eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

Máster y Doctorado en

Ingeniería **C**omputacional y **S**istemas **I**nteligentes

Konputazio Zientziak eta Adimen Artifiziala Saila –
Departamento de Ciencias de la Computación e Inteligencia Artificial

Máster I.C.S.I.

Tesis de Máster

Diseño y Validación de un Método de Análisis de Imagen para Cariotipado Automático

Izaro Goienetxea Urkizu

Directores

Dr. Iñaki Inza
Departamento de Ciencias de la Computación e Inteligencia Artificial
Facultad de Informática

Dr. Grégory Maclair
eSalud & Aplicaciones Biomédicas
Vicomtech-IK4

informatika fakultatea
facultad de informática

KZAA /CCIA

Septiembre 2011

Agradecimientos

En primer lugar me gustaría agradecer a Greg Maclair, Iñaki Inza y Carlos Jauquicoa su apoyo y consejo sin el cuál no habría podido terminar este trabajo.

Agradezco también a Ainhoa, Eider, Isabelle y Carlos todos los buenos momentos que han hecho más fácil el trabajo.

Por último, doy gracias a mis amigos y mi familia, especialmente a mi madre por su humor y apoyo incondicional.

Resumen

Hoy en día, el diagnóstico genético tiene un sitio cada vez más importante en la práctica médica, por ejemplo para buscar y predecir si una persona lleva en su genoma una enfermedad que puede activarse en cualquier momento de su vida. El análisis de los cromosomas permite identificar de manera precoz defectos congénitos, enfermedades neurodegenerativas o algunos tipos de cáncer. En los últimos años, se ha mejorado la resolución a la que se es capaz de identificar las alteraciones genéticas. La etapa del cariotipado es imprediscible en el diagnóstico genético, puesto que permite al genetista ver e interpretar los cromosomas del paciente. Actualmente, la etapa de cariotipado es larga, sobre todo la parte que consiste en segmentar y clasificar los cromosomas por pares. Aunque existen aplicaciones para esta tarea, ninguna es completamente automática y siempre requieren la intervencion del genetista.

El objetivo de este proyecto es proponer y evaluar una solución para automatizar el cariotipado a partir de las imágenes obtenidas con el microscopio, hasta tener los cromosomas clasificados por pares, y proporcionar una solución que hará que el genetista pueda dedicarse a la parte la más importante del diagnóstico genético, que es la interpretación y el análisis del cariotipo.

Abstract

Nowadays, genetic diagnosis is more and more used in medical practise, for example to predict if a person is carrier of a "hidden" genetic disease, which can wake up whenever in his life. The analysis of chromosomes allows the earlier detection and identification of congenital problems, neurodegenerative diseases or some type of cancers. With the evolution of the techniques in the last years, the resolution at which one we are able to detect genetic damage has been improved. The karyotyping step is essential in the genetic diagnosis process, since it allows the genetician to see and interpret patient's chromosomes. Today, this step of karyotiping is a time-cost procedure, especially the part that consists in segmenting and classifying the chromosomes by pairs. Although some applications dedicated to this task are available, most of them are not fully-automatic and always requiere an important intervention of a specialist.

The goal of this project is to propose and evaluate a solution to automate the karyotiping, from the images adquired by the microscope to the obtention of the classified chromosomes, and to deliver a solution which will allow the genetician to focus on the most important part of the genetic diagnosis, namely the interpretation of the karyotype.

Índice

1. Introducción **7**
1.1. Presentación del problema 7
1.2. Objetivos 9

2. Estado del arte **13**
2.1. Segmentación de los cromosomas 13
2.2. Extracción de características 16
2.3. Clasificación 19

3. Material y métodos **21**
3.1. Introducción 21
3.2. Segmentación 21
3.3. Extracción de características 25
3.4. Clasificación 33

4. Resultados **39**
4.1. Clasificación en una fase 39
4.2. Clasificación en dos fases 41
4.2.1. Clasificación en grupos Denver 41
4.2.2. Clasificación en el grupo A 42
4.2.3. Clasificación en el grupo B 42
4.2.4. Clasificación en el grupo C 43
4.2.5. Clasificación en el grupo D 43
4.2.6. Clasificación en el grupo E 44
4.2.7. Clasificación en el grupo F 44
4.2.8. Clasificación en el grupo G 45
4.3. Resultados generales 45

5. Conclusiones y trabajo futuro **47**

Índice de figuras

1. Composición de un cromosoma. 7
2. Esquema de la división celular o mitosis. Fuente: [1] 8
3. Esquema de cariotipo. 9
4. Esquema del proceso de cariotipado. 10
5. Ejemplo de clúster de cromosomas. 13
6. Ejemplo de esqueletización de imagen. 15
7. Ejemplo de segmentación utilizando la transformación watershed. 16
8. Ejemplo de ideograma. 17
9. Representación de los grupos Denver. 18
10. Ejemplo de diferencia entre esqueletización y MAT 18
11. Esquema del pipeline de la segmentación. 21
12. Ejemplo de segmentación de imagen de metafase con Otsu. . 22
13. Cromosomas segmentados. 23
14. Imagen de cromosoma antes y después del closing. 24
15. Diferencia entre cromosoma no solapado y solapado. 24
16. Imagen de cariotipo. 25
17. Ejemplo de problemas en imágenes de cariotipos por solapamiento. 26
18. Ejemplo de enderezado de un cromosoma con ImageJ. 27
19. Ejemplo de cálculo satisfactorio del centrómero. 28
20. Cromosoma con su correspondiente perfil de densidad. 31
21. Cromosoma con su correspondiente perfil de forma. 31
22. Cromosoma con su correspondiente perfil de medias de grises. 32
23. Representación gráfica de funciones WDD. 33
24. Cabecera del archivo arff. 34
25. Resultado del evaluador Chi-cuadrado en las características extraídas. 36
26. Esquema de 4-fold Cross-validation. 37
27. Resultado del estadístico Chi-cuadrado. 39

Glosario

ADN: Ácido DesoxirriboNucleico
ARN: Ácido RiboNucleico
MAT: Medial Axis Transform
WDD: Weighted Density Distribution
ANN: Artificial Neural Networks
ITK: Insight Segmentation and Registration Toolkit
MIST: Medical Imaging Software Toolkit
IC: Índice de Centrómero
CFS: Correlated Feature Selection

1. Introducción

1.1. Presentación del problema

El diagnóstico genético ha ido evolucionando en los últimos años en función del desarrollo tecnológico, lo que ha permitido mejorar la resolución a la que se es capaz de identificar las alteraciones, así como optimizar el tiempo de entrega de los resultados.

El diagnóstico genético permite conocer la base genética de una enfermedad hereditaria mediante el análisis de los cromosomas que se encuentran en los núcleos de las células. En la práctica médica, el estudio de los cromosomas humanos es importante porque los cambios en el número de cromosomas y la estructura pueden dar lugar a defectos congénitos, retraso mental, infertilidad, aborto involuntario y cáncer.

Los cromosomas son estructuras complejas localizadas en el núcleo de las células, compuestos por ADN, histonas y otras proteínas, ARN y polisacáridos. Son básicamente los "paquetes" que contienen el ADN. Normalmente los cromosomas no se pueden ver con un microscopio óptico, pero durante la división celular se condensan lo suficiente como para poder ser fácilmente analizados a 100 aumentos.

Se puede ver la composición de un cromosoma en la Figura 1.

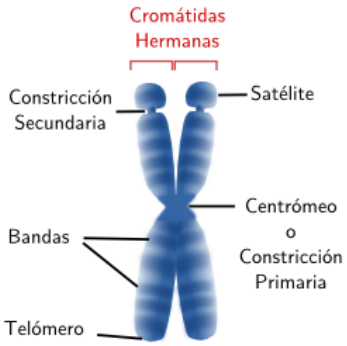

Figura 1: Composición de un cromosoma.

Aunque el número cromosómico del ser humano es 46, pueden darse alteraciones en el número o mutaciones en la estructura de las mismas, debido a errores en la formación de los gametos. Las mutaciones estructurales se dan como consecuencia de roturas cromosómicas seguidas de reconstitución, dando lugar a combinaciones anómalas. Estas combinaciones suponen, en mayor o menor medida, un riesgo para las siguientes generaciones.

Los cambios estructurales no pueden ser estudiados todavía a nivel global por ninguna técnica de biología molecular. Por lo tanto, la técnica de cariotipado continúa siendo insustituible, aunque excesivamente manual y dependiente del factor humano para una adecuada resolución diagnóstica.

La técnica de cariotipado consiste en describir el número de cromosomas en el núcleo de una célula, y cómo se ven a la luz de un microscopio. Los cromosomas son ordenados creando un esquema o ideograma, teniendo en cuenta su morfología (posición del centrómero) y su tamaño. El cariotipo es característico de cada especie, así como el número de cromosomas.

Para observar los cromosomas a ordenar se toman las células en división y se paran en la etapa de metafase, en la que los cromosomas se pueden ver mejor, ya que en este punto el ADN se ha duplicado y ya forma las dos cromátidas. En la Figura 2 puede verse en qué punto se encuentra la metafase en la división celular o mitosis.

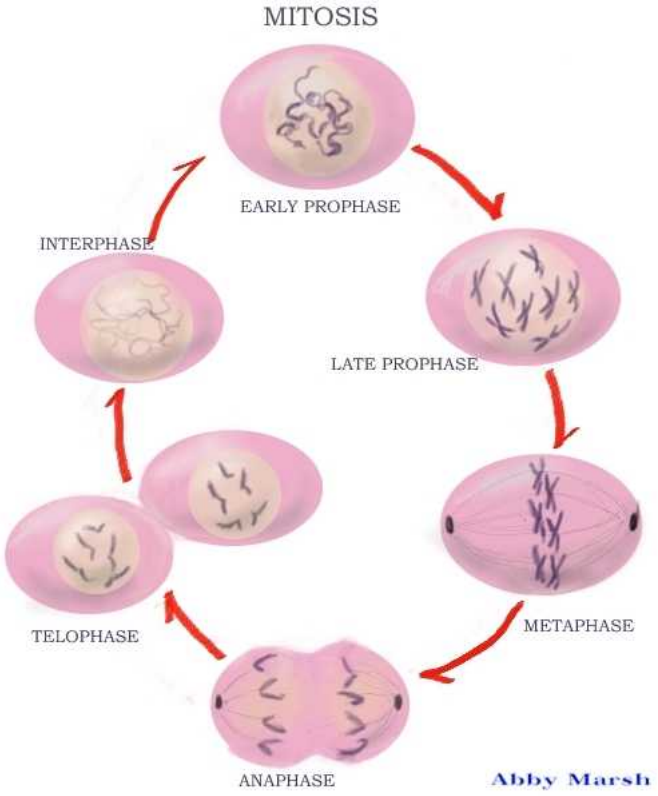

Figura 2: Esquema de la división celular o mitosis. Fuente: [1]

Tras fijar las células en metafase, se usa un colorante (Giemsa) para proceder a la tinción de la célula, que digiere parcialmente ciertas proteínas cromosómicas haciendo posible que se vea el patrón de bandas característico de cada cromosoma.

Teniendo en cuenta la longitud, la posición del centrómero y el patrón de bandas de cada cromosoma se ordenan y se clasifican, como se ve en la Figura 3, haciendo posible la observación de anomalías cromosómicas, tanto numéricas como estructurales.

