# Redes neuronales auto-organizativas basadas en optimización funcional. Aplicación en bioinformática y biología computacional.

## TESIS DOCTORAL

**Alberto Domingo Pascual Montano**

**DIRECTOR DE TESIS:**
**José María Carazo García**

**Escuela Técnica Superior de Informática**
**Departamento de Ingeniería Informática**
**Universidad Autónoma de Madrid**

**Madrid, 2002**

*A mi esposa Blanca y a mi hija Andrea*
*A mis padres Carlos y Olivia*
*A Dora, que se me fue*

# Agradecimientos

Desde estas líneas quisiera manifestar mi más profundo agradecimiento a todas aquellas personas e instituciones que de una forma u otra han apoyado la realización de este trabajo.

En primer lugar, mi más profundo agradecimiento y mi mayor gratitud a José María Carazo, director de esta tesis. Sus enseñanzas científicas y su impresionante capacidad han permitido que llegara hoy hasta aquí. Quisiera también mencionar que su ayuda ha traspasado, en numerosas ocasiones, la frontera de la mera relación profesional para convertirse en un apoyo incalculable para mi estabilidad personal. Muchísimas gracias José María por ayudarme en los momentos en que no veía la luz....

A María (Sra. Calle Gil), por su apoyo y ayuda desde el primer día en que me llamó Sr. Montano. Pero sobre todas las cosas, muchas gracias por ser como eres y muchas gracias por darme tu cariño madrina.....

Al Peter, colega y amigo. Gracias por todos esos cafés en los que siempre te argumentas algo, por los ICPRs, por las visitas a la Yuma, y por supuesto, por el excelente binomio **de Alarcón-Pascual-Montano** que formamos!!!

A Mónica Chagoyen, por su siempre acertados consejos profesionales y por su gran valor científico y humano.

A Carlos Oscar, compañero de lucha de Xmipp!. Gracias por ser una enciclopedia técnica andante que en más de una ocasión tuve que consultar....

A Susana Ayerdi, por personificar el concepto de "buena gente". Ojalá existieran más personas como tu.....

A todos los demás miembros de la Unidad de Biocomputación del CNB: Montse (gracias por tu siempre atenta disposición y por tu G*40*P!!!), Luis Enrique (por tu inestimable ayuda científica siempre que la he necesitado), José Jesús (por estar siempre disponible, no importa cuando ni para qué), Sonieta (por ser la system manager más eficiente que ha dado esta tierra, y como no, por tu siempre rápida y amable atención conmigo), JR, Natalia, Maria Gómez, Rafa, Ernesto, Yola, Diego, Jesús y Javi.

Así mismo quisiera agradecer al Centro Nacional de Biotecnología por permitirme trabajar en sus instalaciones durante todo este tiempo. El presente trabajo de tesis también ha sido posible gracias a la financiación de distintos proyectos a los cuales quisiera agradecer: CICYT (BIO2001-1237), NIH (Ref. 1R01HL67465-01) e IIMS (QLRI-CT-2000-31237).

## TABLA DE CONTENIDO

## Prólogo

En esta memoria se sintetiza el desarrollo del trabajo y las aportaciones realizadas en el campo del análisis exploratorio de datos aplicado al procesamiento de datos biológicos. En ella se presenta un sistema para la organización de datos en una representación de menor dimensión, de manera no lineal y no supervisada. Los tipos de métodos presentados aquí son usualmente conocidos como mapas auto-organizativos y son parecidos, aunque no idénticos, a los bien conocidos mapas auto-organizativos de Kohonen. La idea principal está basada en una combinación de técnicas de agrupamiento de datos con métodos de proyección suave de estos en un espacio de dimensión menor.

Esta tesis inicialmente presenta una revisión de varios métodos clásicos de agrupamiento particional y una descripción detallada de los mapas auto-organizativos de Kohonen. Dentro de la revisión se explican detalladamente las bases teóricas y prácticas de estos algoritmos y sus principales ventajas y desventajas en el análisis de datos.

Una vez finalizada la revisión, se presentarán los nuevos algoritmos desarrollados y que constituyen una de las principales aportaciones de esta tesis doctoral. Estos nuevos métodos tienen como objetivo la obtención de algoritmos de proyección no lineal y de cuantificación vectorial basados en funciones de costo bien definidas.

Finalmente se presentan los resultados obtenidos tras aplicar estos nuevos métodos en el campo de la bioinformática y la biología computacional, utilizando problemas reales de clasificación y modelado de imágenes 2D y 3D obtenidas por microscopía electrónica, así como el análisis exploratorio de datos de expresión génica.

La memoria se encuentra estructurada en 3 capítulos generales que contienen 11 secciones en total. En el capítulo I se presenta una breve introducción al problema de análisis de datos en biología, así como una breve introducción al Análisis Exploratorio de datos. Así mismo se presenta una descripción detallada de los métodos de análisis relacionados con esta tesis.

En el capítulo II se presenta, a través de 3 secciones, los nuevos métodos desarrollados como objetivo principal de esta tesis. En estas secciones se exponen en detalles la motivación y los fundamentos matemáticos de estos métodos, así como su comportamiento con datos simulados.

En el capítulo III aborda, a través de 4 secciones, las aplicaciones de los métodos descritos en el capítulo II a problemas reales de análisis de datos en Biología. En cada sección se describe una aplicación distinta, para la cual se introducirá los fundamentos de las técnicas utilizadas, así como la motivación de su estudio. Finalmente, se mostrará y discutirá los resultados obtenidos en cada una de ellas.

Por último se exponen las conclusiones finales y principales aportaciones de este trabajo.

## Abreviaturas

| | | |
|---|---|---|
| 2D | ..... | Bidimensional |
| 3D | ..... | Tridimensional |
| ADN | ..... | Ácido desoxirribonucleico |
| ARN | ..... | Ácido ribonucleico |
| ARN-m | ..... | Ácido ribonucleico mensajero |
| ARN-t | ..... | Ácido ribonucleico de transferencia |
| CA | ..... | Análisis de correspondencia (Correspondence Analysis) |
| crioEM | ..... | Crio-microscopía electrónica |
| DAFC | ..... | Funciones de Doble Auto Correlación (Double Auto Correlation Functions) |
| EM | ..... | Microscopía Electrónica tridimensional (Electron Microscopy) |
| E-M | ..... | Expectation-Maximization |
| MET | ..... | Microscopio Electrónico de transmisión |
| MSA | ..... | Análisis Estadístico Multivariado (Multivariate Statistical Análisis) |
| NMR | ..... | Resonancia Magnética Nuclear (Nuclear Magnetic Resonance) |
| FCM | ..... | c-medias difuso (Fuzzy c-means) |
| FKCN | ..... | Red de agrupamiento de Kohonen difusa |
| FSOM | ..... | Mapa auto-organizativo difuso (Fuzzy SOM) |
| HAC | ..... | Clasificación jerárquica ascendente (Hierarchical Ascendant Classification) |
| IFM | ..... | Músculo de vuelo de insecto (Insect Flight Muscle) |
| KCM | ..... | c-medias tipo núcleo (Kernel c-Means) |
| KerDenSOM | ..... | Mapa auto-organizativo basado en estimación de la densidad de probabilidad (Kernel Density Estimator Self-Organizing Map) |
| K-means | ..... | K-medias |
| PCA | ..... | Análisis de componentes principales (Principal Component Analysis) |
| PDB | ..... | Banco de Datos de Proteínas (Protein Data Bank) |
| PDF | ..... | Función Densidad de Probabilidad (Probability Density Function) |
| RX | ..... | Rayos X |
| SOM | ..... | Mapa auto-organizativo |
| SV40 | ..... | *Simian Virus 40* |
| UV | ..... | Radiación ultravioleta |

# CAPÍTULO I: INTRODUCCIÓN

# 1. Introducción.

## 1.1. La generación de datos en las ciencias de la vida. El reto de su análisis.

En las últimas décadas las ciencias de la vida han experimentado un avance importantísimo gracias al desarrollo acelerado de nuevas técnicas experimentales automatizadas muy poderosas y a la consecuente acumulación de vastas cantidades de información sobre las moléculas y procesos básicos de la vida. Esto, unido al progreso de las investigaciones en los distintos campos de la biología, ha conllevado al crecimiento explosivo de la información biológica generada por la comunidad científica. Un ejemplo muy claro de estos avances lo ha constituido la compleción del proyecto de secuenciación del genoma humano, el cual ha despertado grandes esperanzas en la sociedad con respecto a sus aplicaciones médicas y a la consecuente mejora de la calidad de vida que esto puede originar.

Actualmente existen más de 500 bases de datos públicas que almacenan información biológica de distintos tipos. La base de datos de secuencias de nucleótidos del Laboratorio Europeo de Biología Molecular (EMBL) [1] es una de las más conocidas y utilizadas por la comunidad científica debido a que almacena casi todas las secuencias de nucleótidos públicas existentes. La figura 1.1 muestra el crecimiento de la misma en los últimos 20 años y como puede apreciarse, su crecimiento ha sido exponencial desde que fue creada, duplicando su tamaño cada año.

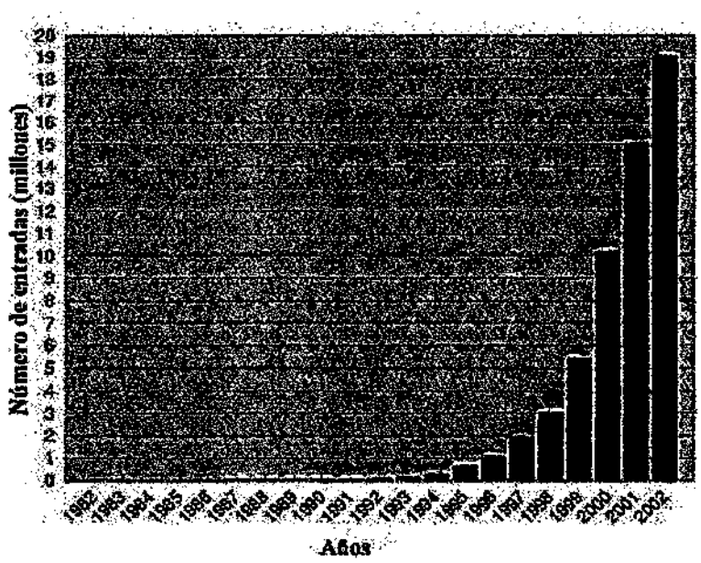



**Figura 1.1** Crecimiento de la base de datos de EMBL en los últimos años.

Otro claro ejemplo de este tipo de bases de datos lo constituye SwissProt [2]. Esta base de datos almacena las secuencias de todas las proteínas que poseen una función conocida. Cada entrada en esta base de datos es anotada manualmente y contiene no solo la información referente a la secuencia de las proteínas, sino también información importante relacionada con la misma, como puede ser descripciones de las funciones con la cual está relacionada, la estructura de sus dominios, modificaciones post-traduccionales, variantes existentes y enlaces a las publicaciones cientfficas relacionadas. La figura 1.2 muestra el ritmo de crecimiento de SwissProt desde que fue creada. Al igual que EMBL, el crecimiento experimentado sigue siendo exponencial.

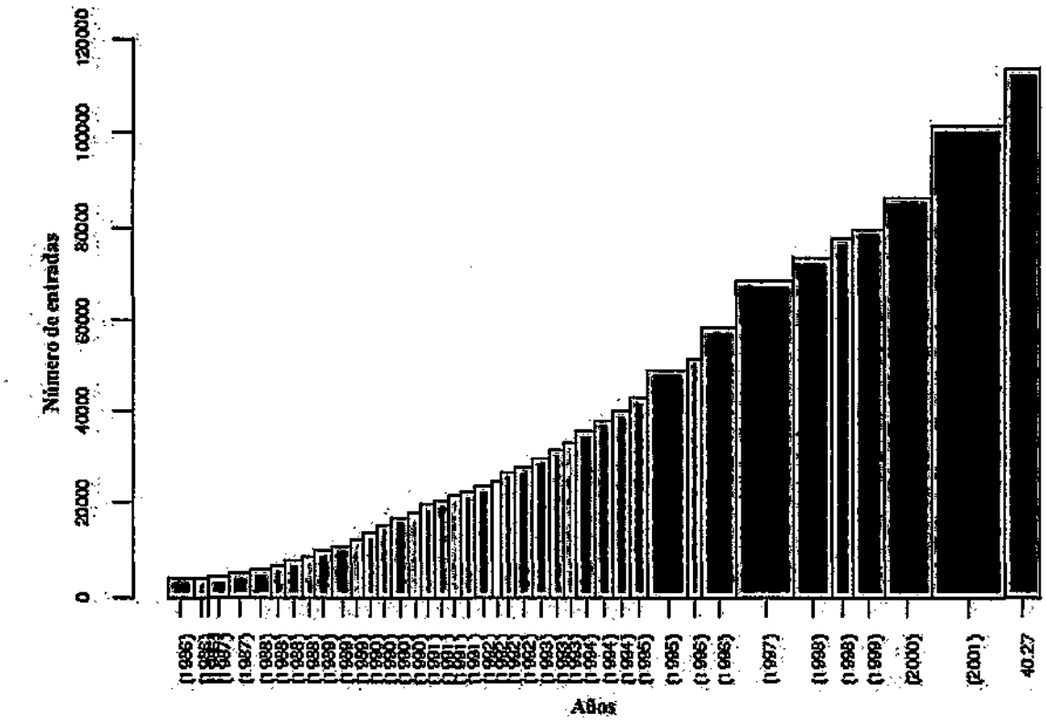



**Figura 1.2** Crecimiento de la base de datos de SwissProt hasta la actualidad..

Todas estas secuencias de ADN y de proteínas almacenadas en bases de datos como EMBL y SwissProt constituyen la base sobre la cual se crean las estructuras moleculares. Es por eso que, paralelamente al desarrollo de las técnicas de secuenciación, se han ido también desarrollando las técnicas de análisis estructural y a pesar de que aún no han llegado a alcanzar la velocidad de análisis de aquellas, también se ha experimentado un avance significativo en los datos producidos.

Del mismo modo que ha venido ocurriendo en el caso de las secuencias, las estructuras moleculares analizadas se han ido recopilando en bases de datos para

permitir a la comunidad científica su consulta y utilización. Un ejemplo de este tipo de repositorio lo constituye el banco de datos de proteínas (Protein Data Bank, PDB) [3]. En esta base de datos se almacenan las coordenadas tridimensionales de los átomos que forman parte de la estructura, así como las interacciones existentes entre ellos.

La principal motivación del análisis estructural radica en intentar comprender mejor los mecanismos físico-químicos por los cuales las moléculas biológicas obtienen su función, así como las diferentes respuestas de las mismas a diversos fármacos con la esperanza de ser capaz de obtener modelos teóricos que faciliten el desarrollo de nuevos medicamentos más efectivos. Debido a la complejidad de las tecnologías experimentales de análisis estructural, la información disponible de estructura de macromoléculas todavía es muy inferior a la de secuencias. Aún así, su crecimiento también ha alcanzado un comportamiento exponencial en los últimos años (figura 1.3).

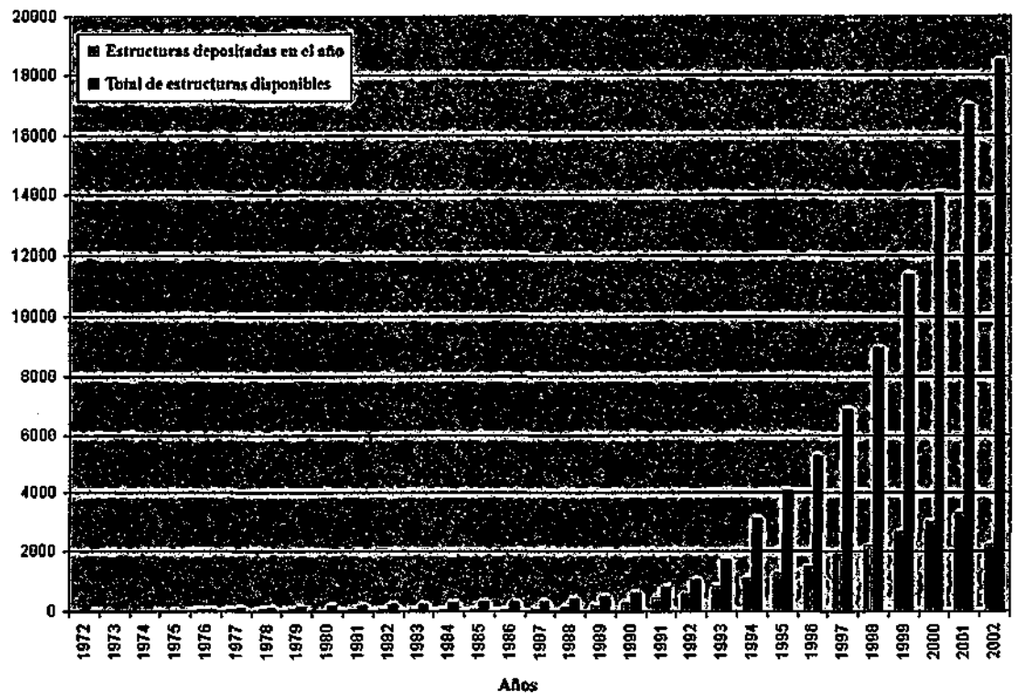



**Figura 1.3** Crecimiento de la base de datos de PDB en los últimos años.

Como se puede apreciar en estos tres ejemplos de fuentes de datos biológicas, el crecimiento experimentado ha sido exponencial y las previsiones son que este ritmo se verá incrementado en varios ordenes de magnitud por la introducción de nuevas tecnologías experimentales que prometen involucrar cambios radicales en la forma de producir los datos. Actualmente existen técnicas experimentales que permiten producir

en un solo experimento la información equivalente a cientos de miles de experimentos tradicionales. Tal es el caso de los microchips de ADN, que ha permitido pasar de analizar genes aislados a trabajar con genomas enteros [4]. Cambios similares están ocurriendo en otros campos, como lo es el análisis de proteínas, en que se está intentando pasar de analizar la función y comportamiento de una proteína a estudiar un gran número de complementos proteicos y enzimáticos de un organismo simultáneamente [5].

El denominador común de estas tecnologías es que generan una figura cada vez más completa de todo el conjunto de interacciones que ocurren simultáneamente en el entorno celular en unas condiciones determinadas. Permitiendo así diferenciar el conjunto de relaciones que ocurren en diferentes tejidos, etapas del desarrollo, fases de una enfermedad, etc. Dadas estas evidentes ventajas es de prever que en un futuro próximo se apliquen de forma generalizada, en especial en la medicina ,en dónde no sólo podrán disponer de un cuadro muchísimo más completo de lo que ocurre dentro de un proceso patológico, sino que además verá incrementada sustancialmente su capacidad diagnóstica y terapéutica.

La bioinformática es una disciplina que se ha desarrollado de forma paralela a la acumulación de la información experimental por los biólogos moleculares para tratar y analizar la dispersa información disponible y se podría definir como una ciencia en la cual la biología, las ciencias de la computación y las tecnologías de la información se unen para formar una sola disciplina. Debido a la complejidad del problema producido por el desbordamiento de datos biológicos que se está generando, los avances de la bioinformática han sido mayores a medida que se ha ido acumulando información que pudiera ser cotejada para extraer significados comprensibles y utilizables. Del mismo modo, a medida que ha ido creciendo la información disponible, ha crecido en complejidad la tarea de compararla e interpretarla, creando la necesidad inmediata de desarrollos en campos de tecnología de la información orientados al almacenamiento, organización e indexado de los datos, así como al desarrollo de herramientas especializadas para su consulta, visualización y análisis. Es por eso que se puede afirmar que en el siglo 21 la biología está sufriendo una transformación de una ciencia puramente experimental hacia una ciencia también de la información.

Los inicios de la bioinformática se relacionan con la creación y el mantenimiento de bases de datos para almacenar la información biológica que se venía produciendo. El desarrollo de este tipo de bases de datos involucraba no solamente

aspectos de diseño, sino también el desarrollo de interfaces complejas a través de las cuales los investigadores pudieran acceder a la información existente, así como actualizar y crear nuevos datos. Sin embargo, con el crecimiento cada vez más acelerado de las bases de datos era de esperar que esta información debía de ser apropiadamente combinada y analizada para formar una imagen global de los procesos biológicos involucrados. Por lo tanto, el campo de la bioinformática ha evolucionado de forma tal que la mayor atención se ha centrado en el análisis y la interpretación de los distintos tipos de datos existentes.

Este proceso de análisis e interpretación de los datos en sí, conocido también como biología computacional, no incluye solamente la aplicación de metodologías de análisis existentes, sino también el desarrollo de nuevas técnicas y métodos que se adapten a la naturaleza compleja de los sistemas biológicos que se estudian. La biología computacional comprende muchas ramas de estudio, entre las cuales podemos destacar las siguientes:

- Análisis de secuencias (tanto de ADN como de proteínas)
- Secuenciación
- Genómica (predicción de estructura genómica, análisis de genoma)
- Análisis de expresión génica
- Proteómica (identificación de proteínas, análisis de expresión)
- Estructura de proteínas (modelado, predicción)
- Interacciones entre proteínas
- Resolución de estructuras tridimensionales por Microscopía Electrónica.
- Análisis filogenético
- Modelado computacional de sistemas biológicos dinámicos (bioinformática integrativa)
- Farmacocinética y Farmacodinámica (PKPD)

Por otra parte, el estudio de estos sistemas biológicos hacen necesario que la biología computacional incluya, además de los campos de estudio mencionados anteriormente, disciplinas tales como la matemática, la estadística, el análisis de imagen, la teoría y el procesamiento de señales, el reconocimiento de patrones, la inteligencia artificial, bases de datos, minería de datos, por solo mencionar algunas.

En el presente trabajo precisamente se pretende estudiar nuevos métodos matemáticos que permitan el análisis masivo de datos biológicos de distintos tipos, con el empeño de ofrecer una aportación al problema descrito anteriormente.

## 1.2. Análisis exploratorio de datos en biología

El presente trabajo de tesis se centra en las tareas de análisis de datos producidos por algunas de las técnicas en el campo de la biología estructural donde el crecimiento y la complejidad de los mismos hace imposible su análisis de manera manual o con metodologías no apropiadas para ello. Esta situación ha motivado nuestro estudio hacia métodos de exploración que permitan, de manera rigurosa, entender la complejidad y variabilidad de la información contenida en estos grandes volúmenes de datos y que permitan extraer información útil para la comprensión de los procesos biológicos que los generan.

Intuitivamente se podría pensar que mientras más datos se posea acerca de un proceso biológico cualquiera, más certeras podrían ser las respuestas a preguntas específicas acerca de la naturaleza estadística de los mismos. Sin embargo, este proceso de análisis no es tan simple cuando los datos no están bien caracterizados, son altamente dimensionales ó cuando el problema a resolver no está bien especificado. En estos casos, el contar con un gran número de datos puede provocar paradójicamente el efecto inverso: mientras más datos se posea, más difícil resulta entenderlos. Este es el caso que ocurre con frecuencia en datos biológicos, donde las técnicas experimentales están generando grandes volúmenes de datos multivariados, con una alta variabilidad y con estructuras cada vez más complejas. Solamente el uso de métodos robustos que sean capaces de descubrir e ilustrar efectivamente las estructuras de estos datos podrían ser utilizados con éxito. Este tipo de métodos, aplicados a grandes conjuntos de datos, es precisamente el tópico de estudio de esta tesis.

Una de las metodologías más utilizadas en los sistemas de análisis y procesamiento es la conocida como Análisis Exploratorio de Datos (EDA), que puede definirse como la búsqueda de evidencias y de modelos estadísticos conducida por los propios datos [6-9]. Los procesos de análisis usualmente comienzan con una etapa de exploración, conducida por los propios datos, seguido de una etapa de confirmación, en la cual la reproducibilidad de los resultados es investigada.

En el campo de las ciencias de la vida, y el especial en la biología, existe una gran variedad de aplicaciones en las cuales el conjunto de datos necesita ser "resumido"

de manera comprensible con el objetivo de obtener información acerca de su estructura. Esto ocurre por la naturaleza de los propios datos, debido a que en la mayoría de las ocasiones no se cuenta con una información a priori sobre la estructura, complejidad, distribución, variación y características de los mismos. Una transformación de los datos de manera que sean fácilmente interpretables, pero a su vez preservando lo mejor posible su estructura y propiedades esenciales es, en muchos casos, un proceso imprescindible. En este tipo de estudio, los EDA pueden jugar un papel muy importante.

Existen distintos métodos de exploración de datos que han sido y todavía son muy utilizados en distintas aplicaciones científicas. A modo de resumen podemos señalar los siguientes:

- *Técnicas gráficas y métodos de visualización de datos multidimensionales* [7, 10]. Estas técnicas están orientadas a la visualización intuitiva de los datos. Como ejemplo podemos señalar: Gráficas de auto-correlación [11], Histogramas, Curvas de Andrews [12], Caras de Chernoff [13], Gráficas de dispersión, etc.
- *Métodos de agrupamiento* [8, 14, 15]. Este tipo de técnicas permiten reducir la cantidad de datos analizados mediante el agrupamiento de los mismos en distintos grupos estructuralmente homogéneos.
- *Métodos de proyección.* La intención de estos métodos es reducir no el número de datos, sino la dimensión de los mismos. El objetivo principal es representar los datos originales que se encuentran en una dimensión elevada en una dimensión mucho menor, pero conservando sus mismas propiedades estadísticas. Estas técnicas de proyección no solo reducen la complejidad del problema, sino que también facilitan las tareas de visualización de los mismos al ser representados en un espacio de bajas dimensiones. A modo de ejemplo podemos señalar los siguientes: Análisis por Componentes Principales (PCA) [16], Projection Pursuit [17, 18], Multidimensional Scaling (MDS) [19], Proyección de Sammon [20], Curvas Principales [21] y los Mapas auto-organizativos (SOM) [22].

Debido a que el campo de análisis y exploración de datos es muy amplio, en este trabajo de tesis hemos centrado nuestro estudio en métodos de exploración basados en redes neuronales auto-organizativas (SOM). La principal motivación para este tipo de

estudios viene dada por la capacidad de este tipo de técnicas para la representación de los datos en espacios de menores dimensiones, pero conservando la estructura y las relaciones entre ellos. Adicionalmente, esta técnica puede ser utilizada tanto como método de agrupamiento para reducir el número de datos como método de proyección no lineal a un espacio de menor dimensión. Estas propiedades lo convierte en una herramienta muy atractiva para el análisis exploratorio. En las sección 2 de esta memoria, se hará una descripción detallada de las características teóricas y prácticas de este método, así como una descripción de los principales problemas de los que adolece. Adicionalmente, debido a la relación con los nuevos algoritmos que se proponen, en la sección 3 y 4 se hará una descripción detallada de algunos de los métodos de agrupamiento más utilizados, así como de técnicas estadísticas de estimación de densidad de probabilidad también relacionadas con el análisis exploratorio de datos.

## 1.3. Planteamiento general de los objetivos

Los mapas auto-organizativos mencionados en el apartado anterior y que serán descrito en detalles en la sección 2, a pesar de que son ampliamente utilizados en análisis exploratorios, sufren de varios problemas importantes debido fundamentalmente a la ausencia de una formulación matemática adecuada que permita el estudio de sus propiedades teóricas. Es por eso que uno de los objetivos propuestos en esta tesis doctoral es el planteamiento de una metodología completamente diferente para construir nuevos mapas auto-organizativos a partir de funciones de costo bien planteadas matemáticamente y que expresen explícitamente sus características fundamentales. De esta forma intentamos resolver varios problemas científicos importantes en este contexto, encontrar una explicación teórica al método de SOM e integrar de manera objetiva los métodos de agrupamiento y proyección pero conservando las propiedades estadísticas de los datos.

Uno de los métodos que se propone en este trabajo consiste en una versión modificada del funcional de un conocido algoritmo de agrupamiento difuso, donde los centros de grupos o vectores representantes se encuentran distribuidos en un espacio de baja dimensionalidad y para lo cual se modifica el funcional para garantizar una distribución suave de los valores de los vectores representantes en ese espacio de baja dimensión. Adicionalmente, se propone otro funcional basado en la estimación no paramétrica de la función densidad de probabilidad, de manera que los vectores

representantes, generados en este caso, tienden a poseer la misma distribución estadística de los datos originales.

Así mismo, se propone también la aplicación de estos nuevos métodos a la resolución de distintos problemas de biología computacional y bioinformática. Específicamente en problemas de clasificación y agrupamiento de imágenes de microscopía electrónica tridimensional, clasificación de volúmenes 3D de tomografía electrónica, análisis y modelado de imágenes 3D de macromoléculas biológicas y análisis de patrones de expresión génica.

A modo de resumen, las contribuciones que aporta esta tesis doctoral son las siguientes:

- Una nueva metodología para la construcción de mapas auto-organizativos basados en optimización funcional.
- Un algoritmo que implementa una nueva red neuronal auto-organizativa difusa.
- Un nuevo algoritmo de cuantificación de vectores basado en la estimación no paramétrica de la función densidad de probabilidad.
- Un nuevo mapa auto-organizativo basado en la estimación no paramétrica de la función densidad de probabilidad.
- Aplicación experimental de los algoritmos propuestos en tareas de clasificación y agrupamiento de imágenes de microscopía electrónica tridimensional.
- Aplicación metodológica de los algoritmos propuestos en tareas de clasificación y agrupamiento de volúmenes obtenidos por tomografía electrónica.
- Creación de una nueva metodológica para el modelado geométrico y topológico de complejos biológicos tridimensionales.
- Aplicación experimental de los mapas auto-organizativos propuestos en el análisis y agrupamiento de datos de expresión génica.

## 2. Mapas auto-organizativos

Las redes neuronales son sistemas muy útiles para la clasificación y el reconocimiento de patrones en grandes grupos de datos. Uno de los tipos de redes neuronales más utilizados son los mapas auto-organizativos (SOM), cuyo propulsor fundamental ha sido Teuvo Kohonen [23]. Este tipo de redes intenta simular el hipotético proceso auto-organizativo que ocurre en el cerebro humano cuando le es presentado un estímulo externo. SOM realiza una proyección de un conjunto de datos de entrada sobre un conjunto de vectores de salida, usualmente distribuidos en una red regular de baja dimensionalidad (generalmente una malla bidimensional), pero esta proyección tiene la peculiaridad de ser ordenada de acuerdo a las características de los datos de entrada, es decir, la vecindad relativa de los datos de entrada se intenta preservar en el espacio de salida.

La estructura de una red neuronal tipo SOM está representada en la figura 2.1 y será descrita con más detalles en secciones posteriores de este trabajo de tesis. Brevemente, la red neuronal está básicamente está formada por dos capas: una de entrada y otra de salida. La capa de entrada está compuesta por un conjunto de neuronas correspondientes a cada variable o componente del vector de entrada y la capa de salida por un conjunto de neuronas de salida interconectadas de forma tal que forme una malla regular de topología arbitraria. Cada neurona contiene un vector de coeficientes asociado y que posee la misma dimensión de los datos de entrada. Este vector asociado a cada neurona se conoce como vector diccionario.

En los SOMs todos los nodos o neuronas del mapa reciben el mismo vector de entradas, y de todos los nodos que forman el mapa, sólo uno será el responsable de generar la salida, y será aquel cuyo vector de pesos sea más parecido a la entrada actual (menor distancia euclídea). En cuanto a la topología de vecindad entre los nodos, esta puede ser muy variada:

- lineal,
- lineal en forma de anillo,
- plana con retículo rectangular,
- plana con retículo hexagonal,
- toroidal,
- etc.

También es posible tener mapas autoorganizados con topologías de dimensiones más altas, pero la utilización, y sobre todo la visualización de los resultados en dimensiones superiores a dos resulta más incómoda ó simplemente impracticable.

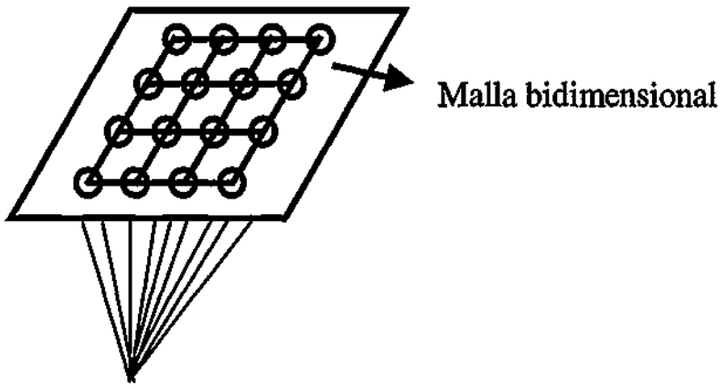



**Figura 2.1** Estructura del mapa auto-organizativo de Kohonen. Las neuronas en la capa de salida están interconectadas entre sí en un espacio de baja dimensionalidad, como por ejemplo una malla. La topología de esta malla puede ser cualquiera: rectangular, hexagonal, toroidal, etc.

Las propiedades prácticas de los mapas auto-organizativos hacen que se conviertan en herramientas poderosas para el análisis de datos en cualquier campo de la ingeniería o las ciencias, permitiendo el proceso, visualización y agrupamiento de grandes cantidades de datos. Las propiedades de preservación de topología y reducción de dimensionalidad hacen del SOM un método imprescindible en la clasificación de entidades donde aparecen grandes números de datos y de clases y donde en muchas ocasiones la transición de una clase a la otra es prácticamente continua, sin separación clara ente ellas.

La funcionalidad de SOM podría ser brevemente descrita de la siguiente manera: cuando se le presenta a la red un dato de entrada, las neuronas en la capa de salida compiten entre sí y la neurona ganadora, cuyo valor sea más parecido al dato de entrada, así como un conjunto de neuronas vecinas actualizan sus valores. Este proceso se repite hasta que se alcanza un criterio de parada, usualmente cuando los valores de las neuronas se estabilizan o cuando se alcanzan un número determinado de iteraciones.

## 2.1. El algoritmo de Kohonen

Matemáticamente, el algoritmo de Kohonen puede ser descrito de la siguiente forma:

Sea $\mathbf{X}_i \in \Re^p, i = 1 \cdots n$ un conjunto de datos de dimensión $p$ y $\mathbf{V}_j \in \Re^p, j = 1 \cdots c$, un conjunto de vectores diccionarios. La regla de actualización de Kohonen es la siguiente:

$$\mathbf{V}_{j,t} = \mathbf{V}_{j,t-1} + \alpha_t h_{r,t} \left( \mathbf{X}_t - \mathbf{V}_{j,t-1} \right) \tag{2.1}$$

donde $\alpha_t$ representa el factor de aprendizaje, el cual es definido como una función decreciente que controla la magnitud de los cambios en cada iteración $t$ y $h_{r,t}$ es una función que controla el tamaño de la vecindad de los nodos a ser actualizados durante el entrenamiento. Ambos parámetros $\alpha_t$ y $h_{r,t}$ decrecen monótonamente durante el entrenamiento con el objetivo de lograr la convergencia. Las formas explícitas mas comúnmente utilizadas para ambos parámetros son:

$$\alpha_t = \alpha_0 (\frac{iter - t}{iter}) \tag{2.2}$$

donde *iter* es el número máximo de iteraciones y $\alpha_0$ es el valor inicial del factor de aprendizaje.

$$h_{rt} = e^{\left( \frac{\|r_j - r_w\|^2}{2\sigma_t^2} \right)} \tag{2.3}$$

y

$$\sigma_t = 1 + (R_0 - 1)(\frac{iter - t}{iter}) \tag{2.4}$$

$R_0$ es el radio inicial de la vecindad a ser actualizada y $\|r_j - r_w\|$ es la distancia en el mapa entre el nodo $j$ (el nodo que está siendo actualizado) y el nodo $w$ (el nodo ganador).

El esquema del mapa auto-organizativo de Kohonen mostrado en la figura 2.1 está formado por dos capas: una capa de entrada con $p$ neuronas, representando cada una de ellas las $p$ variables de los datos de entrada y una capa de salida formado por $c$ neuronas interconectadas entre sí de manera que forman una malla regular. Es importante señalar que los vectores asociados a estas neuronas de la capa de salida suelen denominarse *"vectores diccionarios"* y será el término con que las trataremos a lo largo de esta memoria. La malla puede tener cualquier dimensión, aunque las más utilizadas son las mallas bidimensionales, pues precisamente unos de los objetivos principales de SOM es reducir la dimensionalidad a un espacio menor donde la

visualización pueda ser directa. Durante el desarrollo de la presente memoria asumiremos, a menos que se especifique lo contrario, que la malla de salida será siempre de bidimensional.

El algoritmo de Kohonen es el siguiente:

1. Fijar el tamaño del mapa ($c$ vectores diccionarios);
   Fijar radio inicial $R_0$;
   Fijar factor de aprendizaje inicial $\alpha$;
   Fijar número de iteraciones $t$.

2. Inicializar los vectores diccionarios (V) de manera aleatoria.
3. Presentar un dato de entrada $\mathbf{X}_i$
4. Calcular la distancia del dato $\mathbf{X}_i$ a todos los nodos $\mathbf{V}_j$ en la iteración $t$:

$$d_{ij} = \left\| \mathbf{X}_i - \mathbf{V}_j \right\| \tag{2.5}$$

5. Seleccionar el nodo $w$ con la distancia $d_{ij}$ mínima.

$$w = w(\mathbf{X}) = \arg\min_j \left\{ \left\| \mathbf{X} - \mathbf{V}_j \right\| \right\} \tag{2.6}$$

6. Actualizar el vector diccionario del nodo $w$ y a todos sus vecinos utilizando la ecuación (2.1).
7. Ir al paso 3 hasta que el algoritmo converja.

## 2.2. Propiedades interesantes de los mapas auto-organizativos

En el contexto del análisis exploratorio de datos es muy frecuente ver los métodos más utilizados divididos en dos grandes categorías: métodos de agrupamiento (clustering) y métodos de proyección. Los mapas auto-organizativos representan un caso especial de técnica que puede ser utilizada tanto como método de agrupamiento para reducir el número de datos y como método de proyección no lineal a un espacio de menor dimensión. Esto lo convierte en uno de los métodos más utilizados en distintos campos de la ciencia y la tecnología [24].

Por virtud de su algoritmo de aprendizaje, SOM puede ser interpretado como una regresión no lineal de los vectores diccionarios en el espacio de entrada, formando de esta forma una especie de malla elástica bidimensional que intenta seguir la distribución de los datos en su espacio original. Esta naturaleza ordenada de los vectores de referencia obtenidos por este algoritmo justifica la utilización de los SOM como un visualizador avanzado de datos. La razón es evidente: datos vecinos en el espacio de entrada son proyectados a nodos vecinos en el mapa, por lo tanto esta proyección

ordenada facilita la visualización y comprensión de la estructura de los datos en un espacio reducido. Teuvo Kohonen, el autor de SOM, fue el primero en proponer la utilización de estos mapas como visualizador de datos [25].

Una de las aplicaciones más comunes de SOM como visualizador es la de mostrar la distribución de grupos (clusters) presentes en un conjunto cualquiera de datos. Esto es posible ya que la densidad de los vectores de referencia (vectores diccionario) de un mapa refleja la densidad de los datos de entrada [22, 26]. Este efecto provoca que en zonas de alta densidad de lo datos originales (grupos) los vectores diccionarios del mapa se encontrarán más cerca uno de otros. Por el contrario, en aquellas zonas de baja densidad de los datos, que representan zonas de transición entre grupos, los vectores diccionarios se encontrarán mucho más dispersos. Por lo tanto, visualizar la estructura de grupos en un SOM es relativamente fácil si se muestran las distancias relativas entre vectores diccionarios vecinos en el mapa [27-31].

La figura 2.2 muestra un ejemplo de utilización de SOM con un conjunto de 30 imágenes creadas artificialmente. Las imágenes representan caras con distintas características de color (niveles de grises) en la nariz, ojos y boca. Adicionalmente, a estas imágenes se le agregó ruido aleatorio para demostrar la capacidad de SOM ante un problema de agrupamiento con datos de alta dimensión y en presencia de ruido. La figura 2.2a muestra la galería de imágenes utilizada. En este caso se utilizó un SOM de topología hexagonal de 5x5 que fue inicializado de manera aleatoria con datos del conjunto de entrada (figura 2.2b). Después de 2000 iteraciones (figura 2.2c) comienza a observarse un cierto ordenamiento de las imágenes representantes, hasta que finalmente el mapa converge a una solución donde los conjuntos homogéneos de imágenes de entrada son claramente diferenciables (figura 2.2d). El mapa final refleja la naturaleza del proceso auto-organizativo: vectores diccionarios vecinos representan a imágenes parecidas y vectores diccionarios distantes representan precisamente conjuntos distintos de imágenes. Este agrupamiento, conjuntamente con la posibilidad de una visualización directa de la estructura de los datos en el mapa, hacen de este método una poderosa herramienta para el análisis exploratorio de datos.

Otra de las propiedades más interesantes y atractivas que caracterizan los mapas auto-organizativos son su robustez ante la presencia de puntos atípicos (outliers). Estos puntos atípicos aparecen muy frecuentemente en cualquier proceso físico de medición o toma de datos y se caracterizan por aparecer significativamente distantes del cuerpo principal de los mismos. El tratamiento de este tipo de datos es un tema de investigación

muy importante debido a que mucho de los métodos de análisis se ven sensiblemente afectados por este conjunto de puntos, que a pesar de ser generalmente muy reducido, afecta considerablemente los resultados. SOM, por su propia naturaleza, es un método capaz de tratar eficientemente los puntos atípicos debido al hecho de que estos afectan únicamente a una sola neurona y en menor grado a sus vecinas, mientras que el resto del mapa representa la mayor fuente de varianza de los datos. Inspeccionando estas neuronas afectadas es posible detectar estos datos atípicos al mismo tiempo que el análisis del resto de los elementos se mantiene intacto. Esto permite que estos puntos sean descartados, o en muchos casos, analizados independientemente [32, 33].

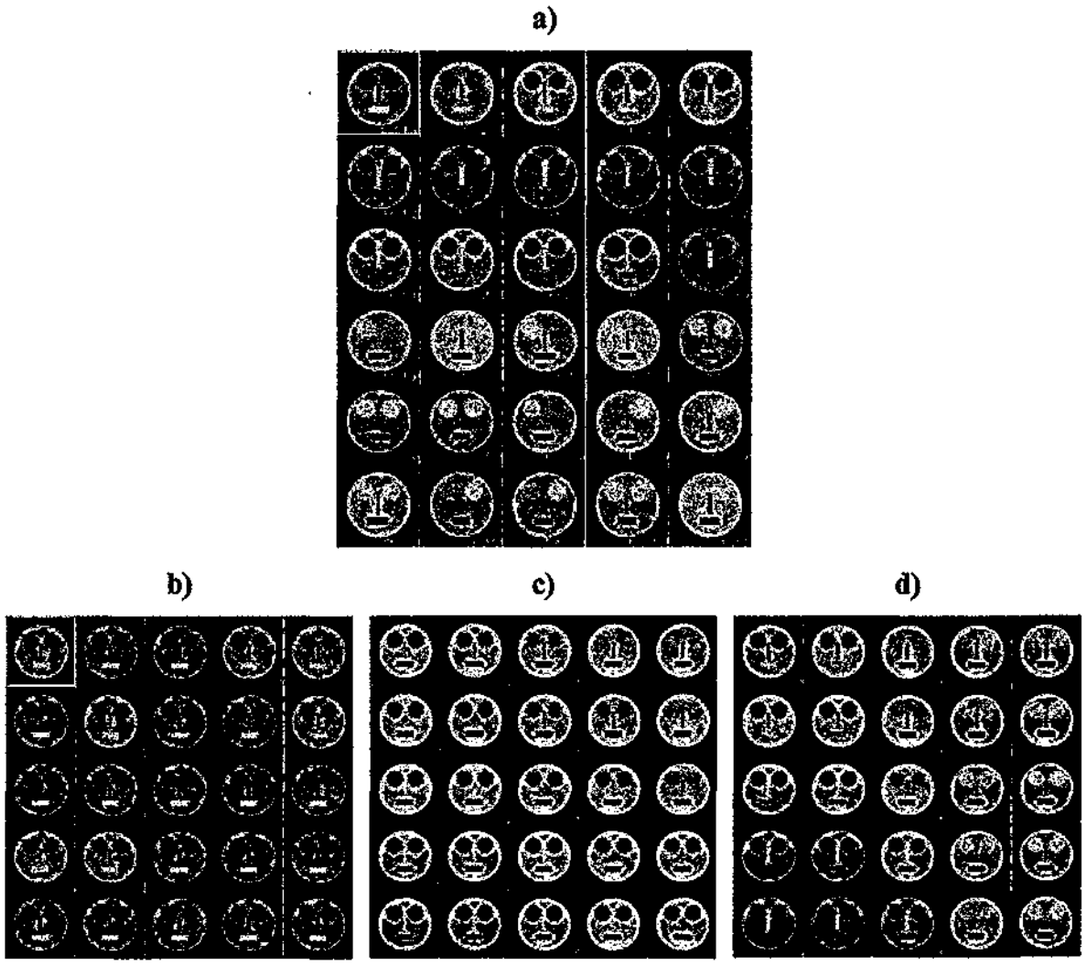



**Figura 2.2** Ejemplo de distintas fases de entrenamiento de SOM. a) Conjunto de 30 imágenes de caras sintéticas y ruidosas con distintas características de color (niveles de grises) en la boca, nariz y ojos. b) SOM inicial (sin entrenar) de 5x5 cuyos vectores diccionarios se inicializaron con imágenes del conjunto original de datos tomadas de manera aleatoria. c) SOM después de 2000 iteraciones. d) Resultado final de SOM después de 5000 iteraciones.

## 2.3. Fundamentos matemáticos del algoritmo de Kohonen

A pesar de que el principio de los mapas auto-organizativos es bastante simple en su descripción e implementación práctica, el comportamiento global del proceso, especialmente en presencia de datos complejos, es muy difícil de describir en términos matemáticos rigurosos. Esto significa que algunas de sus propiedades teóricas todavía son objeto de estudio por numerosos investigadores. Según un trabajo de revisión reciente [34], solo en el caso particular de datos en una dimensión y mapa con topología lineal y abierta (dimensión 1 en forma de cadena) ha sido completamente descrito. Informalmente podríamos decir que *"el algoritmo de SOM usualmente funciona bien, pero no se sabe a fondo por qué"*.

Tal y como se ha descrito en los apartados anteriores, en el algoritmo de SOM se define una estructura de conexión entre las neuronas del mapa y esta estructura se tiene en cuenta durante el proceso de aprendizaje para lograr la conservación de las relaciones espaciales. Los vectores diccionarios asociados a cada neurona de la capa de salida de SOM se van actualizando cada vez que se presenta un nuevo dato de forma tal que puntos vecinos en el espacio de entrada son proyectados sobre la misma neurona o sobre neuronas vecinas en el espacio de salida.

Este proceso práctico descrito por el algoritmo de Kohonen permite definir dos fases claramente identificables:

- *Auto-organización:* El mapa se comienza entrenando con valores de vecindad inicial y factor de aprendizaje elevados
- *Convergencia:* Los valores de los vectores diccionarios son adaptados de manera tal que se cuantifica fielmente el espacio de entrada. En esta fase la vecindad de nodos a ser adaptados se reduce paulatinamente a uno, de manera que cuando el número de iteraciones aumenta mucho, solo el nodo ganador es actualizado. Adicionalmente, el factor de aprendizaje se hace tender a cero para provocar la convergencia.

En este contexto es de esperar que en la fase de convergencia la auto-organización lograda en la primera fase no sea modificada, incluso cuando solamente se actualice el nodo ganador. Estas propiedades fácilmente deducibles del algoritmo deben cumplirse también en los análisis teóricos de este método y precisamente los esfuerzos de demostración de la propiedades matemáticas del algoritmo han ido encaminados en ese sentido. En los trabajos originales de Kohonen [22, 23, 35] se puede encontrar una

demostración preliminar de estos argumentos. Sin embargo, la primera prueba completa de las propiedades de auto-organización y convergencia para distribución uniforme de las entradas y para una función de vecindad de paso simple fue propuesta por Cotrell y Fort [36] en 1987. Este primer trabajo fue también desarrollado para el caso en que tanto los datos como el mapa se encuentran en dimensión 1 y está basado en la aplicación de resultados conocidos de la teoría de los procesos de Markov, ya que el algoritmo de SOM está estrechamente ligado con un tipo especial de estos procesos. La longitud de este trabajo es de unas 40 páginas lo que de cierta forma demuestra la complejidad del problema de demostrar rigurosamente el algoritmo de Kohonen, incluso para el caso más simple de dimensión 1.

Los principales intentos para la comprensión teórica del algoritmo de SOM se han basado en la solución del siguiente problema inverso: "*Encuéntrese el funcional (o función de costo) cuya optimización numérica corresponda al algoritmo de SOM*". La razón de que la demostración de auto-organización y convergencia del algoritmo de SOM exista para el caso unidimensional es obvia, ya que el concepto de ordenamiento en el caso unidimensional es trivial y además es también posible establecer una función de costo para este caso. Por ejemplo, simplificando el problema a modo de clarificar intuitivamente la idea podemos mostrar la siguiente analogía:

Sea $\mu_i \in \Re$, $i = 1...n$, un conjunto de números escalares. La siguiente función:

$$J = \sum_{i=2}^{n} |\mu_i - \mu_{i-1}| - |\mu_n - \mu_1| \tag{2.7}$$

solo puede alcanzar un mínimo (valor cero) si y solo si los valores $\mu_i$ están numéricamente ordenados ascendente o descendentemente.

Este tipo de ordenamiento es el que puede ocurrir en un mapa auto-organizativo si los datos de entrada son unidimensionales y la red está formada por un topología lineal de nodos de una sola dimensión. En una analogía con el ejemplo mostrado, esto implica que los $\mu_i$ se correspondan con los $V_i$ (vectores diccionarios). Este estado donde $J = 0$, se llama estado absorbente y una vez alcanzada esta condición, ya no podría ser cambiada durante la fase de entrenamiento por ningún otro dato de entrada.

En dimensiones mayores que uno, sin embargo, la demostración de la existencia de este estado absorbente es una tarea bastante complicada, especialmente en el caso donde la dimensión de los datos de entrada es mucho mayor que la dimensión del mapa de salida. Muchos autores han intentado demostrar la convergencia del algoritmo en

estas condiciones de dimensionalidad, pero generalmente las demostraciones se han basado en casos particulares de distribuciones específicas de entrada y funciones muy concretas de vecindad [37-42].

Otros intentos en este sentido han propuesto el análisis del algoritmo de SOM en términos de un sistema de funciones de energía para estudiarlo y explicarlo teóricamente. Por esta vía se ha demostrado que el mecanismo de entrenamiento de SOM es equivalente a minimizar un conjunto de funciones de energía sujetas a restricciones, explicándose de esta forma la habilidad de este método en formar mapas topológicamente correctos [39, 43-45]. A pesar del gran esfuerzo de todos estos trabajos, el problema para el caso general sigue sin solución.

## 3. Técnicas de agrupamiento particional

El objetivo de las técnicas de agrupamiento es reducir un conjunto elevado de datos por medio de una categorización en grupos más o menos homogéneos. Este tipo de agrupamiento está motivado por la forma en que los humanos procesamos la información, intentando siempre reducir la cantidad de información a procesar mediante la eliminación de redundancias y la agrupación de objetos con propiedades comunes en una misma categoría. En el contexto de exploración de datos los métodos de agrupamiento siempre han ocupado un lugar especial motivado por la idea de contar con herramientas matemáticas y computacionales que ayuden a la construcción automática de estas categorías [46] [47]. Los métodos de agrupamiento intentan satisfacer, entre otras, las siguientes características:

- Cada grupo es homogéneo y los datos asignados a cada grupo son similares entre sí.
- Cada grupo debe ser diferente del resto de los grupos, lo que implica que los datos o elementos asignados a cada uno de los grupos deben ser también diferentes de aquellos asignados al resto de los grupos.

Actualmente existen muchísimas variantes de técnicas de agrupamiento basadas en distintos criterios y dependiendo de cada tipo, los grupos pueden ser expresados de distintas formas [8, 14, 15, 46-48]. Por ejemplo, las técnicas de agrupamiento pueden ser divididas de manera resumida en una de las siguientes categorías:

- *Particionales:* Estos métodos intentan construir varias particiones de los datos donde cada una de ellas es evaluada por distintos criterios. La función criterio que generalmente tratan estos algoritmos de minimizar refleja la estructura local de los datos.
- *Jerárquicas:* Se crea una descomposición jerárquica de los datos de acuerdo con algún criterio determinado. En este grupo pueden existir métodos divisivos (grandes grupos se dividen sucesivamente de acuerdo a ciertos criterios predeterminados) y aglomerativos (pequeños grupos son unidos según un criterio predeterminado para formar grupos más grandes).

- *Basadas en densidad:* Estos métodos se basan en estimaciones de la densidad de los datos para detectar regiones que sean compactas y densas.
- *Basadas en rejilla (grid):* Cuantifican y dividen el espacio de datos en un conjunto finito de celdas formando una estructura de rejilla. Una vez creada, todas las operaciones de detección de grupos son realizadas sobre la rejilla.
- *Basadas en modelo:* En este tipo de métodos los grupos están basados en un modelo teórico hipotético y el objetivo es encontrar el mejor ajuste de los datos a cada uno de estos modelos.

Adicionalmente, dependiendo del tipo de método utilizado, los grupos pueden clasificarse de las siguientes maneras:

- *Exclusivos:* Los datos pertenecen exclusivamente a uno y solo uno de los grupos.
- *Difusos:* Se permite un cierto grado de solapamiento entre los grupos. Un dato no pertenece exclusivamente a un solo grupo, sino a todos pero con distinto grado de pertenencia.

Debido a que uno de los objetivos principales de esta memoria es presentar un conjunto de métodos de exploración de datos basado en los principios básicos de los mapas auto-organizativos y de la cuantificación vectorial, es importante describir los métodos particionales más conocidos y más comúnmente utilizados que presenten una relación directa con el método de SOM y, por tanto, con los métodos que presentamos en este trabajo. Por esta razón, en los apartados siguientes presentaremos en detalles solamente dos de estos métodos: el bien conocido método de k-medias (k-means) [49] y una generalización difusa del mismo llamada método de c-medias difuso (Fuzzy c-means) [50]. Estos métodos son también conocidos con nombres muy similares: k-medias duro (hard k-means) y k-medias difuso (Fuzzy k-means). Adicionalmente, presentaremos un algoritmo que combina la técnica de agrupamiento de c-medias difuso con los mapas auto-organizativos de Kohonen para conseguir un método de agrupamiento con propiedades interesantes heredadas de ambos métodos. Esta técnica, conocida como redes de agrupamiento difusas de Kohonen (Fuzzy Kohonen Clustering Network, FKCN) [51], evidencia los esfuerzos de la comunidad científica en conseguir

la unión de conceptos aparentemente distintos entre sí en un contexto matemático riguroso.

## 3.1. El método de k-medias (*k-Means*)

El método de k-medias [49] es uno de los métodos de agrupamiento particional más simples de los utilizados en análisis de datos. En esta técnica el problema matemático que se intenta resolver es la minimización de la distancia cuadrática media de los datos a cada uno de los centroides o centros de grupo. Matemáticamente:

Sea $\mathbf{X}_i \in \Re^p, i = 1 \cdots n$ un conjunto de datos de dimensión $p$ y $\mathbf{V}_j \in \Re^p, j = 1 \cdots k$ ($1 < k < n$) un conjunto de centros de grupo o prototipos. Una partición de $\mathbf{X}$ en $k$ conjuntos disjuntos $S_j$ de $N_j$ datos puede ser expresada mediante la siguiente función de costo:

$$J = \sum_{j=1}^{k} \sum_{i \in S_j} \left\| \mathbf{X}_i - \mathbf{V}_j \right\|^2 \tag{3.1}$$

donde $\mathbf{X}_i$ es un vector que representa el $i$-ésimo dato y $\mathbf{V}_j$ es el centroide del conjunto de puntos $S_j$.

Un algoritmo que minimiza la función de costo expresada por la ecuación (3.1) consiste en un proceso alternativo de re-estimación. Primeramente, los datos son asignados aleatoriamente a los $k$ grupos. Seguidamente, los centroides de cada grupo son calculados. Estos dos pasos son repetidos alternadamente hasta que se alcanza algún criterio de parada, usualmente cuando las variaciones de los centroides entre iteraciones sea muy pequeña o cuando se alcanza un número prefijado de iteraciones. De manera resumida el algoritmo es el siguiente:

1. Asignar valores aleatorios iniciales a los $k$ centroides.
2. Asignar cada dato a su centroide más cercano de acuerdo con algún criterio de distancias, formando de esta forma $k$ nuevos grupos de datos $Sj; j = 1..k$.
3. Calcular los nuevos centroides de cada grupo (media de los atributos de los datos de cada grupo)
4. Si la diferencia entre los centroides calculados en la iteración anterior y la actual es significativa (mayor que un valor prefijado) ir al paso 2. De lo contrario el algoritmo termina.

La figura 3.1 muestra un ejemplo del algoritmo de k-medias con un conjunto artifical de 14 datos en dos dimensiones. Es importante señalar que, debido a la

naturaleza del proceso de optimización empleado, el algoritmo señalado anteriormente no garantiza que se alcance un mínimo global de la función, lo que implica que los resultados obtenidos en muchas ocasiones pueden no ser óptimos. Adicionalmente, el algoritmo necesita que el número de grupos a ser extraído sea prefijado de antemano. Esto impone una seria restricción en el análisis de datos, ya que en la mayoría de los casos de análisis reales no se dispone de esta información. Sin embargo, pese a estas evidentes limitaciones, este algoritmo es muy utilizado debido a su simplicidad y a su fácil implementación.

