Dr. Fernando Arámbula Cosío**

**1er. Vocal: Dr. Edgar Garduño Ángeles**

**2do. Vocal: Dr. Gabriel Isaac Corkidi Blanco**

**3er. Vocal: Dr. Carlos Oscar Sorzano Sánchez**

Lugar donde se realizó la tesis:

Ciudad Universitaria, CD. MX.

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas - IIMAS

**TUTOR DE TESIS:**

**DR. EDGAR GARDUÑO ÁNGELES**

---

FIRMA

# Agradecimientos

Quiero agradecer al Posgrado en Ingeniería Eléctrica en su área de Procesamiento Digital de Señales por brindarme la oportunidad de realizar mis estudios.

Al profesor Edgar Garduño Ángeles por su asesoría; por sus valiosos aportes y contribuciones al desarrollo del trabajo.

De igual manera agradezco la participación de los jurados y miembros del comité tutor: el Dr. Caleb Rascón, el Dr. Fernando Arámbula, el Dr. Boris Escalante, el Dr. Gabriel Isaac Corkidi, el Dr. Carlos Oscar Sorzano y el Dr. Edgar Garduño Ángeles.

Este trabajo contó con el apoyo del Centro Nacional de Biotecnología (CNB-CSIC, Madrid España). Allí tuve el valioso apoyo del Dr. Carlos Oscar Sorzano y el Dr. José María Carazo, líderes del grupo de procesamiento de imágenes en el campo de microscopía electrónica; así como de varios de sus investigadores: Amaya Jiménez, Yunior Fonseca y David Strelak. Las estancias que hice allí en Madrid dieron un enorme impulso al desarrollo del trabajo. Dichas visitas no hubieran sido posibles sin la contribución del Programa de Apoyo a los Estudios de Posgrado (PAEP-UNAM).

Aprovecho también para agradecer al Consejo Nacional de Ciencia y Tecnología de México (CONACYT-México) por el apoyo económico sin el cual hubiera sido imposible llevar a cabo mis estudios. El apoyo para mi sostenimiento fue otorgado a través del programa de becas nacionales del segundo periodo de 2017, identificado con el número apoyo 487646. Otra fuente importante de financiación durante mis estudios fue proporcionada por el Ministerio de Ciencias de Colombia (MinCiencias-Colombia) a través de la convocatoria de formación doctoral en el extranjero (Convocatoria 860). A ellos también mi más sincero agradecimiento.

Finalmente, agradezco a tantos y tantos cómplices que en mayor o menor medida hicieron

posible este trabajo, con su apoyo, con su comprensión, con su paciencia, con sus consejos, con sus palabras de ánimo en los momentos difíciles; a todos quienes estuvieron allí para ayudarme a avanzar, muchas gracias!!... Este trabajo está dedicado especialmente a todos ellos: a mi madre, a mi padre, a mis hermanos, a mi familia mexicana (la familia Chamorro, especialmente a Rafa y Charis), a mis parceros Marcos, Iván, Eduard, John y Felipe; al maestro Oscar Montero por sus enseñanzas sobre lo que realmente es importante en la vida ... y a Maribel, mi esposa que además es mi amiga, mi confidente, mi asesora... porque durante estos 6 años que estuvimos en México siempre me brindó su apoyo incondicional.

# Listado de abreviaturas y símbolos

$\text{\AA}$	unidad de medida conocida como ángstrom que equivale a $1 \times 10^{-10}$ metros.
$A$	relación de contraste de amplitud que refleja presencia de absorción en el proceso de adquisición en el microscopio electrónico ( $0 \leq A < 1$ ).
$(\alpha, \beta, \gamma)$	ángulos de Euler que permiten construir la matriz de rotación $\mathbf{R}$ . $\alpha$ y $\beta$ son la rotación y la elevación en el sistema coordenado centrado en el objeto, mientras $\gamma$ es la rotación en el plano de proyección.
$\rho$	función de densidad de los potenciales de Coulomb a través de los cuales se puede describir la estructura de una biomolécula.
$\eta_S, \eta_B$	funciones de ruido que afectan las proyecciones en el microscopio electrónico.
$\varrho(\xi; \Delta z)$	función de perturbación de fase usada en la definición de la CTF.
$\lambda_e$	longitud de onda de los electrones en el haz del microscopio.
$\lambda$	valor propio de una matriz.
$\xi$	coordenada en el espacio de fourier.
$C_s$	constante de aberración esférica del microscopio.
$\phi(f, g)$	coeficiente de correlación de <i>Pearson</i> .
$\psi$	ángulo de rotación en el plano de proyección necesario para alinear rotacionalmente las proyecciones $f$ y $g$ .
CTF	función de transferencia de contraste, $\hat{h}_n$ .

- Crio-EM** técnica de microscopía electrónica de biomoléculas en estado criogénico.
- $d(g_i, g_j)$  distancia angular entre dos proyecciones de referencia que fueron generadas desde dos diferentes direcciones.
- D** matriz que define el grado de los nodos  $V$  del gráfico  $G(V, E)$ ; esto es, la suma de los pesos de los bordes conectados a cada nodo.
- $\hat{\varepsilon}(\xi)$  representación en el espacio de Fourier de la función envolvente del microscopio como sistema óptico.
- $f(x)$  función  $\mathbb{R}^2 \rightarrow \mathbb{R}$  que representa la proyección experimental (imagen experimental) o imagen *class-average*.
- $\hat{f}(\xi) = \mathcal{F}\{f(x)\}$  transformada de Fourier de la función  $f(x)$ .
- f** señal con soporte en los nodos  $V$  del gráfico  $G(V, E)$ .
- $\hat{\mathbf{f}}$  representación espectral de la señal con soporte en los nodos  $V$  del gráfico  $G(V, E)$ .
- $g(x)$  función  $\mathbb{R}^2 \rightarrow \mathbb{R}$  que representa la proyección sintética de un modelo inicial. Esta proyección se usa como imagen referencia en el proceso de refinamiento.
- $G(V, E)$  gráfico que define la relación entre los nodos contenidos en el conjunto  $V$  conectados a través de los bordes  $E$ .
- $\hbar$  constante de planck.
- $h$  psf del microscopio.
- $\hat{h}$  función de transferencia, CTF, del microscopio.
- M** matriz de transformación geométrica.
- L** matriz Laplaciana del gráfico  $G(V, E)$ .
- $\mathcal{P}$  operador que genera la proyección ideal  $[\mathcal{P}\rho]$  como resultado de integrales de línea de las densidades del objeto descrito por  $\rho$ .
- $\mathcal{P}$  conjunto de proyecciones experimentales o *class-average*.

$psf$	función de dispersión puntual que caracteriza un sistema óptico, $h_n$ .
<b>R</b>	matriz de rotación.
$\mathcal{R}$	conjunto de proyecciones de referencia (galería de referencia) usadas en el proceso de asignación angular.
RMN	técnica de Resonancia Magnética Nuclear.
SNR	relación-señal-ruido. Indica la relación de potencias entre la señal de interés y el ruido presente.
SSNR	relación-señal-ruido espectral. Es una medida alternativa de la resolución que puede alcanzar un mapa reconstruido.
SPA	técnica de Análisis de Partículas Individuales.
TEM	técnica de Microscopía Electrónica de transmisión.
$\mathbf{u}_l$	eigenvector de la matriz Laplaciana $L$ asociado al eigenvalor $\lambda_l$ .
$v_1, v_2$	conjuntos tridimensionales de datos (mapas) usados en el calculo de la curva FSC; sus representaciones en el espacio de Fourier son $\hat{v}_1$ y $\hat{v}_2$ , respectivamente.
$v_k$	alguno de los nodos del conjunto $V$ del gráfico $G(V, E)$ .
$V_m$	voltaje de la fuente del microscopio.
<b>W</b>	matriz de pesos del conjunto de nodos $V$ del gráfico $G(V, E)$ .
$x$	vector 2D de coordenadas en el plano de proyección. Es el par ordenado $(x_1, x_2)$ cuyos elementos pertenecen a $\mathbb{R}$ .



# Contenido

<b>1. Introducción</b>	<b>3</b>
1.1. Fuentes de incertidumbre en SPA . . . . .	9
1.2. Motivación del trabajo . . . . .	11
1.3. Objetivo . . . . .	13
<b>2. Marco Conceptual y Estado del Arte</b>	<b>15</b>
2.1. Crio-EM de partículas individuales . . . . .	15
2.1.1. Modelo de formación de imágenes en Microscopía Electrónica . . . . .	16
2.1.2. Corrección de la CTF . . . . .	18
2.1.3. Detectores Directos de Electrones (DEDs) . . . . .	20
2.1.4. Alineamiento de películas . . . . .	21
2.1.5. Selección de partículas . . . . .	22
2.1.6. Alineamiento de partículas y clasificación . . . . .	22
2.1.7. Modelo inicial, refinamiento y asignación angular . . . . .	23
2.1.8. Validación de asignaciones angulares . . . . .	25
2.2. Resolución de mapas en crio-microscopía . . . . .	27
2.3. Procesamiento de señales usando teoría de gráficos . . . . .	29
2.4. Alineamiento de imágenes en frecuencia con correlación cruzada . . . . .	34
<b>3. Diseño Experimental</b>	<b>37</b>
3.1. Resumen del enfoque propuesto . . . . .	46
<b>4. Resultados y Discusión</b>	<b>49</b>

<i>CONTENIDO</i>	IX
4.1. Desempeño de la herramienta de asignación angular . . . . .	50
4.2. Desempeño con proyecciones reales . . . . .	52
4.3. Discusión . . . . .	61
<b>5. Conclusiones y Trabajo Futuro</b>	<b>64</b>
<b>A. Fourier para Señales Indexadas por Gráficos</b>	<b>67</b>
A.1. Propiedades de la Transformada de Fourier . . . . .	67
A.2. Propiedades de la GFT . . . . .	69
<b>B. Algoritmo para Asignación Angular</b>	<b>72</b>
<b>C. Efecto de una Asignación Angular Errónea</b>	<b>74</b>
<b>Referencias</b>	<b>77</b>

# Lista de Tablas

4.1. Resultados de herramienta de asignación usando datos sintéticos. . . . .	51
---	----

# Lista de figuras

1.1. Esquema de microscopio electrónico de transmisión. . . . .	7
2.1. CTF de una micrografía y su perfil unidimensional. . . . .	19
2.2. Ejemplos micrografías. . . . .	20
2.3. Ejemplo de curva FSC . . . . .	29
2.4. Gráfico bi-direccional y sus respectivas matrices. . . . .	31
3.1. Ejemplo de vectores de correlación usando diferentes enfoques. . . . .	41
3.2. Ejemplo señales de correlación indexadas a través de gráficos. . . . .	44
3.3. Resumen del método propuesto para llevar a cabo asignación angular.. . . .	48
4.1. Resultados de mapas reconstruidos usando diferentes algoritmos. . . . .	54
4.2. Histogramas de medidas de calidad de asignaciones angulares previas. . . . .	55
4.3. Resultados de herramienta de asignación usando datos de Crio-EM (Virus). . . . .	57
4.4. Resultados de herramienta de asignación usando datos de Crio-EM ( $\beta$ -gal). . . . .	58
4.5. Resultados de herramienta de asignación usando datos de Crio-EM (apoferritina). . . . .	60
4.6. Comparación contra mapa 3D de alta resolución del BMV. . . . .	61
C.1. Modelo de Ribosoma y ejemplo de proyección ruidosa . . . . .	75
C.2. Curvas FSC para diferentes niveles de asignación angular arrónea. . . . .	76
C.3. Ejemplo de pérdida de detalles de alta resolución. . . . .	76

# Resumen

El análisis de partículas individuales en crio-microscopía electrónica (Crio-EM SPA) es una técnica mediante la cual se pueden obtener mapas 3D que luego permiten resolver las estructuras de biomoléculas; esto resulta relevante, porque en la actualidad existe evidencia de que poseer información de la estructura 3D de las biomoléculas es esencial para entender su funcionamiento y los cambios estructurales y funcionales de éstas.

La técnica consiste en generar proyecciones ortogonales de un conjunto grande de copias (decenas de miles) de una misma biomolécula, orientadas de forma aleatoria; esto es equivalente a “ver” la misma biomolécula desde diferentes puntos de vista. Sin embargo, debido a que las orientaciones o “poses” de las proyecciones experimentales generadas no son conocidas a priori, éstas deben ser determinadas, generalmente comparando las proyecciones experimentales contra las generadas a partir de un volumen de referencia o mapa inicial. La estimación de la orientación para cada proyección experimental no es trivial debido entre otros factores a la baja relación señal-ruido que éstas poseen, generalmente de 10 a 100 veces más ruido que señal en cada proyección.

En el campo de biología estructural se sabe que la calidad de los mapas 3D que se pueden obtener mediante Crio-EM depende de forma crítica de que las posibles orientaciones a partir de las que fueron generadas las proyecciones experimentales sean lo más precisas posible. Sin embargo, la complejidad del problema implica la presencia de múltiples mínimos locales en la función de optimización.

Aunque en la actualidad existen muchos algoritmos para estimar las orientaciones de las proyecciones experimentales, todos ellos haciendo esfuerzos por tener el

menor número de asignaciones erróneas, en menor o mayor medida estos errores se presentan y afectan de forma negativa la resolución de los mapas que pueden alcanzar. Adicionalmente, ninguno de los algoritmos existentes para hacer asignación de orientaciones de proyección hacen una validación de sus propias asignaciones.

En el presente trabajo, propusimos un enfoque para validar las asignaciones de orientaciones hechas por diferentes algoritmos, de manera que se pueda evaluar la cantidad de proyecciones del conjunto final que posiblemente tienen una asignación errónea y que podrían afectar la resolución final del mapa reconstruido. Para esto se utilizó un enfoque basado en el procesamiento de señales con soporte en gráficos, que permite incluir en el proceso, la relación espacial que deberían tener orientaciones vecinas en el entorno de medidas de similitud que resulta de comparar cada proyección experimental contra todas las proyecciones de referencia que se generan a partir del mapa inicial.

# Capítulo 1

## Introducción

Las macromoléculas biológicas (biomoléculas) son esenciales para la vida y son responsables por todas las tareas básicas que se realizan a nivel molecular, tales como la transcripción de material genético, tareas inmunológicas o transmisión y recepción de mensajes. Las biomoléculas se pueden clasificar en las siguientes cuatro categorías: proteínas, fosfolípidos, polisacáridos y el material genético; de éstas, las proteínas tienen un papel relevante debido a la cantidad de tareas de las que son responsables. El tamaño de estas biomoléculas se encuentra en el rango nanométrico, por ejemplo, los anticuerpos tienen una longitud de 10 - 15 nm, el diámetro de una molécula de hemoglobina es de alrededor de 6 nm, el diámetro de una molécula de insulina es de alrededor de 3 - 4 nm y la cadena doble helice de ADN tiene un radio de alrededor 1 nm. Estructuras tan pequeñas requieren de modalidades de imagenología que puedan producir información fiable en el rango atómico.

En la actualidad existe evidencia muy fuerte de que poseer información de la estructura tri-dimensional (3D) de las biomoléculas es esencial para el entendimiento de su dinámica y cambios, tanto estructurales como conformacionales [1]. Además, una vez que se tiene conocimiento de la estructura de las biomoléculas pueden ser determinados los compuestos que pueden unirse a ella (por ejemplo, hormonas y toxinas entre otros), de manera que además de dilucidar los mecanismos de funcionamiento de las biomoléculas se abre paso al desarrollo de aplicaciones tales como diseño de medicamentos e ingeniería de proteínas [1, 2, 3, 4].

Debido a la importancia que tiene una determinación confiable de la estructura de biomoléculas, en las últimas décadas han habido una gran cantidad de esfuerzos orientados al desarrollo de técnicas para dilucidar dichas estructuras y la Microscopía Electrónica en condiciones criogénicas, conocida también como Crio-microscopía Electrónica o Crio-EM, se ha convertido en la más ampliamente usada debido a su flexibilidad y buenos resultados. Antes de que Crio-EM se hiciera popular en la comunidad dedicada a la Biología Estructural, las técnicas más ampliamente usadas para obtener información sobre la disposición estructural de las biomoléculas eran la Cristalografía de rayos-X y la espectroscopía por Resonancia Magnética Nuclear (RMN), también conocida como RMN de proteínas [2]. Hoy en día, a pesar de que estas técnicas siguen siendo usadas, Crio-EM es preferida por muchos investigadores que buscan resolver la estructura de biomoléculas.

La técnica de Cristalografía de rayos-X se basa en la difracción de este tipo de radiación a su paso por sólidos en estado cristalino y tiene su aplicación en el estudio de materiales (incluyendo materiales biológicos). Se puede decir que la mayor limitación de la técnica está en la necesidad de que las muestras se encuentren en estado cristalino, ya que es necesaria la deshidratación de las muestras (lo cual causa cambios estructurales); además, en el caso de muchas proteínas y otras biomoléculas, éstas ni siquiera pueden ser cristalizadas [4].

Por otra parte, la espectroscopía por RMN hace uso de la radiación electromagnética, en el rango de las radiofrecuencias, emitida por los núcleos atómicos que son excitados al ser sometidos a un campo magnético externo (aquellos que tienen número atómico impar o momento magnético distinto de cero), para generar mapas de la distribución del objeto. En Biología Estructural varios de los núcleos más importantes de la química orgánica cumplen con esta condición, sin embargo, esta técnica está principalmente enfocada a proteínas pequeñas y medianas. Esta limitación se debe principalmente a la rapidez con que decae la magnetización en biomoléculas de gran tamaño, dificultando la detección de la señal [4].

Como se mencionó antes, una alternativa a las técnicas mencionadas es Crio-EM, la cual es esencialmente una modalidad tomográfica de imagenología en la que



dependiendo, básicamente, de si se desea analizar una sola biomolécula o un conjunto de ellas se pueden tener dos enfoques: Tomografía Electrónica y Análisis de Partículas Individuales (ET y SPA por sus siglas en inglés, respectivamente); en este proyecto resolvemos un aspecto del proceso de la modalidad de partículas individuales y por ello nos concentraremos en ella. En ambos enfoques se utiliza un microscopio electrónico de transmisión (TEM) que se puede considerar similar al microscopio óptico tradicional pero invertido (uno que tiene el emisor en la parte superior y los detectores al fondo) y que utiliza un haz de electrones para formar las imágenes. En principio la resolución espacial que se podría alcanzar usando electrones acelerados estaría en el orden de los picómetros ( $1 \times 10^{-12}$  m), la cual estaría asociada a la longitud de onda que se podría alcanzar con ciertos potenciales de aceleración [5]. Para enfocar el haz de electrones, el microscopio posee electroimanes que actúan como lentes en un microscopio óptico. El interior del microscopio se mantiene al vacío porque los electrones pueden interactuar con cualquier átomo que se encuentre en su camino entre la fuente emisora y los detectores; esta condición de operación tiene consecuencias importantes para estudiar las biomoléculas. De manera similar a los microscopios ópticos, los microscopios electrónicos de transmisión también sufren de varios tipos de aberraciones, pero algunas de ellas son más complicadas de corregir cuando se usan electrones y electroimanes. Otra característica importante de un TEM es que posee una profundidad de campo infinita (todos los objetos en la trayectoria de los electrones se encuentran en foco).

El hecho de que un TEM opera en condiciones de vacío hace necesario proteger a cualquier material biológico, ya que el alto contenido de agua en éstos hace que sea proclive a evaporarse cuando se encuentra al vacío. Para proteger el material biológico bajo estudio, históricamente, se han implementado varios métodos; por ejemplo, la utilización de partículas metálicas que forman una capa de protección alrededor de la muestra biológica. Esta técnica se conoce como tinción negativa (*negative staining*) y es un método que todavía se utiliza para obtener imágenes de biomoléculas, pero que tiene el inconveniente de que limita la resolución de las imágenes al tamaño de las partículas metálicas, lo cual puede modificar la estructura de las muestras biológicas; el haz de electrones no penetra la estructura interna de las muestras de manera que sólo se crean

imágenes del “molde” de la muestra biológica. En la actualidad, la forma de preservación de especímenes biológicos que afecta menos la resolución y no interfiere con la estructura es la utilización de hielo vítreo (es decir, que no forma cristales) y es por ello que estas modalidades se conocen como Crio-EM.

En la modalidad de partículas individuales se adquiere una sola imagen (micrografía) de la muestra que contiene copias múltiples (generalmente en el rango de miles o decenas de miles) de una biomolécula purificada que dan lugar a imágenes (más pequeñas) de cada copia de la biomolécula que se utiliza para producir el mapa 3D de su densidad de potenciales de Coulomb<sup>1</sup> [6]; aunque la hipótesis es que se purifica una biomolécula, en la práctica no siempre es así y las muestras pueden contener más de una biomolécula o, más importante, biomoléculas en diferentes estados conformacionales (en particular, las proteínas tienen un alto grado de flexibilidad que les permite estar en varios estados funcionales). La idea básica detrás de esta técnica es que todas las copias de la biomolécula se orientan de forma aleatoria en el hielo vítreo durante la preparación y durante la adquisición produciendo múltiples vistas de la misma biomolécula en la micrografía (ver Figura 1.1); este esquema de adquisición permite evitar el daño excesivo a la biomolécula purificada al permitir varias orientaciones en una sola imagen. A pesar de este esquema de adquisición, se utilizan dosis bajas de radiación para evitar daño a la muestra biológica, lo que resulta en micrografías con una baja relación señal-ruido (SNR) (en el orden de 1/10 a 1/100).

Una característica importante en Crio-EM es que cada imagen 2D es considerada una proyección individual de la biomolécula, debido principalmente al reducido espesor de la muestra (del orden 10-100Å) y la baja densidad electrónica de sus elementos químicos constituyentes; estas características proporcionan una débil interacción entre la muestra y el haz de electrones de la fuente, de manera que se considera que muy pocos electrones interactúan con algunos núcleos de la muestra y en caso de hacerlo esta interacción es

---

<sup>1</sup>En el caso de TEM el potencial de Coulomb está relacionado con el efecto que tiene cada núcleo del material sobre los electrones que pasan cerca. Cuando un electrón en su trayectoria del emisor a los detectores pasa cerca del núcleo de un átomo, puede ser desviado debido a que entra en la zona de acción del potencial de Coulomb asociado a la carga neta positiva (en el núcleo), mientras es desviado de su trayectoria original, el electrón puede emitir energía y sufrir un cambio de fase hasta que abandona la zona de acción del potencial del núcleo.

tan débil que el electrón sufre “sólo” un cambio en su fase, es decir que es poco probable que pierdan energía o sean absorbidos [7].

Es importante resaltar que un requisito importante de cualquier modalidad tomográfica es el conocimiento preciso de la orientación de cada proyección; entre más exacta sea esta información, más precisa es la información contenida en el mapa 3D al final. Sin embargo, como se presentará también más adelante en mayor detalle, una característica indeseable en Crio-EM SPA es que la información sobre la orientación de cada proyección de la biomolécula es desconocida. De manera que se hace necesario obtener esta información mediante algún proceso de correspondencia que permita la reconstrucción tomográfica de un mapa 3D que sea compatible con las proyecciones obtenidas en el microscopio.

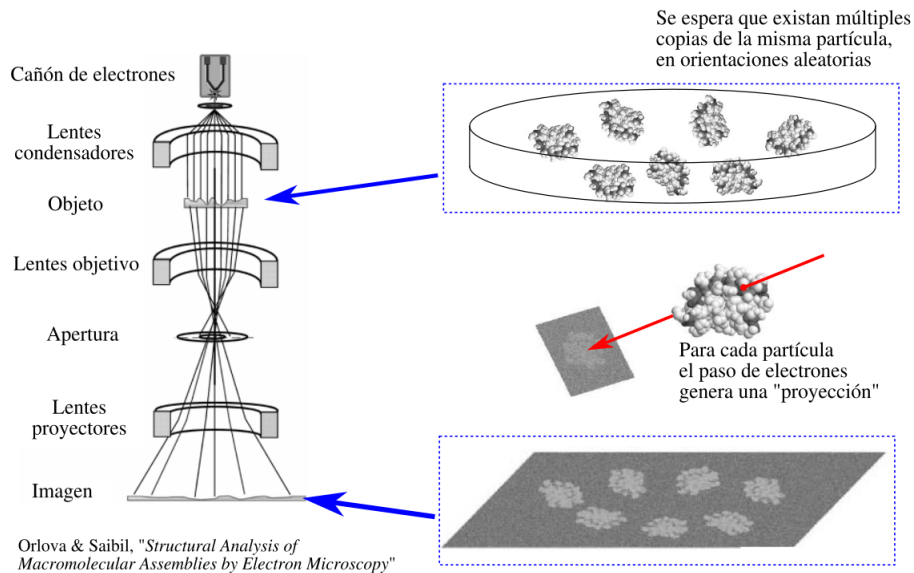


Figura 1.1: Esquema de formación de imagen en microscopio electrónico de transmisión y su uso en Crio-EM SPA.

La popularidad de la que goza la técnica Crio-EM se debe a la forma en la que fue diseñada para sobrellevar las limitaciones de otras técnicas usadas para determinar las estructuras en Biología Estructural [2, 3, 4]. Algunos de los aspectos históricos importantes relacionados con el desarrollo de la técnica se presentan a continuación.

Los doctores Robert Glaeser en [8] y Jacques Dubochet en [9] mostraron que una

de las maneras para conservar las características de alta resolución de biomoléculas en su estado hidratado, además de reducir el daño causado por el haz de electrones, era usar crio-protección (usando hielo vítreo que se forma al utilizar soluciones de congelamiento ultra-rápido). Por otra parte, el desarrollo de los llamados Detectores Directos de Electrones (DEDs, por sus siglas en inglés), gracias a su sensibilidad y velocidad de respuesta, han aumentado por mucho la capacidad de los microscopios electrónicos para detectar eventos asociados al paso de los electrones de la fuente a través de la muestra, permitiendo incluso la captura de una secuencia de imágenes de la misma muestra en los mismos tiempos de adquisición. Adicionalmente, a pesar de la importancia que seguía teniendo la Cristalografía de Rayos-X en los años 80's y 90's, el Dr. Joachim Frank en [10] presentó un método para explotar las características de heterogeneidad y flexibilidad estructural de las biomoléculas al proponer capturar una gran cantidad de proyecciones 2D de éstas en orientaciones aleatorias y someterlas a un proceso de clasificación para posteriormente usarlas en la reconstrucción de la estructura 3D [2]. La combinación de estos desarrollos condujeron al nacimiento de Crio-EM y en particular la técnica de SPA [3, 7]. Como resultado del éxito y la contribución al estudio de biología estructural y molecular, Crio-EM fue elegido como “método del año” en el 2015, y gracias a los aportes a su desarrollo recibieron en el 2017 el Premio Nobel de Química los doctores Joachim Frank, Richard Henderson y Jacques Dubochet [11].

