UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

METODOLOGIAS PARA DETECÇÃO DO CENTRÔMERO NO PROCESSO DE IDENTIFICAÇÃO DE CROMOSSOMOS

DISSERTAÇÃO DE MESTRADO

Guilherme Chagas Kurtz

Santa Maria, RS, Brasil

2011

METODOLOGIAS PARA DETECÇÃO DO CENTRÔMERO NO PROCESSO DE IDENTIFICAÇÃO DE CROMOSSOMOS

Guilherme Chagas Kurtz

Dissertação apresentada ao Curso de Mestrado em Computação do Programa de Pós-Graduação em Informática (PPGI), Área de Concentração em Computação, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Mestre em Informática**.

Orientador: Prof. Giovani Rubert Librelotto

Santa Maria, RS, Brasil

2011

Universidade Federal de Santa Maria
Centro de Tecnologia
Programa de Pós Graduação em Informática

A Comissão Examinadora, abaixo assinada,
Aprova a Dissertação de Mestrado

METODOLOGIAS PARA DETECÇÃO DO CENTRÔMERO NO PROCESSO DE IDENTIFICAÇÃO DE CROMOSSOMOS

elaborada por
Guilherme Chagas Kurtz

com requisito parcial para obtenção do grau de
Mestre em Informática

COMISSÃO EXAMINADORA

Giovani Rubert Librelotto
(Orientador)

Ana Isabela Lopes Sales, Dra. (USP)

Juliana Kaizer Vizzotto, Dra. (UFSM)

Santa Maria, 28 de Outubro de 2011.

RESUMO

Dissertação de Mestrado
Programa de Pós-Graduação em Informática
Universidade Federal de Santa Maria

METODOLOGIAS PARA DETECÇÃO DO CENTRÔMERO NO PROCESSO DE IDENTIFICAÇÃO DE CROMOSSOMOS

Autor: Guilherme Chagas Kurtz
Orientador: Giovani Rubert Librelotto
Data e Local da Defesa: Santa Maria, 28 de outubro de 2011.

Muitas doenças genéticas ou anomalias que podem ocorrer nos cromossomos humanos podem ser descobertas através da análise da forma e das características morfológicas dos cromossomos. A elaboração do cariótipo (organização dos 24 cromossomos de uma célula humana de acordo com o seu tamanho através de um desenho ou de uma fotografia obtida de um microscópio) é geralmente utilizada para alcançar este objetivo. Os primeiros passos para as análises cromossômicas são a definição e extração das características morfológicas e do padrão de bandas dos cromossomos (variações dos níveis de cinza ao longo de seu comprimento). Dentre as características morfológicas, além do seu tamanho, destaca-se a localização do centrômero (local que divide o cromossomo em braço longo e braço curto) e a classificação de acordo com o mesmo. Os avanços ocorridos nas técnicas de cultura celular, bandeamento, coleta e análise dos materiais para a execução do cariótipo possibilitaram grandes progressos no diagnóstico das alterações cromossômicas. Porém, este processo ainda é bastante utilizado de forma manual, pois, apesar da demanda crescente deste tipo de exame, ainda é pequena a oferta de sistemas automáticos que auxiliem o trabalho dos geneticistas na geração do cariótipo. Logo, a automatização deste processo e a possibilidade de se obter resultado em curto espaço de tempo, agilizando condutas terapêuticas ou tranqüilizando familiares tem um valor inestimável. A detecção do centrômero é de grande importância tanto no aspecto do processo manual como no processo automático, pois agilizaria o diagnóstico. No processo manual, a possibilidade de se realizar um agrupamento dos cromossomos em relação ao tamanho e a posição do centrômero auxiliaria o trabalho de um geneticista na parte de identificação e também na segmentação, pois ao se definir a classificação de um cromossomo em relação a sua posição do centrômero, é possível definir a sua polaridade (colocar o cromossomo "em pé"). No processo automático, é um excelente filtro na busca por uma maior taxa de acertos nos sistemas de identificação dos cromossomos, pois cada tipo de cromossomo pertencerá sempre a uma determinada classificação de acordo com o centrômero (metacêntrico, submetacêntrico ou acrocêntrico). Nesta dissertação, portanto, buscou-se desenvolver uma série de métodos para detecção do centrômero, destacando-se a definição de dois algoritmos que utilizam os métodos desenvolvidos no decorrer deste trabalho. Como resultado obtido destaca-se que ao aplicar esta abordagem na base de imagens utilizada do *BioImLab* (Laboratório de Imagem Biomédica, Universidade de Padova, Itália), alcança-se cerca de 94.37% de acertos, uma taxa maior que qualquer trabalho relacionado na literatura.

Palavras Chave: processamento de imagens; cromossomos; identificação de cromossomos; detecção do centrômero; reconhecimento de padrões.

ABSTRACT

Master's Dissertation
Post-Graduate Program in Informatics
Federal University of Santa Maria

METODOLOGIES FOR CENTROMERE DETECTION IN THE PROCESS OF CHROMOSOMES IDENTIFICATION

Author: Guilherme Chagas Kurtz
Advisor: Giovani Rubert Librelotto
Santa Maria, 28 de outubro de 2011.

Many genetic diseases or abnormalities that may occur in human chromosomes can be detected by analyzing the shape and morphology of chromosomes. The elaboration of the karyotype (organization of the 24 chromosomes of a human cell according to its size through a drawing or a photograph obtained from a microscope) is usually used to achieve this goal. The first steps for chromosomal analysis is the definition and extraction of morphology and banding pattern (gray level variations along its length) features of chromosomes. Among the morphological characteristics, in addition to its size, there is the centromere location (a region that divides the chromosome in long arm and short arm) and the classification according to the same. The advances made in cell culture techniques, banding, collecting and analyzing of materials for the implementation of the karyotype allowed great progress in the diagnosis of chromosomal abnormalities. However, this process is still used manually, because despite the growing demand of this type of examination, it is still small the supply of automated systems that help the geneticists work in the katyotype generation. So, the automation of this process and the possibility of obtaining results in a short time speeding therapeutic conduct and reassuring that families are invaluable. Centromere detection is of great importance both in the manual process as the automatic process, for faster diagnosis. In the manual process, the possibility of performing a grouping of the chromosomes in relation to the size and centromere position would help the geneticist work at the identification and also in segmentation, because by defining the chromosome classification in relation to its centromere position, is possible to define their polarity (putting the chromosome "standing"). In the automatic process, it's an excellent filter in the search for a higher correctness rate for chromosomes identification systems, because each type of chromosome always belongs to a particular classification according to the centromere (metacentric, submetacentric or acrocentric). In this dissertation, therefore, sought to develop a series of methods for centromere detection, especially the definition of two algorithms that use the methods developed in this work. As a result it is emphasized that in applying this approach on the image base used from BioImLab (Biomedical Imaging Laboratory, University of Padova, Italy), it achieves about 94.37% of correctness, a higher rate than any work related literature.

Keywords: image processing; chromosomes; chromosomes identification; centromere detection; pattern recognition.

SUMÁRIO

1. INTRODUÇÃO

Cromossomos são estruturas filamentares presentes no interior do núcleo celular que consistem, cada qual, de uma molécula de DNA supercondensada, juntamente com proteínas histonas e não-histonas (VERMA e BABU, 1995). Até 1956 não se sabia exatamente o número de cromossomos da espécie humana. O desenvolvimento de técnicas adequadas, juntamente com o auxílio do microscópio, possibilitou a identificação de cada um deles, bem como a visualização adequada do seu conjunto. É atribuída aos suecos Tjio e Levan (1956) e aos ingleses Ford e Hamerton (1956) a descoberta de que o genoma humano é constituído por 46 cromossomos (ou 23 pares), sendo 44 autossômicos e 2 sexuais, tornando possível a elaboração do cariótipo (classificação dos cromossomos humanos de acordo com o seu tamanho). A partir de então, a Citogenética tem-se desenvolvido enormemente, trazendo grande contribuição não só para o estudo das doenças humanas, como também para estudos das populações normais, sua origem e evolução.

Os cromossomos não se apresentam uniformes ao longo de todo o seu comprimento; cada cromossomo apresenta uma constrição primária denominada centrômero. O centrômero divide o cromossomo em dois braços: o braço curto, designado por *p* (do francês *petit*) e o braço longo por *q* (por ser a letra seguinte do alfabeto). Morfologicamente, os cromossomos são classificados de acordo com a posição do centrômero. Se este estiver localizado centralmente, o cromossomo é denominado metacêntrico; se próximo à extremidade, é acrocêntrico; e se o estiver em uma posição intermediária, o cromossomo é submetacêntrico (VERMA e BABU, 1995).

Entretanto, os cromossomos não diferem somente pela posição dos centrômeros ou pelo seu tamanho, mas também por apresentarem um padrão característico de bandas. Por meio de técnicas de coloração especial, que coram seletivamente o DNA, cada par cromossômico é individualmente identificado; isto ocorre por apenas um breve período, durante a mitose, na metáfase, quando estão condensados ao máximo e quando os genes não podem ser transcritos (BORGES-OSÓRIO e ROBINSON, 2002).

Nas últimas duas décadas, a área da Genética, em especial a Genética Médica, cresceu muito no Brasil e tem atraído um grande número de profissionais. A Citogenética Humana foi uma de suas primeiras subáreas a serem implantadas no país, inicialmente em laboratórios de

pesquisa e, mais recentemente, também em laboratórios de análises clínicas. Atualmente, suas aplicações incluem a caracterização de polimorfismos nas populações e a pesquisa da ação de agentes mutagênicos (que causam mutações ou mudanças na forma do DNA) ou carcinogênicos (que causam mutações e levam a ativação de genes tumorais) em ensaios *in vitro*, além da análise de cariótipo em muitas doenças. Suas técnicas compreendem importantes ferramentas de diagnóstico pré e pós-natal de anomalias congênitas (alterações nas quais os indivíduos já nascem com ela) e de diagnóstico e monitoramento de terapia em casos de neoplasias, principalmente hematológicas (BRUNONI, 1997).

Nos últimos anos, cresceu também o interesse em dados moleculares de doenças genéticas por parte de médicos e profissionais da saúde, acompanhando o progresso da Biologia Molecular. Os testes moleculares que se baseiam na análise dos ácidos nucléicos são sem dúvida métodos fundamentais no diagnóstico das doenças humanas que resultam de lesões do DNA. Esses testes podem ser realizados com amostras oriundas dos mais diversos tipos de tecidos, incluindo aquelas obtidas por amniocentese, biópsia de vilosidade coriônica e outras (FARIA et al., 2004).

O alcance e a precisão dos diagnósticos genéticos trazem responsabilidades clínicas, éticas e legais sem precedentes. Indivíduos identificados como portadores de mutações podem se tornar ansiosos, depressivos e sentirem-se socialmente estigmatizados. Diagnósticos pré-sintomáticos de doenças de manifestação tardia podem ser emocionalmente devastadores. Do mesmo modo que o erro técnico, a compreensão incorreta do significado de um exame é igualmente prejudicial. Assim, todos os profissionais desta área devem ter consciência da responsabilidade de que erros laboratoriais podem causar danos irreparáveis na vida de um paciente (ACOSTA e FERRAZ, 2000).

Os avanços ocorridos nas técnicas de cultura celular, bandeamento, coleta e análise dos materiais para a execução do cariótipo possibilitaram grandes progressos no diagnóstico das alterações cromossômicas, mas apesar da demanda crescente deste tipo de exame, é pequena a oferta de sistemas automáticos que auxiliem o trabalho dos geneticistas na geração do cariótipo (BRUNONI, 1997).

A possibilidade de se obter resultado em curto espaço de tempo, agilizando condutas terapêuticas ou tranqüilizando a família, tem um valor inestimável. Assim, o cariótipo no diagnóstico clínico é uma verdadeira corrida contra o tempo (BRUNONI, 2002).

Portanto, o desenvolvimento de um sistema computadorizado utilizando técnicas de visão por computador servirá de auxílio ao geneticista tanto na otimização do trabalho quanto na execução da análise da forma dos cromossomos humano através da posição do centrômero. Com uma boa técnica de detecção do centrômero, é possível aumentar a homogeneidade dos resultados da análise dos cromossomos permitindo sua análise detalhada, onde a impressão das imagens digitalizadas dos cariótipos poderá ser enviada juntamente com os resultados. Além disso, do ponto de vista dos sistemas automáticos de identificação, o processo de detecção do centrômero pode servir como um grande filtro de agrupamento, tornando possível obter altas taxa de acertos na identificação dos cromossomos e, logo, auxiliar de forma mais confiável o trabalho dos geneticistas.

A meta dos laboratórios que prestam serviços nessa área deve ser o desenvolvimento de uma política de qualidade que faça com que o número de erros tenda a zero e que os laudos emitidos sejam compreensíveis pelos profissionais que atuam junto ao paciente.

Desta forma, o objetivo deste trabalho é o desenvolvimento de uma técnica para detecção do centrômero de cromossomos humanos em imagens digitais de forma a auxiliar o processo de identificação dos mesmos. Para alcançar este objetivo, busca-se definir uma boa sequência de pré-processamento nas imagens obtidas, bem como desenvolver diversas técnicas de detecção do centrômero, e por fim elaborar algoritmos que utilizem o melhor de cada técnica, buscando aumentar e superar as taxas de acertos encontradas na literatura, demonstrando que é possível o desenvolvimento de uma técnica de detecção do centrômero que obtenha altas taxas de acertos.

Assim, esta dissertação está dividida da seguinte forma: o capítulo 2 fará uma revisão bibliográfica sobre as principais características de um cariótipo e seus cromossomos, mostrando as principais características utilizadas no processo de identificação dos mesmos, bem como uma breve revisão bibliográfica a respeito de processamento de imagens. O capítulo 3 irá apresentar os principais trabalhos relacionados encontrados na literatura voltados à detecção do centrômero e o capítulo 4 apresentará a metodologia desenvolvida. Por fim, o capítulo 5 ira mostrar os resultados obtidos nas técnicas desenvolvidas e a seção 6 fará as considerações finais a respeito deste trabalho.

2. REVISÃO BIBLIOGRÁFICA

Este capítulo tem como objetivo realizar uma revisão bibliográfica dos assuntos relevantes que serão tratados nesta dissertação. Desta forma, na seção 2.1 apresentam-se detalhes sobre o cariótipo, os cromossomos, suas formas de classificação e suas principais características tal como o padrão de bandas e o centrômero. Na seção 2.2 é dada visão geral sobre o processo de identificação de cromossomos do ponto te vista manual e automático, destacando onde o trabalho desenvolvido nesta dissertação está encaixado. Por fim, na seção 2.3 será feita uma breve revisão bibliográfica a respeito de processamento de imagens de forma a abordar o que foi utilizado no decorrer deste trabalho.

2.1. Cariótipo, cromossomos e suas características

Durante o ciclo celular, os cromossomos de uma célula passam por estados de menor a maior compactação. O grau mais alto de compactação é alcançado na fase de divisão celular chamada metáfase. É durante a metáfase que se torna possível uma visualização de maior qualidade da quantidade e da morfologia dos cromossomos, sendo possível visualizá-los de forma individual, e desta forma, fotografá-los. Com isto, pode-se realizar o processo de isolamento e ordenação dos cromossomos, denominado cariótipo (ROBERTIS e HIB, 2006).

O cariótipo da espécie humana possui 23 pares de cromossomos, sendo que 22 deles estão presentes tanto nas células masculinas quanto em células femininas, recebendo o nome de autossomos. Além dos cromossomos autossomos, ainda consta um par de cromossomos sexuais, sendo que no homem é constituído por dois cromossomos diferentes: o cromossomo X e o cromossomo Y, e na mulher, o par é formado por dois cromossomos X.

2.1.1. Cromossomos

Os cromossomos podem ser definidos como uma longa cadeia de DNA a qual contém vários genes, e está cadeia está associada a diversas proteínas. Os cromossomos metafásicos

apresentam uma morfologia característica, eles são constituídos por dois filamentos denominados cromátides, as quais são unidas pelo centrômero (região mais condensada do cromossomo, ou seja, a região em que está concentrado o maior número de genes). Nos extremos dos braços estão os telômeros (região constituída por repetitivas seqüências de DNA nos extremos dos cromossomos que protegem os mesmos contra degradação, recombinação e translocação) (ROBERTIS e HIB, 2006). O centrômero divide as cromátides do cromossomo em dois braços onde, em geral, um é mais longo que o outro. O braço curto geralmente é identificado pela letra *p* e o braço longo pela letra *q*. A divisão dos cromossomos pelo centrômero define uma característica importante que é muito utilizada na classificação de cada par, sendo ela denominada de Classificação de Denver.

2.1.2. Classificação de Denver

Os cromossomos humanos, quando classificados com base no tamanho e na posição do centrômero, adotam um esquema estipulado em 1960 em um congresso de citogeneticistas na cidade de Denver, a qual é denominada Classificação de Denver. Nesta classificação, os cromossomos são divididos em 7 grupos identificados pelas letras de A a G em ordem decrescente de tamanho tal como é mostrado na Figura 2.1 (NUSSBAUM et al., 2008).

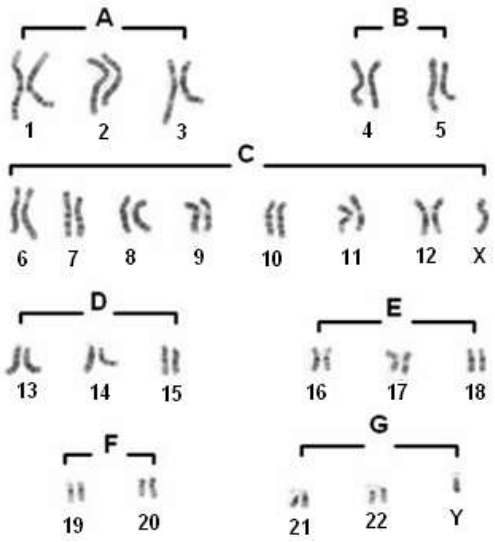


Figura 2.1 – Classificação dos cromossomos de acordo com a Classificação de Denver.

2.1.3. Centrômero

O centrômero, conforme dito anteriormente é a região que divide o cromossomo em braço curto e braço longo, sendo ela a região mais condensada concentrando o maior número de genes. Segundo NUSSBAUM et al. (2008), os centrômeros dos cromossomos são regiões com bandas mais escuras, mas estas bandas acabam sendo imperceptíveis neste tipo de imagens. Porém, também de acordo com NUSSBAUM et al. (2008), as bandas ao redor dos centrômeros tendem a ser mais claras que as demais bandas do cromossomo; logo, os centrômeros além de serem identificados por estarem em uma região mais estreita dos mesmos, geralmente também apresentam bandas cujos píxeis possuem níveis de cinza próximos a 255 (cor branca).

A Figura 2.2 traz exemplos de imagens de cromossomos da base de dados utilizada neste trabalho (ver seção 4.2.1), em que está marcada a região do centrômero. Facilmente nota-se que a afirmação de NUSSBAUM et al. (2008) é verdadeira, com exceção de alguns cromossomos tal como o cromossomo 3 da Figura 2.2.c.

(a) (b) (c) (d) (e) (f)

Figura 2.2 – Exemplos de cromossomos com seus centrômeros marcados.

De acordo com a posição do centrômero ao longo do cromossomo, os cromossomos podem ser classificados em três grupos:

- Metacêntrico: possuem o centrômero em uma posição central, de forma que a diferença de tamanho entre o braço curto e o braço longo seja mínima;
- Submetacêntrico: O centrômero localiza-se longe da região central do cromossomo, de forma que um dos braços seja maior que o outro;

- Acrocêntrico: O centrômero está localizado perto de um dos extremos do cromossomo. Sendo assim, os braços curtos dos cromossomos são muito pequenos (NUSSBAUM et al., 2008).

Os cromossomos sempre irão pertencer a um dos grupos citados acima, sendo que esta característica não muda, ou seja, um cromossomo 1 sempre será metacêntrico.

A Figura 2.3 mostra uma imagem do cariótipo humano em que os cromossomos estão agrupados de acordo com a posição do centrômero. Pode-se ver que os cromossomos acrocêntricos, com exceção do Y, possuem uma pequena massa ligada ao braço por filamentos muito finos. Estas massas, denominadas satélites, são bastante irregulares, o que torna inapropriada a utilização dos mesmos na classificação dos centrômeros individualmente (NUSSBAUM et al., 2008).

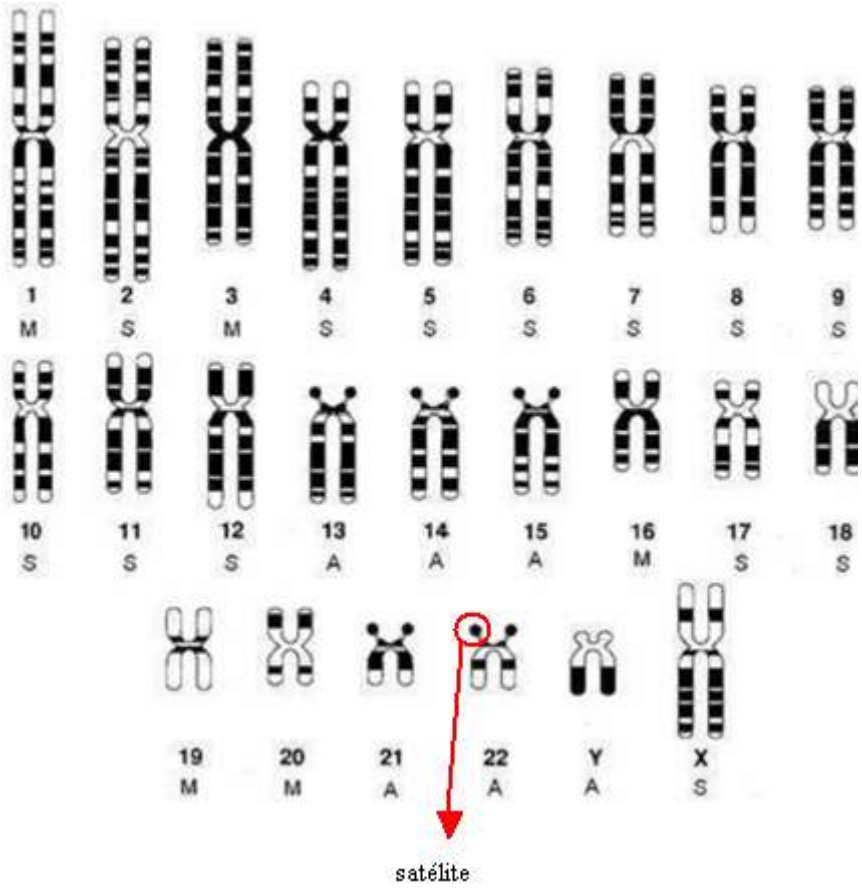


Figura 2.3 - Classificação de acordo com a posição do centrômero, sendo M metacêntrico, S submetacêntrico e A acrocêntrico.

2.1.4. Padrão de bandas

No processo de cariotipagem (classificação dos cromossomos de uma fotografia retirada de uma célula), além da classificação de acordo com o tamanho e a posição do centrômero, os cromossomos podem ser classificados de acordo com seu padrão de bandas. O padrão de bandas de um cromossomo são faixas claras e escuras exibidas ao longo de seu eixo longitudinal tal como é apresentado na Figura 2.4 na forma de ideogramas. A distribuição dessas faixas é diferente e constante em cada um dos cromossomos, o que facilita bastante a sua identificação. Além disso, essas faixas podem constituir um guia muito importante no diagnóstico de transtornos genéticos tais como deleções, duplicações, inversões e translocações cromossômicas (ROBERTIS e HIB, 2006).

Na cariotipagem, o centrômero é geralmente utilizado como um filtro na classificação dos cromossomos, para que então se faça o uso do padrão de bandas para definir quem é quem. Ao tornar este processo automatizado, a detecção do centrômero torna-se uma característica importantíssima para se obter um alto índice de acertos na classificação dos cromossomos de acordo com o padrão de bandas. Portanto, a próxima seção irá apresentar uma visão geral dos processos manuais e automáticos de identificação de cromossomos, sendo que o próximo capítulo apresentará os trabalhos relacionados a este encontrados na literatura que visam o mesmo objetivo do apresentado nesta dissertação, os quais buscam desenvolver técnicas computadorizadas para detecção do centrômero em imagens digitais.

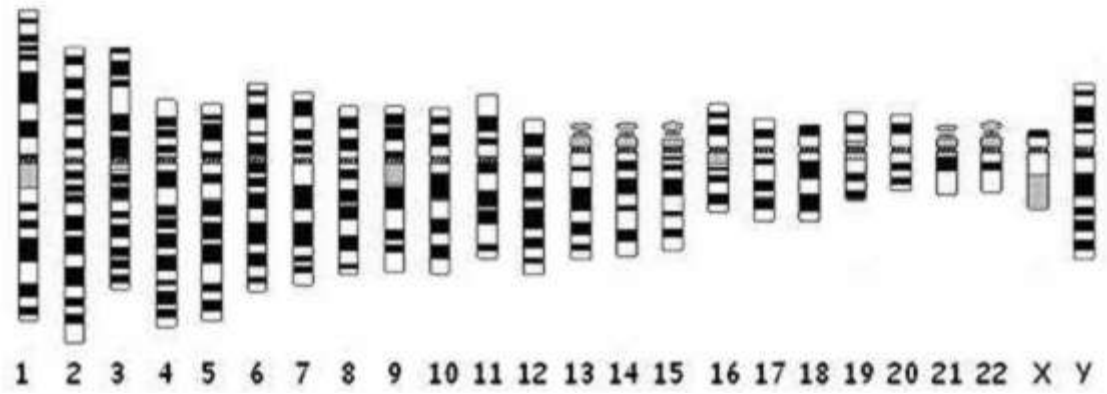


Figura 2.4 – Padrão de bandas do cariótipo humano.

2.2. Visão geral do processo de identificação de cromossomos

O processo de análise e identificação de cromossomos é chamado cariotipagem. A Figura 2.5 traz um exemplo de uma metáfase de uma célula humana (conjunto de cromossomos em uma célula), e a Figura 2.6 um exemplo de metáfase já identificada, em que o processo de identificação dos cromossomos já foi feito (cariotipagem). Este processo é geralmente utilizado como recurso para investigação de alterações cromossômicas que podem ser responsáveis por diversos problemas de saúde. Dessas alterações, podemos citar neoplasias (cânceres), síndromes genéticas, quebras cromossômicas (causadas por radiação, por exemplo). Alem disso, é possível dar um prognóstico ao paciente, ou seja, definir a evolução de uma doença em certo organismo (neste caso, em humanos) e a seguir descrever-lhe a conduta terapêutica necessária (NUSSBAUM et al., 2008).

Figura 2.5 – Exemplo de metáfase de uma célula humana (CAPUTO, 2005).

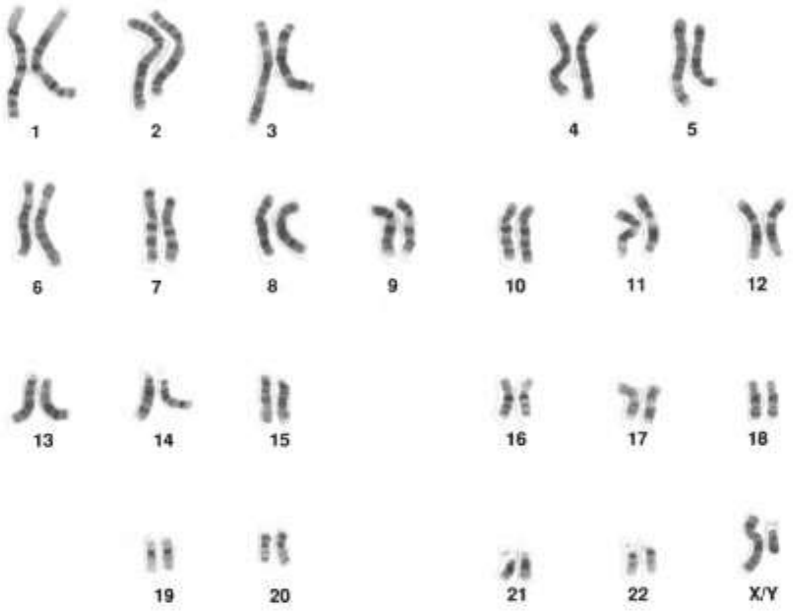


Figura 2.6 – Exemplo de cariótipo humano.

