

**CARLOS HENRIQUE ZANELATO PANTALEÃO**

**CONTRIBUIÇÃO À ANÁLISE E CLASSIFICAÇÃO  
CITOGENÉTICA BASEADA NO  
PROCESSAMENTO DIGITAL DE IMAGENS E NO  
ENFOQUE LÓGICO-COMBINATÓRIO**

**Florianópolis  
2003**

**UNIVERSIDADE FEDERAL DE SANTA CATARINA**

**PROGRAMA DE PÓS-GRADUAÇÃO EM  
ENGENHARIA ELÉTRICA**

**CONTRIBUIÇÃO À ANÁLISE E CLASSIFICAÇÃO  
CITOGENÉTICA BASEADA NO  
PROCESSAMENTO DIGITAL DE IMAGENS E NO  
ENFOQUE LÓGICO-COMBINATÓRIO**

Tese submetida à  
Universidade Federal de Santa Catarina  
como parte dos requisitos para a obtenção do  
grau de Doutor em Engenharia Elétrica.

*por*

**CARLOS HENRIQUE ZANELATO PANTALEÃO**

Florianópolis, Dezembro de 2003

# CONTRIBUIÇÃO À ANÁLISE E CLASSIFICAÇÃO CITOGENÉTICA BASEADA NO PROCESSAMENTO DIGITAL DE IMAGENS E NO ENFOQUE LÓGICO- COMBINATÓRIO

Carlos Henrique Zanelato Pantaleão

Esta Tese foi julgada adequada para obtenção do Título de Doutor em Engenharia Elétrica, Área de Concentração em Sistemas de Informação, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Santa Catarina.

---

Jefferson Luiz Brum Marques, PhD.

Coordenador do Programa de Pós-Graduação em Engenharia Elétrica – UFSC

## **Banca Examinadora:**

---

Jefferson Luiz Brum Marques, PhD.

Orientador

---

Fernando Mendes de Azevedo, DSc.

Co-Orientador

---

Edna Lúcia Flôres, Dra.

UFU

---

João Baptista Tadanobo Yabu-UTI, Dr.

DECOM/UNICAMP

---

Ícaro dos Santos, PhD.

IEB/UFSC

---

Maria Cecília Menks Ribeiro, Dra.

IEB/UFSC

---

Ingrid Tremel Barbato, Esp.

Clínica Neurogene

## **AGRADECIMENTOS**

A todos aqueles que contribuíram de alguma forma, direta ou indiretamente, para a realização deste trabalho, à Coordenação de Curso de Pós-Graduação em Engenharia Elétrica da UFSC, ao Instituto de Engenharia Biomédica da UFSC e ao CNPq pelo auxílio financeiro.

Resumo da Tese apresentada à UFSC como parte dos requisitos necessários para a obtenção do grau de Doutor em Engenharia Elétrica.

# **CONTRIBUIÇÃO À ANÁLISE E CLASSIFICAÇÃO CITOGENÉTICA BASEADA NO PROCESSAMENTO DIGITAL DE IMAGENS E NO ENFOQUE LÓGICO- COMBINATÓRIO**

**Carlos Henrique Zanelato Pantaleão**

Dezembro / 2003

Orientador: Prof. **Jefferson Luiz Brum Marques**, EE, MSc, PhD.

Co-Orientador: Prof. **Fernando Mendes de Azevedo**, EE, MSc, DSc.

Área de Concentração: Sistemas de Informação.

Palavras-chave: Extrapolação, Leucemia Pró-Mielocítica Aguda, Lógico-Combinatório.

Número de Páginas: 149.

## **RESUMO**

Este trabalho apresenta a elaboração de um *Sistema Computadorizado para Análise e Classificação Citogenética*, que constitui uma nova metodologia para reduzir o viés da análise citogenética clínica. Tal sistema foi implementado visando dois objetivos específicos: o reconhecimento automático dos cromossomos através de uma única cena e a análise da leucemia pró-mielocítica aguda. Para alcançar o primeiro objetivo, aplicou-se técnicas de processamento digital de imagens, visando extrair as características de área, fator de forma, distância *Euclidiana* e correlação linear de *Pearson*, com mais precisão. Através destas características, foi possível a identificação automática dos cromossomos homólogos, atingindo 63,04% de índice de acerto quando utilizadas imagens de 276 cromossomos, as quais foram obtidas de diferentes laboratórios. Para atingir o segundo propósito, foi desenvolvido um novo método de prolongamento das extremidades do eixo médio extraído dos cromossomos, baseando-se na técnica de extrapolação pelos mínimos quadrados. Desta forma, foi possível detectar a translocação  $t(15,17)$  que caracteriza a leucemia e, para apresentar o grau de analogia com a doença, utilizou-se o enfoque lógico-combinatório. Os resultados demonstraram um bom desempenho do sistema, obtendo índices entre 63,73 a 99,70% de semelhança com o rearranjo cromossômico, quando utilizadas imagens de células leucêmicas e valores abaixo de 17,30% com o uso de imagens de células normais. Logo, através deste estudo, foi possível comprovar que é possível o reconhecimento automático dos cromossomos através de uma única cena e obter uma ferramenta que permite detectar com maior precisão o rearranjo característico da leucemia pró-mielocítica aguda, oferecendo ao citogeneticista condições para a detecção do rearranjo cromossômico com maior eficiência e confiabilidade.

Abstract of Thesis presented to UFSC as a partial fulfillment of the requirements for the degree of Doctor in Electrical Engineering

# **CONTRIBUTION TO THE CYTOGENESIS ANALYSIS AND CLASSIFICATION BASED ON THE DIGITAL IMAGE PROCESSING AND ON THE LOGICAL COMBINATORIAL APPROACH**

**Carlos Henrique Zanelato Pantaleão**

December/2003

Advisor: Prof. **Jefferson Luiz Brum Marques**, EE, MSc, PhD.

Co-advisor: Prof. **Fernando Mendes de Azevedo**, EE, MSc, DSc.

Area of Concentration: Information Systems.

Keywords: Extrapolation, Acute Promyeloid Leukemia, Logical Combinatorial Approach

Number of Pages: 149.

## **ABSTRACT**

The present work approaches the elaboration of a *Computerized System for Cytogenesis Analysis and Classification*, which constitutes a new methodology to reduce the *bias* of the clinical diagnosis in cytogenetic human. Such system was implemented aiming at two specific objectives: the automatic recognition of the chromosomes through one only scene and the analysis of the acute promyelocytic leukemia disease. To reach the first objective techniques of digital processing of images were applied, aiming at extracting the characteristics of area, factor of form, *Euclidean* distance and linear correlation of *Pearson* with more precision. Through these characteristics it was possible to identify the homologous chromosomes automatically, reaching 63.04% of precision index when using 276 images of chromosomes, which were gotten from different laboratories. To reach the second intention, a new method of prolongation of the extremities of the medial axes extracted of the chromosomes was developed, being based on the technique of extrapolation for the square minimums. Thus, it was possible to detect the *t(15,17)* translocation that characterizes the leukemia and, to present the degree of analogy with the disease, the logical combinatorial approach was used. The results demonstrated a good performance of the system, obtaining indices between 63.73 and 99.70% of similarity with the chromosomal rearrange, when using images of leukemic cells and values below 17.30% when using images of normal cells. Therefore, through this study it was possible to prove that it is possible to recognize the chromosomes automatically through one only scene and to obtain a tool that allows to detect with larger precision the characteristic rearrange of the acute promyelocytic leukemia disease, offering the citogeneticist conditions for the detection of the chromosomal rearrange with more efficiency and confiability.

## SUMÁRIO

<b>Lista de Figuras.....</b>	<b>X</b>
<b>Lista de Tabelas.....</b>	<b>xvi</b>
<b>Lista de Abreviaturas.....</b>	<b>xviii</b>
<b>Capítulo I – Introdução.....</b>	<b>1</b>
1.1 – Introdução.....	1
1.2 – Origem e Importância do Trabalho.....	2
1.3 – Objetivos.....	5
1.3.1 – Objetivo Geral.....	5
1.3.2 – Objetivos Específicos.....	5
1.4 – Hipótese do Trabalho.....	6
1.5 – Metodologia.....	7
1.6 – Apresentação da Tese.....	10
<b>Capítulo II – Estado da Arte.....</b>	<b>11</b>
2.1 – Introdução.....	11
2.2 – Sistemas Automáticos para Análise e Classificação Citogenética.....	11
2.3 – Histórico das Técnicas Utilizadas pelos Sistemas de Cariotipagem.....	13
2.4 – Considerações Finais.....	22
<b>Capítulo III – Citogenética Humana.....</b>	<b>23</b>
3.1 – Introdução.....	23
3.2 – Aspectos Biológicos.....	23
3.3 – O Ciclo Celular.....	24
3.4 – Identificação dos Cromossomos Metafásicos.....	26
3.4.1 – Identificação Clássica dos Cromossomos Metafásicos.....	27
3.4.2 – Identificação dos Cromossomos por Intermédio das Bandas.....	31

## **SUMÁRIO - Continuação**

3.5 – O Cariograma.....	35
3.6 – Preparação das Células Metafásicas como Amostra.....	36
3.7 – Fontes de Incerteza no Domínio do Problema.....	37
3.8 – Leucemia Pró-Mielocítica Aguda.....	39
3.9 – Considerações Finais.....	42
<b>Capítulo IV – Processamento e Análise de Imagens.....</b>	<b>43</b>
4.1 – Introdução.....	43
4.2 – Conceitos Gerais.....	43
4.3 – Sistema de Visão Artificial.....	44
4.4 – Transformação.....	46
4.5 – Pré-Processamento.....	47
4.6 – Segmentação.....	49
4.7 – Descrição.....	55
4.8 – Reconhecimento.....	64
4.9 – Interpretação.....	64
4.10 – Considerações Finais.....	65
<b>Capítulo V – Enfoque Lógico-Combinatório.....</b>	<b>66</b>
5.1 – Introdução.....	66
5.2 – Aspectos Gerais.....	66
5.3 – Formulação do Problema.....	69
5.4 – Formalização do Problema.....	70
5.4.1 – Seleção das Variáveis.....	70
5.4.1.1 – Critério de Analogia.....	70
5.4.1.2 – Critério de Comparação de Valores de uma Variável.....	72
5.4.1.3 – Funções de Semelhança entre Objetos.....	74
5.4.2 – Exposição Formal de um Problema de Seleção de Variáveis.....	75
5.4.3 – Conceitos da Teoria de Testores.....	77

## **SUMÁRIO - Continuação**

5.4.4 – Importância Informacional das Variáveis.....	81
5.5 – Modo de Solução do Problema Expresso em Termos Matemáticos.....	83
5.6 – Considerações Finais.....	84
<b>Capítulo VI – Metodologia Proposta.....</b>	<b>85</b>
6.1 – Introdução.....	85
6.2 – Descrição da Metodologia Proposta.....	85
6.2.1 – Formulação Geral do Problema.....	87
6.2.2 – Formalização do Problema.....	90
6.2.2.1 – Formalização do Problema de Identificação Citogenética.....	91
6.2.2.2 – Formalização da Leucemia Pró-Mielocítica Aguda.....	94
6.2.3 – Modo de Solução do Problema Expresso em Termos Matemáticos.....	106
6.3 – Considerações Finais.....	117
<b>Capítulo VII – Resultados Obtidos.....</b>	<b>118</b>
7.1 – Introdução.....	118
7.2 – Aplicação do Modelo Proposto.....	118
7.2.1 – Característica das Imagens Adquiridas.....	118
7.2.2 – Procedimento Experimental.....	121
7.3 – Resultados Obtidos.....	123
7.4 – Análise e Interpretação dos Resultados.....	129
7.5 – Considerações Finais.....	135
<b>Capítulo VIII – Conclusões, Contribuições e Sugestões para Futuros Trabalhos.....</b>	<b>136</b>
8.1 – Introdução.....	136
8.2 – Conclusões.....	136

## **SUMÁRIO - Continuação**

8.3 – Limitações do Trabalho.....	141
8.4 – Contribuições deste Trabalho.....	141
8.5 – Sugestões para Futuros Trabalhos.....	142
<b>Referências Bibliográficas.....</b>	<b>144</b>

## LISTA DE FIGURAS

Figura 2.1 – Componentes de uso genérico em um sistema de cariotipagem: sistema de aquisição de imagens, processamento, saída e armazenamento.....	13
Figura 3.1 – Ilustração idealizada da forma dos cromossomos metafásicos, de acordo com a posição do centrômero, ou seja, cromossomos metacêntricos, submetacêntricos e acrocêntricos.....	25
Figura 3.2 – Representação esquemática dos cromossomos metafásicos humanos de acordo com o tamanho e a posição do centrômero, identificados através dos 22 pares homólogos, além dos cromossomos sexuais X e Y, nos grupos A, B, C, D, E, F e G.....	29
Figura 3.3 – Comparação entre as bandas G e R do cromossomo número 1.....	33
Figura 3.4 – Representação idealizada dos cromossomos metafásicos humanos mostrando as bandas que podem ser obtidas com o auxílio de diferentes técnicas.....	34
Figura 3.5 – Representação visual dos pares de cromossomos, denominada de cariograma, pelos quais são distribuídos em ordem decrescente de tamanho.....	35
Figura 3.6 – Ilustração dos padrões de banda de um mesmo cromossomo: em (A) metáfase com 400 bandas e em (B) metáfase com 850 bandas.....	39
Figura 3.7 – Ilustração de uma metáfase contendo a translocação nos pares de cromossomos 15 e 17 – $t(15,17)$ , caracterizando a Leucemia Pró-Mielocítica Aguda.....	40
Figura 3.8 – Melhor visualização do prognóstico da Leucemia Pró-Mielocítica Aguda através de quatro exemplos da translocação – $t(15,17)$ .....	41
Figura 4.1 - Principais operações de um sistema de visão artificial de uso genérico.....	44
Figura 4.2 – Ilustração da estrutura funcional de um sistema de processamento e análise de imagens, através das etapas de transformação, pré-processamento, segmentação, descrição, reconhecimento e interpretação.....	46
Figura 4.3 – Máscaras de dimensão 3x3, 5x5 e 7x7 para o cálculo do Filtro de Média.....	48
Figura 4.4 – Ilustração de uma janela 3x3, destacando o pixel central $P_{ij}$ e seus <i>pixels</i> vizinhos.....	48
Figura 4.5 – Determinação do valor do <i>pixel</i> $Q_{ij}$ na nova imagem, definido pela mediana ( $V_5$ ) dos valores dos 9 <i>pixels</i> vizinhos de $P_{ij}$ .....	49

## LISTA DE FIGURAS - Continuação

Figura 4.6 – Ilustração do método de suavização: (A) imagem original de um cromossomo, (B) aplicação do filtro de média e (C) aplicação do filtro mediano.....	49
Figura 4.7 – Ilustração da operação de segmentação considerando o conceito de limiarização utilizando-se o histograma, onde (A) mostra a imagem de uma célula metafásica, (B) seu respectivo histograma com faixa de intensidade de corte em $T_1 = 0$ ( <i>Mín</i> ) e $T_2 = 200$ ( <i>Máx</i> ) e (C) a imagem segmentada.....	51
Figura 4.8 – Exemplo da decomposição de uma imagem colorida utilizando-se o modelo RGB: em (A) a imagem de uma célula metafásica, (B) decomposição das componentes deste modelo e em (C) os histogramas correspondentes de cada componente.....	52
Figura 4.9 – Operadores verticais e horizontais de <i>Sobel</i> e <i>Prewitt</i> aplicados para realçar e detectar bordas.....	53
Figura 4.10 – Técnica de detecção de bordas em uma imagem com 256 níveis de cinza: (A) utilizando operadores de Sobel vertical (B) e horizontal (C), e Prewitt vertical (D) e horizontal (E).....	53
Figura 4.11 – Máscaras para o cálculo do laplaciano: em (A) 3x3 e (B) 5x5.....	54
Figura 4.12 – Exemplo da aplicação de realce e detecção de bordas: (A) imagem monocromática e (B) operador laplaciano 3x3.....	54
Figura 4.13 – Exemplo da aplicação do operador laplaciano 3x3 para a detecção de bordas com fronteiras totalmente fechadas (detalhe do zoom). Em (A) uma imagem contendo cromossomos metafásicos e em (B) imagem resultante com a detecção das bordas e fronteiras totalmente fechadas.....	55
Figura 4.14 – Aplicação da técnica de sucessivas erosões e aberturas: (A) imagem de um cromossomo, (B) imagem binarizada, (C) após a remoção de sucessivas camadas e (D) resultado final do afinamento.....	57
Figura 4.15 – Processo de extração do eixo médio pelos centros de máximos discos.....	59
Figura 4.16 – Janela 3x3, destacando o <i>pixel</i> central (C) e seus <i>pixels</i> vizinhos.....	60

## LISTA DE FIGURAS - Continuação

- Figura 4.17 – Resultados obtidos com o processo de esqueletização utilizando o algoritmo Zhang-Suen: (A) imagem original, (B) esqueleto dos objetos (cromossomos) em que, no detalhe do zoom, é mostrada a limitação deste método em se obter esqueletos de um pixel de largura.....61
- Figura 4.18 – Máscaras utilizadas no método de *staircase removal*, onde o *pixel* central pode ser apagado se um dos valores  $X$  for igual a 0 (branco).....63
- Figura 4.19 – Resultados obtidos com o processo de esqueletização utilizando o algoritmo Holt: (A) imagem original, (B) esqueleto extraído sem o método *staircase removal* e (C) um esqueleto com um melhor afinamento aplicando-se o método *staircase removal*.....63
- Figura 5.1 – Ilustração do processo de modelagem matemática de problemas relacionados com a área da Medicina.....67
- Figura 5.2 – Esquema do processo interativo dos modelos médico e matemático para a formulação e formalização do problema, segundo SHULCLOPER *et al.* (1997).....68
- Figura 6.1 – Diagrama geral das etapas do processo de modelagem matemática para o problema proposto.....86
- Figura 6.2 – Diagrama geral da proposta de solução para o problema, caracterizado pelo desenvolvimento de uma nova sistemática para análise e classificação citogenética, que tem como objetivo minimizar o viés em citogenética clínica.....89
- Figura 6.3 – (A) Cromossomo do par número 17, (B) extração do esqueleto do cromossomo e (C) representação do esqueleto como um gráfico cartesiano.....98
- Figura 6.4 – Posições admissíveis para os sucessivos prolongamentos dos pontos da extremidade superior,  $P_{y-1,x-1}$ ,  $P_{y,x-1}$  e  $P_{y+1,x-1}$ .....103
- Figura 6.5 – Posições admissíveis para os sucessivos prolongamentos dos pontos da extremidade inferior,  $P_{y-1,x+1}$ ,  $P_{y,x+1}$  e  $P_{y+1,x+1}$ .....103

## **LISTA DE FIGURAS - Continuação**

- Figura 6.6 – (A) Cromossomo do par número 17, (B) extração do esqueleto do cromossomo e (C) prolongamento das extremidades utilizando-se uma curva ajustada pelos mínimos quadrados.....104
- Figura 6.7 – Exemplos do método de prolongamento das extremidades dos esqueletos, utilizando-se diversos cromossomos com formatos diferentes.....104
- Figura 6.8 – Janela de aquisição das células metafásicas provenientes do microscópio óptico, implementada no Sistema Computadorizado para Análise e Classificação Citogenética (SCACC).....106
- Figura 6.9 – Ilustração do processo de segmentação implementada ao SCACC, baseado na utilização do histograma do modelo RGB.....107
- Figura 6.10 – Ilustração do processo de segmentação da imagem de uma célula metafásica, com os cromossomos detectados apresentados em preto e suas respectivas bordas em vermelho, sobrepostos à imagem original.....108
- Figura 6.11 – Ilustração da janela principal do SCACC, com uma imagem resultante do processo de detecção de bordas e esqueletização dos cromossomos de uma célula metafásica.....106
- Figura 6.12 – (A) Cromossomo do par número 1 com fator de forma 0,209 e (B) cromossomo do par 21 com fator de forma 0,563.....110
- Figura 6.13 – Exemplo de gráfico de intensidade de nível de cinza, extraído do esqueleto de um cromossomo do par número 7, onde os picos representam as regiões escuras (bandas) e os vales representam regiões claras (regiões entre duas bandas).....111
- Figura 6.14 – Três cromossomos utilizados no cálculo do fator de semelhança para o reconhecimento automático dos mesmos. Os cromossomos (A) e (B) constituem o par número 7 e (C) um cromossomo do par número 14.....112
- Figura 6.15 – Ilustração da janela do cariótipo completo de um indivíduo do sexo feminino, apresentando os cromossomos classificados e pareados automaticamente.....114

## **LISTA DE FIGURAS - Continuação**

- Figura 6.16 – Janela que permite fazer a análise individual dos cromossomos, comparando-os com todos outros cromossomos detectados na metáfase. Também é apresentado o padrão de bandas cromossômicas pelo gráfico de intensidade de nível de cinza.....115
- Figura 6.17 – Detecção da Leucemia Pró-Mielocítica Aguda, disponibilizando um indicador do grau de semelhança com a doença.....116
- Figura 7.1 – Imagem de uma célula metafásica adquirida pelo sistema de aquisição do SCACC, que compõe o banco de dados utilizado nas simulações. Barra de escala 10  $\mu\text{m}$ .....120
- Figura 7.2 – Gráfico comparativo dos resultados obtidos dos experimentos com as variáveis área, fator de forma, distância *Euclidiana* e correlação linear de *Pearson*, testadas separadamente. Para cada experimento foram utilizados 276 cromossomos (138 pares homólogos).....124
- Figura 7.3 – Gráfico comparativo dos resultados obtidos com o 5<sup>o</sup>, 6<sup>o</sup> e 7<sup>o</sup> (quinto, sexto e sétimo) experimentos. Para cada experimento foram utilizados 276 cromossomos (138 pares homólogos).....126
- Figura 7.4 – Gráfico comparativo dos resultados obtidos com as simulações realizadas para a detecção da Leucemia Pró-Mielocítica Aguda (LMA). O eixo horizontal representa as imagens testadas, sendo que as 10 primeiras imagens são de células leucêmicas e as imagens restantes de células normais.....129
- Figura 7.5 – Ilustração de dois cromossomos do par número 9 que apresentam fatores que dificultam sua identificação: (A) apresenta tamanho menor e (B) possui as cromátides levemente separadas.....130
- Figura 7.6 – Identificação dos cromossomos homólogos do par número 7 utilizando-se a variável distância *Euclidiana*, quando na ocorrência de diferença de tamanho e intensidade de nível de cinza entre os homólogos. A identificação ficou na 15<sup>a</sup> ordem.....132

## **LISTA DE FIGURAS - Continuação**

Figura 7.7 – Comparação de um dos cromossomos do par 7 (objeto número 31 nessa figura) com outro cromossomo do par número 9 (objeto homólogo número 11 nessa figura). Devido o cromossomo do par número 7 apresentar-se menor e mais claro, torna-se menos semelhante ao seu homólogo (objeto homólogo número 46 nessa figura).....	133
---	-----

## LISTA DE TABELAS

Tabela 5.1 – Estrutura geral da metodologia do enfoque lógico-combinatório adotada neste trabalho.....	69
Tabela 5.2 – Apresentação dos conceitos de variáveis, domínios e critérios de comparação para a identificação clássica dos pares de cromossomos (ORTIZ-POSADAS <i>et al.</i> , 2003).....	74
Tabela 5.3 – Exemplo de uma matriz de aprendizagem ( $MA$ ), onde $R=\{x_1, x_2, x_3, x_4, x_5\}$ , $m=5$ , $r=2$ e $M=\{0,1\}$ .....	76
Tabela 5.4 - Matriz de aprendizagem do trabalho de ORTIZ-POSADAS <i>et al.</i> (2003), apresentando a descrição de 22 pares de cromossomos, além dos cromossomos sexuais X e Y, em termos da identificação clássica dos mesmos.....	76
Tabela 5.5 – Matriz $T$ formada pela descrição de 5 objetos distribuídos em duas classes ( $T_0$ , $T_1$ ), constatando-se a tipicidade em $\{x_2, x_3\}$ .....	78
Tabela 5.6 – Exemplo de alguns cálculos da matriz diferença $MD_{nm}$ do exemplo da tabela 5.4.....	79
Tabela 5.7 – Caracterização do conceito de testores através da análise do trabalho de ORTIZ-POSADAS <i>et al.</i> (2003) que, adotando o critério de identificação clássica de cromossomos, não se pode caracterizar o conjunto $\tau=\{x_1, \dots, x_7\}$ como um testor típico, pois existem objetos com as mesmas características em classes diferentes.....	80
Tabela 5.8 - Ponderações definidas pelo especialista para as sete características do problema de descrição dos cromossomos, apresentado segundo a tabela 5.4.....	79
Tabela 6.1 – Formalização do modelo matemático para a identificação automática dos cromossomos, descrito através de suas variáveis, seus domínios, critérios de comparação e ponderações.....	94
Tabela 6.2 – Valores ordenados dos pontos do gráfico cartesiano da figura 6.3.....	98
Tabela 6.3 – Valores da tabela 6.1 modificados com o ajuste para $z = \ln[f(x)]$ .....	101

## LISTA DE TABELAS - Continuação

Tabela 6.4 – Formalização do modelo matemático para a detecção da LMA descrito utilizando-se suas variáveis, seus domínios, critérios de comparação e ponderações.....	105
Tabela 6.5 – Medidas extraídas e calculadas pelo SCACC dos cromossomos A, B e C, que são utilizadas no cálculo do fator de semelhança.....	112
Tabela 7.1 – Tipos de imagens que compõem o banco de imagens deste trabalho e a procedência das lâminas das células metafásicas utilizadas para a captura das imagens.....	121
Tabela 7.2 – Resultados obtidos com o 1 <sup>o</sup> , 2 <sup>o</sup> , 3 <sup>o</sup> e o 4 <sup>o</sup> experimento, utilizando-se as variáveis: área ( <i>A</i> ), fator de forma ( <i>FF</i> ), distância <i>Euclidiana</i> ( <i>De</i> ) e correlação linear de <i>Pearson</i> ( <i>CLP</i> ). São apresentados os testes com 276 cromossomos e a ordem de identificação dos homólogos pelo SCACC.....	123
Tabela 7.3 – Resultados obtidos com o 5 <sup>o</sup> , 6 <sup>o</sup> e o 7 <sup>o</sup> experimento, utilizando-se as variáveis, área ( <i>A</i> ), fator de forma ( <i>FF</i> ), distância <i>Euclidiana</i> ( <i>De</i> ) e correlação linear de <i>Pearson</i> ( <i>CLP</i> ), e seus pesos adotados. São apresentados os testes com 276 cromossomos e a ordem de identificação dos homólogos pelo SCACC.....	125
Tabela 7.4 – Resultados obtidos com o 8 <sup>o</sup> (oitavo) experimento, com as suas variáveis utilizadas, área ( <i>A</i> ), fator de forma ( <i>FF</i> ), distância <i>Euclidiana</i> ( <i>De</i> ) e correlação linear de <i>Pearson</i> ( <i>CLP</i> ), seus pesos adotados, e o desempenho geral do SCACC com o número de cromossomos identificados corretamente e o número de cromossomos que não foram classificados corretamente.....	127
Tabela 7.5 – Resultados obtidos com as simulações para a detecção da Leucemia Pró-Mielocítica Aguda (LMA), apresentados como indicadores da doença. E as 10 imagens testadas foram de células leucêmicas.....	127
Tabela 7.6 – Resultados obtidos com as simulações para a detecção da Leucemia Pró-Mielocítica Aguda (LMA), apresentados como indicadores da doença e testados em 10 imagens de células normais.....	128

## **LISTA DE ABREVIATURAS**

UFSC	- Universidade Federal de Santa Catarina
CNPq	- Conselho Nacional de Desenvolvimento Científico e Tecnológico
SCACC	- Sistema Computadorizado para Análise e Classificação Citogenética
LMA	- Leucemia Pró-Mielocítica Aguda
ISCN	- <i>International System for Human Cytogenetic Nomenclature</i>
FDA	- <i>Food and Drug Administration</i>
ASG	- Giemsa Salinoacético
RNAs	- Redes Neurais Artificiais
CCD	- <i>Charge-Coupled Device</i>
MLP	- <i>Multi-Layer Perceptron</i>
RBF	- Função de Base Radial
MAKSYS	- <i>Missouri Automated Karyotyping System</i>
CSSG	- <i>Cross Section Sequence Graphs</i>
AKS	- <i>Automatic Karyotyping System</i>
IKAROS	- <i>Automatic Karyotyping System</i>
DNA	- Ácido Desoxirribonucléico
NaOH	- Hidróxido de Sódio
ERI	- Espaço de Representação Inicial
MD	- Matriz Diferença
CRT	- <i>Cathode Ray Tubes</i>
BEG/UFSC	- Laboratório de Biologia, Embriologia e Genética da UFSC
A	- Área
FF	- Fator de Forma
De	- Distância <i>Euclidiana</i>
CLP	- Correlação Linear de <i>Person</i>

## CAPÍTULO I

### INTRODUÇÃO

#### 1.1 – Introdução

Atualmente, tornou-se comum no estudo citogenético a utilização de determinados tipos de softwares que realizam o reconhecimento automático dos cromossomos para a composição do cariógrama. Em suma, o processo computadorizado apresenta inúmeras vantagens, porém, cabe lembrar que ainda existem alguns pontos não resolvidos neste processo que limitam a identificação automática correta dos cromossomos. Outro ponto também não resolvido é quanto à análise citogenética que é tradicionalmente realizada visualmente, sendo a experiência e a habilidade do profissional fatores determinantes para a correta classificação dos cromossomos. Esse fato resulta muitas vezes em avaliações com alto grau de subjetividade que pode levar a potencializar o viés da análise em citogenética clínica.

E neste trabalho, o que se propõe é a elaboração de uma sistemática, através de uma ferramenta computacional, com a finalidade de compor uma nova metodologia para a análise e classificação citogenética, visando superar ou minimizar estas limitações. Para tanto, houve a necessidade, além da aplicação de técnicas de processamento digital de imagens, do desenvolvimento de um novo método de prolongamento das extremidades do eixo médio extraído dos cromossomos, que permite a extração mais precisa das informações dos mesmos. Foi realizada também a formalização do problema sendo que, a técnica de modelagem matemática utilizada, fundamentou-se no enfoque lógico-combinatório da teoria de reconhecimento de padrões, fazendo possível a classificação dos cromossomos baseando-se no conceito de analogia. Estas técnicas, em conjunto, constituem o *Sistema Computadorizado para Análise e Classificação Citogenética*, elaborado para fins de reconhecimento automático dos cromossomos e também para servir de ferramenta que permite determinar as alterações específicas presentes na Leucemia pró-mielocítica aguda, oferecendo ao citogeneticista melhores condições para a detecção das alterações cromossômicas.

## 1.2 – Origem e Importância do Trabalho

A atividade de análise e classificação citogenética merece atenção prioritária em genética médica, isso porque numerosas condições patológicas são decorrentes de alterações numéricas e/ou estruturais dos cromossomos. O estudo dos cromossomos humanos é, então, um procedimento essencial em muitos campos de investigação que visa detectar alterações genéticas, lesões nos cromossomos decorrentes a fatores ambientais, diagnóstico de câncer, diagnósticos pré-natais, ou infertilidade/esterilidade humana.

Quando a atividade de análise e classificação citogenética, também denominada de *cariotipagem*, é realizada sem o auxílio de um computador, isto é, através de um processo manual, esta atividade exige extrema habilidade dos profissionais técnicos (citogeneticistas). Neste sentido, além de consumir muito tempo, a atividade acaba apresentando elementos significativos de natureza tediosa e repetitiva, o que a caracteriza, portanto, como muito cansativa. Com a possibilidade deste processo ser computadorizado, os citogeneticistas utilizam determinados tipos de softwares, denominados de *Sistemas Automáticos de Cariotipagem*, que realizam o reconhecimento e o pareamento automático dos cromossomos para compor o kariograma. A partir do kariograma torna-se exequível a verificação e o relato, pelo citogeneticista, das possíveis alterações cromossômicas. Além de automatizar os passos de identificação e pareamento dos cromossomos, os sistemas automáticos de kariotipagem apresentam ainda as seguintes vantagens:

- Em alguns casos, permitem melhor análise e interpretação da imagem do cromossomo, além de proporcionar muitas facilidades gráficas e interativas;
- Possibilitam maior agilidade na realização dos exames;
- Um maior número de metáfases podem ser analisadas de um mesmo paciente;
- Podem gerar exames impressos com maior qualidade e auto-explicativos;
- Diminui o tempo de trabalho do citogeneticista no microscópio, minimizando problemas ergonômicos;
- Permitem o armazenamento dos dados e do kariograma do paciente em um banco de dados para futuras análises.
- Possibilitam uma otimização do uso dos instrumentos ópticos (dispendiosos).

Como se pode constatar, estas vantagens dispõem melhores condições ao citogeneticista para realizar o processo de análise e classificação citogenética, tornando-o menos penoso do que pelo processo manual. Contudo, os sistemas automáticos de cariotipagem comerciais existentes ainda são, na verdade, semi-automáticos, pois apresentam muitas limitações na identificação automática dos cromossomos. Estas limitações surgem devido à existência de muitas fontes de incerteza no domínio do problema, inerentes, principalmente, ao processo de cultura das células metafásicas e ao preparo das lâminas que contêm tais células. No que diz respeito ao processo de cultura, apesar de ser padronizado, podem ocorrer variações de um laboratório para outro na quantidade de reagente utilizado e no tempo destinado à cultura. Já na questão da montagem e coloração das lâminas, os procedimentos são totalmente artesanais e o resultado de se ter lâminas com boa qualidade vai depender muito da habilidade e da experiência do profissional. Além desses dois fatores, os equipamentos utilizados, se não forem apropriados, também podem contribuir como fontes de incerteza, levando a grandes variações nas características que descrevem os cromossomos que comprometem, conseqüentemente, a correta identificação dos cromossomos pelos sistemas automáticos. Com isso, há a necessidade de alguma intervenção humana para a composição do cariograma livre de erros.

Buscando superar e/ou contornar os problemas relacionados com as fontes de incerteza, ao longo de mais de 20 anos, cientistas do mundo inteiro passaram a trabalhar em prol da elaboração de inúmeras técnicas computacionais, que permitissem a identificação automática correta dos cromossomos sem a necessidade de qualquer tipo de intervenção humana. Em um primeiro momento, técnicas de processamento digital de imagens foram desenvolvidas e aperfeiçoadas visando a extração mais precisa das características dos cromossomos.

Alguns anos mais tarde, outro fato que também passou a despertar interesse dos cientistas foram os métodos de classificação, sendo que muitas pesquisas voltaram-se para a aplicação de sistemas especialistas, baseados principalmente em redes neurais artificiais, levando a uma expressiva melhoria no desempenho geral dos classificadores (TODESCO, 1995). Doravante, os sistemas automáticos de cariotipagem alcançaram um estágio de aceitação pelos laboratórios de genética humana como ferramentas clinicamente úteis em sua rotina. Mas, apesar de todo esse avanço, os citogeneticistas consideram que o uso

desse sistemas especialistas constitui, ainda, um ponto não totalmente resolvido pelos sistemas automáticos de cariotipagem, apresentando, de modo geral, dois principais inconvenientes: o primeiro é a necessidade do operador ter de armazenar, na base de dados do sistema, um número expressivo de padrões pré-estabelecidos de imagens de cromossomos, que são de rotina do laboratório em questão. Isso ocorre porque as características que descrevem os cromossomos apresentam muitas variações de um laboratório para outro e a precisão do processo de classificação dos cromossomos nestes sistemas torna-se proporcional à quantidade de objetos que foram incluídos na base de dados. E neste caso, nos laboratórios onde se utilizam estes sistemas especialistas, os citogeneticistas consideram que tais sistemas somente conseguem alcançar o estágio de classificação correta dos cromossomos após um período de 2 ou 3 anos de uso.

O segundo inconveniente dos sistemas especialistas utilizados na análise e classificação citogenética é o fato do resultado da classificação ser apresentado apenas em duas alternativas: certo ou errado (classificado e não-classificado). Nesta atividade, como em outras da área das ciências pouco formais, os profissionais freqüentemente são levados a utilizar o conceito de analogia como ferramenta metodológica na caracterização dos problemas. Este conceito de analogia é definido pelo grau de semelhança ou similaridade entre os objetos de estudo e baseia-se no conhecimento adquirido, experiência e observações. Isso significa que a forma como são apresentados estes resultados pelos sistemas especialistas difere muito da forma com que os citogeneticistas identificam os problemas clínicos.

Um outro ponto também considerado pelos citogeneticistas como não solucionado pelos sistemas automáticos de cariotipagem diz respeito aos processos de análise clínica, que ainda ficam, única e exclusivamente, dependentes dos profissionais técnicos. Em citogenética humana a análise do objeto de estudo é feita baseando-se em variáveis *qualitativas* e *quantitativas*, simultaneamente, logo, a presença da subjetividade torna-se marcante nas avaliações clínicas. Um exemplo deste fato pode ser constatado na análise da *Leucemia pró-mielocítica aguda* e que, devido ao alto grau de subjetividade em sua análise, vem a ser prática comum nos laboratórios de citogenética humana a corroboração de vários citogeneticistas para a confirmação desta patologia.

O que se quer mostrar com isso, é que a presença da subjetividade nas avaliações clínicas pode acarretar, em certas ocasiões, equívocos que dificultam a avaliação eficiente

de um determinado problema. Estes equívocos são intrínsecos à toda sistemática desta atividade, levados por indecisões, incertezas, contradições, interpretações diversas, somadas também às fontes de incerteza no domínio do problema que, em geral, induzem à distorções no diagnóstico clínico. De forma prática, é comum o uso do termo ao viés<sup>1</sup> (*bias*) quase como sinônimo de distorções na execução, na análise ou na interpretação de problemas clínicos, como por exemplo, no estudo em citogenética clínica.

Diante dessas limitações apresentadas pelos sistemas automáticos de cariotipagem, buscam-se novas soluções para a atividade de análise e classificação que, de maneira geral, ofereçam melhores condições aos citogeneticistas para reduzir o viés do diagnóstico clínico, permitindo, portanto, avaliações mais eficientes e confiáveis.

## 1.3 – Objetivos

### 1.3.1 – Objetivo Geral

A proposta geral deste trabalho é o desenvolvimento de uma sistemática (método), denominada de *Sistema Computadorizado para Análise e Classificação Citogenética* (SCACC), tendo como finalidade compor uma nova metodologia que permite reduzir o viés do diagnóstico clínico em citogenética humana. Esta sistemática baseia-se na aplicação de técnicas de processamento digital de imagens em conjunto com o enfoque lógico-combinatório da teoria de reconhecimento de padrões.

### 1.3.2 – Objetivos Específicos

O *Sistema Computadorizado para Análise e Classificação Citogenética* proposto foi elaborado visando dois objetivos específicos: o primeiro diz respeito ao reconhecimento automático dos cromossomos através de uma única cena, e o segundo ao

---

<sup>1</sup> Adaptado do conceito de *Murphy*, o termo viés foi descrito por *David L. Sackett* em 1979 como "qualquer processo em qualquer estágio de inferência o qual tende a produzir resultados ou conclusões os quais diferem sistematicamente da verdade" (LOPES, 2002).

desenvolvimento de uma ferramenta para análise da Leucemia pró-mielocítica aguda, conforme detalhados a seguir:

1- *Elaborar um método automático de reconhecimento de cromossomos através de uma única cena*: a meta deste primeiro propósito é a identificação automática dos cromossomos homólogos para compor o kariograma. Apesar dos sistemas de cariotipagem comerciais já cumprirem com este propósito, neste estudo busca-se verificar a possibilidade da realização do reconhecimento automático dos cromossomos através de uma única cena (metáfase), sem precisar comparar entidades específicas com padrões previamente armazenados e classificados. Necessita-se, para tanto, utilizar e desenvolver técnicas avançadas de processamento digital de imagens para a extração mais precisa das principais características que descrevem os cromossomos através do tamanho, forma e padrões de banda. E, em conjunto, aplicar o enfoque lógico-combinatório para o reconhecimento automático dos cromossomos homólogos baseando-se no grau de similaridade entre os objetos comparados.

2- *Desenvolver uma ferramenta para análise da Leucemia pró-mielocítica aguda*: este segundo propósito tem como finalidade apresentar um indicador do grau de semelhança desta doença, dispondo ao citogeneticista uma forma de análise quantitativa da translocação cromossômica que caracteriza tal doença. Esta translocação cromossômica ocorre na porção final dos braços dos cromossomos e, para que seja possível extrair suas informações, necessita-se desenvolver um novo método de prolongamento das extremidades do eixo médio dos cromossomos. Para a formalização das particularidades deste problema, utilizar o enfoque lógico-combinatório como método de modelagem matemática com a finalidade de “traduzir” o conhecimento do especialista e criar uma função de analogia que permita extrair o indicador da patologia.

#### **1.4 – Hipótese do Trabalho**

Espera-se que este novo método (sistemática) possa ser utilizado na realização da atividade de análise e classificação citogenética, não somente como uma ferramenta que permita tornar o processo de reconhecimento dos cromossomos mais ágil, mas também, que elimine o inconveniente dos profissionais terem de incluir, na base de dados, padrões

pré-estabelecidos de imagens de cromossomos, para que o sistema possa classificá-los corretamente. Além disso, espera-se que este novo método se torne uma ferramenta mais adaptada à realidade clínica, disponibilizando resultados na forma como os citogeneticistas analisam e avaliam os problemas. Por fim e em seu ponto mais forte, que possa oferecer melhores condições que permitam reduzir a subjetividade na análise da doença da Leucemia pró-mielocítica aguda, resultando em avaliações clínicas mais eficientes e confiáveis, sem a necessidade da corroboração de outros profissionais para a confirmação da doença.

### 1.5 – Metodologia

A busca de novas e melhores soluções para tentar superar pontos ainda não resolvidos dos sistemas atuais de cariotipagem motiva o desenvolvimento de técnicas avançadas de processamento digital de imagens para a criação de novos modelos computacionais, cujo desempenho tende a aproximar-se o máximo possível das capacidades e sensibilidades do sistema visual humano. Mas, para a elaboração de uma ferramenta computacional que seja eficiente para a finalidade de análise e classificação citogenética, torna-se indispensável também aplicar métodos de modelagem matemática que possam representar os problemas práticos desta atividade. Isso porque a modelagem matemática destes problemas permite a obtenção de regras de decisão que possam ser aplicadas por modelos computacionais, fazendo-se possível o reconhecimento automático dos padrões dos objetos de estudo.

Desse modo, o *Sistema Computadorizado para Análise e Classificação Citogenética* proposto foi elaborado através da implementação das principais operações que um sistema de visão artificial efetua sobre uma imagem. Para algumas destas operações, técnicas simples de processamento digital de imagens resolveram muito bem o problema. Entretanto, para outras operações, foram requeridas técnicas mais complexas, em especial à finalidade de extração dos atributos, pois necessitava-se que a detecção das características que descrevem os cromossomos fosse mais precisa. Tais atributos correspondem à área, perímetro e eixo médio dos cromossomos, sendo que, a partir da área e do perímetro, calculou-se a característica denominada de fator de forma. Para a extração do eixo médio, utilizou-se um algoritmo baseado no método de *Holt*, em conjunto com a

técnica de *Staircase Removal*, conseguindo-se esqueletos com afinamento de apenas um pixel de largura e sem perder a conexão do segmento (FACON, 2003b). Através do esqueleto dos cromossomos, foi possível obter, da posição de cada *pixel* pertencente ao esqueleto, o valor da intensidade de nível de cinza ao longo dos cromossomos. Estes valores de intensidade de nível de cinza foram plotados em um gráfico, onde os picos no gráfico representam as bandas (regiões escuras) e os vales representam as regiões claras (regiões entre as bandas). O reconhecimento dos cromossomos por intermédio das bandas pode ser realizado, então, utilizando-se um indicador de similaridade entre as mesmas através do conceito de distância *Euclidiana*, e também por intermédio da quantificação do grau de associação das bandas, aplicando-se o método de correlação linear de *Pearson*.

De posse das características que descrevem os cromossomos, o passo seguinte compreende o reconhecimento automático dos cromossomos que, para isso, optou-se pela utilização do enfoque lógico-combinatório da teoria de reconhecimento de padrões. Este enfoque permite fazer uma representação formal do conhecimento do especialista, elaborando um modelo matemático que permite a classificação dos cromossomos por conceito de analogia. O uso da lógica-combinatória tem, como principal vantagem, a identificação dos cromossomos baseando-se no grau de semelhança ou similaridade entre os objetos, o que torna o sistema adaptado à forma como os profissionais das áreas das ciências pouco formais analisam o problema.

Assim, buscando atingir o primeiro objetivo específico proposto neste trabalho, que consiste no reconhecimento automático dos cromossomos através de uma única cena, foram realizados, primeiramente, quatro experimentos utilizando-se as características área, fator de forma, distância *Euclidiana* e correlação linear de *Pearson*, separadamente. Para cada uma das quatro características, foram realizadas 276 comparações, ou seja, foram testados 138 pares de cromossomos (6 metáfases) de laboratórios distintos. Nestes experimentos, as características foram utilizadas separadamente com o intuito de identificar aquelas que fossem menos suscetíveis às fontes de incerteza. Em seguida, outros três experimentos foram realizados aplicando-se, agora, as características em conjunto. Contudo, para a formação destes conjuntos, optou-se pela seleção das características, que apresentaram desempenhos melhores nos primeiros experimentos, além de atribuir pesos diferenciados a elas. Estes novos experimentos também foram realizados utilizando-se 276 cromossomos e os melhores resultados foram obtidos quando utilizadas

as seguintes características e seus respectivos pesos: área com peso 2, fator de forma com peso 1, distância *Euclidiana* com peso 3 e correlação linear de *Pearson* com peso 1. O índice de acerto foi de aproximadamente 63,04% e, embora este resultado seja relativamente baixo, equipara-se, na mesma faixa dos resultados obtidos, com outros trabalhos encontrados na literatura como, por exemplo o trabalho de JOHNSTON *et al.* (1993), que aplicou sistemas especialistas na classificação de todos os cromossomos da metáfase e também realizou seus testes com imagens de cromossomos de laboratórios distintos. Outros trabalhos tiveram resultados melhores, porém seus testes foram realizados utilizando-se pares específicos de cromossomos e/ou com imagens de apenas um laboratório em questão, tais como os trabalhos de LUNDSTEEN *et al.* (1991), DENISOV e DUDKIN (1994), TODESCO (1995) e SJAHPUTERA e KELLER (1999).