Según presentó J. M. Carazo, et. al. (2015) en [7], un flujo de trabajo típico usando SPA para determinar la conformación estructural de una biomolécula de interés es como se describe a continuación: una vez que las micrografías han sido capturadas a través de los DEDs, son seleccionadas aquellas sub-micrografías de las biomoléculas que a juicio del usuario presentan buena calidad (usualmente son submuestreadas para mejorar su relación-señal-ruido SNR). Luego, a partir de estas micrografías son seleccionadas de manera manual o automática las biomoléculas presentes. Estas imágenes seleccionadas son sometidas a algoritmos de clasificación para determinar algunas heterogeneidades debidas a diferentes factores, tales como contaminación de la muestra, diferentes conformaciones o diferentes especímenes presentes en la muestra; lo anterior se hace con el fin de descartar algunas imágenes. Con las imágenes seleccionadas en esta

etapa de clasificación, se busca agrupar las imágenes experimentales y a partir de estos grupos más homogéneos generar las llamadas imágenes *class-average*, que son básicamente promedios de las imágenes experimentales alineadas que cuentan con una SNR mejorada. Posteriormente, es usado un primer algoritmo para generar un mapa o mapas iniciales 3D de baja resolución, los cuales son una primera aproximación a la estructura de la biomolécula. Finalmente, este modelo inicial, o uno de los mejores candidatos de la etapa anterior, es sometido a un algoritmo iterativo de refinamiento el cual cuidadosamente asigna direcciones de proyección basándose en cada una de las imágenes de entrada.

## 1.1. Fuentes de incertidumbre en SPA

Dos de los retos principales de la técnica de Análisis de Partículas Individuales SPA respecto de otras técnicas tomográficas, utilizadas principalmente en el ámbito clínico, según [7] son: la flexibilidad estructural de los especímenes de interés y el desconocimiento *a priori* de las orientaciones a partir de las cuales se obtienen las proyecciones, como se mencionó antes. Además, otro reto importante es el de estar en capacidad de resolver mapas con una alta resolución en el menor tiempo posible, debido a la gran cantidad de datos que se utilizan para generar mapas 3D en un flujo de trabajo que involucra varios procesos iterativos que son computacionalmente costosos.

En los sistemas de imagenología para llevar a cabo estudios clínicos, el objeto bajo estudio es único, por ejemplo, el cerebro de un determinado paciente, el cual tiene una determinada distribución estructural o funcional que se quiere determinar. A diferencia de esto, en el caso del estudio de biomoléculas, son obtenidas imágenes de partículas que se consideran en principio idénticas. Sin embargo, ellas pueden ser naturalmente flexibles, alternar entre diferentes estados conformacionales, tener diferentes estados de enlace de sus componentes o, simplemente, que la muestra no sea tan pura como se consideraba [7].

Esta flexibilidad estructural afecta el proceso de reconstrucción, de manera que éste debe ser modificado para tener en cuenta dicha flexibilidad. Uno de los aspectos

importantes en el proceso de reconstrucción está relacionado con la geometría que define la adquisición de las proyecciones, es decir, las orientaciones o “puntos de vista” a partir de los cuales fueron generadas las imágenes de proyección. En SPA, como se mencionó antes, no existe conocimiento *a priori* de dicha geometría, entonces además de ser necesario determinarla *a posteriori*, también es necesario incluir el hecho de que las proyecciones pueden provenir de varios especímenes diferentes.

De manera que en SPA no solo se desconocen las direcciones de proyección de cada partícula y deben ser estimadas *a posteriori*, una de las tareas de mayor complejidad en el proceso de reconstrucción [7], sino que también el volumen a partir del cual se obtienen dichas proyecciones debe ser seleccionado de varios candidatos posibles [12, 13, 14]. Además el número de los posibles candidatos tampoco es conocido *a priori*, de manera que es necesario un trabajo de exploración adicional, el cual también se puede complicar ya que los estados conformacionales de la biomolécula pueden no ser discretos sino estar en un rango continuo de distintas conformaciones [15].

Adicionalmente, existen varias fuentes que afectan la calidad de las proyecciones, haciendo que la SNR típica de este tipo de imágenes se encuentre entre 1/10 y 1/100 como se mencionó antes. Estas fuentes de ruido están asociadas, entre otras cosas, a limitaciones técnicas del microscopio y a la preparación de la muestra, a la naturaleza cambiante de los objetos de estudio así como al daño que puede causar la radiación utilizada sobre éstos, entre otras. A continuación se presentan algunas de las fuentes más representativas. Una de las fuentes de ruido proviene del hielo amorfo en el cual se encuentran inmersas las partículas, ya que no permite distinguir claramente entre éstas y el fondo de la imagen debido a su similar composición, aunque se conoce que el ruido inducido por la estructura amorfa del hielo sigue una distribución Gaussiana [16]. Sin embargo, la utilización de la matriz de hielo amorfo impulsó el desarrollo de la técnica Crio-EM ya que protege en gran medida a las partículas del daño que puede causar el haz de electrones a la vez que permite conservarlas en su estado “hidratado”. Relacionado también a la preservación de la muestra, la intensidad del haz de electrones es baja, aunque el desarrollo instrumental en los microscopios busca conservar una buena relación entre estos aspectos [2]. Por otra parte, el haz de electrones

también puede inducir movimiento sobre las partículas durante la toma de imágenes, lo cual representaba una gran fuente de incertidumbre antes de la introducción de los DEDs más rápidos que permiten hacer seguimiento de las partículas en secuencias de imágenes [7]. Adicionalmente, pueden existir problemas asociados a las lentes electromagnéticas que enfocan el haz sobre la muestra y representan otra fuente de incertidumbre que degrada la calidad de las imágenes que se pueden obtener, esto a pesar de que estas aberraciones ópticas se pueden estimar para luego corregirlas en el proceso de reconstrucción del mapa estructural de las partículas.

Aunque si bien es cierto que el desarrollo instrumental en los últimos 20 años ha permitido el avance vertiginoso de la técnica de Crio-EM SPA, reduciendo muchas de las fuentes de incertidumbre existentes, el estudio de métodos computacionales para obtener mapas 3D de las estructuras presentes en las biomoléculas sigue siendo de gran interés en Biología Estructural, debido principalmente a la cantidad de datos que se requieren para generar un mapa 3D y a la gran cantidad de datos que se generan a diario los cuales necesitan procesarse rápidamente y a las limitaciones en resolución que los métodos actuales poseen [7].

## 1.2. Motivación del trabajo

Como hemos mencionado antes, Crio-EM SPA es básicamente una modalidad tomográfica y por ello requiere de las orientaciones de las micrografías que se usan para producir el mapa 3D de densidades de Coulomb. Sin embargo, estas orientaciones se desconocen y es necesario utilizar métodos que permitan asignar la orientación apropiada a cada una de las proyecciones de las biomoléculas embebidas en la matriz de hielo vítreo. Una de las técnicas predominantes en la actualidad para asignar las posibles direcciones de proyección a partir de las cuales pudieron haber sido generadas las imágenes durante el proceso de adquisición (de ahora en adelante, proyecciones experimentales) es conocida como *projection matching*, un método originalmente propuesto por P. A. Penczek, et. al. (1994) en [17]. Este es un proceso iterativo en el cual cada iteración utiliza un modelo previamente reconstruido para generar a partir

de éste proyecciones teóricas que sirven como referencia para encontrar los parámetros de alineamiento correspondientes para las proyecciones experimentales usando algún método basado en medidas de correlación. Luego, estas proyecciones experimentales y sus correspondientes orientaciones de proyección previamente calculadas, son usadas para generar un nuevo mapa 3D (el mapa refinado) el cual puede ser usado en la siguiente iteración del proceso; este proceso continúa hasta cumplir con algún criterio de parada establecido. Un mapa 3D sometido a un filtrado paso-bajas es una elección frecuente a ser usado como modelo inicial, aunque para este tipo de aproximación han sido presentadas más alternativas para construir modelos iniciales *ab-initio* [18, 19, 20]. En la actualidad, la correcta asignación de las orientaciones a las proyecciones experimentales es quizás una de las tareas más relevantes para producir mapas 3D de alta resolución; este proceso es además computacionalmente muy demandante.

A pesar de que varios algoritmos han sido propuestos para resolver el problema de reconstruir mapas 3D a partir de proyecciones 2D experimentales (o alternativamente las *class-average*) [17, 19, 21, 22, 23, 24, 25, 26, 27], ya sea usando un modelo construido *ab-initio* u obtenido previamente, el problema sigue abierto debido al alto costo computacional de la tarea de asignación de orientaciones a las proyecciones experimentales, además de la alta dimensionalidad del proceso de búsqueda entendiéndolo como un problema de optimización [26]; una consecuencia de esta última característica es la posibilidad de que los métodos para asignar orientaciones de proyección queden atrapados en mínimos locales. Como resultado, se espera observar diferencias entre los resultados producidos por diferentes métodos e incluso entre los resultados del mismo método en diferentes ejecuciones, dicha situación es también el resultado de la dependencia de los resultados con respecto a la función objetivo usada para hallar la solución [28]. Además, aunque existen varios algoritmos para estimar los parámetros de orientación durante el proceso de refinamiento [14, 26, 27, 29, 30, 31], todos ellos haciendo un esfuerzo significativo por evitar los mínimos locales, la fracción de proyecciones experimentales potencialmente mal alineadas usando cualquiera de estos enfoques puede no ser despreciable [32, 33]; lo cual dificulta la posibilidad de usar Crio-EM SPA en conjuntos de datos más complejos, buscando obtener mapas 3D de mayor



resolución. Como se mencionó antes, la función objetivo usada para el alineamiento desempeña un papel importante [22, 26, 28], de manera que considerar varias funciones objetivo de manera simultánea puede resultar ventajoso porque se puede dar el caso de que el mínimo local de una función objetivo no lo sea para las otras.

Adicionalmente, el proceso de refinamiento durante el proceso de *projection matching* tiene asociadas varias fuentes de sesgo tales como las algorítmicas (por ejemplo, la dependencia de la calidad del resultado con respecto al modelo inicial) o el sobre-ajuste (*over-fitting*) y el sesgo estructural (*reference bias*) cuyos orígenes pueden ser asociados a la muestra (por ejemplo, la geometría de proyección no uniforme, la cual conduce a direcciones de proyección sub-representadas). Sin embargo, existen importantes fuentes de sesgo que han recibido menos atención por parte de la comunidad, como por ejemplo una estimación incorrecta de la orientación de proyección o de los parámetros de alineamiento bi-dimensional. Por lo tanto, es altamente recomendable incluir etapas de validación en las primeras iteraciones de los enfoques de refinamiento multi-resolución o, incluso mejor, aplicar tareas de validación en el proceso mismo de resolución de los mapas 3D.

### 1.3. Objetivo

El objetivo del presente trabajo es validar las asignaciones angulares hechas por algoritmos de refinamiento de mapas 3D, bajo la suposición de que la orientación asignada a una proyección en particular debería ser consistente con la asignación angular que se hubiera realizado si hubiera menos ruido en el entorno (*landscape*) de la función objetivo.

Para construir el entorno de medidas de similitud, que usa para validar las asignaciones angulares luego del proceso de refinamiento, se construyó una herramienta basada en los trabajos presentados en [34, 35], la cual permite llevar a cabo una asignación angular cuyos resultados son comparables, en media y baja resolución, con los resultados de otras herramientas ampliamente utilizadas en Crio-EM SPA.

Por otra parte, para llevar a cabo el filtrado del entorno de medidas de similitud,

el cual permite identificar algunas proyecciones experimentales cuyas orientaciones asignadas son poco confiables dentro del proceso de refinamiento, se utilizó la descomposición espectral de señales basada en teoría de gráficos, ya que de esa forma se puede explotar la relación espacial entre orientaciones vecinas con respecto a los parámetros de alineamiento y al criterio de similitud usados para alinear las proyecciones experimentales con las proyecciones de referencia.

## Capítulo 2

# Marco Conceptual y Estado del Arte

### 2.1. Crio-EM de partículas individuales

Esta técnica, como se ha mencionado antes, permite obtener estimaciones de la distribución 3D (mapa 3D) del potencial electrostático que tiene relación directa con las estructuras 3D de biomoléculas en su estado hidratado; dicho mapa 3D se obtiene combinando imágenes de muchas partículas individuales (generalmente, en el rango de miles a decenas de miles). El paso de electrones a través de la muestra, la cual contiene atrapadas en una matriz de hielo amorfo múltiples copias de la misma biomolécula (o partícula) con orientaciones aleatorias (ver Figura 1.1), es afectado principalmente por interacciones de Coulomb lo que permite que se genere una imagen 2D, o micrografía, en los detectores de electrones; durante este proceso, como consecuencia de varias fuentes de incertidumbre, algunas de las cuales se mencionan en la Sección 1.1, las micrografías se ven afectadas por altos niveles de ruido (SNR de 0.1 hasta 0.01). Aunque algunos detalles relacionados con la técnica ya han sido presentados en el capítulo anterior, aquí se hace una presentación más completa que cubre aspectos teóricos y prácticos relacionados con el flujo de procesos de la técnica, buscando que el presente documento sea breve pero auto-contenido.

### 2.1.1. Modelo de formación de imágenes en Microscopía Electrónica

En microscopía electrónica de especímenes biológicos bajo condiciones criogénicas se considera que debido a las débiles interacciones entre el haz de electrones y la muestra, el modelo lineal de formación de imágenes es una buena aproximación cuya validez ha sido probada de forma experimental [36]; este modelo considera que debido al reducido espesor y composición de las muestras (principalmente de átomos “ligeros”) solo se ve ligeramente afectada la fase del haz de electrones incidentes mientras que su amplitud permanece intacta.

Bajo esta aproximación, las imágenes 2D generadas por un microscopio electrónico de transmisión son un conjunto de integrales de línea (proyecciones 2D) del potencial de Coulomb del espécimen modificadas por la Función de Dispersión Puntual del microscopio (psf, por sus siglas en inglés) y degradadas por la presencia de ruido [37]

$$f = h * \varepsilon * ([\mathcal{P}\rho] + \eta_S) + \eta_B, \quad (2.1)$$

donde  $f$  es la imagen medida por los sensores,  $\mathcal{P}\rho$  es la “proyección ideal” de la función de densidad de los potenciales de Coulomb  $\rho$ ,  $h$  representa la psf del microscopio y  $\varepsilon$  es la función envolvente, una función de forma suave que decae con las frecuencias altas y que tiene su origen en la falta de coherencia espacial y temporal en el haz de electrones (el haz contiene electrones ligeramente desviados y con velocidades o energías ligeramente diferentes). En el proceso de formación de una imagen final, ésta es contaminada por ruido al nivel de la muestra, representado por  $\eta_S$ , y al nivel del sistema de lentes, representado por  $\eta_B$ ; ambos tipos de ruido se asumen de media cero, no correlacionados mutuamente, no correlacionados con la señal y pueden ser diferentes para cada proyección [37]. El ruido  $\eta_S$  se atribuye a la dispersión residual de los electrones en el solvente y el carbono presentes en el portamuestras y en muchas ocasiones es posible reducir significativamente este tipo de ruido por medios instrumentales; sin embargo, cuando se toma en cuenta se le considera como ruido blanco.

En este trabajo consideramos que una proyección es un conjunto de integrales de

línea que para una función  $\rho : \mathbb{R}^3 \rightarrow \mathbb{R}$  (la función que describe a una biomolécula) se define como

$$[\mathcal{P}\rho](\vec{o}, \mathbf{x}; \mathbf{R}_{\alpha, \beta, \gamma}, \mathbf{t}) = \int_{\mathbb{R}} \rho(\mathbf{R}_{\alpha, \beta, \gamma}(\tau \vec{o} + \mathbf{x}) + \mathbf{t}) d\tau. \quad (2.2)$$

donde  $\vec{o}$  es un vector de dirección (un vector  $\mathbf{o} \in \mathbb{R}^3$  cuya magnitud  $\|\mathbf{o}\|$  es igual a uno) y  $\mathbf{x} \in \mathbb{R}^3$  es un vector perpendicular a  $\vec{o}$  (el producto interno  $\langle \vec{o}, \mathbf{x} \rangle$  entre ambos vectores es igual a cero). El operador  $\mathcal{P}$  provee todas las densidades de la función  $\rho$  en la dirección de la línea recta  $\mathbf{x} + \tau \vec{o}$  conforme varia  $\tau$ . La posición de la función  $\rho$  es modificada por la matriz de rotación  $\mathbf{R}_{\alpha, \beta, \gamma}$ , de tamaño  $3 \times 3$ , definida por ángulos de Euler [38] (los ángulos  $\alpha$  y  $\gamma$  representan rotaciones sobre la línea recta que contiene al vector  $\vec{o}$  y  $\beta$  es una rotación sobre un eje perpendicular a  $\vec{o}$ ) y el vector  $\mathbf{t} \in \mathbb{R}^3$  que representa un posible desplazamiento.

La Transformada de Fourier es un operador esencial para el método que presentamos en esta tesis y por ello la definimos a continuación, este operador mapea una función  $g$  en  $\mathbb{R}^3$  a una función  $\hat{g}$  en  $\mathbb{R}^3$  definida por

$$\hat{g}(\boldsymbol{\xi}) = \int_{\mathbb{R}^3} g(\mathbf{x}) e^{-i\langle \mathbf{x}, \boldsymbol{\xi} \rangle} d\mathbf{x}, \quad (2.3)$$

para todo  $\boldsymbol{\xi} \in \mathbb{R}^3$ . En el espacio de Fourier, el dominio de la función  $\hat{g}$  y también conocido como el espacio recíproco, la ecuación (2.1) se puede escribir como

$$\hat{f} = \hat{h} \times \hat{\varepsilon} \times \left( [\widehat{\mathcal{P}\rho}] + \hat{\eta}_S \right) + \hat{\eta}_B, \quad (2.4)$$

donde  $\hat{h}$  es la Función de Transferencia de Contraste (CTF por sus siglas en inglés) y que típicamente se define en el espacio recíproco de la siguiente manera

$$\hat{h}_n(\boldsymbol{\xi}; \Delta_z) = \sqrt{1 - A^2} \sin \varrho(\boldsymbol{\xi}; \Delta_z) - A \cos \varrho(\boldsymbol{\xi}; \Delta_z), \quad (2.5)$$

donde  $\Delta_z$  es el valor de desenfoque del microscopio,  $0 \leq A < 1$  es la relación de contraste de amplitud (refleja la presencia de absorción) y  $\varrho(\boldsymbol{\xi}; \Delta_z)$  es conocida como la función de

perturbación de fase [39] definida como

$$\varrho(\boldsymbol{\xi}; \Delta_z) = \pi \left( \frac{1}{2} C_s \lambda_e^3 \xi^4 - \Delta_z \lambda_e \xi^2 \right), \quad (2.6)$$

en donde  $C_s$  es la constante de aberración esférica,  $\lambda_e$  es la longitud de onda de los electrones la cual, ignorando efectos relativistas, se puede obtener como

$$\lambda_e = \sqrt{\frac{\hbar^2}{2m_e e^- V_m}}, \quad (2.7)$$

donde  $\hbar$  es la constante de Planck,  $m_e$  y  $e^-$  son la masa y la carga del electrón, respectivamente, y  $V_m$  es el voltaje de operación del microscopio.

Finalmente, la función envolvente  $\hat{\varepsilon}$  de (2.4) está definida de la siguiente forma

$$\hat{\varepsilon}(\boldsymbol{\xi}; B) = e^{-\frac{B|\boldsymbol{\xi}|^2}{4}}, \quad (2.8)$$

y está caracterizada por el factor  $B$  ( $B$ -factor).

Sin perder generalidad, el modelo presentado en (2.4) se reescribe en la práctica como

$$\hat{f} = \hat{h}_n^* \times [\widehat{\mathcal{P}\rho}] + \hat{\eta}, \quad (2.9)$$

donde  $\hat{h}_n^*$  es la función CTF que incluye la información de la envolvente (por facilidad, haremos referencia en el resto del documento a esta última simplemente como CTF) y  $\hat{\eta}$  contiene la información de ambos tipos de ruido  $\eta_S$  y  $\eta_B$ .

### 2.1.2. Corrección de la CTF

Como se presentó antes, se considera que el microscopio electrónico de transmisión funciona como un sistema lineal y por lo tanto las imágenes o proyecciones obtenidas con dicho instrumento son el resultado de la convolución de la “proyección ideal” (aquella producida por un sistema óptico perfecto) y la  $psf$  que caracteriza al sistema; aquí hay que señalar que en el campo de microscopía electrónica se prefiere usar la CTF, la  $psf$  en el espacio de la frecuencia. Un ejemplo típico de una CTF en microscopía electrónica, así

como su perfil unidimensional, se puede ver en la Figura 2.1, es pertinente hacer notar que una CTF ideal es radialmente simétrica, pero las aberraciones pueden modificar este aspecto de una CTF; en esta figura se puede observar que la CTF es una función oscilatoria en el espacio recíproco, lo cual implica que dependiendo de su contenido frecuencial las imágenes de los especímenes bajo observación (las partículas en Crio-EM SPA) pueden parecer brillantes sobre un fondo oscuro o viceversa. Otro aspecto importante de una CTF es la supresión de la componente de corriente directa (DC).

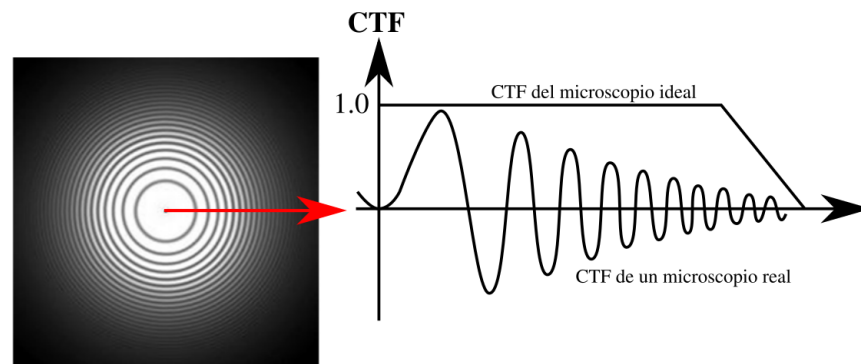


Figura 2.1: Ejemplo de una CTF ideal, la cual es radialmente simétrica, y su perfil unidimensional (fuente [40]).

La estimación de los valores de los parámetros que determinan el comportamiento de una CTF a partir de las imágenes del microscopio es un paso primordial para etapas posteriores de procesamiento. Una vez se estiman los valores de los parámetros que determinan una CTF, se puede realizar un proceso de corrección de las proyecciones adquiridas (un proceso de deconvolución). Para llevar a cabo la estimación de la CTF son necesarios los valores de los parámetros del microscopio usado tales como el voltaje de aceleración usado y la aberración esférica, ver ecuaciones (2.5) y (2.6). Entre algunos de los trabajos en esta dirección se encuentran [41, 42, 43] y en general lo que hacen es ajustar un patrón “ideal” de una CTF al espectro de potencia 2D de las imágenes experimentales del microscopio. En la Figura 2.2 se muestran ejemplos CTFs que ocurren en la práctica, el patrón de la Figura 2.2a se considera una CTF de buena calidad mientras que el patrón de la Figura 2.2b es considerado de mala calidad; normalmente, las micrografías asociadas a CTFs de baja calidad son comúnmente removidas del conjunto de datos que serán utilizados en etapas posteriores del proceso

de reconstrucción.

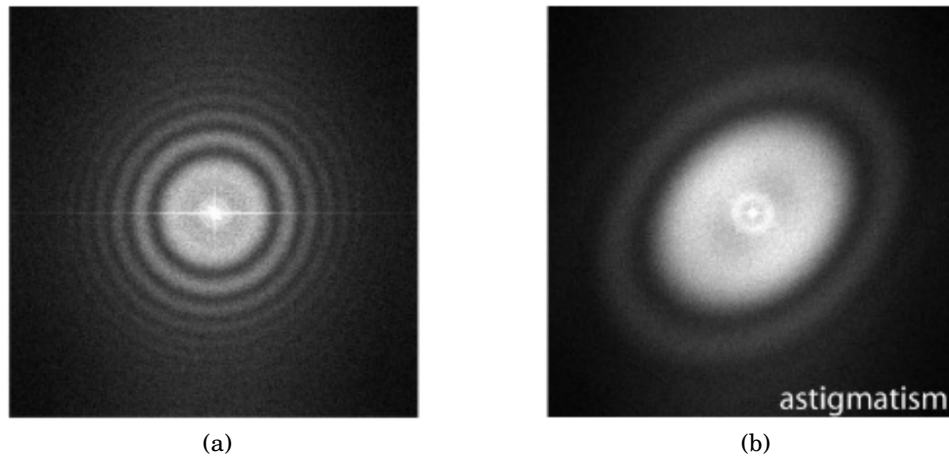


Figura 2.2: Ejemplos de CTFs asociadas e micrografías reales, en la práctica se presentan CTFs (a) de buena calidad y (b) mala calidad, por ejemplo, con astigmatismo. Micrografías con CTFs de mala calidad son comúnmente desechadas en las etapas tempranas del proceso de reconstrucción. (Fuente [44]).

### 2.1.3. Detectores Directos de Electrones (DEDs)

Estos dispositivos, como su nombre lo indica, permiten detectar directamente los electrones que forman la micrografía en un microscopio electrónico a diferencia de los detectores que lo hacen de forma indirecta, estos últimos usan un centelleador (los cuales son transductores que sirven para recibir electrones y producir fotones, típicamente en el rango visible, para ser detectados por un sensor [45]) acoplado por fibra óptica a un sensor, ya sea CCD ó CMOS. Históricamente, en el campo de biología estructural, el desarrollo de estos dispositivos es reconocido como el punto de inflexión a partir del cual inició la “revolución de la resolución” alrededor del año 2014 [46]; el punto a partir del cual ocurre una coyuntura decisiva para el estudio de estructuras biomoleculares, ya que las estructuras que se podían resolver usando Crio-EM empezaron a acercarse cada vez más a resoluciones atómicas y sub-atómicas. Esto es importante porque antes dichas resoluciones sólo se podían lograr con modalidades de imagenología como cristalografía de rayos X, que son menos flexibles que Crio-EM; dicha flexibilidad aunada a la alta resolución alcanzada ha permitido que Crio-EM se haya convertido en una modalidad cada vez más popular en el estudio de estructuras biomoleculares (por cierto,



promoviendo a que muchos cristalografos se muden a esta modalidad).