Geralmente, este processo de análise e identificação dos cromossomos é feito de forma manual por um geneticista, sendo que este processo pode durar em torno de meia hora para cada imagem. O processo de identificação manual envolve as seguintes etapas:

1. Aquisição da imagem – geralmente através de uma câmera acoplada a um microscópio;
2. Segmentação e identificação dos cromossomos – processo feito de forma manual, em que se desenham as metáfases (no mínimo 50 metáfases por paciente), identificando e analisando cada cromossomo, e em seguida as metáfases são impressas e recortadas manualmente (segmentação), como mostram as Figura 2.7 e 2.8.
3. Identificação de anomalias – A partir dos cromossomos já identificados, busca-se por anomalias que podem ocorrer nos mesmos. Tais anomalias podem estar relacionadas tanto ao número de cromossomos quanto a alterações estruturais.

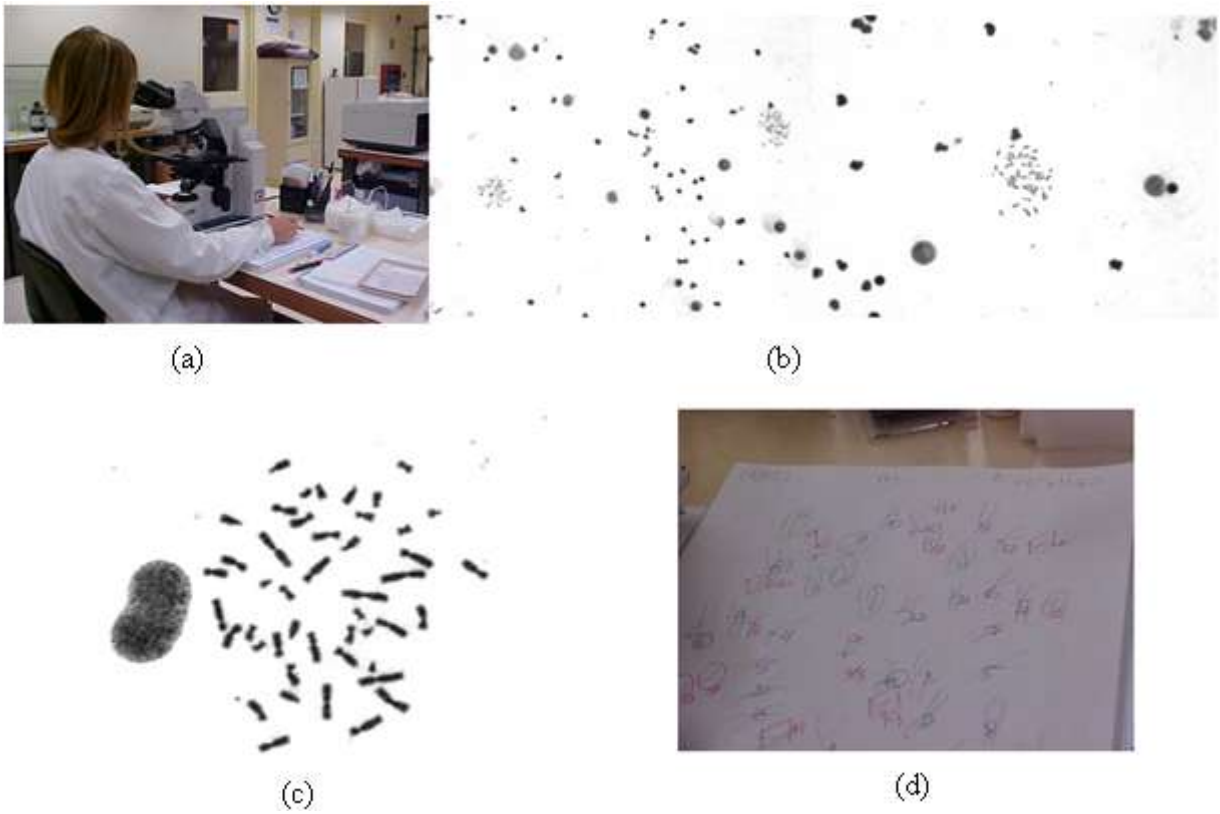


Figura 2.7 – Processo manual de identificação, em que se observa cerca de 50 metáfases através do microscópio e se desenha as mesmas manualmente.

(a) (b)

Figura 2.8 – Processo de segmentação das imagens feita de forma manual. É necessário repetir este passo cerca de 50 vezes para cada paciente.

As Figuras 2.7 e 2.8 demonstram o funcionamento do processo manual na prática. Na Figura 2.7, pode-se observar o processo de aquisição da imagem (a) em que se observa uma metáfase (b) normal e logo após em (c) com um aumento de cerca de 1000 vezes alcançado através do uso de um óleo de imersão, e em seguida cada uma dessas metáfases é desenhada manualmente em (d) e identificada pelo geneticista. Na Figura 2.8 observa-se o processo de segmentação manual (a), em que o geneticista recorta manualmente cada cromossomo obtido nas amostras.

Em relação ao processo de identificação, algumas características são observadas nos cromossomos de forma a facilitar o processo de identificação. As principais características observadas são:

1. Tamanho do cromossomo
2. Padrão de bandas
3. Posição do centrômero

Os cromossomos possuem tamanhos distintos, sendo que a classificação de 1 a 22 (exceto os sexuais X e Y) foi definida de acordo com uma ordem decrescente de tamanho; logo, esta é uma característica muito importante na classificação.

A preparação dos cromossomos é feita através de um método chamado de bandeamento G, onde se utiliza o corante *Giemsa*, o qual é praticado em muitos laboratórios citogenéticos (PAUT, 1993). Este método é um dos mais importantes na detecção de anormalidades cromossômicas, pois a partir dele é possível visualizar os cromossomos através de uma distribuição de bandas claras e escuras: o padrão de bandas.

O padrão de bandas é uma característica distinta em cada par cromossômico, a qual permite uma perfeita classificação dos mesmos (PEREIRA, 1988). O padrão de bandas, no caso de imagens obtidas através do processo de banda G, é a variação de níveis de cinza no decorrer do eixo principal do cromossomo (mais detalhes na seção 2.1.4), sendo que cada tipo de cromossomo irá sempre seguir o mesmo padrão de banda. Entretanto, devido à qualidade da imagem e das diferenças que ocorrem de uma célula para outra, a visualização deste padrão de bandas pode não ser suficiente para a classificação, sendo necessária a utilização de outras características que auxiliem este processo.

Uma das principais características é o centrômero, o qual é o foco deste trabalho. O centrômero é uma das características mais importantes e determinantes na classificação dos cromossomos (BIYANI et al., 2005; CONROY et al., 2000; GREGOR e GRANUM, 1991; LUNDSTEEN et al., 1985; MORADI et al. 2003; PIPER e GRANUM, 1989; SCHWARTZKOPF, et al., 2005; STANLEY et al., 1996; WANG et al., 2008, 2009). Conforme pode-se ver na Figura 2.9, o centrômero divide o cromossomo em dois braços, sendo que o mesmo se destaca como sendo um estrangulamento no interior do cromossomo. Cada tipo de cromossomo terá uma classificação em relação posição do centrômero, e o mesmo sempre será desta classificação (mais detalhes foram apresentados na seção 2.1.3). Ao se obter um alto grau de acerto na classificação do mesmo, é possível filtrar os cromossomos nos três grupos: metacêntrico, submetacêntrico e acrocêntrico. Portanto, ao saber que um cromossomo 1 será sempre metacêntrico (o centrômero divide o cromossomo em dois braços de tamanhos praticamente iguais), e que este geralmente é o maior cromossomo de uma metáfase, facilita a distinção deste cromossomo em relação aos demais.

Figura 2.9 – Exemplo demonstrando a posição do centrômero em relação às 3 classificações possíveis.

A etapa de segmentação e identificação do processo envolve uma tarefa importante também que é a definição da polaridade do cromossomo, ou seja, ao ser recortado manualmente, o mesmo deve ser colocado em “pé” (o braço curto aparece em cima e o longo embaixo). Isso deve ser feito devido ao fato de que o padrão de bandas seguirá um padrão ao longo do eixo principal dos cromossomos em uma ordem específica, logo, é exigência da Comissão do Colégio Americano de Patologia Clínica que isso seja feito dessa forma. Desta forma, o centrômero aparece como a principal característica na definição da polaridade, pois, com a detecção do mesmo, é possível realocá-lo na posição correta.

Devido ao tempo gasto para se realizar todas estas etapas manualmente, e também devido à quantidade de metáfases que chegam aos laboratórios para serem cariotipadas, surge a necessidade de se automatizar este processo. Apesar da demanda crescente, ainda é pequena a oferta de sistemas automáticos que auxiliem o trabalho dos geneticistas na coleta de dados e geração do Cariótipo. O processo automático de identificação de cromossomos envolve as seguintes etapas:

1. Etapa de aquisição – da mesma forma que o processo manual, a imagem da metáfase é obtida através de uma câmera acoplada em um microscópio óptico com um aumento de 1000x;
2. Etapa de pré-processamento – Tem como foco principal a preparação da imagem para a etapa de segmentação, tornando possíveis as operações subseqüentes, a fim de alcançar um resultado final esperado (ou pelo menos próximo dele). Esta etapa envolve a utilização de diversos filtros tal como suavização, detecção de bordas, realce de contraste, entre outros;
3. Etapa de segmentação – etapa em que se extrai e gera uma nova imagem para cada cromossomo da metáfase. Entre a etapa de segmentação e identificação ainda pode ocorrer uma nova etapa de pré-processamento com a aplicação de novos filtros adequados e que auxiliem o processo de identificação e classificação;
4. Etapa de identificação e classificação – Com as imagens segmentadas e preparadas após a etapa de segmentação, é iniciado o processo de identificação utilizando alguma técnica/algoritmo desenvolvido.

Da mesma forma que o processo manual, a etapa de identificação dos cromossomos envolve a extração de informações da imagem de forma a auxiliar e tornar possível a realização desta tarefa. Estas informações geralmente estão também relacionadas ao tamanho do cromossomo, ao padrão de bandas e a posição do centrômero. No trabalho de Moradi e Setarehdan (2006) foi proposto o uso de outras características, mas sem muito sucesso.

Pode-se dizer que o padrão de bandas é a característica que define quem é quem entre os tipos de cromossomo. Porém, apesar de que o trabalho manual feito por especialistas possa se tornar algo rotineiro e fácil, a necessidade de automatizar este processo irá surgir quando estes fatores não estiverem mais superando o tempo gasto para se realizar a cariotipagem de uma grande quantidade de imagens. Daí surge a necessidade de se automatizar, e logo, esta não é uma tarefa fácil, pois, diferente dos exemplos das Figuras 2.2 e 2.5, freqüentemente os

cromossomos aparecem distorcidos, dobrados (dismórficos), sobrepostos e com grandes perdas de qualidade nas imagens, tal como a Figura 2.10, tornando bem mais difícil o processo de classificação utilizando somente o padrão de bandas, talvez não tanto para humanos se for considerada pequenas quantidades de imagens, mas bastante para um sistema automático.

A partir daí sente-se a necessidade de se utilizar outras características dos cromossomos de forma a melhorar estes resultados: o centrômero e o tamanho. Como se pode ver na seção 3.1, os trabalhos voltados à identificação de cromossomos fazem o uso principalmente do padrão de bandas na etapa de identificação, porém, a maior parte destes trabalhos cita a importância de se utilizar um bom algoritmo para detecção do centrômero de forma a classificá-los de acordo com a posição do mesmo, pois acreditam que não é suficiente utilizar somente o padrão de bandas devido ao fato de que, conforme dito anteriormente é freqüente o caso de imagens parecidas com as da Figura 2.10.

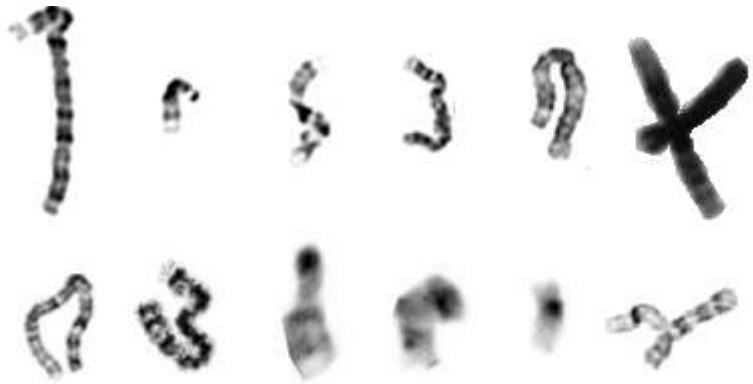


Figura 2.10 – Exemplo de cromossomos distorcidos, dismórficos, sobrepostos e com perda de qualidade nas informações.

Considerando um cromossomo *k* qualquer, o qual não se sabe de que tipo é, mas se sabe que o mesmo é metacêntrico. Ao saber isso, entre os 23 tipos de cromossomos, já descartamos 18 tipos, pois ao saber que este cromossomo é metacêntrico, é fato que ele só poderá ser o cromossomo 1, 3, 16, 19 ou 20. Assim, entende-se que com uma alta taxa de acertos em um algoritmo de detecção e classificação em relação ao centrômero é possível aumentar ainda mais a taxa de acertos na etapa de identificação dos cromossomos, tornando o sistema mais confiável, sendo, portanto este o objetivo deste trabalho.

Desta forma, observa-se a importância de se automatizar o processo de identificação de cromossomos, pois assim é possível obter uma maior quantidade de resultados, resultados mais rápidos e principalmente mais precisos. A detecção do centrômero vem como uma forma de agilizar este processo, tanto do ponto de vista manual quanto do ponto de vista automático. Manual, pois a possibilidade de se realizar um agrupamento dos cromossomos em relação ao tamanho e a posição do centrômero facilitaria bastante o trabalho do geneticista na parte de identificação e também na segmentação, pois ao saber a classificação de determinado cromossomo em relação ao centrômero, é possível saber sua polaridade (se o cromossomo está “de pé”). Do ponto de vista automático, conforme dito anteriormente é um excelente filtro na busca por uma maior taxa de acertos nos sistemas de identificação dos cromossomos.

2.3. Processamento de imagens

O processamento de imagens digitais surge decorrente do interesse de sua aplicação em duas áreas principais de aplicação: melhoria da informação visual para a interpretação humana e o processamento de dados de imagens para percepção automática através de máquinas. Portanto o objetivo do uso do processamento digital de imagens consiste em melhorar o aspecto visual de certas imagens para um analista humano e fornecer subsídios para sua interpretação, extração e processamento de dados de uma imagem. Desta forma, é possível utilizar e interpretar estes dados em sistemas computacionais, além de gerar produtos que possam ser posteriormente submetidos a outros processamentos (GONZALEZ e WOODS, 2000).

A área de processamento digital de imagens tem atraído cada vez mais o interesse nos últimos tempos. A própria evolução da tecnologia e também o desenvolvimento de novos algoritmos para lidar com sinais bidimensionais tem permitido que esta área tenha uma aplicação cada vez maior.

2.3.1. Imagem digital

A expressão imagem monocromática, doravante chamada de imagem, faz referência a função bidimensional de intensidade de luz *f(x,y)*, onde *x* e *y* são as coordenadas espaciais e o valor *f* em qualquer ponto *(x,y)* é proporcional ao brilho (ou nível de cinza) da imagem naquele ponto (GONZALEZ e WOODS; 2000). Ou seja, uma imagem digital contém um número fixo de linhas e colunas de píxeis formando uma matriz. Para imagens monocromáticas, os píxeis geralmente assumem valores inteiros de 8 a 16 *bits* representando o brilho de cada ponto da imagem.

Dependendo de cada aplicação, o valor 0 pode representar os píxeis de cor preta e 255 os píxeis de cor branca para imagens monocromáticas de 8 *bits*. Quanto mais píxeis uma imagem tiver (por exemplo, uma imagem de 1200 x 1000 píxeis), melhor será sua qualidade e resolução. A Figura 2.11 traz um exemplo de uma imagem de 16 linhas e 16 colunas, logo, 256 píxeis. A seção seguinte irá abordar brevemente este tema.

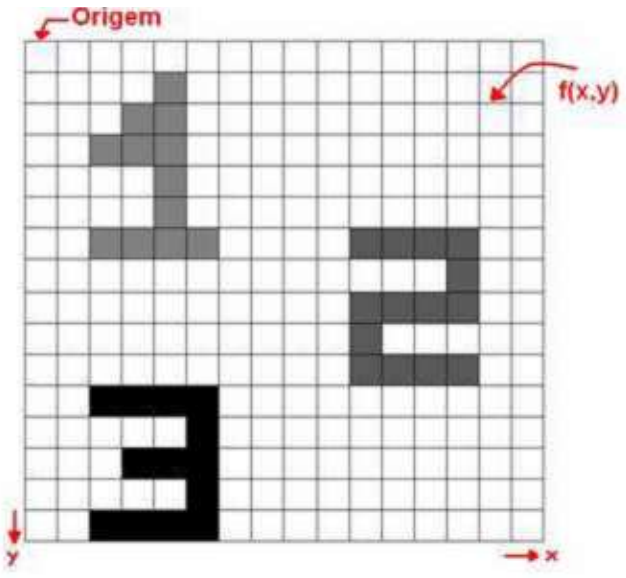


Figura 2.11 – Exemplo de imagem digital com a convenção dos eixos para sua representação.

2.3.2. Resolução espacial e profundidade

A resolução espacial de uma imagem depende do seu destino final, ou seja, de sua aplicação, pois depende da quantidade de detalhes necessários a serem utilizados em determinada aplicação. Por exemplo, considerando uma imagem de 400 linhas e 600 colunas, teremos uma imagem de 240.000. Desta forma, concluímos que a resolução espacial desta imagem é 400 x 600 píxeis. A Figura 2.12 traz um exemplo de variação na resolução espacial de uma imagem.

A profundidade de uma imagem está relacionada a quantidades de níveis de cinza suportadas pela mesma, sendo também conhecido por escala de cinza. Ou seja, uma imagem que suporta 256 níveis de cinza (de 0 a 255) é uma imagem que pode ser descrita com apenas 8 *bits* por pixel. (ou 1 byte por pixel). No caso de uma imagem binária de 1 *bit*, ela poderá assumir somente dois valores: 0 para preto e 1 para branco. A Figura 2.13 traz um exemplo de variação na profundidade de uma imagem.

(a) (b) (c)

Figura 2.12 – Exemplo de variação na resolução espacial de uma imagem. Em (a) temos uma imagem de 240 x 256 píxeis, em (b) 64 x 64 píxeis e em (c) uma imagem de 16 x 16 píxeis.

(a) (b) (c)

Figura 2.13 – Exemplo de variação na profundidade de uma imagem. Em (a) uma imagem de 8 *bits*, em (b) uma imagem de 4 *bits* e em (c) uma imagem de 2 *bits*.

2.3.3. Realce de imagens

O propósito das técnicas de realce é o de desfazer ou reduzir os efeitos de degradação causados na imagem. A crescente necessidade de desenvolver sistemas automatizados para a interpretação de imagens exige que a qualidade da imagem seja livre de ruídos e outras anormalidades. Dessa forma, é importante que se apliquem técnicas de realce de imagens durante o pré-processamento para que a imagem resultante deste processo esteja mais adequada para uma etapa de interpretação (ACHARYA, 2005).

As abordagens sobre realce de imagens se dividem em duas categorias: métodos no domínio espacial e métodos no domínio de freqüência. O domínio espacial refere-se ao plano da imagem e a manipulação direta dos píxeis dessa imagem. Em técnicas no domínio de freqüência a manipulação é feita geralmente através da transformada de Fourier da imagem. É comum também a existência de técnicas de realce baseadas em várias combinações destes dois métodos (GONZALEZ, 2000).

Neste trabalho foram utilizados alguns filtros baseados nestas técnicas, todos eles através da ferramenta *ImageJ* que será apresentada na seção 4.2.1. Portanto, primeiramente será feita uma abordagem referente à matemática utilizada em determinados filtros de processamento de imagens, sendo que tais filtros serão brevemente apresentados nas seções seguintes.

2.3.3.1. Transformada de Fourier

Sendo *f(x)* uma função continua de uma variável real *x*, a Transformada de Fourier de *f(x)* a qual é denotada por F(x), é definida por:

$$\hat{f}(\varepsilon) = \int_{-\infty}^{\infty} f(x) e^{-2\pi x \varepsilon} dx$$

Equação 1

Além disso, pode se obter novamente *f(x)* através da Transformada Inversa de Fourier:

$$f(x) = \int_{-\infty}^{\infty} f(\varepsilon) e^{2\pi x \varepsilon} d\varepsilon$$

Equação 2

Para o uso em computadores é preciso que os valores de *x* sejam discretos. Logo, pode-se usar uma versão discreta da Transformada de Fourier:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}}, k = 0, \dots, N-1$$

Equação 3

Além disso, no caso de imagens bidimensionais, utiliza-se a versão discreta 2D:

$$X_{k,t} = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x_{m,n} e^{-2\pi i \left(\frac{km}{M} + \frac{tn}{N}\right)}, k = 0, \dots, N-1; t = 0, \dots, M-1$$

Equação 4

Da mesma forma que a anterior, é possível obter a Transformada Inversa de Fourier:

$$x_k = \frac{1}{N} \sum_{n=0}^{N-1} X_n e^{\frac{2\pi i}{N}}, n = 0, \dots, N-1$$

Equação 5

O método mais utilizado e preferido para cálculos computacionais é o algoritmo *FFT* (*Fast Fourier Transform*). Este é um algoritmo eficiente para o cálculo da Transformada Discreta de *Fourier*, o qual a complexidade (ou número de operações) é *O(n log n)* contra $O(n^2)$ necessários para o calculo da Transformada Discreta utilizando sua própria definição (BRACEWELL, 1999).

2.3.3.2. Convolução

A convolução é um operador matemático que a partir de duas funções é gerada uma terceira. A convolução de duas funções *f(x)* e *g(x)* é denotada por:

$$f(x) * g(x) = \int_{-\infty}^{\infty} f(\alpha) g(x - \alpha) d\alpha$$

Equação 6

Além disso, sendo *F(u)* a Transformada de Fourier de *f(x)* e *G(u)* a Transformada de Fourier de *g(x)*, a convolução de *f(x)* e *g(x)* pode ser definida como a Transformada Inversa do produto de *F(u)* e *G(u)* (BRACEWELL, 1999):

$$f(x) * g(x) = \hat{f}^{-1}\{F(u) \times G(u)\}$$

Equação 7

No caso de alguns filtros que serão descritos a seguir, *g(x)* será o núcleo da convolução, ou também conhecido como *máscara*, geralmente seguindo a forma de uma matriz, que, no caso deste trabalho, foram utilizadas matrizes 3x3 e 5x5 dependendo do filtro.

2.3.3.3. Suavização

Os filtros de suavização têm como objetivo a remoção (ou redução) de ruídos da imagem através de um borramento da mesma, e assim retirar detalhes desnecessários que não terão significado ou até mesmo atrapalham no resultado final do processamento da mesma. (GONZALEZ e WOODS, 2000).

A suavização geralmente é implementada com a utilização de máscaras de diversos tamanhos. Estas máscaras indicam o novo valor que cada pixel da imagem passará a assumir. Por exemplo, ao utilizar uma máscara 3x3, o novo valor de determinado pixel passará a ser a média dos níveis de cinza de seus vizinhos. Um exemplo de aplicação de um filtro de suavização é mostrado na Figura 2.14.

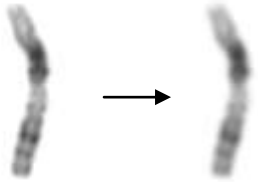


Figura 2.14 – Exemplo de aplicação de um filtro de suavização com uma máscara de 3x3.

2.3.3.4. Mediana

O filtro mediana tem os mesmo objetivos da suavização, porém, ao invés de utilizar a média dos níveis de cinza de seus píxeis vizinhos, o valor de determinado pixel é modificado para a mediana de seus píxeis vizinhos. A Figura 2.15 traz um exemplo de aplicação do filtro mediana (GONZALEZ e WOODS, 2000).

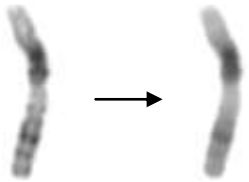


Figura 2.15 – Exemplo de aplicação do filtro mediana com uma máscara 3x3.

2.3.3.5. Dilatação

A finalidade da dilatação é a eliminação de lacunas em imagens binárias (imagens com somente os níveis de cinza 0 - preto - e 1 - branco). A dilatação consiste na união de

todos os pontos *X* (os pontos brancos da Figura 2.16) de uma imagem binária. O elemento estruturante *Bx* intercepta *X*. Os píxeis de cor amarela representam os píxeis que passaram a assumir a cor branca após a dilatação. A cor amarela foi utilizada somente como uma forma de melhor apresentação do resultado.

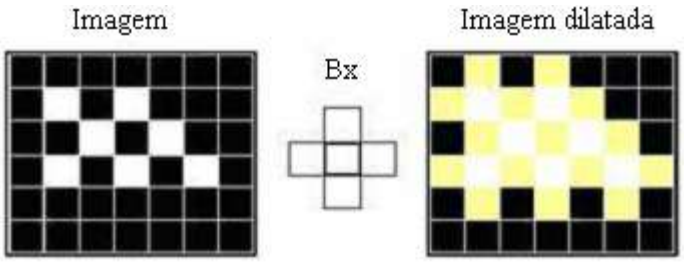


Figura 2.16 – Dilatação de uma imagem (LEITE, 2004).

2.3.3.6. Erosão

A erosão de uma imagem serve para a eliminação de detalhes que são irrelevantes em uma imagem binária. Nesta transformação, a imagem erodida será um conjunto dos pontos de *X*, tal que *Bx* esteja totalmente incluído em *X*, como pode ser visto na Figura 2.17.

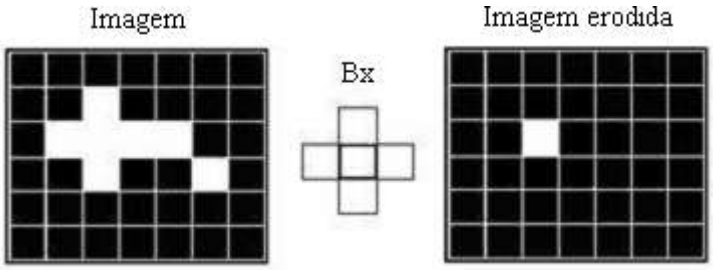


Figura 2.17 – Erosão de uma imagem (LEITE, 2004).

2.3.3.7. Filtro Mínimo

O filtro Mínimo executa uma erosão em escalas de cinza. De forma mais clara, ela atribui o novo valor de cada pixel da imagem com o menor valor dos píxeis vizinhos (RASBAND, 2011). Um exemplo disso está na Figura 2.18.

Figura 2.18 – Aplicação do filtro Mínimo com uma vizinhança de 3x3.

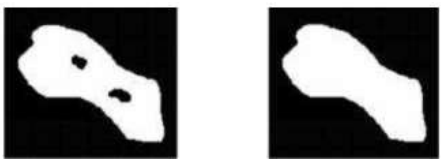
2.3.3.8. Filtro Máximo

O filtro Máximo por sua vez aplica uma dilatação em uma imagem em escalas de cinza fazendo com que o novo valor de cada pixel seja o maior valor dos seus píxeis vizinhos (RASBAND, 2011). A Figura 2.19 mostra um exemplo da aplicação deste filtro.

Figura 2.19 – Aplicação do filtro Máximo com uma vizinhança de 3x3.

2.3.3.9. Remoção de buracos

A remoção de buracos é um procedimento aplicado em imagens binárias no qual toda vez que se encontra a cor do plano de fundo no interior de objetos, esta cor é preenchida pela cor do próprio objeto. Por exemplo, ao ter uma cor de fundo preta e objetos brancos, no momento em que esse filtro for aplicado, todos os buracos de cor preta encontrados no interior desses objetos serão preenchidos com a cor branca (RASBAND, 2011). Isso pode ser visto na Figura 2.20.



(a) (b)

Figura 2.20 – Utilização do filtro de remoção de buracos em (a) com o resultado em (b).

2.3.3.10. Ajuste de contraste

O contraste pode ser definido como a diferença nas propriedades visuais que torna um objeto distinguível de outros objetos e de seu plano de fundo. Em imagens digitais, o contraste é determinado pela diferença na intensidade de níveis de cinza entre as áreas claras e escuras de um determinado objeto (CAMPBELL e ROBSON, 1968).