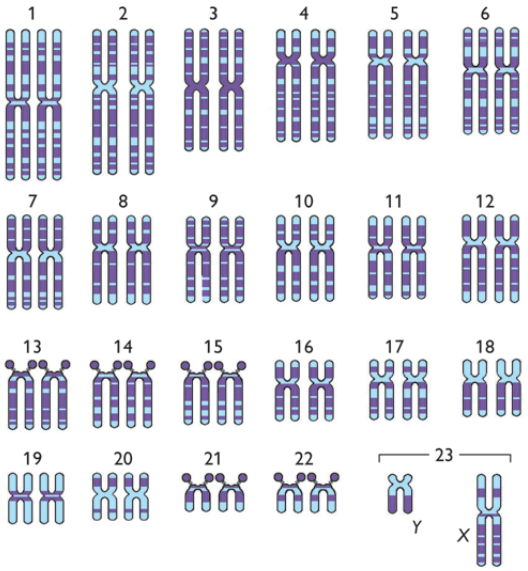

Figura 3: Esquema de cariotipo.

Actualmente existe software para cariotipado, que extrae los cromosomas de una imagen microscópica del núcleo una célula en metafase y realiza la clasificación de cada uno de ellos, pero el proceso no es automático, haciendo que el factor humano sea imprescindible, tanto en la fase de segmentación de los cromosomas como en la fase de clasificación. Esto hace que el proceso sea largo y costoso, haciendo que el proceso completo llegue a durar hasta 3 horas.

En este trabajo se presentará un diseño de un pipeline de procesamiento de imágenes que intentará automatizar el proceso de cartiotipado. Para eso, la primera parte de este trabajo, la sección Estado del arte, ha consistido en un estudio detallado de los métodos de tratamiento de imágenes existentes y relacionados con la problemática del cariotipado, como la segmentación, la extracción de características, y la clasificación. Luego se presentan las soluciones elegidas para este proyecto, en la sección Material y métodos, y las modificaciones aportadas con el fin de agilizar y de mejorar la robustez de las etapas del proceso de cariotipado. Para terminar, se presentan los resultados obtenidos con nuestro pipeline de tratamiento, y se analiza si se ha cumplido el objetivo inicial del proyecto.

1.2. Objetivos

El objetivo principal de este proyecto es analizar los métodos que se utilizan para hacer cariotipado automático (cf. Figura 4), y ver si es viable el diseño de un método para construir un sistema de cariotipado automático eficiente que tomará 10 imágenes de metafase por paciente y creará los cario-

tipos, pudiendo subir la productividad del análisis genético de 9-10 muestras diarias actuales a aproximadamente 100 cariotipos diarios.

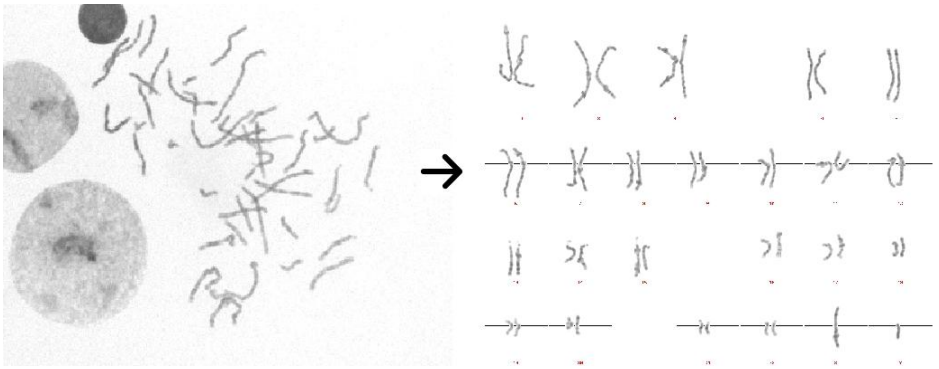

Figura 4: Esquema del proceso de cariotipado.

Para llevar a cabo todo el proceso se han identificado 3 tareas o partes del proyecto:

1. **Segmentación**: El objetivo de esta fase es segmentar los cromosomas de la imagen de metafase de entrada. Para ello hay que idear un algoritmo que permita identificar los cromosomas que estén cruzados o solapados entre ellos y encontrar los puntos de cruce para realizar la división.

2. **Extracción de características:** En esta fase se escogen las características a extraer de las imágenes de los cromosomas segmentados en la fase anterior, para crear la matriz de características que se usará para crear y entrenar el modelo de clasificación.

3. **Clasificación de los cromosomas:** En esta parte del proceso se escogerá un método de clasificación y se aplicará a la matriz de características creada en la fase anterior, pudiendo así crear el cariotipo con los cromosomas clasificados de la imagen de metafase.

Aunque a lo largo del proyecto se analizarán las tres partes, éste se centrará principalmente en la extracción de características y la construcción y entrenamiento de un modelo eficiente para la clasificación de los cromosomas.

El proyecto se realizará en Vicomtech-IK4 con ayuda de Genetadi.

Vicomtech-IK4 es un centro de investigación aplicada que trabaja en el área de gráficos por ordenador interactivos y tecnología multimedia, localizado en el Parque Tecnológico de San Sebastián.

Genetadi Biotech es una compañía surgida en Bizkaia en el año 2005 que desarrolla nuevos servicios y tests de diagnóstico genético humano especializada en ginecología, pediatría y oncología. Impulsada por BEAZ, la Diputación de Bizkaia y el Gobierno Vasco a través del Plan Biobask 2010.

Genetadi ha sido el responsable de introducir al personal de Vicomtech en el contexto del análisis genético y de proporcionar las imágenes de metafases necesarias para realizar este trabajo.

2. Estado del arte

En esta sección se hará un resumen de las técnicas que se utilizan para solucionar el problema que se plantea, aunque se dará una visión más profunda de las técnicas usadas para la solución que se plantea en este trabajo en la sección Material y métodos.

El estado del arte se repartirá en tres secciones, según las tres partes identificadas en la sección 1.2, segmentación, extracción de características y clasificación.

2.1. Segmentación de los cromosomas

La segmentación automática de imágenes de metafase tiene una larga historia, pero en sus inicios se usaban células en estado avanzado de la metafase. En esta fase los cromosomas están contraídos y los cruces y solapamientos son raros, haciendo que la segmentación sea más sencilla.

Más tarde los citogenetistas empezaron a usar imágenes de fases iniciales de la metafase o fases avanzadas de la profase, que ofrecen imágenes en las que se ven más detalles, y cromosomas con más bandas. Esto provoca que los cromosomas se vean más largos, haciendo que sea muy común que haya cruces o solapamientos. No es raro encontrar clústeres de diez o más cromosomas como se ve en la Figura 5, haciendo que sea muy importante un algoritmo para separar esos clústeres.

Figura 5: Ejemplo de clúster de cromosomas.

En métodos como el que se presenta en Wang y col. [2] destacan la importancia de hacer un preprocesado de la imagen, anterior a la segmentación, para realzar la imagen y eliminar el ruido que pueda tener, para obtener una mayor cantidad de detalle de los cromosomas.

Para este realce en [3] se propone la utilización de wavelets para descomponer la imagen, se ordenan los componentes según su contribución a la

imagen y se fijan diferentes coeficientes de realce para ellos. Una vez hechas las modificaciones se reconstruye la imagen.

Los métodos más simples para segmentación de cromosomas están basados en umbrales [4], pero estos no son suficientes para separar los grupos de cromosomas concentrados.

La separación de cromosomas que se tocan sin cruzarse se realiza encontrando una línea de corte entre las dos, para las solapadas, en cambio, es necesario encontrar dos pares de cortes más o menos perpendiculares.

Se desarrollaron más técnicas para abordar este problema como las basadas en descomposiciones de forma fuzzy-logic [5] o las region growing [6], aunque demostraron no ser muy eficaces.

En los últimos años se han hecho grandes avances en este tema, y se han presentado ideas para afrontar el problema como: concavidades de forma, pale paths (entre los cromosomas que se tocan), la esqueletización, validación de forma...

G. Ritter y L. Gao hacen una aproximación al problema en [7]. Aplican una serie de reglas geométricas estrictas para identificar fácilmente y con seguridad los solapamientos y los puntos de contacto entre los cromosomas. Estas reglas geométricas, métodos tradicionales del procesado de imagen, se aplican al problema de la siguiente manera.

En una primera fase de preprocesado se eliminan los componentes grandes, oscuros y redondos, que forman el núcleo celular, y los componentes con un área menor al 1/220 del área total (manchas). Los componentes conectados de la imagen resultante de este preproceso son, en su gran mayoría, cromosomas. Tras el preproceso se suavizan los bordes de los cromosomas con algoritmos morfológicos y se rellenan artefactos que aparecen en los cromosomas como pequeños agujeros.

Algunos de los componentes resultantes son cromosomas aislados, pero el siguiente paso en el proceso es la separación de los clústeres de cromosomas, para realizar este proceso se han seguido los siguientes principios:

- Una ramificación del esqueleto con al menos dos brazos largos indica un clúster de cromosomas.
- Los puntos fronterizos muy cóncavos indican potenciales puntos de corte.
- Una constricción en un componente conecta dos cromosomas, a no ser que ésta sea un centrómero estrecho.
- Un objeto de tamaño menor al 1/220 del área total de la imagen no es un cromosoma, los componentes de este tamaño no se cortan. Particularmente, los objetos con tamaño menor al 1/110 del área total de la imagen no se cortan.

Para llevar a cabo este proceso es necesaria la esqueletización de la imagen.

La esqueletización consiste en extraer el patrón continuo más fino posible que contenga la forma del objeto original, como se puede ver en la Figura 6.

Figura 6: Ejemplo de esqueletización de imagen.

Pueden emplearse varias técnicas para conseguir este propósito, pero éstas se pueden dividir en dos categorías generales:

- Métodos basados en erosión
- Métodos basados en distancias

Los métodos basados en erosión consisten en ir adelgazando el objeto hasta tener una estructura de un único píxel de anchura. Esto puede hacerse iterativamente, como en el método que se presenta en [8] o usando la paralelización como proponen en [9].

En los métodos basados en distancias, como el que se presenta en [10], cada punto del objeto en la imagen toma el valor de la distancia al borde más cercano del objeto, pudiendo usarse varios tipos de distancia para esto (Euclídea, Manhattan...), y utilizan esta matriz de valores para calcular el esqueleto.

Otros autores como P.S. Karvelis y col. proponen usar la transformación watershed de manera recursiva para conseguir separar los clústeres de cromosomas en [11].

En la transformación watershed, la imagen se ve como una representación topológica de un terreno, y a cada píxel se le asocia un valor de altura en el terreno según su nivel de gris. A continuación se simula una inundación de la superficie topológica desde los niveles más bajos de altura (mínimos locales). Este proceso llega a un momento en el que las aguas de cuencas contiguas se unen. Estas líneas de unión, fronteras de regiones homogéneas, representan el resultado de la segmentación (cf. Figura 7).

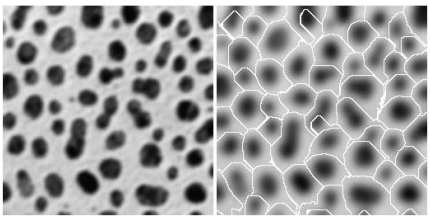

Figura 7: Ejemplo de segmentación utilizando la transformación watershed.

En su trabajo proponen usar la transformación watershed de manera recursiva, afinando más la segmentación en cada iteración, y consiguiendo así mejor separación entre cromosomas pegados o solapados (éxito de segmentación de cromosomas pegados o solapados de entre 90 y 95 %).

Otros autores proponen aproximaciones diferentes, como W. Srisang y col. en [12], que hacen primero una segmentación inicial basada en umbrales, calculan los contornos de los cromosomas usando técnicas de geometría computacional y de las funciones de estos contornos se calculan las funciones de curvatura. Se identifican los posibles puntos de corte entre los cromosomas como los puntos de cambios bruscos en estas funciones de curvatura y se buscan los puntos centrales del área de solapamiento usando diagramas de Voronoi.