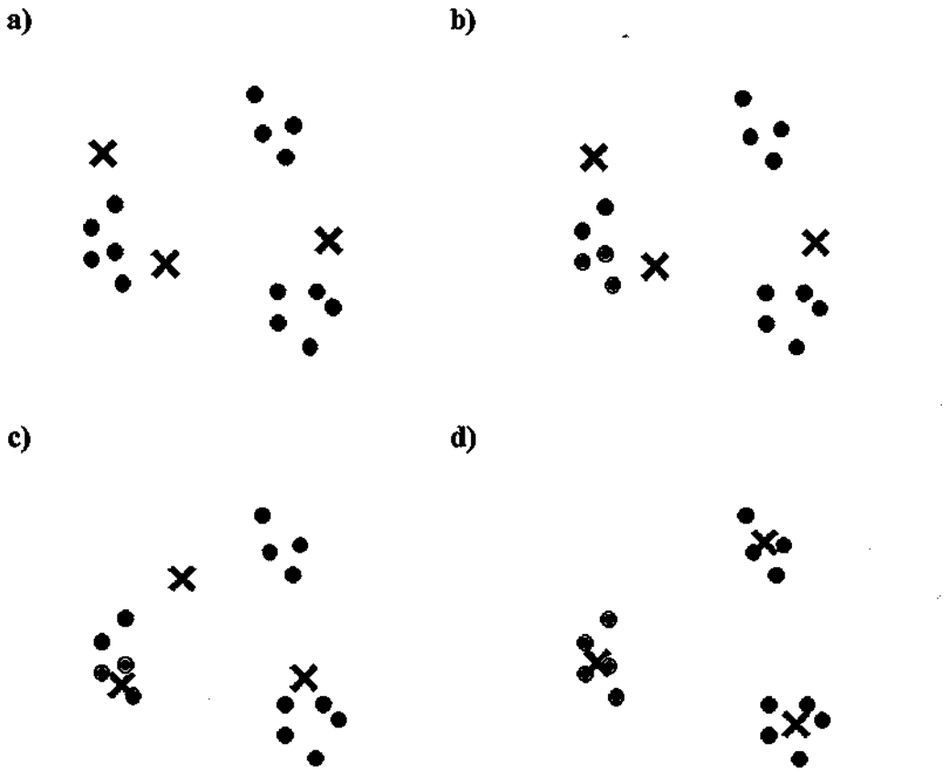



**Figura 3.1** Ejemplo del algoritmo k-medias para k = 3. Los círculos representan datos 2D y las cruces representan los centroides de cada uno de los 3 grupos. a) Los centroides se inicializan aleatoriamente. b) Cada dato se asigna a su centroide más cercano. c) La posición de los centroides se recalcula a partir de los datos asignados al grupo que ellos representan. d) Los pasos b y c se repiten hasta que los centroides no varíen más su posición.

## 3.2. El método de c-medias difuso (*Fuzzy c-Means*)

El método de agrupamiento de c-medias difuso (Fuzzy *c*-means, FCM) [50, 52] es un proceso de agrupación de objetos en clases o grupos, pero la manera de realizar este agrupamiento es difusa, lo que significa que los objetos o datos no son asignados

exclusivamente a una sola clase, como en el caso de *k*-medias, sino parcialmente a todas pero con distinto grado de pertenencia. El objetivo de este tipo de métodos es separar los datos en grupos cuyos miembros posean una gran similitud entre ellos, pero a su vez posean una gran disimilitud con el resto de los miembros de los restantes grupos. Sus bases teóricas son las siguientes:

Sea $\mathbf{X}_i \in \Re^p, i = 1 \cdots n$ un conjunto de datos de dimensión $p$ y $\mathbf{V}_j \in \Re^p, j = 1 \cdots c$ $(1 < c < n)$ un conjunto de centros de grupo o prototipos. Una $c$-partición de $\mathbf{X}$ puede ser representada por $U_{ji}$, que es una función continua en el intervalo [0,1] y representa la pertenencia de $\mathbf{X}_i$ al grupo $j$. En general, los elementos de $U_{ji}$ satisfacen las siguientes restricciones:

$$\left\{ \begin{array}{l} 0 \le U_{ji} \le 1 \\ \sum_{j=1}^{c} U_{ji} = 1 \,, \ \forall i \end{array} \right\} \tag{3.2}$$

El problema a resolver puede ser planteado matemáticamente de la siguiente forma:

$$\min_{\mathbf{U},\mathbf{V}} \sum_{i=1}^{n} \sum_{j=1}^{c} U_{ji}^{m} \left\| \mathbf{X}_i - \mathbf{V}_j \right\|^2 \tag{3.3}$$

Donde el parámetro $m$ es conocido como difusor y determina el grado de difusión para las clases encontradas. Este parámetro toma valores mayores que 1 y cuando es cercano a uno, el algoritmo calcula una solución con clases no difusas, donde la asignación de cada datos a los distintos grupos se realiza de manera exclusiva a uno solo de ellos. Mientras mayor sea $m$ más difusa será la solución, lo que implica que en el extremo todos los datos pertenecerán a todos los grupos con igual grado de pertenencia.

El siguiente algoritmo es capaz de encontrar una solución que converge a un mínimo local del funcional planteado por la ecuación (3.3):

1. Inicializar **V** de manera aleatoria. Inicializar **U** de manera aleatoria, pero satisfaciendo las restricciones dadas en la ecuación (3.2).
2. Fijar un valor para $m$>1.
3. Para $i$=1...$n$, y para $j$=1...$c$, calcular:

$$U_{ji} = \frac{1}{\sum_{k=1}^{c} \frac{\left\|\mathbf{X}_i - \mathbf{V}_j\right\|^{2/(m-1)}}{\left\|\mathbf{X}_i - \mathbf{V}_k\right\|^{2/(m-1)}}} \tag{3.4}$$

4. Para $j$=1…$c$, calcular:

$$\mathbf{V}_j = \frac{\sum_{i=1}^{n} U_{ji}^m \mathbf{X}_i}{\sum_{i=1}^{n} U_{ji}^m} \tag{3.5}$$

5. Parar cuando las diferencias de las $U_{ji}$ entre la iteración actual y la anterior sea más pequeña que un valor $\varepsilon$ dado; en caso contrario ir al paso 3.

Para mayor claridad, el algoritmo ha sido representado en un diagrama de flujo en la figura 3.2. Para demostrar que el algoritmo presentado resuelve el problema definido en la ecuación (3.3), primero se debe tomar la derivada de dicha ecuación con respecto a $\mathbf{V}_j$ y hacerla igual a cero. Esto produce exactamente la ecuación (3.5). El próximo paso es tomar la derivada de la ecuación (3.3) con respecto a **U**, bajo las restricciones planteadas en la ecuación (3.2). Esto produce exactamente la ecuación (3.4). Una demostración completa del proceso de obtención de este algoritmo puede consultarse en [52].

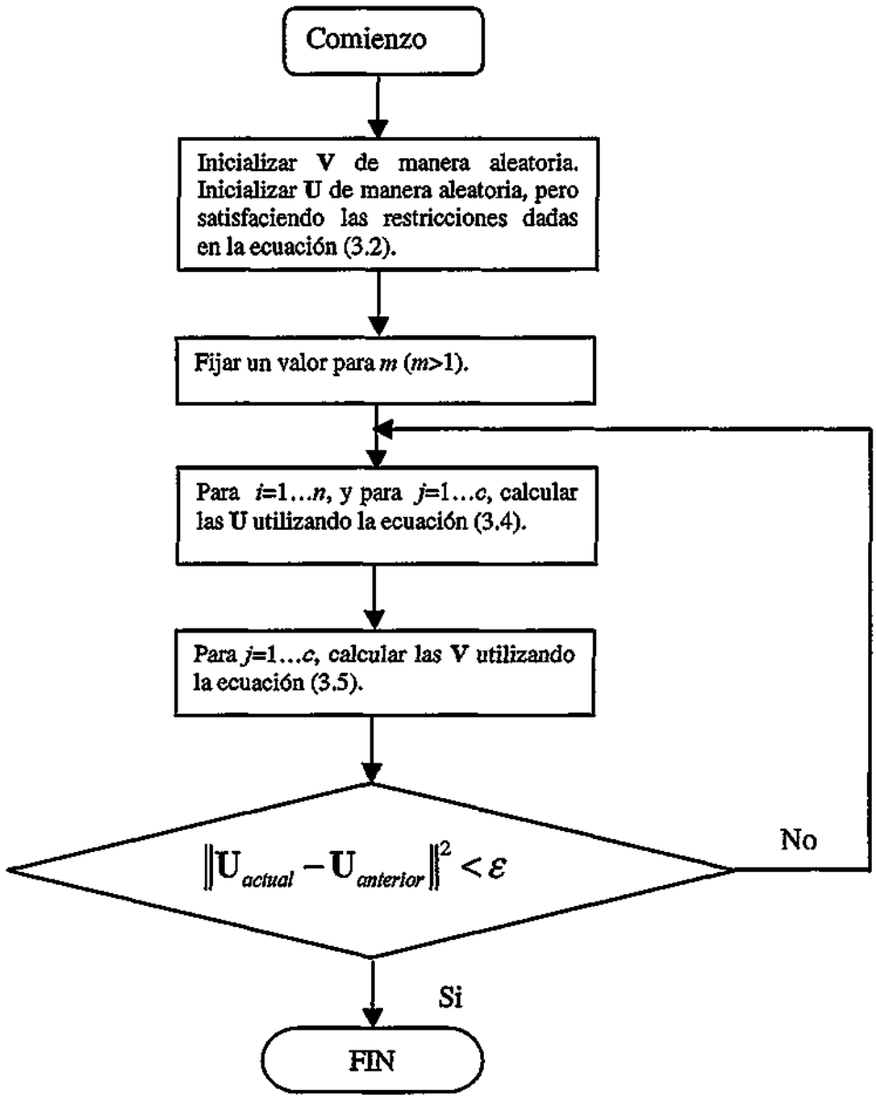



**Figura 3.2.** Diagrama de flujo del algoritmo de agrupamiento *c*-medias difuso (FCM).

## 3.3. Redes de agrupamiento difusas de Kohonen (*FKCN*)

Este método, conocido como redes de agrupamiento difusas de Kohonen (Fuzzy Kohonen Clustering Network, FKCN) [51], es considerado un tipo de red neuronal orientada a la agrupación de datos en conjuntos difusos que combina propiedades de dos de los métodos anteriormente descritos: SOM y FCM. La estructura de esta red es muy similar a la estructura de un SOM lineal y consiste en dos capas: una capa de entrada formada por $p$ nodos correspondientes a cada uno de las variables de los datos de entrada y una capa de salida compuesta por $c$ nodos interconectados entre sí de manera

lineal. Cada nodo tiene asignado, al igual que en SOM, un vector diccionario de la misma dimensión de los datos de entrada.

Cuando la red es estimulada con un dato de entrada, todos los nodos de la capa de salida actualizan sus vectores diccionarios utilizando la siguiente regla de aprendizaje:

$$\mathbf{V}_{i,t} = \mathbf{V}_{i,t-1} + \alpha_{ik,t}\left(\mathbf{X}_k - \mathbf{V}_{i,t-1}\right) \tag{3.6}$$

La novedad de este método es que combina el concepto de pertenencia difusa definido en el algoritmo de FCM en la regla de aprendizaje de SOM dada por la ecuación (3.6). Esta relación viene dada de la siguiente forma:

$$\alpha_{ik,t} = U_{ik,t}^{m_t} \quad (3.7), \quad m_t = \left(m_0 - t\Delta m\right) \quad (3.8) \text{ y } \quad \Delta m = \frac{\left(m_0 - 1\right)}{t_{\max}} \quad (3.9)$$

donde $U_{ik}$ representa los elementos de la matriz de pertenencia difusa y se calculan de la misma manera que en el algoritmo FCM a través de la ecuación (3.4), $m_0$ es una constante positiva mayor que 1 y representa el grado de difusión inicial de los grupos, mientras que $t$ es la iteración actual y $t_{\max}$ representa el número máximo de iteraciones. Este método, posee las siguientes propiedades interesantes:

- El factor de aprendizaje $\alpha$ está expresado en función del número de iteración $t$ y su efecto es distribuir la contribución de cada vector de entrada $\mathbf{X}_k$ a las neuronas de salida en una proporción inversa a sus distancias. Al igual que ocurre en SOM, el nodo ganador (cuyo vector diccionario sea más parecido al vector de entrada) actualiza su valor favorecido por el factor de aprendizaje a medida que el número de iteraciones aumenta. En este sentido el concepto de "vecindad" de Kohonen está implícitamente incluido en esta regla de aprendizaje, haciendo de FKCN un algoritmo auto-organizativo.
- FKCN no es secuencial: el orden de entrada de los datos al algoritmo no afecta su resultado.
- Para un valor de $m_t > 1$ fijo, el algoritmo de FKCN se convierte en el algoritmo de c-medias difuso.
- En el límite cuando $m_t = 1$, el algoritmo se convierte en el algoritmo de agrupamiento de k-medias.

Los pasos del algoritmo de FKCN son los siguientes:

1. Fijar el número de grupos $c$; fijar el criterio de parada $\varepsilon$ a un valor pequeño.
2. Inicializar los vectores diccionarios (centroides) de manera aleatoria. Fijar un valor para el parámetro de difusión inicial $m_0 > 1$. Fijar el número máximo de iteraciones $t_{\max}$
3. Para $t = 1,2,..,t_{\max}$
    a. Calcular todos los factores de aprendizaje $\alpha_{ik,t}$ definidos por la ecuación (3.7).
    b. Actualizar todos los vectores diccionarios $\mathbf{V}_{i,t}$ según la siguiente ecuación:

$$\mathbf{V}_{i,t} = \mathbf{V}_{i,t-1} + \frac{\sum_{k=1}^{n} \alpha_{ik,t} \left( \mathbf{X}_k - \mathbf{V}_{i,t-1} \right)}{\sum_{s=1}^{n} \alpha_{is,t}} \tag{3.10}$$

    c. Calcular el criterio de parada como:

$$E_t = \left\| \mathbf{V}_t - \mathbf{V}_{t-1} \right\|^2 = \sum_i \left\| \mathbf{V}_{i,t} - \mathbf{V}_{i,t-1} \right\|^2 \tag{3.11}$$

    d. Si $E_t \leq \varepsilon$ parar, en caso contrario seguir con la próxima iteración $t$.

# 4. Estimación de la función densidad de probabilidad

La función densidad de probabilidad (pdf) es un concepto estadístico fundamental que nos proporciona una descripción natural de la distribución de una variable aleatoria continua en un intervalo determinado. En otras palabras, es una función que puede ser integrada para obtener la probabilidad de que la variable aleatoria tome un cierto valor en un intervalo dado. Si consideramos la variable aleatoria $X$ con función densidad de probabilidad $f$ y conocemos la descripción exacta de esta función $f$, podemos obtener probabilidades asociadas con $X$ a partir de la siguiente relación:

$$P(a < X < b) = \int_a^b f(x)dx \quad \text{para todo } a < b \tag{4.1}$$

El uso de la función de densidad de probabilidad en el campo de análisis estadístico de datos y reconocimiento de patrones es muy amplio. Especialmente en el caso de clasificación supervisada donde los procesos de decisión de pertenencia de los datos a distintas clases son estudiados de manera probabilística [53, 54]. Adicionalmente, la función densidad de probabilidad es muy útil en el caso en que no se posea ninguna información a priori del conjunto de datos que se quiere analizar, permitiendo un análisis natural de sus propiedades. Así mismo, existe un conjunto elevado de aplicaciones en los que la densidad de probabilidad puede ser utilizada con vistas a entender mejor los datos con los que se trabaja, incluyendo aplicaciones de análisis discriminante [55], análisis de agrupamiento [56-58], simulación y muestreo [59], así como estimación cuantitativa de valores que dependen de la densidad [54], entre otras. Intuitivamente se podría decir que conociendo la función de densidad de probabilidad de la cual provienen los datos que se quieren estudiar, su análisis es relativamente sencillo.

Sin embargo, en la mayoría de los problemas reales de análisis y de exploración de datos, la función de densidad de probabilidad teórica de la cual provienen los mismos raramente es conocida. No obstante, si contamos con un conjunto suficiente de datos que asumimos son muestreados a partir de una función de densidad de probabilidad desconocida, la forma aproximada de esta función puede ser estimada a partir de estas propias observaciones.

Básicamente existen dos metodologías generales para la estimación de la función densidad de probabilidad: la paramétrica y la no paramétrica. La estimación

paramétrica de la pdf asume que los datos provienen de alguna distribución conocida, por ejemplo la distribución normal con media $\mu$ y varianza $\sigma^2$. La función densidad de probabilidad *f* que explica los datos, por lo tanto, se puede obtener a partir de los datos buscando estimaciones razonables de los parámetros $\mu$ y $\sigma^2$ y sustituyendo estos parámetros en la fórmula de la distribución normal.

Los métodos no paramétricos, por el contrario, son menos rígidos en el sentido de que no suponen prácticamente nada acerca de la distribución de los datos. En este caso, se asume que los datos provienen de una función densidad de probabilidad desconocida *f*, y son precisamente los datos quieres "hablarán" por sí mismo para lograr un buen estimador de *f*.

En el contexto de esta memoria destacaremos principalmente las técnicas de estimación no paramétrica de la función densidad de probabilidad por estar estrechamente ligados a los métodos que proponemos como objetivo de esta tesis. En particular, haremos énfasis en los métodos de estimación basados en funciones núcleo (kernel).

## 4.1. Estimadores núcleo de densidad

La estimación de densidad es el proceso de construcción de un estimado de la función de densidad de probabilidad a partir de datos observados. Entre este tipo de estimadores destacan los llamados estimadores núcleo de densidad, conocidos también como estimadores Parzen [60]. Como este tipo de métodos son ampliamente conocidos y existe una literatura abundante sobre ellos [53, 54, 60], solamente los introduciremos brevemente en este apartado.

Sea $\mathbf{X}_i \in \Re^{p*1}$, $i$=1...$n$, un conjunto de datos y $\mathbf{X} \in \Re^{p*1}$ una variable aleatoria. El estimador tipo núcleo de densidad de probabilidad queda definido como:

$$\hat{D}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} K(\mathbf{X} - \mathbf{X}_i; \alpha) \tag{4.2}$$

donde $K$ es una función núcleo y $\alpha > 0$ es el ancho de dicho núcleo, que controla la "suavidad" de la densidad estimada. Este parámetro $\alpha$ es también conocido como parámetro de suavidad ó ancho de banda (bandwidth).

Las características deseables de las funciones núcleo deberían ser las siguientes:

- $K(\mathbf{X}-\mathbf{X}_i;\alpha)$ debería alcanzar el máximo para $X = X_i$.
- $K(\mathbf{X}-\mathbf{X}_i;\alpha)$ debería ser cercano a cero para valores de X muy alejados de Xi
- $K(\mathbf{X}-\mathbf{X}_i;\alpha)$ debería ser una función suave y continua y decrecer monótonamente conforme aumenta la distancia $(X - X_i)$.
- Si $K(\mathbf{X}_1-\mathbf{X}_i;\alpha)=K(\mathbf{X}_2-\mathbf{X}_i;\alpha)$ entonces $X_1$ y $X_2$ deberían tener el mismo grado de similitud con $X_i$.

Un ejemplo típico de una función núcleo comúnmente utilizada es el núcleo Gaussiano:

$$K(\mathbf{Z};\alpha)=\frac{1}{(2\pi\alpha)^{p/2}}\exp\left(-\frac{\|\mathbf{Z}\|^2}{2\alpha}\right) \tag{4.3}$$

o equivalentemente:

$$K(\mathbf{X}-\mathbf{X}_i;\alpha)=\frac{1}{(2\pi\alpha)^{p/2}}\exp\left(-\frac{\|\mathbf{X}-\mathbf{X}_i\|^2}{2\alpha}\right) \tag{4.4}$$

Es importante señalar que el núcleo debe estar normalizado, es decir, que debe cumplir la siguiente condición:

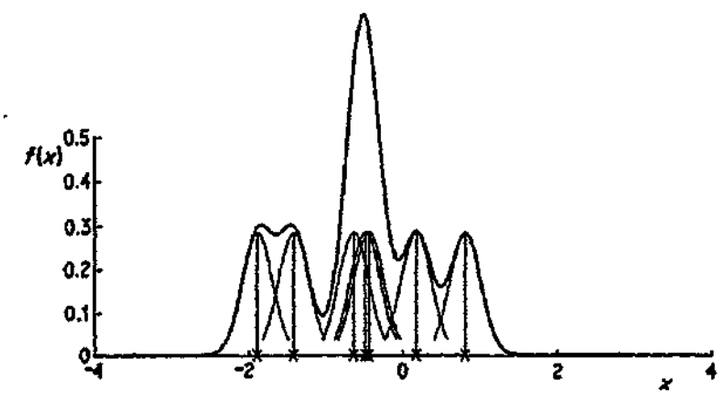
$$\int_{\forall \mathbf{Z}} K(\mathbf{Z};\alpha)\,d\mathbf{Z}=1 \tag{4.5}$$

Intuitivamente el estimador núcleo descrito anteriormente puede verse como una suma de "montículos" ubicados en cada una de las observaciones (datos), donde la función núcleo define la forma del montículo y el parámetro de suavidad $\alpha$ define su ancho. La figura 4.1 muestra una ilustración del proceso de estimación de la densidad utilizando un núcleo Gaussiano con distintos valores de suavidad. Si la suavidad (ancho del núcleo) es muy pequeña (figura 4.1a), la densidad estimada aparece con muchos picos, lo cual no es deseable en muchas aplicaciones al introducir importantes discontinuidades. Por el contrario, si la suavidad utilizada es muy elevada (figura 4.1c), la densidad aparece emborronada, oscureciendo cualquier nivel de detalles.

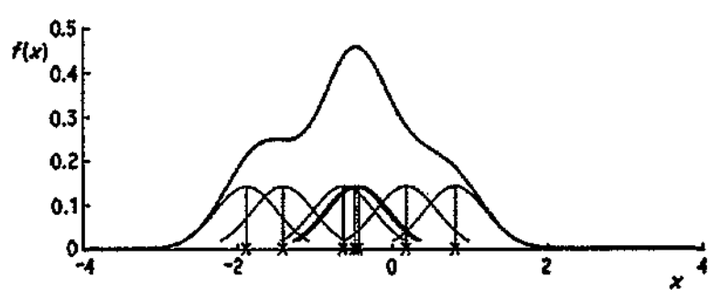
Es importante señalar que la determinación del parámetro de suavidad óptimo es un proceso crítico y muchas veces se fija de manera intuitiva y manual, aunque existen métodos más sofisticados para intentar estimar intervalos de valores razonables para este parámetro [54]. Una regla básica a la hora de escoger el ancho del núcleo podría ser que cuando las muestras estén muy dispersas se debería escoger un ancho de núcleo

elevado. Por el contrario si las muestras están muy agrupadas, el rango de núcleo debería ser menor.

a)




b)




c)

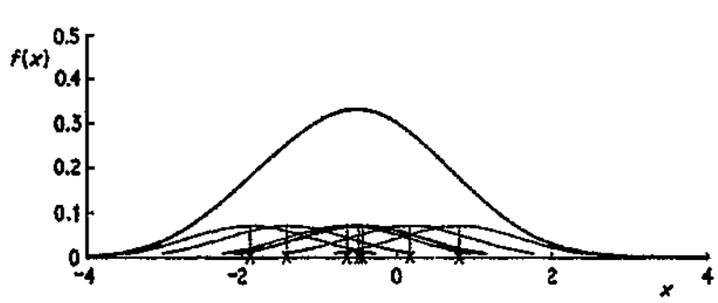



**Figura 4.1** Estimadores de densidad tipo núcleo. Las imágenes muestran la densidad estimada calculada a partir de la suma de los núcleos ubicados sobre cada dato y utilizando distintos valores del parámetro de suavidad: a) $\alpha = 0.2$ b) $\alpha = 0.4$ c) $\alpha = 0.6$

# CAPÍTULO II: NUEVOS ALGORITMOS

## 5. Mapas auto-organizativos basados en optimización funcional

En la sesión 2.3 describimos las propiedades matemáticas del algoritmo clásico de Kohonen. Así mismo se discutió que a pesar de que sus propiedades básicas de auto-organización y cuantificación vectorial son fácilmente reproducibles a través de simulaciones y, a pesar de su simplicidad conceptual y práctica, el algoritmo es sorprendentemente resistente a un estudio matemático completo. Solo el caso particular en que tanto los datos como el mapa se encuentran en dimensión uno ha sido bien caracterizado [34].

Debido a la dificultad de encontrar una sólido fundamento teórico al algoritmo de Kohonen, algunos investigadores han optado por desarrollar otros procedimientos diferentes de SOM pero basados en la misma idea de intentar una proyección de los datos de un espacio de alta dimensión a otro de menor dimensionalidad, conservando la estructura topológica de los mismos. La metodología utilizada para el desarrollo de estos nuevos algoritmos, al contrario de SOM, está basada en la optimización de funciones de costo bien definidas. La idea básica es formular una función de costo que tome su mínimo con respecto a los parámetros que van a ser determinados cuando se alcance el estado deseado del proceso de mapeo. De esta forma la minimización de la función de costo producirá automáticamente el conjunto óptimo de parámetros.

Esta aproximación permite una caracterización matemática completa del proceso de proyección y, por lo tanto, un mayor control sobre el algoritmo. Por ejemplo, Graepel y colaboradores [61], extendiendo un trabajo de Luttrell [62], propusieron distintas funciones de costo como son:

- *Algoritmo de cuantificación de vectores topográficamente suave (The Soft Topographic vector quantization algorithm, STVQ)*:

$$E^{TVQ}\left(\{C_{tr}\},\{m_r\}\right)=\frac{1}{2}\sum_t\sum_r C_{tr}\sum_s h_{rs}\left\|x(t)-m_s\right\|^2 \qquad (5.1)$$

Esta función de costo depende de los siguientes parámetros: $N$ vectores de datos $x(t)\in\Re^n$, $M$ vectores diccionarios $m_r\in\Re^n$, la función de vecindad (similar a la del algoritmo de SOM) $h_{rs}$ y la asignación binaria de variables

$c_{tr} \in \{0,1\}$ que toma el valor $c_{tr} = 1$ si el dato *x(t)* pertenece al nodo *r* y $c_{tr} = 0$ en caso contrario.

Intuitivamente se puede explicar el por qué la minimización de la función de costo dada por la ecuación (5.1) produce un mapa topográficamente correcto si observamos que esa función incurre en un coste para un vector de datos *x(t)* determinado si este es asignado a un nodo *r* (cuando $c_{tr} = 1$). Este coste es el cuadrado de la distancia euclídea entre el vector de datos y su correspondiente vector diccionario $m_s$ ponderada por la función de vecindad $h_{rs}$. Consecuentemente, el coste es mínimo no solo cuando los vectores diccionarios son lo más parecidos a los datos de entrada que representan, sino también cuando sus *s* vecinos en el mapa tienen también asignados vectores de entrada parecidos. Esto es exactamente lo que se pretende con un mapa topográfico, donde las relaciones espaciales de los datos en el espacio de entrada son representadas por las relaciones espaciales de los vectores diccionarios en el mapa.

- *Algoritmo de mapeo topográficamente suave basado en núcleos (The Kernel-based soft topographic mapping, STMK)*: Este nuevo algoritmo es una generalización del método anterior, pero introduce nuevas medidas de distancia basadas en funciones de tipo núcleo. La idea es establecer una función de mapeo del espacio de datos a un espacio de características $\phi : X \mapsto F$ de manera que la cuantificación vectorial no se realiza en el espacio original, sino en el espacio de características. Esta idea se ha venido utilizando con mucho éxito por métodos de clasificación supervisados como son las Máquinas de Vectores Soporte (SVM) [63]. La nueva función de coste quedaría descrita de la siguiente forma:

$$E^{TMK}\left(\{C_{tr}\},\{m_r^{\phi}\}\right) = \frac{1}{2}\sum_t \sum_r C_{tr} \sum_s h_{rs} \left\| \phi(x(t)) - m_s^{\phi} \right\|^2 \qquad (5.2)$$

Este funcional, al igual que STVQ, permite la creación de mapas topológicamente correctos, con la salvedad de que la cuantificación es ahora expresada no en el espacio original, sino en el espacio de características definido por la función de mapeo no lineal $\phi : X \mapsto F$, permitiendo que propiedades que

no pueden ser observadas en el espacio euclídeo original sean reveladas en el espacio de características anteriormente definido.

- *Mapeo topográfico suave para datos de proximidad (The Soft Topographic Mapping for Proximity Data, STMP)*: Este nuevo funcional es muy parecido a los anteriores solo que permite que los datos no estén definidos como vectores en el espacio euclídeo, sino como matrices de diferencia. Esto es especialmente útil en el caso de trabajo con grafos, diccionarios fonéticos, entre otros. El nuevo funcional queda expresado como:

$$E^{TMP}\left(\{C_{tr}\}\right)=\frac{1}{2}\sum_{t,t'}\sum_{r,s,u}\frac{c_{tr}h_{rs}c_{t'u}h_{us}}{\sum_{t''}\sum_{v}c_{t''v}h_{vs}}d_{tt'} \qquad (5.3)$$

Los elementos de la matriz de similitud vienen dados por $d_{tt'}$ e influyen en la función de costo solo cuando dos elementos de datos (en este caso los elementos de datos son los elementos de la matriz de diferencia entre pares de puntos) son asociados a los mismos nodos del mapa. La función de vecindad garantiza, al igual que en los algoritmos anteriores, que datos parecidos en el espacio de entrada sean asignados a nodo vecinos en el espacio de salida, garantizando de esta forma la preservación topológica.

Estos funcionales descritos anteriormente son optimizados con una combinación del algoritmo de E-M (Expectation-Maximization) y técnicas de enfriamiento rápido (Deterministic Annealing), conduciendo de manera natural a sendos algoritmos matemáticamente bien fundamentados.

Por otra parte Bishop y colaboradores [64] también propusieron el algoritmo "Mapeo Topográfico Generativo" (Generative Topographic Mapping, GTM), el cual es una reformulación de SOM que utiliza una función de costo probabilística, optimizada también mediante el algoritmo de E-M. Este método representa un modelo de densidad de probabilidad que describe la distribución de los datos en un espacio de altas dimensiones en términos de un número mucho menor de variables latentes. Utilizando un número de nodos distribuidos en una malla discreta finita en el espacio latente, este método, al igual que SOM, es capaz de establecer una relación no lineal entre el espacio de entrada y el espacio latente, pero manteniendo su formulación matemática tratable.

Aunque todos estos algoritmos son bastantes más complejos y costosos en tiempo de cómputo que el algoritmo de SOM, tienen la gran ventaja de ofrecer un mejor control y una mayor comprensión del proceso de proyección.

Uno de los objetivos principales de este trabajo de tesis es el planteamiento de dos nuevas funciones de coste que expresen, de manera similar a las anteriormente expuestas, las propiedades de los mapas auto-organizativos. La principal motivación de este trabajo ha sido el intentar combinar ideas que han venido utilizándose durante mucho tiempo en el campo del análisis estadístico de datos y en el campo de reconocimiento de patrones. Específicamente, el intentar combinar ideas de agrupamiento difuso, estimación de la función densidad de probabilidad y la exploración de datos con mapas auto-organizativos. Todos estos métodos por separado ofrecen ciertos beneficios y a su vez presentan un conjunto determinado de desventajas. El intentar combinar las mejores propiedades de todos estos ellos supone un reto y una alta motivación científica.

En los apartados siguientes describiremos un nuevo algoritmo basado en la extensión del clásico método de c-medias difuso, descrito en la sección 3.2 de esta memoria, al cual se le han agregado propiedades auto-organizativas. Seguidamente, expondremos una extensión de esta metodología a la creación de mapas auto-organizativos basados en la estimación no paramétrica de la función densidad de probabilidad.

## 5.1. Algoritmo de c-Medias difuso suavemente distribuido

Una de las cualidades más importantes del algoritmo de SOM y de la cual se han creado infinidad de aplicaciones es la de permitir el agrupamiento de datos [24]. Este agrupamiento generalmente no se realiza en el espacio original, sino en el espacio de la malla de salida. Esto es posible ya que el algoritmo de SOM, al intentar preservar la topología, realiza una proyección suave y ordenada de los datos originales en el espacio de salida, por lo tanto, datos de entrada parecidos quedarán asignados a neuronas vecinas durante la proyección. Así mismo, la densidad y parecido de las neuronas en el mapa, reflejarán aproximadamente la densidad de los datos de entrada que ellas representan, permitiendo "visualizar" la estructura de agrupamiento de los mismos.

En este contexto cabe mencionar los distintos intentos que han existido para tratar de combinar las ideas de agrupamiento y proyección. Por ejemplo, Lampinen y Oja [65] demostraron que el algoritmo de SOM está estrechamente relacionado al algoritmo de agrupamiento clásico de k-medias presentado en la sección 3.1. Adicionalmente Y. Cheng [66] demostró que una modificación del algoritmo original de SOM llamada "Batch Map" es también una generalización del bien conocido algoritmo de agrupamiento k-medias.

Por otra parte, la idea de combinar lógica difusa con los mapas auto-organizativos también ha sido objeto de estudio de algunos autores, por ejemplo Vuorimaa [67] propuso una modificación del algoritmo de SOM donde se reemplazan las neuronas por reglas difusas, permitiendo de esta forma un modelado eficiente de funciones continuas. Finalmente, tal y como se expuso en el apartado 3.3, Chen-Kuo Tsao y colaboradores [51] integraron algunos aspectos del clásico algoritmo de agrupamiento de c-medias difuso con el algoritmo de SOM, obteniendo un algoritmo de agrupamiento con ciertas propiedades de ambos métodos.

Como se ha comentado en apartados anteriores, los mapas auto-organizativos deben cumplir dos requisitos fundamentales durante el proceso de entrenamiento: la auto-organización y la convergencia de los valores de las neuronas a un estado donde cuantifique de manera fiel los datos en el espacio de entrada. Una manera de cuantificar fielmente el espacio de entrada es encontrar una partición de los datos en un número finito de grupos, cada uno con un representante o centro del grupo, de forma tal que dentro de un grupo la distancia de los datos a su representante sea lo más pequeña posible y la distancia entre centros o representantes de distintos grupos sea la mayor posible. Uno de los algoritmos más utilizados para este tipo de tareas es precisamente el algoritmo de FCM.

El objeto de esta sección es el planteamiento de una metodología completamente diferente para construir nuevos mapas auto-organizativos parecidos a SOM, a partir de funciones de costo bien planteadas matemáticamente y que expresen explícitamente las dos características fundamentales deseadas de un mapa auto-organizativo: cuantificación del espacio de entrada y proyección suave, ordenada y topológicamente correcta. El sistema que proponemos en este apartado consiste en una versión modificada del funcional del algoritmo de agrupamiento c-medias difuso comentado en

la sección anterior, donde los centros de grupos o vectores diccionarios se encuentran distribuidos en un espacio de baja dimensionalidad (por ejemplo, una malla regular), para lo cual se adiciona al funcional un término de penalización, con el objetivo de garantizar una distribución suave de los vectores diccionarios en ese espacio de baja dimensión. La motivación principal de utilizar esta funcional como base para la creación del nuevo mapa auto-organizativo está basada en que este es un método muy utilizado en el campo de reconocimiento de patrones con excelentes resultados y además está completamente caracterizado matemáticamente [50, 52].

## 5.2. Definición de suavidad

Un ingrediente necesario para conseguir un mapa auto-organizativo correcto sería agregar al funcional de FCM (ecuación (3.3)) un término de penalización que garantice la suavidad de la distribución espacial de los vectores diccionarios en la malla. Intuitivamente la "suavidad" es necesaria aquí para asegurar un mapa ordenado. En otras palabras, se le adiciona una relación de vecindad a los centros de grupos.

Asumamos que los centros de grupo o vectores diccionarios están distribuidos en una malla cuadrada regular como la mostrada en la figura 5.1:

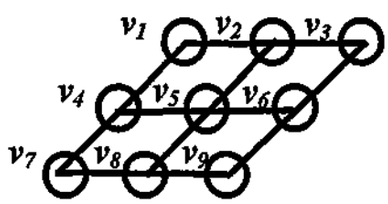



**Figura 5.1.** Malla bidimensional de 3x3 ($c = 9$) vectores diccionarios.

Cabe mencionar que otras topologías son también posibles, como por ejemplo una malla regular en 3D, una malla hexagonal, circular, etc. Una posible implementación de suavidad sería demandar que los valores de un vector diccionario sean parecidos al valor promedio de sus vecinos más cercanos en la malla. Refiriéndonos a la figura 5.1, esto significa que la siguiente medida de no-suavidad debe mantenerse pequeña:

$$tr\left(\mathbf{VCV}^T\right)=\begin{Bmatrix}\left\|\mathbf{V}_1-(\mathbf{V}_2+\mathbf{V}_4)/2\right\|^2+\left\|\mathbf{V}_2-(\mathbf{V}_1+\mathbf{V}_3+\mathbf{V}_5)/3\right\|^2+\left\|\mathbf{V}_3-(\mathbf{V}_2+\mathbf{V}_6)/2\right\|^2\\+\left\|\mathbf{V}_4-(\mathbf{V}_1+\mathbf{V}_5+\mathbf{V}_7)/3\right\|^2+\left\|\mathbf{V}_5-(\mathbf{V}_2+\mathbf{V}_4+\mathbf{V}_6+\mathbf{V}_8)/4\right\|^2+\left\|\mathbf{V}_6-(\mathbf{V}_3+\mathbf{V}_5+\mathbf{V}_9)/3\right\|^2\\+\left\|\mathbf{V}_7-(\mathbf{V}_4+\mathbf{V}_8)/2\right\|^2+\left\|\mathbf{V}_8-(\mathbf{V}_5+\mathbf{V}_7+\mathbf{V}_9)/3\right\|^2+\left\|\mathbf{V}_9-(\mathbf{V}_6+\mathbf{V}_8)/2\right\|^2\end{Bmatrix} \quad (5.4)$$

donde $\|\bullet\|^2$ denota la norma euclídea $L_2$ de un vector. La expresión en el lado izquierdo de la ecuación constituye una manera conveniente de expresar no-suavidad en general, a través del álgebra de matrices, donde $tr(\bullet)$ denota la traza de una matriz cuadrada y el índice superior "$^T$" denota la traspuesta de un vector o una matriz. En la ecuación (5.4) las columnas de la matriz $\mathbf{V}\in\mathbb{R}^{p*c}$ corresponden a los vectores diccionarios y la matriz $\mathbf{C}\in\Re^{c\times c}$ corresponde a un operador diferencial discreto. Esta medida ha venido siendo utilizada con éxito en la teoría de "splines" [68] y aquí haremos una extensión de su uso en el contexto de los mapas auto-organizativos.

Para explicar más detalladamente la medida de suavidad que vamos a utilizar, asumamos que los nodos serán distribuidos en una red regular como la mostrada en la figura 5.1. Para este caso y en términos generales, la "no suavidad" puede ser expresada a través de la siguiente colección de vectores: $\mathbf{W}=(\mathbf{W}_1\ \ \mathbf{W}_2\ \ \ldots\ \ \mathbf{W}_9)\in\Re^{p*9}$, con $\mathbf{W}_1=\mathbf{V}_1-(\mathbf{V}_2+\mathbf{V}_4)/2$, $\mathbf{W}_2=\mathbf{V}_2-(\mathbf{V}_1+\mathbf{V}_3+\mathbf{V}_5)/3$, y así sucesivamente. En notación matricial esto es equivalente a:

$$\mathbf{W}=\mathbf{VB} \quad (5.5)$$

donde, $\mathbf{V}=(\mathbf{V}_1\ \ \mathbf{V}_2\ \ \ldots\ \ \mathbf{V}_9)\in\Re^{p*9}$, $\mathbf{B}\in\Re^{9*9}$, y:

$$B_{ij}=\begin{cases}1\,,\ si\left|\mathbf{r}_i-\mathbf{r}_j\right|=0\\ -1\Big/\sum_{j=1}^{o}I\left(\left|\mathbf{r}_i-\mathbf{r}_j\right|=1\right)\end{cases} \quad (5.6)$$

En la ecuación (5.6), $\mathbf{r}_i$ denota el vector posición, en la malla, del $i^{\text{ésimo}}$ nodo y $I(\bullet)$ es la función indicador.

Finalmente, la medida escalar de "no suavidad" dada por la ecuación (5.4) es simplemente la norma de Frobenius de la matriz W definida por las ecuaciones (5.5) y (5.6):

$$\|\mathbf{W}\|_F=tr\left(\mathbf{WW}^T\right)=tr\left(\mathbf{VBB}^T\mathbf{V}^T\right)=tr\left(\mathbf{VCV}^T\right) \quad (5.7)$$

donde:

$$\mathbf{C} = \mathbf{B}\mathbf{B}^T \tag{5.8}$$

Tomando como ejemplo la malla de la figura 5.1 y la ecuación (5.4), la matriz **B** sería:

$$\mathbf{B} = \begin{bmatrix} 1 & -1/2 & 0 & -1/2 & 0 & 0 & 0 & 0 & 0 \\ -1/3 & 1 & -1/3 & 0 & -1/3 & 0 & 0 & 0 & 0 \\ 0 & -1/2 & 1 & 0 & 0 & -1/2 & 0 & 0 & 0 \\ -1/3 & 0 & 0 & 1 & -1/3 & 0 & -1/3 & 0 & 0 \\ 0 & -1/4 & 0 & -1/4 & 1 & -1/4 & 0 & -1/4 & 0 \\ 0 & 0 & -1/3 & 0 & -1/3 & 1 & 0 & 0 & -1/3 \\ 0 & 0 & 0 & -1/2 & 0 & 0 & 1 & -1/2 & 0 \\ 0 & 0 & 0 & 0 & -1/3 & 0 & -1/3 & 1 & -1/3 \\ 0 & 0 & 0 & 0 & 0 & -1/2 & 0 & -1/2 & 1 \end{bmatrix} \tag{5.9}$$

Es importante enfatizar que para este ejemplo particular definido por las ecuaciones (5.4) a la (5.9), la matriz **B** implementa un Laplaciano discreto con ciertas condiciones de frontera.

Otras variantes son también posibles, por ejemplo, la matriz **B** puede ser definida como un operador derivativo de primer orden (gradiente) y la matriz **C** sería entonces un operador tipo Laplaciano, similar al expresado en la ecuación (5.9). Por ejemplo, los elementos fuera de la diagonal de la matriz **C** se pueden calcular de la siguiente forma:

$$C_{ij} = \begin{cases} 0 \ , \ si \ \left|\mathbf{r}_i - \mathbf{r}_j\right| > 1 \\ -\dfrac{1}{4} \ , \ if \ \left|\mathbf{r}_i - \mathbf{r}_j\right| = 1 \end{cases} \tag{5.10}$$

seguido del cálculo de los elementos de la diagonal:

$$C_{ii} = -\sum_{\substack{j=1 \\ j \neq i}}^{c} C_{ij} \tag{5.11}$$

En la ecuación (5.10), $\mathbf{r}_i \in \mathbb{N}^2$ es el $i^{\text{ésimo}}$ vector de posición de los vectores diccionarios en la malla, expresado en coordenadas enteras. En referencia a la figura 5.1, la matriz **C** sería:

$$C=\frac{1}{4}B^{T}\mathbf{B}=\frac{1}{4}*$$

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2 | -1 | 0 | -1 | 0 | 0 | 0 | 0 | 0 |
| -1 | 3 | -1 | 0 | -1 | 0 | 0 | 0 | 0 |
| 0 | -1 | 2 | 0 | 0 | -1 | 0 | 0 | 0 |
| -1 | 0 | 0 | 3 | -1 | 0 | -1 | 0 | 0 |
| 0 | -1 | 0 | -1 | 4 | -1 | 0 | -1 | 0 |
| 0 | 0 | -1 | 0 | -1 | 3 | 0 | 0 | -1 |
| 0 | 0 | 0 | -1 | 0 | 0 | 2 | -1 | 0 |
| 0 | 0 | 0 | 0 | -1 | 0 | -1 | 3 | -1 |
| 0 | 0 | 0 | 0 | 0 | -1 | 0 | -1 | 2 |

(5.12)

En la ecuación (5.12) la matriz **B** es un operador gradiente discreto definido como:

$$\mathbf{B}=\begin{pmatrix}\mathbf{G}_x\\ \mathbf{G}_y\end{pmatrix}\in\mathbb{R}^{(2p)\cdot p} \tag{5.13}$$

donde $\mathbf{G}_x$ es el operador gradiente discreto a lo largo de la dirección horizontal (de izquierda a derecha) y $\mathbf{G}_y$ es el operador gradiente discreto a lo largo de la dirección vertical (de arriba a abajo):

$G_x=$

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -1 | +1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | -1 | +1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | -1 | +1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | -1 | +1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | -1 | +1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | +1 |

(5.14)

$G_y=$

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -1 | 0 | 0 | +1 | 0 | 0 | 0 | 0 | 0 |
| 0 | -1 | 0 | 0 | +1 | 0 | 0 | 0 | 0 |
| 0 | 0 | -1 | 0 | 0 | +1 | 0 | 0 | 0 |
| 0 | 0 | 0 | -1 | 0 | 0 | +1 | 0 | 0 |
| 0 | 0 | 0 | 0 | -1 | 0 | 0 | +1 | 0 |
| 0 | 0 | 0 | 0 | 0 | -1 | 0 | 0 | +1 |

(5.15)

Nótese que en las ecuaciones de la (5.10) a la (5.15) se utilizó la siguiente condición de frontera: si un vector diccionario está localizado en el borde de la rejilla y su vector diccionario predecesor está fuera de la rejilla, se asume que el valor "virtual" de este vector predecesor es igual al de su vecino en el borde. Esto es equivalente a no tener en cuenta los vectores fuera de la rejilla en el cálculo de la "no suavidad".

## 5.3. El nuevo funcional y su optimización

Haciendo uso de las dos ideas anteriormente expuestas: cuantificación vectorial del espacio de entrada dado por el funcional del algoritmo FCM y distribución suave y ordenada de los vectores diccionarios asociados a las neuronas en el espacio de salida, reflejadas mediante las ecuaciones (5.4) y (5.8), el problema de optimización modificado se puede expresar como una versión regularizada del algoritmo de c-medias difuso de Bezdek [52]:

$$\min_{\mathbf{U},\mathbf{V}} \left\{ \sum_{i=1}^{n}\sum_{j=1}^{c} U_{ji}^{m} \left\| \mathbf{X}_i - \mathbf{V}_j \right\|^2 + \vartheta tr\left(\mathbf{V}\mathbf{C}\mathbf{V}^T\right) \right\} \tag{5.16}$$

**Parte A (fidelidad a los datos)** **Parte B (ordenamiento topológico)**

Esta función está sujeta a las restricciones expresadas en la ecuación (3.2), $m > 1$ es el parámetro de difusión y $\vartheta > 0$ es el parámetro de regularización (también llamado parámetro de suavidad) que controla la magnitud de suavidad a demandar en el mapa.

Una vez planteada una función de costo que explícitamente refleja las características del nuevo mapa auto-organizativo (Parte A + Parte B del funcional de la ecuación (5.16)), el próximo paso es encontrar unos valores apropiados para **V** y **U** que la minimicen.

Para **V** y $\vartheta$ fijos, el problema de la ecuación (5.16) con respecto a **U** es equivalente al siguiente problema utilizando multiplicadores de Lagrange ($\lambda_i$):

$$\min_{\mathbf{U},\Lambda} \left\{ \sum_{i=1}^{n}\sum_{j=1}^{c} U_{ji}^{m} \left\| \mathbf{X}_i - \mathbf{V}_j \right\|^2 + \vartheta tr\left(\mathbf{V}\mathbf{C}\mathbf{V}^T\right) + \sum_{i=1}^{n}\left[ \lambda_i \left( \sum_{j=1}^{c} U_{ji} - 1 \right) \right] \right\} \tag{5.17}$$

Tomando la derivada parcial del funcional de la ecuación anterior con respecto a $U_{ji}$ y haciéndolo cero daría:

$$m U_{ji}^{m-1} \left\| \mathbf{X}_i - \mathbf{V}_j \right\|^2 + \lambda_i = 0 \tag{5.18}$$

Tomando la derivada parcial del funcional de la ecuación (5.17) con respecto a $\lambda_i$ y haciéndolo cero daría:

$$\sum_{j=1}^{c} U_{ji} = 1 \tag{5.19}$$

utilizando la ecuación (5.18) obtendríamos:

$$U_{ji} = \left[-\lambda_i\right]^{\frac{1}{m-1}} \left[\frac{1}{m\left\|\mathbf{X}_i - \mathbf{V}_j\right\|^2}\right]^{\frac{1}{m-1}} \tag{5.20}$$

e insertando la ecuación (5.20) en la ecuación (5.19) obtendríamos:

$$\left[-\lambda_i\right]^{\frac{1}{m-1}} = \frac{1}{\sum_{j=1}^{c}\left(m\left\|\mathbf{X}_i - \mathbf{V}_j\right\|^2\right)^{\frac{1}{1-m}}} \tag{5.21}$$

Finalmente, sustituyendo la ecuación (5.21) en la ecuación (5.20) obtenemos

$$U_{ji} = \frac{\left(\left\|\mathbf{X}_i - \mathbf{V}_j\right\|^2\right)^{\frac{1}{1-m}}}{\sum_{k=1}^{c}\left(\left\|\mathbf{X}_i - \mathbf{V}_k\right\|^2\right)^{\frac{1}{1-m}}} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\left\|\mathbf{X}_i - \mathbf{V}_j\right\|^2}{\left\|\mathbf{X}_i - \mathbf{V}_k\right\|^2}\right)^{\frac{1}{m-1}}} \tag{5.22}$$

Nótese que esta solución para el cálculo de los valores de pertenencia difusa de los datos a los vectores diccionarios es idéntica a la obtenida en el caso del algoritmo de FCM y expresada por la ecuación (3.4)

Por otra parte, para **U** y $\vartheta$ fijos, el problema de la ecuación (5.16) con respecto a $\mathbf{V}_j$, para j = 1..*c*, produce el siguiente sistema de ecuaciones lineales:

$$\mathbf{V}_j \sum_{i=1}^{n} U_{ji}^m + \vartheta \sum_{k=1}^{c} C_{jk}\mathbf{V}_k = \sum_{i=1}^{n} U_{ji}^m \mathbf{X}_i \tag{5.23}$$

ó equivalentemente:

$$\mathbf{V}_j = \frac{\sum_{i=1}^{n} U_{ji}^m \mathbf{X}_i - \vartheta \sum_{\substack{k=1 \\ k \neq j}}^{c} C_{jk}\mathbf{V}_k}{\sum_{i=1}^{n} U_{ji}^m + \vartheta C_{jj}} \tag{5.24}$$

Donde $C_{jk}$ denota los elementos de la matriz **C**. Nótese que si $\vartheta$=0 y para $2 \leq c < n$, entonces la ecuación (5.24) corresponde a la clásica solución de Bezdek del algoritmo FCM dada por la ecuación (3.5).

La ecuación (5.24) se obtiene de manera análoga a como se obtuvo la ecuación (5.22) para los valores de la matriz de pertenencia. Esto es, para U y $\vartheta$ fijos se toma la derivada parcial de la ecuación (5.16) con respecto a $\mathbf{V}_j$, para j = 1..*c*, y se hace cero.

A continuación, y a modo de ejemplo para clarificar el proceso de obtención de esta ecuación, incluiremos su demostración para lo cual se utilizarán algunas de las reglas de derivada de matrices incluidas en el anexo A de esta memoria.

El término derecho de la ecuación (5.16) puede ser rescrito de la siguiente forma:

$$tr\left(\mathbf{V}\mathbf{C}\mathbf{V}^T\right)=\sum_{k=1}^{c}\sum_{l=1}^{c}C_{kl}\mathbf{V}_k^T\mathbf{V}_l \tag{5.25}$$

donde $C_{kl}$ son los elementos de la matriz **C**, y $\mathbf{V}_k$, $\mathbf{V}_l$ son los vectores diccionarios. Esto es cierto, debido a las siguientes igualdades:

$$\left[\mathbf{V}\mathbf{C}\mathbf{V}^T\right]_{ij}=\sum_k\sum_l V_{ik}C_{kl}V_{jl} \tag{5.26}$$

Denotemos "*ij*" los elementos de la matriz $\left[\mathbf{V}\mathbf{C}\mathbf{V}^T\right]$, entonces:

$$tr\left(\mathbf{V}\mathbf{C}\mathbf{V}^T\right)=\sum_i\left[\mathbf{V}\mathbf{C}\mathbf{V}^T\right]_{ii}=\sum_i\sum_k\sum_l V_{ik}C_{kl}V_{il} \tag{5.27}$$

y

$$\sum_i\sum_k\sum_l V_{ik}C_{kl}V_{il}=\sum_k\sum_l C_{kl}\sum_i V_{ik}V_{il}=\sum_k\sum_l C_{kl}\mathbf{V}_k^T\mathbf{V}_l \tag{5.28}$$

Por lo tanto la ecuación (5.16) puede ser rescrita de la siguiente forma:

$$\min_{\mathbf{U},\mathbf{V}}\left\{\sum_{i=1}^{n}\sum_{j=1}^{c}U_{ji}^m\left(\mathbf{X}_i-\mathbf{V}_j\right)^T\left(\mathbf{X}_i-\mathbf{V}_j\right)+\vartheta\sum_{k=1}^{c}\sum_{l=1}^{c}C_{kl}\mathbf{V}_k^T\mathbf{V}_l\right\} \tag{5.29}$$

Tomando la derivada parcial del funcional dado por la ecuación (5.29) con respecto a $\mathbf{V}_j$ y haciéndolo cero, quedaría:

$$-2\sum_{i=1}^{n}U_{ji}^m\left(\mathbf{X}_i-\mathbf{V}_j\right)+2\vartheta\sum_{k=1}^{c}C_{jk}\mathbf{V}_k=\mathbf{0} \tag{5.30}$$

donde $\mathbf{0}\in\mathbb{R}^{p\times 1}$ representa un vector de ceros.

La ecuación (5.30) se obtiene utilizando las reglas de derivada de matrices mencionadas en el anexo A. Por ejemplo, de la ecuación (5.29) se obtiene:

$$\left.\begin{aligned} &\boldsymbol{G} = \left\|\mathbf{X}_i - \mathbf{V}_j\right\|^2 = \left(\mathbf{X}_i - \mathbf{V}_j\right)^T \left(\mathbf{X}_i - \mathbf{V}_j\right) = \\ &= \left(\mathbf{X}_i^T - \mathbf{V}_j^T\right)\left(\mathbf{X}_i - \mathbf{V}_j\right) = \mathbf{X}_i^T\mathbf{X}_i - \mathbf{X}_i^T\mathbf{V}_j - \mathbf{V}_j^T\mathbf{X}_i + \mathbf{V}_j^T\mathbf{V}_j \end{aligned}\right\} \tag{5.31}$$

Tomando el diferencial con respecto a $\mathbf{V}_j$ (ecuación (A.1) del anexo A):

$$d\boldsymbol{G} = -\mathbf{X}_i^T d\mathbf{V}_j - d\mathbf{V}_j^T\mathbf{X}_i + d\mathbf{V}_j^T\mathbf{V}_j + \mathbf{V}_j^T d\mathbf{V}_j \tag{5.32}$$

ó equivalentemente:

$$d\boldsymbol{G} = tr\left[\left(\mathbf{V}_j^T - \mathbf{X}_i^T\right)d\mathbf{V}_j 1 + 1 d\mathbf{V}_j^T\left(\mathbf{V}_j - \mathbf{X}_i\right)\right] \tag{5.33}$$

Donde 1 es la matriz identidad. Nótese también que la traza de un escalar es el propio escalar. Por lo tanto, utilizando las ecuaciones (A.5) y (A.6) del anexo A, la derivada parcial queda:

$$\frac{\partial \boldsymbol{G}}{\partial \mathbf{V}_j} = 2\left(\mathbf{V}_j - \mathbf{X}_i\right) = -2\left(\mathbf{X}_i - \mathbf{V}_j\right) \tag{5.34}$$

Adicionalmente,

$$\left.\begin{aligned} &\boldsymbol{S} = tr\left(\mathbf{V}\mathbf{C}\mathbf{V}^T\right) = \sum_{k=1}^{c}\sum_{l=1}^{c} C_{kl}\mathbf{V}_k^T\mathbf{V}_l = \sum_{l=1}^{c} C_{jl}\mathbf{V}_j^T\mathbf{V}_l + \sum_{\substack{k=1\\k\neq j}}^{c}\sum_{l=1}^{c} C_{kl}\mathbf{V}_k^T\mathbf{V}_l = \\ &= \sum_{l=1}^{c} C_{jl}\mathbf{V}_j^T\mathbf{V}_l + \sum_{l=1}^{c}\sum_{\substack{k=1\\k\neq j}}^{c} C_{kl}\mathbf{V}_k^T\mathbf{V}_l = \sum_{l=1}^{c} C_{jl}\mathbf{V}_j^T\mathbf{V}_l + \sum_{\substack{k=1\\k\neq j}}^{c} C_{kj}\mathbf{V}_k^T\mathbf{V}_j + \sum_{\substack{l=1\\l\neq j}}^{c}\sum_{\substack{k=1\\k\neq j}}^{c} C_{kl}\mathbf{V}_k^T\mathbf{V}_l = \\ &= C_{jj}\mathbf{V}_j^T\mathbf{V}_j + \sum_{\substack{l=1\\l\neq j}}^{c} C_{jl}\mathbf{V}_j^T\mathbf{V}_l + \sum_{\substack{k=1\\k\neq j}}^{c} C_{kj}\mathbf{V}_k^T\mathbf{V}_j + \sum_{\substack{l=1\\l\neq j}}^{c}\sum_{\substack{k=1\\k\neq j}}^{c} C_{kl}\mathbf{V}_k^T\mathbf{V}_l \end{aligned}\right\} \tag{5.35}$$

Tomando el diferencial con respecto a $\mathbf{V}_j$ (utilizando la ecuación (A.1) del anexo A), quedaría:

$$d\boldsymbol{S} = tr\left[C_{jj}d\mathbf{V}_j^T\mathbf{V}_j + C_{jj}\mathbf{V}_j^T d\mathbf{V}_j + \sum_{\substack{l=1\\l\neq j}}^{c} C_{jl}d\mathbf{V}_j^T\mathbf{V}_l + \sum_{\substack{k=1\\k\neq j}}^{c} C_{kj}\mathbf{V}_k^T d\mathbf{V}_j\right] \tag{5.36}$$

y finalmente la derivada parcial quedaría:

$$\frac{\partial \boldsymbol{S}}{\partial \mathbf{V}_j} = 2C_{jj}\mathbf{V}_j + 2\sum_{\substack{l=1\\l\neq j}}^{c} C_{jl}\mathbf{V}_l = 2\sum_{k=1}^{c} C_{jk}\mathbf{V}_k \tag{5.37}$$

Teniendo en cuenta que la matriz $\mathbf{C}$ es simétrica, es decir, que $C_{ij} = C_{ji}$, entonces la ecuación (5.30) queda exactamente igual a la ecuación (5.23), que es lo que se pretendía demostrar.

