"Clasificación y promediado de volúmenes de tomografía electrónica"

Pablo Sprechmann

Tesis de Maestría
en Ingeniería Eléctrica

Directores de tesis:
Prof. Gregory Randall, Universidad de la República, Uruguay
Prof. Guillermo Sapiro, University of Minnesota, E.E. U.U.
Alberto Bartesaghi, National Institutes of Health, E.E. U.U.

Instituto de Ingeniería Eléctrica
Facultad de Ingeniería
Universidad de la República, Uruguay.

Agradecimientos

A mis tres tutores Alberto, Guillermo y Gregory por el apoyo permanente y las tantas cosas que me han enseñado durante estos meses de trabajo.

A todos los compañeros del IIE, en particular a los miembros del grupo de Tratamiento de Imágenes y miembros de la sala 121 por hacer del instituto un lugar tan especial en donde trabajar.

A los compañeros "adoptivos" que tuve en la Universidad de Minnesota y en el National Institutes of Health que alegraron mi estadía en el norte.

A mis amigos, muchos de los cuales les estoy agradeciendo por segunda vez en esta página.

A mi familia.

A Fabi.

Resumen

La tomografía electrónica brinda la posibilidad de determinar estructuras tridimensionales de material biológico a niveles de resolución suficientemente altos como para permitir la identificación de macromoléculas individuales tales como proteínas. Esto ha despertado un enorme interés en la comunidad científica en los últimos tiempos. El procesamiento de dichas imágenes tridimensionales constituye un problema desafiante debido a los muy bajos niveles de la relación señal a ruido que presentan. Esto hace que los volúmenes individuales prácticamente carezcan de valor, debido a que éstos son simplemente demasiado ruidosos como para permitir su correcta visualización y mucho menos su interpretación estructural. Resulta imprescindible entonces la utilización de técnicas de promediado que combinando gran cantidad de volúmenes logren aumentar drásticamente el nivel de señal en la imagen. Este tipo de técnicas ha sido de práctica frecuente en las ultimas décadas en el área de microscopía electrónica conocida con el nombre de análisis de partículas individuales, donde se han alcanzando resultados sorprendentes. La tomografía electrónica tiene sin embargo diferencias sustanciales con las técnicas de partículas individuales, debido sobre todo al hecho de que las imágenes tomográficas son tridimensionales. Se introducen una serie de nuevos problemas cuyo correcto manejo es indispensable para alcanzar una solución satisfactoria. Entre dichas diferencias se destacan el problema de lidiar con el efecto conocido con el nombre de *missing wedge*, característico de las imágenes de tomografía electrónica, así como la necesidad de contar con algoritmos eficientes de registrado y clasificación de volúmenes.

En esta tesis de maestría se presenta un estudio del problema descrito, analizando las soluciones hasta ahora propuestas en la comunidad para luego proponer una solución original. De esta manera se llega a una herramienta poderosa que cumple con todos los requisitos establecidos. Se presta particular atención en compararla con soluciones existente y en realizar experimentos, con datos artificiales y reales, que permitan su validación. A través de dichos experimentos se logra identificar con claridad cuál es el verdadero al-

cance que tiene la herramienta desarrollada y bajo que condiciones es capaz de distinguir diferentes conformaciones de material biológico.

Prefacio

El trabajo documentado en la presente tesis de maestría comenzó en el mes de enero del año 2006 con una pasantía de ocho meses que realicé en la Universidad de Minnesota (University of Minnesota). En aquella oportunidad trabajé como estudiante invitado en el laboratorio de Guillermo Sapiro. Durante dicho período Gregory Randall se encontraba también visitando dicha universidad en el contexto de su año sabático. En aquel momento comencé a trabajar en los temas que en este documento se abordan en el marco de un proyecto de mayor amplitud en el que colaboran los grupos de Guillermo Sapiro y de Sriram Subramaniam (National Cancer Institute, National Institutes of Health) desde hace más de dos años. Por ésta razón realicé en aquel período varias visitas al laboratorio del Dr. Subramaniam donde trabajé orientado por Alberto Bartesaghi.

Desde entonces he trabajado como parte de este grupo en coordinación permanente con los directores de esta tesis. En este contexto, realicé dos nuevas visitas de un mes de duración al National Institutes of Health en enero y mayo de 2007 y otra visita de tres semanas a la Universidad de (la gélida) Minnesota en febrero de 2007.

La totalidad de este proceso, que me permitió estar en contacto directo con un grupo de investigación del primer nivel, ha sido para mi una experiencia extraordinaria. He tenido el agrado de compartir la vida diaria de un laboratorio del prestigio del de Sriram Subramaniam donde se palpita que allí se hace ciencia.

P.S.

Índice general

Capítulo 1

Introducción

1.1. Introducción

La biología estructural es una rama de la biología molecular cuyo objetivo es determinar la estructura tridimensional de complejos moleculares que intervienen en diversos procesos biológicos. Un caso que recibe particular interés es el de las proteínas o agregados protéicos. Este interés proviene de la gran importancia que éstos tienen para el funcionamiento de las células y de la constatación de que la función que las mismas cumplen está muy ligada a la estructura tridimensional que éstas presentan. Lograr un entendimiento de dichas estructuras posibilitaría diseñar vacunas que ayuden a combatir diversas enfermedades, entre otras muchas el SIDA.

Uno de los mecanismos utilizados para la determinación de las estructuras de los agregados protéicos se basa en inferir su conformación tridimensional a partir de muestras o datos experimentales mediante técnicas tales como la cristalografía de rayos X, la resonancia magnética nuclear o la microscopía y tomografía electrónicas. Cada una de ellas tiene sus limitaciones y consecuentemente su rango de operación. Por ejemplo, las muestras utilizadas en la cristalografía de rayos X deben ser previamente cristalizadas cosa que no es para nada sencillo y menos aún cuando la masa de los especímenes aumenta. Sin embargo, en su rango de operación, es posible obtener imágenes de muy alta resolución. El caso de la resonancia magnética nuclear es similar al anterior pero presenta más limitaciones en cuanto a la masa de los especímenes a observar. La microscopía y la tomografía electrónica presentan características opuestas a las dos técnicas anteriores: prácticamente no impone condiciones al espécimen que se quiere observar, pero las imágenes obtenidas son de mucho menor resolución que en las otras técnicas mencionadas debido, entre otros

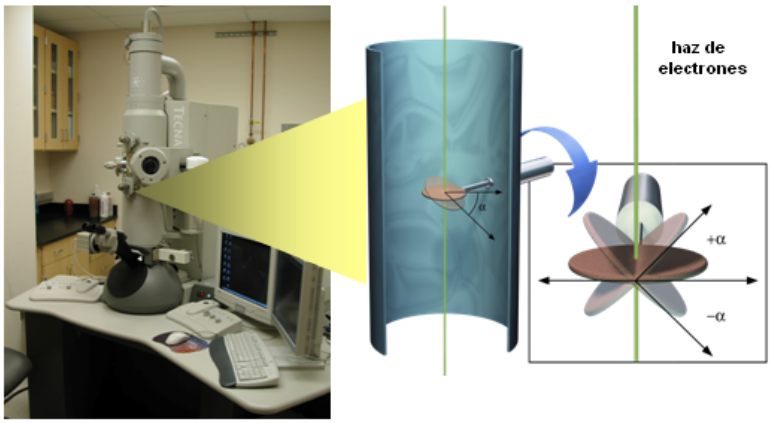


Figura 1.1: En esta figura se muestra un esquema de adquisición de la serie de proyecciones (*tilt series* en inglés). El espécimen se rota para obtener las distintas proyecciones a partir de las cuales se reconstruye un volumen. En la imagen de la izquierda se muestra como existe un máximo ángulo α que puede rotarse el espécimen.

factores, al fuerte ruido que las afecta. Sin embargo asociando esta técnica con algoritmos de procesamiento de imágenes es posible lograr muy altas resoluciones que ya se están empezando a aproximar a los niveles de resolución alcanzados con los otros métodos. En esta tesis de maestría se trabajó en el procesamiento de imágenes obtenidas mediante tomografía electrónica de transmisión o tomografía electrónica (TE) necesarios para producir los mapas de alta resolución mencionados.

La tomografía electrónica es un caso particular de la tomografía axial computarizada, o más comúnmente llamada simplemente tomografía. Esta es un técnica mediante la cual pueden producirse imágenes volumétricas a partir de varias proyecciones en dos dimensiones de un objeto tridimensional. Varios tipos de tomografía son utilizados actualmente en la medicina moderna con mucha frecuencia. Pueden citarse como ejemplos las tomografías basadas en rayos X, rayos γ, resonancia magnética y ultrasonido entre otras. La TE permite alcanzar enormes niveles de magnificación lo que permite producir imágenes con niveles de resolución que varían entre los 5 y los 10 nm.

En el microscópio electrónico de transmisión el espécimen (objeto 3D que se quiere observar) es irradiado con un haz de electrones. Algunos de los electrones del haz se desvían al atravesarlo de acuerdo a la densidad local del mismo y con los restantes se producen las imágenes. Las distintas proyec-

ciones se obtienen rotando el espécimen como se esquematiza en la Figura 1.1. De dichas proyecciones se construye la imagen tridimensional o tomograma utilizando algún algoritmo de reconstrucción como por ejemplo el de retroproyección filtrada (*filtered backprojection*). Como ya se mencionó, mediante el procesado de las imágenes obtenidas con TE pueden lograrse mapas tridimensionales con altísimas resoluciones, del orden de los 50 Å e incluso superiores.

La comunidad internacional está prestando cada vez más atención y esfuerzo al estudio de una modalidad particular de la TE conocida por el nombre de Cryo tomografía electrónica (CryoTE). Esta es una técnica mediante la cual se busca obtener una imágen que refleje a las células en estado natural. El espécimen, embebido en agua, es sometiendo a un enfriamiento extremadamente rápido previo al registrado de la tomografía, este proceso mantiene las estructuras en una conformación muy similar a la que tenían cuando estaban vivas. La posibilidad de observar a los niveles de resolución mencionados a las células en estado vivo, su interacción con distintos tipos de virus, revelar las conformaciones proteínas de distintas macromoléculas y constituye una posibilidad invalorable que se estima revolucionará la disciplina. Esta técnica es particularmente aplicable a ensamblajes macromoleculares con estructuras excesivamente complejas para ser estudiadas con otras técnicas existentes como la cristalografía de rayos X o la resonancia magnética nuclear. Dicha tecnología está siendo utilizada actualmente por varios grupos de investigadores para obtener descripciones de gran precisión de la estructura de distintos virus, lo que posibilitaría aumentar enormemente el entendimiento de la biología de los mismos e incluso el desarrollo de nuevas vacunas. Por otro lado, recientemente se han producido importantes mejoras en las técnicas y tecnologías de adquisición, por lo que no sólo se cuenta imágenes de mejor calidad, sino que estas pueden adquirirse en grandes cantidades diariamente. Se hace entonces urgente la necesidad de contar con herramientas automáticas capaces de procesar imágenes de tomografía electrónica eficazmente y mejorar aún más los niveles de resolución alcanzados.

Las novedosas posibilidades que brinda la CryoTE tienen como costo un considerable incremento en las dificultades tanto en la técnica de adquisición de las imágenes como en su posterior procesamiento. Las imágenes adquiridas por el microscopio y reconstruidas presentan varias dificultades técnicas intrínsecas, sobre las cuales se profundizará en las siguientes secciones. Las más notorias de dichas dificultades son: por un lado el nivel de ruido presente en las imágenes "crudas" debido al bajo nivel de dosis de electrones que se puede suministrar a los especímenes, a fin de minimizar el daño causa-

do a su conformación estructural. Por otro lado en la tomografía electrónica pueden adquirirse una cantidad limitada de proyecciones en un rango que cubre usualmente ±60 grados, no pudiéndose cubrir todas las posibles rotaciones, ver Figura 1.1. Esto implica necesariamente una pérdida de información y se traduce en que las imágenes estén afectadas por una fuente adicional de ruido fuertemente estructurado que se conoce con el nombre de efecto del *missing wedge*. Los dos efectos mencionados se suman al llamado "ruido biológico" que se debe a que las partículas de interés se encuentran inevitablemente rodeadas por material biológico (a veces otras partículas de interés) que dificultan todavía más la correcta visualización del espécimen. Los tomogramas obtenidos por lo tanto están muy lejos de poder ser utilizados directamente para la visualización e interpretación de la muestra. Necesariamente las imágenes deben ser procesadas para lograr obtener los mapas de alta resolución imprescindibles para la interpretación biológica. El desarrollo de técnicas capaces de sobreponerse a dichas dificultades y a la vez lidiar con la gran heterogeneidad del material biológico constituye un verdadero desafío para la comunidad científica. Diversos grupos de investigadores en el mundo trabajan en la actualidad en estos problemas todavía considerados abiertos.

El problema que se aborda en esta maestría es una parte importante del procesamiento requerido para tratar las imágenes obtenidas con TE y CryoTE. Esto incluye estudiar el marco matemático y computacional para diseñar e implementar un algoritmo capaz de producir mapas de alta resolución a partir de un conjunto de imágenes crudas en forma eficiente y computacionalmente viable. Se toma como punto de partida y referencia permanente el exitoso trabajo realizado durante las últimas tres décadas en el campo del análisis de partículas individuales (*Single particle analysis*). Esta disciplina ha logrado dar mapas de hasta 6 $\mathring{A}$ de resolución para diversas partículas, permitiendo reconocer arreglos de proteínas dentro de las mismas. Naturalmente existen grandes diferencias debido a que estas técnicas trabajan sobre proyecciones individuales (imágenes bidimensionales) pero de todas maneras los principios que se han venido utilizando pueden ser tomados y adaptados al caso tridimensional. El espíritu de estas técnicas de procesamiento es el de combinar por medio del promediado un gran número de imágenes representando conformaciones idénticas para de ésta manera mejorar la relación señal a ruido y alcanzar así la resolución deseada. Es claro que esto está basado en las propiedades de aditividad observadas en el ruido que afecta a las imágenes, sobre este y otros aspectos se discute en la presente tesis. En el caso tridimensional será también a través de esta técnica que se intentarán minimizar los problemas introducidos por el efecto del *missing wedge*. Para lograr

este objetivo es necesario realizar ciertas operaciones a los volúmenes crudos. Primeramente, las partículas o estructuras biológicas de interés se encuentran en general en orientaciones arbitrarias dentro del tomograma por lo que es imprescindible desarrollar un algoritmo que permita registrar correctamente un grupo de volúmenes. Por otro lado para garantizar que los promedios sean calculados utilizando conjuntos homogéneos de imágenes es necesario poder realizar exitosamente una clasificación de volúmenes de acuerdo a su conformación. En ambos casos es imprescindible considerar la naturaleza de dichos volúmenes teniendo especial cuidado en considerar el efecto del *missing wedge*. Para ilustrar estos conceptos puede citarse informalmente un ejemplo concreto que será luego debidamente detallado en el texto. Si se intentara registrar volúmenes afectados por el efecto del *missing wedge* con técnicas convencionales, se observaría que en la mayor parte de los casos la alineación obtenida sería tal que se conseguiría un registrado de los ruidos estructurados presentes en las imágenes sin importar la estructura del material biológico que contienen. Para lograr este objetivo resulta imprescindible contar con una medida de similaridad entre volúmenes que tenga en consideración la naturaleza de los mismos. Por otro lado es importante observar que para poder alinear correctamente las partículas entre sí éstas deben estar separadas en grupos homogéneos, en tanto que casi todos los algoritmos de clasificación de imágenes requieren que las partículas se encuentren alineadas entre si. Es por esto que la alineación y la clasificación de las partículas es un problema del tipo del *huevo y la gallina*. Esto hace que el problema no sea para nada trivial.

La solución propuesta en este trabajo propone un esquema iterativo en el que se alternan una etapa de clasificación con otra en la que las volúmenes "crudas" son alineadas a las referencias, ver Figura 1.2. La idea es trabajar con un conjunto reducido de volúmenes que actúen como modelo de referencia. Dichas referencias serán obtenidas promediando grandes subconjuntos de imágenes y serán utilizadas para detectar las posibles conformaciones estructurales presentes en la muestra. Se las utilizará también para ir refinando la orientación de las partículas crudas de manera de mejorar la resolución alcanzada al calcular los promedios. Las referencias iniciales son calculadas a partir de las imágenes crudas en una etapa de inicialización muy similar a la que luego se usará para la clasificación en el "loop" iterativo que le sigue.

Tanto el problema del registrado de volúmenes como el de su clasificación de acuerdo a su conformación, han sido abordados en mayor o menor medida por la comunidad en los últimos años. En ésta tesis se presenta una revisión exhaustiva de los mismos y se proponen métodos originales donde se destacan profundas mejorías respecto de todos los hasta ahora publicados.

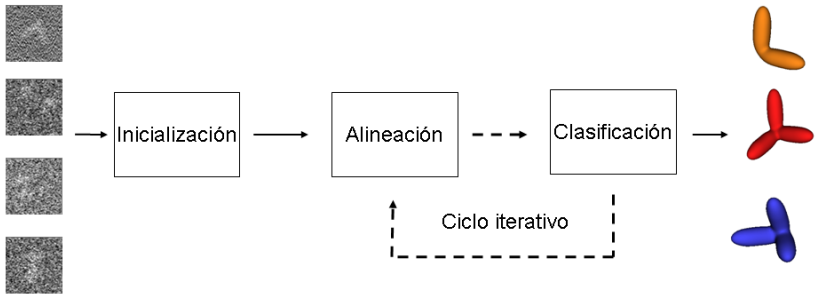


Figura 1.2: Diagrama de bloques general de la solución propuesta. El algoritmo toma volúmenes "crudas" de tomografía electrónica, con altos niveles de ruido para producir mapas tridimensionales de alta resolución. En el diagrama se muestran cortes de los volúmenes de entrada. Se emplea un método iterativo que tras una etapa de inicialización emplea alternadamente un bloque de alineación y otro de clasificación.

El objetivo principal de la solución propuesta pretende atacar el problema de la forma más general posible. Esto implica que la técnica utilizada no se basa o aprovecha ningún tipo de hipótesis que se crea deba cumplir la estructura o forma de la partícula de interés lo cual es práctica común en las soluciones previamente presentadas. Muchos trabajos recientes imponen por ejemplo ciertas propiedades de simetría a las partículas que se estudian durante el procesamiento e incluso a los resultados obtenidos. De esta manera se busca evitar influir el resultado obtenido dotándolo así de mayor credibilidad.

En el área de la biología estructural se han reportado diversos trabajos que utilizan herramientas de procesamiento de imágenes sobre las cuales se cuenta con pocos detalles técnicos. Más aún, no se presenta ninguna serie de pruebas y testeos que sirvan para validarlas. En el presente trabajo se intentó poner especial hincapié en éste punto en particular, por lo que los algoritmos propuestos fueron minuciosamente analizados y testeados utilizando datos artificiales generados para emular a las imágenes reales y datos reales adquiridos especialmente con este fin. La partícula elegida para realizar los experimentos fue una clase de la familia groEL. La elección está fundamentada en que la misma ha sido estudiada durante años y se tiene una idea muy precisa de su estructura a muy altas resoluciones. De esta manera se muestra que la herramienta puede ser aplicada con éxito para el análisis de imágenes reales.

El objetivo final que presigue el proyecto de colaboración mantenido entre los grupos del Prof. Sapiro y del Dr. Subramaniam es, una vez validada la herramienta, utilizarla para analizar material biológico hasta ahora nunca observado en altas resoluciones. En particular se pretende trabajar con muestras del virus de la inmunodeficiencia humana (VIH) así como del virus de la inmunodeficiencia en simios (VIS).

Un reporte intermedio del trabajo descrito en esta tesis fue publicado en el encuentro *IEEE International Symposium on Biomedical Imaging: "From nano to macro"* en abril de 2007 con el título de *Classification and averaging of electron tomography volumes* [3]. El trabajo completo será enviado a la revista *Journal of Structural Biology* en las próximas semanas.

1.2. Organización del texto

La presente tesis de maestría está organizada en ocho capítulos. En el Capítulo 2 se brinda una presentación del problema principal atacado en este trabajo. Dicho capítulo puede pensarse como una extensión de la Sección 1.1 donde ya se bosquejó una introducción de alto nivel para que el lector cuente con un paneo global antes de entrar en el cuerpo del texto. En la Sección 2.2 y en la Sección 2.3 se realiza una explicación detallada de las características técnicas que presentan las imágenes de tomografía electrónica y cryo tomografía electrónica. En particular se describen con precisión conceptos ya presentados de manera conceptual en la introducción, tales como el efecto del *missing wedge*.

El Capítulo 3 está dedicado a realizar una revisión exhaustiva de la literatura en el tema. En éste apartado se analiza primero en la Sección 3.1 los antecedentes en el área del análisis de partículas individuales y luego en la Sección 3.2 los antecedentes de trabajos realizados con imágenes tridimensionales de tomografía electrónica. Como ya fue mencionado, el análisis de las técnicas y resultados obtenidos en el área del análisis de partículas individuales será la piedra fundamental sobre la que se apoyará la solución propuesta en esta tesis. En ambas secciones se discuten las distintas alternativas utilizadas en cada uno de los trabajos allí referenciados y se postulan cuáles son sus ventajas y desventajas haciendo particular hincapié en aquellos aspectos que se intentarán mejorar con la solución aquí propuesta.

El Capítulo 4 está destinado al problema del registrado de volúmenes afectados por el efecto del *missing wedge*. En primer lugar se realiza en la Sección 4.1 una explicación de la medida de disimilaridad entre volúmenes propuesta incluyendo una interpretación intuitiva de la misma así como sus fundamentaciones matemáticas. Luego en la Sección 4.2 se describe la técnica de alineación desarrollada en este trabajo donde se destacan los principales aportes de la misma.

En el Capítulo 5 se ataca el problema de la clasificación de volúmenes de tomografía electrónica de acuerdo a su estructura, algo fundamental para lograr construir mapas de alta resolución. Allí se describe con detalle la solución encontrada.

Los dos capítulos anteriores están dedicados a explicar cada uno de los bloques fundamentales de la solución propuesta en está tesis, de acuerdo al diagrama que se muestra en la Figura 1.2. En el Capítulo 6 se explica al detalle la solución propuesta. En la Sección 6.2 se comentan otras alternativas analizadas durante el proceso de investigación y se explican las razones por las cuales éstas fueron luego descartadas.

Todos los experimentos realizados están concentrados en el Capítulo 7. En la Sección 7.2 se realizan distintas pruebas al algoritmo de registrado de volúmenes de tomografía electrónica propuesto. Se muestran la incidencia de los distintos parámetros con los que cuenta y otros diversos factores que pueden afectar su performance. Se realizan también comparaciones con otras técnicas existentes. Como ya se mencionó, se utilizaron dos grupos de datos: artificiales y reales. En la Sección 7.3 y en la Sección 7.4 se muestran los experimentos realizados a modo de validación de la técnica de clasificación presentada utilizando cada uno de dichos conjuntos y se analizan los resultados obtenidos.

En el Capitulo 8 se presentan las conclusiones finales del trabajo y se describen posibles caminos para continuar el trabajo en el futuro.

Capítulo 2

Tomografía electrónica

2.1. Breve reseña histórica

En los microscopios ópticos el nivel de magnificación alcanzable está limitado por la longitud de onda del haz con el que se irradia al espécimen. Consecuentemente queda también limitado el tamaño de los objetos observables utilizando dicha tecnología. Esto llevó a la búsqueda de radiaciones con longitud de onda inferior al de la luz visible que pudieran utilizarse para la generación de imágenes. El físico teórico francés Louis-Victor de Broglie realizó notables descubrimientos acerca de las propiedades ondulatorias de los electrones en la segunda década del siglo pasado. De Broglie, basándose en los recientes trabajos de Max Plank y Albert Einstein, probó en su tesis doctoral *"Investigaciones sobre la teoría cuántica"*[1] que electrones libres llevados a altas velocidades al atravesar un campo eléctrico se comportan, bajo determinadas condiciones, como una radiación ondulatoria y pasan a adquirir longitudes de onda fijas según la velocidad a la que se desplazan. Por primera vez de Broglie planteaba la dualidad onda-corpúsculo. La comunidad científica, con la casi exclusiva excepción de Einstein, se mostró escéptica frente a los resultados obtenidos por de Broglie, llegando incluso a ser tildada de descabellada por respetables investigadores de la época. Mirándolo retrospectivamente, la aprobación de uno de los más grandes científicos de la Historia inspiraría al menos cierta prudencia a la hora de criticar el trabajo del científico francés. En 1927 los científicos estadounidences Davisson y Germer alcanzan casi involuntariamente la confirmación experimental que probó verdaderas las conjeturas de de Broglie e implicó que se lo galardonara dos años más tarde, a los 36 años de edad, con el Premio Nobel de Física.

[1] El título original de la tesis doctoral de Luis-Victor de Broglie es *"Recherches sur la théorie des quanta"*.

Figura 2.1: (Izquierda) Louis-Victor de Broglie ganador del Premio Nobel de Fisica en el año 1929 por sus aportes a las física teórica. (Derecha) Ernst Ruska ganador del Premio Nobel de Fisica en 1986 por la construcción del primer microscopio electrónico y sus fundamentales aportes en el campo.

Los descubrimientos de de Broglie revelaban la posibilidad de utilizar haces de electrones como radiaciones de longitud de onda órdenes de magnitud menores que la de la luz visible y no tardarían en ser aprovechados para su uso en la microscopía.