Existen varias ventajas de estos dispositivos de detección sobre los detectores indirectos. Primero, los DEDs poseen un espesor menor lo cual se traduce en menos dispersión de la señal y, por lo tanto, en mayor resolución. Por otra parte, debido a que no hay etapas intermedias entre la detección del evento y la formación de la imagen (como transporte por fibra óptica) es que hay una reducción de artefactos. Finalmente, y quizá la característica más importante, es la velocidad a la cual se pueden obtener imágenes a la salida del detector, la cual puede ser útil para diferentes técnicas de procesado de imagen, entre ellas la corrección de movimiento inducido por el haz de electrones [47].

Los DEDs tienen dos modalidades para convertir los electrones que lo alcanzan a valores de intensidad en los píxeles de la imagen: modo de integración y de conteo. En el modo de integración, la carga generada por los electrones que alcanzan a un dispositivo DED es acumulada y leída a la salida cada cierto tiempo para formar la imagen final. En cambio, en el modo de conteo cada evento es detectado de forma independiente [40].

#### **2.1.4. Alineamiento de películas**

Una consecuencia de la rápida velocidad de adquisición de los DEDs es que es posible obtener secuencias de micrografías, o películas, de la muestra. Entonces uno de los primeros pasos en el flujo de trabajo es reducir los efectos asociados al movimiento inducido sobre las partículas por parte del haz de electrones alineando las micrografías que hacen parte de la película. Este alineamiento puede hacerse a nivel de la micrografía [47, 48, 49] o a nivel de las partículas [50]. A nivel de las micrografías, los algoritmos hacen un alineamiento de las micrografías en la secuencia para generar una micrografía promedio, la cual tiene un menor nivel de ruido. En el caso de llevar a cabo el alineamiento a nivel de cada partícula, los algoritmos tratan de rastrear individualmente las partículas y alinearlas, lo cual representa un reto mayor ya que generalmente implica un cambio en el flujo de procesos que tradicionalmente se aplican en Crio-EM [40].

### 2.1.5. Selección de partículas

En Crio-EM se utilizan una gran cantidad de micrografías (cantidades que varían en el rango de decenas de miles a cientos de miles) de muy baja resolución para estimar mapas 3D de alta resolución, de manera que identificar manualmente las regiones de interés (las proyecciones de las biomoléculas o partículas) en tal cantidad de imágenes puede ser una tarea tediosa y, potencialmente, una fuente de errores. Por ello, se han desarrollado varios enfoques para llevar a cabo esta selección de manera automática y semi-automática, pero en general es inevitable que se generen algunos errores en la búsqueda y clasificación de las partículas (ya sea falsos positivos o negativos, aunque los primeros son los que tienen mayor impacto en el mapa 3D final).

Una vez que se ha realizado la selección de las proyecciones de partículas, éstas son sometidas a procesos posteriores que incluyen evaluación de calidad, filtrado, corrección de CTF y clasificación. El propósito es que al final de la etapa de procesamiento se cuente con un “buen” conjunto de imágenes para procesos posteriores, las cuales, aunque presenten una baja SNR, por lo menos representen efectivamente proyecciones de las partículas [40].

### 2.1.6. Alineamiento de partículas y clasificación

A pesar de que las proyecciones, o imágenes, de las partículas seleccionadas han sido procesadas, éstas contienen una baja SNR y normalmente no son usadas de manera directa en la construcción del mapa 3D. En cambio, como mencionó el Dr. Joachim Frank en [10], se pueden promediar muchas de estas imágenes para mejorar su SNR. Este proceso solo es posible si las imágenes corresponden a proyecciones de partículas con la misma orientación dentro de la matriz de hielo amorfo y sólo difieren en una traslación y una rotación en un plano perpendicular al haz de electrones de la fuente. Sin embargo, como se mencionó en la Sección 1.1, existen en una micrografía, proyecciones con diferentes orientaciones, proyecciones de diferentes estados conformacionales de las partículas e incluso impurezas, lo cual hace que los conjuntos de imágenes de las partículas sean heterogéneos. Los problemas de alineamiento de partículas y su

clasificación son problemas complementarios, es decir, para clasificar el conjunto de entrada en conjuntos homogéneos, las partículas deben primero ser alineadas; pero la alineación se llevaría mejor al interior de conjuntos homogéneos. Si bien la clasificación en otros problemas de procesamiento de imágenes se puede llevar a cabo usando características invariantes ante rotación y traslación [51, 52], dichas características en el caso de conjuntos de imágenes con altos niveles de ruido no contienen información suficiente para la clasificación [40]. Como resultado, se han desarrollado dos enfoques que han abordado este problema: el análisis estadístico multivariado y la clasificación-alineación multireferencia. El primero de ellos representa cada imagen del conjunto como una combinación lineal de sus auto-vectores principales para luego, en este espacio de representación, llevar a cabo la clasificación en grupos “más homogéneos” [53]; mientras el segundo, proporciona un enfoque iterativo para llevar a cabo tanto la alineación 2D como la clasificación del conjunto de entrada [54, 55].

### **2.1.7. Modelo inicial, refinamiento y asignación angular**

Después de que se han seleccionado las mejores proyecciones experimentales, aquellas más homogéneas, éstas se utilizan para producir el mapa 3D de densidad por medio de un proceso iterativo; el modelo inicial en este proceso es un mapa de baja resolución para evitar cualquier sesgo. La primera iteración utiliza el modelo inicial para generar proyecciones referencia desde posiciones conocidas que son utilizadas para asignar a cada imagen experimental la posible orientación a partir de la cual pudo haber sido adquirida. Para hacer esto, para cada proyección experimental se busca la proyección referencia más parecida, para lo cual la proyección experimental es alineada y luego comparada (por ejemplo, usando correlación cruzada) con todas las imágenes referencia. Para alinear una proyección experimental a una proyección referencia, éstas se ponen en el mismo sistema de coordenadas y se determinan la rotación y traslación relativa entre ellas. Tomando en cuenta que cada proyección experimental se compara contra todas las proyecciones referencia, la tarea de alineamiento es la que más tiempo consume. Una vez que se ha encontrado la proyección referencia más parecida a una experimental, se asignan los parámetros de orientación de la referencia a la

experimental. El proceso de búsqueda para hallar la proyección referencia más parecida es, en general, uno mal condicionado lo cual da a lugar a posibles errores de alineamiento o asignación, sobre todo, considerando las condiciones de SNR con las que se trabaja. Una asignación de parámetros de orientación errónea afecta de manera brusca la resolución final del mapa 3D reconstruido (ver la Sección 2.2 y el Apéndice C para una discusión más amplia). Una vez que se han asignado parámetros de orientación a todas las proyecciones experimentales, se concluye la iteración produciendo así un nuevo mapa 3D (mapa refinado); esto se logra usando las proyecciones experimentales con sus orientaciones como entrada a un método de reconstrucción a partir de proyecciones. Generalmente, en este punto para la reconstrucción del mapa se usan esquemas que permiten rellenar el espacio recíproco y a partir de allí vía transformada inversa de Fourier, obtener el mapa en el espacio real [56, 57, 58]. El mapa 3D resultante puede ser usado como mapa 3D modelo en una siguiente iteración dependiendo del criterio de paro. En este proceso, la selección o generación del mapa 3D inicial es muy importante y existen varias estrategias que incluyen seleccionar un mapa 3D reconstruido en otros estudios o un mapa 3D reconstruido *ab-initio*. Cuando es usado un modelo previamente construido como punto de partida, éste es sometido previamente a un filtrado pasabajas para reducir sesgos provenientes del modelo. En el caso de usar como punto inicial un modelo *ab-initio*, existen varios métodos y herramientas disponibles que permiten obtener esta primera estimación [18, 19, 20, 26]. Existe también una herramienta WEB que ayuda en su determinación y se puede encontrar en [59]. Más recientemente, fue presentado por E. D. Zhong, et. al. (2020) en [60] un enfoque pionero basado en redes neuronales para llevar a cabo el proceso de reconstrucción de mapas *ab-initio*; generalmente los enfoques propuestos para llevar a cabo este proceso, como se ha mencionado antes, están basados en enfoques de máxima verosimilitud y correlación.

El procedimiento recién descrito para asignar los parámetros de orientación a las proyecciones experimentales es conocido como *projection matching* (originalmente propuesto por P. A. Penczek, et. al. (1994) en [17]) y es una de las técnicas predominantes para asignar las posibles orientaciones de proyección a partir de las cuales pudieron haber sido generadas las proyecciones experimentales durante el proceso de adquisición;

en la actualidad la mayoría de los algoritmos de refinamiento usan este mismo principio, las diferencias con el método original incluyen la forma en que llevan a cabo la asignación de parámetros de orientación a cada proyección experimental, el nivel de control por parte del usuario y la función objetivo utilizada para hallar la solución. Por ejemplo, los autores de [13, 14, 54] presentaron un enfoque Bayesiano y de Máxima Verosimilitud en el cual cada proyección experimental puede contribuir en todas las direcciones de proyección con diferentes pesos que son obtenidos de una función heurísticamente determinada, mientras que en [26] se adoptó un enfoque estadístico en el cual las medidas de similitud entre proyecciones son conducidas a intervalos estadísticamente significativos, introduciendo la noción de significancia de una dirección de proyección para una determinada proyección y viceversa, de esta manera una imagen experimental sólo puede contribuir desde algunas direcciones de proyección.

En el presente trabajo usamos el término *asignación angular* para hacer referencia al proceso de asignar a cada proyección experimental una orientación o “pose” de proyección a partir de la cual ésta pudo haber sido generada durante el proceso de adquisición, además de los parámetros de alineamiento necesarios entre dicha proyección experimental y la proyección teórica del modelo de referencia generada desde esa orientación.

A pesar de que han sido propuestos varios algoritmos para estimar los parámetros de orientación durante el proceso de refinamiento [14, 26, 27, 29, 30, 31], todos ellos haciendo un esfuerzo significativo por evitar los mínimos locales, la fracción de proyecciones experimentales potencialmente mal alineadas usando cualquiera de estos enfoques puede no ser despreciable [32, 33]; lo anterior dificulta la posibilidad de que Crio-EM SPA pueda ser usado en conjuntos de datos cada vez más complejos (que potencialmente permitan obtener mapas 3D de mayor resolución).

### **2.1.8. Validación de asignaciones angulares**

Hasta este punto han sido presentados los elementos básicos de un flujo de trabajo típico en Crio-EM, con los cuales se obtiene un mapa 3D de una determinada estructura. Sin embargo, debido al rápido crecimiento que se ha dado en el campo

de Crio-EM, se han empezado a resolver estructuras de las cuales se tiene muy poca información complementaria (por ejemplo la que se podría obtener de mapas resueltos con cristalografía de rayos-X); de manera que para esos casos en particular (aunque es importante en todos los casos) es necesario proporcionar información sobre la consistencia del mapa 3D reportado respecto al conjunto de proyecciones experimentales disponibles.

Una práctica normalmente usada para llevar a cabo la validación de las asignaciones, requiere de analizar las proyecciones partículas tomadas desde dos inclinaciones diferentes del portamuestras (*tilt-pair analysis*) [61]. Con esta técnica se comparan las diferencias entre las asignaciones angulares hechas para las proyecciones experimentales tomadas desde las dos diferentes inclinaciones, teniendo en cuenta que la inclinación relativa entre ellas es conocida. Sin embargo, para este análisis se requiere de recolectar más datos lo que también implica mayor tiempo de procesamiento; además no son pocos los casos en que no se puede realizar este análisis porque no existe la “versión inclinada” de un determinado conjunto de proyecciones; y finalmente, el mismo haz de electrones de la fuente puede causar desplazamiento de las partículas entre un conjunto inclinado de proyecciones y otro, lo cual implica una fuente extra de discrepancia.

Más recientemente, fue presentado en [32, 33] una metodología estadística, la cual no requiere *tilt-pair analysis*, para evaluar la consistencia entre un mapa reconstruido y el conjunto de proyecciones usado para llegar a dicho mapa. En dicho enfoque se propuso buscar, para cada imagen experimental, una distribución ponderada de orientaciones que se obtienen comparando con las re-proyecciones del mapa reconstruido, con respecto a un determinado valor de significancia estadística.

Aunque es conocido que uno de los factores que afecta fuertemente la resolución de los mapas 3D reconstruidos es una estimación poco precisa de las orientaciones de proyección de las proyecciones experimentales [27], existen pocos trabajos en este sentido y la utilización de éstos no ha ganado popularidad en la comunidad debido a que la práctica común es validar los mapas finales reconstruidos, más que evaluar la calidad de las asignaciones que dan origen a dichos mapas [5, 62].

Dicho esto, en este trabajo presentamos un enfoque para validar las asignaciones

angulares hechas a un conjunto de proyecciones; el enfoque está basado en el procesamiento de señales con soportes en gráficos y busca incluir en el proceso de asignación angular, la relación espacial que deberían tener orientaciones vecinas en el entorno de medidas de similitud que resulta de comparar cada proyección experimental contra todas las proyecciones de referencia que se generan a partir del mapa inicial.

## 2.2. Resolución de mapas en crio-microscopía

En la actualidad, una opción para cuantificar la resolución espacial de un mapa 3D reconstruido por un determinado método, es usar la medida conocida como FSC (*Fourier Shell Correlation*), la cual es una generalización de las funciones de correlación de anillos (FRC, por sus siglas en inglés) usado en el análisis de imágenes 2D [63]. La FSC toma dos mapas 3D y retorna una función unidimensional que da idea sobre la correlación que existe entre los dos mapas 3D a diferentes resoluciones espaciales. En el caso de Crio-EM, la FSC proporciona información sobre el desempeño de un proceso de refinamiento, así como de los métodos que forman parte de éste, además de dar una idea sobre la reproducibilidad de los resultados. El procedimiento común para calcular la FSC consiste en dividir en 2 grupos el conjunto de proyecciones para reconstruir un mapa 3D, con cada uno de éstos se lleva a cabo la reconstrucción de un mapa, luego los dos mapas 3D resultantes  $v_1$  y  $v_2$  son comparados mediante el cálculo de la FSC. En el cálculo de la FSC se comparan los coeficientes de Fourier de ambos volúmenes  $\hat{v}_1$  y  $\hat{v}_2$  que se encuentran en varios “cascarones” concéntricos  $\Omega_r$  de radio  $r$  y espesor  $\Delta r$  mediante la siguiente expresión

$$\text{FSC}(r; \hat{v}_1, \hat{v}_2) = \frac{\text{Re} \left\{ \sum_{\xi \in \Omega_r} \hat{v}_1(\xi) \cdot \hat{v}_2^*(\xi) \right\}}{\left\{ \sum_{\xi \in \Omega_r} |\hat{v}_1(\xi)|^2 \cdot \sum_{\xi \in \Omega_r} |\hat{v}_2(\xi)|^2 \right\}^{1/2}}, \quad (2.10)$$

donde  $\xi$  es la 3-tupla que representa las coordenadas en el espacio de Fourier,  $r$  es el radio del cascarón esférico  $\Omega_r$  y el cual es igual a  $|\xi|$ ,  $\text{Re}\{\cdot\}$  es una función que retorna la parte real de un valor complejo y  $\hat{v}_2^*$  es el complejo conjugado de los coeficientes de Fourier de  $v_2$ . Un ejemplo de la curva que se genera mediante esta expresión se puede ver en la

Figura 2.3. En ésta se puede observar que a bajas resoluciones los mapas siempre son similares (la curva tiene valores cercanos a 1.0), luego la curva tiende a decaer de manera más o menos suave (aunque no necesariamente de forma monótona), indicando que a mayores resoluciones los mapas coinciden en menor medida, para finalmente oscilar en valores cercanos a 0.0 para resoluciones cercanas a la frecuencia de *Nyquist*.

En la práctica se define un umbral para la FSC del mapa 3D reconstruido que define la máxima frecuencia espacial a la cual la información contenida se considera confiable (máxima resolución). Aunque el enfoque de usar un valor fijo para proporcionar la resolución es un tema controversial en el campo de microscopía electrónica, son ampliamente usados los valores de 0.5 y 0.143 cuyas pruebas matemáticas rigurosas se pueden consultar en [64, 65].

Algunas aclaraciones sobre el significado de la FSC y sus propiedades, fueron analizadas y presentadas por J. L. Vilas (2019) en [66] e incluimos a continuación algunas de ellas que son relevantes para el presente trabajo. Debido a que la FSC calculada con (2.10) es una medida de correlación cruzada normalizada que mide la auto-consistencia a diferentes frecuencias entre dos funciones  $v_1$  y  $v_2$ , se dice que la FSC no toma en cuenta errores sistemáticos en el proceso de reconstrucción del mapa (o su refinamiento) y por el contrario los premia; por lo tanto, la FSC no debe ser considerada una medida de calidad para el mapa 3D reconstruido. Por otra parte, la resolución máxima que se puede reportar con el uso de la FSC debe ser entendida como la resolución global del mapa 3D reconstruido, considerando como ya se mencionó antes, que la resolución en los mapas reconstruidos puede variar de manera local debido a la presencia de heterogeneidades (conformacionales o estructurales), flexibilidad de la biomolécula, errores de asignación angular o daño estructural por radiación. Finalmente, la FSC se puede entender como una medida alternativa de la SSNR (SNR espectral) de un mapa 3D reconstruido usando la siguiente relación [65, 67, 68]

$$\text{SSNR} = 2 \frac{\text{FSC}}{1 - \text{FSC}} \quad (2.11)$$



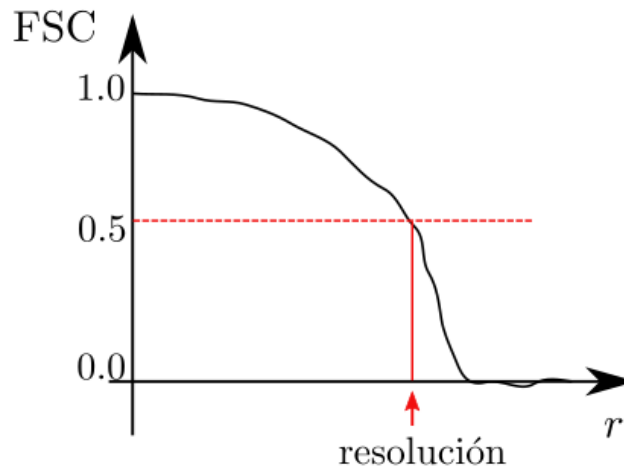


Figura 2.3: Ejemplo de curva FSC

### 2.3. Procesamiento de señales usando teoría de gráficos

Históricamente una gran variedad de técnicas del llamado “procesamiento clásico de señales” han sido desarrolladas para tratar con señales soportadas en los nodos de gráficos simples, por ejemplo, señales discretas unidimensionales e imágenes digitales; sin embargo, en la actualidad hay un creciente interés por hacer uso de estas técnicas en el ámbito de las señales con soporte en los nodos de gráficos más complejos, ya que se han encontrado resultados útiles a problemas modernos como aquellos asociados a redes sociales, redes de información en Internet y redes de monitoreo para adquirir información geoespacial, entre otros [69, 70].

El procesamiento de señales usando representaciones en gráficos (GSP, por sus siglas en inglés) es un enfoque que permite extender el procesamiento clásico de señales discretas a señales que poseen alguna estructura compleja o irregular [69]. Este enfoque es relevante para el presente trabajo ya que permite incorporar la información espacial asociada al proceso de asignación angular durante el refinamiento de modelos iniciales; para esto hay que tener en cuenta que durante la asignación angular se generan señales discretas, de alguna medida de similitud, que resultan de comparar una proyección experimental contra todas las proyecciones en la galería de referencia y que estas últimas son generadas como proyecciones teóricas del modelo inicial desde diferentes orientaciones sobre la esfera unitaria.

Se presentan a continuación algunos conceptos relacionados con el enfoque de GSP (aunque sólo lo necesario para hacer que el presente documento sea auto-contenido). Para las explicaciones de las ideas detrás de este enfoque de procesamiento referimos al lector a [69, 70, 71, 72].

Un gráfico  $G(V, E)$  es una estructura que depende del conjunto finito de nodos, o vértices,  $V$  de cardinalidad  $|V| = N$  y del conjunto de aristas, o arcos,  $E$  que unen dichos nodos; de manera alternativa un gráfico  $G$  se puede definir como una estructura  $G(V, \mathbf{W})$  en donde la matriz simétrica  $\mathbf{W}$ , de dimensiones  $N \times N$ , es la matriz de adyacencia ponderada en donde  $\mathbf{W}_{i,j} > 0$  para el caso de que los nodos  $i$  y  $j$  estén conectados por una arista y  $\mathbf{W}_{i,j} = 0$  en el caso contrario; es pertinente mencionar que la matriz simétrica  $\mathbf{W}$  representa la relación que existe entre nodos, por ejemplo, para dos nodos representando dos puntos o muestras sobre la esfera la arista que los une puede representar la distancia angular que existe entre ellos [73]. Entonces, para representar una señal por medio del gráfico  $G$  basta con indexar sus muestras como nodos para generar el conjunto  $V$  y crear la matriz  $\mathbf{W}$  con la relación existente entre esos valores (por ejemplo, la distancia entre ellos). Alternativamente, la señal que ha sido muestreada y cuyas muestras se representan con el conjunto  $V$  del gráfico también se pueden representar por medio de un vector  $N$ -dimensional  $\mathbf{f}$ .

Un concepto de vital importancia es el concepto de matriz Laplaciana de un gráfico  $G$ . El operador Laplaciano  $\nabla^2 f = \langle \nabla, \nabla f \rangle$  definido para cualquier función  $f : \mathbb{R} \rightarrow \mathbb{R}$  puede ser aproximado para una función  $f : \mathbb{Z} \rightarrow \mathbb{R}$  como  $\nabla^2 f(z) \approx \frac{f(z-\Delta) + 2f(z) + f(z+\Delta)}{\Delta}$  y para un gráfico  $G$  está definido como la matriz simétrica  $\mathbf{L}$ , de tamaño  $N \times N$ , cuyos elementos están dados por

$$l_{i,j} = \begin{cases} \sum_{k=1, k \neq j}^{|V|} w_{i,k}, & \text{si } i = j, \\ -w_{i,j}, & \text{si } i \neq j, \end{cases} \quad (2.12)$$

o de manera alternativa

$$\mathbf{L} = \mathbf{D} - \mathbf{W}, \quad (2.13)$$

donde  $\mathbf{W}$  es la matriz de adyacencia ponderada del gráfico  $G$  y  $\mathbf{D}$  es la matriz de grados la cual es una matriz diagonal que contiene en cada entrada de la diagonal el grado de cada vértice (la suma de los pesos asociados a las aristas que llegan al vértice). En la Figura 2.4 se muestra un ejemplo de un gráfico y sus matrices  $\mathbf{L}$ ,  $\mathbf{D}$  y  $\mathbf{W}$  asociadas.

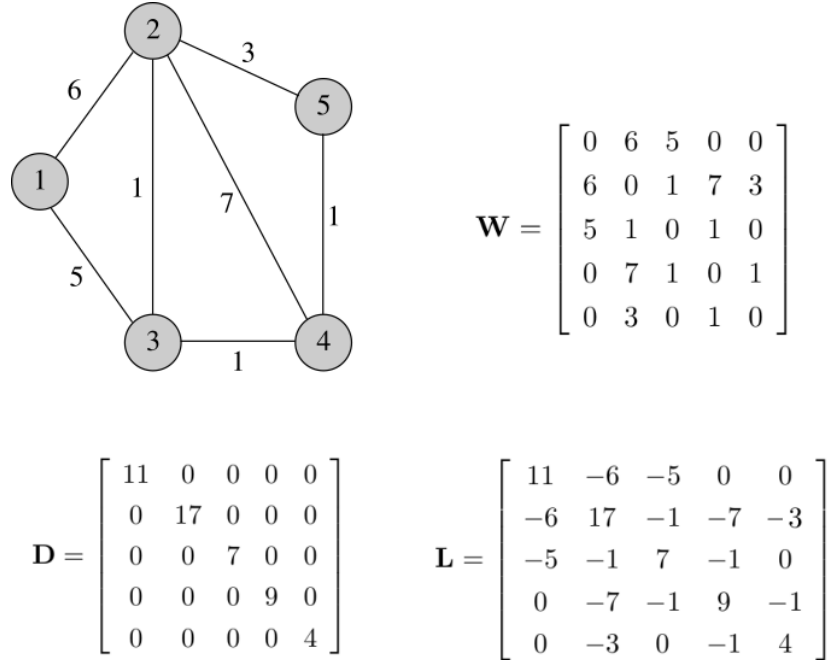


Figura 2.4: Ejemplo de representación en gráfico bi-direccional de una secuencia discreta y sus respectivas matrices. Fuente [69]

Algunas de las propiedades básicas de la matriz Laplaciana  $\mathbf{L}$  de un gráfico  $G$  bi-direccional con matriz  $\mathbf{W}$  no negativa son [69]

- $\mathbf{L}$  es simétrica, lo cual hace que sus valores propios y vectores propios sean reales.
- $\mathbf{L}$  es una matriz singular.
- $\mathbf{L}$  es semi-definida positiva, lo cual hace que sus valores propios asociados  $\lambda_i$  con  $i = 0, 1, \dots, N - 1$ , sean reales no-negativos y que exista al menos un  $\lambda_i = 0$ .
- La suma de cada fila de  $\mathbf{L}$  es 0 en el caso de los gráficos conectados y existe solo un valor propio igual a cero.
- $\mathbf{L}$  tiene un conjunto completo de vectores propios ortonormales, es decir que los vectores propios de esta matriz pueden ser usados como una base ortonormal.

Estas características de la matriz Laplaciana  $L$  hacen que sus vectores propios  $u_i$ , para  $0 \leq i \leq N - 1$ , puedan ser usados como armónicos que permiten hacer análisis en frecuencia de las señales representadas a través del gráfico  $G$ .

Para entender esto último, es necesario retomar algunos conceptos del análisis clásico de Fourier y ver cómo éste se puede relacionar con el análisis de señales discretas indexadas a través de gráficos. Se presentan a continuación algunos elementos y en el Apéndice A una discusión más detallada.