## 5.4. Algoritmo SOM difuso (FuzzySOM)

El algoritmo básico que se deriva del funcional planteado en el apartado anterior es muy parecido al algoritmo de *c*-medias difuso mostrado en la sección 3.2 de esta memoria. La solución será iterativa alternando entre la ecuación (5.22) y la ecuación (5.24). Nótese que la ecuación (5.24) puede ser rescrita de la siguiente forma:

$$\mathbf{V}_j = \frac{\sum_{i=1}^{n} U_{ji}^m \mathbf{X}_i - \vartheta \sum_{\substack{k=1 \\ k \neq j}}^{c} C_{jk} \mathbf{V}_k}{\sum_{i=1}^{n} U_{ji}^m + \vartheta C_{jj}} \tag{5.38}$$

para $j$=1...$c$. De esta forma la ecuación (5.38) queda en la forma conveniente para el algoritmo iterativo de Gauss-Seidel y donde $C_{jk}$ denota los elementos de la matriz **C** explicada anteriormente.

Una simple opción para la matriz **C** es el operador tipo Laplaciano (ecuación (5.12)). En este caso la ecuación (5.38) se simplifica de la siguiente manera:

$$\mathbf{V}_j = \frac{\sum_{i=1}^{n} U_{ji}^m \mathbf{X}_i + \vartheta \bar{\mathbf{V}}_j}{\sum_{i=1}^{n} U_{ji}^m + \vartheta} \tag{5.39}$$

Donde $\bar{\mathbf{V}}_j$ denota el promedio de los vectores diccionarios que son vecinos inmediatos de $\mathbf{V}_j$ en la malla. En este valor promedio $\mathbf{V}_j$ queda excluido. Por ejemplo, refiriéndonos a solo algunas neuronas del mapa de la figura 5.1, quedaría:

$$\left\{ \begin{array}{l} \bar{\mathbf{V}}_1 = (\mathbf{V}_2 + \mathbf{V}_4)/2 \\ \bar{\mathbf{V}}_2 = (\mathbf{V}_1 + \mathbf{V}_3 + \mathbf{V}_5)/3 \\ \bar{\mathbf{V}}_5 = (\mathbf{V}_2 + \mathbf{V}_4 + \mathbf{V}_6 + \mathbf{V}_8)/4 \end{array} \right\} \tag{5.40}$$

La actualización de los vectores diccionarios utilizando la ecuación (5.39) revela la naturaleza del proceso auto-organizativo de este método: un vector diccionario está directamente influenciado tanto por los datos de entrada mas parecidos como por sus vecinos mas cercanos en la malla.

La figura 5.2 muestra el diagrama de flujo de este algoritmo, al cual hemos llamado FuzzySOM (Mapa Auto-organizativo Difuso) y consiste en:

a. Inicializar **V** de manera aleatoria, e inicializar **U** de manera también aleatoria, pero satisfaciendo las restricciones dada por la ecuación (3.2).
b. Fijar un valor para $m$, siendo $m>1$, y un valor para $\vartheta > 0$. Fijar también umbral de parada $\varepsilon$.
c. Calcular las **U**, para i=1....$n$ y para j=1....$c$, utilizando la ecuación (5.22):

$$U_{ji} = \frac{1}{\sum_{k=1}^{c} \frac{\|\mathbf{X}_i - \mathbf{V}_j\|^{2/(m-1)}}{\|\mathbf{X}_i - \mathbf{V}_k\|^{2/(m-1)}}}$$

d. Para $j$=1....$c$, calcular las **V** utilizando la ecuación (5.39):

$$\mathbf{V}_j = \frac{\sum_{i=1}^{n} U_{ji}^m \mathbf{X}_i + \vartheta \bar{\mathbf{V}}_j}{\sum_{i=1}^{n} U_{ji}^m + \vartheta}$$

e. En caso de que no se cumpla la condición $\|\mathbf{V}_{actual} - \mathbf{V}_{anterior}\|^2 < \varepsilon$, se vuelve al paso d), mientras que si se cumple, pero no se cumple la condición $\|\mathbf{U}_{actual} - \mathbf{U}_{anterior}\|^2 < \varepsilon$, entonces se vuelve a repetir el proceso a partir del paso c),
f. Cuando se cumplan las dos condiciones anteriores, el algoritmo finaliza.

Este algoritmo es esencialmente una versión regularizada del algoritmo de FCM, cuya convergencia ha sido demostrada exhaustivamente en [52], por lo tanto, la convergencia de FuzzySOM está garantizada por analogía con el mismo. Así mismo, es bien conocido que algoritmos como el mostrado anteriormente encuentran una solución que converge al menos a un mínimo local de la función de coste. Con el objetivo de ayudar a lograr una convergencia hacia el mínimo global de la función descrita en la ecuación (5.16), y para minimizar el efecto que producen diferentes inicializaciones de **V** y **U**, se puede introducir en el algoritmo una estrategia conocida como enfriamiento determinista (deterministic annealing) aplicado a la variable de difusión [61, 69]. Básicamente la idea sería comenzar el algoritmo con valores altos del parámetro de difusión $m$ (alta temperatura) y hacerlo decrecer gradualmente ("enfriarlo") hasta valores de baja difusión bien cercanos a 1. De esta forma los resultados pueden mejorar sensiblemente. Por lo tanto, en los ejemplos que mostraremos en el apartado siguiente

hemos utilizado una versión modificada del algoritmo anteriormente descrito. En una primera parte, y para cualquier valor inicial de las **V**, los pasos (c) y (d) son repetidos un gran número de veces con una variación lineal de $m$, por ejemplo desde $m = 3$ hasta $m = 1.02$ en 500 pasos. Este sería el paso de enfriamiento determinista. En una segunda fase utilizando los valores actuales de **V** y **U** y con $m = 1.02$ fija, repetimos los pasos (c), (d), (e) y (f) hasta lograr la convergencia.

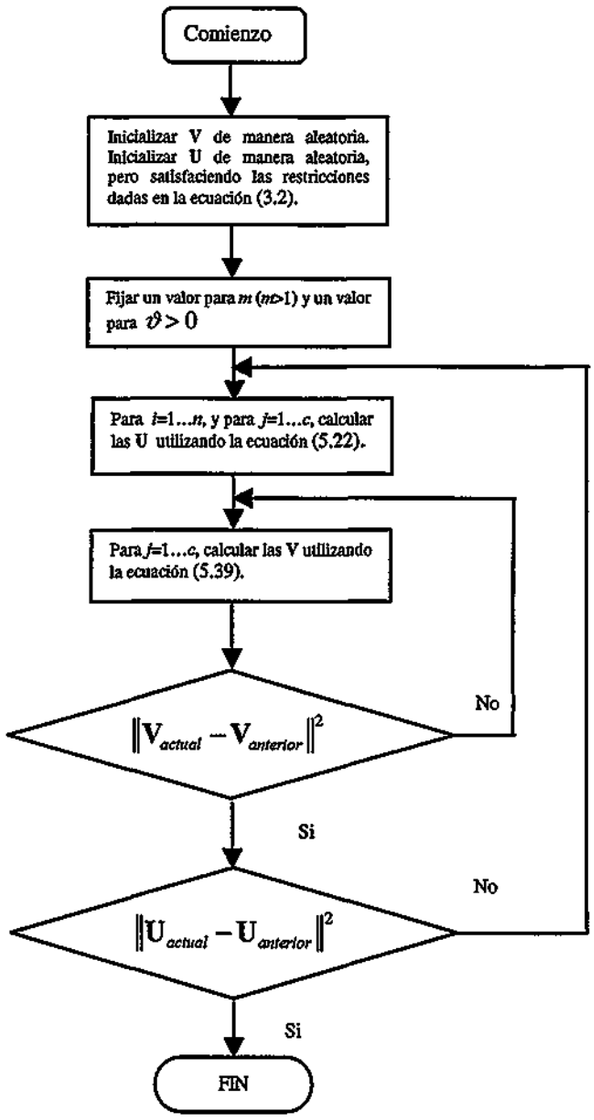



**Figura 5.2.** Diagrama de flujo del algoritmo FuzzySOM.

## 5.5. Ejemplos

En este apartado intentaremos demostrar las propiedades del método anteriormente descrito mediante ejemplos con datos sintéticos. La figura 5.3 muestra un ejemplo interesante de proyección de un conjunto de 855 puntos provenientes de un triángulo en 2D (figura 5.3a) sobre una red en 1D formada por 64 neuronas. Como se observa en la figura 5.3b, los vectores diccionarios tienden a llenar el triángulo de manera ordenada formándose las famosas curvas de "Peano" [22]. En este caso los parámetros utilizados fueron: $\vartheta$=0.5, con $m$ decreciendo desde 3 hasta 1.02 en 500 pasos.

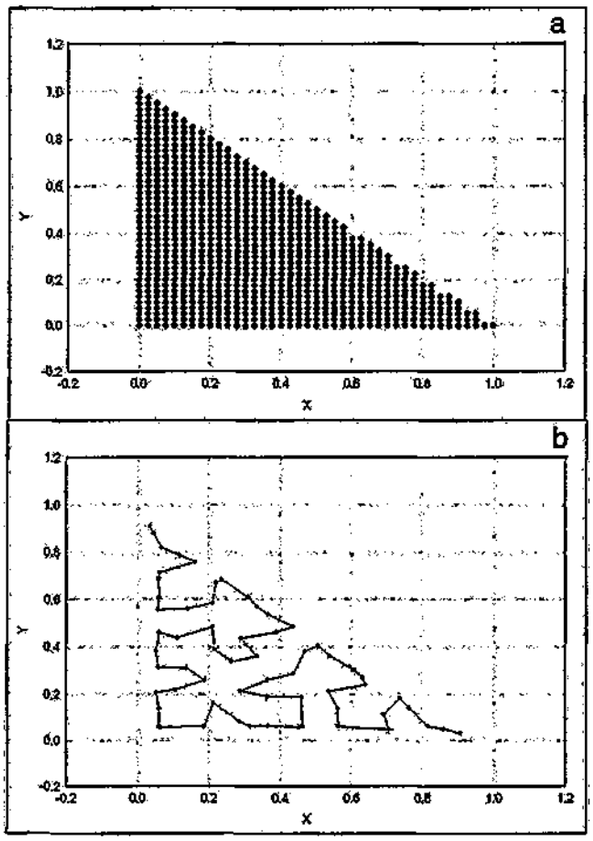



**Figura 5.3** Ejemplo de FuzzySOM con una red lineal. a) Conjunto de 855 puntos en 2D, muestreados a partir de un triángulo. b) Los puntos proyectados en un espacio 1D utilizando un red lineal de 64 neuronas.

El segundo ejemplo, mostrado en la figura 5.4, ilustra el efecto de diferentes valores del parámetro de suavidad $\vartheta$ sobre el nuevo mapa auto-organizativo. En este

caso simple los datos de entrada son un conjunto de 111 puntos en 2D provenientes de una distribución de 3 grupos circulares, como se muestra en la figura 5.4a. El mapa utilizado fue una malla cuadrada de 10x10. En la figura 5.4b ($\vartheta$=0.05), el mapa no está muy organizado. La organización aumenta cuando se aumenta el parámetro de regularización hasta el punto donde comienza a ocurrir una distorsión por excesiva regularización (Fig. 5.4d, $\vartheta$=9). En todos los casos $m$ se hizo decrecer de manera lineal en 400 pasos desde 2 hasta 1.02.

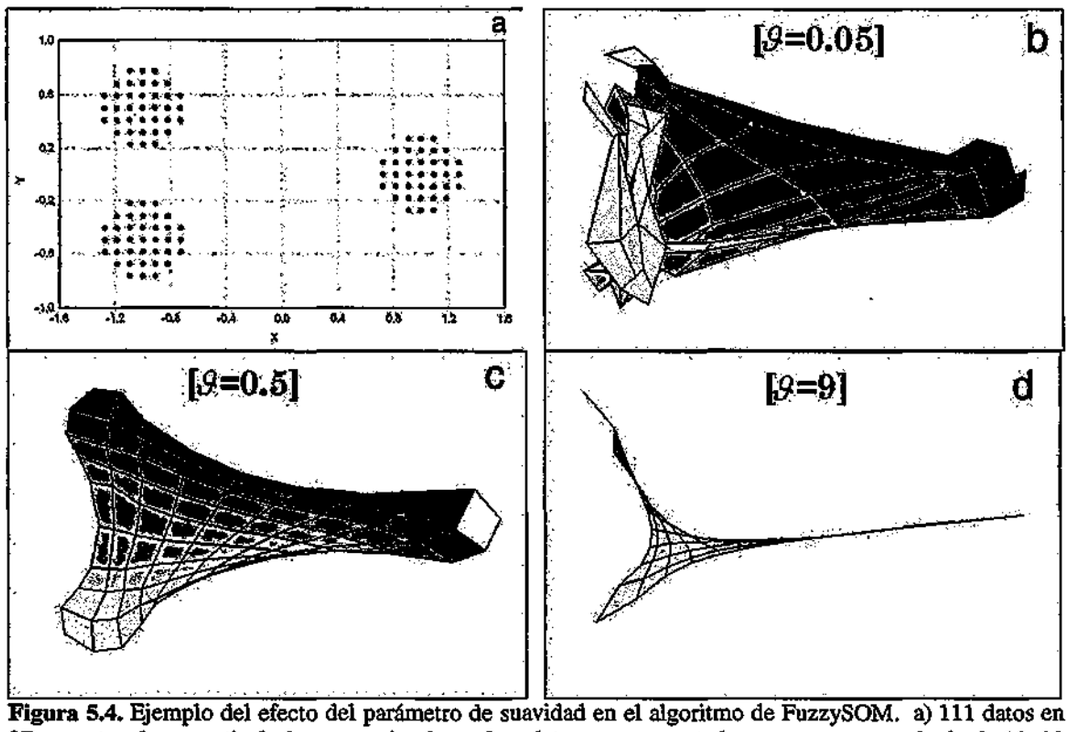



**Figura 5.4.** Ejemplo del efecto del parámetro de suavidad en el algoritmo de FuzzySOM. a) 111 datos en 2D muestreados a partir de 3 grupos circulares. Los datos son proyectados en un mapa cuadrado de 10x10 generado por FuzzySOM. Se muestran distintas proyecciones para diferentes valores del parámetro de suavidad $\vartheta$ en b), c) y d).

La figura 5.5 muestra un nuevo ejemplo correspondiente a la proyección de 93 puntos en 3D muestreados a partir de 3 segmentos ortogonales (figura 5.5a) sobre una malla cuadrada de 15x15. En la figura 5.5b se muestran los vectores diccionarios formando una representación suave de los datos originales. Los parámetros utilizados fueron: $\vartheta$=0.5, con $m$ variando linealmente desde 2 hasta 1.02 en 500 pasos. Este ejemplo ilustra la capacidad de proyección del nuevo método, que a pesar del cierto grado de suavidad presente, es capaz de conservar las características topológicas principales de los datos.

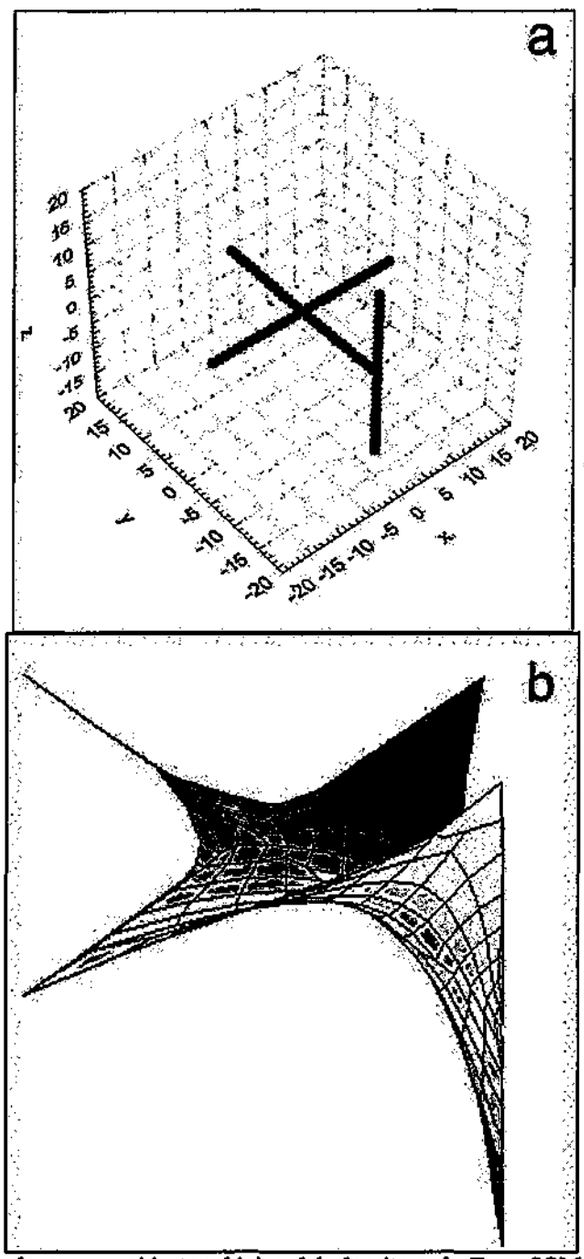



**Figura 5.5** Ejemplo de preservación topológica del algoritmo de FuzzySOM. a) 93 puntos en 3D obtenidos como muestras a partir de 3 segmentos ortogonales. b) Mapa de 15x15 donde los datos fueron proyectados con el algoritmo de FuzzySOM.

La figura 5.6 muestra la proyección de los datos clásicos de "Iris" [70], compuesto por 150 datos en 4D correspondientes a tres especies distintas de flores. Estos datos han sido muy utilizados durante mucho tiempo como conjunto de prueba para métodos de agrupamiento y clasificación. Los resultados de la proyección sobre

una red de 10x15 ($\vartheta$=0.5, con $m$ decreciendo desde 2 hasta 1.02 en 500 pasos) muestran la clara separación de una de las especies (marcada como 1) , mientras que las otras dos (2 y 3) no son claramente separables. Estos resultados están en perfecta concordancia con los obtenidos por la mayoría de métodos de reducción de dimensionalidad, proyección y agrupamiento aplicados a este conjunto de datos.

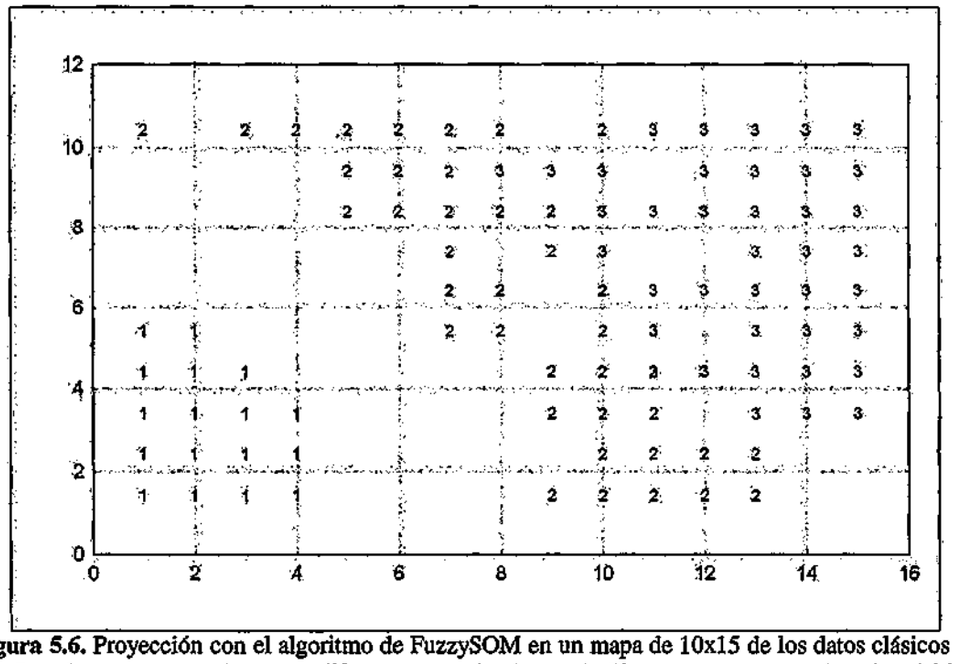



**Figura 5.6.** Proyección con el algoritmo de FuzzySOM en un mapa de 10x15 de los datos clásicos de Iris. Los datos corresponden a tres diferentes especies (grupos), 50 muestras por especie y 4 variables. Los números mostrados como puntos en la figura representan la clase a la que corresponden los datos.

## 5.6. Discusión

El tipo de método presentado en esta sección está basado en una nueva función de costo que expresa, de una manera directa, el diseño de un mapa ordenado que intenta conservar la estructura topológica de los datos. Teóricamente esto se logra a través de dos condiciones fundamentales expresadas por el funcional de la ecuación (5.16). El primer término de este funcional demanda que los vectores diccionarios sean fieles a los datos, es decir, representen de la mejor manera posible los datos originales. Adicionalmente, el segundo término de esta ecuación demanda a su vez que los valores de los vectores diccionarios cambien lo menos posible a través de la malla de salida, es decir, que cambien suavemente. A efectos prácticos esto produce un ordenamiento de los vectores diccionarios en el espacio de salida.

Estas dos propiedades descritas anteriormente: fidelidad a los datos y ordenamiento topológico, son precisamente las dos características que definen a los mapas auto-organizativos. Como se ha mostrado en los ejemplos con datos sintéticos descritos en el apartado anterior, el algoritmo ciertamente produce mapas ordenados que intentan no solamente representar la estructura de los datos, sino también su topología en el espacio original.

Este nuevo método posee ciertas ventajas, tanto teóricas como prácticas, sobre el clásico algoritmo de Kohonen. En primer lugar su planteamiento matemático, a diferencia de SOM, es preciso y claro, ofreciendo un mejor control y un mejor entendimiento del proceso de mapeo. De cierta forma este método constituye un nuevo intento en encontrar una explicación matemática que ayude a entender este algoritmo clásico de SOM. Adicionalmente, el algoritmo de FuzzySOM no solo produce mapas topológicamente correctos, sino que también posee una naturaleza difusa donde la asignación de cada dato original a cada uno de los vectores diccionarios es estimada, por el propio algoritmo, en cada iteración.

Las ventajas de este acercamiento difuso son indiscutibles. Por una parte, la asignación final de los datos al mapa no se realiza a posteriori por mínima distancia como se hace en el caso del SOM clásico, sino que la propia matriz de pertenencia resultante contiene ya esta información. Adicionalmente, la pertenencia difusa es utilizada como factor de ponderación de los datos en cada iteración, de forma tal que datos asignados de igual manera a dos vectores diccionarios distintos, influirán sobre los nuevos valores de estos vectores diccionarios en la misma proporción en que han sido asignados. Esta interesante propiedad hace que este tipo de mapas pueda ser capaz de dilucidar de mejor manera zonas de fronteras entre dos grupos vecinos que tengan cierto grado de solapamiento.

# 6. Método de agrupamiento y cuantificación de vectores basado en la estimación de la densidad de probabilidad.

Uno de los objetivos tradicionales en el campo de la compresión de datos y de codificación es la reducción del tamaño de los mismos de forma tal que se minimice sus requerimientos de almacenamiento a la vez que se preserven de manera fiel sus mismas cualidades para su posterior recuperación. Estos tipos de métodos de compresión son usualmente formalizados a través de la minimización de la distorsión media entre la entrada y la salida, medida por el error cuadrático medio o alguna otra medida similar. Estos métodos son también conocidos como métodos de cuantificación vectorial.

Los métodos de cuantificación vectorial están basados fundamentalmente en la segmentación del espacio vectorial original en un conjunto de grupos diferentes, cada uno de los cuales será representado por un solo vector comúnmente llamado vector diccionario y que tiende a explicar lo mejor posible aquellos datos a los que representa. Estas técnicas se han utilizado principalmente en comprensión de datos y codificación [71] y conceptualmente están muy estrechamente ligados a los métodos de agrupamiento. De hecho, métodos como el k-medias presentado en el apartado 3.1 se han venido utilizando intensamente como técnicas de cuantificación vectorial [69, 71].

Otro criterio igualmente válido para lograr un buena cuantificación vectorial es el de encontrar un conjunto de representantes que preserven de manera fiel la densidad de probabilidad de los datos de entrada [72]. Esta combinación de las ideas de cuantificación vectorial y estimación de densidad ha venido siendo utilizada en el caso de fuentes de datos discretas y están motivadas en el hecho de que un codificador vectorial que obtenga vectores representantes del espacio de entrada a través de la minimización del error de distorsión está, de manera implícita, estimando la densidad de clases de este conjunto de entrada [72].

La motivación principal del desarrollo de este tipo de metodología que combine las ideas de cuantificación vectorial y estimación estadística de la densidad de probabilidad viene dada por el hecho de que la gran mayoría de problemas, tanto en ciencia como en ingeniería, tienen que modelarse irremediablemente de una manera probabilística. Incluso en problemas con una naturaleza inherentemente determinista es frecuente encontrar una formulación probabilística a los mismos como única solución

abordable desde el punto de vista computacional. Esto ha implicado el gran desarrollo de teorías y métodos para lograr modelos estadísticos cada vez más realistas que permitan explicar de la manera más exacta posible los datos con los que se trabaja. De una forma natural, esta metodología requiere necesariamente tratar con la función de densidad de probabilidad cuando se conozca, y en su defecto con estimaciones de la misma a partir de los datos que se estudian. En la sección 4 de esta memoria hemos presentado uno de los métodos no paramétricos más utilizados para la estimación de la función de densidad de probabilidad, donde la densidad es determinada solamente a partir de las observaciones de los datos con que se cuenta. Este tipo de métodos basado en funciones núcleos pueden ser utilizados de manera óptima en la creación de estimadores de densidad que a su vez obtengan una cuantificación vectorial del espacio [71, 72]. Fukunaga y Hayes [73] propusieron a su vez un algoritmo para la estimación de centros de grupos o "representantes" de los datos basado en la idea de que la densidad de probabilidad estimada de estos datos reducidos, utilizando el método de Parzen, sea la más parecidamente posible a la de los datos originales, utilizando para ello la entropía como criterio de similitud entre ambas estimaciones de densidad y todo desarrollado en el contexto de clasificación.

Nuestro trabajo se ha centrado en intentar resolver el problema descrito anteriormente en un marco matemáticamente formal e intenta dar respuesta al siguiente problema: dado un conjunto de datos de entrada, encuéntrese un conjunto reducido de puntos representantes cuya densidad de probabilidad sea lo más parecidamente posible a la densidad de probabilidad de los datos originales, de manera que estos puntos representantes no solo provengan de la misma distribución estadística del espacio original, sino que también lo cuantifiquen fielmente.

## 6.1. El nuevo funcional y su optimización

El nuevo problema matemático podemos plantearlo formalmente de la siguiente manera: dado un conjunto de datos $\mathbf{X}_i \in \Re^p, i = 1 \cdots n$ de dimensión $p$, encuéntrense $c$ datos subrogados, $\mathbf{V}_j \in \Re^p, j = 1 \cdots c$ de forma tal que la densidad de probabilidad estimada:

$$D(\mathbf{X}) = \frac{1}{c}\sum_{j=1}^{c} K(\mathbf{X} - \mathbf{V}_j; \alpha) \tag{6.1}$$

sea lo más parecida posible a la densidad de probabilidad de los datos originales. $K$ es una función tipo núcleo y $\alpha$ es el ancho de la misma que controla la suavidad de la densidad estimada.

En sentido general, sea $D(\mathbf{X};\theta)$ la densidad de probabilidad de una variable aleatoria $\mathbf{X}$, donde $\theta$ representa los parámetros desconocidos. Si $\mathbf{X}_i \in \Re^{p*1}$, $i$=1...$n$ denota los datos, entonces:

$$L = \prod_{i=1}^{n} D(\mathbf{X}_i; \theta) \tag{6.2}$$

es la función de verosimilitud, y el estimador estadístico más común para $\theta$ se obtiene maximizando esa función (ecuación (6.2)). En el caso de estimación paramétrica de la densidad, las medias y las varianzas son estimadas de esta forma [53]. Maximizar la función de verosimilitud es equivalente a maximizar su logaritmo, lo que hace las ecuaciones más tratables matemáticamente. Por lo tanto, combinando las ecuaciones (6.1) y (6.2) obtenemos el nuevo funcional:

$$\max l = \sum_{i=1}^{n} \ln\left(D(\mathbf{X}_i)\right) = \sum_{i=1}^{n} \ln\left(\frac{1}{c}\sum_{j=1}^{c} K(\mathbf{X}_i - \mathbf{V}_j; \alpha)\right) \tag{6.3}$$

Nótese que en este caso en particular el vector de parámetros está formado por los vectores diccionarios y por el ancho de la función núcleo: $\theta = \left\{\{\mathbf{V}_j\}, \alpha\right\}$.

Este nuevo funcional, expresando por la ecuación (6.3), puede ser resuelto de la misma forma en que se resolvió el funcional del algoritmo FuzzySOM mostrado en la sección anterior: tomando la derivada de la función (6.3) con respecto a $\mathbf{V}_j$ y haciéndola cero. De esta forma quedaría:

$$\sum_{i=1}^{n} \frac{K(\mathbf{X}_i - \mathbf{V}_j; \alpha)}{\sum_{k=1}^{c} K(\mathbf{X}_i - \mathbf{V}_k; \alpha)} \left[\mathbf{X}_i - \mathbf{V}_j\right] = \mathbf{0} \tag{6.4}$$

Que es equivalente a:

$$\mathbf{V}_j = \frac{\sum_{i=1}^{n} \mathbf{X}_i U_{ji}}{\sum_{i=1}^{n} U_{ji}} \tag{6.5}$$

Con:

$$U_{ji} = \frac{K(\mathbf{X}_i - \mathbf{V}_j; \alpha)}{\sum_{k=1}^{c} K(\mathbf{X}_i - \mathbf{V}_k; \alpha)} \tag{6.6}$$

Las derivación de este funcional puede demostrarse utilizando las reglas mostradas en el apéndice A, tal y como se hizo en el caso del algoritmo FuzzySOM.

A partir de estos resultados es evidente que la solución a la ecuación (6.4) no es lineal. Sin embargo, cuando se re-escribe en su forma equivalente a través de las ecuaciones (6.5) y (6.6), se aprecia una gran similitud con el algoritmo de FCM ([52], sección 3.2). De hecho, la única diferencia es la fórmula de actualización de los valores de pertenencia (ecuación (6.6) aquí). De esta forma, y a pesar de la notable diferencia en el planteamiento, existe una correspondencia formal directa entre ambos algoritmos dada por el parámetro de difusión $m$ en FCM y el ancho de la función núcleo en este método. Por lo tanto, el algoritmo para obtener la estimación de la función densidad de probabilidad con datos subrogados dado por la ecuación (6.1) es prácticamente igual al algoritmo de FCM (figura 3.2) con la salvedad de que se sustituye la ecuación (3.4) por la ecuación (6.6).

Análogamente al caso de obtención de las $\mathbf{V}_j$, y a diferencia de los métodos clásicos de estimación tipo núcleo de la función densidad de probabilidad, en este método es posible estimar el ancho óptimo del núcleo a partir de la función de coste. En el caso de que la función núcleo utilizada sea la Gaussiana (ecuación (4.4) de la sección 4.1), el funcional de la ecuación (6.3) puede ser re-escrito como:

$$
\begin{aligned}
l &= \sum_{i=1}^{n} \ln\left(\frac{1}{c}\sum_{j=1}^{c} K(\mathbf{X}_i - \mathbf{V}_j;\alpha)\right) \\
&= \sum_{i=1}^{n} \ln\left(\frac{1}{c(2\pi\alpha)^{p/2}}\sum_{j=1}^{c} \exp\left(-\frac{\|\mathbf{X}_i - \mathbf{V}_j\|^2}{2\alpha}\right)\right) \\
&= -n\ln c(2\pi\alpha)^{p/2} + \sum_{i=1}^{n} \ln\left(\sum_{j=1}^{c} \exp\left(-\frac{\|\mathbf{X}_i - \mathbf{V}_j\|^2}{2\alpha}\right)\right)
\end{aligned}
\tag{6.7}
$$

y tomando la derivada con respecto a $\alpha$, y haciéndola cero quedaría:

$$
-\frac{np}{2\alpha} + \sum_{i=1}^{n} \frac{\sum_{j=1}^{c} \frac{\|\mathbf{X}_i - \mathbf{V}_j\|^2}{2\alpha^2} \exp\left(-\frac{\|\mathbf{X}_i - \mathbf{V}_j\|^2}{2\alpha}\right)}{\sum_{j=1}^{c} \exp\left(-\frac{\|\mathbf{X}_i - \mathbf{V}_j\|^2}{2\alpha}\right)} = 0 \tag{6.8}
$$

lo que finalmente produce:

$$
\alpha = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{c} U_{ji}\|\mathbf{X}_i - \mathbf{V}_j\|^2 \tag{6.9}
$$

En este punto se deben señalar dos aspectos importantes de este método: primero, esta metodología que estamos utilizando es puramente estadística y produce un estimador de la densidad de probabilidad de los datos dado por la ecuación (6.1). A su vez, esta técnica resuelve un problema de agrupamiento de datos en el mismo sentido y de forma muy parecida a como lo resuelve FCM.

Adicionalmente es necesario señalar que este método presentado aquí muestra una semejanza importante con el algoritmo de E-M (Expectation-Maximization) para estimación de mezcla de distribuciones normales [53, 74]. Este algoritmo es ampliamente utilizado en reconocimiento de patrones no solo para la estimación no paramétrica de distribución de densidad de probabilidad, sino también para problemas de agrupamiento [75, 76] y está basado en la suposición de que los datos provienen de una población compuesta por una mezcla de distribuciones normales y por lo tanto la tarea de este método es la estimación, por máxima verosimilitud, de los parámetros de estas distribuciones.

El algoritmo que aquí se propone y que será descrito en el próximo apartado corresponde a la maximización por iteraciones de Picard del funcional descrito por la ecuación (6.3) y posee unas propiedades similares a las del algoritmo de E-M para estimación de mezcla de gaussianas [74], con la diferencia de que en el caso aquí propuesto no se estima la proporción de las poblaciones como se hace en el caso de E-M. De hecho, la similitud es lo suficientemente significativa como para poder garantizar la convergencia del algoritmo que se describirá en el apartado siguiente por analogía con el algoritmo de E-M para el cual su convergencia ha sido demostrada.

## 6.2. Algoritmo KCM (Kernel c-Means)

El nuevo algoritmo para estimar la densidad de probabilidad con datos subrogados (dado por el funcional del apartado anterior), es casi igual al algoritmo de FCM expresando en el diagrama de flujo de la figura 3.2, pero cambiando solamente la ecuación (3.4) que actualiza la pertenencia difusa por la ecuación (6.6) e introduciendo el cálculo de $\alpha$ de la ecuación (6.9). El siguiente esquema muestra este nuevo algoritmo al cual llamaremos "c-medias tipo núcleo" (Kernel c-means, KCM). Alternativamente, la figura 6.1 muestra el nuevo diagrama de flujo de dicho algoritmo.

a. Dado un conjunto de datos $\mathbf{X}_i \in \Re^{p*1}$, $i$=1...$n$; dado el número de grupos $c$, donde $n > (c+2)$
b. Inicializar las $U_{ji}$, para $i$=1...$n$ y $j$=1...$c$, satisfaciendo las restricciones dada por la ecuación (3.2)
c. Para $j$=1...$c$, calcular las $\mathbf{V}_j$ a través de la ecuación (6.5):

$$\mathbf{V}_j = \frac{\sum_{i=1}^{n} \mathbf{X}_i U_{ji}}{\sum_{i=1}^{n} U_{ji}}$$

d. Calcular el ancho del núcleo $\alpha$ mediante la ecuación (6.9):

$$\alpha = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{c} U_{ji} \left\| \mathbf{X}_i - \mathbf{V}_j \right\|^2$$

e. Para $i$=1…$n$ y $j$=1…$c$, calcular las $U_{ji}$ con la ecuación (6.6)

$$U_{ji} = \frac{K(\mathbf{X}_i - \mathbf{V}_j;\alpha)}{\sum_{k=1}^{c} K(\mathbf{X}_i - \mathbf{V}_k;\alpha)}$$

f. Ir al paso (c) hasta alcanzar la convergencia (valores pequeños de variación de los vectores diccionarios entre iteraciones)

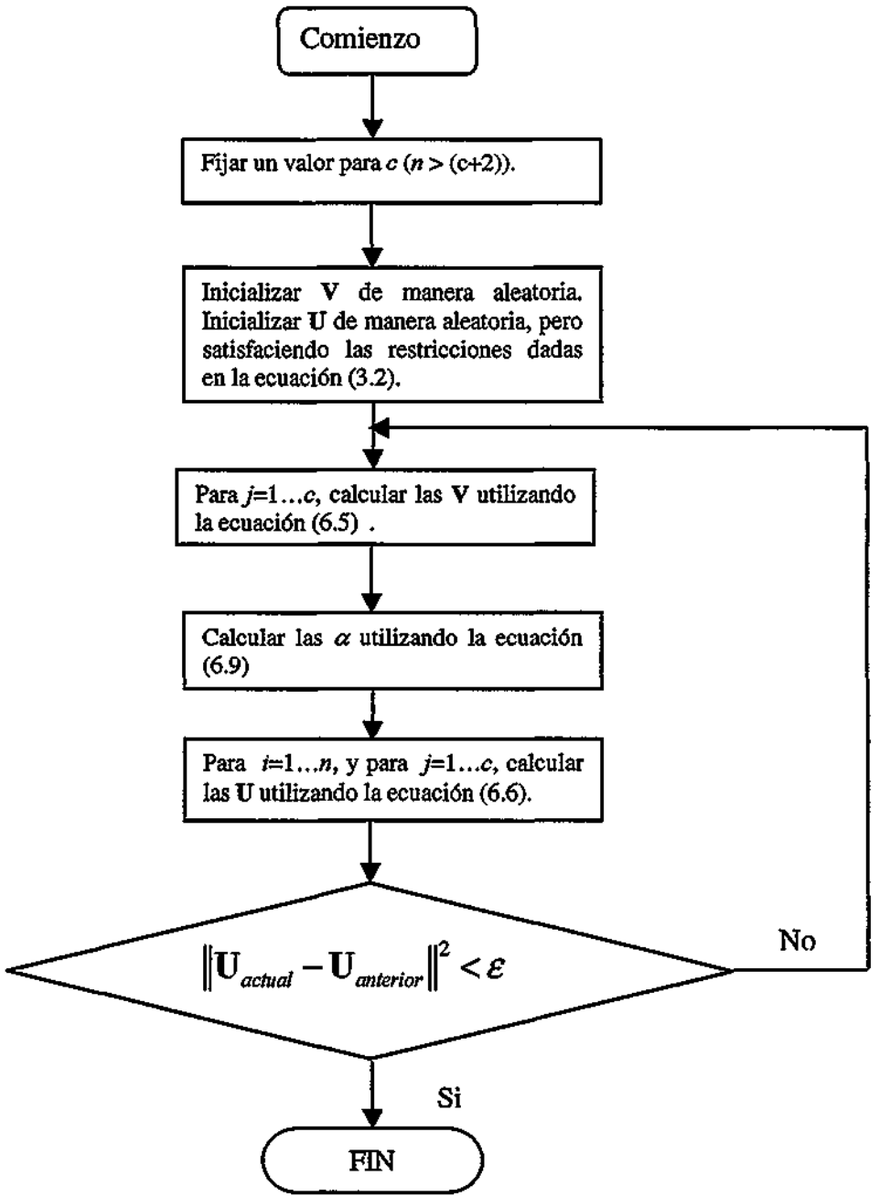



**Figura 6.1.** Diagrama de flujo del algoritmo de c-medias tipo núcleo (Kernel c-Means)

## 6.3. Ejemplos

Para demostrar la funcionalidad del algoritmo de KCM, hemos utilizado una vez más el conjunto de datos clásico de "Iris" [70], compuesto por 150 datos en 4D correspondientes a tres especies distintas de flores. Estos datos han sido muy utilizados durante mucho tiempo como pruebas a métodos de agrupamiento y clasificación (métodos supervisados y no supervisados). La figura 5.6 muestra la proyección de estos datos en 2D utilizando el algoritmo de FuzzySOM y donde se puede apreciar claramente los 3 grupos que componen estos datos. Se observa una clara separación de una de las especies (marcada como 1) , mientras que las otras dos (2 y 3) no son claramente separables. Debido al solapamiento de los grupos 2 y 3, es de esperar que los algoritmo fallen en su clasificación. Típicamente, el número de errores que normalmente cometen los métodos supervisados utilizando estos datos oscila entre 3 y 5. Sin embargo, el número de errores cometidos por los métodos no supervisados (métodos de agrupamiento) oscila entre 10-16 [51, 77].

El algoritmo KCM presentado en este apartado es un método no supervisado y por lo tanto cabe esperar que el número de errores cometidos al intentar separar estas 3 clases se encuentre en este rango. Aplicando el algoritmo para 3 grupos en 200 iteraciones, el número de errores cometidos en separar los grupos 2 y 3 de la figura 5.6 es de 14 elementos. El experimento se ha repetido 10 veces para distintas inicializaciones aleatorias de los vectores diccionarios, siempre obteniendo el mismo resultado que están en perfecta concordancia con los obtenidos por la mayoría de métodos de agrupamiento publicados a los cuales se les ha aplicado este conjunto de datos, lo que demuestra la capacidad de KCM en tareas de agrupamiento.

Adicionalmente al ejemplo anterior, este algoritmo ha sido probado extensamente en tareas de cuantificación vectorial sobre datos reales, y sus resultados han sido comparados con los obtenidos por algoritmos similares. Una descripción extensa de esta aplicación será mostrada en detalles en la sección 10 de esta memoria.

## 6.4. Discusión

El algoritmo de KCM presentado en esta sección será discutido de manera extensiva en la discusión de la siguiente sección, debido a que este método es un caso particular del algoritmo de KerDenSOM que se presentará a continuación.

# 7. Mapas auto-organizativos basados en estimación de densidad de probabilidad.

La idea propuesta para la creación de mapas auto-organizativos basados en el tipo de funciones de costo expuestas en la sección 5 podría ser extensible a otras modificaciones que expresen nuevas características de los mapas. Es también objeto del presente trabajo de tesis utilizar esta metodología para crear nuevos mapas auto-organizativos basados en la construcción de funcionales que demanden fidelidad a los datos y ordenamiento topológico. Combinando las ideas planteadas en la secciones 5 y 6 es posible crear nuevos mapas auto-organizativos siguiendo la misma lógica planteada con relación a los algoritmos de FCM y FuzzySOM: agregar al funcional del algoritmo de KCM descrito por la ecuación (6.3) una restricción que exprese auto-organización, es decir, que los vectores diccionarios generados por KCM se encuentren ubicados en ciertas regiones del espacio de salida con interconexiones explícitas que demanden suavidad o parecido entre ellas. De esta manera es posible obtener una función de coste bien definida que formule de manera rigurosa las dos características principales de los mapas auto-organizativos: fidelidad a los datos, expresado en términos de preservación de la densidad de probabilidad, y ordenamiento topológico en una malla regular.

La combinación de los mapas auto-organizativos con estimación de densidad de probabilidad lleva siendo estudiada desde hace varios años por distintos autores y la motivación principal de la unión de ambas metodologías recae en la carencia de una función de costo en el algoritmo clásico de SOM que explícitamente relacione la dependencia de los vectores diccionarios en el espacio de salida con la distribución estadística de los datos en el espacio de entrada. Si bien es cierto que SOM intenta preservar la densidad de probabilidad de los datos de entrada en el espacio de salida, es bien conocido que esta relación no es lineal, lo que provoca que los vectores diccionarios tiendan a subestimar zonas de alta densidad y sobreestimar zonas de baja densidad [26, 78]. Esta relación de densidades es también conocida como factor de magnificación [22, 79] y la importancia de su efecto debe tenerse en cuenta en aplicaciones en que se intente utilizar la densidad de probabilidad de los vectores diccionarios en el mapa como un estimador de la densidad de los datos originales.

Algunos ejemplos de intento de combinar ambas metodologías son los siguientes:

- *Bayesian Self-Organizing Map (BSOM):* Yin y Allison [80, 81] propusieron este método para resolver el problema de mezcla de gaussianas utilizando un mapa auto-organizativo. Este método ha demostrado ser superior al clásico algoritmo de E-M en cuanto a eficiencia de cómputo y a la presencia de mínimos locales.
- *Probabilistic SOM (PSOM):* Este método, propuesto por Wang y colaboradores [82], ha sido planteado en el contexto de segmentación de imágenes cerebrales cuyos histogramas son modelados como mezcla de gaussianas. Este algoritmo ofrece la ventaja de que optimiza el aprendizaje utilizando el histograma de pixeles de las imágenes, sin embargo, su versión estocástica es muy similar al algoritmo BSOM.
- *Generative Topographic Map (GTM):* Esta técnica puede ser considerada como una reformulación de SOM que utiliza una función de costo probabilística, optimizada también mediante el algoritmo de E-M. [64]. Este método representa un modelo de densidad de probabilidad que describe la distribución de los datos en un espacio de altas dimensiones en términos de un número mucho menor de variables latentes.
- Van Hull [83-85] ha propuesto varios algoritmos que intentan la generación de mapas topográficos a partir de reglas de aprendizaje de máxima entropía basadas en funciones núcleo.
- *Self-Organizing mixture network (SOMN)*: Yin y Allison [86] propusieron este método como una generalización del algoritmo de BSOM a otros tipos de mezclas de distribuciones.
- *Self-Organizing Reduced Kernel Density Estimator (RKDE):* Propuesto incialmente por Holmström y Hämäläinen [87, 88] este método utiliza la idea de la estimación de densidad de probabilidad con ventanas de Parzen (tipo núcleo) tal y como se ha descrito en la sección 4 de esta memoria, pero con la diferencia de que el número de funciones núcleos se reduce significativamente utilizando los vectores diccionarios generados por el algoritmo de SOM como centroides de las funciones núcleo.

Nuestro trabajo se ha centrado en intentar resolver el problema de la combinación de los mapas auto-organizativos con técnicas no paramétricas de estimación de la densidad de probabilidad en un contexto diferente de los métodos mencionados anteriormente mediante la combinación de métodos de estimación tipo núcleo de la pdf de los vectores diccionarios del mapa a la vez que se demanda el ordenamiento topológico del mismo, todo en un marco matemáticamente tratable y formal.

## 7.1. El nuevo funcional y su optimización

Con el objetivo de crear un mapa auto-organizativo basándonos ahora en la estimación de la densidad de probabilidad de los datos, se le adicionará al nuevo funcional dado por la ecuación (6.3) la parte B del funcional del algoritmo "FuzzySOM" (el término de penalización), quedando la nueva función de costo de la manera siguiente:

$$\max l_s = \sum_{i=1}^{n} \ln\left(\frac{1}{c}\sum_{j=1}^{c} K\left(\mathbf{X}_i - \mathbf{V}_j;\alpha\right)\right) - \frac{\vartheta}{2\alpha}\, tr\left(\mathbf{V}\mathbf{C}\mathbf{V}^T\right) \qquad (7.1)$$

**Parte A (fidelidad a los datos)** **Parte B (ordenamiento topológico)**

Siendo $\vartheta > 0$ el parámetro de suavidad para el mapeo y $\alpha$ el ancho de la función núcleo. Nótese la similitud de este nuevo funcional con el funcional del algoritmo de FuzzySOM dado por la ecuación (5.16), donde se conservan las dos partes fundamentales requeridas para formar el mapa auto-organizativo: fidelidad a los datos (en este caso dada por la estimación de la densidad de probabilidad) y ordenamiento topológico sobre una malla de menor dimensión.

Si utilizamos una función núcleo Gaussiana como la planteada por la ecuación (4.4), el funcional sería equivalente a:

$$l_s = -\frac{np}{2}\ln 2c\pi\alpha + \sum_{i=1}^{n}\ln\left(\sum_{j=1}^{c}\exp\left(-\frac{\left\|\mathbf{X}_i - \mathbf{V}_j\right\|^2}{2\alpha}\right)\right) - \frac{\vartheta}{2\alpha} tr\left(\mathbf{V}\mathbf{C}\mathbf{V}^T\right) \qquad (7.2)$$

utilizando la identidad mostrada en la ecuación (5.25) el funcional quedaría:

$$l_s = -\frac{np}{2}\ln 2c\pi\alpha + \sum_{i=1}^{n}\ln\left(\sum_{j=1}^{c}\exp\left(-\frac{\left\|\mathbf{X}_i - \mathbf{V}_j\right\|^2}{2\alpha}\right)\right) - \frac{\vartheta}{2\alpha}\sum_{j=1}^{c}\sum_{k=1}^{c}C_{jk}\mathbf{V}_j^T\mathbf{V}_k \qquad (7.3)$$

El primer paso sería maximizar el funcional con respecto a $\alpha$. Tomando la derivada parcial y haciéndola cero quedaría:

$$-\frac{np}{2\alpha} + \sum_{i=1}^{n}\sum_{j=1}^{c}\frac{\left\|\mathbf{X}_i - \mathbf{V}_j\right\|^2}{2\alpha^2}u_{ji} + \frac{\vartheta}{2\alpha^2}\sum_{j=1}^{c}\sum_{k=1}^{c}C_{jk}\mathbf{V}_j^T\mathbf{V}_k = 0 \qquad (7.4)$$

donde $U_{ji}$ sería idéntica a la de la ecuación (6.6). Sustituyendo queda:

$$\alpha = \frac{1}{np}\left(\sum_{i=1}^{n}\sum_{j=1}^{c}\left\|\mathbf{X}_i - \mathbf{V}_j\right\|^2 U_{ji} + \vartheta\sum_{j=1}^{c}\sum_{k=1}^{c}C_{jk}\mathbf{V}_j^T\mathbf{V}_k\right) \qquad (7.5)$$

Seguidamente se maximiza el funcional con respecto a $\mathbf{V}_j$. Tomando la derivada parcial y haciéndola cero quedaría:

$$\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{V}_j)U_{ji} - \vartheta\sum_{k=1}^{c}C_{jk}\mathbf{V}_k = \mathbf{0} \qquad (7.6)$$

que puede ser rescrita como:

$$\sum_{i=1}^{n}\mathbf{X}_i U_{ji} - \mathbf{V}_j\sum_{i=1}^{n}U_{ji} - \vartheta\sum_{\substack{k=1\\k\neq j}}^{c}C_{jk}\mathbf{V}_k - \vartheta C_{jj}\mathbf{V}_j = \mathbf{0} \qquad (7.7)$$

o equivalentemente:

$$\mathbf{V}_j = \frac{\sum_{i=1}^{n}U_{ji}\mathbf{X}_i - \vartheta\sum_{\substack{k=1\\k\neq j}}^{c}C_{jk}\mathbf{V}_k}{\sum_{i=1}^{n}U_{ji} + \vartheta C_{jj}} \qquad (7.8)$$

Al igual que se hizo con el algoritmo de FuzzySOM, una simple opción para la matriz **C** es el operador tipo Laplaciano (ecuación (5.12)). En este caso la ecuación (7.8) se simplifica de la siguiente manera:

$$\mathbf{V}_j = \frac{\sum_{i=1}^{n}U_{ji}\mathbf{X}_i + \vartheta\bar{\mathbf{V}}_j}{\sum_{i=1}^{n}U_{ji} + \vartheta} \qquad (7.9)$$

Donde $\bar{\mathbf{V}}_j$ denota el promedio de los vectores diccionarios que son vecinos inmediatos de $\mathbf{V}_j$ en la malla. En este valor promedio $\mathbf{V}_j$ queda excluido. Nótese la similitud de esta ecuación con la ecuación (5.39) que calcula los vectores diccionarios en el algoritmo de FuzzySOM.

## 7.2. Algoritmo KerDenSOM

El problema planteado en el apartado anterior puede ser resuelto de manera análoga al algoritmo de FuzzySOM, alternando las ecuaciones de cálculo del ancho del núcleo, de los vectores diccionarios y de la matriz U. El algoritmo propuesto, al que hemos llamado KerDenSOM (*Kernel Probability Density Estimator Self-Organizing Map*) es el siguiente:

1. Dado un conjunto de datos $\mathbf{X}_i \in \Re^{p\bullet 1}$, $i$=1...$n$; dado el número de nodos $c$; dado $\vartheta_1 > 0$ y $\vartheta_0 > 0$; dado $MaxSteps \geq 1$.
2. Inicializar $U_{ji}$ aleatoriamente, para $i$=1...$n$ y $j$=1...$c$, satisfaciendo las restricciones dadas por la ecuación (3.2)
3. Inicializar las $\mathbf{V}_j$: Para $j$=1...$c$, calcular:

$$\mathbf{V}_j = \frac{\sum_{i=1}^{n} \mathbf{X}_i U_{ji}}{\sum_{i=1}^{n} U_{ji}} \tag{7.10}$$

4. Inicializar $\alpha$. Calcular:

$$\alpha = \frac{1}{np}\left( \sum_{i=1}^{n}\sum_{j=1}^{c} \left\| \mathbf{X}_i - \mathbf{V}_j \right\|^2 u_{ji} + \vartheta_1 \sum_{j=1}^{c}\sum_{k=1}^{c} C_{jk} \mathbf{V}_j^T \mathbf{V}_k \right) \tag{7.11}$$

5. Para *Iter*=0 hasta Iter = *MaxSteps* ejecutar los pasos 6 al 10.
6. Calcular: $\vartheta = \exp\left(\ln(\vartheta_1) - \left(\ln(\vartheta_1) - \ln(\vartheta_0)\right) * Iter / MaxIter\right)$
7. Repetir hasta que converja:

   Para $j$=1...$c$, calcular $\mathbf{V}_j$ utilizando la ecuación (7.9):

$$\mathbf{V}_j = \frac{\sum_{i=1}^{n} U_{ji}\mathbf{X}_i + \vartheta \bar{\mathbf{V}}_j}{\sum_{i=1}^{n} U_{ji} + \vartheta}$$

*{Nótese que esta parte del algoritmo es del tipo "Gauss-Seidel", y debe ser repetido hasta que los vectores diccionarios cambien muy poco entre iteraciones}*

8. Calcular $\alpha$ utilizando la (7.5):

$$\alpha = \frac{1}{np}\left( \sum_{i=1}^{n}\sum_{j=1}^{c} \left\| \mathbf{X}_i - \mathbf{V}_j \right\|^2 U_{ji} + \vartheta \sum_{j=1}^{c}\sum_{k=1}^{c} C_{jk} \mathbf{V}_j^T \mathbf{V}_k \right)$$

9. Para $i$=1...$n$ y $j$=1...$c$, calcular $U_{ji}$ utilizando la ecuación (6.6):

$$U_{ji} = \frac{K(\mathbf{X}_i - \mathbf{V}_j; \alpha)}{\sum_{k=1}^{c} K(\mathbf{X}_i - \mathbf{V}_k; \alpha)}$$

10. Ir al paso 7 hasta que converja *(normalmente entendido de manera práctica cuando solo se producen cambios muy pequeños de las $U_{ji}$ entre una iteración y la siguiente)*

Alternativamente, la figura 7.1 muestra el nuevo diagrama de flujo de dicho algoritmo. Similarmente a lo que ocurre en el algoritmo FuzySOM, es bien sabido que algoritmos del tipo propuesto aquí son muy sensibles a las condiciones iniciales: el máximo local hacia el cual el algoritmo converge depende de la selección de los valores iniciales de los vectores diccionarios $\mathbf{V}_j$. Una manera de ayudar a una convergencia hacia el máximo global es utilizar una estrategia de enfriamiento determinista aplicada, en este caso, al factor de regularización $\vartheta$. El paso 5 (a partir del cual se encuentra casi todo el algoritmo) implementa esta estrategia. El algoritmo comienza con un valor grande de $\vartheta$ y una vez que converge, el valor de $\vartheta$ es disminuido y el algoritmo se repite una y otra vez hasta que se alcance el valor de $\vartheta$ deseado (variación de alta suavidad hacia la no-suavidad).

Esta estrategia puede mejorar significativamente los resultados del mapeo, sin embargo, el valor óptimo para $\vartheta$ es todavía una incógnita. Existen varias técnicas para intentar encontrar un valor razonable de regularización. Por ejemplo, una manera posible de estimar el "mejor" valor para $\vartheta$ es una medida de cross-validación [89] sobre los mapas generados.