2.2. Extracción de características

La extracción de características se basa en encontrar un conjunto de parámetros pequeño que mejor describe cada clase, diferenciando a su vez entre ellas las clases diferentes.

La primera característica que buscan los expertos a la hora de realizar el cariotipado es el tamaño. Dependiendo del tamaño son capaces de clasificar los cromosomas en pequeños grupos, [13]. Después buscan características como el índice del centrómero (ratio entre el brazo corto del cromosoma y la longitud total) o la posición de las bandas características. Con este método el experto puede hacer el cariotipado de manera fiable.

También pueden utilizar la ayuda de un ideograma, un esquema en el que se puede ver el patrón de bandas que caracteriza a cada cromosoma, como se ve en la Figura 8. Estos ideogramas pueden tener diferentes resoluciones, mostrando más o menos bandas.

Figura 8: Ejemplo de ideograma.

Hoy en día existen softwares de cariotipado que usan resoluciones de 850 bandas.

Las características morfológicas habituales en cromosomas humanos incluyen parámetros como el área, perímetro, longitud, ratio entre el brazo corto y el largo etc. [14], pero éstos no tienen en cuenta los patrones de bandas de los cromosomas, y algunos experimentos han demostrado la eficacia de los perfiles de bandas a la hora de hacer la clasificación [15].

Generalmente se diferencian dos grupos de características:

- Características morfológicas
- Características de textura

Las características morfológicas aportan información de la forma y tamaño del cromosoma, pero éstas suelen ser variables, por lo que no son suficientes

para realizar la clasificación completa, aunque son usadas para clasificar los cromosomas en los grupos Denver [14].

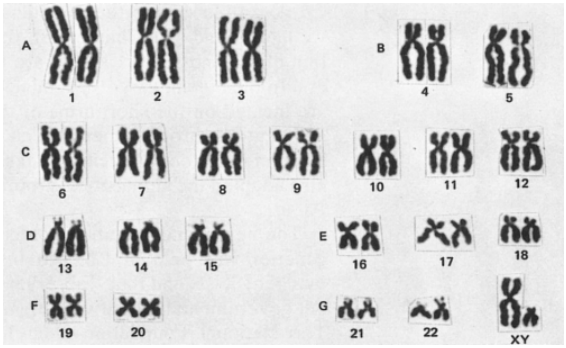

Figura 9: Representación de los grupos Denver.

En la Figura 9 se puede ver la clasificación de los grupos Denver, aunque algunos autores consideran al cromosoma X parte del grupo C por su parecido morfológico.

Las características morfológicas más usadas en la literatura son el eje medio, la longitud y el índice de centrómero [4, 14, 16].

Los cromosomas generalmente no suelen encontrarse completamente rectos, y es necesario encontrar su eje medio para hacer una buena medición de las características [4], para encontrar este eje se utiliza el Medial Axis Transformation (MAT). Esta transformación es muy parecida a la esqueletización, con la diferencia de que el resultado del MAT es una imagen en escala de grises y la de la esqueletización es una imagen binaria. Se puede observar la diferencia de los dos métodos en la Figura 10. En la parte izquierda de la misma se ve el objeto original, en el centro se ve el resultado de la esqueletización y en la derecha está el resultado del MAT.

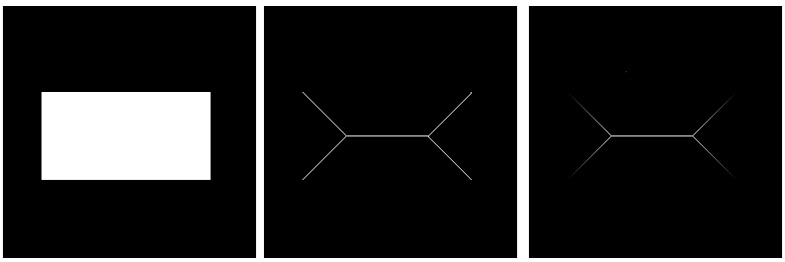

Figura 10: Ejemplo de diferencia entre esqueletización y MAT

Este método es más complejo que la esqueletización, y algunos autores utilizan ésta última [16].

Para calcular el índice de centrómero es necesario buscar la posición del mismo, y se puede pensar la utilidad del perfil de anchura del cromosoma

[17] o el análisis de la curvatura del contorno del mismo [18], ya que el centrómero se ve como una constricción en el cromosoma, pero estas técnicas no son totalmente satisfactorias. Es posible que el centrómero no esté bien representado en la curvatura, o el perfil de anchura sea ruidoso.

En [4], Piper y Granum hacen una aproximación al problema de la localización del centrómero utilizando los perfiles de forma de los cromosomas.

Teniendo en cuenta la dificultad para localizar con seguridad el centrómero, se han desarrollado otras técnicas de clasificación que no tienen en cuenta esta característica, como la de Ritter y col., en la que no se tiene en cuenta la polaridad del cromosoma [19].

Plantean, también, el problema de conocer la orientación y polaridad del cromosoma que se ha extraído de la imagen. Calculan la curvatura del cromosoma y lo rotan hasta que queda vertical.

La polaridad del cromosoma, que tiene en cuenta si el centrómero se encuentra más arriba del centro del cromosoma, se extrae del índice de centrómero, ya que éste nunca se encuentra en una posición más baja que el centro del cromosoma.

Las características de textura se centran en encontrar los parámetros que describen el patrón de bandas de los cromosomas. Los más utilizados son los perfiles.

Los perfiles son gráficos que describen alguna propiedad del cromosoma, y que se calculan en una secuencia de puntos del eje medio, posiblemente curvado. La mayoría de perfiles se calculan con medidas hechas en líneas perpendiculares a los puntos del eje medio [4].

Son muy usados el perfil de densidad [16], el perfil de medias de grises, el perfil de gradiente y el perfil de forma [14]. De estos perfiles se extraen coeficientes de Fourier [20] o coeficientes WDD (Weighted Density Distribution) [4, 14] como parámetros característicos de las funciones.

Se han hecho también, aproximaciones que además de tener en cuenta los perfiles de las bandas, miden también las posiciones en las que se encuentran algunas de las bandas características del patrón [13].

Algunos autores se centran únicamente en el cálculo de perfiles para hacer la clasificación, como J. Kao y col. en [21], pero lo más habitual es hacer una combinación de características morfológicas y de textura, como en [4, 14, 16].

2.3. Clasificación

A la hora de clasificar cromosomas ha sido extendido el uso de redes neuronales supervisadas y no supervisadas, para aumentar la eficiencia y reducir el tiempo de procesado.

El uso de redes neuronales artificiales (ANN, de Artificial Neural Networks) para la clasificación de cromosomas ha sido muy estudiado. Lerner [22] sugirió que las redes neuronales son el mejor clasificador para cromosomas, especialmente cuando el número de clases es limitado. Algunos trabajos

como [13] y [23] se centran en la clasificación dentro de un único grupo con un número de clases muy pequeño, con los cromosomas 16, 17 y 18, por lo que estos métodos pueden resultar muy efectivos.

El Emary [24], propone un método dividido en procesamiento de imagen (realce de los cromosomas, eliminación de ruido y segmentación) en el que extrae las longitudes de los cromosomas, y la clasificación basada en redes neuronales entrenadas con retropropagación. Este método de entrenamiento es muy utilizado en clasificación de cromosomas, pero requiere mucho tiempo.

También se han propuesto otros tipos de redes neuronales, como las redes neuronales wavelet en [25], que en el caso de estudio, la clasificación en el grupo E (cromosomas 16,17 y 18) obtienen mejor resultado que una red neuronal ANN.

Se han utilizado más tipos de redes neuronales para la clasificación de cromosomas, como las redes perceptrón multicapa [21, 26], redes basadas en lógica difusa, redes neuronales recurrentes [27] y redes neuronales probabilísticas [28]. Este tipo de redes tienen ciertas ventajas sobre las redes con retropropagación, ya que su tiempo de entrenamiento es más corto [29].

Para aumentar la eficiencia de estas redes puede optarse por dividir la clasificación en dos fases. En la primera fase los cromosomas se clasifican en uno de los grupos Denver, y en la segunda fase se hace otra clasificación dentro de cada uno de estos grupos, como en [16].

Este método en dos fases se ha utilizado también con redes Bayesianas como clasificadores como en [14], trabajo en el que se demuestra que la clasificación en dos fases y usando redes Bayesianas, es más efectivo que una clasificación simple con una red Bayesiana.

Las redes Bayesianas han demostrado alcanzar la eficacia de las redes neuronales a la hora de la clasificación cuando el número de clases no es muy limitado [13].

Se utilizan otros métodos de clasificación aparte de los mencionados, como relaciones de similitud difusas [30]. En este trabajo se dividen primero los cromosomas en grupos utilizando relaciones de similitud resultantes de grados de pertenencia difusa. En el segundo paso se reordenan los grupos dependiendo de la similitud de las posiciones de sus centrómeros. Como último paso se identifica cada cromosoma del grupo haciendo una correlación de las características de sus bandas con las características de las bandas de la plantilla de cromosomas del grupo correspondiente.

Los modelos ocultos de Markov continuos también han sido utilizados a la hora de clasificar cromosomas, como en [31], que obtienen tasas de error muy parecidas a los clasificadores más típicos como las redes neuronales, así como algunos clasificadores basados en distancias [32] , en kernel de entropía [33] y en la teoría de Bayes [19, 34].

3. Material y métodos

3.1. Introducción

Para llevar a cabo este trabajo se han utilizado imágenes facilitadas por Genetadi. Son imágenes de dos pacientes diferentes, en total 40 imágenes de metafase y sus correspondientes cariotipos realizados por expertos.

Para hacer una mejor clasificación de las imágenes sería preferible que la cantidad de pacientes de las que vienen las imágenes fuese lo más alta posible, para tener más variedad en el modelo de clasificación, pero se han obtenido imágenes de dos únicos pacientes y se ha trabajado con ellas.

Las imágenes de metafase se han obtenido mediante microscopio utilizando una lente de 100 aumentos preparando previamente las células que iban a ser usadas para extraer las imágenes.

Son imágenes de células sanas y sin anomalías cromosómicas, para poder realizar una mejor clasificación.

Se ha introducido como novedad el uso de más de una imagen de metafase de cada paciente (10 imágenes por paciente en este caso) para tener un mejor cariotipo, ya que en muchos casos los cromosomas están muy solapados en las imágenes. Esto hace que partes de algunos cromosomas queden tapadas por otros cromosomas, con lo que se pierde mucha información, sobre todo del patrón de bandas, dificultando la tarea de la clasificación y pudiendo, en el caso de que se oculte una alteración en las bandas, dificultar el posterior diagnóstico genético.

También es posible que al tener varias instancias del mismo cromosoma su clasificación sea más fácil.

En esta sección se explicarán en profundidad los métodos seguidos en la construcción de la solución que se plantea.

3.2. Segmentación

Esta parte del proceso de cariotipado automático aún no ha sido finalizada, pero se ha definido el pipeline que tendrá que seguir y se han desarrollado las primeras aproximaciones.

El pipeline principal puede verse en la Figura 11, y se explicará en detalle a continuación.

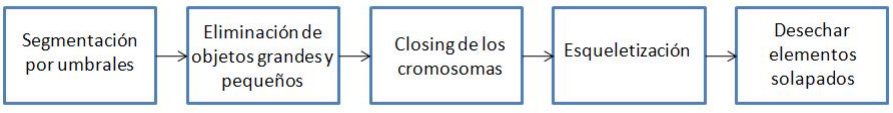

Figura 11: Esquema del pipeline de la segmentación.

Como primer paso del proceso se ha optado por hacer una segmentación

por umbrales, en concreto utilizando el método Otsu. El objetivo de esta segmentación por umbrales es que se identifiquen en la imagen los objetos cuya densidad (valor de gris de sus píxeles) es mayor al umbral calculado. Con este método se consigue una imagen binaria, en la que los objetos identificados toman un valor de 1 mientras que el fondo se queda con valor 0.

