

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/304156342>

Conjunto de herramientas citogenéticas para el trabajo con cromosomas

Thesis · June 2016

DOI: 10.13140/RG.2.1.2840.0247

CITATIONS

0

READS

51

2 authors, including:



[luis alberto mendez-rosado](#)

National Center of Medical Genetic, Habana. Cuba

60 PUBLICATIONS 41 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Noninvasive prenatal testing for aneuploidy using transcervical cell sample collected at 5 -12 weeks of gestation [View project](#)



Characterization of structural aberrations of human chromosomes in Cuba [View project](#)

All content following this page was uploaded by [luis alberto mendez-rosado](#) on 20 June 2016.

The user has requested enhancement of the downloaded file. All in-text references underlined in blue are added to the original document and are linked to publications on ResearchGate, letting you access and read them immediately.

Universidad de La Habana
Facultad de Matemática y Computación



Conjunto de herramientas citogenéticas para el trabajo con cromosomas

Autor: **Liana Beatriz Juliá Roget**
Ricardo Fundora Hernández

Tutores: **Mcs. Luis Alberto Méndez**
Lic. Claudia Paredes Plasencia

Trabajo de Diploma
presentado en opción al título de
Licenciado en Ciencia de la Computación



Junio de 2016

A nuestros padres, familiares y amigos.

Agradecimientos

A nuestros padres, por la comprensión, el apoyo incondicional y por permitirnos dedicarnos por completo a nuestra carrera.

A nuestros compañeros de año, que de una forma u otra nos brindaron ayuda para poder llegar hasta este momento.

A nuestros profesores por enseñarnos a siempre ser mejores profesionales y mejores personas.

A nuestros amigos y familiares por el ánimo y la compañía.

Al tutor Luis Alberto por habernos brindado la oportunidad de realizar este tema de tesis.

A nuestra tutora Claudia, por la comprensión, el apoyo y el cariño que nos brindó.

Al profe Oscar Luis y al colectivo de imágenes por su gran ayuda y preocupación.

Opinión del tutor

La citogenética es la rama de la genética que estudia las enfermedades de herencia cromosómica. Se ha determinado que aproximadamente uno de cada 160 nacidos vivos tienen una alteración cromosómica y que al menos 50% de los abortos espontáneos se deben a anomalías cromosómicas. Dada estas cifras, se hace indispensable la creación del cariotipo con el objetivo de analizar e identificar los cromosomas para la determinación del tipo de alteración que permite la confirmación del diagnóstico de dichas enfermedades y además el correcto asesoramiento genético del paciente y sus familias.

Conformar el cariotipo es una tarea complicada que requiere mucho tiempo del citogenetista. En la actualidad se utilizan sistemas, en su mayoría comerciales, que apoyan la toma de decisiones en la identificación de cada cromosoma. Ricardo y Liana se dieron la tarea de buscar las soluciones existentes con el objetivo de construir un sistema que supla las necesidades del Centro de Genética Médica de Cuba para ayudar a automatizar dicho proceso. A partir de una imagen obtenida del microscopio fueron capaces de segmentar y clasificar los cromosomas de forma semi-automática.

Para lograr el objetivo trazado, Ricardo y Liana tuvieron que asumir un gran cúmulo de conocimientos. Muchos de los cuáles van más allá de los contenidos que se imparten en nuestra carrera. Ambos, tuvieron que dominar los conceptos generales y específicos de citogenética para construir el cariotipo y estudiar como los algoritmos de Visión Computacional pueden adaptarse a la segmentación y clasificación de los cromosomas.

En una primera etapa, Ricardo tuvo que segmentar las imágenes obtenidas del microscopio en cromosomas, no sin antes eliminar el ruido y cualquier otro elemento que no constituya un objeto de interés. Incluso, después de obtener una imagen aparentemente limpia, es común encontrar los cromosomas unidos o solapados, lo cual dificulta mucho la tarea de segmentación. Una vez segmentada la imagen en cromosomas, Liana se dió la difícil tarea de entender las características más importantes de los cromosomas con el objetivo de conformar el vector de características utilizado posteriormente

en la clasificación de cada cromosoma y finalmente, en la construcción del cariotipo.

Incluso con una exitosa investigación y un sistema utilizable por los especialistas de la citogenética, escribir una tesis no es tarea fácil. En particular si se ha de describir y explicar una serie de complejos y novedosos algoritmos. A pesar de ello, Ricardo y Liana no se amilanaron y con gran dedicación lograron entregarnos una meritaria obra escrita que de segura será vastamente consultada por futuros investigadores.

Por todo lo expuesto anteriormente, considero que los estudiantes Liana Beatriz Juliá Roget y Ricardo Fundora Hernández cumplen con todos los requisitos necesarios para obtener el título de Licenciados en Ciencia de la Computación. Para finalizar, quisiera instarlos a no abandonar esta investigación y culminarla con la aplicación del software en el Centro de Genética Médica de Cuba.

Resumen

El análisis e identificación de los cromosomas es un recurso importante en la medicina, para la rápida detección, pronóstico y evaluación de enfermedades genéticas. Este análisis se realiza con la ayuda de una estructura estándar llamada cariotipo. Conformar el cariotipo es una tarea complicada que requiere mucho tiempo de los especialistas del campo. Para facilitar el trabajo de los profesionales se diseñan sistemas que apoyan la toma de decisiones en la identificación de cada cromosoma. En este estudio se propone un sistema para realizar el cariotipo de forma semiautomática, que alcanza una precisión media del 54% en el conjunto de prueba propuesto en la literatura.

Abstract

The analysis and classification of chromosomes is an important task in medicine, applied to the rapid detection, prognosis and assessment of genetic diseases. This analysis is done with the help of a standard structure called karyotype. The karyotype creation is a complicated task that requires a lot of time from the specialist in this field. To facilitate the work of professionals, we propose a system that helps in the decision making process of chromosome classification. The research presented proposes a system for semi-automatic karyotype construction with an average accuracy of 54% on the dataset proposed in the literature.

Índice general

Índice de figuras	xii
Índice de tablas	xiii
Índice de algoritmos	xiii
1. Introducción	1
2. Revisión Bibliográfica	5
2.1. Pre-procesamiento y Segmentación	5
2.1.1. Definición del problema	5
2.1.2. Algoritmos estudiados	5
2.2. Extracción de características	8
2.2.1. Definición del problema	8
2.2.2. Descripción de las características a extraer	8
2.3. Clasificación	9
2.3.1. Definición del problema	10
2.3.2. Algoritmos estudiados	10
2.4. Soluciones conjuntas	10
3. Algoritmos propuestos	12
3.1. Pre-procesamiento y Segmentación	12
3.1.1. Limpieza de la imagen y detección de cromosomas . .	12
3.1.2. Detección de los <i>clusters</i> de los cromosomas parcialmente unidos o solapados	14
3.1.3. Separación de los cromosomas	18
3.2. Extracción de Características	21
3.2.1. Método para enderezar los cromosomas	21
3.2.2. Características extraídas	26
3.3. Clasificación	28

4. Implementación y Experimentación	30
4.1. Experimentación y Resultados	30
4.1.1. Pre-procesamiento	31
4.1.2. Segmentación	33
4.1.3. Extracción de características y clasificación	34
4.2. Diseño de Aplicación	36
4.2.1. Módulo de Segmentación	36
4.2.2. Módulo de Aprendizaje de Máquinas	38
4.2.3. Módulo de la Base de Datos	39
Conclusiones	40
Recomendaciones	41
Bibliografía	42
A. Anexos	45

Índice de figuras

1.1.	Cromátidas unidas por el centrómero	1
1.2.	(a) Muestra de una gota de sangre, (b) Microscopio con una muestra	2
1.3.	(a) Imagen capturada por el microscopio que muestra una metafase celular, (b) Cariotipo de los cromosomas humanos.	3
2.1.	Cromosomas solapados con diagramas de Veronoi y triangulación Delaunay. (a) Posibles puntos de cortes en la linea de contorno, (b) Diagrama de Veronoi, (c) Triangulacion Delaunay.	6
2.2.	Puntos de corte obtenidos	7
2.3.	Ejemplo de captura utilizando la técnica MFISH.	7
2.4.	En azul se muestran las características: (a) linea central, (b) área, (c) perímetro del menor polígono convexo, (d) centrómero.	9
2.5.	(a) MFISH <i>cluster</i> , (b) Segmentacion y clasificacion incorrectas, (c) Segmentacion y clasificacion correctas.	11
3.1.	(a) Imagen original, (b) Método del Umbral aplicado.	13
3.2.	(a) Resultado al aplicar el método del Umbral,(b) Resultado al aplicar el método de Apertura.	13
3.3.	(a) y (b) Cromosomas parcialmente unidos, (c) Cromosomas solapados.	14
3.4.	Se muestra en azul el menor polígono convexo que contiene a cada objeto.	15
3.5.	Se muestra en azul la elipse que contiene a cada objeto.	16
3.6.	(a) Cromosomas solapados, (b) Esqueleto correspondiente a los cromosomas en (a).	18
3.7.	Cromosoma en escala de grises.	21
3.9.	(a) Imagen binaria de un cromosoma, (b) Representación gráfica del vector de proyección horizontal.	22

3.8. (a) Cromosoma en escala de grises, (b) Imagen binaria correspondiente.	22
3.10. (a) Cromosoma con la línea de corte seleccionada anteriormente, (b) Ambos brazos después de enderezados, (c) Resultado después de conectar los dos brazos de los cromosomas.	25
4.1. Metafase celular de uno de los casos de estudio perteneciente al Centro de Genética Médica de Cuba.	30
4.2. Resultado de aplicar el método de apertura con c_i iteraciones.	31
4.3. Resultados para la variación de \minArea y \maxArea . (a) Original, (b) $\minArea = 300$ y $\maxArea = 10000$, (c) $\minArea = 100$ y $\maxArea = 5000$, (d) $\minArea = 500$ y $\maxArea = 30000$,	32
4.4. Resultados de variar el parámetro δ en dos casos de estudio.	33
4.5. Pares de cromosomas, izquierda original, derecha enderezados.	34
4.6. Estructura del vector de características obtenido para la clasificación.	34
4.7. Vista para añadir una nueva imagen a la base de datos.	37
4.8. Vista resultante al identificar los <i>clusters</i> automáticamente	37
4.9. Propuesta de corte brindada por el sistema.	38
4.10. Vista del cariotipo realizado	38
4.11. Vista donde el especialista realiza el cariotipado.	39
A.1. (a) Vista principal de la aplicación, (b) Vista de un cromosoma solapado seleccionado.	46
A.2. (a) Vista para realizar el corte de forma manual, (b) Selección por el usuario de los puntos de corte.	47
A.3. Vista principal del cariotipado.	48
A.4. Vista del cariotipado en una etapa intermedia.	48

Índice de tablas

3.1.	Kernel utilizado en el método de Apertura	13
3.2.	Valores posibles de w_1 y w_2 para el cálculo del índice de rotación.	24
4.1.	Descripción de la base de datos utilizada para la experimentación.	31
4.2.	Resultados obtenidos en la detección de los cromosomas simples y parcialmente unidos o solapados (PU/S).	32
4.3.	Resultados obtenidos en la segmentación de cromosomas parcialmente unidos.	33
4.4.	Mejores soluciones para cada algoritmo.	35
4.5.	Mejores soluciones para cada algoritmo.	35
4.6.	Resultados de la prueba estándar de normalidad Shapiro-Wilk aplicada al corpus de experimentación.	36
A.1.	Resultados de la Segmentación del Caso 1 al Caso 11 de estudio.	49
A.2.	Resultados de la Segmentación del Caso 12 al Caso 22 de estudio.	50
A.3.	Resultados de la Segmentación del Caso 23 al Caso 33 de estudio.	51
A.4.	Parámetros utilizados para cada uno de los algoritmos de clasificación.	52
A.5.	Resultados de los experimentos realizados a los algoritmos.	53

Índice de algoritmos

3.1.	Método del menor polígono convexo	15
3.2.	Método de la Elipse	17
3.3.	Algoritmo para detectar cromosomas parcialmente unidos o solapados	18
3.4.	Algoritmo para remover puntos que no son de corte	20
3.5.	Algoritmo para enderezar un brazo de un cromosoma	24
3.6.	Algoritmo para enderezar cromosomas	26

Capítulo 1

Introducción

Los cromosomas son el material hereditario que carga la información genética de generaciones dentro de las células. Cada cromosoma metafásico¹ está constituido por dos cromátidas² unidas por el centrómero (ver Figura 1.1). Este centrómero divide al cromosoma en dos brazos que se designan *p* para el brazo más corto y *q* para el brazo más largo. En la etapa de la división celular la cromatina³ se condensa para formar los cromosomas, haciéndolos visibles. Los seres humanos presentan 46 cromosomas agrupados en 23 pares, de ellos 22 pares son llamados autosomas⁴ y el par restante cromosomas sexuales.