Através de um histograma é possível descrever a distribuição dos níveis de cinza em relação à quantidade em que eles aparecem em uma imagem. Abaixo, na Figura 2.21, pode-se ver o histograma de uma imagem de alto (a) e baixo contraste (b).

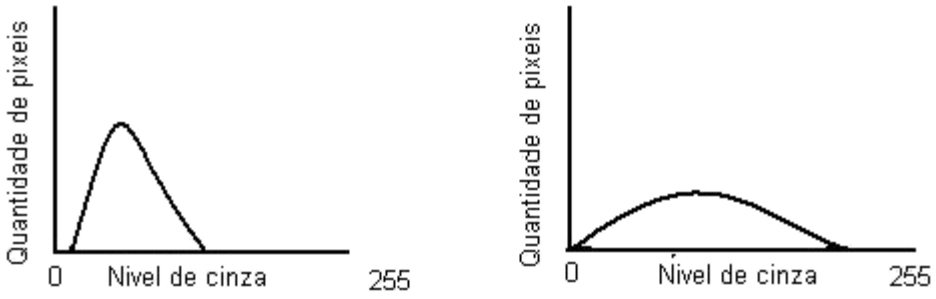


Figura 2.21 – Histograma de uma imagem de alto (a) e baixo (b) contraste.

Além disso, é possível, através do histograma, aplicar técnicas de realce de contraste no qual são utilizadas transformações radiométricas que consistem em mapear as variações da intensidade luminosa em um dado intervalo para outro intervalo desejado, e desta forma aumentar ou diminuir o contraste de uma imagem (CAMPBELL e ROBSON, 1968).

2.3.3.11. Conversão para binário (limiarização)

A limiarização ou conversão para binário é um processo que divide a imagem em objetos e plano de fundo. Ou seja, a partir de um nível de cinza que é passado por parâmetro

(denominado limiar), todos os píxeis que estão com um nível de cinza acima daquele valor terão seu valor trocado para o valor máximo (no caso de uma imagem 8 *bits*, branco, ou 255) e os que tiverem seus valores abaixo desse parâmetro serão alterados para o valor mínimo (preto, ou 0) (GONZALEZ, 2000). Este limiar pode ser determinado através de um histograma dos píxeis de uma imagem, observando seus vales e definido o limiar como sendo o mais próximo ao valor médio da escala de níveis de cinza. A Figura 2.22 traz um exemplo de conversão para binário de uma imagem.

Figura 2.22 – Processo de conversão para binário ou limiarização de uma imagem

2.3.3.12. Esqueletização

O processo de esqueletização de uma imagem binária consiste em reduzir o objeto analisado a uma cadeia simples, com a largura de apenas um pixel. Preservando, no entanto, todas as características importantes da imagem (SOUZA e BANON, 2003).

A definição de esqueleto diz que, um ponto pertence ao esqueleto somente se ele é o centro de um círculo máximo, sendo que este círculo toca a borda do objeto em pelo menos dois pontos distintos. Como existe certa dificuldade em implementar círculos computacionalmente, normalmente utiliza-se de Figuras geométricas mais simples tais como retângulos e losangos (SOUZA e BANON, 2003).

Um dos algoritmos mais utilizados é o desenvolvido em 1984 por Zhang e Suen (ZHANG e SUEN, 1984). Neste trabalho, foi proposto um algoritmo paralelo de esqueletização que trouxe resultados bastante superiores se comparados a outros da mesma época, sendo este um dos trabalhos mais citados até mesmo nos trabalhos mais recentes. A implementação deste algoritmo pode ser encontrado no software *ImageJ* (RASBAND, 2011). O resultado da aplicação deste filtro é apresentado na Figura 2.23.

(a) (b)

Figura 2.23 – Exemplo da aplicação do algoritmo de esqueletização de Zhang e Suen (1984).

2.3.3.13. Detecção de bordas de Sobel

Uma borda é o limite entre duas regiões com propriedades relativamente distintas de nível de cinza (GONZALEZ, 2000). O filtro Sobel calcula o gradiente da intensidade da imagem em cada ponto, dando a direção da maior variação de claro para escuro e a quantidade de variação nessa direção. Quando essas variações de claro-escuro são intensas, elas possivelmente correspondem a fronteiras de objetos (DUDA, 1973).

O detector de bordas de Sobel utiliza um par de máscaras 3x3 que são convoluídas com a imagem original com o objetivo de calcular as aproximações das derivadas. Uma delas estima a variação dos níveis de cinza na direção de x (colunas) e a outra estima a variação na direção de y (linhas). Como resultado, é como se a máscara percorresse toda a imagem manipulando um quadrado 3x3 de píxeis por vez. As máscaras *Gx* (a) e *Gy* (b) utilizadas são as seguintes (DUDA, 1973):

$$\begin{pmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{pmatrix} \quad \begin{pmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix}$$

(a) (b)

A magnitude desta variação é então calculada usando a seguinte fórmula (DUDA, 1973):

$$|G| = \sqrt{Gx^2 + Gy^2}$$

Equação 8

Normalmente uma magnitude aproximada é utilizada usando a seguinte fórmula:

$$|G| = |Gx| + |Gy|,$$

Equação 9

o que torna o processo computacionalmente mais rápido. Desta forma, é possível calcular a direção da variação dos níveis de cinza (DUDA, 1973). A Figura 2.24 traz um exemplo da aplicação do filtro de detecção de bordas de Sobel da ferramenta *ImageJ*.

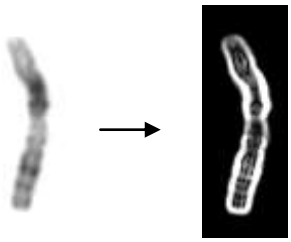


Figura 2.24 – Exemplo de aplicação do filtro de detecção de bordas de Sobel.

2.3.3.14. Suavização Gaussiana

A suavização Gaussiana é um tipo de filtro de suavização de imagens que utiliza uma função Gaussiana para calcular a transformação que será aplicada em cada pixel da imagem. A equação da função Gaussiana em duas dimensões (uma imagem) é:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

Equação 10

na qual x e y são as coordenadas do pixel e σ (sigma) é um parâmetro associado a largura da Gaussiana (SHAPIRO e STOCKMAN, 2001). O parâmetro σ se refere ao desvio padrão da distribuição gaussiana.

Quando esta fórmula é aplicada em uma imagem, ela produz uma superfície cujas curvas de níveis são círculos concêntricos a partir do ponto central. Os valores obtidos a partir

dessa distribuição são utilizados para se construir uma matriz de convolução, esta que é aplicada na imagem original. Dessa forma, o novo valor de cada pixel será a média ponderada dos píxeis vizinhos. Assim, o resultado será uma imagem borrada, que preserva melhor os limites e as bordas do que outros filtros, sendo uma suavização mais uniforme (SHAPIRO e STOCKMAN, 2001).

3. TRABALHOS RELACIONADOS

Esta seção tem como objetivo apresentar os principais trabalhos encontrados na literatura que visam o mesmo objetivo do desenvolvido nesta dissertação. Portanto, são introduzidos os trabalhos relacionados à detecção do centrômero, o qual é o foco desta dissertação, assim como os trabalhos relacionados a identificação dos cromossomos, o qual é utilizado como forma de validação das técnicas desenvolvidas nesta dissertação.

3.1. Detecção do centrômero

Diversas características são consideradas importantes na identificação de cromossomos, tal como o eixo longitudinal, a polaridade, o padrão de bandas, o tamanho, a largura, entre outras. Uma das mais importantes características é o centrômero, pois com uma boa taxa de acertos na detecção do mesmo, é possível utilizá-lo como um filtro muito poderoso na identificação, dividindo os cromossomos nos seus três grupos (metacêntrico, submetacêntrico e acrocêntrico), aumentando consideravelmente a taxa de acertos da identificação.

Porém, são poucos os trabalhos que buscam pesquisar e desenvolver métodos e algoritmos para detecção do centrômero. Segundo vários autores, apesar de ser uma característica importante para identificação, é muito difícil de detectar o centrômero, bem como conseguir uma alta taxa de acertos (KAO et al., 2008; SCHWARTZKOPF et al., 2005; LEGRAND et al., 2008). Uma classificação dos trabalhos encontrados na literatura pode ser feita de acordo com 4 grupos:

1. Trabalhos voltados à identificação de cromossomos, mas que não utilizam/citam a detecção do centrômero (muito poucos);
2. Trabalhos que citam a importância da detecção do centrômero, porém dizem utilizar alguma das técnicas apresentadas por algum dos 3 principais trabalhos a serem apresentados nas seções a seguir;

3. Trabalhos que citam a importância da detecção do centrômero, porém não utilizam por acreditarem ser difícil de programar e principalmente de se obter bons resultados, mas geralmente citam um ou mais dos 3 principais trabalhos ou trabalhos que citam algum desses 3;
4. Trabalhos que realmente apresentam uma técnica de detecção de centrômero, que são os 3 trabalhos a serem apresentados a seguir.

Dentre os trabalhos da classificação 1 pode-se citar os de Guimarães et al. (2003) e Kim et al. (2011); na classificação 2 pode-se citar os de Oskouei e Shanbehzadeh (2010), Nanni (2006), Roshtkhari e Setarehdan (2008), Stanley et al. (1996) e Trimananda (2010); sendo que da classificação 3, pode-se citar os de Kao et al (2008), Schwartzkopf et al. (2005) e Legrand et al. (2008). Os trabalhos da classificação 4 serão apresentados nas seções a seguir.

Portanto, existem poucos trabalhos que realmente desenvolveram uma técnica para detecção dos cromossomos (PIPER e GRANUM, 1989; WANG et al., 2008; MORADI et al., 2003), mas apesar disso, estes trabalhos são freqüentemente citados na bibliografia por diversos autores, sendo que grande parte dos trabalhos desenvolvidos na área de identificação de cromossomos utiliza o que foi desenvolvido por esses autores como método de detecção do centrômero.

3.**1.1. Técnica de Piper e Granum**

Uma das primeiras metodologias propostas e uma das mais citadas e utilizadas por outros autores foi a desenvolvida por Piper e Granum (1989), e a partir dela diversas outras formas de se detectar o centrômero surgiram como formas de aperfeiçoar o método. Esta metodologia tem como objetivo gerar um perfil do cromossomo baseado tanto na sua largura como nos níveis de cinza. Foram utilizadas três bases de dados no estudo, sendo estas divididas em cromossomos sobrepostos (não utilizados no experimento), cromossomos que se encostam um no outro e cromossomos soltos, conforme é apresentado na Tabela 3.1.

A base de dados de Copenhagen apresentada por Philip e Granum (1980) é uma das mais utilizadas pelos autores, tal como Biyani et al. (2005), Castleman (2000), Lundsteen et

al. (1985), Piper e Granun (1989) e Wang et al. (2005, 2008, 2009). Ela apresenta 44 arquivos com informações a respeito dos cromossomos 1 ao 22 (não consta cromossomos X e Y), e as informações são codificadas através de seqüências de *strings*. Cada arquivo contém 100 linhas, sendo que cada linha apresenta informações de um cromossomo, conforme é mostrado no exemplo abaixo:

/ 5467 119 22 27 9 / AA==a==E===d==A==a=Aa=A=a=b

em que o valor 5467 é um identificador daquele cromossomo, 119 é a identificação da metáfase da qual a amostra foi retirada, 22 é o tipo de cromossomo, 27 é o comprimento geral da *string* e 9 é o tamanho do braço curto (que determina a posição do centrômero).

Tabela 3.1: Bases de dados utilizadas no trabalho de Piper e Granum.

Base	Sobrepostos	Encostando	Total
Copenhagen	184	2165	8106
Edimburgo	96	1243	5469
Filadélfia	130	2517	5817

Porém, os dados não são apresentados em forma de imagem, devido ao fato de os dados apresentados por esta base estarem codificados em seqüências e *strings*. Além disso, não foi possível identificar do que se tratam as informações dos perfis apresentadas pelos arquivos deste banco de dados, e estas informações já estão em uma forma normalizada em relação à amplitude (por exemplo, caso as informações fossem em relação aos níveis de cinza), o que acaba muitas vezes tornando estes dados de certa forma imprecisos, pois não se têm acesso aos dados originais, impossibilitando assim a geração de filtros apropriados para o trabalho a ser desenvolvido.

O trabalho original de Philip e Granum (1980) possivelmente apresenta informações a respeito do que se tratam os dados apresentados, além de referências a um possível conjunto original com informações mais completas. Porém, devido ao ano de publicação deste trabalho, não foi possível obter acesso a essas informações. As bases de dados de Edimburgo e de Filadélfia também são utilizadas por alguns autores, mas com uma menor freqüência. Apesar

disso, não foi possível de se obter acesso as mesmas por serem bases bastante antigas e pela dificuldade de entrar em contato com os autores.

Em relação à técnica desenvolvida para detecção do centrômero, os autores citam que alguns trabalhos anteriores utilizam somente a largura do cromossomo (GRAHAM, 1987; GROEN, 1985 apud PIPER e GRANUM, 1989) ou análise da curvatura das bordas do cromossomo como forma de detecção do centrômero (GALLUS e NEURATH, 1970 apud PIER e GRANUM, 1989), mas que não são inteiramente satisfatórias, pois segundo os autores, o estreitamento na região do centrômero pode ser mal representada pelos contornos da borda, e o perfil de larguras do mesmo pode conter ruídos devido as técnicas de preparação das imagens.

Desta forma, uma equação foi proposta para a geração do perfil do cromossomo de forma a utilizar tanto informações a respeito da largura do cromossomo como informações em relação aos níveis de cinza do mesmo. A equação é aplicada em cada ponto do eixo longitudinal do cromossomo, ao longo de uma linha transversal que corta cada ponto:

$$peso = \frac{\sum_1^n mi \times di^2}{\sum_1^n mi}$$

Equação 11

Nesta equação, *mi* é a densidade do pixel (nível de cinza) e *di* é a distância euclidiana do eixo principal.

Um problema existente é em relação a cromossomos acrocêntricos, os quais o centrômero é encontrado na extremidade de um dos braços. O problema é que cromossomos metacêntricos e submetacêntricos geralmente também possuem seus extremos mais estreitos. Além disso, se for desconsiderado os extremos dos cromossomos, ao comparar os cromossomos acrocêntricos com cromossomos metacêntricos e submetacêntricos, os acrocêntricos ainda terão, possivelmente, um mínimo global dos pesos ao decorrer do seu eixo longitudinal, porém não é tão evidente quanto os mínimos dos cromossomos metacêntricos e submetacêntricos. O mesmo vale para os extremos dos cromossomos acrocêntricos se forem comparados a metacêntricos e submetacêntricos, visto que, geralmente, os extremos dos cromossomos acrocêntricos apresentam um mínimo global maior e mais longo em relação aos metacêntricos e submetacêntricos.

Apesar de isto ser uma diferença a ser considerada no momento de decidir a classificação em relação ao centrômero, isso ainda pode atrapalhar, devido ao fato de que os cromossomos de uma célula seguem um padrão morfológico, mas que varia bastante, principalmente em relação à baixa qualidade de imagens geradas dos mesmos, ocasionando erros tal como encontrar os mínimos globais erroneamente nos extremos de metacêntricos e submetacêntricos ou nos píxeis mais internos de cromossomos acrocêntricos. Para evitar isso, utilizou-se uma técnica denominada *reflexão*, tal como é apresentado na seção 4.3.8. Assim, segundo os autores, no perfil final gerado, caso os cromossomos sejam acrocêntricos, os extremos continuarão contendo o mínimo global, e desta forma confirmando que os mesmos são realmente acrocêntricos.

Na técnica de reflexão desenvolvida neste trabalho o resultado não foi semelhante ao apresentado por Piper e Granum (1989), visto que ao invés da proposta do autor de se aumentar o número de acertos em cromossomos acrocêntricos, o algoritmo desenvolvido acabou aumentando a taxa de acertos de cromossomos submetacêntricos e metacêntricos. Detalhes em relação ao algoritmo desenvolvido são apresentados na seção 4.3.8.

A taxa de acertos na detecção do centrômero para cada cromossomo na técnica de Piper e Granum é apresentada na Tabela 5.15 da seção 5.5.3, assim como um comparativo dos mesmos com as técnicas desenvolvidas neste trabalho.

3.1.2. Técnica de Wang

Em (WANG et al., 2008) foi desenvolvida uma técnica que utiliza uma equação semelhante a desenvolvida por Piper e Granum (1989) e apresentada na seção 3.1.1, mas com alguns pequenos detalhes que a diferem da mesma principalmente em relação a uma prévia subdivisão dos cromossomos em três grupos e da base de dados utilizada.

A base de dados utilizada é uma base que consiste de 50 metáfases obtidas de pacientes suspeitos de leucemia. Esta base foi gerada pelo laboratório de genética do Centro de Ciências de Saúde da Universidade de Oklahoma, nas quais 26 dessas metáfases foram classificadas como normais e 24 como anormais, contendo alterações numéricas e estruturais. As imagens foram obtidas utilizando um microscópio óptico Nikon LABOPHOT-2. Mais

detalhes sobre a base de dados são apresentados na Tabela 3.2 e também pode ser obtida em (WANG et al., 2008).

Tabela 3.2: Detalhes da base de dados utilizada por Wang.

Quantidade de células	50
Quantidade de células normais	26
Quantidade de células anormais	24
Quantidade de cromossomos por célula (média)	45.74
Quantidade de cromossomos muito dobrados	134
Quantidade de cromossomos retos e pouco dobrados	2153
Quantidade total de cromossomos	2287

A equação proposta por Wang et al. (2008) é bastante semelhante a de Piper e Granum (1989), porém, com uma mudança no denominador da mesma, tal como é apresentado a seguir:

$$peso = \frac{\sum_1^n gi \times di^2}{\sum_1^n di^2}$$

Equação 12

Nesta equação, *gi* corresponde à densidade do pixel (nível de cinza) e *di* é a distância euclidiana do eixo principal. Os perfis foram gerados e testados nos mais diversos tamanhos, tal como 300, 400, 550, 750 e 800 píxeis, sendo que ao final foi adotado um tamanho padrão de 400 píxeis.

Além disso, após a geração do perfil dos cromossomos, os mesmos são divididos em três grupos, sendo que para cada grupo um diferente critério é aplicado na busca pela posição do centrômero em relação ao perfil gerado (mínimo global dos pesos). Primeiramente é calculado o comprimento de determinado cromossomo em sua célula e também o comprimento médio de todos os cromossomos da mesma célula. Com estas informações os cromossomos são divididos nos seguintes grupos:

1. Caso o comprimento do cromossomo seja maior que a média de comprimento, o cromossomo é considerado como fazendo parte do grupo I. Neste grupo, o mínimo global é buscado truncando 20% do seu comprimento em ambas extremidades. Ou

seja, considerando um perfil de 400 píxeis nas posições 0 a 399, busca-se pelo mínimo global entre os píxeis das posições 79 e 319. Ou seja, o grupo I abrange os cromossomos de maior comprimento em determinada célula, considerando menos os extremos dos cromossomos na busca pelo mínimo global.

2. Caso o comprimento do cromossomo seja menor que a média, ele faz parte do grupo II. Neste grupo, busca-se pelo mínimo global truncando 15% do seu comprimento em ambos os extremos. Ou seja, considerando um perfil de 400 píxeis nas posições 0 a 399, busca-se pelo mínimo global entre os píxeis das posições 59 e 339. Ou seja, o grupo II abrange os cromossomos de menor comprimento daquela célula, e os extremos destes cromossomos são mais considerados na busca do mínimo global destes casos.
3. Caso não se encontre nenhum mínimo global em relação as regras anteriores, busca-se pelo mínimo global em todo o perfil gerado.

Segundo Wang et al. (2008) , os valores dos limiares utilizados para truncar as extremidades dos perfis foram definidos em um estudo anterior (GRAHAM, 1987), o qual infelizmente não obteve-se acesso ao artigo do mesmo.

3.1.3. Técnica de Moradi

No trabalho de Moradi et al. (2003) foi desenvolvida uma técnica bastante simples, mas bastante citada e utilizada em diversos outros trabalhos (WANG et al., 2008; GRISAN et al., 2009; POLETTI et al., 2008; OSKOUEI e SHANBEHZADEH, 2010; TRIMANANDA, 2006, 2010; ROSHTKHARI e SETAREHDAN, 2008). A técnica consiste na mesma idéia anterior que é a de se gerar um perfil do cromossomo, que neste caso é denominado *vetor de projeção horizontal*, para em seguida analisar tais vetores e então definir a posição e a classificação do centrômero baseado nestes dados.

O banco de dados utilizado foi produzido no Laboratório de Citogenética do Instituto do Câncer, do hospital Imam, na cidade de Tehran no Iran. As imagens foram obtidas por um microscópio Leitz ortholux e foram segmentadas por um especialista da área, e então escaneadas com uma resolução de 300dpi, com 256 níveis de cinza.

A técnica aborda uma sequência de passos bastante simples, tal como é apresentado na Figura 3.1. Primeiramente é gerada uma imagem binária dos cromossomos, sendo que para isso é feito o uso de um histograma filtrado da imagem. Para se obter tal histograma, uma série de filtros é aplicada a imagem tal como o filtro de suavização de *Savitzky-Golay* (SAVITZKY e GOLAY, 1964) e filtros medianos, para que em seguida seja gerada uma imagem binária da imagem original do cromossomo.

Com as imagens binárias geradas, é possível então gerar os vetores de projeção horizontais. Como as imagens binárias consistem em píxeis de cor branca (valor 1, background) e píxeis de cor preta (valor 0, cromossomo), a técnica de Moradi simplesmente soma em cada linha do vetor de projeção a quantidade de píxeis de valor 1, ou seja, os vetores de projeção são baseados na distância entre os extremos de cada linha do vetor de projeção.

Com o vetor de projeção em mãos, inicia-se o passo de detecção da posição do centrômero, que consiste em simplesmente buscar a posição de menor distância do vetor de projeção, tal como é apresentado na Figura 3.2 do trabalho de Moradi (2003).

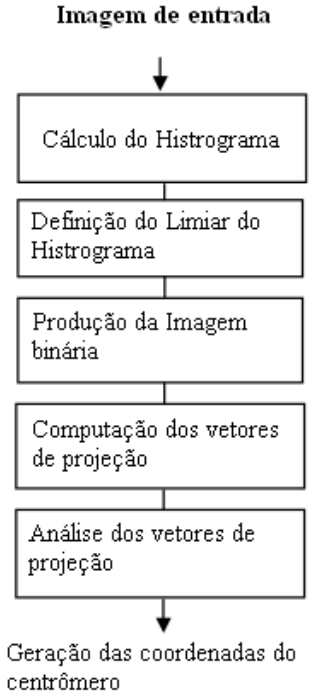


Figura 3.1: Sequência de passos da técnica de Moradi (MORADI et al., 2003).

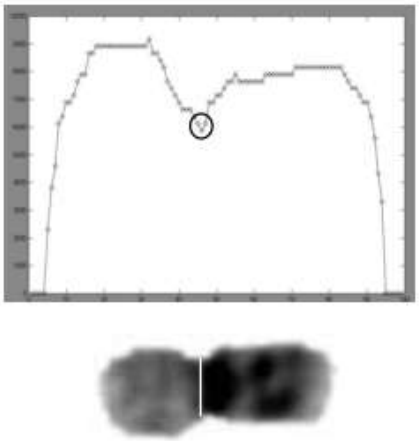


Figura 3.2: Vetor de projeção horizontal com a posição do centrômero marcada tanto no vetor quanto na imagem original (MORADI et al., 2003).

Os resultados desta técnica foram obtidos após a aplicação dos algoritmos em 87 imagens da base de dados. Diferentemente dos trabalhos citados anteriormente, apesar do autor considerar este trabalho, na época, como sendo o de maior taxa de acertos na localização do centrômero, o mesmo não apresenta os resultados na forma de uma taxa de acertos, mas somente em relação a uma visão geral da distância em que foi encontrada a posição do centrômero pelo algoritmo em relação à posição marcada pelo especialista, tal como pode ser visto na Tabela 3.3.

Tabela 3.3: Resultados obtidos pela técnica de Moradi (MORADI et al., 2003).

Valor médio do erro absoluto	4.3 (pixel)
Desvio padrão do erro absoluto	3.8
Valor médio do erro normalizado	0.041
Desvio padrão do erro normalizado	0.03

Apesar de ser um trabalho bastante citado na literatura, o mesmo não apresenta maiores informações sobre como, a partir da imagem binária, é encontrado o eixo longitudinal do cromossomo, para que se torne possível a geração de tal vetor de projeção. Isto pode ser devido ao fato de que Moradi, de acordo com os exemplos apresentados neste trabalho, considera que os cromossomos estão sempre retos ou com pouco encurvamento, o que não

acontece em casos reais de metáfases, pois grande parte dos cromossomos é curvada, conforme pode ser visto nos exemplos da Figura 2.10.

Portanto, a idéia apresentada da uma noção de que a o eixo longitudinal considerado é simplesmente aquele que corta o cromossomo em linha reta, de cima a baixo, o que não pode ser aplicado à grande parte de casos reais, pois geralmente os cromossomos apresentam certo grau de curvatura. Além disso, conforme pode ser visto em uma das técnicas desenvolvidas nesta dissertação (seção 4.3.3), a utilização de somente a distância dos extremos de um ponto do vetor de projeção horizontal não traz resultados satisfatórios, ao contrário do que é afirmado no trabalho de Moradi (Moradi et al., 2003).

4. METODOLOGIA

Este capítulo tem como objetivo apresentar a metodologia desenvolvida neste trabalho para a detecção do centrômero. Primeiramente, na seção 4.1 será feita uma introdução ao processo de detecção do centrômero desenvolvido, e na seção 4.2, serão apresentadas as ferramentas e a base de dados utilizada neste trabalho. Além disso, esta seção também aborda as características dos cromossomos que são extraídas das imagens e que futuramente serão utilizadas nos métodos e nos algoritmos de detecção. A seção 4.3 irá apresentar os 5 métodos desenvolvidos para detecção do centrômero e suas diferenças, e a seção 4.4 irá apresentar os dois algoritmos propostos nesta metodologia, que farão uso dos 5 métodos desenvolvidos para se obter uma técnica com maior taxa de acertos na detecção do centrômero. Por fim, a seção 4.5 ira tratar a respeito do processo e ajuste dos pesos utilizados em cada um dos métodos desenvolvidos.

4.1. Detecção do centrômero

A idéia inicial deste trabalho era o desenvolvimento de um único método de detecção, e a partir daí, efetuar os ajustes necessários para que o mesmo atingisse uma boa taxa de acertos. Porém, com o decorrer do tempo, vários métodos foram sendo desenvolvidos, alguns com taxas de acertos em geral melhores, e também com melhores taxas de acertos em determinado grupo de cromossomos (metacêntricos, acrocêntricos e submetacêntricos). Esta taxa de acertos de cada método varia principalmente em relação ao comprimento dos cromossomos, onde para determinados comprimentos, a utilização de certo método é melhor que a de outro, ou em alguns casos até mesmo a utilização de vários métodos de acordo com o comprimento do cromossomo.

Portanto, ao final desta etapa, foram desenvolvidos cinco métodos para detecção do centrômero, alguns baseados em métodos encontrados na literatura e adaptados para se garantir uma maior taxa de acertos, e outros com algumas modificações a partir de idéias próprias, e, por fim, de forma a utilizar o melhor de cada método desenvolvido, elaborou-se duas metodologias com a criação de dois algoritmos que utilizam estes métodos desenvolvidos. O primeiro algoritmo proposto irá verificar o comprimento do cromossomo, e

de acordo com este comprimento, irá aplicar apenas um dos cinco métodos desenvolvidos. Já o segundo algoritmo aplicará, de acordo com o comprimento do cromossomo, um ou mais métodos.

Para que seja possível a detecção do centrômero a partir destes métodos, primeiramente as imagens devem passar por um pré-processamento em que alguns filtros são aplicados as mesmas. Além disso, algumas informações são extraídas destas imagens de forma que tais informações são utilizadas para determinar o peso de certas variáveis presentes em todos os métodos.

Portanto, esta seção tem o objetivo de apresentar as metodologias utilizadas, adaptadas e criadas para os diferentes métodos de detecção do centrômero.

4.2. Pré-processamento

Antes de colocar em prática os métodos de detecção do centrômero, é interessante e muitas vezes necessária a aplicação de alguns ajustes nas imagens a fim de melhorar a taxa de acertos. Estes ajustes podem variar dependendo do método utilizado, visto que cada um trabalha de uma forma diferente pela busca do centrômero. Além disso, cada método de detecção do centrômero contém variáveis, cada uma com um peso específico para determinado método. Esta seção visa abordar estas duas etapas de pré-processamento, que ocorrem antes da aplicação dos métodos de detecção do centrômero.