El primer microscopio electrónico fue concebido y construido por los físicos alemanes Ernst Ruska y Max Knoll en el año 1931, tan sólo ocho años después que de Broglie presentara su tesis doctoral. El propio Ruska encabezó el grupo de científicos que produjo en 1939 el primer microscopio electrónico comercial para la también alemana compañía Siemens. Casi medio siglo después Ernst Ruska recibiría en Estocolmo el Priemio Nobel de Física por su fundamental contribución a la óptica electrónica. Hoy en día los microscopios electrónicos son de uso corriente en los laboratorios y gracias a ellos han sido posibles notables descubrimientos científicos. La magnificación obtenida con un microscopio óptico es aproximadamente de 1:1.000 mientras que para los microscopios electrónicos ésta aumenta dramáticamente a 1:1.000.000. Esto ha permitido observar material biológico a nivel macromolecular alcanzando resoluciones del orden de los $5 - 20nm$.

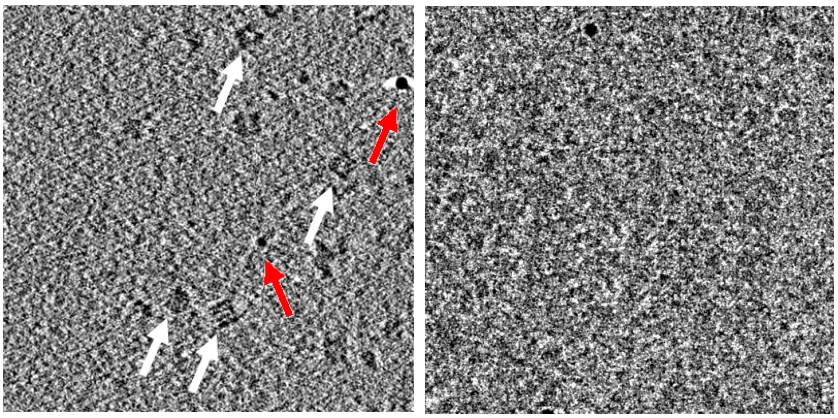


Figura 2.2: *Izquierda* un corte de una imagen tomográfica. Las flechas blancas muestran partículas de la macromolécula groEL y las flechas rojas muestran las partículas de oro utilizadas para lograr alinear las distintas proyecciones y conseguir así reconstruir la imagen tridimensional. *Derecha* Proyección individual correspondiente al tomograma mostrado en la imagen de la izquierda.

2.2. Cryotomografía electrónica

Existen diversos tipos de microscopios electónicos todos ellos basados en los mismos principios utilizados por Ruska y Knoll en su prototipo, que se distinguen por la forma en que se imprime la imagen en el microscopio. El modelo utilizado en este proyecto es el microscopio electrónico de transmisión, que es el de uso más corriente y el que fuera el primero implementado por Siemens. Las imágenes obtenidas son bidimensionales y corresponden a proyecciones del espécimen en la dirección determinada por el haz de electrones. Por lo tanto el nivel de gris de cada pixel de la imagen es proporcional a la integral de la densidad de la materia presente en dicha recta. La tomografía electrónica es una técnica que consiste en combinar una serie de proyecciones de un mismo espécimen sucesivamente rotado para de este modo obtener una imagen tridimensional. Como se mencionó en la sección anterior, para aprovechar las grades ventajas que ofrece la TE, es necesario pagar un costo en el incremento de los esfuerzos de procesamiento. En esta sección se explican brevemente los orígenes y las influencias de los mismos. En la figura 2.2 se muestra un ejemplo un corte en una lámina interior de imagen tomográfica así como una proyección individual.

En la TE el espécimen se rota para obtener las distintas proyecciones en lugar de rotarse el emisor como sucede en otros tipos de tomografía (como por ejemplo en la de rayos X), ver Figura 2.3. Esto produce dos problemas

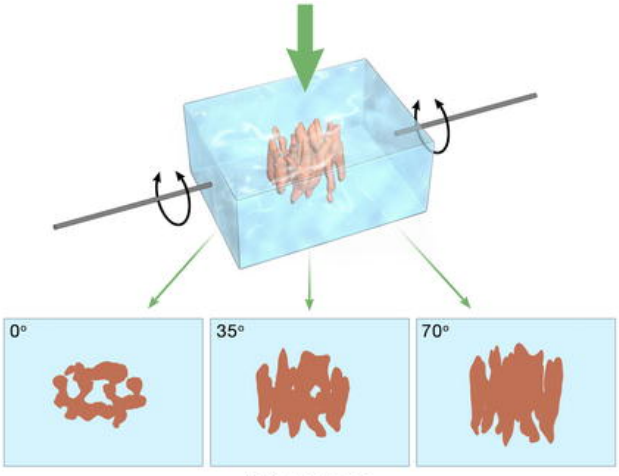


Figura 2.3: En esta figura se muestra gráficamente cómo se obtienen las imágenes en TE y CryoTE. Cada una de las proyecciones está asociada a un ángulo de rotación. Algoritmos de reconstrucción como retroproyección filtrada, SIRT o ART pueden ser usados para reconstruir la imagen tridimensional a partir de las proyecciones. Esta imagen fue tomada de [36].

importantes tanto en la reconstrucción del tomograma como en su posterior procesamiento. Al rotar el espécimen, este se desplaza inevitablemente de su posición original por lo que es preciso alinear las proyecciones obtenidas antes de realizar la reconstrucción. Al nivel de resolución con que se trabaja el más mínimo desplazamiento produce grandes desviaciones. Este problema es resuelto hoy en día mediante la inserción de partículas de oro en la muestra, que actúan como referencias para la alineación de las distintas proyecciones como se muestra en la figura 2.2. En este caso la solución origina otro problema ya que dichas partículas de oro afectan el contraste de la imagen, debido a la gran diferencia de densidad que presentan, y pueden interferir la correcta visualización del espécimen. El segundo problema y que es de mayor importancia, es que el espécimen no puede rotarse más de 60 o 70 grados respecto de la posición horizontal. El diámetro del haz de electrones es considerablemente más pequeño que el espécimen que se quiere analizar, por lo que cuanto mayor es la inclinación, mayor es el espesor de material atravesado por el haz de electrones. Por lo tanto las proyecciones obtenidas difieren de las proyecciones reales. En la Figura 2.5 se muestra una explicación gráfica de este fenómeno. Esto implica que sólo se tomen proyecciones con ángulos entre -60 y 60 grados de inclinación aproximadamente (en lugar de los -90 y 90 teóricamente necesarios). En el momento de la reconstrucción de la im-

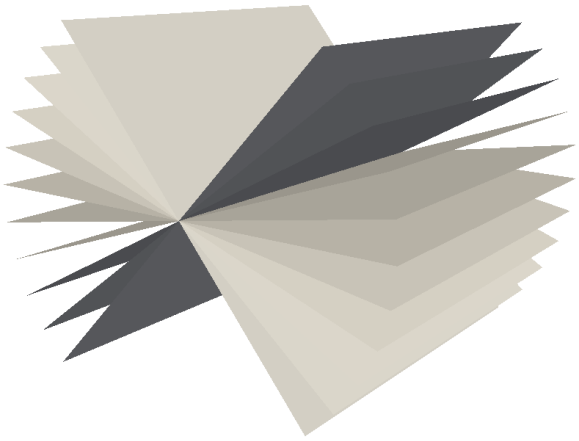


Figura 2.4: Datos faltantes en el dominio de Fourier para imágenes de tomografía electrónica, *missing wedge*. Los distintos planos de la figura corresponden a la transformada de Fourier de las distintas proyecciones tomadas. Puede observarse que faltan proyecciones.

agen 3D a partir de las proyecciones no se cuenta con toda la información necesaria, por lo que la imagen obtenida tiene "vacíos" de información, esto produce además un ruido fuertemente estructurado en la imagen denominado *missing wedge* o en español cuña faltante. Su nombre se debe a la forma que tiene la región de información faltante en el dominio de la transformada de Fourier, ver figura 2.4. De acuerdo al teorema de la sección central (ver por ejemplo [21]), cada proyección aporta información en el dominio de Fourier en un plano que pasa por el origen del espacio recíproco y es ortogonal a la dirección de incidencia del haz. Parte de lo que se busca en este trabajo es la compensación de dicho efecto, indispensable para la obtención de imágenes de alta resolución. En la Figura 2.6 se muestra cómo afecta este efecto a una imagen particular.

Como mencionamos antes, este proyecto se centrará en imágenes de Cryotomografía electrónica (CryoTE). Esta es una modalidad particular de la TE donde el espécimen es embebido en una solución acuosa y enfriado casi instantáneamente a muy bajas temperaturas. Esto permite visualizar a las macromoléculas en su estado natural. Esta técnica presenta sin embargo al menos dos importantes dificultades intrínsecas compartidas en cierto grado también por la TE. La primera de ellas es que las imágenes de CryoTE tienen muy bajo contraste. El origen de este fenómeno se debe principalmente a la poca diferencia que presenta la densidad de masa de los clusters de macro-

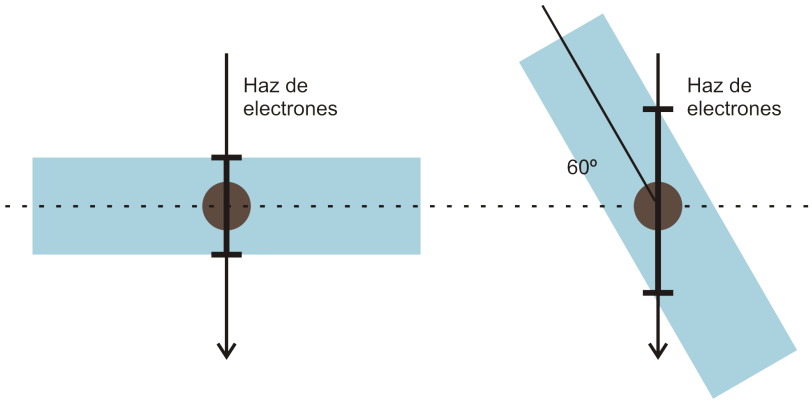


Figura 2.5: En esta figura se muestra un esquema de lo que sucede al rotar el espécimen. La banda celeste representa al espécimen y la flecha de color negro representa el haz de electrones que lo atravieza. A la izquierda se muestra al espécimen en posición horizontal en tanto que a la derecha se lo muestra con una inclinación de 60 grados. El espesor del material atravesado por el haz aumenta al doble.

moléculas y el medio en el que se encuentran. El segundo problema es que las muestras utilizadas son especialmente sensibles al haz de electrones con el que se las irradia ya que con bajos niveles energéticos los cristales formados por el hielo comienzan a romperse, esto limita la dosis total de irradiación que puede tolerar un espécimen.

Para obtener una reconstrucción de mayor resolución es necesario incrementar el número de proyecciones. Dado que se quiere limitar la dosis total de irradiación que afecte el espécimen para evitar su degradación, cuanto mayor es el número de proyecciones, menor es la dosis de electrones que puede utilizarse por proyección. Como consecuencia directa de la disminución de la dosis de electrones en cada corte, se obtiene un detrimento en la relación señal a ruido. Las imágenes tridimensionales obtenidas mediante CryoTE presentan una bajísima relación señal a ruido (como puede observarse en la figura 2.2 izquierda), al punto que en casi todas las ocasiones el ojo humano no logra identificar los detalles estructurales en las mismas. Puede encontrarse en [12, 20, 35, 37] una completa y reciente revisión tanto de TE como de CryoTE. Allí se explican más extensamente los problemas mencionados antes.

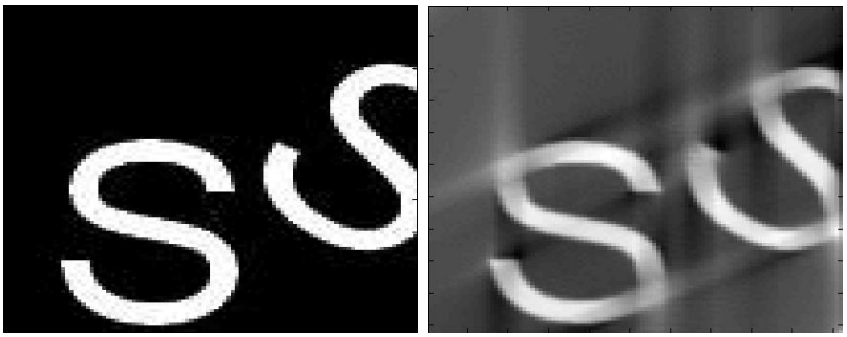


Figura 2.6: En esta figura se muestra como una imagen sintética es afectada por el efecto del *missing wedge*. Del lado derecho se muestra la imagen original, del lado izquierdo se muestra una reconstrucción de la imagen original a partir de proyecciones tomadas cada 2 grados en el intervalo de -60 a 60 grados y reconstruida utilizando el algoritmo ART. La imagen fue seleccionada con el propósito de mostrar claramente la presencia del ruido estructurado. Las imágenes volumétricas están compuestas por una pila de imágenes 2D de este tipo.

2.3. Tomogramas y subtomogramas

Consideraremos que un tomograma es entonces una imagen 3D proveniente en nuestro caso de CryoTE. Cada elemento de dicho bloque tridimensional es llamado voxel. Los voxeles en los tomogramas de TE representan típicamente volúmenes cúbicos de 1 a 4 nm de lado aproximadamente. Estos tienen asociado un valor en una escala de gris que corresponde a la densidad de masa del espécimen en esa región. Por lo tanto un mapa de alta resolución no es más que una imagen 3D libre de ruido que contenga toda la información de su espectro, es decir que no esté afectada por el efecto del *missing wedge*.

Asociado a cada tomograma se tiene un conjunto de información de la adquisición. Este permite conocer los ángulos en los cuales fueron tomadas las distintas proyecciones, lo que permite ubicar en cada imagen cuál es la información con la que no se cuenta, es decir la cuña faltante o *missing wedge*.

En general un tomograma corresponde a un volumen muchas veces mayor que el de la partícula que se quiere estudiar, es decir el objeto de interés. Por lo tanto se trabaja con subtomogramas. Estos son volúmenes mucho más pequeños que el tomograma y contienen a la estructura que se quiere observar (usualmente los tamaños utilizados rondan los $64 \times 64 \times 64$ vóxeles.).

Existen muy diversas técnicas para la detección de las partículas de interés en los tomogramas. La gama de algoritmos cubre todo el espectro, la (todavía muy usada) selección manual, algoritmos semi automáticos [10] y algoritmos de selección automática [8, 22]. En el presente trabajo suponemos que los subtomogramas conteniendo a las partículas se considerarán dados.

Debido a las diferentes posiciones relativas de los subtomogramas en el tomograma original, éstos son afectados de manera distinta por el efecto del *missing wedge*. Es decir, si se tiene dos partículas idénticas dispuestas en posiciones diferentes en el tomograma, estas tendrán un distinto conjunto de información faltante, ya que las proyecciones que se tomaron tienen orientaciones distintas en las coordenadas de referencia de cada volumen particular. Este hecho será explotado para compensar el efecto producido por la falta de proyecciones en la adquisición. Resulta intuitivo pensar que promediando un conjunto de partículas (correctamente alineadas entre si) con distintas regiones de información faltante, puede obtenerse un mapa que tenga información válida en todo el espectro. El cálculo de promedios de subtomogramas es una operación que requiere de cierto cuidado. Sean $V_1, \dots V_n$, subtomogramas, calcular directamente el promedio de los mismos implica,

$$P = \frac{1}{n} \sum_{j=1}^{n} V_j$$

Si así se hiciera se estaría incurriendo en un error. Esto puede verse fácilmente en el dominio de Fourier, haciendo uso de la linealidad tenemos,

$$\mathcal{F}(P) = \frac{1}{n} \sum_{j=1}^{n} F_j$$

donde F_j representa la transformada de Fourier del volumen V_j. En ese caso todos los coeficientes de Fourier de todos los volúmenes estarían aportando de igual modo al promedio lo cual es claramente incorrecto. Las regiones donde no se cuenta con información válida deberían ser dejadas de lado. Por lo tanto la forma adecuada de calcular el promedio es, para cada coeficiente de Fourier, promediar solamente los coeficientes de aquellos volúmenes para los cuales dicho coeficiente cae dentro de la zona de información válida. En esta tesis siempre que se haga referencia a el cálculo de un promedio se estará hablando de promedios calculados debidamente con el método descrito.

La correcta segmentación de las partículas y su extracción del tomograma son temas en los que trabaja el grupo del Dr. Sapiro y el Dr. Subramania

Figura 2.7: Se muestra gráficamente el proceso de enventanado de los subtomogramas.

[2, 1]. Este es un problema fundamental en la automatización del proceso da análisis. En nuestro proyecto consideraremos que la ubicación de las partículas de interés en el tomograma ya ha sido realizado y se cuenta con esa información.

Debido al fuerte ruido que afecta a las imágenes de CryoTE los subtomogramas tienen ser preprocesados antes pensar en realizar cualquier tipo de procesamiento con ellos. El preprocesamiento constituye en la eliminación del ruido presente en los píxeles como también el ruido biológico. El primero se logra aplicando a los subtomogramas un filtro pasabanda de manera de eliminar a la vez las componentes de alta frecuencia (normalmente dominadas por el ruido) y las componentes de muy baja frecuencia. Estas últimas son eliminadas debido a que los tomogramas pueden presentar enorme diferencias en sus niveles de gris presentando valores medios muy distintos. El ruido biológico es eliminado mediante el enventanado de los subtomogramas. Las ventanas utilizadas pueden ser muy diversas. La más comúnmente utilizada el simplemente una ventana esférica centrada en el centro de masas de la partícula de interés, ver Figura 2.7. De aquí en mas se supondrá en todo momento que los subtomogramas están filtrados y enventanados. Cuando sea pertinente se comentará sobre la influencia de los parámetros involucrados en dichos procesos.

Capítulo 3

Antecedentes

En este capítulo se presenta una revisión de la literatura del área del procesamiento de imágenes de tomografía y cryotomografía electrónica. Como se mencionó en la introducción, para entender las técnicas desarrolladas en dicha área, es preciso comentar primero sobre el trabajo hecho en el campo del análisis de partículas individuales. La Sección 3.1 está dedicada a esto.

Los primeros trabajos registrados en el área del análisis de partículas individuales datan de finales de la década del setenta. Desde entonces, un gran número de resultados han sido publicados, por lo que la tarea de realizar una revisión exhaustiva excede el objetivo introductorio que aquí se persigue. Si bien el área mantiene su efervescencia inicial y grandes avances siguen publicándose frecuentemente, hoy en día existen varios textos de estudio que resumen gran parte del trabajo realizado [12, 43]. Allí se encuentran un gran número de referencias donde el lector interesado puede profundizar en el tema. Un claro indicador del nivel de evolución que se ha logrado en esta área a lo largo de los años, es el gran número de paquetes de software que implementan este tipo de técnicas. Existen desde bibliotecas de muy diversas funciones implementadas en `Matlab` [23] hasta programas completos que contienen desde herramientas para la selección de partículas hasta los algoritmos completos de procesamiento [11, 32, 38, 40]. Muchos de ellos han editado ya varias versiones y cuentan interfaces gráficas amigables, lo cual tiene un impacto importante, ya que posibilita al biólogo correr los programas personalmente. Algunos de dichos paquetes de software cuentan con funciones o pequeñas bibliotecas pare el procesamiento de tomogramas, pero en ninguno de ellos se resuelve o se ataca explícitamente el problema que nos concierne en esta tesis. Por último es preciso aclarar que tampoco se comentarán en este texto aquellos trabajos del área de análisis de partículas individuales que se alejan de nuestra aproximación o no son aplicables al caso

tridimensional.

En la Sección 3.2 se realiza una revisión de los trabajos realizados en el área de tomografía y cryo tomografía electrónica. El análisis se concentra exclusivamente en aquellos trabajos que están muy relacionados o donde directamente se ataca el mismo problema que concierne en esta tesis: el de la generación de mapas de alta resolución a partir de imágenes tomográficas. En este caso se intenta hacer un análisis profundo de los principales trabajos publicados, marcando las distintas ventajas y puntos débiles que puedan tener. Se pretende establecer con claridad cuáles son los aspectos en los que se intentará realizar una mejora en la solución planteada.

3.1. Análisis de partículas individuales

3.1.1. Introducción

El número de partículas individuales cuyas estructuras tridimensionales ha sido develada mediante técnicas de cristalografía de rayos X está creciendo considerablemente. Sin embargo como se mencionó en la introducción (Capítulo 1, Sección 1.1) para poder utilizar dichas técnicas es necesario cristalizar los especímenes. Existe una vasta gama de proteínas que no pueden ser cristalizadas o que tienen un peso suficientemente grande como para dificultar seriamente la obtención de imágenes de resonancia magnética nuclear. Esto llevó al desarrollo de técnicas basadas en la microscopía electrónica. Las imágenes obtenidas con los microscopios electrónicos presentan una calidad (y consecuentemente una resolución) muy inferior a la obtenida mediante las técnicas mencionadas. Por esta razón fue necesaria la incorporación de complejos algoritmos de procesamiento de imágenes. Un caso particularmente estudiado, debido a su gran interés biológico asociado, es el de partículas individuales. A continuación se comentan distintos algoritmos de procesamiento de imágenes presentados en este campo.

Todos los algoritmos de procesamiento de imágenes utilizados en el análisis de partículas individuales persiguen el objetivo de construir un mapa tridimensional de alta resolución de una partícula dada. Las imágenes que se utilizan son bidimensionales y supondremos que se cuenta con un conjunto de "subimagenes" que contienen a las distintas partículas en analogía a los subtomogramas presentados en la Sección 2.3. Para evitar ambigüedades, la imagen completa (que contiene un gran número de partículas) se referirá co-

mo micrografo mientras que a las subimágenes (que contienen una única partícula) serán llamadas simplemente imágenes. Cada una de estas imágenes corresponde a una única proyección del tipo de las que se toman en la TE, con la ventaja de que pueden utilizarse dosis mucho mayores ya que justamente se toma solo una proyección. Bajo el supuesto de que las partículas se encuentran en orientaciones aleatorias[1], en el se encontrarán partículas en todas las posibles posiciones. Si se conociera cuál es el ángulo en que fue tomada cada una de dichas proyecciones, podría construirse un volumen tridimensional de la misma manera que se hace en la tomografía. Esto es justamente lo que hacen todos y cada uno de los algoritmos disponibles.

A pesar de poder utilizarse dosis mucho mayores de electrones para generar el micrografo, la relación señal a ruido que se observa en las imágenes es extremadamente mala. Por lo tanto resulta completamente imposible intentar hallar directamente los ángulos de las direcciones a las que corresponden las proyecciones de cada una de las imágenes. Para sortear dicho obstáculo, casi la totalidad de los algoritmos existentes dividen la solución en dos etapas. Primero se realiza una etapa en la que, combinando un gran número de imágenes en promedios, se mejora drásticamente la relación señal a ruido de varias proyecciones. Luego se realiza un segundo paso donde se encuentra a qué dirección corresponde cada una de las proyecciones obtenidas en el paso anterior, de manera de garantizar coherencia entre las mismas. Con esto se logra la reconstrucción del mapa tridimensional buscado.

El problema que se ataca en la primera etapa de los algoritmos utilizados en el análisis de partículas individuales es muy similar al enunciado en la Sección 1.1. Por otro lado, el cómo combinar las distintas proyecciones para así construir el mapa de alta resolución resulta un problema de poco interés en el contexto de esta tesis, ya que en TE y CryoTE los datos crudos son ellos mismos volúmenes individuales (subtomogramas). Por lo tanto en la Sección 3.1.2 se comentan únicamente la primera etapa de cada uno de los distintos algoritmos citados.

Recientemente ha sido publicado una técnica de análisis de partículas individuales [28, 29], inspirada en el notable trabajo presentado por Sigworth hace ya varios años [31], que realiza la reconstrucción sin utilizar el esquema de dos etapas mencionado previamente.. Los autores combinan los dos pasos mencionados en un esquema donde se busca determinar las orientaciones de

[1]Esta suposición no es siempre válida. Excede el objetivo de esta sección profundizar en las técnicas utilizadas en dichos casos.

las imágenes crudas maximizando una función de verosimilitud. El problema se resuelve utilizando el conocido algoritmo EM (*Expectation Maximization*) y se ha mostrado que en general tiene un desempeño superior a los esquemas previamente mencionados, aunque la diferencia es más apreciable en condiciones de ruido extremas [34]. Este tiene sin embargo un costo computacional mucho mayor y en el caso tridimensional se torna impracticable.

3.1.2. Clasificación de proyecciones

En la sección anterior se adelantó que todos los algoritmos en el área del análisis de partículas individuales cuentan con una etapa en la que se identifican y mejoran las diferentes proyecciones de la partícula de interés. Se comentó también que todas ellas utilizan el promediado como técnica de eliminación del ruido. Naturalmente, aquí se asume que el ruido que contamina a las imágenes es aditivo. Poco se sabe a ciencia cierta acerca de las características del ruido en microscopía electrónica [10, 12], principalmente debido a las complejas fuentes que lo generan. Sin embargo los experimentos y resultados obtenidos por la comunidad constituyen una prueba empírica de que la hipótesis de aditividad es sumamente razonable. También en el caso que nos concierne se adoptará esta hipótesis, como se explica en la Sección 5.1. Naturalmente los promedios sólo tienen sentido si las partículas se encuentran debidamente alineadas entre sí. Por otro lado está implícito que las distintas proyecciones deben poder distinguirse unas de otras. La alineación de las imágenes y clasificación de partículas son entonces una parte central de cada uno de los algoritmos que presentaremos a continuación y deja en evidencia la similitud existente entre éste y el problema que se ataca en esta tesis.