El producto interno entre dos funciones  $g, h : \mathbb{R}^n \rightarrow \mathbb{C}$  se define como

$$\langle g, h \rangle = \int_D g(x) h^*(x) dx, \quad (2.14)$$

donde  $h^*$  es el complejo conjugado de la función  $h$  y  $D$  es el dominio de integración. Es por ello que se puede considerar que la transformada de Fourier de (2.3) para una función  $g : \mathbb{R}^3 \rightarrow \mathbb{R}$  es el siguiente producto interno

$$\widehat{g} = \langle g, e^{j\langle \xi, x \rangle} \rangle. \quad (2.15)$$

Por lo tanto, la transformada de Fourier descompone la función  $g$  como la combinación lineal de las exponenciales complejas  $e^{j\langle \xi, x \rangle}$  (bases de Fourier).

Por otro lado, las funciones propias del operador *Laplaciano* unidimensional son, también, las oscilaciones complejas

$$-\nabla^2 (e^{j\xi t}) = -\frac{\partial^2}{\partial t^2} e^{j\xi t} = \xi^2 e^{j\xi t}. \quad (2.16)$$

Por lo que, las funciones propias del operador *Laplaciano* unidimensional resultan en las bases de Fourier. Así, la Transformada de Fourier puede ser entendida como un cambio de base, en donde las nuevas bases son aquellas funciones que se comportan de manera simple cuando son sometidas al operador *Laplaciano*.

De manera análoga al hecho de que las exponenciales complejas son las funciones propias del operador *Laplaciano* unidimensional, los vectores propios de la matriz Laplaciana del gráfico  $G$  pueden ser usados también como base para descomponer las

señales representadas en gráficos para posteriormente hacer análisis en frecuencia [69].

Se sabe que en el análisis clásico de Fourier el conjunto de valores propios  $\{\xi\}_{\xi \in \mathbb{R}}$  contiene la noción de frecuencia: para valores de  $\xi$  cercanos a cero, las funciones propias  $e^{j\xi t}$  son funciones que oscilan poco, mientras que para valores de  $\xi$  más alejados del origen aumentan las oscilaciones de éstas. En el caso de gráficos conectados y no dirigidos ó bi-direccionales  $G(V, W)$  con valores de conectividad  $W_{i,j} \geq 0$ , los valores propios y vectores propios de la matriz Laplaciana  $L$  del gráfico proporcionan una noción similar de frecuencia. El vector propio  $\mathbf{u}_0$  asociado al valor propio  $\lambda = 0$  es constante e igual a  $1/\sqrt{N}$  en cada vértice. En el caso de valores propios cercanos a cero, los vectores propios asociados varían de manera suave a lo largo del gráfico, es decir, si dos nodos están conectados por una arista con un peso alto, es probable que los valores del vector propio en esas ubicaciones sean similares. Los vectores propios asociados a valores propios altos oscilan más rápido y es más probable que tengan valores diferentes en los nodos conectados por aristas con pesos elevados.

Dicho ésto, se puede definir, como lo hicieron X. Zhu y M. Rabbat (2012) en [70], la Transformada de Fourier del Gráfico (GFT, por sus siglas en inglés)  $\hat{\mathbf{f}} = [\hat{f}(\lambda_0), \hat{f}(\lambda_1), \dots, \hat{f}(\lambda_{N-1})]$  de una señal representada con un gráfico  $\mathbf{f} = [f(0), f(1), \dots, f(N-1)]$ , como una expansión de ésta en términos de los vectores propios de la matriz Laplaciana del gráfico de la siguiente manera

$$\hat{f}(\lambda_i) = \langle \mathbf{f}, \mathbf{u}_i \rangle = \sum_{n=0}^{N-1} f(n) u_i^*(n), \quad (2.17)$$

donde  $\hat{f}(\lambda_i)$  es el coeficiente de la GFT asociado al valor propio  $\lambda_i$ . Adicionalmente, la Transformada de Fourier Inversa del Gráfico (IGFT) está dada por

$$f(n) = \sum_{i=0}^{N-1} \hat{f}(\lambda_i) u_i(n). \quad (2.18)$$

Una vez presentados estos elementos, cabe resaltar la utilidad que tiene esta extensión del procesamiento clásico de señales aplicado a señales discretas con alguna relación compleja entre muestras. En particular, para el presente trabajo se puede

incorporar información extra al proceso de asignación angular que proviene de la relación “espacial” existente entre las imágenes referencia.

## 2.4. Alineamiento de imágenes en frecuencia con correlación cruzada

En este trabajo nos referiremos como *alineamiento* al proceso que determina la rotación, traslación y escalamiento entre dos imágenes de la misma modalidad para que las características de ambas imágenes se traslapen de la mejor manera; dejaremos el término *registro* para el alineamiento de series de imágenes o imágenes provenientes de diferentes modalidades. El alineamiento y registro de imágenes son áreas importantes del procesamiento de imágenes y existen varios enfoques para llevar a cabo estas tareas, entre ellos: los que se llevan a cabo en el espacio de las imágenes [74, 75], los que se llevan a cabo en otro espacio de representación de las imágenes [34, 75] y aquellos que se basan en identificación de características de bajo nivel, como esquinas y bordes, y de alto nivel, como zonas características identificadas en los objetos [75].

Debido a las características de las imágenes que son objeto de estudio en el presente trabajo, se presenta a continuación una breve descripción de una de las técnicas ampliamente usada para determinar los parámetros alineamiento entre imágenes.

En el trabajo presentado por B. S. Reddy, et. al (1996) en [34] y antes, en el trabajo presentado por C. D. Kuglin y D. C. Hines (1975) en [76] fue presentado un método en el dominio de la frecuencia para determinar los parámetros de transformación que mejor permiten alinear un par de imágenes, el cual es una extensión del método de *correlación de fase* usado en la determinación de los retrasos de señales unidimensionales [77]. Este método para alinear imágenes hace uso del teorema de traslación de la Transformada de Fourier. Si dos funciones  $f, g : \mathbb{R}^2 \rightarrow \mathbb{R}$  (imágenes) son iguales excepto por un desplazamiento  $t$ , es decir  $f(x) = g(x - t)$ , entonces sus correspondientes representaciones en el espacio de Fourier,  $\hat{f}, \hat{g} : \mathbb{R}^2 \rightarrow \mathbb{C}$  (también denotadas  $\hat{f}(\xi) = \mathcal{F}\{f(x)\}$  y  $\hat{g}(\xi) = \mathcal{F}\{g(x)\}$ , respectivamente), se relacionan por

medio de

$$\widehat{f}(\boldsymbol{\xi}) = \widehat{g}(\boldsymbol{\xi}) e^{-j\langle \boldsymbol{\xi}, \mathbf{t} \rangle}. \quad (2.19)$$

Dicho esto, el espectro cruzado de potencia entre las imágenes  $f_1$  y  $f_2$  se puede definir a través de sus representaciones en el espacio de Fourier como

$$\frac{\widehat{f}(\boldsymbol{\xi}) \odot \widehat{g}^*(\boldsymbol{\xi})}{|\widehat{f}(\boldsymbol{\xi}) \odot \widehat{g}^*(\boldsymbol{\xi})|} = e^{-j\langle \boldsymbol{\xi}, \mathbf{t} \rangle}. \quad (2.20)$$

Esta relación puede ser usada para determinar la diferencia de fase entre las dos imágenes  $f$  y  $g$  ya que el resultado de la transformada inversa de Fourier de (2.20) resulta en una imagen que tiene un valor de aproximadamente cero en todas partes excepto en el valor de desplazamiento necesario para alinear ambas imágenes.

Ahora bien, si las imágenes  $f$  y  $g$  son iguales excepto por una traslación  $\mathbf{t} = (t_x, t_y)$  y una rotación  $\gamma$  sobre el plano, es decir  $f(\mathbf{x}) = g(\mathbf{R}_\gamma \mathbf{x} + \mathbf{t})$  con la rotación definida por la matriz  $\mathbf{R}_\gamma = \begin{bmatrix} \cos \gamma & \sin \gamma \\ -\sin \gamma & \cos \gamma \end{bmatrix}$ , entonces, de acuerdo al teorema de traslación, los espectros se relacionan según

$$\widehat{g}(\boldsymbol{\xi}) = \widehat{f}(\mathbf{R}_\gamma \boldsymbol{\xi}) e^{-j\langle \boldsymbol{\xi}, \mathbf{t} \rangle}. \quad (2.21)$$

Para eliminar la dependencia con la traslación, se calcula la magnitud de los espectros de ambas funciones y se puede ver que

$$|\widehat{g}(\boldsymbol{\xi})| = |\widehat{f}(\mathbf{R}_\gamma \boldsymbol{\xi})|, \quad (2.22)$$

lo que significa que las funciones son iguales excepto por una rotación. Finalmente, si (2.22) es representada en coordenadas polares  $(r, \theta)$  se obtiene

$$|\widehat{g}(r, \theta)| = |\widehat{f}(r, \theta - \gamma)|, \quad (2.23)$$

lo que permite interpretar la rotación  $\gamma$  en el plano de proyección como una traslación lineal de una representación en coordenadas polares de la magnitud de los espectros de

las imágenes. Por lo tanto, para encontrar la rotación relativa  $\gamma$  sobre el plano entre dos proyecciones (imágenes) se realiza un procedimiento similar al que se lleva a cabo para obtener el desplazamiento, pero en este caso primero se lleva a cabo el procedimiento con las imágenes representadas en coordenadas polares; en otras palabras, hallando la coordenada del valor máximo de la imagen que se obtiene al calcular la Transformada de Fourier inversa de (2.20) usando  $f' = |\hat{f}(r, \theta)|$  y  $g' = |\hat{g}(r, \theta)|$ .

El procedimiento presentado permite determinar de forma independiente la rotación y la traslación relativa entre dos imágenes de interés, lo cual representa una ventaja frente a otros métodos, especialmente los que se aplican en el espacio real, los cuales necesitan aplicar varias traslaciones y rotaciones durante el proceso de alineamiento.



## Capítulo 3

# Diseño Experimental

El enfoque conocido como *projection matching*, como se mencionó en la Sección 2.1.7, es uno de los más usados en SPA para asignar las orientaciones de proyección al conjunto de imágenes experimentales (el conjunto  $\mathcal{P}$ ) usando la galería de imágenes referencia (el conjunto  $\mathcal{R}$ ) generadas como proyecciones teóricas de un mapa o modelo 3D inicial. Este procedimiento puede ser descrito como sigue, para cada proyección experimental  $f \in \mathcal{P}$  el método de asignación de orientaciones de proyección encuentra la proyección referencia  $g$  por medio de

$$g = \arg \max_{g_i \in \mathcal{R}} \phi(g_i, M_i(f)), \quad (3.1)$$

donde  $M_i$  es la transformación geométrica óptima para alinear la proyección experimental  $f$  con una proyección  $g_i$  de la galería de referencia y  $\phi$  es la medida de similitud entre proyecciones. Con este enfoque lo que se hace es asignar a la proyección experimental  $f$  la dirección de proyección desde donde fue generada la proyección referencia  $g$ ; dicha información resulta de la búsqueda mediante (3.1) así como los parámetros de alineamiento necesarios. La medida de similitud  $\phi$  elegida en el presente trabajo se conoce como el Coeficiente de Correlación de *Pearson* [78, 79] y se define como

$$\phi(f, g) = \left| \frac{\sum_j^J (f(\mathbf{x}_j) - \bar{f})(g(\mathbf{x}_j) - \bar{g})}{\sqrt{\sum_j^J (f(\mathbf{x}_j) - \bar{f})^2} \sqrt{\sum_j^J (g(\mathbf{x}_j) - \bar{g})^2}} \right|, \quad (3.2)$$

donde  $J$  es el número total de píxeles en cada imagen, mientras  $\bar{f}$  y  $\bar{g}$  son sus respectivos valores promedio.

Esta medida de similitud es ampliamente usada en Crio-EM debido a que se desempeña mejor que medidas como MSE (error cuadrático medio) y MAE (error absoluto medio) ante cambios de iluminación de imágenes que representan las mismas escenas [80]; esto es relevante para el presente trabajo, debido a que como se presentó en la Sección 2.1.2 las proyecciones experimentales en Crio-EM están afectadas por la Función de Transferencia de Contraste (CTF) del microscopio. Además, debido también a la gran cantidad de comparaciones que se hacen en el proceso de alineamiento entre proyecciones experimentales y aquellas contenidas en la galería de referencia, se busca utilizar una medida de similitud cuyo cálculo se ejecute de forma rápida y que sea robusto ante la presencia de ruido en las imágenes; en cuanto a estas últimas características, el Coeficiente de Correlación de *Pearson* ha mostrado ser más robusto y rápido de calcular que otras medidas que han venido ganando popularidad como el Índice de Similitud Estructural (SSIM, por sus siglas en inglés) [81].

La transformación rígida que permite alinear una proyección  $f$  con otra proyección  $g$  puede ser expresada como  $g \approx M(f)$ . La forma básica de esta transformación es la composición  $M(f) = R_\psi \circ T_t(f)$ , donde  $R_\psi$  y  $T_t$  son transformaciones de rotación y traslación, respectivamente, que dependen del ángulo  $\psi$  y de una traslación bi-dimensional en el plano de proyección; cabe recordar que dicha transformación también se puede representar con la siguiente notación matricial  $M(f) = f(\mathbf{R}_\psi \mathbf{x} - \mathbf{t}) = f(\mathbf{M}\mathbf{x})$ .

Aunque existen varias formas de dar solución al problema de optimización presentado en (3.1) en el presente trabajo se propuso un método en tres etapas. Para cada proyección experimental  $f \in \mathcal{P}$ , la primera etapa consiste en el siguiente procedimiento

$$\mathcal{Q} = \mathbf{FirstApprox}(\mathcal{R}, f, \phi), \quad (3.3)$$

el cual compara la proyección experimental  $f$  con la galería de proyecciones referencia (todas las proyecciones  $g \in \mathcal{R}$ ) para producir una lista  $\mathcal{Q}$  cuyos elementos son los pares  $(g_i, \phi(M_i(f), g_i))$  que cumplen con (3.1); recordando que  $\phi$  es la medida de similitud

presentada en (3.2) y  $M_i$  es la transformación rígida apropiada entre la proyección experimental  $f$  y la proyección de referencia  $g_i$ .

Debido a la cantidad de comparaciones que se llevan a cabo en (3.3) la selección de un método rápido y preciso es clave. Una elección frecuente para llevar a cabo el alineamiento entre dos imágenes es el método de correlación de fase presentado en la Sección 2.4. Según la teoría presentada en la Sección 2.4, el método de correlación de fase permitiría determinar de forma independiente la rotación y traslación relativa entre una proyección experimental y una proyección referencia; sin embargo, en la práctica este método no se puede aplicar de manera directa a las proyecciones de Crio-EM SPA porque, en general, en el esquema de *projection matching* existen pocas proyecciones en la galería de referencia que sean similares a cada proyección experimental, las cuales a su vez cuentan con una muy baja SNR como se ha mencionado antes. Como resultado, esta situación conduce a un alto grado de incertidumbre durante el proceso de alineamiento.

Cuando se aplica directamente el método (2.20) basado en correlación de fase, el cual alinea una proyección experimental  $f$  con una proyección referencia  $g$ , a los datos de SPA para buscar el parámetro de rotación relativo entre cada par experimental-referencia, proporciona un gran número de máximos cuya localización podría ser el valor para alinear rotacionalmente las imágenes, ver Figura 3.1a. Por ello, para determinar el mejor valor entre todos los valores máximos, es necesario, en cada caso, buscar la traslación relativa entre las imágenes, una vez que la rotación candidata ha sido aplicada. Teniendo en cuenta lo anterior, se propuso hacer uso de correlación cruzada en lugar de la correlación de fase para reducir el número de candidatos a mejor rotación entre cada par de imágenes, ver Figura 3.1. En dicha figura se puede observar que el número de candidatos a mejor rotación entre cada par de proyecciones a ser alineadas, es menor cuando se usa correlación cruzada en lugar de correlación de fase, esto debido a la naturaleza ruidosa de las proyecciones en Crio-EM SPA.

Para dos funciones  $f, g : \mathbb{R}^2 \rightarrow \mathbb{R}$ , cuyas transformadas de Fourier las mapea a  $\hat{f}, \hat{g} : \mathbb{R}^2 \rightarrow \mathbb{C}$ , respectivamente (denotadas también como  $\hat{f}(\xi) = \mathcal{F}\{f(x)\}$  y  $\hat{g}(\xi) = \mathcal{F}\{g(x)\}$ ),

el Teorema de Correlación Cruzada establece que

$$C_r(f, g) = (f \otimes g)(\mathbf{x}) = \mathcal{F}^{-1} \left\{ \left( \widehat{f} \times \widehat{g}^* \right) (\boldsymbol{\xi}) \right\}, \quad (3.4)$$

donde  $\widehat{g}^*$  es el complejo conjugado de  $\widehat{g}$  y  $\mathcal{F}^{-1}\{\zeta_r\}$  es la Transformada inversa de Fourier de  $\zeta_r = \left( \widehat{f} \times \widehat{g}^* \right) (\boldsymbol{\xi})$ . Hacemos notar que cuando se use la representación en coordenadas polares, nos referiremos a la correlación cruzada como  $C_a$  y al producto de funciones en el espacio de Fourier de (3.4) como  $\zeta_a$ .

En este trabajo, cuando se buscan los parámetros de alineamiento entre una proyección experimental  $f$  y una proyección referencia  $g$ , primero se busca el ángulo de rotación  $\psi$  para alinearlas rotacionalmente usando (3.4) con la representación en coordenadas polares de la magnitud de los espectros de Fourier de las imágenes (en otras palabras,  $f = \left| \widehat{f}(r, \theta) \right|$  y  $g = \left| \widehat{g}(r, \theta) \right|$ , respectivamente). En el caso descrito,  $C_a$  es un arreglo  $\mathbf{C}$  con dimensiones  $m_a \times n_a$  que son las mismas de  $\left| \widehat{f}(r, \theta) \right|$ , o alternativamente las de  $\left| \widehat{g}(r, \theta) \right|$ , donde las variaciones en cada columna representan cambios en la escala, mientras las variaciones en cada fila representan cambios en los ángulos. Para hallar el ángulo que mejor alinea las dos imágenes se construye un vector  $\boldsymbol{\varpi} \in \mathbb{R}^{n_a}$  el cual sirve para determinar los posibles valores para  $\psi$ . Cada elemento  $\varpi_j$ , para  $1 \leq j \leq n_a$ , del vector  $\boldsymbol{\varpi}$  contiene el promedio de la  $j$ -ésima columna de  $\mathbf{C}$  (o sea,  $\frac{1}{m_a} \sum_{i=1}^{m_a} c_{i,j}$ ); no se procede de manera similar para determinar el parámetro de escala debido a la forma en que son adquiridas las proyecciones experimentales en Crio-EM SPA. Después, se encuentran los dos valores máximos en  $\boldsymbol{\varpi}$  los cuales son asignados a  $\psi_1$  y  $\psi_2$ ; además, debido a la simetría en el espacio de Fourier, se suma  $180^\circ$  a  $\psi_1$  y  $\psi_2$ , lo que conduce a dos posibles valores adicionales  $\psi_3$  y  $\psi_4$ , respectivamente (ver Figura 3.1b para un ejemplo del vector de correlaciones y su comparación con el obtenido usando correlación de fase).

Posteriormente, se calculan cuatro vectores de desplazamiento  $t_j$  aplicando (3.4), entre la proyección experimental  $f(\mathbf{x})$  y una versión rotada de la proyección referencia  $g(\mathbf{R}_{-\psi_j} \mathbf{x})$ , para  $1 \leq j \leq 4$ ; es importante observar que  $g$  se alinea con  $f$  aplicando el negativo de los ángulos  $\psi_j$ . En este caso (3.4) típicamente produce una distribución con un sólo pico significativo el cual es elegido como el desplazamiento para cada ángulo  $\psi_j$ .

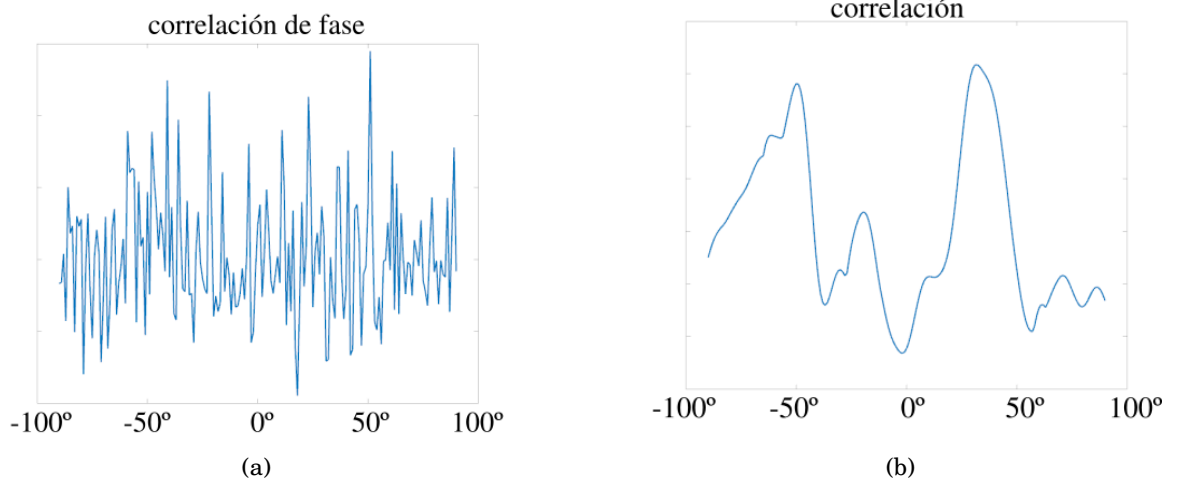


Figura 3.1: Ejemplo de vectores de correlación que usan los enfoques de (a) Transformada de Fase y (b) Correlación cruzada clásica.

El mejor par de valores  $(\psi_j, t_j)$  es el que produce  $\arg \max_{1 \leq j \leq 4} \phi(g(\mathbf{x}), f(\mathbf{R}_{\psi_j} \mathbf{x} + \mathbf{t}_j))$ .

Luego de completar esta primera etapa, es posible conocer los mejores candidatos para proveer información sobre la potencial orientación de proyección a partir de la cual pudo haber sido generada la imagen experimental, seleccionando de entre todas las proyecciones en la galería de referencia aquellas con los valores más altos de  $\phi$  en la lista  $\mathcal{Q}$  producida como resultado de (3.3). Para tal propósito, se eligen todas las proyecciones de referencia  $g \in \mathcal{R}$  que proporcionan un valor  $\phi(g, M_g(f))$  mayor a dos tercios del rango de valores en  $\mathcal{Q}$  y se almacenan en una lista  $\mathcal{O}$ . Dicha lista de proyecciones referencia sirve como entrada al segundo procedimiento

$$\mathcal{U} = \mathbf{SecondApprox}(\mathcal{O}, f, \phi), \quad (3.5)$$

donde  $f$  es la proyección experimental y  $\phi$  es la medida de similitud definida en (3.2). Este procedimiento produce una lista  $\mathcal{U}$  de pares  $(g, \phi(M_g(f), g))$  que cumplen con (3.1), pero a diferencia del primer procedimiento la transformación  $M_g$  ha sido obtenida con el método propuesto en [26]. Este esquema es más intensivo que el usado en la primera etapa debido a que la mejor transformación  $M$  para registrar la proyección experimental  $f$  con cada proyección referencia  $g$ , es elegida entre alguna de las siguientes secuencias o composiciones de transformaciones:  $M'_1 = T_t \circ R_\psi \circ T_t \circ R_\psi \circ T_t \circ R_\psi$  y

$M'_2 = R_\psi \circ T_t \circ R_\psi \circ T_t \circ R_\psi \circ T_t$  que son aplicadas a la proyección experimental  $f$ . Las transformaciones  $T_t$  y  $R_\psi$  son obtenidas usando (3.4), pero en el último caso las proyecciones  $f$  y  $g$  son representadas en coordenadas polares en el espacio real para obtener el correspondiente valor de rotación relativa (esto hace que la búsqueda del parámetro de rotación no sea independiente del desplazamiento o *shift* relativo entre  $f$  y  $g$ ). Luego de obtener las correspondientes transformaciones  $M'_1$  y  $M'_2$ , se calculan imágenes temporales  $f_1 = M'_1(f)$  y  $f_2 = M'_2(f)$  para calcular las medidas de similitud  $\phi(g, f_1)$  y  $\phi(g, f_2)$ , respectivamente. La motivación para llevar a cabo (3.5) para la búsqueda de parámetros es que en el procedimiento (3.3) hay una pérdida de información de la fase de las proyecciones en el espacio de Fourier; sin embargo, sería innecesario llevar a cabo el procedimiento de búsqueda más intensivo presentado en (3.5) con todas las proyecciones en la galería de referencia  $\mathcal{R}$ , ya que como se mencionó antes, en el esquema de *projection matching* son pocas las proyecciones en  $\mathcal{R}$  que realmente se parecen lo suficiente a la proyección experimental  $f$ .

Después de aplicar los procedimientos (3.3) y (3.5) para una proyección experimental  $f$  se obtiene la lista ordenada  $\mathcal{U}$  que contiene proyecciones candidatas para asignar los mejores parámetros de orientación a la proyección  $f$ . Sin embargo, antes de usar los parámetros de proyección correspondientes a la primera proyección  $g$  de la lista  $\mathcal{U}$  se lleva a cabo un procedimiento final, basado en procesamiento de señales soportadas en gráficos, que permite determinar el nivel de confianza de esa posible asignación. Algunas ideas básicas detrás de este enfoque fueron presentadas en la Sección 2.3 de manera que el presente documento pueda ser auto-contenido; aunque para mayor claridad sobre las ideas y conceptos detrás del procesamiento de señales en gráficos se recomienda revisar el Apéndice A y los trabajos presentados en [69, 70, 71, 72].

En este enfoque, cada proyección de referencia  $g_i$  junto con su correspondiente medida de similitud  $\phi$  respecto a la imagen experimental  $f$  luego del proceso de alineamiento ( $\phi(g_i, M_i(f))$ ) representa un nodo  $v_i$  del conjunto que forma el gráfico no dirigido  $G(V, E)$  donde el conjunto de aristas ponderadas y bi-direccionales  $E$ , conecta los nodos en  $V$ . Claramente, esta definición del gráfico permite asociar la función escalar  $\phi : V \rightarrow \mathbb{R}$  que establece el valor para cada nodo  $v_i$ , para  $1 \leq i \leq |V|$  ( $|V|$  es la cardinalidad del

conjunto de nodos), del gráfico  $G$  como  $\phi(v_i) = \phi(g_i, M_i(f))$ ; tal función discreta puede ser representada como  $\Phi = \{\phi(v_1), \dots, \phi(v_{|V|})\}$ .