4.2.1. Base de dados e ferramentas.

Diferente dos principais trabalhos voltados a detecção do centrômero encontrados na literatura, a base de dados utilizada é a disponibilizada pelo Laboratório de Imagem Biomédica (POLETTI et al., 2008). Esta base de imagens é constituída de imagens de 119 metáfases manualmente segmentadas e classificadas por especialistas, sendo ao final 5474

imagens divididas em 119 pastas, cada uma com os 46 cromossomos. Cada uma das pastas representa uma metáfase segmentada.

A opção por utilizar imagens já segmentadas se deve ao fato de que já se desenvolveu em um trabalho anterior (KURTZ et al., 2008) um algoritmo bastante satisfatório para segmentação de cromossomos, e também pelo foco deste trabalho não ser a segmentação, mas sim o estudo e desenvolvimento de novas técnicas de detecção do centrômero. Além disso, a utilização desta base de imagens ao invés das bases utilizadas pelos principais trabalhos relacionados na literatura se deve a dois fatores: ou a base de imagens utilizada pelos autores não estava mais disponível ou uma base era privada. Portanto, a única base encontrada que satisfez as necessidades deste trabalho foi a citada anteriormente.

A linguagem de programação utilizada foi a linguagem Java, bem como a ferramenta *ImageJ* (RASBAND, 2011; ABRAMOFF et al., 2004). O *ImageJ* é uma ferramenta poderosa para análise e processamento de imagens, e também fornece bibliotecas que os usuários possam utilizar tanto em sistemas próprios quanto em *plugins* para o próprio *ImageJ*, estes que também estão disponíveis para os usuários utilizarem em seus projetos. O *ImageJ* é utilizado em todas as etapas de desenvolvimento deste trabalho para manipulação, aplicação de filtros e extração de dados das imagens.

4.2.2. Preparação e geração das imagens

Para ser possível o desenvolvimento das técnicas de detecção do centrômero, é necessário realizar algumas etapas iniciais que envolvem a geração de novas imagens e dados que irão auxiliar no futuro a implementação destas técnicas. As imagens da base de dados estão em um formato BMP monocromático de 8 *bits*, porém, foram transformadas em imagens PNG de 8 *bits* de forma a facilitar seu uso na linguagem Java e com a biblioteca *ImageJ*.

Alguns exemplos de imagens desta base de dados são apresentados na Figura 2.2. A partir destas imagens é gerada então uma nova imagem denominada *máscara* (diferente do termo “máscara” utilizado em convolução). Esta máscara nada mais é que uma imagem binária representando o que é e o que não é cromossomo naquela imagem. Ou seja, os píxeis

de cor branca (valor igual a 1) representam o cromossomo e os de cor preta (valor igual a 0) representam o *background*. Os passos para geração envolvem a aplicação de alguns filtros que são aplicados a todas as imagens bem como são aplicados no trabalho de (KURTZ et al., 2008), que seguem a seguinte sequência:

1. Suavização, utilizando uma média de vizinhança 3x3;
2. Filtro de detecção de bordas de Sobel do *ImageJ*;
3. Ajuste de contraste, de forma que a máscara se torne mais visível;
4. Filtro de remoção de buracos;
5. Filtro mínimo, utilizando uma vizinhança de 3x3;

Assim, é gerada uma imagem binária (a máscara) para cada imagem da base de dados utilizada conforme mostra a Figura 4.1.b. Com a máscara é possível extrair diversas informações tal como a área do cromossomo, a média de níveis de cinza, entre outros, tal como é apresentado na seção 4.2.3.

O próximo passo então é a determinação do eixo longitudinal que corta o cromossomo no meio, tal como é mostrado na Figura 4.1.c. Para isto, é gerado o esqueleto da imagem. O esqueleto de um cromossomo é definido como sendo a linha central que corta o cromossomo ao longo do seu comprimento (eixo longitudinal). A Figura 4.1.a mostra a imagem de um cromossomo 3, em 4.1.b sua máscara e em 4.1.c seu esqueleto.

(a) (b) (c)

Figura 4.1: Cromossomo original (a), sua máscara (b) e seu esqueleto (c).

O esqueleto é a peça chave para a extração das informações dos cromossomos e definição dos pesos das variáveis, visto que ele é utilizado para percorrer o cromossomo longitudinalmente, determinar seu comprimento e o comprimento de seus braços a fim de calcular o valor das razões cromossômicas.

O esqueleto é gerado a partir da máscara da imagem, mas antes de gerá-lo, alguns filtros são aplicados na máscara com o intuito de tornar suas bordas o mais arredondado possível. A sequência foi definida após diversos testes tentando definir a melhor ordem de aplicação dos filtros e a quantidade que cada filtro deve ser aplicado, e assim foi definida como:

1. Suavizar a imagem utilizando uma vizinhança 3x3 e converte-la para binário 10 vezes.
2. Dilatar a imagem utilizando um filtro de vizinhança 3x3.
3. Erodir a imagem, também utilizando uma vizinhança 3x3.
4. Aplicar suavização gaussiana, com sigma de valor 3 e uma precisão de 0.003 (RASBAND, 2011; ABRAMOFF et al., 2004).
5. Converter a imagem para binário.
6. Aplicar filtro da mediana utilizando uma vizinhança de 5x5.
7. Converter a imagem para binário.

A partir deste ponto obtém-se uma nova máscara, com suas bordas arredondadas, ideal para a aplicação do algoritmo de esqueletização da mesma. O algoritmo de esqueletização utilizado é baseado no algoritmo de “afinamento” desenvolvido em (ZHANG e SUEN, 1984). Com o esqueleto da imagem definido é possível então gerar um vetor contendo o caminho que percorre o eixo longitudinal da imagem. Estes dados são extremamente importantes, pois são essenciais na aplicação dos métodos desenvolvidos, pois é a partir deste esqueleto que se percorre a imagem e são aplicadas as equações utilizadas pelos métodos. Além disso, é importante na determinação do comprimento do cromossomo, de seus braços, e outros dados que serão apresentados nas seções a seguir.

4.2.3. Extração de informações e determinação dos pesos das variáveis

Em cada um dos cinco métodos desenvolvidos, algumas informações são extraídas, a fim de auxiliar a detecção do centrômero e a determinação de qual grupo relacionado a posição do centrômero certo cromossomo pertence. Com estas informações é possível iniciar um processo de ajuste de pesos, no qual serão definidos os pesos ideais de tais variáveis

utilizadas pelos métodos. Tais pesos são determinados através de uma exaustiva execução dos métodos de detecção do centrômero, sendo que ao final de cada execução, é realizada uma verificação em relação ao número total de acertos. Ao final, os pesos em que se obteve um maior número de acertos são considerados como os ideais para aquele método. Tais testes são realizados tanto nos métodos individuais como nos dois algoritmos propostos para detecção do centrômero. Este processo é bastante exaustivo devido a grande quantidade de imagens da base de dados e do tempo de processamento para extração destas informações. Mais detalhes sobre a determinação dos pesos pode ser visto na seção 4.5. Após a realização destes testes, obtêm-se definido para cada método os pesos considerados ideais para estas variáveis.

Em relação à extração das informações (que irão definir os pesos das variáveis), a principal informação a ser extraída está relacionada a um perfil da forma do cromossomo, denominado por muitos autores de *Shape Profile* (PIPER; GRANUM, 1989; STANLEY et al., 1996; WANG et al., 2008, 2009; CHO et al., 2004) o qual varia de método para método. Este perfil visa criar um vetor de projeção horizontal do cromossomo (ou seja, baseado no seu esqueleto), no qual o mínimo global é considerado como sendo a posição do centrômero. Este vetor pode estar relacionado somente à largura do cromossomo em cada ponto do esqueleto (MORADI et al., 2003) ou até mesmo ser definido através de equações mais complexas que utilizem tanto a largura em cada ponto do esqueleto quanto os níveis de cinza no de correr dos píxeis das linhas transversais que cortam cada ponto do esqueleto.

Para que seja possível definir este perfil é necessário primeiro extrair outras informações das imagens, relacionadas primeiramente ao próprio esqueleto do cromossomo, ao comprimento dos cromossomos, a proporção de comprimento do braço longo pelo braço curto (razão cromossômica, a partir daqui denominado "razão", ou somente *r*), ao intervalo ao longo do eixo longitudinal no qual é feita a busca pelo centrômero em determinado cromossomo (denominado *k*) e aos intervalos de comprimentos nos quais serão aplicados os valores de *r* e *k*.

4.2.3.1. Razão Cromossômica

A razão cromossômica é determinada, a partir da posição definida como sendo a posição do centrômero, a razão entre o comprimento do braço longo pelo comprimento do braço curto do cromossomo, ou seja:

$$r = \frac{tamBL}{tamBC}$$

onde *tamBL* é o comprimento do braço longo, *tamBC* o comprimento do braço curto e *r* o valor da razão. Dependendo do valor de *r* encontrado para determinado cromossomo em certo método, ele será classificado como metacêntrico, submetacêntrico ou acrocêntrico.

Com o valor de *r* obtido, serão definidos dois pesos para cada método: um valor de *r* inicial, denominado *rIni*, e um valor de *r* final, denominado *rFim*, em que, a partir do valor de *r* encontrado para determinado cromossomo:

- Se *r* for menor que *rIni*, então o cromossomo é classificado como metacêntrico;
- Se *r* for maior ou igual à *rIni* e menor que *rFim*, então o cromossomo é classificado como submetacêntrico;
- Se *r* for maior ou igual à *rFim*, então o cromossomo é classificado como acrocêntrico.

Ou seja, dependendo da proporção do comprimento do braço longo em relação ao braço curto do cromossomo, ele será classificado como metacêntrico, submetacêntrico ou acrocêntrico. Os pesos das variáveis *rIni* e *rFim* variam de um método para outro, e também variam, em cada um dos métodos, de acordo com o comprimento do cromossomo.

4.2.3.2. Intervalo de busca (*k*)

O intervalo de busca pela posição do centrômero é definido como sendo a posição inicial e final ao longo do seu eixo longitudinal no qual é feita a busca pelo centrômero. A posição inicial é definida através da variável *kIni* e a posição final da variável *kFim*. Estes valores são diferentes para cada método e também variam de acordo com o comprimento do cromossomo. Por exemplo, os cromossomos maiores terão um intervalo de busca diferente que o de cromossomos menores (mais detalhes sobre isto na seção 4.2.3.3).

Os de valores de *kIni* e *kFim* indicam um valor percentual em relação ao comprimento do cromossomo para definir a posição inicial e final de busca pelo centrômero. Por exemplo, um cromossomo de comprimento 30 (posições de 0 a 29), com *kIni = 20* e *kFim = 90* indica que a busca é feita entre as posições 5 e 26, e claro, nas duas direções do cromossomo. A

Figura 4.2 abaixo mostra o intervalo de busca de um cromossomo 1 em certo método de detecção do centrômero:

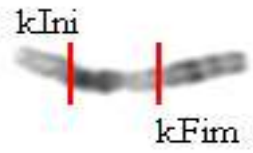


Figura 4.2: intervalo de verificação do centrômero

A definição destes intervalos é importante visto que cromossomos maiores têm a tendência de ser metacêntricos e submetacêntricos em sua maioria, logo, a busca deve ser feita nas regiões mais interiores do cromossomo. Já cromossomos menores tendem a ser acrocêntricos, logo, deve-se buscar pelo centrômero também nas extremidades. Se um método qualquer efetua a busca pelo centrômero ao decorrer de todo o comprimento do cromossomo, a probabilidade de erro é grande visto que inevitavelmente a maioria dos cromossomos apresenta um estreitamento nos seus extremos e, desta forma, classificando-os erroneamente como acrocêntricos. Assim, primeiramente é verificado qual o comprimento do cromossomo, e a partir daí são definidos os valores de *kIni* e *kFim*, iniciando a busca pelo centrômero somente neste intervalo do eixo longitudinal.

4.2.3.3. Intervalo de comprimento

Em cada um dos métodos, os cromossomos são divididos em grupos de acordo com o seu comprimento. Estes grupos são definidos como intervalos de comprimentos. Estes intervalos são utilizados como filtros nos métodos de detecção do centrômero para a determinação das razões cromossômicas (*r*) e dos intervalos de busca (*k*). Ou seja, primeiramente verifica-se o comprimento do cromossomo, e dependendo do intervalo de comprimento que ele pertencer, serão definidos os pesos das variáveis de razão (*r*) e intervalo de busca (*k*).

Foram realizados diversos testes na base de imagens com o intuito de definir quais são e a quantidade de intervalos de comprimento a ser utilizada nos métodos. Porém, devido ao fato do grande número de imagens da base de dados utilizada, e principalmente da quantidade

de variáveis envolvidas e relacionadas umas com as outras, é impossível determinar os valores ideais, logo, optou-se pela busca de um máximo local para cada método em torno de todas as possibilidades, os quais apresentaram as maiores taxas de acertos encontradas.

Ao final desta etapa de ajustes, definiu-se que os melhores resultados são obtidos ao dividir os cromossomos em 4 grupos de comprimentos (considerando que cada cromossomo tem um comprimento proporcional em relação ao seu cariótipo que varia de 0 a 100). Por exemplo, em determinado método, os intervalos de comprimento podem ser definidos como:

- Menor ou igual a 100 e maior que 57 (sendo 57 denominado *tam1*);
- Menor ou igual a 57 e maior que 46 (sendo 46 denominado *tam2*);
- Menor ou igual a 46 e maior que 40 (sendo 40 denominado *tam3*);
- Menor ou igual a 40;

Assim, cada método terá intervalos de comprimento diferente, e para cada intervalo, valores de *kIni*, *kFim*, *rIni* e *rFim* diferentes. Mais detalhes sobre a determinação destes pesos serão apresentados na seção 4.5.

4.2.3.4. Outras informações extraídas

Além das informações descritas anteriormente, algumas outras informações também são retiradas de forma a auxiliar a análise da forma do cromossomo e a detecção do centrômero do mesmo:

- Área do cromossomo: a área é definida simplesmente como a quantidade de píxeis brancos da máscara do cromossomo;
- Comprimento do cromossomo: é o comprimento do cromossomo em número de píxeis.

4.3. Métodos de detecção do centrômero desenvolvidos

Neste trabalho foram desenvolvidos cinco métodos para a geração do perfil da forma/projeção horizontal (*Shape Profile*) e para a detecção do centrômero, cada um utilizando técnicas diferentes. Estes métodos funcionam de forma independente, ou seja, é possível utilizar somente um dos métodos como forma de detecção do centrômero em um processo de identificação de cromossomos. Apesar disso, neste trabalho, além de ser possível utilizar os mesmos de forma independente, foram propostos 2 algoritmos que utilizam estes métodos em conjunto como forma de detecção do centrômero.

Os dois algoritmos irão aplicar determinado método de acordo com o comprimento do cromossomo. O primeiro algoritmo irá verificar o comprimento proporcional do cromossomo, e então aplicar somente um método de detecção do centrômero. Já o segundo, de acordo com o comprimento proporcional do cromossomo, diversos métodos são aplicados (de 1 a 5), verificando qual classificação ocorreu mais vezes (metacêntrico, submetacêntrico ou acrocêntrico).

Os métodos desenvolvidos envolvem principalmente a utilização da largura do cromossomo em cada ponto do esqueleto e da variação dos níveis de cinza nas linhas transversais que cortam cada um destes pontos. Os métodos desenvolvidos serão apresentados nas seções a seguir.

4.3.1 Método da linha perpendicular com níveis de cinza

Este método visa buscar o centrômero utilizando dois critérios: a largura do cromossomo em cada ponto do esqueleto (definida através de uma linha perpendicular que corta cada ponto) e da variação dos níveis de cinza desta linha.

Primeiramente devem ser definidas as linhas perpendiculares que cortam cada ponto do esqueleto. Para a definição destas linhas, é calculado o coeficiente angular da reta que cruza 2 pontos do esqueleto (a). Esses dois pontos são separados por um ponto do esqueleto, ou seja, a iteração para o cálculo do coeficiente angular é feita em incrementos de 2 píxeis.

Ao definir o coeficiente angular desta reta, é possível determinar o coeficiente angular da reta perpendicular (ou coeficiente angular inverso, *aInv*), o qual é definido como:

$$aInv = -\frac{1}{a}$$

Equação 13

Com o coeficiente angular da reta perpendicular, é possível determinar os pontos desta reta perpendicular, e desta forma, calcular a distância entre seus extremos (sendo possível assim a determinação da largura do cromossomo em todos os pontos do esqueleto).

Além disto, este método também utiliza a informação a respeito dos níveis de cinza no decorrer das linhas perpendiculares que cortam cada um destes pontos. Conforme dito na seção 2.1.3, os centrômeros se caracterizam por estarem em uma região mais estreita, porém, outro fator importante é de que esta região apresenta níveis de cinza mais claros. Portanto, este método utiliza também este critério para a definição do centrômero além da largura. Desta forma, para cada ponto do esqueleto, é definido um peso de acordo com a seguinte equação:

$$peso = \frac{\sum_{i=1}^{n}[(255 - g_i) \,.\, (d_i)^3]}{\sum_{i=1}^{n} d_i}$$

Equação 14

Ou seja, em cada ponto do esqueleto, aplica-se esta equação, sendo g_i o nível de cinza em determinado ponto *i* da linha perpendicular, e d_i a distância deste ponto *i* para o ponto do esqueleto que está sendo definido o peso. Esta equação é baseada na utilizada por Stanley et al. (1996), Piper e Granum (1989) e Wang et al. (2008). Nota-se que a distância euclidiana utilizada não se refere a distância dos pontos extremos da linha perpendicular (que definiria a largura), mas sim referente a um somatório das distâncias de cada ponto *i* da linha perpendicular em relação ao ponto do esqueleto em questão. Como nas imagens utilizadas neste trabalho o nível de cinza de valor 255 corresponde à cor branca, a equação utiliza a diferença $255\text{-}g_i$, indicando que quando mais claro for a cor da banda, menor o valor do peso, e logo, maiores as chances do centrômero ser encontrado naquela posição do esqueleto.

Uma versão simplificada deste método também foi implementada, esta utilizando somente a informação referente à distância, baseando-se na idéia apresentada por Moradi

(MORADI et al., 2003). Ou seja, para cada ponto do esqueleto, simplesmente calculava-se a distância entre os pontos mais extremos da sua linha perpendicular, e então definia como sendo a posição do centrômero aquele ponto do esqueleto com a linha de menor largura. Conforme será apresentado na seção 5.1, as taxas de acertos deste método utilizando somente a distância foi bastante baixa em relação à versão que também faz uso dos níveis de cinza. Portanto, a versão que utiliza somente a distância foi descartada, passando a ser utilizada a versão que aplica também os níveis de cinza como critério na definição dos pesos.

Os valores de *k*, *r*, e os intervalos de comprimento para a maior taxa de acerto deste método são apresentados na Tabela 4.1. Os valores de *k*, *r*, e os intervalos de comprimento para este método ao ser aplicado dois algoritmos propostos é apresentado na Tabela 4.2.

Tabela 4.1: Valores de *k*, *r*, e os intervalos de comprimento para o método da linha perpendicular com níveis de cinza individualmente

Intervalo de comprimento	*kIni*	*kFim*	*rIni*	*rFim*
100 – 57	31	47	1.47	2.71
57 – 46	29	39	1	3.23
46 – 40	10	44	1	4.08
40 – 0	12	45	2.76	4.61

Tabela 4.2: Valores de *k*, *r*, e os intervalos de comprimento para o método da linha perpendicular com níveis de cinza nos algoritmos propostos

Intervalo de comprimento	*kIni*	*kFim*	*rIni*	*rFim*
100 – 54	31	41	1	2.58
54 – 46	26	36	1	3.27
46 – 39	10	38	1	3.16
39 – 0	12	48	2.88	3.04

4.3.2 Método da rosa-dos-ventos com níveis de cinza

O segundo método desenvolvido é uma extensão do método das linhas perpendiculares. Este método foi desenvolvido devido ao fato de que a reta gerada pelo método das linhas perpendiculares muitas vezes não é de fato satisfatoriamente perpendicular aos dois pontos levados em consideração; logo, a largura do cromossomo naquele ponto do esqueleto seria definida erroneamente.

Para superar este problema desenvolveu-se este método, que ao invés de tentar definir para cada ponto do esqueleto a sua reta perpendicular através do coeficiente angular inverso, são geradas todas as retas da rosa-dos-ventos que cruzam cada ponto do esqueleto, sendo que cada reta tem como limite a borda da máscara. A Figura 4.3 demonstra o funcionamento do método para melhor entendimento. Para cada ponto do esqueleto, são geradas oito retas que cruzam o mesmo, seguindo a orientação dos pontos cardeais, colaterais e subcolaterais, daí o nome rosa-dos-ventos.

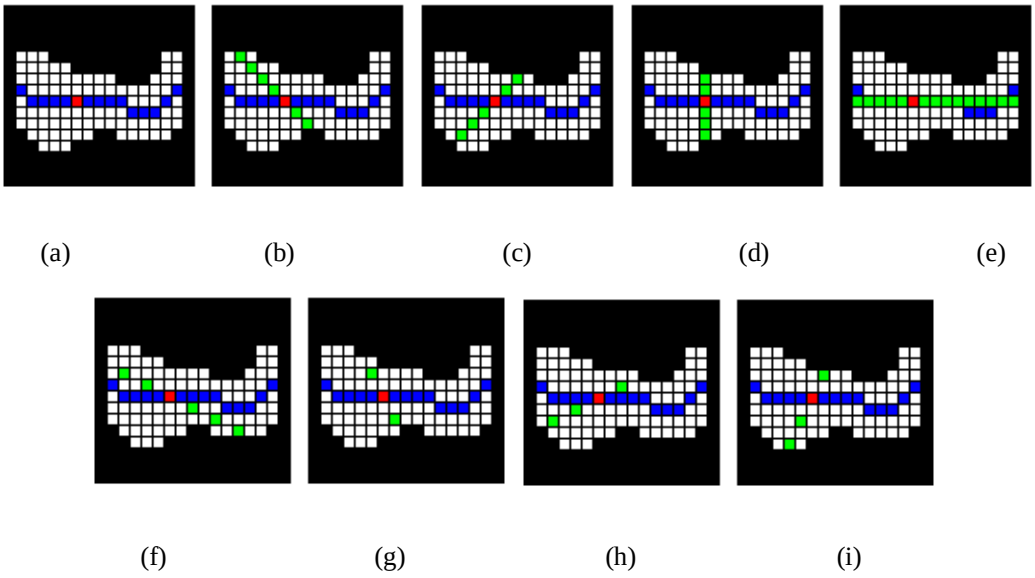


Figura 4.3 - Linhas verificadas no método da rosa-dos-ventos. Ponto do esqueleto a ser verificado em (a), e as linhas seguindo os pontos cardeais (b) e (c), colaterais (d) e (e) e os pontos subcolaterais em (f), (g), (h) e (i).

Da mesma forma que o método da linha perpendicular, aplicou-se no método da rosa-dos-ventos a equação que considera também os níveis de cinza em cada uma das retas da

rosa-dos-ventos. Portanto, para definir os pesos em cada posição do esqueleto, aplica-se a equação 14 em cada uma das retas da rosa-dos-ventos da mesma forma que no método da linha perpendicular, e ao final, a reta da rosa-dos-ventos com o menor peso é considerado o peso daquele ponto do esqueleto. Por fim, ao verificar todos os pontos do esqueleto, o ponto que tiver o menor peso é considerado como sendo a posição do centrômero.

Semelhante ao método da linha perpendicular, também foi testada a idéia de Moradi et al. (2003), implementando-se este método utilizando somente a distância entre os extremos das linhas das rosas-dos-ventos. Apesar disso, conforme será apresentado na seção 5.1, os resultados também não foram bons, passando a ser necessária a aplicação do critério relacionado aos níveis de cinza. As Tabelas a seguir apresentam os resultados e os valores das variáveis *k* e *i* para este método.

Tabela 4.3: Valores de *k*, *r* e os intervalos de comprimento para o método da rosa-dos-ventos com níveis de cinza individualmente

Intervalo de comprimento	*kIni*	*kFim*	*rIni*	*rFim*
100 – 57	32	53	1.32	2.47
57 – 46	23	51	1	3.36
46 – 40	15	62	1	3.16
40 – 0	22	56	1.78	3.47

Os valores para *k*, *r* e dos intervalos de comprimento na utilização deste método nos algoritmos propostos é demonstrado na Tabela x:

Tabela 4.4: Valores de *k*, *r* e os intervalos de comprimento para o método da rosa-dos-ventos com níveis de cinza nos algoritmos propostos

Intervalo de comprimento	*kIni*	*kFim*	*rIni*	*rFim*
100 – 55	32	53	1.32	2.47
55 – 45	22	48	1	3.29
45 – 39	17	62	1	2.74
39 – 0	22	50	2	3.36

4.3.3 Método da rosa-dos-ventos com níveis de cinza por comprimento

Outras variações do método da rosa-dos-ventos com níveis de cinza foram desenvolvidas devido ao fato deste ter apresentado uma maior taxa de acertos que os demais (ver seção 5.2). Esta versão difere na forma como são definidos os valores de *r* e *k* de acordo com o comprimento do cromossomo.

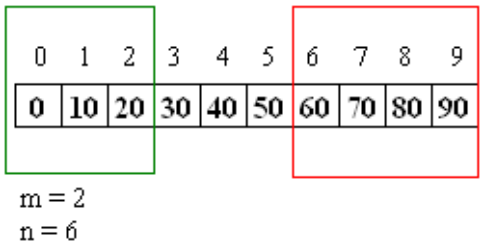
Nos métodos anteriores, os cromossomos são divididos em 4 intervalos de comprimento, e dependendo do intervalo certos pesos de *r* e *k* são aplicados. A diferença desta versão é de que ao invés de aplicar de acordo com um dos 4 intervalos, aplica valores de *k* e *r* diferentes para cada comprimento de 0 a 100 dos cromossomos. Os valores de *r* e *k* utilizados em cada comprimento foram definidos através de uma série de testes visando ajustar estes pesos. Na Tabela 7.6 do anexo F são apresentados os valores de *r* e *k* aplicados para cada comprimento de cromossomo.

4.3.4 Método da rosa-dos-ventos com níveis de cinza refletidos

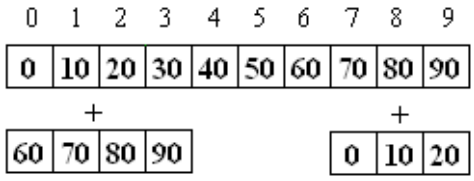
Outra versão do método da rosa-dos-ventos é baseada no desenvolvido por (PIPER e GRANUM, 1989). Esta técnica é desenvolvida de forma a evitar que cromossomos sejam classificados como acrocêntricos erroneamente, visto que, na maior parte das imagens, os cromossomos têm seus extremos mais estreitos. Mais detalhes sobre o problema abordado nesta técnica são apresentados na seção 3.1, e sua implementação neste trabalho é introduzida a seguir.

Para superar isso, é feita uma reflexão do cromossomo perto de seus extremos. Ou seja, sendo o perfil do cromossomo de comprimento *tam*, com as posições variando de *0* a *tam-1*, são feitas duas cópias: uma cópia do perfil da posição *0* até *m* chamada *copia1* e outra cópia da posição *n* até *tam-1* chamada *copia2*. Ao final, os valores de *copia1* são somados nas posições finais do perfil original (*tam-m-1* até *tam-1*) e os valores de *copia2* são somados nas posições iniciais do perfil original (*0* até *tam–n-1*). A Figura 4.4 abaixo mostra a idéia do método para melhor entendimento e a Figura 4.5 traz um exemplo de reflexão, sendo em 4.5.a

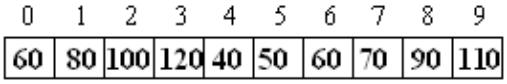
uma representação gráfica do perfil original e em 4.5.b do seu perfil refletido, sendo que o eixo *x* representa a posição e o eixo *y* o peso naquela posição do perfil, mostrando o funcionamento do algoritmo ao elevar os valores das regiões mais extremas do perfil.