De forma similar a lo discutido en la Sección 1.1, el problema de la alineación y de la clasificación de las imágenes representa un problema del tipo del *huevo y la gallina* ya que para alinear las imágenes entre sí es preciso conocer cuáles corresponden a las mismas clases y, en sentido contrario, casi todos los métodos de clasificación exigen como hipótesis que las imágenes se encuentren alineadas.

Los algoritmos de clasificación de proyecciones pueden clasificarse en dos grande grupos, los basados en referencias y los libres de referencias. Como su nombre lo sugiere, en los primeros se cuenta con un conjunto de referencias iniciales mientras que en el segundo las referencias son estimadas utilizando el conjunto de imágenes crudas. Los métodos basados en referencias se utilizan

en muchos casos en los que se cuenta previamente con un mapa aproximado de la partícula individual que se está analizando. Esto puede deberse a que ya se cuente con un mapa tridimensional de media o baja resolución (en general obtenido utilizando otro tipo de tecnología o versiones más viejas de las herramientas de procesamiento de imágenes) y lo que se busca es hallar otro de mayor resolución. Puede también deberse a que se cuente con conocimiento biológico que implique cierta estructura aproximada para la partícula en cuestión. Las referencias utilizadas para la clasificación de proyecciones se obtienen de proyectar dicho mapa 3D según varias proyecciones [38]. Los algoritmos libres de referencias pueden ser de dos tipos, o bien no utilizan referencias en ningún momento o bien las estiman del conjunto de imágenes crudas.

Se han realizado varios estudios sobre la posibilidad de utilizar medidas invariantes respecto a traslaciones y rotaciones de las imágenes para realizar la clasificación [26, 27], en lugar de clasificar a las imágenes crudas directamente. De esta forma se desacoplan el problema de la alineación de las imágenes del de su clasificación. Sin embargo el nivel de ruido presente en las imágenes hace que los invariantes presenten, en muchos casos, muy poca información, afectando seriamente el resultado de la clasificación [12]. Este tipo de algoritmos es intrínsecamente libre de referencias. Una alternativa a esta solución es la iteración de pasos de clasificación y alineación de las partículas. La gran mayoría de los métodos existentes utilizan este tipo de esquema. A continuación se describen dos grupos de algoritmos diferentes pero sumamente ligados entre sí. El primero realiza las etapas de alineación y clasificación. El segundo puede pensarse como una versión sofisticada del primero y presenta una separación más clara entre los dos pasos que componen al algoritmo.

Uno de los primeros intentos para resolver el problema fue la utilización del método llamado "alineación mediante clasificación"(*"Alignment through classification"*) [12, 43, 42]. En este caso la clasificación y la alineación se realizan integradas en un esquema iterativo. En cada iteración se cuenta con un conjunto de imágenes de referencia que representan distintas proyecciones de la partícula tridimensional. Cada una de las imágenes crudas se alinea a cada una de dichas referencias, este proceso se lo refiere con el nombre de "alineación a múltiples referencias" (*"Multi-reference alignment"*). Las imágenes se clasifican de acuerdo a la distancia a la que se encuentran de (o similaridad que tiene respecto a) cada una de las referencias, en forma muy similar a lo que se hace en el conocido algoritmo *k-means*. Cada una de las imágenes crudas es asociada a la referencia que luego de la alineación tiene el puntaje

de similaridad mayor (la más "próxima"), obteniendo entonces un conjunto de imágenes asociadas por cada referencia. Dichos conjuntos son promediados para obtener así el nuevo grupo de referencias que será utilizado en la siguiente iteración. Para calcular el promedio, las partículas se disponen con la orientación que se determinó en el proceso de alineación. El algoritmo se itera hasta que no se observen cambios en la composición de las clases lo cual indica que se alcanzó la convergencia. A continuación se comentan algunos puntos interesantes sobre de éste método que servirán de base para la solución que se propone, para el caso tridimensional, en esta tesis. Como se mencionó antes, la clasificación se realiza de acuerdo a la similitud existente entre las partículas crudas y las referencias. Se han investigado un gran número de medidas de similaridad[2]. No se realizará en este texto un análisis de las mismas para el caso bidimensional, sin embargo se analiza en detalle las medidas de similaridad (o disimilaridad) para el caso tridimensional en el Capítulo 4.

El otro tipo de algoritmos que se comentará, alterna una etapa de alineación de imágenes (del tipo de la alineación a múltiples referencias) y otra de clasificación más claramente definidas que en el caso anterior [12, 13, 43]. La diferencia fundamental es que la clasificación no se limita a mirar únicamente la medida de similaridad entre partículas y referencias, sino que utilizan técnicas de clasificación más complejas. Existe en la literatura una amplia gama de combinaciones de medidas de similaridad y técnicas de clasificación. En esta sección se analiza una técnica en particular que ha resultado una de las más populares en el área.

En cada iteración, una vez realizada la alineación a múltiples referencias, las partículas son posicionadas de acuerdo a la orientación que se obtiene de alinearla a la referencia más cercana. Sobre este conjunto de imágenes (realineadas) es que se realiza la clasificación. En lugar utilizar directamente las medidas de similaridad calculada entre imágenes crudas, se realiza un análisis de componentes principales. Esta técnica se conoce en el área con la sigla MSA que viene de *Multivariate Statistical Analysis*. Cada imagen es expresada como una combinación lineal de los principales valores propios (imágenes propias) del conjunto. Por principales se entiende aquellos vectores propios con mayor valor propio asociado. El número de vectores propios utilizados siempre es mucho menor que el número de píxeles que tienen las imágenes, por lo que con esto se logra comprimir enormemente los datos. Dicho problema es crucial cuando se consideran números de imágenes importantes (del

[2]Se utiliza este término ya que las medidas de similaridad no son en general distancias.

orden del varios miles). Luego se utiliza alguna técnica de clasificación en el espacio de baja dimensión definido. Entre otras varias técnicas utilizadas una que goza de bastante popularidad es el agrupamiento jerárquico.

Para los dos algoritmos presentados, resulta natural preguntarse cómo se define la inicialización del algoritmo, es decir, cómo se obtiene el conjunto de referencias iniciales. La forma en que esto se realiza define si el algoritmo es o no "libre de referencias". Históricamente, el algoritmo de alineación mediante clasificación ha sido más utilizado como algoritmo basado en referencias. A continuación se comentan algunas técnicas empleadas para obtener el conjunto inicial de referencias a partir del conjunto de datos.

Existen diversos algoritmos utilizados para obtener las referencias iniciales a partir del conjunto de imágenes [24, 12, 43]. Inicialmente se consideran las imágenes crudas con su orientación inicial fija. Luego se aplica un algoritmo de clasificación estándar para encontrar subconjuntos de imágenes similares dentro del conjunto. Obsérvese que proyecciones que difieran únicamente en una rotación en el plano serán consideradas diferentes cuando en realidad no lo son y sería deseable incluirlas en el mismo promedio. Promediando los subconjuntos que se obtienen como resultado de esta clasificación es que se hallan las referencias deseadas. Claramente dichas referencias iniciales no tendrán una alta resolución pero servirán de semilla para el proceso iterativo descrito subsiguientemente. La premisa en la que se basan estas técnicas es la siguiente: si el número de imágenes es suficientemente grande, entonces la probabilidad de que varias partículas correspondientes a las mismas proyecciones se encuentren en exactamente la misma orientación será suficientemente alta.

Los algoritmos que usan MSA en su etapa de clasificación en las iteraciones, utilizan este paso de clasificación también en esta etapa inicial. En ese caso, la utilización de algoritmos jerárquicos en el espacio de baja dimensión resulta bastante atractiva ya que pueden obtenerse conjuntos compactos desde el punto de vista de la similaridad.

3.2. Antecedentes en tomografía

En esta sección se comentan los trabajos realizados en el procesamiento de imágenes de TE y CryoTE. Todos ellos presentan características comunes, por ejemplo todos utilizan alguna técnica de registrado de volúmenes. Es por esto que en lugar de analizar los distintos trabajos uno por uno, se analizarán

los distintos bloques que componen al algoritmo por separado, haciendo hincapié en los puntos que se intentarán mejorar en este trabajo. Primeramente se analizarán las diferentes técnicas de alineación de volúmenes que se han venido utilizando en la Sección 3.2.1 y luego en la Sección 3.2.2 se hará un análisis de las distintas aproximaciones a la solución del problema global.

3.2.1. Alineación de subtomogramas

El registrado de dos volúmenes consiste en hallar la rotación y la traslación que hace que éstos sean lo más similares posible. Esto es un problema de optimización en un espacio de seis dimensiones: los tres ángulos necesarios para determinar una rotación tridimensional y las tres componentes del vector de traslación. Por lo tanto resulta claro que la elección de la medida de similaridad entre volúmenes es fundamental para la correcta resolución del problema [9, 10]. Las técnicas de registrado tridimensional estándar, que no tienen en cuenta que los volúmenes en TE están afectados por el *missing wedge*, en general brindan resultados incorrectos [8]. Estas tienden a alinear las regiones de información faltante unas con otras. Los primeros trabajos presentados en el área comenzaron utilizando técnicas de registrado que no consideran el efecto del *missing wedge* [44], más tarde el mismo grupo de investigadores presentaría una versión mejorada de este trabajo [7], la cual también se comenta en esta sección. Walz et al. presentan un algoritmo que compara los volúmenes utilizando simplemente la correlación para determinar el puntaje de similaridad entre los volúmenes en las distintas orientaciones, siendo la que obtiene el mayor puntaje la elegida. La función de correlación se calcula directamente utilizando las imágenes. Sean $V, R \in \Re^3$ dos volúmenes (siempre se considerarán como funciones con dominio en $\Re^3$ y de soporte acotado) considerados en orientaciones fijas, su función de correlación está dada por,

$$CC(V,R) = \frac{\int (V(x) - \bar{V})(R(x) - \bar{R})\, dx}{\sqrt{\int (V(x) - \bar{V})^2\, dx}\sqrt{\int (R(x) - \bar{R})^2\, dx}} \tag{3.1}$$

donde $\bar{V}$ y $\bar{R}$ representan los valores medios de los niveles de gris de V y R respectivamente considerando únicamente la región de interés. El numerador corresponde a la función de correlación entre los volúmenes (considerados con media nula) y en el denominador se multiplican las desviaciones estándar de cada volumen. La normalización se debe a que los volúmenes pertenecientes a diferentes tomogramas (o incluso a distintas regiones dentro de un mismo tomograma) pueden presentar grandes diferencias en la media y la varianza

de su nivel de gris. El término de la media es muchas veces omitido ya que en general, como se comentó en el Capítulo 2, los subtomogramas son filtrados pasabanda para eliminar parte del fuerte ruido que los afecta y esto elimina la componente de continua.

La correlación dada en la Eq. (3.1) puede ser calculada para todas las posibles traslaciones, dada una rotación fija, muy eficientemente en el dominio de Fourier utilizando el teorema de convolución. Sin embargo, no existe un teorema similar que pueda aplicarse para determinar las rotaciones en el proceso de alineación. Esto ha llevado a que la utilización de la búsqueda exhaustiva (también llamado informalmente por el nombre más gráfico de "fuerza bruta") sea el método que goza de mayor popularidad. Simplemente se discretiza el espacio de las posibles rotaciones y para cada uno de los valores se halla la traslación óptima y se guarda la medida de similaridad. La rotación para la cual se tiene el mayor puntaje es la elegida. Si bien esta técnica garantiza la obtención de la terna de ángulos óptimos, tiene como contrapartida un costo computacional enorme que hace impracticable el cálculo de un número medianamente grande de alineaciones. Este fue uno de los principales puntos que se intentó mejorar en la solución propuesta en esta tesis.

Walz et al. proponen dos soluciones diferentes. La primera es muy sencilla y corresponde a realizar una búsqueda exhaustiva en las rotaciones. Este problema tiene las desventajas ya comentadas relativas al costo computacional. Plantean, como posibilidad para mejorar la eficiencia del algoritmo, realizar la búsqueda en dos escalas, primero utilizando una grilla ancha para luego realizar una búsqueda más fina en torno al óptimo encontrado en el primer paso.

La segunda solución propuesta por los autores es conocida por el nombre de "alineación polar" ("*Polar Alignment*"). La hipótesis fundamental asumida en la alineación polar, tiene que ver con el tener cierto conocimiento a priori acerca de la orientación en que se encuentran las partículas. Para ello, las partículas de interés deben ser alineadas manualmente o, en algunos casos particulares, éstas se organizan naturalmente en determinadas orientaciones preferenciales. Este es el caso, por ejemplo, de proteínas que se encuentran dispuestas en la membrana de células o virus. En ese caso se sabe que todas las partículas tendrán su dirección normal aproximadamente igual a la normal a la membrana de la célula o virus en el punto donde se encuentra dicha partícula. Por lo tanto, el problema del registrado tiene menor complejidad, ya que solo es necesario buscar cuál es el ángulo polar en todo el rango de variación, mientras que los restantes pueden solo calcularse en un entorno lim-

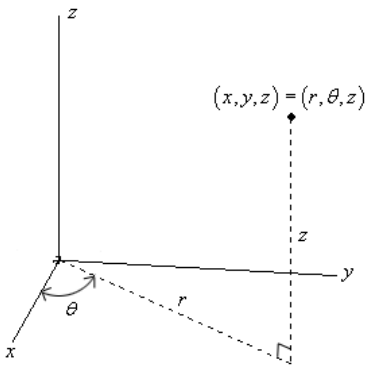


Figura 3.1: Coordenadas cilíndricas. En esta figura se muestra la relación entre las coordenadas cartesianas y las cilíndricas. Estas son usadas en dos de los algoritmos analizados en esta sección [7, 44]. El algoritmo presentado por Förster et al. utiliza un cambio a coordenadas cilíndricas para mejorar la eficiencia al recuperar el ángulo polar θ. Sin embargo esta técnica no considera debidamente en la función de similaridad que los volúmenes están afectados por el efecto del *missing wedge*.

itado de la orientación conocida a priori. El esquema de este nuevo algoritmo es idéntico al de la primera solución, solo que aquí se sustituye la búsqueda exhaustiva de grilla gruesa por una búsqueda alternativa más eficiente. Claramente en la primera solución podría adaptarse la búsqueda exhaustiva y considerar únicamente las regiones relevantes del espacio de rotaciones, dado el conocimiento que se tiene de las orientaciones de las partículas y de hecho es utilizada en otros trabajos [46]. Sin embargo sigue siendo muy costoso computacionalmente.

La idea es realizar una transformación de los volúmenes a coordenadas cilíndricas, como se ilustra en la Figura 3.1. Las rotaciones respecto del ángulo polar θ, en coordenadas cartesianas, corresponden a traslaciones en este nuevo sistema de coordenadas. Por lo tanto, aquí sí puede utilizarse el teorema de convolución para determinar el ángulo polar que mejor alinea a los volúmenes, tal como se hace en el cálculo de la correlación. Esto requiere hacer el cambio de coordenadas correspondiente en la función de correlación definida en la Ecuación 3.2. En este caso, en lugar de mirar en todo el volumen de correlación aquí tiene que considerarse únicamente la recta determinada por la intersección de los planos $z = 0$ y $\rho = 0$ en el nuevo sistema de coordenadas. Una vez hallado el ángulo polar se refina la solución utilizando la misma técnica que en la primera solución planteada por los autores. Si bien

esta solución es mucho más eficiente que la primera, comparte la desventaja de no considerar el efecto del *missing wedge* y requiere conocimiento a priori de la disposición de las partículas o bien la intervención manual.

El grupo de Hanspeter Winkler ha venido trabajando en el procesamiento de imágenes de TE y CryoTE durante algunos años. Recientemente fue publicada una revisión de todo el trabajo que han venido desarrollando [46]. Allí presentan un paquete de software integrado exclusivamente dedicado al procesamiento de imágenes de TE. En cuanto al problema del registrado, comentan una serie de diversas funciones de correlación que pueden ser utilizadas con un método de búsqueda exhaustiva para las rotaciones. Ninguna de dichas funciones considera el efecto del *missing wedge*, sin embargo sí mencionan los problemas en el resultado de la alineación que puede ocasionar el no considerarlo.

Frangakis et al. [8] definen una medida de similaridad que contempla el efecto del *missing wedge*. El caso que ahí se estudia es en realidad diferente al que concierne a esta tesis. El objetivo que persiguen los autores es el de encontrar un tipo de partícula determinado en un tomograma (en general ribosomes) a partir de un modelo conocido a priori. Básicamente es una versión particular del problema conocido con el nombre *template matching*. En dicho problema se comparan subtomogramas con modelos de referencia que tienen el espectro totalmente cubierto con información válida y consecuentemente no están afectados por el efecto del *missing wedge*. Los autores definen una nueva función de correlación que se calcula en el dominio de Fourier utilizando solamente los componentes de frecuencia que caen en la región de información válida del subtomograma. Esto permite calcular adecuadamente la similaridad para volúmenes provenientes de TE y CryoTE. La nueva función de correlación tiene una fórmula muy similar a la que se muestra en la Ecuación (3.1), aquí suponemos que V representa al subtomograma mientras que R es la referencia.

$$CC_f(V,R) = \frac{\int (V(x)-\bar{V})((R(x)-\bar{R}) \star M(x))\,dx}{\sqrt{\int (V(x)-\bar{V})^2\,dx}\sqrt{\int ((R(x)-\bar{R}) \star M(x))^2\,dx}} \tag{3.2}$$

donde M es el volumen cuya transformada de Fourier es una mascara que toma el valor uno en la región donde se tiene información válida y cero en el resto. La convolución de M con la referencia R en el dominio de la imagen implica que en Fourier se tenga el producto de sus transformadas. De esta forma intervienen en el cálculo de la correlación únicamente los valores de los

coeficientes de Fourier de la referencia que caen en la región de información válida que presenta el subtomograma V, asumiendo que son nulos los coeficientes de Fourier de la V en la región de información faltante. Esta función de correlación mantiene la propiedad de poder ser calculada eficientemente para todas las posibles traslaciones dada una rotación fija.

El trabajo presentado por Förster et al. [7] puede pensarse como una continuación de lo presentado por Walz et al. en [44]. Los autores se basan en el trabajo de Frangakis comentado arriba para mejorar el algoritmo de alineación propuesto anteriormente. Utilizan exactamente la función de similaridad definida en la Eq. (3.2) para aplicada al problema de la alineación de volúmenes de CryoTE. El hecho de que puedan hacerlo, como se detallará en la siguiente sección, se debe a que en el algoritmo para obtener el mapa de alta resolución que presentan, los subtomogramas siempre son alineanados a referencias conocidas de ante mano, las cuales no son afectadas por el efecto del *missing wedge*. También en este caso se presentan dos soluciones alternativas, cada una de las cuales corresponde a una mejora de una de las del trabajo previamente publicado. La primera solución consiste únicamente en realizar la búsqueda exhaustiva utilizando la nueva función de correlación. La segunda solución es también una adaptación de la alineación polar antes presentada. Sin embargo, en este sistema de coordenadas la función de correlación definida en la Ecuación (3.2) no puede calcularse eficientemente. Obsérvese que la máscara M que se le aplica a la referencia depende de la rotación. Los autores optan por sacrificar precisión en el cálculo de la medida de similaridad en favor de una mejora en la eficiencia del algoritmo y no utilizan la función de correlación introducida por Frangakis. En la etapa de refinamiento, donde se realiza una búsqueda "rotación por rotación" en torno a la estimación hallada en el primer paso, sí se considera la nueva función de correlación. Si bien este mecanismo es mucho más eficiente que el primero, no considera que los volúmenes tienen datos faltantes para hallar la primera aproximación del ángulo polar, pudiendo ocasionar los mismos errores en la alineación antes mencionados. La idea es que la solución propuesta en esta tesis logre realizar el registrado de subtomogramas eficientemente sin tener que caer en este tipo de problemas y que pueda utilizarse en cualquier caso sin tener que imponer condiciones sobre la orientación de las partículas.

Shmid et al. [30] proponen una medida de similaridad que es más general que la propuesta en la Eq. (3.2). Este caso se contempla la alineación de dos subtomogramas sin restringir a que uno de los volúmenes que participa en la alineación sea una referencia de espectro completo. La presentación de la técnica es muy poco formal y los autores no brindan explícitamente las

fórmulas utilizadas. Comentan que la correlación se calcula únicamente en la región de información valida que presentan en común ambas imágenes. Para determinar cuáles son las regiones de información faltante, los autores simplemente miran qué pixeles de la transformada de Fourier tienen valor cercano a cero. Si bien en general los valores que toman los coeficientes de Fourier en la región de información válida son mayores que los que toman en la región de información faltante, esto puede no ser cierto en todos los casos. Comentan también que la medida de similaridad está escalada por la cantidad de pixeles que caen en la región de información válida común a ambos volúmenes utilizados en el cálculo de la correlación. Esto es muy importante ya que distintas orientaciones relativas entre las partículas presentarán en general un número diferente de pixeles en la intersección de sus regiones de información útil. No considerar este escalado podría resultar en una técnica de alineación que priorice configuraciones con regiones de información válida común pequeñas.

Casi todos los trabajos mencionados utilizan las técnicas de alineación aquí comentadas para producir mapas de alta resolución. En la siguiente sección se analizarán las distintas alternativas que utilizan así como otras aplicaciones relacionadas publicadas recientemente.

3.2.2. Clasificación y promediado de subtomogramas

La generación de mapas de alta resolución en TE y CryoTE se obtiene del promediado de un gran número de subtomogramas precisamente alineados entre sí. Como ya se comentó, para ello es necesario realizar dos procesos íntimamente relacionados: la alineación y la clasificación. Todos los algoritmos que se analizarán en esta sección referencian a los trabajos realizados en el área de análisis de partículas individuales y en mayor o menor medida pueden relacionarse con una o varias de las técnicas presentadas en la Sección 3.1.2. En aquel caso, la clasificación aparecía naturalmente debido a que para lograr la reconstrucción tridimensional de una partícula se necesitan varias proyecciones distintas. En el caso de la TE la clasificación puede no parecer necesaria en algunos casos. Por ejemplo, si se intenta determinar el mapa de alta resolución de una partícula que se sabe tiene una conformación única. Sin embargo, cierto grado de clasificación es imprescindible para alcanzar alta resolución. Siempre existen partículas del conjunto de datos que están dañadas o seriamente afectadas y considerarlas para el cálculo de los promedios en general es contraproducente. En dichos casos, la clasificación puede pensarse como una selección de las mejores partículas. Por otro lado existen

muchas otras aplicaciones en donde no se cuenta con ningún tipo de información previa acerca de la partícula de interés. En ese caso, no incluir una etapa de clasificación en el algoritmo puede inducir a mapas equivocados, si la partícula tiene más de una conformación estructural posible.

De manera similar a lo que ocurre en el caso de la clasificación de proyecciones en el análisis de partículas individuales, los algoritmos de procesamiento de imágenes de TE y CryoTE pueden clasificarse en libres de referencias y basados en referencias. La diferencia entre unos y otros es la misma que en aquel caso los algoritmos basados en referencias requieren de un primer mapa tridimensional inicial mientras que los otros no utilizan más que el conjunto de subtomogramas. A continuación se analizarán varios algoritmos. En todos los casos ya se ha comentado en la sección anterior las técnicas de alineación que éstos utilizan.

Walz et al. [44] plantean un método basado en referencias. La referencia inicial se utiliza para alinear todas las partículas a un modelo común. Luego se realiza un paso de clasificación del estilo de MSA comentado en la Sección 3.1.2. Esta etapa pretende descartar *outliers* y encontar un subconjunto de partículas más "compacto" a partir del cual se calculará el mapa de alta resolución. El análisis de componentes principales allí utilizado no contempla que los volúmenes están afectados por el efecto del *missing wedge*. Por lo tanto, muchas de las imágenes propias (o volúmenes propios) capturan variaciones debidas a dicho efecto y no a variaciones estructurales como sería deseable. Los autores realizan un estudio de la influencia que tiene el resultado de la clasificación y concluyen que para el caso que les compete en esa publicación, no imposibilita alcanzar su objetivo aunque está notoriamente presente. Para mejorar las orientaciones de las partículas pertenecientes al subconjunto seleccionado, se utiliza un algoritmo iterativo, donde todas las partículas se alinean a la referencia con la que se cuenta en esa iteración y con las nuevas orientaciones se recalcula la referencia para la siguiente. Esto es muy utilizado y puede pensarse como el caso particular del método de alineación a múltiples referencias en que se tiene una sola referencia. Notar que en este proceso iterativo, la referencia inicial que se emplea no es la conocida a priori sino el promedio del subconjunto obtenido en el paso MSA. Por último los autores presentan una validación del método utilizando cubos como partículas de prueba.

El trabajo presentado por Winkler et al. [46] también emplea una algoritmo que itera el análisis de múltiples referencias con un paso de clasificación MSA. En este caso, la clasificación se realiza para detectar diversas confor-

maciones y no solo para detectar las partículas buenas como en el algoritmo previamente comentado. Aquí tampoco tienen en cuenta el efecto del *missing wedge* en la clasificación. Una vez que tienen la representación en el espacio de bajas dimensiones proporcionado por el MSA, ejecutan un algoritmo de aglomeración jerárquico mediante el cual se detectan los grupos de subtomogramas pertenecientes a clases diferentes. Esto tiene asociado intrínsecamente la fijación de umbrales de similaridad para diferenciar los grupos. Los autores no comentan mucho acerca de la elección de los mismo.