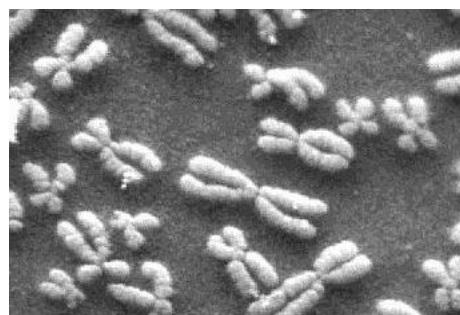


Figura 1.1: Cromátidas unidas por el centrómero

La identificación y el análisis de los cromosomas es necesario para la

¹es una de las etapas presentes en la división celular.

²es la estructura longitudinal que compone al cromosoma.

³es un conjunto de ADN y de proteínas que se encuentran en el núcleo de las células eucariotas, que compone químicamente a los cromosomas

⁴es cualquier cromosoma que no sea sexual

detección de enfermedades genéticas como la leucemia y el síndrome de Down. Se inicia con la extracción de muestras de sangre de los individuos (ver Figura 1.2(a)), asignándole a cada muestra un caso de estudio. Realizar este análisis a cada caso directamente del microscopio es altamente costoso para el citogenetista⁵ (ver Figura 1.2(b)). Toma mucho esfuerzo visual identificar los cromosomas en un espacio tan reducido como es el lente del microscopio; haciéndoles imposible la comparación directa con otros casos estudiados.

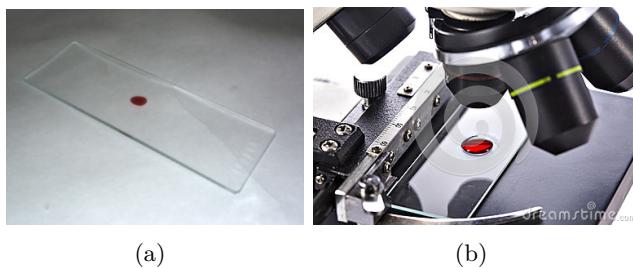


Figura 1.2: (a) Muestra de una gota de sangre, (b) Microscopio con una muestra

Con los avances en el mundo de la computación, surgieron propuestas de sistemas que apoyaban la toma de decisiones en la identificación de cada cromosoma. Los sistemas parten de la captura de una imagen celular tomada por el microscopio a una muestra de sangre.

El cariotipo es la estructura que se emplea para realizar el estudio y análisis del genoma humano, donde se representan las imágenes de los cromosomas agrupados en pares.

En la literatura se propone una secuencia de pasos a seguir para realizar el cariotipado, proceso de elaborar el cariotipo de forma automática: pre-procesamiento y segmentación, extracción de características y clasificación. La imagen de la metafase celular (ver Figura 1.3(a)) tomada por el microscopio contiene ruidos, tiene un contraste bajo y es necesaria la separación de los cromosomas en imágenes individuales. Los métodos de pre-procesamiento y segmentación son los encargados de mejorar la calidad, el contraste de la imagen y la separación de los cromosomas. Una vez obtenidas las imágenes individuales comienza la etapa de extracción de características, donde cada una de las imágenes es procesada con el fin de encontrar un vector numérico que las describa. Luego, con el conjunto de vectores extraídos se realiza la clasificación y el emparejamiento (ver Figura 1.3(b)).

⁵ es el especialista en genética encargado del estudio del genoma

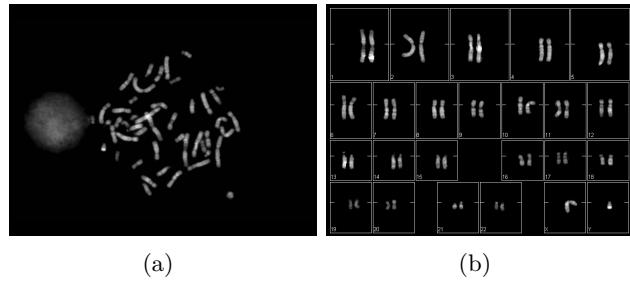


Figura 1.3: (a) Imagen capturada por el microscopio que muestra una metáfase celular, (b) Cariotipo de los cromosomas humanos.

El objetivo principal de esta investigación es desarrollar un sistema que sirva como una herramienta para el análisis del genoma humano. De aquí se derivan los siguientes objetivos específicos:

- Hacer un estudio de los principales resultados teóricos relacionados con el proceso de cariotipado necesarios para la presente investigación.
 - Identificar los pasos para resolver el problema de estudio.
 - Encontrar las principales soluciones presentes en la literatura y seleccionar las que se van a implementar.
- Implementar un sistema que permita elaborar de forma semiautomática el cariotipo, permitiendo el análisis y almacenamiento de los casos de estudio analizados en dicho sistema.
 - Cargar imágenes tomadas por el microscopio.
 - Separar los cromosomas del fondo de la imagen.
 - Identificar los cromosomas simples, parcialmente unidos y solapados.
 - Convertir los cromosomas parcialmente unidos o solapados en simples.
 - Extraer las características que mejor identifiquen al conjunto de datos.
 - Realizar la identificación y agrupación de los cromosomas en el cariotipo.
 - Permitir la comparación de distintos casos de estudio.

La tesis está estructurada de la siguiente forma: en el capítulo 2 se expone una descripción de los principales resultados presentes en la literatura que dan solución a las etapas de pre-procesamiento y segmentación, extracción de características y clasificación. En el capítulo 3 se muestran los algoritmos escogidos para cada una de las etapas anteriores en la conformación del sistema. En el capítulo 4 se presenta el diseño del sistema y los resultados obtenidos en la experimentación. Al finalizar se muestran las conclusiones y se presentan las recomendaciones para trabajos futuros, seguido de las referencias bibliográficas y anexos para más información.

Capítulo 2

Revisión Bibliográfica

En este capítulo se mostrará una descripción de los principales estudios realizados en las etapas de pre-procesamiento y segmentación (2.1), extracción de características (2.2) y clasificación (2.3). Además se exponen métodos donde las tres etapas anteriores se realizan de forma conjunta (2.4).

2.1. Pre-procesamiento y Segmentación

La segmentación es un paso importante en los sistemas de análisis automático de los cromosomas. Es un fenómeno común encontrar cromosomas parcialmente unidos y solapados en las imágenes capturadas durante la metáfase. Cómo separar correctamente estos cromosomas es un problema vital en nuestros días.

2.1.1. Definición del problema

A partir de imágenes en escala de grises se realizará la limpieza, detección y segmentación de los cromosomas. La limpieza de la imagen consiste en la eliminación del ruido. Se identifica como ruido todos aquellos objetos presentes en la captura que no sean cromosomas, dígase núcleo celular, restos de suciedad, entre otros. La detección consiste en identificar los *clusters* que representan cromosomas simples, parcialmente unidos o solapados. La segmentación es la encargada de convertir los *clusters* en objetos simples.

2.1.2. Algoritmos estudiados

Las imágenes captadas por el microscopio pueden presentar diversas dificultades. Los cromosomas pueden estar unidos o solapados y las secciones

consecutivas de claros y oscuros en el cuerpo del cromosoma, mejor conocidas como bandas, pueden propagarse. El primer paso en el análisis de las imágenes es la separación de los cromosomas del fondo. El método más usado en este paso está basado en la evaluación de un umbral global usando como media el método de Otsu [9] o un esquema de *re-thresholding* [10].

En la literatura existen algunos algoritmos para tratar con *clusters* que contienen cromosomas parcialmente unidos pero que no se solapan [12, 13] y para *clusters* solapados pero que no estén parcialmente unidos [1].

En [20], se desarrolla una técnica para identificar los *clusters* simples y compuestos. Se enfoca en diseñar un clasificador usando una red neuronal cuya entrada es una imagen y su salida es binaria, si es cero la imagen es de un solo cromosoma, de lo contrario la imagen contiene varios cromosomas. Se utiliza una red neuronal para la clasificación. El vector de entrada incluye: superficie de la imagen, superficie del cromosoma, número de *pixels* del contorno del cromosoma y seis momentos. El mejor resultado obtenido es de un 73%.

Una de las técnicas existentes para separar los cromosomas solapados hace uso de la geometría computacional. Dicha técnica requiere la identificación de todos los posibles puntos de corte de la línea del contorno de los cromosomas solapados, utilizando Diagramas de Voronoi y Triangulación Delaunay para seleccionar cuatro puntos de corte y separar los cromosomas solapados en cromosomas independientes (ver Imagen 2.1). Este algoritmo fue probado en 35 cromosomas solapados y separó correctamente 28 de los 35 casos de prueba (80%). Tres imágenes fueron separadas incorrectamente (8.6%) y cuatro no pudieron ser separadas por el algoritmo (11.4%) [25].

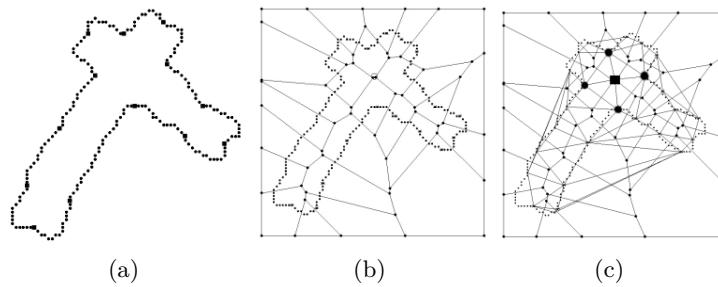


Figura 2.1: Cromosomas solapados con diagramas de Veronoi y triangulación Delaunay. (a) Posibles puntos de cortes en la linea de contorno, (b) Diagrama de Veronoi, (c) Triangulacion Delaunay.

En [16] se propone una técnica que divide la segmentación en dos fases.

En la primera son detectados los cromosomas parcialmente unidos y solapados (*clusters*) usando tres criterios geométricos y en la segunda estos son separados usando una línea de corte 2.2. Si existen más de dos cromosomas solapados en el mismo *cluster* el algoritmo separa a dos y luego, aplica el mismo paso para separar los restantes. Una ventaja de este estudio es que utiliza la geometría de los cromosomas independientemente del tipo de imagen en la que se encuentre por lo que puede aplicarse a cualquier tipo de imágenes, tanto a binarias o multi-espectrales. Este método fue aplicado a una base de datos que contenía 62 cromosomas parcialmente unidos y solapados, obteniendo un 91,9% de acierto.



Figura 2.2: Puntos de corte obtenidos

Hibridación fluorescente *in situ* Multicolor (MFISH) es una técnica para la captura de los cromosomas. Esta produce una imagen en donde cada tipo de cromosoma aparece de un color distinto [24]. MFISH hace que el análisis computarizado de las imágenes sea más fácil. En la Figura 2.3 se observa un ejemplo de esta técnica.

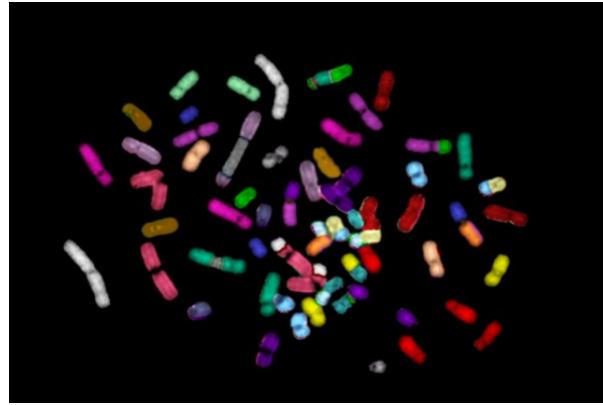


Figura 2.3: Ejemplo de captura utilizando la técnica MFISH.