(a)

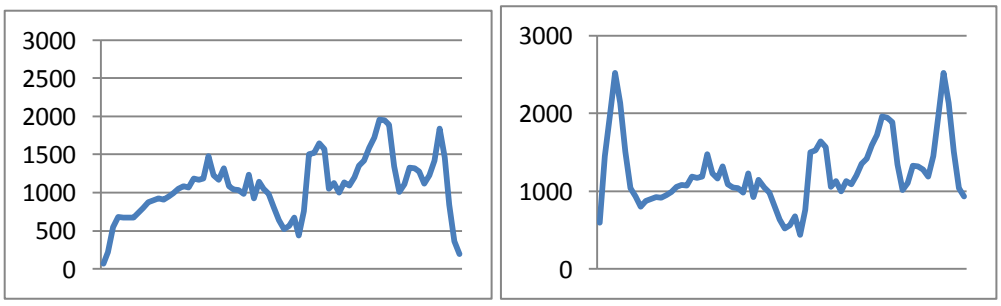


(b)



(c)

Figura 4.4 – Perfil original em (a), com *m=2* e *n=6*, mostrando as partes que serão somadas em (b) e por fim o perfil resultante em (c).



(a) (b)

Figura 4.5 – Exemplo de reflexão de um perfil de um cromossomo 1, sendo em (a) o perfil original e em (b) o perfil refletido.

Nota-se que a idéia do método é elevar os valores das posições mais extremas dos perfis e, desta forma, aumentar as chances de se encontrar o centrômero no meio. Como os cromossomos acrocêntricos tendem a ter um estreitamento maior que os cromossomos submetacêntricos e metacêntricos, o perfil resultante desta técnica na maior parte das vezes acaba não atrapalhando na identificação de cromossomos verdadeiramente acrocêntricos, mas sim evitando que cromossomos metacêntricos e submetacêntricos sejam classificados como acrocêntricos. As Tabelas 4.5 e 4.6 apresentam os valores de *k* e *r* aplicados a este método.

Tabela 4.5: Valores de *k*, *r* e dos intervalos de comprimento para o método da rosa-dos-ventos com níveis de cinza refletido individualmente

Intervalo de comprimento	*kIni*	*kFim*	*rIni*	*rFim*
100 – 57	32	53	1.3	2.47
57 – 46	23	51	1	3.36
46 – 40	15	62	1	3.16
40 – 0	24	56	1.78	3.18

Tabela 4.6: Valores de *k*, *r* e dos intervalos de comprimento para o método da rosa-dos-ventos com níveis de cinza refletidos nos algoritmos propostos.

Intervalo de comprimento	*kIni*	*kFim*	*rIni*	*rFim*
100 – 54	31	41	1	2.34
54 – 44	-	-	-	-
44 – 39	-	-	-	-
39 – 0	29	61	1.78	3.18

4.3.5 Método da rosa-dos-ventos com níveis de cinza médios

Outra versão um pouco diferente deste método também foi desenvolvida. Nesta versão, após ser gerado o perfil do cromossomo de acordo com o método da rosa-dos-ventos com níveis de cinza apresentado na seção 4.3.2, é gerado um novo perfil, sendo que o valor do peso de cada um dos pontos deste perfil será definido através da média dos pesos de 5 pontos

ao redor. Ou seja, Sendo um perfil da rosa-dos-ventos com níveis de cinza de acordo com a Figura 4.6.a, o perfil gerado para este método será conforme é mostrado na Figura 4.6.c.

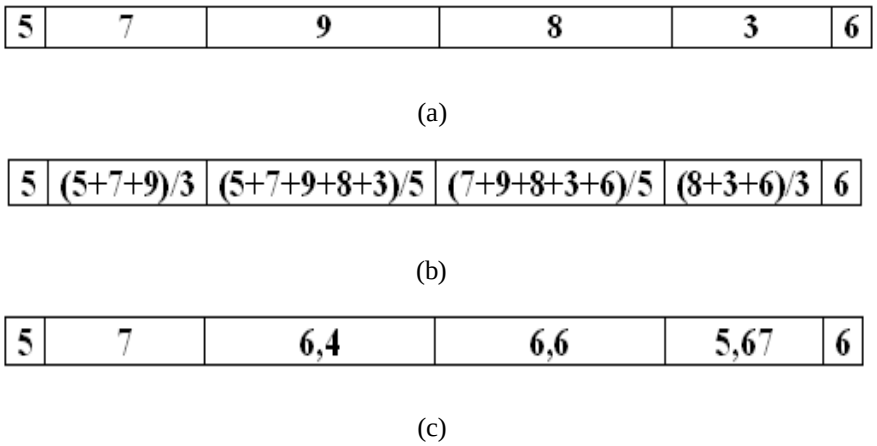


Figura 4.6 – Exemplo da média utilizada no método da rosa-dos-ventos com níveis de cinza médio

A escolha por utilizar 5 pontos também foi feita após uma série de testes, testando a média de 2 a 8 pontos, sendo que a partir de 8 pontos a taxa de acertos passou a cair bastante. As Tabelas 4.7 e 4.8 apresentam os valores de *k* e *r* utilizados neste método.

Tabela 4.7: Valores de *k, r,* e os intervalos de comprimento para o método da rosa-dos-ventos com níveis de cinza médios individualmente

Intervalo de comprimento	*kIni*	*kFim*	*rIni*	*rFim*
100 – 57	35	52	1.26	2.22
57 – 46	23	45	1	3.35
46 – 40	19	78	1	2.99
40 – 0	23	65	1.8	3.33

Tabela 4.8: Valores de *k*, *r*, e os intervalos de comprimento do método da rosa-dos-ventos com níveis de cinza médios nos algoritmos propostos

Intervalo de comprimento	*kIni*	*kFim*	*rIni*	*rFim*
100 – 58	37	55	1.38	1.63
58 – 46	-	-	-	-
46 – 35	20	44	1	3.22
35 – 0	-	-	-	-

4.4 Algoritmos propostos para detecção do cromossomo

Apesar de que a utilização de um só método possa trazer resultados interessantes, principalmente em relação ao método da rosa-dos-ventos com níveis de cinza que alcançou cerca de 89% de acertos (ver capítulo 5), decidiu-se por verificar os resultados obtidos ao tentar utilizar o melhor de cada método. Ou seja, cada método pode apresentar taxa de acertos melhores que os outros dependendo do comprimento dos cromossomos, portanto, o comprimento do cromossomo é utilizado como critério para a elaboração dos novos algoritmos. Estes algoritmos serão apresentados nas seções seguintes.

4.4.1 Primeiro algoritmo proposto para detecção do centrômero

O primeiro algoritmo proposto para detecção do centrômero irá aplicar, de acordo com o comprimento relativo do cromossomo, um dos cinco métodos de detecção descritos anteriormente. Nos testes executados para definição de qual método será aplicado como padrão em determinado comprimento de cromossomo, é verificado com qual dos cinco se obtém maiores acertos para aquele comprimento de cromossomo, e assim, o método que apresentar maiores acertos é tomando este método como fixo para aquele comprimento.

Desta forma, após a determinação dos métodos, o algoritmo funcionará de acordo com o fluxograma da Figura 4.7, que apresenta de forma simplificada o funcionamento deste algoritmo. Primeiramente é lido o comprimento do cromossomo de determinada imagem, e em seguida verifica qual algoritmo deve ser aplicado para aquele comprimento de

cromossomo, e por fim, de acordo com a posição encontrada por determinado método, é definida a classificação de acordo como o centrômero. No anexo C a Tabela 7.3 mostra qual método foi utilizado para determinado comprimento, sendo que os índices correspondentes a cada método são apresentados na Tabela 4.9.

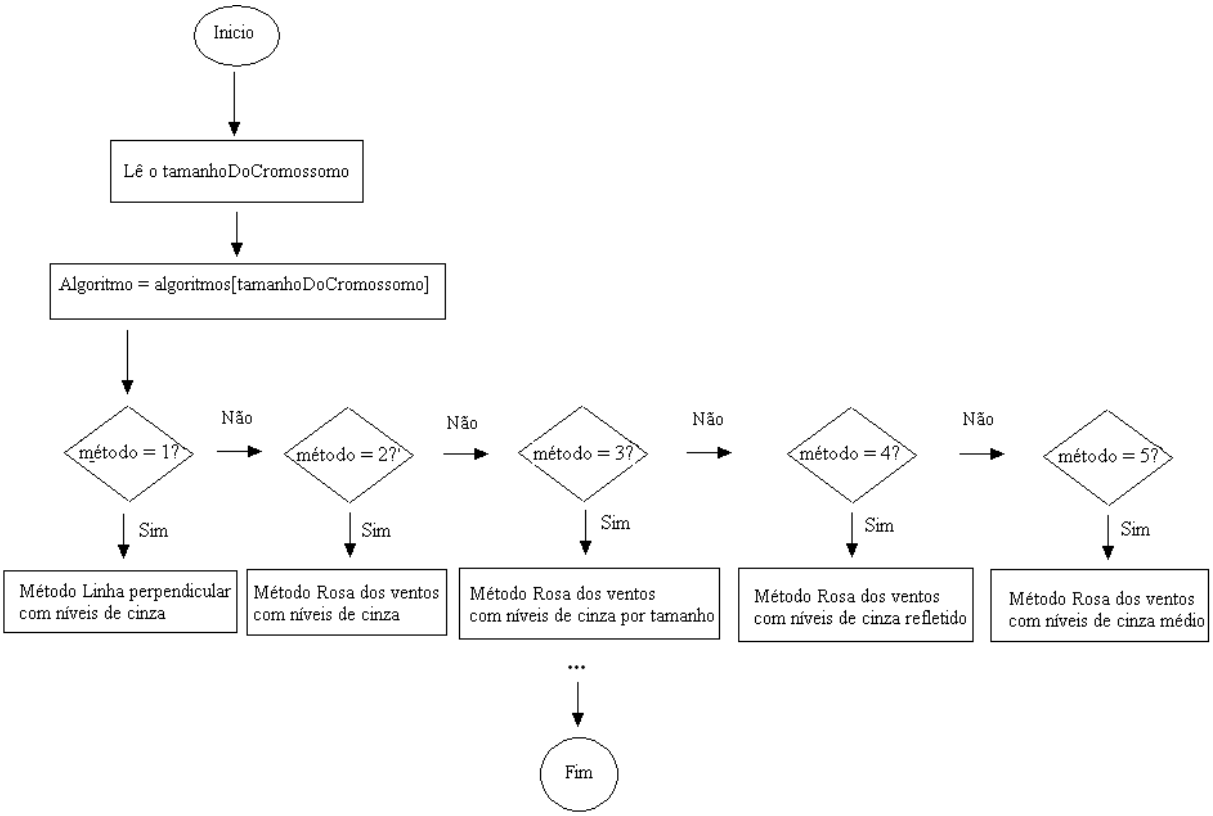


Figura 4.7 – Fluxograma simplificado do funcionamento do primeiro algoritmo proposto.

4.4.2 Segundo algoritmo proposto para detecção do cromossomo

Uma variação o primeiro algoritmo foi criada na tentativa de se buscar por melhores resultados. A idéia do segundo algoritmo é a aplicação de vários métodos para determinado comprimento de cromossomos, diferente do primeiro que aplica somente um dos métodos para determinado comprimento.

Para determinar quais métodos serão e quais não serão aplicados em determinado comprimento de cromossomo, novamente foram realizados testes na forma de um treinamento

para o sistema, verificando qual a melhor combinação de métodos deve ser utilizada para determinado comprimento. A definição das combinações de métodos aplicadas a cada comprimento é feita a partir dos seguintes passos: primeiramente classificam-se todos os cromossomos de acordo com todos os 5 métodos, e em seguida armazena-se o resultado em memória, para que não seja necessário executar os métodos toda a vez que se testar uma nova combinação. Além disso, define-se um índice para cada um dos 5 métodos, conforme mostra a Tabela 4.9.

Tabela 4.9: Índices dos métodos

Índice	**Método**
1	Linha perpendicular com níveis de cinza
2	Rosa-dos-ventos com níveis de cinza
3	Rosa-dos-ventos com níveis de cinza por comprimento
4	Rosa-dos-ventos com níveis de cinza refletido
5	Rosa-dos-ventos com níveis de cinza médio

Com a definição dos índices, inicia-se a segunda etapa, em que se verificam, para cada comprimento relativo dos cromossomo, todas as combinações possíveis de métodos, se eles serão utilizados ou não para determinado comprimento. Ou seja, para cada comprimento relativo de 0 a 100, testam-se todas as combinações possíveis de métodos. Como são cinco métodos, cada um pode assumir dois valores: ser utilizado naquele determinado comprimento (1) ou não (0). Desta forma, para cada comprimento, existem (2^5)-1 combinações, totalizando 31. Nota-se que o total de combinações não é 32, pois não se considera a combinação em que nenhum método é aplicado.

Para representar as combinações, pode-se utilizar tanto um valor inteiro de 1 a 31 como uma *string* de 0's e 1's. Por exemplo, se após testar todas as combinações para cromossomos de comprimento relativo 60 definiu-se que o melhor é utilizar o método da linha perpendicular com níveis de cinza (índice 1), método da rosa-dos-ventos com níveis de cinza (índice 3) e o método da rosa-dos-ventos com níveis de cinza refletido (índice 5), sua combinação pode ser expressa tanto pela *string* binária 10101 quanto pelo valor que ela representa, no caso 21. Na *string* binária, a posição dos valores 0's e 1's indica se o método

será aplicado ou não, conforme pode ser visto na Figura 4.8. A combinação definida nos testes para determinado comprimento relativo é então utilizada, após a etapa de treinamento, em todos os cromossomos do mesmo comprimento.

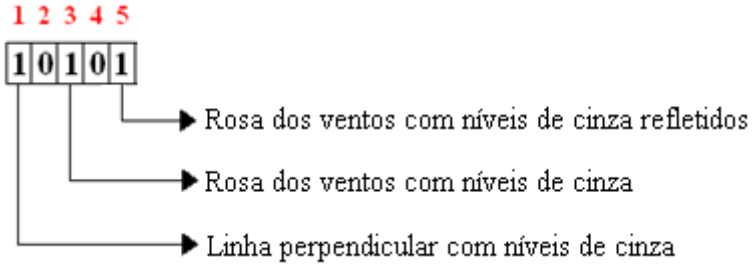


Figura 4.8: Exemplo de combinação de métodos utilizados em cromossomos de comprimento 60.

Para determinar como um cromossomo será classificado de acordo com certa combinação, tem-se o seguinte exemplo: supondo-se que para um cromossomo *k* qualquer de comprimento 60 a classificação do mesmo se dará pelas classificações feitas pelos métodos 1, 3 e 5. Ao executar estes métodos sobre a imagem do cromossomo *k* ele é classificado da seguinte forma:

- Método 1 - Método da linha perpendicular com níveis de cinza – metacêntrico;
- Método 3 - Método da rosa-dos-ventos com níveis de cinza – submetacêntrico;
- Método 5 - Método da rosa-dos-ventos com níveis de cinza refletido – metacêntrico;

Como se pode ver acima nas as classificações dadas por cada método ao cromossomo *k*, ele será classificado como metacêntrico, pois este ocorre em maior quantidade. Nos testes para definição das combinações utilizam-se os mesmos critérios, verificando qual combinação traz a maior taxa de acertos para cada comprimento de cromossomo. O funcionamento do algoritmo é semelhante ao apresentado pelo fluxograma da Figura 4.6, sendo que a diferença é que ao invés de aplicar somente 1 algoritmo, são verificados quais dos 5 algoritmos serão aplicados. As combinações de algoritmos utilizadas em cada comprimento de cromossomo são apresentadas no anexo D na Tabela 7.4.

A seção seguinte tem como objetivo uma análise dos resultados relacionados à detecção do centrômero, bem como uma discussão sobre a grande quantidade de testes realizados para se obter esta taxa de acertos.

4.5. Treinamento e ajuste dos pesos das variáveis

Uma das principais tarefas a ser realizada neste trabalho antes de tornar o sistema pronto para ser utilizado é o processo de ajuste dos pesos das variáveis que fazem parte de cada um dos métodos desenvolvidos. Conforme apresentado anteriormente, todos os métodos individuais seguem um padrão de variáveis a ser utilizada, porém, o processo de ajuste dos pesos que estas variáveis irão assumir na versão final do sistema é uma tarefa difícil e demorada, e que dificilmente se conseguirá um resultado ótimo devido a sua alta complexidade.

Conforme foi visto, os métodos individuais assumem valores diferentes para suas variáveis quando utilizados individualmente ou quando utilizados nos algoritmos propostos. Isso acontece porque quando se utiliza um método individualmente, ou seja, ao ser aplicado a todos os cromossomos da base de dados, os testes para definição dos pesos passam a buscar valores para estas variáveis com o intuito de abranger toda esta base, tentando alcançar um maior índice de acertos em geral. Quando um método é utilizado em conjunto com os demais nos algoritmos propostos, ele passará a ser aplicado somente a certo grupo de cromossomos (dependendo do comprimento do cromossomo) e, portanto, os pesos passam a assumir valores que sejam ideais para se obter uma alta taxa de acertos para aquele grupo no qual o método está sendo aplicado.

Ao observar um método individualmente, ele possui um grande conjunto de variáveis:

- 3 variáveis que determinam os 4 intervalos de comprimento → *tam1*, *tam2* e *tam3;*
- Para cada intervalo de comprimento, 2 variáveis determinando um intervalo de busca → *kIni[N]* e *kFim[N]*, sendo *N* mais uma vez o índice do intervalo de comprimento.
- Para cada intervalo de comprimento, 2 variáveis de razão → *rIni[N]* e *rFim[N]*, sendo *N* o índice do intervalo de comprimento no qual esta variável faz parte;

Ao fim, para cada método há um total de 19 variáveis. Como não se tem valores iniciais definidos para cada uma das variáveis, as possibilidades são muito altas. As variáveis de comprimento podem assumir valores de 0 a 100, sendo que obrigatoriamente *tam1* é maior que *tam2* e *tam2* é maior que *tam3*. As variáveis de intervalo de busca também podem assumir valores de 0 a 100, pois se refere a uma posição relativa de acordo com o comprimento do cromossomo, ou seja, cada par de variáveis *kIni* e *kFim* irá assumir valores de 0 a 100 definindo o intervalo de busca pelo centrômero, sendo que sempre *kFim* é maior que *kIni*. Por fim, as variáveis de razão que assumem valores maiores ou iguais a 1. Como *r* define a razão de comprimento do braço longo pelo braço curto, os testes mostraram que os valores de *rFim* dificilmente passam de 5 nos melhores resultados, portanto, assume-se um limite próximo a este valor. Isso foi definido devido ao fato de serem variáveis de ponto flutuante de duas casas decimais, e logo, o tempo gasto na busca por seus valores se torna maior a media em que se aumenta este intervalo de busca.

Um problema que surge é em relação à inicialização das variáveis. A primeira idéia era de se definir valores médios, como por exemplo, os seguintes valores para intervalos de comprimento: *tam3* igual a 25, *tam2* igual a 50 e *tam1* igual a 75. Logo, os intervalos de comprimento iniciais dos testes:

- Intervalo 1 – Menor ou igual a 100 e maior que 75;
- Intervalo 2 – Menor ou igual a 75 e maior que 50;
- Intervalo 3 – Menor ou igual a 50 e menor que 25;
- Intervalo 4 – Menor ou igual a 25 e maior que 1.

A partir destes valores, passava-se a buscar os valores de *k* e *r* para cada um dos intervalos e definia-se a taxa de acertos, e ao fim de cada etapa realizar pequenas variações nos valores dos intervalos de comprimento e então novamente buscar pelos valores de *k* e *r* ideais para aqueles intervalos de forma a refinar e melhorar a taxa de acertos.

Como se vê, os resultados ficariam de certa forma presos a aqueles intervalos de comprimento ou a pequenas variações dos mesmos, e então surge a necessidade de se buscar por uma alternativa em relação a isto. A alternativa utilizada foi a de reinicio aleatório, ou seja, ao invés de se determinar valores fixos e realizar pequenas variações nos mesmos de forma a refinar os resultados, inicializam-se aleatoriamente os valores dos pesos dos intervalos de comprimentos e então são buscados os valores de *k* e *r*. Ao invés de logo refinar

os primeiros resultados obtidos com os primeiros pesos definidos aleatoriamente, armazenava-se o resultado obtido em relação à taxa de acertos, e então novamente as variáveis de intervalos de comprimento eram reinicializadas. Ao fim de certa quantidade de ciclos, utilizava-se aquele intervalo em que se obteve uma maior taxa de acertos, e então se passava a refinar o mesmo de forma a melhorar os resultados.

Como foi visto, há um ciclo que sempre se repete ao reinicializar os valores dos intervalos de comprimento. Porém, após definir os pesos destas variáveis de comprimento, ainda existem as variáveis de intervalo de busca e da razão a serem definidas. Primeiramente, pode-se pensar que a o ideal seria seguir o seguinte ciclo do quadro 1 para definição dos pesos de *r* e *k*:

```
Para cada intervalo de comprimento N de 1 até 4:
    Para kIni[N] = 0 até kFim[N] faça:
        Para kFim[N] = kIni[N]+1 até 100 faça:
            Para rIni = 1 até rFim faça:
                Para rFim = rIni+0.01 até 5 faça:
                    /* executa ao método de detecção e
                       obtém-se a taxa de acertos e armazena a mesma */
```

Quadro 1 – Exemplo de possível ciclo para determinação dos pesos das variáveis

Pode-se observar que isso desprenderia muito tempo, pois a execução de cada método de detecção leva de 3 a 5 segundos para retornar o resultado da detecção. Logo, é totalmente inviável utilizar este tipo de abordagem para definição dos ciclos. Como os loops para as variáveis *r* são bem mais demorados devido ao fato de serem pontos flutuantes com duas casas decimais, a execução do primeiro ciclo para as variáveis *k* foi mantido, mas ao invés de se tentar buscar os valores de *r* para cada possibilidade de *k*, primeiramente inicializa-se os valores de *r* aleatoriamente, e então se busca pelos valores de *k* ideais para aqueles valores de *r*, e ao fim, refina-se os valores de *r* para aquele intervalo de busca, primeiro utilizando uma casa decimal, mas com uma variação maior, e em seguida, a partir deste resultado obtido, utilizam-se duas casas decimais, refinando o mesmo. Portanto, o ciclo (denominado ciclo 1) passou a ser conforme é demonstrado no quadro 2:

```
Para cada intervalo de comprimento N de 1 até 4:
    Inicializa-se aleatoriamente os valores de rIni[N]
    Inicializa-se aleatoriamente os valores de rFim[N]
Para cada intervalo de comprimento N de 1 até 4:
    /* busca os melhores valores de kIni e kFim para os valores de r inicializados
       antes */
    Para kIni[N] = 0 até kFim[N] faça:
        Para kFim[n] = kIni[n]+1 até 100 faça:
            /* executa ao método de detecção e
               Armazena-se a taxa de acertos e caso seja a maior obtida, armazena-
               se os valores de k em variáveis temporárias kIniTemp e kFimTemp */
    /* as variáveis k assumem os valores das variáveis temporárias, pois são as que
       obteve-se as melhores taxas de acertos até o momento */
    kIni[N] = kIniTemp;
    kFim[N] = kFimTemp;
    /* busca-se agora, para os valores de k definidos, os melhores valores de r. Nota-
       se que a inicialização aleatória no início do método é feita somente para se
       definir valores de k iniciais */
    Para rIni[N] = 1 até 3 incrementando 0.1 faça:
        Para rFim[N] = 2 até 4, incrementando 0.1 faça:
            /* executa ao método de detecção e
               Armazena-se a taxa de acertos e caso seja a maior obtida, armazena-
               se os valores de r em variáveis temporárias rIniTemp e rFimTemp */
    /* ao ter definido os valores de r com 1 casa decimal com a melhor taxa de
       acertos, refina-se r para duas casas decimais, de forma a aumentar ainda mais
       a taxa de acertos */
    Para rIni[N] = rIniTemp-1 até rIniTemp+1 incrementando 0.01 faça:
        Para rFim[N] = rFimTemp-1 até rFimTemp+1 incrementando 0.01 faça:
            /* executa ao método de detecção e
               Armazena-se a taxa de acertos e caso seja a maior obtida, armazena-
               se os valores de r em variáveis temporárias rIniTemp  e rFimTemp*/
    rIni[N] = rIniTemp;
    rFim[N] = rFimTemp;
```

Quadro 2 – Ciclo 1 para determinação dos pesos das variáveis.

Como se pode ver, os valores de *r* inicializados aleatoriamente no início de cada ciclo é feito somente para poder realizar uma busca por valores de *k*, pois após se definir os valores de *k*, novamente busca-se pelos valores de *r* para aqueles intervalos de *k*. Obviamente a variação de *r* na sua segunda busca será pouca, pois ao ter definido aleatoriamente seus valores no começo do algoritmo, os valores de *k* obtidos já estão próximos dos ideais. Este ciclo pode ser executado diversas vezes de forma a sempre refinar os resultados, sendo que leva em torno de 45 minutos para execução de cada um desses ciclos. Portanto, cada reinicio

aleatório das variáveis de intervalo de comprimento faz com que sejam necessários cerca de 45 minutos para definição dos pesos de *k* e *r*, isto para cada método. Como a variação dos intervalos de comprimento é enorme, pois são 3 variáveis, pode-se imaginar o tempo necessário para se obter bons resultados em cada um dos métodos individuais.

No caso dos algoritmos propostos, a idéia do processo de ajuste dos pesos é a mesma, porém, é preciso definir estes pesos para 5 métodos, cada um deles com seus intervalos de comprimento. Por isso ao invés de se realizar uma busca totalmente nova dos pesos para cada método no caso de serem utilizados nos algoritmos propostos, primeiramente é feita uma busca individual pelos pesos de cada método, buscando-se pela maior taxa de acertos ideal de cada método individualmente. Após certo número de testes, os valores dos pesos utilizados nos métodos individuais são passados para os pesos utilizados pelos mesmos nos algoritmos propostos.

Ou seja, ao invés de se realizar um reinicio aleatório nas varáveis de cada método nos algoritmos propostos, estas variáveis assumem os valores encontrados pelos seus métodos individuais ao serem realizados os testes individualmente. A partir daí é possível refinar os resultados executando o ciclo acima citado, mas ao invés de se executar o método individual, é executado o um dos algoritmos propostos (dependendo de qual algoritmo está sendo realizado o ajuste). A definição de qual método será utilizado em cada comprimento de cromossomo é a última etapa a ser realizada, porém, como faz parte de um novo ciclo, ela pode ser repetida infinitamente. Portanto, para os algoritmos propostos um novo ciclo é definido:

1. Executa-se o ciclo 1 para cada um dos métodos, individualmente, utilizando os métodos de detecção individuais, e assim, tentando obter as melhores taxas de acertos para cada método individualmente;
2. Verifica, para cada comprimento de 0 a 100, qual ou quais os melhores métodos a serem aplicados;
3. Refina o resultado, executando o ciclo 1 novamente, mas agora, ao invés de utilizar os métodos de detecção individual, utiliza os algoritmos propostos (mais uma vez, dependendo de qual algoritmo está sendo treinado).
4. Repete os passos 4 e 5 até que os resultado cheguem a um limite de variação.

Desta forma, assumindo-se que para cada um dos métodos já foram realizados, individualmente, diversos testes com o reinicio aleatório dos intervalos de comprimento e com a definição dos valores de *r* e *k*, bem como tendo-se uma taxa de acertos interessante (dependendo do algoritmo, mas geralmente maior que 80%) em cada método, a próxima etapa é realizar um ajuste com o intuito de refinar os pesos para que os mesmos sejam ideais para os algoritmos propostos. Se levar em conta que no na etapa de ajuste dos pesos do sistema cada método leva em torno de 45 minutos para definir os pesos de *r* e *k* para um conjunto de intervalos de comprimentos, e são no total 5 métodos, cada etapa do treinamento para refinamento dos pesos leva em torno de 225 minutos ou cerca de 3.25 horas, isso sem levar em conta os testes realizados na execução do ciclo 1 para cada método individualmente diversas vezes.

Percebe-se, portanto, a dificuldade que se tem em obter altas taxas de acertos (maiores que 90%), pois o tempo gasto é muito grande. Assim, o que acaba sendo priorizado é buscar por uma taxa de acertos relativamente boa num conjunto gigante de possibilidades, e então passar a fazer uma busca local pelos melhores resultados.