El trabajo desarrollado Winkler et al. ha sido utilizado en muy importantes investigaciones biológicas por otro grupo de investigadores [50, 51]. Los detalles que Zhu et al. brindan sobre el procesamiento de imágenes realizado es muy pobre. Los autores incluso utilizan variantes simplificadas del algoritmo propuesto por Winkler et al.. Por ejemplo, las partículas son alineadas manualmente para luego aplicarle una etapa de clasificación MSA con el único objetivo de encontrar un subconjunto que presente características de simetría que los autores intuyen deben tener las partículas. De los subconjuntos obtenidos luego de la clasificación aglomerativa en el esquema MSA, eligen aquel que presente mayor simetría triple. Luego realizan una serie de iteraciones de alineación a una única referencia para mejorar el mapa obtenido en la clasificación. En cada iteración se le impone simetría triple a la referencia que se utiliza en la alineación. En resumen, dichas aplicaciones utilizan herramientas de procesamiento de imágenes que presentan diversos problemas. El hecho de que no se considere el efecto del *missing wedge* para alinear los volúmenes ni para realizar la clasificación y la imposición de la simetría a las referencias hace que los resultados obtenidos hayan sido cuestionados [36].

El algoritmo que presentan Schmid et al. [30] es libre de referencias y está compuesto por dos partes claramente diferenciables. La primera es una etapa de inicialización y la segunda es muy similar a la alineación mediante clasificación. En la etapa de inicialización se busca obtener referencias iniciales para el proceso iterativo de la segunda parte del algoritmo. Debido a que la distancia se calcula utilizando búsqueda exhaustiva, el costo computacional es una verdadera limitante. En particular limita fuertemente el número de subtomogramas que es razonable procesar. En el trabajo citado dicho número es 160, lo cual es extremadamente chico. La etapa de inicialización está compuesta por dos pasos que se explican de manera poco formal. El primero pretende eliminar posibles *outliers*, es decir partículas dañadas o que son muy diferentes al resto de las imágenes del conjunto. Para ello se pretende analizar la matriz de similaridad que en cada componente tiene el valor de la medida de correlación que se obtiene al alinear dos a dos to-

das las partículas. Pero los autores comentan que esta opción es descartada debido al costo computacional que implica. En su lugar, dividen arbitrariamente el conjunto de datos en subconjuntos más pequeños y calculan las submatrices asociadas, estas sí de tamaño reducido (20×20 en la publicación citada). Luego se realiza un proceso iterativo donde se van agrupando progresivamente las partículas de acuerdo a su similaridad. Después de ciertas iteraciones todas las partículas que quedan fuera de los grupos formados son descartadas. Puede verse claramente que la eliminación de *outlier* es muy susceptible a la forma aleatoria en que se subdivide el conjunto de partículas. Por ejemplo, una (o varias) partícula puede(n) ser distinta(s) a las del subgrupo al(a las) que fue(ron) asignada(s), pero muy parecida(s) a muchas otras fuera de él. En el esquema descrito sería descartada. Una vez eliminados los *outliers* se realiza una clasificación de acuerdo al tamaño de las partículas. Esto se hace mediante la definición de un perfil unidimensional calculado directamente a partir del volumen. Si bien Schmid et al. consideran que los volúmenes están afectadas por el efecto del *missing wedge* para realizar el registrado, no lo consideran en este punto crucial de la etapa de inicialización. El valor máximo que alcanza dicho perfil representa una medida del tamaño de cada partícula y es en consecuencia invariante a rotaciones y traslaciones. Dividiendo arbitrariamente el rango en que varía la medida del tamaño, se divide el conjunto de partículas en nuevos subconjuntos. Las referencias se hallan al promediar estos subconjuntos y son utilizadas como referencias iniciales en un proceso del tipo alineación a múltiples referencias. Luego de varias iteraciones se consigue los mapas finales.

Durante el trabajo realizado en esta maestría se obtuvo como resultado intermedio una primera aproximación a la solución del problema [3]. Allí se propone una solución que combina una etapa de inicialización con otra iterativa. La inicialzación puede pensarse como una extensión del método propuesto por Schmid. En lugar del método de agrupación iterativa para la eliminación de *outliers* se utiliza un algoritmo de agrupación aglomerativa. La gran diferencia entre ambas soluciones radica en que, en esta solución las matrices de distancia pueden calcularse completamente llevando a conjuntos iniciales más significativos, debido a que las alineaciones pueden calcularse en forma mucho más eficiente. La técnica de alineación que se usó en [3] es la misma que se describe en el Capítulo 4. Cuando los conjuntos de datos crecen, los tiempos de cálculo necesarios para hallar la matriz de distancias se vuelven prohibitivos, a pesar de que el cálculo de la distancia es sumamente eficiente. Esto llevó a buscar una variación en la solución propuesta como se detallará más adelante. Luego de la etapa de clasificación se realiza un método conceptualmente similar al de “alineación mediante clasificación”, que pre-

senta varias diferencias respecto del propuesto por Schmid. El problema que se pretendía atacar es el más general posible, en donde no se tiene ningún tipo de conocimiento a priori sobre la partícula que se quiere estudiar y es el mismo que el que se ataca en esta tesis. No se pretende describir en detalle dicho algoritmo en esta sección, en la Sección 6.2 se comenta más a fondo sobre esta técnica y porqué se decidió cambiarla.

Capítulo 4

Alineación de subtomogramas

En este capítulo se presenta una técnica de alineación para volúmenes provenientes de tomogramas de TE y CryoTE. Como ya se ha comentado varias veces, estos están intrínsecamente afectados por el efecto del *missing wedge*. La idea es formular formalmente un algoritmo en el que la alineación pueda ser calculada en forma eficiente sin sacrificar precisión en el resultado.

En la Sección 4.1 se presenta una medida de disimilaridad para poder comparar subtomogramas adecuadamente. Poder comparar volúmenes afectados por el *missing wedge* es sumamente útil y su utilidad transciende al problema del registrado. Como se verá en el Capítulo 5 esta medida también será usada para clasificar subtomogramas. En esta sección se establecerá la terminología y nomenclatura que se utilizará en el resto de esta tesis. En la Sección 4.2 se ataca el problema de la alineación propiamente dicha. La solución propuesta está inspirada en el trabajo publicado por Kovacsa y Wriggers recientemente en el área de la cristalografía [16]. Este fue adaptado para así tener en consideración que los volúmenes están afectados por el efecto del *missing wedge*. La presentación se divide en dos partes, primero se presenta formalmente el problema de la alineación en la Sección 4.2.1 y luego en la Sección 4.2.2 se detalla el método de optimización empleado para su resolución.

4.1. Comparación de subtomogramas

En el Capítulo 2 se detallan las características que presentan los volúmenes obtenidos con tomografía de ángulo limitados. Se mostró que, de acuerdo al teorema de la sección central, cada proyección aporta información en un plano que pasa por el origen del espacio recíproco y es ortogonal a la dirección de incidencia del haz. Las proyecciones son tomadas usualmente en un

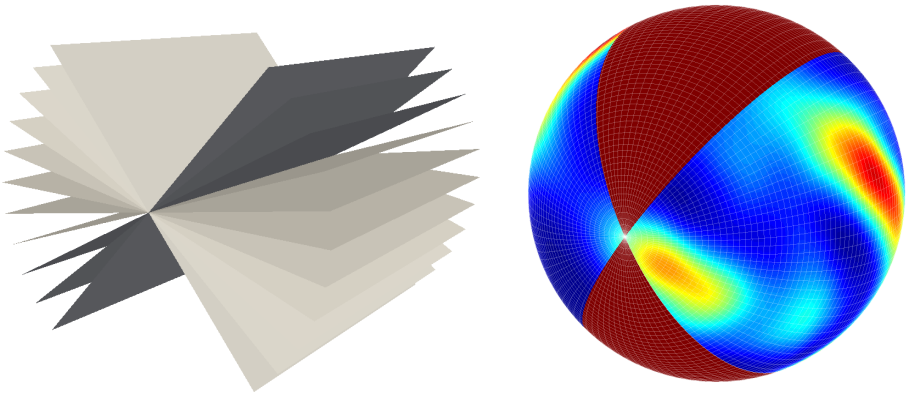


Figura 4.1: Las regiones de información faltante en TE y CryoTE tienen la forma mostrada en la imagen de la derecha. Al calcular su imagen esférica asociada ésta presenta también una región de información faltante claramente delimitada, como se muestra en la imagen de la derecha.

rango de ± 60 grados por lo que no se logra adquirir información en todo el espacio de Fourier. Por lo tanto, la región de información faltante tiene forma de cuña como se muestra en la Figura 4.3. En el dominio de Fourier entonces, resulta muy clara la división entre las regiones de información disponible y no disponible. En la imagen esto ya no es así, ya que la falta información afecta al nivel de gris de cada pixel. Por lo tanto, cualquier medida de similaridad que se calcule utilizando directamente los niveles de gris de los volúmenes está ignorando la presencia del *missing wedge*. Dicho de otro modo, cualquier medida de similaridad significativa tendrá que diferenciar claramente la información disponible de la faltante. En esta sección se describe la técnica que se empleará para llevarlo a cabo. A continuación se ilustrarán primero las ideas intuitivas en un ejemplo unidimensional, ver Figura 4.2, para luego extenderlos al caso de subtomogramas.

Sean r_1 y r_2 dos funciones unidimensionales definidas en un intervalo dado. En analogía a lo que sería el efecto del *missing wedge*, se supone que no es posible observar los valores que éstas toman en todo el intervalo, es decir que presentan regiones de información faltante. Supongamos que no se cuenta con información de la función r_1 en los intervalos $[A, B]$ y $[C, D]$ y para la función r_2, no se cuenta con información en los tramos $[E, F]$ ni $[G, H]$. Claramente estas funciones no pueden ser comparadas directamente. Si por ejemplo se empleara una norma cualquiera de funciones reales, se estarían teniendo en cuenta el valor de las funciones en los intervalos donde no se tiene

información útil lo que implicaría la obtención de una medida equivocada. La región en donde tiene sentido comparar las funciones corresponde a la intersección de los conjuntos donde ambas tienen información válida.

Para establecer un modelo matemático que permita representar estos conceptos, se define una máscara binaria, m_i con $i = 1, 2$, asociada a cada una de las funciones, que toma valor uno en los puntos donde se cuenta con información y cero en el resto, ver Figura 4.2. La región en que se dispone de información para ambas funciones puede representarse con otra máscara definida por la multiplicación de las máscaras de cada una de las funciones, $m_{12}(x) = m_1(x)m_2(x)$. De aquí en más se denominará a dicha región como la *región de solapamiento*.

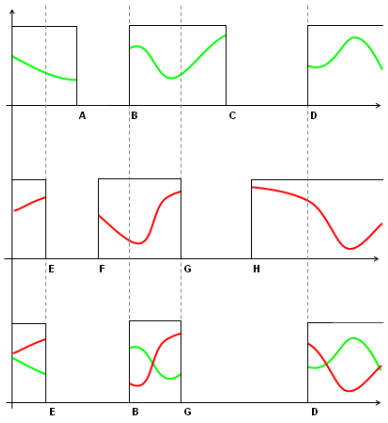
Utilizando esta representación puede utilizarse como medida de similaridad cualquier norma o métrica de funciones que compare a las funciones multiplicadas por la máscara m_{12}. Observar que en ese caso, el tamaño de la región de solapamiento pasa a jugar un papel importante. Si no se aplica ningún factor de normalización, cuanto menor sea la región de solapamiento menor será la medida de similaridad. Por lo cual tiene sentido pensar en una medida que esté normalizada por el tamaño de dicha región. A continuación desarrollaresmos estas ideas intuitivas para el caso de los subtomogramas.

Sea $V \in \Re^3$ un subtomograma con transformada de Fourier tridimensional $F : \Re^3 \to \mathbb{C}$. Este tiene asociada una región de información no disponible que está determinada por su orientación en el tomograma y por la cantidad de proyecciones que se usaron en la adquisición. En forma análoga a lo que se mostraba para el caso unidimensional, ésta puede representarse utilizando una máscara binaria $m : \Re^3 \to \{0, 1\}$ que toma el valor uno en la región de información válida y cero en el resto.

Sean V_1 y V_2 dos subtomogramas con transformadas de Fourier F_1 y F_2 y máscaras asociadas m_1 y m_2 respectivamente. En este caso también la región de solapamiento de información válida puede representarse por otra máscara binaria determinada por el producto de las máscaras individuales, esto es $m_{1,2}(x) = m_1(x)m_2(x)$. Se define como medida de similaridad entre subtomogramas a,

$$d(V_1, V_2) = \frac{1}{O} \int H(|F_1(x) - F_2(x)|)\, m_{1,2}(x)dx \tag{4.1}$$

donde $H : \Re \to \Re$ es cualquier núcleo de similaridad y O es el tamaño de la



$$\hat{r}_1 = r_1 \cdot m_1$$

$$\hat{r}_2 = r_2 \cdot m_2$$

Figura 4.2: Ejemplo unidimensional de funciones con datos faltantes. Las funciones r_1 y r_2 solamente pueden ser comparadas en aquellos intervalos en los que ambas tienen información válida , representados en la parte inferior.

región de solapamiento y está dada por,

$$O = \int B_f(x) m_{1,2}(x) dx$$

donde B_f es la función de transferencia del filtro pasabanda aplicado a los subtomogramas. Como se mencionó para el caso unidimensional, este factor se utiliza para poder comparar en forma justa la similaridad entre volúmenes con diferentes regiones de solapamiento. Conceptualmente al introducir este factor se está midiendo la diferencia que presentan en media las transformadas de Fourier de los volúmenes en la región de información común.

Aquí H puede tomarse de muchas maneras, obteniendo diferentes medidas con cada una de ellas. En esta tesis se eligió trabajar con el núcleo $H(x) = x^2$. En este caso particular, la medida de similaridad pasa a ser igual a la norma 2 del valor absoluto de la diferencia de las transformadas de Fourier en la región de solapamiento. Por lo tanto, la máscara m_{12} puede aplicarse equivalentemente dentro o fuera de la función de similaridad. Observar que la medida de similaridad definida no es una métrica ya que no se cumple la desigualdad triangular. Esto es bastante intuitivo ya que se tienen regiones de información faltante.

La medida de similaridad juega un papel fundamental en todo el procesamiento de subtomogramas. En la siguiente sección se la utilizará para plantear el problema del registrado. La medida de similaridad aquí presenta-

da tiene en cuenta que los mismos están afectados por el *missing wedge*.

4.2. Alineación

4.2.1. Planteo del problema

Dados dos subtomogramas, el problema de la alineación consiste en hallar la transformación rígida que maximize una determinada función de similaridad entre los mismos. En tres dimensiones esto representa un problema de seis grados de libertad ya que para determinar la transformación se necesitan tres ángulos de rotación y un vector de traslación de tres componentes.

Naturalmente, la elección de la función de similaridad juega un papel preponderante en el resultado de la alineación. En la Sección 3.2 se comentó en líneas generales que se ha constatado que las funciones de similaridad que no tienen en cuenta el efecto del *missing wedge* presentan sistemáticamente resultados incorrectos al alinear subtomogramas. En el Capítulo 2 se comentó que los volúmenes reconstruidos en TE y CryoTE no presentan cualquier tipo de valor en las regiones sin datos válidos. Estos valores son comúnmente pequeños, muy cercanos a cero. Al intentar registrar dos subtomogramas estas regiones de información faltante son consideradas como parte de las imágenes por lo que las orientaciones que las alinean entre sí presentarán un alto grado de similaridad. Esto explica la fuerte tendencia a obtener ese tipo de resultados constatada por ejemplo en [8]. Al utilizar la función de similaridad que se introdujo en la Sección 4.1 se estarán teniendo en cuenta solo los valores de los coeficientes de Fourier en la intersección de las regiones de información válida, por lo que este problema será totalmente evitado. En la Sección 7.2 se mostrarán distintos experimentos que se realizaron para estudiar este problema. A continuación se definirá en forma precisa el problema de la alineación de subtomogramas.

Sea $\Lambda_{\mathcal{R},\mathcal{T}}(V)$ el volumen que se obtiene de aplicar una rotación $\mathcal{R}$ perteneciente al grupo de rotacional en $\Re^3$, $SO(3)$, y una traslación $\mathcal{T}$ al volumen V. El problema de la alineación de dos volúmenes V_1 y V_2 es el de hallar la rotación y la traslación que minimice la disimilaridad entre los volúmenes V_1 y $\Lambda_{\mathcal{R},\mathcal{T}}(V_2)$,

$$[\mathcal{R},\mathcal{T}] = \underset{\mathcal{R}\in SO(3),\mathcal{T}\in\Re^3}{\arg\min}\ d(V_1, \Lambda_{\mathcal{R},\mathcal{T}}(V_2)). \tag{4.2}$$

donde d es la función de disimilaridad definida en la Ecuación 4.1.

La forma en la que se resuelve este problema de optimización es fundamental para poder garantizar la obtención de un mínimo global. La gran dificultad que esto lleva asociada ha hecho que la búsqueda exhaustiva sea el método más usado entre los investigadores, como se puede ver en la revisión bibliográfica que se presenta en la Sección 3.2. En la siguiente sección se detallará una técnica para resolver globalmente y en forma muy eficiente el problema de la alineación de subtomogramas haciendo uso de los armónicos esféricos.

4.2.2. Optimizacion

Esta sección está dedicada a explicar el método empleado para la resolución del problema de optimización asociado al del registrado de subtomogramas definido en la Ecuación (4.2).

El problema será atacado en dos pasos resolviendo en cada uno de ellos un problema de tres grados de libertad. La idea es recuperar la rotación en el primero de dichos pasos, trabajando con los módulos de las transformadas de Fourier ya que éstos son invariantes a traslaciones. A los efectos de acelerar dicho procedimiento, se trabajará con proyecciones esféricas de los coeficientes de Fourier de los volúmenes asociándoles de esta manera una función definida en la esfera a cada uno de ellos. De esta manera, el problema de hallar la rotación que mejor alinea un par de volúmenes se reduce al de hallar la rotación que mejor registra dos funciones esféricas. Esto no solo representa una reducción en la dimensión de los datos sino que puede ser resuelto en forma muy eficiente en el marco de la teoría de los armónicos esféricos.

Una vez hallada la rotación que mejor registra a los volúmenes se procederá con el segundo paso del algoritmo. Los parámetros que determinan la traslación óptima serán los que minimicen la función de disimilaridad definida en la Sección 4.1 considerando su orientación fija e igual a la dada por la rotación hallada en el primer paso del algoritmo. Esto puede hacerse en forma muy eficiente utilizando el teorema de convolución. A continuación se explicará en profundidad el método descrito.

Cálculo de la rotación

Sea V_1 una versión rotada y trasladada de otro volumen V_2, esto es

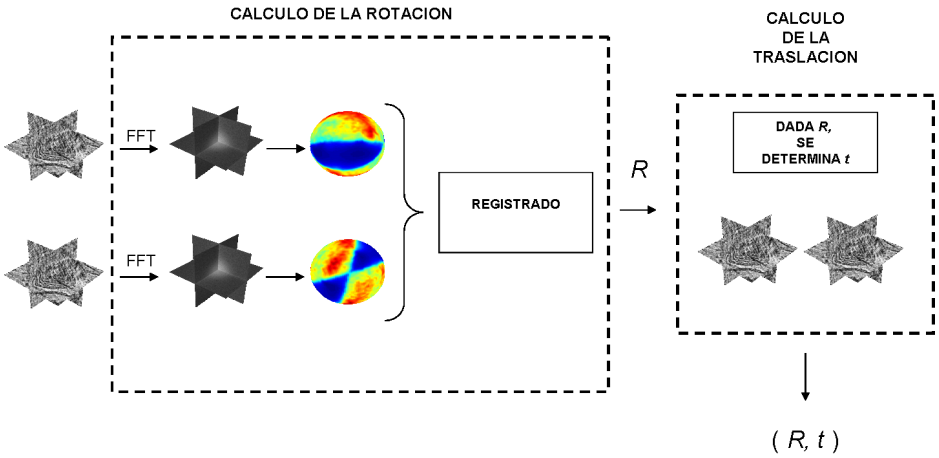


Figura 4.3: En esta figura se muestra un diagrama que muestra los pasos seguidos para la alineación de subtomogramas. Primero la rotación se recupera utilizando las representaciones esféricas. Luego la rotación hallada se considera fija y se utilizan los volúmenes completos para determinar la traslación.

$V_1 = \Lambda_{\mathcal{R}_0,\mathcal{T}_0}(V_2)$. Utilizando las propiedades de invariancia a las traslaciones y de rotación de la trasformada de Fourier puede probarse que los módulos de sus transformadas de Fourier se vincularán a través de exactamente la misma rotación, $|F_1| = \Lambda_{\mathcal{R}_0}(|F_2|)$. Tomando este resultado como base es que se define la forma en la que se recuperarán las rotaciones para el caso de los subtomogramas.

Sean V_1 y V_2 dos subtomogramas, la rotación que mejor los registra será la que minimice la medida de disimilaridad entre los módulos de sus transformadas de Fourier, esto es:

$$\mathcal{R} = \underset{\mathcal{R}\in SO(3)}{\arg\min} \frac{1}{O(\mathcal{R})} \int H(||F_1(x)| - \Lambda_{\mathcal{R}}(|F_2(x)|)|)\, m_1(x)\Lambda_{\mathcal{R}}(m_2(x))\, dx \quad (4.3)$$

con

$$O(\mathcal{R}) = \int B_f(x) m_1(x) \Lambda_{\mathcal{R}}(m_2(x)) dx.$$

Observar que la región de solapamiento depende de la rotación $\mathcal{R}$, por lo que su tamaño debe ser calculado para cada posible rotación. Claramente si se utilizara la búsqueda exhaustiva esto podría ser fácilmente tenido en cuenta.

Para evitar los altos costos computacionales que éste tiene asociado se explicará a continuación el método desarrollado que se basa en la utilización de los armónicos esféricos. A continuación se presentarán los conceptos fundamentales de la teoría de armónicos esféricos.

De ahora en más se referirá como S^2 al conjunto de puntos en $\Re^3$ pertenecientes a la superficie de la esfera unidad. En forma más general puede definirse $S^n = \{p \in \Re^{n+1} | \, ||p|| = 1\}$. Los puntos de la esfera serán parametrizados utilizando las coordenadas esféricas tradicionales (θ, ϕ), con $\theta \in [0, \pi]$ representando el ángulo polar (ángulo de colatitud) y $\phi \in [0, 2\pi)$ el ángulo acimutal (ángulo de longitud). Por simplicidad en la notación se adoptará la expresión $\eta = \eta(\theta, \phi)$ para referir a las coordenadas esféricas. Cada punto en la superficie de la esfera puede expresarse en forma unívoca de la forma,

$$\eta(\theta, \phi) = (\cos(\phi)\sin(\theta), \sin(\phi)\sin(\theta), \cos(\theta)). \tag{4.4}$$

Cada subtomograma V_i será asociado a una función escalar definida en S^2, $f(\eta)$, obtenida de integrar el módulo de la transformada de Foureier de dicho volumen, F_i, a lo largo de rayos trazados desde el origen de coordenadas.

$$f(\eta(\theta, \phi)) = \int_a^b F(\eta(\theta, \phi)t)dt \tag{4.5}$$

donde a, b son constantes positivas que definen el rango de frecuencias de interés, actuando como un filtrado pasabanda. De esta manera se logra eliminar las componentes de alta frecuencia que están muy contaminadas por ruido.

Las nuevas funciones definidas en la esfera serán utilizadas para recuperar la rotación. Esta representación tiene la gran ventaja de que de que las regiones de información disponible y faltante siguen quedando claramente diferenciadas. Dada la geometría de las regiones de información no disponible, si un punto pertenece a dicha región, entonces toda la recta determinada por este punto y el origen también pertenecerá. En la Figura 4.3 en la imagen de la derecha se muestra cómo quedaría la función esférica asociada a un subtomograma con las regiones de información disponible y faltante representada en la imagen de la izquierda de la misma figura.

Las funciones escalares definidas en la esfera serán representadas mediante imágenes esféricas como se muestra en la imagen de la derecha de la Figura

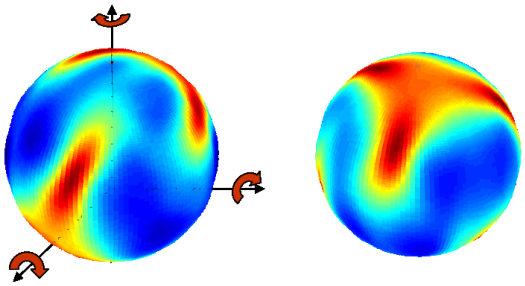


Figura 4.4: Alineación de imágenes esféricas.

4.3, por lo tanto serán referidas en muchas ocasiones como imágenes esféricas.

La región de información faltante en las imágenes esféricas puede también representarse con una máscara binaria definida en S^2 en forma análoga a lo mostrado en la Sección 4.1. Dicha máscara tomará el valor uno en los puntos donde se cuenta con información válida y cero en los demás.

La idea ahora es formular un problema equivalente al definido en la Ecuación (4.3) pero utilizando las funciones esféricas asociadas a los subtomogramas. Para ello es necesario contar con una medida de similaridad (o disimilaridad) que permita comparar funciones definidas en la esfera. En este caso también para poder compararlas en forma adecuada es necesario hacerlo únicamente en la región de solapamiento de información válida. Esto puede definirse de manera análoga a los casos anteriores utilizando otra máscara binaria dada por el producto de las máscaras asociadas a cada una de las funciones, es decir $m_{1,2}(\eta) = m_1(\eta)m_2(\eta)$.