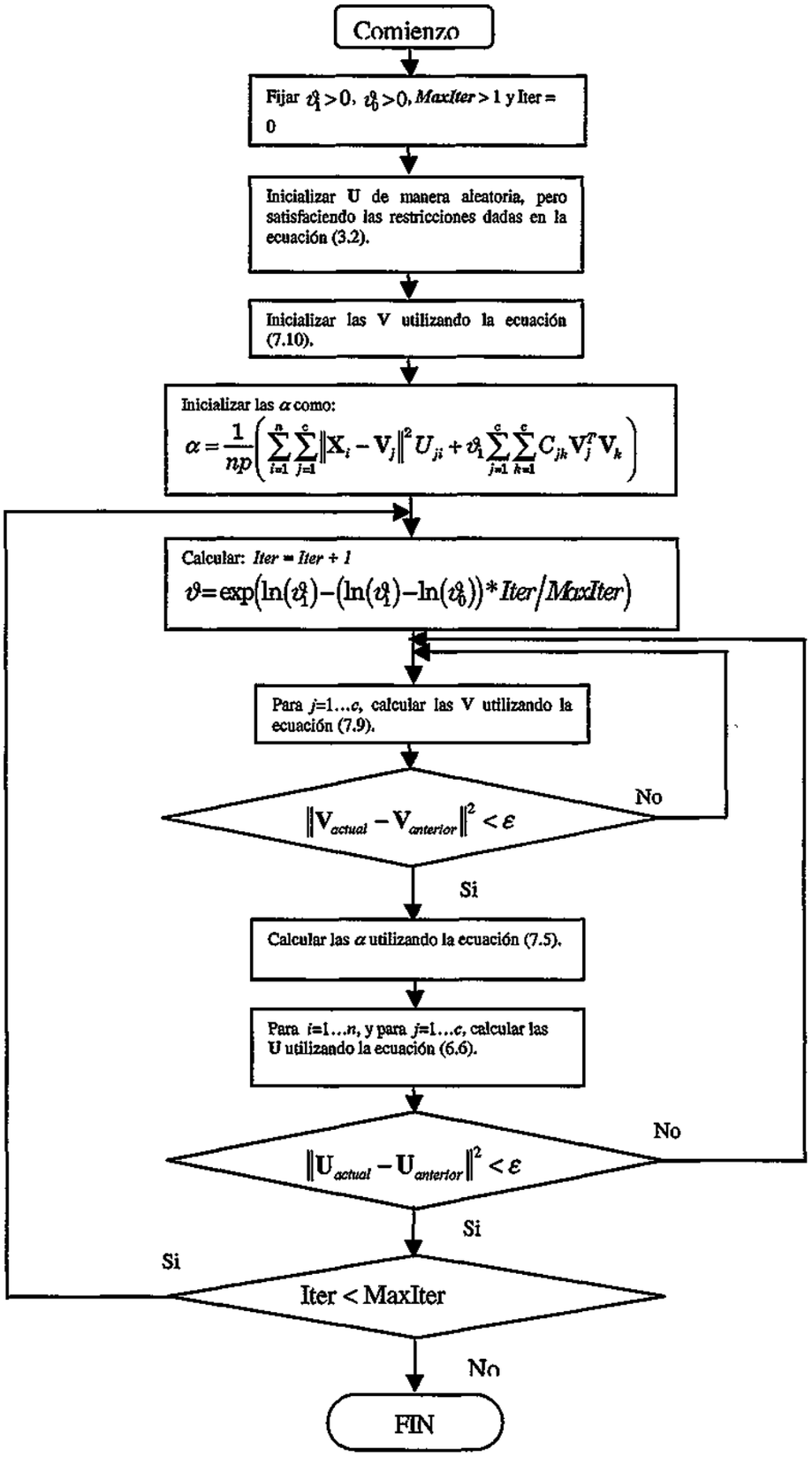



**Figura 7.1.** Diagrama de flujo del algoritmo KerDenSOM.

## 7.3. Ejemplos de mapeo

El algoritmo de KerDenSOM presentado en esta sección también se sometió a pruebas exhaustivas con los mismos datos sintéticos que los utilizados para probar el algoritmo de FuzzySOM y que han sido descritos en detalles en el apartado 5.5. El objetivo de estas pruebas era probar su capacidad para generar mapas suavemente distribuidos de los datos originales. Los resultados utilizando los 4 ejemplos mostrados en ese apartado fueron reproducidos fielmente con el algoritmo de KerDenSOM, obteniéndose los mismos mapas con las mismas propiedades que las mostradas en esa sección. Las figuras de los resultados han sido omitidas por no aportar ninguna información relevante a la ya mostrada en ese apartado. Es importante destacar que estas pruebas fueron realizadas para demostrar la capacidad de KerDenSOM de generar mapas auto-organizativos correctos, pero de ninguna manera se intentaba con ellas demostrar su superioridad con respecto al algoritmo de clásico de SOM o de FuzzySOM. En las secciones 8 a la 11, se mostrarán pruebas de este algoritmo con datos reales, donde se consiguieron resultados superiores a los obtenidos por SOM.

## 7.4. Preservación de la densidad de probabilidad

Como se ha mencionado anteriormente, una de las grandes aportaciones de este método es su capacidad de producir no solo un mapeo no lineal y organizado de los datos de entrada en un espacio de salida de menores dimensiones, sino también una estimación de la función densidad de probabilidad de los datos originales, dada por la ecuación (6.1). Para demostrar la veracidad de esta afirmación hemos realizado un experimento que utiliza estimadores de densidad para construir clasificadores y comparar los resultados de la capacidad de predicción de estos clasificadores en un conjunto de datos sintéticos.

La razón que motiva la utilización del error de predicción de clasificadores basados en densidad como medida de la calidad de los estimadores de densidad es debido a que medir el error de los estimadores de densidad en términos de verosimilitud con datos de prueba es muy poco intuitivo, por el contrario, su comportamiento en

problemas de clasificación suministra una medida muy clara y objetiva de la calidad de la estimación.

los estimadores de densidad se utilizan normalmente en análisis discriminante no paramétrico [54] y su descripción formal es la siguiente:

Sea $D=\{(x^k,l^k),k=1,\ldots,m\}$ un conjunto de $m$ datos con sus correspondientes etiquetas $l^k \in \{1,\ldots,C\}$ que denotan la clase a la que pertenece cada dato. Un clasificador de un nuevo dato $x$ se obtiene asignándole a este la clase $l$ con la máxima probabilidad condicional a posteriori $p(l|x)$. La probabilidad a posteriori puede ser derivada de $p(x|l)$ utilizando el teorema de Bayes:

$$p(l|x)=\frac{p(x|l)p(l)}{p(x)} \tag{7.12}$$

En el caso que nos ocupa tenemos un conjunto de muestras pertenecientes a una población conocida A y tenemos otro conjunto de muestras pertenecientes a otra población conocida B. Estas muestras formarán el conjunto de entrenamiento. Dada una nueva observación $z$ la pregunta a resolver es: ¿pertenece $z$ a la población A o a la población B? La respuesta basada en máxima verosimilitud asignaría la nueva observación $z$ a la población A si

$$p(z|A)p(A)>p(z|B)p(B) \tag{7.13}$$

en caso contrario sería asignado a la población B.

Por lo tanto, la tarea se reduce a estimar las probabilidades condicionadas para las clases A ( $p(A|x)$ ) y B ( $p(B|x)$ ). Las probabilidades a priori de cada clase $p(A)$ y $p(B)$ pueden ser estimadas como el porcentaje de muestras que existen en cada grupo.

Los datos de prueba utilizados para demostrar la capacidad de KerDenSOM de generar una estimación correcta de la función de densidad de probabilidad son un conjunto artificial de datos formado por dos clases en forma de anillo y mostradas en la figura 7.2. Estos datos han sido utilizados previamente como conjunto "estándar" para este tipo de pruebas de validación de estimadores de densidad de probabilidad [90], especialmente en un caso similar de mapa auto-organizativo diseñado como estimador de la pdf. [86].

Estos datos están formados por dos clases distribuidas en forma de círculo, pero con distintos centros y con cierto grado de solapamiento entre ellas (figura 7.2). Precisamente debido a esta estructura y a este grado de solapamiento, este conjunto de datos es interesante como conjunto de pruebas para clasificadores basados en densidad.

En este caso se generaron 200 puntos aleatorios para cada clase a partir de esta distribución. El conjunto original de 200 puntos se dividió en dos grupos distintos, uno para el entrenamiento y otro para las pruebas.

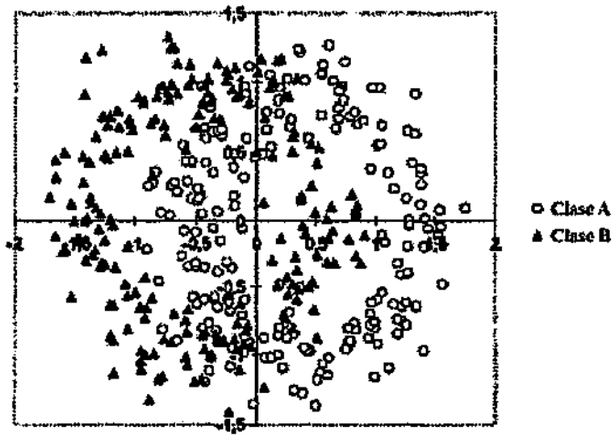



**Figura 7.2** Distribución de características de los datos utilizados para probar la eficacia de KerDenSOM como estimador de densidad de probabilidad. La figura muestra dos clases circularmente distribuidas, pero con distintos centros. Se generaron en total 200 muestras por clases.

A modo de comparación con las pruebas reportadas en [90] donde se utilizaron un conjunto amplio de estimadores de densidad basados en modelos de mezcla de gaussianas, utilizamos un mapa auto-organizativo formado por 20 neuronas distribuidas de manera lineal (mapa de 1x20). El algoritmo de KerDenSOM se ejecutó en 200 iteraciones variando el parámetro de regularización de 10 hasta 0.1 en 30 pasos de enfriamiento determinista. El experimento completo se repitió 50 veces utilizando una función núcleo gaussiana y otras 50 veces utilizando una función núcleo t-Student con 3 grados de libertad. La precisión promedio de la clasificación del conjunto de pruebas en las 50 repeticiones es de 85.8% para el núcleo Gaussiano y 85.42% para el núcleo t-Student.

Estos resultados no solo están en plena concordancia con los obtenidos para este mismo conjunto de datos por otros métodos de estimación de densidad de probabilidad [86, 90], sino que son ligeramente mejores. En [86] se obtuvo un porcentaje de clasificaciones correctas del 85.1% y en [90] la mejor clasificación se obtuvo con un

método Bayesiano y el mayor porcentaje de clasificaciones correctas obtenido fue de 82.7%, lo que demuestra que el método de KerDenSOM que aquí proponemos es un estimador eficiente de la función densidad de probabilidad.

## 7.5. Discusión

En este apartado hemos presentado el método de KerDenSOM, el cual es una versión regularizada del algoritmo de KCM presentado en la sección anterior pero con la diferencia de que en este algoritmo se generan mapas topológicamente correctos. De cualquier forma, y en ambos casos, se producen vectores diccionarios que intentan representar de la mejor manera, la función densidad de probabilidad de los datos originales.

Análogamente a como ocurre con el algoritmo de FuzzySOM el funcional de KerDenSOM, descrito por la ecuación (7.1), refleja las dos cualidades principales de los mapas auto-organizativos: fidelidad a los datos, expresada en este caso desde un punto de vista estadístico a través de la preservación de la pdf, así como ordenamiento topológico demandado por la parte derecha del funcional, donde se exige que las variaciones de los vectores diccionarios en el mapa ocurran lo más suavemente posible. Esta variante de mapa auto-organizativo, al igual que en el caso de FuzzySOM, constituye un nuevo intento de explicar, desde el punto de vista teórico, el proceso auto-organizativo que ocurre en el algoritmo clásico de SOM.

Una de las grandes ventajas de este método radica en que no solo se está obteniendo un mapa auto-organizativo, sino que a la vez se obtiene un estimado de la densidad de probabilidad de los datos dada por la ecuación (6.1). De esta manera este método puede ser utilizado con varios propósitos: análisis exploratorio de datos, agrupamiento y estimación de la densidad de probabilidad. Esta característica representa de manera clara un avance cualitativo con respecto al clásico algoritmo de SOM.

La función de densidad de probabilidad en este contexto podría ser utilizada, por ejemplo, no solo para tareas de análisis discriminante, sino también para separar los vectores diccionarios en el mapa de acuerdo a su estructura. Actualmente, el proceso de agrupamiento sobre el mapa es usualmente llevado a cabo de manera manual agrupando aquellos vectores diccionarios con características similares, sin embargo, utilizando la

pdf estimada por este método, técnicas más avanzadas de agrupamiento podrían ser aplicadas [91].

Adicionalmente, KerDenSOM posee a su vez una naturaleza difusa expresada por la matriz de pertenencia definida por la ecuación (6.6). Esta matriz permite la asignación de valores de probabilidad de pertenencia de los datos a cada uno de los vectores diccionarios, con las consecuentes ventajas que este tipo de planteamiento difuso ofrecen y que ya se discutieron en detalles en la sección 5 de esta memoria.

Es importante destacar también que estos métodos producen una estimación tipo núcleo de la función densidad de probabilidad, sin embargo, tal y como vimos en detalles en la sección 4, en el método original de Parzen [60] el ancho de la función núcleo constituye un parámetro crítico a la hora de realizar las estimaciones. Sin embargo, en el algoritmo de KerDenSOM, y debido a la naturaleza de su funcional, este parámetro es posible estimarlo de manera que el algoritmo seleccionará de manera iterativa el ancho de la función núcleo más adecuado para el conjunto de datos que se está analizando.

Por último quisiéramos destacar el hecho de que si el parámetro de regularización en este método se hace cero, el algoritmo automáticamente se convierte en el método de c-medias tipo núcleo (KCM) descrito en la sección anterior. De hecho, como ya hemos mencionado, KerDenSOM no es más que una versión regularizada de KCM y por ende todas las propiedades de estimación de la función de densidad de probabilidad y de naturaleza difusa discutidas en este apartado, son aplicables también a ese algoritmo. Es por eso que KCM puede entonces interpretarse como un método de cuantificación vectorial que produce vectores representantes que mejor representan la función densidad de probabilidad de los datos originales.

# CAPÍTULO III: APLICACIONES

# 8. Clasificación de Imágenes en Microscopía Electrónica

En este capítulo se pretende mostrar una de las aplicaciones más importantes relacionada con la utilización de los algoritmos de redes neuronales descritos en esta memoria. Primeramente haremos una breve descripción de la microscopía electrónica tridimensional, así como una descripción detallada del problema de clasificación que se pretende resolver. Se muestran, así mismo, resultados de la aplicación del método de KerDenSOM para la resolución de este problema con varios conjuntos de datos, los cuales son típicos ejemplos de los obtenidos comúnmente con este tipo de técnicas experimentales.

## 8.1. Introducción a la Microscopía Electrónica tridimensional

La microscopía electrónica (EM) en estudios biológicos se destaca hoy en día como una metodología muy poderosa que proporciona datos con un alto contenido de información, como son las imágenes. Centrándonos en el área de la biología que estudia los complejos mecanismos de interacción de las moléculas -la "Biología Molecular"-, los estudios de microscopía electrónica que se presentan en esta memoria se enmarcarían en el área conocida como "Biología Estructural de Macromoléculas Biológicas". El objetivo de estos estudios es siempre obtener información sobre la estructura tridimensional de una macromolécula determinada con el objetivo de conocer en detalle su mecanismo de acción: esto es, obtener su estructura tridimensional como un paso hacia la resolución de su función biológica.

Los desarrollos en la Biología moderna pretenden proporcionar una descripción cuantitativa de los complejos químicos que definen los organismos vivos. Para ello se está dedicando un gran esfuerzo a desarrollar modelos detallados de complejos macromoleculares biológicos que permitan estudiar las relaciones estructura-función. La determinación de la estructura de compuestos macromoleculares es en la actualidad uno de los problemas clave de la investigación bioquímica. Muchos procesos biológicos básicos, incluyendo el metabolismo de ácidos nucleicos, la fotosíntesis, la síntesis de proteínas y el ensamblaje de partículas virales, requieren la acción concertada de un gran número de componentes. La comprensión de la organización tridimensional de

estos componentes, así como sus detalles estructurales, si es posible a nivel atómico, es imprescindible para la interpretación de su función.

El microscopio electrónico de transmisión (MET) es en la actualidad una herramienta indispensable en la Biología y la Bioquímica Estructural, ya que proporciona el medio más directo de visualización de una estructura a nivel molecular.

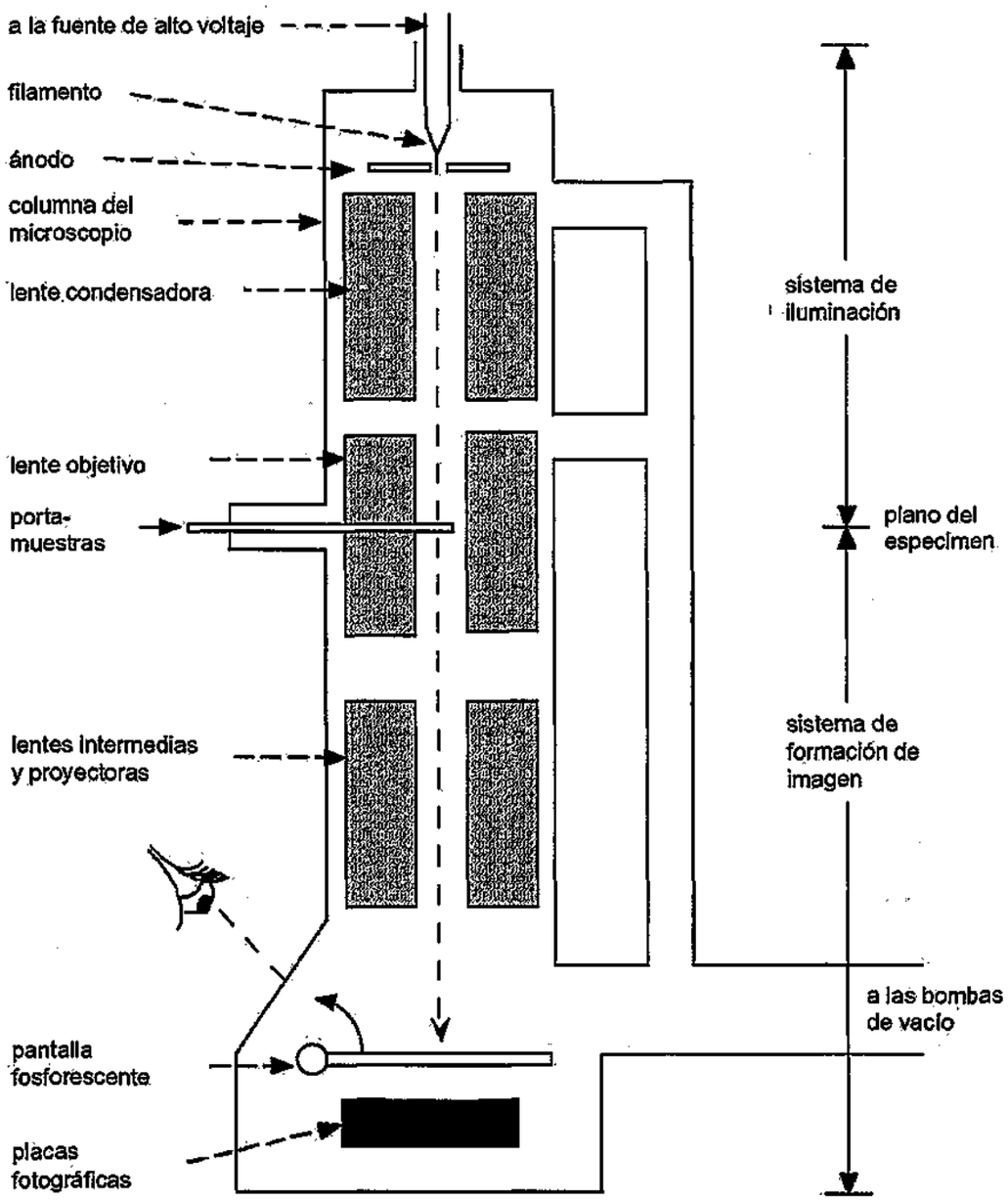



**Figura 8.1** Esquema de un microscopio electrónico de transmisión.

En su diseño general, el MET (figura 8.1) es similar al bien conocido microscopio óptico, aunque sus dimensiones son mucho mayores y está invertido con respecto a éste. La fuente de iluminación es un filamento o cátodo que emite electrones desde lo alto de una columna cilíndrica de unos dos metros de altura. Debido a que los electrones son dispersados por fenómenos de colisión con moléculas de gases, es necesario bombear el aire hacia fuera de la columna para producir vacío. En estas condiciones, los electrones son acelerados desde el filamento mediante un ánodo cercano, pasando por una abertura diminuta formando un fino haz que se dirige hacia la parte inferior de la columna.

Una serie de bobinas electromagnéticas situadas en la columna focalizan el haz de electrones de forma similar a la que las lentes de cristal focalizan la luz en un microscopio óptico. La muestra (montada en un soporte de unos 3 mm de diámetro) se introduce en el entorno de alto vacío de la columna a través de un sistema de esclusas, y se coloca en la trayectoria del haz, en el centro de la bobina que actúa como lente objetivo. Algunos de los electrones que pasan a través del espécimen son dispersados de acuerdo con la densidad local del material; parte de esas dispersiones dan cuenta del contraste de la imagen. Esta imagen se registra, o bien de forma analógica en una placa fotográfica, o sobre una pantalla fluorescente, o bien digitalmente en una cámara CCD.

En un MET, el límite de resolución impuesto por la longitud de onda de la luz visible puede ser superado gracias al empleo de electrones en lugar de fotones, ya que los electrones tienen una longitud de onda mucho más reducida. Para un voltaje de aceleración de 100 kVolts. (valor típico en Biología), el límite de resolución sería de 0.002 nm. Sin embargo, las lentes electromagnéticas sufren de aberraciones mucho más difíciles de corregir que las de las lentes de cristal y, como consecuencia, la resolución que se alcanza en la práctica con estos microscopios es, en el mejor de los casos, 0.1 nm (1 Ang.). Por otro lado, los problemas inherentes a la preparación de las muestras, su bajo contraste, y el daño por la radiación electrónica limitan la resolución para la mayor parte de los especimenes biológicos a un orden de magnitud más (10 Angs.). Afortunadamente, el empleo de nuevas técnicas de preservación de especimenes biológicos junto con la combinación de datos procedentes de imágenes y de difracción electrónica y, finalmente, la utilización de las técnicas de procesamiento de imagen,

están haciendo posible la recuperación de una gran parte del poder resolutivo teórico del MET.

En el MET, la información del espécimen se obtiene a partir de la radiación electrónica que lo ha atravesado. Los electrones tienen un poder muy alto de interacción con la materia y, por tanto, bajo poder de penetración. Consecuentemente, los especimenes que se exponen en un MET han de ser suficientemente finos (típicamente, con un espesor menor de 100 nm) para permitir que los electrones los atraviesen. Por otra parte, un MET estándar tiene una profundidad de foco de varios miles de Angs. [92]. Como consecuencia de esta característica y del rango de espesor visual de las muestras biológicas, las imágenes de MET se forman como superposición de los rasgos estructurales correspondientes a los diferentes niveles de la estructura 3D del espécimen. Tras extensos estudios basados en la teoría de formación de imagen en el MET [93], se ha llegado a la conclusión de que, para aplicaciones biológicas típicas, las imágenes de MET pueden ser consideradas imágenes de proyección del espécimen 3D.

El problema de obtener las relaciones tridimensionales entre las distintas partes del espécimen a partir de las imágenes que proporciona el MET es precisamente el problema de *reconstrucción tridimensional a partir de imágenes de proyección*. Este es un problema que se encuentra con frecuencia en numerosas disciplinas técnicas, médicas y científicas.

## 8.2. El problema de clasificación en Microscopía

La clasificación de imágenes de partículas individuales en microscopía electrónica es esencial como paso previo a la reconstrucción tridimensional del espécimen biológico que se estudia. Todos los métodos de reconstrucción tridimensional utilizados en EM se basan en el requerimiento estricto de que las imágenes de proyección individuales que se van a utilizar en el proceso de reconstrucción tridimensional corresponden a diferentes vistas del mismo espécimen biológico.

La obtención de un conjunto de partículas homogéneas está sujeto a diferentes problemas. En primer lugar, las diferencias entre las imágenes pueden ser realmente genuinas o pueden deberse a factores de posición como un mal alineamiento de rotación

o de traslación. Adicionalmente, la heterogeneidad estructural intrínseca de una población de partículas bioquímicamente homogéneas pertenecientes al mismo espécimen biológico también es una causa importante de diferencias en las imágenes de proyección. Finalmente, la baja relación señal/ruido típica de las imágenes de microscopía electrónica hace que este tipo de análisis sea muy complejo y difícil.

En este contexto, la clasificación de imágenes como un paso de pre-procesamiento es vital. Su objetivo es ordenar y separar la población de imágenes original en diferentes sub-poblaciones en un intento de ayudar a entender el espécimen que se estudia. Estos grupos pueden ser posteriormente utilizados o incluso descartados en el proceso de reconstrucción tridimensional (Figura 8.2)

Debido al hecho de que en la mayoría de los casos reales de estudio no existe información a priori de la macromolécula estudiada, el proceso de clasificación puede ser aún más complicado, por lo tanto, nuevos métodos de clasificación o agrupamiento que sean potentes, robustos y tolerantes al ruido son más que bienvenidos.

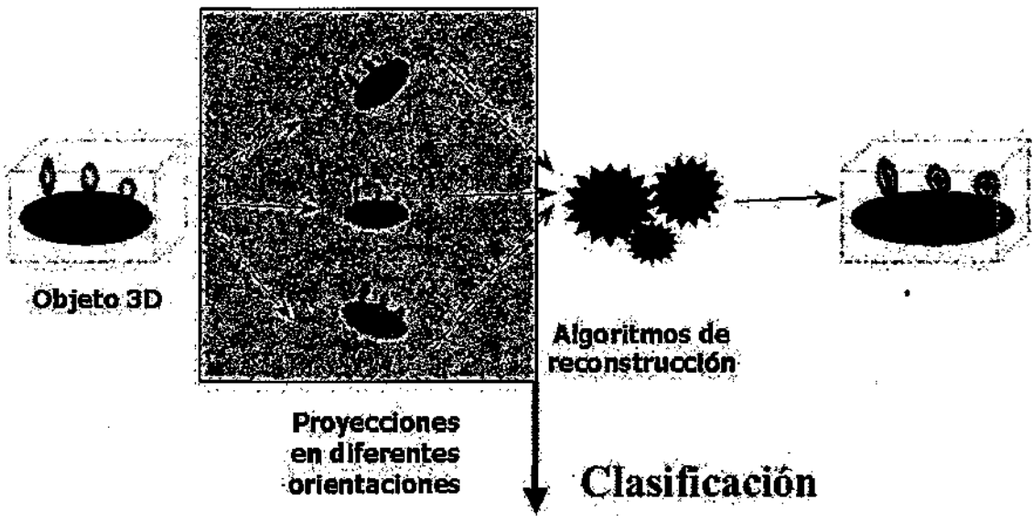



**Figura 8.2** Esquema representativo del proceso de reconstrucción tridimensional en EM, en donde se destaca el punto donde la clasificación de partículas tiene lugar.

En el contexto de clasificación de imágenes de partículas individuales en Microscopía han sido utilizados muchos métodos y técnicas [94, 95]. Por mencionar las más utilizadas y destacadas en este campo podríamos señalar la siguientes:

- **Análisis Estadístico Multivariado (Multivariate Statistical Análisis, MSA)** [96, 97]. Este tipo de técnicas estadísticas clásicas se han venido utilizando

para reducir el número de variables que caracterizan a las imágenes y conseguir de esta forma mayor eficiencia y robustez en la clasificación. Los métodos más utilizados son los de proyección lineal por componentes principales (PCA) y análisis de correspondencia (CA). Ambos métodos tienen en común en que se basan en la descomposición de la varianza total de los datos en componentes mutuamente ortogonales que son ordenados en orden decreciente de acuerdo a su magnitud. El objetivo de este tipo de análisis es encontrar un conjunto de vectores que definan las direcciones de las extensiones principales de la nube de puntos formada por los conjuntos de datos experimentales, en este caso, por el conjunto de imágenes. Intuitivamente estas direcciones principales se construyen de la siguiente manera: (i) Encontrar la máxima extensión de la nube de puntos; (ii) encontrar el vector perpendicular al primero, que apunte en la dirección de la siguiente extensión más grande de la nube de puntos; (iii) encontrar el vector perpendicular al primero y al segundo, que apunte en la dirección de la siguiente extensión más grande de la nube de puntos, y así sucesivamente. Estas mediciones sucesivas que describen la forma de la nube de datos son las componentes de la varianza total inter-imagen y el método para obtenerlas se llama Análisis por Componentes Principales (PCA). CA se distingue del PCA por utilizar una métrica distinta para calcular las distancias entre los datos: en vez de utilizar distancia Euclídea, se utiliza *Chi-Squared* ($\chi^2$). La diferencia principal entre utilizar CA y PCA en imágenes de microscopía electrónica radica en que CA ignora factores multiplicativos entre las diferentes imágenes, lo cual lo hace muy atractivo para trabajar con imágenes obtenidas de distintas micrografías sin necesidad de re-escalarlas.

Este tipo de técnicas estadísticas han sido útiles para clasificar en distintos grupos imágenes heterogéneas. Debido a que PCA y CA son métodos de reducción de dimensionalidad, la representación de estos datos transformados ayuda a evidenciar la separación de imágenes heterogéneas en subgrupos y el algunos casos una simple inspección visual de mapas de factores es suficiente para clasificar distintas vistas de una misma molécula [94]. Sin embargo, a medida que el análisis de partículas individuales en microscopía electrónica se

ha ido extendiendo para incluir nuevos especimenes, se ha detectado que macromoléculas que presentan una heterogeneidad estructural menos evidente son difíciles de analizar utilizando estos métodos. Usualmente este tipo de macromoléculas producen una distribución bastante plana de los valores propios (eigenvalues), imposibilitando la separación en grupos en los mapas de factores. Adicionalmente, cuando se utilizan técnicas de criomicroscopía, las imágenes de proyección presentan mucho menos contraste y una relación señal-ruido mucho menor, dificultando aún mas su clasificación utilizando estos métodos. La figura 8.3 muestra un ejemplo de utilización de CA en clasificación de imágenes.

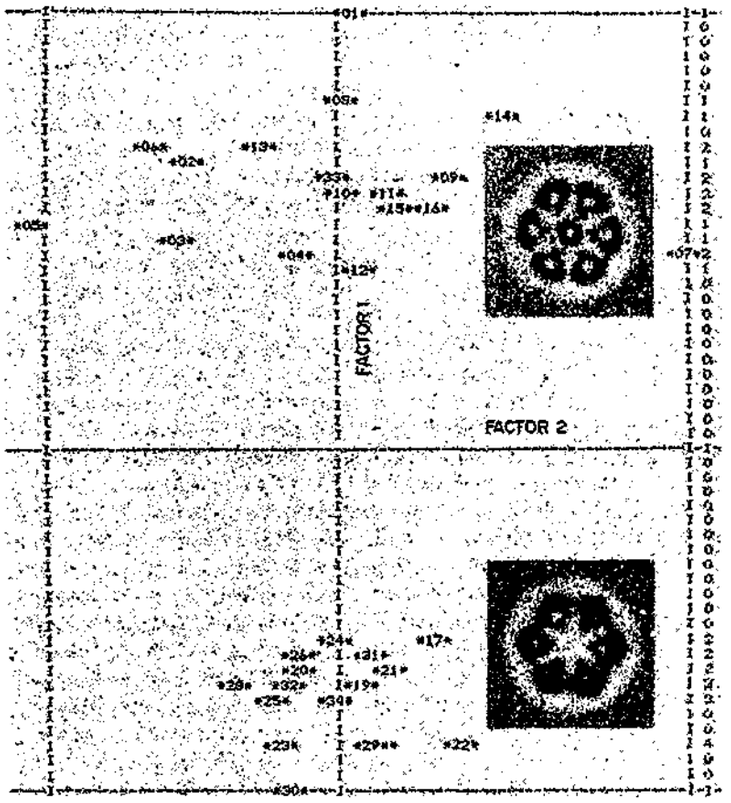



**Figura 8.3** Mapa de factores (Factor 1 vs. Factor 2) resultante de utilizar análisis de correspondencia en un conjunto de moléculas de hemoglobina.

- **Clasificación Jerárquica Ascendente (Hierarchical Ascendant Classification, HAC)** [98]: La clasificación jerárquica ascendente es un método de agrupamiento basado en la construcción de un jerarquía indexada entre los objetos que van a ser clasificados. Estas técnicas se caracterizan por la forma en que generan relaciones jerárquicas anidadas entre grupos distintos.

La jerarquía es creada de manera tal que en el nivel más bajo cada objeto individual es tratado como un grupo, a partir de aquí el método se basa en encontrar los dos grupos más cercanos para unirlos en un nuevo grupo. El proceso se repite hasta que el conjunto completo de datos es aglomerado en un único gran grupo. Los distintos tipos de algoritmos jerárquicos aglomerativos difieren entre sí en cómo calculan la distancia entre grupos, para lo cual se siguen criterios muy distintos como son:

a. Unión simple (single linkage) [99]: mide la distacia entre grupos como la distancia entre aquellos elementos de cada grupos que están más cercanos.
b. Unión completa (complete linkage) [100]: utiliza como criterio la distancia entre los elementos mas alejados de cada grupo.
c. Unión promedio (average linkage) [101]: mide la distancia entre grupos como la distancia media entre cada par de observaciones entre ambos grupos.
d. Unión por centros (centroid linkage) [101]: se basa en la distancia entre los centros de cada grupo.
e. Método de Ward [102]: aglomera grupos que minimicen el error total intra-grupo.

En el caso particular de la clasificación de imágenes de microscopía electrónica, el método más utilizado es el de Ward. La figura 8.4 muestra un ejemplo de este tipo de clasificación. Estos métodos, que han sido muy utilizados en este campo y que son muy simples desde el punto de vista conceptual y computacional, sufren de muchos problemas así como de una falta de robustez en presencia de datos ruidosos. Por ejemplo, las soluciones pueden no ser únicas y dependen en gran medida del orden en que los datos son suministrados al algoritmo. Adicionalmente, la naturaleza determinista de estos métodos y la imposibilidad de re-evaluación de los resultados una vez comenzado el algoritmo pueden producir agrupamientos basados más en características locales que en características globales de la estructura de los datos. La causa principal de estos problemas radica en que estos métodos han sido principalmente desarrollados para ser utilizados en agrupamiento de datos

en sistemas aparentemente jerárquicos, como es el caso de la filogenia en biología, y muy probablemente no estén completamente preparados para trabajar con datos con estructura no jerárquica, altamente dimensional y ruidosos como es el caso de los datos utilizados en microscopía electrónica.

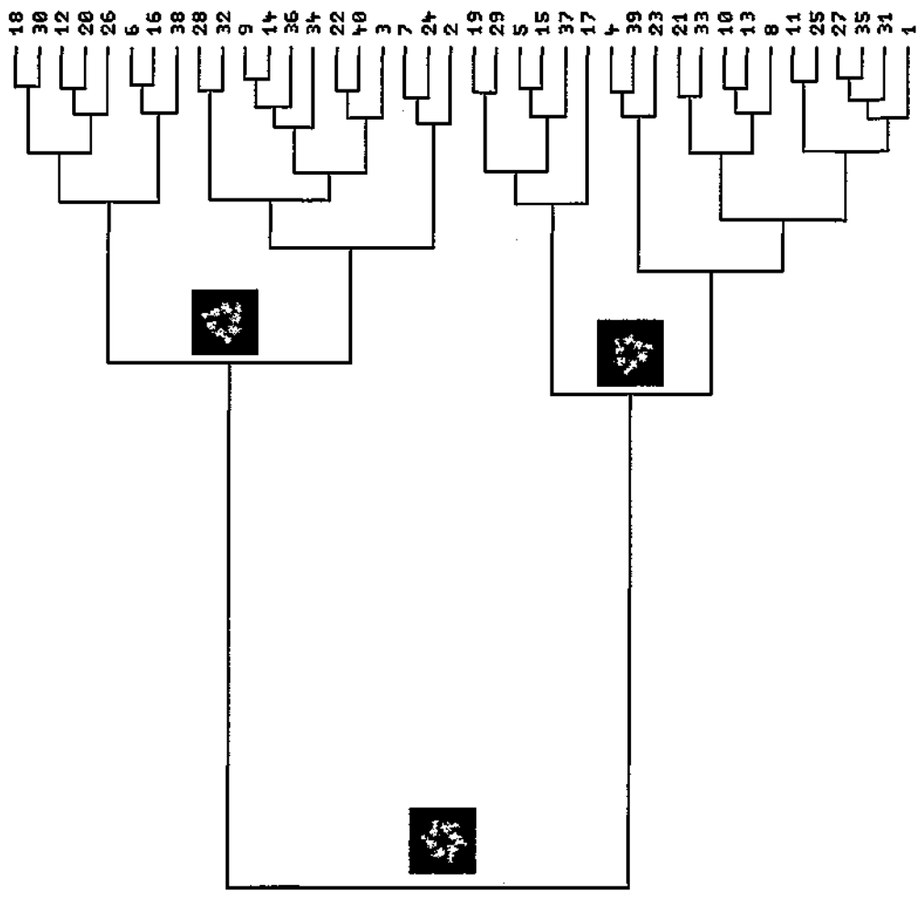



**Figura 8.4** Ejemplo de clasificación jerárquica aplicada a imágenes de microscopía. Las imágenes mostradas pertenecen a la imagen media de los grupos formados a ese nivel de corte del dendograma.

- **Agrupamiento Particional: k medias (k-means)** [14, 53]: Este algoritmo particional es uno de los más comunes y utilizados en el campo del reconocimiento de patrones. Su popularidad radica fundamentalmente en su simplicidad y efectividad. En este tipo de agrupamiento, los datos son divididos iterativamente en un número predeterminado de clases de la siguiente manera:

a. Se toman aleatoriamente $k$ datos del conjunto inicial de datos (imágenes en este caso) y se utilizan como centros de grupos (centroides)
b. Se asigna cada elemento del conjunto de datos (imágenes) al centroide más cercano basado en una medida de distancia cualquiera (usualmente distancia Euclídea).
c. Cuando todos los datos han sido asignados, se recalcula la posición de los centroides.
d. Se repiten los pasos b y c hasta que las posiciones de los centroides dejen de variar.

A pesar de la sencillez conceptual y práctica de este algoritmo, sus numerosas desventajas lo hacen impracticables en aplicaciones complejas como las que se estudian en este capítulo. Por solo mencionar algunas, podemos decir que la solución final se ve afectada por la inicialización de los centroides, lo que provoca que el algoritmo converja a una solución no óptima del espacio de soluciones (mínimo local). Así mismo, la forma geométrica de los grupos extraídos depende de la medida de distancia utilizada. Por ejemplo, si la distancia utilizada es la Euclídea, el algoritmo intentará extraer grupos con forma hiper- esférica. Por último, el número de grupos debe ser seleccionado a priori y en problemas reales esta información casi nunca es conocida a priori.

- **Algoritmo de c-medias difuso (Fuzzy c-means, FCM)** [103]: El método de agrupamiento FCM descrito en la sesión 3 de esta memoria es un proceso de agrupación de objetos en una misma clase o grupo, muy parecido al algoritmo de k-medias, pero la manera de realizar este agrupamiento es difusa, lo que significa que las imágenes no son asignadas exclusivamente a un solo grupo, sino parcialmente a todos pero con distinto grado de pertenencia. Esto permite cierta flexibilidad, principalmente cuando las imágenes a separar tienen características muy similares entre sí que no permiten una separación clara entre ellas.
- **Sistema híbrido (k-means y HCA)** [98, 104-107]: Este tipo de metodología combina los métodos particionales con los jerárquicos. Usualmente se utiliza una de las dos técnicas como paso inicial y la otra como post-procesamiento,

permitiendo de esta manera intentar solventar los problemas de cada una de ellas.

- **Mapas auto-organizativos de Kohonen (SOM):** Las redes neuronales en el caso de clasificación de imágenes de microscopía electrónica fueron introducidas por primera vez por Marabini y Carazo [108]. Estos autores propusieron el uso de los mapas auto-organizativos de Kohonen descritos en el capítulo 2 para realizar las tareas de clasificación de imágenes. Un ejemplo de este tipo de aplicación puede verse en la figura 8.5. En este caso se utilizaron un conjunto de 407 imágenes correspondientes a vistas laterales del complejo macromolecular TCP-1 [109]. Antes de ser procesadas por el algoritmo de SOM, estas imágenes fueron centradas entre sí pero no alineadas rotacionalmente. Como puede observarse en la figura 8.5a, resulta prácticamente imposible detectar patrones de variabilidad en las imágenes de entrada de manera visual, sin embargo el algoritmo ha ordenado las moléculas en el mapa resultante de acuerdo a su orientación en el plano (cambios en la rotación). Los vectores diccionarios ubicados en el borde del mapa (figura 8.5b) muestran claramente versiones rotadas del complejo TCP-1. Es evidente que SOM produce una serie de vectores diccionarios con un bajo ruido en donde es fácil apreciar directamente las variaciones estructurales de los datos. Al mismo tiempo ofrece una clasificación directa de los datos al asignar a cada vector diccionario el conjunto de datos de entrada que más fielmente este representa (figura 8.5d).

  Los mapas auto-organizativos de Kohonen (SOM), han sido y son ampliamente utilizados en diferentes estudios macromoleculares reales [110-115] demostrando sus propiedades de robustez y capacidad de discriminación de datos complejos y de alta dimensión.

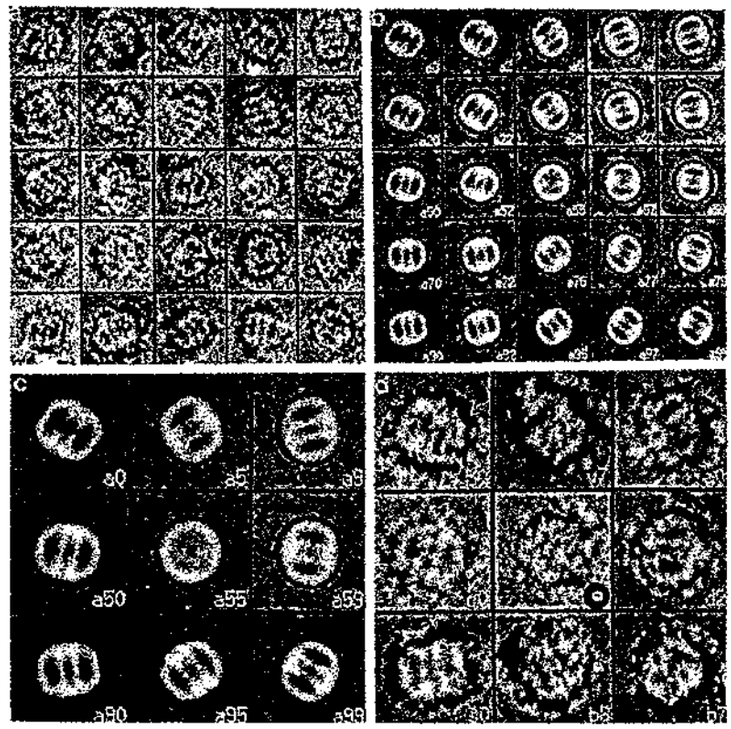


**Figura 8.5** Mapa auto-organizativo de Kohonen aplicado a la clasificación de imágenes de microscopía electrónica del complejo macromolecular TCP-1. a) Ejemplo de algunas de las imágenes de entrada. b) Vectores diccionarios obtenidos por el algoritmo. c) Vista detallada de alguno de los vectores diccionarios. d) Ejemplo de imágenes asignadas a los vectores diccionarios mostrados en c).

- **Redes de agrupamiento difusas de Kohonen (Fuzzy Kohonen Clustering Network, FKCN):** Este tipo de red neuronal descrita en la sección 3.3 es una variante que combina los conceptos de auto-organización y agrupamiento borroso [51]. La parte algorítmica de este método ya ha sido discutida en detalles en capítulos anteriores de esta memoria y solo comentaremos su aplicación a la clasificación de imágenes de microscopía electrónica, la cual se aplicó a imágenes de la helicasa hexamérica G*40*P del bacteriófago SPP1 [116]. En ese caso se intentó reproducir los resultados obtenidos por Bárcena y colaboradores [114], en los cuales se pretendía encontrar la variación estructural de este conjunto de imágenes muy parecidas estructuralmente y que comparten la misma simetría rotacional. La principal motivación de esta aplicación en microscopía electrónica radicaba en intentar combinar las propiedades más generosas de dos métodos conceptualmente distintos pero igualmente eficientes: el FCM y los SOM. Sin embargo, como ya hemos

mencionado en la sección 3.3, este solo ha sido un intento por combinar estas dos ideas que pertenece más al campo de las técnicas de agrupamiento que al campo de los mapas auto-organizativos. De hecho, este tipo de estudios de cierta forma inspiró el desarrollo de los nuevos métodos de clasificación descritos en secciones anteriores.

A pesar de la cantidad de técnicas y metodologías empleado en la clasificación de imágenes en microscopía, el problema general de clasificación sigue sin resolverse completamente debido a la complejidad cada vez más creciente de los datos generados. Esta situación nos motivó en el desarrollo de nuevos métodos, como los mapas auto-organizativos presentados en esta memoria, que ofrecen características teóricas muy sólidas matemáticamente. En las siguientes sesiones presentaremos aplicaciones de estos nuevas métodos a datos reales de microscopía electrónica para así demostrar su eficacia.

## 8.3. Detección de heterogeneidades en Helicasas hexaméricas.

Las helicasas, descubiertas en 1976 [117, 118], son las enzimas responsables de la separación de las hebras de ácidos nucleicos bicatenarios, un proceso que resulta esencial en prácticamente todos los aspectos del metabolismo de los ácidos nucleicos y que hace que estas proteínas sean muy abundantes, de modo que cada organismo parece presentar su propia colección de helicasas, cada una con una labor especializada dentro de la célula [119].

Estas helicasas juegan un papel fundamental en casi todos los procesos de la célula (replicación, recombinación, reparación de ADN, etc.). Recientemente, se ha establecido que algunas enfermedades hereditarias humanas tales como el síndrome de Cockayne o el de Werner están relacionadas con mutaciones en proteínas helicasas [120]. Todo esto hace que el estudio de estas proteínas sea un importante tema de investigación actual.

Dentro de las helicasas hexaméricas, las replicativas ocupan un lugar destacado. Estas enzimas tienen una función esencial ya que son el principal factor responsable de la apertura del ADN de doble cadena en las horquillas de replicación [121]. Sin embargo, la visión de las helicasas como proteínas que únicamente abren la doble hélice

de ADN peca de simplista. Las helicasas hexaméricas replicativas participan activamente en las diferentes etapas de la replicación y establecen múltiples interacciones. Junto con otras proteínas de la maquinaria replicativa, como la polimerasa y la primasa, forman un complejo multiproteico mayor, el replisoma, a cuya integridad funcional contribuyen y al que sirven como motor [122]. En el contexto de esta memoria plantearemos un ejemplo de estudio de heterogeneidad de un tipo específico de helicasa replicativa: la G*40*P del bacteriófago SPP1 de *Bacillus Subtilis*.

### 8.3.1. Procesamiento de Imagen

De entre las micrografías obtenidas por el proceso descrito en la sesión 8.1, se seleccionaron manualmente unas 2560 partículas con forma globular, no solapantes y teñidas homogéneamente. Una vez seleccionadas, se extrajeron las imágenes individuales en ventanas de 50x50 píxeles. La figura 8.6 muestra la micrografía electrónica obtenida y algunas de las imágenes seleccionadas manualmente. En el proceso, la estadística del ruido de cada imagen se normaliza a un valor de media de 0 y una desviación típica de 1 y esta transformación se aplica a todos los píxeles de la imagen [123].

Las imágenes de partículas individuales en microscopía electrónica tienen una relación señal-ruido intrínsecamente baja. La idea fundamental del procesamiento bidimensional consiste en mejorar la relación señal-ruido mediante el promediado de las imágenes. Para ello, las partículas se deben alinear rotacional y traslacionalmente, lo que se realizó según el método general que se describe a continuación. En primer lugar, las imágenes se centraron utilizando correlación cruzada con una máscara anular generada por promediado de todas las partículas. El alineamiento traslacional y rotacional de las imágenes se realizó utilizando métodos de correlación cruzada y el algoritmo PSPC (Pyramidal System for Prealignment Construction) [124], que es una modificación del denominado ``método libre de patrón" [125]. Lo que estos métodos intentan evitar es el sesgo en el resultado que se ha demostrado que produce la elección de un patrón inicial para el alineamiento de las imágenes [126]. En el sistema descrito en [124] se obtiene un patrón a partir de correlacionar y promediar, de manera piramidal, las propias imágenes de la población. A continuación se utilizaron métodos de correlación cruzada para el refinamiento final en la posición y orientación de las

imágenes individuales. se eliminaron del conjunto de la población aquellas imágenes que sufrían desplazamientos y giros superiores a un cierto umbral en esta etapa de refinamiento.

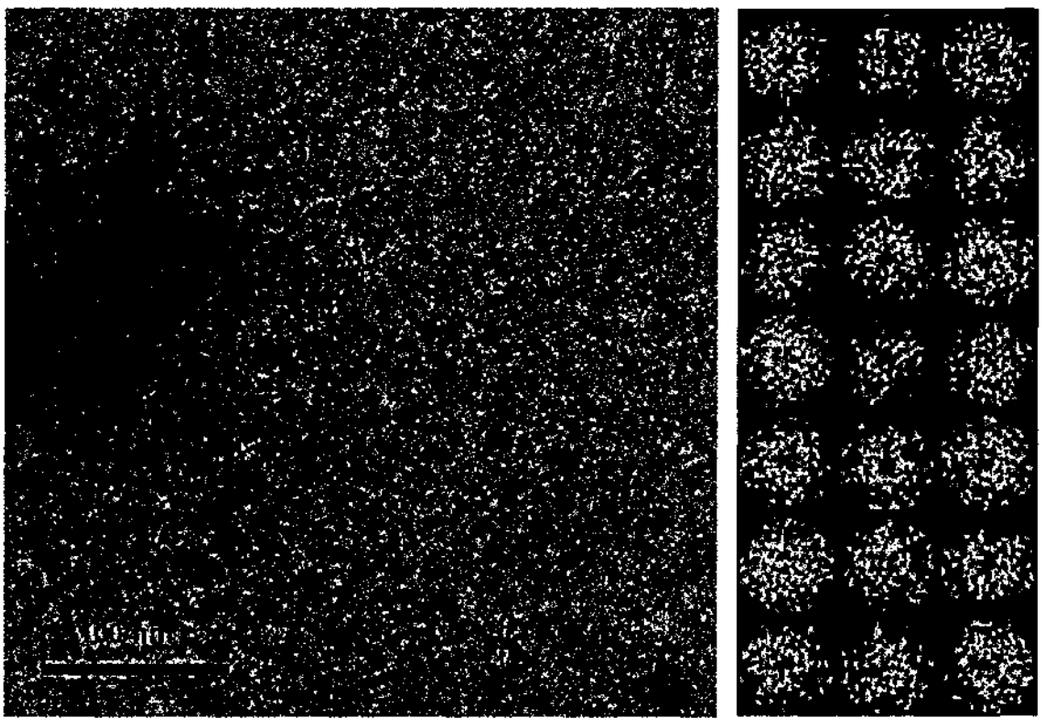


**Figura 8.6** Micrografía que muestra un campo típico de preparaciones de G*40*P teñidas negativamente. Las puntas de flecha señalan algunas de las partículas seleccionadas. A la derecha, se muestra una galería de partículas extraídas manualmente de las micrografías.

Para el procesamiento de imagen se utilizaron los programas integrados en el paquete de software Xmipp [127] que ha sido desarrollado en la Unidad de Biocomputación del Centro Nacional de Biotecnología. Este software, descrito en el apéndice C, es de dominio público y se encuentra accesible en la siguiente dirección: `http://biocomp.cnb.uam.es/Biocomp/public/Software`

### 8.3.2. Clasificación de espectros rotacionales

Las 2458 imágenes resultantes del proceso de alineamiento rotacional y traslacional descrito en la sesión anterior fueron sometidas a un proceso de extracción de características previo a la clasificación. El tipo de característica que fue extraído de las imágenes fue su espectro rotacional [128]. El motivo por el cual se utilizaron los espectros rotacionales está basado en la sospecha previa de que las imágenes de

proyección de esta estructura macromolecular presentaban diferencias con respecto a la simetría de las partículas (simetría de orden 3 y orden 6 respectivamente) [129]. Esto hizo pensar que la mayor fuente de variabilidad correspondía precisamente a esta característica, lo cual justificaba plenamente su estudio directo previo al análisis de los valores de densidad de las imágenes. Esta aproximación, propuesta por primera vez en [114], hace que el análisis de la variabilidad sitúe en un primer plano la simetría de las partículas, perdiendo importancia otros factores de heterogeneidad tales como las diferencias en tinción y otros factores estructurales.

Los espectros rotacionales son funciones derivadas de una transformación de tipo Fourier-Bessel que aportan información sobre el orden de la simetría que tiene una partícula con respecto a un eje predeterminado [128]. En esencia, el proceso consiste en realizar una transformada de Fourier de la imagen en coordenadas polares. Las funciones base de la transformada son circularmente periódicas de tal modo que la componente de orden uno (o armónico uno) se caracteriza por poseer un máximo y un mínimo, la de orden dos (o armónico dos), tiene dos máximos y dos mínimos, y así sucesivamente. El espectro rotacional de la imagen es el módulo al cuadrado de cada una de las funciones base de la transformada. Resulta claro que este tipo de transformación ha de depender críticamente de la elección del origen de coordenadas polares. En este estudio, para calcular este valor, las imágenes se centraron y alinearon previamente a su clasificación. A partir de la imagen media global se calcularon las coordenadas que minimizaban el armónico 1 (ausencia de simetría rotacional). Ese valor se utilizó como origen de coordenadas para la determinación del espectro rotacional de cada una de las imágenes individuales.

Para el cálculo del espectro rotacional se utilizaron solamente los primeros 15 harmónicos, creando de esta forma 2458 vectores de dimensión 15. La figura 8.7 muestra una pequeña galería de este tipo de datos.

Utilizando este conjunto de datos compuesto por espectros rotacionales, hemos aplicado el algoritmo de KerDenSOM para estudiar las posibles variaciones estructurales de las partículas. Para este análisis se empleó un mapa rectangular de 7x7 nodos. Como función núcleo utilizamos un núcleo Gaussiano. El algoritmo se calculó en 200 iteraciones variando el parámetro de suavidad desde 100 hasta 10 en 20 pasos de enfriamiento determinista. La figura 8.8 muestra el mapa resultante.

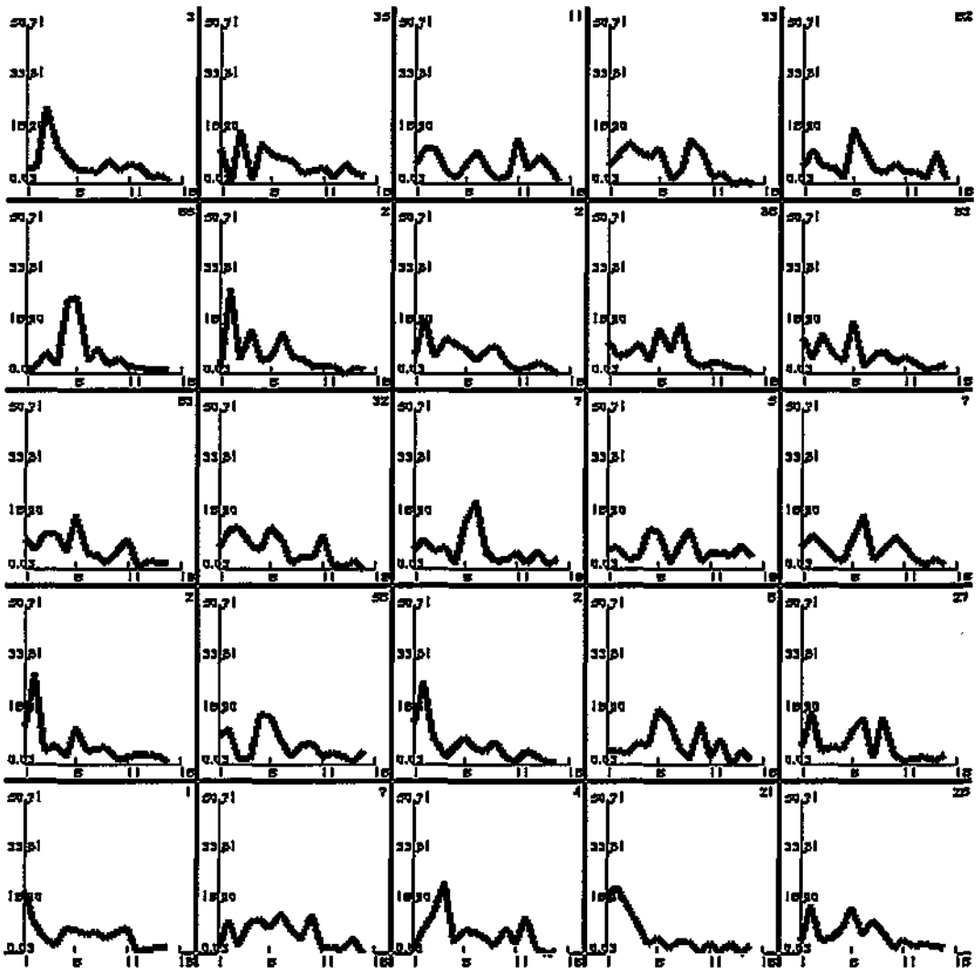


**Figura 8.7** Galería de espectros rotacionales obtenidos a partir de las imágenes de la Helicasa G*40*P.

Como puede observarse en la figura, el análisis de esta población de espectros de la helicasa G*40*P muestra que, efectivamente, existen distintos grupos de partículas caracterizadas por sus diferentes harmónicos predominantes. El mapa puede dividirse a grandes rasgos en seis regiones distintas, señaladas como zonas A, B, C, D, E y F en la figura 8.8. Los dos subgrupos en las esquinas superiores izquierda y derecha del mapa (grupos A y C) son particularmente interesantes, ya que muestran diferencias similares a las descritas para DnaB de *E. coli* [129]. El subgrupo A corresponde a un grupo de imágenes cuya mayor componente rotacional es el armónico de orden 6, mientras que el subgrupo C da cuenta de la existencia de partículas con simetría 3. Estos dos grupos

representan los ya conocidos estados cuaternarios del hexámero de la *G40P* que comparte simetrías de orden 3 y 6 [114, 130].

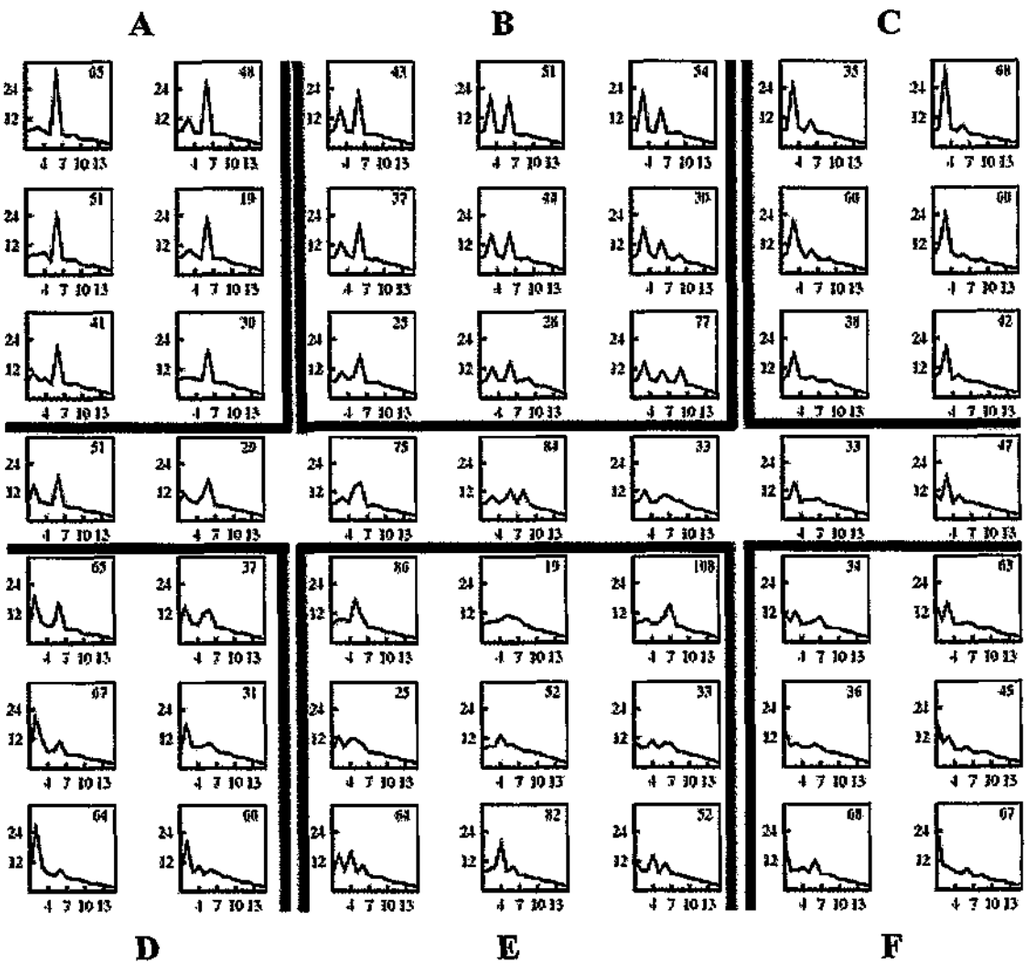



**Figura 8.8** Resultados de KerDenSOM aplicados a espectros rotacionales. Se utilizó un mapa rectangular de 7x7 con una función núcleo Gaussiana. Se aplicaron técnicas de enfriamiento determinista variando el factor de regularización de 100 a 20 en 20 pasos. El número en la esquina superior derecha de cada espectro representa el número de datos originales asignado a cada nodo del mapa.