Quando este processo se tornar desgastado, novamente realiza-se uma busca pelo primeiro grande conjunto de possibilidades, e assim por diante. Vale lembrar que o tempo gasto é alto somente na etapa de ajuste dos pesos, pois os métodos e os algoritmos de detecção são executados milhares de vezes a fim de se buscar por pesos ideais para as variáveis. Desta forma, após a definição destes pesos, os mesmos são armazenados e assumidos toda vez que o sistema é executado, sendo que a partir daí tempo do processo de detecção do centrômero no sistema torna-se praticamente instantâneo.

O próximo capítulo irá abordar os resultados obtidos pelos métodos individuais e pelos algoritmos propostos de acordo com os pesos definidos em cada método/algoritmo.

5. RESULTADOS E ANÁLISE

Este capítulo visa apresentar os resultados obtidos nos 5 métodos individuais e também nos 2 algoritmos propostos desenvolvidos nesta dissertação, bem como realizar uma comparação dos resultados com as técnicas encontradas na literatura que foram apresentadas anteriormente na seção 3.1 e uma análise dos mesmos. As comparações serão feitas em relação aos algoritmos propostos, visto que estes foram os que apresentaram uma maior taxa de acertos na detecção do centrômero, bem como uma análise destes resultados. Em relação aos demais métodos serão feitas análises dos resultados obtidos individualmente neste trabalho.

5.1. Resultados dos métodos individuais baseados somente na distância

Conforme visto anteriormente, antes de se desenvolver os métodos da linha perpendicular com níveis de cinza, da rosa-dos-ventos com níveis de cinza e as variações deste último, foram implementadas versões dos dois primeiros baseando-se somente na distância, sem considerar os níveis de cinza, semelhante ao apresentado por Moradi et al. (2003). Porém, diferente do proposto por este autor, a técnica desenvolvida nesta dissertação já visava superar o problema de cromossomos tortos e dobrados, visto que este é um problema bastante comum, o que dificulta bastante sua implementação, tal como pode ser visto na Figura 2.10.

Ao assumir que todos os cromossomos estão retos ou com dobraduras irrelevantes, identificar o esqueleto do cromossomo torna-se uma tarefa bastante fácil, pois conforme observado em (KURTZ et al., 2008), posicionar o cromossomo “em pé” não gera grandes complicações, logo, o esqueleto do cromossomo passa a ser basicamente a linha central que corta a imagem. Além disso, em ambos os casos, para se calcular a distância entre os extremos de um ponto do esqueleto, é necessário a definição, em cada ponto do esqueleto, da linha perpendicular que o corta, o que em cromossomos retos é uma tarefa bastante simples, diferente de cromossomos tortos em que há a necessidade de se buscar os ângulos de

inclinação para geração dos coeficientes angulares inversos, e assim a geração da linha perpendicular.

Infelizmente, não é possível fazer uma comparação dos resultados obtidos nestes métodos com os resultados apresentados por Moradi et al. (2003), visto que este teve a posição do centrômero marcada nas suas imagens por um especialista, e os resultados apresentados em forma de distância destes pontos marcados. Isso não foi possível devido ao tamanho da base de dados, que no caso de Moradi, eram somente 87 imagens de cromossomos, e nesta dissertação são cerca de 5200 imagens de cromossomos.

Apesar destes dois métodos baseados na distância não terem sido utilizados nos algoritmos propostos, eles são importantes, pois possibilitam uma comparação em relação ao desenvolvido por Moradi et al. (2003) e também devido à evolução que os mesmos trouxeram, pois com o desenvolvimento destes métodos e principalmente de suas variações ficou confirmado que desenvolver métodos que utilizem somente à distância como critério de detecção do centrômero não era suficiente, porém importante, pois a distância é também utilizada na equação 14, e assim os métodos desenvolvidos a seguir passaram a ser baseados nesses. Desta forma, apresenta-se na Tabela 5.1 e 5.2 a taxa de acertos do método da linha perpendicular e do método da rosa-dos-ventos respectivamente, considerando somente a informação da distância.

Tabela 5.1: Taxa de acertos do método da linha perpendicular individualmente

Tipo	Acertos	Taxa de acertos (%)
Metacêntricos	663	58,83%
Submetacêntricos	2390	83,39%
Acrocêntricos	982	81,02%
Acertos em geral	**4035**	**77,54%**

Tabela 5.2: Taxa de acertos do método da rosa-dos-ventos individualmente

Tipo	Acertos	Taxa de acertos (%)
Metacêntricos	762	67,61%
Submetacêntricos	2529	88,24%
Acrocêntricos	1060	87,46%
Acertos em geral	**4351**	**83,61%**

5.2. Resultados dos métodos individuais baseados nos níveis de cinza

Os métodos em que se obtiveram as maiores taxas de acerto são aqueles baseados não somente na geração de um perfil levando em conta somente distância de um extremo ao outro de cada linha perpendicular que corta cada ponto do esqueleto, mas sim considerando também os níveis de cinza de cada ponto desta linha perpendicular, baseando-se no que foi desenvolvido por Piper e Granum (1989) e Wang (2008). No caso desta dissertação, algumas modificações e melhoras foram elaboradas a esta técnica, conforme foi visto na seção 4.3.4, sendo que a partir da idéia original da utilização da equação 11, diversas outras possibilidades foram estudadas principalmente envolvendo em alterações na equação e na forma em como o perfil dos cromossomos é gerado a partir desta equação.

Os primeiros métodos baseados nos níveis de cinza desenvolvidos são o da linha perpendicular com níveis de cinza e o da rosa-dos-ventos com níveis de cinza. Ambos têm a idéia de se buscar pela linha perpendicular em cada ponto do esqueleto e então aplicar a equação 14 em cada linha. Conforme visto anteriormente na seção 5.1, o método da rosa-dos-ventos obteve uma maior taxa de acertos em relação ao método da linha perpendicular.

A aplicação dos níveis de cinza manteve este padrão, aumentando a taxa de acerto em ambos os métodos (cerca de 6% para ambos os métodos), porém, o método da rosa-dos-ventos com níveis de cinza obteve uma taxa de acertos consideravelmente maior em relação ao método da linha perpendicular com níveis de cinza (cerca de 6% maior), conforme pode ser visto nas Tabelas 5.3 e 5.4. As Figuras 5.1 e 5.2 trazem, respectivamente, a taxa de acertos para cada tipo de cromossomo dos métodos da linha perpendicular com níveis de cinza e da rosa-dos-ventos com níveis de cinza.

Tabela 5.3: Taxa de acertos do método da linha perpendicular com níveis de cinza individualmente

Tipo	Acertos	Taxa de acertos (%)
Metacêntricos	829	73,56%
Submetacêntricos	2529	88,24%
Acrocêntricos	1011	83,42%
Acertos em geral	**4369**	**83,95%**

Tabela 5.4: Taxa de acertos do método da rosa-dos-ventos com níveis de cinza individualmente

Tipo	Acertos	Taxa de acertos (%)
Metacêntricos	904	80,21%
Submetacêntricos	2725	95,08%
Acrocêntricos	1026	84,65%
Acertos em geral	**4655**	**89,45%**

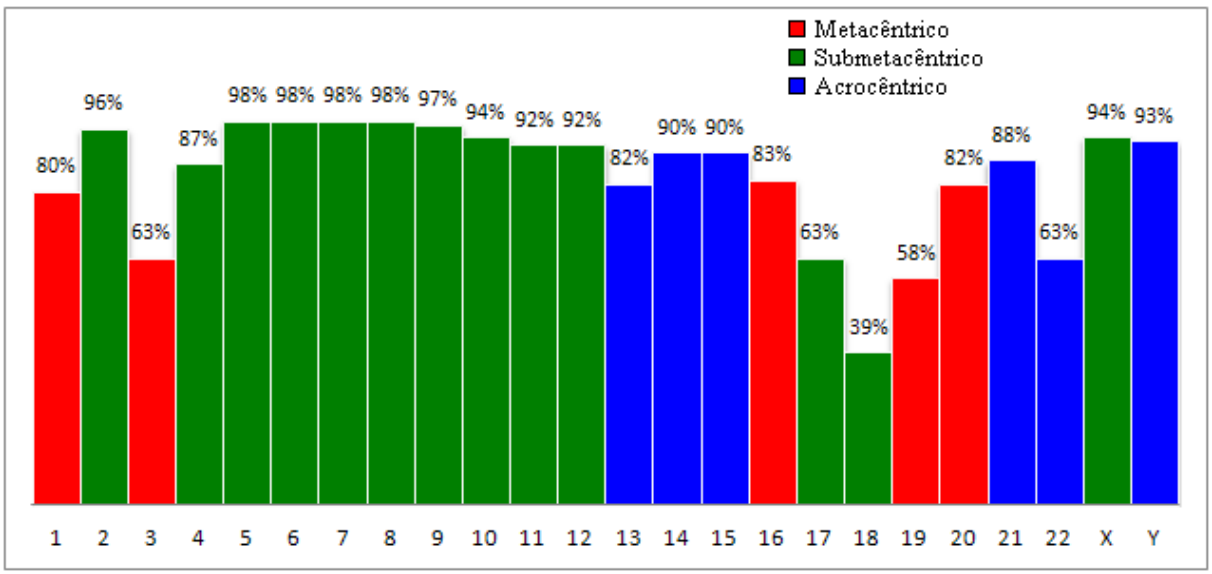


Figura 5.1 - Taxa de acertos para cada tipo de cromossomo do método da linha perpendicular com níveis de cinza.

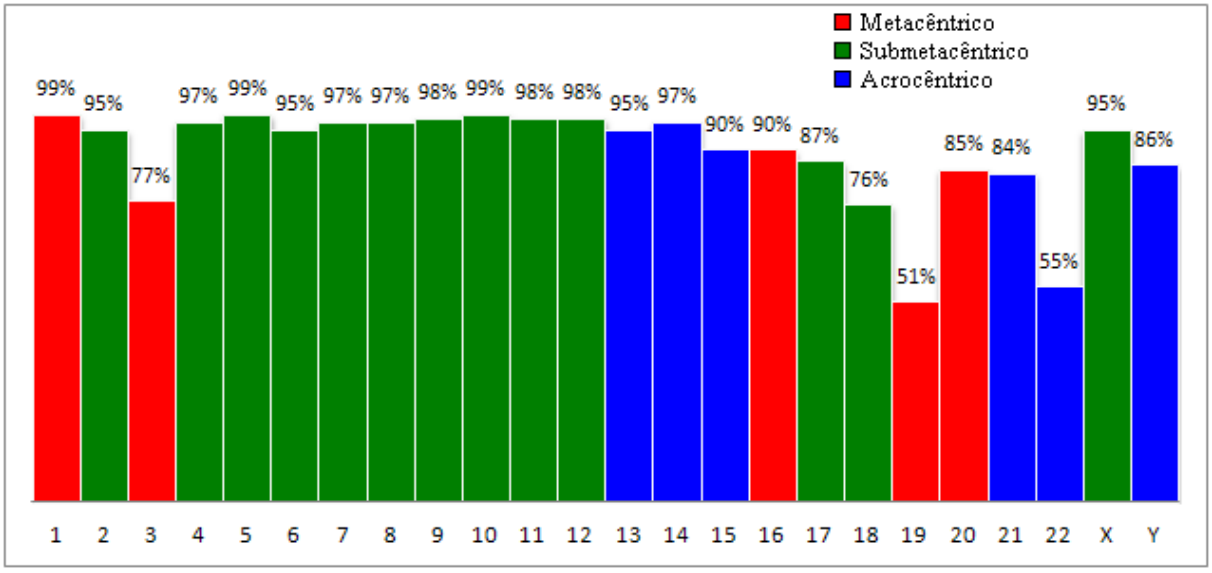


Figura 5.2 – Taxa de acertos para cada tipo de cromossomo do método da rosa-dos-ventos com níveis de cinza.

A partir daí, percebeu-se que o método da rosa-dos-ventos com sua variação utilizando os níveis de cinza era o método que valia a pena buscar por mais opções e começar um estudo mais aprofundado do mesmo, de forma a aumentar sua taxa de acerto. As principais modificações que ocorreram no decorrer deste trabalho com foco no aumento da taxa de acertos foram principalmente em relação à realização de novos testes exaustivos na busca por pesos ideais para este método como também a criação de novos métodos, buscando tanto aumentar a taxa de acertos em geral como a taxa de acertos em grupos específicos de cromossomos.

Para explicar melhor esta evolução, a taxa de acertos em geral da primeira versão do método da rosa-dos-ventos com níveis de cinza era de somente 81.4%, que, neste caso, ao invés de utilizar 4 intervalos de comprimento, ainda utilizava 2 (verificava somente se o cromossomo tinha um comprimento acima ou abaixo da média do seu cariótipo). Logo em seguida, optou-se por utilizar 3 intervalos de comprimentos, e assim houve a necessidade da criação e definição através de testes exaustivos das variáveis *k* e *r*, sendo que após estes testes a taxa de acertos chegou a 83%.

O próximo passou foi adotar a idéia de utilizar 4 intervalos de comprimentos. Além disso, alguns ajustes em relação aos filtros utilizados na geração do esqueleto e da máscara do cromossomo, bem como novos testes exaustivos fizeram com que os resultados passassem a chegar a 84.38% de acertos. A partir daí, as alterações no método da rosa-dos-ventos com

níveis de cinza passaram a ser somente na busca por novos valores dos pesos tanto em relação aos intervalos de comprimento quanto aos valores das variáveis *k* e *r*, e assim, obtendo a taxa de acertos apresentada na Tabela 5.4.

Modificações na forma como são gerados os perfis dos cromossomos levaram a criação de diferentes versões do método da rosa-dos-ventos. Conforme pode ser visto no parágrafo anterior, o aumento do número de intervalos de comprimento causou a ilusão de que quanto mais intervalos, maior a taxa de acertos. Dessa forma, decidiu-se então ao invés de utilizar somente 4 intervalos de comprimento e então gerar valores de *k* e *r* para estes intervalos, definir valores de *k* e *r* para cada comprimento de cromossomo, de 0 a 100. O resultado foi a criação do método da rosa-dos-ventos com níveis de cinza por comprimento, que ao contrário do que se pensava, utilizar diferentes valores de *r* e *k* para cada comprimento não aumentou a taxa de acertos em geral, porém, aumentou em grupos específicos de cromossomos. A Tabela 5.5 traz a taxa de acertos deste método, bem como a Figura 5.3 apresenta a taxa de acertos para cada tipo de cromossomo.

Para exemplificar o porquê de isso acontecer, observa-se a Figura 5.4 em que se apresenta uma situação hipotética de classificação de pontos vermelhos e azuis. Na Figura 5.4.a tem-se um exemplo de definição dos pesos de *r* e *k* por intervalo de comprimento e na Figura 5.4.b a aplicação individual para cada comprimento. Em ambas as Figuras 5.4.a e 5.4.b, para que todos os resultados fossem corretos, todos os pontos vermelhos deveriam estar acima da reta horizontal e os azuis abaixo, o que de fato não acontece, e desta forma deve-se tentar obter o maior número possível de acertos. Percebe-se que, ao utilizar intervalos de comprimento (5.4.a), os métodos priorizam uma quantidade de acertos em grupo, ou seja, da soma de acertos tanto de vermelhos quanto de azuis. Já no caso da Figura 5.4.b em que os valores de *r* e *k* são aplicados individualmente para cada comprimento, o método acaba priorizando somente uma das cores, no caso, no intervalo 1 as cores azuis, e nos intervalos 2 e 3 as cores vermelhas.

Portanto, devido a este problema, a taxa de acertos final ao utilizar pesos individuais para cada comprimento acaba sendo menor do que a utilização de intervalos, porém, para certos comprimentos, os acertos chegam a quase 100%, mas outros a taxa de acertos cai bastante.

Tabela 5.5: Taxa de acertos do método da rosa-dos-ventos com níveis de cinza por comprimento individualmente

Tipo	Acertos	Taxa de acertos (%)
Metacêntricos	615	54,57%
Submetacêntricos	2745	95,78%
Acrocêntricos	1169	96,45%
Acertos em geral	**4529**	**87,03%**

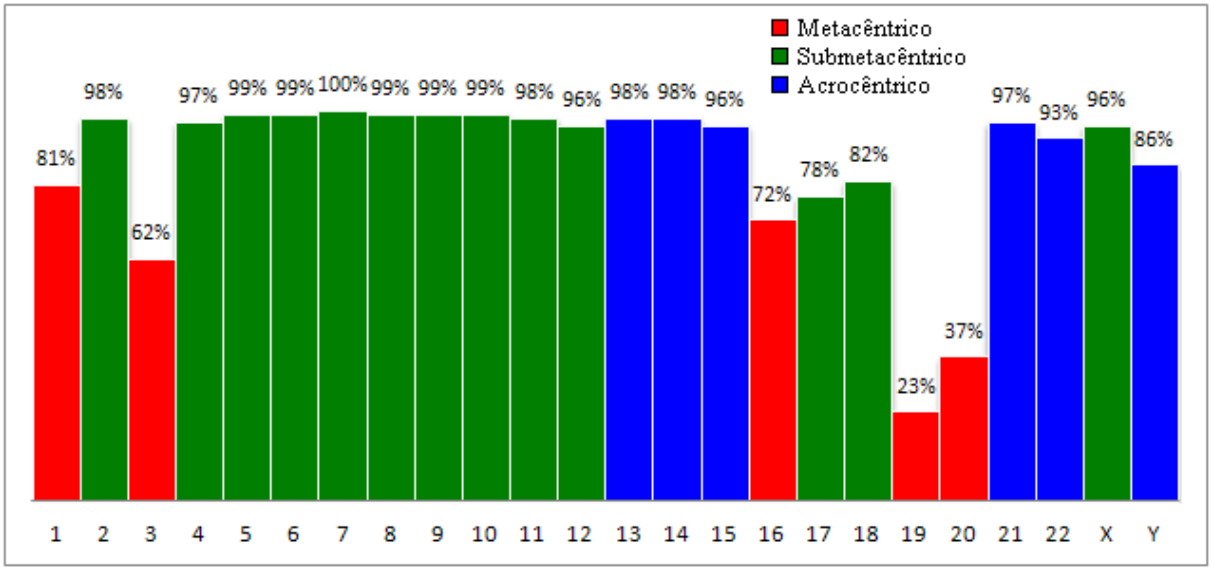
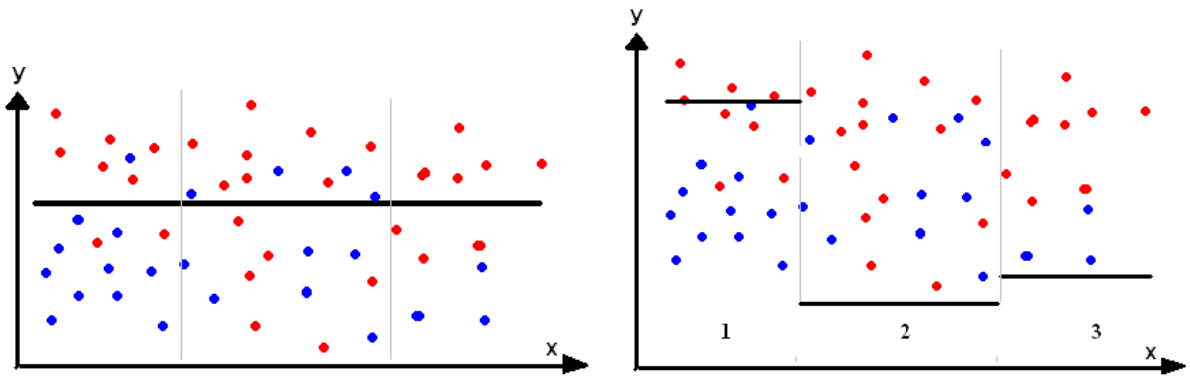


Figura 5.3 – Taxa de acertos para cada tipo de cromossomo do algoritmo da rosa-dos-ventos com níveis de cinza por comprimento.



(a) (b)

Figura 5.4 – Exemplo de aplicação hipotética dos valores de *r* e *k* por intervalo de comprimento em (a) e individual para cada comprimento em (b).

Assim, como se pode ver na Tabela 5.5, os acertos caíram para 87.03%, porém com um aumento significativo em cromossomos submetacêntricos e acrocêntricos. Além disso,

percebe-se na Figura 5.3 que em diversos tipos de cromossomos a taxa de acertos chegou a quase 100%, tornando viável a utilização do mesmo nos algoritmos propostos.

Como se pode ver na Tabela 5.5, a taxa de acertos para os cromossomos metacêntricos caiu bastante, pois os mesmos são os de menor quantidade em todo o cariótipo, que causou a queda na taxa de acertos em geral. Já os demais cromossomos os acertos subiram cerca de 10%, pois estes são encontrados em maior quantidade.

5.3. Resultados dos métodos individuais baseados nos níveis de cinza alternativos

Outras versões do método da rosa-dos-ventos com níveis de cinza foram desenvolvidas tal como os métodos da rosa-dos-ventos com níveis de cinza refletidos e da média, na tentativa de se buscar por uma maior taxa de acertos em geral. Porém, apesar de isto ter acontecido muito suavemente somente no método da rosa-dos-ventos com níveis de cinza refletidos conforme se pode ver nas Tabelas 5.6, a realização dos testes exaustivos mostrou que estes métodos são bastante importantes ao aplicá-los a cromossomos de certos comprimentos. Ou seja, apesar de não aumentarem muito a taxa de acertos em geral, os mesmos aumentaram a taxa de acertos em grupos específicos de cromossomos.

Em relação ao método da reflexão, como se pode ver, a taxa de acertos aumentou cerca de 3% para cromossomos metacêntricos, e diminuiu cerca de 2% em cromossomos acrocêntricos se comparado ao método da rosa-dos-ventos com níveis de cinza. Este método seguiu a proposta de Piper e Granum (1989), pois a idéia era evitar principalmente que cromossomos metacêntricos pequenos fossem classificados erroneamente como acrocêntricos. Devido aos resultados obtidos através deste método, foi possível obter mais acertos em determinados tipos de cromossomos nos quais não se obtivera anteriormente. Assim a utilização do mesmo foi bastante importante devido ao seu grande número de acertos em diferentes grupos de cromossomos. Em geral, a taxa de acertos aumentou 1.5%, tornando-o interessante ao ser utilizado em conjunto com os demais nos algoritmos propostos.

Tabela 5.6: Taxa de acertos do método da rosa-dos-ventos com níveis de cinza refletidos individualmente

Tipo	Acertos	Taxa de acertos (%)
Metacêntricos	939	83,32%
Submetacêntricos	2722	94,98%
Acrocêntricos	999	82,43%
Acertos em geral	**4660**	**89,55%**

Tabela 5.7: Taxa de acertos do método da rosa-dos-ventos com níveis de cinza médios individualmente

Tipo	Acertos	Taxa de acertos (%)
Metacêntricos	814	72,23%
Submetacêntricos	2530	88,28%
Acrocêntricos	1130	93,23%
Acertos em geral	**4474**	**85,97%**

O método da média foi feito na tentativa de suavizar os perfis dos cromossomos gerados. Este, apesar de ter diminuído a taxa de acertos em geral, teve um aumento em cromossomos acrocêntricos, se comparado ao método da rosa-dos-ventos com níveis de cinza. Desta forma, o aumento nos acertos em cromossomos acrocêntricos tornou-o interessante ao ser utilizado nos algoritmos propostos, sendo que o mesmo passou a ser aplicado a cromossomos cujo seu comprimento tende a ser relacionado a um cromossomo acrocêntrico. As próximas seções irão abordar os resultados obtidos nos algoritmos propostos bem como uma análise e comparação dos mesmos com os trabalhos encontrados na literatura.

5.5. Resultados dos algoritmos propostos

A Figura 5.4 traz um gráfico comparativo em relação às taxas de acertos obtidas nos algoritmos propostos e nos principais trabalhos relacionados.

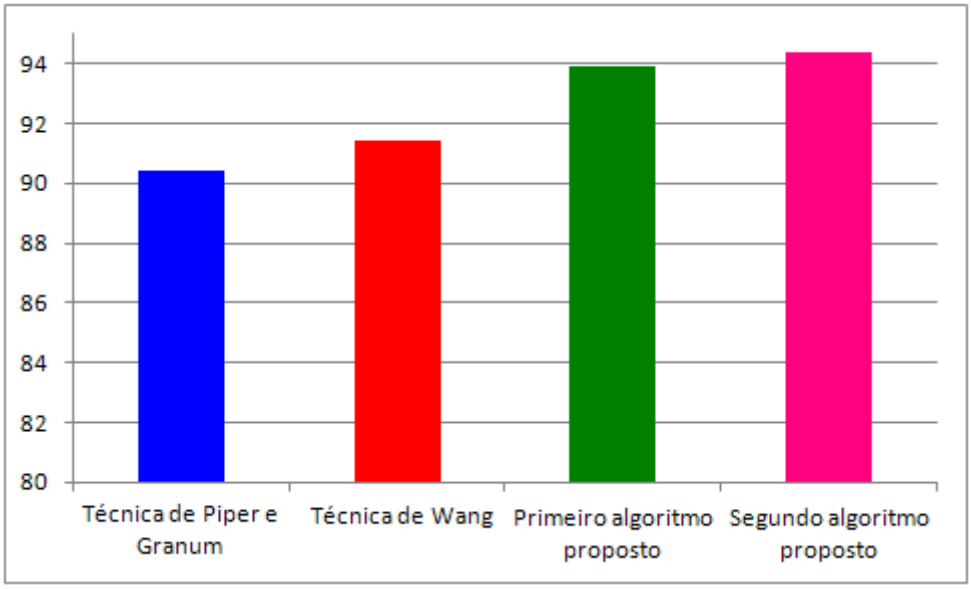


Figura 5.4 – Taxa de acertos percentual dos algoritmos propostos comparadas aos principais trabalhos relacionados.

Neste trabalho, o objetivo principal era o desenvolvimento de um único método que obtivesse uma alta taxa de acertos na detecção do centrômero. Porém, no decorrer do tempo, o foco deste trabalho passou a ser um estudo mais aprofundado das técnicas desenvolvidas na literatura assim como o desenvolvimento destas técnicas, bem como propor métodos alternativos baseados nos desenvolvidos na literatura a fim de verificar quais as melhores maneiras de se desenvolver um método para detecção do centrômero.

Dentre os métodos individuais desenvolvidos, o método da rosa-dos-ventos com níveis de cinza e sua variação com níveis de cinza refletido foram os que se obteve as maiores taxa de acertos, de cerca de 90%. Isto, comparado aos principais trabalhos encontrados na literatura é uma taxa de acertos relativamente alta, pois Wang et al. (2008) obteve uma taxa de 91.43% em geral e Piper e Granum (1989) de 90.4%. Provavelmente a realização de novos ajustes nos pesos das variáveis aumentaria a taxa de acertos destes métodos, porém, devido ao tempo gasto na realização dos mesmos, decidiu-se então partir para uma nova alternativa.

Os algoritmos propostos foram desenvolvidos com o objetivo de se buscar o melhor de cada uma dos métodos implementados, pois se notou que, conforme as taxas de acertos apresentadas nas Tabelas, muitas das técnicas traziam bons acertos em diferentes grupos de cromossomos (geralmente de acordo com o comprimento dos mesmos), porém, dificilmente apresentava um acerto mais uniforme e generalizado. Assim, decidiu-se o desenvolvimento de

um algoritmo que aplicasse um dos métodos de acordo com o comprimento do cromossomo e outro algoritmo que aplicasse de 1 a 5 métodos de acordo com o comprimento. As seções seguintes irão apresentar os resultados obtidos nos algoritmos propostos bem como uma comparação dos mesmos com os trabalhos relacionados.

5.5.1. Resultados dos algoritmos propostos

Desta forma, a taxa de acertos em geral dos algoritmos propostos e para os 3 tipos de classificação do centrômero são apresentadas na Tabela 5.8 e 5.9, e uma apresentação gráfica da taxa de acertos para cada tipo de cromossomo em cada algoritmo é apresentada nas Figuras 5.5 e 5.6.