Como resultado del estudio propuesto en [23], se considera usar la en-

tropía como un criterio para seleccionar líneas de corte para descomponer grupos de cromosomas que estén parcialmente unidos o se solapen. Este algoritmo utiliza la información multi-espectral en imágenes de cromosomas para lograr una mejor segmentación. En el estudio se encuentran las componentes conexas, se clasifican como objetos separados y se calcula su entropía. Si la entropía es menor que el umbral, entonces examinan el objeto para encontrar posibles líneas de corte. Consideran solamente puntos de las líneas de corte que posean 8 vecinos conectados cuyas clases difieran de ellos. Las posibles líneas de corte son todas las posibles combinaciones de las rectas formadas por todos los puntos de corte. Una vez que la división por entropía se realiza, los cromosomas solapados se clasifican como objetos separados. Los resultados muestran que este algoritmo es una buena solución a la segmentación de cromosomas.

2.2. Extracción de características

La extracción de características es crucial para lograr una clasificación adecuada. Este proceso es complicado debido a las múltiples formas que puede presentar un cromosoma, ya que existen infinitas posiciones en que estos pueden ser capturados por la cámara del microscopio en los distintos casos de estudio. En esta sección se abordará el problema que es objeto de estudio y se mostrarán las técnicas usadas actualmente para resolverlo.

2.2.1. Definición del problema

A partir de una imagen en escala de grises se genera un vector de números que represente su estructura. En este vector deben estar presentes los principales rasgos que describan al conjunto de datos. En la actualidad se investigan diferentes rasgos que incluyen características numéricas, morfológicas, de texturas, densidad de perfiles y dominio de frecuencias que se obtienen del resultado de la transformada de Fourier o Wavelet, para lograr optimizar la representación de dichas imágenes. Las características de las texturas incluyen aristas, contrastes, correlación, entropía, entre otras; las morfológicas incluyen longitud, área, perímetro convexo y las numéricas al número de cromosomas.

2.2.2. Descripción de las características a extraer

La longitud se obtiene utilizando el *medial axis*, la línea central del cromosoma (ver Figura 2.4(a)) que pasa por el centrómero, por lo que se con-

sidera la longitud de dicha línea como la longitud del cromosoma [19]. La longitud relativa es la razón entre la longitud del cromosoma y la suma de las longitudes de todos los cromosomas en el conjunto. En [3] se plantea que la longitud relativa es una mejor característica que la longitud.

El área es el número de *pixels* presentes en el cromosoma (ver Figura 2.4(b)), por lo tanto, el área relativa es la razón entre el área del cromosoma y la suma de todas las áreas de los cromosomas en el conjunto. [3].

El perímetro del cromosoma es el perímetro del menor polígono convexo que lo contiene (ver Figura 2.4(c)).

El centrómero es la porción del cromosoma donde la cromátida comienza a separarse durante la división celular (ver Figura 2.4(d))). Usualmente se identifica a través del perfil de intensidad del medial axis [14]. El índice del centrómero es la razón entre la longitud del brazo p y la longitud del brazo q . Esta es una característica importante para la conformación del cariotipo.

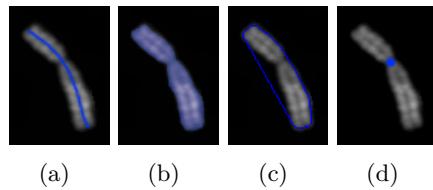


Figura 2.4: En azul se muestran las características: (a) linea cenal, (b) área, (c) perímetro del menor polígono convexo, (d) centrómero.

El perfil de las bandas [19, 11] es la curva del promedio de la intensidad a lo largo de las líneas perpendiculares a los puntos del medial axis por la distancia de dichos puntos a uno de los finales del cromosoma. Esta es la característica más importante para la clasificación de los cromosomas, ya que la estructura de dichas bandas es única para cada cromosoma.

Las distintas transformadas se pueden obtener utilizando el perfil de las bandas de los cromosomas [19], multiplicándola por alguna distribución de densidades predefinida.

2.3. Clasificación

La clasificación es el proceso final en la conformación del cariotipo humano. Es complicada por las múltiples formas en que pueden aparecer los cromosomas. En el proceso de segmentación se pierde información cuando se separan los objetos que se encuentran solapados. No se cuenta con un

método que permita recuperar dicha información en la etapa de extracción de características y esto añade un problema más a la efectividad de los clasificadores. En esta sección se definirá el problema a solucionar y se mostrarán las técnicas usadas actualmente para resolverlo.

2.3.1. Definición del problema

Los cromosomas se dividen en 24 clases que corresponden a los 22 pares de autosomas y a los cromosomas sexuales X y Y . A partir de una lista de vectores de características se identifica la clase a la que pertenecen y se forman los pares homólogos. La clasificación de los cromosomas se puede realizar utilizando distintos métodos, empleando algoritmos de aprendizaje de máquina, que varían en efectividad y eficiencia. Para resolver este problema se utilizan clasificadores de aprendizaje supervisado.

2.3.2. Algoritmos estudiados

El algoritmo más utilizado para la clasificación de cromosomas es *Artificial Neural Networks* (ANN). El número de neuronas de entrada y de salida está determinado por la cantidad de características y de clases respectivamente. La cantidad de capas de neuronas ocultas se puede ajustar, para obtener la mayor eficiencia en la clasificación. ANN es el mejor clasificador de cromosomas conocido.

ANN también se utiliza de forma jerárquica. Primero se dividen los datos en 7 grupos basados en sus características morfológicas. El número de capas ocultas se debe variar de acuerdo a los resultados obtenidos de la clasificación. Luego, se divide cada uno de los 7 grupos en 24 subgrupos a clasificar. En [4] se plantea que utilizando este método se logra disminuir un poco el error.

K-Nearest Neighbour (KNN) es otro de los métodos de clasificación que se utiliza en [8]. Está basado en la cercanía de los datos de entrenamiento en el espacio de características. En la literatura se ha comenzado a emplear *Support Vector Machine* (SVM) [15] para la clasificación, pero a pesar de su eficiencia, se hace muy costoso realizar el entrenamiento.

2.4. Soluciones conjuntas

Los métodos mostrados anteriormente han visto la segmentación y la clasificación como procesos separados, sin embargo, ambos están relaciona-

dos. Cada uno puede mejorarse con la información obtenida del otro. El algoritmo a continuación se apoya en esta hipótesis.

Uno de los algoritmos implementados en la literatura utiliza una función de probabilidad, el tamaño de los cromosomas y la información multi-espectral de los *pixels* para separar y clasificar los cromosomas. Este método está compuesto por tres pasos principales:

- Evaluar la probabilidad de un candidato a cromosoma de pertenecer a un cromosoma de cierta clase.
- Generar el conjunto de cromosomas candidatos.
- Escoger el mejor conjunto de cromosomas candidatos y clases a las que pertenece a partir de una prueba de probabilidad máxima.

Los resultados muestran que la función de probabilidad propuesta puede ser utilizada para comprobar los errores de segmentación, de clasificación, anomalías en los cromosomas, daños por radiación, cáncer y una gran variedad de enfermedades hereditarias .El método tiene un acierto de más de un 90% [26]. En la Figura 2.5 se puede observar un ejemplo de esta técnica.

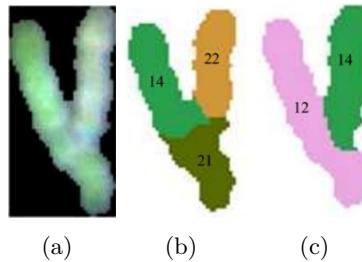


Figura 2.5: (a) MFISH *cluster*, (b)Segmentacion y clasificacion incorrectas, (c) Segmentacion y clasificacion correctas.

Capítulo 3

Algoritmos propuestos

En este capítulo se exponen los algoritmos seleccionados para la confec-
ción del sistema de apoyo en la toma de decisiones en la identificación de
cada cromosoma.

3.1. Pre-procesamiento y Segmentación

El algoritmo propuesto para realizar el pre-procesamiento y la segmen-
tación está dividido en tres fases: la primera, es la limpieza de la imagen,
la detección de los cromosomas y la eliminación del núcleo y otros posibles
ruidos; la segunda es la detección de los *clusters* de los cromosomas
parcialmente unidos o solapados; por último, la tercera fase consiste en la
separación de los cromosomas.

3.1.1. Limpieza de la imagen y detección de cromosomas

Para realizar la limpieza y la detección de cromosomas es necesario con-
vertir la imagen dada a una imagen binaria (ver Figura 3.1). Para ello se
aplica el método de Otsu que calcula el umbral, de forma que la dispersión
dentro de cada segmento sea lo más pequeña posible, pero al mismo tiempo
la dispersión sea lo más alta posible entre segmentos diferentes.

Para la eliminación del ruido es usado el método de apertura con el ker-
nel que se muestra en la Tabla 3.1. Consiste en usar el método de erosión
seguido del método de dilatación (ver Figura 3.2). Además, se toman los
contornos de cada una de las formas en la imagen y se pasa a analizar si
estas son cromosomas o ruido. Las formas cuya área (A_x) esté contenida en
el rango (\minArea, \maxArea), son consideradas cromosomas, de lo contra-

rio es ruido. El mínimo (\minArea) y el máximo (\maxArea) umbral están determinados por las dimensiones de la imagen.

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Tabla 3.1: Kernel utilizado en el método de Apertura

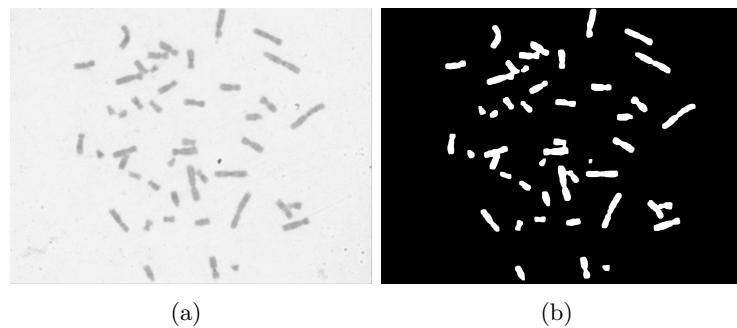


Figura 3.1: (a) Imagen original, (b) Método del Umbral aplicado.

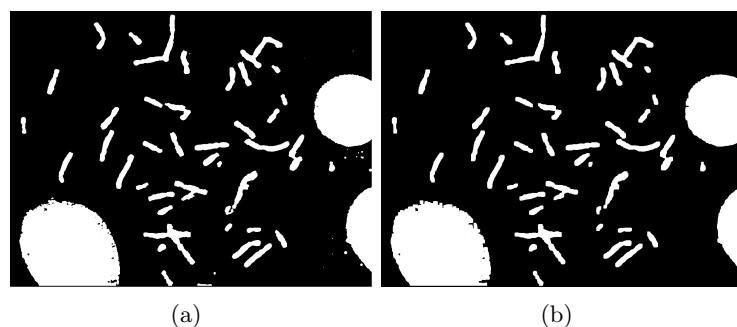


Figura 3.2: (a) Resultado al aplicar el método del Umbral,(b) Resultado al aplicar el método de Apertura.

3.1.2. Detección de los *clusters* de los cromosomas parcialmente unidos o solapados

Para la detección de los *clusters* de cromosomas parcialmente unidos o solapados (ver Figura 3.3), se utilizan tres criterios que se basan en la geometría de los mismos. Estos son:

- Método del menor polígono convexo.
- Método de la elipse que los rodea.
- Método del esqueleto y los puntos finales

Cada forma pasa a través de estos tres criterios y si los satisface todos entonces es considerada como un *cluster* de cromosomas parcialmente unidos o solapados, de lo contrario es considerado un cromosoma simple. Una vez determinados todos los *clusters* compuestos es necesario separarlos para convertirlos en simples.

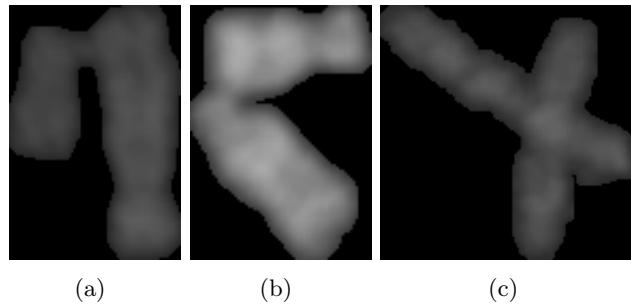


Figura 3.3: (a) y (b) Cromosomas parcialmente unidos, (c) Cromosomas solapados.