Como análise final para o primeiro propósito do *Sistema Computadorizado para Análise e Classificação Citogenética*, considera-se que o método de reconhecimento automático de cromossomos através de uma única cena é totalmente possível, tendo-se como expectativa que o desempenho de identificação possa ser ainda melhorado com o controle de determinados fatores técnicos, tais como a obtenção, através de preparações metafásicas adequadas, metafases com maior resolução e com bandas mais nítidas, e também com o uso de equipamentos ópticos de captação de imagem mais adequados.

Para o segundo propósito da nova sistemática, desenvolveu-se a ferramenta para a análise da Leucemia pró-mielocítica aguda, pela qual houve a necessidade que se identificasse o rearranjo cromossômico decorrente da translocação  $t(15,17)$ . Para tanto, foi preciso realizar a extrapolação dos pontos do esqueleto, ou seja, realizar o prolongamento das extremidades do eixo médio dos cromossomos, até encontrar suas fronteiras. Este método de extrapolação dos pontos do esqueleto foi elaborado baseando-se no ajuste de curvas pela técnica dos mínimos quadrados, em que a idéia foi de tentar encontrar uma curva que melhor se aproximasse ao gráfico do esqueleto, para que pudesse ser feita a extrapolação com uma certa margem de segurança.

Após a realização deste processo de prolongamento das extremidades, foi possível, então, a extração mais precisa das características que descrevem a Leucemia pró-mielocítica aguda, em imagens de cromossomos do par número 17. A determinação da presença do rearranjo cromossômico foi feita baseando-se no fator de semelhança e os resultados apresentados, através de um indicador percentual de similaridade com a

alteração cromossômica característica com a doença. Assim, a detecção desse rearranjo passa a ser feita por meio de uma variável quantitativa, minimizando a subjetividade da avaliação visual. Os resultados obtidos apresentaram indicadores de 63,73 a 99,70% de semelhança com o rearranjo cromossômico, em experimentos com 10 imagens de células leucêmicas e valores abaixo de 17,30%, utilizando-se 10 imagens de células normais. Esses valores mostram uma nítida distinção (distância) entre as classes de células normais e células leucêmicas (com rearranjo), demonstrando um bom desempenho desta nova metodologia. Como conclusão deste segundo propósito, com esse novo método, é possível obter uma ferramenta que permite detectar com maior precisão o rearranjo cromossômico que caracteriza a Leucemia pró-mielocítica aguda, oferecendo ao citogeneticista condições para a realização do diagnóstico clínico com maior eficiência e confiabilidade.

## 1.6 – Apresentação da Tese

O capítulo II será dedicado a um breve histórico dos sistemas automáticos de cariotipagem e seus métodos aplicados, apresentando também os principais pontos não resolvidos por esses sistemas. O capítulo III contém os tópicos referentes à citogenética humana, dando maior ênfase aos tópicos que fazem parte do processo de identificação dos cromossomos, às fontes de incerteza no domínio do problema e também da descrição da Leucemia pró-mielocítica aguda. No capítulo IV, busca-se trazer um caráter descritivo das principais operações de processamento digital de imagens que um sistema de visão artificial deve conter, com o objetivo de conferir o embasamento teórico para a aplicação das técnicas utilizadas para o desenvolvimento da sistemática proposta neste trabalho. Já o capítulo V abrange o conceito sobre a metodologia de modelagem matemática, fundamentada no enfoque lógico-combinatório da teoria de reconhecimento de padrões. Para o capítulo VI, será apresentado o desenvolvimento passo a passo da sistemática proposta, através do detalhamento das técnicas utilizadas na implementação do *Sistema Computadorizado para Análise e Classificação Citogenética*. Já no capítulo VII, são expostas a aplicação do modelo proposto, a apresentação dos resultados obtidos e a análise e interpretação dos resultados. E, por fim, no capítulo VIII, são apresentadas as conclusões finais, contribuições e sugestões para futuros trabalhos.

## **CAPÍTULO II**

### **ESTADO DA ARTE**

#### **2.1 – Introdução**

A análise e classificação citogenética computadorizada tem alcançado o estágio em que muitos laboratórios já aceitam este processo, tanto como clinicamente útil, quanto economicamente justificável. Encontram-se hoje disponíveis muitos sistemas comerciais, proporcionando facilidades gráficas e iterativas para a análise e classificação citogenética, embora havendo a necessidade, algumas vezes, da intervenção do operador para a interpretação de resultados incomuns. Logo, durante os últimos 30 anos, vários experimentos e rotinas de análise e classificação automática de cromossomos têm sido estudados, buscando aperfeiçoar os métodos já existentes e também criando novas técnicas para superar algumas limitações do processo computadorizado.

Com o intuito de conferir um breve histórico dos sistemas automáticos de cariotipagem, suas técnicas e métodos aplicados, neste capítulo serão abordados os seguintes tópicos: sistemas automáticos para análise e classificação citogenética, histórico das técnicas utilizadas pelos sistemas de cariotipagem e, por fim, serão apresentadas as considerações finais sobre o exposto.

#### **2.2 – Sistemas Automáticos para Análise e Classificação Citogenética**

Até a década de 70, a classificação dos cromossomos era feita tomando-se como base somente os atributos tamanho e posição do centrômero. A partir de 1970, foram desenvolvidos processos de coloração dos cromossomos que resultaram na visualização das bandas cromossômicas, possibilitando a caracterização morfológica de cada um dos 24 tipos de cromossomos humanos. Dessa metodologia, foi possível, então, a análise mais precisa de todos os cromossomos metafásicos.

Paralelamente ao desenvolvimento técnico, foram realizadas várias Conferências Internacionais, buscando uniformizar os critérios de análise e propondo uma nomenclatura internacional comum (*An International System for Human Cytogenetic Nomenclature* ou ISCN), com o objetivo de sistematizar e padronizar a classificação dos cromossomos humanos (ISCN, 1995).

Na busca de critérios quantitativos para a avaliação das bandas cromossômicas, *Máximo E. Drets*, em meados da década de 70, foi um dos pioneiros em desenvolver um sistema que avaliava os padrões das bandas-G, ao longo dos braços dos cromossomos, através da análise de curvas geradas por um sistema microdensitométrico de alta resolução (DRETS, 2002). O processo de microdensitometria em fotografias (slides ou negativos) foi inicialmente desenvolvido para localizar a região do centrômero e, posteriormente, utilizado na determinação dos padrões de bandas cromossômicas.

Com a evolução dos sistemas quantitativos e lógicos de análise cromossômica, para obterem a autorização de comercialização, tais sistemas automáticos de cariotipagem necessitam-se adequar às normas prescritas em cada país, em órgãos competentes de fiscalização. Na falta de normas locais, é permitido por estes órgãos fiscalizadores que os fabricantes sigam as normas prescritas pela agência reguladora norte-americana *Food and Drug Administration* (FDA). Essa agência prescreve um critério de avaliação destes dispositivos sugerindo e orientando os fabricantes a desenvolverem sistemas contendo, no mínimo, alguns requisitos necessários aos padrões atuais de prática clínica laboratorial (FDA, 1991).

Em caráter informativo, etapas básicas em uma rotina de análise citogenética utilizando-se um sistema de cariotipagem consistem em, inicialmente, localizar uma célula metafásica através de um microscópio ótico. Se, por decisão do citogeneticista, a célula metafásica encontrada for considerada adequada para a classificação dos cromossomos, a imagem desta célula é capturada através de uma câmera de vídeo acoplada ao microscópio e enviada para um computador. Passa-se, então, para a fase de classificação automática dos cromossomos para, então, formar o kariograma. Através da elaboração completa do kariograma da célula, obtêm-se informações relevantes para o diagnóstico clínico. A Figura 2.1 ilustra um exemplo de sistema automático de cariotipagem, através de seus componentes de uso genérico: sistema de aquisição de imagens, processamento, saída e armazenamento.

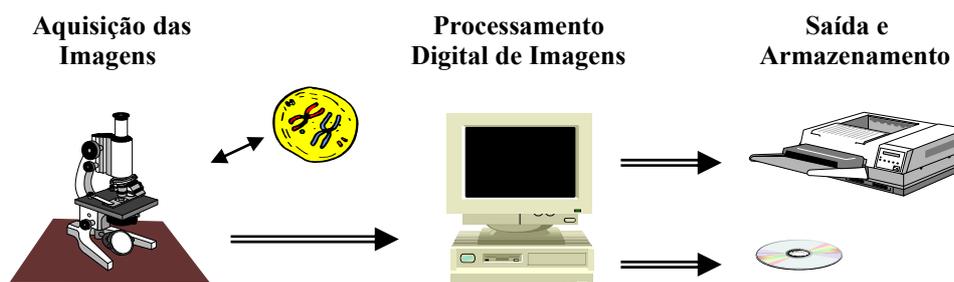


Figura 2.1 – Componentes de uso genérico em um sistema de cariotipagem: sistema de aquisição das imagens, processamento, saída e armazenamento.

Entretanto, mesmo com todos os esforços desempenhados no desenvolvimento de um sistema totalmente automático, sempre existe alguma limitação neste processo que o impossibilite da correta identificação individual dos cromossomos. Devido a estas limitações estes sistemas são denominados de sistemas semi-automáticos, pois sempre requerem a intervenção humana quando o sistema encontra alguma dificuldade ou impossibilidade de classificação. Problemas na identificação dos cromossomos com estes sistemas são muito comuns de ocorrer, levando muitos cientistas a trabalharem em prol do aperfeiçoamento destes sistemas visando superar ou mesmo minimizar tais limitações.

### 2.3 – Histórico das Técnicas Utilizadas pelos Sistemas de Cariotipagem

Com o advento de novas tecnologias para o desenvolvimento de algoritmos, muitos sistemas automáticos de cariotipagem estão sendo disponibilizados, tornando-se uma ferramenta poderosa para o auxílio na avaliação citogenética. Logo, possibilitou-se alternativas ao método tradicional de fotografar as imagens das células metafásicas através de uma câmera fotográfica acoplada ao microscópio, e da montagem do cariótipo a partir de recortes manuais das fotografias dos cromossomos (justificando a denominação de processo manual), separando e colando os pares homólogos em uma folha de papel para compor o cariograma.

Assim, com o intuito de iniciar o processo de quantificação da análise e classificação citogenética, *Ledley*, no ano de 1964, foi o primeiro pesquisador a realizar um experimento publicado de classificação automática de cromossomos, citado com poucas

informações científicas no trabalho de LUNDSTEEN e PHILIP (1989). Novas e grandes contribuições para a era da quantificação e automação citogenética vieram com os trabalhos de *Rutovitz* em 1970 e depois com *Neurath* em 1972. Com a descoberta das bandas cromossômicas, surgiu o protótipo desenvolvido por *Castleman*, em 1976, que podia produzir um cariótipo com a ajuda de um operador, necessitando que o mesmo contasse as bandas dos cromossomos e informasse ao sistema. Entretanto, devido a este protótipo ser muito lento, sua utilização era limitada. Por volta de 1980, surgiu o primeiro sistema comercialmente disponível para ser utilizado em um laboratório de citogenética clínica, denominado de WOLFSON UNIT e desenvolvido pela Universidade de Manchester - Inglaterra. Também não se pode deixar de mencionar o sistema MAGISCAN da Joyce-Loebl, que foi instalado no laboratório em Rigshospitalet – Copenhagen, no ano de 1982 e, depois de inúmeros testes, o sistema pôde ser disponibilizado para uso no diagnóstico clínico em 1983. A partir daí, sucessivas modificações foram feitas neste sistema, possibilitando que os exames fossem realizados ainda com maior rapidez e, além de oferecer uma interface mais “amigável”, o sistema poderia ser disponibilizado para um número maior de laboratórios clínicos (LUNDSTEEN e PHILIP, 1989).

Outro sistema computadorizado para análise e classificação citogenética foi o sistema CYGEN1, desenvolvido por BRUSCHI *et al.* (1981). Esse sistema tinha como objetivo fazer a identificação dos cromossomos, entretanto, o reconhecimento das características das bandas era feito pelo operador que, posteriormente, passava as informações ao sistema para a identificação. A margem de erros desse sistema estava em torno de 10 a 25%, podendo ser processadas 20 metáfases no tempo de 40 a 45 minutos.

Na busca de novos métodos para a classificação dos cromossomos, PIPER (1986) publicou um trabalho em que relatava a identificação de cromossomos através da técnica de máxima probabilidade. A lógica deste trabalho baseava-se na existência de um processo de extração (medida) das características dos cromossomos através de classes de tamanhos fixos, que pudesse resultar na probabilidade de um cromossomo particular pertencer a uma das 24 classes representativas de cromossomos. Logo, a máxima probabilidade de cada cromossomo pertencer a uma classe poderia ser utilizada na classificação, sem necessitar referenciar outro cromossomo na célula. Utilizou-se uma

estratégia de combinação de classificadores já existentes, porém, conseguindo-se pequenas melhorias de classificação em comparação com outros métodos da época.

Mas o sistema de cariotipagem que ganhou mais prestígio pelos laboratórios de citogenética clínica na época foi o sistema semi-automático comercialmente conhecido como ATHENA, desenvolvido por Van VLIET *et al.* (1990). Estes pesquisadores estiveram envolvidos por mais de 15 anos na área de análise automática de imagens digitalizadas de cromossomos metafásicos. Neste sistema, o processo de segmentação das imagens dos cromossomos metafásicos individualizados é automático, assim como os cálculos das bandas cromossômicas e também o processo de classificação. O ponto forte desse trabalho consiste na apresentação da técnica desenvolvida para a segmentação, o método de *correção de sombreamento (shading correction)*, que consistia na combinação das técnicas de filtragem através do limiar de nível de cinza e operações morfológicas binárias de erosão e dilatação. Com este método, era possível separar os cromossomos que se apresentavam ligeiramente encostados.

Contudo, o sistema ATHENA apresentava certas limitações nos processos de segmentação, além de ter dificuldades no processo de localização dos centrômeros e também na classificação, principalmente com os cromossomos do grupo F e G (impossibilidade de reconhecimento). Mas, diante de todas estas limitações, permitia a interação do usuário em cada uma destas etapas, o que justificava a denominação de sistema semi-automático, permitindo ao operador buscar um melhor desempenho na cariotipagem de um modo geral (MAYALL *et al.*, 1990).

Já com o intuito de melhorar a eficiência do sistema de cariotipagem MAGISCAN, os pesquisadores LUNDSTEEN *et al.* (1991) modificaram tanto o software quanto os processos do sistema. Com estas modificações, foi possível a análise de diversas células ao mesmo tempo e a interação com o operador somente passou a ser necessária na etapa de localização das células metafásicas ao microscópio. O processo de cariotipagem é realizado com a escolha de 10 metáfases, sendo que cada cariótipo é representado por uma coluna de 23 pares de cromossomos (24 linhas que representavam os 22 pares de cromossomos homólogos além dos cromossomos X e Y) e as anormalidades cromossômicas são, então, detectadas por comparação através dos pares homólogos. Porém, os cromossomos sobrepostos, encostados, muito tortos ou anormais e aqueles em que não existe a possibilidade de classificação através do software são, então, separados e

apresentados para o operador para a classificação interativa. Segundo os autores, de todos os casos analisados, o sistema conseguiu classificar corretamente 66% dos pares homólogos. Neste estudo, foram utilizadas imagens de cromossomos capturadas de apenas um laboratório de citogenética, ou seja, de um procedimento único de cultura e obtenção de células metafásicas.

Tendo em vista os trabalhos pesquisados até este período e analisando os resultados obtidos, notou-se que uma primeira preocupação era com o desenvolvimento de métodos mais eficientes para a detecção e extração das características morfológicas dos cromossomos. A partir de então, muitos pesquisadores também passaram a dar importância para o processo de classificação dos cromossomos com a aplicação de sistemas especialistas baseados em modelos convencionais da teoria de reconhecimento de padrões. Estes modelos convencionais constituem classificadores paramétricos e não-paramétricos, normalmente utilizando técnicas de probabilidade e estatística. Várias outras técnicas também passaram a serem utilizadas, tais como: algoritmos de busca em grafos, seqüências de transição de bandas, modelos de redes de Markov, algoritmos de transportes, Lógica-Fuzzy e, com maior aplicação, as redes neurais artificiais (RNAs).

No trabalho de TSO *et al.* (1991), por exemplo, implementou-se a técnica denominada de Algoritmo de Transporte, caracterizada por um método de programação linear que fornece uma solução global ótima de mínimo custo. Este método permite obter a priori o número de cromossomos de uma determinada célula antes da realização da classificação propriamente dita. O estudo envolveu a classificação de descrições de três bases de dados de células metafásicas coradas com banda-G: o *Philadelphia*, obtido no *Jefferson Medical College* (Philadelphia) e contendo imagens de células provenientes de vilosidade coriônicas, o *Copenhagen*, obtido no *Rigshospitalet* (Copenhagen) e consistindo de imagens de células de sangue periférico, e o *Edinburgh*, obtido de amostras rotineiras no laboratório clínico *MRC* (Edinburgh) e contendo imagens de células de sangue periférico de indivíduos masculinos. Os testes foram realizados com a utilização das bases de dados separadamente, sendo que os melhores resultados foram obtidos com a base *Copenhagen*, aplicando-se a técnica de máxima probabilidade e aproximação Bayesiana, com 20% das células analisadas tendo todos os cromossomos identificados.

Com o advento das RNAs, novos classificadores de padrões surgiram e novos algoritmos baseados nestes sistemas especialistas foram aplicados à identificação

automática de cromossomos. Um exemplo destes novos classificadores foi apresentado por JOHNSTON *et al.* (1993), que aplicou um algoritmo para identificar os cromossomos através das características das bandas. Estas características eram descritas pela posição, intensidade (extraída através de um limiar densidade óptica de nível de cinza), comprimento e largura das bandas. Na técnica do limiar de densidade óptica, as regiões selecionadas são, então, processadas através de um filtro derivativo de segunda-ordem de Laplace ao longo do eixo principal, tendo a vantagem de detectar os picos e os vales nas imagens com valores de níveis de cinza. Já o algoritmo de classificação foi baseado na máxima probabilidade da distribuição normal multivariada do vetor característico e testado através de duas bases de dados de imagens de cromossomos capturadas através de câmeras CCD 1250x1024 de resolução: *Copenhagen* (179 metáfases) e da Universidade do Texas – *M.D. Anderson Cancer Center* (50 metáfases). Neste estudo, ficou comprovada a excelente qualidade da base de dados *Copenhagen*, conseguindo-se os melhores resultados comparados às outras bases e com índice de acerto de 94,6%. Entretanto, quando os testes eram realizados com as duas bases de dados, *Copenhagen* e *Anderson*, o índice caía para 87,4%. E, quando se testava o classificador com mais duas outras bases de dados (*Edinburgh* e *Philadelphia*), o índice ficava em torno de 60 e 80%.

Também com aplicações em RNAs, ERRINGTON e GRAHAM (1993) publicaram um trabalho de classificação automática de cromossomos através de uma rede *Multi-Layer Perceptron* (MLP), treinada com o algoritmo *Backpropagation*. Utilizou-se como entrada da rede as características de tamanho, índice centromérico e a representação gráfica das bandas cromossômicas. O desempenho do algoritmo na classificação foi favorável comparado aos classificadores paramétricos desenvolvidos em outros trabalhos. A melhoria no desempenho da classificação apresentou índices de 22%, 25% e 49% para as bases de dados de *Copenhagen*, *Edinburgh* e *Philadelphia*, respectivamente, entretanto, segundo os próprios autores, estas melhorias não foram significativas comparadas aos resultados esperados por eles.

Outro trabalho desenvolvido em análise e reconhecimento de cromossomos foi apresentado por DENISOV e DUDKIN (1994), tratando-se de um método que incluía as características de tamanho, índice centromérico e intensidade de nível de cinza para serem analisadas. Um esquema computacional multiestágio de decisão com base na hipótese de construção e verificação foi considerado, testando-o em 20 células metafásicas de um

mesmo laboratório com cromossomos preparados com banda-G. Os resultados obtidos estavam na média em comparação aos outros trabalhos da época, na razão de 66,9-90,32%.

No ano seguinte, também utilizando-se RNAs através da técnica *Multi-Layer Perceptron* (MLP), LERNER *et al.* (1995) implementaram um algoritmo para classificação de cromossomos, juntamente com a técnica de extração de característica baseando-se no método denominado de transformada do eixo médio. Foram estudadas duas aproximações dessa transformada do eixo médio, uma através do processo de *esqueletização* e outra baseada na técnica de *piecewise linear*. Essa primeira resultou em um eixo médio melhor, bem como um melhor desempenho na classificação. As características de intensidade ao longo do eixo foram extraídas e testadas, e o índice de acerto nos testes realizados foi de 98%, através da técnica de esqueletização e 97% com a aproximação *piecewise linear*.

Uma nova classe de classificadores utilizando RNAs também foi apresentada por TODESCO (1995). Neste trabalho, foi realizado um estudo com aplicação de uma função de base radial (RBF) para a classificação de cromossomos em 24 classes. Para o treinamento da rede e a classificação, também foram utilizadas as três bases de dados de células metafásicas mais conhecidas pelo meio científico: *Copenhagen* (180 metáfases padrão banda-G), *Edinburgh* (125 metáfases padrão banda-G) e *Philadelphia* (130 metáfases padrão banda-G). Neste estudo, foram exploradas diferentes combinações de características extraídas de cromossomos, bem como diferentes topologias e parâmetros para a rede. Os resultados obtidos foram comparados a outros trabalhos desenvolvidos, com uma melhoria de 25% na classificação, utilizando-se a base de dados *Copenhagen*, nenhuma melhoria significativa para a base *Philadelphia* e uma melhoria de 28,5% utilizando-se a base *Edinburgh*. Este estudo comprovou que essa classe de RNA fornecia um caminho alternativo com uma taxa de erro de classificação menor.

No cenário dos sistemas automáticos de cariotipagem, também merece destaque o sistema *Missouri Automated Karyotyping System* (MAKSYS), desenvolvido por Popescu e seus colaboradores em 1996, pela Universidade de Engenharia de Missouri - USA. Este sistema identifica os cromossomos pelo tamanho e através de suas bandas para, posteriormente, formar os respectivos pares no kariograma. Foi apresentado com quatro novas características, sendo que a primeira tinha como objetivo encontrar o eixo médio do cromossomo através do método *cross section sequence graphs* (CSSG). Já a segunda característica, na etapa de segmentação não se buscava fazer o corte dos cromossomos

sobrepostos, ao invés disso, o algoritmo extraía um conjunto de fragmentos e comparava-os com outros cromossomos com as mesmas características. Uma terceira característica foi desenvolvida com o intuito de desconectar os cromossomos que se encontram encostados utilizando-se, para tanto, o conceito do “caminho mais claro” para separá-los. A quarta e última característica referia-se aos desempenhos destas novas características, que eram avaliados automaticamente. O sistema MAKSYS também utilizava o método de sistemas especialistas na etapa de classificação, ou seja, a técnica de RNAs com dois algoritmos *Backpropagations* em paralelo (POPESCU *et al.*, 1999). É importante mencionar que, neste estudo, o foco principal foi o reconhecimento automático de cromossomos que se encontravam sobrepostos ou severamente encostados, apresentando taxas de 25 a 90% de acerto. Mais tarde, este sistema foi melhorado, passando a ser denominado *Automatic Karyotyping System* (AKS).

Até o presente momento, todos os métodos desenvolvidos tinham como objetivo tentar identificar cromossomos em células metafásicas de indivíduos normais. Entretanto, na ocorrência de aberrações cromossômicas, a identificação dos pares homólogos tornava-se muito difícil ou praticamente impossível através do processo computadorizado. Assim, o trabalho apresentado por STANLEY *et al.* (1998) teve como intuito desenvolver um algoritmo buscando identificar cromossomos que apresentavam anormalidades estruturais e numéricas. Foram utilizadas como características o índice centromérico e os padrões de banda e, como classificador, foi aplicado o método do algoritmo de transporte para ser comparado com uma RNA de estrutura *Feedforward*. Foram testados os cromossomos nº 17 de 55 metáfases bem espalhadas, sendo que os resultados encontrados foram de 85,5–89,1% de acerto utilizando-se o algoritmo baseado em RNAs e 80% utilizando-se o método do algoritmo de transporte.

Já no ano seguinte, SJAHPUTERA e KELLER (1999) publicaram um trabalho utilizando a técnica de *Lógica-Fuzzy*, em que foi apresentada uma evolução de uma Regra-Base *Fuzzy* que buscava otimizar o desempenho de um classificador automático de cromossomos. Foram utilizadas imagens de células metafásicas de indivíduos normais com cromossomos não sobrepostos, escolhidos por especialistas da *Ellis Fischel Câncer Center*, Columbia – Missouri. As características extraídas foram índice centromérico, comprimento relativo, número de bandas, espaçamento e intensidade das bandas. No conjunto de testes utilizou-se 180 cromossomos: 55 cromossomos nº 16, 54 cromossomos

nº 17 e 71 cromossomos nº 18, e aplicados em um classificador baseado na Regra-Base *Fuzzy* com seis fases de evolução e outro baseado na técnica de RNAs. Os resultados encontrados mostraram que o desempenho do classificador baseado na técnica *Fuzzy* foi superior, com 88,3% de acerto, comparado aos 86,7% da rede neural.

É importante mencionar que, nos trabalhos pesquisados sobre sistemas de cariotipagem, os índices de acerto de cada uma destas abordagens utilizam parâmetros diferentes no tratamento com os dados. Assim, alguns dos sistemas utilizam a abordagem totalmente automática, podendo então, ocorrer erros na determinação do centrômero ou na segmentação, o que afeta o índice de acerto final da classificação. Já nos sistemas que utilizam a abordagem semi-automática, permite-se a interação com o usuário para a correção de erros no decorrer das etapas do processo, por exemplo, na etapa de extração de características, possibilitando-se, portanto, um melhor desempenho na classificação.

Muitos outros sistemas computadorizados de cariotipagem ainda foram pesquisados, sendo que, na grande maioria, as informações foram omitidas sobre as técnicas empregadas devido a estes sistemas serem comerciais. Por conseguinte, as informações levantadas, quando não eram restritas, eram especulativas, sem muito ou quase nada de aproveitamento científico. Entretanto, vale ainda citar os principais sistemas comercialmente conhecidos, dentre eles o LEICA CHANTAL, desenvolvido pela LEICA (2003), sendo uma ferramenta muito útil na determinação de algumas anomalias e patologias clínicas associadas às alterações cromossômicas, utilizando também técnicas de RNAs para a classificação dos cromossomos. Neste contexto também, podem ser citados os sistemas IKAROS, desenvolvido pela META SYSTEMS (2002), o CARIOSYS, desenvolvido pela BIOARGEN (2003), o CYTOVISION e o QUIPS, ambos desenvolvidos pela APPLIED IMAGING (2003).

Sem dúvida nenhuma, pode-se perceber que o processo de reconhecimento automático de cromossomos de células metafásicas tornou-se, nos últimos tempos, um dos maiores desafios para os cientistas, apresentando dificuldade maior em relação a outros processos de análise de imagens para uso em diagnóstico clínico. Isso porque existem muitas limitações dos sistemas, decorrentes da ocorrência de fontes de incerteza no domínio do problema, direcionando as pesquisas para o desenvolvimento de algoritmos baseados em sistemas especialistas, principalmente as RNAs. Estes sistemas especialistas

alavancaram uma melhoria expressiva no desempenho dos classificadores, alicerçando seu uso na grande maioria dos sistemas de cariotipagem comerciais da última década.

Ainda assim, a aplicação dos sistemas especialistas como classificadores minimizou apenas em parte as limitações dos sistemas automáticos de cariotipagem. Nos trabalhos pesquisados baseados nestes sistemas, os testes foram realizados utilizando-se bases de dados de células metafásicas adquiridas em laboratórios de genética humana, tais como o *Copenhagen*, o *Edinburgh* e o *Philadelphia*. Na realização dos testes, quando são utilizadas imagens de um único laboratório, separando-as em grupos de avaliação e treinamento, há a tendência do classificador ter um excelente desempenho, como pode ser constatado nos trabalhos de JOHNSTON *et al.* (1993), DENISOV e DUDKIN (1994), TODESCO (1995) e SJAHPUTERA e KELLER (1999). Isso decorre do fato de que células metafásicas provenientes de um mesmo laboratório são preparadas segundo o mesmo protocolo para obtenção e coloração de células metafásicas, de forma que as características essenciais que descrevem os cromossomos, utilizadas para classificação, sejam mais uniformes. Porém, tais características podem sofrer muitas variações quando são extraídas de imagens de cromossomos de laboratórios distintos. Logo, nos testes em que se treinava o sistema com imagens de cromossomos de um determinado laboratório e se testava com imagens de outros laboratórios, o índice de acerto diminuía consideravelmente. Este fato pode ser muito bem constatado nos testes realizados no trabalho de JOHNSTON *et al.* (1993).

Para contornar este problema, os sistemas comerciais de cariotipagem disponibilizam ao usuário uma forma de armazenamento, na base de dados do próprio sistema, de imagens de cromossomos já classificados pelo usuário e que são de rotina no laboratório em questão. Ainda assim, não há garantia de que o sistema apresente um bom desempenho, uma vez que a eficiência destes sistemas depende da inclusão de um número expressivo de padrões pré-estabelecidos na base de dados. Logo, isso vem a tornar um inconveniente aos citogeneticistas, pois os mesmos são requeridos a sempre armazenarem estas imagens de cromossomos, já classificadas, até que o sistema em uso atinja o estágio de boa eficiência na classificação, o que pode levar, em média, de 2 a 3 anos, dependendo do laboratório.

Além disso, em todos os trabalhos pesquisados, verificou-se que a preocupação dos pesquisadores estava na elaboração de métodos para melhorar o reconhecimento

automático dos cromossomos, desde a produção e a aplicação de novas técnicas de processamento digital de imagens como, também, no desenvolvimento de classificadores mais eficientes. Dessa forma, o objetivo final sempre foi desenvolver um sistema capaz de efetuar o reconhecimento automático dos cromossomos com máxima eficiência, não sendo encontrado na literatura nenhum sistema com características semelhantes disponibilizados para análise em citogenética clínica.

## **2.4 – Considerações Finais**

O principal objetivo deste capítulo, foi fazer uma breve descrição do processo geral de desenvolvimento dos sistemas automáticos de cariotipagem já existentes. Através do entendimento das técnicas aplicadas por estes sistemas, pode-se concluir que ainda existem alguns pontos ainda não resolvidos no processo de identificação automática de cromossomos, o que motiva o desenvolvimento de novas e melhores técnicas para minimizar tais limitações, contribuindo com o citogeneticista a desempenhar melhor a atividade de análise e classificação citogenética.

Nos tópicos citados neste capítulo buscou-se apresentar, então, um breve histórico dos sistemas automáticos de cariotipagem existentes, suas técnicas e métodos aplicados, através dos seguintes aspectos: sistemas automáticos para análise e classificação citogenética e, por fim, o histórico das técnicas utilizadas pelos sistemas de cariotipagem.

## CAPÍTULO III

# CITOGÉNÉTICA HUMANA

### 3.1 – Introdução

O estudo dos cromossomos humanos é um procedimento essencial em muitos campos de investigação, tendo como objetivo a avaliação da integridade numérica e/ou estrutural dos cromossomos, visando detectar anormalidades genéticas. Para tanto, este estudo está obrigatoriamente associado à compreensão dos processos que ocorrem durante a divisão celular (processo este denominado de mitose<sup>1</sup>), devido ao fato dos cromossomos somente poderem ser visualizados como corpúsculos individualizados em núcleos que estão em processo de divisão, e de sua morfologia visando sua análise e identificação.

Assim, este capítulo é dedicado ao entendimento e a compreensão do objeto de estudo, servindo de base e fundamentação para a formulação do problema. Logo, é apresentada uma breve descrição dos tópicos referentes à genética humana e que fazem parte do processo de identificação dos cromossomos humanos, tais como: os aspectos biológicos, o ciclo celular, a identificação dos cromossomos metafásicos, a identificação clássica dos cromossomos metafásicos, a identificação dos cromossomos por intermédio das bandas, a representação visual dos cromossomos denominada de cariograma, a preparação das células metafásicas como amostra, as fontes de incerteza no domínio do problema, a leucemia pró-mielocítica aguda e, finalmente, as considerações finais sobre o que foi exposto neste capítulo.

### 3.2 – Aspectos Biológicos

Os cromossomos estão presentes em todas as células nucleadas de todos os organismos vivos, tendo como função levar as informações genéticas para a reprodução

---

<sup>1</sup> Mitose (do grego, *mitos* = filamento), termo criado em 1880 por *W. Flemming*.

das células e organismos. O nome, *cromossomo*<sup>2</sup>, é derivado do grego "*chroma*" para cor e "*soma*" para corpo, isto porque são facilmente coráveis por corantes básicos. A reprodução das células serve para dois importantes propósitos: o crescimento e a manutenção dos organismos vivos gerando novas células, através da divisão (mitose) e para a reprodução dos organismos pela geração de células reprodutivas especializadas através da redução da divisão, denominada de meiose. Neste trabalho, as células reprodutivas não serão alvo de estudo.

### 3.3 – O Ciclo Celular

O ciclo completo de uma célula é composto por dois estágios: o estágio da intérfase e o estágio da mitose. O estágio da intérfase é caracterizado pela fase de síntese de DNA, isto é, estágio em que o teor de DNA atinge o dobro da quantidade inicial, e neste período podem ser distinguidas três fases: a fase de síntese de DNA, o intervalo pré-sintético e o intervalo que precede a mitose (BEIGUELMAN, 1982). Nas células fixadas e coradas, a cromatina do núcleo interfásico apresenta-se como um emaranhado de filamentos muito finos, de baixa cromafinidade, entre os quais aparecem grumos condensados e fortemente corados. Neste estágio os cromossomos são impossíveis de serem identificados.

Assim, a análise e classificação citogenética somente pode ser realizada no estágio da mitose, justamente porque é nesta fase que os cromossomos apresentam maior individualidade, o que proporciona a possibilidade da identificação cromossômica. O estágio da mitose é o período em que a célula sofre divisão, e o material que constitui os cromossomos é distribuído de modo equitativo para os dois novos núcleos que se originam. Neste processo de divisão, distinguem-se quatro fases definidas e sucessivas, denominadas de prófase, metáfase, anáfase e telófase.

O estágio da prófase pode ser caracterizado pelo fenômeno de condensação da cromatina. Ao final desta fase, pode-se constatar que cada cromossomo apresenta-se constituído por dois filamentos denominados de cromátides. O próximo estágio do ciclo mitótico celular é a metáfase, em que a cromatina torna-se muito condensada ou contraída,

---

<sup>2</sup> *Cromossomo*, termo criado em 1888 por *Waldeyer*.

sendo muito fácil verificar ao microscópio óptico que cada cromossomo metafásico é constituído por duas cromátides ligadas entre si por uma região estrangulada, ou seja, uma constrição denominada de *centrômero*. É a posição do centrômero que determina a forma dos cromossomos metafásicos, ou seja, cromossomos metacêntricos, submetacêntricos e acrocêntricos, como mostra a Figura 3.1.

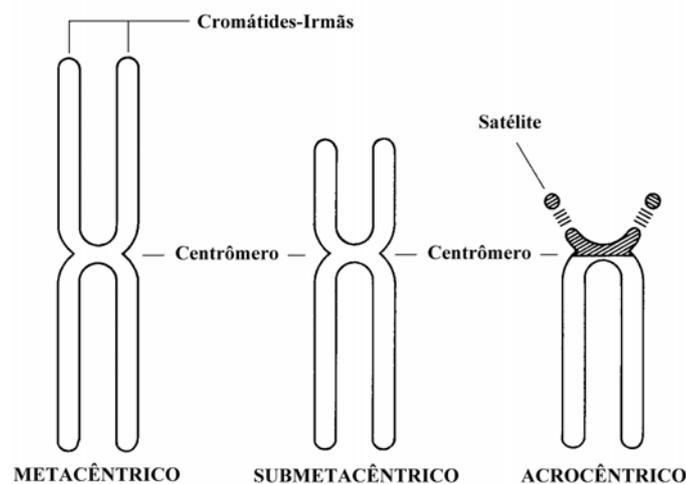


Figura 3.1 – Ilustração idealizada da forma dos cromossomos metafásicos, de acordo com a posição do centrômero, ou seja, cromossomos metacêntricos, submetacêntricos e acrocêntricos.

Logo, quando o centrômero ocupa a posição central ou praticamente central, separando dois braços cromossômicos iguais, então o cromossomo é *metacêntrico*. Entretanto, se o centrômero não tem posição central, de modo a dividir o cromossomo em dois braços de tamanhos desiguais, ele é denominado de *submetacêntrico*, sendo que o braço menor é sempre considerado como o braço superior. Se o centrômero possui posição subterminal ou praticamente terminal, de modo a separar um braço muito curto ou praticamente não-detectável ao microscópio comum, o cromossomo é denominado de *acrocêntrico*<sup>3</sup>. Os cromossomos acrocêntricos podem apresentar o braço curto de forma arredondada, como se estivessem quase destacadas do restante dos cromossomos. Essas formações desse braço são denominadas de *satélites* (THOMPSON *et al.*, 1991).

<sup>3</sup> Acrocêntrico, do grego, *akros* = extremidade.

Após o estágio da metáfase vem a fase da anáfase que se inicia pela divisão dos centrômeros, ou seja, pela separação das cromátides-irmãs em que cada cromátide separada se tornará um novo cromossomo. Ao término da fase da anáfase, inicia-se o último estágio da mitose, a fase da telófase, onde a cromatina passa a descondensar-se formando uma nova membrana nuclear, e os dois núcleos resultantes readquirem o aspecto que tinham durante a intérfase.

A realização de estudos a respeito da constituição cromossômica dos seres humanos normalmente é realizada durante o processo de divisão celular, por intermédio da análise de metáfases em células somáticas. Esta fase é escolhida em virtude da grande condensação sofrida pelo material cromossômico e também devido às cromátides e a posição dos centrômeros ficarem bem evidentes. Portanto, os cromossomos metafásicos constituem o melhor material para a detecção de alterações cromossômicas. Porém, somente este aspecto não minimiza as dificuldades deste tipo de estudo. Ter um grande número de figuras metafásicas em uma determinada população celular não é necessariamente suficiente para que se consiga uma boa análise cromossômica, pois ainda existem muitas fontes de incerteza no domínio do problema tornando-se sérios obstáculos ao estudo individual desses elementos (BEIGUELMAN, 1982).

### 3.4 – Identificação dos Cromossomos Metafásicos

Em 1956 *Tjio e Levan*, através da publicação de um artigo científico, relataram que o número de cromossomos humanos eram 46 e não 48 como se imaginava. Esse trabalho foi confirmado por *Ford e Hamerton*, também em 1956 e, através desses dois trabalhos, houve o estímulo de vários laboratórios engajados no estudo de cromossomos humanos a proporem um sistema de classificação. Foram realizadas várias conferências com o objetivo de trazer propostas para padronização da classificação e uniformização da nomenclatura, dentre as quais pode-se citar: *Denver Conference – 1960*, *London Conference – 1963* (onde foi oficialmente sancionada a classificação dos cromossomos em sete grupos designados por letras de A-G) e *Chicago Conference - 1966* (ISCN, 1995).

Até então, a identificação dos cromossomos humanos era feita, quase que exclusivamente, com base nos critérios de tamanho e forma dessas estruturas, porém esse

tipo de análise era insuficiente para o reconhecimento de todos os pares cromossômicos. A partir de 1970, com o trabalho apresentado por *Torbjörn Casperson* e colaboradores, foram desenvolvidos novos processos de tratamento nos cromossomos, provocando a diferenciação longitudinal das cromátides passando a exibir faixas transversais de espessura variável, denominadas de *bandas cromossômicas*. Assim, várias outras técnicas, que também produziam bandas cromossômicas, foram desenvolvidas com o intuito de também identificar os cromossomos, porém com mais precisão. E, já na conferência internacional do ano seguinte, a *Paris Conference - 1971*, foi publicado, em seus anais, um importante documento propondo um sistema básico, não somente para designação individual de cromossomos, mas também trazendo a composição de bandas (ISCN, 1995).

Outras conferências foram realizadas até que em 1976, em um congresso internacional de genética humana realizado na Cidade do México, foi criado um comitê internacional de padronização em citogenética humana para continuar o trabalho de padronização de identificação de cromossomos humanos. Este comitê reuniu-se em *Stockholm* em 1977, onde foi confeccionado um documento intitulado “An International System for Human Cytogenetic Nomenclature (1978)” sendo abreviado para *ISCN (1978)*. À medida que foram sendo desenvolvidas novas formas de análise cromossômica, foram implementadas complementações destes critérios continuamente, através do comitê em reuniões periódicas. Outras revisões do Sistema Internacional para Nomenclatura de Citogenética Humana foram preparadas em 1984 e 1991, denominadas de *ISCN (1985)* e *ISCN (1991)*, respectivamente, sendo que a mais recente revisão corresponde ao *ISCN (1995)*, no qual este trabalho se baseia.

### **3.4.1 – Identificação Clássica dos Cromossomos Metafásicos**

Como foi mencionado anteriormente neste capítulo, até o ano de 1970 a identificação dos cromossomos humanos era feita com base nos critérios de tamanho e forma dessas estruturas, levando-se em conta, no tocante à forma, além da posição do centrômero, a localização das *constrições secundárias*. Estas constrições secundárias são as constrições cromossômicas que não contêm centrômero e que, em preparações

favoráveis, podem ser detectadas em certos cromossomos. Assim, a identificação dos cromossomos metafásicos com base em tais critérios é denominada de *identificação clássica*, porém este tipo de identificação apresentava algumas limitações pelas quais nem todos os cromossomos podiam ser reconhecidos.

Entretanto, apesar da identificação clássica dos cromossomos apresentar limitações, esse critério é suficiente para classificar os cromossomos quanto ao tamanho, sendo os mesmos facilmente classificados em grandes, médios, relativamente pequenos e pequenos. Não há a necessidade de medi-los ou saber a escala de ampliação utilizada, pois a avaliação do comprimento de um cromossomo metafásico é feita tomando-se por referência o comprimento dos outros cromossomos da mesma metáfase.

Assim, tendo em vista que na espécie humana existem uns totais de 23 pares de cromossomos, sendo que em cada par homólogo um dos cromossomos é de origem paterna e o outro de origem materna. Estes pares são distribuídos em sete grupos, designados pelas sete primeiras letras do alfabeto. Esta classificação é apresentada pela Figura 3.2, onde os cromossomos são dispostos de acordo com o seu tamanho e a posição do centrômero, sendo identificados os 22 pares de cromossomos autossomos homólogos, além dos cromossomos sexuais X e Y, nos grupos *A, B, C, D, E, F* e *G*.

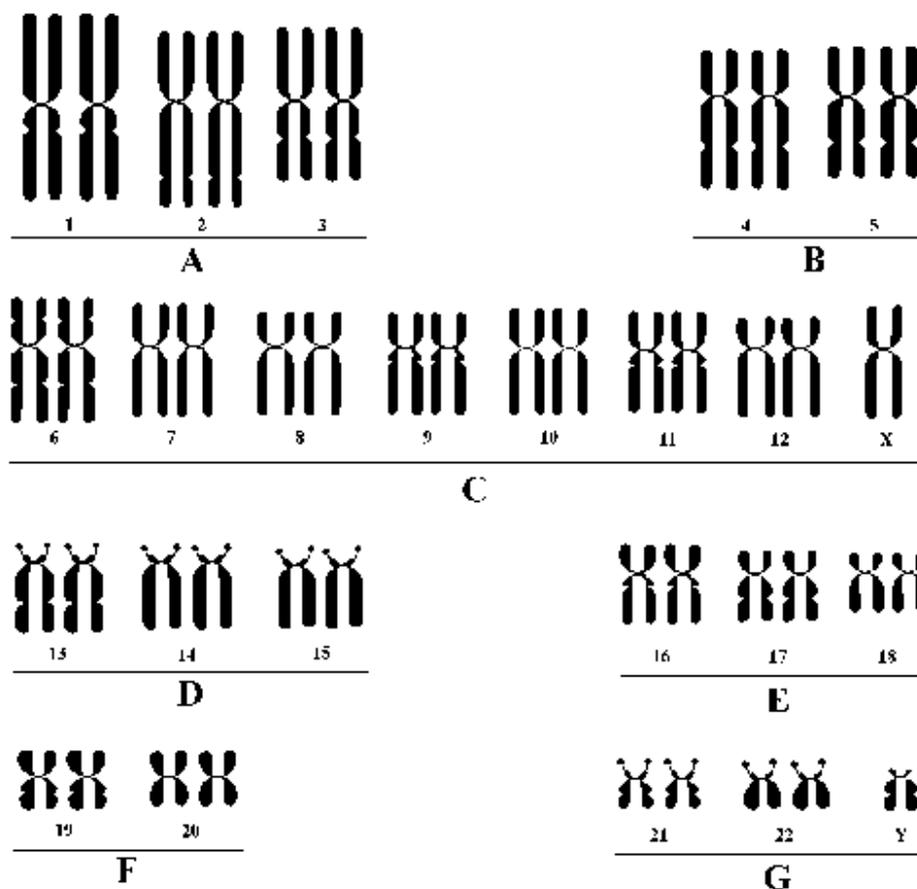


Figura 3.2 – Representação esquemática dos cromossomos metafásicos humanos de acordo com o tamanho e a posição do centrômero, identificados através dos 22 pares homólogos, além dos cromossomos sexuais X e Y, nos grupos A, B, C, D, E, F e G.

Portanto, os cromossomos são classificados em grupos através dos seguintes critérios: A e B para os cromossomos grandes, C (incluindo o cromossomo X) e D para os cromossomos de tamanho médio, E para os cromossomos relativamente pequenos (dependendo da metáfase podem apresentar-se como médios ou como pequenos), e F e G (incluindo o cromossomo Y) para os cromossomos pequenos (ISCN, 1995).