El método Otsu utiliza el histograma de la imagen para encontrar el valor que tomará el umbral, de manera que la varianza de valores dentro de cada segmento sea el mínimo posible a la vez que la varianza de valores entre segmentos diferentes sea máxima.

Se ha utilizado la implementación del Otsu threshold de ITK (Insight Segmentation and Registration Toolkit). En la Figura 12 se puede ver el resultado de aplicar este algoritmo a una imagen de metafase.

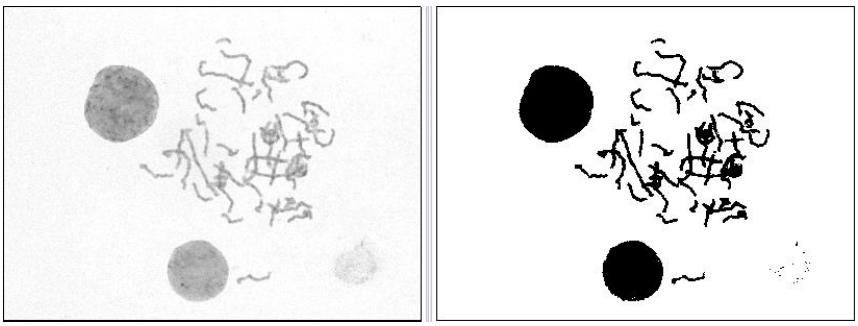

Figura 12: Ejemplo de segmentación de imagen de metafase con Otsu.

El siguiente paso en el proceso es quitar los objetos demasiado grandes y demasiado pequeños de la imagen.

Es muy normal que en las imágenes se vean los núcleos de las células. Suelen aparecer como grandes objetos redondos, y no hay que tenerlos en cuenta en la segmentación, ya que son objetos que no aportan nada al cariotipado.

También es posible que haya ruido o pequeñas partículas de suciedad en la imagen, ya que puede haber polvo en el porta objetos del microscopio. Estas partículas aparecen en la imagen como objetos muy pequeños, que también tendrán que quedar fuera de la segmentación.

Para esto, se han identificado los componentes conectados entre ellos y medido el número de píxeles de cada objeto segmentado. Se han establecido unos límites mínimo y máximo que han de tener los objetos para ser tenidos en cuenta y se han dejado fuera los objetos que no entran en los límites.

En la Figura 13 se puede ver como se han segmentado los cromosomas a la vez que se dejaban fuera los núcleos de la célula.

Figura 13: Cromosomas segmentados.

Hay que tener mucho cuidado estableciendo los límites decidir que un objeto sea o no tenido en cuenta en la segmentación, ya que si se establece un límite mínimo demasiado alto pueden desecharse algunos satélites de cromosomas como si fuesen artefactos no deseables.

El siguiente paso en el proceso sería la esqueletización de los cromosomas, pero para esto es necesario que los cromosomas no tengan ningún agujero, ya que si hay alguno, la esqueletización no se hará bien y en el resultado algunos de los esqueletos que deberían ser continuos aparecerán discontinuos.

Para quitar los posibles agujeros se aplicará un closing, operación morfológica consistente en dilatación y posterior erosión de los objetos de la imagen.

En la Figura 14 se puede ver una imagen binaria de un cromosoma a la izquierda, en el que han quedado algunos pequeños agujeros, y en la imagen de la derecha puede verse cómo ha quedado el mismo cromosoma una vez aplicado el closing, ya sin agujeros.

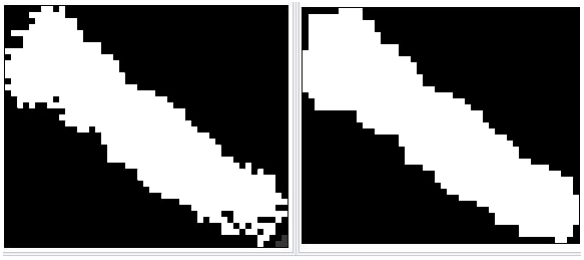

Figura 14: Imagen de cromosoma antes y después del closing.

Una vez hecho el closing hay que esqueletizar la imagen. Se ha mencionado en la parte de estado del arte que hay varias técnicas para la esqueletización. En este trabajo se ha utilizado una implementación basada en erosión, que adelgaza el cromosoma hasta que éste solo tiene un píxel de grosor.

Con la imagen esqueletizada se pueden identificar los cromosomas solapados o cruzados analizando el vecindario de cada píxel del esqueleto usando la siguiente lógica:

Si un píxel tiene dos vecinos con valor 1, ese píxel no hace frontera con otro cromosoma, así que no se solapan en ese punto.

Si el píxel tiene más de dos vecinos con valor 1, ese píxel hace frontera con otro cromosoma, así que hay solapamiento.

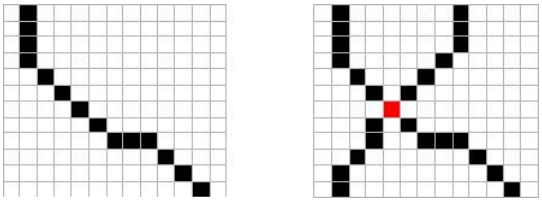

Figura 15: Diferencia entre cromosoma no solapado y solapado.

En la parte izquierda de la Figura 15 se ve un esqueleto sin cruces o solapamientos. Puede verse cómo todos los píxeles del esqueleto tienen en su vecindario únicamente dos píxeles con valor positivo.

En la parte derecha se ven dos esqueletos cruzados. En esta imagen todos los píxeles del esqueleto tienen en su vecindario dos píxeles con valor positivo excepto el píxel marcado en rojo, que es el punto de cruce de dos cromosomas, y en cuyo vecindario hay cuatro píxeles con valor positivo.

Aunque en un futuro trabajo se traten estos solapamientos, en esta primera aproximación al problema se descartarán los cruces y solapamientos de cromosomas.

Como se ha dicho anteriormente, la idea del proyecto es usar unas 10 imágenes por paciente para el cariotipo, por lo que se puede suponer que los cromosomas descartados en una imagen podrán encontrarse en otra sin

cruces ni solapamientos. Esta es una de las ventajas de usar varias imágenes de metafase para cada paciente.

3.3. Extracción de características

Como se ha explicado en la parte de segmentación, ésta todavía no está completa, con lo que no se han podido usar las imágenes de metafase segmentadas para la fase de extracción de características.

Se han utilizado las imágenes de cariotipos facilitadas por Genetadi. Son cariotipos realizados con el software de Leica CW4000, con lo que las imágenes pueden estar más realzadas que las imágenes de metafase.

Se puede ver un ejemplo de imagen de cariotipo en la Figura 16.

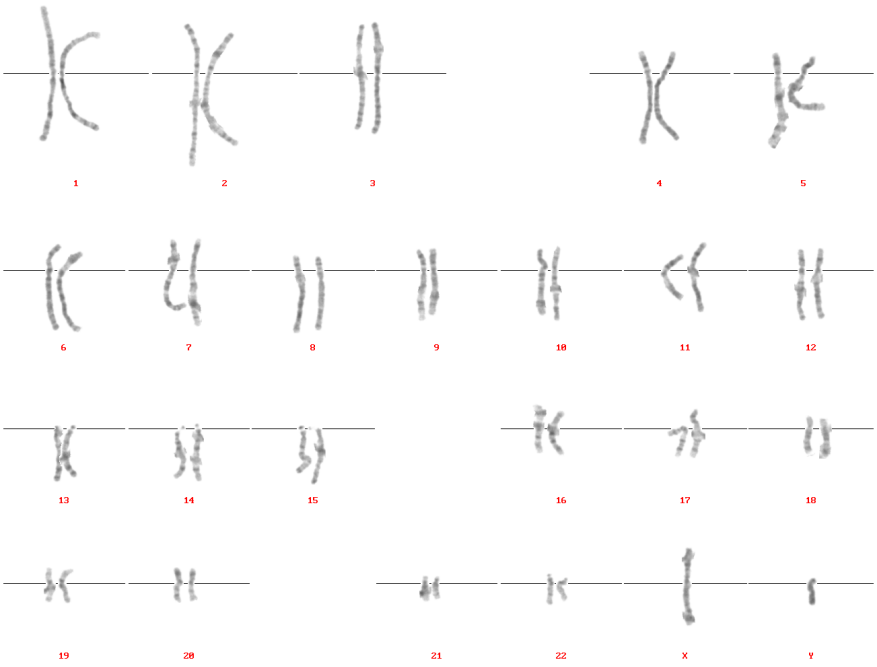

Figura 16: Imagen de cariotipo.

Los problemas como cruces entre cromosomas o solapamientos han sido previamente manejados por el software mencionado, y tienen que ser tenidas en cuenta las medidas que toma para resolver esos problemas.

Por ejemplo en el caso de que un cromosoma aparezca cruzado en la imagen de metafase, la parte que tapa el otro trozo de cromosoma se ve duplicado en la imagen del cariotipo. También es posible, como se ve en la Figura 17, que haya cruces en la imagen de metafase que hacen que en la imagen de cariotipo los cromosomas aparezcan con trozos de bandas correspondientes a otros cromosomas. En la figura puede verse cómo las partes marcadas en la

imagen de la izquierda aparecen como pertenecientes al cromosoma vertical en el cariotipo, cuando en realidad son partes de los cromosomas horizontales.

Esto no afecta a las mediciones de parámetros como la longitud del cromosoma, pero puede hacer que el centrómero no se vea en la imagen o que las medidas relacionadas con el patrón de bandas no sean exactas.

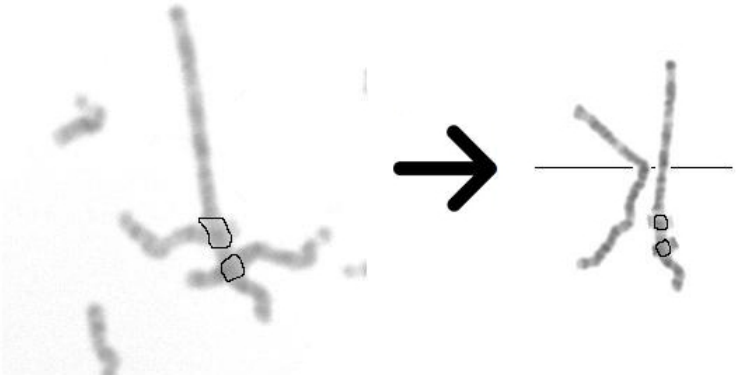

Figura 17: Ejemplo de problemas en imágenes de cariotipos por solapamiento.

En las imágenes de cariotipo no podemos saber si un cromosoma estaba cruzado en la imagen de metafase original, con lo que no se han podido identificar las partes duplicadas, y hay que tener en cuenta que es posible que cierta información que se le dará al clasificador en la siguiente fase no sea del todo correcta.

Teniendo en cuenta que en las imágenes de cariotipos los cromosomas están ordenados y colocados según sus índices de centrómero y que, por esta razón, sus polaridades están claras en casi todas las ocasiones, se ha decidido tener en cuenta las características morfológicas y las basadas en texturas.

Las características morfológicas han demostrado según se ha visto en la sección del estado del arte, ser muy importantes para agrupar los cromosomas en grupos Denver, y se ha decidido medir algunas de estas características. En concreto se tendrán en cuenta la longitud relativa del cromosoma y el índice de centrómero, como hacen en trabajos como [25].

Las características de textura han demostrado ser necesarias para clasificar los cromosomas dentro de los grupos creados a partir de las características morfológicas. Se medirán los perfiles de densidad, perfiles de forma y perfiles de medias de grises.