Sean $f_1, f_2 : S^2 \to \Re$ dos funciones esféricas con máscaras esféricas asociadas $m_1, m_2 : S^2 \to \{0, 1\}$ respectivamente que definen la máscara de solapamiento $m_{1,2}$. Definimos su medida de disimilaridad como,

$$d_s(f_1, f_2) = \frac{1}{O} \int_{\eta \in S^2} (f_1(\eta) - f_2(\eta))^2 m_{1,2}(\eta) d\eta$$

donde O representa el tamaños de la región de solapamiento y está dado por,

$$O = \int_{\eta \in S^2} m_{1,2}(\eta) d\eta.$$

El problema de registrar dichas imágenes esféricas es simplemente hallar la rotación que mejor las alinee. La resolución de este problema tiene tres venta-

jas respecto del problema original utilizando los módulos de las transformadas de Fourier de los volúmenes: las imágenes esféricas son bidimensionales por lo que el costo computacional se reduce notoriamente, el proceso de integración necesario mediante el cual se obtienen las imágenes esféricas hace que las nuevas imágenes sean más robustas al ruido y, finalmente, al trabajar con imágenes definidas en S^2 pueden utilizarse todas las poderosas herramientas que brindan los armónicos esféricos que hacen que la alineación pueda calcularse en forma muy eficiente como se verá a continuación.

Ahora se expresará en forma matemáticamente precisa el problema de alinear dos funciones esféricas. Se denotará $\Lambda_{\mathcal{R}}$ al operador que aplica una rotación $\mathcal{R} \in SO(3)$ a una función esférica f. Utilizando esta notación el problema puede formularse así,

$$\mathcal{R} = \underset{\mathcal{R} \in SO(3)}{\arg\min}\, d_s(f_1, \Lambda_{\mathcal{R}}(f_2)) \tag{4.6}$$

donde las f_i con $i = 1, 2$ representan las funciones esféricas asociadas a los subtomogramas V_i obtenidas haciendo uso de la Ecuación (4.5). Nuevamente en este caso resulta claro que la región de solapamiento cambia con las distintas rotaciones y sin bien en este caso bidimensional la búsqueda por *fuerza bruta* es mucho menos costosa computacionalmente que en el caso tridimensional, ésta sigue representando un costo importante.

Recientemente ha sido publicado un estudio para el registrado de imágenes planas en presencia de máscaras de oclusión [6]. El caso que nos concierne es justamente éste pero tomando en consideración imágenes esféricas. El desarrollo que sigue toma como base y extiende el trabajo presentado por Fitch et al.. En este punto es donde se utilizará la expansión en armónicos esféricos para representar a las imágenes.

Sean f_1 y f_2 dos funciones definidas en S^2, su función de correlación se define como,

$$(f_1 \star f_2) = \int_{\eta \in S^2} f_1(\eta) f_2(\eta)\, d\eta.$$

Resulta útil definir una función que asocie a cada rotación $\mathcal{R}$ en el grupo de las rotaciones tridimensionales el valor que toma la función de correlación si se aplica dicha rotación a uno de los volúmenes, la referiremos como

$SCC : SO(3) \rightarrow \Re$ y está dada por,

$$SCC(\mathcal{R}) = (f_1 \star f_2)(\mathcal{R}) = \int_{\eta \in S^2} f_1(\eta) f_2(\mathcal{R}^T \eta) d\eta. \tag{4.7}$$

En la teoría clásica del análisis de Fourier, toda función real, periódica y cuadrático integrable puede ser escrita como una combinación lineal de las exponenciales complejas $\{e^{jk\theta}\}_{k \in Z}$. Este conjunto forma una base ortonormal de dicho espacio. Las funciones periódicas definidas en la recta real pueden pensarse como funciones complejas definidas en S^1, la circunferencia unidad. Los armónicos esféricos representan la extensión natural de las series de Fourier para mayores dimensiones. De manera análoga, los armónicos esféricos forman una base ortonormal del espacio de las funciones definidas en S^n. En particular, toda función esférica cuadrático integrable, definida en S^2, puede expresarse como series de la forma,

$$f(\eta) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \hat{f}_{lm} Y_l^m(\eta)$$

donde $Y_l^m(\eta)$ son las funciones de harmonizes esféricos y $\hat{f}_{lm}$ se denominan los coeficientes de la transformada de Fourier esférica (TFE) de f. Para una explicación detallada así como las fórmulas de las funciones de armónicos esféricos el lector puede referirse a [18].

Una función definida en la esfera se dice de ancho de banda B, donde B es un entero, si su expansión verifica,

$$f(\eta) = \sum_{l=0}^{B-1} \sum_{m=-l}^{l} \hat{f}_{lm} Y_l^m(\eta) \tag{4.8}$$

donde los coeficientes de la TFE cumplen $\hat{f}_{lm} \simeq 0$ para $l \geq B$.

En el espacio de las funciones definidas en el espacio de las rotaciones $SO(3)$ también puede definirse una transformada de Fourier, conocida por la sigla SOFT que proviene de su nombre en inglés $SO(3)$ *Fourier Transform*. Dichas funciones pueden también expandirse como series de funciones pertenecientes a una base ortonormal. Las constantes que multiplican a dichas funciones determinan los coeficientes de la SOFT de una función definida en $SO(3)$.

Existe un teorema que generaliza al teorema de convolución del análisis de Fourier para funciones reales que permite calcular en forma muy eficiente la función SCC definida en la Ecuación (4.7). Este permite obtener los coeficientes de la SOFT para la función SCC a partir de los coeficientes de la TEF de las funciones f_1 y f_2 que en ella intervienen [19].

Concretamente el teorema establece que los coeficientes de la función de correlación entre dos funciones definidas en la esfera, f_1 y f_2, para todas las posibles rotaciones en $SO(3)$, cumple que los coeficientes que determinan su SOFT se obtienen de multiplicar los coeficientes de la TEF de las funciones f_1 y f_2. Por lo tanto la función SCC definida en la Ecuación (4.7) puede hallarse calculando la TEF de cada función esférica y una SOFT inversa. Estas operaciones pueden calcularse muy eficientemente con algoritmos que serían equivalentes a la FFT para las funciones reales. Para obtener mayores detalles de la implementación mirar [15] y las referencias que allí se indican. En esta tesis se trabajó con la implementación que se encuentra disponible en la página web de los autores.

La idea fundamental será escribir el cálculo de la disimilaridad definida en la Ecuación (4.1) para todas las posibles rotaciones como una combinación de funciones SCC. Esto se obtiene muy fácilmente si consideramos la $d_s(f_1, \Lambda_{\mathcal{R}}(f_2))$ para cada posible rotación en $SO(3)$, esto es,

$$S(\mathcal{R}) = \frac{1}{O(\mathcal{R})} \int_{\eta \in S^2} \left[f_1(\eta) - f_2(\mathcal{R}^t \eta) \right]^2 m_1(\eta) m_2(\mathcal{R}^t \eta)\, d\eta.$$

Ahora, el tamaño de la región de solapamiento, $O(\mathcal{R})$, depende de la rotación considerada y está dada por,

$$O(\mathcal{R}) = \int_{\eta \in S^2} B_f(\eta) m_a(\eta)\, m_b(\mathcal{R}^t \eta)\, d\eta.$$

Desarrollando la integral dentro de $S(\mathcal{R})$ puede expresarse como la suma de tres funciones SCC:

$$S(\mathcal{R}) = \frac{1}{O(\mathcal{R})} \left([a^2 m_a \star m_b](\mathcal{R}) - 2[(a m_a) \star (b m_b)](\mathcal{R}) + (m_a \star b^2 m_b)(\mathcal{R}) \right). \tag{4.9}$$

Aplicando entonces los resultados de armónicos esféricos citados, puede calcularse la función de disimilaridad para todas las posibles soluciones sin

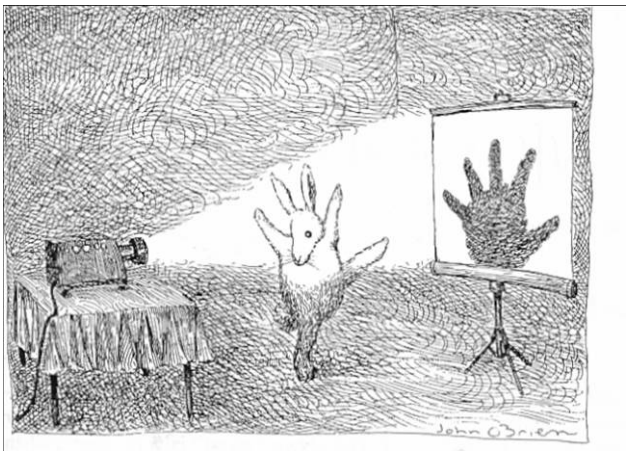


Figura 4.5: Dibujo realizado por John O'Brier, 1991 New Yorker Magazine. Muestra como a partir de una única proyección no puede inferirse el objeto observado. Esto puede suceder en forma menos divertida con las funciones esféricas.

la necesidad de caer en la utilización de la "fuerza bruta". Se obtiene por lo tanto un puntaje de disimilaridad por cada una de las distintas rotaciones en $SO(3)$, cada una de las cuales está asociada a la ternas ángulos que la determina. Puede entonces construirse un volumen de correlación en el cual cada eje corresponde a uno de los ángulos. Por lo tanto la rotación óptima, $\mathcal{R}_0$, se encuentra buscando el valor mínimo dentro del volumen de correlación.

Como se mencionó arriba, el hecho de estar comparando proyecciones de los volúmenes (funciones esféricas) en lugar de los volúmenes completos puede inducir a errores. Podría suceder que una rotación incorrecta produzca una proyección esférica más parecida a la proyección esférica del otro volumen que la real, pero la rotación real inevitablemente se dará en un mínimo local del volumen de correlación. Por lo tanto, para ganar robustez, se calculan los L menores mínimos locales dentro del volumen de correlación. Luego se comparan los puntajes obtenidos para dichas configuraciones, pero utilizando los volúmenes completos siguiendo la fórmula dada en la Ecuación (4.3). L es tomado en general entre 1 y 5, pero en la mayor cantidad de los experimentos realizados la rotación correcta coincidió con el mínimo global del volumen de correlación.

La precisión del resultado obtenido se debe principalmente al muestreo (o grilla) utilizado para definir las funciones esféricas y al número de coeficientes tomados para representar las TEF. En cualquier caso, la solución

obtenida a partir de las funciones esféricas no es tomada como definitiva. Se busca en un entorno de dicha solución para así refinar el resultado obtenido esta vez utilizando los volúmenes completos. Se utiliza el algoritmo de Powell [25], que es una versión mejorada del clásico algoritmo de optimización del gradiente conjugado, para minimizar localmente la función de disimilaridad dada en la Ecuación (4.3). El algoritmo de Powell no garantiza la obtención de un mínimo global, por lo que éste no podría usarse directamente. Una vez que el algoritmo de optimización converge se tiene la rotación óptima. Esta se considerará fija para la determinación de la traslación, como ya se adelantó en la Sección 4.2.

Cálculo de la traslación

La rotación hallada en el primer paso del algoritmo se considerará fija durante la segunda etapa, donde se recuperará la traslación óptima. El método que se empleará es igual al que se usa para hallar la traslación que maximiza la función de correlación de dos volúmenes a partir del teorema de la convolución. Esto se debe a la cercana relación existente entre la función de disimilaridad y la función de correlación.

Ahora se intentará hallar la traslación t que minimice la disimilaridad entre los volúmenes V_1 y $\Lambda_{\mathcal{R}_0,t}(V_2)$, donde $t \in \Re^3$ es el vector de traslación y $\mathcal{R}_0$ es la rotación hallada en el paso anterior. La función de disimilaridad definida en la Sección 4.1 para dichos subtomogramas puede reescribirse de la forma,

$$d(V_1, \Lambda_{\mathcal{R}_0,t}(V_2)) = \frac{1}{O} \int |F_1(x)m_1(x)\Lambda_{\mathcal{R}_0,t}(m_2)(x) - \Lambda_{\mathcal{R}_0,t}(F_2)(x)m_1(x)\Lambda_{\mathcal{R}_0,t}(m_2)(x)|^2 dx.$$

Aquí el tamaño de la región de solapamiento es constante, depende únicamente de la rotación y es considerada fija. Aplicando el teorema de Parsevall puede verse claramente que esta medida es equivalente a calcular,

$$d(V_1, V_2) = \frac{1}{O} \int \left(\tilde{V}_1(x) - \tilde{V}_2(x)\right)^2 dx,$$

donde $\tilde{V}_1$ y $\tilde{V}_2$ son los volúmenes que se obtienen de aplicar a los subtomogramas V_1 y $\Lambda_{\mathcal{R}_0,t}(V_2)$ los filtros con transformada de Fourier igual al producto de las máscaras respectivamente. Por ejemplo $\tilde{V}_1(x) = V_i(x) *$

$\mathcal{F}\{m_1(x)\Lambda_{\mathcal{R}_0,t}(m_2(x))\}$. Por lo tanto la rotación se halla resolviendo el problema,

$$\mathcal{T}_0 = \arg\min_{t\in\Re^3} \frac{1}{O} \int \left(\tilde{V}_1(x) - \Lambda_{\mathcal{R}_0,t}(\tilde{V}_2(x))\right)^2 dx.$$

Este problema de minimización es equivalente a resolver el siguiente problema de maximización,

$$\mathcal{T}_0 = \arg\max_{t\in\Re^3} \frac{1}{O} \int \tilde{V}_1(x)\Lambda_{\mathcal{R},t}(\tilde{V}_2)(x)dx$$

que no es otra cosa que la función de correlación entre los volúmenes $\tilde{V}_1$ y $\Lambda_{\mathcal{R},t}(\tilde{V}_2)$. Dicho problema puede ser resuelto utilizando el teorema de convolución en forma muy eficiente. Al hacerlo se consigue la transformación y la rotación que hacen que los subtomogramas sean lo más parecidos posible.

Capítulo 5

Clasificación

La idea fundamental de este trabajo es brindar un algoritmo de procesamiento de imágenes de TE y CryoTE capaz de producir mapas de alta resolución de partículas o arreglos de proteínas a partir de conjuntos de muestras individuales. Claramente el interés será mayor cuando el material biológico bajo estudio presente una estructura nunca antes visualizada en alta resolución. Por lo tanto como se mencionó en la Sección 2, es fundamental que el algoritmo no asuma propiedades de ningún tipo sobre las partículas que se estén estudiando. En particular el número de conformaciones que pueda presentar una determinada partícula no puede suponerse conocido. Muchos de los análisis realizados recientemente en la comunidad [50, 51] asumen que el objeto de estudio tiene una única conformación relevante cuando esto en realidad es desconocido [36]. El algoritmo propuesto en esta tesis consta de una etapa de clasificación donde se pretende poder separar a los subtomogramas de acuerdo a la conformación estructural que presenten.

La técnica de clasificación elegida considera a los subtomogramas en una orientación fija. En la Sección 3.2 se comentaron varias publicaciones recientes en las que se presentaron diversos tipos de algoritmos de clasificación de volúmenes de TE y CryoTE. La gran mayoría de ellos no tiene en cuenta las propiedades que estos tienen asociadas. En particular no tienen en cuenta el efecto del *missing wedge*. Por lo que muchas veces los resultados tienden a agrupar volúmenes de acuerdo a la posición relativa del missing wedge y no de su conformación estructural. En la Sección 5.1 se presenta un algoritmo de clasificación que se basa en la medida de la disimilaridad presentada en la Sección 4.1 la cual sólo tiene en cuenta la información comprendida en la región donde ambos tomogramas tienen información válida. Dicho algoritmo se basa en una técnica de clasificación jerárquica.

El algoritmo de clasificación propuesto requiere de un gran número de cálculos de la medida de disimilaridad presentada en la Sección 4.1. Si bien este no tiene asociado un costo computacional elevado, la combinación de muchos de ellos puede llegar a tener un costo muy importante. Una forma inteligente de reducir el costo computacional implicado en el cálculo del valor de la función de disimilaridad para un par de subtomogramas dado puede lograrse simplemente cambiando la forma en la que estos se almacenan. La idea será arreglar los datos de los subtomogramas para pasar de una representación volumétrica a una vectorial. En la Sección 5.2 se desarrollará esta idea y se describirán los distintos detalles de implementación.

5.1. Clasificación de subtomogramas

La idead del algoritmo de clasificación que aquí se presenta es dado un conjunto de partículas hallar subconjunto que presenten un mayor grado de similaridad. Por lo tanto el algoritmo tendrá como entrada un conjunto de subtomogramas que serán considerados con su orientación fija. Esto implica por ejemplo que si se cuenta con dos volúmenes idénticos que se encuentran orientados de manera diferente su medida de similaridad será baja y muy probablemente no serán agrupados juntos. El algoritmo que se presenta es del tipo jerárquico ascendente, este tipo de algoritmos son conocidos en la comunidad por la sigla que tiene su nombre en inglés (HAC) (*Hierarchical Ascendant Classifiaction*). La idea del algoritmo se detalla a continuación.

Supongamos que se tiene N subtomogramas. Inicialmente se tiene una partición del conjunto en N clases, $C_1, \ldots, C_N$, donde cada subtomograma pertenece a su propia clase unitaria. En cada iteración, los conjuntos más "cercanos" de la partición son agrupados, obteniéndose así una nueva partición con una clase menos. Luego de $N - 1$ iteraciones la partición queda compuesta por una única clase conteniendo a todos los subtomogramas. El orden en que los distintos conjuntos fueron agrupados a lo largo de las iteraciones permite contar con un orden de similaridad jerárquico del conjunto de datos.

Pueden utilizarse muchos criterios distintos para definir la distancia o similaridad entre conjuntos de subtomogramas. La elección de una determinada función de distancia representa un criterio de agrupación de clases y tienen un rol fundamental en el resultado obtenido. En general estos son definidos a partir de una métrica o función de disimilaridad definida entre los puntos

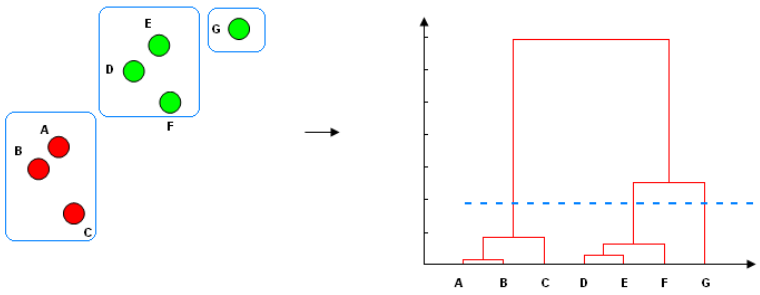


Figura 5.1: En esta figura se muestra un ejemplo de clasificación jerárquica para un caso artificial en $\Re^2$. Los puntos A, B, C, D, E, F y G (izquierda) fueron comparados con la distancia euclídea. La jerarquía obtenida en el proceso se muestra en el gráfico de la derecha. Al cortar el árbol a la altura indicada por la línea punteada se obtienen las clases $\{A, B, C\}$, $\{D, E, F\}$ y $\{G\}$

(en este caso subtomogramas). En este caso la medida utilizada es la función de similaridad definida por la Ecuación (4.1). Los criterios más populares para la agrupación de conjuntos en los algoritmos de agrupación jerárquica ascendente son: enlace simple o sencillo, enlace completo y método de Ward entre otros. Cada uno de dichos métodos tiene asociada una energía y el par de clases que se une es aquel que hace crecer menos al valor energético global [14].

El enlace simple o el del vecino más próximo simplemente define la di entre dos conjuntos C_i y C_j como la menor disimilaridad encontrada al compara los elementos de cada uno de los conjuntos, en este caso sería,

$$d_1(C_i, C_j) = \min_{V_h \in C_i, V_k \in C_j} d(V_h, V_k)$$

donde d es la medida de disimilaridad definida en la Sección 4.1. El criterio del enlace completo es opuesto al mencionado. La disimilaridad entre los conjuntos se define como la máxima disimilaridad existente entre elementos de cada uno de los conjuntos, es decir,

$$d_2(C_i, C_j) = \max_{V_h \in C_i, V_k \in C_j} d(V_h, V_k).$$

Ward [45] define una medida de qué tan buena es una determinada partición de un conjunto de acuerdo a qué tan parecidos son los miembros de los subconjuntos a su centroide. Está medida está dada por,

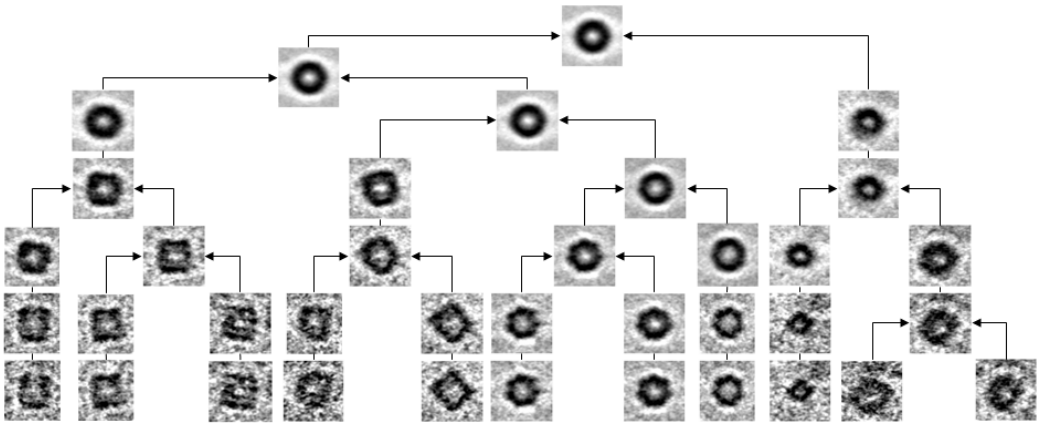


Figura 5.2: En esta figura se muestra una versión simplificada del árbol jerárquico obtenido al analizar un conjunto de datos de groEL. Este forma parte de uno de los experimentos realizados en el Capítulo 7.

$$SSE = \sum_{j} \sum_{V_i \in C_j} d(V_i, \bar{C}_j)$$

donde $\bar{C}_j$ representa el promedio de los elementos del subconjunto C_j. En tanto que el costo de juntar dos subconjuntos dados está dado por,

$$\Delta = \sum_{h} (V_h \in C_i U C_j) d(V_h, \bar{C}_{ij}) - \sum_{h} (V_h \in C_i) d(V_h, \bar{C}_i) - \sum_{h} (V_h \in C_j) d(V_h, \bar{C}_j)$$

que representa el incremento en el valor de la función SSE asociado a agrupar los subconjuntos C_i y C_j.

En el algoritmo diseñado el criterio de unión de clases puede ser variado a gusto del usuario. En general los resultados obtenidos muestran que la utilización de el método del enlace sencillo tiende a generar una clase mayoritaria que domina a las demás. Para los métodos de Ward y de enlace completo se obtuvieron los mejores resultados, siendo el primero el más comúnmente empleado.

Los resultados obtenidos en la clasificación jerárquica pueden observarse en un tipo de gráficos llamados dendrogramas. Este tipo de representación permite visualizar el "árbol" de una jerarquía dada. Cada nodo al pie de dicho árbol representa a una partícula individual, Cuando dos partículas o grupos de partículas son agrupados los trazos se unen formando una única

línea. La altura a la que se dibujan las uniones corresponde al valor de la disimilaridad que presentan los grupos unidos. Al cortar el árbol en un nivel dado se obtienen conjuntos que presentan una disimilaridad interna menor al umbral utilizado. En la Figura 5.1 se muestra un ejemplo artificial donde se agrupan puntos del plano. Con la línea punteada se representa el nivel de disimilaridad al que se cortó el árbol. Pueden utilizarse otros criterios alternativos para cortar el árbol. Por ejemplo puede decidirse cortar el árbol jerárquico cuando se tenga una partición con un cierto número de subconjuntos. En el Capítulo 7 se muestran varios experimentos en los que se evaluó el desempeño de este algoritmo de clasificación.

Obsérvese que el algoritmo de clasificación requiere del cálculo de la disimilaridad para todos los posibles pares de subtomogramas, lo que significa $N(N-1)/2$ cálculos. En la siguiente sección se describe la técnica utilizada para acelerar el proceso de cálculo de estas distancias.

5.2. Reducción de dimensiones

El costo computacional asociado a la clasificación de subtomogramas es en general elevado debido al gran número de cálculos de distancias que éste implica. Esto llevó a la búsqueda de representaciones alternativas de los datos que permitieran calcular las distancias en forma más eficiente (lo cual tiene sentido siempre que el cálculo de dicha representación tenga menor complejidad que el cálculo de las disimilaridades). En el área del análisis de partículas individuales se utilizó ampliamente el análisis de componentes principales o MSA [12, 43]. Sin embargo esto no puede aplicarse para el caso de los subtomogramas ya que estos estan afectados por el efecto del *missing wedge*.