Após a separação dos cromossomos nos sete grupos, o próximo passo, então, é a identificação dos pares de cromossomos dentro destes grupos, ou seja, a ordenação dos *pares homólogos*, de acordo com critérios propostos pela Conferência de Paris. Os pares de cromossomos homólogos do grupo A são numerados de 1 a 3. O par de número 1 é composto pelos maiores cromossomos e são metacêntricos, já os cromossomos do par

número 2, de tamanho intermediário entre os pares 1 e 3, são submetacêntricos, e o par de cromossomos número 3, de tamanho menor que os pares 1 e 2, são metacêntricos.

Em oposição ao que ocorre com os cromossomos do grupo *A*, a ordenação dos pares homólogos do grupo *B*, pares 4 e 5, oferece certa dificuldade para sua separação, isto porque, apesar de serem cromossomos grandes, eles mostram grande similaridade quanto à posição do centrômero. A única possibilidade para a diferenciação dos pares de cromossomos do grupo *B* é, então, o tamanho dos braços cromossômicos inferiores, atribuindo-se o número 4 ao par que apresenta tais braços maiores.

O número de cromossomos pertencentes ao grupo *C* varia de acordo com o sexo. No conjunto representativo de cromossomos da Figura 3.2, verifica-se a presença de 7 pares de cromossomos submetacêntricos de tamanho médio além do cromossomo X (sem homólogo), ou seja, um cariótipo masculino normal. Em um cariótipo feminino normal, além dos 7 pares de cromossomos submetacêntrico, o cromossomo *X* possui homólogo. Os cromossomos do grupo *C* oferecem muita dificuldade no seu pareamento devido à grande semelhança de tamanho. A única diferença pode ser vista nos cromossomos dos pares número 6 e nos cromossomos *X* do restante do grupo *C*, devido ser os maiores neste grupo. Para a distinção dos pares homólogos de números 7, 8, 9, 10, 11 e 12 - pode-se classificá-los por ordem decrescente de tamanho, começando-se pelo número 7 que tem tamanho ligeiramente inferior ao par de número 6. Os pares de número 8 e 9 têm tamanho praticamente igual, mas os cromossomos número 9 podem ter os braços inferiores ligeiramente menores do que os de número 8. Os cromossomos dos pares 10, 11 e 12 têm tamanho praticamente igual, mas os do par número 11 tendem a ser mais metacêntricos. Por outro lado, os cromossomos do par número 12 são mais acentuadamente submetacêntricos que os do par número 10.

Já os pares de cromossomos do grupo *D*, são médios em tamanho e acrocêntricos. Comparados com os outros cromossomos acrocêntricos, eles são os maiores do cariótipo humano. Através da identificação clássica, é muito difícil distinguir os pares de cromossomos que fazem parte do grupo *D*. Algumas vezes, o par de número 15 apresenta-se ligeiramente menor do que os pares 13 e 14.

Os cromossomos do grupo *E* são de fácil distinção devido ao tamanho relativamente pequeno, e separados em pares homólogos numerados de 16 a 18. Os cromossomos do par 16 são submetacêntricos ou, muitas vezes, quase metacêntricos, e os

de maior tamanho apresentam-se dentro desse grupo. Já os cromossomos do par número 17 possuem tamanho intermediário, comparado aos outros dois pares do grupo, são submetacêntricos e os cromossomos do par número 18 são submetacêntricos.

Existem também, no cariótipo humano, cromossomos que são pequenos e metacêntricos, constituindo os pares de número 19 e 20 do grupo *F*. A distinção entre os cromossomos desse grupo é extremamente difícil pela identificação clássica, a não ser quando os cromossomos do par número 19 são ligeiramente maiores.

Assim como ocorre com os cromossomos do grupo *C*, o número de cromossomos do grupo *G* varia segundo o sexo. Portanto, em um conjunto representativo de cromossomos do cariótipo feminino normal, verifica-se a existência de apenas dois pares de cromossomos acrocêntricos pequenos (pares homólogos de número 21 e 22), enquanto que, em um conjunto representativo de cromossomos do cariótipo masculino normal, constata-se a presença dos dois pares de cromossomos número 21 e 22, além de um cromossomo sem homólogo denominado Y. A distinção entre os cromossomos 21 e 22 é quase impossível, além de mostrarem grande variabilidade de forma, às vezes eles se apresentam com aspecto submetacêntrico. Entretanto, se entre eles for notada diferença de tamanho, atribui-se o número 21 ao par de cromossomos menores. Já o cromossomo Y é mais facilmente distinguível entre os do grupo *G* porque é mais acrocêntrico. Porém, o cromossomo Y pode apresentar grande variabilidade de comprimento, mas que afeta apenas a parte distal de seu braço inferior.

Assim, pode-se concluir que os cariótipos masculinos e femininos diferem apenas em relação a um único par cromossômico. No cariótipo feminino, os cromossomos constituem 23 pares homólogos, incluindo o par de cromossomos X, denominados de *cromossomos sexuais*, enquanto que, no cariótipo masculino, apenas 22 dentre os 23 pares são homólogos, sendo que o par não homólogo é o par de cromossomos sexuais constituído por um cromossomo X (grupo *C*) e um cromossomo Y (grupo *G*) de origem paterna.

### 3.4.2 – Identificação dos Cromossomos por Intermédio das Bandas

A primeira técnica empregada para auxiliar a identificação dos cromossomos humanos foi a auto-radiografia de metáfases de culturas expostas a pulsos de *desoxitimidina triciada*. O emprego da auto-radiografia, além de ser difícil e demorado, não permitia a diferenciação de todos os cromossomos humanos. Já em 1970, *Caspersson* e seus colaboradores verificaram que os cromossomos metafásicos humanos tratados com *mostarda de quimacrina* (*quimacrine mustard* ou *quinacrine dihydrochloride*) apresentavam uma diferenciação longitudinal em suas cromátides, as quais mostravam faixas transversais de fluorescência variável. Assim, empregando esse fluorocromo, demonstrou-se que todos os pares cromossômicos podiam ser perfeitamente distinguidos de acordo com o tamanho, localização dos centrômeros e também através da distribuição e intensidade de fluorescência das faixas que apresentavam. Estas faixas ou *bandas cromossômicas* foram, posteriormente, denominadas de *bandas Q* (inicial de quimacrina) e apresentam-se através de listas escuras ou claras. Logo, os cromossomos são visualizados como série de bandas escuras alternando com regiões claras (área entre duas bandas).

Paralelamente, *Pardue* e *Gall*, também em 1970, realizaram a hibridação *in situ* em cromossomos de camundongos tratando-os com uma base forte (NaOH) demonstrando que o DNA satélite dos cromossomos dos camundongos estava confinado à região dos centrômeros (ISCN, 1995). Além disso, depois de tratar as preparações citológicas com solução de *Giemsa*, as áreas cromossômicas com DNA satélite coravam-se fortemente ao contrário das outras. Estimulados por este trabalho, *Sumner* e colaboradores em 1970 demonstraram o DNA satélite dos cromossomos humanos, revelando blocos de cromatina bem corada, não apenas na região centromérica de todos os cromossomos, mas também, na região distal do braço longo do cromossomo Y, na região do braço menor e dos satélites dos cromossomos dos grupos *D* e *G*, e nas constrições secundárias proximais ao centrômero dos cromossomos 1, 9 e 16. Tais regiões heterocromáticas foram denominadas de *bandas C*.

Modificando a técnica de *Pardue* e *Gall* em 1971, *Drets* e *Shaw* demonstraram a existência de faixas ao longo dos cromossomos, as quais são evidenciadas pelo corante *Giemsa*, e apresentam correspondência com as bandas Q. Outros pesquisadores utilizando

diversos procedimentos, obtiveram resultados similares, denominando-se este processo de bandas G (inicial de *Giemsa*).

Após o aparecimento da técnica de identificação cromossômica através das bandas, numerosas variantes técnicas têm aparecido para a produção de bandas Q, C e G, entre elas o trabalho de *Dutrillaux e Lejeune*, em 1971, que após incubar as preparações citológicas em solução de *Earle* aquecida, evidenciaram um padrão de bandas denominado *bandas R* (derivação da denominação inglesa *reverse bands*), as quais correspondem ao inverso das bandas G (ISCN, 1995).

A Figura 3.3 ilustra a comparação entre as bandas G e R do cromossomo número 1. Ainda nesta figura, pode-se visualizar a subdivisão dos cromossomos em braços, regiões e bandas, relacionando-os com símbolos para facilitar a indicação das áreas cromossômicas (VOGEL e MOTULSKY, 1997). Por convenção, a letra *p* designa o braço superior dos cromossomos, enquanto a letra *q* designa o braço inferior deles. Já as regiões são designadas em regiões numeradas em ordem crescente, no sentido distal a partir do centrômero (designado por 10), ou seja, as regiões próximas à constrição primária, tanto o braço superior quanto o inferior são indicadas pelo número 1, e as bandas que seguem o sentido distal, quando existem, são numeradas em ordem crescente.

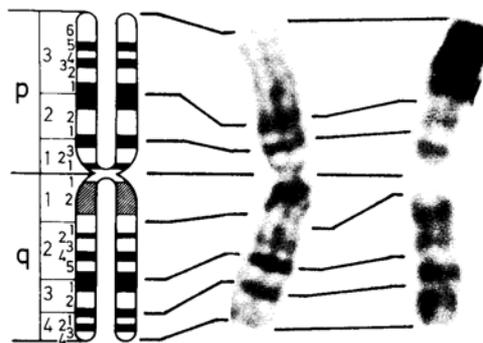


Figura 3.3 – Comparação entre as bandas G e R do cromossomo número 1.

Logo, para designar a posição de uma determinada banda, é necessário: o número do cromossomo, o símbolo do braço, o número da região e o número da banda. Por exemplo, *1p31* indica o cromossomo número 1, braço superior, região 3 e banda 1.

A Figura 3.4 ilustra uma representação idealizada dos cromossomos metafásicos humanos, mostrando as bandas que podem ser obtidas com o auxílio de diferentes técnicas. As faixas escuras correspondem às bandas Q ou G; as faixas claras às bandas R e as faixas hachuradas (regiões polimórficas) às bandas cromossômicas que mostram muita variabilidade de tamanho na população.

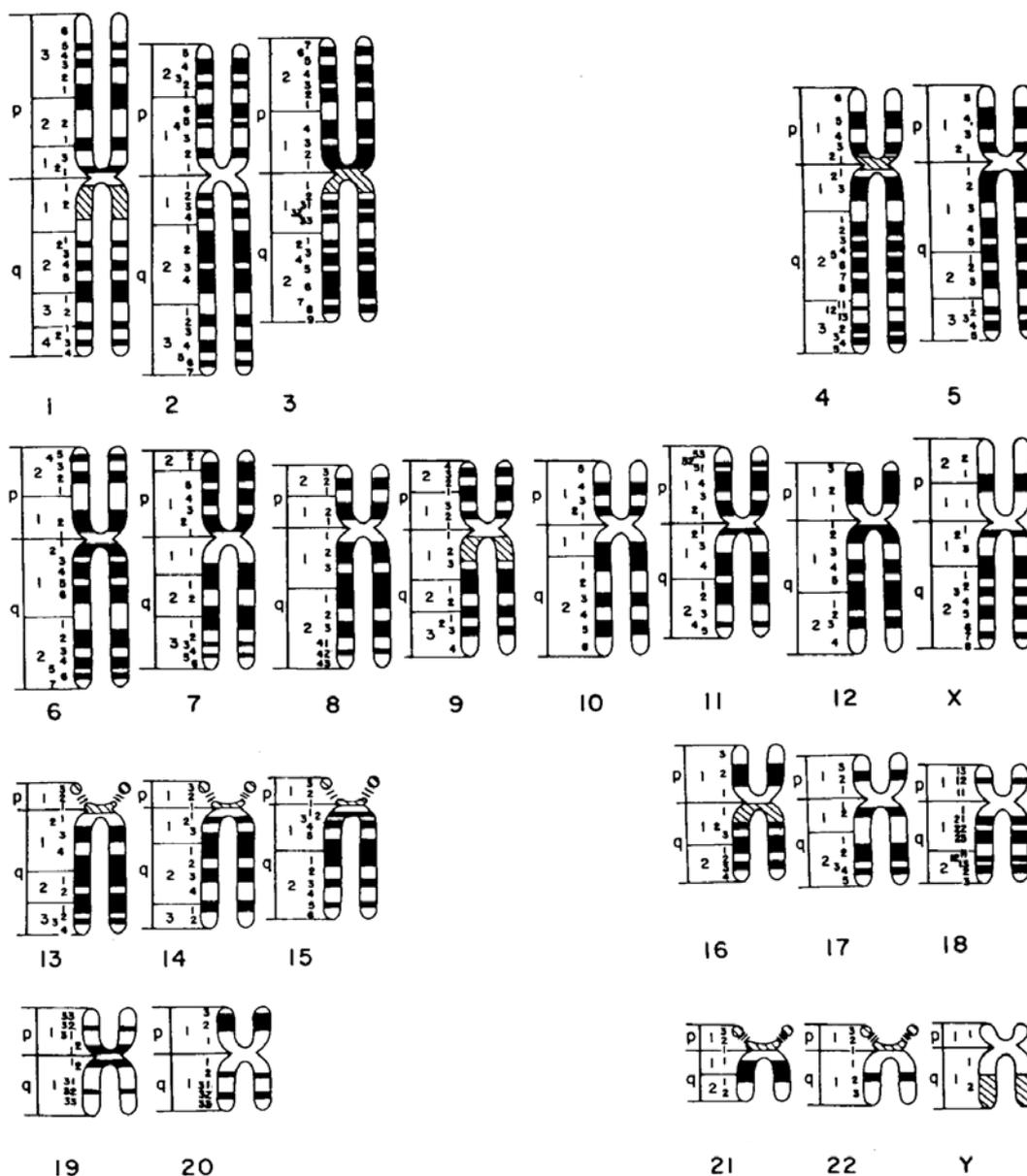


Figura 3.4 – Representação idealizada dos cromossomos metafásicos humanos mostrando as bandas que podem ser obtidas com o auxílio de diferentes técnicas.

### 3.5 – O Cariograma

Para fins de diagnóstico clínico, é necessário que se faça, então, a análise quanto a normalidade ou anormalidade dos cromossomos. Após a identificação individual e dos pares dos cromossomos da metáfase, é necessário que se tenha um arranjo, ou melhor, uma representação visual desses pares de cromossomos, organizados em grupos pela ordem decrescente de tamanho para possibilitar a comparação entre os cromossomos homólogos, como mostra a Figura 3.5. Esta representação visual dos cromossomos é denominada de *cariograma* ou *cariótipo*, tradução do termo *karyotype*<sup>4</sup> e, por esse motivo, o processo de classificação e análise dos cromossomos é denominado de *cariotipagem*.



Figura 3.5 – Representação visual dos pares de cromossomos, denominada de cariograma, pelos quais são distribuídos em ordem decrescente de tamanho.

<sup>4</sup> *Karyotype*, termo derivado da palavra grega *karyon* = nó e *typos* = forma.

### 3.6 – Preparação das Células Metafásicas como Amostra

Como citado anteriormente neste capítulo, o melhor material para a investigação da ocorrência de alterações numéricas e/ou estruturais são os cromossomos metafásicos. Para obter-se um maior número de células em divisão, é realizada uma cultura de células, seguindo-se procedimentos para interrupção da divisão celular na metáfase. Para a cultura de células para o estudo citogenético, podem ser utilizados sangue periférico, fibroblastos e ainda fluido amniótico, *vilosidade coriônica* ou sangue fetal para realização de diagnóstico pré-natal. Alternativamente, podem-se utilizar preparações diretas de medula óssea. Os procedimentos utilizados para cultura e obtenção de metáfases são similares para todos os tipos de material biológico utilizado, como pode ser visto no procedimento a seguir:

- 1- A divisão celular é bloqueada no estágio da metáfase adicionando-se ao material uma solução diluída de *colchicina* ou um agente similar, que inibe a formação das fibras do fuso central retardando a separação dos centrômeros.
- 2- Em seguida, adiciona-se ao material uma solução *hipotônica* para inchar o núcleo da célula e facilitar o espalhamento dos cromossomos e, logo depois, adiciona-se uma solução de *metanol* e *ácido acético* com o objetivo de fixar a proteína básica da cromatina.
- 3- A membrana do núcleo celular é fisicamente rompida no momento em que o material é gotejado (lançado a certa distância) nas lâminas e quando se aplica um "fluxo" de ar, ou então cobrindo-as em alta umidade. Logo em seguida, as lâminas são aquecidas por 6 horas a 45° C.
- 4- Para o bandeamento G, as células são tratadas com uma solução de *trypsin*a para a desnaturação das proteínas cromossômicas e, para a coloração, utiliza-se o corante *Giemsa* ou *Wright* (solução de *eosina-azul de metileno*). Após a coloração, a metáfase pode ser visualizada ao microscópio de luz comum.

A análise em bandas G, há mais de 20 anos, tornou-se a metodologia padrão em praticamente todos os laboratórios de citogenética. Apesar dos procedimentos de cultura serem padronizados, podem ocorrer variações de um laboratório para outro na quantidade de reagentes utilizados e no tempo de coloração das lâminas. Além disso, a preparação da

lâmina é totalmente artesanal e o resultado de se ter boas metáfases vai depender muito da habilidade e da experiência do profissional.

### 3.7 – Fontes de Incerteza no Domínio do Problema

Concluída a preparação da amostra, as células no estágio da metáfase mesmo em circunstâncias ideais, são uma pequena minoria das células presentes em uma lâmina. O mais importante é encontrar, entre todas as células da lâmina, um conjunto de metáfases adequadas que permitam avaliar criteriosamente sua constituição cromossômica, utilizando as características que identificam os cromossomos: tamanho, posição dos centrômeros e padrões de banda. Entretanto, existem alguns fatores técnicos que podem comprometer a acuracidade da análise, ou mesmo impedir a análise dessas características comprometendo, conseqüentemente, a correta identificação dos cromossomos. Tais fatores são denominados aqui de *Fontes de Incerteza* no domínio do problema, sendo inerentes, principalmente, ao processo de cultura das células metafásicas e preparação das lâminas, além também da falta de correções ópticas das lentes de aumento. Dentre as principais fontes de incerteza e suas causas, pode-se citar:

1- *Ocorrência de sobreposições, toques e ou torções severas*: considerando que os cromossomos estão concentrados em um espaço relativamente pequeno dentro da célula, as sobreposições são quase que inevitáveis, tornando-se um dos principais obstáculos para a identificação cromossômica. Existem vários fatores que podem influenciar a ocorrência destas sobreposições, dentre eles a solução hipotônica utilizada. Se for pouco concentrada, essa solução faz com que os cromossomos fiquem menos espalhados na lâmina, dando margem a um número maior de sobreposições. Outro fator de grande influência surge no processo de preparação da lâmina, quando o material é gotejado na lâmina e, em seguida, aplicado um “fluxo” de ar para gerar o espalhamento. A boa qualidade do vidro da lâmina permite um melhor espalhamento dos cromossomos e o clima também pode apresentar uma certa influência devido ao fator de secagem da lâmina.

2- *Forma de espalhamento aleatório dos cromossomos na lâmina*: há uma tendência dos cromossomos, que se localizam na periferia da metáfase (ou periferia das imagens), apresentarem-se maiores (mais inchados) e com coloração mais clara do que os

cromossomos que se localizam na região mais central. Esse fator pode ser causado devido às falhas ópticas produzidas pelas objetivas, gerando aberrações nas imagens. Tais aberrações compreendem as distorções através da perda de cor e/ou nitidez da imagem e quanto maior a magnitude das objetivas, maior será a percepção das falhas. Essas aberrações ópticas podem ser divididas em: aberrações cromáticas, que surgem devido à dispersão da luz produzida pela lente, e as aberrações esféricas que aparecem devido a não convergência dos raios luminosos. Cores com comprimento de onda menor sofrem maiores desvios, logo, objetivas especiais foram desenvolvidas visando corrigir os problemas quanto às aberrações cromáticas, denominadas de objetivas *acromáticas* (correções de apenas duas cores) e *apocromáticas* (correções das demais cores do espectro). Já as aberrações esféricas geram deformações geométricas na imagem e, mesmo as melhores objetivas, apresentam o campo visual sensivelmente curvado, de maneira que existe uma diferença de nitidez entre o centro e a periferia das imagens. As correções para essas aberrações têm como finalidade o nivelamento do campo visual e, os métodos para se chegar a uma imagem plana, dependem da utilização das objetivas denominadas de *plan*.

3- *Diferença de tamanho dos cromossomos quando comparados em metáfases distintas*: quando as culturas são expostas a *colchicina*, a divisão celular é interrompida na metáfase. Sabe-se que o tamanho do cromossomo varia de acordo com o tempo de exposição e a etapa da metáfase, logo, no momento em que é aplicada a solução, podem aparecer, em uma mesma cultura, células com diferentes estágios de metáfase e, conseqüentemente, cromossomos de tamanhos diferentes. Porém, entre os cromossomos de uma mesma célula (metáfase) o tamanho é proporcional.

4- *Ocorrência de alterações estruturais*: a ocorrência de alterações cromossômicas pode dificultar a classificação dos cromossomos, quando analisados o tamanho e o padrão das bandas, pois são decorrentes da falta ou excesso de algum segmento cromossômico, que resulta na alteração da morfologia do mesmo.

5- *Falta de nitidez das bandas*: a definição das bandas cromossômicas depende de cromossomos mais alongados e corados. Conseguem-se cromossomos mais alongados com uma quantidade menor de *colchicina*, porém desespiralização em excesso pode dar margem à ocorrência de sobreposições e torções. Já a coloração depende do tratamento e processo de coloração, obtendo-se maior nitidez com bandas mais escuras, contudo, se o

tempo de coloração for muito grande, as bandas claras podem também se apresentar escuras, perdendo-se parte da informação. É importante mencionar também que o número de bandas produzidas varia de acordo com o grau de espiralização do cromossomo e, quanto mais longo o cromossomo, maior o número de bandas que podem ser visualizadas. Neste caso, o exame é caracterizado por níveis de resolução: o cariótipo simples é aquele em que o nível de resolução apresenta aproximadamente 400 bandas, já o cariótipo de alta resolução é caracterizado por apresentar aproximadamente 800 bandas. O mais comum na prática dos laboratórios de citogenética humana é a análise em cariótipos com resolução na faixa entre 450 a 550 bandas. De forma comparativa, a Figura 3.6 ilustra o padrão de bandas de um mesmo cromossomo em duas metáfases, uma com resolução de 400 bandas e outra com resolução de 850 bandas.



Figura 3.6 – Ilustração dos padrões de banda de um mesmo cromossomo: em (A) metáfase com 400 bandas e em (B) metáfase com 850 bandas.

### 3.8 – Leucemia Pró-Mielocítica Aguda

A leucemia pode ser definida como síndrome *mieloproliferativa* ou, mais simplesmente, de *leucose*. É uma patologia que acomete a medula óssea e se caracteriza pela proliferação medular anárquica, progressiva no tempo e no espaço. É um câncer agressivo e, muitas vezes, fatal, podendo ser classificados como agudos (predominam células imaturas) e crônicos (predominam células maduras) (MOTTA JÚNIOR, 2003).

Em particular, a Leucemia Pró-Mielocítica Aguda (LMA – não-linfocítica tipo M3) representa 20% de todos os casos de leucemia, superando taxas de várias outras doenças genéticas (LOPES, 2003). A etiologia pode ser devido a exposição à radiação, substâncias químicas, fármacos, por motivos hereditários (Neurofibromatose, síndrome de *Down*,

anemia de *Fanconi*, síndrome de *Bloom*), além de outros. É uma doença fulminante que, na ausência da intervenção médica (terapia agressiva), provoca a morte. O tratamento visa exterminar as células leucêmicas com quimioterapia intensa, provocando uma aplasia de medula e que dura em torno de 2-3 semanas para se recuperar.

Os exames para leucemia pró-mielocítica aguda são realizados através da análise da morfologia dos blastos da medula óssea e de exames de citoquímica com colorações específicas destes blastos (LOPES, 2003). Além destes exames, a constituição cromossômica tem se revelado como um poderoso fator prognóstico deste tipo de leucemia, sendo observada nas células leucêmicas um rearranjo cromossômico característico. Esta anormalidade cromossômica pode ser detectada em metáfases adequadas, verificando-se a translocação de um segmento de um dos cromossomos do par número 17 para um dos cromossomos do par número 15 -  $t(15,17)$ . A Figura 3.7 mostra uma metáfase de um indivíduo do sexo masculino com leucemia. Através da análise dos pares de cromossomos 15 e 17, pode ser verificado que em um dos cromossomos do par 17 ocorreu a perda de um segmento no braço inferior  $q$  ( $17q^-$ ), enquanto que em um cromossomo do par número 15, ocorreu a soldadura deste segmento cromossômico ( $15q^+$ ).

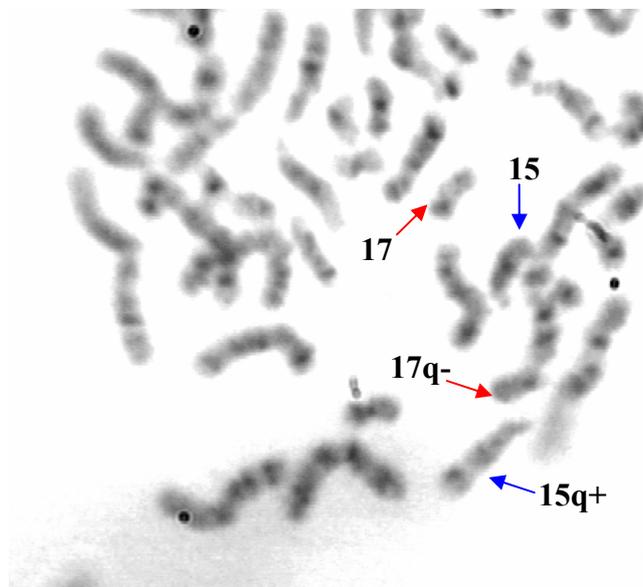


Figura 3.7 – Ilustração de uma metáfase contendo a translocação nos pares de cromossomos 15 e 17 -  $t(15,17)$ , caracterizando a Leucemia Pró-Mielocítica Aguda.

Para melhor visualização do rearranjo observado na Leucemia Pró-Mielocítica Aguda, a Figura 3.8 mostra quatro exemplos de pares de cromossomos 15 e 17 de células com este tipo de leucemia, pelos quais pode-se constatar mais facilmente a translocação  $t(15,17)$  (HURET e CHOMIENNE, 2003).

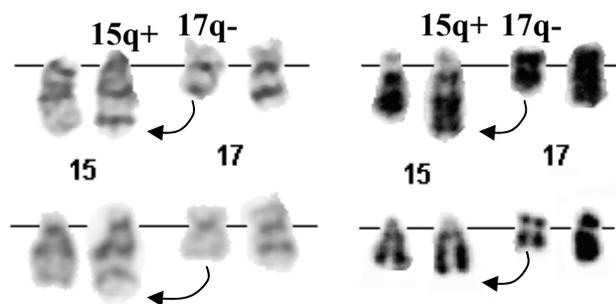


Figura 3.8 – Melhor visualização do prognóstico da Leucemia Pró-Mielocítica Aguda através de quatro exemplos da translocação –  $t(15,17)$ .

Na análise citogenética, algumas anormalidades estruturais são muito difíceis de serem detectadas através das técnicas convencionais sendo que, na grande maioria das vezes, somente as anomalias estruturais “maiores” são detectadas. Já na ocorrência de anormalidades estruturais discretas, a detecção se faz através do uso de técnicas de citogenética molecular. Para a análise da Leucemia Pró-Mielocítica Aguda não é diferente, exigindo muita habilidade e experiência técnica do profissional para a detecção do segmento translocado entre os cromossomos dos pares 15 e 17. Uma técnica muito utilizada para a detecção desta translocação consiste na análise de dois fatores simultâneos: na diferença de tamanho entre os cromossomos destes dois pares e na verificação de uma banda a mais em um dos cromossomos do par número 15 (15q+). A diferença no tamanho dos cromossomos homólogos dos pares 15 e 17 por si só não garante que exista uma translocação cromossômica, pois tal diferença pode ser causada também devido às fontes de incerteza (formas de espalhamento aleatório dos cromossomos na lâmina). Logo, busca-se também a verificação da presença de uma banda a mais em um dos cromossomos do par número 15, provinda do segmento translocado de um cromossomo do par número 17 (17q-), como pode ser constatado pelos exemplos mostrados na Figura 3.8. Entretanto, a detecção desta banda que aparece a mais em um dos cromossomos do par número 15

pode se tornar muito difícil, pois ela pode variar de intensidade às vezes mais escura ou mais clara.

Neste contexto, a avaliação citogenética da Leucemia Pró-Mielocítica Aguda pode vir a ser muito subjetiva, pois depende da análise de variáveis qualitativas e quantitativas, simultaneamente, permitindo que ocorra equívocos na sua caracterização. Para evitar que isso aconteça, quando uma possível translocação é detectada por um profissional técnico, é muito comum a necessidade da corroboração de outros citogeneticistas para a confirmação da doença, pois, se ocorrer um diagnóstico falso-positivo, o paciente será submetido a um tratamento agressivo (quimioterapia) sem necessidade. Sendo assim, é de suma importância a necessidade de encontrar novas técnicas para a detecção da Leucemia Pró-Mielocítica Aguda, que permita a sua análise com o mínimo de subjetividade.

### **3.9 – Considerações Finais**

Tendo em vista o exposto, este capítulo foi dedicado ao entendimento e a compreensão do objeto de estudo, servindo de base e fundamentação para a formulação do problema. Logo, foi apresentada uma breve descrição dos tópicos referentes à genética humana, que fazem parte do processo de identificação dos cromossomos humanos, tais como: os aspectos biológicos, o ciclo celular, a identificação dos cromossomos metafásicos, a identificação clássica dos cromossomos metafásicos, a identificação dos cromossomos por intermédio das bandas, a representação visual dos cromossomos denominada de cariograma, a preparação das células metafásicas como amostra, as fontes de incerteza no domínio do problema e, por fim, a Leucemia Pró-Mielocítica Aguda.

## **CAPÍTULO IV**

### **PROCESSAMENTO E ANÁLISE DE IMAGENS**

#### **4.1 – Introdução**

Pode-se dizer que o sentido da visão é o meio mais eficiente que o ser humano possui para absorver informações do ambiente que o cerca (FACON, 2002). A compreensão de como reage o sistema visual humano, através da capacidade para processar e interpretar imensas quantidades de dados de natureza visual, motiva o desenvolvimento de novas metodologias que possam dar suporte à criação de ferramentas artificiais de visão. Em aplicações onde se utilizam imagens na forma digital torna-se possível seu processamento computacional, e o uso das ferramentas artificiais de visão tem como objetivo aprimorar a imagem à aplicação, além de permitir a análise automática de uma determinada cena.

Assim, neste capítulo, busca-se trazer um caráter descritivo de algumas técnicas de processamento digital de imagens, com o objetivo de conferir o embasamento teórico para o desenvolvimento do método proposto neste trabalho. Os tópicos citados neste capítulo seguem a estrutura geral de um sistema de visão artificial, dentre eles: conceitos gerais, sistema de visão artificial, transformação, pré-processamento, segmentação, descrição, reconhecimento e interpretação. Ainda neste capítulo, serão apresentadas as considerações finais sobre o exposto.

#### **4.2 – Conceitos Gerais**

A área de processamento e análise de imagens age através dos recursos do computador, simulando experiências em uma cena de maneira a resolver um problema proposto, além de comparar entidades específicas com padrões previamente armazenados. Este campo da ciência vem sendo objeto de crescente interesse, pois, além de melhor adequar as informações pictóricas para interpretação humana, também permite a análise

automática por computador de informações extraídas de uma cena. Nestes dois casos, alguns autores, tais como MARQUES FILHO e VIEIRA NETO (1999), adotam apenas a expressão “processamento de imagens” para o primeiro e, “análise de imagens”, “visão por computador” e “reconhecimento de padrões”, para o segundo caso.

Pode-se verificar que as primeiras aplicações nesta área de processamento e análise de imagens remontam ao começo do século XX, visando transmitir imagens digitalizadas por cabo submarino entre Londres e Nova Iorque. Porém, o grande impulso para esta área viria com o advento dos primeiros computadores digitais de grande porte e o início do programa espacial norte-americano. A partir de então, a área de processamento e análise de imagens vem mostrando crescimento expressivo, englobando diversas áreas de interesse (biologia, medicina, comunicação visual, eletrônica, óptica, inteligência artificial, ciência da computação, matemática, entre outras), buscando desenvolver novas técnicas para implementar e explorar mais eficientemente os *sistemas de visão artificial*.

### 4.3 – Sistema de Visão Artificial

Um sistema de visão artificial, de uso genérico, deve conter determinados elementos que abrangem as principais operações que se podem efetuar sobre uma imagem, tais como: *aquisição*, *processamento*, *armazenamento* e *saída*, como mostra a Figura 4.1.

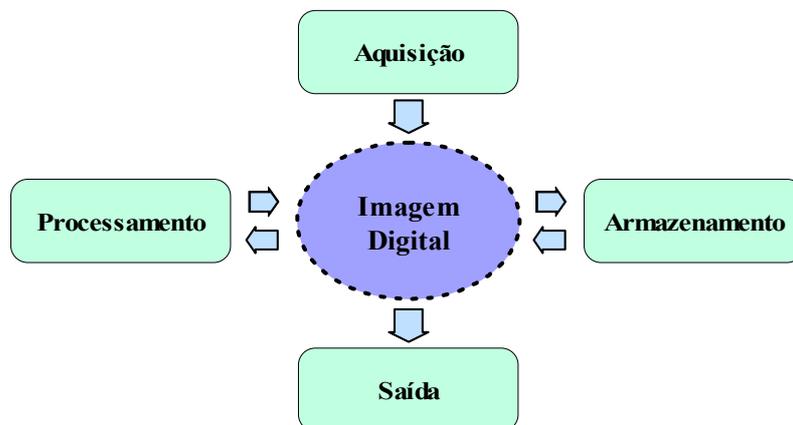


Figura 4.1 - Principais operações de um sistema de visão artificial de uso genérico.

O ponto central de discussão em processamento digital de imagens, como pode ser visto, é a imagem digital. Entretanto, não é possível que a imagem de uma cena real tridimensional seja colocada diretamente no computador e analisada, pois o computador trabalha com dados numéricos. Logo, há a necessidade da conversão desta imagem real em uma representação numérica adequada, que permita o processamento digital subsequente. Função esta desempenhada por dispositivos de aquisição, tais como câmeras digitais, scanners, etc. Já na etapa de saída, o monitor de vídeo é um elemento fundamental em um sistema de visão artificial.

Por definição, um sistema de visão artificial é um sistema computadorizado capaz de adquirir, processar e interpretar imagens correspondentes a cenas reais (MARQUES FILHO e VIEIRA NETO, 1999). Logo, a etapa de processamento permite modificar, analisar e manipular imagens digitais envolvendo procedimentos normalmente expressos sob forma algorítmica e, em função disto, a maioria das funções de processamento de imagens pode ser implementada via software. Assim, os sistemas de visão artificial dispõem de algoritmos de processamento de imagens, que são procedimentos passo a passo para otimizar as operações de tratamento e análise de imagens.

Segundo FACON (2002), toda a etapa de processamento digital de imagens pode ser estruturada, quanto ao grau de abstração, em três níveis distintos de processamento: *baixo*, *médio* e *alto*. Adotou-se, neste trabalho, que cada um destes níveis de abstração está ainda sub-dividido em duas etapas. O nível baixo, por exemplo, é estruturado através das etapas de *transformação* e *pré-processamento* da imagem, em que a entrada corresponde à imagem real original que é transformada na forma de valores numéricos associados a cada *pixel*, sendo que os dados de saída representam propriedades desta imagem já adequada à aplicação. No processamento de nível médio, este conjunto de valores produz como resultado uma lista de características, estruturado em *segmentação* e *descrição* da imagem. Já o processamento de nível alto produz, a partir destas características extraídas, a identificação e a compreensão do conteúdo da imagem, ou seja, estágio este constituído pelas etapas de *reconhecimento* e *interpretação*. A Figura 4.2 mostra esta estrutura funcional de um sistema de processamento e análise de imagens, através das etapas de transformação, pré-processamento, segmentação, descrição, reconhecimento e interpretação.

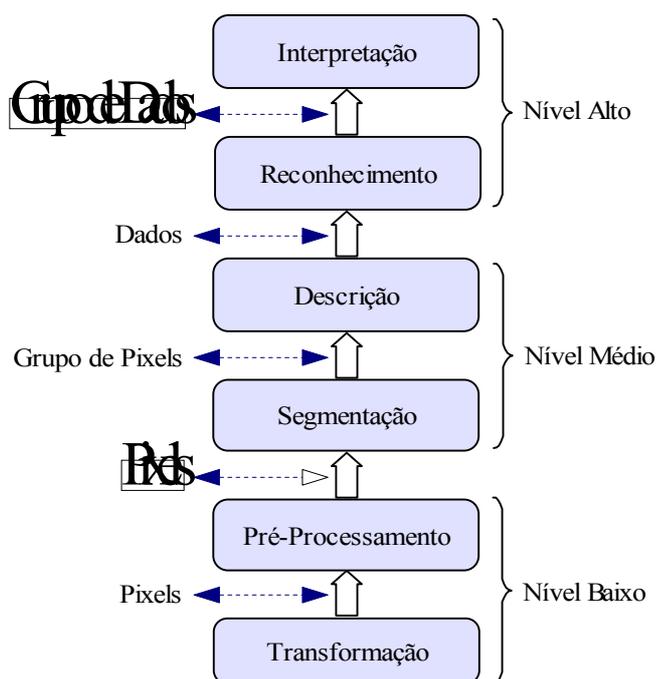


Figura 4.2 – Ilustração da estrutura funcional de um sistema de processamento e análise de imagens, através das etapas de transformação, pré-processamento, segmentação, descrição, reconhecimento e interpretação.

Os níveis de abstração baixo, médio e alto, através de suas respectivas etapas, serão apresentados com maiores detalhes e com embasamento teórico, dando maior ênfase aos aspectos que servirão de base para o desenvolvimento da metodologia de análise e classificação citogenética proposta neste trabalho.

#### 4.4 – Transformação

O estágio da transformação da imagem consiste na conversão de uma cena real tridimensional em uma imagem digital e, para tanto, são imprescindíveis duas etapas: a *captura* da imagem e sua *digitalização*. Os dispositivos que fazem esta conversão da imagem são compostos por dois elementos principais: o *sensor* e o *digitalizador*. O sensor é um dispositivo físico constituído por uma matriz de células semicondutoras fotossensíveis que transforma a energia luminosa (fótons) em um sinal elétrico analógico proporcional ao nível de energia detectado. Já o segundo elemento, o digitalizador, é um

dispositivo que converte o sinal elétrico analógico (imagem analógica) em representações de bits de 0's e 1's (processo de digitalização ou discretização). A imagem analógica é submetida, então, a uma discretização espacial (amostragem) e em amplitude (quantização) para ficar no formato desejável ao processamento computacional. A amostragem converte a imagem analógica em uma matriz de pontos de endereços  $(x,y)$ , denominados de *pixels*, pelos quais a claridade é medida e quantificada, e esse processo (quantização) gera um número inteiro representando a intensidade luminosa naquele ponto. Assim, uma imagem monocromática  $f(x,y)$  é uma imagem digital (matriz de *pixels*), sendo  $f$  uma função que atribui um valor (intensidade) de nível de cinza para cada par de coordenadas. Essa matriz de dados digitais está, então, pronta para o processamento no computador (MARQUES FILHO e VIEIRA NETO, 1999).

Atualmente, muitas câmeras digitais para captura de imagens estão utilizando como sensores a tecnologia *Charge-Coupled Device* (CCD), com o digitalizador na própria câmera, disponibilizando para o computador as imagens já no formato digital (TAKAHASHI, 2001).

#### 4.5 – Pré-Processamento

A imagem adquirida pelo sistema de aquisição é então recebida pelo computador para ser processada. A etapa de pré-processamento permite compensar distorções específicas da imagem que podem ocorrer na etapa de aquisição devido a *pixels* ruidosos e/ou brilho inadequado, tornando-a mais adequada do que a imagem original para as etapas subsequentes. E, para a melhoria destas imperfeições, são aplicados métodos para *suavizar* uma imagem. Tais métodos trabalham no domínio espacial (filtragem espacial) e atuam diretamente sobre a matriz de *pixels*  $f(x,y)$  utilizando operações orientadas à vizinhança. Logo, a suavização de imagens baseia-se em operações de convolução com máscaras definidas a uma certa vizinhança de um *pixel* de referência  $(x,y)$ , denominado de *pixel* central. Este processo provoca um leve borramento da imagem e a remoção de ruídos presentes nela.

A suavização da imagem é considerada como o efeito da filtragem passa-baixas, pois atenuam ou eliminam as componentes de alta frequência no domínio das

transformadas de Fourier, que correspondem às regiões de bordas, ruídos e/ou detalhes finos da imagem (ao contrário dos filtros passa-altas que realçam as bordas e as regiões de alto contraste da imagem). Dentre as técnicas mais conhecidas de suavização estão os *Filtros de Média e Mediano*.

A máscara para o Filtro de Média é construída com todos os coeficientes iguais a 1, dividindo o resultado da convolução por um fator de normalização, ou seja, cada *pixel* da imagem é substituído pela média ponderada dos *pixels* de sua vizinhança. A Figura 4.3 ilustra este conceito de máscaras de dimensão 3x3, 5x5 e 7x7 para o cálculo do Filtro de Média, sendo que quanto maior o tamanho da máscara, maior o grau de borramento da imagem resultante.

$$\frac{1}{9} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad \frac{1}{25} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \quad \frac{1}{49} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Figura 4.3 – Máscaras de dimensão 3x3, 5x5 e 7x7 para o cálculo do Filtro de Média.

Uma das principais limitações do Filtro de Média está na incapacidade dele preservar bordas e detalhes finos da imagem. Para contornar esta limitação, uma alternativa é o Filtro Mediano, onde o nível de cinza do *pixel* central da janela, mostrada pela Figura 4.4, é substituído pela mediana dos *pixels* que constituem a janela ao seu redor. Sendo assim, o filtro mediano pode ser construído a partir de uma imagem  $P$ , onde cada *pixel*  $Q_{ij}$  de uma nova imagem  $Q$  é calculado em função do *pixel*  $P_{ij}$  e seus *pixels* vizinhos.

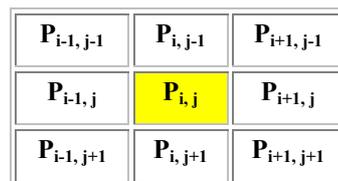


Figura 4.4 – Ilustração de uma janela 3x3, destacando o *pixel* central  $P_{ij}$  e seus *pixels* vizinhos.

O valor do *pixel*  $Q_{ij}$ , na nova imagem, é definido pela mediana dos valores ( $V_n$ ) dos 9 *pixels* da janela 3x3 apresentada pela Figura 4.4, ou seja, dos *pixels*  $P_{ij}$  e seus 8 *pixels* vizinhos, ordenados em ordem crescente ou decrescente em um vetor, pelo qual toma-se o valor do *pixel* que está na posição central dele. No exemplo da Figura 4.5, o valor do *pixel*  $Q_{ij}$ , na nova imagem, será o valor de  $V_5$ :

-	-	-	-	$Q_{i,j}$	-	-	-	-
$V_1$	$V_2$	$V_3$	$V_4$	$V_5$	$V_6$	$V_7$	$V_8$	$V_9$

Figura 4.5 – Determinação do valor do *pixel*  $Q_{ij}$  na nova imagem, definido pela mediana ( $V_5$ ) dos valores dos 9 *pixels* vizinhos de  $P_{ij}$ .

O filtro mediano apresenta desempenho particularmente bom em situações nas quais a imagem é contaminada por ruído impulsivo (*salt and pepper*) e, para situações onde o ruído é do tipo gaussiano, seu desempenho é apenas satisfatório, se comparado ao filtro de média. A Figura 4.6 ilustra o método de suavização da imagem de um cromossomo com a aplicação dos filtros de média e mediano, ambos com máscara 3x3.

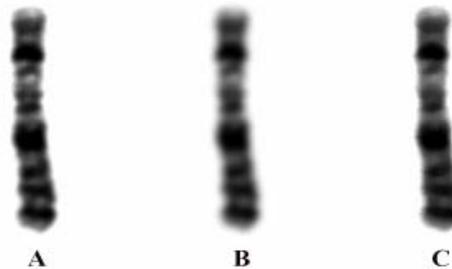


Figura 4.6 – Ilustração do método de suavização: (A) imagem original de um cromossomo, (B) aplicação do filtro de média e (C) aplicação do filtro mediano.

#### 4.6 – Segmentação

Em processamento digital de imagens torna-se muito comum a necessidade de se obter dados relacionados com os objetos presentes na imagem. Sendo assim, a operação

de segmentação procura dividir a imagem em partes construtivas, ou melhor, busca-se identificar *pixels* contíguos (regiões) distinguindo-os do fundo. Em geral, a separação da região de interesse é realizada através de duas estratégias: por descontinuidade, que são representadas pelas mudanças bruscas nos níveis de cinza como linhas e bordas, e por similaridade, que se baseia na semelhança entre os *pixels* que levam em conta limiares de níveis de cinza, subdivisão da imagem em regiões homogêneas e crescimento de regiões (FACON, 2002).

Não existe um modelo formal para a segmentação, o processo é essencialmente empírico e deverá se ajustar a diferentes tipos de imagens. O método mais comum de segmentação é a operação de limiarização (*thresholding*) por um tom de corte. É computacionalmente simples e utiliza uma determinada faixa de intensidade de cinza ( $T_1$  a  $T_2$ ) para identificar os *pixels*. O resultado é uma imagem binarizada, atribuindo-se a todos os *pixels* com nível de cinza dentro dessa faixa de intensidade como pertencente ao objeto, e todos os *pixels* com nível de cinza fora dessa faixa como não pertencentes ao objeto.