Tabela 5.8: Taxa de acertos do primeiro algoritmo

Tipo	Acertos	Taxa de acertos (%)
Metacêntricos	960	85.18%
Submetacêntricos	2788	97,28%
Acrocêntricos	1138	93.89%
Acertos em geral	**4886**	**93.89%**

Tabela 5.9: Taxa de acertos do segundo algoritmo

Tipo	Acertos	Taxa de acertos (%)
Metacêntricos	968	85.89%
Submetacêntricos	2803	97.80%
Acrocêntricos	1140	94.06%
Acertos em geral	**4911**	**94.37%**

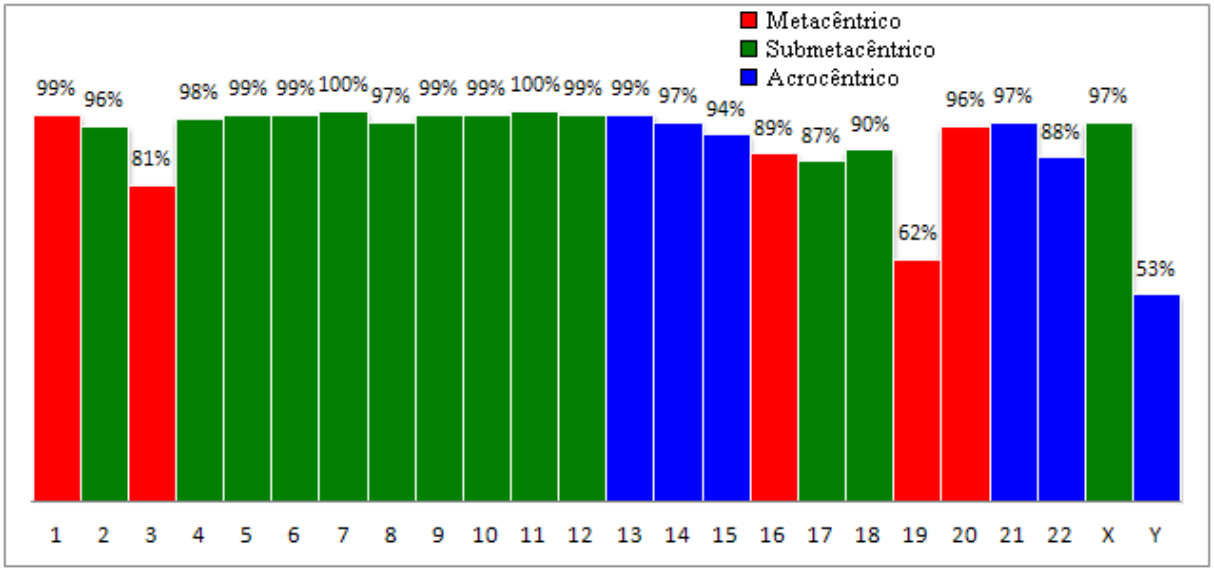


Figura 5.5 – Taxa de acertos para cada cromossomo do primeiro algoritmo proposto.

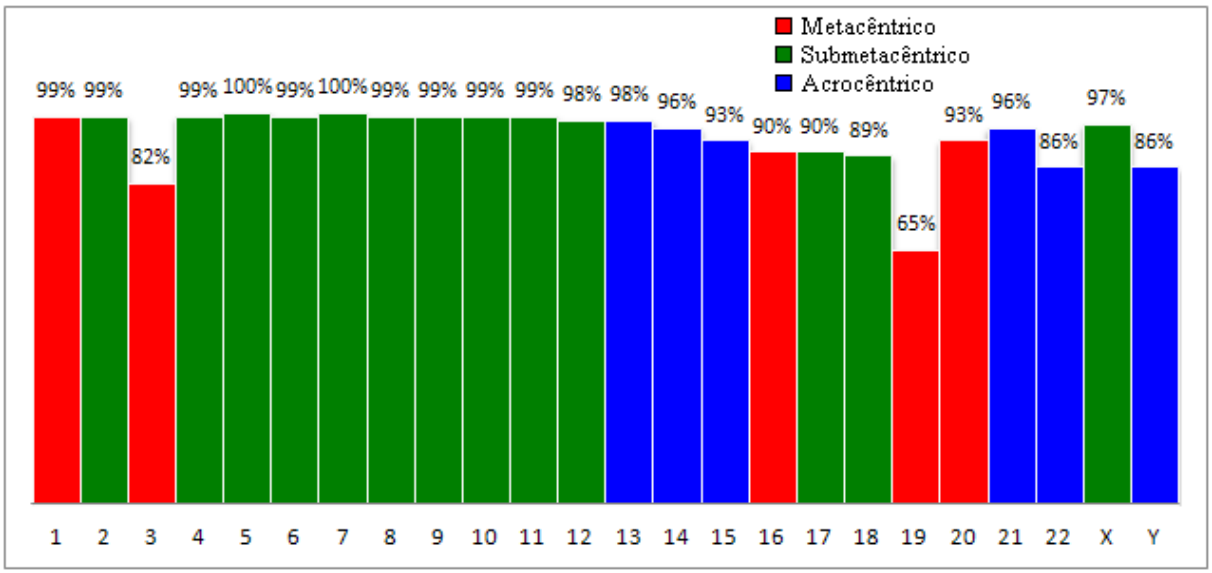


Figura 5.6 – Taxa de acertos para cada cromossomo do segundo algoritmo proposto.

Em suma, pode-se observar que em relação ao método da rosa-dos-ventos com níveis de cinza, o qual se obteve a maior taxa de acertos individualmente de cerca de 90%, a opção de utilizar determinados métodos para cada comprimento de cromossomo tornou-se interessante visto que os acertos chegaram a mais de 94% utilizando esta técnica.

O segundo algoritmo proposto apresenta resultados semelhantes ao primeiro, porém, alguns pontos devem ser observados. Pode-se ver um aumento importante na taxa de acertos

de cromossomos Y, que subiu de 58% a 86%, mostrando que para este cromossomo a utilização de diversos métodos como critério de detecção tem suas vantagens.

Obviamente a execução de novos testes exaustivos aumentaria ainda mais a taxa de acertos de ambos algoritmos, logo, resultados maiores ou iguais a 94.37% de acertos são possíveis. Porém, devido à dificuldade e ao tempo gasto para elaboração e execução destes testes, acabou tornando-se inviável esta prática devido ao tempo restante para elaboração desta dissertação, mas que não diminui sua validade. Mais detalhes sobre isso podem ser vistos na seção 4.6.

No decorrer deste trabalho, alguns resultados adicionais também foram gerados no com o intuito de se tentar analisar e entender os resultados obtidos, a fim de se buscar alternativas e melhoras nas técnicas desenvolvidas. A Tabela 5.10 apresenta a taxa de acertos para os dois algoritmos propostos em relação à classificação de Denver.

Tabela 5.10: Taxa de acertos em relação aos grupos de Denver dos algoritmos propostos.

	1º Algoritmo proposto	2º Algoritmo proposto
Grupo	Acertos (%)	Acertos (%)
A	92	93
B	98	99
C	99	99
D	97	96
E	89	90
F	79	79
G	89	91

Esta classificação apresentada na Tabela 5.10 reforça o que foi dito anteriormente. Em ambos os algoritmos, os grupos A, B, C e D são os grupos que abrangem cromossomos maiores, logo, a taxa de acertos para estes grupos é maior que as demais. Uma exceção ocorre no grupo A, pois o cromossomo 3 faz parte deste grupo, e devido a dificuldades em relação a este cromossomo que podem ser vistas na seção 5.5.2., diminui assim a quantidade de acertos. Os grupos E e F, por abrangerem cromossomos menores, tem uma taxa de acertos um pouco mais baixa, com exceção grupo G que também obteve uma alta taxa de acertos.

Outra forma de visualização observada é em relação à classificação dada pelos algoritmos propostos em cada um dos cromossomos. Ou seja, a Tabela 7.2 do anexo B mostra a porcentagem geral de classificações dada a cada um dos cromossomos, do 1 ao 22 e dos cromossomos X e Y, bem como a classificação correta do mesmo.

Por fim, a porcentagem geral de classificações dada para cada comprimento de cromossomo (relativo) é apresentada na Tabela 7.1 do anexo A, pois o comprimento é o que define qual ou quais métodos serão utilizados para detecção nos algoritmos propostos. As Figuras 5.7, 5.8 e 5.9 apresentam, respectivamente, a variação da quantidade geral (percentual) de cromossomos metacêntricos, submetacêntricos e acrocêntricos no decorrer do aumento do comprimento relativo dos mesmos, sendo que a Figura 5.7 está de acordo com a classificação do primeiro algoritmo proposto, a Figura 5.8 para o segundo algoritmo proposto e a Figura 5.9 de acordo com a classificação correta. Pode-se ver na Tabela 7.1 do anexo A que cromossomos menores tendem a ser acrocêntricos, e a medida que o comprimento aumenta, ocorre uma divisão entre acrocêntricos e metacêntricos. Em seguida, conforme se aumenta o comprimento, esta divisão passa a ser entre metacêntricos e submetacêntricos, sendo que a partir de certo comprimento a maior porcentagem passa a ser em submetacêntricos e por fim submetacêntricos e metacêntricos.

A Tabela 5.11 traz a correlação entre os dados dos gráficos das Figuras 5.7 e 5.8 ao serem comparado com a classificação correta, demonstrando que há um alto grau de correlação entre os mesmos, mostrando que a divisão dos cromossomos por comprimento é um bom critério de filtragem nos métodos de detecção do centrômero. Detalhes sobre o cálculo do coeficiente de correlação podem ser encontrados no anexo G.

Tabela 5.11: Correlação de acordo com o comprimento dos cromossomos entre os algoritmos propostos e a classificação correta.

Algoritmo	Metacêntricos	Submetacêntricos	Acrocêntricos
1º Alg. vs. correta	0,898	0,984	0,954
2º Alg. vs. correta	0,904	0,986	0,955

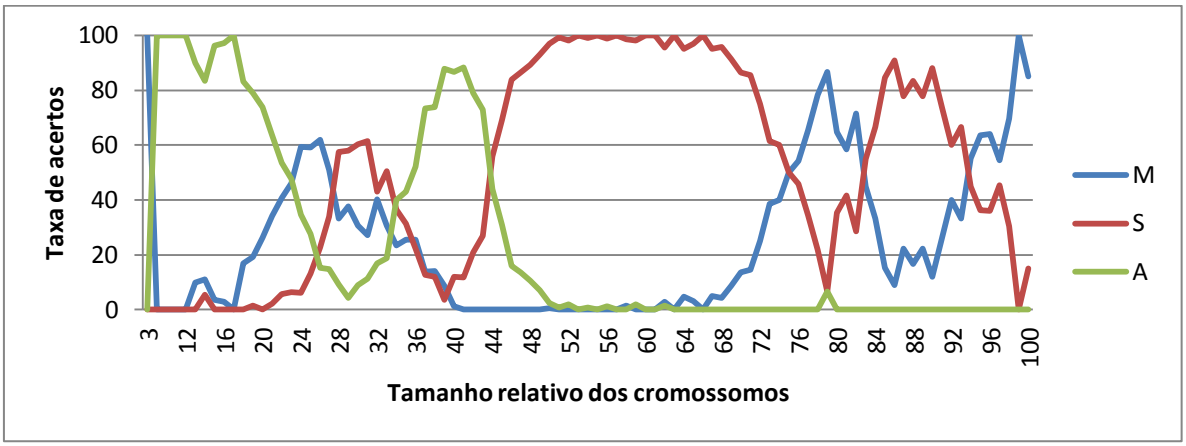


Figura 5.7 – Variação da quantidade (percentual) geral de cromossomos metacêntricos, submetacêntricos e acrocêntricos de acordo com a classificação dada pelo primeiro algoritmo proposto em relação ao seu comprimento relativo.

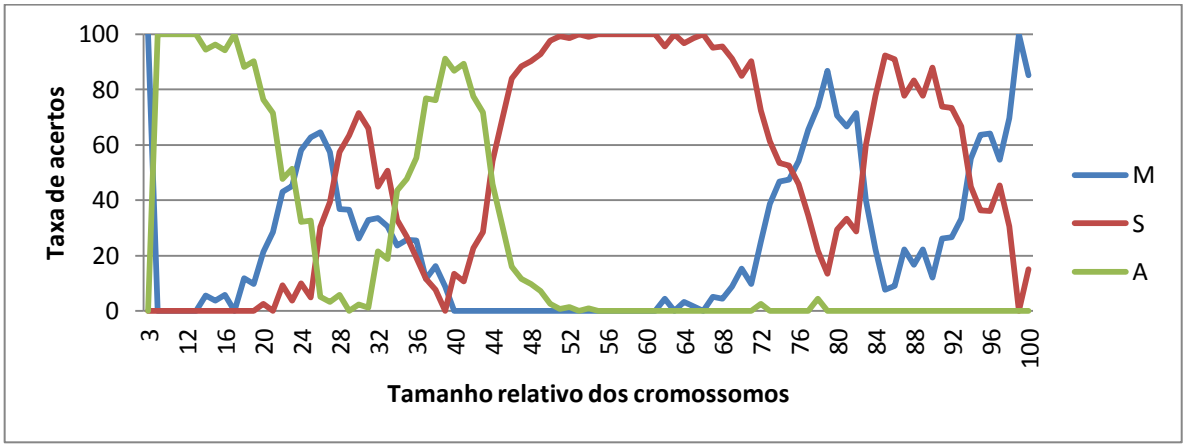


Figura 5.8 – Variação da quantidade (percentual) geral de cromossomos metacêntricos, submetacêntricos e acrocêntricos de acordo com a classificação dada pelo segundo algoritmo proposto em relação ao seu comprimento relativo.

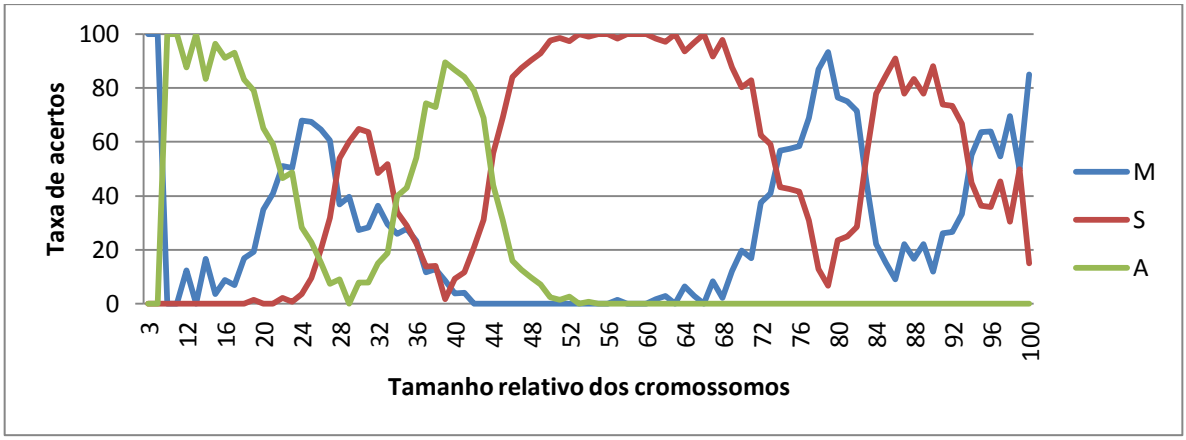


Figura 5.9 – Variação da quantidade (percentual) geral de cromossomos metacêntricos, submetacêntricos e acrocêntricos de acordo com a correta classificação em relação ao seu comprimento relativo.

5.5.2. Dificuldades encontradas

Como se pode ver na Tabela 7.2 do anexo B, os algoritmos propostos tendem a trazer um alto índice de acertos em cromossomos maiores, com exceção dos cromossomos 3. Isso se deve principalmente em relação à qualidade das imagens, pois não apresentam boas resoluções para que se possa realizar testes mais precisos (as imagens geralmente seguem uma resolução de cerca de 50 píxeis de largura por 100 de altura). Esta baixa resolução faz com que na medida em que o comprimento dos cromossomos diminui, o estreitamento causado pelo centrômero vai perdendo sua real forma, posição e nitidez, ou até mesmo confundindo com os estreitamentos que ocorrem nas pontas dos cromossomos.

A Figura 5.10 traz um exemplo disto, onde a baixa resolução do cromossomo 19 em 5.10.a fez com que o cromossomo se tornasse praticamente um círculo. Na Figura 5.10.b, um outro cromossomo 19, que apesar de ser metacêntrico, a posição do centrômero aparece deslocada devido a qualidade da imagem, e assim o algoritmo classificou-o erroneamente como submetacêntrico. A Figura 5.10.c mostra um cromossomo 20, sendo um exemplo de que, ao diminuir o comprimento dos cromossomos, eles passam a assumir uma forma mais uniforme, e assim gerando erros na classificação. O cromossomo 19 é um dos cromossomos que apresentam as menores taxas de acertos em ambos os algoritmos propostos, possivelmente pelo fato de este ser muito pequeno e também por ter um contraste baixo, o que dificulta a aplicação da equação 7.

(a) (b) (c)

Figura 5.10 – Posição do centrômero encontrada em cromossomos 19 (a, b) e 20 (c).

Além disso, os acertos em submetacêntricos e acrocêntricos em todos os casos foram maiores que os cromossomos metacêntricos, em que a taxa de acertos não passou de 85.89%. A resolução das imagens mais uma vez é o principal fator gerador desta baixa taxa de acertos, pois, como foi mostrada na Figura 5.10, principalmente em cromossomos menores, a posição

do centrômero de cromossomos metacêntricos perde sua nitidez, ou abrange uma região muito grande, fazendo com que se torne difícil a real definição da posição do centrômero, e assim causando um erro na classificação.

O cromossomo 3, apesar de ser um cromossomo de tamanho grande, trouxe uma taxa de acertos baixa se comparado a cromossomos de comprimento semelhante. A Figura 5.11 traz imagens de alguns cromossomos 3 e a posição do centrômero encontrada. Em geral, a posição do centrômero é marcada por uma região mais clara, com um estreitamento notável. Já em cromossomos 3 isso não acontece, pois a região branca tem um tamanho muito pequeno, ou muitas vezes nem aparece devido a resolução da imagem. Como os métodos desenvolvidos geralmente utilizam os níveis de cinza como critério de definição do centrômero, valorizando as regiões claras, isso se tornou um problema para cromossomos do tipo 3.

(a) (b) (c)

Figura 5.11 – Posição do centrômero encontrada em cromossomos 3. Em (a) e (b), a parte clara do centrômero é imperceptível, sendo que em (c) ela apresenta um tamanho bastante pequeno.

Assim, conforme é mostrado na Tabela 7.2 do anexo B, pode-se perceber que as maiores dispersões ocorrem nos cromossomos 3 e 19 em ambos os algoritmos e nos cromossomos Y no primeiro algoritmo proposto. O que se observa é que no caso dos cromossomos 3, que são metacêntricos, são classificados erroneamente em submetacêntricos, pois por geralmente serem cromossomos grandes, a busca pelo centrômero é feita mais no seu interior, logo, quando os erros ocorrem, é por que a possível posição do centrômero encontrada é ainda na região mais interior do cromossomo, porém não mais no meio, que seria o correto. Já o 19, que também é metacêntrico, é classificado como acrocêntrico.

Uma vez mais, o que explica isso é o comprimento do cromossomo, pois da mesma forma que, pelo fato de o cromossomo 3 ser um cromossomo grande, as chances de se encontrar o centrômero em sua região mais interior é maior; já o cromossomo 19, que devido

ao seu comprimento, busca-se o centrômero também nas extremidades, aumentando as chances de defini-lo erroneamente como acrocêntrico. Portanto, um padrão é observado:

- Cromossomos grandes metacêntricos quando classificados errados, geralmente é em submetacêntrico;
- Cromossomos pequenos metacêntricos quando classificados errados geralmente é em acrocêntrico;
- Cromossomos submetacêntricos e acrocêntricos quando classificados errados tendem a variar a classificação.

Uma questão que surge é em relação a metodologia proposta estar muito especializada somente nesta base de imagens. Como não foi possível de se ter acesso a outras bases de imagens, foram realizados alguns testes alternativos. Estes testes têm como objetivo retirar alguns cromossomos da base de imagens na hora de realizar a detecção do centrômero de cada um deles, e por fim, verificar a taxa de acertos.

A idéia é tentar mostrar que a metodologia proposta em relação aos métodos, aos algoritmos e aos pesos das variáveis podem ser aplicadas a qualquer base de imagens. Portanto, a Tabela 5.12 apresenta diversos testes aplicados ao segundo algoritmo proposto em que se retiraram certa quantidade de imagens da base de forma aleatória, e a partir daí são mostrados os resultados obtidos. É importante ressaltar que não são retirados cariótipos inteiros, mas somente certas imagens destes cariótipos de forma aleatória. A Figura 5.12 traz uma representação gráfica dos dados da Tabela 5.12.

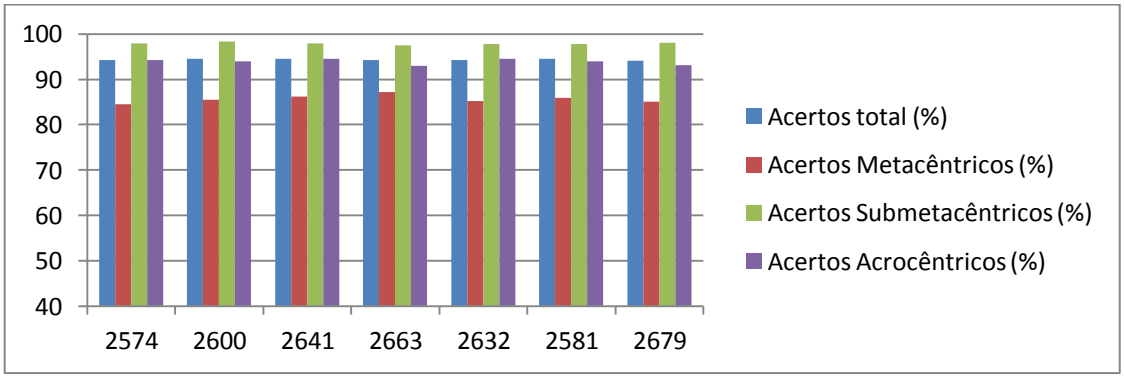


Figura 5.12 – Representação gráfica da Tabela 5.12 em que se retira uma certa quantidade aleatória de cromossomos.

Tabela 5.12: Taxa de acertos na detecção do centrômero do segundo algoritmo proposto ao retirar uma quantidade aleatória de cromossomos da base de imagens.

Retirados	Acertos total (%)	Acertos M (%)	Acertos S (%)	Acertos A (%)
2574	94,21	84,52	97,94	94,19
2600	94,54	85,51	98,38	93,95
2641	94,59	86,18	97,94	94,51
2663	94,25	87,22	97,53	92,96
2632	94,30	85,22	97,80	94,46
2581	94,46	85,98	97,76	93,90
2679	94,10	85,03	98,03	93,11
Média:	**94,35**	**85,66**	**97,91**	**93,87**
Desvio Padrão	**0,18**	**0,89**	**0,26**	**0,62**

Como se percebe na Figura 5.11 e na Tabela 5.12, ao retirar certa quantidade aleatória de cromossomos a taxa de acertos em geral e em cada classificação em relação ao centrômero continua seguindo um mesmo padrão, mostrando que possivelmente em outras bases de imagens do mesmo tipo este padrão de acertos continuará sem grandes diferenças. Como a base de imagens utilizada é dividida em 119 pastas, sendo que cada pasta é referente a uma metáfase/cariótipo, a Tabela 7.5 do anexo E mostra a quantidade de acertos em cada metáfase/cariótipo para o segundo algoritmo proposto, mostrando que raramente ocorre uma metáfase/cariótipo em que os acertos sejam menores que 90%, sendo que o desvio padrão obtido é de somente 4,19%.

5.5.3. Comparação dos resultados

A Tabela 5.13 traz um comparativo das técnicas com o desenvolvido por Wang et al. (2008) e Piper e Granum (1989) em relação aos acertos em cada tipo de cromossomo. Conforme dito anteriormente, apesar de o trabalho de Moradi et al. (2003) ser bastante citado na literatura não foi possível realizar uma comparação da taxa de acertos apresentada pelo mesmo devido ao seu formato, conforme visto na seção 3.1.3. Além disso, também não foi

possível realizar os testes com as mesmas bases de imagens utilizadas pelos autores por se tratarem de bases privadas.

Tabela 5.13: Comparação dos acertos dos algoritmos com as técnicas da literatura.

Cromossomo	1º algoritmo (%)	2º algoritmo (%)	Piper e Granum (%)	Wang(%)
1	99	99	85.1	98
2	96	99	91.5	97
3	81	82	92.1	90
4	98	99	94.8	97
5	99	100	96.5	96
6	99	99	91.7	92.9
7	100	100	93.4	96
8	97	99	91.7	88
9	99	99	92.6	97
10	99	99	93.8	88.9
11	100	99	95.7	96
12	99	98	94.9	84
13	99	98	81.9	93
14	97	96	84.8	95
15	94	93	85.6	99
16	89	90	93.2	86.1
17	87	90	97.2	77.2
18	90	89	91.8	76.8
19	62	65	95.8	85.9
20	96	93	97.2	80.9
21	97	96	70.3	99
22	88	86	77.3	96
X	97	97	91.6	89.6
Y	53	86	90.4	100
Total	**93.89**	**94.37**	**90.4**	**91.43**

Desta forma, o primeiro algoritmo proposto obteve uma taxa de acertos de 93.89%, já o segundo de 94.37%, sendo maiores que qualquer trabalho encontrado na literatura no decorrer destes estudos, sendo que os principais trabalhos de Wang et al. (2008) e Piper e Granum (1989) obtiveram 91.43% e 90.4% de acertos respectivamente. Além disso, uma diferença importante é em relação ao número de acertos em cromossomos maiores. No algoritmo aqui desenvolvido, os acertos na maior parte dos cromossomos chegam a quase 100%, o que não ocorre nos algoritmos de Piper e Granum, e somente em alguns casos de Wang. Porém, em cromossomos menores, no algoritmo desenvolvido os acertos passam a cair

semelhantemente aos resultados de Wang, sendo que os de Piper e Granum ainda mantêm um mesmo padrão.

6. CONCLUSÃO

Nas últimas décadas, a Genética Médica tornou-se uma das áreas da Genética que mais cresceram no Brasil. Apesar disso, ainda é pequeno o número de profissionais qualificados que trabalhem nesta área. Dentro desta área, surgiu a Citogenética, que veio com o intuito de pesquisar as alterações que ocorrem em nível de DNA nas populações, bem como realizar um estudo a respeito de agentes mutagênicos. Dentre as técnicas mais importantes se encaixa a análise do cariótipo, sendo uma técnica importantíssima no diagnóstico de doenças ligadas a alterações cromossômicas.

No decorrer dos anos, diversos avanços ocorreram nesta área, principalmente em relação ao bandeamento de cromossomos, a coleta dos materiais e a análise do cariótipo com o intuito de se melhorar o processo de identificação, agilizar o trabalho dos geneticistas, e também aumentar a confiabilidade dos diagnósticos relacionados a possíveis alterações cromossômicas que possam ser encontradas, podendo assim definir condutas terapêuticas mais rápidas e com mais confiabilidade. Porém, sistemas automáticos voltados a estes objetivos ainda são poucos, devido a uma série de fatores que dificultam a elaboração dos mesmos, tanto pela pouca quantidade de profissionais especialistas na área, aos custos que estes tipos de pesquisas envolvem, como também a respeito da dificuldade na elaboração destes sistemas que envolvem diversas etapas e abrangendo diversas áreas da computação tal como processamento de imagens, análise de dados, inteligência artificial, entre outras.

Neste trabalho, buscou-se realizar um estudo a respeito de técnicas de detecção do centrômero, que, além do padrão de bandas, é uma das principais características utilizadas no processo de identificação dos cromossomos. O processo de detecção do centrômero vem com dois principais objetivos: ao ser utilizado isoladamente, agiliza o processo de identificação de cromossomos e arranjo dos mesmos aos geneticistas; ao ser utilizado em sistemas automáticos, cria filtros que diminuem o espaço de busca no processo de identificação de um cromossomo, e desta forma aumenta ainda mais a confiabilidade do sistema, melhorando a taxa de acertos do processo de identificação de cromossomos.

A maior parte dos trabalhos relacionados cita a importância de se ter uma boa técnica de detecção do centrômero. Apesar disso, são poucos que realmente desenvolvem esta abordagem por considerarem uma tarefa difícil de ser implementada e que dificilmente se

conseguirá altas taxas de acertos na detecção do centrômero e classificação dos cromossomos em relação ao mesmo.

Desta forma, esta dissertação buscou, a partir de idéias apresentadas por trabalhos encontrados na literatura, desenvolver algumas técnicas de detecção do centrômero e demonstrar que é possível desenvolver uma técnica relativamente simples, que pode ser acoplada a qualquer sistema de identificação de cromossomos e que principalmente obtém uma alta taxa de acertos na classificação dos cromossomos em relação à posição do centrômero. Assim, foram desenvolvidos alguns métodos de detecção do centrômero e, por fim, foram apresentados dois algoritmos que fazem o uso destes métodos de forma a buscar o melhor de cada um, e assim obter uma taxa de acertos maior.