Para varias de las medidas necesarias para la extracción de características es una ayuda que el cromosoma esté enderezado. En las imágenes de cariotipos, aunque muchos estén bastante rectos, pocos cromosomas lo están totalmente, por lo que se han enderezado. Para ello se ha usado el software

de procesamiento de imágenes open source ImageJ.

Antes de enderezar los cromosomas se ha utilizado este programa para realzar las bandas de los cromosomas ajustando el contraste de la imagen.

Para el enderezado en ImageJ es necesario que el usuario introduzca unos puntos que marquen la línea central del cromosoma. Una vez que el usuario ha introducido los puntos el programa dibuja una spline utilizando estos puntos. El programa también necesita el número de píxeles que se tomarán en cuenta a los lados de la spline dibujada para enderezar la región definida. En la Figura 18, se puede ver el resultado del enderezamiento realizado con ImageJ.

Figura 18: Ejemplo de enderezado de un cromosoma con ImageJ.

ImageJ permite guardar todas las imágenes enderezadas por separado, con lo que se puede guardar cada cromosoma independientemente para luego usarlo en la extracción de características.

Los métodos para extraer cada característica han sido los siguientes:

Índice de Centrómero (IC): Esta característica mide el ratio entre el brazo corto del cromosoma y la longitud total.

Dependiendo del valor del índice de este parámetro los cromosomas humanos pueden clasificarse en 3 tipos diferentes:

- Metacéntricos: El centrómero se encuentra en el centro del cromosoma, dando lugar a brazos de igual longitud.
- Submetacéntricos: La ubicación del centrómero causa que un brazo sea ligeramente más corto que el otro.
- Acrocéntricos: El centrómero está muy cerca de uno de los telómeros, haciendo que un brazo sea muy corto en comparación con el otro.

Como se ha visto en la sección de Estado del arte, la extracción automática de esta característica es complicada.

Algunos autores buscan el mínimo del perfil de forma (más adelante se explica cómo se extrae este perfil), pero esta técnica no siempre es válida, sobre todo cuando el centrómero no está en el centro del cromosoma (cromosoma metacéntrico). En los cromosomas acrocéntricos no se puede observar un mínimo real en el perfil de forma, con lo que la técnica no puede usarse.

Se ha programado este método y se ha observado que el resultado es satisfactorio por ejemplo en el caso de la Figura 19, pero en muchos otros casos no funciona.

Figura 19: Ejemplo de cálculo satisfactorio del centrómero.

Como se ha mencionado anteriormente, la polaridad y la localización aproximada de los centrómeros es conocida en las imágenes de cariotipo, por lo que se ha localizado manualmente y se ha medido el índice con la siguiente fórmula.

$$IC = \frac{L_S}{L_T} \tag{1}$$

Siendo L_S la longitud del brazo corto del cromosoma y L_T la longitud total del cromosoma.

Se ha usado como ayuda la tabla que adjuntan Vogel y Motulsky en [35] (ver Tabla 1) en la que se pueden ver las medias de longitudes e índices de centrómero de todos los cromosomas, aunque en algunos de los casos las medidas no han estado dentro del baremo al que deberían haber correspondido.

Longitud: Los cromosomas enderezados se han guardado en imágenes independientes, por lo tanto para obtener la longitud del cromosoma solo se ha tenido que tener en cuenta la longitud en el eje X del cromosoma. Cada imagen de metafase que se utiliza para hacer el cariotipo es de momentos diferentes, por lo tanto la distribución y el tamaño de los cromosomas varía según la imagen. Por esto es muy importante normalizar la longitud que se ha medido mediante la siguiente fórmula.

Cromosome	Relative Length	IC
1	9.11 (4.43 : 4.68)	48.36±1.166
2	8.61 (3.35 : 5.26)	38.23±1.824
3	6.97 (3.30 : 3.67)	46.95±1.557
4	6.49 (1.80 : 4.69)	29.07±1.867
5	6.21 (1.66 : 4.55)	29.25±1.739
6	6.07 (2.30 : 3.77)	39.05±1.665
7	5.43 (2.01 : 3.42)	39.05±1.771
8	4.94 (1.62 : 3.32)	34.08±1.975
9	4.78 (1.56 : 3.22)	35.43±2.559
10	4.80 (1.55 : 3.25)	33.95±2.243
11	4.82 (1.95 : 2.87)	40.14±2.328
12	4.50 (1.23 : 3.27)	30.16±2.339
13	3.87 (0.64 : 3.23)	17.08±3.227
14	3.74 (0.69 : 3.05)	18.74±3.596
15	3.30 (0.58 : 2.72)	20.3±3.702
16	3.14 (1.33 : 1.81)	41.33±2.74
17	2.97 (0.94 : 2.03)	33.86±2.771
18	2.78 (0.74 : 2.04)	30.93±3.044
19	2.46 (1.10 : 1.36)	46.54±2.299
20	2.25 (1.03 : 1.22)	45.45±2.526
21	1.70 (0.49 : 1.21)	30.89±5.002
22	1.80 (0.51 : 1.29)	30.48±4.932
X	5.16 (1.94 : 3.22)	40.12±2.117
Y	2.21 (0.51 : 1.70)	27.17±3.182

Tabla 1: Tabla de Vogel y Motulsky.

$$L_N = \frac{L_R}{L_T} \tag{2}$$

Siendo L_R la longitud relativa del cromosoma y L_T la longitud del cromosoma más largo de la metafase.

Perfil de densidad: Es un gráfico de una dimensión de la propiedad del patrón de bandas perpendicular al eje medio del cromosoma. Se calcula en una secuencia de puntos del eje medio, posiblemente curvado, del cromosoma. Se hacen mediciones en líneas transversales, perpendiculares a las tangentes del eje medio. Cada valor del perfil (I_i) resulta de la suma de propiedades de puntos espaciados en unidades de distancia en cada línea transversal.

Para reducir la varianza de los valores de densidad causada por la diferencia de condiciones de las células, es necesario normalizar el perfil de densidad.

En las siguientes ecuaciones, m es el número de píxeles en la línea perpendicular a la tangente del eje medio, *d(i,j)* es el valor del pixel en las coordenadas *(i,j)*, y *w(i)* es la anchura del cromosoma en el punto i-ésimo.

$$I_i = \sum_{j=0}^{m-1} d(i,j) \quad (i = 0, 1, \ldots, n-1) \tag{3}$$

$$d_w(i) = \frac{I_i}{w(i)} \quad (i = 0, 1, \ldots, n-1) \tag{4}$$

Finalmente, el perfil se normaliza con la siguiente ecuación, en la que $d_{wMIN}(i)$ es el valor de densidad mínimo y $d_{wMAX}(i)$ es el valor de densidad máximo.

$$d_N(i) = \frac{d_w(i) - d_{wMIN}(i)}{d_{wMAX}(i)} \quad (i = 0, 1, \ldots, n-1) \tag{5}$$

Como se ha explicado anteriormente, se ha trabajado con cromosomas enderezados, por lo que ha sido más fácil calcular los perfiles de densidad, ya que no ha habido que calcular las tangentes de cada punto del eje medio de los cromosomas.

En la Figura 20 se puede ver la imagen de un cromosoma y su correspondiente perfil de densidad.

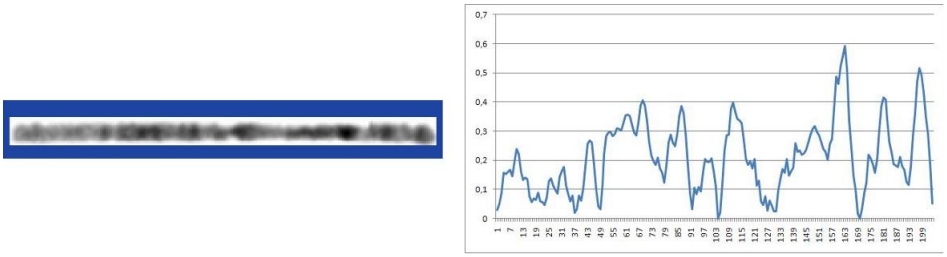

Figura 20: Cromosoma con su correspondiente perfil de densidad.

Perfil de forma: Para obtener el perfil de forma, como en el perfil de densidad, hay que hacer mediciones en las líneas transversales perpendiculares a las tangentes del eje medio.

Este perfil se define como el ratio entre el segundo momento y el momento 0 de cada línea transversal, y puede calcularse con la siguiente ecuación.

$$s_w = \frac{\sum_{j=0}^{m-1}(d(i,j)dist(i,j)^2)}{\sum_{j=0}^{m-1} d(i,j)} \quad (i = 0, 1, \ldots, n-1) \tag{6}$$

En la ecuación *d(i,j)* es el valor de gris del píxel en las coordenadas *(i,j)* y *dist(i,j)* es la distancia euclídea entre el punto i del eje medio y las coordenadas *(i,j)*.

En la Figura 21 se puede ver la imagen de un cromosoma con su correspondiente perfil de forma.

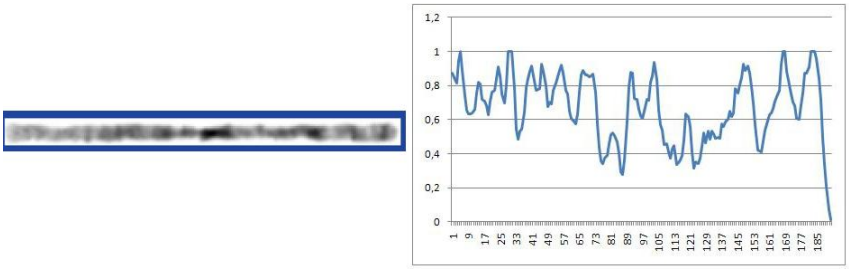

Figura 21: Cromosoma con su correspondiente perfil de forma.

Perfil de media de grises: Este perfil mide la media de los niveles de gris en los puntos de la línea transversal de cada punto del eje medio del cromosoma. Se ha medido con la siguiente ecuación.

$$g_w = \frac{\sum_{j=0}^{m-1} d(i,j)}{n} \quad (i = 0, 1, \ldots, n-1) \tag{7}$$

En la ecuación *d(i,j)* hace referencia al valor de gris del píxel en las coordenadas *(i,j)* y n es el número de píxeles de la línea transversal en el punto i del eje medio.

En la Figura 22 se puede ver un cromosoma con su correspondiente perfil de medias de grises.

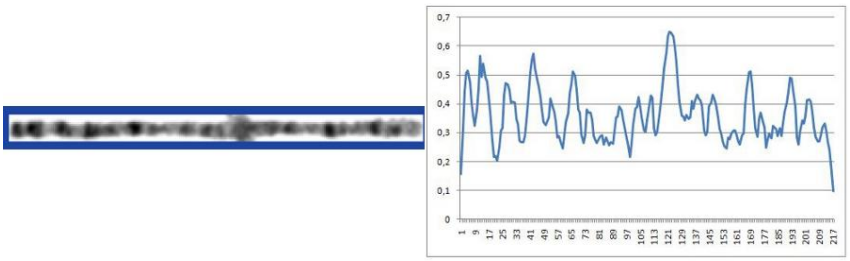

Figura 22: Cromosoma con su correspondiente perfil de medias de grises.

Para reducir la dimensionalidad de estos perfiles y poder usar en la clasificación la información que contienen, se ha decidido utilizar los coeficientes WDD, y se han extraído los 6 primeros coeficientes de cada perfil.

Para la extracción de coeficientes WDD es necesario calcular primero las funciones WDD y multiplicar estas funciones con el perfil del que se quieren extraer los coeficientes.