A operação de segmentação, através do método de limiarização, é baseada na utilização do *histograma*. O histograma é simplesmente um conjunto de números, indicando o percentual de *pixels* naquela imagem que apresentam um determinado nível de cinza. Estes valores são normalmente representados por um gráfico de barras que fornece, para cada nível de cinza, o número (ou o percentual) de *pixels* correspondentes na imagem. Esse número é calculado através da Equação 4.1:

$$p_r(r_k) = n_k / n \quad (4.1)$$

onde  $n_k$  é o número de *pixels* para o nível de cinza  $r_k$ ,  $n$  o número total de *pixels* na imagem e  $p_r(r_k)$  a probabilidade do  $k$ -ésimo nível de cinza. Para ilustrar um exemplo da operação de segmentação, considerando o conceito de limiarização utilizando-se o histograma, a Figura 4.7 apresenta uma imagem de uma célula metafásica e seu respectivo histograma, com faixa de intensidade de corte em  $T_1 = 0$  (*Min*) e  $T_2$  igual 200 (*Máx*), obtendo-se como resultado uma imagem segmentada.

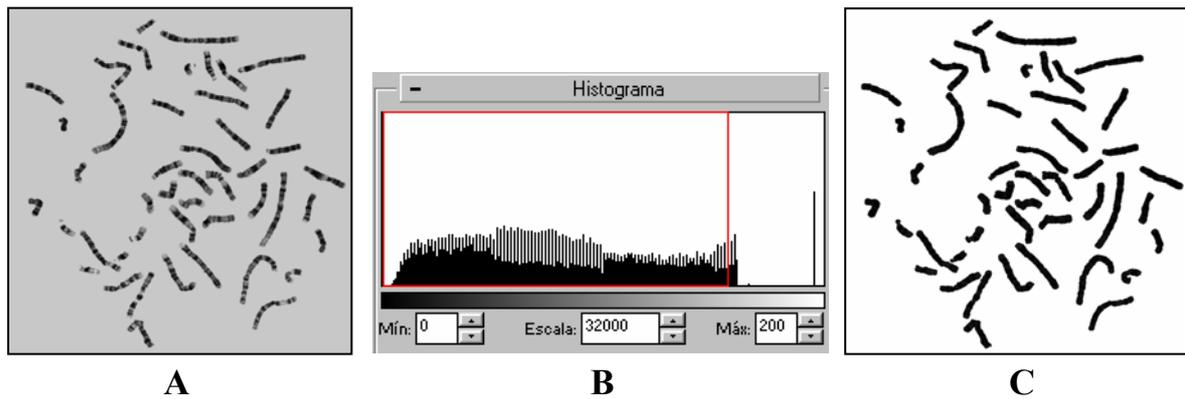


Figura 4.7 – Ilustração da operação de segmentação considerando o conceito de limiarização utilizando-se o histograma, onde (A) mostra a imagem de uma célula metafásica, (B) seu respectivo histograma com faixa de intensidade de corte em  $T_1 = 0$  (*Mín*) e  $T_2 = 200$  (*Máx*) e (C) a imagem segmentada.

Em algumas situações, regiões fisicamente distintas podem ter tonalidades de níveis de cinza muito próximas, impossibilitando sua segmentação por limiarização comum. Caso estas regiões possuam cores distintas, pode-se utilizar um dos modelos de cores existentes, como por exemplo, o modelo aditivo RGB (cores primárias *Red*, *Green* e *Blue*). Neste caso, a imagem é decomposta em suas componentes de cor e a segmentação é realizada através do histograma de uma dessas componentes, ou então pelos histogramas das três componentes simultaneamente. A escolha de limiares no histograma de cada componente primária selecionará uma faixa de cores que poderá segmentar a região de interesse.

A Figura 4.8 mostra o exemplo de decomposição de imagens coloridas utilizando-se o modelo RGB, em que a imagem de uma célula metafásica foi decomposta em suas componentes primárias desse modelo e, em seguida, foram extraídos os histogramas correspondentes de cada componente.

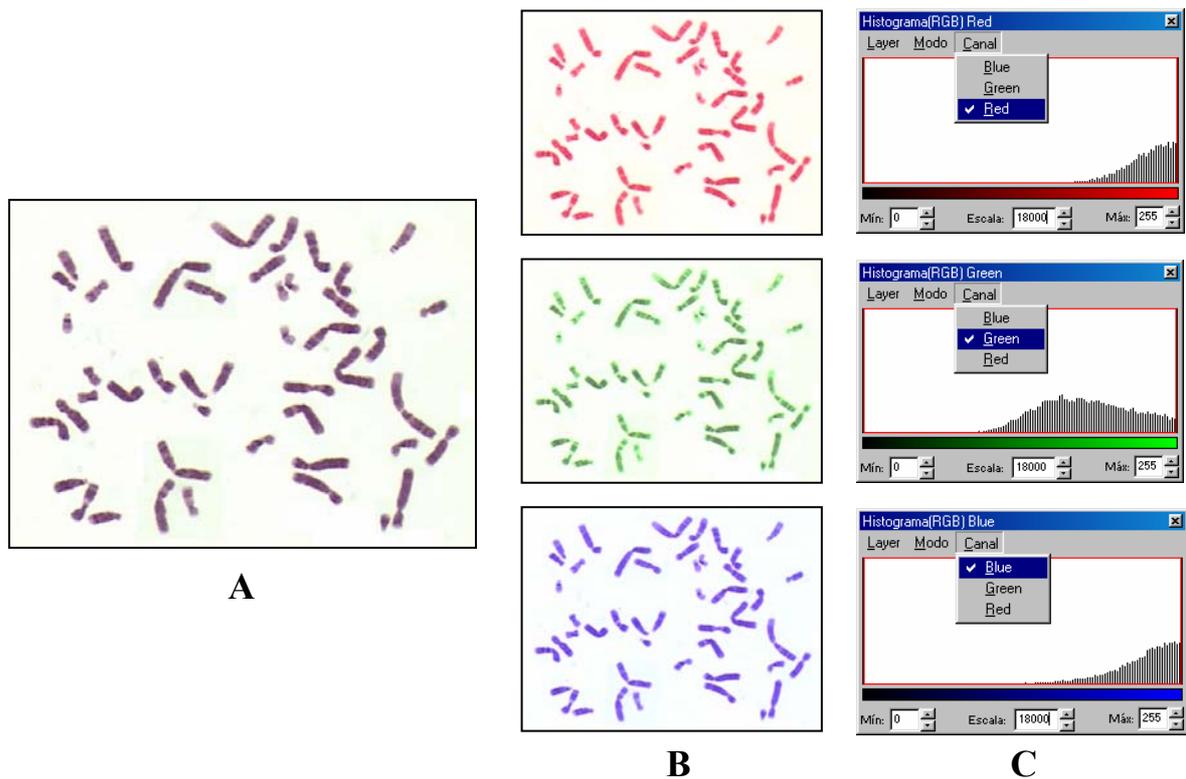


Figura 4.8 – Exemplo da decomposição de uma imagem colorida utilizando-se o modelo RGB: em (A) a imagem de uma célula metafásica, (B) decomposição das componentes deste modelo e em (C) os histogramas correspondentes de cada componente.

Uma outra técnica que também permite o reconhecimento dos objetos em uma imagem é a operação de realce e detecção de bordas (*edge*). As bordas são regiões da imagem, indicando o limite entre duas regiões (objeto e o fundo), cujos níveis de cinza predominantes, nestas regiões, são razoavelmente diferentes, ou seja, ocorre uma mudança no nível de cinza (descontinuidade) em um certo intervalo do espaço e em uma certa direção, ou quando o gradiente de uma imagem tem uma variação abrupta (FACON, 2002).

Os filtros utilizados para o realce e a detecção de bordas permitem gerar uma imagem em que as bordas aparecem com alta intensidade, ao passo que as regiões internas ao objeto aparecerão com baixa intensidade. Aplicam-se habitualmente filtros espaciais lineares de dois tipos: os filtros baseados no *gradiente* da função de luminosidade da imagem e os filtros baseados no *laplaciano*, e ambos costumam ser aproximados por máscaras de convolução 3x3 (MARQUES FILHO e VIEIRA NETO, 1999). Exemplos

destas máscaras utilizadas para estimar a amplitude do gradiente de uma borda são os operadores verticais e horizontais de *Sobel* e *Prewitt*, como mostrado na Figura 4.9.

	Vertical	Horizontal
<b>Sobel</b>	$\frac{1}{4} \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$	$\frac{1}{4} \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$
<b>Prewitt</b>	$\frac{1}{3} \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$	$\frac{1}{3} \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$

Figura 4.9 – Operadores verticais e horizontais de *Sobel* e *Prewitt* aplicados para realçar e detectar bordas.

A Figura 4.10 ilustra os resultados da aplicação dos operadores de *Sobel* e *Prewitt* para o realce e a detecção de bordas em uma imagem com 256 níveis de cinza.

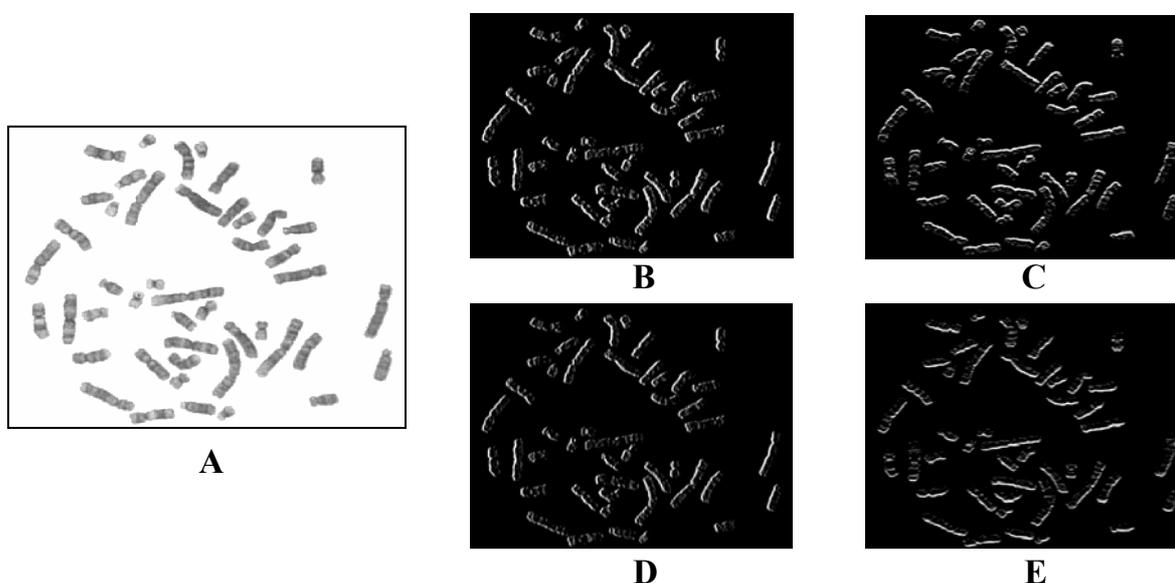


Figura 4.10 – Técnica de detecção de bordas em uma imagem com 256 níveis de cinza: (A) utilizando operadores de Sobel vertical (B) e horizontal (C), e Prewitt vertical (D) e horizontal (E).

Os operadores verticais e horizontais de Sobel e Prewitt podem ainda serem combinados por meio de uma operação lógica OR, o que resulta em apenas uma iteração.

No caso dos filtros baseados no laplaciano, são operadores derivativos de segunda ordem definidos pela Equação 4.2 (GONZALEZ e WOODS, 1992):

$$\nabla^2 f(x,y) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \quad (4.2)$$

e que pode ser aproximado pelas máscaras 3x3 e 5x5 da Figura 4.11.

$$\begin{array}{c} \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \\ \mathbf{A} \end{array} \quad \begin{array}{c} \begin{bmatrix} -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 24 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \end{bmatrix} \\ \mathbf{B} \end{array}$$

Figura 4.11 – Máscaras para o cálculo do laplaciano: em (A) 3x3 e (B) 5x5.

O operador baseado no laplaciano é insensível à rotação, ou seja, é capaz de realçar ou detectar bordas em qualquer direção. A Figura 4.12 ilustra um exemplo da aplicação do operador laplaciano 3x3 para uma imagem monocromática.

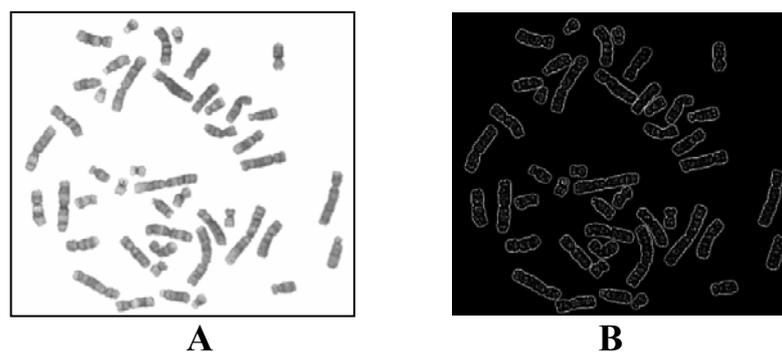


Figura 4.12 – Exemplo da aplicação de realce e detecção de bordas: (A) imagem monocromática e (B) operador laplaciano 3x3.

Existem várias outras técnicas para a detecção de bordas, tais como os operadores de *Roberts*, *Robinson*, *Canny*, *Marr-Hildreth*, etc., e a eficiência de uma sobre outra

depende da aplicação. Um fato que pode gerar limitações em algumas técnicas é que, para determinadas aplicações, não se pode garantir que as fronteiras identificadas serão curvas fechadas. Logo, a extração das características e o cálculo das medidas dos objetos como área, perímetro, forma, etc., ficam impossibilitadas. A Figura 4.13 mostra um exemplo da aplicação do operador laplaciano 3x3, onde podem ser vistas bordas com fronteiras totalmente fechadas (detalhe através do zoom) em uma imagem contendo cromossomos metafásicos que, para tanto, houve a necessidade da aplicação do método de limiarização, seguido do filtro mediano e, por fim, a inversão da escala de níveis de cinza.

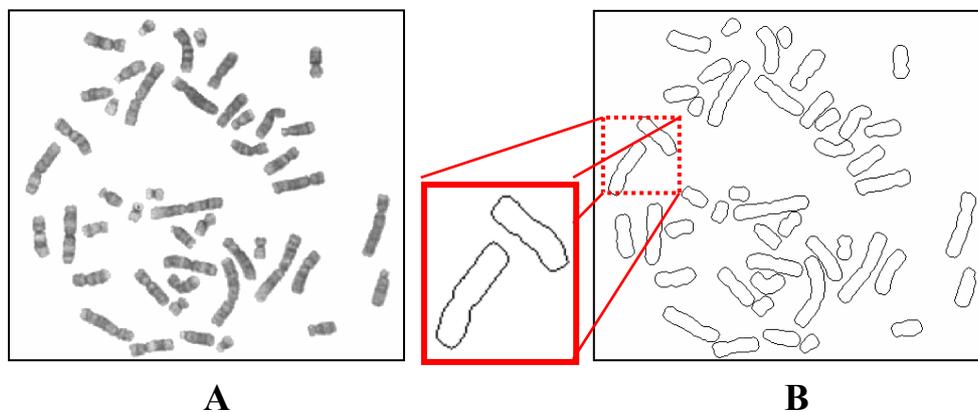


Figura 4.13 – Exemplo da aplicação do operador laplaciano 3x3 para a detecção de bordas com fronteiras totalmente fechadas (detalhe do zoom). Em (A) uma imagem contendo cromossomos metafásicos e em (B) imagem resultante com a detecção das bordas e fronteiras totalmente fechadas.

Como pode ser visto, o processo de detecção dos objetos de interesse na imagem é um dos passos mais importantes do processamento de imagens, pois dos objetos identificados serão extraídos e calculados uma série de atributos que os descreverão e que servirão de base para o seu reconhecimento.

#### 4.7 – Descrição

O objetivo da etapa de descrição é extrair características (descritores ou atributos) das regiões de interesse (objetos) resultantes do processo de segmentação. E tais atributos, extraídos dos objetos, devem ser representados por uma estrutura de dados adequada ao

algoritmo de reconhecimento. Observa-se que a entrada ainda é uma imagem ou determinadas regiões da imagem (grupo de *pixels*), mas a saída é um conjunto de dados correspondentes àquela imagem (MARQUES FILHO e VIEIRA NETO, 1999).

Essa etapa constitui, então, o processo de mensuração das “formas” resultantes da segmentação (FACON, 2002). Existem basicamente duas classes de medidas: as medidas de campo (*field features*), que se referem ao campo como um todo (número de objetos na imagem, área total dos objetos, etc.), e as medidas de região (*region features*), que se referem aos objetos independentemente (tamanho, forma, intensidade, etc.). Para estas últimas medidas, podem ser aplicadas técnicas muito simples como também métodos muito sofisticados, permitindo uma nova separação dos objetos em classes de similaridades, em função dos parâmetros medidos.

Existem centenas de parâmetros que podem ser extraídos e utilizados para definir um objeto e, a partir destes, podem-se ainda obter outras medidas que poderão descrevê-los mais precisamente. A forma mais simples de se extrair o tamanho de um objeto, por exemplo, é através da área ( $A$ ), que pode ser facilmente calculada pela contagem do número de *pixels* pertencentes ao objeto. Outro parâmetro de tamanho também pode ser extraído através do perímetro ( $P$ ) e suas medidas não são tão simples de serem obtidas como no caso da área, pois depende da análise dos *pixels* vizinhos na direção horizontal, vertical e diagonal, definida pela Equação 4.3 (PACIORNIK, 2003).

$$P = N + \sqrt{2 * N_d} \quad (4.3)$$

onde  $N$  é o número de passos horizontais ou verticais e  $N_d$  é o número de passos diagonais.

Na representação de um objeto quanto à sua forma estrutural, pode-se aplicar uma medida de similaridade de um objeto com um corpo específico (*Fator de Forma*), como também reduzir o objeto para uma versão mais simplificada (*Afinamento*). Para exemplificar o primeiro caso, pode-se extrair um fator de forma circular ( $FF$ ) através da Equação 4.4 (PACIORNIK, 2003).

$$FF = \frac{4\pi * A}{(P)^2} \quad (4.4)$$

A segunda alternativa para se obter parâmetros de forma dos objetos, o afinamento (*thinning*), mostra a aplicabilidade em reduzir padrões a uma fina linha representativa, podendo ser atribuído não apenas à análise estrutural mais simples dos objetos, bem como à necessidade de se reduzir a quantidade de informação a ser processada. Neste trabalho somente serão citados os processos de afinamento direcionados para imagens binárias (valores de *pixel* 1 pertencentes ao fundo e 0 para *pixels* pertencentes a um objeto).

Em determinadas aplicações, o método de afinamento pode ser realizado utilizando-se técnicas morfológicas. Assim, o afinamento de uma região  $A$ , denotado por  $S(A)$ , pode ser obtido através da união do subconjunto  $S_k(A)$ , ou seja, em termos de erosões e aberturas, de acordo com a Equação 4.5 (GONZALEZ e WOODS, 1992).

$$S(A) = \bigcup_{k=0}^K S_k(A) \quad (4.5)$$

com

$$S_k(A) = \bigcup_{k=0}^K \{(A \ominus kB) - [(A \ominus kB) \circ B]\}$$

onde  $B$  é um elemento estruturante,  $(A \ominus kB)$  indica  $k$  erosões sucessivas de  $A$ ,  $(\circ B)$  abertura com o elemento estruturante, e  $K$  o último passo iterativo antes de  $A$  resultar em um conjunto vazio. A Figura 4.14 ilustra este processo de afinamento utilizando técnicas morfológicas, onde sucessivas camadas do revestimento externo de um objeto (cromossomo) são removidas, resultando em uma estrutura mais fina e reduzida.



Figura 4.14 – Aplicação da técnica de sucessivas erosões e aberturas: (A) imagem de um cromossomo, (B) imagem binarizada, (C) após a remoção de sucessivas camadas e (D) resultado final do afinamento.

A imagem resultante desse tipo de afinamento, utilizando operações de erosões e aberturas, é mais espessa do que deveria ser, além de remover a conectividade dos elementos da imagem original (MARQUES FILHO e VIEIRA NETO, 1999). Resultados de afinamento melhores foram conseguidos com o método de transformação *Hit-or-Miss* (*hom*), sendo uma operação morfológica genérica e também executada pela translação da origem de um elemento estruturante por todos os pontos da imagem. Esse elemento utilizado na técnica *hom* é uma ligeira extensão do tipo que é introduzido para o processo de erosão e dilatação, sendo que este pode conter pixels do objeto e do fundo, ao invés do objeto apenas (MIRA JÚNIOR, 2003). Ainda assim, novas técnicas foram e estão sendo desenvolvidas na busca de resultados mais eficientes, levando em conta satisfazer três requisitos:

- a) remover todos os *pixels* redundantes, produzindo uma nova imagem com largura de um único *pixel*;
- b) não violar a conectividade da imagem original;
- c) manter as propriedades significativas dos padrões (geométricas e topológicas).

Conseguir esses requisitos utilizando um algoritmo simples e rápido ainda é um desafio. O termo afinamento é geralmente utilizado para denotar a operação de redução de um determinado objeto para um grafo (coleção de arcos e curvas finas), entretanto, alguns autores utilizam termos diferentes para este processo, eixo médio (*medial axis*) e esqueletização (*skeletonization*) tornaram-se os mais utilizados (GONZALEZ e WOODS, 1992). Já o termo “esqueleto” é utilizado para referir-se ao resultado do algoritmo.

De acordo com FRANÇA e Von WANGENHEIM (2003), os algoritmos de esqueletização procedem de forma similar, ou seja, excluem, de forma sucessiva, diversas camadas da extremidade (borda) de um padrão até que apenas um esqueleto permaneça. A exclusão de um ponto  $p$  dependerá dos *pixels* na vizinhança deste ponto. De acordo com o modo de como se examinam os *pixels*, estes algoritmos podem ser classificados como seqüenciais ou paralelos. Nos algoritmos seqüenciais, os *pixels* são examinados para exclusão em uma seqüência fixa em cada iteração, e a exclusão de  $p$  na  $n$ -ésima iteração depende de todas as operações que tenham sido realizadas até aquele momento, isto é, do resultado da  $(n-1)$ -ésima iteração, bem como dos *pixels* já processados na  $n$ -

ésima iteração. Por outro lado, nos algoritmos paralelos, a exclusão na  $n$ -ésima iteração depende apenas dos *pixels* da iteração  $n-1$ . Por isso, todos os *pixels* podem ser analisados independentemente de forma paralela a cada iteração.

De modo geral, os algoritmos seqüenciais geram esqueletos melhores (em termos da conectividade e da conservação da topologia). Um exemplo de algoritmo seqüencial é o método através da transformação do eixo médio (*Medial Axis Transformation*), proposto em 1967 por *Blum* (CHOI *et al.*, 1997). Segundo REMY e THIEL (2002), o eixo médio pode ser definido como o conjunto de centros de discos máximos inclusos no objeto e obtido através da análise de cada *pixel* ( $p$ ) pertencente ao mesmo. Para cada *pixel*  $p$ , encontra-se o vizinho mais próximo (Distância Euclidiana) nos limites do objeto (borda). Se  $p$  tem mais de um *pixel* vizinho, ele pertence, então, ao eixo médio, como mostrado na Figura 4.15.

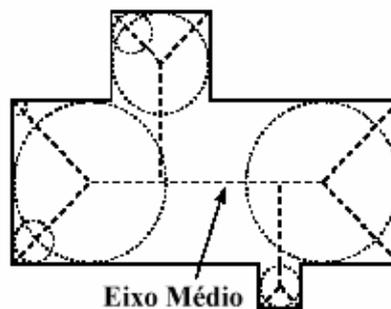


Figura 4.15 – Processo de extração do eixo médio pelos centros de máximos discos.

Contudo, o algoritmo baseado na transformação do eixo médio requer um grande esforço computacional, pois sua implementação envolve o cálculo da distância de todos os pontos interiores de uma imagem para todos os pontos da borda da mesma.

Já no processo de afinamento através dos algoritmos paralelos, os *pixels* são examinados para a exclusão baseados apenas na iteração anterior e podem ser excluídos ao mesmo tempo. Os algoritmos paralelos caracterizam-se pela velocidade, mas os algoritmos completamente paralelos podem ter dificuldades em conservar a conectividade das imagens. Em 1984, *Zhang* e *Suen* propuseram um novo tipo de algoritmo paralelo de esqueletização, o algoritmo *Zhang-Suen*, trazendo resultados muito superiores comparados

a outros da sua época. A partir de então, foram propostos novos métodos com base na técnica de *Zhang-Suen*, contribuindo ainda mais para melhorar os resultados no processo de esqueletização e, ainda hoje, vários métodos ainda são desenvolvidos baseados no trabalho original desses pesquisadores (FACON, 2003a).

A idéia básica do método *Zhang-Suen* é decidir se um determinado *pixel* central (*C*), de uma janela 3x3 apresentada pela Figura 4.16, será removido analisando somente seus oito *pixels* vizinhos. Os *pixels* que devem ser removidos são primeiramente marcados e posteriormente removidos, sendo que este processo é repetido até que não existam mais *pixels* redundantes, ou seja, até que os *pixels* remanescentes sejam aqueles que pertençam ao esqueleto do objeto. Para tanto, o algoritmo é dividido em duas *sub-iterações*: na primeira *sub-iteração*, o *pixel* central (*C*) é marcado para a deleção dependendo do valor verdade das seguintes condições:

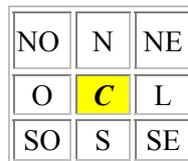


Figura 4.16 – Janela 3x3, destacando o *pixel* central (*C*) e seus *pixels* vizinhos.

1. *Se o número de conectividade for igual a 1* – ele é definido como sendo o número de transições de branco (fundo) para preto (objeto) nos *pixels* que circundam o *pixel* central, iniciando em *N* e terminando em *NO*, que deve ser *exclusiva* uma. Isto significa que o *pixel* central está conectado somente a uma única região e, se o número de conectividade for igual a 2, então, duas regiões estarão sendo conectadas pelo *pixel* central. Logo, esta regra assegura que, se o *pixel* em questão for eliminado, não fará com que diferentes regiões ligadas por ele passem a ficar separadas;

2. *Existem, ao menos, dois pixels vizinhos pretos, e não mais do que seis* - o número de *pixels* vizinhos refere-se também aos *pixels* na faixa *N,...,NO*. Esta regra assegura que os *pixels* resultantes serão retirados sucessivamente das bordas do objeto, e não de suas partes internas;

3. *Ao menos um dos pontos N, L e O sejam fundo da imagem (branco), e;*

4. *Ao menos um dos pontos N, S e O sejam fundo da imagem (branco).*

No final desta primeira sub-iteração, os pixels marcados são removidos. A próxima sub-iteração é a mesma para os passos 1 e 2, e diferencia para os passos 3 e 4, logo:

1. O número de conectividade é 1;
2. Existem ao menos dois pixels vizinhos pretos, e não mais do que seis;
3. Ao menos um dos pontos  $N$ ,  $L$  e  $S$  sejam fundo da imagem (branco);
4. Ao menos um dos pontos  $L$ ,  $S$  e  $O$  sejam fundo da imagem (branco);

E também no final desta segunda sub-iteração, cada um dos *pixels* marcados é removido. Quando não houver mais *pixels* a serem removidos, a esqueletização está completa. Apesar da simplicidade, este algoritmo não gera esqueletos de um *pixel* de largura, linhas diagonais de dois *pixels* de largura podem ser seriamente erodidas, e quadrados de 2x2 são completamente eliminados do esqueleto (FRANÇA e Von WANGENHEIM, 2003). A Figura 4.17 mostra os resultados obtidos com o processo de esqueletização utilizando o algoritmo *Zhang-Suen*, e também destaca a limitação deste método em se obter esqueletos de um *pixel* de largura.

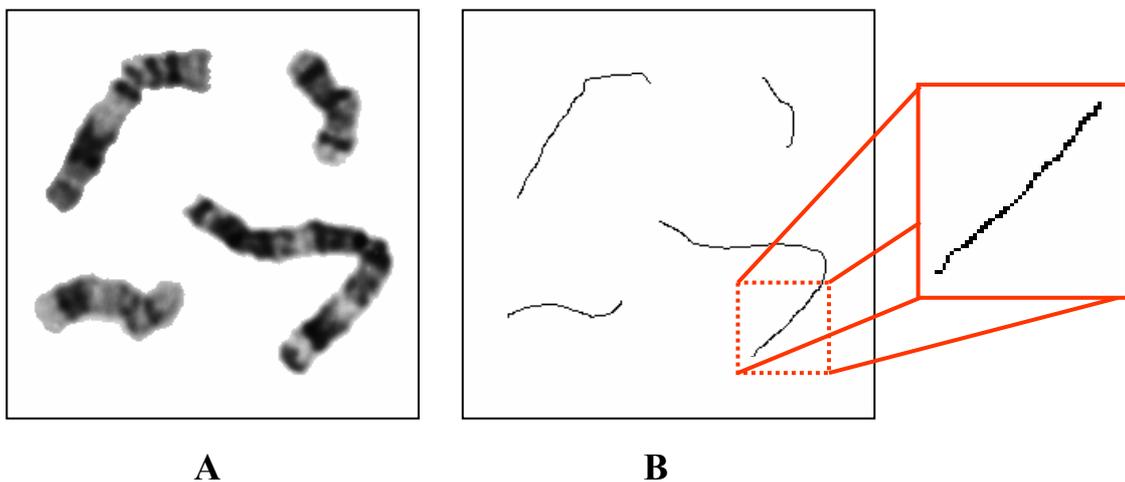


Figura 4.17 – Resultados obtidos com o processo de esqueletização utilizando o algoritmo Zhang-Suen: (A) imagem original, (B) esqueleto dos objetos (cromossomos) em que, no detalhe do zoom, é mostrada a limitação deste método em se obter esqueletos de um pixel de largura.

A partir da dedução do algoritmo de *Zhang-Suen*, *Holt* em 1987 sugeriu um algoritmo mais rápido e que não envolve iterações, em que as duas sub-iterações do

algoritmo de *Zhang-Suen* são escritas como expressões lógicas utilizando-se a vizinhança 3x3 sobre um ponto corrente (FACON, 2003b). Logo, a primeira sub-iteração foi reescrita utilizando-se a Equação 4.6, e a segunda pela Equação 4.7:

$$V(C) \wedge (\sim \text{borda}(C) \vee (V(L) \wedge V(S) \wedge (V(N) \vee V(O)))) \quad (4.6)$$

$$V(C) \wedge (\sim \text{borda}(C) \vee (V(O) \wedge V(N) \wedge (V(S) \vee V(L)))) \quad (4.7)$$

onde a função  $V()$  representa o valor do ponto e o resultado é verdadeiro se o ponto pertence ao objeto, falso, caso contrário. Já a função  $\text{borda}()$ , o resultado é verdadeiro se o ponto estiver na borda do objeto e, falso, caso contrário. O símbolo  $(\wedge)$  assume a função do operador lógico AND,  $(\vee)$  o operador lógico OR e  $(\sim)$  o operador de negação. Um ponto de borda é um ponto que atende às condições de conectividade igual a 1 e possui de dois a seis *pixels* vizinhos pretos. Se o resultado das expressões lógicas for falso, o ponto é apagado, caso contrário ele permanece. *Holt* combinou as duas expressões lógicas acima e obteve uma única expressão, definida pela Equação 4.8:

$$V(C) \wedge \left( \begin{array}{l} \sim \text{borda}(C) \vee (\text{borda}(L) \wedge V(N) \wedge V(S)) \vee \\ \vee (\text{borda}(S) \wedge V(O) \wedge V(L) \vee (\text{borda}(L) \wedge \text{borda}(SE) \wedge \text{borda}(S))) \end{array} \right) \quad (4.8)$$

O resultado do algoritmo *Zhang-Suen* ganha em rapidez e pode ser ainda melhorado quanto à limitação de se obter o esqueleto com apenas um pixel de largura, utilizando-se o método *staircase removal* (remoção de escada). O processo de *staircase removal* consiste da seguinte observação: metade dos pontos que apresentam uma forma semelhante a uma escada pode ser removida, sem afetar o formato ou a conectividade do objeto. Sendo assim, o *pixel* central das quatro máscaras, ilustradas através da Figura 4.18, pode ser apagado se um dos valores  $X$  for 0 (branco) (FACON, 2003b). E a expressão lógica para as máscaras 1 e 2 é definida pela Equação 4.9, e para as máscaras 3 e 4 pela Equação 4.10:

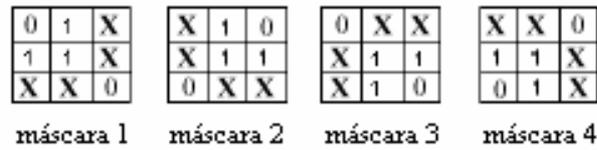


Figura 4.18 – Máscaras utilizadas no método de *staircase removal*, onde o *pixel* central pode ser apagado se um dos valores *X* for igual a 0 (branco).

$$V(C) \wedge \sim \left( V(N) \wedge \left( V(L) \wedge \sim V(NE) \wedge \sim V(SO) \wedge (\sim V(O) \vee \sim V(S)) \vee \vee (\sim V(O) \wedge \sim V(NO) \wedge \sim V(SE) \wedge (\sim V(L) \vee \sim V(S))) \right) \right) \quad (4.9)$$

$$V(C) \wedge \sim \left( V(S) \wedge \left( V(L) \wedge \sim V(NE) \wedge \sim V(SO) \wedge (\sim V(O) \vee \sim V(N)) \vee \vee (\sim V(O) \wedge \sim V(NO) \wedge \sim V(SE) \wedge (\sim V(L) \vee \sim V(N))) \right) \right) \quad (4.10)$$

A Figura 4.19 mostra uma comparação dos resultados obtidos com a utilização do algoritmo *Holt*, em que, num dos casos, não foi utilizado o método *staircase removal* resultando em um esqueleto com mais de um *pixel* de largura e em outro caso em que foi utilizado o método *staircase removal*, resultando em um esqueleto com um melhor afinamento.

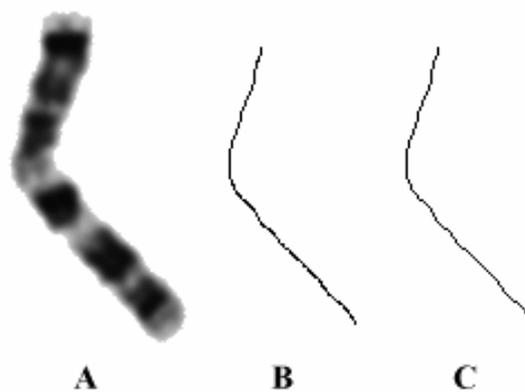


Figura 4.19 – Resultados obtidos com o processo de esqueletização utilizando o algoritmo *Holt*: (A) imagem original, (B) esqueleto extraído sem o método *staircase removal* e (C) um esqueleto com um melhor afinamento aplicando-se o método *staircase removal*.

Apesar de gerarem um esqueleto mais perfeito, devido ao tempo de processamento os algoritmos seqüenciais são pouco utilizados. Na prática, os algoritmos paralelos são bem mais aplicados, como no caso do algoritmo *Zhang-Suen e Holt*, pois com a utilização de outras técnicas em conjunto (*staircase removal*), pode-se conseguir excelentes resultados (esqueletos com um afinamento melhor) superando as limitações dos algoritmos paralelos tradicionais.

#### **4.8 – Reconhecimento**

Dado um conjunto de objetos segmentados, a partir dos quais podem-se extrair e medir diversos parâmetros (padrões), busca-se, nesta etapa, distinguir (atribuir um rótulo) cada objeto encontrado na imagem de forma automática. Logo, o reconhecimento de padrões (classificação) tem como finalidade o agrupamento por classes de similaridade dos diversos objetos obtidos na segmentação, baseando-se nos padrões que descrevem os objetos.

Este processo é muito complexo e existem diversos níveis de automação. Os mais simples implicam em processos de agrupamento estatístico, para os quais a decisão humana é fundamental. Os mais sofisticados permitem ao computador reconhecer diversos objetos através de técnicas de inteligência artificial com pouca ou nenhuma intervenção humana. Neste trabalho em particular, será utilizado, como método de reconhecimento de padrões, o enfoque Lógico-Combinatório, descrito com um detalhamento maior no capítulo seguinte.

#### **4.9 – Interpretação**

Realizada a etapa de reconhecimento dos objetos, a partir de seus padrões extraídos e medidos, pode-se agora tomar decisões e relatar fatos relacionados com os objetos do mundo real, ponderando sempre através de uma medida de erro de classificação. Esta etapa constitui, então, a interpretação e a compreensão do fenômeno inicial, ou melhor, consiste em atribuir um significado a um determinado conjunto ou classe de objetos

reconhecidos. Faz uso do conhecimento a priori do caso estudado e do conhecimento adquirido durante as fases precedentes.

#### **4.10 – Considerações Finais**

Tendo em vista o exposto, o principal objetivo desse capítulo foi fazer uma breve descrição da estrutura geral de um sistema de visão artificial, baseando-se nos elementos que abrangem as principais operações que podem ser efetuadas sobre uma imagem. Logo, esta descrição consiste na fundamentação teórica dos métodos de processamento digital de imagens utilizados para o desenvolvimento deste trabalho. Os tópicos descritos foram os conceitos gerais, o sistema de visão artificial, a transformação, o pré-processamento, a segmentação, a descrição, o reconhecimento e, por fim, o tópico interpretação.

## CAPÍTULO V

### ENFOQUE LÓGICO-COMBINATÓRIO

#### 5.1 – Introdução

Na solução de problemas práticos de determinadas áreas, tais como medicina, biologia, psicologia, etc., que constituem as áreas das ciências pouco formais, tornou-se muito comum pelos profissionais o uso do conceito de analogia como ferramenta metodológica para caracterizar os objetos de estudo. Contudo, para que se possa descrever eficientemente um problema, é necessária a utilização de técnicas de modelagem matemática para a criação de uma função de analogia que permita representar tais objetos. Logo, muitas pesquisas estão sendo desenvolvidas visando criar novas metodologias de modelagem matemática, para que possam ser utilizadas como ferramentas para a obtenção de uma regra de decisão ótima na identificação e na solução de problemas práticos.

Portanto, o objetivo deste capítulo é de apresentar uma breve descrição sobre a metodologia de modelagem matemática, baseando-se no enfoque lógico-combinatório da teoria de reconhecimento de padrões. Esta descrição está estruturada nos seguintes tópicos: aspectos gerais, formulação do problema, formalização do problema, seleção de variáveis, de critério de analogia, critério de comparação dos valores de uma variável e funções de semelhança entre objetos, exposição formal de um problema de seleção de variáveis, conceitos da teoria de testes, importância informacional das variáveis, modo de solução do problema expresso em termos matemáticos e, por fim, as considerações finais sobre o que foi exposto neste capítulo.

#### 5.2 – Aspectos Gerais

O termo reconhecimento de padrões, como muitos outros da ciência moderna, recebe inúmeras interpretações devido ao fato de englobar diferentes ramos das áreas de Engenharia, Computação e Matemática. Logo, pode-se encontrar aplicações da teoria de

reconhecimento de padrões em processamento de imagens, processamento de sinais, análise e interpretação de imagens, análise e interpretação de sinais, percepção remota, redes neurais para reconhecimento de padrões, algoritmos genéticos em reconhecimento de padrões, técnicas de inteligência artificial para reconhecimento de padrões, morfologia matemática, reconhecimento estatístico, reconhecimento sintático-estrutural e reconhecimento lógico-combinatório (SHULCLOPER *et al.*, 1997).

Do ponto de vista histórico, a origem da lógica-combinatória da teoria de reconhecimento de padrões deve-se à fusão de experiências de algumas comunidades científicas, tais como a Filial Siberiana do Centro de Cálculo da Academia de Ciências da antiga União Soviética, o Grupo de Reconhecimento de Padrões do Instituto de Cibernética, Matemática e Física de Cuba além, também, de um importante foco de desenvolvimento no México. É uma ciência de caráter multidisciplinar, que se ocupa com processos relacionados com objetos físicos e abstratos, com o propósito de extrair determinadas informações que permitam estabelecer propriedades e/ou vínculos entre os conjuntos de objetos de estudo.

Portanto, este trabalho visa apresentar a interpretação de forma ampla deste enfoque, que tem como objetivo principal formalizar um modelo matemático que possa representar eficientemente os problemas práticos das áreas das ciências pouco formais através de suas particularidades. Tornam-se, então, modelos muito úteis com fim expreso de permitir o desenvolvimento de ferramentas computacionais que possam gerar uma regra de decisão ótima para um determinado classificador, auxiliando, portanto, a análise e identificação dos problemas. De forma ilustrativa, o processo de modelagem matemática para problemas relacionados, por exemplo, com a área da Medicina, consiste na interação de pelo menos três modelos, como mostra a Figura 5.1.

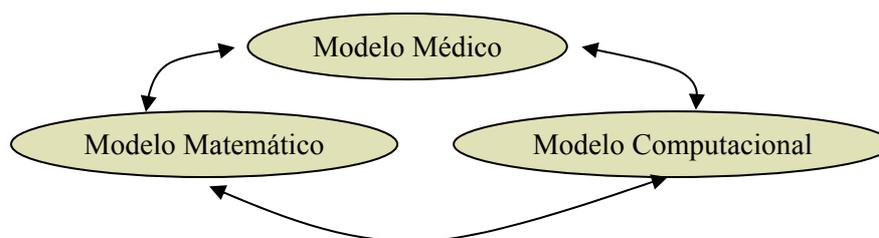


Figura 5.1 – Ilustração do processo de modelagem matemática de problemas relacionados com a área da Medicina.

No modelo médico, o objeto de estudo é, na maioria das vezes, abstrato, ou seja, descrito em termos de variáveis quantitativas e qualitativas simultaneamente, sendo que, algumas vezes, as descrições (informações) dos objetos podem se apresentar incompletas. Neste modelo a subjetividade é marcante, sendo expressa através das peculiaridades do paciente e de uma determinada enfermidade, o que leva os profissionais que atuam nesta área a freqüentemente empregar o conceito de *analogia* entre as descrições como base metodológica para as decisões. É importante também citar que a análise das informações é realizada por partes e com ponderações diferenciadas em determinadas características, além do tratamento das informações ser essencialmente qualitativo.

Em oposição ao que ocorre no modelo médico, o modelo matemático traz como resultado um rigor formal para a solução dos problemas. Logo, deve-se levar em conta a necessidade de um esforço interativo destes dois modelos para a formulação e formalização do problema. Segundo SHULCLOPER *et al.* (1997), o processo interativo destes dois modelos pode ser ilustrado através do esquema mostrado pela Figura 5.2. Este esquema de interação entre os modelos médico e matemático compreende os processos de formulação e formalização do problema, que serão descritos no transcórre deste capítulo.

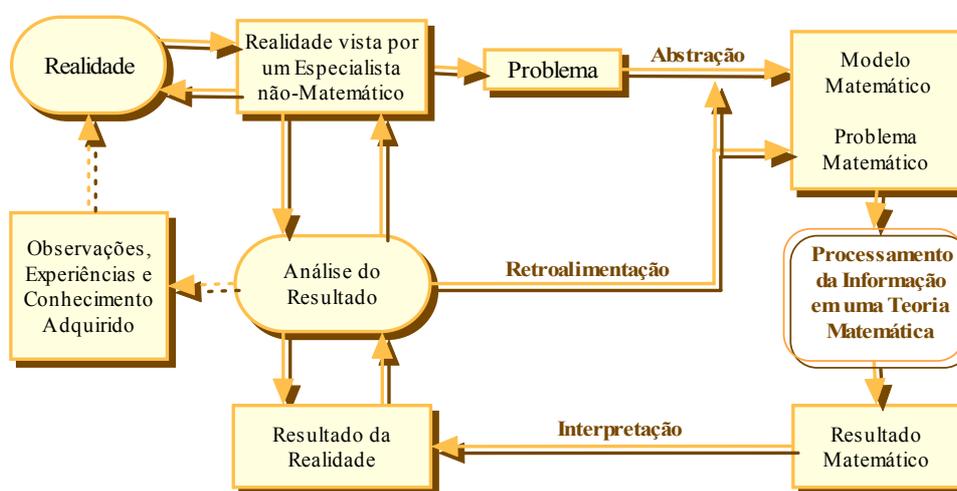


Figura 5.2 – Esquema do processo interativo dos modelos médico e matemático para a formulação e formalização do problema, segundo SHULCLOPER *et al.* (1997).

Já o modelo computacional rege a realização do sistema propriamente dito, sendo o resultado da interação dos modelos médico e matemático e a qualidade de resposta do

sistema dependente da qualidade de informação que se obtenha na formulação do problema. A incorporação do sistema na prática médica deve levar em conta as formas operativas da mesma, porém é importante ficar claro que qualquer que seja a qualidade de resposta do sistema, a habilidade do especialista em suas investigações não poderá ser substituída.

Logo, pode-se notar que o enfoque lógico-combinatório é mais que um conjunto de técnicas, é uma maneira de solucionar problemas de reconhecimento de padrões a partir de uma determinada metodologia de modelagem matemática destes problemas. A Lógica Matemática, a Teoria Clássica de Conjuntos e Subconjuntos Difusos, a Teoria Combinatória e a Matemática Discreta, em geral, constituem o embasamento teórico-matemático em que se denomina o “Enfoque Lógico-Combinatório”. Com o intuito de estruturar esta metodologia através de etapas, neste trabalho, adotou-se alguns dos tópicos da estrutura citada por SHULCLOPER *et al.* (1997), como mostra a Tabela 5.1. É necessário analisar, para tanto, que cada área do conhecimento tem suas peculiaridades e esta estrutura do enfoque lógico-combinatório pode sofrer algumas modificações dependendo do caso.

Tabela 5.1 – Estrutura geral da metodologia do enfoque lógico-combinatório adotada neste trabalho.

<b>1 - Formulação do Problema</b>
<b>2 - Formalização do Problema</b>
<b>3 - Modo de Solução do Problema Expresso em Termos Matemáticos</b>

### 5.3 – Formulação do Problema

A etapa de formulação do problema visa definir e estruturar os objetivos que devem ser alcançados sobre a manifestação do problema que se pretende resolver. Baseia-se no processo de entendimento e definição do objeto de estudo, que será a base para as etapas subsequentes. Neste processo, é fundamental a presença do especialista da área de aplicação, porque são eles que passam a essência do fenômeno e, expressando-se em sua linguagem, tomam as decisões para a caracterização do problema a resolver.