Los subtomogramas tienen por lo general un número de vóxeles mucho mayor a la dimensionalidad de los datos. Esto puede verse muy claramente en el dominio de Fourier. Como se comentó en la Sección 3.2 los volúmenes deben ser filtrados pasabanda de manera de eliminar las componentes de ruido dominantes en las altas frecuencias. Al no considerar a muchos de los coeficientes de Fourier implícitamente se está reduciendo la dimensionalidad de los datos. Por lo tanto operar con los volúmenes completos implica un gasto de recursos computacionales innecesarios. En la Figura 5.3 se muestra gráficamente dicho proceso. A cada subtomograma se le asocia un vector que contiene un arreglo lineal de los coeficientes útiles del espectro. La dimensión del vector asociado es en general mucho menor que la cantidad de píxeles en

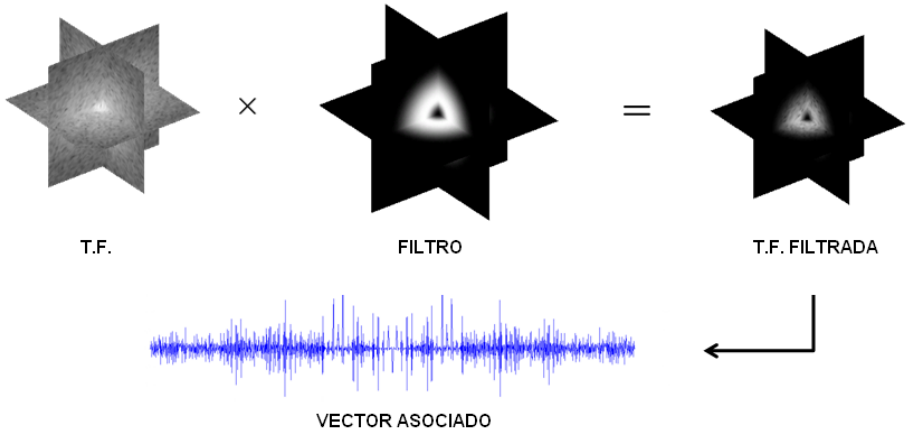


Figura 5.3: Método de reducción de dimensiones. Este simplemente se basa en construir un vector con los componentes relevantes del espectro del tomograma. La transformada de Fourier (T.F.) de un subtomograma es filtrada con un filtro pasabanda. La transformada de Fourier del volumen filtrado consta de una cantidad de coeficientes distintos de cero mucho menor al tamaño del volumen original.

los subtomogramas lográndose reducir hasta 100 veces la cantidad de componentes. Esto tiene asociado un incremento importante de la eficiencia del algoritmo de clasificiación.

Naturalmente los vectores asociados a los subtomogramas también tendrán datos faltantes. En este caso, la disimilaridad entre vectores se calcula también empleando únicamente los elementos que caen en la región de solapamiento con la ayuda de máscaras binarias (vectoriales), de la misma forma en la que se hacía antes. De hecho la situación es muy parecida al ejemplo unidimensional utilizado para ilustrar los conceptos en la Sección 4.1. El valor de la disimilaridad calculada utilizando su representación vectorial o volumétrica es exactamente la misma.

Capítulo 6

Solución propuesta

En este capítulo se presentará la solución propuesta para el problema que se estudia en esta tesis, la cual se describe en la Sección 6.1. El algoritmo implementado tendrá como entrada un conjunto de subtomogramas crudos y dará como salida un conjunto de mapas de alta resolución con una cantidad de elementos en principio desconocida. Se intentó que el presente capítulo sea autocontendido, es decir que muchos de los conceptos que aquí se desarrollarán ya fueron mencionados a lo largo de esta tesis y serán nuevamente comentados en busca de una mayor claridad en la explicación.

Durante el desarrollo del trabajo que aquí se documenta se evaluaron una gran cantidad de alternativas diferentes. En particular se llego a proponer una solución global [3] que mostró tener un muy buen desempeño al ser evaluada en experimentos con datos artificiales. Sin embargo al testearla con los datos reales su desempeño no fue el esperado, los resultados eran fuertemente dependientes de los parámetros escogidos y demandaba un tiempo de computación excesivo. En la Sección 6.2 se profundizará sobre esta alternativa y las razones por las cuales fue descartada.

6.1. Algoritmo final

Cuando se quiere visualizar por primera vez la estructura de una cierta partícula de la cual se quiere entender su función, en general se tiene poca información previa. En base a otros casos estudiados muchas veces puede intuirse algún tipo de características de la misma. Muchas veces el asumir que la partícula de interés presenta ciertas propiedades puede simplificar mucho la dificultad del problema de procesamiento de imágenes asociado. En

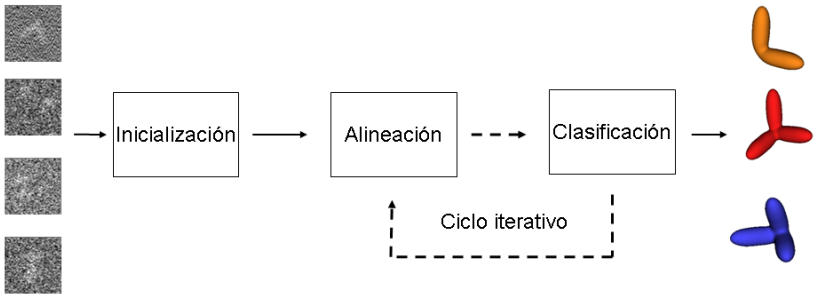


Figura 6.1: Diagrama de bloques general de la solución propuesta. El algoritmo toma imágenes "crudas" de tomografía electrónica, con altos niveles de ruido para producir mapas tridimensionales de alta resolución. Se emplea un método iterativo que tras una etapa de inicialización emplea alternadamente un bloque de alineación y otro de clasificación.

este punto existe un compromiso delicado. Se han hecho diversas pruebas que muestran que muchas veces los algoritmos de procesamiento de imágenes que buscan obtener mapas de alta resolución haciendo uso de información a priori (como la imposición de ciertos tipos de simetría) muchas veces encuentran lo que buscan aunque eso no esté presente en los datos [31]. Otro tipo de simplificación común es asumir una única conformación para la partícula de interés. Para evitar estas posibles fuentes de error se presentará un algoritmo totalmente general que ataque el problema más amplio posible. Esto no necesariamente implica que todo el algoritmo sea completamente automático. Dada la importancia de la aplicación con la que se trabaja, es preferible desarrollar un algoritmo que garantice un buen funcionamiento y requiera de una muy moderada supervisión por parte del biólogo. A continuación se detallarán las características del mismo.

El problema de la clasificación y la alineación de subtomogramas representa un problema del tipo del *huevo y la gallina*. La clasificación de los volúmenes en conjuntos homogéneos es crucial para poder asegurar que éstos sean calculados a partir de subconjuntos de partículas idénticas[1] para lograr así alcanzar los mapas de alta resolución buscados. Si se contara con dicha clasificación de antemano, alinear a los distintos subtomogramas para obtener estos promedios sería un problema fácil de resolver. Por otro lado clasificar

[1]Siempre algún grado de heterogeneidad se pierde en pro de aumentar la relación señal a ruido.

los volúmenes una vez que estos están debidamente alineados unos con otros puede realizarse con grandes esperanzas de éxito empleando, por ejemplo, el método de clasificación presentado en la Sección 5.1. Por lo tanto el algoritmo presentado hará uso de un esquema iterativo que alterne pasos de clasificación y alineación sucesivamente como es detallará en la Sección 6.1.2.

En el paso de alineación se hace uso de un conjunto de modelos de referencia contra los cuales se registran los volúmenes crudos. Dichas referencias serán obtenidas mediante un algoritmo de inicialización directamente a partir del conjunto de subtomogramas crudos. Luego éstas serán utilizadas como semillas en el "loop" iterativo subsiguiente. El paso de alineación será muy similar a la técnica de alineación a múltiples referencias visto en el Capítulo 3.1 mientras que el paso de clasificación será el presentado en el Capítulo 5. En la Figura 6.2 se muestra un diagrama de alto nivel de la solución propuesta. En la Sección 6.1.1 se describe el paso de inicialización empleado y luego en la Sección 6.1.2 se describe el ciclo iterativo.

6.1.1. Inicialización "libre de referencias"

El objetivo de la etapa de inicialización es brindar al ciclo iterativo un conjunto de volúmenes de referencia que puedan ser utilizados en el la alineación a múltiples referencias que se describe en la siguiente sección.

Los subtomogramas se considerarán en su posición original, tal cual fueron extraídos del tomograma. Con lo cual éstos estarán orientados arbitrariamente. Estos serán pasados como conjunto de entrada al algoritmo de clasificación jerárquico descrito en el Capítulo 5. Por lo tanto volúmenes que contengan partículas con idéntica conformación pero ubicados en orientaciones diferentes serán considerados distintos.

Este algoritmo se basa en que si se cuenta con un gran número de partículas la probabilidad de que por coincidencia muchas de ellas estén en la misma posición relativa crece. Por lo general el criterio utilizado en el paso de inicialización para cortar el árbol jerárquico que se obtiene de la clasificación es diferente del usado en el paso de clasificación del ciclo iterativo. Si se elige cortar el árbol utilizando un cierto umbral de disimilaridad éste no puede ser muy exigente, ya que es de esperar que las partículas no estén muy próximas unas de otras. Si se optara por cortar el árbol cuando se alcance un número de clases dado, este debe elegirse prudentemente grande. Si no se hiciera así es de esperar que el algoritmo iterativo requiera de un número mayor de

iteraciones para alcanzar la convergencia.

Si el número de partículas con las que se cuenta es reducido, podría llegar a suceder que los subconjuntos de partículas que se encontraban originalmente en una misma posición sean muy pequeños. La experiencia muestra que en ese caso resulta mejor utilizar referencias iniciales que contengan un mayor número de volúmenes mejorando la relación señal a ruido (por la disminución de la varianza del ruido aditivo presente) y perdiendo en homogeneidad. Esta fue una de las razones por las que se decidió modificar la primera solución propuesta y se comentará nuevamente en la Sección 6.2. Se realizaron diversos experimentos en este sentido y serán comentados en el Capítulo 7.

Muchos algoritmos utilizan una inicialización en la que se construyen las referencias tomando subconjuntos aleatoriamente seleccionados y se corre el algoritmo de clasificación varias veces para distintos conjuntos de inicialización sorteados. Este es el caso por ejemplo del algoritmo *k-means*. En el presente problema esto no es viable dado el alto costo computacional que implica cada ejecución del algoritmo.

6.1.2. Ciclo iterativo

La parte iterativa del algoritmo está compuesta por dos bloques principales: uno de alineación y otro de clasificación. A continuación se describe cómo está compuesto dicho método así como algunos detalles de su implementación.

Sea $C = \{V_1, \ldots, V_N\}$ un conjunto de subtomogramas. Se supone que éstos son copias (rotadas, trasladadas y afectadas por el *missing wedge* en forma aleatoria) de una serie de modelos ideales $M_1, \ldots, M_K$. La cantidad de modelos K es desconocida. El objetivo es a la vez determinar los K subconjuntos de C formados por aquellos subtomogramas que son copias de un mismo modelos y determinar las transformaciones que mejor los alinea entre sí.

Como los problemas de clasificación y alineación están acoplados se propone un algoritmo iterativo. En cada iteración se supone que se cuenta con un conjunto de referencias $R_1, \ldots, R_{K'}$. Todos los volúmenes son alineados a cada una de las referencias y se le asigna la transformación obtenida como resultado del registrado contra la referencia que resultó más próxima. De la misma forma que se hace en el método de alineación a múltiples refer-

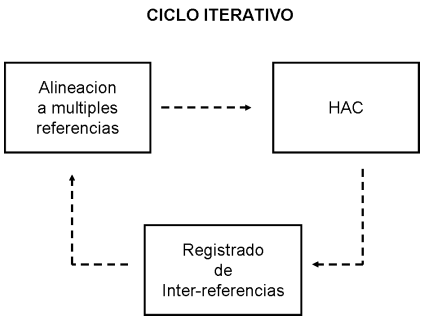


Figura 6.2: En esta figura se muestra un diagrama de bloques del "loop" iterativo. HAC es la sigla en inglés para clasificación jerárquica ascendente. El proceso iterativo combina la etapa de alineación a múltiples referencias, la clasificación presentada en el Capítulo 5 y una etapa donde todas las referencias son alineadas entre sí.

encias antes mencionado. Para calcular las alineaciones se utiliza la técnica presentada en el Capítulo 4. Los volúmenes son utilizados con sus nuevas orientaciones en un paso de clasificación empleando el algoritmo de clasificación jerárquica ascendente presentado en el Capítulo 5. De cortar el árbol jerárquico obtenido se hallan nuevos subconjuntos de volúmenes a partir de los que se calcularán las nuevas referencias. Para la primera iteración se utilizan las referencias obtenidas al promediar los volúmenes contenidos en cada una de las clases iniciales. Estas referencias se hallan a partir del conjunto de subtomogramas por lo que el algoritmo entra dentro de la categoría "libre de referencias".

En el algoritmo de clasificación todos los volúmenes se consideran en una orientación fija. Antes de calcular las referencias se emplea un algoritmo iterativo con el objetivo de corregir errores o incongruencias que puedan existir en las orientaciones. De esta manera se pretende conseguir promedios más precisos. El algoritmo utilizado es uno muy popular en el área del análisis de partículas individuales [24, 12] (conocido con el nombre de *bundle alignment*). Dicho algoritmo presenta un esquema iterativo que permite refinar las orientaciones de un conjunto de volúmenes. Supongamos que se cuenta con tres partículas estructuralmente idénticas afectadas por ruido y con distinto *missing wedge*. Si se las alinea dos a dos es probable que para una partícula dada se obtengan transformaciones que difieran ligeramente dependiendo de a cual de las dos restantes se la alineó. Intuitivamente puede pensarse que lo que habría que hacer es cambiar la función a minimizar en el problema de la alineación. El algoritmo introducido por Penczek et al. halla las trans-

formaciones que minimizan la suma de las disimilaridades dos a dos de un conjunto que se quiere alinear a partir de alineaciones de pares de partículas.

En el paso de clasificación, el criterio utilizado para cortar el árbol puede ser elegido por el usuario. En general en la gran mayoría de los algoritmos de este tipo, tanto en tomografía como en el caso del análisis de partículas individuales, se trata de evitar los umbrales de decisión automática. Como se comentará en el Capítulo 7 el método comúnmente utilizado es el de elegir un número de clases mayor al esperado y dejar evolucionar al algoritmo. En caso de que el número de clases elegido excede al número de clases real se observarán dos fenómenos distintos. El primero es que de las clases obtenidas la mayor parte de las partículas se agrupa en clases correspondientes asociadas a las reales quedando las otras con volúmenes dañados o con una minoría mal clasificada. El segundo efecto puede ser que más de una clase con similar número de elementos correspondan a una misma conformación con pequeñas variaciones (claramente si se cortara el árbol en un nivel mayor estas clases se unirían). El usuario puede visualmente detectar cualquiera de las dos situaciones fácilmente y descartar manualmente las clases que desee.

Antes de pasar a una nueva iteración, una vez realizada la etapa de clasificación y construidas las nuevas referencias, éstas son alineadas entre sí. En este caso también se utiliza la técnica de *bundle alignment* descrita. De está manera las referencias quedan todas puestas en un mismo marco. Si hubiera más de una referencias representando a la misma conformación estructural estas quedarian alineadas entre sí. Lo que implica que sean unidas en el siguiente paso de clasificación.

Por último cabe observar que De esto se desprende que las referencias podrán también tener regiones en el espacio recíproco donde no cuenten con información válida. Es decir que las referencias también presentarán el efecto del *missing wedge*. La idea es que al ir agrupando subtomogramas que originalmente se encontraran en orientaciones diferentes se logrará ir achicando el tamaño de dicha región, lo que hace que luego de pocas iteraciones el efecto mencionado sea cada vez menos perceptible.

6.2. Otros caminos explorados

Cerca de un año después de haber comenzado la presente maestría se llegó a elaborar una solución inicial para el problema de la construcción

de un mapa de alta resolución a partir de subtomogramas [3]. La solución allí propuesta es conceptualmente muy similar a la descrita en la sección anterior pero presenta algunas diferencias de implementación que se traducen en la performance de las mismas.

La solución propuesta utilizaba también un algoritmo iterativo que combinaba una etapa de alineación de múltiples referencias y otra de clasificación luego de aplicar una etapa de inicialización. El bloque de alineación era exactamente el mismo que el utilizado en la solución final, por lo que utilizaba también la medida de similaridad y la estrategia de alineación presentadas en el Capítulo 4. Las diferencias se encuentran en los bloques de inicialización y clasificación.

La etapa de clasificación era mucho más sencilla que la que se emplea ahora. Por cada referencia se obtenía una clase asociada a partir de la cual se generaba una referencia para el siguiente iteración. Cada subtomograma se asociaba a la referencia que resultara más próxima en la alineación a múltiples referencias. Luego se eliminaban de los conjuntos aquellos subtomogramas que distaran en más de un cierto umbral del promedio. Los promedios eran también debidamente calculados con la técnica que se explicó en la sección anterior. Es decir que es muy similar al método de análisis de partículas individuales conocido como "alineación mediante clasificación". Esta metodología es peor que la que se utiliza en la solución actual. En la nueva solución si se tiene algún error de asignación de una partícula a una referencia en el bloque de alineación (debido por ejemplo al ruido) el hecho de que las partículas sean comparadas entre sí en un mismo marco de orientación (ya que las referencias se encuentran alineadas) permite que éste sea corregido. Sin embargo las diferencias más grandes vienen dadas en la etapa de inicialización.

En la inicialización el método empleado para generar las primeras referencias se basaba en la premisa de que es mejor generar referencias construidas utilizando pocos volúmenes muy parecidos entre sí a utilizar promedios de muchos volúmenes heterogéneos. Para ello se utilizó un método en el que se registraban dos a dos todos los volúmenes del conjunto usando el método presentado en esta tesis. La medida de disimilaridad hallada para la transformación óptima era guardada, obteniendo así una matriz de disimilaridad. A esta matriz se le aplicaba un algoritmo de clasificación jerárquica ascendente igual al presentado en el Capítulo 5 pero utilizando la medida de disimilaridad guardada en la matriz. Luego se cortaba el árbol a un nivel de disimilaridad muy bajo de manera de garantizar que los conjuntos fueran homogéneos. Con

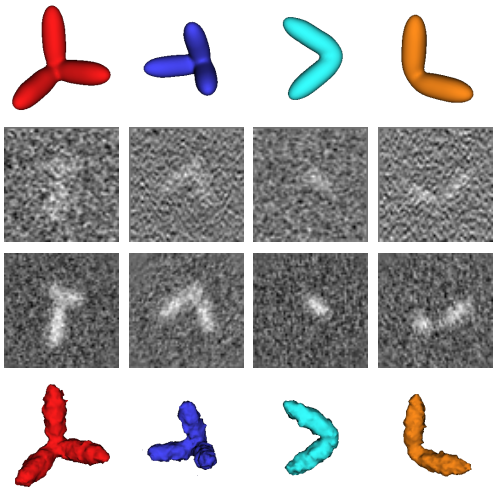


Figura 6.3: Resultados de un experimento exitoso para testear al algoritmo propuesto en la primera solución formulada. Se generaron 20 volúmenes por cada una de las clases afectados por ruido y el efecto del *missing wedge*. *Arriba*: Cuatro modelos artificiales utilizados. *Centro*: Secciones centrales para cuatro partículas crudas y sus correspondientes en los volúmenes de referencia finales. *Abajo*: Mapas tridimensionales reconstruidos para cada clase sin filtrar.

el correr de las iteraciones se iban agregando paulatinamente más volúmenes a las distintas clases hasta alcanzar la convergencia.

La diferencia fundamental entre las soluciones puede tener que ver con el hecho de que la solución propuesta en una primera instancia utilizara conjuntos de alta similaridad y pocos elementos mientras que la solución actual utiliza promedios de un gran número de volúmenes no tan parecidos entre sí. Utilizar muchos volúmenes en el promediado tiene la ventaja de la reducción de la varianza del ruido aditivo que las afecta. Por otro lado el hecho de que dos volúmenes sean muy parecidos no necesariamente implica que estos sean de idéntica conformación. Valores de disimilaridad muy baja pueden deberse a factores externos como por ejemplo el ruido. Por lo tanto si se tenían errores en las referencias iniciales éstas perdían su significado por estar formadas a partir de muy pocos volúmenes. Las diferencias en el desempeño fueron principalmente apreciables cuando se trabajó con los datos reales adquiridas unos meses después de publicada la solución anterior.

La solución propuesta en una primera instancia mostró excelentes resultados cuando se trabajó con datos artificiales. En la Figura 6.3 se muestra

el resultado de uno de los muchos experimentos realizados. En ese caso se trabajó con partículas artificiales generadas de acuerdo a la técnica que se detallará en la Sección 7.1. Se generaron cuatro clases de partículas diferentes para construir un conjunto de un total de 80 subtomogramas igualmente divididos entre las clases. Todas las partículas están afectadas por el efecto del *missing wedge* de forma de simular lo que sucede a las imágenes de CryoTE, en este ejemplo particular se tomaron proyecciones en un rango de ± 60 grados. Luego de algunas iteraciones todos los volúmenes fueron asignados a sus respectivas clases y los mapas de alta resolución obtenidos fueron altamente satisfactorios. En la Figura 6.3 se muestra también una sección central tanto de los volúmenes crudos como de las referencias halladas al término de la iteración. Puede verse claramente la mejora en la relación señal a ruido a pesar del pequeño tamaño de la muestra. Este ejemplo tiene sin dudas muy pocos elementos. Básicamente el costo computacional asociado al cálculo de la matriz de disimilaridad dificultaba seriamente el hecho de trabajar con conjuntos de datos grandes. Si no hubiera habido diferencias en la performance, esta razón hubiera sido suficiente por si sola para justificar el cambio realizado a la solución.

Capítulo 7

Resultados

Este capítulo está dedicado a mostrar los resultados obtenidos. La primera sección está dedicada al análisis del algoritmo de registrado de volúmenes propuesto en el Capítulo 4. Este está concentrado en la Sección 7.2. Allí se mostrarán dos experimentos fundamentales para medir su desempeño y a su vez entender cuáles son sus alcances y limitaciones. En las secciones 7.3 y 7.4 se muestran los resultados del algoritmo de reconstrucción completo. En la primera se muestran los resultados obtenidos con datos artificiales mientras que en la otra sección se trabajó con datos reales adquiridos en el laboratorio de Dr. Sriram Subramaniam por el Dr. Jun Lui.

Todo el código fue programado en C++. Para realizar los experimentos comentados en esta sección el código fue paralelizado y ejecutado haciendo uso de las capacidades computacionales de alta performance del *Biowulf Linux cluster* en el National Institutes of Health, Bethesda, MD.

7.1. Generación de los datos artificiales

Todos los volúmenes artificiales con los que se trabajó fueron generados de forma de simular las características que presentan los volúmenes de CryoTE.

Dado que los distintos factores que afectan a los volúmenes de TE y CryoTE son tan diversos resultan muy difíciles de cuantificar. En muchos de los trabajos del área, los algoritmos se evalúan (cuando se evalúan) utilizando datos que son claramente irreales. Esto puede dejar ciertas dudas respecto del desempeño real que dichos algoritmos tendrán cuando se los utilice con

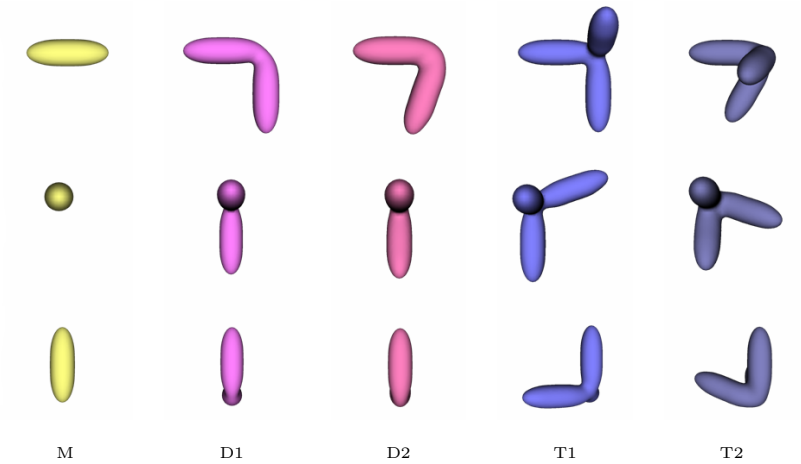


Figura 7.1: Cinco modelos aritificiales utilizados para generar conjuntos de datos para los experimentos de la Sección 7.3. En cada fila se los muestra desde tres direcciónes ortogonales. Algunos de ellos fueron generados para tener configuraciones con altos grados de similitud de manera de testear severamente al algoritmo de clasificación.

datos reales.

En una primera instancia, los volúmenes de prueba eran simplemente partículas artificiales a las cuales se le sumaba ruido aditivo gaussiano y luego se le aplicaba un filtro cuya transferencia fuera igual a la máscara binaria que se deseaba. Visualmente dichos volúmenes no eran del todo parecidos a los adquiridos con el microscopio, por lo que se decidió modificar la técnica de generación. El segundo paso fue simular el proceso de reconstrucción de los tomogramas, es decir, tomar proyecciones de las partículas artificiales y a éstas agregarles ruido para luego utilizarlas con el mismo algoritmo de reconstrucción que se utiliza con los datos reales (Algebraic Reconstruction Technique - ART). El resultado obtenido mostró grandes mejorías ya que visualmente los datos de prueba comenzaron a parecerse mucho más a los reales.

Como se describió en la Sección 2.2 las distintas proyecciones son alineadas manualmente unas a otras para realizar la reconstrucción. Esto implica que muchas de dichas proyecciones presenten pequeños errores de alineación respecto a las demás. Por lo tanto se agregó al algoritmo de generación de volúmenes artificiales un error aleatorio en las orientaciones de las proyecciones previo a la reconstrucción. De esta manera los volúmenes artificiales están bastante cerca de lo que se tiene en el caso real. Naturalmente existen

muchos otros factores que afectan a los volúmenes reales y no a los artificiales, siendo el más notorio el ruido biológico debido al material presente alrededor de las partículas.

El diseño de las partículas artificiales, si bien no es de complejidad excesiva, está basado en estructuras biológicas reales y fue supervisado por el Dr. Subramaniam. En la Figura 7.1 se muestra un conjunto de cinco partículas artificiales utilizados en los experimentos de la Sección 7.3. Como se explicará más adelante, algunos de ellos fueron generados para tener configuraciones con altos grados de similitud de manera de testear severamente al algoritmo de clasificación.