El mapa auto-organizativo generado por KerDenSOM proporciona, sin embargo, cuatro grupos adicionales de gran importancia como son el grupo B, D, E y F. El grupo B está formado por imágenes con fuertes componentes de orden 6 y orden 3. Nótese que en este caso 3 no es un armónico de 6, lo que induce a pensar que este conjunto de partículas presenta propiedades estructurales diferentes a las del grupo A y C ya que la aparición de esta componente de orden 3 no es esperable en partículas con simetría rotacional de orden 6.

Adicionalmente, el grupo D representa partículas con fuerte simetría de orden 2 con un componente también notable en el armónico 6. La imagen media de las partículas asignadas a este grupo aparece de forma ligeramente elipsoidal, lo cual ocasiona que aparezca este pico de orden 2 en su espectro rotacional, junto con su correspondiente pico en la componente 6. El grupo F encontrado por KerDenSOM representa imágenes sin ninguna simetría claramente apreciable. Estas imágenes pueden representar partículas extremadamente ruidosas sin ninguna interpretación biológica aparente y usualmente son eliminadas de la población.

El grupo E, sin embargo, revela tres nuevas clases significativamente pobladas con partículas que representan fundamentalmente simetrías de orden 4, 5 y 7. Es muy probable que estas nuevas clases de partículas pertenezcan a grupos de partículas que fueron mal alineadas traslacionalmente y que en realidad pertenezcan a grupos de simetría 6 6 3. Usualmente el procedimiento a seguir en estos casos sería repetir los pasos de alineamiento para estos subgrupos de partículas y comprobar posteriormente su orden de simetría.

Al analizar imágenes de proyección de tinción negativa hay que tener en cuenta las limitaciones metodológicas impuestas por la propia técnica. Entre estas limitaciones se encuentran las deformaciones que sufre la proteína y la posibilidad de que esta no esté homogéneamente teñida. Este tipo de fenómenos pueden generar a partir de una partícula intrínsecamente simétrica, imágenes individuales sin una simetría dominante, con un alto grado de heterogeneidad entre sí. Esto es precisamente lo que sucede con el subgrupo F.

Por otro lado, es necesario contemplar la posibilidad de que imágenes de proyección diferentes correspondan a la misma estructura tridimensional observada desde una perspectiva distinta. En las condiciones del presente estudio, las orientaciones que adopta el oligómero de G*40*P sobre el soporte del carbón parecen estar significativamente restringidas, ya que en todos los casos se obtuvieron vistas de la partícula con forma de anillo. A este tipo de vistas les daremos el nombre genérico de vistas frontales. Aunque en las muestras de microscopía es frecuente encontrar vistas preferentes de las macromoléculas, es prácticamente inevitable tener, en lugar de una única orientación, un rango de orientaciones que corresponden a pequeñas variaciones respecto a la orientación preferida. Además de los cambios menores en la orientación de

la partícula sobre el carbón, el propio soporte puede sufrir ligeras deformaciones o no ser perfectamente plano en la escala de la macromolécula. Todos estos factores producirían imágenes de proyección que corresponderían a vistas frontales relativamente inclinadas de la partícula. En el caso que nos ocupa, este tipo de vistas parecen estar representadas por la subpoblación D. El fuerte componente de simetría 2 corresponde a las características que cabría esperar de vistas frontales inclinadas de una partícula con forma de anillo. En consecuencia, de las seis clases de imágenes separadas, ni el subgrupo C ni el D parecen representar estructuras intrínsecamente diferentes del oligómero de G*40*P.

Comparando los resultados obtenidos con KerDenSOM con los obtenidos con el algoritmo clásico de Kohonen [114], las diferencias son notables. El algoritmo de SOM aplicado sobre este mismo conjunto de datos produjo un mapa donde solo se observaban claramente 4 grupos: simetría 3, simetría 6, simetría 2 y falta de simetría (simetría de orden 1), es decir, los resultados obtenidos por SOM muestran muchos menos detalles que los obtenidos por KerDenSOM, mezclando partículas con evidentes diferencias en simetría en los 4 grupos mayoritarios mencionados anteriormente e imposibilitando la detección de patrones de variabilidad minoritarios y sutiles como los mostrados por los grupos B y E encontrados por KerDenSOM. La razón fundamental de esta diferencia de comportamiento entre ambos algoritmos radica en la naturaleza intrínseca del algoritmo de KerDenSOM, que al intentar reproducir fielmente posible la densidad de probabilidad de los datos de entrada es capaz de detectar las pequeñas pero todavía significativas fuentes de variaciones. Así mismo, KerDenSOM ofrece un mejor control del proceso de proyección a través del control de la suavidad del mapa obtenido, esto permite obtener mapas, que sin dejar de ser suaves y ordenados, ofrezcan una mayor preservación topológica de los datos de entrada.

### 8.3.3. Clasificación de imágenes

Hasta ahora sólo se han detallado aspectos relacionados con la simetría de las imágenes. No obstante, diferentes tipos de imágenes pueden compartir simetrías similares aun siendo marcadamente diferentes. Estas imágenes no se separarán, por tanto, si únicamente se consideran sus espectros rotacionales. Es necesario, entonces, examinar paralelamente otro tipo de heterogeneidades, lo que hace necesario realizar el

análisis de clasificación utilizando como entrada directamente las imágenes de las partículas. Este era el caso del grupo B obtenido mediante la clasificación de espectros rotacionales, el cual mostraba un interés particular por estar compuesto por partículas con una mezcla poco común de simetrías (componentes significativos en 6 y 3 simultáneamente). Un análisis preliminar utilizando SOM detectó que este conjunto estaba en realidad constituido por una mezcla de partículas de forma similar pero de quiralidad opuesta [114]. Las diferencias observadas resultaban más acusadas en la zona externa de las partículas y, por tanto, con el objetivo de conseguir una mejor separación de las clases, se aplicó una máscara binaria para extraer los pixeles de interés solo en una corona circular externa de las imágenes. La figura 8.9 muestra la máscara tipo corona aplicada sobre la imagen media de estas partículas.

**Figura 8.9** Máscara tipo corona utilizada para seleccionar el área de interés en la clasificación de imágenes de la helicasa G*40*P.

Una vez aplicada la máscara al conjunto de imágenes formadas por este grupo de simetría 3 y 6, se obtuvo un conjunto de 338 vectores de dimensión 780 (338 imágenes de 780 pixeles dentro de la corona seleccionada). Este nuevo conjunto de datos, caracterizados por su alta dimensionalidad y muy baja relación señal/ruido, se utilizó como datos de prueba de los algoritmos propuestos en esta memoria con el objetivo de comprobar su robustez y tolerancia en la detección de sutiles heterogeneidades en condiciones extremas.

#### *8.3.3.1. Aplicación del algoritmo clásico de SOM*

La figura 8.10 muestra los resultados obtenidos por Barcena y colaboradores [114] al aplicar el algoritmo clásico de SOM al conjunto de imágenes pertenecientes al grupo de partículas con mezcla de simetría 3 y 6. El mapa ha sido dividido en dos áreas que constituyen el grupo de los vectores diccionarios en los que puede observarse la

quiralidad opuesta. Ambas clases fueron alineadas independientemente para obtener las imágenes medias de proyección que se presenta en la misma figura.

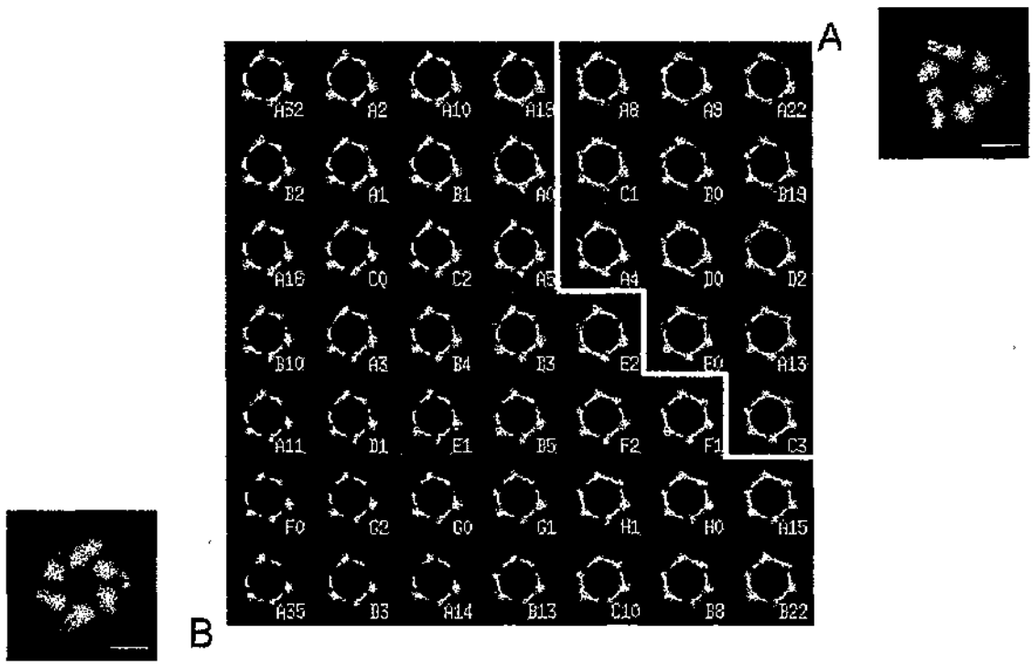



**Figura 8.10** Resultados de aplicar el algoritmo clásico de SOM sobre el conjunto de imágenes de la helicasa hexamérica G*40*P. El mapa utilizado es de 7x7 neuronas y ha sido dividido manualmente en dos grandes grupos de partículas divididos por una línea blanca. A cada lado del mapa se muestra la imagen media de las partículas asignadas a cada grupo. El número de imágenes asignadas a cada vector diccionario también es mostrado en la esquina inferior derecha de cada nodo.

Según el análisis de las medias obtenidas, es interesante examinar las dos clases de imágenes halladas dentro de esta subpoblación. Con la resolución alcanzada, la única diferencia significativa entre ambos grupos de imágenes parece ser la de su opuesta quiralidad. Esto evidencia que estas clases corresponden a vistas frontales desde caras opuestas de un mismo tipo de arquitectura macromolecular. Así pues, las dos clases finales en realidad representarían únicamente dos tipos de estructuras distintas del hexámero de G*40*P.

### *8.3.3.2. Aplicación del algoritmo Kernel c-means*

Con el objetivo de intentar reproducir los resultados descritos anteriormente utilizando la técnica de SOM clásica [114], se procedió a ejecutar el algoritmo de agrupamiento Kernel c-means (KCM) para obtener dos grupos distintos. Era de esperar que este método reprodujera fielmente los mismos resultados observados por Barcena

[114] detectando básicamente estos dos grandes grupos de partículas, mostradas en la figura 8.10. Sin embargo, los resultados obtenidos por KCM distaban de ser los mismos que los obtenidos por SOM. La figura 8.11 muestra el resultado del agrupamiento para dos grupos.

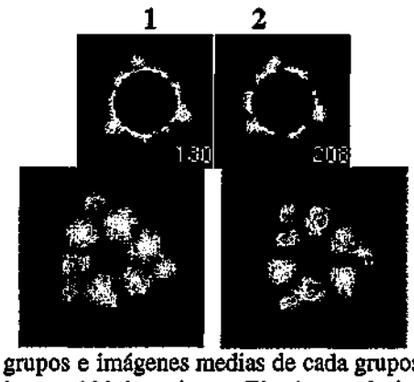



**Figura 8.11** Centros de grupos e imágenes medias de cada grupos obtenidos por KCM. Se utilizó un núcleo Gaussiano y 200 iteraciones. El número de imágenes asignadas a cada vector diccionario aparece en la esquina inferior derecha de cada nodo.

La figura 8.11 evidencia que los resultados obtenidos por SOM no han podido reproducirse adecuadamente utilizando dos grupos en KCM. A pesar de que el grupo 2 muestra claramente partículas con orientación a favor de las manecillas del reloj, la imagen media del grupo 1 parece estar compuesta por una mezcla de partículas con diferente quiralidad que carece de explicación biológica. Esto hizo suponer que esta dos grupos no eran suficiente para explicar toda la variabilidad presente en esa población, por lo que se repitió el experimentos utilizando ahora tres grupos. La figura 8.12 muestra los resultados de este nuevo agrupamiento.

Analizando los resultados del nuevo agrupamiento en tres grupos es evidente que KCM correctamente separó las partículas con diferente quiralidad. El grupo 1 en la figura 8.12 representa al conjunto de imágenes con orientación en contra de las manecillas del reloj, resultado también obtenido por SOM. Sin embargo, KCM ha necesitado dos grupos (2 y 3) para representar la totalidad de partículas con orientación a favor de las manecillas del reloj. La pregunta que inmediatamente surge es, por qué se han necesitado 3 grupos para reproducir los mismo resultados obtenidos por SOM? La respuesta puede encontrarse analizando las partículas asignadas a cada uno de estos tres grupos obtenidos por KCM. La figura 8.12b muestra las imágenes medias de cada grupo y sus correspondientes espectros rotacionales. El grupo1 muestra una clara orientación en contra de las manecillas del reloj que está en plena concordancia con los resultados

obtenidos por SOM. Los grupos 2 y 3, sin embargo, a pesar de mostrar una clara orientación a favor de las manecillas del reloj tal y como se observó en SOM también muestran una diferencia sutil pero significativa en cuanto a su simetría: ambas presentan una componente predominante de orden 6, pero el grupo 3 a diferencia del grupo 2 está influenciado por una componente significativa de orden 3. Estas pequeñas diferencias en cuanto a simetría no se detectaron visualmente con SOM (figura 8.10) y probablemente explican el por qué KCM no fue capaz de detectar variaciones de quiralidad utilizando solo dos grupos, debido a que existen en realidad dos grandes fuentes de variación en estos datos: simetría y quiralidad. KCM, sin embargo fue capaz de detectarlas cuando se utilizó tres grupos, demostrando su eficiencia para detectar pequeñas variaciones en condiciones extremas de alta dimensionalidad y alto nivel de ruido cuando se especifica a priori el número de grupos.

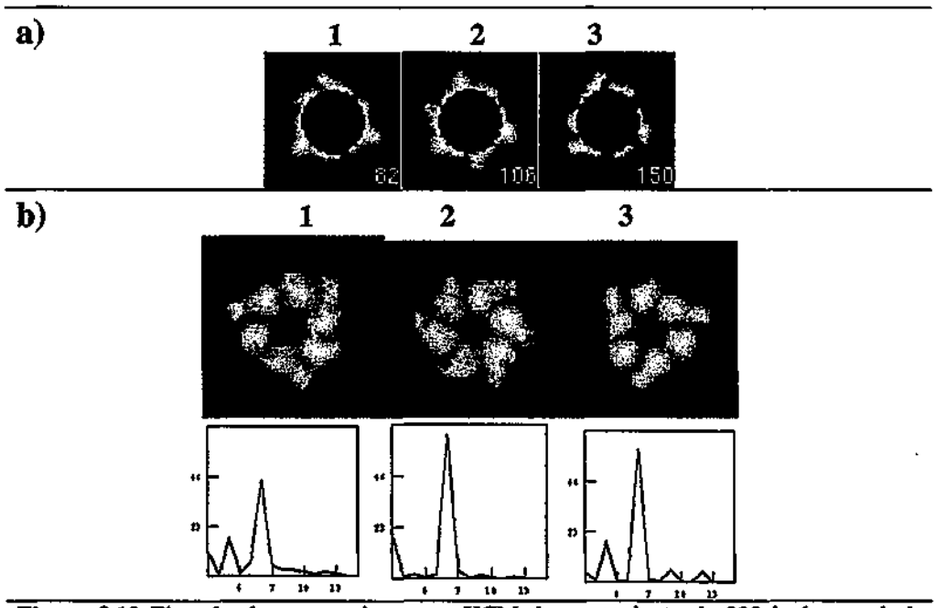



**Figura 8.12** Ejemplo de agrupamiento por KCM de un conjunto de 338 imágenes de la helicasa hexamérica G40P del bateriófago SPP1. a) Resultados del algoritmo utilizando 3 grupos. El número en la esquina inferior derecha representa el número de imágenes asociadas a cada grupo. b) Imágenes medias de las partículas asignadas a cada grupo y su correspondiente espectro rotacional.

#### *8.3.3.3. Aplicación del algoritmo KerDenSOM*

En esta sesión expondremos como el algoritmo de KerDenSOM es capaz de detectar también las pequeñas variaciones observadas utilizando KCM. La razón

principal de intentar utilizar un mapa auto-organizativo está motivada por el hecho de que, a pesar de que KCM es capaz de encontrar heterogeneidades importantes en este tipo de datos característicos de la microscopía electrónica, presenta una gran desventaja práctica y es que se necesita saber exactamente el número de grupos presentes en el conjunto de datos originales. Esta condición hace que este tipo de métodos particionales sean poco utilizados en la práctica en situaciones donde no se conoce a priori las características de los datos que estamos analizando. Es por eso que un método eficiente y robusto que ayude a la exploración de datos desconocidos es más que necesario.

KerDenSOM puede ser clasificado como uno de estos métodos, debido principalmente al hecho de que, al igual que SOM, no es necesario prefijar anticipadamente el número de clases a extraer. Si bien es cierto que el tamaño (número de vectores diccionarios) y la topología del mapa puede influenciar el número de grupos a extraer, este parámetro no es tan crítico como lo es el número de grupos en un algoritmo de agrupamiento particional como KCM. Aunque desafortunadamente no existen reglas adecuadas para seleccionar un tamaño de mapa, se debe escoger un tamaño no muy pequeño o se corre el riesgo de que el algoritmo no sea capaz de acomodar toda la varianza de los datos. En este sentido KerDenSOM es mucho mas flexible que KCM. La figura 8.13 muestra los resultados de aplicar este algoritmo al mismo conjunto de datos utilizando un mapa de 10x5 vectores diccionarios organizados en una topología rectangular.

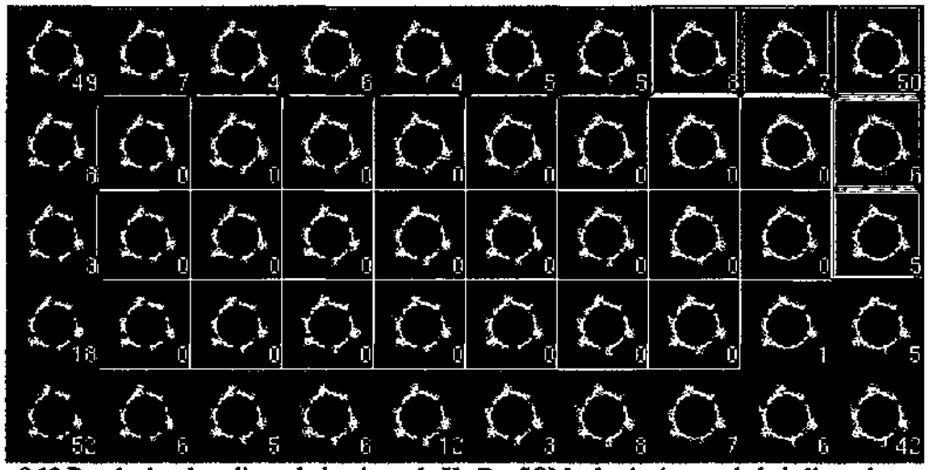



**Figura 8.13** Resultados de aplicar el algoritmo de KerDenSOM a las imágenes de la helicasa hexamérica G*40*P del bateriófago SPP1. Se utilizó un núcleo Gaussiano y 5 pasos de enfriamiento determinista variando el parámetro de suavidad desde 2000 hasta 200. El mapa ha sido separado manualmente en 3 grupos diferenciados por los colores verde, rojo y azul. El número en la esquina inferior derecha representa el número de imágenes asociadas a cada nodo.

En este caso se utilizó un núcleo gaussiano y el algoritmo se ejecutó en 5 pasos de enfriamiento determinista variando el parámetro de suavidad desde 2000 hasta 200.

Como se puede observar en la figura 8.13, el algoritmo de KerDenSOM ha sido capaz de encontrar 3 grupos distintos de partículas distribuidas alrededor de los bordes del mapa. Los vectores diccionarios pertenecientes a cada grupo han sido marcados en el mapa como pertenecientes a los grupos 1, 2 y 3 que concuerdan perfectamente con los obtenidos por el algoritmo de KCM mostrados en la sesión anterior. Si bien es cierto que la selección de estos subconjuntos en el mapa se realiza de manera manual, es también cierto que existen evidencias sólidas y objetivas para hacerlo. En primer lugar, y como se ha mostrado en esta memoria, el algoritmo de KerDenSOM intenta obtener un conjunto de vectores diccionarios que reflejen de la manera más fielmente posible la distribución de densidad estadística de los datos. A nivel práctico esto puede traducirse en que aquellos vectores diccionarios que presenten una mayor densidad (número de imágenes asociadas a él) son grandes candidatos a ser centros de grupos. En el caso de la figura 8.13, se evidencia la naturaleza de este tipo de mapa auto-organizativo: vectores diccionarios caracterizados por poseer una alta densidad y que se diferencian gradualmente de sus vecinos de manera suave pasando por zonas de baja densidad (áreas con pocos o ningún dato asociado). De esta forma es relativamente sencillo separar el mapa en diferentes grupos no solo teniendo en cuenta la apariencia del vector diccionario, sino también su valor de densidad. Por ejemplo, el grupo 1 que se corresponde fielmente al grupo 1 obtenido por KCM muestra una evidente orientación en contra de las manecillas del reloj y posee su máximo de densidad en el vector diccionario ubicado justo en la esquina superior derecha del mapa. Similarmente, los grupos 2 y 3 han sido claramente diferenciados no solo por su apariencia sino por la zona central de baja densidad (sin datos asociados) que los separa. Esta zona central corresponde a un área de transición entre estos dos grupos.

Este efecto tan evidente no ha sido observado por SOM y es una de las características que hacen de KerDenSOM un algoritmo robusto y eficiente para la clasificación de este tipo de datos. Las propiedades de preservación de la densidad de probabilidad que posee este método, combinado con las propiedades de proyección suave y ordenada características de los mapas auto-organizativos, lo hacen una

herramienta interesante e importante para resolver los problemas de clasificación de partículas individuales en microscopía electrónica.

## 8.4. Aplicación a imágenes del Antígeno T del virus SV40

En este apartado mostraremos una nueva aplicación de los métodos de clasificación sobre imágenes de proyección de partículas de otro espécimen biológico de especial relevancia: el Antígeno T del virus SV40. Esta pfoteína está estrechamente relacionada con la proliferación de células cancerígenas y su estudio, tanto bioquímico como estructural, es de vital importancia para entender los complejos procesos biológicos asociados a esta mortal enfermedad. En las secciones siguientes presentaremos una breve descripción de esta proteína así como los estudios de heterogeneidad estructural llevados a cabo con el algoritmo KerDenSOM, objeto de esta memoria.

### 8.4.1. Información general acerca del Antígeno T del Virus SV40: Su funcionalidad y relevancia.

La perpetuación de todos los seres vivos requiere de un proceso fundamental cual es la duplicación del material genético parental, que constituirá la dotación genética de la progenie. Este proceso recibe el nombre de replicación del ADN. En las células eucariotas la replicación del ADN ocurre (y debe ocurrir solamente) una vez por ciclo celular. Para duplicar el genoma de forma eficiente se requiere la adecuada coordinación de, por un lado, las proteínas implicadas en la replicación en la propia célula, y, por otro lado, de la replicación con otros procesos celulares (como la mitosis y la citocinesis), y con la replicación del ADN de las células vecinas. El modelo experimental más utilizado para el estudio de la replicación de la cromatina y del ADN de mamíferos es el del virus SV40 [131], también se emplea para los estudios del desarrollo tumoral y la regulación del ciclo celular.

El virus SV40 (del inglés Simian Virus 40) es un virus de la familia polioma que se identificó por vez primera durante la década de los años 50 durante los ensayos que culminaron con el desarrollo de una vacuna eficaz frente al virus de la poliomielitis humana. SV40 produce enfermedades diversas en monos, induce tumores en roedores e infecta a una gran variedad de células de mamíferos, aunque la infección solo es productiva en primates. En células de mamíferos diferentes a los primates la infección o

es abortiva o conduce a la inmortalización de la línea celular [132]. La replicación del ADN de SV40 se puede reconstituir in vitro con sólo una proteína de origen viral y diez proteínas provenientes de la célula infectada [133, 134]. De entre los componentes de SV40 hay una proteína, localizada en la cápsida viral, que, como se detallará más adelante, cobra especial relevancia: el denominado antígeno de tumorogenicidad, abreviado T-Ag, la única proteína de origen viral necesaria para la replicación del cromosoma de SV40 [135-137].

El T-Ag es una fosfoproteína que desempeña múltiples funciones. Aparte de en la replicación del ADN de SV40 también participa en la regulación del ciclo infectivo y en la estimulación de la proliferación celular y el control del ciclo celular. Para ello es capaz de interaccionar con una gran diversidad de ligandos, desde nucleótidos y ácidos nucleicos hasta proteínas celulares, entre ellas la proteína supresora de tumores p53 (un factor de transcripción crítico en los mecanismos celulares que responden a condiciones de estrés genotóxico mediante la detención del progreso del ciclo celular o la inducción de apoptosis).

En la replicación del ADN el T-Ag actúa como iniciador de la replicación, mediante el reconocimiento del origen de replicación viral y unión específica a esta región del cromosoma de SV40, y como helicasa, una actividad enzimática que cataliza el desenrollamiento de la doble hélice del ADN, acontecimiento indispensable para que el resto de las proteínas de la maquinaria de replicación puedan acceder a la hebra de ADN que ha de ser copiada y ejercer su labor de síntesis de las cadenas de ADN hijas.

### 8.4.2. Estudios estructurales de los hexámeros del T-Ag en el origen de replicación viral.

Los complejos macromoleculares grandes, y los que se forman durante la replicación del ADN lo son, poseen unas características de flexibilidad y tamaño que dificultan considerablemente, cuando no imposibilitan, su análisis estructural mediante técnicas resolutivas como la cristalografía de rayos X o la espectroscopia de resonancia magnética nuclear. La microscopía electrónica tridimensional de especimenes embebidos en hielo vítreo (crioEM), que proporciona unos mapas de densidad electrónica obtenidos a una resolución media, constituye una alternativa sumamente adecuada para el estudio estructural de los complejos anteriormente mencionados. Los mapas de densidad electrónica se pueden complementar, mediante las denominadas

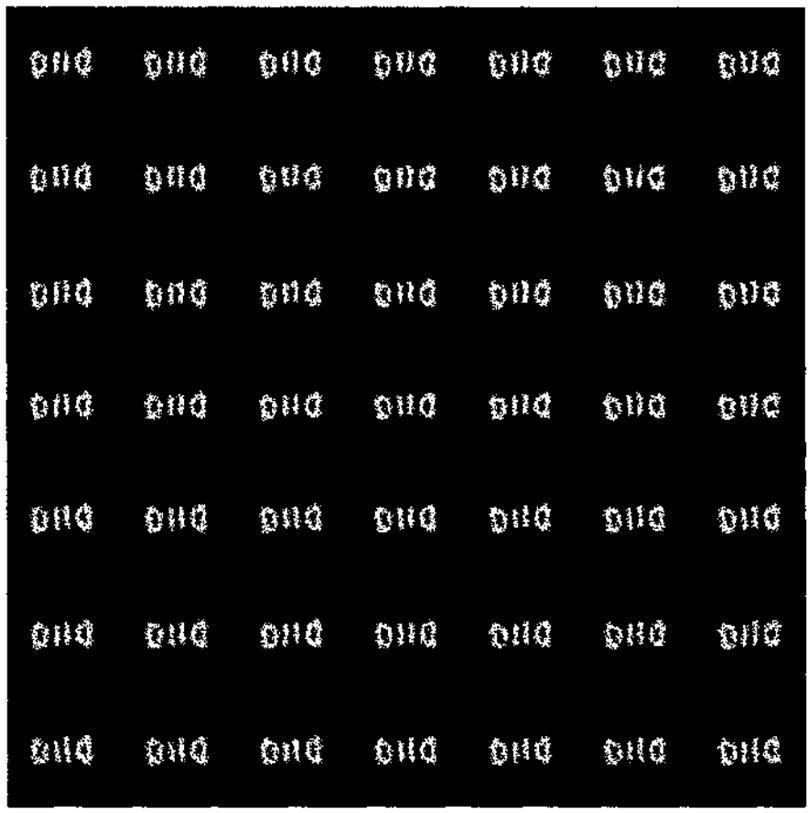
técnicas de multirresolución, con datos provenientes de otras técnicas de elucidación estructural y obtenidos a un nivel de resolución atómica, lo que permite conseguir una visión de conjunto, pero extremadamente rica en detalles, del espécimen objeto de estudio.

La crioEM sigue una aproximación metodológica idéntica a la de la tomografía médica y se basa en el promediado y la combinación de miles de imágenes de proyección del espécimen objeto de estudio obtenidas en el microscopio electrónico. El cálculo de una reconstrucción tridimensional veraz y ajustada precisa, pues, de un conjunto homogéneo de imágenes iniciales. Las heterogeneidades pueden ser de carácter extrínseco, que la muestra sea en realidad una mezcla de componentes de diferente composición química, o intrínseco, que una única muestra de lugar a distintas imágenes de proyección. Sea como fuere, la detección de estas heterogeneidades, y su posterior clasificación en grupos, es sumamente crítica en los procesos de reconstrucción tridimensional tal y como hemos visto en apartados anteriores.

Para nuestros estudios de reconstrucción tridimensional de los dobles hexámeros del T-Ag ensamblado sobre el origen de replicación viral ha sido necesario el empleo de unos complejos nucleoproteícos (cuyas características no procede detallar) que preveíamos exhibiesen ciertas heterogeneidades, provenientes de la propia preparación de la muestra, de complicada detección. Constatamos que esto era así cuando haciendo uso de los algoritmos de SOM clásico al comienzo de nuestros estudios y tras la separación de las imágenes de criomicroscopía iniciales en grupos presuntamente homogéneos (ver figura 8.14a) se obtuvo la reconstrucción tridimensional que se muestra en la figura 8.14b de aspecto completamente artefactual tras simple inspección visual.

Debido a lo anteriormente expuesto, nos propusimos el estudio de la variabilidad estructural de estas imágenes utilizando el algoritmo de KerDenSOM. Para ello tomamos unas 3022 partículas de las micrografías electrónicas obtenidas por crioEM. Las partículas fueron previamente alineadas traslacional y rotacionalmente antes del análisis. En este proceso de alineamiento, 200 imágenes fueron descartadas por su imposibilidad de ser alineadas correctamente, indicando que pertenecen a imágenes de ruidos donde no aparece información estructural útil.

a)



b)

**Figura 8.14** Clasificación de las partículas del T-Ag utilizando el algoritmo clásico de Kohonen. a) Mapa auto-organizativo de 7x7. b) Reconstrucción tridimensional.

Para el análisis de heterogeneidad se utilizó un mapa de 10x7 de topología rectangular con un núcleo Gaussiano. El algoritmo se ejecutó en 300 iteraciones variando el factor de suavidad desde 300 hasta 250 en 20 pasos de enfriamiento determinista. La figura 8.15 muestra un esquema del proceso.

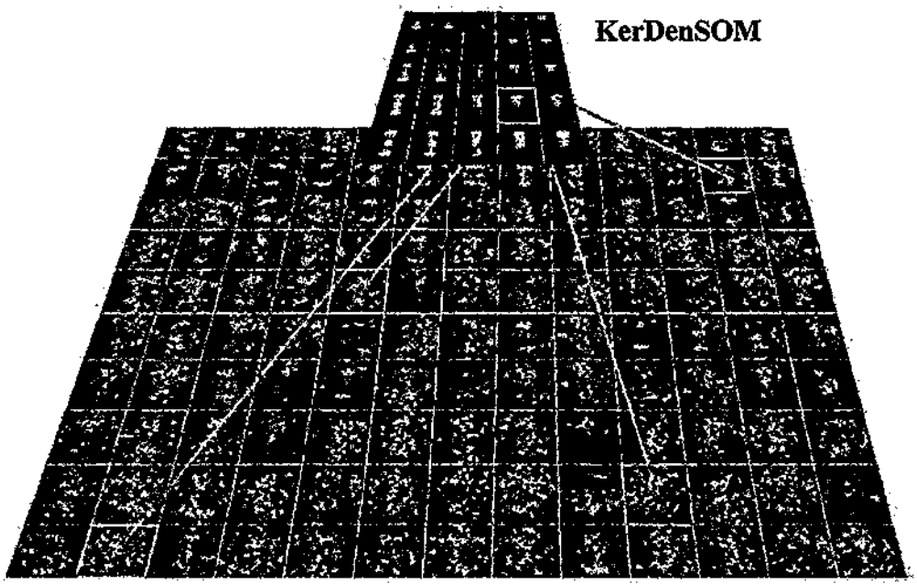



**Figura 8.15** Representación esquemática del proceso de clasificación por KerDenSOM de las imágenes del Antígeno T del virus SV40. Las imágenes originales (imagen inferior) son "proyectadas" en el mapa auto-organizativo (parte superior de la imagen). Las imágenes marcadas muestran la naturaleza del proceso de proyección donde un vector diccionario del mapa representa un conjunto de imágenes originales.

La aplicación del nuevo algoritmo de KerDenSOM condujo a la detección de diferencias tan sutiles como relevantes en las imágenes de proyección que anteriormente habían pasado desapercibidas, ilustradas en la figura 8.16, siendo posible además encontrar una explicación para estas diferencias así como inferir una estrategia de clasificación, respetuosa con ciertas particularidades de tipo bioquímico acerca de la interacción ADN-T-Ag en el origen de replicación viral conocidas de antemano.

Una inspección visual del mapa de la figura 8.16 revela la existencia de distintos conjuntos de partículas con diferencias estructurales significativas. El grupo marcado como A en la figura representa partículas que poseen fuertes variaciones de la curvatura axial a lo largo de su eje longitudinal y que están formadas aparentemente por tres grupos de masas no muy bien definidas y en las cuales la interfaz hexámero-hexámero del Antígeno no es observada.

Este tipo de partículas fueron también observadas previamente en estudios realizados con técnicas de tinción negativa [138] y representan partículas que se desvían

del complejo macromolecular canónico debido a diferentes razones: la flexibilidad estructural del dominio de unión con el ADN, plegamiento local del fragmento de ADN, desensamblaje parcial de la partícula o una combinación de las tres causas.

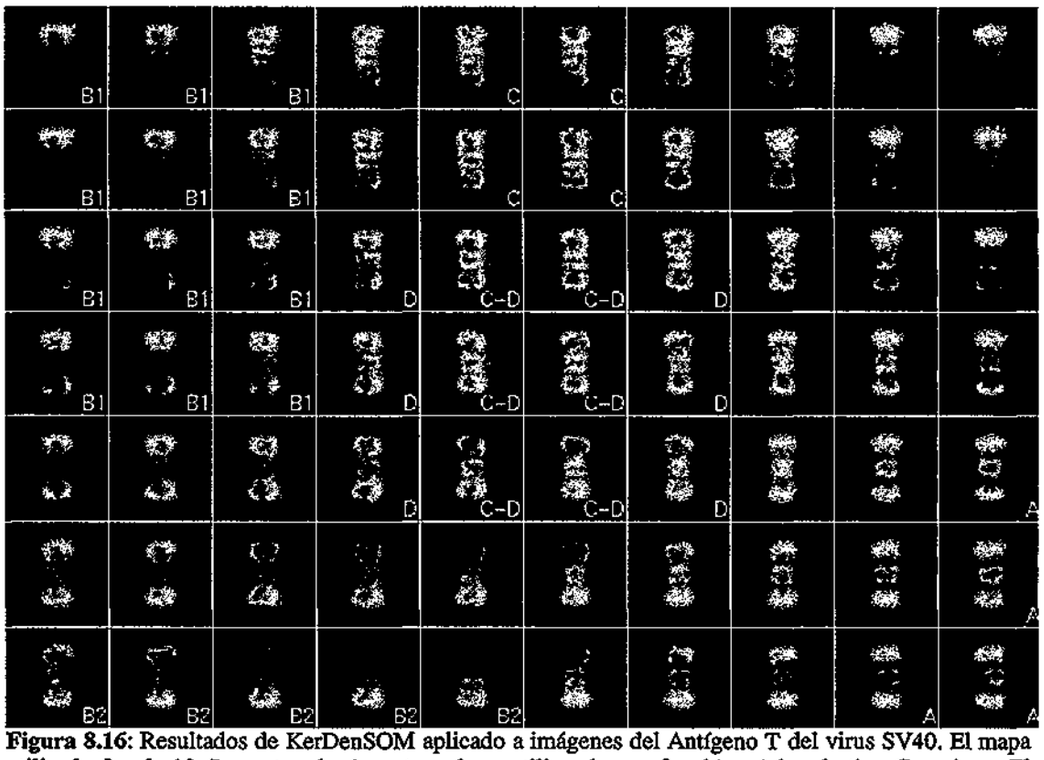



**Figura 8.16**: Resultados de KerDenSOM aplicado a imágenes del Antígeno T del virus SV40. El mapa utilizado fue de 10x7 con topología rectangular y utilizando una función núcleo de tipo Gaussiano. El algoritmo se calculó en 300 iteraciones variando el parámetro de suavidad desde 300 a 250 en 20 pasos de enfriamiento determinista.

En la esquina superior e inferior izquierda del mapa se puede observar otro conjunto de imágenes dividido en dos subgrupos de partículas similares pero con fuertes diferencias estructurales relacionadas con una rotación de 180° con respecto a su eje longitudinal principal (grupos B1 y B2 en la figura 8.16). Estos dos grupos representan complejos macromoleculares dodecamericos con dos mitades significativamente diferentes, una de ellas más brillante que la otra. Adicionalmente, se puede apreciar que los hexámeros correspondientes a la parte superior e inferior de estas partículas no solo difieren por sus valores de densidad, sino por sutiles variaciones estructurales que aparecen en la región más ancha de los hexámeros: en algunos casos aparece una disminución significativa de densidad en forma de cavidad y en otros esta característica no se observa.

Las posibles razones que explican estas variaciones estructurales son bastante complejas e involucran diferencias en la masa total entre ambos hexámeros debido a la sonda de ADN utilizada así como posibles rotaciones entre los dos hexámeros debido a la posición en que estas partículas se depositaron en la rejilla, que por algún motivo físico, corresponden a vistas preferentes en distintas orientaciones. Por ejemplo, la cola del fragmento de ADN no cubierto por el doble hexámero en uno de los extremos del complejo, podría ser una de estas causas debido a que esta pequeña porción de ADN podría adherirse de manera no específica alrededor de la superficie externa de uno de los hexámeros en los extremos del complejo, provocando de esta forma una diferencia significativa de la masa detectada en las imágenes de proyección.

Un tercer grupo de partículas, marcado como grupo C en la figura 8.16 también ha sido detectado por KerDenSOM. Este grupo se caracteriza por presentar una interfaz hexámero-hexámero muy claramente diferenciable a la vez que los hexámeros en los extremos de la partícula presentan una apariencia prácticamente idéntica. Estas características indican que en este caso los hexámeros no están rotados uno con respecto al otro.

Hay un cuarto grupo de imágenes que son idénticas a las del grupo anterior con la salvedad de presentar además una pequeña masa de densidad, que se localiza únicamente en uno de los laterales del complejo (vectores diccionarios indicados con D en la figura 8.16). Esta masa adicional parece provenir de la interfaz inter-hexámeros. En trabajos anteriores [135] se ha propuesto que la unión del T-Ag a la región conocida como del palíndromo temprano (EP, una de las tres zonas que conforman el origen de replicación viral) induce el despareamiento local de la doble hebra del ADN en ausencia de la hidrólisis de ATP (es decir, sin necesidad de un aporte energético). Este proceso conduciría a la aparición de una hebra sencilla de ADN desplazada, que presumiblemente se colocaría fuera del hexámero del T-Ag, mientras que la otra cadena de ADN permanecería unida a la cara interna del otro oligómero de T-Ag. La cadena desplazada abandonaría, pues, el complejo a través de un punto fijo, la región situada entre los dos hexámeros, mientras que el extremo libre de esta hebra desplazada disfrutaría de una mayor libertad de movimiento. Así pues, la masa de densidad adicional que se observa en el subgrupo D encajaría plenamente dentro de la descripción anterior.

Como conclusión podemos señalar que el análisis llevado a cabo con imágenes del Antígeno T del virus SV40 ha mostrado diferencias estructurales poco evidentes en este conjunto de partículas tan ruidosas, especialmente diferencias relacionadas con la simetría y la composición estructural de los hexámeros en ambos extremos de este complejo macromolecular así como sutiles factores de posición (rotación e inclinación) que provocan las variaciones estructurales principales de este conjunto de partículas. Este análisis ha permitido desarrollar nuevas hipótesis sobre el significado biológico de las variaciones detectadas permitiendo a su vez un estudio más amplio de esta estructura.

# 9. Clasificación de volúmenes de Tomografía Electrónica

El objetivo principal de este capítulo es presentar una aplicación del algoritmo de KerDenSOM en el contexto de clasificación de estructuras tridimensionales obtenidas mediante tomografía electrónica. La tomografía electrónica es una metodología muy poderosa para determinar arquitecturas complejas de especimenes biológicos. Sin embargo, en muchos casos, cuando la estructura estudiada está compuesta por un conjunto extenso de especimenes individuales, comienzan a aparecer patrones estructurales que necesitan ser estudiados individualmente para lograr una mejor comprensión del espécimen biológico y en algunos casos para aumentar por técnicas de promediado de estructuras similares la calidad de la reconstrucción. Este tipo de aplicación difiere fundamentalmente de la clasificación de partículas individuales en EM presentada en la sección anterior, en que las estructuras bajo estudio que la componen son imágenes tridimensionales, lo cual aumenta considerablemente la complejidad del problema. En esta sección presentaremos una breve descripción de la tomografía electrónica y presentaremos un caso particular de estudio de ciertas estructuras tridimensionales relacionadas con el proceso contracción muscular presente en algunos tipos de tejidos de los animales.

## 9.1. Breve introducción a la tomografía electrónica

La Tomografía Electrónica se define como cualquier técnica que emplee el microscopio electrónico para obtener proyecciones de un objeto con el objetivo de obtener la reconstrucción tridimensional de dicho objeto [139]. En la sección anterior comentamos el análisis de un conjunto de imágenes de proyecciones de un agregado macromolecular como paso previo para el proceso de reconstrucción tridimensional de su estructura. Este análisis es conocido como análisis de partículas individuales y se basa fundamentalmente en la utilización de un número elevado de proyecciones provenientes de muchas partículas del mismo espécimen con el objetivo de aumentar la resolución de la reconstrucción a través del efecto de promediado de partículas similares. El término de partículas individuales se presta a confusión ya que precisamente la reconstrucción final de la estructura macromolecular no proviene de una partícula individual, sino de miles de ellas.

Sin embargo, cuando se quiere realizar la reconstrucción tridimensional de estructuras individuales mucho más grandes, gruesas ó donde no sea posible la recolección de múltiples copias del mismo espécimen, el proceso de recolección de proyecciones previo a la reconstrucción tridimensional difiere del usualmente utilizado en análisis de partículas individuales. Este es el caso de estudio no solamente de estructuras asiladas sino también de células completas, secciones de tejidos ó estructuras polimorfas como las mitocondrias u otros orgánulos. En estos casos la metodología utilizada es la recolección de proyecciones de la estructura bajo estudio en el rango más amplio posible de ángulos de inclinación con incrementos lo más pequeños posibles. Esta técnica está muy relacionada con las utilizadas en tomografía axial computarizada en medicina [140].

La figura 9.1 describe el principio básico de la tomografía electrónica. La figura 9.1a muestra el esquema de adquisición de serie de eje único (single axis tilt series) donde una única rejilla se inclina sucesivamente sobre un eje tomando en cada paso una micrografía del mismo espécimen. La figura 9.1b muestra un esquema muy simplificado del proceso de reconstrucción tridimensional del objeto original a partir de sus imágenes de proyecciones. Esta reconstrucción se realiza reproyectando cada proyección de manera que la suma de todas reconstruye el objeto original.

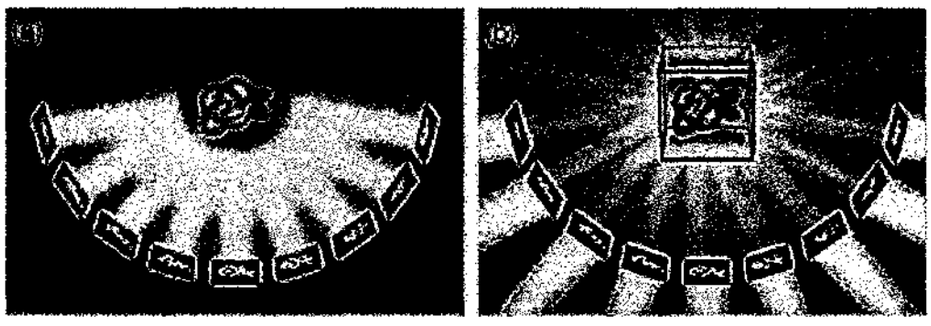


**Figura 9.1** a) Esquema de adquisición de serie de eje único (single axis tilt series) b) Esquema simplificado del proceso de reconstrucción tridimensional del objeto original a partir de sus imágenes de proyecciones.

Esta metodología, sin embargo, presenta un problema intrínseco que genera una situación contradictoria: por una parte, para obtener reconstrucciones tridimensionales a alta resolución se necesita tener la mayor cantidad de proyecciones posibles, esto implica que se deben tomar proyecciones en el mayor rango angular posible con el

mayor número de incrementos posible, pero al mismo tiempo se tiene que garantizar que la dosis de voltaje total suministrada a la estructura no sea tan alta que dañe su composición ni tan baja que se pierdan los detalles importantes [141, 142]. El grosor de las estructuras constituye también un límite para este tipo de técnicas y depende fundamentalmente del voltaje de aceleración del microscopio electrónico utilizado. Independientemente de los problemas encontrados en tomografía electrónica y la baja resolución obtenida debido a ellos (20-70 Ángstrom), esta técnica es ampliamente utilizada, especialmente en el análisis de células completas y tejidos donde la información producida en el tomograma es enorme [143].

## 9.2. Un caso de estudio: Músculo de vuelo de un insecto

Los músculos voluntarios, también conocidos como esqueléticos o estriados, son uno de los tres tipos de músculos presentes en los mamíferos al igual que lo son los músculos involuntarios y los músculos cardiacos. Cada una de estas categorías de músculos cumple una función muy importante en los organismos: los músculos cardiacos son los responsables de los complejos movimientos producidos en el corazón, los músculos involuntarios o lisos son los que intervienen en el alineamiento de las paredes de las arterias para controlar la presión arterial, o controlar la digestión de los alimentos a través del movimiento del intestino y por último, los músculos estriados (voluntarios) son los responsables del movimiento, la locomoción ó el vuelo.

Este último tipo de músculo en seres humanos y en otros animales está bajo control directo del sistema nervioso central por lo que se le denomina músculo voluntario (controlado). Debido a su importancia en la comprensión de los mecanismos de contracción muscular, este tipo de músculos ha sido y es objeto de muchos estudios. La figura 9.2 muestra una micrografía tomada por microscopía electrónica de un músculo estriado humano. El tejido del músculo está compuesto por paquetes de células llamadas fibras musculares dentro de las cuales se encuentran las miofibrillas, que también aparecen distribuidas en paquetes. El término estriado viene dado por la apariencia en forma de rayas o estrías provocadas por la disposición de las zonas claras y oscuras de los sarcómeros que son elementos que se repiten a lo largo de las miofibrillas y que forman la unidad estructural y funcional de las células musculares estriadas.

El análisis de la estructura y composición molecular del sarcómero (figura 9.3), permite entender el mecanismo de contracción de las fibras musculares estriadas, basado en el deslizamiento de los filamentos gruesos sobre los filamentos finos. Los filamentos gruesos (de 15 nm de ancho y 1.6 μm de largo) están formados principalmente por miosina y se localizan a lo largo de la banda A (figura 9.3). Los filamentos finos (de 8 nm de ancho y 1.0 μm de largo) corresponden a microfilamentos de F-actina. Estos anclan en la línea Z, luego cursan a lo largo de la banda I y penetran la banda A, donde corren paralelos a los filamentos gruesos, terminando a nivel de la banda H que contiene sólo filamentos gruesos. En la banda A se observan puentes que se extienden desde los filamentos gruesos hacia los filamentos finos y que corresponden a las cabezas de las moléculas de miosina. A nivel de la línea M, cada filamento grueso se asocia a 6 filamentos gruesos adyacentes, a través de puentes proteicos dispuestos radialmente. Durante el proceso de contracción muscular, los filamentos finos de los sarcómeros adyacentes son empujados hacia el centro de la banda A, lo que produce el acortamiento del sarcómero. Como consecuencia de este proceso, se oblitera la banda H y disminuye la longitud de la banda I, sin que se modifique la longitud de la banda A. El grado de solapamiento entre filamentos gruesos y finos explica este fenómeno.

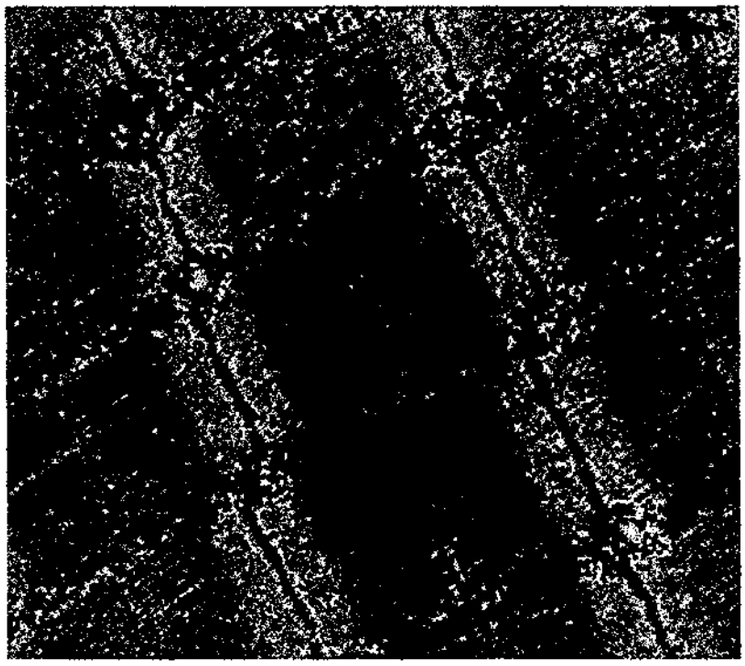


**Figura 9.2** Micrografía electrónica de un músculo estriado humano.

Uno de los objetos de estudio basados en el tipo de células musculares como las descritas anteriormente es entender el mecanismo molecular de producción de fuerzas en el músculo a través de la visualización tridimensional de los estados de los puentes que son las cabezas enzimáticas de la miosina que actúa como motor molecular produciendo la fuerza durante la contracción muscular. El estudio de la estructura de estos puentes en diferentes condiciones es el objetivo principal de muchos proyectos de investigación actuales y constituye el fin principal de la aplicación que aquí se propone.

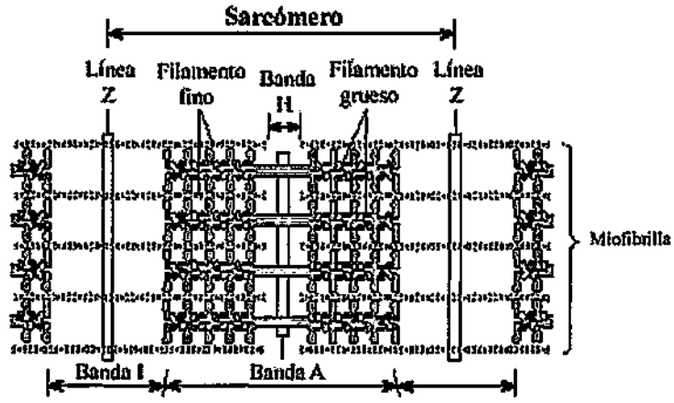
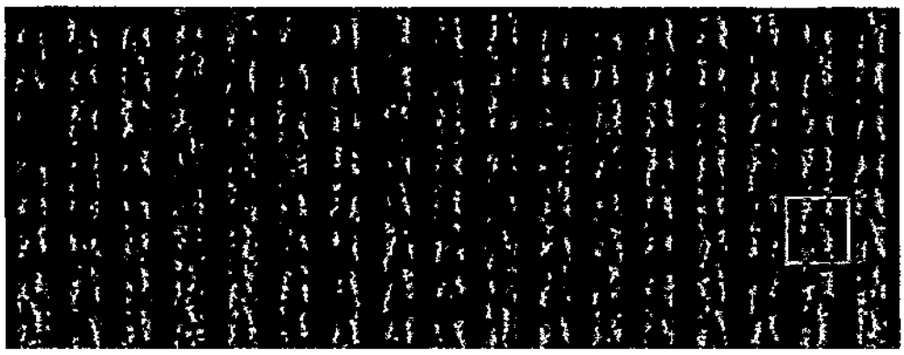



**Figura 9.3** Representación esquemática de la estructura de la miofibrilla.

El músculo de vuelo del insecto (Insect Flight Muscle, IFM) del *Lethocerus* sp. es un espécimen ideal para este tipo de estudio debido a que su malla de filamentos es la más ordenada que existe en el reino animal. Este ordenamiento permite el agrupamiento de muchos de estos puentes en estructuras similares facilitando de esta forma la obtención de su estructura tridimensional a más alta resolución. La figura 9.4 muestra un ejemplo de reconstrucción tridimensional del IFM utilizando tomografía electrónica, donde se pueden observar las disposiciones espaciales ordenadas de estas estructuras. El patrón repetitivo observado en esta figura y marcado con un rectángulo rojo en la figura 9.4a y representado como una malla transparente en la figura 9.4b contiene los puentes que conectan o unen el filamentos fino (actina) con los dos filamentos gruesos (miosina). A este patrón se le llama motivo.

a)



b)

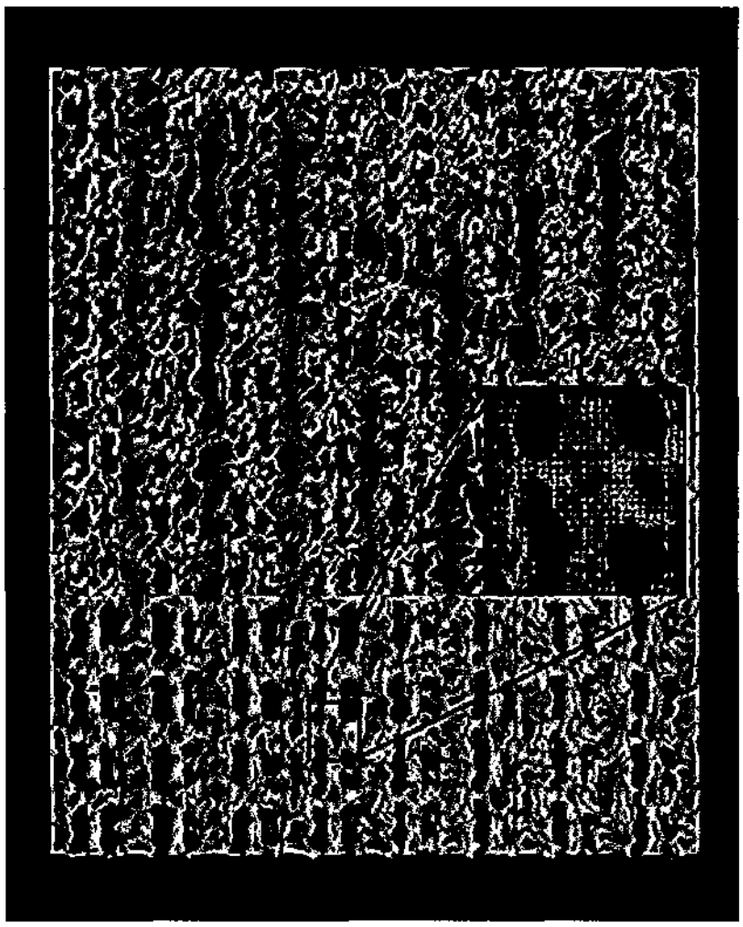


**Figura 9.4** Tomograma 3D del IFM congelado repentinamente durante una contracción activa. a) Sección central de la reconstrucción. La imagen muestra los ejes de los filamentos finos y gruesos. El cuadro marcado en rojo indica el tamaño de un motivo (50x56 nm). b) Imagen compuesta que muestra vistas de la superficie de un tomograma 3D. La parte superior resaltada en color naranja muestra la reconstrucción no promediada y la parte inferior color oro muestra la reconstrucción utilizando técnicas de promediado axial.