Neste sentido, no que diz respeito à proposta de desenvolvimento de uma ferramenta computacional que visa auxiliar o citogeneticista a minimizar o viés do diagnóstico em citogenética humana, para a formulação deste problema, houve a necessidade da interação com especialistas em genética médica (médicos e biólogos), que caracterizaram o problema fundamentando-se em padrões internacionalmente aceitos (ISCN, 1995). Esta primeira fase foi descrita através do capítulo II deste trabalho, servindo, portanto, de base para a etapa de formalização do problema.

#### **5.4 – Formalização do Problema**

Considera-se, nesta etapa, a exposição formal do problema, ou seja, a tradução da linguagem do especialista da área de aplicação para a linguagem formal da Matemática. Essa tradução é feita de tal maneira que, da etapa anterior, continuam sendo contemplados os mesmos objetivos. Nesta etapa, foram aplicados apenas alguns tópicos da metodologia do enfoque lógico-combinatório, sendo suficientes para a formalização do problema proposto e apresentados através da *seleção de variáveis* e suas propriedades, da *exposição formal de um problema de seleção de variáveis*, da aplicação dos *conceitos da teoria dos Testores* e, finalmente, a definição da *importância informacional das variáveis*.

##### **5.4.1 – Seleção de Variáveis**

A etapa de seleção de variáveis tem como objetivo selecionar o espaço de representação dos objetos de investigação, encontrando as características que os descrevem, de modo eficiente e que incidem no problema de maneira determinante. Essa etapa compreende a definição do *critério de analogia*, *critério de comparação de valores de uma variável* e *funções de semelhança entre objetos*.

###### **5.4.1.1 – Critério de Analogia**

*Analogia* pode ser conceitualizada como a semelhança que tem os objetos, dependendo das características que os descrevem. O conceito de analogia ou semelhança

entre objetos (pacientes) se encontra na base de quase todas as conclusões de especialistas que trabalham na área de Medicina, baseando-se no conhecimento adquirido, experiências e observações. O conhecimento adquirido é denominado de *precedentes*, que podem ser *parciais* ou *totais*, sendo os primeiros àqueles que levam em conta parte das descrições dos objetos relacionando-os com alguma classe, e ao segundo aqueles que levam em conta as descrições em sua totalidade.

Para muitos modeladores matemáticos, a semelhança entre os objetos pode ser representada por meio de uma função de distância entre as respectivas descrições dos objetos citados, e que a aproximação pode ser sinônimo de similaridade. Esta é a melhor representação matemática, pois a distância  $d$  é uma função real, simétrica e deve cumprir a desigualdade triangular (SHULCLOPER *et al.*, 1997).

$$\begin{aligned} d: M \times M &\rightarrow \mathfrak{R}^+ \\ \forall x \forall y \ d(x,y) &= 0 \Leftrightarrow x=y; \\ \forall x \forall y \ d(x,y) &= d(y,x); \\ \forall x \forall y \forall z \ d(x,y) &\leq d(x,z) + d(z,y); \end{aligned}$$

onde  $\mathfrak{R}^+$  é o conjunto dos números reais positivos. Este conceito de distância, para caracterizar a semelhança entre objetos, levanta algumas discussões como, por exemplo, quanto ao espaço necessário para a representação desses objetos. Tais objetos podem ser caracterizados como objetos físicos (variáveis quantitativas e qualitativas) ou abstratos (quadro clínico de um paciente, estado psicológico, etc), o que pode gerar dificuldades para representá-los através de uma distância comparativa. Outro fato é que nem todas as características são numéricas, e nem as comparações entre as representações dos objetos em termos das características devem ter uma distância.

Para exemplificar este tópico, pode-se constatar na caracterização do objeto de estudo, no capítulo III deste trabalho, que a identificação dos pares de cromossomos depende de três variáveis: do tamanho (grande, médio, relativamente pequeno e pequeno), que, para o especialista em genética médica, tais variáveis apresentam-se como qualitativas; através da posição do centrômero (metacêntrico, submetacêntrico e acrocêntrico), que se caracterizam por variáveis quantitativas; e pelo padrão de bandas, sendo identificados pelo número (quantitativa) e posição (qualitativa) das bandas ao longo

dos cromossomos. Já o diagnóstico clínico de um paciente com Leucemia Pró-Mielocítica Aguda, por exemplo, a avaliação pelo citogeneticista pode se tornar extremamente abstrata quando esta análise é realizada através do processo manual. Este fato decorre devido a inúmeros fatores inerentes ao processo de análise e classificação citogenética, que podem dificultar ou impossibilitar o profissional a fazer a correta avaliação desta divergência.

Logo, as características ou variáveis, em termos das quais são utilizadas para comparar os objetos em um processo de modelagem matemática, determinam-se em função dos objetivos que se busca nos estudos em questão. Assim, formalizar a analogia entre as características dos objetos é importante e não é sempre fácil, necessitando da ajuda de especialistas dos objetos em questão.

#### 5.4.1.2 – Critério de Comparação de Valores de uma Variável

Na etapa de comparação de valores de uma variável, determina-se qual o tipo de valor que uma variável assume, por exemplo, considere  $M$  o universo dos objetos admissíveis e o subconjunto  $K_1, \dots, K_r$ , onde  $K_t \subset M$  e  $t=1, \dots, r$ , sendo  $M$  o conjunto da união dos subconjuntos  $K_1, \dots, K_r$ . Seja  $R = \{x_1, \dots, x_n\}$  o conjunto de características em termos do objeto estudado. Cada característica  $x_i$  tem associado um conjunto  $M_i$  que se denomina de conjunto de valores admissíveis da característica  $x_i$ ,  $i=1, \dots, n$ . Esse conjunto pode ser denominado também de *domínio* de  $x_i$ . Logo, o domínio  $M_i$  pode ser descrito como:

- $\{0,1\}$  trata-se de variáveis *binárias* ou *booleanas*.
- $\{0, 1, \dots, k-1\}$ ,  $k > 2$ ; a característica considera diferentes gradações de valores tratando-se de uma variável com *k-valores*.
- $[a, b]$ ,  $[a, b)$ ,  $(a, b]$ ,  $(a, b)$ ;  $a$  e  $b$  considerados como números reais quaisquer.

Neste sentido, também, pode-se ter por definição a *descrição* de um objeto como uma *n-ésima*  $I(O) = (x_1(O), \dots, x_n(O))$ , onde  $x_i(O) \in M_i$  e  $i=1, \dots, n$ , sendo o valor da característica  $x_i$  no objeto  $O$ . E, quando consideram-se para a descrição de um objeto somente as características que aparecem em um subconjunto  $\Omega = \{x_{i_1}, \dots, x_{i_s}\} \subseteq R$ , diz-se que  $\Omega I(O) = (x_{i_1}(O), \dots, x_{i_s}(O))$  é uma *subdescrição* de  $O$  em termos das variáveis de  $\Omega$ .

Tomando-se como exemplo as características de um determinado cromossomo do par número 1, segundo os padrões de identificação clássica (tamanho, posição do centrômero e existência de constrições secundárias), a descrição deste objeto é feita em termos de três características, cujos valores são:

$$I(O)=(\text{grande, metacêntrico, sim})$$

$$x_1(O)=\text{grande}; x_2(O)=\text{metacêntrico}; x_3(O)=\text{sim}$$

onde  $x_1(O)$  é a variável que descreve o tamanho,  $x_2(O)$  a posição do centrômero e a variável  $x_3(O)$  a existência de constrições secundárias. Sendo  $\Omega=\{x_1, x_2\}$  variáveis qualitativas, então:  $\Omega I(O)=(x_1(O), x_2(O))=(\text{grande, metacêntrico})$ . Assim também, considera-se que os objetos se descrevem em um *espaço de representação inicial* (ERI), determinados pelo produto cartesiano dos conjuntos  $M_i$ :

$$I(O)=(x_1(O), \dots, x_n(O)) \in (M_1 \times M_2 \times \dots \times M_n).$$

Baseando-se nesses conceitos, sobre o domínio  $M_i$  definiu-se uma função denominada de *critério de comparação de valores de  $x_i$* , apresentada pela Equação 5.1:

$$C_i: M_i \times M_i \rightarrow L_i \quad (5.1)$$

onde  $L_i$  é a imagem definida como um conjunto totalmente ordenado. Se  $L_i=\{0,1,\dots,k-1\}$ ,  $C_i$  é um critério de comparação assumindo *k-valores*, e se  $L_i \subseteq \mathcal{R}$ ,  $C_i$  será um critério de comparação *real*. Existem ainda outros critérios de comparação de valores de uma variável, tais como os difusos, lingüísticos, entre outros. Assim também, se:

$$C_i(x_i(O), x_i(O)) = \min_{y \in L_i} \{y\} \quad - C_i \text{ é um critério de comparação de } \textit{mínima semelhança}$$

entre os valores de  $x_i$ , para  $i=1, \dots, n$ .

$$C_i(x_i(O), x_i(O)) = \max_{y \in L_i} \{y\} \quad - C_i \text{ é um critério de comparação de } \textit{máxima semelhança}$$

entre os valores de  $x_i$ , para  $i=1, \dots, n$ .

Portanto,  $C_i$  nos oferece uma avaliação do grau de similaridade (ou de diferença) entre dois valores de uma mesma variável, para  $i=1, \dots, n$ . Através do cálculo do critério de comparação, é possível um tratamento diferenciado, não uniforme das variáveis distintas com as quais os objetos são descritos (SHULCLOPER *et al.*, 1997).

Para ilustrar a definição dos conceitos de variáveis, domínios e critérios de comparação, tomou-se, como exemplo, um trabalho prévio realizado por ORTIZ-POSADAS *et al.* (2003), no qual seus autores desenvolveram um modelo matemático para a identificação clássica dos cromossomos. Nesse trabalho, foram definidas sete variáveis, seus domínios e seus respectivos critérios de comparação, apresentados na Tabela 5.2.

Tabela 5.2 – Apresentação dos conceitos de variáveis, domínios e critérios de comparação para a identificação clássica dos pares de cromossomos (ORTIZ-POSADAS *et al.*, 2003).

Variável ( $x_i$ )	Domínio ( $M_i$ )	Critério de comparação ( $C_i$ )			
$x_1$ - Tamanho	Grande – <b>G</b> Médio – <b>M</b> Relativamente Pequeno – <b>RP</b> Pequeno - <b>P</b>	<b>G</b>	<b>M</b>	<b>RP</b>	<b>P</b>
		0	0,33	0,66	1
		<b>M</b>	0	0,33	0,66
		<b>RP</b>		0	0,33
		<b>P</b>			0
$x_2$ - Posição dos Centrômeros	Metacêntrico – <b>M</b> SubMetacêntrico – <b>SM</b> Acrocêntrico - <b>A</b>	<b>M</b>	<b>M</b>	<b>SM</b>	<b>A</b>
			0	0,5	1
		<b>SM</b>		0	0,5
		<b>A</b>			0
$x_3$ - Sem Condições Secundárias $x_4$ - Com Condição Secundária Simples Distal $x_5$ - Com Condição Secundária Simples Mediana $x_6$ - Com Condição Secundária Simples Próxima $x_7$ - Com Condição Secundária Dupla	Sim, Não	<b>Sim</b>	<b>Sim</b>	<b>Não</b>	
			0	1	
		<b>Não</b>		0	

#### 5.4.1.3 – Funções de Semelhança entre Objetos

A analogia entre os objetos foi formalizada por meio do conceito de *função de semelhança* (ORTIZ-POSADAS *et al.*, 1998). Esta função permite a comparação entre as descrições de dois objetos, podendo ser construída a partir dos critérios de comparação antes mencionados neste capítulo. Então, seja  $M_{i_1} \times \dots \times M_{i_s}$ , para qualquer  $\Omega = \{x_{i_1}, \dots, x_{i_s}\} \subseteq R$ , define-se uma função  $\beta: (M_{i_1} \times \dots \times M_{i_s})^2 \rightarrow L$ , onde  $L$  é um conjunto

totalmente ordenado e de qualquer natureza, sendo  $\beta$  uma *função de semelhança parcial* (ou de comparações parciais). O termo “parcial” está fundamentado no conceito de *precedências parciais*, pois, quando  $s = n$ , então,  $\Omega=R$  e  $\beta$  será, portanto, uma *função de semelhança total* (relativo às precedências totais, que em determinadas ocasiões depende das funções de semelhança parciais).

Observe que  $\beta$  é uma avaliação do grau em que se assemelham as descrições dos objetos, sendo que a função de semelhança total é definida pela Equação 5.2:

$$\beta(I(O_i), I(O_j)) = \beta((C_1(x_1(O_i), x_1(O_j)), \dots, C_n(x_n(O_i), x_n(O_j)))) \rightarrow L \quad (5.2)$$

onde  $I(O_i)$  é a descrição do objeto  $O_i$ ,  $C_i$ , para  $i=1, \dots, n$ , representa o critério de comparação gerado por cada variável  $x_i$ ,  $L$  representa a imagem, podendo ser de qualquer natureza (booleana, *k-valores*, real, difusa, etc.) dependendo do tipo da variável, e  $\beta$  a função de semelhança total.

#### 5.4.2 – Exposição Formal de um Problema de Seleção de Variáveis

Do ponto de vista formal do problema que se quer resolver, pode-se resumir esta etapa na seguinte forma: supondo os subconjuntos  $K_1, \dots, K_r$  de  $M$ , sabendo-se que em  $M$  todos os objetos se distribuem em  $r$  classes. É dado um conjunto  $M_i = \{I(O_1), \dots, I(O_m)\} \subseteq M$ , do qual se conhece sua distribuição em classes, e que os objetos de  $M_i$  estão distribuídos entre  $K'_1, \dots, K'_r$ , onde os  $K'_i \subset K_i$ . Em outras palavras, tem-se uma tabela de objetos, denominada de *matriz de aprendizagem (MA)*, constituída por todas as classes  $K_i$ ,  $i=1, \dots, r$ . Logo, por  $I_0(K'_1, \dots, K'_r)$ , denota-se a informação relativa à descrição dos objetos de  $M_i$  e a sua distribuição em  $r$  classes antes mencionadas neste capítulos, expressa na forma de matriz, em que, em cada fila, aparece a descrição de um dos  $m$  objetos. A Tabela 5.3 ilustra um exemplo dessa matriz (*MA*), onde  $R = \{x_1, x_2, x_3, x_4, x_5\}$ ,  $m=5$  (número de objetos),  $r=2$  (classes  $K_1$  e  $K_2$ ) e o domínio  $M = \{0, 1\}$ .

Tabela 5.3 – Exemplo de uma matriz de aprendizagem (MA), onde  $R = \{x_1, x_2, x_3, x_4, x_5\}$ ,  $m=5$ ,  $r=2$  e  $M = \{0,1\}$ .

MA	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>4</sub>	x <sub>5</sub>
O <sub>1</sub>	1	1	0	1	1
O <sub>2</sub>	0	1	0	1	1
O <sub>3</sub>	1	0	1	0	1
O <sub>4</sub>	1	1	1	1	1
O <sub>5</sub>	0	0	0	1	1

K <sub>1</sub>	X <sub>1</sub>	x <sub>2</sub>	X <sub>3</sub>	x <sub>4</sub>	x <sub>5</sub>
O <sub>1</sub>	1	1	0	1	1
O <sub>2</sub>	0	1	0	1	1
O <sub>3</sub>	1	0	1	0	1

K <sub>2</sub>	X <sub>1</sub>	x <sub>2</sub>	X <sub>3</sub>	x <sub>4</sub>	x <sub>5</sub>
O <sub>4</sub>	1	1	1	1	1
O <sub>5</sub>	0	0	0	1	1

A título de ilustração prática deste tópico, no trabalho de ORTIZ-POSADAS *et al.* (2003), é elaborada uma matriz de aprendizagem com a descrição de 22 pares de cromossomos, além dos cromossomos sexuais X e Y, em termos da identificação clássica dos cromossomos, sendo apresentada pela Tabela 5.4.

Tabela 5.4 - Matriz de aprendizagem do trabalho de ORTIZ-POSADAS *et al.* (2003), apresentando a descrição de 22 pares de cromossomos, além dos cromossomos sexuais X e Y, em termos da identificação clássica dos mesmos.

Par de Cromossomos - O <sub>n</sub> -	Variáveis						
	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>4</sub>	x <sub>5</sub>	x <sub>6</sub>	x <sub>7</sub>
O <sub>1</sub>	Grande	Metacêntrico	Não	Não	Não	Sim	Não
O <sub>2</sub>	Grande	Submetacêntrico	Não	Sim	Não	Não	Não
O <sub>3</sub>	Grande	Metacêntrico	Não	Não	Sim	Não	Não
O <sub>4</sub>	Grande	Submetacêntrico	Não	Sim	Não	Não	Não
O <sub>5</sub>	Grande	Submetacêntrico	Não	Sim	Não	Não	Não
O <sub>6</sub>	Médio	Submetacêntrico	Não	Não	Não	Não	Sim
O <sub>7</sub>	Médio	Submetacêntrico	Sim	Não	Não	Não	Não
O <sub>8</sub>	Médio	Submetacêntrico	Sim	Não	Não	Não	Não
O <sub>9</sub>	Médio	Submetacêntrico	Não	Não	Não	Sim	Não
O <sub>10</sub>	Médio	Submetacêntrico	Sim	Não	Não	Não	Não
O <sub>11</sub>	Médio	Submetacêntrico	Não	Não	Não	Sim	Não
O <sub>12</sub>	Médio	Submetacêntrico	Sim	Não	Não	Não	Não
O <sub>X</sub>	Médio	Submetacêntrico	Sim	Não	Não	Não	Não
O <sub>13</sub>	Pequeno	Acrocêntrico	Não	Não	Sim	Não	Não
O <sub>14</sub>	Pequeno	Acrocêntrico	Sim	Não	Não	Não	Não
O <sub>15</sub>	Pequeno	Acrocêntrico	Sim	Não	Não	Não	Não
O <sub>16</sub>	Pequeno	Submetacêntrico	Não	Não	Não	Sim	Não
O <sub>17</sub>	Pequeno	Submetacêntrico	Não	Não	Sim	Não	Não
O <sub>18</sub>	Pequeno	Submetacêntrico	Sim	Não	Não	Não	Não
O <sub>19</sub>	Muito Pequeno	Metacêntrico	Não	Não	Sim	Não	Não
O <sub>20</sub>	Muito Pequeno	Metacêntrico	Sim	Não	Não	Não	Não
O <sub>21</sub>	Muito Pequeno	Acrocêntrico	Não	Não	Sim	Não	Não
O <sub>22</sub>	Muito Pequeno	Acrocêntrico	Sim	Não	Não	Não	Não
OU	Muito Pequeno	Acrocêntrico	Não	Não	Sim	Não	Não

### 5.4.3 – Conceitos da Teoria de Testores

A teoria de Testores foi formulada em 1954-1955 na área científica da Cibernética Matemática – na URSS. Sua origem estava vinculada à utilização de métodos lógicos matemáticos para a localização de defeitos em circuitos elétricos, pelos especialistas soviéticos *Yablonskii* e *Cheguis*, abrindo uma importante linha de investigações denominada de "Teoria de Testores para Esquemas Lógicos". Entretanto, o trabalho desses especialistas ainda não havia sido aplicado em problemas de reconhecimento de padrões, até que em 1966, *Zhuravliov* publicou um trabalho fora dos limites da síntese dos esquemas lógicos e em um terreno pouco formalizado, a Geologia, constituindo o primeiro trabalho da intersecção da teoria de Testores em reconhecimento de padrões. A partir de então, foi o início de um novo enfoque, unindo também a Lógica Matemática e a Matemática Discreta (SHULCLOPER *et al.*, 1997).

A partir da publicação do trabalho de *Zhuravliov*, a teoria de Testores trilhou o caminho para aplicações de algoritmos de reconhecimento em ciências pouco formais, em que se necessitava encontrar uma determinada relevância das características dos objetos. Essa teoria pode ser definida baseando-se no seguinte conceito: seja a tabela  $T$  formada pelas descrições dos objetos em termos de suas características  $R = \{x_1, \dots, x_n\}$ . Assume-se que a tabela seja dividida em duas subtabelas  $T_0$  e  $T_1$ , sendo que os primeiros objetos cumprem uma determinada propriedade que os caracteriza em  $T_0$  e os objetos restantes satisfazem outra propriedade, que os caracteriza em  $T_1$ . Supõe-se que ambas subtabelas são disjuntas e não coincidem com as descrições dos objetos que estão em diferentes subtabelas.

Através da definição de *Zhuravliov*, o conjunto  $\tau = \{i_1, \dots, i_s\}$  de colunas da tabela  $T$  denomina-se *testor* para  $(T_0, T_1) = T$ , se depois de eliminar de  $T$  todas as colunas exceto as de  $\tau$ , cumpre-se não existir nenhuma fila (linha) em  $T_0$  igual a nenhuma fila de  $T_1$ , ou seja, não existem filas iguais em classes diferentes. Dessa forma, um testor é um conjunto de características que *permitem se diferenciar* entre classes, porque nenhum objeto da classe  $T_0$  (em termos dos valores de suas características) se confunde, em termos dos mesmos valores, com nenhum objeto de outra classe  $T_1$ .

Neste contexto também, um testor se chama *irredutível (típico)* se, ao eliminar uma coluna qualquer, deixa de ser testor para  $(T_0, T_1)$ . O termo "irredutível" reflete a idéia de

que não se podem eliminar mais colunas (pois as demais colunas perderiam tal propriedade), e o termo "típico" constitui a idéia de tipicidade de uma classe de objetos que possuem um conjunto de características que, de um certo sentido, tipificam uma classe de objetos das demais. A Tabela 5.5 apresenta uma matriz  $T$  formada pela descrição de 5 objetos distribuídos em duas classes ( $T_0$ ,  $T_1$ ), sendo que os 3 primeiros objetos (sombreados) pertencem a primeira classe ( $T_0$ ) e os restantes à segunda ( $T_1$ ). O conjunto das variáveis  $\{x_2, x_3\}$  é um testor típico para  $T_0'$  e  $T_1'$ , pois, após a eliminação das variáveis  $x_1, x_4$  e  $x_5$  de  $T$ , pode-se constatar que não existem mais filas em  $T_0'$  iguais às filas de  $T_1'$ .

Tabela 5.5 – Matriz  $T$  formada pela descrição de 5 objetos distribuídos em duas classes ( $T_0$ ,  $T_1$ ), constatando-se a tipicidade em  $\{x_2, x_3\}$ .

$T$	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
$O_1$	1	1	0	1	1
$O_2$	0	1	0	1	1
$O_3$	1	0	1	0	1
$O_4$	1	1	1	1	1
$O_5$	0	0	0	1	1

$T_0'$	$x_2$	$x_3$
$O_1$	1	0
$O_2$	1	0
$O_3$	0	1

$T_1'$	$x_2$	$x_3$
$O_4$	1	1
$O_5$	0	0

Como pode ser visto, através da obtenção dos testores típicos, é possível reduzir o *espaço de representação inicial* (ERI) das características visando obter subespaços de representação discriminantes que cumprem com certas propriedades necessárias. A determinação dos testores típicos é realizada através de uma busca exaustiva entre todos os subconjuntos de características que são  $2^{|R|}$ , sendo  $R$  o conjunto das características, devendo haver um custo de tempo um pouco elevado para alguns casos.

Uma estratégia para a determinação de todos os testores típicos é feita através do cálculo binário, que pode apresentar uma redução no tempo do processo de cálculo de todos os testores típicos de uma matriz de dados  $T_{nmr}$ . Assim, chama-se *matriz de diferença* (ou de *comparação*) de  $T_{nmr}$  a matriz *booleana*  $MD_{nm}$ , que possui somente zeros e uns. Para exemplificar este tópico, toma-se o problema de identificação clássica dos

cromossomos, apresentado pela Tabela 5.4, onde são descritos 24 tipos de objetos distribuídos em 15 classes. Calcula-se a matriz diferença da seguinte forma:

1- Considera-se a descrição do primeiro objeto que aparece na primeira classe e, através do critério de comparação de valores da variável  $x_1$  compara-se o valor  $x_1(O_1)$  com o valor da variável do primeiro objeto da classe seguinte -  $C_1(x_1(O_1), x_1(O_4))$ , tal que, atribui-se 0 para  $x_1(O_1)=x_1(O_4)$  e 1 para  $x_1(O_1) \neq x_1(O_4)$ . A comparação se repete com os valores das demais características desses objetos comparados e os resultados são colocados na posição  $(1,i)$ , com  $i=1, \dots, n$  (onde  $n$  é o número de características).

2- O processo de critério de comparação é feito com os valores das características de todos os objetos da primeira classe sendo comparados com todos os valores das características dos objetos das classes restantes. Alguns cálculos da matriz diferença  $MD_{nm}$  do exemplo da Tabela 5.4, são apresentados através da Tabela 5.6.

Tabela 5.6 – Exemplo de alguns cálculos da matriz diferença  $MD_{nm}$  do exemplo da Tabela 5.4.

$MD_{nm}$	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$
$S_{12}$	0	1	1	1	0	1	0
$S_{13}$	0	0	0	0	1	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$S_{23}$	0	1	0	1	1	0	0
$S_{24}$	0	0	0	0	0	0	0
$S_{25}$	0	0	0	0	0	0	0
$S_{26}$	1	0	0	1	0	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

A caracterização do conceito de testor pode ser vista através da análise da matriz diferença  $MD_{nm}$ . Busca-se obter, então, a matriz  $MD/\tau$  em que o conjunto  $\tau = \{x_{j_1}, \dots, x_{j_s}\}$  pode apresentar-se com  $s=n$ , sendo  $s$  o número de características do testor e  $n$  o número de características de  $T_{nmr}$ , ou  $s < n$ , tornando-se um subconjunto de colunas de  $T_{nmr}$ . Para este segundo caso, a matriz  $MD/\tau$  é obtida de  $MD$  ao eliminar todas as colunas, exceto as de  $\tau$ . Logo,  $\tau$  é um testor se e somente se  $MD/\tau$  não possui nenhuma fila de zeros. A demonstração desta propriedade pode ser constatada tomando-se o testor  $\tau = \{x_1, \dots, x_7\}$  da matriz  $MD$  da Tabela 5.6, onde foi verificado que da comparação do objeto  $O_2$  com os

objetos  $O_4$  e  $O_5$  resultaram em filas de zeros. Este fato demonstra que tais objetos (pares de cromossomos 2, 4 e 5) comparados, possuem as mesmas características (grandes, submetacêntricos e com constrição secundária simples distal) pertencendo a classes diferentes. E isso também ocorre para os pares de cromossomos 7, 8, 10, 12 e X, para os pares 9 e 11, além do 14 e 15, e também para o 21 e Y. Portanto, o conjunto  $\tau$  não pode ser caracterizado como um testor típico, pois as sete características utilizadas no trabalho de ORTIZ-POSADAS *et al.* (2003) não permitem distinguir as 24 classes de cromossomos, como mostrado a Tabela 5.7.

Tabela 5.7 – Caracterização do conceito de testores através da análise do trabalho de ORTIZ-POSADAS *et al.* (2003) que, adotando o critério de identificação clássica de cromossomos, não se pode caracterizar o conjunto  $\tau = \{x_1, \dots, x_7\}$  como um testor típico, pois existem objetos com as mesmas características em classes diferentes.

Par de Cromossomos - $O_n$ -	Variáveis						
	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$
$O_1$	Grande	Metacêntrico	Não	Não	Não	Sim	Não
$O_3$	Grande	Metacêntrico	Não	Não	Sim	Não	Não
$O_2, O_4, O_5$	Grande	Submetacêntrico	Não	Sim	Não	Não	Não
$O_6$	Médio	Submetacêntrico	Não	Não	Não	Não	Sim
$O_7, O_8, O_{10}, O_{12}, O_X$	Médio	Submetacêntrico	Sim	Não	Não	Não	Não
$O_9, O_{11}$	Médio	Submetacêntrico	Não	Não	Não	Sim	Não
$O_{13}$	Pequeno	Acrocêntrico	Não	Não	Sim	Não	Não
$O_{14}, O_{15}$	Pequeno	Acrocêntrico	Sim	Não	Não	Não	Não
$O_{16}$	Pequeno	Submetacêntrico	Não	Não	Não	Sim	Não
$O_{17}$	Pequeno	Submetacêntrico	Não	Não	Sim	Não	Não
$O_{18}$	Pequeno	Submetacêntrico	Sim	Não	Não	Não	Não
$O_{19}$	Muito Pequeno	Metacêntrico	Não	Não	Sim	Não	Não
$O_{20}$	Muito Pequeno	Metacêntrico	Sim	Não	Não	Não	Não
$O_{21}, O_Y$	Muito Pequeno	Acrocêntrico	Não	Não	Sim	Não	Não
$O_{22}$	Muito Pequeno	Acrocêntrico	Sim	Não	Não	Não	Não

Através do conjunto de variáveis utilizado por ORTIZ-POSADAS *et al.* (2003), mostrado na Tabela 5.2, foi possível distinguir 15 classes de cromossomos. Entretanto, era necessário identificar novas características que permitissem a distinção de todos os cromossomos em 24 classes diferentes, o que motivou o desenvolvimento de uma nova sistemática para a análise e classificação citogenética proposta neste trabalho.

Como pode ser visto, a teoria de Testores torna-se, então, muito útil para a comunicação com especialistas não matemáticos, servindo como ferramenta para a

determinação da relevância das características dos objetos no marco da etapa de classificação.

#### 5.4.4 – Importância Informacional das Variáveis

Existem vários enfoques para definir uma medida de relevância das características dos objetos. A idéia básica consiste no seguinte: ao passar um testor para um testor típico chega-se a uma combinação irreduzível, em que cada característica resulta em manter as diferenças entre as classes. É natural supor que, se uma característica aparece em muitas combinações irreduzíveis, deduz-se que tal característica descreve de maneira diferencial as classes. Tomando-se como base esta idéia, *Zhuravliov* formulou a definição de *peso informacional*  $\rho(x)$  de uma característica, também denominado de *parâmetro de ponderação*, como a frequência relativa de aparição desta característica na família de todos os testores típicos. Além disso, também é natural considerar como relevância através da medida da quantidade de características que são necessárias em um testor típico para diferenciar os objetos que estão em classes diferentes, isto é, quanto uma característica é mais diferencial que as outras. Logo, também se associa a cada característica uma magnitude de acordo com a menor longitude dos testores típicos, sendo o testor  $\tau = \{i_1, \dots, i_s\}$  e suas respectivas características  $x_{i_1}, \dots, x_{i_s}$ , o número  $s$  denomina-se longitude do testor.

Neste sentido, existem cálculos para a definição do peso informacional utilizando-se equações para o cálculo da relevância das variáveis, segundo a frequência relativa de aparição das características na família de todos os testores típicos, e também, segundo a longitude dos testores, porém o peso informacional  $\rho(x)$ , associado a cada característica  $x$ , poderá ser definido pelo especialista, e a função de semelhança entre objetos apresentada pela Equação 5.2, pode ser completamente definida pela equação 5.3:

$$\beta(I(O_i), I(O_j)) = 1 - \frac{\sum_{t=1}^k \rho_t C_t(x_t(O_i), x_t(O_j))}{n} \quad (5.3)$$

onde  $I(O)$  corresponde à descrição do objeto,  $k$  é o número de variáveis,  $\rho_t$  é o parâmetro de ponderação definido pelo especialista,  $C_t$  o critério de comparação dos objetos, e  $n$  é o

fator de estandarização calculado pela somatória de todas as ponderações máximas. Esta estandarização é necessária para garantir que o valor da semelhança se encontre no intervalo  $[0, 1]$ , e que, quando o fator é igual a um, significa uma semelhança total, ou seja, que os dois objetos em comparação são iguais. Entretanto, se a semelhança for igual a zero, os objetos são totalmente diferentes.

Tomando-se como exemplo a definição do parâmetro de ponderação das variáveis, considerando-se o problema das variáveis segundo a Tabela 5.4, foram definidas pelo especialista as seguintes ponderações (relevâncias) de acordo com as sete características de descrição dos cromossomos, como mostra a Tabela 5.8.

Tabela 5.8 - Ponderações definidas pelo especialista para as sete características do problema de descrição dos cromossomos, apresentado segundo a Tabela 5.4.

Variável ( $x_i$ )	Ponderação ( $\rho_i$ )
$x_1$ - Tamanho	1
$x_2$ - Posição dos Centrômeros	1
$x_3$ - Sem Condições Secundárias	0,2
$x_4$ - Com Condição Secundária Simples Distal	0,2
$x_5$ - Com Condição Secundária Simples Mediana	0,2
$x_6$ - Com Condição Secundária Simples Próxima	0,2
$x_7$ - Com Condição Secundária Dupla	0,2

Definidos os pesos informacionais para cada característica, pode-se aplicar, então o cálculo completo da semelhança entre diferentes cromossomos. Para exemplificar, escolheram-se as descrições dos pares de cromossomos 1, 13 e 22, assim como no trabalho de ORTIZ-POSADAS *et al.* (2003). Substituindo-se os valores na Equação 5.3, o fator de semelhança entre  $O_1$  e  $O_{13}$  e para  $O_1$  e  $O_{22}$  resultou :

$$\beta_{cromossomos}(I(O_1), I(O_{13})) = 1 - \left[ \frac{1(0.66) + 1(1) + 0.2(0) + 0.2(0) + 0.2(1) + 0.2(1) + 0.2(0)}{3} \right] \cong 0.31$$

$$\beta_{cromossomos}(I(O_1), I(O_{22})) \cong 0.75$$

Levando-se em conta as semelhanças obtidas, pode-se concluir que o cromossomo  $O_1$  é mais semelhante ao cromossomo  $O_{22}$  do que ao cromossomo  $O_{13}$ .

### 5.5 – Modo de Solução do Problema Expresso em Termos Matemáticos

O processo de formalização, muitas vezes, restringe fortemente a busca por métodos e técnicas de solução. Decide-se, nesta etapa, qual o enfoque do reconhecimento de padrões a ser utilizado, ou a combinação deles para se determinar qual o algoritmo ou a família de algoritmos de classificação que, de melhor maneira, atende às condições do problema matemático formalizado a se resolver.

No enfoque lógico-combinatório, esta etapa tem um papel decisivo no desempenho da análise da matriz de aprendizagem em que foram selecionadas as características para a classificação e a representação dos objetos. Logo, o problema da classificação em termos desse enfoque consiste em ter um algoritmo  $\zeta$  tal que:

1-  $\zeta(I_0(K'_1, \dots, K'_r)) = \left\| x_i(O_j) \right\|_{mp}$ , que representa uma matriz, onde, em cada fila, tem-se  $\Omega I(O_j) = (x_{i_1}(O_j), \dots, x_{i_p}(O_j))$ , sendo  $p < n$  e  $j = 1, \dots, m$ , que são as subdescrições dos  $O_j$  objetos com um número menor de variáveis, ou seja, o subconjunto  $\Omega = \{x_{i_1}, \dots, x_{i_p}\}$ .

2- Dado um certo critério de classificação  $K$  e um objeto qualquer  $O$  da matriz  $I_0(K'_1, \dots, K'_r)$ , cumpre-se que:

$$K((x_1(O), \dots, x_n(O))) = K((x_{i_1}(O), \dots, x_{i_p}(O))),$$

Como visto na teoria dos Testores, ao reduzir a dimensão  $n$  (número de variáveis) não mudará a pertinência da classe. Sendo assim, por "critério de classificação", está entendido o conhecimento, a experiência, as razões pelas quais os objetos que confirmam a matriz  $I_0(K'_1, \dots, K'_r)$ , e que pertencem a uma certa classe.

Tomando-se como base os dados formalizados e o tipo de algoritmo a utilizar, elabora-se o sistema computadorizado e se obtém a solução do problema matemático. Analisa-se, então, a concordância do resultado matemático alcançado com os objetivos formalizados do problema matemático. Nesta etapa, tem-se como ferramenta fundamental a formalização dos critérios para a avaliação dos resultados da segunda etapa.

Para finalizar, é importante mencionar que a aplicação da metodologia de modelagem matemática para a solução de problemas de reconhecimento de padrões para as ciências pouco formais é somente um ponto de partida na busca de novas e melhores soluções para se tentar simular o pensamento humano.

## 5.6 – Considerações Finais

Tendo em vista o exposto, este capítulo teve como principal objetivo fazer uma breve descrição sobre a metodologia de modelagem matemática, baseando-se no enfoque lógico-combinatório da teoria de reconhecimento de padrões. Esta descrição foi estruturada nos seguintes tópicos: aspectos gerais, formulação do problema, formalização do problema, seleção de variáveis, critério de analogia, critério de comparação de valores de uma variável e funções de semelhança entre objetos, exposição formal de um problema de seleção de variáveis, conceitos da teoria de testes, importância informacional das variáveis e, por fim, o modo de solução do problema, expresso em termos matemáticos. Este estudo é, portanto, de suma importância servindo de fundamentação teórica para a proposta de um novo método de análise e classificação citogenética, uma vez que serão utilizados alguns dos tópicos da metodologia do enfoque Lógico Combinatório para proporcionar uma alternativa eficaz para classificação dos cromossomos.

## CAPÍTULO VI

### **METODOLOGIA PROPOSTA**

#### **6.1 – Introdução**

À luz dos preceitos aqui expostos, pode-se verificar que ainda existem muitas limitações dos sistemas automáticos de cariotipagem que, junto com outros aspectos inerentes desta atividade, potencializam o viés da análise em citogenética clínica. Um dos fatores limitantes desses sistemas de cariotipagem diz respeito à correta identificação dos cromossomos e o outro diz respeito à presença marcante da subjetividade na caracterização dos problemas e na análise citogenética. Assim, o desenvolvimento de técnicas de processamento digital de imagens, visando a extração mais precisa dos cromossomos e a aplicação em conjunto com técnicas de modelagem matemática para a formalização desses problemas, torna-se, então, o grande desafio deste trabalho para a redução do viés da análise em citogenética clínica.

Em síntese, o objetivo deste capítulo será de apresentar o desenvolvimento passo a passo de uma sistemática, através de uma ferramenta computacional, que visa propor uma nova metodologia de análise e classificação citogenética, baseando-se no enfoque lógico-combinatório da teoria de reconhecimento de padrões e em técnicas de processamento digital de imagens. Os tópicos que contemplam a metodologia proposta são: descrição da metodologia proposta, formulação geral do problema, formalização do problema e modo de solução do problema expresso em termos matemáticos. Ainda neste capítulo, serão apresentadas as considerações finais sobre o conteúdo apresentado.

#### **6.2 – Descrição da Metodologia Proposta**

A determinação da composição cromossômica das células de um indivíduo é um procedimento essencial para a análise citogenética. É um processo que exige extrema habilidade dos citogeneticistas e consome muito tempo para sua execução. Logo, com a

finalidade de disponibilizar ao citogeneticista melhores condições para a execução desta atividade, os sistemas comerciais de cariotipagem foram desenvolvidos visando automatizar o processo de identificação dos cromossomos para a composição do cariógrama, facilitando sobremaneira a execução de tarefas laboratoriais com alto grau de precisão e repetibilidade, além de proporcionarem muitas outras vantagens.

Diante disso, o processo computadorizado para a análise e classificação dos cromossomos está se tornando uma ferramenta muito útil aos laboratórios de genética humana, porém, nem todos os problemas para o citogeneticista foram resolvidos. Os sistemas de cariotipagem comerciais apresentam, ainda, muitas limitações que precisam ser superadas para que o profissional, no uso de uma ferramenta computacional, possa efetuar um diagnóstico citogenético com maior segurança e confiabilidade.

Surge, então, a proposta deste trabalho, tendo como meta principal aplicar e desenvolver novos métodos computacionais visando minimizar as limitações apresentadas pelos sistemas automáticos de cariotipagem atuais. Para tanto, o processo de modelagem matemática para o problema proposto está estruturado em três etapas: *formulação geral do problema*, *formalização do problema* e *modo de solução do problema expresso em termos matemáticos*, como mostrado no diagrama geral da Figura 6.1. No transcorrer deste capítulo serão apresentadas detalhadamente estas três etapas.

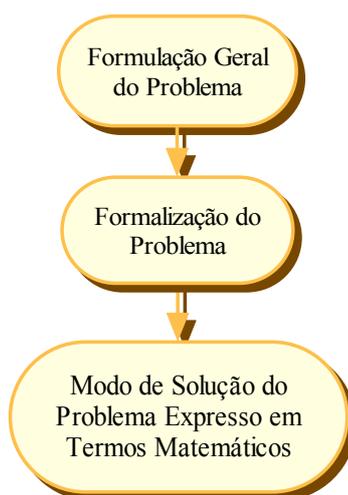


Figura 6.1 – Diagrama geral das etapas do processo de modelagem matemática para o problema proposto.

### 6.2.1 – Formulação Geral do Problema

Sabe-se que um aspecto fundamental na solução de problemas depende de como entender e definir o problema que se quer resolver. Neste contexto, ENSSLIN *et al.* (2001) procurou soluções mais adequadas para os problemas aplicando, em trabalhos científicos, um método que depende da resposta de quatro questões básicas: *qual o problema que se quer resolver, de quem é o problema, o que se pode fazer para solucionar ou melhorar tal problema e se o modo de solução do problema é factível* com a realidade em que se encontra.

Trilhando neste sentido, para responder a primeira questão, buscou-se relembrar dois pontos ainda não resolvidos pelos sistemas automáticos de cariotipagem já mencionados no decorrer deste trabalho. O primeiro ponto diz respeito a algumas limitações nesses sistemas decorrentes de algumas fontes de incerteza no domínio do problema. Tais fontes de incerteza dificultam a extração correta das características dos cromossomos comprometendo sua classificação e, conseqüentemente, impossibilitando a apresentação automática de kariogramas livres de erros. Logo, tais sistemas são rotulados de sistemas semi-automáticos de cariotipagem, disponibilizando sempre uma forma iterativa ao usuário para que este possa intervir no processo de classificação final dos cromossomos.

Também com a finalidade de se tentar superar as limitações decorrentes a tais fontes de incerteza, muitos sistemas automáticos passaram a utilizar algoritmos baseados nos sistemas especialistas, principalmente as RNAs. Conseguiu-se uma melhoria expressiva no desempenho dos classificadores, entretanto, não existe ainda garantia de que o sistema apresente um bom desempenho (TODESCO, 1995). Isso ocorre devido ao fato de que a eficiência desses sistemas é diretamente proporcional ao número de imagens de cromossomos que são apresentadas (armazenadas) à base de dados. Assim, para que o sistema melhore seu desempenho de classificação, é necessário que o profissional sempre inclua imagens de cromossomos previamente classificadas, tornando-se um processo inconveniente à rotina do laboratório. Além disso, também existe a desvantagem de que o resultado dessa classificação é apresentado na forma objetiva do certo ou errado (classificado e não-classificado), diferindo muito da forma como os citogeneticistas analisam e avaliam os problemas em citogenética clínica, ou seja, do conceito de analogia.

Levando em conta esses fatores, pode-se dizer que a aplicação dos sistemas especialistas minimizou apenas em parte as limitações dos sistemas de cariotipagem.

O segundo ponto limitante nos sistemas automáticos de cariotipagem vem da análise e avaliação clínica, que ainda fica, única e exclusivamente, dependente do citogeneticista. Como pode ser visto, a caracterização do objeto de estudo em citogenética humana é feita através da análise de variáveis quantitativas e qualitativas, simultaneamente, como no caso da detecção da leucemia pró-mielocítica aguda. Dessa forma, a presença da subjetividade é marcante nos diagnósticos clínicos finais, podendo acarretar, em certas ocasiões, equívocos que dificultam a detecção de uma determinada anormalidade. Tais equívocos são intrínsecos a toda sistemática da atividade de citogenética humana, constituídas por indecisões, incertezas, contradições ou interpretações diversas que, em geral, induzem ao viés da análise em citogenética clínica.

Levando-se em conta estes dois pontos ainda não resolvidos, define-se, então, o problema que se quer melhorar que, no sentido geral, visa auxiliar a realização da análise e classificação citogenética para minimizar o viés da análise em citogenética clínica. Para tanto, é necessário e de suma importância, que se pense no processo computacional não somente como uma ferramenta que permite tornar o processo mais ágil, mas também que busque auxiliar o citogeneticista a ter cariogramas livres de erros e, em seu ponto mais forte, que possibilite a detecção mais eficiente de anomalias, resultando em um diagnóstico clinicamente mais confiável. Surge implícito, neste contexto, o autor do problema que nesta etapa de formulação geral do problema, descreve-o em sua linguagem (modelo do citogeneticista).

Diante desses fatos, a alternativa encontrada para a solução desse problema foi a proposta de uma nova sistemática, denominada de *Sistema Computadorizado para Análise e Classificação Citogenética* (SCACC), elaborada para dois propósitos distintos: o primeiro diz respeito à elaboração de um método de reconhecimento automático de cromossomos através de uma única cena, utilizando-se técnicas avançadas de processamento digital de imagens para a extração e detecção dos cromossomos e, em conjunto, aplicando o enfoque lógico-combinatório da teoria de reconhecimento de padrões para a classificação automática dos mesmos. Para o segundo propósito, busca-se o desenvolvimento de um método computacional que possa ser utilizado como ferramenta para detecção mais confiável da anormalidade cromossômica estrutural da leucemia pró-

mielocítica aguda. Para melhor entendimento, a Figura 6.2 ilustra um diagrama geral da proposta de desenvolvimento da nova sistemática em citogenética humana.