Los pesos para cada arista  $E$  que conectan los nodos  $v_i$  y  $v_j$  están dados por

$$\mathbf{W}_{i,j} = \begin{cases} e^{-\frac{d(g_i, g_j)}{D_{\max}}}, & \text{if } d(g_i, g_j) < D_{\max}, \\ 0, & \text{otro caso,} \end{cases} \quad (3.6)$$

donde  $g_i$  y  $g_j$  son proyecciones de referencia en  $\mathcal{R}$ ,  $d$  es la distancia angular entre éstas y  $D_{\max}$  es el umbral de distancia predefinido que determina la topología del gráfico  $G$ .

Como se presentó antes en la Sección 2.3 (ver también el Apéndice A), la Transformada de Fourier de la función soportada en el gráfico  $G$  es

$$\hat{\phi}(\lambda_i) = \sum_{j=1}^{|V|} \phi(v_j) u_{i,j}, \quad (3.7)$$

donde  $u_i$ , con  $1 \leq i \leq |V|$ , es uno de los vectores propios de la matriz Laplaciana del gráfico  $G$  (ver (2.12) de la Sección 2.3) y  $\lambda_i$  su respectivo valor propio. Para complementar la definición de la Transformada de Fourier, la definición de la Transformada Inversa de Fourier es

$$\phi(v_i) = \sum_{j=1}^{|V|} \hat{\phi}(\lambda_j) u_{j,i}. \quad (3.8)$$

Es importante señalar que como sucede también con la Transformada de Fourier estándar, en el caso de la GFT se tiene la siguiente identidad de *Parseval*

$$\sum_{i=1}^{|V|} (\phi(v_i))^2 = \sum_{i=1}^{|V|} (\hat{\phi}(\lambda_i))^2, \quad (3.9)$$

la cual relaciona la energía de la señal soportada en el gráfico  $G$  y su respectiva representación en el dominio espectral del gráfico. Lo anterior es relevante a la hora de proponer el filtrado espectral de las señales soportadas en gráficos.

Una forma más intuitiva de entender los elementos que se han presentado es la siguiente. El gráfico  $G$  representa las medidas de similitud  $\phi$ , calculadas con (3.2), de la proyección experimental  $f$  respecto a todas las proyecciones de referencia en  $\mathcal{R}$

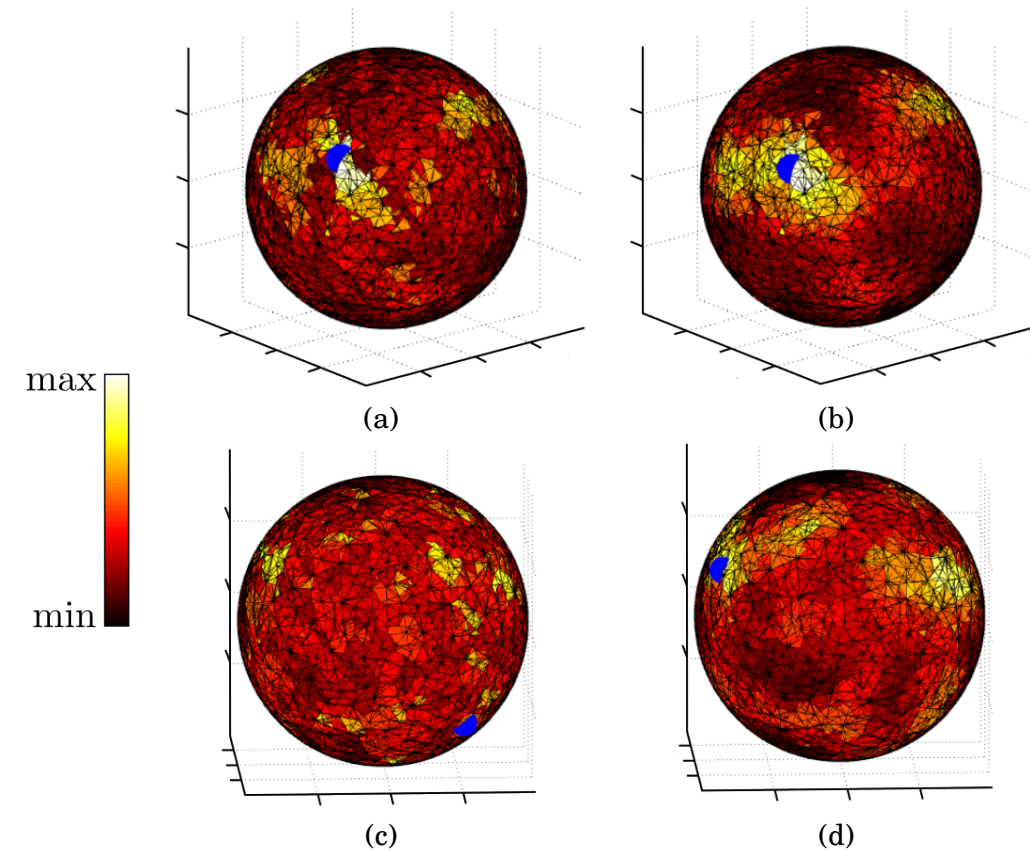


Figura 3.2: (*columna izquierda*) Representación gráfica de las medidas de similitud entre una imagen experimental  $f$  y todas la proyecciones referencia en  $\mathcal{R}$  calculadas con (3.2) (es decir, la función  $\Phi$ ) y (*columna derecha*) la correspondiente versión pasa-bajas (es decir, la función  $\Phi'$ ) calculadas con el enfoque de procesamiento de señales soportadas en gráficos. La barra de colores en la izquierda muestra el rango de valores de la medidas de similitud. Los puntos azules muestran el máximo valor en los gráficos, la distancia entre el punto azul en el gráfico original y el que se encuentra en su respectiva versión filtrada proporciona un criterio de confianza durante el proceso de asignación angular. De manera que la *fila superior* muestra un ejemplo en el que hay mayor confianza en la asignación que aquel representado en la *fila inferior*.



luego del correspondiente alineamiento. Adicionalmente, esta representación también contiene la relación espacial entre dichas proyecciones de referencia. En el caso ideal, cuando las proyecciones experimentales no están contaminadas con ruido, el gráfico debería poseer una sola región donde se encuentra un valor máximo de  $\Phi$  rodeado de una distribución que decae de dicho valor máximo a cero. En este escenario sería fácil identificar la proyección referencia  $g$  cuya orientación debería ser asignada a la proyección experimental  $f$ . Sin embargo, la naturaleza ruidosa de las proyecciones experimentales se refleja en la posible aparición de regiones que tienen valores altos de  $\phi$  rodeados de valores bajos (en este caso, correlaciones espurias luego de alinear  $f$  contra la proyección referencia  $g$  ubicada en esa dirección). Para encontrar la región “suave” con valores altos de  $\phi$  se deben “aplanar” aquellas regiones espurias en la función  $\Phi$ . El aplanamiento se logra usando la Transformada de Fourier de (3.7) y seleccionando los coeficientes con menor variación, aquellos que contienen el 95 % de la energía espectral de  $\Phi$ , para calcular la Transformada Inversa de Fourier con (3.8) y obtener una versión suavizada de la función  $\Phi$  (o sea, la función  $\Phi'$ ). La función  $\Phi'$  sirve para revisar si la dirección de proyección a partir de la cual fue generada  $g$  (aquella que luego de las dos primeras etapas de alineamiento proporciona el mayor valor de similitud  $\phi(g, M(f))$ ) se ubica en la región donde se encuentra el pico de valores de similitud; este comportamiento es un indicador de confianza de la asignación de la información de proyección de  $g$  a la proyección experimental  $f$ . La idea detrás de este enfoque es que la información espacial de las proyecciones de referencia en  $\mathcal{R}$  es relevante cuando se calculan los valores de similitud, usando (3.2), entre las proyecciones en el conjunto  $\mathcal{P}$  y las proyecciones alineadas en  $\mathcal{R}$ , porque estos valores corresponden en si mismos a una variable aleatoria debido a la naturaleza ruidosa de las proyecciones experimentales (ver ejemplo en la Figura 3.2).

Cabe anotar que este método de procesamiento de señales en gráficos puede ser usado para medir la calidad de la asignación angular llevada a cabo por cualquier método y compararla contra la realizada por el método propuesto en este trabajo. Esto se realiza comparando la asignación hecha por cualquier otro método con la asignación que lleva a cabo el método propuesto en este trabajo, considerando, como se mostrará en el

siguiente capítulo, que este último se desempeña de forma similar, como herramienta de asignación angular, a otros métodos ampliamente usados en el campo de procesamiento de imágenes en Crio-EM. De esta manera, dada la asignación angular hecha a una proyección experimental  $f$  usando un método A y por el método propuesto en este trabajo, p. e. método R, se puede proporcionar información sobre la discrepancia entre la orientación proporcionada por los métodos A y R calculando la correlación ( $\varepsilon_{ref}$ ) entre las proyecciones de referencia asociadas a  $f$  por ambos métodos. También, es posible medir la distancia ( $d_{peak}$ ) de la proyección de referencia  $g_A$ , asignada a la proyección experimental  $f$  por el método A, hasta la región del pico de similitudes de la versión suavizada de correlaciones de Pearson indexadas a través del gráfico  $G$  (es decir, la función  $\Phi'$ ). Finalmente, el método propuesto en este trabajo puede proporcionar también una medida de similitud,  $\gamma_{ori}$ , entre las proyecciones de referencia que corresponden a las orientaciones asignadas por el método A y las proyecciones de referencia asociadas a las orientaciones que corresponden a los valores más altos del gráfico suavizado  $\Phi'$ .

### 3.1. Resumen del enfoque propuesto

Un resumen de la herramienta propuesta se puede encontrar en la Figura 3.3 y el correspondiente algoritmo en el Apéndice B. El método que se propone y se ilustra en la Figura 3.3 recibe una proyección de entrada  $f$ , ya sea una proyección experimental o *class-average*, que se compara, vía correlación de Pearson (Ecuación (3.2)), con cada una de las proyecciones referencia  $g$  en el conjunto  $\mathcal{R}$  (galería de proyecciones de referencia generadas a partir del modelo inicial). El objetivo es asignar una dirección de proyección a partir de la cual pudo haber sido generada la proyección experimental  $f$  durante el proceso de adquisición, dado el conjunto  $\mathcal{R}$  y sus respectivas direcciones de proyección asociadas; sin embargo, antes de asignar una dirección de proyección a cada  $f$  es necesario registrarla con cada  $g_i \in \mathcal{R}$ . En el primer ciclo de búsqueda se calcula un conjunto de parámetros de alineamiento entre  $f$  y cada  $g_i \in \mathcal{R}$ , así como su respectiva medida de similitud usando (3.2) luego de ser alineadas usando la transformación  $M_i$ . Posteriormente, la proyección experimental es sometida a un segundo ciclo de búsqueda

de parámetros, pero esta vez solo con respecto a las proyecciones  $g_i \in \mathcal{R}$  cuyos valores de  $\phi$ , calculados por (3.2), sean altos después del primer ciclo de búsqueda. Este segundo ciclo de búsqueda es un “refinamiento” de los parámetros de alineamiento entre cada par de proyecciones, como se mencionó antes. Finalmente, la dirección de proyección asignada a la proyección experimental  $f$  es aquella que corresponde a la proyección de referencia  $g_i$  con respecto a la cual se obtuvo el mejor valor de  $\phi(g_i, M_i(f))$  luego del alineamiento durante los dos ciclos de búsqueda antes mencionados. Adicionalmente, usando la función  $\Phi$ , el entorno de valores para  $\phi$  indexados a través del gráfico  $G$  que resultan de comparar la proyección experimental  $f$  con todas las  $g_i \in \mathcal{R}$ , así como su versión filtrada (usando el enfoque espectral basado en gráficos), se calculan varios parámetros que permiten validar la confiabilidad de la asignación angular hecha a cada proyección experimental  $f$  (ver ejemplo en la Figura 3.2).

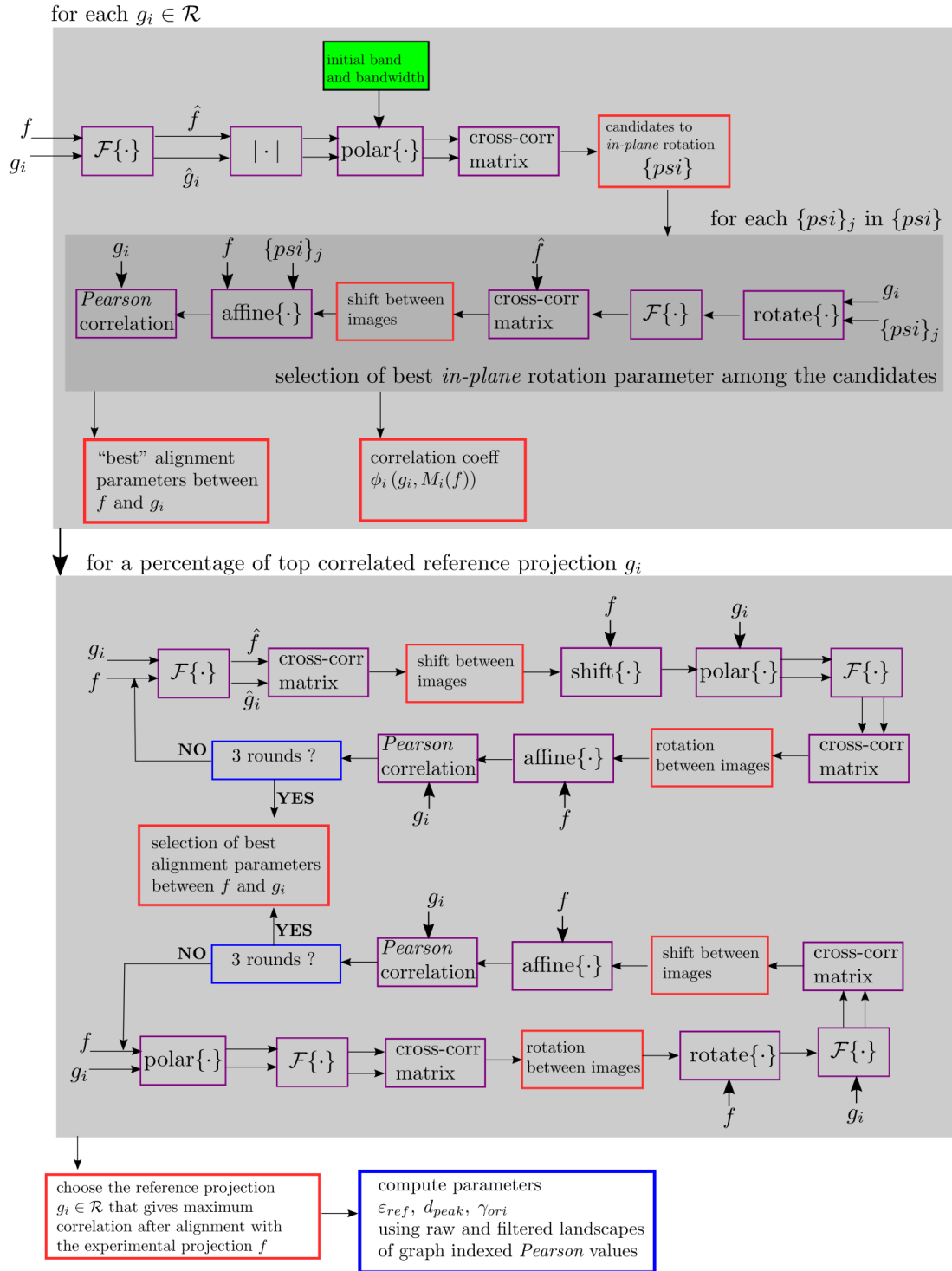


Figura 3.3: Resumen del método propuesto en este trabajo para llevar a cabo asignación angular.

## Capítulo 4

# Resultados y Discusión

En este capítulo se reportan los resultados de los experimentos que demuestran el desempeño del método propuesto para llevar a cabo la validación de asignaciones angulares en la etapa de refinamiento. El primer experimento pretende mostrar únicamente el desempeño de la herramienta de asignación angular usando datos sintéticos mientras el segundo muestra el desempeño como alternativa para medir la confianza en las asignaciones angulares hechas, ya sea que éstas hayan sido dadas con la herramienta propuesta o con otro método.

Para los experimentos, nuestro método fue implementado dentro de la infraestructura del paquete para procesamiento de imágenes de Crio-EM conocido como Scipion [59] usando también las bibliotecas de Xmipp [82] (ambos paquetes pueden ser obtenidos desde [83] y [84], respectivamente). Uno de los grandes beneficios de Scipion es que permite incorporar programas y bibliotecas de varios grupos como *plugins* lo cual permite a los usuarios utilizar métodos variados para producir un mapa 3D y a los desarrolladores comparar los diferentes métodos y sus implementaciones. Para la siguiente presentación es importante mencionar que un protocolo es una secuencia de diferentes métodos implementados como programas independientes. La herramienta propuesta en esta tesis para llevar a cabo asignación angular se incluyó en el protocolo de refinamiento conocido como Highres en Scipion como alternativa al método que este protocolo tiene por defecto, conocido como Significant [26]. Por otra parte, la implementación del procedimiento de validación (el tercer procedimiento descrito en el Capítulo 3), fue implementada como

un protocolo independiente dentro del entorno de Scipion de manera que éste pudiera recibir un conjunto de proyecciones y validar su asignación angular.

#### 4.1. Desempeño de la herramienta de asignación angular

Primero se describe el experimento con datos sintéticos para evaluar el método propuesto y su funcionamiento como herramienta para hacer asignación angular, para esto se usaron proyecciones ruidosas sintéticas generadas desde un mapa 3D de ribosoma [85]; dicho mapa está disponible dentro de los datos de prueba de Scipion para el *plugin* de Relion [56], la cual es una de las herramientas disponibles para el refinamiento de biomoléculas en Crio-EM y una de las más utilizadas en la comunidad.

El conjunto de proyecciones de referencia  $\mathcal{R}$  para este experimento consta de 4,412 proyecciones sin ruido del mapa 3D representando el ribosoma; las proyecciones de este conjunto cubren la esfera de proyección de la manera más homogénea posible (un método de muestreo que usa una rejilla triangular basada en el trabajo presentado en [86]) resultando en una separación de  $\sim 3.85^\circ$  entre cada par de proyecciones. Todas las proyecciones del conjunto tienen dimensiones  $128 \times 128$  y una resolución de  $3.3 \text{ \AA}/\text{píx}$ . Para crear el conjunto sintético de proyecciones experimentales  $\mathcal{P}$ , se seleccionaron de forma aleatoria 1,000 proyecciones del conjunto  $\mathcal{R}$  y se les aplicó una transformación rígida  $T_t \circ R_\theta(g)$ . El valor de  $\theta$  fue elegido de una distribución uniforme continua con límites  $[-179, 180]$  mientras que los elementos del vector de desplazamiento  $t$  fueron elegidos de una distribución uniforme continua con límites  $[-12.8, 12.8]$  (esto es,  $\pm 10\%$  del tamaño de las proyecciones); este rango para el desplazamiento fue elegido porque son los valores reportados para conjuntos reales de proyecciones en Crio-EM después de las etapas de selección de partículas y pre-procesamiento. Finalmente, cada proyección  $f \in \mathcal{P}$  fue corrompida con ruido cuya función de densidad es una distribución Normal con media nula y cuatro diferentes varianzas  $\sigma_1^2$ ,  $\sigma_2^2$ ,  $\sigma_3^2$  y  $\sigma_4^2$  de forma tal que la SNR tuviera valores promedio de -7.53 dB, -10.24 dB, -12.36 dB, y -15.06 dB, respectivamente; de esta manera, se construyeron cuatro conjuntos  $\mathcal{P}_{\sigma_1}$ ,  $\mathcal{P}_{\sigma_2}$ ,  $\mathcal{P}_{\sigma_3}$  y  $\mathcal{P}_{\sigma_4}$  de 1,000 proyecciones experimentales creadas de forma sintética.

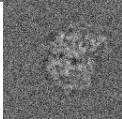
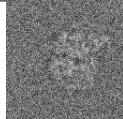
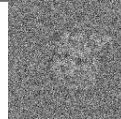
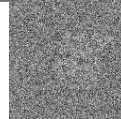
		SNR [dB]			
		-7.53	-10.24	-12.36	-15.06
					
		$\sigma_1$	$\sigma_2$	$\sigma_3$	$\sigma_4$
Propuesto	Exactitud [%]	99.00	98.00	97.50	89.40
	Dif. angular [deg]	1.21±0.73	1.25±0.74	1.33±0.77	1.69±0.97
	Dif. <i>shift</i> [píx]	0.29±0.13	0.30±0.14	0.32±0.16	0.47±0.24
Relion	Exactitud [%]	99.80	99.10	98.50	97.30
	Dif. angular [deg]	0.69±0.29	0.61±0.26	1.12±0.34	1.27±0.47
	Dif. <i>shift</i> [píx]	0.14±0.08	0.15±0.07	0.20±0.10	0.30±0.15
Highres	Exactitud [%]	98.70	97.30	96.70	94.50
	Dif. angular [deg]	1.23±0.75	1.29±0.76	1.38±0.83	1.49±0.86
	Dif. <i>shift</i> [píx]	0.35±0.16	0.36±0.17	0.38±0.18	0.40±0.20

Tabla 4.1: Resultados de exactitud y diferencias entre parámetros reales y estimados producidos por los métodos Relion, Highres y el propuesto en esta tesis para la clasificación de orientación de proyecciones creadas de forma sintética a partir de proyecciones de un mapa 3D de ribosoma; se usaron 4 diferentes niveles de ruido  $\sigma_1$ ,  $\sigma_2$ ,  $\sigma_3$ , y  $\sigma_4$ .

Las proyecciones de referencia en  $\mathcal{R}$  y los cuatro conjuntos de proyecciones experimentales sintéticas  $\mathcal{P}_{\sigma_1}$  a  $\mathcal{P}_{\sigma_4}$ , fueron usados como entrada al método de asignación angular propuesto (ver Capítulo 3). Durante el procesamiento de los conjuntos experimentales fueron calculadas las diferencias entre la orientación y parámetros de alineamiento estimados y los valores reales usados para crear las proyecciones experimentales sintéticas.

Los resultados en la Tabla 4.1 muestran que el método propuesto en esta tesis alcanza una exactitud comparable con la alcanzada por Relion y Highres detectando la proyección de referencia, o alguna de sus vecinas, usada para generar la proyección experimental sintética; aunque se destaca el menor desempeño alcanzado en el caso de los datos con mayor cantidad de ruido. Además, para las proyecciones sintéticas cuyas orientaciones fueron estimadas de manera correcta, la diferencia entre los parámetros de alineamiento estimados y la transformación aplicada a la hora de generarlas es también similar a las producidas por estos métodos. Finalmente, el tiempo de ejecución promedio para el método propuesto fue de 1.1 horas, mientras Relion y Highres tardaron 0.6 y 4.2 horas, respectivamente; estos valores de tiempo son para arreglos de ejecución similares (4 procesos paralelos usando el protocolo de comunicación MPI - *Message Passing Interface*) en el mismo computador.

Es importante mencionar en este punto que si bien la herramienta de asignación angular no superó en robustez frente al ruido y en tiempo de ejecución a los otros métodos comparados, en todos los escenarios simulados, sí mostró una consistencia suficiente para que pueda ser usada como parte del proceso de validación, ya que normalmente el proceso de validación de asignación angular en estos casos se hace en media-baja y baja resolución.

## 4.2. Desempeño con proyecciones reales

Adicionalmente, se llevaron a cabo otros tres experimentos usando ahora datos reales, es decir, datos adquiridos durante un proceso real de Crio-EM SPA; de manera similar al experimento anterior, se compararon los resultados de la herramienta



propuesta con los producidos por los algoritmos Relion y Highres.

El primer conjunto de datos proviene de un virus conocido como BVM (*Brome Mosaic Virus*) [87, 88], el segundo proviene de una macromolécula conocida como  $\beta$ -galactosidasa ( $\beta$ -gal) [89, 90] y el tercer conjunto de datos proviene de la apoferritina cuyas proyecciones experimentales fueron adquiridas en las instalaciones del Centro Nacional de Biotecnología en Madrid, España (colaboradores en el desarrollo del presente trabajo). Las proyecciones del BMV fueron adquiridas con un voltaje de 300 kV, magnificación de  $50,000\times$ , aberración esférica de 4.1 mm, contraste de amplitud de 0.1 y un tamaño de píxel de  $0.99 \text{ \AA}/\text{píx}$ . Las proyecciones para el  $\beta$ -gal fueron adquiridas con un voltaje de 300 kV, magnificación de  $50,000\times$ , aberración esférica de 2.7 mm, contraste de amplitud de 0.1 y un tamaño de píxel de  $0.32 \text{ \AA}/\text{pix}$ . Mientras las proyecciones de la apoferritina fueron adquiridas con un voltaje de 300 kV, magnificación de  $50,000\times$ , aberración esférica de 2.7 mm, contraste de amplitud de 0.1 y un tamaño de píxel de  $0.94 \text{ \AA}/\text{píx}$ . Para todos los experimentos fueron usadas proyecciones cuyas CTFs son normales (p. e., sin defectos de astigmatismo); resultando en 21,244 proyecciones para el BMV, 10,000 proyecciones para la  $\beta$ -gal y 46,182 proyecciones en el caso de la apoferritina. Estas proyecciones fueron procesadas para normalizar sus valores de píxel (sustrayendo una función rampa y garantizando valores de fondo con una distribución normal de media cero y desviación estándar unitaria), también fueron removidos valores inusualmente altos y reemplazados por valores de una distribución normal de media cero y desviación estándar unitaria (técnica conocida como *dust removal*) y finalmente se les realizó un proceso de *phase-flipping* (inversión de fase a las frecuencias que generan contraste invertido en las proyecciones según la CTF estimada, ver la Sección 2.1.2).