7.2. Performance de la rutina de registrado

7.2.1. Alcances y limitaciones

En esta sección se evaluará el desempeño de la rutina de registrado presentada en el Capítulo 4. Es imposible realizar un juicio categórico sobre la rutina de alineación del tipo bueno/malo. En general el éxito del registrado en volúmenes provenientes de TE o CryoTE depende de varios factores, unos fácilmente medibles y otros de tipo cualitativos que resultan difíciles de cuantificar. La relación señal a ruido y el tamaño de la región de información faltante son factores que pueden controlarse en ejemplos sintéticos. El éxito en el resultado de la rutina de alineación depende también de factores como el tamaño relativo de las regiones de información faltante respecto de la estructura que presenten las partículas contenidas en los subtomogramas. Podría llegar a suceder, bajo ciertas circunstancias, que dadas las regiones de información disponible de los volúmenes que se quiere alinear no sea posible encontrar la transformación correcta incluso en casos libres de ruido.

Sin embargo es posible comparar la performance de la rutina de alineación desarrollada contra otras técnicas que se utilizan en el análisis de imágenes de TE y CryTE. En particular se trabajará contrastando los resultados con los que se obtendrían si se utiliza la correlación de volúmenes para realizar el registrado. Como ya se comentó en el Capítulo 3, esta técnica sigue siendo comúnmente utilizada.

Para intentar obtener una idea de los alcances y las limitaciones que presenta la técnica de registrado propuesta se realizaron muy diversas pruebas variando las condiciones en las que se realiza el registrado. Todas las prue-

bas fueron realizadas con un modelo artificial que no presenta simetría de manera de que exista una única solución correcta. Este modelo es el $T1$ que se muestra en la Figura 7.1. Se pretende estudiar cómo afectan al resultado de la alineación dos factores: el nivel de ruido que los afecta y el tamaño de la región de información faltante. En el resto de la sección nos referiremos a dicho tamaño como *wedge* y estará medido en grados, representando el rango de proyecciones tomados, ± 90 indicaría que el volumen no tiene regiones sin información válida. Se contrastara el desempeño de la solución propuesta en esta tesis contra el resultado de alinear mediante la utilización de la correlación.

Para cada par *(ruido, wedge)* se generó un conjunto de 50 volúmenes con orientaciones y posiciones relativas de la región de información faltante aleatorias. Luego se alineó cada uno de los volúmenes de dichos conjuntos a una copia de referencia (también afectada por las mismas condiciones de ruido y *wedge*). Para medir el error se utilizó la norma de Frobenius entre las matrices asociadas a la transformación real y la transformación recuperada en el proceso de alineación. En la Figura 7.2 se muestra el gráfico de una función bidimensional, donde se muestra para cada una de las condiciones los errores medios obtenidos. En la imagen de la izquirda se muestra el desempeño obtenido por el alineado via correlación, en el gráfico de la derecha se muestran los resultados arrojados por la técnica de registrado presentada en esta tesis.

En líneas generales puede verse que el error en la alineación es para casi todos los puntos superior en el primer caso. Se observa claramente que en el primer caso la performance de los resultados se degradan rápidamente. Para para valores de *wedge* menores a ± 70 e incluso muy modestos niveles de ruido presenta errores muy grandes. En el otro caso pueden conseguirse buenos resultados hasta valores *wedge* de ± 55 y presenta una tolerancia al ruido mucho mayor prácticamente para todos los valores. Bajo ciertas condiciones de fuerte ruido y tamaños de *wedge* grandes no es posible obtener un resultado que en media sea aceptable. Para aportar una medida cuantitativa de lo que se observa a las claras en la Figura 7.2 se calculó el error cuadrático medio para toda la grilla de *(ruido, wedge)*. Esta medida global dio para el algoritmo basado en correlación 3.72 mientras que para el otro algoritmo fue de 1.37.

Debe recalcarse sin embargo que la evaluación que aquí se está haciendo es muy estricta ya que se están alineando directamente volúmenes crudos. Los resultados obtenidos al alinear volúmenes crudos contra referencias con mejor relación señal a ruido y posiblemente menor tamaño de *wedge*(o incluso

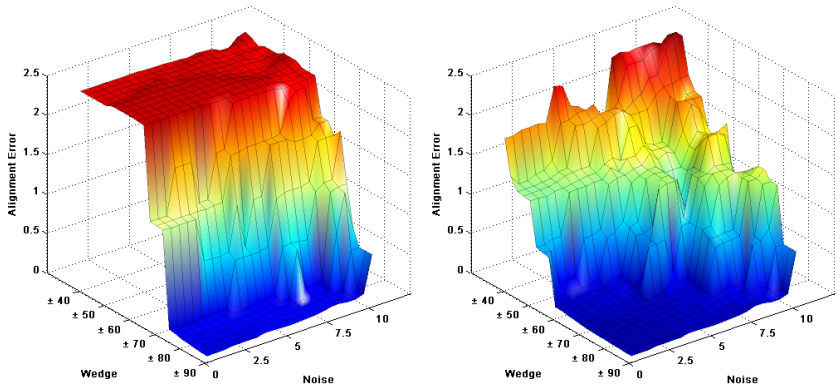


Figura 7.2: Error promedio obtenido al evaluar las técnicas de registrado sobre un conjunto de 50 volúmenes presentando orientaciones y regiones de información faltante aleatorias. Se probaron distintas condiciones de ruido y tamaño de región de infromación disponible distintos. *Izquierda*: Método basado en la correlacion, valor cuadrático medio total 3.72. *Derecha*: Utilizando la técnica de registrado presentada en la Sección 4.2, error cuadrático medio total 1.37.

nulo) mejoran drásticamente. Es por esta razón que en general los algoritmos diseñados para alinear conjuntos de subtomogramas evitan realizar muchos registrados de volúmenes crudos.

Por último cabe recalcar que la forma del gráfico obtenido depende naturalmente del modelo de partícula artificial que se esté empleando. Si se utilizara por ejemplo una esfera, los gráficos para ambos algoritmos resultarían muy superiores a los mostrados en la Figura 7.2. Con la partícula elegida se intentó buscar un modelo que fuera biológicamente fundado y que presente un grado de complejidad tal que haga la comparación interesante y a la vez no fuera un caso aislado o extremadamente complejo.

El costo computacional asociado a cada una de las técnicas presenta diferencias enormes. El tiempo requerido para calcular el registrado de dos volúmenes de tamaño $64 \times 64 \times 64$ en una computadora Pentium 4 de 2.13MhZ es de menos de 30 segundos para la rutina que aquí se propone y varios minutos en el caso de la búsqueda exhaustiva.

7.2.2. Promediado de subtomogramas

En esta sección se pretende ilustrar con un ejemplo artificial, pero biológicamente fundado, los errores en los que se puede incurrir por no considerar

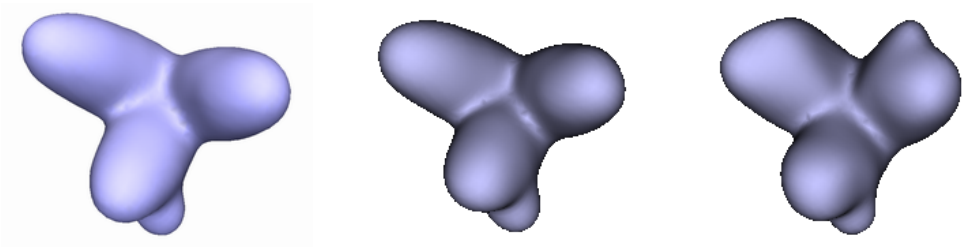


Figura 7.3: Esta figura muestra los posibles efectos del error en la alineación en el promediado volúmenes. *Izquierda*: Modelo original. *Centro*: Promedio de 70 copias aleatorias del volumen modelo alineadas utilizando la técnica presentada en la Sección 4.2. Se aprecia claramente que el resultado es muy similar al modelo original. *Derecha*: Mismo promedio pero en este caso registrando los volúmenes utilizando correlación. Los errores en la alineación hacen que la pequeña irregularidad presente en una de las extremidades del modelo original aparezca en todas ellas.

apropiadamente que los volúmenes están afectados por el efecto del *missing wedge*. En la Figura 7.3 (izquierda) se muestra el modelo utilizado para este experimento que pretende simular la interacción entre macromoléculas de diferentes tamaños comúnmente observadas con CryoTE. Este es exactamente el tipo de situación que se pretende encontrar en los estudios la interacción de anticuerpos adosados a las proteínas presentes en las membranas de células o virus.

La idea fue plantear un escenario sumamente sencillo en donde las diferencias puedan ser aún visibles. Se pretendió buscar un ejemplo en el cual pudiera aislarse el factor introducido por considerar o no las regiones de información faltante. Por estas razones se generaron 70 copias del volumen modelo nuevamente en posiciones y con orientación del *missing wedge* aleatorias. Se tomaron proyecciones en un rango de ± 60 y se utilizó el algoritmo ART para su reconstrucción. Todos los volúmenes fueron alineados al modelo original para determinar su orientación. Posteriormente dichos volúmenes fueron promediados para producir los mapas asociados.

En los gráficos del centro y derecha de la Figura 7.3 se muestran los resultados obtenidos. El mapa construido de promediar volúmenes alineados utilizando la correlación no es correcto. Puede notarse claramente que la pequeña irregularidad que aparece en solo una de las patas del modelo original puede verse en las tres patas del mismo. Esto se debe a frecuentes errores en la alineación. En tanto, el mapa construido de alinear utilizando

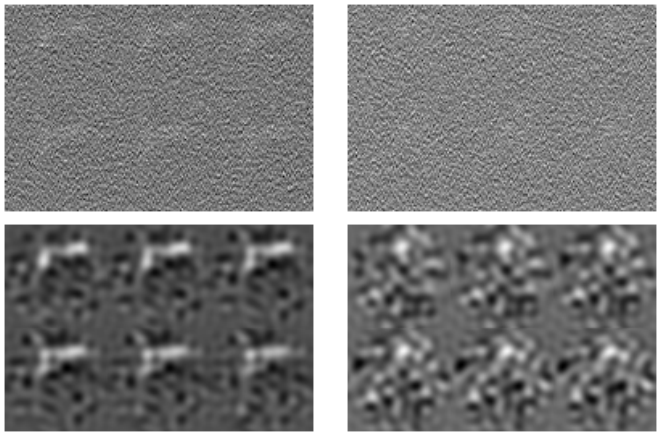


Figura 7.4: Volúmenes reconstruidos utilizando la técnica explicada en la Sección 7.1 a partir de los modelos mostrados en la Figura 7.1. Se muestran seis secciones centrales antes y después del filtrado pasabanda (arriba y abajo respectivamente) para dos niveles de ruido: $SNR = 0{,}04$ (*Izquierda*) y $SNR = 0{,}018$ (*Derecha*). Se generaron un total de 550 volúmenes para las cinco configuraciones M, D1, D2, T1 y T2 y un conjunto extra conteniendo imágenes de ruido.

el método presentado en la Sección 4.2 es muy similar al original.

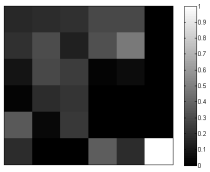
7.3. Validación utilizando datos artificiales

En esta Sección se estudiará un experimento donde se intentará testear el algoritmo en su totalidad. Se utilizarán datos artificiales creados empleando la técnica descrita en la Sección 7.1 a partir de los modelos M, D1, D2, T1, y T2 mostrados en la Figura 7.1. Dichos modelos fueron creados especialmente presentando conformaciones muy similares para testear fuertemente a la capacidad de clasificación del algoritmo. Los modelos D1 y D2 son particularmente parecidos, ambos están formados por dos componentes dispuestas en un caso formando un ángulo recto mientras que en otro un ángulo de 70 grados. Se desea estudiar cuál es la capacidad real que tiene el algoritmo de producir mapas de alta resolución cuando se tiene más de una conformación a diferentes intensidades de ruido.

Los volúmenes fueron obtenidos utilizando el algoritmo de reconstrucción ART a partir de proyecciones tomadas en un rango de ± 60 grados. El efecto del *missing wedge* puede constituir un estorbo severo no solo en la etapa de

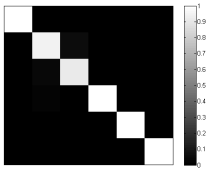
SNR = 0.04
Error = 2.5
Iteración 1

Clase	M	D1	D2	T1	T2	N	
I	12	22	19	29	37	0	119
II	15	39	13	32	59	0	54
III	6	36	24	2	6	0	46
IV	2	23	21	0	0	0	74
V	26	5	23	0	0	0	158
VI	14	0	0	37	23	25	99
Total	75	125	100	100	125	25	550



SNR = 0.04
Error = 2.5
Iteración 5

Clase	M	D1	D2	T1	T2	N	
I	75	0	1	0	0	0	76
II	0	119	6	0	0	0	125
III	0	4	92	0	1	0	97
IV	0	2	1	100	0	0	103
V	0	0	0	0	124	0	124
VI	0	0	0	0	0	25	25
Total	75	125	100	100	125	25	550



SNR = 0.018
Error = 5.0
Iteración 9

Clase	M	D1	D2	T1	T2	N	
I	74	1	2	0	0	0	77
II	0	64	42	2	3	0	110
III	0	55	49	2	1	0	107
IV	0	5	5	94	4	0	108
V	0	0	3	2	117	0	122
VI	1	0	0	0	0	25	26
Total	75	125	100	100	125	25	550

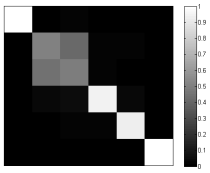


Figura 7.5: En esta figura se muestran los resultados obtenidos al realizar experimentos del algoritmo completo utilizando datos artificiales generados según los criterios establecidos en la Sección 7.1. Se utilizaron las clases de modelos que se muestran en la Figura 7.1. *Arriba*: Resultados de la clasificación luego de la primera iteración para el caso de $SNR = 0{,}04$ y ± 2.5 píxeles de error en la alineación de las proyecciones. Las clases todavía no están claramente separadas. *Centro*: Resultado final de la clasificación para el mismo caso que la gráfica de arriba. El resultado luego de nueve iteraciones muestra que todas las clases fueron separadas con éxito. *Abajo*: Resultado de la clasificación para el caso de $SNR = 0{,}018$ y errores en la alineación de ± 5 píxeles. Casi todas las clases pueden recuperarse a pesar de los niveles de ruido extremos. Sin embargo, las clases más similares entre sí del conjunto de prueba (D1 y D2) no pueden separarse adecuadamente y resultan en dos clases con aproximadamente igual número de partículas de cada una.

alineación del algoritmo sino también en la de clasificación. Observar que el modelo T1 se construye de agregar una extremidad al modelo D1. Por lo tanto, en determinadas posiciones relativas de la región de información faltante, dicha extremidad puede resultar seriamente afectadas y entonces, copias de D1 y T1, pasar a ser muy parecidos.

Se generaron un total de 550 volúmenes. Se puso particular atención en generar clases que tuvieran números de elementos considerablemente distintos de manera de testear también el comportamiento del algoritmo en ese sentido. La cantidad de volúmenes generados para de cada clase fue de: 75 M, 100 D1, 125 D2, 125 T1, 100 T2. Se agregaron también 25 volúmenes de ruido. Se realizaron varios experimentos variando el nivel de ruido y el error en la alineación de las proyecciones (explicado en la Sección 7.1). Para cada uno de dichos factores se utilizaron dos niveles distintos de complejidad. El error de alineación en las proyecciones se tomó aleatorio (de distribución uniforme) variando en un caso entre $\pm$2,5 píxeles y en el otro $\pm$5 píxeles. Se utilizaron dos niveles de ruido afectando a los volúmenes. En el primer caso éstos fueron generados para tener una relación señal a ruido (luego de la reconstrucción) de $SNR = 0{,}04$, mientras que en el segundo experimento se emplearon volúmenes con un nivel de ruido mucho mayor, $SNR = 0{,}018$. En la Figura 7.4 se muestran seis secciones para volúmenes creados con cada uno de los niveles de ruido. Estas se muestran antes y después de ser filtradas pasabanda. Las imágenes crudas en el segundo caso hacen prácticamente imposible la visualización.

Se corrió el algoritmo para los cuatro pares posibles de ruido y error en la alineación de las proyecciones. En todos los casos se impuso seis clases y se dejó fijo dicho valor a lo largo de las iteraciones. Para ambos niveles de ruido y un error en la alineación de las proyecciones de $\pm$2,5 píxeles se observó que luego de algunas iteraciones las clases fueron recuperadas correctamente y los promedios reflejan cada una de las correspondientes clases. Sin embargo, en el experimento en que las condiciones fueron las más severas se constató que las clases D1 y D2 no pudieron ser correctamente separadas. El hecho de que se haya impuesto el número de clases a ser cinco en el algoritmo de clasificación probablemente es lo que hace que haya una división de las mismas. En este caso si un usuario observara las clases las consideraría iguales aunque no los son. Un resultado más natural sería que dichas clases estuvieran unidas formando una única clase. Esto puede verse con claridad en la Figura 7.5 (Abajo), donde se muestra la matriz de confusión para el caso de $SNR = 0{,}018$ y error en las alineaciones de $\pm$5 píxeles. Si bien en este caso se observan algunos errores más que en los otros casos, si se dejan

de lado los errores mencionados entre las clases D1 y D2, puede decirse que la clasificación resultó muy exitosa.

Al mirar como varía la matriz de confusión con las iteraciones, puede apreciarse cómo las partículas van mejorando su alineación y agrupándose con las de su grupo. En las dos imágenes de más arriba de la Figura 7.5 se muestran las matrices de confusión luego de la primera y última iteración para el experimento con $SNR = 0{,}04$ y un error en la la alineación de las proyecciones de $\pm 2{,}5$. Puede apreciarse con claridad como inicialmente las clases no representan a ninguna de las clases originales mientras que al final se constantan sólo muy pocos errores.

Los resultados obtenidos por el algoritmo son considerados excelentes ya que en todos los casos se logró obtener mapas de alta resolución para las distintas clases incluso bajo condiciones de ruido extremas. Se detectaron errores únicamente en el experimento de mayor dificultad, combinando el ruido más severo y grandes errores en la alineación de las proyecciones. Las dos clases más parecidas entre sí no pudieron separarse. Sin embargo, el resultado puede aún considerarse bueno, ya que el algoritmo pudo separar con éxito las restantes conformaciones. Esto muestra que bajo ciertas circunstancias, el algoritmo presenta un límite en la similitud de las conformación llevó a detectar ciertos errores entre clases muy similares entres peor el resto de las clases pudo ser separadas con éxito.

7.4. Validación utilizando datos reales

Realizar una cuidadosa validación utilizando datos artificiales es fundamental para la evaluación de algoritmos de este tipo, ya que cuando se trabaja con datos artificiales pueden controlarse muchos factores y de este modo medir con precisión el desempeño del algoritmo. Sin embargo, para poder tener la certeza de que el algoritmo será capaz de procesar adecuadamente datos reales no hay otro camino que trabajar con ellos. En dichos casos, sin embargo, resulta difícil evaluar su performance.

El experimento que se decidió realizar fue el de aplicar el algoritmo propuesto a un conjunto de datos reales de alguna partícula para la cual se cuente a priori con información precisa acerca de su confromación estructural. Este es el caso de la partícula groEL. Dicha partícula consta de una única conformación y viene estudiándose desde hace años en el área de análisis de

partículas individuales. Se ha logrado obtener mapas de hasta $6\mathring{A}$ de resolución [17]. Por lo tanto el mapa obtenido mediante el algoritmo que se propone en esta tesis puede ser validado mirando el mapa conocido a priori.

capitulos/capitulo2/ejemplo

El conjunto de datos con el que se trabajó en este experimento está compuesto por 345 partículas de groEL provenientes de seis tomogramas distintos con proyecciones variando en rangos distintos pero todos ellos en las proximidades de ± 65 grados. Este conjunto es sumamente chico para obtener mapas de resoluciones importantes pero ese no es el objetivo de este experimento. En la Figura 7.6 se muestran los datos utilizados, mientras que en la Figura 7.9 se muesta una una sección de la reconstrucción de un tomograma. Todas las partículas fueron extraídas de sus respectivos tomogramas a mano. La orientación relativa de cada subtomograma en su respectivo tomograma fue guardada en el momento de la adquisición, para así poder determinar cuál es la la región de información faltante que tiene asociada, tal como lo requiere la rutina de registrado presentada en la Sección 4.2.

Primero se corrió la etapa de inicialización imponiendo cinco clases. Antes de comenzar a correr la etapa iterativa del algoritmo, se analizaron los resultados obtenidos por la clasificación jerárquica para evaluar el desempeño de este paso de clasificación utilizando la función de disimilaridad definida en la Sección 4.1. En la Figura 7.10 se muestran el árbol (simplificado) de clasificación obtenido para el conjunto de datos de groEL. En el árbol se muestran las secciones horizontales centrales de los volúmenes obtenidos al promediar los subtomogramas contenidos en cada una de las clases del árbol. Puede verse que el resultado obtenido es muy bueno, quedan claramente separadas en grupos homogéneos las vistas desde arriba y las vistas laterales, mientras que un grupo de partículas dañadas o mal seleccionadas es también agrupado. Puede verse cómo en esta etapa inicial, partículas que se encuentran en orientaciones distintas son consideradas diferentes, hay más de un grupo de vistas laterales que difieren en una rotación planar y fueron agrupadas en clases distintas.

Al igual que lo que sucedió en el caso de los datos artificiales, las clases obtenidas con los datos de groEL fueron refinándose con el correr de las iteraciones. En la Figura 7.7 se muestra la evolución de las clases obtenidas. Los promedios de cada una de dichas clases se muestran proyectados en dos direcciones ortogonales. Claramente en la primera iteración las partículas dejan ver las incongruencias esperables, resultado de promediar volúmenes que no están adecuadamente registrados. Por otro lado es muy notorio el efecto del

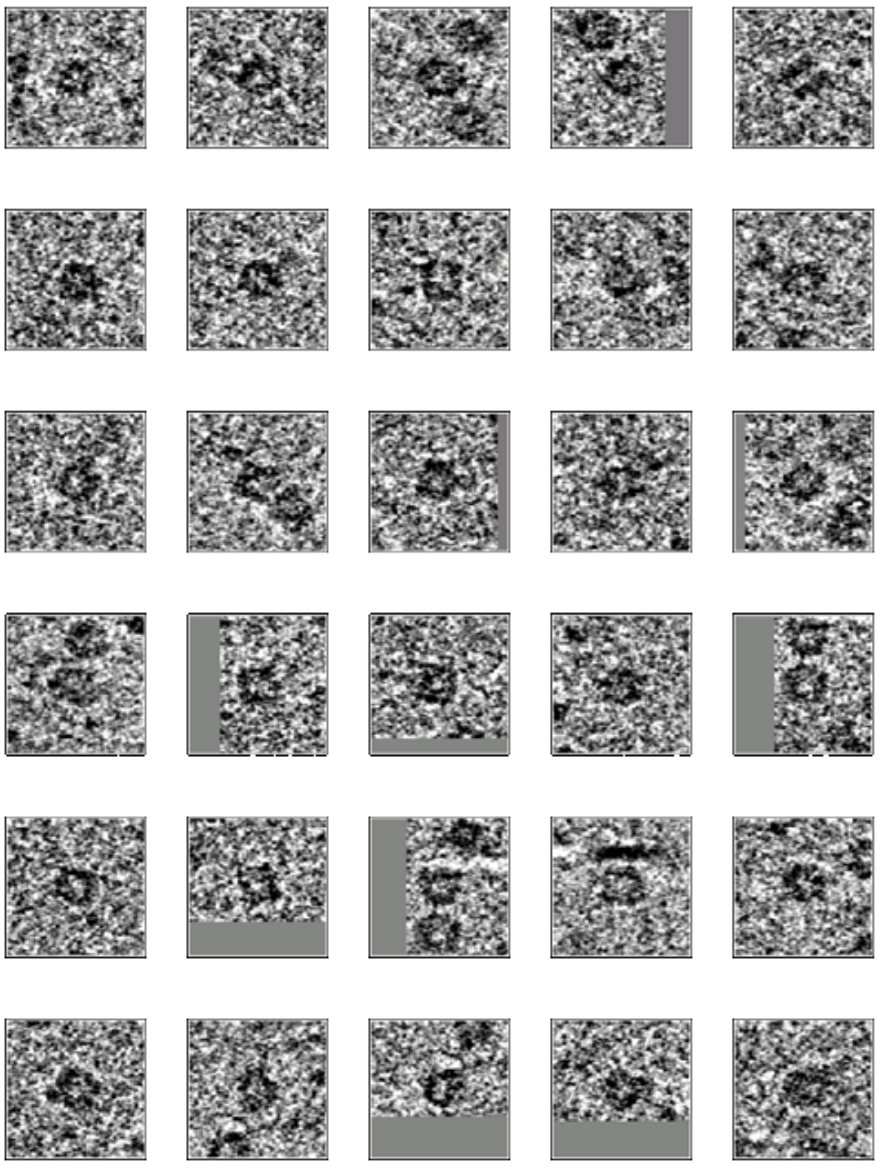


Figura 7.6: Subtomogramas extraídos a mano para su procesamiento del algoritmo propuesto. Se muestra la sección horizontal central de cada partícula.