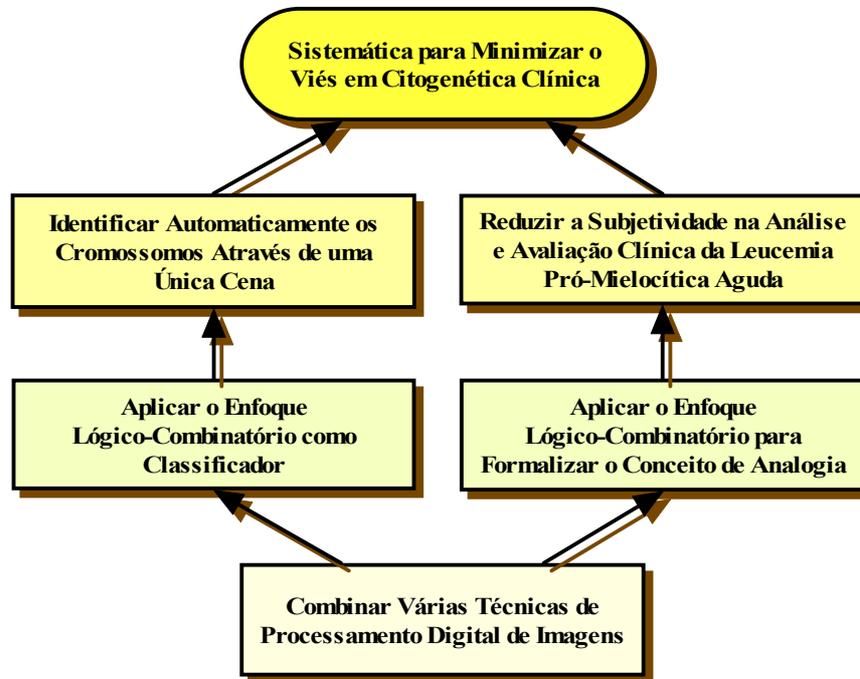


Figura 6.2 – Diagrama geral da proposta de solução para o problema, caracterizado pelo desenvolvimento de uma nova sistemática para análise e classificação citogenética, que tem como objetivo minimizar o viés em citogenética clínica.

A elaboração de um método de reconhecimento automático de cromossomos através de uma única cena, isto é, através da análise de uma única metáfase, justifica-se por buscar desenvolver um método de identificação que não precisa comparar entidades específicas com padrões previamente armazenados. Para tanto, busca-se a implementação de um conjunto de técnicas de processamento digital de imagens para a extração mais precisa das características dos cromossomos que sejam menos suscetíveis às fontes de incerteza. A aplicação do enfoque lógico-combinatório como classificador, baseia-se na função de semelhança entre objetos e tem como principal vantagem a identificação dos cromossomos por ordem de semelhança, ou seja, pelo grau de similaridade entre os objetos e à classe a que pertencem.

Quanto ao propósito de formalizar o conceito de analogia para reduzir a subjetividade na detecção da leucemia pró-mielocítica aguda, deve-se ao fato de ser uma anormalidade cromossômica de difícil detecção e que depende de muita habilidade técnica do profissional, muitas vezes necessitando da corroboração de outros técnicos para sua confirmação. Logo, desenvolver uma ferramenta computacional que possibilite caracterizar mais eficientemente o rearranjo presente na leucemia, permite ao citogeneticista prescrever um diagnóstico clínico de forma mais convicta e segura.

Finalmente, para tornar exequível o desenvolvimento da sistemática proposta, ou seja, tornar o problema factível para a sua realização, há a necessidade de se utilizar os mesmos dispositivos que constituem os sistemas automáticos de cariotipagem, porém diferenciando-os em algumas características específicas no capítulo posterior. Quanto ao desenvolvimento da sistemática proposta que compreende o *Sistema Computadorizado para a Análise e Classificação Citogenética* (SCACC), baseou-se na linguagem de programação C++Builder 6.0 Enterprise Suíte - da *Borland Software Corporation*, para ambiente Microsoft Windows 98. Após a etapa de formulação geral do problema (problema descrito no modelo do citogeneticista), o passo seguinte é traduzir o conhecimento do citogeneticista para um modelo computacional. Essa tradução é caracterizada pela formalização do problema, que nada mais é do que a elaboração de um modelo matemático que represente eficientemente o problema.

### 6.2.2 – Formalização do Problema

Para a elaboração de um modelo matemático, que contemple todos os objetivos propostos no modelo do citogeneticista, houve a necessidade de subdividi-lo em duas etapas: a *formalização do problema de identificação automática* e a *formalização do problema de detecção da Leucemia Pró-Mielocítica Aguda (LMA)*. Estas duas etapas serão detalhadas separadamente a seguir:

### 6.2.2.1 – Formalização do Problema de Identificação Citogenética

Na etapa de formulação do problema de identificação citogenética, busca-se fazer uma representação do conhecimento através da criação de um modelo matemático, que permite a implementação de uma rotina computacional que possa apresentar, automaticamente, cariogramas livres de erros a partir de uma única metáfase. O primeiro passo para essa formalização é a escolha das variáveis (características) dos cromossomos a serem utilizadas na classificação, sendo fundamental a utilização daquelas que sejam menos suscetíveis às fontes de incerteza. Pode ser visto, no trabalho de PIPER (1986) que, para a identificação dos cromossomos unicamente por classes de tamanho fixo, não eram suficientes para o reconhecimento correto dos cromossomos. Também pode ser demonstrado, em um trabalho prévio de ORTIZ-POSADAS *et al.* (2003), que, a partir das características escolhidas quanto à identificação clássica dos cromossomos, tamanho, posição dos centrômeros e constrições secundárias, não foi possível a identificação de todos os pares de cromossomos, ou seja, das 24 classes de cromossomos apenas 15 poderiam ser detectadas, logo, não sendo possível a identificação de todos os pares homólogos.

A partir da inclusão dos padrões de bandas no grupo de variáveis dos cromossomos, passou-se a conseguir melhores resultados de identificação, como pode ser constatado nos trabalhos de JOHNSTON *et al.* (1993), LERNER (1995) e POPESCU (2001). Entretanto, estes trabalhos utilizaram RNAs como classificadores, não sendo possível a classificação precisa dos cromossomos através de uma única metáfase havendo, então, a necessidade de incluir padrões pré-estabelecidos para o sistema.

Tomando-se como base os trabalhos pesquisados, constatou-se, neste estudo que as variáveis tamanho e padrões de bandas seriam imprescindíveis para a identificação dos cromossomos. Além dessas duas variáveis, optou-se por uma terceira, relacionada à forma, muito utilizada em trabalhos de avaliação morfológica de objetos (PACIORNIK, 2003). A variável tamanho pode ser extraída de várias maneiras, mas uma das técnicas que tem sido mais utilizada é através da determinação do eixo longitudinal dos cromossomos, porém esta técnica deixa de ser precisa quando se trata de cromossomos severamente tortos (TODESCO, 1995). Como a ocorrência de torções em cromossomos não é difícil de acontecer na rotina dos laboratórios, optou-se, neste trabalho, por

determinar o tamanho dos cromossomos pela extração da área. De posse desta variável e extraindo o perímetro dos cromossomos, calculou-se também o fator de forma utilizando-se a Equação 4.6.

Já a variável que diz respeito às bandas cromossômicas, caracterizadas por regiões escuras, tornou-se prática comum o cálculo do perfil de densidade média ao longo do eixo longitudinal extraído através de um limiar de densidade óptica (JOHNSTON *et al.*, 1993). Nesta pesquisa, em particular, aplicou-se a extração do eixo médio, ou esqueleto, baseando-se no método de *Holt* em conjunto com a técnica de *staircase removal*. Este método de esqueletização apresenta esqueletos com melhor afinamento e sem desconexão, permitindo, então, detectar a intensidade do nível de cinza dos pontos ao longo desse eixo e, a partir destes valores, construir um gráfico. Pode-se constatar que a semelhança entre os dois gráficos de cromossomos que pertencem à mesma classe (par) é nítida, com picos e vales muito próximos. Portanto, o reconhecimento dos cromossomos por intermédio das bandas pode ser realizado pela identificação da similaridade entre os gráficos de intensidade de nível de cinza, e um indicador de similaridade utilizado freqüentemente é baseado no conceito de distância *Euclidiana* ( $De$ ) (HAYKIN, 2001). Considere que  $x_i$  represente um vetor  $m$ -por-1;

$$x_i = [x_{i1}, x_{i2}, \dots, x_{im}]^T$$

cujos elemento são todos números reais e o índice superior  $T$  indica o vetor transposto matricial. O vetor  $x_i$  define um ponto em um espaço de dimensão  $m$ , chamado espaço euclidiano, e a distância entre um par de vetores  $m$ -por-1,  $x_i$  e  $x_j$ , é definida utilizando-se a Equação 6.1 (HAYKIN, 2001).

$$De(x_i - x_j) = \left[ \sum_{k=1}^m (x_{ik} - x_{jk})^2 \right]^{1/2} \quad (6.1)$$

onde  $x_{ik}$  e  $x_{jk}$  são os  $k$ -ésimos elementos dos vetores de entrada  $x_i$  e  $x_j$ , respectivamente. Assim, a similaridade entre os vetores  $x_i$  e  $x_j$  é definida como o recíproco da distância euclidiana  $De(x_i - x_j)$ . Quanto mais próximos entre si estiverem os elementos

individuais dos vetores  $x_i$  e  $x_j$ , menor será a distância euclidiana e maior será a similaridade entre os vetores.

Outro indicador de similaridade entre os gráficos de intensidade de nível de cinza pode ser calculado através da quantificação do grau de associação entre eles. Para realizar esse processo de quantificação, utiliza-se a função de correlação linear de *Pearson*, que gera um índice de associação entre  $-1$  e  $+1$ , sendo que o valor  $0$  indica nenhuma associação,  $-1$  indica uma associação máxima com sinais opostos (os sinais variam em sentidos opostos) e  $+1$  indica uma associação máxima com variações dos sinais no mesmo sentido (BARBETTA, 2002). A correlação linear de *Pearson* entre dois sinais pode ser calculada utilizando-se a Equação 6.2.

$$r_{ij} = \frac{S_{ij}}{S_i \cdot S_j} \quad (6.2)$$

onde  $S_i$  e  $S_j$  são os desvios padrão dos dois sinais e  $S_{ij}$  é a covariância entre eles, que, para sinais discretos, pode ser calculada utilizando-se a Equação 6.3.

$$S_{ij} = \frac{1}{N-1} \sum_{n=1}^N [(i(n) - \mu_i) \cdot (j(n) - \mu_j)] \quad (6.3)$$

onde  $N$  é o número de amostras utilizadas e  $\mu$  é a média do sinal.

Buscando formalizar um modelo matemático que descreva essas características escolhidas para a identificação automática dos cromossomos, utilizou-se o enfoque lógico-combinatório da teoria de reconhecimento de padrões. Na Tabela 6.1, são apresentadas as variáveis escolhidas, seus domínios, critérios de comparação e ponderações. Na caracterização do domínio  $M_1$ , a variável tamanho pode admitir valores de  $T_{min}$  até  $T_{max}$ ,  $M_2$  valores de  $FF_{min}$  até  $FF_{max}$ ,  $M_3$  pode admitir valores de distância *Euclidiana* da intensidade de nível de cinza de  $De_{min}$  até  $De_{max}$  e  $M_4$  admitir valores da correlação linear de *Pearson* da associação entre os gráficos de  $CLP_{min}$  até  $CLP_{max}$ . Já para o critério de comparação ( $C_i$ ),  $O_i$  representa o objeto fixo de comparação e  $O_j$  o objeto a ser comparado.

Tabela 6.1 – Formalização do modelo matemático para a identificação automática dos cromossomos, descrito através de suas variáveis, seus domínios, critérios de comparação e ponderações.

Variável ( $x_i$ )	Domínio ( $M_i$ )	Critério de comparação ( $C_i$ )	Ponderação ( $\rho_i$ )
$x_1$ – Tamanho ( $T$ )	$[T_{\min}, T_{\max}]$	$\left  \frac{x_1(O_i) - x_1(O_j)}{T_{\max} - T_{\min}} \right $	2
$x_2$ – Fator de Forma ( $FF$ )	$[FF_{\min}, FF_{\max}]$	$\left  \frac{x_2(O_i) - x_2(O_j)}{FF_{\max} - FF_{\min}} \right $	1
$x_3$ – Distância Euclidiana ( $De$ )	$[De_{\min}, De_{\max}]$	$\left  \frac{De[x_3(O_{i_k}) - x_3(O_{j_k})] - De_{\min}}{De_{\max} - De_{\min}} \right $	3
$x_4$ – Correlação Linear de Pearson ( $CLP$ )	$[CLP_{\min}, CLP_{\max}]$	$\left  \frac{CLP[x_4(O_{i_k}) - x_4(O_{j_k})] - CLP_{\min}}{CLP_{\max} - CLP_{\min}} \right $	1

Os valores das  $n$  ponderações  $\rho_i$  são adotados pelo citogeneticista e pelo especialista matemático de acordo com os resultados dos experimentos realizados.

#### 6.2.2.2 – Formalização da Leucemia Pró-Mielocítica Aguda

Para a caracterização da leucemia pró-mielocítica aguda (LMA) sabe-se, de antemão, que ocorre uma anormalidade estrutural através da translocação  $t(15;17)$ . Uma das formas de se detectar este problema é através da determinação da diferença de tamanho entre os cromossomos do par 17, que pode ser extraído pelo comprimento de seus esqueletos. Outra forma é através da desigualdade de intensidade de nível de cinza, na região terminal do braço inferior  $q$ , de um dos cromossomos desse mesmo par. Esta diferença pode ser constatada devido aos cromossomos do par 17, quando normais, apresentarem ambas as extremidades terminadas em regiões claras e, quando na ocorrência da doença, a quebra do segmento cromossômico acontecer em uma região escura ( $17q22$ ). Esta última característica, pode ser determinada através da extração da média de intensidade de nível de cinza dos últimos pontos dos esqueletos dos cromossomos deste par (braço  $q$ ).

Mas, para a detecção desses dois fatores, existe a necessidade da extração do esqueleto de uma extremidade à outra, o que não ocorre através do processo de esqueletização, em que se tem uma perda nas extremidades dos cromossomos, justamente onde está a informação do segmento de translocação. Para contornar este problema é necessário um método que permita fazer o prolongamento das extremidades dos esqueletos, até encontrar as bordas dos cromossomos. Dentre as técnicas existentes que permitem realizar esta tarefa, pode-se citar o método de *Veinerization*, proposto por DESEILLINGNY *et al.* (1998), e o método da extração do *Grafo do Eixo Mediano*, proposto por CABRAL (2003), que permitem produzir o prolongamento, mas com a desvantagem da enorme demanda computacional.

Neste trabalho, optou-se pelo desenvolvimento de um método próprio de prolongamento das extremidades do esqueleto, utilizando-se o processo de extrapolação. Este método baseia-se no ajuste das curvas pela técnica dos mínimos quadrados, em que a idéia é de tentar encontrar uma curva que seja uma boa aproximação dos dados disponíveis. O ajuste das curvas permite, então, que a partir de dados conhecidos, possam ser feitas extrapolações com uma certa margem de segurança (RUGGIERO e LOPES, 1988). Considere uma Tabela de pontos  $(x_1, f(x_1)), (x_2, f(x_2)), \dots, (x_m, f(x_m))$  com  $x_1, x_2, \dots, x_m$ , pertencentes a um intervalo  $[a,b]$ , o ajuste das curvas consiste em, escolhidas  $n$  funções  $g_1(x), g_2(x), \dots, g_n(x)$ , obter  $n$  constantes  $\alpha_1, \alpha_2, \dots, \alpha_n$  tais que  $\varphi(x)$ , utilizando-se a Equação 6.4, se aproxime ao máximo de  $f(x)$ .

$$\varphi(x) = \alpha_1 g_1(x) + \alpha_2 g_2(x) + \dots + \alpha_n g_n(x) \quad (6.4)$$

Pode-se dizer que  $\varphi(x)$  é um modelo matemático linear porque os coeficientes  $\alpha_1, \alpha_2, \dots, \alpha_n$  são lineares, embora as funções  $g_1(x), g_2(x), \dots, g_n(x)$  possam ser funções não-lineares. Aplicando-se o método dos *mínimos quadrados* (ou soma dos desvios ao quadrado), pode-se estabelecer um conceito de máxima proximidade entre as funções  $f(x)$  e  $\varphi(x)$ , em que o objetivo é encontrar os  $n$  coeficientes  $\alpha_1, \alpha_2, \dots, \alpha_n$  de tal forma que o erro  $D$  (resíduo), definido pela Expressão 6.5, seja mínimo (CLÁUDIO e MARINS, 1989).

$$D = \sum_{k=1}^m d_k^2 = \sum_{k=1}^m [f(x_k) - \varphi(x_k)]^2 = F(\alpha_1, \alpha_2, \dots, \alpha_n) \quad (6.5)$$

onde  $d_k$  é o desvio da amostra conhecida no ponto  $x_k$ ,  $m$  é o número total de pontos e  $k = 1, 2, 3, \dots, m$ . No critério dos mínimos quadrados,  $D$  pode ser expresso utilizando-se a Equação 6.6.

$$D = F(\alpha_1, \alpha_2, \dots, \alpha_n) = \sum_{k=1}^m [f(x_k) - \alpha_1 g_1(x_k) - \alpha_2 g_2(x_k) - \dots - \alpha_n g_n(x_k)]^2 \quad (6.6)$$

Do cálculo diferencial, sabe-se que, para se obter um ponto de mínimo de  $F(\alpha_1, \alpha_2, \dots, \alpha_n)$ , deve-se, inicialmente, encontrar seus pontos críticos, ou seja, os coeficientes  $(\alpha_1, \alpha_2, \dots, \alpha_n)$ . Para isso, encontram-se as derivadas parciais e igualam-se a zero, como ilustra a Equação 6.7.

$$\frac{\partial F}{\partial \alpha_j} \Big|_{(\alpha_1, \alpha_2, \dots, \alpha_n)} = 0, \quad \text{para } j = 1, 2, \dots, n \quad (6.7)$$

Derivando a função  $F$ , obtém-se a Equação 6.8.

$$\frac{\partial F}{\partial \alpha_j} = 2 \sum_{k=1}^m [f(x_k) - \alpha_1 g_1(x_k) - \alpha_2 g_2(x_k) - \dots - \alpha_n g_n(x_k)] [-g_j(x_k)] \quad j = 1, 2, \dots, n \quad (6.8)$$

E, impondo-se a condição necessária para o mínimo (Equação 6.7 igual a zero):

$$\sum_{k=1}^m [f(x_k) - \alpha_1 g_1(x_k) - \alpha_2 g_2(x_k) - \dots - \alpha_n g_n(x_k)] [g_j(x_k)] = 0, \quad j = 1, 2, \dots, n$$

De forma explícita, tem-se:

$$\begin{aligned} \sum_{k=1}^m [f(x_k) - \alpha_1 g_1(x_k) - \alpha_2 g_2(x_k) - \dots - \alpha_n g_n(x_k)] g_1(x_k) &= 0 \\ \sum_{k=1}^m [f(x_k) - \alpha_1 g_1(x_k) - \alpha_2 g_2(x_k) - \dots - \alpha_n g_n(x_k)] g_2(x_k) &= 0 \\ \vdots \\ \sum_{k=1}^m [f(x_k) - \alpha_1 g_1(x_k) - \alpha_2 g_2(x_k) - \dots - \alpha_n g_n(x_k)] g_n(x_k) &= 0 \end{aligned}$$

Separando-se os somatórios e isolando os termos com variáveis dos termos constantes, obtém-se a Expressão 6.9.

$$\begin{aligned} \left[ \sum_{k=1}^m g_1(x_k) g_1(x_k) \right] \alpha_1 + \left[ \sum_{k=1}^m g_2(x_k) g_1(x_k) \right] \alpha_2 + \dots + \left[ \sum_{k=1}^m g_n(x_k) g_1(x_k) \right] \alpha_n &= \sum_{k=1}^m f(x_k) g_1(x_k) \\ \left[ \sum_{k=1}^m g_1(x_k) g_2(x_k) \right] \alpha_1 + \left[ \sum_{k=1}^m g_2(x_k) g_2(x_k) \right] \alpha_2 + \dots + \left[ \sum_{k=1}^m g_n(x_k) g_2(x_k) \right] \alpha_n &= \sum_{k=1}^m f(x_k) g_2(x_k) \quad (6.9) \\ \vdots \\ \left[ \sum_{k=1}^m g_1(x_k) g_n(x_k) \right] \alpha_1 + \left[ \sum_{k=1}^m g_2(x_k) g_n(x_k) \right] \alpha_2 + \dots + \left[ \sum_{k=1}^m g_n(x_k) g_n(x_k) \right] \alpha_n &= \sum_{k=1}^m f(x_k) g_n(x_k) \end{aligned}$$

As equações acima formam um sistema de equações lineares com  $n$  equações e  $n$  incógnitas que, de forma matricial, pode ser representado por  $A\underline{\alpha} = \underline{b}$ , em que chega-se à Equação 6.10.

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \quad \underline{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \quad \underline{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \quad (6.10)$$

cujos valores dos elementos da matriz de coeficientes e do vetor independente são determinados pelas Equações 6.11 e 6.12.

$$a_{ij} = a_{ji} = \sum_{k=1}^m g_i(x_k) g_j(x_k) \quad \text{para } i=1, \dots, n \text{ e } j=1, \dots, n; \quad (6.11)$$

$$b_i = \sum_{k=1}^m f(x_k)g_i(x_k) \quad \text{para } i = 1, \dots, n; \quad (6.12)$$

onde  $n$  é o número de termos da função  $\varphi(x)$  a ser ajustada e  $m$  é o número de pontos da amostra conhecida.

Entretanto, o maior problema do ajuste de curvas consiste em definir qual tipo de função  $\varphi(x)$  melhor se aproxima à função  $f(x)$ . Para resolver este problema, aplica-se o método da soma dos desvios ao quadrado, para extrair, dentre os tipos de funções existentes, aquela que apresenta o menor desvio  $D$  definido pela Equação 6.5. Logo, a função que apresenta o menor desvio é a função que melhor se aproxima à função  $f(x)$ .

Para demonstrar o conceito de ajuste de curvas, tomou-se como exemplo um cromossomo do par número 17, e dele foi extraído seu esqueleto e sua borda, como mostra a Figura 6.3. Considerando-se o esqueleto extraído como uma representação do gráfico cartesiano, pode-se ordenar, em uma tabela, seus pontos também conhecidos, como mostrado na Tabela 6.2.

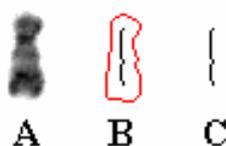


Figura 6.3 – (A) Cromossomo do par número 17, (B) extração do esqueleto do cromossomo e (C) representação do esqueleto como um gráfico cartesiano.

Tabela 6.2 – Valores ordenados dos pontos do gráfico cartesiano da Figura 6.3.

X	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
f(x)	35	34	34	34	34	34	34	34	35	35	34	34	34	34	34	35

De acordo com análise do gráfico cartesiano (esqueleto) da Figura 6.3, percebe-se claramente a semelhança com uma reta, entretanto, há casos em que os cromossomos estão muito tortos, assemelhando-se mais a uma parábola ou exponencial. Para ilustrar este conceito, através dos valores apresentados na Tabela 6.2, busca-se ajustá-los,

primeiramente, a uma reta e a uma parábola e, em seguida, compara-se os valores destas duas funções para saber qual delas apresenta o menor desvio.

Logo, a equação geral de uma *reta* pode ser definida pela função  $\varphi(x) = \alpha_1 + \alpha_2 x$ , onde as funções  $g_1(x) = 1$ ,  $g_2(x) = x$ . Deve-se, primeiramente, elaborar o sistema linear  $A\underline{\alpha} = \underline{b}$ :

$$a_{11} = \sum_{k=1}^{16} 1^2 = 16, \quad a_{12} = a_{21} = \sum_{k=1}^{16} 1 \cdot x_k = 552 \quad \text{e} \quad a_{22} = \sum_{k=1}^{16} x_k^2 = 19384$$

$$b_1 = \sum_{k=1}^{16} 1 \cdot f(x_k) = 548 \quad \text{e} \quad b_2 = \sum_{k=1}^{16} x_k f(x_k) = 18908$$

Através destes valores calculados, chega-se ao sistema linear  $A\underline{\alpha} = \underline{b}$ :

$$\begin{bmatrix} 16 & 552 \\ 552 & 19384 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 548 \\ 18908 \end{bmatrix}$$

Pelo método de *Gauss*, encontra-se os coeficientes  $\alpha_1$  e  $\alpha_2$ , e o resultado é a equação da reta ajustada para  $\varphi(x) = 34,04 + 0,0059x$ , onde  $F(\alpha_1, \alpha_2)$  atinge seu ponto mínimo. Aplicando-se os mesmos valores da Tabela 6.2, para o caso da parábola, a equação geral é definida por  $\varphi(x) = \alpha_1 + \alpha_2 x + \alpha_3 x^2$ , com  $g_1(x) = 1$ ,  $g_2(x) = x$ , e  $g_3(x) = x^2$ . Para elaborar o sistema linear  $A\underline{\alpha} = \underline{b}$ , tem-se:

$$a_{11} = \sum_{k=1}^{16} 1^2 = 16, \quad a_{12} = a_{21} = \sum_{k=1}^{16} 1 \cdot x_k = 552 \quad \text{e} \quad a_{22} = a_{13} = a_{31} = \sum_{k=1}^{16} x_k^2 = 19384$$

$$a_{23} = a_{32} = \sum_{k=1}^{16} x_k^3 = 692208 \quad \text{e} \quad a_{33} = \sum_{k=1}^{16} x_k^4 = 25108168$$

$$b_1 = \sum_{k=1}^{16} 1 \cdot f(x_k) = 548, \quad b_2 = \sum_{k=1}^{16} x_k f(x_k) = 18908 \quad \text{e} \quad b_3 = \sum_{k=1}^{16} x_k^2 f(x_k) = 664070$$

Resultando no sistema linear  $A\underline{\alpha} = \underline{b}$ :

$$\begin{bmatrix} 16 & 552 & 19384 \\ 552 & 19384 & 692208 \\ 19384 & 692208 & 25108168 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 548 \\ 18908 \\ 664070 \end{bmatrix}$$

Pelo método de *Gauss*, encontra-se os coeficientes  $\alpha_1$ ,  $\alpha_2$  e  $\alpha_3$ , e o resultado é a equação da parábola ajustada para  $\varphi(x) = 68,63 - 2,03x + 0,029x^2$ , onde  $F(\alpha_1, \alpha_2, \alpha_3)$  atinge seu ponto mínimo. Para a verificação do melhor ajuste de curva, calcula-se a soma dos desvios quadráticos para a equação da reta e para a equação da parábola, logo:

$$\text{Para a reta: } \sum_{k=1}^{16} d_k^2 = 0,04898$$

$$\text{E para a parábola: } \sum_{k=1}^{16} d_k^2 = 0,08833$$

De acordo com os valores obtidos entre a reta e a parábola, a reta se ajusta melhor aos pontos tabelados. Para esses dois casos, a função a ser ajustada torna-se linear com relação aos coeficientes  $\alpha$ 's. O método dos mínimos quadrados aplicado deve ser utilizado somente para funções lineares, porém, há casos em que os pontos do gráfico cartesiano necessitam do ajuste para uma função não-linear, isto é, uma curva exponencial. A solução, neste caso, é linearizar o problema através de uma transformação conveniente com o uso de logaritmos (CAMPAGNOLO, 2003).

Para demonstrar este conceito, deseja-se ajustar a curva exponencial  $\varphi(x) = \alpha_1 e^{-\alpha_2 x}$ , com  $\alpha_1$  e  $\alpha_2$  positivos, também para o gráfico cartesiano  $f(x)$  apresentado através da Figura 6.3 do esqueleto do cromossomo. Deve-se, primeiramente, efetuar a linearização  $\Phi(x) = \ln(\varphi(x))$ , resultando na Equação 6.13.

$$z = \ln[f(x)] \approx \ln[\varphi(x)] = \ln(\alpha_1) - \alpha_2 x = \Phi(x) \quad (6.13)$$

Definindo:  $\gamma_1 = \ln(\alpha_1)$  e  $\gamma_2 = \alpha_2$ , obtém-se a Equação 6.14.

$$\Phi(x) = \gamma_1 - \gamma_2 x \tag{6.14}$$

A função  $\Phi(x)$  é linear com relação aos coeficientes  $\gamma$ 's, desta forma, pode-se determinar os valores de  $\gamma_1$  e  $\gamma_2$  pelo método dos mínimos quadrados e os valores de  $\alpha_1$  e  $\alpha_2$  são obtidos pelas relações definidas anteriormente. Os valores de  $\gamma_1$  e  $\gamma_2$  ajustam a função  $\Phi(x)$  à função  $z = \ln(f(x))$  no sentido dos mínimos quadrados. Não se pode afirmar que os parâmetros  $\alpha_1$  e  $\alpha_2$  obtidos a partir de  $\gamma_1$  e  $\gamma_2$  ajustam  $\varphi(x)$  a  $f(x)$  no mesmo sentido. Para exemplificar, considera-se os valores da Tabela 6.1, buscando ajustar para a curva geométrica  $\varphi(x) = \alpha_1 e^{-\alpha_2 x}$ , que utilizando-se a Equação 6.14 tem-se  $g_1(x) = 1$  e  $g_2(x) = x$ . O ajuste será feito para  $z = \ln[f(x)]$ , logo, os valores da Tabela 6.2 são modificados para os valores da Tabela 6.3.

Tabela 6.3 – Valores da Tabela 6.1 modificados com o ajuste para  $z = \ln[f(x)]$ .

x	27	28	29	30	.....	38	39	40	41	42
$z = \ln[f(x)]$	3,555	3,526	3,526	3,526	.....	3,526	3,526	3,526	3,526	3,555

Para resolver o sistema linear  $A\underline{\alpha} = \underline{b}$ , tem-se:

$$a_{11} = \sum_{k=1}^{16} 1^2 = 16, \quad a_{12} = a_{21} = \sum_{k=1}^{16} 1 \cdot x_k = 552 \quad \text{e} \quad a_{22} = \sum_{k=1}^{16} x_k^2 = 19384$$

$$b_1 = \sum_{k=1}^{16} z(x_k) \cdot 1 = 56,537 \quad \text{e} \quad b_2 = \sum_{k=1}^{16} z(x_k) \cdot x_k = 1950,412$$

$$\begin{bmatrix} 16 & 552 \\ 552 & 19384 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} 56,537 \\ 1950,412 \end{bmatrix}$$

E, pelo método de *Gauss*, encontra-se:

$$\underline{\gamma} = \begin{bmatrix} 3,548 \\ -0,00041 \end{bmatrix}$$

Portanto,

$$\alpha_1 = e^{\gamma_1} = e^{3,548} = 34,743$$

$$\alpha_2 = -\gamma_2 = 0,00041$$

$$\varphi(x) = \alpha_1 e^{-\alpha_2 x} = 34,743 e^{-0,00041x}$$

Para a verificação do ajuste da curva, calcula-se a soma dos desvios quadráticos para a equação da curva exponencial, logo:

$$\sum_{k=1}^{16} d_k^2 = 0,06833$$

De acordo com os valores obtidos da soma dos desvios quadráticos da reta, da parábola e da curva exponencial, constata-se que a reta ajusta-se melhor aos pontos tabelados do gráfico cartesiano da Figura 6.3 devido a seu desvio ser menor.

Doravante, a realização do processo de prolongamento das extremidades depende da utilização da curva com menor desvio quadrático, para a extrapolação dos pontos extremos do esqueleto até encontrar a borda dos cromossomos. Para demonstrar este conceito, foi considerado o exemplo do gráfico cartesiano (esqueleto), mostrado através da Figura 6.3, na qual é realizado o prolongamento das extremidades superior e inferior utilizando-se a função da reta  $\varphi(x) = 34,04 + 0,0059x$ . Considere os pontos finais do esqueleto, extremidade superior e inferior, como o ponto central de uma máscara de dimensão 3x3. Para o prolongamento da extremidade superior, as posições admissíveis para os sucessivos prolongamentos são  $P_{y-1,x-1}$ ,  $P_{y,x-1}$  e  $P_{y+1,x-1}$ , como mostrado na máscara apresentada na Figura 6.4.

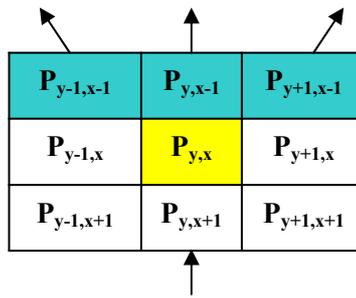


Figura 6.4 – Posições admissíveis para os sucessivos prolongamentos dos pontos da extremidade superior,  $P_{y-1,x-1}$ ,  $P_{y,x-1}$  e  $P_{y+1,x-1}$ .

No sentido oposto, o prolongamento da extremidade inferior é realizado considerando-se as posições admissíveis para os sucessivos prolongamentos em  $P_{y-1,x+1}$ ,  $P_{y,x+1}$  e  $P_{y+1,x+1}$ , como mostrado na máscara apresentada na Figura 6.5.

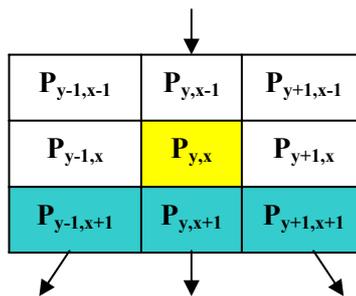


Figura 6.5 – Posições admissíveis para os sucessivos prolongamentos dos pontos da extremidade inferior,  $P_{y-1,x+1}$ ,  $P_{y,x+1}$  e  $P_{y+1,x+1}$ .

Os valores obtidos da equação da reta ajustada  $\varphi(x)$  são os sucessivos pontos extrapolados do esqueleto no eixo  $x$  (linhas), resultando nos valores do eixo  $y$  (coluna). Apesar dos valores obtidos de  $x$  serem inteiros, os resultados obtidos em  $y$  são pontos flutuantes, necessitando-se, portanto, de uma aproximação para o valor inteiro mais próximo. E este processo de extrapolação dos pontos do esqueleto é realizado até que uma posição da máscara, das três admissíveis, encontre algum ponto da borda. Para apresentar esse exemplo do prolongamento das extremidades do esqueleto pela curva

ajustada pelo método dos mínimos quadrados, a Figura 6.6 ilustra este processo utilizando-se o mesmo cromossomo apresentado pela Figura 6.3.

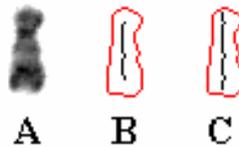


Figura 6.6 – (A) Cromossomo do par número 17, (B) extração do esqueleto do cromossomo e (C) prolongamento das extremidades utilizando-se uma curva ajustada pelos mínimos quadrados.

A Figura 6.7 mostra mais exemplos de prolongamento das extremidades, utilizando cromossomos com diversos formatos para ilustrar a eficiência do ajuste de curvas.

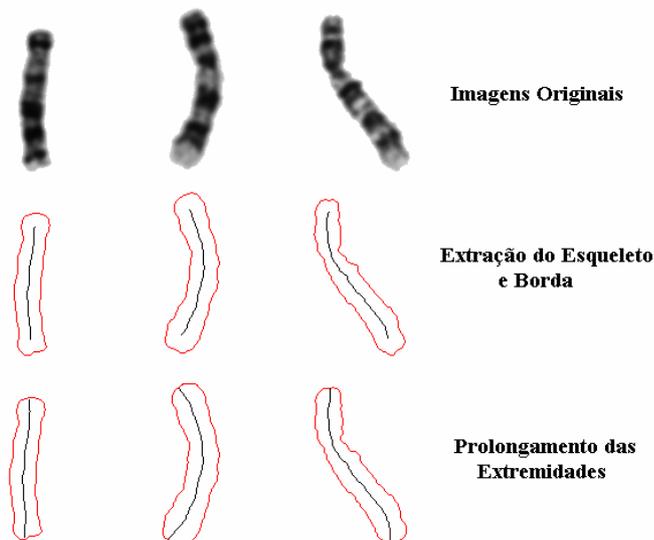


Figura 6.7 – Exemplos do método de prolongamento das extremidades dos esqueletos, utilizando-se diversos cromossomos com formatos diferentes.

Com as extremidades dos eixos prolongadas, tem-se agora o eixo médio completo dos cromossomos sendo possível, então, a extração das informações nas extremidades dos cromossomos. O passo seguinte constitui a formalização de um modelo matemático para a caracterização da leucemia pró-mielocítica aguda (LMA), utilizando-se o enfoque lógico-

combinatório da teoria de reconhecimento de padrões. A Tabela 6.4 apresenta o modelo matemático das variáveis escolhidas para a caracterização da LMA, seus respectivos domínios, critérios de comparação e ponderações. A caracterização do domínio  $M_5$  da variável *comprimento do esqueleto do cromossomo* ( $CE$ ) admite somente dois valores,  $CE_1$  e  $CE_2$ , porque existe somente dois cromossomos em comparação, sendo que  $x_5(O_{maior})$  representa o maior valor do comprimento medido entre  $CE_1$  e  $CE_2$ , e  $S_{perc}$  representa o valor máximo percentual do segmento translocado em relação ao tamanho normal do cromossomo do par 17. Já o domínio  $M_6$  da variável *diferença de intensidade de nível de cinza na região terminal do braço q* ( $DI$ ) admite somente dois valores,  $DI_1$  e  $DI_2$ , que constituem os valores dos dois cromossomos em comparação, sendo que  $V_{dicmáx}$  representa o valor máximo de diferença da intensidade de nível de cinza que um cromossomo translocado do par 17 pode ter em relação ao outro normal do mesmo par.

Tabela 6.4 – Formalização do modelo matemático para a detecção da LMA descrito utilizando-se suas variáveis, seus domínios, critérios de comparação e ponderações.

Variável ( $x_i$ )	Domínio ( $M_i$ )	Critério de comparação ( $C_i$ )	Ponderação ( $\rho_i$ )
$x_5$ – Comprimento do Esqueleto do Cromossomo ( $CE$ )	$[CE_1, CE_2]$	$1 - \frac{ x_5(O_1) - x_5(O_2) }{S_{perc}/100 * x_5(O_{maior})}$	1
$x_6$ – Diferença de Intensidade de Cinza da Região Terminal do Braço q ( $DI$ )	$[DI_1, DI_2]$	$1 - \frac{ x_6(O_1) - x_6(O_2) }{V_{dicmáx}}$	1

Os valores  $S_{perc}$ ,  $V_{dicmáx}$  e  $\rho_i$  são adotados pelo citogeneticista e pelo especialista matemático de acordo com os resultados constatados nos experimentos com vários exemplos de imagens de indivíduos portadores de leucemia pró-mielocítica aguda.

### 6.2.3 – Modo de Solução do Problema Expresso em Termos Matemáticos

Esta etapa constitui o detalhamento do desenvolvimento do modelo computacional da sistemática propriamente dita, isto é, a implementação do *Sistema Computadorizado para Análise e Classificação Citogenética* (SCACC). Tal sistemática deve constar de determinadas operações que um sistema de visão artificial efetua sobre uma imagem: transformação, pré-processamento, segmentação, descrição, reconhecimento e interpretação. Essas operações são descritas na seqüência de suas implementações.

A primeira operação implementada no SCACC foi uma rotina de aquisição de imagens de células metafásicas. Esta rotina de aquisição permite selecionar o *driver* de qualquer tipo de equipamento de vídeo instalado no computador e adquirir as imagens em três níveis de resolução: 640x480, 320x240 e 160x120. O padrão de vídeo utilizado é o NTSC-M, e estão disponíveis para a configuração três formatos de imagens: 32, 24 ou 15 *bits* RGB. A imagem proveniente do microscópio pode ser vista em tempo real por uma janela (janela de aquisição) e, após o citogeneticista escolher a metáfase mais adequada para o estudo, a imagem é transferida para a janela principal do SCACC para ser, então, processada, com mostrado na Figura 6.8.

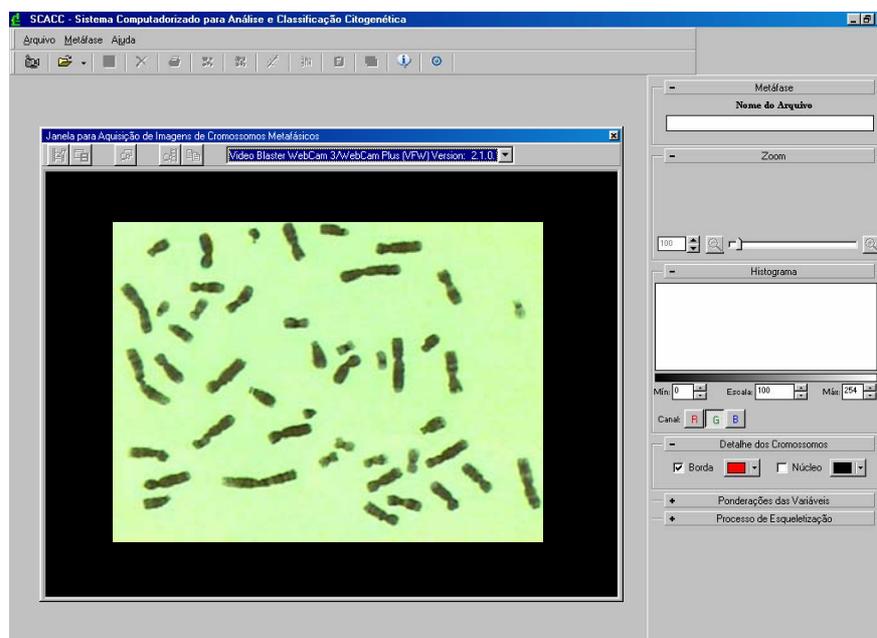


Figura 6.8 – Janela de aquisição das células metafásicas provenientes do microscópio óptico, implementada no Sistema Computadorizado para Análise e Classificação Citogenética (SCACC).

Na própria janela de aquisição das imagens dos cromossomos metafásicos, é permitido que se faça um pré-processamento das imagens para correções de cor, através da compensação de brilho, contraste, saturação e matiz. Ainda antes da etapa de segmentação das imagens, outra operação de pré-processamento é necessária, aplicando-se um filtro mediano 3X3 para atenuar as componentes de alta frequência e provocar uma leve suavização nas imagens.

Para a detecção dos cromossomos na imagem, implementou-se, no SCACC, uma operação de limiarização (*thresholding*) no modelo RGB, baseada na utilização do histograma. A imagem é decomposta em suas componentes primárias de cor e uma rotina permite a escolha de uma destas componentes para que a segmentação seja aplicada utilizando-se seu histograma, como mostrado na janela da Figura 6.9. Nesta janela, é permitido também a escolha da faixa de intensidade de corte, através dos valores mínimos (Mín) e máximos (Máx), e a escala utilizada.

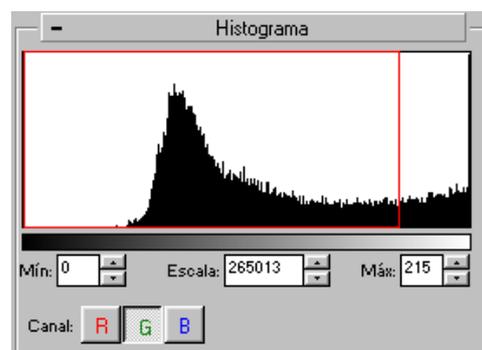


Figura 6.9 – Ilustração do processo de segmentação implementada ao SCACC, baseado na utilização do histograma do modelo RGB.

A imagem resultante desse processo é uma imagem binária, com fundo branco e os objetos detectados em pretos. O processo seguinte é a detecção das bordas dos objetos, pelo qual optou-se pela técnica do operador baseado no laplaciano 3X3, que permitiu extrair as bordas com fronteiras totalmente fechadas. Em seguida, os objetos e as bordas são sobrepostos à imagem original, permitindo ao usuário a escolha de cores diferentes para os objetos detectados e também a borda. A Figura 6.10 ilustra este processo de segmentação da imagem de uma célula metafásica, com os cromossomos detectados

apresentados em preto e suas respectivas bordas em vermelho, sobrepostos à imagem original.

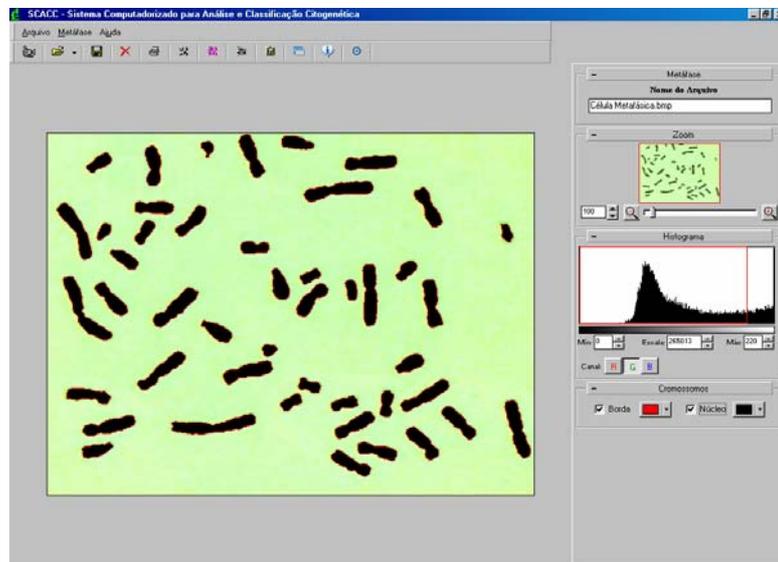


Figura 6.10 – Processo de segmentação da imagem de uma célula metafásica, com os cromossomos detectados apresentados em preto e suas respectivas bordas em vermelho, sobrepostos à imagem original.

Com o processo de detecção dos objetos de interesse realizado, pode-se extrair os atributos necessários para a identificação dos cromossomos. A primeira medida implementada no SCACC diz respeito à medida de campo, que tem a função de contar o número de objetos pertencentes à imagem total. É importante mencionar que os objetos são detectados na imagem de baixo para cima e recebendo um número, porém esta numeração ainda não significa sua classificação. Quanto às medidas de regiões, foram extraídas a área dos objetos pela contagem do número de *pixels*, e o perímetro é calculado pela contagem do número de passos horizontais, verticais e diagonais dos vizinhos da borda como definido na Equação 4.5. Por fim, extrai-se também o eixo médio ou o esqueleto através do processo de esqueletização, fundamentando-se no método de *Holt*, em conjunto com a técnica de *staircase removal*. Entretanto, para a implementação do algoritmo *Holt*, utilizou-se as duas expressões lógicas definidas pelas Equações 4.9 e 4.10, ao invés da combinação das duas, definida pela Equação 4.11, pois o esqueleto resultante desta última equação apresentava bifurcações nas extremidades dos esqueletos. O

processo de afinamento implementado pelas duas expressões lógicas de *Holt* pode ter ficado um pouco mais lento, mas aplicado em conjunto com a técnica *staircase removal*, conseguiu-se esqueletos com melhor afinamento. O resultado dessa implementação pode ser visto na janela principal do SCACC, apresentada na Figura 6.11, com a imagem de uma célula metafásica. Além do esqueleto, as bordas dos cromossomos são também apresentadas e destacadas em vermelho, para a análise comparativa do tamanho original dos cromossomos.