Método del menor polígono convexo

Normalmente los cromosomas simples tienen una forma relativamente convexa, por lo que el menor polígono convexo que lo contenga tendrá aproximadamente la misma cantidad de *pixels* que el original. Mientras que en los *clusters* de cromosomas compuestos, el menor polígono convexo que lo rodea tendrá muchos más *pixels* que el original (ver Figura 3.4).

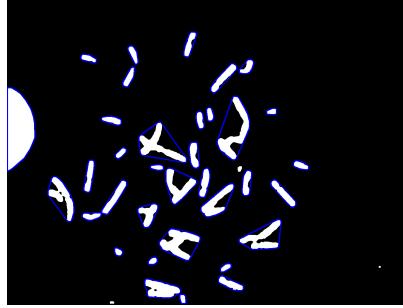


Figura 3.4: Se muestra en azul el menor polígono convexo que contiene a cada objeto.

Los *clusters* compuestos se pueden detectar usando la información anterior. Se halla la razón entre la cantidad de *pixels* del objeto original y el número de *pixels* en el menor polígono convexo que lo contiene. Si este valor es menor que un umbral se pasa la forma a la siguiente fase para seguir comprobando. El umbral puede ser determinado como el promedio de todas las razones de todas las formas. El procedimiento se muestra en el Algoritmo 3.1.

Algoritmo 3.1: Método del menor polígono convexo

```

chromos  $\leftarrow [chr_1, \dots, chr_t]$  lista de cromosomas
ratio_list  $\leftarrow []$ 
for chr in chromos do
    polygone  $\leftarrow$  calcular menor polígono convexo que contiene a chr
    pix_chromo  $\leftarrow$  cantidad de pixels en chr
    pix_polygone  $\leftarrow$  cantidad de pixels en polygone
    ratio  $\leftarrow \frac{pix\_chromo}{pix\_polygone}$ 
    add ratio to ratio_list
end for
threshold  $\leftarrow$  promedio ratio_list
chromos_result  $\leftarrow []$ 
for i in range(0,t) do
    if ratio_list[i]  $\geq$  threshold then
        add chromos[i] to chromos_result
    end if
end for
return chromos_results

```

Este método es efectivo identificando cromosomas simples, pero tiene problema con los curvados ya que los puede identificar como un *cluster* compuesto. Esta deficiencia no causa ningún problema al algoritmo general que determina los *clusters* con cromosomas parcialmente unidos o solapados, ya que para ser clasificado de esta forma debe satisfacer los tres criterios.

Método de la elipse que lo rodea

Los cromosomas normalmente son largos y delgados, excepto por los cromosomas que pertenecen a los pares 20,21 y 22; por lo que la elipse que los rodea debe ser larga. Mientras que los cromosomas solapados tendrán una elipse cercana a un círculo (ver Figura 3.5). Siguiendo esta idea se pueden detectar los *clusters* compuestos.

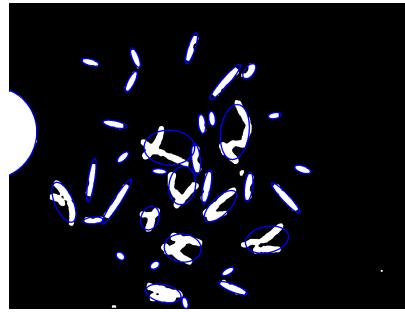


Figura 3.5: Se muestra en azul la elipse que contiene a cada objeto.

El primer paso es determinar las longitudes de la menor y la mayor abscisa de la elipse que rodea a la forma y hallar la razón entre estas. Para los *clusters* compuestos es de esperar que este valor sea cercano a 1, ya que la elipse que lo rodea debe ser cercana a un círculo, pero si es un cromosoma simple debe tener un radio menor, por lo que debe determinarse un umbral para distinguir entre los simples y los parcialmente unidos o solapados. Este umbral puede ser tomado como el promedio de todas las razones. Para cada *cluster* se compara la razón con el valor obtenido, si es menor, se puede tomar como un cromosoma simple, de lo contrario se pasa el *cluster* a la

siguiente fase. Este método se puede ver en Algoritmo 3.2.

Algoritmo 3.2: Método de la Elipse

```

chromos ← [chr1,...,chrt] lista de cromosomas
ratio_list ← []
for chr in chromos do
    ellipse ← calcular elipse que contiene a chr
    min_ratio ← radio menor de ellipse
    max_ratio ← radio mayor de ellipse
    ratio ←  $\frac{\min\_ratio}{\max\_ratio}$ 
    add ratio to ratio_list
end for

threshold ← promedio ratio_list
chromos_result ← []
for i in range(0,t) do
    if ratio_list[i] >= threshold then
        add chromos[i] to chromos_result
    end if
end for

return chromos_results
  
```

Este método es efectivo, pero tiene problemas con dos tipos de cromosomas: los cromosomas pequeños y los cromosomas curvados. El criterio del menor polígono convexo identifica los cromosomas pequeños por lo cual no representan un problema. Para identificar los curvados es utilizado el esqueleto y puntos finales en el próximo paso.

Esqueleto y puntos finales

El último método del algoritmo detecta el esqueleto de cada forma restante obtenida de los pasos anteriores. El esqueleto de una forma es una versión más delgada de esta, que es equidistante de los contornos de la forma original. Normalmente, tienen un *pixel* de amplitud (ver Figura 3.6). Para hallar el esqueleto se va erosionando la imagen hasta que tenga un *pixel* de amplitud. Luego de obtener el esqueleto se buscan los puntos finales, que no son más que los últimos *pixels* que se encuentran en cada lado de las líneas del esqueleto. Si la cantidad de puntos obtenidos es mayor que dos entonces se puede decir que el *cluster* contiene cromosomas parcialmente unidos o solapados, de lo contrario es un cromosoma simple. El método propuesto se puede observar en el Algoritmo 3.3.

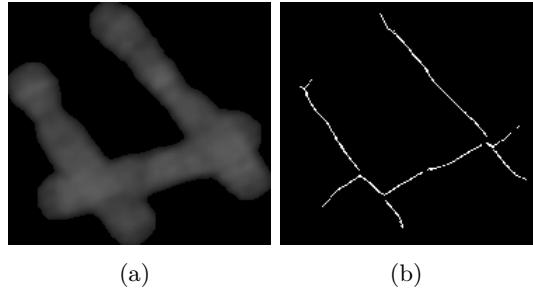


Figura 3.6: (a) Cromosomas solapados, (b) Esqueleto correspondiente a los cromosomas en (a).

Algoritmo 3.3: Algoritmo para detectar cromosomas parcialmente unidos o solapados

```

Chromosomes_Image  $\leftarrow [img_1, \dots, img_t]$  lista de imágenes de cromosomas

clusters  $\leftarrow []$ 
for img in Chromosomes_Image do
    esc  $\leftarrow$  Encontrar el esqueleto de img
    c  $\leftarrow$  Número de puntas del esqueleto
    if c > 2 then
        add img to clusters
    end if
end for

return clusters

```

Este método tiene una alta efectividad, aunque es costoso debido a las varias iteraciones necesarias para hallar el esqueleto.

3.1.3. Separación de los cromosomas

El método descrito a continuación es utilizado para separar los cromosomas parcialmente unidos. Este consiste en encontrar dos puntos de corte por los cuales se pueda trazar una línea recta que divida al *cluster* en dos cromosomas independientes. Si el *cluster* contiene más de dos cromosomas el algoritmo es aplicado múltiples veces.

En primer lugar, se obtienen los puntos del borde del *cluster*; se ordenan estos *pixels* en sentido de las manecillas del reloj. Luego se utilizan dos criterios para determinar cuáles de estos puntos son los puntos de corte. Todos los puntos del contorno son evaluados según los criterios siguientes:

- Variación en el ángulo de la dirección del movimiento (VAMD *Variations in the Angle of Motion Direction*)
- Suma de las distancias entre el total de puntos (SDTP *Sum of Distances among Total Points*)

Variación en el ángulo de la dirección del movimiento (VAMD)

El ángulo de la dirección del movimiento θ_i es el ángulo entre la abscisa horizontal y la dirección en la que se recorren los vértices en el i -ésimo *pixel*. Sea $P = \{P_1, P_2, \dots, P_n\}$ los puntos del borde del cromosoma ordenados en sentido de las manecillas del reloj. Por cada *pixel* P_i se tiene un punto $(X_i; Y_i)$ que representa las coordenadas $(x; y)$ en la imagen. El ángulo θ_i puede ser calculado utilizando el vector formado por P_i y P_{i+1} y el eje horizontal, como se muestra a continuación:

$$\theta_i = \tan^{-1} \frac{Y_{i+1} - Y_i}{X_{i+1} - X_i} \quad (3.1)$$

Para una mejor estimación de este ángulo se pueden utilizar P_i y P_j donde $j = i + k$ con $k > 1$, considerando que los puntos pertenecen a una lista circular.

$$\theta_i^j = \tan^{-1} \frac{Y_j - Y_i}{X_j - X_i} \quad (3.2)$$

Según [16], una buena aproximación para θ_i seria:

$$\theta_i = \frac{1}{2} * (\theta_i^{i+4} + \theta_i^{i+5}) \quad (3.3)$$

Luego de calcular θ_i para cada *pixel* del contorno, es necesario hallar la variación del ángulo en el i -ésimo punto, este ángulo se puede obtener de la siguiente forma:

$$\Delta\theta_i = \theta_{i+1} - \theta_i \quad (3.4)$$

Es de esperar que los puntos de corte tengan un $\Delta\theta_i$ elevado comparado con otros *pixels* en el contorno del *cluster*. Para eliminar algunos puntos innecesarios que no sean posibles puntos de corte, es utilizado el Algoritmo

3.4.

Algoritmo 3.4: Algoritmo para remover puntos que no son de corte

```

points  $\leftarrow [p_1, \dots, p_k]$  lista de puntos del contorno
points_angle  $\leftarrow []$ 
for pt in points do
     $\theta \leftarrow$  calcular  $\theta$  asociado a pt
    add  $\theta$  to points_angle
end for

points_result  $\leftarrow []$ 
 $\Delta\theta \leftarrow$  promedio de points_angle
for pt in points do
    if pt  $> \Delta\theta$  then
        add pt to points_result
    end if
end for

return points_result

```

Los puntos restantes son analizados por el otro criterio con la intención de determinar los dos puntos de corte.

Suma de las distancias entre el total de puntos (SDTP)

Primero, se halla la suma de las distancias entre un punto P_i y todos los demás puntos del borde del *cluster*:

$$Dist(i) = \sum_{j=1, j \neq i}^M d(P_i, P_j) \text{ para } i = \overline{1, M} \quad (3.5)$$

donde M es la cantidad de puntos del criterio anterior y $d(P_i, P_j)$ es la distancia euclíadiana entre P_i y P_j . Finalmente, se define el costo del i -ésimo punto del contorno como:

$$Cost(i) = Dist(i) - \delta * \Delta\theta_i \quad (3.6)$$

El parámetro δ es un número positivo el cual controla el efecto de $\Delta\theta_i$ en la función de costo. Para seleccionar los puntos con menor costo en la función, es necesario minimizar la función $Cost(i)$, por lo que se deben minimizar la $Dist(i)$ y maximizar $\Delta\theta_i$.

Luego de aplicar esta función a cada punto del contorno del *cluster* se almacenan los dos menores costos y estos se seleccionan como puntos de

corte. El *cluster* es separado a través de una línea entre los dos puntos de corte.

3.2. Extracción de Características

Para poder realizar la clasificación de los cromosomas, es necesario extraer de las imágenes un conjunto de características que permitan decidir la clase de un cromosoma, de la forma más efectiva posible.

3.2.1. Método para enderezar los cromosomas

Para aumentar la efectividad y facilitar la extracción de las características se decide, previo a la extracción, implementar una técnica para enderezar los cromosomas [21]. De esta forma se elimina una de las mayores dificultades presentes en esta etapa, los cromosomas curvos. Dicha técnica consiste en localizar el centrómero del cromosoma, en una representación binaria de la imagen original en escala de grises. La imagen de un cromosoma puede ser considerada como una imagen bimodal, en donde un objeto está situado sobre un fondo de un único color, conformando una imagen con diferentes escalas de grises, como se muestra en la Figura 3.7.