El IFM tiene como particularidad estructural la característica de que posee filamentos de actina (filamentos finos) posicionados en medio de pares de filamentos de miosina (filamentos gruesos). Esta característica es diferente a la que poseen los músculos estriados de los vertebrados cuya distribución es triangular [144]. La disposición de los filamentos finos permite una mejor expresión de la simetría espiral de orden 2 de la actina, sin embargo, estudios previos utilizando difracción de rayos X han demostrado que esta simetría de orden 2 sólo aparece por la rotación aleatoria en 180° del filamento de la actina sobre su eje helicoidal [145]. Adicionalmente, esta estructura posee un desorden intrínseco debido al intervalo de aparición de los filamentos finos y gruesos. Estos dos factores en su conjunto introducen una gran heterogeneidad en todo el sistema.

El músculo de vuelo del insecto ha sido estudiado extensivamente utilizando métodos de reconstrucción tridimensional de imágenes. Por ejemplo, reconstrucciones medias del IFM han sido obtenidas utilizando adaptaciones de esquemas de reconstrucción cristalográficas [146-148]. Sin embargo, el desorden intrínseco de esta estructura limita mucho el tipo de información que puede ser obtenida de tales reconstrucciones debido fundamentalmente a que estas técnicas calculan la media de objetos estructuralmente diferentes. La tomografía electrónica también ha sido aplicada para estudiar la estructura del IFM [149] aunque las reconstrucciones no utilizan el promediado de los especimenes y por lo tanto la imagen tridimensional obtenida se caracteriza por su alto contenido en ruido. Sin embargo, las reconstrucciones tomográficas conservan la variación estructural inherente que se encuentra en el IFM, haciendo posible el uso de métodos de clasificación para identificar motivos parecidos que puedan ser promediados para mejorar la relación señal/ruido de las imágenes.

En este contexto, Winkler y Taylor propusieron por primera vez el uso de técnicas estándares de clasificación ya utilizadas en el análisis de partículas individuales 2D para la clasificación de motivos 3D extraídos del tomograma del IFM [150]. Para realizar esta clasificación, utilizaron una combinación de métodos que incluían el análisis de correspondencia (CA) y clasificación jerárquica ascendente (HCA). Ambos métodos han sido importados del campo de la clasificación de partículas individuales en EM y han sido explicados en detalle en la sección anterior de esta memoria. Sin embargo, el tipo de problemas presente en la clasificación de motivos en tomografía es

ligeramente diferente a la clasificación de partículas 2D en EM . El conjunto de estructuras aquí tratadas constituye un conjunto muy heterogéneo de datos que difieren entre sí de varias formas, incluyendo la orientación y las diferencias en la periodicidad axial de los filamentos de la actina y la miosina. Esto implica que mientras más cercanos se encuentren los filamentos de actina y miosina en el enrejado, más complicado es el proceso de clasificación utilizando imágenes de proyecciones 2D de la serie de inclinación. Esta clasificación utilizando solamente las imágenes de proyección se convierte en un proceso complicado debido a la contaminación de los motivos adyacentes a medida que el espécimen es inclinado. Por lo tanto, la clasificación no puede hacerse a partir de imágenes de proyección sino a partir de las imágenes 3D de los motivos, lo que provoca un incremento drástico de la dimensionalidad del problema.

Winkler y Taylor utilizaron 423 motivos tridimensionales que fueron cortados y extraídos del mapa tomográfico. Los motivos extraídos fueron sometidos a un proceso iterativo que alterna un paso de reducción de dimensionalidad utilizando CA, un paso de alineamiento con múltiples referencias y finalmente un paso de agrupamiento jerárquico utilizando HAC. La figura 9.5 muestra una vista de la superficie de cuatro motivos típicos de esta población. Como ese puede observar, estas imágenes 3D se caracterizan por un nivel alto de ruido que impide discriminar visualmente los patrones presentes en el puente que une el filamento de la actina con los dos filamentos de miosina.

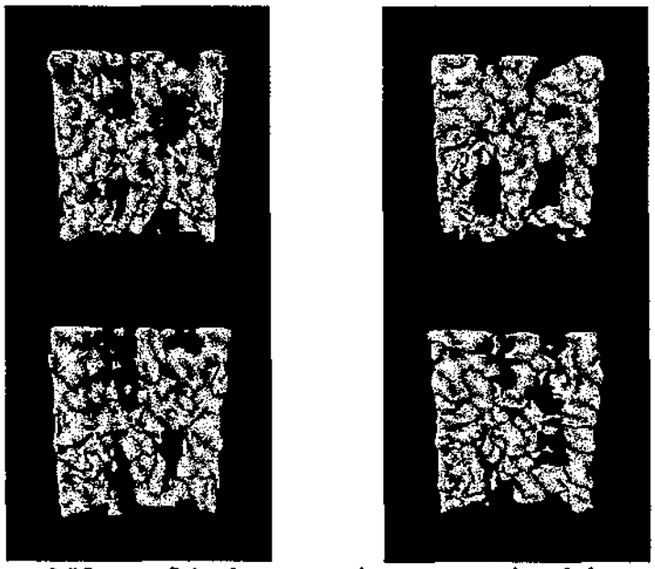


**Figura 9.5** Isosuperficies de cuatro motivos representativos de la población de motivos extraídos del tomograma del IFM.

Inicialmente se realizó un proceso a agrupamiento utilizando como vector de características funciones invariantes a la traslación e invariantes a la rotación en 180° con respecto al eje del filamento de la actina. Estas funciones que se utilizaron como paso preliminar para un agrupamiento jerárquico, son funciones de doble auto correlación (DACF) [151]. Dicha clasificación inicial permitió obtener un conjunto de imágenes medias de referencia que no estaban sesgadas por una selección manual de motivos. Estas referencias iniciales fueron utilizadas posteriormente en un proceso de alineamiento de múltiples referencias utilizando las imágenes medias de los grupos extraídos por HAC [150].

Una vez que las imágenes 3D fueron alineadas inicialmente, se comenzó el proceso iterativo que repetía los pasos de alineamiento con múltiples referencias seguido del análisis de factores por CA y finalmente un proceso de agrupamiento por HAC utilizando solamente los primeros 8 factores extraídos por CA. Este proceso de agrupamiento produce un conjunto de imágenes medias de grupos homogéneos refinadas por el alineamiento, las cuales a su vez fueron utilizadas de nuevo como referencias en el primer paso de alineamiento constituyendo un nuevo ciclo en el proceso de análisis. La figura 9.6 muestra los resultados obtenidos en [150] utilizando dos ciclos de refinamiento. En este caso, el árbol jerárquico se dividió manualmente para producir 16 grupos. Un experimento parecido con el mismo conjunto de imágenes pero utilizando 25 clases también ha sido recientemente reportado en [152].

Este método de clasificación ha permitido obtener nueva información acerca de la mezcla de configuraciones de los puentes, muchos de los cuales están específicamente adosados a la actina [150]. Sin embargo, a pesar de que la combinación de estos métodos ha demostrado su capacidad para extraer información relevante de este conjunto de datos complejos, también son bien conocidas sus desventajas. La clasificación jerárquica ascendente [47] es una técnica muy conocida que produce un orden lineal de los datos organizados en una representación de árbol en la cual los patrones más similares son agrupados en una jerarquía de subconjuntos anidados. A pesar de la simpleza conceptual de este método, es también bien conocido que sufre de muchos problemas y falta de robustez cuando se trabaja con datos de muy alta dimensión y elevado nivel de ruido, como es el caso de las imágenes 3D de tomografía

aquí tratadas. Estos métodos han sido desarrollados en el contexto de aplicaciones donde los datos seguían una estructura más o menos jerárquica, como es el caso de los datos de filogenia. Es por eso que quizás este tipo de métodos no estén completamente adaptados para trabajar con datos de otra naturaleza, especialmente si presentan un alto contenido en ruido y alta dimensión. Por tanto, esta metodología propuesta puede no ser la mejor cuando se trabaje con datos aún más complejos que los de la presente aplicación de IFM.

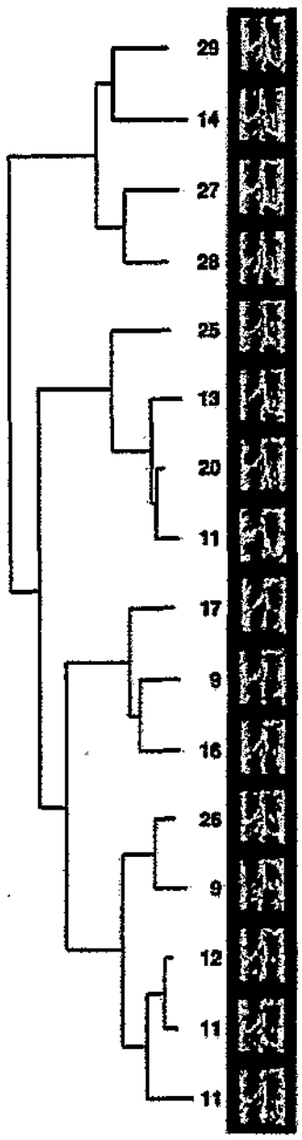



**Figura 9.6** Resultados de aplicar HCA al conjunto de motivos extraídos del tomograma del IFM.

En el contexto de la presente memoria proponemos la utilización de la nueva técnica de mapas auto-organizativos (KerDenSOM) como alternativa para la clasificación de las imágenes 3D presentadas en esta sección. El conjunto de datos utilizados para mostrar la eficacia de KerDenSOM son los mismos utilizados por Winkler y Taylor [150] una vez alineados de la manera descrita anteriormente. Como ya se ha descrito, las estructuras utilizadas en esta aplicación están compuestas por filamentos gruesos y finos dispuestos de manera alternativa y conectados por pares de puentes cruzados. Por lo tanto el interés fundamental del proceso de clasificación es determinar las heterogeneidades presentes únicamente en la estructura del puente y no en toda la imagen tridimensional. Es por eso que se utilizó una máscara binaria para extraer solamente aquellos voxeles presentes en la zona donde reside los pares de puentes conectando el filamentos de actina con los dos filamentos de miosina. La máscara utilizada se muestra en la figura 9.7. El resultado de la aplicación de la máscara produjo vectores de dimensión 4807, quedando de esta forma el conjunto de datos compuesto por 423 vectores de 4807 componentes (número de voxeles) cada uno.

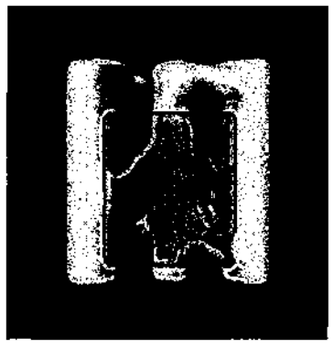


**Figura 9.7** Máscara utilizada para extraer los voxeles de interés sobre la imagen media calculada sobre toda la población.

Utilizando este conjunto de datos, se ejecutó el algoritmo de KerDenSOM utilizando un mapa de 10x5 con topología rectangular. El núcleo utilizado para la estimación de la densidad de probabilidad fue el núcleo Gaussiano. Para intentar garantizar la convergencia del método, se ejecutó el algoritmo en cinco pasos de enfriamiento determinista variando la constante de suavidad desde 300 hasta 50. El proceso completo duró aproximadamente 12 minutos en una estación Silicon Graphics SGI Origin 200 con procesadores R12000 a 360MHz y 1.5 GB de memoria RAM. La figura 9.8 muestra el mapa resultante. Nótese que las imágenes 3D mostradas en esta

figura representan la superficie de las imágenes formada por cada vector diccionario, pero únicamente en la zona marcada por la máscara utilizada, que es precisamente el área de interés en el caso que se está analizando.

El mapa resultante ha condensado toda la variación detectada en los motivos originales en un conjunto reducido de elementos representativos. En este caso 50 vectores diccionarios distribuidos en una malla de 10x5. Estos vectores diccionarios representantes del conjunto inicial de datos, sin embargo, poseen propiedades estadísticas muy similares a este conjunto inicial de datos.

Una inspección visual del mapa obtenido y mostrado en la figura 9.8 manifiesta la naturaleza propia de KerDenSOM, el cual ha organizado los vectores diccionarios en el plano de salida de forma tal que las variaciones en este plano se realizan de manera suave y ordenada, garantizando que la proximidad geométrica de los vectores diccionarios en el mapa reflejen lo mas fielmente posible la similitud de los datos asignados a cada uno de ellos evidenciando de esta manera la estructura de grupos de los datos originales.


**Figura 9.8** Resultados del algoritmo de KerDenSOM en la clasificación de los motivos extraídos del tomograma del IFM. Se utilizó un mapa de 10x5 con topología rectangular y núcleo Gaussiano. El algoritmo se ejecutó en cinco pasos de enfriamiento determinista decrementando el valor de la constante de suavidad desde 300 hasta 50. El número de motivos asignados a cada grupo se muestran en la esquina superior derecha de cada vector diccionario. Los seis grupos que representan las regiones más pobladas del mapa han sido marcados y etiquetados.

Este tipo de organización es posible porque cuando un dato original es presentado a la red el vector diccionario más parecido y un conjunto de sus vecinos más cercanos adaptan sus valores para representar este dato original de la manera más fielmente posible, creando regiones donde los vectores diccionarios son muy parecidos y provocando que datos originales similares sean proyectados hacia estas áreas de vectores diccionarios similares. De esta forma el efecto de agrupamiento de los datos queda fielmente reflejada en el mapa.

Es importante señalar que las áreas que contienen vectores diccionarios sin datos asignados (áreas de baja densidad marcadas en la figura 9.8 con un 0 en la esquina superior derecha de cada vector) representan zonas de transición entre grupos aparentes formados por zonas de alta densidad de datos. Este efecto ocurre debido a la naturaleza intrínseca de los mapas auto-organizativos que intentan garantizar transiciones suaves a lo largo del mapa, ayudando de esta forma la identificación de la estructura de grupos presente en los datos originales. Por consiguiente, una regla simple para realizar el agrupamiento de los vectores diccionarios puede ser la de segmentar aquellas zonas con apariencia similar y que representen máximos de densidad separadas por zonas de más baja densidad.

Desde el punto de vista estructural, el mapa mostrado en la figura 9.8 revela la existencia de varios grupos de motivos con diferente composición de los puentes cruzados representados por los seis grupos de vectores diccionarios más poblados en el mapa. Estos resultados coinciden con los obtenidos previamente utilizando la combinación de CA y HAC [150, 152]. Estos seis grupos marcados como A, B, C, D, E y F en la figura 9.8 representan variaciones de tres tipos clásicos de estructuras de puentes cruzados: simple, doble y doble incompleta. Los puentes cruzados dobles han sido asignados en su totalidad a los grupos B, D y F. Así mismo los puentes cruzados dobles incompletos se agruparon en los grupos A y C y finalmente los puentes simples quedaron clasificados en el grupo E. La figura 9.9 muestra las imágenes medias calculadas a partir de los motivos originales que estos grupos representan.

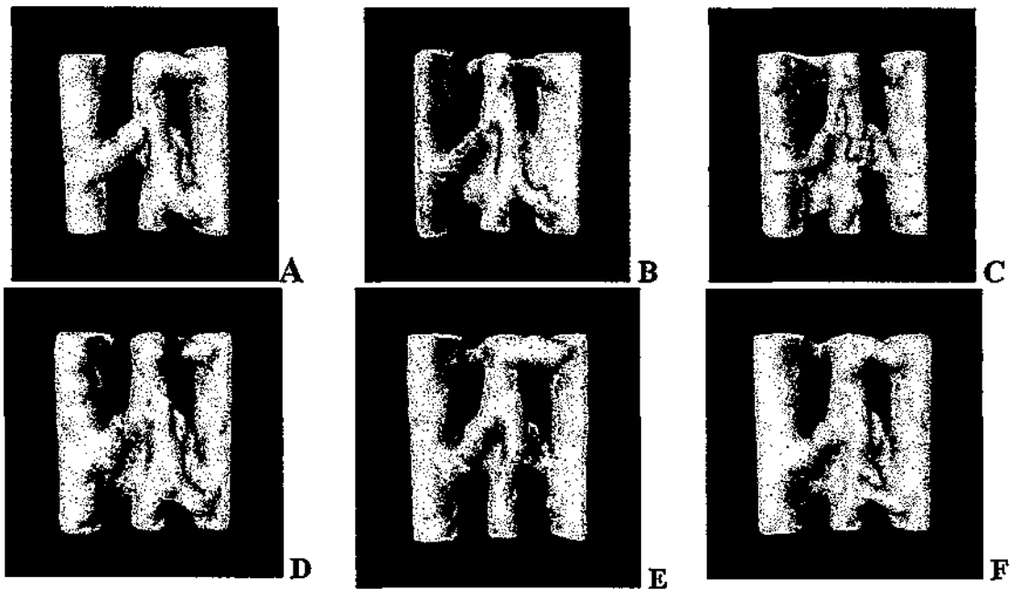



**Figura 9.9** Imágenes medias de los motivos originales asignados a cada grupo marcado en el mapa de la figura 9.8. El número de motivos asignados a cada grupo son los siguientes: A = 87, B = 86, C = 29, D = 35, E = 43 y F = 75.

La observación directa de los grupos encontrados por KerDenSOM permite extraer conclusiones en cuanto a las estructuras de los motivos analizados que tienen un marcado interés biológico. Por ejemplo, aunque la apariencia de los puentes cruzados dobles representados por los grupos B, D y F es ligeramente similar, estos grupos representan motivos diferentes en cuanto a la estructura del puente principal (parte superior) y el puente trasero (parte inferior). Esta es la principal razón por la cual estos motivos fueron asignados a lugares relativamente distantes en el mapa. En los grupos B y D, el puente principal a la derecha del filamento fino (actina) es más ancho axialmente que el puente principal a la izquierda de este mismo filamento. Sin embargo, en el caso del grupo F, ocurre justo lo contrario, el puente principal a la izquierda del filamento fino es más ancho que el puente principal a su derecha. Adicionalmente, las diferencias principales entre los grupos B y D radican en la extensión del puente trasero. Este puente trasero a la derecha del filamento fino en el grupo D es mas extenso que su homólogo en los grupos B y F que a su vez presentan este puente trasero a la izquierda del filamento fino mas extenso que en el caso del grupo D.

A pesar de que el número real de cabezas de miosina que pueden ser acomodadas en cada puente cruzado no puede ser calculado exactamente sin la existencia de un modelo, es conocido por trabajos previos que en general cada uno de los puentes cruzados principales contienen dos cabezas de miosina. Así mismo cada

puente cruzado trasero contiene una cabeza simple de miosina. Calculando la diferencia relativa en cuanto a extensión entre los puentes principales y traseros presentes en los seis grupos encontrados por KerDenSOM se ha demostrado que al parecer los puentes cruzados ciertamente cumplen esta regla. Utilizando los 6 grupos obtenidos se calculó el número promedio de cabezas de miosina por motivos que aparecen en la reconstrucción tomográfica. Experimentalmente deben aparecer aproximadamente 5.44 cabezas de miosina por motivo [153-155] y el número representado por los grupos extraídos por KerDenSOM es 5.43. Este dato evidencia la precisión en el agrupamiento que produce este algoritmo, ya que al intentar preservar la densidad de probabilidad de los datos originales, es posible cuantificar fielmente los valores de densidad de cada grupo extraído.

Como característica adicional observada en el mapa de la figura 9.8 podemos señalar una pequeña pero significativa fuente de variación evidenciada por el hecho de que varios vectores diccionarios "perdieron" parte de la densidad de la columna derecha (filamento grueso). Esto significa que existe un ligero problema de alineamiento de las estructuras originales. Estas variaciones provocadas por deficiencias del proceso de alineamiento pueden ser observadas por el mapa generado por KerDenSOM aunque su efecto no resulta dominante con respecto a las variaciones estructurales de los motivos. Sin embargo, la presencia de estas diferencias de alineamiento es una evidencia de que KerDenSOM es capaz de representar no solo fuertes variaciones de los datos, sino también aquellas menos significativas y por lo tanto más difíciles de detectar por métodos tradicionales en presencia de alta dimensionalidad y alto nivel de ruido..

Como ventaja adicional en esta aplicación sobre datos tomográficos podemos señalar que KerDenSOM no necesita conocer a priori el número de clases a extraer. Esta información es posible observarla en el mapa resultante sin necesidad de imponerle al algoritmo esta condición. Si bien es cierto que el tamaño del mapa está relacionado con el número de grupos que será capaz de obtener, este parámetro no resulta tan crítico como en el caso de los métodos de agrupamiento tradicionales. En este sentido este algoritmo resulta una poderosa herramienta para la clasificación de motivos 3D, donde no se tiene información previa de la estructura y las fuentes de variación de los datos que permita predecir el número de clases.

## 10. Modelado de forma y topología en imágenes 3D

En las dos secciones anteriores de esta memoria hemos mostrado aplicaciones en distintas áreas de biología estructural donde el principal objetivo es conseguir las estructuras tridimensionales de complejos macromoleculares. En los últimos años ha habido un incremento constante en el número de estructuras que han sido ya resueltas por la comunidad científica, abriéndose una nueva área conocida como genómica (ó proteómica estructural).

Este rápido crecimiento de información estructural tridimensional está suponiendo un reto importante en campos de las tecnología de la información tales como las bases de datos, necesarias para la manipulación de los grandes volúmenes de información que estás técnicas generan. Este problema se agrava aún mas por la complejidad cada vez más creciente de los datos en sí, siendo necesario desarrollar nuevas técnicas específicas para analizar y representar esta compleja información.

Entre la amplia variedad de esfuerzos dedicados al manejo y mantenimiento de bases de datos de estructuras tridimensionales, podemos señalar el Banco de Datos de Proteínas (Protein Data Bank, PDB) [156]. Esta base de datos ha sido diseñada para almacenar y manipular estructuras tridimensional de proteínas resueltas a resolución atómica por cristalografía de rayos X (RX), por resonancia magnética nuclear (NMR) ó por microscopía electrónica tridimensional.

La utilidad de esta base de datos ha quedado evidenciada por la gran cantidad de estudios científicos que la han utilizado, fundamentalmente en trabajos relacionados con similitudes estructurales y propiedades bioquímicas [157-160]. También es importante destacar otros tipos de estudios basados principalmente en la información estructural relacionada con la forma y la geometría de las macromoléculas, como lo son los estudios de acoplamiento entre proteínas, interacciones ligandos-proteínas, etc. [161-163].

Desde el punto de vista biológico el mayor interés, en el contexto de esta nueva aplicación, es la caracterización de la topología y la superficie de las macromoléculas biológicas a partir de datos de media resolución, como son los datos producidos por la microscopía electrónica y que han sido tratados en secciones anteriores. La razón fundamental que justifica el interés en técnicas como la microscopía electrónica es que

no se requiere que los especimenes estudiados formen cristales, como es el caso de técnicas como la difracción de RX. Asimismo, las estructuras obtenidas a baja resolución complementan cada vez mas los datos a resolución atómica [142, 164-166]. Ejemplo de ello son los esfuerzos dedicados a encajar estructuras resueltas a resolución atómica en estructuras mas grandes resueltas a baja y media resolución por microscopía electrónica [167-170].

La integración de información a alta y baja resolución, sin embargo, impone un serio reto técnico a nivel de base de datos. La razón fundamental es que los datos a media resolución se representan de manera completamente distinta a los datos a resolución atómica. Estos últimos están formados por las coordenadas precisas de los átomos que constituyen la estructura molecular. Por el contrario, los datos a media resolución son representados como mapas de densidad en una malla tridimensional discreta (imagen 3D), en las cuales cada punto (voxel) tiene asociado un valor de densidad. Adicionalmente, debido al hecho de que los datos a media resolución resuelven estructuras mas grandes, el tamaño de estos conjuntos de datos (número de voxeles) usualmente es muchísimo mayor que los datos de estructura atómica, lo que implica la necesidad de contar con sistemas de manipulación y consulta mas complejos y eficientes.

Una manera de entender correctamente el amplio espectro de características presentes en las estructuras resueltas a media resolución (imágenes 3D) puede ser a través de sus propiedades geométricas, por ejemplo, sus forma. Sin embargo, utilizando solamente la información de densidad proporcionada por los puntos que la definen (voxeles) esto no es posible debido a que la forma geométrica de un conjunto de puntos no conectados no está definida. Es por eso que la mayoría de los esfuerzos realizados para tratar con la forma de este tipo de datos han ido encaminados de alguna manera hacia la definición de su superficie.

En el caso de las estructuras a resolución atómica la propia naturaleza de los datos hace posible la definición de un modelo de superficie teóricamente preciso [171-173]. Este proceso, sin embargo, no es válido para el caso de datos a media resolución, en los cuales se deben utilizar algoritmos de segmentación para extraer los contornos del objeto 3D. En este último caso la obtención de la superficie externa de una

macromolécula no es tarea sencilla, de forma que los resultados obtenidos presentan cierta dependencia con el algoritmo de segmentación utilizado.

La resolución de estas estructuras macromoleculares introduce otro problema adicional debido al hecho de que características estructurales a distintos niveles de resolución no tienen por qué preservarse. Esto implica que características importantes tales como depresiones y canales pueden cambiar su forma y tamaño llegando incluso hasta desaparecer por completo con el cambio de resolución. Por lo tanto, la resolución es un parámetro crítico que debe ser tratado cuidadosamente cuando se comparan datos volumétricos.

Por lo tanto, el objetivo de la aplicación que aquí se propone es el desarrollo de una metodología eficiente de representación de datos volumétricos a baja y media resolución que puedan ser almacenados, manipulados y comparados entre sí de manera eficaz en el contexto de bases de datos. Esto implica la utilización de técnicas de compresión combinada con la creación de un modelo de representación que preserve las características de forma y topología presentes en las estructuras tridimensionales y que permitan posteriormente el acceso a su información estructural.

## 10.1. Representación de formas: Alfa- Formas (Alpha-Shapes)

El concepto de alfa-formas ($\alpha$-shapes), introducido por primera vez por Herbert Edelsbrunner [174], es una metodología para formalizar la noción intuitiva de forma de un conjunto de puntos espaciales. Las alfa-formas representan un concepto geométrico concreto, matemáticamente bien definido, que constituye una generalización de la envolvente convexa (convex hull) y un subgrafo de la triangulación de Delaunay. Utilizando esta teoría es posible asociar una familia de formas a un conjunto finito de puntos en un espacio euclídeo de *n* dimensiones. Cada forma constituye un polítopo (sólido n-dimensional con caras planas) derivado de la triangulación de Delaunay de un conjunto de datos y donde el parámetro $\alpha \in \Re$ controla el nivel de detalles deseado.

Matemáticamente podemos definir la triangulación de Delaunay de la manera siguiente: dado un punto en el espacio con un peso asignado $P=(p, w_p)$ donde $p \in \Re^n$, la distancia ponderada desde un punto cualquiera $x \in \Re^n$ a P, se define como $\Pi_P = \|p-x\|^2 - w_p$, siendo $\|p-x\|^2$ la distancia euclídea entre p y x. Adicionalmente, dado un conjunto $\{P_i\}$ de puntos con peso asignado, el diagrama ponderado de Voronoi es la

partición del espacio en regiones convexas (celdas) donde la i-ésima celda es el conjunto de puntos más cercanos a $P_i$. (según la métrica dada por la distancia ponderada). La triangulación ponderada de Delaunay es el grafo de adyacencia entre caras construido a partir del diagrama ponderado (dual). Existe una conexión entre un par de vértices de la triangulación siempre que sus celdas correspondientes en el Diagrama Ponderado compartan una cara. La triangulación de Delaunay de un conjunto de puntos define su envolvente convexa que está compuesta por elementos lineales de orden k (k-simplices), para k=0,1,2,3:

0-simplex: puntos en el espacio n-dimensional.

1-simplex: segmento que une dos puntos.

2-simplex: triángulo formado por tres puntos.

3-simplex: tetraedro formado entre cuatro puntos.

Estos conceptos ya fueron introducidos hace tiempo y aplicados al campo de la biología estructural para la medida de volumen y área de macromoléculas representadas con distintos modelos de superficie [171, 173]. La figura 10.1 muestra dos ejemplos de este tipo de modelos. Para ellos, la triangulación del espacio en regiones de Voronoi es la base topológica para construir la superficie. Los átomos son considerados como puntos ponderados, esto es, esferas $B(p, r_{vw})$ en $\Re^3$ donde p es la localización del átomo y $r_{vw}$ el peso del correspondiente radio de Van der Waals [171]. Esto es, la triangulación ponderada de Delaunay definida sobre el conjunto de átomos de la molécula dada, proporciona su estructura topológica subyacente (conectividad).

La teoría de formas alfas extiende todos estos conceptos mediante al introducción de un nuevo parámetro $\alpha$. Supongamos que el radio de todos los átomos (esferas) de la molécula empieza a crecer simultáneamente en un incremento $\alpha$. Así, cada átomo se redefine como una esfera $B_\alpha=(p, r_\alpha)$ donde $r_\alpha=\sqrt{r_{vw}^2+\alpha^2}$. Conforme $\alpha$ se incrementa (ver figura 10.2) las esferas crecen gradualmente de modo que en algún momento empezarán a solaparse entre sí. En el momento en el que el borde de dos esferas se tocan aparece un nuevo simplex 1-dimensional (segmento) y se añade al complejo de simplices correspondiente a ese valor de $\alpha$. Cuando se interceptan 3 esferas entre sí, añadimos un triángulo e igualmente un tetraedro cuando son 4. El complejo de simplices para un valor concreto de $\alpha$ es un subconjunto del complejo de Delaunay y se llama complejo alfa. La forma alfa es la parte del espacio euclídeo ocupada por el

complejo alfa. Cuando $\alpha$=0 (zero-shape) obtenemos la topología de la molécula a partir del radio de Van der Waals. En cambio cuando $\alpha$ tiende a $\infty$, el complejo alfa es la envolvente convexa del conjunto inicial de puntos.

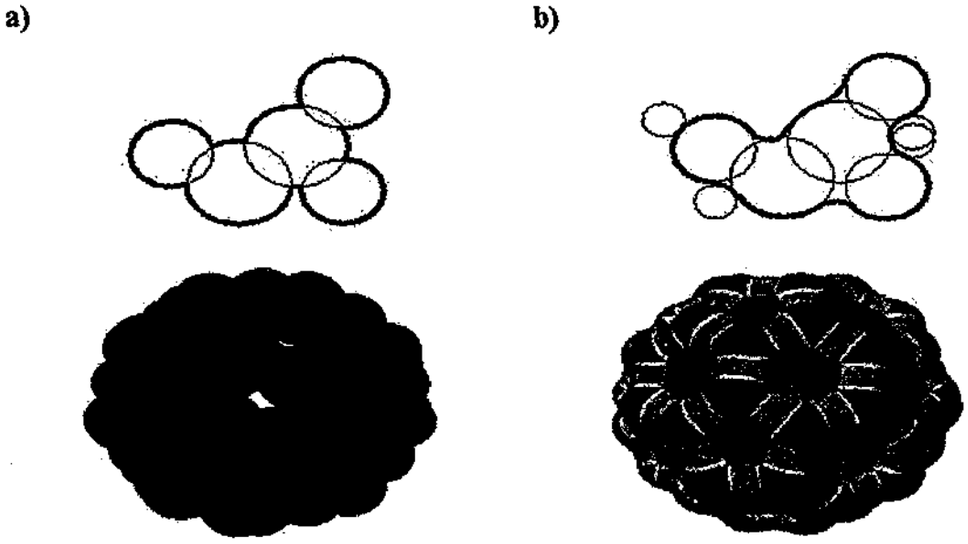



**Figura 10.1** Modelos de superficie para datos de moléculas a alta resolución. a) Superficie de van der Waals. b) Superficie molecular.

Aquellos simplices que están en un complejo alfa para un valor de $\alpha_1$, también están en todos los complejos para $\alpha_2$, con $\alpha_1<\alpha_2$ (propiedad de inclusión). A su vez, todos los complejos alfa son subcomplejos del complejo de Delaunay (envolvente convexa).

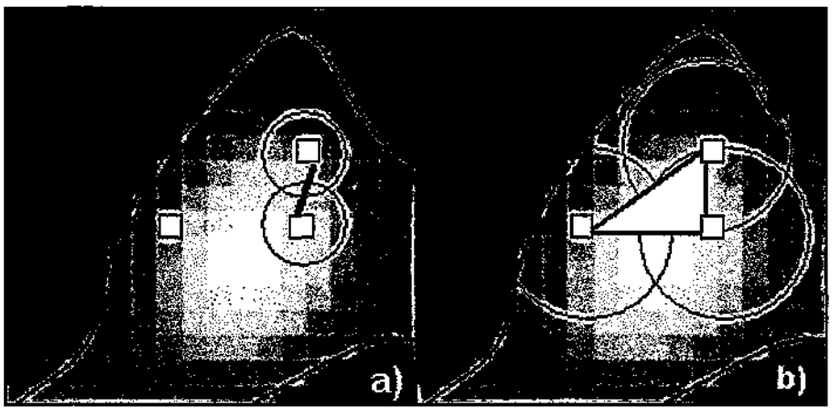



**Figura 10.2** Esquema del funcionamiento de la teoría de formas alfa. Dados tres puntos dentro de un objeto, a) ilustramos como el incremento del radio en las esferas hace que dos de ellas se solapen, apareciendo así un nuevo 1-simplex (segmento) que será añadido al complejo alfa. b) Cuando el radio es lo suficientemente grande para que se intercepten las tres esferas se genera un 2-simplex (triángulo).

El complejo alfa codifica información combinatoria, topológica y métrica sobre la molécula. De ahí que la teoría de las formas alfa sea una poderosa herramienta de estudio en el campo de la biología estructural. Su justificación matemática es sólida y ya ha sido exitosamente probado con datos de moléculas a alta resolución.

El aporte del presente trabajo radica en la aplicación de la teoría de alfa formas a imágenes 3D de macromoléculas a media/baja resolución de las que no se dispone de su estructura atómica. Los mapas de densidad están formados por un conjunto relativamente elevado de puntos (voxeles) que están distribuidos espacialmente de manera equidistante, con lo cual el uso directo sobre estos datos de la teoría de las alfa-formas carece de sentido. El objetivo es la creación de un modelo compacto e informativo de estas imágenes que permita extraer información importante sobre su forma y topología, pero la posición de los voxeles solamente no es información suficiente para poder modelar la forma del objeto que ellos codifican. Para esto hay que tener también en cuenta los valores de densidad presentes en la imagen.

Es por ello que para la utilización efectiva de esta teoría se necesita un paso previo de extracción de características que codifiquen de manera eficiente la densidad de los objetos contenidos en la imagen 3D. Esta extracción de características se lleva a cabo a través de un proceso de cuantificación vectorial utilizando el algoritmo KCM (Kernel c-means) propuesto en esta memoria.

## 10.2. Cuantificación vectorial de la densidad

La idea principal de la cuantificación vectorial está basada en segmentar el espacio vectorial original en un conjunto de grupos, cada uno de los cuales será representado por un solo vector, usualmente llamado vector diccionario, y que tiende a explicar lo mejor posible aquellos datos a los que representa. Muchas formas de cuantificación vectorial han sido propuestas, diferenciándose principalmente entre sí por la manera en que se trata la relación entre los datos originales y el conjunto reducido de salida [69].

Uno de los algoritmos más utilizados para realizar cuantificación vectorial es el llamado algoritmo LBG, cuyo nombre viene dado por las iniciales de sus creadores (Yoseph Linde, Andrés Buzo y Robert Gray) [175]. Este algoritmo se basa en la minimización de la distorsión entre los vectores diccionarios y los datos que estos

representan y por lo tanto tiende a encontrar una partición donde el error de distorsión es mínimo. Sin embargo, a pesar de que es conocido que el conjunto de elementos representantes (vectores diccionarios) producidos mediante técnicas de cuantificación vectorial tienden a aproximar la función de distribución de probabilidad del conjunto de datos originales, esta relación no queda reflejada de forma clara en la formulación matemática de dichos métodos.

En el caso que nos proponemos aplicar el método de cuantificación vectorial, la preservación de la densidad de probabilidad de los datos originales es muy importante, precisamente es este tipo de información la que queremos codificar. Es decir, en los mapas de densidad obtenidos por microscopía electrónica, la información de densidad asociada a cada voxel de la imagen 3D es la que contiene la información de la estructura macromolecular que esta imagen representa. Es por ello que la utilización de un método de cuantificación vectorial debe garantizar de forma explícita que los puntos subrogados encontrados representen de manera fiel la densidad original de la estructura que se estudia.

La aplicación de métodos de cuantificación vectorial a datos de microscopía electrónica fue propuesta por primera vez por Willy Wriggers [169, 170], el cual aplica exitosamente este tipo de técnicas sobre datos de crio-microscopia orientado al acoplamiento (docking) de datos de coordenadas atómicas en mapas de media resolución.

En este apartado proponemos el uso del algoritmo de KCM, que es una red neuronal basada en una función de costo específicamente diseñada para seleccionar un conjunto de vectores representativos cuya densidad de probabilidad se asemeje lo mejor posible a la distribución de probabilidad de los datos de entrada. Para nuestra aplicación, los objetos de entrada $\mathbf{X}_i \in \Re^{3\times 1}$ se corresponden con los voxeles del volumen de 3D-EM, seleccionados de forma que representen su densidad. Los vectores diccionarios ó elementos representantes resultantes son $\mathbf{V}_j \in \Re^{3\times 1}$ , y los llamaremos pseudo-átomos ya que para nosotros desempeñarán un papel análogo a los átomos en una estructura a alta resolución. Los pseudo-átomos conforman la base en nuestra representación de la geometría de la macromolécula. La figura 10.3 muestra de forma esquemática un conjunto de 12 pseudo-átomos calculados para un mapa de media resolución de la macromolécula Gal6.

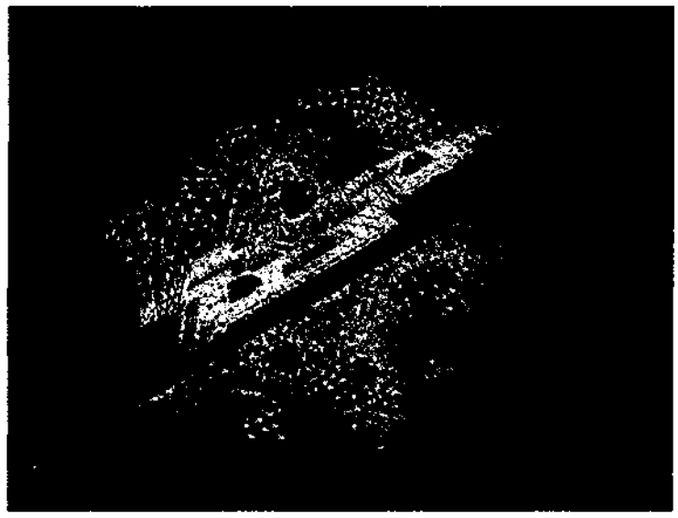


**Figura 10.3** Representación esquemática del resultado de la cuantificación vectorial. En este caso, hemos usado el algoritmo Kernel c-means para situar 12 pseudo-átomos (esferas rojas) dentro del mapa de densidad de baja resolución de la bleomicín hidrolasa Gal6. La estructura es un hexámero que puede considerarse como un trímero de dímeros. Esta simetría se conserva en la disposición espacial de los pseudo-átomos debido a la buena estimación de la pdf original.

El uso que hacemos de los complejos alfa se basa en este método de cuantificación vectorial que nos proporciona el conjunto de 0-simplices (puntos) a partir de los cuales construimos el complejo alfa.

Un aspecto importante a considerar es la determinación del valor alfa apropiado para construir el complejo. El requisito básico es que el complejo tenga las mismas propiedades topológicas que la macromolécula original. En este trabajo hemos seleccionado el máximo del error de cuantificación disponible para cada pseudo-átomo como valor alfa. El error de cuantificación representa la media de las distancias desde el pseudo-átomo a cada uno de los datos de los cuales es representante. Esta medida es usualmente usada en la evaluación de la calidad de la cuantificación e intuitivamente, puede interpretarse como el radio de la región de Voronoi definida para cada pseudo-átomo, lo que justifica su utilización como el parámetro alfa más adecuado para obtener un complejo alfa topológicamente correcto.

### 10.2.1. Estabilidad y eficiencia de la cuantificación vectorial

Al igual que la mayoría de los métodos de cuantificación de vectores existentes, el algoritmo de KCM puede mostrar ciertas variaciones en dependencia de los valores

iniciales de los parámetros utilizados. Esto significa que el máximo local al cual el algoritmo converge depende sensiblemente de los valores iniciales de los vectores diccionarios (pseudo-átomos en este contexto). Generalmente esta variación depende directamente de la forma de la distribución 3D de densidades con la que se trabaja y del número de pseudo-átomos que se pretende obtener.

Con el objetivo de medir la estabilidad de este algoritmo en distintas condiciones de inicialización, hemos ejecutado el algoritmo 10 veces sobre un mismo conjunto de datos pero utilizando distintas inicializaciones de los vectores diccionarios (inicialización aleatoria). Para realizar estas pruebas utilizamos el mapa de densidades de la bleomicín hidrolasa Gal6 a 20 Å de resolución (figura 10.3). Para medir el efecto de estabilidad en el caso extremo se utilizaron 1500 vectores diccionarios y se calculó el error cuadrático medio (RMSE) en 10 repeticiones estadísticamente independientes. La variación del RMSE de los vectores diccionarios obtenida fue de 1.79 Å, lo que indica que la posición de estos vectores en diferentes ejecuciones es lo suficientemente estable para este tipo de aplicación.

Sin embargo, debemos enfatizar que este método de cuantificación de vectores tiene como objetivo preservar la densidad de probabilidad de los datos de entrada, lo que se traduce en términos prácticos a que el algoritmo intentará posicionar más vectores diccionarios en aquellas áreas donde los valores de densidad sean mayores y por consiguiente ubicará menos vectores diccionarios en áreas de baja densidad. Esto significa que en zonas de alta densidad la separación de los vectores diccionarios puede ser muy pequeña mientras que en las zonas de baja densidad esta separación tiende a ser mucho mayor. Este hecho explica el por qué la posición de los vectores diccionarios no debe considerarse como un único parámetro para demostrar la estabilidad del algoritmo y por lo tanto otros métodos alternativos deben ser utilizados. En este caso, debido a que el algoritmo produce una estimación de la función de densidad de probabilidad, hemos medido la estabilidad del modelo calculando el logaritmo de la verosimilitud de la densidad estimada en cada una de las 10 ejecuciones estadísticamente independientes, obteniendo una media de 233970.8 y una desviación estándar de 4.18, lo que demuestra de manera fehaciente que el modelo generado es muy estable.

Por último es importante también mencionar que el algoritmo propuesto se ha comparado exhaustivamente con otros métodos clásicos de cuantificación vectorial que

han sido utilizado previamente para este tipo de problemas. Entre ellos podemos señalar k-medias (K-means), c-medias difuso (FCM), TRN (Topology representing network) [176], GNG (Growing Neural Gas) [177, 178] y Growing Cell Structures (GCS) [179, 180]. En todos estos casos el algoritmo KCM propuesto aquí consiguió valores de estabilidad superiores y mejores tiempo de ejecución en muchos casos, excepto en el caso de k-medias y c-medias difuso, que a pesar de manifestar mayor eficiencia computacional, en la mayoría de las repeticiones quedaron atrapados en algún mínimo local, obteniéndose resultados erróneos no deseables.

## 10.3. Algoritmo para la construcción del modelo

En este apartado describimos detalladamente los pasos a seguir para construir el complejo alfa a partir de la reconstrucción del espécimen representada en forma de imagen 3D. Una variante directa de visualización del complejo alfa como una imagen tridimensional puede hacerse a través del mapeo inverso del complejo alfa en una versión binaria y discreta del mismo correspondiente al espacio ocupado en $\Re^3$. De esta forma proporcionamos un mecanismo para la traducción del complejo alfa (distintos simplices que lo forman) a una imagen binaria. El procedimiento es sencillo: para cada voxel de la imagen se comprueba si está contenido dentro de algún simplex del complejo. Si es así se le asigna valor 1, en otro caso se asigna 0. En principio, el parecido de la imagen sintetizada a partir del complejo alfa con la imagen original depende de dos parámetros:

a) La cardinalidad del conjunto de pseudo-átomos. Tal y como se mostrará más adelante, cuanto mayor es el número de pseudo-átomos usados para construir el complejo alfa mejor se aproxima la forma en el modelo resultante.
b) La selección particular del valor $\alpha$ para un conjunto dado de pseudo-átomos debe preservar la topología del volumen original así como aproximar la forma lo mejor posible.

En la práctica, obtenemos el valor $\alpha$ de forma automática como el máximo de los errores de cuantificación para cada pseudo-átomo. Así pues, el único parámetro libre que debe ser optimizado es la cardinalidad del conjunto de pseudo-átomos, lo cual incrementa la eficiencia en el cálculo del modelo. La optimización es guiada por un

número de criterios encaminados a asegurar la preservación de la topología y la geometría (forma).

Las medidas topológicas son proporcionadas directamente por el modelo de formas alfa y como criterio de calidad del modelo podemos utilizar la preservación en la topología, lo cual implica que:

1) Sólo se obtiene una componente conexa, esto es, no queda desconectada ninguna parte del volumen.
2) El complejo alfa presenta igual número de cavidades y canales que el volumen original.

La evaluación del parecido en la geometría, sin embargo, es más complicada. Para poder realizar la comparación, el complejo alfa y el volumen original deben estar en el mismo espacio de representación. Así, el complejo es traducido a una imagen 3D binaria la cual será comparada con una versión binarizada de la imagen original. La máscara binaria del volumen original es obtenida automáticamente mediante la aplicación del filtro de Deriche [181]. Una vez extraído el contorno del objeto con dicho filtro, obtenemos una media de los valores que presenta la imagen en los puntos del contorno. A continuación realizamos una binarización por umbral usando dicha media.

Hemos utilizado un total de 6 descriptores de forma de un objeto tridimensional. La idea es que, teniendo varios descriptores, tenemos distintas visiones desde las que analizar el parecido entre el modelo y el original. Hay que notar que para el cálculo de alguno de los descriptores fue necesario definir un sistema de referencia situado en el centroide del objeto. La razón es sencilla: algunos de los descriptores no son invariantes a rotación (por ejemplo, histograma de normales), por tanto hay que colocar los objetos en la misma orientación antes de calcular la característica. Los descriptores pueden ser dividido en dos categorías:

- **Características de forma basadas en el contorno del objeto**
    - *Espectro de curvatura (Image Shape Spectrum)* [182]: Consiste en un histograma calculado a partir del índice de curvatura.
    - *Histograma de normales* [183]: El conjunto de 2-simplices (triángulos) de la parte externa del complejo alfa define una triangulación de la superficie de la macromolécula. Para cada triángulo es posible definir una normal al mismo.

A continuación, se construye un histograma a partir del ángulo que forma la normal con los dos ejes principales del objeto.

- **Características basadas en el volumen encerrado dentro de la superficie.**
  - *Proporción entre los ejes principales* [184]: Sea $(\lambda_0, \lambda_1, \lambda_2)$ los tres autovalores calculados a partir de la matriz de inercia del objeto. Están ordenados por magnitud, de modo que $\lambda_0$ corresponde al mayor autovector y $\lambda_2$ al menor. La proporción que existe entre ellos nos da una idea de lo plano, alargado o redondeado del objeto.
  - *Correlación cruzada.* Para imágenes binarias, la correlación cruzada entre dos objetos la definimos como la diferencia normalizada entre sus áreas una vez los objetos han sido rotados/trasladados de acuerdo con sus ejes principales.
  - *Tamaño:* Medimos el largo alto y ancho de la mínima caja que contiene al objeto (esta caja está orientada de acuerdo con los ejes principales del objeto).
  - *Distribución circular de la masa.* Este descriptor fue usado en [185] con datos de alta resolución. Se construye un histograma a partir de una partición del espacio en celdas concéntricas y/o sectores radiales que parten del centro geométrico de la macromolécula. El número de puntos que cae en cada celda y/o sector es contabilizado y acumulado en la forma de histograma.

De esta forma, el algoritmo para la construcción del modelo queda descrito como sigue:

***Entrada:*** Imagen 3D (mapa de densidad) de la macromolécula (V), Segmentación del contorno del objeto (filtro de Deriche), Umbral de parecido entre el modelo y el volumen original.

***Salida:*** Complejo alfa que aproxima la geometría del volumen.

1. Obtención de la máscara binaria *MK* a partir del mapa de densidad V.
2. Inicialización de $n$=n0 como la cardinalidad inicial del conjunto de pseudo-átomos.
3. Ejecución del algoritmo de KCM para cuantificación vectorial. Llamamos $\{S_i\}$ al conjunto de $n$ puntos 3D resultante.
4. Selección de $\alpha$ como el máximo del error de cuantificación.

5. Obtención del complejo alfa $A_\alpha = C \cup \{S_i\}$ . Donde C es el conjunto de k-simplices (k=1,2,3) (conectividad asociada a $\{S_i\}$).
6. Traducción de $A_\alpha$ en un modelo discreto binario M en el espacio euclídeo 3D. M = Minner ∪ Mcontour . Minner corresponde a aquellos voxeles generados a partir de 3-simplices (tetraedros) del complejo alfa. Mcontour se obtiene a partir de los 2-simplices (triángulos) situados en la parte exterior del complejo alfa.
7. Si la topología de $A_\alpha$ es igual que la de *MK*, entonces calcular los 6 descriptores de forma y realizar la media sim(M,MK,n, α) de las medidas de similitud para cada uno de ellos comparados con el volumen original.
8. Si sim(M,MK,n, α)<umbral entonces incrementar *n* e ir al paso 3. En otro caso ir a 9 (la topología y geometría satisface el grado de similitud establecido por el usuario).
9. Almacenamiento del complejo alfa $A_\alpha$ como el modelo de aproximación de la macromolécula original. Los descriptores de forma calculados previamente son almacenados también para la facilitar la posterior búsqueda por contenido en la base de datos.

La figura 10.4 muestra de manera esquemática la esencia de este algoritmo.

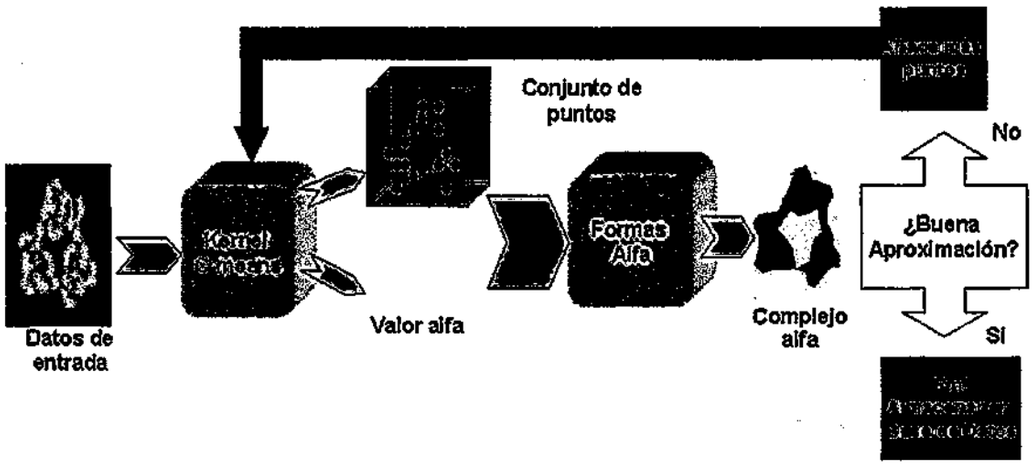



**Figura 10.4** Esquema del algoritmo para la creación de modelos de imágenes 3D utilizando cuantificación vectorial y alfa-formas.

## 10.4. Aplicación a imágenes de macromoléculas biológicas

En este apartado mostraremos la aplicación del algoritmo anterior en dos aplicaciones biológicas reales: la primera será la representación compacta de la macromolécula descrita en una imagen tridimensional a través de una aproximación a su geometría, mientras que la segunda consistirá en la segmentación y calculo automático de medidas de cavidades y canales.

Como ya hemos introducido, el incremento de información estructural producida en el contexto de estructuras resueltas a media resolución no cesa. Además, su complejidad hace inviable la anotación manual. Así pues, la posibilidad de aislar y medir automáticamente partes locales de la macromolécula con métodos como el descrito en esta sección es de una especial importancia en biología estructural. En este contexto se realizaron varios experimentos para demostrar la viabilidad y eficiencia del método propuesto, intentando cubrir al mismo tiempo diversas posibles aplicaciones.

Como primer ejemplo se utilizó la información cristalográfica disponible de la Bleomicin hidrolasa para generar mapas a distintas resoluciones con el objetivo de poder comparar los resultados para un mismo volumen en un contexto multiresolución. La bleomicin hidrolasa [186] (número de acceso pdb: 1GCB) es una cistein-proteasa que hidroliza el agente anticáncer bleomicin. Para este experimento usamos su homólogo en levadura Gal6. La estructura de Gal6 fue resuelta por cristalografía de rayos X a 2.2 Ángstrom de resolución (figura 10.5). Gal6 presenta una estructura hexamérica con un prominente canal central. Las dimensiones globales del hexámero son 125 x 125 x 85 Ángstrom. Debido a la gran interacción existente entre los dímeros, el hexámero puede considerarse como un trímero de dímeros. Los sitios activos están situados dentro del canal central. El tamaño y la forma, así como el potencial electroestático positivo de este canal, sugieren que representa la zona de interacción con el DNA.

A partir de la estructura de rayos X generamos mapas a distintas resoluciones: 7, 10, 20, 30 Ángstrom usando un espaciado de 3.409 Ángstrom /pixel. El objetivo de este experimento era analizar el impacto de la cardinalidad del conjunto de pseudo-átomos obtenidos por el algoritmo de cuantificación vectorial en el modelo final.

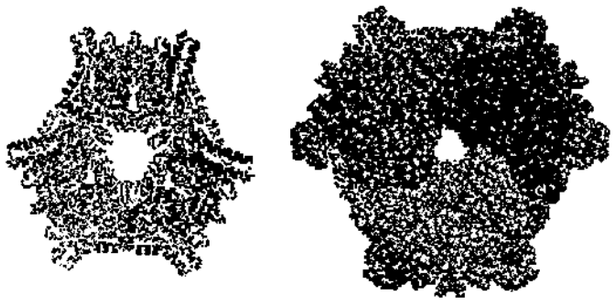


**Figura 10.5** Representación de cinta (ribbons) y de ocupación espacial de la estructura atómica de Gal6.

Para cada volumen construimos complejos alfa con 1000 y 1500 pseudo-átomos. Los resultados muestran que la representación final de estos complejos es siempre compacta y nunca excede unas cuantas decenas de kilobytes (mínimo de 41496 bytes para 7 Ángstrom de resolución y máximo de 64604 bytes para 30 Ángstrom). Adicionalmente, la similitud del modelo aproximado con el original para las distintas resoluciones muestra, tal y como se esperaba, que un incremento en el número de pseudo-átomos se traduce en una mejora de la precisión de la representación. Sin embargo, los resultados sugieren que un incremento de 1000 a 1500 pseudo-átomos no mejora de forma muy acusada la precisión en la representación. Este hecho revela el hecho de que existe un punto a partir del cual la bondad de la aproximación se mantiene estable.

Teniendo en cuenta las evidencias anteriores se diseñó un nuevo experimento enfocado a la búsqueda de un número óptimo de pseudo-átomos que refleje el compromiso entre fidelidad de la representación y compactación de la misma. Para esto realizamos una ejecución completa del algoritmo para el mapa de Gal6 a 20 Ángstrom de resolución.. La figura 10.6 muestra una gráfica de la variación de la correlación cruzada entre el complejo alfa y el mapa original conforme varia el número de pseudo-átomos utilizado. Esta relación puede interpolarse claramente mediante una función logarítmica. Tal y como se esperaba, el coeficiente de correlación no mejora significativamente a partir de un número aproximado de 600 pseudo-átomos.

Como segunda aplicación para demostrar la eficacia de este algoritmo de representación de formas se utilizó como ejemplo la segmentación y cálculo automático de medidas de cavidades y canales presente en las macromoléculas. En este caso escogimos una reconstrucción real obtenida mediante criomicroscopía electrónica del complejo DnaB-DnaC (helicasas replicativas). Esta estructura 3D del hexámero de

DnaB en complejo con DnaC ha sido resuelta por microscopía electrónica muy recientemente [115]. La resolución alcanzada usando técnicas de criomicroscopía y procesamiento de imagen fue de 26 Ángstrom. Este complejo macromolecular presenta una forma toroidal (ver figura 10.7), con un canal central que tiende a cerrarse en uno de los extremos del volumen. Su diámetro máximo es de 13.8 nanómetros y la altura 12.4 nm. Esta reconstrucción también permite identificar las partes correspondiente a DnaB y a DnaC así como las correspondientes superficies de contacto (interfaces) entre ellas.

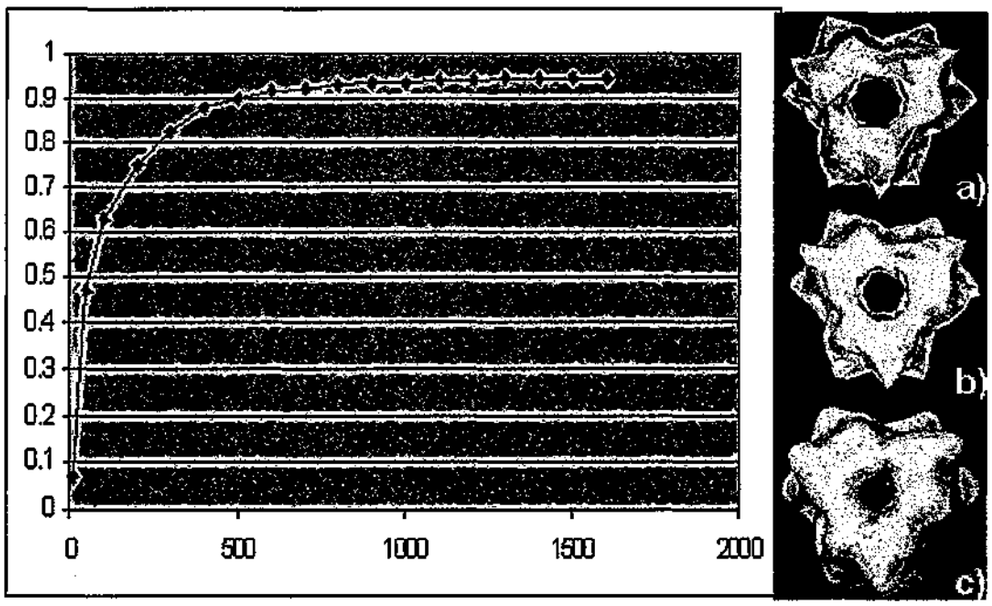



**Figura 10.6** Evolución de la correlación cruzada entre el complejo alfa y el modelo original de Gal6 conforme se incremente el número de pseudo-átomos (eje X). La relación puede ser interpolada por una función logarítmica. Con 600 pseudo-átomos parece alcanzarse una buena aproximación. A la derecha se muestran los complejos alfa con 600 (a) y 1500 (b) pseudo-átomos. La correlación cruzada no varia significativamente entre los dos modelos. La superficie del volumen original se muestra en c).