Os resultados obtidos nesta dissertação chegaram a uma taxa de acertos de 94.37%, que ao ser comparado aos trabalhos relacionados, é a maior taxa de acertos encontrada na literatura. Estes resultados mostram que, apesar de diversos autores afirmarem o contrário, é possível desenvolver uma técnica que traga um alto índice de acertos na detecção e classificação dos cromossomos em relação ao centrômero. Além disso, um fator importante dos algoritmos apresentados é que os resultados relacionados à taxa de acertos para cada tipo de cromossomo tende a ser uniforme, com exceção de poucos casos.

Outro fator relevante é em relação aos acertos obtidos em cada uma das metáfases da base de imagens utilizada, que também tende a seguir uma taxa de acertos bastante uniforme, gerando um desvio padrão de apenas 4.19%, sendo que na maior parte destas metáfases a taxa de acertos foi maior que 90% (ou seja, de 46 cromossomos, os acertos geralmente são maiores que 41). Com este alto índice de acertos é possível definir tanto um sistema isolado de auxilio ao geneticista, visando agilizar o processo de identificação do cariótipo, como também incorporar estes algoritmos a um sistema de identificação de cromossomos, que certamente aumentaria a confiabilidade e o índice de acertos na classificação.

Portanto, a maior contribuição deste trabalho é o estudo e a caracterização de diversas técnicas de detecção do centrômero, e principalmente aos algoritmos desenvolvidos, mostrando que ao contrário do que é visto em grande parte dos trabalhos encontrados na literatura, é possível obter um alto índice de acertos na detecção do centrômero. Além disso, este trabalho contribui para sistemas de identificação de cromossomos futuros ou até mesmo já existentes que podem fazer o uso tanto dos algoritmos como de algum dos métodos

desenvolvidos, e desta forma, tornando as tarefas desenvolvidas pelos geneticistas em laboratório mais ágeis e confiáveis.

REFERÊNCIAS

ACHARYA, T.; RAY, A. K. Image Processing: Principles and Applications. [S.l.]: Wiley-Interscience, 2005.

ACOSTA, A. X.; FERRAZ, V. E. F. **Exercício profissional em genética médica**: genética clínica no Brasil. In: Anais do XII Congresso Brasileiro de Genética Clínica. Teresópolis: Sociedade Brasileira de Genética Clínica; 2000. p. 10.

ABRAMOFF, M. D., MAGALHAES, P. J., RAM, S. J. **Image Processing with *ImageJ*.** Biophotonics International, v. 11, p. 36-42, 2004.

BIYANI, P.; WU, X.; SINHA, A. **Joint classification and pairing of human chromosomes**, IEEE/ACM Trans. Comput. Biol. Bioinformatics 2 (2005), p. 102–109.

BORGES-OSÓRIO, M.R.; ROBINSON, W.M. **Genética Humana**. Editora Artmed, 2ª edição, Porto Alegre, 2002.

BRACEWELL, R. **The Fourier Transform and Its Applications**. [S.l.]: McGraw-Hill Science, 1999.

BRUNONI, D. **Estado atual do desenvolvimento dos serviços de genética médica no Brasil**. Rev Bras Genet 1997; 20 Suppl:11-23.

BRUNONI, D. **Aconselhamento genético**. Ciênc Saúde Coletiva 2002; 7:101-7.

CAMPBELL, F. W.; ROBSON, J. G. **Application of fourier analysis to the visibility of gratings**. Journal of Physiology, 1968.

CAPUTO, L. Z. **Aplicação do cariótipo nas investigações de doenças onco-hematológicas**. Medicina em Foco, p.1-2, 2005.

CASTLEMAN, K. **Automated chromosome classification using wavelet-based band pattern descriptors.** Proceedings 13th IEEE Symposium on Computer-Based Medical Systems. CBMS, p. 189-94, 2000.

CHO, J.; RYU, S. Y.; WOO, S. H. **A study for the hierarchical artificial neural network model for Giemsa-stained human chromosome classification.** Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. V. 6, p. 4588-4591, 2004.

CONROY, J. M.; KOLDA, T.G.; O'LEARY, D.P. **Chromosome identification using hidden markov models: comparison with neural networks, singular value decomposition, principal components analysis, and Fisher discriminant analysis**, Lab. Invest. 80, p. 1629–1641, 2000.

DUDA, P. E. H. R. O. **Pattern Classi¯cation and Scene Analysis**. [S.l.]: John Wiley and Sons Inc, 1973.

FARIA, A. P. M.; FERRAZ, V. E. F.; ACOSTA, A. X.; BRUNONI, D. **Clinical genetics in the developing countries**: the Brazilian situation. Community Genet 2004; 7:95-105.

FORD, C. E.; HAMETON, J. L. **A Colchicine, Hypotonic Citrate, Squash Sequence for Mammalian Chromosomes.** v. 31, n. 6, p. 247-251, 1956.

GALLUS G.; NEURATH, P. W. **Improved computer chromosome analysis incorporating preprocessing and boundary analysis**. Phys Med Biol v. 15, p.435-45, 1970.

GONZALEZ, R. C.; WOODS, R. E. **Processamento de imagens digitais**. [S.l.]: Edgard BlÄucher, 2000.

GRAHAM, J. **Automation of routine clinical chromosome analysis.** Analytical and quantitative cytology and histology, v. 9, p. 383–390, 1987.

GREGOR, J.; GRANUM, Erik. **Finding chromosome centromeres using band pattern information,** Computers in Biology and Medicine, Vol. 21, No. 1/2, p. 55-67, 1991.

GRISAN, E., POLLETI, E., RUGGERI, A. **Automatic segmentation and disentangling of chromosomes in Q-band prometaphase images.** IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society, v.13, p. 575-81, 2009.

GROEN F. C. A. **Prophase banding classifiers**. Preliminary Report on EEC Work Group Meeting on Automated Chromosome Analysis, Leiden, pp 82-93, 1985.

GUIMARAES, L.; SCHUK, A.; ELBERN, A. **Chromosome classification for karyotype composing applying shape representation on wavelet packet transform.** Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, p. 941-43, 2003.

KAO, J.; CHUANG, J.; WANG, T. **Chromosome classification based on the band profile similarity along approximate medial axis.** Pattern Recognition, v. 41, p. 77-89, 2008.

KIM, T.; BHATTACHARYYA, D., BANDYOPADHYAY, S. K. **Supervised chromosome clustering and image classification.** Future Generation Computer Systems, v. 27, p. 372-76, 2011.

KURTZ, G. C. ; BONINI, T. ; PERLES, L. A. ; SAGRILLO, M. R. ; LIBRELOTTO, Giovani R. . **Identificação automática de cromossomos humanos**. In: SIBGRAPI, 2008, Campo Grande. XXI Brazilian Symposium on Computer Graphics and Image Processing, 2008. p. 33-36.

LEGRAND, B.; CHANG, C.; ONG, S. **Chromosome classification using dynamic time warping.** Pattern Recognition Letters, v. 29, p.215 – 222, 2008.

LEITE, N. J. **Introdução ao Processamento de Imagens Digitais**. Campinas: Instituto de Computação - Unicamp, 2004. Disponível em <http://www.ic.unicamp.br/~afalcao/sensremoto/processamento.ppt> Acesso em 1 nov. 2011.

LUNDSTEEN, C.; GERDES, C.; MAAHR, J. **Automatic classification of chromosomes as part of a routine system for clinical analysis,** Cytometry, p. 1-7, 1985.

MORADI, M.; SETAREHDAN. **New features for automatic classification of human chromosomes: A feasibility study,** Pattern Recognition Letters, v. 27, p. 19-28, 2006.

MORADI, M.; SETAREHDAN, S.K.; GHAFFARI, S.R. **Automatic locating the centromere on human chromosome pictures,** Proceedings of the 16th IEEE conference on Computer-based medical systems (CBMS'03), Marina Krol, Sunanda Mitra, and D. J. Lee (Eds.). IEEE Computer Society, Washington, DC, USA, p. 56-61, 2003.

NANNI, L. **A reliable method for design an automatic karyotyping system.** Neurocomputing, v.69, p.1739-42, 2006.

NUSSBAUM, R. L.; MCINNES, R. R.; WILLARD, H. F. **Thompson & Thompson**: **Genética Médica,** 7 ed., Elseiver, 2008.

OSKOUEI, B.; SHANBEHZADEH, J. **Chromosome Classification Based on Wavelet Neural Network.** International Conference on Digital Image Computing: Techniques and Applications, p. 605-610, 2010.

PAUT, A. **Gene and chromosome analysis**. Academic Press, Inc, San Diego, CA, 1993.

PEREIRA, E. T. **Citogenética**: Fundamentos e Aplicação Clínica. Arq. Cat. Med. - v. 17, No. 2 - Abril/Junho de 1988.

PHILIP, J.; GRANUM, E. **Quantitative analysis of 6985 digitized trypsin {G}-banded human metaphase chromosomes.** Clinical Genetics, v. 18, p.355-370, 1980.

PIPER, J.; GRANUM, E. **On fully automatic feature measurement for banded chromosome classification,** Cytometry, v. 10, p. 242-255, 1989.

POLETTI, E.; GRISAN, E.; RUGGERI, A. **Automatic classification of chromosomes in Q-band images.** 30th Annual International Conference of the IEEE-EMBS, Vancouver, British Columbia, Canada, p. 20-24, 2008.

RASBAND, W. S. ***ImageJ*.** U. S. National Institutes of Health, Bethesda, Maryland, USA, http://*ImageJ*.nih.gov/ij/, 2011.

ROBERTIS, E.M.D.; HIB,J. **Bases da biologia celular e molecular**. [S.l.]: Guanabara Koogan, 2006.

RODGERS, J. L.; NICEWANDER, A. W. **Thirteen ways to look at the correlation coefficient**. The American Statistician, 1988.

ROSHTKHARI, M. J.; SETAREHDAN, S. K. **A novel algorithm for straightening highly curved images of human chromosome.** Pattern Recognition Letters, v. 29, p.1208-17, 2008.

SAVITZKY, A.; GOLAY, M. J. E. **Smoothing and Differentiation of Data by Simplified Least Squares Procedures.** Analytical Chemistry 36 (8), p. 1627-39, 1964.

SCHWARTZKOPF, W. C.; BOVIK, A.C.; EVANS, B. L. **Maximum-likelihood techniques for joint segmentation–classification of multispectral chromosome images**, IEEE Trans. Med. Imaging 24, p. 1593–1610, 2005.

SHAPIRO, L.; STOCKMAN, G. **Computer Vision**. [S.l.]: Prentice Hall, 2001.

SOUZA, A. F. de; BANON, G. J. F. **Um algoritmo simples de esqueletização.** Workshop dos Cursos de Computação Aplicada do INPE, 2003

STANLEY, R.J.; KELLER, J.M.; CALDWELL, C.W.; GADER, P. **Centromere attribute integration based chromosome polarity assignment,** Conference of the American Medical Informatics Association, p. 284-288, 1996.

TIJO, J.H.; LEVAN, A. **The chromosome number of man**. Hereditas, v. 42, p. 1–6, 1956.

TRMANANDA, R. **A reliable method for designing an automatic karyotyping system.** Neurocomputing, v.69, p.1739-42, 2006.

TRIMANANDA, R. **Chromosome Centromere and Chromatid's Banding Identification Using Pattern Vector.** Second International Conference on Advances in Computing, Control, and Telecommunication Technologies, p. 132-134, 2010.

VERMA R.S.; BABU A. **Human Chromosomes**: Principles and Techniques. McGraw-Hill, Inc, 2a edição, New York, 1995.

WANG, X.; ZHENG, B.; LI, S.; CHEN, W.; WOOD, M. C.; LIU, H. **Development and evaluation of automated systems for detection and classification of banded chromosomes**: current status and future perspectives. Journal of Physics D: Applied Physics, v.38, p. 2536-42. 2005

WANG, X.; ZHENG, B.; LI, S.; MULVIHILL, J. J.; WOOD, M. C.; LIU, H. **A rule-based computer scheme for centromere identification and polarity assignment of metaphase chromosomes**. Comput. Methods Prog. Biomed. 89, 1, p. 33-42, 2008.

WANG, X.; ZHENG, B.; LI, S.; MULVIHILL, J. J.; WOOD, M. C.; LIU, H. **Automated classification of metaphase chromosomes: optimization of an adaptive computerized scheme.** Journal of biomedical informatics, v. 42, p 22-31, 2009.

ZHANG, T. Y.; SUEN, C. Y. **A fast parallel algorithm for thinning digital patterns**. Commun. ACM, v. 27, n. 3, p. 236-239, 1984.

ANEXOS

ANEXO A – TABELA COM CLASSIFICAÇÕES DE ACORDO COM O COMPRIMENTO

Tabela 7.1: Porcentagem geral de classificação de acordo com o comprimento.

	1º algoritmo proposto			2º algoritmo proposto		
Tipo	M	S	A	M	S	A
3	0	0	100	0	0	100
6	0	0	100	0	0	100
10	0	0	100	0	0	100
11	0	0	100	0	0	100
12	0	0	100	12,5	0	87,5
13	10	0	90	0	0	100
14	11,11	0	88,89	5,56	0	94,44
15	3,7	0	96,3	3,7	0	96,3
16	5,88	0	94,12	5,88	0	94,12
17	0	0	100	0	0	100
18	11,86	0	88,14	6,78	0	93,22
19	20,97	1,61	77,42	8,06	1,61	90,32
20	23,75	1,25	75	20	1,25	78,75
21	29,55	1,14	69,32	28,41	1,14	70,45
22	34,88	2,33	62,79	26,74	2,33	70,93
23	43,24	8,11	48,65	32,43	0,9	66,67
24	55,56	6,17	38,27	48,15	3,7	48,15
25	56,63	14,46	28,92	45,78	16,87	37,35
26	60,76	24,05	15,19	48,1	24,05	27,85
27	52,13	32,98	14,89	53,19	29,79	17,02
28	29,89	55,17	14,94	27,59	50,57	21,84
29	32,26	60,22	7,53	12,9	79,57	7,53
30	27,27	65,91	6,82	30,68	67,05	2,27
31	23,86	68,18	7,95	21,59	57,95	20,45
32	38,32	44,86	16,82	33,64	38,32	28,04
33	28,24	50,59	21,18	23,53	51,76	24,71
34	24,71	36,47	38,82	23,53	27,06	49,41
35	32,56	29,07	38,37	24,42	31,4	44,19
36	18,09	26,6	55,32	26,6	21,28	52,13
37	11,63	12,79	75,58	13,95	5,81	80,23
38	13,04	11,96	75	15,22	10,87	73,91
39	5,26	3,51	91,23	8,77	0	91,23
40	0	13,33	86,67	2,67	8	89,33
41	0	13,83	86,17	0	11,7	88,3
42	0	20,97	79,03	0	20,97	79,03
43	0	29,73	70,27	0	29,73	70,27
44	0	55,93	44,07	0	55,93	44,07
45	0	70,59	29,41	0	69,12	30,88
46	0	83,95	16,05	0	81,48	18,52

47	0	88,35	11,65	0	93,2	6,8
48	0	91,07	8,93	0	91,07	8,93
49	0	92,86	7,14	0	94,05	5,95
50	0	97,52	2,48	0	98,76	1,24
51	0	99,22	0,78	0	100	0
52	0	97,3	2,7	0	97,3	2,7
53	0	100	0	0	100	0
54	0	99,08	0,92	0	100	0
55	0	100	0	0	100	0
56	0	100	0	0	100	0
57	0	100	0	0	100	0
58	0	98,55	1,45	0	100	0
59	1,89	96,23	1,89	0	100	0
60	0	100	0	0	100	0
61	0	100	0	0	100	0
62	4,48	95,52	0	1,49	98,51	0
63	0	100	0	0	100	0
64	3,23	96,77	0	3,23	96,77	0
65	4,62	95,38	0	1,54	98,46	0
66	0	100	0	0	100	0
67	8,33	91,67	0	6,67	93,33	0
68	8,7	91,3	0	2,17	97,83	0
69	7,02	92,98	0	5,26	94,74	0
70	16,67	83,33	0	16,67	83,33	0
71	12,2	87,8	0	9,76	90,24	0
72	20	80	0	17,5	82,5	0
73	25	75	0	34,09	65,91	0
74	43,33	56,67	0	43,33	56,67	0
75	50	50	0	42,5	57,5	0
76	54,17	45,83	0	50	50	0
77	65,52	34,48	0	62,07	37,93	0
78	60,87	39,13	0	69,57	30,43	0
79	86,67	13,33	0	80	20	0
80	76,47	23,53	0	52,94	47,06	0
81	50	50	0	58,33	41,67	0
82	71,43	28,57	0	71,43	28,57	0
83	40	60	0	40	60	0
84	22,22	77,78	0	22,22	77,78	0
85	15,38	84,62	0	7,69	92,31	0
86	9,09	90,91	0	9,09	90,91	0
87	22,22	77,78	0	22,22	77,78	0
88	16,67	83,33	0	16,67	83,33	0
89	22,22	77,78	0	22,22	77,78	0
90	12	88	0	12	88	0
91	26,09	73,91	0	26,09	73,91	0
92	26,67	73,33	0	26,67	73,33	0
93	38,1	61,9	0	33,33	66,67	0
94	55,17	44,83	0	55,17	44,83	0

95	63,64	36,36	0	63,64	36,36	0
96	64	36	0	64	36	0
97	54,55	45,45	0	54,55	45,45	0
98	69,57	30,43	0	69,57	30,43	0
99	100	0	0	50	50	0
100	85,98	14,02	0	85,05	14,95	0

ANEXO B – TABELA COM CLASSIFICAÇÕES DE ACORDO COM O TIPO DE CROMOSSOMO

Tabela 7.2: Porcentagem de classificação em metacêntricos (M), submetacêntricos (S) e acrocêntricos (A) para cada tipo de cromossomo nos dois algoritmos propostos.

	Primeiro Algoritmo			Segundo Algoritmo			
Tipo	M (%)	S (%)	A (%)	M (%)	S (%)	A (%)	Correta
1	**99,02**	0,98	0	**99,51**	0,49	0	M
2	3,32	**96,68**	0	0,47	**99,53**	0	S
3	**81,28**	17,81	0,91	**82,65**	16,44	0,91	M
4	1,9	**98,1**	0	0,95	**99,05**	0	S
5	0,93	**99,07**	0	0	**100**	0	S
6	0	**99,56**	0,44	0,44	**99,11**	0,44	S
7	0	100	0	0	**100**	0	S
8	0,89	**97,77**	1,34	0	**99,11**	0,89	S
9	0,44	**99,56**	0	0,44	**99,56**	0	S
10	0,44	**99,12**	0,44	0,44	**99,56**	0	S
11	0	**100**	0	0,88	**99,12**	0	S
12	0	**99,14**	0,86	0	**98,71**	1,29	S
13	0,43	0,43	**99,14**	0,43	1,29	**98,28**	A
14	0,87	1,3	**97,84**	1,73	1,3	**96,97**	A
15	0,86	4,31	**94,83**	1,29	5,17	**93,53**	A
16	**89,18**	8,23	2,6	**90,04**	6,93	3,03	M
17	10,3	**87,98**	1,72	6,01	**90,99**	3	S
18	0,86	**90,56**	8,58	0,43	**89,7**	9,87	S
19	**62,03**	4,22	33,76	**65,4**	7,17	27,43	M
20	**96,17**	0,85	2,98	**93,62**	1,7	4,68	M
21	2,1	0,84	**97,06**	0,42	2,94	**96,64**	A
22	9,36	2,13	**88,51**	6,81	6,38	**86,81**	A
X	0	**97,83**	2,17	0	**97,83**	2,17	S
Y	2,33	44,19	**53,49**	0	13,95	**86,05**	A

ANEXO C – TABELA COM OS MÉTODOS UTILIZADOS PARA CADA COMPRIMENTO NO PRIMEIRO ALGORITMO PROPOSTO

Tabela 7.3: Métodos utilizados em cada comprimento relativo de cromossomo no primeiro algoritmo proposto.

Comprimento	Algoritmo	Comprimento	Algoritmo	Comprimento	Algoritmo
0	3	33	1	67	2
1	3	34	1	68	1
2	3	35	3	69	1
3	3	36	1	70	1
4	3	37	1	71	1
5	3	38	1	72	1
6	3	39	5	73	1
7	2	40	5	74	1
8	2	41	3	75	1
9	1	42	3	76	1
10	3	43	1	77	1
11	3	44	1	78	1
12	2	45	1	79	3
13	3	46	1	80	1
14	2	47	3	81	1
15	2	48	1	82	1
16	2	49	1	83	1
17	2	50	3	84	1
18	2	51	1	85	1
19	2	52	1	86	1
20	2	53	1	87	5
21	2	54	1	88	1
22	1	55	3	89	1
23	4	56	3	90	1
24	4	57	2	91	3
25	4	58	1	92	1
26	4	59	4	93	1
27	4	60	1	94	1
28	4	61	2	95	1
29	4	62	1	96	1
30	4	63	1	97	1
31	4	64	2	98	1
32	3	65	1	99	1
		66	2	100	3

ANEXO D – TABELA COM A COMBINAÇÃO DE MÉTODOS UTILIZADOS PARA CADA COMPRIMENTO NO SEGUNDO ALGORITMO PROPOSTO

Tabela 7.4: Combinação de métodos utilizados em cada comprimento relativo de cromossomo no segundo algoritmo proposto.

Comprimento	Algoritmos	Comprimento	Algoritmos	Comprimento	Algoritmos
1	11111	35	11111	69	11001
2	11111	36	10011	70	11101
3	00100	37	11011	71	11001
4	11111	38	11111	72	11001
5	11111	39	11100	73	10000
6	11111	40	11110	74	11001
7	11111	41	11110	75	11101
8	11111	42	11101	76	11111
9	11111	43	10100	77	11001
10	11111	44	11101	78	10111
11	11111	45	10001	79	10000
12	11111	46	11101	80	00100
13	11101	47	11110	81	00100
14	11101	48	11110	82	11001
15	10101	49	11110	83	11101
16	11111	50	11110	84	11101
17	10101	51	11110	85	11111
18	01000	52	11110	86	11001
19	11111	53	11000	87	11101
20	01110	54	11000	88	11001
21	01110	55	11111	89	11101
22	01110	56	10100	90	11001
23	01110	57	11100	91	11001
24	11010	58	11111	92	11101
25	11010	59	11111	93	11111
26	00010	60	11111	94	11111
27	10110	61	11110	95	11101
28	11110	62	11111	96	11111
29	00110	63	11111	97	11111
30	11110	64	11111	98	11111
31	00010	65	11111	99	11111
32	11110	66	11111	100	11111
33	10111	67	11101		
34	11110	68	11111		

ANEXO E – TABELA COM A TAXA DE ACERTOS DO SEGUNDO ALGORITMO PROPOSTO PARA CADA METÁFASE

Tabela 7.5: Taxa de acertos na detecção do centrômero do segundo algoritmo proposto em cada metáfase/cariótipo da base de imagens. *Met* indica o índice da metáfase.

Met.	**Acertos (%)**	**Met.**	**Acertos (%)**	**Met.**	**Acertos (%)**	**Met.**	**Acertos (%)**
1	93,00	31	97,00	61	88,00	91	95,00
2	100,00	32	95,00	62	97,00	92	95,00
3	93,00	33	100,00	63	93,00	93	95,00
4	93,00	34	100,00	64	95,00	94	90,00
5	93,00	35	97,00	65	100,00	95	93,00
6	93,00	36	93,00	66	100,00	96	97,00
7	90,00	37	93,00	67	95,00	97	95,00
8	92,00	38	97,00	68	97,00	98	93,00
9	90,00	39	90,00	69	95,00	99	95,00
10	93,00	40	95,00	70	95,00	100	95,00
11	86,00	41	84,00	71	97,00	101	97,00
12	97,00	42	95,00	72	91,00	102	97,00
13	97,00	43	90,00	73	90,00	103	97,00
14	100,00	44	93,00	74	97,00	104	97,00
15	95,00	45	97,00	75	97,00	105	83,00
16	88,00	46	100,00	76	95,00	106	93,00
17	78,00	47	93,00	77	94,00	107	93,00
18	90,00	48	93,00	78	90,00	108	93,00
19	89,00	49	97,00	79	94,00	109	86,00
20	93,00	50	100,00	80	86,00	110	97,00
21	95,00	51	97,00	81	89,00	111	97,00
22	100,00	52	92,00	82	97,00	112	97,00
23	81,00	53	95,00	83	93,00	113	89,00
24	93,00	54	93,00	84	97,00	114	90,00
25	91,00	55	97,00	85	86,00	115	95,00
26	95,00	56	97,00	86	88,00	116	100,00
27	95,00	57	95,00	87	93,00	117	100,00
28	93,00	58	97,00	88	93,00	118	95,00
29	97,00	59	93,00	89	100,00	119	97,00
30	85,00	60	97,00	90	95,00		

ANEXO F – TABELA COM VALORES DE *k* E *r* PARA O ALGORITMO DA ROSA-DOS-VENTOS COM NÍVEIS DE CINZA POR COMPRIMENTO

Tabela 7.6 – Valores de *k* e *r* para o algoritmo da rosa-dos-ventos com níveis de cinza por comprimento.

Comprimento	*kIni*	*kFim*	*rIni*	*rFim*
10 até 21	10	20	1	2
22	10	26	1	3,49
23	10	20	1	2
24	10	20	1	2
25	15	34	1	2,87
26	23	53	1,73	3,89
27	50	72	1,27	2
28	56	66	1,17	2
29	30	57	1,62	2,96
30	34	51	1,56	2,51
31	22	34	1	3,58
32	19	53	1,75	3,68
33	19	43	1,85	3,11
34	16	56	1,8	3,55
35	22	50	1,57	2,66
36	20	35	2,8	3,05
37	24	37	2,19	2,62
38	22	45	1,55	2,43
39	10	20	1	2
40	17	53	1	2,44
41	15	58	1	2,69
42	12	51	1	2,66
43	18	62	1	2,56
44	13	46	1	3,13
45	15	49	1	2,88
46	14	53	1	2,96
47	21	38	1	3,09
48	14	68	1	2,99
49	15	47	1	3,02
50	23	33	1	3,55
51	21	43	1	3,07
52	21	37	1	3,57
53	25	35	1	3,13
54	22	40	1	3,5
55	25	35	1	3,42
56	24	35	1	3,29

57	24	41	1	3,01
58	25	35	1	3,31
59	27	37	1	2,88
60	24	34	1	3,48
61	26	40	1,01	3,3
62	67	77	1	3,08
63	26	42	1	3,22
64	25	35	1,55	3,76
65	43	57	1	2
66	24	38	1	3,44
67	25	37	1	3,22
68	26	36	1	3,12
69	24	42	1	3,3
70	29	53	1,23	2,98
71	15	37	1	3,44
72	30	68	1,31	2,55
73	24	34	1	3,46
74	38	51	1,42	2,02
75	27	65	1,7	2,69
76	30	70	2,09	2,56
77	30	50	1,29	2,64
78	32	61	1,91	2,19
79	10	25	1	3,37
80	41	58	1,27	2
81	38	52	1,25	2
82	32	47	1,36	2,35
83	10	52	1,24	2
84	10	40	1	2
85	10	42	1	2,93
86	10	41	1	2
87	10	38	1	2
88	10	51	1,04	2
89	10	44	1	2
90	18	51	1,07	2
91	24	34	1,97	3,32
92	10	40	1	2
93	23	55	1,4	3,2
94	18	49	1,23	2
95	10	39	1	2
96	15	52	1,24	2
97	25	53	1,14	2
98	14	53	1,12	2
99	10	47	1	2
100	11	54	1,23	2

ANEXO G – DETALHES SOBRE O COEFICIENTE DE CORRELAÇÃO

Coeficiente de Correlação

A coeficiente de correlação utilizado é definido como:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

Equação 15

As variáveis x e y correspondem aos valores do vetor referente à quantidade percentual de cromossomos de acordo com o seu comprimento e *n* é o tamanho desses vetores (que deve ser o mesmo). O retorno deste cálculo será um numero de -1 a 1 que quanto maior for o seu valor, maior deve ser a semelhança entre os dados. No caso se esta comparação tivesse sido feita por uma amostra idêntica de dados, o valor retornado seria igual a 1 (RODGERS e NICEWANDER, 1988).