Figura 3.7: Cromosoma en escala de grises.

Un histograma de dichas imágenes contiene dos máximos, uno de ellos corresponde a los *pixels* del fondo y el otro corresponde a los *pixels* del objeto. Se puede determinar el valor de un umbral efectivo que separe al objeto del fondo utilizando el método de Otsu. Obtenemos de esta forma la representación binaria de la imagen necesaria para realizar dicha técnica. Un ejemplo de la imagen de un cromosoma y de su representación binaria es mostrado en la Figura 3.8.

Teóricamente, los dos vectores de proyección ortogonal (horizontal y vertical) de la imagen binaria contienen toda la información morfológica del

objeto y pueden utilizarse para la extracción de las características. Para calcular el vector de la proyección horizontal, se suma el valor de los *pixels* de cada fila de la imagen binaria. Teniendo en cuenta que dicha representación binaria, solo posee 0s (*pixels* negros) y 1s (*pixels* blancos), los elementos en el vector de proyección representan la cantidad de *pixels* blancos (1s) en la fila correspondiente (ver Figura 3.9).

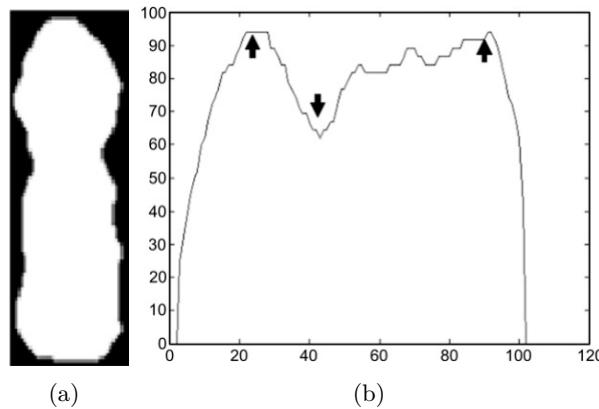


Figura 3.9: (a) Imagen binaria de un cromosoma, (b) Representación gráfica del vector de proyección horizontal.

Como se puede ver, en el caso de los cromosomas derechos, el mínimo global que se encuentra entre los dos puntos de máximo global, en la parte central del vector de proyección, corresponde al centrómero del cromosoma, donde el cromosoma es más estrecho. Localizando los dos máximos y el mínimo global en el vector de proyección horizontal, se puede encontrar de forma automática la posición del centrómero del cromosoma.

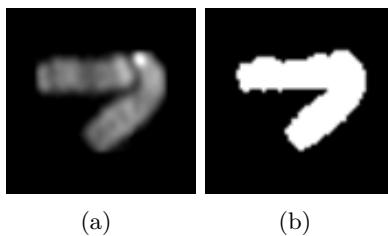


Figura 3.8: (a) Cromosoma en escala de grises, (b) Imagen binaria correspondiente.

Una idea similar se puede utilizar para localizar el centrómero en los cromosomas curvos. Dependiendo del grado de rotación de la imagen, se obtienen vectores de proyección, que contienen puntos y cantidades de máximos y mínimos distintas. Es precisamente en esta idea en la que se basa el método, realizar una serie de rotaciones de la imagen binaria, analizando los vectores de proyección correspondientes a las mismas. Se calculan todos los vectores de la imagen binaria, comenzando desde 0 grados hasta 180 grados, con un paso de 10 grados entre cada una. Entre los vectores de proyección se seleccionan aquellos que posean dos puntos de máximo, con valores próximos y que contengan un punto de mínimo global entre ellos, el cual va a corresponder con el centro del cromosoma curvo. Puede existir más de un vector que cumpla dichas condiciones, por lo tanto, se selecciona de ellos el que posea el menor punto de mínimos, entre los mejores puntos de máximo, con longitudes similares.

Basada en estas observaciones y para identificar de forma automática el grado de rotación de la imagen necesario para obtener el mejor vector de proyecciones, se define el índice de rotaciones (S) de la forma siguiente:

$$S = w_1 * R_1 + w_2 * R_2 \quad (3.7)$$

donde:

$$R_1 = \frac{P_1 - P_2}{P_1 + P_2} \quad (3.8)$$

$$R_2 = \frac{P_3}{P_1 + P_2} \quad (3.9)$$

Donde P_1 es el mayor valor de los dos puntos de máximo, P_2 es el valor del segundo punto de máximo y P_3 es el valor del mínimo global que se encuentra entre P_1 y P_2 . R_1 representa el hecho de que las longitudes de los dos puntos de máximo sean similares, en dicho caso R_1 tendría un valor pequeño. R_2 representa la amplitud del punto de mínimo, con respecto a los dos puntos de máximo. Por lo tanto, la imagen resultante será aquella que posea el menor índice S . El mínimo global en el vector de proyección horizontal de la imagen resultante, es el que corresponde con el centro del cromosoma curvo.

Los coeficientes w_1 y w_2 son los que se utilizan para controlar el peso de cada término en el índice de rotación S , donde $w_1 < 1$, $w_2 < 1$ y $w_1 + w_2 = 1$. En estudios realizados anteriormente se definen los valores aconsejables para cada uno de estos parámetros (ver Tabla 3.2)

	Valor mínimo	Valor máximo	μ	σ
w_1	0.42	0.46	0.434	0.031
w_2	0.54	0.58	0.566	0.028

Tabla 3.2: Valores posibles de w_1 y w_2 para el cálculo del índice de rotación.

Una vez localizado el centro de los cromosomas, se puede enderezar el cromosoma. Primero se separa la imagen en dos sub-imágenes, cortando por la posición en que se encuentra el punto de mínimo. Cada una de estas imágenes contiene un brazo del cromosoma, los cuales son usualmente objetos que ya están derechos. Cada imagen se debe rotar hasta quedar en una posición vertical, buscando siempre maximizar la cantidad de ceros en el vector de proyección vertical (ver Algoritmo 3.5).

Algoritmo 3.5: Algoritmo para enderezar un brazo de un cromosoma

```

i ← 10*i
    rotate_img ← calcular la imagen rotada en angi grados
    proyect_v ← calcular vector de proyección vertical
    count ← cantidad de ceros en proyect_v
    if max_zeros < count then
        angle ← angi
        max_zeros ← count
    end if
end for
return angle
  
```

Una vez que se calculen los ángulos de rotación necesarios para que los dos brazos queden derechos, se repite el mismo procedimiento con los valores ya determinados a la imagen original en escala de grises. Luego de tener las dos sub-imágenes ya derechas, se conectan los dos brazos para formar una imagen con un cromosoma ya enderezado (ver Figura 3.10). Este método se muestra en el Algoritmo 3.6.

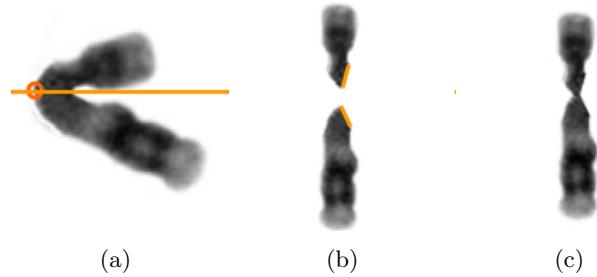


Figura 3.10: (a) Cromosoma con la línea de corte seleccionada anteriormente, (b) Ambos brazos después de enderezados, (c) Resultado después de conectar los dos brazos de los cromosomas.

Algoritmo 3.6: Algoritmo para enderezar cromosomas

```

for i in range(0,18) do
    angi ← 10*i
    rotate_img ← calcular la imagen rotada en angi grados
    proyect_v ← calcular vector de proyección horizontal
    p1,p2 ← calcular los dos valores máximos de proyect_v
    p3 ← calcular mínimo entre p1 y p2
    Si ← calcular utilizando p1, p2 y p3
    if Si < S then
        S ← Si
        index ← posición de p3
        angle ← angi
    end if
end for

img1,img2 ← picar binary_img por index
angle1 ← calcular ángulo para enderezar un brazo de img1
angle2 ← calcular ángulo para enderezar un brazo de img2

img ← rotar img en angle grados
img_b1,img_b2 ← picar img por index
img_b1 ← rotar img_b1 en angle1 grados
img_b2 ← rotar img_b2 en angle2 grados
img_result ← concatenar img_b1 con img_b2

return img_result

```

3.2.2. Características extraídas

Una vez que se tienen solo imágenes parcial o totalmente derechas, se hace un poco más simple la extracción de los rasgos característicos. Las características seleccionadas se computan de la forma siguiente:

- Longitud relativa

Partiendo de que las imágenes de los cromosomas están parcial o totalmente derechas, se puede obtener la longitud de los cromosomas como

la razón de la suma del largo y el ancho de la imagen que lo contiene y la suma de los anchos y largos de todas las imágenes del caso de estudio.

- Área relativa

Para calcular el área de los cromosomas basta con calcular cualquiera de los vectores de proyección y sumar todos los valores. De esta forma se obtienen todos los *pixels* que forman parte del objeto en la imagen. El valor resultante es la razón entre la suma de los valores de la proyección del cromosoma y la suma de las proyecciones de los cromosomas en el caso de estudio.

- Posición del centrómero

Para determinar la posición del centrómero, se utiliza la idea expuesta en la sección anterior, determinando los dos puntos de máximo en el vector de proyección horizontal y luego, se determina el mínimo global que se encuentra entre los dos puntos de máximo. La posición del punto de mínimo se corresponde con la posición del centrómero. Luego se calculan las longitudes de los extremos al centrómero (longitud de los brazos de los cromosomas) y el valor de dicha característica se corresponde con la razón entre la longitud del menor brazo y la longitud del mayor brazo.

- Perímetro convexo

Se calcula el perímetro del menor polígono convexo que contiene al cromosoma. Se utilizan las funciones *convexHull* y *arcLength* para hallar el polígono convexo y su perímetro.

- Información respectiva a las bandas

La primera característica que se extrajo referente a las bandas fue el número de bandas oscuras y el número de bandas claras presentes en la imagen del cromosoma. Se calcula contando los cambios en la monotonía del eje central del objeto.

Otra característica utilizada fue el color de la banda con que inicia y el color de la banda con que finaliza el cromosoma. Cada par de cromosomas empieza y termina con una combinación determinada de bandas claras y oscuras. En caso de que la banda sea clara la característica vale 0 y en el caso contrario vale 1. Además, se calcula un vector que representa el mayor valor de los vectores horizontales de *pixels* que conforman la imagen del cromosoma. Se cuenta la cantidad

de posiciones del vector que forman parte de las bandas claras y la cantidad que forma las bandas oscuras del cromosoma.

Finalmente se conforma un vector que representa rangos de intensidades de igual tamaño de 0 a 255 y se cuenta la cantidad de *pixels* que pertenecen a cada una de estas regiones.

3.3. Clasificación

Una vez realizada la etapa de extracción de características, comienza la búsqueda del mejor clasificador que determine la clase de cada uno de los cromosomas.

En la búsqueda del mejor método se probaron varios algoritmos de aprendizaje de máquina, los cuales se describen a continuación:

- ***K-Nearest Neighbors:***

Clasifica los nuevos ejemplos basándose en las clases de los k vecinos más cercanos a ellos. Se representan los datos en un espacio métrico y se define una distancia, respetando lo más posible la topología del espacio original [6].

- ***Decision Tree:***

Aproximan una función a partir de realizar pruebas a los valores de las características que definen a los datos. ID3 es el algoritmo que se utiliza para construir árboles de decisión, generando de forma iterativa el árbol, seleccionando en cada nodo la característica que mejor particione el conjunto de entrenamiento [22].

- ***Logistic Regression:***

Conocido también como clasificador de máxima entropía. Está basado en un modelo lineal que minimiza el costo de *hit or miss* de la función, en vez de la suma de las raíces de sus residuales, como una regresión ordinaria [17].

- ***Ridge Classifier:***

Se basa en un modelo lineal. Utiliza la regresión de Ridge sobre el problema de mínimos cuadrados, imponiendo una penalización cuadrática al tamaño de los coeficientes [17].