El cálculo de las funciones se hace con la siguiente fórmula.

$$w_n(x) = \left[2.floor\left(n.\frac{2.x+1}{2.L}\right) + 1 - 2.n.\frac{2.x+1}{2.L}\right](-1)^{floor(\frac{2.x+1}{2.L})-1} \quad 0 \leq x < L \tag{8}$$

donde n es el número de coeficiente seleccionado y L es la longitud del perfil.

Una vez calculadas las funciones WDD, los coeficientes se calculan con la siguiente ecuación.

$$WDD_n = \sum_{x=0}^{L-1} w_n(x)p(x) \tag{9}$$

Gráficamente las primeras 6 funciones WDD son como se representan en la Figura 23.

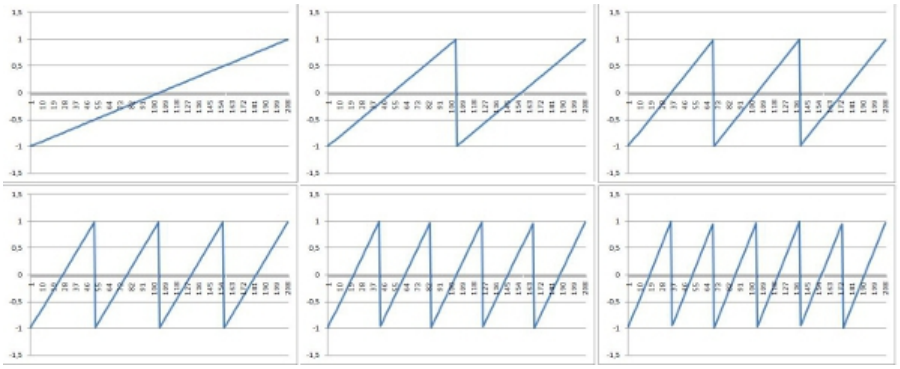

Figura 23: Representación gráfica de funciones WDD.

En la Tabla 2 se ve un resumen de las características que se usan en este trabajo.

Número de característica	Descripción
1	Índice de centrómero (IC)
2	Longitud relativa
3-8	Seis coeficientes WDD del perfil de densidad del cromosoma
9-14	Seis coeficientes WDD del perfil de forma del cromosoma
15-20	Seis coeficientes WDD del perfil de media de grises del cromosoma

Tabla 2: Resumen de las características utilizadas.

Se ha desarrollado una aplicación donde cargar las imágenes de los cromosomas y hacer la extracción de sus características automáticamente. Esta aplicación escribe las características en un archivo .arff que se le podrá pasar a Weka para la clasificación.

La aplicación se ha desarrollado utilizando la librería de imagen médica de Vicomtech-IK4, MIST.

3.4. Clasificación

En la anterior sección se ha explicado que la extracción de características se ha hecho directamente de imágenes de cariotipos, por tanto podemos saber con seguridad que las clases que tienen los cromosomas son correctas, pudiendo usar esta información como entrenamiento del clasificador.

Para la clasificación se ha utilizado el programa para aprendizaje automático y minería de datos open source Weka [36]. Este programa tiene algunos puntos fuertes para este trabajo, como una interfaz gráfica que hace que sea fácil de usar, muchas técnicas de procesamiento de datos y su licencia GNU.

Para utilizar el programa se ha creado un fichero arff con las características extraídas de las imágenes de los cromosomas (cf. Figura 24).

```
@RELATION chromosomes

@ATTRIBUTE CI REAL
@ATTRIBUTE Size REAL
@ATTRIBUTE Density1 REAL
@ATTRIBUTE Density2 REAL
@ATTRIBUTE Density3 REAL
@ATTRIBUTE Density4 REAL
@ATTRIBUTE Density5 REAL
@ATTRIBUTE Density6 REAL
@ATTRIBUTE Shape1 REAL
@ATTRIBUTE Shape2 REAL
@ATTRIBUTE Shape3 REAL
@ATTRIBUTE Shape4 REAL
@ATTRIBUTE Shape5 REAL
@ATTRIBUTE Shape6 REAL
@ATTRIBUTE Gray1 REAL
@ATTRIBUTE Gray2 REAL
@ATTRIBUTE Gray3 REAL
@ATTRIBUTE Gray4 REAL
@ATTRIBUTE Gray5 REAL
@ATTRIBUTE Gray6 REAL
@ATTRIBUTE class {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, Y}
```

Figura 24: Cabecera del archivo arff.

Se han creado dos archivos arff diferentes, cambiando entre ellos la variable clase. En el primer archivo la variable clase hará referencia al número de cromosoma *(1,2,...,Y)*, mientras que en el segundo archivo la variable clase hará referencia al grupo Denver al que pertenece el cromosoma *(A,B,...,G)*. Este segundo archivo se utilizará en la clasificación en dos fases que se explica más adelante.

Se ha decidido hacer pruebas de clasificación con los siguientes métodos de clasificación:

- **Redes Bayesianas:** Las redes Bayesianas son estructuras que capturan las relaciones de dependencia que existen entre los atributos de los datos observados. Describen la distribución de probabilidad que tiene un conjunto de variables especificando suposiciones de independencia condicional con probabilidades condicionales [37].
- **Naive Bayes:** Este clasificador se basa en la teoría de Bayes suponiendo que hay una independencia entre los atributos de los individuos del modelo. Se calculan las distribuciones de probabilidad de cada clase para establecer la relación entre los atributos y la clase (variable dependiente).
- **Perceptrón multicapa:** Es una red neuronal formada por varias capas, característica que le permite resolver problemas linealmente no

separables [38].

- **J48:** Se crea un árbol de decisión tomando para cada nodo del árbol el atributo, no utilizado, cuya entropía es menor, haciendo que el nodo aporte la mayor cantidad de información posible.
- **Random Forest:** Se construye un bosque de random trees. Estos random trees son árboles de clasificación independientes que incorporan aleatoriedad. Hay diferentes tipos de Random Forests dependiendo de cómo se incorpora la aleatoriedad [39].

Se han seleccionado los clasificadores basados en la teoría de Bayes y las redes neuronales debido a su frecuente uso en literatura sobre clasificación de cromosomas.

En la práctica se ha visto que las redes neuronales no son una técnica muy adecuada para el problema que se plantea, ya que necesitan mucho tiempo para el entrenamiento, y el objetivo del proyecto es conseguir un cariotipo automático y rápido. De todas maneras se han mantenido en la experimentación para probar si las técnicas alternativas llegan a su tasa de acierto o son inferiores.

No se han encontrado en la literatura referencias a la clasificación de cromosomas utilizando árboles de decisión, por eso se ha decidido incorporarlos a las pruebas, para ver si pueden ser técnicas con mejor rendimiento que las más populares.

También se analizará la posibilidad de reducir la dimensionalidad quitando algunas características que pueden no estar muy correlacionadas con la clase, haciendo así que el paso de extracción de características sea algo más rápido. Para ello se ha usado la funcionalidad "Select attributes" del mismo Weka, que crea un listado de los atributos usados en la clasificación según su importancia. Esto se explicará en la sección de resultados.

En este caso se ha usado el estadístico Chi-cuadrado para medir la relación entre cada atributo y la clase. El resultado de este evaluador es una lista de atributos ordenada según su correlación con la clase, como se ve en la Figura 25.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
        Chi-squared Ranking Filter

Ranked attributes:
13445.0894    1 CI
 6400.062     2 Size
  898.4266    7 Density5
  877.4706   16 Gray2
  787.8556    4 Density2
  652.1547   19 Gray5
  646.3386   17 Gray3
  633.7126    5 Density3
  554.7397    3 Density1
  447.2756    8 Density6
  377.5661   10 Shape2
  375.1647   14 Shape6
  369.9133   13 Shape5
  337.3138   15 Gray1
  334.7028    6 Density4
  301.5275   11 Shape3
  300.987     9 Shape1
  279.8757   18 Gray4
  218.3338   12 Shape4
  180.119    20 Gray6

Selected attributes: 1,2,7,16,4,19,17,5,3,8,10,14,13,15,6,11,9,18,12,20 : 20
```

Figura 25: Resultado del evaluador Chi-cuadrado en las características extraídas.

Se compararán los resultados obtenidos en una clasificación en una sola fase y los resultados de una clasificación en dos fases.

Para la clasificación en dos fases se hará una primera clasificación teniendo en cuenta únicamente las características morfológicas, la longitud relativa y el IC. Con esto se clasificarán los cromosomas en sus grupos Denver correspondientes.

Después se hará una segunda clasificación dentro de cada grupo, siendo posible usar métodos de clasificación diferentes al usado en la primera fase.

Para evaluar el clasificador se utilizará la técnica 10-fold cross-validation, que primero parte el conjunto de datos en 10 hojas. Utiliza luego, cada una de esas hojas como test, tomando como conjunto de entrenamiento las 9 hojas restantes.

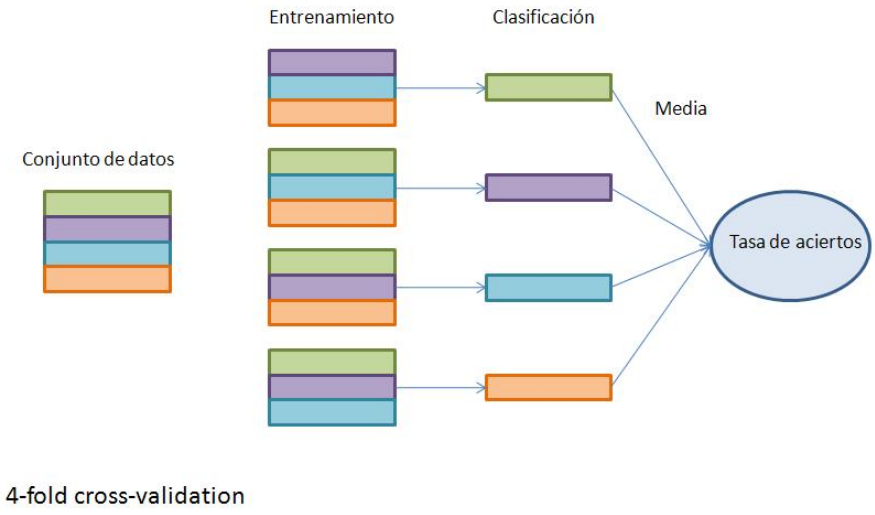

Figura 26: Esquema de 4-fold Cross-validation.

En la Figura 26 se ve un ejemplo de un método 4-fold cross-validation, en la que se parte el conjunto de datos inicial en 4 hojas. Para testear cada una de las partes se entrena el clasificador con las partes restantes y se hace una media de las tasas de acierto de cada test realizado.

Se hará el cross-validation en 10 ocasiones para cada clasificador, utilizando diferentes números de semilla para que la tasa de acierto sea lo más realista posible.

4. Resultados

A continuación se compararán los resultados obtenidos en la clasificación en una sola fase y las obtenidas en la clasificación en dos fases.

4.1. Clasificación en una fase

La clasificación en una fase se ha hecho con un set de entrenamiento de 1288 casos, de 24 clases (1,2,...,X,Y). El set tiene 56 casos de clases 1 a 22, 44 casos de clase X y 12 casos de clase Y. Se han utilizado todos los atributos en esta clasificación, puesto que se ha aplicado el estadístico chi-cuadrado al dataset y se ha obtenido el siguiente resultado.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
        Chi-squared Ranking Filter

Ranked attributes:
13445.0894    1 CI
 6400.062     2 Size
  898.4266    7 Density5
  877.4706   16 Gray2
  787.8556    4 Density2
  652.1547   19 Gray5
  646.3386   17 Gray3
  633.7126    5 Density3
  554.7397    3 Density1
  447.2756    8 Density6
  377.5661   10 Shape2
  375.1647   14 Shape6
  369.9133   13 Shape5
  337.3138   15 Gray1
  334.7028    6 Density4
  301.5275   11 Shape3
  300.987     9 Shape1
  279.8757   18 Gray4
  218.3338   12 Shape4
  180.119    20 Gray6

Selected attributes: 1,2,7,16,4,19,17,5,3,8,10,14,13,15,6,11,9,18,12,20 : 20
```

Figura 27: Resultado del estadístico Chi-cuadrado.