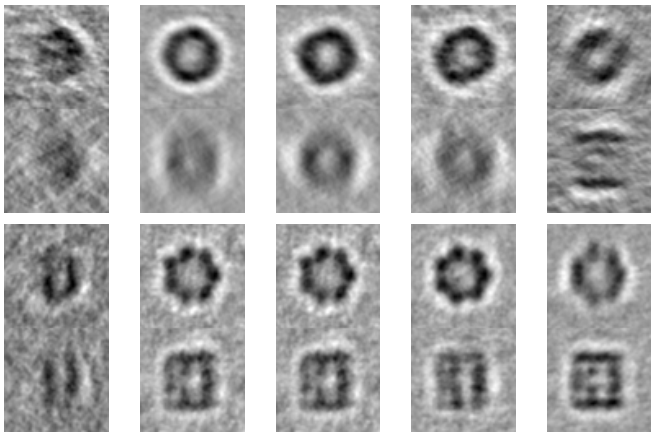


Figura 7.7: Resultados obtenidos de correr el algoritmo propuesto con los datos de la partícula groEL. *Arriba*: Dos Proyecciones ortogonales de las cinco clases iniciales utilizadas en el algoritmo. El promediado incoherente asociado produce resultados de baja resolución y con una notoria incidencia del efecto *missing wedge*. *Centro*: Las mismas proyecciones ortogonales para las cinco clases luego de seis iteraciones.

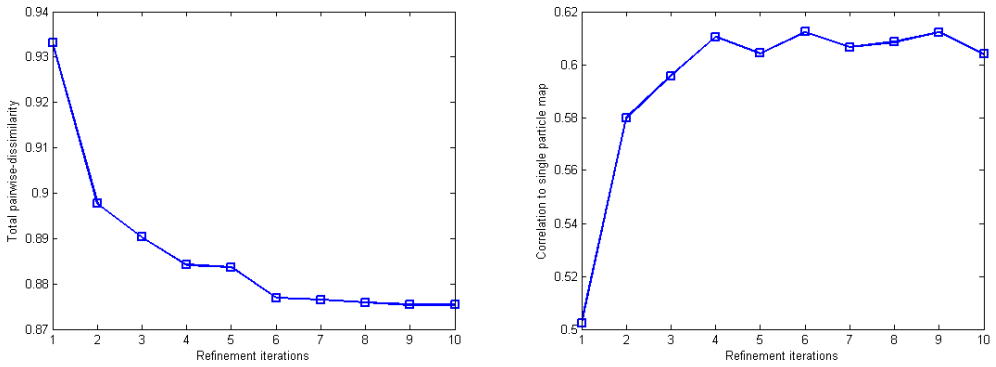


Figura 7.8: Gráficas de la performance del algoritmo con los datos de groEL. *Izquierda*: Varianza intra-clase para las cinco clases estudiadas a lo largo de las iteraciones. *Derecha*: Correlación entre el mapa construido por el algoritmo y el mapa conocido a priori.

missing wedge en algunas de ellas. Luego de varias iteraciones los promedios que representan a las clases toman una forma mucho más definida y puede verse que las distintas vistas pueden alinearse correctamente. Esto muestra que cualitativamente las clases mejoran con el correr de las iteraciones. Las dos clases que se muestran más a la derecha concentran la mayor parte de los subtomogramas. Resulta claro que todas las clases menos la que se muestra más a la izquierda representan (al nivel de resolución alcanzado) a una misma conformación estructural. La primera de las clases agrupó a una serie de parículas dañadas de tamaño claramente menor que las de groEL. Finalmente las cuatro clases relevantes fueron unidas para construir un mapa de la partícula.

A los efectos de encontrar una medida cuantitativa que refleje las mejoras cualitativas observadas se calculó la varianza media intra-clases media obtenida durante las iteraciones. Esta medida refleja qué tan "compactas" se hacen las clases al refinarse los parámetros de alineación de los distintos subtomogramas o, dicho de otro modo, qué tan bien el promedio representa a los miembros de la clase. El resultado esperado, siempre que el algoritmo esté funcionando correctamente, es que dicho valor decrezca a medida que los volúmenes van colocándose en posiciones más precisas con las tandas de alineación sucesivas. Esto es exactamente lo que sucede, tal como se muestra en la Figura 7.8.

Como se mencionó, la partícula groEL fue elegida porque se cuenta con mapas de alta resolución confiables que pueden compararse con el resultado obtenido. Cualitativamente los resultados son muy parecidos. En la Figura 7.11 se muestran gráficas comparativas entre el mapa obtenido utilizando el algoritmo propuesto, el mapa conocido a priori y una partícula cruda individual. Se muestran en mosaico varios cortes horizontales de los subtomogramas. Al comparar el mapa obtenido con la partícula cruda queda al descubierto el mejoramiento en la calidad de la imagen obtenido. En la partícula cruda puede advertirse una densidad en forma de anillo en algunas de las secciones centrales, por otro lado el mapa presentado muestra claramente que la partícula presenta (o está muy cerca de presentar) simetría propias de la partícula groEL. Al comparar el mapa conocido de ante mano con el obtenido, pueden verse muchas características en común.

Pueden definirse también medidas cuantitativas para la comparación contra el mapa conocido previamente y el obtenido en el experimento. Una forma de hacerlo es mirar como varía su similaridad a lo largo de las iteraciones. Cabe aclarar que el mapa con el que se cuenta a priori fue obtenido en

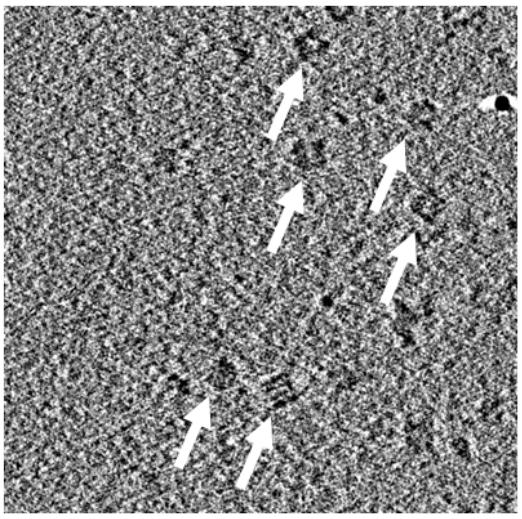


Figura 7.9: Sección de un tomograma conteniendo partículas de groEL. Las flechas blancas ubican diferentes partículas de groEL dispuestas en el tomograma.

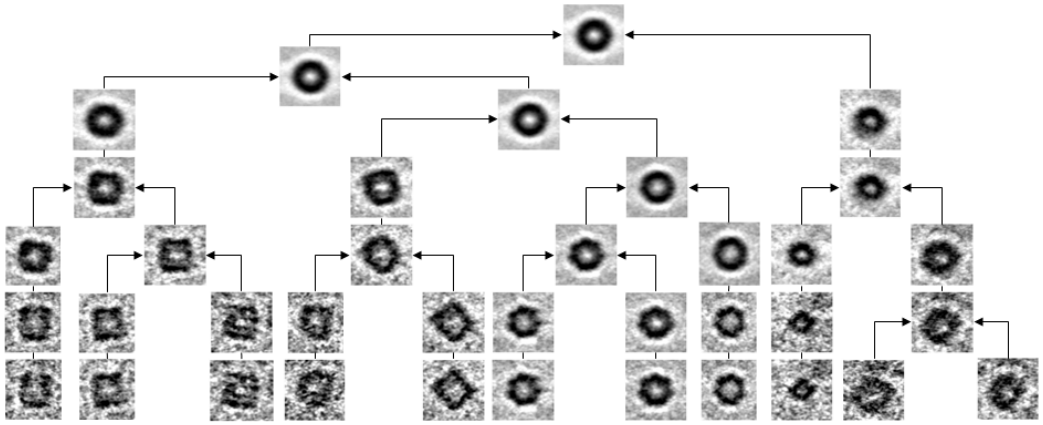


Figura 7.10: Clasificación jerárquica obtenida para un conjunto de 345 partículas de groEL. Los subtomogramas son llevados a la representación de baja dimensión explicada en la Sección **??** y utilizados en el cálculo de la matriz de distancias utilizada por el algoritmo de clasificación jerárquica. Se muestra una versión simplificada del árbol obtenido conteniendo solamente los nodos relevantes. Las partículas fueron clasificadas con éxito en diversos grupos correspondientes a partículas vistas de arriba y vistas laterales. Las partículas dañadas o mal seleccionadas también fueron agrupadas y constituyen una rama separada del árbol.

forma totalmente independiente. En la Figura 7.8 se muestra el resultado obtenido. Como era de esperar a medida que transcurren las iteraciones el mapa obtenido se parece cada vez más al mapa con el que se cuenta. Lógicamente, también en este caso la similitud no puede crecer indefinidamente, por lo que se acerca a un valor máximo.

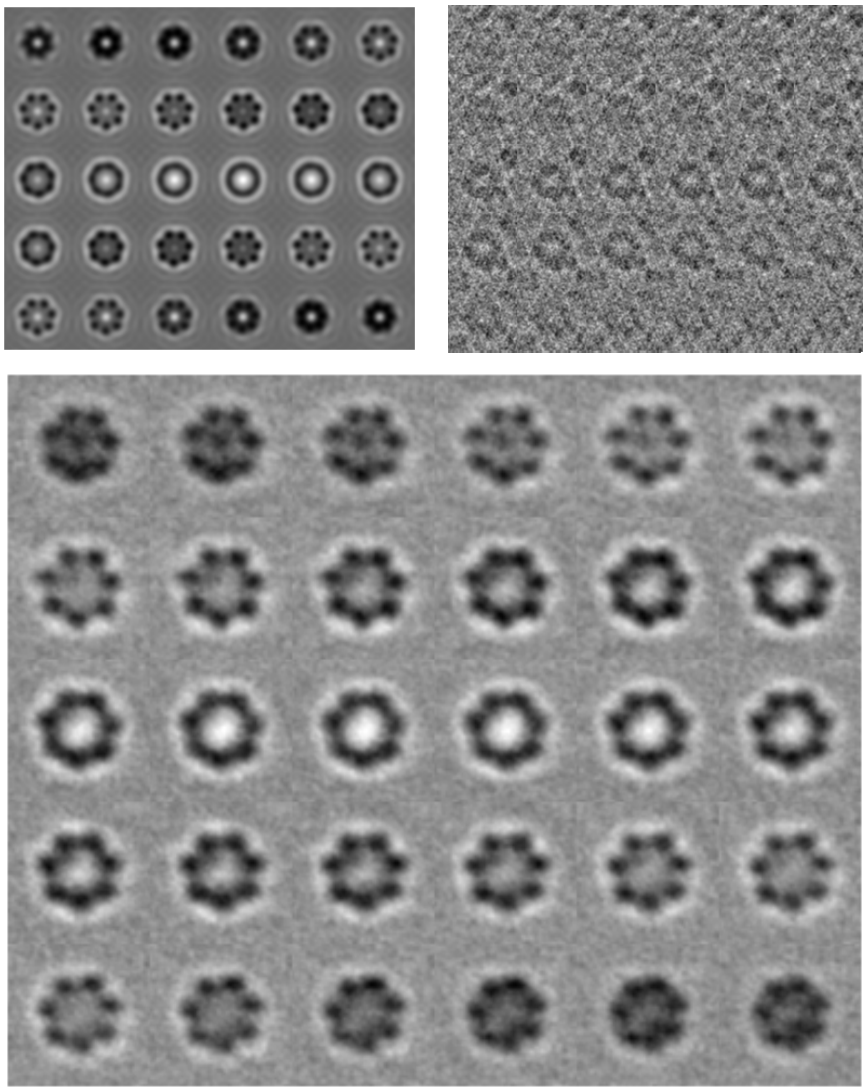


Figura 7.11: Comparación del mapa obtenido. Se muestran cortes horizontales sucesivos de: *Arriba*: El mapa conocido a priori filtrado a 32$\mathring{A}$ de resolución. *Abajo* El mapa reconstruido utilizando el algoritmo presentado en esta tesis.

Capítulo 8

Conclusiones y trabajo a futuro

En esta tesis se presentó un algoritmo para el registrado de volúmenes de tomografía electrónica, intrínsecamente afectados por el efecto del *missing wedge*. Para ello se introdujo un modelado de los volúmenes que mediante el uso de máscaras permite atacar el problema en forma matemáticamente precisa y formal. Por primera vez se realizó un estudio detallado que muestra claramente que ignorar el hecho de que los volúmenes tienen regiones de información faltante (en determinados componentes frecuenciales) puede conducir a errores sistemáticos en las orientaciones recuperadas, incluso en situaciones aparentemente sencillas. Se realizó un análisis exahustivo en el que se analizó cómo afectan al resultado de la alineación diversos factores, como ser el nivel de ruido que afecta a las imágenes o el tamaño del *missing wedge*. Se concluyó que dentro de determinados rangos, el algoritmo propuesto tiene una muy buena performance. Por otro lado, la rutina presentada es computacionalmente eficiente y su costo es muy inferior al de todas las técnicas hasta ahora publicadas que también garanticen el hallazgo de la solución óptima.

Se presentó un algoritmo de clasificación basado en la experiencia previa desarrollada en el campo del análisis de partículas individuales. Este método es fiel a la filosofía utilizada para desarrollar la técnica de registrado de volúmenes. Por primera vez se presentó un algoritmo que en todo momento tiene en consideración que las imágenes están afectadas por el *missing wedge*. Se realizó un análisis minucioso utilizando datos artificiales especialmente construidos para emular a los datos reales. Se construyeron modelos biológicamente relevantes de distintos grados de complejidad y se verificó un excelente desempeño bajo niveles de ruido razonables.

Para validar un algoritmo de esta naturaleza es imprescindible mostrar

que es capaz de procesar con éxito imágenes reales. Con este fin, se procesaron tomogramas adquiridos de la partícula groEL ya que están publicamente disponibles mapas de alta resolución que fueron utilizados como *ground truth*. Se obtuvieron muy buenos resultados y se alcanzó a construir un mapa cuya resolución fue acorde a lo esperable, dado el número de partículas con las que se trabajó.

Podemos concluir entonces, que el resultado de este año y medio de trabajo es una herramienta que permite la creación de mapas de alta resolución a partir de imágenes de tomografía electrónica. Resulta natural pensar en que un paso a futuro inmediato es utilizarla para procesar imágenes de material biológico desconocido. Esto está sucediendo mientras se escriben estas páginas y será objeto de otras publicaciones.

Sin embargo la solución encontrada no es más que el punto de partida para nuevas investigaciones. Existen muy diversos puntos donde pueden realizarse nuevos aportes. Para citar algunos de los ejemplos mencionados en el texto, puede profundizarse en la técnica utilizada al calcular los promedios, la reducción de dimensiones y el algoritmo de clasificación utilizado en la etapa de clasificación del *loop* entre muchos otros que seguramente surgirán del estudio de los nuevos conjuntos de datos que están siendo analizados.

Por otro lado recientemente han sido reportados algoritmos en el área del análisis de partículas individuales donde se plantea la clasificación y el promediado como un problema de maximización de la verosimilitud. Resulta una idea muy atractiva estudiar las posibilidades de extenderlo para el caso tridimensional, manteniendo siempre la filosofía de considerar que se está trabajando con volúmenes afectados por el *missing wedge*.

Bibliografía

[1] A. Bartesaghi, G. Sapiro, and S. Subramaniam, "An energy-based three-dimensional segmentation approach for the quantitative interpretation of electron tomograms." *IEEE Transactions on Image Processing*, vol. 14, no. 9, pp. 1314–1323, 2005.

[2] A. Bartesaghi, G. Sapiro, S. Lee, J. Lefman, S. Wahl, S. Subramaniam, and J. Orenstein, "A new approach for 3d segmentation of cellular tomograms obtained using three-dimensional electron microscopy." in *ISBI*, 2004, pp. 5–8.

[3] A. Bartesaghi, P. Sprechmann, G. Randall, G. Sapiro, and S. Subramanian, "Classification and averaging of electron tomography volumes," *ISBI*, 2007.

[4] J. Böhm, A. S. Frangakis, R. Hegerl, S.Ñickell, D. Typke, and W. Baumeister, "Toward detecting and identifying macromolecules in a cellular context: Template matching applied to electron tomograms," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 26, pp. 14 245–14 250, 2000.

[5] J. Fernandez, J. Carazo, and I. García, "Three-dimensional reconstruction of cellular structures by electron microscope tomography and parallel computing," *J. Parallel Distrib. Computing*, vol. 64, pp. 285–300, 2004.

[6] A. Fitch, A. Kadyrov, and J. W.J. Christmas, J. Kittler, "Fast robust correlation," *IEEE Transactions on Image Processing*, vol. 14, no. 8, pp. 1063–1073, 2005.

[7] F. Förster, O. Medalia, N. Zauberman, W. Baumeister, and D. Fass, "Retrovirus envelope protein complex structure in situ studied by cryo-electron tomography." *Proc. Natl. Acad. Sci. USA*, vol. 102, no. 13, pp. 4729–4734, 2005.

[8] A. S. Frangakis, J. Bohm, F. Forster, S.Ñickell, D.Ñicastro, D. Typke, R. Hegerl, and W. Baumeister, "Identification of macromolecular complexes in cryoelectron tomograms of phantom cells," *Proc. Natl Acad. Sci.*, vol. 99, no. 22, pp. 14 153–14 158, 2002.

[9] J. Frank, "The role of correlation techniques in computer image processing." *In Hawkes, P. W. (Ed.), Computer Processing of Electron Microscope Images*, pp. 187–222, 1980.

[10] ——, *Electron Tomography: : Three-dimensional Imaging with the Transmission Electron Microscope.* New York: Plenum Prexss, 1992.

[11] J. Frank, M. Radermacher, P. Penczek, J. Zhu, Y. Li, M. Ladjadj, and A. Leith, "Spider and web: processing and visualization of images in 3d electron microscopy and related welds." *Journal of Structural Biology*, vol. 116, pp. 190–199, 1996.

[12] J. Frank, *Three-Dimensional Electron Microscopy Of Macromolecular Assemblies : Visualization Of Biological Molecules In Their Native State.* Oxford University Press, 2006.

[13] G. Harauz, E. Boekema, and M. van Heel, "Statistical image analysis of electron micrographs of ribosomal subunits." *Methods Enzymol*, vol. 164, pp. 35–49, 1988.

[14] A. K. Jain, M. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, September 1999.

[15] P. Kostelec and D. Rockmore, "FFTs on the rotation group," Santa Fe Institute, Tech. Rep., 2003.

[16] J. A. Kovacsa and W. Wriggers, "Fast rotational matching," *Acta Crystallographica*, vol. 58, no. 8, pp. 1282–1286, 2002.

[17] S. Ludtke, D. Chen, J. Song, D. Chuang, and W. Chiu, "Seeing groel at 6 a resolution by single particle electron cryomicroscopy," *Structure*, vol. 12, no. 7, pp. 1129–36, 2004.

[18] T. MacBobert, "Spherical harmonics; an elementary treatise on harmonic functions with applications," 1927.

[19] A. Makadia and K. Daniilidis, "Rotation recovery from spherical images without correspondences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1170–1175, 2006.

[20] R. McIntosh, D.Ñicastro, and D. Mastronarde., "New views of cells in 3d: an introduction to electron tomography," *Trends Cell Biol.*, vol. 15, no. 1, pp. 43–51, 2005.

[21] F.Ñatterer and F. Wuebbeling, *Mathematical Methods in Image Reconstructions*. Philadelphia: SIAM, 2001.

[22] W. V. Nicholson and R. M. Glaeser, "Automatic particle detection in electron microscopy," *J. Struct. Biol.*, vol. 133, no. 2-3, pp. 90–101, 2001.

[23] S.Ñickell, F. Förster, A. Linaroudis, W. D. Net, F. Beck, R. Hegerl, W. Baumeister, and J. M. Plitzko, "Tom software toolbox: acquisition and analysis for electron tomography," *J. of Struct. Biology*, vol. 149, no. 3, pp. 227–234, 2005.

[24] P. Penczek, M. Radermacher, and J. Frank, "Three-dimensional reconstruction of single particles embedded in ice," *Ultramicroscopy*, vol. 40, no. 1, pp. 33–53, 1992.

[25] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, 2nd ed. Cambridge (UK) and New York: Cambridge University Press, 1996.

[26] M. Schatz and M. van Heel, "Invariant classification of molecular views in electron micrographs." *Ultramicroscopy*, vol. 32, pp. 255–264, 1990.

[27] ——, "Invariant recognition of molecular projections in vitreous ice preparations." *Ultramicroscopy*, vol. 45, pp. 15–22, 1992.

[28] S. Scheres, M. Valle, R.Ñuñez, C. Sorzano, R. Marabini, G. Herman, and J. Carazo, "Maximum-likelihood multi-reference refinement for electron microscopy," *J Mol Biol.*, vol. 348, no. 1, pp. 139–149, 2005.

[29] S. Scheres, H. Gao, M. Valle, G. Herman, P. Eggermont, J. Frank, and J. Carazo, "Disentangling conformational states of macromolecules in 3d-em through likelihood optimization," *Nature Methods*, vol. 4, pp. 27–29, 2007.

[30] M. F. Schmid, A. M. Paredes, H. A. Khant, F. Soyer, H. C. Aldrich, W. Chiu1, and J. M. Shively, "Structure of halothiobacillus neapolitanus carboxysomes by cryo-electron tomography," *J. Mol. Biol.*, vol. 364, pp. 526–535, 2006.

[31] F. Sigworth, "A maximum-likelihood approach to single-particle image refinement." *J Struct Biol.*, vol. 122, no. 3, pp. 328–339, 1998.

[32] C. Sorzano, R. Marabini, J. Velazquez-Muriel, J. Bilbao-Castro, S. Scheres, J. 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, pp. 194–204, 2004.

[33] C. Sorzano, "Algoritmos iterativos de tomografía tridimensional en microscopía electrónica de transmisión." Ph.D. dissertation, Univ. Politécnica de Madrid, España, 2002.

[34] A. Stewart and N. Grigorieff, "Maximum-likelihood multi-reference refinement for electron microscopy," *Ultramicroscopy*, vol. 102, pp. 67–84, 2004.

[35] S. Subramaniam, "Bridging the imaging gap: Visualizing suncellular architecture with electron tomography," *Curr. Opin. Microbiology*, vol. 8, pp. 316–322, 2005.

[36] ——, "The siv surface spike imaged by electron tomography: One leg or three?" *PLoS Pathog*, vol. 2, no. 8, 2006.

[37] S. Subramaniam and J. L. S. Milne, "Three dimensional electron microscopy at molecular resolution," *Annal Review of Biophysics and Biomolecular Structure*, vol. 33, 2004.

[38] G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees, and S. J. Ludtke, "Eman2: An extensible image processing suite for electron microscopy." *J Struct Biol.*, vol. 157, pp. 38–46, 2007.

[39] M. van Heel, "Detection of objects in quantum noise limited images," *Ultramicroscopy*, vol. 8, pp. 331–342, 1982.

[40] M. van Heel, G. Harauz, E. Orlova, R. Schmidt, and M. Schatz, "A new generation of the imagic image processing system." *Journal of Structural Biology*, vol. 116, pp. 17–24, 1996.

[41] M. van Heel and M. Schatz, "Fourier shell correlation threshold criteria." *J Struct Biol.*, vol. 151, no. 3, pp. 250–262, 2005.

[42] M. van Heel and M. Stoffler-Meilicke, "çharacteristic views of e. coli and b. stearothermophilus 30 s. ribosomal subunits in the electron microscope."," *EMBO J.*, vol. 4, pp. 2389–2395, 1985.

[43] M. van Heel, B. Gowen, R. Matadeen, E. V. Orlova, R. Finn, T. Pape, D. Cohen, H. Stark, R. Schmidt, M. Schatz, and A. Patwardhan, "Single-particle electron cryo-microscopy: towards atomic resolution," *Quarterly Reviews of Biophysics*, vol. 33, no. 4, pp. 307–369, 2000.

[44] J. Walz, D. Typke, M.Ñitsch, A. J. Koster, R. Hegerl, and W. Baumeister, "Electron tomography of single ice-embedded macromolecules: Three-dimensional alignment and classification," *J. Struct. Biol.*, vol. 20, pp. 387–395, 1997.

[45] J. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, pp. 236–244, 1963.

[46] H. Winkler, "3D reconstruction and processing of volumetric data in cryo-electron tomography," *J. Struct. Biol.*, vol. 157, pp. 126–137, 2006.

[47] H. Winkler and K. A. Taylor, "Multivariate statistical analysis of three-dimensional cross-bridge motifs in insect flight muscle," *Ultramicroscopy*, vol. 77, pp. 141–152, 1999.

[48] X. Wu, J. Milne, M. Borgnia, A. Rostapshov, S. Subramaniam, and B. Brooks, "A core-weighted fitting method for docking atomic structures into low-resolution maps: application to cryo-electron microscopy." *J. Struct. Biol.*, vol. 141, no. 1, pp. 63–76, 2003.

[49] A. J. Yezzi and S. Soatto, "Deformotion: Deforming motion, shape average and the joint registration and approximation of structures in images." *International Journal of Computer Vision*, vol. 53, no. 2, pp. 153–167, 2003.

[50] P. Zhu, E. Chertova, J. Bess, J. Lifson, L. Arthur, J. Liu, K. Taylor, and K. Roux, "Electron tomography analysis of envelope glycoprotein trimers on hiv and simian immunodeficiency virus virions." *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 15 812–15 817, 2003.

[51] P. Zhu, J. Liu, J. Bess, E. Chertova, J. D. Lifson, H. Grise, G. A. Ofek, K. A. Taylor, and K. H. Roux, "Distribution and three-dimensional structure of aids virus envelope spikes," *Nature*, vol. 441, pp. 847–852, 2006.

Índice de figuras