**Figura 10.7** Representación de isosuperficie del complejo DnaBC. Están representados el 100% (malla) y el 50% de la masa.

En la figura 10.8, mostramos los complejos alfa de aproximación a lo largo de varias etapas del algoritmo: con 200, 600,1000 y 1500 pseudo-átomos. Tal y como se muestra en las figuras la forma del modelo con 1500 pseudo-átomos es prácticamente idéntica a la original. Sin embargo, alguna de las características mencionadas anteriormente ya pueden apreciarse con 600 pseudo-átomos (ver figura 10.8). Nosotros sugerimos que 1000 pseudo-átomos es el mejor compromiso entre la eficiencia de la representación y su precisión, sin embargo esta decisión depende de las particularidades de cada aplicación y constituye un tema de estudio abierto en la actualidad.

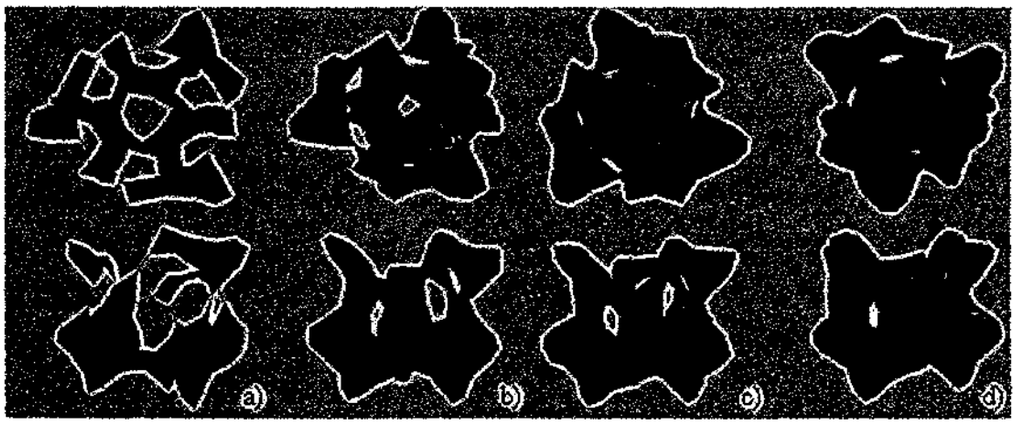



**Figura 10.8** Vistas frontales y laterales de los complejos alfa para DnaBC con 200 (a), 600 (b), 1000(c) y 1500(d) pseudo-átomos.

Una vez construido el complejo alfa, resulta inmediato aislar y realizar medidas sobre aquellos simplices que ocupan el espacio de cavidades internas y canales así como otras características estructurales como pueden ser protuberancias en la superficie. La figura 10.9 muestra la cavidad abierta de la estructura de la DnaBC, la cual ha sido asilada y su volumen medido en 15329 Amgstron$^3$.

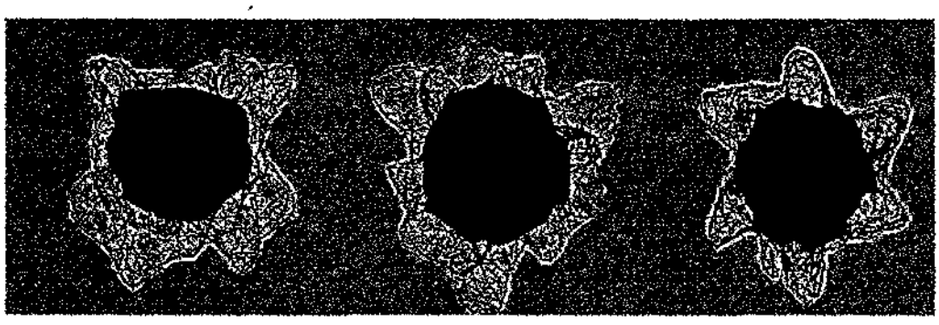


**Figura 10.9** Varias vistas de la cavidad abierta del complejo DnaBC. El volumen medido automáticamente para dicha cavidad es de 15329 Amgstron$^3$.

Estos experimentos expuestos anteriormente constituyen una demostración práctica de la eficacia y utilidad de este nuevo algoritmo de análisis de formas en imágenes 3D. Es primera vez que se propone la aplicación de la teoría de alfa formas en el contexto de imágenes tridimensionales de densidad electrónica, constituyendo una poderosa metodología con amplia aplicación en biología estructural, al abrir un nuevo camino en las tareas de extracción automática de conocimiento en bases de datos de macromoléculas biológicas. La técnica de cuantificación vectorial a través del algoritmo de KCM en combinación con la teoría de alfa formas constituyen un excelente binomio para la codificación de forma y topología en imágenes tridimensionales en general.

## 11. Análisis de datos de expresión génica

En esta sección se describen los resultados que se han obtenido tras aplicar los procedimientos de análisis exploratorio de datos presentados en los capítulos anteriores en un nuevo campo de aplicación de la biología molecular: el análisis de la expresión génica. En esta memoria hemos presentado varias aplicaciones relacionadas con el análisis de datos estructurales de baja y media resolución provenientes de técnicas como la microscopía electrónica. Sin embargo, en esta nueva aplicación describiremos un tipo de aplicación completamente diferente aplicando una nueva técnica experimental que ha provocado un cambio radical en el campo de la genómica funcional: la técnica de los microchips de ADN.

La esencial novedad de esta técnica está basada en su capacidad de medir la expresión de decenas de miles de genes simultáneamente frente a distintas condiciones experimentales [4]. El conjunto de genes de un organismo se comporta de forma totalmente distinta según las condiciones a que está sometida cada célula, el tejido del que forma parte, y el momento concreto del ciclo celular en que se encuentra. Esto hace que cada célula fabrique un conjunto específico de proteínas para cada situación que rige su comportamiento. Si bien hasta hace pocos años esta actividad se tenía que analizar de forma individual para cada gen, los microchips de DNA permiten observar de forma global que genes son más o menos activos (cuales están codificando proteínas y en que medida) en cada una de las situaciones de la célula, situación que hace solo una década atrás era algo completamente impensable. Sin embargo, este súbito desarrollo tecnológico está generando volúmenes de datos de varios ordenes de magnitud mayor de los que se venían manejando con los métodos existentes, con lo cual el almacenamiento, la manipulación y el análisis de esta nueva información se han convertido a su vez en el mayor cuello de botella para la utilización de este tipo de tecnología. Es por ello que muchos investigadores en los campos de análisis y minería de datos están dedicando grandes esfuerzos a la obtención de técnicas robustas capaces de analizar eficientemente toda esta cantidad de información [4, 187-192].

A pesar de la extensa batería de técnicas y metodologías utilizadas para analizar datos de expresión génica de microchips, no existe una única técnica óptima capaz de manipular estas grandes cantidades de datos de manera eficiente y que responda a todas

las preguntas biológicas que sobre estos datos se puede formular, es por eso que la investigación en este campo todavía está en su infancia [193].

En este apartado pretendemos introducir brevemente los fundamentos teóricos de la genética molecular, así como los aspectos teóricos y prácticos de la tecnología de los microchips de ADN. Posteriormente presentaremos una pequeña revisión de los métodos bioinformáticos más utilizados para el análisis de los datos que esta técnica produce y finalmente presentaremos la aplicación de los algoritmos objetos de estudio de esta memoria sobre este conjunto de datos con el objetivo de demostrar su capacidad de proporcionar nueva y relevante información.

## 11.1. Breve introducción a la genética molecular

La unidad básica de los organismos vivos es una maquinaria muy compleja llamada célula, de las cuales están constituidos todos los organismos superiores. Estas células presentan típicamente, un núcleo en el cual se almacena toda la información necesaria para su funcionamiento y para la creación de otro ser igual. Esta información almacenada dentro de los núcleos de las células se encuentra en forma de ADN (Ácido desoxiribonucleico).

El ADN está constituido por dos secuencias muy largas procedentes de la combinación de cuatro moléculas llamadas nucleótidos (bases): Adenina (A), Guanina, (G) Citosina (C) y Timina (T). Las dos hebras que forma esta molécula de ADN se unen mediante enlaces débiles de hidrógeno entre pares de base, creando una forma parecida a la de una escalera torcida. En circunstancias normales la adenina solo se puede unir con la timina y la citosina solo puede formar un enlace con la guanina. Es por eso que se usualmente se dice que las dos hebras se complementan entre sí (ver figura 11.1). Dada la gran cantidad de información que se debe almacenar en el ADN, este no se encuentra en forma lineal sino se compacta de manera muy eficiente formando los llamados cromosomas.

La unidad básica y funcional de la herencia es conocida como gen, el cual no es más que una secuencia específica de bases de nucleótidos que porta la información necesaria para la construcción de las proteínas. Estas moléculas son cadenas de polímeros constituidos por aminoácidos y son las que forman los componentes

estructurales de las células y los tejidos, así como enzimas necesarias para reacciones bioquímicas esenciales en los organismos.

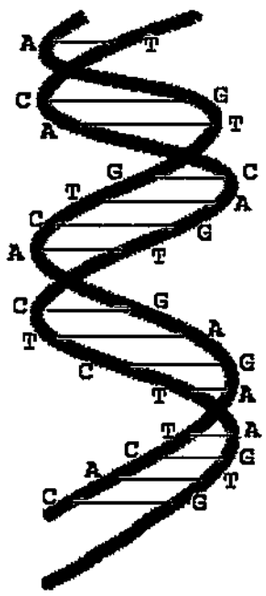



**Figura 11.1** Estructura de la doble hélice de ADN.

Para poder llevar acabo las síntesis de estos polímeros la información contenida en los genes debe llegar la exterior del núcleo de la célula, al citoplasma, que es donde están los orgánulos en los que se produce el ensamblaje de los aminoácidos para constituir las proteínas. Pero el ADN no puede salir al citoplasma porque si lo hiciese sería degradado. Por lo tanto para transmitir la información necesaria para la síntesis de proteínas se necesita de una molécula capaz de almacenar la información y transportarla desde el núcleo hasta el citoplasma para la síntesis proteica. La molécula encargada de llevar a cabo esta función es el ARN mensajero. El ARN (Ácido Ribonucleico) es muy parecido al ADN, también constituido por cuatro bases que son Adenina (A), Guanina (G), Citosina (C) y Uracilo (U). El ARN mensajero es una copia en ARN de una de las cadenas de ADN que sirve como molde, de esta forma se logra transmitir la información contenida en los genes a una molécula capaz de salir del núcleo y que va a servir para la síntesis de proteínas en unos orgánulos celulares llamados ribosomas. A este proceso de transformación ADN-ARN-Proteína se le conoce como "El Dogma Central de la Biología Molecular". Buscando un símil con las ciencias computacionales, se podría comparar el ADN con el código fuente presente en los ordenadores, mientras las

proteínas serían los programas ejecutables. La figura 11.2 muestra una representación esquemática de este proceso de transformación ADN-ARN-Proteína.

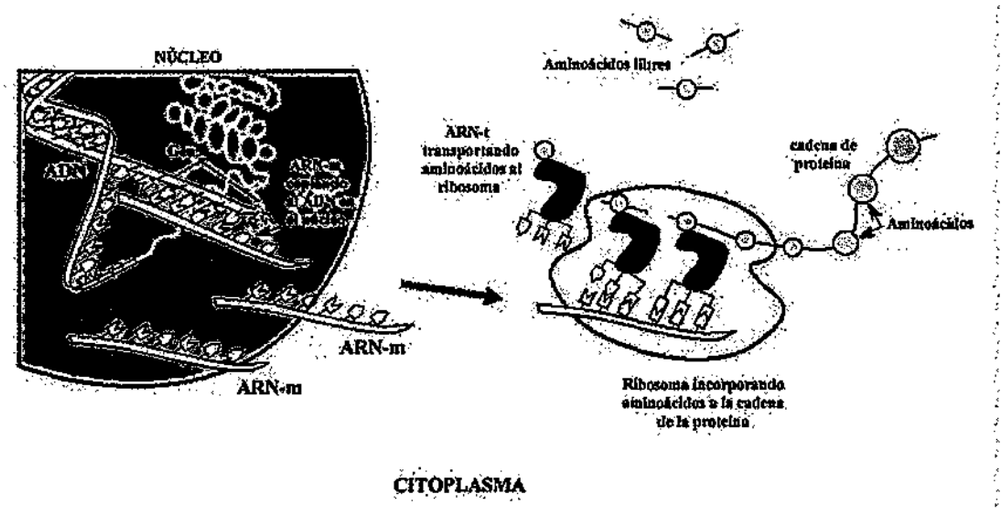



**Figura 11.2** Esquema del proceso de transformación ADN-ARN-Proteína. Cuando los genes se expresan la información genética del ADN (secuencias de bases) es transcrita (copiada) al ARN mensajero en un proceso muy similar al ocurrido en la replicación del ADN. Estas moléculas de ARN abandonan el núcleo y salen al citoplasma donde ocurre el proceso de traducción que es llevado a cabo por el ribosoma, el cual lee el código genético contenido en el RNA mensajero en forma de tripletas de bases llamadas codones que codifican a un aminoácido específico. El ribosoma, con la información del RNA mensajero y del RNA de transferencia, crea las cadenas de aminoácidos que forman a la proteína.

La rama de la biología dedicada al estudio de los anteriores procesos se la conoce como genética molecular, y se encarga de estudiar los mecanismos mediante los cuales son realizados estos pasos. La Genómica ha surgido como una evolución de la genética molecular fundamentada en las nuevas tecnologías y en las nuevas aproximaciones de estudios masivos que estas permiten. La palabra Genómica surge de fusión entre gen y el sufijo "–omica" que significa conjunto. Las técnicas de microchips de DNA es un ejemplo claro de este tipo de estudios masivos y está basada directamente en varios de los conceptos aquí introducidos.

## 11.2. Introducción a las técnicas de microchips de ADN.

El proceso de medición de la expresión de un gen no es un concepto nuevo y de hecho se ha venido utilizando durante muchísimos años sobre la base de "un gen, un experimento". La novedad de la técnica que aquí se presenta es fundamentalmente

tecnológica, permitiendo la medición de expresión de miles de genes simultáneamente en la misma condición experimental. La figura 11.3 muestra un esquema descriptivo del proceso.

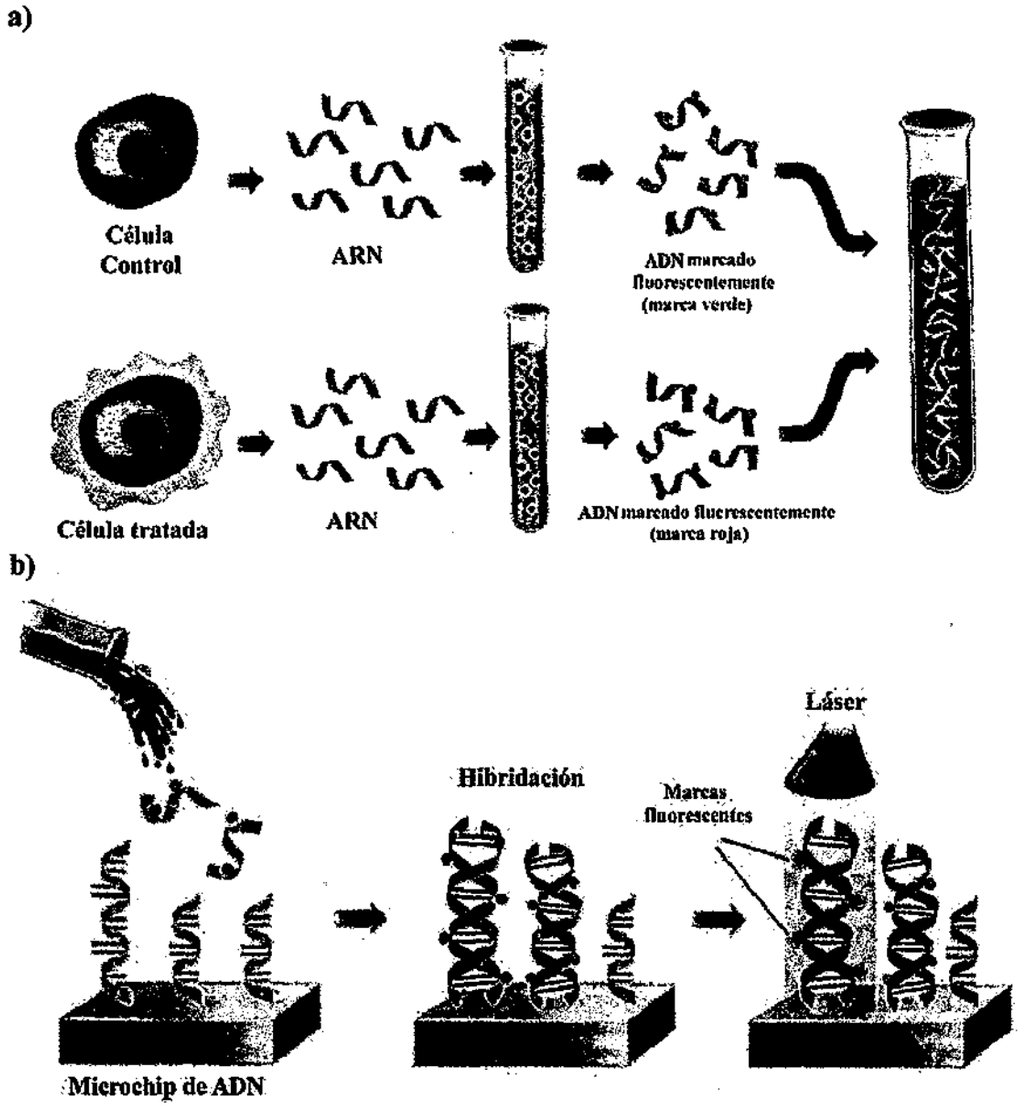



**Figura 11.3** Esquema representativo de la técnica de microchip de ADN. La expresión de los genes en una célula normal (células control) es comparada con la expresión de los genes en una célula sometida a una condición experimental. a) Cuando las células comienzan a sintetizar proteínas, aparece el proceso de transcripción de los genes para el cual múltiples copias de ARN mensajero salen al citoplasma. El ARN es extraído y por un proceso de transcripción inversa es convertido a ADN de simple cadena (llamado ADN complementario) y marcado con una sustancia fluorescente (verde para la célula control y roja para la célula tratada). b) El ADN de simple cadena marcado se pone en contacto con el microchip, el cual contiene ADN de simples cadena de genes conocidos y en posiciones conocidas. Por un proceso llamado hibridación, aquellas cadenas de ADN de los genes expresados en las células, tenderán a unirse a su copia complementaria en el microchip. Una vez terminado este proceso, el microchip es sometido a un estímulo luminoso para provocar la fluorescencia y medir las marcas de las cadenas de ADN unidas al chip.

Básicamente un microchip consiste en una matriz de pocillos miniaturizados sobre un substrato de vidrio en donde se implantan, utilizando diversas técnicas, cadenas simples de oligonucleótidos. El procedimiento de fabricación de un microchip empieza con la preparación de muestras de ADN de cada uno de los genes que se quieren analizar de un organismo. Un robot se encarga de la deposición y fijación de cantidades muy pequeñas de cada una de estas muestras sobre un substrato de cristal, de tal manera que la muestra correspondiente a cada gen ocupa un lugar específico en una minúscula matriz. El proceso requiere mucha precisión y la ausencia total de polvo y contaminantes, ya que la separación entre los depósitos de ADN que corresponden a cada gen son del orden de entre 200 y 300 micrómetros (milésimas de milímetro). Una vez fabricado el microchip, los investigadores lo ponen en contacto con ADN que se ha generado in vitro a partir de preparaciones de ARN mensajero proveniente de las células del organismo que se quiere estudiar. Las cadenas de ADN generadas, que corresponden solo a aquellos genes que son activos bajo las condiciones en que se encuentran las células, se combinan solo con los depósitos complementarios de la matriz (proceso de hibridación). Los puntos de la matriz que se han combinado destacan sobre un fondo oscuro al ser excitados mediante un detector confocal de fluorescencia y procesados por un sistema informático que permite su interpretación. Así, las coordenadas de los puntos brillantes de la matriz informan sobre que genes son activos en las células analizadas y en que medida se han activado.

Uno de los objetivos principales de los estudios basados en esta tecnología está relacionado con la estimación de la activación de las proteínas en ciertas condiciones, sin embargo, es evidente que estas técnicas lo que miden el nivel de ARN mensajero producido durante el proceso de transcripción en las células y no el nivel de proteínas producidas. No obstante, aunque el ARN mensajero no es el producto final de un gen, la transcripción es el primer paso en el proceso de producción de las proteínas. Es importante destacar que aunque la correlación entre el nivel de ARN mensajero y el nivel de abundancia de las proteínas en las células no sigue una relación directa, es altamente probable que una ausencia de ARN mensajero en un proceso celular implique niveles muy bajos de abundancia de la respectiva proteína que ese gen codifica y por lo tanto los niveles cuantitativos del proteoma pueden ser estudiados a partir de la información del *transcriptoma* [189].

Actualmente existen muchas variedades de fabricación de microchips, desde los artesanales realizados en el laboratorio a los ofrecidos por compañías de biotecnología. Sin embargo, el principio que rige su diseño es el mismo: la complementariedad de las cadenas aisladas de ADN. Actualmente podría decirse que son dos las tecnologías de chips que imperan en el mercado: la técnica de cDNA (ADN complementario), desarrollada por Pat Brown de la Universidad de Stanford [4] y la técnica de oligonucleótidos, también conocida como Affymetrix, por ser esta la empresa que los comercializa (www.affymetrix.com). La fabricación de ambos tipos de chips difieren significativamente en cuanto a la tecnología que utilizan para fijar la secuencia de nucleótidos de los genes en los chips.

La técnica de cDNA es ampliamente utilizada por su relativo bajo costo y su flexibilidad para la creación de chips "a la carta". Esta técnica utiliza un brazo robotizado para fijar al substrato la secuencia del gen que corresponda. Usando impresión por contacto (contact printing) o impresión por inyección (inkjet) la solución con miles de copias de la secuencia de ADN es fijada en la superficie de contacto [194]. La figura 11.4 muestra un esquema representativo de esta técnica.

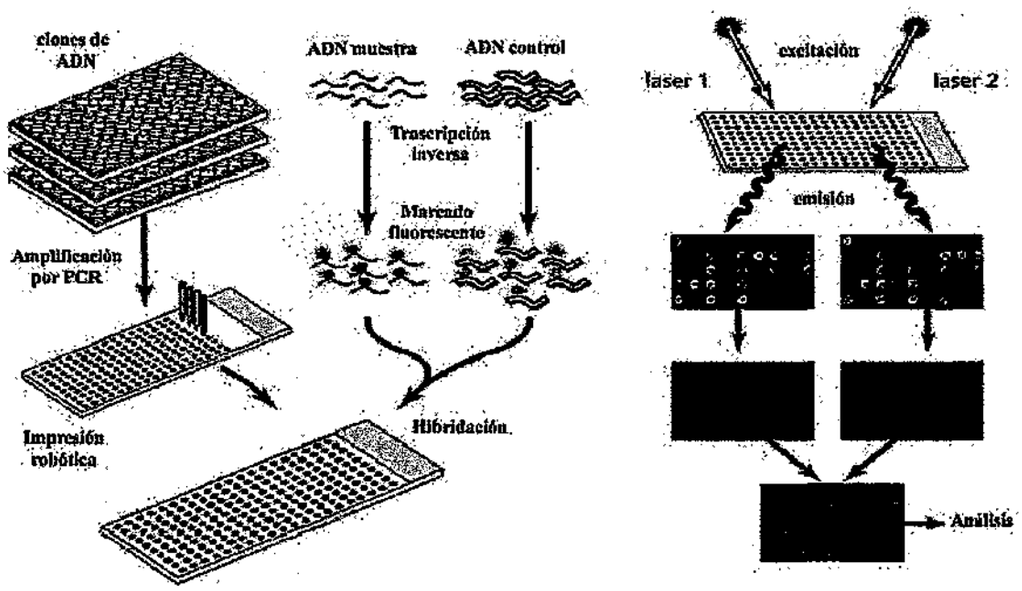



**Figura 11.4** Representación esquemática del método de cDNA.

Es importante señalar que con los chips de cDNA lo que se mide es la abundancia relativa de ARN mensajero en dos muestras diferentes (generalmente

control y muestra), por lo que los valores son expresados como la razón de expresión de la muestra sobre el control (rojo/verde). Si el color del pocillo resultante es verde, implica que solo se ha expresado ese gen en la célula control. Si por el contrario el color resultante es el rojo, implica la expresión únicamente en la muestra. Así mismo, si el color es de tonalidad amarilla, eso implica una mezcla de abundancia del gen en ambas condiciones (control y muestra). El color negro es indicativo de que el gen no se ha expresado en ninguna de las dos condiciones experimentales.

La técnica de Affymetrix difiere significativamente en su proceso de fabricación. Esta sintetiza en las propias celdas del chip, mediante ciclos sucesivos de síntesis, los oligonucleótidos correspondientes mediante un método de fotolitografía muy similar al utilizado en la industria de semiconductores. La figura 11.5 muestra una representación de este proceso.

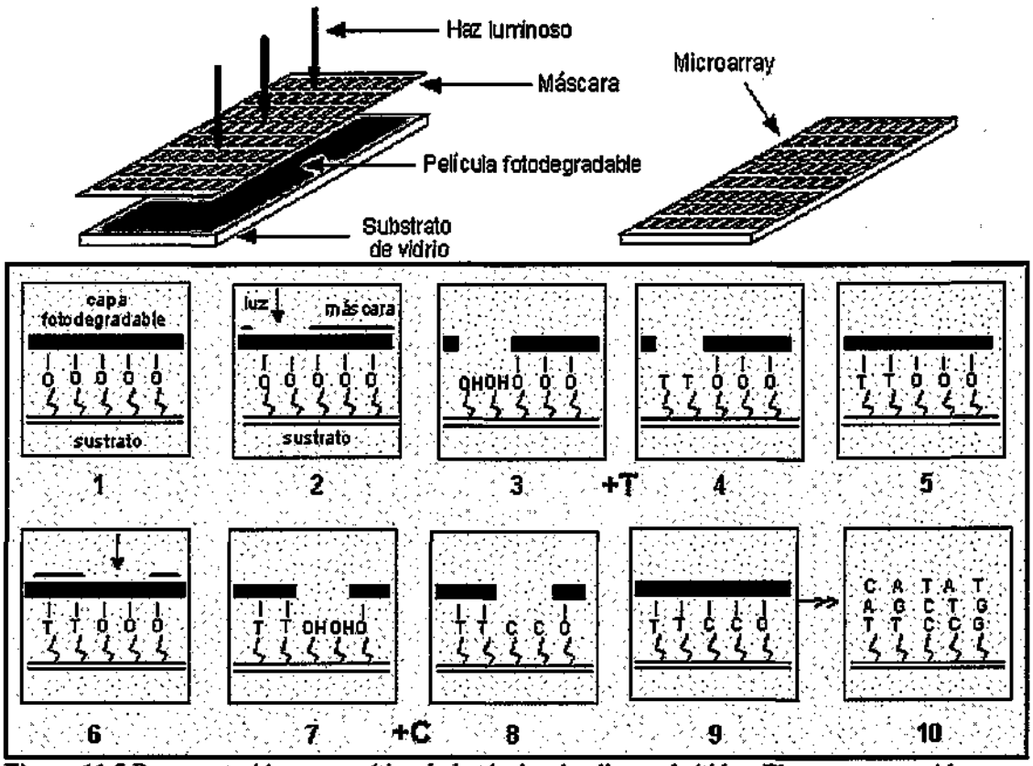



**Figura 11.5** Representación esquemática de la técnica de oligonucleótidos. El proceso conocido como fotolitografía permite la fijación "in situ" de la secuencia de nucleótidos uno a uno en el substrato. El ejemplo muestra la adición de bases de Timina y Citosina en posiciones específicas del chip.

El proceso de síntesis "in situ" de los nucleótidos en el chip se realiza en las siguientes fases:

- Elaboración previa de un mapa de distribución del tipo de oligonucleótido correspondiente a cada celda del chip.
- Preparación del sustrato y deposición de una película fotodegradable.
- Aplicación de una máscara que permita eliminar selectivamente la película protectora en las zonas del chip correspondiente a un determinado nucleótido.
- Incubación química y acoplamiento del nucleótido previsto.
- Una nueva capa fotodegradable es aplicada sobre el chip.
- Se repiten los pasos anteriores para cada nucleótido hasta obtener la secuencia prevista .
- Se elimina definitivamente la película fotodegradable.

Este método utilizado por Affymetrix tiene un nivel de integración mucho más elevado que el de la técnica de cDNA, ofreciendo la mayor densidad de sondas (secuencias) por unidad de área. Adicionalmente permite la síntesis de ologonucleótidos de hasta 25 bases. En los chip producido mediante esta tecnología se incluyen, para cada gen, unos 20 oligonucleótidos llamados de correspondencia perfecta (perfect match) que son trozos de ciertas partes del gen de unas 25 bases y otros 20 oligonucleótidos llamados de incongruencias (mismatch) que son los mismos trozos utilizados en la correspondencia perfecta pero alterándole en valor de una de las bases. La utilización de múltiples oligos favorece el aumento de la relación señal-ruido en las mediciones y también permite que las hibridaciones cruzadas entre genes parecidos sean detectadas y tratadas adecuadamente. Estas características hacen que esta técnica, a pesar de ser mucho mas cara, sea mucho más precisa e informativa que la de cDNA.

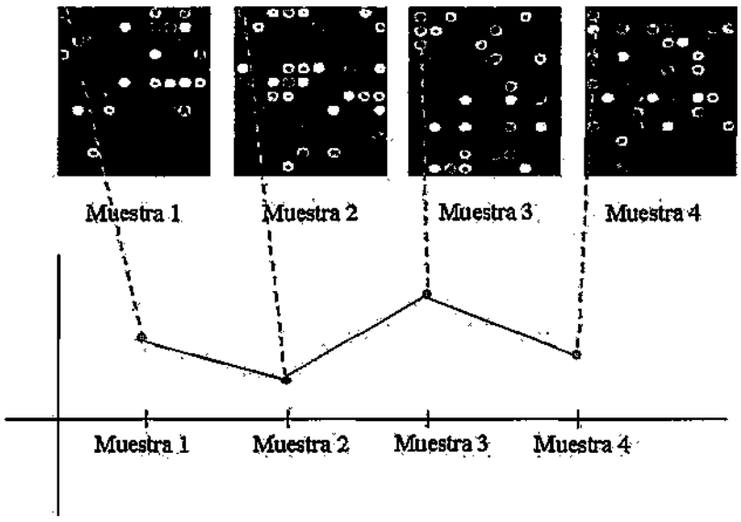
## 11.3. Análisis de expresión génica

El tipo de datos producidos por las técnicas descritas anteriormente inicialmente están formados por grandes imágenes monocromas formadas por puntos brillantes que expresan la expresión de los genes utilizados en cada chip. El primer paso en el análisis de esta información es convertir estas imágenes a matrices de valores de expresión. Este proceso se realiza con técnicas clásicas de procesamiento de imágenes. En primer lugar los puntos brillantes en la imagen deben de ser identificados y segmentados y su nivel de intensidad medido y comparado con la intensidad del fondo y con la intensidad de

los demás canales medidos para el mismo experimento. El objetivo es obtener una matriz de niveles de expresión normalizada y confiable, eliminando el efecto producido por los artefactos intrínsecos del proceso de adquisición de las imágenes.

Una vez obtenida esta matriz de expresión, en la cual cada fila generalmente se corresponde con un gen (o punto en la imagen) y cada columna se corresponde con una condición experimental (o valor del mismo punto en los distintos chips que forman el experimento) se puede proceder al análisis de los niveles de expresión (ver figura 11.6).

**a)**




**b)**

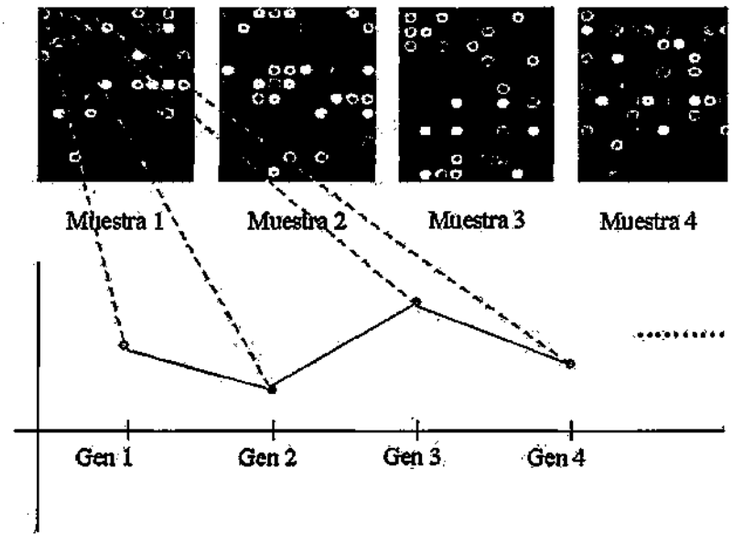



**Figura 11.6** Esquema de obtención de la matriz de expresión a partir de las imágenes de los chips de ADN. a) Análisis de niveles de expresión de los genes: cada gen (punto en la imagen) forma un vector (fila) y cada chip (muestra) es una componente del vector. b) Análisis fenotípico: cada muestra o chip forma un vector y cada punto del chip (gen) es una componente de este vector.

En este contexto este análisis podría responder preguntas biológicas como las siguientes:

- ¿Cuál es el papel funcional de un grupo de genes y en qué procesos celulares participan?
- ¿De que manera están los genes regulados?
- ¿Cómo interaccionan los genes y sus productos?
- ¿Cómo difiere la expresión de los genes en distintos tipos de células y en presencia de distintas condiciones, como puede ser enfermedades o tratamiento con medicamentos?
- ¿El cambio de expresión de un gen está correlacionado con condiciones externas a la célula?
- ¿En cuánto ha cambiado los niveles de expresión de un gen en distintas condiciones experimentales?
- ¿Se puede predecir la función de un gen desconocido a partir de las funciones de otros genes con similar nivel de expresión?

Para este tipo de análisis generalmente se utilizan dos formas posibles de estudio:

- La comparación de los perfiles de expresión de los genes comparando las filas de la matriz de expresión.
- La comparación de los perfiles de expresión de las muestras comparando las columnas de la matriz.

Por ejemplo, si utilizando este tipo de análisis encontramos que dos filas son similares, se podría plantear la hipótesis de que los respectivos genes son co-regulados y posiblemente relacionados a nivel funcional. Adicionalmente, en el caso de comparar muestras podemos encontrar qué genes están expresados diferencialmente y estudiar el efecto de expresión de distintos componentes químicos sobre las células (fármacogenómica).

Este tipo de comparación global de genes o muestras es posible llevarlo a cabo a través de métodos de agrupamiento como los mencionados en este memoria. El objetivo final es agrupar aquellos genes o muestras que tengan propiedades de expresión

similares. Sin embargo, antes de comenzar el proceso de análisis, los datos deben ser preprocesados convenientemente para poder ser utilizados por los algoritmos de agrupamiento. Los métodos de procesamiento más comúnmente utilizados en el análisis de perfiles de expresión son los siguientes:

***Transformación logarítmica:*** Los datos de expresión generalmente muestran distribuciones asimétricas respecto a la expresión o inhibición, lo cual dificulta el uso de medidas de distancias para establecer diferencias entre ellos. Para compensar estas diferencias, se utiliza generalmente la transformación logarítmica. Por ejemplo, en cDNA, genes expresados ocupan la escala de 1 a infinito (o al menos 1000-fold), pero los genes inhibidos ocupan solamente la escala de 0 a 1. La transformación logarítmica (usualmente de base 2) hace la escala simétrica alrededor del cero. Por ejemplo, supongamos que los valores de expresión del control y la muestra son los mostrados:

| **Muestra/Control:** | Logaritmo base 2: |
|---|---|
| 100/1 = 100 | 2 |
| 10/1 = 10 | 1 |
| 1/1 = 1 | 0 |
| 1/10 = 0.1 | -1 |
| 1/100 = 0.01 | -2 |

***Filtros:*** Los chips de ADN usualmente contienen miles de genes y en muchos casos, especialmente en chips comerciales, no todos los genes están relacionados con los procesos que se miden. Esto provoca que una gran cantidad de genes no se expresen ni se inhiban en ninguna de las condiciones testadas (perfiles planos) y por lo tanto, estos genes no son interesantes para el análisis y deben ser eliminados. Los filtros permiten la eliminación de aquellos genes o muestras que no sufran ninguna variación estadísticamente significativa en todas las condiciones medidas.

***Datos incompletos:*** Debido al hecho de que los niveles de expresión en estas técnicas son obtenidos a partir de imágenes, estas a pueden contener zonas dañadas por diversos motivos: insuficiente resolución, contaminación con polvo, zonas deterioradas, etc. A nivel de la matriz de expresión obtenida de estas imágenes, esto se traduce en valores que no existen. Este tipo de valores no definidos usualmente son excluidos del análisis o marcados para que los algoritmos no los tengan en cuenta. Sin embargo, existen métodos más sofisticados para estimar lo mejor posible los valores de expresión perdidos, como pueden ser el método de los K vecinos más cercanos ponderado (weighted KNN) [195].

***Normalización:*** Unos de los pre-procesamientos más comunes en análisis de datos es la normalización. La idea de la normalización en los datos de expresión es intentar transformar los datos de manera que los valores de expresión en distintas condiciones experimentales sean comparables entre sí. La normalización puede ser por filas o columnas. Lo más común es normalizar cada columna (variable) a media cero y desviación estándar uno y normalizar la longitud del vector que forma cada fila a uno (dividiendo cada componente del vector por su norma). Esto permite que la medida de distancia utilizada por los métodos de agrupamiento no se vea afectada por la magnitud de los vectores y de esta forma se le de más importancia a la tendencia de variación de los perfiles de expresión que al valor absoluto de la misma [189].

Una vez procesados los perfiles de expresión los datos se encuentran listos para ser estudiados adecuadamente por los distintos métodos de clasificación, tanto supervisados como no supervisados, en dependencia de la pregunta biológica que se quiera responder. Los métodos supervisados asumen que para cada caso de estudio (gen o muestra) existe una información adicional que permita crear clasificadores basados en ella. Un ejemplo de este tipo de información puede ser las clases funcionales de los genes, o atributos de las muestras (muestra normal, tipos de tumores, enfermedades, etc.). Los clasificadores supervisados tienen como objetivo aprender la distribución de clases de estos datos para poder predecir la clase a la que pertenece un dato nuevo no clasificado. Por ejemplo, Brown y colaboradores [196] aplicaron varias técnicas de clasificación supervisada, entre ellas la llamada "Máquinas de Vectores Soporte" (Support Vector Machines, SVM) [63], sobre datos de expresión de 6 grupos funcionales de genes de levadura medidos en 79 muestras intentando crear un clasificador que aprendiera a diferenciar aquellas clases funcionales que están co-reguladas. Golub y colaboradores [197], aplicaron también métodos clasificados para construir clasificadores para la identificación de dos tipos de leucemias (acute myeloid y acute lymphoblastic) a partir de la expresión de un conjunto de 50 genes y 38 muestras de este tipo de leucemias.

Los métodos no supervisados, por el contrario, tienen como uno de sus objetivos el agrupamiento de genes o muestras con propiedades de expresión similares, sin tener el cuenta ningún otro tipo de información definida a priori. Básicamente los métodos más utilizados en estas aplicaciones son las clásicas técnicas de agrupamiento como las

descritas en las distintas aplicaciones presentadas en esta memoria. Entre los más utilizados podemos señalar la clasificación jerárquica ascendente (HAC) [188, 198, 199], el método de K-medias [200] y los mapas auto-organizativos de Kohonen [187, 201]. Otras técnicas menos conocidas, pero igualmente eficaces como son las técnicas basadas en grafos también han sido propuestas y utilizadas en este contexto [202, 203]. Para ver una revisión más extensa sobre la gran cantidad de trabajos publicados en el contexto de agrupamiento de perfiles de expresión de genes se puede consultar el trabajo de Brazma y Vilo [189].

En el siguiente apartado proponemos la utilización del algoritmo de KerDenSOM como una nueva metodología para el análisis exploratorio de datos de expresión génica. La principal motivación de aplicación de este método de análisis en este contexto viene dada por la pobre utilización que hasta ahora se le ha venido dando a los mapas auto-organizativos sobre este tipo de datos. Usualmente el método clásico de SOM es utilizado como un método de agrupamiento particional, utilizando cada vector diccionario como el centroide de un único grupo (por ejemplo, un mapa de 3x3 es utilizado para extraer 9 grupos independientes). De esta manera los mapas auto-organizativos no son utilizados en su mayor potencial, ya que las bondades de las propiedades de mapeo suave y ordenado no son explotadas. La propuesta de utilización de un mapa auto-organizativo de bases teóricas sólidas como lo es KerDenSOM puede ayudar en el descubrimiento de información importante oculta en los datos.

## 11.4. Un caso de estudio: análisis de la respuesta de células de la piel a la irradiación de luz ultravioleta.

La radiación ultravioleta (UV) es el agente cancerígeno más importante que se encuentra en el medio ambiente y la piel constituye su principal objetivo. La radiación UV puede provocar efectos muy dañinos y crónicos en la piel como lo son el eritema ó quemadura solar [204], el envejecimiento prematuro de la piel [205, 206] y tumores malignos [207, 208]. Recientemente se ha experimentado un incremento en la importancia y la preocupación de la comunidad científica por los efectos de las radiaciones UV debido fundamentalmente al deterioro de la capa de ozono [204] que ha provocado un incremento notable del número de casos de cáncer de piel que se han reportado [209].

La luz ultravioleta afecta la piel de distintas maneras en dependencia de su longitud de onda. Del total de las radiaciones ultravioleta que llega a la superficie de la tierra proveniente del sol un pequeño porcentaje de ellas corresponde a radiaciones con longitudes de onda entre los 290 y los 320 nanómetros y es llamada UVB, que al contrario de los conocidas radiaciones UVA, es considerado el agente causante de muchos de los efectos nocivos atribuidos a las radiaciones ultravioletas, provocando mutaciones en al ADN y modificando el patrón de expresión génica que se produce en las células de la piel como mecanismo de defensa.

La piel constituye una barrera fisiológica que protege al organismo contra los agentes patógenos, así como otro tipo de agresiones tanto físicas como químicas. Este tejido está compuesto por una capa superior llamada epidermis cuyas células principales son los queratinocitos (encargados de crear las queratinas) y una capa más profunda llamada dermis, básicamente compuesta por fibroblastos.

Con vistas a entender los mecanismos de respuesta a las radiaciones UV, en esta aplicación nos proponemos extender el estudio realizado por Sesto y colaboradores [210] sobre los perfiles de transcripción de los queratinocitos primarios humanos después de ser sometidos a radiaciones UVB. En este estudio se realizaron experimentos que incluían queratinocitos tratados con tres dosis distintas de UVB(10, 20 y 40 $mJ/cm^2$) y medidos en diferentes intervalos de tiempo (4 y 24 horas). Estas dosis representan exposiciones a las radiaciones UVB consideradas como bajas, medias y altas respectivamente. Así mismo, los tiempos de medición de los niveles de expresión se consideran como eventos regulatorios tempranos (4h) y tardíos (24h). Los experimentos fueron llevados a cabo utilizando la técnica de Affymetrix con un chip comercial (HuGeneFL) que contiene unos 6000 genes aproximadamente y todos los experimentos fueron realizados en duplicado para aumentar la fiabilidad de los resultados.

El conjunto de datos resultante del proceso de hibridación quedó compuesto por más de 6000 vectores (representando los aproximadamente 6000 genes presentes en el chip ) de 7 componentes cada uno (6 condiciones experimentales más el chip de la muestra control). A pesar de la gran cantidad de datos obtenidos por el uso de un chip comercial estándar, solamente se tenía interés en aquellos genes afectados por la radiación UVB, por lo tanto, un paso previo de filtrado fue necesario para eliminar

aquellos genes que no variaron su expresión en ninguna de las 6 condiciones experimentales con respecto a la muestra control. Utilizando este tipo de filtro clásico se mantuvieron los genes cuyos valores de expresión duplicaron al valor de su expresión en la muestra control.

Una vez aplicado el filtro, el conjunto de datos resultante quedó compuesto por 539 vectores de 7 dimensiones correspondientes a 539 genes en 7 condiciones experimentales: control, 10mJ/cm$^2$ a las 4 horas, 20mJ/cm$^2$ a las 4 horas, 40mJ/cm$^2$ a las 4 horas, 10mJ/cm$^2$ a las 24 horas, 20mJ/cm$^2$ a las 24 horas y 40mJ/cm$^2$ a las 24 horas. Posteriormente la magnitud de los vectores se normalizó a uno con el objetivo de eliminar las diferencias en magnitud entre vectores con la misma tendencia de expresión.

A este conjunto de datos se le aplicó posteriormente el algoritmo de KerDenSOM utilizando un mapa de 20x10 con topología rectangular y utilizando un núcleo Gaussiano en 200 iteraciones. El parámetro de suavidad fue variado entre 150 y 15 en 5 pasos de enfriamiento determinista. El mapa resultante puede verse en la figura 11.7.

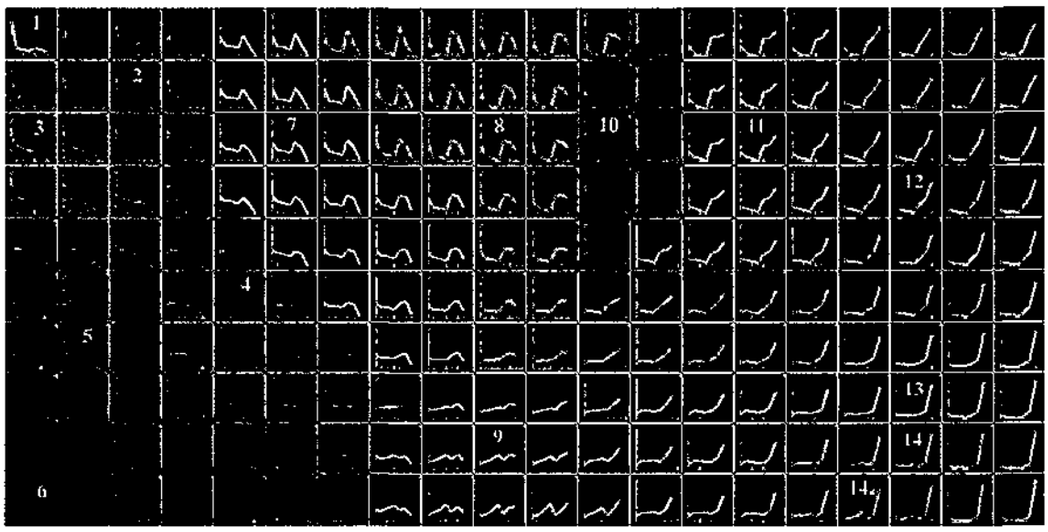



**Figura 11.7** Resultados de algoritmo de KerDenSOM aplicado a los perfiles de expresión de genes regulados por UVB. El mapa ha sido dividido y numerado en regiones homogéneas (grupos).

El mapa generado por el algoritmo de KerDenSOM fue dividido (agrupado) en 14 zonas homogéneas basado en la similitud de los vectores diccionarios y en su valor de densidad (número de perfiles de expresión originales) que estos vectores diccionarios

representan. Estos 14 grupos encontrados por KerDenSOM representan patrones de regulación muy específicos que pueden ser resumidos de la siguiente forma:

- *Grupo 1*: Este grupo está constituido por un solo vector diccionario que representa 11 genes que fueron completamente reprimidos después de la irradiación con UVB, independientemente de la dosis y del tiempo.
- *Grupos 3 y 4*: Estos grupos representan 51 genes con un patrón generalizado de represión, pero los niveles de represión entre ambos grupos difieren en dependencia de la dosis y del tiempo.
- *Grupos 2, 7, 8, 10 y 11*: Estos grupos que representan unos 190 genes muestran una característica en común: todos ellos presentan una represión de los niveles de expresión a las 4 horas, sin embargo, difieren entre sí en su comportamiento a las 24 horas, mostrando inducciones, represiones y ausencia de cambios.
- *Grupos 5 y 6*: Estos grupos incluyen 83 genes que muestran una inducción dependiente de la dosis a las 4 horas y una evidente represión a las 24 horas.
- *Grupo 9*: Contiene 23 genes que muestran como característica principal una inducción a las 4 horas que se mantiene también a las 24.
- *Grupos 12, 13 y 14*: Estos grupos engloban más de un tercio de los vectores diccionarios en el mapa y representan también una cantidad importante de genes (172 en total). Estos genes tienen como característica común una inducción a las 24 horas aunque curiosamente dependen de la dosis de radiación.

Incluido en el grupo 14 aparece un vector diccionario, denominado 14a en la figura, cuyo patrón de expresión, al igual que el resto de los vectores diccionarios del grupo 14, muestra una alta inducción en la dosis de 40mJ/cm$^2$ a las 24 horas. Sin embargo, a diferencia del resto de las neuronas incluidas en este grupo, esta en particular muestra una clara inducción a las 4 horas. Este subgrupo es altamente notable debido al hecho de que los 8 genes representados por él son los únicos genes que muestran una respuesta específica a las altas dosis de UVB tanto a las 4 como a las 24 horas. Estos genes, que son mostrados en la tabla 11.1, podrían representar un papel muy importante en la respuestas a las radiaciones de UVB activándose solamente cuando el daño provocado por las radiaciones sobrepasa ciertos límites. La figura 11.8 muestra la media y las desviaciones estándar de los perfiles de expresión asignados a cada grupo.

| Acceso a GeneBank | Gen | Función biológica |
|---|---|---|
| M28130 | Interleukin 8 | Respuesta inflamatoria |
| M60278 | Heparin-binding EGF-like growth factor | Transducción de señales |
| M69043 | MAD-3 encoding IKB-like activity | Factor de transcripción |
| U18062 | TFIID subunit TAFII55 | Factor de transcripción |
| U65093 | Msg1-related 1 (mrg1) | Factor de transcripción |
| U89505 | Hlark | Factor de transcripción |
| X78687 | Neuraminidase 1 | Lisosomal Hidrolase |
| Z34974 | Plakophilin | Adhesión |

**Tabla 11.1:** Descripción y función biológica de los genes encontrados en el grupo 14a.

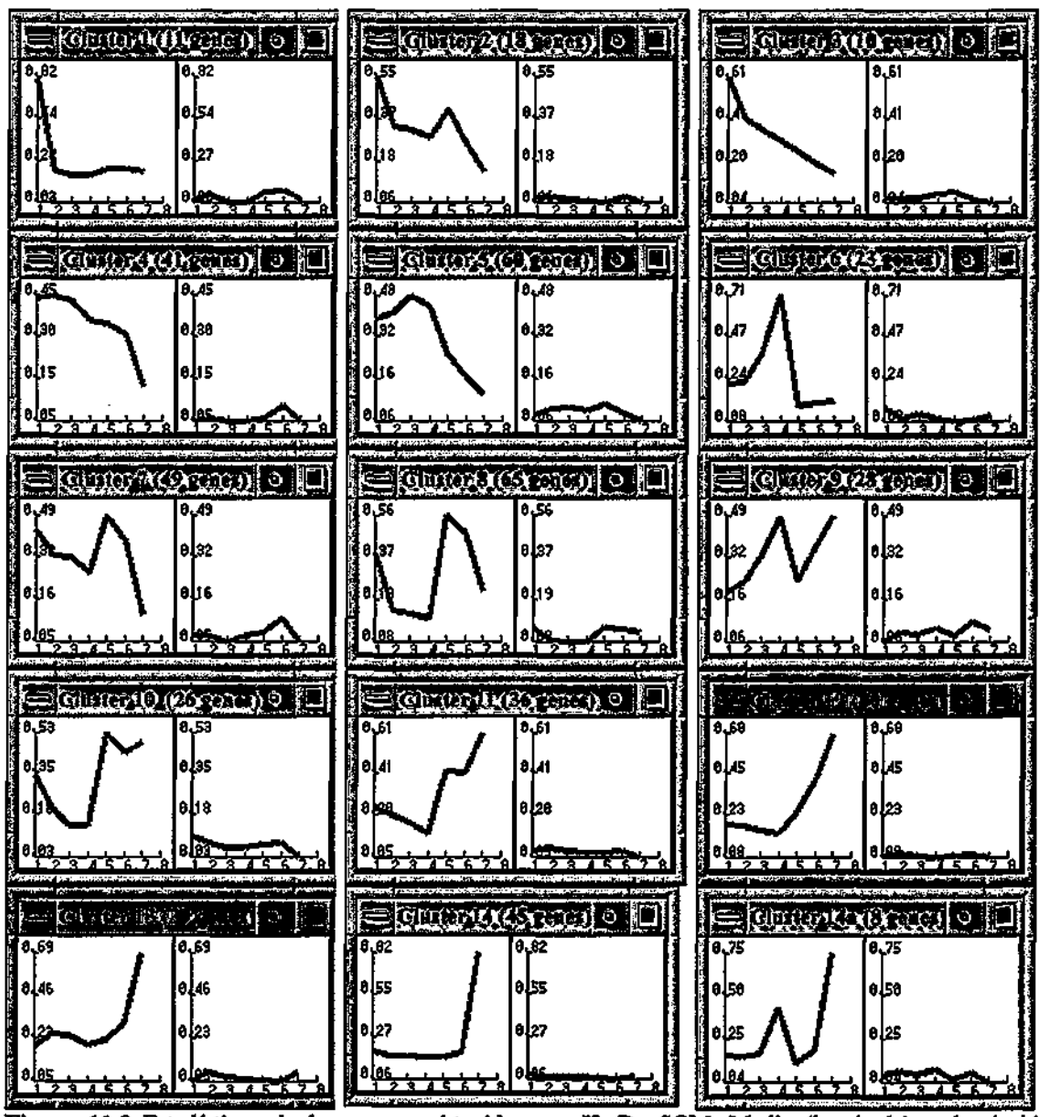


**Figura 11.8** Estadísticos de los grupos obtenidos por KerDenSOM. Media (izquierda) y desviación estándar (derecha) de los perfiles de expresión asignados a cada grupo de la figura 11.7. Puntos experimentales en el eje X: **1** Control, **2** 10 mJ/cm$^2$ 4h, **3** 20 mJ/cm$^2$ 4h **4** 40 mJ/cm$^2$ 4h, **5** 10 mJ/cm$^2$ 24h, **6** 20 mJ/cm$^2$ 24h y **7** 40 mJ/cm$^2$ 24h.

Los resultados obtenidos por KerDenSOM en esta aplicación muestran patrones de expresión que se corresponden consecuentemente con el comportamiento esperado para este tipo de genes regulados por las radiaciones UV y cuyo patrón de expresión era ya conocido, como por ejemplo genes involucrados con el estrés y los procesos inflamatorios, así como factores de transcripción relacionados con la respuesta a las radiaciones UV [210].

Estos resultados permiten la extracción de conocimiento biológico de estos datos estableciendo una correlación entre los patrones de expresión de los grupos y las funciones biológicas de los genes asignados a cada uno de ellos. Por ejemplo, analizando el mapa de la figura 11.7 y la información funcional de los genes extraídas de la base de datos GeneCards (http://bioinformatics.weizmann.ac.il/cards/) se ha detectado que genes involucrados en transcripción basal, empalme (splicing), traslación y degradación mediada por proteasomas muestran un patrón generalizado de inducción (grupos 9, 12, 13 y 14). Por el contrario, genes funcionalmente relacionados con el metabolismo y la adhesión presentan una fuerte represión (grupos 1, 2, 3, 4 y 7).

Estos resultados ofrecen una perspectiva global de los procesos involucrados en la respuesta a las radiaciones UVB, permitiendo no solo conocer qué genes son afectados por estas radiaciones, sino también cómo y en qué medida lo están. Adicionalmente mediante esta metodología es posible la identificación de funciones biológicas que se ven afectadas por estos estímulos.

Estos resultados han sido comparados con los previamente obtenidos sobre estos mismos datos por el método de SOM clásico y reportados previamente [210]. Si bien los patrones generales extraídos por SOM coinciden con los mostrados en esta memoria, no ha sido posible observar el alto nivel de detalles mostrado por KerDenSOM. La comparación, sin embargo, no es completamente justa debido a que las dimensiones del mapa utilizado en esos trabajos previos era de 3x3 lo cual limita considerablemente el nivel de detalles que el algoritmo puede extraer.

Con vistas a realizar una comparación más equitativa, hemos repetido este experimento utilizando el método de SOM clásico sobre un mapa de 20x10, variando los parámetros de ancho de vecindad, iteraciones y razón de aprendizaje para obtener diferentes mapas con los que poder comparar nuestro algoritmo. Los resultados, a pesar de ser parecidos, no mostraban el nivel de detalles obtenidos por KerDenSOM,

especialmente en lo referente al grupo 14a, que por su baja densidad, y su poca diferencia con respecto a los genes asignados al grupo 14, no fue detectado por SOM. Creemos que la razón fundamental de estas diferencias radican en la naturaleza intrínseca de ambos métodos. KerDenSOM tiende a seguir los detalles de la densidad de probabilidad de los datos de la mejor manera posible, sin perder por eso su capacidad de generalización.

Adicionalmente, creemos que una de las ventajas prácticas de este método con respecto a SOM reside en su control sobre el parámetro de suavidad del mapa generado, que permite una selección muy amplia de niveles de suavidad de la proyección. A efectos prácticos esto se traduce en un mejor control del proceso de mapeo, permitiendo encontrar mapas que, sin perder su nivel de organización, reflejen de manera mucho más clara la diferencia entre grupos vecinos.

A pesar de la gran cantidad de métodos de exploración y agrupamiento de datos que han sido propuestos para el análisis de perfiles de expresión génica, creemos que este algoritmo aporta un número considerable de ventajas que lo hacen un buen candidato para ser considerado una herramienta rutinaria en este tipo de análisis, no solo por los aspectos teóricos en los que está basado, sino también por la calidad de los resultados que produce, como ha quedado demostrado en este apartado.