Se ve claramente que los atributos que más información ofrecen son la longitud relativa y el IC, lo que era previsible, y algunos otros como la cuarta componente del perfil de forma o el sexto componente del perfil de media de gris ofrecen mucha menos información.

No se ha visto claro un punto de corte para descartar atributos viendo sus valores de correlación con la variable clase, así que se han hecho varias pruebas.

Se ha utilizado el método Correlated Feature Selection (CFS), que ha seleccionado los 8 atributos que según su algoritmo mejor describen la variable clase.

Este algoritmo propone el set de atributos que mejor describe la clase teniendo en cuenta el nivel de correlación medido mediante el cálculo de la información mutua entre la variable predictora y la clase a predecir, y la redundancia entre las variables predictoras [40].

Se ha probado a utilizar únicamente las variables devueltas para la clasificación (longitud relativa, IC, componentes cuarta y quinta del perfil de densidad, sexta componente del perfil de forma y primera, segunda y tercera componente del perfil de media de grises) y se ha visto que la tasa de acierto disminuye.

También se ha probado a quitar los últimos 7 atributos en la lista de chi-cuadrado (atributos con baja correlación con la variable clase), y el resultado también ha sido peor que el resultado del set de atributos completo.

Se ha decidido viendo estos resultados, utilizar todos los atributos en la clasificación.

Los resultados obtenidos en la clasificación con los clasificadores mencionados han sido los siguientes:

Clasificador	Media	Desviación Típica
Perceptrón multicapa	0.8271	0.00546
Naive Bayes	0.8617	0.00322
Red Bayesiana	0.8383	0.00629
J48	0.8261	0.00686
Random Forest	0.8672	0.00712

Tabla 3: Tasa de aciertos de diferentes clasificadores en una fase.

Para la clasificación con redes Bayesianas se ha utilizado el algoritmo de búsqueda K2 con el parámetro "InitAsNaiveBayes" en falso y el número máximo de padres de cada nodo fijado en 3.

Se ha optado por fijar el parámetro "InitAsNaiveBayes" en falso, debido a que si no se hace así el sistema utiliza una estructura inicial de aprendizaje de Naive Bayes. Fijándolo en falso el sistema utiliza una red vacía como estructura inicial.

Se ha fijado el número máximo de padres de los nodos en 3 viendo que es necesario que este valor sea más grande que 1 para evitar entrenar la red como Naive Bayes, y que con el parámetro fijado en 3 se obtienen los mejores resultados.

Para la clasificación con Random Forest se ha fijado el valor del parámetro m (número de parámetros tomados aleatoriamente para la construcción de los árboles) en 15 y el número de árboles en 10, ya que son los parámetros que mejor resultado han ofrecido. Además de la semilla del cross-validation,

en este caso también se ha cambiado la semilla para la generación de los árboles.

4.2. Clasificación en dos fases

Esta clasificación se ha hecho en dos etapas. En la primera etapa se han clasificado los cromosomas en sus correspondientes grupos Denver (grupos A, B,..., G) y una vez hecho esto se ha hecho una clasificación por cada grupo.

En las clasificaciones de cada grupo se han hecho 5 pruebas con 10-fold cross-validation para cada clasificador (cambiando la semilla en cada prueba).

Cuando se utilizan redes Bayesianas se ha puesto el algoritmo de búsqueda en K2, con el parámetro "InitAsNaiveBayes" en falso en todos los casos.

En el caso de las pruebas con Random Forests también se ha cambiado la semilla de la generación de árboles en las pruebas.

4.2.1. Clasificación en grupos Denver

Para esta primera fase de la clasificación se han tenido en cuenta únicamente las características morfológicas, la longitud relativa y el IC, y se ha contado con un set de datos de 1288 casos.

Se ha hecho un 10-fold cross-validation como en la clasificación en una fase, con 10 pruebas, con 10 números de semilla diferentes.

Clasificador	Media	Desviación Típica
Perceptrón multicapa	0.9337	0.00294
Naive Bayes	0.9681	0.00098
Red Bayesiana	0.962	0.00273
J48	0.9643	0.00171
Random Forest	0.9698	0.00261

Tabla 4: Tasa de aciertos de diferentes clasificadores en clasificación en Grupos Denver.

Tras esta primera fase se ha llevado a cabo una segunda clasificación en cada uno de los grupos obtenidos en la primera fase.

La clasificación en esta fase no ha sido exacta, aunque se han alcanzado buenas tasas de acierto cercanas al 97 % (tasas de acierto más altas en el caso de los Random Forests). Esto puede deberse a que los centrómeros han sido localizados aproximadamente, como se ha dicho en un apartado anterior, y los índices de centrómero son aproximados, lo que puede hacer que la clasificación falle en algunos casos.

4.2.2. Clasificación en el grupo A

Para la clasificación de los cromosomas del grupo A se ha contado con un grupo de datos de 168 casos, de cromosomas de clase 1, 2 y 3 (56 casos por cada clase). Se ha hecho inicialmente un análisis de los atributos, para ver si alguno de éstos no aporta información para esta clasificación.

Se ha aplicado el método CFS con el método de búsqueda Best First y ha seleccionado 6 atributos (la longitud relativa, el IC, los componentes segundo y cuarto del perfil de densidad y los componentes tercero y quinto del perfil de media de grises).

Se ha aplicado también el estadístico chi-cuadrado, y se ha visto que atributos como el sexto componente del perfil de media de grises, el primer, tercer y quinto componente del perfil de forma y el sexto componente de densidad no están correlacionados con la variable clase, y por tanto, se pueden descartar para la clasificación en este grupo.

Tras varias pruebas, y ver que la tasa de aciertos no baja haciendo una reducción de atributos, se ha decidido coger solo las seis variables devueltas por el método CFS para la clasificación en este grupo.

Clasificador	Media	Desviación Típica
Perceptrón multicapa	0.9762	0
Naive Bayes	0.9869	0.00498
Red Bayesiana	0.9726	0.00677
J48	0.9595	0.00498
Random Forest	0.9786	0.00905

Tabla 5: Resultados de clasificación en el grupo A.

4.2.3. Clasificación en el grupo B

En esta clasificación se ha utilizado un grupo de datos de 112 casos de cromosomas de las clases 4 y 5 (56 casos por clase). Se ha seguido el mismo procedimiento que en la clasificación del grupo A. Se ha hecho un análisis de los atributos y se han descartado los que no aportan información a la clasificación.

El método CFS ha seleccionado 7 atributos (el primer, segundo y cuarto componente del perfil de densidad, el primer componente del perfil de forma y el segundo cuarto y quinto componente del perfil de media de gris) y se han tomado éstos para la clasificación tras comprobar que la tasa de acierto no baja al disminuir la cantidad de atributos.

Clasificador	Media	Desviación Típica
Perceptrón multicapa	0.8857	0.01467
Naive Bayes	0.8947	0.00748
Red Bayesiana	0.8857	0.01597
J48	0.8857	0.01466
Random Forest	0,8982	0.01621

Tabla 6: Resultados de clasificación en el grupo B.

4.2.4. Clasificación en el grupo C

Como en las clasificaciones de los otros grupos, aquí también se ha hecho el análisis de atributos para descartar los que no aportan información a la clasificación.

Se ha utilizado un set de datos de 436 casos de cromosomas de las clases 6-12 y de la clase X (56 casos de clase 6-12 y 44 casos de clase X).

Se ha decidido tomar 10 atributos, los 10 con mayor correlación con la clase según el estadístico chi-cuadrado (la longitud relativa, el IC, el primer, segundo, tercer, quinto y sexto componente del perfil de densidad y el primer, tercer y quinto componente del perfil de media de gris).

Clasificador	Media	Desviación Típica
Perceptrón multicapa	0.806	0.00733
Naive Bayes	0.8358	0.00349
Red Bayesiana	0.8128	0.00807
J48	0.7702	0.00833
Random Forest	0.822	0.01061

Tabla 7: Resultados de clasificación en el grupo C.

4.2.5. Clasificación en el grupo D

En la clasificación de este grupo se ha trabajado con un set de datos de 166 casos de cromosomas de clases 13, 14 y 15 (56 casos por clase).

Para reducir la cantidad de atributos en la clasificación, se ha probado el método CFS que ha seleccionado 4 atributos (el IC, tercer componente del perfil de densidad, el quinto componente del perfil de forma y el segundo componente del perfil de media de grises). El método chi-cuadrado ha demostrado que hay algunos de los atributos no están correlacionados con la clase, y se ha decidido tomar solo los seleccionados por el método CFS.

Clasificador	Media	Desviación Típica
Perceptrón multicapa	0.888	0.00686
Naive Bayes	0.9133	0.00689
Red Bayesiana	0.8807	0.00893
J48	0.8747	0.00782
Random Forest	0.9048	0.01158

Tabla 8: Resultados de clasificación en el grupo D.

4.2.6. Clasificación en el grupo E

Para la clasificación de este grupo se ha utilizado un set de datos de 166 casos de cromosomas de clases 16, 17 y 18 (56 casos por clase).

Para la reducción de la cantidad de atributos en la clasificación, se ha usado el método CFS que ha seleccionado 7 atributos (el IC, longitud relativa, primer, cuarto y sexto componente del perfil de densidad y primer y segundo componente del perfil de forma). Tras hacer algunas pruebas y ver que reduciendo el número de atributos la tasa de acierto del clasificador no baja, se ha decidido tomar únicamente los atributos seleccionados por el método evaluador de atributos.

Clasificador	Media	Desviación Típica
Perceptrón multicapa	0.9735	0.00684
Naive Bayes	0.9843	0.00329
Red Bayesiana	0.9783	0.00684
J48	0.9711	0.00502
Random Forest	0.9856	0.00689

Tabla 9: Resultados de la clasificación en el grupo E.

4.2.7. Clasificación en el grupo F

En la clasificación de este grupo se ha trabajado con un set de datos de 112 casos de cromosomas de clases 19 y 20 (56 casos por clase).

Para la reducción del número de atributos usados en la clasificación, el método CFS ha seleccionado solo 3 atributos (el IC, primer componente del perfil de densidad y el quinto componente del perfil de forma). El método chi-cuadrado ha demostrado, también, que muchos de los atributos no están correlacionados con la clase.

Se ha decidido tomar los 4 primeros atributos en la lista del chi-cuadrado (el IC, primer y quinto componente del perfil de densidad y el primer componente del perfil de media de grises).

Clasificador	Media	Desviación Típica
Perceptrón multicapa	0.8901	0.01173
Naive Bayes	0.8577	0.00757
Red Bayesiana	0.8811	0.01731
J48	0.8558	0.01426
Random Forest	0.8919	0.01102

Tabla 10: Resultados de la clasificación en el grupo F.

4.2.8. Clasificación en el grupo G

En la clasificación del grupo se ha trabajado con un set de datos de 112 casos de cromosomas de clases 21 y 22 (56 casos por clase).

Para reducir el número de atributos usados en la clasificación, el resultado del método CFS ha sido de 6 atributos (el IC, tercer y cuarto componente del perfil de densidad y primer y sexto componente del perfil de medias de grises), y el estadístico chi-cuadrado ha demostrado, también en este caso, que muchos de los atributos no están correlacionados con la clase.

Tras quedarnos solo con las variables devueltas por el método CFS las tasas de acierto de los clasificadores son las siguientes.