El mapa inicial para generar las proyecciones de referencia  $\mathcal{R}$  para el experimento con el BVM fue creado usando Significant como se presentó en [26] con el conjunto de proyecciones *class-average* de “buena calidad” según el método de clasificación CL2D [55] que usa un método de *clustering* para dividir las proyecciones en un número predeterminado de clases. De forma similar, CL2D fue usado para generar un “buen conjunto” de proyecciones *class-average* para obtener el mapa inicial utilizado en los proyectos de la  $\beta$ -gal y la apoferritina, pero en el caso de la  $\beta$ -gal la reconstrucción del

mapa inicial fue llevada a cabo usando Relion [14, 56] mientras que en el caso de la apoferritina la reconstrucción fue llevada a cabo usando cryoSparc [91].

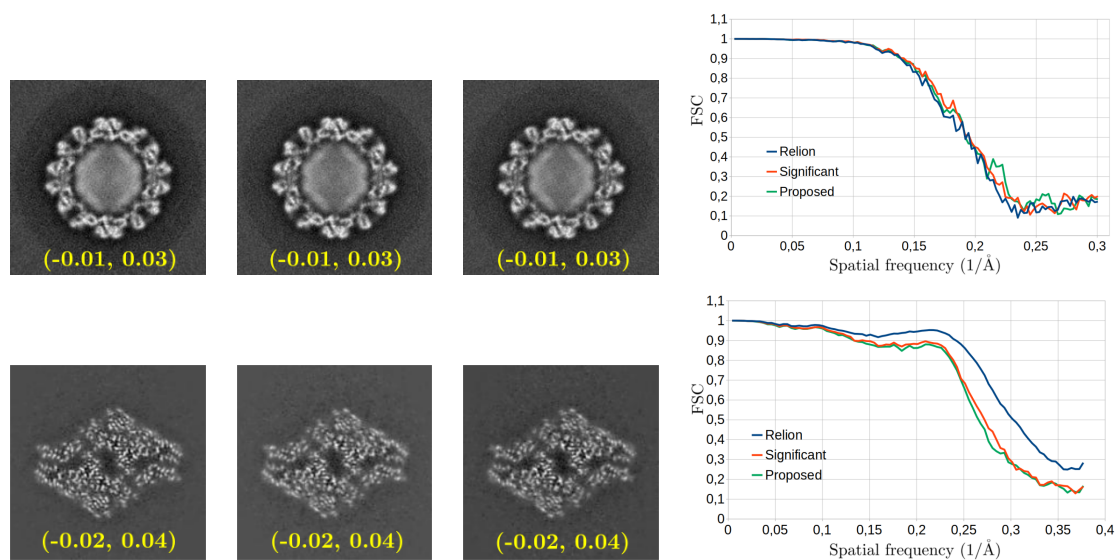


Figura 4.1: Secciones centrales de reconstrucciones de alta resolución producidas por refinamiento de mapas usando (de izquierda a derecha) Relion, Highres y el método propuesto para (*fila superior*) los datos de BMV y (*fila inferior*)  $\beta$ -gal, respectivamente. En amarillo está el rango de niveles de gris usado para desplegar las imágenes. Adicionalmente, en el extremo derecho, se presentan las curvas FSC que dan idea de la resolución alcanzada por cada método.

De la misma forma que en la sección anterior, se usó para los tres experimentos el método de [86] para generar el conjunto de proyecciones referencia  $\mathcal{R}$ , el cual garantiza una distancia angular de aproximadamente  $3.85^\circ$  entre cada par de proyecciones, aunque el número de proyecciones es diferente en cada experimento porque este número depende de la simetría de la biomolécula. Para el caso del BMV, el cual tiene simetría icosaédrica, cada proyección tiene 60 vistas equivalentes de manera que el número total de elementos en  $\mathcal{R}$  se reduce de 4,412 a 74 proyecciones referencia; por otra parte, en el caso de la  $\beta$ -gal, que tiene simetría diédrica, hay 4 vistas equivalentes por cada proyección y el número de elementos en  $\mathcal{R}$  se reduce a 1,103 proyecciones referencia. Finalmente, para el caso de la apoferritina que tiene simetría octaédrica el conjunto  $\mathcal{R}$  queda con 183 proyecciones referencia.

Los conjuntos de proyecciones experimentales junto con sus correspondientes galerías de referencia fueron procesados con Relion y Highres dentro de Scipion para

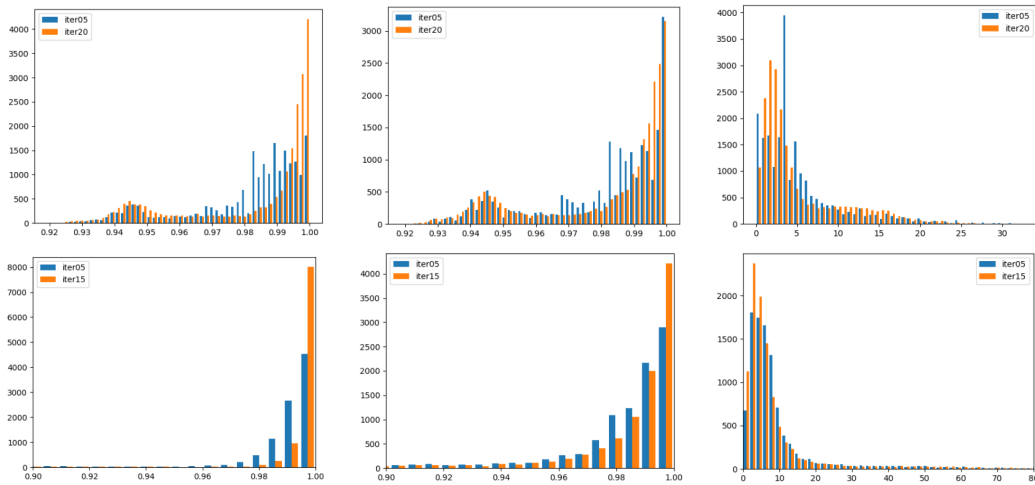


Figura 4.2: Histogramas de las medidas (izquierda)  $\epsilon_{ref}$ , (centro)  $\gamma_{ori}$ , y (derecha)  $d_{peak}$  producidas en diferentes iteraciones durante la asignación angular de Relion para las proyecciones experimentales de (fila superior) el BMV y (fila inferior) la  $\beta$ -gal. Los histogramas de color azul fueron producidos en la 5<sup>a</sup> iteración para los dos conjuntos de datos mientras los histogramas de color naranja fueron producidos en la 20<sup>a</sup> iteración para el BMV y en la 15<sup>a</sup> iteración para la  $\beta$ -gal.

producir los correspondientes mapas 3D. En la Figura 4.1 se muestran secciones centrales de los mapas 3D obtenidos con los tres métodos y se puede observar que las imágenes presentan características similares, por lo menos visualmente con la ventana de visualización seleccionada manualmente. Además, para tener un panorama menos “subjetivo” del desempeño de los tres métodos se incluyen en esta misma figura las curvas FSC producidas por cada uno de los métodos; estas curvas muestran que los tres métodos producen resoluciones similares después de su respectiva asignación angular.

Sin embargo, para ganar información adicional acerca de la calidad de la asignación de orientaciones por los tres métodos, se aplicó el procedimiento, basado en GFT, que permite determinar el nivel de confianza de la asignación angular efectuada por algún método y que fue presentado en el capítulo anterior. Las medidas  $\epsilon_{ref}$ ,  $d_{peak}$  y  $\gamma_{ori}$  proporcionadas por nuestro método basado en teoría de gráficos permiten evaluar la calidad de las asignaciones llevadas a cabo por algún método. La Figura 4.2 muestra un ejemplo de cómo se comportan esas medidas cuando se usan para validar las asignaciones angulares hechas por otros métodos. Se pueden ver los histogramas correspondientes a esas medidas obtenidas para dos iteraciones diferentes durante la

asignación hecha por Relion a las proyecciones experimental es del BMV y la  $\beta$ -gal. Estos histogramas demuestran que valores más altos para  $\varepsilon_{ref}$  y  $\gamma_{ori}$  y valores menores de  $d_{peak}$  proporcionan mejor indicación de calidad en la asignación de orientaciones hecha por algún método. Para decidir qué proyecciones experimentales corresponden a los valores más altos de  $\varepsilon_{ref}$  y  $\gamma_{ori}$  y los valores más bajos para  $d_{peak}$  a partir de las distribuciones de estas medidas, se pueden obtener umbrales que indiquen en dónde se encuentran estos rangos de valores; dichos umbrales pueden ser obtenidos con un método como el propuesto por Otsu [92]. De esta manera, las proyecciones experimentales por fuera del “buen rango” de calidad son aquellas que no contribuyen de manera significativa en el mapa 3D final; estas proyecciones pueden ser consideradas como potencialmente mal alineadas. Para probar esta hipótesis se seleccionaron las proyecciones en los rangos antes mencionados y se almacenaron en un conjunto C, el resto de proyecciones se ubicaron en un conjunto W. Debido a que estos conjuntos no tienen la misma cantidad de elementos, se igualan eligiendo aleatoriamente el número adecuado de proyecciones del conjunto más grande. Finalmente, ambos conjuntos de proyecciones son usados para producir las curvas FSC que a su vez pueden ser utilizadas para comparar la resolución alcanzada por los mapas 3D correspondientes.

Para examinar el desempeño del método que hemos propuesto como herramienta de validación de asignaciones angulares previamente hechas por otros programas, en este caso por Relion y Highres, utilizamos los conjuntos de datos disponibles; cada conjunto de proyecciones experimentales con su respectiva asignación angular fueron divididos en el conjunto C de proyecciones “bien” asignadas y el conjunto complementario W de proyecciones “no-significativas” usando el criterio de medidas  $\varepsilon_{ref}$ ,  $\gamma_{ori}$  y  $d_{peak}$  antes mencionado.

En la Figura 4.3 se presentan los resultados obtenidos con las proyecciones experimentales del BMV. La Figura 4.3a muestra, con diferentes colores, las curvas FSC de los mapas 3D producidos con Relion cuando se usaron las proyecciones clasificadas en los conjuntos C y W para la asignación angular en las 10<sup>a</sup> y 15<sup>a</sup> iteraciones, respectivamente. La Figura 4.3b muestra, con diferentes colores, las curvas FSC de los mapas 3D producidos con Highres usando las proyecciones experimentales

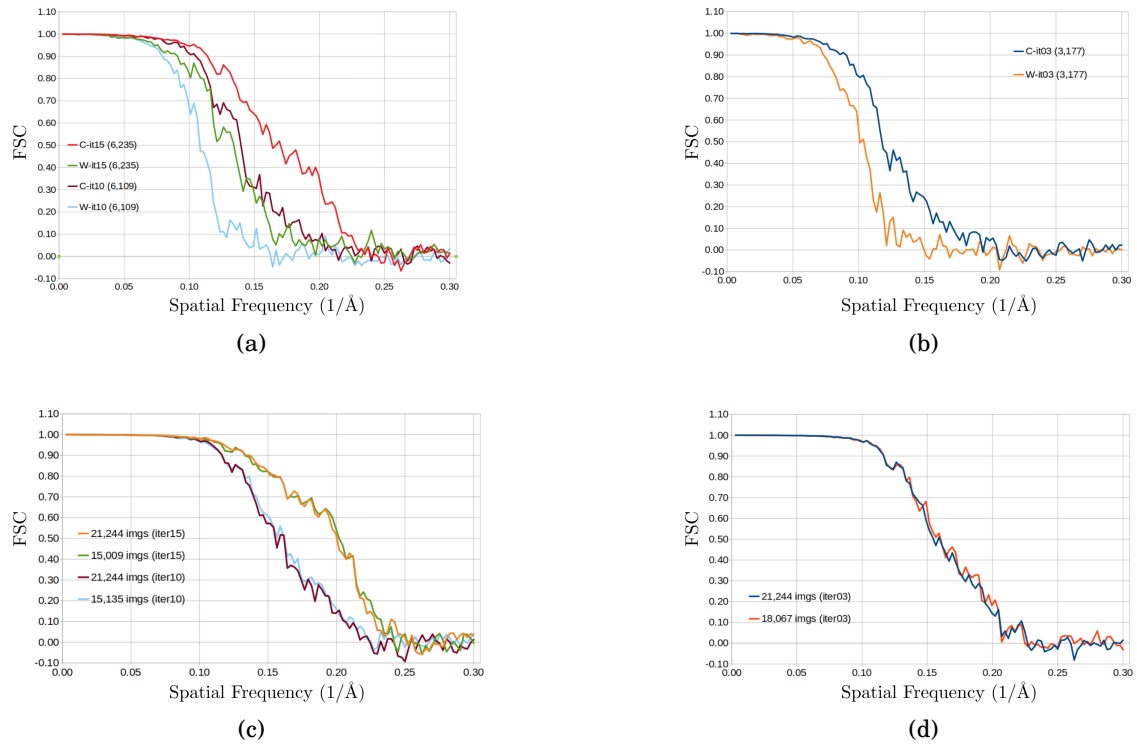


Figura 4.3: Diferentes curvas FSC que corresponden a los mapas 3D producidos con varios métodos usando las proyecciones experimentales del BMV. Cada imagen en la fila superior muestra las FSC correspondientes a mapas 3D obtenidos con proyecciones en los conjuntos C y W usando (a) Relion hasta la 10<sup>a</sup> y 15<sup>a</sup> iteración, respectivamente, y (b) Highres hasta la 3<sup>a</sup> iteración; para todas las curvas el número de proyecciones en W es el valor en paréntesis. Adicionalmente, cada imagen en la fila inferior presenta las curvas FSC correspondientes a los mapas 3D obtenidos con las proyecciones en los conjuntos C y  $\mathcal{R}$  usando (c) Relion para diferentes iteraciones (mostradas en paréntesis) y (d) Highres hasta la 3<sup>a</sup> iteración.

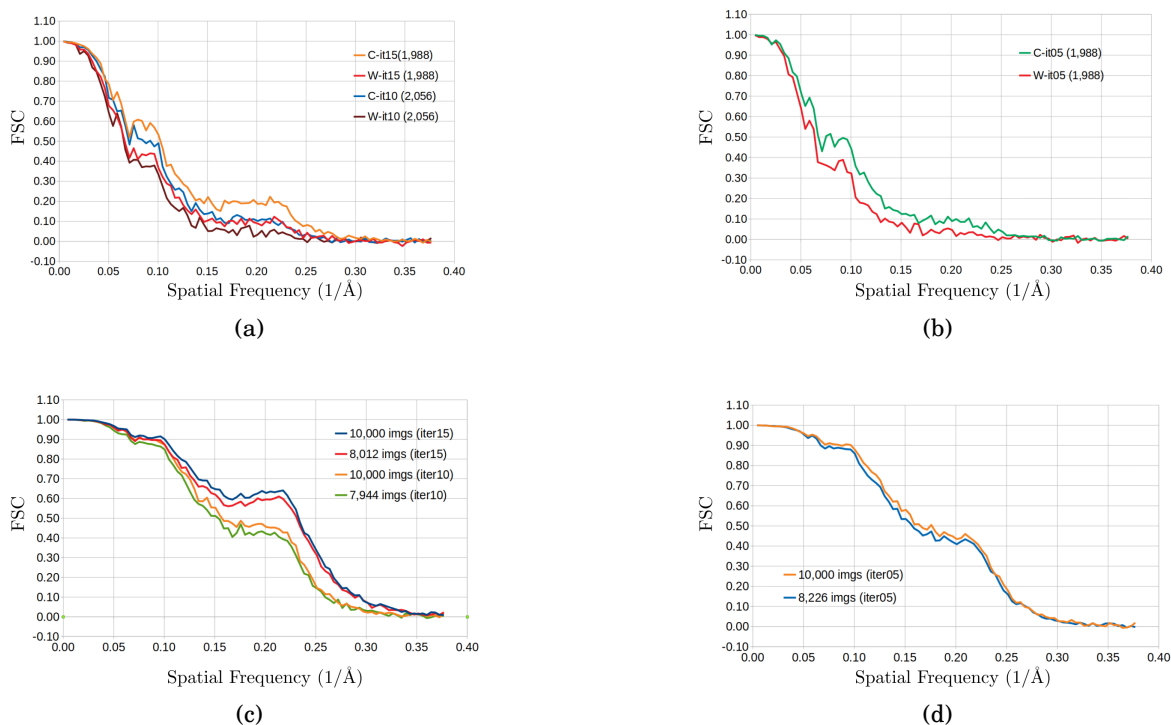


Figura 4.4: Diferentes curvas FSC asociadas a los mapas 3D producidos con varios métodos usando las proyecciones experimentales de la  $\beta$ -gal. Cada imagen muestra la FSC que corresponde a los mapas obtenidos con las proyecciones en los conjuntos C y W usando (a) Relion hasta las iteraciones 10<sup>a</sup> y 15<sup>a</sup> iteración, respectivamente; y (b) Highres hasta la 5<sup>a</sup> iteración; para todas las curvas el número de proyecciones en W es el valor en paréntesis. Adicionalmente, cada imagen en la fila inferior presenta las curvas FSC correspondientes a los mapas 3D obtenidos con las proyecciones en los conjuntos C y  $\mathcal{R}$  usando (c) Relion para diferentes iteraciones (mostradas en paréntesis) y (d) Highres hasta la 5<sup>a</sup> iteración.

en los conjuntos C y W. Estas figuras muestran en todos los casos que los mapas 3D producidos con las proyecciones clasificadas en los conjuntos C tienen mejor resolución que sus contrapartes producidas con las proyecciones en los conjuntos W. Para tener comparaciones adicionales, se presentan en las Figuras 4.3c y 4.3d las curvas FSC que corresponden a los mapas 3D reconstruidos con todas las proyecciones experimentales en el conjunto  $\mathcal{P}$  después de la asignación angular junto con las curvas FSC correspondientes a los mapas 3D reconstruidos con las proyecciones experimentales asociadas al conjunto C por el método de validación propuesto. La Figura 4.3c corresponde a las curvas FSC asociadas a los mapas 3D producidos por Relion usando las proyecciones experimentales asignadas en las 10<sup>a</sup> y 15<sup>a</sup> iteraciones. Por otra parte, la Figura 4.3d corresponde a las curvas FSC asociadas a los mapas 3D producidos con Highres.

Para la  $\beta$ -gal y la apoferritina se repitió el experimento presentado en la Figura 4.3 y se presentan los resultados en las Figuras 4.4 y 4.5, respectivamente. De forma similar a los experimentos con el BMV, los resultados para la  $\beta$ -gal y la apoferritina también sugieren en cada caso que los mapas producidos con las proyecciones y sus respectivas asignaciones angulares clasificadas en los conjuntos C tiene una mejor resolución que sus contrapartes producidas con las proyecciones en los conjuntos W.

Finalmente, de manera que se pueda observar más claramente la utilidad del método propuesto como herramienta de validación, se incluyen los resultados para un experimento con el BMV en el cual se estudia el efecto de incluir proyecciones identificadas como potencialmente mal alineadas durante la reconstrucción de un mapa 3D. Para esto, se usó el mapa de alta resolución EMD-6000 del virus en cuestión, obtenido del EMDB (siglas del banco de datos de microscopía electrónica) [88, 93], como referencia para propósitos de comparación. El método de validación propuesto identificó 1,383 proyecciones experimentales como potencialmente mal alineadas en el conjunto de datos del BMV; basados en este resultado, se construyeron dos conjuntos de proyecciones de igual tamaño, el conjunto C que contiene solo proyecciones identificadas como correctamente asignadas y el conjunto W conteniendo porcentajes complementarios de proyecciones correcta e incorrectamente asignadas. Luego, estos

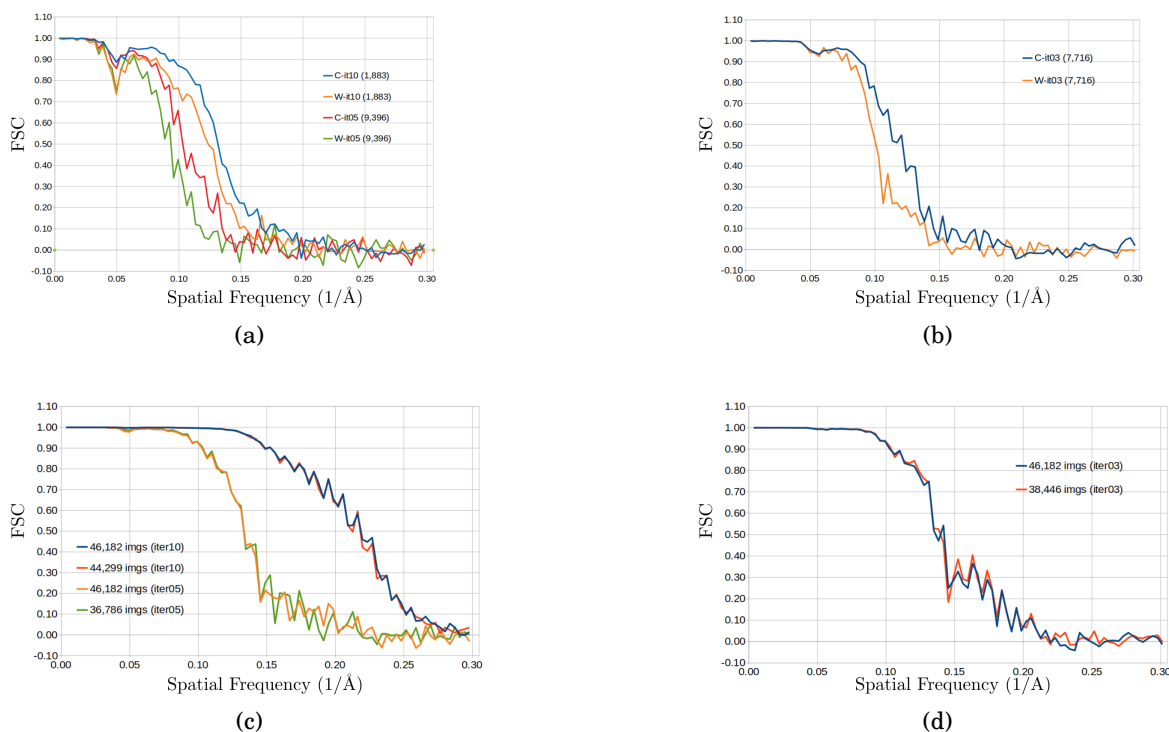


Figura 4.5: Diferentes curvas FSC asociadas a los mapas 3D producidos con varios métodos usando las proyecciones experimentales de la apoferritina. Cada imagen muestra la FSC que corresponde a los mapas obtenidos con las proyecciones en los conjuntos C y W usando (a) Relion hasta las iteraciones 5<sup>a</sup> y 10<sup>a</sup> iteración, respectivamente; y (b) Highres hasta la 3<sup>a</sup> iteración; para todas las curvas el número de proyecciones en W es el valor en paréntesis. Adicionalmente, cada imagen en la fila inferior presenta las curvas FSC correspondientes a los mapas 3D obtenidos con las proyecciones en los conjuntos C y  $\mathcal{R}$  usando (c) Relion para diferentes iteraciones (mostradas en paréntesis) y (d) Highres hasta la 3<sup>a</sup> iteración.



conjuntos de proyecciones y sus respectivas asignaciones angulares fueron usados para reconstruir dos mapas a ser comparados contra el mapa de referencia de alta resolución antes mencionado. Este experimento se llevó a cabo cuando el segundo conjunto de datos contiene 50 %, 30 % y 10 % de proyecciones identificadas como mal alineadas, respectivamente. Los resultados de este experimento se reportan en la Figura 4.6 donde se puede ver que a mayor porcentaje de proyecciones potencialmente mal alineadas en un conjunto, mayor es la diferencia entre las curvas FSC.

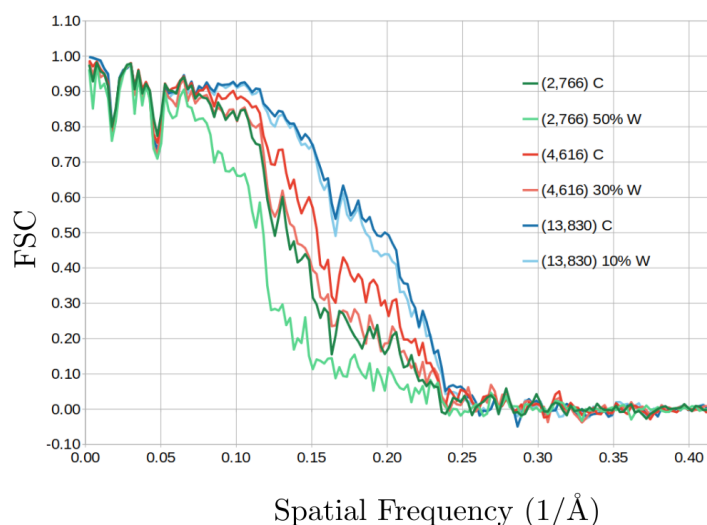


Figura 4.6: Comparación de curvas FSC producidas cuando se compara el mapa 3D de alta resolución del BMV (EMD-6000 del EMDB) contra los mapas obtenidos de los dos conjuntos de datos C y W, cuando W contiene (verde) 50 %, (rojo) 30 %, y (azul) 10 % de proyecciones identificadas como mal alineadas, respectivamente.

### 4.3. Discusión

La calidad de un mapa 3D reconstruido se puede ver afectada de manera negativa, debido a la posibilidad de cometer errores durante el proceso de refinamiento cuando se usa el enfoque de *projection matching*; estos errores se deben principalmente a la falta de exactitud en la estimación de la dirección de proyección (derivada de los errores en el alineamiento entre la proyección experimental y las proyecciones de referencia), así como a la presencia de heterogeneidad (ya sea estructural o composicional) en las proyecciones involucradas en el proceso las cuales no fueron detectadas durante el

proceso previo de clasificación. Adicionalmente, la falta de exactitud en la asignación de orientaciones a las proyecciones experimentales puede ser persistente como resultado del fenómeno *reference bias* [62, 94], es decir, el alineamiento incorrecto de las proyecciones referencia con el ruido presente en las proyecciones experimentales (donde puede no haber realmente señal de la biomolécula); o al fenómeno de *overfitting* [28, 95], es decir, el registro de ruido de alta frecuencia que es interpretado de manera errónea como características estructurales de alta resolución presentes en la biomolécula de interés.

En la actualidad, la práctica establecida en Crio-EM es validar los mapas 3D finales. Aunque se han presentado algunos trabajos que buscan validar la asignación angular durante el proceso de refinamiento [32, 33], estas técnicas no han sido hasta ahora adoptadas de forma amplia de manera que el problema sigue abierto a nuevas propuestas.