- ***Random Forest:***

Consiste en un conjunto de árboles de decisión. Cada árbol se construye de un ejemplo extraído con reemplazo del conjunto de entrenamiento.

Al dividir un nodo, el corte realizado se escoge de forma óptima en un subconjunto aleatorio de las características. Como resultado generalmente aumenta ligeramente el sesgo del clasificador, disminuyendo en cambio la varianza. A menudo la disminución de la varianza compensa el aumento del sesgo, resultando un clasificador más preciso [2].

■ **LDA:**

Modela la distribución condicional de los datos, $P(X|y = k)$, para cada clase k . La predicción puede obtenerse utilizando la regla de Bayes. La probabilidad $P(X|y)$ es modelada como una distribución Gaussiana, asumiendo la misma matriz de covarianza para cada clase [17].

Capítulo 4

Implementación y Experimentación

En este capítulo se muestran los principales resultados obtenidos en la experimentación. Además, se presenta el sistema propuesto de apoyo en la toma de decisiones en la identificación de cada cromosoma.

4.1. Experimentación y Resultados

Para valorar la efectividad de los algoritmos implementados se utilizaron tres conjuntos de prueba. El primero se utilizó para analizar la correcta eliminación del ruido y la identificación de los cromosomas simples, parcialmente unidos y solapados. Contiene 33 imágenes en metafase celular que pertenecen a casos reales del Centro de Genética Médica de Cuba. Las imágenes estaban en formato *bmp* y con una resolución de 1504x1144 *pixels* (ver Figura 4.1).

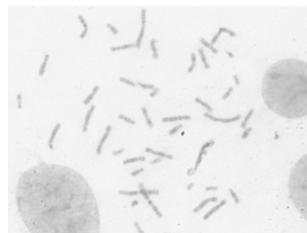


Figura 4.1: Metafase celular de uno de los casos de estudio perteneciente al Centro de Genética Médica de Cuba.

El segundo conjunto de prueba se utilizó para validar la segmentación de los cromosomas parcialmente unidos. Este contiene 58 imágenes en formato bmp y las resoluciones variaban de acuerdo al tamaño del cromosoma.

Para evaluar el desempeño de los algoritmos de clasificación se modificó un conjunto de prueba recomendado en la literatura [7]. Con el objetivo de simplificar el proceso de la extracción de características se aplicó el método para enderezar los cromosomas. La base de datos resultante contiene 5474 imágenes de cromosomas simples, divididas en 119 casos de estudio con 46 capturas cada uno. Los resultados se validaron de forma automática utilizando las bibliotecas *SciKit-Learn* y *SciPy*. En la Tabla 4.1 se muestra la estructura de los datos por clases.

Clases	No. Cromosomas
De la Clase 1 a la Clase 22	238 x 22
Clase 23	193
Clase 24	45
TOTAL	5474

Tabla 4.1: Descripción de la base de datos utilizada para la experimentación.

4.1.1. Pre-procesamiento

El método de apertura se utilizó para eliminar el ruido. Se varió el número de iteraciones para obtener mejores resultados. El rango de la variación fue de 1 a 5, el mejor valor obtenido fue 3. En la Figura 4.2 se puede apreciar los distintos resultados de variar la cantidad de iteraciones en uno de los casos de prueba.

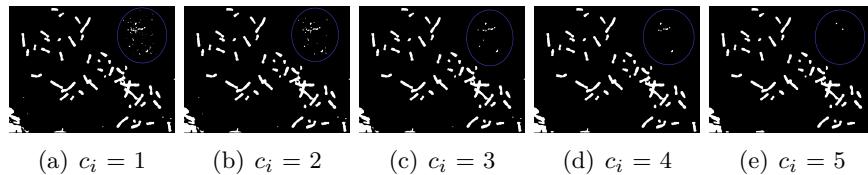


Figura 4.2: Resultado de aplicar el método de apertura con c_i iteraciones.

Para ajustar los parámetros *minArea* y *maxArea* se fueron variando los valores para el área mínima y máxima. Los rangos utilizados fueron $[100, 1000]$ para la menor y $[10000, 100000]$ para la mayor en *pixels*². Estos

rangos para cada parámetro de área se comprobaron en imágenes con resolución de 1504x1144. Los mejores resultados obtenidos fueron 500 pixels^2 y 30000 pixels^2 para la menor y mayor área respectivamente. Debido a que la aplicación puede recibir imágenes de cualquier tamaño, se generalizó el resultado calculándose con la Regla de Tres de la siguiente forma:

$$\minArea = \frac{h * w}{3441}; \quad \maxArea = \frac{h * w}{57} \quad (4.1)$$

donde h es la altura y w es el ancho de la imagen. En la Figura 4.3 se pueden observar los resultados de eliminar el núcleo y otros ruidos en la imagen.

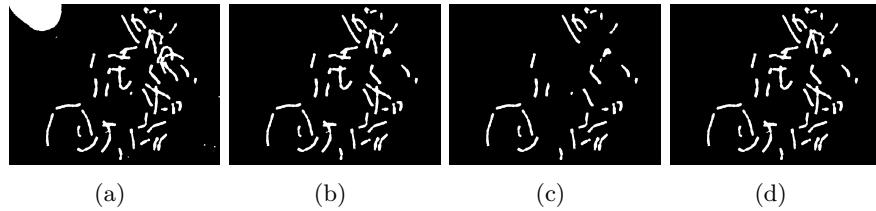


Figura 4.3: Resultados para la variacion de \minArea y \maxArea . (a) Original, (b) $\minArea = 300$ y $\maxArea = 10000$, (c) $\minArea = 100$ y $\maxArea = 5000$, (d) $\minArea = 500$ y $\maxArea = 30000$,

Cada caso de prueba se analizó detalladamente como se puede observar en el Anexo A, en las Tablas A.1, A.2 y A.3, mostrando por cada caso la cantidad de cromosomas presentes en la imagen, los detectados por el algoritmo y la cantidad de aciertos. Los 33 casos de estudios presentan 1192 cromosomas. Después de aplicar el experimento, se detectaron 1204 *clusters*, de los cuales 1091 eran correctos. Obteniendo un 91,52% de efectividad en la detección de los *clusters* con cromosomas simples y parcialmente unidos o solapados. Para más información ver Tabla 4.2.

	No. Cromosomas	<i>Clusters</i> Detectados	<i>Clusters</i> Correctos
Simples	1008	947	917
PU/S	184	257	174

Tabla 4.2: Resultados obtenidos en la detección de los cromosomas simples y parcialmente unidos o solapados (PU/S).

4.1.2. Segmentación

El método implementado en la segmentación separa los cromosomas parcialmente unidos. En el algoritmo SDTP se ajustó el valor del parámetro δ , comprobando la segmentación para cada uno de los siguientes valores: $\{100, 500, 1000, 1500, 2000, 2500, 10000\}$. El mejor valor obtenido fue 1500. En la Figura 4.4 se muestran los resultados al variar el valor del parámetro δ .

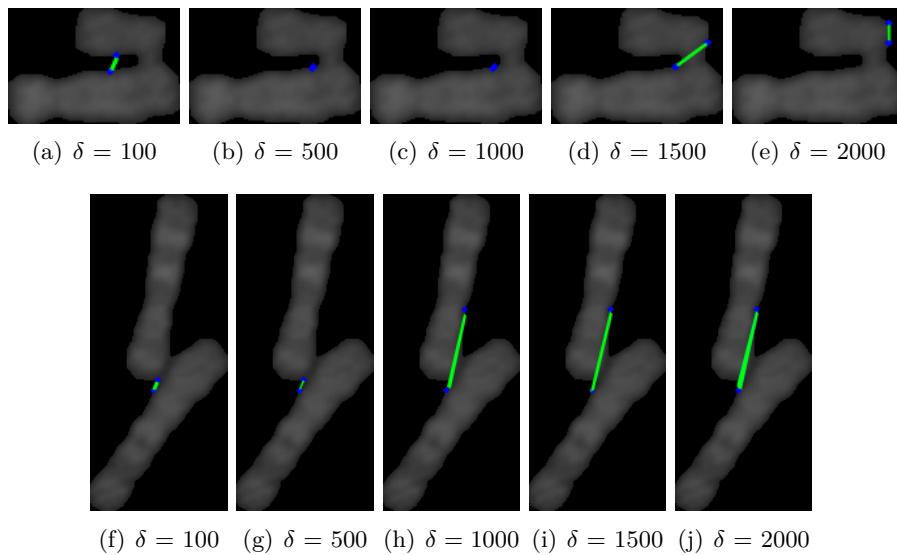


Figura 4.4: Resultados de variar el parámetro δ en dos casos de estudio.

Para mostrar los resultados de la segmentación de los cromosomas parcialmente unidos, se elaboró un conjunto de prueba que contiene 48 imágenes de este tipo. El método separó correctamente un total de 16 cromosomas, dando un 33,33% de acierto. En la Tabla 4.3 se muestran los resultados obtenidos.

	No. Cromosomas	Correctos	Incorrectos
Parcialmente unidos	48	16	32

Tabla 4.3: Resultados obtenidos en la segmentación de cromosomas parcialmente unidos.

4.1.3. Extracción de características y clasificación

La extracción de características se realizó a partir de los cromosomas enderezados. En la Figura 4.5 se pueden ver algunos resultados de las imágenes procesadas por el Algoritmo 3.6.

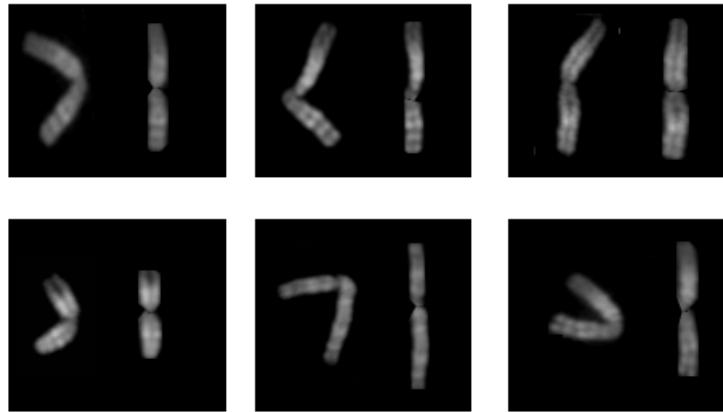


Figura 4.5: Pares de cromosomas, izquierda original, derecha enderezados.

La extracción se realizó utilizando los criterios expuestos anteriormente (ver Sección 3.2.2). En la Figura 4.6 se muestra la estructura final del vector de rasgos obtenido.

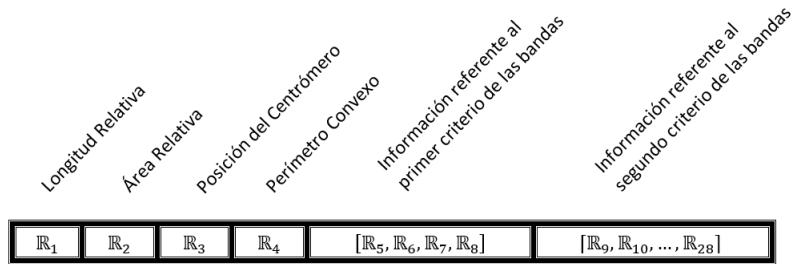


Figura 4.6: Estructura del vector de características obtenido para la clasificación.

Para la clasificación se utilizaron 6 algoritmos de aprendizaje supervisado (ver Sección 3.3). Los parámetros seleccionados se pueden ver en el Anexo A, en la Tabla A.4.

Se diseñaron dos experimentos para validar las características y los clasificadores. En el primero se evalúa la efectividad de las características rea-

lizando 30 corridas de validación cruzada para cada algoritmo y los parámetros seleccionados. En cada corrida se tomó aleatoriamente un 80% de las imágenes como conjunto de entrenamiento y el 20% restante de prueba. El algoritmo con mejores resultados fue *Random Forest*, alcanzando un 54% de precisión. En la Tabla 4.4 se pueden ver los mejores resultados obtenidos para cada algoritmo con los parámetros seleccionados.