## Conclusiones y principales aportaciones

Los resultados obtenidos en el presente trabajo de tesis nos permiten extraer las siguientes conclusiones:

- Se ha propuesto una nueva metodología para la creación de mapas auto-organizativos basados optimización funcional mediante la combinación de términos matemáticos que expresan dos de las cualidades básicas de los mapas auto-organizativos: fidelidad a los datos y ordenamiento topológico. Para ellos se han fusionado ideas utilizadas durante mucho tiempo en el campo del análisis estadístico de datos y en el campo de reconocimiento de patrones, como son agrupamiento difuso, estimación de la función de densidad de probabilidad y proyección no lineal de datos.
- Se ha propuesto un nuevo algoritmo para obtener mapas auto-organizativos basado en una función de coste matemáticamente bien definida a partir de la extensión del funcional del algoritmo de c-medias difuso, obteniéndose, por primera vez, un mapa auto-organizativo difuso.
- Se desarrolló un nuevo algoritmo de cuantificación vectorial basado en la obtención de vectores representantes que preservan, de la mejor manera posible, la función de densidad de probabilidad de los datos originales objeto de estudio.
- Se propuso un nuevo algoritmo para obtener mapas auto-organizativos basados en la obtención de vectores diccionarios distribuidos suavemente en un espacio de baja dimensión y que preservan, de la mejor manera posible, la función densidad de probabilidad de los datos originales.
- Los algoritmos presentados en este trabajo de tesis han sido desarrollados en un marco matemáticamente formal y tratable, permitiendo no solo su aplicación práctica en problemas de clasificación reales, sino también ofreciendo una mejor comprensión teórica de los procesos de cuantificación vectorial y de proyección que estos algoritmos llevan a cabo.
- Los algoritmos de FuzzySOM, KCM y KerDenSOM propuestos en esta memoria producen, no solo un conjunto de vectores representantes que tienden a seguir de manera fiel la distribución de los datos que se estudian, sino también una matriz de pertenencia de cada dato a cada vector representante, concediéndole una naturaleza

difusa a los mismos, con las consecuentes ventajas prácticas que esta información produce.

- El algoritmo de KerDenSOM propuesto en esta tesis produce no solo un mapa auto-organizativo topológicamente correcto, sino que también produce una estimación no paramétrica de la función densidad de probabilidad de los datos que se analizan.
- La aplicación de los algoritmos desarrollados en esta tesis a datos reales de microscopía electrónica tridimensional han permitido la extracción de información relevante difícil de observar con algoritmos de clasificación tradicionales, incluyendo los clásicos mapas auto-organizativos de Kohonen.
- Se propuso, por primera vez, la aplicación de mapas auto-organizativos a datos de tomografía electrónica, para la clasificación de tomogramas. Estos estudios han permitido obtener nuevas evidencias sobre la variación estructural de los puentes de unión de los filamentos de actina y miosina en músculos estriados, los cuales juegan un papel clave en el proceso contracción muscular.
- Se ha propuesto una nueva y eficiente metodología para la representación de datos volumétricos a baja y media resolución que puedan ser almacenados, manipulados y comparados entre sí de manera eficaz en el contexto de bases de datos. Esta metodología ha sido basada en la utilización de técnicas de cuantificación vectorial combinada con la creación de un modelo de representación que preserva las características de forma y topología presentes en las estructuras tridimensionales.
- La aplicación de los mapas auto-organizativos basados en la estimación de la función densidad de probabilidad presentados en esta memoria han sido también aplicados de manera exitosa a datos reales de expresión génica. Esta aplicación ha permitido la creación de una nueva metodología para el análisis masivo de este tipo de datos.
- Se ha desarrollado una jerarquía de clases en C++ que contiene las clases y funciones necesarias para la utilización y programación de los nuevos algoritmos presentados en este trabajo de tesis. Así mismo, se ha desarrollado dos paquetes de programas para el análisis de imagen, agrupamiento y clasificación de datos de microscopía electrónica tridimensional, tomografía electrónica y perfiles de expresión génica.

## Trabajo futuro

Los resultados presentados en este trabajo constituyen un punto de partida para un estudio más completo y una extensión metodológica tanto de los algoritmos propuestos como de las aplicaciones que, debido a la generalidad de los mismos, pudieran derivarse en otras áreas de las ciencias o de la tecnología. Algunas de las líneas de estudio futuro son las que se proponen a continuación:

- Los mapas auto-organizativos presentados en esta tesis podrían ser planteados de distintas maneras utilizando la misma metodología propuesta de combinación de los términos de fidelidad en la cuantificación y el ordenamiento topológico. Por ejemplo, cualquier combinación monotónica de ambos términos (parte A y parte B de los funcionales propuestos) podría ser válida y quizás produciría algoritmos diferentes, que manteniendo el mismo objetivo, mejore los resultados o simplemente sean más eficiente computacionalmente. El estudio de estas variantes podría ser una línea de investigación futura.
- El parámetro de suavidad en los funcionales propuestos desafortunadamente no es estimable y la utilización de valores incorrectos de este parámetro puede conducir irremediablemente a la obtención de mapas auto-organizativos erróneos o que no representen de manera clara la variabilidad estructural de los datos. Una línea de trabajo futura es el estudio de métodos que permitan estimar, lo mejor posible, el rango de valores que sea más adecuado para este parámetro en dependencia de los datos que se estudien.
- Los algoritmos de KCM y KerDenSOM están basados en la estimación de la densidad de probabilidad mediante funciones núcleo. Usualmente se trabaja con una función gaussiana, pero cualquier otra función núcleo puede ser válida. Una línea de estudio inmediata es la comparación de distintas funciones núcleo y su impacto en la clasificación y el agrupamiento de datos reales como los utilizados en esta memoria.
- Los algoritmos presentados en este trabajo producen una matriz de pertenencia difusa que está relacionada con la probabilidad de que un dato pertenezca a un vector diccionario determinado. Esta información puede ser utilizada de muchas maneras, aunque en las aplicaciones propuestas en esta tesis su utilización ha sido limitada exclusivamente a la asignación del dato al vector diccionario para el cual el valor de pertenencia sea mayor. Sin embargo, la matriz de pertenencia podría ser

utilizada de manera más eficiente permitiendo una valoración de regiones de solapamiento entre grupos (vectores diccionarios) cercanos para los cuales se pueda identificar valores de pertenencia parecidos. El estudio de estas variantes es también una línea de trabajo futura.

- La utilización explícita de la función de densidad de probabilidad es otro de los elementos de estudio propuestos a corto plazo. Esta información podría ser muy bien combinada con la que proyección que producen los mapas auto-organizativos para detectar zonas de alta densidad y poder realizar una separación automática del mapa en grupos distintos a partir de la detección de picos y valles en la función de densidad.
- En el caso de las aplicaciones en Microscopía Electrónica, uno de los puntos de estudio más directo, es la utilización de los valores de densidad en el caso de la cuantificación vectorial para la obtención de modelos de forma y topología, presentado en la sección 10 de esta memoria. Actualmente los datos utilizados son 3D (correspondientes a la posición de los voxeles con más densidad dentro del volumen). Sin embargo, la utilización del valor de intensidad de los voxeles podría ser utilizada explícitamente en este algoritmo si se aplica directamente como factor de ponderación de los datos. Esto permitiría que cada pseudo-átomo se representaría no solo por su posición en el espacio, sino también por el valor de intensidad que representa.
- Por último, una de las líneas de trabajo de más prioridad para el futuro sería la paralelización de los algoritmos propuestos en esta memoria. Debido al enorme número de datos que usualmente se procesan y su alta dimensionalidad, estos métodos pueden tardar varias horas de procesamiento. Una implementación paralela de estos algoritmos aliviaría considerablemente el tiempo dedicado al análisis y exploración de los datos, a la vez que se aprovecharía eficazmente los recursos computacionales.

## Apéndice A: Derivadas de matrices

Tomado como referencias textos clásicos de derivadas de matrices [211] podemos resumir algunas reglas de derivación necesarias para la optimización de los funcionales descritos en esta memoria:

Sean $\mathbf{Y}$ y $\mathbf{Z}$ dos matrices de valores complejos. Sus diferenciales son descritos por las siguientes ecuaciones:

$$d(\mathbf{YZ}) = (d\mathbf{Y})\mathbf{Z} + \mathbf{Y}(d\mathbf{Z}) \tag{A.1}$$

$$d(\mathbf{Y}^{-1}) = -\mathbf{Y}^{-1}(d\mathbf{Y})\mathbf{Y}^{-1} \tag{A.2}$$

$$d(tr(\mathbf{Y})) = tr(d\mathbf{Y}) \tag{A.3}$$

$$d(\ln|\mathbf{Y}|) = tr(\mathbf{Y}^{-1}(d\mathbf{Y})) \tag{A.4}$$

donde $tr(\mathbf{M})$ denota la traza y $|\mathbf{M}|$ denota el determinante de la matriz $\mathbf{M}$.

Sea $\Psi$ una función escalar de la matriz $\mathbf{Y}$. Si el diferencial de $\Psi$ con respecto a $\mathbf{Y}$ puede ser expresado como:

$$d\Psi(\mathbf{Y}) = tr\left[\mathbf{A}(d\mathbf{Y})\mathbf{B} + \mathbf{C}(d\mathbf{Y}^*)\mathbf{D}\right] \tag{A.5}$$

Entonces:

$$\frac{\partial}{\partial \mathbf{Y}}\Psi(\mathbf{Y}) = \mathbf{A}^*\mathbf{B}^* + \mathbf{DC} \tag{A.6}$$

Donde las matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, y $\mathbf{D}$ pueden depender de $\mathbf{Y}$, y donde $\mathbf{M}^*$ denota el complejo conjugado y traspuesto de $\mathbf{M}$.

Sea $\Psi$ una función escalar de la matriz variable $\mathbf{A}$, donde cada elemento de $\mathbf{A}$ es una función escalar de la variable real $z$. Entonces:

$$\frac{\partial}{\partial z}\Psi(\mathbf{A}(z)) = tr\left[\frac{\partial}{\partial \mathbf{A}}\Psi(\mathbf{A})\cdot\frac{\partial}{\partial z}\mathbf{A}\right] \tag{A.7}$$

## Apéndice B: Publicaciones

- P.A. de-Alarcon, A. **Pascual-Montano, A.,** J.M. Carazo. Spin Images and Neural Networks for Efficient Content-Based Retreival in 3D Object Database. *Lecture Notes on Computer Science. Image and Video Retrieval,* ed. M.S. Lew, Sebe, N., Eakins, J.P. 2002, Springer. 225-234.
- **Pascual-Montano, A.,** Taylor, K.A., Winkler, H., Pascual-Marqui, R.D, Carazo, J.M., Quantitative Self-Organizing Maps for Clustering Electron Tomograms, *Journal of Structural Biology.* 2002, 138, pp. 114-122.
- P.A. de-Alarcon, **A. Pascual-Montano,** A. Gupta and J.M. Carazo. Modeling Shape and Topology of Low-resolution Density Maps of Biological Macromolecules. *Biophysical Journal* 2002, 83(2): 619-632.
- P.A. de-Alarcon, A. **Pascual-Montano, A.** Gupta, J.M. Carazo. Modeling Shape and Topology of 3D Images of Biological Specimens., *Proceedings of the International Congress of Pattern Recognition.* 2002, 1, 79-82.
- **Pascual-Montano, A.,** Donate, L.E., Valle, M., Bárcena, M., Pascual-Marqui, R.D, Carazo, J.M., A Novel Neural Network Technique for Analysis and Classification of EM Single Particle Images, *Journal of Structural Biology.* 2001, Feb;133(2-3):233-45. (Medline PMID: 11472094)
- Pascual-Marqui, R., **Pascual-Montano, A.,** Kochi, K., Carazo, J.M, Smoothly Distributed Fuzzy c-Means: a New Self-Organizing Map, *Pattern Recognition.* 2001, 34:2395-2402.
- **Pascual, A.,** Bárcena, M., Merelo, J.J., Carazo, J.M., Mapping and Fuzzy Classification of Macromolecular Images using Self-Organizing Neural Networks. *Ultramicroscopy* 84, 2000, 85-99.
- **Pascual, A.,** Barcena, M., Merelo, J.J., Carazo, J.M., Self-Organizing networks for mapping and clustering biological macromolecules images, *Proceedings of the conference Artificial Neural Networks in Medicine and Biology (ANNIMAB-1, Göteborg, Sweden),* 2000, pp. 283-288.
- **Pascual, A.,** Barcena, M., Carazo, J.M., Application of the Fuzzy Kohonen Clustering Network to Biological Macromolecules Image Classification. *Pattern Recognition and Image Analysis. M.I. Torres & A. Sanfeliu (eds.). Proceedings of the VIII Symposium Nacional de Reconocimientos de Formas y Análisis de Imágenes. (Bilbao, 1999)* pp. 531-538.
- **Pascual, A.,** Barcena, M., Merelo, J.J., Carazo, J.M., Application of the Fuzzy Kohonen Clustering Network to Biological Macromolecules Image Classification, *Lecture Notes on Computer Science* (1607), pp. 331- 340, 1999.
- Merelo, J.J., Rivas, V., Romero, G., Castillo, P., **Pascual, A.,** Carazo, J.M., Improved automatic classification of biological particles from electron-microscopy images using genetic neural nets. *Lecture Notes on Computer Science* (1607), pp. 373-382, 1999.

**PATENTES:** Sistema para el mapeo no lineal de datos y reducción de dimensionalidad. Número: PCT/ES00/00466. Diciembre 2000.

## Apéndice C: Software desarrollado

**XMIPP:**

Los métodos desarrollados en este trabajo de tesis han sido implementados en C++ para sistemas operativos UNIX (principalmente Linux e IRIX) y forman parte del paquete de programas **Xmipp** (**X**-Windows-based **M**icroscopy **I**mage **P**rocessing **P**ackage), principalmente orientado para el tratamiento de imágenes de Microscopía Electrónica. Este software es de dominio público y se puede descargar gratuitamente en la siguiente dirección web: http://www.cnb.uam.es/~bioinfo

El desarrollo de este paquete de programas incluye dos niveles fundamentales de trabajo:

- Nivel de usuario formado por más de 20 programas que incluyen operaciones de pre-procesamiento, procesamiento de imágenes y una amplia galería de programas para clasificación y agrupamiento.
- Nivel de programador donde se proporciona una una jerarquía de clases en C++, también de libre acceso, que permite la implementación y desarrollo de aplicaciones utilizando estas técnicas.

Un esquema simplificado de esta jerarquía es el siguiente:

- xmippCDSet
  - xmippCB
    - xmippMap
    - xmippFCB
      - xmippFuzzyMap
- Descent
- xmippDistance
  - xmippMDistance
  - xmippEDistance
- Layout
  - RECTLayout
  - HEXALayout
- xmippNorm
- xmippPlanes
- xmippSammon
- xmippCTSet
  - xmippCTVectors
  - xmippCB
    - xmippMap
    - xmippFCB
      - xmippFuzzyMap

- xmippUmatrix
- xmippUniform
- xmippBaseAlgo
    - xmippKerDenSOM
    - xmippFuzzySOM
    - xmippSOM
        - xmippBatchSOM
    - xmippFCMeans
        - xmippFKCN

Estas clases contiene todas las funciones y estructuras de datos necesarias para implementar nuevos algoritmos de clasificación y agrupamiento con un mínimo de esfuerzos.

**ENGENE:**

Adicionalmente al paquete XMIPP descrito anteriormente, se ha implementado un sistema de análisis de datos de microchips de ADN (descrito en la sección 11 de esta memoria de tesis).

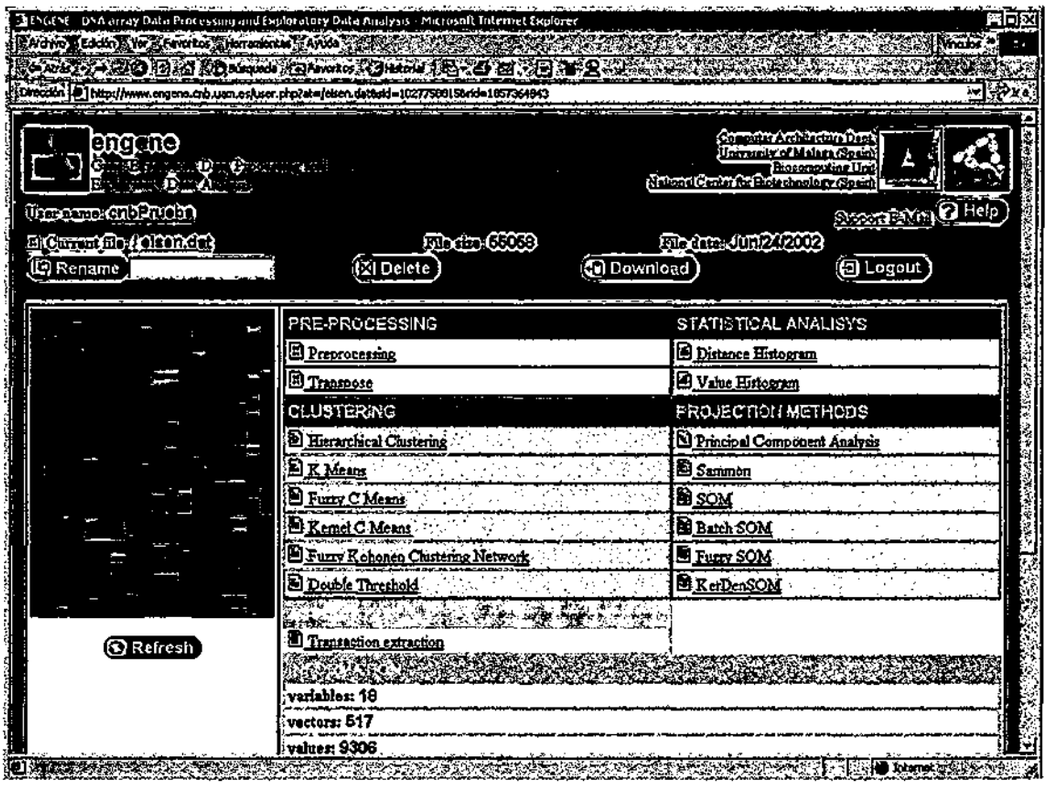



**Figura C.1** Imagen de la página principal de Engene.

Este software fue diseñado siguiendo una arquitectura cliente-servidor con una interfaz de usuario desarrollada sobre un navegador web. El sistema, llamado **engene™** (***"gene engine"***), permite el almacenamiento, pre-procesamiento, análisis de agrupamiento y visualización de datos de expresión génica.

El motor de algoritmos utilizados por este sistema está basado en una estructura de clases en C++ parecida a la del sistema XMIPP con algunas funciones extras de análisis propias para los microchips de ADN. Su utilización para fines académicos es gratuita y puede accederse, previo registro, en la siguiente dirección: www.engene.cnb.uam.es. La figura C.1 muestra una imagen de la página principal de **engene™** donde se puede observar la galería de algoritmos disponibles. Al igual que en XMIPP, los algoritmos propuestos en esta tesis se encuentran también disponibles para en este sistema para el análisis y agrupamiento de datos de microchips de ADN.

## Bibliografía


[1] Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Redaschi, N., Stoehr, P., Tuli, M.A., Tzouvara, K. and Vaughan, R., *The EMBL Nucleotide Sequence Database.* Nucleic Acids Research, 2002. **30**: p. 21-26.

[2] Appel, R.D., Bairoch, A. and Hochstrasser, D.F., *A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server.* Trends Biochem. Sci., 1994. **19**: p. 258-260.

[3] Bhat, T.N., Bourne, P.E., Feng, Z., Gilliland, G., Jain, S., Ravichandran, V., Schneider, B., Schneider, K., Thanki, N., Weissig, H., Westbrook, J. and Berman, H.M., *The PDB data uniformity project.* Nucleic Acids Res., 2001. **29**: p. 214-218.

[4] Brown, P.O. and Botsein, D., *Exploring the new world of the genome with DNA microarrays.* Nature Biotechnol., 1999. **14**: p. 1675-1680.

[5] Wilkins, M.R., Williams, K.L., Appel, R.D. and Hochstrasser, D.F., *Proteome research: new frontiers in functional genomics.* 1997: Springer Verlag.

[6] Hoaglin, D.C., *Exploratory data analysis*, in *Encyclopedia of Statistical Sciences*, S. Kotz, N.L. Johnson, and C.B. Read, Editors. 1982, Wiley: New York. p. 579-583.

[7] Tukey, J.W., *Exploratory Data Analysis.* 1977: Addison-Wesley, Reading, MA.

[8] Jain, A.K. and Dubes, R.C., *Algorithms for Clustering Data.* 1988, New York: Prentice Hall, Englewood Cliffs.

[9] Velleman, P.F. and Hoaglin, D.C., *Applications, Basics, and Computing of Exploratory Data Analysis.* 1981, Boston, MA.: Duxbury Press.

[10] Fayyad, U., Grinstein, G.G. and Wierse, A., eds. *Information visualization in Data Mining and Knowledge Discovery.* 2002, Morgan Kaufmann.

[11] Box, G.E.P. and Jenkins, G., *Time Series Analysis: Forecasting and Control.* 1976: Holden-Day.

[12] Andrews, D.F., *Plots of high-dimensional data.* Biometrics, 1972. **28**: p. 125-136.

[13] Chernoff, H., *The use of faces to represent points in k-dimensional space graphically.* Journal of the American Statistical Association, 1973. **68**: p. 361-368.

[14] Anderberg, M.R., *Cluster Analysis for Applications.* 1973, London: Academic Press.

[15] Hartigan, J., *Clustering Algorithms.* 1975, New York: Wiley.

[16] Hotelling, H., *Analysis of a complex of statistical variables into principal components.* Journal of Educational Psychology, 1933. **24**: p. 417-441, 498-520.

[17] Friedman, J.H., *Exploratory projection pursuit.* Journal of the American Statistical Association, 1987. **82**: p. 249-266.

[18] Friedman, J.H. and Tukey, J.W., *A projection pursuit algorithm for exploratory data analysis.* IEEE Transactions on Computers, 1974. **23**: p. 881-890.

[19] Kruskal, J.B. and Wish, M., *Multidimensional Scaling*, in *Paper series on Quantitative Applications in the Social Sciences.* 1978, Sage University: 07-011. Sage Publications, Newbury Park, CA.

[20] Sammon, J.W., *A nonlinear mapping for data structure analysis.* IEEE Transactions on Computers, 1969. **18**: p. 401-409.
[21] Hastie, T. and Stuetzle, W., *Principal curves.* Journal of the American Statistical Association, 1989. **84**: p. 502-516.
[22] Kohonen, T., *Self-Organizing maps.* Second ed. 1997: Springer-Verlag.
[23] Kohonen, T., *Self-organized formation of topologically correct feature maps.* Biol. Cybernet, 1982. **43**: p. 59-69.
[24] Kaski, S., Kangas, J. and Kohonen, T., *Bibliography of Self-Organizing Map (SOM) Papers: 1981--1997.* Neural Computing Surveys, 1998. **1**: p. 102-350.
[25] Kohonen, T., *Construction of similarity diagrams for phonemes by a self-organizing algorithm.* 1981, Report TKK-F-A463. Helsinki University of Technology, Espoo, Finland.
[26] Ritter, H., *Asymptotic level density for a class of vector quantization processes.* IEEE Transactions on Neural Networks, 1991. **2**: p. 173-175.
[27] Kraaijveld, M.A., Mao, J. and Jain, A.K. *A non-linear projection method based on Kohonen's topology preserving maps.* in *11th International Conference on Pattern Recognition.* 1992. Los Alamitos, CA.: IEEE Computer Society Press.
[28] Kraaijveld, M.A., Mao, J. and Jain, A.K., *A nonlinear projection method based on Kohonen's topology preserving maps.* IEEE Transactions on Neural Networks, 1995. **6**: p. 548-559.
[29] Ultsch, A., *Self-organizing neural networks for visualization and classification,* in *Information and Classification,* O. Opitz, B. Lausen, and R. Klar, Editors. 1993, Springer-Verlag: Berlin. p. 307-313.
[30] Ultsch, A. and Siemon, H.P. *Kohonen's self organizing feature maps for exploratory data analysis.* in *International Neural Network Conference.* 1990: Kluwer, Dordrecht.
[31] Kaski, S., *Data exploration using self-organizing maps,* in *Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series.* 1997, Helsinki University of Technology: Helsinki, Finland.
[32] Muñoz, A. and Muruzábal, J., *Self-Organizing Maps for Outlier Detection.* Neurocomputing, 1998. **18**(1-3): p. 33-60.
[33] Muruzábal, J. and Muñoz, A., *On the Visualization of Outliers via Self-Organizing Maps.* Journal of Computational and Graphical Statistics, 1997. **6**(4): p. 355-382.
[34] Cottrell, M.F., J.C., Pagès, G., *Theoretical aspects of the SOM algorithm.* Neurocomputing, 1998. **21**: p. 119-138.
[35] Kohonen, T., *Analysis of a simple self-organizing process.* Biol. Cybernet., 1982. **44**: p. 135-140.
[36] Cottrell, M. and Fort, J.C., *Etude d'un algorithme d'auto-organisation.* Ann. Inst. Henri Poincare´, 1987. **23**(1): p. 1–20.
[37] Bouton, C. and Pagès, G., *Self-organization of the one-dimensional Kohonen algorithm with non-uniformly distributed stimuli.* Stochastic Process. Appl., 1993. **47**: p. 249-274.
[38] Bouton, C. and Pagès, G., *Convergence in distribution of the one-dimensional Kohonen algorithm when the stimuli are not uniform.* Adv. Appl. Probab., 1994. **26**: p. 80-103.

[39] Erwin, E., Obermayer, K., Schulten, K., *Self-organizing maps: ordering, convergence properties and energy functionals.* Biol. Cybernet, 1992. **67**: p. 47-55.

[40] Erwin, E., Obermayer, K. and Shulten, K., *Self-organizing maps: stationary states, metastability and convergence rate.* Biol. Cybernet., 1992. **67**: p. 35-45.

[41] Fort, J.C. and Pagès, G. *About the convergence of the generalized Kohonen algorithm.* in *ICANN'94*. 1994. Berlin: Springer.

[42] Fort, J.C. and Pagès, G., *On the a.s. convergence of the Kohonen algorithm with a general neighborhood function.* Ann. Appl. Probab, 1995. **5**(4): p. 1177-1216.

[43] Ritter, H., Schulten, K., *Convergence properties of Kohonen's topology conserving maps: fluctuations, stability and dimension selection.* Biol. Cybern., 1988. **60**: p. 59-71.

[44] Kohonen, T., *Self-organizing maps: optimization approaches*, in *Self-organizing Maps*. 1991, Springer Verlag: Berlin.

[45] Tolat, V.V., *An analysis of Kohonen's self-organizing maps using a system of energy functionals.* Biol. Cybernet., 1990. **64**: p. 155-164.

[46] Jardine, N. and Sibson, R., *Mathematical Taxonomy*. 1971, London: Wiley.

[47] Sneath, P.H.A. and Sokal, R.R., *Numerical Taxonomy*. 1973, San Francisco, CA.: Freeman.

[48] Tryon, R.C. and Bailey, D.E., *Cluster Analysis*. 1973, New York: McGraw-Hill.

[49] MacQueen, J. *Some methods for classfication and analysis of multivariate observations.* in *Fifth Berkeley Symposium on Mathematical Statistics and Probability*. 1967. Berkeley and Los Angeles, CA.: University of California Press.

[50] Bezdek, J.C., *Fuzzy Mathematics in Pattern Classification*, in *Ph.D. dissertation. Dept. Appl. Math.* 1973, Cornell Univ.: Ithaca, N.Y.

[51] Chen Kuo Tsao, E., Bezdek, J.C. and Pal, N.R., *Fuzzy kohonen clustering networks.* Pattern Recognition, 1994. **27**: p. 757-764.

[52] Bezdek, J.C., *Pattern Recognition with Fuzzy Objective Function Algorithms*. 1981: Plenum, New York.

[53] Duda, R.O. and Hart, P.E., *Pattern Classification and Scene Analysis*. 1973, New York: John Wiley & Sons.

[54] Silverman, B.W., *Density Estimation for Statistics and Data Analysis*. 1986, London: Chapman and Hall.

[55] Fix, E. and Hodges, J.L., *Discriminatory analysis, nonparametric discrimination: Consistency properties*. 1951, Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Air Field, TX, Report No. 4.

[56] Koontz, W.L.G., Narendra, P.M. and Fukunaga, K., *A Graph-Theoretic Approach to Nonparametric Cluster Analysis.* IEEE Transactions on Computers, 1976. **25**(9): p. 936-944.

[57] Fukunaga, K. and Hostetler, L.D., *Estimation of the gradient of a density function with applications in pattern recognition.* IEEE Transactions on Information Theory, 1975. **IT-21**: p. 32-40.

[58] Kittler, J., *A locally sensitive method for cluster analysis.* Pattern Recognition, 1976. **8**: p. 23-33.

[59] Ripley, B.D., *Computer generation of random variables: a tutorial.* Int. Stat. Rev., 1983. **51**: p. 301-319.

[60] Parzen, E., *On the estimation of a probability density function and the mode.* Annals of Mathematical Statistics, 1962. **33**: p. 1065-1076.
[61] Graepel, T.B., M., Obermayer, K., *Self-organizing maps: Generalizations and new optimization techniques.* Neurocomputing, 1998. **21**: p. 173-190.
[62] Luttrell, S.P., *A Bayesian analysis of self-organizing maps.* Neural Computing, 1994. **6**: p. 767-794.
[63] Cristianini, N. and Shawe-Taylor, J., *An introduction to Support Vector Machines (and other kernel-based learning methods).* 2000: Cambridge University Press.
[64] Bishop, C.M., Svensén, M., Williams, C.K.I., *GTM: the generative topographic mapping.* Neural Computing, 1998. **21**: p. 215-234.
[65] Lampinen, J. and Oja, E., *Clustering properties of hierarchical self-organizing maps.* J. Math. Imaging and Vision, 1992. **2**: p. 261-272.
[66] Cheng, Y., *Convergence and ordering of Kohonen's batch map.* Neural Computing, 1997. **9**: p. 1667-1676.
[67] Vuorimaa, P., *Fuzzy self-organizing map.* Fuzzy Stes and Systems, 1994. **66**: p. 223-231.
[68] Wahba, G. *Spline Models for Observational Data.* in *SIAM.* 1990. Philadelphia.
[69] Gersho, A. and Gray, R.M., *Vector Quantization and Signal Compression.* 1992, Boston: Kluwer Academic Publishers.
[70] Fisher, R.A., *The use of multiple measurements in taxonomic problems.* Ann. Eugen, 1936. **7**: p. 179-188.
[71] Jain, A.K., Duin, P.W. and Mao, J., *Statistical Pattern Recognition: A Review.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000. **22**(1): p. 4-37.
[72] Gray, R.M. and Olshen, R.A. *Vector Quantization and Density Estimation.* in *Int'l Conf. Compression and Complexity of Sequences. Available at http://www-isl.stanford.edu/~gray/positano.pdf.* 1997.
[73] Fukunaga, K. and Hayes, R.R., *The Reduced Parzen Classifier.* IEEE Trans. on Pattern Analysis and Machine Intelligence, 1989. **11**(4): p. 423-425.
[74] Redner, R.A. and Walker, H.F., *Mixture densities, maximum likelihood and the EM algorithm.* SIAM Review, 1984. **26**: p. 195-202.
[75] Xu, L. and Jordan, M.I. *Unsupervised learning by EM algorithm based on finite mixture of Gaussians.* in *World Congr. Neural Networks (II).* 1993.
[76] Xu, L. and Jordan, M.I., *On convergence properties of the EM algorithm for Gaussian mixtures.* Neural Computation, 1996. **8**: p. 129–151.
[77] Bezdek, J.C. and Pal, N.R., *Two soft relatives of learning vector quantization.* Neural Networks, 1995. **8**(5): p. 729-743.
[78] Dersch, D.R. and Tavan, P., *Asymptotic level density in topological feature maps.* IEEE Trans. Neural Networks, 1995. **6**: p. 230-236.
[79] Ritter, H. and Schulten, K., *On the stationary state of Kohonen's self-organizing sensory mapping.* Biol. Cybernet., 1986. **54**: p. 99-106.
[80] Yin, H. and Allinson, N.M. *Comparison of a Bayesian SOM with the EM algorithm for Gaussian mixtures.* in *Workshop Self-Organizing Maps.* 1997.
[81] Yin, H. and Allinson, N.M., *Bayesian learning for self-organizing maps.* Electron. Lett., 1997. **33**: p. 304-305.

[82] Wang, Y., Adali, T., Kung, S.-Y. and Szabo, Z., *Quantification and segmentation of brain tissues from MR images: A probabilistic neural network approach.* IEEE Trans. Image Processing, 1998. **7**.

[83] Van Hulle, M.M., *Kernel-based equiprobabilistic topographic map formation.* Neural Computation, 1998. **10**(7): p. 1847-1871.

[84] Van Hulle, M.M. *Nonparametric density estimation and -regression achieved with a learning rule for equiprobabilistic topographic map formation.* in *IEEE Workshop on Neural Networks for Signal Processing.* 1996. Kyoto.

[85] Van Hulle, M.M., *Topographic map formation by maximizing unconditional entropy: a plausible strategy for "on line" unsupervised competitive learning and non-parametric density estimation.* IEEE Trans. Neural Networks, 1996. **7**(5): p. 1299-1305.

[86] Yin, H. and Allinson, N.M., *Self-Organizing Mixture Networks for Probability Density Estimation.* IEEE Transactions On Neural Networks, 2001. **12**(2): p. 405-411.

[87] Holmström, L. and Hämäläinen, A. *The self-organizing reduced kernel density estimator.* in *IEEE International Conference on Neural Networks.* 1993. San Francisco, California.

[88] Hämäläinen, A., *Self-Organizing Map and Reduced Kernel Density Estimation,* in *PhD thesis, University of Jyväskylä.* 1995: Jyväskylä, Finland.

[89] Wahba, G., Johnson, D.R., Gao, F. and Gong, J., *Adaptive Tuning of Numerical Weather Prediction Models: Part I: Randomized GCV and Related Methods in Three and Four Dimensional Data Assimilation.* 1994, TR 920. Department of Statistics. University of Wisconsin-Madison. http://www.stat.wiscedu/~whaba.

[90] Ormoneit, D. and Tresp, V., *Averaging, Maximum Penalized Likelihood and Bayesian Estimation for Improving Gaussian Mixture Probability Density Estimates.* IEEE Transactions On Neural Networks, 1998. **9**(4): p. 639-650.

[91] Herbin, M., Bonnet, N. and Vautrot, P., *A Clustering Method Based on the Estimation of the Probability Density Function and on the Skeleton by Influence Zones. Application to Image Processing.* Pattern Rec. Lett., 1996. **17**: p. 1141-1150.

[92] Meek, G.A., *Practical Electron Microscopy for Biologists.* 1982: John Wiley and Sons.

[93] Hawkes, P.W., *The electron microscope as a structure projector,* in *Electron Tomography,* J. Frank, Editor. 1992, Plenum. p. 17-38.

[94] Frank, J., *Three-Dimensional Electron Microscopy of Macromolecular Assemblies.* 1996, San Diego: Academic Press.

[95] Bonnet, N., *Artificial Intelligence and Pattern Recognition Techniques in Microscope Image Processing and Analysis.* Advances in Imaging and Electron Physics, 2000. **114**: p. 1-77.

[96] van Heel, M. and Frank, J., *Use of multivariate statistics in analysing the images of biological macromolecules.* Ultramicroscopy, 1981. **6**(2): p. 187-94.

[97] Frank, J. and van Heel, M., *Correspondence analysis of aligned images of biological particles.* J Mol Biol, 1982. **161**(1): p. 134-7.

[98] van Heel, M., *Multivariate statistical classification of noisy images (randomly oriented biological macromolecules).* Ultramicroscopy, 1984. **13**(1-2): p. 165-83.

[99] McQuitty, L.L., *Elementary linkage analysis for isolating orthogonal and oblique types of typal relevancies.* Educational and Psychological Measurement, 1957. **17**: p. 297-329.
[100] Horn, D., *A study of personality syndromes.* Character and Personality, 1943. **12**: p. 257-274.
[101] Sokal, R.R. and Michener, C.D., *A statistical method for evaluating systematic relationships.* University of Kansas Science Bulleting, 1958. **38**: p. 1409-1438.
[102] Ward, J.H., *Hierarchical grouping to optimize an objective function.* Jornal of the American Statistical Association, 1963. **58**: p. 236-244.
[103] Carazo, J.M., Rivera, F.F., Zapata, E.L., Radermacher, M. and Frank, J., *Fuzzy sets-based classification of electron microscopy images of biological macromolecules with an application to ribosomal particles.* J Microsc, 1990. **157**(Pt 2): p. 187-203.
[104] Frank, J., Bretaudiere, J.P., Carazo, J.M., Verschoor, A. and Wagenknecht, T., *Classification of images of biomolecular assemblies: a study of ribosomes and ribosomal subunits of Escherichia coli.* J Microsc, 1988. **150**(Pt 2): p. 99-115.
[105] Harauz, G., Boekema, E. and van Heel, M., *Statistical image analysis of electron micrographs of ribosomal subunits.* Methods Enzymol, 1988. **164**: p. 35-49.
[106] van Heel, M., *Classification of very large electron microscopial image data sets.* Optik, 1989. **82**: p. 114-126.
[107] Wong, M.A., *A hybrid clustering method for identifying high-density clusters.* Am. Stast. Assoc. J., 1982. **77**: p. 841-847.
[108] Marabini, R. and Carazo, J.M., *Pattern recognition and classification of images of biological macromolecules using artificial neural networks.* Biophys J, 1994. **66**(6): p. 1804-1814.
[109] Gao, Y., Vainberg, I.E., Chow, R.L. and Cowan, N.J., *Two cofactors and cytoplasmic chaperonin are required for the folding of alpha- and beta-tubulin.* Mol Cell Biol, 1993. **13**(4): p. 2478-85.
[110] San Martin, C., Radermacher, M., Wolpensinger, B., Engel, A., Miles, C.S., Dixon, N.E. and Carazo, J.M., *Three-dimensional reconstructions from cryoelectron microscopy images reveal an intimate complex between helicase DnaB and its loading partner DnaC.* Structure, 1998. **6**(4): p. 501-9.
[111] Llorca, O., Martin-Benito, J., Ritco-Vonsovici, M., Grantham, J., Hynes, G.M., Willison, K.R., Carrascosa, J.L. and Valpuesta, J.M., *Eukaryotic chaperonin CCT stabilizes actin and tubulin folding intermediates in open quasi-native conformations.* Embo J, 2000. **19**(22): p. 5971-9.
[112] Llorca, O., McCormack, E.A., Hynes, G., Grantham, J., Cordell, J., Carrascosa, J.L., Willison, K.R., Fernandez, J.J. and Valpuesta, J.M., *Eukaryotic type II chaperonin CCT interacts with actin through specific subunits.* Nature, 1999. **402**(6762): p. 693-6.
[113] Llorca, O., Smyth, M.G., Carrascosa, J.L., Willison, K.R., Radermacher, M., Steinbacher, S. and Valpuesta, J.M., *3D reconstruction of the ATP-bound form of CCT reveals the asymmetric folding conformation of a type II chaperonin.* Nat Struct Biol, 1999. **6**(7): p. 639-42.
[114] Barcena, M., Martin, C.S., Weise, F., Ayora, S., Alonso, J.C. and Carazo, J.M., *Polymorphic quaternary organization of the Bacillus subtilis bacteriophage SPP1 replicative helicase (G40 P).* J Mol Biol, 1998. **283**(4): p. 809-19.

[115] Barcena, M., Ruiz, T., Donate, L.E., Brown, S.E., Dixon, N.E., Radermacher, M. and Carazo, J.M., *The DnaB.DnaC complex: a structure based on dimers assembled around an occluded channel.* Embo J, 2001. **20**(6): p. 1462-8.

[116] Pascual, A., Barcena, M., Merelo, J.J. and Carazo, J.M., *Mapping and fuzzy classification of macromolecular images using self- organizing neural networks.* Ultramicroscopy, 2000. **84**(1-2): p. 85-99.

[117] Abdel-Monem, M., Durwald, H. and Hoffmann-Berling, H., *Enzymic unwinding of DNA. 2. Chain separation by an ATP-dependent DNA unwinding enzyme.* ur. J. Biochem., 1976. **65**: p. 441-449.

[118] Abdel-Monem, M. and Hoffmann-Berling, H., *Enzymic unwinding of DNA. 1. Purification and characterization of a DNA-dependent ATPase from Escherichia coli.* Eur. J. Biochem., 1976. **65**: p. 431-440.

[119] Matson, S.W., Bean, D.W. and George, J.W., *DNA helicases enzymes with essential roles in all aspects of DNA metabolism.* BioEssays, 1994. **16**: p. 16-32.

[120] Ellis, N.A., *DNA helicases in inhirited human disorders.* Curr. Opin. Genet. & Dev., 1997. **7**: p. 354-363.

[121] Kornberg, A. and Baker, T.A., *DNA replication. 2da edición.* 1992, San Francisco, California. USA.: Freeman.

[122] Baker, T.A. and Bell, S.P., *Polymerases and the replisome: machines within machines.* Cell, 1998. **92**: p. 295-305.

[123] Boisset, N., Penczek, P., Pochon, F., Frank, J. and Lamy, J., *Three-dimensional architecture of human alpha 2-macroglobulin transformed with methylamine.* J Mol Biol, 1993. **232**(2): p. 522-9.

[124] Marco, S., Chagoyen, M., de la Fraga, L.G., Carazo, J.M. and Carrascosa, J.L., *A variant to the ``random approximation'' of the reference-free alignment algorithm.* Ultramicroscopy, 1996. **66**: p. 5-10.

[125] Penczek, P., Radermacher, M. and Frank, J., *Three-dimensional reconstruction of single particles embedded in ice.* Ultramicroscopy, 1992. **40**(1): p. 33-53.

[126] van Heel, M., Schatz, M. and Orlova, E., *Correlation functions revisited.* Ultramicroscopy, 1992. **46**: p. 307-316.

[127] Marabini, R., Masegosa, I.M., San, M., iacute, n, M.C., Marco, S., Fern, aacute, ndez, J.J., de la Fraga, L.G., Vaquerizo, C. and Carazo, J.M., *Ximipp: An Image Processing Package for Electron Microscopy.* J Struct Biol, 1996. **116**(1): p. 237-40.

[128] Crowther, R.A. and Amos, L.A., *Harmonic analysis of electron microscope images with rotational symmetry.* J. Mol. Biol., 1971. **60**: p. 123-130.

[129] Yu, X., Jezewska, M.J., Bujalowski, W. and Egelman, E.H., *The hexameric E. coli DnaB helicase can exist in different quaternary states.* J. Mol. Biol., 1996: p. 7-14.

[130] Bárcena, M., *Análisis Estructural del Polimorfismo Cuaternario en las Helicasas Replicativas y del Complejo DnaB-DnaC de Escherichia coli*, in *Tesis doctoral. Departamento de Biología Molecular, Facultad de Ciencias.* 2000, Universidad Autónoma de Madrid: Madrid, España.

[131] DePamphilis, M.L., *DNA replication in eukaryotic cells.* 1996, New York.: Cold Spring Harbor.

[132] Levine, A.J. and Burger, M.M., *A working hypotheis explaining the maintenance of the transformed state by SV40 and polyoma.* J Theor Biol, 1972. **37**: p. 436-446.

[133] Waga, S., Bauer, G. and Stillman, B., *Reconstitution of complete SV40 DNA replication with purified replication proteins.* J Biol Chem, 1994. **269**: p. 10923-10934.
[134] Waga, S. and Stillman, B., *Anatomy of a DNA replication fork revealed by reconstitution of SV40 DNA replication in vitro.* Nature, 1994. **369**: p. 207-212.
[135] Fanning, E. and Knippers, R., *Strcuture and function of simian virus 40 large tumour antigen.* Ann. Rev. Biochem, 1992. **61**: p. 55-85.
[136] Bullock, P.A., *The initiation of simina virus 40 DNA replication in vitro.* Crit Rev Biochem Mol Biol, 1997. **32**: p. 503-568.
[137] Simmons, D.T., *SV40 large T antigen functions in DNA replication and transformation.* Adv Virus Res, 2000. **55**: p. 75-134.
[138] Valle, M., Gruss, C., Halmer, L., Carazo, J.M. and Donate, L.E., *Large T-antigen double hexamers imaged at the simian virus 40 origin of replication.* Mol Cell Biol, 2000. **20**(1): p. 34-41.
[139] Frank, J., ed. *Electron Tomography: Three-Dimensional Imaging With The TEM.* 1992, Plenum Press: New York.
[140] Koster, A.J., Grimm, R., Typke, D., Hegerl, R., Stoschek, A., Walz, J. and Baumeister, W., *Perspectives of molecular and cellular electron tomography.* J Struct Biol, 1997. **120**(3): p. 276-308.
[141] Baumeister, W., Grimm, R. and Walz, J., *Electron tomography of molecules and cells.* Trends Cell Biol, 1999. 9(2): p. 81-5.
[142] Baumeister, W. and Steven, A.C., *Macromolecular electron microscopy in the era of structural genomics.* Trends Biochem Sci, 2000. **25**(12): p. 624-31.
[143] Auer, M., *Three-dimensional electron cryo-microscopy as a powerful structural tool in molecular medicine.* J Mol Med, 2000. **78**(4): p. 191-202.
[144] Reedy, M.K. and Reedy, M.C., *Rigor crossbridge structure in tilted single filament layers and flared- X formations from insect flight muscle.* J Mol Biol, 1985. **185**(1): p. 145-76.
[145] Holmes, K.C., Tregear, R.T. and Barrington Leigh, J., *Interpretation of the low angle x-ray diffraction from insect muscle in rigor.* Proc. Roy. Soc. (London) - Series B: Biological, 1980. **207**: p. 13-33.
[146] Taylor, K.A., Reedy, M.C., Cordova, L. and Reedy, M.K., *Three-dimensional image reconstruction of insect flight muscle. I. The rigor myac layer.* J Cell Biol, 1989. **109**(3): p. 1085-102.
[147] Taylor, K.A., Reedy, M.C., Cordova, L. and Reedy, M.K., *Three-dimensional reconstruction of rigor insect flight muscle from tilted thin sections.* Nature, 1984. **310**(5975): p. 285-91.
[148] Taylor, K.A., Reedy, M.C., Reedy, M.K. and Crowther, R.A., *Crossbridges in the complete unit cell of rigor insect flight muscle imaged by three-dimensional reconstruction from oblique sections.* J Mol Biol, 1993. **233**(1): p. 86-108.
[149] Schmitz, H., Reedy, M.C., Reedy, M.K., Tregear, R.T., Winkler, H. and Taylor, K.A., *Electron tomography of insect flight muscle in rigor and AMPPNP at 23 degrees C.* J Mol Biol, 1996. **264**(2): p. 279-301.
[150] Winkler, H. and Taylor, K.A., *Multivariate statistical analysis of three-dimensional cross-bridge motifs in insect flight muscle.* Ultramicroscopy, 1999. **77**: p. 141-152.
[151] Schatz, M. and van Heel, M., *Invariant recognition of molecular projections in vitreous ice preparations.* Ultramicroscopy, 1992. **45**: p. 15-22.

[152] Chen, L.F., Winkler, H., Reedy, M.K., Reedy, M.C. and Taylor, K.A., *Molecular modeling of averaged rigor crossbridges from tomograms of insect flight muscle: A range of strongly-bound structures for the late-stage power stroke.* J. Struct.Biol., 2002. **En prensa.**

[153] Goody, R.S., Reedy, M.C., Hofmann, W., Holmes, K.C. and Reedy, M.K., *Binding of myosin subfragment 1 to glycerinated insect flight muscle in the rigor state.* Biophys J, 1985. **47**(2 Pt 1): p. 151-69.

[154] Lovell, S.J., Knight, P.J. and Harrington, W.F., *Fraction of myosin heads bound to thin filaments in rigor fibrils from insect flight and vertebrate muscles.* Nature, 1981. **293**(5834): p. 664-6.

[155] Thomas, D.D., Cooke, R. and Barnett, V.A., *Orientation and rotational mobility of spin-labelled myosin heads in insect flight muscle in rigor.* J. Muscle Res. Cell Motil., 1983. **4**: p. 367-378.

[156] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., *The Protein Data Bank.* Nucleic Acids Research, 2000. **28**: p. 235-242.

[157] Artymiuk, P.J., *Similarity searching in databases of three-dimensional molecules and macromolecules.* J. Chem. Inf. Comput. Sci., 1992. **32**(6): p. 617-630.

[158] Holm, L. and Sander, C., *Protein Structure comparison by alignment of distance matrices.* J. Mol. Biol., 1993. **233**: p. 123-138.

[159] Shindialov, I.N. and Bourne, P.E., *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.* Protein Engineering, 1998. **11**(9): p. 739-747.

[160] Westhead, D.R., Slidel, T.W., Flores, T.P. and Thornton, J.M., *Protein structural topology: automated analysis, diagrammatic representation and database searching.* Protein Sci, 1999. **8**: p. 897-904.

[161] Edelsbrunner, H., Liang J. and Woodward, C., *Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design.* Protein Science, 1998. **7**: p. 1884-1897.

[162] Norel, R., Petrey, D., Wolfson, H.J. and Nussinov, R., *Examination of Shape Complementarity in Docking of Unbound Proteins.* Proteins, 1999. **36**: p. 307-317.

[163] Morris, G.M., Olson, A.J. and Goodsell, D.S., *Protein-Ligand Docking.* Evolutionary Algorithms in Molecular Design, ed. D.E.Clark. 2000, Weinheim, Germany: Wiley-VCH. 31-48.

[164] Grimes, J.M., Fuller, S.D. and Stuart, D.I., *Complementing crystallography: the role of cryo-electron microscopy in structural biology.* Acta Crystallogr D Biol Crystallogr., 1999. **10**: p. 1742-1749.

[165] Kalko, S.G., Chagoyen , M., Jimenez-Lozano, N., Verdaguer, N., Fita, I. and Carazo, J.M., *The need for a shared database infrastructure: combining X-ray crystallography and electron microscopy.* Eur Biophys J., 2000. **29**(6): p. 457-462.

[166] Bohm, J., Frangakis, A.S., Hegerl, R., Nickell, S., Typke, D. and Baumeister, W., *Toward detecting and identifying macromolecules in a cellular context: Template matching applied to electron tomograms.* Proc Natl Acad Sci U S A, 2000. **97**(26): p. 14245-14250.

[167] Volkmann N. and Hanein., D., *Quantitative fitting of atomic models into observed densities derived by electron microscopy.* J. of Struct Biol., 1999. **125**(2/3): p. 176-184.
[168] Wriggers, W. and Birmanns, S., *Using Situs for Flexible and Rigid-Body Fitting of Multiresolution Single-Molecule Data.* J. of Struct. Biol., 2001. **133**(2/3): p. 193-202.
[169] Wriggers, W., Milligan, R.A. and McCammon, A., *Situs: A Package for Docking Crystal Structures into Low-Resolution Maps from Electron Microscopy.* J. of Struct. Biol., 1999. **125**: p. 185-195.
[170] Wriggers, W., Milligan, R.A., Schulten, K. and McCammon, A., *Self-Organizing Neural Networks Bridge the Biomolecular Resolution Gap.* J. Mol. Biol., 1998. **184**: p. 1247-1254.
[171] Connolly, M.L., *Solvent-accessible surfaces of proteins and nucleic acids.* Science, 1983. **221**: p. 709-713.
[172] Connolly, M.L., *Analytical molecular surface calculation.* Journal of Applied Crystallography, 1983. **16**: p. 548-558.
[173] Connolly, M.L., O'Donnell, T. and Warde, S., *Special issue on molecular surfaces.* Network Science, 1996. **2, 4**.
[174] Edelsbrunner, H. and Mucke, E.P., *Three-dimensional alpha shapes.* ACM Trans. Graphics, 1994. **13**: p. 43-72.
[175] Linde, Y., Buzo, A. and Gray, R.M., *An algorithm for vector quantiser design.* IEEE Transactions on Communications, 1980. **COM-28**: p. 84--95.
[176] Martinetz, T. and Schulten, K., *Topology representing networks.* Neural Networks, 1994. **7**(3): p. 507-522.
[177] Martinetz, T. and Schulten, K., *A neural-gas network learns topologies,* in *Artificial Neural Networks,* T. Kohonen, et al., Editors. 1991, Elsevier: Amsterdam. p. 397-402.
[178] Martinetz, T., Berkovich, S. and Schulten, K., *Neural-gas network for vector quantization and its application to time series prediction.* IEEE Transactions on Neural Networks, 1993. **4**(4): p. 558-569.
[179] Fritzke, B., *Let it grow - self-organizing feature maps with problem dependent cell structure.*, in *Artificial Neural Networks*, T. Kohonen, et al., Editors. 1991, Elsevier: Amsterdam. p. 397-402.
[180] Fritzke, B., *Growing cell structures - a self-organizing network for unsupervised and supervised learning.* Neural Networks, 1994. **7**(9): p. 1441-1460.
[181] Deriche, R., *Using Canny's criteria to derive a recursively implemented optimal edge detector.* Image and Vision Computing, 1987. **1**(2): p. 167-187.
[182] Nastar, C., *The Image Shape Spectrum for Image Retrieval.* 1997, Research Report 3206. INRIA Rocquencort.
[183] Paquet, E. and Rioux, M. *Content-based access of VRML Libraries.* in *IAPR-International Workshop on Multimedia Information Analysis and Retrieval.* 1998. August 13-14. Hong Kong, China.: Lecture Notes in Computer Sciences-Springer.
[184] Lohmann, G., *Volumetric Image Analysis.* 1998: Wiley-Teubner.
[185] Ankerst, M., Kastenmuller, G., Kriegel, H.P. and Seidl, T. *Nearest Neighbor Classification in 3D protein databases.* in *Proceedings ISMB'99.* 1999.

[186] Joshua-Tor, L., Xu, E.H., Johnston, S.A. and Reeds, D.C., *Crystal structure of a conserved protease that binds DNA: The blomycin hydrolase, Gal6.* Science, 1995. **269**: p. 945-950.

[187] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R., *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.* Proc. Natl. Acad. Sci. USA, 1999. **96**: p. 2907-2912.

[188] Eisen, M., Spellman, P.L., Brown, P.O. and Botsein, D., *Cluster analysis and display of genome-wide expression patterns.* Proc. Natl. Acad. Sci. USA, 1998. **95**: p. 14863-14868.

[189] Brazma, A. and Vilo, J., *Gene expression data analysis.* FEBS Lett, 2000. **480**(1): p. 17-24.

[190] Brazma, A., Robinson, A., Cameron, G. and Ashburner, M., *One-stop shop for microarray data.* Nature, 2000. **403**(6771): p. 699-700.

[191] Törönen, P., Kolehmainen, M., Wong, G. and Castrén, E., *Analysis of gene expression data using self-organizing maps.* FEBS letters, 1999. **451**: p. 142-146.

[192] Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L. and Somogyi, R., *Large-scale temporal gene expression mapping of central nervous system development.* Proc. Natl. Acad. Sci. USA, 1998. **95**: p. 334-339.

[193] Zhang, M.Q., *Large-scale gene expression data analysis: a new challenge to computational biologists.* Genome Res, 1999. **9**(8): p. 681-8.

[194] Moore, S.K., *Making chips to probe genes.* IEEE Spectrum, 2001. **38**(3): p. 54-60.

[195] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B., *Missing value estimation methods for DNA microarrays.* Bioinformatics, 2001. **17**(6): p. 520-525.

[196] Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Jr. and Haussler, D., *Knowledge-based analysis of microarray gene expression data by using support vector machines.* Proc Natl Acad Sci U S A, 2000. **97**(1): p. 262-7.

[197] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-537.

[198] Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Staudt, L.M. and et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, 2000. **403**(6769): p. 503-11.

[199] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J., *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.* Proc Natl Acad Sci U S A, 1999. **96**(12): p. 6745-6750.

[200] Tavazoie, S., Hughes, D., Campbell, M.J., Cho, R.J. and Church, G.M., *Systematic determination of genetic network architecture.* Nature Genetics, 1999. **22**: p. 281-285.
[201] Törönen, P., Kolehmainen, M., Wong, G. and Castrén, E., *Analysis of gene expression data using self-organizing maps.* FEBS Lett., 1999. **451**: p. 142-146.
[202] Hartuv, E., Schmitt, A.O., Lange, J., Meier-Ewert, S., Lehrach, H. and Shamir, R., *An algorithm for clustering cDNA fingerprints.* Genomics, 2000. **66**(3): p. 249-56.
[203] Ben-Dor, A., Shamir, R. and Yakhini, Z., *Clustering Gene Expression Patterns.* Journal of Computational Biology, 1999. **6**(3/4): p. 281-297.
[204] Young, A.R., *The biological effects of ozone depletion.* Br. J. Clin. Pract., 1997. **Suppl. 89**: p. 10-15.
[205] Gilchrest, B.A., ed. *Skin and Ageing Processes.* 1989, CRC Press: Boca Raton, FL.
[206] Fisher, G.J., Datta, S.C., Talwar, H.S., Wang, Z.Q., Varani, J., Kang, S. and Voorhees, J.J., *Molecular basis of sun-induced premature skin ageing and retinoid antagonism.* Nature, 1996. **379**: p. 335-339.
[207] Lee, J.A., Frederick, J.E., Haywood, E.K. and Stevens, R.G., *Skin cancers and ultraviolet radiation.* Med. J. Aust., 1989. **150**.
[208] Rogers, G.S. and Gilchrest, B.A., *The senile epidermis: environmental influences on skin ageing and cutaneous carcinogenesis.* Br. J. Dermatol., 1990. **122**(Suppl. 35): p. 55-60.
[209] Green, A., Whiteman, D., Frost, C. and Battistutta, D., *Sun exposure, skin cancers and related skin conditions.* J. Epidemiol., 1999. **9**: p. 7-13.
[210] Sesto, A., Navarro, M., Burslem, F. and Jorcano, J.L., *Analysis of the UVB response in primary human keratinocytes using oligonucleotide microarrays.* Proc Natl Acad Sci U S A, 2002. **99**(5): p. 2965-2970.
[211] Joreskog, K.G., *Factor analysis by least-square and maximum likelihood methods,* in *Statistical Methods for Digital Computers,* A.R.R. K. Enslein, and R. S. Wilf, Editor. 1977, John Wiley & Sons, Inc: New York. p. 125-153.