En el presente trabajo se propuso un método para llevar a cabo el proceso de asignación angular durante el refinamiento de mapas de biomoléculas, así como la posibilidad de hacer también la validación de dicha asignación. Aunque la etapa de asignación angular no supera la robustez frente al ruido de otros métodos comunmente usados, como se mostró en la Sección 4.1, presenta una consistencia suficiente en sus asignaciones, sobre todo en media-baja y baja resolución, de manera que puede ser usado como parte del método de validación de asignaciones presentado en la Sección 4.2. En dicha sección se mostró que el enfoque de validación de asignaciones permite identificar proyecciones que contribuyen poco a mejorar la resolución del mapa, posiblemente porque exista un error en su asignación angular de acuerdo con los criterios de calidad propuestos. Este procedimiento de validación puede ser usado junto con cualquier método de asignación de orientaciones como parte de algún método de refinamiento, en donde puede servir como herramienta para excluir proyecciones pobremente registradas (que de otra forma podrían continuar siendo mal asignadas durante el proceso de refinamiento), facilitando así la reconstrucción de mapas 3D con mayor resolución. Teniendo en cuenta el creciente interés en la comunidad dedicada a hacer Crio-EM SPA para proporcionar mayor validez a los mapas reportados, el método propuesto podría complementar los enfoques existentes, ya sea durante el proceso mismo de refinamiento

de los mapas o cuando se lleva a cabo el proceso de validación cruzada de los mapas reconstruidos.

La calidad de las estructuras determinadas a través de Crio-EM SPA puede seguir mejorando si se introducen nuevos métodos de procesamiento de imagen junto con herramientas de validación más robustas que permitan identificar diferencias sutiles en mapas 3D de alta resolución, de manera que SPA pueda ser aplicado con especímenes más complejos o heterogéneos [96, 97]. Aunque también es importante mencionar que, para alcanzar dicho objetivo, las herramientas computacionales necesitan trabajar de manera conjunta con otros procesos involucrados en SPA tales como la preparación de los especímenes y la adquisición de proyecciones.

## Capítulo 5

# Conclusiones y Trabajo Futuro

La calidad de los mapas 3D en Crio-EM SPA dependen críticamente de asignar orientaciones de forma precisa a las proyecciones experimentales previo al proceso de reconstrucción. Como ya hemos visto, el problema básico en Crio-EM SPA es uno de reconstrucción tomográfica en el que se tienen proyecciones tomadas desde orientaciones desconocidas generadas a partir de múltiples copias de una biomolécula que, en principio se consideran idénticas, pero que en la práctica presentan heterogeneidades composicionales y estructurales; además, dichas proyecciones cuentan con una muy baja relación señal-ruido a pesar de todos los desarrollos instrumentales para adquisición. Estas características de las proyecciones obtenidas en Crio-EM SPA implican la presencia de muchos mínimos locales en las funciones de optimización, lo cual a su vez limita la calidad de los mapas reconstruidos.

En el presente trabajo se propuso un método, basado en correlación cruzada, para llevar a cabo el proceso de asignación angular durante el refinamiento de mapas 3D de biomoléculas y otro parahacer la validación de dicha asignación. Esta última característica es importante y novedosa, ya que ningún método de asignación angular usado en el campo de biología estructural hace una validación sobre sus propias asignaciones. Aunque el método propuesto para la etapa de asignación angular no supera la robustez frente al ruido de otros métodos comúnmente usados, sí presenta una consistencia suficiente en sus asignaciones, sobre todo en media-baja y baja resolución, lo cual permite que pueda ser usado como parte del método propuesto para validar las

asignaciones angulares hechas durante el proceso de refinamiento. Para el proceso de validación de asignaciones angulares previas, se presentó una metodología basada en el procesamiento de señales con soporte en gráficos de nodos y aristas que permite analizar el entorno de correlación como una función de la orientación a partir de la cual fueron generadas las proyecciones usadas como referencia, este es un enfoque que permite estimar la confiabilidad de las orientaciones asignadas buscando un consenso entre la información calculada a partir del vecindario de proyecciones de referencia. Usando este método, se pueden identificar proyecciones experimentales con baja confiabilidad cuya asignación angular pueda afectar de manera negativa la reconstrucción final del mapa 3D.

Una gran ventaja de la implementación de los métodos propuestos, tanto de asignación angular como de validación de asignaciones, es que fueron desarrollados como parte de la plataforma Scipion la cual integra varios paquetes de *software* utilizados en el ámbito de Crio-EM y biología estructural. Lo anterior es relevante porque facilita su difusión y aumenta la posibilidad de que sea probado por los usuarios en diferentes proyectos, con proyecciones de diferentes características y diferentes niveles de complejidad asociados principalmente a la presencia de heterogeneidades estructurales y composicionales. Adicionalmente, facilita también la construcción de otras herramientas de procesamiento dentro de Scipion que usen como base los programas construidos en el presente trabajo.

La idea del método propuesto para validar una asignación angular previa ofrece la posibilidad de extenderse en la clasificación o detección de heterogeneidades en el conjunto de proyecciones experimentales, esto se hace utilizando el mismo esquema de validación de asignaciones angulares pero esta vez utilizando dos o más mapas 3D iniciales a partir de los cuales se genera la galería de proyecciones referencia. Con esto se busca, al final de cada etapa de asignación angular (una diferente para cada modelo o mapa inicial), sacar del conjunto total de proyecciones experimentales aquellas que según el criterio de validación propuesto no se ajusten de manera confiable a cada mapa inicial. Hemos implementado esta idea en Scipion y a la fecha hemos diseñado una serie de experimentos para hacer pruebas preliminares.

Finalmente, aunque los programas incluidos en el presente trabajo permiten ejecutar varios procesos al tiempo, éstos están limitados por el número de procesadores que tenga la máquina donde se ejecutan; de manera que se espera poder construir versiones que funcionen en paralelo sobre tarjetas gráficas (GPUs, por sus siglas en inglés).

## Apéndice A

# Fourier para Señales Indexadas por Gráficos

### A.1. Propiedades de la Transformada de Fourier

La Transformada de Fourier es una operación matemática que permite descomponer una función en sus frecuencias constituyentes. Esta herramienta desempeña un papel muy importante en el procesamiento de señales así como en otros campos de la ciencia y la ingeniería. Se presentan a continuación algunas características asociadas al cambio de representación de señales usando la Transformada de Fourier, para luego presentar en la Sección A.2 el análisis relacionado con este cambio de representación para señales indexadas a través de gráficos.

**Definición A.1.** Para una función continua y diferenciable  $f$ , la variación total está definida como  $\|f\| = \int_{-\infty}^{\infty} |f'(t)| dt$ , donde  $f'(t)$  es su derivada. Para el caso discreto, la variación total se define típicamente como  $\|f\| = \sum_n |f(n) - f(n-1)|$ . Además se dice que  $f$  tiene variación acotada si  $\|f\| < +\infty$ .

La variación total permite medir la amplitud total de las oscilaciones de una función, lo cual es relevante para el estudio de técnicas de procesamiento de señales ya que la amplitud total de las oscilaciones impacta en la tendencia a decaer de los coeficientes de Fourier que representan la función en este espacio. El producto interno de dos funciones

$f, g \in L_2(\Omega \subseteq \mathbb{R})$  es  $\langle f, g \rangle = \int_{\Omega} f(t) g^*(t) dt$ , por lo que  $\|f\|^2 = \langle f, f \rangle = \int_{\Omega} |f(t)|^2 dt$ .

**Proposición A.1.** en [98]. Si  $f(t)$  es diferenciable y  $\hat{f}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t} dt$  es su transformada de Fourier, entonces  $|\hat{f}(\omega)| \leq \frac{\|f\|}{|\omega|}$ .

En el análisis de teoría de aproximación, se considera que una función  $f$  de cuadrado integrable en el intervalo  $[0, 1]$  se puede descomponer como  $f(t) = \sum_{m=-\infty}^{\infty} |\langle f(u), e^{i2\pi mu} \rangle| e^{i2\pi mt}$  con  $\langle f(u), e^{i2\pi mu} \rangle = \int_0^1 f(u) e^{-i2\pi mu} du$ , de manera que el término de aproximación de Fourier de la función  $f(t)$  con  $M$  términos es  $f_M = \sum_{|m| < M/2} |\langle f(u), e^{i2\pi mu} \rangle| e^{i2\pi mt}$ .

**Definición A.2.** El término de error de la aproximación de Fourier con  $M$  términos:

$$\epsilon_l(M, f) = \sum_{|m| > M/2} |\langle f(u), e^{i2\pi mu} \rangle|^2.$$

Es decir que la aproximación lineal mantiene las  $M$  componentes de frecuencia menor mientras descarta el resto.

**Teorema A.1** en [98]. Si  $\|f\| < +\infty$ , entonces  $\epsilon_l(M, f) = O(\|f\|_V M^{-1})$ .

**Teorema A.2** en [98]. Para cualquier  $s > 1/2$ , si  $\sum_{m=0}^{\infty} |m|^{2s} |\langle f, g_m \rangle|^2 < +\infty$  donde  $g_m$  es el  $m$ -ésimo vector de una cierta base ortogonal, entonces  $\epsilon_l(M, f) \sim o(M^{-2s})$ .

Los teoremas de arriba describen la tasa de decaimiento de los coeficientes de Fourier y el comportamiento del error de aproximación lineal. Cabe anotar que la Proposición A.1 es consistente con el hecho de que una función suave es candidata a ser comprimida cuando se usa su representación en el espacio de Fourier, es decir que ésta puede ser fielmente representada usando solo una porción de sus coeficientes de Fourier. El Teorema A.1 muestra que el error de aproximación está acotado por el valor de variación total, es decir que señales con poca variación total resultan en valores bajos de error de aproximación lineal. El Teorema A.2 resalta que el error de aproximación lineal depende de la tasa de decaimiento de  $|\langle f, g_m \rangle|$ . En la próxima subsección se mostrará como estas características de la transformada de Fourier tienen versiones similares en la llamada GFT (sigla en inglés de Transformada de Fourier del Gráfico).



## A.2. Propiedades de la GFT

Un gráfico  $G(V, E)$  se compone de un conjunto  $V$  de vértices, o nodos, y un conjunto de arcos que unen a los vértices. Una función  $f$  se dice que está definida sobre una gráfica  $G$  si  $f : V \rightarrow \mathbb{R}$  o  $f \in \mathbb{R}^V$ . Para gráficos no dirigidos, o bidireccionales, que constan del conjunto de vértices  $V$  con cardinalidad  $|V| = N$ , se puede definir la matriz de pesos  $\mathbf{W}_{N \times N}$  con entradas

$$w_{i,j} = \begin{cases} dist(i, j) \in \mathbb{R}^+, & \text{si existe un borde entre los nodos,} \\ 0, & \text{c. o. c.,} \end{cases}$$

donde el grado del  $i$ -ésimo vértice, denotado como  $d_i$  corresponde a la suma  $\sum_j w_{i,j}$ , es decir que corresponde a la suma de los pesos de los bordes que ingresan al nodo  $i$ . La función  $dist(i, j)$  puede seleccionarse según la aplicación, por ejemplo, puede usarse la distancia de muestreo que separa a los dos vértices  $i$  y  $j$  o la distancia entre vectores de características asociados a esos vértices. Esta información se puede usar para definir la matriz  $\mathbf{D}$  como una matriz diagonal que contiene en cada entrada el grado del vértice; mientras la matriz Laplaciana  $\mathbf{L}$ , la cual es indispensable para llevar a cabo el análisis de señales indexadas a través de gráficos, se puede obtener como  $\mathbf{L} = \mathbf{D} - \mathbf{W}$ .

Para tener un primer acercamiento a la relación entre la transformada de Fourier y el análisis espectral que se puede realizar sobre las señales indexadas a través de gráficos bidireccionales, se puede usar el hecho de que los vectores propios de la matriz Laplaciana de cualquier gráfico circulante (p.e., muestras ordenadas a través de un anillo circular o una rejilla regular 2-D como la usada para representar imágenes digitales) son iguales a las bases de la transformada discreta de Fourier DFT de éstos [70, 99]. Además como se presentará a continuación, las características mencionadas antes para la transformada de Fourier se mantienen en gráficos bidireccionales con estructuras más complejas.

Por otra parte, un concepto vital asociado a la Transformada de Fourier es la “suavidad de las funciones” que se asocia a funciones suaves que poseen coeficientes de Fourier compresibles, es decir, que una función se puede aproximar con una porción

de los coeficientes de mayor magnitud mientras se descartan los otros. La “suavidad” de una función representada en un gráfico se puede entender en términos de que el valor de la función asociado a un determinado vértice sea similar a los valores de la función en los vértices vecinos. Esto se puede definir también en términos de la variación total del gráfico  $\|f\|_G$  la cual describe la suavidad de una función indexada a través de los vértices de un gráfico.

**Definición A.3.** La variación total de un gráfico: Dada una función  $f \in \mathbb{R}^V$ ,  $\|f\|_G = (f^T \mathbf{L} f)^{1/2} = \left( \sum_{i=1}^N \sum_{j=1}^N w_{ij} (f(i) - f(j))^2 \right)^{1/2}$ .

**Definición A.4.** Se dice que  $f \in \mathbb{R}^V$  tiene una variación acotada si es posible encontrar un número positivo  $c \ll \lambda_{N-1}$  tal que  $\|f\|_G^2 \leq c \|f\|^2$ , donde  $\lambda_{N-1}$  es el valor propio más grande de la matriz Laplaciana  $\mathbf{L}$  del gráfico  $G$ .

A continuación se muestra que el error de aproximación lineal para la GFT tiene propiedades similares a la de la Transformada de Fourier.

**Definición A.5.** El error de aproximación lineal usando  $M$  términos:  $\epsilon_l(M, f) = \sum_{i=M}^{N-1} |\hat{f}(\lambda_i)|^2$ , donde  $\hat{f}(\lambda_i) = \langle f, u_i \rangle$  denota el  $i$ -ésimo coeficiente de la GFT de la función  $f$  y  $u_i$  es el  $i$ -ésimo vector propio de la matriz Laplaciana  $\mathbf{L}$  del gráfico  $G$ .

Los siguientes teoremas describen las propiedades de la GFT.

**Teorema A.3.** Dada la función  $f : V \rightarrow \mathbb{R}$  sobre los vértices de un gráfico  $G(V, E)$ , y sea  $\lambda_i$  el  $i$ -ésimo valor propio de la matriz Laplaciana  $\mathbf{L}$  y  $\hat{f}(\lambda_i)$  el  $i$ -ésimo coeficiente GFT de la función  $f$ ; entonces  $|\hat{f}(\lambda_i)| \leq \frac{\|f\|_G}{\sqrt{\lambda_i}}$ .

*Bosquejo de la prueba:* Es directo que  $\lambda_i |\hat{f}(\lambda_i)|^2 \leq \sum_{i=0}^{N-1} \lambda_i |\hat{f}(\lambda_i)|^2 = \sum_{i=0}^{N-1} \lambda_i (\langle f, u_i \rangle) (\langle f, u_i \rangle)^T = \sum_{i=0}^{N-1} \lambda_i (f^T u_i) (u_i^T f) = f^T \left( \sum_{i=0}^{N-1} \lambda_i u_i u_i^T \right) f = f^T \mathbf{L} f = \|f\|_G^2$ .

Comparado con la Proposición A.1, el Teorema A.3 implica que los valores propios del Laplaciano del gráfico desempeñan el mismo papel que las “frecuencias” en el procesamiento clásico de señales. El conjunto de valores propios  $\{\lambda_0, \dots, \lambda_{N-1}\}$  corresponde a los coeficientes de “Fourier” del gráfico desde bajas frecuencias hasta altas frecuencias. Conforme a lo anterior, los vectores propios del Laplaciano son de hecho las componentes de “frecuencia” del gráfico.

**Teorema A.4.** *Considere un gráfico  $G$  con una función  $f : V \rightarrow \mathbb{R}$  soportada en él. Si  $f$  tiene varianza acotada, entonces para un  $M$  suficientemente grande se tiene que:*

$$\epsilon_l(M, f) \leq \|f\|_G^2 \lambda_M^{-1}.$$

*Bosquejo de la prueba.* Nótese que  $\sum_{i=M}^{N-1} \lambda_i |\widehat{f}(\lambda_i)|^2 \leq \sum_{i=0}^{N-1} \lambda_i |\widehat{f}(\lambda_i)|^2 = \|f\|_G^2$  y  $\epsilon_l(M, f) = \sum_{i=M}^{N-1} |\widehat{f}(\lambda_i)|^2$ . Además para interpretar este último término se puede considerar el siguiente problema de optimización:

$$\text{máx} \sum_{i=M}^{N-1} x_i^2 \text{ s.t. } \sum_{i=M}^{N-1} \lambda_i x_i^2 \leq \|f\|_G^2, \quad (\text{A.1})$$

cuya solución es  $x_M^* = \|f\|_G^2 \lambda_M^{-1}$  y  $x_i^* = 0$  para todo  $i = M + 1, \dots, N - 1$ . De manera que  $\sum_{i=M}^{N-1} (x_i^*)^2$  es un límite superior para  $\epsilon_l(M, f)$  así como también lo es  $\|f\|^2$ . Entonces  $\epsilon_l(M, f) \leq \min \left\{ \|f\|^2, \|f\|_G^2 \lambda_M^{-1} \right\}$  y debido a la condición de variación acotada presentada en la Definición A.4 se tiene que  $\|f\|_G^2 \lambda_M^{-1} \leq \frac{c}{\lambda_M} \|f\|^2$  y debido a que  $c \ll \lambda_N$ , siempre se puede hallar un  $\lambda_M > c$  para un  $M$  suficientemente grande tal que la condición  $\|f\|_G^2 \lambda_M^{-1} \leq \|f\|^2$  se cumpla.

La anterior afirmación corresponde al Teorema A.1 para la Transformada de Fourier. Del Teorema A.4, el límite superior del error de aproximación lineal está relacionado tanto al Mésimo valor propio (por la condición  $\epsilon_l(M, f) \leq \|f\|_G^2 \lambda_M^{-1}$ ) y a la variación total de gráfico  $\|f\|_G$ . Esto implica que si los valores propios tienen tendencia a crecer, el error de aproximación lineal tiene la tendencia a decrecer. Además, dado que el error de aproximación lineal también se ve afectado por la variación total del gráfico  $\|f\|_G$ , entonces si  $\|f\|_G$  es acotado entonces el error de aproximación lineal se espera que sea pequeño. Lo anterior es consistente con la intuición de que una función suave puede ser aproximada de buena manera incluso cuando se usan pocos coeficientes.

## Apéndice B

# Algoritmo para Asignación

## Angular

---

**Algorithm B.1** Angular assignment for Cryo-EM SPA

---

- 1: Read experimental  $\mathcal{P}$  and reference  $\mathcal{R}$  projections
  - 2: neighboring graph  $G(V, E)$
  - 3: Laplacian matrix  $\mathbf{L}$  of graph  $G$
  - 4: eigen-values  $\{\lambda_i\}_{1 \leq i \leq |V|}$  and eigen-vectors  $\{\mathbf{u}_i\}_{1 \leq i \leq |V|}$  of  $\mathbf{L}$
  - 5: **for**  $f_k(\mathbf{x}) \in \mathcal{P}$  **do**
  - 6:      $\hat{f}_k(\boldsymbol{\xi}) \leftarrow \mathcal{F}\{f_k(\mathbf{x})\}$
  - 7:     **for**  $g_i(\mathbf{x}) \in \mathcal{R}$  **do**
  - 8:          $\hat{g}_i(\boldsymbol{\xi}) \leftarrow \mathcal{F}\{g_i(\mathbf{x})\}$
  - 9:          $\{\psi_j\}_{1 \leq j \leq 4} \leftarrow \text{find peaks of } \mathcal{F}^{-1} \left\{ \left( \mathcal{F} \left\{ \left| \hat{f}_k(\rho, \theta) \right| \right\} \odot \left( \mathcal{F} \{ |\hat{g}_i(\rho, \theta)| \} \right)^* \right) (\boldsymbol{\xi}) \right\}$
  - 10:          $\{\psi_i, \mathbf{t}_i, \phi_i\} \leftarrow \arg \max_{1 \leq j \leq 4} \phi(g_i(\mathbf{x}), M_j(f_k(\mathbf{x}))) = \arg \max_{1 \leq j \leq 4} \phi(g_i(\mathbf{x}), f_k(\mathbf{R}_{\psi_j} \mathbf{x} + \mathbf{t}_j))$
  - 11:         **for**  $g_i(\mathbf{x}) : \phi(g_i(\mathbf{x}), M_i(f_k(\mathbf{x}))) \geq \text{Threshold}$  **do**
  - 12:              $\{\psi_i, \mathbf{t}_i, \phi_i\} \leftarrow \max(\phi_{TR}(g_i(\mathbf{x}), TR(f_k(\mathbf{x}))), \phi_{RT}(g_i(\mathbf{x}), RT(f_k(\mathbf{x}))))$
  - 13:         graph-signal  $\Phi = \{\phi(g_1), \dots, \phi(g_{|V|})\}$  with  $\phi(g_l) = \phi(g_l(\mathbf{x}), M_l(f_k(\mathbf{x})))$
  - 14:         spectral graph-signal  $\hat{\phi}(\lambda_m) = \sum_{l=1}^{|V|} \phi(g_l) u_{ml}$
  - 15:         filter spectral graph-signal  $\tilde{\phi}(g_l) = \sum_{m=1}^M \hat{\phi}(\lambda_m) u_{ml}$
  - 16:         filtered graph-signal  $\tilde{\Phi} = \{\tilde{\phi}(g_1), \dots, \tilde{\phi}(g_{|V|})\}$
  - 17:         max. directions;  $\mathbf{v}_{f_k} = \max(\Phi)$  and  $\tilde{\mathbf{v}}_{f_k} = \max(\tilde{\Phi})$
  - 18:          $d_{f_k} = \cos^{-1}(\langle \mathbf{v}_{f_k}, \tilde{\mathbf{v}}_{f_k} \rangle)$
  - 19:          $\tilde{\phi}_{f_k} = \phi(g_i(\mathbf{x}), \tilde{M}_i(\tilde{g}_i(\mathbf{x})))$
- 

En las líneas 1-4 se hace la lectura de los conjuntos de datos, se crea el gráfico  $G$  y su respectiva matriz Laplaciana  $\mathbf{L}$ ; y de esta última se calculan sus valores y vectores propios. Luego cada proyección experimental  $f_k(\mathbf{x})$  se compara contra todas las proyecciones del conjunto de referencia  $\mathcal{R}$ ; primero usando el método tiene como entrada la representación en polares de la magnitud de los espectros de las proyecciones

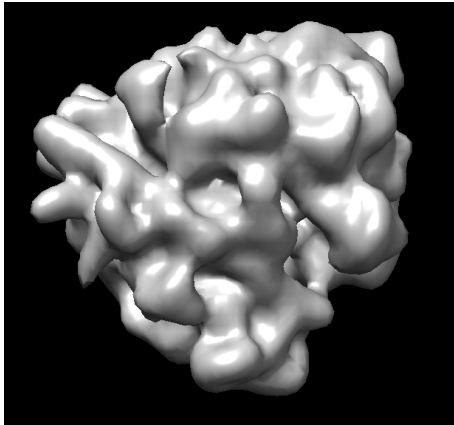
a ser alineadas (líneas 9 y 10) y luego con un alineamiento más intensivo que usa la representación de las imágenes en el espacio real, pero esta vez solo para un subconjunto de las proyecciones de referencia que tienen mayores valores de similitud  $\phi(g_i(\mathbf{x}), M_i(f_k(\mathbf{x})))$  luego del primer ciclo de búsqueda. En este punto se tendría la proyección referencia  $g_i$  que más se parece a la proyección experimental; de manera que la orientación a partir de la cual fue generada la proyección referencia  $g_i$  podría ser asignada a  $f_k$ ; pero antes de eso, se evalúa la confiabilidad de dicha asignación, primero construyendo la representación de los valores de similitud mediante el gráfico  $G$  (línea 13) y su respectiva versión filtrada (líneas 14 y 15) para posteriormente usar estas representaciones para calcular las medidas de confiabilidad propuestas (líneas 17-19).

## Apéndice C

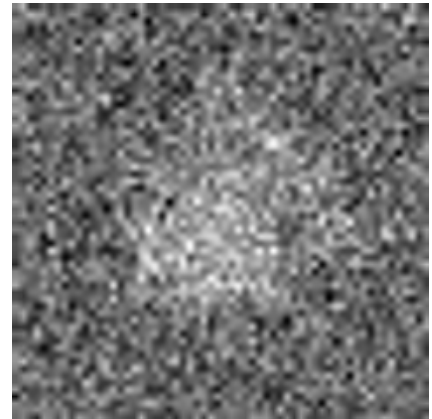
# Efecto de una Asignación Angular Errónea

Una vez presentados los elementos teóricos sobre el uso de la FSC como herramienta para determinar la resolución de un mapa reconstruido en el ámbito de Crio-EM (ver Sección 2.2), se presenta a continuación la descripción y resultados de un experimento que permite observar el efecto que tiene sobre el proceso de refinamiento una correcta asignación angular.

A partir de un modelo suavizado de un ribosoma (ver Figura C.1a) se generaron proyecciones desde direcciones conocidas, luego a cada imagen se agregó ruido gaussiano de parámetros conocidos (ver Figura C.1b) para que éstas tuvieran una SNR de aproximadamente  $-10.0$  dB. Una primera reconstrucción se llevó a cabo a partir de las proyecciones ruidosas construidas, usando para cada imagen la misma orientación dada por los parámetros de rotación y elevación que se usaron en el proceso de generación de la proyección, simulando una asignación angular correcta para cada imagen. Posteriormente, se hizo una segunda reconstrucción, pero en este caso se agregó a los valores reales de rotación y elevación un valor aleatorio de una distribución uniforme con valores mínimo y máximo simétricos al rededor de cero, simulando así varios niveles asignación angular errónea. Las distribuciones uniformes usadas para simular dichos niveles de asignación angular



(a) Modelo de Ribosoma sometido a filtrado paso-bajas.



(b) Ejemplo de proyección ruidosa.

Figura C.1:

errónea fueron:  $\mathcal{U}(0, 0)$ ,  $\mathcal{U}(-1, 1)$ ,  $\mathcal{U}(-2, 2)$ ,  $\mathcal{U}(-3, 3)$ ,  $\mathcal{U}(-5, 5)$  y  $\mathcal{U}(-10, 10)$ . Finalmente, para cada nivel de asignación angular errónea, fueron usados los dos volúmenes reconstruidos (asignación correcta vs asignación incorrecta) para calcular la curva de FSC. Los resultados son como se muestran en la Figura C.2; en ésta se puede observar la disminución de la resolución del mapa final reconstruido al aumentar el rango de los valores aleatorios añadidos a los parámetros reales de rotación y elevación. Para el ejemplo presentado se puede afirmar que muy poco de los detalles de alta resolución se conservan a partir de  $\mathcal{U}(-3, 3)$  (ver Figura C.3).