Algoritmo	μ	min	max	σ
KNN	0.410	0.373	0.432	0.185
Decision Tree	0.502	0.473	0.533	0.171
Logistic Regression	0.372	0.347	0.399	0.213
Ridge Classifier	0.241	0.208	0.269	0.242
Random Forest	0.540	0.499	0.566	0.168
LDA	0.524	0.489	0.556	0.171

Tabla 4.4: Mejores soluciones para cada algoritmo.

En el segundo experimento se tuvo en cuenta la estructura del conjunto de prueba. Aprovechando la división en casos de los datos, se hizo validación cruzada con *K-fold*, tomando como K los 119 de casos de estudio. Se realizaron K corridas de los algoritmos con cada uno de los parámetros seleccionados. El mayor porcentaje de precisión fue 55%, y lo obtuvo *Random Forest*. En la Tabla 4.5 se muestra para cada algoritmo el mejor resultado obtenido con los parámetros seleccionados.

Algoritmo	μ	min	max	σ
KNN	0.423	0.086	0.782	0.194
Decision Tree	0.510	0.304	0.673	0.171
Logistic Regression	0.379	0.130	0.586	0.213
Ridge Classifier	0.244	0.108	0.369	0.244
Random Forest	0.555	0.326	0.782	0.168
LDA	0.529	0.304	0.739	0.171

Tabla 4.5: Mejores soluciones para cada algoritmo.

Para comparar los resultados de los algoritmos, se realizó para cada una de estas muestras una prueba de normalidad Shapiro-Wilk, en la Tabla 4.6 se muestran los resultados. En todos los casos se valida la normalidad de la distribución.

Algoritmo	W	p_value
KNN	0.9670	0.4614
Decision Tree	0.9836	0.9119
Logistic Regression	0.9800	0.8273
Ridge Classifier	0.9853	0.9431
Random Forest	0.9513	0.1835
LDA	0.9439	0.1160

Tabla 4.6: Resultados de la prueba estándar de normalidad Shapiro-Wilk aplicada al corpus de experimentación.

4.2. Diseño de Aplicación

El sistema que se diseñó para brindar apoyo en la toma de decisiones para la identificación de los cromosomas, consta de tres módulos:

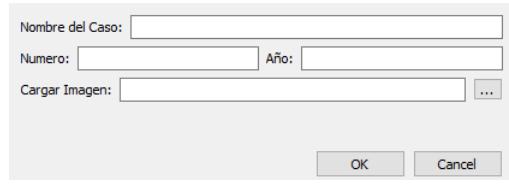
- Segmentación: contiene los métodos de pre-procesamiento y segmentación de las imágenes obtenidas del microscopio.
- Aprendizaje de Máquina: está formado por los algoritmos de extracción de características y de clasificación de los cromosomas individuales.
- Base de Datos: contiene las funciones de acceso a los datos almacenados en el sistema.

Estos módulos fueron implementados en el lenguaje Python 2.7 utilizando el IDE Pycharm 4.5.1.

4.2.1. Módulo de Segmentación

El objetivo del módulo de segmentación es procesar la imagen capturada por el microscopio hasta transformarla en un conjunto de cromosomas simples, que puedan ser utilizadas por el módulo de Aprendizaje de Máquina. Entre las principales funcionalidades se encuentran:

- Añadir una nueva imagen a la base de datos para su estudio.



(a)

Figura 4.7: Vista para añadir una nueva imagen a la base de datos.

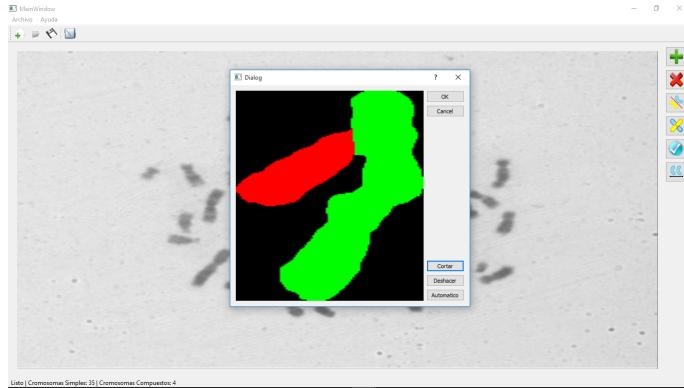
- Realizar de forma automática la separación de los cromosomas del fondo, e identificar los *clusters* de objetos simples y compuestos.



(a)

Figura 4.8: Vista resultante al identificar los *clusters* automáticamente

- Brindar al especialista la opción de separar los cromosomas eligiendo los puntos de corte en la imagen, o de solicitar una propuesta de corte al sistema.



(a)

Figura 4.9: Propuesta de corte brindada por el sistema.

Por la importancia que tiene el trabajo eficiente con imágenes, se utilizaron los métodos implementados en la biblioteca OpenCV.

4.2.2. Módulo de Aprendizaje de Máquinas

El encargado de conformar el cariotipo con las imágenes obtenidas por el Modulo de Segmentación es el Módulo de Aprendizaje de Máquinas. Las funcionalidades brindadas son:

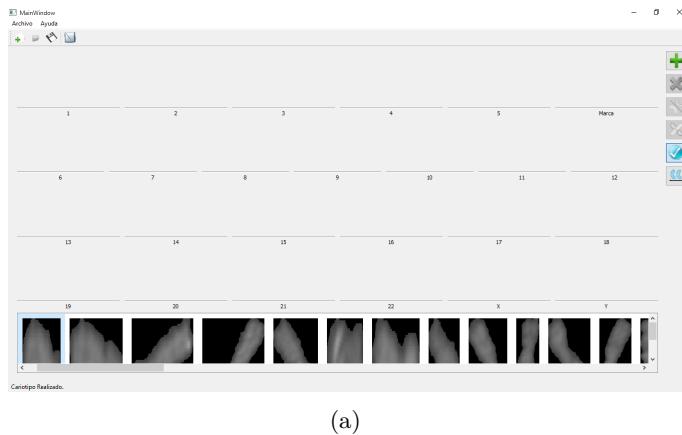
- Realizar de forma automática el cariotipo.



(a)

Figura 4.10: Vista del cariotipo realizado

- Permitir la edición del cariotipo propuesto.



(a)

Figura 4.11: Vista donde el especialista realiza el cariotipado.

En este módulo se utilizan los algoritmos de aprendizaje de máquinas implementados en la biblioteca *SciKit-Learn* [18].

4.2.3. Módulo de la Base de Datos

El objetivo del Módulo de la Base de Datos es almacenar y gestionar los casos de estudio procesados por el sistema en la base de datos SQLite.

Está compuesta por dos tablas. La primera contiene la información necesaria para realizar el entrenamiento del clasificador elegido. En la segunda se almacenan los casos de estudio que el sistema ha analizado, y las anotaciones hechas por los especialistas.

Conclusiones

Como resultado de este trabajo se implementó un sistema de apoyo en la toma de decisiones para la clasificación y análisis de cromosomas. En la etapa de pre-procesamiento se aplicó el método de apertura. Fueron utilizados tres criterios para la detección de los *clusters* compuestos. Para la segmentación de los cromosomas parcialmente unidos se utilizaron los algoritmos *VAMD* y *SDTP*. En la clasificación se emplearon un conjunto de 6 clasificadores, con un grupo específico de parámetros de entrada, resultando *Random Forest* el mejor clasificador.

Recomendaciones

Para futuras investigaciones se recomienda implementar una técnica para separar los cromosomas solapados. Estudiar diferentes conjuntos de prueba, con un mayor número de imágenes, que permitan obtener mejores resultados estadísticos. Se recomienda estudiar más a fondo las características de los cromosomas para lograr mejores descripciones del conjunto de datos. Sería aconsejable probar con otros clasificadores, con un mayor grupo de parámetros que permita extender los resultados de este estudio.

Bibliografía

- [1] Agam and Dinstein. Geometric separation of partially overlapping non-rigid objects applied to automatic chromosome segmentation. Technical report, 1997. (Citado en la página 6).
- [2] Leo Breiman. Random forests. *Machine learning*, 2001. (Citado en la página 29).
- [3] Cho, Ryu, and Woo. A study for the hierarchical artificial neural network model for giemsa stained human chromosome classification. Technical report, 2004. (Citado en la página 9).
- [4] J. Cho. A hierarchical artificial neural network model for giemsa-stained human chromosome classification. Technical report, 2008. (Citado en la página 10).
- [5] Choi, Bovik, and Castleman. Maximum-likelihood decomposition of overlapping and touching m-fish chromosomes using geometry, size and color information. Technical report, 2006.
- [6] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. 2nd edition edition, 2012. (Citado en la página 28).
- [7] E. Grisan, E. Poletti, and A. Ruggeri. Automatic segmentation and disentangling of chromosome in q-band prometaphase images. *IEEE Trans Inf Technol B*, 2009. (Citado en la página 31).
- [8] Janani, Nandakumar, and Nirmala. Feature extraction and paring of g banded chromosome image using k nearest neighbour classifier. Technical report, 2012. (Citado en la página 10).
- [9] Ji. Intelligent splitting in the chromosome domain. Technical report, 1989. (Citado en la página 6).

- [10] Ji. Fully automatic chromosome segmentation. Technical report, 1994. (Citado en la página 6).
- [11] Khmelinskii, Venture, and Sanches. A novel metric for bone marrow cells chromosome pairing. Technical report, 2010. (Citado en la página 9).
- [12] Lerner. Toward a completely automatic neural-network-based human chromosome analysis. Technical report, 1998. (Citado en la página 6).
- [13] Lerner, Guterman, and Dinstein. A classification-driven partially occluded object segmentation (cpoos) method with application to chromosome analysis. Technical report, 1998. (Citado en la página 6).
- [14] Madian and Jayanthi. Analysis of human chromosome classification using centromere position. Technical report, 2014. (Citado en la página 9).
- [15] Markou, Maramis, and Depoulos. Automatic chromosome classification using support vector machine. Technical report. (Citado en la página 10).
- [16] Minaee, Fotouhi, and Khalaj. A geometric approach to fully automatic chromosome segmentation. Technical report, 2014. (Citado en las páginas 6 y 19).
- [17] Kevin Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, Massachusetts, 2012. (Citado en las páginas 28 y 29).
- [18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. (Citado en la página 39).
- [19] Piper and Granum. On fully automatic feature measurement for banded chromosome classification. Technical report, 1989. (Citado en la página 9).
- [20] Rahimi, Amirkhattabi, and Ghaderi. Design of a neural network classifier for separation of images with one chromosome from images with several chromosomes. *Information Technology and Biomedical Applications*, 2008. (Citado en la página 6).

- [21] Roshtkhari and Setarehdan. A novel algorithm for straightening highly curved images of human chromosome. Technical report, 2007. (Citado en la página 21).
- [22] S.j.a. Russell, P.a. Norvig, and R.b. Gutiérrez. *Inteligencia Artificial: Un Enfoque Moderno*. Colección de Inteligencia Artificial. Prentice Hall Hispanoamericana, S.A., 1996. (Citado en la página 28).
- [23] Schwartzkopf, Evans, and Bovik. Minimum entropy segmentation applied to multi-spectral chromosome images. Technical report, 2001. (Citado en la página 7).
- [24] Speicher, Ballard, and Ward. *Karyotyping Human Chromosomes by Combinatorial Multifluor FISH*. Nature Genetics, 1996. (Citado en la página 7).
- [25] Wacharapong, Krisanadej, and Mullica. Segmentation of overlapping chromosome images using computational geometry. Technical report, 2006. (Citado en la página 6).
- [26] Wade, Alan, and Brian. Maximum-likelihood techniques for joint segmentation-classification of multispectral chromosome images. Technical report, 2005. (Citado en la página 11).