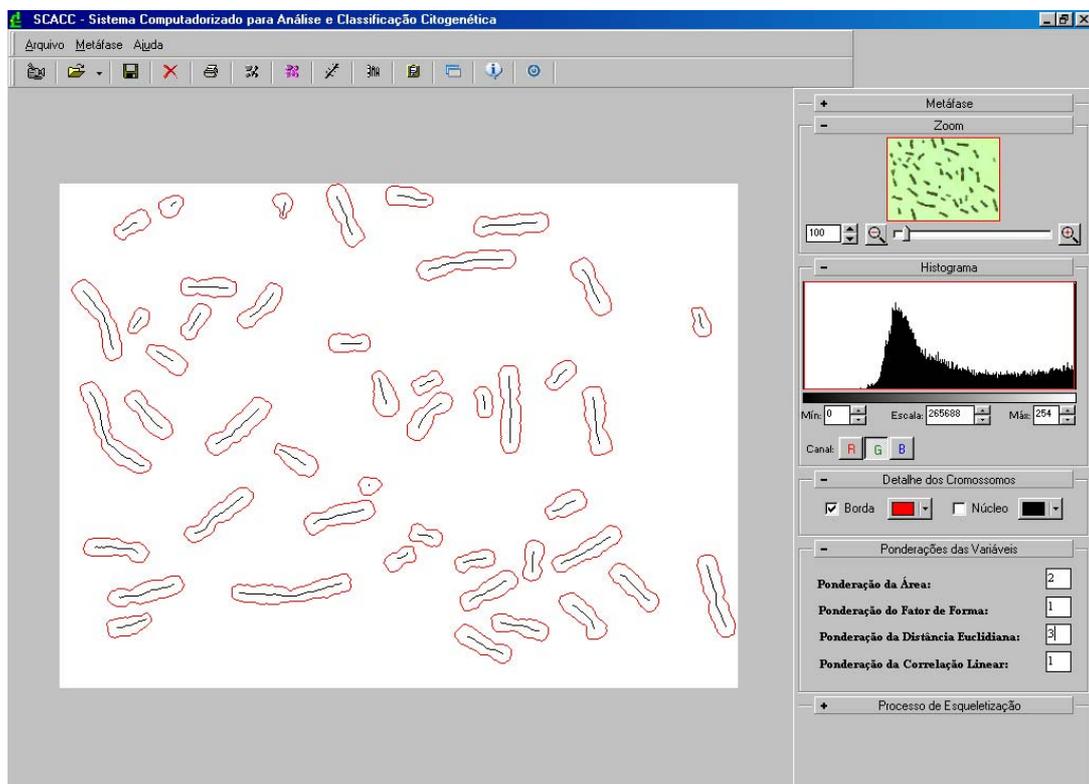


Figura 6.11 – Ilustração da janela principal do SCACC, com uma imagem resultante do processo de detecção de bordas e esqueletização dos cromossomos de uma célula metafásica.

A partir das características extraídas dos cromossomos, área, perímetro e esqueleto, pode-se, então, calcular diversas medidas que descrevem os cromossomos. Da característica área, pode-se ter a identificação do tamanho dos cromossomos e, da área e do perímetro, pode-se calcular seu fator de forma definido pela Equação 4.6. O fator de forma é apresentado em uma escala [0,1]. Se o cromossomo apresenta-se com um

formato tende a um círculo perfeito (cromossomos menores), seu fator de forma tende a 1 (um), e se o cromossomo apresentar-se de forma mais alongada (cromossomos maiores), seu fator de forma tende a 0 (zero). Logo, essa característica também pode ser utilizada como identificação dos cromossomos, buscando aqueles que são mais similares ou diferenciando-os, como mostra a Figura 6.12, ilustrando um cromossomo grande, do par número 1, com fator de forma 0,209 e um outro cromossomo pequeno, do par número 21, com fator de forma 0,563.



Figura 6.12 – (A) Cromossomo do par número 1 com fator de forma 0,209 e (B) cromossomo do par 21 com fator de forma 0,563.

A necessidade de extração do esqueleto dos cromossomos ocorre devido ao fato de se tentar identificar as características das bandas dos cromossomos. De posse do eixo médio do cromossomo é possível obter, da posição de cada *pixel* pertencente a este esqueleto, o valor da intensidade de nível de cinza dos cromossomos. Logo, percorrendo o esqueleto de uma ponta a outra, no sentido do braço *p* para *q*, consegue-se construir um gráfico com valores de intensidade de nível de cinza. A Figura 6.13 ilustra um exemplo desse tipo de gráfico, extraído do esqueleto de um cromossomo do par número 7, onde os picos no gráfico representam as regiões escuras do cromossomo (bandas) e os vales representam as regiões claras, isto é, as regiões entre duas bandas. Os pontos do esqueleto de cada cromossomo são, então, armazenados em um vetor com seus respectivos valores de intensidade de nível de cinza para o processamento subsequente.

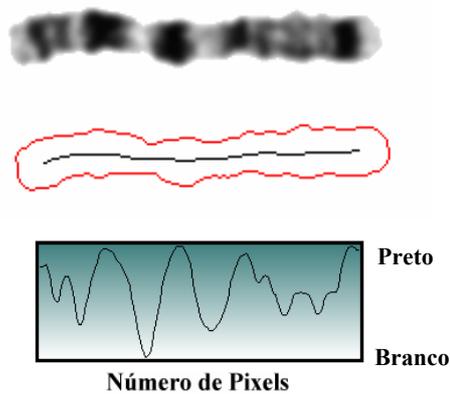


Figura 6.13 – Exemplo de gráfico de intensidade de nível de cinza, extraído do esqueleto de um cromossomo do par número 7, onde os picos representam as regiões escuras (bandas) e os vales representam regiões claras (regiões entre duas bandas).

A partir do conjunto de cromossomos detectados e seus parâmetros medidos, o passo seguinte é distinguí-los de forma automática, por classes de similaridade, para compor o cariógrama. Para tanto, aplica-se o modelo matemático apresentado na Tabela 6.1 pela função semelhança definida na Equação 5.3, para calcular o fator de semelhança entre os cromossomos, como mostrado na Equação 6.15.

$$\beta_{cromossomos} = 1 - \left[ \frac{\rho_A(C_A) + \rho_{FF}(C_{FF}) + \rho_{De}(C_{De})}{\rho_A + \rho_{FF} + \rho_{De}} \right] \quad (6.15)$$

onde  $\beta$  é definido como o fator de semelhança entre os cromossomos, as ponderações  $\rho_A$ ,  $\rho_{FF}$  e  $\rho_{De}$  são os pesos adotados para as características área ( $A$ ), fator de forma ( $FF$ ) e diferença euclidiana ( $De$ ), respectivamente, e  $C_A$ ,  $C_{FF}$  e  $C_{De}$  os critérios de comparação para as características anteriormente citadas.

Para ilustrar a aplicação do fator de semelhança para o reconhecimento automático dos cromossomos, foram calculadas as semelhanças entre três cromossomos, denominados aqui de A, B e C, e mostrados através da Figura 6.14. Os cromossomos A e B constituem o par número 7 e C um cromossomo do par número 14.

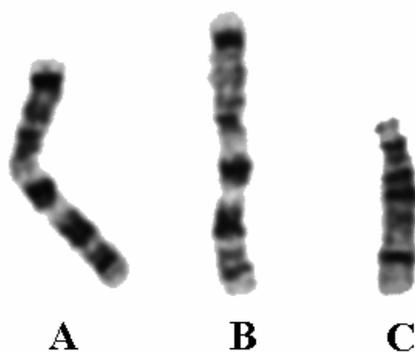


Figura 6.14 – Três cromossomos utilizados no cálculo do fator de semelhança para o reconhecimento automático dos mesmos. Os cromossomos (A) e (B) constituem o par número 7 e (C) um cromossomo do par número 14.

Cabe lembrar que o resultado do cálculo do fator de semelhança encontra-se no limite  $[0,1]$  devido ao fator de estandarização. Logo, se o resultado desse cálculo for igual a um, significa semelhança total, ou seja, que os dois cromossomos comparados são iguais. Entretanto, se a semelhança for igual a zero, os cromossomos são totalmente diferentes. Os valores intermediários representam, então, o grau de similaridade entre os mesmos. Para ilustrar este conceito, utilizou-se o SCACC para extrair as medidas dos cromossomos mostrados pela Figura 6.14, e essas medidas são apresentadas na Tabela 6.5:

Tabela 6.5 – Medidas extraídas e calculadas pelo SCACC dos cromossomos A, B e C da Figura 6.14, que são utilizadas no cálculo do fator de semelhança.

<b>Cromossomos</b>	<b>Tamanho (Área)</b>	<b>Fator de Forma</b>	<b>Distância Euclidiana</b>
<b>CR<sub>A</sub></b>	3594	0,285	-----
<b>CR<sub>B</sub></b>	3890	0,259	entre CR <sub>A</sub> e CR <sub>B</sub> 0
<b>CR<sub>C</sub></b>	2490	0,395	entre CR <sub>A</sub> e CR <sub>C</sub> 1

Substituindo os valores na Equação 6.10, os fatores de semelhança entre CR<sub>A</sub> e CR<sub>B</sub> e entre CR<sub>A</sub> e CR<sub>C</sub> são:

$$\beta_{\text{cromossomos}}(I(CR_A), I(CR_B)) = 1 - \left[ \frac{1(0,21) + 1(0,19) + 2(0)}{4} \right] \cong 0,89$$

$$\beta_{\text{cromossomos}}(I(CR_A), I(CR_C)) = 1 - \left[ \frac{1(0,78) + 1(0,80) + 2(1)}{4} \right] \cong 0,10$$

Levando em conta o grau de semelhança obtido entre  $A-B$  e  $A-C$ , pode-se concluir que o cromossomo  $CR_A$  tem um grau de semelhança maior com o cromossomo  $CR_B$  do que com o cromossomo  $CR_C$ .

Neste sentido, foi implementada uma rotina de classificação no SCACC que processa este cálculo para todos os cromossomos encontrados na imagem, com a finalidade de encontrar seus pares homólogos. Em seguida, buscou-se implementar outra rotina para determinar as 24 classes de cromossomos, ou seja, os 22 pares além dos cromossomos sexuais X e Y. De acordo com o que foi exposto sobre os padrões internacionais de nomenclatura citogenética, ISCN (1995), a ordem dos pares de cromossomos para a composição do cariógrama está disposta de forma decrescente de tamanho. A análise do tamanho de um cromossomo metafásico é feita tomando-se por referência o tamanho de todos os outros presentes na mesma metáfase, dessa forma, tentou-se ordenar os pares de cromossomos homólogos de forma decrescente aos valores da área calculada. Um dos sistemas que compõe o cariógrama por ordem decrescente de tamanho é o sistema MAKSYS, desenvolvido por POPESCU (2001). O resultado da rotina de classificação dos cromossomos é apresentado em uma janela do SCACC, mostrando os cromossomos da metáfase do cariótipo de um indivíduo do sexo feminino, classificados e pareados automaticamente, ou seja, o cariógrama completo é mostrado na Figura 6.15.

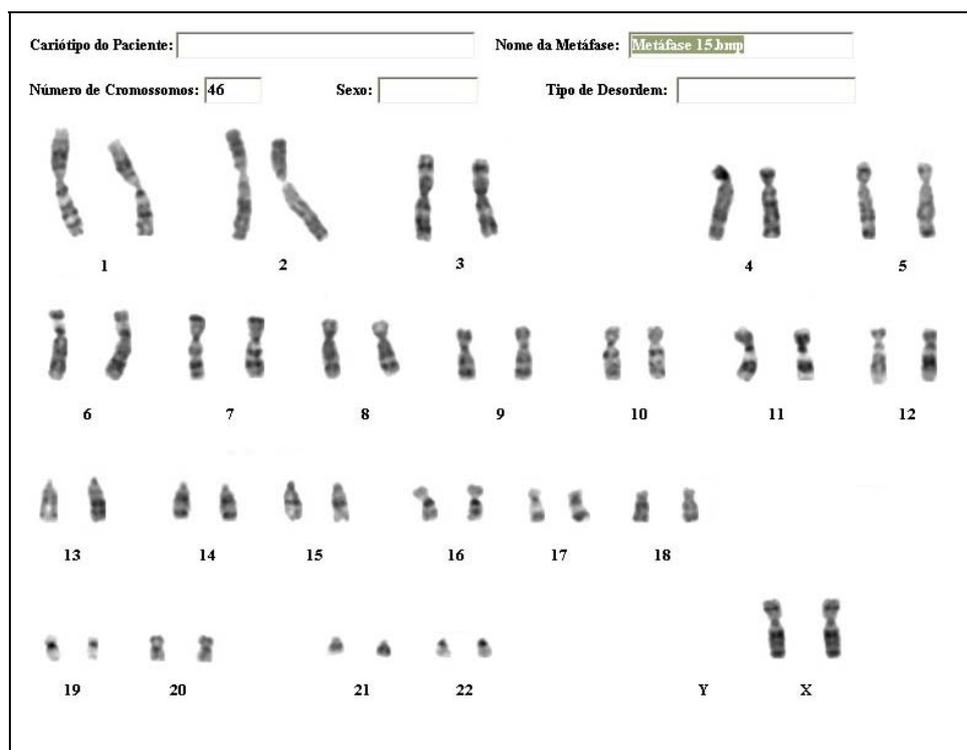


Figura 6.15 – Janela do kariótipo completo de um indivíduo do sexo feminino, apresentando os cromossomos classificados e pareados automaticamente.

Uma outra rotina que foi implementada no SCACC apresenta as características extraídas e medidas dos cromossomos, como mostrado na janela da Figura 6.16. Nesta rotina, o usuário pode fazer a análise individual dos cromossomos, selecionando-os através de um menu denominado de *Número do Objeto* (no exemplo da Figura 6.16 foi selecionado o objeto número 2 que corresponde a um cromossomo do par número 11) e comparar com todos os outros cromossomos detectados na metáfase. Logo, escolhendo-se um cromossomo para análise, o sistema realiza o cálculo do fator de semelhança desse cromossomo selecionado com todos os demais. Disponibilizou-se, nesta janela, uma listagem com os resultados obtidos do cálculo do fator de semelhança, por ordem decrescente de valor e de cima para baixo. Ao lado dessa lista, é apresentada uma outra listagem, denominada de *Objeto Homólogo*, em que o sistema apresenta o cromossomo correspondente ao fator de semelhança calculado, sendo que o cromossomo apresentado no topo da lista é o que possui mais semelhança com o cromossomo selecionado. Pode-se fazer uma análise comparativa da morfologia do cromossomo selecionado com todos os

demais e também ter uma representação visual do padrão de bandas, pelo gráfico de intensidade de níveis de cinza extraído dos pontos do esqueleto.

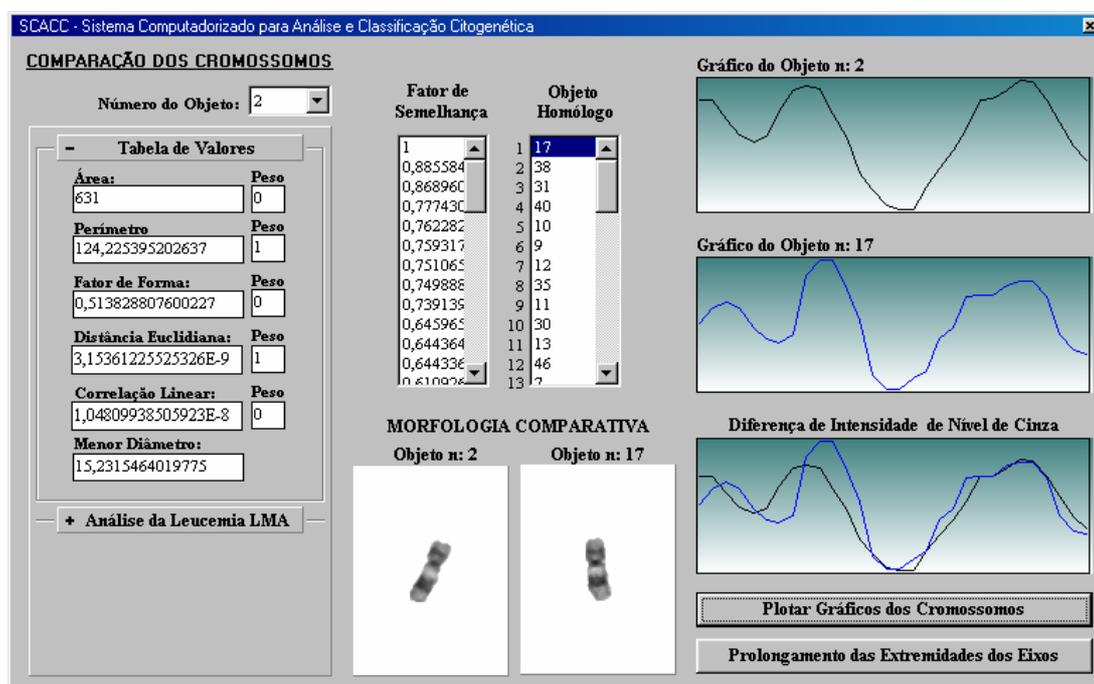


Figura 6.16 – Janela que permite fazer a análise individual dos cromossomos, comparando-os com todos outros cromossomos detectados na metáfase. Também é apresentado o padrão de bandas cromossômicas pelo gráfico de intensidade de nível de cinza.

A idéia da implementação dessa rotina se justifica pela possibilidade do usuário interagir no sistema, caso haja erros de classificação de algum cromossomo. Assim, selecionando o cromossomo que foi classificado errado, o usuário pode ter a opção de analisar o segundo maior fator de semelhança que aparece na listagem e, automaticamente, os cromossomos serão novamente ordenados corrigindo a classificação e o cariógrama.

Ainda na janela da Figura 6.16, foi desenvolvida uma rotina para servir de ferramenta ao citogeneticista na detecção da leucemia pró-mielocítica aguda, como mostrado na Figura 6.17. Esta ferramenta permite uma redução na subjetividade da análise desta doença, pois é disponibilizado pelo sistema um indicador da probabilidade de ocorrência da LMA e também gráficos representativos da intensidade de nível de cinza dos cromossomos, permitindo uma visualização comparativa do padrão de bandas dos mesmos. Na aba denominada de *Análise da Leucemia LMA*, é selecionado o par de cromossomos

número 17, no qual o sistema faz a análise da LMA. O citogeneticista poderá intervir no sistema alterando três valores: o valor da porcentagem do tamanho do segmento em relação ao tamanho normal do cromossomo do par número 17, a quantidade de pontos finais utilizados para extrair a média de intensidade de nível de cinza e também a faixa do valor máximo de intensidade de nível de cinza dos pontos finais dos dois cromossomos. Para a detecção dessa doença, é necessário também que se faça a extrapolação dos pontos do esqueleto, isto é feito utilizando-se o botão *Prolongamento das Extremidades do Eixo*. Plotando-se o gráfico de intensidade de nível de cinza dos esqueletos com as extremidades prolongadas, torna-se, então, nítida a visualização da diferença de tamanho, causado pela perda do segmento em um dos cromossomos, e da diferença de intensidade de nível de cinza no final do gráfico, pois o cromossomo normal tem as extremidades dos cromossomos terminadas em regiões claras, enquanto que o cromossomo que perde o segmento transladado possui a extremidade finalizada em uma região escura (banda). No botão do *Cálculo da Semelhança*, é apresentada a probabilidade da existência da doença LMA, baseando-se no cálculo do fator de semelhança. Utiliza-se, para este cálculo, os valores apresentados pelo citogeneticista e as características extraídas dos cromossomos.

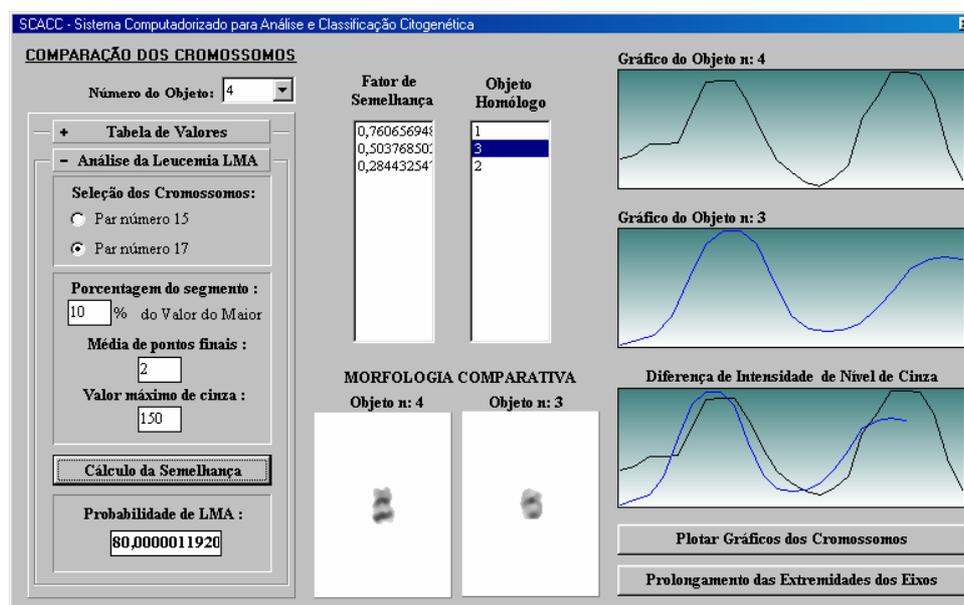


Figura 6.17 – Detecção da leucemia pró-mielocítica aguda, disponibilizando um indicador do grau de semelhança com a doença.

Reforçando o que foi mencionado, através da ferramenta mostrada na Figura 6.17, é disponibilizado ao citogeneticista não somente um indicador de similaridade com a doença, mas também uma representação visual das características das bandas dos cromossomos, que poderá auxiliar o profissional na detecção e avaliação do rearranjo mais eficiente da leucemia pró-mielocítica aguda.

### **6.3 – Considerações Finais**

Em síntese, o objetivo deste capítulo foi o de apresentar a descrição passo a passo do desenvolvimento da sistemática, baseando-se em uma ferramenta computacional, que visa propor uma nova metodologia de análise e classificação citogenética, fundamentando-se no enfoque lógico-combinatório da teoria de reconhecimento de padrões e nas técnicas de processamento digital de imagens. Os tópicos que contemplam a metodologia proposta foram: descrição do método de pesquisa, formulação geral do problema, formalização do problema e, finalmente, o modo de solução do problema expresso em termos matemáticos.

## **CAPÍTULO VII**

### **RESULTADOS OBTIDOS**

#### **7.1 – Introdução**

Uma etapa fundamental no desenvolvimento de novas metodologias confere a aplicação do modelo proposto neste trabalho em experimentos reais. Isso porque os resultados obtidos nestes experimentos precisam ser analisados pelos especialistas na matéria para a verificação da possibilidade de validação da nova metodologia desenvolvida. Portanto, o objetivo deste capítulo dispõe da aplicação em simulações do novo modelo computacional desenvolvido, o *Sistema Computadorizado para Análise e Classificação Citogenética* (SCACC), da apresentação dos resultados obtidos com estas simulações, da análise e interpretação destes resultados e, por fim, das considerações finais deste capítulo.

#### **7.2 – Aplicação do Modelo Proposto**

As etapas da aplicação do modelo proposto são apresentadas através dos seguintes tópicos: *Características das Imagens Adquiridas e Procedimentos Experimentais*.

##### **7.2.1 – Características das Imagens Adquiridas**

Várias imagens de células metafásicas foram adquiridas, com o intuito de compor o banco de dados a ser utilizado nas simulações com o SCACC. Estas imagens foram capturadas pelos seguintes dispositivos: um *fotomicroscópio*, uma *câmera de vídeo digital colorida CCD*, uma *placa de captura* e um *microcomputador*.

- *Fotomicroscópio óptico*: com iluminação halógena, da marca *Olympus* e modelo *BX41*. O primeiro dispositivo de aumento da imagem inclui uma objetiva de imersão Plan, da marca *Olympus*, com 100 vezes de magnitude com 1,25 de abertura numérica. O segundo dispositivo de aumento constitui uma fotoculuar PE de 3,3 vezes 125 *Olympus*, inserida no adaptador *C-Mount* da câmera de vídeo U-PMTVC da mesma marca.
- *Câmera de vídeo digital colorida CCD*: marca Sony Exwave HAD, modelo SSC-DC54A. Esta câmera apresenta alta sensibilidade (mínimo de 0.8 Lux de iluminação), resolução de até 752 (horizontal) X 582 (vertical) linhas, e relação sinal/ruído do vídeo de 50dB ou mais.
- *Placa de captura*: é uma placa PROLINK, para slot PCI e modelo PV-BT878P.
- *Microcomputador*: possui um processador *Intel Pentium III* – 750 MHz e 128 MB de memória RAM. O sistema de saída possui um monitor de vídeo 15", de tecnologia CRT (*Cathode Ray Tubes*), capaz de exibir as imagens dos cromossomos e a tela do kariograma final com resolução de até 1.024 x 768 pixels e 32 *bits* de cores distintas. Para o armazenamento das imagens, tem-se disponível 20 GB de memória de massa.

Além desses dispositivos, o sistema completo de captura consta ainda da rotina de aquisição de imagens implementada no SCACC, que permite que as imagens provenientes do microscópio sejam analisadas, selecionadas e armazenadas no computador para compor o banco de dados utilizado neste trabalho. As imagens adquiridas com este sistema são imagens reais e foram padronizadas, nesta pesquisa, com as seguintes características:

- resolução da imagem: 640x480 *pixels*;
- tamanho do arquivo: 0,901 *Mbytes*;
- extensão do arquivo: *bmp (Bitmap)*;
- padrão de cores: modelo RGB – com 24 *bits*;
- filtro de cor: verde escuro;

A Figura 7.1 mostra uma das imagens de células metafásicas que compõem o banco de dados do SCACC, com as características descritas anteriormente e escala métrica comparativa (barra de escala 10 $\mu$ m).



Figura 7.1 – Imagem de uma célula metafásica adquirida pelo sistema de aquisição do SCACC, que compõe o banco de dados utilizado nas simulações. Barra de escala 10  $\mu$ m.

As imagens das células metafásicas adquiridas passaram por uma triagem, selecionando-se apenas as dos indivíduos femininos, que não apresentavam sobreposições nem anormalidades estruturais e numéricas. Entretanto, as lâminas contendo as células metafásicas, que foram extraídas tais imagens, tiveram procedência de dois laboratórios de citogenética distintos, cada qual com processos de cultura e preparação das lâminas diferentes. Um destes é o Laboratório de Citogenética do Departamento de Biologia Celular, Embriologia e Genética da Universidade Federal de Santa Catarina (BEG/UFSC), e o outro é a Clínica de Genética Humana *NeuroGene*, de Florianópolis. A Tabela 7.1 apresenta os tipos de imagens que compõem o banco de dados adquirido para este trabalho, indicando também a procedência das lâminas das células metafásicas utilizadas para a captura das imagens.

Tabela 7.1 – Tipos de imagens que compõem o banco de imagens deste trabalho e a procedência das lâminas das células metafásicas utilizadas para a captura das imagens.

<b>Imagens</b>	<b>Lâminas - Laboratórios</b>
12 Imagens de Células Metafásicas de Indivíduos Normais Femininos ( <i>46 Cromossomos</i> ).	BEG/UFSC e <i>NeuroGene</i>
10 Imagens, cada qual contendo o par de cromossomos de número 17 de indivíduos normais.	BEG/UFSC e <i>NeuroGene</i>
10 Imagens, cada qual contendo o par de cromossomos de número 15 de indivíduos normais.	BEG/UFSC e <i>NeuroGene</i>
10 Imagens, cada qual contendo o par de cromossomos de número 17 de indivíduos com Leucemia Pró-Mielocítica Aguda.	<i>NeuroGene</i>
10 Imagens, cada qual contendo o par de cromossomos de número 15 de indivíduos com Leucemia Pró-Mielocítica Aguda.	<i>NeuroGene</i>

As lâminas preparadas no BEG/UFSC apresentavam nível de resolução de cariótipo na faixa de 350 a 450 bandas, já as lâminas preparadas na Clínica *NeuroGene* apresentavam nível de resolução de cariótipo na faixa de 350 a 400 bandas (esses dados são aproximados).

### 7.2.2 – Procedimento Experimental

A série de experimentos ocorreu em três fases: *reconhecimento automático dos pares de cromossomos, composição automática do cariótipo e detecção da Leucemia Pró-Mielocítica Aguda*.

A primeira fase compreende os testes para verificação do desempenho do novo modelo computacional, quanto ao reconhecimento automático dos cromossomos homólogos. Foram utilizadas, para estes testes, as quatro variáveis extraídas e medidas pelo SCACC: a área, o fator de forma, a distância *Euclidiana* e a correlação linear de *Pearson*. Essas variáveis foram selecionadas por uma rotina do SCACC, ilustrada pela Figura 6.16 do capítulo 6 deste trabalho, e os cálculos foram feitos baseando-se na comparação entre os cromossomos através da seleção individual dos mesmos. A partir de um cromossomo selecionado pelo usuário, foram realizadas as comparações com todos os outros cromossomos da mesma célula, com a finalidade de encontrar seu homólogo. O

sistema foi desenvolvido para apresentar, em ordem decrescente, os cromossomos mais semelhantes ao cromossomo selecionado e a rotina possibilita ao usuário verificar em que ordem de classificação o sistema conseguiu identificar o par. Sendo assim, o primeiro cromossomo que aparece na lista é caracterizado pelo SCACC como o mais semelhante ao cromossomo selecionado e, se estes dois cromossomos formam o par, então, a identificação é considerada de 1ª ordem. Entretanto, se o primeiro cromossomo da lista não for o homólogo do cromossomo selecionado e este for o segundo que aparece na lista, a identificação é considerada de 2ª ordem e assim sucessivamente.

Para os experimentos dessa primeira fase, optou-se em simular o reconhecimento automático dos pares homólogos utilizando-se, de início, as quatro variáveis separadamente. Assim, foram realizados quatro experimentos, um para cada variável extraída e medida, aplicados em 276 cromossomos (138 pares homólogos). Em seguida, foram realizados mais três experimentos, utilizando-se as variáveis em conjunto, porém diferenciando-as por pesos (ponderações). No 5º (quinto) experimento, foi feito o seguinte arranjo das variáveis e seus respectivos pesos: área (peso 1), fator de forma (peso 1) e distância *Euclidiana* (peso 1). Para o 6º (sexto) experimento, área (peso 2), fator de forma (peso 1) e distância *Euclidiana* (peso 3). E, no 7º (sétimo) experimento, área (peso 2), fator de forma (peso 1), distância *Euclidiana* (peso 3) e correlação linear de *Pearson* (peso 1). A quantidade de cromossomos testados foi a mesma em relação aos quatro primeiros experimentos.

Para as simulações da segunda fase, isto é, para a verificação do desempenho do SCACC quanto à composição automática do cariótipo, foi realizado o experimento de número 8 (oito), o qual foram utilizadas as mesmas variáveis e os mesmos pesos adotados em relação ao 7º (sétimo) experimento. Nesta fase, foram utilizadas nas simulações 12 metáfases de indivíduos normais do sexo feminino.

Por último, foram realizadas as simulações para a detecção da Leucemia Pró-Mielocítica Aguda, onde foram utilizadas nos testes 10 imagens de cromossomos do par número 17 de células leucêmicas. Em comparação com células normais, foram utilizadas 10 imagens de cromossomos do par número 15 e 10 imagens do par número 17 destas células. Nessas simulações, o objetivo era que o sistema apresentasse, para cada imagem testada, um indicador da alteração cromossômica, considerando, para todas as imagens

testadas, o tamanho do segmento translocado com o valor de 20% do tamanho do cromossomo normal e a faixa de diferença de intensidade de nível de cinza igual a 100.

### 7.3 – Resultados Obtidos

Os resultados obtidos com os quatro primeiros experimentos, da fase de reconhecimento automático dos pares homólogos, são mostrados na Tabela 7.2. Nesta Tabela aparecem as quatro variáveis área ( $A$ ), fator de forma ( $FF$ ), distância *Euclidiana* ( $De$ ) e correlação linear de *Pearson* ( $CLP$ ) utilizadas para o 1º, 2º, 3º e o 4º experimentos, sendo que os resultados são apresentados na seqüência por ordem de identificação dos homólogos pelo SCACC. Para cada simulação, foram testados 276 cromossomos (138 pares homólogos).

Tabela 7.2 – Resultados obtidos com o 1º, 2º, 3º e o 4º experimento, utilizando-se as variáveis: área ( $A$ ), fator de forma ( $FF$ ), distância *Euclidiana* ( $De$ ) e correlação linear de *Pearson* ( $CLP$ ). São apresentados os testes com 276 cromossomos e a ordem de identificação dos homólogos pelo SCACC.

Experimentos e Variáveis Utilizadas	Número de Cromossomos Reconhecidos por Ordem de Identificação										
	1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	10ª	Acima
1º - A	102	43	35	30	18	20	17	5	1	3	2
2º - FF	82	30	36	17	20	25	18	6	10	15	17
3º - De	139	41	23	22	10	16	11	7	5	2	0
4º - CLP	85	37	12	33	21	10	7	7	12	10	42

Pode-se constatar que a variável distância *Euclidiana* e a área (3º e 1º experimentos, respectivamente) apresentaram, em geral, desempenhos melhores, seguidas pela correlação linear de *Pearson* (4º experimento) e o fator de forma (2º experimento). A correlação linear de *Pearson* apresentou resultado muito próximo em relação à variável fator de forma na identificação de 1ª ordem, entretanto, acima da décima ordem esta variável apresentou um número significativo de cromossomos, o que pode acarretar uma certa dificuldade na identificação automática dos pares homólogos. Para melhor análise desses experimentos, o gráfico apresentado pela Figura 7.2 ilustra a comparação das curvas

dos resultados obtidos para cada variável testada. Neste gráfico, são apresentados os resultados até a décima nona ordem (19ª ordem) de identificação, sendo que as identificações que ocorreram acima desta ordem foram consideradas para a vigésima ordem (20ª ordem).

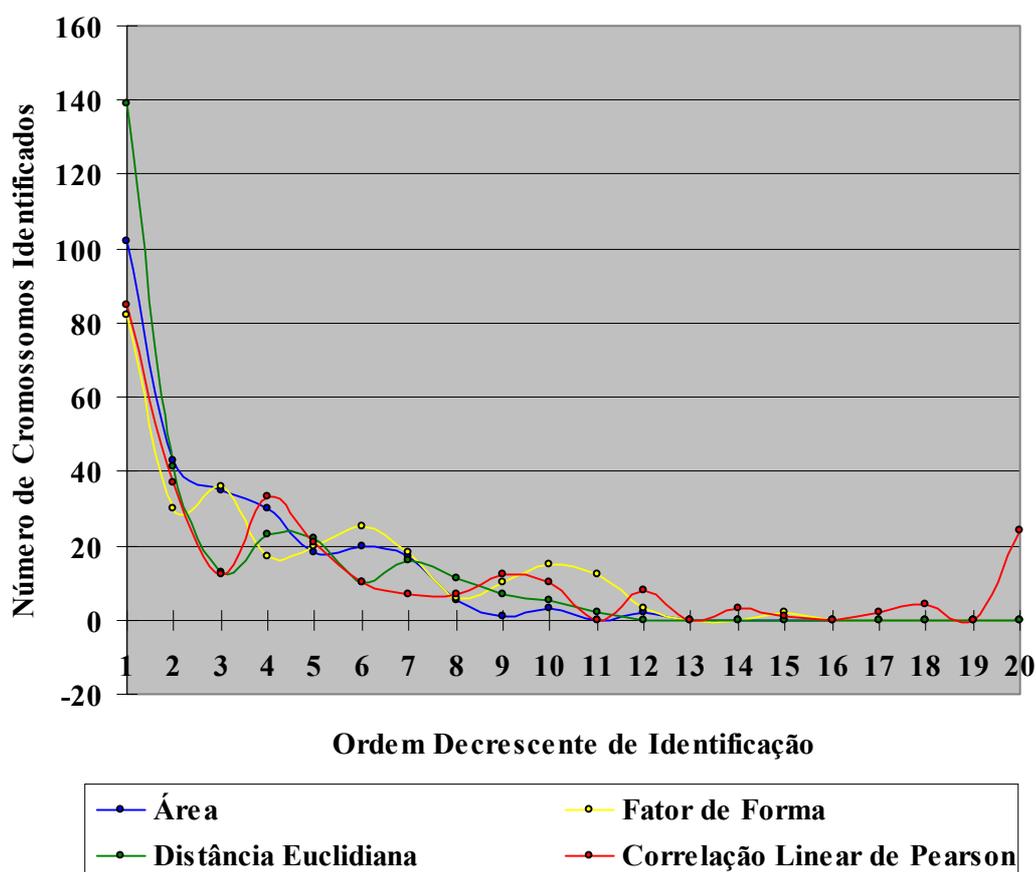


Figura 7.2 – Gráfico comparativo dos resultados obtidos dos experimentos com as variáveis área, fator de forma, distância *Euclidiana* e correlação linear de *Pearson*, testadas separadamente. Para cada experimento foram utilizados 276 cromossomos (138 pares homólogos).

Os experimentos, utilizando as variáveis separadamente, são importantes para que se possa identificar aquelas variáveis que apresentam melhores desempenhos no reconhecimento automático dos cromossomos, demonstrando, portanto, que são menos suscetíveis quanto às fontes de incerteza no domínio do problema. Logo, nos três experimentos com o uso das características em conjunto, pode-se, então, atribuir pesos maiores no cálculo do fator de semelhança para as características que apresentaram

desempenhos melhores. Dessa forma, explica-se a opção dos pesos diferenciados nos três experimentos seguintes realizados para o reconhecimento automático dos pares de cromossomos. A Tabela 7.3 apresenta os resultados obtidos com o 5<sup>o</sup>, 6<sup>o</sup> e 7<sup>o</sup> (quinto, sexto e sétimo) experimentos realizados para os conjuntos das variáveis utilizadas. Para o 5<sup>o</sup> (quinto) experimento, utilizou-se em conjunto a área (*A*) com peso 1, fator de forma (*FF*) com peso 1 e distância *Euclidiana* (*De*) também com peso 1. Para o 6<sup>o</sup> (sexto) experimento, utilizou-se a área (*A*) com peso 2, fator de forma (*FF*) com peso 1 e distância *Euclidiana* (*De*) com peso 3. E o 7<sup>o</sup> (sétimo) experimento, utilizou-se a área (*A*) com peso 2, fator de forma (*FF*) com peso 1, distância *Euclidiana* (*De*) com peso 3 e correlação linear de *Pearson* (*CLP*) com peso 1. Em seguida, é mostrada a ordem em que o SCACC conseguiu encontrar os cromossomos homólogos. Foram utilizados 276 cromossomos (138 pares homólogos) para cada experimento.

Tabela 7.3 – Resultados obtidos com o 5<sup>o</sup>, 6<sup>o</sup> e o 7<sup>o</sup> experimento, utilizando-se as variáveis, área (*A*), fator de forma (*FF*), distância *Euclidiana* (*De*) e correlação linear de *Pearson* (*CLP*), e seus pesos adotados. São apresentados os testes com 276 cromossomos e a ordem de identificação dos homólogos pelo SCACC.

Experimentos e Variáveis Utilizadas	Pesos	Número de Cromossomos Simulados e sua Ordem de Identificação								
		1 <sup>a</sup>	2 <sup>a</sup>	3 <sup>a</sup>	4 <sup>a</sup>	5 <sup>a</sup>	6 <sup>a</sup>	7 <sup>a</sup>	8 <sup>a</sup>	Acima
5 <sup>o</sup> - A, FF, De	1, 1, 1	127	53	29	24	17	8	6	3	9
6 <sup>o</sup> - A, FF, De	2, 1, 3	161	38	16	17	7	10	6	7	14
7 <sup>o</sup> - A, FF, De, CLP	2,1,3,1	174	31	12	5	10	6	1	3	34

Pode-se constatar que o 7<sup>o</sup> (sétimo) experimento com a variável área (2), fator de forma (1), distância *Euclidiana* (3) e correlação linear de *Pearson* (1), apresentou-se com desempenho um pouco melhor do que o 6<sup>o</sup> (sétimo) experimento, mesmo apresentando um número significativo de cromossomos identificados acima da 8<sup>a</sup> ordem de identificação. Logo, este conjunto de variáveis foi, no geral, mais eficiente para o reconhecimento automático dos pares homólogos. Para melhor análise, a Figura 7.3 ilustra a comparação das curvas dos resultados obtidos para o 5<sup>o</sup>, 6<sup>o</sup> e 7<sup>o</sup> (quinto, sexto e sétimo) experimentos. Neste gráfico, como no caso anterior, são apresentados os resultados até a décima nona ordem de identificação (19<sup>a</sup> ordem), sendo que as identificações que ocorreram acima desta ordem foram consideradas para a vigésima ordem (20<sup>a</sup> ordem).

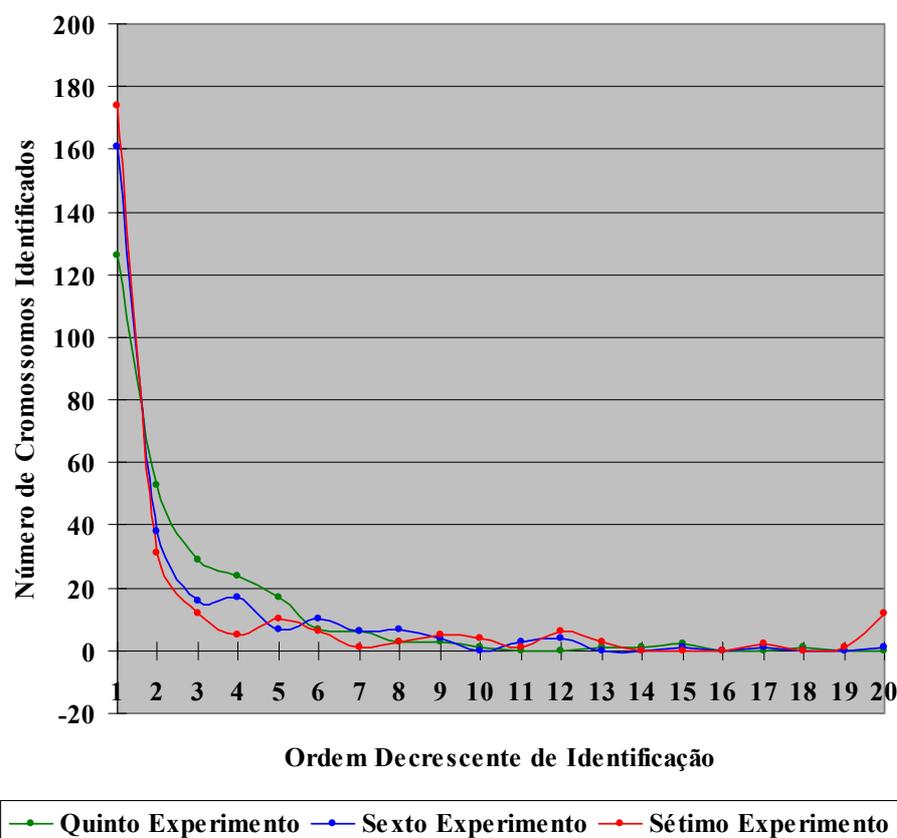


Figura 7.3 – Gráfico comparativo dos resultados obtidos com o 5º, 6º e 7º (quinto, sexto e sétimo) experimentos. Para cada experimento foram utilizados 276 cromossomos (138 pares homólogos).

Estas três simulações, com as variáveis em conjunto e com os pesos diferenciados, torna-se a base para as simulações subseqüentes, onde é verificado o desempenho geral do SCACC para a composição automática do cariótipo (segunda fase). Nesta segunda fase dos experimentos, foi utilizado o conjunto de variáveis que apresentou melhores resultados para o reconhecimento dos pares homólogos, isto é, o arranjo das variáveis e dos pesos que constituem o 7º (sétimo) experimento, como mostrado na Tabela 7.4. Nesta tabela, é apresentado o 8º experimento, com suas variáveis utilizadas, seus pesos adotados e o desempenho geral do SCACC com o número de cromossomos identificados corretamente e o número de cromossomos que não foram possíveis de serem classificados corretamente em suas classes. Nestas simulações, foram utilizadas as 12 metáfases de indivíduos do sexo feminino normais.

Tabela 7.4 – Resultados obtidos com o 8<sup>o</sup> (oitavo) experimento, com as suas variáveis utilizadas: área (*A*), fator de forma (*FF*), distância *Euclidiana* (*De*) e correlação linear de *Pearson* (*CLP*), seus pesos adotados, e o desempenho geral do SCACC com o número de cromossomos identificados corretamente e o número de cromossomos que não foram classificados corretamente.

<b>Características Utilizadas</b>	<b>Pesos</b>	<b>Metáfases Utilizadas</b>	<b>Cromossomos Classificados Corretamente</b>	<b>Cromossomos não Classificados Corretamente</b>
8 <sup>o</sup> – A, FF, De, CLP	2, 1, 3, 1	12 metáfases (552 cromossomos)	279	273

E, para a última fase dos experimentos, foram realizadas as simulações para a detecção da Leucemia Pró-Mielocítica Aguda (LMA), como mostrado na Tabela 7.5. Os resultados, para cada imagem testada, foram apresentados como indicadores percentuais da doença, sendo que os valores próximos ou iguais a 100% (cem) caracterizam a presença da leucemia e, os valores próximos ou iguais a 0% (zero), caracterizam indivíduos normais. Considerou-se, nos testes, o tamanho do segmento translocado com o valor de 20% do tamanho do cromossomo normal e a faixa de diferença de intensidade de nível de cinza em 100. As primeiras simulações para a detecção da leucemia foram realizadas com 10 imagens cromossomos com o par número 17 de células leucêmicas.

Tabela 7.5 – Resultados obtidos com as simulações para a detecção da Leucemia Pró-Mielocítica Aguda (LMA), apresentados como indicadores da doença. E as 10 imagens testadas foram de células leucêmicas.