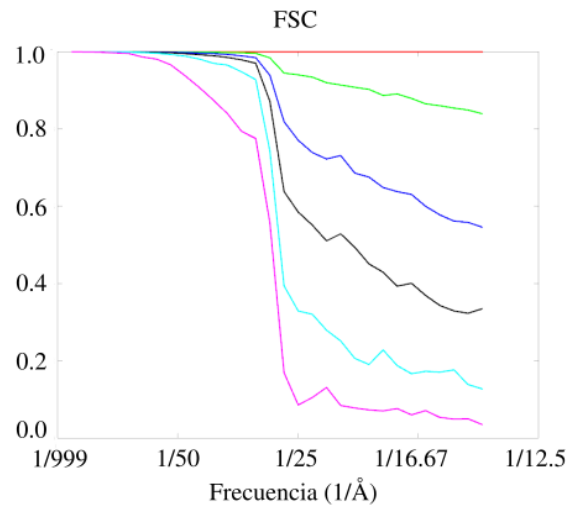


Figura C.2: Curvas de FSC que resultan de comparar un mapa reconstruido con asignación angular correcta contra mapas reconstruidos con diferentes niveles de asignación angular errónea. Desde la “mejor” FSC (curva de color rojo) hasta la “peor” (curva de color magenta) se agregó a los valores reales de rotación y elevación un valor aleatorio de una distribución uniforme con valores mínimo y máximo simétricos al rededor de cero:  $\mathcal{U}(0, 0)$ ,  $\mathcal{U}(-1, 1)$ ,  $\mathcal{U}(-2, 2)$ ,  $\mathcal{U}(-3, 3)$ ,  $\mathcal{U}(-5, 5)$  y  $\mathcal{U}(-10, 10)$ .

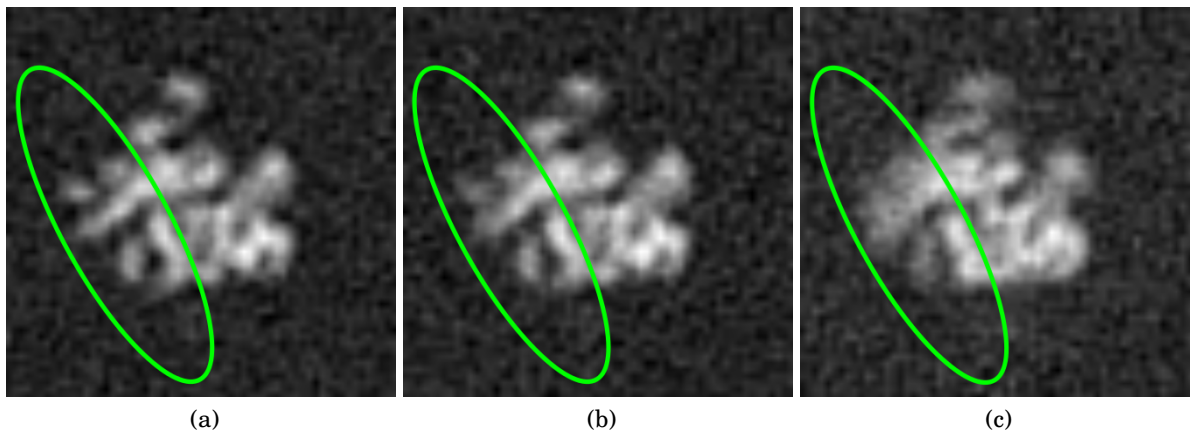


Figura C.3: Ejemplo de pérdida de detalles de alta resolución. Cortes transversales del volumen reconstruido usando (a) asignación angular correcta, (b) asignación angular cuya incertidumbre se asocia a una distribución  $\mathcal{U}(-3, 3)$  y (c) usando distribución  $\mathcal{U}(-5, 5)$ .



# Referencias

- [1] L. J. Banaszak, *Foundations of structural biology*. Academic Press, 2000.
- [2] M. Eisenstein, “The field that came in from the cold,” *Nature Methods*, vol. 13, no. 1, pp. 19–22, 2016.
- [3] E. Nogales, “The development of cryo-EM into a mainstream structural biology technique,” *Nature Methods*, vol. 13, no. 1, pp. 24–27, 2016.
- [4] R. M. Glaeser, “How good can cryo-EM become?,” *Nature Methods*, vol. 13, no. 1, pp. 28–32, 2016.
- [5] R. Henderson, “Overview and future of Single Particle Electron Cryomicroscopy,” *Archives of Biochemistry and Biophysics*, vol. 581, no. 1, pp. 19–24, 2015.
- [6] C. O. S. Sorzano, J. Vargas, J. Otón, J. M. de la Rosa-Trevín, J. L. Vilas, M. Kazemi, R. Melero, L. del Caño, J. Cuenca, P. Conesa, J. Gómez-Blanco, R. Marabini, and J. M. Carazo, “A survey of the use of iterative reconstruction algorithms in electron microscopy,” *BioMed Research International*, vol. 2017, no. 1, 2017. Article ID 6482567, 17 pages.
- [7] J. M. Carazo, C. O. S. Sorzano, J. Otón, R. Marabini, and J. Vargas, “Three-dimensional reconstruction methods in Single Particle Analysis from transmission electron microscopy data,” *Archives of Biochemistry and Biophysics*, vol. 581, no. 1, pp. 39–48, 2015.
- [8] R. M. Glaeser and K. A. Taylor, “Electron diffraction of frozen, hydrated protein crystals,” *Science*, vol. 13, no. 1, pp. 1036–1042, 1974.

- [9] J. Dubochet and A. W. McDowell, “Vitrification of pure water for electron microscopy,” *Journal of Microscopy*, vol. 124, no. 1, pp. 3–4, 1981.
- [10] J. Frank, “Averaging of low exposure electron micrographs of non-periodic objects,” *Ultramicroscopy*, vol. 1, no. 1, pp. 159–162, 1975.
- [11] P. J. Shen, “The 2017 Nobel Prize in Chemistry: Cryo-EM comes of age,” *Analytical and Bioanalytical Chemistry*, vol. 1, no. 1, pp. 2053–2057, 2018.
- [12] A. E. Leschziner and E. Nogales, “Visualizing flexibility at molecular resolution: Analysis of heterogeneity in Single-Particle Electron Microscopy reconstructions,” *Annual Review of Biophysics and Biomolecular Structure*, vol. 36, no. 1, pp. 43–62, 2007.
- [13] S. H. W. Scheres, H. Gao, M. Valle, G. T. Herman, P. P. B. Eggermont, J. Frank, and J. M. Carazo, “Maximum likelihood refinement of electron microscopy data with normalization errors,” *Nature Methods*, vol. 4, no. 1, pp. 27–29, 2007.
- [14] S. H. W. Scheres, “A Bayesian view on Cryo-EM structure determination,” *Journal of Molecular Biology*, vol. 415, no. 2, pp. 406–418, 2012.
- [15] Q. Jin, C. O. S. Sorzano, J. M. de la Rosa-Trevín, J. R. Bilbao-Castro, R. Núñez Ramírez, O. Llorca, F. Tama, and S. Jonić, “Iterative elastic 3D-to-2D alignment method using normal modes for studying structural dynamics of large macromolecular complexes,” *Structure*, vol. 22, no. 3, pp. 496–506, 2014.
- [16] C. O. S. Sorzano, L. G. de La Fraga, R. Clackdoyle, and J. M. Carazo, “Normalizing projection images: A study of image normalizing procedures for single particle three-dimensional electron microscopy,” *Ultramicroscopy*, vol. 101, no. 4, pp. 129–138, 2004.
- [17] P. A. Penczek, R. A. Grasucci, and J. Frank, “The ribosome at improved resolution: New techniques for merging and orientation refinement in 3D cryo-electron microscopy of biological particles,” *Ultramicroscopy*, vol. 53, no. 1, pp. 251–270, 1994.

- [18] M. van Heel, “Angular reconstitution: A posteriori assignment of projection directions for 3D reconstruction,” *Ultramicroscopy*, vol. 21, no. 2, pp. 111–123, 1987.
- [19] A. Singer, R. R. Coifman, F. J. Sigworth, D. W. Chester, and Y. Shkolnisky, “Detecting consistent common lines in Cryo-EM by voting,” *Journal of Structural Biology*, vol. 169, no. 3, pp. 312–322, 2010.
- [20] E. Sanz-García, A. B. Stewart, and D. M. Belnap, “The random-model method enables ab initio 3D reconstruction of asymmetric particles and determination of particle symmetry,” *Journal of Structural Biology*, vol. 171, no. 2, pp. 216–222, 2010.
- [21] P. A. Penczek, J. Zhu, and J. Frank, “A common-lines based method for determining orientations for  $N > 3$  particle projections simultaneously,” *Ultramicroscopy*, vol. 63, no. 3-4, pp. 205–218, 1996.
- [22] S. Jonić, C. O. S. Sorzano, P. Thévenaz, C. El-Bez, S. De Carlo, and M. Unser, “Spline-based image-to-volume registration for three-dimensional electron microscopy,” *Ultramicroscopy*, vol. 103, no. 4, pp. 303–317, 2005.
- [23] A. Singer and Y. Shkolnisky, “Three-dimensional structure determination from Common Lines in Cryo-EM by eigenvectors and semidefinite programming,” *SIAM Journal on Imaging Sciences*, vol. 4, no. 2, pp. 543–572, 2011.
- [24] H. Elmlund, D. Elmlund, and S. Bengio, “PRIME: Probabilistic initial 3D model generation for Single-Particle Cryo-Electron Microscopy,” *Structure*, vol. 21, no. 8, pp. 1299–1306, 2013.
- [25] J. Vargas, A. L. Álvarez Cabrera, R. Marabini, J. M. Carazo, and C. O. S. Sorzano, “Efficient initial volume determination from electron microscopy images of single particles,” *Bioinformatics*, vol. 30, no. 20, pp. 2891–2898, 2014.
- [26] C. O. S. Sorzano, J. Vargas, J. de la Rosa-Trevín, J. Otón, A. Álvarez Cabrera, V. Abrishami, E. Sesmero, R. Marabini, and J. M. Carazo, “A statistical approach to the initial volume problem in Single Particle Analysis by Electron Microscopy,” *Journal of Structural Biology*, vol. 189, no. 3, pp. 213–219, 2015.

- [27] C. O. S. Sorzano, J. Vargas, J. M. de la Rosa-Trevín, A. Jiménez, D. Maluenda, R. Melero, M. Martínez, E. Ramírez-Aportela, P. Conesa, J. L. Vilas, R. Marabini, and J. M. Carazo, “A new algorithm for high-resolution reconstruction of single particles by electron microscopy,” *Journal of Structural Biology*, vol. 204, no. 2, pp. 329–337, 2018.
- [28] A. Stewart and N. Grigorieff, “Noise bias in the refinement of structures derived from single particles,” *Ultramicroscopy*, vol. 102, no. 1, pp. 67–84, 2004.
- [29] A. Punjani, M. A. Brubaker, and D. J. Fleet, “Building proteins in a day: Efficient 3D molecular structure estimation with electron cryomicroscopy,” vol. 39, pp. 706–718, 2017.
- [30] T. Grant, A. Rohou, and N. Grigorieff, “cisTEM, user-friendly software for single-particle image processing,” *eLife*, vol. 7, 2018.
- [31] C. F. Reboul, M. Eager, D. Elmlund, and H. Elmlund, “Single-particle cryo-EM-improved *ab initio* 3d reconstruction with simple/prime,” *Protein science*, vol. 27, pp. 51–61, Aug. 2018.
- [32] J. Vargas, J. Oton, R. Marabini, J. M. Carazo, and C. O. S. Sorzano, “Particle alignment reliability in single particle electron cryomicroscopy: a general approach,” *Scientific Reports*, vol. 6, p. 21626, 2016.
- [33] J. Vargas, R. Melero, J. Gomez-Blanco, J. M. Carazo, and C. O. S. Sorzano, “Quantitative analysis of 3d alignment quality: its impact on soft-validation, particle pruning and homogeneity analysis,” *Scientific Reports*, vol. 7, p. 6307, 2017.
- [34] B. S. Reddy and B. N. Chatterji, “An FFT-based technique for translation, rotation, and scale-invariant image registration,” *IEEE Transactions on Image Processing*, vol. 5, no. 8, pp. 1266–1271, 1996.
- [35] N. A. Anoshina, A. S. Krylov, and D. V. Sorokin, “Correlation-based 2D registration method for single particle cryo-EM images,” in *2017 Seventh International*

- Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6, 2017.
- [36] R. H. Wade, “A brief look at imaging and contrast transfer,” *Ultramicroscopy*, vol. 46, no. 1, pp. 145–156, 1992.
- [37] G. Jensen, ed., *Cryo-EM, Part B: 3-D Reconstruction*, vol. 482 of *Methods in Enzymology*. Academic Press - Elsevier, 2010.
- [38] P. R. Baldwin and P. A. Penczek, “The transform class in SPARX and EMAN2,” *Journal of Structural Biology*, vol. 157, no. 1, pp. 250–261, 2007.
- [39] O. Scherzer, “The theoretical resolution limit of the electron microscope,” *Journal of Applied Physics*, vol. 20, no. 1, pp. 20–29, 1949.
- [40] V. Abrishami, *New computational methods toward atomic resolution in single particle cryo-electron microscopy*. PhD thesis, Computer Science Department, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 2016.
- [41] A. Rohou and N. Grigorieff, “Ctffind4: Fast and accurate defocus estimation from electron micrographs,” *Journal of Structural Biology*, vol. 192, no. 2, pp. 216–221, 2015.
- [42] C. O. S. Sorzano, S. Jonic, R. Núñez Ramírez, N. Boisset, and J. M. Carazo, “Fast, robust, and accurate determination of transmission electron microscopy contrast transfer function,” *Journal of Structural Biology*, vol. 160, no. 2, pp. 249–262, 2007.
- [43] J. Vargas, J. Otón, R. Marabini, S. Jonic, J. M. de la Rosa-Trevín, J. M. Carazo, and C. O. S. Sorzano, “Fastdef: Fast defocus and astigmatism estimation for high-throughput transmission electron microscopy,” *Journal of Structural Biology*, vol. 181, no. 2, pp. 136–148, 2013.
- [44] S. H. W. Scheres, R. Núñez Ramírez, C. O. S. Sorzano, J. M. Carazo, and R. Marabini, “Image processing for electron microscopy single-particle analysis using xmipp,” *Nature Protocols*, vol. 3, no. 1, pp. 977–990, 2008.

- [45] W. R. Leo, *Techniques for nuclear and particle physics experiments*. Springer, 2014.
- [46] W. Kühlbrandt, “Cryo-EM enters a new era,” *eLife*, vol. 3, p. e03678, aug 2014.
- [47] S. Wu, J. Armache, and Y. Cheng, “Single-particle cryo-em data acquisition by using direct electron detection camera,” *Microscopy*, vol. 65, no. 1, pp. 35–41, 2015.
- [48] V. Abrishami, J. Vargas, X. Li, Y. Cheng, R. Marabini, C. O. S. Sorzano, and J. M. Carazo, “Alignment of direct detection device micrographs using a robust optical flow approach,” *Journal of Structural Biology*, vol. 189, no. 3, pp. 163–176, 2015.
- [49] X. Li, P. Mooney, S. Zheng, C. R. Booth, M. B. Braunfeld, S. Gubbens, D. A. Agard, and Y. Cheng, “Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-em,” *Nature Methods*, vol. 10, no. 6, pp. 584–590, 2013.
- [50] J. L. Rubinstein and M. A. Brubaker, “Alignment of cryo-em movies of individual particles by optimization of image translations,” *Journal of Structural Biology*, vol. 192, no. 2, pp. 188–195, 2015.
- [51] C. W. Chong, P. Raveendran, and R. Mukundan, “Translation invariants of Zernike moments,” *Pattern Recognition*, vol. 36, no. 1, pp. 1765–1773, 2003.
- [52] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 1, pp. 91–110, 2004.
- [53] J. Frank, “Classification of macromolecular assemblies studied as «single particles»,” *Quarterly Reviews of Biophysics*, vol. 23, no. 1, pp. 281–329, 1990.
- [54] S. H. W. Scheres, M. Valle, R. Núñez, C. O. S. Sorzano, R. Marabini, G. T. Herman, and J. M. Carazo, “Maximum-likelihood multi-reference refinement for electron microscopy images,” *Journal of Molecular Biology*, vol. 348, no. 1, pp. 139–149, 2005.
- [55] C. O. S. Sorzano, J. R. Bilbao-Castro, Y. Shkolnisky, M. Alcorlo, R. Melero, G. Caffarena-Fernández, M. Li, G. Xu, R. Marabini, , and J. M. Carazo, “A clustering

- approach to multireference alignment of single-particle projections in electron microscopy,” *Journal of Structural Biology*, vol. 171, no. 1, pp. 197–206, 2010.
- [56] S. H. W. Scheres, “RELION: Implementation of a Bayesian approach to cryo-EM structure determination,” *Journal of Structural Biology*, vol. 180, no. 1, pp. 519–530, 2012.
- [57] V. Abrishami, J. R. Bilbao-Castro, J. Vargas, R. Marabini, J. M. Carazo, and C. O. S. Sorzano, “A fast iterative convolution weighting approach for gridding-based direct Fourier three-dimensional reconstruction with correction for the contrast transfer function,” *Ultramicroscopy*, vol. 157, no. 1, pp. 79–87, 2015.
- [58] P. Penczek, R. Renka, and H. Schomberg, “Gridding-based direct Fourier inversion of the three-dimensional ray transform,” *Journal of the Optical Society of America A. Optics, Image Science, and Vision*, vol. 21, no. 4, pp. 499–509, 2004.
- [59] J. de la Rosa-Trevín, A. Quintana, L. del Cano, A. Zaldívar-Peraza, I. Foche, J. Gutiérrez, J. Gómez-Blanco, J. Burguet-Castells, J. Cuenca-Alba, V. Abrishami, J. Vargas, J. Otón, G. Sharov, J. Vilas, J. Navas, P. Conesa, M. Kazemi, R. Marabini, C. O. S. Sorzano, and J. Carazo, “Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy,” *Journal of Structural Biology*, vol. 195, no. 1, pp. 93–99, 2016.
- [60] E. D. Zhong, T. Bepler, J. H. Davies, and B. Berger, “Reconstructing continuous distributions of 3D protein structure from cryo-EM images,” in *International Conference on Learning Representations*, 2020. <https://openreview.net/forum?id=SJxUjlBtwB>.
- [61] R. Henderson, S. Chen, J. Z. Chen, N. Grigorieff, L. A. Passmore, L. Ciccarelli, J. L. Rubinstein, R. A. Crowther, S. P. L., and P. B. Rosenthal, “Tilt-pair analysis of images from a range of different specimens in single-particle electron cryomicroscopy,” *Journal of Molecular Biology*, vol. 413, no. 5, pp. 1028–1046, 2011.

- [62] R. Henderson, “Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise,” in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, pp. 18037–18041, 2013.
- [63] G. Harauz and M. van Heel, “Exact filters for general geometry three dimensional reconstruction,” *Optik*, vol. 73, no. 4, pp. 146–156, 1986.
- [64] M. van Heel and M. Schatz, “Fourier Shell Correlation threshold criteria,” *Journal of Structural Biology*, vol. 151, no. 3, pp. 250–262, 2005.
- [65] C. O. S. Sorzano, J. Vargas, J. Otón, V. Abrishami, J. M. de la Rosa-Trevín, J. Gómez-Blanco, J. L. Vilas, R. Marabini, and J. M. Carazo, “A review of resolution measures and related aspects in 3D Electron Microscopy,” *Progress in Biophysics and Molecular Biology*, vol. 124, no. 1, pp. 1–30, 2017.
- [66] J. L. Vilas, *Local quality assessment of Cryo-EM reconstructions and its applications*. PhD thesis, Facultad de Ciencias - Universidad Autónoma de Madrid, Biocomputing Unit - Centro Nacional de Biotecnología, 2019.
- [67] N. Bershad and A. Rockmore, “On estimating signal-to-noise ratio using the sample correlation coefficient,” *IEEE Transactions on Information Theory*, vol. 20, no. 1, pp. 112–113, 1974.
- [68] M. Unser, B. L. Trus, and A. C. Steven, “A new resolution criterion based on spectral Signal-To-Noise ratios,” *Ultramicroscopy*, vol. 23, no. 1, pp. 39–51, 1987.
- [69] B. S. Manoj, A. Chakraborty, and R. Singh, *Complex Networks: A Networking and Signal Processing Perspective*. Prentice Hall, 2018.
- [70] X. Zhu and M. Rabbat, “Approximating signals supported on graphs,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3921–3924, 2012.
- [71] W. Waheed, G. Deng, and B. Liu, “Discrete Laplacian operator and its applications in Signal Processing,” *IEEE Access*, vol. 8, no. 1, pp. 89692 – 89707, 2020.



- [72] A. Sandryhaila and J. M. F. Moura, “Discrete signal processing on graphs: Graph Fourier Transform,” in *ICASSP 2013 - 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6167–6170, 2013.
- [73] D. Petar and R. Cédric, *Cooperative and graph signal processing : principles and applications*. Academic Press, 2018.
- [74] P. J. Besl and N. D. McKay, “A method for registration of 3D shapes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.
- [75] N. Sayan, “Image registration techniques: A survey.” DOI: 10.17605/OSF.IO/RV65C, 2017.
- [76] C. D. Kuglin and D. C. Hines, “The phase correlation image alignment method,” in *IEEE International Conference on Cybernetics and Society*, (San Francisco), pp. 163–165, IEEE, September 1965.
- [77] M. Liang, L. Xi-Hai, Z. Wan-Gang, and L. Dai-Zhi, “The Generalized Cross-Correlation method for time delay estimation of infrasound signal,” in *2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*, pp. 1320–1323, Sept 2015.
- [78] K. Pearson, “Notes on regression and inheritance in the case of two parents,” *Proceedings of the Royal Society of London*, vol. 58, no. 1, pp. 240–242, 1895.
- [79] J. L. Rodgers and W. A. Nicewanders, “Thirteen ways to look at the correlation coefficient,” *The American Statistician*, vol. 42, no. 1, pp. 59–66, 1988.
- [80] H. B. Mitchell, *Image fusion: theories, techniques and applications*. Springer, 2010.
- [81] V. V. Starovoitov, E. E. Eldarova, and K. T. Iskakov, “Comparative analysis of the SSIM index and the pearson coefficient,” *Eurasian Journal of Mathematical and Computer Applications*, vol. 8, no. 1, pp. 76–90, 2020.

- [82] C. O. S. Sorzano, R. Marabini, J. Velázquez-Muriel, J. R. Bilbao-Castro, S. H. Scheres, J. M. Carazo, and A. Pascual-Montano, “Xmipp: A new generation of an open-source image processing package for electron microscopy,” *Journal of Structural Biology*, vol. 148, no. 2, pp. 194–204, 2004.
- [83] Scipion, “Scipion: Cryo EM image processing framework.” <http://scipion.i2pc.es>.
- [84] Xmipp, “X-window-based microscopy image processing package.” <http://xmipp.i2pc.es>.
- [85] T. Bharat and S. Scheres, “Resolving macromolecular structures from electron cryotomography data using subtomogram averaging in RELION,” *Nature Protocols*, vol. 11, no. 1, pp. 2054 – 2065, 2016. doi: 10.1038/nprot.2016.124.
- [86] J. Baumgardner and P. Frederickson, “Icosahedral discretization of the two-sphere,” *SIAM Journal on Numerical Analysis*, vol. 22, no. 6, pp. 1107–1115, 1985.
- [87] EMPIAR-10010, “EMPIAR project 10010.” <https://www.ebi.ac.uk/pdbe/emdb/empiar/entry/10010/>.
- [88] Z. Wang, C. F. Hryc, B. Bammes, P. V. Afonine, J. Jakana, D. H. Chen, X. Liu, M. L. Baker, C. Kao, S. J. Ludtke, M. F. Schmid, P. D. Adams, and W. Chiu, “An atomic model of brome mosaic virus using direct electron detection and real-space optimization,” *Nature Communications*, vol. 5, no. 1, pp. 4808 – 4808, 2014.
- [89] EMPIAR-10061, “EMPIAR project 10061.” <https://www.ebi.ac.uk/pdbe/emdb/empiar/entry/10061/>.
- [90] A. Bartesaghi, A. Merk, S. Banerjee, D. Matthies, X. Wu, J. L. Milne, and S. Subramaniam, “2.2 Å resolution cryo-EM structure of  $\beta$ -galactosidase in complex with a cell-permeant inhibitor,” *Science*, vol. 348, no. 6239, pp. 1147 – 1151, 2015.
- [91] A. Punjani, D. J. Fleet, J. L. Rubinstein, and M. A. Brubaker, “CryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination,” *Nature Methods*, vol. 14, no. 3, pp. 290 – 296, 2017.

- [92] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [93] EMDB-6000, "ID: EMD-6000. Full virus map of Brome Mosaic Virus," 2014.
- [94] M. van Heel, "Finding trimeric HIV-1 envelope glycoproteins in random noise," in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, pp. E4175–E4177, 2013.
- [95] S. Chen, G. McMullan, A. R. Faruqi, G. N. Murshudov, J. M. Short, S. H. W. Scheres, and R. Henderson, "High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy," *Ultramicroscopy*, vol. 135, no. 1, pp. 24–35, 2013.
- [96] P. B. Rosenthal and J. L. Rubinstein, "Validating maps from single particle electron cryomicroscopy," *Current Opinion in Structural Biology*, vol. 34, no. 1, pp. 135–144, 2015.
- [97] P. B. Rosenthal, "Chapter nine - testing the validity of single-particle maps at low and high resolution," in *The resolution revolution: recent advances in cryoEM* (R. A. Crowther, ed.), vol. 579 of *Methods in Enzymology*, pp. 227–253, Academic Press, 2016.
- [98] S. G. Mallat, *A Wavelet Tour of Signal Processing*. San Diego: Academic Press, 3rd ed. ed., 1999.
- [99] R. M. Gray, *Toeplitz and circulant matrices: A review*. Boston: now publishers, 2006.