Apéndice A

Anexos

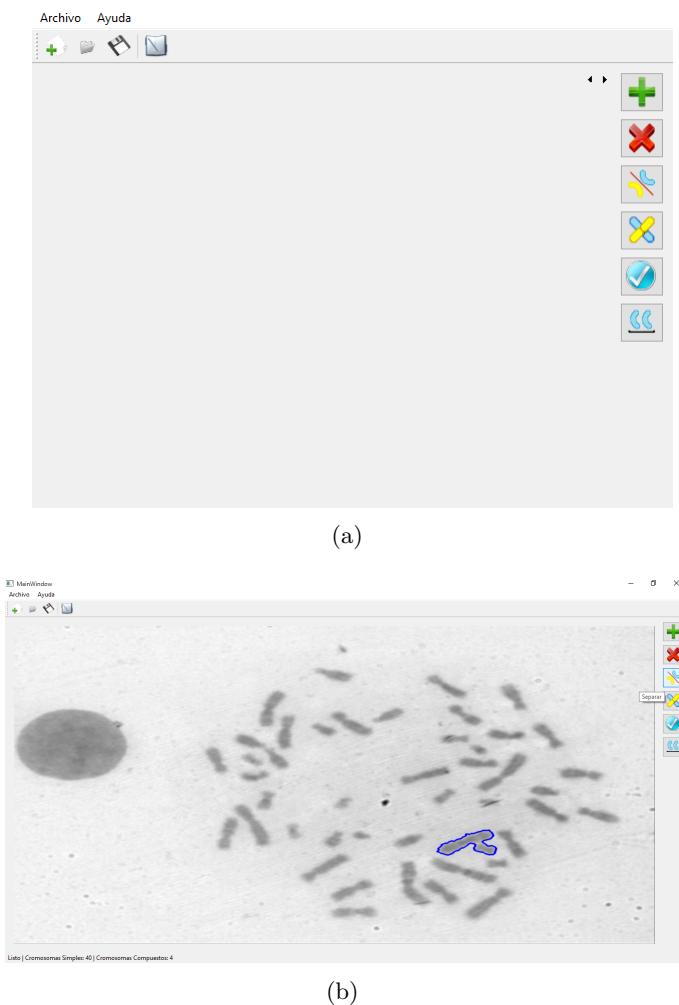


Figura A.1: (a) Vista principal de la aplicación, (b) Vista de un cromosoma solapado seleccionado.

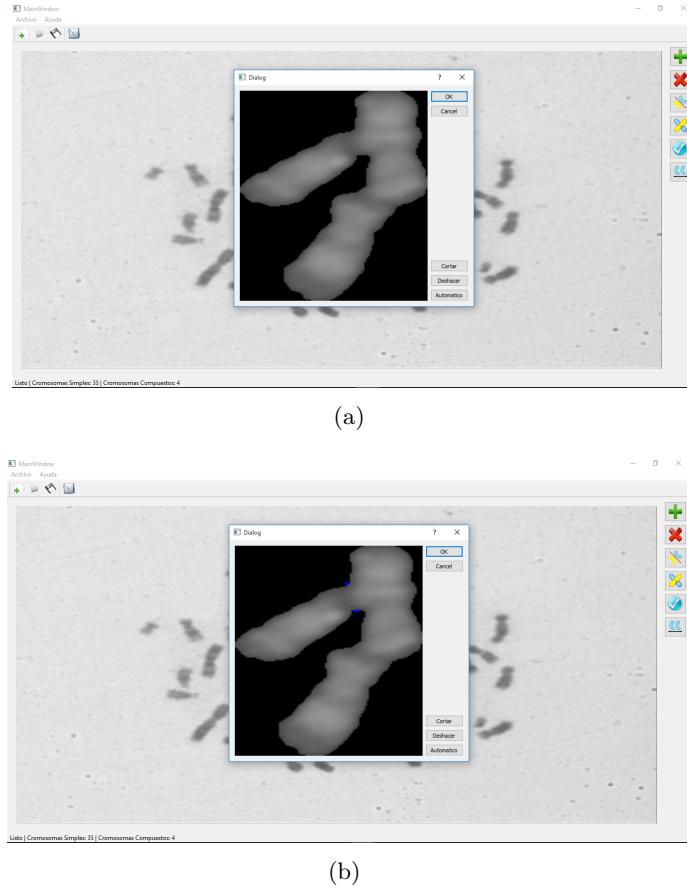
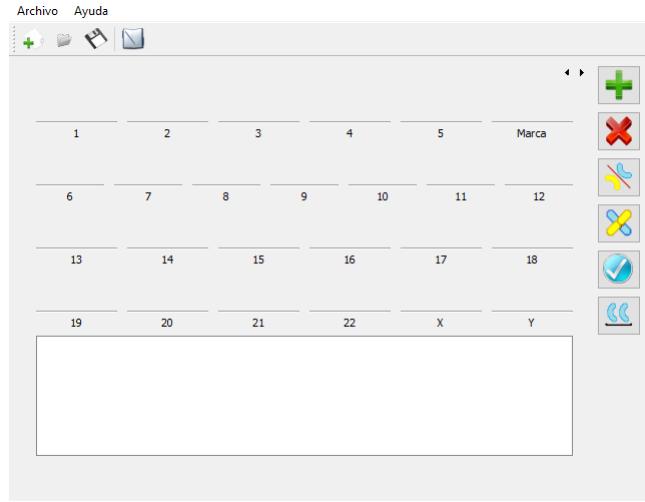
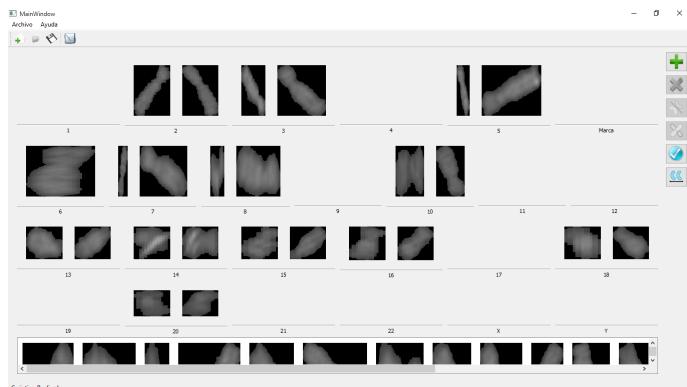


Figura A.2: (a) Vista para realizar el corte de forma manual, (b) Seleccion por el usuario de los puntos de corte.



(a)

Figura A.3: Vista principal del cariotipado.



(a)

Figura A.4: Vista del cariotipado en una etapa intermedia.

	No. Cromosomas	Detectados	Correctos
Caso 1			
Simples	35	32	32
Compuestos	4	7	4
Caso 2			
Simples	38	38	37
Compuestos	2	2	1
Caso 3			
Simples	32	30	29
Compuestos	5	8	5
Caso 4			
Simples	33	30	30
Compuestos	6	9	6
Caso 5			
Simples	34	32	32
Compuestos	6	8	6
Caso 6			
Simples	32	30	28
Compuestos	7	12	7
Caso 7			
Simples	34	30	30
Compuestos	6	10	6
Caso 8			
Simples	40	39	38
Compuestos	3	5	3
Caso 9			
Simples	36	34	34
Compuestos	3	5	3
Caso 10			
Simples	39	23	23
Compuestos	7	13	7
Caso 11			
Simples	44	40	40
Compuestos	1	5	1

Tabla A.1: Resultados de la Segmentación del Caso 1 al Caso 11 de estudio.

	No. Cromosomas	Detectados	Correctos
Caso 12			
Simples	25	25	24
Compuestos	9	9	8
Caso 13			
Simples	24	23	23
Compuestos	8	9	8
Caso 14			
Simples	24	24	23
Compuestos	9	9	8
Caso 15			
Simples	25	26	24
Compuestos	8	9	8
Caso 16			
Simples	28	28	27
Compuestos	6	10	6
Caso 17			
Simples	33	33	31
Compuestos	5	7	5
Caso 18			
Simples	30	28	28
Compuestos	4	6	4
Caso 19			
Simples	33	28	28
Compuestos	5	7	5
Caso 20			
Simples	31	30	28
Compuestos	7	9	7
Caso 21			
Simples	9	8	7
Compuestos	2	4	2
Caso 22			
Simples	17	18	17
Compuestos	8	7	7

Tabla A.2: Resultados de la Segmentación del Caso 12 al Caso 22 de estudio.

	No. Cromosomas	Detectados	Correctos
Caso 23			
Simples	44	39	38
Compuestos	3	6	2
Caso 24			
Simples	30	28	25
Compuestos	8	12	8
Caso 25			
Simples	42	37	37
Compuestos	2	7	2
Caso 26			
Simples	14	15	13
Compuestos	10	9	8
Caso 27			
Simples	26	25	25
Compuestos	7	8	7
Caso 28			
Simples	38	33	33
Compuestos	3	8	3
Caso 29			
Simples	37	35	35
Compuestos	4	6	4
Caso 30			
Simples	29	28	28
Compuestos	4	6	4
Caso 31			
Simples	23	26	23
Compuestos	8	5	5
Caso 32			
Simples	27	27	26
Compuestos	7	8	7
Caso 33			
Simples	32	25	21
Compuestos	5	11	5

Tabla A.3: Resultados de la Segmentación del Caso 23 al Caso 33 de estudio.

Algoritmo	Parámetros
KNN	n-neighbors=21, weights=distance
KNN	n-neighbors=21, weights=uniform
KNN	n-neighbors=17, weights=distance
KNN	n-neighbors=17, weights=uniform
KNN	n-neighbors=11, weights=distance
KNN	n-neighbors=11, weights=uniform
Decision Tree	criterion=entropy, max-depth=10
Decision Tree	criterion=entropy, max-depth=5
Decision Tree	criterion=gini, max-depth=10
Decision Tree	criterion=gini, max-depth=5
Decision Tree	max-features=log2, criterion=entropy, max-depth=10
Decision Tree	max-features=sqrt, criterion=gini, max-depth=5
Logistic Regression	penalty=l1, C=0.1
Logistic Regression	penalty=l1, C=1
Logistic Regression	penalty=l1, C=0.5
Logistic Regression	penalty=l2, C=0.1
Logistic Regression	penalty=l2, C=0.5
Logistic Regression	penalty=l2, C=1
Logistic Regression	penalty=l2, C=10
Logistic Regression	penalty=l1, C=10
Ridge	alpha=0.001
Ridge	alpha=0.1
Ridge	alpha=0.5
Ridge	alpha=10
Ridge	tol=0.01
Ridge	alpha=1
Random Forest	criterion=entropy, max-depth=10
Random Forest	criterion=entropy, max-depth=5
Random Forest	criterion=gini, max-depth=10
Random Forest	criterion=gini, max-depth=5
Random Forest	max-features=log2, criterion=entropy, max-depth=5
Random Forest	max-features=sqrt, criterion=gini, max-depth=10
LDA	n-components=100

Tabla A.4: Parámetros utilizados para cada uno de los algoritmos de clasificación.

Algoritmo	μ	min	max	σ
KNN	0.408	0.379	0.431	0.194
KNN	0.403	0.379	0.428	0.184
KNN	0.410	0.373	0.432	0.185
KNN	0.405	0.365	0.431	0.177
KNN	0.41	0.38	0.433	0.186
KNN	0.405	0.369	0.431	0.184
Decision Tree	0.485	0.459	0.517	0.188
Decision Tree	0.458	0.44	0.478	0.17
Decision Tree	0.502	0.473	0.533	0.171
Decision Tree	0.464	0.44	0.492	0.182
Decision Tree	0.402	0.35	0.46	0.238
Decision Tree	0.287	0.225	0.345	0.357
Decision Tree	0.32	0.29	0.347	0.224
Logistic Regression	0.333	0.31	0.358	0.233
Logistic Regression	0.325	0.303	0.354	0.227
Logistic Regression	0.372	0.347	0.399	0.213
Logistic Regression	0.311	0.276	0.335	0.236
Logistic Regression	0.322	0.3	0.347	0.243
Logistic Regression	0.321	0.294	0.343	0.239
Logistic Regression	0.326	0.303	0.352	0.228
Ridge Classifier	0.241	0.208	0.269	0.242
Ridge Classifier	0.228	0.195	0.252	0.245
Ridge Classifier	0.224	0.191	0.253	0.245
Ridge Classifier	0.218	0.188	0.245	0.249
Ridge Classifier	0.222	0.194	0.25	0.244
Ridge Classifier	0.232	0.195	0.252	0.249
Random Forest	0.531	0.513	0.553	0.176
Random Forest	0.461	0.413	0.506	0.216
Random Forest	0.540	0.499	0.566	0.168
Random Forest	0.452	0.417	0.5	0.22
Random Forest	0.538	0.502	0.568	0.184
Random Forest	0.452	0.394	0.502	0.216
LDA	0.524	0.489	0.556	0.171

Tabla A.5: Resultados de los experimentos realizados a los algoritmos.