<b>Imagens Testadas</b>	<b>Indicador da LMA - %</b>
Imagem 01	97,11
Imagem 02	99,70
Imagem 03	62,55
Imagem 04	70,73
Imagem 05	63,73
Imagem 06	99,29
Imagem 07	81,46
Imagem 08	65,84
Imagem 09	98,95
Imagem 10	91,50

Em seguida, foram realizados os mesmos testes agora com 10 imagens do par número 17 de células normais. Os resultados dos indicadores percentuais são mostrados na Tabela 7.6.

Tabela 7.6 – Resultados obtidos com as simulações para a detecção da Leucemia Pró-Mielocítica Aguda (LMA), apresentados como indicadores da doença e testadas em 10 imagens células normais.

<b>Imagens Testadas</b>	<b>Indicador da LMA - %</b>
Imagem 11	8,54
Imagem 12	7,59
Imagem 13	6,00
Imagem 14	13,64
Imagem 15	8,57
Imagem 16	15,85
Imagem 17	17,30
Imagem 18	15,54
Imagem 19	15,65
Imagem 20	13,37

Para a análise comparativa dos resultados obtidos com as simulações para a detecção da Leucemia Pró-Mielocítica Aguda (LMA), a Figura 7.4 apresenta um gráfico com os valores calculados da probabilidade da doença mostrados nas Tabelas 7.5 e 7.6. O eixo horizontal indica as imagens testadas, sendo que as 10 primeiras são as imagens dos pares de cromossomos número 17 de células leucêmicas e as imagens restantes de pares de cromossomos número 17 de células normais.

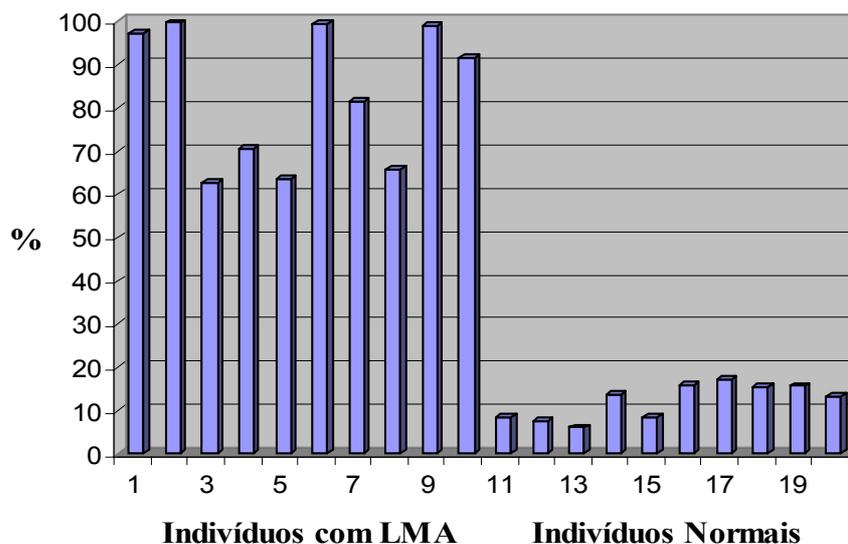


Figura 7.4 – Gráfico comparativo dos resultados obtidos com as simulações realizadas para a detecção da Leucemia Pró-Mielocítica Aguda (LMA). O eixo horizontal representa as imagens testadas, sendo que as 10 primeiras imagens são de células leucêmicas e as imagens restantes de células normais.

É importante mencionar também que foram realizadas simulações com imagens de cromossomos do par número 15, entretanto, devido ao tamanho desses cromossomos ser maior do que os cromossomos 17, em alguns casos o segmento translocado não representava um aumento considerável no comprimento dos cromossomos. Neste sentido, era comum a ocorrência de cromossomos do par número 15, de comprimento normal, apresentar-se praticamente do mesmo tamanho que o seu homólogo com o segmento translocado, o que dificultava a detecção precisa da doença utilizando-se esses cromossomos. Dessa forma, optou-se apenas pela realização da detecção da leucemia em cromossomos do par número 17.

#### 7.4 – Análise e Interpretação dos Resultados Obtidos

A análise e a interpretação dos resultados obtidos utilizando-se o modelo computacional proposto é a "tradução" da linguagem matemática para a linguagem do

especialista, de forma contrária ao que se fez anteriormente neste capítulo. Este estágio é importante porque serve de base para comprovação do modelo computacional, ou seja, se este modelo soluciona o problema formulado.

Portanto, de acordo com os resultados obtidos na primeira fase dos experimentos, para o reconhecimento automático dos pares homólogos, pode-se observar que no 1º experimento, utilizando-se a variável área, os resultados da identificação de 1ª ordem em geral foram satisfatórios, e muito bom para a identificação dos cromossomos grandes e pequenos (grupos A, B, E, F e G), porém, o uso apenas dessa variável dificultou a classificação dos cromossomos médios e relativamente pequenos (grupo C e D), pois o tamanho dos cromossomos destes grupos são muito próximos. Essa variável também apresentou muita dificuldade na identificação dos cromossomos homólogos quando na ocorrência de diferença dos tamanhos dos cromossomos causada pela forma de espalhamento aleatório na lâmina, isto é, os cromossomos que se encontram mais na periferia da metáfase são maiores e com coloração mais clara. Outro fator que impossibilitou a identificação precisa pela área é quando os cromossomos aparecem com as cromátides separadas, isso resultou em medidas muito diferentes em relação a seus cromossomos homólogos. A Figura 7.5 mostra dois cromossomos do par número 9 apresentando a diferença de tamanhos devido ao fator de espalhamento aleatório na lâmina e também porque um deles possui as cromátides levemente separadas.

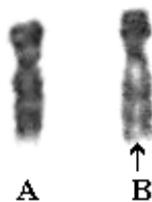


Figura 7.5 – Ilustração de dois cromossomos do par número 9 que apresentam fatores que dificultam sua identificação: (A) apresenta tamanho menor e (B) possui as cromátides levemente separadas.

O método de medir o tamanho dos cromossomos pela área, apesar dos fatores que dificultam a correta identificação dos homólogos, acredita-se que foi a melhor alternativa

comparada à extração do comprimento dos cromossomos pelo eixo médio. Nesta extração, além de estar sujeito ao fator de espalhamento aleatório dos cromossomos na lâmina, este processo também apresenta imprecisão quanto às torções severas de cromossomos quando comparado com o seu homólogo apresentando-se normal (reto). Independentemente do uso desses dois métodos, a variável área ainda é indispensável para a identificação dos cromossomos.

Para as simulações com a variável fator de forma (2<sup>o</sup> experimento), houve grande dificuldade quando na ocorrência de torções ou irregularidades no contorno dos objetos, além de falhar na identificação dos cromossomos médios e relativamente pequenos (grupo C e D). Nos demais grupos, a classificação de 1<sup>a</sup> ordem foi satisfatória.

Como pode ser constatado, o reconhecimento dos cromossomos utilizando-se a área e o fator de forma apresentou maior dificuldade na identificação dos cromossomos médios e relativamente pequenos (grupo C e D). Na tentativa de superar este problema, utilizou-se também a variável distância *Euclidiana* (3<sup>o</sup> experimento) para que fosse possível a identificação dos cromossomos por intermédio das características das bandas. Em cromossomos que apresentavam cromátides separadas, ficava impossibilitada a identificação, pois os cálculos para essa variável dependem da extração do eixo médio e da intensidade de nível de cinza ao longo dos pontos do esqueleto. Assim, quando as cromátides apresentam-se separadas, o eixo médio passa entre as mesmas não sendo possível adquirir as informações das bandas. Além disso, os dois parâmetros extraídos, eixo médio e intensidade de nível de cinza, apresentam-se muito suscetíveis quanto aos problemas causados pelas fontes de incerteza. Logo, quando não havia a influência desses problemas, os resultados com a identificação apresentavam-se muito bons, mas o reconhecimento dos cromossomos homólogos ficava comprometido quando ocorria grande diferença nos tamanhos dos cromossomos e na intensidade de nível de cinza (cromossomo mais claro que o homólogo). Para ilustrar um exemplo da ocorrência dos problemas de diferença de tamanho e diferença de intensidade de nível de cinza, a Figura 7.6 mostra a comparação de dois cromossomos do par número 7, o objeto selecionado número 31 na figura, menor e mais claro, e seu homólogo (objeto homólogo número 46 na figura). Eles somente resultaram na identificação de 15<sup>a</sup> ordem.

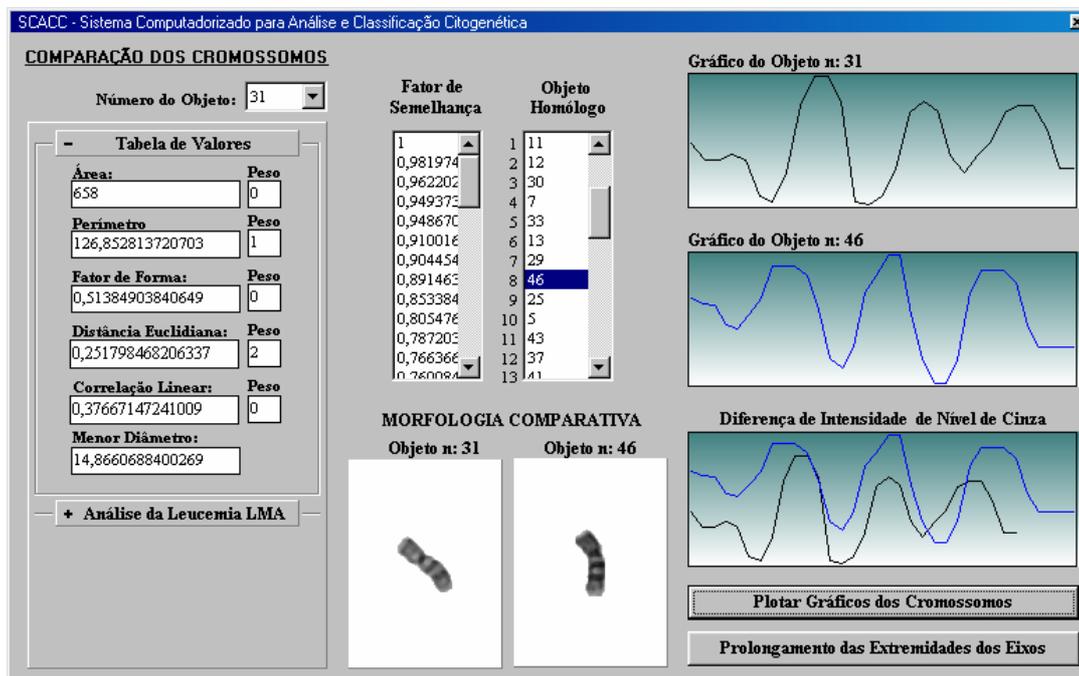


Figura 7.6 – Identificação dos cromossomos homólogos do par número 7 utilizando-se a variável distância *Euclidiana*, quando na ocorrência de diferença de tamanho e intensidade de nível de cinza entre os homólogos. A identificação ficou na 15ª ordem.

Nos casos como no exemplo da Figura 7.6, em que o cromossomo selecionado do par número 7, apresenta-se menor e mais claro, os demais cromossomos tendem a ficarem mais semelhantes a este cromossomo que sofreu influência com os problemas referentes às fontes de incerteza, como mostra a Figura 7.7. Pode-se observar que o cromossomo do par número 7 selecionado (objeto denominado de 31 nessa figura) é mais semelhante ao cromossomo do par número 9 (objeto homólogo número 11 nessa figura) do que o seu homólogo (objeto homólogo número 46 nessa figura). Isto pode ser comprovado também observando-se os gráficos de intensidade de nível de cinza mostrados nas Figura 7.6 e 7.7.

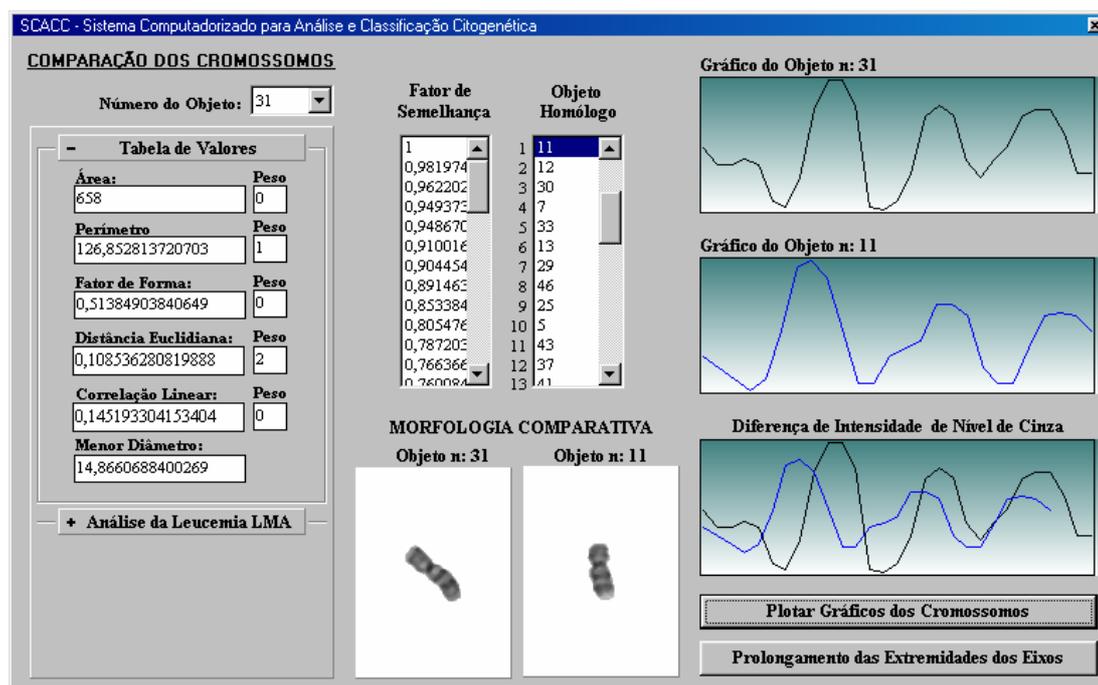


Figura 7.7 – Comparação de um dos cromossomos do par 7 (objeto número 31 nessa figura) com outro cromossomo do par número 9 (objeto homólogo número 11 nessa figura). Devido o cromossomo do par número 7 apresentar-se menor e mais claro, torna-se menos semelhante ao seu homólogo (objeto homólogo número 46 nessa figura).

Além da dificuldade quanto à diferença de tamanho e a diferença de intensidade de nível de cinza, é importante mencionar também que os resultados da classificação dos pares homólogos, utilizando-se a variável distância *Euclidiana*, são melhores em imagens de células metafásicas com nível de resolução do cariótipo maior (cromossomos mais longos) e com bandas mais nítidas (mais definidas). O nível de resolução do cariótipo das imagens utilizadas apresentava-se em uma faixa intermediária, e com bandas pouco definidas.

Considerando as fontes de incerteza, as simulações, utilizando a variável medida de correlação linear de *Pearson* (4<sup>o</sup> experimento), tiveram os mesmos problemas apresentados quando foi utilizada a variável distância *Euclidiana*, sendo que as maiores deficiências foram notadas nos pares de cromossomos com grandes diferenças de tamanho (comprimento dos eixos). Esta variável apresentou muita eficiência na identificação dos cromossomos pequenos, já para os cromossomos médios e grandes o método não foi muito eficiente. Também foi notado, no uso dessa variável, que na maioria dos cromossomos simulados, quando estes não eram classificados na 1<sup>a</sup>, 2<sup>a</sup> ou 3<sup>a</sup> ordem, a identificação

ficava acima da 20<sup>a</sup> ordem. Quando foi utilizada a variável distância *Euclidiana*, nas simulações com imagens de cromossomos com bandas um pouco mais definidas, a variável correlação linear de Pearson apresentou uma grande melhoria nos resultados.

No 5<sup>o</sup>, 6<sup>o</sup> e 7<sup>o</sup> experimentos, onde foram utilizadas, as variáveis em conjunto, constatou-se que os melhores resultados foram obtidos no 7<sup>o</sup> experimento, por ter incluído no conjunto a variável correlação linear de *Pearson*. Considerando-se apenas a identificação de 1<sup>a</sup> ordem, o índice de acerto foi de aproximadamente de 63,04%. Comparando-se esse resultado com os resultados do 6<sup>o</sup> experimento pode-se constatar que a melhoria obtida com o 7<sup>o</sup> experimento, apesar de ser muito pequena no que diz respeito à identificação de 1<sup>a</sup> ordem, ela foi mais expressiva nas ordens subseqüentes.

Utilizando-se o mesmo conjunto de variáveis do 7<sup>o</sup> experimento para a simulação da segunda fase, isto é, para a verificação do desempenho do SCACC quanto à composição automática do cariótipo, foi realizado o 8<sup>o</sup> experimento, que apresentou desempenho de aproximadamente 50% de índice acerto. Este índice baixo decorreu do fato da composição automática do cariótipo não ser muito eficiente para os cromossomos sexuais do par X, desencadeando uma série de erros nos demais cromossomos médios (grupo C). Uma forma de tornar mais eficiente a composição automática dos cromossomos seria a identificação de uma ou mais características que pudessem descrever todas as 24 classes de cromossomos. Logo, além da classificação dos pares homólogos, seria necessário também a classificação das 24 classes possíveis de cromossomos por características pré-definidas destas classes, ao invés de classificá-los por ordem decrescente de tamanho. Algumas destas características seriam: a posição das bandas principais ao longo dos braços dos cromossomos, a largura dessas bandas, a intensidade, etc. Somente dessa forma seria possível também a classificação de células de indivíduos do sexo masculino devido tais indivíduos apresentam o cromossomo Y sem homólogo. Contudo, isso não garantiria total eficiência, pois alguns pares de cromossomos do grupo C apresentam grandes similaridades quanto aos padrões de banda, necessitando ainda de outras características para descrevê-los.

Por último, foram realizadas as simulações para a detecção da Leucemia Pró-Mielocítica Aguda, tendo como objetivo extrair de cada imagem testada um indicador da doença (próximo de 100 seriam de imagens de células leucêmicas e próximos de 0 células normais). Nas simulações com as imagens de pacientes com leucemia pode ser constatado

que a porcentagem de acerto foi acima de 62,55% e, com as imagens normais, a porcentagem foi abaixo de 17,30%. O distanciamento dessas porcentagens caracteriza, então, que as peculiaridades extraídas dos cromossomos do par número 17 e, utilizadas para o cálculo dessas porcentagens, realmente comprovam que o método consegue identificar com eficiência o rearranjo cromossômico que caracteriza a doença.

### **7.5 – Considerações Finais**

O objetivo deste capítulo foi apresentar os resultados obtidos ao utilizar o *Sistema Computadorizado para Análise e Classificação Citogenética (SCACC)* e, por fim, descrever detalhadamente a análise e a interpretação desses resultados que irão servir de base para a verificação da validação desse novo sistema desenvolvido para a análise e classificação citogenética proposta.

## **CAPÍTULO VIII**

# **CONCLUSÕES, CONTRIBUIÇÕES E SUGESTÕES PARA FUTUROS TRABALHOS**

### **8.1 – Introdução**

Com a análise e a interpretação dos resultados obtidos, é possível ter uma avaliação entre a interpretação dos resultados e o problema formulado pelo especialista da área de aplicação. Assim, é verificado se o diagnóstico apresentado pelo modelo computacional é válido e se soluciona o problema formulado. Este capítulo apresenta as conclusões sobre o trabalho exposto, as limitações apresentadas pela nova metodologia, as contribuições deste trabalho e, por fim, as sugestões para futuros trabalhos.

### **8.2 – Conclusões**

Foi possível constatar, no transcorrer deste trabalho, que o desenvolvimento de novos métodos computacionais para análise e classificação citogenética ainda encontra muitas dificuldades para a realização da atividade de cariotipagem automática sem que haja qualquer tipo de intervenção humana.

Com o intuito de buscar soluções para reduzir e/ou eliminar estas dificuldades, foi proposto o desenvolvimento de uma nova sistemática, baseando-se em uma ferramenta computacional para análise e classificação citogenética, com dois propósitos distintos: tentar realizar a identificação automática dos cromossomos através de uma única cena e desenvolver uma nova metodologia que pudesse ser utilizada como ferramenta para análise mais confiável da anormalidade cromossômica da leucemia pró-mielocítica aguda.

No que diz respeito ao primeiro propósito desta nova sistemática, foi possível a elaboração de um novo método para o reconhecimento automático de cromossomos homólogos utilizando-se uma única metáfase, sem a necessidade da comparação com entidades específicas e previamente armazenadas. Para tanto, houve a necessidade da

aplicação de técnicas de processamento digital de imagens para a extração mais precisa das principais características que descrevem os cromossomos. Uma dessas técnicas baseava-se na extração do eixo médio dos cromossomos, que permitiu obter o esqueleto dos cromossomos com excelentes resultados de afinamento. A partir destes esqueletos, foi possível o cálculo de duas características para a descrição dos cromossomos, a distância *Euclidiana* e a correlação linear de *Person*. Essas duas características, juntamente com outras técnicas mais tradicionais, a área e o fator de forma, formaram o conjunto de características utilizadas para a identificação dos cromossomos. Nos experimentos realizados, o uso das características em conjunto resultou em melhores resultados de identificação do que com a aplicação dessas separadamente. Isso porque não é difícil a ocorrência de grandes variações nas características extraídas devido às várias fontes de incerteza mencionadas neste trabalho, o que dificulta ou impossibilita a identificação dos cromossomos por apenas uma única característica. Logo, se no processo de identificação dos cromossomos de uma determinada metáfase uma das características utilizada apresenta-se muito suscetível às fontes de incerteza, o reconhecimento automático ainda poderá ser assegurado por intermédio das demais características.

No início deste trabalho, foi comentado que uma das causas que poderiam gerar variações das características que descrevem os cromossomos diz respeito ao preparo das lâminas contendo as células metafásicas, sendo que este processo é muito difícil de se controlar para que não ocorra o surgimento de sobreposições, toque e torções severas. Entretanto, existem outros aspectos que podem ser controlados para que se eliminem ou minimizem tais fontes de incerteza como, por exemplo, o processo de cultura das células. Neste caso, consegue-se controlar a questão da nitidez das bandas através do tempo de coloração aplicado e o nível de resolução do cariótipo. À questão da nitidez das bandas, foi constatada que é fundamental para o cálculo da variável distância *Euclidiana* e correlação linear de *Pearson*, sendo que, nos experimentos realizados, os melhores desempenhos de identificação foram conseguidos em imagens com bandeamento mais nítido e nível de resolução de cariótipo maior.

No decorrer dos experimentos verificou-se, que o fato que mais gerou dificuldades para o SCACC identificar os cromossomos homólogos em 1ª ordem ocorria devido a grandes diferenças de tamanho e intensidade de nível de cinza entre os cromossomos homólogos, acarretando, portanto, em variações em todas as características

utilizadas. Para esses problemas, também existe um outro aspecto que pode ser controlado para que se tenha características mais fiéis na descrição dos cromossomos, ou então, que permita minimizar as variações nas características extraídas. Este aspecto depende da escolha de objetivas adequadas para a análise e classificação citogenética, pois objetivas com pouca tecnologia em correções, de nivelamento de campo (esféricas) e de cores, podem colocar em xeque toda a credibilidade de qualquer sistema computadorizado. Em microscópios para tal finalidade, o uso de objetivas planacromáticas ou planapocromáticas torna-se muito comum, justificando-as pela necessidade de uma magnitude superior a 1250 vezes para este tipo de estudo, ao invés somente das objetivas *plan* utilizadas neste trabalho.

Também é importante o comentário de que seria necessário ainda a análise de novas e melhores características que possam descrever os objetos de estudo com mais eficiência e que sejam menos suscetíveis às fontes de incerteza, buscando melhorar os índices de acerto de 1<sup>a</sup> ordem na classificação. Uma destas características diz respeito ao cálculo do índice centromérico, que depende da localização precisa da posição do centrômero.

Diante de todos esses fatos, com o uso de imagens de células metafásicas preparadas com nível de resolução do cariótipo maiores e com bandas mais nítidas, o uso de objetivas mais adequadas (planacromáticas ou planapocromáticas), além também da extração e do cálculo de novas características, acredita-se que, de modo geral, poder-se-ia obter melhorias significativas no desempenho da identificação automática dos cromossomos homólogos, através de uma única metáfase.

Já no contexto da aplicação do enfoque lógico-combinatório, o método apresentou-se muito eficiente para tal finalidade, pois permitiu a elaboração de uma função de analogia, considerando todas as variáveis (qualitativas e quantitativas) simultaneamente e adotando ponderações diferenciadas para a descrição do problema, da mesma forma como os citogeneticistas fazem. Com isso, formalizou-se o problema através de um modelo matemático, tornando possível a identificação automática dos pares homólogos de cromossomos. Outro ponto forte da utilização desse enfoque está no fato da apresentação dos resultados por ordem de semelhança, o que permitia analisar a eficiência do sistema em vários estágios de identificação.

De acordo com os resultados obtidos pelo SCACC no reconhecimento automático de cromossomos, considerando-se apenas a identificação de 1ª ordem, o índice de acerto foi de aproximadamente de 63,04% em seu 7º experimento. Embora este resultado seja relativamente baixo, equipara-se aos resultados obtidos com outros trabalhos encontrados na literatura como, por exemplo, o trabalho de JOHNSTON *et al.* (1993), que aplicou sistemas especialistas na classificação de todos os cromossomos da metáfase e também realizou seus testes com imagens de cromossomos de laboratórios distintos. Outros trabalhos tiveram resultados melhores, porém seus testes foram realizados utilizando-se pares específicos de cromossomos e/ou imagens de apenas um laboratório em questão, tais como os trabalhos de LUNDSTEEN *et al.* (1991), DENISOV e DUDKIN (1994), TODESCO (1995) e SJAHPUTERA e KELLER (1999).

Com o método de reconhecimento automático dos cromossomos homólogos elaborado, a etapa seguinte consiste na composição automática do cariótipo, ou seja, a identificação dos pares por ordem decrescente de tamanho. O método criado para esta finalidade apresentou algumas deficiências, principalmente porque na ocorrência de erros nas primeiras classes (pares), desencadeia-se uma série de outros erros na seqüência da composição do cariótipo. Houve falhas também para a classificação dos pares do grupo C, em especial nos cromossomos sexuais X. A eficiência da composição automática do cariótipo poderia ser melhorada através da classificação, primeiramente pela área, dos cromossomos pertencentes aos sete grandes grupos e, em seguida, pela identificação dos mesmos através dos padrões de bandas, índice centromérico, etc. Outra alternativa seria a busca de novas características que pudessem descrever todas as 24 classes de cromossomos, tais como a posição das bandas principais ao longo dos cromossomos, a intensidade e a largura das principais bandas seriam as soluções mais viáveis, tal como no trabalho de JOHNSTON *et al.* (1993).

Como conclusão deste primeiro propósito, ter um método de reconhecimento automático de cromossomos através de uma única cena, sem necessitar comparar com entidades específicas previamente armazenadas, como pode ser constatado, é totalmente possível, apesar da composição do cariótipo ainda apresentar falhas. Finalmente, considera-se que este propósito foi em parte solucionado, com a expectativa de que alguns aspectos possam ser controlados e outros precisam ainda serem explorados para melhorar os índices de classificação.

Para o segundo propósito que consiste na elaboração de uma nova metodologia para a detecção mais confiável da leucemia pró-mielocítica aguda (LMA), houve a necessidade de se desenvolver uma nova técnica de prolongamento das extremidades do eixo resultante do processo de esqueletização, um método eficaz que permitiu extrair com precisão as características que descreviam a doença. A detecção da LMA somente foi possível utilizando-se o par número 17, apresentando-se mais confiável do que ao utilizar o par de cromossomos número 15, explicando-se o motivo pelos quais os experimentos foram realizados apenas com o par 17. Já o uso do enfoque lógico-combinatório como método de modelagem matemática teve fundamental importância, pois possibilitou a “tradução” eficiente do conhecimento do especialista formalizando as particularidades do problema, além de dispor ao profissional os resultados por um indicador do rearranjo cromossômico. E os resultados obtidos com os experimentos, demonstraram um excelente desempenho dessa nova metodologia, com valores acima de 62,55% para as imagens testadas de células leucêmicas e abaixo de 17,73% para imagens testadas de células normais.

Da conclusão deste segundo propósito, pode-se afirmar que o SCACC, como ferramenta para auxílio na detecção da leucemia pró-mielocítica aguda, permitiu reduzir a subjetividade na análise e avaliações desta doença, disponibilizando aos citogeneticistas melhores condições para a detecção de anomalias cromossômicas com maior eficiência e confiabilidade.

Do que foi exposto neste capítulo, tendo em vista que o objetivo geral deste trabalho, tinha-se como finalidade a elaboração de uma nova metodologia visando reduzir o viés da análise em citogenética clínica, considera-se, então, que os objetivos para o reconhecimento automático dos cromossomos foram atingidos parcialmente e os objetivos para a criação de uma ferramenta computacional para a detecção da leucemia pró-mielocítica aguda foram totalmente atingidos. Portanto, conclui-se que, em seu ponto mais forte, a nova sistemática pode se tornar uma ferramenta que permite reduzir o viés da análise em citogenética clínica.

### 8.3 – Limitações do Trabalho

O trabalho de elaboração da nova sistemática para análise e classificação citogenética apresentou algumas limitações no contexto geral do modelo computacional desenvolvido, as quais são descritas a seguir:

1- O sistema apresenta limitações na rotina de rotacionamento dos objetos detectados. A rotina desenvolvida apenas calcula os valores máximos e mínimos nos eixos  $x$  e  $y$ , e apresenta os cromossomos individualizados e rotacionados em uma nova janela.

2- Na composição do cariótipo, não foi implementada a rotina para a classificação automática do cromossomo sexual Y (células de indivíduos do sexo masculino), pois este cromossomo não possui par homólogo. Haveria, para tanto, a necessidade de se identificar características que o descrevessem em sua classe, dessa forma optou-se, primeiramente, em utilizar somente imagens de células metafásicas de indivíduos normais do sexo feminino.

3- Não foram feitos estudos para a constatação da qualidade dos dispositivos de aumento (objetivas) utilizados e a comparação com outros existentes no mercado, isto é, não houve uma pesquisa da eficiência da objetiva utilizada na aquisição das imagens, utilizadas neste trabalho com outras para uma análise comparativa da qualidade das imagens quanto às correções ópticas.

4- O cálculo do índice centromérico não foi possível de ser realizado porque o método aplicado para extrair o menor diâmetro dos cromossomos apresentou muitas falhas.

5- Para a identificação do número do par, necessita-se da comparação feita com as imagens de células de indivíduos normais, com os 46 cromossomos da metáfase, não sendo possível identificá-los separadamente.

### 8.4 – Contribuições deste Trabalho

A nova sistemática desenvolvida apresenta várias contribuições para a atividade de análise e classificação citogenética, dentre as quais:

- permite tornar o processo de reconhecimento dos cromossomos mais ágil;
- em alguns casos, permite melhor análise e interpretação da imagem do cromossomo;
- um maior número de metáfases podem ser analisadas de um mesmo paciente;
- permite o armazenamento dos dados e do kariograma do paciente em um banco de dados para futuras análises.
- Possibilitam uma otimização do uso dos instrumentos ópticos (dispendiosos).

Estas vantagens são comuns a todos os sistemas de cariotipagem desenvolvidos, contudo, além destas vantagens, a nova sistemática elimina ainda o inconveniente dos profissionais terem de incluir na base de dados padrões pré-estabelecidos de imagens de cromossomos para que o sistema possa classificá-los corretamente. Este fato é decorrente ao método utilizado que não é baseado na utilização de sistemas especialistas, ou seja, o reconhecimento automático de cromossomos é realizado através de uma única cena. Além disso, o SCACC é uma ferramenta mais adaptada à realidade clínica, disponibilizando os resultados na forma com que os citogeneticistas analisam e avaliam os problemas clínicos.

Por fim e em seu ponto mais forte, o SCACC permite reduzir a subjetividade na determinação das alterações específicas que caracterizam a leucemia pró-mielocítica aguda, auxiliando os citogeneticistas na análise e na avaliação clínica mais eficientes e confiáveis, sem a necessidade da corroboração de outros profissionais para a confirmação da doença.

### **8.5 – Sugestões para Futuros Trabalhos**

Como recomendações para futuros trabalhos, é de suma importância a elaboração de um método de extração do índice centromérico. Esta característica, além de permitir a identificação dos cromossomos em determinadas classes, permite também a identificação dos braços  $p$  e  $q$ , para que se possa realizar a correta rotação dos cromossomos.

Também é de fundamental importância que o processo de composição automática do cariótipo seja realizado de acordo com a identificação dos cromossomos às suas classes (pares). Necessita-se, para tanto, que sejam identificadas e extraídas as características

próprias de cada par, que possam distingui-los uns dos outros. Para esta tarefa, a aplicação da teoria dos Testores poderia auxiliar muito na definição das características que descrevem cada classe.

O desenvolvimento de técnicas para a identificação de cromossomos que se encontram sobrepostos ou ligeiramente encostados estão sendo muito exploradas, devido estes fatores serem muito comuns nas metáfases analisadas.

E, no advento do desenvolvimento de métodos para o auxílio ao diagnóstico clínico, as anormalidades estruturais apresentam muitas dificuldades para o citogeneticista detectar. Dessa forma, a criação de novas ferramentas, como a apresentada neste trabalho para a detecção da leucemia pró-mielocítica aguda, muito contribuiria para a área de genética humana.

## REFERÊNCIAS BIBLIOGRÁFICAS<sup>1</sup>

APPLIED IMAGING. **CytoVision System and Quips Karyotyping Analysis System**. Disponível em: <<http://www.cytovision.com>> Acesso em 22 de abril de 2003.

BARBETTA, P.A. **Estatística Aplicada às Ciências Sociais**. 5. ed. rev. Florianópolis: Ed. da UFSC, 2002. 340p.

BEIGUELMAN, B. **Citogenética Humana**. Rio de Janeiro: Guanabara Koogan, 1982. 328 p.

BIOARGEM. **Cariosys: Software para Análisis de Cariotipos**. Disponível em: <<http://www.cariosys.com.ar>> Acesso em 20 de abril de 2003.

BRUSCHI, C.; TEDESCHI, F.; PUGLISI, P.P.; MARMIROLI, N. Computer-assisted Karyotyping system of banded chromosomes. **Cytogenetics and Cell Genetics**, [S.l.], v. 29, n. 1, p. 1-8, 1981.

CABRAL, R.B. **Home Page: Retificação e Análise de Características de Imagens de Nematóides - Grafo do Eixo Mediano**. Disponível em <<http://www.eps.ufsc.br/~cabral/graf.html>> Acesso em 15 de Maio 2003.

CAMPAGNOLO, J.M. **Home Page de EEL 7031 Computação Científica II: Aulas 15 e 16 - Ajuste de Curvas**. Dep. de Engenharia Elétrica – UFSC. Disponível em <<http://www.labspot.ufsc.br/~campagno/numerico.html>> Acesso em 10 de junho de 2003.

CHOI, H.I.; CHOI, S.W.; MOON, H.P. Mathematical Theory of Medial Axis Transform. **Pacific Journal of Mathematics**, [S.l.], v. 181, n. 1, 1997.

CLÁUDIO, D.M. e MARINS, J.M. **Cálculo Numérico Computacional**. São Paulo - SP: ATLAS/S.A., 1989. 464 p.

---

<sup>1</sup> ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 6023**: informação e documentação: referências – elaboração. Rio de Janeiro, 2002.

**REFERÊNCIAS BIBLIOGRÁFICAS - Continuação**

DENISOV, D.A.; DUDKIN, A.K. Model-Based Chromosome Recognition via Hypotheses Construction/Verification. **Pattern Recognition Letters**, North-Holland, n. 15, p. 299-307, 1994.

DESEILLINGNY, M.P.; STAMON, G.; SUEN, C.Y. Veinerization: A New Shape Description for Flexible Skeletonization. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [S.l.], v. 20, n. 5, Maio, 1998.

DRETS, M.E. Una Saga Citogenética: El descubrimiento de los métodos de bandeó cromosómico. Significado y proyección bio-médica. **EL SMU**, Uruguay, v. 18, n.2, p. 107-121, set. 2002.

ENSSLIN, L., MONTIBELLER NETO, G., NORONHA, S.M. **Apoio à Decisão: Metodologia para Estruturação de Problemas e Avaliação Multicritério de Alternativas**. 1ª ed., Florianópolis: Insular, 2001. 296 p.

ERRINGTON, P.A.; GRAHAM, J. Application of Artificial Neural Networks to Chromosome Classification. **Cytometry**, [S.l.], v. 14, p. 627-639, 1993.

FACON, J. **Processamento e Análise de Imagens**. Pontifícia Universidade Católica do Paraná, Apostila do Curso de Mestrado em Informática Aplicada, Curitiba – PR: 2002.

FACON, J. **Home Page de Jacques Facon – Processamento e Análise de Imagens: Algoritmo de Afinamento Zhang-Suen**. Pontifícia Universidade Católica do Paraná, Programa de Pós-Graduação em Informática Aplicada. Disponível em: <<http://www.ppgia.pucpr.br/~facon/Afinamento/AfinamentoZhangSuen.pdf>> Acesso em 11 de julho de 2003a.

FACON, J. **Home Page de Jacques Facon – Processamento e Análise de Imagens: Algoritmo de Afinamento Holt**. Pontifícia Universidade Católica do Paraná, Programa de Pós-Graduação em Informática Aplicada, Disponível em: <<http://www.ppgia.pucpr.br/~facon/Afinamento/AfinamentoHolt.pdf>> Acesso em 11 de julho de 2003b.

**REFERÊNCIAS BIBLIOGRÁFICAS - Continuação**

FDA. **Review Criteria form Assessment of Cytogenetic Analysis Using Automated and Semi-Automated Chromosome Analyzers**. U.S. Food and Drug Administration – Center for Devices and Radiological Health, Draft Document for Manufacturers, 1991.

FRANÇA, J.A.; Von WANGENHEIM, A. **Home Page de Visão Computacional, Esqueletização: a Linha Essencial**. Curso de Pós-Graduação em Ciências da Computação. Disponível em: <<http://www.inf.ufsc.br/~visao/.html>> Acesso em 07 de junho de 2003.

GONZALEZ, R.C.; WOODS, R.E. **Digital Image Processing**. USA: Addison-Wesley, 1992. 716 p.

HAYKIN, S. **Redes Neurais: princípios e práticas**. 2ª ed., Porto Alegre – RS: Bookman, 2001. 900 p.

HURET, J.L.; CHOMIENNE, C. **Home Page Atlas of Genetics and Cytogenetics in Oncology and Haematology: t(15;17)(q22;q21)**. Disponível em <<http://www.infobiogen.fr/services/chromcancer/Anomalies/t1517ID1035.html>> Acesso em 01 de junho de 2003.

ISCN (1995). **An International System for Human Cytogenetic Nomenclature**. Mitelman F (ed); S. Karger, Basel, 1995.

JOHNSTON, K.S.; TANG; ZIMMERMAN, S. Band Features as Classification Measures for G-Banded Chromosome Analysis. **Computers in Biology and Medicine**, Great Britain, v. 23, n. 2, p. 115-129, 1993.

LEICA. **Leica Chantal**. Disponível em: <<http://www.leica.com>> Acesso em 25 de abril de 2003.

LERNER, B.; GUTERMAN, H; DINSTEIN, I.; ROMEM, Y. Medial Axis Transform-Based Features and a Neural Network for Human Chromosome Classification. **Pattern Recognition**, [S.l.], v. 28, n. 11, p. 1673-1683, 1995.

**REFERÊNCIAS BIBLIOGRÁFICAS - Continuação**

LOPES, A.C. **Viés, esse ilustre desconhecido**. Psychiatry On-line Brazil. Part of the International Journal of Psychiatry - ISSN 1359 7620. Disponível em: <<http://www.polbr.med.br/arquivo/evba0499.htm>>. Acesso em: 11 de dezembro de 2002.

LOPES, L.F. **Home Page do Centro de Tratamento e Pesquisa do Hospital do Câncer. Leucemia na Infância**. Departamento de Pediatria – São Paulo. Disponível em <<http://www.hcanc.org.br/outrasinfs/ensaios/leukids.html>> Acesso em julho de 2003.

LUNDSTEEN, C.; GERDES, T.; MAAHR, J. Automated Multiple-Cell Karyotyping: a clinical feasibility study. **Journal Clinical Genetics**, [S.l.], v. 39, n. 5, p. 338-346, 1991.

LUNDSTEEN, C.; PHILIP, J. Automated Cytogenetic Analysis: accomplishments, present status and practical future possibilities. **Clinical Genetics**, [S.l.], v. 36, p. 386-391, 1989.

MARQUES FILHO, O.; VIEIRA NETO, H. **Processamento Digital de Imagens**. Rio de Janeiro – RJ: Brasport Livros e Multimídia, 1999. 406 p.

MAYALL, B.H.; TUCKER, J.D.; CHRISTENSEN, M.L.; Van VLIET, L.J.; YOUNG, I.T. Experience With The Athena Semi-Automated Karyotyping System. **Cytometry**, [S.l.], v. 11, p. 59-72, 1990.

META SYSTEMS. **Ikaros**. Disponível em: <<http://www.metasystems.de/products/ikaros/ikaros3.htm#Karyotyping>> Acesso em 18 de outubro de 2002.

MIRA JÚNIOR, J. **Identificação de Indivíduos através do Processamento Morfológico da Íris**. 2003. 116 f. Dissertação (Mestrado em Engenharia Elétrica). Universidade Federal de Santa Catarina, Florianópolis.

MOTTA JÚNIOR, M.C.M. **Home Page de Hematologia: Leucemia**. Faculdade de Medicina da EMESCAM, Vitória-ES, Disponível em <<http://ioh.medstudents.com.br/hemato7.htm>> Acesso em julho de 2003.

**REFERÊNCIAS BIBLIOGRÁFICAS - Continuação**

ORTIZ-POSADAS, M.R.; PANTALEÃO, C.H.Z.; BOSE, R. **A Mathematical Model for Classical Chromosome Identification Using the Logical Combinatory Approach.** Proceedings 25<sup>th</sup> Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Cancún, México, p. 1342-1345, 2003.

ORTIZ-POSADAS, M.R.; VEGA-ALVARADO, L.; JIMÉNEZ-JACINTO, V.; LAZO-CORTÉS, M. **El concepto de analogía en medicina. Una función de semejanza para pacientes con fisura de paladar.** III Taller Iberoamericano de Reconocimiento de Patrones, D. F. México, p. 247-255, 1998.

PACIORNIK, S. **Home Page do Laboratório de Processamento Digital de Imagens. Curso de Pós-Graduação em Ciência dos Materiais e Metalurgia, MET2431.** Pontifícia Universidade Católica do Rio de Janeiro. Disponível em <<http://www.dcomm.puc-rio.br/lpdi/estrutura/cursos/cursos.html>> Acesso em 07 de junho de 2003.

PIPER, J. Classification of chromosomes constrained by expected class size. **Pattern Recognition Letters**, North-Holland, v. 4, p. 391-395, 1986.

POPESCU, M.; GARDER, P.; KELLER, J.; KLEIN, C; STANLEY, J.; CALDWELL, C. Automatic Karyotyping of Metaphase Cells With Overlapping Chromosomes. **Computers in Biology and Medicine**, [S.l.], v. 29, n. 1, jan., 1999.

REMY, E.; THIEL, E. Medial Axis for Chamfer Distances: Computing Look-Up Tables and Neighbourhoods in 2D or 3D. **Pattern Recognition Letters**, [S.l.], v. 23, p. 649-661, 2002.

RUGGIERO, M.A.G; LOPES, V.L.R. **Cálculo Numérico: Aspectos Teóricos e Computacionais.** São Paulo - SP: McGraw-Hill, 1988, 295 p.

SHULCLOPER, J.R.; ARENAS, A.G. e TRINIDAD, J.F.M. **Enfoque Lógico Combinatorio al Reconocimiento de Patrones: I - Selección de Variables y Clasificación Supervisionada.** Série Avances em Reconocimiento de Patrones: Teoría y Práctica, Instituto Politécnico Nacional – Centro de Investigación em Computación, v.1, México – DF, 1997.

**REFERÊNCIAS BIBLIOGRÁFICAS - Continuação**

SJAHPUTERA, O.; KELLER, J.M. Evolution of a Fuzzy Rule-Based System for Automatic Chromosome Recognition. In: IEEE INTERNATIONAL FUZZY SYSTEMS CONFERENCE, 1999, Seoul – Korea, **Proceedings**, August 22-25, 1999, p. 129-134.

STANLEY, R.J.; KELLER, J.M.; GADER, P.; CALDWELL, C.W. Data-Driven Homologue Matching for Chromosome Identification. **IEEE Transactions on Medical Imaging**, [S.l.], v. 17, n. 3, June, 1998.

TAKAHASHI, A. O Estado das Webcams USB, **Revista PCs: memórias DDR**. São Paulo: Lucano, n. 24, p. 39-46, 2001.

THOMPSON, M. W.; McINNES, R. R.; WILLARD, H. F. **Thompson & Thompson: genetics in medicine**. 5. ed. Philadelphia: Saunders, 1991. 500 p.

TODESCO, J.L. **Reconhecimento de Padrões Usando Rede Neuronal Artificial com uma Função de Base Radial: uma aplicação na classificação de cromossomos humanos**. 1995. 105 f. Tese (Doutorado em Engenharia de Produção), Universidade Federal de Santa Catarina, Florianópolis.

TSO, M.; KLEINSCHMIDT, P; MITTERREITER, I.; GRAHAM, J. An Efficient Transportation Algorithm for Automatic Chromosome Karyotyping. **Pattern Recognition Letters**, North-Holland, v. 12, p. 117-126, 1991.

Van VLIET, L.J.; YOUNG, I.T.; MAYALL, B.H. The Athena Semi-Automated Karyotyping System. **Cytometry**, [S.l.], v. 11, p. 51-58, 1990.

VOGEL, F.; MOTULSKY, A. G. **Human Genetics: problems and approaches**. 3. ed. New York : Springer, 1997. 851 p.