Clasificador	Media	Desviación Típica
Perceptrón multicapa	0.8696	0.01331
Naive Bayes	0.8783	0.00907
Red Bayesiana	0.8978	0.0124
J48	0.8739	0.02501
Random Forest	0.9043	0.01416

Tabla 11: Resultados de la clasificación en el grupo G.

4.3. Resultados generales

Se ha visto en las tablas que la clasificación en una fase no ofrece muy buen resultado, siendo la mejor tasa de acierto media de un 86.72 % en el caso de los Random Forests.

En el caso de la clasificación en dos fases se pueden ver tasas de acierto muy diferentes entre los distintos grupos, teniendo por ejemplo en el caso de los grupos A y E tasas cercanas al 98 % y en el caso del grupo C una tasa de aciertos de 83 %.

Esto puede deberse a problemas que se han descrito en secciones anteriores con las características de los patrones de bandas, y que el grupo C es el más grande de todos con 7 cromosomas, lo que puede dificultar más la clasificación.

Exceptuando la clasificación del grupo C, todos los demás grupos ofrecen mejores resultados que la clasificación en una fase, siendo algunos de los resultados bastante satisfactorios (alrededor del 98 %).

5. Conclusiones y trabajo futuro

En este trabajo se ha presentado un problema de búsqueda de un método capaz de realizar el cariotipado de los cromosomas de imágenes de metafase de manera automática y ágil, que facilite el trabajo a los genetistas.

Se ha analizado la viabilidad de crear un sistema que cree cariotipos a partir de 10 imágenes de metafase del mismo paciente. Con esto se pretende mejorar la clasificación y la calidad del cariotipo, aportando información que en el caso de cariotipos a partir de una sola imagen puede quedar oculta.

Tras las pruebas de clasificación realizadas, se ha demostrado que la clasificación de cromosomas en dos fases ofrece mejores resultados que la clasificación en una sola fase, exceptuando la clasificación del grupo C, que solo ha tenido una tasa de acierto del 83.58 %. Puede que haya que buscar otras características para la clasificación de ese grupo, puesto que es posible que tenga problemas con las características escogidas en este trabajo.

Una de las conclusiones más importantes del trabajo es que cada grupo Denver utiliza su propio subgrupo de las características calculadas para la clasificación, reduciendo las dimensiones de los atributos desde las 20 características originales hasta 3 ó 4 en algunos grupos. Teniendo esto en cuenta, a la hora de la extracción de características, habría que extraer todas las características de todos los cromosomas, pero tras hacer la clasificación en grupos Denver, podría usarse el set de atributos correspondiente para clasificar cada grupo.

A la hora de hacer el cariotipo, aparte de la fiabilidad del resultado, también es importante el rendimiento del sistema. Tiene que ser un sistema ágil que no puede pasar mucho tiempo en la clasificación, y se ha podido ver que la reducción de las dimensiones de los atributos provoca que la clasificación se haga más rápido.

También se ha probado que los árboles pueden ser muy eficientes en la clasificación de cromosomas, sobre todo los Random Forest. Ofrecen tasas de acierto parecidas a las ofrecidas por las redes neuronales, o incluso más altas en algunos casos, y son mucho más rápidos que éstas.

Teniendo en cuenta que no se ha integrado ningún algoritmo que detecte automáticamente el centrómero para calcular el IC, y que los valores utilizados en las pruebas han sido cálculos aproximados, y que puede que haya información errónea en el clasificador debido a la forma de tratar los cromosomas solapados del software de Leica, los resultados obtenidos en las clasificaciones son bastante satisfactorios, y puede decirse que la construcción de un sistema que realice el cariotipado automáticamente a partir de 10 imágenes de metafase de manera ágil (haciendo la clasificación según las características seleccionadas para cada grupo Denver) es viable.

Queda como trabajo futuro encontrar un algoritmo de detección automática del centrómero para probar si efectivamente los resultados que se obtienen en la clasificación son mejores que los obtenidos en este trabajo.

También ha quedado por resolver el problema encontrado en la segmentación de los clústeres de cromosomas, en los que hay que separar los cromosomas y hay que encontrar una forma de tratar las partes solapadas, muy importante para evitar duplicar información de las bandas en diferentes cromosomas.

En este proyecto se ha trabajado con imágenes de solo dos pacientes. Sería necesario utilizar imágenes de más pacientes para mejorar el clasificador.

También es necesario pensar que se ha trabajado solo con células sanas, pero no todas las imágenes que habrá que cariotipar lo serán. Algunos cromosomas podrán tener anomalías, y habrá que entrenar al sistema para que pueda detectar y clasificar esos cromosomas.

Por último, habrá que dar al usuario la posibilidad de corregir la clase de un cromosoma que se haya clasificado mal, facilitando así que el sistema "aprenda" y mejore para la siguiente clasificación.

Referencias

[1] http://www.dartmouth.edu/~cbbc/courses/bio4/bio4-lectures/theCell.html

[2] X. Wang, B. Zheng, M. Wood, S. Li, W. Chen and H. Liu. Development and evaluation of automated systems for detection and classification of banded chromosomes: current status and future perspectives. *J. Phys. D: Appl. Phys. 38 (2005) 2536–2542*

[3] Y. Wang, Q. Wu, K. Castleman and Z. Xiong. Cromosome Image Enhancement Using Multiscale Differential Operators. *EEE Transactions on Medical Imaging,* vol. 22, no. 5, may 2003

[4] J. Piper and E. Granum. On Fully Automatic Feature Measurement fot Banded Chromosome Classification. *Cytometry* 10:242-255 (1989)

[5] L. Vanderheydt, F. Dom, A. Oosterlinck and H. Berghe. Two-dimensional shape decomposition using fuzzy subset theory applied to automated chromosome analysis. *Pattern Recognition 13 147–57,* 1981

[6] C. Gaybay. Image structure representation and processing: a discussion of some segmentation methods in cytology. *IEEE Trans. Pattern Anal. Mach. Intell. 8 140–6*, 1986

[7] G. Rittery and L. Gao. Automatic segmentation of metaphase cells based on global context and variant analysis.

[8] V. Vijaya Kumar, A. Srikrishna, D. V. L. N. Somayajulu, B.Raveendra Babu. An Improved Iterative Morphological Decomposition Approach for Image Skeletonization. *ICGST-GVIP Journal, ISSN: 1687-398X* , Volume 8, Issue 1, June 2008.

[9] G. Rakesh and K. Rajpreet. Skeletonization Algorithm for Numeral Patterns. *International Journal of Signal Processing, Image Processing and Pattern Recognition*

[10] S.Chang. Extracting Skeletons from Distance Maps. *IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.7*, July 2007

[11] P.S. Karvelis, D.I. Fotiadis, M. Syrrou and I. Georgiou. Segmentation Of Chromosome Images Based On a Recursive Watershed Transform. *The 3rd European Medical and Biological Engineering Conference*, 2005

[12] W. Srisang, K. Jaroensutasinee and M. Jaroensutasinee. Segmentation of Overlapping Chromosome Images Using Computational Geometry. *Walailak J Sci & Tech 2006; 3(2):181-194.*

[13] M. Moradi, S. Kamaledin Setarehdan. New features for automatic classification of human chromosomes: A feasibility study. *Pattern Recognition Letters 27, 19–28* , 2006

[14] D. Ming and J. Tian. Automatic Pattern Extraction and Classification for Chromosome Images. *J Infrared Milli Terahz Waves 31:866–877*, 2010

[15] B. Lerner, H. Guterman and I. Dinstein. A Classification-Driven Partially Occluded Object Segmentation (CPOOS) Method with Application to Chromosome Analysis. *IEEE Transactions On Signal Processing 46 (10), 2841–2847*, 1998

[16] J. Cho. A Hierarchical Artificial Neural Network Model for Giemsa-Stained Human Chromosome Classification. *Biomed 06, IFMBE Proceedings 15, pp. 12-15*, 2007

[17] J. Graham. Automation of routine clinical chromosome analysis I-karyotyping by machine. *Anal Quant Cytol Histol9:383-390*, 1987

[18] G. Gallus,PW. Neurath. Improved computer chromosome analysis incorporating preprocessing and boundary analysis. *Phys Med Biol 15:435-445*, 1970

[19] G. Ritter and C. Pesch. Polarity-free automatic classification of chromosomes. *Computational Statistics & Data Analysis 35 351–372,* 2001

[20] G. Ritter and G. Schreib. Profile and feature extraction from chromosomes. *ICPR00(Vol II: 287-290)*, 2000

[21] J. Kao, J. Chuang, T. Wang. Chromosome classification based on the band profile similarity along approximate medial axis. *Pattern Recognition 41 (2008) 77 – 89*

[22] B. Lerner. Towards a Completely Automatic Neural Network-Based Human Chromosome Analysis. *IEEE Trans. On Sys., Man and Cybernetics, Vol. 28, No. 4, pp. 544-522*, 1998

[23] M. Javan-Roshtkhari and S. Kamaledin Setarehdan. A New Approach to Automatic Classification of the Curved Chromosomes. *Proceedings of the 5th International Symposium on image and Signal Processing and Analysis,* 2007

[24] I. M. M. El Emary. On the Application of Artificial Neural Networks in Analyzing and Classifying the Human Chromosomes. *Journal of Computer Science 2 (1): 72-75*, 2006

[25] B. C. Oskouei and J. Shanbehzadeh. Chromosome Classification Based on Wavelet Neural Network. *DOI 10.1109/DICTA.2010.107*

[26] B. Lerner. Toward A Completely Automatic Neural-Network-Based Human Chromosome Analysis. *IEEE transactions on systems, man, and cybernetics—part b: cybernetics, VOL. 28, NO. 4*, august 1998

[27] C. Martínez, A. Juan, and F. Casacuberta. Using Recurrent Neural Networks for Automatic Chromosome Classification. *J.R. Dorronsoro (Ed.): ICANN 2002, LNCS 2415, pp. 565–570,* 2002.

[28] S. Rungruangbaiyok and P. Phukpattaranont. Chromosome image classification using a two-step probabilistic neural network.*Songklanakarin J. Sci. Technol. 256 32 (3), 255-262*, 2010

[29] P.D. Wasserman. Advanced Methods in Neural Computing, *Van, Nostrand Reinhold, New York, pp. 35-55.* 1993.

[30] M. Elif Karsligil, M. Yahya Karsligil. Fuzzy Similarity Relations for Chromosome Classification and Identification. *F. Solina and A. Leonardis (Eds.): CAIP'99, LNCS 1689, pp. 142-148*, 1999

[31] C. Martínez, H. García, A. Juan, and F. Casacuberta. Chromosome Classification Using Continuous Hidden Markov Models. *F.J. Perales et al. (Eds.): IbPRIA 2003, LNCS 2652, pp. 494–501,* 2003.

[32] P. Kleinschmidt and I. Mitterreiter. Improved Chromosome Classification Using Monotonic Functions of Mahalanobis Distance and the Transportation Method. ZOR - *Mathematical Methods of Operations Research (1994) 40:305-323*

[33] Z. Liu and D. Chen. Classification of Chromosome Sequences with Entropy Kernel and LKPLS Algorithm. *ICIC 2005, Part I, LNCS 3644, pp. 543–551*, 2005

[34] Q. Wu, Z. Liu, T. Chen, Z. Xiong and K. R. Castleman. Subspace-Based Prototyping and Classification of Chromosome Images. *IEEE transactions on image processing, VOL. 14, NO. 9,* September 2005

[35] F. Vogel, A. G. Motulsky. Human genetics: problems and approaches. ISBN 978-3-540-37653-8

[36] http://www.cs.waikato.ac.nz/ml/weka/

[37] I. Ben-Gal. Bayesian Networks. *Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons*, 2007

[38] L. Noriega. Multilayer Perceptron Tutorial

[39] L. Breiman. Random Forests.

[40] M. A. Hall. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning.